Robust estimation in total variation distance under a shape constraint

Yannick Baraud, Hélène Halconruy, Guillaume Maillard
University of Luxembourg

logo_uni.jpg

Density estimation in Total Variation

Let $(X_1, \dots, X_n)$ be an independent sample with marginals $P_1^\star, \dots, P_n^\star$. Let

$$\overline{P}^\star = \frac{1}{n}\sum_{i=1}^{n} P_i^\star.$$

The aim is to estimate $\overline{P}^\star$ using the sample. Error is quantified using total variation distance

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{B}(\mathbb{R})} |(P-Q)(A)|.$$

A *shape constraint* is assumed for $\overline{P}^\star$ and imposed on the estimator $\hat{P}$, associated to a density model such as

$$\mathcal{M}^{cv}(I) = \{\text{p is convex on } I = \text{supp}(p)\}$$
$$\mathcal{M}^{cv} = \bigcup \{\mathcal{M}^{cv}(I) : I \text{ interval}\}$$
$$\mathcal{M}^{dec}(I) = \{\text{p is non-increasing on } I = \text{supp}(p)\}$$
$$\mathcal{M}^{dec} = \bigcup \{\mathcal{M}^{cv}(I) : I \text{ interval}\}$$
$$\mathcal{M}^{lcv} = \{p = e^{\phi}, \phi : \mathbb{R} \to \mathbb{R} \cup \{-\infty\} \text{ concave}\}.$$

Motivation

Maximum likelihood (ML) is by far the most popular approach in the field. However, ML

- is **undefined** on $\mathcal{M}^{cv}, \mathcal{M}^{dec}$ (support must be known)
- Fails to be robust
  - to model mis-specification
  - or to outliers.

Our estimator overcomes these shortcomings of maximum likelihood.

TV-estimation

For any two pdfs p, q, let

$$t_{p,q} = 1\!\!1_{q>p} - \int_{\{q>p\}} p(x)dx.$$

Given a sample $\mathbf{X} = (X_1, \dots, X_n)$, let

$$T(\mathbf{X}, p, q) = \frac{1}{n}\sum_{i=1}^{n} t_{p,q}(X_i).$$

Given in addition a countable *density model* $\mathcal{M}$, let

$$T_{\mathcal{M}}(\mathbf{X}, p) = \sup_{q \in \mathcal{M}} T(\mathbf{X}, p, q).$$

A TV-estimator associated with the model $\mathcal{M}$ and precision ε is some $\hat{p}_{\mathcal{M}} \in \mathcal{M}$ s.t.

$$T_{\mathcal{M}}(\mathbf{X}, \hat{p}_{\mathcal{M}}) \le \inf_{p \in \mathcal{M}} T_{\mathcal{M}}(\mathbf{X}, p) + \varepsilon.$$

Let also $\hat{P}_{\mathcal{M}} : A \mapsto \int_A \hat{p}_{\mathcal{M}}(x)dx$.

Extremal point of a model

$\overline{p} \in \mathcal{M}$ is *extremal in* $\mathcal{M}$ with degree at most D ($\overline{p} \in \mathcal{O}(D)$) if

- $\{\{q > \overline{p}\} : q \in \mathcal{M}\}$ is VC with dimension $\le D$
- $\{\{q < \overline{p}\} : q \in \mathcal{M}\}$ is VC with dimension $\le D$.

For example,

- In $\mathcal{M}^{dec}$, if p is piecewise constant with D pieces, then $p \in \mathcal{O}(3D+9)$
- In $\mathcal{M}^{cv}$, if p is piecewise affine with D pieces, then $p \in \mathcal{O}(2D+10)$.

Properties of $\hat{p}_{\mathcal{M}}$

Theorem 1. *For any $p \in \mathcal{M}$, let*

$$\mathbb{B}_(n)(P) = \inf_{D \ge 1}\left\{3 \inf_{P \in \mathcal{O}(D)} d_{TV}(\overline{P}^\star, P) + 48\sqrt{\frac{D}{n}}\right\}.$$

For any $\xi > 0,$ with probability greater than $1 - e^{-\xi}$,

$$\mathbb{E}\left[d_{TV}\left(\hat{P}_{\mathcal{M}}, \overline{P}^\star\right)\right] \le \mathbb{B}_n(\overline{P}^\star) + \sqrt{\frac{2(\log 2 + \xi)}{n}} + \frac{\varepsilon}{n}.$$

Hence, $\hat{P}_{\mathcal{M}}$

- converges at parametric rate $\sqrt{\frac{D}{n}}$ if $\overline{P}^\star \in \mathcal{M}$ (*Adaptivity*)
- Is robust
  - To mis-specification: if $\overline{P}^\star \notin \mathcal{M}, \overline{P} \in \mathcal{M}$,
  $$\mathbb{B}_n(\overline{P}^\star) \le \mathbb{B}_n(\overline{P}) + 3d_{TV}(\overline{P}^\star, \overline{P}).$$
  - To outliers: let $P_i^\star = \overline{P}, i \in I$ and $P_i^\star = Q_i, i \in O$, then
  $$d_{TV}\left(\overline{P}^\star, \overline{P}\right) \le \frac{1}{n}\sum_{i \in O} d_{TV}(Q_i, \overline{P}) \le \frac{|O|}{n}.$$
  It follows that
  $$\mathbb{B}_n(\overline{P}^\star) \le \mathbb{B}_n(\overline{P}) + 3\frac{|O|}{n},$$
  i.e the upper bound increases at most by $3\frac{|O|}{n}$.

Convergence rates

Let $\overline{P}^\star$ have pdf $\overline{p}^\star$. Let

$$I = \text{supp}(\overline{p}^\star), L = |I|, V = \sup_I \overline{p}^\star - \inf_I \overline{p}^\star.$$

Assuming $\overline{p}^\star \in \mathcal{M}$, we obtain the following rates.

Model $\mathcal{M}$	$\mathbb{E}\left[d_{TV}\left(\hat{P}_{\mathcal{M}}, \overline{P}^\star\right)\right] \le$	Optimal?
$\mathcal{M} = \mathcal{M}^{dec}$	$c_1\left(\frac{\log(1+VL)}{n}\right)^{1/3} + \frac{c_2}{\sqrt{n}}$	Yes
$\mathcal{M} = \mathcal{M}^{cv}$	$c_1\left(\frac{\log(1+\sqrt{VL})}{n}\right)^{2/5} + \frac{c_2}{\sqrt{n}}$	Best known
$\mathcal{M} = \mathcal{M}^{lcv}$	$\frac{c_1}{n^{2/5}}$	Yes

where c_1, c_2 are numerical constants.

Conclusion

The following table shows how TV-estimation compares to alternatives on the different models.

Property	TV-estimator	MLE	$\rho-$estimator[1]	Piec. Poly [3]
Defined	All	$\mathcal{M}^{dec}(I), \mathcal{M}^{lcv}$	All	$\mathcal{M}^{lcv}$
Robust	All	$\mathcal{M}^{dec}(I)$ no, $\mathcal{M}^{lcv}$?	All	$\mathcal{M}^{lcv}$
Adaptive	All	$\mathcal{M}^{dec}(I)$[2], $\mathcal{M}^{lcv}$[5]	$\mathcal{M}^{dec}, \mathcal{M}^{lcv}$	No
Optimal	All	$\mathcal{M}^{dec}(I)$[2], $\mathcal{M}^{lcv}$ [4]	$\mathcal{M}^{dec}$ (rate)	$\mathcal{M}^{lcv}$

Bibliography

References

[1] Baraud,Y. and Birgé,L. (2016). Rho-estimators for shape-restricted density estimation. *Stochastic process. Appl.*

[2] Birgé,L. (1989). The Grenander estimator: a non-asymptotic approach. *Annals of Statistics.*

[3] Chan, S. O., Diakonikolas, I., Servedio, R. A., and Sun, X. (2014). Efficient density estimation via piecewise polynomial approximation. *Proceedings of the forty-sixth annual ACM symposium on Theory of computing.*

[4] Kim, A. K. H. and Samworth, R. J. (2016). Global rates of convergence in log-concave density estimation. *Annals of Statistics*

[5] Kim, A. K. H., Guntuboyina, A., and Samworth, R. J. (2018). Adaptation in log-concave density estimation. *Annals of Statistics.*