Shape-restricted Univariate Smooth Density Estimation

Hanzhang Qin [1]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

Dec 12, 2016

How to estimate a univariate density?

- Histogram.
- Kernel density estimation (KDE).
- Maximum likelihood estimation (MLE) from a parametric family.
- Mixture model estimation: the expectation-maximization (EM) algorithm.
- ...

We want a general approach.

- We might not have prior information about the parametric family of data (MLE).
- We might not know how to choose the bandwidth optimally in a histogram or KDE (they are heuristics).
- We might not know how good is the EM algorithm (it converges to a local minimum).
- Sometimes we want some control on the shape of our estimation function (e.g., data is corrupted).

Overview of our approach: MIQP & first order methods.

We develop: (MIQP: mixed-integer quadratic programming)

(1) Mixed-integer formulations for monotonicity/convexity/concavity/k-modality/log-concavity shape restriction;

(2) First-order optimal algorithms (FO-k) for k-modal density estimation, based on mirror descent method and prefix isotonic regression;

We suggest an algorithmic framework as follows:

(1) Monotonicity and convexity/concavity: QP formulation.

(2) k-modality: if $k = 0$ or 1, use FO-k; if $k \geq 2$, use MIQP with FO-0 warmstart or simply FO-0 if samples are of good quality.

(3) Log-concavity: MIQP (taking binary expansions) with FO-1 warmstart or simply FO-1 if samples are of good quality.

All codes written in Julia. MIQP implemented in JuMP with Gurobi.

Example: convex & nonincreasing estimation

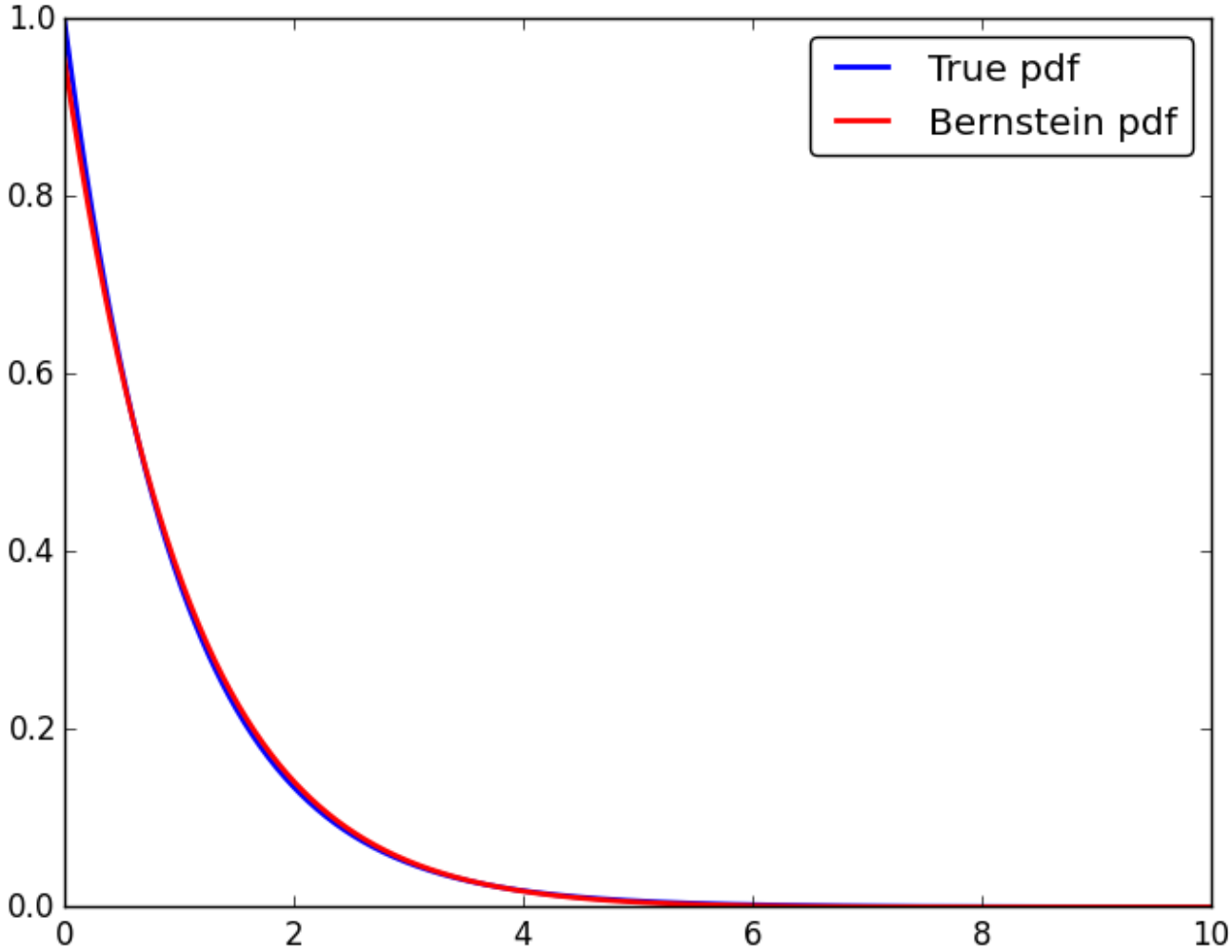

Figure: 100 samples $\sim$ Exponential(1); Time: 0.0089s

Example: unimodal estimation

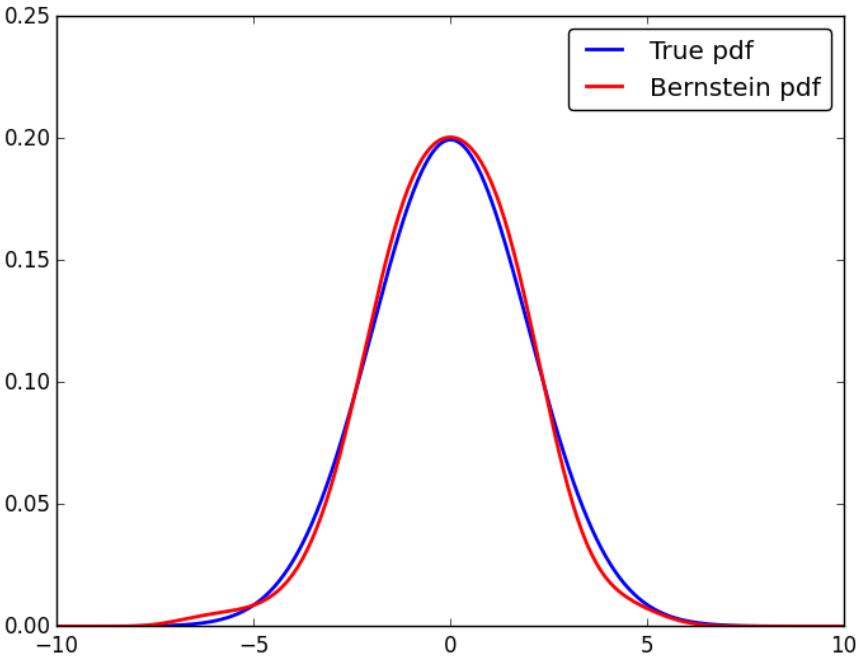

Figure: 100 samples $\sim$ Normal(0,2); Time: 0.4205s

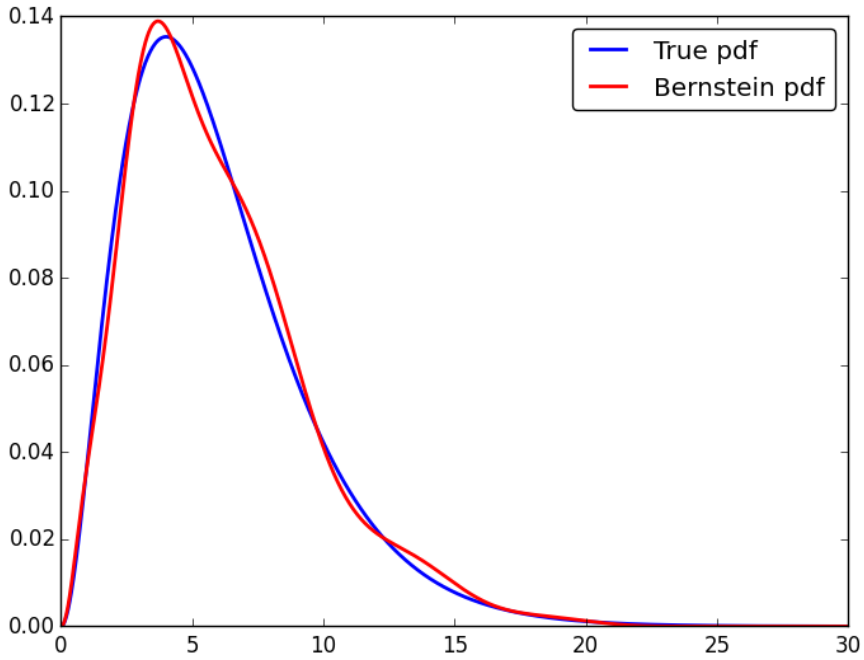

Figure: 100 samples $\sim$ Gamma(3,2); Time: 0.6789s

Example: bimodal estimation

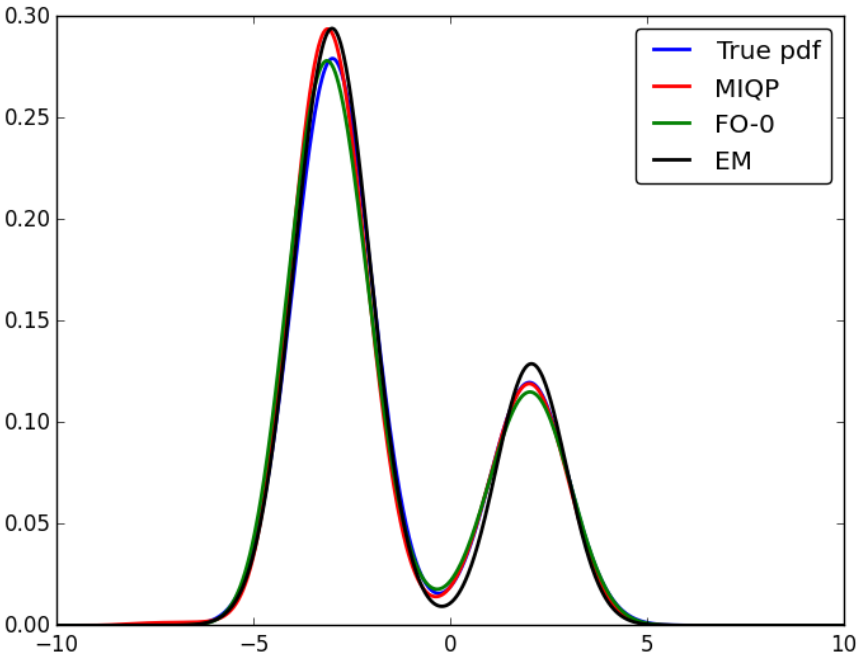

Figure: 500 samples $\sim$ 0.3N(2,1)+0.7N(-3,1); Time: 1.2108s for MIQO+FO-0 (warmstart)

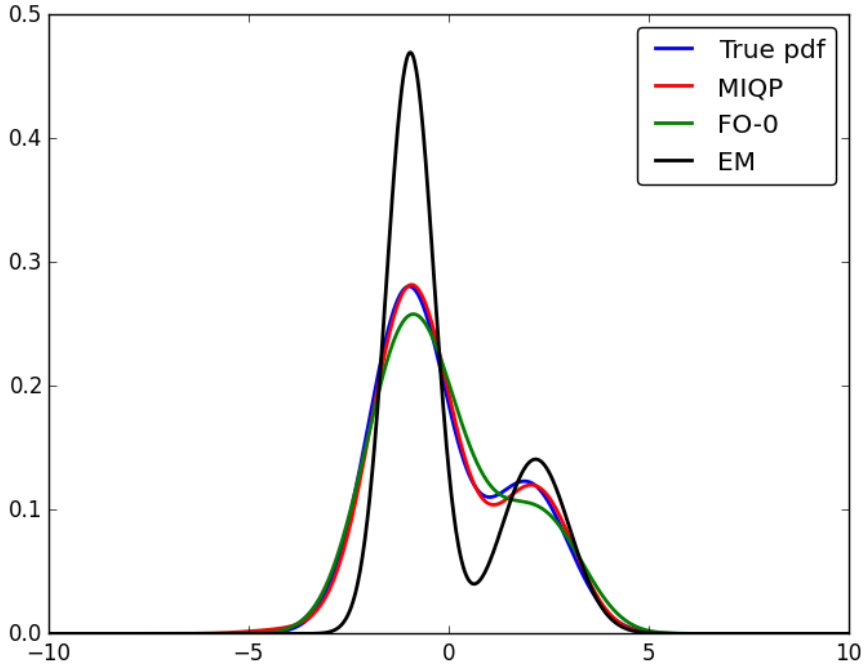

Figure: 500 samples $\sim$ 0.3N(2,1)+0.7N(-1,1); Time: 1.4516s for MIQO+FO-0 (warmstart)

Example: inferring a bimodal density from noisy data

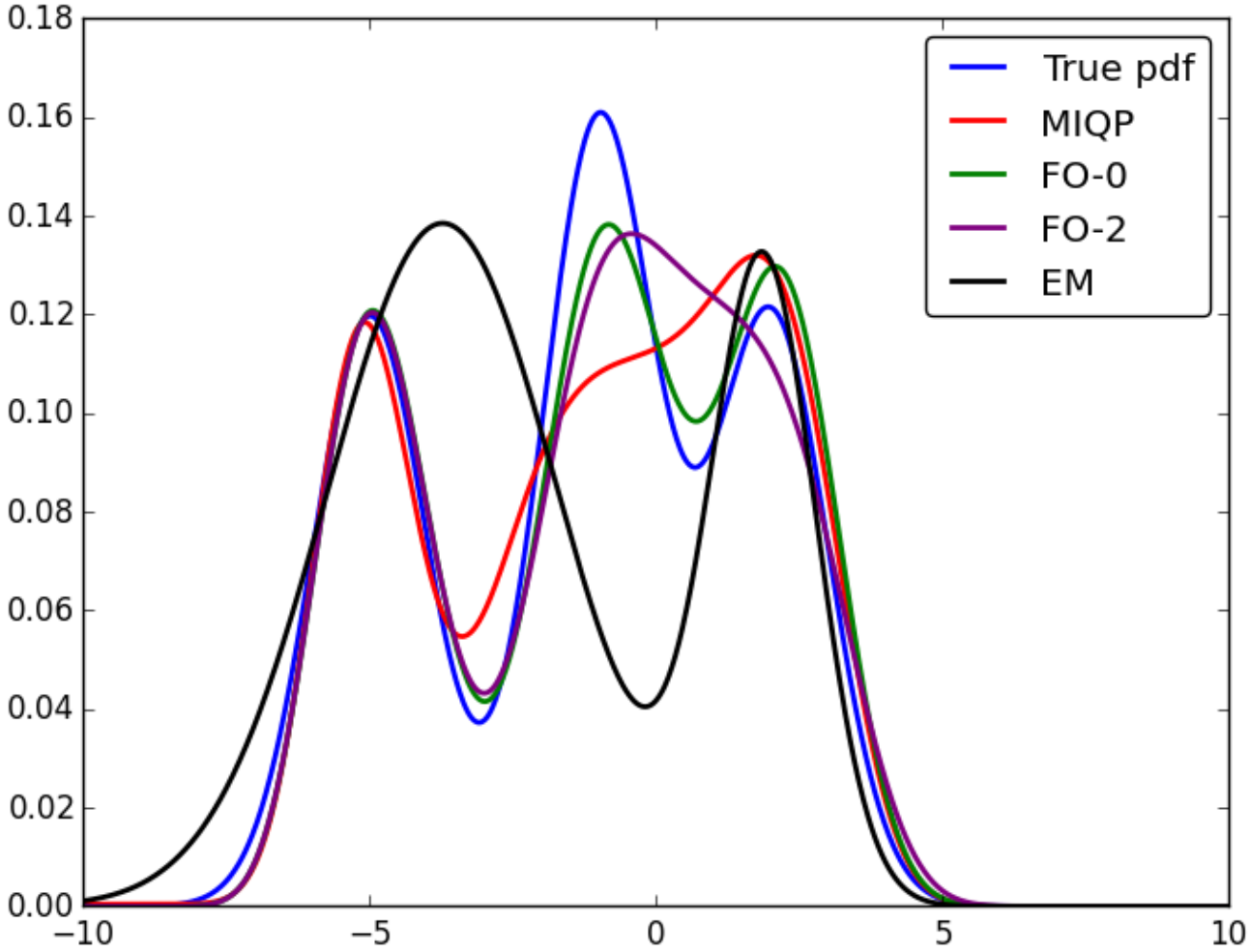

Figure: 500 samples $\sim$ 0.3N(2,1)+0.4N(-1,1)+0.3N(-5,1); Time: 1.4699s

Our method is based on a polynomial mixture model.

- Let $x_1, \ldots, x_n \in (a, b)$ be IID random samples drawn from some distribution with an unknown density f.
- We want to find a polynomial estimate of f, in the form of

$$B_m(x, f) = \sum_{k=0}^{m} f\left(a + \frac{k(b-a)}{m-1}\right) \binom{m-1}{k} \left(\frac{x-a}{b-a}\right)^k \left(\frac{b-x}{b-a}\right)^{m-k+1}$$

- $||B_m(\cdot, f) - f(\cdot)||_\infty \to 0$ on $[a, b]$ as $m \to \infty$. (Bernstein polynomials)
- Since we want $B_m(x, f)$ to be a pdf, it must be a Beta mixture

$$f_m(x, \mathbf{w}) = \sum_{j=0}^{m} w_j f_b(x; j+1, m-j+1)$$

for some $\mathbf{w}$, where $f_b(x; j+1, m-j+1)$ is the scaled Beta pdf with shape parameters k and $m-k+1$ defined on (a, b).

The objective is maximizing the log-likelihood function or minimizing the mean squared CDF error.

- We want to maximize the log-likelihood function

$$\sum_{i=1}^{n} \log \left[\sum_{j=0}^{m} w_j f_b(x_i; j+1, m-j+1) \right].$$

- Or minimizing the mean squared CDF error (Anderson-Darling test)

$$\sum_{i=1}^{n} \frac{n[F_n(x_i) - F_b(x_i, \mathbf{w})]^2}{(F_n(x_i) + \epsilon_n)(1 + \epsilon_n - F_n(x_i))}.$$

The MIQP formulation for univariate shape-restricted density estimation.

- The optimization problem we consider is of the following form.

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{r}} \sum_{i=1}^{n} r_i^2 \\
&\text{s.t. } \mathbf{r} = L^{1/2}(\mathbf{f} - B\mathbf{w}), \\
&\sum_{j=0}^{m} w_j = 1, \\
&w_j \geq 0, \forall\ j = 0, \ldots, m, \\
&\text{[Shape-restricted constraints]}
\end{aligned}$$

where $L^{1/2}_{n\times n} = \text{Diag}(n^{1/2}((F_n(x_i) + \epsilon_n)(1 + \epsilon_n - F_n(x_i)))^{-1/2})$, $B_{n\times m} = (F_b(x_i, \mathbf{w}, j, m - j + 1))$, $\mathbf{f}_{n\times 1} = (F_n(x_i))$. $\mathbf{r}$ is introduced mainly to sustain the numerical stability of the problem.

We can reduce shape restrictions on $f_m(x, \mathbf{w})$ to $\mathbf{w}$.

- Among the polynomial family, Bernstein polynomials are known to be the best "shape-preserving".
- We can sufficiently model the monotonicity, concavity, convexity, k-modality, log-concavity of $f_m(x, \mathbf{w})$ by constraining $\mathbf{w}$ to be monotone, concave, convex, having at most k modes, log-concave.

Monotonicity/concavity/convexity shape restrictions are "simple".

Theorem

(Monotonicity, concavity, convexity)

(i) *If* $w_0 \geq w_1 \geq \ldots \geq w_m$, *then* $f_m(x, \mathbf{w})$ *is monotonically nonincreasing in (0,1); If* $w_0 \leq w_1 \leq \ldots \leq w_m$, *then* $f_m(x, \mathbf{w})$ *is monotonically nondecreasing in (0,1).*

(ii) *If* $2w_j \geq w_{j-1} + w_{j+1}$ *for all* $j = 1, \ldots, m-1$, *then* $f_m(x, \mathbf{w})$ *is concave in* $(0, 1)$*; If* $2w_j \leq w_{j-1} + w_{j+1}$ *for all* $j = 1, \ldots, m-1$, *then* $f_m(x, \mathbf{w})$ *is convex in* $(0, 1)$.

- These shape restrictions can thus be modeled as linear inequalities.

$\mathbf{w}$ having at most k modes is sufficient for $f_m(x, \mathbf{w})$ to have at most k modes.

Theorem

(k-modality) If the weight sequence $\mathbf{w}$ satisfies the following condition:

$$w_0 \leq \ldots \leq w_{j_1^*} \geq \ldots \geq w_{j_2^*} \ldots w_{j_{2k-1}^*} \geq \ldots \geq w_m \tag{1}$$

or

$$w_0 \geq \ldots \geq w_{j_1^*} \leq \ldots \leq w_{j_2^*} \ldots w_{j_{2k-1}^*} \leq \ldots \leq w_m \tag{2}$$

holds for some $j_1^*, \ldots, j_{2k-1}^*$. Then $f_m(x, \mathbf{w})$ has at most k modes in $(0, 1)$.

Mixed integer formulation for the k-modality constraint.

- The k-modality constraint can be written sufficiently as (M is a big constant)[1]

$$
\begin{aligned}
& w_j \leq w_{j+1} + M z_{j+1}, \forall\, j = 0, \ldots, m-1, \\
& w_j \geq w_{j+1} - M(1 - z_{j+1}), \forall\, j = 0, \ldots, m-1, \\
& z_{j+1} - z_j \leq d_j, \forall j = 1, \ldots, m-2, \\
& -z_{j+1} + z_j \leq d_j, \forall j = 1, \ldots, m-2, \\
& \sum_{j=0}^{m-2} d_j \leq 2k - 1, z_1 = 0, \\
& \text{All } z_j \in \{0, 1\}.
\end{aligned}
$$

[1] The SOS-1 constraint denotes at most one of σ_j^+ and σ_j^- is nonzero.

Mixed integer formulation for the log-concavity constraint.

Theorem

(Log-concavity) If the nonnegative weight sequence $\mathbf{w}$ *is log-concave, i.e.,* $w_j^2 \geq w_{j-1} w_{j+1}$ *for* $1 \leq j \leq m-1$. *Then* $f_m(x, \mathbf{w})$ *is log-concave in* $(0, 1)$.

- So the log-concavity constraint can be written sufficiently as ($p \in \mathbf{N}$)

$$\sum_{j=0}^{m} w_j^{(p)} = 1, w_j^{(p)} = \sum_{s=0}^{p} u_{j,j,s+1,s+1} 2^{-s-1}, \forall\, j = 0, \ldots, m,$$

$$u_{i,j,s,l} \leq u_{i,i,s,s}, \;\; u_{i,j,s,l} \leq u_{j,j,l,l}, \;\; u_{i,j,s,l} \geq u_{i,i,s,s} + u_{j,j,l,l} - 1,$$

$$\sum_{s,l=0}^{p} 2^{-s-l-2} u_{j,j,s+1,l+1} - \sum_{s,l=0}^{p} 2^{-s-l-2} u_{j-1,j+1,s+1,l+1} \geq 0,$$

All $u_{i,j,s,l} \in \{0, 1\}$.

The first order method for k-modal density estimation.

- The algorithm is a combination of the mirror descent method and prefix isotonic regression.
- Use the negative entropy function as the mirror map

$$\Phi(\mathbf{w}) = \sum_{j=0}^{m} w_j \log(w_j).$$

- In iteration $t+1$, we perform
  - (Gradient update) $\nabla\Phi(\mathbf{v}^{t+1}) = \nabla\Phi(\mathbf{w}^t) - \frac{1}{L}\nabla l(\mathbf{w}^t|\mathbf{x})$.
  - (Projection) $\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}\in\Delta_{m+1}:k\text{-modal}} \sum_{j=0}^{m} w_j \log\frac{w_j}{v_j^{t+1}}$.
- Δ_{m+1} is the $m+1$ dimensional simplex, l is the log-likelihood function. The key observation: the projection can be finished in $O(m^k)$ steps, so the algorithm solves the problem in $O(m^k \log(m)/\epsilon)$ steps, very fast when k is small.

Monotonicity/concavity/convexity constraints are "simple".

- By "simple" we mean it is easy to compute.

Distribution	m	n	Time	Error
Exponential(1)	10	10	0.0049	0.0059
	100	10	0.0034	0.0113
	10	100	0.0101	0.0176
	100	100	0.0109	0.0230
Exponential(2)	10	10	0.0052	0.0973
	100	10	0.0089	0.0897
	10	100	0.0118	0.0834
	100	100	0.0340	0.0922

Table: Computational results for convex nonincreasing density estimation

FO-1 is more efficient for unimodal density estimation.

			MIQP		MIQP +FO-0		FO-1	
Distribution	n	m	Time	Error	Time	Error	Time	Error
Normal(0,2)	100	100	0.3039	0.0079	0.3758	0.0079	0.4205	0.0064
		500	3.6330	0.0021	3.3494	0.0021	0.5298	0.0030
	250	100	1.3189	0.0061	1.0522	0.0061	1.3912	0.0046
		500	24.228	0.0038	18.936	0.0038	7.8491	0.0023
	500	100	8.2476	0.0099	6.0760	0.0099	3.3779	0.0105
		500	72.362	0.0041	62.928	0.0041	15.860	0.0025
Gamma(3,2)	100	100	0.2421	0.0045	0.2538	0.0045	0.6788	0.0034
		500	3.9238	0.0026	0.0027	0.0026	3.6238	0.0026
	250	100	0.7336	0.0047	0.6752	0.0047	0.0048	0.0039
		500	12.382	0.0024	11.244	0.0024	7.8209	0.0012
	500	100	2.6342	0.0110	2.5338	0.0110	3.0447	0.0095
		500	39.604	0.0029	37.877	0.0029	15.572	0.0030

Table: Computational results for unimodal density estimation

FO-k is good for bimodal density estimation, MIQP+FO-0 scales better in k. They both outperform EM.

				MIQP		MIQP +FO-0		FO-k		EM-k	
k	Distribution	m	n	Time	Error	Time	Error	Time	Error	Time	Error
2	0.3N(2,1)+0.7N(-3,1)	100	100	0.265	0.0056	0.383	0.0056	5.582	0.0085	0.006	0.0107
			500	2.450	0.0029	1.450	0.0029	6.324	0.001	0.009	0.0031
		250	100	17.62	0.0104	11.32	0.0104	60.56	0.0100	0.006	0.0107
			500	78.49	0.0038	70.32	0.0038	60.83	0.0031	0.009	0.0031
		500	100	517.8	0.0204	464.8	0.020	441.8	0.0196	0.006	0.0107
			500	>600	0.0078	>600	0.0040	447.0	0.0068	0.009	0.0031
	0.3N(2,1)+0.7N(-1,1)	100	100	0.138	0.0094	0.521	0.0094	7.035	0.0105	0.030	0.0139
			500	0.162	0.0029	1.151	0.0029	8.193	0.0035	0.026	0.0119
		250	100	24.04	0.0061	23.44	0.0061	63.87	0.0090	0.030	0.0139
			500	123.1	0.0040	89.27	0.0040	66.85	0.0036	0.026	0.0119
		500	100	>600	0.0110	556.1	0.0103	468.6	0.0154	0.030	0.0139
			500	>600	0.0053	>600	0.0045	476.1	0.0039	0.026	0.0119
3	0.3N(2,1)+0.4N(-1,1) +0.3N(-5,1)	100	100	7.525	0.0125	0.432	0.0125	186.6	0.0126	0.010	0.0280
			500	21.61	0.0038	1.469	0.0038	199.2	0.0041	0.025	0.0255
		250	100	33.87	0.0123	31.25	0.0123	>600	0.0231	0.010	0.0280
			500	423.6	0.0062	325.1	0.0052	>600	0.0096	0.025	0.0255
		500	100	>600	0.0121	558.3	0.0194	>600	0.0225	0.010	0.0280
			500	>600	0.0063	>600	0.0077	>600	0.0326	0.025	0.0255

Table: Computational results for k-modal density estimation

MIQP+FO-1 is good for log-concave density estimation.

- Log-concave density estimation is a well-studied problem, but the optimal solution is generally nonsmooth. Our approach, instead, provides log-concave smooth density estimates.

				MIQP		MIQP+FO-1	
Distribution	m	p	n	Time	Error	Time	Error
Normal(0,2)	10	5	10	0.2581	0.0203	0.5379	0.0203
			100	0.2526	0.0172	0.7124	0.0172
		10	10	0.6122	0.0190	1.2937	0.0190
			100	0.8271	0.0168	1.3573	0.0168
	50	5	10	45.952	0.0201	12.753	0.0201
			100	95.19	0.0063	56.381	0.0063
		10	10	74.535	0.0211	22.169	0.0211
			100	106.54	0.0049	62.649	0.0049
	100	5	10	166.35	0.0253	133.25	0.0174
			100	>600	0.0099	184.82	0.0086
		10	10	418.98	0.0239	116.45	0.0136
			100	>600	0.0104	227.34	0.0055

Table: Computational results for log-concave density estimation

Conclusions

- We develop a computational framework for univariate smooth density estimation with general shape restrictions: monotonicity, convexity/concavity, log-concavity, k-modality, etc.
- Through numerical experiments we show it provides not only an efficient but also a general model for univariate density estimation.
- Some future directions might include generalization to higher dimensional models (with a sparsity/robustness consideration).