Inference, Models and Simulation for Complex Systems
CSCI 7000-001, Fall 2011
Prof. Aaron Clauset
Problem Set 1, due 9/13

This is a long problem set. I encourage you to converse with your fellow students about the trickier parts. But remember that you must write up your solutions alone and don't forget to show your work!

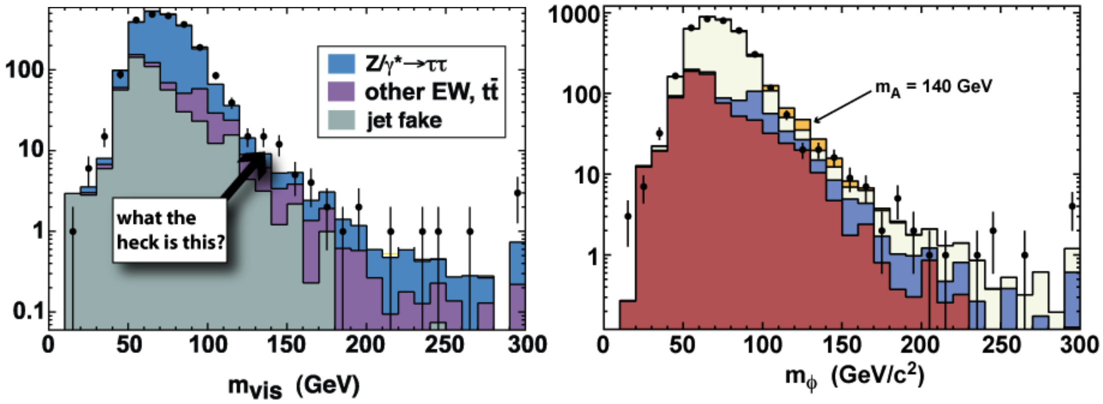

Figure 1: `http://tinyurl.com/3t8gx5x`

1. **The Sciences of the Artificial** (15 points)

   Read the "Preface to the 2nd Edition" and Chapter 1. Write no more than two paragraphs that clearly and concisely explain (i) what Simon means by "inner" and "outer" environments and (ii) why this conceptualization is useful for a science of complex systems.

2. **Stretched exponential distributions** (45 points)

   Recall that an exponential distribution is defined as $\Pr(x) \propto \mathrm{e}^{-\lambda x}$ with $x \geq 0$, which corresponds to a dynamical (Poisson) 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, i.e., if $\partial q / \partial t \neq 0$, generating process is no longer a Poisson process and the distribution of waiting times will deviate from the exponential form.

Clearly, deviations from the exponential form could be arbitrarily complicated. However, there are a few simple generalizations; one is called a *stretched exponential* (SE) or Weibull distribution and has the form $\Pr(x) = Cx^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$, where C is the normalization constant, $x \geq x_{\min} \geq 0$, and $\lambda, \beta > 0$.

We'll use the SE distribution as a tool to explore several general ideas about probabilistic models, synthetic data, and model fitting. That is, the SE model is simple and tractable, but the ideas can easily be adapted to much more complicated probabilistic models.

(a) (5pts) Derive and simplify the log-likelihood function $\ln \mathcal{L}(\{x_i\} \mid \lambda, \beta)$ for the stretched exponential distribution on the interval $[x_{\min}, \infty)$. Note that you'll need to derive the normalization constant C first.

(b) (10pts) Using your result from (2a), derive a maximum likelihood estimator (MLE) for λ and a transcendental equation the represents the MLE for β. Report simplified expressions.

Hint: Look up what a transcendental equation is and your answers should be in simplified form. Note that each of these equations will contain $x_{\min}$.

(c) (10pts) Many numerical applications call for synthetic data with a particular distribution. For instance, it's a good idea to verify the accuracy of your procedure using synthetic data with known structure before you apply it to 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 pseudorandom 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 $[x_{\min}, \infty)$, onto the unit interval $[0, 1]$. Thus, the *inverse* cdf is a function that does the reverse: it maps values from the unit interval onto values in our distribution's range. This is called the *transformation method*. (That being said, some distributions are not susceptible to the transformation method and other methods are necessary.)

Analytically derive a generator that produces stretched exponential deviates on the interval $[x_{\min}, \infty)$. Note that the generator will take a single uniform deviate r, along with the parameters $x_{\min}$, λ and β, and produce a single deviate of the target distribution x.

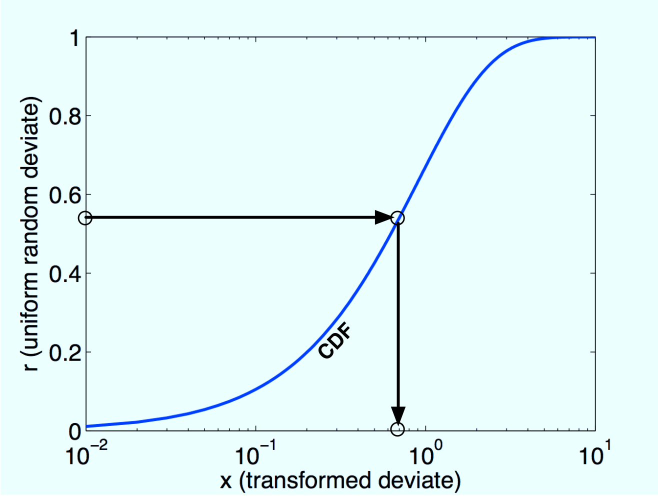

Figure 2: An illustration of the transformation method.

(d) (5pts) Let $x_{\min} = 1$. 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 $\Pr(X \geq x)$. No credit if you don't label your axes and provide a legend.

(e) (10pts) Set up and run numerical experiments to demonstrate that the MLEs we derived in (2b) for β and λ are *asymptotically consistent* when applied to data generated using the method we derived in (2c). 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. No credit if you don't label your axes and provide a legend.

Hint: for each estimator, produce one log-log plot showing the mean squared error (MSE) $\langle(\hat{\theta} - \theta)^2\rangle$ (where θ is the true and $\hat{\theta}$ is the estimated parameter) as a function of n, for a dozen logarithmically spaced choices of n over several orders of magnitude. For each particular value of n, average the error over a few hundred independent samples to get a clear trend.

(**10pts extra credit**) Derive and plot alongside your simulated results an analytic expression for the MSE as a function of n, $\hat{\beta}$ and $\hat{\lambda}$. *Hint*: *Fisher information*.

(f) (5pts) A *hazard function* $h(x)$ describes the relationship between the probability q that a "death" or failure event happens and the lifetime of the corresponding object. It's defined as the fraction of objects with lifetime x or greater that

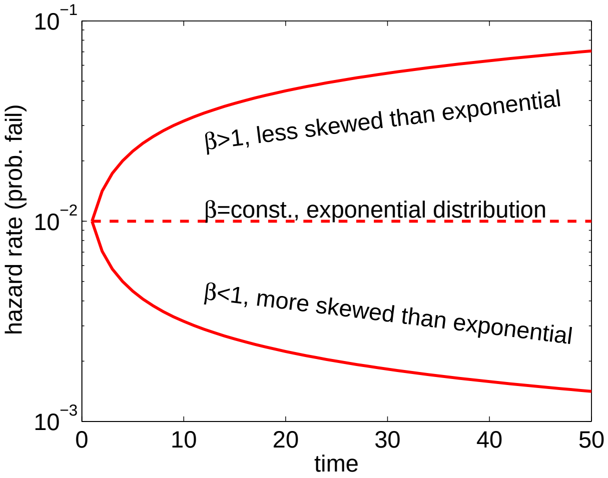

Figure 3: An illustration of hazard functions and the types of distributions they produce.

have lifetime exactly x; mathematically, we say $h(x) = \Pr(X)/\Pr(X \geq x) = \mathrm{pdf}(x)/(1 - \mathrm{cdf}(x))$. 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 inhomogeneous processes.

For the SE, 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. (See Figure 3.)

Derive an analytic form for $h(x)$ for the stretched exponential distribution. Simplify the expression and report your result.

(**10pts extra credit**) Using your result for (2f), fix $x_{\min} = 0$ and choose values for λ and β. Then, design and conduct a numerical experiment in which the probability of "death" q varies according to your hazard rate expression. Generate $n = 1000$ deviates using the stochastic process and fit the stretched exponential distribution to your data. Report the values of λ,β and of $\hat{\lambda}$,$\hat{\beta}$. Produce a plot that shows the simulated distribution of lifetimes and the fitted stretched exponential distribution. No credit if you don't label your axes and provide a legend.

3. **Power-law distributions** (20 points)

   As we saw in class, a power-law (PL) 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) (5pts) Derive via the transformation method an expression for generating random deviates from a PL distribution on the interval $[x_{\min}, \infty)$. Report the simplified expression.

   (b) (5pts) Choose a few values $1 < \alpha \leq 4$, generate $n = 10000$ deviates for each setting, and plot the resulting edfs together on a single log-log graph as ccdfs. No credit if you don't label your axes and provide a legend.

   (c) (10pts) Set up and run a numerical experiment to demonstrate that the MLE we derived in class for α is asymptotically consistent. No credit if you don't label your axes and provide a legend.

4. **Data analysis** (20 points)

   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) (5pts) Choose one data set from among data sets 1A, 1B, 1C and 1D on the course webpage. For both models, set $x_{\min} = \min_i x_i$ and fit to these data (using your MLEs) both the PL and SE distributions. Report your parameter estimates.

   (b) (5pts) 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. No credit if you don't label your axes and provide a legend.

   (c) (10pts) 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 PL and SE distributions ("non-nested" means that one is not a subset of the other), the significance of Vuong's test statistic can be measured by computing a standard p-value, using the analytic expression

$$p = \operatorname{erfc}\left(|\mathcal{R}| \middle/ \sigma\sqrt{2n}\right) \quad ,$$

where erfc(.) is the "error function," which is simply the cdf for the Normal distribution. There is no analytic expression for erfc(.), but most numerical or scientific programming environments include a function that approximates it very closely.

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 the data set you chose in part (4a), construct a LRT to decide whether a power-law or stretched exponential distribution is a better fit.