`scdensity`: an R Package for Shape-Constrained Kernel Density Estimation

Mark A. Wolters
Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China

download the poster!

Summary

Adding ***shape constraints*** to a nonparametric estimator:

- eliminates unrealistic waves and bumps in the estimate
- maintains more shape flexibility than parametric families
- improves statistical performance in small samples

The `scdensity` package implements two related methods for enforcing constraints on a kernel density estimator (KDE):

1. The ***weighted KDE*** method (Hall and Huang, 2002)
2. The ***adjusted KDE*** method (Wolters and Braun, 2018)

It unifies the methods under a ***common optimization scheme*** and makes estimation ***numerically stable***.

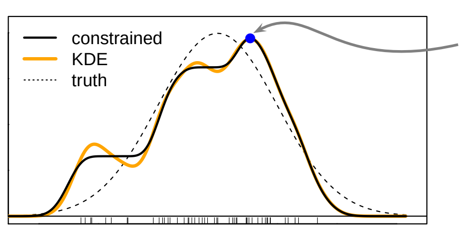
The package makes it ***easy*** to get estimates with ***many different constraints***, using ***familiar kernel methods***.

Which Constraints Can it Handle?

In these examples, set the bandwidth to `h <- bw.SJ(x)/2`.

Unimodal

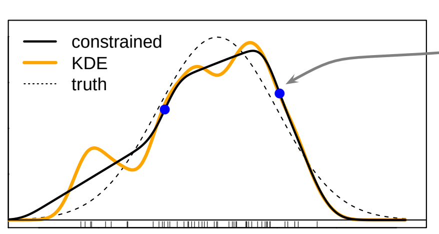
```
scdensity(x, bw=h, constraint="unimodal")
```



Derivative sign changes are "important points." If their locations are *known*, the optimization problem is a quadratic program (QP).

Two inflection points

```
scdensity(x, bw=h, constraint="twoInflections")
```

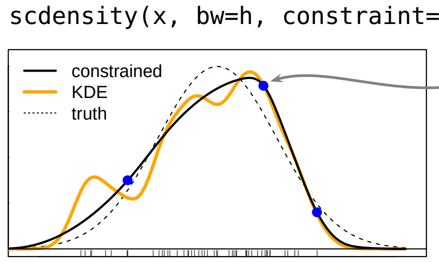


Sign changes of f''.

If important points are not known, the QP is run inside a search routine to find them.

Three inflection points in f'

```
scdensity(x, bw=h, constraint="twoInflections+")
```

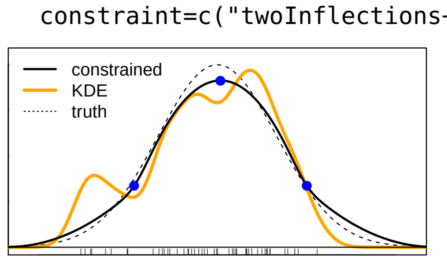


Sign changes of f'''.

Restricting sign changes of higher derivatives enforces greater smoothness.

Symmetric

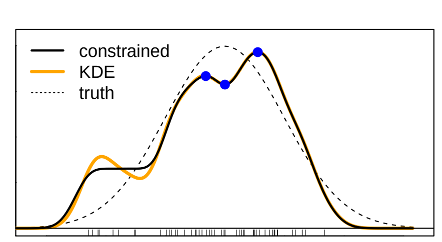
```
scdensity(x, bw=h,
  constraint=c("twoInflections+","symmetric"))
```



multiple constraints can be combined.

Bimodal

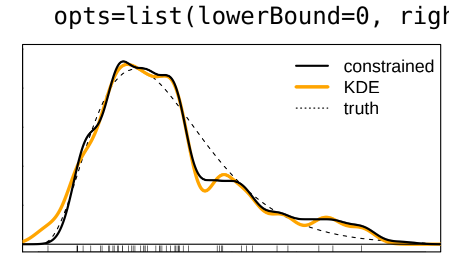
```
scdensity(x, bw=h, constraint="bimodal")
```



In this case the important points are the locations of the modes and their intervening antimode.

Monotone, and/or bounded

```
scdensity(x, bw=h,
    constraint=c("monotoneRightTail","boundedLeft"),
    opts=list(lowerBound=0, rightTail=50))
```



This estimate has:

- negligible probability mass to the left of zero
- monotonicity to the right of the original estimate's median.

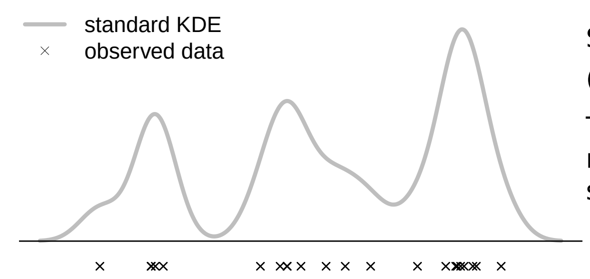
How does it work?

A weighted KDE is $f_{\mathbf{s}}(x|\mathbf{p}) = \frac{1}{h}\sum_{i=1}^{r} p_i K\left(\frac{x-s_i}{h}\right)$, where $\mathbf{s}$ are the ***kernel centers*** and $\mathbf{p}$ are the ***weights***.

The data are $\mathbf{x}$. We do not require kernel centers to be located at $\mathbf{x}$!

We can express the integrated squared error (ISE) between any two weighted KDEs as a quadratic form in the weights.

1 *Initial estimate*

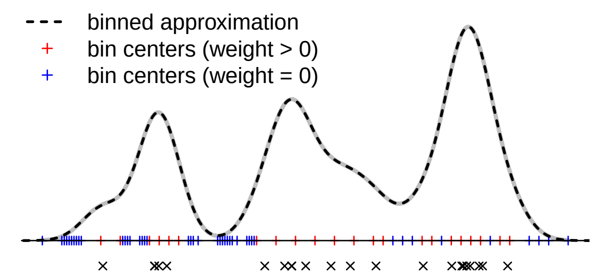


Start with the standard KDE, $f_{\mathbf{x}}(x|\frac{1}{n}\mathbf{p}_{\text{unif}})$ (where $\mathbf{p}_{\text{unif}} = \frac{1}{n}\mathbf{1}$)

This is the chondrite data (percent Si in 22 meteorites, Good and Gaskins, 1980). We seek a bimodal estimate.

2 *Binning step*

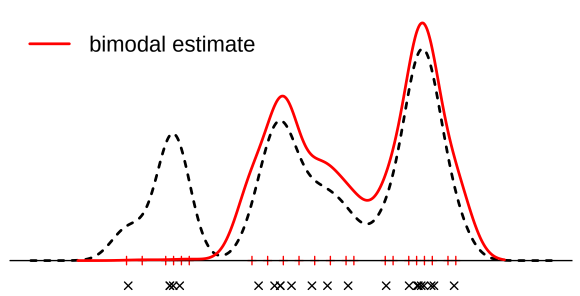
1. Set up a uniform grid of kernel centers, $\mathbf{s}$.
2. Find $\tilde{\mathbf{w}}$ that minimizes $ISE(f_{\mathbf{s}}(x|\mathbf{w}), f_{\mathbf{x}}(x|\mathbf{p}_{\text{unif}}))$ (this is a QP with no shape constraints)
3. Subdivide any intervals where $f_{\mathbf{x}}$ is poorly approximated.
4. Repeat 2 & 3 until approximation is good.



End result:

- $f_{\mathbf{s}}(x|\tilde{\mathbf{w}})$ closely approximates $f_{\mathbf{x}}(x|\mathbf{p}_{\text{unif}})$.
- $\tilde{\mathbf{w}}$ contains both zero and nonzero weights.

3a *Estimation step (weighted KDE method)*

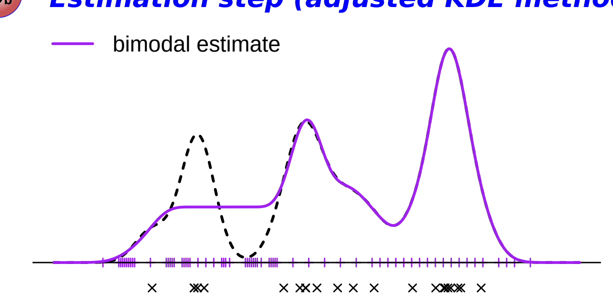


Find new weights, $\hat{\mathbf{v}}$ that minimize $ISE(f_{\mathbf{s}}(x|\mathbf{v}), f_{\mathbf{s}}(x|\tilde{\mathbf{w}}))$, subject to shape constraints

If we use only the nonzero-weighted centers, we have the weighted KDE method.

The method cannot modify density shape away from $\mathbf{x}$ points.

3b *Estimation step (adjusted KDE method)*



If we use all centers, we have the adjusted KDE method.

This method is more flexible and can handle more constraints.

The QPs in the binning and estimation steps have the same form.
The estimation step is run inside a search to find the best mode locations.

Examples

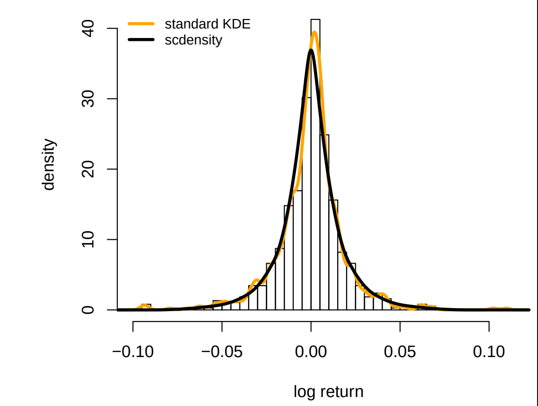
S&P 500 log returns (Turnbull & Ghosh, 2014).

Constraints: `twoInflections+`, symmetric around zero.

Smooth tails

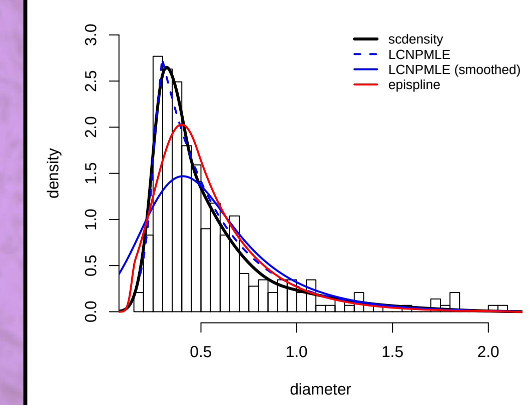
No restriction on tail weight

Fixed mode location



Axon diameters (Sepehrband *et al.*, 2016).

Constraint: `twoInflections+`.



Original authors compared 16 parametric options

`scdensity` provides a "parametric-like" shape

Also shown: log-concave estimates from `logcondens`, smooth unimodal estimate from `episplineDensity`.

Q & A

Is it fast?

- A fraction of a second for $N(0,1)$ data with `unimodal` constraint.
- Several seconds for t_5 data with `twoInflections+`.

Is it robust?

- The QP problem is convex, but can be ill-conditioned. The package checks for problems and remedies them.
- Constraint systems are occasionally infeasible. The package checks feasibility and handles problems gracefully.

What about asymptotics?

- Because we use the usual kernel density estimator, we can borrow its asymptotic behavior. If the constraints are valid, necessary shape adjustments should shrink to zero.

References

Good, I & Gaskins, R (1980), "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," *Journal of the American Statistical Association*, **75**(369), 42–56

Hall, P & Huang, LS (2002), "Unimodal density estimation using kernel methods," *Statistica Sinica*, **12**, 965–990.

Sepehrband, F, Alexander, DC, Clark, KA, Kurniawan, ND, Yang, Z, & Reutens, DC (2016), "Parametric probability distribution functions for axon diameters of corpus callosum," *Frontiers in Neuroanatomy*, **10**(59).

Turnbull, BC & Ghosh, SK (2014), "Unimodal density estimation using Bernstein polynomials," *Computational Statistics and Data Analysis*, **72**, 13–29

Wolters, MA & Braun, WJ (2018), "Enforcing shape constraints on a probability density estimate using an additive adjustment curve," *Communications in Statistics—Simulation and Computation*, **47**(3), 672–691.

contact: mark@mwolters.com