# Chapter 6.

# Parametric Models and Bayesian Statistics

This chapter briefly introduces some fundamental results of the parametric models, including exponential families of disbributions, the fisher information, the maximum likelihood estimation, and likelihood ratio tests, and generalized linear models. The basic framework and philosophy of Bayesian statistics are reviewed and illustrated with the influential results such as regularized estimation and naive Bayes approach in classification.

## 6.1. Parametric models and exponential families.

In general, data are viewed as observations of random variables coming from a population distribution, which is determined by an unknown parameter $\theta$. The population is the goal of data analysis or the goal of learning from data. In the extreme case, $\theta$ may be the distribution itself, meaning that nothing is assumed about the population distribution. A simple case is that the samples are taken repeatedly with replacement from a population. We view the observed data as realized values of a colleciton of independent, identically distributed random variables that follow the common population distribution, $F(\cdot;\theta)$. For simplicity, assume $F(\cdot,\theta)$ has density function $f(\cdot;\theta)$ (or probability function, if discrete). If $\theta$ is of finite dimension, we call this a parametric model.

The exponential family of distributions are widely used in statistical modeling and they form the basis of parametric distributions. An exponential family of distribution refers to the distributions with density

$$f(x;\theta) = \exp\{g(\theta)^T T(x) + S(x) + h(\theta)\} I(x \in A) \tag{6.1}$$

where $A$ is irrelevant with $\theta$. For iid random variables $X_i$, $i = 1, .., n$, if the common distribution is is an exponential family, the density of $X_1, ..., X_n$ is

$$f(x_1, ..., x_n; \theta) = \exp\Big\{ g(\theta) \sum_{i=1}^{n} T(x_i) + \sum_{i=1}^{n} S(x_i) + nh(\theta) \Big\} I(x_1 \in A, ..., x_n \in A),$$

which also belongs to an exponential family of distribution.

Most of the commonly used distributions, such as normal distribution, exponential distribution, Gamma distribution, Poisson distribution, geometric distribution, binomial distribution, with relevant parameters, are exponential family of distributions. However, Cauchy distributon centering at $\theta$ and uniform distribution on $[0, \theta]$ are not. As to be seen, exponential family of distribution has certain desired analytical properties.

The concept of sufficient statistics can be used to measure no-loss of information in data reduction. Suppose we wish to reduce the original data set of large size, say $n$, to a statistic, say $T^*$, of possibly a smaller dimension, and does not wish to lose any information about the parameter. This statistic is called sufficient statistic. Its definition is that the conditional distribution of the full data given $T^*$ is irrelevant with $\theta$. Once $T$ is fixed, the variation of the data are simply noise that carries no information about what $\theta$ should be. The full data itself, viewed as a statistic, is certainly sufficient, though it does not have any effect of data reduction. For exponential family of distribution, the $T$ statistic in (6.1) is a sufficient statistic. If we have $n$ iid observations, $X_1, ..., X_n$, from an exponential family, then $\sum_{i=1}^{n} T(X_i)$ as in (6.2) is a sufficient statistic. For example, suppose the common distribution is $N(\mu, \sigma^2)$ with two unknown parameters $\mu$ and $\sigma^2$. Then $T(x) = (x, x^2)^T$ and $g(\mu, \sigma^2) = (\mu, 1/(2\sigma^2))^T$. If we have $n$ iid observations $X_1, ..., X_n$, then $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient statistics for $(\mu, \sigma^2)$. In other words, knowing sample mean $\bar{X}$ and sample variance $s^2$, the rest of the variations over the data $X_1, ..., X_n$ are irrelevant with the target parameter $(\mu, \sigma^2)$ of interest. Sufficiency is a top benchmark for data reduction in parametric models.

## 6.2. The Fisher information

It is often useful to quantify the amount of information about the parameter $\theta$ contained in the data. Using a generic $\mathcal{D}$ to stand for data, or, more precisedly, the random variables with observed values in data, and let $f(\mathcal{D};\theta)$ be its density function. The fisher information is defined as

$$I(\theta) = -E_\theta\Big(\frac{\partial^2}{\partial\theta^2}\log f(\mathcal{D};\theta)\Big) = \mathrm{var}_\theta\Big(\frac{\partial^2}{\partial\theta^2}\log f(\mathcal{D};\theta)\Big).$$

For simplicity, we confine our attention on one dimensional parameter. The larger the $I(\theta)$, the more information about $\theta$ contained in $\mathcal{D}$. It measures the sentivity of the variation of data with respect to the variation of the parameters. For example, if $X \sim N(\mu, 1)$, its Fisher information, with some simple calculation, is 1. On the other hand, $X \sim N(\mu, 1000)$, its Fisher information is $1/1000$. For large variance 1000, variation of $X$ is largely "noise", those unrelated with the change of $\mu$, and thus not sensitive to the change of $\mu$, while for small variance, $X$ is much more "sensitive" to the change of the parameter $\mu$.

Independent random variables should carry different or non-overlapping information about the parameter. Their total amount of information should be the sum of each. This is indeed the case by the definition of Fisher information. If one random variable $X_1$ has Fisher information $I(\theta)$, then $n$ iid random variables $X_1, ..., X_n$ together will have Fisher information $nI(\theta)$.

One primary target in statistical analysis is the estimation of the population distribution, which is the same as estimation of the parameters in the case of parametric models. Developing estimators or estimation methods with certain optimality is one of the main tasks for data analysis. It may be tempting to find an estimator that always dominates all others in some common sense criterion. The mean squared error (MSE) is a widely used criterion for measuring estimation accuracy. No such ultra-optimal estimator exists, since, by taking $\theta$ as one fixed value blindly, it will result in 0 MSE if that fixed value happens to be the true parameter value. However, if one only focus on unbiased estimators based on data $\mathcal{D}$, then, under regularity conditions, the lowest possible MSE is the inverse of its Fisher information. The more information the data contains about the parameter, the more accurate the parameter could be estimated.

Fisher information can be extended to any dimension $d$, in which case it is a $d \times d$ matrix, or even infinite dimensional parameters. With infinite dimensional parameters in the model, we call it semiparametric or nonparametric models.

## 6.3. The maximum likelihood estimation

In statistics, the likelihood principle states that all evidence in the data about the assumed model is in the likelihood function. Given the assumed model being correct, the likelihood based statistical procedure, such as estimation or testing, indeed has certain superiority over other competing methods. The likelihood function of the data is actually just its joint density function or, if discrete, probability function, but viewed as a function of the parameter $\theta$. The higher the likelihood function is at certain $\theta$ value, the more "likely" that value of $\theta$ is the true value of $\theta$. Often the log-likelihood is much more convenient to work with. The log function, denoted as $l(\theta, \mathcal{D})$, is a monotone increasing function and similar interpretation holds for log-likelihood. The maximum likelihood estimation then identifies the value of that such that $l(\theta, \mathcal{D})$ achieves maximum. That is

$$\hat{\theta}_{\mathrm{MLE}} = \ \mathrm{argmax}_\theta l(\theta, \mathcal{D}).$$

It is not surprising that the MLE enjoys many advantages. Under regularity conditions, MLE can be proved to be asymptotically following normal distribution with mean $\theta$ and variance being the inverse of the Fisher information of the data. Moreover, no other regular estimators would have asymptotic distribution with mean $\theta$ and smaller variance than that of the MLE.

Numerically, the MLE is often the solution of

$$\dot{l}(\theta, \mathcal{D}) = 0.$$

We use $\dot{l}$ and $\ddot{l}$ to denote the first and second derivatives with respect to $\theta$. The left hand side is often called the score function. Note that the score function has mean 0. This is a special case of a general methodology based on estimating functions. The estimating function approach equates a function of the data and the parameter, say $g(\theta, \mathcal{D})$, to 0:

$$g(\theta, \mathcal{D}) = 0.$$

And the solution, denoted as $\hat{\theta}$, is the estimator of $\theta$. The choice of the function $g$ determines the quality of the estimator. The least requirement of $g$ is its mean is 0 or close to 0. If the model is correct, one of the top choices of $g$ is $\dot{l}$. With large data and some regularity conditions, we can approximate

$$g(\theta, \mathcal{D}) = g(\theta, \mathcal{D}) - g(\hat{\theta}, \mathcal{D}) \approx -\dot{g}(\hat{\theta}, \mathcal{D})(\hat{\theta} - \theta)$$

If $g(\theta, \mathcal{D})$ is mean 0 and its variance is estimated by $\hat{V}$, then the variance of $\hat{\theta}$ is estimated by a sandwich formula

$$\dot{g}(\hat{\theta}, \mathcal{D})\hat{V}\dot{g}(\hat{\theta}, \mathcal{D})^T.$$

If the data consist of iid observations $X_1, .., X_n$, $g(\theta, \mathcal{D})$ is often the sum of $h(\theta, X_i)$ with mean 0. The estimating equation is

$$\sum_{i=1}^{n} h(\theta, X_i) = 0$$

and the variance of the estimator is estimated by the sandwhich formula:

$$\sum_{i=1}^{n} \dot{h}(\theta, X_i)\Big(\sum_{i=1}^{n} h(\theta, X_i)\{h(\theta, X_i)\}^T\Big)\{\sum_{i=1}^{n} \dot{h}(\theta, X_i)\}^T.$$

And approximate inference can be obtained through the central limit theorm. For example, for a given constant vector $\mathbf{b}$, the confidence interval for $\mathbf{b}^T\theta$ at approximate level $(1-\alpha)\times 100\%$ is

$$\mathbf{b}^T\hat{\theta} + z(\alpha/2)\sqrt{\mathbf{b}^T\hat{V}_{\hat{\theta}}\mathbf{b}},$$

where $z(\alpha/2)$ is the percentile of $N(0,1)$ at level $1-\alpha/2$, and $\hat{V}_{\hat{\theta}}$ is the estimated variance of $\hat{\theta}$.

Under the likelihood principle, it is natural to evaluate the relative evidence in the data in favor of one value of the parameter, say $\theta_1$, against that in favor of another value of the parameter, say $\theta_0$, by the likelihood ratio or, equivalently, the log-likelihood ratio:

$$l(\theta_1; \mathcal{D}) - l(\theta_0; \mathcal{D}).$$

This gives rise to the well known likelihood ratio test, whose optimality is theoretically supported by the Neyman-Pearson Lemma.When testing $\theta \in \mathcal{H}_0$ against $\theta \notin \mathcal{H}_0$, the score test is based on

$$-\dot{l}(\hat{\theta}_{\text{MLE}}; \mathcal{D})\{\ddot{l}(\hat{\theta}_{\text{MLE}}; \mathcal{D})\}^{-1}\dot{l}(\hat{\theta}_{\text{MLE}}; \mathcal{D}) \sim \chi^2_k \qquad \text{approximately}$$

where $k$ is the difference of the dimension of $\theta$ and the dimension of $\mathcal{H}_0$ and $\hat{\theta}_{\text{MLE}}$ is the MLE under $\mathcal{H}_0$

## 6.4. Generalized linear models

The linear regression model detailed in Chapter 2 assumes the conditional mean of the response given the covariates is linear function of the covariates. This in many cases can be too rigid and

can be readily extended. The generalized linear model relates the conditional mean of the response given covariate with a linear function of the covariates through a link function.

$$g(\mu) = \mathbf{x}^T\beta,$$

where $\mu = E(Y|X = \mathbf{x})$, and $g$ is the *link function*. The commonly used link functions can be derived from the forms of the exponential family of distribution in (6.1). For example, the normal, exponential, binomial and Poisson distributions respectively correspond to the identity, inverse, log, and logit link functions.

In particular, if the response is binary and categorical, such as Yes/No, we can code it into numerical 1 and 0. Then, logistic regression model assumes

$$\log\Big(\frac{\mu}{1-\mu}\Big) = \mathbf{x}^T\beta.$$

With iid observations $(y_i, \mathbf{x}_i), i = 1, ..., n$. The maximum likelihood estimates of $\beta$ is

$$\hat{\beta}_{\text{MLE}} = \text{argmax}_\beta l(\beta, \mathcal{D}) = \sum_{i=1}^{n}(y_i\beta^T\mathbf{x}_i - \log(1 + \exp(\beta^T\mathbf{x}_i))),$$

which is the solution of

$$\sum_{i=1}^{n}\mathbf{x}_i(y_i - 1/(1 + \exp(-\beta^T\mathbf{x}_i))) = 0.$$

Suppose we wish to classify a new observation with covariates $\mathbf{x}_0$ into either class 1 or class 0. With the estimated $\hat{\beta}_{\text{MLE}}$, the classification can be carried out by comparing $P(Y = 1|X = \mathbf{x}_0)$ against $P(Y = 0|X = \mathbf{x}_0)$. In other words, the classification can be made based on the estimated $\log[P(Y = 1|X = \mathbf{x}_0)/P(Y = 0|X = \mathbf{x}_0)] = \beta^T\mathbf{x}_0$, leading to

$$\text{classify into } \begin{cases} \text{class 1} & \text{if } \mathbf{x}_0^T\hat{\beta}_{\text{MLE}} \geq c; \\ \text{class 0} & \text{if } \mathbf{x}_0^T\hat{\beta}_{\text{MLE}} < c. \end{cases}$$

It can be seen that the separating boundary is a linear function of $\mathbf{x}_0$. Another linear classification is the linear discriminant analysis (LDA), which is based on the normality assumption of the covariates.

## 6.5. The Bayesian estimates and inference

As opposed to the frequentist interpretation of probability as the frequency of events, the Bayesian interpret it as a degree of belief. In frequentist's framework, parameters are fixed but unknown vectors or functions, while in the Bayesian framework, the parameters are also random variables. The parameters $\theta$ as random variables assume a *prior* distribution $p(\cdot)$. The parameters determine the distribution of the variables that we observe in data. Thus data contains the evidence of values of the parameters and reshape the distribution of the parameters, which is called *posterior* distribution. Thus

$$f(\theta|\mathcal{D}) = \frac{f(\theta; \mathcal{D})}{f(\mathcal{D})} = \frac{f(\mathcal{D}|\theta)p(\theta)}{f(\mathcal{D})} \propto f(\mathcal{D}|\theta)p(\theta).$$

The simplest version of the Bayesian theorm is often expressed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $B$ represents the evidence event, $P(A)$ is our prior belief, and $P(A|B)$ represents our posterior belief after observing the evidence.

The Bayesian estimation and inference is then straightforward both conceptually and computationally. The Bayesian estimator of parameter $\theta$ is the posterior mean. Another is maximum a

posterior probability (MAP). Note that MAP maximizes $f(\mathcal{D}|\theta)p(\theta)$ which is different from MLE which maximizes $f(\mathcal{D}|\theta)$.

For example, suppose the conditional distribution of $X$ given $\theta$ is is normal, i.e., $X \sim N(\theta, \sigma^2)$ and the prior is normal, $\theta \sim N(\mu, \tau^2)$, then the posterior is also normal and the Bayes estimator is given by

$$\widehat{\theta}(x) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x.$$

If $x_1, ..., x_n$ are iid Poisson random variables $x_i|\theta \sim P(\theta)$, and if the prior is Gamma distributed $\theta \sim G(a, b)$, then the posterior is also Gamma distributed, and the Bayes estimator as the posterior mean is given by

$$\widehat{\theta}(X) = \frac{n\overline{X} + a}{n + \frac{1}{b}}.$$

In general, it is not true that the prior and posterior would fall into the same family of distributions.

There exists a striking difference between Bayesian posterior intervals or regions for parameters and the frequentiests' confidence intervals or regions for parameters. The frequentists's statement of confidence is the probability that the random interval would cover the fixed parameter, while in reality the interval constructed, after data are collected, is actually fixed. The Bayesian posterior interval is much more natural to interprete: the parameters are random and follows the posterior distribution, which is actually the conditional distribution given data. With the posterior distribution, in principle, one can choose the subset of $\theta$ with highest posterior density such that this subset has posterior probability $1 - \alpha$. However, in general, this may involve computational complexity. And a more convenient way is appealing to the central limit theory. In this case, the Bayesian posterior interval and the frequentiest confidence interval often differ little, especially for large sample size.

## 6.6. General Bayesian framework

In general, a Bayesian decision rule can be understood loosely as an estimator $\hat{\theta} = \hat{\theta}(\mathcal{D})$ of the parameter $\theta$. A loss function $L = L(\cdot, \cdot)$ is used to gauge the "loss" incurred by the estimation error. A commonly used loss function is the square loss: $L(x, y) = \|x - y\|^2$. Then the risk of the function of the decision rule is is the mean loss:

$$R(\theta, \hat{\theta}) = E\{L(\theta, \hat{\theta})|\theta\},$$

where the expectation is with respect to both the random $\theta$ and $\hat{\theta}$. The Bayes risk is $E\{(R(\theta, \hat{\theta})\}$. It is then natural to find the estimator $\hat{\theta}$ to minimize the Bayes risk. Notice that

$$E\{(R(\theta, \hat{\theta})\} = E\{L(\theta, \hat{\theta})\} = E\{E[L(\theta, \hat{\theta})|\mathcal{D}]\})$$

As a result, the Baysian estimator with loss function $L$ is

$$\hat{\theta} = \text{argmax}_{\hat{\theta}} E[L(\theta, \hat{\theta})|\mathcal{D}] = \text{argmax}_{\hat{\theta}} \int L(\theta, \hat{\theta}) f(\theta|\mathcal{D}) d\theta.$$

Recall that $f(\theta|\mathcal{D})$ is the posterior density.

Consider one dimensional $\theta$ for simplicity. It is then easy to see that, if $L$ is the square loss ($l^2$ loss), the the Bayes estimator is posterior mean; if $L$ is the absolute value loss ($l_1$ loss), then the Bayes estimate is the posterior median. If $L(x, y) = I(|x - y| > c)$, then the Bayesian estimator approaches the posterior mode as $c \to 0$.

## 6.7. The regularized estimation from Bayesian pointview

For linear regression models in Chapter 2:

$$y_i = \beta_0 + \mathbf{x}_i^T \beta + \epsilon_i,$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ are the inputs/covariates, $\beta = (\beta_1, ..., \beta_p)^T$ and $\epsilon_i \sim N(0, \sigma^2)$. Assume $\beta$ has a normal prior:

$$p(\beta) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\tau^2}} \exp\{-\beta_j^2/(2\tau^2)\}$$

(Here, $\beta_0$ has the inproper flat prior.) Then, the posterior density of $\beta$ is

$$f(\beta|\mathcal{D}) \propto \exp\Big(-\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i^T\beta)^2}{2\sigma^2} - \frac{\sum_{j=1}^{p}\beta_j^2}{2\tau^2}\Big),$$

which is multivariate normal distribution. The posterior mean is the well known ridge estimator of $\beta$.

$$\hat{\beta}_{\text{ridge}} = \text{argmin}_\beta \sum_{i=1}^{n}(y_i - \beta_0 - \beta^T\mathbf{x}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{y}$$

where $I_p$ is $I_p$ is the $(p+1)\times(p+1)$ diagonal matrix $\text{diag}(0, 1, 1, ..., 1)$ and $\lambda$ is a tuning parameter, which can be determined by cross-validation.

Assume $\beta$ has a double exponential prior:

$$p(\beta) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\tau}} \exp\{-\beta_j^2/\tau\}.$$

Then the MAP is the seminal LASSO estimator:

$$\hat{\beta}_{\text{LASSO}} = \text{argmin}_\beta \sum_{i=1}^{n}(y_i - \beta_0 - \beta^T\mathbf{x}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

When $\lambda = 0$, the regularized estimators reduce to the ordinary least squares estimator. When $\lambda$ increases, the estimator is shrunk towards 0. As $\lambda$ tends to $\infty$, the regualarized estimators tend to 0.

It may not be immediately clear to see the advantages of the least squares estimation under Bayesian framework or the regularized estimation. In fact, the oridinary least squares can also be viewed as special case of Bayesian estimates with flat prior. It implies there is no *prior* preference of the values of the parameter $\beta$. An astronomically large value of $\beta$ is equally likely as the value as ordinary as 0. In reality, this should not be the case. Constraining the scale of the parameters to be small in fact is contraining the model to be small or restrictive. A large model can easily lead to small training error and very good fit, which in the case of ordinary least squares, is high $R^2$ and small RSS. But it is likely to give large test error or prediction error, because of the tendency to overfit the training data. An overfitted model would have large variance, meaning that, if the collection of training data is replicated, the resulting fitted model would be very different. Constraining a large model to be a reasonably smaller one would control the variance at the cost of increased bias. This is the phenomenon of the *bias variance dilemma*. Through finding the best tuning parameter, usually by cross validation, we hope that the best model would balance the variance and bias and result in minimal mean squared error.

We note that the regularization can be phrased into the "Loss + Penalty" criterion and can be applied to other loss functions, such as the deviance function, which is minus 2 times the log-likelihoood.

## 6.8. Classification from Bayesian pointview

The logistic regression in Section 6.4 is one of the classical way for classification. Classification is one of the common tasks in statistics and machine learning. Assume that there are two classes in the population, such as virus carrier/non-carrier or buyer/non-buyer, with prior probability $p_1 = P(Y = 1)$ and $p_0 = P(Y = 0) = 1 - p_1$. Assume that the distribution of $X$ in class 1 is $f_1(\cdot)$ and that in class 0 is $f_0(\cdot)$. Then,

$$\delta_1(\mathbf{x}) = P(Y = 1|X = \mathbf{x}) = \frac{f_1(\mathbf{x})p_1}{f_X(\mathbf{x})} \quad \text{and} \quad \delta_0(\mathbf{x}) = P(Y = 0|X = \mathbf{x}) = \frac{f_0(\mathbf{x})p_0}{f_X(\mathbf{x})}$$

The Bayesian classfier is

$$\text{classify into} \begin{cases} \text{class 1} & \text{if } \delta_1(\mathbf{x}) \geq \delta_0(\mathbf{x}) \\ \text{class 0} & \text{if } \delta_1(\mathbf{x}) < \delta_0(\mathbf{x}). \end{cases}$$

Bayesian classifier is optimal in some sense but may not be available since $f_1$ and $f_0$ may not be entirely known. A common model is that $f_1$ and $f_0$ are both multivariate normal with mean $\mu_1$ and $\mu_2$ and common variance $\Sigma$. The parameters $\mu_1$, $\mu_2$ and $\Sigma$ are generally assumed unknown, and are estimated by sample means $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and pooled estimator of the variance $\mathbf{S}$, as in the classical two sample problem. This results in the estimated discrimination function

$$\hat{\delta}_k(\mathbf{x}) = \bar{\mathbf{x}}_k^T \mathbf{S}^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_k \mathbf{S}^{-1}\bar{\mathbf{x}} + \log p_k, \qquad k = 0, 1$$

and the decision rule is, classify the observation with covariate $\mathbf{x}$ into class $k$ if $\hat{\delta}_k(\mathbf{x})$ is the largest of $\hat{\delta}_1(\mathbf{x})$ and $\hat{\delta}_0(\mathbf{x})$. This is the linear discriminant analysis (LDA) approach.

When $p$ is of high dimension, particularly when $p > n$, it is inconvenient to model $f_1$ and $f_0$ with normal distribution, in which case $\mathbf{S}$ could be singular. The method of *naive Bayes* is often employed. The naive Bayes assumes The inputs $X_1, ..., X_p$ conditioning on $Y = 1$ or 0 are conditionally independent. Namely $f_k(\mathbf{x})$ can be written as products of functions, each being a function of a single $x_j$ alone. Then the Bayes discrimination functions are additive functions of $x_1, ..., x_p$. The resulting model can be analyzed by using techniques in generalized additive model (GAM). The naive Bayes is often practically accurate, even though the assumption of conditional marginal independence is untrue.

## 6.9. Examples

**Data Explanation**

In this part, we consider the problem of predicting daily returns of CSI 300 index. The period of data ranges from April 7th, 2011 to June 13th, 2016, totally 1261 trading days. We choose the last 100 trading days as the testing set (from January 13th, 2016 to June 13th, 2016) and train models in the previous 1161 trading days.

In finance, technical analysis, aside from the fundamental analysis, is a major class of methodologies for forecasting future prices. In technical analysis, historical market data, primarily price and volume, are used to create various indicators which may capture some important features of security price and have the potential to predict future price movement.

Here we choose 49 widely-used indicators as covariates to predict future daily returns of CSI 300 index. Their names are listed in Table 1. For example, MA5 means price moving average over the last 5 trading days, MA10.AMT means moving average over the last 10 trading days for trading amount. CCI, short for commodity channel index, is an oscillator originally introduced by Donald Lambert in 1980. It measures a securitys variation from the statistical mean. It can be calculated by the following formula,

$$CCI = \frac{1}{0.015}\frac{p_t - MA(p_t)}{\sigma(p_t)}$$

Here $p_t = (High + Low + Close)/3$ is called typical price, which can treated as the average price of a single day, sometimes more useful than widely-used close price. $MA(p_t)$ is the moving average of

$p_t$ (here we choose the period as 14 trading days, which is the default setting in Lambert's study). $\sigma(p_t)$ is the standard deviation. The indicator is scaled by an inverse factor of 0.015 to provide more readable numbers.

| | | | | | | |
|---|---|---|---|---|---|---|
| ADTM | CDP | SLOWKD | ROC | TAPI | WVAD | MA60 |
| ATR | DMA | MACD | RSI | TRIX | VOL_RATIO | MA5.AMT |
| BBI | DMI | MIKE | SAR | VHF | MA5 | MA10.AMT |
| BBIBOLL | DPO | MTM | SI | VMA | MA10 | MA15.AMT |
| BIAS | ENV | PRICEOSC | SOBV | VMACD | MA15 | MA20.AMT |
| BOLL | EXPMA | PVT | SRMI | VOSC | MA20 | MA30.AMT |
| CCI | KDJ | RC | STD | VSTD | MA30 | MA60.AMT |

Table 1: Covariate Names

For other indicators, readers can find their meaning and formula in this webpage http://www.incrediblecharts.com/topic/Indicator_Guide.

Since the values of these indicators have different scales and the change of values are usually more important, we consider their daily changes as the true covariates in our models. For example, we will not use the values of $CCI$ but $CCI_{today}/CCI_{yesterday} - 1$ to predict tomorrow's daily return (that is, $CLOSE_{tomorrow}/CLOSE_{today} - 1$). The covariates must not contain any concurrent or future information of the responses to be predicted. For example, one cannot use today's CCI to predict today's daily return.

**Example 1**

Consider a relatively simpler problem of just predicting prices "increase" or "decrease", i.e. daily returns are positive or negative. Here we transfer daily return to categorical day and apply the logistic regression on the training set. Part of the results of the model fit are shown in Table 2.

We found that only a few covariates, such as MA5.AMT and MA10.AMT, are statistically significant. But does this model give the training a good fit? One can use pseudo R-squareds to measure the goodness of fit. For example, McFadden's R-squared, which is defined as following:

$$R^2_{McFadden} = 1 - \frac{ln(L_M)}{ln(L_0)}$$

Here $L_M$ means the likelihood for the model being estimated. $L_0$ means the value of likelihood function for a null model (i.e. model with no predictors, only intercept). It plays a role analogous to the residual sum of squares in linear regression. For our model, we can calculate $R^2_{McFadden} = 0.04351681$, which is a quite small number. As the definition of McFadden's R-squared, a small value near to 0 means the model has little improvement beyond the null model.

To check the predictive performance on the test data, we set the decision boundary as 0.5 (i.e. if $P(y = 1|X) > 0.5$ then $y = 1$ otherwise $y = 0$). Then we calculated that the hit accuracy on the testing set is only 0.48, which is even slightly worse than random guess. In fact, in the test data, there are 52 out of 100 days which CSI 300 has positive return. But in the prediction result, there are only 14 out of 100 days which has $P(y = 1|X) > 0.5$. This phenomenon implies our model contains huge bias. The likely reasons are the nonstationarity of the price process and the linearity of the components contribution to the outcomes.

| | *Dependent variable:* |
|---|---|
| | RETURN |
| ADTM | 0.016 |
| | (0.015) |
| ATR | −0.163 |
| | (0.111) |
| BBI | 144.236* |
| | (81.062) |
| ... | ... |
| MA60 | −40.208 |
| | (47.530) |
| MA5.AMT | −13.475** |
| | (5.471) |
| MA10.AMT | 7.895*** |
| | (3.018) |
| MA15.AMT | −4.747 |
| | (4.000) |
| MA20.AMT | −6.249 |
| | (5.074) |
| MA30.AMT | −5.820 |
| | (6.535) |
| MA60.AMT | −0.908 |
| | (10.056) |
| Constant | 0.037 |
| | (0.067) |
| Observations | 1,161 |
| Log Likelihood | −769.691 |
| Akaike Inf. Crit. | 1,637.381 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 2: Logistic Regression Results

**Example 2**

Continued with Example 1, we compared the prediction performance of Lasso regression and ridge regression. Recall that ridge doesn't have the ability of variable selection while lasso usually has a sparse solution.

It turns out that Lasso only selected two variables (MA5.AMT and MA20.AMT, with coefficients -1.074059 and -4.259734 respectively) out of total 49 covariates, implying that most technical indicators in our dataset may not be suitable for linear model although they do have sophisticated meaning. Note that here we use 10 folds cross validation to choose the tuning parameter $\lambda$ and picked the parameter which minimizes the CV error.

For ridge regression, although there is no variable selection, most coefficients are rather small. Among them, MA20 and MA60 have relatively large coefficients (-1.078728 and -1.293058 respectively).

| | training | testing |
|---|---|---|
| Lasso | 0.5443583 | 0.47 |
| Ridge | 0.543497 | 0.55 |

Table 3: Hitting rate of Lasso and Ridge

In Table 3, we compare the hitting rates for these two types of regression. From the result, we observed that in the training set, the hitting rates of Lasso and ridge are very closed. But ridge has a higher hitting rate in the testing set, which is 55%, better than random guessing. However, when we consider the predicted probability instead of predicted class (i.e. 0 or 1) in Table 4, we found that most prediction values are closed to 0.5, implying that the ridge regression may not be as good as it appears if other criterion the prediction error rate is used.

| date | predicted probability | true result |
|---|---|---|
| 2016-01-13 | 0.486 | 1 |
| 2016-01-14 | 0.497 | 0 |
| 2016-01-15 | 0.477 | 1 |
| 2016-01-18 | 0.477 | 1 |
| 2016-01-19 | 0.490 | 0 |
| ... | ... | ... |
| 2016-05-27 | 0.515 | 1 |
| 2016-05-30 | 0.496 | 1 |
| 2016-05-31 | 0.536 | 0 |
| 2016-06-01 | 0.514 | 1 |
| 2016-06-02 | 0.506 | 1 |
| 2016-06-03 | 0.522 | 0 |
| 2016-06-06 | 0.511 | 0 |
| 2016-06-07 | 0.497 | 0 |
| 2016-06-08 | 0.502 | 0 |
| 2016-06-13 | 0.515 | 1 |

Table 4: Predicted probability of ridge regression in the testing set