Noname manuscript No.
(will be inserted by the editor)

Applications of Hyperellipsoidal Prediction Regions

David J. Olive

Received: date / Accepted: date

Abstract Olive (2013) developed a large sample $100(1-\delta)\%$ nonparametric prediction region for a future $m \times 1$ test vector $\boldsymbol{y}_f$ given past training data $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$. Consider predicting an $m \times 1$ future test response vector $\boldsymbol{y}_f$, given $\boldsymbol{x}_f$ and past training data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$. For the multivariate linear regression model $\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$, let the pseudodata $\boldsymbol{w}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$ where the $\hat{\boldsymbol{\epsilon}}_i$ are the residual vectors. Under mild regularity conditions, applying the Olive (2013) prediction region to the pseudodata gives a large sample $100(1-\delta)\%$ nonparametric prediction region for $\boldsymbol{y}_f$.

Suppose there is an $m \times 1$ statistic T_n such that $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$. Under regularity conditions, applying the Olive (2013) prediction region to the bootstrap sample $T_1^*, ..., T_B^*$ gives a large sample $100(1-\delta)\%$ confidence region for the parameter vector $\boldsymbol{\mu}$.

Keywords Bagging · Bootstrap · Highest density region · Prediction interval · Multivariate linear regression

1 Introduction

This paper shows that the Olive (2013) nonparametric prediction region computed from pseudodata can result in a prediction region for the multivariate linear regression model, while the nonparametric prediction region computed from a bootstrap sample can result in a confidence region.

A multivariate regression model has an $m \times 1$ response vector $\boldsymbol{y}_i$ and a $p \times 1$ vector of predictor variables $\boldsymbol{x}_i$ with $\boldsymbol{w}_i = (\boldsymbol{y}_i, \boldsymbol{x}_i)$. Let $\boldsymbol{w}_i = \boldsymbol{y}_i$ for the

David J. Olive
Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA.
Tel.: 618-453-6566
Fax: 618-453-5300
E-mail: dolive@siu.edu

multivariate location and dispersion model that has no $\boldsymbol{x}_i$. Given training data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ and $\boldsymbol{x}_f$, a large sample $100(1-\delta)\%$ *prediction region* for an $m \times 1$ future test random vector $\boldsymbol{y}_f$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1-\delta$, while a large sample $100(1-\delta)\%$ *confidence region* for an $m \times 1$ vector of parameters $\boldsymbol{\mu}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \rightarrow 1-\delta$ as the sample size $n \rightarrow \infty$. When $m = 1$, a large sample $100(1-\delta)\%$ *prediction interval* (PI) $[\hat{L}_n, \hat{U}_n]$ satisfies $P(\hat{L}_n \leq W_f \leq \hat{U}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$.

For the multivariate location and dispersion model, there is a some literature for prediction regions that may perform well for small m. Following Hyndman (1996), when unique, the $100(1-\delta)\%$ highest density region is $R(f_{1-\delta}) = \{\boldsymbol{z} : f(\boldsymbol{z}) \geq f_\delta\}$ where f_δ is the largest constant such that $P[\boldsymbol{y} \in R(f_{1-\delta})] \geq 1-\delta$ and $f(\boldsymbol{z})$ is the probability density function (pdf) of $\boldsymbol{y}$. Let $\hat{f}_{(1)}, ..., \hat{f}_{(n)}$ be the order statistics of $\hat{f}(\boldsymbol{y}_1), ..., \hat{f}(\boldsymbol{y}_n)$. Hyndman (1996) used the estimated highest density region

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

where $d > 0$ can be any constant, $h = max(1, \lfloor n\delta \rfloor)$, and $\lfloor x \rfloor$ is the integer part of x. (Often $f(\boldsymbol{z}) = kg(\boldsymbol{z})$ and $d = 1/k > 0$.) Also see Lei, Robins, and Wasserman (2013), who estimate $f(\boldsymbol{z})$ with a kernel density estimator.

For $m = 1$ and positive integer c, the shorth(c) estimator is a useful estimator of the highest density region when the region is an interval. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{2}$$

Let

$$k_n = \lceil n(1-\delta) \rceil \tag{3}$$

where $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$ when the nominal coverage is $1-\delta$, and used the shorth(c) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{4}$$

Mohie El–Din and Shafay (2013) also derived prediction intervals based on order statistics. Olive (2007, 2013) used the shorth of the pseudodata $Z_i = \hat{Y}_f + \hat{e}_i$ to make prediction intervals for multiple linear regression and the additive error regression model $Y_i = g(\boldsymbol{x}_i) + e_i$ where $g(\boldsymbol{x}_i)$ is known up to a set of unknown parameters and $\hat{e}_i$ is the ith residual for $i = 1, .., n$. Also see Cai, Tian, Solomon, and Wei (2008) and Lei and Wasserman (2014).

Some notation is needed to describe the Olive (2013) nonparametric prediction region that performs well even if m is large. Suppose $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are iid $m \times 1$ random vectors with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma_y}$. Let $(\overline{\boldsymbol{y}}, \boldsymbol{S})$ be the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \ \text{ and } \ \boldsymbol{S} = \boldsymbol{S_y} = \frac{1}{\text{n}-1}\sum_{\text{i}=1}^{\text{n}} (\boldsymbol{y}_\text{i} - \overline{\boldsymbol{y}})(\boldsymbol{y}_\text{i} - \overline{\boldsymbol{y}})^\text{T}. \tag{5}$$

Then the ith *squared sample Mahalanobis distance* is the scalar

$$D^2_{\boldsymbol{w}} = D^2_{\boldsymbol{w}}(\overline{\boldsymbol{y}}, \boldsymbol{S}) = (\boldsymbol{w} - \overline{\boldsymbol{y}})^T \boldsymbol{S}^{-1} (\boldsymbol{w} - \overline{\boldsymbol{y}}). \quad (6)$$

Let $D_i^2 = D^2_{\boldsymbol{y}_i}$ for each observation $\boldsymbol{y}_i$. Let $D_{(c)}$ be the cth order statistic of $D_1, ..., D_n$. Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{y}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{w} : D_{\boldsymbol{w}}(\overline{\boldsymbol{y}}, \boldsymbol{S}) \leq D_{(c)}\}. \quad (7)$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

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

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

Let $D_{(U_n)}$ be the $100q_n$th percentile of the D_i. For example, use $U_n = c = \lceil nq_n \rceil$. Then the Olive (2013) large sample $100(1-\delta)\%$ nonparametric prediction region for a future value $\boldsymbol{y}_f$ given iid data $\boldsymbol{y}_1, ..., , \boldsymbol{y}_n$ is

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{y}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \quad (9)$$

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

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{y}}, \boldsymbol{S}) \leq \chi^2_{m,1-\delta}\}. \quad (10)$$

See Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). Here the population percentile $u_{1-\delta}$ of a random variable U satisfies $P(U \leq u_{1-\delta}) = 1-\delta$ where often $U \sim \chi^2_m$, a chi–square distribution with m degrees of freedom. Di Bucchianico, Einmahl, and Mushkudiani (2001) used the minimum volume ellipsoid to compute small volume covering regions for $m \leq 2$.

Olive (2013) showed that (9) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. If $m = 1$ and $n \geq 20$, the correction factor $\lceil nq_n \rceil$ closely tracks the Frey (2013) correction factor (4). The volume of the hyperellipsoid

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{\boldsymbol{y}})^T \boldsymbol{S}^{-1} (\boldsymbol{w} - \overline{\boldsymbol{y}}) \leq h^2\} \text{ is equal to } \frac{2\pi^{\mathrm{m}/2}}{\mathrm{m}\Gamma(\mathrm{m}/2)} \mathrm{h}^{\mathrm{m}} \sqrt{\det(\boldsymbol{S})}, \quad (11)$$

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

The ratio of the volumes of prediction regions (10) and (9) is

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

which can become close to zero rapidly as m gets large. Hence if the data distribution is not the multivariate normal $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma y})$ distribution, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension m increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality,

$P(D^2_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) \leq \gamma) \geq 1 - m/\gamma > 0$ for $\gamma > m$. See Budny (2014), Chen (2011), and Navarro (2014a, 2016). Navarro (2014b) makes a prediction region for $\boldsymbol{y}_f$ based on the multivariate Chebyshev inequality where the cutoff tends to be larger than $D^2_{(U_n)}$. For example, replace $D^2_{(U_n)}$ by $\gamma = m/\delta$ in (9).

Section 2 derives a nonparametric prediction region for the multivariate linear regression model. Section 3 shows that applying the Olive (2013) nonparametric prediction region on a bootstrap sample gives a confidence region.

2 Prediction Regions for Multivariate Regression

This section will derive a prediction region for multivariate regression models of the form $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = g(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ where the function $g(\boldsymbol{x})$ is known up to a set of unknown parameters, but the distribution of $\boldsymbol{\epsilon}_i$ may not be known. The multivariate linear regression model satisfies the regularity conditions.

The following technical theorem will be needed to prove Theorem 2, which shows how to obtain a practical prediction region using pseudodata. The ith residual vector $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$.

Theorem 1 *Let $a > 0$ and assume that $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$.* Then

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

Proof Let B_n denote the subset of the sample space on which $\hat{\boldsymbol{\Sigma}}_n$ has an inverse. Then $P(B_n) \to 1$ as $n \to \infty$. Now

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

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

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

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

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

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

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

$$+ \frac{1}{a} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T [a \hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)$$

on B_n, and the last three terms are $o_P(1)$. □

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

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

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

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{12}$$

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

b) *If* $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ *is a consistent estimator of* $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, *then (12) 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* $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, *then the prediction region (12) 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1-\delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1-\delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ exists, Theorem 1 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in distribution, 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ at continuity points of the distribution of $D(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

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). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then approximately $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region with $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Of course the volume of the prediction region could be large if a poor estimator is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured

distribution. Olive, Pelawa Watagoda, and Rupasinghe Arachchige Don (2015) suggest that the residual, response, and DD plots described below can be used to check model assumptions. They considered tests for the multivariate linear regression model, but did not develop prediction regions.

Prediction region (12) can be used for the Su and Cook (2012) inner envelopes estimator and the seemingly unrelated regressions model. Theorem 3 shows that prediction region (12) is the prediction region (9) applied to pseudodata for the *multivariate linear model*

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i \tag{13}$$

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

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

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

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

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

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

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. For Theorem 3, if $D_{1-\delta}$ is a continuity point of the distribution of D, then (12) will be a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ if the $\boldsymbol{\epsilon}_i$ are iid with fourth moments and a nonsingular covariance matrix,

$\max_{1\leq i\leq n} h_i \xrightarrow{P} 0$ and $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$ as $n \to \infty$. The ith leverage h_i is the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let the $m \times 1$ column vector T be a multivariate location estimator, and the $m \times m$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator.

Theorem 3 *For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (12) is the prediction region (9) applied to the $\hat{\boldsymbol{z}}_i$.*

Proof Multivariate linear regression with least squares satisfies Theorem 2 by Su and Cook (2012). Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (5) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\boldsymbol{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. □

These prediction regions can be displayed with the Rousseeuw and Van Driessen (1999) DD plot of $MD_i = D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $RD_i = D_i(T, \boldsymbol{C})$. For $(T, \boldsymbol{C})$, we will use the Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ for a large class of elliptically contoured distributions, where $a = 1$ for the $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Also see Olive (2016b, ch. 4) and Zhang, Olive, and Ye (2012). For iid data and large n, Olive (2002) showed that plotted points in the DD plot scatter tightly about a line through the origin for a large class of elliptically contoured distributions, and about the identity line with unit slope and zero intercept if the data are multivariate normal. Simulations suggest that the DD plot of the residual vectors can be used in a similar way.

Olive (2013) used three prediction regions that can be extended to multivariate linear regression. The regions have the form

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T\boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}. \quad (14)$$

Let (12) be the nonparametric region with $h = D_{(U_n)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{m,q_n}$ where $P(W \leq \chi^2_{m,q_n}) = q_n$ if $W \sim \chi^2_m$. The semiparametric and parametric regions are only conjectured to be large sample prediction regions for the multivariate regression model, but are useful as diagnostics. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D^2_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T\boldsymbol{S}_r^{-1}(\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, ..., n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}, \quad (15)$$

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

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi^2_{m,1-\delta}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \sqrt{\chi^2_{m,1-\delta}}\}. \quad (16)$$

The nonparametric prediction region (15) has simple geometry. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$. Then R_r is a

hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

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

Example 1. Cook and Weisberg (1999a, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$, and $X_4 = H$: the shell length, log(width), and height. Figures 1 and 2 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships, and highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. Figure 3 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Cases 8, 48, and 79 have especially large distances. The horizontal line $RD \approx 3$ corresponds to the semiparametric region. These two lines were also the 95th percentiles of the MD_i and RD_i. The horizontal line $RD \approx 2.45$ corresponds to the parametric MVN region. A vertical line $MD \approx 2.45$ (not shown) corresponds to a large sample classical region.

Suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots (not shown) for Y_2 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 4 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error vector distribution is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$, and the prediction regions still contain 95% of the training data $\boldsymbol{y}_i$.

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

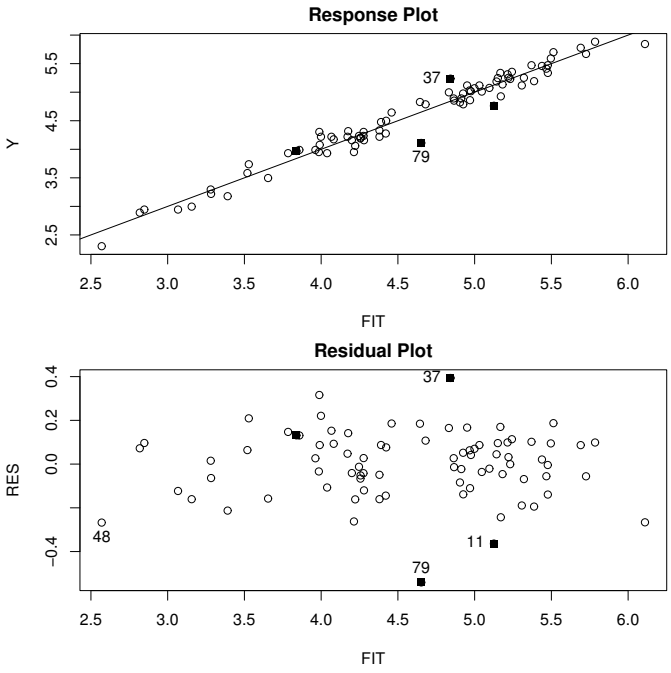

Fig. 1 Plots for $Y_1 = \log(S)$.

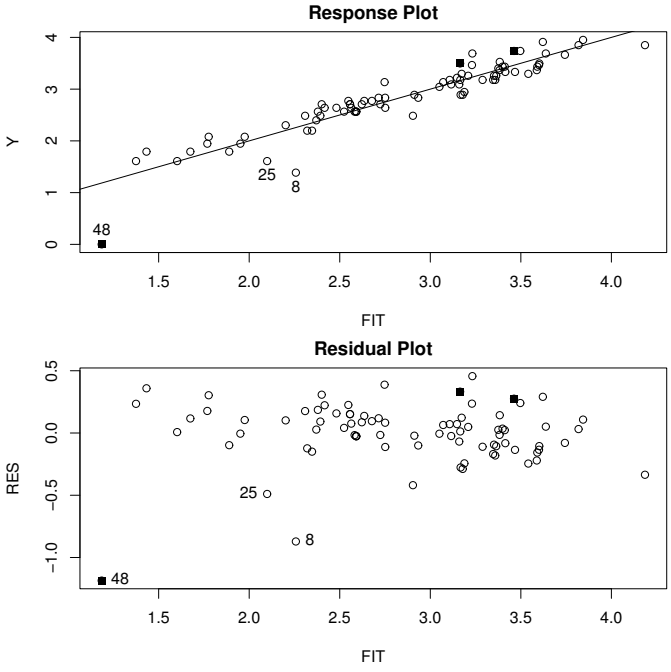

Fig. 2 Plots for $Y_2 = \log(M)$.

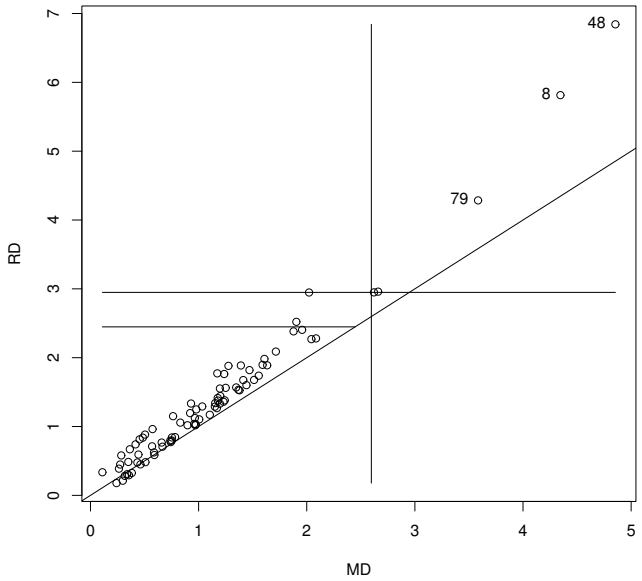

Fig. 3 DD Plot of the Residual Vectors for the Mussel Data.

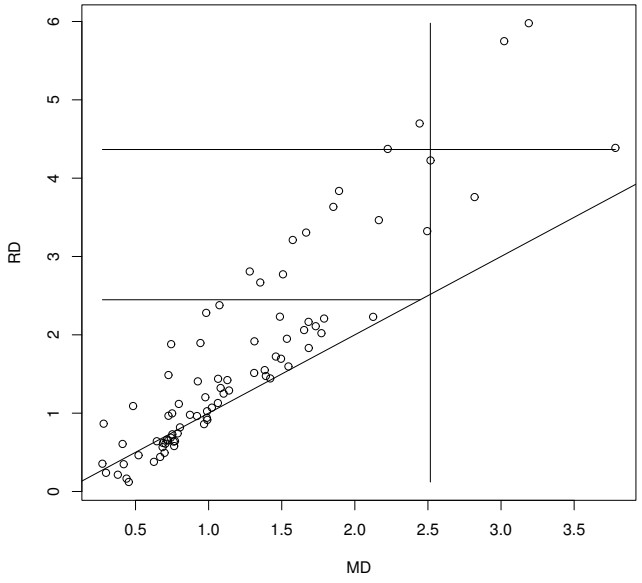

Fig. 4 DD Plot if $Y_2 = M$.

Table 1 Simulated Coverages for 90% Prediction Regions.

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

where $\psi = 0.10$. Hence the correlations are $(2\psi+(m-2)\psi^2)/(1+(m-1)\psi^2)$. As ψ gets close to 1, the data clusters about the line in the direction of $(1, ..., 1)^T$. We used $\boldsymbol{w}_i \sim N_m(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\boldsymbol{0}, \boldsymbol{I}) + \tau N_m(\boldsymbol{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

Then 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. As in Olive (2013), the ratio of the prediction region volumes

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

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

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

As in Olive (2013), for iid $\boldsymbol{y}_i$ from an elliptically contoured distribution, coverage was often near the nominal value for $n \geq 10p$, and voln was often near 1 for $n \geq 50p$. In the simulations, we took $n = 3(m + p)^2$ and $m = p$. Table 1 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three

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

3 Bootstrapping Confidence Regions and Hypothesis Tests

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $m \times 1$ vector. For example, let $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{A}$ is a known full rank $m \times p$ matrix with $1 \leq m \leq p$. Let the statistic T_n be an estimator of $\boldsymbol{\mu}$ based on a sample of size n. This section shows that under regularity conditions, applying the large sample $100(1-\delta)\%$ prediction region (9) to the bootstrap sample $T_1^*, ..., T_B^*$ gives a large sample $100(1-\delta)\%$ confidence region for the $m \times 1$ parameter vector $\boldsymbol{\mu}$. We call this bootstrap technique the prediction region method.

For $m = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the $T_{i,n}^*$ from a bootstrap sample $T_{1,n}^*, ..., T_{B,n}^*$ where the statistic T_n is an estimator of μ based on a sample of size n. Often the n is suppressed in the double subscripts. Let $T_{(1)}^*, T_{(2)}^*, ..., T_{(B)}^*$ be the order statistics of the bootstrap sample. Then one version of the percentile method discards the largest and smallest $\lceil B\delta/2 \rceil$ order statistics, resulting in an interval $[\hat{L}_B, \hat{R}_B]$. We recommend using the Frey (2013) shorth (c) interval when $m = 1$. Hall (1988) discusses the shortest bootstrap interval based on all bootstrap samples.

The following theorem shows that the hyperellipsoid R_c centered at the statistic T_n is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, but the hyperellipsoid centered at known $\boldsymbol{\mu}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$.

Theorem 4 *Let the $100(1-\delta)$th percentile $D_{1-\delta}^2$ be a continuity point of the distribution of D^2. Assume that $D_{\boldsymbol{\mu}}^2(T_n, \boldsymbol{\Sigma}_T) \xrightarrow{D} D^2$, $D_{\boldsymbol{\mu}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) \xrightarrow{D} D^2$, and $\hat{D}_{1-\delta}^2 \xrightarrow{P} D_{1-\delta}^2$ where $P(D^2 \leq D_{1-\delta}^2) = 1-\delta$. i) Then $R_c = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, and if $\boldsymbol{\mu}$ is known, then $R_p = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$. ii) Region R_c contains $\boldsymbol{\mu}$ iff region R_p contains T_n.*

Proof i) Note that $D_{\boldsymbol{\mu}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) = D_{T_n}^2(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T)$. Thus the probability that R_c contains $\boldsymbol{\mu}$ is $P(D_{\boldsymbol{\mu}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2) \to 1-\delta$, and the probability that R_p contains $T_{f,n}$ is $P(D_{\boldsymbol{\mu}}^2(T_{f,n}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2) \to 1-\delta$, as $n \to \infty$.

ii) $D_{\boldsymbol{\mu}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2$ iff $D_{T_n}^2(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}_{1-\delta}^2$. □

Motivated by Theorem 4, the prediction region method applies the nonparametric prediction region (9) to the bootstrap sample to get a confidence

region. The rather simple theory follows. Let $\overline{T^*} = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ and $\boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T^*})(T_i^* - \overline{T^*})^T$ be the sample mean and sample covariance matrix of $T_1^*, ..., T_B^*$. Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_T$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_T$ and $\boldsymbol{S}_T^*$ are nonsingular $m \times m$ matrices, and T_n is an estimator of $\boldsymbol{\mu}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U} \tag{17}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_T^{-1/2}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{U} = \boldsymbol{Z},$$

$$n\ (T_n - \boldsymbol{\mu})^T\ \hat{\boldsymbol{\Sigma}}_T^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Z}^T\boldsymbol{Z} = D^2$$

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

$$(T_n - \boldsymbol{\mu})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} D^2 \tag{18}$$

as $n, B \to \infty$. Assume $P(D^2 \le D^2_{1-\delta}) = 1 - \delta$.

If the distribution of D^2 is known, then a common bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{1-\delta}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{1-\delta}\}. \tag{19}$$

Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$, and $D^2 \sim \chi^2_m$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_T^{-1}$. Machado and Parente (2005) provide sufficient conditions and references for when $n\boldsymbol{S}_T^*$ is a consistent estimator of $\boldsymbol{\Sigma}_T$.

Bickel and Ren (2001) use $n\hat{\boldsymbol{\Sigma}}_T^{-1}$ instead of $[\boldsymbol{S}_T^*]^{-1}$, and replace the D^2 cutoff in (19) by $D^2_{(k_B)}$ where $D^2_{(k_B)}$ is computed from $D_i^2 = n(T_i^* - T_n)^T\hat{\boldsymbol{\Sigma}}_T^{-1}(T_i^* - T_n)$ for $i = 1, ..., B$. If $n\boldsymbol{S}_T^* = \hat{\boldsymbol{\Sigma}}_T$, the (modified) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \tag{20}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$ for $i = 1, ..., B$.

The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \tag{21}$$

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

Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{m,1-\delta}$ cutoff by $D^2_{(U_B)}$.

Given (17) and (18), a sufficient condition for (20) to be confidence region is

$$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}, \tag{22}$$

while sufficient conditions for (21) to be confidence region are

$$\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}, \tag{23}$$

and

$$\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}. \tag{24}$$

Note (23) and (24) follow from (22) and (17) if $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{P} \boldsymbol{0}$, so $T_n - \overline{T}^* = o_P(n^{-1/2})$.

As in Bickel and Ren (2001), let $\boldsymbol{\mu} = T(F)$, $T_n = T(F_n)$, and $T^* = T(F_n^*)$ where F is the cumulative distribution function (cdf) of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (Hadamard differentiable with a well behaved Hadamard derivative $\dot{T}(F)$), then (17) and (22) hold with $\boldsymbol{U} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size k is drawn with replacement from the empirical distribution, then $\sqrt{k}(T(F_k^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $k = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{U}$.

To justify the prediction region method, assume that (17) and (22) hold where $\boldsymbol{U} \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$, an $m \times 1$ multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}_T$. Use $\boldsymbol{Z}_n \sim AN_m(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_m(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let $T_i^* = T_{i,n}^*$. Then $T_i^* \sim AN_m\left(T_n, \frac{\boldsymbol{\Sigma}_T}{n}\right)$. Fix n temporarily and let $\boldsymbol{W}_i = \sqrt{n}(T_i^* - T_n)$. Then with respect to the bootstrap distribution (so conditional on the data), $\boldsymbol{W}_1, ..., \boldsymbol{W}_B$ are iid, and $\sqrt{n}(\overline{T}^* - T_n) = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{W}_i \sim AN_m\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_T}{B}\right)$ is a normal approximation. Hence $\sqrt{nB}(\overline{T}^* - T_n) \sim AN_m(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$. Now unfix n. Since the same normal approximation holds for n and B large (and $AN_m(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$ does not depend on n or B), it follows that $\overline{T}^* - T_n = o_P(n^{-1/2})$.

The prediction region method should often work if $E(\overline{T}^*) - T_n = o_P(n^{-1/2})$ and the asymptotic covariance matrix of $\overline{T}^*$ is $\frac{\boldsymbol{\Sigma}_T}{nB}$ as $n, B \to \infty$. As in Efron (2014), $\overline{T^*}$ is the bagging or smoothed bootstrap estimator of $\boldsymbol{\mu}$, which often outperforms T_n for inference. See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

These results suggest that under reasonable conditions, (17), (22), (23), and (24) hold: $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}$, and $\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$. Stronger conditions are needed for $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_T$. The regularity conditions for the prediction region method are weaker when $m = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed: the prediction region method is the closed interval centered at $\overline{T}^*$ just long enough to contain U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method when $m = 1$. Efron (2014) also used a confidence interval centered at the bagging estimator $\overline{T}^*$.

The prediction region method is simple. Let $\hat{\boldsymbol{\mu}}$ be a consistent estimator of $\boldsymbol{\mu}$ and make a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, ..., B$. Using (21) applied to the $\boldsymbol{w}_i$ as a large sample $100(1-\delta)\%$ confidence region, fail to reject H_0 if $\boldsymbol{0}$ is in the confidence region (if $D_{\boldsymbol{0}} \leq D_{(U_B)}$), and reject H_0 otherwise.

As an example, consider variable selection for the linear model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Let $\hat{\boldsymbol{\beta}}_{I_{min}}$ correspond to the submodel that minimized the C_p criterion, and form $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ by adding 0s corresponding to the omitted variables. Then the residual bootstrap method can be applied: instead of computing the least squares estimator from regressing $\boldsymbol{Y}_i^*$ on $\boldsymbol{X}$, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$. Then test $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ using the prediction region method for $m > 1$ and using the Frey (2013) shorth(c) interval if $m = 1$.

Table 2 Bootstrapping the All Subsets Variable Selection Model

variable	$\hat{\beta}_{I_{min},0}$	OLS SE	shorth intervals
constant	-0.9573	0.1519	[−2.769, 0.460]
L	0		[−0.004, 0.004]
logW	0		[−0.595, 0.869]
H	0.0072	0.0047	[0.000, 0.016]
logS	0.6530	0.1160	[0.324, 0.913]

Example 2. Consider the data from Example 1, but let $Y = \log(M)$, and let the other variables be the predictors with β_1 and $x_1 \equiv 1$ corresponding to the constant. The minimum all subsets C_p model I_{min} used a constant, H, and $\log(S)$. Table 2 shows results for this model including the shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap. The OLS SE would only be correct if I_{min} was selected before looking at the data. Note that the interval for H is right skewed and contains 0 when closed intervals are used instead of open intervals. Consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = (\beta_2, \beta_3, \beta_4)^T = \boldsymbol{0}$. Using the prediction region method with the I_{min} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\boldsymbol{0}} = 1.134$. So fail to reject H_0. Hence $\log(S)$ appears to be the important predictor.

A small simulation study was done in R using $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$, $p = 4$, and various zero mean iid error distributions. The design matrix $\boldsymbol{X}$ consisted

of iid N(0,1) random variables. Hence the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance σ^2. The simulation computed the shorth(c) interval for each β_i and used the prediction region method to test $H_0 : \beta_3 = \beta_4 = 0$. Observed coverage between 0.94 and 0.96 suggests the actual coverage is close to the nominal coverage 0.95.

Table 3 Bootstrapping Regression and Variable Selection

model	cov/len	β_1	β_2	β_3	β_4	test
reg	cov	0.9496	0.9430	0.9440	0.9454	0.9414
	len	0.3967	0.3996	0.3997	0.3997	2.4493
vs	cov	0.9482	0.9486	0.9974	0.9974	0.9896
	len	0.3965	0.3990	0.3241	0.3257	2.6901

The regression models used the residual bootstrap on the full model least squares estimator and on the all subsets variable selection estimator for the model I_{min}. The residuals were from least squares applied to the full model in both cases. Results are shown for when the iid errors $e_i \sim N(0, 1)$. Table 3 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for the all subsets variable selection. The column for the "test" gives the length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ where $D_{(U_B)}$ is the cutoff for the confidence region. The volume of the confidence region will decrease to 0 as $n \to \infty$. The cutoff will often be near $\sqrt{\chi^2_{m,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is very close to 2.4493 for the full model regression bootstrap test. The coverages were near 0.95 for the regression bootstrap on the full model. For I_{min} the coverages were near 0.95 for β_1 and β_2, but higher for the other 3 tests since zeroes often occurred for $\hat{\beta}^*_j$ for $j = 3, 4$. The average lengths and coverages were similar for the full model and all subsets variable selection I_{min} for β_1 and β_2, but the lengths were shorter for I_{min} for β_3 and β_4. Volumes of the hyperellipsoids were not computed, but the average cutoff of 2.69 for the variable selection test suggests that the test statistic was not asymptotically normal, which is not surprising since many zeroes were produced for $\hat{\beta}^*_j$ for $j = 3, 4$.

See Olive (2016a) for more information about the prediction region method. Schomaker (2012) suggests bootstrap estimates of the standard error of $\hat{\beta}_i$ for shrinkage estimators. Firinguetti and Bobadilla (2011) suggest confidence intervals for β_i for ridge regression.

4 Conclusion

Under regularity conditions, the Olive (2013) nonparametric prediction region computed from pseudodata gives a prediction region, while the nonparametric prediction region computed from a bootstrap sample gives a confidence region that can be used for hypothesis testing for complicated models such as variable selection models. The Olive (2013) prediction intervals can also be used for some variable selection models, and may be useful for cross validation.

Applications of the prediction region method are numerous, but we may need $n \geq 50m$ and $B \geq \max(100, n, 50m)$ if the test statistic has an approximate multivariate normal distribution. Sample sizes may need to be much larger for other limiting distributions. A similar technique can be used to estimate the $100(1-\delta)\%$ Bayesian credible region for $\boldsymbol{\theta}$. Generate $B = \max(100000, n)$ values of $\hat{\boldsymbol{\theta}}$ from the posterior distribution, and compute the prediction region (9). Use prediction region (1) with $\hat{f} = f$ if the posterior pdf f is known. Olive (2014, pp. 283, 364) used the shorth(k_B) estimator to compute shorter bootstrap confidence intervals, and to estimate the highest density region corresponding to a known posterior pdf for Bayesian inference. Mohie El–Din and Shafay (2013) consider Bayesian prediction intervals.

The practical method for making prediction regions does not need the error distribution to be known. Other methods, such as (1), may not be competitive for m much larger than two. Obtaining prediction regions when the errors are not additive is a difficult problem. See Cai, Tian, Solomon, and Wei (2008) for some useful results. Plots and simulations were done in *R*. See R Core Team (2015). Programs are in the collection of functions *mpack* available at (http://lagrange.math.siu.edu/Olive/mpack.txt). The function `mpredsim` was used to simulate the prediction regions (15), `mregddsim` simulated the residual vector DD plots for various distributions, and the function `ddplot4` makes the DD plots. Functions `hdr2` and `predrgn` can be used to simulate (1) and (9). The functions `regbootsim` and `vsbootsim` can be used to simulate the bootstrap tests for multiple linear regression and for the all subsets variable selection model that minimizes C_p.

Acknowledgements The author thanks the Editor and two referees for their work.

References

1. Bickel PJ, Ren JJ (2001) The bootstrap in hypothesis testing. In State of the art in probability and statistics: festschrift for William R. van Zwet, eds. de Gunst M, Klaassen C, van der Vaart A, The Institute of Mathematical Statistics, Hayward, CA, 91-112.
2. Büchlmann P, Yu B (2002) Analyzing bagging. Ann Stat 30:927-961.
3. Budny K (2014) A generalization of Chebyshev's inequality for Hilbert-space-valued random variables. Stat Probab Lett 88:62-65.
4. Cai T, Tian L, Solomon SD, Wei, LJ (2008) Predicting future responses based on possibly misspecified working models. Biometrika 95:75-92.
5. Chen X (2011) A new generalization of Chebyshev inequality for random vectors. See arXiv:0707.0805v2.

6. Chew V (1966) Confidence, prediction and tolerance regions for the multivariate normal distribution. J Am Stat Assoc 61:605-617.
7. Cook RD, Weisberg S (1999a) Applied regression including computing and graphics. Wiley, New York, NY.
8. Cook RD, Weisberg S (1999b) Graphs in statistical analysis: is the medium the message? Am Stat 53:29-37,
9. Di Bucchianico A, Einmahl JHJ, Mushkudiani NA (2001) Smallest nonparametric tolerance regions. Ann Stat 29:1320-1343.
10. Efron B (2014) Estimation and accuracy after model selection (with discussion). J Am Stat Assoc 109:991-1007.
11. Firinguetti L, Bobadilla G (2011) Asymptotic confidence intervals in ridge regression based on the Edgeworth expansion. Stat Pap 52:287-307.
12. Frey J (2013) Data-driven nonparametric prediction intervals. J Stat Plan Inference 143:1039-1048.
13. Friedman JH, Hall P (2007) On bagging and nonlinear estimation. J Stat Plan Inference 137:669-683.
14. Hall P (1988) Theoretical comparisons of bootstrap confidence intervals (with discussion). Ann Stat 16:927-985.
15. Hyndman RJ (1996) Computing and graphing highest density regions. Am Stat 50:120-126.
16. Johnson RA, Wichern DW (1988) Applied multivariate statistical analysis, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
17. Lei J, Robins J, Wasserman L (2013) Distribution free prediction sets. J Am Stat Assoc 108:278-287.
18. Lei J, Wasserman L (2014) Distribution free prediction bands. J Roy Stat Soc B 76:71-96.
19. Machado JAF, Parente P (2005) Bootstrap estimation of covariance matrices via the percentile method. Econometrics Journal 8:70–78.
20. Mohie El–Din MM, Shafay AR (2013) One– and two–sample Bayesian prediction intervals based on progressively type–II censored data. Stat Papers 54:287-307.
21. Navarro J (2014a) Can the bounds in the multivariate Chebyshev inequality be attained? Stat Probab Lett 91:1-5.
22. Navarro J (2014b) A note on confidence regions based on the bivariate Chebyshev inequality. Applications to order statistics and data sets. Istatistik J Turkish Stat Assoc 7:1-14.
23. Navarro J (2016) A very simple proof of the multivariate Chebyshev's inequality. Commun Stat Theory Meth 45:3458-3463.
24. Olive DJ (2002) Applications of robust distances for regression. Technom 44:64-71.
25. Olive DJ (2007) Prediction intervals for regression models. Computat Stat Data Anal 51:3115-3122.
26. Olive DJ (2013) Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. Internat J Stat Probab 2:90-100.
27. Olive DJ (2014) Statistical theory and inference. Springer, New York, NY.
28. Olive DJ (2016a) Bootstrapping hypothesis tests and confidence regions. Unpublished manuscript. http://lagrange.math.siu.edu/Olive/ppvselboot.pdf
29. Olive DJ (2016b) Robust Multivariate Analysis, Springer, New York, NY, to appear.
30. Olive DJ, Hawkins DM (2005) Variable selection for 1D regression models. Technom 47:43-50.
31. Olive DJ, Hawkins DM (2010) Robust multivariate location and dispersion. Preprint. http://lagrange.math.siu.edu/Olive/pphbmld.pdf
32. Olive DJ, Pelawa Watagoda LCR, Rupasinghe Arachchige Don HS (2015) Visualizing and testing the multivariate linear regression model. Internat J Stat Probab 4:126-137.
33. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
34. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technom 41:212-223.
35. Schomaker M (2012) Shrinkage averaging estimation. Stat Pap 53:1015-1034.
36. Su Z, Cook RD (2012) Inner envelopes: efficient estimation in multivariate linear regression. Biometrika 99:687-702.
37. Zhang J, Olive DJ, Ye P (2012) Robust covariance matrix estimation with canonical correlation analysis. Internat J Stat Probab 1:119-136.