Chapter 4
Prediction Regions and Bootstrap Confidence Regions

In chapter two, it was shown that applying certain prediction intervals to the bootstrap sample results in confidence intervals. In this chapter, it will be shown that applying the nonparametric prediction region to the bootstrap sample results in a confidence region. Prediction intervals are a special case of prediction regions when $p = 1$ so the $p \times 1$ random vector is a random variable.

4.1 Prediction Regions

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 $\boldsymbol{x}_f$.

Definition 4.1. 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 $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ is from a distribution with a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. The following definition makes sense when the highest density region is unique. Section 2.4 discussed the highest density region for a random variable where $p = 1$. Then nonzero flat spots in the pdf can cause the region to have higher than nominal coverage. For example, the highest density region of a uniform(θ_1, θ_2) random variable is not unique. See Figure 2.1 where the area under the pdf from 0 to 1 gives the 36.8% highest density region. Figure 3.1 shows the highest density regions for two bivariate normal distributions.

Definition 4.2. When unique, the $100(1-\delta)\%$ *highest density region* $R(f_{1-\delta}) = \{\boldsymbol{z} : f(\boldsymbol{z}) \geq f_\delta\}$ where f_δ is the largest constant such that $P[\boldsymbol{x} \in R(f_{1-\delta})] \geq 1-\delta$ and $f(\boldsymbol{z})$ is the probability density function (pdf) of $\boldsymbol{x}$.

Highest density regions are usually hard to estimate for p not much larger than four, but for elliptically contoured distributions with continuous decreasing g, the highest density region is the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \leq u_{1-\delta}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq u_{1-\delta}\} \tag{4.1}$$

where $P(U \leq u_{1-\delta}) = 1-\delta$, and U is given by (3.9). If $HDR_Y(1-\delta)$ is the $100(1-\delta)\%$ highest density region for a random variable Y, and $X \sim U(0,\theta) \perp\!\!\!\perp Y$ (meaning X is independent of Y), then the $100(1-\delta)\%$ highest density region for (X_f, Y_f) is

$$\{(x,y) : x \in (0,\theta), y \in HDR_Y(1-\delta)\}.$$

There is a moderate amount of literature for prediction regions that may perform well for small p. Let $\hat{f}_{(1)}, ..., \hat{f}_{(n)}$ be the order statistics of $\hat{f}(\boldsymbol{x}_1), ..., \hat{f}(\boldsymbol{x}_n)$. Hyndman (1996) used the estimated highest density region

$$\hat{R}(f_{1-\delta}) = \{\boldsymbol{z} : d\hat{f}(\boldsymbol{z}) \geq d\hat{f}_{(h)}\} \tag{4.2}$$

where $d > 0$ can be any constant, $h = max(1, \lfloor n\delta \rfloor)$, and $\lfloor x \rfloor$ is the integer part of x. Here $\hat{f}$ is a kernel density estimator. See Remark 4.3, and see Lei et al. (2013) for references.

Let $D^2_{(c)}$ be the cth order statistic of $D^2_1, ..., D^2_n$, and consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{4.3}$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. Olive (2013a) showed that (4.3) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. See Theorem 11.29. Let $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. 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 cumulative distribution function (cdf) of D.

A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.3) 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*. Increasing c will improve the coverage for moderate samples. Empirically for many distributions, for $n \approx 20p$, the prediction

region (4.3) applied to iid data using $k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{4.4}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i where

$$D_i^2 = D_{\boldsymbol{x}_i}^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}).$$

Definition 4.3. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(U_n)}^2\}, \tag{4.5}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi_{p,1-\delta}^2\}. \tag{4.6}$$

Remark 4.1. The nonparametric prediction region (4.5) 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 nonparametric prediction region is asymptotically optimal on a large class of elliptically contoured distributions in that the prediction region's volume converges in probability to the volume of the highest density region (4.1). The asymptotic coverage is $1-\delta$ if the $100(1-\delta)$th percentile $D_{1-\delta}$ of D is a continuity point of the distribution of D, although prediction regions with smaller volume may exist. If $D_{1-\delta}$ is not a continuity point of the distribution of D, then the asymptotic coverage tends to be $\geq 1-\delta$ since a sample percentile with cutoff q_n that decreases to $1-\delta$ is used and a closed region is used. Often D has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point a, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to a.) The nonparametric prediction region (4.5) 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 $\boldsymbol{x}_f$.

Remark 4.2. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. The ratio of the volumes of regions (4.5) and (4.6) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}} \right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension p increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$ for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (4.7) usually results in prediction regions with volume and coverage that is too large.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D^2_i(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m$. The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151).

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \quad (4.7)$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \quad (4.8)$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.7) if $D_{\boldsymbol{x}_f} \leq h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (4.7) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i where q_n is defined near (4.4). For example, use $U_n = c = \lceil nq_n \rceil$.

Remark 4.3. There may not yet be any practical competing prediction regions that do not have the form of (4.7) if p is much larger than two and the distribution of the $\boldsymbol{x}_i$ is unknown. The prediction region of Section 4.3 also has this form. Remark 4.1 suggests that the nonparametric prediction region (4.5) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Of course for any n there are error distributions that will have severe undercoverage. Prediction regions that estimate the pdf $f(\boldsymbol{z})$ with a kernel density estimator quickly become impractical as p increases since large sample sizes are needed for good estimates. See Silverman (1986, p. 129).

For example, the Hyndman nominal 95% prediction region (4.2) was computed for iid $N_p(\boldsymbol{0}, \boldsymbol{I})$ data with 1000 runs. Let the coverage be the observed proportion of prediction regions that contained the future value $\boldsymbol{x}_f$. For $p = 1$, the coverage was 0.933 for $n = 40$. For $p = 2$, the coverage was 0.911 for $n = 50$, and 0.930 for $n = 150$. For $p = 4$, the coverage was 0.920 for $n = 250$. For $p = 5$ the coverage was 0.866 for $n = 200$ and 0.934 for $n = 2000$. For $p = 8$, the coverage was 0.735 for $n = 125$. For the multivariate lognormal distribution with $n = 20p$, the Olive (2013a) large sample nonparametric 95% prediction region (4.5) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200,$ and 500. Some R code is below.

```
nruns=1000 #p = 1
count<-0
for(i in 1:nruns){
x <- rnorm(40)
xff <- rnorm(1)
count <- count + hdr2(x,xf=xff)$inr}
count #933/1000

count<-0 #p = 5
for(i in 1:nruns){
x <- matrix(rnorm(1000),ncol=5,nrow=200)
xff <- as.vector(rnorm(5))
count <- count + hdr2(x,xf=xff)$inr}
count #886/1000

#lognormal, p = 100
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000
```

Olive (2013a) used three prediction regions (4.7) that can be displayed with the DD plot. The nonparametric prediction region (4.5) uses the classical

estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(U_n)}$. The other two prediction regions are defined below.

Definition 4.4. The *semiparametric prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The *parametric MVN prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,\delta}) = \delta$ if $W \sim \chi^2_p$.

All three prediction regions are asymptotically optimal for MVN distributions with nonsingular $\boldsymbol{\Sigma}$. The first two prediction regions are asymptotically optimal for a large class of $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distributions given by Assumption (E1) used in Theorem 3.20, provided g is continuous and decreasing. For distributions with nonsingular covariance matrix $c_X \boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region, but regions with smaller volume may exist.

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100 q_n\%$ of the n cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$ and $q_n \to 1-\delta$ as $n \to \infty$. If $q_n \equiv 1-\delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (4.7) is a large sample prediction region, but taking q_n given by (4.4) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for small n the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (4.4) helped reduce undercoverage for small n due to the unknown variability of $(T, \boldsymbol{C})$.

Example 4.1. An artificial data set consisting of 100 iid cases from a

$$N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.49 & 1.4\\ 1.4 & 1.49\end{pmatrix}\right)$$

distribution and 40 iid cases from a bivariate normal distribution with mean $(0, -3)^T$ and covariance $\boldsymbol{I}_2$. Figure 4.1 shows the classical ellipsoid (with $MD \leq \sqrt{\chi^2_{2,0.95}}$) that uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The symbol "1" denotes the data while the symbol "2" is on the border of the covering ellipse. There is an R package that makes an ellipse. Notice that the classical parametric ellipsoid covers almost all of the data. Figure 4.2 displays the robust ellipsoid (using $RD \leq \sqrt{\chi^2_{2,0.95}}$) which contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 4.5 recreates similar figures with the classical and RMVN estimators using $q_n = 0.95$.

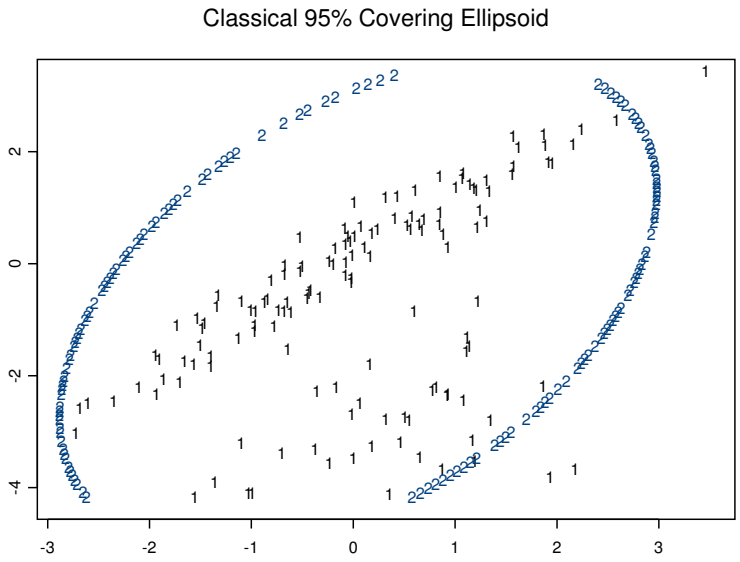

Fig. 4.1 Artificial Bivariate Data

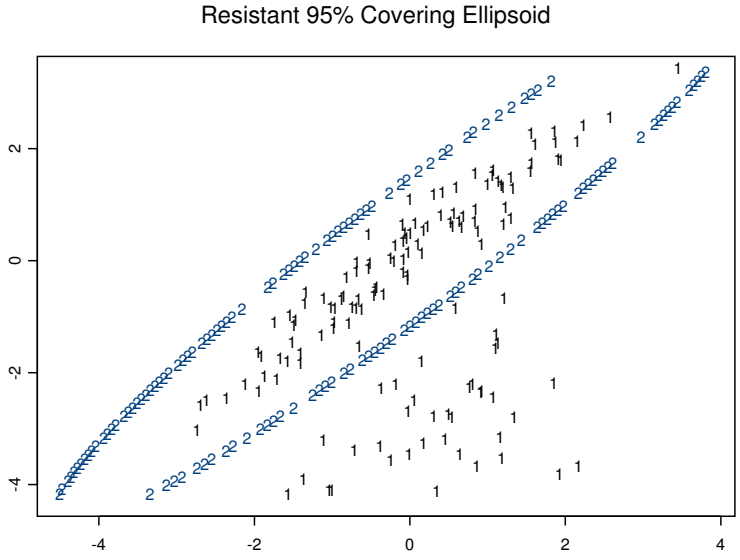

Fig. 4.2 Artificial Data

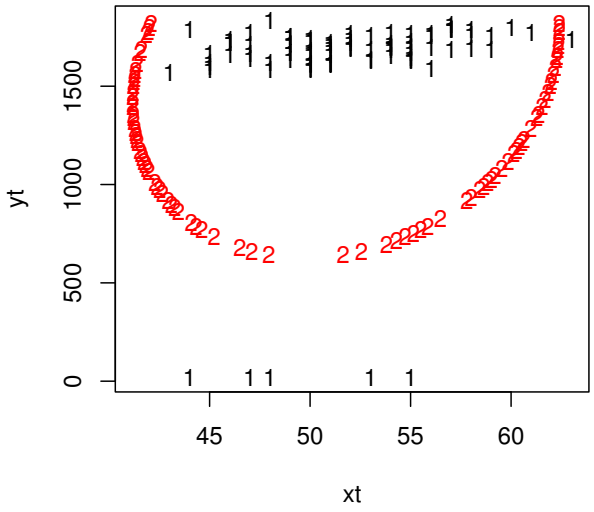

Fig. 4.3 Ellipsoid is Inflated by Outliers

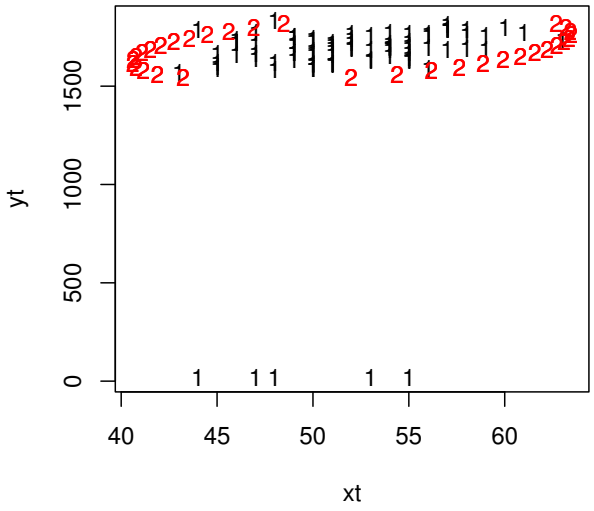

Fig. 4.4 Ellipsoid Ignores Outliers

Example 4.2. Buxton (1920) gave various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth*, and *cephalic index*. Five *heights* were recorded to be about 19mm (and the actual heights for these cases were recorded as the head lengths) and are massive outliers. First *height* and *nasal height* were used with $q_n = 0.95$. Figure 4.3 shows that the classical parametric prediction region (using $MD \leq \sqrt{\chi^2_{2,.95}}$) is quite large but does not include any of the outliers. Figure 4.4 shows that the parametric MVN prediction region (using $RD \leq \sqrt{\chi^2_{2,.95}}$) is not inflated by the outliers.

Next all 87 cases and 5 predictors were used. Figure 4.5 shows the RMVN DD plot with the identity line added as a visual aid. Points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The nonparametric and semiparametric 90% regions blow up unless the outlier proportion is small.

Figure 4.5 can be made with the following *R* commands, assuming source commands for `pack` and `robdata` have been performed. See the Preface or Section 11.2. Right click `Stop` to get the cursor.

```
x <- cbind(buxy,buxx)
ddplot4(x)  #right click Stop
```

Figure 4.6 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff again at 3.33, slightly below the semiparametric region cutoff of 3.44. Cases to the left of the vertical line $MD = 3.33$ (not shown since you can mentally drop down a vertical line where the horizontal line ends at the identity line), correspond to a (modified) classical prediction region.

Figure 4.6 can be made with the following *R* commands. Right click `Stop` to get the cursor and the output following the two commands.

```
zx <- x[-c(61:65),]
ddplot4(zx) #right click Stop
$cuplim
     95%
3.086005
$ruplim
     95%
3.438821
$mvnlim
[1] 3.327236
```

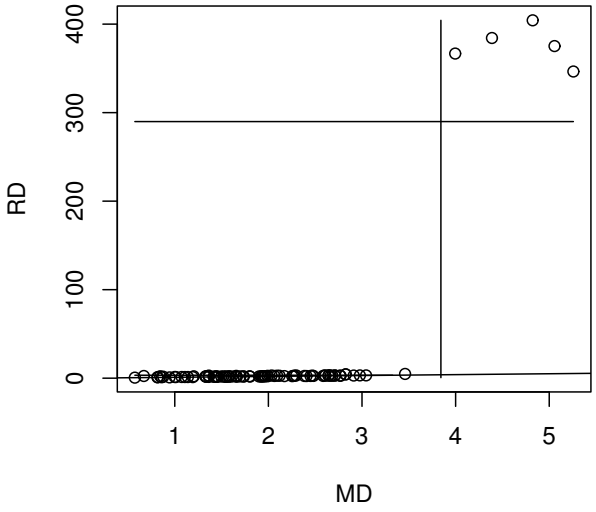

Fig. 4.5 Prediction Regions for Buxton Data

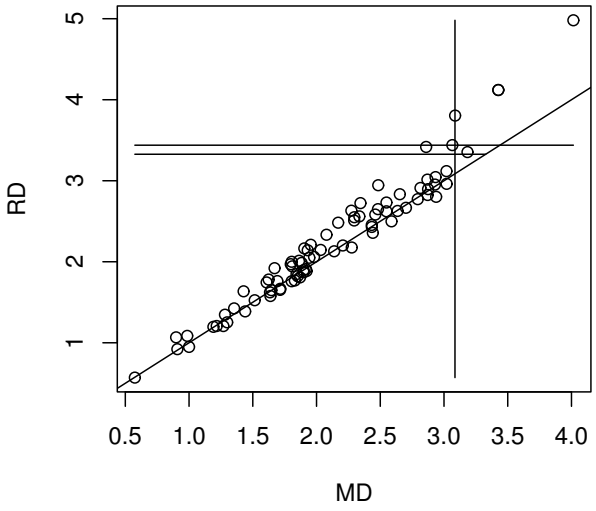

Fig. 4.6 Prediction Regions for Buxton Data without Outliers

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ (MVN), $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\delta = 0.1$.

Often the coverage for the semiparametric region was better than that of the nonparametric region for n near $10p$. The nonparametric covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq D^2_{(n)}(\overline{\boldsymbol{x}}, \boldsymbol{S})\}$ uses all of the data, but for small n, data is sparse, and the covering region overfits and hence the volume is too small. The nonparametric prediction region is a hyperellipsoid that is concentric with the covering region (that replaces $D^2_{(U_n)}$ with $D^2_{(n)}$). The semiparametric region is based on the RMVN half set of data. This region is not a good estimator of the population 50% covering region for small n. Hence when it is blown up to cover 95% of the training data, the region is quite large, so it is likely that a future $\boldsymbol{x}_f$ is in the region.

For large n, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ if Assumption (E1) holds. The parametric MVN region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage (and volume ratio near 0.5 for MVN data).

Table 4.1 Coverages for 90% Prediction Regions

$\boldsymbol{w}$ dist	n	p	ncov	scov	mcov	voln	volm
MVN	600	30	0.906	0.919	0.902	0.503	0.512
MVN	1500	30	0.899	0.899	0.900	1.014	1.027
LN	1000	10	0.903	0.906	0.567	0.659	0+
MVT(1)	1000	10	0.914	0.914	0.541	22634.3	0+

Simulations and Table 4.1 suggest that for MVN data, the coverages (ncov, scov, and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than 5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN

data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

4.2 Bootstrap Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.1 to a bootstrap sample results in a confidence region. The volume of a confidence region $\rightarrow 0$ as $n \rightarrow 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1 - \delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal. If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$.

Definition 4.5. 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 $n \rightarrow \infty$.

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 4.6. 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 4.3. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 4.6. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

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

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

Example 2.8 was similar to Example 4.3, and see Example 2.9 for the empirical cdf of a random variable. 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 $j = 1, ..., B$ where the n is often suppressed.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B} \sum_{i=1}^{B} T_i^* \ \text{ and } \ \boldsymbol{S}_T^* = \frac{1}{B-1} \sum_{i=1}^{B} (T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{4.9}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$.

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1 - \delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $k_B = \lceil B(1-\delta) \rceil$.

Remark 4.4. A useful fact for the F and chi-square distributions is $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq \chi^2_{g,1-\delta}) = 1 - \delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1 - \delta$ if $X \sim F_{g,d_n}$.

Definition 4.7. a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{4.10}$$

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 $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A / n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{4.11}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A/n]^{-1}(T_i^* - T_n) = n(T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A]^{-1}(T_i^* - T_n)$.

Confidence region (4.10) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies the nonparametric prediction region (4.5) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2019). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{4.12}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (4.12) as a correction factor for finite $B \geq 50p$.

Definition 4.8. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{4.13}$$

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 $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{4.14}$$

where the cutoff $D^2_{(U_B,T)}$ 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 $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. c) Shift region (4.13) to have center T_n, or equivalently, change the cutoff of region (4.14) to $D^2_{(U_B)}$ to get the hybrid large sample $100(1 - \delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{4.15}$$

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

Hyperellipsoids (4.13) and (4.15) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (4.13) and (4.14) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g . \tag{4.16}$$

The volume of confidence region (4.14) tends to be greater than that of (4.13) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

Next we review the Section 2.5 confidence intervals corresponding to the three confidence regions if $g = 1$. Suppose the parameter of interest is θ, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $g = 1$. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI corresponding to (4.14) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (4.15). Note that the three CIs corresponding to (4.13)–(4.15) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The shorth(c) CI (2.13) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 2.5 for some theory for bootstrap CIs.

Remark 4.5. Suppose the $p \times 1$ vector $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\theta} = \boldsymbol{A}\hat{\boldsymbol{\beta}}$ is $g \times 1$. We will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}$ is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.4) to compensate for finite n. The bootstrap confidence regions (4.13)–(4.15) and the shorth CI use the correction factors (4.12) and (2.14) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

4.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (4.14) is a large sample confidence region by Bickel and Ren (2001), (4.13) and (4.15) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular. Olive (2017b: § 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, but the above Pelawa Watagoda and Olive (2019) proof is simpler. Remark 2.5 gave theory for bootstrap confidence intervals.

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.17}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{4.18}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (4.10) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$.

Remark 4.6. Under reasonable 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) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Then

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

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_n - \boldsymbol{\theta}),$$

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

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

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T\boldsymbol{\Sigma}_A^{-1}\boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T(n\boldsymbol{S}_T^*)^{-1}\boldsymbol{u}$ for large n, and the confidence regions (4.13), (4.14), and (4.15) will have coverage near $1 - \delta$. The regularity conditions for (4.13)–(4.15) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following Pelawa Watagoda and Olive (2019) theorem is very useful. Let $D^2_{(U_B)}$ be the cutoff for the nonparametric prediction region (4.5) computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence n is replaced by B. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$. In the following theorem, note that we an replace $\sqrt{n}$ by n^δ where $0 < \delta \leq 1$.

Theorem 4.1: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Assume $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Then the large sample

$100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1-\delta_B \to 1-\delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1-\delta_B$ which is eventually bounded below by $1-\delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 11.26 and 11.27, and see Example 11.12.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma u}}{B}\right)$$

by Theorem 11.29. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

Remark 4.7. Theorem 4.1 is useful for explaining why a correction factor is needed. R_c contains $\overline{T}$ with probability $1-\delta_B$ and there is a hyperellipsoid $R_{\overline{T}}$ about $\overline{T}$ that contains $\boldsymbol{\theta}$ with high probability where the volume of $R_{\overline{T}}$ goes to 0 as $B \to \infty$. For finite $B \approx 50g$, the volume of the hyperellipsoid $R_{\overline{T}}$ is small compared to that of R_c but is not zero. As B increases, covering $\overline{T}$ also covers most of $R_{\overline{T}}$, and the confidence region coverage gets near the nominal level $1-\delta$. When B is near $50g$, covering $\overline{T}$ does not necessarily cover most of $R_{\overline{T}}$, and hence there may be undercoverage for $\boldsymbol{\theta}$ if $U_B = \lceil B(1-\delta)\rceil$. Using the correction factor (4.12) increases the coverage of $\overline{T}$, $R_{\overline{T}}$, and $\boldsymbol{\theta}$ when B is near $50g$.

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (4.13) applies the prediction region to the bootstrap. Then the hybrid region (4.15) centers that region at T_n. Hence (4.15) is a confidence region by the geometric argument, and (4.13) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence

the coverage and volume of (4.14) tend to be at least as large as the coverage and volume of (4.15).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1} (\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1} (T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 3.1 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 3.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 4.1, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 3.1. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 4.8. Even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the prediction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 4.1 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jg$ where $J = 20$ or 50. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have

better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

4.4 Data Splitting

Data splitting can be used to get prediction regions using estimators such as $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ or $(T, \boldsymbol{C}) = (\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into two sets H and V where H has $n_H \geq n/2$ of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $(T_H, \boldsymbol{C}_H)$ is computed using the data set H. Then the squared validation distances $D_j^2 = D_{\boldsymbol{x}_{i_j}}^2(T_H, \boldsymbol{C}_H) = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Let $D_{(U_V)}^2$ be the U_Vth order statistic of the D_j^2 where

$$U_V = \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil). \tag{4.19}$$

Definition 4.9. The large sample $100(1 - \delta)\%$ data splitting prediction region for $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}. \tag{4.20}$$

To show that (4.20) is a prediction region, suppose the $\boldsymbol{x}_i$ are iid for $i = 1, ..., n, n + 1$ where $\boldsymbol{x}_f = \boldsymbol{x}_{n+1}$. Compute $(T_H, \boldsymbol{C}_H)$ from the cases in H. Consider the squared validation distances D_k^2 for $k = 1, ..., n_V$ and the squared validation distance $D_{n_V+1}^2$ for case $\boldsymbol{x}_f$. Since these $n_V + 1$ cases are iid, the probability that D_t^2 has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $D_{(j)}^2$ be the ordered squared validation distances using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown squared validation distance $D_{n_V+1}^2$. Then $D_{(i)}^2$ has rank i if $D_{(i)}^2 < D_{n_V+1}^2$ but rank $i+1$ if $D_{(i)}^2 > D_{n_V+1}^2$. Thus $D_{(U_V)}^2$ has rank $U_V + 1$ if $D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2$ and

$$P(\boldsymbol{x}_f \in \{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}) = P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) \geq U_V/(1 + n_V) \rightarrow$$

$1 - \delta$ as $n_V \rightarrow \infty$. If there are no tied ranks, then

$$P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) = P(D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2) = P(\text{rank of } \mathrm{D}_{\boldsymbol{x}_\mathrm{f}}^2 \leq \mathrm{U_V}) = \mathrm{U_V}/(1+\mathrm{n_V}).$$

Note that we can get coverage close to $1 - \delta$ for $n_V \geq 20$ for $\delta = 0.05$ even if $(T_H, \boldsymbol{C}_H)$ is a bad estimator. The volume of the prediction region tends to be much larger than that of the highest density region, even if $\boldsymbol{C}_H$ is well conditioned. We likely need $U_V \geq 50$ for $D_{(U_V)}^2$ to approximate the population percentile of $D_j^2 = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$.

As an example, consider using $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Then the prediction region is a hypersphere centered at the coordinatewise median. The prediction region is good if the iid $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$, but if the $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that highest density region is a hyperellipsoid tightly clustered about a vector in the direction of $\mathbf{1}$, then the prediction region (4.20) has huge volume compared to the highest density region.

If $p > n$, prediction region (4.20) can be used as long as $\boldsymbol{C}$ is nonsingular. Then $\boldsymbol{C} = \boldsymbol{I}_p$, $\boldsymbol{C} = diag(S_1^2, ..., S_p^2)$, or

$$\boldsymbol{C} = diag([MAD(x_{11}, ..., x_{n1})]^2, ..., [MAD(x_{1p}, ..., x_{np})]^2)$$

could be used. Regularized covariance matrices or precision matrices could also be used.

If $n \geq 20p$, using $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ might result in a prediction region with smaller volume then using $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ since the robust estimator attempts to estimate a small volume hyperellipsoid. Also, if $D^2_{(U_V)} \approx D^2_{(U_n)}$ in Definition 4.4, then the semiparametric region using all n cases should have good coverage.

4.5 Summary

4) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} =$
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. 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 $n \to \infty$ where $0 < \delta < 1$. 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 $n \to \infty$.

5) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n =$
$\min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i. The large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume.

6) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let q_B and U_B be as in 5) with n replaced by B and p replaced by g. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$ 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 $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or D^2 for a), b), and c).

For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model if n and B are large enough. Let $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_O = \boldsymbol{0}$. If $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{O,i} = \boldsymbol{0}$ for greater than $B\delta$ of the bootstrap samples $i = 1, ..., B$, then fail to reject H_0. (If $\boldsymbol{S}_T^*$ is nonsingular, the $100(1-\delta)\%$ prediction region method confidence region contains $\boldsymbol{0}$.)

7) **Theorem 4.1: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

8) Applying the nonparametric prediction region (4.24) to the iid data $T_1, ..., T_B$ results in the $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{S}_T^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}(T_n, \boldsymbol{S}_T)\}$ where $D^2_{(U_B)}(T_n, \boldsymbol{S}_T)$ is computed from the $(T_i - T_n)^T \boldsymbol{S}_T^{-1} (T_i - T_n)$ provided the $\boldsymbol{S}_T = \boldsymbol{S}_{T_n}$ are "not too ill conditioned." For OLS variable selection, assume there are two or more component clouds. The bootstrap component data clouds have the same asymptotic covariance matrix as the iid component data clouds, which are centered at $\boldsymbol{\theta}$. The jth bootstrap component data cloud is centered at $E(T_{ij}^*)$ and often $E(T_{jn}^*) = T_{jn}$. Confidence region (4.32) is the prediction region (4.24) applied to the bootstrap sample, and (4.32) is slightly larger in volume than (4.24) applied to the iid sample, asymptotically. The hybrid region (4.34) shifts (4.32) to be centered at T_n. Shifting the component clouds slightly and computing (4.24) does not change the axes of the prediction region (4.24) much compared to not shifting the component clouds. Hence by the geometric argument, we expect (4.34) to have coverage at least as high as the nominal, asymptotically, provided the $\boldsymbol{S}_T^*$ are "not too ill conditioned." The Bickel and Ren confidence region (4.33) tends to have higher coverage and volume than (4.34). Since $\overline{T}^*$ tends to be closer to $\boldsymbol{\theta}$ than T_n, (4.32) tends to have good coverage.

9) Suppose m independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\delta_n)$ where $1-\delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta$. A prediction region with $1-\delta_n < 1-\delta$ is liberal and a region with $1-\delta_n > 1-\delta$ is conservative. It is better to be conservative by 3% than liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

4.6 Complements

There are few practical competitors for the prediction regions in Sections 4.1 and 4.3. Parametric regions such as the classical region for multivariate normal data tend to have severe undercoverage because the data rarely follows the parametric distribution. Procedures that use brand name high breakdown multivariate location and dispersion estimators take too long to compute for $p > 2$. An interesting idea is to estimate the pdf of the data, then use the pdf to find small prediction regions. The problem with these regions is that nonparametric pdf estimators do not work well for $p > 4$. See Lei et al. (2013). A useful application of prediction regions is Mykland (2003).

Bickel and Ren (2001) have interesting sufficient conditions for (4.11) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016), Hesterberg (2014), and Rajapaksha and Olive (2021). One of the sufficient conditions for the bootstrap confidence region is that T has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986b, 2000), Fernholtz (1983), Gill (1989),

Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

For bootstrapping robust estimators, see Olive (2017b), Rupasinghe Arachchige Don and Olive D.J. (2019) and Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

4.7 Problems

R Problems Use the command *source("G:/rpack.txt")* **to download the functions** and the command *source("G:/robdata.txt")* **to download the data. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *covmba*, will display the code for the function. Use the `args` command, e.g. *args(covmba)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/robRhw.txt) into *R*.

4.1. Use the *R* source commands and then type *ddplot4(buxx, alpha=0.2)* and put the plot in *Word*. The Buxton data has 5 outliers, $p = 4$, and $n = 87$, so the 80% prediction region uses the $100(1 - \delta + p/n) = 84.6$th percentile. The output shows that the cutoffs are 2.527, 2.734, and 2.583 for the nonparametric, semiparametric, and robust parametric prediction regions. The two horizontal lines that correspond to the robust distances are obscured by the identity line. (Right click *Stop* once on the plot.)

4.2. Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$. Were the three coverages near 90%?

4.3. The function `predsim2` computes the data splitting prediction region. The output gives cvr = observed coverage, up $\approx$ actual coverage, and mnhsq = mean cutoff $D^2_{(U_V)}$. With 5000 runs, expect observed coverage $\in [0.94, 0.96]$ if the actual coverage is close to 0.95.

a) When xtype=3 and dtype=1, $(T, C) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ where $\boldsymbol{x}_i \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$. If $n \geq \max(20p, 200)$ and $n_V = 100$, then $D^2_{(U_V)}$ should estimate the population percentile $\chi^2_{p,0.95}$. Copy and paste the commands for this problem into *R*. Include the output in *Word*.

i) Was the observed coverage near the actual coverage?

ii) Was the mnhsq near 18.3?

b) When xtype = 1, $\boldsymbol{x}_i \sim N_p(\mathbf{0}, diag(1, ..., p))$ and the χ^2 approximation no longer holds. Copy and paste the commands for this problem into *R*. Include the output in *Word*.

i) Was the observed coverage near the actual coverage?

ii) Was the mnhsq a lot larger than 18.3? (If so, then the volume of the prediction region is much larger than that in a).)

c) Copy and paste the commands for this problem into *R*. Include the output in *Word*. Now $p > n$. Were the observed and actual coverages close?