Asymptotically Optimal Prediction Regions

David Olive
Southern Illinois University

I. Introduction

Prediction is done all of the time. Want to predict the weather, what investments will make money (for stocks about 45% of wealth gains are from dividends and 55% from the stock increasing in value), how large the national debt will be next year (if every home owner sold their house and gave the money to the government, would that pay off the debt?), et cetera. When President Clinton left office, the prediction was that there would be budget surpluses for the next ten years, instead there was record setting debt for the next 12 years. In Iowa City, Iowa, there have been two "100 year floods" in the past decade. If you are diagnosed with a cancer where 80% of the patients die of something else, then the standard treatment is probably effective, but if you are told that you have 6 to 18 months to live, the standard treatment is probably ineffective, and you may want to change the population by trying to get in a clinical trial for a new treatment.

Statistical prediction regions try to give a guarantee on the prediction. Suppose you are trying to predict a $k \times 1$ vector $\boldsymbol{y}_f$, for example, corn and soybean yield in Illinois for 2013. For a large sample 90% prediction region, there should be about a 90% chance that $\boldsymbol{y}_f$ lies in the prediction region. So if 100 data sets are independently gathered, about 90 times $\boldsymbol{y}_{f,j}$ should lie in the prediction region, and about 10 times $\boldsymbol{y}_{f,j}$ should fail to lie in the region. Hence a prediction region is a region of typical values of $\boldsymbol{y}_f$. Prediction regions are actually used, so the user gets mad at the Statistician when the prediction region fails. Unfortunately, statistical prediction regions with a nominal $100(1-\delta)\%$ coverage, e.g. 90%, typically have much smaller actual coverage. In Statistics, 90%, 95% and 99% nominal regions are often used, but several fields use 50% nominal regions since the high nominal coverage regions work so poorly. Prediction regions generally estimate percentiles of the underlying distribution, which is often assumed to be normal. The normal distribution is rarely a good approximation to the data since the normal distribution has small variability. Hence the estimated percentiles assuming normality tend to underestimate the true percentiles of the underlying distribution, and the true coverage is smaller than the nominal coverage.

Suppose a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ is available for predicting $\boldsymbol{y}_f$ and there is data $\boldsymbol{y}_i \equiv \boldsymbol{y}_i|\boldsymbol{x}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ where the zero mean $\boldsymbol{\epsilon}_i$ are independent and identically distributed (iid). Asymptotically optimal prediction regions can be derived for many additive error models of this form, where the distribution of the $\boldsymbol{\epsilon}_i$ is unknown but from a large class of distributions.

A large sample $(1-\delta)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1-\delta$, and is asymptotically optimal if the volume of the region converges in probability to the volume of the population minimum volume covering region. Want asymptotically optimal prediction regions that perform well for moderate sample size n but with few assumptions on the data distribution.

Statistics is the science of extracting useful information from data, and a statistical model is used to provide a useful approximation to some of the important characteristics of the population which generated the data. For a *parametric statistical model*, the

distribution of the data (and often that of the statistic) is known except for k unknown parameters. Want to avoid parametric (frequentist or Bayesian) models for prediction regions since parametric regions tend to perform poorly unless there is strong graphical evidence supporting the parametric model.

Example 1. Cook and Weisberg (1999, p. 351, 433, 447) give a data set on 82 mussels sampled off the coast of New Zealand. The variables are $y_1 = L, y_2 = \log(W), y_3 = H, y_4 = log(S)$ and $y_5 = \log(M)$ where L is the *length*, W is the *shell width*, H is the *height* of the shell in mm, S is the *shell mass*, and M is the *muscle mass* in grams. Might want to predict a) y_5 (location model), b) y_5 given $y_1, ..., y_4$ (regression model, often the multiple linear regression model), c) $\boldsymbol{y}$ (multivariate location and dispersion model), or d) y_4 and y_5 given y_1, y_2, and y_3 (multivariate linear regression model).

A case or observation consists of the p random variables measured for one person or thing. The ith case consists of the p measurements on the ith object (mussel) for $i = 1, ..., n$. The cases are collected into an $n \times p$ data matrix.

```
 L  log(W)   H  log(S) log(M)  p = 5
318 4.220  158  5.844  3.850  case 1
312 4.025  148  5.670  3.951  case 2
265 3.829  124  5.118  3.296  case 3
 .    .     .    .      .
 :    :     :    :      :
220 3.584  105  4.159  2.773  case 82=n
```

Things that can go wrong with using a training data set to predict a future data set:

i) the training set and future set may have different distributions, possibly due to a change in population. Population drift occurs when the population changes over time.

For example, suppose there are several variables being used to produce greater yield of a crop or a chemical. If one journal paper out of 50 (the training set) finds a set of variables and variable levels that successfully increases yield, then the next 25 papers (the future set) are more likely to use variables and variable levels similar to the one successful paper than variables and variable levels of the 49 papers that did not succeed. Also Reagan, Bush, Bush and Obama spent massive amounts of money to make the economy appear better, Clinton froze spending.

ii) Training set or future set could be distorted from the population if outliers (cases far from the bulk of the data) are present or if one data set is not a random sample from the population.

For example, the training data set could be drawn from three hospitals, and the future data set could be drawn from two more hospitals. These two data sets may not represent random samples from the same population of hospitals.

iii) Other model assumptions could be wrong (why multiple linear regression?, why iid normal errors?). Check model assumptions with response plots, residual plots, QQ plots and DD plots.

iv) The sample size n of the training set may not be large enough. Often want $n > 10p$.

II. Prediction Intervals and Regions for Additive Error Models

Asymptotically Optimal Prediction Intervals for the Location Model

The *location model* is $Y_i = \mu + e_i$, for $i = 1, ..., n$. Let $Y_1, ..., Y_n, Y_f$ be iid from the location model. In the location model there is one variable so the ith case is Y_i. Given $Y_1, ..., Y_n$, want an asymptotically optimal prediction interval (PI) $(\hat{L}_n, \hat{U}_n)$ for Y_f. Hence want $P(Y_f \in (\hat{L}_n, \hat{U}_n)) \to 1 - \delta$ and want the PI length $(\hat{U}_n - \hat{L}_n) \xrightarrow{P} D = U - L$ where D is as small as possible. U and L will be upper and lower percentiles of the unknown distribution of the Y_i. For asymptotically optimal prediction regions, we will assume that the distribution has a unimodal probability density distribution (pdf), and that the population PI is unique. If the distribution was symmetric, then the population 90% PI discards the top and bottom 5% of the distribution: L is the 0.05 percentile and U is the 0.95 percentile.

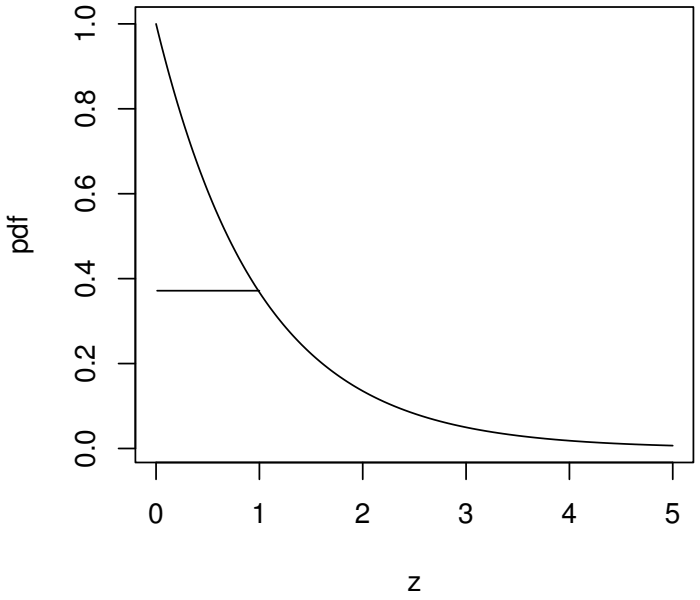

Figure 1: Highest 36.8% Density Region is (0,1)

Suppose $Y_1, ..., Y_n$ are iid from a unimodal pdf that has interval support, and that the pdf $f(y)$ decreases rapidly as y moves away from the mode. Let (a, b) be the shortest interval such that $F(b) - F(a) = 1 - \delta$ where the cumulative distribution function $F(y) = P(Y \leq y)$. Then the interval is the highest density region containing $1 - \delta$ of the mass. To find the $(1-\delta)100\%$ 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 $1 - \delta$. See Figure 1 where the area under the pdf from 0 to 1 gives the 36.8% highest density region. Will often have $f(a) = f(b)$, e.g., if the support where $f(y) > 0$ is $(-\infty, \infty)$.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. For the location model, consider intervals that contain c cases: $(Y_{(1)}, Y_{(c)})$, $(Y_{(2)}, Y_{(c+1)}), ..., (Y_{(n-c+1)}, Y_{(n)})$. Denote the set of c cases in the ith interval by J_i, for $i = 1, 2, ..., n - c + 1$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth$(c) = (Y_{(d)}, Y_{(d+c-1)})$ is the interval with the shortest length.

Example 2, votes for preseason 1A basketball poll Nov. 22, 2011 WSIL News:

```
111   89   778   78   76

get shorth(3)

order data 76 78 89 111 778

                 13 = 89 - 76

                     33 = 111 - 78

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

The correct value of 778 was 78 giving shorth(3) = (76,78).

The shorth$(c = \lceil n(1-\delta) \rceil)$ estimator estimates (a, b), and can be used as a $100(1-\delta)\%$ asymptotically optimal prediction interval for a future observation Y_f, but the PI has slight undercoverage for moderate sample sizes.

Olive (2013a) recommends the following asymptotically optimal PI. Let $a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+1}{n-1}}$. Let $c = \lceil n(1-\delta) \rceil$. Let shorth$(c) = (Y_{(d)}, Y_{(d+c-1)})$. Let MED$(n)$ be the sample median = sample $50th$ percentile. If $Y_1, ..., Y_n$ are iid, then the recommended large sample $100(1-\delta)\%$ PI for Y_f is the closed interval $[\hat{L}_n, \hat{U}_n] = [(1 - a_n)\text{MED}(n) + a_n Y_{(d)}, (1-a_n)\text{MED}(n) + a_n Y_{(d+c-1)}]$. This PI is a special case of the PI for multiple linear regression using the least absolute deviations estimator, but with a closed interval. The PI inflates the length of the shorth, reducing undercoverage. Plot the $Y's$ to examine the data and to check for outliers.

Asymptotically Optimal Prediction Intervals for the Multiple Linear Regression Model

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. Regression prediction intervals are for a future response Y_f given a vector $\boldsymbol{x}_f$ of predictors when the regression model has the form

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{1}$$

for $i = 1, ..., n$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid from a continuous unimodal distribution. Many of the most important regression models have this form, including the multiple linear regression model and many time series, nonlinear, nonparametric and semiparametric models (neural networks, partial least squares, vector support

machines, kriging, additive models, etc.). If $\hat{m}$ is an estimator of m, then the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$.

Also, conditional on $\boldsymbol{x}_i$, Y_i follows the location model with $\mu_i = m(\boldsymbol{x}_i)$. For example, if the e_i are iid $N(0, \sigma^2)$, then $Y_i|\boldsymbol{x}_i \sim N(m(\boldsymbol{x}_i), \sigma^2)$. If there was a lot of data at $\boldsymbol{x}_f$, then a location model PI could be used; however, often there is no data at $\boldsymbol{x}_f$. If m was known and there were n e_i's, then we could generate $Y_{f,j} = m(\boldsymbol{x}_f) + e_j$ for $j = 1, ..., n$ and use a location model PI on the $Y_{f,j}$. Since $\hat{m}$ estimates m and r_i estimates e_i, the basic idea is to use a location model type PI on a pseudosample $\tilde{Y}_{f,j} = \hat{m}(\boldsymbol{x}_f) + r_j$ for $j = 1, ..., n$. Since $\hat{m}$ is based on the past or training data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$, Y_f and $\hat{m}$ are independent. Hence, conditional on $\boldsymbol{x}_f$, the variance $V(Y_f - \hat{m}(\boldsymbol{x}_f)) = V(Y_f) + V(\hat{m}(\boldsymbol{x}_f)) = \sigma^2 + V(\hat{m}(\boldsymbol{x}_f))$ where $\sigma^2 = V(e_i)$ and $V(\hat{m}(\boldsymbol{x}_f)) \to 0$ as $n \to 0$.

Olive (2007) showed how to form asymptotically optimal prediction intervals for model (1), but for many regression models and estimators, large n is needed for the intervals to perform well. Prediction intervals derived for multiple linear regression did perform well. Olive (2013a) derives asymptotically optimal prediction intervals that perform well for many models for moderate n.

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 $n \to \infty$. Let ξ_δ be the δ percentile of the error e, i.e., $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. Consider predicting a future observation Y_f given a vector of predictors $\boldsymbol{x}_f$ where $(Y_f, \boldsymbol{x}_f)$ comes from the same population as the past data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let $1 - \delta_2 - \delta_1 = 1 - \delta$ with $0 < \delta < 1$ and $\delta_1 < 1 - \delta_2$ where $0 < \delta_i < 1$. Then $P[Y_f \in (m(\boldsymbol{x}_f) + \xi_{\delta_1}, m(\boldsymbol{x}_f) + \xi_{1-\delta_2})] = 1 - \delta$.

Assume that $\hat{m}$ is consistent: $\hat{m}(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ as $n \to \infty$. Then $r_i - e_i = Y_i - \hat{m}(\boldsymbol{x}_i) - (Y_i - m(\boldsymbol{x}_i)) \xrightarrow{P} 0$ and, under "mild" regularity conditions, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. If $a_n \xrightarrow{P} 1$ and $b_n \xrightarrow{P} 1$, then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + a_n \hat{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\delta_2}) \tag{2}$$

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

According to regression folklore, the percentiles of the residuals are consistent estimators, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$, under "mild" regularity conditions, and this consistency is the basis for using QQ plots. The folklore is true for linear models: sufficient conditions are $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and the $\boldsymbol{x}_i$ are bounded. See Olive and Hawkins (2003).

The multiple linear regression model is $Y_i \equiv Y_i|\boldsymbol{x}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the zero mean e_i are iid. Then $m(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. In matrix form, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown iid zero mean errors e_i with variance σ^2. Let 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 $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$.

If least squares is used and $e_i \sim N(0, \sigma^2)$, then the classical parametric $100(1-\delta)\%$ PI is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}\sqrt{MSE}\sqrt{(1 + h_f)} \tag{3}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom and MSE $= \hat{\sigma}$. If $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$, then for many distributions, MSE $\xrightarrow{P} \sigma^2$ and $V(\hat{Y}_f) \approx MSE \;\; h_f$. Also for iid $N(0, \sigma^2)$ errors, the pivotal quantity

$$T = \frac{Y_f - \hat{Y}_f}{\sqrt{MSE(1+h_f)}} \sim t_{n-p}.$$

Notice that the PI

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}\sqrt{MSE}\sqrt{(1+h_f)} =$$

$$\hat{Y}_f \pm z_{1-\delta/2}\sqrt{MSE}\frac{t_{n-p,1-\delta/2}}{z_{1-\delta/2}}\sqrt{(1+h_f)}.$$

Thus the quantity

$$a_n = b_n = \frac{t_{n-p,1-\delta/2}}{z_{1-\delta/2}}\sqrt{(1+h_f)}$$

can be regarded as a finite sample correction factor if $e_i \sim N(0, \sigma^2)$ where $P(Z \leq z_\delta) = \delta$ if $Z \sim N(0,1)$.

Let $1-\gamma$ be the asymptotic coverage of the classical nominal $(1-\delta)100\%$ PI (3). Then $1 - \gamma = P(-\sigma z_{1-\delta/2} < e < \sigma z_{1-\delta/2})$

$$\geq 1 - \frac{1}{z_{1-\delta/2}^2} \tag{4}$$

where the inequality follows from Chebyshev's inequality. For a 95% PI, $z_{1-\delta/2} \approx 2$, so actual coverage could be as low as 75%.

For the multiple linear regression model, let $\hat{\xi}_\delta$ be the sample percentile of the residuals. Following Olive (2007), let

$$a_n = b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{5}$$

Then a large sample semiparametric $100(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}). \tag{6}$$

PI (6) is very similar to PI (3) except $\hat{\xi}_\delta$ is used instead of $\hat{\sigma} z_\delta$ to estimate the error percentiles ξ_δ. A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. The PI (6) is asymptotically optimal on a large class of unimodal continuous symmetric error distributions. For more general distributions, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Following Olive (2007), a 100 $(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n\tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n\tilde{\xi}_{1-\delta_2}) \tag{7}$$

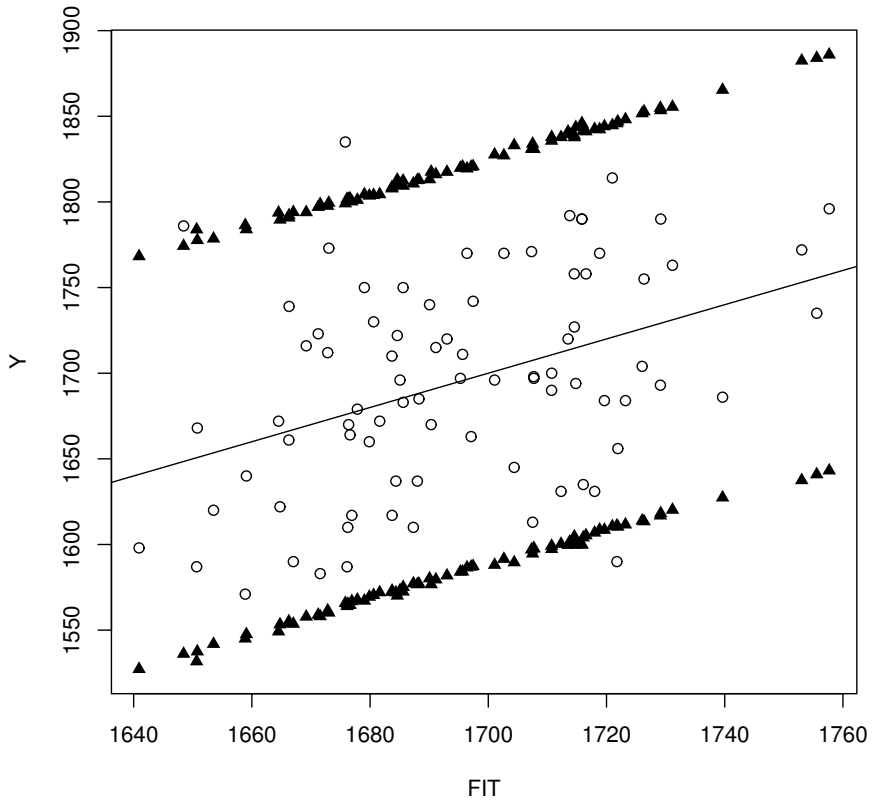

Figure 2: 95% PI Limits for Buxton Data

where a_n is given by (5). This prediction interval performs well for moderate n for multiple linear regression and several estimators, including least squares (several regression estimators have asymptotic variances that are of the same order as that of least squares).

Example 3. For the Buxton (1920) data suppose that the response $Y = height$ and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five outliers were deleted leaving 82 cases. Figure 2 shows a fit response plot of the fitted values versus the response Y with the identity line added as a visual aid. If the model is good then the plotted points should scatter about the identity line in an evenly populated band. The triangles represent the upper and lower limits of the semiparametric 95% PI (5). Notice that 79 (or 96%) of the Y_i fell within their corresponding PI while 3 Y_i did not.

Asymptotically Optimal Prediction Intervals for the Regression Model $Y = m(\boldsymbol{x}) + e$

A problem with prediction intervals is choosing a_n and b_n so that the intervals have short length and coverage close to or higher than the nominal coverage for a wide variety of regression models when n is moderate.

The idea for finding the asymptotically optimal prediction intervals and regions is simple. Find the target population $100(1-\delta)\%$ covering region. For small n, the coverage of the training data will be higher than that for the future case to be predicted. In simulations for a large group of models and distributions, the undercoverage could be as high as $\min(0.05, \delta/2)$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{8}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then use the prediction interval or region that covers $100q_n\%$ of the training data. The coverage of the training data is $100q_n\%$ and converges to $100(1 - \delta)\%$ as $n \to \infty$, even if the model assumptions fail to hold.

This technique is used to produce asymptotically optimal PIs that perform well for moderate samples. Find $\hat{Y}_f$ and the residuals from the regression model. Since the leverage of $\boldsymbol{x}_i$ is closely related to the Mahalanobis distance of $\boldsymbol{x}_i$ from the sample mean $\overline{\boldsymbol{x}}$ of the n predictor vectors, leverage and extrapolation are useful for a wide range of regression models. For a wide range of regression models, extrapolation occurs if $h_f > 2p/n$: if $\boldsymbol{x}_f$ is too far from the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then the model may not hold and prediction can be arbitrarily bad. This result suggests replacing (5) by

$$a_n = b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n + 2p}{n - p}}. \tag{9}$$

Let $\delta_n = 1 - q_n$ where q_n is given by (8). Then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{\delta_n/2}, \hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{1-\delta_n/2}) \tag{10}$$

is a large sample $100(1 - \delta)\%$ PI for Y_f that is similar to (2) and (6).

Let $c = \lceil nq_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Then the asymptotically

optimal 100 $(1-\delta)\%$ large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}), \tag{11}$$

and is similar to (7).

For asymptotic optimality, can not have extrapolation. Also, even if the coverage converges to the nominal coverage, the length of the PI need not be asymptotically shortest unless the highest $1-\delta$ density region of the probability density function of the iid errors is an interval. The highest density region is an interval for unimodal distributions, but need not be an interval for multimodal distributions for all δ.

Notice that the technique computes a PI for coverage $q_n \geq 1-\delta$ which converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% PI uses $q_n = 0.975$ while the nominal 50% PI uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $\hat{m}$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate the errors for small n. For small n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the length of the PI for small n, compensating for the unknown variability of $\hat{m}$.

The geometry of the "asymptotically optimal prediction region" is simple. The region is the area between two parallel lines with unit slope. Consider a plot of $m(\boldsymbol{x}_i)$ versus Y_i on the vertical axis. The identity line with zero intercept and unit slope is $E(Y_i) = m(\boldsymbol{x}_i)$. Let (L_i, U_i) be the asymptotically optimal population 95% prediction interval containing $m(\boldsymbol{x}_i)$. For example, if the errors are iid $N(0, \sigma^2)$, then $Y_i | m(\boldsymbol{x}_i) \sim N(m(\boldsymbol{x}_i), \sigma^2)$, and $(L_i, U_i) = (m(\boldsymbol{x}_i) - 1.96\sigma, m(\boldsymbol{x}_i) + 1.96\sigma)$. Then the upper line has unit slope and passes through $(m(\boldsymbol{x}_i), U_i)$ while the lower line has unit slope and passes through $(m(\boldsymbol{x}_i), L_i)$.

The geometry of the "prediction region" for PI (11) is a natural sample analog of the population "asymptotically optimal prediction region." A response plot of $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ versus Y_i has identity line $\hat{E}(Y_i) = \hat{m}(\boldsymbol{x}_i)$. The region corresponding to pointwise prediction intervals is between two lines with unit slope passing through the points $(\hat{m}(\boldsymbol{x}_i), \hat{U}_i)$ and $(\hat{m}(\boldsymbol{x}_i), \hat{L}_i)$, respectively, where $(\hat{L}_i, \hat{U}_i)$ is the asymptotically optimal prediction interval (9) for Y_f if $\boldsymbol{x}_f = \boldsymbol{x}_i$. For the multiple linear regression model, expect the points in the response plot to scatter in an evenly populated band for $n > 5p$. Other regression models, such as additive models, may need a much larger sample size n.

Example 4. Chambers and Hastie (1993, p. 251, 516) examine an environmental study that measured the four variables $Y =$ *ozone concentration*, $x_1 =$ *solar radiation*, $x_2 =$ *temperature*, and $x_3 =$ *wind speed* for $n = 111$ consecutive days. Figure 3 shows the response plot made in *Splus* with the pointwise large sample 95% PI bands for the additive model $Y = m(\boldsymbol{x}) + e$ where the additive predictor $m(\boldsymbol{x}) = \alpha + \sum_{j=1}^{3} S_j(x_j)$ for some functions S_j to be estimated. Here $\hat{m}(\boldsymbol{x}) =$ estimated additive predictor (EAP). Note that the plotted points scatter about the identity line in a roughly evenly populated band, and that 3 of the 111 PIs (11) corresponding to the observed data do not contain Y.

The "theory section" shows that PI (11) and the shorth of the residuals behave well when the sample percentiles are consistent. Even if these assumptions do not hold, the

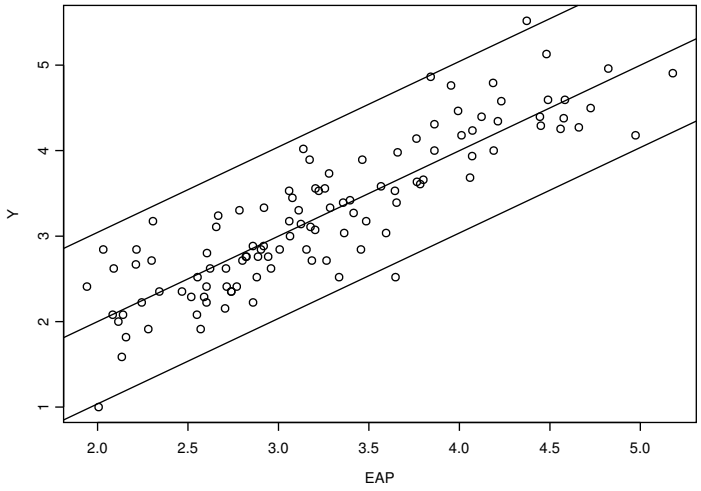

Figure 3: Pointwise Prediction Interval Bands for Ozone Data

PI covers $100q_n\%$ of the training data, and often the coverage of the future case will be close to $100(1-\delta)$ if the future case Y_f is similar to the training data.

Asymptotically Optimal Prediction Regions for the Multivariate Location and Dispersion Model

Asymptotically optimal prediction regions use ideas similar to those in the previous subsection. Some notation is needed. Let the ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}. \tag{12}$$

Some important joint distributions for $\boldsymbol{x}$ are completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important model is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function $f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})]$ 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.

Let the $p \times 1$ column vector $T(\boldsymbol{X})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{X})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar $D_i^2 = D_i^2(T(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) =$

$$(\boldsymbol{x}_i - T(\boldsymbol{X}))^T \boldsymbol{C}^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{13}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{X})$ is $D_i(T(\boldsymbol{X}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Often the data $\boldsymbol{X}$ will

be suppressed. Then the classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Following Johnson (1987, p. 107-108), the population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{14}$$

and for elliptically contoured distributions, U has probability density function (pdf)

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{15}$$

The volume of the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq h^2\}$ is equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{S})}, \tag{16}$$

see Johnson and Wichern (1988, p. 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 $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of

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

For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

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

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

$$= \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \tag{17}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)} \tag{18}$$

by (16). A future observation (random vector) $\boldsymbol{x}_f$ is in region (17) if $D_{\boldsymbol{x}_f} \leq h$.

The Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ is an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ under regularity conditions (E1) that include a large class of elliptically contoured distributions, and $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Also see Zhang, Olive and Ye (2012). The RMVN estimator also gives a useful estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data even when certain types of outliers are present.

A large sample $(1-\delta)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1-\delta$. Let q_n be given by (8). If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (17) 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. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then region (17) is asymptotically optimal on a large class of elliptically contoured distributions in that its volume converges in probability to the volume of the minimum volume covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq u_{1-\delta}\}$ where $P(U \leq u_{1-\delta}) = 1 - \delta$ and U has pdf given by (15). The classical parametric multivariate normal large sample prediction region uses $D_{\boldsymbol{x}_f}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \equiv MD_{\boldsymbol{x}_f} \leq \sqrt{\chi^2_{p,1-\delta}}$. Olive (2013a) gives three new prediction regions. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(up)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(up)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,q_n}) = q_n$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions with nonsingular $\boldsymbol{\Sigma}$. The first two regions are asymptotically optimal for a large class of elliptically contoured distributions. For distributions with nonsingular covariance matrix $c_X \boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region, but regions with smaller volume may exist.

Notice that for the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $100q_n\%$ of the n cases are in the prediction region, and $q_n \to 1 - \delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1 - \delta + 0.05$ for $n \leq 20p$ and $q_n \to 1 - \delta$ as $n \to \infty$. If $q_n \equiv 1 - \delta$, then (17) is a large sample prediction region, but taking q_n given by (8) improves the finite sample performance of the region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for moderate n the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (8) helped reduce undercoverage for moderate n due to the unknown variability of $(T, \boldsymbol{C})$.

Rousseeuw and Van Driessen (1999) introduce the DD plot of the classical Mahalanobis distances MD versus the robust distances RD. Olive (2002) shows that if consistent estimators are used and n is large, then the plotted points will follow the identity line with unit slope and zero intercept if the data distribution is multivariate normal, and the plotted points will follow some other line through the origin if the data distribution is from a large class of elliptically contoured distributions but not multivariate normal.

Figure 4 was made with the *Arc* software of Cook and Weisberg (1999). The 10%, 30%, 50%, 70%, 90% and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 4a),

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

Note that the ellipsoids are narrow with high positive correlation. In Figure 4b),

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

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.

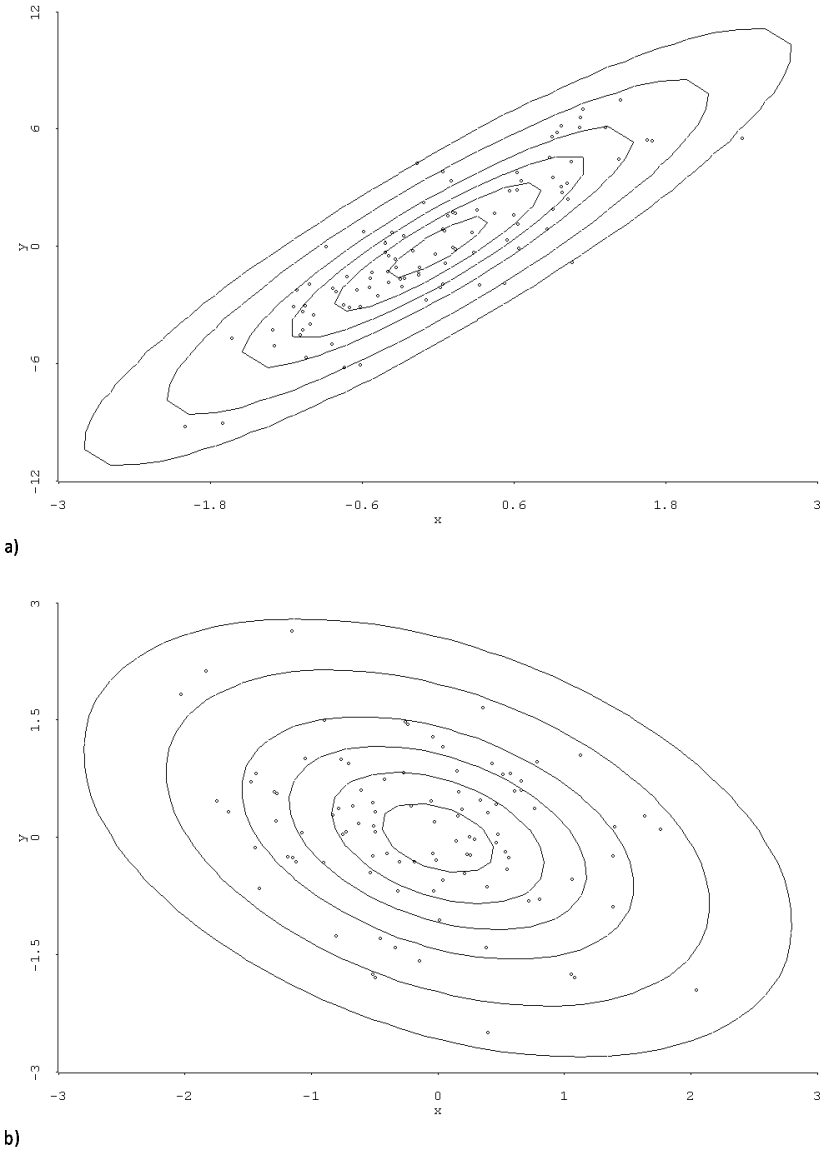

Figure 4: Highest Density Regions for 2 MVN Distributions

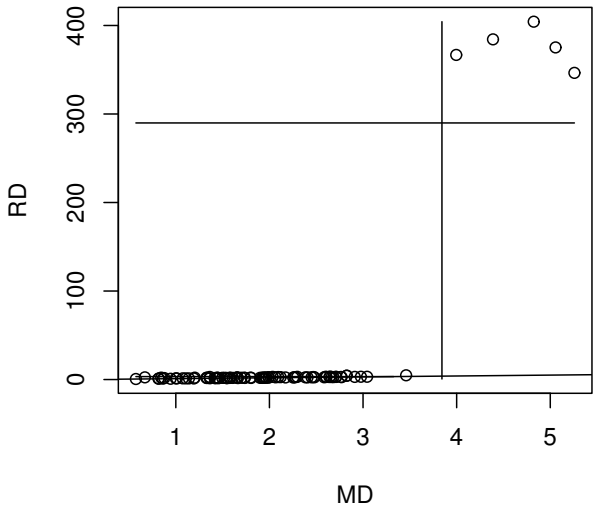

Figure 5: Prediction Regions for Buxton Data

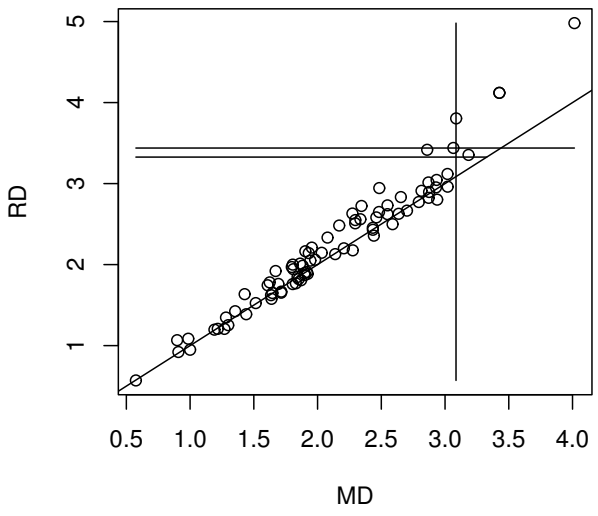

Figure 6: Prediction Regions for Buxton Data without Outliers

Ex. 5. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm and are massive outliers. All 87 cases and 5 predictors were used. Figure 5 shows the RMVN DD plot with the identity line added as a visual aid. Points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The semiparametric 90% region blows up unless the outlier proportion is small.

Figure 6 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff again at 3.33, slightly below the semiparametric region cutoff of 3.44.

Example 6. Consider the mussel data set from Example 1 with $p = 5$ and $n = 82$. Figure 5 shows a DD plot of the data with multivariate prediction regions added. This plot suggests that the data may come from an elliptically contoured distribution that is not multivariate normal. The semiparametric and nonparametric 90% prediction regions consist of the cases below the $RD = 5.86$ line and to the left of the $MD = 4.41$ line. These two lines intersect on a line through the origin that is followed by the plotted points. The parametric MVN prediction region is given by the points below the $RD = 3.33$ line and does not contain enough cases. Points to the left of a vertical line $MD = 3.33$ would give a modified classical MVN prediction region. Parametric prediction regions for multivariate normal data tend to have severe undercoverage if the data is not multivariate normal. This undercoverage problem becomes worse as p increases, since if the cutoff h is too small, then the volume of the prediction region depends on h^p by (16).

Asymptotically Optimal Prediction Regions for the Multivariate Linear Regression Model

The *multivariate linear regression model* $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the constant $x_{i1} = 1$ could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\mathrm{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \ldots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

If $\boldsymbol{v}_1 = \boldsymbol{1}$, the $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

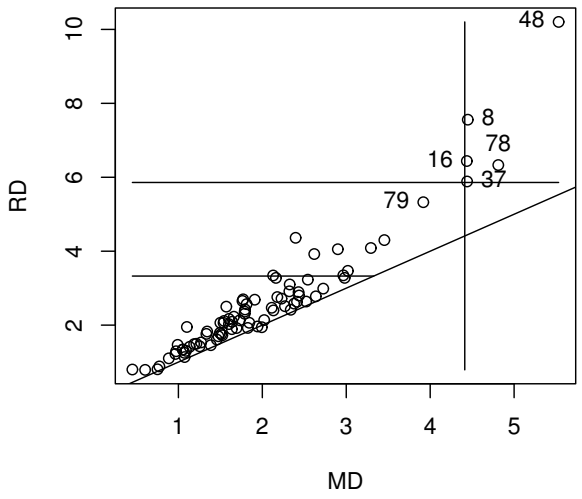

Figure 7: DD Plot of the Mussels Data.

The $p \times m$ matrix $\boldsymbol{B} = \left[\begin{array}{cccc} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{array} \right]$.
The $n \times m$ matrix

$$\boldsymbol{E} = \left[\begin{array}{cccc} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \ldots & \boldsymbol{e}_m \end{array} \right] = \left[\begin{array}{c} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{array} \right].$$

Least squares is the classical method for fitting multivariate linear regression. The *least squares estimators* are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\begin{array}{cccc} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \ldots & \hat{\boldsymbol{\beta}}_m \end{array} \right]$. The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{cccc} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{array} \right].$$

The *residuals*

$$\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{c} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{array} \right] = \left[\begin{array}{cccc} \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \ldots & \hat{\boldsymbol{r}}_m \end{array} \right].$$

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 $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ 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 $\mathbf{0}$. Let $\hat{\boldsymbol{\Sigma}}_\epsilon = \hat{\boldsymbol{\Sigma}}_{\epsilon,p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid.

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could make a prediction region for $\boldsymbol{z}_i$ using one of the three multivariate location and dispersion model prediction regions with p replaced by m. Instead, use the nonparametric region on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T\boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f - (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. The theory section will show that the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ have the same quantiles, asymptotically.

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and continuous decreasing g, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\delta}) = 1 - \delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

The "theory section" shows that applying the nonparametric prediction region on the $\hat{\boldsymbol{z}}_i$ results in a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ given the vector of predictors $\boldsymbol{x}_f$. The prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid from an $EC_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, g)$ distribution for a large class of elliptically contoured distributions.

The nonparametic region uses the sample mean and sample covariance matrix $(T, \boldsymbol{C})$ of the $\hat{\boldsymbol{z}}_i$. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{y} : (\boldsymbol{y} - T)^T\boldsymbol{C}^{-1}(\boldsymbol{y} - T) \le h^2\} =$$

$$\{\boldsymbol{y} : D^2_{\boldsymbol{y}} \le h^2\} = \{\boldsymbol{y} : D_{\boldsymbol{y}} \le h\}. \tag{19}$$

A future observation (random vector) $\boldsymbol{y}_f$ is in the region (19) if $D_{\boldsymbol{y}_f} \le h$.

Since the least squares residuals have sample mean $\mathbf{0}$, $\boldsymbol{S}_r$ is the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and of the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and the sample mean of the $\hat{\boldsymbol{z}}_i$ is $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ using the $\hat{\boldsymbol{z}}_i$ are used to compute $D_{(up)}$. Set up the nonparametric prediction region (19) using $h = D_{(up)}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ using m instead of p.

The nonparametic prediction region has some interesting properties. If there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, only need to update $\hat{\boldsymbol{y}}_{jf}$ for $j = 1, ..., 100$, do not need to update the covariance matrix $\boldsymbol{S}_r$.

The geometry of the nonparametric region is simple. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$, and let (19) be the nonparametric prediction region using $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ when the multivariate regression is fit by least squares.

Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region (19) is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$.

It is common practice to examine how well the prediction regions work on the data. That is, for $i = 1, ..., n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \delta$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (19) for the data since the prediction region (19) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i = \hat{\boldsymbol{\epsilon}}_i$. So $D_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \boldsymbol{S}_r) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \boldsymbol{S}_r^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = (\hat{\boldsymbol{\epsilon}}_i - \mathbf{0})^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{\epsilon}}_i - \mathbf{0}) = D_{\hat{\boldsymbol{\epsilon}}_i}(\mathbf{0}, \boldsymbol{S}_r)$. Thus $100q_n\%$ of prediction regions corresponding to the training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \rightarrow 100(1 - \delta)\%$. Hence the prediction regions work well on the training data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the training data. Of course simulation should be done for $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to training data cases.

This result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n training data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is very robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response, residual and DD plots. Coverage can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

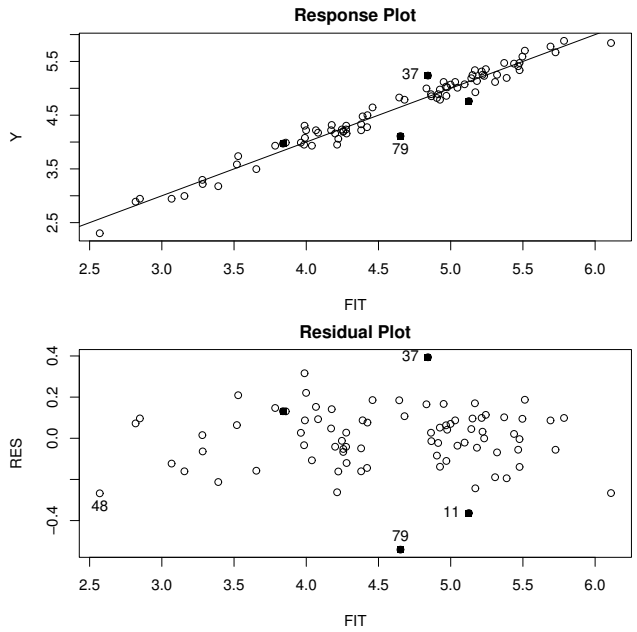

Figure 8: Plots for $Y_1 = \log(W)$.

Example 7. Consider the Cook and Weisberg (1999) mussels data with $Y_1 = \log(S)$ 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, width and height. (Example 6 and Figure 5 discussed the multivariate prediction regions.) Figures 8 and 9 give the response and residual plots for Y_1 and Y_2. For Y_2, cases 8, 25 and 48 are not fit well. A

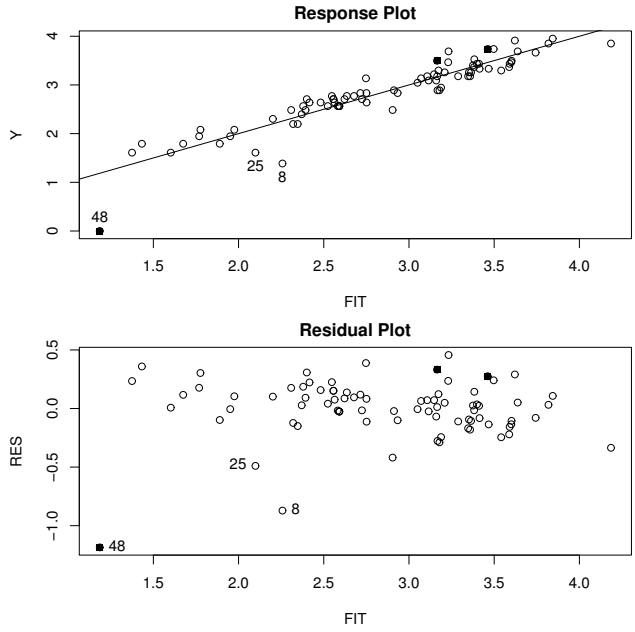

Figure 9: Plots for $Y_2 = \log(M)$.

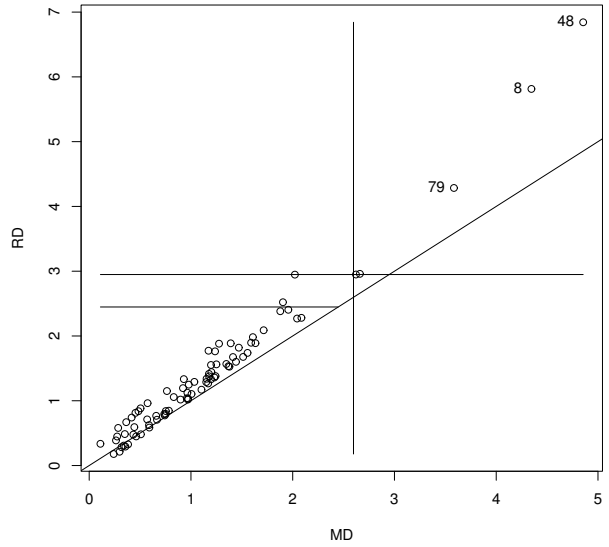

Figure 10: DD Plot of the Residual Vectors.

residual vector $\boldsymbol{r} = (\boldsymbol{r} - \boldsymbol{e}) + \boldsymbol{e}$ is a combination of $\boldsymbol{e}$ and a discrepancy $\boldsymbol{r} - \boldsymbol{e}$ that tends to have an approximate multivariate normal distribution. The $\boldsymbol{r} - \boldsymbol{e}$ term can dominate for small to moderate n when $\boldsymbol{e}$ is not multivariate normal, incorrectly suggesting that the distribution of the error $\boldsymbol{e}$ is closer to a multivariate normal distribution than is actually the case. Figure 10 shows the DD plot of the residual vectors. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Comparing Figures 5 and 10, the residual distribution is closer to a multivariate normal distribution. Cases 8, 48 and 79 have especially large distances. Note that cases to the right of the vertical line correspond to cases that are not in their prediction region. Also adding a constant does not change the distance, so the DD plot for the residuals is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

III. Theory

Location Model:

Grübel (1988) shows that, under regularity conditions, the length of the shorth converges in probability to D at a $\sqrt{n}$ rate but the midpoint of the shorth converges in probability to the midpoint of the population shorth at an $n^{1/3}$ rate. Since the location model is a special case of the regression models, the results for regression suggest that a PI based on the shorth should work well, asymptotically, if the sample percentiles are consistent estimators of the population percentiles and the distribution is continuous.

For a discrete distribution, the asymptotically optimal region will be a set of points rather than an interval. The shorth should produce short intervals, but odd behavior can occur. For example, if $P(Y = 0) = 0.1, P(Y = 1) = 0.8$ and $P(Y = 7) = 0.1$, then the shortest population 90% PI is [0,1], but the sample shorth covering 90% of the cases will use (0,1), (1,7) and (0,7).

Multiple Linear Regression:

Assume that the predictors are bounded. Hence $\|\boldsymbol{x}\| \leq M$ for some constant M. Let $0 < \gamma < 1$, and let $0 < \epsilon < 1$. Since $\hat{\boldsymbol{\beta}}_n$ is consistent, there exists an N such that $P(A) =$

$$P(\hat{\beta}_{j,n} \in [\beta_j - \frac{\epsilon}{4pM}, \beta_j + \frac{\epsilon}{4pM}], j = 1, ..., p)$$

$$\geq 1 - \gamma$$

for all $n \geq N$. If $n \geq N$, then on set A,

$$\max_{i=1,...,n} |r_i - e_i| = \max_{i=1,...,n} |\sum_{i=1}^{p} x_{i,j}(\beta_j - \hat{\beta}_{j,n})| \leq \frac{\epsilon}{2}.$$

Since ϵ and γ are arbitrary,

$$\max_{i=1,...,n} |r_i - e_i| \xrightarrow{P} 0.$$

Hence the sample percentiles of the residuals are consistent estimators of the population percentiles of the population distribution of the $\boldsymbol{e}_i$. The following result suggests that the shorth of the residuals will perform well.

Regression where $Y = m(\boldsymbol{x}) + e$:

To see that the PI (11) is asymptotically optimal, assume that the sample percentiles of the residuals converge to the population percentiles of the iid unimodal errors: $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$.

Also assume that the population shorth $(\xi_{\delta_1}, \xi_{1-\delta_2})$ is unique and has length L. Since $b_n \to 1$, $\hat{m}(\boldsymbol{x}_f) \xrightarrow{P} m(\boldsymbol{x}_f)$, and $q_n = 1-\delta$ for large enough n, it is enough to show that the shorth of the residuals converges to the population shorth of the e_i: $(\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}) \xrightarrow{P} (\xi_{\delta_1}, \xi_{1-\delta_2})$. Let L_n be the length of the shorth $(\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$. Since $r_i - e_i \xrightarrow{P} 0$, and since consistent sample percentiles imply that the proportion of r_i that do not get arbitrarily close to the e_i goes to 0, the shorth is an interval that asymptotically covers $100(1-\delta)\%$ of the cases (e_i) and the population shorth is the shortest such interval. Hence $P(L_n < L) \to 0$. But $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \leq \hat{\xi}_{1-\delta_2} - \hat{\xi}_{\delta_1} \xrightarrow{P} L$ as $n \to \infty$ since the sample percentiles are consistent and the shorth is the smallest sample interval covering 100 $(1-\delta)\%$ of the data (define the sample percentiles to be order statistics, e.g. $\hat{\xi}_{\delta_1} = r_{(\lceil n\delta_1 \rceil)}$, and $c = c_n \approx n(1-\delta)$ so that c_n is the number of residuals in $[\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2}]$). Hence $L_n \xrightarrow{P} L$, and since the population shorth is unique, the shorth of the residuals converges in probability to the population shorth (can't have two intervals of length L that asymptotically cover $100(1-\delta)\%$ of the data).

Multivariate Location and Dispersion:

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (17) 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. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then region (17) is asymptotically optimal on a large class of elliptically contoured distributions in that its volume converges in probability to the volume of the minimum volume covering region $\{\boldsymbol{z} : (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \leq u_{1-\delta}\}$ where $P(U \leq u_{1-\delta}) = 1-\delta$ and U has pdf given by (15).

$\boldsymbol{T}_n$ converges in probability to $\boldsymbol{\theta}$, written $\boldsymbol{T}_n \xrightarrow{P} \boldsymbol{\theta}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{T}_n - \boldsymbol{\theta}\| > \epsilon) \to 0$ as $n \to \infty$. $\boldsymbol{T}_n$ is a consistent estimator of $\boldsymbol{\theta}$ if $\boldsymbol{T}_n \xrightarrow{P} \boldsymbol{\theta}$ for every $\boldsymbol{\theta} \in \Theta$. $\boldsymbol{T}_n$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\theta}$ if $\sqrt{n}\|\boldsymbol{T}_n - \boldsymbol{\theta}\|$ is bounded in probability. Then $n^{0.499}\|\boldsymbol{T}_n - \boldsymbol{\theta}\| \xrightarrow{P} 0$.

Collect the n cases into a data matrix $\boldsymbol{W}$ with ith row $\boldsymbol{x}_i^T$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$. Then the multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{20}$$

and

$$\boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{21}$$

The assumption below gives the class of distributions for which RMVN has been shown to be $\sqrt{n}$ consistent. Distributions where the minimum covariance determinant (MCD) functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Theorem 1, Lopuhaä (1999). Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant

a depends on the positive constants s, h^2, p and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying the above theorem iteratively for a fixed number k of steps produces a sequence of estimators
$(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j \boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, p, h and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{22}$$

be the population median of the population squared distances. Olive and Hawkins (2010) show that if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is given by Equation (22), and that $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for each j. Then Olive and Hawkins (2010) show that the DGK and MCD estimators are estimating the same quantity where the DGK estimator uses the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. See Devlin, Gnanadesikan and Kettenring (1981). Note that the DGK estimator is practical to compute but has a much lower breakdown value than the impractical MCD estimator.

Theorem 2, Olive and Hawkins (2010). Assume (E1) holds. a) Then the DGK estimator and MCD estimator are $\sqrt{n}$ consistent affine equivariant estimators of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

b) The FCH, RFCH and RMVN estimators are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i\boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for multivariate normal data. If the clean data are in general position, then T_{FCH} is a high breakdown estimator and $\boldsymbol{C}_{FCH}$ is nonsingular even if nearly half of the cases are outliers.

Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

This is the fastest estimator of multivariate location and dispersion that has been shown to be both consistent and high breakdown with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$. See Bernholt and Fischer (2004).

For nearly 20 years, the elemental basic resampling algorithm was the main way for computing "high breakdown multivariate robust estimators." Randomly select h cases and compute the classical estimator $(T_i, \boldsymbol{C}_i)$ (or $T_i = \hat{\boldsymbol{\beta}}_i$ for MLR) for these cases. Here $h = p$ for multiple linear regression and $h = p+1$ for multivariate location and dispersion. The estimator uses K elemental sets as trial fits.

Theorem 3: The elemental basic resampling algorithm estimators are inconsistent and zero breakdown.

Proof: Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency. Change each of the Kh cases to $d\mathbf{1}$ where $\mathbf{1}$ is a vector of ones and d is some constant. Then the classical OLS estimator can not be computed and the classical dispersion estimator is singular with rank 1 if $d \neq 0$. Hence the breakdown value is bounded by $Kh/n \to 0$. QED

The Rousseeuw Yohai paradigm for high breakdown multivariate robust statistics is to approximate an impractical brand name estimator by computing a fixed number of easily computed trial fits and then use the brand name estimator criterion to select the trial fit to be used in the final robust estimator. The resulting estimator will be called an F-brand name estimator where the F indicates that a fixed number of trial fits was used. For example, generate 500 easily computed estimators of multivariate location and dispersion as trial fits. Then choose the trial fit with the dispersion estimator that has the smallest determinant. Since the minimum covariance determinant (MCD) criterion is used, call the resulting estimator the FMCD estimator. Hubert, Rousseeuw, and Verdonck (2012) claim to compute MCD with two FMCD estimators Fast-MCD and Det-MCD, but the claim is false because sometimes Fast-MCD has the smallest determinant and sometimes Det-MCD does. If both were finding the half set corresponding to the sample covariance matrix with the smallest determinant, then the two determinants would always be equal.

Multivariate Linear Regression:

Theorem 4, (Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of the multivariate linear model holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, ..., p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \mathbf{0}$ and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Su and Cook (2012) show that if the iid errors have 4th moments, $\max h_i \to 0$ and $\boldsymbol{X}^T\boldsymbol{X}/n \xrightarrow{P} \boldsymbol{C}^{-1}$, then $\hat{\boldsymbol{B}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent and asymptotically normal. The following result is also useful.

Theorem 5. $\boldsymbol{S}_r = \boldsymbol{\Sigma}_\epsilon + O_P(n^{-1/2})$ if $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$,

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{-1/2})$$

and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$.

Proof. Note that $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\epsilon}}_i$. Hence $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{B} - \hat{\boldsymbol{B}})^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Thus

$$\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \sum_{i=1}^n (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)^T =$$

$$\sum_{i=1}^n [\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i(\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)^T + (\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)\hat{\boldsymbol{\epsilon}}_i^T] =$$

$$\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T + (\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}})$$

$$+(\boldsymbol{B} - \hat{\boldsymbol{B}})^T (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\epsilon}_i^T) +$$

$$(\boldsymbol{B} - \hat{\boldsymbol{B}})^T (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}).$$

Thus $\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T +$

$$O_P(1)O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(1) +$$

$$O_P(n^{-1/2})O_P(n^{1/2})O_P(n^{-1/2}),$$

and the result follows since $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T.$$

The following technical theorem will be needed to prove Theorem 7.

Theorem 6. 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 $n \to \infty$.

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

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

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

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

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

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

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

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

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

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \tag{23}$$

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

Theorem 7. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i)+\boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i+\hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. 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 $q_n = \min(1-\delta+0.05, 1-\delta+m/n)$ for $\delta > 0.1$ and

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

If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. Consider the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$

$$\{\boldsymbol{z} : (\boldsymbol{z}-\hat{\boldsymbol{y}}_f)^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{z}-\hat{\boldsymbol{y}}_f) \le D_{(U_n)}^2\} =$$

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

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}\}. \tag{24}$$

a) Consider the n prediction regions for the 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 (24) 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 (24) is asymptotically optimal.

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

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

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

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 $(1-\delta)$ percentile of the D_i asymptotically, $U_n/n \to 1-\delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1-\delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 6 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1-\delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$.

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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). QED

IV. Simulations Data and programs are in the collections of functions *rpack* and *mpack* are available at (www.math.siu.edu/olive/ol-bookp.htm) and (www.math.siu.edu/olive/multbk.htm).

Location Model Use *rpack* function *lpisim*.

Multiple Linear Regression See Olive (2007) and *rpack* function *pisim*.

Regression where $Y = m(\boldsymbol{x}) + e$

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for PIs (10) and (11). Values for PI (10) were denoted by scov and slen while values for PI (11) were denoted by ocov and olen. The five error distributions in the simulation were 1) N(0,1), 2) t_3, 3) exponential(1) -1, 4) uniform$(-1, 1)$ and 5) $0.9N(0,1) + 0.1N(0,100)$. The value $n = \infty$ gives the asymptotic coverages and lengths and does not depend on the model. So these values are same for multiple linear and nonlinear regression as well as additive models.

The regression model was $Y_i = m(\boldsymbol{x}_i) + e_i$, $E(Y_i) = m(\boldsymbol{x}_i) = \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i3} + \beta_6 x_{i3}^2$. This model was fit as an additive model in x_1, x_2, and x_3 with additive predictor $m(\boldsymbol{x}_i) = \alpha + \sum_{j=1}^{3} S_j(x_{ij})$ where the S_j are unknown. The additive model had mean function $m(\boldsymbol{x}_i) = x_{i1} + x_{i1}^2$. Thus $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0)^T$, $\alpha = 0$, $S_1(x_{i1}) = x_{i1} + x_{i1}^2$, $S_2(x_{i2}) = 0$ and $S_3(x_{i3}) = 0$. For this model, the vectors $(x_1, x_2, x_3)^T$ were iid $N_3(\boldsymbol{0}, \boldsymbol{I}_3)$.

The PIs for the additive model were computed using the R function `gam`. See Hastie and Tibshirani (1990) and Wood (2006). The PI (10) is not asymptotically optimal with error type 3. It is not known whether $\hat{m}$ is a consistent estimator of m, but the prediction intervals appear to have the correct asymptotic coverage and length. Some consistency results for the additive model and models of the form $Y = m(\boldsymbol{x}) + e$ where m is smooth are given in Müller, Schick and Wefelmeyer (2012) and Wang, Liu, Liang, and Carroll (2011).

The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\delta)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1 - \tau_n)$ distribution where $1 - \tau_n$ converges to the asymptotic coverage $(1 - \tau)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0031$ and 0.0071 for $p = 0.05$ and 0.5, respectively. Hence an observed coverage $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.479, .521)$ for 50% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Table 1 shows that for $n = 1000$, the coverages and lengths are near the asymptotic $n = \infty$ values. For the 95% PI (11), the coverages were in or near (.94,.96) while the

Table 1: PIs for Additive Models

error type	n	95% slen	PI olen	95% scov	PI ocov	50% slen	PI olen	50% scov	PI ocov
1	50	5.126	4.998	0.959	0.950	1.862	1.674	0.596	0.520
1	100	4.691	4.515	0.968	0.957	1.662	1.528	0.570	0.516
1	1000	3.994	3.944	0.954	0.949	1.379	1.351	0.514	0.505
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.444	8.630	0.951	0.943	2.385	2.153	0.576	0.512
2	100	8.245	7.596	0.962	0.954	2.042	1.878	0.577	0.532
2	1000	6.523	6.388	0.950	0.946	1.584	1.553	0.499	0.489
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.186	4.823	0.958	0.948	1.573	1.275	0.611	0.526
3	100	4.677	4.156	0.967	0.955	1.382	1.063	0.603	0.533
3	1000	3.771	3.227	0.954	0.952	1.112	0.774	0.509	0.512
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	2.634	2.598	0.961	0.958	1.237	1.087	0.593	0.506
4	100	2.318	2.272	0.972	0.968	1.155	1.028	0.561	0.480
4	1000	1.936	1.926	0.959	0.954	1.014	0.969	0.499	0.486
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.689	17.747	0.944	0.935	2.976	2.693	0.608	0.548
5	100	18.754	16.230	0.955	0.946	2.352	2.164	0.580	0.534
5	1000	13.855	12.930	0.946	0.943	1.602	1.569	0.510	0.504
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

50% PI (9) was sometimes slightly conservative. The coverage for the 50% PI (10) was near 60% for $n = 50$. PI (11) is recommended since its asymptotic optimality does not depend on the symmetry of the error distribution.

Multivariate Location and Dispersion

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, \sqrt{2}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\delta = 0.1$.

Table 2: Coverages for 90% Prediction Regions

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

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

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

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

Simulations and Table 2 suggest that for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, the coverages (ncov, scov and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than 5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

Multivariate Linear Regression

See *mpack* function *mpredsim*.

Summary

These prediction regions cover $100q_n\%$ of the training data where $q_n \to 1 - \delta$ as $n \to \infty$. For multiple linear regression and multivariate location and dispersion, need $n > 10p$ so that the estimators start giving good estimates. In the multivariate location

and dispersion model, data is very sparse as p increases. So for $p+1 \leq n < 10p$ could have a low volume covering hyperellipsoid that does not approximate the population covering hyperellipsoid. Then the prediction region could have serious undercoverage.

The prediction region for the multivariate model given in Theorem 7 and (24) can be used for many multivariate regression models and estimators, including the Su and Cook (2012) estimators for the multivariate linear regression model, seemingly unrelated regression, and some multivariate time series models. The model is $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. When OLS is used in multivariate linear regression with $\boldsymbol{S}_r$, then the prediction region becomes the nonparametric prediction region applied to the residuals. In general this result will not hold since the sample mean of the residuals will usually not be $\mathbf{0}$.

What to plot for regression: if $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$ where the real valued function $h : \mathcal{R}^p \to \mathcal{R}$, make a *response plot* of $\hat{h}(\boldsymbol{x})$ versus Y, and a residual plot of $\hat{h}(\boldsymbol{x})$ versus r. See Olive (2013b)

What to plot for multivariate location and dispersion: make a DD plot.

What to plot for multivariate regression: make a response and residual plot for each of the m response variables and make a DD plot of the residuals. See Olive (2013c).

There is not much competition for these regions. The PI $\hat{m}(\boldsymbol{x}) \pm z_{1-\delta}\sqrt{\hat{\sigma}^2 + \hat{V}(\hat{m})}$ needs normality, a consistent estimator $\hat{m}$ and an estimate $\hat{V}(\hat{m})$ of the variance of $\hat{m}$ that goes to 0 as $n \to \infty$. Bootstrap regions tend to take too long to compute and are not backed by theory. An interesting idea is to estimate the pdf of the data, then use the pdf to find small prediction regions. The problem with these regions is that nonparametric pdf estimators do not work well for $p > 4$. See Lei, Robins and Wasserman (2011), Lei and Wasserman (2012), and Vovk, Nouretdinov and Gammerman (2009).

Future Work a) Univariate and Multivariate Time Series. b) Prediction regions for $\boldsymbol{y}_1, ..., \boldsymbol{y}_k$ such that the probability that $\boldsymbol{y}_i \in \mathcal{A}_i$ for all k regions $\to 1-\delta$. So if $1-\delta = 0.9$, gather 100 data sets. In about 90 data sets, all k of the $\boldsymbol{y}_i \in \mathcal{A}_i$ but in about 10 data sets, at least one $\boldsymbol{y}_j$ is not in $\mathcal{A}_j$. c) Prediction regions so that at least m of k $\boldsymbol{y}_j$ are in the $\mathcal{A}_j$.

References

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Chambers, J.M., and Hastie, T.J. (editors) (1993), *Statistical Models in S*, Chapman & Hall, New York, NY.

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

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

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

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall, London, UK.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

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.

Lei, J., Robins, J., Wasserman, L. (2011), "Efficient Nonparametric Conformal Prediction Regions," see
(http://arxiv.org/pdf/1111.1418.pdf).

Lei, J., and Wasserman, L. (2012), "Distribution Free Prediction Bands," see
(http://arxiv.org/pdf/1203.5422.pdf).

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Müller, U. U., Schick, A., and and Wefelmeyer, W. (2012), "Estimating the Error Distribution Function in Semiparametric Additive Regression Models," *Journal of Statistical Planning and Inference,* 142, 552-566.

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 and Data Analysis,* 51, 3115-3122.

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

Olive, D.J. (2013b), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 41, to appear.

Olive, D.J. (2013c), "Plots, Prediction and Testing for the Multivariate Linear Model," preprint at (www.math.siu.edu/olive/ppmultreg.pdf).

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

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," unpublished manuscript, online at
(www.math.siu.edu/olive/pphbmld.pdf).

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.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009), "On-line Predictive Linear Regression," *The Annals of Statistics,* 37, 1566-1590.

Wang, L., Liu, X., Liang, H., and Carroll, R.J. (2011), "Estimation and Variable Selection for Generalized Additive Partial Linear Models," *The Annals of Statistics,* 39, 1827-1851.

Wood, S.N. (2006), *Generalized Additive Models: an Introduction with R*, Chapman & Hall/CRC, Boca Rotan, FL.

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.