Bayesian prediction with streaming data

Sonia Petrone

Bocconi University, Milan

MLSS 2020

first of all: **congratulations** for this summer school from remote, you have done an amazing job!! It's being great!

congratulations and thank you !

Thank you also for having me - a statistician! - taking part in it :-)

Bayesian prediction with streaming data

1. **Bayesian prediction**
   - bit of history & foundations
   - the predictive approach
2. **Streaming data**

   Ex: unsupervised learning and classification via mixture models.

   The Bayesian approach can provide an optimal solution, but computations are involved, especially with streaming data and online learning.

   $\rightarrow$ A fast, **recursive algorithm**. But, **how about uncertainty?**
3. **From an algorithm to a statistical method**

   Taking a predictive approach, i.e., *regarding the algorithm as a probabilistic predictive rule*, we can find that it underlies a *quasi-Bayes* statistical model: thus, we can quantify the uncertainty

Part I: Bayesian prediction

the job of Statistics

Think of COVID-19.....

1. talk with experts, THINK... formulate questions
2. collect info (experts', data, ...)
3. organize info (explore, set algorithms/statistical models, ...)
4. learn ("learn from experience")
5. predict
6. evaluate risk
7. support decisions under incomplete information (uncertainty)
8. also, communicate with the society..

where does ML fit?

where does your work fit?

where does Bayesian Stats fit? – all over! :-)

zooming into learn and predict

- Statistics: learn from experience
  sample $(X_1, Y_1), \ldots, (X_n, Y_n) \sim$ prob model $f_\theta(\cdot)$
  $\rightarrow$ estimate parameters
  $\rightarrow$ predict & provide uncertainty quantification

- Machine learning: predict
  training sample $(x_1, y_1), \ldots, (x_n, y_n)$
  $\rightarrow$ train the algorithm (fix parameters..)
  $\rightarrow$ predict

zooming into learn and predict

- Statistics: learn from experience
  sample $(X_1, Y_1), \ldots, (X_n, Y_n) \sim$ prob model $f_\theta(\cdot)$
  $\rightarrow$ estimate parameters
  $\rightarrow$ predict & provide uncertainty quantification

- Machine learning: predict
  training sample $(x_1, y_1), \ldots, (x_n, y_n)$
  $\rightarrow$ train the algorithm (fix parameters..)
  $\rightarrow$ predict & quantify the uncertainty?

Bayesian approach

Core of the Bayesian approach:

- incomplete information (uncertainty) is formalized through probability
- learning is solved through conditional probabilities

Prediction: given $(x_1, \ldots, x_n)$, predict X_{n+1}:

compute the predictive distribution

$$P(X_{n+1} \in \cdot \mid X_1 = x_1, \ldots, X_n = x_n),$$

The predictive density $p(x_{n+1} \mid x_{1:n})$ gives a full description of the information on X_{n+1} given the data. From it, we can compute a point forecast (ex, w.r.t. quadratic loss)

$$\hat{x}_{n+1} = E(X_{n+1} \mid x_{1:n})$$

and credible intervals

$$P(\hat{x}_{n+1} - q < X_{n+1} < \hat{x}_{n+1} + q \mid x_{1:n}) = 1 - \alpha$$

A flash of history

Often, the Bayesian approach is regarded as 'yet another tool', having optimal properties.

But it comes from a different interpretation on the role of probability and prediction..

The origins go back to Bayes, Bernoulli, Laplace...

'Modern Bayesian statistics' : Ramsey, de Finetti, Savage, ...

Let's think of the historical environment where they lived and worked .. back to 1920'-50'...

Perhaps not a coincidence that they were mathematicians, logicians, probabilists also involved in actuarial science, economics, information, risk and decision..

F.P. Ramsey (1903-1930)

Frank P. Ramsey (Cambridge, 1903, London 1930)

TRUTH AND PROBABILITY

& "Further Considerations" and "Probability and Partial Belief"

by

Frank Plumpton Ramsey, M.A.
Fellow and Director of Studies in Mathematics at King's College,
Lecturer in Mathematics in the University of Cambridge

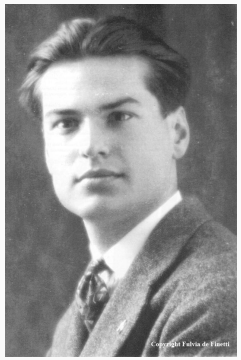
Cambridge of the 1920's, center of knowledge, Camdridge of the economists Piero Sraffa and John Maynard Keynes….

Ramsey starts his "Truth and Probability" arguing on Keynes vision on probability. His interest is in logic as the science of rationa thought…

B. de Finetti (1906 - 1985)

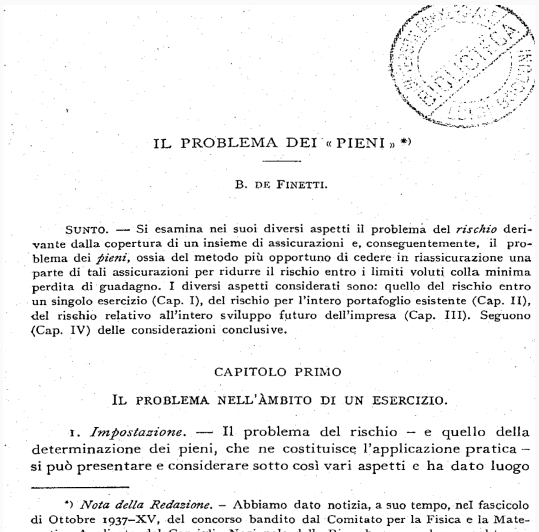
Bruno de Finetti (Innsbruck 13 June 1906 -- Roma 1985)

He worked at the Italian Central Statistical Institute, until 1931, and at the Assicurazioni Generali insurance company in Trieste



He worked on actuarial problems , life insurance mathematics, credibility theory and theory of risk….

His work in 1940 on the mean-variance approach for portoflio selection, largely anticipates Harry Markowitz's(Nobel Prize in 1990).



IL PROBLEMA DEI « PIENI » *)

B. DE FINETTI.

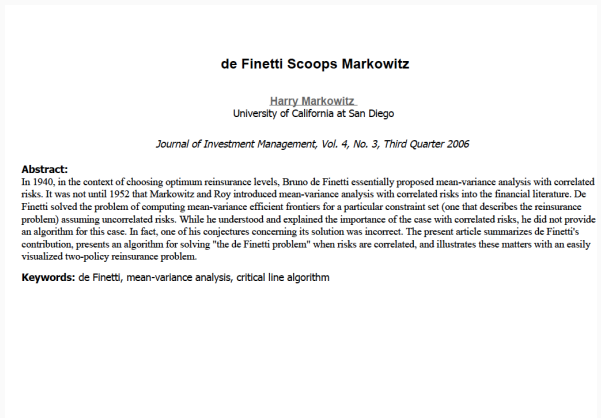
SUNTO. — Si esamina nei suoi diversi aspetti il problema del *rischio* derivante dalla copertura di un insieme di assicurazioni e, conseguentemente, il problema dei *pieni*, ossia del metodo più opportuno di cedere in riassicurazione una parte di tali assicurazioni per ridurre il rischio entro i limiti voluti colla minima perdita di guadagno. I diversi aspetti considerati sono: quello del rischio entro un singolo esercizio (Cap. I), del rischio per l'intero portafoglio esistente (Cap. II), del rischio relativo all'intero sviluppo futuro dell'impresa (Cap. III). Seguono (Cap. IV) delle considerazioni conclusive.

CAPITOLO PRIMO

IL PROBLEMA NELL'AMBITO DI UN ESERCIZIO.

1. *Impostazione.* — Il problema del rischio – e quello della determinazione dei pieni, che ne costituisce l'applicazione pratica – si può presentare e considerare sotto così vari aspetti e ha dato luogo

*) *Nota della Redazione.* – Abbiamo dato notizia, a suo tempo, nel fascicolo di Ottobre 1937–XV, del concorso bandito dal Comitato per la Fisica e la Mate-



de Finetti Scoops Markowitz

Harry Markowitz
University of California at San Diego

Journal of Investment Management, Vol. 4, No. 3, Third Quarter 2006

Abstract:
In 1940, in the context of choosing optimum reinsurance levels, Bruno de Finetti essentially proposed mean-variance analysis with correlated risks. It was not until 1952 that Markowitz and Roy introduced mean-variance analysis with correlated risks into the financial literature. De Finetti solved the problem of computing mean-variance efficient frontiers for a particular constraint set (one that describes the reinsurance problem) assuming uncorrelated risks. While he understood and explained the importance of the case with correlated risks, he did not provide an algorithm for this case. In fact, one of his conjectures concerning its solution was incorrect. The present article summarizes de Finetti's contribution, presents an algorithm for solving "the de Finetti problem" when risks are correlated, and illustrates these matters with an easily visualized two-policy reinsurance problem.

Keywords: de Finetti, mean-variance analysis, critical line algorithm

L.J. Savage (1917 – 1971)

Leonard Jimmie Savage (1917, Detroit – 1971, New Haven, USA

Savage initially worked at the Institute for Advanced Study at Princeton….. and interacted with von Neuman, Milton Friedmar Paul Samuelson..

…..And Abraham Wald…. Decision theory, and Sequential Analysis (1943)

This was just to accompany the change of vision.. from frequencies, to prediction, risk, decisions...

* $\rightarrow$ provide an axiomatization of rational decisions – preferences among actions – and of how they can reveal our beliefs (subjective probability measure) and utility...

* $\rightarrow$ prediction has a central role.

(de Finetti, 1937)

we do not express information (probability) on unobservable events (parameters)

The model [the algorithm?] *is just a ring of the chain from past events to the prediction of future events*

Bayesian prediction

Prediction: given $(x_1, \ldots, x_n)$, predict X_{n+1}

- Classical stats - and ML?
  * random sample: $X_i \overset{iid}{\sim} f_\theta(x)$
  * Then $X_{n+1} \mid x_{1:n} \overset{d}{=} X_{n+1} \sim f_\theta(\cdot)$!
  * One estimates the parameter θ, e.g. by MLE $\hat{\theta}_n$, then plugs it in:

$$X_{n+1} \sim f_{\hat{\theta}_n}(\cdot).$$

- Bayes: we want to learn on future events: thus, the X_i are *probabilistically* dependent!

* $(X_1, X_2, \ldots)$ exchangeable : labels do not matter

For any $n \geq 1$, $p(x_1, \ldots, x_n)$ is invariant to permutations of the labels

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)}).$$

* Then $X_{n+1} \mid x_{1:n} \sim p(x_{n+1} \mid x_{1:n})$ predictive dist.

The predictive distribution gives a full description of the information on X_{n+1} given the data.

but.. how assigning the predictive distribution?

$(X_n)_{n\geq 1}$ exchangeable

- The sequence of empirical distributions $\hat{F}_n$ and the sequence of predictive distr. P_n converge (weakly, a.s.), to the same random dist. F (*directing random measure*).
  In Bayesian stats, F is the model, its probability law is the prior.
- *de Finetti's representation theorem*. (X_n) exchangeable implies that

$$X_i \mid F \overset{iid}{\sim} F.$$

  (the reverse implication is obvious).

Thus, to assign the predictive rule, we can

1 *hypothetical approach*: give a prior law on F; usually, through a parametric model

$$X_i \mid \theta \overset{iid}{\sim} F(\cdot \mid \theta), \quad \theta \sim p(\theta).$$

Then

$$p(x_{n+1} \mid x_{1:n}) = \int f(x_{n+1} \mid \theta) p(\theta \mid x_{1:n}) d\theta.$$

Predictive approach

2 *predictive approach*: Assign the predictive rule: $X_1 \sim P_1$ and fpr $n \geq 1$,

$$X_{n+1} \mid X_{1:n} \sim P_n.$$

My point here will be: P_n could be a predictive algorithm!

Of course, (1) implies (2). But the reverse is also true!

(Ionescu-Tulcea theorem) The sequence of predictive distributions characterizes the probability law P of the process (X_n)

If P is exchangeable, then, by the representation theorem, it characterizes the random directing measure i.e. the model and the prior law.

They are given by the random limit of the predictive distributions P_n.

$\rightarrow$ *This is what we are going to exploit, when we will regard a predictive algorithm as a probabilistic predictive rule P_n, to find the implicit model and prior law!*

Example: Unsupervised sequential learning and classification

example: Unsupervised sequential learning and classification

Recursively classify observations X_n into one of k groups (patterns, signals,..), with no feedback on previous classifications.

A finite mixture model

$$X_n \sim \sum_{j=1}^{k} w_j f_j(x)$$

Here, the components $f_j(\cdot)$ are known.

The Bayesian solution is clear:

* assign a prior law at the unknown weights: $w = (w_1, \ldots, w_k) \sim p(w)$

$*$ learning: posterior distribution of $w \mid x_{1:n}$

$*$ classification: predictive probability that $X_{n+1} \sim f_j \mid x_{1:n}$.

But computations are involved - the more so with streaming data! when prediction and inference have to be updated as a new observation becomes available..

Heterogeneous data: common, classical problems

Mixture models, multiple experiments, individual random effects: many applications, and classical problems in statistics.

Example: Binomial experiments. e.g. clinical trails in different centers, $X_i =$ number of success in M_i trials in center i

$$X_i \mid \theta_i \sim Binomial(M_i, \theta_i)$$

Typical goal: estimate the θ_i, borrowing information across centers/patients..

Hierarchy of nested populations: groups and units within groups

* hospitals, and patients within hospitals
* schools, and students within schools
* genes withing a group of animals,

$\vdots$

Bayesian hierarchical models

Problems of this nature posed the well known surprising results for MLE in the 1950's.

- (Stein, 1956). For estimating the mean vector $(\theta_1, \ldots, \theta_n)$ of a multivariate Gaussian dist, the MLE is not an admissible estimator.
- (Kiefer & Wolfowitz, 1956). For estimating a common parameter in presence of the infinitely many nuisance parameters θ_i, the MLE is inconsistent.
- $\vdots$

The Bayesian approach is quite powerful in modeling heterogeneous data and complex dependence structures!

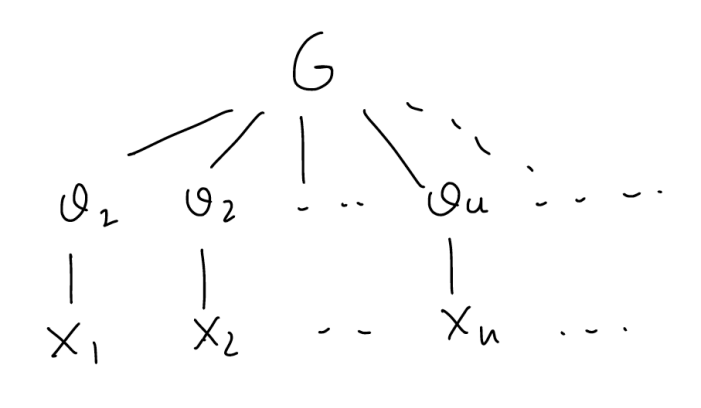
- conditional independence
- hierarchical models

Bayesian hierarchical models

The problems of MLE suggest a hierarchical formulation, where the θ_i are *random variables*, randomly sampled from a latent distribution G.

The hierarchical model is

$$\begin{aligned} X_i \mid \theta_i &\overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma^2) \\ \theta_i \mid G &\overset{iid}{\sim} G \\ G &\sim \pi \qquad \text{prior law} \end{aligned}$$



The $\theta_{1:n-1}$ carry information about the unknown G, thus about θ_n: therefore, they are probabilistically dependent. The θ_i are exchangeable.

Mixture models

Integrating the θ_i out, the hierarchical model

$$\begin{aligned} X_i \mid \theta_i &\overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma^2) \\ \theta_i \mid G &\overset{iid}{\sim} G \\ G &\sim \pi \end{aligned}$$

gives an exchangeable mixture model

$$X_i \mid G \overset{iid}{\sim} f_G(x) = \int f(x \mid \theta) dG(\theta),$$

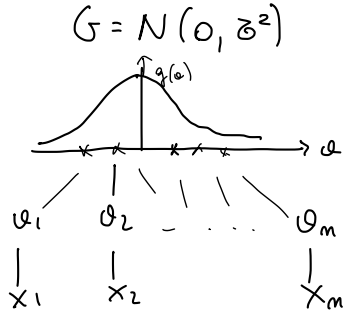
where one assigns a prior law to the latent mixing distribution G.

The prior on G may select discrete or absolutely continuous distributions; which give different types of mixture models.

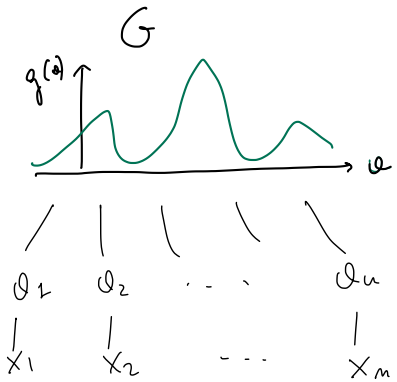
Modeling the latent distribution G

We model the latent distribution through its prior law

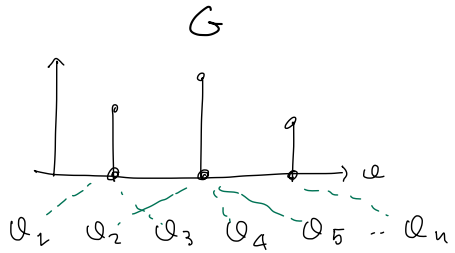
- parametric



- multimodal



- discrete



clusters

Finite mixtures

If G is discrete, with atoms θ_j^* and masses w_j, $j = 1, \ldots, k$, $G = \sum_{j=1}^{k} w_j \delta_{\theta_j^*}$, the model reduces to a finite mixture

$$X_i \mid w, \theta^* \overset{iid}{\sim} \sum_{j=1}^{k} w_j f(x \mid \theta_j^*).$$

Bayesian inference: Assign a prior law on $w = (w_1, \ldots, w_k)$, typically

$$w \sim Dirichlet(\alpha/k, \ldots, \alpha/k), \quad \text{where } E(w_j) = 1/k.$$

$*$ learning: posterior distribution

$$w \mid x_{1:n} \sim Dirichlet(\alpha/k + n_1, \ldots, \alpha/k + n_k)$$

$*$ classification: predictive probability

$$Pr((X_{n+1} \sim f_j) \mid x_{1:n}), \quad j = 1, \ldots, k.$$

Yet, computations are involved..

Dirichlet Process mixtures

How choosing k? Let it be unbounded a priori

$$X_i \mid w, \theta \overset{iid}{\sim} \sum_{j=1}^{\infty} w_j f(x \mid \theta_j^*)$$

where $w = (w_1, w_2, \ldots)$.

What prior on w? The limit of $Dirichlet(\alpha/k, \ldots, \alpha/k)$ does not work... BUT, if we order the w_j: $\mathbf{w}'_k = (w'_1 < w'_2 < \cdots < w'_k)$, then $\mathbf{w}'_k$ converges in distribution to $\mathbf{w}' \sim$ Poisson-Dirichlet(α) (Kingman, 1975).

A countable mixture model with $\mathbf{w}' \sim$ Poisson-Dirichlet(α) and $\theta_j^* \overset{iid}{\sim} G_0$, independently on w', is a Dirichlet process mixture model where

$$G = \sum_{j=1}^{\infty} w'_j \delta_{\theta_j^*} \sim DP(\alpha G_0).$$

DP mixture models are widely used, as the discrete nature of G implies a clustering structure of the $\theta_1, \theta_2, \ldots$ sampled from it. (Chinese Restaurant Process, ...)

Continuous G

G may have a parametric form, e.g. a Gaussian distribution $\mathsf{N}(0, \tau)$; i.e.

$$\theta_i \mid \tau \overset{iid}{\sim} \mathsf{N}(0, \tau).$$

Then the prior distribution on G reduces to the prior on τ: $\tau \sim \pi(\tau)$.

Conditionally on τ,

$$E(\theta_i \mid \tau, x_{1:n}) = \frac{\frac{1}{\sigma^2} x_i + \frac{1}{\tau}\,\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau}}$$

and the Bayesian point estimate is

$$E(\theta_i \mid x_{1:n}) = E\Big(\frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau}} \mid x_{1:n}\Big)\, x_i$$

One has shrinkage towards $\mu = 0$, with the posterior distribution of τ given $x_{1:n}$ providing a data-driven choice of the shrinkage effect.

Nonparametric G

But why a Normal distribution?

Borrowing of information across subjects is quite sensitive to departures from this assumption.

The Normal distribution has light tails and does not allow some subjects to be very different from other subjects or to have groups of subjects that cluster close together. Hence, outlying subjects tend to have their means over-shrunk towards the population mean..

We could use a heavy-tail distribution. But, it would still be unimodal.

$\rightarrow$ Nonparametric: multiple modes of g suggest the type of multiple shrinkage.

Inference on G

Estimating the mixing distribution G, or its density g, in mixture models, is known to be a difficult problem.

Still, Bayesian inference is in principle easy: assign a prior law on G. Then, inference is solved through the posterior distribution of $G \mid X_{1:n}$.

However, computations become involved..

DP mixtures

Given $(\theta_1, \ldots, \theta_n)$, the law of G is still a DP

$$G \mid \theta_{1:n}, x_{1:n} \sim DP(\alpha G_0 + \sum_{1=1}^{n} \delta_{\theta_i}),$$

and

$$E(G \mid \theta_{1:n}, x_{1:n}) = \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{1=1}^{n} \delta_{\theta_i}}{n},$$

a weighted average between the prior guess G_0 and the empirical distribution of the θ_i.

Yet, the θ_i are not observed, and one has to integrate them out, w.r.t. their posterior distribution

$$G \mid x_{1:n} \sim \int DP(\alpha G_0 + \sum_{1=1}^{n} \delta_{\theta_i}) dP(\theta_{1:n} \mid x_{1:n}).$$

inference on G

The Bayesian point estimate of G, w.r.t. quadratic loss, is

$$\begin{aligned} & \mathbf{G_n^{(Bayes)}}(\theta) = E(G(\theta) \mid x_{1:n}) = \\ = \quad & E(\frac{\alpha G_0(\theta) + \sum_{i=1}^{n} \delta_{\theta_i}(\theta)}{\alpha + n}) \mid x_{1:n}) \\ = \quad & (1 - \frac{1}{\alpha + n}) \frac{\alpha G_0(\theta) + \sum_{i=1}^{n-1} P(\theta_i \leq \theta \mid x_{1:n})}{\alpha + n - 1} + \frac{1}{\alpha + n} P_{G_{n-1}^{Bayes}}(\theta_n \leq \theta \mid x_n) \end{aligned}$$

Efficient, in exploiting the info from related experiments for inference on θ_i

but analytic computations become involved...

On learning on the mixing density

In principle, just compute the posterior distribution!

But ..

– Computations are involved. One can resort to MCMC (usually slow) or variational Bayes or ABC.. (yet, just approximations of the Bayesian solution).

– with streaming data, one may use sequential MC, or recent sequential versions of variational Bayes (Lin (2013, Broderick et al (2013), Naesseth et al (2018), ...). Still, slow or approximate or not clearly motivated....

Underlying problem: approximate Bayesian solutions

The Bayesian solution is clear, but computationally demanding, especially with streaming data and online learning, when inference and prediction have to be sequentially updated, as new data become available.

In nowadays trade-off between statistical and computational efficiency, slightly less efficient but more tractable statistical methods are an attractive compromise.

In fact, one may have algorithms; possibly designed as approximations of an optimal, but computationally intractable, Bayesian solution. **Can we go from algorithms to methods?**

A recursive algorithm

1. Unsupervised sequential learning and classification for finite mixtures

Recursively classify observations X_n into one of k groups (patterns, signals,..), with no feedback on previous classifications.

A finite mixture model

$$X_n \sim \sum_{j=1}^{k} w_j f_j(x)$$

The Bayesian solution is clear:

* assign a prior law at the unknown weights: $w = (w_1, \ldots, w_k) \sim p(w)$

$*$ learning: posterior distribution of $w \mid x_{1:n}$

$*$ classification: predictive probability that $X_{n+1} \sim f_j \mid x_{1:n}$.

But computations are involved - the more so with streaming data!

2. Sequential learning on the mixing density

Online estimation of the latent density g in mixture models

$$X_n \mid g \overset{iid}{\sim} \int f(x \mid \theta) g(\theta) d\theta$$

Example: Location mixtures of Gaussian (deconvolution)

$$X_n \mid g \overset{iid}{\sim} \int \mathrm{N}(x \mid \theta, \sigma^2) g(\theta) d\theta$$

Example: Binomial experiments

$$\begin{aligned} X_i \mid \theta_i &\overset{ind}{\sim} Binomial(M, \theta_i) \\ \theta_i \mid g &\overset{iid}{\sim} g(\theta) \end{aligned}$$

θ_i student's ability, and $g(\theta)$, or its d.f. $G(\theta)$, is the latent distribution of ability in the class.

Suppose data arrive sequentially..

A recursive algorithm

M. Newton et al. (Newton and Zhang, 1999; Newton, Quintana & Zhang (1998); Newton (2002). A review: Martin (2018)) proposed a simple recursive algorithm for estimating the latent density g:

*** Start at a prior guess g_0 and for $n \geq 1$ recursively compute**

$$g_n(\theta) = (1 - \alpha_n) g_{n-1}(\theta) + \alpha_n p_{g_{n-1}}(\theta \mid x_n)$$

where

$$p_{g_{n-1}}(\theta \mid x_n) = \frac{f(x_n \mid \theta) g_{n-1}(\theta)}{\int f(x_n \mid \theta) g_{n-1}(\theta) d\theta}$$

is the posterior density of θ_n given x_n, using the previous estimate g_{n-1} as the prior,
and the α_n are weights $\in (0, 1)$. Usually, $\sum_i \alpha_i = \infty$ and $\sum_i \alpha_i^2 < \infty$, a *default choice* being $\alpha_n = \frac{1}{\alpha + n}$.

*** Finally, return $g_N(\cdot)$ as the estimate of g**

Newton's algorithm

The so-called *Newton's algorithm* extends the sequential procedure by Smith & Makov (1978) for recursive learning and classification in finite mixtures.

Interestingly, it is a *stochastic approximation* sequence (Martin & Ghosh (2008))

It is quite attractive: simple and computationally fast.

But, what is the algorithm actually doing?

It was motivated by computations in DP mixtures; yet, examples show that results may disagree with those of DP mixtures.

The sequence (X_n) is no longer exchangeable: order matters.

Also, what about uncertainty around the point estimate provided by the algorithm?

A mysterious algorithm?

People have been trying to give an answer (see Martin & Ghosh, *Stat. Science*, 2008; Martin, 2018).. but these remain open questions.

Results are available on frequentist properties: as an estimator of the latent density g, the recursive g_n is consistent for the true g_{true}, under restrictions on g_{true} (by using properties of stochastic approximation: Smith & Makov (1978), Martin and Ghosh (2008)), and Ghosh and Todkar (2006); Tokdar, Martin & Ghosh (*Ann Stat*, 2009)).

But the original motivation, providing a fast approximation of an optimal but computationally involved Bayesian procedure, gets lost. Given the huge popularity of DP and BNP mixture models, one would like to know **if**, and, in case, **what** Bayesian model one is approximating.

PART II: From an algorithm to a quasi-Bayes method

A recursive algorithm

M. Newton et al. (Newton and Zhang, 1999; Newton, Quintana & Zhang (1998); Newton (2002). A review: Martin (2018)) proposed a simple recursive algorithm for estimating the latent density g:

*** Start at a prior guess g_0 and for $n \geq 1$ recursively compute**

$$g_n(\theta) = (1 - \alpha_n) g_{n-1}(\theta) + \alpha_n p_{g_{n-1}}(\theta \mid x_n)$$

where

$$p_{g_{n-1}}(\theta \mid x_n) = \frac{f(x_n \mid \theta) g_{n-1}(\theta)}{\int f(x_n \mid \theta) g_{n-1}(\theta) d\theta}$$

is the posterior density of θ_n given x_n, using the previous estimate g_{n-1} as the prior,
and the α_n are weights $\in (0, 1)$. Usually, $\sum_i \alpha_i = \infty$ and $\sum_i \alpha_i^2 < \infty$, a *default choice* being $\alpha_n = \frac{1}{\alpha + n}$.

A recursive predictive rule

A new approach

The key of our developments is to frame the recursive algorithm in a probabilistic setting, and to read it as a probabilistic **predictive rule**.

Having a probabilistic framework allows to understand the modeling assumptions implicitly made by the recursive rule, and develop the algorithm into a clear statistical method.

Without such understanding, this all would remain mysterious and extensions and applications would be at most developed as heuristic (although clever) procedures.

Recursive PREDICTION

Let's recall that a mixture model can be rephrased as a hierarchical model

$$\begin{aligned} X_i \mid \theta_i &\overset{ind}{\sim} f(x \mid \theta_i) \\ \theta_i \mid G &\overset{iid}{\sim} G \end{aligned}$$

where G is an unknown latent distribution which, in a Bayesian approach, is given prior law (e.g. a Dirichlet Process, $\mathrm{DP}(\alpha, G_0)$).

Integrating the θ_i out, one gets back the mixture model

$$X_i \mid G \overset{iid}{\sim} f_G(x) = \int f(x \mid \theta) dG(\theta).$$

Recursive PREDICTION

Bayesian inference on the mixing distribution G is solved through the posterior distribution of $G \mid x_{1:n}$.

The Bayesian point estimate w.r.t. quadratic loss is

$$\hat{G}^{Bayes}(\cdot) = E(G(\cdot) \mid x_{1:n}) = P(\theta_{n+1} \in \cdot \mid x_{1:n})$$

and coincides with the **predictive distribution** of θ_{n+1} given $x_{1:n}$

Thus, the problem of recursively estimating G (or its density g) can be rephrased as **recursive prediction**.

Newton algorithm as a PREDICTIVE RULE

For DP mixtures, one has

$$\mathbf{G_n^{(Bayes)}}(\theta) = E(G(\theta) \mid x_{1:n}) = P(\theta_{n+1} \leq \theta \mid x_{1:n})$$

$$= (1 - \frac{1}{\alpha + n}) \frac{\alpha G_0(\theta) + \sum_{i=1}^{n-1} P(\theta_i \leq \theta \mid x_{1:n})}{\alpha + n - 1} + \frac{1}{\alpha + n} P_{G_{n-1}^{Bayes}}(\theta_n \leq \theta \mid x_n)$$

Newton's algorithm gives a different point estimate

$$G_n(\theta) = (1 - \alpha_n) G_{n-1}(\theta) + \alpha_n P_{G_{n-1}}(\theta \mid x_n)$$

Our point is that G_n should be read as a different specification of the predictive rule for $\theta_{n+1} \mid x_{1:n}$. That is,

$$\mathbf{G_n}(\theta) = \mathbf{P}(\theta_{n+1} \leq \theta \mid x_{1:n}), \quad n \geq 1$$

As a probabilistic learning rule, it implies an underlying statistical model. When using the recursive algorithm, a researcher should be aware of such a model, i.e. of the assumption he is implicitly making on the data.

The underlying statistical model

By an *underlying statistical model* we mean a probability law, P say, for the process $((X_n, \theta_n))_{n \geq 1}$, such that the predictive rule for θ_n is given by Newton's recursions.

In BNP mixture model, the law P is such that

$$\begin{aligned} X_i \mid \theta_i &\overset{ind}{\sim} f(x \mid \theta_i) \\ \theta_i \mid G &\overset{iid}{\sim} G \end{aligned}$$

i.e., the θ_i are exchangeable; consequently, the X_i are exchangeable too.

For Newton's recursions, read as predictive distributions, we have

$$X_i \mid \theta_i \overset{ind}{\sim} f(x \mid \theta_i)$$

but the θ_i, and the X_i, are no longer exchangeable: order matters!

However, taking a predictive approach, we find that they are such asymptotically: the implicit model is, asymptotically, an exchangeable mixture model !.

Results: quasi-Bayes properties

We regard Newton's rule as a predictive learning rule:

$$\begin{aligned} X_n \mid \theta_n &\overset{ind}{\sim} f(\cdot \mid \theta_n) \\ \theta_{n+1} \mid X_{1:n} &\sim G_n, \quad n \geq 1, \quad \text{with } \theta_1 \sim G_0. \end{aligned}$$

What model are we actually using?

Results: quasi-Bayes properties

We regard Newton's rule as a predictive learning rule:

$$\begin{aligned} X_n \mid \theta_n &\overset{ind}{\sim} f(\cdot \mid \theta_n) \\ \theta_{n+1} \mid X_{1:n} &\sim G_n, \quad n \geq 1, \quad \text{with } \theta_1 \sim G_0. \end{aligned}$$

What model are we actually using?

The (θ_n), thus the (X_n), are no longer exchangeable. Thus, there is no G such that $\theta_n \mid G \overset{iid}{\sim} G$.

Results: quasi-Bayes properties

We regard Newton's rule as a predictive learning rule:

$$\begin{aligned} X_n \mid \theta_n &\overset{ind}{\sim} f(\cdot \mid \theta_n) \\ \theta_{n+1} \mid X_{1:n} &\sim G_n, \quad n \geq 1, \quad \text{with } \theta_1 \sim G_0. \end{aligned}$$

What model are we actually using?

The (θ_n), thus the (X_n), are no longer exchangeable. Thus, there is no G such that $\theta_n \mid G \overset{iid}{\sim} G$. BUT

Results

The (θ_n), thus the (X_n), are no longer exchangeable. Thus, there is no G such that $\theta_n \mid G \overset{iid}{\sim} G$.
BUT

Result: such G exists asymptotically: G_n converges to a random distribution G, a.s.
If the mixture is identifiable, the sequence (θ_n) is asymptotically exchangeable, with $\theta_n \mid G \overset{iid}{\approx} G$ for large n.

Results

The (θ_n), thus the (X_n), are no longer exchangeable. Thus, there is no G such that $\theta_n \mid G \overset{iid}{\sim} G$.
BUT

Result: such G exists asymptotically: G_n converges to a random distribution G, a.s.
If the mixture is identifiable, the sequence (θ_n) is asymptotically exchangeable, with $\theta_n \mid G \overset{iid}{\approx} G$ for large n.

Result: the (X_n) are c.i.d.. The predictive density converges pointwise to a mixture $f_G(x) = \int k(x \mid \theta) dG(\theta)$, and (X_n) is asymptotically exchangeable. Informally, for n large,

$$X_n \mid G \overset{iid}{\approx} f_G(x) = \int f(x \mid \theta) dG(\theta).$$

In this sense, Newton's rule is a **quasi-Bayes mixture model**, with prior distribution on G induced by the law P of (X_n).

Idea of proofs: predictive constructions

(*how can we go from predictive distributions to models and priors?*)

I'm skipping the proofs here, but I kept this part in the slides

Predictive construction of P

In Bayesian statistics, predictive constructions are powerful tools for characterizing a model and a prior distribution.

In BNP, a well known predictive construction is given by Blackwell & MacQueen's Pólya sequences (or Chinese Restaurant Process):
$X_1 \sim F_0$ and for any $n \geq 1$

$$P_n = \frac{\alpha F_0 + \sum_{i=1}^{n} \delta_{x_i}}{\alpha + n} = (1 - \frac{1}{\alpha + n}) P_{n-1} + \frac{1}{\alpha + n} \delta_{x_n}.$$

This predictive rule characterizes an exchangeable law P for (X_n).

Exchangeability implies that $P_n \Rightarrow \tilde{F}$, P-a.s. and $X_i \mid \tilde{F} \overset{iid}{\sim} \tilde{F}$.

The prior law is implicitly defined as the law of the limit $\tilde{F}$. Here, $\tilde{F} \sim DP(\alpha F_0)$.

What happens for the recursive *predictive rule*?

Interestingly, it defines a novel Polya-urn scheme, and a novel prior!

A measure-valued Polya urn

We find that the predictive rule $P_n(x) \equiv P(X_{n+1} \leq x \mid x_{1:n})$ implied by Newton's recursions is a novel *measure-valued Polya urn scheme* (Bandyopadhyay & Thacker (2016), Mailler and Marckert(2017), Janson, 2019) defined as

$X_1 \sim F_{G_0}(x) = \int f(x \mid \theta) dG_0(\theta)$ and for any $n \geq 1$

$$P_n(x) = \int F(x \mid \theta) dG_n(\theta) = (1 - \alpha_n) P_{n-1}(x) + \alpha_n F_{G_{n-1}}(x \mid x_n)$$

where $F_{G_{n-1}}(x \mid x_n) = \int F(x \mid \theta) dP_{G_{n-1}}(\theta \mid x_n)$.

Comparing with Blackwell & MacQueen's Pólya sequence, it replaces the empirical point masses with absolutely continuous distributions.

This measure-valued Pólya urn predictive rule gives a predictive characterization of the probability law P of the process (X_n). Such P is not exchangeable, but it is c.i.d.

Exchangeable and c.i.d. sequences

Thm(from Kallenberg, 1988) *A predictive rule* (P_n) *characterizes an exchangeable law* P *for* (X_n) *iff*

1. *the sequence* (P_n) *is a measure-valued martingale*
2. (X_n) *is stationary.*

Then, a.s., P_n converges weakly to a random dist. $\tilde{F}$, and $X_i \mid \tilde{F} \overset{iid}{\sim} \tilde{F}$.

Def. A sequence (X_n) is **c.i.d.** if

$$X_{n+k} \mid X_{1:n} \overset{d}{=} X_{n+1} \mid X_{1:n}, \quad \text{for any } n \geq 0 \text{ and } k \geq 1,$$

Easy to see that **c.i.d.** is equivalent to (1).

We prove that Newton's recursive predictive rule satisfies the c.i.d. property (1) but not (2). Still, c.i.d. guarantees that (2), therefore exchangeability, holds asymptotically.

asymptotic exchangeability

Thm (Aldous, 1983) *If P_n converges, to a random probability measure $\tilde{F}$, then (X_n) is asymptotically exchangeable*

$$(X_{n+1}, X_{n+2}, \ldots) \xrightarrow{d} (Z_1, Z_2, \ldots)$$

where (Z_n) is exchangeable with directing random measure $\tilde{F}$.

Roughly speaking, for large n, $X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F}$.

For Newton's model, the c.i.d. property implies $P_n \Rightarrow \tilde{F}$, a.s.
We prove that $\tilde{F}$ is a mixture F_G. Thus, asymptotically,

$$X_i \mid G \overset{iid}{\approx} F_G.$$

In this sense, Newton's model is a **quasi-Bayes** mixture model.

a novel prior on densities

Again, if (X_n) is (X_n) is exchangeable, then $X_n \mid \tilde{F} \overset{iid}{\sim} \tilde{F}$.
If (X_n) is c.i.d., same holds asymptotically.

When is $\tilde{F}$ absolutely continuous, a.s.?

Thm (Berti, Pratelli, Rigo, Ann.Prob. 2013). *If (X_n) is c.i.d., then*

$$\tilde{F} << \lambda \ P\text{-a.s} \quad \text{iff} \quad P_n << \lambda \text{ and } P_n \to \tilde{F} \text{ in TV, } P\text{-a.s.}$$

Newton's recursive predictive rule satisfies the assumptions of the above theorem and characterizes a novel prior on absolutely continuous distributions.

Thm *Under fairly mild conditions, if G_0 is abs continuous then G is abs continuous, P-a.s.*

Remarks

* Here, the c.i.d. model is just a convenient assumption for computational tractability. We have shown that it is an asymptotic exchangeable mixture model.

* Yet, in some problems, stationarity may actually be broken by forms of selection, competition, or by interventions whose effect tends to vanish.

  A class of c.i.d. hierarchical models: Airoldi, Costa, Bassetti, Leisen, Guindani (*JASA*, 2014).

A time-varying mixture model

A time-varying mixture model

In Newton's model, the θ_i are no longer exchangeable. Thus, there is no G such that $\theta_i \mid G \overset{iid}{\sim} G$.

Yet, we can give a hierarchical specification in terms of a latent sequence of distributions $\tilde{G}_n$

$$\begin{aligned} X_n \mid \theta_n &\overset{ind}{\sim} f(x \mid \theta_n) \\ \theta_n \mid \tilde{G}_n &\overset{ind}{\sim} \tilde{G}_n \end{aligned}$$

that is,

$$X_n \mid \tilde{G}_n \overset{ind}{\sim} f_{\tilde{G}_n}(x) = \int f(x \mid \theta) d\tilde{G}_n(\theta).$$

Example: let $X_n \mid \theta_n \sim Bernoulli(\theta_n)$ with θ_n representing student's ability. Here $\tilde{G}_n$ is the latent distribution of ability at time n, and one is assuming an evolution of such ability distributions over time.
Thus, in general, (X_n) is not stationary.

Dynamics of the latent distributions $\tilde{G}_n$

To complete the model

$$X_n \mid \tilde{G}_n \overset{ind}{\sim} f_{\tilde{G}_n}(x) = \int f(x \mid \theta) d\tilde{G}_n(\theta).$$

we have to specify the dynamics of the $\tilde{G}_n$.

* The model assumes an *unpredictable dynamics*, such that for any $n \geq 1$, $E(\tilde{G}_n) = G_0$ and

$$E(\tilde{G}_{n+k} \mid x_{1:n}) = E(\tilde{G}_{n+1} \mid x_{1:n}), \quad k > 1$$

* The conditional law of $\tilde{G}_n$ at time n is a DP centered on the current estimate G_{n-1}

$$\tilde{G}_n \mid X_{1:n-1}, \theta_{1:n-1} \sim DP(\frac{1-\alpha_n}{\alpha_n} G_{n-1})$$

For this model, Newton's one-step-ahead update is exact: it is the Bayesian estimate of $\tilde{G}_n$ from the DP prior.

Thus, the time-varying mixture model is

$$X_n \mid \tilde{G}_n \overset{ind}{\sim} f_{\tilde{G}_n}(x) = \int f(x \mid \theta) d\tilde{G}_n(\theta)$$

and (X_n) is not stationary, thus, not exchangeable.

Yet, $\tilde{G}_n$ converges to G and we have exchangeability, asymptotically.

$$X_n \mid G \overset{iid}{\approx} f_G(x) = \int f(x \mid \theta) dG(\theta) \quad \text{for } n \text{ large.}$$

Informally: the process converges to an exchangeable steady state, with a new prior on G.

Prior on G

Although popularly used as an approximation of the Bayesian solution in Dirichlet Process mixture models, Newton's model implies a different, novel prior on absolutely continous G.

Thm (Berti, Pratelli, Rigo, Ann.Prob. 2013). *If* (X_n) *is c.i.d., then*

$$\tilde{F} << \lambda \ P\text{-a.s} \quad \text{iff} \quad P_n << \lambda \text{ and } P_n \to \tilde{F} \text{ in TV, } P\text{-a.s.}$$

Newton's recursive predictive rule satisfies the assumptions of the above theorem and characterizes a novel prior on absolutely continuous distributions.

Thm *Under fairly mild conditions, if* G_0 *is abs continuous then* G *is abs continuous,* P*-a.s.*

Asymptotic posterior distribution & credible intervals

Asymptotic posterior distribution

Although the prior law is only implicitly defined, we can give results on the induced asymptotic posterior distribution.

Indeed, a Bayesian mixture model would offer more than the point estimate provided by the algorithm: it would describe the uncertainty through the posterior distribution of $G \mid x_{1:n}$.

We can provide an a.s. asymptotic Gaussian approximation of the posterior law of $(G(A_1), \ldots, G(A_k))$, given $x_{1:n}$. For brevity, here $k = 1$.

Results

Thm. Under conditions, which hold if $\alpha_n = 1/(\alpha + n)^\beta$ with $1/2 < \beta \leq 1$ and $\alpha > 0$, then for almost all $\omega = (x_1, x_2, \ldots)$ such that $V_A \neq 0$ we have

$$P(\sqrt{r_n}(G(A) - G_n(A)) \leq t \mid x_{1:n}) \to \Phi(t \mid 0, V_A).$$

where $r_n = (2\beta - 1)n^{2\beta - 1}$, and

$$V_A = \int_{\{x: f_G(x) \neq 0\}} P_G(A \mid x)^2 dF_G(x) - G(A)^2.$$

Remark: $V_A = 0$ if and if $G(A)$ is either zero or one.

Results

In the previous result, the limit variance V_A is unknown, depending on G. Replacing the random V_A with the convergent estimate $V_{n,A}$ is not so direct as using Cramér-Slutzky Theorem in standard i.i.d. settings; but we get

Let

$$V_{n,A} = \int P_{G_n}(A \mid x)^2 dF_{G_n}(x) - G_n(A)^2$$

Lemma: $V_{n,A} \to V_A$ a.s. for $n \to \infty$.

Theorem *Under the same assumptions as before, for almost all* $\omega = (x_1, x_2, \dots)$ *such that* $V_A(\omega) > 0$,

$$P(\sqrt{r_n}\frac{G(A) - G_n(A)}{\sqrt{V_{A,n}}} \leq t \mid x_{1:n}) \to \Phi(t \mid 0, 1).$$

Asymptotic credible intervals for $G(A)$

From the previous results, we have in particular that for $\alpha_n = 1/(\alpha + n)$, then $r_n = n$ and for n large

$$G(A) \mid x_{1:n} \approx \mathbf{N}(G_n(A), \frac{V_{A,n}}{n})$$

We can thus obtain **asymptotic credible intervals** for $G(A)$, given $x_{1:n}$. If $V_A \neq 0$,

$$P(G_n(A) - z_{1-\gamma/2}\sqrt{V_{n,A}/n} < \mathbf{G(A)} < G_n(A) + z_{1-\gamma/2}\sqrt{V_{n,A}/n} | \mathbf{x_{1:n}}) \approx 1 - \gamma$$

Remark These results differ from Berstein-von-Mises types of theorems, which are stated a.s.-$P^{\infty}_{G_{true}}$.

They inform about the rate of convergence of $G_n(A) = G_n(\cdot)(x_{1:n})$ to the limit $G(A)(x_1, x_2, \ldots)$, a.s. P.

Yet they give (novel) hints on frequentist coverage in a simulation study: fast convergence of G_n may mean that the predictive rule does not learn enough from the data..

Tuning the weights α_n

The lack of exchangeability implies that the estimates obtained by Newton's algorithm depend on the ordering of the observations .

* We can give practical insights on how tuning the algorithm to attenuate the effect of the ordering.

** Another popular remedy (Newton et al; Todkar et al (2009); Dixit & Martin (2019) is to use use an average over a number of random permutations of the sample $x_{1:n}$.

This improves symmetry and is still fast but one loses the recursive feature of computations. And, what model is one approximating ?

Interestingly, we can also study the permutation-based algorithm in our predictive approach, as defining a new process and a new prior...

Simulation study

We simulate data from a location mixture of Gaussian kernels

$$X_i \overset{iid}{\sim} \int \mathrm{N}(x \mid \theta, \sigma^2) g_{true}(\theta) d\theta$$

and consider different values of σ^2, different shapes of g, and different choices of the weights. In the plots here, $n = 1000$, and

- $\alpha_n = 1/(\alpha + n)$, with $\alpha = 1$;
- $\alpha_n = 1/(\alpha + n)$, with $\alpha = 100$;
- $\alpha_n = 1/(\alpha + n)^{2/3}$, with $\alpha = 100$;
- split-sample weights: $\alpha_n = 1/(\alpha + n)$ for $n \leq N$, and $\alpha_n = 1/(\alpha + n)^{3/4}$ for $n > N$, with $N = 500$ and $\alpha = 100$.

Bimodal mixing density

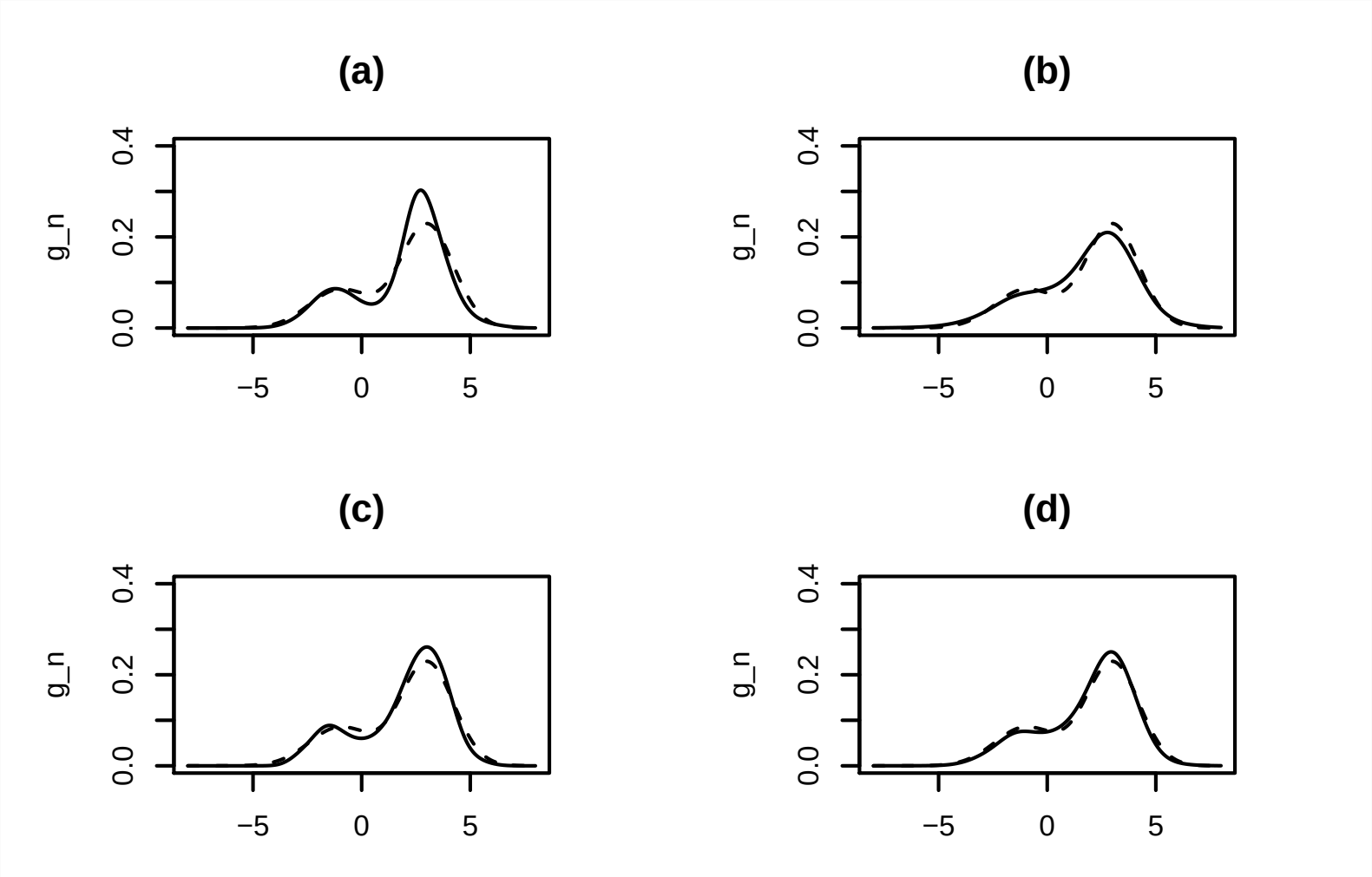


Figure 1: Mixing density estimate g_n. Panel (a): $\alpha_n = 1/(\alpha + n)$, with $\alpha = 1$. Panel (b): $\alpha_n = 1/(\alpha + n)$, $\alpha = 100$. Panel (c): $\alpha_n = 1/(\alpha + n)^{2/3}$, $\alpha = 100$. Panel (d): split-sample weights, $N = 500$, $\gamma = 3/4$; $\alpha = 100$.

Bimodal mixing density

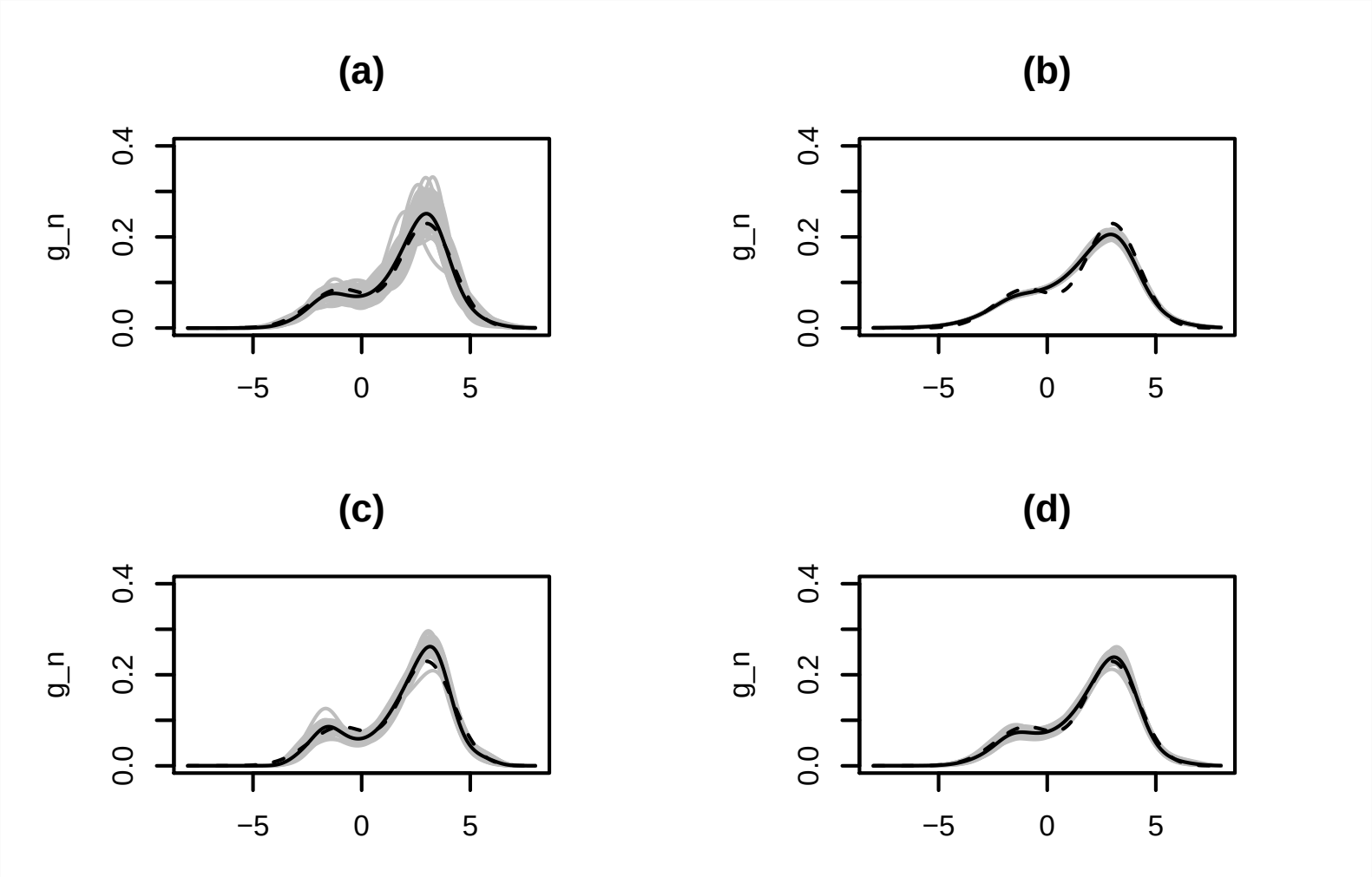


Figure 2: Location mixture of Gaussians; $\sigma^2 = 1$; $n = 1000$. Mixing density estimate g_n (mean over 50 random permutations of $x_{1:n}$, plotted in gray). Dashed: true mixing density.

Asymptotic credible intervals for $G(t)$

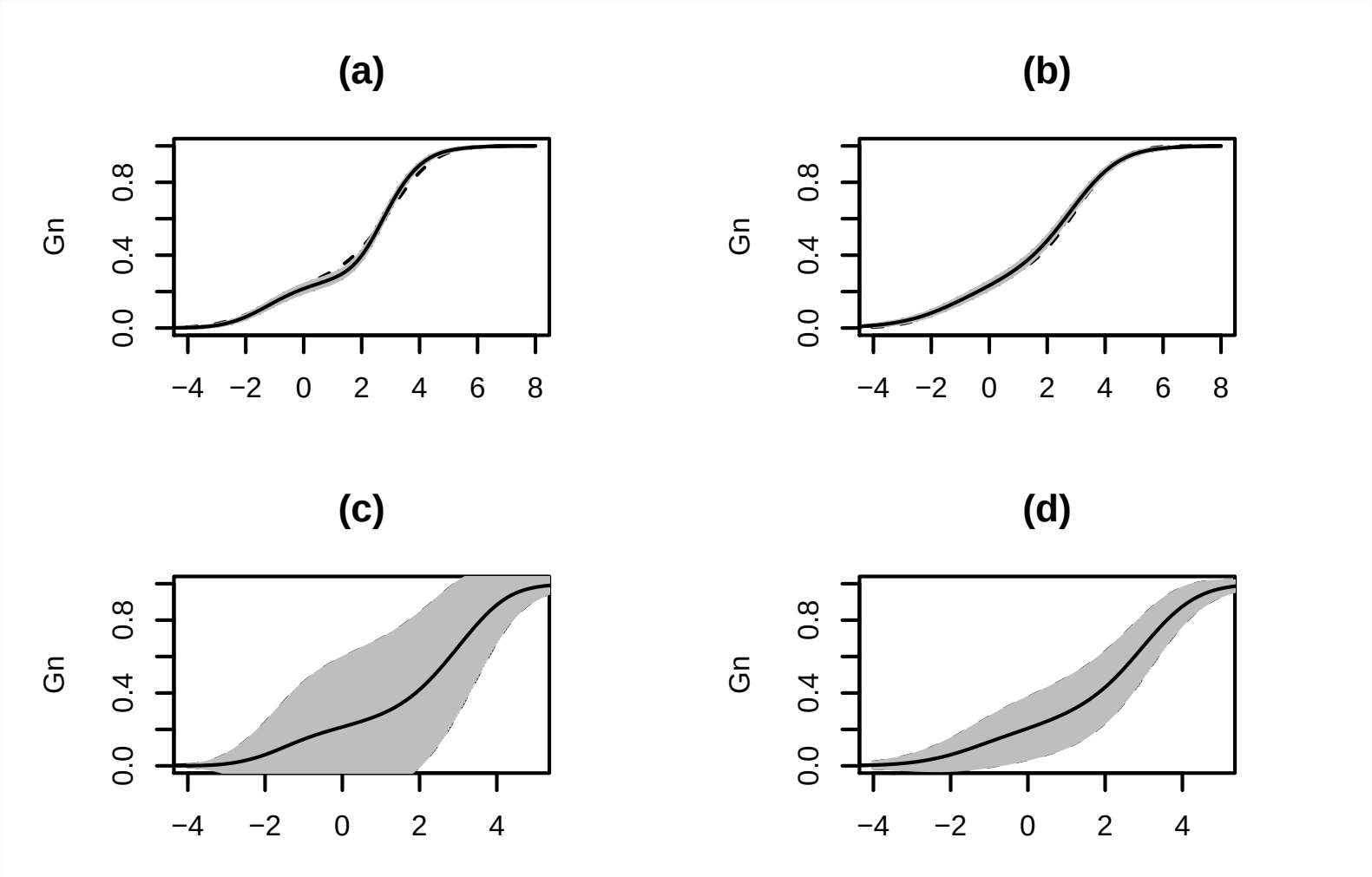


Figure 3: Recursive estimate of the mixing distribution and marginal asymptotic 0.95 credible intervals for $G(t_j)$, for t_j as evidenced in the plot. Dashed: true mixing distribution.

closing

summary (1)

- Prediction is central in Bayesian statistics, and in ML
  We have seen a bit of foundations, and insisted on the *predictive approach*: move from the predictive learning rule, that, in principle, characterizes the model and prior.
- As an example, we looked at (simple) mixture models, aiming at Bayesian learning and classification with streaming data.
  Yet, computations are involved, and most algorithms for approximation are developed for batch data..

summary (2)

- The so-called *M. Newton algorithm* gives a recursive point estimate of the latent distribution G. But the implicit model was unknown, and we didn't have uncertainty quantification.
  Regarding the algorithm as a Bayesian probabilistic predictive rule, we could find the implicit model, and provide the (asymptotic) *posterior distribution* of G, given $x_{1:n}$: the estimate provided by the algorithm is just the posterior expected value $G_n(\cdot) = E(G(\cdot) \mid x_{1:n})$, but we can also provide credible intervals.
- This approach could be extended to other predictive algorithms....

summary (2)

- The so-called *M. Newton algorithm* gives a recursive point estimate of the latent distribution G. But the implicit model was unknown, and we didn't have uncertainty quantification.
  Regarding the algorithm as a Bayesian probabilistic predictive rule, we could find the implicit model, and provide the (asymptotic) *posterior distribution* of G, given $x_{1:n}$: the estimate provided by the algorithm is just the posterior expected value $G_n(\cdot) = E(G(\cdot) \mid x_{1:n})$, but we can also provide credible intervals.
- This approach could be extended to other predictive algorithms....

- Thanks to you all, and to the organizers again: I hope to see you at some future meeting or school! .... (perhaps, at our 'old fashion' school :)

Bocconi Summer School in Advanced Statistics &Probability

2020 edition in collaboration with University of Oxford and Imperial College London

Reproducibility in Data Science, 6-17 July 2020

Home Program Schedule Deadlines Application Registration Contact us Venue and Accomodation Course Material

Photo & Video

Home

NEWS! Because of the COVID-19 outbreak in Italy, the school has been cancelled and will be rescheduled in 2021. New dates will be announced soon!

The Bocconi Summer School in Advanced Statistics and Probability is hosted by the Lake Como School of Advanced Studies at Villa del Grumello, on the shores of the Lake of Como, usually in July. The School continues the tradition of the Summer Schools in Statistics and Probability that Università Bocconi had been organizing since the early '90s, and held in Torgnon, Val d'Aosta, until 2008.

The aim of the Bocconi Summer School in Advanced Statistics and Probability is to establish a track of high level courses on advanced and cutting-edge topics in Statistics and Probability. The Summer School offers lectures delivered by internationally leading scholars on the specific designated topic, and supervised tutorials.

sonia.petrone@unibocconi.it

ciao