Maximum Likelihood Estimation for q-Exponential (Tsallis) Distributions

Cosma Rohilla Shalizi

*Statistics Department, Carnegie Mellon University**

(Dated: Begun 28 December 2006, last updated 31 January 2007)

This expository note describes how to apply the method of maximum likelihood to estimate the parameters of the "q-exponential" distributions introduced by Tsallis and collaborators. It also describes the relationship of these distributions to the classical Pareto distributions.

In a series of papers beginning with [1], Constantino Tsallis and collaborators introduced what have come to be called q-exponential probability distributions. These can be defined through their "complementary" ("upper", "upper cumulative") distribution functions, also called "survival" functions:

$$P_{q,\kappa}(X \geq x) = \left(1 - \frac{(1-q)x}{\kappa}\right)^{1/(1-q)} \qquad (1)$$

Tsallis *et al.* proposed these distributions to handle statistical-mechanical systems with long-range interactions, necessitating (it is claimed) a non-extensive generalization of the ordinary Gibbs-Shannon entropy. Following Jaynes's procedure of maximizing an entropy subject to constraints on expectation values [2], they got the q-exponential distributions, in which κ enforces the constraints, and q measures the departure from extensivity, Boltzmann-Gibbs statistics being recovered as $q \to 1$.

Tsallis's ideas about non-extensive entropy and its possible applications, in and out of statistical mechanics, have attracted intense (not to say "extensive") interest in physics; the bibliography at `http://tsallis.cat.cbpf.br/biblio.htm` has over 2000 entries. They are also quite controversial (see, e.g., Refs. [3, 4, 5, 6, 7, 8], the replies by Tsallis and others, and in some cases the replies to the replies). Whether or not the critics are correct, however, q-exponentials are still valid probability distributions, and can usefully describe some empirical phenomena. To this end, in a recent paper Douglas R. White *et al.* pose the problem of estimating the parameters q and κ from data by the method of maximum likelihood [9]. This note solves that problem.

I first reparameterize Eq. 1 to simplify estimation and emphasize links to Pareto distributions. I then rehearse the math of finding the maximum likelihood estimator (MLE) for the q-exponential distribution, discussing its accuracy and precision, and adjustments for data in which samples below a fixed threshold are all dropped ("censoring"). I compare maximum-likelihood estimates to those found by the current practice of curve-fitting; the latter are inferior. Finally, I discuss testing the assumption that the data are q-exponentially distributed. Code implementing the MLE for q-exponentials is available at `http://bactra.org/research/tsallis-MLE/`, written in R, a free, open-source programming language for statistical computing (`http://www.r-project.org/`). This code also calculates probabilities and quantiles, generates random numbers, etc.

a. Reparameterization as Generalized Pareto Distributions While it is possible to find the MLE for q-exponentials in the form given in Eq. 1, the algebra is needlessly messy. It is simpler to reparameterize, and change back to the original parameter system at the end, if desired. (Under a 1-1 change of parameters, an MLE for the old parameters must, under the transformation, be an MLE for the new parameters, and vice versa.) Thus, define the new parameters $\theta \equiv -\frac{1}{1-q}$ and $\sigma \equiv \theta\kappa$, from which the original parameters can be recovered:

$$q = 1 + \frac{1}{\theta}, \ \kappa = \frac{\sigma}{\theta} \qquad (2)$$

In the new parameter system, the survival function becomes

$$P_{\theta,\sigma}(X \geq x) = (1 + x/\sigma)^{-\theta} \qquad (3)$$

Hence the probability density is

$$p_{\theta,\sigma}(x) = \frac{\theta}{\sigma}(1 + x/\sigma)^{-\theta-1} \qquad (4)$$

The code mentioned above uses both parameterizations.

Y has a Pareto distribution with scaling exponent α and cut-off y_0 if $p(y) = 0$ when $y < y_0$, and otherwise $p(y) \propto (y/y_0)^{-\alpha-1}$. Hence when X has a q-exponential distribution, $1 + x/\sigma$ has a Pareto distribution with cut-off 1 and scaling exponent θ. Following the classification given in Arnold's monograph on Pareto distributions [10], this is an instance of a "type II generalized Pareto", often used in operations research on failure times and other reliability problems, the standard form of which is $P(X \geq x) = [1 + (x - \mu)/\sigma]^{-\alpha}$. The q-exponentials come from taking $\mu = 0$ and $\alpha = \theta$; the ordinary Pareto distribution is recovered by taking $\sigma = x_0$ and $\mu = \sigma$. According to Ref. [10, pp. 13–14, 208–210], the type II generalized Pareto was introduced in Refs. [11, 12, 13], and the latter two also derived the MLE.[1] The calculations below are a special case of their results, except for the treatment of censoring, which may be new.

---

*Electronic address: cshalizi@cmu.edu

[1] Arnold [10, p. 48] observes that a mixture of exponentials can

b. Derivation of the MLE for q-Exponentials Under the q-exponential model with parameters θ, σ, the log-probability density of a sequence of independent, identically-distributed samples $X_1 = x_1$, $X_2 = x_2$, $\ldots X_n = x_n$, for short $X_1^n = x_1^n$, is

$$\log p_{\theta,\sigma}(x_1^n) = -n \log \sigma + n \log \theta \tag{5}$$

$$-(\theta + 1) \sum_{i=1}^{n} \log 1 + x_i/\sigma$$

$$\equiv \ell(\theta, \sigma) , \tag{6}$$

the log-likelihood of the parameter combination θ, σ.

To find the MLEs, take the first derivatives of the log-likelihood with respect to the parameters and set them equal to zero. First, the shape parameter θ:

$$\frac{\partial \ell}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log 1 + x_i/\sigma \tag{7}$$

$$\hat{\theta} = n \left[\sum_{i=1}^{n} \log 1 + x_i/\sigma \right]^{-1} \tag{8}$$

Similarly for the scale parameter σ:

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\theta + 1}{\sigma^2} \sum_{i=1}^{n} \frac{x_i}{1 + x_i/\sigma} \tag{9}$$

$$\hat{\sigma} = \frac{\theta + 1}{n} \sum_{i=1}^{n} \frac{x_i}{1 + x_i/\hat{\sigma}} \tag{10}$$

Eqs. 8 and 10 give the MLEs for θ and σ, respectively, if the other parameter is known. The former gives the value of $\hat{\theta}$ explicitly[2], while the latter does so implicitly, through the solution of an equation. Implicitly-defined MLEs like this occur in several generalizations of the exponential distribution, such as the ones known to physicists as "stretched exponentials" and to statisticians as "Weibull distributions" (after the physicist who introduced them) [17, ch. 20]. The lack of a closed form is only a small annoyance, since such equations can generally be rapidly solved numerically, to a precision much smaller than the uncertainty inherent in the data.

If neither θ nor σ is known (i.e., neither q nor κ), then the simultaneous solution of Eqs. 8 and 10 gives the joint maximum likelihood estimator. Substituting the former equation into the latter gives a single equation in $\hat{\sigma}$ and

---

produce a type II generalized Pareto. If the distribution of $X - \mu$, given Z, is an exponential with mean σ/Z, and Z has a $\Gamma(\alpha, 1)$ distribution, then X has a type II generalized Pareto distribution with parameters μ, σ and α. He assigns priority for this result to [11]. It would appear to be equivalent to C. Beck's "superstatistics" approach to Tsallis statistics (reviewed in [14]).

[2] Cf. the well-known MLE for the scaling exponent in a Pareto distribution [10, 15, 16], $\hat{\alpha} = n/\left[\sum_{i=1}^{n} \log x/x_0\right]$.

the data:

$$\hat{\sigma} = \frac{1}{n} \left(1 + n \left[\sum_{i=1}^{n} \log 1 + x_i/\hat{\sigma} \right]^{-1} \right) \sum_{i=1}^{n} \frac{x_i}{1 + x_i/\hat{\sigma}} \tag{11}$$

This does not seem to simplify, but, again, can be solved numerically. (Eq. 11 is transcendental, whereas Eq. 10 is rational, but no worse than the equation for the MLE of the Weibull distribution, which also contains a sum of logarithms, etc.) Substituting the solution into Eq. 8 gives $\hat{\theta}$, and then Eq. 2 give $\hat{q}, \hat{\kappa}$.

c. Accuracy and Precision of the MLE An estimator $\hat{\psi}(X_1^n)$ of a parameter ψ of a statistical distribution is *consistent* when $\hat{\psi}$ converges in probability to ψ, i.e., for any $\epsilon > 0$ and any $\delta > 0$, for sufficiently large n, $P\left(\left\| \hat{\psi}(X_1^n) - \psi \right\| \geq \epsilon \right) \leq \delta$. In other words, a consistent estimator is "probably $(1-\delta)$ approximately (ϵ) correct", for arbitrarily small δ and ϵ. Under quite general conditions, met here, maximum likelihood estimators are consistent [18].

Consistency alone is not enough to calculate standard errors or confidence regions. However, under conditions only mildly more restrictive than those needed for consistency, MLEs are *asymptotically normal* and *unbiased*. That is, $\hat{\psi}(X_1^n) - \psi$ has, for large n, a multidimensional Gaussian distribution with mean zero and covariance matrix $(1/n)I^{-1}(\psi)$, where $I(\psi)$ is the *Fisher information matrix*,

$$I_{ij}(\psi) \equiv - \int \frac{\partial^2 \log p_\psi(x)}{\partial \psi_i \partial \psi_j} p_\psi(x) dx \tag{12}$$

By the famous Cramér-Rao inequality [19], any consistent unbiased estimator has a covariance at least equal to $I^{-1}(\psi)$; the MLE is *asymptotically efficient* because it attains this bound. Since the true value of ψ is unknown, $I(\psi)$ cannot give us standard errors or confidence regions, but $I(\hat{\psi})$ is a consistent estimator of $I(\psi)$, and can be used for those purposes. Another consistent estimator of the Fisher information is the *observed information matrix*, $J_{ij}(\psi) \equiv -n^{-1}\partial^2\ell(\psi)/\partial\psi_i\partial\psi_j$, and $J(\hat{\psi})$ also gives asymptotically-correct error estimates. Ref. [20] treats these standard results in detail.

For q-exponential distributions, it is easy to verify that the standard conditions for the asymptotic normality of the MLE hold. In the θ, σ parameterization, simple but lengthy calculus yields

$$I(\theta, \sigma) = \begin{bmatrix} \frac{1}{\theta^2} & -\frac{1}{(\theta+1)\sigma} \\ -\frac{1}{(\theta+1)\sigma} & \frac{\theta}{\sigma^2(\theta+2)} \end{bmatrix} \tag{13}$$

Either $I(\hat{\theta}, \hat{\sigma})$ or the observed information matrix could be used to find standard errors and Gaussian confidence regions. Propagation of errors can then carry these to estimates on q and κ.

For small samples, asymptotic approximations should be avoided in favor of *parametric bootstrapping* [21, sec.

9.11]. Having obtained an estimate $\hat{\psi} = \hat{\psi}(x_1^n)$, make up a "bootstrap" sample of random numbers $Y_1, Y_2, ...Y_n$ with the density $p_{\hat{\psi}}$, and calculate $\hat{\psi}(Y_1^n)$. The distribution of $\hat{\psi}(Y_1^n) - \hat{\psi}$ is approximately the same as that of $\hat{\psi}(X_1^n) - \psi$, so by taking many bootstrap samples one can estimate standard errors and confidence regions, without making Gaussian approximations. (For more on bootstrapping, see, e.g., [21, ch. 8].) The code mentioned above finds bootstrapped biases, standard errors and confidence intervals.

d. Censored Data In many applications, only measurements exceeding some known lower threshold x_0 are available, i.e., only values of $X \geq x_0$ become data. Parameters estimation from such *left-censored* data must take account of the threshold. Specifically, rather than maximizing the unconditional likelihood, $\ell(\theta, \sigma)$, one should maximize the likelihood conditional on being in the right tail, $\ell_C(\theta, \sigma, x_0)$. It is easily shown that the censored density is 0 when $x < x_0$, and otherwise

$$p_{\theta, \sigma, x_0}(x) = (1 + x_0/\sigma)^{\theta} p_{\theta, \sigma}(x) \tag{14}$$

$p_{\theta, \sigma}(x)$ being given by Eq. 4. The censored likelihood thus equals $\ell(\theta, \sigma)$ plus a term involving only θ, σ and x_0:

$$\ell_C(\theta, \sigma, x_0) = \ell(\theta, \sigma) + n\theta \log 1 + x_0/\sigma \tag{15}$$

The likelihood estimating equations become

$$\hat{\theta}_C = n \left[\sum_{i=1}^{n} \log \frac{1 + x_i/\sigma}{1 + x_0/\sigma} \right]^{-1} \tag{16}$$

$$\hat{\sigma}_C = -\theta \frac{x_0}{1 + x_0/\hat{\sigma}_C} + \frac{\theta + 1}{n} \sum_{i=1}^{n} \frac{x_i}{1 + x_i/\hat{\sigma}_C} \tag{17}$$

Eqs. 16 and 17 reduce to Eqs. 8 and 10 when $x_0 = 0$ (no censoring), and can be solved in the same way. The MLE remains consistent, and asymptotically normal and efficient. The Fisher information matrix, after an even longer calculation, ends up being $I(\theta, \sigma + x_0)$; explicitly,

$$I_C(\theta, \sigma, x_0) = \begin{bmatrix} \frac{1}{\theta^2} & -\frac{1}{(\theta+1)(\sigma+x_0)} \\ -\frac{1}{(\theta+1)(\sigma+x_0)} & \frac{\theta}{(\sigma+x_0)^2(\theta+2)} \end{bmatrix} \tag{18}$$

Bootstrapping, however, is even more strongly recommended than with uncensored data. Simulated values should be drawn from the tail only.

e. Comparison to Curve-Fitting Hitherto, attempts to estimate the parameters of q-exponential distributions have been based on curve-fitting. (Ref. [22] is an unusually careful example.) Taking the log of both sides of Eq. 3,

$$\log P_{\theta, \sigma}(X \geq x) = -\theta \log (1 + x/\sigma) \tag{19}$$

Write $S_n(x)$ for the empirical distribution function, i.e., the fraction of points in $x_1, x_2, \ldots x_n$ which are $\geq x$. Then we expect that, at least for large sample sizes n,

$$\log S_n(x_i) \approx -\theta \log (1 + x_i/\sigma) \tag{20}$$

Estimates of q Values
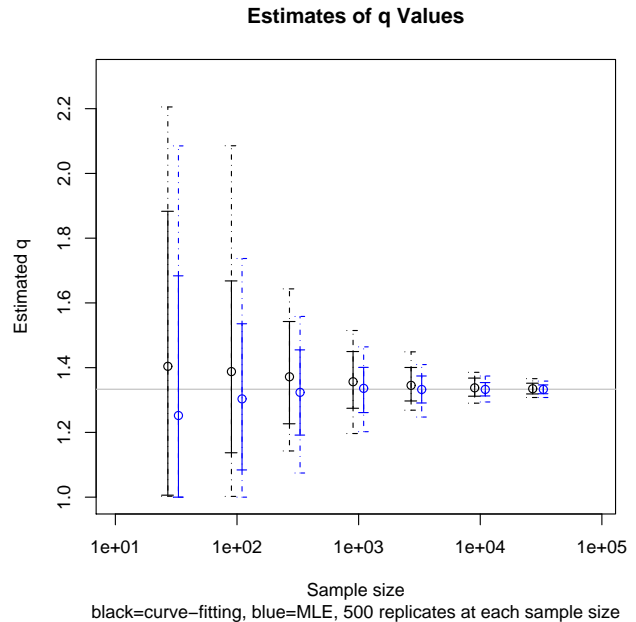
FIG. 1: (Color online) Comparison of estimates of q using the MLE and curve-fitting. All data generated using $q = 4/3, \kappa = 200/3$ ($\theta = 3, \sigma = 200$), with varying sample sizes, and 500 independent replications are each sample size. Curve-fitting estimates are plotted in black, displaced slightly to the left, and MLEs in blue, displaced to the right. Solid bars show the 5th and 95th percentiles of sample estimates, circles the median estimate, and dashed lines the sample extrema. Note that the MLE is always less biased and more precise.

for all x_i. A least-squares approach to estimating the parameters minimizes the squared difference between the two sides of Eq. 20, summed over all x_i, i.e.,

$$(\tilde{\theta}, \tilde{\sigma}) \equiv \operatorname*{argmin}_{\theta, \sigma} \sum_{i=1}^{n} \left(\log S_n(x_i) + \theta \log \left(1 + \frac{x_i}{\sigma} \right) \right)^2 \tag{21}$$

It is possible to show that this estimator is consistent.

The analogous procedure for Pareto distributions was the one originally used by Pareto in the 1890s [10], and still widespread in physics. For Pareto distributions, however, statisticians have known since the 1950s that such estimation-by-regression is much more biased, and much less precise, than the maximum likelihood estimator [10, 15]. (In particular, the standard errors are much larger than blind use of the ordinary regression formulas suggest.) The same is true of the least-squares estimate of q-exponentials (Fig. 1). Fitting curves to binned estimates of the probability density, rather than to the cumulative distribution, is even less accurate. Neither approach should be used.

f. Validation All the claims of consistency, efficiency, etc., made above assume that the data really do come from a q-exponential distribution. In statistical terminology, the assumption is that the q-exponential model is correctly *specified*, as opposed to being *mis-specified*.

In applications to empirical data, it is crucial to check this assumption. Rigorous mis-specification tests are too complicated to go into here [23, 24], but some remarks are in order.

The most common test of specification in the literature on Tsallis statistics is to look at the fraction, R^2, of the variance in $\log S_n$ accounted for by the fitted distributional curve. Unfortunately, this popularity is not based on any reliability; it is easy to construct examples where $1 + x/\sigma$ has, say, a log-normal distribution, but R^2 is always close to 1. Rather than looking at R^2, one should either test q-exponentials against alternative distributions such as the Pareto, the log-normal, etc., or do general goodness-of-fit tests, adjusting for the way parameters are estimated from the data [21, ch. 10]. The latter must be interpreted with caution: failing a goodness-of-fit test provides strong evidence against a model, but passing one may give only very weak evidence in its favor, depending on the severity of the test [25].

Two heuristic checks for mis-specification deserve mention. One compares the parametric bootstrap, described above, with a *non-parametric bootstrap*, in which the values $Y_1, \ldots Y_n$ come from resampling the data $x_1, \ldots x_n$ with replacement, not from the fitted distribution. If parametric and non-parametric bootstrap estimates of bias, standard error, etc., differ substantially, this is a sign that the model is mis-specified. Similarly, if the expected Fisher information at the MLE, $I(\hat{\theta}, \hat{\sigma})$ is very different from the observed information, $J(\hat{\theta}, \hat{\sigma})$, this again suggests the statistical model poorly describes the data-generating process. The comparison of information matrices can be turned into a formal test for mis-specification [23].

Conclusion   Tsallis q-exponentials are legitimate possible models of heavy-tailed data. Under other names, they have been so used in operations research and statistics for half a century, without any entropic origin story. To model data with q-exponentials, their parameters must be estimated accurately. The estimators currently used by physicists are inferior to the MLE, which is asymptotically efficient. If physicists want to describe data with q-exponentials, they should stop fitting curves and start maximizing likelihoods. Whether using Tsallis statistics is a good idea in the first place is another matter, beyond the scope of this note.

[1] C. Tsallis, Journal of Statistical Physics **52** 479 (1988).

[2] E. T. Jaynes, *Essays on Probability, Statistics, and Statistical Physics* (Reidel, London, 1983).

[3] B. R. La Cour and W. C. Schieve, Physical Review E **62** 7494 (2000), cond-mat/0009216.

[4] D. H. Zanette and M. M. Montemurro, Physics Letters A **316**, 184 (2003), cond-mat/0212327.

[5] D. H. Zanette and M. M. Montemurro, Physics Letters A **324**, 383 (2004), cond-mat/0305070.

[6] F. Bouchet, T. Dauxois, and S. Ruffo, Europhysics News **37**, 9 (2006), cond-mat/0605445.

[7] B. H. Lavenda and J. Dunning-Davies, Journal of Applied Sciences **5** 315–322 (2005), physics/0310117.

[8] M. Nauenberg, Physical Review E **67**, 036114 (2003), cond-mat/0210561.

[9] D. R. White, N. Kejzar, and L. Tambayong, in *Globalization as Evolutionary Process: Modeling, Simulating, and Forecasting Global Change*, edited by G. Modelski, T. Devezas, and W. Thompson (Routledge, London, 2007).

[10] B. C. Arnold, *Pareto Distributions* (International Cooperative Publishing House, Fairland, Maryland, 1983).

[11] B. A. Maguire, E. S. Pearson, and A. H. A. Wynn, Biometrika **39**, 168 (1952), JSTOR.

[12] H. Silcock, Journal of the Royal Statistical Society A **117** (1954), JSTOR.

[13] C. M. Harris, Operations Research **16**, 307 (1968), JSTOR.

[14] C. Beck, in *Complexity, Metastability and Nonextensiviity*, edited by C. Beck, G. Benedek, A. Rapisarda, and C. Tsallis (World Scientific, Singapore, 2005), cond-mat/0502306.

[15] A. N. M. Muniruzzaman, Bulletin of the Calcutta Statistical Association **7**, 115 (1957).

[16] M. E. J. Newman, Contemporary Physics **46**, 323 (2005), cond-mat/0412004.

[17] N. L. Johnson and S. Kotz, *Continuous Univariate Distributions — 1* (Wiley, New York, 1970).

[18] E. J. G. Pitman, *Some Basic Theory for Statistical Inference* (Chapman and Hall, London, 1979).

[19] H. Cramér, *Mathematical Methods of Statistics* (Almqvist and Wiksells, Uppsala, 1945).

[20] O. E. Barndorff-Nielsen and D. R. Cox, *Inference and Asymptotics* (Chapman and Hall, London, 1995).

[21] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer-Verlag, Berlin, 2003).

[22] D. R. White, N. Kejzar, C. Tsallis, D. Farmer, and S. White, Physical Review E **73** (2006), cond-mat/0508028.

[23] H. White, *Estimation, Inference and Specification Analysis* (Cambridge University Press, Cambridge, 1994).

[24] A. Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data* (Cambridge University Press, Cambridge, 1999).

[25] D. G. Mayo and D. R. Cox, in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo (Institute of Mathematical Statistics, Bethesda, Maryland, 2006), pp. 77–97, math.ST/0610846.