Scientific Research Publishing

Improvement of Misclassification Rates of Classifying Objects under Box Cox Transformation and Bootstrap Approach

Mst Sharmin Akter Sumy¹, Md Yasin Ali Parh¹, Ajit Kumar Majumder², Nayeem Bin Saifuddin³

¹Department of Statistics, Islamic University, Kushtia, Bangladesh
²Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh
³Dhaka Nursing College, Dhaka, Bangladesh
Email: sh.ak.sumy@gmail.com, yali@isrt.ac.bd, ajit@juniv.edu, nayeem.dnc.bd.@gmail.com

Abstract

Discrimination and classification rules are based on different types of assumptions. Also, all most statistical methods are based on some necessary assumptions. Parametric methods are the best choice if it follows all the underlying assumptions. When assumptions are violated, parametric approaches do not provide a better solution and nonparametric techniques are preferred. After Box-Cox transformation, when assumptions are satisfied, parametric methods provide fewer misclassification rates. With this problem in mind, our concern is to compare the classification accuracy of parametric and non-parametric approaches with the aid of Box-Cox transformation and Bootstrapping. We carried Support Vector Machines (SVMs) and different discrimination and classification rules to classify objects. The attention is to critically compare the SVMs with Linear discrimination Analysis (LDA), and Quadratic discrimination Analysis (QDA) for measuring the performance of these techniques before and after Box-Cox transformation using misclassification rates. From the apparent error rates, we observe that before Box-Cox transformation, SVMs perform better than existing classification techniques, on the other hand, after Box-Cox transformation, parametric techniques provide fewer misclassification rates compared to nonparametric method. We also investigated the performances of classification techniques using the Bootstrap approach and observed that Bootstrap-based classification techniques significantly reduce the classification error rate than the usual techniques of small samples. Thus, this paper proposes to apply classification techniques under the Bootstrap approach for classifying objects in case of small sample. A real and simulated datasets application is carried out to see the performance.

Keywords

Misclassification Rate, SVM, Box Cox Transformation, Bootstrapping

1. Introduction

Most of the statistical decisions can be made with the aid of explanatory data analysis, test of hypothesis, and from the suitable statistical model. In real life, most of the models are nonlinear. Linear models are very simple to estimate, test, as well as forecast. But nonlinear or generalized linear models are appropriate in real-life situations. Nonlinear models can be transformed into linear models to avail the benefits of the linear model by different transformation techniques. To overcome this difficulty of transformation techniques, several nonparametric procedures can be used. SVMs are one of the popular nonparametric methods to handle nonlinear model [1].

In data mining, nonparametric model is one that is data-driven. No explicit equations are used to determine the model. Parametric methods are the best choice if it follows all the underlying assumptions of the assumed models [2]. Also, all most statistical models are based on some assumptions. When assumptions are violated, parametric approaches can't be better work, in that case, we should use nonparametric techniques to obtain better performance. In practice, researchers often carry out the experiment without checking assumptions and apply parametric methods. If most of the assumptions are violated, then without checking it is better to use nonparametric and robust methods or some other data mining techniques.

Violation of these assumptions can seriously increase the chances of the researcher committing misleading classification rules. Naturally, question arises, when observations are not independently distributed, observations are not normally distributed, and observations have unequal variance-covariance matrix [3]. According to the above situations, can we observe better performances of all discrimination and classification rules?

In order to overcome the above problem, we can transform non-normal data to near normal by using Box-Cox transformation [4]. It is not always necessary or desirable to transform a data set to resemble a normal distribution. Our aim is to identify the best technique of classification in all situations. When assumptions are not satisfied then without checking we can use SVMs as a nonparametric method.

Classification techniques cannot usually provide an error-free method of assignment [3]. This is because there may not be a clear distinction between the measured characteristics of the populations: that is the groups which may overlap. Classification accuracy rates can be improved by using Bootstrapping. We addressed this issue of classification errors over small samples, investigated the performances of classification techniques, and observed that Bootstrap-based

classification techniques significantly reduce the classification error than the usual techniques of small samples. Thus, this paper proposes to apply classification techniques under the Bootstrap approach for classifying objects in case of small sample.

2. Data Information

The data was obtained from Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer. The dataset consists of post-operative life expectancy in lung cancer patients from 470 distinct individuals across 17 distinct attributes. The response variable has two classes, class 1, death within one year after surgery, class 2, survival [5]. We split the dataset into training with two-thirds of observations and rest of the observations as test dataset. We split the dataset as a rule of thumb. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we evaluated the performance of our models.

3. Methods

There were two main phases of the analysis. First, parametric, and nonparametric classification methods were used to see the classification accuracy before and after Box-Cox transformation. Finally, we applied these classification techniques under the Bootstrap approach for classifying objects in the case of small and large samples to see the performance.

3.1. Non-Parametric Classification Algorithm

Support Vector Machines (SVMs)

SVMs [6], are a classification method which has drawn tremendous attention in machine learning. They belong to a family of generalized linear models which achieves a classification decision based on the value of the linear combination of the features. They are also said to belong to "kernel methods". A special property of SVMs is, SVMs simultaneously minimize the empirical classification error and maximize the geometric margin.

Linear SVMs

Let us consider a binary classification task with data points x_i $(i = 1, \cdots, m)$, having corresponding labels $y_i = \pm 1$ and let the decision function be:

$$f(x) = \text{sign}(\mathrm{w} \cdot \mathrm{x} + b) \tag{1}$$

where, $\cdot$ is the scalar or inner product (so $\mathrm{w} \cdot \mathrm{x} \equiv \mathrm{w}^\mathrm{T} \mathrm{x}$). From the decision function we see that the data is correctly classified if $y_i(\mathrm{w} \cdot \mathrm{x}_i + b) > 0 \; \forall i$ since $(\mathrm{w} \cdot \mathrm{x}_i + b)$ should be positive when $y_i = +1$, and it should be negative when $y_i = -1$. This leads the concept of distance or margin. Hence, we define a scale for (w, b) by setting $\mathrm{w} \cdot \mathrm{x} + b = 1$ for the closest points on one side and $\mathrm{w} \cdot \mathrm{x} + b = -1$ for the closest on the other side. The hyperplanes passing through $\mathrm{w} \cdot \mathrm{x} + b = 1$ and $\mathrm{w} \cdot \mathrm{x} + b = -1$ are called canonical hyperplanes, and the region between these canonical hyperplanes are called the margin band [6].

Let x_1 and x_2 be two points inside the canonical hyperplanes on both sides. If $w \cdot x_1 + b = 1$ and $w \cdot x_2 + b = -1$, we deduce that $w \cdot (x_1 - x_2) = 2$. For the separating hyperplane $w \cdot x + b = 0$, the normal vector is $w / \|w\|_2$ (where $\|w\|_2$ is the square root of $w^T w$). Thus, the distance between the two canonical hyperplanes is equal to the projection of $x_1 - x_2$ onto the normal vector $w / \|w\|_2$, which gives $(x_1 - x_2) \cdot w / \|w\|_2 = 2 / \|w\|_2$ maximizing the margin is therefore equivalent to minimizing:

$$\frac{1}{2} \|w\|_2^2 \tag{2}$$

subject to the constraints:

$$y_i (w \cdot x_i + b) \geq 1 \quad \forall i \tag{3}$$

This is a constraint optimization problem in which we minimize an objective function (2) subject to the constraints (3), [6]. As a constrained optimization problem, the above formulation can be reduced to minimization of the following Lagrange function, consisting of the sum of the objective function and the *m* constraints multiplied by their respective Lagrange multipliers. We can call this the primal formulation:

$$L(w,b) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{m} \alpha_i \left(y_i (w \cdot x_i + b) - 1 \right) \tag{4}$$

where, α_i are Lagrange multipliers, and thus $\alpha_i \geq 0$ at the minimum, we can take the derivatives with respect to *b* and w, and set there to zero.

Kernel Selection

SVMs sensitive to the proper choice of parameters, so we checked the range of parameter combinations. In order to improve the performance of the support vector classification, we needed to select the best parameters for the model. To find misclassification rates, we used Radial Basis kernel, because of its good general performance and the few numbers of parameters (*C* and γ). First, we used cross validation to find better *C* and finally decided which γ values need to be used for the better *C*.

3.2. Transforming Univariate Observations

Multivariate transformation is accomplished by applying a possibly different univariate transformation to each of the components of the multivariate data. The most used univariate transformation family is the Box-Cox power transformation family. A convenient analytical is available for choosing a power transformation. In the case of univariate analysis Box and Cox consider the slightly modified family of power transformations

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{5}$$

which, is continuous in λ for $x > 0$.

Manly (1976) suggests applying Box-Cox transformation to the exponential data. Thus, Manly's transformation is,

$$h(x;\lambda) = \begin{cases} \dfrac{\exp(x^\lambda) - 1}{\lambda} & \lambda \neq 0 \\ x & \lambda = 0 \end{cases} \tag{6}$$

Box-Cox suggests that a power transformation can be selected by maximum likelihood or Bayesian estimation. Given the observations $x_1, x_2, \cdots, x_n$, the Box-Cox solution for the choice of an appropriate power λ is the solution that maximizes the expression below:

$$l(\lambda) = -\frac{n}{2} \ln\left[\frac{1}{n} \sum_{j=1}^{n} \left(x_j^{(\lambda)} - \bar{x}^{(\lambda)}\right)^2\right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j \tag{7}$$

where,

$$\bar{x}^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_j^\lambda - 1}{\lambda}\right) \tag{8}$$

This estimation procedure can be applied to the exponential data to estimate the parameter in Manly's transformation family [3].

3.3. Linear Discriminant Analysis (LDA)

LDA models the distribution of the predictors separately in each of the response classes and then uses Bayes' theorem to flip them around into estimates. For more than one predictor, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate gaussian distribution which has a class-specific mean and common variance. Class means and common variance must be estimated from the data, and once obtained are then used to create linear decision boundaries in the data. LDA then simply classifies an observation according to the region in which it is located [7].

The simplest procedure is to calculate a linear discriminant for each class, this discriminant being just the logarithm of the estimated probability density function for the appropriate class, with constant terms dropped. Where the prior class proportions are unknown, they would be estimated by the relative frequencies in the training set. Suppose the prior probability of class A_i is π_i, and that $f(x_i)$ is the probability density of x in A_i class and is the normal density equation.

$$\frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^{\text{T}} \Sigma^{-1}(x - \mu)\right) \tag{9}$$

The joint probability of observing class A_i and attribute x is $\pi_i f_i(x)$ and the logarithm of the probability of observing class A_i and attribute x is,

$$\log \pi_i + x^{\text{T}} \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^{\text{T}} \Sigma^{-1} \mu_i \tag{10}$$

to within an additive constant. So, the coefficients β_i are given by the coeffi-

cients of *x*.

$$\beta_i = \Sigma^{-1}\mu_i \tag{11}$$

and the additive constant α_i by,

$$\alpha_i = \log \pi_i - \frac{1}{2}\mu_i^{\mathrm{T}}\Sigma^{-1}\mu_i \tag{12}$$

though these can be simplified by subtracting the coefficients for the last class.

The above formulae are stated in terms of the (generally unknown) population parameters Σ, μ_i and π_i. To obtain the corresponding "plug-in" formulae, substitute the corresponding sample estimators: S for Σ; $\bar{x}_i$ for μ_i; and p_i for π_i, where p_i is the sample proportion of class A_i examples [8].

3.4. Quadratic Discriminant Analysis (QDA)

QDA is very similar to LDA but does not assume constant variance across classes. Heterogenous class variances change the decision boundaries from linear to quadratic, thus changing the behavior of the classifier. LDA is a simpler model with a higher bias but less variation. QDA is a more flexible model that has a lower bias but higher variance. LDA will outperform QDA when the decreases in bias are outweighed by the increases in variance [9].

The quadratic discriminant function is most simply defined as the logarithm of the appropriate probability density function, so that one quadratic discriminant is calculated for each class. Taking the logarithm and allowing for differing prior class probabilities π_i we obtain

$$\log \pi_i f_i(x) = \log(\pi_i) - \frac{1}{2}\log(|\Sigma_i|)\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i) \tag{13}$$

as the quadratic discriminant for class A_i.

In classification, the quadratic discriminant is calculated for each class and the class with the largest discriminant is chosen. To find the posteriori class probability explicitly, the exponential is taken of the discriminant and the resulting quantities normalized to sum to unity. Thus, the posterior class probabilities $P(A_i|x)$ are given by,

$$P(A_i|x) = \exp\left[\log(\pi_i) - \frac{1}{2}\log(|\Sigma_i|)\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i)\right] \tag{14}$$

The most frequent problem with quadratic discriminants is caused when some attribute has zero variance in one class, for then the covariance matrix cannot be inverted. One way of avoiding this problem is to and a small positive constant term to the diagonal terms in the covariance matrix (this corresponds to adding random noise to the attributes). Another way, adopted in our own implementation, is to use some combination of the class covariance and the pooled covariance. [10] [11].

3.5. Cross Validation (CV)

Cross validation [12] involves randomly dividing the set of observations into K

"folds" of approximately equal size. The first fold is treated as a validation set, and the model is trained on the remaining $K - 1$ folds of data. This trained model is then used to predict the target in the Kth fold, and an accuracy metric, ACC_1, is computed. This procedure is repeated K times where a new validation set is used during each iteration. This process results in K estimates of the test error: $ACC, ACC_2, \cdots, ACC_k$. The K fold CV estimate is computed by averaging these values,

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{n} ACC_i \tag{15}$$

This article required classifying observations into one of two categories; therefore, accuracy, misclassification rate, and the kappa statistics were used as accuracy measures in CV.

3.6. Kappa Statistic

Accuracy can be a misleading metric since it is possible to make correct classifications simply by chance alone. The following is the formula for calculating the kappa statistic:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{16}$$

In this formula, $Pr(a)$ refers to the proportion of actual agreement and $Pr(e)$ refers the probability of making a correct classification purely by chance. Kappa values range from 0 to a maximum of 1. 1 indicates perfect agreement, a value of 0 indicating no agreement, and values between 0 and 1 indicating varying degrees of agreement. Depending on how a model is to be used, the interpretation of the kappa statistic might vary. Traditional metrics such as precision, recall, and specificity can still be calculated with multiple classes, but the objective of this analysis was overall accuracy, not a specific error rate [12] [13].

3.7. Bootstrap Methods

The Bootstrap is a computer intensive re-sampling technique introduced by Efron [14] where theoretical statistics are difficult to obtain. The ability to do a lot of computation extremely fast has led to the use of techniques that provide "new" sets of data by re-sampling numbers generated from a single data set [15]. The bootstrap estimate of the sampling distribution is generally better than the normal approximation based on the central limit theorem [16], even if the statistic is not following the distribution F with mean, μ and variance σ^2. The standard bootstrap procedure is to draw with replacement a random sample of size n from $X_1, X_2, \cdots, X_n$. Denote the bootstrap sample by $X_1^*, X_2^*, \cdots, X_n^*$ and denote their mean and standard deviation by $\bar{\bar{X}}_n^*$ and, S_n^*. Suppose F_n indicate the empirical distribution of $X_1, X_2, \cdots, X_n$, then the sampling distribution of $\left(\bar{\bar{X}}_n^* - \bar{\bar{X}}_n \right)$ under F_n is the bootstrap approximation of the sampling distribution of $\left(\bar{\bar{X}}_n - \mu \right)$ under F. The bootstrap technique provides the mean $\hat{\theta}_B$

of all the bootstrap estimators $\hat{\theta}_B = \dfrac{\sum_{i=1}^{B} \hat{\theta}_i}{B}$, where $\hat{\theta}_i$ is the estimate using the ith bootstrap sample and B is the number of Bootstraps. The idea behind the bootstrap is very simple, namely that (in the absence of any other information), the sample itself offers the best guide of the sampling distribution. By re-sampling with replacement from the original sample, we can create a bootstrap sample, and use the empirical distribution of our estimator in a large number of such bootstrapped samples to construct confidence intervals and tests for significance.

4. Findings

It is important to evaluate how well SVMs perform in comparison with Linear Discrimination Analysis (LDA), and Quadratic Discrimination Analysis (QDA). This section compares the misclassification rates of SVMs with the existing classification techniques.

Table 1 represents that the SVMs have been successfully trained, the results produced by it are often superior to those obtained from the usual discrimination and classification rules before Box-Cox transformation. On the other hand, after Box-Cox transformation when assumptions are satisfied, although apparent error rate of SVMs is decreased but apparent error rates of existing methods are decreased significantly.

Table 2 shows the same performance of SVMs when we applied these procedures in simulated data. We then apply these techniques to the original data set under Bootstrapping. The results are given in the following subsequent tables.

From Table 3, it is clear that, SVMs, LDA, and QDA give the apparent error rates for initial data sets are 22.2%, 30.0%, and 25.5% respectively, whereas under Bootstrapping these classification techniques reduces this misclassification rates significantly and reaches to 16.20%, 24.51%, and 19.20% respectively for 2000 Bootstrapping. Also, we observe that apparent error rates decrease if the number of Bootstrapping increases but after a certain number of Bootstrapping the difference of the apparent error rates between current and previous Bootstrapping is negligible.

From Table 4 we also observe that misclassification error rates decrease if the number of Bootstrapping increases but after a certain number of Bootstrapping

Table 1. Misclassification error rates before and after box-cox transformation of thoracic surgery data.

Classification methods	APER in % Before Box-Cox transformation			APER in % After Box-Cox transformation		
	SVM	LDA	QDA	SVM	LDA	QDA
Cross Validation Kappa Statistic	40.5	52.0	53.1	38.1	32.2	34.0
Classification error rate	40.0	51.0	53.0	37.0	32.1	33.0

Table 2. Misclassification error rates before and after box-cox transformation of simulated data.

Classification methods	APER in % Before Box-Cox transformation			APER in % After Box-Cox transformation		
	SVM	LDA	QDA	SVM	LDA	QDA
Cross Validation Kappa Statistic	22.2	30.0	25.5	21.0	10.2	8.0
Classification error rate	21.1	29.9	23.0	21.0	10.2	7.5

Table 3. Results for thoracic surgery data under bootstrapping.

	Apparent Error Rate (APER) in %		
Initial Sample (Without Bootstrap)	SVM	LDA	QDA
	22.2	30.0	25.5
(No. of Bootstrap)	Apparent Error Rate (APER) in %		
50	17.99	25.75	20.57
250	17.34	25.21	20.25
500	17.99	25.20	20.01
Bootstrap results 750	16.29	25.12	19.65
1000	16.21	24.52	19.21
2000	16.20	24.51	19.20

Table 4. Results for simulated data under bootstrapping.

	Apparent Error Rate (APER) in %		
Initial Sample (Without Bootstrap)	SVM	LDA	QDA
	15.2	25.00	16.70
(No. of Bootstrap)	Apparent Error Rate (APER) in %		
50	10.99	21.75	12.57
250	10.34	21.21	12.25
500	9.99	20.79	12.01
Bootstrap results 750	9.29	20.25	11.65
1000	8.21	20.02	11.21
2000	8.20	18.99	11.20

the difference of the misclassification rates between current and previous Bootstrapping is negligible. It is clear that, SVMs, LDA, and QDA give the apparent error rates for initial data sets are 15.2%, 25.0%, and 16.70% respectively, whereas under Bootstrapping these classification techniques reduces this misclassification rates significantly and reaches to 8.20%, 18.99%, and 11.20% respectively

for 2000 Bootstrapping samples.

5. Summary and Conclusions

The success of this research relies on correctly classifying patients into two classes of response variables. Many different classification methodologies were used to obtain classification accuracies or misclassification rates. However, the misclassification rates of the classification methods present several conclusions about the integrity of the results. Support vector machines (SVMs) are special kernel-based nonparametric methods that don't depend on different types of necessary assumptions. When assumptions are violated the Support Vector Machines perform better results than other existing parametric methods. But in such a case, when assumptions are satisfied (by applying Box-Cox transformation) parametric methods give better results than SVMs. SVMs require some sophisticated computer programming, which is not easily accessible. So, without checking assumptions or prior to using existing parametric methods we can use support vector machines.

From our analysis, we also investigated that, in the case of small samples classification techniques significantly reduce the classification errors under Bootstrapping. It is clear from our analysis that, classification techniques under Bootstrapping perform better than the usual techniques for small samples. Thus, we may conclude that, in the case of small samples, we propose to apply classification techniques under Bootstrapping for classifying objects.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xie, W., She, Y. and Guo, Q. (2021) Regression on Classification Based on Improved SVM Algorithm for Balanced Binary Decision Tree. *Scientific Programming*, **2021**, Article ID: 5560465. https://doi.org/10.1155/2021/5560465

[2] Conover, W.J. (1980) Practical Nonparametric Statistics. Wiley Series in Probability and Statistics, 2nd Edition. Wiley, Hoboken.

[3] Johnson, R.A. and Wichern, D.W. (2002) Applied Multivariate Statistical Analysis. 5th Edition, Pearson Education (Singapore) Pte. Ltd., Singapore.

[4] Reddy, B.V.R., Pagadala, B. and Rayalu, G.M. (2011) Analysis of Transformations and Their Applications in Statistics: Extended Box and Cox Transformation Regression. LAP LAMBERT Academic Publishing, Saarbrücken, Germany.

[5] UCI (n.d.) Thoracic Surgery Data Set. https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

[6] Campbell, C. and Ying, Y. (2011) Learning with Support Vector Machines (Synthesis Lectures on Artificial Intelligence and Machine Learning).1st Edition, Morgan & Claypool Publishers, San Rafael.

[7] Izenman, A.J. (2008) Modern Multivariate Statistical Techniques: Regression, Clas-

sification, and Manifold Learning. Springer, New York.
https://doi.org/10.1007/978-0-387-78189-1

[8]   Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994) Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York.
http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf

[9]   Camilo, L.M.M., Lima, K.M.G., Martin, F.L. (2019) Uncertainty Estimation and Misclassification Probability for Classification Models Based on Discriminant Analysis and Support Vector Machines. *Analytica Chimica Acta*, **1063**, 40-46.
https://doi.org/10.1016/j.aca.2018.09.022

[10] Sumy, M.S.S, Parh, M.Y.A. and Hossain, M.S. (2021) Identifying and Classifying Traveler Archetypes from Google Travel Reviews. *International Journal of Statistics and Applications*, **11**, 61-69.

[11] Wahl, P. and Kronmal, R. (1977) Discriminant Functions When Covariances Are Unequal and Sample Sizes Are Moderate. *Biometrics*, **33**, 479-484.
https://doi.org/10.2307/2529362

[12] Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: With Applications in R. Springer, New York.
https://doi.org/10.1007/978-1-4614-7138-7

[13] Delgado, R. and Tibau, X. (2019) Why Cohen's *Kappa* Should Be Avoided as Performance Measure in Classification. *PLoS ONE*, **14**, Article ID: e0222916.
https://doi.org/10.1371/journal.pone.0222916

[14] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26. https://doi.org/10.1214/aos/1176344552

[15] Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. 1st Edition, Chapman and Hall, London.

[16] Bickel, P.J. and Freedman, D.A. (1981) Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, **9**, 1196-1217. https://doi.org/10.1214/aos/1176345637