Puzzling model choice for overdispersed samples: a Bayesian approach using bibliometric data

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

September 2002

Preliminary version, do not quote without authors' permission

Abstract

The paper explores model selection in the presence of overdispersion in the data. Overdispersion 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 overdispersion 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

Over-dispersion may have at least two meanings in the statistical literature. Cox (1983) and Fitzmaurice (1997) for instance consider univariate sampling processes such as exponential, Poisson or Binomial which predict a strict relation between mean and variance. There is over-dispersion in the data when the observed variance is greater than that predicted by the model. This may be due to correlated sampling or clustering. When model choice is considered, over-dispersion favours models with more parameters than necessary. Fitzmaurice (1997) proposes to estimate the degree of over-dispersion in a sample and to use this estimation for correcting Akaike and Bayesian information criteria.

In this paper, we shall not consider over-dispersion by reference to a given model. A sample presents over-dispersion when its empirical standard deviation is greater than its empirical mean or in other terms when the coefficient of variation is greater than one. In the case of a positive random variable, over-dispersion also means that the empirical density is skewed with most of the probability for low values of the observations. Over-dispersion can occur in different settings. In the financial domain, Bauwens and Giot (2000) analyse the duration between two successive transactions using high frequency data. A large fraction financial transactions occur at relatively short intervals of time, but a significant fraction of the transaction are subject to long delays and this creates overdispersion in the durations. Bauwens and Giot (2000) check the ability of their model to reproduce this type of behaviour.

In a totaly different domain, some authors, starting with Lotka (1926), analyse the statistical pattern of publication data. A large number of scientific authors have a small number of papers, while few authors publish most of the papers in their related domain. Lotka (1926) plotted on a logarithmic paper the number of authors against the number of their respective contributions in the field of chemistry and physics. He found a straight line with a negative slope of $\alpha = 2$ and proposed his famous empirical law. According to Lotka's law, the number a_n of authors publishing n papers over a given period is given by

$$\log(a_n) = \log(a_1) - \alpha \log(n)$$

This empirical law implies that on average 60% of the authors publish only one article. It has received much attention in the bibliometric literature, see for instance the references given in the paper of Cox and Chung (1991). Several theoretical justifications have been given to Lotka's law, see the references given in Cox and Chung (1991), but to our knowledge, it was not noticed until Lubrano and Protopopescu (2002) that the Pareto distribution (or more precisely, the discrete Zypf distribution, often used to model the occurrence of words in a text) is behind. The implicit Pareto assumption implies that the coefficient of variation is greater than 1 if the shape parameter α lies between 2 and 2.41.

The purpose of this paper is to analyse model selection when the data present possibly some degree of over-dispersion and publication data will be used as our reference case. The above mention Lotka's law implicitly assume that the data follow a Pareto distribution. The most immediate question is whether this density is not too restrictive, and if a more versatile distribution could help to represent the publication data. If so, how could such a density be selected. The approach used in this paper is Bayesian in its spirit. It draws on the density matching approach proposed by Lubrano (2000).

2 Data characteristics and available densities

Modern bibliometric research no longer proceeds by simply counting the number of publications of an author, but tries to measure his performance by taking into account the quality of the journal where his paper is published, and also the number of co-authors involved. An index is thus constructed which acts as a proxy for an underlying continuous and positive random variable. We therefore have a bivariate data set: a discrete component which counts for each author the number n of published articles, and an approximately continuous component p, the performance index. We are interested in analysing the marginal distribution of p.

The data set was earlier analysed in Lubrano and Protopopescu (2002). It concerns publication data in the field of economics collected from the JEL CDROM and covering the period 1991-2000. It concerns the state of California and the countries France, the Netherlands and the United Kingdom.

2.1 Descriptive statistics

For each country, we have a list of authors having published at least one article in one of the referenced journals. For each author, we know what papers where published in which journal, and how many co-authors were involved. From this data we construct an index which takes into account the quality of the journal where each paper is published and the number of co-authors as follows: If i is the index of the authors ($i = 1, 2, ...$), an element of the bivariate data set is

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

where n_i is the number of papers an author i has published over the period, v_j is a numerical quality index at value in [1,10] indicating the quality of the journal where a paper is published and co_j the number of coauthors. Lotka (1928) was interested in modelling the distribution of n_i. We are interested in modelling the marginal distribution of p_i and eventually the conditional distribution of $(p_i|n_i \geq z)$ where z is a minimum level of publications (whatever their ranking) over the period.

In Table 1, the main descriptive statics of our four samples are presented. These figures call for several comments. The distribution of the p_i is over-dispersed for all four regions, with large skewness and kurtosis. Between 56 and 60% of the authors have published less than 2 papers, which is in accordance with Lotka's law. In this respect, all four samples

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.

display identical publishing patterns, but looking at the conditional moments of the data the samples are quite different. There are large differences in conditional mean, variance, skewness and kurtosis. Only after excluding authors with less than 4 papers in their name, do we obtain the more general under- dispersion. In the last part of this paper we therefore choose to focus on the distribution of $p_i|n_i \geq 4$. This will provide a coherent point of comparison to asses the impact of over-dispersion on model choice.

We can conclude that Lotka's law, which relies extensively on the Pareto distribution, corresponds to one of the characteristics of the samples, but ignores others like moment conditions. Would there be a different distributional assumption which would be more in line with the data?

2.2 The generalised Beta-II and related densities

The generalised Beta-II distribution (GB2) of McDonald (1984) is one of the most general distributions which have been proposed in the literature for fitting income data. Its main characteristic is that it encompasses most of the distributions with positive support that can be found in the literature, including the Pareto. The expression of this four-parameter density is:

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

with $y > 0$ and where $B(p,q)$ is the Beta function. The parameters a, b, p and q have to be positive[1]. The analytical expression of the h^{th} un-centred moments is

$$\begin{aligned} \mathrm{E}(Y^h) = \mu'_h &= b^h \frac{B(p+h/a, q-h/a)}{B(p,q)} \\ &= b^h \frac{\Gamma(p+h/a)\Gamma(q-h/a)}{\Gamma(p)\Gamma(q)} \end{aligned} \tag{2}$$

where $\Gamma(p)$ is the Gamma function. The h^{th} moment exists provided $aq > h$. Note that b is a scale parameter which does not influence the moment ratios. McDonald (1984) uses the hypergeometric function to give the expression of the cumulative distribution. However, it may appear convenient to exploit the relation between the hypergeometric function and the incomplete Beta function. The incomplete Beta function is defined as

$$IB_{x,p,q} = \int_0^x \frac{u^{p-1}(1-u)^{q-1}}{B(p,q)}$$

From Abramowitz and Stegun (1972, p. 263), we have

$$I_{x,p,q} = \frac{x^p}{pB(p,q)} \times {}_2F_1(p, 1-q; p+1; x)$$

so that the expression given in McDonald (1984) for the cumulative distribution can be simplified to

$$F(y; a, b, p, q) = IB\left(\frac{(y/b)^a}{1+(y/b)^a}, p, q\right).$$

A procedure for the numerical evaluation of the incomplete Beta function is available in most standard programming languages while that of the hypergeometric function usually is not.

McDonald (1984) presents in a graph all the possible reductions of this family of densities. Four densities will be useful for us: the Burr XII, the Generalised Gamma (GG), the Weibull and the Pareto density.

The Burr XII density, from Burr (1942) (also named the Singh-Maddala (1976) density function) is obtained as as simple restriction of the GB2, with $p = 1$:

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

[1] In the most general writing, there is no restriction on the sign of a because it appears with an absolute value in the numerator of f_{GB2}. But in all the empirical applications we saw, a was positive, so we finally adopted this more simple notation of the density.

Its cumulative distribution has the analytical form:

$$F_{B12}(x) = 1 - \frac{1}{(1 + (x/b)^a)^q}$$

and its moments are

$$\mathrm{E}(X^h) = b^h \frac{\Gamma(1 + h/a)\Gamma(q - h/a)}{\Gamma(q)}$$

The moment of order h exists provided $a \times q > h$. This distribution has been widely studied in the statistical literature, see e.g. Rodriguez (1977) for an interesting discussion.

The Generalised Gamma density is obtained as a limit of the GB2 when $q \to \infty$:

$$\begin{aligned} f_{GG}(y|a, \beta, p) &= \lim_{q \to \infty} f_{GB2}(y|a, b = \beta q^{1/a}, p, q) \\ &= \frac{a y^{ap-1}}{\beta^{ap}\Gamma(p)} \exp[-(y/\beta)^a] \end{aligned} \tag{3}$$

The cumulative distribution function is

$$F(x) = G[(x/\beta)^a; p] \tag{4}$$

where G is the incomplete Gamma function

$$G(x; p) = \frac{1}{\Gamma(p)} \int_0^x t^{p-1} e^{-t} dt \tag{5}$$

The un-centred moments are easily derived using the properties of the gamma distribution[2]

$$\mathrm{E}[Y^h] = \beta^h \frac{\Gamma(p + h/a)}{\Gamma(p)} \tag{6}$$

There are no parametric restrictions for the existence of the moments, except the positivity restrictions mentioned before.

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

$$f_W(y|a, \beta) = f_{GG}(y|a, \beta, p = 1) = \frac{a}{\beta}(y/\beta)^{a-1} \exp(-(y/\beta)^a) \tag{7}$$

with a and β positive. The cumulative distribution function is

$$F_W(x) = 1 - \exp(-(y/\beta)^a) \tag{8}$$

[2] An alternative way to define the Generalised Gamma density is to start from a random variable Z with a standard gamma distribution and make the change of variable $Y = \beta Z^{1/a}$ with Jacobian $(a/\beta)(y/\beta)^{a-1}$.

The un-centred moments reduce to

$$\mathrm{E}(Y^h) = \beta^h \Gamma(1 + h/a). \tag{9}$$

There are no constraints on the parameters to guarantee the existence of the moments, except positivity constraints.

The Pareto density corresponds to a completely different line of restricting the GB2 as now q is finite and becomes the sole parameter while $a = b = p = 1$

$$\begin{aligned} f_{P2}(y|\alpha) &= f_{GB2}(y|a = 1, b = 1, p = 1, q = \alpha) \\ &= \alpha[1 + y]^{-(\alpha+1)} \end{aligned} \tag{10}$$

This is the Pareto of the second kind (or Lomax distribution). The usual Pareto is obtained after a change in variable $x/s = 1 + y$ (with Jacobian $1/s$) so that

$$f(x) = \alpha s^\alpha x^{-(\alpha+1)} \mathbb{I}(x > \sigma) \tag{11}$$

The cumulative distribution function is

$$F(x) = 1 - (s/x)^\alpha. \tag{12}$$

The un-centred moments are

$$\mathrm{E}(X^h) = s^h \frac{\alpha}{\alpha - h} \qquad \text{if } \alpha > h \tag{13}$$

They are obtained as a particular case of the moments of GB2 using the previous change of variable and a recursion over h, or, even simpler, they can be computed directly from the expression of the probability density. The moment of order h exists if and only if $\alpha > h$.

3 Classical inference

3.1 Traditional methods

Singh and Maddala (1976) suggest estimating the parameters of the Burr XII by minimizing the least squares distance between a natural estimate of the distribution function and its analytical expression, by means of a nonlinear least squares procedure. They use grouped data. The estimation of Lotka's law relies on the same principle. However, the minimization of a least squares distance instead of a χ^2 distance leads to biaised etimtes as noted by McDonald and Ransom (1979).

McDonald and Ransom (1979) consider explicitly the case of grouped data. For k groups, the empirical probability of each group is given by n_i/n with $\sum n_i = n$. If I_i is the interval length of group i and $f(x)$ the density function, then the probability of group

i is given by $p_i = \int_{I_i} f(x)dx$. The Pearson minimum chi-squared estimator is obtained by minimizing the chi-squared distance

$$n \sum_{i=1}^{k} \frac{(n_i/n - p_i)^2}{p_i}$$

The chi-squared distance is distributed as $\chi^2(k-p-1)$ where p is the dimension of the parameter space. It can be used to test the adequacy of the proposed density with respect to the sample at hand.

Raw data contains more information than grouped data, but the application of the maximum likelihood principle is in this case impeded by severe numerical difficulties. We have reviewed five possible distributions and will have to select one of them: GB2, Burr XII, GG, Weibull and Pareto. While the one-parameter Pareto and the two-parameter Weibull models do not cause many problems for inference on raw data, this is not the case for the richer distributions. The econometric literature is not very talkative on this question. For instance Singh and Maddala (1976) McDonald (1984) or McDonald and Mantrala (1995) all report results based on grouped data and do not mention the problems that appear when one tries to maximise directly a likelihood function based on the use of raw data. We can just note that Cowell (2000) suggests that two parameter densities may be preferable for ease of interpretation and estimation.

There is on the contrary abundant statistical literature concerning the numerical difficulties applying the maximum likelihood method on individual observations with probability densities with more than two parameters. For instance, Johnson, Kotz and Balakrishnan (1994) report that the maximum likelihood estimator may fail to exist for the Generalised Gamma (GG) distribution. Cohen and Whitten (1988) do not recommend the GG for modelling purposes unless there are enough observations to group data in a frequency table. These difficulties can be expected to occur similarly with the GB2 distribution and the Burr XII, as they are generalisation of the GG.

Cowell (2000) suggests using a method of moments estimator, choosing the vector of parameters θ such that the theoretical moments of the distribution are as close as possible to their empirical counterparts. However, for a four-parameter distribution this entails using moments up to the 4th order which cannot be estimated precisely when the size of the sample is limited. Moreover, higher order moments may not exist for some of the parameter configurations.

Cowell (2000) also suggests using the robust M-estimator of Huber (1972). We would like to build on this idea to propose a numerically stable estimator which is valid for raw data. The basic idea is simple. Some of the estimators proposed for grouped data are based on the minimisation of a particular distance between an empirical histogram and the theoretical probability of each cell. We would like to replace the empirical histogram by a kernel estimate of the empirical density and minimise a distance between the kernel estimate and the theoretical density. The minimum Hellinger distance estimator of Beran (1977) is one solution to this problem.

3.2 Minimum Hellinger distance estimators

For a given density f, the L^2 norm is defined by

$$||f||_2 = \left[\int f^2 dx\right]^{1/2}$$

where the integral is taken with respect to the dominating measure dx. The Hellinger distance between two densities f and g is an L^2 norm defined as

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

This distance is invariant with respect to the choice of the dominating measure. Simple calculations show that the square of the Hellinger distance can be written as

$$D_H^2(f,g) = 2[1 - \rho(f,g)] \qquad \text{with } \rho(f,g) = \int \sqrt{f \times g}\, dx$$

The quantity $\rho(f,g)$ is called the affinity between f and g and is equal to 1 when f and g assign the same probability to every measurable set. It is equal to zero when f and g are mutually singular.

Suppose that we have n independent observations $X_1, \cdots, X_n$ supposedly coming from a common density f_θ with $\theta \in R^p$. A minimum Hellinger distance estimator of θ is a minimiser of $D^2(f_\theta, \hat{f}_n)$

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

or equivalently the value maximising $\rho(f_\theta, \hat{f}_n)$. In both cases f_n is a nonparametric kernel estimate of the density f.

This estimator, first proposed by Beran (1977) is asymptotically equivalent to the maximum likelihood estimator. We have under some regularity conditions

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, V^{-1})$$

with

$$V = 4 * \int \frac{\partial f^{1/2}}{\partial \theta} \frac{\partial f^{1/2\prime}}{\partial \theta}$$

V is in fact the information matrix associated with f.

Beran (1977) considers $D^2(f_{\hat{\theta}}, \hat{f}_n)$ evaluated at the optimum as a goodness-of-fit statistic. When it is appropriately centred and scaled, this statistic is asymptotically distributed as a N(0,1) like many non-parametric goodness-of-fit statistics (see e.g. Pagan and Ullah 1999). More formally, if K is the kernel smoother, h_n the window length, let us define R_n as the maximum interval between observations, and

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

where $*$ denotes convolution. Then

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

This statistic is a function of both the sample size n and of the window size h.

Following these first results, Simpson (1989) develops a Hellinger deviance test which allows to test for restrictions on the parameter space. Remember that $\theta \in R^p$ and let θ_0 be a restriction of θ which belongs to a proper subset of R^p of dimension $q < p$. Then

$$d_n = 4n[D^2(f_{\theta_0}, \hat{f}_n) - D^2(f_\theta, \hat{f}_n)]$$

is distributed as a $\chi^2(p-q)$and is independent of the window size h.

3.3 The choice of the kernel and of the window size

The shape of the distribution of our four samples is not standard. As we noted previously, there is systematic over-dispersion in the data. A coarse histogram displays an L shaped density, whereas a finer grid suggests an increasing, then a decreasing graph, but with a density which is never zero at the smallest observation. A traditional Gaussian kernel is appropriate for an unbounded support and may be not so appropriate here. Using a Gaussian kernel would give a positive mass to potentially negative observations. To escape such endpoints effects, a truncated Gaussian kernel can be used, with truncation at zero. The corresponding estimator of the data density would be

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

where K is a Gaussian kernel and F the cumulative Gaussian.

In order to find the optimal window width h_n, the usual practise consists in minimizing the approximate mean integrated squared error (AMISE) a formula of which is given by (see e.g. Pagan and Ullah 1999):

$$AMISE = \frac{1}{4}\lambda_1 h^4 + \frac{\lambda_2}{nh}$$

where

$$\lambda_1 = (\int \psi^2 K(\psi)\, d\psi)^2 \int (f''(x))^2\, dx \qquad \lambda_2 = \int K^2(\psi)\, d\psi$$

The bias and the variance of the kernel smoother are of the same order of magnitude if $h \propto n^{-1/5}$. The AMISE is minimized for $h = cn^{-1/5}$ where

$$c = (\lambda_2/\lambda_1)^{-1/5}$$

When both the kernel and the density to estimate are normal, then the optimal $c = 1.06$. When both the kernel and the density are truncated Gaussians, it can be shown that the value of c does not change. **Give the case for the gamma density.** Let us now suppose that the true density is a Pareto with parameter a. For different plausible values of a, we get the corresponding values of c in Table 2. So for the value of a given by Lotka's law, the

Table 2: Optimal window size
for a Pareto density

α	0.2	0.5	1.0	1.5	2
c	0.38	0.53	0.65	0.79	1.08

smoothing parameter is very similar to the normal case. But, anticipating on forthcoming results, the estimated value of α will be well below 2, which suggests much lower values for c. If the true distribution is an iso-dispersed exponential, then $c = 0.98$. Over-dispersion would mean lower values for c.

In order to compute the Beran (1977) test, we have to compute $||K||^2$ and $||K * K||^2$ for the truncated normal. In the case of the normal distribution, the convolution is defined as

$$K(t) * K(t) = \int_{-\infty}^{\infty} K(t-r)K(r)\,dr$$

So

$$\int K(t-r)K(r)dr = 0.5/\sqrt{\pi}\exp(-0.25t^2)$$

and

$$\int (K * K)^2 = 0.25\sqrt{2}/\sqrt{\pi} = 0.19947$$

and

$$\int K^2(x)\ dx = 0.5/\sqrt{\pi} = 0.28209$$

The truncated Gaussian kernel is a Gaussian kernel normalised for the range $[0, \infty]$:

$$K(t) = \frac{2}{\sqrt{2\pi}}\exp(-t^2/2)$$

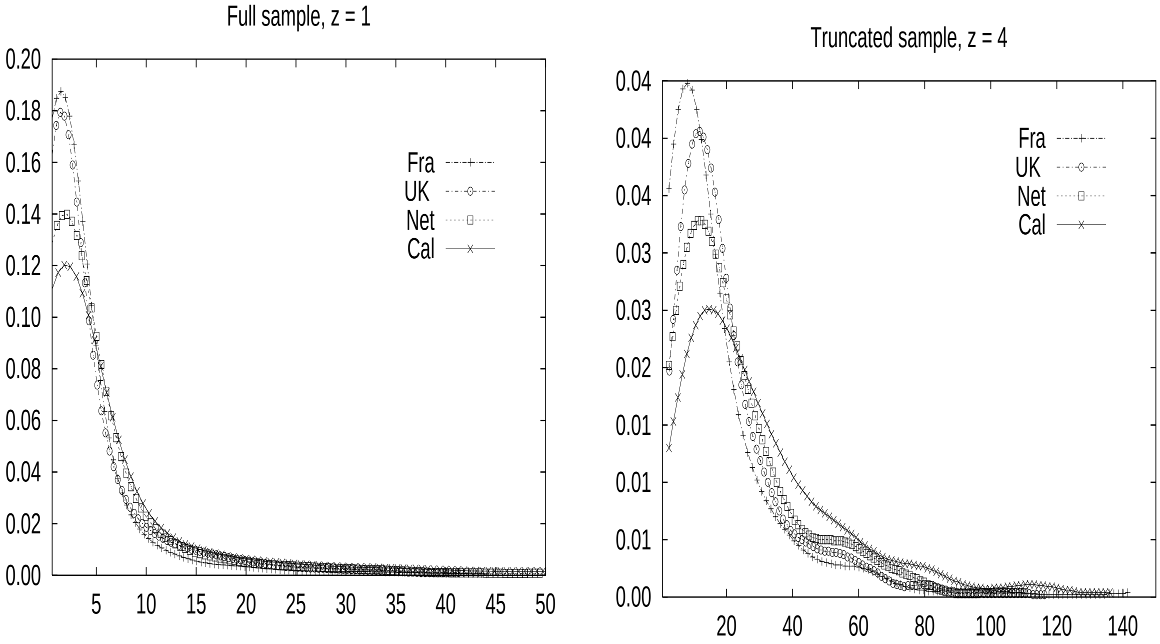
so that $||K||^2$ becomes

$$\int_0^{\infty} K^2\,dr = \frac{1}{\sqrt{\pi}} = 0.31837$$

It is in a way natural to choose the same range of integration for the convolution

$$\int_0^{\infty} (K * K)^2\,dr = 2\frac{\sqrt{2}}{\sqrt{\pi}} = 1.59576$$

As an illustration, we give in Figure 2 the graph of the non-parametric estimate of the density for the four regions, both including all other (left panel) and excluding authors with 3 or less publications (right panel). The window size is computed with $c = 0.9$.



The right tail is not fully displayed.

Figure 1: Non parametric density estimate
using a truncated Gaussian kernel

4 First classical empirical results

The practise reported in Simpson (1989) consists in evaluating the integral present in the definition of the Hellinger distance by means of a Simpson rule with equidistant points. Simpson (1989) recommends integrating beyond the range of the observations namely using $[y_{min} - h, y_{max} + h]$. However, in Simpson (1989), the theoretical range is $[-\infty, \infty]$. Here we have a very particular sample configuration which is illustrated both in Table 1 and Figure 2: over-dispersion and a very long right tail. Beside being very long, this right tail is also extremely thin and hard to estimate non-parametrically. Therefore we decided to limit our analysis to the range where the non-parametric density estimate was no smaller than $1/250\times$ the maximum of the density. Otherwise numerical errors appear. Truncation points together with observed maxima are reported in Table 3. For every sample the

Table 3: Right truncation points

	Fra	UK	Net	Cal
Truncation	40	56	64	83
Maximum	215	159	187	190

observed minimum is 0.577. But there is some kind of censoring in the observations. First, papers with more than 3 authors are reported as one author papers. Second, potential authors, those having published 0 papers are ignored. Finally, the grading scale for journals starts at 1. Very low quality journals which could have been graded below 1 are ignored. All these facts are an incentive to start the integration range below the observed minimum. We have fixed it at 0.001. When we started the integration at the observed minimum, some of the parameters of the GB2 went to infinity, or convergence was not achieved. In short, starting at 0.001 produces a better fit. However, the Pareto distribution must be treated differently, because the support of the distribution is not $[0, +\infty)$, but $[s, +\infty)$. A consistent estimator of s is the minimum of the observed sample and all moments are functions of s. So we started the integration at that point. We used 501 equidistant points and a Simpson rule. We selected for the complete samples a window size c equal to 0.9.

There is not much to say about the estimates reported in Table 4. They have relatively low standard errors, except for the scale parameter of the GG. Lotka's law is not verified on this data set as the reported values for α are well below 2 in the Pareto case. Recall however that these data do not represent the number of publications but only weighted index. For the GB2, the mean always exists, but the variance may fail to exist for The Netherlands and California, and skewness and kurtosis exist for none of the regions. The situation is slightly better for the Burr XII. For the Pareto, no moments exist at all.

We can now use the adequacy test of Beran (1977) to check if the GB2 is a valid starting point for representing the data. From Table 5, the GB2 does not fit to any of the four reported samples despite the fact that the reported squared Hellinger distances are very

Table 4: 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. The Hel line gives the square of the Hellinger distance between the NP estimate of the density and the parametric density at the optimal θ.

low and in accordance with the values reported in the original paper of Beran. This is due to the large sample sizes (from 1793 for the Netherlands to 6656 for the UK) we have here and is thus taken as a manifestation of Lindley's paradox. When the data are truncated with $z = 4$, the size of the samples is greatly reduced (to 335 for The Netherlands and up

Table 5: 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.

to 1292 for the UK). The adequacy of the GB2 is not rejected for three out of 4 regions at the 1% level. GB2 does not fit the UK data which has by far the greatest sample size.

If we decide to accept the GB2 as a valid starting point, the χ^2 test of Simpson indicates if one of the proposed reductions is valid. The Burr XII and Generalised Gamma correspond

Table 6: 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: $\chi^2(1) = 6.64$, $\chi^2(2) = 9.12$, $\chi^2(3) = 11.34$

to one restriction, the Weibull to two restrictions and the Pareto to three restrictions. Table 6 indicates that the Burr XII is valid in 3 regions. All the other distributions (in particular the Pareto) are rejected. When we exclude authors having published less than 4 papers, it is far more easy to find parsimonious parametric densities to fit the data. Table 7 indicates

Table 7: 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: $\chi^2(1) = 6.64$, $\chi^2(2) = 9.12$, $\chi^2(3) = 11.34$

that the Burr XII is never rejected at 1%. The Generalised Gamma is also very successful while the Weibull is rejected only once. The Pareto continues to be rejected all the time.

The use of individual data together with rather complicated distributions makes it necessary to use a non-standard estimation technique. The fact that the data are overdispersed implies that these techniques have to be adapted: a particular range of integration to compute the Hellinger distance and a truncated Gaussian kernel.

The goodness-of-fit tests are biased by the very large sample sizes. The restriction tests of Simpson (1989) on the contrary seem to be less sensitive to this problem. However, with a reduced sample size obtained by excluding less productive authors, seems necessary to be able to select a parsimonious distribution. In the rest of the paper, we show how the Bayesian approach may give some answers to these questions.

5 Bayesian model selection: the posterior predictive approach

One of the many possible approaches to model selection in a Bayesian framework relies on the old idea of comparing actual data y_d (a sample of size T) to hypothetical replicated samples y_{rep} (each of size T) obtained by simulating the model. If the model under inspection fits the data in a satisfactory way, the observed sample and the simulated samples should not look too different. To apply the idea, it is common to define a statistics $g(y_d)$ and to compute a tail area probability

$$\Pr\{g(y) \geq g(y_d)|M\} \tag{14}$$

which quantifies the extremeness of the observed value $g(y_d)$ with reference to the distribution of the transformation $g(.)$ of the random variable y which represents the possible outcomes of the model. A small tail area probability casts doubts on the model. For Box (1980), this probability is computed in reference to the predictive distribution attached to model M

$$p(y|M) = \int f(y|\eta, M)\, \varphi(\eta|M)\, d\eta \tag{15}$$

which represents all the possible outcomes of the data y if model M (data density and prior distribution) were true. In this writing, $f(y|\eta, M)$ is the data density, η the vector of parameters and $\varphi(\eta|M)$ the prior density. Rubin (1984) develops the posterior predictive approach arguing that if in some situations it may be interesting to operate unconditional frequency checks, in other cases it may be preferable "to regard η as a fixed feature of the replications" and then to average over the posterior distribution of η. This is the posterior predictive approach. Rubin (1984) says "posterior meaning conditional on observed values and predictive meaning the distribution of a future observable quantity". This last approach knew a significant development in the statistical literature with Meng (1994), Gelman and Meng (1996) or Gelman, Meng and Stern (1996). With the posterior predictive approach, interest is focused on the fit of the model to the observed data. Whereas with the pure predictive approach, concern is stressed on reliability of the model for every possible outcome, at the cost of specifying an informative prior for η. See Gelman, Meng and Stern (1996) for more details concerning the comparison of the prior predictive approach and the posterior predictive approach.

The posterior predictive approach is particularly convenient in a MCMC setting. Suppose that we have already obtained N draws from the posterior density. Given a draw η_i, we draw an hypothetical sample y_i^{rep} from the sampling distribution. We compute the $p-$value as the proportion of cases in which the simulated transformation $g(y_i^{rep})$ exceed the realised value $g(y_d)$

$$\frac{1}{N}\sum_{i=1}^{N} ID[g(y_i^{rep}) > g(y_d)] \tag{16}$$

where ID is the indicator function.

5.1 Moment matching

The function $g(.)$ is simple to select if we are interested in the ability of the model to reproduce correctly some of the sample moments. The mean comes first. The coefficient of variation adds to that information some indication on the shape of the density as well as the skewness and kurtosis. However, this approach may be limited. For instance Rubin (1984) notes that the first two moments of two densities may be equal while the gap between order statistics may still be different. Consequently if $g(.)$ is the empirical mean or variance, the method will not reject the normal model when the true sampling model is Cauchy.

The method is also handicapped by the restrictions imposed on the parameters to insure existence of moments. Some densities such as the Generalised Gamma or the Weibull always have all their moments. This is not the case for the GB2, the Burr XII and the Pareto.

5.2 Measuring distances between densities

Lubrano (2000) advocates that it is more fruitful to compare directly a non-parametric estimate of the sample density to the posterior predictive density. The ϕ divergence is popular in the Bayesian literature, especially the Kullback-Leibler divergence and the χ^2 distance to compare densities. However both are asymmetric. If f and g are two densities to be compared, the distance between f and g is not the same as that between g and f when measured with the above two instruments. On the contrary, half the square of the Hellinger distance defined in the previous sections is a symmetric divergence. If $\hat{f}_n$ is the nonparametric estimate and f_θ the candidate density, then for each value of θ, say θ_i, we get a point

$$D_H^2(\theta_i) = (1 - \int \sqrt{\hat{f}_n(x)\, f_{\theta_i}(x)}\, dx) \tag{17}$$

corresponding to a draw in the posterior Hellinger distance between the model density and the non parametric estimate of the data density. For N draw of θ, we get N draws of D_H^2 which can be used to compute several quantities. We can decide for instance to select the model producing the smallest average distance if we compute the mean of the distribution. Or we can compute the probabilities that two models have the same fit.

6 Bayesian inference

To our knowledge, there does not exist a Bayesian treatment for inference with the generalised Beta-II. Tsionas (2001) has proposed a Bayesian analysis for the generalised gamma. He introduces a reparameterisation of the density which provides an analytical expression for the conditional posterior density of the scale parameter. He then builds a Gibbs sampler algorithm. In Tsionas (2000) a similar approach is proposed for the Weibull density. However this reparameterisation does not appear to be so convenient as it does not manage to kill the high negative correlation between a and p. Consequently, the Gibbs sampler may easily remain stuck is a corner of the parameter space. So we prefer to use the original parameterisation given in the previous section and a Metropolis algorithm to integrate the posterior density.

6.1 Posterior results for overdispersed samples

We used an independent Metropolis with a Student importance function centred on the minimum Hellinger distance estimator and with 4 degrees of freedom and an inflating factor of 1.2 for the variance. We iterated the method several times with 20 000 draws each time until convergence was achieved. Convergence was checked in three different ways: checking if the graph of the posterior was smooth and unimodal; using CUMSUM graphs; checking the stability of posterior moments between two iterations. Final posterior means and standard deviations are presented in Table 8. It also gives the posterior mean of the Hellinger distance.

There is extreme positive and negative correlation between the parameters for the GB2, the Burr XII and the GG. This correlation disappears for the Weibull model. That correlation destroys any possibility of getting sound results with a non informative prior. We had to use a Normal prior centred on the classical estimates to get convergence. On the contrary, a diffuse prior was possible with the remaining models.

In general, posterior standard deviation are greater than their classical counterparts, except for the GG where they are much smaller. There are marked differences between the posterior expectation of the parameters and their classical robust counterpart. These difference are explained by the fact that the classical procedure is robust, when the present Bayes' estimates rely on the likelihood function and are thus prone to the influence of outliers. We should develop a robust Bayesian approach to treat these data.

The graph of the posterior density of the Hellinger distance is particularly interesting to examine. As Bayesian inference is based on the likelihood principle, we no longer have the classical property that a richer model displays a smaller Hellinger distance. For all countries, the Hellinger distance corresponding to the Weibull model is smaller than that of the GB2 despite the fact that the Weibull is a restriction of the GB2.

When we compare these results to the classical ones given in Table 6, the only restriction which was accepted was the Burr XII (but it was rejected for France). All the other restrictions were rejected, in particular the Weibull, due to the large sample size. The only common point between classical and Bayesian results is that the Pareto model is all the

Table 8: 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. Results obtained with a Normal prior centred on the classical estimates for the GB2, B12 and GG to achieve convergence. Diffuse prior for the other models.

time rejected. We have not displayed the posterior density of the Hellinger distance for that model as it was outside the range of the presented graphs. We can conclude that the Bayesian approach solves the Lindley paradox for this case.

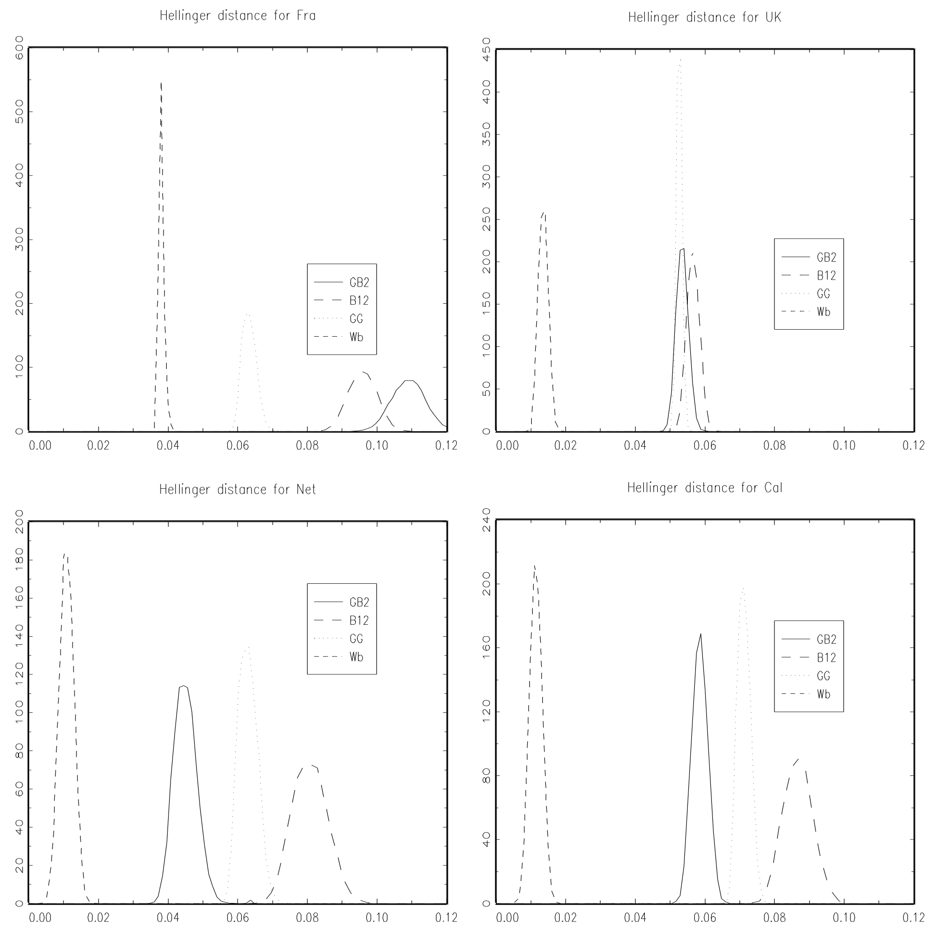


Figure 2: Posterior density of the square Hellinger distance

Let us now investigate the ability of the different models to reproduce the sample moments. We are going to evaluate the theoretical formulae given in section 2 for each draw of θ and see if the sample equivalent is contained in a given probability interval. From Table 9, it is evident that we cannot moments of order greater than one because for two models they simply fail to exist (three models if we include the Pareto for which even the mean does not exist). In most cases, all models managed to reproduce correctly the sample mean . However, we are on the border line for the B12 and France and UK and the Weibull for UK. A good point is that all models kept the ordering of the mean between countries. This point is essential if stochastic ordering is of interest.

Table 9: Posterior probability moment existence conditions

	Mean	Variance	Skewness	Kurtosis
France				
GB2	1.00	1.00	0.96	0.72
B12	1.00	0.81	0.97	0.73
UK				
GB2	1.00	0.00	0.00	0.00
B12	1.00	0.00	0.00	0.00
Netherlands				
GB2	1.00	0.00	0.00	0.00
B12	1.00	0.00	0.00	0.00
California				
GB2	1.00	0.00	0.00	0.00
B12	1.00	0.00	0.00	0.00

6.2 What is the effect of overdispersion for model selection

We now investigate the properties of the conditional distribution of $(p|z = 4)$ obtained by considering the samples where the authors with less than four publications have been excluded. We have seen that the main effect of considering these conditional distribution was to reduce greatly the coefficient of variation. It is still greater than one for France, but become lower than one for the other countries and goes down to 0.82 for California.

The case of France should be put aside, as there is still a slight overdispersion in the data. From Figure 3, the Weibull is still a good candidate, but the GG is equivalent. The other models are all very near. For the three other samples, Figure 3 indicates the 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. The classical approach concluded that all these models were equivalent, except for the Weibull with the UK data set.

All models manage to reproduce correctly the sample mean with this time no border line case. It is apparent from Table 10 that the variance exists in most cases (except for France) so it can be used as another testing device. For the GB2 and the Burr12, the sample coefficient of variation belongs to a 95% confidence interval of its posterior predictive equivalent. This is not the case for the GG and the Weibull, except the GG for California. As more moments exist, the procedure of moment matching allows to discriminate among models (elimination of GG and Weibull) while the density matching procedure eliminated only the Weibull in three cases out of four.

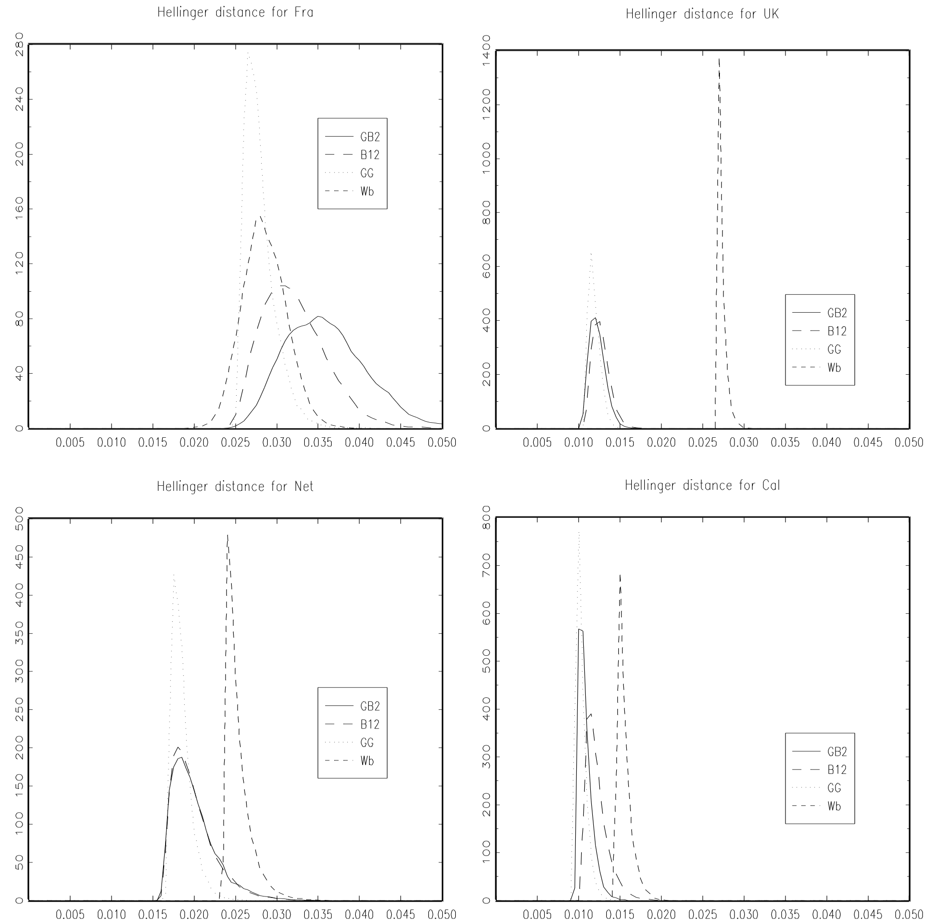


Figure 3: Posterior density of the square Hellinger distance for truncated samples

7 Conclusion

The fact that a sample presents over-dispersion is found to have severe consequences on model selection. In the bivariate sample concerning bibliometric data the amount of over-dispersion could be controlled by choosing to consider only authors with a certain minimum amount of publications.

It was found that the model with lowest Hellinger distance in the case of over-dispersion becomes the worse when an under-dispersed dataset concerning authors with at least 4 papers is considered. Over-dispersion makes inference delicate and a robust method of inference is needed. After a careful analysis, the Weibull density was found to fit the data well.

In a classical framework, it is difficult to maximise the likelihood function of a versatile

Table 10: Posterior probability moment existence conditions for truncated samples

	Mean	Variance	Skewness	Kurtosis
France				
GB2	1.00	0.06	0.00	0.00
B12	1.00	0.24	0.00	0.00
UK				
GB2	1.00	1.00	0.51	0.01
B12	1.00	1.00	0.09	0.00
Netherlands				
GB2	1.00	0.99	0.60	0.18
B12	1.00	1.00	0.55	0.11
California				
GB2	1.00	1.00	1.00	1.00
B12	1.00	1.00	0.99	0.68

density especially when there is over-dispersion. We had to use a robust method based on the minimum Hellinger distance. In a Bayesian framework, it appears easier (but more computationally demanding) to integrate over the parameter space. As often happens in such cases, the Metropolis algorithm was sensitive to the tuning of its parameters (e.g. the specification of the importance function, variance inflating factor). Finally, the Bayesian approach helps in solving the model selection problem even with large sample size, when the classical approach could not provide a satisfactory answer.

8 References

Abramowitz and Stegun (1972) *Handbook of Mathematical Functions.* Dover Publications, New-York.

Bauwens L. and P. Giot (2000) The logarithmic ACD model: an application to the bid/ask quote process of three NYSE stocks. *Annales d'Economie et de Statistique*, Vol. 60, 117-149.

Beran R. (1977) Minimum Hellinger distance estimators for parametric models. *The Annals of Statistics* 5, 445-463.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society* A 143, 383–430

Burr I.W. (1942) Cumulative frequency functions. *Annals of Mathematical Statistics* 13, 215-235.

Cohen A.C. and B.J. Whitten (1988) Modified moment estimation for the three parameter gamma distribution. *Journal of Quality Technology* 17, 147-154.

Cowell F.A. (2000) Measurement of inequality. In A. Atkinson and F. Bourguignon (2000) Income Distribution, North Holland, Amsterdam

Cox D.R. (1983) Some remarks on overdispersion. *Biometrika* 70 (1), 269-274.

Cox R.A.K. and K.H. Chung (1991) Pattern of research output and author concentration in the economic literature. *The Review of Economics and Statistics* 73(4), 740-747

Fitzmaurice G.M. (1997) Model selection with overdispersed data. *The Statistician* 46 (1), 81-91.

Gelman A. and Meng X.L. (1996) Model checking and model improvement. *Practical Markov Chain Monte Carlo* (W. Gilks, S. Richardson and S. Spiegelhalter, eds.). London: Chapman and Hall, 189–201.

Gelman A., Meng X.L. and Stern H. (1996) Posterior predictive assessment of model fitness via realised discrepancies. *Statistica Sinica* **6**, 733-807.

Huber P.J. (1972) Robust statistics: a review. *Annals of Mathematical Statistics* 43, 1041-1067.

Johnson N.L., S. Kotz and N. Balakrishnan (1994) *Continuous Univariate Distributions.* New-York, Wiley

Lotka A.J. (1926) The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16 (12), 317-323.

Lubrano M. and C. Protopopescu (2002) Density Inference for Ranking European Research Systems in the Field of Economics. Mimeo, GREQAM Marseille.

McDonald J.B. and M.R Ransom (1979) Functional forms, Estimation techniques and the distribution of income. *Econometrica* 47(6), 1513-1525.

McDonald J.B. (1984) Some generalised functions for the size distribution of income. *Econometrica* 52, 647-663.

McDonald J.B. and A. Mantrala (1995) The distribution of personal income revisited. *Journal of Applied Econometrics* 10, 201-204.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics* 22, 1142–1160.

Rodriguez R.N. (1977) A guide to the Burr type XII distributions. *Biometrika* 64, 129-134.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151–1172.

Simpson D.G. (1989) Hellinger deviance tests: efficiency, breakdown points and examples. *Journal of the American Statistical Association* 84, 107-113.

Singh S.K. and Maddala G.S. (1976). A function for size distribution of incomes. *Econometrica*, 44 (5), 963–970.

Tsionas E.G. (2001) Exact inference in four-parameter Generalized Gamma distributions. *Communications in Statistics (Theory and Methods)* 30 (4), 747-756.