Economic Well-Being and Inequality: Papers from the Fifth ECINEQ Meeting

Tournaments and Superstar Models: A Mixture of Two Pareto Distributions

Abdoul Aziz Junior Ndoye Michel Lubrano

TOURNAMENTS AND SUPERSTAR MODELS: A MIXTURE OF TWO PARETO DISTRIBUTIONS

Abdoul Aziz Junior Ndoye and Michel Lubrano

ABSTRACT

We provide a Bayesian inference for a mixture of two Pareto distributions which is then used to approximate the upper tail of a wage distribution. The model is applied to the data from the CPS Outgoing Rotation Group to analyze the recent structure of top wages in the United States from 1992 through 2009. We find an enormous earnings inequality between the very highest wage earners (the "superstars"), and the other high wage earners. These findings are largely in accordance with the alternative explanations combining the model of superstars and the model of tournaments in hierarchical organization structure. The approach can be used to analyze the recent pay gaps among top executives in large firms so as to exhibit the "superstar" effect.

Keywords: Pareto distribution; superstars; tournament theory; wage inequality

JEL classifications: D03; D33; D41; J31; J33

Economic Well-Being and Inequality: Papers from the Fifth ECINEQ Meeting
Research on Economic Inequality, Volume 22, 449–479

ISSN: 1049-2585/doi:10.1108/S1049-258520140000022015

INTRODUCTION

During the last three decades, the U.S. economic growth has experienced episodes of rapidly changing wage differentials characterized by a marked increase in wage inequality in the upper tail of the distribution. Various competing explanations of these earning inequalities have recently generated a heated debate. Common explanations among others attribute these changes to the increase in the wage premium for skilled relative to unskilled workers, to the changes in labor market institutions and in wage setting norms (Bound & Johnson, 1992; DiNardo, Fortin, & Lemieux, 1996; Katz & Autor, 1999; Mincer, 1993; Murphy & Welch, 1992). The recent empirical literature on the changing distribution of wages,[1] and, on the competitive market for Chief Executive Officers (CEOs) and on high compensations in financial markets[2] are contradicting these textbook explanations. These various explanations suggest that each part of the wage distribution is governed by a different logic, each part requiring different explanations. Burkhauser, Feng, Jenkins, and Larrimore (2008) emphasize that "if income inequality has been substantially increasing since 1993 in the U.S., such increases have been confined to the very upper tail of the income distribution." This makes alternative explanations of inequalities in the upper tail of the U.S. earning distribution a considerable challenge.

Atkinson (2008) suggests to combine the superstars model of Rosen (1981) with the hierarchical model of Simon (1957) and Lydall (1959). In addition, tournament theory[3] gives a supplementary motivation for explaining both the increase of the span of control in large firms and the wage gap of top performers.

In the tournament theory in hierarchical organizations, salaries depend on individual performance and are individually negotiated, so that each worker earns a constant multiple of the salary of the worker in the rank below him. This generates approximately a Pareto tail for the earnings distribution (as shown by Lydall, 1968). Tournament theory leads to high increment of salaries in the top rank making the changing of positions costly and difficult. As a consequence, only a small number of competitors will share prizes generated in multi-period tournaments leading to the "superstar" effect of Rosen (1981).

In the economics of superstars of Rosen (1981), individuals differ in their talents, which are assumed fixed, and small differences in talent may imply large differences in earnings. Adler (1985) extends the "superstar" model of Rosen (1981) and argues that factors other than talent, like popularity and charisma, may also expand the services of superstars. Frank and

Cook (1995) emphasize that the distribution of earnings of superstars stems largely from the growing prevalence of winner-take-all markets, which are in many cases the result of competitive forces. The distribution of earnings is then given by the maximum values generated by the results of many separate competitions. Exceeding a given high threshold value, the distribution takes a generalized Pareto form with a Pareto tail (Embrechts, Kluppelberg, & Mikosch, 1997).

Each type of earnings distribution in both models has a Pareto form, the combination of the two models leads naturally to a mixture of two Pareto distributions. This paper provides Bayesian inference for a mixture of two Pareto distributions in order to approximate the upper tail of a wage distribution. This mixture model is applied to the data from the CPS Outgoing Rotation Group to analyze the recent structure of top wages in the United States from 1992 through 2009. Our results show a rising wage inequality between the very highest wage earners (0.95 quantile) and the other high wage earners. These findings are largely in accordance with the explanations combining the model of superstars and the model of tournaments in hierarchical organization structure.

The remainder of the paper is organized as follows. The next section reviews the tournament theory in the hierarchical organization model of Simon (1957) and Lydall (1959), the superstar model of Rosen (1981) and shows how a mixture of two Pareto distributions emerges. The section "Mixture of Pareto Distributions" provides Bayesian inference for the model represented by a mixture of two Pareto distributions. The section on "Empirical Application" presents a truncated sample procedure and illustrates our approach using the CPS-ORG sample from 1992 to 2009. The last section concludes and points out the inferential statistical problems of a mixture of Pareto.

TOURNAMENTS, SUPERSTARS, HIERARCHIES, AND PARETO DISTRIBUTIONS

Simon (1957) and Lydall (1959) model a pyramidal employment structure where each worker has subordinates and where pay increases by a constant increment as one advances up on the ladder. The ratio between the number of supervisors in each rank and the number of persons below this rank is constant. The structure and the number of jobs are relatively fixed and remunerations are closely dependent on the rank $t \in [1,n]$ in the hierarchy

rather than being dependent on the individual performance. If s is a fixed span of control and N_t the total number of subordinates at each rank t, we have the recurrence relation $N_{t+1} = sN_t$. And, the salary of a supervisor of any level in the hierarchy is directly related to the aggregate salary of the persons whom he controls. If the salary of each grade of employee is w_t and i a fixed increment of salary when moving up to the next step of the ladder, the wage structure is given by $w_{t+1} = (1+i)w_t$.

With these assumptions, Lydall (1968) shows that the distribution of earnings in a hierarchical organization can be approximated by a Pareto distribution with exponent α corresponding to the slope of the plot of the logarithm of the number of persons in each grade ($\log N_t$) against the logarithm of the salary appropriate to that grade ($\log w_t$):

$$\alpha = \frac{\log(N_{t+1}/N_t)}{\log(w_{t+1}/w_t)} = \frac{\log s}{\log(1+i)} \tag{1}$$

The Gini index I_G has an analytical expression for the Pareto distribution and depends only on the Pareto coefficient α, so that $I_G = 1/(2\alpha - 1)$, $\alpha > 0.5$. This suggests that, a rise in wage inequality in a hierarchical organization corresponds to a decrease of the Pareto coefficient, α. This can be caused either by an increase in the increment prize i or a decrease in the span of control s.

Tournament Theory and the Hierarchical Model

Rosen (1981) and Lazear and Rosen (1981) show that an optimal tournament dominates other forms of remuneration schemes adopted by firms. In tournament theory, firms are willing to pay extremely large compensation differences to those situated in the top ranks of the hierarchy in order to create adequate incentives for employees. These large pay differences are supposed to induce greater efforts and performances for employees who seek to increase their chances of promotion (Brian, Charles, & James, 1993; Rosen, 1986). The highest rank corresponds to the highest salary, so the main prize of this competition is the top executive's job.

We now introduce the simple assignment model (see Sattinger, 1993; Tervio, 2008 or Gabaix & Landier, 2008) for top rank managers pay in the previous model of large hierarchical organizations. In this tournament model, we consider a set of n ordered positions labeled by t, and

determined by the expected talent of managers. The market capitalization of the firm is characterized by its size denoted by S. The manager indexed by its rank t has an expected talent $T(t)$ and receives a total compensation $w(T(t))=w(t)$. If C is the effect of talent on earnings, then $CS^{\gamma}T(t)$ is the firm value generated by the overhead t. For a fixed high compensation, the firm chooses its potential CEO so as to maximize his net impact:

$$\max_{t} \; CS^{\gamma}T(t) - w(t) \tag{2}$$

In a competitive equilibrium, the manager t who is expected to be the most talented is ranked at the top, so that $t=n$. The first order condition implies that the marginal cost $w'(n)$ of the better manager is equal to the marginal benefit that he generates, that is:

$$w'(n) = CS^{\gamma}T'(n) \tag{3}$$

where a small difference in talent, $\mathrm{d}n$, is such that $T'(n)\mathrm{d}n = T(n+\mathrm{d}n) - T(n)$.

In a hierarchical organization, tournaments suggest high compensation differences among the executive ranks of the structure. As a consequence, if the compensation is kept largely fixed at the top, this makes any organizational change costly and any position modification in the structure difficult. This leads to an increase in both the span of control s, and the salary of the top executives.

Gabaix and Landier (2008) develop an assignment model for CEOs' pay across different hierarchical organizations. In their model, CEOs have different talents and are matched to firms in a competitive assignment model. There is a set of N firms, each firm $n \in [1,N]$ has size $S(n)$. It is assumed that each firm has a potential CEO who is optimally chosen using the maximization program 2. In a competitive equilibrium, the classical assignment Eq. (3) holds.

Following extreme value theory, Gabaix and Landier (2008) emphasize that the size of the firm, $S(n)$ and the spacings in the upper tail of the talents, $T'(n)$ follow a Pareto distribution. That is, $S(n)=An^{-\alpha}$, and $T'(n) = -Bn^{\beta-1}$. They solve Eq. (3) and find that the reservation wage of at least one talented CEO is given by:

$$w(n) = \frac{A^{\gamma}BC}{\alpha\gamma - \beta} \times n^{-(\alpha\gamma+\beta)} \tag{4}$$

The reservation wage of talented CEO follows a Pareto distribution with coefficient $(\alpha\gamma + \beta)$. Then, the enormous CEO pay in large pyramidal organizations is consistent with "the superstar effect" of Rosen (1981).

However, Lazear and Rosen (1981)'s tournament theory is different from the superstar theory of Rosen (1981).

Superstars Model and Its Pareto Approximation

The economics of "superstars" has long been employed to explain wages of professional sport players or opera singers. However, in recent years, the superstar model has been employed in many other fields including academic research, scientific work, arts and architecture, financial markets, executives, and others.

Building on the insights of Marshall (1947),[4] Rosen (1981) introduced the superstar model to explain large earnings differentials induced by small differences in talent among performers. According to Rosen (1981) and Adler (1985), the development in communication, manufacturing technology, trade liberalization, and globalization allows to the most talented and popular performers to expand their services and to extract a rent that is related to the extent of the served market. For every superstar who receives a high earning, advanced payment in accordance to his talent or popularity which impacts to the vast audience he is able to reach, there are the others, many of them being nearly as talented, who never manage to even support themselves. This ensures a small number of top performers with extremely high earnings relative to their rivals.

The superstar model of Rosen (1981) is re-interpreted by Frank and Cook (1995) as a winner-take-all market. In a winner-take-all market, Frank and Cook (1995) suggest that runaway professional wages are the result not of a breakdown of competition but of the spread of markets in which the value of production depends primarily on the efforts of only a handful of top players, "superstars" who are paid accordingly. Frank and Cook (1995) emphasized that the distribution of earnings of top performers stems largely from the growing prevalence of winner-take-all markets, which are in many cases the result of competitive forces. In this case, the distribution of earnings is given by the maximum values generated by the results of many separate competitions. Exceeding a given high threshold value w_m, the distribution of earnings takes the form of a Pareto as shown by Embrechts et al. (1997) (or Gabaix & Landier, 2008), with coefficient α

such that the mean income higher than the threshold w_m, $E(w|w \geq w_m)$ is given by:

$$E(w) = w_m \frac{\alpha}{\alpha - 1} \Pi(w \geq w_m), \quad \alpha > 1$$

Atkinson (2008) suggests that a decline in the exponent of the Pareto distribution is the basis for the superstar explanation of rising earnings dispersion.

In the winner-takes-all market, the presence of superstars with significantly higher talent and popularity will reduce the efforts of other competitors. Brown (2011) studies the effects of incentives and strategies created by the presence of a superstar among golf professionals. He finds that the presence of *Tiger Woods* has a negative effect on the top half of the competitor field in terms of ability. This means that the earnings of the second best player depend on the reach of *Tiger Woods* and so on down the range of talents.

Similarly, tournament prizes in hierarchical organizations can create negative incentives, the presence of a talented executive can have a negative impact on the performance of many workers and thus commands a very high salary. This configuration leads to a decrease of the Pareto exponent in the superstar model.

What Do We Observe?

The distribution of earnings in the model of "superstars" (Rosen, 1981) and in the model of tournament in hierarchical organization (Lydall, 1959; Simon, 1957) has a Pareto form and their combination leads to a mixture of two Pareto. In the tournament theory in hierarchical organizations, salaries depend on individual performance and are individually negotiated. Lydall (1968) shows that for a fixed span of control, c, and a fixed increment of salary, i, the distribution of earnings can be approximated by a Pareto distribution with an exponent equal to $\ln(c)/\ln(1+i)$. The superstars theory of Rosen (1981) is re-interpreted by Frank and Cook (1995) as the "winner-takes-all" markets. In this model, extreme values theory suggests that for a sufficiently high threshold, the distribution of earnings takes a generalized Pareto form which has a Pareto tail (see for instance Embrechts et al., 1997). Using a reparametrization, a random variable of the generalized Pareto distribution can be transformed into that of the Pareto

distribution. In addition, Gabaix and Landier (2008) provide a competitive assignment equilibrium model of CEO pay. Matching CEOs with different talents, they find a very small dispersion in CEO talent, which nonetheless justifies large pay differences. This exhibits the "superstars" effect. They demonstrate that the distribution of CEO pay takes a Frechet form which behaves like a Pareto for very high earnings. However, the superstar models leads in theory to a lower Pareto coefficient than the hierarchical model. For a given dataset, we observe two types of top wage formation. We can successfully test these models using a mixture of two Pareto distributions with significantly different Pareto parameters to distinguish the high wage earners and the very highest wage earners.

MIXTURE OF PARETO DISTRIBUTIONS

The Pareto density is very well documented in the literature (see, e.g., the review by Arnold, 2008, and the references quoted in). Mixtures of Pareto densities are on the contrary very scarcely covered and most of the time only particular cases are considered. Nadarajah (2006) considers the case of a mixture of two Pareto with a common scale parameter which is in fact equivalent to the double Pareto of Reed and Jorgenson (2004). Nair (2007) explores the properties of Pareto II mixtures for modelling income distribution and reliability analysis. However, he has constrained parameters for inference (see also Noor & Aslam, 2012). Bee, Benedetti, and Espa (2013) consider an unconstrained model of two Pareto I mixture. They report that the EM algorithm does not work in this case, confirming that inference for Pareto mixtures is not a simple task. We shall first review the Pareto process and Bayesian inference for this process. We will then investigate how Bayesian inference can be implemented for a mixture of two Pareto distributions.

Pareto Processes

A random variable Y is said to be distributed according to a Pareto law if its density is equal to:

$$f(y|\alpha, y_m) = \alpha y_m^{\alpha} y^{-(\alpha+1)} \Pi(y > y_m), \quad y_m > 0, \quad \alpha > 0$$

where $\Pi(\cdot)$ is the indicator function, y_m is a scale parameter and α a shape parameter, which is known as the tail index. The support of the Pareto density is defined over $[y_m, \infty)$, which means that the support of the density depends on a parameter. The first two moments are:

$$E(y) = \frac{\alpha}{\alpha - 1} y_m \quad \text{Var}(y) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} y_m^2$$

and exist only for $\alpha > 1$ and $\alpha > 2$, respectively. The cumulative distribution is:

$$F(y) = (1 - y_m^{\alpha} y^{-\alpha}) \Pi(y > y_m)$$

Two sufficient statistics are provided by $\text{Min}(y)$ and $\sum \log(y_i/y_m)$. Classical estimates are obtained by taking $\hat{y}_m = \text{Min}(y)$ and $\hat{\alpha} = n / \sum \log(y_i/y_m)$.[5]

This is the Type I Pareto. The Pareto family is rather rich. Four variants are commonly reported and are particular cases of the Feller-Pareto distribution $\text{FP}(\mu, \sigma, \gamma, \alpha)$ which is obtained as the ratio of two Gamma distributions U_1 and U_2 with:

$$W = \mu + \sigma \left(\frac{U_1}{U_2} \right)^{\gamma}$$

where $U_1 \sim \Gamma(1,1)$ and $U_2 \sim \Gamma(\alpha, 1)$. The Pareto I corresponds to $\text{FP}(y_m, y_m, 1, \alpha)$. Simple Bayesian inference is not available for the Pareto II−IV as underlined by Arnold (2008). So that in practice, only the Pareto I is commonly used.

Bayesian Inference for the Pareto I Process

Bayesian inference for the Pareto process was treated in a number of papers, starting with Lwin (1972) and Arnold and Press (1983). Arnold (2008) details a Gibbs sampler after a re-parametrization in $\tau = 1/y_m$, where τ is called a precision parameter. This parametrization is convenient for Bayesian inference because both a Pareto prior on τ and a Gamma prior on α are natural conjugates when the other parameter is considered as fixed. The conditional posterior of τ is itself a Pareto density, while the conditional posterior of α is a Gamma density. However, a Pareto prior on τ is

difficult to interpret while a prior on y_m has a natural sample interpretation. It is possible to keep the usual parametrization of the Pareto process if we choose the prior on y_m as a power function density.

Power Functions

A random variable X is said to have a power function distribution if its probability density function is defined as

$$p(x) = \alpha x_m^{-\alpha} x^{\alpha-1} \Pi(x < x_m), \quad \alpha > 0, \quad x_m > 0$$

It is an increasing function of x when $\alpha > 1$ (decreasing when $0 < \alpha < 1$) and is defined over $[0, x_m]$. Its moments are:

$$E(x) = \frac{\alpha}{\alpha+1} x_m \quad \text{Var}(x) = \frac{\alpha}{(\alpha+1)^2(\alpha+2)} x_m^2$$

and always exist, contrary to the Pareto process. The cumulative distribution function is:

$$F(x) = x_m^{-\alpha} x^{\alpha} \Pi(x < x_m)$$

Two sufficient statistics are provided by Max(y) and $\sum \log(y_i/y_m)$.[6] If x has a power function distribution in (α, x_m), then $y = 1/x$ is distributed according to a Pareto(α, y_m) where $y_m = 1/x_m$. We have chosen to present separately this distribution even if it corresponds to a simple transformation of the Pareto I because we shall use its properties and moments to elicit a prior information.

Likelihood Function and Prior Densities

Let us consider a sample $(y_1, \ldots, y_n)$ of n observations of the Pareto random variable Y. The likelihood function of this sample is:

$$L(y; \alpha, y_m) = \alpha^n y_m^{\alpha n} \prod y_i^{-(\alpha+1)} \Pi(y_{(1)} > y_m)$$

where $y_{(1)}$ is the first order statistics, that is, the minimum of the sample. It is convenient to rewrite this likelihood function as:

$$L(y; \alpha, y_m) = \alpha^n \ \exp\left\{ -(\alpha+1) \sum \log(y_i) + \alpha n \log(y_m) \right\} \Pi(y_{(1)} > y_m)$$

It is clear that the Pareto distribution does not belong to the exponential family when its two parameters are unknown, just because the support depends on one of the parameters. However, from this writing, we can find that $y_{(1)}$ and $\sum \log(y_i)$ are two sufficient statistics. And, conditionally on y_m known, the Pareto does belong to the exponential family. Following Arnold and Press (1983), we propose to use an independent prior $p(\alpha,y_m)=p(\alpha)p(y_m)$. We shall discuss and justify this choice in a separate subsection.

When y_m is known, $\log(y/y_m)$ is distributed according to an exponential distribution. In this case, the natural conjugate prior for α is the Gamma density with ν_0 degrees of freedom and as scale parameter α_0:

$$p(\alpha|\nu_0,\alpha_0)\propto\alpha^{\nu_0-1}\exp(-\alpha\alpha_0),\quad E(\alpha)=\nu_0/\alpha_0,\quad \mathrm{Var}(\alpha)=\nu_0/\alpha_0^2$$

A non-informative prior corresponds to letting the prior parameters go to the limit of their domain of definition with $\alpha_0=0$ and $\nu_0=0$:

$$p(\alpha)\propto\frac{1}{\alpha}$$

When α is known, it is also possible to find a convenient conjugate prior for y_m. As we adopt the same parameterization as given by Arnold and Press (1983), the conjugate prior is a Power function distribution with shape parameter α_0 and scale parameter y_{m0}:

$$p(y_m|\gamma_0,y_{m0})=\gamma_0 y_{m0}^{-\gamma_0} y_m^{\gamma_0-1}\Pi(y_m<y_{m0})$$

A non-informative prior is obtained for $\gamma_0=0$, and letting y_{m0} go to infinity:

$$p(y_m)\propto\frac{1}{y_m}$$

Alternative Prior Distributions

From a probabilistic point of view, the natural way of specifying a prior is to consider a joint prior on all the parameters. A joint prior in a natural conjugate prior is a convenient tool as it combines easily with the sufficient

statistics of the sample. It leads to simple posteriors in the linear regression model. An independent prior on the regression parameter might lead to more complicated posteriors in this simple model (see, e.g., Drèze, 1977). However, a joint natural conjugate prior has a tendency to hide possible conflicts between the sample and the prior information, whereas an independent prior would not. Following Bauwens (1991), a joint natural conjugate prior can lead to some pathologies in the case of partial non-informativeness.

We are here in a special case with the Pareto I process. It has two parameters, a shape parameter α and a scale parameter y_m which is in fact a minimum bound of the observations. Since this parameter depends on the support of the density, the Pareto process does not belong to the exponential family, except when the minimum bound is fixed. Despite this pathology, Lwin (1972) managed to propose a joint prior density for (α,y_m) which combines nicely with the sample. However, this prior has some particular features that can make it not attractive. First, it is parameterized as $p(\alpha, y_m) = p(\alpha)p(y_m|\alpha)$. In the Pareto process, it is more natural to start eliciting a prior on y_m in terms of a boundary and then independently eliciting a prior on the shape parameter α, for instance in terms of a Gini coefficient, $1/(2\alpha - 1)$. Second, as amply underlined by Arnold and Press (1983), when using this joint prior, the marginal posterior density of y_m is unbounded with no sample information being able to remove this feature. On the contrary, with independent priors on α and y_m, first the marginal posteriors have simpler expressions (even if their evaluation requires numerical integration) and second, a slight prior information on y_m removes the unboundedness of the marginal posterior of y_m as we shall see below. Another advantage of these independent priors on α and y_m is that they combine easily with the likelihood function, leading to simple conditional posterior densities leading to a simple Gibbs sampler algorithm for inference in a mixture of two Pareto densities. Finally, we should note that Arnold, Castillo, and Sarabia (1998) have proposed for many models and in particular for the Pareto process a specific bivariate prior that has the independent priors of Arnold and Press (1983, 1989) and the bivariate prior of Lwin (1972), as particular cases (see also Arnold, Castillo, & Sarabia, 1999; Arnold, Castillo, & Sarabia, 2001). This prior is very convenient for comparing the influence of the different parametrizations. From the application reported by Arnold and Press (1989), there does not seem to be a major difference.

Alternative Parametrizations

The parametrization for inference in the Pareto process is not a settled matter. Arnold and Press (1983) adopt the parameterization α, y_m and note that with this parametrization the joint prior of Lwin (1972) leads to an unbounded marginal posterior of y_m at the boundary zero. They note that a reparametrization in $\tau = 1/y_m$ would solve the problem. However, they argue that the natural parametrization in y_m is much easier both to elicit and to interpret. As a matter of fact, and especially when in a mixture of two Pareto densities, a marginal prior on y_m is easy to elicit in terms of a minimum bound first and then in terms of a point of discontinuity between the two Pareto members. Once the support of the two Pareto members is elicited, it is easier to elicit the two α's in terms of a Gini coefficient for each member. For the single Pareto I, Arnold and Press (1989) adopt the parametrization $\tau = 1/y_m$ together with a modified Lwin's (1972) prior in $\alpha|y_m$ and τ and note that using this parametrization, the pathological behavior of the original Lwin's (1972) prior disappears. However, they maintain their criticism against the use of a joint prior on α and τ instead of two independent priors on each parameter. Arnold (2008) adopts the parametrization in $x > y_m$ for presenting the process and detailing Bayesian inference with independent priors using a Gibbs sampler. Introducing the conditionally conjugate priors of Arnold et al. (1998), they turn to the parametrization in terms of τ without much discussion. We could conclude that the parametrization in τ is more convenient for the modified Lwin's prior (Arnold & Press, 1989) while the parametrization in y_m is more convenient when using independent marginal priors in the context of mixtures.

Joint Posterior Densities

Arnold and Press (1983) have conducted Bayesian inference when both α and y_m are unknown. They derived analytical expressions for the two marginal densities of α and y_m, using independent informative priors. Those marginals do not belong to any class of known densities as we have:

$$p(\alpha|y) \propto \left(\alpha + \frac{\gamma_0}{n}\right)^{-1} \alpha^{n+\nu_0-1} \exp\left(-\alpha\left(\alpha_0 + \sum \log(y_i) - n \log \min(y_{m0}, y_{(1)})\right)\right)$$

$$p(y_m|y) \propto y_m^{\gamma_0 - 1}\left(1 - \frac{n}{\alpha_0 + \sum \log y_i} \log y_m\right)^{n+\nu_0} \Pi(y_m < \min(y_{(1)}, y_{m0})$$

The sole point of interest of these posteriors is to point out that the marginal posterior of y_m is ill-behaved when $\gamma_0 \leq 1$. So, we are obliged to have an informative prior on y_m.

Integrating these two posteriors for finding moments for instance is a cumbersome task (Arnold & Press, 1983, provide approximations). So, as it is possible to find well-defined conditional posteriors, a Gibbs sampler is the privileged route.

Conditional Posteriors

The conditional posterior of α given y_m is:

$$p(\alpha|y_m, y) \propto \alpha^{n+\nu_0-1} \exp\Big(-\alpha\Big(\sum \log(y_i) + \alpha_0 - n\log(y_m)\Big)\Big)$$

This is a Gamma density $G(a_*, v_*)$ where:

$$\nu_* = \nu_0 + n \quad \alpha_* = \alpha_0 + \sum \log\left(y_i/y_m\right)$$

The conditional posterior of y_m given α is obtained by neglecting all the elements which are independent of y_m in the product of the likelihood function times the prior:

$$p(y_m|y, \alpha) \propto y_m^{\alpha n+\gamma_0-1} \Pi(y_m < y_i) \Pi(y_m < y_{m0})$$

We identify a Power function density $PF(g_*, y_{m*})$ with parameters:

$$\gamma_* = \gamma_0 + n\alpha \quad y_{m*} = \text{Max}(\text{Min}(y_i), y_{m0})$$

We note that the support of the conditional posterior density y_{m*} depends on the minimum value of the sample and on the value of y_{m0}. Collecting these results, inference on α and y_m is conducted using a Gibbs sampler. If y_m were given, inference would rely only on the Gamma posterior density $p(\alpha|y_m, y)$.

Motivations for Considering a Mixture of Two Pareto Distributions

We now analyze the statistical characterization of a mixture of two Pareto I distributions. As we have seen earlier, this model is justified by the use of

a combination of tournament and superstars theories for high wage formation. This is the joint model sketched by Atkinson (2008, pp. 93–95). Of course, a Pareto concerns only the upper tail of the income distribution (which is here our main interest). We have truncated the lower incomes and even more. We could have considered hybrid mixtures of lognormal and Pareto distributions. Harrison (1981) for instance considers both the lognormal and the Pareto distribution to analyze the earning distribution in the United Kingdom for 1972 coming from the New Earnings Survey. But he estimates separately these two distributions, considering truncated data. We have tried to implement an hybrid lognormal-Pareto mixture, but we could identify only one Pareto tail. So, the resulting statistical model was not in coherence with our economic model, certainly because one of the lognormal tails played the role of a Pareto tail. Mitzenmacher (2004) shows that a lognormal distribution can exhibit a Pareto tail by expanding its shape parameter, σ^2.

Mixtures of Two Pareto Distributions

A mixture of two distributions for the random variable Y consists in considering two random variables: the observed values of Y and an unobserved random variable Z which is assumed to follow a binomial process that allocates the observations between the two regimes. The conditional distribution of observation y_i given the value of z_i is simplify:

$$y_i|z_i = j \sim f_j(y|\theta_j)$$

If we now consider the marginal distribution of Y, when Z is integrated out, we have the usual mixture formulation:

$$f_Y(y|\theta) = pf_1(y|\theta_1) + (1-p)f_2(y|\theta_2)$$

with a probability mixing weight p. The conditional probability that $z_i = j$, with $j = 1, 2$, given each observation y_i is obtained as a normalized ratio:

$$P(z_i = 1|Y = y) = \frac{pf_1(y|\theta_1)}{pf_1(y|\theta_1) + (1-p)f_2(y|\theta_2)}$$

This formulation corresponds to the general case where every observation has the same marginal probability p of belonging to the first member. It is

used to build a Gibbs sampler in a Bayesian framework or an EM algorithm in a classical framework.

A Pareto mixture is a slightly different case, as the support of each member depends on a parameter. In a Bayesian framework, Noor and Aslam (2012) propose to analyze the following two-member rescaled Pareto II mixture:

$$f(y|\alpha_1,\alpha_2,\lambda,p)=p\alpha_1\frac{\lambda^{\alpha_1}}{(\lambda+y)^{(\alpha_1+1)}}\Pi(y>0)+(1-p)\alpha_2\frac{\lambda^{\alpha_2}}{(\lambda+y)^{(\alpha_2+1)}}\Pi(y>0)$$

But we note that in this formulation the two members have the same support where y and λ are just restricted to be positive. Moreover, Noor and Aslam (2012) assume that λ is known, so that their final model is a particular case of a mixture of two Pareto I densities.

The type of mixture we are interested in is slightly different and more general as we consider the case with different shape and scale parameters as given by Bee et al. (2013):

$$f(y|\alpha_1,\alpha_2,y_{m1},y_{m2},p)=p\alpha_1\frac{y_{m1}^{\alpha_1}}{y^{(\alpha_1+1)}}\Pi(y>y_{m1})+(1-p)\alpha_2\frac{y_{m2}^{\alpha_2}}{y^{(\alpha_2+1)}}\Pi(y>y_{m2}) \quad (5)$$

In Eq. (5) the two components have a different support, so it is natural to assume for instance that $y_{m2}>y_{m1}$, even if y_{m2} could be as close as possible from y_{m1} which thus corresponds to the minimum observation of the complete sample (this is also the assumption made by Bee et al., 2013). In this framework, the first member is concerned with observations greater than y_{m1} while the second component corresponds to observations greater than y_{m2}. So, any observation y_i such that $y_{m1}<y_i<y_{m2}$ belongs to the first regime with probability 1 and not with probability p. It is not surprising that under these conditions, Bee et al. (2013) report that the usual EM algorithm does not work for estimating the five parameters of Eq. (5) and works only when y_{m2} is known. Bee et al. (2013) proved that in the bivariate case, the EM algorithm does not update the value found for the estimation of y_{m2}. In the M step, the objective function is not differentiable in y_{m2}.

This motivates our concern for presenting first a Gibbs sampler when y_{m2} is fixed and given a priori. We shall then try to investigate the case where both y_{m1} and y_{m2} are unknown in order to see if a prior information on y_{m2} is enough in order to get convergence of the Gibbs sampler.

Remark:

The mixture we consider is that of Bee et al.'s (2013) and implicitly also that of Atkinson's (2008, pp. 93–95). However, as we suppose that $y_{m2} > y_{m1}$, we could have also consider a hybrid mixture of a Pareto I and a generalized Pareto. The argument might be derived from Pickands (1975) and extreme value theory. The generalized Pareto that Pickands (1975) (see also Arnold, 2008) considers is:

$$f(x|\beta,\sigma) = \frac{1}{\sigma}\left(1+\beta\frac{x}{\sigma}\right)^{-(1+\beta)/\beta}\Pi(x>0,\beta x>\sigma), \quad \beta, \quad \sigma>0$$

It also corresponds to a particular case of the Pareto II which is:

$$f(x|\alpha,\sigma,y_{m2}) = \frac{\alpha}{\sigma}\left[1+\frac{x-y_{m2}}{\sigma}\right]^{-(\alpha+1)}\Pi(x>y_{m2})$$

and that it collapses down to the Pareto I for $\sigma = y_{m2}$. This density is not convenient for inference as it is not possible to find sufficient statistics as underlined by Arnold and Press (1983). So, a mixture combining a Pareto I and a Pareto II would not be a convenient model. We prefer to stick to the original model of Bee et al. (2013).

For the moment, let us give in Fig. 1 a graphical representation of Eq. (5) with $p = 0.80$, $\alpha_1 = 3$, $y_{m1} = 30$, $\alpha_2 = 1$, and $y_{m2} = 80$. A mixture of two densities is usually a smooth function of y. The very particular shape of the Pareto and the fact that its support is parameter dependent produce a discontinuity in the graph which is well present here at $y = y_{m2}$. The height of the discontinuity depends on the value of $1-p$. So, a small jump is coherent with a large p.

Bayesian Inference for y_{m2} *Known*

If y_{m2} is known, we have a conceptual simplification of the problem. We are in one of the cases investigated by Bee et al. (2013) in a classical framework. Let us consider an observation y_i. If $y_i < y_{m2}$, then for sure, this observation belongs to the first component of the mixture. If $y_i > y_{m2}$, there

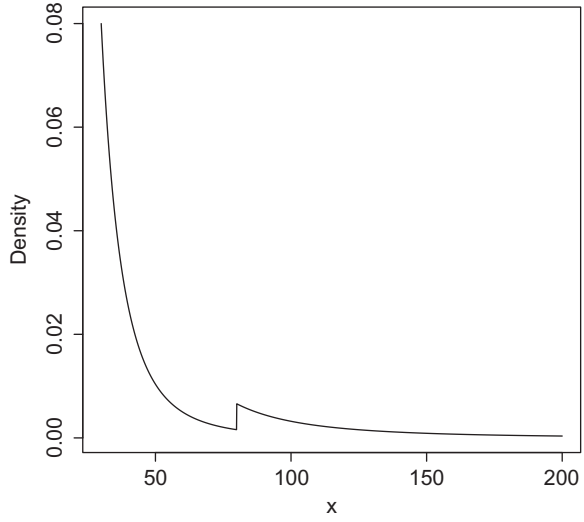

Fig. 1. Mixture of Two Pareto.

is a certain probability that it belongs to the first regime and a complementary probability that it belongs to the second regime. The probability $P(z_i = j|Y = y_i)$ can be computed only for the observations for which $y_i > y_{m2}$. Consequently, the random allocation of the observations between the two regimes which is at the heart of the usual Gibbs sampler in mixture problems can only be applied for that part of the sample. Because the support of the density depends on the parameters, we have a different sample allocation, different from the usual algorithm.

As in the usual framework, we introduce a Beta prior on p with prior parameters n_{o1} and n_{o2}. The prior moments of p are:

$$E(p) = \frac{n_{o1}}{n_{o1} + n_{o2}} \quad \text{Var}(p) = \frac{n_{o1} n_{o2}}{(n_{o1} + n_{o2})^2 (n_{o1} + n_{o2} + 1)}$$

The posterior density of p is also a Beta density with posterior parameters $n_{o1} + n_1$ and $n_{o2} + n_2$. n_1 and n_2 are the number of observations that conditionally on z fall into each regime. n_{1r} represents the number of observations that are randomly allocated to the first regime as a function of z, while n_{1s} represents the number of observations that are for sure allocated

to the first regime because they are lower than y_{m2}, so that $n_1 = n_{1s} + n_{1r}$. We propose the following algorithm:

Gibbs Sampler Algorithm with y_{m2} *Known*

1. Fix a value for the total number of draws m, fix a value for y_{m2}, select a starting value for p, and compute the following starting values $y_{m1} = y_{(1)}$, $\alpha_1 = \hat{\alpha}(y_{m1})$, $\alpha_2 = \hat{\alpha}(y_{m2})$.
2. Determine the vector of observations $y_{1s}|y < y_{m2}$ that belong for sure to the first regime. The Gibbs sampler will run only for the remaining vector of observations $y_{12}|y > y_{m2}$ as y_{m2} is given and fixed.
3. Start the loop on j, the Gibbs iterations.
4. For the remaining observations, simulate the sample allocation $z^{(j)}$ where each element is drawn according to a Binomial$(z_i|p_i)$ with base probability:

$$p_i = \frac{p^{(j-1)} f_1(y_{12i}|\alpha_1^{(j-1)}, y_{m1}^{(j-1)})}{p^{(j-1)} f_1(y_{12i}|\alpha_1^{(j-1)}, y_{m1}^{(j-1)}) + (1 - p^{(j-1)}) f_2(y_{12i}|\alpha_2^{(j-1)}, y_{m2}^{(j-1)})}$$

5. Select the sub-sample separation $y_{1r}^{(j)}$ and $y_{2r}^{(j)}$ among the y_{12}.
6. Form the first regime allocation $y_1^{(j)} = (y_{1s}, y_{1r}^{(j)})$ which is partially fixed and the second regime allocation $y_2^{(j)} = y_{2r}^{(j)}$ which is random.
7. Compute $n_1^{(j)} = n_{1s}^{(j)} + n_{1r}^{(j)}$ and $n_2^{(j)}$.
8. Draw $p^{(j)} \sim \text{Beta}(n_1^{(j)} + n_{o1}, n_2^{(j)} + n_{o2})$.
9. Draw $y_{m1}^{(j)} \sim \text{Power}(n_1^{(j)} * \alpha_1^{(j-1)} + \gamma_1^o, \text{Max}(\text{Min}(y_1^{(j)}), y_{mo_1}))$.
10. Draw $\alpha_1^{(j)} \sim \text{Gamma}(\alpha_1^o + \sum \log(y_1^{(j)}/y_{m1}^{(j)}), \nu_{01} + n_1)$.
11. Draw $\alpha_2^{(j)} \sim \text{Gamma}(\alpha_2^o + \sum \log(y_2^{(j)}/y_{m2}^{(j)}), \nu_{02} + n_2)$.
12. $j = j + 1$ End of loop.

As a check, we have simulated a sample of size $n = 1{,}000$ with $y_{m1} = 30$, $y_{m2} = 80$, $\alpha_1 = 3$, $\alpha_2 = 2$, and $p = 0.80$. We have chosen a rather weak prior information with $y_{mo} = (30, 80)$ and $\gamma_o = (1, 1)$ for the Power prior on y_m and $\nu_o = (5, 5)$, $\alpha_o = (1, 2)$ for the Gamma prior on α and $n_o = (1, 1)$ for the Beta prior on p. We took $3{,}000 + 500$ draws for the Gibbs. We get $E(\alpha_1|y) = 2.98$, (0.15), $E(\alpha_2|y) = 2.10$, (0.17), $E(y_{m1}|y) = 30.01$, (0.012), and $E(p|y) = 0.83$, (0.016), which all are close to the values used for the generating process. The algorithm converges without any problem, as checked with CUMSUM graphs.

Bayesian Inference for Both y_{m1} *and* y_{m2} *Unknown*

Considering y_{m2} as a random parameter modifies slightly the previous algorithm. The observations that belong for sure to the first component of the mixture, $y_{1s}|y<y_{m2}$ have to be determined at each iteration, for each random value of y_{m2}. So once y_{m2} is drawn, the algorithm follows the same logic.

General Gibbs Sampler Algorithm

1. Fix a value for the total number of draws m, fix a value for y_{m2}, select a starting value for p, and compute the following starting values $y_{m1} = y_{(1)}$, $\alpha_1 = \hat{\alpha}(y_{m1})$, $\alpha_2 = \hat{\alpha}(y_{m2})$.
2. Start the loop on j, the Gibbs iterations.
3. Determine the observations $y_{1s}|y<y_{m2}$ that belong for sure to the first regime for a given draw of y_{m2}. Determine the remaining observations $y_{12}|y>y_{m2}$.
4. For the remaining observations y_{12}, simulate the sample allocation $z^{(j)}$ where each element is drawn according to a Binomial$(z_i|p_i)$ with base probability:

$$p_i = \frac{p^{(j-1)} f_1(y_{12i}|\alpha_1^{(j-1)}, y_{m1}^{(j-1)})}{p^{(j-1)} f_1(y_{12i}|\alpha_1^{(j-1)}, y_{m1}^{(j-1)}) + (1-p^{(j-1)}) f_2(y_{12i}|\alpha_2^{(j-1)}, y_{m2}^{(j-1)})}$$

5. Select the sub-sample separation $y_{1r}^{(j)}$ and $y_{2r}^{(j)}$ among the y_{12}.
6. Form the first regime allocation $y_1^{(j)} = (y_{1s}, y_{1r}^{(j)})$ and the second regime allocation $y_2^{(j)} = y_{2r}^{(j)}$.
7. Compute $n_1^{(j)} = n_{1s}^{(j)} + n_{1r}^{(j)}$ and $n_2^{(j)}$.
8. Draw $p^{(j)} \sim \text{Beta}(n_1^{(j)} + n_{o1}, n_2^{(j)} + n_{o2})$.
9. Draw $y_{m1}^{(j)} \sim \text{Power}(n_1^{(j)} * \alpha_1^{(j-1)} + \gamma_1^o, \text{Max}(\text{Min}(y_1^{(j)}), y_{mo_1}))$.
10. Draw $y_{m2}^{(j)} \sim \text{Power}(n_2^{(j)} * \alpha_2^{(j-1)} + \gamma_2^o, \text{Max}(\text{Min}(y_2^{(j)}), y_{mo_2}))$.
11. Draw $\alpha_1^{(j)} \sim \text{Gamma}(\alpha_1^o + \sum \log(y_1^{(j)}/y_{m1}^{(j)}), \nu_{01} + n_1)$.
12. Draw $\alpha_2^{(j)} \sim \text{Gamma}(\alpha_2^o + \sum \log(y_2^{(j)}/y_{m2}^{(j)}), \nu_{02} + n_2)$.
13. $j = j + 1$ End of loop.

Using the same parameters to draw the random sample and the same prior information as before, we manage to have a similar convergence of the algorithm. Again with $3{,}000 + 500$ draws for the Gibbs, we get

$E(\alpha_1|y) = 3.09$, (0.16), $E(\alpha_2|y) = 2.13$, (0.16), $E(y_{m1}|y) = 29.99$, (0.012), $E(y_{m2}|y) = 80.05$, (0.41), and $E(p|y) = 0.80$, (0.015). Of course, there is more uncertainty for the estimation of y_{m2} than for y_{m1}. This is why at least a weak prior information is needed.

We have now to explore the role of the prior information to insure a proper behavior of the posterior density of the scale parameters y_m. First of all, in order to have a proper posterior, we must have $\gamma_1^o, \gamma_2^o > 1$, so as to avoid having a sample configuration where the Power function could have a shape parameter lower than 1, leading to a bimodal posterior (see Arnold & Press, 1983 for a similar result). Second, apparently any value for y_{mo1} between 0 and $\text{Min}(y_i)$ can be chosen without any influence on the results. This is due to the way one part of the observations in regime 1 is chosen and to the expression for the conditional posterior density of y_{m1}. Choosing a $y_{mo1} > \text{Min}(y_i)$ does have an influence on the posterior results and eventually leads to an abnormal termination of the algorithm. The case of y_{mo2} is of course more delicate. For values of y_{mo2}, greater than the value used for generating the data, the algorithm either does not converge or converges to wrong values. For values of y_{mo2} lower than that used for generating the data, the algorithm converges to reasonable values and seems to be rather insensitive to the choice of y_{mo2}, provided this choice is not too far from the truth (between 50 and 80 in our case). The other priors seem to play only a minor role. These simulation examples are rather encouraging, especially when comparing them to the usual EM algorithm (see Fig. 2 below).

Eliciting Prior Information

It is very difficult to estimate a mixture without prior information and the usual practice consists in computing sample moments and using them to provide identical information for each mixture component, at the cost of favoring label switching. For a Pareto mixture, we have some extra tools, one of which being the graph of $\log(1-\hat{F})$ versus $\log(y)$. A natural estimate for the cumulative distribution F is very easy to obtain. The Pareto distribution belongs to the power family and thus its log is an affine function of $\log(y)$. The regression of $\log(1-\hat{F})$ over $\log(y)$ is a common device to estimate the Pareto coefficient. The situation is slightly more complex here as the complementary cumulative distribution is given by:

$$1 - F(y) = p(y_{m1}^{\alpha_1} y^{-\alpha_1})\Pi(y > y_{m1}) + (1-p)(y_{m2}^{\alpha_2} y^{-\alpha_2})\Pi(y > y_{m2})$$

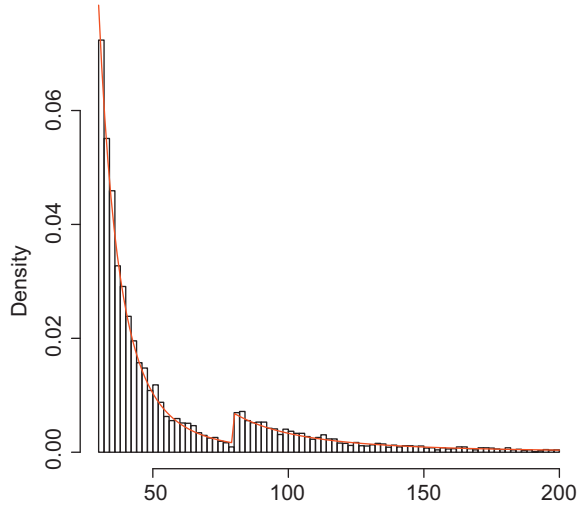

Fig. 2. Fitting a Mixture of Two Pareto Distributions.

Taking the log of this expression of course is not similar to taking the log of the complementary distribution of a single Pareto because we have a sum of two elements. So, we cannot infer the same type of information as we could derive from the same graph corresponding to a single Pareto process. However, we can infer information on the localization of y_{m2}. This can be checked with our simulated sample as shown in Fig. 3. The vertical line corresponds to $x = \log(80)$. There is a clear kink in the curve corresponding to y_{m2}. So, this type of graph can be used to determine a prior value for y_{m2} which is a crucial information as we have shown before. But it cannot be used to extract further information on the value of the Pareto coefficients.

EMPIRICAL APPLICATION: TOP EARNINGS INEQUALITY IN THE UNITED STATES 1992–2009

Over the past two decades, the United States experienced a sharp rise in wage inequality accompanied by large increase in wage differentials in the upper tail. We illustrate the approach developed above by approximating the upper tail of the U.S. wage distribution by a mixture of two Pareto distributions.

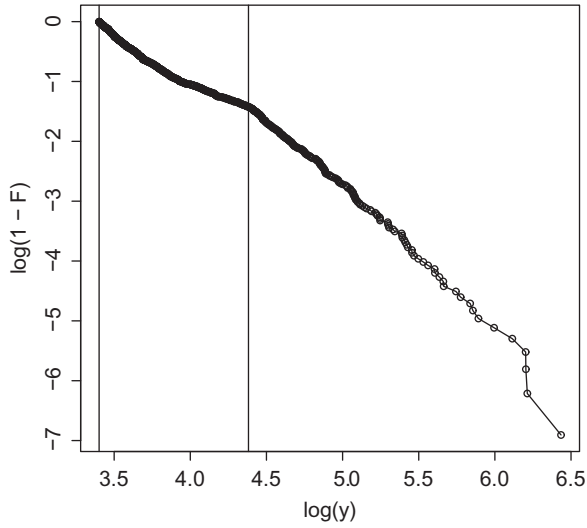

Fig. 3. A Log-Log Plot of the Complementary Cumulative Distribution of Y.

The Data

This paper uses the Current Population Surveys (CPS),[7] Outgoing Rotation Groups (ORG).[8] We take the monthly earnings files for January 1992 through May 2009. We decide to focus our attention on three years (1992, 2001, 2009) to cover the main features of the recent period and their evolution. We use the weekly wage divided by the number of hours worked in order to get an homogeneous definition of hourly wages.[9] We deflate these wages by the annual average CPI which values are respectively 140.2, 177.1, and 214.5 for these three years.

Between 1992 and 2009, we have a constant rise of real wage together with a sharp increase in inequality in the upper tail at the end of the period. As displayed in Table 1, we should notice that the number of respondents in 2009 is much lower (by 20%) than in 1992 and 2001. This might create a selection bias. However, we shall see in Table 2 that when we restrict our attention to the upper tail, the number of respondents becomes quite comparable over the different samples.

In Fig. 4, we display a non-parametric estimate of the wage density, which is characterized by a heavy right tail. Even if it is difficult to have a precise estimate of this tail, we can have a glance at the huge difference contained in the 2009 data.

Table 1. Hourly Real Wage Dispersion for the U.S. Recent Period.

	1992	2001	2009
Mean	18.00	20.05	24.49
Median	14.52	15.53	15.62
$q_{0.75}$	22.21	24.22	26.49
$q_{0.90}$	31.37	36.01	46.12
$q_{0.95}$	39.75	46.57	67.25
$q_{0.99}$	63.32	77.62	144.23
Gini	0.352	0.369	0.455
N	62,107	63,409	47,837
% of Female	53.03	52.43	52.34

Table 2. Top Wage Dispersion in the Upper Tail of the U.S. Wage Distribution.

	1992		2001		2009	
Min	25.80	(0.828)	24.55	(0.764)	22.20	(0.673)
q_{25}	28.67	(0.868)	28.52	(0.823)	27.00	(0.755)
Mean	39.83	(0.951)	41.42	(0.931)	49.26	(0.913)
Median	33.10	(0.912)	33.76	(0.881)	34.60	(0.836)
q_{75}	41.99	(0.958)	43.66	(0.939)	50.48	(0.918)
q_{90}	55.17	(0.985)	58.23	(0.977)	75.00	(0.967)
q_{95}	65.33	(0.992)	72.76	(0.988)	110.55	(0.984)
q_{99}	122.31	(0.998)	141.30	(0.998)	278.05	(0.997)
Gini	0.214		0.240		0.352	
N	11,079		14,945		15,634	
% of Female	46.65		45.10		45.60	

Note: In parenthesis, we have given the position of the left number in the original un-truncated distribution of hourly wages.

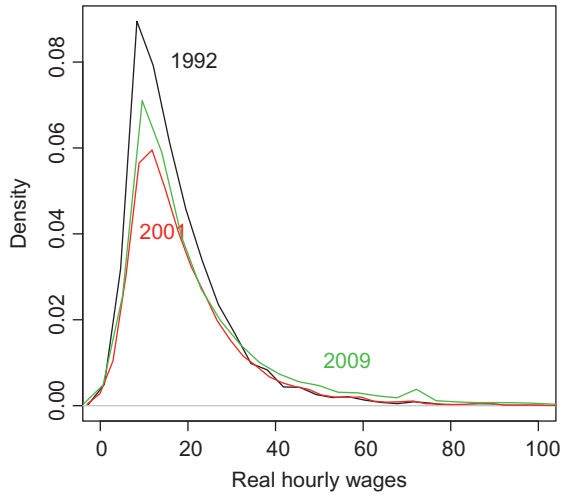

Fig. 4. Real Wage Density Estimates.

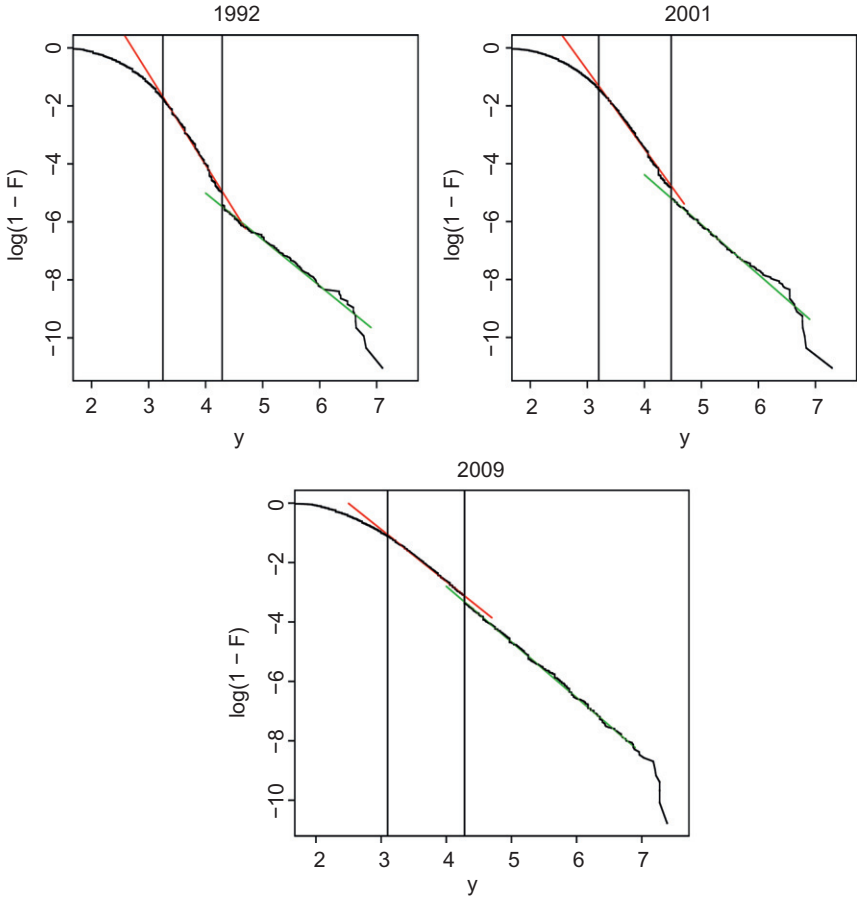

Fig. 5. Power Tails in the ORG Wages for 1992, 2001, and 2009.

In Fig. 5, we have plotted the complementary cumulative distribution function $\log(1-F)$ on the $\log(y)$ for the same three years. From these graphs, we can extract two pieces of information: the first point at which we have to cut our sample (y_{m1}) in order to use a Pareto model and then at which point there can be a change in the Pareto coefficients. This graphical information is translated into numerical values in Table 3. We note that with this graphical method, the number of superstars as reported in the last column of Table 3 is very small, compared to the total sample size, except

Table 3. Cutting Points for the Pareto Tails in the Complete Distribution.

Years	y_{m1}	$Pr(y < y_{m1})$	y_{m2}	$Pr(y < y_{m2})$	n_2
1992	exp(3.25)	0.83	exp(4.29)	0.996	273
2001	exp(3.20)	0.76	exp(4.47)	0.995	346
2009	exp(3.10)	0.67	exp(4.28)	0.966	1,637
2009	exp(3.10)	0.67	exp(4.58)	0.980	936

perhaps for 2009. However, we propose for this year two different prior for y_{m2} as we were obliged to modify it during the estimation due to a contradiction between the sample and the prior information.

At the discontinuous point, Fig. 5 shows a large number of observations (high wage earners) which could illustrate the impact of tournaments in hierarchical organizations, that is, the difficulty of changing a position in the hierarchy.

Modeling the Upper Tail of the U.S. Wage Distribution

We summarize the information contained in the truncated samples that we shall use (dropping all the observations lower than y_{m1}) in Table 2. The first quartile (in the truncated distribution) and the median remain constant over time. The distribution is changing over time only after $q_{0.75}$, for the highest hourly wages.

We elicit prior information in the following way. We assume that the Pareto coefficient is lower for the superstar model than for the hierarchical model which means, $\alpha_{o2} < \alpha_{o1}$. The number of superstars is assumed to be lower than the number of overheads while the mean wage of the former is greater than that of the latter. So we assume that y_{mo2} is strictly greater than y_{mo1}. We have chosen the value of y_{mo2} following the indications provided in Fig. 5. The role of p is now ambiguous as it is both the total proportion of observations in regime 1 and it plays also a role in determining the proportion of observations in regime 1 that are greater than y_{m2}. So we put a prior expectation of 0.50. All this motivates the selection of the following prior information. The prior value for y_{mo} is different among the three samples and given in Table 3. We have in contrast taken the same prior information for the different samples concerning the other parameters (see Table 4 below).

Table 4. Hyperparameters and Resulting Prior Moments.

	Hyperparameter	γ_{o1}	γ_{o2}	α_{o1}	ν_{o1}	α_{o2}	ν_{o2}	n_{o1}	n_{o2}
	Value	100	100	10	30	15	30	100	100
	Parameter	y_{m1}	y_{m2}	α_1		α_2		p	
1992	Mean	25.54	72.24	3.00		2.00		0.50	
	S.D.	(0.25)	(0.72)	(0.55)		(0.37)		(0.035)	
2001	Mean	24.29	86.49	3.00		2.00		0.50	
	S.D.	(0.24)	(0.86)	(0.55)		(0.37)		(0.035)	
2009	Mean	21.98	71.53	3.00		2.00		0.50	
	S.D.	(0.22)	(0.71)	(0.55)		(0.37)		(0.035)	
2009	Mean	21.98	94.64	3.00		2.00		0.50	
	S.D.	(0.22)	(0.94)	(0.55)		(0.37)		(0.035)	

Note: Prior moments for y_{m1} and y_{m2} were computed using the information contained in Table 3 together with the values of γ_{o1} and γ_{o2}.

Table 5. Posterior Inference for Mixture of Two Pareto Distributions and Corresponding Gini Indices.

	y_{m1}	y_{m2}	α_1	α_2	p	π	I_{G_1}	I_{G_2}
1992	25.80	73.18	3.08	1.50	0.978	0.013	0.194	0.509
	(0.001)	(0.25)	(0.032)	(0.14)	(0.002)	(0.001)	(0.0024)	(0.076)
2001	24.55	87.36	2.46	1.23	0.988	0.006	0.255	0.691
	(0.001)	(−)	(0.021)	(0.13)	(0.001)	(0.001)	(0.003)	(0.12)
2009	22.19	95.91	2.42	1.76	0.980	0.014	0.268	0.400
	(0.001)	(0.16)	(0.31)	(0.016)	(0.002)	(0.002)	(0.043)	(0.0049)

Notes: π is the proportion of superstars computed as the mean number of superstars n_2 divided by the sample size n. Results for 2001 were obtained with a fixed y_{m2}. Otherwise, the algorithm did not converge.

We use 5,000 draws for the main chain plus 500 draws to warm up the chain. The algorithm seems still to converge quite well, but we were obliged to consider y_{mo2} as known for 2001. The posterior mean of each parameter with the posterior standard error in parentheses are summarized in Table 5. Using the Gibbs output, we have computed a Gini index for each group of the mixture, corresponding to each of the α:

$$I_{G_k} = \sum_{t=1}^{m} \frac{1}{2\alpha_k^{(t)} - 1} \tag{6}$$

Table 6. Posterior Mean Wages for the Two Groups.

	High Wage Earners	Superstars
1992	37.58	210.59
	(0.14)	(11.51)
2001	39.97	299.62
	(0.16)	(22.72)
2009	47.59	163.15
	(0.28)	(14.05)

Notes: Standard deviations are given between parenthesis. These figures are computed inside the Gibbs loop, using the sample separation. If it were computed using the posterior draws of α and y_m, for values of α getting close to 1, the resulting estimates would have been very unstable and thus unreliable.

imposing the restriction $\alpha_k^{(t)} > 1$ for the evaluation. Inequality is greater in the second group (superstars) for 1992 and 2001. The implied Gini steadily increases over time, which is coherent with the literature. The mean wage of the first group increases as reported in Table 6. When we turn to the superstars group, the situation is more contrasted. First, their number, as detected in the sample, is very small, less than 2% of the truncated sample representing 145, 84, and 224 persons. Their mean wages has increased a lot in 2001, but has dropped in 2009. Inequality which was much higher than in the first group and also increasing in 2001, suddenly dropped to a much lower level in 2009 after the financial crisis. Table 6 shows a remarkable differences of the mean wages between the two groups.

CONCLUSION AND DISCUSSION

The distribution of earnings takes a Pareto form in the model of tournament in hierarchical organizations and in the model of "superstars." This paper has provided a Bayesian inference for a mixture of two Pareto distributions to approximate the upper tail of a wage distribution. This mixture model is applied to the data from the CPS Outgoing Rotation Group to analyze the recent structure of top wages in the United States from 1992 through 2009. We find enormous wage disparities between two groups: the very highest wage earners ("superstars") and the other high wage earners. These findings are largely in line with the recent literature explaining wage

inequalities in the upper tail of wage distributions, as the recent pay differences of executives among large firms.

Inference for mixture of Pareto distributions is very scarcely covered, as underlined by Bee et al. (2013) the EM algorithm cannot be applied if the largest of the two parameters is not known. Bayesian inference for a mixture of Pareto is sensitive to the choice of prior information. This paper has provided a careful elicitation procedure on the parameter's priors of the mixture.

NOTES

1. Atkinson (2008) argues that "a constantly rising demand for more educated and more skilled workers does not lead to a constantly rising wage premium, but to a stable wage differential, and countries that adjust more rapidly will see, on a continuing basis, smaller wage differences". Lubrano and Ndoye (2014) find that most of the recent changes in U.S. between 1992 to 2009 have occurred at the top decile, and, the decline in unionization had a weak impact.
2. Tervio (2008) shows that the rise in earnings dispersion in large U.S. firm is largely due to the increase in the span of control of firms and to the tremendous rise in CEOs pay (see also Fox, 2009; Gabaix & Landier, 2008).
3. Tournament theory was first introduced by Lazear and Rosen (1981). It describes a competition between executives. Their pay is determined by their rank in the hierarchy which is determined by their personal performances. High wages are seen as a prize for the winners of the tournament.
4. Marshall (1947) first pointed out the idea of superstar.
5. In the sequel, we shall use the notation $\hat{\alpha}(y_m)$. We mean that $\hat{\alpha}$ is computed using $\sum \log(y_i/y_m)$ where y_i is restricted to the sub-sample $y_i > y_m$.
6. It is simple to draw random numbers using the inverse transform method with $x = x_m u^{1/\alpha}$ and $u \sim U(0,1)$. For a Pareto process, we have $y = y_m u^{-1/\alpha}$.
7. The CPS is the monthly household survey conducted by the Bureau of Labour Statistics to measure labour force participation and employment. 50-60,000 households per month are queried. This is not really a panel survey since households are not followed if they move. They include the March CPS file and the Outgoing Rotation Group (ORG) files.
8. The ORG files correspond to the set of every household that enters the CPS interviewed each month for 4 consecutive months, and then ignored for 8 months.
9. The ORG files are often used because they include a direct observation of the hourly wage, which thus has not to be computed as the ratio between the weekly wage and the number of worked hours. However, many individuals did not answer to that question, so we prefer to compute a ratio in order to keep the maximum number of observations. And anyway, apart from a few aberrant values, our ratio series gave similar figures as the one given by the hourly series. Atkinson (2008, p. 401) reports different bibliographic sources showing that the ORG data are the most accurate to study the evolution of the U.S. wage structure.

ACKNOWLEDGMENTS

We are grateful to Anthony Atkinson and Stephen Bazen for pointing out several references, data sources and providing stimulating comments. We benefited also from comments of Karim Abadir and Christian Robert. We thank an anonymous referee for constructive comments. Of course, the usual disclaimers apply.

REFERENCES

Adler, M. (1985). Stardom and talent. *American Economic Review*, *75*(1), 208–212.

Arnold, B. C. (2008). Pareto and generalized Pareto distributions. In D. Chotikapanich (Ed.), *Modeling income distributions and Lorenz curves* (pp. 145–199). New York, NY: Springer.

Arnold, B., Castillo, E., & Sarabia, J. (1998). Bayesian analysis for classical distributions using conditionally specified priors. *Sankhya, Series B*, *60*, 228–245.

Arnold, B., Castillo, E., & Sarabia, J. (1999). *Conditional specification of statistical models*. Springer Series in Statistics. New York, NY: Springer Verlag.

Arnold, B., Castillo, E., & Sarabia, J. (2001). Conditionally specified distributions: An introduction (with discussion). *Statistical Science*, *16*, 151–169.

Arnold, B., & Press, S. (1983). Bayesian inference for Pareto populations. *Journal of Econometrics*, *21*, 287–306.

Arnold, B., & Press, S. (1989). Bayesian estimation and prediction for Pareto data. *Journal of the American Statistical Association*, *84*, 1079–1084.

Atkinson, A. B. (2008). *The changing distribution of earnings in OECD countries*. Oxford: Oxford University Press.

Bauwens, L. (1991). The "pathology" of the natural conjugate prior density in the regression model. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, *23*, 49–64.

Bee, M., Benedetti, R., & Espa, G. (2013). On maximum likelihood estimation of a Pareto mixture. *Computational Statistics*, *28*(1), 161–178

Bound, J., & Johnson, G. (1992). Changes in the structure of wages in the 1980's: An evaluation of alternative explanations. *American Economic Review*, *82*(3), 371–392.

Brian, G., Charles, A. O. I., & James, W. (1993). Top executive pay: Tournament or teamwork? *Journal of Labor Economics*, *11*(4), 606–628.

Brown, J. (2011). Quitters never win: The (adverse) incentive effect of competing with superstars. *Journal of Political Economy*, *119*(5), 982–1013.

Burkhauser, R. V., Feng, S., Jenkins, S. P., & Larrimore, J. (2008). *Estimating trends in US income inequality using the current population survey: The importance of controlling for censoring*. Working Paper No. 14247, NBER.

DiNardo, J., Fortin, N., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, *64*(5), 1001–1044.

Drèze, J. H. (1977). Bayesian regression analysis using poly-t densities. *Journal of Econometrics*, *6*(3), 329–354.

Embrechts, P., Kluppelberg, C., & Mikosch, T. (1997). *Modelling extremal events*. Berlin: Springer Verlag.
Fox, J. T. (2009). Firm-size wage gaps, job responsibility, and hierarchical matching. *Journal of Labor Economics*, *27*(1), 83–126.
Frank, R. H., & Cook, P. J. (1995). *The winner take-all society*. New York, NY: Free Press.
Gabaix, X., & Landier, A. (2008). Why has CEO pay increased so much? *Quarterly Journal of Economics*, *123*, 49–100.
Harrison, A. (1981). Earnings by size: A tale of two distributions. *The Review of Economic Studies*, *48*(4), 621–631.
Katz, L. F., & Autor, D. H. (1999). Changes in the wage structure and earnings inequality. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1463–1555). Amsterdam: Elsevier.
Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, *89*, 841–864.
Lubrano, M., & Ndoye, A. A. J. (2014). Bayesian unconditional quantile regression: An analysis of recent expansions in wage structure and earnings inequality in the U.S. 1992–2009. *Scottish Journal of Political Economy*, *61*(2), 129–153.
Lwin, T. (1972). Estimating the tail of the Paretian law. *Scandinavian Actuarial Journal*, *55*, 170–178.
Lydall, H. F. (1959). The distribution of employment incomes. *Econometrica*, *27*(1), 110–115.
Lydall, H. F. (1968). *The structure of earnings*. Oxford: Clarendon Press.
Marshall, A. (1947). *Principles of economics* (8th ed.). New York, NY: MacMillan.
Mincer, J. (1993). Human capital, technology, and the wage structure: What do time series show? In J. Mincer (Ed.), *Studies in human capital*. Brookeld, VT: Edward Elgar Publishing.
Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*(2), 226–251.
Murphy, K., & Welch, F. (1992). The structure of wages. *Quarterly Journal of Economics*, *107*, 285–326.
Nadarajah, S. (2006). Information matrices for laplace and pareto mixtures. *Computational Statistics and Data Analysis*, *50*, 950–966. Retrieved from www.elsevier.com/locate/csda
Nair, M. T. (2007). *On finite mixture of Pareto and Beta distributions*. PhD thesis, Department of Statistics, Cochin University of Science and Technology, Cochin, Kerala.
Noor, F., & Aslam, M. (2012). Bayesian analysis of two parameter pareto mixture using censoring. *International Journal of Physical Sciences*, *7*(44), 5878–5891.
Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, *3*(1), 119–131.
Reed, W. J., & Jorgenson, M. (2004). The double pareto-lognormal distribution – A new parametric model for size distributions. *Communications in Statistics – Theory and Methods*, *33*(8), 1733–1753.
Rosen, S. (1981). The economics of superstars. *The American Economic Review*, *71*(5), 845–858.
Rosen, S. (1986). Prizes and incentives in elimination tournaments. *American Economic Review*, *76*, 701–715.
Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, *31*, 831–880.
Simon, H. A. (1957). The compensation of executives. *Sociometry*, *20*, 32–35.
Tervio, M. (2008). The difference that CEOs make: An assignment model approach. *American Economic Review*, *98*(3), 642–648.