A Data Splitting Prediction Region

Lingling Zhang and David J. Olive*
Southern Illinois University

September 20, 2022

Abstract

This paper derives nonparametric data splitting prediction regions that have very simple theory. Some of the prediction regions can be used when the data distribution does not have first moments, and some can be used for high dimensional data where the number of predictors is larger than the sample size.

KEY WORDS: Conformal Prediction, High Dimensional Data, Prediction Interval.

1 Introduction

This section reviews data splitting and prediction regions. Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. A common method of data splitting randomly divides the training data into the two sets H and V. Often $n_H \approx \lceil n/2 \rceil$ where $\lceil x \rceil$ is the ceiling function = least integer function, e.g $\lceil 7.7 \rceil = 8$. See Zhang (2022) for references.

Consider predicting a future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where the $p \times 1$ random vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are independent and identically distributed (iid). A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. A prediction region is asymptotically optimal if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

Prediction regions often use an estimator of multivariate location and dispersion $(T, \boldsymbol{C})$. For example, let a) $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the sample mean and sample covariance matrix, b) $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median of the $\boldsymbol{x}_i$ and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, or c) let $(T, \boldsymbol{C})$ be a robust estimator such as the RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ given by Olive (2017, 2022a), Olive and Hawkins (2010), and Zhang, Olive, and Ye (2012).

*Lingling Zhang is a recent Ph.D. and David J. Olive is a Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

Olive (2013) developed large sample prediction regions using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ if the data distribution has a nonsingular population covariance matrix, and using $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ if the data distribution comes from a large class of elliptically contoured distributions. Also see Olive (2018, 2022a). If $n \geq 20p$, using $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ might result in a prediction region with smaller volume than using $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ since the robust estimator attempts to estimate a small volume hyperellipsoid. The smaller volume can also occur if outliers are present or if the data distribution is highly skewed, and the new data splitting prediction region, given in the following section, does not need the elliptically contoured distribution assumption.

Section 2 gives a data splitting prediction region for multivariate data, and section 3 gives simulations.

2 A Data Splitting Prediction Region

Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into two sets: H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $(T_H, \boldsymbol{C}_H)$ is computed using the data set H. Then the squared validation distances $D_j^2 = D_{\boldsymbol{x}_{i_j}}^2(T_H, \boldsymbol{C}_H) = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1} (\boldsymbol{x}_{i_j} - T_H)$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Let $D_{(U_V)}^2$ be the U_Vth order statistic of the D_j^2 where

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

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

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

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

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

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

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

Note that we can get the actual coverage $U_V/(1 + n_V)$ close to $1 - \delta$ for $n_V \geq 20$ for $\delta = 0.05$ even if $(T_H, \boldsymbol{C}_H)$ is a bad estimator. The volume of the prediction region tends to be much larger than that of the highest density region, even if $\boldsymbol{C}_H$ is well

conditioned. We likely need $U_V \geq 50$ for $D^2_{(U_V)}$ to approximate the population percentile of $D^2_j = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1} (\boldsymbol{x}_{i_j} - T_H)$.

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

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

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

could be used where MAD is the median absolute deviation. Regularized covariance matrices or precision matrices could also be used.

3 SIMULATIONS

The theory for the new prediction regions is simple, so Tables 1-3 are more of a check that the programs work than that the theory works. The output gives cov = observed coverage, up $\approx$ actual coverage, and mnhsq = mean cutoff $D^2_{(U_V)}$. With 5000 runs, expect observed coverage $\in [0.94, 0.96]$ if the actual coverage is close to 0.95. The random vector $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{x} = \boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ for xtype = 3, and $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$ for xtype = 1. For xtype = 2, $\boldsymbol{w}$ has the w_i iid lognormal(0,1) with $\boldsymbol{A} = diag(1, \sqrt{2}, ..., \sqrt{p})$. The dispersion matrix types are dtype = 1 if $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ and dtype = 2 if $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

When xtype=3 and dtype=1, $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ where $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Then $D^2_{(U_V)}$ should estimate the population percentile $\chi^2_{p,0.95}$ if $n \geq \max(20p, 200)$ and $n_V = 100$. This result did occur in the simulations.

4 CONCLUSIONS

Only a few nonparametric prediction regions have appeared in the literature. See Olive (2013, 2018) for references. The new prediction regions can be used for distributions that do not have an expected value if appropriate $(T, \boldsymbol{C})$ is used, e.g. $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Pelawa Watagoda and Olive (2021a) and Lei et al. (2018) use data splitting to obtain prediction intervals for the multiple linear regression model.

Prediction regions have some nice applications besides prediction. Applying a prediction region to data generated from a posterior distribution gives an estimated credible region for Bayesian Statistics. See Chen and Shao (1999). Certain prediction regions applied to a bootstrap sample result in a confidence region. See Pelawa Watagoda and Olive (2021b), Rajapaksha and Olive (2022), and Rathnayake and Olive (2021). Mykland (2003) converts prediction regions into investment strategies. The new prediction regions can be used for the Haile and Olive (2022) prediction regions for random walks.

SOFTWARE

Table 1: Data Splitting Nominal 95% Prediction region

n	p	nv	xtype	dtype	cov
50	10	20	1	1	0.9538
50	10	20	2	1	0.9550
50	10	20	3	1	0.9538
50	10	20	1	2	0.9492
50	10	20	2	2	0.9578
50	10	20	3	2	0.9554
50	50	20	1	1	0.9490
50	50	20	2	1	0.9584
50	50	20	3	1	0.9538
50	50	20	1	2	0.9512
50	50	20	2	2	0.9532
50	50	20	3	2	0.9532
50	100	20	1	1	0.9560
50	100	20	2	1	0.9466
50	100	20	3	1	0.9504
50	100	20	1	2	0.9558
50	100	20	2	2	0.9508
50	100	20	3	2	0.9522
100	10	50	1	1	0.9572
100	10	50	2	1	0.9582
100	10	50	3	1	0.9656
100	10	50	1	2	0.9664
100	10	50	2	2	0.9620
100	10	50	3	2	0.9584
100	10	25	1	1	0.9620
100	10	25	2	1	0.9628
100	10	25	3	1	0.9564
100	10	25	1	2	0.9608
100	10	25	2	2	0.9620
100	10	25	3	2	0.9624
100	50	50	1	1	0.9638
100	50	50	2	1	0.9602
100	50	50	3	1	0.9614
100	50	50	1	2	0.9620
100	50	50	2	2	0.9634
100	50	50	3	2	0.9590
100	50	25	1	1	0.9620
100	50	25	2	1	0.9646
100	50	25	3	1	0.9654
100	50	25	1	2	0.9646
100	50	25	2	2	0.9620
100	50	25	3	2	0.9580

Table 2: Data Splitting Nominal 95% Prediction region

n	p	nv	xtype	dtype	cov
100	100	50	1	1	0.9620
100	100	50	2	1	0.9622
100	100	50	3	1	0.9596
100	100	50	1	2	0.9638
100	100	50	2	2	0.9578
100	100	50	3	2	0.9638
100	100	25	1	1	0.9588
100	100	25	2	1	0.9658
100	100	25	3	1	0.9568
100	100	25	1	2	0.9622
100	100	25	2	2	0.9672
100	100	25	3	2	0.9662
200	10	100	1	1	0.9498
200	10	100	2	1	0.9470
200	10	100	3	1	0.9476
200	10	100	1	2	0.9544
200	10	100	2	2	0.9494
200	10	100	3	2	0.9504
200	10	50	1	1	0.9606
200	10	50	2	1	0.9592
200	10	50	3	1	0.9606
200	10	50	1	2	0.9632
200	10	50	2	2	0.9602
200	10	50	3	2	0.9610
200	50	100	1	1	0.9494
200	50	100	2	1	0.9552
200	50	100	3	1	0.9502
200	50	100	1	2	0.9472
200	50	100	2	2	0.9544
200	50	100	3	2	0.9550
200	50	50	1	1	0.9564
200	50	50	2	1	0.9656
200	50	50	3	1	0.9646
200	50	50	1	2	0.9624
200	50	50	2	2	0.9574
200	50	50	3	2	0.9646

Table 3: Data Splitting Nominal 95% Prediction region

n	p	nv	xtype	dtype	cov
200	100	100	1	1	0.9516
200	100	100	2	1	0.9488
200	100	100	3	1	0.9518
200	100	100	1	2	0.9540
200	100	100	2	2	0.9538
200	100	100	3	2	0.9492
200	100	50	1	1	0.9596
200	100	50	2	1	0.9602
200	100	50	3	1	0.9600
200	100	50	1	2	0.9532
200	100	50	2	2	0.9568
200	100	50	3	2	0.9584

Simulations were done in *R*. See R Core Team (2020). The collection of Olive (2022b) *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The function `predsim2` simulates the data splitting prediction region.

Acknowledgments

The authors thank the referees for their work.

REFERENCES

Chen, M.H., and Shao, Q.M. (1999), "Monte Carlo Estimation of Bayesian Credible and HPD Intervals, *Journal of Computational and Graphical Statistics,* 8, 69-92.

Haile, M.G. and Olive, D.J. (2022), "Prediction Intervals and Regions for Random Walks, and Renewal Processes," preprint at (http://parker.ad.siu.edu/Olive/pprwalk.pdf).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094-1111.

Mykland, P.A. (2003), "Financial Options and Statistical Prediction Intervals," *The Annals of Statistics*, 31, 1413-1438.

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

Olive, D.J. (2017), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

Olive, D.J. (2022a), *Robust Statistics*, online course notes, see (http://parker.ad.siu.edu/Olive/robbook.htm).

Olive, D.J. (2022b), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

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

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021a), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers*, 62, 2407-2431.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers*, 62, 681-700.

R Core Team (2020), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2022), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, to appear.

Rathnayake, R.C., and Olive, D.J. (2021), "Bootstrapping Some GLMs and Survival Regression Models after Variable Selection," *Communications in Statistics: Theory and Methods*, to appear. Preprint at (http://parker.ad.siu.edu/Olive/ppbootglm.pdf).

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.

Zhang, L. (2022), "Data Splitting Inference," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/slinglingphd.pdf).