Active Set Algorithms for Estimating Shape-Constrained Density Ratios*

Lutz Dümbgen, Alexandre Mösching and Christof Strähl

Institute of Mathematical Statistics and Actuarial Science, University of Bern

Introduction

Data. Summarized as a distribution

$$\widehat{P} \;:=\; \sum_{i=1}^{n} w_i \delta_{x_i},$$

with

- weights $w_1, w_2, \ldots, w_n > 0$;
- support points $x_1 < x_2 < \cdots < x_n$ in $\mathbb{R}$.

Model. $\widehat{P}$ estimates distribution

$$P(dx) \;=\; e^{\theta(x)}\, M(dx),$$

with

- a given measure M on $\mathbb{R}$;
- an unknown function θ in given family Θ_1.

Goal. Estimate $\theta \in \Theta_1$ via MLE

$$\widehat{\theta} \;\in\; \operatorname*{arg\,max}_{\theta \in \Theta_1} \int \theta \, d\widehat{P}.$$

Modification. Suppose that for a larger function family Θ

$$\Theta_1 \;:=\; \Big\{ \theta \in \Theta : \int e^\theta \, dM = 1 \Big\}$$
$$\theta + c \;\in\; \Theta \quad \text{for all } \theta \in \Theta, c \in \mathbb{R}.$$

Then

$$\widehat{\theta} \;\in\; \operatorname*{arg\,max}_{\theta \in \Theta} L(\theta),$$

with the modified log-likelihood

$$L(\theta) \;:=\; \int \theta \, d\widehat{P} - \int e^\theta \, dM + 1.$$

Estimation problems

Setting 1: Log-concave densities. P has log-concave density on $\mathbb{R}$:

- $M =$ Lebesgue measure;
- $\Theta = \big\{\theta : \mathbb{R} \to [-\infty, \infty) \text{ concave and u.s.c.}\big\}$.

Setting 2A: Tail inflation. P has log-convex density w.r.t. distribution P_o on $\mathbb{R}$:

- $M = P_o$;
- $\Theta = \big\{\theta : \mathbb{R} \to \mathbb{R} \text{ convex}\big\}$.

Setting 2B: Tail inflation (McCullagh and Polson (2012))**.** P has log-convex and isotonic density w.r.t. distribution P_o on $[0, \infty)$:

- $M = P_o$;
- $\Theta = \big\{\theta : [0, \infty) \to \mathbb{R} \text{ convex and isotonic}\big\}$.

Existence and uniqueness of $\widehat{\theta}$

Lemma 1. In Setting 2A, there exists a unique maximizer $\widehat{\theta}$ of $L(\theta)$ over all $\theta \in \Theta$, provided that $\operatorname{supp}(P_o) = \mathbb{R}$. Precisely, either $\widehat{\theta}$ is linear, or there exists $m \in \{1, \ldots, n-1\}$ points

$$\tau_1 < \cdots < \tau_m$$

in $[x_1, x_n] \setminus \{x_1, \ldots, x_n\}$ such that:

- $\widehat{\theta}$ is piecewise linear;
- $\widehat{\theta}$ changes of slope at these τ_j's;
- Sequence of slopes $\big(\widehat{\theta}'(\tau_j+)\big)_{j=1}^m$ is strictly increasing;
- Each (x_i, x_{i+1}) contains at most one τ_j.

Similar results hold in Settings 1 and 2B.

Characterization of $\widehat{\theta}$

Lemma 1 implies that $\widehat{\theta} \in \mathbb{V} \cap \Theta$ with

$$\mathbb{V} \;:=\; \big\{\text{linear splines on } \mathbb{R}\big\}.$$

For a spline $v \in \mathbb{V}$, define

$$D(v) \;:=\; \big\{\tau \in \mathbb{R} : v'(\tau-) \neq v'(\tau+)\big\},$$

the set of deactivated (equality) constraints.
For a finite set $D \subset \mathbb{R}$, further define the linear space

$$\mathbb{V}_D \;:=\; \big\{v \in \mathbb{V} : D(v) \subset D\big\}$$

of dimension $2 + \#D$.

Finally, the directional derivative of the target functional L along v at a point θ is

$$DL(\theta, v) \;:=\; \lim_{t \to 0+} \frac{L(\theta + tv) - L(\theta)}{t}.$$

This yields the following characterization of $\widehat{\theta}$.

Lemma 2 (Global optimality). $\theta \in \mathbb{V} \cap \Theta$ equals $\widehat{\theta}$ if, and only if,

$$DL(\theta, v) \;\leq\; 0 \quad \text{whenever } \theta + tv \in \Theta \text{ for some } t > 0.$$

Example in Setting 2A: Gaussian mixture

Let $X_1, X_2, \ldots, X_n$ be a sample of the Gaussian mixture

$$P \;:=\; (1-\varepsilon)\mathcal{N}(\mu_1, 1) + \varepsilon \mathcal{N}(\mu_2, 1).$$

For $P_o := \mathcal{N}(0,1)$, the log-density ratio

$$\theta(x) \;:=\; \log \frac{dP}{dP_o}(x) \;=\; \log\big((1-\varepsilon)e^{\mu_1(x-\mu_1)} + \varepsilon e^{\mu_2(x-\mu_2)}\big)$$

is a convex function. In the plots below, $n = 500$, $\varepsilon = 0.7$, $\mu_1 = -1.2$ and $\mu_2 = 1.2$.

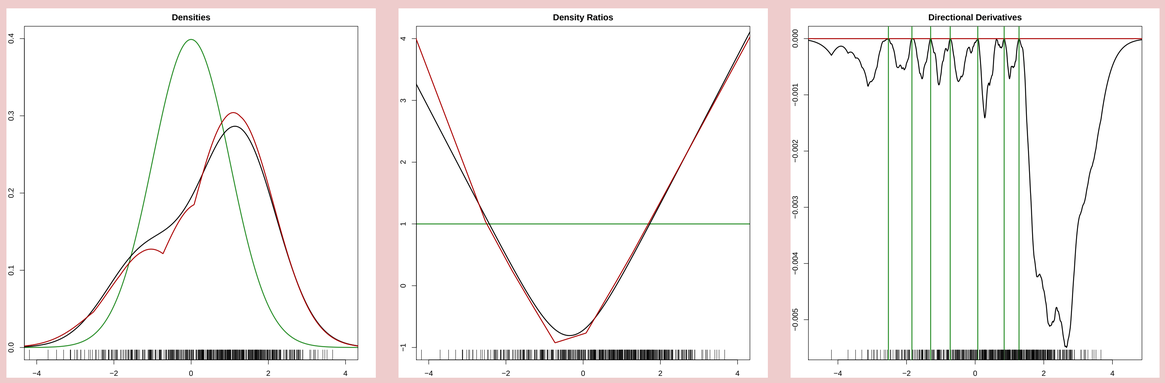

Figure 1: Densities of P (black) and P_o (green) and the estimator (red) w.r.t. Lebesgue measure. Log-density ratios of P (black), P_o (green) and the estimator (red) w.r.t. P_o. The function $\tau \mapsto DL(\widehat{\theta}, V_\tau)$ shows that $\widehat{\theta}$ is globally optimal.

The algorithm (Setting 2A)

Initialize. Start with

$$\theta \;=\; \operatorname*{arg\,max}_{\eta \text{ linear}} L(\eta).$$

Check optimality. Let $\theta \in \mathbb{V} \cap \Theta$ be locally optimal:

$$\theta \;=\; \operatorname*{arg\,max}_{\eta \in \mathbb{V}_{D(\theta)}} L(\eta).$$

Then θ is globally optimal, i.e. $\theta = \widehat{\theta}$ if, and only if,

$$\max_{\tau \in \mathbb{R} \setminus D(\theta)} DL(\theta, V_\tau) \;\leq\; 0,$$

where $V_\tau(x) := (x - \tau)^+$.

Deactivate constraint. If θ is locally but not globally optimal, determine $\tau(\theta) \in \mathbb{R} \setminus D(\theta)$ such that

$$0 \;<\; DL(\theta, V_{\tau(\theta)}) \;\approx\; \max_{\tau \in \mathbb{R} \setminus D(\theta)} DL(\theta, V_\tau).$$

Procedure $\theta_{\text{new}} \leftarrow \text{LocalSearch}(\theta, D, \delta_o)$

```
θ_new ← Newton(θ, V_D)
δ ← DL(θ, θ_new − θ)
while (δ > δ_o) do {
  while (L(θ_new) < L(θ) + δ/3) do {
    θ_new ← (θ + θ_new)/2
    δ ← δ/2 }
  if (θ_new ∉ Θ) do {
    t_o ← max_{t∈(0,1]} {(1 − t)θ + tθ_new ∈ Θ}
    θ_new ← (1 − t_o)θ + t_o θ_new }
  θ ← θ_new
  θ_new ← Newton(θ, V_{D(θ)})
  δ ← DL(θ, θ_new − θ) }
```

Local search (Shape-constrained Newton). Perform Newton steps in $\mathbb{V}_{D(\theta) \cup \tau(\theta)}$ with step size corrections ensuring

$$\theta_{\text{new}} \;\in\; \Theta \qquad \text{and} \qquad L(\theta_{\text{new}}) \;>\; L(\theta).$$

Stop when θ_{new} is ess. locally optimal.

Complete algorithm. Let $\widehat{P}$ be the data and $\delta_o > 0$ a small number.

Procedure $\theta \leftarrow \text{ActiveSetMLE}(\widehat{P}, \delta_o)$

```
θ ← arg max_{η ∈ V_∅ ∩ Θ} L(η)
τ_o ← arg max_{τ ∈ ℝ \ D(θ)} DL(θ, V_τ)
δ ← DL(θ, V_{τ_o})
while (δ > δ_o) do {
  θ ← LocalSearch(θ, D(θ) ∪ τ_o, δ_o)
  τ_o ← arg max_{τ ∈ ℝ \ D(θ)} DL(θ, V_τ)
  δ ← DL(θ, V_{τ_o}) }
```

Remarks.

- After the initialization, θ is locally optimal with all constraints active.
- In the local search procedure, $L(\theta)$ is increasing and $D(\theta)$ is decreasing. Eventually, θ is ess. locally optimal.
- After finitely many iterations, the algorithm will stop at $\theta \approx \widehat{\theta}$.
- Replacing the simple kink functions $V_\tau = (\cdot - \tau)^+$ with "localized versions" leads to better numerical precision.
- The function $\tau \mapsto DL(\theta, V_\tau)$ is strictly concave on each (x_i, x_{i+1}); see Figure 1.

Comparing statistical power

Consider

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P.$$

We want to test

$$H_o: \; P = \mathcal{N}(0,1) \quad \text{vs.} \quad H_A: \; P \neq \mathcal{N}(0,1).$$

Our test statistic:

$$T_n \;:=\; n \cdot \max_{\theta \in \Theta} L(\theta).$$

"Higher criticism": Donoho and Jin (2004) or Gontscharuk et al. (2016) test statistic:

$$B_n \;:=\; \min_{k=1,\ldots,n} \min\Big(B_{k,n+1-k}\big(\Phi(X_{(k)})\big), B_{n+1-k,k}\big(\Phi(-X_{(k)})\big)\Big),$$

with $B_{k,l}$ the cdf of Beta(k,l).

Power comparison for contiguous alternatives

$$P \;=\; (1-\varepsilon_n)\mathcal{N}(0,1) + \varepsilon_n \mathcal{N}(\mu, 1),$$

where $\varepsilon_n := \varepsilon_0 n^{-1/2}$, for a selection of μ and ε_0.

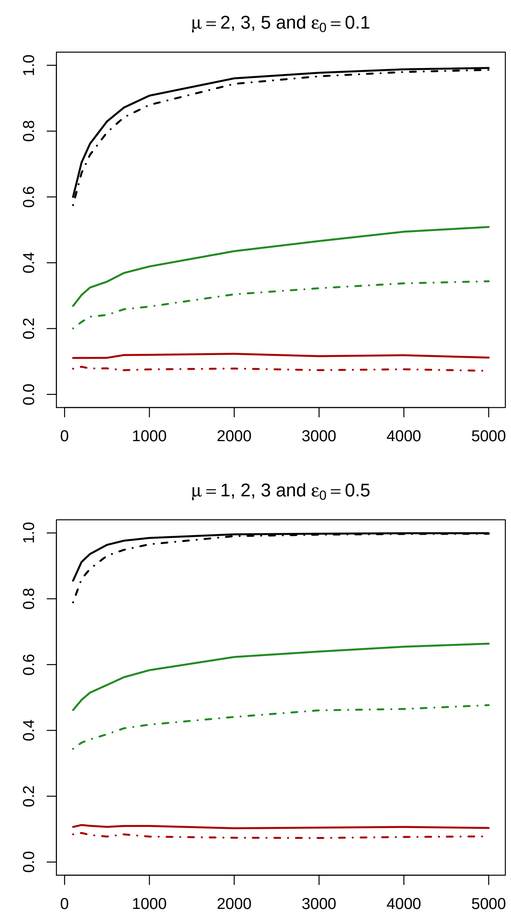

Figure 2: Power comparison of two statistical tests: one based on the MLE of the tail inflation parameter θ (solid) and one as in Gontscharuk et al. (dot-dash). The colors red, green and black correspond to the smallest, medium and largest μ, respectively. The power increases with μ for both tests.

References

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004. URL `arxiv.org/pdf/math/0410072`.

L. Dümbgen, A. Mösching, and C. Strähl. Active set algorithms for estimating shape-constrained density ratios. Preprint, 2018. URL `arxiv.org/abs/1808.09340`.

V. Gontscharuk, S. Landwehr, and H. Finner. Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli*, 22(3):1331–1363, 2016. URL `arxiv.org/abs/1603.05461`.

P. McCullagh and N. G. Polson. Tail inflation. Preprint, 2012.

Acknowledgements

* This work was supported by SNSF.