Lecturer: Geoff Nicholls

Lecture 13: Exchangeability and its implications

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/

and via the MSc weblearn pages.
de Finetti’s theorem

Consider an infinite sequence of random variables \( \{X_i\}_{i=1}^{\infty} \). The sequence is exchangeable if

\[
(X_1, X_2, \ldots, X_n) \sim (X_{\sigma_1}, X_{\sigma_2}, \ldots, X_{\sigma_n})
\]

for each \( n \geq 1 \) and any permutation \( \sigma \in \mathcal{P}_n, \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) of the numbers \( (1, 2, \ldots, n) \).

Example: an iid sequence of binary random variables \( X_i \sim \text{Bern}(\theta) \) is clearly exchangeable since

\[
\pi(x_1, \ldots, x_n) = \prod_i \theta^{x_i}(1 - \theta)^{1-x_i}
\]

and if we permute the indices we just shuffle the order of the indices in the product - they all appear once.
Polya urn: $b$ black and $w$ white balls in urn. Sample a ball. Let $X_i = \mathbb{I}_{\text{$i$th ball is black}}$. Ball returned plus $A$ balls of same color.

Variables $X_1, X_2, \ldots$ are not independent but they are exchangable. Consider the probability for sequences $0, 0, 1, 1$ and $1, 1, 0, 0$

\[
\pi(0011) = \frac{w}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{w+A}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{b}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{b+A}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)}
\]

\[
\pi(1100) = \frac{b}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{b+A}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{w}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)} \frac{w+A}{(w+b)(w+b+A)(w+b+2A)(w+b+3A)}
\]

Permuting arrivals shuffles the numerator factors, they all still appear exactly once. Generalises to sequences of arbitrary length.

If $\{X_i\}_{i=1}^{\infty}$ are iid they are exchangeable, but not conversely.
Hierarchical model: For \( n = 1, 2, \ldots \) let \( 0_n \) be a vector of \( n \) zeros, \( \Sigma_{i,i}^{(n)} = 1 \) and \( \Sigma_{i,j}^{(n)} = \rho \) with \( 1 > \rho \geq 0 \). The distribution

\[
\pi(x_1, \ldots, x_n) = N(x; 0_n, \Sigma^{(n)})
\]
gives an exchangeable sequence. For \( \sigma \in \mathcal{P}_n \) a permutation, we have \( E(X_{\sigma_i}) = \mu \) and \( \text{cov}(X_{\sigma_i}, X_{\sigma_j}) = \rho \), so that

\[
\pi(x_{\sigma_1}, \ldots, x_{\sigma_n}) = N(x_{\sigma}; 0_n, \Sigma^{(n)}) = N(x; 0_n, \Sigma^{(n)}).
\]
Here \( 0 \leq \rho < 1 \) is N&S for \( \Sigma^{(n)} \) positive definite for all \( n \geq 1 \).

The eigenvalues* are

\[
\lambda_1 = 1 + (n - 1) \rho, \quad \text{and} \quad \lambda_2 = \ldots = \lambda_n = 1 - \rho
\]
so \( -1/(n - 1) < \rho < 1 \) is N&S for PD \( \Sigma^{(n)} \) at fixed \( n \), and \( 0 \leq \rho < 1 \) is N&S to make this work for all \( n \geq 1 \).

* eigenvectors \((1, 1, 1, \ldots, 1)\) and anything orthogonal for eg \((1, -1, 0, 0, \ldots, 0)\), \((0, 1, -1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1, -1)\)
Theorem (de Finetti): Let $X_1, X_2, \ldots X_n, \ldots$ be an infinite sequence of binary random variables. The sequence is exchangeable if and only if there exists a distribution function $F(\theta)$ on $[0, 1]$ such that

$$p(x_1, \ldots, x_n) = \int_0^1 \left[ \prod_{i=1}^n p(x_i|\theta) \right] dF(\theta) \quad (*)$$

with

$$F(\theta) = \Pr(\Theta \leq \theta) \quad \text{where} \quad \Theta = \lim_{N \to \infty} N^{-1} \sum_{i=1}^N X_i.$$ 

and $p(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$. It further holds that the conditioned distribution is Bernoulli,

$$p(x_1, \ldots, x_n|\Theta = \theta) = \prod_{i=1}^n p(x_i|\theta).$$
Remarks In short, “an infinite exchangeable sequence is distributed as a mixture of iid random variables”. The theorem says $F$ and $\theta$ must exist to make this hold.

The theorem extends to cover infinite exchangeable sequences of random vectors (ie $X_i$ continuous multivariate random variables) with a multivariate parameter $\theta$.

The expression $dF(\theta)$ may be off-putting. If $F(\theta)$ is the cdf of a pdf $\pi$ then $dF(\theta) = \pi(\theta)d\theta$, and

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^n p(x_i|\theta)\pi(\theta)d\theta.$$

In general $F$ is just some unknown distribution which puts probability mass on sets $d\theta$ in parameter space $\Omega$.

It isnt always obvious what $F$ and $\theta$ are. Here is an example.
Hierarchical model: If $\Sigma_{ii}^{(n)} = 1$ and $\Sigma_{ij}^{(n)} = \rho$ with $0 \leq \rho < 1$ then

$$\pi(x_1, \ldots, x_n) = N(x; 0_n, \Sigma^{(n)})$$

gives an exchangeable sequence. Write it in de Finetti form.

Simulate $\theta \sim N(0, v)$ and set

$$X_i = \theta + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, w).$$

Marginally $E(X_i) = 0$, $\text{var}(X_i) = v + w$ and $\text{cov}(X_i, X_j) = v$ so $w = 1 - \rho$ and $v = \rho$ gives $X \sim N(0_n, \Sigma^{(n)})$. Conditionally,

$$\pi(x_1, \ldots, x_n|\theta) = \prod_i N(x_i; \theta, 1 - \rho)$$

and unconditionally

$$\pi(x_1, \ldots, x_n) = \int \prod_i N(x_i; \theta, 1 - \rho)N(\theta; 0, \rho)d\theta,$$

the de Finetti form, and equal to $N(x; 0_n, \Sigma^{(n)})$. 
Proof (of deF): we begin by looking at \( n \) of the first \( N \) outcomes in the sequence. We write down the joint pmf of \( X_1, \ldots, X_n \) and show that it converges to the RHS of Equation \((*)\) as \( N \to \infty \). Let

\[
S_n = X_1 + X_2 + \ldots + X_n, \quad n = 1, 2, \ldots
\]

and let \( r, s \) be two integers \( 0 \leq r \leq s \leq N \). By exchangeability, the conditional distribution of \( S_n \) given \( S_N \) is hypergeometric,

\[
\Pr(S_n = r | S_N = s) = \frac{{s \choose r} \left( \frac{N-s}{n-r} \right)}{{N \choose n}}
\]

since this is the probability to draw \( r \) 1's in \( n \) draws without replacement from an urn containing \( s \) 1's and \( N - s \) 0's.
When $S_n = r$ the urn contains at least $n - r$ 0's so

$$\Pr(S_n = r) = \sum_{s=r}^{N-(n-r)} \Pr(S_n = r | S_N = s) \Pr(S_N = s)$$

$$= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r | S_N/N = \theta(s)) \Pr(S_N/N = \theta(s))$$

where $\theta(s) \equiv s/N$.

The random variable $\Theta_N \sim S_N/N$ has CDF a $F_N(\theta)$ which is

$$F_N(\theta) = \Pr(\Theta_N \leq \theta),$$

$$= \sum_{s=0}^{N} \Pr(S_N = N\theta(s))\mathbb{1}_{\theta(s) \leq \theta}.$$
As $\theta$ increases past $\theta(s)$, $F_N(\theta)$ jumps up by $\Pr(S_N = N\theta(s))$. It is not differentiable at discontinuities, but we can write down a density

$$f_N(\theta) = \sum_{s=0}^{N} \Pr(S_N = N\theta)\delta(\theta - \theta(s))$$

which assigns the correct probability to sets when integrated. In this expression

$$\delta_{\theta(s)}(\theta) = \delta(\theta - \theta(s))$$

is a Dirac delta function which places a unit point mass at $\theta = \theta(s)$. This function is defined by its action in integrals: if $g(\theta)$ is continuous at $\theta(s)$ then

$$\int_{0}^{1} g(\theta)\delta(\theta - \theta(s))d\theta = g(\theta(s)).$$
Returning to our density we calculate

\[ \Pr(a \leq \Theta_N \leq b) = \int_a^b f_N(\theta)d\theta \]

\[ = \int_a^b \left[ \sum_{s=0}^{N} \Pr(S_N = N\theta)\delta(\theta - \theta(s)) \right] d\theta \]

\[ = \sum_{s=0}^{N} \Pr(S_N = N\theta(s))\mathbb{I}_{a\leq\theta(s)\leq b} \]

\[ = F_N(b) - F_N(a). \]

Point masses \( \delta_{\theta(s)} \) in \( f_N \) are associated with the discontinuities at \( \theta(s) \) in \( F_N \). The derivative

\[ dF(\theta) = f(\theta)d\theta, \]

is defined in terms of its action under an integral.

Exercise: show that \( \int_0^1 f_N(\theta)d\theta = 1 \).
Now write $\Pr(S_n = r)$ in terms of $F_N$. Let

$$g_N(\theta) = \mathbb{I}_{r \leq N \theta \leq N-(n-r)} \Pr(S_n = r|S_N = N\theta).$$

In terms of $F_N$, $f_N$ and $g_N$,

$$\int_0^1 g_N(\theta)dF_N(\theta) = \int_0^1 g_N(\theta)f_N(\theta)d\theta$$

$$= \sum_{s=0}^{N} \int_0^1 g_N(\theta) \Pr(S_N = N\theta)\delta(\theta - \theta(s))d\theta$$

$$= \sum_{s=0}^{N} g_N(\theta(s)) \Pr(S_N = N\theta(s))$$

$$= \sum_{s=r}^{N-(n-r)} [\Pr(S_n = r|S_N = N\theta(s)) \times \Pr(S_N = N\theta(s))]$$

$$= \Pr(S_n = r) \ldots \text{so we can write...}$$
... so we can write

\[ Pr(S_n = r) = \int_0^1 g_N(\theta) dF_N(\theta) \]

\[ = \int_0^1 \mathbb{I}_{r \leq N\theta \leq N - (n-r)} Pr(S_n = r|S_N = N\theta) dF_N(\theta) \]

\[ = \int_{r/N}^{1-(n-r)/N} Pr(S_n = r|S_N = N\theta) dF_N(\theta) \]

We now assume two limiting results. The hypergeometric \( Pr(S_n = r|S_N = N\theta) \) converges uniformly to the binomial,

\[ \frac{\binom{N\theta}{r} \binom{N(1-\theta)}{n-r}}{\binom{N}{n}} \rightarrow \binom{n}{r} \theta^r (1 - \theta)^{n-r} \]

as \( N \to \infty \) at fixed \( \theta \). I leave this to you to show. It makes sense because we have \( N\theta \) 1's and \( N(1 - \theta) \) 0's, and there is little difference between sampling with and without replacement when we sample a small number from a large population.
Helly's Theorem (Feller (1966) *Probability Theory and Applications* II) “an infinite sequence of probability distributions $F_N$ on a finite interval contains a convergent subsequence”.

It follows that a limit $F_N(\theta) \to F(\theta)$ exists, that is

$$
\lim_{N \to \infty} \Pr(N^{-1}S_N \leq \theta) = \Pr(\Theta \leq \theta)
$$

where $\Theta = \lim_{N \to \infty} N^{-1}S_N$. Assuming these two convergence results hold, then for the infinite exchangeable sequence,

$$
\Pr(S_n = r) = \binom{n}{r} \int_0^1 \theta^r (1 - \theta)^{n-r} dF(\theta) \quad (A)
$$

Finally then,

$$
\Pr(S_n = r) = \binom{n}{r} p(x_1, \ldots, x_n) \quad (B)
$$
for any $x_1, \ldots, x_n$ summing to $r$, so subbing (B) in (A),

$$p(x_1, \ldots, x_n) = \int_0^1 \theta^r (1 - \theta)^{n-r} dF(\theta) \quad (C)$$

which is the result of de Finetti.

The representation in terms of $F$ is unique. A distribution on a bounded interval is uniquely determined by its moments, and since (C) fixes all the moments $E(\theta^n), n = 1, 2, \ldots$ of $F$ (take $r = n$ and $n = 1, 2, \ldots$) it follows that $F$ exists and is unique. [EOP]
Bayesian inference: If the data are $x_1, \ldots, x_n$ then

$$p(x_1, \ldots, x_n|\theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i}$$

is the likelihood and $F(\theta)$ is the CDF of the prior. de Finetti gives the form for the prior predictive distribution of the data

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n|\theta) dF(\theta).$$

If the observations are an infinite exchangeable sequence, then all these objects exist.

Suppose we have seen $x_1, \ldots, x_m$ and we wish to predict $x_{m+1}, \ldots, x_n$. The predictive distribution is

$$p(x_{m+1:n}|x_{1:m}) = p(x_{1:n})/p(x_{1:m})$$

$$= \int p(x_{m+1:n}|\theta) \frac{p(x_{1:m}|\theta)dF(\theta)}{p(x_{1:m})}.$$
We see that

$$dF(\theta|x_1, \ldots, x_m) \propto p(x_1, \ldots, x_n|\theta)dF(\theta)$$

is our updated prior, in other words this is the posterior for $\theta$ given data $x_1, \ldots, x_m$. de Finetti tells us that in this exchangeable setting we should be doing Bayesian inference... if we can.
De Finetti and the Polya urn

It may be shown that, for a Polya urn,

\[ p(x_1, \ldots, x_n) = \int \prod_i \theta^{x_i} (1 - \theta)^{1-x_i} \text{Beta}(\theta; b/A, w/A) d\theta. \]

ie, without using de Finetti.

Exercise: suppose \( \sum_{i=1}^n x_i = k \). Show for the polya urn that

\[ p(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{k-1} (b + iA) \prod_{j=1}^{n-k-1} (w + jA)}{\prod_{i=1}^{n-1} b + w + iA}. \]

and write this in terms of \( \Gamma \)-functions using the identity \( x\Gamma(x) = \Gamma(x+1) \). Do the integral above and hence show directly the LHS and RHS are equal.