10-602: Statistical Approaches to Learning and Discovery

Problem Set 5 Solutions

---

1. This problem comes from Vardi, Shepp, and Kaufman, "A Statistical Model for Positron Emission Tomography," Journal of the American Statistical Association, Volume 80, pp.8–20, 1985.

   (a) If we start with a Poisson distribution $Po(\lambda_j)$ and filter the arrivals with probability p_{ij}, then the result is $Po(p_{ij}\lambda_j)$. Since the sum of independent Poissons is Poisson, $n_i \sim Po(\sum_j p_{ij}\lambda_j)$.

   (b)
   $$l(\boldsymbol{\lambda}) = \sum_i n_i \log(\sum_j p_{ij}\lambda_j) - \sum_{ij} p_{ij}\lambda_j \tag{1}$$

   The first term is concave down and the second term has zero curvature, so the log-likelihood is concave down unless all $n_i = 0$. There is a unique maximum for $\boldsymbol{\lambda}$.

   (c)
   $$\begin{aligned}
   \log(\sum_j p_{ij}\lambda_j) &\geq \sum_j q_{ij} \log(p_{ij}\lambda_j / q_{ij}) && (2)\\
   \text{where } \sum_j q_{ij} &= 1 && (3)\\
   l(\boldsymbol{\lambda}) &\geq g(\boldsymbol{\lambda}) = \sum_i n_i \sum_j q_{ij} \log(p_{ij}\lambda_j/q_{ij}) - \sum_{ij} p_{ij}\lambda_j && (4)
   \end{aligned}$$

   In the E-step, we choose q_{ij} to make contact at $\boldsymbol{\lambda}^{old}$:

   $$q_{ij} = \frac{p_{ij}\lambda_j^{old}}{\sum_j p_{ij}\lambda_j^{old}} \tag{5}$$

   In the M-step, we choose $\boldsymbol{\lambda}$ to maximize the bound:

   $$\begin{aligned}
   \frac{dg(\boldsymbol{\lambda})}{d\lambda_j} &= \frac{\sum_i n_i q_{ij}}{\lambda_j} - \sum_i p_{ij} && (6)\\
   \lambda_j &= \frac{\sum_i n_i q_{ij}}{\sum_i p_{ij}} && (7)
   \end{aligned}$$

   (d) Before observing n_i, we know that $n_{ij} \sim Po(p_{ij}\lambda_j)$. Therefore the joint distribution is

   $$p(n_i | n_{i1}, ..., n_{iJ}) p(n_{i1}, ..., n_{iJ} | \boldsymbol{\lambda}) = \delta(n_i - \sum_j n_{ij}) \prod_j e^{-p_{ij}\lambda_j} (p_{ij}\lambda_j)^{n_{ij}} / (n_{ij}!) \tag{8}$$

   Dividing this by the known marginal for n_i (part (a)) gives

   $$\begin{aligned}
   p(n_{i1}, ..., n_{iJ} | n_i, \boldsymbol{\lambda}) &= \delta(n_i - \sum_j n_{ij}) \binom{n_i}{n_{i1} \cdots n_{iJ}} \prod_j \left(\frac{p_{ij}\lambda_j}{\sum_j p_{ij}\lambda_j} \right)^{n_{ij}} && (9)\\
   &\sim \text{Multinomial}(n_i, q_{i1}, ..., q_{iJ}) && (10)
   \end{aligned}$$

(e) Let $\mathbf{h} = \{n_{ij}\}$ be the hidden variables. The complete-data likelihood is

$$p(D, \mathbf{h}|\boldsymbol{\lambda}) = \prod_{ij} Po(n_{ij}; p_{ij}\lambda_j) = \prod_{ij} e^{-p_{ij}\lambda_j} (p_{ij}\lambda_j)^{n_{ij}} / (n_{ij}!) \tag{11}$$

EM maximizes

$$\int_{\mathbf{h}} p(\mathbf{h}|D, \boldsymbol{\lambda}^{old}) \log p(D, \mathbf{h}|\boldsymbol{\lambda}) d\mathbf{h} = -\sum_{ij} p_{ij}\lambda_j + \sum_{ij} E[n_{ij}] \log(p_{ij}\lambda_j) \tag{12}$$

$$E[n_{ij}] = n_i q_{ij} \tag{13}$$

This gives the same EM algorithm we derived in part (c).

2. The problem reduces to a short bit of matlab code:

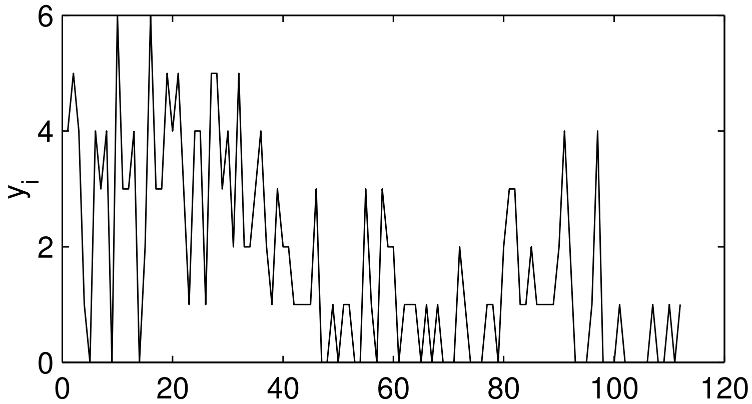
```
c = 2;
n = 10000;
y = tan(pi*rand(1,n));
f = (y > c);
mean(f)
std(f)/sqrt(n)

f = (abs(y) > c)/2;
mean(f)
std(f)/sqrt(n)

x = c./rand(1,n);
f = (1/c/pi)*x.^2./(1+x.^2);
mean(f)
std(f)/sqrt(n)
```

For a typical run, the results are:

Proposal	$\hat{I}$	std($\hat{I}$)
Cauchy	0.1451	0.0035
Half-Cauchy	0.1456	0.0023
Pareto	0.1476	0.0001

The Pareto proposal is best by far, because it focuses on the nonzero part of $f(y)$. Its advantage increases with increasing c. It is important to address sparsity in statistical integrals.

3. (a) It is clear that something changed around $t = 40$, but to pin down the time exactly, we need statistical methods.



(b) The main trick in changepoint analysis is to condition on k and reduce to a simple problem. The likelihood given k is

$$p(D|\lambda_1, \lambda_2, k) = \prod_{i=1}^{k} Po(y_i; \lambda_1) \prod_{i=k+1}^{n} Po(y_i; \lambda_2) \tag{14}$$

Therefore the MLEs for λ_1 and λ_2 given k are

$$\hat{\lambda}_1 = \bar{y}_1 = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad \hat{\lambda}_2 = \bar{y}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n} y_i \tag{15}$$

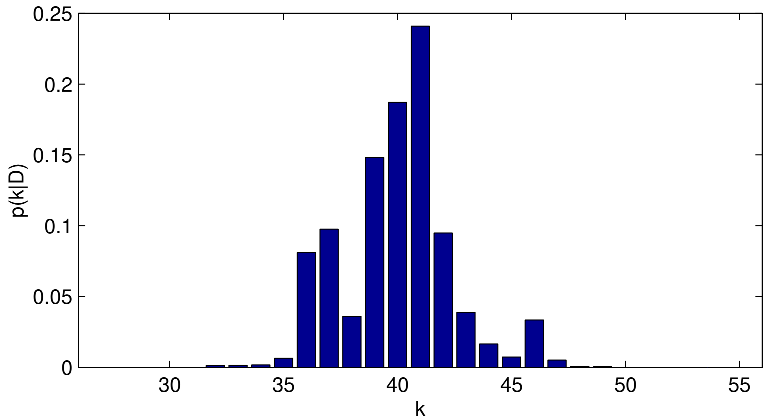
To find the joint MLE, just enumerate all k and pick the place where the likelihood $p(D|k, \hat{\lambda}_1, \hat{\lambda}_2)$ is maximum. This occurs at $(k = 41, \lambda_1 = 3.0976, \lambda_2 = 0.9014)$.

(c) Define $n_1 = k$, $n_2 = n - k$.

$$\begin{aligned}
p(\lambda_j|D, k, b_j) &\propto p(D|\lambda_1, \lambda_2, k)p(\lambda_j|b_j) && (16)\\
&\sim \mathcal{G}(a_j + n_j\bar{y}_j, 1/(1/b_j + n_j)) && (17)\\
p(b_j|D, k, \lambda_j) &\propto p(\lambda_j|b_j)p(b_j) && (18)\\
&\sim \mathcal{IG}(a_j + c_j, \lambda_j + d_j) && (19)\\
p(k|D, \lambda_1, \lambda_2) &\propto p(D|\lambda_1, \lambda_2, k) \propto (\lambda_1/\lambda_2)^{\sum_{i=1}^{k} y_i} e^{k(\lambda_2-\lambda_1)} && (20)
\end{aligned}$$

(d) Matlab's `gamrnd(a,b)` will return a sample from $\mathcal{G}(a, b)$. To sample from the inverse Gamma distribution $\mathcal{IG}(c, d)$, just use `1/gamrnd(c,1/d)`. Besides k, we also need a starting value for b_j or for λ_j. The MLE for λ_j will do, as well as sampling b_j from its prior.

(e) The estimated posterior is shown below:



It matches the figure in the book and is stable over multiple runs, even with bad initialization like $k = 2$. It isn't even necessary to sample over b_j; setting $b_j = 1$ throughout the run gives the same estimate. A histogram of k over the 10 samples shows that the changepoint lies between 37–42 but contains none of the detailed structure of the above figure.

Note that we could also have approximated $p(k|D)$ by enumerating k and approximating the marginal likelihood $p(D|k) = \int_{\lambda_1,\lambda_2} p(D|k, \lambda_1, \lambda_2)p(\lambda_1)p(\lambda_2)\ d\lambda_1\ d\lambda_2$. For example, using the maximized likelihood $p(D|k) \approx p(D|k, \hat{\lambda}_1, \hat{\lambda}_2)$ works pretty well for this problem.