Recursive methods for predictive Bayes

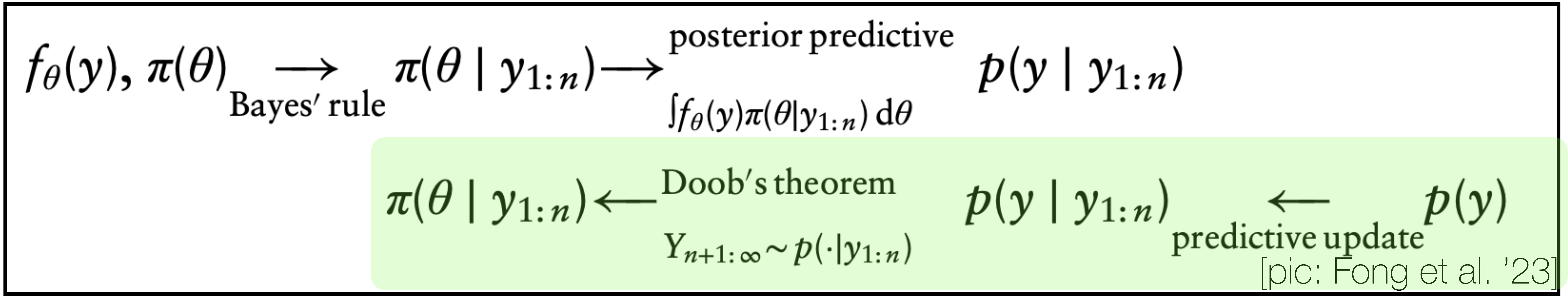
Lorenzo Cappello

Post-Bayes Seminar Series - 01/07/2025

Bayesian predictive inference

[Fortini Petrone '20, Fong et al. 23', Berti et al. '23, & others]

Let θ be a parameter of interest



[pic: Fong et al. '23]

- Notation: $P_n := P(Y_{n+1} \mid y_{1:n})$ and p_n the corresponding density
- $Y_{n+1:\infty} \sim P(\cdot \mid y_{1:n})$ may be tricky to elicit
- The sequence of one step ahead (P_n) is often easier to elicit
  - e.g., via a **learning rule**, i.e., a map $(P_{n-1}, y_n) \mapsto P_n$

Predictive Resampling

The mechanism that enables moving from prediction to posterior distribution

Bootstrap/Predictive Resampling [Rubin'81, Fong et al. '23]

Input:

- An estimator $\hat{\theta}_n = f(y_{1:n})$
- A learning rule: i.e., a map $(P_{n-1}, y_n) \mapsto P_n$
- Data $y_{1:n}$
- B (#bootstrap samples) and N (#synthetic samples)

Given $y_{1:n}$, compute P_n

For j in 1 to B repeat

For i in 1: N repeat

- Sample synthetic data $Y_{n+i} \sim P_{n+i-1}$
- Update the predictive $(P_{n+i-1}, y_{n+i}) \mapsto P_{n+i}$

Freedom in the choice of which predictive to use

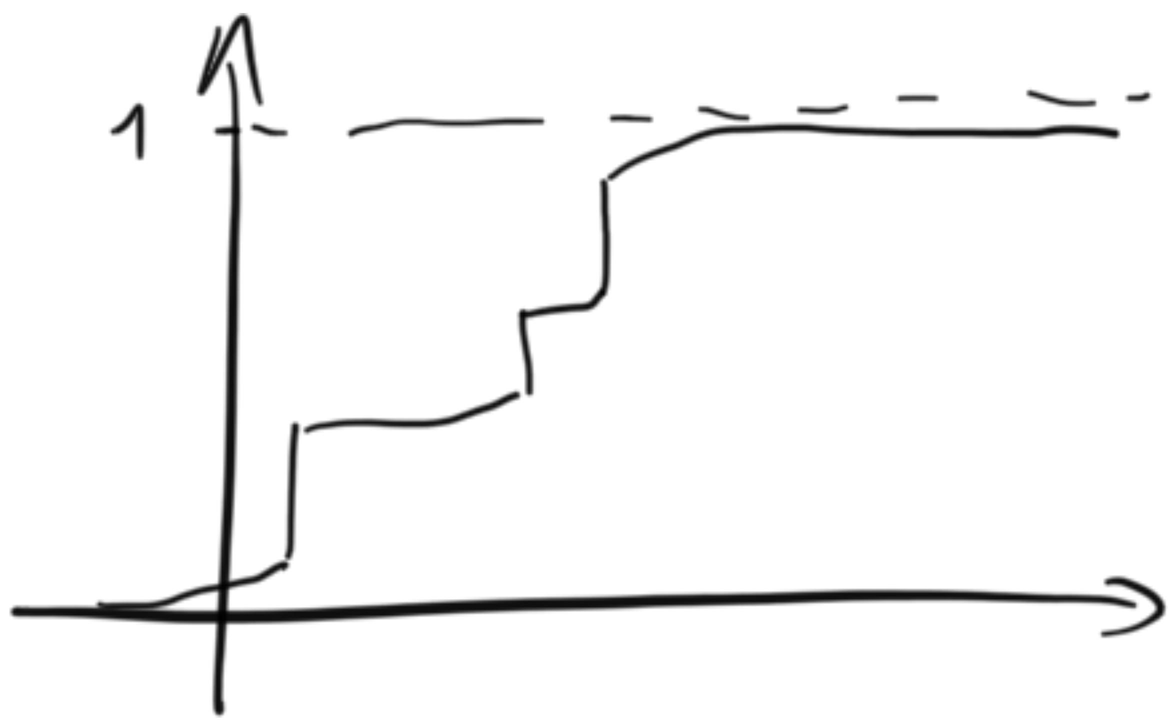
Compute $\hat{\theta}^{(i)}_{n+N} = f(y_{1:n}, y_{n+1:n+N})$

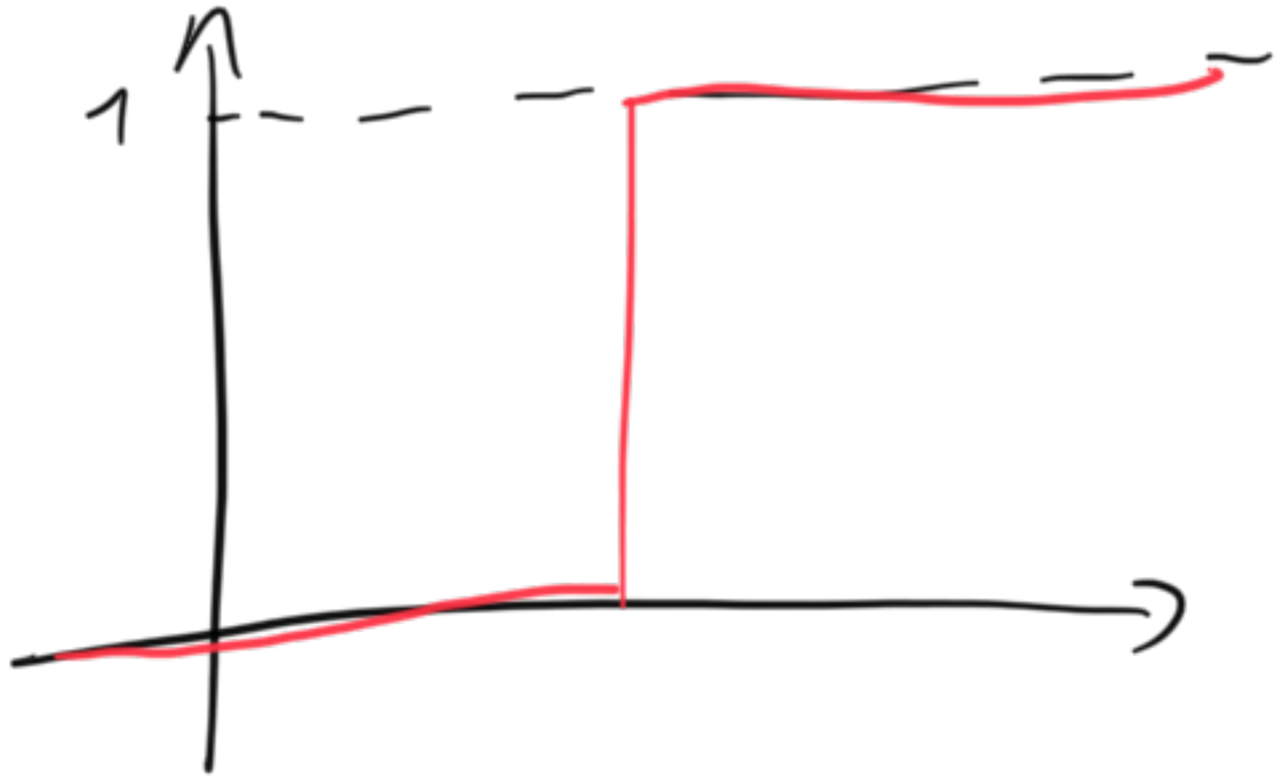
Output: $\hat{\theta}^{(1)}_{n+N}, \ldots, \hat{\theta}^{(B)}_{n+N}$

Examples

Different proposals differ by how you define $(P_{n-1}, y_n) \mapsto P_n$

1. $$P_n(\cdot) = \left(1 - \frac{1}{n}\right) P_{n-1}(\cdot) + \frac{1}{n}\delta_{Y_n}(\cdot)$$

$1 - \frac{1}{n}$ 

$+\frac{1}{n}$ 

Here, the update is a point mass. i.e.
It's the predictive of the **Dirichlet Process**
[≈ Bayesian bootstrap]
[Ferguson '73, Rubin '81]

2. …

Question 1 of the tutorial

How to come up $(P_{n-1}, y_n) \mapsto P_n$?

A. $(P_{n-1}, y_n) \mapsto P_n$ is ***given by a Bayesian model*** [e.g., Dirichlet Process]
- ***More interestingly,*** it "mimics" the predictive distribution of a Bayesian model [e.g. Dirichlet Process Mixtures]

  Today: Predictive Recursion, Gaussian Copula Algorithm

B. $(P_{n-1}, y_n) \mapsto P_n$ is ***new/almost fully new***

  Today: Any-Copula Algorithm

C. $(P_{n-1}, y_n) \mapsto P_n$ is ***taken from (neighbouring) literature***

  Today: Partition Estimators, Kernel methods, Parametric Boostrap

Question 2 of the tutorial

Once we have $(P_{n-1}, y_n) \mapsto P_n$, can we use it for predictive resampling?

Two desiderata for the resulting (P_n): (recall that $Y_n \sim P_{n-1}$)

- "Valid": the sequence of synthetic data $(Y_N)_{N\geq n}$ should satisfy some properties
  - We will require ***asymptotically exchangeable***

  Today: - (P_n) is a **martingale** → (Y_n) is **c.i.d.** [Berti et al. '04] (Trival for A., not B. C.)

  - (P_n) is NOT a martingale → **Beyond martingales** ((Y_n) is NOT **c.i.d.**)

- "Practical": we need to be able to implement this algorithm
  - We need to sample $Y_{n+1} \sim P_n$ quicky
  - We need to update $(P_{n-1}, y_n) \mapsto P_n$ quickly
  - Other numerical properties that affect computations (grid? Order dependence? Etc.)

  Today: not so much, except focusing on **recursive algorithm**

Constructing predictive recursion

Dirichlet Process Mixtures (DPMs)

Consider the mixture

$Y_i | \theta_i \sim k(\cdot | \theta_i)$

$\theta_i | G \overset{iid}{\sim} G$

$G \sim DP(cG_0)$

[Lo '84]

- **Case** $n = 1$: suppose you are given a single data y_1, the Bayes estimate of the mixing measure

$$E(G(B) | Y_1) = P(\theta_2 \in B | Y_1) = G_1(B)$$

$$= E(P(\theta_2 \in B | \theta_1) | Y_1)$$

$$= E\left(\frac{c}{c+1} G_0(B) + \frac{1}{c+1} 1_B(\theta_1) \middle| Y_1\right)$$

$$= \frac{c}{c+1} G_0(B) + \frac{1}{c+1} P(\theta_1 \in B | Y_1)$$

Weights

$$= \frac{\int_B k(Y_i | \theta) G_0(d\theta)}{p(Y_i)}$$

- It mixes prior guess and an "update"
- The update is where Bayes rule is used

- **Case** $n > 2$: usually we approximate it!

Predictive Recursion (PR) [Newton et al. '98, Newton Zhang '99]

Algo. Start with a prior guess $G_0(\cdot)$, choose a deterministic sequence (α_n), given $y_{1:n}$

for a set B **repeat**

$$G_i(B) = (1 - \alpha_i)G_{i-1}(B) + \alpha_i \frac{\int_B k(y_i \mid \theta) G_{i-1}(d\theta)}{p_{i-1}(y_i)}$$

Output: $G_n(\cdot)$

Weights

Best guess up to $i-1$

Update applying Bayes rule to a single obs.

Predictive Recursion (PR) [Newton et al. ’98, Newton Zhang ’99]

Algo. Start with a prior guess $g_0(\theta)$, choose a deterministic sequence (α_n), given $y_{1:n}$

repeat: $$g_i(\theta) = (1-\alpha_i)g_{i-1}(\theta) + \alpha_i \frac{k(y_i \mid \theta) g_{i-1}(\theta)}{p_{i-1}(y_i)}$$

Output: $g_n(\theta)$

Weights — Best guess up to $i-1$ — Update applying Bayes rule to a single obs.

- It is exact predictive when $n = 1$
- It is not when $n > 1$
- Some numerical considerations (**order dependent**, grid, numerical integration required,…)
- **As an estimator** It can be every accurate [Newton ’02, Tokdar et al. ’09, Zuanetti et al. ’19, C. Walker, ’18, Dixit Martin ’24, … and many more]
- [Fortini Petrone ’20]: if taken as a **predictive rule**, it can be used for predictive inference

Observables? What about if we want a predictive scheme for P

Algo. Start with a prior guess $g_0(\theta)$, choose a deterministic sequence (α_n), given $y_{1:n}$ repeat:

$$g_i(\theta) = (1 - \alpha_i) g_{i-1}(\theta) + \alpha_i \frac{k(Y_i \mid \theta) g_{i-1}(d\theta)}{p_{i-1}(Y_i)}$$

Then

$$p_n(y) = \int_\theta k(y \mid \theta) g_n(\theta) d\theta$$

- It is a valid predictive rule (satisfies the martingale property), it can be used for a resampling scheme [Fortini Petrone '20]
- Lots of numerical approximations involved
- Somewhat convoluted way of getting to the sequence of (P_n)

Copula-based algorithms

Copula characterization [Hahn et al. '18]

A way to update $(P_{n-1}, y_n) \mapsto P_n$

Let $p_n(y) = \int f(y \mid \theta)\pi_n(d\theta)$ be the predictive density of Y_{n+1} given $y_{1:n}$

$$\frac{p_n(y)}{p_{n-1}(y)} = \frac{\int f(y \mid \theta) f(y_n \mid \theta)\pi_{n-1}(d\theta)}{p_{n-1}(y)p_{n-1}(y_n)} = c_n\Big(P_{n-1}(y), P_{n-1}(y_n)\Big).$$

Copula density ($P_{n-1}(y)$ is the CDF)

A. It depends only on the sample size

B. It must converge to the independence copula (i.e. no update)

Example: Gaussian (unknown μ)

- $f(y \mid \theta) = \mathrm{N}(y \mid \theta, 1)$ and $\pi(\theta) = \mathrm{N}(0, \tau^{-1})$
- $p_n(y) = \mathrm{N}(y \mid \mu_n, \sigma_n^2)$
- $\dfrac{p_n(y)}{p_{n-1}(y)} = \dfrac{\int \mathrm{N}(y \mid \theta, 1)\mathrm{N}(y_n \mid \theta, 1)\pi_{n-1}(d\theta)}{p_{n-1}(y)p_{n-1}(y_n)}$

Gaussian copula density with $\rho_n = (n + \tau)^{-1}$

Back to DPMs

$$f(y\,|\,G) = \int N(y\,|\,\theta, \sigma^2)\, dG(\theta) \quad \text{and} \quad G \sim DP(c\, G_0)\,.$$

Case $n = 1$: suppose you are given y_1,

$$p_1(y) = (1-w)p_0(y) + w\,\frac{\int N(y\,|\,\theta,\sigma^2)N(y_1\,|\,\theta,\sigma^2)\, dG_0(\theta)}{p_0(y_1)}$$

& w is available in closed form

$$\frac{p_1(y)}{p_0(y)} = (1-w)1 + w\,\frac{\int N(y\,|\,\theta,\sigma^2)N(y_1\,|\,\theta,\sigma^2)\, dG_0(\theta)}{p_0(y)p_0(y_1)}$$

Independence copula

Gaussian Copula

Copula density

It is a copula mixture!

$$p_1(y) = (1-w)p_0(y) + wc_{\rho_1}(P_0(y), P_0(y_1))p_0(y)$$

- Similar to PR derivations
- We can use weights to drive convergence to independence copula

Gaussian copula algorithm [Hahn et al. '18]

Algo. Start with a prior guess $p_0(y)$, choose a deterministic sequence (α_n), given $y_{1:n}$ repeat: $p_i(y) = (1-\alpha_i)p_{i-1}(y)+\alpha_i p_{i-1}(y)c_\rho(P_{i-1}(y), P_{i-1}(y_i))$

Output: $p_n(y)$

- Note, α_i drive the convergence the independence copula
- We bypass the need to compute G_n and do numerical integration
- It is exact predictive when $n = 1$
- It is not when $n > 1$
- Some numerical considerations (**order dependent**, grid, NO numerical integration required,…)
- **As an estimator** It can be every accurate [Hahn et al. '18, Fong et al. '23]
- [Fong et al. '23]: if taken as a **predictive rule**, it can be used for predictive inference

Same as PR

Nothing special about location-DPM

Let $f(y\,|\,G) = \int N(y\,|\,\theta, \sigma^2)\, dG(\theta, \sigma^2)$ and $G \sim DP(c\, G_0)$. Given y_1,

$$\frac{p_1(y)}{p_0(y)} = (1-w)1 + w\, c_{\rho_n,\nu}(P_0(y), P_1(y))$$ Student t copula density

Algo. Start with a prior guess $p_0(y)$, choose a deterministic sequence (α_n), given $y_{1:n}$ repeat: $p_i(y) = (1-\alpha_i)p_{i-1}(y) + \alpha_i p_{i-1}(y) c_{\rho,\nu}(P_{i-1}(y), P_{i-1}(y_i))$

Output: $p_n(y)$

- If taken as a sequence of predictives, it defines a martingale
- Same reasoning can be extended to more complex models
    - Multivariate/Regression [Fong et al. '23]
    - DPM of Linear Regression [C. Walker '25a]

Intermediate "lessons"

- Starting from a Bayesian model, it is fairly easy to get a martingale
- Practical algorithms
- Can we avoid taking a Bayesian model as starting point?

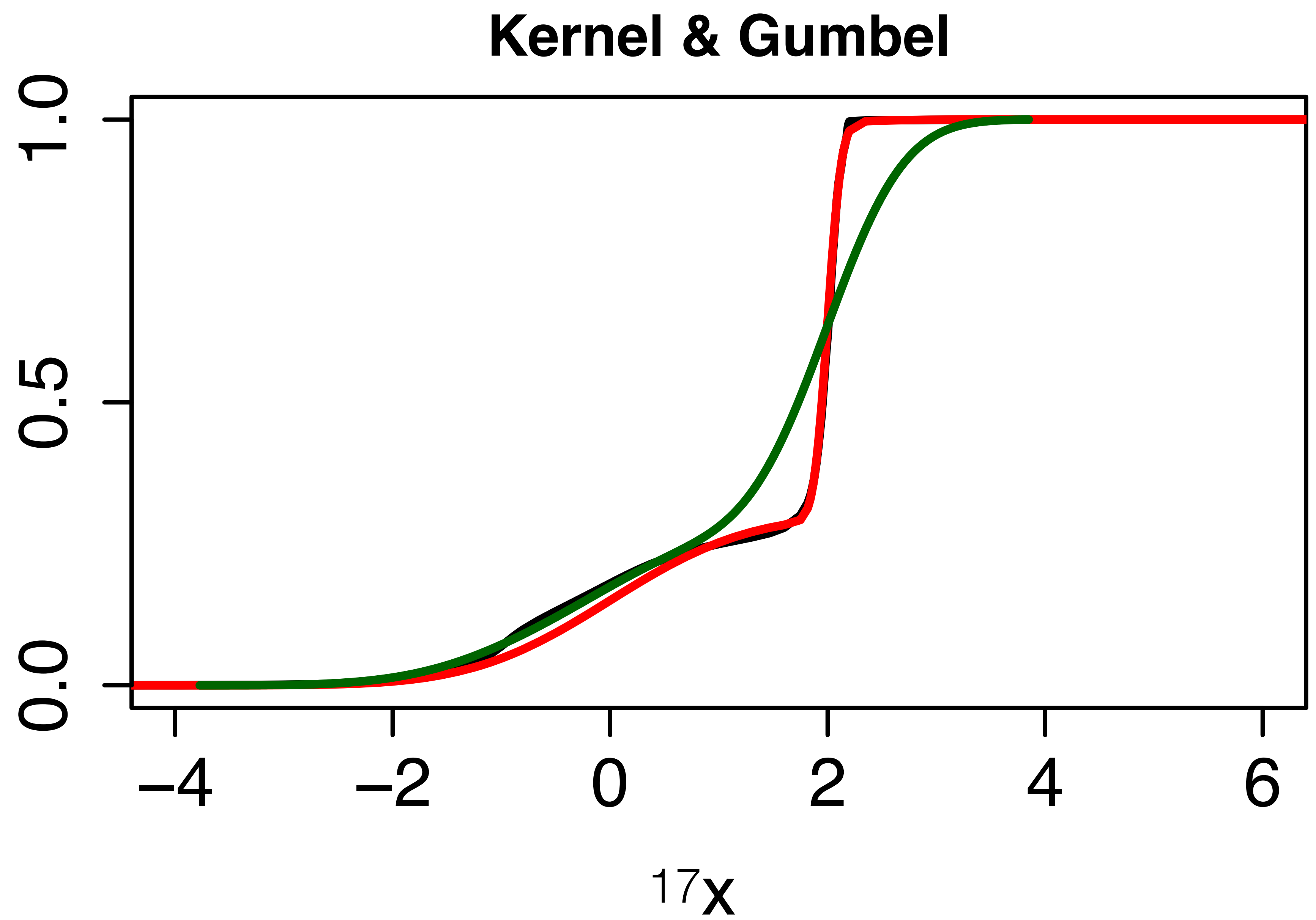
Any-copula

Algo. Start with a prior guess $p_0(y)$, choose a deterministic sequence (α_n), given $y_{1:n}$ repeat: $p_i(y) = (1 - \alpha_i)p_{i-1}(y) + \alpha_i p_{i-1}(y)c_\theta(P_{i-1}(y), P_{i-1}(y_i))$

Output: $p_n(y)$

- We did not start from a (known) Bayesian model
- We chose a copula c_θ and use the weights as before
- Example: Gumbel copula [C. Walker '25b]

Black: truth
Green: kernel density estimator (Silverman's rule)
Red: Algo. With Gumbel Copula



Can we use it for predictive resampling?

Validity: **Martingale condition**

$$\mathbb{E}(P_{n+1}(B) \mid Y_{1:n}) = (1 - \alpha_{n+1})P_n(B) + \alpha_{n+1}\mathbb{E}\left(\int_B p_n(y)c_\theta(P_n(y), P_n(Y_{n+1}))dy \,\middle|\, Y_{1:n}\right)$$

$$\mathbb{E}\left(\int_B p_n(y)c_\theta(P_n(y), P_n(Y_{n+1}))dy \,\middle|\, Y_{1:n}\right) = \iint_B p_n(y)p_n(y_{n+1})c_\theta(P_n(y), P_n(y_{n+1}))dydy_{n+1}$$

$$= \iint_B p_n(y)p_n(y_{n+1})\frac{p(y, y_{n+1})}{p_n(y)p_n(y_{n+1})}dydy_{n+1} = P_n(B)$$

- The resulting sequence is (Y_n) is c.i.d.
- The connection between copulas and c.i.d. sequences is deeper [Bissiri Walker '25]

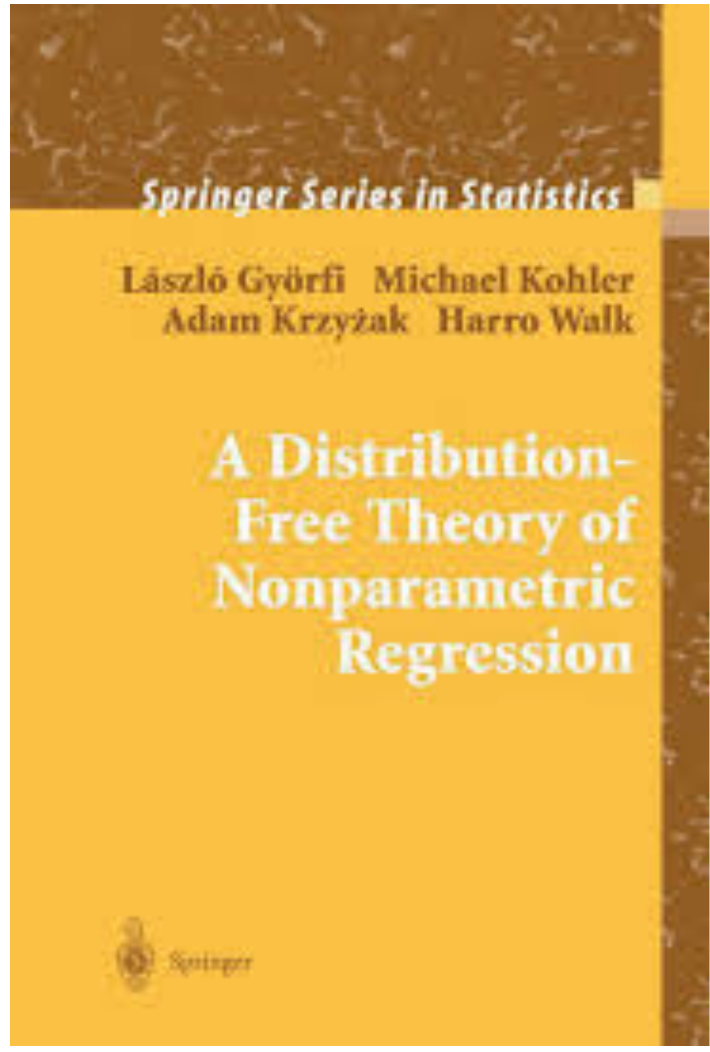
- We can sample from copula-based predictive algorithms [Fong et al. '23]
- It is clearly very fast

Copulas define a "valid" $(P_{n-1}, y_n) \mapsto P_n$

$Y_{n+1} \sim P_n$ is "practical"

Beyond Bayesian models

(Recursive) Partition estimator [Tukey '47,'61, Colomb '77,…]



[Gyorfi et al. '12,Ch. 4,25]

Consider

- $(\mathscr{P}_n)$: sequence of nested dyadic partitions of $[0,1]$, with $\mathscr{P}_0 = \{A_0 = [0,1]\}$, $\mathscr{P}_1 = \{A_{1,0} = [0,1/2), A_{1,1} = [1/2,1]\}$ …
- $A_n(x)$: the cell containing x

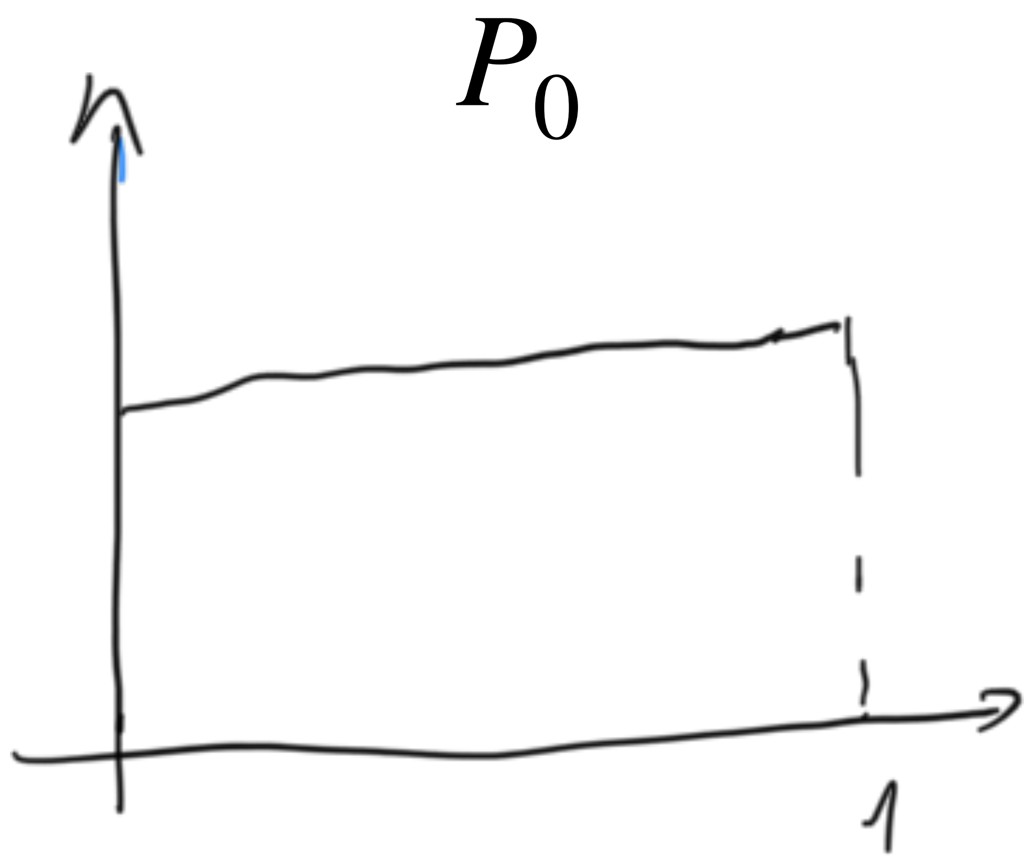
Recursive predictive scheme

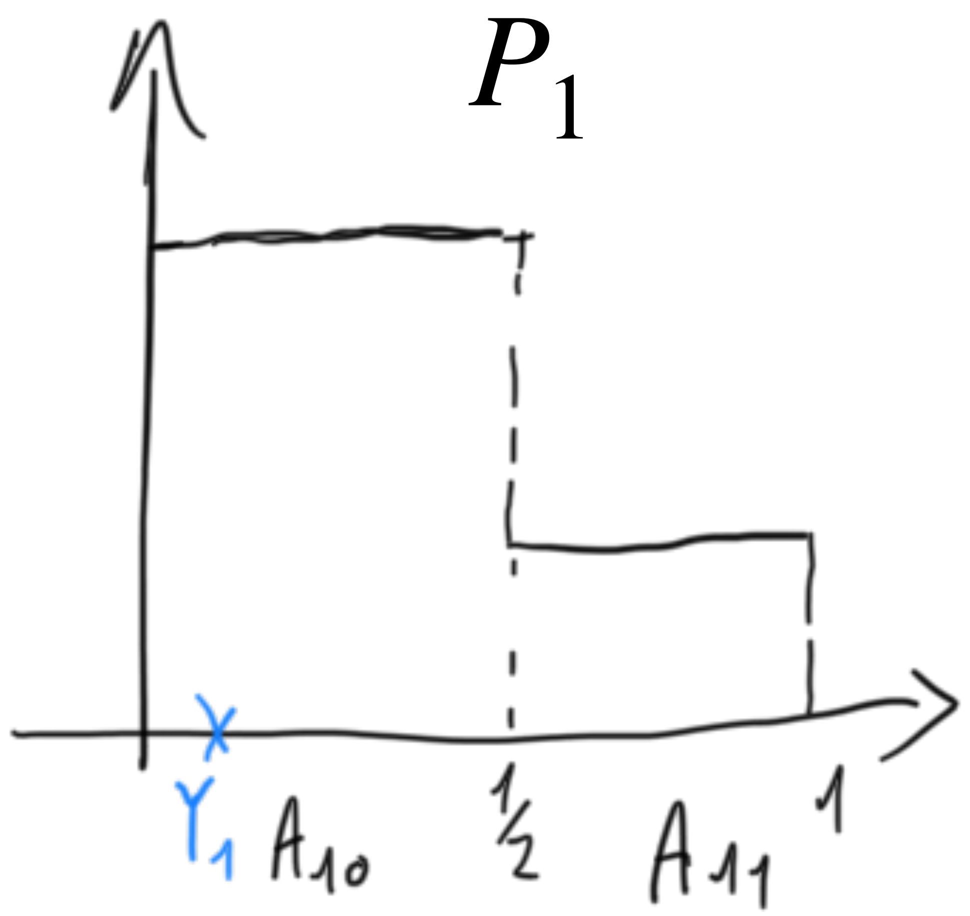
$$Y_1 \sim \text{Unif}(A_0) := P_0$$

$$Y_2 \,|\, Y_1 \sim \frac{1}{2}\text{Unif}(A_0) + \frac{1}{2}\text{Unif}(A_1(Y_1)) := P_1$$

…

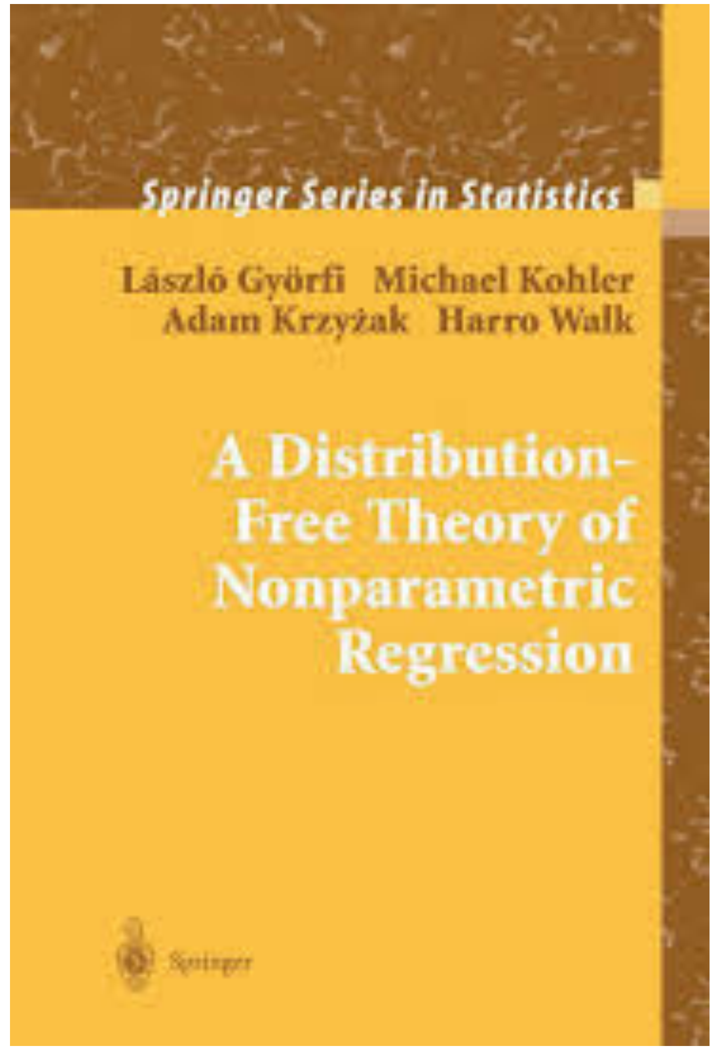
$$Y_{n+1} \,|\, Y_{1:n} \sim \left(1 - \frac{1}{n}\right) P_n + \frac{1}{n}\text{Unif}(A_n(Y_n)) := P_{n+1}$$





- Update $(P_{n-1}, y_n) \mapsto P_n$ is "valid": (P_n) is martingale.
- It can be used for nonparametric regression etc.
- Not sure how "practical" it is. (Definitely order dependent etc.)

(Recursive) Kernel estimator [Parzen '62, Rosenblatt'56, Akaike '54,…]



Consider

- Kernel $K : \mathbb{R} \mapsto \mathbb{R}^+$ such that $\int K(u)du = 1$
- Bandwidth (h_n) with $h_n \to 0$

[Gyorfi et al. '12,Ch. 5,25]

Recursive predictive scheme

$Y_1 \sim K := P_0$

- $Y_{n+1} \mid Y_{1:n} \sim P_n^c$ with $p_n^c(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - Y_i}{h_n}\right)$ (classical kernel estimator)
- $Y_{n+1} \mid Y_{1:n} \sim P_n^r$ with $p_n^r(x) = \left(1 - \frac{1}{n}\right) P_{n-1}^r + \frac{1}{nh_n} K\left(\frac{x - Y_n}{h_n}\right)$ (recursive kernel estimator)

- (P_n^c) is a martingale if and only iff Kernel is Laplace with a specific scale [Thm1 West '91]
- (P_n^r) is not a martingale [Battiston & C. '25]

Beyond Martingales

Parameter updates

Several proposals [Holmes Walker '23, Garelli et al. '24, Fong Yiu '24, Fortini Petrone '25…] moved the recursive update to a parametric family. With appropriate initialization, **repeat**

$$Y_{n+1} \mid \hat{\theta}_n \sim P_{\hat{\theta}_n}$$

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \eta_n f(\hat{\theta}_{n-1}, Y_n, \dots)$$

Example [Holmes Walker '23]: Gaussian

- $Y_{n+1} \sim \mathrm{N}(\hat{\theta}_n, 1)$
- $\hat{\theta}_n = \bar{Y}_n$ (sample mean)

Is this "valid" update?

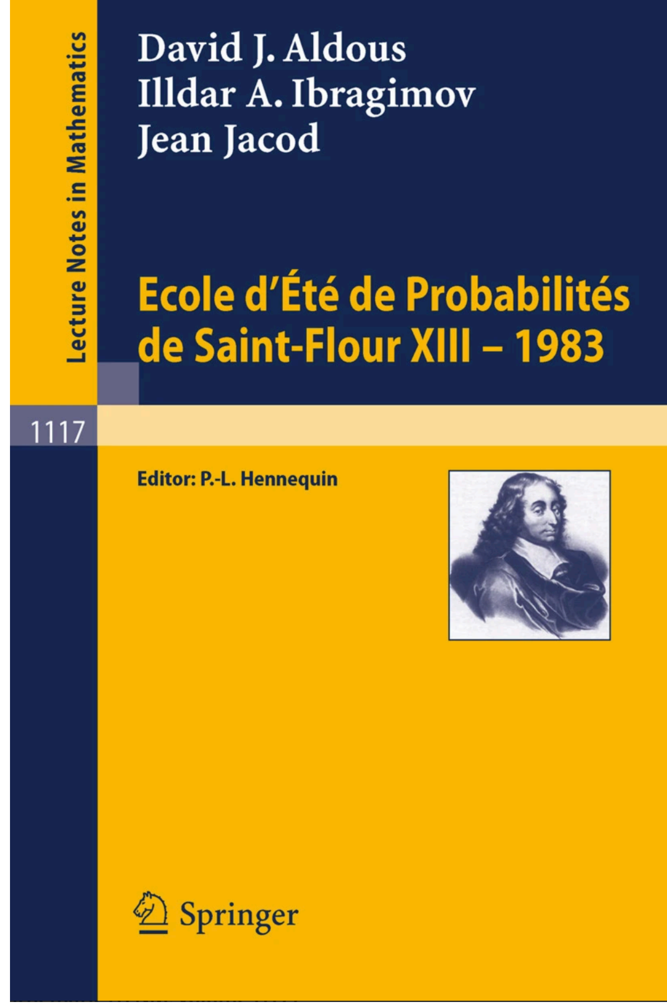
Validity: **Martingale condition**

$$\mathbb{E}[P_{n+1}(B) \mid Y_{1:n}] = \mathbb{E}[\mathrm{N}(B \mid \bar{Y}_{n+1}, 1) \mid Y_{1:n}] = \int\int_B \phi\left(y \middle| \frac{n}{n+1}\bar{Y}_n + \frac{1}{n+1}y_{n+1}, 1\right) \phi\left(y_{n+1} \middle| \bar{Y}_n, 1\right) dy dy_{n+1}$$

$$= \mathrm{N}\left(B \middle| \bar{Y}_n, 1 + \frac{1}{(n+1)^2}\right)$$

Not a martingale!

What other options we have?



[Aldous '85]

For a general sequence (P_n), if all we are trying to establish is that the resulting (Y_n) sequence is **asymptotic exchangeability**, we need to show [Aldous '85, Lemma 8.2]

$$\mathbb{P}(\omega \in \Omega : P_n(\,\cdot\,,\omega) \xrightarrow{w} \tilde{F}(\,\cdot\,,\omega)) = 1$$

- If (P_n) not a martingale, the c.i.d. property is lost
- This makes establishing the above somewhat non-trivial, as we don't even have this conditional identical distributed property

Convergence of parameters/moments

- $(\hat{\theta}_n)$ **is martingale:** Prove parameter convergence using martingales property, and then one is pratically done
    - Parametric Bayesian Boostrap [Holmes Walker '23, Fong Yiu '24]
    - Logistic regression (Sandra's talk) [Fortini Petrone '25]

- **Convergence of sample mean and sample variance for non i.i.d. random variables** appropriately constructed
    - Predictive distributions driven by 1st and 2nd sample moments [Garelli et al. '25]

Both approaches cover the Gaussian example and much more general examples

Larger classes of a.s. weakly convergent r.m.s.

- (P_n) belongs to a class of random measures a.s. weakly convergent
  - [Battiston & C. '25] Informally, if (P_n) satisfies $|P(X_{n+2} \in B \mid Y_{1:n}) - P(X_{n+1} \in B \mid Y_{1:n})| \leq \xi_n$ and $\sum \xi_n < \infty$ a.s.

- The martingale property is acquired asymptotically
- It covers the Gaussian example
- It also covers recursive kernel estimators (not-order dependent in the data, no grid required, easy to sample)
- Talk on YouTube - Post Bayes Workshop (same channel as the seminar series)

Discussion

Discussion

- We have reviewed several possible constructions of algorithms that can be used for these **bootstrap-type schemes**
  - They can come mimicking $n = 1$ update of existing Bayesian models (Predictive Recursion, Gaussian Copula,…)
  - They can be new (Any-copula, Predictive sequence driven by moments, SGD driven predictives,…)
  - They can be borrowed from other literature (Recursive Partion estimators,…)

- We have discussed ways of checking their suitability for these schemes
  - "Valid" update (Martingale conditions, …)
  - "Practical" (Being able to sample from, fast update,…)

Themes not touched (Maybe simply because there is not enough out there)

- Which scheme to use?
  - For specific applications, it will depend on "*suitability to the available data*" [Garelli et al. '24]
  - Still, maybe something can be done to understand the properties. I like a "*forward* $(n+1:\infty)$-*backward* $(1:n)$ view in [Fong Yiu '24]

- More practical ways of establishing validity
  - All approaches require some form of mathematical tractability

Thanks

lorenzo.cappello@upf.edu comments welcome!

Funding:

- Ramon y Cajal (Ministerioo de cienca, innovacion y universidades) (RYC2022-038467-I)
- Spanish Ministry of Economy and Competitiveness grant PID2022-138268NB-I00

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

References (In the order at which they appeared)

- Fortini, S., S. Petrone, et al. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. Journal of the Royal Statistical Society Series B 82(4), 1087–1114.
- Fong, E., C. Holmes, and S. G. Walker (2023). Martingale posterior distributions. Journal of the Royal Statistical Society Series B 85(5), 1357—-1391.
- Berti, P., Dreassi, E., Pratelli, L. & Rigo, P. (2021). A class of models for Bayesian predictive inference. Bernouilli, 27, 702–726.
- Rubin, D. B. (1981). The Bayesian bootstrap. The Annals of Statistics, 9(1), 130–134.
- Newton, M. A., Zhang, Y., & Régnier, D. (1998). Nonparametric Bayes methods using predictive updating. In J. M. Bernardo et al. (Eds.), Bayesian Statistics 6 (pp. 651–660). Oxford University Press.
- Newton, M. A., & Zhang, Y. (1999). A recursive algorithm for nonparametric Bayes classification. Biometrika, 86(2), 239–249.
- Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. Sankhyā: The Indian Journal of Statistics, Series A, 64, 306–322.
- Tokdar, S. T., Martin, R., & Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. The Annals of Statistics, 37(5B), 2502–2522.
- Zuanetti, D.A., M¨uller, P., Zhu, Y., Yang, S., Ji, Y.: Bayesian nonparametric clustering for large data sets. Statistics and Computing 29, 203–215 (2019)
- Cappello. Walker (2018). A Bayesian motivated Laplace inversion for multivariate probability distributions. Methodology and Computing in Applied Probability
- Dixit, V., & Martin, R. (2023). Revisiting consistency of a recursive estimator of mixing distributions. *Electronic Journal of Statistics*, *17*(1), 1007-1042.

References (In the order at which they appeared)

- Hahn, P.R., Martin, R., Walker, S.G. (2018 On recursive predictive distributions. Journal of the American Statistical Association 113, 1085–1093
- Cappello, L, Walker, S.G. (2025a). Recursive nonparametric predictive of a regression model.
- Cappello, L, Walker, S.G. (2025b). Recursive estimation of probability distributions
- Bissiri, P. G., & Walker, S. G. (2025). Bayesian analysis with conditionally identically distributed sequences. *Electronic Journal of Statistics*, *19*(1), 1609-1632.
- Tukey, J. W. (1947). Non-parametric estimation II. Comparison of distributions. Annals of Mathematical Statistics, 18(4), 501–507.
  Tukey, J. W. (1961). Curves as parameters, and touch estimation. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.
- Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2012). A distribution-free theory of nonparametric regression.Springer Science & Business Media.
- Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33(3), 1065–1076.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3), 832–837.
- Akaike, H. (1954). An approximation to the density function. Annals of the Institute of Statistical Mathematics, 6, 127–132.
- West, M. (1991). Kernel density estimation and marginalization consistency. Biometrika 78(2), 421–425.

References (In the order at which they appeared)

- Battiston, M., Cappello, L. (2025) New (and old) predictive schemes with a.c.i.d. sequences
- Holmes, C. C. and S. G. Walker (2023). Statistical inference with exchangeability and martingales. Philosophical Transactions of the Royal Society A 381(2247), 20220143.
- Garelli, S., F. Leisen, L. Pratelli, and P. Rigo (2024). Asymptotics of predictive distributions driven by sample means and variances. arXiv preprint arXiv:2403.16828.
- Fong, E. and A. Yiu (2024). Asymptotics for parametric martingale posteriors. arXiv preprint arXiv:2410.17692.
- Fortini, S. and S. Petrone (2025). Exchangeability, prediction and predictive modeling in bayesian statistics. Statistical Science 40(1), 40–67.