Puzzling model choice for over-dispersed samples: a Bayesian approach using bibliometric data

Charles S. Bos [1]   Michel Lubrano[2]

December 2002

Preliminary version, do not quote without authors' permission

Abstract

The paper explores model selection in the presence of over-dispersion in the data. Over-dispersion can occur in different settings such as the analysis of financial data. In this paper a bivariate bibliometric data set is used which represents both the number of publications and the quality of the journals in which the publication appeared. The minimum Hellinger distance estimator of Beran (1977) is used to fit the data using a whole range of probability densities going from the Generalised Beta II to the Pareto distribution.

The classical approach to model selection, using the asymptotic distribution of the Hellinger adequation test, suffers from the large number of observations, leading to a rejection of most cases (Lindley's paradox). The Bayesian approach developed in the paper solves the question. It appears that in the presence of over-dispersion the two parameter Weibull model delivers the best explanation of the data. When the coefficient of variation is below one, the Weibull is no longer the best model and more versatile distributions such as the four parameter Generalised Beta II provide a better fit.

JEL Classification: C11, C14, C51
Keywords: Bayesian inference, Model choice, Hellinger distance, bibliometry

[1]Nuffield College, Oxford University and Dept. of Econometrics, *Vrije* Universiteit *Amsterdam*; email:cbos@feweb.vu.nl.

[2]GREQAM-CNRS, Centre de la Vieille Charité, 2 rue de la Charité, F-13002 Marseille, France; email:lubrano@ehess.cnrs-mrs.fr.

Support of the European Economic Association through contract "Ranking Economic Departments in Europe" is gratefully acknowledged. While remaining responsible for any error, the authors wish to thank Costin Protopopescu for useful remarks and suggestions.

1 Introduction

What is over-dispersion:

- Cox (1983), Fitzmaurice (1997): the observed variance is greater than that predicted by the model (exponential, Poisson or Binomial). For model choice, over-dispersion favours models with more parameters than necessary.

- Over-dispersion when the sample coefficient of variation is greater than one. No reference to a particular model. Duration data in finance, Bauwens and Giot (2000). What are the consequences on model selection?

Bibliometric data :

- Lotka (1926), analyse the statistical pattern of publication data. Lotka's law: $\log(a_n) = \log(a_1) - \alpha \log(n)$. Implicit Pareto assumption and large implied over-dispersion (eventually infinite).

- The Pareto assumption is very restrictive (infinite moments, a single shape parameter).

Purpose of the paper: model selection in a Bayesian framework accounting for over-dispersion. Density matching approach proposed in Lubrano (2000). Check the ability of the selected model to reproduce sample moments.

2 Data characteristics and available densities

Publication data in the field of economics collected from the JEL CDROM and covering the period 1991-2000.

2.1 Descriptive statistics

- number of papers for each author : n_i

- quality index : p_i. Involves the number of co-authors and the quality of the journal

- Bivariate data set: discrete and continuous

$$[n_i \quad , \quad p_i = \sum_{j=1}^{n_i} v_j / \sqrt{co_j}]$$

- modelling p_i and $p_i | n_i > z$

Comments:

- Large but varying over-dispersion

- Identical pattern (last column)

- Different moments

Table 1: Conditional sample moments of p_i

Region	N	Mean	C.V.	Skew.	Kurt.	$\Pr(n_i \geq z)$
			$z = 1$			
California	3419	9.57	1.78	3.96	23.60	1.00
France	2698	5.42	2.38	7.70	86.01	1.00
Netherlands	1793	7.35	1.91	5.45	46.59	1.00
UK	6656	6.78	1.69	4.51	32.05	1.00
			$z = 2$			
California	1380	19.86	1.16	2.55	11.44	0.40
France	1101	10.96	1.71	5.21	40.17	0.41
Netherlands	778	14.29	1.34	3.86	24.74	0.43
UK	2920	12.88	1.18	3.20	17.58	0.44
			$z = 4$			
California	696	32.53	0.82	1.90	7.62	0.20
France	452	20.88	1.25	3.59	20.03	0.17
Netherlands	335	25.81	0.95	2.85	14.65	0.19
UK	1292	22.20	0.85	2.37	10.88	0.19

N is the sample size and $\Pr(n_i \geq z)$ indicates the proportion of the sample which is kept after truncation over z. C.V. is the coefficient of variation defined as the standard deviation divided by the sample mean. Skewness and kurtosis follow.

2.2 The generalised Beta-II and related densities

GB2 of McDonald (1984): encompasses most of the distributions with positive support including the Pareto. Four-parameter density :

$$f_{GB2}(y|a,b,p,q) = \frac{ay^{ap-1}}{b^{ap}B(p,q)[1+(y/b)^a]^{p+q}} \tag{1}$$

The h^{th} moment exists provided $aq > h$.

The Burr XII density: simple restriction of the GB2, with $p = 1$:

$$\begin{aligned} f_{B12}(x|a,b,q) &= f_{GB2}(x|a,b,p=1,q) \\ &= b^{-a}qax^{a-1}[1+(x/b)^a]^{-(q+1)} \end{aligned}$$

The h^{th} moment exists provided $aq > h$.

The Generalised Gamma density: limit of the GB2 when $q \to \infty$:

$$\begin{aligned} f_{GG}(y|a,\beta,p) &= \lim_{q\to\infty} f_{GB2}(y|a, b=\beta q^{1/a}, p, q) \\ &= \frac{a y^{ap-1}}{\beta^{ap}\Gamma(p)} \exp[-(y/\beta)^a] \end{aligned} \tag{2}$$

No parametric restrictions for moments. Gamma density corresponds to $a = 1$.

The Weibull density: a restriction of the Generalised Gamma for $p = 1$:

$$f_W(y|a,\beta) = f_{GG}(y|a,\beta,p=1) = \frac{a}{\beta}(y/\beta)^{a-1}\exp(-(y/\beta)^a) \tag{3}$$

No parametric restrictions for moments.

The Pareto density: a restriction of the GB2 with $a = b = p = 1$, plus a change of variable

$$\begin{aligned} f_{P2}(y|\alpha) &= f_{GB2}(y|a=1, b=1, p=1, q=\alpha) \\ &= \alpha[1+y]^{-(\alpha+1)} \\ f_P(x|\alpha,s) &= \alpha s^{\alpha} x^{-(\alpha+1)} \mathbb{1}(x>s) \quad x/s = 1+y \end{aligned} \tag{4}$$

The h^{th} moment exists provided $\alpha > h$.

3 Classical inference

3.1 Traditional methods

- Minimization of least squares or χ^2 distances for grouped data.
- The MLE is untractable with raw data for GG. Johnson, Kotz and Balakrishnan (1994), Tsionnas (2001). Cohen and Whitten (1988) propose a modified method of moment.
- Cowell (2000) prefers two parameters densities (Lognormal, gamma).

3.2 Minimum Hellinger distance estimator: Beran (1977)

The Hellinger distance between two densities f and g is an L^2 norm defined as

$$D_H(f,g) = ||f^{1/2} - g^{1/2}||_2 = \left[\int (f^{1/2} - g^{1/2})^2 dx\right]^{1/2}$$

A minimum Hellinger distance estimator of θ is a minimiser of $D^2(f_\theta, \hat{f}_n)$

$$\hat{\theta}_n = \text{Argmin}||f_\theta^{1/2} - \hat{f}_n^{1/2}||_2^2$$

Asymptotic equivalence with the MLE. Under some regularity conditions

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, V^{-1}) \qquad \text{with } V = 4 * \int \frac{\partial f^{1/2}}{\partial \theta} \frac{\partial f^{1/2}}{\partial \theta}'$$

Goodness-of-fit statistic asymptotically distributed as a N(0,1)

$$[nh_n D^2(f_{\hat{\theta}}, \hat{f}_n) - \mu_n]/\sigma_n \xrightarrow{\mathcal{L}} \mathrm{N}(0,1)$$

Defining R_n as the maximum interval between observations

$$\mu_n = 0.25 R_n ||K||^2 \qquad \sigma_n^2 = 0.125 h_n R_n ||K * K||^2$$

where $*$ denotes convolution. Simpson (1989) : Hellinger deviance test to test for restrictions on the parameter space:

$$d_n = 4n[D^2(f_{\theta_0}, \hat{f}_n) - D^2(f_\theta, \hat{f}_n)] \xrightarrow{\mathcal{L}} \chi^2(p-q)$$

independent of the window size.

3.3 The choice of the kernel and of the window size

A Gaussian kernel would give a positive mass to potentially negative observations. A truncated Gaussian kernel to deal with positivity

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{K\left((x - X_i)/h\right)}{1 - F\left(-X_i/h\right)}$$

where K is a Gaussian kernel and F the cumulative Gaussian. Optimal window width h_n minimizes the approximate mean integrated squared error (AMISE) for

$$h = c\,\sigma(x)\,n^{-1/5} \qquad c = 0.9 \text{ optimal when overdispersion instead of } 1.06$$

4 Classical empirical results

- Second moments fail to exist for the GB2, Burr XII and Pareto.
- Estimators sensitive to the range of integration of the Hellinger distance

Table 2: Minimum Hellinger distance estimates

	Fra	UK	Net	Cal
		GB2		
a	2.66 [0.31]	1.24 [0.089]	2.43 [0.40]	1.96 [0.23]
b	5.09 [0.26]	7.95 [0.70]	6.31 [0.43]	7.56 [0.46]
p	0.40 [0.0.54]	1.00 [0.097]	0.43 [0.088]	0.54 [0.78]
q	0.85 [0.14]	2.19 [0.30]	0.79 [0.18]	0.93 [0.16]
Hel	0.0299	0.0256	0.0254	0.0291
		Burr XII		
a	1.30 [0.027]	1.24 [0.016]	1.29 [0.034]	1.24 [0.026]
b	7.96 [0.73]	7.94 [0.39]	8.54 [0.86]	9.48 [0.72]
q	2.73 [0.23]	2.18 [0.083]	2.09 [0.19]	1.90 [0.12]
Hel	0.0318	0.0256	0.0266	0.0296
		GG		
a	0.58 [0.036]	0.36 [0.022]	0.38 [0.041]	0.27 [0.029]
β	0.58 [0.19]	0.028 [0.018]	0.049 [0.53]	0.0012 [0.0020]
p	2.92 [0.34]	6.13 [0.74]	5.75 [1.22]	10.04 [2.15]
Hel	0.0358	0.0270	0.0305	0.0.0322
		Weibull		
a	1.02 [0.012]	0.91 [0.0064]	0.94 [0.013]	0.86 [0.0084]
b	4.53 [0.065]	5.75 [0.059]	6.43 [0.13]	8.07 [0.12]
Hel	0.0418	0.0365	0.0401	0.0424
		Pareto		
a	0.59 [0.0068]	0.57 [0.0041]	0.560 [0.0074]	0.490 [0.0050]
Hel	0.181	0.177	0.219	0.227

Standard deviations are between brackets. Hellinger distance at the optimum.

Table 3: Beran goodness of fit statistics for the GB2

	France	UK	Netherlands	California
$z = 1$	26.77	51.55	18.58	46.22
$z = 4$	2.44	5.40	1.61	2.35

The 1% critical value of the N(0,1) is 2.326.

Is the GB2 a valid starting point?

- From Table 3, the GB2 does not fit to any of the four countries. Lindley's paradox because of large sample sizes.
- On truncated and shorter samples, the GB2 fit for three countries. UK does not fit because it has the largest sample size.

Table 4: Simpson deviance tests for complete samples

	France	UK	Netherlands	California
Burr XII	35.88	3.34	5.01	1.35
Gen Gam	97.78	16.06	26.42	25.43
Weibull	188.21	209.87	81.26	140.13
Pareto	2166.00	5200.00	1674.00	3243.00

1% critical values: $\chi^2(1) = 6.64$, $\chi^2(2) = 9.12$, $\chi^2(3) = 11.34$

Admissible reductions

- From Table 4,The Burr XII is a valid restriction for three countries out of four.
- All other models are rejected.
- From Table 5 (truncated samples), the Burr XII and the GG are never rejected at 1%.
- The Weibull is rejected only once.
- The Pareto is always rejected.

Problems with large samples.

Table 5: Simpson deviance tests for truncated samples

	France	UK	Netherlands	California
Burr XII	0.19	4.55	0.48	2.26
Gen Gam	0.20	0.17	0.10	0.29
Weibull	4.05	12.99	0.09	0.27
Pareto	433	1696	457	1128

1% critical values: $\chi^2(1) = 6.64$, $\chi^2(2) = 9.12$, $\chi^2(3) = 11.34$

5 Bayesian model selection: the posterior predictive approach

The posterior predictive approach compares actual data y_d (a sample of size T) to hypothetical replicated samples y_{rep} (each of size T) obtained by simulating the model. Let us define a statistics $g(y_d)$ and compute a tail area probability

$$\Pr\{g(y) \geq g(y_d)|M\}$$

A small tail area probability casts doubts on the model. This probability is computed in reference to the posterior predictive distribution attached to model M Rubin (1984)

$$p(y|M) = \int f(y|\eta, M)\, \varphi(\eta|y_d, M)\, d\eta \tag{5}$$

and then to average over the posterior distribution of η. This is the posterior predictive approach. Useful to check if the model reproduces correctly the over-dispersion. Problems as for some models, the existence of moments is subject to restrictions.

Lubrano (2000) advocates that it is more fruitful to compare directly a non-parametric estimate of the sample density to the posterior predictive density using the square of the Hellinger distance. For each draw of the MCMC integration

$$D_H^2(\theta_i) = 2(1 - \int \sqrt{\hat{f}_n(x)\, f_{\theta_i}(x)}\, dx) \tag{6}$$

Model selection by comparing posterior Hellinger distances.

6 Bayesian inference

Bayesian inference is made difficult because of the tricky shape of the likelihood function for three models: GB2, Burr XII and GG. Iterated Metropolis plus a Gaussian informative prior centred on the classical estimate for these three models. Diffuse prior for the other models: Weibull and Pareto.

Table 6: Bayesian posterior moments

	Fra	UK	Net	Cal
	GB2			
a	3.06 [0.07]	1.46 [0.020]	1.61 [0.092]	1.23 [0.035]
b	1.59 [0.07]	3.35 [0.14]	5.017 [0.41]	6.69 [0.38]
p	0.48 [0.012]	0.90 [0.021]	0.69 [0.025]	0.87 [0.020]
q	0.33 [0.016]	1.00 [0.043]	1.28 [0.075]	1.32 [0.047]
Hel	0.109	0.0535	0.0450	0.0586
	Burr XII			
a	1.55 [0.015]	1.41 [0.010]	1.52 [0.020]	1.45 [0.016]
b	1.75 [0.081]	3.08 [0.13]	2.46 [1.17]	2.76 [0.16]
q	0.83 [0.032]	1.69 [0.12]	0.82 [0.044]	0.78 [0.033]
Hel	0.0958	0.0566	0.0807	0.0866
	GG			
a	0.48 [0.0056]	0.34 [0.0022]	0.35 [0.0034]	0.26 [0.0013]
b	0.26 [0.017]	0.019 [0.0014]	0.026 [0.0025]	0.0006 [0.000]
p	3.37 [0.046]	6.35 [0.071]	6.07 [0.091]	10.18 [0.10]
Hel	0.0629	0.0524	0.0623	0.0711
	Weibull			
a	0.76 [0.012]	0.81 [0.009]	0.77 [0.014]	0.74 [0.010]
b	4.28 [0.14]	5.86 [0.13]	6.04 [0.21]	7.58 [0.20]
Hel	0.0380	0.0135	0.0108	0.0111
	Pareto			
a	0.71 [0.014]	0.59 [0.007]	0.58 [0.014]	0.53 [0.010]
Hel	0.243	0.211	0.250	0.251

Posterior standard deviations are indicated between brackets. The Hel line gives the posterior mean of the Hellinger distance.

6.1 Posterior results for overdispersed samples

- Posterior expectations rather different from classical estimates despite the prior information.
- Posterior standard deviation greater than their classical counterparts, except for the GG where they are much smaller.

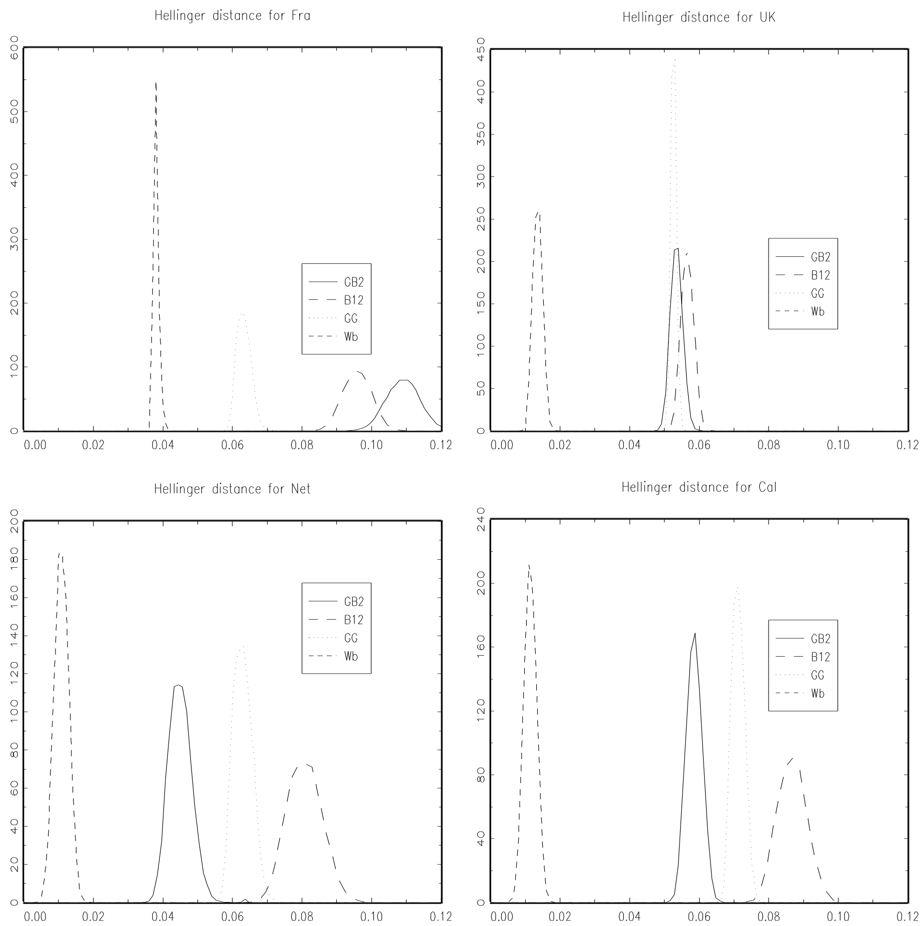

Figure 1: Posterior density of the square Hellinger distance

Model selection

- Richer models do not necessarily displays a smaller Hellinger distance.
- With over-dispersed samples, the two parameter Weibull model becomes the preferred model.
- The only common point between classical and Bayesian results is that the Pareto model is all the time rejected. (Graph not given because it is outside of the scale).
- All the models reproduce correctly the ordering between the empirical means. This is essential for stochastic dominance analysis. But difficulties to reproduce the absolute value of the sample mean.

What is the effect of over-dispersion for model selection

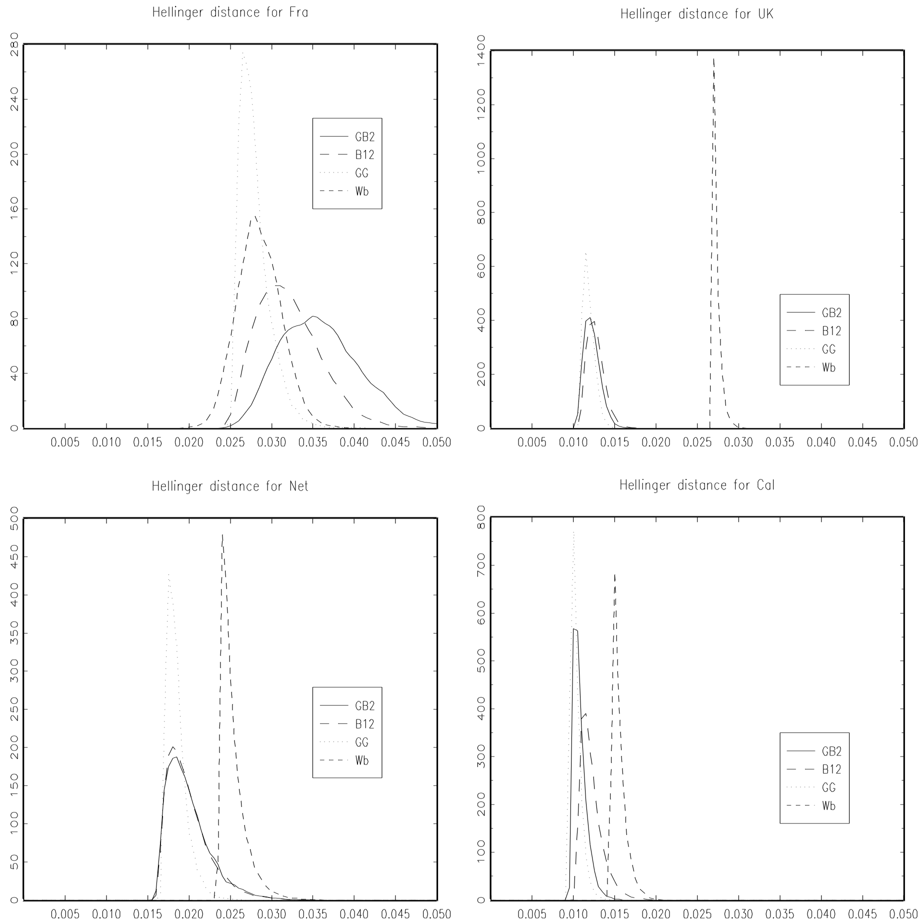

Figure 2: Posterior density of the square Hellinger distance
for truncated samples

- The Weibull is no longer the best fitting density and is dominated by the other models.
- The GB2, GG and Burr12 appear essentially equivalent with perhaps a slight advantage for the GG.
- All models manage to reproduce correctly the absolute value of the sample mean.
- The GB2 and the Burr12 reproduce correctly the coefficient of variation (which now exists).
- The GG and the Weibull do not reproduce correctly the coefficient of variation.

7 Conclusion

- Fitting a density is not an easy task in the presence of over-dispersion. Difficult to use the likelihood function.
- The classical selection procedure is impeded by the large sample sizes
- Classical and Bayesian selection procedure do not give the same answers
- The Weibull model appears to have nice properties with over-dispersion (in a Bayesian setting)
- The Weibull model is no longer selected with the truncated sample which has no over-dispersion

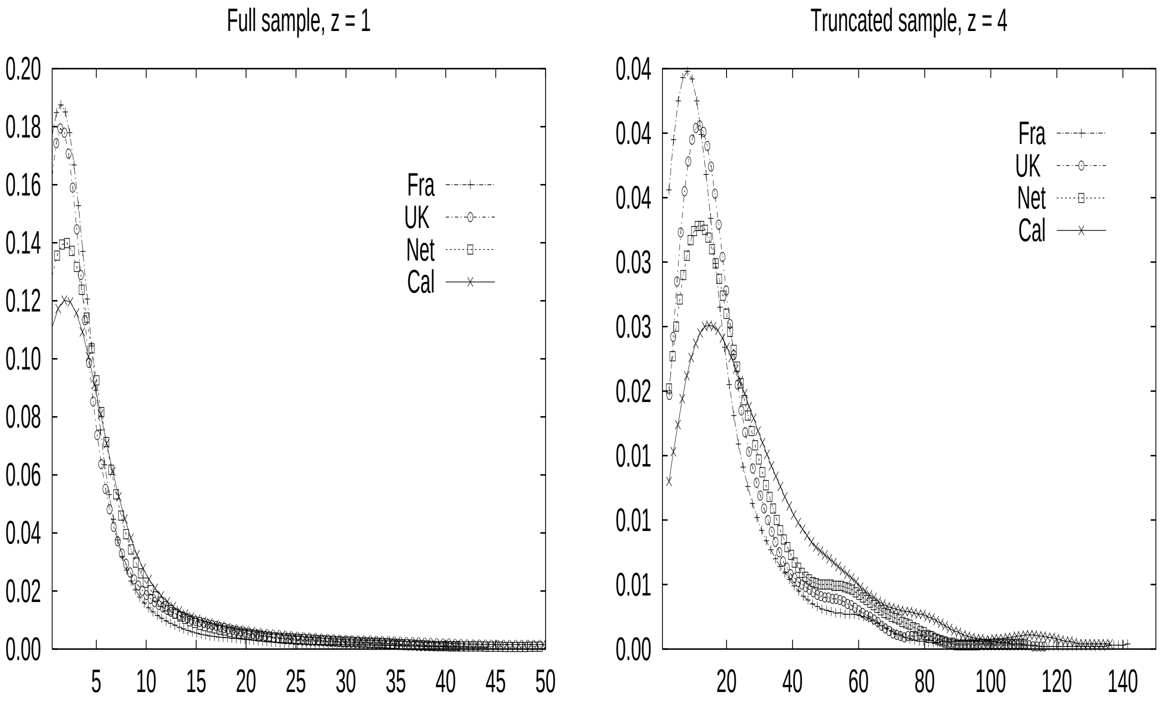

Figure 3: Non parametric density estimate
using a truncated Gaussian kernel