cemmap
centre for microdata methods and practice

Exact and robust conformal inference methods for predictive machine learning with dependent data

Victor Chernozhukov
Kaspar Wüthrich
Yinchu Zhu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP16/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Exact and Robust Conformal Inference Methods for Predictive Machine Learning With Dependent Data

Victor Chernozhukov* Kaspar Wüthrich† Yinchu Zhu‡

February 17, 2018

Abstract

We extend conformal inference to general settings that allow for time series data. Our proposal is developed as a randomization method and accounts for potential serial dependence by including block structures in the permutation scheme. As a result, the proposed method retains the exact, model-free validity when the data are i.i.d. or more generally exchangeable, similar to usual conformal inference methods. When exchangeability fails, as is the case for common time series data, the proposed approach is approximately valid under weak assumptions on the conformity score.

Keywords: Conformal inference, permutation and randomization, dependent data, groups.

1 Introduction

Suppose that we observe a times series $\{Z_t\}_{t=1}^{T_0}$, where each $Z_t = (X_t, Y_t)$ is a random variable in $\mathbb{R}^p \times \mathbb{R}$. Y_t is a response variable and X_t is a p-dimensional vector of features. We want to predict future responses $\{Y_t\}_{t=T_0+1}^{T_0+T_1}$ from future feature values $\{X_t\}_{t=T_0+1}^{T_0+T_1}$. For a pre-specified miscoverage level, we consider the problem of constructing a prediction set for $\{Y_t\}_{t=T_0+1}^{T_0+T_1}$.

The goal and main contribution of this paper is to provide prediction sets, for which performance (coverage accuracy) bounds can be obtained in a wide range of situations, including time series data. While it is possible to design a prediction set for each problem/model, the proposed framework can be used to obtain one unified method with performance guarantees across different settings.

Our method is built on a carefully-designed randomization approach. Under the proposed methodology, we randomize the data based on a certain (algebraic) group of permutations. Note that the standard conformal prediction approach can be viewed as choosing the group to be the set of all permutations. The key idea is to choose a group of permutations that preserve the dependence structure in the data. We do so by randomizing blocks of observations. If exchangeability does not hold, finite-sample performance bounds can still be obtained under weak conditions on the conformity score as long as transformations of the data serve as meaningful approximations for a stationary series.

Our work is closely related to the literature on randomization inference via permutations (Fisher, 1935; Rubin, 1984; Romano, 1990; Lehmann and Romano, 2005) and conformal inference (Vovk et al., 2005, 2009; Lei et al., 2013; Vovk, 2013; Lei and Wasserman, 2014; Burnaev and Vovk, 2014; Balasubramanian et al., 2014; Lei et al., 2015, 2017). These papers typically exploit the i.i.d assumption to obtain the exchangeability condition under all permutations and establish model-free validity of procedures that randomize the data for

*email: vchern@mit.edu
†email: kwuthrich@ucsd.edu
‡email: yzhu6@uoregon.edu

general algorithms. The properties of these methods in the absence of exchangeability are unknown in general. Our work makes contributions in this direction by establishing theoretical guarantees for randomization inference in the non i.i.d. case, covering most common types of time series models. In particular, our results cover strongly mixing processes as a special case, thereby delivering a conformal prediction method to predictive machine learning with dependent data. The very recent work Chernozhukov et al. (2017) explores permutations of residuals obtained from specific regression or factor models in a longitudinal data context, focusing on inference for counterfactual policy evaluation. By contrast, our work deals with randomization of the data and aims to robustify the conformal inference method by extending its validity to settings with dependent data.

The remainder of the paper is as follows. In Section 2, we present the setup, describe a general algorithm for constructing prediction sets, introduce the permutation schemes, and discuss two specific examples. In Section 3, we present the main theoretical properties of the proposed prediction sets. Section 4 concludes. All proofs are collected in the appendix.

2 Conformal Inference for Dependent Data

2.1 Conformal Inference by Permutations

Our approach is based on testing candidate values for $(Y_{T_0+1}, \ldots, Y_{T_0+T_1})$. Prediction sets are then constructed via test inversion. Let $y = (y_{T_0+1}, \ldots, y_{T_0+T_1})$ be a hypothesized value for $(Y_{T_0+1}, \ldots, Y_{T_0+T_1})$. Define the augmented data set $Z_{(y)} = \{Z_t\}_{t=1}^T$, where

$$Z_t = \begin{cases} (Y_t, X_t) & \text{if } 1 \leq t \leq T_0 \\ (y_t, X_t) & \text{if } T_0 + 1 \leq t \leq T_0 + T_1. \end{cases} \tag{1}$$

Similar to the typical conformal inference, we adopt a conformity score measure, also known as nonconformity measure (Vovk et al., 2009), which is a measurable function that maps the (augmented) data $Z_{(y)}$ to a real number. In this paper, $S(Z_{(y)})$ denotes the conformity score and can contain general machine learning algorithms. We shall suppress the subscript (y) and write Z to simplify the notation. Computing S usually involves an estimator, for example, a regression estimator or a joint density. Notice that the estimators embedded in the conformity score S can be either estimated in an online manner or in the typical batch framework in statistics. Concrete examples for S are provided in Section 2.3.

Let $T = T_0 + T_1$. Under our general setup, let $Z = \{Z_t\}_{t=1}^T$ be arbitrary stochastic process indexed by $t \in \{1, ..., T\}$ taking values in a sample space $\mathcal{Z}_T$. A permutation π is bijection from $\{1, ..., T\}$ to itself. Let $Z^\pi = \{Z_{\pi(t)}\}_{t=1}^T$ with $\pi \in \Pi$ be an indexed collection of arbitrary stochastic processes indexed by $t \in \{1, ..., T\}$ taking values in $\mathcal{Z}_T$. We regard these processes as randomized versions of Z. We assume that the index Π includes an identity element $\mathbb{I}$ so that $Z = Z^{\mathbb{I}}$. We denote $n = |\Pi|$ and define the randomization p-value

$$\hat{p} = \hat{p}(y) := \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1}(S(Z^\pi) \geq S(Z)).$$

Given $\alpha \in (0, 1)$, the predictor generates the set of y with corresponding p-values larger than α:

$$\mathcal{C}_{1-\alpha} = \{y : \hat{p}(y) > \alpha\}. \tag{2}$$

We summarize this general method in Algorithm 1.

Note that the p-value can also be stated in terms of order statistics. Let $\{S^{(j)}(Z)\}_{j=1}^n$ denote the nondecreasing rearrangement of $\{S(Z^\pi) : \pi \in \Pi\}$. Call these randomization quantiles. Observe that

$$\mathbf{1}\{\hat{p} \leq \alpha\} = \mathbf{1}\{S(Z) > S^{(k)}(Z)\},$$

where $k = k(\alpha) = n - \lfloor n/\alpha \rfloor = \lceil n(1-\alpha) \rceil$.

Algorithm 1: Generalized Conformal Inference

Input: Data $\{(Y_t, X_t)\}_{t=1}^{T_0}$, $\{X_t\}_{t=T_0+1}^{T_0+T_1}$, miscoverage level $\alpha \in (0,1)$, conformity score $S(\cdot)$, permutation scheme Π

Output: $(1-\alpha)$ confidence set $\mathcal{C}_{1-\alpha}$

for $y \in \mathbb{R}^{T_1}$ **do**

  define $Z_{(y)}$ as in (1)

  compute $\hat{p}(y)$ as in (2)

end

Return the $(1-\alpha)$ confidence set $\mathcal{C}_{1-\alpha} = \{y : \hat{p}(y) > \alpha\}$.

2.2 Designing Permutations Π for Dependent Data

To construct valid prediction sets, we need to take into account the dependence structure in the data. We therefore design Π to have a block structure which preserves the dependence. These blocks are allowed to be overlapping or non-overlapping.

We start with the non-overlapping blocking scheme. Let b be an integer between T_1 and T. We split the data into $K = T/b$ blocks with each block having b consecutive observations. (Here, and henceforth, we assume that the T/b is an integer, for simplicity, as only very minor changes are needed when T/b is not integer-valued.)

We divide the data into K non-overlapping blocks and each block contains b observations. We adopt the convention of labeling the last b observations as the first block. Therefore, the j-th block contains observations for $t \in \{T-jb+1, \ldots, T-(j-1)b\}$. For $1 \le j \le K$, we define the j-th non-overlapping block (NOB) permutation $\pi_{j,\text{NOB}} : \{1, \ldots, T\} \to \{1, \ldots, T\}$ via

$$t \mapsto \pi_{j,\text{NOB}}(t) = \begin{cases} t + (j-1)b & \text{if } 1 \le t \le T-(j-1)b \\ t + (j-1)b - T & \text{if } T-(j-1)b+1 \le j \le T \end{cases} \Bigg| \; t = 1, \ldots, T. \tag{3}$$

The collection of all permutation is given by $\Pi_{\text{NOB}} = \{\pi_{j,\text{NOB}} : 1 \le j \le K\}$. Clearly, Π_{NOB} is a group (in the algebraic sense) and contains the identity map.

We also consider an overlapping blocking scheme. We construct the permutation as a composition of elements in Π_{NOB} and cyclic sliding operation (CSO) permutations. We first define CSO permutations. For $1 \le j \le T$, consider permutations defined by:

$$t \mapsto \pi_{j,\text{CSO}}(t) = \begin{cases} t + (j-1) & \text{if } 1 \le t \le T-(j-1) \\ t + (j-1) - T & \text{if } T-(j-1)+1 \le t \le T \end{cases} \Bigg| \quad t = 1, \ldots, T.$$

The set of cyclic sliding operations is then $\Pi_{\text{CSO}} = \{\pi_{j,\text{CSO}} : 1 \le j \le T\}$. Our overlapping block scheme can be represented by the Minkowski composition of the two groups:

$$\Pi_{\text{OB}} = \Pi_{\text{CSO}} \circ \Pi_{\text{NOB}} = \{\pi_{j_1,\text{CSO}} \circ \pi_{j_2,\text{NOB}} : 1 \le j_1 \le T,\ 1 \le j_2 \le K\}. \tag{4}$$

The set of overlapping block permutations, Π_{OB}, also forms a group and contains the identity map.

2.3 Examples

2.3.1 Penalized Regression

Assume that the data are drawn from the model

$$Y_t = X_t'\beta + \varepsilon_t, \quad 1 \le t \le T, \tag{5}$$

where ε_t is mean-zero stationary stochastic process and $\beta \in \mathbb{R}^p$ is a coefficient vector. We estimate β based on the augmented dataset Z_y using penalized regression

$$\hat{\beta}(Z) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - X_t'\beta\right)^2 + \text{pen}(\beta),$$

where $\text{pen}(\cdot)$ is a penalty function. Popular penalty functions include ℓ_1-norm (LASSO), ℓ_2-norm (ridge regression) and non-convex penalties (e.g., SCAD). We define the fitted residual as

$$\hat{\varepsilon}_t(Z) = Y_t - X_t'\hat{\beta}(Z), \quad t = 1, \ldots, T.$$

Consider the following residual-based conformity score, which operates on the last T_1 elements of the residual vector:

$$S(Z) = \left(\sum_{t=T_0+1}^{T_0+T_1} |\hat{\varepsilon}_t(Z)|^p \right)^{1/p}. \tag{6}$$

If $T_1 = 1$ and $p = 1$, this conformity score corresponds to the absolute value of the last residual, $S(Z) = |\hat{\varepsilon}_T(Z)|$ as in Lei et al. (2017). An natural choice for the block size is $b = T_1$ such that $K = T/T_1$. If $\hat{\beta}(Z)$ is invariant to permutations of the data (which is the case for most regression methods), p-values based non-overlapping and overlapping block permutations can be computed by permuting the fitted residuals.

2.3.2 Autoregressive Models and Neural Networks

Assume that the data are generated by a K-th order linear autoregressive model

$$Y_t = \sum_{k=0}^{K} \rho_k \mathrm{L}^k(Y_t) + \varepsilon_t, \quad t = 1, ..., T, \quad Y_0, \ldots, Y_{-K+1} \text{ given.}$$

where L^k is the lag operator. We use least squares to obtain an estimate of the vector of autoregressive coefficients $\rho = (\rho_0, \ldots, \rho_K)$ based on the augmented data Z. Denote this estimator as $\hat{\rho} = (\hat{\rho}_0, \ldots, \hat{\rho}_K)$ and define the fitted residuals as

$$\hat{\varepsilon}_t(Z) = Y_t - \sum_{k=0}^{K} \hat{\rho}_k \mathrm{L}^k(Y_t), \quad 1 \le t \le T,$$

More generally, we can consider nonlinear autoregressive models

$$Y_t = \rho\left(Y_{t-1}, \ldots, Y_{t-K}\right) + \varepsilon_t, \quad 1 \le t \le T, \quad Y_0, \ldots, Y_{-K+1} \text{ given,}$$

where ρ is a nonlinear function. Such models arise when using neural networks for predictive time series modeling (e.g., Chen and White, 1999; Chen et al., 2001). We allow ρ to be parametric, nonparametric or semi-parametric. Let $\hat{\rho}$ be a suitable estimator for ρ, obtained based on the augmented data Z. Define the fitted residuals as

$$\hat{\varepsilon}_t(Z) = Y_t - \hat{\rho}\left(Y_{t-1}, \ldots, Y_{t-K}\right), \quad 1 \le t \le T.$$

For both the linear and the nonlinear models, we choose a residual-based conformity score

$$S(Z) = \left(\sum_{t=T_0+1}^{T_0+T_1} |\hat{\varepsilon}_t(Z)|^p \right)^{1/p}.$$

A natural choice for the block size is $b = T_1$ such that there are $K = T/T_1$ blocks in total.

3 Theory: General Results on Exact and Approximate Conformal Inference

We now provide theoretical guarantees for the proposed method. When the data are exchangeable, the proposed approach exhibits model-free and exact finite sample validity, similar to the existing conformal inference methods. When the data are serially dependent and exchangeability is violated, our method retains approximate finite sample validity under weak assumptions on the conformity score as long as transformations of the data serve as meaningful approximations for a stationary series.

3.1 Exact Validity

The key insight from the randomization inference literature is to exploit the exchangeability in the data. Since we can cast conformal inference approaches as randomizing in the set Π, we can analyze the proposed generalized conformal inference method (Algorithm 1) by examining the exchangeability and the quantile invariance property (implied by Π being a group).

Theorem 1 (General Exact Validity) *Suppose that $\{Z^\pi\}$ has an exchangeable distribution under permutations $\pi \in \Pi$. Consider any fixed Π such that the randomization α-quantiles are invariant surely, namely*

$$S^{(k(\alpha))}(Z^\pi) = S^{(k(\alpha))}(Z), \text{ for all } \pi \in \Pi.$$

The latter condition holds when Π is a group. Or, more generally, suppose that surely

$$S^{(k(\alpha))}(Z^\pi) \geq S^{(k)}(Z), \text{ for all } \pi \in \Pi. \tag{7}$$

Then

$$P(\hat{p} \leq \alpha) = P(S(Z) > S^{(k)}(Z)) \leq \alpha \quad \text{and} \quad P((Y_{T_0+1}, \ldots, Y_{T_0+T_1}) \in \mathcal{C}_{1-\alpha}) \geq 1 - \alpha.$$

This result follows from standard arguments for randomization inference, see Romano (1990). To the best of our knowledge, this is the weakest condition under which one can obtain model-free validity of conformal inference. A sufficient condition for exchangeability is that the data is i.i.d. (exchangeable) and that Π is a group.

3.2 Approximate Validity

When a meaningful choice of S is available, we can relax the exchangeability condition and expect to achieve certain optimality. Let S_* be an oracle score function, which is typically an unknown population object. For example, S_* can be a transformation of the true population conditional distribution of $(y_{T_0+T_1}, ..., y_{T_0+T_1})$ given $(X_{T_0+T_1}, ..., X_{T_0+T_1})$. In a regression setup, S_* might be measuring the magnitude of the error terms; in the example of Section 2.3.1, S_* would be the analogous of S defined in (6) with true residuals:

$$S_*(Z) = \left(\sum_{t=T_0+1}^{T_0+T_1} |\varepsilon_t(Z)|^p \right)^{1/p}, \tag{8}$$

where $\varepsilon_t(Z) = Y_t - X_t'\beta$. We show that when S consistently approximates the oracle score S_*, the resulting confidence set is valid and approximately equivalent to inference using the oracle score.

For approximate results, assume that the number of randomizations becomes large, $n = |\Pi| \to \infty$ (in examples above, this is caused by $T_0 \to \infty$). Let $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ be sequences of numbers converging to zero, and assume the following conditions.

(E) With probability $1-\gamma_{1n}$: the randomization distribution

$$\tilde{F}(x) := \frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}\{S_*(Z^\pi) < x\},$$

is *approximately ergodic* for $F(x) = P\left(S_*(Z) < x\right)$, namely

$$\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F\left(x\right)\right| \leq \delta_{1n},$$

(A) With probability $1-\gamma_{2n}$, estimation errors are small:

(1) the mean squared error is small, $n^{-1}\sum_{\pi\in\Pi}\left[S(Z^\pi) - S_*(Z^\pi)\right]^2 \leq \delta_{2n}^2$;

(2) the pointwise error at $\pi = \text{Identity}$ is small, $|S(Z) - S_*(Z)| \leq \delta_{2n}$;

(3) The pdf of $S_*(Z)$ is bounded above by a constant D.

Condition (A) states the precise requirement for the quality of approximating the oracle $S_*(Z^\pi)$ by $S(Z^\pi)$. When we view S as an estimator for S_*, we merely require pointwise consistency and consistency in the prediction norm. This condition can be easily verified for many estimation methods under appropriate model assumptions. For example, in sparse high-dimensional linear models, we can invoke well-known results such as Bickel et al. (2009). For linear autoregressive models, sufficient conditions follow from standard results in Hamilton (1994) and Brockwell and Davis (2013). For neural networks, sufficient conditions can be derived from results in Chen and White (1999).

Condition (E) is an ergodicity condition, which states that permuting the oracle conformity scores provides a meaningful approximation to the unconditional distribution of the oracle conformity score. In Section 3.3, we show that Condition (E) holds for strongly mixing time series using the groups of blocking permutations defined in Section 2.2. For regression problems, S_* is typically constructed as a transformation of the regression errors.

The next theorem shows that, under conditions (A) and (E), the proposed generalized conformal inference method is approximately valid.

Theorem 2 (Approximate General Validity of Conformal Inference) *Under the approximate ergodicity condition (E) and the small error condition (A), the approximate conformal p-value is approximately uniformly distributed, that is, it obeys for any* $\alpha \in (0,1)$

$$|P\left(\hat{p} \leq \alpha\right) - \alpha| \leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n}$$

and the conformal confidence set has approximate coverage $1-\alpha$*, namely*

$$|P((Y_{T_0+1},\ldots,Y_{T_0+T_1}) \in \mathcal{C}_{1-\alpha}) - (1-\alpha)| \leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n}.$$

Under further stronger conditions on the estimation quality, the generalized conformal prediction $\mathcal{C}_{1-\alpha}$ achieves an oracle property in volume. Let $\mathcal{C}^*_{1-\alpha} = \{w : 1 - F(S_*(Z)) \geq \alpha\}$ be the oracle prediction set. Let $\mu(\cdot)$ denote the Lebesgue measure. When such oracle prediction set has continuity in the sense that $\mu(\{y : \ |F(S_*(y)) - (1-\alpha)| \leq \varepsilon\}) \to 0$ as $\varepsilon \to 0$, we can show that shrinking errors in approximating $\{S_*(Z^\pi)\}_{\pi\in\Pi}$ by $\{S(Z^\pi)\}_{\pi\in\Pi}$ implies that $\mu(\mathcal{C}_{1-\alpha} \triangle \mathcal{C}^*_{1-\alpha})$ decays to zero, where $\triangle$ denotes the symmetric difference of two sets. Such results have been established by Lei et al. (2013) among others for specific models under i.i.d. data.

3.3 Approximate Ergodicity for Strongly Mixing Time Series with Blocking Permutations Π

In the blocking schemes discussed in Section 2.2, we can view $\{S_*(Z^\pi)\}_{\pi\in\Pi}$ as a time series $\{u_t\}$. Recall $S_*(Z)$ defined in (8) for the penalized regression example in Section 2.3.1. By non-overlapping block permutations defined in (3) with $b = T_1$, we can see that $\{S_*(Z^\pi)\}_{\pi\in\Pi_{\text{NOB}}}$ can be rearranged as $\{u_t\}_{t=1}^K$, where

$$u_t = \left(\sum_{s=T_0+(1-t)b+1}^{T_0+(2-t)b} |\varepsilon_s|^p \right)^{1/p}$$

and ε_s is the true regression residuals in (5).

The situation with overlapping permutations is more complicated. Let $b = T_1$. We observe that for $1 \le k \le K$, $\pi_{k,\text{NOB}} = \pi_{b(k-1)+1,\text{CSO}}$; for $1 \le j_1, j_2 \le T$, $\pi_{j_1,\text{CSO}} \circ \pi_{j_2,\text{CSO}} = \pi_{j_1+j_2-1-T\mathbf{1}\{j_1+j_2>T+1\},\text{CSO}}$. Therefore, for a fixed $1 \le j \le T$, we define the integer $q = \lfloor (T+1-j)/b \rfloor$ and rearrange $\{S_*(Z^{\pi_{j,\text{CSO}}\circ\pi_{k,\text{NOB}}})\}_{k=1}^K$ as follows:

$$\left\{ \left(\sum_{s=T_0+2-j-b(k-1)}^{T+1-j-b(k-1)} |\varepsilon_s|^p \right)^{1/p} \right\}_{k=1}^{q}, \left\{ \left(\sum_{s=T+T_0+2-j-b(k-1)}^{2T+1-j-b(k-1)} |\varepsilon_s|^p \right)^{1/p} \right\}_{k=q+2}^{K} \tag{9}$$

and

$$\left(\sum_{1\le s\le T+2-j-bq \text{ or } T+T_0+2-j-bq\le s\le T} |\varepsilon_s|^p \right)^{1/p}.$$

Therefore, for any fixed $1 \le j \le T$, we can rearrange $\{S_*(Z^{\pi_{j,\text{CSO}}\circ\pi_{k,\text{NOB}}})\}_{k=1}^K$ to be two segments of stationary process in (9) and one extra term.

With this setup in mind, the following result gives a mild sufficient condition for the ergodicity condition (E).

Lemma 1 (Mixing implies Approximate Ergodicity) *We consider both overlapping blocks and non-overlapping blocks.*

1. *Let $\Pi = \Pi_{NOB}$ the set of non-overlapping blocks defined in (3). Suppose that there exists $\{u_t\}_{t=1}^K$ a rearrangement of $\{S_*(Z^\pi)\}_{\pi\in\Pi}$ such that $\{u_t\}_{t=1}^K$ is stationary and strong mixing with $\sum_{k=1}^\infty \alpha_{mixing}(k) \le M$ for a constant M. Then there exists a constant $M' > 0$ depending only on M such that*

$$P\left(\sup_{x\in\mathbb{R}} \left| \tilde{F}(x) - F(x) \right| \le \delta_{1n} \right) \ge 1 - \gamma_n,$$

   where $\gamma_n = M'(\log K)^2/(K\delta_{1n})$.

2. *Let $\Pi = \Pi_{OB}$ the set of non-overlapping blocks defined in (4). Suppose that for each $\pi \in \Pi_{CSO}$, there exist a further permutation $\{u_t^\pi\}_{t=1}^K = \{S_*(Z^{\pi\circ\tilde{\pi}})\}_{\tilde{\pi}\in\Pi_{NOB}}$ such that $\{u_t^\pi\}_{t=1}^{K_\pi}$ and $\{u_t^\pi\}_{t=K_\pi+2}^K$ are stationary and strong mixing with $\sum_{k=1}^\infty \alpha_{mixing}(k) \le M$ for a constant M that does not depend on π. Then there exists a constant $M' > 0$ depending only on M such that*

$$P\left(\sup_{x\in\mathbb{R}} \left| \tilde{F}(x) - F(x) \right| \le \delta_{1n} \right) \ge 1 - \gamma_n,$$

   where $\gamma_n = M'(\log K)/(\sqrt{K}\delta_{1n})$.

Strong mixing is a mild condition on dependence and is satisfied by many stochastic processes. For example, it is well known that any stationary Markov chains that are Harris recurrent and aperiodic are strong mixing. Many common serially dependent processes such as ARMA with i.i.d. innovations can also be shown to be strong mixing.

4 Conclusion

This paper extends the applicability of conformal inference to general settings that allow for time series data. Our results are developed within the general framework of randomization inference. Our method is based on a carefully-designed randomization approach based on groups of permutations, which exhibit a block structure to account for the potential serial dependence in the data. When the data are i.i.d. or more generally exchangeable, our method exhibits exact, model-free validity. When the exchangeability condition does not hold, finite-sample performance bounds can still be obtained under weak conditions on the conformity score as long as transformations of the data serve as meaningful approximations for a stationary series.

A Proof of Theorem 1

The proof essentially follows by standard arguments, see, e.g. Romano (1990). We have by (7)

$$\sum_{\pi\in\Pi} \mathbf{1}(S(Z^\pi) > S^{(k)}(Z^\pi)) \leq \sum_{\pi\in\Pi} \mathbf{1}(S(Z^\pi) > S^{(k)}(Z)) \leq \alpha n.$$

Since $\mathbf{1}(S(Z) > S^{(k)}(Z))$ is equal in law to $\mathbf{1}(S(Z^\pi) > S^{(k)}(Z^\pi))$ for any $\pi \in \Pi$ by the exchangeability hypothesis, we have that

$$\alpha \geq E \sum_{\pi\in\Pi} \mathbf{1}(S(Z^\pi) > S^{(k)}(Z^\pi))/n = E\mathbf{1}(S(Z) > S^{(k)}(Z)) = E\mathbf{1}(\hat{p} \leq \alpha).$$

B Proof of Theorem 2

Since the second claim (bounds on the coverage probability) is implied by the first claim, it suffices to show the first claim. Define

$$\hat{F}(x) = \frac{1}{n} \sum_{\pi\in\Pi} \mathbf{1}\{S(Z^\pi) < x\}.$$

The rest of the proof proceeds in two steps. We first bound $\hat{F}(x) - F(x)$ and then derive the desired result.

Step 1: We bound the difference between the p-value and the oracle p-value, $\hat{F}(S(Z)) - F(S_*(Z))$.

Let $\mathcal{M}$ be the event that the conditions (A) and (E) hold. By assumption,

$$P(\mathcal{M}) \geq 1 - \gamma_{1n} - \gamma_{2n}. \tag{10}$$

Notice that on the event $\mathcal{M}$,

$$\begin{aligned}
\left|\hat{F}(S(Z)) - F(S_*(Z))\right| &\leq \left|\hat{F}(S(Z)) - F(S(Z))\right| + |F(S(Z)) - F(S_*(Z))| \\
&\overset{\text{(i)}}{\leq} \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - F(x)\right| + D\,|S(Z) - S_*(Z)| \\
&\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| + D\,|S(Z) - S_*(Z)|
\end{aligned}$$

$$\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\left|S(Z) - S_*(Z)\right|$$
$$\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\delta_{2n}, \tag{11}$$

where (i) holds by the fact that the bounded pdf of $S_*(Z)$ implies Lipschitz property for F.

Let $A = \left\{\pi \in \Pi : \ |S(Z^\pi) - S_*(Z^\pi)| \geq \sqrt{\delta_{2n}}\right\}$. Observe that on the event $\mathcal{M}$, by Chebyshev inequality

$$|A|\delta_{2n} \leq \sum_{\pi\in\Pi} \left(S(Z^\pi) - S_*(Z^\pi)\right)^2 \leq n\delta_{2n}^2$$

and thus $|A|/n \leq \delta_{2n}$. Also observe that on the event $\mathcal{M}$, for any $x \in \mathbb{R}$,

$$\begin{aligned}
&\left|\hat{F}(x) - \tilde{F}(x)\right| \\
&\leq \frac{1}{n}\sum_{\pi\in A} \left|\mathbf{1}\left\{S(Z^\pi) < x\right\} - \mathbf{1}\left\{S_*(Z^\pi) < x\right\}\right| + \frac{1}{n}\sum_{\pi\in(\Pi\setminus A)} \left|\mathbf{1}\left\{S(Z^\pi) < x\right\} - \mathbf{1}\left\{S_*(Z^\pi) < x\right\}\right| \\
&\overset{\text{(i)}}{\leq} 2\frac{|A|}{n} + \frac{1}{n}\sum_{\pi\in(\Pi\setminus A)} \mathbf{1}\left\{|S_*(Z^\pi) - x| \leq \sqrt{\delta_{2n}}\right\} \leq 2\frac{|A|}{n} + \frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}\left\{|S_*(Z^\pi) - x| \leq \sqrt{\delta_{2n}}\right\} \\
&\leq 2\frac{|A|}{n} + P\left(|S_*(Z) - x| \leq \sqrt{\delta_{2n}}\right) + \sup_{z\in\mathbb{R}} \left|\frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}\left\{|S_*(Z^\pi) - z| \leq \sqrt{\delta_{2n}}\right\} - P\left(|S_*(Z) - z| \leq \sqrt{\delta_{2n}}\right)\right| \\
&= 2\frac{|A|}{n} + P\left(|S_*(Z) - x| \leq \sqrt{\delta_{2n}}\right) \\
&\quad + \sup_{x\in\mathbb{R}} \left|\left[\tilde{F}\left(z + \sqrt{\delta_{2n}}\right) - \tilde{F}\left(z - \sqrt{\delta_{2n}}\right)\right] - \left[F\left(z + \sqrt{\delta_{2n}}\right) - F\left(z - \sqrt{\delta_{2n}}\right)\right]\right| \\
&\leq 2\frac{|A|}{n} + P\left(|S_*(Z) - x| \leq \sqrt{\delta_{2n}}\right) + 2\sup_{z\in\mathbb{R}} \left|\tilde{F}(z) - F(z)\right| \\
&\overset{\text{(ii)}}{\leq} 2\frac{|A|}{n} + 2D\sqrt{\delta_{2n}} + 2\delta_{1n} \overset{\text{(iii)}}{\leq} 2\delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}},
\end{aligned} \tag{12}$$

where (i) follows by the boundedness of indicator functions and the elementary inequality of $|\mathbf{1}\{S(Z^\pi) < x\} - \mathbf{1}\{S_*(Z^\pi) < x\}| \leq \mathbf{1}\{|S_*(Z^\pi) - x| \leq |S(Z^\pi) - S_*(Z^\pi)|\}$, (ii) follows by the bounded pdf of $S_*(Z)$ and (iii) follows by $|A|/n \leq \delta_{2n}$. Since the above display holds for each $x \in \mathbb{R}$, it follows that on the event $\mathcal{M}$,

$$\sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| \leq 2\delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}}. \tag{13}$$

We combine (11) and (13) and obtain that on the event $\mathcal{M}$,

$$\left|\hat{F}(S(Z)) - F(S_*(Z))\right| \leq 3\delta_{1n} + 2\delta_{2n} + D(\delta_{2n} + 2\sqrt{\delta_{2n}}). \tag{14}$$

Step 2: Here we derive the desired result. Notice that

$$\begin{aligned}
&\left|P\left(1 - \hat{F}(S(Z)) \leq \alpha\right) - \alpha\right| \\
&= \left|E\left(\mathbf{1}\left\{1 - \hat{F}(S(Z)) \leq \alpha\right\} - \mathbf{1}\left\{1 - F(S_*(Z)) \leq \alpha\right\}\right)\right| \\
&\leq E\left|\mathbf{1}\left\{1 - \hat{F}(S(Z)) \leq \alpha\right\} - \mathbf{1}\left\{1 - F(S_*(Z)) \leq \alpha\right\}\right| \\
&\overset{\text{(i)}}{\leq} P\left(|F(S_*(Z)) - 1 + \alpha| \leq \left|\hat{F}(S(Z)) - F(S_*(Z))\right|\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq P\left(|F(S_*(Z)) - 1 + \alpha| \leq \left|\hat{F}(S(Z)) - F(S_*(Z))\right| \text{ and } \mathcal{M}\right) + P(\mathcal{M}^c) \\
&\overset{\text{(ii)}}{\leq} P\left(|F(S_*(Z)) - 1 + \alpha| \leq 3\delta_{1n} + 2\delta_{2n} + D(\delta_{2n} + 2\sqrt{\delta_{2n}})\right) + P\left(\mathcal{M}^c\right) \\
&\overset{\text{(iii)}}{\leq} 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n},
\end{aligned}
$$

where (i) follows by the elementary inequality $|\mathbf{1}\{1 - \hat{F}(S(Z)) \leq \alpha\} - \mathbf{1}\{1 - F(S_*(Z)) \leq \alpha\}| \leq \mathbf{1}\{|F(S_*(Z)) - 1 + \alpha| \leq |\hat{F}(S(Z)) - F(S_*(Z))|\}$, (ii) follows by (14), (iii) follows by the fact that $F(S_*(Z))$ has the uniform distribution on $(0, 1)$ and hence has pdf equal to 1, and by (10). The proof is complete.

C Proof of Lemma 1

Proof of the first claim. By assumption,

$$
\tilde{F}(x) - F(x) = \frac{1}{K}\sum_{t=1}^{K}\left(\mathbf{1}\{u_t < x\} - F(x)\right).
$$

Applying Proposition 7.1 of Rio (2017), we have that

$$
E\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right|^2\right) \leq \frac{1 + 4M}{K}\left(3 + \frac{\log K}{2\log 2}\right)^2.
$$

Therefore, the first result follows by Markov's inequality

$$
\begin{aligned}
P\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right| > \delta_{1n}\right) &\leq \frac{E\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right|^2\right)}{\delta_{1n}^2} \\
&\leq \frac{1 + 4M}{K\delta_{1n}^2}\left(3 + \frac{\log K}{2\log 2}\right)^2.
\end{aligned}
$$

Proof of the second claim. For any $\pi \in \Pi_{\text{CSO}}$, define

$$
G_\pi(x) = \frac{1}{K}\sum_{\tilde{\pi}\in\Pi_{\text{NOB}}}\left(\mathbf{1}\{S_*(Z^{\pi\circ\tilde{\pi}}) < x\} - F(x)\right).
$$

Notice that

$$
\tilde{F}(x) - F(x) = \frac{1}{T}\sum_{\pi\in\Pi_{\text{CSO}}} G_\pi(x).
$$

It follows that

$$
E\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right| \leq \frac{1}{T}\sum_{\pi\in\Pi_{\text{CSO}}} E\sup_{x\in\mathbb{R}}|G_\pi(x)|. \tag{15}
$$

We now bound $E\sup_{x\in\mathbb{R}}|G_\pi(x)|$. For a fixed $\pi \in \Pi_{\text{CSO}}$, we have

$$
G_\pi(x) = \frac{1}{K}\sum_{t=1}^{K}\left(\mathbf{1}\{u_t^\pi < x\} - F(x)\right).
$$

We can further decompose

$$G_\pi(x) = \frac{1}{K}\left[K_\pi G_\pi^{(1)}(x) + (K - K_\pi - 1)G_\pi^{(2)}(x) + \left(\mathbf{1}\{u^\pi_{K_\pi+1} < x\} - F(x)\right)\right],$$

where

$$G_\pi^{(1)}(x) = K_\pi^{-1}\sum_{t=1}^{K_\pi}\left(\mathbf{1}\{u_t^\pi < x\} - F(x)\right),$$

$$G_\pi^{(2)}(x) = (K - K_\pi - 1)^{-1}\sum_{t=K_\pi+2}^{K}\left(\mathbf{1}\{u_t^\pi < x\} - F(x)\right).$$

By the same argument as in part 1, we can show that

$$E\left(\sup_{x\in\mathbb{R}}\left|G_\pi^{(1)}(x)\right|^2\right) \le \frac{1+4M}{K_\pi}\left(3 + \frac{\log K_\pi}{2\log 2}\right)^2,$$

$$E\left(\sup_{x\in\mathbb{R}}\left|G_\pi^{(2)}(x)\right|^2\right) \le \frac{1+4M}{K-K_\pi-1}\left(3 + \frac{\log(K-K_\pi-1)}{2\log 2}\right)^2.$$

Let $z \mapsto f(z)$ be defined by

$$f(z) = \frac{1+4M}{z}\left(3 + \frac{\log(z)}{2\log 2}\right)^2.$$

It is not difficult to verify that $d^2 f(z)/dz^2 < 0$ for $z \ge 1$. Therefore, $f(z)$ is concave on $[1, K-1]$. Therefore,

$$f(K_\pi) + f(K - K_\pi - 1) \le 2f((K-1)/2) = \frac{4+16M}{K-1}\left(3 + \frac{\log((K-1)/2)}{2\log 2}\right)^2.$$

It follows that

$$E\left(\sup_{x\in\mathbb{R}}\left|G_\pi^{(1)}(x)\right|^2\right) + E\left(\sup_{x\in\mathbb{R}}\left|G_\pi^{(2)}(x)\right|^2\right) \le \frac{4+16M}{K-1}\left(3 + \frac{\log((K-1)/2)}{2\log 2}\right)^2.$$

Therefore,

$$\begin{aligned}
E\sup_{x\in\mathbb{R}}|G_\pi(x)|^2 &= E\sup_{x\in\mathbb{R}}\left|\frac{1}{K}\left[K_\pi G_\pi^{(1)}(x) + (K-K_\pi-1)G_\pi^{(2)}(x) + \left(\mathbf{1}\{u^\pi_{K_\pi+1} < x\} - F(x)\right)\right]\right|^2 \\
&\le E\left(\frac{1}{K}\left[K_\pi \sup_{x\in\mathbb{R}}\left|G_\pi^{(1)}(x)\right| + (K-K_\pi-1)\sup_{x\in\mathbb{R}}\left|G_\pi^{(2)}(x)\right| + 2\right]\right)^2 \\
&\overset{\text{(i)}}{\le} 4\frac{K_\pi^2}{K^2}E\sup_{x\in\mathbb{R}}\left|G_\pi^{(1)}(x)\right|^2 + 4\frac{(K-K_\pi-1)^2}{K^2}E\sup_{x\in\mathbb{R}}\left|G_\pi^{(2)}(x)\right|^2 + \frac{8}{K^2} \\
&\le 4\left(E\sup_{x\in\mathbb{R}}\left|G_\pi^{(1)}(x)\right|^2 + E\sup_{x\in\mathbb{R}}\left|G_\pi^{(2)}(x)\right|^2\right) + \frac{8}{K^2} \\
&\le \frac{16+64M}{K-1}\left(3 + \frac{\log((K-1)/2)}{2\log 2}\right)^2 + \frac{8}{K^2},
\end{aligned}$$

where (i) follows by the elementary inequality $(a+b+c)^2 \le 4a^2 + 4b^2 + 2c^2$.

Notice that the above bound does not depend on π. In light of (15), it follows that

$$E\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right| \le \sqrt{\frac{16+64M}{K-1}\left(3 + \frac{\log((K-1)/2)}{2\log 2}\right)^2 + \frac{8}{K^2}}.$$

The second claim of the lemma follows by Markov's inequality.

References

Vineeth N. Balasubramanian, , Shen-Shyang Ho, , and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning*. Morgan Kaufmann, Boston, 2014.

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

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 605–622, 2014.

Xiaohong Chen and H. White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, Mar 1999. ISSN 0018-9448. doi: 10.1109/18.749011.

Xiaohong Chen, Jeffrey Racine, and Norman R Swanson. Semiparametric arx neural-network models with an application to forecasting inflation. *IEEE Transactions on neural networks*, 12(4):674–683, 2001.

V. Chernozhukov, K. Wüthrich, and Y. Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. arXiv:1712.09089, 2017.

R.A. Fisher. *The Design of Experiments*. Oliver & Boyd, 1935.

James D. Hamilton. *Time series: theory and methods*. Springer Science & Business Media, 1994.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Allesandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, (just-accepted), 2017.

Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*. Springer, 2017.

Joseph P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990. ISSN 01621459.

Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. 12 (4):1151–1172, 12 1984. doi: 10.1214/aos/1176346785.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2):349–376, Sep 2013. ISSN 1573-0565. doi: 10.1007/s10994-013-5355-6.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590, 2009.