A lookback, based on the Lecture Note [1]

May 14, 2007

1 Revisiting Wicksell's problem

Recall that in Wicksell's corpuscule problem, the distribution function F to be estimated was written as

$$F(x) = 1 - \frac{V(x)}{V(0)} \ , \tag{1}$$

where V is the integral

$$V(x) = \int_x^\infty \frac{1}{\sqrt{(z-x)}} dG(z) \tag{2}$$

with respect to the distribution G from which samples are available. Naïve plug-in estimators $\tilde{V}_n$ and $\tilde{F}_n$ arise from substituting the empirical distribution G_n for G in these formulae. As 1 and 2 were derived under the assumption that G is absolutely continuous with respect to Lebesgue measure (i.e. that G has a density g), and the G_n are discrete measures, the plug-in estimators have bad properties and need to be corrected. One way of performing this correction is by forcing the plug-in estimator of V (and thus also that of F) to be monotone, as V (and F) should be. This was elaborated on last week, and is illustrated in Figure 1.

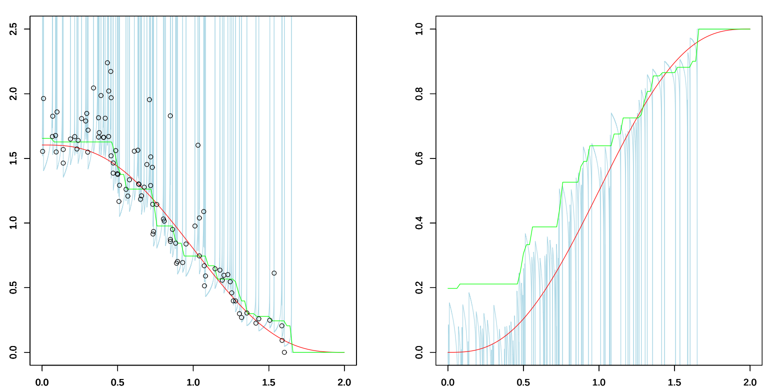

Figure 1: The functions V (left) and F (right) with their plug-in estimators, their isotonic inverse estimators via least concave majorants, and – for V – the data points used for the computation of the latter

Correcting the plug-in estimators in Wicksell's problem at the level of G

Another way of obtaining a good estimate of F is to use estimators $\tilde{G}_n$ of G that are themselves absolutely continuous, ie. by estimators g_n of the density g of G:

$$V_n(x) = \int_x^\infty \frac{dG_n(z)}{\sqrt{z-x}} = \int_x^\infty \frac{g_n(z)}{\sqrt{z-x}} dz$$

One way of performing this is by kernel estimation using a kernel $k(\cdot)$ and a bandwidth h. This yields kernel density estimates

$$g_n(z) = \frac{1}{nh} \sum_{i=1}^n k\Big(\frac{z - Z_i}{h}\Big) \ ,$$

where $Z_1, \ldots, Z_n$ are the samples from G underlying the estimation. Thus V is estimated by

$$\begin{aligned} V_n(x) &= \int_x^\infty \frac{1}{\sqrt{z-x}} \cdot \frac{1}{nh} \sum_{i=1}^n k\Big(\frac{z-Z_i}{h}\Big) dz \\ &= \frac{1}{nh} \sum_{i=1}^n \int_x^\infty \frac{k\big(\frac{z-Z_i}{h}\big)}{\sqrt{z-x}} dz \ . \end{aligned}$$

To carry out this procedure, then, the essential numbers to compute are

$$I_k(x, z_0, h) := \int_x^\infty \frac{k\big(\frac{z-z_0}{h}\big)}{\sqrt{z-x}} dz \ ,$$

where z_0 runs through all the samples Z_i and x is non-negative. However, although we can simplify this expression to

$$I_k(x, z_0, h) := 2 \int_0^\infty k\Big(\frac{z^2 + x - z_0}{h}\Big) dz \ ,$$

it remains fairly non-trivial to compute for the usual kernels, and numerical techniques may be required.

2 Isotonic Inverse Estimation for Deconvolution Problem

Let Z_i denote an observation equals the sum of two independent random variables X_i and Y_i. We assume that Y_i has an known density k and X_i has an unknown distribution function F. We know that the density of Z_i is given by the convolution of the k and F. That is,

$$g(z) = \int_{\mathbb{R}} k(z-x) \ dF(x) \ ,$$

or, equivalently, $G(x) = \int_{\mathbb{R}} K(x-z) dF(z)$, with K the distribution function of Y. Since we are interested in then unknown F and k is known, the problem is then the deconvolution of G with k. This is a type of the inverse problems where the relation between F and G is available explicitly as an inverse of the convolution. Having an explicit inverse relation of the distribution of interest F in terms of the sampling distribution G, we can construct a plug-in estimator for F via the empirical distribution function. Typically a plug-in estimator would be based on kernel estimation of g without taking into account the monotonicity of the function F given by the inverse relation. Hence, we consider an isotonic version as an estimator for F.

An explicit inverse relation of F in terms of G depends on the density k. The three most simplest cases are the exponential deconvolution, the uniform deconvolution and the laplace deconvolution. Below we show that for the above three kernels we obtain an explicit inverse relation of F.

Exponential deconvolution
Let X be a positive random variable and Y has density $k(y) = e^{-y}$, for all $y \geq 0$. That is, k is an standard exponential density. Then, from the Lecture Note,

$$F(x) = g(x) + G(x) \ .$$

Uniform deconvolution
Let X be a positive random variable and Y has density $k(y) = 1$, for all $y \in [0, 1]$. Then,

$$g(z) = \int_o^\infty k(z-x)dF(x) = \int_{z-1}^z dF(x) = F(z) - F(z-1) \ .$$

Laplace deconvolution
Let F has a support on $\mathbb{R}$ and Y has density $k(y) = \frac{1}{2}\exp(-|x|)$ for all $x \in \mathbb{R}$. Then,

$$F(x) = G(x) - g'(x) \ , \text{ at the point where } F \text{ is differentiable} \ .$$

To see this, note that the standard Laplace has distribution function $K(y)$ equal to $1 - \frac{1}{2}\exp(-x)$ for all $x > 0$ and to $\frac{1}{2}\exp(x)$ for all $x < 0$. Thus,

$$G(x) = \int_{\mathbb{R}} K(x-z)dF(z) = \int_{-\infty}^x (1 - \frac{1}{2}e^{-x+z} \ dF(z) \ + \ \int_x^\infty \frac{1}{2}e^{x-z} \ dF(z)$$

and

$$\begin{aligned}
-g'(x) &= -\frac{d}{dx}\Big[\int_{-\infty}^x (1/2) \ e^{-x+z}dF(z) \ + \ \int_x^\infty (1/2) \ e^{x-z}dF(z)\Big] \\
&= -\frac{d}{dx}\Big[(1/2)e^{-x}\int_{-\infty}^x e^z \ dF(z) \ + \ (1/2) \ e^x \int_x^\infty e^{-z} \ dF(z)\Big] \\
&= \frac{1}{2}e^{-x}\int_{-\infty}^x e^z dF(z) - \frac{1}{2}e^x \int_x^\infty e^{-z}dF(z) \ .
\end{aligned}$$

Hence, $G(x) - g'(x) = F(x)$. The above arguments can be found in, e.g., [2].

Example and Simulation

As an example we consider the exponential deconvolution, where we have the explicit inverse relation $F(x) = g(x) + G(x)$. Define the convex function

$$U(x) = \int_0^x F(y)dy = G(x) + \int_0^x G(y)dy \ .$$

As an estimation for $U(x)$ we define its empirical counterpart

$$\begin{aligned} U_n(x) &= G_n(x) + \int_0^x G_n(y)dy \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i \le x) \; + \; \frac{1}{n}\sum_{i=1}^n \int_0^x \mathbb{1}(Z_i \le y)dy \;, \end{aligned}$$

where G_n is the empirical sampling distribution. The function U_n is an increasing function that is linear between successive data points. At these points it has jumps of size $1/n$ and after each jump its the slope is increased by $1/n$. Clearly U_n is not differentiable.

One could consider the piecewise linear function that connects the points $(z_i, U_n(z_i))$. The derivative of that function equals to

$$\frac{1}{n}(i + \frac{1}{z_{i+1} - z_i})$$

for $x \in [z_i, z_{i+1}]$, $i = 1, \ldots, n$, and to $1/(nz_1)$ for $x < z_1$. In general this derivative will not be monotone. In this case, the isotonic inverse estimator $\hat{F}_n$ is defined as the right derivative of the greatest convex minorant of the function U_n. Here, $\hat{F}_n(0) = 0$, $\lim_{x\to\infty} \hat{F}_n(x) = 1$, $\hat{F}_n$ is monotone and right continuous.

We compute an estimate of the distribution function F based on standard exponential deconvolution. The true distribution is chosen as the standard exponential distribution.

References

[1] Geurt Jongbloed, (1999). *Inverse Problems in Statistics*. A lecture note on AIO-Course. Vrije Universiteit, Amsterdam.

[2] A.J. van Es and A.R. Kok, (1998). Simple kernel estimators for certain nonparametric deconvolution problems. *Statistics and Probability Letters*, **39**, 151–160.