12-2013

L1 methods for shrinkage and correlation

Jie Shen
Clemson University, jieshen.discover@gmail.com

L_1 METHODS FOR SHRINKAGE AND CORRELATION

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Jie Shen
December 2013

Accepted by:
Dr. Colin Gallagher, Committee Chair
Dr. Christopher McMahan
Dr. Robert Lund
Dr. Peter Kiessler
Dr. Patrick Gerard

Abstract

This dissertation explored the idea of L_1 norm in solving two statistical problems including multiple linear regression and diagnostic checking in time series. In recent years L_1 shrinkage methods have become popular in linear regression as they can achieve simultaneous variable selection and parameter estimation. Their objective functions containing a least squares term and an L_1 penalty term which can produce sparse solutions (Fan and Li, 2001). Least absolute shrinkage and selection operator (Lasso) was the first L_1 penalized method proposed and has been widely used in practice. But the Lasso estimator has noticeable bias and is inconsistent for variable selection. Zou (2006) proposed adaptive Lasso and proved its oracle properties under some regularity conditions. We investigate the performance of adaptive Lasso by applying it to the problem of multiple undocumented change-point detection in climate. Artificial factors such as relocation of weather stations, recalibration of measurement instruments and city growth can cause abrupt mean shifts in historical temperature data. These changes do not reflect the true atmospheric evolution and unfortunately are often undocumented due to various reasons. It is imperative to locate the occurrence of these abrupt mean shifts so that raw data can be adjusted to only display the true atmosphere evolution. We have built a special linear model which accounts for long-term temperature change (global warming) by linear trend and is featured by $p = n$ (the number of variables equals the number of observations). We apply adaptive Lasso to estimate the underlying sparse model and allow the trend parameter to be unpenalized in the objective function. Bayesian Information Criterion (BIC) and the CM criterion (Caussinus and Mestre, 2004) are used to select the finalized model. Multivariate t simultaneous confidence intervals can post-select the change-points detected by adaptive Lasso to attenuate overestimation.

Considering that the oracle properties of adaptive Lasso are obtained under the condition of linear independence between predictor variables, adaptive Lasso should be used with caution since

it is not uncommon for real data sets to have multicollinearity. Zou and Hastie (2005) proposed elastic net whose objective function involves both L_1 and L_2 penalties and claimed its superiority over Lasso in prediction. This procedure can identify a sparse model due to the L_1 penalty and can tackle multicollinearity due to the L_2 penalty. Although Lasso and elastic net are favored over ordinary least squares and ridge regression because of their functionality of variable selection, in presence of multicollinearity ridge regression can outperform both Lasso and elastic net in prediction. The salient point is that no regression method dominates in all cases (Fan and Li, 2001, Zou, 2006, Zou and Hastie, 2005). One major flaw of both Lasso and elastic net is the unnecessary bias brought by constraining all parameters to be penalized by the same norm. In this dissertation we propose a general and flexible framework for variable selection and estimation in linear regression. Our objective function automatically allows each parameter to be unpenalized, penalized by L_1, L_2 or both norms based on parameter significance and variable correlation. The resulting estimator not only can identify the correct set of significant variables with a large probability but also has smaller bias for nonzero parameters. Our procedure is a combinatorial optimization problem which can be solved by exhaustive search or genetic algorithm (as a surrogate to computation time). Aimed at a descriptive model, BIC is chosen as the model selection criterion.

Another application of the L_1 norm considered in this dissertation is portmanteau tests in time series. The first step in time series regression is to determine if significant serial correlation is present. If initial investigations indicate significant serial correlation, the second step is to fit an autoregressive moving average (ARMA) process to parameterize the correlation function. Portmanteau tests are commonly used to detect serial correlation or assess the goodness-of-fit of the ARMA model in these two steps. For small samples the commonly employed Ljung-Box portmanteau test (Ljung and Box, 1978) can have low power. It is beneficial to have a more powerful small sample test for detecting significant correlation. We develop such a test by considering the Cauchy estimator of correlation. While the usual sample correlation is estimated through L_2 norm, the Cauchy estimator is based on L_1 norm. Asymptotic properties of the test statistic are obtained. The test compares very favorably with the Box-Pierce/Ljung-Box statistics in detecting autoregressive alternatives.

Dedication

To my parents.

Acknowledgments

I would like to express my heartfelt gratitude to Dr. Colin Gallagher, my PhD advisor, whose expert guidance greatly helps the completion of this dissertation. He has been playing an indispensable role in that I grow from a beginner to a confident and passionate scientific researcher. While I tend to look at the downsides if my research comes into dilemma he is always optimistic and can help me find a solution. I appreciate all his time, ideas, good advices and enthusiasm.

I am truly grateful to Dr. Chris McMahan, Dr. Robert Lund, Dr. Peter Kiessler, and Dr. Patrick Gerard for being my committee members. Special thanks goes to Dr. Chris McMahan for his valuable suggestions on my research. Every time I ask him a question puzzling me for days he can always give a quick and straight answer and explains the details very patiently. He also encouraged me to never give up on problems just because they appear complicate. Sincere thanks also goes to Dr. Robert Lund. I enjoy reading all his well-written change-point papers, which provide good and complete resources for learning the origin, background and techniques of change-points. I also acknowledge Dr. Peter Kiessler for being so nice with students. I enjoyed all the three courses that I took with him. His teaching philosophy relevant to instructing materials and drawing students' attention is my favorite.

I would like to extend my special thanks to Department of Mathematical Science at Clemson University for the financial assistance during my PhD study. I greatly value the teaching experience with my undergraduate students who have been so kind, sincere and friendly to me. I am also very grateful to the Palmetto Cluster at Clemson University. Without it I cannot imagine how long it takes for my simulations to finish.

Furthermore, I am very grateful to all my friends for their care and support. I am particularly thankful to Ms. Shuai Li and Mr. Hao An for being such good friends during the past twelve years. Their friendship is a treasure to me. I also acknowledge Ms. Yu Meng, my roommate, for being a

pleasant friend during her stay at Clemson and continuous love after graduation. Wish her family and her the best.

Finally, I would like to thank my parents, aunt, uncle, sister and my lovely niece. Without their love and support, this dissertation would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Change-Point Analysis

1.1.1 Background

Retrospective change-point analysis is an active area of statistics. Observations measured in order may be divided into subsamples while all subsamples are collected from the same distribution family but with different parameters. Change-points occur at unknown times when the parameters before and after them differ. They appear under a variety of synonyms across various scientific fields such as segmentation, structural breaks/changes, regime shifts and detecting disorder, etc. Change-points analysis can only be performed once all the data are collected and provides insight into historical data.

A well-credited article introducing the problem of change-points detection is Page (1954). One motivating example was given in industry quality control, where change-point detection is used for detecting a deterioration in quality measures during manufacturing process. A detected deterioration may lead to suspension of production and a machine reset. Change-point analysis also has various applications in many other areas of scientific research. In speech recognition, automatic change-point analysis is needed for detecting the change of state from noise to speech or from speech to noise (Tahmasbi and Rezaei, 2008). In finance, it is important to detect sudden shifts in volatility of stock market returns (Aggarwal et al., 1999). In computer science, efficient algorithms of change-point analysis have been applied to detect network intrusion in traffic data to avoid server

saturation and malicious attacks (Lévy-Leduc and Roueff, 2009). In geophysical sciences, it is of vital importance to detect sudden shifts in the structure of a seismogram. In biology and medicine the genomic profiles of a group of cancer patients may share the same sudden changes at precise places which can point to a direction of cancer treatment (Bleakley and Vert, 2011). In the social sciences, researchers are interested in whether the enforcement of seat belt legislation indeed reduces the mean number of monthly deaths and serious injuries as desired (Davis et al., 2006). In climatology, sudden mean shifts in instrumental climate data may not be due to real atmospheric evolution but artificial factors such as city growth or station relocation, which need to be isolated by change-point analysis (Caussinus and Lyazrhi, 1997, Menne and Williams, 2005). Visual detection of those sudden changes is only tenable in some rare cases and significance of the suspected change-points should still be checked. Automatic detection can save labor costs and is necessary under many situations such as quality control and network intrusion.

Change-point analysis can be used to detect shifts in the mean, variance, quantile, etc. The majority of research efforts have been put into detecting shifts in the mean. The typical questions to ask in a change-point analysis are: (1) Has there been a sudden shift? (2) How many change-points are there? (3) Where did they occur? (4) How confident are we about that they are true change-points?

Change-point detection techniques can vary with the nature of the application. For example, a network intrusion is detected when a sudden mean shift occurs in the sum of received SYN packets by all the destination IP addresses which consists of a multi-dimensional times series. Thus the task of detecting network intrusion requires change-point analysis techniques for multi-dimensional signals. A similar situation in biological application is to detect sudden changes in genomic profiles of a group of cancer patients with genomic profiles viewed as multi-dimensional signals). In speech recognition, a change-point in a noisy speech signal refers to the occurrence of a state change from speech to noise or noise to speech. A noisy speech signal can be modeled by a GARCH(p,q) process. Hence change-point analysis for detecting shifts in variance or GARCH processes are needed. If the task is to detect changes in the number of tropical cyclones, techniques for discrete or categorical data are needed (Robbins et al., 2011b).

The frequency of change-point analysis being implemented also varies with applications and depends on the speed of the data generating mechanism. An annual company report might include change-point analysis to detect improvements in inventory returns. A manager of a customer service

department may implement change-point analysis to look for changes in customer complaints on a monthly basis. In product control, the analysis can be implemented once a week to monitor deteriorations in manufacturing. In contrast, the analysis should be done more frequently in stock market where very large amount of data are generated in a day.

1.1.2 Change-point Detection in Climate Data

Our work will focus on undocumented change-point detection in historical climate data, which has been an important and active research area for decades. Many long historical instrument records of temperature greatly benefit the research on atmospheric evolution. Influences on secular temperature data mainly fall into three categories: (1) the local-environmental, such as city growth; (2) the observational, such as station relocation and instrumental recalibration; (3) the large-scale truly climatic (Mitchell, 1953). Climatologists are devoted to monitor temperature changes corresponding to the third category through instrumental records. Unfortunately, artificial factors from the first two categories usually cause sudden mean shifts in those records which are definitely no reflection of real atmospheric changes. Mitchell (1953) showed that United States temperature data over a century are subjected to an average of six station relocation and instrumentation changes. City growth is also a well-known cause for local temperature increases, which again is not a reflection of atmospheric changes. Although some of these change-point times are documented in metadata, others are unknown for a variety of reasons. Climate research frequently depends on "continuity of the measurement process". Direct climatic analysis on the raw data disturbed by artificial factors can generate misleading conclusions about atmospheric evolution. Thus, it is necessary to detect and correct the sudden mean shifts before any valid conclusions are made.

Climate data have their own features and require some special techniques for change-point detection. Concerns include correlation and periodicity. The independent and identically distributed (IID) assumption may be tenable for yearly temperature data while correlation and periodicity effects are typical when daily/monthly temperature series are considered. Another concern is that temperature series can show an increasing long-term trend due to the global warming effect. The trend is usually modeled as linear in time for simplicity. Unlike in some applications where a piecewise constant approximation to the data serves the purpose, an annal temperature series may have to be approximated by a piecewise linear function under the IID assumption. Although the trend has important meaning in practice as reflection of secular atmospheric changes, it is not directly related

to change-points detection hence is a nuisance parameter. Additionally linear trend estimates are trustworthy only when the series is homogeneous in time (in absence of change-points) (Easterling and Peterson, 1995). However, the trend must be accounted for in the process of change-point detection. Otherwise, the trend may diminish the significance of some change-points. In summary, to detect change-point in a temperature series with trend one will assume a homogeneous linear trend first (Wang, 2003). Trend estimates will be obtained later from the readjusted data which is change-point free. One more special feature of climate data is that they can have reference series from neighboring cities which share similar trends and periodicity and fortunately happen to be change-point free. Researchers can diminish the trend and periodic effects by differencing the target and reference series and apply change-points analysis to the differences. A good reference series can enhance the change-point detection. However, good reference series are not available in some situations; they may have unknown change-points and complicated trends themselves. Hence it is quite possible to bring in additional change-points by using such inappropriate references series.

Under the at most one change-point (AMOC) setting, hypothesis tests have been proposed to test the null hypothesis of no changes (Caussinus and Mestre, 2004, Menne and Williams, 2005, Menne and Williams, 2009). As the change-point community in climate thrives, literature reviews of available hypothesis tests are also provided for ease of comparison (Lund et al., 2007, Rodionov, 2005). Methods for IID data include CUSUM tests (Macneill, 1974), likelihood ratio tests (Csörgö and Horváth, 1997), etc. Non-parametric methods include Mann-Whitney U-test (Mauget, 2003), Mann-Kendall test (Goossens and Berger, 1987) and Pettitt test (Pettitt, 1979). These nonparametric tests can only be used after the trend and periodic effect are diminished by homogeneous references series. Assuming IID error with linear trend effect Hinkley (1969, 1971) proposed a two-phrase regression (TPR) method which was recently revised by Lund and Reeves (2002) and Wang (2003). Another method is the generalized likelihood ratio test (GLRT) proposed by Kim and Siegmund (1989). For correlated data Robbins et al. (2011a) proposed an adjusted CUSUM test. For data with both correlation and periodicity one can use the test proposed by Lund et al. (2007).

For the problem of multiple change-point detection, hierarchical splitting algorithms and successive AMOC tests are typically used. The choice of tests is based on the assumptions. This hierarchical strategy is well-known for introducing too many false alarms (Lavielle and Teyssiere, 2006, Scott and Knott, 1974, Vostrikova, 1981). A refinement of the hierarchical splitting algorithm is a possible re-merging step after a new change-point is introduced in a splitting step. But the

refined method still cannot produce the optimal change-points reliably (Hawkins, 2001). Model selection methods are also employed to solve the multiple change-point problem. For IID and Gaussian data without trend, a least absolute shrinkage and selection operator (Lasso) method using a total variation penalty was proposed by Harchaoui and Levy-Leduc (2010). To deal with autocorrelation Davis et al. (2006) proposed an minimum description length (MDL) approach to approximate time series by piecewise AR processes. Lu et al. (2010) extended their work to allow for both autocorrelation and periodicity. They also assume a time-homogeneous linear trend for simplicity. In Chapter 2 we will use adaptive Lasso Zou (2006) to tackle multiple change-point detection in annual temperature series with linear trend. Our method is fast and easy to implement. Although we assume IID data for simplicity, the validity of this assumption will be checked in model diagnosis.

1.2 Shrinkage Methods in Linear Regression

Statistical modeling describes how predictor variables relate to response variables. It is usually a good and informative alternative when using a deterministic function to describe the relationship is impossible or too complicated. Linear regression models are the first type of statistical models that have been studied rigorously. Our work will focus on where this type of models is appropriate as the underlying data generating mechanism.

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the response variable and $\mathbf{x_i} = [x_{i1}, \ldots, x_{in}]^T$, $i = 1, \ldots, p$ be the predictor variables. A linear regression model assumes that the response variable is linear in the predictor variables, that is, $\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is the error vector and $\beta_1, \cdots, \beta_p, \sigma^2$ are unknown parameters. The unknown parameters will be estimated from data. The linear assumption has been be criticized for being unrealistic in many settings. Standard goodness-of-fit tests do not help confirm its validity because they often fail to reject the null hypothesis of linearity until the non-linearity becomes extreme (Bickel et al., 2006, Breiman, 2001). Although cautions and critics against their misuse have been around for years, linear regression models are extensively involved in many fields of practical applications mainly because of their simple structure, interpretability of the **x**-**y** relationship and easily attainable statistical properties of estimates compared to non-linear models.

Statistical modeling aims at either prediction or interpretability based on applications. In

practice a large number of predictors usually have to be introduced at the initial stage of modeling for two reasons: (1) to attenuate possible modeling bias; (2) the number of predictors is still large even after excluding some evidently useless predictors by domain knowledge. Only the predictors relevant to $\mathbf{y}$ are beneficial in obtaining better interpretability and prediction. Irrelevant and redundant predictors in the model cannot not contribute useful information but just adding in noise leading to poor interpretation and prediction. The well-known example is ordinary least squares (OLS) defined as:

$$\hat{\boldsymbol{\beta}}(\text{ols}) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_i \mathbf{x_i} \beta_i \right\|^2 . \tag{1.1}$$

OLS keeps all the given predictor variables in the model and can cause "overfitting", that is, providing good fit to the training data but poor prediction on test data. This is mainly due to the large variance of parameter estimates when the data have high multicollinearity (some $\mathbf{x}$'s are linearly related). Ridge regression can provide parameter estimates with small variance for data with multicollinearity, hence can do a better job in prediction. Ridge regression minimize an objective function containing a least squares error term and a L_2 penalty term. They are defined as:

$$\hat{\boldsymbol{\beta}}(\text{ridge}) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_i \mathbf{x_i} \beta_i \right\|^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i|^2 , \tag{1.2}$$

where $\lambda_2 \geq 0$ is the tuning parameter that controls the amount of shrinkage. The subscript "2" refers to the L_2 norm appearing in the penalty term. With all variables being kept in the model, interpretability remains an issue for ridge regression. For better interpretability it is desirable to perform *variable selection*: select only those significant variables into the finalized model, i.e., a subset of the given predictor variables. For clarity it is necessary to point out that there are some non-statistical criteria of variable selection, such as measurement cost, size of memory storage, ease of use, etc. For instance, measurement costs may be extremely high for some variable but more economical for another variable and both variables provide similar information for $\mathbf{y}$. Mathematically, it does not matter which variable is to be selected. The practitioner may prefer or have to use one particular variable. Our work will not consider those non-statistical criteria, but use model selection criteria involving interpretability and prediction.

Best-subset selection (Beale et al., 1967, Garside, 1971) is a natural way of selecting a subset of variables, where all possible subsets of p predictor variables are considered and ideally the one with the best value of some model selection criterion is chosen as the finalized model.

If p predictor variables are considered, there are 2^p possible regressions. This method becomes computationally expensive when p gets large. Stepwise selection (Efroymson, 1965), namely, forward stepwise selection and backward elimination, is proposed as a computational surrogate to best-subset selection. Take forward stepwise selection for example. It starts with the null model (no variables) and adds one single variable that best improves the model each time. Variables already in the model may be dropped if their significance becomes lower than a thresholding value whenever one new variable enters the model. The procedure continues until no more variables can improve the model. Backward elimination implements similarly but in the reverse direction, that is, starting with the full model including all variables and dropping one variable each time until no other variables can be dropped. Although those subset selection methods provide sparse and interpretable models, they are criticized for being extremely variable due to the stochastic error in the sense that variables are either retained or dropped from the model. A little disturbance in the data can result in selecting a very different model (Fan and Li, 2001, Tibshirani, 1996, Zou, 2006).

In contrary to the discreteness of subset selection, continuous shrinkage/penalized methods were proposed and became popular rapidly. The first method is Lasso which was proposed by Tibshirani (1996). To allow all predictor variables to be compared on the same footing, we suppose that each $\mathbf{x_i}$ is centered and standardized such that $\sum_j x_{ij} = 0$ and $\sum_j x_{ij}^2 = 1$ for $i = 1, \ldots, p$. Also let $\mathbf{y}$ be centered such that $\sum_j y_j = 0$. Lasso is the solution that minimizes the following convex objective function:

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_i \mathbf{x_i} \beta_i \right\|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \,, \tag{1.3}$$

where λ is a nonnegative regularization parameter and the second term is referred to as the L_1 penalty. Lasso seems to have started a revolution in statistical modeling by first applying the L_1 norm. Lasso can provide a sparse model because it continuously shrinks some parameters to be exactly zero if λ is sufficiently large, more intrinsically, because the L_1 norm is singular at the origin which causes some β_i's to be 0 when the objective function is minimized. Although Lasso is popular in many applications, it has been proved to not satisfy the oracle properties of a good procedure (Fan and Li, 2001). First of all, the lasso shrinkage produces biased estimates for significant coefficients. Second, the optimal λ for prediction gives inconsistent variable selection results; in fact, many noise features are included in the selected predictive model (Meinshausen and Bhlmann, 2006). Despite the problems Lasso remains a popular procedure in practice for simultaneous variable selection and

parameter estimation. The question raised for statisticians was whether the L_1 penalty could produce an oracle procedure. Zou (2006) answered this question by proposing adaptive Lasso which was proved to enjoy oracle properties under some assumptions including linear independence. However, linear independence is not realistic in many situations because multicollinearity is often encountered given many predictors. Adaptive Lasso also has to start with some initial estimates from OLS, marginal regression or Lasso. Zou and Hastie (2005) proposed elastic net which inherits good properties of both ridge regression and Lasso. Elastic net estimates are defined as below:

$$\hat{\boldsymbol{\beta}}((\text{naïve enet})) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_i \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i|^2 + \lambda \sum_{i=1}^{p} |\beta_j| \,. \tag{1.4}$$

This procedure can select a sparse model due to the L_1 penalty and is claimed to provide better prediction than Lasso in presence of multicollinearity due to the L_2 penalty. While Lasso can only select one arbitrary variable among a group of highly correlated variables, elastic net can select the whole group into the finalized model. One potential application is in the setting of microarray genomic profiles where genes sharing the same pathway are correlated and are expected to be selected together.

In practice when a practitioner fits a model he may think that a few variables are relevant and important based on domain knowledge and should always be kept in the model. Those predictors are then left unpenalized (or called "mandatory variables" in biostatistics), otherwise are at the risk of being shrank to 0. The objective functions of shrinkage methods can easily be modified to allow for unpenalized variables, i.e. omit them in the penalty.

The algorithms available for solving the Lasso problem dating back to 1996 (when Lasso was proposed) include quadratic programming. It is however inefficient in partitioning the L_1 penalty into 2^p constraints. An efficient algorithm named least angle regression (LARS) was proposed in 2003 and can be easily modified to solve the Lasso problem. LARS remained the state-of-the-art until another more efficient algorithm named path-wise coordinate descent was credited (Friedman et al., 2008a, Wu and Lange, 2008). Both algorithms can also solve elastic net. But they are not the the only options for solving Lasso. For a detailed description of a variety of algorithms see Schmidt (2005) and `http://www.di.ens.fr/~mschmidt/Software/lasso.html` (the latter includes 15 strategies).

For comparison between the estimators of OLS, ridge regression, Lasso and elastic net, let

us assume orthogonality of $\mathbf{x}_i$'s. Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_i(\text{ridge}) &= \frac{\hat{\boldsymbol{\beta}}_i(\text{ols})}{1+\lambda_2}, \\
\hat{\boldsymbol{\beta}}_i(\text{lasso}) &= \text{sign}(\hat{\boldsymbol{\beta}}_i(\text{ols}))(\hat{\boldsymbol{\beta}}_i(\text{ols}) - \lambda/2)_+, \\
\hat{\boldsymbol{\beta}}_i(\text{naïve enet}) &= \text{sign}(\hat{\boldsymbol{\beta}}_i(\text{ols}))\frac{(\hat{\boldsymbol{\beta}}_i(\text{ols}) - \lambda/2)_+}{1+\lambda_2},
\end{aligned} \tag{1.5}$$

where $\hat{\boldsymbol{\beta}}(\text{ols}) = \mathbf{X}^\mathbf{T}\mathbf{y}$ and $\hat{\boldsymbol{\beta}}(\text{lasso})$ is called a "soft-threshold" estimator with

$$z_+ = \begin{cases} z & \text{if } z \geq 0, \\ 0 & \text{if } z < 0. \end{cases}$$

It is easy to see that the ridge estimator is a constant scaled OLS estimator, the Lasso estimator is a constant translated OLS estimator and elastic net can be viewed as a two-stage procedure: a ridge-type direct shrinkage followed by a lasso-type thresholding (Tibshirani, 1996, Zou and Hastie, 2005). It is known that the OLS estimator is unbiased but has large variance in presence of multicollinearity and the estimators of ridge regression, Lasso or elastic net are all biased and aim at prediction accuracy by a bias-variance trade-off. The Lasso and elastic net estimators do not have a clear form as in (1.5) in presence of correlated variables, hence comparison becomes difficult. The salient point is that no methods dominate others in all settings. In Chapter 3 we propose a flexible and general procedure which works as least well as OLS, ridge regression, Lasso or elastic net and can often outperform them.

1.2.1 Other Lasso Extensions

Besides adaptive Lasso and elastic net, Lasso also has some other variants motivated by problem-specific situations. Yuan and Lin (2006) proposed a group Lasso to solve the problem of selecting a group of variables for accurate prediction. Once a variable is selected into the model, the whole group which it belongs to are selected. The multi-factor analysis-of-variance problem is one important example of its practical applications. As an improvement of group Lasso, Simon et al. (2013) proposed a sparse group Lasso which involves two penalties and can achieve sparsity both on a group and within group level. Fused Lasso (Tibshirani et al., 2005) was proposed to encourage sparsity and smoothness by penalizing the L_1-norm of both the parameters and their

successive differences. Thus it encourages sparsity of the parameters and their differences, i.e. local constancy of the parameter profile. Friedman et al. (2008b) proposed graphic Lasso to estimate a sparse graph by a L_1 penalty applied to the inverse covariance matrix. Additionally, shrinkage methods have also been generalized to logistic regression (Meier et al., 2008), multinomial regression, Poisson regression, Cox model (Simon et al., 2011), penalized linear discrimination analysis (Witten and Tibshirani, 2011), sparse principle component analysis (Zou et al., 2006), etc.

Variable selection and shrinkage methods have also been extended to high dimensional problems where the number of variables p is much larger than the number of observations n. For example, a typical task in gene selection problem is to separate healthy patients from cancer patients based on their gene expression profiles. The number of predictor variables could range from $6,000$ to $60,000$ while the number of patients available from most public sources is fewer than 100. Lasso and elastic net still work in high-dimensional situations. But Lasso can only select at most n variables into the model, which is not adequate for the gene selection problem. Elastic net can select more than n variables due to its utility of an augmented matrix (Zou and Hastie, 2005).

1.3 Genetic Algorithm

Genetic algorithm (GA) is a stochastic search that can be applied to a variety of combinatorial optimization problems (COPs). Given a COP, such methods as enumeration and exhaustive search guarantee to find the best solution but sometimes we can not afford to wait, while GA search may quickly find a near optimal solution. A near optimal solution may work fairly well for most applications and may be preferable to a time-consuming optimal solution. So far GA has been successfully applied in many real world problems, such as antenna design, drug design, chemical classification, electronic circuits, satellite design, cell phone factory tuning, etc. See Goodman (2009) presented in *the 2009 World Summit on Genetic and Evolutionary Computation, Shanghai, China*. One application of GA in variable selection is to solve the best-subset regression problem (Wasserman and Sudjianto, 1994).

Glover and Kochenberger (2003) gave a good review about GA regarding its history and methods in detail. The emergence of GA can date back to the 1960s when Rechenberg (1973), Schwefel (1977) from Germany and Bremermann, Fogel and others from USA implemented the idea of mutation and selection separately. Then John Holland published his book named *Adaptation in*

Natural and Artificial Systems (Holland, 1975), which first introduced crossover to GA in addition to selection and mutation. Again in 1975, Holland's graduate student, Ken Dejong (Jong, 1975), finished his doctoral thesis which was also viewed important in GA development as the first to provide a thorough treatment of GA in optimization problems. The evolutionary computing of GA finally took off in 1980s thanks to the advance of computer science.

GA was originally motivated by the "survival of the fittest" concept in biology: off-springs with good features adaptable to the environment survive. In GA, similar to genetic crossover and mutation, strings or "chromosomes" with high fitness are preserved in the next generation and are used to breed off-spring. A good solution is identified as the particular chromosome with highest fitness when GA is finished. Specifically, the main parts of GA algorithm include initialized population, mutation, crossover, new generation and stopping rule.

Initialized Population: An initial population of chromosomes of size l are randomly selected. The selected population is a subset of the full combinatorial space. Intuitively, a larger population size can decelerate the algorithm significantly; a smaller population may not be able to represent the whole search space and to maintain population diversity. But an optimal population size was hard to determine (some suggested $O(log(l))$).

Crossover: Two chromosomes are randomly selected from the population and mate to breed an off-spring. Crossover methods can be 1-point (1X), 2-point (2X), uniform (UX), nonlinear. In 1X crossover, only one crossover point is selected, an offspring copies genes from one parent chromosome up to this point and the other after this point. In 2X crossover, genes between two crossover points are exchanged between two parent chromosomes and one resulting chromosome is selected as the off-spring. In UX crossover each pair of alleles on parent chromosomes are swapped with a Bernoulli probability. Empirical experience has shown that 1X is too inflexible and not optimal hence 2X and UX are suggested. Adaptive crossover rates rather than a fixed value were also proposed (Booker, 1987).

Mutation: Each gene on a chromosome mutates with a probability from standard uniform distribution. The gene would mutate if its mutation probability is less than a pre-specified mutation rate, which is a small thresholding value and typically taken to be $1/l$ so that on average there is one mutation per chromosome. This mutation strategy may be inefficient if population size and chromosome size are large. An alternative is to draw a random number m from a Poisson(λ) distribution with $\lambda = 1/l$, then m genes are selected without replacement to mutate (Bremermann

et al., 1966) from a chromosome. Other adaptive mutation rates are also seen in some applications (Fogarty, 1989, Reeves, 1995). If the algorithm is trapped in a local optimum, mutation can help it get out.

New Generation: Methods of selecting chromosomes for the next generation include Roulette Wheel selection, tournament and ranking selection. Among them, ranking selection is the most popular. In each generation, a new population is created using *elitism* and *population overlaps* (Jong, 1975). Elitism will preserve a percentage of chromosomes with top fitness in the next generation. The rest of the new generation consist of off-springs from crossover between elitism chromosomes. All chromosomes are allowed to mutate. The percentage of elitism chosen can vary and typical values are 20% and 50%.

Stopping Rule: GA could continue to search forever in principle. Termination criteria include iteration times, computer clock time, estimate closeness and diversity measures.

Although GA has been successfully applied to many practical applications, the relevant theory of GA is debatable. Some claim that mutation is only to preserve population diversity and subordinate to crossover, while others like Holland advocate for crossover. As GA may be trapped to local optimum, it is advisable to run several times and compare different solutions. In Chapter 3 we will use GA to solve our COP to save computation time.

1.4 Portmanteau Tests in Times Series

In time series autoregressive moving average (ARMA) models (Jenkins and Box, 1970) are applied to describe the observations ordered by time and forecast future values. They are defined as

$$x_t - \phi_1 x_{t-1} \cdots - \phi_p x_{t-p} = z_t - \theta_1 z_{t-1} \cdots - \theta_q z_{t-q}, \quad t = 1, \cdots, n, \tag{1.6}$$

where x_t is the observed IID series, z_t is a random process with mean 0 and constant variance σ^2 and $\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, \sigma^2$ are unknown parameters. ARMA(p, q) models incorporate the natural idea that observations close together tend to be more related than those further away. This type of model has been widely applied in climate, economics, finance, social science and signal processing.

Effective fitting of an ARMA model mainly involves the iterative use of three steps: identification, estimation and diagnostic checking (Box and Pierce, 1970). Specifically, the first step is to remove trend and seasonality if applicable by differencing x_t to identify a stationary time series.

Diagnostic tests (portmanteau tests) are used to check for autocorrelations in the stationary process together with sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. They play an important role in determining whether we should use a sophisticated ARMA model. In presence of autocorrelation, an ARMA(p, q) model will be fit to the stationary time series and estimated where p and q can be selected based on the ACF and PACF. The model is checked for adequacy by studying the residuals. If the model fits the data adequately, the residuals should behave as random noise. Zero autocorrelation is one fundamental feature of random noise. Thus it is natural to check the ACF of the residuals and diagnostic tests (portmanteau tests) are used again. Significant nonzero autocorrelation in the residuals leads to careful reexamination of the model fit and iterative use of the above three steps. One can see that the diagnostic tests can either be applied to the the original x_t sequence for serial correlation or to the residuals to check model adequacy. Denote $\rho_k = E[z_t z_{t-k}]/E[z_t^2]$. Denote the residuals from the fitted ARMA(p, q) model $\hat{z}_1, \cdots, \hat{z}_n$. An estimator of ρ_k, the autocorrelations of the residuals, is given by $\hat{\rho}_k = \sum_{t=k+1}^{n} \hat{z}_t \hat{z}_{t-k} / \sum_{t=1}^{n} \hat{z}_t^2$, $k = 1, \cdots, m$. Several portmanteau tests have been proposed based on $\hat{\rho}_k$ to help check adequacy of the model fitting in addition to the ACF and PACF plots. Typically, we are interested in the hypothesis test

$$H_0 : \text{All } \rho_k\text{'s equal to } 0 \quad \text{vs} \quad H_\alpha : \text{At least one } \rho_k \text{ is not } 0, \;\; k = 1, \cdots, m. \tag{1.7}$$

Instead of testing the significance of one single ρ_k, portmanteau tests access the overall significance of the first m autocorrelations and are named for this reason. The limiting distribution of $\rho = (\rho_1, \cdots, \rho_m)$ is proved to be a multivariate normal distribution with mean zero and a diagonal covariance matrix with diagonal elements $\text{var}(\rho_k) = (n-k)/(n(n+2))$ (Anderson, 1942, Anderson and Walker, 1964). Box and Pierce (1970) noted that the quantity

$$\tilde{Q}(r) = n(n+2) \sum_{k=1}^{m} \frac{\rho_k^2}{n-k} \tag{1.8}$$

converges to a χ^2 distribution with degrees of freedom m for large n. For further simplicity, they suggested that

$$Q(r) = n \sum_{k=1}^{m} \rho_k^2 \tag{1.9}$$

can be approximated by a χ^2_m distribution and proposed the following Box-Pierce test statistic

$$Q(\hat{\rho}) = n \sum_{k=1}^{m} \hat{\rho}_k^2 \tag{1.10}$$

to check goodness-of-fit. They also proved that this test statistic converged to a χ^2 distribution with degrees of freedom $m - p - q$ asymptotically under the null hypothesis. However the Box-Pierce test performs poorly in application mainly because it is oversimplified in replacing $\tilde{Q}(\rho)$ with $Q(\rho)$ and the $Q(\hat{\rho})$ values can seriously deviate from a χ^2_{m-p-q} distribution. For better finite-sample performance Ljung and Box (1978) then proposed the Ljung-Box test statistic given by

$$\tilde{Q}(\hat{r}) = n(n+2) \sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{n-k}, \tag{1.11}$$

which can be approximated by the χ^2_{m-p-q} distribution under the null hypothesis. It substantially improves the Box-Pierce test and may be adequate in most applications despite of the fact that the variance of $\tilde{Q}(\hat{\rho})$ is larger than that of the χ^2_{m-p-q} variable.

As for choosing the value of m, Ljung (1986) claimed large m may impair the performance of the Ljung-Box test because the finite-sample significance levels of $\tilde{Q}(\hat{\rho})$ for large m tend to exceed the nominal significance levels of the χ^2_{m-p-q} distribution. A smaller m value can help improve the performance of the Ljung-Box test. This follows because if the model is inadequately specified, the problem would be exposed within a few lags, corresponding to a small value for m.

As an analog to portmanteau tests based on ACF, Monti (1994) proposed a portmanteau test to check lack-of-fit based on PACF. If the model adequately fits the data and the residuals behave like random noise, the PACF should not be significantly different from 0. Let π_k be the PACF of z_t and $\hat{\pi}_k$ be the estimated PACF from the residuals $\hat{z}_t$. Similar to the Ljung-Box test, the Monti test statistic is given by

$$Q(\hat{\pi}) = n(n+2) \sum_{k=1}^{m} \frac{\hat{\pi}_k^2}{n-k}, \tag{1.12}$$

which can be approximated by a χ^2 distribution with degrees of freedom $m - p - q$ for large n. Though Monti (1994) claimed that Monte's test outperforms Ljung-Box when the fitted model understates the order of the moving average component, the results in Andy and Wu (1997) showed that the empirical power of the Monti and the Ljung-Box tests are actually similar if m is properly chosen (motivated by the discussion about choosing m in Ljung (1986)). Pena and Rodriguez (2002)

proposed another portmanteau statistic,

$$D_m = n[1 - |\hat{R}_m|^{1/m}], \tag{1.13}$$

where R_m is the residual correlation matrix of order m with the $(i,j)^{th}$ off-diagonal element being $\hat{\rho}_{i-j}$. They also showed that $|\hat{R}_m|^{1/m} = \prod_{k=1}^{m}(1-\hat{\pi}^2)^{(m+1-i)/m}$, implying $|\hat{R}_m|^{1/m}$ is a weighted function of the first m partial autocorrelation coefficients of the residuals. Under the null hypothesis, D_m can be approximated by a Gamma($\alpha = b/2$, $\beta = 1/2\alpha$) distribution where both α and β are functions of m, p and q. For a better approximation, they suggested replacing $\hat{R}_m$ with $\tilde{R}_m$ which uses the standardized autocorrelation coefficient $(n-2)/(n-k)\hat{\rho}_{i-j}$. This test may outperform the Ljung-Box and Monti tests because it is highly asymmetric with respect to ρ_k and puts more weight on lower-order ρ_k, while Ljung-Box and Monti treat all ρ_k equally. Because $\tilde{R}_m$ may be ill-conditioned for large m relative to n and may not have a determinant (McLeod and Lin, 2006), Pena and Rodriguez (2006) proposed a modified test statistic,

$$D_m^{\star} = -n\sum_{k=1}^{m}\frac{m+1-k}{m+1}\log(1-\hat{\pi}_k^2), \tag{1.14}$$

which has the same asymptotic distribution as D_m and does not require calculating the determinant. However the logarithm calculation in $D_m^{\star}$ may cause computation instability. The portmanteau test in McLeod and Mahdi (2012) also has the same issue of calculating $|\hat{R}_m|$.

A recent weighted portmanteau test proposed by Pena and Rodriguez (2006) is derived from the trace of the square of the m_{th} order autocorrelation matrix. A weighted Ljung-Box statistic is defined as

$$\tilde{Q}_W = n(n+2)\sum_{k=1}^{m}\frac{m+1-k}{m}\frac{\hat{\rho}_k^2}{(n-k)}, \tag{1.15}$$

and similarly a weighted Monti statistic is given by

$$\tilde{M}_W = n(n+2)\sum_{k=1}^{m}\frac{m+1-k}{m}\frac{\hat{\pi}_k^2}{(n-k)}. \tag{1.16}$$

It is easy to see that more weight is put on smaller lags. Their asymptotic distributions can be approximated by a Gamma(α, β) process with α and β being specified through p, q and m. This weighted portmanteau test can outperform Ljung-Box and Monte tests.

Portmanteau tests also have application in linear regression models where the residuals from the fitted model should be checked for serial correlation/autocorrelation. If the residuals do not behave as random noise, the variance of parameter estimates may be inflated, which renders the relevant t-test, F-test or chi-squared test invalid and can lead to incorrect decisions in applications. In the presence of nonzero autocorrelations in the residuals, one can resort to the Cochrane-Orcutt procedure. Firstly a linear regression model is fit, then an ARMA(p, q) model is fit to the residuals and parameters are estimated. Next the original regression model is modified to incorporate the ARMA(p, q) error. Unless the user is confident that the autocorrelation might be described by an AR(1) model that the Durbin-Watson statistic (Durbin and Watson, 1950, 1951) was specially designed for, the Ljung-Box test or other portmanteau tests are recommended.

We should also point out that the power of portmanteau tests may be impaired by conditional heteroskedasticity which are usually modeled by GARCH(p, q) processes. In presence of conditional heteroskedasticity, bootstrapped Box-Pierce test (Horowitz et al., 2006) applied blocks-of-blocks bootstrap to approximate the distribution of the Box-Pierce statistic under the null hypothesis. The tests in McLeod and Mahdi (2012) and Fisher and Gallagher (2012) can be modified to test for heteroskedasticity. But these considerations will beyond the scope of the dissertation. In Chapter 4 we will study goodness-of-fit tests for checking the adequacy of ARMA models from the L_1-norm perspective. Instead of estimating the sample autocorrelation with the usual $\hat{\rho}_k$, we will use the Cauchy estimator quantified by the L_1 norm. A new portmanteau test statistic will be proposed and its asymptotic distribution derived.

Chapter 2

Multiple Change-Point Detection

Many long instrumental records of atmospheric temperature greatly benefit climate research. However these records are often systematically altered by recalibration of an instrument, relocation of a weather station, etc. These artificial factors can cause abrupt mean shifts in measured temperature. Failure to find such abrupt shifts in the records could lead to wrong conclusions about temperature development, while unveiling them helps readjust the raw series to reflect underlying changes in actual temperatures. The problem of detecting multiple undocumented change-points in temperature series has been well studied (Caussinus and Mestre, 2004, Lund and Reeves, 2002, Menne and Williams, 2005, Menne and Williams, 2009). Many of the proposed methods use reference series (typically from spatially-close measuring stations) to isolate and remove the effect of inhomegenieties. When only a single long instrumental series is used, a linear model can be used to give an indication of the long-term change of the temperature series. In this paper we propose a new method of detecting multiple undocumented change-points in univariate temperature sequences. For ease of exposition, we follow the recommendation of Wang (2003) and use a model with (one) simple linear trend to reflect long-term temperature change. Trend-type change-points are suggested to be tested after the underlying undocumented change-points are detected as suggested in Wang (2003).

As pointed out by Caussinus and Lyazrhi (1997), multiple change-point detection can be viewed as a model selection problem and tackled by testing multiple hypotheses or by optimizing some model fit criterion. To detect a single change-point in a sequence with simple linear trend, various hypothesis test statistics have been proposed, for example, a generalized likelihood ratio test (GLRT) (Kim and Siegmund, 1989). Basically, a GLRT statistic, T_k, is calculated at each potential

change-point time k, then the statistic $\max_k\{T_k\}$ and its asymptotic distribution are used to test for statistical significance. The asymptotic distribution of $\max_k\{T_k\}$ can be expressed as a function of a Gaussian process. Although no closed-form formulas for the quantiles are available, these quantiles can be found via numerical integration (Kim and Siegmund, 1989). For a more recent treatment of GLRT in climate setting see Gallagher et al. (2013).

To tackle the problem of multiple change-point detection, one of the most established methods is to combine hierarchic binary segmentation of the series and successive hypothesis tests (Scott and Knott, 1974, Vostrikova, 1981). We firstly test the null hypothesis H_0: *No change-point exists in the sequence* against the alternative H_1: *At least one change-point is in the sequence.* If the calculated statistic, $\max_k\{T_k\}$, is statistically significant, the sequence is divided into two subsequences. The one-change test is then applied to the two subsequences. The same procedure continues until no statistics calculated for the current subsequences are statistically significant. But this method is intrinsically problematic. Although one hypothesis test has a type I error rate of α, applying it to m subsequences would result in an accumulative type I error rate of $1-(1-\alpha)^m$ assuming independence of the tests. Note that $1-(1-\alpha)^m$ increases with m. The direct consequence is that many false change-points are probably detected (Lavielle and Teyssiere, 2006, Scott and Knott, 1974, Vostrikova, 1981).

An alternative to sequential testing is the model selection technique. We begin it by writing a linear model which includes a dummy variable for each potential change-point and a trend parameter. Suppose that Y_t, $t = 1, \ldots, n$, are independently and normally distributed with finite variance σ^2. The number and positions of change-points in the sequence are unknown. Let μ be the detrended mean in absence of change-points. Denote the simple linear trend parameter by α_1. For parameter identifiability, we assume that no change-point would occur at the first observation and the last observation of the sequence. Let $\beta_j,\ j = 2, \ldots, n-1$, be the amplitude of the mean shift from time $j-1$ to j. Using matrix notation, let $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_{n-1})'$ with each element corresponding to a potential change-point. Let $\mathbb{I} = (1, \ldots, 1)' \in \mathbb{R}^n$, $\mathbf{T} = (1, \ldots, n)^{'}$ and $\mathbf{X} = [\mathbf{X}_2, \ldots, \mathbf{X}_{n-1}]$, where $\mathbf{X}_j = (0, \ldots, 1)'$ with the first $j-1$ elements being 0 and thereafter 1 for $j = 2, \ldots, n-1$. Let $\boldsymbol{\theta} = (\mu, \beta_2, \ldots, \beta_{n-1},\ \alpha_1)^{'}$ be the parameter vector and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{'}$ be the error vector. The linear model is then displayed as

$$\mathbf{Y} = [\mathbb{I}, \mathbf{X}, \mathbf{T}]\ \boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{2.1}$$

Note that the trend parameter α_1, though of great interest in application, is not directly related to the number or positions of change-points. Our goal is to estimate those change-point parameters $\boldsymbol{\beta}$ and determine which are statistically nonzero.

Typically one would expect to have a few actual change-points within the set of potential changes. For example, in the U.S. temperature record we expect between 5 and 7 change-points per century (Menne et al., 2009). In terms of the full model (2.1) which contains a parameter for each potential change-point, we expect to have a sparse model with a few significant nonzero change-point parameters. In recent years there have been several methods proposed for simultaneous model selection and parameter estimation in sparse models, for example, ordinary Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006) and SCAD (Fan and Li, 2001). By adding an ℓ_1 penalty to an objective function, a sparse model can be selected while the parameters are simultaneously estimated; the ℓ_1 penalty forces small magnitude parameter estimates to be *shrunk* to identically zero (Tibshirani, 1996). Several variants of Lasso have been used for the purpose of multiple change-point detection, for example, ordinary Lasso via a total variation penalty (Harchaoui and Levy-Leduc, 2010), fused Lasso (Tibshirani, 1996) and group fused Lasso (Bleakley and Vert, 2011). Since they focus on applications such as signal EEG, speech processing, network intrusion detection and hot spot detection for CGH data, where the piecewise-constant mean approximation is more appropriate, they do not take trend into account. Detecting multiple change-points in temperature sequences with simple linear trend is nontrivial because one needs to estimate the sequence with a piecewise-linear approximation. Hence a procedure that can estimate the trend and change-points simultaneously is useful.

In this paper we show how adaptive Lasso can estimate the change-point times in temperature series with simple linear trend. The model identified by regular model selection criteria may overestimate the number of change-points. Assuming normality, we are able to apply multivariate t simultaneous confidence intervals to the model selected by adaptive Lasso. Confidence intervals containing zero are associated with non-significant change-points, hence reducing the number of change-points detected.

The rest of this paper is organized as follows. In Section 2 we describe the adaptive Lasso procedure together with post-selection. Section 3 shows simulation results. Section 4 applies the methods to a temperature series. We close the paper with discussion in Section 5.

2.1 Methods

2.1.1 Adaptive Lasso

Zou (2006) proposed adaptive Lasso as a method of simultaneous estimation and variable selection. Adaptive Lasso can shrink some parameters to be exactly zero and hence generates a sparse model. Assuming the number of parameters p_n is fixed, under some regularity conditions, Zou (2006) proved that adaptive Lasso enjoys the oracle properties in the sense of Fan and Li (2001), that is, consistency for variable selection and asymptotic normality of estimators. Huang et al. (2008) studied adaptive Lasso for sparse high-dimensional regression models which allow p_n to depend on the sample size and even be larger than the sample size. They proved that under regularity conditions adaptive Lasso still is consistent for variable selection and that the estimators of nonzero parameters have the asymptotical normal distribution if the zero parameters were known in advance. The number of parameters in our model increases with sample size, hence we are involved in the high-dimensional regression context.

The objective function in Huang et al. (2008) or Zou (2006) penalizes every parameter in the linear model. In the temperature context, we will not penalize the detrended mean μ. As global warming is well recognized, we don't want adaptive Lasso to shrink the trend parameter α_1 to be zero either. Thus $\boldsymbol{\theta}$ is estimated as

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \left\| \mathbf{Y} - \mu\mathbb{I} - \sum_{j=2}^{n-1} \mathbf{X}_j \beta_j - \alpha_1 \mathbf{T} \right\|^2 + \lambda_n \sum_{j=2}^{n-1} \hat{w}_j \left| \beta_j \right|, \tag{2.2}$$

where $(\hat{w}_2, \ldots, \hat{w}_{n-1})'$ is a weight vector for the change-point parameters selected prior to the minimization.

2.1.2 Implementation of Adaptive Lasso

For good interpretability we will not standardize our variables. To deal with the unpenalized parameters, we apply the idea presented in Tibshirani and Taylor (2012a). Let $\mathbf{Z} = [\mathbb{I}, \mathbf{T}]$ and $\mathbf{M}$ be the projection matrix onto $\mathbf{Z}$. Thus $\mathbf{M} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, where $\mathbf{I}$ is an $n \times n$ identity matrix. First we estimate

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{M}\mathbf{Y} - \mathbf{M} \sum_{j=2}^{n-1} \mathbf{X}_j \beta_j \right\|^2 + \lambda_n \sum_{j=2}^{n-1} \hat{w}_j \left| \beta_j \right|. \tag{2.3}$$

We then regress $\mathbf{Y} - \sum_{j=2}^{n-1} \mathbf{X}_j \hat{\beta}_j$ on $\mathbf{Z}$ to estimate μ and α_1.

In order to implement adaptive Lasso, a proper initial set of estimates for $\boldsymbol{\beta}$ is needed to construct each weight (the reciprocal of the magnitude of an initial parameter estimate) in the objective function (2.3). One choice is the ordinary least squares (OLS) estimates proposed in Zou (2006), but they do not exist for our model since $p = n$. Another choice is the marginal regression estimates suggested in Huang et al. (2008), but according to our simulation study (not shown here), this initial set of estimates prevents adaptive Lasso from returning a sparse model in our situation. A third choice is the ordinary Lasso estimates. Assuming the number of change-points is known and the distance between two consecutive change-points is at least $O(log(n))^2$, Harchaoui and Levy-Leduc (2010) proved that ordinary Lasso can estimate the underlying change-points consistently (though they did not consider the trend effect). Ordinary Lasso return a sparse model and the weights based on its estimates do not cause adaptive Lasso to overestimate the number of change-points as much as the marginal regression weights. Therefore, we choose the ordinary Lasso estimates as the initial set of estimates for $\boldsymbol{\beta}$, which is

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{MY} - \mathbf{M} \sum_{j=2}^{n-1} \mathbf{X}_j \beta_j \right\|^2 + \lambda_n \sum_{j=2}^{n-1} |\beta_j| \,. \tag{2.4}$$

Then β_j in (2.3) is automatically set to zero if $\hat{\beta}_j(\text{lasso})$ is exactly zero (weights for those β_j's are no longer necessary). Each of the remaining β_j's has weight $\hat{w}_j = 1/\hat{\beta}_j(\text{lasso})$.

Both of the optimizations, (2.3) and (2.4), can be implemented via the least angle regression selection (LARS) algorithm (Efron et al., 2004, Zou, 2006), which provides a sequence of candidate models. Each model has degrees of freedom given by

$$\text{df}(\hat{\mu}\mathbb{I} + \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\alpha}_1 \mathbf{T}) = 2 + \text{E}[\text{Rank}(\mathbf{MX_d})],$$

where $\mathbf{d}$ is the active set and $\text{E}[\text{Rank}(\mathbf{MX_d})]$ is estimated by the number of nonzero parameters (Tibshirani and Taylor, 2012a). Each model is also associated with some value of the regulation parameter λ_n. A model selection criterion is then minimized to determine the optimal model, which is tantamount to selecting a value for λ_n. An estimator $\hat{\beta}_j$ shrunk to zero indicates no change-point at time j, otherwise, a change-point is detected. A *false alarm* occurs if a change-point is detected at a position where no underlying change-point exists. As we either detect a true change-point in a

sequence or not, we call the probability of detecting a true change-point the *detection rate*. A proper model selection criterion should not detect many false alarms while maintaining high detection rates for underlying change-points. The Bayesian information criterion (BIC) is defined as

$$-2\ln(L) + \ln(n)K,$$

where L is Gaussian Likelihood function and K is the number of change-points detected in a candidate model. Another selection criterion with a heavier penalty term proposed by Caussinus and Mestre (2004) in the change-point context is

$$-2\ln(L) + 2\frac{n}{n-1}\ln(n)K.$$

Denote this criterion by CM. Based on our simulation (not shown here), BIC tends to overfit the model while CM underfit the model. See Hannart and Naveau (2012) for a literature review concerning the criterion choice for multiple change-point detection. Here we use BIC to select the Lasso weights and CM to determine the final model because BIC avoids loss of potential change-points while CM prevents overfit.

2.1.3 Post-selection of Change-points

Multiple change-point selection methods of Lasso-type may find a large set of change-points (Harchaoui and Levy-Leduc, 2010). A post-selection step can be applied to the preselected change-points by these methods, for instance, a reduced version of dynamic programming (DP) in Harchaoui and Levy-Leduc (2010) or subjective thresholding as in Tibshirani and Wang (2008). We provide an example in the simulation section below showing that the reduced version of DP may not diminish overestimation from ordinary Lasso effectively. Instead of DP, we post-select change-points from adaptive Lasso using multivariate t simultaneous confidence intervals, which turns out to be effective in reducing overestimation.

We now construct the simultaneous confidence intervals for change-point parameters based on a fitted linear model. First adaptive Lasso selects a linear model that includes K nonzero change-point parameters. Let $\mathbf{d} = (d_1, \ldots, d_K)$ be the index set of the selected change-point parameters. Denote the set of the corresponding change-point parameters by $\boldsymbol{\beta}_d = (\beta_{d_1}, \ldots, \beta_{d_K})'$. The selected

model excludes many zero change-point parameters but may still contain a few more parameters than necessary. Note that μ and α_1 stay in the selected model. Let $\boldsymbol{\theta_d} = (\mu, \beta_{d_1}, \ldots, \beta_{d_K}, \alpha_1)'$. Under normality, we obtain that

$$\hat{\boldsymbol{\theta}}_d \sim N(\boldsymbol{\theta_d},\ \sigma^2(\mathbf{X_d'X_d})^{-1}), \tag{2.5}$$

where $\mathbf{X_d}$ is a subset of columns (including $\mathbb{I}$ and $\mathbf{T}$) corresponding to $\boldsymbol{\theta_d}$ from the design matrix in (2.1). We estimate σ^2 with

$$s^2 = \frac{||\mathbf{Y} - \mathbf{X_d}\hat{\boldsymbol{\theta}}_{\boldsymbol{d}}||^2}{n-K-2}.$$

It is known that $(n-K-2)s^2/\sigma^2$ follows a chi-square distribution with degrees of freedom $n-K-2$. Denote by $t_\nu(\boldsymbol{u}, \Sigma)$ a multivariate t-distribution with degrees of freedom ν, location vector $\boldsymbol{u}$ and covariance matrix Σ. It follows that

$$\frac{(\hat{\boldsymbol{\theta}}_{\boldsymbol{d}} - \boldsymbol{\theta_d})}{s} \sim t_{n-K-2}\big(\mathbf{0}, (\mathbf{X_d'X_d})^{-1}\big). \tag{2.6}$$

Since our goal is to determine the significance of the change-point parameters $\boldsymbol{\beta_d}$, what we exactly need is the following distribution:

$$\frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{d}} - \boldsymbol{\beta_d})}{s} \sim t_{n-K-2}(\mathbf{0}, \mathbf{V}), \tag{2.7}$$

where $\mathbf{V}$ is the covariance matrix obtained by deleting the columns and rows corresponding to μ and α_1 from the matrix $(\mathbf{X_d'X_d})^{-1}$. Then the simultaneous confidence intervals of $\boldsymbol{\beta_d}$ are given by

$$P\left\{\begin{array}{c} \left|\frac{\hat{\beta}_{d_1}-\beta_{d_1}}{s\sqrt{\mathbf{V}[1,1]}}\right| \le t_{\alpha/2:n-K-2}(\mathbf{0},\mathbf{V}), \\ \vdots \\ \left|\frac{\hat{\beta}_{d_K}-\beta_{d_K}}{s\sqrt{\mathbf{V}[K,K]}}\right| \le t_{\alpha/2:n-K-2}(\mathbf{0},\mathbf{V}) \end{array}\right\} = 1-\alpha, \tag{2.8}$$

where $t_{\alpha/2:n-K-2}(\mathbf{0}, \mathbf{V})$ is the $100(1-\alpha/2)$th percentile of multivariate t-distribution. Any confidence interval of β_{d_i} containing zero means that no change-point occurs at time d_i. Otherwise, a change-point is detected at time d_i with simultaneous confidence level at least $100(1-\alpha)\%$.

Note that post-selection by simultaneous confidence intervals may not be effective here if one

uses the marginal regression weights. Since the marginal regression weights prevent adaptive Lasso from returning a sparse model in our situation, the resulting simultaneous confidence intervals have to include bounds for every parameter β_{d_i}. Bounding type I error rate by α causes the confidence interval of each parameter to be wide enough to contain zero. Hence, the probability of detecting actual change-points becomes quite small. On the contrary, simultaneous $100(1-\alpha)\%$ confidence intervals based on the *sparse model* selected by adaptive Lasso using the ordinary Lasso weights contain bounds for only a few parameters β_{d_i}, thus are narrow enough to test significance of $\hat{\beta}_{d_i}$ and further reduce the dimension of the selected model.

2.2 Simulation Results

This section investigates the performance of adaptive Lasso and its post-selection. We consider three methods including adaptive Lasso (M_1), adaptive Lasso with post-selection by multivariate t confidence intervals (M_2), and successive GLRT (M_3). Percentiles of the multivariate t-distribution used to construct confidence intervals are obtained via the `qmvt` function in R package **mvtnorm**. We simulate random sequences from the true models: $\mathbf{Y} = \mathbf{h} + \alpha_1 \mathbf{T} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$.

Example 1: $\mathbf{h} = \mathbf{0}$, $\alpha_1 = 0.1$ and $\sigma = 1$.

Example 2: We uses the simulation setup from Caussinus and Mestre (2004), where six change-points are positioned at 20%, 40%, 50%, 70%, 75% and 85% of the sequence and the jump size of each change-point is constant. Let

$$\mathbf{h} = (\underbrace{0,\ldots,0}_{0.2n},\ \underbrace{2,\ldots,2}_{0.2n},\ \underbrace{4,\ldots,4}_{0.1n},\ \underbrace{2,\ldots,2}_{0.2n},\ \underbrace{0,\ldots,0}_{0.05n},\ \underbrace{2,\ldots,2}_{0.1n},\ \underbrace{4,\ldots,4}_{0.15n}).$$

Then jump size is 2. Set $\alpha_1 = 0.1$ and $\sigma = 1$.

Example 3: This is the same with *Example 2* except that the jump size is 3.

Example 4: We simulate random sequences with three change-points at 39%, 56% and 87%. Their jump sizes are -1, 0.75 and 0.5, respectively. Let

$$\mathbf{h} = (\underbrace{0,\ldots,0}_{0.38n},\ \underbrace{-1,\ldots,-1}_{0.27n},\ \underbrace{0.75,\ldots,0.75}_{0.31n},\ \underbrace{0.5,\ldots,0.5}_{0.24n}).$$

Set $\alpha_1 = 0.02$ and $\sigma = 0.5$.

Note that *Example 1* demonstrates performance when no change-points are present, *Example 2* and *Example 3* are similar to the simulation setting in Caussinus and Mestre (2004), and *Example 4* is motivated by our data application.

2.2.1 Type I Error Rate

A good technique of detecting multiple change-points should not give too many false alarms when there are no change-points in a random sequence. For each sample size $n = 100$, 200 and 1000 we simulate 1000 sequences from *Example 1*. For the null hypothesis H_0: *No change-points exist in the sequence* against the alternative H_1: *At least one change-point exists in the sequence*, we are interested in

$$\text{type I error rate} = \frac{\#\text{ of sequences in which change-points are detected}}{\#\text{ of simulated sequences}}.$$

We use a significance level 0.05 to get the quantiles of GLRT statistic. The simulated type I error rate of each method is displayed in Table 2.1. In general, we see that adaptive Lasso (M_1) is not likely to detect false alarms compared to successive GLRT (M_3). The simulated type I error rate of adaptive Lasso decreases from 0.046 to 0.007 as sample size increases from 100 to 1000. Post-selection (M_2) does not rule out any false alarms detected by adaptive Lasso (M_1) under this no-change-points scenario, but the type I error rate of adaptive Lasso (M_1) alone is already sufficiently low. Successive GLRT has a simulated type I error rate above 0.3. This is due to the high accumulative type I error of applying the GLRT successively to subsequences generated by binary segmentation. The simulated type I error rate also increases with the sample size. As the sample size grows, more subsequences tend to be generated by segmentation, and more GLRTs are implemented, resulting in an increasing accumulative type I error rate.

Table 2.1: Type I Error Rate

Method	$n = 100$	$n = 200$	$n = 1000$
M_1	0.046	0.026	0.007
M_2	0.046	0.026	0.007
M_3	0.307	0.330	0.391

2.2.2 Multiple Change-point Detection

We now investigate the performance of adaptive Lasso and multivariate t post-selection in detecting multiple underlying change-points. To start off we present some results from ordinary Lasso to convince us of the adaptive Lasso necessity. We implement ordinary Lasso on 1000 simulated sequences from *Example 3* when $n = 200$. Figure 2.1 includes histograms of detected change-points. Assuming that we know the maximum number of change-points $K = 6$ (a): ordinary Lasso can detect only the middle change-point with a high probability. Assuming that we misspecified the maximum number of change-points $K = 10$ (b): DP does not help reduce the number of change-points detected by Lasso. Assuming that the number of change-points is unknown (c), we use the CM selection criterion to choose the optimal model. We see that ordinary Lasso detects too many false alarms and can seriously overestimate the number of change-points. This implies that even for fairly large sample size, ordinary Lasso does not work well at least for our simulated example.

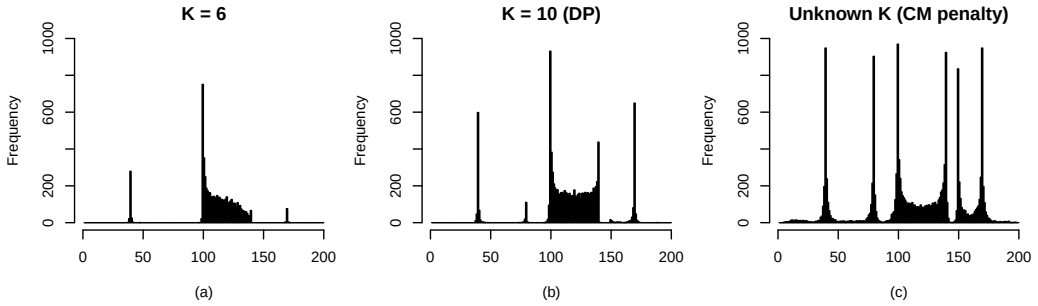


Figure 2.1: Histograms of positions of detected change-points by ordinary Lasso. Six change-points are simulated at 20%, 40%, 50%, 70%, 75% and 85% of each sequence.

Next we will compare adaptive Lasso and post-selection by multivariate t confidence intervals with the successive GLRT procedure. We simulate 1000 random sequences for each sample size $n = 100$, 200 and 1000 from *Example 2* and *Example 3*, respectively.

Table 2.2 shows the percentage of sequences in which K change-points are detected. In terms of false alarms, adaptive Lasso overestimates the number of simulated change-points ($K \geq 7$) except when jump size = 2 and $n = 100$. Post-selection by multivariate t confidence intervals (M_2) can lower down the false alarm rate significantly, and can identify the correct number of change-points with the highest probability among the three methods considered (See the row marked by "$\star$"). Under both jump size scenarios, successive GLRT (M_3) seems to do fairly well in controlling

false alarm rates when $n = 100$, but its false alarm rates grow fast with the sample size.

Figures 2.2 and 2.3 provide histograms of positions of change-points detected by each method and clearly show the detection rate of each underlying change-point. Each row in Figures 2.2 and 2.3 corresponds to a section of the same sample size and jump size in Table 2.2. When jump size is 2, adaptive Lasso and its post-selection (M_1, M_2) gain more power in picking up the correct change-points as sample size gets large. When jump size is 3, they can detect all the change-points with high probability.

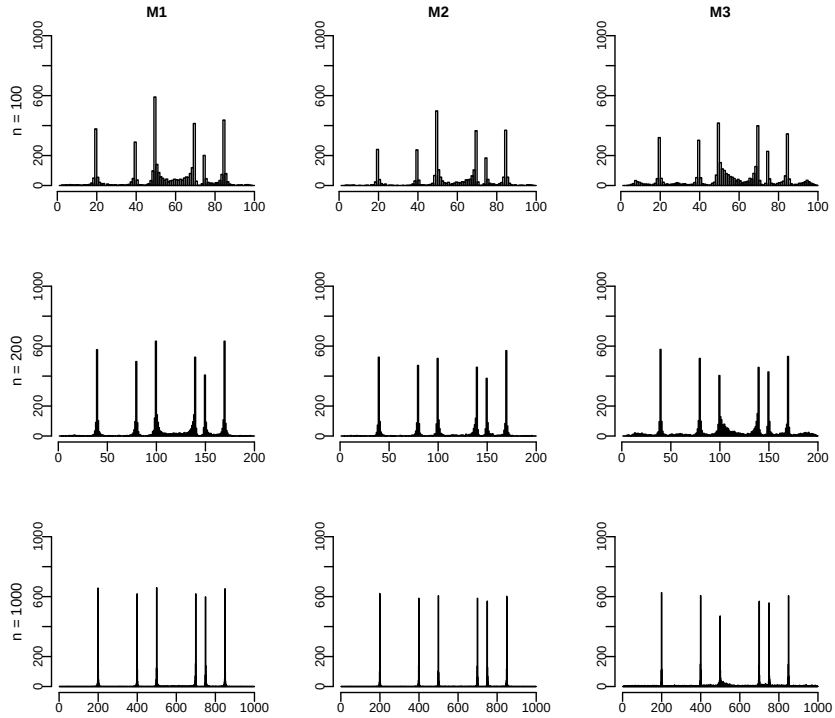


Figure 2.2: Histograms of positions of detected change-points by the three methods. Jump size = 2.

We also have examined the performance of each method in detecting change-points of different jump sizes. We simulate 1000 random sequences of length $n = 100$ from *Example 4*. This example is motivated by the real data set in Section 4 which contains three documented change-points at 39, 56 and 87 and has a rough estimated standard deviation 0.5 and trend 0.02. Simulation results are presented in Table 2.3 and Figure 2.4. We obtain consistent information as conveyed by the six change-points study. Both adaptive Lasso (M_1) and successive GLRT (M_3) tend to overestimate the number of change-points; Post-selection (M_2) can identify the correct number of change-points

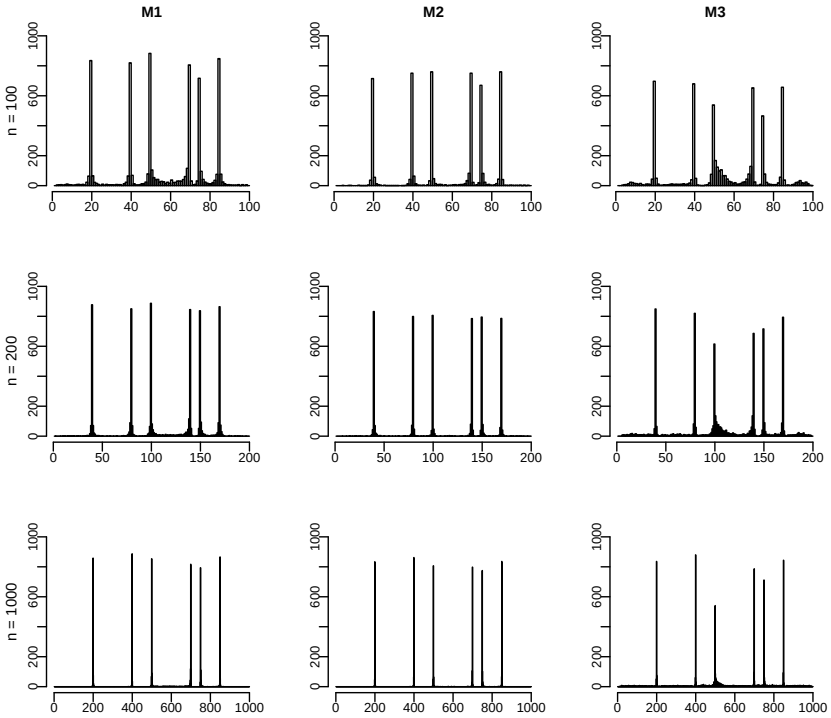


Figure 2.3: Histograms of positions of detected change-points by the three methods. Jump size = 3.

Table 2.2: Percentage of sequences in which K change-points are detected

Jump = 2	$n = 100$			$n = 200$			$n = 1000$		
K	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
0	0.004	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.040	0.065	0.000	0.000	0.001	0.000	0.000	0.000	0.000
2	0.170	0.267	0.094	0.025	0.045	0.002	0.000	0.000	0.000
3	0.184	0.273	0.194	0.054	0.102	0.010	0.000	0.002	0.000
4	0.170	0.226	0.249	0.083	0.199	0.034	0.000	0.038	0.000
5	0.146	0.104	0.224	0.118	0.288	0.100	0.001	0.231	0.000
⋆6	**0.113**	**0.056**	**0.159**	**0.237**	**0.324**	**0.235**	**0.410**	**0.640**	**0.013**
7	0.085	0.003	0.066	0.252	0.038	0.271	0.365	0.086	0.082
8	0.053	0.000	0.012	0.146	0.003	0.230	0.155	0.003	0.121
≥ 9	0.035	0.000	0.002	0.085	0.000	0.118	0.069	0.000	0.784
Jump = 3	$n = 100$			$n = 200$			$n = 1000$		
K	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.006	0.013	0.002	0.000	0.000	0.000	0.000	0.000	0.000
3	0.009	0.045	0.032	0.000	0.003	0.000	0.000	0.000	0.000
4	0.035	0.130	0.114	0.000	0.032	0.005	0.000	0.014	0.000
5	0.051	0.279	0.182	0.001	0.240	0.007	0.002	0.147	0.000
⋆6	**0.277**	**0.485**	**0.348**	**0.345**	**0.648**	**0.140**	**0.605**	**0.776**	**0.015**
7	0.298	0.044	0.246	0.377	0.070	0.327	0.259	0.061	0.074
8	0.198	0.003	0.060	0.189	0.007	0.282	0.087	0.002	0.135
≥ 9	0.126	0.001	0.014	0.088	0.000	0.239	0.047	0.000	0.776

⋆: The number of detected change-points equals that of the simulated change-points.

with a higher probability (0.639 as shown in Table 2.3). All the methods considered have similar detection rates for the three underlying change-points.

Table 2.3: Percentage of sequences in which K change-points are detected

	n=100		
K	M_1	M_2	M_3
0	0.000	0.001	0.000
1	0.042	0.056	0.000
2	0.113	0.233	0.021
⋆3	**0.525**	**0.639**	**0.586**
4	0.224	0.064	0.290
≥ 5	0.096	0.007	0.103

⋆: The number of detected change-points equals that of the simulated change-points.

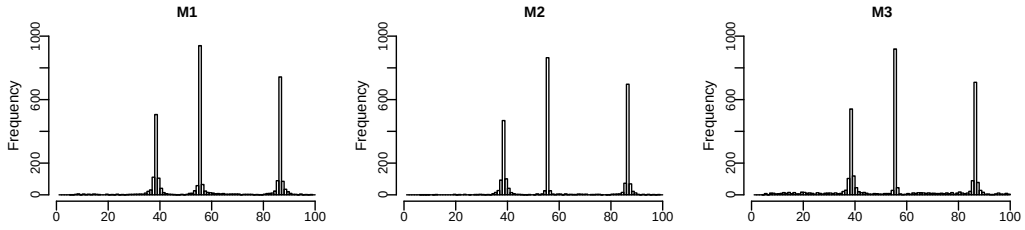


Figure 2.4: Histograms of positions of detected change-points by the three methods. Three change-points are simulated at 39, 56 and 87. Jump size $= (-1, 0.75, -0.5)$. $\sigma = 0.5$.

In general, adaptive Lasso with post-selection by confidence intervals is a promising method of detecting multiple change-points for the following reasons: (1) it provides high detection rates for change-points; (2) it has relatively low false alarm rates; (3) it identifies the true number of change-points with high probability.

2.3 Application

Figure 2.5 contains plots of the yearly temperature data from Tuscaloosa, Alabama, which ranges from year 1902 to 2000 ($n = 99$). This temperature series is known to have three documented station changes at year 1939, 1956 and 1987 (Reeves et al., 2007) and hence can serve as a proving ground for methods of multiple change-points detection.

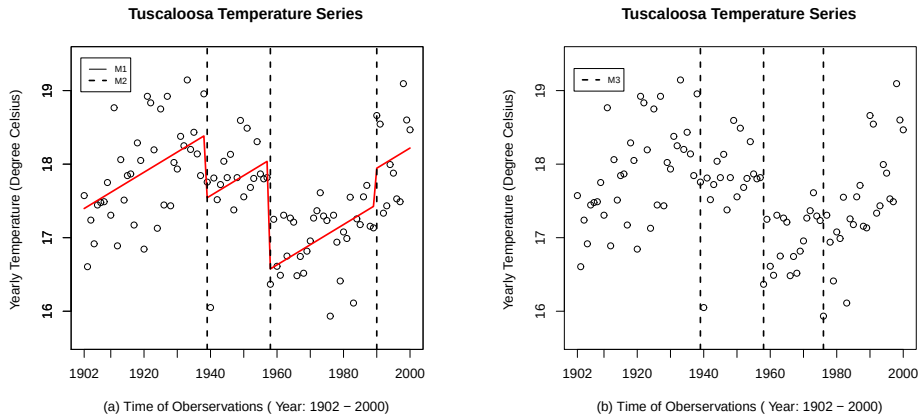


Figure 2.5: (a) Change-points detected by adaptive Lasso (M_1) and post-selection (M_2); (b) Change-points detected by successive GLRT (M_3).

We detect the change-points in the Tuscaloosa sequence by adaptive Lasso (M_1), adaptive Lasso with post-selection (M_2) and successive GLRT (M_3), respectively. The solid line in Figure 2.5(a) displays the fitted result by adaptive Lasso. The linear trend rate underlying this temperature series is estimated around 0.027^oC/year. It detects three change-points at year 1939, 1958 and 1990. The change-point at 1939 is an accurate hit. The 1958 and 1990 change-points are two and three years later than the corresponding documented change-points, respectively. The estimated jump size of each change-point is -0.867^oC, -1.149^oC and 0.492^oC, respectively. The estimated standard error is 0.516^oC. Post-selection by multivariate t simultaneous confidence intervals (M_2), indicated by the dashed line, identifies all three change-points as statistically significant. For comparison, we also include the ordinary Lasso result: six change-points are detected at year 1939, 1940, 1955, 1956, 1958 and 1990. The dashed line in Figure 2.5(b) displays the change-points detected by successive GLRT (M_3). Three change-points are found at year 1939, 1958 and 1976. The first two change-points are exactly the same found by adaptive Lasso. The 1976 change-point is not part of the documented change-points set and could be a false alarm.

Figure 2.6 shows the standardized residuals (Stand. Res) plots from the adaptive Lasso fit. There is no evident violation of the normal assumption based on the Q-Q plot. Applying the weighted Ljung-Box test (Fisher and Gallagher, 2012) to the residuals, we obtain a p-value of 0.2636 at lag 5 indicating no significant autocorrelation. There might be some concern about heterocedasticity from the first residual plot (Stand.Res vs Time). The standard deviation between 1940 and 1975

seems to be smaller than that in the remaining observations. Our method can be generalized to allow for nonconstant variance by replacing the least squares part in the objective function with weighted least squares. This will be left as future work.

Overall, adaptive Lasso post-selected by multivariate t confidence intervals (M_2) detects all the three documented change-points; successive GLRT (M_3) detects two documented change-points, but also detects one potential false alarm.

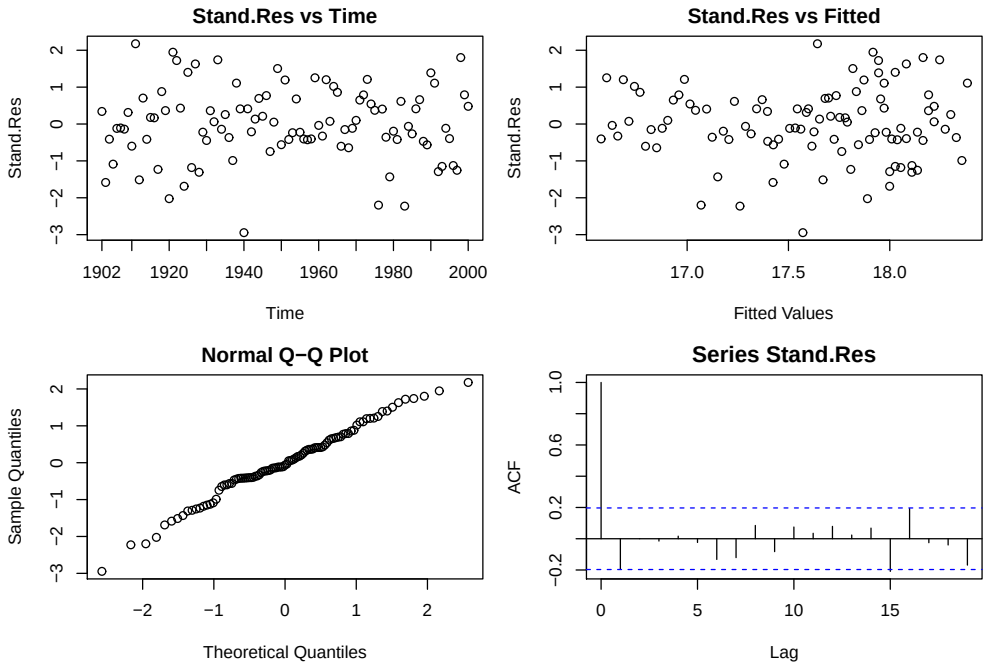


Figure 2.6: Model diagnosis of adaptive Lasso fit.

2.4 Conclusion and Discussion

In this paper we have explored the idea of using adaptive Lasso to simultaneously estimate parameters and select change-point times in temperature series with simple linear trend. At least in the specific examples considered above, adaptive Lasso has high detection rate of the true change-points but may overestimate the number of change-points. Applying multivariate t simultaneous confidence intervals to a model selected by adaptive Lasso can significantly reduce overestimation while still maintaining reasonable detection rates for the underlying change-points. Comparison is made with successive GLRT. As sample size grows, overestimation of the successive GLRT procedure gets much worse. Generally adaptive Lasso posted-selected by multivariate t simultaneous confidence intervals outperforms successive GLRT in terms of both detection rate and overestimation.

The methods in this paper can be extended in several ways. For monthly or even daily temperature data, the constant variance assumption could be relaxed to allow for periodic variance which is natural in many regions. The periodic variance can be estimated by the method of moments or be parameterized by linear combinations of trigonometric functions. The least squares part of the objective function could then be replaced with a weighted least squares based on the variance. A second extension would be to allow for autocorrelation in the error terms. In the at-most-one-change-point context, Robbins et al. (2011a) summarized hypothesis tests for detecting change-points in correlated data. The methods in Lund et al. (2007) allow for both periodicities and autocorrelation. These methods could be modified by adding an ℓ_1 penalty and use shrinkage methods to detect multiple change-points. Also, nonparametric regression has been used to allow for flexible trend rather than linear trend (Bates et al., 2012, Bowman et al., 2006). Under our setting we can apply kernelized Lasso (Gao et al., 2010, Wang et al., 2007) to incorporate flexible trend.

Chapter 3

Shrinkage Methods via Genetic Algorithm

Statistical modeling in linear regression mainly includes ordinary least squares (OLS), ridge regression, least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005). The OLS estimator has the minimum variance among all the unbiased linear estimators. In presence of multicollinearity the OLS estimator can have very large variance due to computational instability of inverting $X^{'}X$, which leads to poor prediction. The ridge regression estimator, defined as $(X^{'}X + \lambda_2 I)^{-1}X'\mathbf{y}$, stabilizes matrix inversion through the nonnegative regularization parameter λ_2. It is often shown to provide better prediction when OLS is hurt by multicollinearity. But ridge regression retains all predictor variables and does not give a good interpretable model. Best subset selection and stepwise regression are able to perform variable selection and parameter estimation simultaneously. But they are criticized for ignoring the stochastic error in the discrete process of variable selection, and a little disturbance in the data may lead to a very different model.

Another approach, often referred to as L_1 shrinkage/regularization/penalized methods, has become very popular. This technique was proposed by Lasso. Because the resulting estimator is a thresholding rule, Lasso is able to provide a sparse model (Fan and Li, 2001). It is also credited for being much stable than subset selection through a continuous shrinkage process. But the bias in the Lasso estimator is noticeably large (Fan and Li, 2001). Later Zou and Hastie (2005) proposed

the elastic net which relies on the L_1 penalty to achieve model sparsity and the L_2 penalty to address multicollinearity. While Lasso arbitrarily chooses only one variable among a group of highly correlated variables, elastic net selects the whole group through the "grouping" effect, and is claimed to improve Lasso in prediction whenever ridge regression improves OLS.

The salient point is that no regression method is shown to absolutely dominate in all cases (Tibshirani, 1996, Zou, 2006, Zou and Hastie, 2005). If all predictor variables are uncorrelated and highly significant, OLS can provide good enough parameter estimation and good prediction. In this setting ridge regression, Lasso and elastic net should reduce to OLS, otherwise, they introduce extra bias in the parameter estimates. When the true underlying model is sparse, Lasso and elastic net are favored because of their sparse model representation. However ridge regression has been shown to outperform Lasso and elastic net in terms of prediction. We presume that this is because that the heavy impact of extra bias in parameter estimates of Lasso or elastic net discounts the fruits of a sparse model. Estimators from ridge regression, Lasso and elastic net are all biased. They aim to trade estimation bias for smaller mean squared error (MSE) resulting better prediction. Since Lasso and elastic net restrict all parameters to be penalized by the same norm, some parameter estimates can suffer from unnecessary bias. Assume that x_1 is uncorrelated with other independent variables but correlated with the response. Then penalizing x_1 by either the L_1 or L_2 norm would cause extra bias. An ideal modeling procedure would leave x_1 unpenalized. Further if x_2, x_3 and x_4 are highly correlated and only β_2 is significant, one would think that β_2 should be subjected to the L_2 penalty for a bias-variance trade-off while β_3 and β_4 are penalized by the L_1 penalty. Having β_2 penalized by the L_1 norm would shrink it towards zero. Generally speaking, an ideal modeling procedure would automatically treat parameters differently, i.e., penalize based on the significance and correlation of the predictor variables.

In this chapter we propose a new procedure which can simultaneously perform variable selection and parameter estimation. If some true parameters are zero, our procedure can shrink them to zero. Nonzero parameters can be estimated as well as if the true sub-model were known in advance. Our procedure works at least as well as the canonical methods, and often result in better model fits by allowing flexible penalties for different parameters. The resulting nonzero estimates have smaller bias and variance. Specifically, in our objective function we allow each variable to be unpenalized or selectively penalized by L_1, L_2 or both L_1 and L_2 norms. The canonical methods of OLS, ridge regression, Lasso and elastic net simply become special cases of our framework. The

proposed procedure turns out to be a large combinatorial optimization problem. It can be solved by an exhaustive search if one can afford the time and the number of independent variables p is reasonably small. We also describe the genetic algorithm as an efficient alternative in large p scenarios.

3.1 The Objective Function

Suppose that $\mathbf{y} = [y_1, \ldots, y_n]'$ is the response vector. Let $\mathbf{x}_1, \cdots, \mathbf{x}_p$ denote the predictor variables. Suppose that each $\mathbf{x_i}$ is centered and standardized such that $\sum_j x_{ij} = 0$ and $\sum_j x_{ij}^2 = 1$ for $i = 1, \ldots, p$. Also let $\mathbf{y}$ be centered such that $\sum_j y_j = 0$. We assume that $E[\mathbf{y}] = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p$. Denote

- $G_1 = \{\beta_i \text{ unpenalized}\}$;
- $G_2 = \{\beta_i \text{ penalized by the } L_2 \text{ norm}\}$;
- $G_3 = \{\beta_i \text{ penalized by the } L_1 \text{ norm}\}$;
- $G_4 = \{\beta_i \text{ penalized by both } L_1 \text{ and } L_2 \text{ norm}\}$.

For simplicity we use $i \in G_1$ to indicate that x_i belongs to category G_1 and the same for other categories. All variables could possibly be classified into any of the four categories G_1, G_2, G_3, G_4. We refer to the vector of their categories as "grouping". That is, a grouping is a p-dimensional vector which takes values from G_1, G_2, G_3 and G_4. Given a grouping, our objective function is written as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i=1}^{p} \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 \sum_{i \in G_2 \text{ or } G_4} |\beta_i|^2 + \lambda \sum_{i \in G_3 \text{ or } G_4} |\beta_i|, \tag{3.1}$$

where λ_2 and λ are nonnegative regularization parameters, the second term on the right provides the L_2 penalty and the third the L_1 penalty. Each grouping determines a distinct way of fitting the model and generates a set of parameter estimates. It is easy to see that the canonical methods of OLS, ridge regression, elastic net and Lasso correspond to special groupings. If a grouping happens to consist of G_1 for each variable, the objective function (3.1) reduces to OLS. If a grouping happens to consist of G_2 for all variables, it reduces to ridge regression. Similarly, Lasso and elastic net also correspond to special groupings and are covered by our objective function (3.1). A regular grouping is also easy to comprehend. Assume that $\mathbf{x_1}$ is known to be an irrelevant variable in

predicting $\mathbf{y}$ beforehand, we might assign it to G_3 to be shrunk to zero by the L_1 norm. But it is typically unknown what variables are irrelevant to $\mathbf{y}$ and what variables are correlated in practice. A predetermined grouping is usually difficult to pursue. It is not uncommon for real data sets to have irrelevant variables, correlation or independence simultaneously. Our conjecture is that the optimal grouping for an real data set should often be a mixture of G_1, G_2, G_3 and G_4 rather than a vector of a single category. Therefore, the objective function in (3.1) has created a more practical framework, which is also very general and flexible and makes it possible to explore the universe of methods of fitting linear models.

Since each grouping determines a set of model estimates, the problem of finding the best model fit amounts to finding an optimal grouping. For a multiple linear model with p variables, the total number of possible grouping is 4^p. This number grows exponentially as p increases. For a model with eight variables, the total number of grouping is $4^8 = 65,536$. Searching for an optimal grouping among 4^p possibilities is a nontrivial task especially for large p. To solve this problem we will introduce Bayesian information criterion (BIC) as our model selection criteria and illustrate how to calculate the degrees of freedom for a model fit under its grouping. This allows us to compare all the 4^p fitted models on the same footing. The model fit ranking first by BIC is considered to be an optimal model with its corresponding grouping being an optimal grouping. For simplicity we will denote a grouping just by its subscripts (1, 2, 3, 4) of the categories and ignore the capital letter "G" henceforth.

3.2 Degrees of Freedom

In statistical modeling the researchers often need to select a model among a family of candidate models. Some model selection criteria have been proposed to help accomplish this task. The most common criteria include Akaike Information Criterion (AIC), BIC or Schwarz criterion (SBC, SBIC), Mallow's Cp, minimum description length (MDL), cross-validation (CV), etc. In determining which criteria to use for scoring the candidate models, first one should be clear about the objective of the analysis: whether to obtain a model with best prediction or to identify the true model among the candidates. AIC is *asymptotically efficient* in the sense of selecting the candidate model which minimizes prediction mean squared error (Shibata, 1999). BIC has been proven to be *asymptotically consistent* for model selection, i.e., assuming that the true model belongs to a

family of candidate models, the probability that BIC will select the true model approaches one as the sample size gets large (Nishii, 1984). But AIC is not asymptotically consistent and BIC is not asymptotically efficient. Therefore, if prediction is the objective, one should choose AIC or CV as the model selection criterion; if identifying the true model is the main concern, BIC should be used. In this chapter we will consider identifying the true model as the primary concern and hence BIC is the favored selection criterion.

A linear model of p variables would be fitted 4^p times based on all the possible groupings. Next the BIC score is calculated for each model fit. All the groupings (or the model fits) are ranked by BIC. The grouping with the lowest BIC score is considered as the optimal grouping. Two distinct groupings probably share the same lowest BIC score. For example, a parameter can be shrunk to exactly zero whether the associated variable is classified into G_3 or G_4. In such case both groupings share the same model fit.

Obtaining the BIC score amounts to finding the effective degrees of freedom for each model fit. It is unfair to compare two models ignoring their model complexity: one using three variables and another using ten variables. Degrees of freedom allows one to compare models on the same footing. We now determine the degrees of freedom under different groupings. Let $\hat{\mathbf{y}} = g(\mathbf{y})$ be a procedure of estimating $\mathbf{y}$. Then

$$(\hat{y}_j - \mu_j)^2 = (\hat{y}_j - \mu_j)^2 - (y_j - \mu_j)^2 + 2(\hat{y}_j - \mu_j)(y - \mu_j). \tag{3.2}$$

Summing over j yields

$$E\left\{ \frac{||\hat{\mathbf{y}} - \mu||^2}{\sigma^2} \right\} = E\left\{ \frac{||\mathbf{y} - \mu||^2}{\sigma^2} - n \right\} + 2\sum_{j=1}^{n} cov(\hat{y}_j, y_j)/\sigma^2 \tag{3.3}$$

Effective degrees of freedom (Efron et al., 2004) is defined as

$$df = \sum_{i=1}^{m} cov(\hat{y}_j, y_j)/\sigma^2. \tag{3.4}$$

Denote a linear smoother $\hat{\mathbf{y}} = H\mathbf{y}$. Then

$$df(H\mathbf{y}) = \text{trace}(H). \tag{3.5}$$

For example, the OLS estimator is $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Then the degrees of freedom in OLS is

$$df(\text{ols}) = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{rank}(\mathbf{X}). \tag{3.6}$$

The second equality is due to a fact in linear algebra: the trace of a projection matrix is equal to the rank of $\mathbf{X}$. This expression implies that $\mathbf{X}$ is of full rank. Otherwise, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ can be replaced with the generalized matrix inverse $\mathbf{X}^{+}$. The ridge regression estimates are $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda_2\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$. Then the degrees of freedom in ridge regression is

$$df(\text{ridge}) = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}))^{-1}\mathbf{X}'). \tag{3.7}$$

The degrees of freedom for Lasso and elastic net do not have an analytical form because they are not linear smoothers. Bootstrap can be used to estimate $df^{\text{lasso}}(\lambda)$ (Efron et al., 2004) by using definition (3.4), but it is computationally expensive. Zou et al. (2007) provided an unbiased estimator for $df^{\text{lasso}}(\lambda)$. Given any λ, they showed that

$$df^{\text{lasso}}(\lambda) = E[\text{trace}(\mathbf{X}_A(\mathbf{X}_A^{'}\mathbf{X}_A)^{-1}\mathbf{X}_A)] = E|A|, \tag{3.8}$$

where $A = \{\mathbf{x}_i : \beta_i \neq 0\}$ and $|A|$ is the number of variables in A. Then $|A|$ is an unbiased estimator for $df^{\text{lasso}}(\lambda)$. As the entire solution of Lasso can be calculated by the least angle regression (LARS) algorithm, $|A|$ is attainable without extra effort. Another result showed by Zou et al. (2007) for BIC is that

$$\lambda(\text{optimal}) = \text{argmin}_\lambda \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n\sigma^2} + \frac{log(n)}{n}\widehat{df}(\lambda) \tag{3.9}$$

and

$$k^\star = \text{argmin}_k \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n\sigma^2} + \frac{log(n)}{n}\widehat{df}(\lambda_k) \tag{3.10}$$

coincide such that $\lambda(\text{optimal}) = \lambda_{k^\star}$, where k is the step number in the LARS algorithm. The degrees of freedom of elastic net does not have an analytic expression either, but using a similar argument to Lasso an unbiased estimator (Tibshirani and Taylor, 2012b) can also be obtained from

$$df^{\text{enet}}(\lambda_2) = E[\text{trace}(\mathbf{X}_A(\mathbf{X}_A^{'}\mathbf{X}_A + \lambda_2 I)^{-1}\mathbf{X}_A^{'})], \tag{3.11}$$

Next we will illustrate how to calculate the degrees of freedom for a model fit with a grouping consisting of at least two categories from G_1, G_2, G_3, and G_4. All scenarios can fall into one of the following cases:

Case 1: Have G_1 and G_2 variables. Then the estimates are defined as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i \in G_1} \mathbf{x}_i \beta_i - \sum_{i \in G_2} \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 \sum_{i \in G_2} \beta_i^2. \tag{3.12}$$

For example, $\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_4\beta_4 + \boldsymbol{\epsilon}$. One possible grouping might be $(1, 1, 2, 2)$. Then (3.12) becomes

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y}_\star - \mathbf{X}_\star \boldsymbol{\beta} \right\|^2, \tag{3.13}$$

where $\mathbf{y}_\star = \begin{pmatrix} \mathbf{y} \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{X}_\star = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \\ 0 & 0 & \sqrt{\lambda_2} & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_2} \end{pmatrix}$. Since $\hat{\mathbf{y}}_\star = \mathbf{X}_\star(\mathbf{X}_\star^{'}\mathbf{X}_\star)^{-1}\mathbf{X}_\star^{'}\mathbf{y}_\star$, one can easily show $\hat{\mathbf{y}} = H\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{J})^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{J} = \text{diag}(0,\ 0,\ 1,\ 1)$. This shows that

$$\text{df}(\lambda_2) = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{J})^{-1}\mathbf{X}'). \tag{3.14}$$

Case 2: Have G_1 and G_3 variables. Then the estimates are defined as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i \in G_1} \mathbf{x}_i \beta_i - \sum_{i \in G_3} \mathbf{x}_i \beta_i \right\|^2 + \lambda \sum_{i \in G_3} |\beta_i|. \tag{3.15}$$

Revisit the example under Case 1. Consider the grouping $(1, 1, 3, 3)$ instead. Let $\mathbf{X}_s = [\mathbf{x}_1, \mathbf{x}_2]$ and $\mathbf{X}_t = [\mathbf{x}_3, \mathbf{x}_4]$. Let $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_t$ be the corresponding parameters. Let $\mathbf{M} = \mathbf{I} - \mathbf{X}_s(\mathbf{X}_s^{'}\mathbf{X}_s)^{-1}\mathbf{X}_s^{'}$. Note that $\mathbf{y} - \mathbf{X}_s\hat{\boldsymbol{\beta}}_s - \mathbf{X}_t\hat{\boldsymbol{\beta}}_t = \mathbf{M}\mathbf{y} - \mathbf{M}\mathbf{X}_t\boldsymbol{\beta}_t$. Then (3.15) becomes

$$\widehat{\boldsymbol{\beta}}_t = \text{argmin}_{\boldsymbol{\beta}_t} \left\| \mathbf{M}\mathbf{y} - \mathbf{M}\mathbf{X}_t\boldsymbol{\beta}_t \right\|^2 + \lambda|\boldsymbol{\beta}_t|. \tag{3.16}$$

Then

$$\text{df}(\mathbf{X}_s\hat{\boldsymbol{\beta}}_s + \mathbf{X}_t\hat{\boldsymbol{\beta}}_t) = 2 + E[\text{rank}(M\mathbf{X}_{tA})], \tag{3.17}$$

where 2 is the dimensionality of $\mathbf{X}_s$ and tA indicates the current active variables from $\mathbf{X}_t$.

Case 3: Have G_3 and G_4 variables but no G_1 and G_2 variables. Then the estimates are defined as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i \in G_3} \mathbf{x}_i \beta_i - \sum_{i \in G_4} \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 \sum_{i \in G_4} \beta_i^2 + \lambda \sum_{i \in (G_3, G_4)} |\beta_i|. \tag{3.18}$$

Now consider the grouping $(4, 4, 3, 3)$. Similarly to elastic net (Zou and Hastie, 2005), the above can also be rewritten as a Lasso problem:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y}_\star - \mathbf{X}_\star \boldsymbol{\beta} \right\|^2 + \lambda |\boldsymbol{\beta}|, \tag{3.19}$$

where $\mathbf{y}_\star = \begin{pmatrix} \mathbf{y} \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{X}_\star = \begin{pmatrix} \frac{1}{1+\sqrt{\lambda_2}} \mathbf{x}_1 & \frac{1}{1+\sqrt{\lambda_2}} \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \\ \frac{1}{1+\sqrt{\lambda_2}} \sqrt{\lambda_2} & 0 & 0 & 0 \\ 0 & \frac{1}{1+\sqrt{\lambda_2}} \sqrt{\lambda_2} & 0 & 0 \end{pmatrix}$ with $\frac{1}{1+\sqrt{\lambda_2}}$ being the normalizer. Using the similar argument for the degrees of freedom of elastic net, one can show that

$$\text{df}(\mathbf{X}\hat{\boldsymbol{\beta}}) = E\left[\text{trace}\left(\mathbf{X}_A (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{J})^{-1} \mathbf{X}_A^{'} \right) \right], \tag{3.20}$$

where $\mathbf{J} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$.

Case 4: Have variables from (G_1 or G_2), (G_2 or G_3) and (G_3 or G_4). The estimates are defined as

$$\widehat{\boldsymbol{\beta}}(\lambda_2, \lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i \in (G_1, G_2, G_3, G_4)} \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 \sum_{i \in (G_2, G_4)} \beta_i^2 + \lambda \sum_{i \in (G_3, G_4)} |\beta_i|. \tag{3.21}$$

Consider the example $\mathbf{y} = \mathbf{x}_1 \beta_1 + \ldots + \mathbf{x}_8 \beta_8 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ and the grouping $(1, 2, 3, 3, 2, 3, 4, 4)$. Then the objective function in (3.21) becomes

$$\widehat{\boldsymbol{\beta}}(\lambda_2, \lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{i=1}^{8} \mathbf{x}_i \beta_i \right\|^2 + \lambda_2 (|\beta_2|^2 + |\beta_5|^2 + |\beta_7|^2 + |\beta_8|^2) + \lambda (|\beta_3| + |\beta_4| + |\beta_6|). \tag{3.22}$$

Rewrite (3.22) as

$$\widehat{\boldsymbol{\beta}}(\lambda_2, \lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y}_\star - \mathbf{X}_\star \boldsymbol{\beta} \right\|^2 + \lambda (|\beta_3| + |\beta_4| + |\beta_6|), \tag{3.23}$$

where $\mathbf{y}_\star = \begin{pmatrix} \mathbf{y} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{X}_\star = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 & \mathbf{x}_6 & \mathbf{x}_7 & \mathbf{x}_8 \\ 0 & \sqrt{\lambda_2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\lambda_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\lambda_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\lambda_2} \end{pmatrix}$. Denote $x_{i\star}$ each column of $\mathbf{X}_\star$. Let $\mathbf{X}_{s\star} = [\mathbf{x}_{1\star},\ \mathbf{x}_{2\star},\ \mathbf{x}_{5\star},\ \mathbf{x}_{7\star},\ \mathbf{x}_{8\star}]$ and $\mathbf{X}_{t\star} = [\mathbf{x}_{3\star},\ \mathbf{x}_{4\star},\ \mathbf{x}_{6\star}]$. Also let $\mathbf{M} = \mathbf{I} - \mathbf{X}_{s\star}(\mathbf{X}_{s\star}^{'}\mathbf{X}_{s\star})^{-1}\mathbf{X}_{s\star}^{'}$. Since

$$\mathbf{y}_\star - \mathbf{X}_{s\star}\boldsymbol{\beta}_s - \mathbf{X}_{t\star}\boldsymbol{\beta}_t = \mathbf{M}\mathbf{y}_\star - \mathbf{M}\mathbf{X}_{t\star}\boldsymbol{\beta}_t, \tag{3.24}$$

we have

$$\widehat{\boldsymbol{\beta}}_t = \text{argmin}_{\boldsymbol{\beta}_t}||\mathbf{M}\mathbf{y}_\star - \mathbf{M}\mathbf{X}_{t\star}\boldsymbol{\beta}_t||^2 + \lambda|\boldsymbol{\beta}_t|. \tag{3.25}$$

After obtaining $\widehat{\boldsymbol{\beta}}_t$, one can solve for $\boldsymbol{\beta}_s$ by applying OLS to

$$\mathbf{y}_\star - \mathbf{X}_{t\star}\widehat{\boldsymbol{\beta}}_t = \mathbf{X}_{s\star}\boldsymbol{\beta}_s + \boldsymbol{\epsilon}. \tag{3.26}$$

Then, using (3.24), (3.25) and (3.26), the degrees of freedom in (3.23) becomes

$$\widehat{df}(\lambda_2, \lambda) = \text{trace}(\mathbf{X}_s(\mathbf{X}_{s\star}^{'}\mathbf{X}_{s\star})^{-1}\mathbf{X}_s^{'}) + E\left[\text{trace}\left((\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}')[1{:}n, 1{:}n]\right)\right], \tag{3.27}$$

where $\mathbf{Q} = \mathbf{M}\mathbf{X}_{t\star A}$.

3.3 Genetic Algorithm

Obviously one can carry out an exhaustive search over all the possible groupings. The fitted model with the smallest BIC is chosen as the optimal model. But this strategy quickly becomes infeasible to implement for even a small number of variables. Take a toy example of an eight-variable model. It can take over five hours to implement the exhaustive search over 4^8 groupings. The computation time increases exponentially fast with p. To resolve this intractable question we now introduce genetic algorithm (GA), which is a popular heuristic approach of finding a promising solution to a large combinatorial optimization problem. Generally speaking, GA mimics the process of species evolution in nature: chromosomes with good adaptability to the environment are preserved

in the population by natural selection, and have a higher chance of breeding off-springs who will carry on good features of their parents' chromosomes.

Chromosome Representation. Each model fitting is identified by its corresponding grouping, which in turns serves as the chromosome in GA. For example, if $p = 8$, an example of a regular chromosome might be represented as $(1, 1, 3, 3, 2, 2, 3, 4)$. The chromosome of the OLS fit is eight straight 1's and the ridge fit is eight straight 2's. The chromosome representation of a model fitting is identical to its grouping.

Population Initialization. The total number of possible groupings is 4^p. The initialized population in a GA search is a subset of the 4^p groupings. In our simulation and data analysis we take the subset to be of size 200. We also supply the initialized population with the chromosomes of the four canonical methods. The rest chromosomes are randomly selected from the possible groupings. Each bit on a chromosome is randomly selected from 1, 2, 3, and 4 with equal probability 1/4 and with replacement.

Selection. The BIC score is evaluated for each model fit in the current generation. The models are then sorted by the BIC score in increasing order. We retain the top 20% fitted models in the next generation. The other 80% groupings are discarded and are refilled with off-spring who are created by crossover between chromosomes in the current population.

Crossover. Two parent chromosomes in the current generation are randomly chosen to generate an off-spring for the next generation. The probability of being chosen as the parent chromosomes to crossover is proportional to their rank. The new off-spring is generated through uniform crossover. Specifically, the off-spring chromosome can inherit each gene from either the parent chromosome or the father chromosome with equal chance. Although no satisfactory theories have been developed for uniform crossover, it is said to have good empirical performance.

Mutation. Each gene in a chromosome has a chance of mutation. For example, if a gene on a chromosome is 3 it can be mutated to 1, 2 or 4. A relatively high mutation rate helps maintain the diversity of the population. But a high mutation rate incurs a lot of variation to the population and the genetic algorithm would reduce to pure random search. So a small mutation rate is recommended, but not too small to lose diversity. In our study it is set to be $1/p$ so that one mutation per chromosome is expected on average.

Stopping Rule. The algorithm terminates after a fixed number of iterations. In our simulation we stop the algorithm after 20 iterations which can return a fitted model close to that

obtained from the exhaustive search.

GA takes much longer than a single Lasso or elastic net fit, but it is worthwhile to run if one aims at a sparse model representation with less biased estimates than those from Lasso or elastic net fits. The model chosen by GA works at least as well as a lasso or elastic net fit in terms of BIC score or any other preferred model selection criteria.

3.4 Simulation Study

In this section we investigate the performance of OLS, ridge regression, elastic net, Lasso, exhaustive search and genetic algorithm. For each of the five examples considered below we simulate 150 data sets each consisting 100 observations. Models identified by groupings are estimated from the data. To save computation time, we only consider a few λ_2 values: 0, 0.01, 0.1, 0.5, 1, 5, 10, 100, 1000 (the same as in Zou and Hastie (2005)). We select the tuning parameters k and λ_2 using BIC. All the methods considered are compared based on the mean squared error $\sum_{i=1}^{p} |\hat{\beta}_i - \beta_i|^2$ and variable selection. The number of significant variables selected by each method is reported as frequencies. We generate data from the following model

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}),$$

where $p = 8$ and $\sigma = 1$. We pick examples that we expect the canonical methods to do well and compare how we do in comparison. The five examples simulated are listed as follows:

1. Let $\beta = (20, 20, 20, 20, 20, 20, 20, 20)$. Any pair of x_i and x_j $(i \neq j)$ are uncorrelated.
2. Let $\beta = (2, 2, 3, 2, 2.5, 2, 3, 2)$. The pairwise correlation is set to be corr(i, j)= $0.5^{|i-j|}$.
3. Let $\beta = (3, 5, 0, 0, 5, 3, 0, 0)$. The pairwise correlation is set to be corr(i, j)= 0.9.
4. Let $\beta = (3, 5, 0, 0, 5, 3, 0, 0)$. The pairwise correlation is set to be corr(i, j)= 0.
5. Let $\beta = (4, 6, 0, 0, 3, 3, 0, 0)$. The pairwise correlation is set to be corr(i, j)= 0.85 for i, $j = 3, \ldots, 7$, $i \neq j$ and 0 for other pairwise correlation.

Example 1 is created to have all variables uncorrelated with each other and all parameters extremely significant compared to the noise. Under this scenario OLS can be expected to fit the data

sets well enough. There may be not much room left for improvement in the model fit so that ridge regression, Lasso and elastic net almost reduce to OLS by setting the regularization parameters to be close to zero. We want to show that our procedure can pick the OLS grouping as the optimal grouping most frequently. In Example 2 all predictors are correlated and coefficients are significant. This example is designed for ridge regression to stand out since the true model is not sparse so that no L_1 shrinkage is necessary while L_2 shrinkage can help handle correlation. The data generating mechanism in Examples 3 and 4 is a sparse model with all predictor variables being correlated and uncorrelated, respectively. Both examples are where either Lasso or elastic net is expected to work the best since L_1 shrinkage is needed to identify the underlying sparse model. Example 5 is designed to have $\mathbf{x}_i, i = 3, \ldots, 7$ correlated with each other and each of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_8$ uncorrelated with the other seven predictor variables. Also because $\beta_i, i = 1, 2, 5, 6$ are nonzero, the optimal grouping under this example may not come from any of the canonical methods but a mixture of different categories.

Tables 3.1 - 3.5 summarize the simulation results of these examples. We compute the BIC score and $\sum_{i=1}^{p} |\hat{\beta}_i - \beta_i|^2$ for each of the 150 simulated data sets. The means of those statistics are reported together with the standard deviations (shown in the parenthesis immediately after mean). In Table 3.1, since each coefficient is 20 times as large as the noise and all predictor variables are uncorrelated under Example 1, OLS works fairly well as expected and the fitted models from penalized methods perform almost the same as OLS in terms of BIC and parameter estimates. The exhaustive search and genetic algorithm select the OLS grouping in 150 out of the 150 runs. Similarly, Table 3.2 shows that ridge regression has the smallest BIC score among the four canonical methods. Meanwhile, the ridge estimates have small mean squared error compared to other canonical methods. Full search and GA successfully select the ridge regression grouping in 140 out of the 150 runs. It is expected that L_1 shrinkage methods should do well in Examples 3 and 4. As shown in Table 3.3, elastic net has the smallest BIC among the canonical methods followed closely by Lasso and far lower than those of OLS and ridge regression. Note that elastic net has a slightly high mean squared error compared to Lasso. This is probably due to the double shrinkage of elastic net though attenuated by rescaling (Zou and Hastie, 2005). Apart from the canonical methods, the good news for us is that the full search finds the same optimal grouping $(2, 2, 3, 3, 2, 2, 3, 3)$ in 104 out of the 150 runs. The BIC corresponding to this optimal grouping is much lower than that of elastic net or Lasso, and the absolute bias and mean squared error are consistently lower as well. This optimal

grouping shrinks the zero coefficients (β_3, β_4, β_7, β_8) to zero by assigning the corresponding variables to category G_3 to be penalized by the L_1 norm. Because of correlation among all predictor variables, the nonzero coefficients are assigned to category G_2 and be penalized by the L_2 norm. Similarly to Example 3, elastic net also performs the best in terms of BIC in Example 4. This time an optimal grouping $(1, 1, 3, 3, 1, 1, 3, 3)$ is selected in 78 out of the 150 runs. Because all predictor variables are uncorrelated, the zero coefficients are assigned to category G_1 and unpenalized to avoid unnecessary bias. The results shown in Table 3.5 identifies $(1, 1, 3, 3, 2, 2, 3, 3)$ as the optimal grouping in 106 out of the 150 runs, which further support our claim about grouping. Because $\mathbf{x_1}$ and $\mathbf{x}_2$ are uncorrelated with any other variables and have nonzero coefficients, they are assigned to category G_1. The other variables are dealt in the same way as in Examples 3 and 4. Thus the resulting optimal grouping consists variables from three categories. One can also notice that the full search can identify the correct number of significant variables with a large probability.

Table 3.1: Example 1 – methods comparison

Methods	BIC	$\sum_{i=1}^{p} \lvert\hat{\beta}_i - \beta_i\rvert^2$	Time (mins)	# of Vars Selected (freq.)	FavoredGrouping
OLS	30.08 (14.34)	0.07 (0.04)	0	8(150)	
Ridge	30.08 (14.34)	0.07 (0.04)	0	8(150)	
Lasso	30.08 (14.34)	0.07 (0.04)	0	8(150)	
ElasticNet	30.08 (14.34)	0.07 (0.04)	0	8(150)	
FullSearch	30.08 (14.34)	0.07 (0.04)	227	8(150)	(1,1,1,1,1,1,1,1)-150
GA	30.08 (14.34)	0.07 (0.04)	10	8(150)	(1,1,1,1,1,1,1,1)-150

Table 3.2: Example 2 – methods comparison

Methods	BIC	$\sum_{i=1}^{p} \lvert\hat{\beta}_i - \beta_i\rvert^2$	Time (mins)	# of Vars Selected (freq.)	FavoredGrouping
OLS	29.49 (16.67)	0.12 (0.09)	0.0	8(150)	
Ridge	29.06 (16.64)	0.12 (0.08)	0.0	8(150)	
Lasso	29.49 (16.67)	0.12 (0.09)	0.0	8(150)	
ElasticNet	29.48 (16.60)	0.12 (0.09)	0.0	8(150)	
FullSearch	29.06 (16.64)	0.12 (0.09)	218	8(150)	(2,2,2,2,2,2,2,2)-140
GA	29.06 (16.64)	0.12 (0.09)	9	8(150)	(2,2,2,2,2,2,2,2)-140

Since all methods almost perform the same with OLS, no further exploration is needed. For Examples 2 - 5 we also display the boxplots of parameter estimates from the canonical method with the smallest BIC and the GA method in Table 3.6. The estimates for each parameter are put side by side for ease of comparison. Because ridge regression performs the best among all canonical

Table 3.3: Example 3 – methods comparison

Methods	BIC	$\sum_{i=1}^{p} \lvert\hat{\beta}_i - \beta_i\rvert^2$	Time (mins)	# of Vars Selected (freq.)	FavoredGrouping
OLS	32.66 (13.74)	0.79 (0.44)	0	8(150)	
Ridge	32.49 (13.62)	0.97 (0.62)	0	8(150)	
Lasso	23.82 (14.28)	0.56 (0.42)	0	4(23) 5(48) 6(58) 7(17) 8(4)	
ElasticNet	23.81 (14.27)	0.58 (0.43)	0	4(23) 5(48) 6(58) 7(17) 8(4)	
FullSearch	17.46 (13.59)	0.43 (0.39)	255	4(115) 5(25) 6(10)	(2,2,3,3,2,2,3,3)–109
GA	17.54 (13.59)	0.40 (0.34)	13	4(111) 5(29) 6(10)	

Table 3.4: Example 4 – methods comparison

Methods	BIC	$\sum_{i=1}^{p} \lvert\hat{\beta}_i - \beta_i\rvert^2$	Time (mins)	# of Vars Selected (freq.)	FavoredGrouping
OLS	31.70 (13.65)	0.08 (0.04)	0	8(150)	
Ridge	31.70 (13.65)	0.08 (0.04)	0	8(150)	
Lasso	21.57 (13.81)	0.07 (0.04)	0	4(63) 5(45) 6(30) 7(10) 8(2)	
ElasticNet	20.98 (13.92)	0.08 (0.05)	0	4(60) 5(47) 6(30) 7(11) 8(2)	
FullSearch	16.98 (13.56)	0.05 (0.03)	287	4(91) 5(56) 6(3)	(1,1,3,3,1,1,3,3)-78
GA	16.98 (13.55)	0.05 (0.03)	12	4(93) 5(53) 6(4)	

Table 3.5: Example 5 – methods comparison

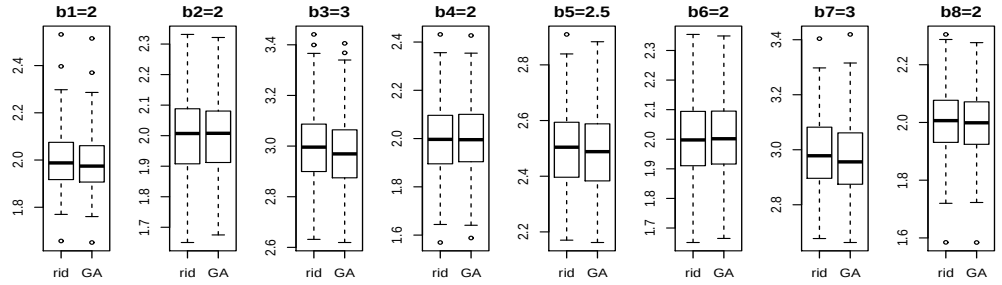
Methods	BIC	$\sum_{i=1}^{p} \lvert\hat{\beta}_i - \beta_i\rvert^2$	Time (mins)	# of Vars Selected (freq.)	FavoredGrouping
OLS	27.57 (15.32)	0.24 (0.18)	0	8(150)	
Ridge	27.36 (15.22)	0.28 (0.19)	0	8(150)	
Lasso	19.18 (15.35)	0.15 (0.15)	0	4(33) 5(55) 6(47) 7(15)	
ElasticNet	18.56 (15.30)	0.17 (0.20)	0	4(33) 5(54) 6(47) 7(14) 8(2)	
FullSearch	12.74 (15.06)	0.06 (0.10)	270	4(109) 5(30) 6(10) 7(1)	(1,1,3,3,2,2,3,3)-106
GA	12.76 (15.06)	0.06 (0.10)	12	4(113) 5(27) 6(9) 7(1)	

methods in Example 2, we compare ridge estimates and GA estimates. The two types of estimates are similar as GA selects the ridge grouping most frequently in the 150 runs. For Examples 3-5 comparison is made between elastic net and GA. In Example 3 we see that $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_7$, and $\hat{\beta}_8$ by our procedure are exactly equal to zero in most runs while those of elastic net have more variability. Our mean estimates of the remaining nonzero coefficients are closer to the true values than those of elastic net in general (except that the β_1 is estimated better by elastic net). In Example 4, although elastic net performs as well as our procedure in identifying the zero coefficients, our procedure gives almost unbiased estimates for the nonzero coefficients compared to biased estimates from elastic net. In Example 5, our procedure obviously outperforms elastic net in terms of both zero and nonzero parameter estimates.

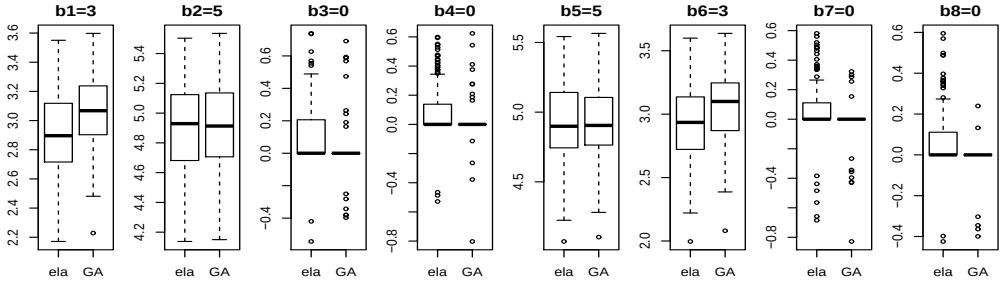
In summary, if a canonical method does the best among all possible groupings both full search and GA can find this canonical grouping successfully with a large chance, as shown in Examples 1 and 2. On the other hand, if one grouping can give a better fitted model than the canonical methods, our procedure can find this particular grouping with a large chance. As verified by Examples 3, 4 and 5 where elastic net has the smallest BIC among all the canonical methods and Lasso has smaller bias than elastic net, the optimal grouping selected by full search gives an even better fitted model than either elastic net or Lasso in terms of BIC, bias and mean squared error. All the simulation results convey a message that it is worthwhile to run the full search or GA for a sparse model which captures the salient predictor variables related to a response variable.

Table 3.6: Model estimates from the best canonical method and full search

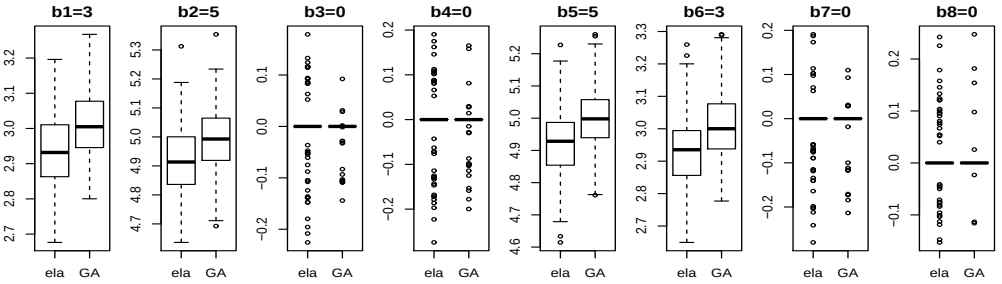
Model 2: Ridge regression and full search



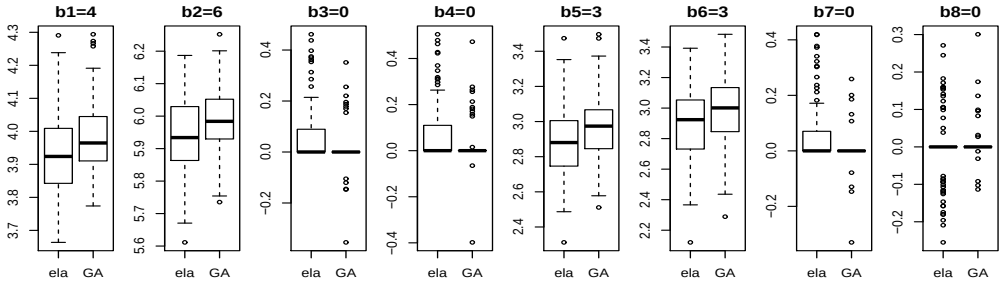
Model 3: Elastic net and full search



Model 4: Elastic net and full search



Model 5: Elastic net and full search



3.5 Data Analysis

We demonstrate our methods on two data sets. The prostate cancer data set in Section 3.5.1 was previously analyzed in Zou and Hastie (2005) and has eight predictor variables. We provide the model fits from both the exhaustive search and GA. The diabetes data set in Section 3.5.2 has been analyzed by Efron et al. (2004). This data set has ten variables. The exhaustive search becomes time-consuming, thus we only provide the model fit from GA.

3.5.1 Prostate cancer data

Prostate cancer is one of the top five cancers causing death in men. The data set analyzed in this section comes from a study of prostate cancer in Stamey et al. (1989). It is accessible from `http://rss.acs.unt.edu/Rdoc/library/ElemStatLearn/html/prostate.html`. Table 3.7 shows a small part of the data. Preoperative serum prostate specific antigen (PSA) was measured for 102 men before radical prostatectomy and reported as nanograms of PSA per milliliter (ng/mL) of blood. PSA usually exists in small quantities in the serum if a man lives with a healthy prostate, but can be elevated due to the occurrence of prostate cancer or other prostate abnormality. The PSA test has been a widely applied method of screening men for prostate cancer for years. The problem is, even though a man with a higher PSA level has a larger chance of having prostate cancer, prostate cancer is not the only factor causing the PSA level to rise. Thus some people are against the use of PSA test for screening considering misdiagnose as prostate cancer causes panic in patients.

Table 3.7: Prostate cancer data

ID	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{y}$
1	-0.5798	2.7695	50	-1.3863	0	-1.3863	6	0	-0.4308
2	-0.9943	3.3196	58	-1.3863	0	-1.3863	6	0	-0.1625
3	-0.5108	2.6912	74	-1.3863	0	-1.3863	7	20	-0.1625
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
95	2.9074	3.3962	52	-1.3863	1	2.4639	7	10	5.1431
96	2.8826	3.7739	68	1.5581	1	1.5581	7	80	5.4775
97	3.4720	3.9750	68	0.4383	1	2.9042	7	20	5.5829

The goal here is to examine which variables could possibly elevate serum PSA in addition to prostate cancer. The response is the logarithm of preoperative serum PSA ($\mathbf{y}$). The predic-

tor variables are log(cancer volume) ($\mathbf{x}_1$), log(prostate weight) ($\mathbf{x}_2$), age($\mathbf{x}_3$), the logarithm of the amount of benign prostatic hyperplasia ($\mathbf{x}_4$), seminal vesicle invasion ($\mathbf{x}_5$), log(capsular penetration) ($\mathbf{x}_6$), gleason score ($\mathbf{x}_7$) and percentage gleason score 4 or 5 ($\mathbf{x}_8$). Prostate cancer volume is a primary determinant of serum prostate specific antigen levels as indicated in the analysis of Stamey et al. (1989). Prostate weight is another risk factor of a high preoperative serum PSA level clinically. Seminal vesicle invasion by prostate cancer has been considered a sign of a poor prognosis and is an indicator variable with 1 indicating being invaded by prostate cancer and 0 indicating not invaded. Capsular penetration is an even more aggressive sign than seminal vesicle penetration. Benign prostatic hyperplasia may cause the PSA level to rise but the increment could be too slight to be significant. The gleason score takes value from 2 to 10, with 10 having the worst prognosis. Percentage gleason score 4 or 5 measures the proportion of tumors ranking 4 or 5 on the gleason score scale in a prostate specimen. These predictor variables are viewed as major determinants of prostate cancer progression. We want to explore whether they can elevate serum PSA.

Because prostate weight is missing for five patients, the regression analysis was done on 97 patients. Simply by looking at the correlation matrix, prostate cancer volume is most correlated with the PSA level (correlation $r = 0.734$), followed by seminal vesicle invasion ($r = 0.566$), capsular penetration ($r = 0.549$), weight ($r = 0.433$), percentage Gleason score 4 or 5 ($r = 0.422$) and gleason score ($r = 0.369$). These predictor variables are also correlated among themselves with the highest correlation being 0.75. The correlations between PSA level and age, the amount of benign prostatic hyperplasia are low (0.170 and 0.180, respectively). We compare the model fits from OLS, ridge regression, Lasso, elastic net, full search and GA on this data set. Table 3.8 displays the fitted results.

Table 3.8: Prostate cancer data – methods comparison

Methods	BIC Score	Parameters	Variables Selected
OLS	-37.61	Null	All
Ridge	-41.61	$\lambda_2 = 0.1$	All
Lasso	-45.51	$k = 3$	(1, 2, 5)
ElasticNet	-51.12	$k = 5$, $\lambda_2 = 1.0$	(1, 2, 5, 6, 8)
FullSearch	-53.78	$k = 0$, $\lambda_2 = 0.1$	(1, 2, 5)
GA	-53.78	$k = 0, \lambda_2 = 1.0$	(1, 2, 5)

$\star$: The groupings chosen by full search and GA are $(2, 2, 3, 3, 2, 3, 3, 3)$ and $(2, 2, 4, 3, 2, 4, 4, 4)$, respectively.

From Table 3.8 OLS has the largest BIC score. Ridge regression dominates OLS by a lower

BIC score. From Table 3.9 both OLS and ridge regression have a negative parameter estimate for age, which can generate a misleading conclusion that PSA levels decreases with age. Such unreasonable estimates should result from multicollinearity. Lasso benefits from a sparse model including $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_5$ and outperforms ridge regression in terms of BIC. The model estimated by elastic net has the smallest BIC among the four canonical methods. It also includes the predictor $\mathbf{x}_6$ and $\mathbf{x}_8$ in addition to those selected by Lasso. This may be due to the "grouping" fact that elastic net tends to select highly correlated variables into the model together. The model chosen by full search and GA have the same BIC which is smaller than any of the four canonical methods. The corresponds optimal groupings are $(2, 2, 3, 3, 2, 3, 3, 3)$ and $(2, 2, 4, 3, 2, 4, 4, 4)$, respectively. Although the two groupings differ, the model fits are exactly the same. See Table 3.9. The parameters of predictor variables in category 3 or 4 are shrunk to exactly zero. Assigning an insignificant variable into category G_3 or G_4 does not change the model fit for this data set since its coefficient is shrunk to exactly zero anyway.

Table 3.9: Prostate cancer data – model estimates

Methods	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
OLS	0.5643	0.6220	-0.0212	0.0967	0.7617	-0.1061	0.0492	0.0045
Ridge	0.4725	0.5964	-0.0155	0.0829	0.6658	-0.0238	0.0666	0.0032
Lasso	0.4730	0.4010	0.0000	0.0000	0.4419	0.0000	0.0000	0.0000
ElasticNet	0.4057	0.4103	0.0000	0.0000	0.5476	0.0995	0.0000	0.0019
FullSearch	0.4792	0.6331	0.0000	0.0000	0.6744	0.0000	0.0000	0.0000
GA	0.4792	0.6331	0.0000	0.0000	0.6744	0.0000	0.0000	0.0000

$\star$: The groupings chosen by full search and GA are $(2, 2, 3, 3, 2, 3, 3, 3)$ and $(2, 2, 4, 3, 2, 4, 4, 4)$, respectively.

3.5.2 Diabetes Data

This diabetes data set was first introduced in Efron et al. (2004) and is accessible from `http://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt`. One objective of the statistical analysis was to find a descriptive model explaining which predictor variables are important in disease progression. Ten baseline measurements ($p = 10$), age ($\mathbf{x}_1$), sex ($\mathbf{x}_2$), body mass index ($\mathbf{x}_3$), average blood pressure ($\mathbf{x}_4$), and six blood serum measurements ($\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_7$, $\mathbf{x}_8$, $\mathbf{x}_9$, $\mathbf{x}_{10}$) were obtained from 442 diabetes patients ($n = 442$) in addition to the response variable, a measure of disease progression one year after baseline. The correlation matrix of the predictor variables displays some medium correlation with the highest magnitude of correlation being 0.738.

Since $p = 10$, it becomes time-consuming to run the full search and we only provide the results of GA in addition to those of the four canonical methods.

Table 3.10: Diabetes data

ID	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	response
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{y}$
1	59	2	32.1	101	157	93.2	38	4	4.8598	87	151
2	48	1	21.6	87	183	103.2	70	3	3.8918	69	75
3	72	2	30.5	93	156	93.6	41	4	4.6728	85	141
4	24	1	25.3	84	198	131.4	40	5	4.8903	89	206
5	50	1	23	101	192	125.4	52	4	4.2905	80	135
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1299	85	220
442	36	1	19.6	71	250	133.2	97	3	4.5915	92	57

Table 3.11 displays the BIC scores, tuning parameters and the selected variables of the methods considered. For the BIC score, OLS > ridge regression > Lasso > elastic net. The model chosen by GA is not from any of those four canonical methods but one with the grouping $(3, 2, 2, 2, 4, 4, 2, 3, 2, 4)$. This is the grouping that GA finds the best in 40 minutes (the initialized population size is 200 and the algorithm is stopped after 20 iterations). The selected variables are $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_7$, $\mathbf{x}_9$ which are all classified into category G_2 because of correlation. The rest variables are either classified into category G_3 or G_4 and are shrunk to exactly zero. Lasso selects six variables compared to seven by elastic net. GA seems to like selecting a more parsimonious model (with five variables). Both Lasso and elastic net agree on the demographic variables of sex, body mass index, average blood pressure ($\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$) as important risk factors of disease progression. They also agree on three blood measurements $\mathbf{x}_7$, $\mathbf{x}_9$ and $\mathbf{x}_{10}$ but disagree on the blood measure $\mathbf{x}_5$. The model selected by GA only keeps $\mathbf{x}_7$ and $\mathbf{x}_9$ among the six blood measurements in addition to the common demographic variables selected by both Lasso and elastic net. We also have provided the model estimates in Table 3.12. Note that variables selection among the six blood measurements not only helps obtain better interpretability through a parsimonious model but also can save the costs of collecting irrelevant measurements from patients.

Table 3.11: Diabetes data – methods comparison

Methods	BIC	Parameters	Variables Selected
OLS	3585	Null	All
Ridge	3575	$\lambda_2 = 0.1$	All
Lasso	3570	$k = 7$	(2, 3, 4, 5, 7, 9, 10)
ElasticNet	3568	$k = 6$, $\lambda_2 = 0.1$	(2, 3, 4, 7, 9, 10)
GA	3560	$k = 0$, $\lambda_2 = 0.1$	(2, 3, 4, 7, 9)

$\star$: The grouping chosen by GA is (3, 2, 2, 2, 4, 4, 2, 3, 2, 4).

Table 3.12: Diabetes data – model estimates

Methods	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
OLS	-0.04	-22.86	5.60	1.12	-1.09	0.75	0.37	6.53	68.48	0.28
Ridge	0.0048	-19.75	5.28	1.04	-0.11	-0.11	-0.69	4.27	40.46	0.36
Lasso	0.00	-18.85	5.63	1.02	-0.14	0.00	-0.82	0.00	46.92	0.23
ElasticNet	0.00	-11.69	5.62	0.95	0.00	0.00	-0.82	0.00	42.02	0.21
GA	0.00	-18.98	5.33	1.07	0.00	0.00	-1.00	0.00	41.03	0.00

$\star$: The grouping chosen by GA is $(3, 2, 2, 2, 4, 4, 2, 3, 2, 4)$.

3.6 Conclusion

In this chapter we have proposed a new procedure of model selection and parameter estimation. The objective function of Lasso or elastic net constrains all predictor variables to be regularized by the same penalty and the estimates can suffer from unnecessary bias. To resolve this problem, our objective function allows each independent variable to be either unpenalized or penalized; if penalized, it allows the variable to be selectively penalized by L_1, L_2 or both L_1 and L_2 norms. Such flexibility is the significant reason for the success of our method. We introduce the grouping definition which makes it fairly convenient to identify all the possible model fits. The resulted estimates from our procedure have smaller bias and variance than those from Lasso or elastic net. Since none of the canonical methods dominate under all circumstances, our procedure naturally becomes superior by providing a general and flexible framework covering the universe of all linear models. The comprehensive simulation study further supports the necessity of searching among the possible linear models. Our procedure is a combinatorial optimization problem and becomes time-consuming for even moderate p. As an efficient alternative, genetic algorithm can save computation time tremendously and is able to return an nearly optimal solution within a reasonably amount of time. Given any model selection criteria our procedure can provide an optimal model allowing sparsity and low bias. The chosen model fit is as least good as those provided by the canonical methods and most often can outperform them greatly in accuracy.

Chapter 4

Cauchy Test

In many applications of time series analysis, we model correlation to improve prediction accuracy and to avoid erroneous inferences regarding regression parameters. However, if no correlation is present we add unnecessary model complication and add error to the model fitting process. It is therefore an important step to first determine if significant serial correlation is present in the process under study. For small samples, say only a few years of monthly data, the commonly employed Ljung-Box portmanteau test (Ljung and Box, 1978) for autocorrelation can have low power. It might be helpful to have a more powerful small sample test for detecting significant correlation. In this paper we develop such a test by considering the Cauchy estimator of correlation, rather than the usual sample autocorrelation estimator. The Cauchy estimator has been considered previously in the literature for parameter estimation (Gallagher, 2001), model identification (Gallagher, 2002) and to create confidence intervals for autoregressive parameters (Gallagher and Tunno, 2008, So and Shin, 1999). The proposed test compares favorably with the Box-Pierce/Ljung-Box statistics in detecting autoregressive alternatives.

If initial investigations indicate significant serial correlation, an autoregressive moving average (ARMA) process is typically used to parameterize the correlation function. It is common practice to assess the goodness-of-fit of the ARMA model via a portmanteau test for significant correlation in the ARMA residuals. A plethora of portmanteau tests for correlation and associated ARMA goodness-of-fit tests have been proposed. A brief review of some of the most relevant literature is provided. The sample correlation function of the observed residuals is used in the commonly employed tests due to Box and Pierce (1970) and Ljung and Box (1978). Monti (1994) proposed

a test based on the sample partial correlation function, which performs well in detecting missing moving average components. Recently, other authors have proposed asymmetric tests based on the determinant (McLeod and Mahdi, 2012, Pena and Rodriguez, 2002, 2006) or the trace (Fisher and Gallagher, 2012) of the sample autocorrelation matrix. In this paper we modify the Cauchy test for serial correlation and develop an ARMA goodness-of-fit test. The asymptotic behavior of the test statistic is quantified.

The remainder of the paper is organized as follows. In the next section, we introduce the Cauchy estimator, derive the proposed statistic and give its asymptotic distribution under the null hypothesis of independent identically distributed (iid) data. In Section 4.2 we consider the ARMA goodness-of-fit problem and describe a Cauchy test based on sample residuals. We use simulation to investigate the small sample behavior of the proposed methodology in Section 4.3. Section 4.4 applies our test to monthly asset returns for Facebook, Incorporated. We close the paper with some remarks in Section 4.5.

4.1 Testing for correlation

Let $\{X_t\}$ be a weakly stationary time series observed at times $t = 1, 2, \ldots, n$. Since we can always subtract off the sample mean, without loss of generality we will assume that $\mathrm{E}(X_t) = 0$. Consider the equation

$$X_t = \rho(k)X_{t-k} + Z_t^{(k)} \qquad k = 1, 2, \ldots, m,$$

where $\rho(k)$ is the lag k correlation, and $Z_t^{(k)}$ has mean 0 and variance $\sigma^2 = \mathrm{E}(X_t - \rho(k)X_{t-k})^2$. For example if $\{X_t\}$ follows a first order autoregressive difference equation:

$$X_t = \phi X_{t-1} + \epsilon_t,$$

with $\{\epsilon_t\}$ an iid $N(0, \sigma^2)$ sequence, then $Z_t^{(k)} = X_t - \phi^k X_{t-k}$. If on the other hand $\{X_t\}$ is an iid sequence then for each k, $Z_t^{(k)} = Z_t = X_t$.

Notice that by construction $Z_t^{(k)}$ and X_{t-k} are uncorrelated and $\rho(k)X_{t-k}$ is the best linear predictor of X_t based on X_{t-k}. If $\{X_t\}$ is Gaussian, then

$$E(X_t|X_{t-k}) = \rho(k)X_{t-k}. \tag{4.1}$$

Notice from (4.1), that for any measurable function g,

$$\mathrm{E}\left(X_t g(X_{t-k})\right) = \rho(k) E\left(X_{t-k} g(X_{t-k})\right).$$

In particular, let $g(X_{t-k}) = |X_{t-k}|^r S_{t-k}$, where $r \geq 0$ and $S_t = \mathrm{sign}(X_t)$. Note

$$E\left(X_t |X_{t-k}|^r S_{t-k}\right) = \rho(k) E|X_{t-k}|^{r+1},$$

or

$$\rho(k) = \frac{\mathrm{E}\left(X_t |X_{t-k}|^r S_{t-k}\right)}{E|X_{t-k}|^{r+1}}. \tag{4.2}$$

Taking $r = 1$ corresponds to the usual lag k autocorrelation and other values give correlation between X_t and the signed power, $|X_{t-k}|^r S_{t-k}$. For any independent sequence, $\rho(k) = 0$ for every non-zero k.

In this paper we consider the special case of $r = 0$ which corresponds to the correlation between $\{X_t\}$ and the sequence of signs $\{S_t\}$. For non-Gaussian processes (and for the remainder of this paper) we take $\rho(k)$ to be given by (4.2) with $r = 1$. For any time reversible ARMA process with iid innovations, the function $\rho(k)$ given by (4.2) will satisfy the same set of recursions–and thus takes the same values–as the autocorrelation function. For ARMA processes which are not reversible, we consider correlation between present values of the process and past values of the sign process. To test for significant correlation in any of the first m lags, the set of statistical hypotheses we test is

$$H_0 : \rho(k) = 0 \quad k = 1, 2, \ldots, m \qquad H_1 : \rho^2(k) > 0 \quad \text{for some} \quad 1 \leq k \leq m.$$

Equivalently we can test if the Euclidean norm of the vector $\boldsymbol{\rho} = (\rho(1), \ldots, \rho(m))^t$ is zero. Below we derive a new test statistic for the above set of hypotheses and give its asymptotic distribution under the iid null hypothesis.

To estimate $\rho(k)$ from data we can use moment estimators

$$\hat{\rho}_r(k) = \frac{\sum_{t=k+1}^{n} X_t |X_{t-k}|^r S_{t-k}}{\sum_{t=k+1}^{n} |X_{t-k}|^{r+1}},$$

where $r = 1$ corresponds to the usual sample autocorrelation function and $r = 0$ gives the Cauchy estimator of lag k autocorrelation. Unlike the usual covariance the Cauchy version is not symmetric in its arguments. For the correlation test described here we use only positive values of k, but for ARMA goodness-of-fit test of Section 4.2 we use both positive and negative values of k. Our simulations not included in this paper indicate that using negative values of k does not significantly improve the correlation test, but can provide improvement for the test described in the next section.

Now consider

$$\sum Z_t^{(k)} S_{t-k} = \sum |X_{t-k}| \left(\hat{\rho}_0(k) - \rho(k)\right).$$

For iid data (null hypothesis), the summands on the left hand side form a k-dependent sequence so that

$$(n-k)^{-1/2} \sum Z_t S_{t-k} \Rightarrow N(0, \sigma_k^2),$$

where $\Rightarrow$ denotes weak convergence and $\sigma_k^2 = \sigma^2$ for $k = 1, 2, \ldots, m$. For symmetric iid sequences, the convergence rate is quite fast as the joint distribution of $(Z_{k+1}S_1, \ldots, Z_n S_{n-k})$ is the same as the distribution of $(Z_{k+1}, \ldots, Z_n)$ (see Theorem 3.1 from Gallagher and Tunno (2008)). Furthermore for Gaussian data the distribution of the sum is exactly normal for every n. It is easy to see that under the null hypothesis of iid data, that for each k

$$\sum Z_t S_{t-k} = \hat{\rho}_0(k) \sum |X_{t-k}| = \sum X_t S_{t-k},$$

and since for any t, s,

$$\mathrm{Cov}(Z_t S_{t-k}, Z_s S_{s-l}) = \begin{cases} \mathrm{Var}(Z_t) & t = s; k = l \\ 0 & \text{otherwise,} \end{cases}$$

the Cramér-Wold device provides the following result.

Theorem 4.1.1. *If $\{X_t\}$ is an iid sequence, H_0 holds and*

$$\left(\frac{\sum X_t S_{t-1}}{\sqrt{n-1}}, \ldots, \frac{\sum X_t S_{t-m}}{\sqrt{n-m}}\right) \Rightarrow N_m(\mathbf{0}, \mathbf{D}),$$

where $\mathbf{D}$ *is a diagonal matrix with* k^{th} *element*

$$\sigma_k^2 = \sigma^2.$$

We note here that if the data is iid and Gaussian, then each component of the vector is marginally normally distributed and the components are uncorrelated. However, the components are not independent and the joint normality only holds in the limit even in that case. To test our set of hypothesis we can use the Euclidean norm of the vector of the standardized sample Cauchy correlations,

$$\lambda(k) = \sum X_t S_{t-k} / \sqrt{(n-k)\hat{\sigma}_k^2},$$

where $\hat{\sigma}_k$ is any consistent estimator of σ_k. Using standard Slutsky and mapping arguments, we have the following result.

Corollary 4.1.2. *If $\{X_t\}$ is iid, H_0 holds and*

$$\sum_{k=1}^{m} (\lambda(k))^2 \Rightarrow \chi_m^2,$$

where χ_m^2 denotes a chi-square random variable with m degrees of freedom.

The variance σ_k^2 could be estimated under the null hypothesis by using the observed values of $\{X_t\}$, but to gain power in our simulations below we estimate σ_k under the alternative hypothesis using

$$\hat{\sigma}_k^2 = \sum_{k+1}^{n} (X_t - \hat{\rho}_1(k) X_{t-k})^2 / (n-k). \tag{4.3}$$

Here we use the sample correlation to estimate ρ as it has smaller mean squared error than the Cauchy estimator. We propose using the test statistic from Corollary 4.1.2 to test an iid null hypothesis against a general portmanteau alternative of serial correlation.

To investigate the theoretical power of the proposed test consider a stationary linear process

satisfying

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}; \quad \{\epsilon_t\} \sim iid(0, \sigma^2); \quad \sum |\psi_j| < \infty. \tag{4.4}$$

Recall that all causal invertible ARMA processes satisfy (4.4).

Theorem 4.1.3. *Let $\{X_t\}$ satisfy (4.4). For each fixed positive integer m, as $n \to \infty$,*

$$n^{-1/2} \left(\sum_{t=2}^{n} X_t S_{t-1} - n\rho(1)E|X_0|, \ldots, \sum_{t=m}^{n} X_t S_{t-m} - n\rho(m)E|X_0| \right)^t \Rightarrow N_m(\mathbf{0}, \mathbf{V}),$$

where N_m denotes a multivariate normal on m dimensions and the matrix $\mathbf{V}$ has elements

$$V_{i,j} = E(X_i^2 S_0 S_{i-j} - \rho(i)\rho(j)) \left(E|X_0|\right)^2.$$

Proof. For arbitrary vector $\mathbf{a} = (a_1, \ldots, a_m)^t$, let $\varepsilon_t = \sum_{j=1}^{m} a_j \left(X_t S_{t-j} - \rho(j)E|X_0|\right)$. Also, let $\mathcal{F}_t$ denote the sigma field generated by past values of the innovations $\{\epsilon_t, \epsilon_{t-1}, \ldots\}$. The sequence $\{\varepsilon_t\}$ is a (stationary and ergodic) square integrable martingale-difference sequence with respect to filtration $\{\mathcal{F}_t\}$, with

$$\eta^2 = E(\varepsilon^2) = \sum_{i,j \le m} a_i a_j E\left((X_t S_{t-i} - \rho(i)E|X_0|)(X_t S_{t-j} - \rho(j)E|X_0|)\right) = \mathbf{a^t V a}.$$

Using Theorem 18.3 in Billingsley (1999) we conclude that

$$n^{-1/2} \sum_{t=1}^{n} \varepsilon_t / \eta \Rightarrow N(0, 1).$$

Our result now follows from the Cramér-Wold device. □

From the above theorem, we see that under ARMA alternatives, the summands in the proposed test behave like

$$\frac{n-k}{n} \hat{\sigma}^2 \lambda_k^2 = n^{-1} \left(\sum X_t S_{t-k} - n\rho(k)\mu + n\rho(k)\mu \right)^2 = W^2 + 2W\sqrt{n}\rho(k)\mu + n\rho^2(k)\mu^2,$$

where $\mu = E|X_t|$, and W converges weakly to a mean zero normal random variable. It follows that for $\rho_n(k) = cn^{\gamma}$ with $c \neq 0$, the proposed test statistic will diverge for any $\gamma > -1/2$, i.e., the proposed test has non-trivial asymptotic power in detecting local alternatives converging to the null

at the $n^{-1/2}$ rate.

4.2 Detecting ARMA underfit

The test described in the previous section can be modified to detect the goodness-of-fit of a causal and invertible ARMA model:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q}, \tag{4.5}$$

where $\{z_t\}$ is an iid sequence with zero mean and variance σ^2 with $P(z_t = 0) = 0$. Toward this end, consider fitting model (4.5) using Gaussian maximum likelihood or any estimation procedure with the same asymptotic behavior as the least squares estimation. Now let $\{\hat{z}_t\}$ be the residuals from a fitted ARMA model and $\hat{s}_t = \text{sign}(\hat{z}_t)$. For $k > 0$, consider

$$\hat{\lambda}(k) = \frac{\sum_{t=k+1}^{n} \hat{z}_t \hat{s}_{t-k}}{\hat{\sigma}\sqrt{n-k}} \quad \text{and} \quad \hat{\lambda}(-k) = \frac{\sum_{t=k+1}^{n} \hat{s}_t \hat{z}_{t-k}}{\hat{\sigma}\sqrt{n-k}},$$

where $\hat{\sigma}$ is a consistent estimator of σ.

Theorem 4.2.1. *If $\{X_t\}$ follows model (4.5), then as $n \to \infty$,*

$$\left(\hat{\lambda}(-1), \ldots, \hat{\lambda}(-m), \hat{\lambda}(1), \ldots, \hat{\lambda}(m)\right)' \Rightarrow N_m(\mathbf{0}, \mathbf{W}).$$

Here

$$\mathbf{W} = \begin{pmatrix} \mathbf{I} & \delta(\mathbf{I} - \mathbf{H}) \\ \delta(\mathbf{I} - \mathbf{H}) & \mathbf{I} - \delta\mathbf{H} \end{pmatrix},$$

where $\mathbf{I}$ denotes the $m \times m$ identity matrix, $\mathbf{H}$ is a rank $p+q$ projection matrix, and $\delta = (\mathrm{E}|\mathrm{z_t}|)^2/\mathrm{E}(\mathrm{z_t^2})$.

Proof. The result follows from arguments similar to those in McLeod (1978). To demonstrate the idea, we give the proof for causal autoregressive models. The proof for general ARMA models is similar but requires additional notation as in McLeod (1978). Let $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ denote vectors of the true and estimated parameters, respectively. Below we use the causal representation

$$x_t = \sum_{j=0}^{\infty} \psi_j z_{t-j}.$$

For $s_t = \text{sign}(z_t)$ and $k > 0$, consider

$$\lambda(k) = \frac{\sum_{t=k+1}^{n} z_t s_{t-k}}{\sigma\sqrt{n-k}} \quad \text{and} \quad \lambda(-k) = \frac{\sum_{t=k+1}^{n} s_t z_{t-k}}{\sigma\sqrt{n-k}}.$$

The idea of the proof is to approximate $\hat{\lambda}$ with λ. Simple Taylor expansion arguments and the assumption that $P(z_t = 0) = 0$, indicate that as $n \to \infty$, $\hat{s}_t$ can be replaced with s_t. The consistency of $\hat{\sigma}$ and Slutsky arguments justify replacing $\hat{\sigma}$ with σ. Combining these approximations with algebra results in

$$\hat{\lambda}(k) \approx \lambda(k) - \sigma^{-1}\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})'\hat{\boldsymbol{\gamma}}, \tag{4.6}$$

where the $p \times 1$ vector $\hat{\boldsymbol{\gamma}}$ has i^{th} component $\hat{\gamma}_i = \sum x_{t-i}s_{t-k}/\sqrt{n(n-k)}$, which converges in the almost sure sense to $\gamma_i = \text{E}(x_{t-i}s_{t-k}) = \psi_{k-i}E|z_t|$. We note here that for $\ell < 0$, $\psi_{\ell-i} = 0$, so that

$$\begin{pmatrix} \hat{\lambda}(-1) \\ \vdots \\ \hat{\lambda}(-m) \end{pmatrix} = \begin{pmatrix} \lambda(-1) \\ \vdots \\ \lambda(-m) \end{pmatrix} + o_p(1). \tag{4.7}$$

We now turn our attention to $\hat{\lambda}(k)$ for $k > 0$. It is well known (**?**) that

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) = \sigma^{-2}\mathbf{V}^{-1}\mathbf{v} + O_p(1/\sqrt{n}), \tag{4.8}$$

where $\mathbf{V}^{-1}/n$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\phi}}$ and vector $\mathbf{v}$ has j^{th} component

$$v_j = n^{-1/2}\sum z_t x_{t-j}.$$

Combining (4.6) and (4.8) we have

$$\begin{pmatrix} \hat{\lambda}(1) \\ \vdots \\ \hat{\lambda}(m) \end{pmatrix} = \begin{pmatrix} \lambda(1) \\ \vdots \\ \lambda(m) \end{pmatrix} - \frac{\sqrt{\delta}}{\sigma^2}\mathbf{X}\mathbf{V}^{-1}\mathbf{v} + o_p(1), \tag{4.9}$$

where the matrix $\mathbf{X}_{m\times p}$ is as in Box and Pierce (1970) and McLeod (1978) and only depends on the sequence $\{\psi_j\}$. Thus ignoring terms that are $o_p(1)$, any linear combination of $\hat{\lambda}(-m), \ldots, \hat{\lambda}(m)$ can

be shown to converge to a normal distribution using martingale difference central limit theory. It follows that the vector in Theorem 4.2.1 converges in distribution to a multivariate normal random variable. It remains to determine the asymptotic covariance matrix. An easy calculation shows that for $|k| \neq |l|$, $\text{cov}(\lambda(k), \lambda(l)) = 0$, and $\text{cov}(\lambda(-k), \lambda(k)) = \delta$. From (4.7) it follows that

$$\text{var}\left((\hat{\lambda}(-m), \ldots, \hat{\lambda}(-1))'\right) \to \mathbf{I}.$$

Using (4.9) and calculations very similar to McLeod (1978),

$$\text{cov}\left((\lambda(1), \ldots, \lambda(m))', \sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right) \to \sqrt{\delta}\mathbf{X}\mathbf{V}^{-\mathbf{1}}.$$

The term δ arises in the equation because

$$\text{cov}(\sum z_t s_{t-k}, \sum z_t x_{t-i}) = n\sigma^2 \psi_{k-i} E|z_t|.$$

For $\boldsymbol{\lambda}_+ = (\lambda(1), \ldots, \lambda(m))'$, we have

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\lambda}}_+) &= \text{var}(\boldsymbol{\lambda}_+) - \text{cov}(\boldsymbol{\lambda}_+, \sqrt{\delta}\sigma^{-2}\mathbf{X}\mathbf{V}^{-1}\mathbf{v}) \\
&\quad -\text{cov}(\sqrt{\delta}\sigma^{-2}\mathbf{X}\mathbf{V}^{-1}\mathbf{v}, \boldsymbol{\lambda}_+) + \delta\text{var}(\mathbf{X}\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \\
&= \mathbf{I} - 2\delta\mathbf{X}\mathbf{V}^{-1}\mathbf{X}' + \delta\mathbf{X}\mathbf{V}^{-1}\mathbf{X}' \\
&= \mathbf{I} - \delta\mathbf{X}\mathbf{V}^{-1}\mathbf{X}'.
\end{aligned}$$

As pointed out in Box and Pierce (1970) and McLeod (1978), for large m the matrix $\mathbf{X}\mathbf{V}^{-1}\mathbf{X}'$ can be approximated by projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, which has rank $p + q$. Thus

$$\text{var}\left((\hat{\lambda}(1), \ldots, \hat{\lambda}(m))'\right) = \mathbf{I} - \delta\mathbf{H}.$$

Finally, for $\boldsymbol{\lambda}_- = (\lambda(-m), \ldots, \lambda(-1))'$ arguments similar to those above yield

$$\text{cov}\left(\hat{\boldsymbol{\lambda}}_-, \hat{\boldsymbol{\lambda}}_+\right) \approx \text{cov}\left(\boldsymbol{\lambda}_-, \boldsymbol{\lambda}_+ - \sqrt{\delta}\mathbf{X}\mathbf{V}^{-1}\mathbf{v}/\sigma^2\right) \to \delta(\mathbf{I} - \mathbf{H}).$$

□

The following corollary can be used to develop a goodness-of-fit test using $\hat{\lambda}(k)$.

Corollary 4.2.2. *If $\{X_t\}$ follows model (4.5), then as $n \to \infty$,*

$$\sum_{k=-m}^{-1} \hat{\lambda}(k)^2 + \sum_{k=1}^{m} \hat{\lambda}(k)^2 \Rightarrow Y_1 + (1+\delta)Y_2 + (1-\delta)Y_3,$$

where Y_1, Y_2, and Y_3 are independent chi-square distributed random variables with $p+q$, $m-p-q$, and $m-p-q$ degrees of freedom, respectively.

Proof. Using Theorem 4.2.1 and Box (1954), the quadratic form $\sum_{k=-m}^{-1} \hat{\lambda}(k)^2 + \sum_{k=1}^{m} \hat{\lambda}(k)^2$ has asymptotic distributional representation $\sum_{i=1}^{2m} c_i \chi_i^2$, where $\{\chi_i^2\}$ are iid chi-square distributed random variables and $c_1, \ldots, c_{2m}$ are eigenvalues of $\mathbf{W}$. The projection matrix $\mathbf{H}$ has $p+q$ eigenvalues which are 1 and $m-p-q$ eigenvalues of 0. Let $\mathbf{u}$ and $\mathbf{e}$ denote eigenvectors with $\mathbf{Hu} = \mathbf{0}$ and $\mathbf{He} = \mathbf{e}$. For each choice of $\mathbf{u}$ and $\mathbf{e}$, we can create the following orthonormal eigenvectors for $\mathbf{W}$:

$$(\mathbf{0}, \mathbf{e})'; \quad (\mathbf{e}, \mathbf{0})'; \quad (\mathbf{u}/\sqrt{2}, -\mathbf{u}/\sqrt{2})'; \quad (\mathbf{u}/\sqrt{2}, -\mathbf{u}/\sqrt{2})'.$$

The first two eigenvectors have eigenvalues of 1, the third eigenvector has an associated value of $1-\delta$ and the fourth vector has an eigenvalue of $1+\delta$. Note that using the Cauchy-Schwarz inequality it is easy to see that $1-\delta > 0$, so that all eigenvalues are positive. □

4.3 Simulation Study

In this section we summarize simulation results which give an idea of the small sample behavior of the proposed Cauchy portmanteau test. We present simulation results demonstrating situations for which the Cauchy test can provide some improvement over other methods. Typically in statistical software, decision rules for a test are based on quantiles of the relevant asymptotic distribution. Here we study our statistic using both the the asymptotic distribution and Monte Carlo methods (see McLeod and Lin (2006)). Our simulations were executed in a parallel framework utilizing the `snow` package within the R-project.

We compare the proposed method to the commonly used Ljung-Box and Monti statistics. Ljung Box is chosen over Box Pierce since it is known to have better small sample performance and is typically utilized by practitioners. Monti is chosen since it is based on the partial autocorrelation

function, is known to be more powerful in detecting underfit moving-average process, and provides another comparison. Recently, several asymmetric portmanteau test (weighted variants) have been introduced to the the statistics literature. In theory, an asymmetric version of the Cauchy estimator test can be developed as ultimately the statistic is a quadratic form. The asymptotic distribution would be similar to that of McLeod and Mahdi (2012) and Fisher and Gallagher (2012). As discussed in the literature, the results of Pena and Rodriguez (2002) and McLeod and Mahdi (2012) are essentially asymmetric Monti tests and the results of Fisher and Gallagher (2012) are an asymmetric Ljung Box; for brevity we exclude comparisons with the asymmetric tests. The main theme is that for small samples the proposed statistic can more easily detect autocorrelation than portmanteau methods based on sample autocorrelation. Also, the Cauchy-based test statistic performs comparably well as the sample size increases.

4.3.1 Empirical Size

We begin by studying the empirical size of the various statistics in finite samples. One crucial aspect of testing for correlation in a time series is the selection of the lag m at which to test. No optimal lag is known and many suggestions have been made in the literature. In these studies, we utilize $m = \sqrt{n}$ and $m = n/3$ based on some rules of thumb in the literature. As discussed around Corollary 4.1.2, any consistent estimator for the standard deviation can be used in our Cauchy test statistic. For white noise, the sample mean squared error (MSE) on the series works well, but we can gain power using the estimator given in (4.3). Table 4.1 provides the empirical size for the statistics under study when the data follows an iid $N(0, 1)$ process of different sample sizes ranging from one-year of monthly data to eight-years. We see that when the variance term is measured

Table 4.1: Empirical size, based on 10,000 simulated series, of tests when data follows a white noise process and p-values are calculated using the asymptotic distribution

	$n = 12$		$n = 24$		$n = 36$		$n = 48$		$n = 96$	
	m=3	m=4	m=5	m=8	m=6	m=12	m=7	m=16	m=10	m=32
Cauchy Test (MSE)	.0418	.0476	.0503	.0573	.0553	.0636	.0549	.0676	.0577	.0775
Cauchy Test (4.3)	.1145	.1299	.0817	.1000	.0789	.0979	.0754	.0969	.0678	.0954
Ljung Box	.0513	.0602	.0547	.0660	.0579	.0762	.0574	.0835	.0618	.0910
Monti	.0616	.0678	.0533	.0528	.0583	.0556	.0523	.0491	.0530	.0447

with (4.3), the statistic has fairly liberal Type I error performance. When the variance is estimated using the sample MSE, the empirical size is more satisfactory. Overall the size of the Cauchy testing

utilizing (4.3) improves as the sample size increases. The simulation results for Ljung Box follow its well known behavior, slightly liberal both for small samples and as m increases. The Monti statistic appears to have a more consistent size but is slightly liberal at smaller sample sizes.

In the case of a fitted ARMA model, it makes sense to utilize (4.3) to estimate the variance so we do so here. Table 4.2 provides the empirical size for properly fit AR(1) models under a variety of sample sizes. The statistics from Ljung Box and Monti are liberal for small sample sizes but generally

Table 4.2: Empirical size, based on 10,000 simulated series, of tests under properly fit AR(1) models utilizing the respective asymptotic distributions

	$n=12$		$n=24$		$n=36$		$n=48$		$n=96$	
AR(1) Models	m=3	m=4	m=5	m=8	m=6	m=12	m=7	m=16	m=10	m=32
$\phi = .3$										
Cauchy Test (CT)	.0784	.0996	.0575	.0798	.0527	.0779	.0534	.0765	.0509	.0834
Ljung Box (LB)	.0599	.0561	.0494	.0542	.0478	.0604	.0512	.0661	.0508	.0739
Monti (MT)	.0784	.0826	.0614	.0606	.0614	.0553	.0603	.0502	.0544	.0423
$\phi = .6$										
Cauchy Test (CT)	.0735	.0916	.0627	.0810	.0537	.0813	.0517	.0781	.0484	.0841
Ljung Box (LB)	.0644	.0598	.0582	.0631	.0544	.0622	.0552	.0646	.0534	.0774
Monti (MT)	.0780	.0862	.0672	.0599	.0571	.0522	.0556	.0480	.0561	.0419
$\phi = .9$										
Cauchy Test (CT)	.0849	.0959	.0751	.0905	.0593	.0882	.0540	.0881	.0554	.0867
Ljung Box (LB)	.0655	.0623	.0636	.0696	.0584	.0681	.0576	.0721	.0600	.0775
Monti (MT)	.0822	.0913	.0661	.0611	.0652	.0512	.0601	.0472	.0538	.0387

improve. The proposed Cauchy test appears to be a little more liberal in some cases but overall seems comparable. The performance of all the tests improve as the sample size increases, however all the statistics exhibit some sensitivity for larger values of m. Fortunately, modern computing techniques makes this issue nearly moot as critical points and p-values can be determined from the true distribution via Monte Carlo methods. Furthermore, by utilizing the Monte Carlo distribution a practitioner does not have to be concerned with deciding whether to use (4.3) or the MSE in the white noise null hypothesis test. To improve power performance, we recommend (4.3) be used in the Cauchy estimator test.

As aforementioned, computer power has continued to grow at an exponential rate over the years. If a practitioner is not satisfied with the Type I performance based on the asymptotic null distribution, Monte Carlo methods can be utilized using the algorithm in McLeod and Lin (2006). Here we study the performance of the statistics when the p-values are determined using the Monte Carlo algorithm. For the sake of practicality in our simulations, we perform 1000 replicates where each p-value is calculated using $T = 1000$ iterations in the Monte Carlo algorithm. As expected, we see in Table 4.3 that all the statistics have satisfactory Type I error when p-values are determined

Table 4.3: Empirical size, based on 1,000 simulated series, of tests under study for White Noise and properly fit AR(1) models when the Monte Carlo distribution is utilized for p-values

		$n = 12$		$n = 24$		$n = 36$		$n = 48$		$n = 96$	
		m=3	m=4	m=5	m=8	m=6	m=12	m=7	m=16	m=10	m=32
	CT	.048	.055	.045	.046	.046	.040	.050	.051	.056	.049
WN	LB	.053	.064	.050	.041	.047	.039	.054	.053	.059	.054
	MT	.056	.065	.049	.046	.041	.045	.062	.049	.047	.047
AR(1)	CT	.042	.048	.049	.049	.043	.045	.060	.050	.045	.036
$\phi = .3$	LB	.046	.042	.039	.050	.049	.052	.060	.053	.048	.041
	MT	.039	.050	.037	.044	.049	.055	.057	.050	.054	.040
AR(1)	CT	.043	.043	.053	.055	.050	.040	.049	.061	.053	.059
$\phi = .6$	LB	.055	.044	.052	.056	.054	.049	.049	.050	.052	.054
	MT	.044	.039	.048	.061	.054	.051	.048	.046	.046	.051
AR(1)	CT	.050	.053	.056	.054	.058	.061	.061	.061	.046	.051
$\phi = .9$	LB	.046	.061	.055	.059	.057	.049	.059	.055	.059	.052
	MT	.046	.046	.045	.043	.055	.057	.056	.048	.067	.039

through Monte Carlo methods. Since when utilizing the asymptotic distribution, all the statistics under study have Type I error concerns in small samples, we will utilize the Monte Carlo algorithm throughout the remainder of this article and recommend it in practice.

4.3.2 Empirical Power

In general our simulations show the proposed Cauchy test is comparable to the Ljung-Box test and can be more powerful in some cases. In our first study we attempt to detect a generated AR(1) with $\phi = 0.9$. We let the sample size n grow from 6 to 36 and calculate the test statistics at lag $m = n/3$. The p-value of each statistic is calculated using the algorithm of McLeod and Lin (2006) and the empirical power is determined as the proportion of rejections at $\alpha = 0.05$. Figure 4.1 provides a plot of relative power in detecting correlation in first-order autoregressive data as a function of the sample size. We see that the proposed test provides more power at small samples than the other portmanteau procedures and that they are comparable at larger sample sizes. We note here that this example is not selected as all encompassing or even typical, but rather is an example for which the proposed test performs well. For example, for underfit moving average models, tests based on the partial correlations (such as Monti's test) generally will outperform the proposed Cauchy-based test. However, a partial Cauchy correlation test might compete in this case although we do not pursue this here.

A similar study is displayed in Figure 4.2. Here, an AR(2) with $\phi_1 = 0.1$ and $\phi_2 = 0.8$ is inadequately fit with an AR(1). The power of the three statistics is studied as a function of n with lag $m = n/3$. We see that the Cauchy test is comparable to the Ljung Box statistic with slight

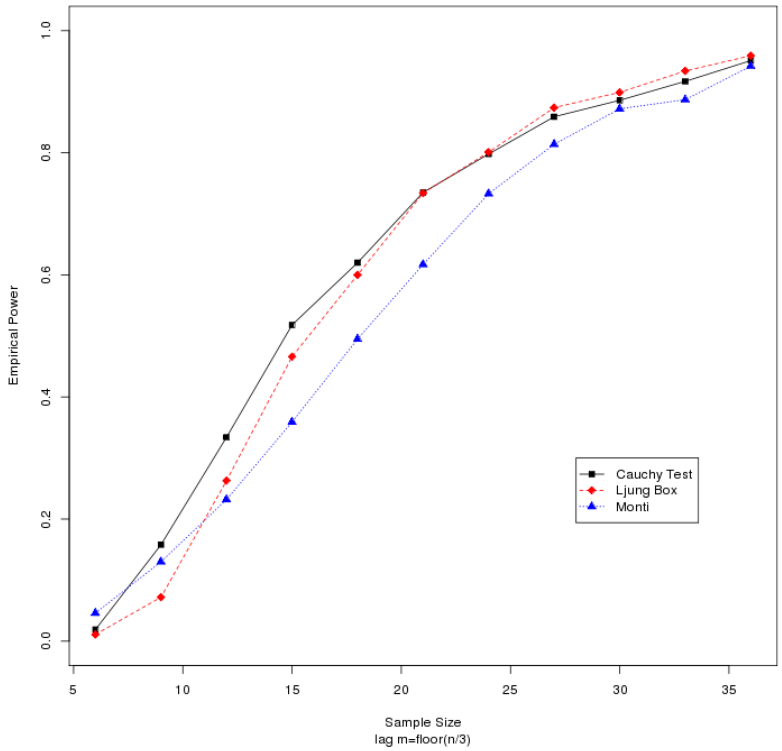


Figure 4.1: Empirical power at $\alpha = 0.05$ for detecting AR(1) with $\phi = 0.9$

improvement in a few cases. Both statistics appear to dominate the Monti test which is generally better at detecting underfit moving average components.

Lastly, we explore the power performance on a larger order ARMA model. Data is generated from an ARMA(2,1) with parameters $\phi_1 = 0.2$, $\phi_2 = 0.7$ and $\theta = -0.5$ and underfit with an AR(1). The statistics are calculated using the Monte Carlo algorithm at $m = \sqrt{n}$ and the sample size increases from $n = 12$ to $n = 48$ in increments of 3. Figure 4.3 displays the empirical power of the statistics under study. As demonstrated in the figure, the proposed Cauchy estimator test provides the most powerful method for some small n values, is comparable to Ljung Box for moderate sample sizes and is eventually surpassed by the Ljung Box statistic as n-grows.

The small sample behavior of the proposed test is exemplified in Table 4.4. Here a series of size $n = 18$ from an AR(2) with $\phi_1 = 0.3$ and $\phi_2 = 0.6$ is generated. For each simulated sample three 0.05-level tests are performed. First, we test for correlation in the simulated process; these power results are in the first column of the table. Second, we improperly fit an AR(1) model to

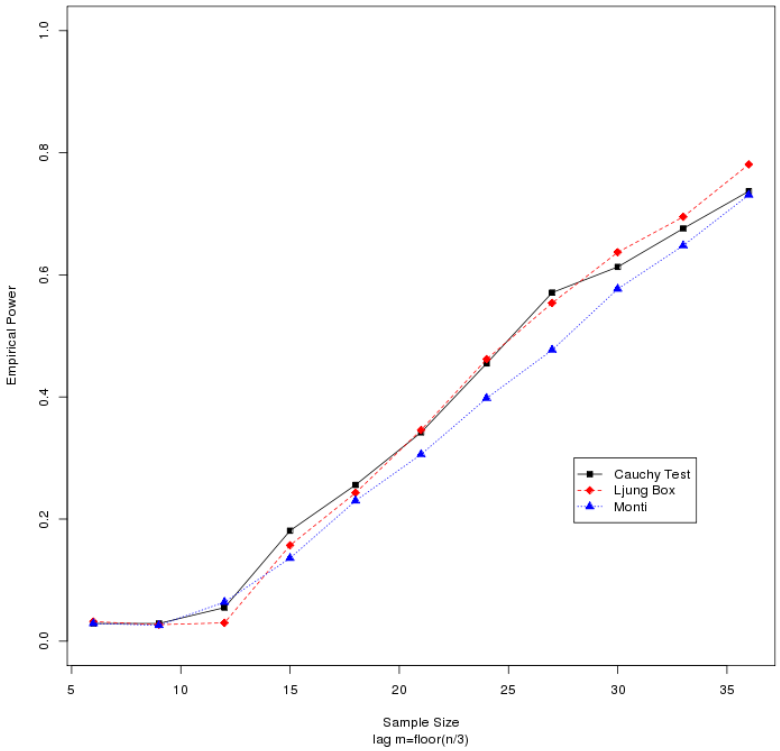


Figure 4.2: Empirical power at $\alpha = 0.05$ for AR(2) underfit as AR(1), $\phi_1 = 0.1$, $\phi_2 = 0.8$

the data; these power results appear in the second column of the table. Third, we fit the correct order AR(2) model; these Type I error results appear in the last column of the table. We see

Table 4.4: Empirical power and size at $\alpha = 0.05$ in fitting an AR(2) model, $n = 18$, $m = 6$

Method	No fit	AR(1) fit	Correct fit
Cauchy Test	.300	.118	.054
Ljung Box	.276	.121	.054
Monti	.201	.103	.057

that the Cauchy-based test provides the highest power in detecting the correlation in the simulated autoregressive process. The proposed test also compares favorably to the other methods when the model is underfit with an AR(1).

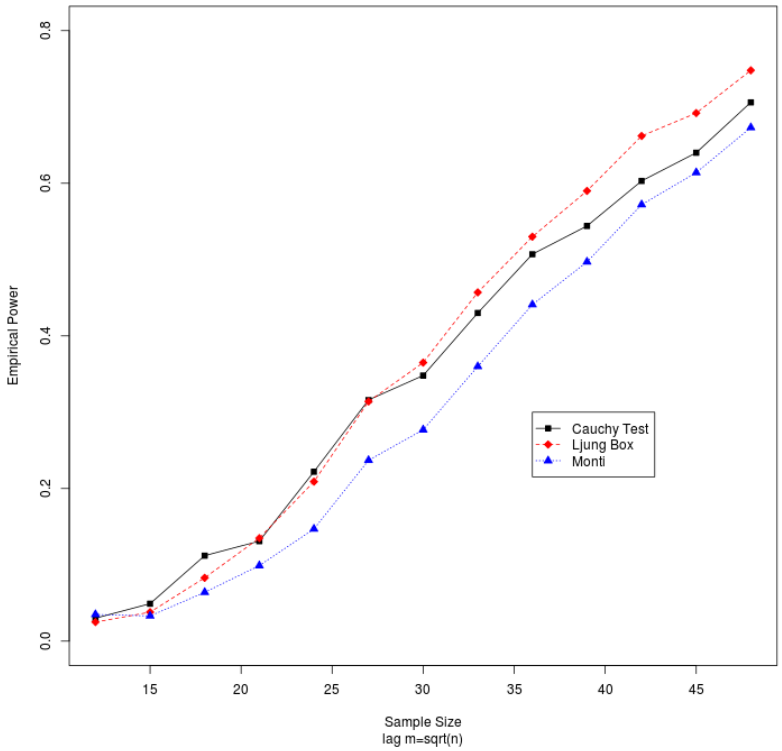


Figure 4.3: Empirical power at $\alpha = 0.05$ for ARMA(2,1) underfit with AR(1), $\phi_1 = 0.2$, $\phi_2 = 0.7$, $\theta = -0.5$

4.4 Data Analysis

The simulation results indicate the proposed Cauchy test can be more powerful in detecting autocorrelation in small sample time series. Here we utilize the statistics to check for serial correlation in the log monthly returns of Facebook, Inc. (FB) stock. Facebook is a very popular social media website that had a highly publicized Initial Public Offering (IPO) on May 18, 2012. We consider the IPO to be the start of the series and look at the first trading day of each month since to comprise our monthly series. The log monthly returns comprise only $n = 11$ observations at the time of the authoring of this article. Table 4.5 provides the p-values (based on the Monte Carlo distribution) for the proposed statistic, Ljung-Box and Monti statistics at lags 3 (roughly $\sqrt{n}$) and 4 (roughly $n/3$). We see all the statistics are insignificant at $\alpha = 0.05$ and both lags. Given the results from the simulation study, this suggests that the Facebook monthly data does not contain serial correlation.

Table 4.5: Detecting serial correlation in Facebook, Inc. monthly returns, p-values via Monte Carlo methods

	m=3	m=4
Cauchy Test	0.3906	0.5624
Ljung Box	0.8317	0.7303
Monti	0.8703	0.7033

4.5 Summary

This paper introduces a new portmanteau test for correlation and a new ARMA goodness-of-fit test. Both tests are based on the Cauchy estimator of correlation. The asymptotic behavior of the tests are quantified and studied via simulation. Our simulations indicate, like that of the statistics in the literature, the asymptotic distribution can lead to inflated Type I error rates for smaller sample sizes. As such, we recommend Monte Carlo methods like those described in McLeod and Lin (2006) be utilized to determine critical points and p-values. To improve the power performance, we suggest the variance term of our proposed method be estimated with equation (4.3).

Like tests based on the residual sample autocorrelation, the proposed test is better at picking up underfit autoregressive processes than detecting underfit moving average components. Perhaps a version with Cauchy partial correlations, similar to Monti (1994), would provide more power in detecting missing moving average components. It is possible that an asymmetric test could be constructed based on the matrix of Cauchy correlations similar to McLeod and Mahdi (2012), Pena and Rodriguez (2002, 2006) or Fisher and Gallagher (2012). A test of this type might provide higher power than the tests proposed in this paper.

Bibliography

Aggarwal, R., Inclan, C., and Leal, R. (1999). Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis*, 34(1):33–55.

Anderson, R. L. (1942). Distribution of the serial correlation coefficients. *The Annals of Mathematical Statistics*, 13(1):1–13.

Anderson, T. W. and Walker, A. M. (1964). On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process. *The Annals of Mathematical Statistics*, 35(3):1296–1303.

Andy, C. C. K. and Wu, Y. (1997). Further results on the finite-sample distribution of Monti's portmanteau test for the adequacy of an ARMA(p, q) model. *Biometrika*, 84(3):733–736.

Bates, B. C., Chandler, R. E., and Bowman, A. W. (2012). Trend estimation and change point detection in individual climatic series using flexible regression methods. *Journal of Geophysical Research: Atmospheres*, 117(D16):n/a–n/a.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366.

Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2):721–741.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability & Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition.

Bleakley, K. and Vert, J.-P. (2011). The group fused Lasso for multiple change-point detection. Technical Report HAL-00602121, Bioinformatics Center.

Booker, L. B. (1987). Improving search in genetic algorithms. In *Genetic Algorithms and Simulated Annealing*. Morgan Kaufmann Publishing.

Bowman, A., Pope, A., and Ismail, B. (2006). Detecting discontinuities in nonparametric regression curves and surfaces. *Statistics and Computing*, 16(4):377–390.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25(2):290–302.

Box, G. E. P. and Pierce, D. A. (1970). Distribution of residual autocorrelation in autoregressive-integrated moving average time series models. *Journal of the American Statistical Assocciation*, 65(332):1509–1526.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Sciences*, 16(3):199–231.

Bremermann, H. J., Rogson, M., and Salaff, S. (1966). Global properties of evolution processes. In *Natural Automata and Useful Simulations*, pages 3–41. Spartan Books, Washington, DC.

Caussinus, H. and Lyazrhi, F. (1997). Choosing a linear model with a random number of change-points and outliers. *Annals of the Institute of Statistical Mathematics*, 49(4):761–775.

Caussinus, H. and Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C*, 53(3):405–425.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability & Statistics: Probability and Statistics. Wiley.

Davis, R. A., Lee, T. C. M., and Rodriguez-yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika*, 37(3/4):409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika*, 38(1/2):159–179.

Easterling, D. R. and Peterson, T. C. (1995). A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15(4):369–377.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Efroymson, M. A. (1965). *Multiple regression analysis*. New York: Wiley.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fisher, T. J. and Gallagher, C. M. (2012). New weighted portmanteau statistics for time series goodness of fit testing. *Journal of the American Statistical Association*, 107(498):777–787.

Fogarty, T. C. (1989). Varying the probability of mutation in the genetic algorithm. In *Proceedings of the third international conference on genetic algorithms*, pages 104–109, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Friedman, J., Hastie, T., and Tibshirani, R. (2008a). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J., Hastie, T., and Tibshirani, R. (2008b). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441.

Gallagher, C., Lund, R., and Robbins, M. (2013). Changepoint detection in climatic time series with long-term trend. *Journal of Climate*, 26:4994–5006.

Gallagher, C. M. (2001). A method for fitting stable autoregressive models using the autocovariation function. *Statist. Probab. Lett.*, 53(4):381–390.

Gallagher, C. M. (2002). Order identification for gaussian moving averages using the covariation. *J. Stat. Comput. Simul.*, 72(4):279–284.

Gallagher, C. M. and Tunno, F. (2008). A small sample confidence interval for autoregressive parameters. *J. Statist. Plann. Inference*, 138(12):3858–3868.

Gao, J., Kwan, P. W., and Shi, D. (2010). Sparse kernel learning with Lasso and Bayesian inference algorithm. *Neural Networks*, 23(2):257–264.

Garside, M. J. (1971). Some computational procedures for the best subset problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 20(1):8–15.

Glover, F. and Kochenberger, G. A. (2003). *Handbook of Metaheuristics*. Kluwer Academics.

Goodman, E. D. (2009). Introduction to genetic algorithms.

Goossens, C. and Berger, A. (1987). How to recognize an abrupt climatic change? In *Abrupt Climatic Change*, volume 216 of *NATO ASI Series*, pages 31–45. Springer Netherlands.

Hannart, A. and Naveau, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics*, 54(3):256–268.

Harchaoui, Z. and Levy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, 37(3):323–341.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, 56(3):495–504.

Hinkley, D. V. (1971). Inference in two phase regression. *Journal of American Statistics Association*, 66(336):736–743.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Horowitz, J. L., Lobato, I. N., Nankervis, J. C., and Savin, N. E. (2006). Bootstrapping the Box-Pierce Q test: A robust test of uncorrelatedness. *Journal of Econometrics*, 133(2):841–862.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.

Jenkins, G. M. and Box, G. E. P. (1970). *Time Series Analysis: Forecasting and Control*. Wiley series in probability and statistics. Wiley.

Jong, K. D. (1975). *An analysis of the behavior of a class of genetic adaptive systems*. PhD thesis, University of Michigan.

Kim, H. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3):409–423.

Lavielle, M. and Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306.

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662.

Ljung, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, 73(3):725–730.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lu, Q., Lund, R., and Lee, T. C. M. (2010). An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1):299–319.

Lund, R. and Reeves, J. (2002). Detection of undocumented changepoints: a revision of the two-phrase regression model. *Journal of Climate*, 15:2547–2554.

Lund, R., Wang, X. L., Lu, Q., Reeves, J., Gallagher, C., and Feng, Y. (2007). Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, 20:5178–5190.

Macneill, I. (1974). Tests for change of parameter at unknown times and distribtions of some related functionals of Brownian Motion. *The Annals of Statistics*, 2(5):950–962.

Mauget, S. A. (2003). Multidecadal regime shifts in U.S. streamflow, precipitation, and temperature at the end of the twentieth century. *Journal of Climate*, 16(23):3905–3916.

McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):296–302.

McLeod, A. I. and Lin, J. W. (2006). Improved Pena-Rodriguez portmanteau test. *Computational Statistics and Data Analysis*, 51(3):1731–1738.

McLeod, I. A. and Mahdi, E. (2012). Improved multivariate portmanteau test. *Journal of Time Series Analysis*, 33(2):211–222.

Meier, L., Van De Geer, S., and Bhlmann, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71.

Meinshausen, N. and Bhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462.

Menne, M. J. and Williams, C. N. (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18:4271–4286.

Menne, M. J. and Williams, C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7):1700–1717.

Menne, M. J., Williams, C. N., and Vose, R. S. (2009). The U.S. historical climatology network monthly temperature data, Version 2. *Bulletin of the American Meteorological Society*, 90(7):993–1007.

Mitchell, J. M. (1953). On the causes of instrumentally observed secular temperature trends. *Journal of Applied Meteorology*, 10:244–261.

Monti, A. C. (1994). A proposal for a residual autocorrelation test in linear models. *Biometrika*, 81(4):776–800.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 2:758–765.

Page, E. S. (1954). Continuous insepection schemes. *Biometrika*, 41(1/2):100–115.

Pena, D. and Rodriguez, J. (2002). A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association*, 97(458):601–610.

Pena, D. and Rodriguez, J. (2006). Testing goodness of fit in time series. *Journal of Statistical Planning and Inference*, 136(8):2706–2718.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(2):126–135.

Rechenberg, I. (1973). *Evolutions strategic: Optimierung technischer Systeme nach Prinzipen der biologischen Evolution*. Frommmann-Holzboog Verlag.

Reeves, C. R. (1995). A genetic algorithm for flowshop sequencing. *Computers & Operations Research*, 22(1):5–13.

Reeves, J., J. Chen, X. L. W., Lund, R., and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteoroology and Climatology*, 46:900–915.

Robbins, M., Gallagher, C., Lund, R., and Aue, A. (2011a). Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32(5):498–511.

Robbins, M., Lund, R., Gallagher, C., and Lu, Q. (2011b). A change-point analysis of atlantic tropical cyclones. *Journal of the American Statistical Association*, 106(493):89–99.

Rodionov, S. N. (2005). A brief overview of the regime shift detection methods. In *Large-scale disturbances (regime shifts) and recovery in aquatic ecosystems: challenges for management toward sustainability. UNESCO-ROSTE/BAS Workshop on Regime Shifts, Varna, Bulgaria*, pages 17–24.

Schmidt, M. (2005). Least squares optimization with L_1-norm regularization. *CS542B Project Report*.

Schwefel, H. P. (1977). *Numerische Optimierung von Computer-modellen mittelsder Evolution-sstrategie*. Birkhuser Verlag.

Scott, A. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512.

Shibata, R. (1999). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35:415–423.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

So, B. S. and Shin, D. W. (1999). Cauchy estimators for autoregressive processes with application to unit root tests and confidence intervals. *Economic Theory*, 15(02):165–176.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. *Journal of Urology*, 16:1076–1083.

Tahmasbi, R. and Rezaei, S. (2008). Change point detection in garch models for voice activity detection. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1038–1046.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108.

Tibshirani, R. and Taylor, J. (2012a). Degress of freedom in Lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused Lasso. *Biostatistics*, 9(1):18–29.

Tibshirani, R. J. and Taylor, J. (2012b). Degrees of freedom in Lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Vostrikova, L. J. (1981). Detection of disorder in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59.

Wang, G., Yeung, D.-Y., and Lochovsky, F. H. (2007). The kernel path in kernelized Lasso. *Journal of Machine Learning Research - Proceedings Track*, 2:580–587.

Wang, X. L. (2003). Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model". *Journal of Climate*, 16(20):3383–3385.

Wasserman, G. S. and Sudjianto, A. (1994). All subsets regression using a genetic search algorithm. *Computers and Industrial Engineering*, 27(1-4):489–492.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Series B*, 73(5):753–772.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *The Annals of Statistics*, 35(5):2173–2192.