Parametric Modelling of Income Distributions

Econometric Seminar, LS 2026

Michal Kuba

MFF CUNI

March 4, 2026

Roadmap

Our goal

- We model income distribution across the United States and Central/Eastern Europe
- We make use of the **generalized beta distribution**
- Estimate the distribution over the years and from different-sized households to infer systematic differences

Why parametric income distributions?

- Summarize entire income distribution with few parameters
- Provide analytically tractable inequality/poverty measures (e.g., Gini, GE indices).
- Compare countries/years and test distributional change (e.g., Lorenz dominance).

Two papers, two angles

- **Łukasiewicz & Orłowski (2004):** fit models to *individual incomes*; derive *two-earner household* distribution via convolution; compare USA vs Poland.
- **Brzeziński (2013):** fit **GB2** to CEE disposable incomes over time; model selection via likelihood ratio; inequality/polarization and Lorenz dominance inference.

Data format

- **Individual/one-earner income** X vs. **two-earner household income** $Z = X + Y$ (L&O 2004).
- **Equivalized disposable household income** (post-tax/post-transfer) with weights to obtain personal distributions (Brz. 2013).

Notation

- Income random variable $X \in \mathbb{R}_+$ with density $f_X(x)$ and CDF $F_X(x)$.
- Samples $\{x_i\}_{i=1}^n$ (possibly weighted) used for MLE.
- **Two-earner household:** $Z = X + Y$ where X, Y are individual incomes.

Convolution

If $X \perp Y$ with densities f_X, f_Y, then

$$f_Z(z) = (f_X * f_Y)(z) = \int_{\mathbb{R}} f_X(t)\, f_Y(z-t)\, dt.$$

Models used in Łukasiewicz & Orłowski (2004)

Candidate densities for individual incomes:

Dagum

$$f_D(x) = \frac{abc\, x^{b-1}}{(1 + ax^{-b})^{c+1}}, \quad a > 1,\ b > 0,\ c > 0.$$

Singh–Maddala

$$f_{SM}(x) = \frac{abc\, x^{b-1}}{(1 + ax^{b})^{c+1}}, \quad a > 0,\ b > 0,\ c > 0,\ bc > 1.$$

Exponential & Weibull

$$f_E(x) = \frac{1}{a} e^{-x/a},\ a > 0, \qquad f_W(x) = \frac{b}{a} x^{b-1} \exp\left(-\frac{x^b}{a}\right),\ a, b > 0.$$

GB2 as a unifying family (Brzeziński 2013)

Generalized Beta of the Second Kind (GB2)

Four-parameter model with scale b and shapes $a, p, q > 0$:

$$f_{\mathsf{GB2}}(x; a, b, p, q) = \frac{a\, x^{ap-1}}{b^{ap}\, B(p, q)\, \left(1 + (x/b)^a\right)^{p+q}}, \quad x > 0.$$

- b: scale (linked to mean via moments when they exist).
- a: overall tail-thickness control; p affects left tail; q affects right tail.

Nested models

- Singh–Maddala: GB2 with $p = 1$.
- Dagum: GB2 with $q = 1$.

Estimation: Maximum Likelihood

- Fit parameters $\hat{\theta}$ by maximizing $\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$ (or weighted MLE in survey settings).
- Use of some numerical optimization methods.
- The goal is to find a good balance between fit and complexity via likelihood ratio tests.

Fit diagnostics in Łukasiewicz & Orłowski (2004)

They evaluate fit using error summaries between empirical and model-implied bin probabilities:

$$\mathrm{SSE} = \sum_{i=1}^{k} \left(\frac{n_i}{n} - p_i(\hat{y}) \right)^2, \qquad \mathrm{SAE} = \sum_{i=1}^{k} \left| \frac{n_i}{n} - p_i(\hat{y}) \right|.$$

Also report agreement measures (e.g., W_r, W_p) and quantile agreement via correlation r^2.

Table 1: Individual incomes in the USA (2000)

Model	$\hat{a}$	$\hat{b}$	c	SSE	SAE	W_r (%)	W_p (%)
Exponential	28.591	—	—	0.0008	0.1268	93.66	93.63
Weibull	26.535	0.9806	—	0.0010	0.1392	93.04	92.97
Dagum	29474.0	2.7434	0.3138	0.0003	0.0709	96.45	96.17
Singh–Maddala	0.0053	1.1228	5.1245	0.0005	0.0899	95.50	95.67

Results of estimation of parameters and adjustment of models to empirical data (USA, 2000).

Table 2: Individual incomes in Poland (2000)

Model	$\hat{a}$	$\hat{b}$	c	SSE	SAE	W_r (%)	W_p (%)
Dagum	12.7312	2.8498	1.4202	0.0008	0.0992	95.04	95.04
Singh–Maddala	0.0369	1.6582	0.6955	0.0010	0.1039	94.81	94.80

Results of estimation of parameters and adjustment of models to empirical data (Poland, 2000).

Fit & selection in Brzeziński (2013)

- **Goodness-of-fit:** q–q plots + numerical comparison of implied vs empirical statistics.
- **Model selection (nested):** likelihood ratio test

$$LR = 2(\hat{\ell}_{\mathsf{GB2}} - \hat{\ell}_{\mathsf{nested}}) \ \sim \ \chi^2_{(1)}.$$

- Note: large samples often reject all models in strict EDF-based tests; hence emphasis on graphical/numerical checks.

From individuals to two-earner households

Assume two earners with incomes X and Y:

$$Z = X + Y.$$

If independence holds

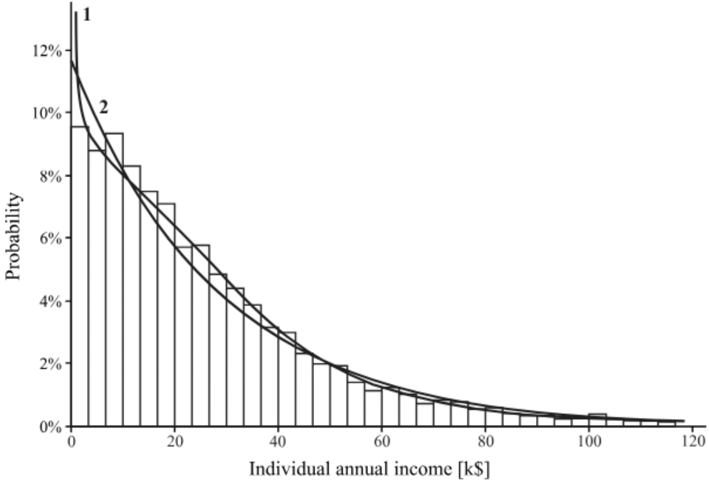
Compute f_Z as convolution of fitted individual-income distributions.

Interpretation: Good fit of f_Z to empirical two-earner distribution supports (approx.) statistical independence of earners' incomes.

If independence fails

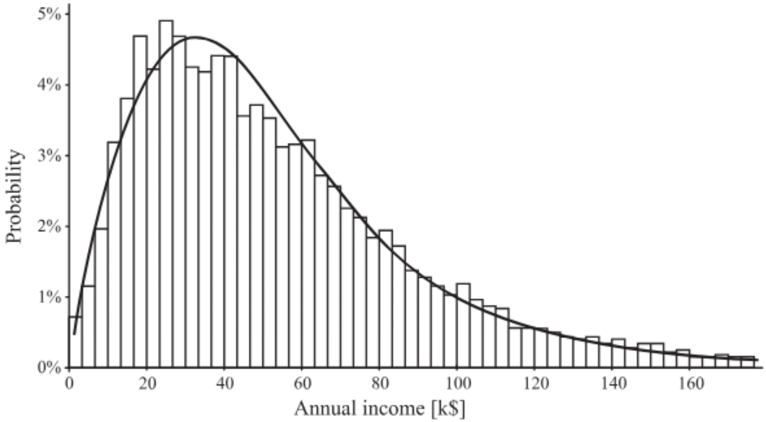
Mismatch between convolution and data suggests dependence (e.g., assortative matching, labor market structure, household formation effects).

Individual earnings - USA



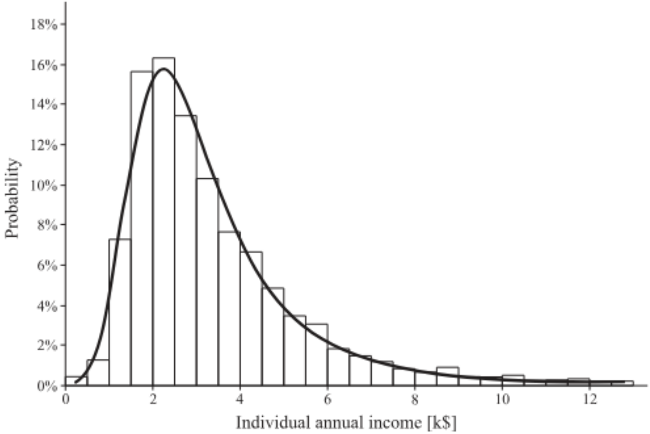
US single income distribution, taken from [1].

Double income earnings - USA



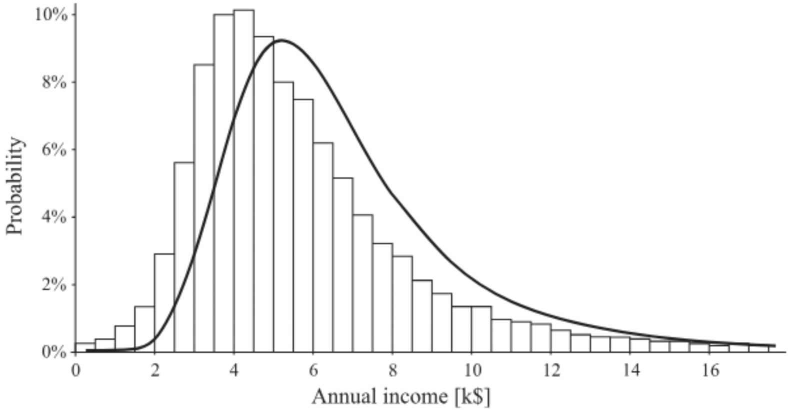
US dual income distribution with covolution, taken from [1].

Individual earnings - Poland



Polish single income distribution, taken from [1].

Double income earnings - Poland



Polish dual income distribution with covolution, taken from [1].

Key empirical result (Ł&O 2004)

- Among candidates, **Dagum** fits individual incomes best (USA and Poland, year 2000).
- **USA:** convolution of two Dagum-fitted individual distributions matches two-earner household income well $\Rightarrow$ earners approximately independent.
- **Poland:** convolution fit is much worse $\Rightarrow$ strong dependence between earners' incomes.

Lorenz curve: definition and intuition

Let $X \geq 0$ denote income with CDF F, quantile function $Q(u) = F^{-1}(u)$, and mean $\mu = \mathbb{E}[X]$.

Lorenz curve

$$L(p) = \frac{1}{\mu} \int_0^p Q(u)\, du, \qquad p \in [0, 1].$$

- $p =$ population share (poorest $100p\%$).
- $L(p) =$ income share owned by the poorest $100p\%$.
- $L(0) = 0$, $L(1) = 1$, L is increasing and convex.
- Perfect equality: $L(p) = p$. More inequality $\Rightarrow$ Lorenz curve lies further below the 45° line.

Lorenz dominance: ordering distributions

Consider two income distributions A and B with Lorenz curves L_A and L_B.

Definition (Lorenz dominance)

Distribution B *Lorenz-dominates* A (write $B \succeq_L A$) if

$$L_B(p) \;\geq\; L_A(p) \quad \text{for all } p \in [0,1],$$

and strict inequality holds for some p.

- If $B \succeq_L A$, then B is (weakly) **more equal** than A.
- Interpretation: for *every* bottom-p group, the cumulative income share is at least as large in B.

Why Lorenz dominance is powerful

- Lorenz dominance is a **robust** inequality ranking: it does not require choosing a specific inequality index (Gini, Theil, GE(α), Atkinson, ...).
- If $B \succeq_L A$, then **all** inequality measures consistent with the Pigou–Dalton transfer principle rank B as less unequal than A.
- If Lorenz curves **cross**, there is **no** Lorenz dominance: different inequality indices may disagree.

Tail behavior and (bi-)polarization intuition

For GB2:

- Left tail thickness controlled by ap (smaller $\Rightarrow$ fatter left tail).
- Right tail thickness controlled by aq (smaller $\Rightarrow$ fatter right tail).
- Simultaneous decreases in ap and aq can be interpreted as fatter tails on both ends (related to income *bi-polarization*).

Lorenz dominance in parametric families (GB2 context)

In Brzeziński (2013), each year–country distribution is fitted parametrically (often GB2), and Lorenz dominance is assessed via parameter restrictions.

Sufficient conditions used for GB2 comparisons

For two GB2 distributions (1) and (2), if

$$a_1 \leq a_2, \qquad a_1 p_1 \leq a_2 p_2, \qquad a_1 q_1 \leq a_2 q_2,$$

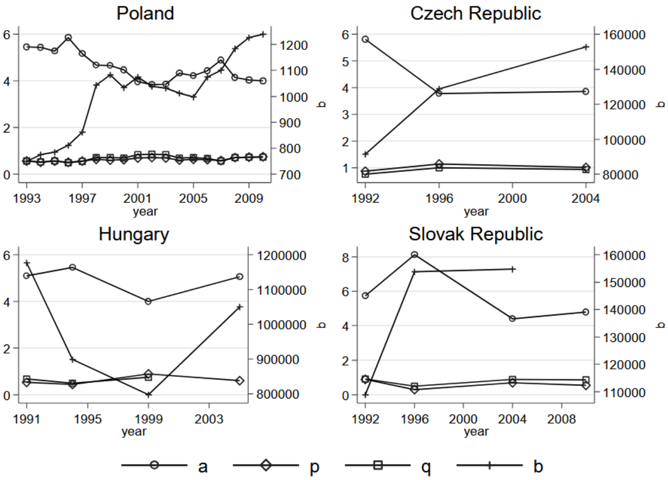
then (2) Lorenz-dominates (1), i.e. distribution (2) is less unequal.

- We therefore estimate (a, p, q, b) by MLE each year, then test these inequalities statistically (e.g., Wald/chi-square-type tests).
- If conditions fail and Lorenz curves plausibly cross, avoid definitive dominance claims.

Key empirical results (Brzeziński 2013)

- Countries: Czech Rep., Hungary, Poland, Slovak Rep. (1990s–2000s).
- **Model fit:** GB2 often best overall; especially strong for **Poland** (and largely for **Hungary**).
- **Dagum vs GB2:** for Czech and Slovak, **Dagum** frequently performs as well as GB2 $\Rightarrow$ pragmatic preference for simpler form.
- **Inequality dynamics:**
  - Czech, Poland, Slovak: Lorenz dominance indicates **inequality increased** over transformation period.
  - Hungary: no Lorenz dominance; direction depends on inequality measure.
- **Tails:** evidence of fatter tails over time in Czech/Poland/Slovak (interpreted as growing bi-polarization).

Parameter evolution - Central Europe



GB2 parameter evolution over a span of years, taken from [2].

Conclusion

- Dagum/GB2 families provide strong empirical fits in income distributions.
- Convolution approach is a good, direct empirical test of (in)dependence between earners.
- GB2 framework supports systematic cross-country/time comparisons and inequality inference.

References I

P. Łukasiewicz and A. Orłowski.
Probabilistic models of income distributions.
Physica A 344 (2004) 146–151.

M. Brzeziński.
Parametric modelling of income distribution in Central and Eastern Europe.
Working Papers No. 31/2013 (116), Faculty of Economic Sciences, University of Warsaw, 2013.