Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 41 (2005) 1317–1329

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Symmetry and other transformation features of Lorenz/Leimkuhler representations of informetric data

Quentin L. Burrell *

Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB, UK

Received 27 September 2004; accepted 3 March 2005
Available online 3 May 2005

Abstract

In this paper we develop in particular the use of Lorenz/Leimkuhler concentration curves in an informetric context. Many of the features to be presented are akin to, or are adaptations of, ones that have featured in the econometric literature but not in informetrics. We acknowledge in particular our debt to Lambert [Lambert, P. J. (2001). *The distribution and redistribution of income*. Manchester: Manchester University Press] and Kleiber and Kotz [Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: Wiley] for source material in the econometrics literature. Although the development is purely theoretical, the aim is to provide additional and more incisive analytic tools for the practising informetrician.

1. Introduction

We will be concerned with the graphical representation of informetric data, in particular with relation to concentration aspects. It can be argued that Bradford (1934) provided the original spur to such graphical methods in bibliometric analysis in the presentation of his data sets with the aim of identifying a core journal collection. Certainly Bradford's approach is well represented in the literature and we will not pursue it here. Indeed, we will argue that it has been superseded by the methods described here. Another notable contribution was made by Trueswell (1966, 1969, 1976), in this case seeking to identify a core library collection, using graphical representations that are essentially equivalent to the Lorenz/Leimkuhler approach to be adopted here but predating its informetric application. See also Burrell (1985). There have been other

* Tel.: +44 1624824638; fax: +44 1624665095.
E-mail address: q.burrell@ibs.ac.im

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.03.016

recent examples of graphical approaches to the analysis of informetric data sets, such as Burrell (2002), but our concern here will be with the widespread representation via Lorenz/Leimkuhler curves of concentration.

Once again, let us note the general framework for stochastic modelling in informetrics, namely that of a "population of sources" producing "items" in some random fashion over time, typical examples being "papers/authors/journals receiving citations" and, in the academic library context, "books accumulating loans". In the terminology of Egghe (2004), these are examples of so-called two-dimensional informetrics studies. In this particular study we will not be concerned with the cumulative productivity of sources over time, simply with a set of sources that have produced numbers of items during a period of study and their graphical representation. (For theoretical and empirical examples of such time evolution, refer to Burrell, 1991, 1992b.) Later in this study we will also consider other sorts of informetric data but the above will be sufficient to motivate the approach.

In the field of economics, a simple way of illustrating inequalities in such as income and wealth distribution is via the so-called Lorenz curve of concentration. See, for instance, Lorenz (1905), Atkinson (1970), Lambert (2001) and Kleiber and Kotz (2003). In informetrics, where the interest usually focuses on the most productive sources, an equivalent graphical representation is the Leimkuhler curve; see Burrell (1991, 1992c). The difference between the two constructions is that for the Lorenz curve in economics one arranges the sources (individuals) in increasing order of productivity (income), while for the Leimkuhler curve they are arranged in decreasing order. In both cases we plot the cumulative proportion of total productivity against the cumulative proportion of sources. In this paper we will use these distinctions to differentiate between Lorenz and Leimkuhler curves. See Burrell (1991) for a demonstration of the equivalence between the two approaches and a justification for the attribution to Leimkuhler.

More formally, let X denote the number of items produced or, more generally, the *productivity*, of a randomly chosen source and suppose that the distribution of X in the population is given by the probability density function (*pdf*) $f(x)$. Note that although in the "source-item" formulation, X is a non-negative integer-valued variable, it is often useful to invoke a continuous approximation. Purely to simplify the discussion and some of the derivations, we will mostly restrict attention to the continuous formulation in what follows. (With a few exceptions, the carryover to the discrete set-up is straightforward.)

1.1. Notation/definitions

(i) $\mu = E[X] = \text{mean of } X = \int_0^\infty f(x)\,\mathrm{d}x$.

(ii) The *tail distribution function* of $X = \Phi(x) = P(X \geqslant x) = \displaystyle\int_x^\infty f(y)\,\mathrm{d}y.$ (1)

(iii) The *tail-moment distribution function* of $X = \Psi(x) = \left(\displaystyle\int_x^\infty yf(y)\,\mathrm{d}y\right)\bigg/\mu = \displaystyle\int_x^\infty g(y)\,\mathrm{d}y,$ (2)

where $g(y) = yf(y)/\mu$ is the *tail-moment density function*, and note that g is a legitimate pdf on the positive half-line.

[*Note:* The investigation of the relationship between the probability density functions f and g constitute what Egghe (2003) termed "type/token-taken" informetrics but as Burrell (2003) pointed out this is most naturally done via Leimkuhler curves, as we shall be doing in what follows.]

The Leimkuhler curve is then given by $\Psi = L(\Phi)$ which is the plot of Ψ on the vertical axis against Φ on the horizontal. Note that this is plotted via the implicit variable x. If X is continuous then, as x varies, $(\Phi(x), \Psi(x))$ determines a smooth curve. (In the discrete formulation, the "curve" is a piece-wise-continuous polygonal line.) In all cases, the curve lies within the unit square, is concave, starting at the origin and ending at $(1, 1)$.

Example 1 (*Exponential distribution*). The exponential distribution is a continuous distribution with a single (scale) parameter, $\lambda > 0$, and pdf $f(x) = \lambda e^{-\lambda x}$, $x > 0$. It is well known (see, e.g. Burrell, 1992c) that the Leimkuhler curve is given by

$$\Psi = L(\Phi) = \Phi[1 - \ln \Phi]. \tag{3}$$

Note that this is independent of the parameter λ. This is to be expected since, as noted, it is a scale parameter and it is well known that the Leimkuhler curve is scale invariant, i.e. if $Y = cX$ for some constant c then the Leimkuhler curves of X and Y coincide. Thus the exponential family has a single Leimkuhler representation, as in Fig. 1(a). To assist later considerations, also included are the *line of equality* $\Phi = \Psi$ and the *line of reflection/symmetry* $\Phi + \Psi = 1$.

[*Aside:* It is interesting to note that Basu (1992), in an investigation of Bradford's law, derives (3) as a Leimkuhler curve—but without identifying it as such—by a limiting argument concerning "random partition of the unit line", giving rise to a so-called random hierarchical distribution. See below.]

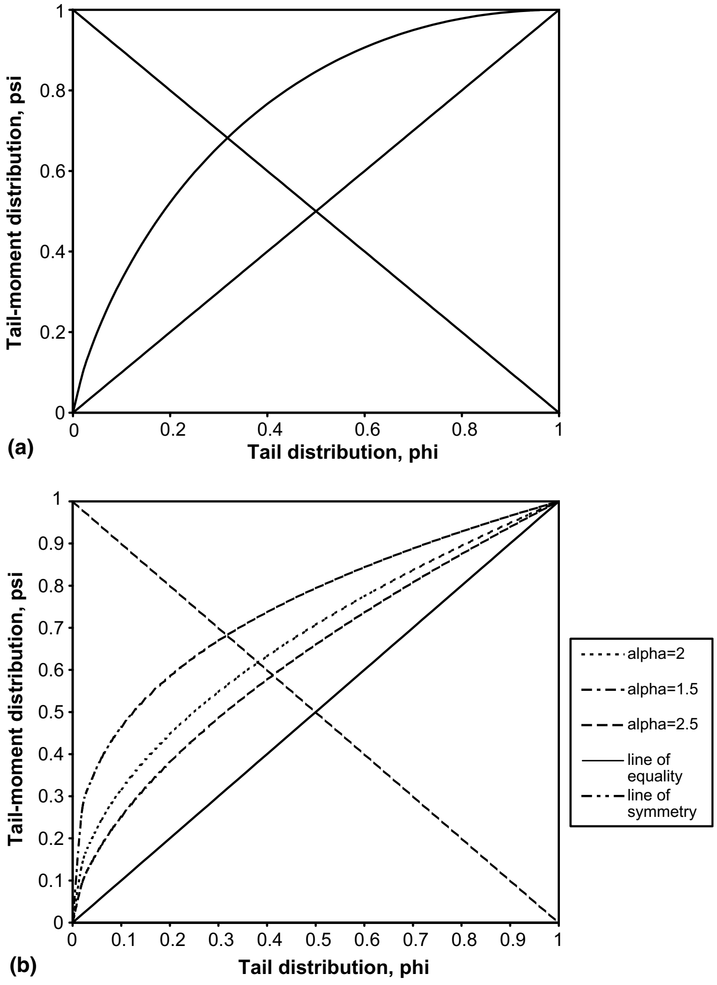

Fig. 1. (a) Leimkuhler curve, exponential, (b) Leimkuhler curve, Pareto.

Example 2 (*Pareto distribution*). In its simplest, original form the Pareto pdf can be written

$$f(x) = \frac{\alpha}{x^{1+\alpha}} \quad \text{for } x > 1.$$

Routine calculus (see, e.g. Burrell, 1992c) then gives the Leimkuhler curve as

$$\Psi = L(\Phi) = \Phi^{(\alpha-1)/\alpha}. \tag{4}$$

Note that here the curve does depend on the parameter $\alpha > 1$. This again should be expected since the Pareto parameter is not a scale parameter but a shape parameter. Hence each member of the Pareto family gives rise to a different Leimkuhler curve. For examples see Fig. 1(b). Again we have included the line of equality and the line of symmetry.

We have already remarked that the Leimkuhler curve is scale invariant. The converse is also true, i.e. if X and Y have the same Leimkuhler curves then their distributions are the same up to a change of scale, as can be seen from the following result, possibly originally demonstrated by Thompson (1976) but see also Arnold (1987).

Proposition 1. *The Leimkuhler curve determines the productivity distribution* (*up to a change of scale*).

Proof (*Adapted from* Lambert (2001)).

$$\begin{aligned} L'(\Phi) &= \frac{\mathrm{d}L}{\mathrm{d}\Phi} = \frac{\mathrm{d}[L(\Phi(x))]/\mathrm{d}x}{\mathrm{d}[\Phi(x)]/\mathrm{d}x} \text{ by the chain rule} \\ &= \frac{\mathrm{d}[\Psi(x)]/\mathrm{d}x}{\mathrm{d}[\Phi(x)]/\mathrm{d}x} = \frac{-xf(x)/\mu}{-f(x)} \text{ by differentiating (1) and (2)} \\ &= \frac{x}{\mu}, \end{aligned} \tag{5}$$

$$\begin{aligned} L''(\Phi) &= \frac{\mathrm{d}^2L}{\mathrm{d}\Phi^2} = \frac{\mathrm{d}^2\Psi}{\mathrm{d}\Phi^2} = \frac{\mathrm{d}}{\mathrm{d}\Phi}\left(\frac{\mathrm{d}\Psi}{\mathrm{d}\Phi}\right) = \frac{\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}\Psi}{\mathrm{d}\Phi}\right)}{\frac{\mathrm{d}\Phi}{\mathrm{d}x}} \text{ by the chain rule} \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x}{\mu}\right)}{-f(x)} = -\frac{1}{\mu f(x)} \text{ from (3) and (2).} \end{aligned}$$

The result follows.

(Actually, in the above we need to be careful about the support of the distribution and its end-points but such technical points need not concern us here. From the practical/intuitive point of view, the important thing is that productivity distributions and their Leimkuhler representations are essentially equivalent.)

Notes: (a) The striking—and delightful!—result of the Proposition is that it is the *shape* of the Leimkuhler curve, as summarised by its first two derivatives, that determines the distribution (up to scale) from which it derives.

(b) An interesting corollary of the result follows from (5) by noting that $L'(\Phi) = 1 \Longleftrightarrow x = \mu$, i.e. the Leimkuhler curve is parallel to the diagonal line of equality $\Phi = \Psi$ precisely at the mean of the distribution. Note also that at this point the curve is also at its greatest (perpendicular) distance from the line of equality $\Phi = \Psi$. □

Example 3. From the form of the Leimkuhler curve given by (3), we find

$$L'(\Phi) = [1 - \ln \Phi] + \Phi[-1/\Phi] = -\ln \Phi.$$

But from Proposition 1 this is equal to x/μ where $\mu > 0$ is the (arbitrary) mean. Hence we have $\Phi(x) = \mathrm{e}^{-x/\mu}$ and then $f(x) = -\Phi'(x) = \frac{1}{\mu}\mathrm{e}^{-x/\mu}$, i.e. an exponential distribution with parameter $1/\mu$.

The reader is invited to do the confirmatory calculations using the second derivative but it is interesting to note that in this example, the distribution can be recovered directly from knowledge of the slope of the Leimkuhler curve.

Example 4. To recover the distribution from the Leimkuhler curve of the Pareto distribution given in (4), we again use the Proposition. Note first that from (4) we find

$$L'(\Phi) = \frac{(\alpha - 1)}{\alpha}\Phi^{-1/\alpha}.$$

Now using the first part of the Proposition we find, after a little manipulation, that

$$\Phi = \Phi(x) = \left(\frac{\alpha}{(\alpha - 1)\mu}\right)^{\alpha} x^{-\alpha},$$

and then it follows that

$$f(x) = -\Phi'(x) = \left(\frac{\alpha}{(\alpha - 1)\mu}\right)^{\alpha} \frac{\alpha}{x^{\alpha+1}}.$$

The requirement that f is a pdf on $(1, \infty)$, so that its total integral is equal to 1, then fixes μ, the mean, as $\alpha/(\alpha - 1)$. Again we leave consideration of the second derivative to the reader.

Remark. Actually, the proof of the Proposition implies rather more than the Proposition itself states, namely that given any curve satisfying the requirements of a Leimkuhler curve does allow, at least in principle, a productivity distribution—in fact a scale-parameter family of distributions—which would give the required Leimkuhler curve. (On the other hand, it should be noted that, except in very special cases, even a simple functional form for the Leimkuhler curve does not necessarily allow the underlying form of the distribution to be recovered in an explicit or "elementary" form. We shall encounter this difficulty shortly.)

2. Power transformations

The above discussion suggests a simple method to generate "new productivity distributions from old" by choosing a suitable transformation of a known Leimkuhler curve. For instance:

Proposition 2. *Let $\Psi = L(\Phi)$ be a given Leimkuhler curve, then $L_\beta(\Phi) = L(\Phi)^\beta$ is also a Leimkuhler curve, provided*

$$0 < \beta < 1 - \frac{L \cdot L''}{(L')^2} \quad \forall 0 < \Phi < 1. \tag{6}$$

Proof

(i) Clearly, $L_\beta(0) = 0$, $L_\beta(1) = 1$, so the curve passes through the origin and $(1, 1)$.

(ii) $L'_\beta(\Phi) = \beta L(\Phi)^{(\beta-1)} L'(\Phi) > 0$, since L is increasing, and hence the curve is monotone increasing if $\beta > 0$.

(iii) $$\begin{aligned}L''_\beta(\Phi) &= \beta(\beta-1)L(\Phi)^{(\beta-2)}(L'(\Phi))^2 + \beta L(\Phi)^{(\beta-1)}L''(\Phi)\\ &= \beta L(\Phi)^{(\beta-2)}[(\beta-1)(L'(\Phi)^2) + L(\Phi)L''(\Phi)]\end{aligned}$$

For concavity, we require this second derivative to be negative. The term outside the brackets on the RHS is positive so, after a little algebra on the term within the brackets, we find the condition to be

$$\beta < 1 - \frac{L(\Phi)L''(\Phi)}{(L'(\Phi))^2}.$$

Combining this with the result of (ii) gives the required expression.

Note: If we want a general result, independent of the particular Leimkuhler curve, note that $\beta < 1$ always suffices. This follows since $L''(\Phi) < 0$ so that the upper allowable limit for β is always greater than 1. □

Example (*Pareto distribution*). As shown above, the general form for the Leimkuhler curve for a Pareto distribution is $\Psi = L(\Phi) = \Phi^{(\alpha-1)/\alpha}$. Thus the power transformation is

$$\Psi = L_\beta(\Phi) = L(\Phi)^\beta = [\Phi^{(\alpha-1)/\alpha}]^\beta = \Phi^{\beta(\alpha-1)/\alpha}.$$

This is a power function and hence is the Leimkuhler curve for a Pareto distribution of order γ if

$$(\gamma-1)/\gamma = \beta(\alpha-1)/\alpha < 1.$$

A little algebra shows that this requires firstly that

$$\gamma = \alpha/(\alpha - \alpha\beta + \beta),$$

and, since we must have $\gamma > 1$, then $\beta > 0$. The upper limit given by (6) reduces to $\beta < \alpha/(\alpha - 1)$.

Remarks. (a) Let us consider the same power transformation for an exponential distribution, i.e.

$$L_\beta(\Phi) = L(\Phi)^\beta = \Phi^\beta[1 - \ln\Phi]^\beta. \tag{7}$$

This Leimkuhler curve is not of an immediately recognisable form, indeed if we try to apply Proposition 1 to recover the underlying distribution we get a functional equation that cannot be solved by elementary means. All we can say is that it is a valid Leimkuhler curve for an appropriate range of values of β. To find this range, note that

$$L(\Phi) = \Phi[1-\ln\Phi], \quad L'(\Phi) = -\ln\Phi, \quad L''(\Phi) = -1/\Phi.$$

Using Proposition 2 we find that we must have, using (6), that

$$0 < \beta < 1 - \frac{L(\Phi)L''(\Phi)}{(L'(\Phi))^2} = 1 - \frac{\Phi[1-\ln\Phi](-1/\Phi)}{(-\ln\Phi)^2} = 1 + \frac{1-\ln\Phi}{(\ln\Phi)^2}.$$

The function on the right hand side is increasing in $\ln\Phi$ and hence approaches its minimum value as $\ln\Phi \to -\infty$. The minimum value is then 1 and we have that the power transformation is valid for the exponential distribution only for $0 < \beta < 1$.

(b) In a pair of papers that warrant further consideration, Basu (1992, 1995) introduced what she termed the random hierarchical distribution and its generalisation. The former is, as remarked earlier, nothing more than the exponential distribution expressed via its Leimkuhler curve representation. The latter, a two-parameter generalisation, is essentially a power transformation of a geometric distribution expressed in Leimkuhler terms, see Burrell (1985, 1992c), and this includes as a special case the power transformation of the exponential distribution as given by (3). Unfortunately, Basu did not realise that the choice of power is subject to the restrictions as derived above and includes a graphical example (Basu, 1995, Fig. 1(1)) with, in our notation, $\beta = 2$. It is clear that this is not valid since the resulting curve is not concave.

(c) The power transformation discussed above parallels a similar approach suggested by Sarabia, Castillo, and Slottje (1999) for generating new Lorenz curves via power (and other) transformations of given Lorenz curves. These ideas carry over and give new methods for generating Leimkuhler curves from known ones including the power transformation described above. It is interesting to note that although this sort of construction has recently received much attention in econometrics, a particular case had already been considered in informetrics by Basu (1992, 1995).

3. Leimkuhler ordering

If we have two Leimkuhler curves, $L(\Phi)$ and $L^*(\Phi)$ say, such that $L(\Phi) \geqslant L^*(\Phi)$ for all $0 \leqslant \Phi \leqslant 1$, we say that the first curve dominates the second. For instance, if we have two Pareto distributions then the one with the smaller parameter dominates the other (see Fig. 1(b)). (We will come across other families having this Leimkuhler dominance ordering property in later sections.) The Pareto family can be considered to be just a special case of a power transformation family. Starting with any Leimkuhler curve, $L(\Phi)$, we have that for any (allowable) power transformation $L^\beta(\Phi)$, either L or L^β is dominant depending on whether $\beta > 1$ or $\beta < 1$.

Note that this idea of one curve dominating another does not determine an ordering of all Leimkuhler curves since the curves for two different distributions can intersect, for instance the exponential and Pareto (see Fig. 1(a) and (b)).

Another sort of transformation that can be considered is that of the original variable X. We have the following:

Theorem (Jakobsson–Fellman). *Let g be a real function which is non-negative and monotone increasing and let $Y = g(X)$ be such that X and Y both have finite means. Then if $g(x)/x$ is monotone on the range of X, the Leimkuhler curve of Y dominates that of X, or is dominated by that of X according as $g(x)/x$ is increasing or decreasing.*

This result is originally due Fellman (1976) although in the economics literature it is often referred to as the Jakobsson–Fellman Theorem in deference to Jacobsson (1976). See also Arnold (1987). In informetrics it was first used by Rousseau (1992a) and an elementary proof was given by Burrell (1992a). It has recently been used by Egghe (2004) in a study of what is sometimes termed three-dimensional informetrics (Egghe & Rousseau, 1990) but has also previously been referred to as "linked" informetric processes; see Rousseau (1992a) and Burrell (1992d). Simple and natural examples for g are power, exponential and logarithmic functions; see Burrell (1992a). It would seem that there are several lines of enquiry that could usefully be pursued in this area.

4. "Self-symmetry" of the Leimkuhler curve

When looking at a Leimkuhler curve, the first consideration is how it compares with the 45°-degree line, i.e. the line $\Psi = \Phi$, since this is how it indicates "inequality" or concentration of productivity within the population of sources, for instance by means of the Gini coefficient which looks at the area between the Leimkuhler curve and the 45°-degree line. A different perspective is offered by consideration of the reflection/rotation of the Leimkuhler curve in/around the other diagonal $\Phi + \Psi = 1$. As an example, Fig. 2 shows the Leimkuhler curve of a Pareto distribution of index $\alpha = 1.5$ together with its reflection. We say that the curve is *self-symmetric* if the original curve coincides with its reflection. Clearly in this example we do not have the self-symmetric property. Consideration of the examples in Fig. 1(a) and (b) would suggest that the exponential and Pareto families all fail. In fact this can be easily demonstrated by using the following:

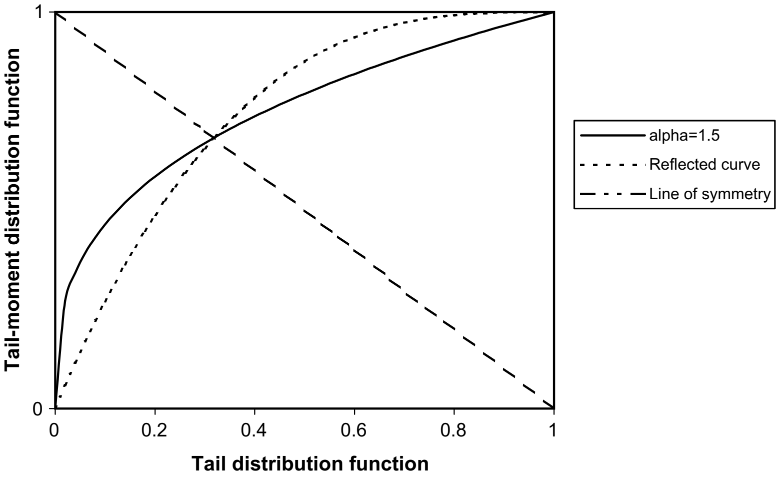

Fig. 2. Reflected Leimkuhler curve—Pareto distribution, $\alpha = 1.5$.

Lemma. *A necessary condition for self-symmetry is* $\Phi(\mu) + \Psi(\mu) = 1$.

Proof. Clearly for self-symmetry we must have that the curve is parallel to the main diagonal $\Phi = \Psi$ where the curve crosses the other diagonal $\Phi + \Psi = 1$. Thus we need $L'(\Phi) = 1$ when $\Phi + \Psi = 1$. But we have already seen that $L'(\Phi) = 1 \iff x = \mu$. Thus it is necessary that $\Phi(x) + \Psi(x) = 1$ at $x = \mu$, as required.

As an example, consider the exponential distribution with mean $\mu = 1$. Then (see Burrell, 1992c),

$$\Phi(x) + \Psi(x) = \mathrm{e}^{-x} + (1 + x)\mathrm{e}^{-x} = (2 + x)\mathrm{e}^{-x},$$

and when $x = \mu = 1$, RHS $= 3\mathrm{e}^{-1} > 1$, which confirms that the exponential family does not have the self-symmetry property. Similar calculations settle the case for the Pareto family, where we find $\Phi(\mu) + \Psi(\mu) < 1$. □

Remark. Damgaard and Weiner (2000) suggest using $S = \Phi(\mu) + \Psi(\mu)$ as an asymmetry coefficient for Lorenz/Leimkuhler curves. The basic idea uses the previously mentioned result that the slope of the curve equals 1 at $x = \mu$. If this occurs "above" the line $\Phi + \Psi = 1$ then this suggests concentration is towards the lower end of the distribution, if "below" then towards the upper end. (Note how this interpretation applies to the exponential and Pareto families.) Although this is an easily implemented approach, it is weakened by the fact that the authors assume that the condition $\Phi(\mu) + \Psi(\mu) = 1$ is both necessary and sufficient for self-symmetry, whereas it is easy to construct counter-examples to show that in fact it is not sufficient. For instance, take two different self-symmetric curves crossing the line of symmetry at the same point. From these, we can construct a new Leimkuhler curve that coincides with the first below the line of symmetry and with the second above the line of symmetry.

The formal requirement for self-symmetry is given by the following:

Definition. The Leimkuhler curve $\Psi = L(\Phi)$ is self-symmetric if and only if

$$L(1 - L(\Phi)) = 1 - \Phi \quad \text{for all } 0 \leqslant \Phi \leqslant 1. \tag{8}$$

This follows from the geometry of self-symmetry where reflection in the line $\Phi + \Psi = 1$ takes the point (Φ, Ψ) to the point $(1 - \Psi, 1 - \Phi)$.

5. Examples of self-symmetric distributions

5.1. Singh–Maddala distribution

The Singh–Maddala is a flexible three-parameter distribution that has been widely used in the economics literature after its first introduction by Singh and Maddala (1975, 1976). See also Kleiber and Kotz (2003). We believe that it may in future have useful applications in informetrics, particularly in the study of citation age data. The general pdf of the Singh–Maddala distribution is given by

$$f(x) = \frac{aqx^{a-1}}{b^a[1 + (x/b)^a]^{1+q}}, \quad \text{for } x > 0. \tag{9}$$

The parameter b is a scale parameter and, as we are here only concerned with the Leimkuhler curve, we can take $b = 1$. In the special case where, in addition, $q = (a + 1)/a$, the pdf reduces to $f(x) = \frac{(a+1)x^{a-1}}{(1+x^a)^{2+1/a}}$ and it is not too difficult to derive the Leimkuhler curve as

$$\Psi = L(\Phi) = 1 - [1 - \Phi^{a/(a+1)}]^{(a+1)/a} = 1 - [1 - \Phi^{1/q}]^q. \tag{10}$$

(This can also be derived as a special case of a family of curves derived by Rasche, Gaffney, Koo, and Obst (1980).)

Examples of the Leimkuhler curve as in (10) are given in Fig. 3. Note that this reduced Singh–Maddala family satisfies the Leimkuhler dominance ordering, with dominance increasing with q. Note also that these curves certainly appear self-symmetric. This is confirmed by the following:

Proposition 3. *The family of Leimkuhler curves given by (10) above are self-symmetric.*

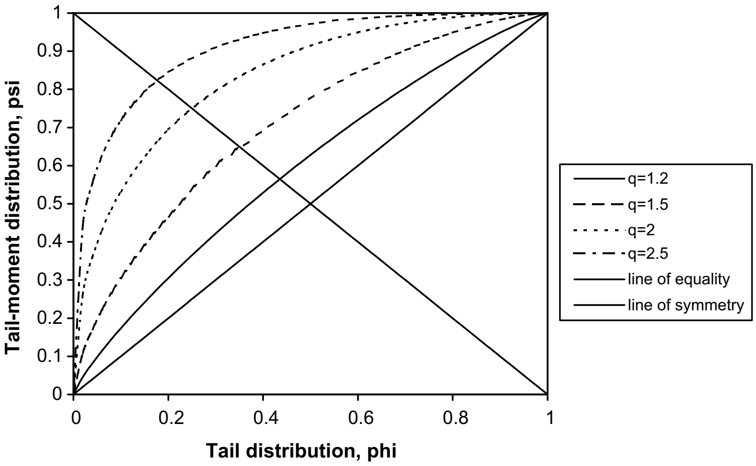

Fig. 3. Leimkuhler curve, Singh–Maddala.

Proof. We will use the definition of self-symmetry given by (8) so consider

$$\begin{aligned}L(1-L(\Phi)) &= L([1-\Phi^{1/q}]^q) \text{ making use of (10)}\\ &= 1-\{1-([1-\Phi^{1/q}]^q)^{1/q}\}^q \text{ again using (10)}\\ &= 1-\{1-[1-\Phi^{1/q}]\}^q\\ &= 1-\{\Phi^{1/q}\}^q = 1-\Phi.\end{aligned}$$

Hence the curve is self-symmetric according to (8). □

5.2. Lognormal distribution

The random variable X is said to have a lognormal distribution with parameters θ and σ if $\ln X$ has a Normal distribution with mean θ and variance σ^2. Thus the pdf of X is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp[-(\ln x-\theta)^2/2\sigma^2] \quad \text{for } x>0, \tag{11}$$

and we write $X \sim LN(\theta,\sigma^2)$. It is well known that the mean of this distribution is given by $E[X] = \mu = \exp(\theta + \sigma^2/2)$.

The lognormal distribution has been used in various ways in informetric modelling, most notably in citation age studies. See Egghe and Ravichandra Rao (1992), Gupta (1998) and Burrell (2002).

From its definition, we have that the tail-moment pdf is given by

$$\begin{aligned}g(x) = \frac{xf(x)}{\mu} &= \frac{x\left(\frac{1}{x\sigma\sqrt{2\pi}}\right)\exp\left[-(\ln x-\theta)^2/2\sigma^2\right]}{\exp(\theta+\sigma^2/2)}\\ &= \left(\frac{1}{x\sigma\sqrt{2\pi}}\right)\exp\left[\ln x-(\ln x-\theta)^2/2\sigma^2-\theta-\sigma^2/2\right]\\ &= \left(\frac{1}{x\sigma\sqrt{2\pi}}\right)\exp\left[-\left\{(\ln x)^2-2\ln x(\theta+\sigma^2)+(\theta+\sigma^4)\right\}/2\sigma^2\right]\\ &= \left(\frac{1}{x\sigma\sqrt{2\pi}}\right)\exp\left[-(\ln x-(\theta+\sigma^2))/2\sigma^2\right],\end{aligned}$$

i.e. the pdf of $LN(\theta + \sigma^2,\sigma^2)$ so that the mean is

$$\mu^* = \exp((\theta+\sigma^2)+\sigma^2/2) = \mu\exp(\sigma^2) = E[X]\exp(\sigma^2).$$

Remark. Although we have explicit expressions for the density functions, these do not lead to explicit formulae in terms of x for $\Phi(x)$ and $\Psi(x)$. Hence there is not a simple expression in closed form for the Leimkuhler curve. However, many readily available computer packages allow the computation of the cumulative distribution function—and hence the tail distribution function—of any lognormal distribution with specified parameter values. In particular, it can be shown that $\exp(\theta)$ is a scale parameter and so far as the Leimkuhler curve is concerned there is no loss of generality in taking it equal to 1, equivalently take $\theta = 0$. It is then straightforward to compute both $\Phi(x)$ and $\Psi(x)$ and hence plot the Leimkuhler curve. For examples, using the Lorenz formulation, see Kleiber and Kotz (2003, p. 116).

From such graphs it is apparent that the lognormal family satisfies the Leimkuhler dominance ordering, with increased dominance corresponding to increasing σ^2. It would also appear that the curves are self-symmetric, but in the absence of a closed form for the Leimkuhler curve, we cannot use the condition (8) of the definition to check this. Instead we make use of the following:

Theorem (Kendall). *A pdf f(x) gives rise to a self-symmetric Leimkuhler curve if and only if it can be expressed in the form*

$$f(x) \propto x^{-3/2} h\left(\ln \frac{x}{\mu}\right), \text{ where } h(y) \text{ is an even function of } y, \text{ i.e. } h(y) = h(-y).$$

Proof. See Kendall (1956).

To apply this to the lognormal case, we can write the pdf (11) as

$$\begin{aligned} f(x) &= \frac{1}{x\sigma\sqrt{2\pi}} \exp[-(\ln x - \theta)^2/2\sigma^2] \propto x^{-1} \exp[-(\ln x - \theta)^2/2\sigma^2] \\ &\propto x^{-3/2} \exp[\ln x^{1/2} - (\ln x - \theta)^2/2\sigma^2] \\ &\propto x^{-3/2} \exp[(\sigma^2 \ln x - (\ln x)^2 + 2\theta \ln x)/2\sigma^2] \\ &\propto x^{-3/2} \exp[-\{(\ln x)^2 - 2 \ln \mu \ln x\}/2\sigma^2] \\ &\propto x^{-3/2} \exp[-(\ln x - \ln \mu)^2/2\sigma^2] \\ &\propto x^{-3/2} \exp[-\{\ln(x/\mu)\}^2/2\sigma^2] \\ &\propto x^{-3/2} h\left(\ln \frac{x}{\mu}\right) \text{ where } h \text{ is an even function.} \end{aligned}$$

This satisfies the requirements of Kendall's Theorem and we can conclude that the lognormal distribution has a self-symmetric Leimkuhler curve. □

5.3. Dagum distribution

Another three-parameter distribution widely used to describe income distributions is one introduced by Dagum (1977). Again this is a distribution that may be useful in modelling citation age data, but we leave consideration of this to another time. The pdf of the Dagum distribution may be written

$$f(x) = \frac{apx^{ap-1}}{b^{ap}[1 + (x/b)^a]^{p+1}}, \quad \text{for } x > 0.$$

At first sight, this might seem to be simply a reparametrisation of the Singh–Maddala distribution in (9). This is not so. They are in fact related, however, since it can be shown that if X has a Dagum distribution, then $1/X$ has a Singh–Maddala distribution and vice versa. They are both members of the family of distributions originally introduced by Burr (1942).

If we consider the special case where $p = 1 - 1/a$ and $b = 1$. (Again b is a scale parameter so there is no loss of generality when we consider the Leimkuhler curve.) The pdf then reduces to $f(x) = \frac{(a-1)x^{a-2}}{(1+x^a)^{2-1/a}}$. Unfortunately the Leimkuhler curve cannot be expressed in terms of elementary functions so we cannot use the definition of self-symmetry (8) but again the use of Kendall's Theorem shows that this pdf has a self-symmetric Leimkuhler curve. The proof is left as an exercise for the reader.

Remark. In the early days of informetrics, Leo Egghe (1990) propounded a formal structure in terms of what he called information production processes (IPPs) describing the relationship between items and sources; see Egghe and Rousseau (1990) for a summary. Although this has now been mostly supplanted by the stochastic approach, it is interesting to note the inter-relationship between IPPs and Lorenz/Leimkuhler representations as brought out, somewhat indirectly, in Egghe (1992), Rousseau (1992b) and Burrell (1993). In particular, the last two papers both highlighted the formula (8) for self-symmetry given in the definition. As an aside, Burrell (1993) further showed that, if we write $H(\Phi) = L(1 - \Phi)$, then the self-symmetry condition (8) is equivalent to $H(H(\Phi)) = \Phi$, so that H equals its own inverse. Although it has an undoubted aesthetic appeal, it is hard to say what, if any, is the practical value of self-symmetry. Although Damgaard and Weiner (2000) argued in favour of its use as supporting a lognormal model, we would prefer rather more direct support!

6. Concluding remarks

In this paper we have focussed on the geometric rather than the concentration aspects of the Leimkuhler curve. Most of the results are well known in the econometrics literature but many have received little attention in informetrics. Two that we would draw particular attention to are the uses of the power transformation and applications of the Singh–Maddala and Dagum distributions. For the former, we have noted that Basu (1992, 1995) has already used essentially this approach. In Basu's work she was trying to gain an improved fit via the Leimkuhler curve to data that would traditionally have been presented via a Bradford curve (Bradford, 1934). It is worth recalling that Burrell (1992b) pointed out that a Bradford curve is essentially equivalent, after standardisation, to a Leimkuhler curve but using a logarithmic scale for the horizontal scale. In view of this equivalence, and the fact that the Leimkuhler approach is more flexible than that of Bradford we would argue that the former should hold sway. In particular, it would seem that Basu's power transformation idea warrants further investigation.

In this paper we have been concerned with the graphical presentation, via the Leimkuhler curve, of productivity distributions. In other areas of informetrics, such as the analysis of citation age data, or ageing of information sources, various standard families of probability distributions—including exponential, Pareto, lognormal, Weibull and log-logistic—have been considered; see e.g. Egghe and Ravichandra Rao (1992), Gupta (1998) and Burrell (2002). Given their flexible nature and their success in modelling income distributions, consideration of the Singh–Maddala and Dagum families could well prove fruitful in modelling such distributions also. Leimkuhler representations of such distributions appear not to have been considered in the literature.

One important application of the Leimkuhler curve—as a representation of productivity concentration—has not been considered directly here. Such applications, via the Gini coefficient or index, can be found discussed in such as Burrell (1991, 1992c, 2005).

References

Arnold, B. C. (1987). Majorization and the Lorenz order. *Lecture Notes in Statistics* (43). Berlin & New York: Springer.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory, 2*, 244–263, Reprinted In A. B. Atkinson (Ed.) (1973). *Wealth, income and inequality* (pp. 46–68). Harmondsworth: Penguin.

Basu, A. (1992). Hierarchical distributions and Bradford's law. *Journal of the American Society for Information Science, 43*, 494–500.

Basu, A. (1995). Concentration measures in random hierarchical distributions. *JISSI: The International Journal of Scientometrics and Informetrics, 1*, 39–48.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–86.

Burr, I. W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics, 13*, 215–232.

Burrell, Q. L. (1985). The 80/20 rule: library lore or statistical law. *Journal of Documentation, 41*, 24–39.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics, 21*, 181–194.

Burrell, Q. L. (1992a). A note on a result of Rousseau for concentration measures. *Journal of the American Society for Information Science, 43*, 452–454.

Burrell, Q. L. (1992b). The dynamic nature of bibliometric processes: A case study. In I. K. Ravichandra Rao (Ed.), *Informetrics—91: Selected papers from the third international conference on informetrics* (pp. 97–129). Bangalore: Ranganathan Endowment.

Burrell, Q. L. (1992c). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing and Management, 28*, 19–33.

Burrell, Q. L. (1992d). A simple model for linked informetric processes. *Information Processing and Management, 28*, 637–645.

Burrell, Q. L. (1993). A remark on the geometry of Egghe's dual IPPs. *Information Processing and Management, 29*, 515–521.

Burrell, Q. L. (2002). Modelling citation age data: Simple graphical methods from reliability theory. *Scientometrics, 55*, 273–285.

Burrell, Q. L. (2003). Type/token-taken informetrics: Some comments and further examples. *Journal of the American Society for Information Science and Technology, 54*, 1260–1263.

Burrell, Q. L. (2005). Measuring similarity of concentration between different informetric distributions: Two new approaches. *Journal of the American Society for Information Science and Technology, 56*, 704–714.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. *Economie Appliquée, 30*, 413–437.

Damgaard, C., & Weiner, J. (2000). Describing inequality in plant size or fecundity. *Ecology, 81*, 1139–1142.

Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16*, 17–27.

Egghe, L. (1992). Duality aspects of the Gini index for general information production processes. *Information Processing and Management, 28*, 35–44.

Egghe, L. (2003). Type-token taken informetrics. *Journal of the American Society for Information Science and Technology, 54*, 603–610.

Egghe, L. (2004). Positive reinforcement and 3-dimensional informetrics. *Scientometrics, 60*, 497–509, Corrigendum, 2004b.

Egghe, L., & Ravichandra Rao, I. K. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information Processing and Management, 28*, 201–217.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Fellman, J. (1976). The effect of transformations on Lorenz curves. *Econometrica, 44*, 823–824.

Gupta, B. M. (1998). Growth and obsolescence of literature in theoretical population genetics. *Scientometrics, 42*, 335–347.

Jacobsson, U. (1976). On the measurement of the degree of progression. *Journal of Public Economics, 5*, 161–168.

Kendall, M. G. (1956). Discussion of Hart and Prais. *Journal of the Royal Statistical Society (A), 119*, 184–185.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: Wiley.

Lambert, P. J. (2001). *The distribution and redistribution of income* (3rd ed.). Manchester: Manchester University Press.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Journal of the American Statistical Association, 9*, 209–219.

Rasche, R. H., Gaffney, J., Koo, A. Y. C., & Obst, N. (1980). Functional forms for estimating the Lorenz curve. *Econometrica, 48*, 1061.

Rousseau, R. (1992a). Concentration and diversity of availability and use in information systems. *Journal of the American Society for Information Science, 43*, 391–395.

Rousseau, R. (1992b). Two remarks on the preceding paper by L. Egghe. *Information Processing and Management, 28*, 45–51.

Sarabia, J.-M., Castillo, E., & Slottje, D. J. (1999). An ordered family of Lorenz curves. *Journal of Econometrics, 91*, 43–60.

Singh, S. K., & Maddala, G. S. (1975). A stochastic process for income distribution and tests for income distribution functions. *ASA Proceedings of the Business and Economic Statistics Section*, 551–553.

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

Thompson, W. A. (1976). Fisherman's luck. *Biometrics, 32*, 265–271.

Trueswell, R. W. (1966). Determining the optimal number of number of volumes for a library's core collection. *Libri, 16*, 49–60.

Trueswell, R. W. (1969). Some behavioral patterns of library users: the 80/20 rule. *Wilson Library Bulletin, 43*, 458–461.

Trueswell, R. W. (1976). Growing libraries: Who needs them? In D. Gore (Ed.), *Farewell to Alexandria: Solutions to space, growth and performance problems of libraries* (pp. 72–104). Westport, Connecticut: Greenwood Press.