Skilling's Nests

Carlos C. Rodríguez

May 31, 2006

Problem:

Given $q(\theta) \geq 0$ with $\theta \in R^k$. I want estimates for,

1.
$$Z = \int q(\theta) d\theta$$

2.
$$E_p[g(\theta)] = \int g(\theta) p(\theta) d\theta$$

where $p(\theta) = q(\theta)/Z$ and $g(\theta)$ is an arbitrary function of θ for which the expectation exists.

Assumption:

We have an approximation $\hat{p}(\theta)$ for $p(\theta)$ and an algorithm for generating samples from $\hat{p}$. For example $\hat{p}$ could be uniform or obtained by fitting gaussians at the peaks of q or using Minka's EP or just the prior when q is an unnormalized posterior which is Skilling's original recommendation.

Rearrangement of Probability

Let,
$$r(\theta) = \frac{q(\theta)}{\hat{p}(\theta)}$$
(r is for ratio and for radius) and define for $t \geq 0$,
$$v(t) = \int_{\{r>t\}} \hat{p}(\theta) d\theta$$

Note: v is not velocity but volume and t is not time but radius and I am not sure if the pun should be intended. Notice also that $1 - v(r)$ is the cdf of

the non-negative scalar $r(\theta)$ when $\theta \sim \hat{p}$. Thus, $v(r)$ decreases from $v(0) = 1$ to $v(\infty) = 0$ and,

$$\langle r \rangle_{\hat{p}} = \int_0^\infty v(r)dr = \int r\hat{p}\, d\theta = Z$$

hence, integrating the other way,

$$Z = \int_0^1 r(v)dv$$

where I harass the notation a little by identifying r (which was defined as a function of θ) with v^{-1} which is a function (of volume) defined on $[0, 1]$.

Two propositions and one corollary

Prop1: If $\theta \sim \hat{p}$ then $v = v(r(\theta)) \sim \text{unif}(0, 1)$.
proof:

$$P\{v \le c\} = \hat{P}\{\theta : r(\theta) \ge r_c\}$$

where $v(r_c) = c$. Hence,

$$P\{v \le c\} = v(r_c) = c$$

and v has the uniform$(0, 1)$ cdf. **qed**.

Now let,

$$\hat{p}_t(\theta) = \frac{1}{v(t)}\hat{p}(\theta)$$

provided $r(\theta) > t$ and define $\hat{p}_t(\theta) = 0$ when $r(\theta) \le t$. In other words, $\hat{p}_t$ is the conditional pdf of $\theta \sim \hat{p}$ given that $r > t$.

The algorithm to sample from $\hat{p}$ can be trivially used to sample from $\hat{p}_t$ for any $t \ge 0$. This however, becomes increasingly expensive as t increases and needs (special) MCMC for t large (see my Metric MonteCarlo 07).
Prop2: If $\theta_1, \theta_2, \ldots, \theta_N$ are iid from $\hat{p}_t(\theta)$, then $v_1, \ldots, v_N$ are iid unif$(0, v(t))$. Where $v_j = v(r(\theta_j))$ for $j = 1, \ldots, N$.
proof: The θ_j are iid so the v_j are also iid. Thus, we only need to check the distribution of one v_j. We have,

$$P\{v_j \le c\} = \hat{P}_t\{r(\theta_j) \ge r_c\} = \frac{c}{v(t)}.$$

qed.

An immediate consequence of this last proposition is:
Corollary1: If $\theta_1, \theta_2, \ldots, \theta_N$ are iid $\hat{p}_{t_0}(\theta)$, then

$$v_N^* = \max\{v_1, \ldots, v_N\}$$

has pdf

$$p_N^*(v) = \frac{N}{v_0^N} v^{N-1} \text{ for } 0 \le v \le v_0$$

where $v_0 = v(t_0)$.
proof:

$$P\{v_N^* \le v\} = \prod_{j=1}^{N} \hat{P}_{t_0}\{v_j \le v\} = \left(\frac{v}{v_0}\right)^N .$$

The derivative with respect to v is p_N^*. **qed**.

Using the density p_N^* we compute (just do it!) the mean,

$$E[v_N^*] = \frac{N}{N+1} v_0$$

and the standard deviation,

$$\sigma[v_N^*] = \frac{1}{N+1}\sqrt{\frac{N}{N+2}}\ v_0 = \sigma_N$$

Harvesting the diamond eggs

Here is John `skilling@eircom.ie` home run (sans the dogma): When $\theta \sim \hat{p}_{t_0,N}$ (explicitly defined below) having the property that $v(r(\theta)) \sim p_N^*$ we have available its radius $r(\theta) = q(\theta)/\hat{p}(\theta)$ and an estimate of the volume that it encloses,

$$v(r(\theta)) = \frac{N}{N+1} v_0 \pm \sigma_N$$

with $\sigma_N = O(1/N)$ so it can be made arbitrarily precise by just increasing N.

The Algorithm

Starting from $(r_0, v_0) = (0, 1)$, Skilling's algorithm sequentially builds an stochastic sequence $(r_1, v_1), (r_2, v_2), \ldots, (r_m, v_m)$ to approximate the curve $v = v(r)$. It then estimates the enclosed area by,

$$\frac{1}{2}\sum_{j=1}^{m}(v_{j-1} + v_j)(r_j - r_{j-1}) \approx Z.$$

Explicitly,

SETUP: Define $r_0 = 0$ and $v_0 = 1$ which can be thought as effectively starting from $\theta_0 = \infty$. Pick a set $S^0 = \{\theta_1^0, \theta_2^0, \ldots, \theta_N^0\}$ of N iid $\hat{p}$ (which is the same as $\hat{p}_0$) points.
LOOP: For $j = 1, 2, \ldots, m$ sequentially define:

$$\theta_j \sim \hat{p}_{r_{j-1},N}$$

by choosing,

$$\theta_j = \arg\min_{\theta \in S^0} \{r(\theta)\}$$

$$r_j = r(\theta_j)$$

$$v_j = v(r_j) \sim p_N^*$$

peel off θ_j from S^0 replacing it by a new

$$\theta \sim \hat{p}_{r_j}$$

i.e.,

$$S^0 \leftarrow S^0 \setminus \{\theta_j\} \cup \{\theta\}$$

END.

In the algorithm, $\hat{p}_{t,N}(\theta)$ denotes the pdf of a θ^* obtained as the outer-most (i.e. minimum radius or equivalently maximum volume) point from the set S^0 of N iid $\hat{p}_t$ points. We have,

$$\theta^* = \arg\max_{\theta \in S^0} \{v(r(\theta))\}.$$

The distribution of θ^* is readily obtained by noticing that,

$$P\{\theta^* \in A | N, t\} = \int_0^1 P\{\theta^* \in A | v_N^* = v\} p_N^*(v) dv$$

and using the expression for p_N^*,

$$P\{\theta^* \in A | N, t\} = \int_0^{v_0} \lambda(A; v) \frac{N v^{N-1}}{v_0^N} \, dv$$

where $\lambda(A; c)$ denotes the normalized volume of the set A in the manifold,

$$M_c = \{\theta : v(r(\theta)) = c\}$$

of equiprobable points at fix radius r_c so that $v(r_c) = c$. Denoting the uniform density on M_v by $\lambda(\theta; v)$ we can write,

$$\hat{p}_{t,N}(\theta) = \int_0^{v_0} \lambda(\theta; v) d \left(\left(\frac{v}{v_0} \right)^N \right)$$

Thus,

$$\hat{p}_{t,N}(\theta) = \int_0^1 \lambda(\theta; v(t) u^{1/N}) \, du$$

and the joint density of the sequence $\theta_1, \ldots, \theta_m$ generated by the algorithm is,

$$\hat{p}^*(\theta_1, \ldots, \theta_m | N) = \prod_{j=1}^m \hat{p}_{r(\theta_{j-1}),N}(\theta_j).$$

Therefore, for finite N, the θ_j follow the above complicated distribution. However,

$$\lim_{N\to\infty} \hat{p}_{t,N}(\theta) = \lambda(\theta; v(t))$$

and we can write for N large that,

$$\hat{p}^*(\theta_1, \ldots, \theta_m | N) \approx \prod_{j=1}^{m} \lambda(\theta_j; v_{j-1})$$

which shows that $\theta_1, \ldots, \theta_m$ are the rearrangement of m iid r^* with

$$r^*(\theta) = \frac{1}{Z_0} \frac{q(\theta)}{\hat{p}(\theta)}$$

and

$$Z_0 = \int r(\theta) d\theta.$$

Expectations under $p(\theta)$ are then obtained for any (proper) function $g(\theta)$ as,

$$E_p[g(\theta)] = \frac{Z_0}{Z} E^*[g(\theta)\hat{p}(\theta)]$$

where E^* denotes expectation with respect to r^*. The ratio of normalizing constants Z/Z_0 is independent of g so by taking $g = 1$ for all θ we can write,

$$\frac{Z}{Z_0} = E^*[\hat{p}(\theta)] \approx \frac{1}{m} \sum_{j=1}^{m} \hat{p}(\theta_j)$$

and,

$$E_p[g(\theta)] = \frac{E^*[g(\theta)\hat{p}(\theta)]}{E^*[\hat{p}(\theta)]} \approx \frac{\sum_{j=1}^{m} g(\theta_j)\hat{p}(\theta_j)}{\sum_{j=1}^{m} \hat{p}(\theta_j)}$$

references

For more information and nicer illustrations just Google: “nested sampling”