Optimization Heuristics for Unimodal Kernel Density Estimation by Data Sharpening

Mark A. Wolters and W. John Braun

Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada

Summary

Problem. How can additional qualitative information about a density's shape be used to obtain an improved estimate?

Our main focus. Rendering a kernel density estimate unimodal, using data sharpening.

Difficulty. Performing data sharpening involves a hard optimization problem. Standard methods (e.g. sequential quadratic programming) often fail to converge.

Our contribution. A greedy algorithm that always finds a feasible solution. Solutions are typically of good quality.

Impact. This work applies generally, and could lead to a better optimizer for handling qualitative constraints in other nonparametric settings.

Background

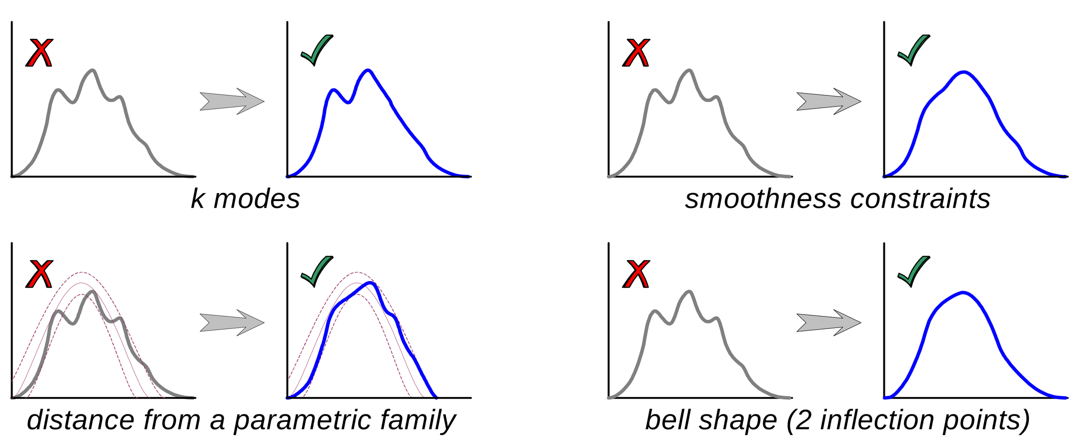
Constrained Estimation

Common examples: –isotonic regression
–monotone or unimodal density estimation

Qualitative Constraints

Natural-language shape constraints on an estimator. Such constraints are not always expressible in the standard form of mathematical programming.

Examples for density estimation:



Data Sharpening

A general strategy for constrained estimation [2], where the observed data points (not the estimator) are modified to achieve constraint satisfaction. Sharpening is done by moving points the minimum amount, as measured by some *sharpening distance*, necessary to satisfy the constraint.

Framework:

$\mathbf{x}$	the original data, length n
$\mathbf{y}$	the perturbed (sharpened) data vector
$\hat{f}_{\mathbf{y}}$	the estimator based on data $\mathbf{y}$
$\delta(\mathbf{y},\mathbf{x})$	a measure of distance between $\mathbf{y}$ and $\mathbf{x}$
$\mathcal{C}$	the set of feasible $\mathbf{y}$'s

GOAL: find $\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ \delta(\mathbf{y},\mathbf{x})$

Sequential Quadratic Programming (SQP)

A deterministic, gradient-based nonlinear programming method suitable for constrained optimization with a convex objective function and convex constraints.

Difficulties with SQP

Some constraints (like unimodality) are hard to express in a form suitable for SQP codes. The problem has dimension n and could have a non-convex constraint set or multiple optima.

⇒ **SQP can be slow, return poor solutions, or fail to converge.**

The Sharpening Optimization Problem

Unimodal Kernel Density Estimation

The kernel density estimator (KDE) with kernel function K_h is: $\hat{f}_{\mathbf{y}}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - y_i)$.

Using the L_1 distance, we want to find: $\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ \sum_{i=1}^{n} |y_i - x_i|$,

Where $\mathcal{C}$ is the set of $\mathbf{y}$'s that produce a **unimodal** KDE:

$$\mathcal{C} = \left\{ \mathbf{y} : \exists m \text{ s.t. } \begin{array}{l} \hat{f}'_{\mathbf{y}}(x) \geq 0 \text{ if } x \leq m \\ \hat{f}'_{\mathbf{y}}(x) \leq 0 \text{ if } x \geq m \end{array} \right\}$$

Note:

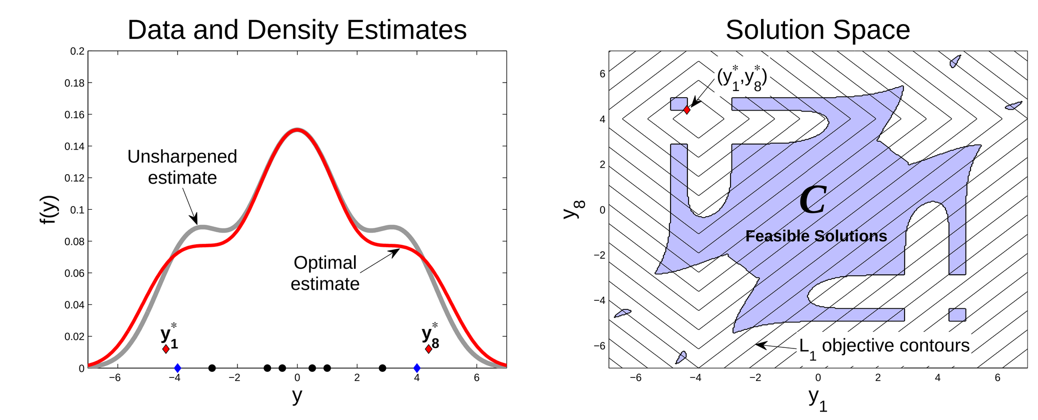
◇ SQP only works when m is known. To solve by SQP, must iterate over a range of candidate m values.

◇ There is some evidence that the L_1 objective has better statistical performance than the L_2 distance for unimodal density estimation [2,4].

An Illustration

Consider a case where sharpening is done by moving only two of eight points. Unsharpened data: $\mathbf{x} = [x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8]$,
Sharpened data: $\mathbf{y} = [y_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ y_8]$.

Then the solution space is composed of (y_1, y_8) pairs.



Even for this simplified example, the feasible set has complex, non-convex structure, making it hard to find the optimum by standard methods.

A Greedy Algorithm—`improve( )`

Main Idea

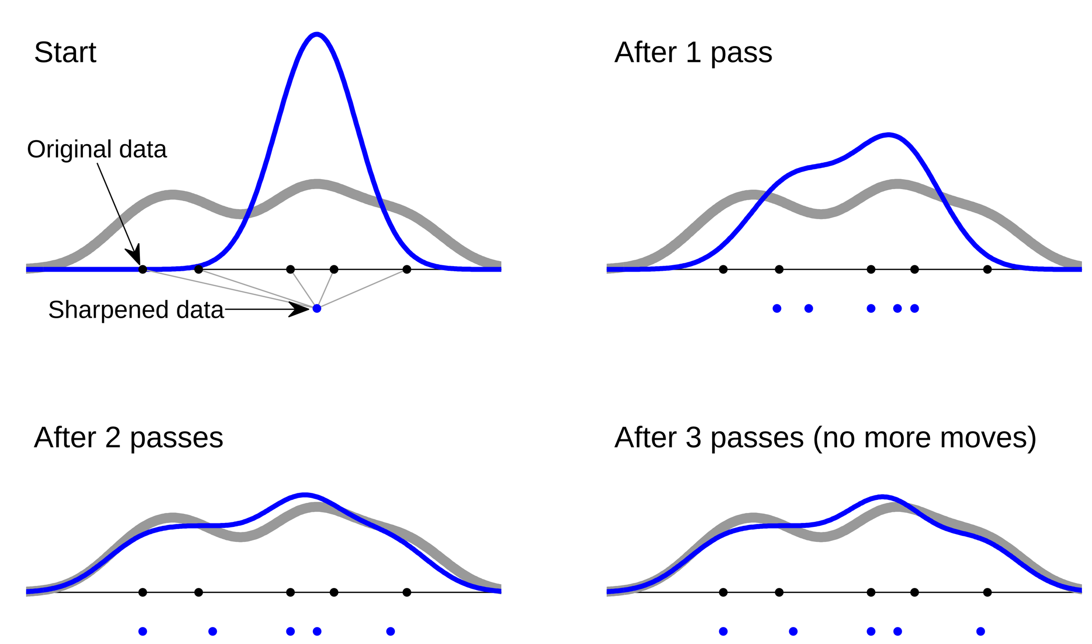
Use a heuristic approach to find good solutions while avoiding SQP's optimization difficulties.

Starting with a naïve feasible solution $\mathbf{y}_0$, repeatedly pass through the data, moving points one at a time to get closer to $\mathbf{x}$. **Preserve feasibility throughout.**

Stop when no points are moveable, i.e., when no $\mathbf{y}$ point can be moved closer to its corresponding $\mathbf{x}$ point without violating the constraint.

Optimization Function: $\mathbf{y} = \texttt{improve}(\mathbf{y}_0, \mathbf{x})$

```
Let y be the initial guess: y := y0.
Set the number of grid search steps to S := 1.
WHILE there are still moveable points:
    FOR each point in y, in decreasing order of |yi − xi|:
        If feasible, move yi closer to xi.
        (Use a grid search with S steps).
    END FOR
    IF no points were moved:
        Double the number of grid-search steps: S := 2S.
    END IF
END WHILE
```



◇ As a starting value, we put all points at the highest unsharpened mode.

◇ During each pass, begin with a coarse search grid so large moves are made first. Shrinking the grid gradually helps to fine-tune the solutions.

Performance: Greedy vs. SQP

A Simulation Study

Methods compared:	Greedy, SQP.
True distribution:	t distribution with 5 degrees of freedom.
Replicates:	500 random samples of size $n = 25$.
Estimator:	Gaussian KDE, unimodality constraint, bandwidth 1/2 of the normal-scale bandwidth.
Objective function:	L_1 sharpening distance, $\sum_{i=1}^{n} \lvert y_i - x_i\rvert$

Results

1) Greedy algorithm always returns a solution, and does so quickly.

Runs Returning a Feasible Solution (%)	
SQP	70
Greedy	100

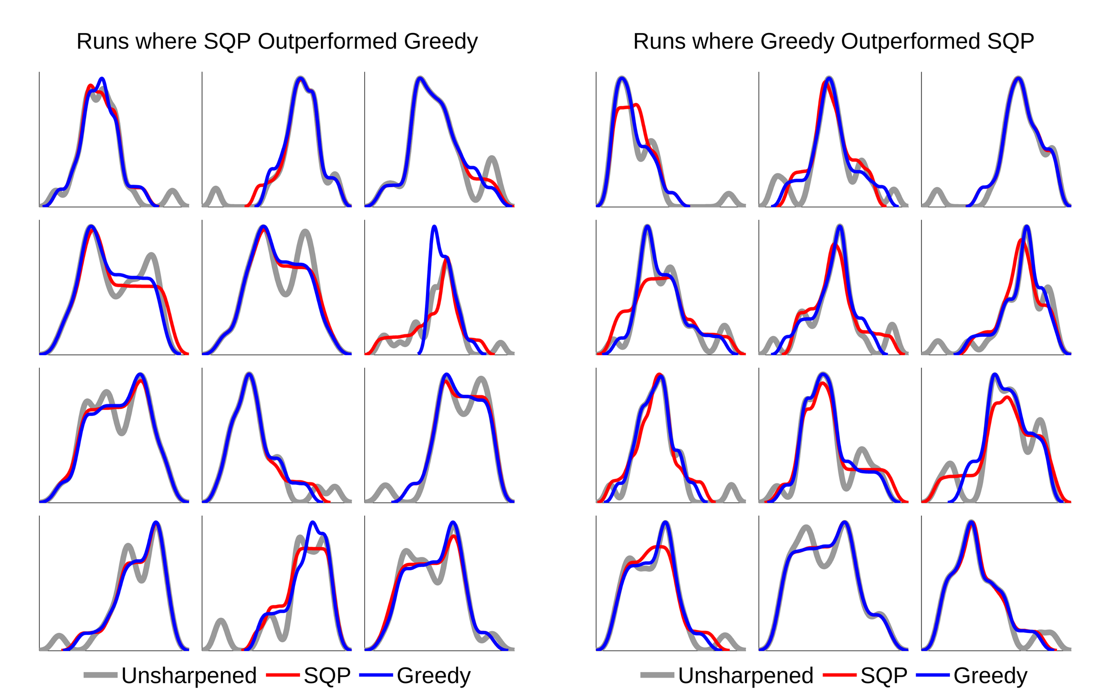
Mean Run Time (s)	
SQP	89
Greedy	0.32

2) Greedy solutions often have better objective function values than SQP.

Percent of runs where greedy found a better solution than SQP:	68
Mean sharpening distance:	SQP 3.65 Greedy 2.86
95% confidence interval for mean improvement in sharpening distance:	$(0.54, 1.05)$

Example Estimates

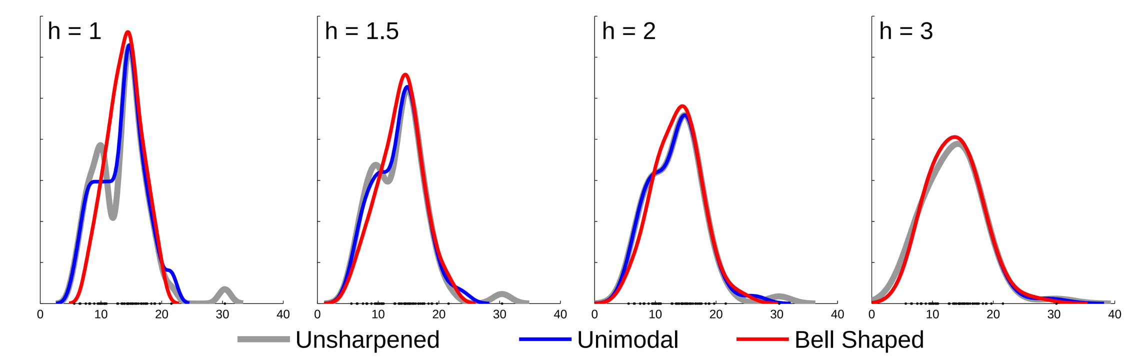
A random sample of estimates from simulation runs:



Examples and Extensions

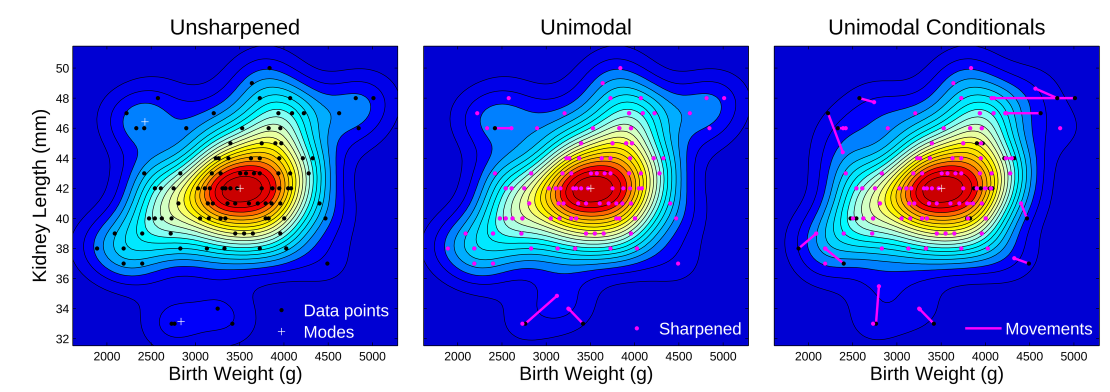
Extreme Wind Speed Events [1]

Estimator: Gaussian KDE, bandwidth h, sample size 57.
Case 1: unsharpened (no constraint).
Case 2: unimodal constraint with non-negative support.
Case 3: bell-shaped constraint with non-negative support.



Birth Weights and Kidney Lengths in Infants [3]

Estimator: Bivariate Gaussian KDE, product kernel, normal-scale bandwidths, sample size 102.
Case 1: unsharpened (no constraint).
Case 2: unimodal constraint (only one local maximum, no local minima).
Case 3: constrained to have unimodal conditional distributions.



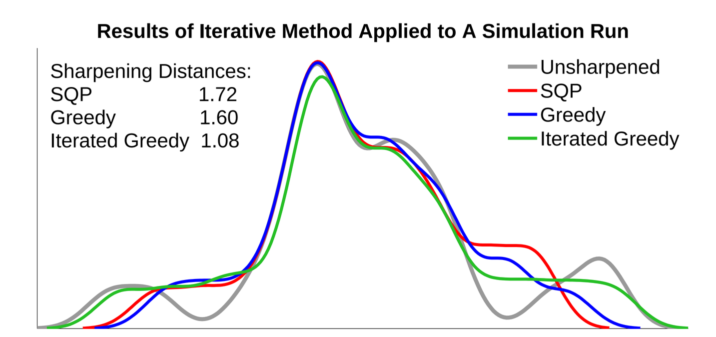
New Metaheuristics for Improved Optimization

The greedy algorithm lends itself to various iterative or population-based schemes to further improve solution quality.

E.g.: 1) Start with standard greedy solution, $\mathbf{y}$.
2) Perturb it by adding noise: $\mathbf{y}_\epsilon = \mathbf{y} + \epsilon$.
3) Use `improve` to make it feasible: $\mathbf{y}_f = \texttt{improve}(\mathbf{y}, \mathbf{y}_\epsilon)$.
4) Use `improve` to move it toward $\mathbf{x}$: $\mathbf{y}_{new} = \texttt{improve}(\mathbf{y}_f, \mathbf{x})$.
5) Repeat steps 2-4, keeping the better of $\mathbf{y}, \mathbf{y}_{new}$ each time.

This simple algorithm was run on the simulation data described at left:

93% of runs found a better solution than SQP.



References

[1] Alibrandi and Ricciardi (2008), Efficient Evaluation of the PDF of a Random Variable through the Kernel Density Maximum Entropy Approach, *International Journal for Numerical Methods in Engineering*, vol. 75, pp. 1511-1548.

[2] Braun and Hall (2001), Data Sharpening for Nonparametric Inference Subject to Constraints, *Journal of Computational and Graphical Statistics*, vol. 10, no. 4, pp. 786-806.

[3] Fitzsimons (1983), Kidney Length in the Newborn Measured by Ultrasound, *Acta Pediatrica Scandinavica* vol. 72, pp. 885-887.

[4] Hall and Kang (2005), Unimodal Kernel Density Estimation by Data Sharpening, *Statistica Sinica* vol. 15, pp. 73-98.

Acknowledgements

Funding from the following sources is gratefully acknowledged:

◇ Natural Sciences and Engineering Research Council of Canada

◇ Ontario Student Assistance Program