Estimation of a density living near a manifold

by Judith Rousseau

Laura Ventura

Department of Statistical Sciences
University of Padova, Italy
ventura@stat.unipd.it

Statistical Methods and Models for Complex Data
Padova, September 21–23, 2022

- Thanks for the opportunity to explore the great work of Judith and cohautors
- An excellent paper, stimulating, rigorous and actual on non-parametric Bayesian density estimation in high-dimensional statistical problems

BUT a necessary premise

- I'm parametric with my central model (only small deviations are allowed)
- I'm in love with small-sample problems (there is a paper by Tang and Reid on modifed likelihood root in high dimensions, but with a likelihood)
- I'm more often frequentist than Bayesian (or better hybrid)

The context

- The common feature of the talks of this conference is that nowadays data reflects the complexity of reality. Data may be:
  - functional
  - hierarchical
  - high-dimensional
  - networks
  - multiple-sourced
  - complex dependencies structures
- From Complexity to Simplicity
  In high-dimensional statistical problems, it is common to consider that the observations actually live on a smaller dimensional structure: the effective dimension of the problem is "simpler" if one can take advantage of the geometry of the data.

The framework

- In this framework, recently, there has been a growing interest in the so-called manifold hypothesis were the data $X_i \in \mathbb{R}^D$ is believed to be (or near) supported on a low dimensional submanifold M of an ambient space.
- When the manifold is known prior to the experiment, nonparametric density estimation dates back to 1985.
- If the submanifold M itself is unknown, getting closer in spirit to a dimension reduction approach, the situation becomes drastically different!! M hence its geometry is unknown, and it is considered as a nuisance parameter.

Small deviations from M

- $X_1, \ldots, X_n$ are a n sample from P_0 in $\mathbb{R}^D$ with density f_0.
- f_0 has support concentrated near a low dimensional manifold M of the ambient space $\mathbb{R}^D$ and the goal is to to study the estimation of the density in the vicinity of M.
- Data is assumed to belong to a neighbourhood

$$M^\delta = \{x \in \mathbb{R}^D : d(x, M) \leq \delta\}$$

  where everything is unknown: M, d, f_0 and also the width $\delta > 0$ of the tube M^δ.
- So Judith plays in an unusual framework where the support of a distribution is unknown while the aim is to recover the density at a point $x \in \mathbb{R}^D$ which is known to be on the support.

The goal

- Posterior concentration rates strengthen the notion of Bayesian consistency, quantifying the speed at which the posterior distribution concentrates on arbitrarily small neighborhoods of the true model, with probability tending to 1 or almost surely, as n goes to infinity.
- Goal: study of the posterior concentration rates, i.e. the smallest possible ε_n such that

$$\Pi(d_H(f_0, f) \leq \varepsilon_n | X_1, \ldots, X_n) \to 1 \quad \text{(in probability)}$$

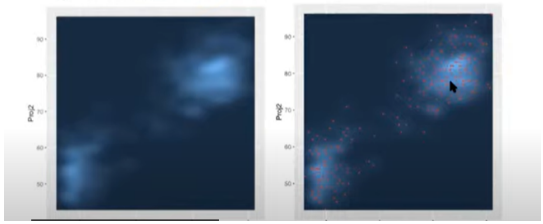
for a class of priors based on specific families of location-scale mixture of Gaussian distributions.

The great contributions

1. General setting for studying density supported near a submanifold: definition of general manifold anisotropic Hölder functions (regularity properties of densities defined on M^{δ}) with smoothness parameter vector β.
2. New family of versatile priors: location-scale of Dirichlet mixture of Gaussian priors.
3. The posterior concentration rate depends on the smoothnesses of the density (β), the dimensions D and d, and the thickness of the support around the manifold (δ).

Some (robust) curiosities

1. It's remarkable that the priors do not depend on β, δ or M ... in particular on d (since conditions on the prior depends on d)! Other possible mixtures, for instance Fisher-Gaussian kernels?
2. When the data naturally lie on M^δ? and – thinking about my deviations from a parametric model – what about the values of δ in practice? what about $\delta \to 0$?
3. May your procedures, with respect to some values of δ, which may give "signals" of wrong assumptions/identifiability problems? may your methods may be used to test these hypotheses? for example, with particularly concentrated densities, which suggest a change in the curvature or an actual extra dimension of the manifold



Two other curiosities

1. With the Hellinger distance, the posterior concentration rates induces also a convergence rate for the posterior mean. Some consequences also for interval estimation?
2. What about the interest in practice of thinking that the data sits near a submanifold or a subdimensional space? and what it may mean in terms of dimension reduction?