Chapter 4
Prediction Intervals and Prediction Regions

This chapter considers prediction intervals and prediction regions for iid data. In later chapters, prediction intervals for regression and prediction regions for multivariate regression are derived. Inference after variable selection will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. See Chapter 5.

4.1 Prediction Intervals

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.

Definition 4.1. 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.

If Y_f has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as $n \to \infty$. 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.

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.

Consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid with the same distribution as Y. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the iid training data $Y_1, ..., Y_n$. Then the unknown future value Y_f is the test data. Suppose the sample percentiles $[\hat{L}_n, \hat{U}_n]$ of the training data $Y_1, ..., Y_n$ are consistent estimators of the population percentiles $[L, U]$ of the distribution where $P(Y \in [L, U]) = 1 - \delta$. Then $P(Y_f \in [\hat{L}_n, \hat{U}_n] \rightarrow P(Y_f \in [L, U]) = 1 - \delta$ as $n \rightarrow \infty$. Three common choices are a) $P(Y \leq U) = 1 - \delta/2$ and $P(Y \leq L) = \delta/2$, b) $P(Y^2 \leq U^2) = P(|Y| \leq U) = P(-U \leq Y \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) such that $P(Y \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions.

If the cdf F_Y of Y has jumps, then it may not be possible to find L and U such that $P(Y \in [L, U]) = 1 - \delta$, but it is possible to find L and U such that $P(Y \in [L, U]) \geq 1 - \delta$ for $0 < \delta < 1$. For example, if $P(Y = c) = 1$, then $P(Y \in [c, c]) = 1 \geq 1 - \delta$ for $0 < \delta < 1$. For $Y_1, ..., Y_n$ iid BIN$(n = 1, \rho)$, useful PIs are [0,0], [0,1], and [1,1]. Using open intervals would give 0% coverage.

Let $0 < \alpha < 1$, and let Y_α be a number such that $P(Y \leq Y_\alpha) = \alpha$ if Y_α is a continuity point of the cdf $F_Y(y)$. Let $F(y-) = P(Y < y)$. If Y_α is not a continuity point of $F_Y(y)$, let $F(Y_\alpha -) = \alpha_1 \leq \alpha \leq \alpha_2 = F(Y_\alpha)$ where $0 \leq \alpha_1 < \alpha_2 \leq 1$. Suppose $\alpha_1 < \alpha < \alpha_2$. For example, let $\alpha_1 = 0.89 < \alpha = 0.9 < \alpha_2 = 0.92$. Let $\lceil x \rceil$ be the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then $\sum_{i=1}^{n} I(Y_i \leq Y_{(\lceil n\alpha \rceil)}) \geq \lceil n\alpha \rceil$ with equality unless there are ties: at least two $Y_i = Y_{(\lceil n\alpha \rceil)}$. Thus if $Y_{(\lceil n\alpha \rceil)} < Y_\alpha$, not enough $Y_i \leq Y_{(\lceil n\alpha \rceil)}$, while if $Y_{(\lceil n\alpha \rceil)} > Y_\alpha$, too many $Y_i \leq Y_{(\lceil n\alpha \rceil)}$. Hence $P(Y_{(\lceil n\alpha \rceil)} = Y_\alpha) \rightarrow 1$, $P(Y_f < Y_{(\lceil n\alpha \rceil)}) \rightarrow \alpha_1 < \alpha$, and $P(Y_f \leq Y_{(\lceil n\alpha \rceil)}) \rightarrow \alpha_2 > \alpha$ as $n \rightarrow \infty$. Similarly, if $\alpha_2 = \alpha$, then $P(Y_{(\lceil n\alpha \rceil)} \geq Y_\alpha) \rightarrow 1$ as $n \rightarrow \infty$. If $\alpha_1 = \alpha$ and $F_Y(y)$ is strictly increasing on the interval $(Y_\alpha - \epsilon, Y_\alpha]$ for some $\epsilon > 0$, then $P(Y \leq Y_{(\lceil n\alpha \rceil)})$ gets arbitrarily close to or higher than α as $n \rightarrow \infty$. If Y_m is the smallest value of y such that $P(Y leq Y_m) = \alpha$, $\alpha_1 = \alpha$, and $Y_m < Y_\alpha$, then $P(Y_{(\lceil n\alpha \rceil)} \geq Y_m) \rightarrow 1$ as $n \rightarrow \infty$. Hence $P(Y \leq Y_{(\lceil n\alpha \rceil)})$ gets arbitrarily close to or higher than α in all cases. Hence closed intervals have coverage eventually bounded below by $1 - \delta$.

Remark 4.1. Confidence intervals, prediction intervals, confidence regions, and prediction regions should used closed sets not open sets. The closed sets have the same volume as as the open sets, but have coverage at least as high as the open sets with weaker regularity conditions. In particular, confidence and prediction intervals should be closed intervals, not open intervals.

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

Theorem 4.1. 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

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{4.1}$$

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. For the uniform distribution, the population shorth is not unique. Of course the length of the population shorth is unique.

Definition 4.2. Let the shortest closed interval containing at least c of the $Y_1, ..., Y_n$ be

$$\text{shorth(c)} = [\text{Y}_{(\text{s})}, \text{Y}_{(\text{s+c}-1)}]. \tag{4.2}$$

Theorem 4.2, Frey (2013). Let $Y_1, ..., Y_n$ be iid. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{4.3}$$

For large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. The maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions.

Theorem 4.3, Frey (2013). 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. The large sample $100(1-\delta)\%$ shorth(c) prediction interval for Y_f is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \text{c} = \min(\text{n}, \lceil \text{n}[1 - \delta + 1.12\sqrt{\delta/\text{n}}\ ]\ \rceil). \tag{4.4}$$

Theorem 4.4. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(n)}$ be the order statistics of the squared training data $W_1, ..., W_n$ where $W_i = Y_i^2$ for $i = 1, ..., n$. Let k_n be given by Equation (4.3). Let $L_n = -U_n$ and $U_n = \sqrt{W_{(k_n)}}$. Then $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ PI for Y_f.

Note that $P(0 \leq W_f \leq U_n^2)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

By Chebyshev's inequality, for $k > 1$,

$$P(\mu - k\sigma \leq Y \leq \mu + k\sigma) \geq P(\mu - k\sigma < Y < \mu + k\sigma) \geq 1 - \frac{1}{k^2}. \quad (4.5)$$

Note that $k = 5$ gives 96% asymptotic coverage. The value $k = 1.96$ gives 95% coverage for the $N(\mu, \sigma^2)$ distribution, but the coverage could be as low as 74%. Use $\hat{\mu} = \overline{Y}$ and $\hat{\sigma} = S$, the square root of the unbiased sample variance estimator.

Theorem 4.5. Let $Y_1, ..., Y_n, Y_f$ be iid. Suppose that $E(Y) = \mu$ and the standard deviation $SD(Y) = \sigma$. Let $\hat{\mu}$ and $\hat{\sigma}$ be consistent estimators of μ and σ. Let $1 - 1/k^2 \geq 1 - \delta$. Let $\mu \pm k\sigma$ be continuity points of $F_Y(y)$. Then

$$[L_n, U_n] = [\hat{\mu} - k\hat{\sigma}, \hat{\mu} + k\hat{\sigma}]$$

is a large sample $100(1 - \delta)\%$ *Chebyshev PI* for Y_f.

A problem with the prediction intervals that cover $\approx 100(1 - \delta)\%$ of the training data cases Y_i (such as (4.2) using $c = k_n$ given by (4.3)), 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. For iid data, Frey (2013) used (4.4) to correct for undercoverage.

Remark 4.2. a) The Chebyshev PIs tend to be too long, and need second moments. b) The shorth PI (4.4) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval, respectively. The correction factor also increases the length. For a unimodal and symmetric error distribution, the nonparametric PI (4.1), shorth PI (4.4), and Theorem 4.4 PI are asymptotically equivalent, but PI (4.1) can be the shortest PI. c) The nonparametric PI and Theorem 4.4 PI can be much longer than the shorth PI if the data distribution is skewed. The Theorem 4.4 PI can very long if Y is a nonnegative random variable.

Example 4.1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76
```

```
                    33 = 111 - 78

                        689 = 778 - 89
shorth(3) = [76,89]
```

Remark 4.3. The large sample $100(1-\delta)\%$ shorth PI (4.4) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F_Y(y)$ is not strictly increasing in intervals $(L_s - \epsilon, L_s + \epsilon)$ and $(U_s - \epsilon, U_s + \epsilon)$ for some $\epsilon > 0$. To see the issue, suppose Y has probability mass function (pmf) $f(0) = 0.4$, $f(1) = 0.3$, $f(2) = 0.2$, $f(3) = 0.06$, and $f(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$ population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq y) = 1$ if $Y_i \leq y$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq y)$$

is the sample proportion of $Y_i \leq y$. If $Y_1, ..., Y_n$ are iid, then for fixed y, $n\hat{F}_n(y) \sim binomial(n, F(y))$. Thus $\hat{F}_n(y) \sim AN(F(y), F(y)(1-F(y))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (4.4) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (4.4) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

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 4.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

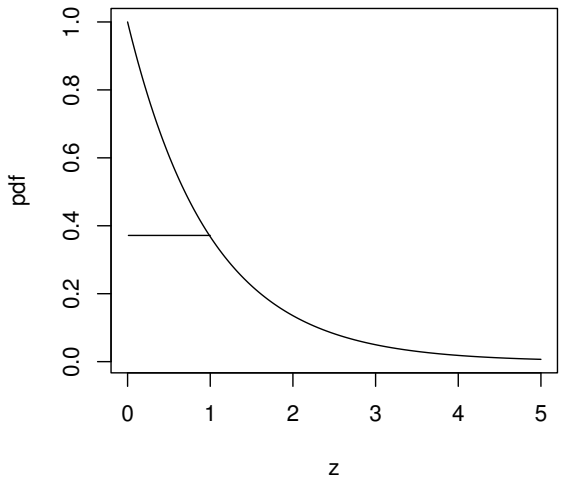

Fig. 4.1 The 36.8% Highest Density Region is [0,1]

Remark 4.4. Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to \infty$. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and confidence regions improves the performance for moderate sample size n.

4.2 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 4.3. 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 $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$. A PI is a prediction region where $p = 1$. 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 (4.13). For more about highest density regions, see Olive (2017b, pp. 148-155) and Hyndman (1996).

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 $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 4.4. 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 $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, 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 $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 4.5. 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*

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

where $\mathbf{1}$ is the $n \times 1$ vector of ones. 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 $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\, \overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\mathbf{1}\mathbf{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{G}\boldsymbol{W}$.

Definition 4.6. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (4.6)$$

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

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

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

See Definition 1.29 for the population mean and population covariance matrix. The Mahalanobis distance in Definition 4.6 is a random variable that estimates the population Mahalanobis distance of Definition 1.49. 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. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance from Definition 1.49 is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (4.7)$$

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.

Theorem 4.6. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\,\boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D_{\boldsymbol{x}}^2(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T-\boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T-\boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: i) $\boldsymbol{A}\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{W}$ by Theorem 3.13 iii), and the result follows.

ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x}-T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}-T) =$
$(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)^T[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)$
$= (\boldsymbol{x}-\boldsymbol{\mu})^T[s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{x}-T)^T[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-T)$
$+(\boldsymbol{x}-\boldsymbol{\mu})^T[s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu}-T) + (\boldsymbol{\mu}-T)^T[s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\boldsymbol{\mu})$
$+(\boldsymbol{\mu}-T)^T[s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu}-T) = s^{-1}D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1}D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T-\boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T-\boldsymbol{\mu}) =$
$n(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T-\boldsymbol{\mu}) = n(T-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) +$
$n(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T-\boldsymbol{\mu}) = n(T-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since $\sqrt{n}(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}]\sqrt{n}(T-\boldsymbol{\mu}) = O_P(1)o_P(1)O_P(1) = o_P(1)$. □

Next, we derive a prediction region for $\boldsymbol{x}_f$ if $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, $\boldsymbol{\mu} = E(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$ is nonsingular. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$ by Theorem 4.6. 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 (2013b) showed that the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\} \tag{4.8}$$

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 (4.10). A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.8) 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. Also see Remark 4.4. Empirically for many distributions, for $n = 20p$, the prediction region (4.8) 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

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{4.9}$$

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

$$c = \lceil nq_n \rceil \tag{4.10}$$

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

The nonparametric prediction region is due to Olive (2013b). 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 (4.11) 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 (4.11) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (4.11) 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. See Section 1.7.) 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 $D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \xrightarrow{D} D^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{\mu})$.

Theorem 4.7. 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

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\} \tag{4.11}$$

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

Theorem 4.8. 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* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{4.12}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (4.4) and (4.10) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 4.2 where ol = (Eq. 4.10)$/n$ is plotted versus fr = (Eq. 4.4)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows

that the correction factors do not differ greatly. R code to create Figure 4.2 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

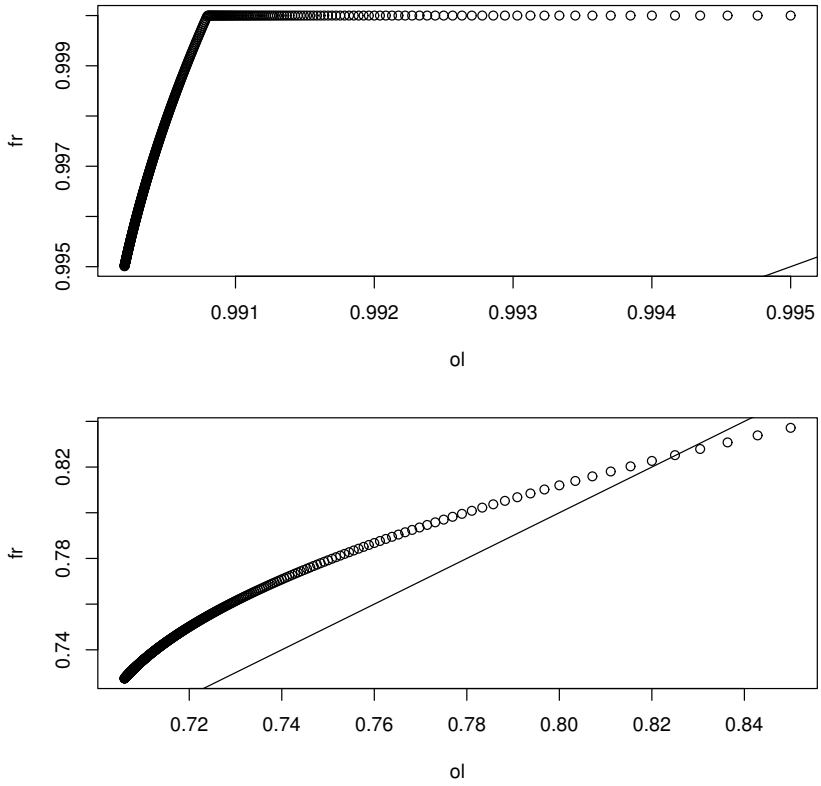

Fig. 4.2 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

Remark 4.5. The nonparametric prediction region (4.11) 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 (4.11) 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 $\boldsymbol{x}_f$.

Theorem 4.9, Chen (2011). Multivariate Chebyshev's Inequality: Let $E(\boldsymbol{x}) = \boldsymbol{\mu}$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$ be nonsingular. Then

$$P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$$

for $\gamma > p$.

For more on the above theorem, see Budny (2014) and Navarro (2014, 2016). For $h > 0$, consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}. \quad (4.13)$$

Using $\gamma = h^2 = p/\delta$ in (4.13) usually results in prediction regions with volume and coverage that is too large. Using $\gamma = h^2 = \chi^2_{p,1-\delta}$ in (4.13) gives the classical prediction region (4.12), which usually has volume and coverage that is too low, although bounded above 0 by Theorem 4.9 asymptotically if $0 < \delta < 0.25$. (The median of a chi-square χ^2_p distribution is $\chi^2_{p,0.5} \approx p - 2/3$.) Using $h^2 = D^2(U_n)$ tends to give better volume and coverage.

Remark 4.6. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. It can be shown that the ratio of the volumes of regions (4.12) and (4.11) 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 (4.12) is used, and the undercoverage tends to get worse as the dimension p increases.

Remark 4.7. The nonparametric prediction region (4.11) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Olive (2013b) 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.

For the multivariate lognormal distribution with $n = 20p$, the large sample nonparametric 95% prediction region (4.11) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some R code is below.

```
nruns=1000 #lognormal, p = 100, n = 20p = 2000
```

```
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, may take a few minutes
```

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 4.3 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 $\boldsymbol{\mu} = \mathbf{0}$. In Figure 4.3a),

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

Note that the ellipsoids are narrow with high positive correlation. In Figure 4.3b),

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

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.

4.3 Prediction Regions If n/p Is Small

See Zhang and Olive (2022).

4.4 Summary

4.5 Complements

See Frey (2013) for references about nonparametric PIs. For large sample theory for the shorth, see Chen and Shao (1999), Einmahl and Mason (1992), and Grübel (1988). A method for obtaining an asymptotically optimal PI from a parametric distribution, possibly with right censored data, is given by Olive, Rathnayake, and Haile (2021).

Prediction intervals and prediction regions can be used to estimate Bayesian credible intervals and Bayesian credible regions.

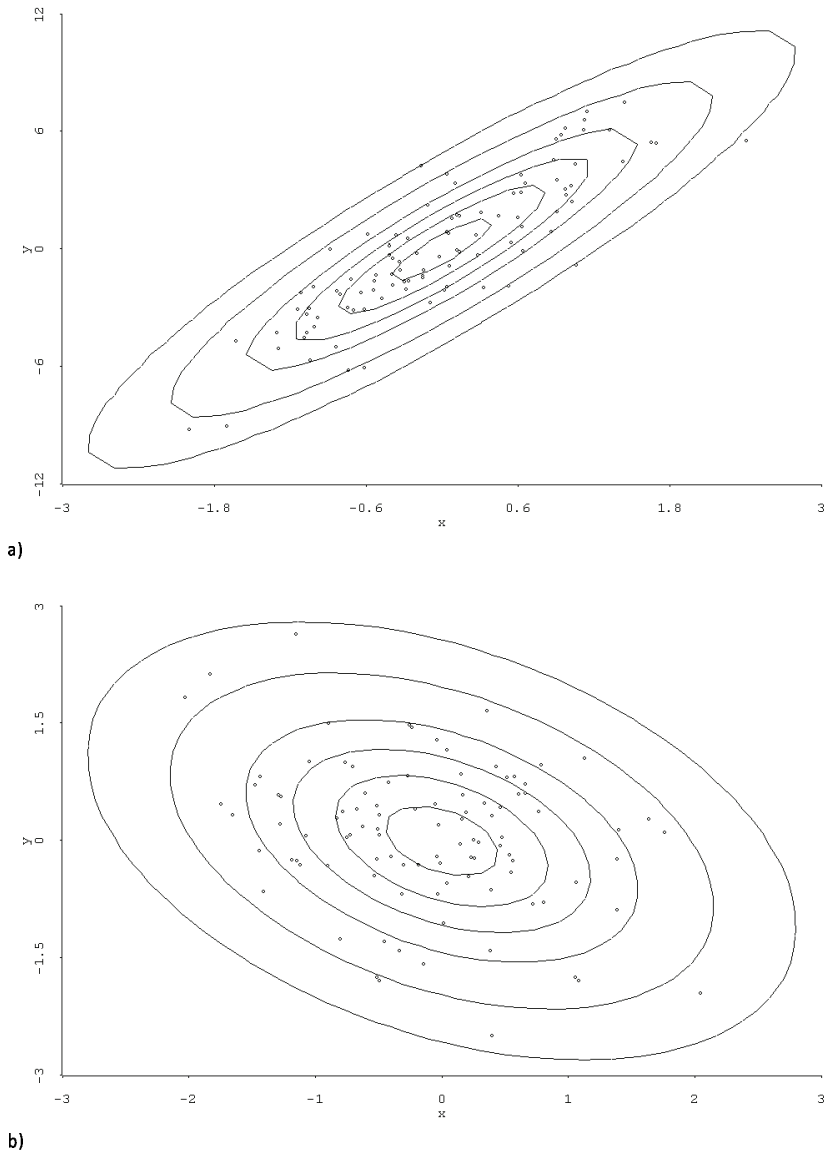

Fig. 4.3 Highest Density Regions for 2 MVN Distributions

Software. The simulations were done in *R*. See R Core Team (2016). The function `predrgn` makes the nonparametric prediction region and determines whether $\boldsymbol{x}_f$ is in the region. The function `predreg` also makes the nonparametric prediction region, and determines if $\mathbf{0}$ is in the region. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$.

4.6 Problems

4.1. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

```
0.0  0.8  1.0  1.2  1.3  1.3  1.4  1.8  2.4  4.6
```

4.2. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

4.3. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

4.4. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number      1     2     3     4   ...  97   98   99   100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

R Problems

Use the command *source("G:/lsamppack.txt")* **to download the functions** and the command *source("G:/lsampdata.txt")* **to download the data. See Preface.** Typing the name of the `lsamppack` function, e.g. *predsim*, will display the code for the function. Use the `args` command, e.g. *args(predsim)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/lsamphw.txt) into *R*.

4.5. a) 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 $n \to \infty$.

b) Were the three coverages near 90%?