Discussion of **Bayesian nonparametric density estimation for data living close to an unknown manifold**, by Judith Rousseau

Sonia Petrone
Bocconi University

Statistical methods and models for complex data.
800 years of research to understand a complex world.
University of Padova, 21-23 Sept 2022

Impressive work!

- Problem interesting, and timely
- Approach - density estimation - statistically neat, and many potential uses
- Work: impressive!! methodological, theoretical, and results of independent interest!
  - work to formalize the problem! (new elegant anisotropic Hölder smoothness,...)
  - new and careful prior law
  - results! rates, also adaptive
  - Highly technical but well explained and illustrated...

really, congratulations!

Impressive work!

- Problem interesting, and timely
- Approach - density estimation - statistically neat, and many potential uses (comment 3)
- Work: impressive!! methodological, theoretical, and results of independent interest!
  - work to formalize the problem! (new elegant anisotropic Hölder smoothness,...)
  - new and careful prior law (comment 1)
  - results! rates, also adaptive (comment 2)
  - Highly technical but well explained and illustrated...

really, congratulations!

comment 1: prior law

$X_i \mid f \overset{iid}{\sim} f$ and we assign a prior law on f.
Gaussian mixtures – scale-location

$$X_i \mid P \sim f_P(x) = \int N(x \mid \mu, \Sigma) dP(\mu, \Sigma)$$

with a prior law on the mixing distribution P.

Comment This is in fact a model (for f_0) + prior law on its parameters (the mixing distribution P).
Thus, care is needed in the choice of the **model**, too.

comment 1: prior law

$X_i \mid f \overset{iid}{\sim} f$ and we assign a prior law on f.
Gaussian mixtures – scale-location

$$X_i \mid P \sim f_P(x) = \int N(x \mid \mu, \Sigma) dP(\mu, \Sigma)$$

with a prior law on the mixing distribution P.

Comment This is in fact a model (for f_0) + prior law on its parameters (the mixing distribution P).
Thus, care is needed in the choice of the **model**, too.

The idea is that a mixture of Gaussian dist. can *approximate* any density: i.e., it is a *flexible model*. But is the approximation *fast*, *efficient*?
(e.g. Bernstein-Dirichlet prior or more general Feller priors (P.+Veronese, 2010): simpler! but efficient approximation?..)

Sect 3.2 in Berenfeld, Rosa and Rousseau: "*The main difficulty in our setup is in proving the Kullback-Leibler prior mass condition. To do so, we need to construct an efficient approximation of f_0 by mixtures of Gaussian.*

How much is due to the model, or to the prior?

prior on the mixing dist P

Very popular in many fields: P discrete, $P \sim DP(aH)$ or a finite-DP (if Dirichlet$(a/k, \ldots, a/k)$, it converges to a DP(aH) for $k \to \infty$).

It is shown that the choice of the base measure H (that generates the atoms) is crucial.
Here, two new clever specifications: *partial or hybrid location-scale mixtures*

prior on the mixing dist P

Very popular in many fields: P discrete, $P \sim DP(aH)$ or a finite-DP (if Dirichlet$(a/k, \ldots, a/k)$, it converges to a DP(aH) for $k \to \infty$).

It is shown that the choice of the base measure H (that generates the atoms) is crucial.
Here, two new clever specifications: *partial or hybrid location-scale mixtures*

Usually, people focus on the scale parameter a (as it controls the induced random partition). But also H is crucial - especially in multivariate settings.
Here, indeed, care is - interestingly - given to H. Why?

Is it because H is in fact "part of the model" for f? or as it merges components, or..?

Comments 2. intrinsic dimension d

Interest in estimating the intrinsic dimension d

Comments 2. intrinsic dimension d

Interest in estimating the intrinsic dimension d

From Antonietta Mira

- Do you need to know d? no, because the rate is adaptive? Or would you estimate it for quantifying the rate?
- In some cases, the data are generated by, say, k distributions, each of which lives on a manifold with different dimensions $d_1, \ldots, d_k$. Did you consider this extension?

Comments 2. intrinsic dimension d

Interest in estimating the intrinsic dimension d

From Antonietta Mira

- Do you need to know d? no, because the rate is adaptive? Or would you estimate it for quantifying the rate?
- In some cases, the data are generated by, say, k distributions, each of which lives on a manifold with different dimensions $d_1, \dots, d_k$. Did you consider this extension?
  or, can your inference on f_0 hint that?

- Allegra, Facco, Denti, Laio, Mira (2020). Data segmentation based on the local intrinsic dimension. *Nature Scientific Reports.*
- Denti, Doimo, Laio, Mira (2022+) Distributional Results for Model-Based Intrinsic Dimension Estimators.

comment 3: inference on f_0, or prediction?

Manifold learning is of interest in many fields and studied with different approaches.
For ex, from Berenfeld, Rosa and Rousseau, about generative adversarial networks:
The theoretical results associated to these approaches control the error between the true generative process and the estimated generative models typically under adversarial losses such as the Wasserstein distance, since the focus is more on generating interesting samples than on estimating the distribution per se.

generating samples.. or **prediction**: $X_{n+1} \mid X_{1:n}$

Different approaches, or related?

More conceptually: A genuine Bayesian statistician would not look at frequentist properties under $P_{f_0}^{\infty}$ i.e. $X_i \overset{iid}{\sim} f_0$. No model is true.. for a genuine Bayesian, it is just a useful tool for **prediction**.
However:

- the Bayesian density estimate $\hat{f}(\cdot) \equiv E(f(\cdot) \mid x_{1:n})$ is the predictive density of $X_{n+1} \mid x_{1:n}$
- In terms of d.f., and real r.v.'s, for n large

$$F(t) \approx N(\hat{f}(t), U(t)/n))$$

where the asymptotic variance $U(t)/n$ depends of the *predictive updates* (Fortini+ P., 2022).

The latter result implies that asymptotic Bayesian credible intervals (on $F(t)$) depend on the "efficiency" of the predictive distribution.

More conceptually: A genuine Bayesian statistician would not look at frequentist properties under $P_{f_0}^{\infty}$ i.e. $X_i \overset{iid}{\sim} f_0$. No model is true.. for a genuine Bayesian, it is just a useful tool for **prediction**.
However:

- the Bayesian density estimate $\hat{f}(\cdot) \equiv E(f(\cdot) \mid x_{1:n})$ is the predictive density of $X_{n+1} \mid x_{1:n}$
- In terms of d.f., and real r.v.'s, for n large

$$F(t) \approx N(\hat{f}(t), U(t)/n))$$

  where the asymptotic variance $U(t)/n$ depends of the *predictive updates* (Fortini+ P., 2022).

The latter result implies that asymptotic Bayesian credible intervals (on $F(t)$) depend on the "efficiency" of the predictive distribution.

If the X_i are indeed generated independently from a true F_0, a "non efficient" predictive distribution might imply asymptotic credible intervals with poor frequentist coverage..

More conceptually: A genuine Bayesian statistician would not look at frequentist properties under $P_{f_0}^{\infty}$ i.e. $X_i \overset{iid}{\sim} f_0$. No model is true.. for a genuine Bayesian, it is just a useful tool for **prediction**.
However:

- the Bayesian density estimate $\hat{f}(\cdot) \equiv E(f(\cdot) \mid x_{1:n})$ is the predictive density of $X_{n+1} \mid x_{1:n}$
- In terms of d.f., and real r.v.'s, for n large

$$F(t) \approx N(\hat{f}(t), U(t)/n))$$

  where the asymptotic variance $U(t)/n$ depends of the *predictive updates* (Fortini+ P., 2022).

The latter result implies that asymptotic Bayesian credible intervals (on $F(t)$) depend on the "efficiency" of the predictive distribution.

If the X_i are indeed generated independently from a true F_0, a "non efficient" predictive distribution might imply asymptotic credible intervals with poor frequentist coverage..

Coverage – and for manifold-anisotropic Hölder densities! - is beyond the scope here. Still, rates are a crucial first step.
Are rates (and adaptive!) related to the efficiency of the predictive rule??

Thank you, Judith, and congratulations again!