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 18, 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.

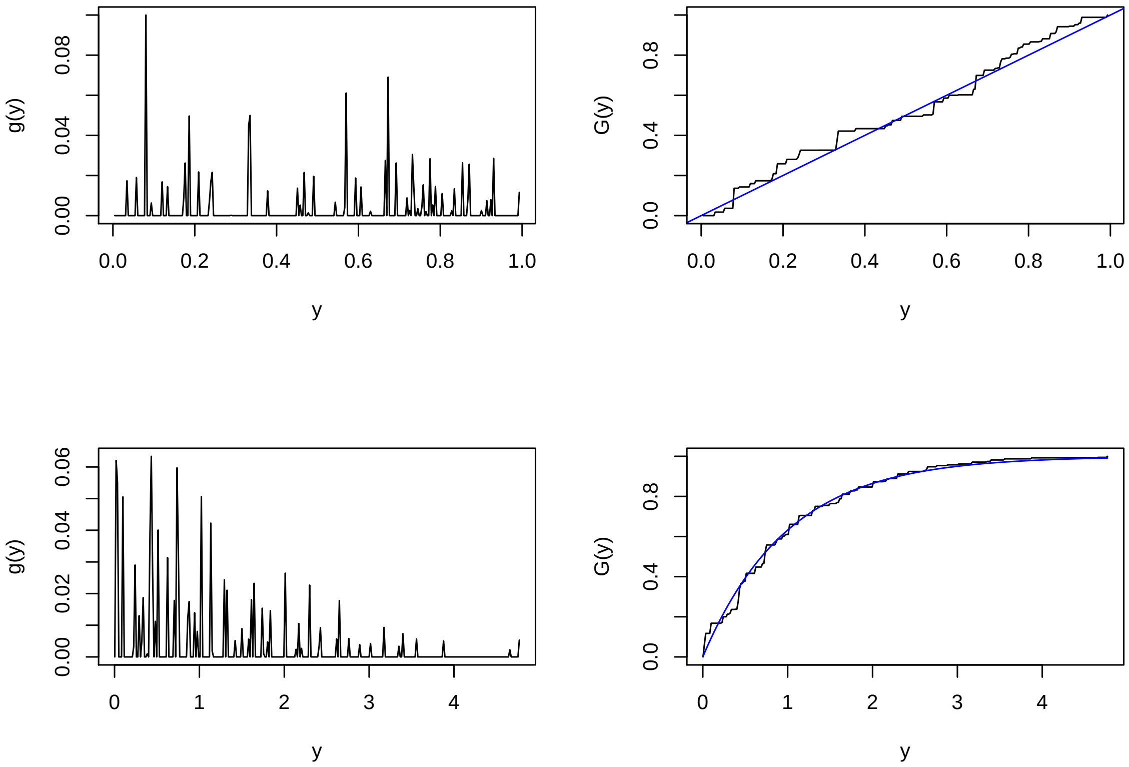

Figure 2. NPMLEs for two Wicksell experiments

4. Ill-behaviour of NPMLE and the tuning of grid size

The fit of NPMLE in the above two examples using grid size equals to 300 seem to work like a charm. However a close inspection of the likelihood function suggests that in the deconvolution formulation of the Wicksell problem, the grid size in fact acts like a regularization. This curiously contrasts with mixture models with exponential family kernel (i.e. Gaussian mixture of location with known variance, Poisson mixture and many other examples discussed in Koenker and Gu (2017)) where typically grid size is only a reflection of data resolution: once the grid is refined enough, increasing its fineness does not increase the likelihood further. In contrast, the kernel in the Wicksell problem has a singularity at zero. By choosing $h(y)$ to have a mass point arbitrarily close to any of the data points would lead to an explosion of the likelihood. In light of this, the grid size controls the smallest distance between any grid point and data point, hence acting as a regularization device. In a similar spirit, Jongbloed (2001) suggests[1] restricting the class of function $\mathcal{H}$ to have $h(y)$ taking non-zero values for $y \in \{z_1, \dots, z_n\}$. To be more explicitly, let $z_{(1)} < z_{(2)} < \cdots < z_{(n)}$ be the order statistics, and

[1] Jongbloed used a different parameterzation of the problem. Here we stick to the formulation in Groeneboom and Jongbloed (2014 book section 4.1, c.f. ex 4.8 & 4.9).

let

$$\bar{\mathcal{H}} := \{h(y), \int_0^{+\infty} \sqrt{y}h(y)dy = 1, h(y) \geq 0, \forall y \in [0,\infty) \text{ and } h(y) > 0 \text{ for } y \in \{z_1,\dots,z_n\}\}$$

For $h \in \bar{\mathcal{H}}$ let $(h_1,\dots,h_n)$ be the vector of weights associated to the grid points which are simply the order statistics of z, then the log-likelihood can be written as $\sum_{i=1}^n \log f(z_{(i)})$ with

$$\begin{aligned} f(z_{(i)}) &= \int_{(z_{(i)},+\infty)} \frac{1}{2}\frac{1}{\sqrt{x - z_{(i)}}}h(x)dx \\ &= \frac{1}{2}\frac{1}{\sqrt{z_{(i+1)} - z_{(i)}}}h_{i+1} + \frac{1}{2}\frac{1}{\sqrt{z_{(i+2)} - z_{(i)}}}h_{i+2} + \dots \frac{1}{2}\frac{1}{\sqrt{z_{(n)} - z_{(i)}}}h_n \text{ for } i = 1,\dots,n-1 \\ f(z_{(n)}) &= \int_{(z_{(n)},+\infty)} \frac{1}{2}\frac{1}{\sqrt{x - z_{(n)}}}h(x)dx = 0 \end{aligned}$$

Clearly the density for the largest sample point would lead to $-\infty$ for the log-likelihood , one simple remedy is to exclude it. Given the fact that $P(z_i = z_j) = 0$ for any $i \neq j$ since Z is a continuous random variable, for all $h \in \bar{\mathcal{H}}$, $\sum_{i=1}^{n-1} \log f(z_{(i)}) < \infty$ hence the NPMLE for $g(y)$ is well-defined when we restrict the class of function to $\bar{\mathcal{H}}$. The Wicksell function can be easily modified to accommodate the discussion above and Figure 2 illustrates the NPMLE when restricted to only allowing $h(y)$ to have jump points on data points.

But one may wonder, why is that an optimal class of function to consider? Our previous experience with NPMLE of mixture models suggests that optimal jump points of NPMLE is not necessarily data points (with the exception of the one-dimensional Cosslett problem), hence restricting mass points at data points as in Jongbloed (2001) may also lead to some sort of efficiency loss?

We know that for a given grid size, there is a unique solution for $h(y)$ and hence $g(y)$ by maximizing the log-likelihood. One consideration might be when the grid size increases, the corresponding NPMLE for $h(y)$ may lead to over-fitting for the marginal density $f(z)$. By constructing some goodness-of-fit criteria for the density of the observed data z may provide some guidance for how to choose grid size. The log product spacing method considered in Roeder (1992) may become useful. See also Roeder (1990).

References

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

Jongbloed, G. (2001), 'Sieved maximum likelihood estimation in Wicksell's problem and related deconvolution problems', *Scandinavian Journal of Statistics* **28**, 161–183.

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

Roeder, K. (1990), 'Density estimation with confidence sets exemplified by superclusters and voids in the galaxies', *Journal of the American Statistical Association* **85**, 617–624.

Roeder, K. (1992), 'Semiparametric estimation of normal mixture densities', *The Annals of Statistics* **20**, 929–943.

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

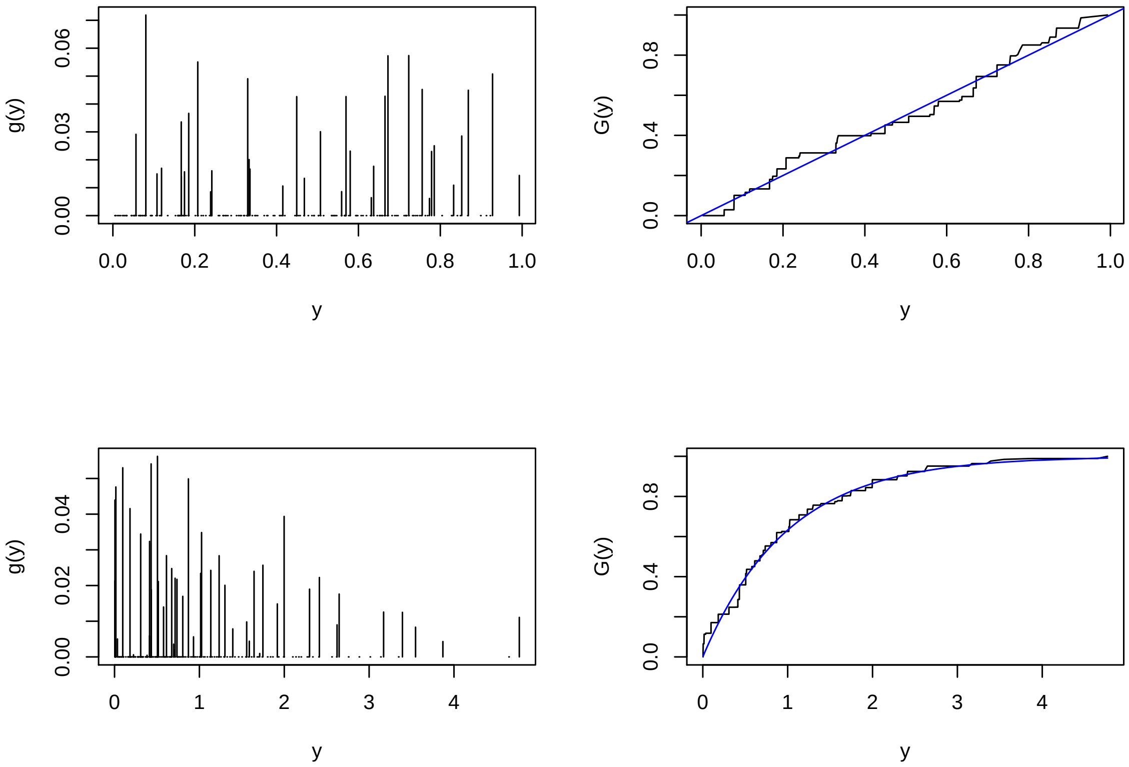

Figure 3. NPMLEs for two Wicksell experiments