the review of income and wealth

Review of Income and Wealth
Series 57, Number 3, September 2011
DOI: 10.1111/j.1475-4991.2010.00425.x

A GENERAL METHOD FOR CREATING LORENZ CURVES

BY ZUXIANG WANG

Wuhan University

YEW-KWANG NG AND RUSSELL SMYTH*

Monash University

A general method to construct parametric Lorenz models of the weighted-product form is offered in this paper. Initially, a general result to describe the conditions for the weighted-product model to be a Lorenz curve, created by using several component parametric Lorenz models, is given. We show that the key property for an ideal component model is that the ratio between its second derivative and its first derivative is increasing. Then, a set of Lorenz models, consisting of a basic group of models, along with their convex combinations, is proposed, and it is shown that any model in the set possesses this key property. We introduce the concept of balanced fit, which provides a means of assigning weights, according to the preferences of the practitioner, to two alternative objectives for developing Lorenz curves in practice. These objectives are generating an acceptable Lorenz curve and improving the accuracy of the density estimation. We apply the balanced fit approach to income survey data from China to illustrate the performance of our models. We first show that our models outperform other popular traditional Lorenz models in the literature. Second, we compare the results generated by the balanced fit approach applied to one of the Lorenz models that we develop with those generated by the kernel method to show that the approach proposed in the paper generates plausible density estimates.

1. INTRODUCTION

The parametric Lorenz model is an important tool in income distribution analysis. Many researchers have contributed to the literature on Lorenz models. Normally, each contribution provides an individual model with test results applied to some empirical data. Schader and Schmid (1994) give an exhaustive list of the models until the mid-1990s. More recent models include those proposed by Ogwang and Rao (1996, 2000), Ryu and Slottje (1996), and Sarabia *et al*. (1999, 2001). Overall, there have been about two dozen Lorenz models proposed in the literature. For a comparison of existing models, see Cheong (2002) and Schader and Schmid (1994).

The shortcomings of existing models in the literature include the following. First, they fail to explain why a specific functional form can be used to model income data for a variety of sources (Ryu and Slottje, 1996). Second, some models do not give a global approximation to the actual data. Specifically, they may fit the data well at some parts of the distribution, but are poor fits elsewhere (Basmann *et al*., 1990; Ryu and Slottje, 1996; Ogwang and Rao, 2000). Third, some models do not satisfy the definition of the Lorenz curve. We address these limitations by providing a general method to construct Lorenz models. There are three important

Note: We thank three referees for helpful comments on an earlier version of this paper.

*Correspondence to: Russell Smyth, Department of Economics, Monash University, 3800 Australia (Russell.Smyth@BusEco.monash.edu.au).

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA, 02148, USA.

features of the models which we provide: they each satisfy the definition of the Lorenz curve; the efficiency of some of the models has never before been demonstrated in the literature; and several well-known models in the literature are included as special cases of these models.

The general method we propose entails constructing weighted-product models by using a special set of parametric Lorenz models. The simplest weighted-product model is the multiplicative form of two-component Lorenz models. We first provide general conditions for this simplest form to satisfy the definition of the Lorenz curve and find that an ideal component for the multiplicative form is that the ratio between its second derivative and its first derivative is increasing. Equipped with this result, we provide a general theorem which sets forth the conditions for a weighted-product model of finite Lorenz models to satisfy the definition of the Lorenz curve. We then suggest a special set X of parametric Lorenz models with this ideal property. The set X consists of a few simple Lorenz models as well as their convex combinations. These simple models can be understood as generalizations of the Lorenz curve associated with the classical Pareto distribution. With the aid of the general theorem, and the set X, we can generate millions of weighted-product models.

In addition, we propose the method of balanced fit as a compromise between two different, though related, objectives when developing Lorenz curves in practice. On the one hand, the key objective may be to obtain an overall measure of income inequality such as the Gini index where the acceptability of the Lorenz curve is important. On the other hand, the major objective may be to estimate the poverty index, making the accuracy of the density estimates over the relevant range important. We therefore have two criteria to evaluate the Lorenz curve estimation. As both objectives may be important for different purposes, we propose the idea of a balanced fit with different weights (summing to one) given to these two different objectives. This produces a more general method to determine the Lorenz curve.

To illustrate the performance of our models, and the concept of balanced fit, we use data on income distribution from two sources. We initially use data on income distribution in the United States, previously used by Basmann *et al.* (1993) to demonstrate the performance of several of our proposed Lorenz models. The rationale for using the data for this purpose is that it provides continuity with several others who have used these data to test the performance of their proposed Lorenz models. We then proceed to apply the balanced fit approach to rural and urban income survey data collected by the State Statistical Bureau in Hubei province in China in 2006 to compare the performance of our main models with popular existing models in the literature and to show that our approach generates plausible density estimates.

China has undergone large-scale economic transition since market reforms in the late 1970s, which has resulted in a high rate of economic growth. Rapid economic growth, however, has been accompanied by a sharp increase in income inequality (see, e.g. Chotikapanich *et al.*, 2007). Rising income inequality threatens China's ability to maintain sustainable growth and potentially impinges on political and social stability (Wan and Zhou, 2005). The latter has been of particular concern to the Chinese government, with income distribution a central platform of constructing a harmonious society as first enunciated by the Hu-Wen administration during the 2005 National People's Congress. Hence, we use data from China to illustrate our models because income inequality in China is such an important policy issue, and despite the plethora of studies on income inequality in China there is an urgent need for further advancements in measuring income inequality in that country. In particular, most studies of income inequality in China have used household survey data (see, e.g. Meng, 2004). Grouped data are more readily available in China than household survey data, but since the income data are in grouped form, some acceptable Lorenz model is needed to approximate the underlying Lorenz curve. Income inequality in China also has implications that extend beyond its national boundaries. As noted by Chotikapanich *et al.* (2007, pp. 127–8): "As China accounts for about a quarter of the world's population, changes in income and income inequality in China have important implications [for] global income inequality . . . This means that any advancement in the measurement of income inequality within China is not only important for understanding the economic development and well-being of people inside the 'Middle Kingdom,' but also important in the global context."

The structure of the paper is as follows. Sufficient conditions for the weighted-product model to satisfy the definition of the Lorenz curve are set out in the next section. The basic group of Lorenz models is proposed in Section 3. The special set *X* of parametric Lorenz models is provided in Section 4, together with some selected examples of the weighted-product models created from *X*. The concept of the balanced fit is proposed in Section 5, while the test results of our new models are reported in Section 6. The final section offers some suggestions for future research.

2. The General Method for Creating Lorenz Models

We call $L(p)$ a Lorenz curve if $L(p)$, defined on [0,1], possesses a continuous third derivative and satisfies the conditions that $L(0) = 0$, $L(1) = 1$, $L'(p) \geq 0$, and $L''(p) \geq 0$. To commence, consider the function of the multiplicative form:

$$\tilde{L}(p) = f(p)^{\alpha} g(p)^{\upsilon},\ \alpha \geq 0 \text{ and } \upsilon \geq 0$$

where both the component functions $f(p)$ and $g(p)$ are parametric Lorenz curves. It follows that $\tilde{L}(p)$ is a Lorenz curve if $\tilde{L}'(p) \geq 0$ and $\tilde{L}''(p) \geq 0$. But

$$\tilde{L}'(p) = \alpha f(p)^{\alpha-1} f'(p) g(p)^{\upsilon} + \upsilon g(p)^{\upsilon-1} g'(p) f(p)^{\alpha} \geq 0$$

is true, therefore we only have to consider the condition for $\tilde{L}''(p) \geq 0$. Since

$$\begin{aligned}\tilde{L}''(p) = {} & \alpha(\alpha-1) f(p)^{\alpha-2} f'(p)^2 g(p)^{\upsilon} + \alpha f(p)^{\alpha-1} f''(p) g(p)^{\upsilon} \\ & + \alpha\upsilon f(p)^{\alpha-1} f'(p) g(p)^{\upsilon-1} g'(p) + \upsilon(\upsilon-1) g(p)^{\upsilon-2} g'(p)^2 f(p)^{\alpha} \\ & + \upsilon g(p)^{\upsilon-1} g''(p) f(p)^{\alpha} + \upsilon\alpha g(p)^{\upsilon-1} g'(p) f(p)^{\alpha-1} f'(p),\end{aligned}$$

it follows that $\tilde{L}''(p) \geq 0$ if both $\alpha \geq 1$ and $\upsilon \geq 1$ (see Ogwang and Rao, 2000).

We can consider other cases. Denote the sum of the first three terms on the right-hand side of the above equation as $h(p)$ and the sum of the remaining three terms as $t(p)$. Thus, we need only find the condition for both $h(p) \geq 0$ and $t(p) \geq 0$. Since

$$(1)\qquad \frac{h(p)}{\alpha f(p)^{\alpha-2} g(p)^{\upsilon-1}} = (\alpha-1)f'(p)^2 g(p) + f(p)f''(p)g(p) + \upsilon f(p)f'(p)g'(p),$$

we can conclude that $h(p) \geq 0$ if $\alpha \geq 1/2$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, and $f'''(p) \geq 0$. Furthermore, we also have $h(p) \geq 0$ if $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, and $f_1(p) \equiv f''(p)/f'(p)$ is increasing.[1]

Note further:

$$\frac{t(p)}{\upsilon g(p)^{\upsilon-2} f(p)^{\alpha-1}} = (\upsilon-1)g'(p)^2 f(p) + g(p)g''(p)f(p) + \alpha g(p)g'(p)f'(p).$$

The right-hand side of this equation is exactly the same as that of (1), if we exchange the position of $g(p)$ and $f(p)$, and the position of α and υ. Thus we have $t(p) \geq 0$ if $\alpha \geq 0$, $\upsilon \geq 1/2$, $\alpha + \upsilon \geq 1$, and $g'''(p) \geq 0$. Furthermore, we also have $t(p) \geq 0$ if $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, and $g_1(p) \equiv g''(p)/g'(p)$ is increasing.

To synthesize the discussion, we have the following lemma:

Lemma 1. Assume both $f(p)$ and $g(p)$ are Lorenz curves. It follows that $\tilde{L}(p) = f(p)^{\alpha} g(p)^{\upsilon}$ is a Lorenz curve if any of the following conditions holds:

(i) $\alpha \geq 1$ and $\upsilon \geq 1$.

(ii) $\alpha \geq 1/2$, $\upsilon \geq 1$, and $f'''(p) \geq 0$ on [0,1].

(iii) $\alpha \geq 0$, $\upsilon \geq 1$, and $f''(p)/f'(p)$ is increasing on [0,1].

(iv) $\alpha \geq 1/2$, $\upsilon \geq 1/2$, and both $f'''(p) \geq 0$ and $g'''(p) \geq 0$ on [0,1].

(v) $\alpha \geq 0$, $\upsilon \geq 1/2$, $\alpha + \upsilon \geq 1$, $f''(p)/f'(p)$ is increasing and $g'''(p) \geq 0$ on [0,1].

(vi) $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, and both $f''(p)/f'(p)$ and $g''(p)/g'(p)$ are increasing on [0,1].

By symmetry, under the assumption that g''/g' is increasing and $f'''(p) \geq 0$ on [0,1], statement (v) of the lemma implies that $\tilde{L}(p)$ is a Lorenz curve if $\upsilon \geq 0$, $\alpha \geq 1/2$, and $\alpha + \upsilon \geq 1$.[2] For a pair of fixed component Lorenz curves $f(p)$ and $g(p)$, the ideal situation is that both f''/f' and g''/g' are increasing. Statement (vi) then asserts that the admissible range of α and υ is $\{(\alpha,\upsilon)|\alpha \geq 0,\ \upsilon \geq 0,$

[1] If we write the right-hand side of (1) as $\psi(p)$, we find that $\psi(0) = 0$, and $\psi'(p) \geq 0$ for any $p \in [0,1]$. Moreover, assume $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$ and that $f_1(p) \equiv f''(p)/f'(p)$ is increasing, which means $f_1'(p) \geq 0$. Rewrite the right-hand side of (1) as

$$[(\alpha-1)f'(p)g(p) + f(p)f_1(p)g(p) + \upsilon f(p)g'(p)]f'(p).$$

Let the function between the braces be $\varphi(p)$, we can verify that $\varphi(0) = 0$ and $\varphi'(p) \geq 0$ for any $p \in [0,1]$. Consequently, we can again conclude that $h(p) \geq 0$.

[2] Note that the condition $\alpha + \upsilon \geq 1$ cannot be relaxed. If, to the contrary, $\alpha \geq 0$, $\upsilon \geq 0$, and $\alpha + \upsilon < 1$, then by letting $f(p) = g(p) = p$, we get $\tilde{L}(p) = p^{\alpha+\upsilon}$, which is not a Lorenz curve. According to Lemma 1, the stricter the condition imposed upon a component function, the larger the admissible range of the corresponding exponential parameter.

$\alpha + \upsilon \geq 1\}$, which achieves a state of maximum.[3] An important special case of the multiplicative model of two component Lorenz models is $L_S(p) = p^{\alpha}L(p)^{\upsilon}$ (Sarabia *et al.*, 1999). We have the following result by Lemma 1:

Corollary. Assume $L(p)$ is a Lorenz curve. Then $L_s(p)$ is a Lorenz curve if any one or more of the following conditions holds:

(i) $\alpha \geq 0$ and $\upsilon \geq 1$;
(ii) $\alpha \geq 0$, $\upsilon \geq 1/2$, $\alpha + \upsilon \geq 1$, and $L'''(p) \geq 0$;
(iii) $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, and $L''(p)/L'(p)$ is increasing.

Sarabia *et al.* (1999) provide statement (i) of the corollary, but they impose the condition $L'''(p) \geq 0$. The first two statements are also provided, and elaborated on, in Wang *et al.* (2007). Let

$$X_0 = \{L(p)|L(p) \text{ be a Lorenz curve with increasing } L''(p)/L'(p)\}.$$

Consider a series of component Lorenz models $\{L_i(p)\}_{i=1}^{m} \subset X_0$. Denote the weighted-product model

$$\tilde{L}(p) = L_1(p)^{\alpha_1}L_2(p)^{\alpha_2}\cdots L_m(p)^{\alpha_m}, \alpha_1 \geq 0, \alpha_2 \geq 0, \cdots, \alpha_m \geq 0.$$

Furthermore, let

$$Y_0 = \{L(p)|L(p) \text{ be a Lorenz curve with } L'''(p) \geq 0\},$$
$$Z_0 = \{L(p)|L(p) \text{ be a Lorenz curve}\}.$$

Therefore, Z_0 contains all possible parametric Lorenz curves. We have $X_0 \subset Y_0 \subset Z_0$. Our general method of creating Lorenz models is described in the following theorem, which follows from statements (iii), (v), and (vi) of Lemma 1:

Theorem 1. We have three statements:

(i) Let $L(p) \in Z_0$. Then $\tilde{L}(p)L(p)^{\upsilon}$ is a Lorenz curve if $\upsilon \geq 1$.
(ii) Let $L(p) \in Y_0$ and assume that there exists an exponent, say, $\alpha_i \in \{\alpha_1, \ldots, \alpha_m\}$, such that $\alpha_i + \upsilon \geq 1$. Then $\tilde{L}(p)L(p)^{\upsilon}$ is a Lorenz curve if $\upsilon \geq 1/2$.
(iii) Let $\{L_i(p)\}_{i=1}^{m} \subset X_0$ and assume that there is a pair of exponents within $\{\alpha_1, \ldots, \alpha_m\}$, say, α_i and α_j with $\alpha_i + \alpha_j \geq 1$. $\tilde{L}(p)$ itself is a Lorenz curve.[4]

A weighted-product model can also be called a Cobb–Douglas model. Whether our general method is feasible depends on whether we can find the set X_0.

[3] We regard the fact that $A_L(p) = L''(p)/L'(p)$ is increasing as a purely technical condition in this paper. However, $A_L(p)$ can be a measure of the curvature of $L(p)$. Based on the Arrow–Pratt measure of absolute risk aversion (see Pratt, 1964), it can be easily verified that for Lorenz curves $L_I(p)$ and $L_{II}(p)$, $A_{L_I}(p) \geq A_{L_{II}}(p)$ for any $p \in [0,1]$, if and only if there exists a Lorenz curve H such that $L_I(p) = H(L_{II}(p))$. That H is a Lorenz curve implies $p \geq H(p)$ for any $p \in [0,1]$. We thus have $L_{II}(p) \geq H(L_{II}(p))$, and consequently, $L_{II}(p) \geq L_I(p)$ for any $p \in [0,1]$, implying that $L_I(p)$ is Lorenz dominated by $L_{II}(p)$ and that the distribution underlying $L_{II}(p)$ is unambiguously more equal than the distribution underlying $L_I(p)$.

[4] First note that $L_i(p)^{\alpha_i}L(p)$ is a Lorenz curve for any $L_i(p) \in X_0$ and $\alpha_i \geq 0$ as long as $L(p)$ is a Lorenz curve, as implied by statement (iii) of Lemma 1. Since $L(p)^{\upsilon}$ is a Lorenz curve for any $L(p) \in Z_0$ and $\upsilon \geq 1$, the statement implies that $L_m(p)^{\alpha_m}L(p)^{\upsilon}$ is a Lorenz curve. This implies that $L_{m-1}(p)^{\alpha_{m-1}}L_m(p)^{\alpha_m}L(p)^{\upsilon}$ is a Lorenz curve. Hence, statement (i) of the theorem is true by induction. If $L(p) \in Y_0$ and $\alpha_i + \upsilon \geq 1$ with $\upsilon \geq 1/2$ and $\alpha_i \geq 0$, statement (v) of Lemma 1 implies that $L_i(p)^{\alpha_i}L(p)^{\upsilon}$ is a Lorenz curve. Hence, statement (ii) of the theorem follows in a similar manner to the verification of statement (i). The same applies to statement (iii) of the theorem by using statement (vi) of Lemma 1.

If so, we can, for example, create new Lorenz models combining $\tilde{L}(p)$ and any $L(p)$ extant in the literature, according to the first statement of Theorem 1. Unfortunately, we are not able to find the entire set, X_0. However, we can consider a less general alternative by finding a subset $X \subset X_0$ and construct weighted-product models $\tilde{L}(p)$ with the elements of X as components. In the next section we suggest such a set.

3. Generalized Pareto Lorenz Models

Consider the set of basic models:[5]

$$L_1(p) = p, \tag{2}$$

$$L_2(p) = 1-(1-p)^{\beta}, \beta \in (0,1], \tag{3}$$

$$L_{\lambda}(p) = \frac{e^{\lambda p}-1}{e^{\lambda}-1}, \lambda > 0, \tag{4}$$

$$L_3(p) = 1 - L_{\lambda_1}(1-p)^{\beta_1}, \beta_1 \in (0,1], \lambda_1 \in (-\infty, 0) \cup (0, \ln \beta_1^{-1}], \tag{5}$$

$$L_4(p) = 1-\left(1-L_{\lambda_2}(p)\right)^{\beta_2}, \beta_2 \in (0,1], \lambda_2 \in [\ln \beta_2, 0) \cup (0, +\infty). \tag{6}$$

These functions possess the derivative of any order. $L_2(p)$ is the Lorenz curve associated with the classical Pareto distribution. $L_\lambda(p)$ is the Lorenz curve suggested by Chotikapanich (1993) with λ its unique parameter. $L_\lambda(p)$ is satisfied for any $\lambda \neq 0$,

$$L_\lambda(p) \geq 0 \text{ and } L_\lambda'(p) \geq 0 \text{ on } [0,1],$$
$$L_\lambda^{(n)}(p) = \lambda^{n-1} L_\lambda'(p), n = 2, 3, \cdots.$$

To avoid confusion in the ensuing discussion, note that unlike the parameter λ in the model $L_\lambda(p)$ or $L_\lambda(1-p) = (e^\lambda - 1)^{-1}(e^{\lambda(1-p)} - 1)$, the symbol i in $L_i(p)$ does not represent a parameter of the model. $L_1(p)$ is a special case of $L_2(p)$ because it can be obtained by letting $\beta = 1$ in the latter. $L_\lambda(p)$ is equal to $L_1(p)$ when $\lambda \to 0$. $L_3(p)$ is the Lorenz model provided by Wang and Smyth (2007) and is a generalization of $L_2(p)$. $L_4(p)$ is a new model and is also a generalization of $L_2(p)$. We call these basic models generalized Pareto (GP) models.

Note that we have the following two inequalities:

$$(1-\beta_1)L_{\lambda_1}'(1-p) - \lambda_1 L_{\lambda_1}(1-p) \geq 0, \tag{7}$$

[5] Strictly speaking, $L_1(p) = p$ is not the Lorenz curve associated with complete equality. As everyone has the same income level, strictly speaking, no one can be said to be at the lowest or highest 20 percent (or any other figure) of the population. The associated Lorenz curve then exists only at the origin and the termination point by the definition of the curve. To overcome this point, we may adopt, for the practically non-existent case of complete equality only, the convention of allocating any fraction $0 < x < 1$ of the population to be the lowest/highest x percent. This convention then allows the 45 percent line through the origin to be associated with complete equality as usually loosely taken to be so. This allows us to use $L_1(p) = p$ here and it can be a useful component in the creation of Lorenz curves.

(8) $$(1-\beta_2)L'_{\lambda_2}(p)+\lambda_2\big(1-L_{\lambda_2}(p)\big)\geq 0,$$

with the parameters defined in (5) and (6), respectively. By the definition of $L_\lambda(p)$, the inequality (7) is equivalent to $\lambda_1(e^{\lambda_1}-1)^{-1}\big(1-\beta_1 e^{\lambda_1(1-p)}\big)\geq 0$. This inequality holds because $\lambda_1(e^{\lambda_1}-1)^{-1}\geq 0$ for any $\lambda_1 \neq 0$ and $1-\beta_1 e^{\lambda_1(1-p)}\geq 0$ if β_1 and λ_1 are defined by (5). Meanwhile, (8) is equivalent to $\lambda_2(e^{\lambda_2}-1)^{-1}\big(e^{\lambda_2}-\beta_2 e^{\lambda_2 p}\big)\geq 0$. It also holds if β_2 and λ_2 are defined by (6).[6] Furthermore, we have:

Lemma 2. Every GP model $L(p)$ is a Lorenz curve with increasing $L''(p)/L'(p)$.[7]

Employing only the GP models and the third statement of Theorem 1, we can create many weighted-product models. For example, all the following are Lorenz curves:

(9) $$\big(1-(1-p)^{\beta}\big)^{\upsilon},\ \beta\in(0,1],\ \upsilon\geq 1,$$

(10) $$p^{\alpha}\big[1-(1-p)^{\beta}\big]^{\upsilon},\ \beta\in(0,1],$$

(11) $$p^{\alpha}L_{\lambda}(p)^{\upsilon},\ \lambda>0,$$

(12) $$p^{\alpha}\big[1-L_{\lambda}(1-p)^{\beta}\big]^{\upsilon},\ \beta\in(0,1],\ \lambda\in(-\infty,0)\cup(0,\ln\beta^{-1}],$$

(13) $$p^{\alpha}\big[1-(1-L_{\lambda}(p))^{\beta}\big]^{\upsilon},\ \beta\in(0,1],\ \lambda\in[\ln\beta,0)\cup(0,+\infty),$$

where $\alpha\geq 0$, $\upsilon\geq 0$, and $\alpha+\upsilon\geq 1$ for the models specified in (10)–(13); (9) is the model provided by Rasche *et al.* (1980); (10) and (11) are models proposed by Sarabia *et al.* (1999, 2001, respectively), but with $\upsilon\geq 1$ imposed; and (12) is suggested by Wang and Smyth (2007), but with $\upsilon\geq 1/2$ imposed. Since $L_\lambda(x)\to x$ when $\lambda\to 0$, (12) includes (9)–(10) as special cases and (13) includes (9)–(11) as

[6] To see that $L_3(p)$ is a Lorenz curve we have only to verify both $L'_3(p)\geq 0$ and $L''_3(p)\geq 0$. However, $L'_3(p)=\beta_1 L_{\lambda_1}(1-p)^{\beta_1-1}L'_{\lambda_1}(1-p)$. Thus $L'_3(p)\geq 0$. Moreover,

$$L''_3(p)=\beta_1 L_{\lambda_1}(1-p)^{\beta_1-2}[(1-\beta_1)L'_{\lambda_1}(1-p)-\lambda_1 L_{\lambda_1}(1-p)]L'_{\lambda_1}(1-p).$$

Therefore, $L''_3(p)\geq 0$ by (7). Using (8) and the same deviation we can verify $L_4(p)$ is also a Lorenz curve.

[7] The statement is evident for $L_1(p)$, $L_\lambda(p)$ and $L_2(p)$. Denote

$$h(p)=L''_3(p)/L'_3(p)=[(1-\beta_1)L'_{\lambda_1}(1-p)-\lambda_1 L_{\lambda_1}(1-p)]/L_{\lambda_1}(1-p).$$

Thus, we have $h'(p)=(1-\beta_1)L'_{\lambda_1}(1-p)[L'_{\lambda_1}(1-p)-\lambda_1 L_{\lambda_1}(1-p)]/L_{\lambda_1}(1-p)^2$. The right-hand side is non-negative by (7). Therefore $L''_3(p)/L'_3(p)$ is increasing. Using (8) and the same steps we can verify that $L''_4(p)/L'_4(p)$ is also increasing.

special cases. Note that Sarabia *et al.*'s (1999) model $p^{\alpha}\left(1-(1-p)^{\beta}\right)^{\upsilon}$ with $\alpha \geq 0$ and $\upsilon \geq 1$ may be significantly inferior to the same model with $\alpha \geq 0$, $\upsilon \geq 0$, and $\alpha + \upsilon \geq 1$. The former is a sub-model of the latter. The latter may be useful in practice, but we do not test it since, for the data used, we get parameter estimates $\alpha = 0$ and $\upsilon > 1$; namely, it is equivalent to the Rasche *et al.* (1980) model $\left(1-(1-p)^{\beta}\right)^{\upsilon}$ for the data tested.

Assume again that $\alpha \geq 0$, $\upsilon \geq 0$, and $\alpha + \upsilon \geq 1$. By Theorem 1, a more sophisticated Lorenz model is as follows:

$$p^{\alpha}\left[1-L_{\lambda_1}(1-p)^{\beta_1}\right]^{\alpha_1}\left[1-\left(1-L_{\lambda_2}(p)\right)^{\beta_2}\right]^{\upsilon}, \tag{14}$$

where $\alpha_1 \geq 0$ should be imposed. Of course, we can also impose $\alpha + \alpha_1 \geq 1$ or $\alpha_1 + \upsilon \geq 1$ instead. Other parameters are defined in (5)–(6). We have avoided using a GP member repeatedly in any model above. However, Theorem 1 implies that

$$\left(1-(1-p)^{\beta_1}\right)^{\alpha}\left(1-(1-p)^{\beta}\right)^{\upsilon} \tag{15}$$

is also a Lorenz curve, if $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, $\beta \in (0,1]$, and $\beta_1 \in (0,1]$. Model (15) nests

$$p^{\alpha}\left(1-(1-p)^{\beta}\right) \tag{16}$$

where $\alpha \geq 0$ and $\beta \in (0,1]$, suggested by Ortega *et al.* (1991) and the models defined by (9)–(10). Clearly, (14) nests all other models presented here and should outperform these other models.

4. The Weighted-Product Lorenz Models

While we have obtained a number of Lorenz models in the last section, better options still exist. Define

$$X = \{L(p) | L(p) \text{ is a convex combination of the GP models}\}.$$

Every element of X can be used as a Lorenz model. (For $L_1(p) = p$, see footnote 4.) Note that the requirement that $h(p) = L''(p)/L(p)'$ is increasing is equivalent to $h'(p) \geq 0$ or $L'''L' - L''^2 \geq 0$ under the continuity assumption of the derivatives. This implies $L'''(p) \geq 0$ in turn, because a Lorenz curve $L(p)$ must satisfy $L'(p) \geq 0$. Let $x(p)$ and $y(p)$ be Lorenz curves with increasing L''/L'. A sufficient condition for the weighted sum $L(p) = \delta x(p) + (1-\delta)y(p)$ to have increasing L''/L', where δ is the weight coefficient and satisfies $\delta \in [0,1]$, is

$$x'''y' + y'''x' - 2x''y'' \geq 0. \tag{17}$$

First we give a simple result.

Lemma 3. Let $X_1 \subset X_0$ be a set of Lorenz models with (17) being satisfied for any pair $x(p)$ and $y(p)$ in X_1, where X_0 is defined above. Then, any convex combination of the elements of X_1 belongs to X_0.

Theorem 2. Every element of X is a Lorenz curve with increasing L''/L'.

We postpone proofs of Lemma 3 and Theorem 2 to the online Appendices 1 and 2.[8] Using the five GP models, X contains 31 linearly independent elements. Therefore, millions of weighted-product Lorenz models can be created by using the elements of X inclusively and the third statement of Theorem 1, even if we refrain from using an element of X repeatedly in building a model. The method can be reinforced by adding even a single new member to the GP group. Theorem 2 implies that X will contain 63 elements, increasingly significantly the availability of the weighted-product models. Alternatively, we can use less desirable models. For example, $L(p) = pA^{p-1}$, where $A > 0$, is a Lorenz curve with $L'''(p) \geq 0$, which is suggested by Gupta (1984). Wang *et al.* (2007) provide another option:

$$H(p) = 1 - e^{-\gamma p}(1-p)^{\beta}$$

with $H'''(p) \geq 0$, where $\beta \in (0,1]$ and $0 \leq \gamma + \beta \leq \sqrt{\beta}$. Adding even a single such model to the GP group, we can create about 2^{63} weighted-product models by the second statement of Theorem 1. However, the admissible range of the exponential parameter α_i for the component with, say, $H(p)$, no longer satisfies $\alpha_i \geq 0$. For example,

$$p^{\alpha}\{\delta[1 - e^{-\gamma p}(1-p)^{\beta}] + (1-\delta)[1 - L_{\lambda_1}(1-p)^{\beta_1}]\}^{\upsilon} \tag{18}$$

is a Lorenz curve, where $\upsilon \geq 1/2$ must be imposed by Theorem 1. Other parameter ranges for (18) are $\alpha \geq 0$, $\alpha + \upsilon \geq 1$, $\beta \in (0,1]$, $0 \leq \gamma + \beta \leq \sqrt{\beta}$, $\delta \in [0,1]$, $\beta_1 \in (0,1]$, and $\lambda_1 \in (-\infty, 0) \cup (0, \ln \beta_1^{-1}]$. Therefore, Theorems 1 and 3 suggest many Lorenz curves. The following are a few examples with only GP members involved.

$$p^{\alpha}\{\delta L_{\lambda}(p) + (1-\delta)[1 - (1 - L_{\lambda_2}(p))^{\beta_2}]\}^{\upsilon}, \tag{19}$$

$$p^{\alpha}\{\delta_1(1 - (1-p)^{\beta}) + \delta_2 L_{\lambda}(p) + \delta_3(1 - L_{\lambda_1}(1-p)^{\beta_1})\}^{\upsilon}, \tag{20}$$

$$(1 - L_{\lambda_1}(1-p)^{\beta_1})^{\alpha}\{\delta p + (1-\delta)[1 - (1 - L_{\lambda_2}(p))^{\beta_2}]\}^{\upsilon}, \tag{21}$$

$$[\delta p + (1-\delta)L_{\lambda}(p)]^{\alpha}\{\delta_1(1 - L_{\lambda_1}(1-p)^{\beta_1}) + (1-\delta_1)L_{\lambda_0}(p)\}^{\upsilon}, \tag{22}$$

$$p^{\alpha}[\delta p + (1-\delta)L_{\lambda}(p)]^{\alpha_1}(1 - (1-p)^{\beta})^{\upsilon}, \tag{23}$$

$$p^{\alpha}\{\delta L_{\lambda}(p) + (1-\delta)(1 - L_{\lambda_1}(1-p)^{\beta_1})\}^{\upsilon}. \tag{24}$$

[8] The appendices can be downloaded from the website of the journal (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1475-4991).

The parameter ranges are at their maximum for all these models; namely, $\alpha \geq 0$, $\upsilon \geq 0$, $\alpha + \upsilon \geq 1$, $\alpha_1 \geq 1$, $\lambda_0 > 0$, δ, δ_1, δ_2, $\delta_3 \in [0,1]$, and $\delta_3 = 1 - \delta_1 - \delta_2$. Those not mentioned are defined in (3)–(6). Since the models are all non-linear functions, a non-linear least squares (NLS) algorithm must be used to estimate the parameters. While the parameter ranges seem complicated, they can be enforced by analogous parameter transformations to those used in Wang *et al*. (2007). Such transformations allow us to use the unconstrained non-linear least squares (UNLS) algorithm which is generally more efficient than its constrained counterpart. For example, the condition for the three weight coefficients in (20) can be enforced by parameter transformations

$$\delta_1 = \sin^2\theta_1, \delta_2 = \cos^2\theta_1 \sin^2\theta_2, \text{and } \delta_3 = \cos^2\theta_1 \cos^2\theta_2 = 1 - \delta_1 - \delta_2$$

where θ_1 and θ_2 are two new parameter variables.[9]

We can expect that over-parameterization will occur in general when we use too many elements of X as components in a model. There are three guiding lessons in creating the weighted-product models. One is that we should include different models of X, rather than use specific instances repeatedly in creating a single model as done, for example, in the model specified in (15). We find that (15) performs only slightly better than (10). The second is that models with convex combination components perform better. For example, (22) with two convex combination components performs very well, while (14) performs relatively poorly considering the number of parameters involved. Third, components with (4), (5), (6), or $H(p)$ involved tend to be more satisfactory when constructing the models. For instance, (19) or (24) is very satisfactory. There is another explanation. $L_\lambda(x) \to x$ when $\lambda \to 0$. This implies that $L_\lambda(1 - p)^\beta$ or $(1 - L_\lambda(p))^\beta$ is much more flexible than $(1 - p)^\beta$. Therefore (12) or (13) are generalizations of (9) or (10). Since (9) and (10) perform quite satisfactorily in many of the instances amongst the traditional Lorenz models, it is therefore reasonable to expect that (12) or (13) or the models which have (5) or (6) as components will also generate good results.

One of the drawbacks of some of the above models is that they are complicated. An alternative to the above complicated models is trying simple models, such as:

$$\delta p^\alpha[1-(1-p)^\beta] + (1-\delta)[1-(1-p)^{\beta_1}]^\upsilon \tag{25}$$

in applications. One of the advantages of (25) is that its convexity is clear, given the findings of Ortega *et al*. (1991) and Rasche *et al*. (1980), respectively.

[9] The condition $\alpha + \upsilon \geq 1$ with $\alpha \geq 0$ and $\upsilon \geq 1/2$ can be enforced by

$$\alpha = (1/2 + \zeta^2)\sin^2\theta, \upsilon = 1/2 + (1/2 + \zeta^2)\cos^2\theta$$

with ζ and θ two new parameter variables. $\lambda \in [\ln\beta, 0) \cup (0, +\infty)$ with $\beta \in [0,1]$ can be enforced by $\lambda = \ln\beta + \zeta^2$ with ζ a new parameter. Since $L_\lambda(p) = p$ as $\lambda \to 0$, we do not have to avoid $\lambda = 0$ when estimating the models.

5. The Lorenz Curve of Balanced Fit

With the variety of the Lorenz models developed above, we are able to consider more sophisticated fitting applications for grouped data. Assume that we have mean income μ and income ranges $x_0 < x_1 < \dots < x_n < x_{n+1}$, where $x_0 \geq 0$ and x_{n+1} is a sufficiently large number. Moreover, assume that we have grouped data $(p_i, L_i)_{i=0}^{n+1}$ with $p_0 = L_0 = 0$ and $p_{n+1} = L_{n+1} = 1$, where p_i is the cumulative proportion of income units whose incomes are less than x_i and L_i is the income share owned by the population. The Lorenz curve is denoted as $l(p)$. $l'(p)$ is equal to the p-quantile of the underlying distribution, divided by μ. $l(p)$ satisfies $l(p_i) = L_i$ and $l'(p_i) = x_i/\mu$ for $i = 1,2, \dots, n$. Specifically, all the information given is contained in:

$$(p_i, L_i), \text{ for } i = 1, 2, \cdots, n \tag{26}$$

$$(p_i, x_i/\mu), \text{ for } i = 1, 2, \cdots, n. \tag{27}$$

Kakwani (1976) uses both (26) and (27) to create polynomial functions to approximate the Lorenz curve. However, many authors do not use (27) in developing Lorenz models. Instead, they only require their estimated Lorenz curve to be as close as possible to (26), normally, by minimizing the objective function $\sum_{i=1}^{n}(L(p_i) - L_i)^2$, assuming that the estimated Lorenz curve will generate an acceptable approximation to (27). This may not necessarily be the case, where $L(p)$ is the proposed parametric Lorenz model. One can take the opposite approach and attempt to get the derivative of the estimated Lorenz curve as close as possible to (27) by minimizing $\sum_{i=1}^{n}(L'(p_i) - x_i/\mu)^2$, while assuming that the estimated Lorenz curve can produce an acceptable approximation to (26). This may also not necessarily be the case, but improved estimation of the derivative suggests it should be possible to better estimate the relative frequency, since the solver of $\mu L'(p) - x = 0$ is the relative frequency at x.

The above two approaches to estimating Lorenz curves represent two extremes, which are useful in practice. If our main objective is to obtain the Gini index, where the approximation of the Lorenz curve is important, the former approach is useful. On the other hand, if what we are mainly concerned with is the density estimate, for example, in poverty index estimation where a plausible density estimate is essential, the latter approach is useful. We can consider a third approach between the two extremes. This can be achieved by a trade-off between the two extremes, namely, choosing $b \in [0,1]$ and then minimizing the balanced objective function

$$b\sum_{i=1}^{n}(L(p_i) - L_i)^2 + (1-b)\sum_{i=1}^{n}(L'(p_i) - x_i/\mu)^2 \tag{28}$$

to find an estimate of the Lorenz curve. If one's key focus is on improving the approximation quality of the estimated Lorenz curve, choosing a larger value of b < 1 is more appropriate. Alternatively, if one is more concerned with the accuracy of the density estimate, selecting a smaller $b > 0$ is more appropriate. In this sense, b can be adjusted as a balance according to the objectives of the practitioner.

We find that

$$(29) \qquad b\sum_{i=1}^{n}(L(p_i)-L_i)^2+(1-b)\sum_{i=1}^{n}\left(\hat{F}(x_i)-p_i\right)^2$$

is a better form, where $\hat{F}(x_i)$ is the root of $\mu L'(p) - x_i = 0$ for each x_i and is a relative frequency estimate at x_i. Minimizing (29) may not necessarily yield the same solution as minimizing (28). However, not only is (29) better numerically when $b \neq 1$ since all the numbers involved in the function are then in the interval [0,1], but the objective function in (28) will also be small at a solution that minimizes (29) if the solution makes the objective function in (29) small. We call the resultant Lorenz curve from minimizing (28) or (29), the Lorenz curve of balanced fit.

To require the fitted curve to take account of the function values as well as derivatives is not new. The well-known Hermite polynomial interpolation is widely used in approximation theory, where piecewise polynomials are used to interpolate both (26) and (27). Better tools, such as splines, can also be used to interpolate the two sets of conditions. One difficulty of such interpolations for the grouped income data is that the approximation is cumbersome in the income intervals at the two ends of the entire income range. Another difficulty is that the computation is quite complicated. Cowell and Mehta (1982) and Kakwani (1976) thoroughly study these methods.

The balanced fit cannot be implemented without satisfactory Lorenz models. For example, the non-Lorenz-curve functions used to model Lorenz curves in the literature (see, e.g. Kakwani and Podder, 1976; Kakwani, 1980; Basmann *et al.*, 1990,) cannot be used to form the second term of (28) or (29), since $\hat{L}'(p_i)$ may not be positive, or $\mu\hat{L}'(p)-x_i=0$ may have multiple solvers or no solvers at all.

6. Empirical Calculations

In the tests performed in this section, we use UNLS to estimate the parameters of the Lorenz models developed in this paper. Two examples are presented to illustrate the performance of our models. The first uses data on income distribution for the United States, which has previously been used by Basmann *et al.* (1993), to illustrate the performance of models (14) and (18)–(23). The second uses data from the 2006 income survey of Hubei province, China to illustrate the concept of balanced fit.

6.1. *U.S. Income Data Estimation*

The United States income distribution data, used by Basmann *et al.* (1993), consists of grouped data for seven years over the period 1977–83. In total, there are 99 points on the empirical Lorenz curve for each year; that is, $(p_i, L(p_i))_{i=1}^{99}$ with $p_i = 0.01i$, where $L(p)$ denotes the empirical Lorenz curve. We fit the models given in (14) and (18)–(23), respectively, to the 99 points by minimizing (28) with $b = 1$, since we do not have the associated income interval information.

TABLE 1

Error Measures and Gini Estimates for U.S. 1977 Income Data

	(14)	(18)	(19)	(20)	(21)	(22)	(23)
MSE × 10^6	0.3835	0.0246	0.0308	0.0291	0.0298	0.0067	0.3929
MAE	0.0005	0.0001	0.0001	0.0001	0.0002	0.0001	0.0005
MAXABS	0.0028	0.0005	0.0007	0.0006	0.0005	0.0002	0.0026
Gini	0.3685	0.3683	0.3683	0.3683	0.3683	0.3683	0.3681
	0.0216	**0.0211**	**0.0229**	**0.0225**	**0.0210**	**0.0211**	**0.0222**

Note: Figures in bold below the Gini indices are their bootstrap standard errors.

Our estimation results for 1977 are presented in Table 1, with estimated parameters in Table A1 in the online Appendix 3, where error measures

$$(30)\qquad \begin{cases} \text{MAXABS} = \max\limits_{1\le i\le n}\left|\hat{L}(p_i) - L(p_i)\right| \\ \text{MSE} = n^{-1}\sum_{i=1}^{n}\left(\hat{L}(p_i) - L(p_i)\right)^2 \\ \text{MAE} = n^{-1}\sum_{i=1}^{n}\left|\hat{L}(p_i) - L(p_i)\right| \end{cases}$$

are used to compare the models, where $n = 99$. Sarabia *et al.* (1999, 2001) also used these three measures in the development of their Lorenz models.

From the MAXABS measure in Table 1, models (14) and (23) are inferior to the others, while model (22) performs best. The MAXABS value is only about 0.02 percent, while the MSE measure is only 0.0067×10^{-6}. The other four models are not distinguishable by MAE. Apart from models (14) and (23), each model is a good global approximation to the data. Their MAXABS values are not larger than 0.07 percent, implying that the error of the estimated Lorenz curve begins to occur at most at the fourth digit after the decimal point. Our estimated Gini indices listed in Table 1 are only slightly different from the empirical Gini provided by Cheong (2002), which is 0.3682. The empirical Gini can be understood as the lower limit of the Gini indices, since it is calculated from the Lorenz curve obtained as the piecewise linear interpolation over the 99 data points.

The numbers which are emboldened in Table 1 are the bootstrapped standard errors with 200 repetitions. We use the re-sampling bootstrapping method to estimate the standard errors, employing the detailed procedure given by Efron and Tibshirani (1993, pp. 45–9). We use a procedure called bootstrapping the pairs. The basic requirement of bootstrapping is that the data re-sampled should be independent and identically distributed. For the grouped data given in (26) and (27), the average income of the income units whose incomes are in $[x_{i-1},x_i]$ is $\mu_i = \mu\Delta l_i/\Delta p_i$, where $\Delta l_i = l(p_i) - l(p_{i-1})$, $\Delta p_i = p_i - p_{i-1}$, and μ is the average income of all the income units. Let $X_i = (\Delta p_i,\mu_i)$.[10] We draw B random samples of size n with replacement from the set $\{1,2, \ldots ,n\}$. Let $\{i_1,i_2, \ldots ,i_n\}$ be such a sample. We then apply the Lorenz curve models to sample $\{X_{i_j}\}_{j=1}^{n}$, and obtain new estimates of the parameters and Gini indices. Finally, the standard errors are computed according to Efron and Tibshirani (1993).

[10] For the United States data, $\Delta p_i = 0.01$ for all $i = 1,2, \ldots ,99$ if setting $p_0 = 0$. Therefore, if setting $\mu = 1$, $X_\text{i} = (\Delta p_i,\Delta l_i)$ can be used.

The performance of model (23) is much poorer if $L_\lambda(p)$ is replaced with the model specified in (3) in the second component of (23), so as to result in a model which nests (15). This implies that (15) does not satisfactorily cope with the data configuration here. Thus, we can conclude that there may be many models created from X which are superior to some traditional models currently in the literature. We do not reproduce our results for the United States income distribution data for 1978–83 in order to conserve space, but these are available on request. Our results for 1978–83 paint a similar picture to those for 1977; i.e. models (18)–(22) perform satisfactorily for almost all the years and these are superior to models (14) and (23).

6.2. *An Application to the Data of Hubei Province, China*

We next use data from the 2006 income survey of Hubei province in China. We use samples from rural and urban Hubei. Hubei is located in central China, with a rural population of 32 million and an urban population of 28.3 million. The rural sample size is 13,232 and the urban sample size is 5317. The survey was conducted by a survey team operating under the auspices of the State Statistical Bureau.

After obtaining the Lorenz curve estimate $\hat{L}(p)$ by using a Lorenz model $L(p)$ to fit the grouped data, we can find a density estimate $\hat{f}(x) = 1/\mu\hat{L}''(p)$ for any x, where p is obtained by solving $\mu\hat{L}'(p) - x = 0$. Given any $x > 0$, the solver denoted by $\hat{F}(x)$ is the estimate of the ratio of population whose income is less than x. Therefore, the relative frequency estimates of any income interval can be calculated. However, it can be difficult to find formulae of close form for the density function with a weighted product model, given the complexity of $\hat{L}'(p)$.

One well-known method to estimate the density function when sample data is available is the kernel method. Assume the sample of size m is $\{y_1, y_2, \ldots, y_m\}$. In our kernel density estimates, the standard normal density function $K(t) = e^{-t/2}/\sqrt{2\pi}$ is used as a kernel function with window width $h = 1.06\hat{\sigma}m^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation (see Silverman, 1998 for an explanation.). The kernel estimate is then

$$\hat{f}(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x - y_i}{h}\right).$$

The kernel method is satisfactory because $\hat{f}(x)$ converges in probability to the true density underlying the sample as $m \to \infty$ under some conditions, but the shape of the density estimate for the given finite sample can vary for different window widths.

We arrange the incomes of the urban and rural samples in increasing order, divide the income range containing all the incomes into equal length intervals, and then form grouped data $\{(p_i, L_i), (p_i, x_i/\mu)\}_{i=1}^{n}$. Since it is common in practice to have grouped data with 10 or so groups, we use data with a small number of groups. The grouped data with 10 groups is used as shown in Table 2. We use the balanced fit to obtain Lorenz estimates where (29) is minimized. We consider $b = 1$, $b = 0.5$, and $b = 0$.

TABLE 2

Grouped Income Data for 2006 Hubei Province, China

Urban			Rural		
Income Range	Income Units	Class Average	Income Range	Income Units	Class Average
0–1999	20	1,524.95	0–899	584	672.06
2000–3999	393	3,231.74	900–1799	1,834	1,401.49
4000–5999	998	5,056.20	1800–2699	3,002	2,239.60
6000–7999	1,192	7,001.41	2700–3599	2,845	3,117.22
8000–9999	877	8,920.33	3600–4499	2,016	4,015.28
10000–11999	553	11,075.19	4500–5399	1,264	4,911.59
12000–13999	478	12,936.17	5400–6299	680	5,803.99
14000–15999	296	14,939.05	6300–7199	390	6,723.93
16000–17999	171	16,971.20	7200–8099	267	7,619.88
18000–19999	130	18,921.53	8100–8999	143	8,540.94
Above 20000	209	25,135.19	Above 9000	207	12,686.77

We present the results in three parts. The first part is a comparison of the performance of some popular traditional Lorenz models in the literature with our models, using urban grouped data. The second part gives density estimates for both the kernel method and for model (22) using data for both rural and urban areas. The third part discusses the implications of the Gini coefficients from the sample.

6.2.1. Comparing Models Using the Balanced Fit Approach

We compare the performance of models (9), (16), and (25) to that of models (19), (22), and (23), since (9) and (16) are among the most well-known in the literature and (25) is a satisfactory model in practice. Table 3 contains the estimated errors of these models in terms of (30). We note that the frequency estimates when $b = 0.5$ are inferior to when $b = 0$, but are better than when $b = 1$ for all the models. Meanwhile, the Lorenz curve estimates when $b = 0.5$ are inferior to when $b = 1$, but are better than when $b = 0$. As b gets smaller, we sacrifice accuracy of the Lorenz curve estimates in exchange for the increased accuracy of the frequency estimates.

There are three observations concerning the performance of models (9) and (16) vis-à-vis the other four models for all three values of b. First, the performance of models (9) and (16) is inferior to that of the other four models, with (9) being slightly better than (16). Second, the frequency estimate errors of (9) and (16) are much larger than that of the other models. Third, in comparison with the other models, (9) and (16) may not be adequate in estimating densities, since the errors of the frequency estimates here are rather large, even if $b = 0$.

With respect to the other four models, we also make three observations. First, the performance of (19) and (23) is very similar, both in terms of the Lorenz curve approximation and frequency estimates. Meanwhile, (25) is only slightly inferior to (19) and (23), and as such has much to commend it because of its simplicity. Second, (22) performs best in all situations. It yields the best estimate of the Lorenz curve when $b = 1$ and the best estimates of frequency when $b = 0$. Furthermore,

TABLE 3

Estimated Errors for the Urban Data of Hubei Province, China

	Lorenz Curve Approximation			Frequency Approximation			
	MSE × 10^5	MAE	MAXABS	MSE × 10^5	MAE	MAXABS	Gini
$b = 1$							
(9)	0.9569	0.0026	0.0050	24.7315	0.0125	0.0283	0.2859
(16)	1.3677	0.0031	0.0060	31.9309	0.0141	0.0318	0.2863
(25)	0.0172	0.0003	0.0009	3.3032	0.0049	0.0111	0.2837
(19)	0.0172	0.0004	0.0008	3.0447	0.0048	0.0104	0.2838
(22)	0.0014	0.0001	0.0002	0.9789	0.0024	0.0073	0.2838
(23)	0.0143	0.0003	0.0007	2.4373	0.0041	0.0101	0.2837
$b = 0.5$							
(9)	2.9235	0.0043	0.0092	15.2387	0.0108	0.0227	0.2895
(16)	4.0576	0.0050	0.0113	19.6440	0.0121	0.0254	0.2905
(25)	0.0367	0.0005	0.0010	2.6617	0.0041	0.0098	0.2834
(19)	0.0479	0.0006	0.0011	2.4282	0.0041	0.0096	0.2838
(22)	0.0322	0.0005	0.0008	0.1673	0.0010	0.0027	0.2832
(23)	0.0378	0.0005	0.0010	2.2536	0.0038	0.0091	0.2838
$b = 0$							
(9)	6.6528	0.0068	0.0129	12.2656	0.0103	0.0194	0.2941
(16)	9.5357	0.0081	0.0155	15.5204	0.0116	0.0214	0.2962
(25)	0.0646	0.0007	0.0015	2.6591	0.0041	0.0100	0.2838
(19)	0.1513	0.0010	0.0020	2.3549	0.0040	0.0096	0.2847
(22)	0.1037	0.0009	0.0015	0.1576	0.0009	0.0027	0.2839
(23)	0.1241	0.0009	0.0017	2.2201	0.0038	0.0090	0.2846

Note: Gini index from sample is 0.2836.

both the Lorenz curve and the frequency estimates of this model are quite satisfactory when $b = 0.5$. Third, observing the MAXABS we find that better performing models like (22) are needed, if better density estimates are considered desirable. Traditional models such as (9), (16), and (25) may not perform satisfactorily in this respect.

6.2.2. Results of Density Estimation

Figure 1 (A) presents our density estimates with model (22) for the urban sample, where the observed frequencies are shown by the background histogram with bin length $h_0 = 2000$. The kernel density estimate for the sample is also given in the figure which exhibits a little upward bias near the origin. The figure shows that the densities corresponding to $b = 1$, $b = 0.5$, and $b = 0$, respectively are close to the kernel density. The bimodality of densities corresponding to $b = 0.5$ and $b = 0$, shown in Figure 1 (A), are reasonable based on histograms with smaller bin length, which are not given to conserve space. The possibility of generating such densities is desirable in practice. Most, if not all, of the probability density functions for modeling the size distribution of income in the literature are unimodal, as noted by Lambert (2001). Figure 1 (B) presents the counterpart estimates to the urban area for the rural sample.

It is difficult to distinguish the densities generated by the two methods for the urban and rural samples from the figures. We list the relative frequency approximations in Table 4 with the same error measures of (22) as listed in Table 3 used

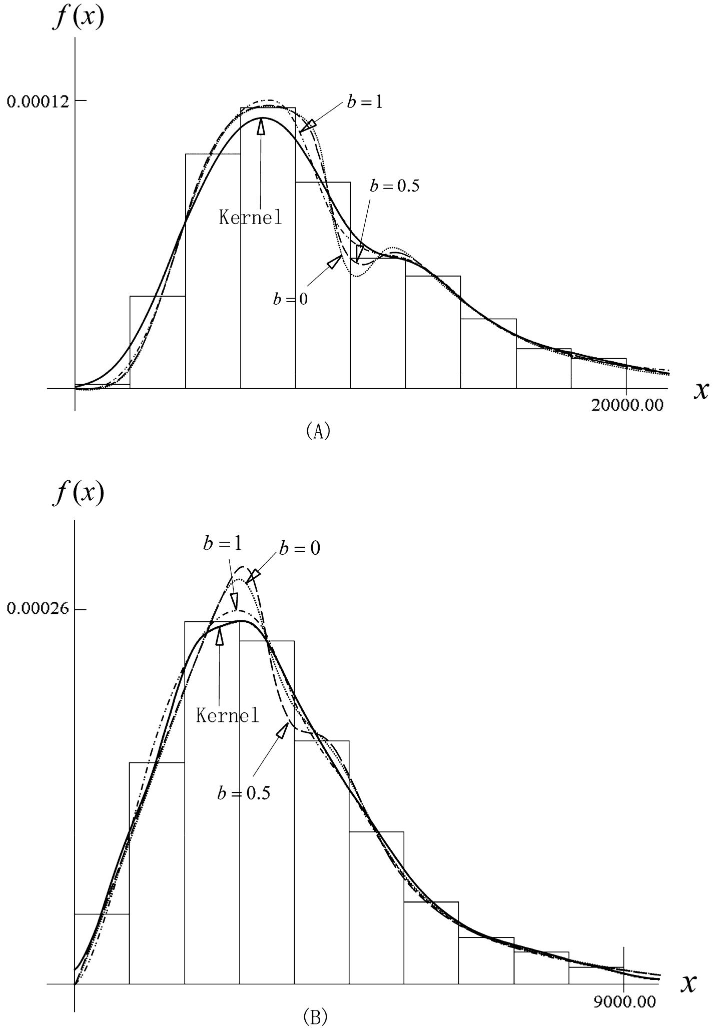

Figure 1. Densities of the urban area (A) and rural area (B) of Hubei Province, China

to facilitate comparison, where $\Delta F_i = p_i - p_{i-1}$ is the observed relative frequency in the income interval $[x_{i-1}, x_i]$. The other four columns are estimated frequencies with the balanced fit and kernel methods, respectively. For both the urban and rural samples, with b decreasing from 1 to 0, the frequency estimates improve for the three balanced fit results in terms of the three error measures. For the rural sample, the kernel density is marginally better than the density estimate of the balanced fit with $b = 1$. But with $b = 0.5$ and $b = 0$, this result is reversed. Compared with the estimates from the kernel method, the frequency estimates from the balanced fit for (22) are closer to the empirical data in terms of the measures considered. Hence, the density estimates with (22) appear plausible for the Hubei income distribution data.

TABLE 4
Relative Frequency Estimation for the 2006 Data of Hubei Province, China

	Actual	Balanced Fit with Model (22)			
	ΔF_i	$b = 1$	$b = 0.5$	$b = 0$	Kernel
Urban:					
	0.0038	0.0059	0.0045	0.0046	0.0129
	0.0739	0.0727	0.0725	0.0727	0.0804
	0.1877	0.1865	0.1882	0.1882	0.1758
	0.2242	0.2255	0.2236	0.2240	**0.2117**
	0.1649	**0.1576**	0.1650	0.1650	0.1629
	0.1040	0.1091	0.1043	0.1041	0.1106
	0.0899	0.0902	0.0895	0.0896	0.0897
	0.0557	0.0558	0.0558	0.0555	0.0562
	0.0322	0.0335	0.0342	0.0341	0.0349
	0.0244	0.0224	**0.0217**	**0.0217**	0.0241
	0.0393	0.0407	0.0407	0.0406	0.0411
MSE × 10^5		0.9789	0.1673	0.1576	4.8196
MAE		0.0024	0.0010	0.0009	0.0054
MAXABS		0.0073	0.0027	0.0027	0.0125
Rural:					
	0.0441	0.0388	0.0448	0.0441	0.0478
	0.1386	**0.1515**	0.1388	0.1386	0.1436
	0.2269	0.2225	0.2270	0.2269	**0.2188**
	0.2150	0.2149	0.2149	0.2150	0.2106
	0.1524	0.1490	0.1520	0.1523	0.1503
	0.0955	0.0955	0.0961	0.0957	0.0964
	0.0514	0.0516	0.0501	0.0509	0.0529
	0.0295	0.0300	0.0300	0.0302	0.0303
	0.0202	0.0189	0.0195	**0.0192**	0.0200
	0.0108	0.0118	**0.0122**	0.0118	0.0107
	0.0156	0.0156	0.0147	0.0152	0.0186
MSE × 10^5		2.2934	0.0621	0.0290	1.4081
MAE		0.0029	0.0007	0.0004	0.0030
MAXABS		0.0129	0.0014	0.0010	0.0081

Note: Characters in bold indicate where the MAXABS turns up.

Table 5 provides estimates of Lorenz curves for the rural and urban samples. The results show that all three balanced fit estimates are better than those with the kernel method, where the same error measures of (22) as displayed in Table 3 are repeated for convenience of comparison. The estimates of the balanced fit with $b = 1$ are closest to the empirical values. When b decreases from $b = 1$ to $b = 0$, the error measures get larger. The Gini indices generated from the two methods are very close to each other. But the Gini estimates of the kernel method have a little upward bias. Therefore, the estimates from the Lorenz model, which uses much less information, seem to be better than those of the kernel method in the estimation of Lorenz curves. Based on the results reported in Tables 4 and 5, the estimated Lorenz curves for the rural and urban areas produced by the balanced fit with $b = 0.5$ and $b = 0$ yield a global approximation to the empirical data both in Lorenz curve values and relative frequencies. The parameter estimation of models for the 2006 data of Hubei province, China, is in Table A2 in the online Appendix 3.

TABLE 5

Lorenz Estimates for the 2006 Data of Hubei Province, China

		Actual	Balanced Fit with Model (22)			
	p	$L(p)$	$b = 1$	$b = 0.5$	$b = 0$	Kernel
Urban:						
	0.0038	0.0006	0.0006	0.0006	0.0006	0.0003
	0.0777	0.0261	**0.0259**	0.0264	0.0264	0.0230
	0.2654	0.1273	0.1274	0.1281	0.1279	**0.1222**
	0.4896	0.2947	0.2946	**0.2955**	0.2953	0.2899
	0.6545	0.4516	0.4517	0.4515	0.4508	0.4491
	0.7585	0.5744	0.5742	0.5738	0.5733	0.5722
	0.8484	0.6984	0.6985	0.6979	0.6972	0.6973
	0.9041	0.7871	0.7872	0.7866	0.7859	0.7867
	0.9362	0.8453	0.8452	0.8446	**0.8438**	0.8453
	0.9607	0.8946	0.8947	0.8940	0.8931	0.8951
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
MSE × 10^5			0.0014	0.0322	0.1037	0.7069
MAE			0.0001	0.0005	0.0009	0.0020
MAXABS			0.0002	0.0008	0.0015	0.0051
Gini		0.2836*	0.2838	0.2832	0.2839	0.2907
Rural:						
	0.0441	0.0087	0.0087	0.0077	0.0078	0.0074
	0.1827	0.0654	0.0654	**0.0643**	0.0645	0.0627
	0.4096	0.2138	0.2138	0.2148	0.2152	0.2110
	0.6246	0.4095	0.4095	0.4097	0.4110	0.4082
	0.7770	0.5882	0.5882	0.5891	**0.5903**	0.5883
	0.8725	0.7252	0.7252	0.7256	0.7270	0.7268
	0.9239	0.8123	**0.8124**	0.8127	0.8142	0.8149
	0.9534	0.8702	0.8701	0.8705	0.8719	0.8735
	0.9735	0.9151	0.9151	0.9154	0.9168	0.9192
	0.9844	0.9420	0.9420	0.9422	0.9437	**0.9469**
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
MSE × 10^5			0.0001	0.0478	0.2547	0.7905
MAE			0.0000	0.0006	0.0015	0.0025
MAXABS			0.0001	0.0011	0.0021	0.0048
Gini		0.3063*	0.3064	0.3061	0.3045	0.3124

Note: Characters in bold indicate where the MAXABS turns up for each model. Gini indices with asterisks are calculated from the samples.

6.2.3. Results for the Gini Coefficients

The Gini coefficients for urban and rural Hubei for 2006 calculated from the sample are 0.2836 and 0.3063, respectively. These Gini coefficients are fairly similar in magnitude to those reported in previous studies that have examined income inequality in China using grouped data (see, e.g. Bramall, 2001; Chotikapanich *et al.*, 2007; Wang *et al.*, 2009). In particular, consistent with the results in Chotikapanich *et al.* (2007), we find that inequality in rural China has been higher than in urban China. This result is significant for the following reasons. First, the Chinese government has allocated a considerable amount of funds to support poverty reduction each year since 1986 (see Park *et al.*, 2002). However, several studies suggest that high rural income inequality has undermined the Chinese government's attempts to reduce poverty in China (see, e.g. Ravallion and Chen, 2007). Second, rural income inequality has been a particular concern of the

Hu-Wen administration who are concerned about the potential adverse effects on China's economic growth. There is a voluminous theoretical and empirical literature on the effects of income inequality on growth. From a theoretical perspective there are arguments going both ways (Wan *et al.*, 2006). The empirical evidence on the effect of inequality on growth for a range of countries and empirical specifications has been mixed (Banerjee and Duflo, 2003). The only study that examines this issue for rural China is Ravallion and Chen (2007). Their findings point to a negative effect of inequality on growth. In particular, they find that periods of most rapid growth were not associated with more rapid increases in inequality, while periods of falling inequality had the highest growth in household income.

This said, while income inequality in urban China has increased dramatically in recent years, most of the growth in income inequality in rural China dates to the 1980s (see Wang *et al.*, 2009). The emergence of the non-agricultural sector in the 1980s and first half of the 1990s, particularly the collective township and village enterprise (CTVE, *xiang-zhen qiye*) sector, changed the composition of rural income and generated higher inequality. Decollectivization gave rural households more discretion in their production decisions. With this new found freedom and the small land-to-person ratio available in many rural areas, it was natural for rural labor to move into CTVEs. The emergence of CTVEs was also related to fiscal decentralization. Fiscal decentralization placed pressure on local governments to raise revenue and sub-provincial governments invested in CTVEs, the taxes from which became an important source of revenue. Chotikapanich *et al.* (2007) found that rural income inequality starts to stabilize from the mid-1990s, following the increase in rural income inequality in the 1980s, and that it has largely plateaued since 2003. One explanation for stabilization in rural income inequality is that since the mid-1990s participation rates in non-farm activity among low-income rural households increased, resulting in a more equal distribution of income (Zhu and Luo, 2006).

7. Conclusion

We have presented a general method for creating Lorenz models of the weighted-product form. The method rests on finding a set of parametric Lorenz models. We find that an ideal situation is when, for each element of the set, the ratio of its second derivative to its first derivative is increasing. We have presented a set X of Lorenz models which possess this property. Hence, we can create a large number of parametric Lorenz models. Moreover, our results provide evidence that we can have models with good global approximation to the actual data. Since all the models developed satisfy the definition of the Lorenz curve, they can be used to generate underlying densities. We also introduced the concept of balanced fit. The balanced fit approach provides a means of assigning weights when developing the Lorenz curve according to whether the practitioner wants to put more emphasis on using the Lorenz curve as an overall measure of inequality or as a targeted poverty index.

To illustrate the performance of our models and the concept of balanced fit, we used data on income distribution for the United States for 1977–83, previously used by Basmann *et al.* (1993) as well as income survey data collected by the State

Statistical Bureau in Hubei province in China in 2006. We use Chinese data to illustrate our models, given the pressing policy importance of income inequality in China and the associated urgent need for further advancements in measuring income inequality using grouped data. We find that our proposed models perform well, particularly compared to popular existing Lorenz models in the literature, and that our approach generates plausible density estimates. Our results suggest that rural income inequality in China has been higher than urban income inequality. We have discussed some of the reasons for, as well as implications of, high rural income inequality in China.

The most significant feature of our method is that we can increase the power of the method by increasing the set X found. This could be an interesting topic for further research. Another further research subject could be finding methods to determine the most favorable model/models among the ones created in X. It could be possible to find completely new sets, with elements possessing the same property as the models in X, so as to obtain other sets of weighted-product Lorenz models. Given that the main objective of this study was to examine the feasibility of using new methods to fit income distributions to grouped data, we have not examined a number of interesting aspects of income inequality and poverty in China. Specifically, we have used data from one province for a single year. Future research could apply the balanced fit approach developed in this study to a broader cross-section of data as well as examine trends in income inequality and poverty over time.

References

Banerjee, A. and E. Duflo, "Inequality and Growth. What Can the Data Say?" *Journal of Economic Growth*, 8, 267–99, 2003.

Basmann, R. L., K. J. Hayes, D. J. Slottje, and J. D. Johnson, "A General Functional Form for Approximating the Lorenz Curve," *Journal of Econometrics*, 43, 77–90, 1990.

Basmann, R. L., K. J. Hayes, and D. J. Slottje, *Some New Methods for Measuring and Describing Economic Inequality*, JAI Press, Greenwich, CT, 1993.

Bramall, C., "The Quality of China's Household Income Surveys," *The China Quarterly*, 167, 689–705, 2001.

Cheong, K. S., "An Empirical Comparison of Alternative Functional Forms for the Lorenz Curve," *Applied Economics Letters*, 9, 171–6, 2002.

Chotikapanich, D., "A Comparison of Alternative Functional Forms for the Lorenz Curve," *Economics Letters*, 3, 187–92, 1993.

Chotikapanich, D., D. S. P. Rao, and K. K. Tang, "Estimating Income Inequality in China Using Grouped Data and the Generalized Beta Distribution," *Review of Income and Wealth*, 53, 127–47, 2007.

Cowell, F. A. and F. Mehta, "The Estimation and Interpolation of Inequality Measures," *Review of Economic Studies*, 49, 273–90, 1982.

Efron, B. and R. J. Tibshirani, *An Introduction to Bootstrapping*, Chapman and Hall, New York, 1993.

Gupta, M. R., "Functional Form for Estimating the Lorenz Curve," *Econometrica*, 52, 1313–4, 1984.

Kakwani, N. C., "On the Estimation of Income Inequality Measures from Grouped Observations," *Review of Economic Studies*, 43, 483–92, 1976.

———, "On a Class of Poverty Measures," *Econometrica*, 48, 437–46, 1980.

Kakwani, N. C. and N. Podder, "Efficient Estimation of Lorenz Curves and Associated Inequality Measures from Grouped Observations," *Econometrica*, 41, 137–48, 1976.

Lambert, P. J., *The Distribution and Redistribution of Income*, Manchester University Press, Manchester, 2001.

Meng, X., "Economic Restrictions and Income Inequality in Urban China," *Review of Income and Wealth*, 56, 357–79, 2004.

Ogwang, T. and U. L. G. Rao, "A New Functional Form for Approximating the Lorenz Curve," *Economics Letters*, 52, 21–9, 1996.

———, "Hybrid Models of the Lorenz Curve," *Economics Letters*, 69, 39–44, 2000.

Ortega, P., G. Martin, A. Fernandez, M. Ladoux, and A. Garcia, "A New Functional Form for Estimating Lorenz Curves," *Review of Income and Wealth*, 37, 447–52, 1991.

Park, A., S. Wang, and G. Wu, "Regional Poverty Targeting in China," *Journal of Public Economics*, 86, 123–53, 2002.

Pratt, J. W., "Risk Aversion in the Small and in the Large," *Econometrica*, 32, 122–36, 1964.

Rasche, R. H., J. Gaffney, A. Y. C. Koo, and N. Obst, "Functional Forms for Estimating the Lorenz Curve," *Econometrica*, 48, 1061–2, 1980.

Ravallion, M. and S. Chen, "China's (Uneven) Progress Against Poverty," *Journal of Development Economics*, 82, 1–42, 2007.

Ryu, H. K. and D. J. Slottje, "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve," *Journal of Econometrics*, 72, 251–74, 1996.

Sarabia, J., E. Castillo, and D. J. Slottje, "An Ordered Family of Lorenz Curves," *Journal of Econometrics*, 91, 43–60, 1999.

———, "An Exponential Family of Lorenz Curves," *Southern Economic Journal*, 67, 748–56, 2001.

Schader, M. and F. Schmid, "Fitting Parametric Lorenz Curves to Grouped Income Distribution—A Critical Note," *Empirical Economics*, 19, 361–70, 1994.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1998.

Wan, G. and Z. Zhou, "Income Inequality in Rural China: Regression-Based Decomposition Using Household Data," *Review of Development Economics*, 9, 107–20, 2005.

Wan, G., M. Lu, and Z. Chen, "The Inequality-Growth Nexus in the Short and Long Run: Empirical Evidence from China," *Journal of Comparative Economics*, 34, 654–67, 2006.

Wang, Z. X. and R. Smyth, "Two New Exponential Families of Lorenz Curves," Discussion Paper No. 20/07, Department of Economics, Monash University, 2007.

Wang, Z. X., Y-K. Ng, and R. Smyth, "Revisiting the Ordered Family of Lorenz Curves," Discussion Paper No. 19/07, Department of Economics, Monash University, 2007.

Wang, Z. X., R. Smyth, and Y-K. Ng, "A New Ordered Family of Lorenz Curves with an Application to Measuring Income Inequality and Poverty in Rural China," *China Economic Review*, 20, 218–35, 2009.

Zhu, N. and X. Luo, "Nonfarm Activity and Rural Income Inequality: A Case Study of Two Provinces in China," World Bank Policy Research Working Paper 3811, 2006.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix 1: Proof of Lemma 3.

Appendix 2: Proof of Theorem 2.

Appendix 3: Parameter Estimations for U.S. 1977 Data and for the 2006 Data of Hubei Province, China.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.