 

THE "TOMATO SALAD PROBLEM" PROBLEM: AN R VINAIGRETTE

ROGER KOENKER AND JIAYING GU

1. Introduction

A fundemental problem in stereology initially considered by Wicksell (1925) involves estimating the distribution of 3d spherical radii from a sample of 2d cross-sectional radii. The problem may be viewed as an idealization of a microsopy setting in which opaque spherical objects embedded in a translucent medium are observable only from 2d slices, or more mundanely as inferring the distribution of the radii of some idealized, spherical tomatoes from a sample of slices. To add an element of verisimulitude to the problem it is convenient to assume that radii, y, are bounded above by $\bar{y}$ and below by $\underline{y}$. Wicksell showed that the relationship between the density of the radii of the slices, f, and the density of the radii of the tomatoes, g, is given by,

$$f(x) = C \int_{\underline{y}}^{\bar{y}} I_{[x,\bar{y}]}(y) x (y^2 - x^2)^{-1/2} dG(y).$$

In an effort to make this look as much like a conventional mixture problem as possible, one can write this as,

$$f(x) = \int_{\underline{y}}^{\bar{y}} \varphi(x, y) dG(y).$$

where $\varphi(x, y) = (x/y)(y^2 - x^2)^{-1/2}$ can be interpreted as a conditional density for x given y, supported on the interval $[0, y]$. To accomodate the truncation we can renormalize the conditional density as done for the missing species Poisson mixture problem.

Date: January 17, 2019.

2. Reparameterization

For a moment let's simplify by setting $\underline{y} = 0$ and $\bar{y} = +\infty$ and use the reparameterization in Groeneboom and Jongbloed (2014) to consider squared radii of both the balls and the circles. Abusing notations slightly we have the following density function of the observed squared radii of the circles

$$f(z) = C \int_z^{+\infty} (y - z)^{-1/2} g(y) dy$$

with $C = 2m_F := 2\int_0^{+\infty} \sqrt{y} g(y) dy$. Absorbing the constant into $g(y)$ we can reformulate $f(z)$ as

$$f(z) = \int_z^{+\infty} \varphi(z, y) h(y) dy$$

with $\varphi(z, y) = \frac{1}{2}(y - z)^{-1/2}$ and $h(y) := g(y)/\int_0^{+\infty} \sqrt{y} g(y) dy$. This leads to a slight modification of the nonparametric maximum likelihood problem in Koenker and Gu (2017):

$$\min_{h \in \mathcal{H}} \{-\sum_{i=1}^n \log f(z_i) | f(z_i) = \int \varphi(z_i, y) h(y) dy, i = 1, \ldots, n\}$$

where $\mathcal{H}$ denotes the set of functions satisfying

$$\mathcal{H} := \{h(y) : \int_0^{+\infty} \sqrt{y} h(y) dy = 1, h(y) \geq 0 \quad \forall y \in [0, \infty)\}$$

This is again a convex optimization problem and the final estimator for the density g can be obtained as

$$g(y) = h(y)/\int_0^{+\infty} h(y) dy$$

3. Two Examples

We consider two simple examples here. The first assumes the true distribution G is standard uniform. In this case, as discussed in Chapter 4.1 of Groeneboom and Jongbloed (2014), we have a closed form for the density $f(z)$ as

$$f(z) = \frac{3}{2}\sqrt{1 - z} 1\{0 \leq z \leq 1\}$$

The second example assumes the true distribution G is standard exponential. Again we have a nice closed form for the density of z that $C = 2m_F = \sqrt{\pi}$ and

$$f(z) = \frac{1}{\sqrt{\pi}} \int_z^{+\infty} \frac{1}{\sqrt{y - z}} e^{-y} dy = e^{-z} 1\{z \geq 0\}$$

In each case we generate 200 realizations. Figure 1 illustrates the NPMLE estimates, the left panel plots the estimates for $g(y)$ and the right panel plots the estimated cumulative distribution $G(y)$ against its true distribution curve in blue.

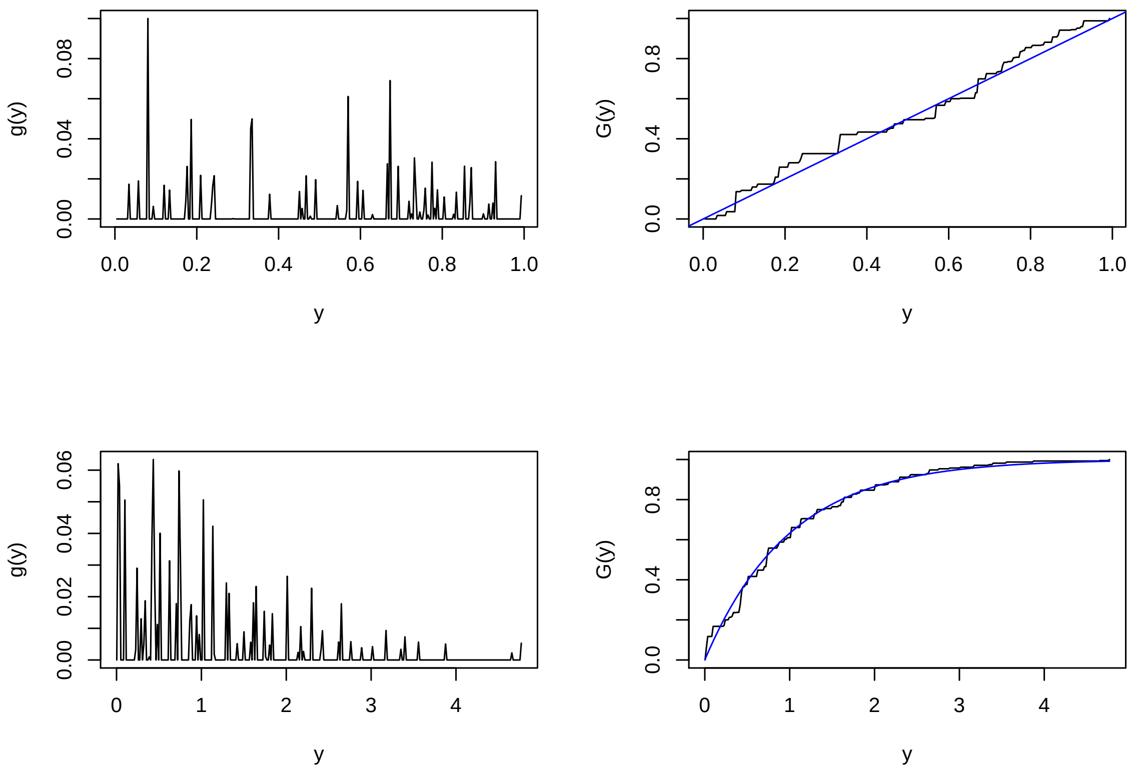

FIGURE 2. NPMLEs for two Wicksell experiments

REFERENCES

Groeneboom, P. and Jongbloed, G. (2014), *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*, Cambridge University Press.

Koenker, R. and Gu, J. (2017), 'Rebayes: An R package for empirical Bayes mixture methods', *Journal of Statistical Software* **82**, 1–26.

Wicksell, S. D. (1925), 'The corpuscle problem: A mathematical study of a biometric problem', *Biometrika* **17**, 84–99.