BvM for Mixtures

Stefan Franssen — *Delft Institute of Applied Mathematics, TU Delft* — s.e.m.p.franssen@tudelft.nl

Joint work with Jeanne Nguyen and Aad van der Vaart

Mixtures

Suppose we observe a random sample from a mixture density of the form

$$x \mapsto p_{\theta,F}(x) := \int p_\theta(x|\,z)\,dF(z). \tag{1}$$

The kernel $x \mapsto p_\theta(x|\,z)$ is the density of a random variable X given a latent random variable Z, which has marginal distribution F. The kernel is indexed by a parameter $\theta \in \mathbb{R}^d$ that we wish to estimate.

A prior

- We endow q with a prior π;
- We endow F with a Dirichlet process prior.

Earlier literature

Frequentist

- Consistency is understood in some generality [1];
- In a few models MLE is asymptotically efficient [2–4].

Frequentist Bayes

- Semiparametric BvM [5, 6];
- Parametric Bvm [6];

Two consistency theorems

Lemma 1. *Assume the assumptions as in [1] and suppose that (θ_0, F_0) belongs to the Kullback-Leibler support of Π. Then the posterior distribution is consistent at (θ_0, F_0).*

Lemma 2. *Suppose that $\{\log(p_{\theta,F})\}$ is Glivenko-Cantelli, (θ_0, F_0) is identifiable and that (θ_0, F_0) belongs to the Kullback-Leibler support of Π. Then the posterior distribution is consistent at (θ_0, F_0).*

The Bernstein-von Mises theorem for semiparametric mixtures

Theorem 3. *If $\Pi_n(\theta \in \Theta_n, F \in \mathcal{F}_n|\,X_1,\dots,X_n) \to 1$, in P_0^n-probability, and (2)-(5) hold, then*

$$\sqrt{n}(\theta - \hat\theta_n)|X^{(n)} \rightsquigarrow N(0, \tilde I_0^{-1}).$$

Verifying the assumptions

The likelihood condition

For given $(\theta, F) \in \Theta \times \mathcal{F}$, assume that there exists a map $t \mapsto F_t(\theta, F)$ from a given neighbourhood of $0 \in \mathbb{R}^d$ to $\mathcal{F}$ such that, for given measurable subsets $\Theta_n \subset \Theta$ and $\mathcal{F}_n \subset \mathcal{F}$,

$$\begin{aligned}\ell_n\Big(\theta + \frac{t}{\sqrt{n}}, F_{t/\sqrt{n}}(\theta,F)\Big) - \ell_n(\theta,F)\\ = t^T\mathbb{G}_n\tilde\ell_0 - t^T\big(\tilde I_0 + R_{n,1}(\theta,F)\big)\sqrt{n}(\theta-\theta_0)\\ - \frac{1}{2}t^T\tilde I_0 t + R_{n,2}(\theta,F),\end{aligned}$$

for a matrix-valued process $R_{n,1}$ and and scalar process $R_{n,2}$ such that

$$\sup_{\theta\in\Theta_n, F\in\mathcal{F}_n} \big\|R_{n,1}(\theta,F)\big\| + |R_{n,2}(\theta,F)| \xrightarrow{P_0^n} 0. \tag{2}$$

Rewriting condition (2)

Condition (2) can be split in a random and a deterministic part:

$$\mathbb{G}_n\Big[\sqrt{n}\log\frac{p_{\theta+n^{-1/2}t,F_{n^{-1/2}t}(\theta,F)}}{p_{\theta,F}} - t^T\tilde\ell_0\Big] = o_P(1), \tag{3}$$

$$\begin{aligned} nP_0\log\frac{p_{\theta+n^{-1/2}t,F_{n^{-1/2}t}(\theta,F)}}{p_{\theta,F}} =\\ -t^T\big(\tilde I_0 + o_P(1)\big)\sqrt{n}(\theta-\theta_0) - \tfrac{1}{2}t^T\tilde I_0 t + o_P(1).\end{aligned} \tag{4}$$

Verifying (2)

Lemma 4. *Suppose that the map $t \mapsto \ell(t;\theta,F)(x) := \log p_{\theta+n^{-1/2}t,F_{n^{-1/2}t}(\theta,F)}(x)$ is twice continuously differentiable in a neighbourhood of zero, for every $(\theta,F) \in \Theta_n \times \mathcal{F}_n$ and $x \in \mathcal{X}$.*

- *If the classes of functions $\{\dot\ell(t/\sqrt{n};\theta,F) : \|t\| < 1, (\theta,F) \in \Theta_n\times\mathcal{F}_n\}$ are contained in a given P_0-Donsker class and $P_0\|\dot\ell(t_n/\sqrt{n};\theta_n,F_n) - \tilde\ell_0\|^2 \to 0$, for every $\|t_n\| < 1$ and $(\theta_n,F_n) \in \Theta_n\times\mathcal{F}_n$, then (3) is valid.*
- *If $P_0\|\ddot\ell(t_n/\sqrt{n};\theta_n,F_n) - \tilde I_0\| \to 0$, for every $\|t_n\| \le 1$ and $(\theta_n,F_n) \in \Theta_n\times\mathcal{F}_n$, then (4) is satisfied if also*

$$\sup_{\theta\in\Theta_n,F\in\mathcal{F}_n}\frac{\|P_0\dot\ell(0;\theta,F) + \tilde I_0(\theta-\theta_0)\|}{\|\theta-\theta_0\| + n^{-1/2}} \xrightarrow{P_0^n} 0.$$

(Under some Donsker conditions and smoothness assumptions we can verify the likelihood condition)

The change of measure condition

$$\frac{\int_{\Theta_n\times\mathcal{F}_n} e^{\ell_n\left(\theta-t/\sqrt{n},F_{-t/\sqrt{n}}(\theta,F)\right)}\,d\Pi(\theta,F)}{\int_{\Theta_n\times\mathcal{F}_n} e^{\ell_n(\theta,F)}\,d\Pi(\theta,F)} \xrightarrow{P_0^n} 1. \tag{5}$$

Verifying (5) for DP mixtures

Lemma 5. *Let G be a atomless probability measure with a density g. Assume that we model $(\theta,F) \sim \pi \times \mathrm{DP}(MG)$ and let π have a density h. Assume the posterior is consistent of p_{θ_0,F_0} Let T be an open neighbourhood of 0. Denote $\Pi_t = \Pi\circ(\theta_t, F_t)$. Assume that $(\theta_t,F_t) = (\theta+t, F\circ\phi_t^{-1}(\theta))$. In addition, suppose that there exists constants $c_t > 0$ such that for all $\theta\in\Theta_n, F\in\mathcal{F}_n, t\in T$ the following two smoothness conditions hold:*

- *Smoothness in G: $\|\frac{g(\phi^{-1}_{\frac{t}{\sqrt{n}}}(\theta)(z))\phi^{-1'}_{\frac{t}{\sqrt{n}}}(\theta)(z)}{g(z)} - 1\| \le \frac{C_t}{\sqrt{n}}$;*
- *Smoothness in h: $\|\frac{h(\theta+\frac{t}{\sqrt{n}})}{h(\theta)} - 1\| \le \frac{C_t}{\sqrt{n}}$.*

Finally assume that for all $t \in T$:

$$\frac{e^{-C\sqrt{n}\log(n)}}{\Pi_{\frac{t}{\sqrt{n}}}(B_{n,k}((\theta_0,F_0),\epsilon_n))} = o(e^{-2n\epsilon_n^2}). \tag{6}$$

Then the change of measure condition (5) is satisfied.

Proof sketch

- Mixtures can be seen as a hierarchical model with latent variables;
- Each latent variable under the model is distributed according to F;
- Which under the prior is Dirichlet distributed;
- Using the Chinese restaurant representation we can compute prior expectations;

$$\mathbb{E}_F\left[\prod_{i=1}^n p_\theta(X_i|Z_i)\right] = \frac{1}{M^{[n]}}\sum_{S\in\mathcal{S}_n}\prod_{j=1}^{\#S}\left[(\#S_j-1)!\int\prod_{l\in S_j}p_\theta(X_l|z)\,\mathrm{d}G(z)\right].$$

- Note that partitions correspond to having the same atom in F;
- This is rare under the prior, so under KL assumptions we can apply the remaining mass theorem;
- This means we can restrict to sums over partitions which have less than $\sqrt{n}$ distinct latent variables;
- For those terms, we can bound the error in each product by $O(\frac{1}{\sqrt{n}})$;
- And since there are less than $\sqrt{n}$ terms, total error becomes $o(1)$.

Example: Symmetric location mixtures

Suppose we observe a sample from a distribution $f(x-\theta)$ where f is some symmetric density around 0 and θ is the parameter of interest.
Least favourable submodel is known explicitly, $F_t = F$.
Tricky part becomes consistency. Need to find a good prior, symmetrized DP mixtures do not work!

Example: Exponential frailty

We observe a sample from the distribution of $(X,Y)^T$, where given Z the variables X and Y are independent and exponentially distributed with intensities Z and θZ. Hence F is a distirbution on $(0,\infty)$ and $\theta > 0$.
The least favourable submodel is known explicitly [4]:

$$\mathcal{F}_t(\theta)(B) = F(B(1-\frac{t}{2\theta})^{-1})$$

Hence $\phi_t^{-1}(\theta)(z) = z(1-\frac{t}{2\theta})^{-1}$.
Kullback-Leibler balls can be constructed using the stick-breaking construction for the Dirichlet process, bounding prior mass via [6, Lemma G.13], showing approximation via [6, Lemma B.2].

References

1. Kiefer, J. & Wolfowitz, J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. **27,** 887–906. ISSN: 0003-4851 (1956).
2. van der Vaart, A. W. Estimating a Real Parameter in a Class of Semiparametric Models. *The Annals of Statistics* **16,** 1450–1474 (Dec. 1988).
3. Murphy, S. & van der Vaart, A. Likelihood Inference in the Errors-in-Variables Model. *Journal of Multivariate Analysis* **59,** 81–108. ISSN: 0047-259X (Oct. 1996).
4. van der Vaart, A. W. Efficient Maximum Likelihood Estimation in Semiparametric Mixture Models. *The Annals of Statistics* **24,** 862–878 (Apr. 1996).
5. Castillo, I. & Rousseau, J. A Bernsteinvon Mises Theorem for Smooth Functionals in Semiparametric Models. *The Annals of Statistics* **43,** 2353–2383. ISSN: 0090-5364 (2015).
6. Ghosal, S. & van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference* xxiv+646. ISBN: 978-0-521-87826-5 (Cambridge University Press, Cambridge, 2017).
7. Franssen, S. E. M. P., Nguyen, J. & van der Vaart, A. W. *Bernstein-von Mises for Semiparametric Mixtures*