

Conformal Prediction via Transported Beta Laws

Predição conforme, leis beta de estatística de ordem e transporte ótimo

Thiago R. Ramos Helton Graziadei Luben M. C. Cabezas

Universidade Federal de São Carlos

Roteiro

- 1 Predição conforme e a garantia marginal
- 2 Exemplo prático: regressão conforme
- 3 Marginal vs. condicional na calibração
- 4 A lei beta e a má calibração
- 5 Controlar o condicional dá a cobertura
- 6 O objeto de estudo: desvio da beta
- 7 Transporte ótimo
- 8 Dualidade e o bound de cobertura
- 9 Aplicações

1. O que é predição conforme (split)

Cenário. Treina-se uma regra de predição em uma amostra; uma amostra de **calibração** separada $S_1, \dots, S_n$ escolhe o limiar (escores: maior = pior).

Limiar conforme no nível $1 - \alpha \in (0, 1)$:

$$\hat{q}_{n,1-\alpha} = S_{(k_{1-\alpha})}, \quad k_{1-\alpha} = \lceil (n+1)(1-\alpha) \rceil,$$

onde $S_{(1)} \leq \dots \leq S_{(n)}$ são as estatísticas de ordem.

Garantia marginal (Teorema 1)

Sob trocabilidade e sem empates,

$$1 - \alpha \leq \mathbb{P}(S_{n+1} \leq \hat{q}_{n,1-\alpha}) < (1 - \alpha) + \frac{1}{n+1}$$

Crucial: essa probabilidade é **marginal** — ela faz média sobre o escore de teste e sobre a amostra de calibração aleatória.

Exemplo prático: regressão conforme heterocedástica

Dados. $y = f(x) + \sigma(x)\varepsilon$, com ruído $\sigma(x) = 0.15 + 0.45x$ **crescente em x** (heterocedástico). A amostra é dividida em três partes disjuntas:

Treino

ajusta a média $\hat{\mu}(x)$ e o modelo de dispersão $\hat{\sigma}(x)$ (floresta sobre $|y - \hat{\mu}(x)|$).

Calibração

escores $S_i \rightarrow$ corte $\hat{q} = S_{(k)}$, $k = \lceil (n+1)(1-\alpha) \rceil$.

Teste

conjunto $\{y : S(x, y) \leq \hat{q}\}$; mede-se a cobertura empírica em pontos *novos*.

Dois escores. A diferença está só em como medimos o “erro de conformidade”:

(A) Escore absoluto (quantil fixo)

$$S = |y - \hat{\mu}(x)| \Rightarrow [\hat{\mu}(x) \pm \hat{q}]$$

largura **constante**.

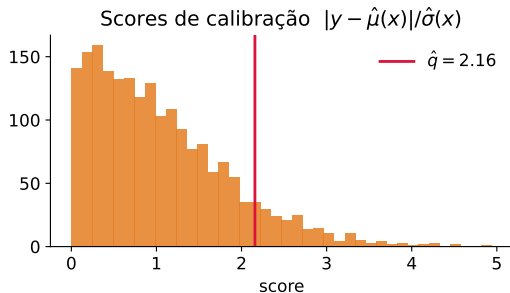
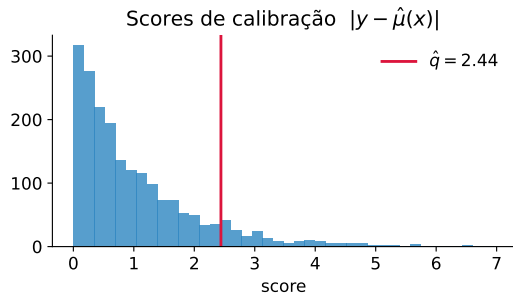
(B) Escore normalizado (modela a variância)

$$S = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x)} \Rightarrow [\hat{\mu}(x) \pm \hat{q} \hat{\sigma}(x)]$$

largura **adaptativa** ao x .

Exemplo prático: os scores de calibração e o corte $\hat{q}$

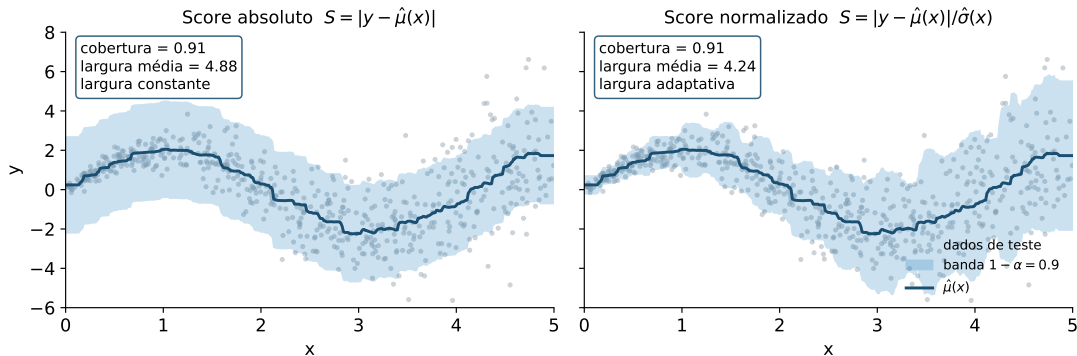
Na **calibração**, cada escore vira um número; o corte $\hat{q}$ é a k -ésima estatística de ordem ($\approx$ quantil $(1 - \alpha)$ empírico). É este corte que define a banda usada no teste.



O escore normalizado divide o resíduo por $\hat{\sigma}(x)$: pontos em regiões ruidosas deixam de dominar o corte.

Exemplo prático: as duas bandas no teste ($1 - \alpha = 0.9$)

Mesma cobertura (90%), geometria diferente sob heterocedasticidade



Avaliação no teste: para cada x novo, verifica-se se $y \in \text{banda}$. A **cobertura** é a fração de acertos.

Ambos atingem $\approx 90\%$ **marginal**, mas o escore normalizado dá bandas mais estreitas onde o ruído é baixo e mais largas onde é alto — cobertura mais *uniforme* em x .

2. Marginal vs. condicional na calibração

Na prática: *uma* calibração é sorteada, *um* limiar $S_{(k)}$ é fixado e usado para todas as predições futuras. A cobertura realizada é então uma **variável aleatória**.

Experimento marginal
(muitos pares calibração+teste)

$$\text{Cov} = \mathbb{P}(S_{n+1} \leq \hat{q}_{n,1-\alpha})$$

média sobre *tudo*.

Condicional na calibração
(calibração fixada, muitos testes)

$$C_{n,k} := \mathbb{P}(S_{n+1} \leq S_{(k)} \mid S_1, \dots, S_n)$$

Propriedade da torre

$$\mathbb{P}(S_{n+1} \leq \hat{q}_{n,1-\alpha}) = \mathbb{E}[\mathbb{P}(S_{n+1} \leq \hat{q}_{n,1-\alpha} \mid S_1, \dots, S_n)] = \mathbb{E}[C_{n,k_{1-\alpha}}].$$

A garantia usual controla apenas a **média** de $C_{n,k}$ — não sua distribuição.

3. A cobertura realizada é exatamente uma Beta — prova

Proposição 2 — lei da estatística de ordem

No caso i.i.d. contínuo, $S \sim F_S$, $\{S_i\} \stackrel{iid}{\sim} F_S$:

$$C_{n,k} = F_S(S_{(k)}) \sim \text{Beta}(k, n+1-k) =: \beta_{n,k}.$$

Esboço da prova (transformada integral de probabilidade):

- ❶ Como $\tilde{T} := S_{n+1}$ é independente da calibração e tem c.d.f. F_S ,

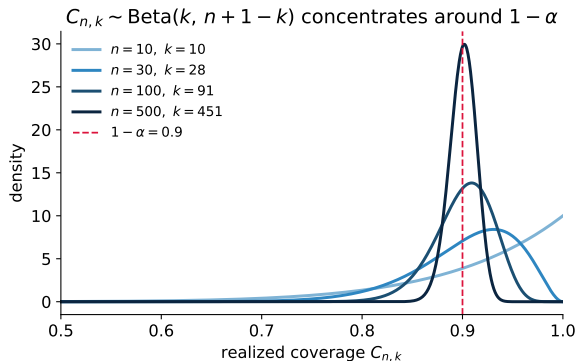
$$C_{n,k} = \mathbb{P}(S_{n+1} \leq S_{(k)} \mid S_1, \dots, S_n) = F_S(S_{(k)}).$$

- ❷ **PIT**: $U_i := F_S(S_i) \sim \text{Unif}(0, 1)$ i.i.d., e F_S monótona $\Rightarrow F_S(S_{(k)}) = U_{(k)}$, a k -ésima estatística de ordem uniforme.
- ❸ A densidade de $U_{(k)}$ é

$$f_{U_{(k)}}(u) = \frac{n!}{(k-1)!(n-k)!} u^{k-1} (1-u)^{n-k} = \text{Beta}(k, n+1-k). \quad \square$$

A lei *não* depende de F_S : vale para qualquer escore contínuo (universalidade).

3. A cobertura realizada é exatamente uma Beta — consequência



A garantia marginal é simplesmente a **média** desta beta:

$$\mathbb{E}[C_{n,k_{1-\alpha}}] = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \in \left[1-\alpha, (1-\alpha) + \frac{1}{n+1} \right].$$

Conclusão

A beta não é só uma rota para o limite marginal — é a **lei em amostra finita** da cobertura realizada (Vovk 2012; Angelopoulos–Bates 2021). Tudo (cobertura, caudas, variabilidade) se lê dela.

A lei da cobertura realizada com $k_{1-\alpha} = \lceil (n+1)(1-\alpha) \rceil$.

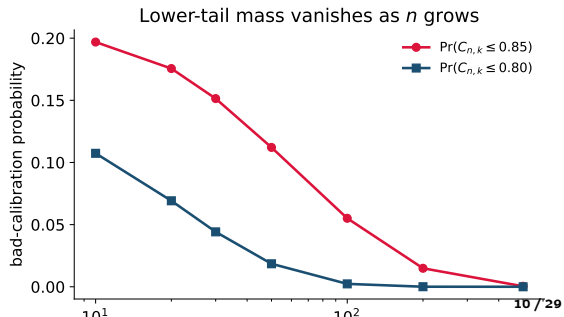
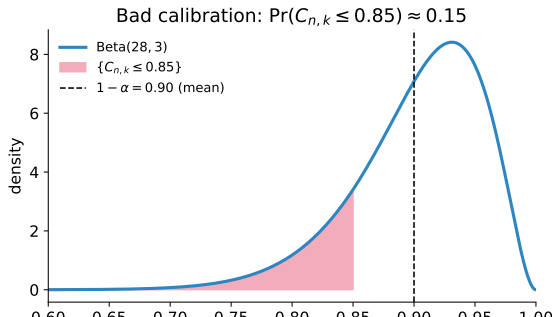
3. Má calibração: a cauda inferior da beta

Evento de má calibração: a calibração sorteada produz cobertura realizada bem abaixo do nominal. Invisível para uma garantia que controla só a média.

Teorema 3 — probabilidades de má calibração

Com $B_{n,k} \sim \beta_{n,k}$, para todo $t \in [0, 1]$:

$$\mathbb{P}(C_{n,k} \leq t) = \mathbb{P}(B_{n,k} \leq t).$$



3. Intuição: por que a marginal “funciona” nos dois casos

- A cobertura realizada $C_{n,k} \sim \text{Beta}(k, n+1-k)$ tem média $\frac{k}{n+1} \approx 1-\alpha$ e variância $\frac{k(n+1-k)}{(n+1)^2(n+2)} = O(\frac{1}{n})$.
- Quando n não é pequeno, a beta **se concentra** em torno de $1-\alpha$: a média (garantia marginal) passa a representar bem o caso típico, e a cauda de má calibração se torna desprezível.
- Quando n é pequeno, média e típico divergem: a garantia marginal continua válida *em média*, mas uma calibração específica pode estar longe.

A garantia marginal é a média de uma lei que, para n grande, é quase um ponto em $1-\alpha$.

4. Controlar a lei condicional $\Rightarrow$ controlar a cobertura

Se sabemos a lei de $C_{n,k}$ (a beta no caso ideal), sabemos **tudo**:

- A **cobertura marginal** é a média: $\text{Cov}(k) = \mathbb{E}[C_{n,k}]$.
- As **probabilidades de má calibração** são as caudas: $\mathbb{P}(C_{n,k} \leq t)$.
- A **variabilidade entre deployments** é o espalhamento da beta.

Ideia central do paper

A lei beta $\beta_{n,k}$ é o **objeto de referência**. Saber controlar a lei condicional da cobertura realizada é saber controlar a cobertura — e seus desvios.

Pergunta: e quando os dados *não* são i.i.d. (shift, dependência, escores não idênticos)? A lei condicional deixa de ser exatamente beta. **Quão longe ela vai?**

5. O que queremos estudar: a marginal da beta vs. de outra lei

Sob uma lei conjunta geral P , calibração $\mathcal{C}_n = (T_1, \dots, T_n)$, teste $\tilde{T}$:

$$D_{n,k} := \mathbb{P}(\tilde{T} \leq T_{(k)} \mid \mathcal{C}_n), \quad \nu_{n,k} := \mathcal{L}(D_{n,k}).$$

No caso i.i.d.: $\nu_{n,k} = \beta_{n,k}$. Em geral, $\nu_{n,k}$ é uma versão **deformada / transportada** da beta.

O alvo: diferença de duas esperanças

O *gap* de cobertura é a diferença entre a marginal da lei realizada e a da referência beta:

$$\text{Cov}(k) - \frac{k}{n+1} = \mathbb{E}_{\nu_{n,k}}[D] - \mathbb{E}_{\beta_{n,k}}[B] = \int_0^1 u d\nu_{n,k}(u) - \int_0^1 u d\beta_{n,k}(u)$$

Queremos **limitar** essa diferença. Ferramenta natural: **transporte ótimo**.

6. Transporte ótimo: o problema de Monge (1781)

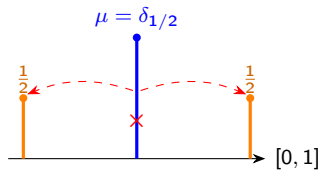
Origem histórica. Qual a forma mais barata de mover uma pilha de terra (μ) para um buraco (ν)? Monge busca um **mapa determinístico** T que leva cada partícula x a um destino $T(x)$:

$$T: \inf_{T_{\#}\mu=\nu} \int d(x, T(x)) d\mu(x), \quad T_{\#}\mu = \nu \text{ (empurra } \mu \text{ em } \nu).$$

Problema: T pode nem existir. Um mapa *não divide* massa — cada x vai a um único ponto. Tome

$$\mu = \delta_{1/2}, \quad \nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

A única partícula em $\frac{1}{2}$ teria que ir a 0 e a 1 ao mesmo tempo. Qualquer T manda $\frac{1}{2}$ num só ponto $\Rightarrow T_{\#}\mu$ é um único átomo $\neq \nu$. **Não existe mapa de Monge.**



Massa indivisível $\Rightarrow$ nenhum T satisfaz $T_{\#}\mu = \nu$.

6. Relaxação de Kantorovich: acoplamentos

Ideia. Em vez de um mapa, permita **dividir** massa: um *acoplamento* π é uma lei conjunta em $\mathcal{X} \times \mathcal{X}$ com marginais μ e ν .

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int d(x, y)^p d\pi(x, y) \right)^{1/p}, \quad \Pi(\mu, \nu) = \{\text{acoplamentos}\}.$$

Acoplamentos **sempre existem** (p. ex. o produto $\mu \otimes \nu$) — o problema é sempre bem posto.

O exemplo anterior resolvido. O acoplamento

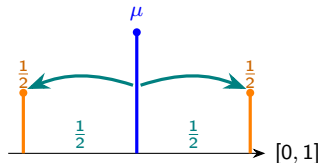
$$\pi = \frac{1}{2} \delta_{(1/2, 0)} + \frac{1}{2} \delta_{(1/2, 1)}$$

manda *metade* da massa de $\frac{1}{2}$ para 0 e metade para 1.

Custo:

$$W_1(\mu, \nu) = \frac{1}{2} \left| \frac{1}{2} - 0 \right| + \frac{1}{2} \left| \frac{1}{2} - 1 \right| = \frac{1}{2}.$$

O que Monge proíbe, o acoplamento permite.



Aqui $\mathcal{X} = [0, 1]$ (a **escala de cobertura**), $d(u, v) = |u - v|$.

7. Dualidade $W_1 \Rightarrow$ bound do erro de cobertura

Proposição 6 — dualidade de Kantorovich–Rubinstein

$$W_1(\mu, \nu) = \sup_{\|f\|_{\text{Lip}} \leq 1} \left| \int f d\mu - \int f d\nu \right|.$$

Tomando $f(u) = u$ (que é 1-Lipschitz em $[0, 1]$):

$$\left| \int u d\mu - \int u d\nu \right| \leq W_1(\mu, \nu).$$

Aplicando às leis de cobertura $\nu_{n,k}$ e $\beta_{n,k}$:

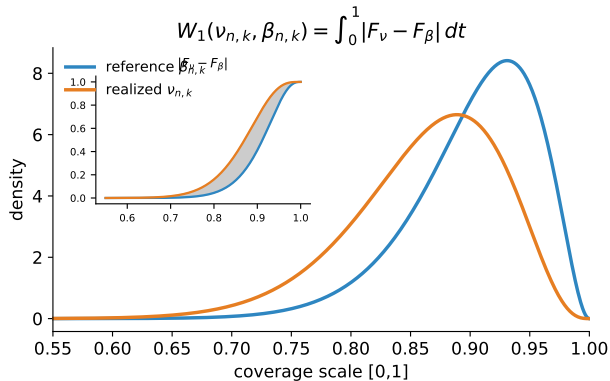
Teorema 11 — gap de cobertura via Wasserstein

$$\left| \text{Cov}(k) - \frac{k}{n+1} \right| \leq W_1(\nu_{n,k}, \beta_{n,k})$$

Se $\nu_{n,k} \in \mathcal{B}_\rho(\beta_{n,k}, \rho)$ então $|\text{Cov}(k_{1-\alpha}) - (1-\alpha)| \leq \rho + \frac{1}{n+1}$.

Controlar W_1 entre a lei realizada e a beta ideal $\Rightarrow$ controlar o erro de cobertura.

7. Visualizando: W_1 é a área entre as c.d.f.s



Fórmula 1-D (Prop. 8)

$$W_1(\nu, \beta) = \int_0^1 |F_\nu(t) - F_\beta(t)| dt.$$

Em $[0, 1]$, W_1 é **exatamente** a área cinza entre as duas funções de distribuição (inset).

Mapa monótono ótimo. O transporte ótimo em 1-D é o **rearranjo crescente**

$$T_{\text{mon}} = F_\nu^{-1} \circ F_\beta,$$

que leva o quantil- p da beta no quantil- p de ν . É o acoplamento de menor custo — preserva a ordem, nunca cruza “massa”.

Aplicações — uma taxonomia dos cenários

Antes das aplicações, organizamos os casos por **duas perguntas**:

- 1 O **teste** é independente da calibração? (*desacoplado* vs. *acoplado*)
- 2 A **calibração** é i.i.d. entre si?

	Teste $\perp$ calibração (desacoplado)	Teste acoplado (depende da calib.)
Calib. i.i.d.	mesma lei: caso ideal $\nu = \beta$ lei diferente: shift (a) $\nu = h_{\#}\beta$	(não natural)
Calib. dependente	deforma a lei da ordem	série temporal mixing (b)

Dois mecanismos distintos:

- (a) **shift no teste** $\Rightarrow$ *transporta* a beta por um mapa h na escala de cobertura;
- (b) **dependência na calibração** $\Rightarrow$ *deforma* a própria lei da estatística de ordem.

(a) Shift de distribuição — o problema

Cenário (Sec. 4.1). Calibração i.i.d. de F_T ; teste **independente** da calibração, mas de uma lei **diferente** $F_{\tilde{T}}$ (covariate/label shift).

- O limiar $T_{(k)}$ é construído sob a lei de calibração F_T ...
- ...mas a cobertura é avaliada sob a lei de teste $F_{\tilde{T}}$.

Pergunta

A garantia $\text{Beta}(k, n + 1 - k)$ era exata sob a mesma lei. **Quanto a cobertura se degrada** quando a lei de teste muda? Em que direção (sub ou sobre cobertura)?

(a) Shift de distribuição — a aplicação

Mecanismo. A cobertura realizada se escreve na escala beta via o mapa $h = F_{\tilde{T}} \circ F_T^{-1}$:

$$D_{n,k} \stackrel{d}{=} h(B_{n,k}), \quad \nu_{n,k} = h_{\#} \beta_{n,k}.$$

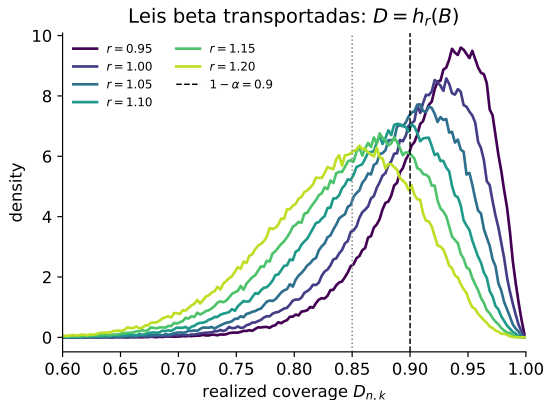
O shift não altera o mecanismo da estatística de ordem — só o transporta por h .

Exemplo concreto (half-normal scale shift). Erro Gaussiano, escore $|y - \hat{\mu}(x)|$ é half-normal. Com razão de escala $r = \sigma_{\tilde{T}}/\sigma_T$ e $q(u) = \Phi^{-1}(\frac{u+1}{2})$:

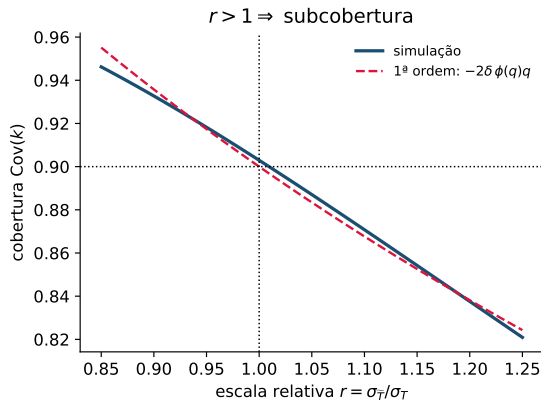
$$h_r(u) = 2\Phi\left(\frac{q(u)}{r}\right) - 1.$$

$r > 1$ (teste mais disperso) $\Rightarrow h_r(u) < u \Rightarrow$ subcobertura; $r < 1 \Rightarrow$ sobrecobertura.

(a) Shift de distribuição — o resultado



Esquerda: a lei realizada $D = h_r(B)$ é a beta transportada; $r > 1$ a desloca para a esquerda.



Direita: $\text{Cov}(k)$ vs r . A simulação dá 0.871 em $r = 1.1$ (perda ≈ 0.03), batendo com a aproximação de 1ª ordem $-2\delta \phi(q)q$.

(a) Shift — o que o Teorema diz sobre o erro de cobertura

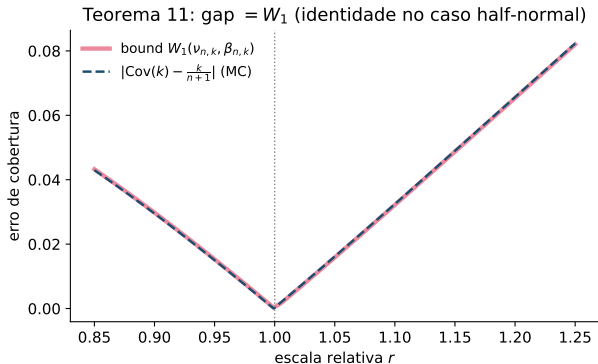
Teorema 11 (gap de cobertura)

$$\left| \text{Cov}(k) - \frac{k}{n+1} \right| \leq W_1(\nu_{n,k}, \beta_{n,k}).$$

No caso half-normal o mapa h_r é **monótono**, então o transporte ótimo é o acoplamento monótono e o bound vira **identidade exata**:

$$\left| \text{Cov}(k) - \frac{k}{n+1} \right| = \mathbb{E} |h_r(B) - B| = W_1.$$

Em $r = 1.1$: gap = 0.033 e $W_1 = 0.033$ (Monte Carlo) — coincidem.



Validação MC: as duas curvas se sobrepõem em todo r , confirmando que o bound é *atingido* (não há folga) neste modelo.

(b) Dependência temporal (mixing) — o problema

Cenário (Sec. 4.3). Escore é um **processo estacionário** (T_i) (série temporal); calibração $\mathcal{C}_n = (T_1, \dots, T_n)$ **não é i.i.d.**, e o teste $\tilde{T} = T_{n+\ell}$ vem ℓ passos à frente.

- $U_i = F_T(T_i)$ são marginalmente uniformes, mas **dependentes**.
- A lei de $U_{(k)}$ *não* é mais exatamente beta — a dependência deforma a própria estatística de ordem.

Pergunta

Quão longe da beta a lei realizada vai, em função da força da dependência?

(b) Dependência temporal (mixing) — a aplicação

Mecanismo. Coeficientes de mixing (α -, ϕ -, β -mixing) medem o decaimento da dependência. Sob mixing, a lei de $U_{(k)}$ admite uma aproximação **Berry–Esseen**:

$$\sup_x \left| \mathbb{P}(\sqrt{n}\{U_{(k_{1-\alpha})} - (1-\alpha)\} \leq x) - \Phi\left(\frac{x}{\tau_{1-\alpha}}\right) \right| \leq a_n \rightarrow 0.$$

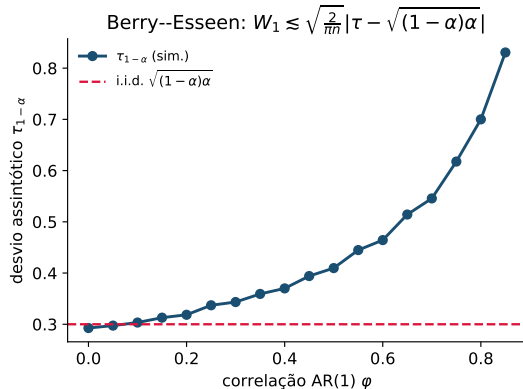
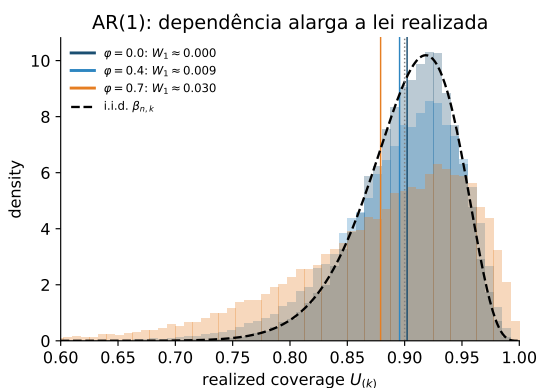
Isso basta para limitar a distância de Wasserstein à beta de referência:

Proposição 15 — transporte quantil-a-beta

$$W_1(\mathcal{L}(U_{(k_{1-\alpha})}), \beta_{n, k_{1-\alpha}}) \leq a_n + \sqrt{\frac{2}{\pi n}} |\tau_{1-\alpha} - \sqrt{(1-\alpha)\alpha}| + O(n^{-1/2}).$$

A i.i.d. tem $\tau_{1-\alpha} = \sqrt{(1-\alpha)\alpha}$; o excesso mede o custo da dependência.

(b) Dependência temporal (mixing) — o resultado



Esquerda: num AR(1) latente, a correlação φ alarga a lei realizada em torno de $1 - \alpha$; o W_1 empírico à beta cresce $0.00 \rightarrow 0.03$.

Direita: o desvio assintótico $\tau_{1-\alpha}$ sobe acima do valor i.i.d. $\sqrt{(1-\alpha)\alpha}$ — exatamente o termo que o bound de Berry–Esseen prevê.

(b) Mixing — o erro de cobertura (1/2): o que o Teorema diz

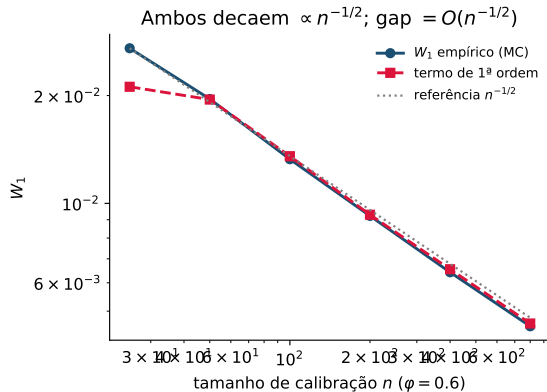
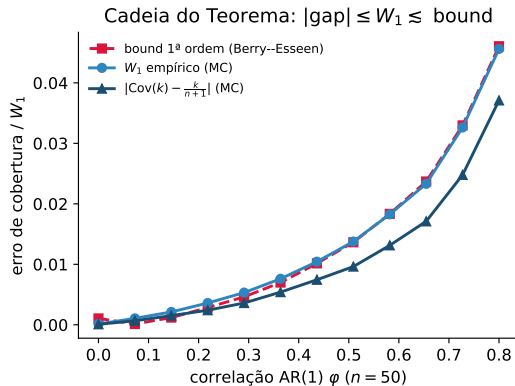
Cadeia de cobertura

$$\left| \text{Cov}(k) - \frac{k}{n+1} \right| \underbrace{\leq}_{\text{Teorema 11}} W_1(\nu_{n,k}, \beta_{n,k}) \underbrace{\lesssim}_{\text{Berry-Esseen}} \sqrt{\frac{2}{\pi n}} |\tau_{1-\alpha} - \sqrt{(1-\alpha)\alpha}|.$$

- **1ª desigualdade (Thm 11):** o W_1 entre a lei realizada e a beta controla o erro de cobertura marginal.
- **2ª desigualdade (Berry-Esseen):** sob mixing, esse W_1 é limitado por um termo explícito que mede o *excesso de desvio* $\tau_{1-\alpha} - \sqrt{(1-\alpha)\alpha}$.

Ao contrário do shift, aqui o bound *não* é exato: a dependência transporta “massa” nos dois sentidos, então parte se cancela na média e $|\text{gap}|$ pode ser $< W_1$.

(b) Mixing — o erro de cobertura (2/2): validação por simulação



Esquerda ($n = 50$): a cadeia $|\text{gap}| \leq W_1 \lesssim \text{bound}$ vale em todo φ ; o termo de 1ª ordem *gruda* no W_1 real (em $\varphi = 0.6$: $\text{gap} = 0.014 \leq W_1 = 0.020 \approx \text{bound} = 0.020$).

Direita ($\varphi = 0.6$): W_1 e o bound decaem ambos $\propto n^{-1/2}$; o erro do bound é $O(n^{-1/2})$, desaparecendo com mais calibração.

- 1 Predição conforme split dá cobertura marginal $\geq 1 - \alpha$ em amostra finita.
- 2 Mas a garantia é **marginal**: faz média sobre a calibração aleatória.
- 3 A cobertura realizada (condicional na calibração) é exatamente $\text{Beta}(k, n + 1 - k)$; a garantia marginal é a sua **média**, e as caudas dão a má calibração.
- 4 Para n grande a beta concentra $\Rightarrow$ marginal $\approx$ típico.
- 5 Fora do i.i.d., a lei realizada $\nu_{n,k}$ é uma **beta transportada**.
- 6 Via dualidade de Kantorovich–Rubinstein, $|\text{Cov}(k) - \frac{k}{n+1}| \leq W_1(\nu_{n,k}, \beta_{n,k})$: **controlar W_1 controla o erro de cobertura**.

A lei beta permanece a referência; o não-i.i.d. é entendido por como ele a transporta.

Obrigado!

Conformal Prediction via Transported Beta Laws

Thiago R. Ramos Helton Graziadei Luben M. C. Cabezas

- Preprint: [arXiv:2605.19024](https://arxiv.org/abs/2605.19024)
- <https://arxiv.org/abs/2605.19024>

Aponte a câmera para o QR code →



arXiv:2605.19024