Among the things I learned as a student at Madison are
1) think as Bayesian -- George Box & George Tiao
2) think in function spaces -- Grace Wahba

Today’s talk will show that I haven’t forgotten these lessons

Optional Polya Tree & Bayesian Inference

Wing H. Wong
(Wong & Li, Annals of Statistics, 2010, June issue)
Example 1: modeling flow cytometry data by a density in $\mathbb{R}^k$
<table>
<thead>
<tr>
<th>Genotype of subject</th>
<th>Disease Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>Example 2:</td>
</tr>
<tr>
<td>no</td>
<td>Modeling joint status of markers by a $3^k$ table</td>
</tr>
<tr>
<td>no</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>
The Bayesian nonparametric problem

- \( x_1, x_2, \ldots x_n \) are independent r.v. on a space \( \Omega \)
- drawn from a common distribution \( Q \) on \( \Omega \)
- \( Q \) is unknown but assumed to have a prior distribution \( \pi \).

Our task is to construct a class of priors for the distribution \( Q \) so that Bayesian Inference on \( Q \) is feasible

Want this to work well in moderate dimensions, e.g. \( k=10 \)
Ferguson’s conditions (1973)

- Support of $\pi$ should be large
- The corresponding posterior should be tractable

Dirichlet process prior with parameter $\alpha$:

$Q(\cdot)$ is a stochastic process indexed by subsets of $\Omega$ such that for disjoint sets $A_i$’s,

$$(Q(A_i), i=1,...k) \sim \text{Dirichlet}\left(\alpha(A_i), i=1,...k\right)$$

where $\alpha$ is a measure on $\Omega$.

Ferguson shows that this prior satisfies the two conditions. However, when $\Omega$ is Euclidean the random distribution $Q$ does not possess a density
Density is needed in many applications

- Estimating Kullback-Leibler divergence
  \[ \int \log(g/f) f(x)dx \]

- Error distribution for regression
  \[ y_i = g(x_i ; \theta) + \varepsilon_i \quad \text{where } \varepsilon \text{ has density } q( ) \]

Likelihood \((\theta \mid q) = \prod q( y_i - g(x_i ; \theta) )\)
Partitioning scheme

Suppose $\Omega$ can be partitioned in one of $M$ ways:
   For $j=1, 2, \ldots, M$,
   \[
   \Omega = \bigcup_{k=1}^{K^j} \Omega^j_k
   \]
   where $\Omega^j_k$’s are disjoint

Each level-1 elementary region $\Omega^j_k$ can be further partitioned in one of several ways into level-2 elementary regions:
   \[
   \Omega^j_{k_1} = \bigcup_{k_2=1}^{K^{j_1j_2}} \Omega^{j_1j_2}_{k_1k_2}
   \]

Let $A^k$ be the set of elementary regions with level $= k$,
   $A^{(k)}$ be the set of elementary regions with level $\leq k$
In general, $A^j_k = k^{th}$ part of the $j^{th}$ way to partition $A$
Some Examples

Example 1.

\[ \Omega = \{x = (x_1, \ldots, x_p) : x_i \in \{1, 2\}\} \]
\[ \Omega^j_k = \{x : x_j = k\}, \quad k = 1 \text{ or } 2 \]
\[ \Omega^{j_1j_2}_{k_1k_2} = \{x : x_{j_1} = k_1, x_{j_2} = k_2\}, \text{ etc.} \]

In this example, the number of ways to partition a level-$k$ elementary region decreases as $k$ increases.

Example 2.

\[ \Omega = \{(x_1, x_2, \ldots, x_p) : x_i \in [0, 1]\} \subset \mathbb{R}^p \]

If $A$ is a level-$k$ elementary region (a rectangle) and $m_j(A)$ is the midpoint of the range of $x_j$ for $A$, we set $A^j_1 = \{x \in A : x_j \leq m_j(A)\}$ and $A^j_2 = A \setminus A^j_1$. There are exactly $p$ ways to partition each $A$, regardless of its level.
Piecewise constant density

• \( S \leftarrow \text{Ber} (\rho) \), if \( S=1 \), \( Q(1) \leftarrow \text{uniform on } \Omega \), stop.

• Else,
  
  draw \( J=j \) with probability=\( \lambda_j \)

  use the \( j^{\text{th}} \) partition of \( \Omega \), i.e.,

  \[
  \Omega = \bigcup_{k=1}^{K} \Omega_k^j
  \]

• \( \theta^j = (\theta_1^j, \ldots, \theta_K^j) \leftarrow \text{Dirichlet} \ (\alpha_1^j, \ldots, \alpha_K^j) \)

• \( Q(1) (\Omega_k^j) \leftarrow \theta_k^j \)

• \( Q(1) (\mid \Omega_k^j) \leftarrow \text{uniform} \)
Piecewise constant density on partitions of finite depth

• Suppose we have drawn $Q^{(k)}$ supported on a partition composing of regions from $A^{(k)}$
• For each region not yet stopped, repeat the partitioning process
• This gives a random distribution $Q^{(k+1)}$ with a density $q^{(k+1)}$ that is piecewise constant on a partition with regions from $A^{(k+1)}$
• Note: this is just a random recursive partitioning process
Definition of Optional Polya Tree (OPT)

**Theorem 1.** Suppose there is a $\delta > 0$ such that with probability 1, $\rho(A) > \delta$ for any region $A$ generated during any step in the recursive partitioning process. Then with probability 1, $Q^{(k)}$ converges in variational distance to a probability measure $Q$ that is absolutely continuous with respect to $\mu$.

\[
i.e. \quad P \left\{ \int |q^{(k)}-q|dx \to 0 \text{ for some density } q \right\} = 1
\]

This random probability measure $Q$ is said to have an Optional Polya Tree distribution with parameters $\rho$ (stopping rule), $\lambda$ (selection probabilities) and $\alpha$ (probability assignment weights).
OPT prior has large support in $L_1$

**Theorem 2.** Let $\Omega$ be a bounded rectangle in $\mathbb{R}^p$. Suppose that the condition of Theorem 1 holds and that the selection probabilities $\lambda_i(A)$, the assignment probabilities $\alpha_i^j(A)/(\sum_l \alpha_l^j(A))$ for all $i, j$ and $A \in \mathcal{A}^{(\infty)}$, are uniformly bounded away from 0 and 1. Let $q = dQ/d\mu$, then for any density $f$ and any $\tau > 0$, we have

$$P \left( \int |q(x) - f(x)| \, d\mu < \tau \right) > 0$$

**Remark:** A useful choice for $\alpha$ is

$$\alpha_i^j(A) = \mu \left( A_i^j \right) / \mu(A) \quad \text{for} \ A \in \mathcal{A}^k$$
Theorem 3:

The posterior distribution \( \pi(Q \mid x_1, \ldots x_n) \) is also OPT with

1. **Stopping probability:**

\[
\rho(A \mid x) = \rho(A) \Phi_0(A) / \Phi(A)
\]

2. **Selection probabilities:**

\[
P(J = j \mid x) \propto \lambda_j \frac{D(n_j + \alpha_j)}{D(\alpha_j)} \prod_{i=1}^{K_j} \Phi \left( A_i^j \right) \quad j = 1, \ldots, M
\]

3. **Allocation of probability to subregions:** the probabilities \( \theta_i^j \) for subregion \( A_i^j, i = 1, \ldots, K_j \) are drawn from Dirichlet \( (n_j + \alpha_j) \).

where

\[
\Phi(A) = \int q(x(A) \mid A) \, d\pi_A(q)
\]

\[
\Phi_0(A) = u(x(A) \mid A)
\]
Computation of $\Phi(A)$ by recursion

If $A = \bigcup_{i=1}^{K_j} A_i^j$

then

$$\Phi(A) = \rho \Phi_0(A) + (1 - \rho) \sum_{j=1}^{M} \lambda_j \frac{D(n_j^j + \alpha_j^j)}{D(\alpha_j^j)} \prod_{i=1}^{K_j} \Phi\left(A_i^j\right)$$

where $D(t) = \Gamma(t_1) \ldots \Gamma(t_k) / \Gamma(t_1 + \cdots + t_k)$
Termination rule for Recursion (case of $2^p$ table)

1. $A$ contains no observation. In this case, $\Phi(A) = 1$.
2. $A$ is a single cell (in the $2^p$ table) containing any number of observations. In this case, $\Phi(A) = 1$.
3. $A$ contains exactly one observation and $A$ is a region where $M$ of the $p$ variables are still available for splitting. In this case,

   $$\Phi(A) = 2^{-M}$$

Similarly, termination rules exist for the continuous case

Thus, Ferguson’s second condition is also satisfied.
OPT prior leads to asymptotically consistent inference

**Theorem 4.** Let \( x_1, x_2, \ldots \) be independent, identically distributed variables from a probability measure \( Q \), \( \pi(\cdot) \) and \( \pi(\cdot|x_1, \ldots, x_n) \) be the prior and posterior distributions for \( Q \) as defined in Theorem 3. Then, for any \( Q_0 \) with a bounded density, it holds with \( Q_0^{(\infty)} \) probability equal to 1 that

\[
\pi(U|x_1, \ldots, x_n) \longrightarrow 1
\]

for all weak neighborhoods \( U \) of \( Q_0 \).

**Remark 1:**

\[
U = \left\{ Q : \left| \int g_i(\cdot) \, dQ - \int g_i(\cdot) \, dQ_0 \right| < \epsilon_i, \quad i = 1, 2, \ldots, K \right\}
\]

where \( g_i(\cdot) \) is a bounded continuous function on \( \Omega \).

**Remark 2:** It should be possible to get rates in Hellinger distance
Example 2

(a) Sample size = 100

(b) Sample size = 500
Example 2 (continued)

(c) Sample size = 1000

(d) Sample size = 5000
Comparison of two samples

$x_0$ is a sample from distribution $Q_0$

$x_1$ is a sample from distribution $Q_1$

The OPT can be used to derive tests statistics for the equality of the two distributions.

This is not an easy problem in the multivariate case
One approach based on OPT

• Given $x_0$, $Q_0$ has an OPT as it posterior
• We want to learn a partition for $Q_1$ that tells us on which parts of the sample space is $Q_1$ different from $Q_0$
• When deciding whether to stop or continue to divide $A$, replace $\Phi_0(A) = u(x_1(A) | A)$ by $\Phi_0(A|x_0) = \int q_0(x_1(A) | A) \pi_A(dq_0|x_0)$
• This can be computed by repeating the basic OPT posterior computation twice
Two simulations of case-control samples

1. $X_1, X_2, \ldots, X_{15} \sim_{i.i.d.} \text{Bernoulli}(0.5)$

2. $X_1, X_2, \ldots X_8$ as a Markov Chain with $X_1 \sim \text{Bernoulli}(0.5)$, and $P(X_t = X_{t-1} | X_{t-1}) = 0.7$, while $X_9, X_{10}, \ldots X_{15} \sim_{i.i.d} \text{Bernoulli}(0.5)$ and are independent of $X_1, \ldots, X_8$.

$$Y \sim \begin{cases} 
\text{Bernoulli}(0.3) & \text{if } X_3 = 1 \text{ and } X_7 = 1 \\
\text{Bernoulli}(0.2) & \text{if } X_7 = 0 \text{ and } X_{10} = 0 \\
\text{Bernoulli}(0.1) & \text{otherwise}
\end{cases}$$
For more examples, go to Li Ma’s oral, May 19, 2010
An example of partition learned from data