Visualizing Some Bootstrap Confidence Regions

Welagedara, W.A.D.M. and David J. Olive *
Southern Illinois University

January 5, 2023

Abstract

The DD plot for visualizing prediction regions can also be used to visualize three bootstrap confidence regions.

KEY WORDS: Prediction Regions.

1 INTRODUCTION

This section reviews prediction intervals, prediction regions, confidence intervals, confidence regions, and visualizing prediction regions. Several bootstrap confidence intervals and regions are obtained by applying prediction intervals and regions to the bootstrap sample. Notation: $P(A_n)$ is "eventually bounded below" by $1-\delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1-\delta$ as $n \to \infty$. Hence $P(A_n) > 1-\delta-\epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1-\delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1-\delta$. The actual coverage is $1-\gamma_n = P(Y_f \in [L_n, U_n])$, the nominal coverage is $1-\delta$ where $0 < \delta < 1$. The 90% and 95% large sample prediction intervals and prediction regions are common. This section follows Olive (2014, ch. 9; 2017, $\oint$ 5.3) closely. Also see Olive (2023abcd).

Section 2 reviews the bootstrap while section 3 shows how to visualize some bootstrap confidence regions.

1.1 Prediction Intervals and Regions

Consider predicting a future test value Y_f given training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

*

Let the data $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ have joint pdf or pmf $f(\boldsymbol{y}|\theta)$ with parameter space Θ and support $\mathcal{Y}$. Let $L_n(\boldsymbol{Y})$ and $U_n(\boldsymbol{Y})$ be statistics such that $L_n(\boldsymbol{y}) \leq U_n(\boldsymbol{y})$, $\forall \boldsymbol{y} \in \mathcal{Y}$. Then $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a 100 $(1-\delta)$ % *confidence interval* (CI) for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1 - \delta$$

for all $\theta \in \Theta$. The interval $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a large sample $100(1-\delta)$ % CI for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y}))$$

is eventually bounded below by $1 - \delta$ for all $\theta \in \Theta$ as the sample size $n \to \infty$.

A *large sample* $100(1-\delta)$% *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

A *large sample* $100(1-\delta)$% *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

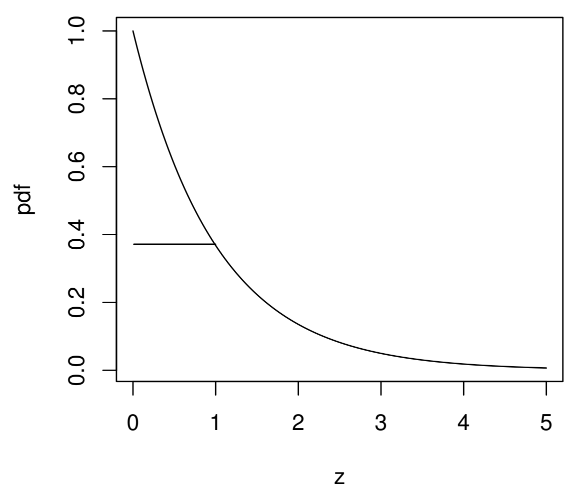


Figure 1: The 36.8% Highest Density Region is [0,1]

For a random variable Y, the $100(1-\delta)$% highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)$% highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1 - \delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

The interpretation of a 100 $(1 - \delta)$% PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times. If Y_f has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated.

In the following theorem, if the open interval $(Y_{(k_1)}, Y_{(k_2)})$ was used, we would need to add the regularity condition that $Y_{\delta/2}$ and $Y_{1-\delta/2}$ are continuity points of $F_Y(y)$.

Definition 1. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1 - \delta/2) \rceil$ where $0 < \delta < 1$. The large sample $100(1 - \delta)$% *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{1}$$

The bootstrap percentile confidence interval given by Equation (2) is obtained by applying the percentile prediction interval to the bootstrap sample $T_1, ..., T_B$.

Definition 2. The large sample $100(1 - \delta)$% *bootstrap percentile confidence interval* for θ is an interval $[T^*_{(k_L)}, T^*_{(K_U)}]$ containing $\approx \lceil B(1 - \delta) \rceil$ of the T^*_i. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1 - \delta/2) \rceil$. A common choice is

$$[T^*_{(k_1)}, T^*_{(k_2)}]. \tag{2}$$

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T_n = T_n(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}_n = \boldsymbol{C}_n(\boldsymbol{W})$ be a dispersion estimator.

Definition 3. Let $x_{1j},...,x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\text{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$.

Definition 4. Let $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1,...,\overline{x}_p)^T.$$

The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T_n, \boldsymbol{C}_n) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

Definition 5. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T_n(\boldsymbol{W}), \boldsymbol{C}_n(\boldsymbol{W})) = (\boldsymbol{x}_i - T_n(\boldsymbol{W}))^T \boldsymbol{C}_n^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T_n(\boldsymbol{W})) \tag{3}$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T_n, \boldsymbol{C}_n) = (T_n(\boldsymbol{W}), \boldsymbol{C}_n(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T_n, \boldsymbol{C}_n) = (\boldsymbol{x} - T_n)^T \boldsymbol{C}_n^{-1}(\boldsymbol{x} - T_n).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. If $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{4}$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0,1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Next, we derive a prediction region for $\boldsymbol{x}_f$ if $(T_n, \boldsymbol{C}_n) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, $\boldsymbol{\mu} = E(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$ is nonsingular. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cdf of D, and the sample percentiles of the D_i^2 are consistent estimators of the population percentiles of D^2 at continuity points of the cdf of D^2. Let $c = k_n = \lceil n(1-\delta) \rceil$. Then Olive (2013) showed that the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\} \tag{5}$$

is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist.

To improve performance, we will use a correction factor $c = U_n$ where U_n decreases to k_n. U_n is defined under Equation (7). A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (5) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data than on training data. Empirically for many distributions, for $n = 20p$, the prediction region (5) applied to iid data using $c = k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as $\min(0.05, \delta/2)$. The undercoverage decreases rapidly as n increases. (Referring to the next paragraph, taking $q_n \equiv 1-\delta$ does not take into account the unknown variability of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$, which is another reason for undercoverage and the need for a correction factor.)

Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{6}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Using

$$c = \lceil n q_n \rceil \tag{7}$$

in (4.8) decreased the undercoverage. Let $D_{(U_n)}$ be the $100 q_n$th sample quantile of the D_i.

The nonparametric prediction region is due to Olive (2013). For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). A future observation (random vector) $\boldsymbol{x}_f$ is in the region (8) if $D_{\boldsymbol{x}_f} \leq D^2_{U_n})$. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, the nonparametric prediction region (8) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (8) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance.) Refer to the above paragraph for $D_{(U_n)}$. Let $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$ if $D^2_{1-\delta}$ is a continuity point of the cdf $F_{D^2}(y)$ and $D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \xrightarrow{D} D^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}} (\boldsymbol{x}-\boldsymbol{\mu})$.

Definition 6. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with mean $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\} \tag{8}$$

if $D^2_{1-\delta}$ is a continuity point of the cdf $F_{D^2}(y)$.

Highest density regions are usually hard to estimate for p not much larger than four, but many elliptically contoured distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (8). For more about highest density regions, see Olive (2017, pp. 148-155). If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. A PI is a prediction region where $p = 1$.

Definition 7. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then the large sample $100(1-\delta)\%$ *classical prediction region* for multivariate normal data is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{9}$$

The nonparametric prediction region (8) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1-\delta$ if D has a pdf, although prediction regions with smaller volume may exist. The nonparametric prediction region (8) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1-\delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$. Olive (2013) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are distributions that will have severe undercoverage.

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq h^2\}$. Figure 2 was made with the *Arc* software of Cook and Weisberg (1999). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 2a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

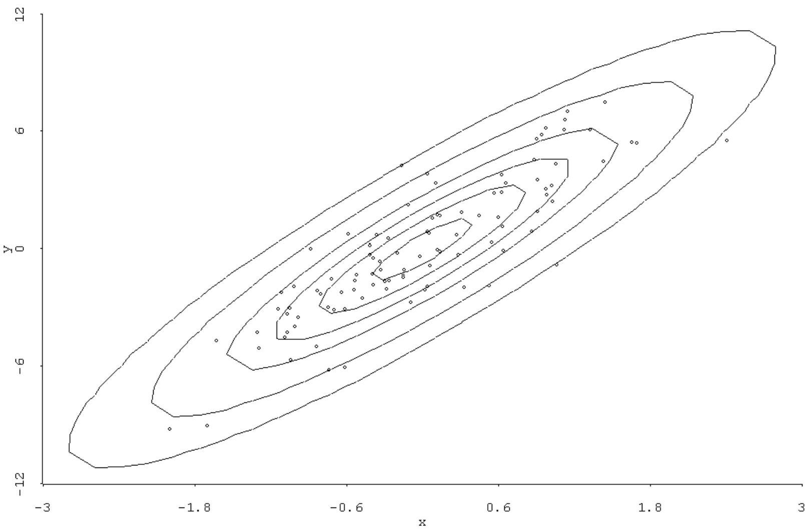
Note that the ellipsoids are narrow with high positive correlation. In Figure 2b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

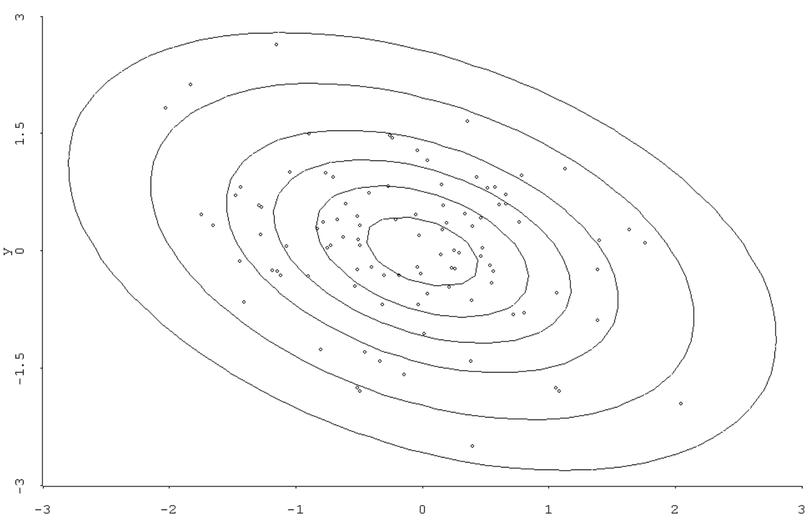
Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

1.2 Bootstrap Confidence Regions

For bootstrap confidence regions, if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_n^* - T_n) \xrightarrow{D} \boldsymbol{u}$, then the percentiles of $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}_0)$ can be estimated with the sample percentiles of $n(T_n^* - T_n)^T \boldsymbol{C}_n^{-1}(T_n^* - T_n)$. Let $\boldsymbol{\theta}$ be a $g \times 1$ vector. For the correction factor below, and a nominal 95% confidence region, instead of using $D^2_{(\lceil 0.95B \rceil)}$ as the cutoff where $D^2_{(c)}$



a)



b)

Figure 2: Highest Density Regions for 2 MVN Distributions

is the cth order statistic of the D_i^2, the $100q_B$th sample quantile of the D_i^2, denoted by $D^2_{(U_B)}$, is used where $0.95B \leq U_B \leq 0.975B$ and $U_B \to 0.95B$ as B increases. Let $q_B = \min(1-\delta+0.05, 1-\delta+g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1-\delta/2, 1-\delta+10\delta g/B), \quad \text{otherwise}. \tag{10}$$

If $1-\delta < 0.999$ and $q_B < 1-\delta+0.001$, set $q_B = 1-\delta$. This correction factor helps reduce undercoverage when $B \geq 50b$.

The following three confidence regions can be used for inference. The Olive (2017ab, 2018) prediction region method confidence region applies prediction region (8) to the bootstrap sample. Let the bootstrap sample be $T_1^*, ..., T_B^*$. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample.

Definition 8. The large sample $100(1-\delta)\%$ prediction region method confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{11}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Olive (2017ab, 2018) also gave the modified Bickel and Ren (2001) confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$.

Definition 9. The large sample $100(1-\delta)\%$ modified Bickel and Ren confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_{BT})}\} \tag{12}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_{BT})}$.

The hybrid confidence region is due to Pelawa Watagoda and Olive (2021).

Definition 10. Shift region (11) to have center T_n, or equivalently, change the cutoff of region (12) to $D^2_{(U_B)}$ to get the large sample $100(1-\delta)\%$ hybrid confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{13}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Rajapaksha and Olive (2022) gave the following two confidence regions. The names of these confidence regions were chosen since they are similar to the Bickel and Ren and prediction region method confidence regions.

Definition 11. The large sample $100(1-\delta)\%$ BR confidence region is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - T_n)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{C}_n/n) \leq D^2_{(U_{BT})}\} \tag{14}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = n(T_i^* - T_n)^T \boldsymbol{C}_n^{-1}(T_i^* - T_n)$ where q_B is found from (3) with $\boldsymbol{z}_i = T_i^*$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_{BT})}$.

Definition 12. The large sample $100(1-\delta)\%$ PR confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - \overline{T}^*)^T \boldsymbol{C}_n^{-1}(\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{C}_n/n) \leq D^2_{(U_B)}\} \tag{15}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = n(T_i^* - \overline{T}^*)^T \boldsymbol{C}_n^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(\overline{T}^* - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

The standard bootstrap confidence region is similar to what would be obtained if the classical prediction region (9) for multivariate normal data was applied to the bootstrap sample.

Definition 13. The large sample $100(1-\delta)\%$ standard bootstrap confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{16}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$.

Much of the theory for the above confidence and prediction region appears in Olive (2023d, ch. 4, 5). If $n\boldsymbol{C}_n^{-1} = [\boldsymbol{S}_T^*]^{-1}$, then (14) and (15) are the modified Bickel and Ren (2001) and Olive (2017ab, 2018) prediction region method large sample $100(1-\delta)\%$ confidence regions for $\boldsymbol{\theta}$. Under regularity conditions, Bickel and Ren (2001) and Olive (2017b, 2018) proved that (11) and (12) are large sample confidence regions. Pelawa Watagoda and Olive (2021) gave simpler proofs. Pelawa Watagoda and Olive (2021) showed that under reasonable regularity conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Usually i) and ii) are proven using large sample theory. If $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$ with $\boldsymbol{\Sigma_u}$ nonsingular, then Pelawa Watagoda and Olive (2021) showed $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{P} \boldsymbol{0}$. Thus iii) and iv) hold if i) and ii) hold. If T_n is the sample mean or sample coordinatewise median, then see Bickel and Freedman (1981) and Rupasinghe Arachchige Don and Olive (2019). Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{C}_n/n) = \sqrt{n}(T_i^* - \overline{T}^*)^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{C}_n/n) = \sqrt{n}(T_n - \boldsymbol{\theta})^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{C}_n/n) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T \boldsymbol{C}_n^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{C}_n/n) = \sqrt{n}(T_i^* - T_n)^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_i^* - T_n),$$

are well behaved. If $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, and (14) and (15) are large sample confidence regions. If $\boldsymbol{C}_n^{-1}$ is "not too ill conditioned," then $D_j^2 \approx \boldsymbol{u}^T \boldsymbol{C}_n^{-1} \boldsymbol{u}$ for large n, and the confidence regions (14) and (15) will have coverage near $1 - \delta$.

1.3 Visualizing the Nonparametric Prediction Region

Olive (2013) showed how to visualize the nonparametric prediction region (8) with the Rousseeuw and Van Driessen (1999) DD plot of classical distances versus robust distances on the vertical axis. See Section 3 where the exact same method will be used to visualize the bootstrap confidence region (11).

2 The Bootstrap

This section illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 14. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 1. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 14. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^n \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \ \square$$

Example 2. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n}\sum_{i=1}^n I(W_i \le y)$$

where the indicator $I(W_i \le y) = 1$ if $W_i \le y$ and $I(W_i \le y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 3. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

3 Visualizing Some Bootstrap Confidence Regions

As mentioned in Section 1.3, the DD plot will be used to visualize some bootstrap confidence regions. If a good robust estimator is used, Olive (2002) showed that the plotted points in a DD plot cluster about the identity line with zero intercept and unit slope if the $\boldsymbol{x}_i$ are iid from a multivariate normal distribution with nonsingular covariance matrix, while the plotted points cluster about some other line through the origin if the $\boldsymbol{x}_i$ are iid from a large family of nonnormal elliptically contoured distributions. For the robust estimator of multivariate location and dispersion, we recommend the RFCH or RMVE estimator. See Olive (2017b, 2023c), Olive and Hawkins (2010), and Zhang, Olive, and Ye (2012). These two estimators $(T_n, \boldsymbol{C}_n)$ are such that $\boldsymbol{C}_n$ is a $\sqrt{n}$ consistent estimator of $a\text{Cov}(\boldsymbol{x})$ for a large class of elliptically contoured distributions where the constant $a > 0$ depends on the elliptically contoured distribution and the estimator RFCH or RMVN, and $a = 1$ for the multivariate normal distribution with nonsingular covariance matrix. We will use the RMVN estimator in the software.

Example 4. We generated $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, \boldsymbol{I})$ for $i = 1, ..., 250$. The coordinatewise median was the statistic T_n. The nonparametric bootstrap was used with $B = 1000$. The DD plot of the bootstrap sample is shown in Figure 3. The plotted points cluster about the identity line. The vertical line MD = 2.9098 is the cutoff for the prediction region method confidence region (11). The long horizonal line RD = 3.0995 is the cutoff using the robust estimator. When $\sqrt{n}(T_n - \boldsymbol{\theta}) \stackrel{D}{\to} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$, then under mild regularity conditions, $\sqrt{n}(T_n - \overline{T}^*_n) \stackrel{P}{\to} \boldsymbol{0}$. Thus the short horizontal line is RD = 2.8074. Thus MD = 2.8074 is approximately the cutoff that would be used by the standard bootstrap confidence region (mentally drop a vertical line from where the short horizontal line ends at the identity line). Variability in DD plots increases as RD increases. The *R* commands for making the plot are shown below.

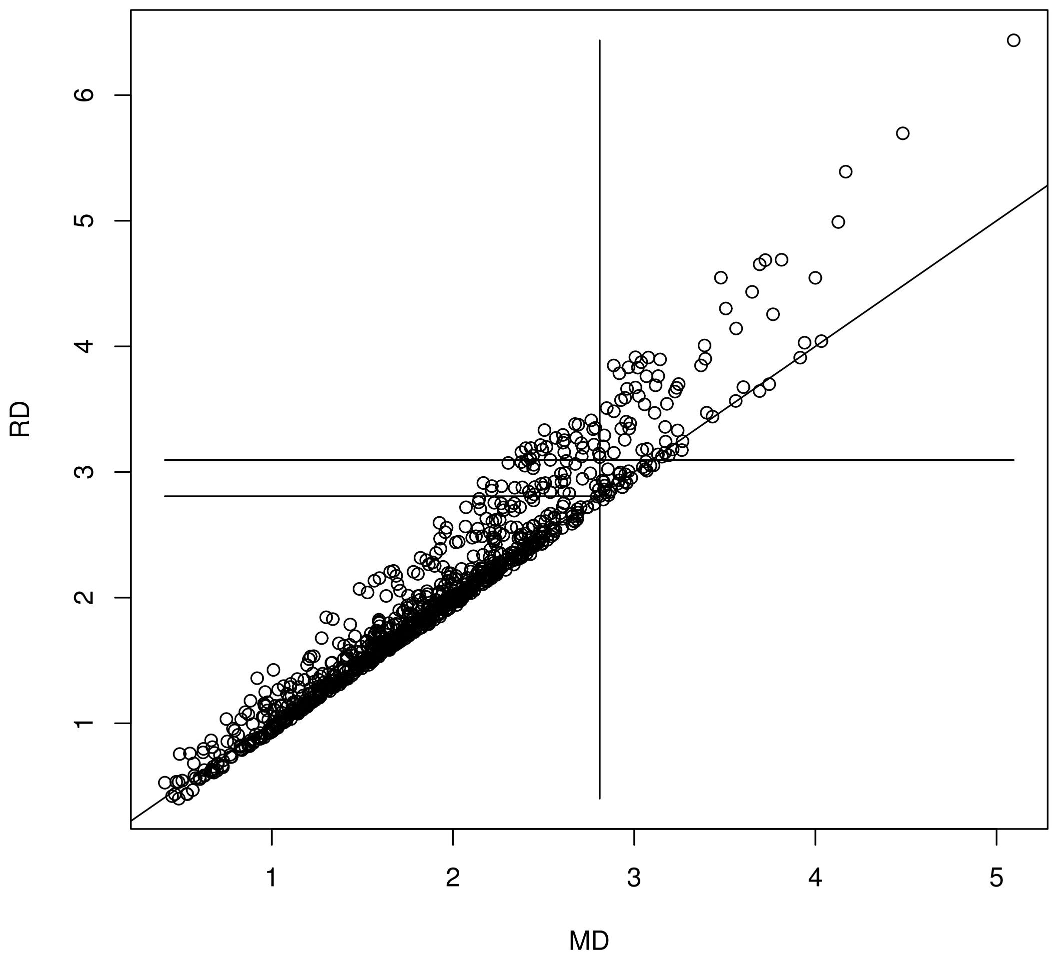


Figure 3: Visualizing the confidence region with a DD plot

```
source("http://parker.ad.siu.edu/Olive/mpack.txt")
```

```
x <-matrix(rnorm(1000),nrow=250,ncol=4)
out<-rhotboot(x)
ddplot4(out$mus)
$cuplim
   90.4%
2.809824
$ruplim
   90.4%
3.095542
$mvnlim
[1] 2.807479
```

Some *R* functions for bootstrapping several statistics are shown below.

```
source("http://parker.ad.siu.edu/Olive/slpack.txt")
args(bicboot)  #bootstrap min BIC model forward selection regression
#function (x, y, B = 1000)
args(FDAboot) #Bootstraps FDA betahat = first eigenvector.
#function (x, group, B = 1000)
args(fselboot2) #bootstrap min Cp model forward selection regression
#function (x, y, B = 1000, c = 0.01, aug = F)
args(lassoboot2) #bootstrap lasso or ridge regression for MLR
#function (x, y, B = 1000, regtype = 1, c = 0.01, aug = F)
args(LPHboot)  #bootstraps the Cox regression lasso, takes a few minutes
#function (x, time, status, B = 1000)
args(LRboot)  #bootstrap logistic regression  full model
#function (x, y, mv = c(1, 1), B = 1000, bin = T)
args(pcaboot) #Bootstraps PCA. Likely only accurate for positive eigenvalues
#function (x, corr = T, rob = F, B = 1000)
args(PHboot)  #bootstraps the Cox PH regression full model
#with the nonparametric bootstrap
#function (x, time, status, B = 1000)
args(PRboot)  #bootstraps the Poisson regression full model
#function (x, y, B = 1000)
args(regboot) #residual bootstrap for MLR
function (x, y, B = 1000)
args(rowboot) #nonparametric bootstrap for MLR
#function (x, y, B = 1000)

source("http://parker.ad.siu.edu/Olive/mpack.txt")
args(corboot)  #rowwise nonparametric bootstrap of the correlation matrix
#function (x, B = 1000) #stacks entries above the diagonal into a vector beta
args(rhotboot) #Bootstraps  RMVN center (med=F) or coordinatewise median.
#function (x, B = 1000, med = T)
```

```
source("http://parker.ad.siu.edu/Olive/tspack.txt")
args(arboot) Bootstraps AR(p) model selection using the parametric bootstrap
#function (Y, B = 100, pmax = 10, c = 0.01)
args(arboot2) #Bootstraps AR(p) model selection using the residual bootstrap.
#function (Y, B = 100, pmax = 10, c = 0.01)
args(maboot) #Bootstraps MA(q) model selection using the parametric bootstrap.
#function(Y,B=100,qmax=10,c=0.05)
args(maboot2) #Bootstraps MA(q) model selection using the residual bootstrap.
#function(Y,B=100,qmax=10,c=0.05)
```

4 CONCLUSIONS

The bootstrap is due to Efron (1979). Also see Efron (1982) and Olive (2017ab, 2023abcd). Rathnayake and Olive (2021) show how to bootstrap many variable selection estimators. Haile and Olive (2023) show how to bootstrap AR, MA, and ARMA time series model selection estimators. See Zhang and Olive (2023) for prediction regions if n/p is small.

The `rpack` function ddplot4 applied to the bootstrap sample can be used to visualize the bootstrap prediction region method confidence region.

References

Bickel, P.J., and Freedman, D.A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196-1217.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Chew, V. (1966), "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association,* 61, 605-617.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Efron, B. (1979), "Bootstrap Methods, Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Haile, M.G., and Olive, D.J. (2023), "Bootstrapping ARMA Time Series Models after Model Selection" is at (http://parker.ad.siu.edu/Olive/pptsboot.pdf).

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

Olive, D.J. (2023a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J. (2023b), *Theory for Linear Models*, online course notes, (http://parker.ad.siu.edu/Olive/linmodbk.htm).

Olive, D.J. (2023c), *Robust Statistics*, online course notes, (http://parker.ad.siu.edu/Olive/robbook.htm).

Olive (2023d) *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," unpublished manuscript, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers,* 62, 681-700. See (http://parker.ad.siu.edu/Olive/ppboottest.pdf).

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2022), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, to appear.

Rathnayake, R.C., and Olive, D.J. (2021), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, to appear.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation with Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Zhang, L., and Olive, D.J. (2023), "A Data Splitting Prediction Region," is at (http://parker.ad.siu.edu/Olive/ppdsplitpred.pdf).