Inference, Models and Simulation for Complex Systems
CSCI 7000-003, Fall 2010
Prof. Aaron Clauset
Problem Set 1, due 9/15

This is a long problem set and thus I encourage you to strike up conversations with your fellow students about the trickier parts.

1. **Stretched exponential distributions**

   Recall that an exponential distribution is defined as $\Pr(x) \propto \mathrm{e}^{-\lambda x}$, which corresponds to a dynamical process in which the probability q of some event is independent and identically distributed (iid). However, if q depends in some way on this history of the system, for example, decreasing with time, the distribution of lifetimes will deviate from an exponential. One such generalization is called a *stretched exponential* distribution and has the form $\Pr(x) = Cx^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$, where $C = \beta\lambda$, $x \geq 0$, and $\lambda, \beta > 0$.

   (a) Derive and simplify the log-likelihood function $\ln \mathcal{L}(\{x_i\} \,|\, \lambda, \beta)$ for the stretched exponential distribution.

   (b) Using our result from (1a), derive a maximum likelihood estimator (MLE) for λ and a transcendental equation the represents the MLE for β.

   (c) Many numerical applications call for synthetic data with a particular distribution. For instance, we might want to test a procedure on synthetic data with known structure before running it on empirical data with unknown structure.

   In simple situations, this can be accomplished by first generating n random numbers distributed uniformly on the unit interval $[0, 1)$ (e.g., using a good pseudo-random number generator like the Mersenne twister) and then transforming those numbers into the target distribution. (We'll look at more complicated data-generation procedures later in the class).

   For many distributions of interest, the transformation step can be done like so: recall that the cdf of the target distribution is a function that maps the distribution's range, say $[1, \infty)$, onto the unit interval. Thus, the *inverse* cdf is a function that can transform a uniform random deviate into a deviate from our target distribution. This is called the *transformation method*. (That being said, some distributions are not susceptible to the transformation method and other methods are necessary to handle these.)

Analytically derive a generator that produces stretched exponential deviates. Note that the generator will take a single uniform deviate r, along with the parameters λ and β, and produce a single deviate of the target distribution x.

(d) Choose a few values for λ and β, generate $n = 1000$ deviates for each setting, and plot the resulting empirical distribution functions (edfs) together on a single log-log graph as ccdfs, that is, as $\Pr(X \geq x)$. Label your axes and provide a legend.

(e) Set up and run numerical experiments to demonstrate that the MLEs we derived in (1b) for β and λ are *asymptotically consistent* when applied to data generated using the method we derived in (1c). That is, show that the estimated parameter $\hat{\lambda}$, for fixed β, converges on the true value λ as $n \to \infty$. Repeat this for β, with fixed λ.

Hint: for each estimator, produce a log-log plot showing the mean absolute error (MAE) $\langle|\hat{\theta} - \theta|\rangle$ as a function of n, for logarithmically spaced choices of n. For each particular value of n, average the error over many independent samples to get a clear trend. Discuss the apparent functional form of the MAE.

(f) A *hazard function* $h(x)$ describes the relationship between the probability q that a "death" event happens and the lifetime of the corresponding object. It's defined as the fraction of objects with lifetime x or greater that have lifetime exactly x; mathematically, we say $h(x) = \Pr(X = x)/\Pr(X \geq x)$. This kind of analysis is used to estimate component failure rates, for example, in computers and nuclear bombs, but it has also been applied to the death rates of terrorist groups and biological organisms, and can be used to model a variety of growth processes.

For a stretched exponential, when $\beta > 1$ the tail decays more quickly than an exponential. This implies that $h(x)$ is an increasing function and that the probability of death q increases with the value of x. Similarly, if $\beta < 1$ the tail decays more slowly than exponential, which implies that $h(x)$ is a decreasing function and the probability of death decreases with size.

Derive an analytic expression for $h(x)$ for the stretched exponential distribution.

(g) Given empirical data for some quantity, we can calculate the *empirical hazard function* directly, once we've tabulated the empirical pdf and cdf.

Set up and run a numerical experiment to verify that the empirical hazard function for synthetic data from a stretched exponential [see (1c)] matches the analytic function we derived in (1f). Choose any values of λ and β, but choose a value of n large enough to get a good experimental result. Hint: averaging the empirical hazard function across independent samples will improve your results.

2. **Power-law distributions**

As we saw in class, a power-law distribution is a special kind of distribution because some of their moments don't exist. Recall that a power-law distribution follows the form $\Pr(x) = Cx^{-\alpha}$, for $\alpha > 1$ and $x \geq x_{\min} > 0$.

(a) Derive a random deviate generator that produces numbers distributed according to a power-law distribution with range $[x_{\min}, \infty)$.

(b) Choose a few values for α, generate $n = 10000$ deviates for each setting, and plot the resulting edfs together on a single log-log graph as ccdfs.

(c) Set up and run a numerical experiment to demonstrate that the MLE we derived in class for α is asymptotically consistent. [Hint: see (1e)].

3. **Data analysis**

In this section, we will analyze two real-world data sets using the tools we constructed above. We'll also learn something about comparing statistical models and and something about the difficulty of making clear inferences with real data.

(a) Choose one data set from among data sets 1A, 1B, 1C and 1D on the course webpage and fit (using your MLEs) both the power-law distribution and the stretched exponential distribution to it.

(b) Make a log-log plot showing the ccdf of your chosen data set, along with the two maximum likelihood fitted models. Comment on the visual quality of the fits. Give the parameter estimates.

(c) A *likelihood ratio test* (LRT) is a powerful way to decide whether model A or model B is a better fit to some empirical data. Note that a LRT is only an exercise in model comparison as it can't tell us if either or both of A and B is itself a plausible explanation of our data. A LRT works by computing the likelihood ratio statistic

$$\mathcal{R} = \ln\left(\frac{\mathcal{L}_A(\{x_i\} \mid \hat{\theta}_A)}{\mathcal{L}_B(\{x_i\} \mid \hat{\theta}_B)}\right)$$

which is defined as the logarithm of the ratio of the likelihoods, evaluated at their respective maximum likelihood parameters, of the empirical data under models A and B. The sign of $\mathcal{R}$ tells us whether model A or B is favored. If the sign is positive, the model represented by the numerator is a better fit; if the sign is negative, the model represented by the denominator is a better fit.

Note, however, that because the data $\{x_i\}$ are considered a random variable, $\mathcal{R}$ is itself a random variable and thus the particular sign we observe could be a chance occurrence, implying that we cannot say which model is a better fit. Before we can interpret the sign of $\mathcal{R}$, we need to try to eliminate this possibility by performing a simple hypothesis test for $\mathcal{R} = 0$. Following a procedure described by Vuong (*Econometrica* **57**, 307–333 [1989]), we can do this by computing the normalized log-likelihood ratio $n^{-1/2}\mathcal{R}/\sigma$, where σ is the estimated standard deviation of $\mathcal{R}$ and is defined as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(\ell_i^A - \ell_i^B) - (\overline{\ell}^A - \overline{\ell}^B)\right]^2$$

where

$$\overline{\ell}^A = \frac{1}{n}\sum_{i=1}^{n}\ell_i^A \qquad\qquad \overline{\ell}^B = \frac{1}{n}\sum_{i=1}^{n}\ell_i^B$$

where ℓ_i^A is the likelihood of the ith observation under model A.

For non-nested models like the power-law distribution and the stretched exponential distribution, the significance of Vuong's test statistic can be measured by computing a standard p-value, using the analytic expression

$$p = \operatorname{erfc}\Bigl(|\mathcal{R}| \Big/ \sigma\sqrt{2n}\Bigr) \quad .$$

If $p < 0.1$, we reject the null hypothesis that the sign of $\mathcal{R}$ is ambiguous, and proceed with interpreting the sign of $\mathcal{R}$.

For one of the data sets you chose, construct a LRT to decide whether a power-law or stretched exponential distribution is a better fit.

(d) Download data set 1E and plot it. Write a paragraph about what makes this data set difficult to model as either a stretched exponential or power law. Be specific, both about the structure of the data set itself and about the structure and assumptions of the models.

(e) (**optional**) Download data from the web, e.g., from the US Census 2000 website, on the age distribution of individuals in the United States. Plot the data. Estimate the empirical hazard function from the data. Discuss whether the data is better modeled using an exponential or stretched exponential distribution, and comment on the implication of that insight on the risk of dying at different ages. Finally, comment on the implicit assumptions in this kind of analysis of this kind of data. (Hint: think about iid random variables.)