A Review of Classical and Modern Model Selection Methods

ZONGWU CAI

March 6, 2025

1 Introduction

Given a possibly large set of potential predictors, which ones do we include in our model? Suppose X_1, X_2, $\cdots$ is a *pool* of potential predictors. The model with all predictors is given by

$$Y_t = \boxed{\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots} + u_t,$$

is the most general model. It holds even if some of the individual β_j's are zero. But, if some β_j's zero or close to zero, it is better to omit those X_j's from the model. Reasons about why you should omit variables whose coefficients are close to zero:

(a) Parsimony principle:

Given two models that perform equally well in terms of prediction, one should choose the model that is more parsimonious (simple).

(b) Prediction principle:

The model should give predictions that are as accurate as possible, not just for current observation, but for future observations as well. Including unnecessary predictors can apparently improve prediction for the current data, but can harm prediction for future data. Note that the sum of squares errors (SSE) or sum of squared residuals (SSR) never increases as we add more predictors.

Next, we discuss all possible methods available in the literature, which attempt to choose the best of all models considered, according to some criterion.

2 Subset Approaches

The all-possible-regressions procedure calls for considering all possible subsets of the pool of potential predictors and identifying for detailed examination a few good sub-sets according to some criterion. The purpose of all-possible-regressions approach is identifying a small group of regression models that are *good* according to a specified criterion (summary statistic) so that a detailed examination can be made of these models leading to the selection of the final regression model to be employed. The main problem of this approach is computationally expensive. For example, with 10 predictors, we need to investigate $2^{10} = 1024$ potential regression models. With the aid of modern computing power, this computation is possible. But still the number of 1024 possible models to examine carefully would be an overwhelming task for a data analyst.

Different criteria for comparing the regression models may be used with the all-possible-regressions selection procedure. We discuss several summary statistics:

(i) R_p^2 (or SSR_p)

(ii) $R_{adj;p}^2$ (or MSE_p)

(iii) C_p

(iv) PRESS_p

(v) Sequential Methods

(vi) AIC type criteria

We shall denote the number of all potential predictors in the pool by $K-1$. Hence including an intercept parameter β_0, we have K potential parameters. The number of predictors in a subset will be denoted by $p-1$, as always, so that there are p parameters in the regression function for this subset of predictors. Thus we have $1 \leq p \leq K$. Now, we discuss each one in detail.

1. R_p^2 (or SSR_p)

R_p^2 indicates that there are p parameters (or, $p-1$ predictors) in a regression model. The coefficient of multiple determination R_p^2 is defined as

$$R_p^2 = 1 - \frac{\text{SSR}_p}{\text{SST}_c},$$

where $\text{SST}_c = \sum_{t=1}^{n}(Y_t - \bar{Y})^2$ and $\text{SSR}_p = \sum_{t=1}^{n} \widehat{u}_t^2$.

- It measures the proportion of variance of Y explained by $p-1$ predictors.
- R_p^2 always goes up as we add more predictors.
- R_p^2 varies inversely with SSR_p because SST_c is constant for all possible regression models. That is, choosing the model with the largest R_p^2 is equivalent to choosing the model with smallest SSR_p.
- R_p^2 might not be a good criterion. **WHY?**

2. $R_{adj;p}^2$ (or MSE_p)

One often considers models with a large R_p^2 value. However, R_p^2 always increases with the number of predictors. Hence it cannot be used to compare models with different sizes. The adjusted coefficient of multiple determination $R_{adj;p}^2$ has been suggested as an alternative criterion:

$$R_{adj;p}^2 = 1 - \frac{\text{SSR}_p/(n-p)}{\text{SST}_c/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSR}_p}{\text{SST}_c} = 1 - \frac{\text{MSE}_p}{\text{SST}_c/(n-1)},$$

where $\text{MSE}_p = \text{SSR}_p/(n-p)$.

- It is like R_p^2 but with a penalty for adding unnecessary variables. R_p^2 can go down when a useless predictor is added. It can be even negative.
- $R_{adj;p}^2$ varies inversely with MSE_p because $\text{SST}_c/(n-1)$ is constant for all possible regression models. That is, choosing the model with the largest $R_{adj;p}^2$ is equivalent to choosing the model with smallest MSE_p.
- R_p^2 is useful when comparing models of the same size, while $R_{adj;p}^2$ (or C_p, defined later) is used to compare models with different sizes.

- $R^2_{adj:p}$ is better than R^2_p.

3. Mallows C_p

The Mallows C_p, proposed by Mallows (1973), is concerned with the total mean squared error of the n fitted values for each subset regression model. The mean squared error concept involves the total error in each fitted value:

$$\widehat{Y}_t - \mu_t = \underbrace{\widehat{Y}_t - E(\widehat{Y}_t)}_{\text{random error}} + \underbrace{E(\widehat{Y}_t) - \mu_t}_{\text{bias}},$$

where μ_t is the true mean response at tth observation. The means squared error for $\widehat{Y}_t$ is defined as the expected value of the square of the total error in the above. It can be shown that

$$\text{MSE}(\widehat{Y}_t) = E\left\{(\widehat{Y}_t - \mu_t)^2\right\} = \text{Var}(\widehat{Y}_t) + \left[\text{Bias}(\widehat{Y}_t)\right]^2,$$

where $\text{Bias}(\widehat{Y}_t) = E(\widehat{Y}_t) - \mu_t$. The total mean square error for all n fitted values $\widehat{Y}_t$ is the sum over the observation t:

$$\sum_{t=1}^{n} \text{MSE}(\widehat{Y}_t) = \sum_{t=1}^{n} \text{Var}(\widehat{Y}_t) + \sum_{t=1}^{n} \left[\text{Bias}(\widehat{Y}_t)\right]^2.$$

It can be shown that

$$\sum_{t=1}^{n} \text{Var}(\widehat{Y}_t) = p\,\sigma^2 \qquad \text{and} \qquad \sum_{t=1}^{n} \left[\text{Bias}(\widehat{Y}_t)\right]^2 = (n-p)[E(S_p^2) - \sigma^2],$$

where S_p^2 is the MSE from the current model. Using this, we have

$$\sum_{t=1}^{n} \text{MSE}(\widehat{Y}_t) = p\,\sigma^2 + (n-p)[E(S_p^2) - \sigma^2], \tag{1}$$

Dividing (1) by σ^2, we make it scale-free:

$$\sum_{t=1}^{n} \frac{\text{MSE}(\widehat{Y}_t)}{\sigma^2} = p + (n-p)\frac{E(S_p^2) - \sigma^2}{\sigma^2},$$

If the model does not fit well, then, S_p^2 is a biased estimate of σ^2. We can estimate $E(S_p^2)$ by MSE_p and estimate σ^2 by the MSE from the maximal model (the largest model we can consider), i.e., $\widehat{\sigma}^2 = \text{MSE}_{K-1} = \text{MSE}(X_1, \ldots, X_{K-1})$. Using the estimators for $E(S_p^2)$ and σ^2 gives

$$C_p = p + (n-p)\frac{\text{MSE}_p - \text{MSE}(X_1, \ldots, X_{K-1})}{\text{MSE}(X_1, \ldots, X_{K-1})} = \frac{\text{SSR}_p}{\text{MSE}(X_1, \ldots, X_{K-1})} - (n - 2p).$$

- Small C_p is a good thing. A small value of C_p indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses. This precision will not improve much by adding more predictors. Look for models with small C_p.

- If we have enough predictors in the regression model so that all the significant predictors are included, then, $\text{MSE}_p \approx \text{MSE}(X_1, \ldots, X_{K-1})$ and it follows that $C_p \approx p$.

- Thus C_p close to p is evidence that the predictors in the pool of potential predictors $(X_1, \ldots, X_{K-1})$ but not in the current model, are not important.

- Models with considerable lack of fit have values of C_p larger than p.

- The C_p can be used to compare models with different sizes.

- If we use all the potential predictors, then, $C_p = K$.

4. PRESS_p

The PRESS (prediction sum of squares) is defined as

$$\text{PRESS} = \sum_{t=1}^{n} \widehat{u}_{-t}^2,$$

where $\widehat{u}_{-t}$ is called PRESS (prediction sum of squares) residual for the the tth observation. The PRESS residual is defined as $\widehat{u}_{-t} = Y_t - \widehat{Y}_{-t}$, where $\widehat{Y}_{-t}$ is the fitted value obtained by leaving the tth observation. Models with small PRESS_p fit well in the sense of having small prediction errors. PRESS_p can be calculated without fitting the model n times, each time deleting one of the n cases. One can show that

$$\widehat{u}_{-t} = \frac{\widehat{u}_t}{1 - h_{tt}},$$

where h_{tt} is the tth diagonal element of $H = X(X^T X)^{-1} X^T$ (hat matrix).

3 Sequential Methods

1. Forward selection

(i) Start with the *null* model.

(ii) Add the significant variable if p-value is less than p_{enter}, (equivalently, F is larger than F_{enter}).

(iii) Continue until no more variables enter the model.

2. Backward elimination

(i) Start with the *full* model.

(ii) Eliminate the least significant variable whose p-value is larger than p_{remove}, (equivalently, F is smaller than F_{remove}).

(iii) Continue until no more variables can be discarded from the model.

3. Stepwise selection

(i) Start with any model.

(ii) Check each predictor that is currently in the model. Suppose the current model contains $X_1, \ldots, X_k$. Then, F statistic for X_j is

$$F = \frac{\text{SSR}(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k) - \text{SSR}(X_1, \ldots, X_k)}{\text{MSE}(X_1, \ldots, X_k)} \sim F(1; n-k-1),$$

which is similar to t-test statistic. Eliminate the least significant variable whose p-value is larger than p_{remove}, (equivalently, F is smaller than F_{remove}).

(iii) Continue until no more variables can be discarded from the model.

(iv) Add the significant variable if p-value is less than p_{enter}, (equivalently, F is larger than F_{enter}).

(v) Go to step (ii)

(vi) Repeat until no more predictors can be entered and no more can be discarded.

4 Likelihood Based-Criteria

The following is based on Akaike's approach Akaike (1973) and subsequent papers; see also the book by Burnham and Anderson (2003).

Suppose that $f(y)$: true model (unknown) giving rise to data (is a vector of data) and $g(y, \theta)$: candidate model (parameter vector). Want to find a model $g(y, \theta)$ "close to" $f(y)$. The Kullback-Leibler discrepancy (K-L distance):

$$K(f, g) = E_f \left[\log \left(\frac{f(Y)}{g(Y, \theta)} \right) \right].$$

This is a measure of how "far" model g is from model f (with reference to model f) with the following properties:

$$K(f, g) \geq 0 \qquad K(f, g) = 0 \iff f(\cdot) = g(\cdot).$$

Of course, we can never know how far our model g is from f. But Akaike (1973) showed that we might be able to estimate something almost as good.

Suppose we have two models under consideration: $g(y, \theta)$ and $h(y, \phi)$. Akaike (1973) showed that we can estimate $K(f, g) - K(f, h)$. It turns out that the difference of maximized log-likelihoods, corrected for a bias, estimates the difference of K-L distances. The maximized likelihoods are, $\widehat{L}_g = g(y, \widehat{\theta})$ and $\widehat{L}_h(y, \widehat{\phi})$, where $\widehat{\theta}$ and $\widehat{\phi}$ are the ML estimates of the parameters. Akaike's result: $[\log(\widehat{L}_g) - q] - [\log(\widehat{L}_h) - r]$ is an asymptotically unbiased estimate (i.e. bias approaches zero as sample size increases) of $K(f, g) - K(f, h)$. Here p is the number of parameters estimated in θ (model g) and r is the number of parameters estimated in ϕ (model h). The price of parameters: the likelihoods in the above expression are penalized by the number of parameters.

The Akaike Information Criterion (AIC) for model g:

$$\text{AIC} = -2 \log(\widehat{L}_g) + 2\,p. \tag{2}$$

A biased correction version of AIC was proposed by Hurvich and Tsai (1989), called AIC_c, defined by

$$\text{AIC}_c = \text{AIC} + 2\,q(p+1)/(n-p-1) = -2 \log(\widehat{L}_g) + 2\,pn/(n-q-1). \tag{3}$$

The difference between AIC and AIC_c is the penalty term. Intuitively, one can think of $2pn/(n-p-1)$ in (3) as a penalty term to discourage over-parameterization. Shibata (1976) suggested that the AIC has a tendency to overestimate parameter p. By comparing the penalty terms in (2) and (3), we can see that the factors, $2pn/(n-p-1)$ and $2p$, for the AIC_c and AIC statistics are asymptotically equivalent as $n \to \infty$. The AIC_c statistic however has more extreme penalty for larger-order models, which counteracts the over-fitting tendency of the AIC.

Another approach is given by the much older notion of Bayesian statistics. In the Bayesian approach, we assume that a priori uncertainty about the value of model parameters is represented by a prior distribution. Upon observing the data, this prior is updated, yielding a posterior distribution. In order to make inferences about the model (rather than its parameters), we integrate across the posterior distribution. Under the assumption that all models are a priori equally likely (because the Bayesian approach requires model priors as well as parameter priors), Bayesian model selection chooses the model with highest marginal likelihood. The ratio of two marginal likelihoods is called a Bayes factor (BF), which is a widely used method of model selection in Bayesian inference. The two integrals in the Bayes factor are nontrivial to compute unless they form a conjugated family. Monte Carlo methods are usually required to compute BF, especially for highly parameterized models. A large sample approximation of BF yields the easily-computable Bayesian information criterion (BIC)

$$\text{BIC} = -2\,\log(\widehat{L}_g) + p\,\log n. \tag{4}$$

In a sum, both AIC and BIC as well as their generalizations have a similar form as

$$LC = -2\,\log(\widehat{L}_g) + \lambda\, p, \tag{5}$$

where λ is fixed constant. From (2), (3), and (4), we can see that the BIC statistic has much more penalty when n is large than AIC and AIC_c to overcome the over-fitting. Also, from (5), it is easy to see that LC includes AIC, AIC_c and BIC as a special case.

Some developments suggest the use of a data adaptive penalty to replace the fixed penalties λ. See, Bai, Rao and Wu (1999) and Shen and Ye (2002). That is to estimate λ by data

in a complexity form based on a concept of generalized degree of freedom. Note that the above AIC methods can be extended to a nonparametric setting by Cai and Tiwari (2000), which can be learned in detail in ECON818.

5 Cross-Validation and Generalized Cross-Validation

The cross validation (CV) is the most commonly used method for model assessment and selection. The main idea is a direct estimate of extra-sample error. The general version of CV is to split data into K roughly equal-sized parts and to fit the model to the other K-1 parts and calculate prediction error on the remaining part.

$$\text{CV} = \sum_{t=1}^{n} (Y_t - \widehat{Y}_{-t})^2 \tag{6}$$

where $\widehat{Y}_{-t}$ is the fitted value computed with t-th part of data removed.

A convenient approximation to CV for linear fitting with squared error loss is generalized cross validation (GCV). A linear fitting method has the following property: $\widehat{Y} = H\,Y$, where $\widehat{Y}_t$ is the fitted value with the whole data and $H = (h_{tt})_{n\times n}$ is the smoothing (hat) matrix. For many linear fitting methods with leave-one-out ($k = 1$), it can be showed easily that

$$\text{CV} = \sum_{t=1}^{n} (Y_t - \widehat{Y}_{-t})^2 = \sum_{t=1}^{n} \left(\frac{Y_t - \widehat{Y}_t}{1 - h_{tt}} \right)^2 .$$

Due to the intensive computation, the CV can be approximated by the GCV, defined by

$$\text{GCV} = \sum_{t=1}^{n} \left(\frac{Y_t - \widehat{Y}_t}{1 - \text{trace}(H)/n} \right)^2 = \frac{\sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2}{(1 - \text{trace}(H)/n)^2}. \tag{7}$$

It has been shown that both the CV and GCV methods are very appalling to nonparametric modeling; see the book by Hastie and Tishirani (1990). It follows from (7) that

$$\text{GCV} \approx \sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2 \left(1 + \text{trace}(H)/n\right)^2 \approx \widehat{\sigma}^2 \left[\text{SSR}/\widehat{\sigma}^2 + 2\,\text{trace}(H)\right], \tag{8}$$

where $\widehat{\sigma}^2 = \sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2/n$. Therefore, under the normality assumption, GCV is asymptotically equivalent to AIC since $\text{trace}(H) = p$.

Later, the leave-one-out cross-validation method was challenged by Shao (1993). Shao (1993) claimed that the popular leave-one-out cross-validation method, which is asymptotically equivalent to many other model selection methods such as the AIC, the C_p, and the bootstrap, is asymptotically inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$ and he showed that the inconsistency of the leave-one-out cross-validation can be rectified by using a leave-n_ν-out cross-validation with n_ν, the number of observations reserved for validation, satisfying $n_\nu/n \to 0$ as $n \to \infty$. Please see some papers on how to choose n_ν in practice.

6 Penalized Methods

1. **Bridge and Ridge:**

   Frank and Friedman (1993) proposed the $L_q(q > 0)$ penalized least squares as

   $$\sum_{t=1}^{n}(Y_t - \sum_j \beta_j X_{ij})^2 + \lambda \sum_j |\beta_j|^q,$$

   which results in the estimator which is called the **bridge** estimator. If $q = 2$, the resulting estimator is called the **ridge** estimator given by $\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$.

2. **LASSO:**

   Tibshirani (1996) proposed the so-called shrinkage and selection operator (LASSO) or L_1 regularization, which is the minimizer of the following constrained least squares

   $$\sum_{t=1}^{n}(Y_t - \sum_j \beta_j X_{ij})^2 + \lambda \sum_j |\beta_j|,$$

   which results in the soft threshing rule $\widehat{\beta}_j = \text{sign}(\widehat{\beta}_j^0)(|\widehat{\beta}_j^0| - \lambda)^+$.

3. **Non-concave Penalized LS:**

   Fan and Li (2001) proposed the non-concave penalized least squares

   $$\sum_{t=1}^{n}(Y_t - \sum_j \beta_j X_{ij})^2 + \sum_j p_\lambda(|\beta_j|),$$

where the hard threshing penalty function $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 | (|\theta| < \lambda)$, which results in the hard threshing rule $\widehat{\beta}_j = \widehat{\beta}_j^0 \; I(|\widehat{\beta}_j^0| > \lambda)$. Finally, Fan and Li (2001) proposed the so-called the smoothly clipped absolute deviation (SCAD) model selection criterion with the penalized function defined as

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) - \frac{(a\,\lambda - \theta)^+}{(a-1)\lambda} I(\theta > \lambda) \right\} \qquad \text{for some} \quad a > 2 \quad \text{and} \quad \theta > 0,$$

which results in the estimator

$$\widehat{\beta}_j = \begin{cases} \text{sign}(\widehat{\beta}_j^0)(|\widehat{\beta}_j^0| - \lambda)^+ & \text{when } |\widehat{\beta}_j^0| \leq 2\lambda, \\ \left\{ (a-1)\widehat{\beta}_j^0 - \text{sign}(\widehat{\beta}_j^0)\, a\, \lambda \right\} / (a-2) & \text{when } \quad 2\lambda \leq |\widehat{\beta}_j^0| \leq a\,\lambda, \\ \widehat{\beta}_j^0 & \text{when } \quad |\widehat{\beta}_j^0| > a\,\lambda. \end{cases}$$

Empirically, as for how choose a and λ for real applications, Fan and Li (2001) suggested that the GCV approach should be used. Also, Fan and Li (2001) showed theoretically that the SCAD estimator satisfies three properties: (1) unbiasedness, (2) sparsity, and (3) continuity, and Fan and Peng (2004) considered the case that the number of regressors p can depend on the sample size and goes to infinity in a certain rate.

Finally, for more methods, please read the survey paper by Fan and Lv (2010), which discusses the properties of non-concave penalized likelihood and its roles in high dimensional statistical/econometric modeling and also reviews some recent advances in **ultra-high dimensional** variable selection, with emphasis on **independence screening** and two-scale methods. This topic is still hot in statistics and econometrics, particular, with applications in economics and finance.

Exercise: Please do simulation studies to compare the classical model selection methods such as AIC or AIC_c and the modern variable selection approaches such as LASSO or SCAD.

7 Applications in Economics and Finance

We can apply the foregoing ideas to applications in economics and finance for selecting variables and models. Please read the papers by Cai and Wang (2014), Cai, Juhl and Yang (2015), among others, for details. More importantly, please read some papers about how to impose penalty on selecting functionals.

8 Implementation in R

8.1 Classical Models

To fit a multiple regression in **R**, one can use **lm()** or **glm()**; see the followings for details

```
lm(formula, data, subset, weights, na.action,
     method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
     singular.ok = TRUE, contrasts = NULL, offset, ...)


glm(formula, family = gaussian, data, weights, subset,
      na.action, start = NULL, etastart, mustart,
      offset, control = glm.control(...), model = TRUE,
      method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)
```

to fit a regression model without intercept, you need to use

```
fit1=lm(y~-1+x1+...x9)
```

where fit1 is called the objective function containing all outputs you need. If you want to model diagnostic checking, you need to use

```
plot(fit1)
```

For multivariate data, it is usually a good idea to view the data as a whole using the pairwise scatter plots generated by the **pairs()** function:

```
pairs(data)
```

To drop or add one variable from or to a regression model, you use the command **drop1()** or **add1()**, for example,

```
drop1(fit)
add1(fit1,~x10+x11+...+x20)
```

The last command means that you choose the "best" one from X_{10} to $X20$ to add it into the model. Adding and dropping terms using **add1()** and **drop1()** is useful method for selecting a model when only a few terms are involved, but it can quickly become tedious. Functions **add1()** and **drop1()** are based on the C_p criterion. The **step()** function provides an automatic procedure for conducting stepwise model selection. The **step()** function requires an initial model, often constructed explicitly as an intercept-only model. For example, suppose that we want to find the "best" model involving $X_1, \cdots, X_{10}$, we could create an intercept-only model and then, call **step()** as follows:

```
fit0=lm(y~1)
fit2=step(fit0,~x1+x2+...+x10, trace=F)


step(object, scope, scale = 0,
        direction = c("both", "backward", "forward"),
        trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

With "trace=T" or "trace=1", **step()** displays the output of each step of the selection process. The **step()** function is based on AIC or BIC by specifying k in the function. Also, one can use the function **stepAIC()** in the package **MASS** for a wider range of object classes.

8.2 LASSO Type Methods

The package **lasso2**, which, unfortunately, is not available in the 4.4.3 (2025-02-28) version of R (outdated), provides many features for solving regression problems while imposing L_1 constraints on the estimates and the package **lars** provides efficient procedures for an entire LASSO with the cost of a single least squares. In **lars**, you can use the function **lars()**

```
lars(x, y, type = c("lasso", "lar", "forward.stagewise"),trace = FALSE,
 Gram, eps = .Machine$double.eps, max.steps, use.Gram = TRUE)
```

and the function **cv.lars()** to compute the K-fold cross-validated mean squared prediction error for least angle regressions (lars), LASSO, or forward stagewise.

```
cv.lars(x, y, K = 10, fraction = seq(from = 0, to = 1, length = 100),
    trace = FALSE, plot.it = TRUE, se = TRUE, ...)
```

In the package **lasso2**, the function **l1ce()** is for regression fitting with L_1-constraint on the parameters.

```
l1ce(formula, data = sys.parent(), weights, subset, na.action,
     sweep.out = ~ 1, x = FALSE, y = FALSE,
     contrasts = NULL, standardize = TRUE,
     trace = FALSE, guess.constrained.coefficients = double(p),
     bound = 0.5, absolute.t = FALSE)
```

or the function **gl1ce()** for fitting a generalized regression problem while imposing an L_1 constraint on the parameters

```
gl1ce(formula, data = sys.parent(), weights, subset, na.action,
      family = gaussian, control = glm.control(...), sweep.out = ~ 1,
      x = FALSE, y = TRUE, contrasts = NULL, standardize = TRUE,
      guess.constrained.coefficients = double(p), bound = 0.5, ...)
```

Also, you can use the function **gcv()** to extract the generalized cross-validation score(s) from fitted model objects, such as **r**gcv(object, ...).

Now, you can use the **glmnet** package in order to perform ridge regression and the lasso, which is widely used and well-maintained. This package provides extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, Poisson regression, Cox model, multiple-response Gaussian, and the grouped multinomial regression. The main function in this package is **glmnet()**, which can be used to fit ridge regression models, lasso models, and more. If alpha = 0 then a ridge regression model is fit, and if alpha = 1 then a lasso model is fit.

```
glmnet(x,y, family = c("gaussian", "binomial", "poisson", "multinomial", "cox",
"mgaussian"), weights = NULL, offset = NULL,
 alpha = 1,
```

```
  nlambda = 100,
  lambda.min.ratio = ifelse(nobs < nvars, 0.01, 1e-04),
  lambda = NULL,
  standardize = TRUE,
  intercept = TRUE,
  thresh = 1e-07,
  dfmax = nvars + 1,
  pmax = min(dfmax * 2 + 20, nvars),
  exclude = NULL,
  penalty.factor = rep(1, nvars),
  lower.limits = -Inf,
  upper.limits = Inf,
  maxit = 1e+05,
  type.gaussian = ifelse(nvars < 500, "covariance", "naive"),
  type.logistic = c("Newton", "modified.Newton"),
  standardize.response = FALSE,
  type.multinomial = c("ungrouped", "grouped"),
  relax = FALSE,
  trace.it = 0,
  ...
)
```

You might find more R packages for implementing the LASSO type methods.

8.3 A Real Example

We begin by introducing several environmental and economic as well as financial time series to serve as illustrative data for time series methodology. Figure 1 shows monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated recruitment (number of new fish) computed from a model by Pierre Kleiber, Southwest Fisheries Center, La Jolla, California. This data set is provided by Shumway (2006) and it can be downloaded

from https://www.ncei.noaa.gov/access/monitoring/enso/soi, which can be updated to today. Both series from Shumway (2006) are for a period of 453 months ranging over the years 1950-1987. The SOI measures changes in air pressure that are related to sea surface temperatures in the central Pacific. The central Pacific Ocean warms up every three to seven years due to the El Niño effect, which has been blamed, in particular, for foods in the midwestern portions of the U.S.

Both series in Figure 1 tend to exhibit repetitive behavior, with regularly repeating (stochastic) **cycles** that are easily visible. This **periodic behavior** is of interest because underlying processes of interest may be regular and the rate or frequency of oscillation characterizing the behavior of the underlying series would help to identify them. One can also remark that the cycles of the SOI are repeating at a faster rate than those of the recruitment series. The recruit series also shows several kinds of oscillations, a faster frequency that seems to repeat about every 12 months and a slower frequency that seems to repeat about every 50 months. The study of the kinds of cycles and their strengths will be discussed later. The two series also tend to be somewhat related; it is easy to imagine that somehow the fish population is dependent on the SOI. Perhaps there is even a lagged relation, with the SOI signaling changes in the fish population.

The study of the variation in the different kinds of cyclical behavior in a time series can be aided by computing the power spectrum, which shows the variance as a function of the frequency of oscillation. Comparing the power spectra of the two series would then give valuable information relating to the relative cycles driving each one. One might also want to know whether or not the cyclical variations of a particular frequency in one of the series, say the SOI, are associated with the frequencies in the recruitment series. This would be measured by computing the correlation as a function of frequency, called the **coherence**. The study of systematic periodic variations in time series is called **spectral analysis**. See Shumway (1988), Shumway (2006), and Shumway and Stoffer (2000) for details.

We will need a characterization for the kind of stability that is exhibited by the environmental and fish series. One can note that the two series seem to oscillate fairly regularly

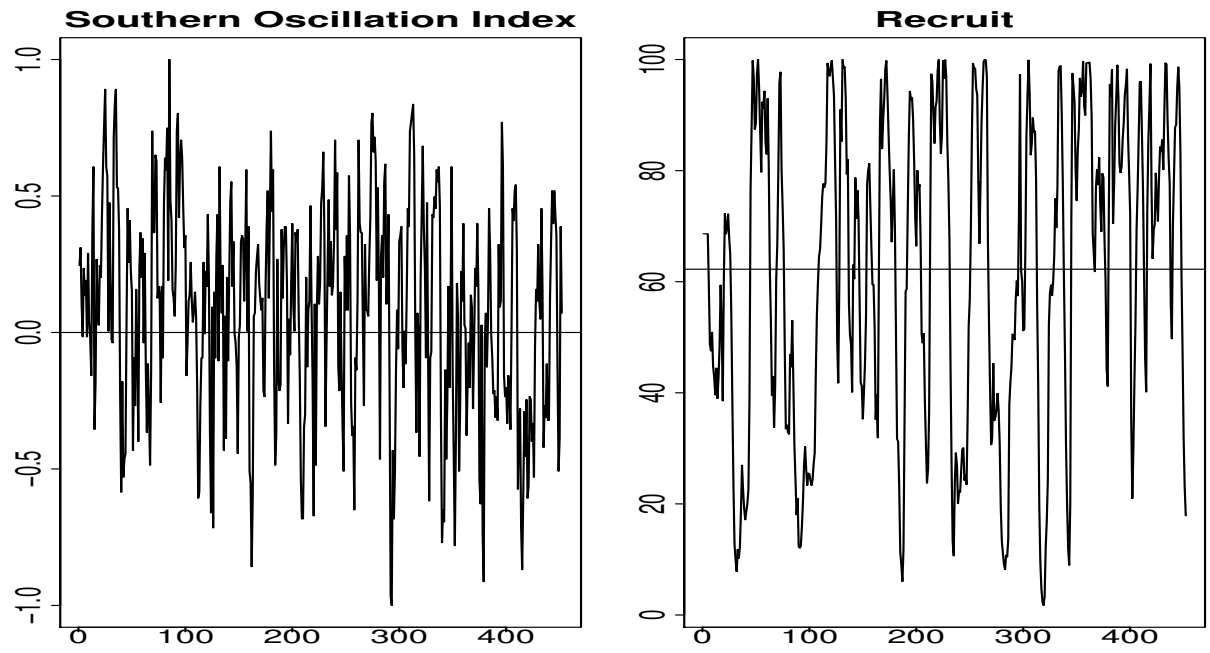

Figure 1: Monthly SOI (left panel) and simulated recruitment (right panel) from a model (n=453 months, 1950-1987).

around central values (0 for SOI and 64 for recruitment). Also, the lengths of the cycles and their orientations relative to each other do not seem to be changing drastically over the time histories.

We consider the twelve month moving average $a_j = 1/12$, $j = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6$ and zero otherwise. The result of applying this filter to the SOI index is shown in Figure 2. It is clear that this filter removes some higher oscillations and produces a smoother series. In fact, the yearly oscillations have been filtered out (see the right panel in Figure 2) and a lower frequency oscillation appears with a cycling rate of about 42 months. This is the so-called El Niño effect that accounts for all kinds of phenomena. This filtering effect will be examined further later on spectral analysis since it is extremely important to know exactly how one is influencing the periodic oscillations by filtering.

In Figure 3, we have made a lagged scatterplot of the SOI series at time $t + h$ against the SOI series at time t and obtained a high correlation, 0.412, between the series x_{t+12} and the series x_t shifted by 12 years. Lower order lags at $t - 1$, $t - 2$ also show correlation.

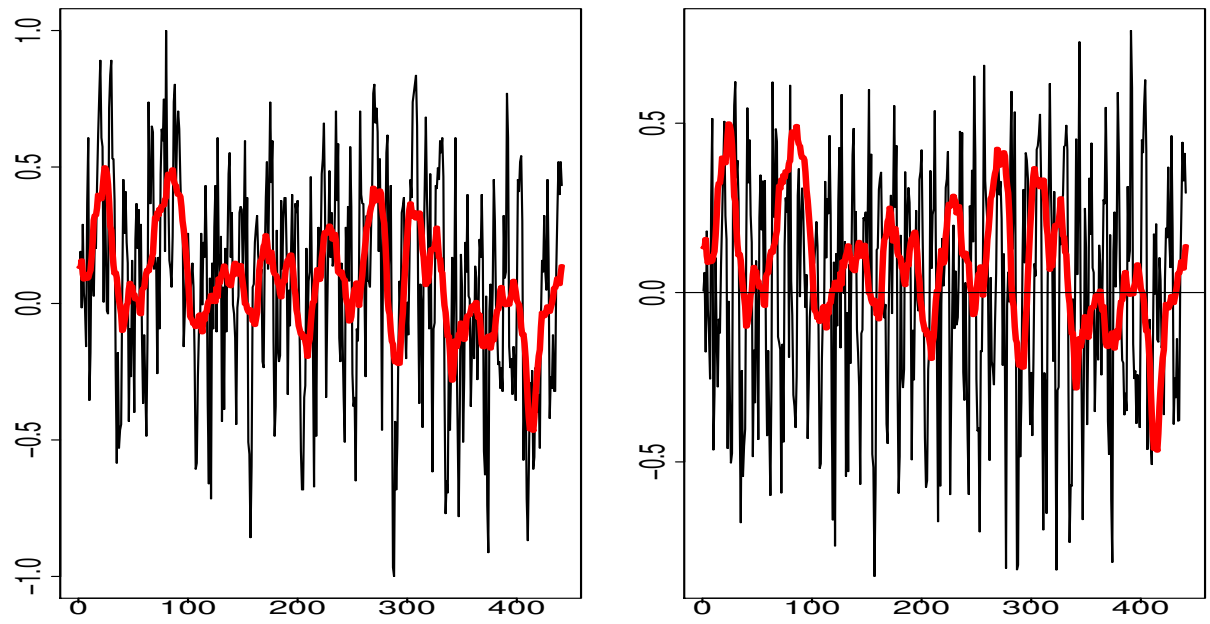

Figure 2: The SOI series (black solid line) compared with a 12 point moving average (red thicker solid line). The left panel: original data and the right panel: filtered series.

The scatterplot shows the direction of the relation, which tends to be positive for lags 1, 2, 11, 12, 13, and tends to be negative for lags 6, 7, 8. The scatterplot can also show no significant nonlinearities to be present. In order to develop a measure for this self correlation or autocorrelation, we utilize a sample version of the scaled auto-covariance function, say

$$\widehat{\rho}_x(h) = \widehat{\gamma}_x(h)/\widehat{\gamma}_x(0),$$

where

$$\widehat{\gamma}_x(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x}),$$

which is the sample counterpart with $\overline{x} = \sum_{t=1}^{n} x_t/n$. Under the assumption that the underlying process x_t is white noise, the approximate standard error of the sample ACF is $\sigma_\rho = 1/\sqrt{n}$. That is, $\widehat{\rho}_x(h)$ is approximately normal with mean 0 and variance $1/n$.

As an illustration, consider the autocorrelation functions computed for the environmental and recruitment series shown in the top two panels of Figure 4. Both of the autocorrelation functions show some evidence of periodic repetition. The ACF of SOI seems to repeat at

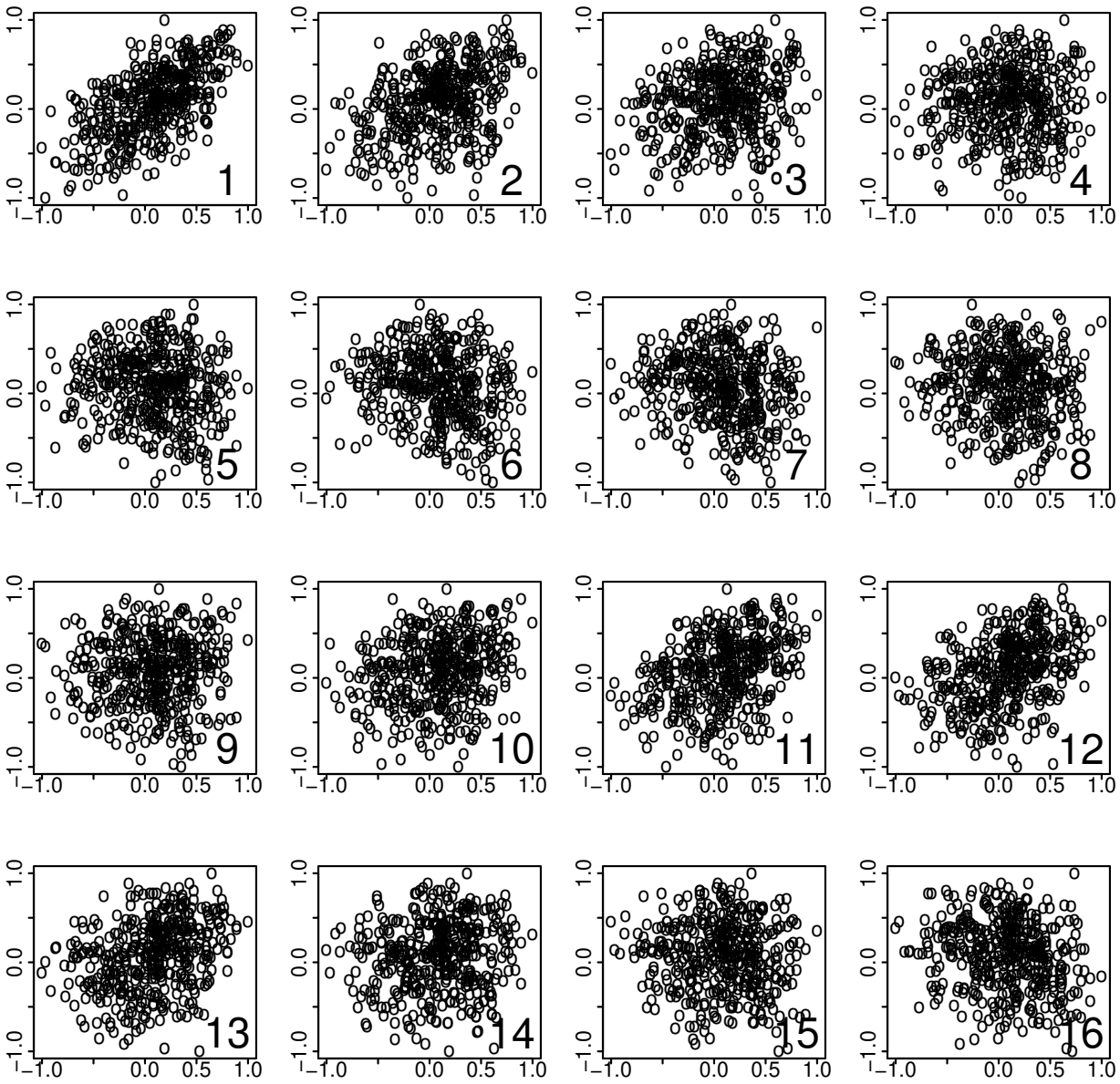

Figure 3: Multiple lagged scatterplots showing the relationship between SOI and the present (x_t) versus the lagged values (x_{t+h}) at lags $1 \leq h \leq 16$.

periods of 12 while the recruitment has a dominant period that repeats at about 12 to 16 time points. Again, the maximum values are well above two standard errors shown as dotted lines above and below the horizontal axis.

In order to examine this possibility, consider the lagged scatterplot matrix shown in

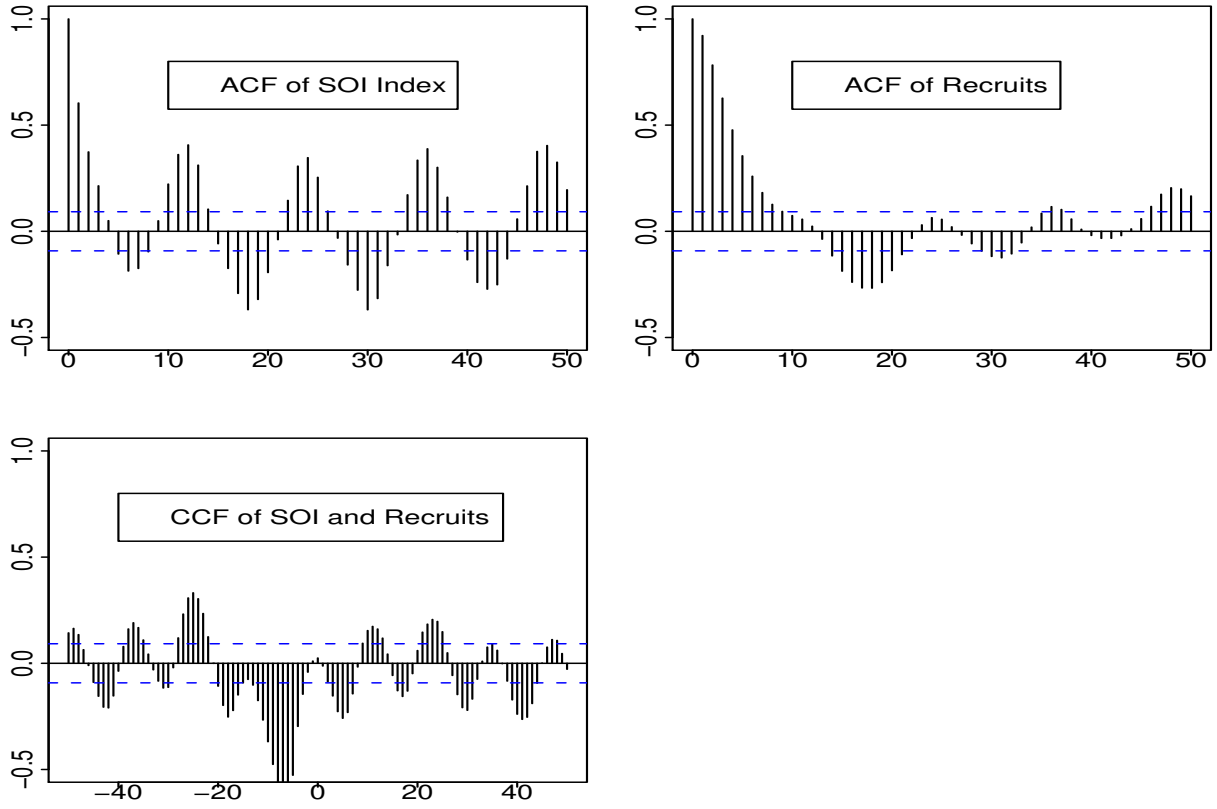

Figure 4: Autocorrelation functions of SOI and recruitment and cross correlation function between SOI and recruitment.

Figures 5 and 6, respectively. Figure 5 plots the SOI at time $t+h$, x_{t+h}, versus the recruitment series y_t at lag $0 \leq h \leq 15$ in Figure 5. There are no particularly strong linear relations apparent in this plots, i.e. future values of SOI are not related to current recruitment. This means that the temperatures are not responding to past recruitment. In Figure 6, the current SOI values, x_t are plotted against the future recruitment values, y_{t+h} for $0 \leq h \leq 15$. It is clear from Figure 6 that the series are correlated negatively for lags $h = 5, \ldots, 9$. The correlation at lag 6, for example, is -0.60 implying that increases in the SOI lead decreases in number of recruits by about 6 months. On the other hand, the series are hardly correlated (0.025) at all in the conventional sense, measured at lag $h = 0$. The general pattern suggests that predicting recruits might be possible using the El Niño at lags of 5, 6, 7, ... months.

We show in the right panels panel of Figure 7 the partial autocorrelation functions (PAC)

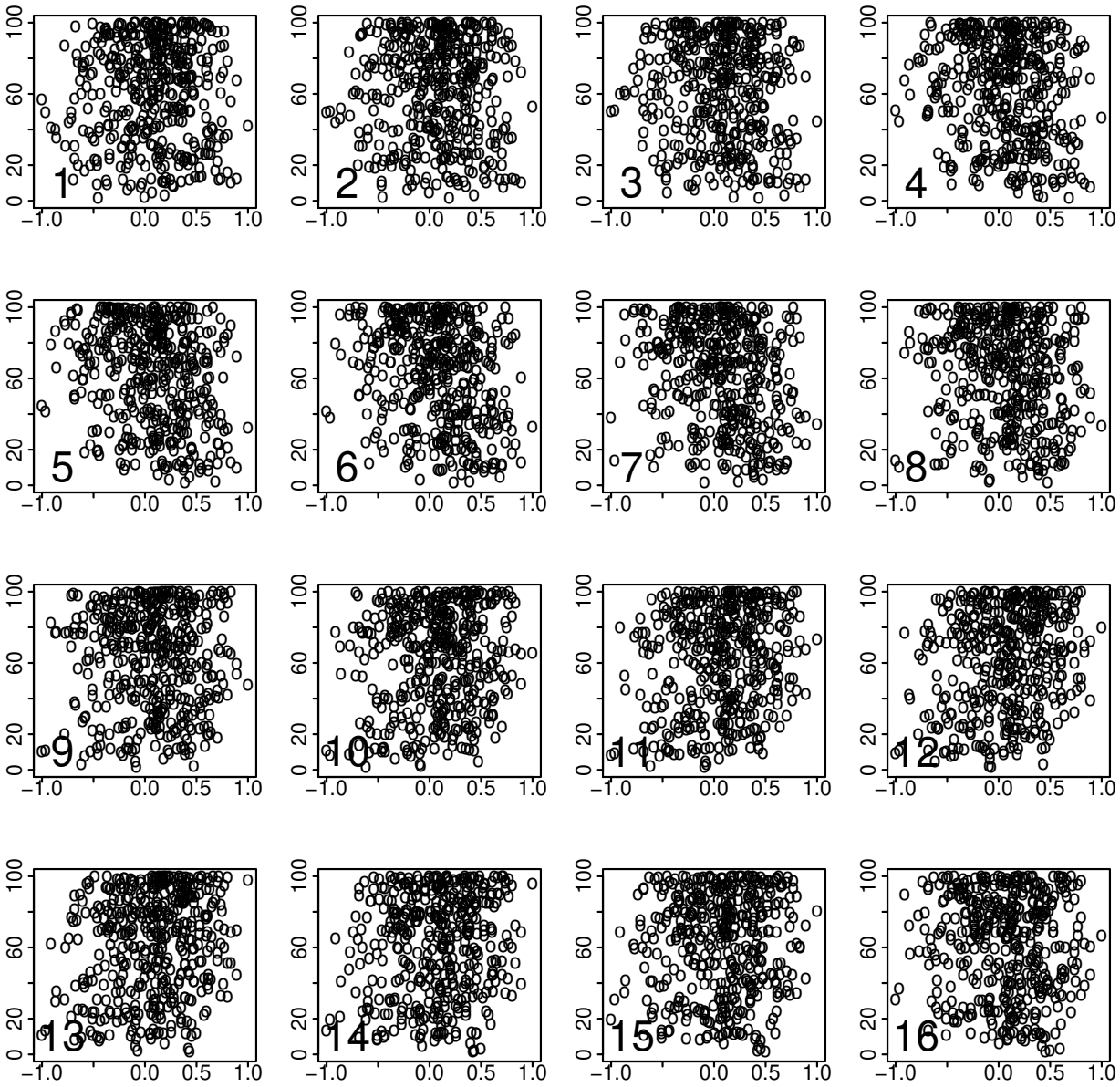

Figure 5: Multiple lagged scatterplots showing the relationship between the SOI at time $t + h$, say x_{t+h} (x-axis) versus recruits at time t, say y_t (y-axis), $0 \leq h \leq 15$.

of the SOI series (left panel) and the recruits series (right panel). Note that the PACF of the SOI has a single peak at lag $h = 1$ and then relatively small values. This means, in effect, that fairly good prediction can be achieved by using the immediately preceding point and that adding further values does not really improve the situation. Hence, we might try an autoregressive model with $p = 1$. The recruits series has two peaks and then small values,

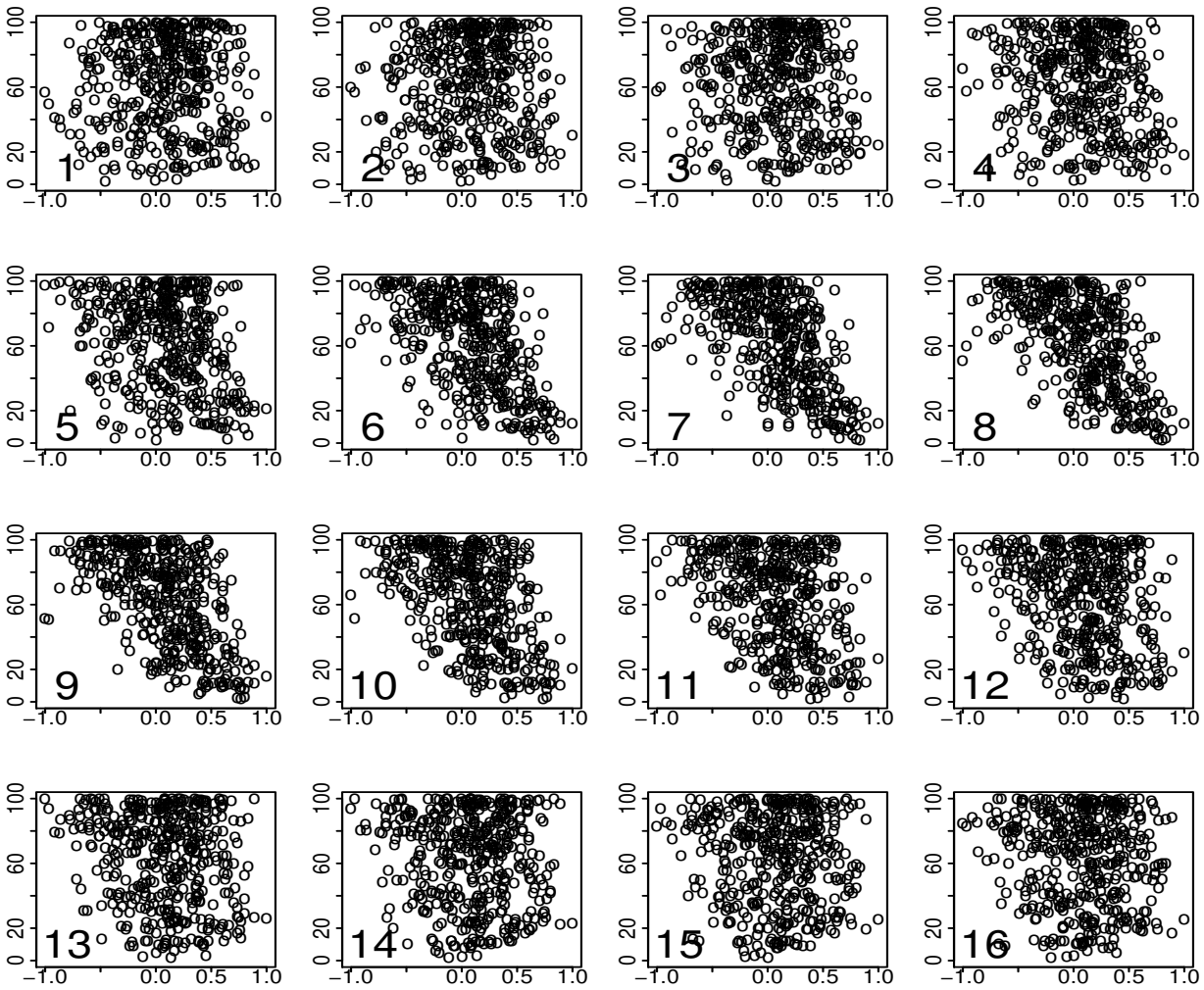

Figure 6: Multiple lagged scatterplots showing the relationship between the SOI at time t, say x_t (x-axis) versus recruits at time $t+h$, say y_{t+h} (y-axis), $0 \leq h \leq 15$.

implying that the pure correlation between points is summarized by the first two lags.

We consider the simple problem of modeling the recruit series shown in the right panel of Figure 1 using an autoregressive model. The top right panel of Figure 4 and the right panel of Figure 7 shows the autocorrelation and partial autocorrelation functions of the recruit series. The PACF has large values for $h = 1$ and 2 and then is essentially zero for higher order lags. This implies by the property of an autoregressive model that a second order $(p = 2)$ AR model might provide a good fit. Running the regression program for an AR(2)

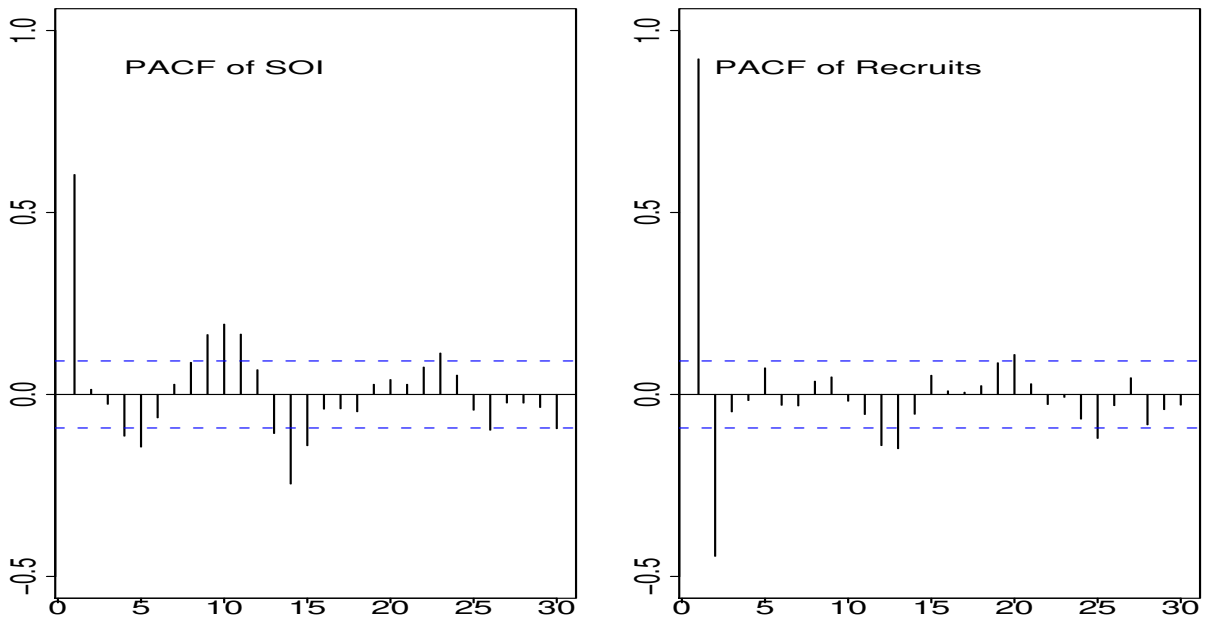

Figure 7: Partial autocorrelation functions for the SOI (left panel) and the recruits (right panel) series.

Table 1: AIC_c values for ten models for the recruits series

p	1	2	3	4	5	6	7	8	9	10
AIC_c	5.75	**5.52**	5.53	5.54	5.54	5.55	5.55	5.56	5.57	5.58

model with intercept

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

leads to the estimators $\widehat{\phi}_0 = 61.8439(4.0121)$, $\widehat{\phi}_1 = 1.3512(0.0417)$, $\widehat{\phi}_2 = -0.4612(0.0416)$ and $\widehat{\sigma}^2 = 89.53$, where the estimated standard deviations are in parentheses. To determine whether the above order is the best choice, we fitted models for $1 \leq p \leq 10$, obtaining corrected AIC_c values summarized in Table 1 using (3). This shows that the minimum AIC_c obtains for $p = 2$ and we choose the second order model.

The previous example uses various autoregressive models for the recruits series, for example, one can fit a second-order regression model. We may also use this regression idea to fit the model to other series such as a de-trended version of the SOI given in previous

discussions. We have noted in our discussions of Figure 7 from the partial autocorrelation function that a plausible model for this series might be a first order autoregression of the form given above with $p = 1$. Again, putting the model above into the regression framework $x_t = \phi_0 + \phi_1 x_{t-1} + w_t$ for a single coefficient leads to the estimators $\widehat{\phi}_1 = 0.59$ with standard error 0.04, $\widehat{\sigma}^2 = 0.09218$ and $\text{AIC}_c(1) = -1.375$. The ACF of these residuals shown in the left panel of Figure 8, however, will still show cyclical variation and it is clear that they still have a number of values exceeding the $1.96/\sqrt{n}$ threshold. A suggested procedure is to try higher order

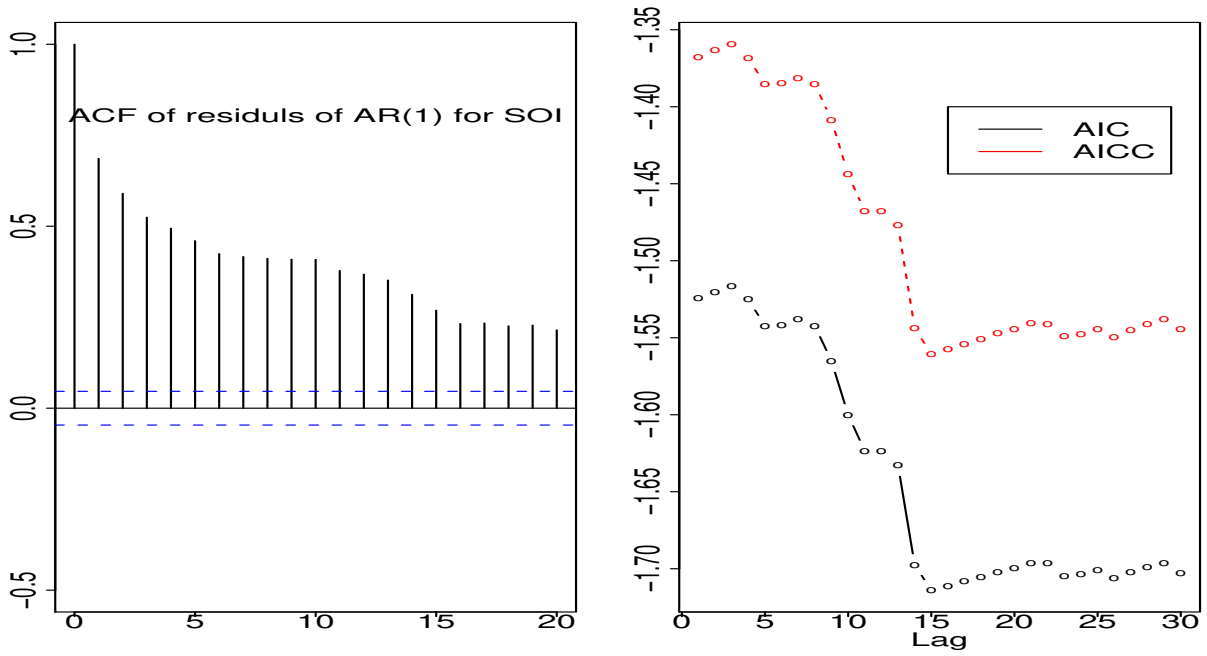

Figure 8: ACF of residuals of AR(1) for SOI (left panel) and the plot of AIC and AIC_c values (right panel).

autoregressive models and successive models for $1 \leq p \leq 30$ were fitted and the AIC_c values are plotted in the right panel of Figure 8. There is a clear minimum for a $p = 16$ order model. The coefficient vector is $\boldsymbol{\phi}$ with components and their standard errors in the parentheses 0.4050(0.0469), 0.0740(0.0505), 0.1527(0.0499), 0.0915(0.0505), −0.0377(0.0500), −0.0803(0.0493), −0.0743(0.0493), −0.0679(0.0492), 0.0096(0.0492), 0.1108 (0.0491), 0.1707(0.0492), 0.1606(0.0499), 0.0281(0.0504), −0.1902(0.0501), −0.1283(0.0510), −0.0413(0.0476), and $\widehat{\sigma}^2 = 0.07166$.

Exercise: Please download the SOI until now and re-analyze this data set to see if there is

any change in the last 4 decades.

9 Computer Codes

```
#########################################################################
 # This is the example for Southern Oscillation Index and Recruits data
 #########################################################################

 y<-read.table(file="~/Desktop/ECON817/data/soi.txt")
 # read data file
 x<-read.table(file="~/Desktop/ECON817/data/recruit.txt")
 y=y[,1]
 x=x[,1]

 postscript(file="~/Desktop/ECON817/materials/figs/fig_2_1.eps",
  horizontal=F,width=6,height=6)
 #win.graph()
 par(mfrow=c(1,2),mex=0.4,bg="yellow")
 # save the graph as a postscript file
 ts.plot(y,type="l",lty=1,ylab="",xlab="")
 # make a time series plot
 title(main="Southern Oscillation Index",cex=0.5)
 # set up the title of the plot
 abline(0,0)
 # make a straight line
 #win.graph()
 ts.plot(x,type="l",lty=1,ylab="",xlab="")
 abline(mean(x),0)
 title(main="Recruit",cex=0.5)
```

```
dev.off()

n=length(y)
n2=n-12
yma=rep(0,n2)
for(i in 1:n2){yma[i]=mean(y[i:(i+12)])}        # compute the mean
yy=y[7:(n2+6)]
yy0=yy-yma

postscript(file="~/Desktop/ECON817/materials/figs/fig_2_2.eps",
 horizontal=F,width=6,height=6)
par(mfrow=c(1,2),mex=0.4)
ts.plot(yy,type="l",lty=1,ylab="",xlab="")
points(1:n2,yma,type="l",lty=1,lwd=3,col=2)
ts.plot(yy0,type="l",lty=1,ylab="",xlab="")
points(1:n2,yma,type="l",lty=1,lwd=3,col=2)   # make a point plot
abline(0,0)
dev.off()

m=17
n1=n-m
y.soi=rep(0,n1*m)
dim(y.soi)=c(n1,m)
y.rec=y.soi
for(i in 1:m){
y.soi[,i]=y[i:(n1+i-1)]
y.rec[,i]=x[i:(n1+i-1)]}
text_soi=c("1","2","3","4","5","6","7","8","9","10","11","12","13",
 "14","15","16")
```

```
postscript(file="~/Desktop/ECON817/materials/figs/fig_2_3.eps",
 horizontal=F,width=6,height=6)
par(mfrow=c(4,4),mex=0.4,bg="light blue")
for(i in 2:17){
plot(y.soi[,1],y.soi[,i],type="p",pch="o",ylab="",xlab="",
 ylim=c(-1,1),xlim=c(-1,1))
text(0.8,-0.8,text_soi[i-1],cex=2)}
dev.off()

text1=c("ACF of SOI Index")
text2=c("ACF of Recruits")
text3=c("CCF of SOI and Recruits")
SOI=y
Recruits=x
postscript(file="c:/res-teach/xiamen12-06/figs/fig_2_4.eps",
 horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light pink")
acf(y,ylab="",xlab="",ylim=c(-0.5,1),lag.max=50,main="")
 # make an ACF plot
legend(10,0.8, text1)                        # set up the legend
acf(x,ylab="",xlab="",ylim=c(-0.5,1),lag.max=50,main="")
legend(10,0.8,text2)
ccf(y,x, ylab="",xlab="",ylim=c(-0.5,1),lag.max=50,main="")
legend(-40,0.8,text3)
dev.off()

postscript(file="~/Desktop/ECON817/materials/figs/fig-2.5.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(4,4),mex=0.4,bg="light green")
for(i in 1:16){
```

```
plot(y.soi[,i],y.rec[,1],type="p",pch="o",ylab="",xlab="",
ylim=c(0,100),xlim=c(-1,1))
text(-0.8,10,text_soi[i],cex=2)}
dev.off()

postscript(file="~/Desktop/ECON817/materials/figs/fig_2_6.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(4,4),mex=0.4,bg="light grey")
for(i in 1:16){
plot(y.soi[,1],y.rec[,i],type="p",pch="o",ylab="",xlab="",
ylim=c(0,100),xlim=c(-1,1))
text(-0.8,10,text_soi[i],cex=2)}
dev.off()

postscript(file="~/Desktop/ECON817/materials/figs/fig_2_7.eps",
 horizontal=F,width=6,height=6)
par(mfrow=c(1,2),mex=0.4,bg="light blue")
pacf(y,ylab="",xlab="",lag=30,ylim=c(-0.5,1),main="")
text(10,0.9,"PACF of SOI")
pacf(x,ylab="",xlab="",lag=30,ylim=c(-0.5,1),main="")
text(10,0.9,"PACF of Recruits")
dev.off()
#####################################################################

x<-read.table(file="~/Desktop/ECON817/data/ex2-1a.txt")
x.soi=x[,1]
n=length(x.soi)
aicc=0

if(aicc==1){
```

```
aic.value=rep(0,30)                              # max.lag=30
aicc.value=aic.value
sigma.value=rep(0,30)
for(i in 1:30){
fit3=arima(x.soi,order=c(i,0,0))                 # fit an AR(i)
aic.value[i]=fit3$aic/n-2                        # compute AIC
sigma.value[i]=fit3$sigma2
 # obtain the estimated sigma^2
aicc.value[i]=log(sigma.value[i])+(n+i)/(n-i-2)  # compute AICC
print(c(i,aic.value[i],aicc.value[i]))}
data=cbind(aic.value,aicc.value)
write(t(data),file="~/Desktop/ECON817/materials/soi_aic.dat",ncol=2)
}else{
data<-matrix(scan(file="~/Desktop/ECON817/materials/soi_aic.dat"),byrow=T,ncol=2)
}
text4=c("AIC", "AICC")

fit11=arima(x.soi,order=c(1,0,0))
resid1=fit11$residual
postscript(file="~/Desktop/ECON817/materials/figs/fig_2_8.eps",
 horizontal=F,width=6,height=6)
par(mfrow=c(1,2),mex=0.4,bg="light yellow")
acf(resid1,ylab="",xlab="",lag.max=20,ylim=c(-0.5,1),main="")
text(10,0.8,"ACF of residuals of AR(1) for SOI")
matplot(1:30,data,type="b",pch="o",col=c(1,2),ylab="",xlab="Lag",cex=0.6)
legend(16,-1.40,text4,lty=1,col=c(1,2))
dev.off()

#fit2=arima(x.soi,order=c(16,0,0))
#print(fit2)
```

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceeding of 2nd International Symposium on Information Theory* (V. Petrov and F. Csáki, eds.) 267281. Akadémiai Kiadó, Budapest.

Bai, Z., C.R. Rao and Y. Wu (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inferences*, **77**, 103-117.

Burnham, K.P. and D. Anderson (2003). *Model Selection And Multi-Model Inference: A Practical Information Theoretic Approach*, 2nd edition. New York: Springer-Verlag.

Cai, Z., T. Juhl and B. Yang (2015). Functional index coefficient models with variable selection. *Journal of Econometrics*, **189**, 272-284.

Cai, Z. and R.C. Tiwari (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**(3), 341-350.

Cai, Z. and X. Wang (2014). Selection of mixed copula model via penalized likelihood. *Journal of The American Statistical Association*, **109**, 788-801.

Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (L. LeCam and J. Neyman, eds.), University of California Press, Berkeley.

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

Fan, J. and J. Lv (2010). A selective review of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101-148.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, **32**, 928-961.

Frank, I.E. and J.H. Friedman (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics*, **35**, 109-148.

Hastie, T.J. and R.J. Tishirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

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

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**(4), 661675

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

Shao (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486-494.

Shen, X.T. and J.M. Ye (2002). Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210-221.

Shibata (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**(1), 117-126.

Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Shumway, R.H. (2006). *Lecture Notes on Applied Time Series Analysis*. Department of Statistics, University of California at Davis.

Shumway, R.H. and D.S. Stoffer (2000). *Time Series Analysis & Its Applications*. New York: Springer-Verlag.

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