Nonparametric estimation:

$s-$concave and log-concave densities:

alternatives to maximum likelihood

Jon A. Wellner

University of Washington, Seattle

Statistics Seminar, York

October 15, 2015

Statistics Seminar, York University

Based on joint work with:

- **Qiyang (Roy) Han**
- **Charles Doss**
- Fadoua Balabdaoui
- Kaspar Rufibach
- Arseni Seregin

Outline

- A: Log-concave and $s-$concave densities on $\mathbb{R}$ and $\mathbb{R}^d$
- B: $s-$concave densities on $\mathbb{R}$ and $\mathbb{R}^d$
- C: Maximum Likelihood for log-concave and $s-$concave densities
  - ▷ 1: Basics
  - ▷ 2: On the model
  - ▷ 3: Off the model
- D. An alternative to ML: Rényi divergence estimators
  - ▷ 1. Basics
  - ▷ 2. On the model
  - ▷ 3. Off the model
- E. Summary: problems and open questions

A. Log-concave densities on $\mathbb{R}$ and $\mathbb{R}^d$

If a density f on $\mathbb{R}^d$ is of the form

$$f(x) \equiv f_\varphi(x) = \exp(\varphi(x)) = \exp\left(-(-\varphi(x))\right)$$

where φ is concave (so $-\varphi$ is convex), then f is **log-concave**. The class of all densities f on $\mathbb{R}^d$ of this form is called the class of *log-concave* densities, $\mathcal{P}_{log-concave} \equiv \mathcal{P}_0$.

Properties of log-concave densities:

- Every log-concave density f is unimodal (quasi concave).
- $\mathcal{P}_0$ is closed under convolution.
- $\mathcal{P}_0$ is closed under marginalization.
- $\mathcal{P}_0$ is closed under weak limits.
- A density f on $\mathbb{R}$ is log-concave if and only if its convolution with any unimodal density is again unimodal (Ibragimov, 1956).

- Many parametric families are log-concave, for example:
  - ▷ Normal (μ, σ^2)
  - ▷ Uniform(a, b)
  - ▷ Gamma(r, λ) for $r \geq 1$
  - ▷ Beta(a, b) for $a, b \geq 1$
- t_r densities with $r > 0$ are not log-concave.
- Tails of log-concave densities are necessarily sub-exponential.
- $\mathcal{P}_{log-concave}$ = the class of "Polyá frequency functions of order 2", PF_2, in the terminology of Schoenberg (1951) and Karlin (1968). See Marshall and Olkin (1979), chapter 18, and Dharmadhikari and Joag-Dev (1988), page 150. for nice introductions.

B. $s-$ concave densities on $\mathbb{R}$ and $\mathbb{R}^d$

Let $s < 0$. If a density f on $\mathbb{R}^d$ is of the form

$$f(x) \equiv f_\varphi(x) = \begin{cases} (\varphi(x))^{1/s}, & \varphi \ \ \mathit{convex}, \ \text{if } s < 0 \\ \exp(-\varphi(x)), & \varphi \ \ \mathit{convex}, \ \text{if } s = 0 \\ (\varphi(x))^{1/s}, & \varphi \ \ \mathit{concave}, \ \text{if } s > 0, \end{cases}$$

then f is **s-concave**.

The classes of all densities f on $\mathbb{R}^d$ of these forms are called the classes of $s-$*concave* densities, $\mathcal{P}_s$. The following inclusions hold: if $-\infty < s < 0 < r < \infty$, then

$$\mathcal{P}_r \subset \mathcal{P}_0 \subset \mathcal{P}_s \subset \mathcal{P}_{-\infty}$$

Properties of s-concave densities:

- Every $s-$concave density f is quasi-concave.
- The Student t_ν density, $t_\nu \in \mathcal{P}_s$ for $s \le -1/(1+\nu)$. Thus the Cauchy density ($= t_1$) is in $\mathcal{P}_{-1/2} \subset \mathcal{P}_s$ for $s \le -1/2$.
- The classes $\mathcal{P}_s$ have interesting closure properties under convolution and marginalization which follow from the Borell-Brascamp-Lieb inequality: let $0 < \lambda < 1$, $-1/d \le s \le \infty$, and let $f, g, h : \mathbb{R}^d \to [0, \infty)$ be integrable functions such that

$$h((1-\lambda)x + \lambda y) \ge M_s(f(x), g(x), \lambda) \quad \text{for all} \quad x, y \in \mathbb{R}^d$$

where

$$M_s(a, b, \lambda) = ((1-\lambda)a^s + \lambda b^s)^{1/s}, \qquad M_0(a, b, \lambda) = a^{1-\lambda} b^\lambda.$$

Then

$$\int_{\mathbb{R}^d} h(x)dx \ge M_{s/(sd+1)} \left(\int_{\mathbb{R}^d} f(x)dx, \int_{\mathbb{R}^d} g(x)dx, \lambda \right).$$

C. Maximum Likelihood:

0-concave and s-concave densities

MLE of f and φ: Let $\mathcal{C}$ denote the class of all concave function $\varphi : \mathbb{R} \to [-\infty, \infty)$. The estimator $\widehat{\varphi}_n$ based on $X_1, \ldots, X_n$ i.i.d. as f_0 is the maximizer of the "adjusted criterion function"

$$
\begin{aligned}
\ell_n(\varphi) &= \int \log f_\varphi(x) d\mathbb{F}_n(x) - \int f_\varphi(x) dx \\
&= \begin{cases} \int \varphi(x) d\mathbb{F}_n(x) - \int e^{\varphi(x)} dx, & s = 0, \\ \int (1/s) \log(-\varphi(x))_+ d\mathbb{F}_n(x) - \int (-\varphi(x))_+^{1/s} dx, & s < 0, \end{cases}
\end{aligned}
$$

over $\varphi \in \mathcal{C}$.

1. Basics

- The MLE's for $\mathcal{P}_0$ exist and are unique when $n \geq d + 1$.
- The MLE's for $\mathcal{P}_s$ exist for $s \in (-1/d, 0)$ when

$$n \geq d \left(\frac{r}{r - d} \right)$$

  where $r = -1/s$. Thus $n \to \infty$ as $-1/s = r \searrow d$.

- Uniqueness of MLE's for $\mathcal{P}_s$?
- MLE $\widehat{\varphi}_n$ is piecewise affine for $-1/d < s \leq 0$.
- The MLE for $\mathcal{P}_s$ does not exist if $s < -1/d$. (Well known for $s = -\infty$ and $d = 1$.)

2. On the model

- The MLE's are Hellinger and L_1- consistent.
- The log-concave MLE's $\widehat{f}_{n,0}$ satisfy

$$\int e^{a|x|}|\widehat{f}_{n,0}(x) - f_0(x)|dx \rightarrow_{a.s.} 0.$$

  for $a < a_0$ where $f_0(x) \le \exp(-a_0|x| + b_0)$.

- The $s-$concave MLE's are computationally awkward; log is "too aggressive" a transform for an $s-$concave density. [Note that ML has difficulties even for location $t-$ families: multiple roots of the likelihood equations.]
- Pointwise distribution theory for $\widehat{f}_{n,0}$ when $d = 1$;
  no pointwise distribution theory for $\widehat{f}_{n,s}$ when $d = 1$;
  no pointwise distribution theory for $\widehat{f}_{n,0}$ or $\widehat{f}_{n,s}$ when $d > 1$.
- Global rates? $H(\widehat{f}_{n,s}, f_0) = O_p(n^{-2/5})$ for $-1 < s \le 0$, $d = 1$.

3. Off the model

Now suppose that Q is an arbitrary probability measure on $\mathbb{R}^d$ with density q and $X_1, \ldots, X_n$ are i.i.d. q.

- The MLE $\widehat{f}_n$ for $\mathcal{P}_0$ satisfies:

$$\int_{\mathbb{R}^d} |\widehat{f}_n(x) - f^*(x)| dx \to_{a.s.} 0$$

where, for the Kullback-Leibler divergence

$$K(q, f) = \int q \log(q/f) d\lambda,$$

$$f^* = \text{argmin}_{f \in \mathcal{P}_0(\mathbb{R}^d)} K(q, f)$$

is the "pseudo-true" density in $\mathcal{P}_0(\mathbb{R}^d)$ corresponding to q. In fact:

$$\int_{\mathbb{R}^d} e^{a\|x\|} |\widehat{f}_n(x) - f^*(x)| dx \to_{a.s.} 0$$

for any $a < a_0$ where $f^*(x) \leq \exp(-a_0\|x\| + b_0)$.

- The MLE $\widehat{f}_n$ for $\mathcal{P}_s$ does not behave well off the model. Retracing the basic arguments of Cule and Samworth (2010) leads to negative conclusions. (How negative remains to be pinned down!)

Conclusion: Investigate alternative methods for estimation in the larger classes $\mathcal{P}_s$ with $s < 0$! This leads to the proposals by Koenker and Mizera (2010).

D. An alternative to ML:
Rényi divergence estimators

0. Notation and Definitions

- $\beta = 1 + 1/s < 0$, $\alpha^{-1} + \beta^{-1} = 1$.
- $\mathcal{C}(\underline{X}) =$ all continuous functions on conv$(\underline{X})$.
- $\mathcal{C}^*(\underline{X}) =$ all signed Radon measures on $\mathcal{C}(\underline{X}) =$ dual space of $\mathcal{C}(\underline{X})$.
- $\mathcal{G}(\underline{X}) =$ all closed convex (lower s.c.) functions on conv$(\underline{X})$.
- $\mathcal{G}(\underline{X})^\circ = \{G \in \mathcal{C}^*(\underline{X}) : \int g dG \le 0$ for all $g \in \mathcal{G}(\underline{X}\}$, the polar (or dual) cone of $\mathcal{G}(\underline{X})$.

Primal problems: $\mathcal{P}_0$ and $\mathcal{P}_s$:

- $\mathcal{P}_0$: $\min_{g \in \mathcal{G}(\underline{X})} L_0(g, \mathbb{P}_n)$ where

$$L_0(g, \mathbb{P}_n) = \mathbb{P}_n g + \int_{\mathbb{R}^d} \exp(-g(x))dx.$$

- $\mathcal{P}_s$: $\min_{g \in \mathcal{G}(\underline{X})} L_s(g, \mathbb{P}_n)$ where

$$L_s(g, \mathbb{P}_n) = \mathbb{P}_n g + \frac{1}{|\beta|} \int_{\mathbb{R}^d} g(x)^{\beta} dx.$$

Dual problems: $\mathcal{P}_0$ and $\mathcal{P}_s$:

- $\mathcal{D}_0$: $\max_f \{-\int f(y)\log f(y)dy\}$ subject to

$$f(y) = \frac{d(\mathbb{P}_n - G)}{dy} \quad \text{for some} \quad G \in \mathcal{G}(\underline{X})^\circ.$$

- $\mathcal{D}_s$: $\max_f \int \frac{f(y)^\alpha}{\alpha} dy$ subject to

$$f(y) = \frac{d(\mathbb{P}_n - G)}{dy} \quad \text{for some} \quad G \in \mathcal{G}(\underline{X})^\circ.$$

Why do these make sense?

- Population version of $\mathcal{P}_0$: $\min_{g\in\mathcal{G}} L_0(g, f_0)$ where

$$L_0(g, f_0) = \int \{g(x) f_0(x) + e^{-g(x)}\} dx.$$

  Minimizing the integrand pointwise in $g = g(x)$ for fixed $f_0(x)$ yields $f_0(x) - e^{-g} = 0$ if $e^{-g} = e^{-g(x)} = f_0(x)$.

- Population version of $\mathcal{P}_s$: $\min_{g\in\mathcal{G}} L_s(g, f_0)$ where

$$L_s(g, f_0) = \int \{g(x) f_0(x) + \frac{1}{|\beta|} g^{\beta}(x)\} dx.$$

  Minimizing the integrand pointwise in $g = g(x)$ for fixed $f_0(x)$ yields $f_0(x) + (\beta/|\beta|) g^{\beta-1} = f_0(x) - g^{\beta-1} = 0$, and hence $g^{1/s} = g^{1/s}(x) = f_0(x)$.

1. Basics for the Rényi divergence estimators:

- (Koenker and Mizera, 2010) If conv($\underline{X}$) has non-empty interior, then strong duality between $\mathcal{P}_s$ and $\mathcal{D}_s$ holds. The dual optimal optimal solution exists, is unique, and $\widehat{f}_n = \widehat{g}_n^{1/s}$.
- (Koenker and Mizera, 2010) The solution $f = g^{1/s}$ in the population version of the problem when $Q = P_0$ has density $p_0 \in \mathcal{P}_s$ is Fisher-consistent; i.e. $f = p_0$.

2. Off the model: Han & W (2015)

Let

$$\mathcal{Q}_1 \equiv \{Q \text{ on } (\mathbb{R}^d, \mathcal{B}^d): \int \|x\| dQ(x) < \infty\},$$
$$\mathcal{Q}_0 \equiv \{Q \text{ on } (\mathbb{R}^d, \mathcal{B}^d): \text{int}(\text{csupp}(Q)) \neq \emptyset\}.$$

- Theorem (Han & W, 2015): If $-1/(d+1) < s < 0$ and $Q \in \mathcal{Q}_0 \cap \mathcal{Q}_1$, then the primal problem $\mathcal{P}_s(Q)$ has a unique solution $\tilde{g} \in \mathcal{G}$ which satisfies $\tilde{f} = \tilde{g}^{1/s}$ where $\tilde{g}$ is bounded away from 0 and $\tilde{f}$ is a bounded density.

- Theorem (Han & W, 2015): Let $d = 1$. If $\widehat{f}_{n,s}$ denotes the solution to the primal problem $\mathcal{P}_s$ and $\widehat{f}_{n,0}$ denotes the solution to the primal problem $\mathcal{P}_0$, then for any $\kappa > 0$, $p \geq 1$,

$$\int (1 + |x|)^\kappa |\widehat{f}_{n,s}(x) - \widehat{f}_{n,0}(x)|^p dx \to 0 \quad \text{as} \quad s \nearrow 0.$$

- Theorem (Han & W, 2015): Suppose that:
  (i) $d \geq 1$,
  (ii) $-1/(d+1) < s < 0$, and
  (iii) $Q \in \mathcal{Q}_0 \cap \mathcal{Q}_1$.
  If $f_{Q,s}$ denotes the (pseudo-true) solution to the primal problem $\mathcal{P}_s(Q)$, then for any $\kappa < r - d = (-1/s) - d$,

$$\int (1 + |x|)^{\kappa} |\widehat{f}_{n,s}(x) - f_{Q,s}(x)| dx \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty.$$

3. On the model: Q has density $f \in \mathcal{P}_s$; $f = g^{1/s}$ for some g convex.

- Consistency: Suppose that: (i) $d \geq 1$ and $-1/(d+1) < s < 0$. Then for any $\kappa < r - d = (-1/s) - d$,

$$\int (1 + |x|)^\kappa |\widehat{f}_{n,s}(x) - f(x)| dx \to_{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

  Thus $H(\widehat{f}_{n,s}, f) \to_{a.s.} 0$ as well.

- Pointwise limit theory: (paralleling the results of Balabdaoui, Rufibach, and W (2009) for $s = 0$)

  Assumptions:

  ▷ (A1) $g_0 \in \mathcal{G}$ and $f_0 \in \mathcal{P}_s(\mathbb{R})$ with $-1/2 < s < 0$.

  ▷ (A2) $f_0(x_0) > 0$.

  ▷ (A3) g_0 is locally C^2 in a neighborhood of x_0 with $g_0''(x_0) > 0$.

Theorem 1. (Pointwise limit theorem; Han & W (2015))
Under assumptions (A1)-(A3), we have

$$\begin{pmatrix} n^{\frac{2}{5}}\Big(\widehat{g}_n(x_0) - g_0(x_0)\Big) \\ n^{\frac{1}{5}}\Big(\widehat{g}_n'(x_0) - g_0'(x_0)\Big) \end{pmatrix} \to_d \begin{pmatrix} -\left(\frac{g_0^4(x_0) g_0^{(2)}(x_0)}{r^4 f_0(x_0)^2 (4)!}\right)^{1/5} H_2^{(2)}(0) \\ -\left(\frac{g_0^2(x_0)\Big[g_0^{(2)}(x_0)\Big]^3}{r^2 f_0(x_0)^3 \Big[(4)!\Big]^3}\right)^{1/5} H_2^{(3)}(0) \end{pmatrix},$$

and . . .

... furthermore

$$\begin{pmatrix} n^{\frac{2}{5}}\Big(\widehat{f}_n(x_0) - f_0(x_0)\Big) \\ n^{\frac{1}{5}}\Big(\widehat{f}'_n(x_0) - f'_0(x_0)\Big) \end{pmatrix} \to_d \left(\begin{array}{c} \left(\frac{r f_0(x_0)^3 g_0^{(2)}(x_0)}{g_0(x_0)(4)!}\right)^{1/5} H_2^{(2)}(0) \\ \left(\frac{r^3 f_0(x_0)^4 \Big(g_0^{(2)}(x_0)\Big)^3}{g_0(x_0)^3 \Big[(4)!\Big]^3}\right)^{1/5} H_2^{(3)}(0) \end{array} \right),$$

where H_2 is the unique lower envelope of the process Y_2 satisfying

1. $H_2(t) \leq Y_2(t)$ for all $t \in \mathbb{R}$;
2. $H_k^{(2)}$ is concave;
3. $H_2(t) = Y_2(t)$ if the slope of $H_2^{(2)}$ decreases strictly at t.
4. $Y_2(t) = \int_0^t W(s)ds - t^4$, $t \in \mathbb{R}$ where W is two-sided Brownian motion started at 0.

- Estimation of the mode for $d = 1$.

  Theorem 2. (Estimation of the mode) Assume (A1)-(A4) hold. Then

$$n^{1/5}\left(\widehat{m}_n - m_0\right) \to_d \left(\frac{g_0(m_0)^2 (4)!^2}{r^2 f_0(m_0) g_0^{(2)}(m_0)^2}\right)^{1/5} M(H_2^{(2)}), \quad (1)$$

  where $\widehat{m}_n = M(\widehat{f}_n), m_0 = M(f_0)$.

- What is the price of assuming $s < 0$ when the truth $f \in \mathcal{P}_0$?

Assume $-1/2 < s < 0$ and $k = 2$. Let $f_0 = \exp(\varphi_0)$ be a log-concave density where $\varphi_0 : \mathbb{R} \to \mathbb{R}$ is the underlying concave function. Then f_0 is also s-concave.

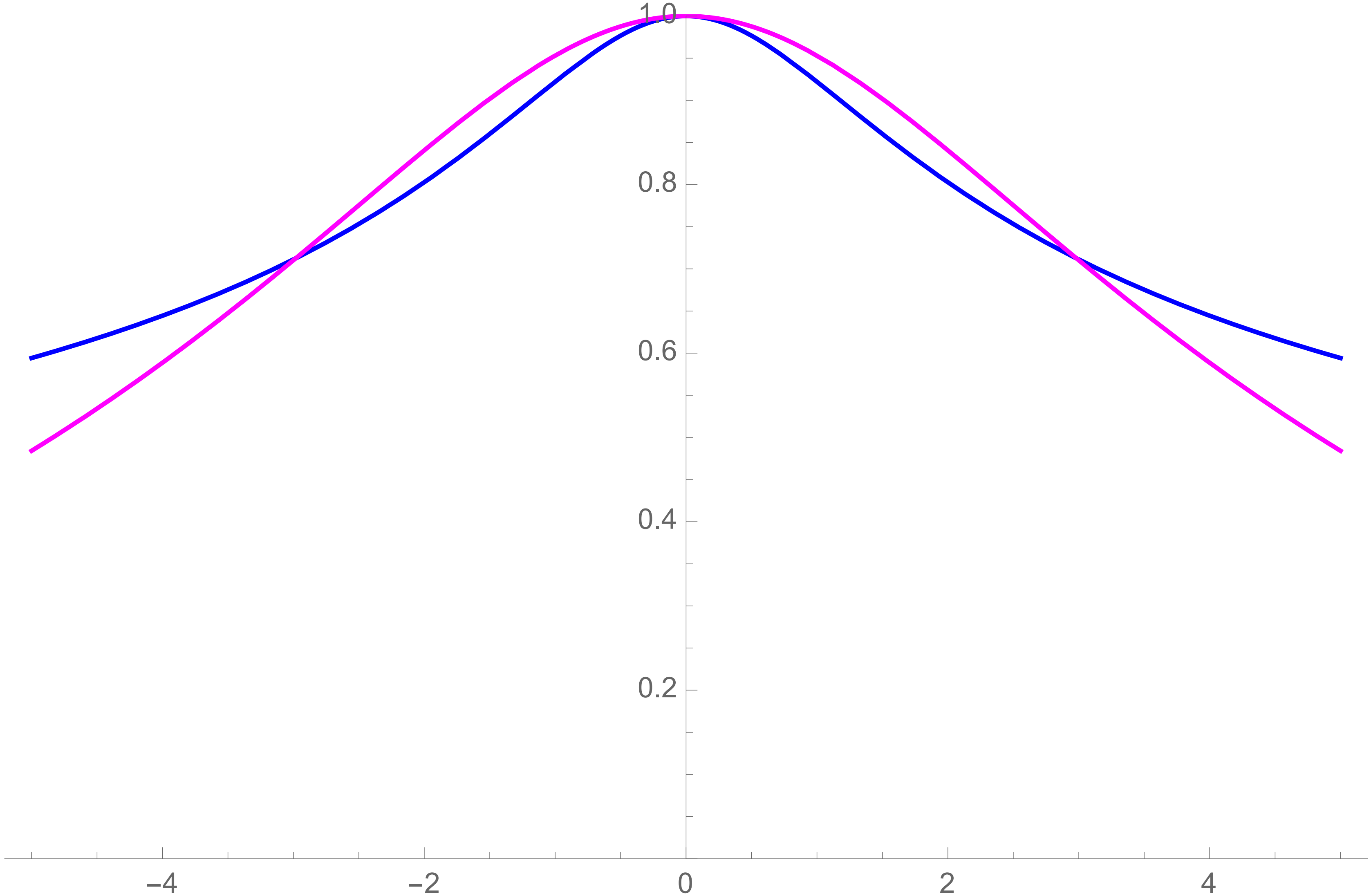
Let $g_s := f_0^{-1/r} = \exp(-\varphi_0/r)$ be the underlying convex function when f_0 is viewed as an s-concave density. Calculation yields

$$g_s^{(2)}(x_0) = \frac{1}{r^2} g_s(x_0) \left(\varphi_0'(x_0)^2 - r\varphi_0''(x_0) \right).$$

Hence the constant before $H_2^{(2)}(0)$ appearing in the limit distribution for $\widehat{f}_n$ becomes

$$\left(\frac{f_0(x_0)^3 \varphi_0'(x_0)^2}{4!r} + \frac{f_0(x_0)^3 |\varphi_0''(x_0)|}{4!} \right)^{1/5}.$$

The second term is the constant involved in the limiting distribution when $f_0(x_0)$ is estimated via the log-concave MLE: (2.2), page 1305 in Balabdaoui, Rufibach, & W (2009). The ratio of the two constants (or asymptotic relative efficiency) is shown for f_0 standard normal (blue) and logistic (magenta) in the figure:


1.0
0.8
0.6
0.4
0.2
−4
−2
0
2
4

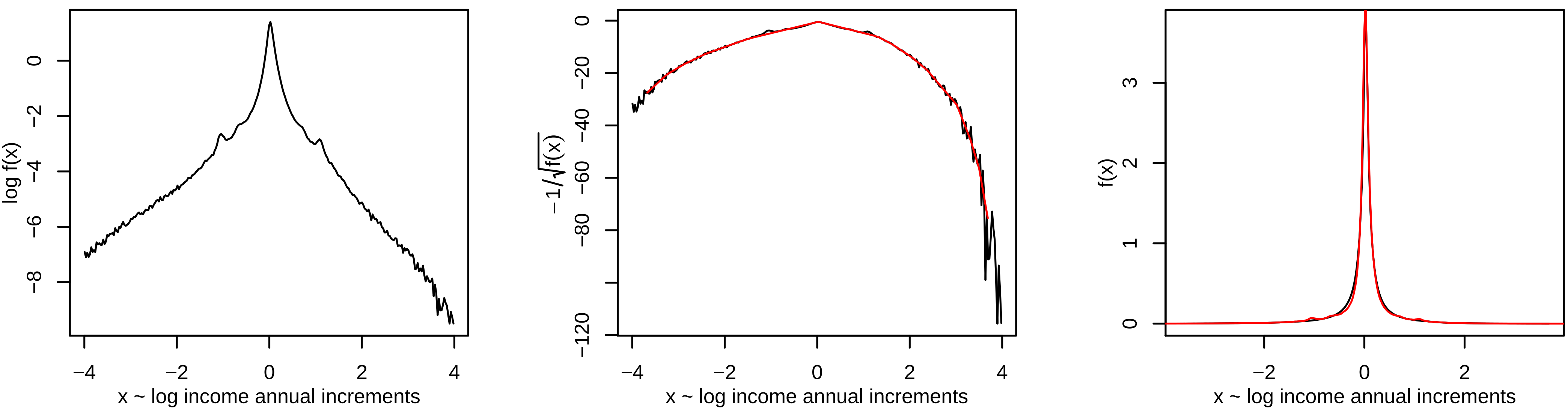
- The first term is non-negative and is the price we pay by estimating a true log-concave density via the Rényi divergence estimator over a larger class of s-concave densities.
- Note that the first term vanishes as $r \to \infty$ (or $s \nearrow 0$).
- Note that the ratio is 1 at the mode of f_0.
- For estimation of the mode, the ratio of constants is always 1: nothing is lost by enlarging the class from $s = 0$ to $s < 0$!

E. Summary: problems and open questions

- Global rates of convergence?
- Limiting distribution(s) for $d > 1$? (n^r with $r = 2/(4+d)$?)
- MLE (rate-) inefficient for $d \geq 4$ (or perhaps $d \geq 3$)? How to penalize to get efficient rates?
- Can we go below $s = -1/(d+1)$ with other methods?
- Multivariate classes with nice preservation/closure properties and smoother than log-concave?
- Algorithms for computing $\widehat{f}_n \in \mathcal{P}_s$?
- Related results for convex regression on $\mathbb{R}^d$: Seijo and Sen, *Ann. Statist.* (2011).

Guvenen et al (2014)

have estimated models of income dynamics using very large (10 percent) samples of U.S. Social Security records linked to W2 data The density is not log-concave, but $s-$concave density with $s = -1/2$ fits well:



Courtesy Roger Koenker

F. Selected references

- Dümbgen and Rufibach (2009).
- Cule, Samworth, and Stewart (2010)
- Cule and Samworth (2010).
- Dümbgen, Samworth, and Schuhmacher (2011).
- Balabdaoui, Rufibach, and W (2009)
- Seregin & W (2010), *Ann. Statist.*
- Koenker and Mizera (2010), *Ann. Statist.*
- Han & W (2015): arXiv:1505.00379v3.
- Doss & W (2013-15): arXiv:1306.1438v2.
- Guntuboyina and Sen (2015), *Prob. Theor. Rel. Fields*

Many thanks!

Skiing toward the Nisqually Glacier