Conformal prediction for multi-dimensional time series by ellipsoidal sets

Chen Xu [*1] **Hanyang Jiang** [*1] **Yao Xie** [1]

Abstract

Conformal prediction (CP) has been a popular method for uncertainty quantification because it is distribution-free, model-agnostic, and theoretically sound. For forecasting problems in supervised learning, most CP methods focus on building prediction intervals for univariate responses. In this work, we develop a sequential CP method called `MultiDimSPCI` that builds prediction *regions* for a multivariate response, especially in the context of multivariate time series, which are not exchangeable. Theoretically, we estimate *finite-sample* high-probability bounds on the conditional coverage gap. Empirically, we demonstrate that `MultiDimSPCI` maintains valid coverage on a wide range of multivariate time series while producing smaller prediction regions than CP and non-CP baselines.

1. Introduction

Conformal prediction (CP) has been a popular distribution-free technique to quantify uncertainty in modern machine learning (Volkhonskiy et al., 2017). In building predictive algorithms, CP can enhance trained machine learning estimators to output not just point estimates but also provide uncertainty sets that contain the unobserved ground truth with user-specified high probability. As a result, CP has been applied successfully to many applications, such as anomaly detection (Xu & Xie, 2021a), classification (Angelopoulos et al., 2021; Xu & Xie, 2022), regression (Barber et al., 2021), and so on. In a nutshell, CP methods work as wrappers that take in three components: a black-box predictive model f, an input feature X, and a potential output Y. Then, it designs a so-called "non-conformity" score that measures how *non-conforming* the potential output is. Naturally, the uncertainty set conditioning on the input feature and the predictor model would contain all potential outputs that are *conforming* to the past.

Most successful applications of CP have considered Y as an univariate variable. In the regression setting (Kim et al., 2020), CP methods thus output *prediction intervals* while in the classification setting (Romano et al., 2020), these methods produce *prediction sets*. With mild assumptions, such as assuming that data (X, Y) are exchangeable, these one-dimensional uncertainty sets can theoretically guarantee high coverage probability. Recent works have also extended such guarantees to non-exchangeable observations and either quantify the coverage gap in finite training samples (Barber et al., 2023) or show asymptotic coverage convergence (Xu & Xie, 2023b).

Despite the success of CP on scalar outputs Y, effective use of CP on multi-dimensional outputs is considerably less studied, *especially* when data are non-exchangeable as in the case of multivariate time-series forecasting. Moreover, there can be complex dependence between the multiple dimensions of the time series, making the problem more interesting and important. Specifically, the goal is not just to provide a prediction interval for each dimension of Y but to produce an uncertainty region that captures the correlation within Y and jointly contains all coordinates of Y. While uncertainty quantification methods for this problem have existed outside CP, as in vector auto-regressive models (Salinas et al., 2020), these approaches often have strong modeling or data assumption and lack rigorous theoretical justifications. On the other hand, various multi-dimensional CP methods have been proposed. Yet, they are either repeated use of one-dimensional CP methods (Stankevičiūtė et al., 2021) or fail to work beyond exchangeability (Messoudi et al., 2022).

We highlight the differences against copula-based CP methods, which have been developed for multi-dimensional forecasting. The initial approach developed in (Messoudi et al., 2021) assumed exchangeability, which is unsuitable for time series. The recent development by (Sun & Yu, 2024) proposes copula CP for multi-step time-series forecasting. However, their theory assumed that each data sample of an entire time series is drawn i.i.d. from an unknown distribution, hence ignoring temporal dependency. We further introduce copula and its use in CP in Section 3.2, with

[*]Equal contribution [1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. Correspondence to: Yao Xie <yao.xie@isye.gatech.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

additional comparison against baselines in Section 5.2.

Hence, the central focus of this work is to advance CP in the context of multivariate time-series forecasting. Specifically, we build ellipsoidal prediction sets whose size is adaptively and efficiently calibrated during test time. We also provide rigorous theoretical guarantees and extensive experiments to showcase the utility of the proposed method. In summary, our contributions are

- We propose a sequential conformal prediction method that builds ellipsoidal prediction regions for multivariate time series. In particular, sizes of the ellipsoids are sequentially re-estimated during test time to ensure adaptiveness and accuracy.
- We provide finite-sample high-probability bounds for the coverage gap of constructed prediction regions, which do not depend on the exchangeability of the observations.
- We empirically verify on multivariate time-series (up to dimension 20) that `MultiDimSPCI` can yield smaller prediction regions than baseline CP and non-CP methods without losing coverage.

For clarity, a taxonomy of existing CP methods is in Table 1 to highlight our role within the CP literature. In this paper, we assume that the noise sequence in the data-generating process (see Eq (11)) is stationary and strongly mixing, where the original data sequence does not need to be exchangeable. Meanwhile, we impose some standard assumptions on the tail behavior of the distribution to derive a non-asymptotic bound on the conditional guarantee. We highlight that our guarantees differ from existing multi-dimensional CP works that assume exchangeability (Messoudi et al., 2021; 2022) or i.i.d. data (Sun & Yu, 2024).

We adopt the standard notations. For a random process $\{X_n\}_n$, $X_n = o_p(a_n)$ means that $|X_n|/a_n \xrightarrow{p} 0$. For function $f(n)$ and $g(n)$, $f(n) = \tilde{O}(g(n))$ means that $f(n) = O(g(n)\log(g(n))^k)$ for some k. Besides, for event A and B, the notation $A\big|B$ means A under the condition B. For vector $u, v \in \mathbb{R}^p$, $u \otimes v$ is the outer product of u and v.

1.1. Literature review

CP with exchangeable data. The field of CP started in (Vovk et al., 2005) and has been widely used for uncertainty quantification due to its flexibility and robustness. On a high level, we define a "non-conformity" score and evaluate such scores on a hold-out calibration set. Then, uncertainty sets include all potential observations whose non-conformity scores are within the empirical quantiles of the calibration scores. Assuming nothing but that input data are exchangeable, CP methods have been successfully developed in different applications (Wisniewski et al., 2020; Xu & Xie, 2021a; 2022), in addition to the active research on proper designs of non-conformity scores (Angelopoulos et al., 2021; Huang et al., 2023; Gui et al., 2023). Comprehensive reviews of conformal prediction can be found in (Fontana et al., 2023; Angelopoulos & Bates, 2023).

CP for one-dimensional time series. Two general trends of extending beyond exchangeability work well for univariate Y. The first considers adaptively adjusting the significance level α during test time to account for mis-coverage. Such works include (Gibbs & Candes, 2021; Zaffran et al., 2022; Lin et al., 2022). The recent work (Angelopoulos et al., 2024) extends such framework through the lens of control theory to prospectively model non-conformity scores in online settings. The second considers weighing the past non-conformity score non-equally so that scores more similar to the present are given higher weights. Such works include (Tibshirani et al., 2019; Xu & Xie, 2021b; 2023b; Xu et al., 2023; Nair & Janson, 2023), some of which have successfully been applied to univariate time series. The recent work (Barber et al., 2023) also suggests that re-weighting can be a general scheme to account for non-exchangeability. Our `MultiDimSPCI` is similar to the second line of approaches but works in high dimensions.

CP for multi-dimensional data. Numerous works have been on this topic. (Stankevičiūtė et al., 2021) builds coordinate-wise prediction intervals for multi-horizon time-series prediction using Bonferroni correction of the significance level. For multivariate functional data, a similar idea of building prediction *bands* was studied in (Diquigiovanni et al., 2022), where this idea was further developed for time series (Diquigiovanni et al., 2021). In addition, (Messoudi et al., 2021) develops a principled way to determine the length of coordinate-wise intervals by using copula, resulting in hyper-rectangular prediction regions. The extension of copula for time-series forecasting was later studied in (Sun & Yu, 2024). However, it is important to note that the use of hyper-rectangles can be sub-optimal in many cases, even in the two-dimensional instances when the true conditional distribution $Y|X$ is $N(f(X), \Sigma)$ with a non-zero off-diagonal entry in Σ. To overcome this, (Messoudi et al., 2022) considers ellipsoidal uncertainty sets that rely on data exchangeability. A more exact quantification of the uncertainty set is studied in (Johnstone & Ndiaye, 2022), which, however, strongly depends on the underlying predictive model of Y. As a result, extending CP for multivariate time-series forecasting beyond using hyper-rectangles still needs to be explored.

Uncertainty quantification beyond CP. The task of building uncertainty set for unobserved response has been widely studied beyond CP. There has been a long history of using copula to capture the joint distribution of multivariate re-

Table 1. A 2×2 taxonomy of conformal prediction approaches (not an exhaustive list), categorized based on the dimension of the response variable Y (rows) and data assumptions (columns).

	Exchangeable	Non-exchangeable
Univariate Y	(Volkhonskiy et al., 2017) (Barber et al., 2021; Kim et al., 2020)	(Zaffran et al., 2022; Xu & Xie, 2023a) (Xu & Xie, 2023b; Barber et al., 2023)
Multivariate Y	(Messoudi et al., 2021; Diquigiovanni et al., 2022) (Johnstone & Ndiaye, 2022; Feldman et al., 2023)	**Ours** (Stankevičiūtė et al., 2021; Sun & Yu, 2024)

sponse by relating the joint cumulative distribution function (CDF) with each marginal CDF (Sklar, 1959; Elidan, 2013). Meanwhile, (Dobriban & Lin, 2023) uses conditional pivots to construct joint coverage regions for parameters and observations, extending beyond CP. However, its utility beyond exchangeable data remains unclear. On the other hand, there has been extensive development in the *probabilistic forecasting* literature, popular examples of which include the DeepAR (Salinas et al., 2020) and Temporal Fusion Transformer (Lim et al., 2021). Such approaches optimize (variants of) the pinball loss to estimate quantiles of multivariate responses but typically require extensive hyper-parameter tuning and return hyper-rectangular uncertainty sets. We will show experimentally that their performances are often worse than their CP counterparts. Lastly, (Feldman et al., 2023) uses CP in the representation space learnt by a deep generative model, allowing general prediction sets for multivariate data. Coverage guarantee for exchangeable data is proved, and extension to time series remains unexplored.

2. Problem setup

We consider a multi-dimensional time-series regression setup: for time index $t = 1, 2, \ldots$, assume observations $Z_t = (X_t, Y_t)$ are sequentially revealed, where $Y_t \in \mathbb{R}^p$ are p-dimensional vector variables and $X_t \in \mathbb{R}^d$ are d-dimensional features. The features X_t may be the history of Y_t or contain other variables that help predict Y_t. In particular, we allow arbitrarily unknown correlation among the observations Z_t. Let $\{Z_t\}_{t=1}^T$ be the training data.

Our goal is to sequentially construct prediction *regions* $\widehat{C}_{t-1}(X_t, \alpha)$ starting from $t = T+1$, which depends on past observations, the current feature X_t, and a user-specified significance level $\alpha \in [0, 1]$. In particular, we desire the prediction regions to contain the true observations Y_t with a probability at least $1 - \alpha$. Mathematically, there are two types of coverage guarantees to be satisfied by $\widehat{C}_{t-1}(X_t, \alpha)$. The first is the weaker *marginal* coverage:

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t, \alpha)) \geq 1 - \alpha, \forall t, \tag{1}$$

while the second is the stronger *conditional* coverage:

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t, \alpha) | X_t) \geq 1 - \alpha, \forall t. \tag{2}$$

If $\widehat{C}_{t-1}(X_t, \alpha)$ satisfies (1) or (2), it is called marginally or conditionally valid, respectively. When $\widehat{C}_{t-1}(X_t, \alpha)$ satisfies the coverage guarantees, we aim to build regions that are as small as possible to quantify uncertainty precisely.

3. Method

In this section, we first propose the ellipsoidal uncertainty set that effectively quantifies multi-dimensional prediction. We then discuss several benefits of the proposed approach against alternatives. We finally suggest alternative forms of the uncertainty set beyond using ellipsoids.

3.1. Ellipsoidal uncertainty set

We build the prediction regions in the shape of ellipsoids and calibrate the radius of ellipsoids using conformal prediction for univariate time series. Recall that we have access to T training data $Z_t = (X_t, Y_t)$ for $t = 1, \ldots, T$. Assume we have been given an algorithm $\hat{f}$, trained on a separate set I to perform point prediction for Y. Meanwhile, we collect *prediction* residuals $\hat{\varepsilon}_t \in \mathbb{R}^p$ $\hat{\varepsilon}_t = Y_t - \hat{f}(X_t),\ t = 1, \ldots, T$. This approach is similar to the split conformal prediction (Volkhonskiy et al., 2017) or leave-one-out techniques (Barber et al., 2021; Xu & Xie, 2023a).

To define the ellipsoidal uncertainty set, we first denote the covariance estimator over the prediction residuals as

$$\widehat{\Sigma} = \frac{1}{T-1} \sum_{t=1}^{T} (\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_t - \bar{\varepsilon})^T, \tag{3}$$

where $\bar{\varepsilon} = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t$ is the sample mean vector over the residuals. As the definition of an ellipsoid will rely on the inverse of $\widehat{\Sigma}$ in (3), which may not be invertible, we consider a low-rank approximation of $\widehat{\Sigma}$ as follows. Let the singular value decomposition of $\widehat{\Sigma}$ be $\widehat{\Sigma} = USV^T$, where $S = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of singular values satisfying $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, and U and V satisfy $UU^T = VV^T = I_p$. Given a small positive threshold $\rho > 0$, the low-rank approximation $\widehat{\Sigma}_\rho$ is

$$\widehat{\Sigma}_\rho = U_\rho S_\rho V_\rho^T, \tag{4}$$

where $S_\rho = \text{diag}(\lambda_1, \ldots, \lambda_k)$ for which $\lambda_k \geq \rho$, and U_ρ and V_ρ^T contain the first k columns and rows of U and V^T, respectively. The pseudo-inverse $\widehat{\Sigma}_\rho^{-1}$ is thus written as

$$\widehat{\Sigma}_\rho^{-1} = V_\rho S_\rho^{-1} U_\rho^T, \tag{5}$$

where $S_\rho^{-1} = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_k)$. Using $\widehat{\Sigma}_\rho^{-1}$, which is always well-defined, an ellipsoid with radius r is thus $\mathcal{B}(r, \bar{\varepsilon}, \widehat{\Sigma}_\rho) = \{x \in \mathbb{R}^p : (x - \bar{\varepsilon})^T \widehat{\Sigma}_\rho^{-1} (x - \bar{\varepsilon}) \leq r\}$.

We then find an appropriate radius r using time-series conformal prediction methods. First, given a new residual $\hat{\varepsilon} = Y - \hat{f}(X)$ and the pseudo-inverse $\widehat{\Sigma}_\rho^{-1}$ in (5), we define the scalar non-conformity score $\hat{e}(Y)$ as

$$\hat{e}(Y) = (\hat{\varepsilon} - \bar{\varepsilon})^T \widehat{\Sigma}_\rho^{-1} (\hat{\varepsilon} - \bar{\varepsilon}) \in \mathbb{R}. \tag{6}$$

We then compute the non-conformity scores on the training set (X_t, Y_t) for $t = 1, \ldots, T$ to obtain the set $\mathcal{E}_T = \{\hat{e}(Y_t)\}_{t=1}^T$. Note that non-conformity scores in $\mathcal{E}_T$ can be sequentially dependent due to the inherent dependency among the original data (X_t, Y_t). We take this into account by using `SPCI` (Xu & Xie, 2023b), a sequential conformal inference method for univariate time series. Specifically, rather than directly taking the *empirical* quantile over $\mathcal{E}_T$, we fit a *quantile regression estimator* $\widehat{Q}_t$ on $\mathcal{E}_T$, where $\widehat{Q}_t(\alpha)$ aims to predict the α-quantile of the unseen non-conformity score. There is no specific restriction on the quantile regression method used here. For example, `SPCI`(Xu & Xie, 2023b) uses the quantile random forest.

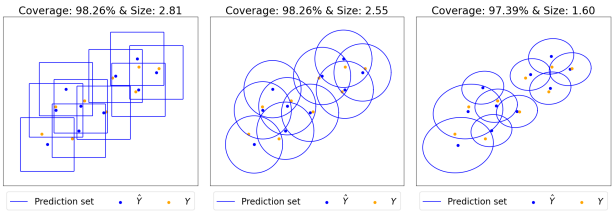
We first define the set difference of two sets A and B as $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$. Thus, the prediction set $\widehat{C}_{t-1}(X_t, \alpha) \subset \mathbb{R}^p$ for a given confidence level α is

$$\widehat{C}_{t-1}(X_t, \alpha) = \{Y : \widehat{Q}_t(\hat{\beta}) \leq \hat{e}(Y) \leq \widehat{Q}_t(1 - \alpha + \hat{\beta})\} \tag{7}$$

$$\hat{\beta} = \underset{\beta \in [0, \alpha]}{\arg\min}\, V(\widehat{\Sigma}_\rho, \widehat{Q}_t(1 - \alpha + \beta)) - V(\widehat{\Sigma}_\rho, \widehat{Q}_t(\beta)) \tag{8}$$

The prediction region (7) is further simplified as $\hat{f}(X_t) + \mathcal{B}(\sqrt{\widehat{Q}_t(1 - \alpha + \hat{\beta})}, \bar{\varepsilon}, \widehat{\Sigma}_\rho) \setminus \mathcal{B}(\sqrt{\widehat{Q}_t(\hat{\beta})}, \bar{\varepsilon}, \widehat{\Sigma}_\rho)$, which contains all Y such that their non-conformity scores $\hat{e}(Y)$ are within the respective quantiles of the ellipsoid centers at the prediction $\hat{f}(X_t)$. In (8), $V(\widehat{\Sigma}_\rho, r)$ denotes the volume of the ellipsoid with radius r, and we find $\hat{\beta}$ empirically as the tightest significance level at which the volume of the prediction region is smallest. Note that optimizing β further allows us to consider asymmetry in the distribution of non-conformity scores. When the optimal $\hat{\beta}$ is zero, it reduces to an ellipsoid as follows, which is also shown in Figure 1(c): $\widehat{C}_{t-1}(X_t, \alpha) = \{Y : \hat{e}(Y) \leq \widehat{Q}_t(1 - \alpha)\} = \hat{f}(X_t) + \mathcal{B}\left(\sqrt{\widehat{Q}_t(1 - \alpha)}, \bar{\varepsilon}, \widehat{\Sigma}_\rho\right)$ In Appendix A.1, we further discuss the use of local ellipsoids to define the prediction set $\widehat{C}_{t-1}(X_t, \alpha)$, which empirically can lead to improved performances.

We propose `MultiDimSPCI` in Algorithm 1 as a multi-dimensional generalization of the original `SPCI` (Xu & Xie, 2023b) method. The main benefits lie in the extension to quantify uncertainty in multi-dimensional prediction. The method we propose is simple and uses an ellipsoid uncertainty set. However, we will show later that this method can achieve conditional coverage. Besides, even though our method only uses the information of the first two moments, it outperforms the copula method, which is the state of the art. Figure 1 illustrates the benefit of `MultiDimSPCI` over existing methods in yielding smaller uncertainty sets for multi-dimensional UQ time series.



(a) (b) (c)

Figure 1. Comparison of multivariate CP method on real two-dimensional wind data (see Section 5.2). Left (a): Empirical copula (Messoudi et al., 2021) which constructs coordinate-wise prediction intervals. Middle (b): Spherical confidence set introduced in (Sun & Yu, 2024). Right (c): our proposed ellipsoidal confidence set via `MultiDimSPCI`. While all methods yield coverage at least above the target 95% on test data, our method yields the smallest average size.

Algorithm 1 Multi-dimensional `SPCI` (`MultiDimSPCI`)

Require: Training data $\{(X_t, Y_t)\}_{t=1}^T$, prediction algorithm $\mathcal{A}$, significance level α, quantile regression algorithm $\mathcal{Q}$, positive threshold $\rho > 0$.
Ensure: Prediction intervals $\widehat{C}_{t-1}(X_t, \alpha), t > T$
1: Obtain $\hat{f}$ and residuals $\{\hat{\varepsilon}_t\}_{t=1}^T \subset \mathbb{R}^p$ (computed on the holdout set) with $\mathcal{A}$ and $\{(X_t, Y_t)\}_{t=1}^T$
2: Compute non-conformity scores $\mathcal{E}_T$ from $\{\hat{\varepsilon}_t\}_{t=1}^T$ and $\widehat{\Sigma}_\rho$ using (6)
3: **for** $t > T$ **do**
4: Use quantile regression to obtain $\widehat{Q}_t \leftarrow \mathcal{Q}(\mathcal{E}_T)$
5: Obtain uncertainty set $\widehat{C}_{t-1}(X_t, \alpha)$ as in (7)
6: Obtain new residual $\hat{\varepsilon}_t$
7: Update residual set $\{\hat{\varepsilon}_t\}_{t=1}^T$ by adding $\hat{\varepsilon}_t$ and removing the oldest one and update $\mathcal{E}_T$
8: **end for**

3.2. Comparison with copula

We briefly introduce copula and explain how copula has been utilized in multivariate conformal prediction. We then highlight the key differences of the copula-based CP method with our `MultiDimSPCI`.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a generic p-dimensional continuous random vector with the joint CDF F and marginal CDFs F_j of X_j for $j = 1, \ldots, p$. We remark that in this

subsection, for notation convenience, the subscript $_j$ in X_j denotes the j-th component of $\boldsymbol{X}$ rather than the j-th feature vector of the original time series (i.e., X_j in the sequence $\{(X_t, Y_t)\}$). Define $U_j := F_j(X_j)$, where for $u \in [0,1]$, $\mathbb{P}(U_j \leq u) = \mathbb{P}(X_j \leq F_j^{-1}(u)) = F_j(F_j^{-1}(u)) = u$. Hence, $U_j \sim \text{Unif}[0,1]$ is a uniform random variable on $[0,1]$. Now, the joint CDF of $(U_1, \ldots, U_p)$ is the *copula* C of $(X_1, \ldots, X_p)$:

$$\begin{aligned} C(u_1, \ldots, u_p) &= \mathbb{P}(U_1 \leq u_1, \ldots, U_p \leq u_p) \\ &= \mathbb{P}(X_1 \leq F_1^{-1}(u_1), \ldots, X_p \leq F_p^{-1}(u_p)) \\ &= F(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)). \end{aligned} \tag{9}$$

Hence, the copula C links p marginal CDFs $\{F_j\}$ to the joint CDF F. For instance, consider bivariate Gaussian copula as an example, where we can explicitly write down the copula C. Let $(X_1, X_2) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \kappa$ for $\kappa \in [-1, 1]$. Then,

$$\begin{aligned} C(u_1, u_2) &= \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2) \\ &= \mathbb{P}(X_1 \leq \Phi^{-1}(u_1), X_2 \leq \Phi^{-1}(u_2)) \\ &= \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \kappa), \end{aligned} \tag{10}$$

where Φ is the CDF of $\mathcal{N}(0,1)$ and $\Phi_2(\cdot; \kappa)$ is the joint CDF of $\mathcal{N}(\mathbf{0}, \Sigma)$. Note that the bivariate Gaussian copula is parametric, assuming the marginal and joint distributions follow Gaussian distributions.

In conformal prediction, copula has been used to calibrate the coordinate-wise quantile of prediction residuals. Let $|\hat{\varepsilon}_{tj}|$ be the j-th coordinate of the t-th prediction residual in absolute value, and let F_{tj} be its marginal distribution. Then, past works (Messoudi et al., 2021) fit a copula C_t to the p-dimensional random vector $(|\hat{\varepsilon}_{t1}|, \ldots, |\hat{\varepsilon}_{tp}|)$. Specifically, they find $(u_{t1}, \ldots, u_{tp}) \in [0,1]^p$ so that $\mathbb{P}(|\hat{\varepsilon}_{t1}| \leq F_{t1}^{-1}(u_{t1}), \ldots, |\hat{\varepsilon}_{tp}| \leq F_{tp}^{-1}(u_{tp})) = C_t(u_{t1}, \ldots, u_{tp}) = 1 - \alpha$, where α is a pre-specified significance level (e.g., $\alpha = 0.05$). In practice, F_{tj} is unknown so it is replaced by $\hat{F}_{tj}$, the empirical distribution defined using past residuals, and the values $(u_{t1}, \ldots, u_{tp})$ are found under special assumptions (e.g., $u_{t1} = \ldots = u_{tp}$ (Messoudi et al., 2021)) or searched via stochastic gradient descent (Sun & Yu, 2024).

We remark two main differences between copula conformal prediction and our proposed `MultiDimSPCI`. First, the use of copula CP requires searching for multi-dimensional vectors $\boldsymbol{u}_t = (u_{t1}, \ldots, u_{tp})$ at each t, whose efficiency and accuracy also highly depends on the choice of copula C_t. How to design copula and search for the best $\boldsymbol{u}_t$ remains unclear. In contrast, our `MultiDimSPCI` requires much less design effort, as it only uses an estimation of the covariance matrix of residuals $\{\hat{\varepsilon}_t\}$. Second, note that copula CP returns *hyper-rectangular* prediction sets, as the method constructs one prediction interval at each p coordinates. Such hyper-rectangular sets can be too large compared to ellipsoidal sets, as we experimentally find ours are significantly smaller without affecting test coverage (see Section 5.2).

3.3. Benefits of the proposed approach

We further discuss the benefits of `MultiDimSPCI` against other approaches.

Against coordinate-wise use of `SPCI` (Xu & Xie, 2023b): Rather than building ellipsoidal uncertainty sets, a naive but perhaps more intuitive approach is to apply `SPCI` p times, once per dimension of $Y \in \mathbb{R}^p$. The resulting uncertainty sets are hyper-rectangles, which can be unnecessarily large in many cases. In addition, the significance values for `SPCI` at different dimensions need to be adjusted appropriately to achieve valid coverage of Y. Computationally, such use of `SPCI` is also more expensive than `MultiDimSPCI` because $(p-1)T_1$ additional quantile regression models are fitted (T_1 is the length of the test set).

Against copula-based CP methods (Messoudi et al., 2021; Sun & Yu, 2024): Besides the limitation above of returning hyper-rectangular uncertainty sets, these copula-based methods fail to account for the sequential dependency of non-conformity scores when taking the empirical quantile over scores. In contrast, the proposed `MultiDimSPCI` explicitly takes dependency into account by adaptively re-estimating the quantile of non-conformity scores.

Against probabilistic forecasting methods (Salinas et al., 2020; Lim et al., 2021): There are two main benefits of `MultiDimSPCI`. First, our proposed method is compatible with any user-specified prediction model $\hat{f}$ of Y. In contrast, such probabilistic forecasting methods often require specifically designed deep neural networks to predict the quantiles of Y directly. Second, we can provide coverage guarantees for the proposed method, whereas those methods often lack sound justifications.

4. Theoretical Analysis

In this section, we will present theoretical results for bounding the conditional coverage of our method. The result is based on the case where the sample covariance matrix is invertible. We first recall and define the notations and then give out the assumptions required, which are general and identifiable. After that, we will present our coverage guarantees when using the empirical quantile function as the quantile regression predictor. The norm $\|\cdot\|$ used in the paper is the spectral norm (2-norm). The proof details will be in Appendix C. The main idea of the proof consists of two parts. The first part is the convergence of the empirical CDF to the true CDF of the residual. The second part is to control the estimation error of the sample covariance matrix.

We assume that $Y_t \in \mathbb{R}^p$ is generated from a true model with unknown additive noise:

$$Y_t = f(X_t) + \varepsilon_t, \quad t = 1, 2, \ldots, \tag{11}$$

where f is an unknown function and ε_t represents the process noise, whose marginal distribution is not necessarily Gaussian, the process noise may have temporal dependence. For the simplicity of notation, we assume $\bar{\varepsilon} = 0$ here so that the non-conformity score simplifies to $\hat{\varepsilon}_t^T \widehat{\Sigma}^{-1} \hat{\varepsilon}_t$. Besides, without loss of generality, we can assume $\mathbb{E}[\varepsilon] = 0$. Otherwise, we can subtract the mean from the noises and add to the function f.

We define $\hat{\varepsilon}_t = y_t - \hat{f}(x_t)$ as the *vector prediction residual*, and the *scalar non-conformity score* $\hat{e}_t = \hat{\varepsilon}_t^T \widehat{\Sigma}^{-1} \hat{\varepsilon}_t \in \mathbb{R}$. Moreover, $e_t = \varepsilon_t^T \Sigma^{-1} \varepsilon_t, \quad \Delta_t = \hat{\varepsilon}_t - \varepsilon_t$. Here $\varepsilon_t \in \mathbb{R}^p$ is the true noise in model (11) and $\Sigma \in \mathbb{R}^{p \times p}$ is the true covariance matrix of ε. Besides, we define the empirical CDF $\widehat{F}_{T+1}(x) = \frac{1}{T} \sum_{t=1}^{T} 1\{\hat{e}_t \le x\}$, $\widetilde{F}_{T+1}(x) = \frac{1}{T} \sum_{t=1}^{T} 1\{e_t \le x\}$.

We also use $F_e(x) = \mathbb{P}\{e \le x\}$ to represent the CDF of the nonconformity score. Since we consider the case where the marginal distributions of e_t are identical, their CDF is the same, and we can define it as $F_e(x)$. In our method, we use the empirical distribution of non-conformity score $\hat{e}$ to approximate the distribution of e. A new observation Y_{T+1} being covered by the conformal interval with given coverage is equivalent to $\hat{e}_{T+1}$ falling in a given quantile in empirical distribution $\widehat{F}_{T+1}$.

From the property of CDF, we know that $F_e(e_{T+1}) \sim \mathrm{Unif}[0, 1]$. If we can show that $\widehat{F}_{T+1}$ approximates F_e well, then it will follow that $\widehat{F}_{T+1}$ approximately covers a region of $1 - \alpha$ probability. Comparing $\hat{e}$ and e, we see that

$$\hat{e}_t - e_t = \hat{\varepsilon}_t^T \widehat{\Sigma}^{-1} \hat{\varepsilon}_t - \varepsilon_t^T \Sigma^{-1} \varepsilon_t, \tag{12}$$

where $\widehat{\Sigma}$ is estimated from $\{\hat{\varepsilon}_t\}_{t=1}^T$. We would need $\hat{\varepsilon}_t$ to be close to ε_t. Otherwise, this approximation would not hold.

Assumption 4.1 (i.i.d. and Lipschitz)**.** Assume $\{\varepsilon_t\}_{t=1}^{T+1}$ are independent and identically distributed (i.i.d.). Meanwhile, $F_e(x)$ (the CDF of the true non-conformity score) is assumed to be Lipschitz continuous with constant $L_{T+1} > 0$.

Remark 4.2. We first assume that the error process $\{\varepsilon_t\}_{t=1}^{T+1}$ is i.i.d. In fact, this assumption is not necessary, and we will extend this assumption to cases beyond exchangeability. The result for stationary and strong mixing sequences will be presented in Corollary 4.18.

Assumption 4.3 (Estimation quality)**.** There exists a sequence $\{\delta_T\}_{T \ge 1}$ such that

$$\frac{1}{T} \sum_{t=1}^{T} \|\Delta_t\|^2 \le \delta_T^2, \ \ \|\Delta_{T+1}\| \le \delta_T. \tag{13}$$

Remark 4.4. The assumption requires that the square sum of the prediction error be bounded by δ_T^2. For many estimators, there exists a sequence $\{\delta_T\}_{T \ge 1}$ that goes to zero. For example, $\delta_T = o_p(T^{-1/4})$ for general neural networks sieve estimators when f is sufficiently smooth (Chen & White, 1999). When f is a sparse high-dimensional linear model, $\delta_T = o_p(T^{-1/2})$ for the Lasso estimator and Dantzig selector (Bickel et al., 2009).

Assumption 4.5 (Covariance eigenvalue)**.** There exists a $\lambda > 0$ s.t. $\|\Sigma\| \ge \lambda$ and $\|\widehat{\Sigma}\| \ge \lambda$.

Remark 4.6. It is a common assumption to require both the covariance matrix and the estimated covariance matrix to be strictly positive definite. The condition holds for the covariance matrix as long as there is no linear dependency between variables, which is true for the errors ε_t. Besides, the assumption is also satisfied by the sample covariance matrix because our algorithm uses the pseudo-inverse, which ensures positive eigenvalues.

Assumption 4.7 (Tail behavior)**.** There exist some constants $q > 4$, $K_1, K_2 > 0$ and $L \ge 1$, such that $\max_{t \le T} \|\varepsilon_t\| \le \sqrt{K_1 p}$ almost surely and $\mathbb{E}|\langle \varepsilon, x \rangle|^q \le L^q$ for $x \in S^{p-1}$. There exists a constant K_2 such that $\mathrm{Var}[\|\varepsilon\|^2] \le K_2 p$.

Remark 4.8. The assumption is required in Theorem 1.2 (Vershynin, 2012) so that the sample covariance matrix converges to the true covariance matrix in the operator norm. Other assumptions in literature ensure a $O(T^{-1/2})$ convergence. For example, (Koltchinskii & Lounici, 2017) requires random variables ε to be weakly square-integrable, sub-Gaussian, and pre-Gaussian. Our method can use covariance estimators other than the classic sample covariance matrix. The sample covariance matrix is a natural choice, but it is a poor estimator when the dimension is very high unless there are some nice tail behaviors (Lugosi & Mendelson, 2019). There are a lot of results in the literature focusing on covariance estimation under different conditions. The Assumption 4.7 can be easily switched to other requirements if we change the estimator. (Cai et al., 2016) offers an overview of covariance estimators with their optimal rates. The choice and analysis of the covariance matrix is not the main focus of our paper, so we use the sample covariance matrix here for simplicity.

With the i.i.d. assumption, we can show that the empirical distribution of e_t approximates the true CDF well in the following sense.

Lemma 4.9 (Convergence of empirical CDF of $\{\varepsilon_t\}_{t=1}^T$ under i.i.d.)**.** *Under Assumption 4.1, for any training size T, there is an event A_T which occurs with probability at least $1 - \sqrt{\frac{\log(16T)}{T}}$, such that conditioning on A_T,*

$$\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \le \sqrt{\frac{\log(16T)}{T}}. \tag{14}$$

Remark 4.10. The i.i.d. assumption is not a must, and we can easily extend it to the case where $\{e_t\}_{t=1}^T$ is stationary and strong mixing. We will show a similar result in Corollary C.11.

With the assumptions, we can also show that the empirical distribution of $\hat{e}$ approximates the empirical distribution of e well in the following sense:

Lemma 4.11 (Distance between the empirical CDF of $\{\varepsilon_t\}_{t=1}^T$ and $\{\hat{\varepsilon}_t\}_{t=1}^T$ under i.i.d.)**.** *Under Assumption 4.1, 4.3, 4.5 and 4.7, with a high probability* $1-\delta$*,*

$$\begin{aligned}&\sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \\ \leq &(L_{T+1}+1)C_S + 2\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)|,\end{aligned}$$

where $C_S = (\frac{\delta_T^2}{\lambda} + \frac{K_3}{\lambda^2}C(\frac{1}{\delta})^{20/9q}\log(\frac{1}{\delta})(\log\log p)^2 \frac{p^{3/2-2/q}}{T^{1/2-2/q}} + 22K_3\max\{\frac{p^3}{T\delta}, p^{3/2}\delta_T\}])^{1/2}$ *and* C *is a constant that depends only on* K_1, q, L *and* $K_3 = K_1 + \sqrt{3K_2}$ *is a constant. Thus*

$$C_S = \tilde{O}\left(\max\left\{\frac{p^{3/4-1/q}}{T^{1/4-1/q}}, p^{3/4}\delta_T^{1/2}\right\}\right). \tag{15}$$

Our main theorem is the following Theorem 4.12, which establishes the asymptotic conditional coverage as a result of Lemma 4.9 and 4.11.

Theorem 4.12 (Conditional guarantee under i.i.d. assumption)**.** *Under Assumption 4.1, 4.3, 4.5 and 4.7, with a high probability* $1-\delta$*, for any training size* T *and* $\alpha \in (0,1)$*, we have*

$$\begin{aligned}&|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1} = x_{T+1}) - (1-\alpha)| \\ \leq &12\sqrt{\frac{\log(16T)}{T}} + 4(L_{T+1}+1)(C_S + \delta_T),\end{aligned} \tag{16}$$

where C_S *is defined in* (15)*.*

Remark 4.13. The bound is controlled by the sample size T and the coefficient δ_T. This vanishes when $T \to \infty$ and $\delta_T \to 0$, which means that when the sample size is large enough, and the estimator $\hat{f}$ is accurate enough, the conditional coverage will converge to $1-\alpha$.

Corollary 4.14 (Guarantee with true covariance matrix, and under i.i.d.)**.** *If the true covariance matrix* Σ *is known, we can use* $\widehat{\Sigma} = \Sigma$*. Under Assumption 4.1, 4.3, 4.5 and 4.7, for any training size* T *and* $\alpha \in (0,1)$*, we have*

$$\begin{aligned}&|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1} = x_{T+1}) - (1-\alpha)| \\ \leq &12\sqrt{\frac{\log(16T)}{T}} + 4(L_{T+1}+1)\left(\frac{\delta_T}{\sqrt{\lambda}} + \delta_T\right).\end{aligned} \tag{17}$$

Remark 4.15. When the true covariance matrix is known, we have a tighter and simpler bound than Theorem 4.12. Although the true covariance is usually unknown in reality, the same bound can be reached if the sample covariance matrix is estimated from another independent training set.

Then, we would present a corollary that extends our guarantee to the case where $\{\varepsilon_t\}_{t=1}^T$ is a stationary and strong mixing sequence.

Definition 4.16. A sequence of random variables $\{X_n\}$ is said to be *strictly stationary* if for every $k \geq 1$, any integers $n_1, \ldots, n_k$, and any integer h, the joint distribution of the random variables $(X_{n_1}, \ldots, X_{n_k})$ is the same as the joint distribution of $(X_{n_1+h}, \ldots, X_{n_k+h})$.

Definition 4.17. A sequence of random variables $\{X_n\}$ is said to be *strongly mixing* (or α-mixing) if the mixing coefficients α_k defined by $\alpha_k = \sup_{n\in\mathbb{N}} \sup_{A\in\mathcal{F}_1^n, B\in\mathcal{F}_{n+k}^{\infty}} |P(A\cap B) - P(A)P(B)|$ tend to zero as $k \to \infty$, where $\mathcal{F}_a^b$ denotes the σ-algebra generated by $\{X_a, \ldots, X_b\}$.

Using a similar technique, we can prove the following result for the case where $\{\varepsilon_t\}_{t=1}^T$ is a stationary and strong mixing sequence. Here, we assume that the true covariance matrix Σ is known for simplicity, but we will discuss in Remark 4.19 how to extend the result when true covariance is unknown.

Corollary 4.18 (Guarantee with true covariance matrix, under stationarity and strong mixing)**.** *Assume* $\{\varepsilon_t\}_{t=1}^T$ *is a stationary and strong mixing sequence with mixing coefficient* $0 < \sum_{k>0}\alpha_k < M$*. Under Assumption 4.3, 4.5 and 4.7, for any training size* T *and* $\alpha \in (0,1)$*, we have*

$$\begin{aligned}&|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1} = x_{T+1}) - (1-\alpha)| \\ \leq &12\frac{(\frac{M}{2})^{1/3}(\log T)^{2/3}}{T^{1/3}} + 4(L_{T+1}+1)\left(\frac{\delta_T}{\sqrt{\lambda}} + \delta_T\right)\end{aligned} \tag{18}$$

Remark 4.19. The first term in the convergence rate is of order $\tilde{O}(T^{-1/3})$, which is slighter bigger than the order $\tilde{O}(T^{-1/2})$ in Theorem 4.12 under i.i.d. case. The second term is the same. The essence of generalizing the outcome to scenarios where the true covariance matrix Σ remains unknown lies in the convergence properties of the sample covariance matrix. There are works directed towards these convergence properties within the context of stationary time series. For instance, (Chen et al., 2013) presented an asymptotic convergence result for a threshold sample covariance estimator. Utilizing methodologies akin to those employed in Theorem 4.12, a similar result can be substantiated, and there is also the possibility to apply a variety of estimators. However, the focus of this work is not on the convergence of the covariance matrix estimator, which is left as future works. Moreover, as previously indicated, independent training data can be leveraged to estimate the covariance matrix and achieve the same bound in Corollary 4.18.

Table 2. Simulation results by both methods. Target coverage is 90%. Standard deviation is computed over ten independent trials in which training and test data are regenerated.

(a) Independent AR(w)

p	2		4		8		10		16		20	
Method	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI
Coverage	90.0% (0.26)	90.0% (0.29)	90.0% (0.25)	89.9% (0.14)	90.0% (0.31)	89.9% (0.30)	89.8% (0.25)	89.8% (0.27)	89.9% (0.24)	89.9% (0.23)	90.0% (0.26)	89.8% (0.30)
Size	1.45e+1 (9.34e-2)	1.52e+1 (8.73e-2)	3.00e+2 (2.62e+0)	3.94e+2 (3.38e+0)	1.30e+5 (1.43e+3)	3.68e+5 (6.44e+3)	2.65e+6 (4.79e+4)	1.22e+7 (1.61e+5)	2.23e+10 (5.61e+8)	5.84e+11 (1.39e+10)	9.15e+12 (2.97e+11)	8.67e+14 (2.90e+13)

(b) VAR(w)

p	2		4		8		10		16		20	
Method	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI	MultiDim SPCI	SPCI
Coverage	90.0% (0.26)	92.7% (0.25)	90.2% (0.21)	91.5% (0.22)	90.0% (0.23)	91.6% (0.18)	89.9% (0.23)	90.7% (0.31)	89.9% (0.20)	91.0% (0.19)	90.0% (0.25)	90.9% (0.19)
Size	2.73e+0 (1.36e-2)	6.46e+0 (5.84e-2)	3.89e+1 (2.25e-1)	4.94e+2 (7.49e+0)	7.16e+4 (7.25e+2)	9.27e+6 (1.46e+5)	3.63e+7 (4.79e+5)	3.24e+9 (6.09e+7)	8.55e+12 (1.45e+11)	1.91e+17 (5.38e+15)	1.14e+16 (2.11e+14)	7.41e+22 (1.68e+21)

5. Experiments

We test Algorithm 1 on both simulated and real-time series to show that `MultiDimSPCI` reaches valid coverage with smaller prediction regions. In all our experiments, the value of ρ used in (4) is set to be 0.001. For simplicity, we only consider the global covariance matrix in (3) rather than its local variant (A.1), which would bring further improvements. Code is available at `https://github.com/hamrel-cxu/MultiDimSPCI`.

5.1. Simulation

We simulate two types of stationary time series. The first case considers independent AR(w) sequences, and the second case considers VAR(w) sequences. We want to show that compared to `SPCI` applied independently to each dimension (equivalent to using independent copula (Messoudi et al., 2021, See Sec. 3.3.1)), `MultiDimSPCI` yields significantly smaller prediction regions with valid coverage.

Data generation. Denote $Y_t = [Y_{i1}, \ldots, Y_{ip}]^T \in \mathbb{R}^p$ for $p \geq 2$. We generate Y_t as

$$Y_t = \sum_{l=1}^{w} \boldsymbol{\alpha}_l Y_{i-l} + \boldsymbol{\varepsilon}_t, \;\; \boldsymbol{\varepsilon}_t \sim N(0, \Sigma). \tag{19}$$

In (19), $\boldsymbol{\alpha}_l \in \mathbb{R}^{p\times p}$ contains the set of coefficients, where we further construct them so that the sequences $\{Y_t\}$ are stationary. In the first case of independent AR(w) sequences, we have $\Sigma = I_p$. In the second case of VAR(w) sequences, we design $\Sigma = BB^T$ to be a positive definite covariance matrix, where $B_{ij} \overset{i.i.d.}{\sim} \text{Unif}[-1, 1]$.

Setup. In both cases of AR and VAR time series following (19), we let $w = 5$ and vary $p \in \{2, 4, 8, 10, 16, 20\}$. The initial 80K samples $\{Y_t\}$ are training data; the remaining 20K samples are test data. Because `SPCI` assumes independence across different univariate sequence, we let $\tilde{\alpha} = 1-(1-\alpha)^{1/p}$ and apply `SPCI` on individual sequences with the corrected $\tilde{\alpha}$. The multivariate linear regression method is used as the point predictor.

Results. Table 2 examines the empirical coverage and average size of prediction regions in $\mathbb{R}^p$ by both methods on the two cases of data generation. Both methods can maintain valid coverage around the target 90% in two cases. Nevertheless, it is clear that as dimension p increases, the average size of prediction regions by the proposed `MultiDimSPCI` is significantly smaller (for several magnitudes) than that by `SPCI` applied to individual sequences. In Figure A.1, we visualize the non-critical regions in both cases to demonstrate why `MultiDimSPCI` provides smaller prediction regions.

5.2. Real-data

We now compare `MultiDimSPCI` with existing methods designed for multivariate uncertainty quantification. The three CP baselines are CopulaCPTS (Sun & Yu, 2024), Local ellipsoid (Messoudi et al., 2022), and Copula (Messoudi et al., 2021). The two probabilistic forecasting baselines are temporal fusion transformers (TFT) (Lim et al., 2021) and DeepAR (Salinas et al., 2020).

We consider three real multivariate time-series datasets. The first *wind* dataset considers wind speed in meters per second at different wind farms (Zhu et al., 2021), with 764 observations in total. The second *solar* dataset considers solar radiation in Diffused Horizontal Irradiance (DHI) units at different solar sensors (Zhang et al., 2021), with 8755 observations in total. The third *traffic* dataset considers traffic flow collected at different traffic sensors (Xu & Xie, 2021a), with 8778 observations in total. On each dataset, we randomly select $p \in \{2, 4, 8\}$ locations (same for all methods) and examine the test coverage and average size on the p-dimensional time series. The first 85% data are used

Table 3. Real-data comparison of test coverage and average prediction set size by different methods. The target coverage is 0.95, and at each p, the smallest size of prediction sets is in **bold**. Our `MultiDimSPCI` yields the narrowest confidence sets without sacrificing coverage for two reasons. First, it explicitly captures dependency among coordinates of Y_t by forming ellipsoidal prediction sets. Second, it captures temporal dependency among non-conformity scores upon adaptive re-estimation of score quantiles.

(a) Wind data

Method	$p = 2$ coverage	$p = 2$ size	$p = 4$ coverage	$p = 4$ size	$p = 8$ coverage	$p = 8$ size
`MultiDimSPCI`	0.97	**1.60**	0.96	**7.02**	0.96	**72.10**
CopulaCPTS (Sun & Yu, 2024)	0.98	2.55	0.97	10.23	0.97	252.67
Local ellipsoid (Messoudi et al., 2022)	0.96	3.51	0.97	13.07	0.98	1.09e+3
Copula (Messoudi et al., 2021)	0.98	2.81	0.98	10.32	0.97	1.60e+3
TFT (Lim et al., 2021)	0.94	10.61	0.75	159.39	0.94	2.91e+4
DeepAR (Salinas et al., 2020)	0.96	7.07	0.76	67.97	0.96	1.79e+5

(b) Solar data

Method	$p = 2$ coverage	$p = 2$ size	$p = 4$ coverage	$p = 4$ size	$p = 8$ coverage	$p = 8$ size
`MultiDimSPCI`	0.96	1.68	0.96	**2.89**	0.97	**4.97**
CopulaCPTS (Sun & Yu, 2024)	0.99	4.36	0.99	37.56	0.99	3.28e+3
Local ellipsoid (Messoudi et al., 2022)	0.97	**1.32**	0.97	3.20	0.97	43.07
Copula (Messoudi et al., 2021)	0.99	4.11	0.99	27.73	0.99	1.42e+3
TFT (Lim et al., 2021)	0.99	13.68	0.99	71.72	0.93	1.19e+3
DeepAR (Salinas et al., 2020)	0.97	10.76	0.98	157.09	0.74	31.82

(c) Traffic data

Method	$p = 2$ coverage	$p = 2$ size	$p = 4$ coverage	$p = 4$ size	$p = 8$ coverage	$p = 8$ size
`MultiDimSPCI`	0.96	**1.31**	0.96	**1.93**	0.96	**2.98**
CopulaCPTS (Sun & Yu, 2024)	0.95	1.70	0.94	3.15	0.95	14.10
Local ellipsoid (Messoudi et al., 2022)	0.95	1.36	0.94	2.08	0.95	4.13
Copula (Messoudi et al., 2021)	0.95	1.44	0.95	3.90	0.94	40.60
TFT (Lim et al., 2021)	0.89	9.07	0.93	87.92	0.88	9.69e+2
DeepAR (Salinas et al., 2020)	0.87	13.53	0.88	57.20	0.82	9.89e+3

for training, and the remaining 15% are used for testing.

Table 3 shows that the test coverage of `MultiDimSPCI` and two CP methods is always valid by yielding coverage greater than or equal to the target 95%. In contrast, the two probabilistic forecasting baselines may incur severe under-coverage, where TFT coverage is generally better than DeepAR's. Regarding the average size of prediction regions, we also note that the average size by `MultiDimSPCI` is consistently smaller than those by baselines (except against Local ellipsoid on solar data when $p = 2$), demonstrating that our proposed method quantifies prediction uncertainty more precisely. We believe these benefits come from using ellipsoidal rather than hyper-rectangular prediction sets and the adaptive re-estimation of quantiles of non-conformity scores. Additionally, Table A.3 shows comparisons on higher-dimensional wind data, on which the benefits of `MultiDimSPCI` persist. Figure A.2 further analyzes the rolling performance of different methods. We see that the rolling coverage of `MultiDimSPCI` and the CP baselines all center around the target 95% coverage value with reasonably small variations. Meanwhile, `MultiDimSPCI` has a smaller rolling width than the CP baselines, indicating that our proposed method almost always yields smaller prediction regions. Lastly, as seen in Figure A.3, the estimated correlation between residuals from two different locations can be as high as 0.92 (see solar data). Thus, it is indeed necessary to consider such correlation when constructing prediction regions to quantify prediction uncertainty effectively.

6. Discussions

In this work, we proposed `MultiDimSPCI`, a general sequential conformal prediction method for multivariate time series. Specifically, `MultiDimSPCI` extends sequential univariate CP method to construct ellipsoids during test time. Extensions using local ellipsoids that are adaptive in shape are also discussed. Theoretically, we bound the coverage gap in finite samples without assuming data exchangeability. Empirically, on both simulation and real time series, we show `MultiDimSPCI` yields significantly smaller prediction sets than baselines and maintains coverage.

In the future, we will explore constructing prediction regions beyond ellipsoids. One such possibility is using convex hulls, which are irregular in shape but could lead to the tightest fit as prediction sets. We discuss such possibility and preliminary results in Appendix A.2). We will also further study the theoretical properties of CP in high dimensions, leveraging existing results on multivariate quantile estimation.

Acknowledgement

The authors would like to thank Jonghyeok Lee and Bo Dai for their helpful discussions and comments. This work is partially supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, DMS-1830210, and the Coca-Cola Foundation.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, especially in the context of uncertainty quantification. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Angelopoulos, A., Candes, E., and Tibshirani, R. J. Conformal pid control for time series prediction. *Advances in Neural Information Processing Systems*, 36, 2024.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL http://dx.doi.org/10.1561/2200000101.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Cai, T. T., Ren, Z., and Zhou, H. H. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10:1–59, 2016.

Chen, X. and White, H. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

Chen, X., Xu, M., and Wu, W. B. Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021, 2013.

Diquigiovanni, J., Fontana, M., and Vantini, S. Distribution-free prediction bands for multivariate functional time series: an application to the italian gas market. *arXiv preprint arXiv:2107.00527*, 2021.

Diquigiovanni, J., Fontana, M., and Vantini, S. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Dobriban, E. and Lin, Z. Joint coverage regions: Simultaneous confidence and prediction sets. *arXiv preprint arXiv:2303.00203*, 2023.

Elidan, G. Copulas in machine learning. In *Copulae in Mathematical and Quantitative Finance: Proceedings of the Workshop Held in Cracow, 10-11 July 2012*, pp. 39–60. Springer, 2013.

Feldman, S., Bates, S., and Romano, Y. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Fontana, M., Zeni, G., and Vantini, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Gui, Y., Barber, R., and Ma, C. Conformalized matrix completion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6f320HfMeS.

Huang, K., Jin, Y., Candes, E., and Leskovec, J. Uncertainty quantification over graph with conformalized graph neural networks. *NeurIPS*, 2023.

Johnstone, C. and Ndiaye, E. Exact and approximate conformal inference in multiple dimensions. *arXiv preprint arXiv:2210.17405*, 2022.

Kim, B., Xu, C., and Barber, R. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020.

Koltchinskii, V. and Lounici, K. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=PM5gVmG2Jj.

Lugosi, G. and Mendelson, S. Near-optimal mean estimators with respect to general norms. *Probability theory and related fields*, 175(3-4):957–973, 2019.

Messoudi, S., Destercke, S., and Rousseau, S. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

Messoudi, S., Destercke, S., and Rousseau, S. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pp. 294–306. PMLR, 2022.

Nair, Y. and Janson, L. Randomization tests for adaptively collected data. *arXiv preprint arXiv:2301.05365*, 2023.

Rio, E. et al. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Sklar, M. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pp. 229–231, 1959.

Stankevičiūtė, K., Alaa, A., and van der Schaar, M. Conformal time-series forecasting. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Rx9dBZaV_IP.

Sun, S. H. and Yu, R. Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ojIJZDNIBj.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, pp. 2530–2540, 2019.

Vershynin, R. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Volkhonskiy, D., Burnaev, E., Nouretdinov, I., Gammerman, A., and Vovk, V. Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pp. 132–153. PMLR, 2017.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wisniewski, W., Lindsay, D., and Lindsay, S. Application of conformal prediction interval estimations to market makers' net positions. In Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., and Cherubin, G. (eds.), *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pp. 285–301. PMLR, 09–11 Sep 2020.

Xu, C. and Xie, Y. Conformal anomaly detection on spatio-temporal observations with missing data. *arXiv preprint arXiv:2105.11886*, 2021a. ICML 2021 Workshop on Distribution-Free Uncertainty Quantification.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021b.

Xu, C. and Xie, Y. Conformal prediction set for time-series. *arXiv preprint arXiv:2206.07851*, 2022. ICML 2022 Workshop on Distribution-Free Uncertainty Quantification.

Xu, C. and Xie, Y. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Xu, C. and Xie, Y. Sequential predictive conformal inference for time series. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38707–38727. PMLR, 23–29 Jul 2023b.

Xu, C., Xie, Y., Vazquez, D. A. Z., Yao, R., and Qiu, F. Spatio-temporal wildfire prediction using multi-modal data. *IEEE Journal on Selected Areas in Information*

Theory, 4:302–313, 2023. doi: 10.1109/JSAIT.2023.3276054.

Zaffran, M., Dieuleveut, A., F'eron, O., Goude, Y., and Josse, J. Adaptive conformal predictions for time series. In *ICML*, 2022.

Zhang, M., Xu, C., Sun, A., Qiu, F., and Xie, Y. Solar radiation ramping events modeling using spatio-temporal point processes. *arXiv preprint arXiv:2101.11179*, 2021.

Zhu, S., Zhang, H., Xie, Y., and Van Hentenryck, P. Multi-resolution spatio-temporal prediction with application to wind power generation. *arXiv preprint arXiv:2108.13285*, 2021.

A. Additional technical details

A.1. Improvements using local ellipsoids

In practice, constructing the scalar non-conformity scores in (6) based on a global covariance matrix in (3) fails to capture local variation in data. We can improve our `MultiDimSPCI` through using local ellipsoids proposed in (Messoudi et al., 2022).

Specifically, given a test data feature X_t for $t > T$, we first consider its k nearest neighbors among previous T samples $\{X_{t-1}, X_{t-2}, \ldots, X_{t-T}\}$. Let the index set of neighbors be N_t with $|N_t| = k$. Then, we denote $\widehat{\mathrm{Cov}}_t$ as the sample covariance estimator using $\{\hat{\varepsilon}_t\}_{t\in N_t}$ (see Eq. (3) for the definition using $\{\hat{\varepsilon}_t\}_{t=1}^T$). As a result, given a parameter $\lambda \in [0, 1]$, the local covariance estimator at time t is written as the weighted average

$$\widehat{\Sigma}_t = \lambda \widehat{\mathrm{Cov}}_t + (1-\lambda)\widehat{\Sigma}, \tag{A.1}$$

where $\widehat{\Sigma}$ is the global empirical covariance matrix in (3). We recommend setting $k = 0.1T$ and $\lambda = 0.95$ to capture local variations effectively. Lastly, as $\widehat{\Sigma}_t$ in (A.1) may not be invertible, we use its low-rank approximation in (4) and the corresponding pseudo-inverse in (5), which would be used to compute the non-conformity score (6) at time t. We empirically find that compared to using (3), the use of (A.1) can lead to up to 25% reduction in the average size of prediction sets.

A.2. Prediction set using convex hulls

Currently, the ellipsoid shape is utilized for the prediction region. This method is robust and guarantees coverage accuracy, as demonstrated by experiments. However, considering the distribution shape of the residuals may further enhance its performance. The distribution might not conform to an ellipsoidal shape in high-dimensional cases, potentially being irregular. As a result, the ellipsoidal shape may not be the most optimal or tight fit. What if we could allow our prediction set to adapt to any shape? In doing so, the new region would likely be much tighter in scenarios where the true residuals do not follow an ellipsoidal distribution.

In almost all instances, a convex hull can cover a set of points more compactly than an ellipsoid. We could achieve a significantly tighter fit by adopting a convex hull for the prediction region. The primary distinction between the two methods lies in the control parameters: the ellipsoid requires only the radius adjustment, whereas the convex hull necessitates control over all vertices. Ideally, we would select a set of data points that optimally balances coverage and minimizes the region size. However, this becomes computationally infeasible as the dataset size increases. Rather than optimizing the convex hull, we require it to cover exactly all the training data encompassed by the ellipsoid method.

The experimental results for time series in $\mathbb{R}^2$ and $\mathbb{R}^4$ are presented in Appendix B. With the same training size, the convex hull method produces a significantly smaller prediction region but with lower than required coverage. This issue can be mitigated by using a larger training dataset, allowing the convex hull to approximate the true distribution more accurately. However, computing a convex hull in high dimensions involves a worst-case time complexity of $O(T^{\lfloor p/2 \rfloor})$ using standard methods (Barber et al., 1996). Another challenge arises as dimensions increase: the convex hull method demands a substantially larger training dataset to capture the distribution, which may be impractical adequately. This aspect will be explored in future research.

B. Additional experiments

Comparison of non-critical regions. From Figure A.1, we see that `MultiDimSPCI` almost always yields smaller prediction sets than the coordinate-wise use of `SPCI`, whose prediction regions are squares that tend to over-cover the test samples. In contrast, `MultiDimSPCI` can well capture the dependency within Y to enable accurate uncertainty quantification.

Results using convex hulls. The convex hull method selects the points in the training set that are covered by the ellipsoid method and uses the convex hull of these points as the prediction regions. It has a smaller region than the ellipsoid method because of how it is constructed. As shown in Table A.1, the convex hull method on time series in $\mathbb{R}^2$ reaches valid coverage with a smaller prediction set. However, as shown in Table A.2, getting a region with valid coverage in higher dimensions requires much more training data. The computational cost in higher dimensions becomes unaffordable if we want to reach a reasonable coverage.

Table A.1. Accuracy and size of the prediction sets on two independent $AR(w)$ sequences.

	MultiDimSPCI	Copula	Convex hull	Baseline
Coverage	90.2	90.6	90.0	89.8
Size	3.92	3.59	3.64	3.58

C. Proof

Lemma C.1. $F_e(e_{T+1}) \sim \text{Unif}[0,1]$.

Proof. This holds for random variable e as long as the CDF F_e is continuous and strictly increasing. □

Lemma C.2. *Under Assumption 4.1, for any training size T, there is an event A_T which occurs with probability at least $1-\sqrt{\log(16T)/T}$, such that conditioning on A_T,*

$$\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \le \sqrt{\frac{\log(16T)}{T}}. \tag{A.2}$$

Proof. The proof follows the proof of Lemma 1 in (Xu & Xie, 2023a). When the error process is i.i.d., the famous Dvoretzky-Kiefer-Wolfowitz inequality (Kosorok, 2008) implies that

$$\mathbb{P}\left(\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| > s_T\right) \le 2e^{-2Ts_T^2}. \tag{A.3}$$

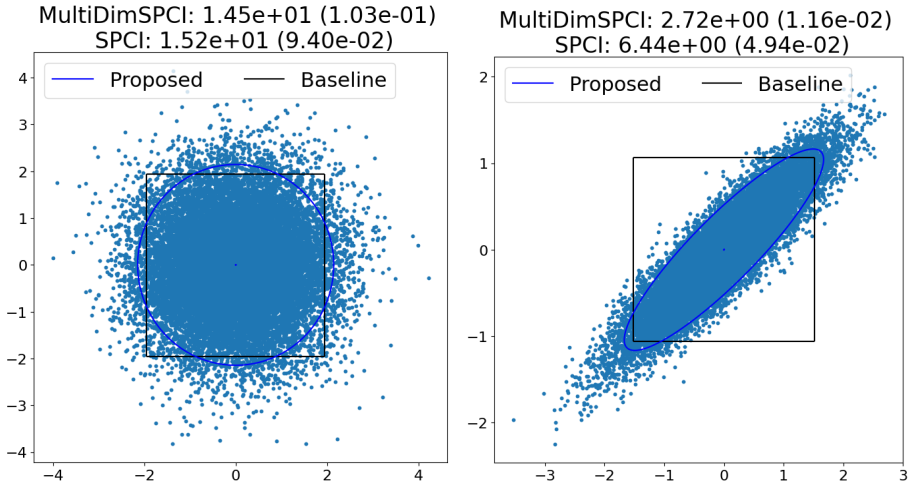


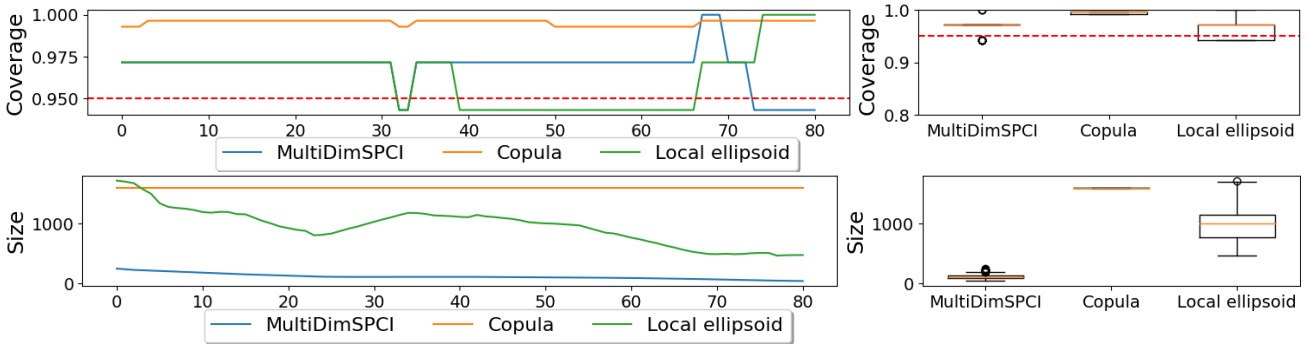
Figure A.1. Non-critical regions on independent $\text{AR}(w)$ (left) and $\text{VAR}(w)$ (right) in $\mathbb{R}^2$. The average size of prediction regions is shown in captions. The size of the non-critical region by the proposed method is smaller, especially on $\text{VAR}(w)$. As a result, the prediction regions by `MultiDimSPCI` are smaller than those by `SPCI`.

Table A.2. Accuracy and size of the prediction set for error uniformly spread in $[-1,1]^4$.

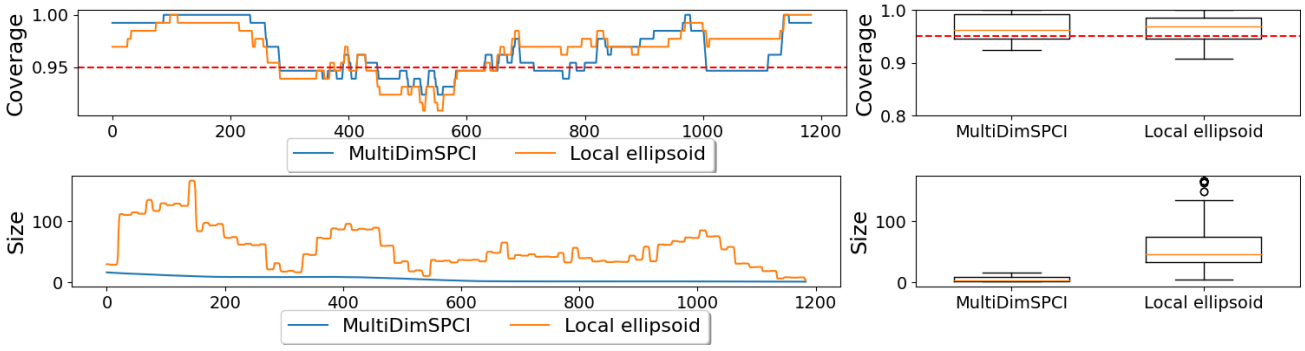
	MultiDimSPCI	Copula	Convex hull	Baseline
Coverage (80000 training samples)	89.6	90.5	85.9	89.9
Size (80000 training samples)	22.2	14.5	14.2	14.4
Coverage (800000 training samples)	90.2	90.5	88.6	89.9
Size (800000 training samples)	22.2	14.5	14.3	14.4

Table A.3. Comparison on wind data when dimension $d = 25$. The setup is identical to that in Table 3.

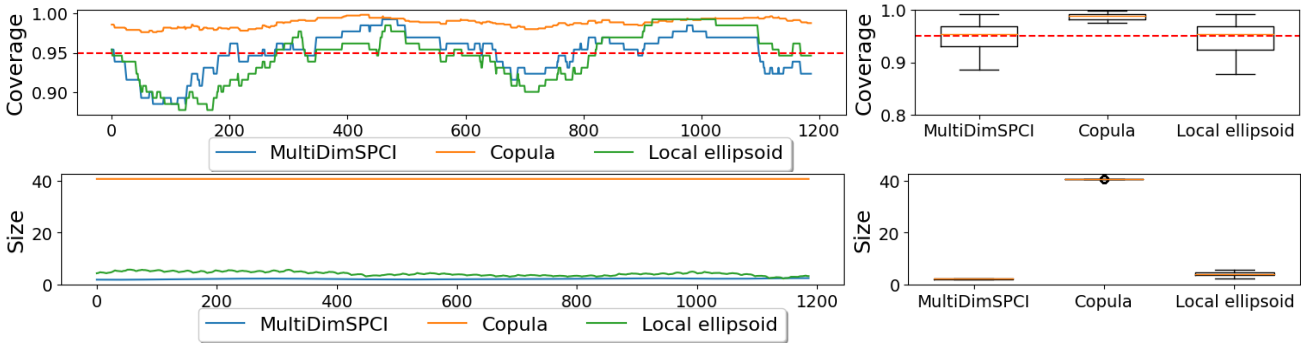
	`MultiDimSPCI`	CopulaCPTS	Local ellipsoid	Copula
Coverage	0.95	0.98	0.98	0.94
Size	3.55e+7	1.13e+14	4.93e+16	1.20e+13



(a) Wind data



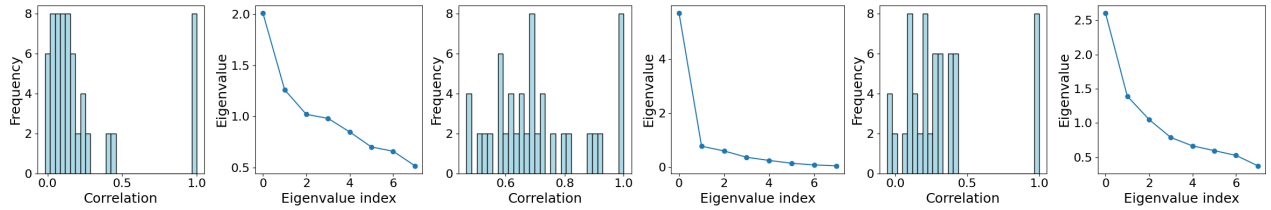
(b) Solar data



(c) Traffic data

Figure A.2. Real-data comparison of rolling coverage (target coverage is 95%) and size of prediction sets at $p = 8$. In each subplot of (a)-(c), the top row plots rolling coverage over prediction time indices (red dashed line is the target coverage) and as boxplots, and the bottom row shows results for rolling sizes. We only visualize the comparison of `MultiDimSPCI` with selected CP baselines, which have comparable average size of prediction regions in Table 3.

Pick $s_T = \sqrt{W(16T)}/(2\sqrt{T})$, where $W(T)$ is the Lambert W function that satisfies $W(T)e^{W(T)} = T$. We see that



(a) Wind data (b) Solar data (c) Traffic data

Figure A.3. Distribution of estimated correlation and the eigenvalues of corresponding correlation matrices on real-time series. We visualize the results using p-dimensional prediction residuals with $p = 8$.

$s_T \leq \sqrt{\log(16T)/T}$. Thus, define the event A_T on which $\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \leq \sqrt{\log(16T)/T}$, whereby we have

$$\begin{aligned} \sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \big| A_T \leq \sqrt{\frac{\log(16T)}{T}}, \\ P(A_T) > 1 - \sqrt{\frac{\log(16T)}{T}}. \end{aligned} \tag{A.4}$$

□

Lemma C.3 (Theorem 1.2, (Vershynin, 2012))**.** *Consider a random vector $\varepsilon \in \mathbb{R}^p (p \geq 4)$ which has zero mean and satisfies moment assumption 4.7 for some $q > 4$ and some K_1, L. Let $\delta > 0$. Then, with probability at least $1 - \delta$, the covariance matrix Σ of ε can be approximated by the sample covariance matrix $\frac{1}{T}\sum_{t=1}^T \varepsilon_t \otimes \varepsilon_t$ as*

$$\left\| \Sigma - \frac{1}{T}\sum_{t=1}^T \varepsilon_t \otimes \varepsilon_t \right\| \leq C \left(\frac{1}{\delta}\right)^{20/9q} \log\left(\frac{1}{\delta}\right) (\log\log p)^2 \left(\frac{p}{T}\right)^{1/2-2/q}, \tag{A.5}$$

where C is a constant that depends only on parameters q, K_1, L.

Lemma C.4. *Under Assumption 4.1, 4.3, 4.5 and 4.7, with high probability $1 - \delta$,*

$$\sum_{t=1}^T |\hat{e}_t - e_t| \leq \frac{T}{\lambda}\delta_T^2 + \frac{K_3 T}{\lambda^2}\left[C\left(\frac{1}{\delta}\right)^{20/9q} \log\left(\frac{1}{\delta}\right) (\log\log p)^2 \frac{p^{3/2-2/q}}{T^{1/2-2/q}} + 22K_3 \max\left\{\frac{p^3}{T\delta}, p^{3/2}\delta_T\right\}\right], \tag{A.6}$$

where C is a constant that depends only on parameters q, K_1, L and $K_3 = K_1 + \sqrt{3K_2}$.

Proof. For any test conformity score $\hat{e}$ and the corresponding e, we drop subscript t here for notation simplicity.

$$\begin{aligned} |\hat{e} - e| &= |\hat{\varepsilon}^T \widehat{\Sigma}^{-1}\hat{\varepsilon} - \varepsilon^T \Sigma^{-1}\varepsilon| \\ &\leq |\hat{\varepsilon}^T \widehat{\Sigma}^{-1}\hat{\varepsilon} - \varepsilon^T \widehat{\Sigma}^{-1}\varepsilon| + |\varepsilon^T \widehat{\Sigma}^{-1}\varepsilon - \varepsilon^T \Sigma^{-1}\varepsilon| \\ &\leq \|\widehat{\Sigma}^{-1}\|\|\Delta\|^2 + |\varepsilon^T(\widehat{\Sigma}^{-1} - \Sigma^{-1})\varepsilon| \\ &\leq \frac{1}{\lambda}\|\Delta\|^2 + \|\varepsilon\|^2 \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\| \\ &= \frac{1}{\lambda}\|\Delta\|^2 + \|\varepsilon\|^2 \|\widehat{\Sigma}^{-1}(\Sigma - \widehat{\Sigma})\Sigma^{-1}\| \\ &\leq \frac{1}{\lambda}\|\Delta\|^2 + \|\varepsilon\|^2 \|\widehat{\Sigma}^{-1}\|\|\Sigma^{-1}\|\|\Sigma - \widehat{\Sigma}\| \\ &\leq \frac{1}{\lambda}\|\Delta\|^2 + \frac{1}{\lambda^2}\|\varepsilon\|^2 \|\Sigma - \widehat{\Sigma}\|. \end{aligned} \tag{A.7}$$

Then the problem becomes bounding the spectral norm $\|\Sigma - \widehat{\Sigma}\|$. Recall that $\hat{\varepsilon}_t = \varepsilon_t + \Delta_t$, and we define $\bar{\varepsilon}_* = (\sum_{t=1}^T \hat{\varepsilon}_t)/T$,

$\bar{\Delta} = (\sum_{t=1}^{T} \Delta_t)/T$. The sample covariance matrix can be represented as

$$\begin{aligned}
\widehat{\Sigma} &= \frac{1}{T-1}\sum_{t=1}^{T}(\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_t - \bar{\varepsilon})^T \\
&= \frac{1}{T-1}\sum_{t=1}^{T}[(\varepsilon_t + \Delta_t) - (\bar{\varepsilon}_* + \bar{\Delta})][(\varepsilon_t + \Delta_t) - (\bar{\varepsilon}_* + \bar{\Delta})]^T \\
&= \frac{1}{T-1}\sum_{t=1}^{T}[(\varepsilon_t - \bar{\varepsilon}_*) + (\Delta_t - \bar{\Delta})][(\varepsilon_t - \bar{\varepsilon}_*) + (\Delta_t - \bar{\Delta})]^T \\
&= \underbrace{\frac{1}{T-1}\sum_{t=1}^{T}(\varepsilon_t - \bar{\varepsilon}_*)(\varepsilon_t - \bar{\varepsilon}_*)^T}_{①} + \underbrace{\frac{1}{T-1}\sum_{t=1}^{T}(\Delta_t - \bar{\Delta})(\Delta_t - \bar{\Delta})^T}_{②} + \\
&\quad \underbrace{\frac{1}{T-1}\sum_{t=1}^{T}\left[(\varepsilon_t - \bar{\varepsilon}_*)(\Delta_t - \bar{\Delta})^T + (\Delta_t - \bar{\Delta})(\varepsilon_t - \bar{\varepsilon}_*)^T\right]}_{③}.
\end{aligned} \tag{A.8}$$

The first term is the sample covariance matrix of ε_t, which typically converges to the true covariance matrix when the dimension p is negligible compared to T. From the assumption 4.3, the magnitude of the second term and the third term should be bounded by δ_T, which is the accuracy of prediction. This means that

$$\|\widehat{\Sigma} - \Sigma\| \leq \|① - \Sigma\| + \|②\| + \|③\|. \tag{A.9}$$

We can bound each spectral norm respectively.

$$\begin{aligned}
\|① - \Sigma\| &= \left\| \frac{1}{T-1}\left(\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^T - T\bar{\varepsilon}_*\bar{\varepsilon}_*^T\right) - \Sigma \right\| \\
&= \left\| \left(\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^T - \Sigma\right) + \frac{1}{T(T-1)}\left(\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^T - T^2\bar{\varepsilon}_*\bar{\varepsilon}_*^T\right) \right\| \\
&\leq \left\| \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^T - \Sigma \right\| + \frac{1}{T(T-1)}\left(\left\|\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^T\right\| + \|T^2\bar{\varepsilon}_*\bar{\varepsilon}_*^T\|\right) \\
&\overset{(i)}{\leq} C\left(\frac{3}{\delta}\right)^{\frac{20}{9q}}\log\left(\frac{3}{\delta}\right)(\log\log p)^2\left(\frac{p}{T}\right)^{\frac{1}{2}-\frac{2}{q}} + \frac{1}{T(T-1)}\left(\sum_{t=1}^{T}\|\varepsilon_t\varepsilon_t^T\| + T^2\|\bar{\varepsilon}_*\bar{\varepsilon}_*^T\|\right) \\
&\overset{(ii)}{\leq} 4C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}}\log\left(\frac{1}{\delta}\right)(\log\log p)^2\left(\frac{p}{T}\right)^{\frac{1}{2}-\frac{2}{q}} + \frac{1}{T(T-1)}\sum_{t=1}^{T}\|\varepsilon_t\|^2 + 2\|\bar{\varepsilon}_*\|^2,
\end{aligned}$$

where (i) holds with high probability $1 - \frac{\delta}{3}$ under Assumption 4.7 according to the Lemma C.3 from Theorem 1.2 in (Vershynin, 2012) and (ii) holds because $3^{20/9q} \leq 2$, $\log(3/\delta) \leq 2\log(1/\delta)$ when $\delta \leq 1/3$ and $T/(T-1) \leq 2$. We can put the constant 4 into the constant C which simplifies the notation. From now on, we use C to represent the whole constant $4C$ in the expression. For the other term $\|\bar{\varepsilon}_*\|^2 = \|(\sum_{t=1}^{T}\varepsilon_t)/T\|^2$, we can bound it with Chebshev's inequality. Since $\varepsilon_t \in \mathbb{R}^p$, we use $\varepsilon_{ti}(1 \leq i \leq p)$ to denote the ith entry of random vector ε_t.

$$\|\bar{\varepsilon}_*\|^2 = \left\|\frac{\sum_{t=1}^{T}\varepsilon_t}{T}\right\|^2 = \sum_{i=1}^{p}\left|\frac{\sum_{t=1}^{T}\varepsilon_{ti}}{T}\right|^2. \tag{A.10}$$

Using Chebshev inequality, we have that

$$\mathbb{P}\left(\left|\frac{\sum_{t=1}^{T}\varepsilon_{ti}}{T}\right| \geq \sqrt{\frac{3p\mathrm{Var}(\varepsilon_{ti})}{T\delta}}\right) \leq \frac{\mathrm{Var}(\varepsilon_{ti})}{T(\frac{3p\mathrm{Var}(\varepsilon_{ti})}{T\delta})} = \frac{\delta}{3p} \tag{A.11}$$

This means that

$$\begin{aligned}\mathbb{P}\left(\|\bar{\varepsilon}_*\|^2 \leq \frac{3p\sum_{i=1}^p \mathrm{Var}(\varepsilon_{ti})}{T\delta}\right) &= \mathbb{P}\left(\sum_{i=1}^p \left|\frac{\sum_{t=1}^T \varepsilon_{ti}}{T}\right|^2 \leq \frac{3p\sum_{i=1}^p \mathrm{Var}(\varepsilon_{ti})}{T\delta}\right) \\ &\geq \mathbb{P}\left(\bigcap_{i=1}^p \left\{\left|\frac{\sum_{t=1}^T \varepsilon_{ti}}{T}\right|^2 \leq \frac{3p\mathrm{Var}(\varepsilon_{ti})}{T\delta}\right\}\right) \\ &\geq \left(1-\frac{\delta}{3p}\right)^p \geq 1-\frac{\delta}{3}.\end{aligned} \tag{A.12}$$

Considering that

$$\sum_{i=1}^p \mathrm{Var}(\varepsilon_{ti}) = \mathbb{E}(\|\varepsilon_t\|^2) \leq K_1 p.$$

We have that with probability higher than $1-\delta/3$,

$$\begin{aligned}\|① - \Sigma\| &\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2\left(\frac{p}{T}\right)^{\frac{1}{2}-\frac{2}{q}} + \frac{1}{T(T-1)}\sum_{t=1}^T \|\varepsilon_t\|^2 + 2\|\bar{\varepsilon}_*\|^2 \\ &\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2\left(\frac{p}{T}\right)^{\frac{1}{2}-\frac{2}{q}} + \frac{1}{T(T-1)}\sum_{t=1}^T \|\varepsilon_t\|^2 + \frac{6K_1p^2}{T\delta}.\end{aligned} \tag{A.13}$$

For the second term, we have

$$\begin{aligned}\|② \| &= \left\|\frac{1}{T-1}\left(\sum_{t=1}^T \Delta_t\Delta_t^T - T\bar{\Delta}\bar{\Delta}^T\right)\right\| \\ &\leq \frac{1}{T-1}\left(\left\|\sum_{t=1}^T \Delta_t\Delta_t^T\right\| + T\|\bar{\Delta}\bar{\Delta}^T\|\right) \\ &\leq \frac{2}{T-1}\sum_{t=1}^T \|\Delta_t\|^2 \\ &\leq \frac{2T\delta_T^2}{T-1}.\end{aligned} \tag{A.14}$$

For the third term, we have

$$\begin{aligned}
\|③\| &= \frac{1}{T-1}\left\|\sum_{t=1}^{T}(\varepsilon_t\Delta_t^T+\Delta_t\varepsilon_t^T)-T(\bar{\varepsilon}_*\bar{\Delta}^T+\bar{\Delta}\bar{\varepsilon}_*^T)\right\| \\
&\le \frac{2}{T-1}\left(\left\|\sum_{t=1}^{T}\varepsilon_t\Delta_t^T\right\|+T\|\bar{\Delta}\|\|\bar{\varepsilon}_*\|\right) \\
&= \frac{2}{T-1}\left(\left\|\sum_{t=1}^{T}\varepsilon_t\Delta_t^T\right\|+\sqrt{\left(\frac{\|\sum_{t=1}^{T}\varepsilon_t\|^2}{T}\right)\left(\frac{\|\sum_{t=1}^{T}\Delta_t\|^2}{T}\right)}\right) \\
&\le \frac{2}{T-1}\left(\sum_{t=1}^{T}\|\varepsilon_t\|\|\Delta_t\|+\sqrt{\left(\sum_{t=1}^{T}\|\varepsilon_t\|^2\right)\left(\sum_{t=1}^{T}\|\Delta_t\|^2\right)}\right) \\
&\overset{(i)}{\le} \frac{4}{T-1}\sqrt{\left(\sum_{t=1}^{T}\|\varepsilon_t\|^2\right)\left(\sum_{t=1}^{T}\|\Delta_t\|^2\right)} \\
&\le \frac{4\delta_T}{T-1}\sqrt{T\sum_{t=1}^{T}\|\varepsilon_t\|^2}.
\end{aligned}$$

The inequality (i) holds because of Cauchy-Schwarz inequality $(\sum_{t=1}^{T}\|\varepsilon_t\|^2)(\sum_{t=1}^{T}\|\Delta_t\|^2)\ge(\sum_{t=1}^{T}\|\varepsilon_t\|\|\Delta_t\|)^2$. From Assumption 4.7, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\varepsilon_t\|^2\right]=T\mathbb{E}(\|\varepsilon_t\|^2)\le TK_1p. \tag{A.15}$$

Using Chebshev's inequality we have

$$\mathbb{P}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\|\varepsilon_t\|^2-\mathbb{E}[\|\varepsilon_t\|^2]\right|\ge\sqrt{\frac{3\mathrm{Var}[\|\varepsilon_t\|^2]}{T\delta}}\right\}\le\frac{\mathrm{Var}[\|\varepsilon_t\|^2]}{T(\frac{3\mathrm{Var}[\|\varepsilon_t\|^2])}{T\delta})}=\frac{\delta}{3}, \tag{A.16}$$

which means with probability higher than $1-\delta/3$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\varepsilon_t\|^2 &\le \mathbb{E}[\|\varepsilon_t\|^2]+\sqrt{\frac{3\mathrm{Var}[\|\varepsilon_t\|^2]}{T\delta}} \\
&\le K_1p+\sqrt{\frac{3K_2p}{T\delta}} \\
&\le (K_1+\sqrt{3K_2})p\coloneqq K_3p.
\end{aligned} \tag{A.17}$$

The last inequality holds because of Assumption 4.7 and $pT\delta\ge1$. Without loss of generality, we can assume $K_3\ge1$. Define $S_T=\frac{1}{T}\sum_{t=1}^{T}\|\varepsilon_t\|^2$. Overall, with probability higher than $1-\delta$, we have inequality A.17 and the following

inequality holds when $T \geq p$

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\| &\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + \frac{1}{T(T-1)}\sum_{t=1}^{T}\|\varepsilon_t\|^2 \\
&\quad + \frac{6K_1p^2}{T\delta} + \frac{2T\delta_T^2}{T-1} + \frac{4\delta_T}{T-1}\sqrt{T\sum_{t=1}^{T}\|\varepsilon_t\|^2} \\
&= 2C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{3}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + \frac{2}{T-1}S_T + \frac{6K_1p^2}{T\delta} + \frac{2T\delta_T^2}{T-1} + \frac{4T\delta_T}{T-1}\sqrt{S_T} \\
&\overset{(i)}{\leq} C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + \frac{4}{T}S_T + \frac{6K_1p^2}{T\delta} + 4\delta_T^2 + 8\delta_T\sqrt{S_T} \\
&\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + 16\max\left\{\frac{S_T}{T}, \delta_T^2, \delta_T\sqrt{S_T}\right\} + \frac{6K_1p^2}{T\delta} \\
&\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + 16\max\left\{\frac{K_3p}{T}, \delta_T^2, \sqrt{K_3p}\delta_T\right\} + \frac{6K_3p^2}{T\delta} \\
&\leq C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + 22K_3\max\left\{\frac{p}{T}, \delta_T^2, \sqrt{p}\delta_T, \frac{p^2}{T\delta}\right\} \\
&\overset{(ii)}{\leq} C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + 22K_3\max\left\{\frac{p^2}{T\delta}, \sqrt{p}\delta_T\right\}.
\end{aligned}$$

where (i) holds because $\frac{T}{T-1} \leq 2$ and (ii) holds because $p \geq \delta_T^2$. Then the following inequality holds with high probability $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}|\hat{e}_t - e_t| &\leq \frac{1}{\lambda}\sum_{t=1}^{T}\|\Delta_t\|^2 + \frac{1}{\lambda^2}\sum_{t=1}^{T}\|\varepsilon_t\|^2\|\widehat{\Sigma} - \Sigma\| \\
&\leq \frac{T}{\lambda}\delta_T^2 + \frac{K_3Tp}{\lambda^2}\left[C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2(\frac{p}{T})^{\frac{1}{2}-\frac{2}{q}} + 22K_3\max\left\{\frac{p^2}{T\delta}, \sqrt{p}\delta_T\right\}\right] \qquad \text{(A.18)} \\
&= \frac{T}{\lambda}\delta_T^2 + \frac{K_3T}{\lambda^2}\left[C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2\frac{p^{3/2-2/q}}{T^{1/2-2/q}} + 22K_3\max\left\{\frac{p^3}{T\delta}, p^{3/2}\delta_T\right\}\right].
\end{aligned}$$

□

Corollary C.5. *If the true covariance matrix Σ is known and we use $\widehat{\Sigma} = \Sigma$, then*

$$\sum_{t=1}^{T}|\hat{e}_t - e_t| \leq \frac{T}{\lambda}\delta_T^2. \qquad \text{(A.19)}$$

Proof. This is because when we use $\hat{\Sigma} = \Sigma$, the second term in Equation A.18 is zero. Then the bound is simply the first term. □

Lemma C.6. *Under Assumption 4.1, 4.3, 4.5 and 4.7, with a high probability $1-\delta$,*

$$\sup_x|\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \leq (L_{T+1}+1)C_S + 2\sup_x|\widetilde{F}_{T+1}(x) - F_e(x)|, \qquad \text{(A.20)}$$

where

$$C_S = \left(\frac{\delta_T^2}{\lambda} + \frac{K_3}{\lambda^2}\left[C\left(\frac{1}{\delta}\right)^{\frac{20}{9q}} \log\left(\frac{1}{\delta}\right)(\log\log p)^2\frac{p^{3/2-2/q}}{T^{1/2-2/q}} + 22K_3\max\left\{\frac{p^3}{T\delta}, p^{3/2}\delta_T\right\}\right]\right)^{1/2}$$

and $K_3 = K_1 + \sqrt{3K_2}$.

Proof. Using lemma C.4 and 4.3, we have that with probability $1-\delta$

$$\begin{aligned}\sum_{t=1}^{T}|e_t-\hat{e}_t| &\le \frac{T}{\lambda}\delta_T^2+\frac{K_3T}{\lambda^2}\left[C\left(\frac{1}{\delta}\right)^{20/9q}\log\left(\frac{1}{\delta}\right)(\log\log p)^2\frac{p^{3/2-2/q}}{T^{1/2-2/q}}+22K_3\max\left\{\frac{p^3}{T\delta},p^{3/2}\delta_T\right\}\right] \\ &= TC_S^2.\end{aligned} \tag{A.21}$$

Let $S=\{t:|e_t-\hat{e}_t|\ge C_S\}$. Then

$$|S|C_S\le\sum_{t=1}^{T}|e_t-\hat{e}_t|\le TC_S^2. \tag{A.22}$$

So $|S|\le TC_S$. Then

$$\begin{aligned}|\widehat{F}_{T+1}(x)-\widetilde{F}_{T+1}(x)| &\le \frac{1}{T}\sum_{t=1}^{T}|\mathbf{1}\{\hat{e}_t\le x\}-\mathbf{1}\{e_t\le x\}| \\ &\le \frac{1}{T}\left(|S|+\sum_{t\notin S}|\mathbf{1}\{\hat{e}_t\le x\}-\mathbf{1}\{e_t\le x\}|\right) \\ &\overset{(i)}{\le} \frac{1}{T}\left(|S|+\sum_{t\notin S}\mathbf{1}\{|e_t-x|\le C_S\}\right) \\ &\le \frac{1}{T}\left(|S|+\sum_{t=1}^{T}\mathbf{1}\{|e_t-x|\le C_S\}\right) \\ &\le C_S+\mathbb{P}(|e_{T+1}-x|\le C_S)+ \\ &\quad \sup_x\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{|e_t-x|\le C_S\}-\mathbb{P}(|e_{T+1}-x|\le C_S)\right| \\ &= C_S+[F_e(x+C_S)-F_e(x-C_S)]+\sup_x\Big|[\widetilde{F}_{T+1}(x+C_S)-\widetilde{F}_{T+1}(x-C_S)] \\ &\quad -[F_e(x+C_S)-F_e(x-C_S)]\Big| \\ &\overset{(ii)}{\le} (L_{T+1}+1)C_S+2\sup_x|\widetilde{F}_{T+1}(x)-F_e(x)|,\end{aligned}$$

where (i) is because $|\mathbf{1}\{a\le x\}-\mathbf{1}\{b\le x\}|\le\mathbf{1}\{|b-x|\le|a-b|\}$ for $a,b\in\mathbb{R}$ and (ii) is because the Lipschitz continuity of $F_e(x)$. □

Corollary C.7. *If the true covariance matrix Σ is known and we use $\widehat{\Sigma}=\Sigma$. Under Assumption 4.1, 4.3, 4.5 and 4.7, with a high probability $1-\delta$,*

$$\sup_x|\widehat{F}_{T+1}(x)-\widetilde{F}_{T+1}(x)|\le(L_{T+1}+1)\frac{\delta_T}{\sqrt{\lambda}}+2\sup_x|\widetilde{F}_{T+1}(x)-F_e(x)|. \tag{A.23}$$

Theorem C.8. *Under Assumption 4.1, 4.3, 4.5 and 4.7, with a high probability $1-\delta$, for any training size T and $\alpha\in(0,1)$, we have*

$$\begin{aligned}&|\mathbb{P}(Y_{T+1}\in\widehat{C}_{T+1}^{\alpha}\mid X_{T+1}=x_{T+1})-(1-\alpha)| \\ &\le 12\sqrt{\log(16T)/T}+4(L_{T+1}+1)(C_S+\delta_T).\end{aligned} \tag{A.24}$$

Proof. First, we recall the notation here. We define $\hat{\varepsilon}_t=y_t-\hat{f}(x_t)$ as the prediction residual, $\hat{e}_t=\hat{\varepsilon}_t^T\widehat{\Sigma}^{-1}\hat{\varepsilon}_t$ as the

non-conformity score, $e_t = \varepsilon_t^T \Sigma^{-1} \varepsilon_t$ and $\Delta_t = \hat{\varepsilon}_t - \varepsilon_t$. Besides, we define the empirical CDF

$$\begin{aligned} \widehat{F}_{T+1}(x) &= \frac{1}{T}\sum_{t=1}^{T} 1\{\hat{e}_t \le z\} \\ \widetilde{F}_{T+1}(x) &= \frac{1}{T}\sum_{t=1}^{T} 1\{e_t \le z\}. \end{aligned} \tag{A.25}$$

For any $\beta \in [0, \alpha]$, following the idea in Section 4:

$$\begin{aligned} &\left|\mathbb{P}\left(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \middle| X_{T+1} = x_{T+1}\right) - (1-\alpha)\right| \\ =& \left|\mathbb{P}\left(\hat{e}_{T+1} \in [\beta \text{ quantile of } \widehat{F}_{T+1}, 1-\alpha+\beta \text{ quantile of } \widehat{F}_{T+1}] \middle| X_{T+1} = x_{T+1}\right) - (1-\alpha)\right| \\ =& \left|\mathbb{P}\left(\beta \le \widehat{F}_{T+1}\left(\hat{e}_{T+1}\right) \le 1-\alpha+\beta\right) - \mathbb{P}\left(\beta \le F_e\left(e_{T+1}\right) \le 1-\alpha+\beta\right)\right|. \end{aligned} \tag{A.26}$$

The last equality is a result of Lemma C.1. We can further rewrite the equation A.26 as follows:

$$\begin{aligned} &|\mathbb{P}(\beta \le \widehat{F}_{T+1}(\hat{e}_{T+1}) \le 1-\alpha+\beta) - \mathbb{P}(\beta \le F_e(e_{T+1}) \le 1-\alpha+\beta)| \\ &\le \mathbb{E}|\mathbf{1}\{\beta \le \widehat{F}_{T+1}(\hat{e}_{T+1}) \le 1-\alpha+\beta\} - \mathbf{1}\{\beta \le F_e(e_{T+1}) \le 1-\alpha+\beta\}| \\ &\le \mathbb{E}(|\mathbf{1}\{\beta \le \widehat{F}_{T+1}(\hat{e}_{T+1})\} - \mathbf{1}\{\beta \le F_e(e_{T+1})\}| + \\ &\quad |\mathbf{1}\{\widehat{F}_{T+1}(\hat{e}_{T+1}) \le 1-\alpha+\beta\} - \mathbf{1}\{F_e(e_{T+1}) \le 1-\alpha+\beta\}|). \end{aligned} \tag{A.27}$$

The last inequality is because that for any constants a, b and univariates x, y,

$$|\mathbf{1}\{a \le x \le b\} - 1\{a \le y \le b\}| \le |\mathbf{1}\{a \le x\} - \mathbf{1}\{a \le y\}| + |\mathbf{1}\{x \le b\} - \mathbf{1}\{y \le b\}|.$$

Then we have

$$\begin{aligned} \mathbb{E}|\mathbf{1}\{\beta \le \widehat{F}_{T+1}(\hat{e}_{T+1})\} - \mathbf{1}\{\beta \le F_e(e_{T+1})\}| \le \mathbb{P}(|F_e(e_{T+1}) - \beta| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|) \\ \mathbb{E}|\mathbf{1}\{\widehat{F}_{T+1}(\hat{e}_{T+1}) \le 1-\alpha+\beta\} - \mathbf{1}\{F_e(e_{T+1}) \le 1-\alpha+\beta\}| \le \\ \mathbb{P}(|F_e(e_{T+1}) - (1-\alpha+\beta)| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|), \end{aligned}$$

which holds since $|\mathbf{1}\{a \le x\} - \mathbf{1}\{b \le x\}| \le \mathbf{1}\{|b-x| \le |a-b|\}$ for any constant a, b and univariate x and $\mathbb{E}[\mathbf{1}\{A\}] = \mathbb{P}(A)$. Recall in Lemma 4.9, we defined A_T as the event on which

$$\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \Big| A_T \le \sqrt{\frac{\log(16T)}{T}},$$

where $\mathbb{P}(A_T) > 1 - \sqrt{\frac{\log(16T)}{T}}$. Let A_T^C denote the complement of the event A_T. For any $\gamma \in [0, 1]$, we have that

$$\begin{aligned} &\mathbb{P}(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|) \\ &\le \mathbb{P}(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})| \big| A_T) + \mathbb{P}(A_T^C) \\ &\le \mathbb{P}(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})| \big| A_T) + \mathbb{P}(A_T^C) \\ &\le \mathbb{P}(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| + |F_e(\hat{e}_{T+1}) - F_e(e_{T+1})| \big| A_T) + \sqrt{\frac{\log(16T)}{T}}. \end{aligned} \tag{A.28}$$

To bound the conditional probability above, we note that with a high probability $1-\delta$, conditioning on the event A_T,

$$\begin{aligned} &|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| + |F_e(\hat{e}_{T+1}) - F_e(e_{T+1})| \big| A_T \\ &\le \sup_x |\widehat{F}_{T+1}(x) - F_e(x)| \big| A_T + L_{T+1}|\hat{e}_{T+1} - e_{T+1}| \\ &\le \sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \big| A_T + \sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \big| A_T + L_{T+1}|\hat{e}_{T+1} - e_{T+1}| \\ &\le (L_{T+1}+1)C_S + 3\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \big| A_T + L_{T+1}\delta_T \\ &\le 3\sqrt{\frac{\log(16T)}{T}} + (L_{T+1}+1)(C_S + \delta_T). \end{aligned} \tag{A.29}$$

which follows the result from Lemma 4.9 and 4.11.

Therefore, because $F_e(e_{T+1}) \sim \text{Unif}[0,1]$, we have

$$\begin{aligned}&\mathbb{P}\left(|F_e(e_{T+1}) - \gamma| \leq |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| + |F_e(\hat{e}_{T+1}) - F_e(e_{T+1})| \big| A_T\right) \\ \leq & 6\sqrt{\frac{\log(16T)}{T}} + 2(L_{T+1}+1)(C_S + \delta_T).\end{aligned} \tag{A.30}$$

As a result, by letting $\gamma = \beta$ and $1 - \alpha + \beta$, we have

$$\begin{aligned}&|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \mid X_{T+1} = x_{T+1}) - (1-\alpha)| \\ &\leq 12\sqrt{\frac{\log(16T)}{T}} + 4(L_{T+1}+1)(C_S + \delta_T).\end{aligned} \tag{A.31}$$

□

Now we have an asymptotic coverage guarantee for the case where $\{\varepsilon_t\}_{t=1}^T$ is i.i.d., and we can extend the result to the case where $\{\varepsilon_t\}_{t=1}^T$ is stationary and strong mixing.

Assumption C.9. Assume $\{\varepsilon_t\}_{t=1}^{T+1}$ is stationary and strong mixing with the mixing coefficients $\sum_{k>0} \alpha_k < M$. Meanwhile, $F_e(x)$ (the CDF of true non-conformity score) is Lipschitz continuous with constant $L_{T+1} > 0$.

The properties of stationary and strong mixing can be imparted to the sequence $\{e_t\}_{t=1}^T$.

Lemma C.10. *Under Assumption C.9, $\{e_t\}_{t=1}^T$ is stationary and strong mixing with coefficients $\sum_{k>0} \alpha_k(\{e_t\}_{t\geq 1}) < M$, where $\alpha_k(\{e_t\}_{t\geq 1})$ represents the $\alpha-$mixing coefficients of the random sequence $\{e_t\}_{t\geq 1}$.*

Proof. The relationship between e_t and ε_t is that

$$e_t = \varepsilon_t^T \Sigma^{-1} \varepsilon_t. \tag{A.32}$$

Define $f(x) = x^T \Sigma^{-1} x$. Since $f(x)$ is a Borel-measurable function, we have that $f^{-1}(B)$ is also Borel-measurable for any Borel set $B \subset \mathbb{R}$. Thus we have

$$\begin{aligned}\mathbb{P}(e_{n_1} \in B_{n_1}, \cdots, e_{n_k} \in B_{n_k}) &= \mathbb{P}(\varepsilon_{n_1} \in f^{-1}(B_{n_1}), \cdots, \varepsilon_{n_k} \in f^{-1}(B_{n_k})) \\ &= \mathbb{P}(\varepsilon_{n_1+h} \in f^{-1}(B_{n_1}), \cdots, \varepsilon_{n_k+h} \in f^{-1}(B_{n_k})) \\ &= \mathbb{P}(e_{n_1+h} \in B_{n_1}, \cdots, e_{n_k+h} \in B_{n_k}),\end{aligned} \tag{A.33}$$

which shows the stationarity of $\{e_t\}_{t=1}^T$. Besides, the σ-algebra generated by $f(\varepsilon_t)$ is contained in the σ-algebra generated by ε_t; consequently for all $I \subset \mathbb{Z}$ (possibly infinite),

$$\sigma(f(\varepsilon_t), t \in I) \subset \sigma(\varepsilon_t, t \in I). \tag{A.34}$$

Since the definition of the mixing coefficient is the maximum over the sub-sigma algebra generated by the sequence, it follows that for all k,

$$\alpha_k(\{f(\varepsilon_t)\}_{t\geq 1}) \leq \alpha_k. \tag{A.35}$$

□

As a result, we have that $\{e_t\}_{t=1}^T$ is strong mixing with mixing coefficients $\sum_{k>0} \alpha_k(\{e_t\}_{t\geq 1}) < M$.

Lemma C.11. *Under Assumption C.9, with a high probability $1 - (\frac{M(\log T)^2}{2T})^{\frac{1}{3}}$,*

$$\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \leq \frac{(\frac{M}{2})^{1/3}(\log T)^{2/3}}{T^{1/3}}. \tag{A.36}$$

Proof. The proof follows the proof of Corollary 2 in (Xu & Xie, 2023a). Define $v_T(x) := \sqrt{T}(\widetilde{F}_{T+1}(x) - F_e(x))$. Then, Proposition 7.1 in (Rio et al., 2017) shows that

$$\mathbb{E}\left(\sup_x |v_T(x)|^2\right) \leq \left(1 + 4\sum_{k=0}^{T} \alpha_k\right)\left(3 + \frac{\log T}{2\log 2}\right)^2, \tag{A.37}$$

where α_k is the kth mixing coefficient. Since we assumed that the coefficients are summable with $\sum_{k\geq 0}\alpha_k < M$ (for example, $\alpha_k = \mathcal{O}(n^{-s}), s > 1$), Markov's Inequality shows that

$$\begin{aligned}\mathbb{P}(\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \geq s_T) &\leq \frac{\mathbb{E}(\sup_x |v_T(x)|^2/T)}{s_T^2} \\ &\leq \frac{1+4M}{Ts_T^2}\left(3 + \frac{\log T}{2\log 2}\right)^2.\end{aligned} \tag{A.38}$$

Thus, we let

$$s_T := \left(\frac{1+4M}{T}\left(3 + \frac{\log T}{2\log 2}\right)^2\right)^{1/3} \approx \left(\frac{M(\log T)^2}{2T}\right)^{1/3}, \tag{A.39}$$

and see that

$$\begin{aligned}&\mathbb{P}\left(\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \leq \left(\frac{M(\log T)^2}{2T}\right)^{1/3}\right) \\ &\geq 1 - \left(\frac{M(\log T)^2}{2T}\right)^{1/3}.\end{aligned} \tag{A.40}$$

Hence, the event A_T is chosen so that conditioning on A_T,

$$\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \leq \frac{(\frac{M}{2})^{1/3}(\log T)^{2/3}}{T^{1/3}}. \tag{A.41}$$

□

Corollary C.12. *Assume $\{\varepsilon_t\}_{t=1}^T$ is a stationary and strong mixing sequence with mixing coefficient $0 < \sum_{k>0}\alpha_k < M$. Under Assumption 4.3, 4.5 and 4.7, for any training size T and $\alpha \in (0,1)$, we have*

$$\begin{aligned}&|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \mid X_{T+1} = x_{T+1}) - (1-\alpha)| \\ &\leq 12\frac{(\frac{M}{2})^{1/3}(\log T)^{2/3}}{T^{1/3}} + 4(L_{T+1}+1)\left(\frac{\delta_T}{\sqrt{\lambda}} + \delta_T\right).\end{aligned} \tag{A.42}$$

When the true covariance matrix Σ is known, lemma C.5 also holds for the stationary and strong mixing process, and the proof can be directly used. Combining C.5 and C.11 with the same technique in Theorem C.8 yields the bound in Corollary 4.18.

When the true covariance matrix Σ is unknown, we only need to prove a similar result in Lemma C.4. The difference is that we require the covariance estimator to converge to the true covariance matrix at a certain speed. As mentioned in Remark 4.19, there is work presenting covariance estimators with guarantee in the stationary case, like (Chen et al., 2013). As long as we plug in certain estimators, the proof will follow, and the bound will depend on the guarantee of the estimator.