Predicting Random Walks and a Data Splitting Prediction Region

Mulubrhan G. Haile, Lingling Zhang, and David J. Olive *
Southern Illinois University

December 20, 2023

Abstract

Perhaps the first nonparametric asymptotically optimal prediction intervals are given for univariate random walks with applications for renewal processes. Perhaps the first nonparametric prediction regions are given for vector valued random walks. This paper also 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. The prediction regions can make use of many estimators of multivariate location and dispersion.

KEY WORDS: Conformal Prediction, High Dimensional Data, Renewal Processes, Shorth

1 Introduction

This paper suggests prediction intervals and regions for univariate and vector valued random walks. This section reviews the random walk, renewal processes, nonparametric prediction intervals, and nonparametric prediction regions. Section 2 gives new nonparametric data splitting regions.

A random walk (with drift) $Y_t = Y_{t-1} + e_t$ where the e_t are independent and identically distributed (iid). Suppose there is a sample $Y_1, ..., Y_n$ and we want a prediction interval (PI) for Y_{n+h}. Then $Y_t = Y_{t-2} + e_{t-1} + e_t = Y_{t-h} + e_{t-h+1} + \cdots + e_t = Y_0 + e_1 + \cdots + e_t$, or $Y_{n+h} = Y_n + e_{n+1} + e_{n+2} + \cdots + e_{n+h} = Y_n + \epsilon_{n,h}$. Let $e_j = Y_j - Y_{j-1}$ for $j = 2, ..., n$. Divide $e_2, ..., e_n$ into blocks of length h and let ϵ_i be the sum of the e_i in each block. Hence $\epsilon_1 = e_2 + \cdots + e_{h+1}$, $\epsilon_2 = e_{h+2} + \cdots + e_{2h+1}$, and $\epsilon_i = e_{(i-1)h+2} + e_{(i-1)h+3} + \cdots + e_{(i-1)h+h+1}$ for $i = 1, ..., m = \lfloor n/h \rfloor$. These ϵ_i are iid from the same distribution as $\epsilon_{n,h}$. The same

*Mulubrhan G. Haile is Assistant Professor, Westminster College, Fulton, MO. Lingling Zhang is Visiting Assistant Professor at the University at Albany, and David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901-4408 (E-mail: *dolive@siu.edu*).

decomposition can be made for a vector valued random walk, $\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, where the vectors are $p \times 1$. Thus $\boldsymbol{\epsilon}_i = \boldsymbol{e}_{(i-1)h+2} + \boldsymbol{e}_{(i-1)h+3} + \cdots + \boldsymbol{e}_{(i-1)h+h+1}$ for $i = 1, ..., m$.

The random walk can be written as $Y_t = Y_0 + \sum_{i=1}^{t} e_i$ where $Y_0 = y_0$ is often a constant. A stochastic process $\{N(t) : t \geq 0\}$ is a counting process if $N(t)$ counts the total number of events that occurred in time interval $(0, t]$. Let e_n be the interarrival time or waiting time between the $(n-1)$th and nth events counted by the process, $n \geq 1$. If the nonnegative e_i are iid with $P(e_i = 0) < 1$, then $\{N(t), t \geq 0\}$ is a *renewal process*. Let $Y_n = \sum_{i=1}^{n} e_i =$ the time of occurrence of the nth event = waiting time until the nth event. Then Y_n is a random walk with $Y_0 = y_0 = 0$. Let the expected value $E(e_i) = \mu > 0$. Then $E(Y_n) = n\mu$ and the variance $V(Y_n) = nV(e_i)$ if $V(e_i)$ exists. A Poisson process with rate λ is a renewal process where the e_i are iid exponential EXP(λ) with $E(e_i) = 1/\lambda$. See Ross (2014) for the Poisson process and renewal process. Given $Y_1, ..., Y_n$, then n events have occurred, and the 1-step ahead PI is for the time until the next event, the 2-step ahead PI is for the time until the next 2 events, and the h-step ahead PI is for the time for the next h events.

For forecasting, predict the test data $Y_{n+1}, ..., Y_{n+L}$ given the past training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ prediction interval for Y_{n+h} is $[L_n, U_n]$ where the coverage $P(L_n \leq Y_{n+h} \leq U_n) = 1 - \delta_n$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. We often want $1 - \delta_n \to 1 - \delta$ as $n \to \infty$. A large sample $100(1-\delta)\%$ PI is asymptotically optimal if it has the shortest asymptotic length: the length of $[L_n, U_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

The shorth estimator of the population shorth will be defined as follows. If the data are $Z_1, ..., Z_n$, let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the order statistics. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$). Consider intervals that contain c cases $[Z_{(1)}, Z_{(c)}], [Z_{(2)}, Z_{(c+1)}], ..., [Z_{(n-c+1)}, Z_{(n)}]$. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Then the estimator shorth$(c) = [Z_{(s)}, Z_{(s+c-1)}]$ is the interval with the shortest length.

For a large sample $100(1-\delta)\%$ PI, the nominal coverage is $100(1-\delta)\%$. Undercoverage occurs if the actual coverage is below the nominal coverage. For example, if the actual coverage is 0.93 when $n = 100$, then for a large sample 95% PI, the undercoverage is 0.02 =2%. Suppose the data $Z_1, ..., Z_n$ are iid and a large sample $100(1-\delta)\%$ PI is desired for a future value Z_f. The shorth(c) interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1-\delta$ as $n \to \infty$, that often has the asymptotically shortest length. Frey (2013) showed that for large $n\delta$ and iid data, the shorth$(k_n = \lceil n(1-\delta) \rceil)$ prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the large sample $100(1-\delta)\%$ PI shorth$(c) =$

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

The shorth PI (1) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed that for iid data, the length and center of the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval, respectively. The correction factor also increases the length of PI (1). Einmahl and Mason (1992) give large sample theory for the shorth

under slightly milder conditions than Grübel (1982). Chen and Shao (1999) show that the shorth PI converges to the population shorth under mild conditions for ergodic data.

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

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

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

To describe the Olive (2013) nonparametric prediction region, Mahalanobis distances will be useful. Let the $p \times 1$ column vector T be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D_{\boldsymbol{w}_i}^2(T, \boldsymbol{C}) = (\boldsymbol{w}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{w}_i - T) \tag{2}$$

for each observation $\boldsymbol{w}_i$, where $i = 1, ..., n$. Notice that the Euclidean distance of $\boldsymbol{w}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance D_i uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{w}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{w}_\mathrm{i} - \overline{\boldsymbol{w}})(\boldsymbol{w}_\mathrm{i} - \overline{\boldsymbol{w}})^\mathrm{T}. \tag{3}$$

Consider predicting a future test value $\boldsymbol{w}_f$, given past training data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ where $\boldsymbol{w}_1, ..., \boldsymbol{w}_n, \boldsymbol{w}_f$ are iid. Prediction intervals are a special case of prediction regions with $p = 1$ so the $\boldsymbol{w}_i$ are random variables.

A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{w}_f \in \mathcal{A}_n) \geq 1-\delta$ asymptotically. 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 $\boldsymbol{w}_f$.

Like prediction intervals, prediction regions often need correction factors. For iid data from a distribution with a $p \times p$ nonsingular covariance matrix, it was found that the

simulated maximum undercoverage of prediction region (5) without the correction factor was about 0.05 when $n = 20p$. Hence the correction factor (4) is used to give better coverage for small n. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{4}$$

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 sample quantile of the D_i where $i = 1, ..., n$.

The large sample $100(1 - \delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{w}_f$ given iid data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{w}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{w}}) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}. \tag{5}$$

The nonparametric prediction region is a large sample prediction region if the iid $\boldsymbol{w}_i$ have a nonsingular covariance matrix, and is asymptotically optimal for a large class of elliptically contoured distributions, including multivariate normal distributions with nonsingular covariance matrices. Regions with smaller asymptotic volumes can exist if the distribution is not elliptically contoured. From Olive (2018), simulated coverage was often near the nominal for $n \geq 20p$, but simulated volumes behaved better for $n \geq 50p$. The shorth PIs do not need the mean or variance of the e_t to exist.

There are many prediction intervals and regions in the literature. See Beran (1990, 1993), Fontana, Zeni, and Vantini (2023), Guan (2023), Olive (2013, 2018), Steinberger and Leeb (2023), and Tian, Nordman, and Meeker (2022), 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. $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. 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 (2023). Mykland (2003) converts prediction regions into investment strategies.

New data splitting prediction regions that do not need the nonsingular covariance matrix to exist are given in section 2, section 3 describes the prediction intervals and regions for the random walk, while section 4 gives two examples and simulations.

2 A Data Splitting Prediction Region

Some of the new data splitting prediction regions, described in this section, can handle $\boldsymbol{\epsilon}_i$ from a distribution where the population mean does not exist. Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ 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 $n_H \approx \lceil n/2 \rceil$.

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

$$U_V = \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil). \tag{6}$$

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

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}. \tag{7}$$

To show that (7) 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 $D_{(U_V)}^2$ to approximate the population percentile of $D_j^2 = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$.

The above prediction region coverage theory did not depend on the dimension p as long as $\boldsymbol{C}$ is nonsingular. If $\boldsymbol{C} = \boldsymbol{I}_p$ or $\boldsymbol{C} = diag(S_1^2, ..., S_p^2)$, then prediction region (7) can be used for high dimensional data where $p > n$. Regularized covariance matrices or precision matrices could also be used.

3 Prediction Intervals and Regions for the Random Walk

To our knowledge, asymptotically optimal nonparametric prediction intervals for the random walk have not previously been proposed. The nonparametric prediction regions described in this section may be the first ones proposed for vector valued random walks, and are asymptotically optimal if the $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{i,h}$ are iid from a large class of elliptically contoured distributions. The random walk with drift is an AR(1) model with unit root and an ARIMA(0,1,0) model since $Y_t - Y_{t-1} = e_t$. Parametric prediction intervals are give

by Niwitpong and Panichkitkosolkul (2009) and Panichkitkosolkul and Niwitpong (2011). Wolf and Wunderli (2015) consider time series prediction regions for $(Y_{n+1}, ..., Y_{n+L})^T$. Parametric prediction regions have been given for vector autoregression (VAR) models. See Kim (1999, 2004) for details and references.

The new prediction intervals and regions for the random walks are simple. First consider the random walk $Y_t = Y_{t-1} + e_t$ where the e_t are iid. Find the ϵ_i for $i = 1, ..., m = \lfloor n/h \rfloor$. Assume $n \geq 50h$ and let $[L, U]$ be the shorth(c) PI (1) for a future value of ϵ_f based on $\epsilon_1, ..., \epsilon_m$ with $m \geq 50$. Then the large sample $100(1-\delta)\%$ PI for Y_{n+h} is $[Y_n + L, Y_n + U]$. This PI tends to be asymptotically optimal as along as the e_t are iid. This PI is equivalent to applying the shorth(c) PI (1) on $Y_n + \epsilon_1, ..., Y_n + \epsilon_m$.

For the vector valued random walk $\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, find $\boldsymbol{\epsilon}_{1,h}, ..., \boldsymbol{\epsilon}_{m,h}$. The nonparametric $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{\epsilon}_{f,h}$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{\epsilon}})^T \boldsymbol{S}_h^{-1} (\boldsymbol{z} - \overline{\boldsymbol{\epsilon}}) \leq D^2_{(U_m)}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{\epsilon}}, \boldsymbol{S}_h) \leq D^2_{(U_m)}\} \tag{8}$$

where $\boldsymbol{S}_h$ is the sample covariance matrix of the $\boldsymbol{\epsilon}_{i,h}$ and $D_i^2 = (\boldsymbol{\epsilon}_{i,h} - \overline{\boldsymbol{\epsilon}})^T \boldsymbol{S}_h^{-1} (\boldsymbol{\epsilon}_{i,h} - \overline{\boldsymbol{\epsilon}})$. This prediction region is a hyperellipsoid centered at the sample mean $\overline{\boldsymbol{\epsilon}}$. The following large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{Y}_{n+h}$ shifts the hyperellipsoid (8) to be centered at $\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}}$:

$$\{\boldsymbol{z} : [\boldsymbol{z} - (\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}})]^T \boldsymbol{S}_h^{-1} [\boldsymbol{z} - (\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}})] \leq D^2_{(U_m)}\}. \tag{9}$$

Since $\boldsymbol{Y}_{n+h}$ has the same distribution as $\boldsymbol{Y}_n + \boldsymbol{\epsilon}_{f,h}$, $P(\boldsymbol{Y}_{n+h} \in (9)) = P(\boldsymbol{\epsilon}_{f,h} \in (8)) = 1 - \delta_n$ which is bounded below by $1-\delta$, asymptotically. The prediction region (9) is equivalent to applying the nonparametric prediction region (5) to $\boldsymbol{Y}_n + \boldsymbol{\epsilon}_{1,h}, ..., \boldsymbol{Y}_n + \boldsymbol{\epsilon}_{m,h}$. The prediction region (9) is similar to the Olive (2018) prediction region for the multivariate regression model.

Since the $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{i,h}$ are iid, alternative prediction intervals and regions, such as those in section 2 or Hyndman (1995) for small p, could be used.

4 Examples and Simulations

Example 1. Common examples of random walks are stock prices. The EuStockMarkets data set, available from the *R* software, is a multivariate time series with 1860 observations on 4 variables. The observations are the daily closing prices of major European stock indices: Germany DAX, Switzerland SMI, France CAC, and UK FTSE. The data are sampled in business time, i.e., weekends and holidays are omitted. If we consider Y_t =DAX, the plot of the random walk $e_t = Y_t - Y_{t-1}$ is rectangular about the $e = 0$ line for cases 1-1460. Cases 1461-1800 scatter about the $e = 0$ line, but have much more variability (not shown but Figure 9.1 in Haile (2022)). Let cases 1-1450 be the training data, and let cases 1451-1460 be the test data. Figure 1 shows a plot of Y_{t-1} versus Y_t on the vertical axis for $t = 2$ to 1450. The two parallel lines correspond to the one step ahead 95% prediction intervals, which cover slightly more than 95% of the training data.

Example 2. The Wisseman, Hopke, and Schindler-Kaudelka (1987) pottery data consists of a chemical analysis on pottery shards. The data set has 36 cases and 5 groups

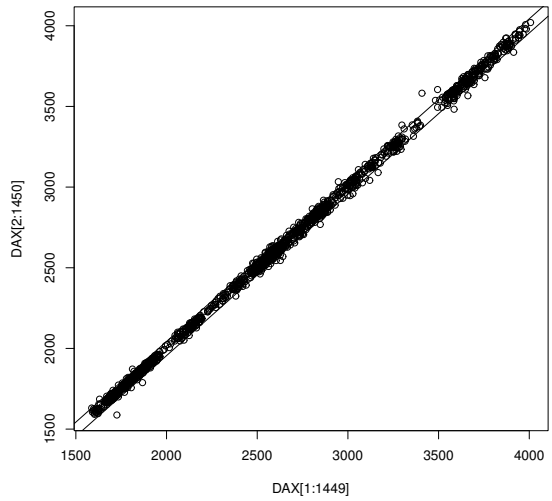

Figure 1: PI Plot of the DAX Data Set

corresponding to types of pottery shards. The variables $x_1, ..., x_{20}$ correspond to the $p = 20$ chemicals analyzed. Consider the $n = 18$ group 1 cases where the pottery shards were Arretine, a type of Roman pottery. We randomly selected case 4 from group 1 to be $\boldsymbol{x}_f$ and computed the 88.89% data splitting prediction region with the remaining 17 cases, $n_V = 8$, and $(T, \boldsymbol{C}) = (MED(\boldsymbol{W}), \boldsymbol{I}_p)$ where $MED(\boldsymbol{W})$ is the coordinatewise median computed from the 9 cases in H. The cutoff $D^2_{(U_V)} = 612.2$ and $D^2(\boldsymbol{x}_f) = 353.8$. Hence $\boldsymbol{x}_f$ was in the 88.89% prediction region. Next, we made $\boldsymbol{x}_f$ equal to each of the 36 cases. Then 8 cases $\boldsymbol{x}_f$ were not in the above prediction region, including 7 of the 18 cases that were not from group 1.

The remainder of this section gives simulations for the prediction intervals and regions. More simulations and tables are in Haile (2022). With 5000 runs, coverages between 0.94 and 0.96 suggest that there is no reason to believe that the nominal coverage is not 0.95.

A small random walk simulation was done for the large sample 95% PIs using 5000 runs with $Y_0 = 1$. The errors e_t were iid from four distributions: i) N(1,1), ii) Cauchy(1,1), iii) EXP(1), and iv) uniform$(0, 2)$. Only distribution iii) is not symmetric. We computed the h-step ahead 95% PIs for $h = 1, 2, 3, 4 = L$. We want $n \geq 50L$, but simulations may use smaller n such as $n = 25L$. The asymptotic optimal lengths are i) 3.92, 5.54, 6.79, 7.84, ii) 25.41, 50.82, 76.24, 101.65, iii) 3.00, 4.72, 6.11, 7.22, iv) 1.90, 3.11, 3.87, 4.48.

Let the population forecast error be $e(h)$. For type 1, the asymptotic optimal lengths of the large sample 95% PIs are $3.92\sqrt{h}$ where $e(h) \sim N(h, \sigma^2 = h)$. For type 2, $e(h) \sim C(h, \sigma = h)$: a Cauchy distribution. For type 3, $e(h) \sim G(h, 1)$: a Gamma distribution. For type 4, $e(2) \sim$ triangular(0,4). The distribution of the sum of n iid U(0,1) random variables is known as the Irwin-Hall distribution. See Gray and Odell (1966), Marengo, Farnsworth, and Stefanic (2017), and Roach (1963).

Results are shown in Table 1. Roughly need $n \geq 50h$ for good coverage. Thus $n = 100$ was too small for the h-step ahead PIs with $h = 3$ and $h = 4$. The Cauchy

Table 1: Random Walk 95% PI,parentheses:sd(length)

n	dist	h=1	h=2	h=3	h=4
100	N	0.9528	0.9578	0.9456	0.9220
100		4.1683(0.3923)	6.3504(0.9390)	7.2516(1.2066)	7.8247(1.4372)
100	C	0.9606	0.9656	0.9472	0.9262
100		47.33(39.38)	1075.43(41234.9)	1079.36(41233.0)	1065.19(41233.7)
100	EXP	0.9552	0.9562	0.9408	0.9242
100		3.6615(0.6325)	6.3141(1.4891)	7.1391(1.6336)	7.6647(1.8121)
100	U	0.9486	0.9584	0.9408	0.9212
100		1.9023(0.0408)	3.2878(0.2577)	3.9791(0.5093)	4.4074(0.6977)
400	N	0.9526	0.9506	0.9556	0.9508
400		4.0646(0.1868)	5.7753(0.3813)	7.2431(0.6028)	8.3282(0.7921)
400	C	0.9600	0.9622	0.9654	0.9632
400		32.7277(8.3139)	71.7138(28.29)	133.9884(79.20)	188.3578(146.52)
400	EXP	0.9582	0.9598	0.9602	0.9578
400		3.3131(0.2598)	5.1497(0.4369)	6.7619(0.6877)	7.9367(0.8970)
400	U	0.9542	0.9534	0.9568	0.9558
400		1.9028(0.0193)	3.1602(0.1268)	4.0569(0.2564)	4.7092(0.3808)
800	N	0.9514	0.9520	0.9536	0.9514
800		4.0205(0.1334)	5.7498(0.2720)	7.0086(0.4012)	8.1579(0.5338)
800	C	0.9520	0.9550	0.9516	0.9522
800		29.7122(4.9301)	65.2292(16.21)	98.9266(31.08)	144.3277(57.72)
800	EXP	0.9564	0.9550	0.9518	0.9596
800		3.2000(0.1727)	5.0514(0.3100)	6.4202(0.4333)	7.6747(0.5787)
800	U	0.9506	0.9522	0.9522	0.9518
800		1.9014(0.0132)	3.1666(0.0908)	3.9651(0.1835)	4.6357(0.2693)

Table 2: Random Walk 95% Prediction Regions, p=8

n	ψ	type	h=1	h=2	h=3	h=4
400	0	1	0.9426	0.9438	0.9370	0.9214
400	0	2	0.9490	0.9502	0.9444	0.9270
400	0	3	0.9466	0.9530	0.9476	0.9392
400	0	4	0.9416	0.9446	0.9388	0.9216
400	0.354	1	0.9514	0.9446	0.9456	0.9186
400	0.354	2	0.9450	0.9572	0.9460	0.9290
400	0.354	3	0.9556	0.9546	0.9496	0.9314
400	0.354	4	0.9416	0.9412	0.9340	0.9182
400	0.9	1	0.9484	0.9462	0.9424	0.9198
400	0.9	2	0.9524	0.9502	0.9480	0.9310
400	0.9	3	0.9482	0.9576	0.9546	0.9392
400	0.9	4	0.9458	0.9376	0.9346	0.9228
800	0	1	0.9458	0.9450	0.9460	0.9484
800	0	2	0.9516	0.9554	0.9514	0.9506
800	0	3	0.9494	0.9508	0.9480	0.9544
800	0	4	0.9432	0.9408	0.9438	0.9418
800	0.354	1	0.9456	0.9464	0.9478	0.9450
800	0.354	2	0.9474	0.9550	0.9540	0.9488
800	0.354	3	0.9534	0.9516	0.9532	0.9536
800	0.354	4	0.9494	0.9466	0.9480	0.9518
800	0.9	1	0.9436	0.9482	0.9478	0.9450
800	0.9	2	0.9500	0.9494	0.9512	0.9514
800	0.9	3	0.9552	0.9520	0.9514	0.9484
800	0.9	4	0.9474	0.9450	0.9494	0.9464
1600	0	1	0.9506	0.9516	0.9476	0.9464
1600	0	2	0.9522	0.9534	0.9532	0.9514
1600	0	3	0.9496	0.9530	0.9524	0.9522
1600	0	4	0.9418	0.9428	0.9414	0.9430
1600	0.354	1	0.9506	0.9472	0.9504	0.9502
1600	0.354	2	0.9440	0.9520	0.9488	0.9502
1600	0.354	3	0.9506	0.9572	0.9574	0.9570
1600	0.354	4	0.9488	0.9418	0.9444	0.9462
1600	0.9	1	0.9510	0.9496	0.9476	0.9458
1600	0.9	2	0.9492	0.9500	0.9532	0.9474
1600	0.9	3	0.9524	0.9558	0.9548	0.9540
1600	0.9	4	0.9450	0.9508	0.9452	0.9500

distribution needs huge n before the average PI lengths get close to the asymptotically optimal lengths. Two lines are given for each distribution-sample size combination. The first line gives the coverages while the second line gives the average PI lengths with the standard deviation of the lengths in parentheses. The coverage is the proportion of the 5000 PIs that contained the test data case $Y_f = Y_{fi}$ for $i = 1, ..., 5000$. The last two lines of Table 1 correspond the uniform(0,2) distribution with $n = 800$. The $h = i$ label corresponds to the i-step ahead 95% prediction interval with $i = 1, 2, 3$ and 4. The coverages were near 0.95 and the simulated average lengths (1.9014, 3.1666, 3.9651, 4.6357) were near the asymptotically optimal lengths (1.90, 3.11, 3.87, 4.48).

A small vector valued random walk simulation was also done for the large sample 95% prediction regions using 5000 runs. We used distributions with nonsingular population covariance matrices. Let $\boldsymbol{u}_t = (u_{t1}, ..., u_{tp})^T$ where the u_{ti} are iid from a type 1) $N(1, 1)$, 2) $1 + t_5$, 3) EXP(1), or 4) U(0,2) distribution. Then $\boldsymbol{e}_t = \boldsymbol{A}\boldsymbol{u}_t$ where the $p \times p$ matrix $\boldsymbol{A} = (a_{ij})$ with the diagonal elements $a_{ii} = 1$, and $a_{ij} = \psi$ for $i \neq j$.

Table 2 shows some results when $p = 8$, giving the coverages. Roughly need $n \geq 20ph$ to get good coverages near 0.95. Thus $n = 400$ was too small for $p = 8$ with $h = 3$ or $h = 4$, although undercoverage was small for $h = 3$. Note that $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, ..., \epsilon_{8t})^T$. The value $\psi = 0$ makes the ϵ_{it} uncorrelated. Increasing ψ increases the correlation $\rho = cor(\epsilon_{it}, \epsilon_{jt})$ where $i \neq j$. The prediction regions are hyperellipsoids, which have volumes (not given), instead of lengths.

Simulations for the Data Splitting Prediction Region

Table 3: Data Splitting Nominal 95% Prediction region

n	p	nv	xtype	dtype	cov
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	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

The theory for the new prediction regions is simple, so Table 3 is more of a check that the programs work than that the theory works. See Zhang (2022) for more simulations. 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 $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ where MED($\boldsymbol{W}$) is the coordinatewise median of the $\boldsymbol{x}_i$.

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.

Table 3 gives n, p, n_V, a number xtype corresponding to the distribution of $\boldsymbol{x}$, and a number dtype corresponding to $(T, \boldsymbol{C})$ used in prediction region (7). High dimensional data was used since $p \geq n$. With $n_V = 20$, the actual coverage is $20/21 = 0.9524$, $n_V = 25$ has actual coverage $25/26 = 0.9615$, and $n_V = 50$ has actual coverage $49/51 = 0.9608$. The observed coverages were close to the actual coverages in Table 3.

5 Discussion

The new nonparametric asymptotically optimal h-step ahead prediction intervals for the random walk appear to perform well if $n \geq 50h$. The new nonparametric h-step ahead 95% prediction regions for the vector valued random walk appear to have coverages near 0.95 for $n \geq 20ph$. The new nonparametric prediction regions are fast with simple theory, and have coverage $\geq \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil)/(n_V + 1)$.

Data sets where the future data does not behave like the past data are common, and then the prediction intervals and regions tend to perform poorly. In Example 1, cases 1-1460 appear to follow one random walk, while cases 1461-1800 follow another random walk with more variability.

Some prediction intervals for stochastic processes include Pan and Politis (2016), Vidoni (2004), and Vit (1973). Makridakis et al. (1987) noted that a PI for the random walk, derived assuming normal errors, often failed to give good coverage. Pankratz (1983, p. 106) noted that the random walk model has been found to be a good model for many stock price time series.

Conformal prediction gives precise levels of coverage for one future observation, and prediction region (7) is a conformal prediction region that can have large volume. 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 $\boldsymbol{w}_i \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$, but if the $\boldsymbol{w}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that highest density region is a hyperellipsoid tightly clustered about a vector in the direction of $\boldsymbol{1} = (1, 1, ..., 1)^T$, then the prediction region (7) has huge volume compared to the highest density region.

There are many methods where prediction is useful. For example, Garg, Aggarwal, et al. (2022) used support vector machines while Garg, Belarbi, et al. (2022) used Gaussian process regression. Olive (2018) shows how to get prediction intervals when the model

is $Y_i = m(\boldsymbol{x}_i) + e_i$ if the errors are iid. If heterogeneity is present, and there are enough cases $\boldsymbol{x}_i$ with $\hat{m}(\boldsymbol{x}_i)$ near $\hat{m}(\boldsymbol{x}_f)$, make a prediction interval using the Y_i corresponding to those $\boldsymbol{x}_i$. Graphically, in a plot of $\hat{m}(\boldsymbol{x}_i)$ versus Y_i (on the vertical axis), make a narrow vertical slice centered at $\hat{m}(\boldsymbol{x}_f)$, and then make the PI from the Y_i in the slice.

Plots and simulations were done in *R*. See R Core Team (2020). Programs are in the collection of functions *tspack.txt*. See (http://parker.ad.siu.edu/Olive/tspack.txt). Tables 1 and 2 used the functions `rwpisim` and `rwprsim` for the random walk simulations. The function `predsim2` simulates the data splitting prediction region for Table 3. The function `predrgn2` computes the prediction region (7) using $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. The pottery data is available from (http://parker.ad.siu.edu/Olive/sldata.txt).

REFERENCES

Beran, R. (1990), "Calibrating Prediction Regions," *Journal of the American Statistical Association*, 85, 715-723.

Beran, R. (1993), "Probability-Centered Prediction Regions," *The Annals of Statistics*, 21, 1967-1981.

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.

Einmahl, J.H.J., and Mason, D.M. (1992), "Generalized Quantile Processes," *The Annals of Statistics*, 20, 1062-1078.

Fontana, M., Zeni, G., and Vantini, S. (2023), "Conformal Prediction: a Unified Review of Theory and New Challenges," *Bernoulli*, 29, 1-23.

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

Garg, A., Aggarwal, P., Aggarwal, Y., Belarbi, M.O., Chalak, H.D., Tounsi, A., and Gulia, R. (2022), "Machine Learning Models for Predicting the Compressive Strength of Concrete Containing Nano Silica," *Computers and Concrete*, 30, 1, 33-42.

Garg, A., Belarbi, M.-O., Chalak, H.D., Tounsi, A., Li, L., Singh, A., and Mukhopadhyay, T. (2022), "Predicting Elemental Stiffness Matrix of FG Nanoplates Using Gaussian Process Regression Based Surrogate Model in Framework of Layerwise Model," *Engineering Analysis with Boundary Elements*, 143, 779-795.

Gray, H.L., and Odell, P.L., (1966), "On Sums and Products of Rectangular Variates," *Biometrika*, 53, 615-617.

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

Guan, L. (2023), "Localized Conformal Prediction: a Generalized Inference Framework for Conformal Prediction," *Biometrika*, 110, 33-50.

Haile, M.G. (2022), "Inference for Time Series after Variable Selection," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/shaile.pdf).

Hyndman, R.J. (1995), "Highest Density Forecast Regions for Non-Linear and Non-Normal Time Series Models, *Journal of Forecasting*, 14, 431-441.

Kim, J.H. (1999), "Asymptotic and Bootstrap Prediction Regions for Vector Autoregression," *International Journal of Forecasting*, 15, 393-403.

Kim J.H. (2004), "Bias-Corrected Bootstrap Prediction Regions for Vector Autoregression," *Journal of Forecasting*, 23 (2), 141-154.

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.

Makridakis, S., Hibon, M., Lusk, E., and Belhadjali, M. (1987), "Confidence Intervals: an Empirical Investigation of the Series in the M-Competition," *International Journal of Forecasting*, 3, 489-508.

Marengo, J.E., Farnsworth, D.L., and Stefanic, L. (2017), "A Geometric Derivation of the Irwin-Hall Distribution," *International Journal of Mathematics and Mathematical Sciences*, 2017, online.

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

Niwitpong, S., and Panichkitkosolkul, W. (2009), "Prediction Interval for an Unknown Mean Gaussian AR(1) Process Following Unit Root Test," *Journal of Management Science and Statistical Decisions*, 6, 43-51.

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

Pan, L., and Politis, D.N. (2016), "Bootstrap Prediction Intervals for Markov Processes," *Computational Statistics & Data Analysis*, 100, 467-494.

Panichkitkosolkul, W., and Niwitpong, S. (2011), "On Multistep-Ahead Prediction Intervals Following Unit Root Tests for a Gaussian AR(1) Process with Additive Outliers," *Applied Mathematical Sciences*, 5, 2297-2316.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, NY.

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. (2023), "Bootstrapping Some GLMs and Survival Regression Models after Variable Selection," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

Roach, S.A. (1963), "The Frequency Distribution of the Sample Mean Where Each Member of the Sample is Drawn from a Different Rectangular Distribution," *Biometrika*, 50, 508-513.

Ross, S.M. (2014), *Introduction to Probability Models*, 11th ed., Academic Press, San Diego, CA.

Steinberger, L., and Leeb, H. (2023), "Conditional Predictive Inference for Stable Algorithms," *The Annals of Statistics*, 51, 290-311.

Tian, Q., Nordman, D.J., and Meeker, W.Q. (2022), "Methods to Compute Prediction Intervals: a Review and New Results," *Statistical Science,* 37(4): 580-597.

Vidoni, P. (2004), "Improved Prediction Intervals for Stochastic Process Models," *Journal of Time Series Analysis*, 25, 137-154.

Vit, P. (1973), "Interval Prediction for a Poisson Process," *Biometrika*, 60, 667-668.

Wisseman, S.U., Hopke, P.K., and Schindler-Kaudelka, E. (1987), "Multielemental and Multivariate Analysis of Italian Terra Sigillata in the World Heritage Museum, University of Illinois at Urbana-Champaign," *Archeomaterials*, 1, 101-107.

Wolf, M., and Wunderli, D. (2015), "Bootstrap Joint Prediction Regions," *Journal of Time Series Analysis*, 36, 352-376.

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