Chasing the Tail: A Generalized Pareto Distribution Approach to Estimating Wealth Inequality

OeNB

11 November 2022

Arthur Kennickell
Stone Center, CUNY Graduate Center

Acknowledgements

Thanks to the Stone Center at CUNY for early computer access and other support.

Thanks for comments from Pirmin Fessler, Bertrand Garbinti, Andrea Neri, Martin Schürz and Philip Vermeulen. I alone am responsible for any errors or oversights.

Motivation

- Kennickell (2019): Even small difference in effective coverage of the upper tail of the wealth distribution in a survey can yield large biases in estimates of inequality (and even the precision on the estimates)
- Comparisons across surveys may be even worse, if there are differences in effective coverage
- Vermeulen [2018] introduced idea of using “rich lists” to supplement survey data, for purposes of estimating a Pareto approximation of the upper tail
    - A positive step, but rich lists are generally opaque in their construction and replete with possibilities for error
    - Overstated/understated wealth, family vs. individual, actual country of residence, etc.
    - Typically, approach uses only extreme observed tail in estimation
    - By now, extensive application of this method
- Can we forge a principled and more flexible alternative?

A very brief nod at the literature

- By now, a large literature using Pareto methods with rich list data
  - Results appear highly dependent on such opaque data
- Used in work on distributional national accounts
  - But details matter(!): e.g., paper with Peter and Martin on DFA
- Much less attention to the generalized Pareto distribution (GPD)
  - Recent work for Austria by Ines Heck, Jakob Kapeller and Rafael Wildauer using Austrian data incorporating rich list data with GPD
    - Alas for me, only available now in German, so I have been unable to read it

- This paper focuses entirely on a GPD approach
- (Note: Ignoring all other sources of measurement bias besides tail bias)

Preview of results

- Survey data generally understate aggregate wealth
- Generalized Pareto, as implemented here allows possibility of:
  1. Better fitting of observed data as result of additional parameter
  2. Treating some data as unrepresentative
  3. Explicitly addressing an unobserved right tail
  4. Constraining parameter estimates to reproduce an aggregate total
  5. *[Directly introducing data from "rich lists" (<u>not done here</u>)]*
- Only #4 is effective in closing the gap with aggregate data
  - For Austria, yields much higher level of wealth concentration

The data

- Wealth data from 2017 HFCS for Austria and 2016 SCF for US
- Many conceptual similarities
- Principal difference for current purpose is effective coverage of the upper tail
  - HFCN:AT lacks a means of explicitly sampling the upper tail
    - 28 observations represent top 1%
  - SCF uses transformation of tax data to sample the upper tail and perform post-survey adjustments
    - 563 observations represent top 1%
    - Already explains the great majority of aggregate wealth
- **NOTE:** For most estimates, only first implicate is used

Descriptive statistics for 2016 HFCS:AT and 2017 SCF

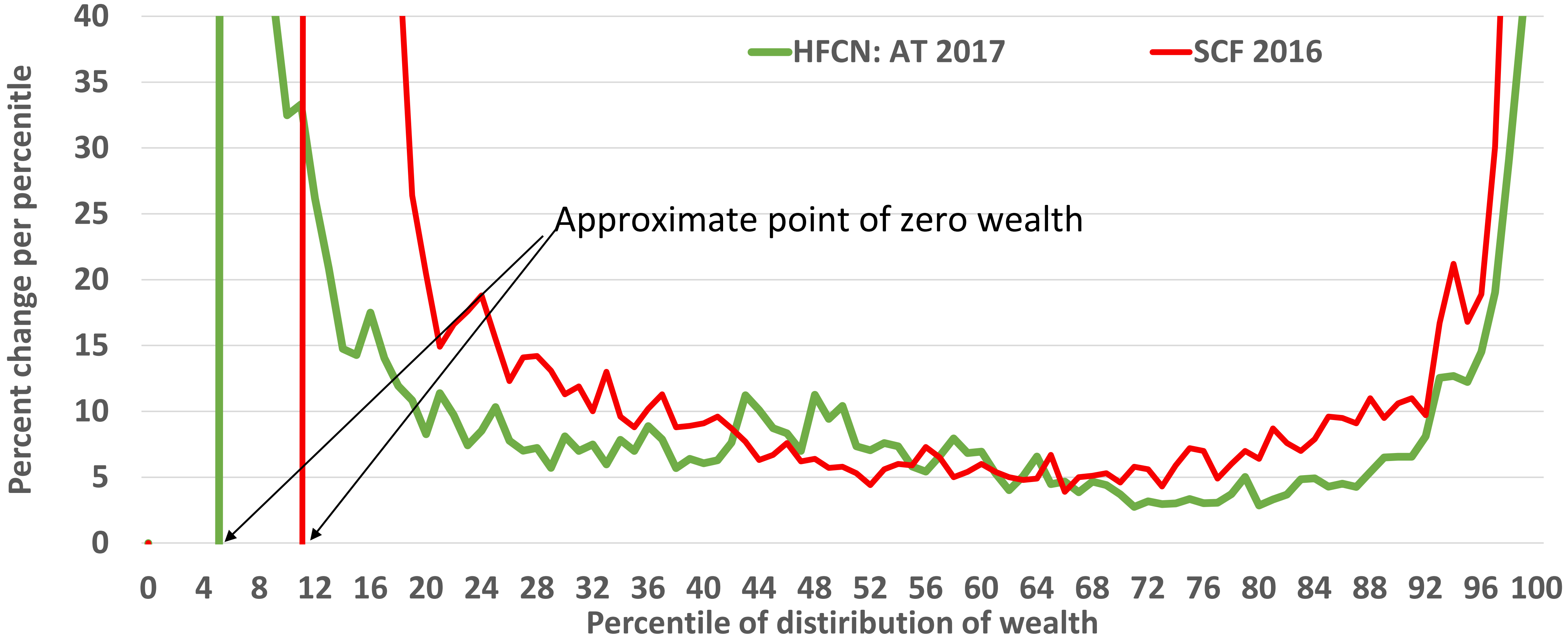
Item	2017 Austrian HFCS		2016 SCF	
Mean	237500		689200	
	11200	18400	7100	12700
Median	74800		97300	
	1180	4040	422	2700
Share top 1%	22.9		38.6	
	3.01	4.65	0.16	0.68
Gini coefficient	0.723		0.850	
	0.011	0.019	0.001	0.003
P90-P25 ratio	44.9		116.8	
	0.66	2.37	1.75	4.86
Number of observations	3072		6248	
"Population mean"	332692		708536	

Mean and median figures are given in home currency in each case: 2 Jan 2017 exchange rate 1 Euro = 1.05 USD.

Str error wrt IMPUTATION

Std error wrt IMPUTATION AND SAMPLING

Percent change in net worth per percentile



Estimation approach here

- Use generalized Pareto distribution (GPD): **ASSUMPTION!**
  - Flexibility of additional parameter beyond simple Pareto offers hope of integrating better with more than just the extreme tail
- Reinterpretation of Castillo and Hadi [1997] estimation method to apply to survey data
- Extension of method to allow for
  - Errors in regions of data
    - *(Note: Ignoring other sources of reporting error)*
  - Effective undercoverage at the top of the wealth distribution
  - Incorporation of external aggregate as a constraint on the estimation

Generalized Pareto distribution (GPD)

$$F(\mathrm{V}) = F(\lambda - \lambda^0 | \lambda > \lambda^0 > 0) = 1 - \left(1 - \frac{k(\lambda - \lambda^0)}{\sigma}\right)^{\frac{1}{k}}$$

V: vector of wealth values

λ^0: "location parameter": value above which GPD taken to apply

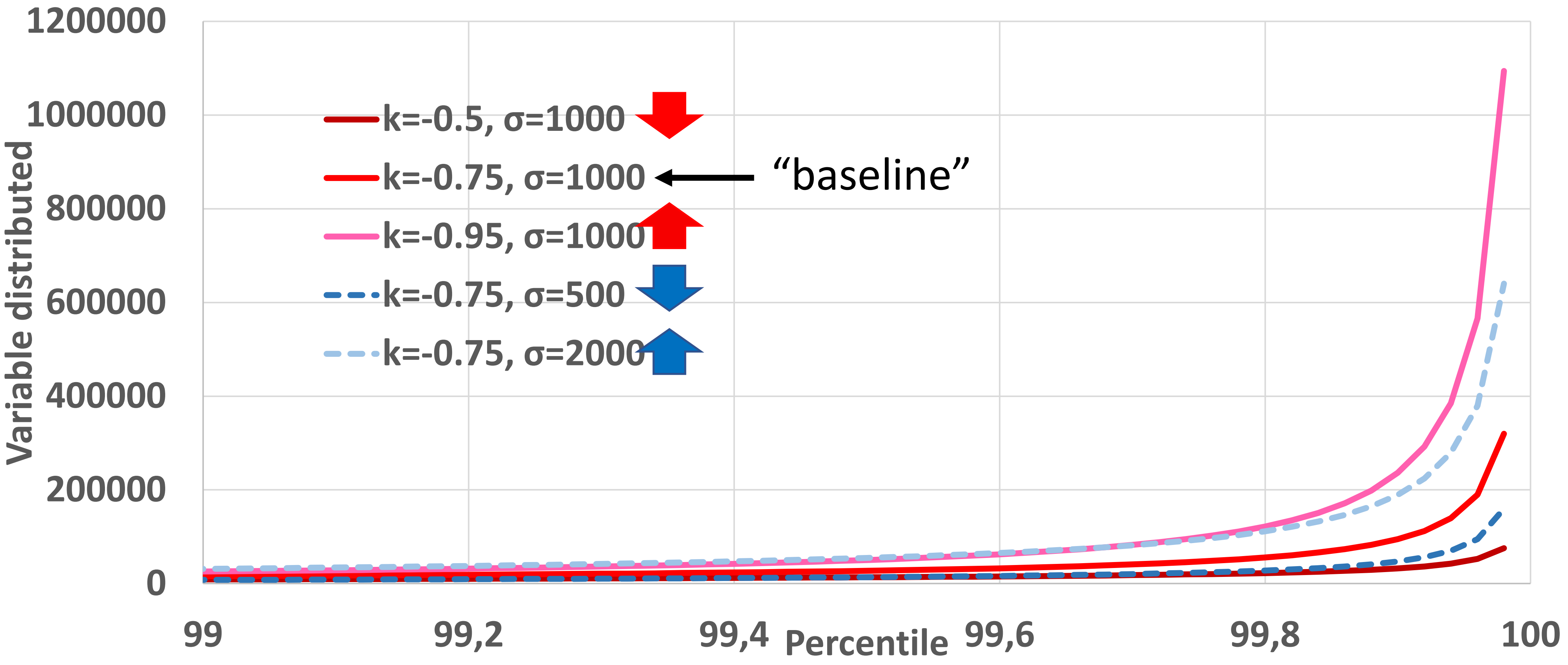
k: "shape parameter"

σ: "scale parameter"

$\lambda - \lambda^0$: vector of "exceedances"

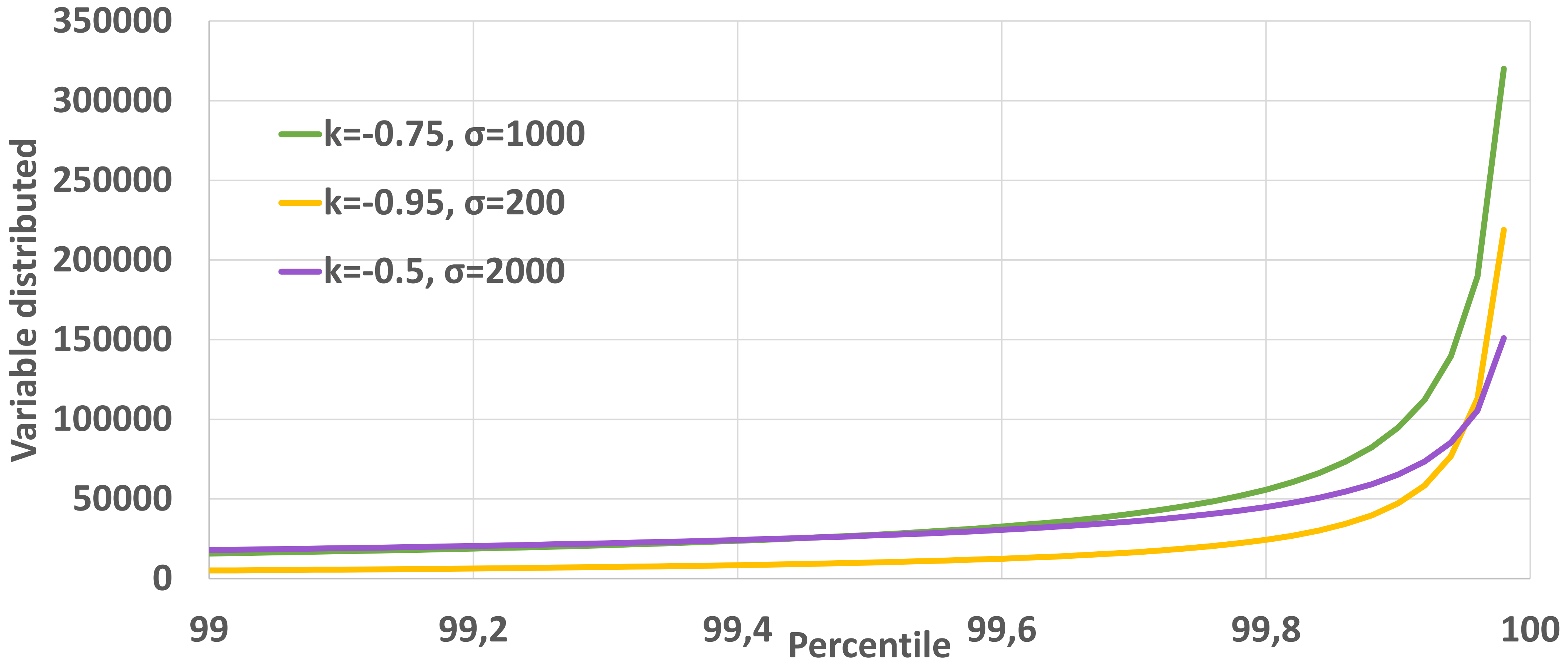
Simple Pareto (F$(\lambda) = 1 - \left({\lambda^0}/{\lambda}\right)^{\alpha}$*) is a special case of GPD*

GPD for various parameter values



For simplicity, taking λ^0 =0, and focusing on only the top 1%

Selected GPDs with same mean



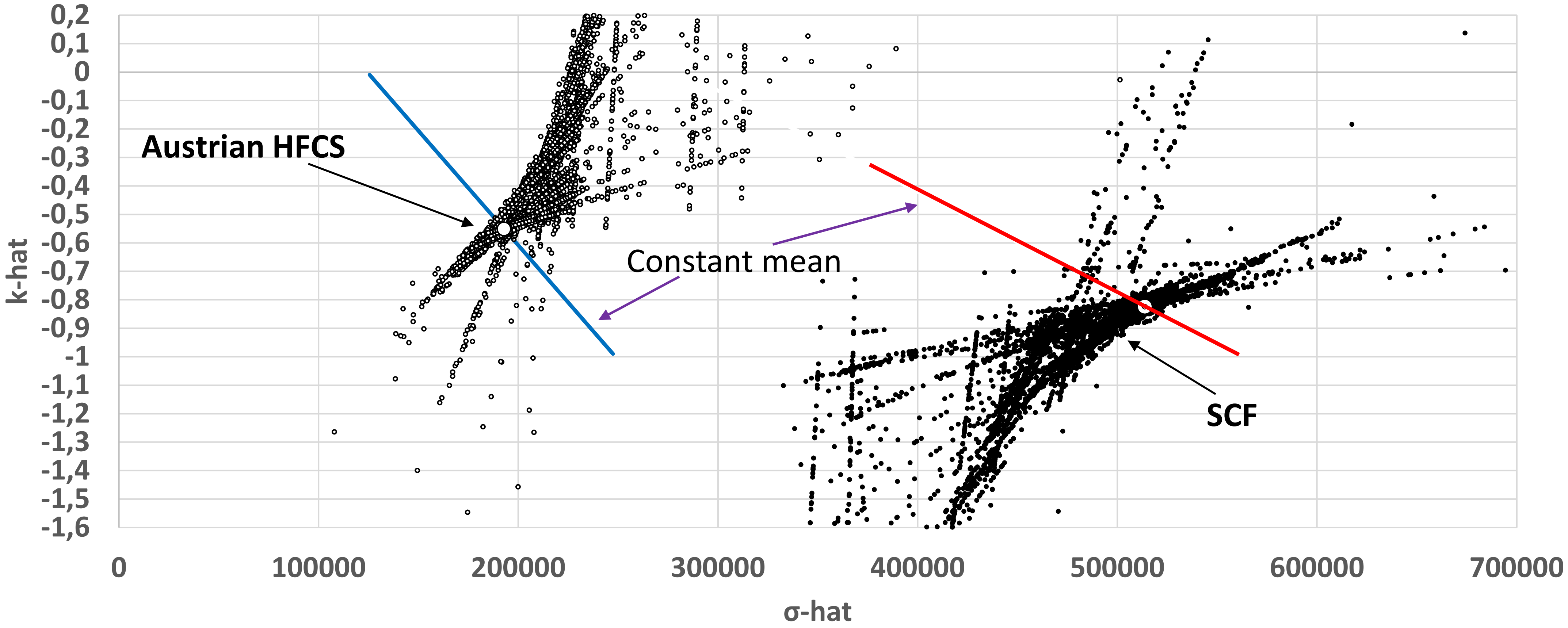
Basic estimation method (Castillo & Hadi [1997])

- Substitute $\delta = \sigma/k$ (for k^=0): $\mathrm{F}(\lambda - \lambda^0) = 1 - (1 - (\lambda - \lambda_0)/\delta)^{\frac{1}{k}}$
- For observed (λ_i , p_i): $k = ln(1 - (\lambda_i - \lambda_0)/\delta)/ln(1 - p_i)$
- Take ratio for *i* and *j*: $\frac{ln\left(1-\left(\lambda_i-\lambda_0\right)/\delta\right)}{ln\left(1-\left(\lambda_j-\lambda_0\right)/\delta\right)} = \frac{ln\left(1-p_j\right)}{ln\left(1-p_i\right)}$
- Compute $\hat{\delta}$ by search, and use data to compute $\hat{k}$ and $\hat{\sigma}$
- (p is percentile of distribution as defined above λ_0)
- (See paper for other technical details)

Basic implementation

- (More on selecting λ^0 later: **take as given for now**)
- Choose many data pairs (λ_i , λ_j) and corresponding (p_i , p_j)
  - In principle, could choose any set of pairs
  - In this implementation, 5,100 pairs used, "stratified" to ensure broad distribution
  - Maps out the range of k and σ compatible with the data
    - In practice, a very broad range

Estimates of (k ,σ)



For lowest formally plausible value of λ^0

Stage 1: Select “best” $\hat{k}$ and $\hat{\sigma}$

- Select $(\hat{k}, \hat{\sigma})$ to minimize (modified) Anderson-Darling right-tail criterion
- $AD\_RT\left(\lambda^0, \hat{k}, \hat{\sigma}\right) = \sum_{\lambda > \lambda^0} W(\lambda) \left\{ \frac{\hat{p}\left(\lambda - \lambda^0 | \hat{k}, \hat{\sigma}\right) - p\left(\lambda - \lambda^0\right)}{1 - min\left(0.99, \hat{p}\left(\lambda - \lambda^0 | \hat{k}, \hat{\sigma}\right)\right)} \right\}^2 / \sum_{\lambda > \lambda^0} W(\lambda)$
- Weighted sum of $\left(\frac{\text{“Predicted percentile”} - \text{Actual percentile}}{100 - \min(99, \text{"Predicted percentile"})}\right)^2$ over range above λ^0

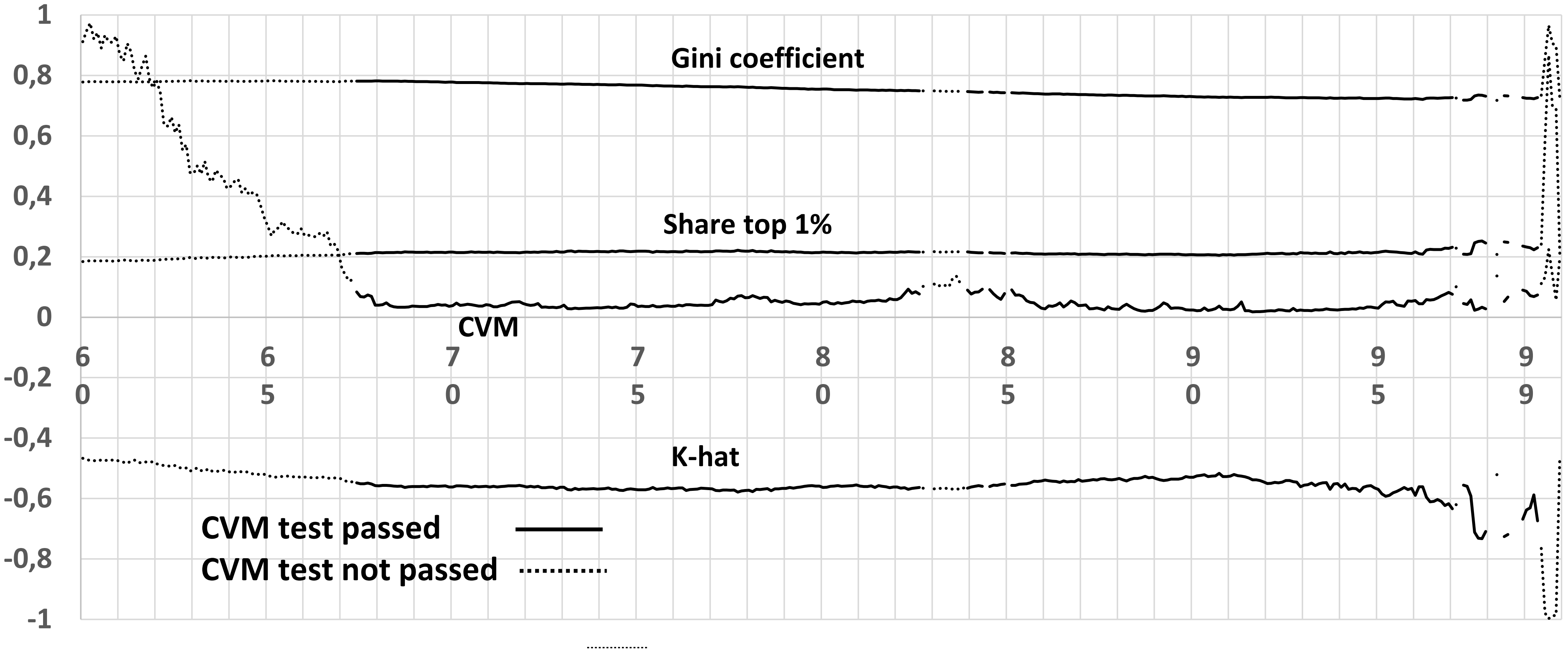
Stage 2: determine whether plausibly GPD

- Use Cramer-von Mises test critical value (Choulakian and Stephens [2001]) to assess plausibility of result as GPD
- $CVM\left(\lambda^0, \widehat{k}, \widehat{\sigma}\right) = \sum_{\lambda>\lambda^*} \varpi(\lambda) \left[\hat{p}\left(\lambda - \lambda^0 | \widehat{k}, \widehat{\sigma}\right) - p\left(\lambda - \lambda^0\right)\right]^2 + \frac{1}{12ESS\left(\lambda^0\right)}$
  - $\varpi(\lambda) = \frac{w(\lambda)}{\sum_{L>\lambda^0} w(L)} \frac{\left(\sum_{L>\lambda^0} w(L)\right)^2}{\sum_{L>\lambda^0} w(L)^2} = \frac{w(\lambda)}{\sum_{L>\lambda^0} w(L)} ESS\left(\lambda^0\right)$
- Sum of (“Predicted percentile”−Actual percentile)^2, weighted over range above λ^0, with correction for the effective sample size

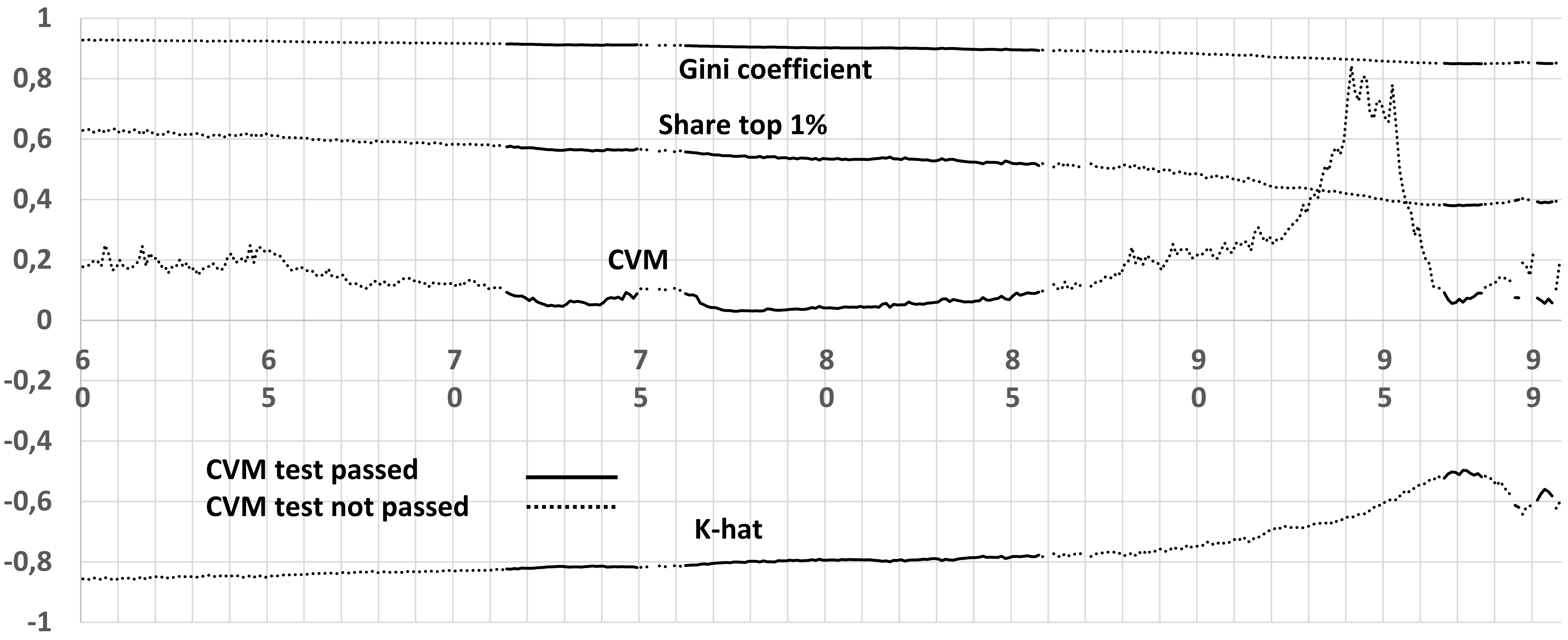
Wait a minute! How select $\boldsymbol{\lambda^0}$?

- Theory: If the data above any λ^0 are GPD, then the distribution from any point λ^+ above λ^0 is also GPD with the same *k* and with σ given by:
  - $\sigma\left[\lambda - \lambda^+ | \lambda > \lambda^+ > \lambda^0\right] = \sigma^0 - k\left(\lambda^+ - \lambda^0\right)$
- Estimation efficiency argues for selecting lowest value of λ^0 for which $\left(\hat{k}, \hat{\sigma}\right)$ pass CVM test
  - **But…**

HFCS: AT: Gini coeff., share top 1%, CVM and $\hat{k}$, by %-ile corresponding to $\boldsymbol{\lambda}^{0}$



SCF: Gini coeff., share top 1%, CVM and k̂, by %-ile corresponding to λ^0



How select $\boldsymbol{\lambda^0}$?: An alternative

- Smooth over the region with "acceptable" estimates
- $\hat{k}_{mean} = \frac{\sum_{\lambda \geq \lambda^{0*}}^{\lambda^{max}} ESS(\lambda)W(\lambda)\left(CVM(\lambda, \hat{k}_\lambda, \hat{\sigma}_\lambda) \leq \text{TCVM}(\hat{k}_\lambda)\right)\hat{k}_\lambda}{\sum_{\lambda \geq \lambda^{0*}}^{\lambda^{max}} ESS(\lambda)W(\lambda)\left(CVM(\lambda, \hat{k}_\lambda, \hat{\sigma}_\lambda) \leq \text{TCVM}(\hat{k}_\lambda)\right)}$
- But cannot do same for δ , because δ depends on λ through σ
  - "Standardize" σ at some λ^s before smoothing: $\sigma(\lambda, \lambda^s) \equiv \hat{\sigma}_\lambda - \hat{k}_\lambda(\lambda^s - \lambda)$
  - $\hat{\delta}_{mean} = \frac{\sum_{\lambda \geq \lambda^*}^{\lambda^{max}} ESS(\lambda)W(\lambda)\left(CVM(\lambda, \hat{k}_\lambda, \hat{\sigma}_\lambda) \leq \text{TCVM}(\hat{k}_\lambda)\right)(\sigma(\lambda, \lambda^s)/\hat{k}_\lambda)}{\sum_{\lambda \geq \lambda^*}^{\lambda^{max}} ESS(\lambda)W(\lambda)\left(CVM(\lambda, \hat{k}_\lambda, \hat{\sigma}_\lambda) \leq \text{TCVM}(\hat{k}_\lambda)\right)}$
- $\hat{k}_\lambda^{norm} = \left(\hat{k}_\lambda - \hat{k}_{mean}\right)/\hat{k}_{std}$ and $\hat{\delta}_\lambda^{norm} = \left(\hat{\delta}_\lambda - \hat{\delta}_{mean}\right)/\hat{\delta}_{std}$
- Choose λ and $(\hat{k}, \hat{\sigma})$ corresponding to minimum of $\left({\hat{k}_\lambda^{norm}}^2 + {\hat{\delta}_\lambda^{norm}}^2\right)$

“Top break”

- Breaks in the fit of the data
  - Region below about 85 %-ile for HFCS: AT
  - Region below about 95 %-ile for SCF
- Estimates considered:
  - Lowest acceptable value of λ^0
  - Smoothed estimates
  - Smoothed estimates above break
- **Note:** (Largely) omitting confidence intervals (until end)

HFCS: AT: Basic estimates

	1^{st} plausible λ^0	λ^0 for k_{mean}	λ^0 for k_{mean}>break	Unaltered data
%-ile of λ^0	67.4	79.5	93.6	NA
$\hat{k}$	-0.548	-0.560	-0.547	NA
Top 1% share	**21.1%**	**21.4%**	**21.1%**	**22.9%**
Gini coefficient	0.781	0.756	0.724	0.723
Mean/"Pop_mean"	**73.4%**	**73.7%**	**73.4%**	**73.7%**
N for estimates	**830**	**527**	**167**	**3072**

In this and all charts that follow, the %-ile of λ^0 given is defined in terms of the original data.

SCF: Basic estimates

	1st plausible λ^0	λ^0 for k_{mean}	λ^0 for k_{mean}>break	Unaltered data
%-ile of λ^0	71.4	79.9	96.6	NA
$\hat{k}$	-0.823	-0.794	-0.523	NA
Top 1% share	**57.4**	**53.5**	**38.3**	**38.6%**
Gini coefficient	0.915	0.902	0.851	0.850
Mean/"Pop_mean"	**133.4%**	**125.7%**	**94.7%**	**96.0%**
N for estimates	**2690**	**2247**	**1042**	**6248**

Estimates

- For AT, apparently the model together with the observed data is insufficient to alter the results substantially
- For the SCF, the situation is more strange
  - The most straightforward estimates imply highly implausible population results
  - Estimating only over the part of the distribution above the “break area” yields results close to the unaltered results

- Consider two alternatives:
  - Exclude some data at the top under assumption of substantial misreporting
  - Formally treat top of the distribution as entirely unobserved

Exclude “bad” data?

- Some data may be measured with error, and errors in the right tail may be particularly damaging
  - (As noted earlier, ignoring reporting error other than in right tail!)
- Because estimation method relies on pairs of data, it is straightforward to exclude any region of data
  - Focus on top values here
- But CVM test for GPD plausibility relies on comparisons of actual/predicted across whole range above given λ location value
  - Approximate omitted range by inflating CVM difference elsewhere
    - ½ omitted distance (in percentiles) below truncation point
    - ½ omitted distance (in percentiles) above location value
    - (Note: CVM values tend to zero at either end)

HFCS: AT: Omit some data

	Omit top 1%	Omit top 2%	Unaltered data
%-ile of λ^0	67.5	76.1	NA
$\hat{k}$	-0.550	-0.622	NA
Top 1% share	**21.1%**	**25.0%**	**22.9%**
Gini coefficient	0.781	0.778	0.723
Mean/"Pop_mean"	**73.4%**	**77.7%**	**73.7%**
N for estimates	**799**	**558**	**3072**

Reporting only estimates for k_{mean}

SCF: Omit some data

	Omit top 0.5%	Omit top 1%	Unaltered data
%-ile of λ^0	97.3	97.6	NA
$\hat{k}$	-0.242	-0.506	NA
Top 1% share	**31.3%**	**39.6%**	**38.6%**
Gini coefficient	0.833	0.853	0.850
Mean/"Pop_mean"	**85.8%**	**95.9%**	**96.0%**
N for estimates	**387**	**277**	**6248**

Reporting only estimates for k_{mean} above break

Estimates

- For AT, omitting top 1% makes little difference over the estimate using all data
  - Omitting top 2% increases measures of concentration, but fraction of aggregate still far below 1
- For SCF, omitting top ½ percent lowers both top share and the mean
  - Omitting entire top 1% raises top share, and leaves implied aggregate nearly same
- Appears to be not a useful alternative on its own

Possibly unmeasured top of distribution?

- Very likely that 100th %-ile in survey is not population 100th %-ile
  - Similarly likely for some range below that
- For example, if top 1% is not observed, then:
  - Observed 100th %-ile is true 99th %ile, observed 50th %-ile is true 49.5th %-ile
- Let λ^{0*} be GPD location parameter (**$\Pi\left(\lambda^{0*}\right)$ %-ile of observed data**) and let ρ be the percent unobserved
  - Then adjust observed p_i in parameter estimates by $\frac{\left(\left(100-\Pi\left(\lambda^{0*}\right)\right)/100\right)}{((100-\Pi(\lambda^{0*})+\rho)/100)}$
  - (Also need to make same CVM approximation as in previous case)
- (A little more later on a path toward specifying ρ)

HFCS: AT: Unobserved region

	Missing top 0.1%	Missing top 0.5%	Unaltered data
%-ile of λ^0	73.0	70.6	NA
$\hat{k}$	-0.575	-0.628	NA
Top 1% share	**22.2%**	**25.8%**	**22.9%**
Gini coefficient	0.774	0.785	0.723
Mean/"Pop_mean"	**74.8%**	**81.0%**	**73.7%**
N for estimates	**678**	**749**	**3072**

Reporting only estimates for k_{mean}

SCF: Unobserved region

	Missing top 0.1%	Missing top 0.5%	Unaltered data
%-ile of λ^0	96.7	96.6	NA
$\hat{k}$	-0.543	-0.544	NA
Top 1% share	**39.5%**	**39.0%**	**38.6%**
Gini coefficient	0.852	0.858	0.850
Mean/"Pop_mean"	**95.1%**	**110.3%**	**96.0%**
N for estimates	**1030**	**1042**	**6248**

Reporting only estimates for k_{mean} above break

Estimates

- Fraction of implied aggregate mean explained by survey mean appears to increase with size of area of distribution omitted
  - But no guarantee that result is monotonic or that 100% is obtainable
- Next alternative is more direct

Impose an external constraint?

- A key property of GPD:
  - $E\left[\lambda - \lambda^0 | \lambda > \lambda^0 > 0\right] = \frac{\sigma}{1+k}$, if $k > -1$
- Given an external total estimate A_0, define the mean value above some λ_i (location parameter for GPD estimate) as follows:
- $A_{\lambda_i}/N_{\lambda_i} \equiv \mu_{\lambda_i} = \left(A_0 - \sum_{\lambda<\lambda_i} W(\lambda)\lambda\right)/N_{\lambda_i}$
- Use earlier search technique to solve for $\hat{\delta}$
- Then $\hat{k} = \frac{\mu_{\lambda_i} - \lambda_i}{\hat{\delta} - \left(\mu_{\lambda_i} - \lambda_i\right)}$ and $\hat{\sigma} = \left(\mu_{\lambda_i} - \lambda_i\right)\left(1 + \hat{k}\right)$
- (External estimate may be questionable, but unlike "rich list" data it (usually) results from a transparent process)

HFCS:AT: Constrained

Constraint:	Survey total	External aggregate	External aggregate	Unaltered data
Unobserved:	**0%**	**0%**	**0.2%**	**NA**
%-ile of λ^0	68.8	97.9	[redacted]	NA
$\hat{k}$	-0.558	-0.902	[redacted]	NA
Top 1% share	**21.4%**	**41.7%**	[redacted]	**22.9%**
Gini coefficient	0.782	0.793	[redacted]	0.723
Mean/"Pop_mean"	**73.7%**	**100%**	[redacted]	**73.7%**
N for estimates	**815**	**56**	[redacted]	**3072**

Reporting only estimates for k_{mean}

SCF: Constrained

Constraint:	**Survey total**	**External aggregate**	**External aggregate**	**Unaltered data**
Unobserved:	**0%**	**0%**	**0.01%**	**NA**
%-ile of λ^0	96.4	96.4	96.5	NA
$\hat{k}$	0.535	-0.535	0.532	NA
Top 1% share	**38.7%**	**38.8**	**38.7**	**38.6%**
Gini coefficient	0.852	0.852	0.852	0.850
Mean/"Pop_mean"	**96.0%**	**100%**	**100%**	**96.0%**
N for estimates	**1067**	**1067**	**1055**	**6248**

Reporting only estimates for k_{mean} above break (same as for basic k_{mean} here)

Estimates

- For HFCS: AT and SCF, using a measure of aggregate wealth
  - At this stage, not critical to question its precise reliability
    - (Beyond my competence in any case!_
- Constraining the total to equal the observed total reasonably approximates direct estimates from the data
  - Bit better for the SCF
- Constraining the total to equal the aggregate yields:
  - Dramatic increase in top share for HFCS:AT
  - Little change (as expected) for the SCF
- But sample size for HFCS:AT very small
  - **Possible to do better?**

Combine with missing mass approach?

- By sample definition, SCF is missing at least the *Forbes* 400 wealthiest
- Austrian data do not contain cases present in rich lists, due either to their not being sampled or their decision not to participate
- Therefore, prima facie case for treating at least some mass as missing

- Need some criterion for selecting a value of missing mass
- Up to a point, when applying the aggregate constraint, increasing fraction of assumed missing mass appears to yield increasing or nearly flat sample size used in estimation
- Beyond that point, sample size declines notably
- Searched over relatively fine gradations to find a maximal sample size for each survey

HFCS:AT: Constrained + missing mass

Constraint:	Survey total	External aggregate	External aggregate	Unaltered data
Unobserved:	**0%**	**0%**	**0.2%**	**NA**
%-ile of λ^0	68.8	97.9	83.3	NA
$\hat{k}$	-0.558	-0.902	-0.806	NA
Top 1% share	**21.4%**	**41.7%**	**41.2%**	**22.9%**
Gini coefficient	0.782	0.793	0.817	0.723
Mean/"Pop_mean"	**73.7%**	**100%**	**100%**	**73.7%**
N for estimates	**815**	**56**	**429**	**3072**

Reporting only estimates for k_{mean}

SCF: Constrained + missing mass

Constraint:	Survey total	External aggregate	External aggregate	Unaltered data
Unobserved:	**0%**	**0%**	**0.01%**	**NA**
%-ile of λ^0	96.4	96.4	96.5	NA
$\hat{k}$	0.535	-0.535	0.532	NA
Top 1% share	**38.7%**	**38.8**	**38.7**	**38.6%**
Gini coefficient	0.852	0.852	0.852	0.850
Mean/"Pop_mean"	**96.0%**	**100%**	**100%**	**96.0%**
N for estimates	**1067**	**1067**	**1055**	**6248**

Reporting only estimates for k_{mean} above break (same as for basic k_{mean} here)

Estimates

- For HFCS AT, top share unchanged (Gini somewhat higher), but sample size used for estimation greatly increased if top 2/10th percent treated as unobserved
- For SCF, little difference in estimates, but sample size declines notably beyond region of 1/100th percent treated as unobserved

Top share: StdErr wrt imputation and sampling

- (**Note:** Earlier estimates for last two columns used only 1st implicate)

	Raw data	Basic Estimate	Constrainted + missing mass
HFCS: AT	22.9	19.8	41.1
	(4.65)	(3.27)	(1.94)
SCF	38.6	38.5	38.7
	(0.68)	(1.51)	(0.69)

- No improvement for SCF: probably reflects high sampling rate at top
- Large improvement for HFCS:AT: much thiner sampling at top

Conclusions

- SCF provides a reasonably workable measure of the wealth distribution even without adjustment
- For HFCS:AT, GPD unaided is not “magic” enough to conjure estimates that align with wealth aggregate
  - Even with tweaks to address bad reporting in the upper tail, or allow for omission of the extreme upper tail
  - Assuming(!) reality is GPD, observed curvature is too thin and contains too little information about the upper tail
  - Constraining the estimates to reproduce the aggregate, especially when combined with an allowance for effective under coverage at the top may be helpful
- Research with other data and further technical development are needed

What we hope we are NOT doing!

Thanks for your attention!
Questions/Comments?

Arthur.Kennickell@gmail.com

Paper available at:

stonecenter.gc.cuny.edu/research/chasing-the-tail-a-generalized-pareto-distribution-approach-to-estimating-wealth-inequality/