Highest Density Region Prediction

David J. Olive [*]
Southern Illinois University

November 3, 2015

Abstract

Practical large sample prediction regions for an $m \times 1$ future response vector $\boldsymbol{y}_f$, given $\boldsymbol{x}_f$ and past training data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$, are developed for models of the form $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{e}_i = m(\boldsymbol{x}_i) + \boldsymbol{e}_i$ where the distribution of $\boldsymbol{e}_i$ may not be known. The classical prediction regions assume that the $\boldsymbol{e}_i$ are iid from a multivariate normal distribution, do not perform well if the normality assumption is violated, and the performance decreases as the dimension m increases.

The new $100(1-\delta)\%$ prediction regions have a parameter c such that c of the training data cases $\boldsymbol{y}_i$ are in their prediction regions, even if the model is wrong or misspecified.

KEY WORDS: bootstrap, cross validation, prediction interval, prediction region, time series, multivariate regression, statistical learning.

[*] David J. Olive is Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (E-mail: *dolive@siu.edu*).

1. Introduction

This paper gives a practical method for obtaining a large sample prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ if the data are from a model of the form

$$\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{e}_i = m(\boldsymbol{x}_i) + \boldsymbol{e}_i \tag{1}$$

where the iid zero mean error vectors $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ may come from an unknown distribution, and $\boldsymbol{x}_i$ is a $p \times 1$ vector of predictors. Examples of such models are the *location model* $Y_i = \mu + e_i$, the *multivariate location and dispersion model* where $\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{e}_i$ and the $\boldsymbol{e}_i$ have nonsingular covariance matrix $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_{\boldsymbol{e}}$, the *multiple linear regression model* $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$, the *additive error regression model* $Y_i = m(\boldsymbol{x}_i) + e_i$ (which includes many nonlinear and nonparametric regression models), many *time series models*, and the *multivariate linear regression model* discussed in Section 3.

Consider predicting a future test value $\boldsymbol{y}_f$, given $\boldsymbol{x}_f$ and past training data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ...,$ $(\boldsymbol{x}_n, \boldsymbol{y}_n)$. A *large sample* $(1-\delta)100\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \xrightarrow{P} 1-\delta$ 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 $\boldsymbol{y}_f|\boldsymbol{x}_f$. As an example, a large sample $100(1-\delta)\%$ *prediction interval* (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\delta$ as the sample size $n \to \infty$. If the highest density region is an interval, then a PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage.

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 $\hat{\boldsymbol{y}}_f$ of the future random vector $\boldsymbol{y}_f$. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. These classical regions do not perform well if the normality assumption is violated, and the performance decreases as the dimension m increases, as will be shown below.

The highest density region can be constructed, in principle, if the probability density function (pdf) $f(\boldsymbol{z})$ of $\boldsymbol{y}_f$ is known. See Hyndman (1996). The method of construction will first be illustrated for a random variable Y_f with pdf $f(z)$. To find the $(1-\delta)100\%$ highest density region corresponding to 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 $1-\delta$. Then the highest density region is the union of the k intervals. Often the pdf is unimodal and decreases rapidly as z moves away from the mode. Then $k = 1$ and the highest density region is an interval. If Y_f has an exponential distribution, the highest density region is $(0, \xi_{1-\delta})$ where $P(Y_f \leq \xi_\alpha) = \alpha$. For a symmetric unimodal distribution, the highest density region is $(\xi_{\delta/2}, \xi_{1-\delta/2})$. In general, slice the pdf $f(\boldsymbol{z})$ with a horizontal hyperplane.

An important multivariate distribution with a simple highest density region is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with pdf

$$f(\boldsymbol{z}) = k_p|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})]$$

where $k_p > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case. Following Johnson (1987, pp. 107-108), $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = c_{\boldsymbol{x}} \boldsymbol{\Sigma}$ for some constant $c_{\boldsymbol{x}} > 0$ if second moments exist. The population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{2}$$

and for elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{3}$$

If g is continuous and decreasing, then the highest density region is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \le u_{1-\delta}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le u_{1-\delta}\} \tag{4}$$

where $P(U \le u_{1-\delta}) = 1 - \delta$.

Typically the pdf of $\boldsymbol{y}_f$ is not known. Then there is a moderate amount of literature for prediction regions that may perform well for small m. See Lei, Robins, and Wasserman (2013), who estimate $f(\boldsymbol{z})$ with a kernel density estimator, for references.

There are two practical methods for obtaining prediction regions: use the shorth estimator if Y_f is a random variable, and use sample Mahalanobis distances if $\boldsymbol{y}_f$ is a random vector. Let $Z_1, ..., Z_n$ be random variables, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics, and let c be a positive integer. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Let

$$\text{shorth(c)} = (\text{Z}_{(\text{d})}, \text{Z}_{(\text{d+c}-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}) \tag{5}$$

correspond to the interval with the smallest distance. Let

$$k_n = \lceil n(1-\delta) \rceil \tag{6}$$

where $\lceil x \rceil$ is the smallest integer $\ge x$, e.g., $\lceil 7.7 \rceil = 8$. Under mild conditions, the shorth(k_n) estimator is a consistent estimator of the highest density (interval) region if the $Z_i = Y_i$ are iid. See Grübel (1988).

To describe the second practical prediction region, let the $p \times 1$ column vector T be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D^2_i = D^2_i(T, \boldsymbol{C}) = D^2_{\boldsymbol{x}_i}(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{7}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\text{n}-1} \sum_{\text{i}=1}^{\text{n}} (\boldsymbol{x}_{\text{i}} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\text{i}} - \overline{\boldsymbol{x}})^{\text{T}}. \tag{8}$$

The volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq h^2\} \text{ is equal to } \frac{2\pi^{\mathrm{p}/2}}{\mathrm{p}\Gamma(\mathrm{p}/2)} \mathrm{h}^{\mathrm{p}} \sqrt{\det(\boldsymbol{S})}, \tag{9}$$

see Johnson and Wichern (1988, pp. 103-104).

Note that if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$, then

$$D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) = (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

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

Thus the sample percentiles of $D^2_i(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_p$.

The second practical $100(1 - \delta)\%$ prediction region is the hyperellipsoid

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{10}$$

Olive (2013) showed that this prediction region estimates the highest density region for a large class of EC distributions if $c = k_n$ given by (6). Di Bucchianico, Einmahl, and Mushkudiani (2001) used the minimum volume ellipsoid to compute small volume covering regions for $m \leq 2$.

A problem with the prediction regions (5) and (10) is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n if $c = k_n = \lceil n(1-\delta) \rceil$. Note that both prediction regions cover $k_n \approx 100(1 - \delta)\%$ of the training data cases $\boldsymbol{y}_i$, and empirically statistical methods perform worse on test data. Increasing c will improve the coverage for moderate samples. Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1 - \delta)\%$ PI where $c = \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil$.

The practical method for producing a prediction region for $\boldsymbol{y}_f$ from model (1) is to create pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{e}}_n$ using the residuals $\hat{\boldsymbol{e}}_i$ and the predicted value $\hat{\boldsymbol{y}}_f$. Then apply one of the two practical prediction regions (5) or (10) to the pseudodata but modify $c = k_n = \lceil n(1 - \delta) \rceil$ so that the coverage is better for moderate samples.

Let *df* be the model degrees of freedom. Then empirically for many models, for $n \approx 20df$, the two prediction regions (5) and (10) applied to iid data or pseudodata using $k_n = \lceil n(1-\delta) \rceil$ tend to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.} \tag{11}$$

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

$$c = \lceil n q_n \rceil \tag{12}$$

in (5) or (10) decreased the undercoverage.

There are at least two reasons to use pseudodata. If there was an iid sample $\boldsymbol{z}_1, ..., \boldsymbol{z}_k$ from the same distribution as $\boldsymbol{y}_f$, then the prediction region could be applied to $\boldsymbol{z}_1, ..., \boldsymbol{z}_k$. If $E(\boldsymbol{y}_f | \boldsymbol{x}_f) = m(\boldsymbol{x}_f)$ was known, and there was an iid sample $\boldsymbol{e}_1, ..., \boldsymbol{e}_k$ from the error distribution, then $\boldsymbol{z}_i = m(\boldsymbol{x}_f) + \boldsymbol{e}_i$. The pseudodata uses $\hat{\boldsymbol{y}}_f = \hat{m}(\boldsymbol{x}_f)$ and $\hat{\boldsymbol{e}}_i$ in place of $m(\boldsymbol{x}_f)$ and $\boldsymbol{e}_i$ with $k = n$.

The second reason is the relationship between the pseudodata and the bootstrap. One way to get a bootstrap distribution is to draw a sample of size k with replacement from the n residuals $\hat{\boldsymbol{e}}_i$ to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{e}}_1^B, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{e}}_k^B$. As $k \to \infty$ the bootstrap sample will take on n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{e}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Olive (2013) showed that one of the sufficient conditions for the shorth PI to be large sample $100(1-\delta)\%$ PI is that the sample quantiles of the residuals be consistent estimators of the population quantiles of the error distribution. Then the shorth of the residuals is a consistent estimator of the population shorth of the error distribution. For multiple linear regression and consistent estimators of $\hat{\boldsymbol{\beta}}$, the residuals behave well if the vectors of predictors $\boldsymbol{x}_i$ are bounded in probability. See Olive and Hawkins (2003) and Rousseeuw and Leroy (1987, pp. 127-129).

The next four examples show that the above ideas have been used to create prediction regions for multiple linear regression, additive error regression, the location model, and the multivariate location and dispersion model. These ideas will be used to develop new prediction regions for time series and multivariate regression models in Sections 2 and 3.

Example 1. Consider the multiple linear regression model $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith *leverage* and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. Let

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{13}$$

Following Olive (2007), apply the shorth estimator to the residuals: let $c = k_n$ and $\text{shorth}(c) = (r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$. Then a large sample 100 $(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n\tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n\tilde{\xi}_{1-\delta_2}). \tag{14}$$

This practical PI is asymptotically optimal if the $\boldsymbol{x}_i$ are bounded in probability and the iid e_i come from a large class of zero mean unimodal distributions. Also see Cai, Tian, Solomon, and Wei (2008).

The 100 $(1-\delta)\%$ classical PI for Y_f is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred) \tag{15}$$

where $se(pred) = \sqrt{MSE\ (1+h_f)}$ and $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom. This PI may not perform well if the e_i are not iid $N(0,\sigma^2)$ since the normal quantiles are not the correct quantiles for other error distributions.

Example 2. Olive (2013) derived an asymptotically optimal PI (for a large class of zero mean unimodal error distributions) for the additive error regression model, provided that $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$ is a consistent estimator of $m(\boldsymbol{x}_f)$ and the shorth of the residuals is a consistent estimator of the population shorth of the error distribution. Let

$$b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n+2p}{n-p}}. \tag{16}$$

Let $c = \lceil n q_n \rceil$ where q_n is given by (11) with m replaced by p. Let shorth$(c) = (r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ be the shorth of the residuals. Then the 100 $(1-\delta)$% large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}), \tag{17}$$

and is similar to (14).

Geometrically, plot $\hat{Y}_i$ versus Y_i on the vertical axis. Then the PIs are given by two parallel lines with unit slope that contain c of the training data cases Y_i where $\hat{Y}_f$ is on the identity line with unit slope and zero intercept. Hence c of the training data are within their PIs even if the additive error regression model is wrong. If the plotted points do not scatter about the identity line in an evenly populated band, the method of Lei and Wasserman (2014) may be useful.

Example 3. The location model is a special case of both the regression model and of the multivariate location and dispersion model. Let $a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+1}{n-1}}$. Let $c = k_n = \lceil n(1-\delta) \rceil$. Let shorth$(c) = (Y_{(d)}, Y_{(d+c-1)})$ be the shorth of the Y_i. Let MED(n) be the sample median. Following Olive (2013), if $Y_1, ..., Y_n$ are iid, then the recommended large sample $100(1-\delta)$% PI for Y_f is the closed interval $[L_n, U_n] = [(1-a_n)MED(n) + a_n Y_{(d)}, (1-a_n)MED(n) + a_n Y_{(d+c-1)}]$. This PI is (14) using the least absolute deviations estimator, but with a closed interval. Compare Frey (2013). This PI also works for discrete data where a good PI should be short with coverage $\geq 1-\delta$, asymptotically.

Example 4. Olive (2013) derived a prediction region for the multivariate location and dispersion model where the $\boldsymbol{x}_i$ are iid random vectors. 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. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \tag{18}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \tag{19}$$

A future observation (random vector) $\boldsymbol{x}_f$ is in region (18) if $D_{\boldsymbol{x}_f} \leq h$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (18) is a large sample $(1-\delta)100$% prediction region if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i where q_n is given by (11) with $m = p$. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then region (18) is asymptotically optimal on a large class of elliptically contoured distributions in that its volume converges in probability to the volume of the highest density region (4).

The recommended prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the large sample $100(1-\delta)$% prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ...,, \boldsymbol{x}_n$ is

$$\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(up)}\}, \tag{20}$$

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

$$\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le \chi^2_{p,1-\delta}\}. \tag{21}$$

See Chew (1966) and Johnson and Wichern (1988, pp. 134, 151).

The prediction region (20) contains q_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. Also, (20) is a large sample $100(1-\delta)\%$ prediction region if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$ provided the covariance matrix is nonsingular, although prediction regions with smaller volume may exist.

The ratio of the volumes of regions (21) and (20) is

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

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 $D^2_{(up)} \approx D^2_{\boldsymbol{x},0.5} = 6$. 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 inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \le \gamma) \ge 1 - p/\gamma > 0$ for $\gamma > p$. See Budny (2014) and Navarro (2014, 2015).

The following two sections will illustrate how to develop new prediction regions for more models of the form (1), and the above examples will be useful. Obtaining prediction regions when the errors are not additive is a difficult problem. See Cai, Tian, Solomon, and Wei (2008) for some useful results.

2. Time Series Prediction Intervals

Many time series models have the form

$$Y_t = \tau + \sum_i \psi_i Y_{t-ik_i} + \sum_j \nu_j e_{t-jk_j} + e_t$$

where the e_t are iid with 0 mean and variance σ^2_e. For example, the Box, Jenkins, and Reinsel (1994) multiplicative seasonal ARIMA$(p,d,q) \times (P,D,Q)_s$ time series models have this form. Then the l step ahead forecast for a future value Y_{t+l} is $\hat{Y}_t(l)$ and the l step ahead forecast residual is $\hat{e}_t(l) = Y_{t+l} - \hat{Y}_t(l)$. For example, a common choice is

$$\hat{Y}_t(l) = \hat{\tau} + \sum_i \hat{\psi}_i Y^*_{t+l-ik_i} + \sum_j \hat{\nu}_j \hat{e}^*_{t+l-jk_j}$$

where $\hat{e}_t$ is the tth residual, $Y^*_{t+l-ik_i} = Y_{t+l-ik_i}$ if $l - ik_i \le 0$, $Y^*_{t+l-ik_i} = \hat{Y}_t(l - ik_i)$ if $l - ik_i > 0$, $\hat{e}^*_{t+l-jk_j} = \hat{e}_{t+l-jk_j}$ if $l - jk_j \le 0$, and $\hat{e}^*_{t+l-jk_j} = 0$ if $l - jk_j > 0$, and the forecasts $\hat{Y}_t(1), \hat{Y}_t(2), ..., \hat{Y}_t(L)$ are found recursively if there is data $Y_1, ..., Y_t$. Typically the residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ are the 1 step ahead forecast residuals and the fitted or predicted values $\hat{Y}_t = \hat{Y}_{t-1}(1)$ are the 1 step ahead forecasts.

In the simulations, a moving average MA(2) = ARIMA(0,0,2)$\times(0,0,0)_1$ model, $Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$, was used. Suppose data $Y_1, ..., Y_n$ from this model is available. The R software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = 1, ..., n$ where $\hat{Y}_t = \hat{Y}_{t-1}(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_{t-1} + \hat{\theta}_2 \hat{e}_{t-2}$ for $t = 3, ..., n+1$, and $\hat{e}_t(1) = Y_{t+1} - \hat{Y}_t(1)$ for $t = 3, ..., n-1$. Hence there are n 1 step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available. Then $\hat{Y}_t(2) = \hat{\tau} + \hat{\theta}_2 \hat{e}_t$ for $t = 1, ..., n$.

Often time series PIs assume normality and are similar to equation (22) below. There is a large literature on alternative PIs, especially for AR(p) models. See Clements and Kim (2007), Kabaila and He (2007), Pan and Politis (2015ab), Panichkitkosolkul and Niwitpong (2012), Thombs and Schucany (1990), and Vidoni (2009) for references. For many time series models, a large sample $100(1-\delta)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = \hat{Y}_t(l) \pm t_{1-\delta/2, n-p-q} SE(\hat{Y}_t(l)). \tag{22}$$

Applying a PI similar to the one in Example 3 to $\overline{e}_t = Y_t - \overline{Y}$ ignores the time series structure of the data. Let shorth$(k_n) = (\tilde{L}_n, \tilde{U}_n)$ be computed from the $\overline{e}_t$. Then the large sample shorth(k_n) $100(1-\delta)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = (\overline{Y} + a_n \tilde{L}_n, \overline{Y} + a_n \tilde{U}_n) \tag{23}$$

where a_n is given in Example 3. For stationary invertible ARMA(p, q) models, this PI is too long for l near 1, but should have short length for large l and if $l > q$ for an MA(q) model.

The following PI is new and takes into account the time series structure of the data. A similar idea in Masters (1995, p. 305) is to find the l step ahead forecast residuals and use percentiles to make PIs for Y_{t+l} for $l = 1, ..., L$. For ARIMA(p, d, q) models, let $c = \lceil n q_n \rceil$ where q_n is given by (11) with m replaced by $p + q$. Compute shorth$(c) = (\tilde{L}_n, \tilde{U}_n)$ of the l-step ahead forecast residuals $\hat{e}_t(l)$. Then a large sample $100(1-\delta)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = (\hat{Y}_n(l) + \tilde{L}_n, \hat{Y}_n(l) + \tilde{U}_n). \tag{24}$$

This PI is similar to that of Example 2.

Figure 1 shows a simulated MA(2) time series with $n = 100$, $L = 7$ and $U(-1, 1)$ errors. The horizontal lines correspond to the 95% PI (23). Two of the one hundred time series points $Y_1, ..., Y_{100}$ lie outside of the two lines. All seven of the future cases $Y_{101}, ..., Y_{107}$ lie within their large sample 95% PI.

The simulations used the MA(2) model where the distribution of the white noise e_t is N(0,1), t_5, $U(-1, 1)$, or (EXP(1) - 1). All these distributions have mean 0, but the fourth distribution is not symmetric. The simulation generated 5000 time series of length $n + L$ and PIs are found for $Y_{n+1}, ..., Y_{n+L}$. The simulations used $L = 7$ and 95% and 50% nominal PIs. The two types of PI used were the normal PI (22), and the possibly asymptotically optimal PI used (23) for Y_{t+j} where $j > 2$ and (24) for $j = 1, 2$. These two types of PIs are denoted by N and A respectively in the tables. The simulated coverages and average lengths of the PI are shown using two rows in the tables. With 5000 runs, coverages between 0.94 and 0.96 suggest that there is no reason to believe that the nominal coverage is not 0.95.

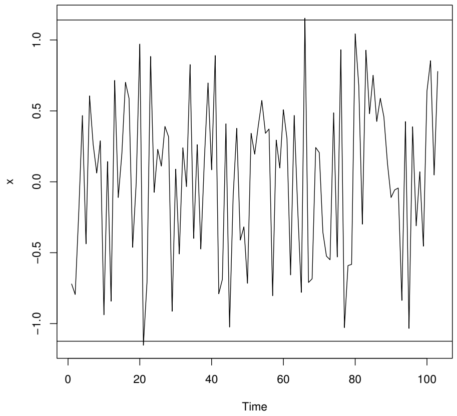

Figure 1: PIs for an MA(2) Time Series with Uniform$(-1, 1)$ Errors

Table 1: Normal Errors

δ	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9396	0.9432	0.9444	0.9436	0.9486	0.9498	0.9462
0.05	100		3.889	4.072	4.198	4.198	4.198	4.198	4.198
0.05	100	A	0.9482	0.9582	0.9550	0.9496	0.9556	0.9590	0.9532
0.05	100		4.143	4.509	4.461	4.461	4.461	4.461	4.461
0.05	1000	N	0.9520	0.9464	0.9476	0.9474	0.9496	0.9524	0.9474
0.05	1000		3.919	4.080	4.179	4.179	4.179	4.179	4.179
0.05	1000	A	0.9520	0.9488	0.9482	0.9446	0.9478	0.9500	0.9482
0.05	1000		3.913	4.086	4.170	4.170	4.170	4.170	4.170

Table 2: Uniform Errors

α	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9904	0.9796	0.9820	0.9794	0.9780	0.9818	0.9800
0.05	100		2.254	2.359	2.433	2.433	2.433	2.433	2.433
0.05	100	A	0.9816	0.9756	0.9756	0.9702	0.9730	0.9776	0.9754
0.05	100		2.132	2.342	2.388	2.388	2.388	2.388	2.388
0.05	1000	N	1.0000	0.9898	0.9826	0.9830	0.9834	0.9822	0.9844
0.05	1000		2.263	2.357	2.416	2.416	2.416	2.416	2.416
0.05	1000	A	0.9548	0.9486	0.9494	0.9512	0.9514	0.9506	0.9478
0.05	1000		1.913	2.094	2.182	2.182	2.182	2.182	2.182

Some results are shown in Tables 1 and 2 for 95% PIs. From Table 1 for normal errors, note that for $n = 1000$, the coverages and lengths of PIs (23) and (24) were very similar to the those of PI (22). PIs (23) and (24) were longer than the normal PI (22) for $n = 100$ and normal errors. From Table 2 for uniform errors, the normal PIs (22) were too long and the coverage was too high for 95% PIs. The alternative PIs (23) and (24) had coverage closer to the nominal level with good coverage for $n = 1000$.

For t_5 errors, the 95% normal PI (22) worked well, but the nominal 50% normal PI (22) had coverage that was too high and the average lengths were too large. The alternative PIs had coverage near 50% with shorter average lengths. For EXP(1) - 1 errors, for 95% PIs the normal PIs (22) were longer than the alternative PIs (23) and (24). For the 50% PIs, the normal PIs (22) were too long with coverage that was too high. The alternative PIs (23) and (24) were shorter with good coverage.

3. Prediction Regions for Multivariate Regression

This section will derive a prediction region for model (1), and then consider the multivariate linear regression model as a special case. The following technical theorem will be needed to prove Theorem 2, which shows how to obtain a practical prediction region using pseudodata.

Theorem 1. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j\hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Proof: Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$. Now

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) +$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \, \hat{\boldsymbol{\Sigma}}_j^{-1})(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \; +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b). □

Theorem 2. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is positive definite, and the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $0 < \delta < 1$ and $D_{(U_n)}$ be the q_nth sample quantile of the D_i. Let the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}\}. \tag{25}$$

a) Consider the n prediction regions for the training data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (25) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}\}$, then the prediction region (25) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1 - \delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ exists, Theorem 1 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances also converge in probability, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1 - \delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region, and $q_n \to 1 - \delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Of course the volume of the prediction region could be large if a poor estimator is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. Olive, Pelawa Watagoda, and Rupasinghe Arachchige Don (2015) suggest that the residual, response, and DD plots described below can be used to check model assumptions.

Prediction region (25) can be used for the Su and Cook (2012) inner envelopes estimator and the seemingly unrelated regressions model. Theorem 3 shows that prediction region (25) is the Example 4 prediction region applied to pseudodata for the *multivariate linear model*

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i \tag{26}$$

for $i = 1, ..., n$ that has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. Multivariate linear regression and MANOVA models are special cases. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\begin{array}{cccc} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{array} \right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. Subscripts are needed for the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \boldsymbol{I}_n$ for $i, j = 1, ..., m$.

The $n \times m$ matrix of response variables and $n \times m$ matrix of errors are

$$\boldsymbol{Z} = [\boldsymbol{Y}_1 \ \boldsymbol{Y}_2 \ \ldots \ \boldsymbol{Y}_m] = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix} \quad \text{and} \quad \boldsymbol{E} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \ldots \ \boldsymbol{e}_{\text{m}}] = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix},$$

while the $n \times p$ design matrix of predictor variables is $\boldsymbol{X}$.

Least squares is the classical method for fitting the multivariate linear model. The *least squares estimators* are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = [\hat{\boldsymbol{\beta}}_1 \ \hat{\boldsymbol{\beta}}_2 \ \ldots \ \hat{\boldsymbol{\beta}}_m]$. The matrix of *predicted values* or *fitted values* $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = [\hat{\boldsymbol{Y}}_1 \ \hat{\boldsymbol{Y}}_2 \ \ldots \ \hat{\boldsymbol{Y}}_m]$. The matrix of *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} = [\boldsymbol{r}_1 \ \boldsymbol{r}_2 \ \ldots \ \boldsymbol{r}_m]$. These quantities can be found from the m multiple linear regressions of Y_j on the predictors: $\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$ and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$.

Theorem 3. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (25) is the Example 4 prediction region applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 2 by Su and Cook (2012). Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (8) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\boldsymbol{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. □

These prediction regions can be displayed with the Rousseeuw and Van Driessen (1999) DD plot of $MD_i = D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $RD_i = D_i(T, \boldsymbol{C})$. For $(T, \boldsymbol{C})$, we will use the Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for a large class of elliptically contoured distributions, where $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Also see Zhang, Olive, and Ye (2012). For iid data and large n, Olive (2002) showed that plotted points in the DD plot scatter tightly about a line through the origin for a large class of elliptically contoured distributions, and about the identity line with unit slope and zero intercept if the data are multivariate normal. Simulations suggest that the DD plot of the residuals can be used in a similar way.

Three regions (18) used by Olive (2013) for the multivariate location and dispersion model can be extended to multivariate linear regression. Let (25) be the nonparametric region with $h = D_{(U_n)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{m,q_n}$ where $P(W \leq \chi^2_{m,q_n}) = q_n$ if $W \sim \chi^2_m$. The semiparametric and parametric regions are only conjectured to be large sample prediction regions for the multivariate regression model, but are useful as diagnostics. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, ..., n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}, \tag{27}$$

while the (Johnson and Wichern 1988: p. 312) classical large sample $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi^2_{m,1-\delta}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \sqrt{\chi^2_{m,1-\delta}}\}. \tag{28}$$

The nonparametric prediction region (27) has simple geometry. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$. Then R_r is a hyperellipsoid with center $\boldsymbol{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(up)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

Two other plots are useful for checking the model. A *response plot* for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij} where $i = 1, ..., n$. The identity line is added to the plot as a visual aid. A *residual plot* corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}. Suppose the multivariate linear regression model is good, the error distribution is not highly skewed, and $n \geq 10p$. Then the plotted points should cluster about the identity line or $r = 0$ line in each of the m response and residual plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression. See Olive and Hawkins (2005) and Cook and Weisberg (1999a, p. 432; 1999b).

Example 5. Cook and Weisberg (1999a, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where

S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$ and $X_4 = H$: the shell length, log(width) and height. Figures 2 and 3 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25 and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. Figure 4 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Cases 8, 48 and 79 have especially large distances. The horizontal line $RD \approx 3$ corresponds to the semiparametric region. These two lines were also the 95th percentiles of the MD_i and RD_i. The horizontal line $RD \approx 2.45$ corresponds to the parametric MVN region. A vertical line $MD \approx 2.45$ (not shown) corresponds to a large sample classical region.

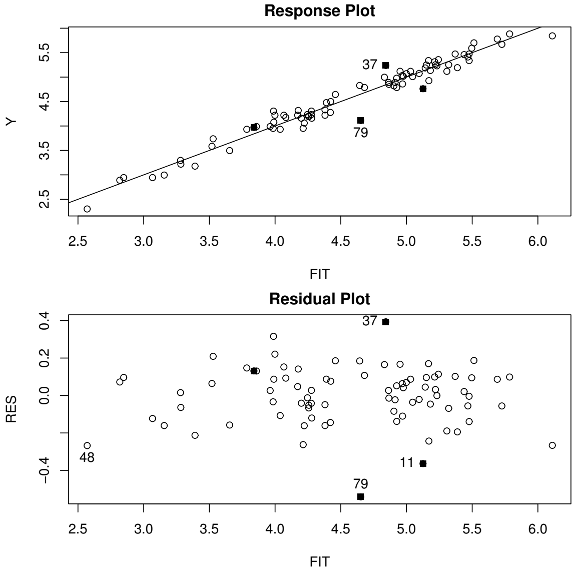

Figure 2: Plots for $Y_1 = \log(S)$.

Suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots (not shown) for Y_2 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 5 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error vector distribution is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$, and still contains 95% of the data.

A small simulation was used to study the prediction regions. First $m \times 1$ error vectors $\boldsymbol{w}_i$ were generated such that the m errors are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations are $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the data clusters about the line

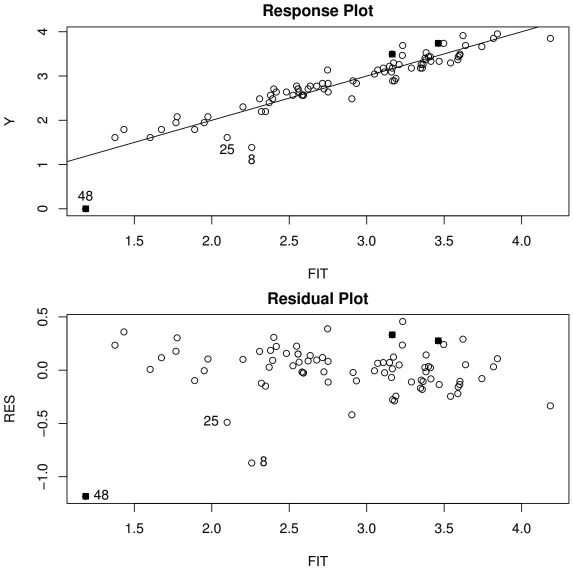

Figure 3: Plots for $Y_2 = \log(M)$.

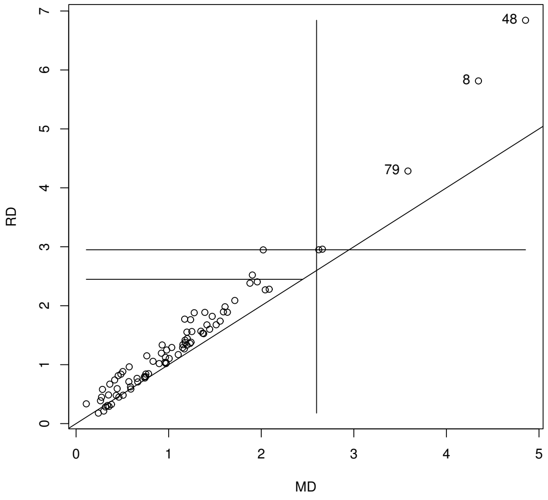

Figure 4: DD Plot of the Residual Vectors for the Mussel Data.

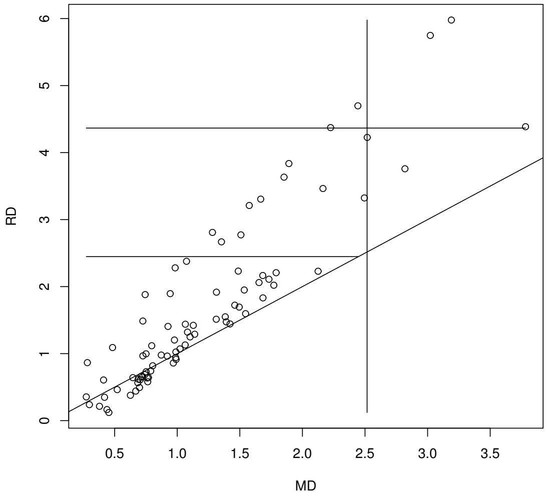

Figure 5: DD Plot if $Y_2 = M$.

in the direction of $(1,...,1)^T$. Used $\boldsymbol{w}_i \sim N_m(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\boldsymbol{0}, \boldsymbol{I}) + \tau N_m(\boldsymbol{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

Then 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. Following Olive (2013), consider the prediction region $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. Then the ratio of the prediction region volumes

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

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. Here h_1 and h_2 were the cutoffs $D_{(U_n)}(T_i, \boldsymbol{C}_i)$ for $i = 1, 2$, and $h_3 = \sqrt{\chi^2_{m,q_n}}$.

If, as conjectured, the RMVN estimator is a consistent estimator when applied to the residual vectors instead of iid data, then the volume ratios converge in probability to 1 if the iid zero mean errors $\sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the volume ratio converges to 1 for $i = 1$ for a large class of elliptically contoured distributions. These volume ratios were denoted by voln and volm for the nonparametric and parametric MVN regions. The coverage was the proportion of times the prediction region contained $\boldsymbol{y}_f$ where ncov, scov and mcov are for the nonparametric, semiparametric and parametric MVN regions.

In the simulations, took $n = 3(m + p)^2$ and $m = p$. Table 3 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three elliptically contoured distributions. Since the volume of

Table 3: Coverages for 90% Prediction Regions.

$\boldsymbol{w}$ dist	n	$m = p$	ncov	scov	mcov	voln	volm
MVN	48	2	0.901	0.905	0.888	0.941	0.964
MVN	300	5	0.889	0.887	0.890	1.006	1.015
MVN	1200	10	0.899	0.896	0.896	1.004	1.001
MIX	48	2	0.912	0.927	0.710	0.872	0.097
MIX	300	5	0.906	0.911	0.680	0.882	0.001
MIX	1200	10	0.904	0.911	0.673	0.889	0+
MVT(7)	48	2	0.903	0.910	0.825	0.914	0.646
MVT(7)	300	5	0.899	0.909	0.778	0.916	0.295
MVT(7)	1200	10	0.906	0.911	0.726	0.919	0.061
LN	48	2	0.912	0.926	0.651	0.729	0.090
LN	300	5	0.915	0.917	0.593	0.696	0.009
LN	1200	10	0.912	0.916	0.593	0.679	0+

the prediction region is proportional to h^m, the volume can be very small if h is too small and m is large. Parametric prediction regions usually give poor estimates of h when the parametric distribution is misspecified. Hence the parametric MVN region only performed well for multivariate normal data.

4. Three Applications

One application is bootstrap tests. See Olive (2015). A similar technique can be used to estimate the $100(1-\delta)\%$ Bayesian credible region for $\boldsymbol{\theta}$. Generate $B = \max(100000, n)$ values of $\boldsymbol{\theta}$ from the posterior distribution, and compute the prediction region (20). Olive (2014, pp. 283, 364) used the shorth(k_n) estimator to compute shorter bootstrap confidence intervals, and to estimate the highest density region corresponding to a known posterior pdf for Bayesian inference.

A third application is for cross validation. In addition to large sample theory, want the prediction regions to work well on a single data set as future observations are gathered, but only have the training data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$. Following James, Witten, Hastie, and Tibshirani (2013, pp. 181-186), to perform k-fold cross validation randomly divide the data set into k groups of approximately equal size. For $i = 1, ..., k$, compute the model from $k - 1$ groups other than the ith group, and use the ith group as a validation set. Much like k-fold cross validation for classification, compute the prediction region $\mathcal{R}_i$ for $\boldsymbol{y}_f = \boldsymbol{y}_j$ for each j in the ith group. Compute the proportion of times $\boldsymbol{y}_i$ was not in its prediction region $\mathcal{R}_i$ for $i = 1, ..., n$ and compute the average volume of the n prediction regions. Want the proportion near the nominal proportion δ and small average volume if two or more models or prediction regions are being considered. Hence $CV_{(n)}(PR) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} I(\boldsymbol{y}_i \notin \mathcal{R}_i)$. For additive error regression, the average volume is just the average length of the n PIs. These two statistics can be used to augment the traditional cross validation estimates such as $CV_{(k)}(MSE) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$ where $MSE_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_j - \hat{Y}_j)^2$ when an additive error regression model is used and n_i is the number of cases in the ith group.

Statistical Learning methods for the additive error regression model often have a parameter controlling the "flexibility" of the estimator. As the flexibility increases, the estimator overfits the training data, eventually using interpolation. The overfit data will have residuals that under estimate the errors. Hence the average length of the PIs will be small, but the $CV_{(n)}(PR)$ become 1 when there is interpolation. If the flexibility is too low, the average length of the PIs should be large since underfit data will have residuals that over estimate the errors.

5. Discussion

This paper suggests a practical method for making prediction regions when the error distribution may be unknown. Plots and simulations were done in *R*. See R Development Core Team (2011). Programs are in the collections of functions *tspack* and *mpack*. See (http://lagrange.math.siu.edu/Olive/tspack.txt) and (http://lagrange.math.siu.edu/Olive/mpack.txt). The function `pimasim` was used to simulate the time series prediction intervals. The functions `mpredsim` was used to simulate the prediction regions (25), `mregddsim` simulated the residual vector DD plots for various distributions, and the function `ddplot4` makes the DD plots.

REFERENCES

Box, G., Jenkins, G.M., and Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ.

Budny, K. (2014), "A Generalization of Chebyshev's Inequality for Hilbert-Space-Valued Random Variables," *Statistics & Probability Letters,* 88, 62-65.

Cai, T., Tian, L., Solomon, S.D., and Wei, L.J. (2008), "Predicting Future Responses Based on Possibly Misspecified Working Models," *Biometrika*, 95, 75-92.

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

Clements, M.P., and Kim, N. (2007), "Bootstrapping Prediction Intervals for Autoregressive Time Series," *Computational Statistics & Data Analysis*, 51, 3580-3594.

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

Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

Di Bucchianico, A., Einmahl, J.H.J., and Mushkudiani, N.A. (2001), "Smallest Nonparametric Tolerance Regions," *The Annals of Statistics*, 29, 1320-1343.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician,* 50, 120-126.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer, New York, NY.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

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

Kabaila, P., and He, Z. (2007), "Improved Prediction Limits for AR(p) and ARC(p) Processes," *Journal of Time Series Analysis*, 29, 213-223.

Lei, J., Robins, J., and Wasserman, L. (2013), "Distribution Free Prediction Sets," *Journal of the American Statistical Association*, 108, 278-287.

Lei, J., and Wasserman, L. (2014), "Distribution Free Prediction Bands," *Journal of the Royal Statistical Society, B*, 76, 71-96.

Masters, T. (1995), *Neural, Novel, & Hybrid Algorithms for Time Series Prediction*, Wiley, New York, NY.

Navarro, J. (2014), "Can the Bounds in the Multivariate Chebyshev Inequality be Attained?" *Statistics & Probability Letters*, 91, 1-5.

Navarro, J. (2015), "A Very Simple Proof of the Multivariate Chebyshev's Inequality," *Communications in Statistics: Theory and Methods,* to appear.

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

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

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

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

Olive, D.J. (2015), "Bootstrapping Hypotheses Tests," unpublished manuscript at (http://lagrange.math.siu.edu/Olive/ppvselboot.pdf).

Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics & Probability Letters,* 63, 259-266.

Olive, D. J., and Hawkins, D. M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).

Olive, D.J., Pelawa Watagoda, L.C.R., and Rupasinghe Arachchige Don, H.S. (2015), "Visualizing and Testing the Multivariate Linear Regression Model," *International Journal of Statistics and Probability*, 4, 126-137.

Pan, L., and Politis, D.N. (2015a), "Bootstrap Prediction Intervals for Linear, Nonlinear, and Nonparametric Autoregressions," *Journal of Statistical Planning and Inference,* (with discussion), to appear.

Pan, L., and Politis, D.N. (2015b), "Bootstrap Prediction Intervals for Markov Processes," Working Paper, Department of Economics, UCSD. URL: (https://escholarship.org/uc/item/7555757g).

Panichkitkosolkul, W., Niwitpong, S.-A. (2012), "Prediction Intervals for the Gaussian

Autoregressive Processes Following the Unit Root Tests," *Model Assisted Statistics and Applications*, 7, 1-15.

R Development Core Team (2011), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

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

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Thombs, L.A., and Schucany, W.R. (1990), "Bootstrap Prediction Intervals for Autoregression," *Journal of the American Statistical Association*, 85, 486-492.

Vidoni, P. (2009), "A Simple Procedure for Computing Improved Prediction Intervals for Autoregressive Models," *Journal of Time Series Analysis*, 30, 577-590.

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