Contents

A Theoretical proof

A.1 Proof of Theorem 1

Theorem 1 (Validity)**.** *Under the assumptions stated above, the proposed method satisfies the following conditional coverage guarantee:*

$$|\mathbb{P}(Y_{T+1} \in \mathcal{C}_{T+1}(\alpha)|X_{T+1} = x) - (1-\alpha)| \leq (4L + 2L\sqrt{\omega} + 2)\sqrt{\omega} + 6\sqrt{\frac{\log(16n)}{n}} + \frac{\log(16n)}{n},$$

where

$$\omega = \nu_n^2 r + 2r\nu_n\sqrt{(\kappa_1 + \sqrt{\kappa_2})I} + o(g(n))(\kappa_1 + \sqrt{\kappa_2})I.$$

For easy notation, denote $\hat{A} = A_n, \Delta_t = \hat{\epsilon}_t - \epsilon_t$ and sometimes we drop subscript t.

Lemma 1. *For any test conformity score $\hat{s}_t = \hat{\epsilon_t}^T \hat{A} \hat{\epsilon_t}$ and the true conformity score $s_t = \epsilon_t^T A \epsilon_t$, with probability at least $1 - \delta$,*

$$\sum_{t=T-n+1}^{T} |\hat{s}_t - s_t| \leq \omega n, \tag{1}$$

where

$$\omega = \nu_n^2 r + 2r\nu_n\sqrt{(\kappa_1 + \sqrt{\kappa_2})I} + o(g(n))(\kappa_1 + \sqrt{\kappa_2})I.$$

Proof. We have:

$$\begin{aligned}
&|\hat{s}_t - s_t| = |\epsilon_t^\top A \epsilon_t - \hat{\epsilon_t}^\top \hat{A} \hat{\epsilon_t}| \leq |\epsilon_t^\top A \epsilon_t - \epsilon_t^\top \hat{A} \epsilon_t| + |\epsilon_t^\top \hat{A} \epsilon_t - \hat{\epsilon_t}^\top \hat{A} \hat{\epsilon_t}| \\
\leq &|\Delta^\top \hat{A} \Delta| + 2|\Delta^\top \hat{A} \epsilon_t| + |\epsilon_t^\top (A - \hat{A}) \epsilon_t| \\
\leq &\|\hat{A}\|\|\Delta\|^2 + 2\|\hat{A}\|\|\Delta\|\|\epsilon\| + \|\epsilon\|^2 \|A - \hat{A}\| \qquad \text{(i)} \\
\leq &r\|\Delta\|^2 + 2r\|\Delta\|\|\epsilon\| + o(g(n))\|\epsilon\|^2. \qquad (2)
\end{aligned}$$

The inequality (i) exists because of the Cauchy-schwartz inequality and Assumption 2. Hence, by Assumption 1, we have

$$\begin{aligned}
&\sum_{t=T-n+1}^{T} |\hat{s}_t - s_t| \leq rn\nu_n^2 + 2r \sum_{t=T-n+1}^{T} \|\Delta_t\|\|\epsilon_t\| + o(g(n)) \sum_{t=T-n+1}^{T} \|\epsilon_t\|^2 \\
\leq &rn\nu_n^2 + 2r\sqrt{(\sum_{t=T-n+1}^{T} \|\Delta_t\|^2)(\sum_{t=T-n+1}^{T} \|\epsilon_t\|^2)} + o(g(n)) \sum_{t=T-n+1}^{T} \|\epsilon_t\|^2 \\
\leq &rn\nu_n^2 + 2r\sqrt{n\nu_n^2(\sum_{t=T-n+1}^{T} \|\epsilon_t\|^2)} + o(g(n)) \sum_{t=T-n+1}^{T} \|\epsilon_t\|^2. \qquad (3)
\end{aligned}$$

From Assumption 3, we have

$$\mathbb{E}\left[\frac{1}{n} \sum_{t=T-n+1}^{T} \|\epsilon_t\|^2\right] = \frac{1}{n} \sum_{t=T-n+1}^{T} \mathbb{E}[\|\epsilon_t\|^2] \leq \kappa_1 I. \tag{4}$$

Using Chebyshev's inequality, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{t=T-n+1}^{T} \|\epsilon_t\|^2 - \mathbb{E}[\|\epsilon_t\|^2] \geq \sqrt{\frac{\mathrm{Var}[\|\epsilon_t\|^2]}{n\delta}}\right) \leq \frac{\mathrm{Var}[\|\epsilon_t\|^2]}{n \cdot \frac{\mathrm{Var}[\|\epsilon_t\|^2]}{n\delta}} = \delta, \tag{5}$$

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

$$\begin{aligned}\frac{1}{n}\sum_{t=T-n+1}^{T}\|\epsilon_t\|^2 &\leq \mathbb{E}[\|\epsilon_t\|^2]+\sqrt{\frac{\mathrm{Var}[\|\epsilon_t\|^2]}{n\delta}}\\ &\leq \kappa_1 I+\sqrt{\frac{\kappa_2 I}{n\delta}}\\ &\leq (\kappa_1+\sqrt{\frac{\kappa_2}{n\delta I}})I \leq (\kappa_1+\sqrt{\kappa_2})I. \end{aligned} \tag{6}$$

The last inequality is because we can set δ such that $\delta n I < 1$. Plug into eq. (3), we have with probability higher than $1-\delta$, we obtain eq. (1) and the lemma follows.

□

Denote the empirical CDF: $\widehat{F}_{n+1}(x)=\frac{1}{n}\sum_{i=T-n+1}^{T}1_{\hat{s}_i\leq x}$, $\widetilde{F}_{n+1}(x)=\frac{1}{n}\sum_{i=T-n+1}^{T}1_{s_i\leq x}$ and true CDF of score function $F_s(x)=P(s\leq x)$.

Lemma 2. *Under Assumption 3, for any n, there exists an event A_n which occurs with probability at least $1-\sqrt{\frac{\log(16n)}{n}}$, such that, conditioning on A_n,*

$$\sup_x\left|\widetilde{F}_{n+1}(x)-F(x)\right|\leq\sqrt{\frac{\log(16n)}{n}}.$$

Proof. The proof follows Lemma 1 in [16] that utilizes Dvoretzky-Kiefer-Wolfowitz inequality in [7]. □

Lemma 3. *Under Assumption 1, Assumption 3,with high probability,*

$$\sup_x\left|\widehat{F}_{n+1}(x)-\widetilde{F}_{n+1}(x)\right|\leq(2L+1)\sqrt{\omega}+2\sup_x\left|\widetilde{F}_{n+1}(x)-F_e(x)\right|.$$

Proof. The proof is similar to Lemma $B.6$ in [15], and is written here for completeness.

Using Lemma 1 we have that with probability $1-\delta$,

$$\sum_{t=T-n+1}^{T}|s_t-\hat{s}_t|\leq n\omega. \tag{7}$$

Let $S=\{t:|s_t-\hat{s}_t|\geq\sqrt{\omega}\}$. Then,

$$|S|\sqrt{\omega}\leq\sum_{t=T-n+1}^{T}|s_t-\hat{s}_t|\leq n\omega. \tag{8}$$

So $|S| \leq n\sqrt{\omega}$. Then,

$$
\begin{aligned}
|\widehat{F}_{n+1}(x) - \widetilde{F}_{n+1}(x)| &\leq \frac{1}{n} \sum_{t=T-n+1}^{T} |1\{\hat{s}_t \leq x\} - 1\{s_t \leq x\}| \\
&\leq \frac{1}{n}|S| + \sum_{t \notin S} |1\{\hat{s}_t \leq x\} - 1\{s_t \leq x\}| \\
&\leq \frac{1}{n}|S| + \frac{1}{n} \sum_{t=T-n+1}^{T} 1\{|s_t - x| \leq \sqrt{\omega}\} \qquad \text{(i)} \\
&\leq \sqrt{\omega} + P(|s_{T+1} - x| \leq \sqrt{\omega}) \\
&\quad + \sup_x \left| \frac{1}{n} \sum_{t=T-n+1}^{T} 1\{|s_t - x| \leq \sqrt{\omega}\} - P(|s_{T+1} - x| \leq \sqrt{\omega}) \right| \\
&= \sqrt{\omega} + [F_s(x + \sqrt{\omega}) - F_s(x - \sqrt{\omega})] \\
&\quad + \sup_x \left[\widetilde{F}_{n+1}(x + \sqrt{\omega}) - F_{n+1}(x - \sqrt{\omega}) \, - \big(F_s(x + \sqrt{\omega}) - F_s(x - \sqrt{\omega})\big) \right] \\
&\qquad\qquad (ii) \\
&\leq (2L+1)\sqrt{\omega} + 2 \sup_x \left| \widetilde{F}_{n+1}(x) - F_s(x) \right|, \qquad (9)
\end{aligned}
$$

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

Proof of Theorem 1

Proof. Look at the conditional coverage of Y_{T+1} given X_{T+1}:

$$
\begin{aligned}
&\left| \mathbb{P}\left(Y_{T+1} \in \mathcal{C}^{\alpha}_{T+1} \mid X_{T+1} = x_{T+1}\right) - (1-\alpha) \right| \qquad (10) \\
=&\left| \mathbb{P}\left(\widehat{s}_{T+1} \leq 1-\alpha \text{ quantile of } \widehat{F}_{n+1} \mid X_{T+1} = x\right) - (1-\alpha) \right| \\
=&\left| \mathbb{P}\left(\widehat{F}_{n+1}(\widehat{s}_{T+1}) \leq 1-\alpha\right) - \mathbb{P}\left(F_s(s_{T+1}) \leq 1-\alpha\right) \right| \\
=&\left| \mathbb{E}[1\{\widehat{F}_{n+1}(\widehat{s}_{T+1}) \leq 1-\alpha\} - 1\{F_s(s_{T+1}) \leq 1-\alpha\}] \right| \\
\leq&\mathbb{P}\left(|F_s(s_{T+1}) - (1-\alpha)| \leq |\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(s_{T+1})| \right). \qquad (11)
\end{aligned}
$$

Based on Lemma 2, we can define the event $A_n, \mathbb{P}(A_n) \geq 1 - \frac{\log(16n)}{n}$, conditional on A_n, we have:

$$
\sup_x \left| \widetilde{F}_{n+1}(x) - F_s(x) \right| \leq \sqrt{\frac{\log(16n)}{n}}, \qquad (12)
$$

Hence, we can write eq. (11) as

$$
\begin{aligned}
&\mathbb{P}(|F_s(s_{T+1}) - (1-\alpha)| \leq |\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(s_{T+1})|) \\
\leq&\mathbb{P}(|F_s(s_{T+1}) - (1-\alpha)| \leq |\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(s_{T+1})||A_n) + \mathbb{P}(A_n^c) \\
\leq&\mathbb{P}(|F_s(s_{T+1}) - (1-\alpha)| \leq |\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(\widehat{s}_{T+1})| + |F_s(\widehat{s}_{T+1}) - F_s(s_{T+1})||A_n) + \frac{\log(16n)}{n}. \qquad (13)
\end{aligned}
$$

Conditional on A_n:

$$
\begin{aligned}
&|\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(\widehat{s}_{T+1})| + |F_s(\widehat{s}_{T+1}) - F_s(s_{T+1})| \\
\leq& \sup_x |\widehat{F}_{n+1}(x) - F_s(x)| + L|\widehat{s}_{T+1} - s_{T+1}| \\
\leq&(2L+1)\sqrt{\omega} + 3\sqrt{\frac{\log(16n)}{n}} + L\omega. \qquad (14)
\end{aligned}
$$

The last equation exists because of Lemma 3 1 and eq. (12).

Note that $F_s(s_{T+1}) \sim Unif(0,1)$, we have

$$\begin{aligned}&\mathbb{P}(|F_s(s_{T+1}) - (1-\alpha)| \leq |\widehat{F}_{n+1}(\widehat{s}_{T+1}) - F_s(\widehat{s}_{T+1})| + |F_s(\widehat{s}_{T+1}) - F_s(s_{T+1})||A_n)\\ \leq&(4L+2)\sqrt{\omega} + 6\sqrt{\frac{\log(16n)}{n}} + 2L\omega. \end{aligned} \tag{15}$$

Plug into eq. (11), we have

$$\begin{aligned}&\left|\mathbb{P}\left(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \mid X_{T+1} = x_{T+1}\right) - (1-\alpha)\right|\\ \leq&(4L+2)\sqrt{\omega} + 6\sqrt{\frac{\log(16n)}{n}} + 2L\omega + \frac{\log(16n)}{n}. \end{aligned} \tag{16}$$

□

A.2 Proof of Theorem 2

Theorem 2 (Efficiency)**.** *There exists $0.1 < \alpha_0 < 1$, such that when $\alpha < \alpha_0$, the optimal solution to the minimization problem is given by:*

$$A_* := \arg\min_{A \succ 0} V(A, Q_{1-\alpha}(\epsilon^\top A \epsilon)), \tag{17}$$

where $\epsilon \sim \mathcal{N}(0, A_*^{-1})$.

To prove the theorem, we firstly introduce second-order stochastic dominance to compare the magnitude of the $1-\alpha$ quantile under different A.

Definition 1 (Second–Order Stochastic Dominance)**.** *Let ρ and ν be probability measures on $\mathbb{R}$ such that $\mathbb{E}_{X\sim\rho}[\,|X|\,] < \infty \quad$ and $\quad \mathbb{E}_{Y\sim\nu}[\,|Y|\,] < \infty$. The following statements are equivalent; whenever any of them hold we say that ρ* second-order stochastically dominates *ν and write $\rho \succeq_{ssd} \nu$.*

- ***Utility criterion.*** *For every non-decreasing and concave function $u : \mathbb{R} \to \mathbb{R}$,*
$$\mathbb{E}_{X\sim\rho}\big[u(X)\big] \;\geq\; \mathbb{E}_{Y\sim\nu}\big[u(Y)\big]. \tag{18}$$
- ***Integrated distribution-function criterion.*** *Denoting by F_ρ and F_ν the cumulative distribution functions of ρ and ν, respectively,*
$$\int_{-\infty}^{t} F_\rho(x)\,dx \;\leq\; \int_{-\infty}^{t} F_\nu(x)\,dx, \qquad \forall\, t \in \mathbb{R}. \tag{19}$$
- ***Coupling criterion.*** *There exist random variables $X \sim \rho$ and $Y \sim \nu$ such that*
$$Y = X - \delta + \varepsilon, \qquad \delta \geq 0, \quad \mathbb{E}\big[\varepsilon \mid X - \delta\big] = 0. \tag{20}$$

The equivalence of the definitions is proved in [8, 11]. Second-order stochastic dominance between X and Y implies the the quantile functions of X and Y exhibit certain tail behavior, which would be further utilized to compare the $1-\alpha$ quantile for a given small α.

Lemma 4 (Tail-behavior of quantile function)**.** *Let X and Y be integrable random variables with cumulative distribution functions F_X, F_Y and continuous quantile functions $Q_X(\alpha) = \inf\{t \in R, F_X(t) >= \alpha\}, Q_Y(\alpha) = \inf\{t \in R, F_X(t) >= \alpha\}$. If $X \succeq_{ssd} Y$, then there exists $\alpha^* \in [0,1]$such that*

$$Q_X(\alpha) \leq Q_Y(\alpha), \;\; \textit{for } \alpha^* < \alpha < 1.$$

Proof. For integrable X and Y, the distribution function criterion for second stochastic dominance (19) is equivalent to the quantile–integral condition[13].

$$\Phi(\alpha) := \int_0^\alpha \big(Q_X(u) - Q_Y(u)\big)\,du \;\geq\; 0, \quad \forall \alpha \in [0,1], \qquad \Phi(0) = \Phi(1) = 0. \tag{21}$$

Set $\Delta(\alpha) := Q_X(\alpha) - Q_Y(\alpha)$. Now we prove by contradiction. Assume that there exists a b^* such that $\int_{b^*}^1 \Delta > 0$. Then $\Phi(b^*) = \Phi(1) - \int_{b^*}^1 \Delta < 0$, contradicting (21). By the continuity of the quantile function, there exists an α^* such that $\Delta(\alpha) \leq 0$ for $\alpha > \alpha^*$. □

Lemma 5. *Let* $Z_1, Z_2, Z_3, \ldots, Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ *and*

$$S_{\boldsymbol{\lambda}} \;=\; \sum_{i=1}^{d} \lambda_i \, Z_i^2, \qquad \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_d) \in [0,1]^d, \quad \sum_{i=1}^{d} \lambda_i = 1.$$

Write $\boldsymbol{\lambda}^\star = (\frac{1}{d}, \frac{1}{d}, \cdots, \frac{1}{d})$. *Then*

$$S_{\boldsymbol{\lambda}} \;\preceq_{ssd}\; S_{\boldsymbol{\lambda}^\star} \qquad \textit{for every } \boldsymbol{\lambda}.$$

Proof. Consider the special case in the coupling criterion (20) when $\delta = 0$, which is also known as Mean-preserving spread (MPS). Let $T = \sum_{i=1}^{d} Z_i^2 \sim \chi_d^2$. Conditioned on $T = t$, the vector $(Z_1^2, Z_2^2, \cdots, Z_d^2)$ is exchangeable, i.e. we have $\mathbb{E}[Z_1^2 \mid T] = \mathbb{E}[Z_2^2 \mid T] = \cdots = \mathbb{E}[Z_d^2 \mid T]$. Rewrite $S_{\boldsymbol{\lambda}} = S_{\boldsymbol{\lambda}^*} + \varepsilon$, where $\varepsilon = \sum_{i=1}^{d} (\lambda_i - \frac{1}{d}) Z_i^2$. Then

$$\mathbb{E}(\varepsilon | S_{\boldsymbol{\lambda}^*}) = \sum_{i=1}^{d} \mathbb{E}\left[(\lambda_i - \frac{1}{d}) Z_i^2 | T \right] = \sum_{i=1}^{d} (\lambda_i - \frac{1}{d}) \mathbb{E}[Z_i^2 | T] = 0.$$

By the coupling criterion (20), $S_{\boldsymbol{\lambda}} \;\preceq_{ssd}\; S_{\boldsymbol{\lambda}^*}$. In other words, the uniform weight vector minimises risk for every risk-averse preference, completing the proof. □

Proof of Theorem 2

Proof. By the definition of ellipsoid volume, we have

$$V(A, Q_{1-\alpha}(\epsilon^\top A \epsilon)) = Constant \times Q_{1-\alpha}(\epsilon^\top A \epsilon)^{I/2} [\det(A)]^{-1/2} \tag{22}$$

Since the optimization problem is invariant to the rescaling of the scalar (that is, A and any positive scalar multiple cA, where $c > 0$, produce the same mathematical solution), additional constraints, such as the bounding of the matrix norm $A \leq 1$, can be imposed without loss of generality. Note that $\epsilon \sim N(0, A_*^{-1})$ and $A \succ 0$ by Cholesky decomposition, we can write $A_*^{-1} = LL^\top$ where L is a lower triangular matrix. Define the matrix $B = L^\top A L$, we can rewrite the eq. (22) as

$$\min_{B \succ 0} Q_{1-\alpha}^{I/2}(x^\top B x) \det(L) [\det(B)]^{-1/2}, \tag{23}$$

where $x \sim N(0, \mathrm{Id})$and $\det(L)$ is a constant independent of B.

To further solve the optimization problem, we look at the eigenvalue of B, suppose $B = O\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_I) O^\top$ and O is an orthogonal matrix. Since the optimization value is invariant to different scaling of B, we are imposing additional constrains on λ_i.

$$\min_{\lambda_i > 0, \sum_{i=1}^{I} \lambda_i = 1} Q_{1-\alpha}^{I} \left(\sum_{i=1}^{I} \lambda_i x_i^2 \right) \prod_{i=1}^{I} \lambda_i^{-\frac{1}{2}}, \tag{24}$$

and $\{x_i\}_{1 \leq i \leq I}$ are i.i.d. random variables $x_i \sim N(0,1)$.

Now we would like to prove that the above optimization problem is solved when $\lambda_1 = \lambda_2 \cdots = \lambda_d = \frac{1}{I}$. Note that by Cauchy-Schwartz Inequality $\prod_{i=1}^{I} \lambda_i^{-\frac{1}{2}}$ is minimized when $\lambda_1 = \lambda_2 \cdots = \lambda_d = \frac{1}{I}$. By Lemma 4 5, there exists an α_0 s.t. when $\alpha < \alpha_0$, $Q_{1-\alpha}(\sum_{i=1}^{I} \lambda_i x_i^2)$ is minimized when $\lambda_1 = \lambda_2 \cdots = \lambda_d = \frac{1}{I}$. By comparing the quantile function of most majorized(risky) combinations of $\boldsymbol{\lambda} = (1, 0, \cdots, 0)$ with the quantile function of the least risky combinations $\boldsymbol{\lambda} = (\frac{1}{I}, \frac{1}{I}, \cdots, \frac{1}{I})$, emperically we find that $\alpha_0 > 0.1$[9].

□

A.3 Tail-up model

Lemma 6. *The spatial covariance* Σ *of any node pair* (u, v) *is:*

$$\Sigma(u,v) = \int_{\wedge u \cap \wedge v} m(r-u) m(r-v) \frac{w(r)}{\sqrt{w(u) w(v)}} dr. \tag{25}$$

Proof. Note that

$$\Sigma(u,v) = \text{Cov}(\int_{\wedge u} m(s-u)\sqrt{\frac{w(s)}{w(u)}}dB(s), \int_{\wedge v} m(r-v)\sqrt{\frac{w(r)}{w(v)}}dB(r))$$

Due to the independence of increments for Brownian motion, only when $r = s \in \wedge u \cap \wedge v$, the covariance is non-zero. Note that for Brownian motion, we have $\text{Cov}(dB(r), dB(s)) = 1_{r=s}drds$. Hence Lemma 6 follows. □

Lemma 7. *If we set the moving function* $m(r-u) = \beta \exp\Big(-\frac{d(r,u)}{\phi}\Big)$, *with parameters* $\beta > 0$ *(a scale factor) and* $\phi > 0$ *(a range or decay parameter), then the covariance matrix between two locations* u, v *can be expressed as:*

$$\Sigma(u,v) = \sigma^2 \sqrt{\frac{w(u)}{w(v)}} \exp\big(-d(u,v)/\phi\big)\, \mathbf{1}_{\{u \to v\}}. \tag{26}$$

Proof. By Lemma 6 and substitute $m(r-u) = \beta\, e^{-d(r,u)/\phi}$ and $m(r-v) = \beta\, e^{-d(r,v)/\phi}$, we have

$$\Sigma(u,v) \;=\; \int_{\wedge u \cap \wedge v} \beta^2 \;\exp\Big(-\tfrac{d(r,u)}{\phi}\Big)\; \exp\Big(-\tfrac{d(r,v)}{\phi}\Big)\; \frac{w(r)}{\sqrt{w(u)\,w(v)}}\; dr.$$

The set $\wedge u \cap \wedge v$ are the segments of networks that flow into *both* u and v. Consider the following cases:

- *If* u *and* v *are not flow-connected*, then $\wedge u \cap \wedge v = \emptyset$ and hence $\Sigma(u,v) = 0$.
- *If* u *and* v *are flow-connected*, without loss of generality assume v is downstream of u. Then $\wedge u \cap \wedge v = \wedge u$, and for each r in $\wedge u$, $d(r,v) = d(r,u) + d(u,v)$. Hence we have

$$\exp\Big(-\frac{d(r,u)+d(r,v)}{\phi}\Big) \;=\; \exp\Big(-\frac{d(u,v)}{\phi}\Big)\, \exp\Big(-\frac{2\,d(r,u)}{\phi}\Big).$$

  leads to a remaining integral over $r \in \wedge u$. We can write

$$\Sigma(u,v) \;=\; \beta^2 \sqrt{\frac{w(u)}{w(v)}}\, \exp\Big(-\frac{d(u,v)}{\phi}\Big) \int_{\wedge u} \exp\Big(-\tfrac{2d(r,u)}{\phi}\Big)\; \frac{w(r)}{w(u)}\; dr,$$

Note that $\int_{\wedge u} \exp\Big(-\frac{2d(r,u)}{\phi}\Big)\; \frac{w(r)}{w(u)}\; dr = \Sigma(u,u)$ is a constant, since the additivity constrain on $w(u)$ assures the constant variance of site u. Thus, the tail-up exponential model yields a covariance of the form

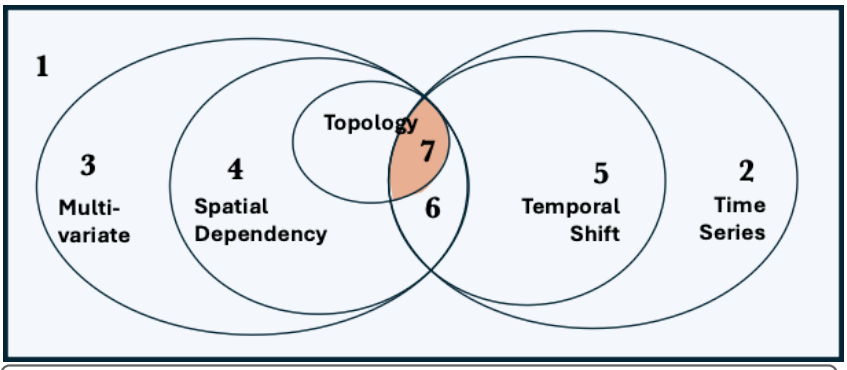
$$\Sigma(u,v) \;=\; \begin{cases} \Big(\text{constant factors}\Big)\; \exp\Big(-\frac{d(u,v)}{\phi}\Big), & \text{if } u \text{ and } v \text{ are flow-connected,} \\ 0, & \text{otherwise.} \end{cases}$$

□

B Taxonomy for Related Works

Figure 1 provides an overview of the conformal prediction (CP) literature, complementing the discussion in the related work section. The Venn diagram categorizes existing CP methods based on the type of data they are designed for—time series, multivariate data, or both—and the nature of the assumptions they make, particularly regarding exchangeability, temporal stationarity, and spatial independence.

Traditional CP methods, such as split conformal prediction for i.i.d. regression[14], assume full exchangeability, and thus cannot be directly applied to time series or spatially structured data without modification. In the time series domain, recent works such as [1, 18, 5, 17] relax the



1. [14] 3. [3] 4. [10]
5. [5, 18, 16, 1, 17]
6. [15] 7. Our method, `STACI`

Figure 1: Taxonomy of works in conformal prediction. Among studies that account for both spatial dependency and temporal shift—without assuming spatial and temporal exchangeability—our work is the first to incorporate topology information.

exchangeability assumption via online calibration, sliding window methods, scorecasters or dynamic programming. These methods handle distribution shift over time but typically operate in univariate or low-dimensional settings. In contrast, CP methods for multivariate or high-dimensional data [10, 3] often focus on constructing prediction sets that exploit geometry or sparsity, but generally assume no temporal structure.

Our proposed method, *STACI*, is positioned at the intersection of these axes in the Venn diagram. It is designed for high-dimensional time-indexed multivariate data arising from spatio-temporal stream networks. Our method explicitly accounts for both spatial dependencies and temporal shifts, leveraging the underlying topological structure of the network to enhance predictive performance.

C Adaptive Uncertainty Set Construction

We construct a spatio-temporally adaptive prediction set for a new observed history X_{T+1} using our proposed nonconformity score as follows:

$$\mathcal{C}(X_{T+1}; \alpha) = \{y : s(X_{T+1}, y) \leq \hat{Q}_{1-\alpha}\},$$

where $\hat{Q}_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the empirical cumulative distribution function of $\{s(X_t, Y_t), t \in \mathcal{D}_{\text{cal}}\}$. The complete `STACI` algorithm is outlined in Algorithm 1.

To account for potential temporal distribution shifts in the predictive error, we adopt the Adaptive Conformal Inference (ACI) framework proposed in [5]. This approach dynamically updates the confidence level α_t over time, ensuring that the prediction set remains responsive to evolving data distributions.

Specifically, we iteratively update α_t, and reconstruct the prediction set $\mathcal{C}(X_t, \alpha_t)$ accordingly. At the initial test time $T+1$, the confidence level is set as $\alpha_{T+1} = \alpha$. For subsequent time steps $t > T+1$, we update α_t with a step size $\gamma \geq 0$ as follows:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathbb{1}\{Y_t \notin \mathcal{C}(X_t; \alpha_t)\}), \quad \forall t \geq T+1. \tag{27}$$

The rationale behind ACI is that if the prediction set fails to cover the true value at time t, the effective error level is reduced, leading to a wider prediction interval at time $t+1$, thereby increasing the likelihood of coverage. A larger step size γ makes the method more responsive to observed distribution shifts but also introduces greater fluctuations in α_t. When $\gamma = 0$, the method reduces to standard (non-adaptive) conformal prediction.

We present the analysis of the average coverage guarantee of `STACI` without any assumption about ϵ_t. The proof follows from Proposition 4.1 in [5].

Algorithm 1 STACI

Input: Data $\mathcal{D}$; Network topology $\mathcal{G}$; Model $f(\cdot)$; Hyper-parameters λ; Confidence level α.
Output: Prediction set $\mathcal{C}(X_{T+1}; \alpha)$.

// Training
$\hat{f} \leftarrow$ Fit f using $\mathcal{D} \setminus \mathcal{D}_{\text{cal}}$;
// Calibration
$\mathcal{E} \leftarrow \{\hat{\epsilon}_t = Y_t - \hat{f}(X_t) - \frac{\sum_{t \in \mathcal{D}_{\text{cal}}} (Y_t - \hat{f}(X_t))}{|\mathcal{D}_{\text{cal}}|}\}_{t \in \mathcal{D}_{\text{cal}}}$;
$\hat{\Sigma}_n \leftarrow \sum_{t \in \mathcal{D}_{\text{cal}}} \hat{\epsilon}_t \hat{\epsilon}_t^\top / (n-1)$ given $\mathcal{E}$;
$\hat{\Sigma}_{\mathcal{G}} \leftarrow$ Compute (26) for $(\ell_i, \ell_{i'}), \forall i, i' \in \mathcal{I}$ given $\mathcal{G}$;
$A \leftarrow \lambda \hat{\Sigma}_{\mathcal{G}}^{-1} + (1-\lambda)\hat{\Sigma}_n^{-1}$;
$\mathcal{S} \leftarrow \{\hat{\epsilon}_t^\top A \hat{\epsilon}_t\}_{t \in \mathcal{D}_{\text{cal}}}$ given $\mathcal{E}$;
$\hat{Q}_{1-\alpha} \leftarrow$ Compute $\frac{\lceil (1-\alpha)(n+1) \rceil}{n}$-th quantile given $\mathcal{S}$;
// Testing
$\mathcal{C}(X_{T+1}; \alpha) \leftarrow \{y : s(X_{T+1}, y) \leq \hat{Q}_{1-\alpha}\}$;

Proposition 1. *Consider n' test data points as the n' realizations of (X_{T+1}, Y_{T+1}), denoted by $\mathcal{D}_{test}$. We have the asymptotic coverage guarantee:*

$$\lim_{n' \to \infty} \sum_{t \in \mathcal{D}_{test}} \mathbb{1}\{Y_t \notin \mathcal{C}(X_t; \alpha_t)\} / n' = \alpha.$$

While Proposition 1 provides a weaker coverage guarantee compared to Theorem 1, it offers broader applicability, remaining valid even in adversarial online settings. Empirical results suggest that when the error process exhibits minimal distribution shift and the assumptions of Theorems 2 and 1 are only slightly violated, STACI maintains the predefined coverage level ($\gamma = 0.01$) while achieving efficient prediction sets. However, when $\gamma > 0$, Proposition 1 does not ensure a finite-sample coverage gap. Understanding this limitation and developing methods to control the finite-sample coverage gap presents an interesting direction for future research. While Proposition 1 provides a weaker coverage guarantee compared to Theorem 1, it offers broader applicability and maintains validity even under adversarial online setting. Our empirical observations indicate that when the error process exhibits minimal distribution shift and the assumptions of Theorems 2 and 1 are only slightly violated, STACI achieves efficient prediction sets while maintaining the predefined coverage level with $\gamma = 0.01$. When $\gamma > 0$, Proposition 1 does not guarantee a finite coverage gap. Investigating this limitation further presents a promising direction for future research.

D Additional Experiment Details and Results

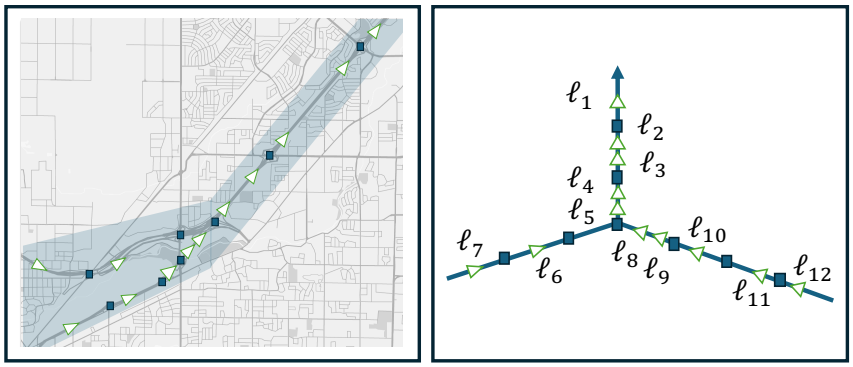
D.1 Computational Resource

Experiments were conducted on a single NVIDIA GeForce RTX 4080 Super GPU, an AMD Ryzen 9 7950X 16-Core Processor CPU, 64GB Memory and 2TB SSD.

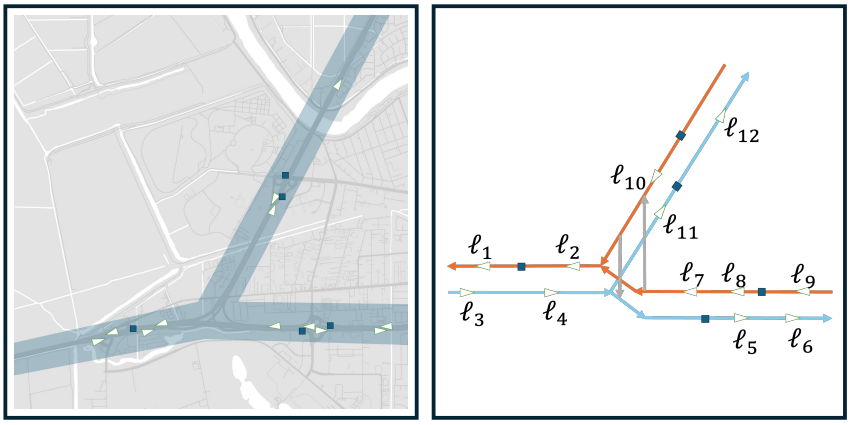
D.2 Datasets and Codes

The PeMS03 dataset used in this paper is collectd by California Transportation Agencies (CalTrans) Performance Measurement System (PeMS). It contains three months of statistics on traffic flow every 5 minutes ranging from Sept. 1st 2018 to Nov. 30th 2018, including 358 sensors. The datasets are from `https://github.com/guoshnBJTU/ASTGNN/tree/main` without available license.

The first backbone ST-Graph model is AGCRN [2], which we adapted the official implementation from `https://github.com/LeiBAI/AGCRN` under MIT license. ASTGCN [6] was implemented with PyTorch Geometric Temporal package [12] from `https://github.com/benedekrozemberczki/pytorch_geometric_temporal` under MIT license. STGODE [4] was adapted from the official implementation `https://github.com/square-coder/STGODE` under Apache-2.0 license.



(a) PEMS-G1



(b) PEMS-G2

Figure 2: Real-world road network structures and their abstraction for PEMS-G1 and PEMS-G2. In each sub-figure, the left map displays the road network, where freeways are bold gray lines in blue shade, and ramps off the freeway are represented by blue squares. Based on these ramps and road junctions, the network is divided into different segments. Traffic flow monitoring sensors from ℓ_1 to ℓ_{12} are placed exclusively on those northbound freeways, marked with green transparent triangles. The right map provides an abstract representation of the road network and sensor locations, using the same symbols for consistency.

D.3 Simulation Data Generation Details

Segment r_1 and r_2 starts with (0, 1) and (0.5, 0.8), respectively, and both end with (0.3, 0.5). The next segment r_3 also starts with (0.3, 0.5), and end with (0.2, 0.1). Segment r_4 start from (0.6, 0.6), and ends at the same location as r_3. Starting from this location, r_5 ends at (0.4, 0). The weights for segment $1-5$ are set as $0.35, 0.5, 0.85, 0.15$ and 1, respectively. Each segment has two observation locations – one at the start point, another at the middle point.

To approximate the integral, each segment is uniformly divided into 300 smaller sub-intervals. For segments without parent nodes (r_1, r_2 and r_4 in our example), the source nodes are treated as infinitely distant. In implementation, the source node of each segment is extended 10 times in the same direction to simulate infinity.

D.4 Real-world Data Details

As shown in Figure 2, we construct two subgraphs from PeMS03 dataset. Note that road network structure are required in stream networks. Among all PeMS datasets, PeMS03 is the only one that contains a mapping to real sensor identification number in PeMS. Therefore, other datasets are not available for our setting.

D.5 Running Time

On a single Nvidia Geforce 4080S, the training time of AGCRN with PEMS-G1 is $1419s$. In table 1, we show the computation time of our and baseline CP methods on AGCRN model and PEMS-G1 dataset. Our method `STACI` has comparable running time with other methods.

Table 1: Computation Time (seconds) for Different CP Methods with Different Calibration Set Size n

Calibration Set Size n	Sphere	Sphere-ACI ($\gamma = 0.01$)	Square	*MultiDimSPCI*	**STACI** ($\gamma = 0$)	**STACI** ($\gamma = 0.01$)
100	142	144	24	24	25	25
200	152	153	24	24	26	25
300	135	129	24	24	25	26
400	132	135	25	24	24	25
500	135	136	23	24	23	24

D.6 Method Details

We provide pseudo-codes for our proposed method in 1. This applies all experiments in this paper, excluding offline setting detailed in Section D.8.

D.7 Hyperparameter Study

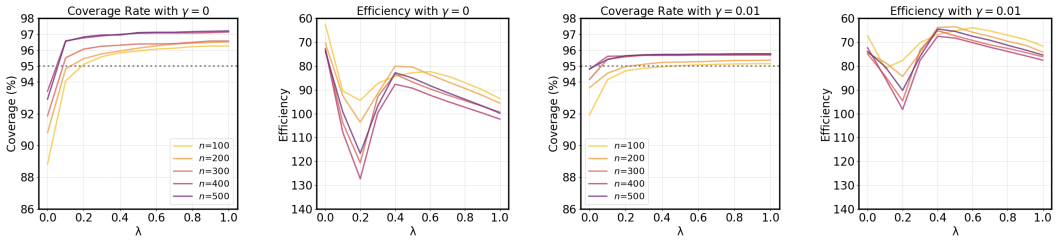
The studies for hyperparameter are presented in Figure 3.

D.8 Additional Ablation Study: Offline Experiment

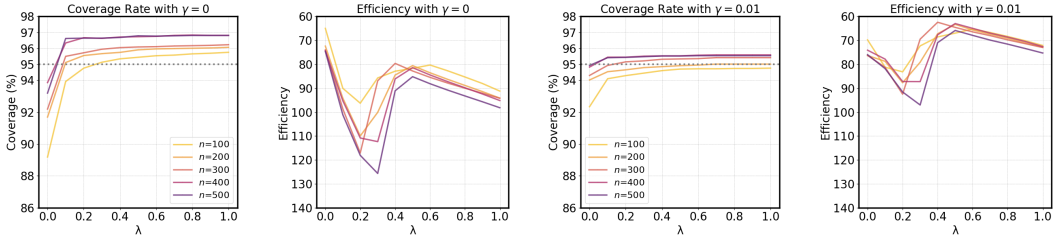
In both synthetic and real-world data, MultiDimSPCI achieve the closest to our proposed method, `STACI`, in efficiency. Therefore, we focus our comparison on four specific variants: vanilla MultiDimSPCI($\gamma = 0$), MultiDimSPCI($\gamma = 0.01$), `STACI`($\gamma = 0$), and `STACI`($\gamma = 0.01$). In the offline setting, `STACI` does not update the covariance matrix estimation. To ensure a fair comparison, we similarly fix the covariance matrix for MultiDimSPCI methods at the beginning of the test phase.

The results are illustrated in Figure 4. As seen in the left figure, fixing the covariance matrix significantly improves the coverage rates of all methods, bringing them close to the desired 95% level. However, despite having the same γ, `STACI` consistently outperforms MultiDimSPCI in efficiency. Notably, when ACI is not applied ($\gamma = 0$), both methods tend to be overly conservative, resulting in coverage rates well above the desired 95%. Therefore, since `STACI` ($\gamma = 0$) achieves a higher coverage rate, MultiDimSPCI ($\gamma = 0.01$) and `STACI`($\gamma = 0$) exhibit similar efficiency.

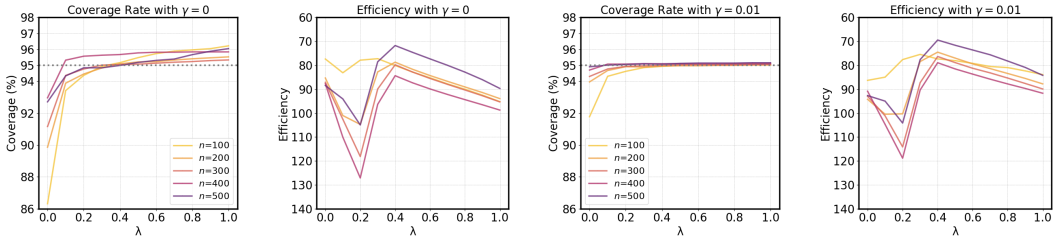
In conclusion, regardless of whether the covariance matrix is fixed or not, `STACI` consistently surpasses MultiDimSPCI in both coverage and efficiency. Furthermore, to achieve an exact coverage rate, incorporating ACI ($\gamma = 0.01$) is recommended.



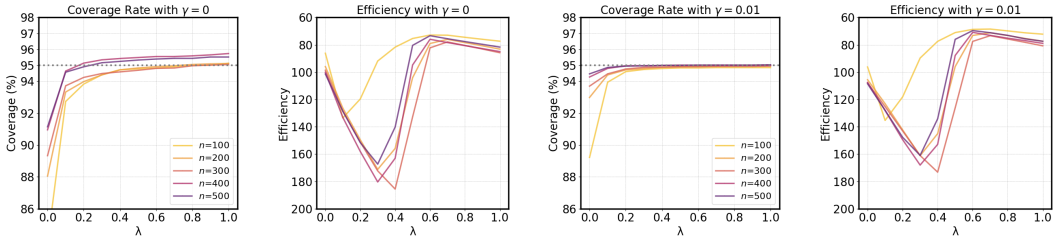
(1a)-(1d): AGCRN on PeMS-G1



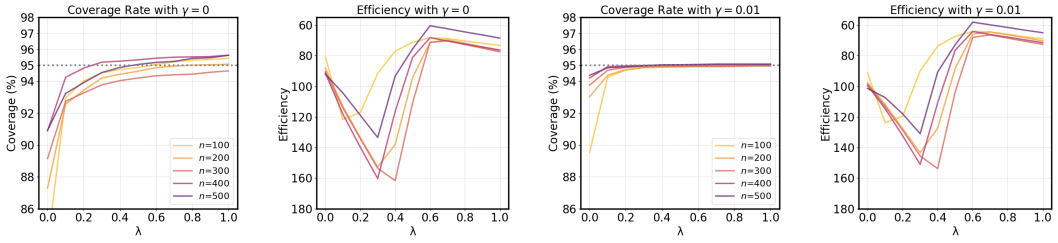
(2a)-(2d): ASTGCN on PeMS-G1



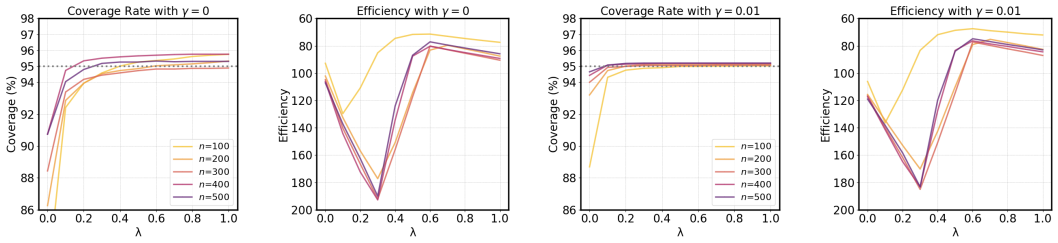
(3a)-(3d): STGODE on PeMS-G1



(4a)-(4d): AGCRN on PeMS-G2



(5a)-(5d): ASTGCN on PeMS-G2



(6a)-(6d): STGODE on PeMS-G2

Figure 3: Comparison of Coverage and Efficiency for all PeMS experiments with different backbone GNN models and datasets. Each figure show results for different belief weight λ and calibration set size n, with adaptive step size $\gamma = 0$ (a&b) and 0.01 (c&d). Pre-determined coverage threshold of 95% is shown by horizontal gray dotted lines.

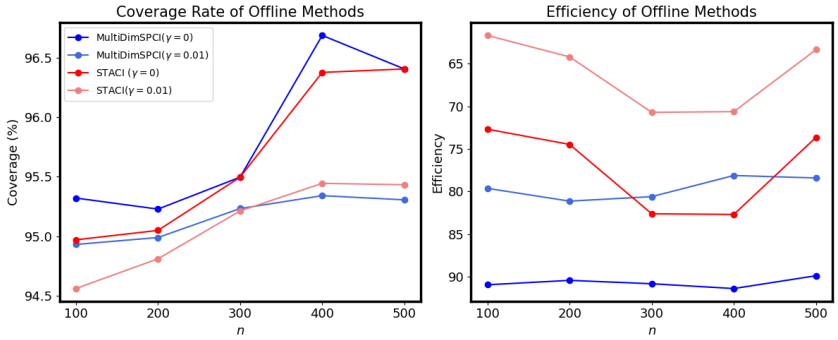


Figure 4: Coverage and efficiency of different methods with different calibration set size n. MultiDimSPCI methods are in blue, and our methods are in red. Methods with $\gamma = 0$ are in darker colors; while those with adaptive coverage, $\gamma = 0.01$, are shown in shallow colors.

References

[1] Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

[2] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

[3] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Distribution-free prediction bands for multivariate functional time series: an application to the italian gas market. *arXiv preprint arXiv:2107.00527*, 2021.

[4] Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 364–373, 2021.

[5] Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[6] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929, Jul. 2019.

[7] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

[8] Michael Landsberger and Isaac Meilijson. Mean-preserving portfolio dominance. *The Review of Economic Studies*, 60(2):479–485, 1993.

[9] Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications. 1979.

[10] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

[11] Michael Rothschild and Joseph E Stiglitz. Increasing risk: I. a definition. In *Uncertainty in economics*, pages 99–121. Elsevier, 1978.

[12] Benedek Rozemberczki, Paul Scherer, Yixuan He, George Panagopoulos, Alexander Riedel, Maria Astefanoaei, Oliver Kiss, Ferenc Beres, Guzman Lopez, Nicolas Collignon, and Rik Sarkar. PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, page 4564–4573, 2021.

[13] Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007.

[14] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[15] Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Forty-first International Conference on Machine Learning*, 2024.

[16] Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11575–11587, 2023.

[17] Zitong Yang, Emmanuel Candès, and Lihua Lei. Bellman conformal inference: Calibrating prediction intervals for time series. *arXiv preprint arXiv:2402.05203*, 2024.

[18] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.