10/15: Estimation

Reminder

- In class exam next Tuesday, Oct. 22
- Open book/notes

Recall: Maximum Likelihood

The most commonly used method for point estimation. Given a family $\mathcal{F} = \{f(x \mid \theta)\}$ and data $X_1, X_2, \ldots, X_n$, the *likelihood function* is defined as

$$\mathcal{L}_n(\theta) = \prod_i f(X_i \mid \theta)$$

and the *log-likelihood function* is given by

$$\begin{aligned} \ell_n(\theta) &= \log \mathcal{L}_n(\theta) \\ &= \sum_i \log f(X_i \mid \theta) \end{aligned}$$

Maximum Likelihood

The *maximum likelihood estimator* is

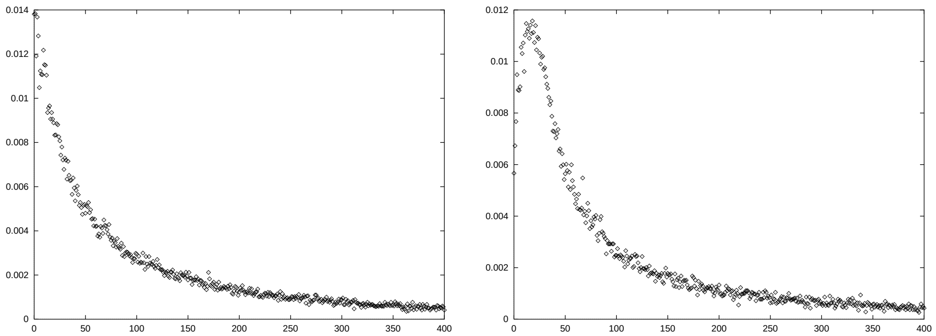
$$\hat{\theta} = \arg\max_{\Theta} \ell_n(\theta)$$

Maximum Likelihood

Examples (on the board): binomial, multinomial (last time) normal, and uniform.

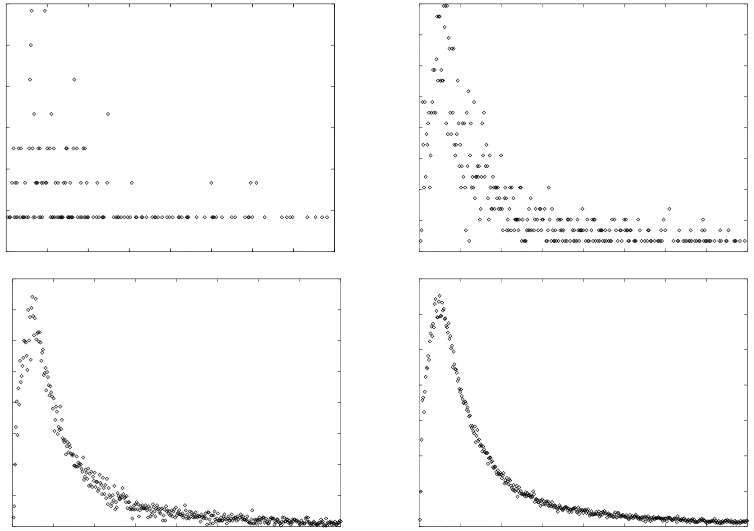
Today we'll do a fairly complicated and elaborate example.

Recall Example – Word Spacings



For a given distance k on the x-axis, the value on the y-axis is the empirical probability that two "trigger" words within the group are separated by exactly $k+2$ words, conditional on the event that they co-occur within a 400 word window.

Word Spacings



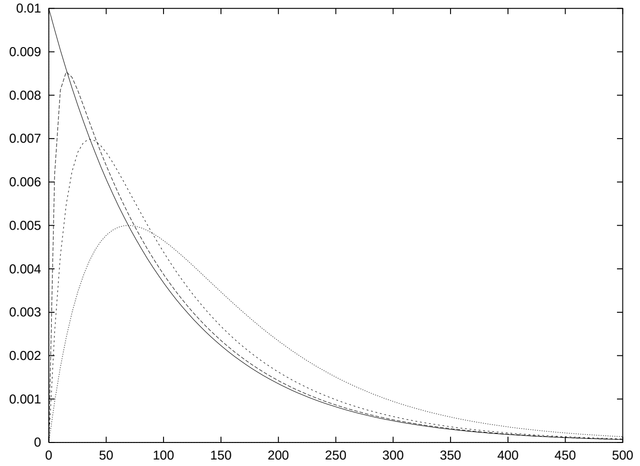
Word Spacings

Three characteristics:

1. Mode away from zero
2. Exponential fall off
3. Long tail

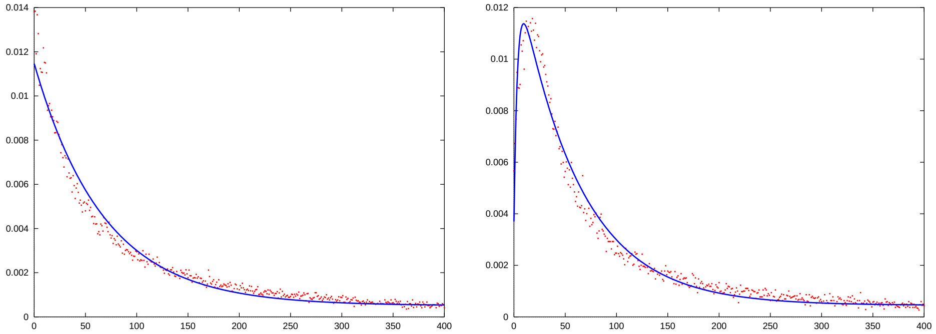
What type of distribution will fit this data?

Word Spacings



A sequence of two-stage exponential models $p_{\mu_1,\mu_2}(x)$ with $\mu_1 = 0.01,\ 0.02,\ 0.06,\ 0.2,\ \infty$ and $\mu_2 = 0.01$.

Word Spacings



The empirical distance distributions fit to a three-parameter mixture model $p_{\mu_1,\mu_2,\alpha}$.

How is this done?

Exponential

First property: decay of the probability of a word t as the distance k from the most recent occurrence of its mate s increases.

Suggestive of single-parameter exponential family

$$p_\mu(x) = \mu\, e^{-\mu x}\,.$$

Exponential

We'll begin by showing the continuous analogues of the discrete formulas we actually use, since they are simpler in appearance.

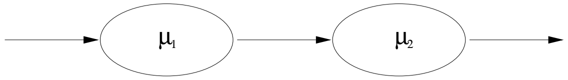
This family is uniquely characterized by the *memoryless property* that the probability of waiting an additional length of time Δt is independent of the time elapsed so far.

Distribution p_μ has mean $1/\mu$ and variance $1/\mu^2$.

Convolution of Exponentials

Second characteristic: bump between 0 and 25 words when same word is repeated.

Behavior appears when two exponential distributions are arranged in serial



Convolution of Exponentials

The time it takes to travel through two service facilities arranged in serial, where the first provides exponential service with rate μ_1 and the second provides exponential service with rate μ_2, is simply the convolution of the two exponentials:

$$\begin{aligned} f_{\mu_1,\mu_2}(x) &= \mu_1\mu_2 \int_0^x e^{-\mu_1 t} e^{-\mu_2(x-t)}\, dt \\ &= \frac{\mu_1\mu_2}{\mu_2-\mu_1}\left(e^{-\mu_1 x} - e^{-\mu_2 x}\right) \qquad \mu_1 \neq \mu_2\,. \end{aligned}$$

Convolution of Exponentials

Note: when $\mu_1 = \mu_2$ this results in the second order *Erlang distribution* (related to Gamma).

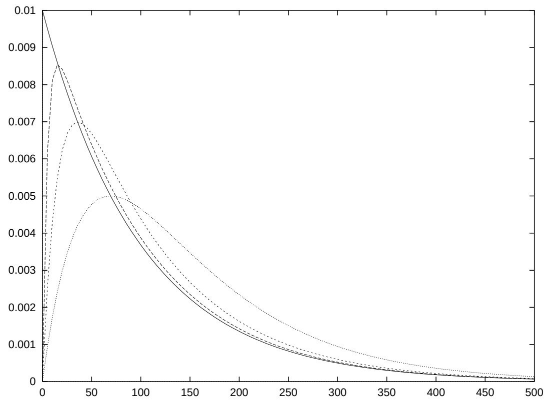
These models are useful in queueing/network theory because they relax the memoryless and Markov properties of the exponential while they maintain many of its simple mathematical properties.

Convolution of Exponentials

The mean and variance of the two-stage exponential f_{μ_1,μ_2} are $1/\mu_1 + 1/\mu_2$ and $1/\mu_1^2 + 1/\mu_2^2$ respectively.

As μ_1 (or, by symmetry, μ_2) gets large, the peak shifts towards zero and the distribution approaches the single-parameter exponential f_{μ_2} (by symmetry, f_{μ_1}).

Convolution of Exponentials



A sequence of two-stage exponential models $f_{\mu_1,\mu_2}(x)$ with $\mu_1 = 0.01,\ 0.02,\ 0.06,\ 0.2,\ \infty$ and $\mu_2 = 0.01$.

Interpretation

two-stage exponential is a good candidate for distance modeling because of its mathematical properties — also well-motivated for "linguistic" reasons.

The first queue in the two-stage model represents the stylistic and syntactic constraints that prevent a word from being repeated too soon.

After this waiting period, the distribution falls off exponentially, with the memoryless property.

Third Property

The third property of empirical data is the tendency of the curves to approach a constant, positive value for large distances.

While the exponential distribution quickly approaches zero, the empirical data settles down to a nonzero "steady-state" value.

Parametric Model

Together these three features suggest modeling distance with a three-parameter family of distributions:

$$f_{\mu_1,\mu_2,c}(x) = \gamma^{-1}(f_{\mu_1,\mu_2}(x) + c)$$

where $c > 0$ and γ is a normalizing constant.

Discrete Time

Rather than a continuous-time exponential, we use the discrete-time analogue

$$f_\mu(k) = (1 - e^{-\mu})\, e^{-\mu k}\,.$$

In this case, the two-stage model becomes the discrete-time convolution

$$f_{\mu_1,\mu_2}(k) = \sum_{t=0}^{k} f_{\mu_1}(t)\, f_{\mu_2}(k-t)\,.$$

Alternatives

There is another parametric family that is an excellent candidate for distance models, based on the first two features noted above.

Gamma distribution

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\,.$$

Mean α/β and variance α/β^2.

Note, for Bayesian analysis, this distribution is the "conjugate prior" for the exponential parameter μ

Estimating the Parameters

Using maximum likelihood: let $\tilde{p}(k)$ is the empirical probability that two words appear a distance of k words apart.

Objective function:

$$\ell(\theta) = \sum_{k \geq 0} \tilde{p}(k) \log f_\theta(k)\,.$$

Estimating the Parameters

First suppose that $\{f_\theta\}_{\theta\in\Theta}$ is the family of continuous one-stage exponential models $f_\mu(k) = \mu e^{-\mu k}$.

In this case the maximum likelihood problem is straightforward: the mean is the sufficient statistic for this exponential family. Maximum likelihood is

$$\mu = \frac{1}{\sum_{k\geq 0} k\, \tilde{p}(k)} = \frac{1}{E_{\tilde{p}}\,[k]}\,.$$

Discrete Case, Single Geometric

In the case where we instead use the discrete model $f_\mu(k) = (1 - e^{-\mu})\, e^{-\mu k}$, a little algebra shows that the maximum likelihood estimate is

$$\mu = \log\left(1 + \frac{1}{E_{\tilde{p}}[k]}\right).$$

Two-Queue Case

Now suppose that our parametric family $\{f_\theta\}_{\theta\in\Theta}$ is the collection of two-stage exponential models.

The log-likelihood becomes

$$\ell(\mu_1, \mu_2) = \sum_{k\geq 0} \tilde{p}(k) \log\left(\sum_{j=0}^{k} f_{\mu_1}(j)\, f_{\mu_2}(k-j)\right).$$

Here it is not obvious how to proceed! Idea?

Two-Queue Case

$$\ell(\mu_1, \mu_2) = \sum_{k\geq 0} \tilde{p}(k) \log\left(\sum_{j=0}^{k} f_{\mu_1}(j)\, f_{\mu_2}(k-j)\right).$$

The difficulty is that there is a sum inside the logarithm, and direct differentiation results in coupled equations for μ_1 and μ_2.

Two-Queue Case

Solution: view the convolving index j as a *hidden variable* and apply the *EM algorithm*.

Recall that the interpretation of j is the time used to pass through the first "queue"

This value is hidden given only the total time k required to pass through both queues.

Applying EM

The difference in log-likelihood for two parameter pairs (μ_1', μ_2') and (μ_1, μ_2) can be bounded from below.

Applying EM

$$
\begin{aligned}
\ell(\mu') - \ell(\mu) &= \sum_{k \geq 0} \tilde{p}(k) \log \left(\frac{p_{\mu_1', \mu_2'}(k)}{p_{\mu_1, \mu_2}(k)} \right) \\
&\geq \sum_{k \geq 0} \tilde{p}(k) \sum_{j=0}^{k} p_{\mu_1, \mu_2}(j \,|\, k) \log \left(\frac{p_{\mu_1', \mu_2'}(k, j)}{p_{\mu_1, \mu_2}(k, j)} \right) \\
&\equiv \mathcal{A}(\mu', \mu)
\end{aligned}
$$

where

$$f_{\mu_1, \mu_2}(k, j) = f_{\mu_1}(j)\, f_{\mu_2}(k - j)$$

and

$$f_{\mu_1, \mu_2}(j \,|\, k) = \frac{f_{\mu_1, \mu_2}(k, j)}{f_{\mu_1, \mu_2}(k)} .$$

Applying EM

Thus, the *auxiliary function* $\mathcal{A}$ can be written as

$$
\begin{aligned}
\mathcal{A}(\mu', \mu) \;=\; & \log\left(1 - e^{-\mu_1'}\right) + \log\left(1 - e^{-\mu_2'}\right) \\
& - \mu_1' \sum_{k \geq 0} \tilde{p}(k) \sum_{j=0}^{k} j\, p_{\mu_1, \mu_2}(j \,|\, k) \\
& - \mu_2' \sum_{k \geq 0} \tilde{p}(k) \sum_{j=0}^{k} (k - j)\, p_{\mu_1, \mu_2}(j \,|\, k) \\
& + constant(\mu) .
\end{aligned}
$$

Applying EM

Differentiating $\mathcal{A}(\mu', \mu)$ with respect to μ_i', we get the EM updates

$$\mu_1' = \log \left(1 + \frac{1}{\sum_{k \geq 0} \tilde{p}(k) \sum_{j=0}^{k} j\, p_{\mu_1, \mu_2}(j \,|\, k)} \right)$$

$$\mu_2' = \log \left(1 + \frac{1}{\sum_{k \geq 0} \tilde{p}(k) \sum_{j=0}^{k} (k - j)\, p_{\mu_1, \mu_2}(j \,|\, k)} \right)$$

Computation

Note: It appears that these updates require $O(N^2)$ operations if a window of distance N is used.

However, using tricks for the geometric series, such as $\sum_{k=0}^{\infty} kx^k = x/(1-x)^2$, we can write the expectation $\sum_{j=0}^{k} j\, p_{\mu_1,\mu_2}(j \mid k)$ in closed form.

Thus, the updates can be calculated in linear time $O(N)$.

Three Parameter Model

Finally, suppose that our parametric family $\{f_\theta\}_{\theta\in\Theta}$ is the three-parameter collection of two-stage exponential models together with an additive constant:

$$f_{\mu_1,\mu_2,c}(k) = \gamma^{-1}(f_{\mu_1,\mu_2}(k) + c)\,.$$

Here again, the maximum likelihood problem can be solved by introducing a hidden variable. In particular, by setting $\alpha = \frac{c}{c+N^{-1}}$ we can express this model as a *mixture* of a two-stage exponential and a uniform distribution:

$$f_{\mu_1,\mu_2,\alpha}(k) = (1-\alpha)\, f_{\mu_1,\mu_2}(k) + \alpha \left(\frac{1}{N}\right)\,.$$

Result from all of that Work