10. Exchangeability and hierarchical models

Objective

Introduce exchangeability and its relation to Bayesian hierarchical models. Show how to fit such models using fully and empirical Bayesian methods.

Recommended reading


Exchangeability

Suppose that we have a sequence of variables $X_1, X_2, \ldots, X_n$. Then, in many cases, we might wish to assume that the subscripts of each individual variable are uninformative. For example, in tossing a coin three times, it is natural to assume that $P(0, 0, 1) = P(0, 1, 0) = P(1, 0, 0)$. This idea underlines the concept of *exchangeability* as developed by De Finetti (1970, 1974).

**Definition 27**
A sequence of random variables $X_1, \ldots, X_n$ is said to be (finitely) *exchangeable* if the distribution of any permutation is the same as that of any other permutation, that is if

$$P(\bigcap_{i=1}^n X_{\pi(i)}) = P(\bigcap_{i=1}^n X_i)$$

for all permutation functions $\pi(\cdot)$.

The definition of exchangeability can be extended to infinite sequences of variables.
**Definition 28**

An infinite sequence, $X_1, X_2, \ldots$ is said to be (infinitely) exchangeable if every finite subsequence is judged to be exchangeable in the above sense.

Thus, a sequence of variables that are judged to be independent and identically distributed is exchangeable. However, exchangeability is clearly a weaker concept more related to symmetry.

For example, if $X_1, X_2, \ldots, X_5$ are the results of 5 draws without replacement from a pack of cards, then this sequence is exchangeable but clearly, the variables are not independent.

Typical non exchangeable variables are Markov chains or other time varying sequences.
De Finetti’s theorem for 0-1 random variables

Assume that we have an infinitely exchangeable sequence of 0-1 variables. For example, we may believe that an infinite sequence of tosses of the same coin is exchangeable. Then, De Finetti derived the following theorem.

**Theorem 41**
If $X_1, X_2, \ldots$ is any infinitely exchangeable sequence of 0-1 variables with probability measure $F$ then there exists a distribution function $P$ such that

$$f(x_1, \ldots, x_m) = \int_0^1 \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{m-x_i} \, dP(\theta)$$

where $P(\theta) = \lim_{n \to \infty} F(Y_n/n \leq \theta)$, $Y_n = \sum_{i=1}^n X_i$ and $\lim_{n \to \infty} Y_n/n = \theta$.

**Proof** See Bernardo and Smith (1994).
Interpretation of De Finetti’s theorem

If a sequence $X_1, X_2, \ldots$ of 0-1 variables is judged to be exchangeable, then we may interpret this as if

- The $X_i$ are judged to be Bernoulli variables given some random variable $\theta$.
- $\theta$ is given a probability distribution $P$.
- Using the strong law of large numbers, $\theta = \lim_{n \to \infty} Y_n/n$ which implies that we can interpret $P$ as representing our beliefs about the limiting frequency of 1’s.

Thus, $P$ may be interpreted as a prior distribution.

http://en.wikipedia.org/wiki/De_Finetti%27s_theorem
An immediate consequence of Theorem 41 is that if we define $Y_n = \sum_{i=1}^{n} X_i$, then automatically, the distribution of $Y_n$ can be represented as

$$f(Y_n = y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1 - \theta)^{n-y_n} P(\theta) d\theta.$$ 

Thus, if we are expressing our beliefs about $Y_n$, then we are justified in acting as if the likelihood were binomial and with a prior distribution $P(\theta)$.

However, a much stronger general representation theorem is available for any infinitely exchangeable sequence of variables.
De Finetti’s general theorem

**Theorem 42**
Let $X_1, X_2, \ldots$ be an infinitely exchangeable sequence of variables with probability measure $F$. Then, their exists a probability measure $P$ such that the joint distribution of $X_1, \ldots, X_n$ can be represented as

\[
F(x_1, \ldots, x_n) = \int_G \prod_{i=1}^n G(x_i) dP(G)
\]

where $G$ is the space of distribution functions, $P(G) = \lim_{n \to \infty} F(G_n)$ and $G_n$ is the empirical distribution function defined by $X_1, \ldots, X_n$.

**Proof** See Bernardo and Smith (1994).

The theorem implies that if the $X_i$ are judged to be exchangeable, then their exists a variable $\theta$ such that

\[
F(x) = \int_\Theta \prod_{i=1}^n F(x_i | \theta) dP(\theta).
\]
De Finetti’s theorems provide a theoretical justification of Bayesian inference based on the assumptions of exchangeability. However, they are generally not very useful in the practical determination of the form of the prior distribution.

Certain extensions can be used to justify more specific distributional models. For example if we believe that the sequence $X_1, X_2, \ldots$ is exchangeable and spherically symmetric, i.e. that the distribution of $X_n$ is the same as the distribution of $AX_n$ for any orthogonal matrix $A$, then this implies that the $X$’s may be interpreted as normally distributed given a prior distribution on the precision.

See Bernardo and Smith (1994) for details and extensions to other models.
Hierarchical models

In many models we are unclear about the extent of our prior knowledge. Suppose we have data \( x \) with density \( f(x|\theta) \) where \( \theta = (\theta_1, \ldots, \theta_k) \). Often we may make extra assumptions about the structural relationships between the elements of \( \theta \).

Combining such structural relationships with the assumption of exchangeability leads to the construction of prior density, \( p(\theta|\phi) \), for \( \theta \) which depends upon a further, unknown hyperparameter \( \phi \).

In such cases, following Good (1980), we say that we have a hierarchical model.
Examples of hierarchical models

Example 74
Various individuals $i = 1, \ldots, n$, take an IQ test where it is supposed that the result is

$$Y_i | \theta_i, \phi \sim \mathcal{N} \left( \theta_i, \frac{1}{\phi} \right)$$

where the outcome for subject $i$ is supposed to depend on his or her true IQ $\theta_i$. Now if we suppose that the true IQ’s of the people in the study are exchangeable then we might reasonably assume that

$$\theta_i | \mu, \psi \sim \mathcal{N} \left( \mu, \frac{1}{\psi} \right)$$

where the unknown hyperparameters are $\mu$, representing the mean true IQ in the population, and $\psi$. 

Bayesian Statistics
Example 75
George et al (1993) analyzed data concerning failures of 10 power plant pumps. The number of failures $X_i$ at plant $i$ was assumed to follow a Poisson distribution

$$X_i | \theta_i \sim \mathcal{P}(\theta_i t_i) \text{ for } i = 1, \ldots, 10,$$

where $\theta_i$ is the failure rate for pump $i$ and $t_i$ is the length of operation time of the pump (in 1000s of hours).

It is natural to assume that the failure rates are exchangeable and thus we might model

$$\theta_i | \alpha, \beta \sim \mathcal{G}(\alpha, \beta)$$

where $\alpha$ and $\beta$ are the unknown hyperparameters.
Fitting hierarchical models

The most important problem in dealing with hierarchical models is how to treat the hyperparameters $\phi$. Usually, there is very little prior information available with which to estimate $\phi$.

Thus, two main approaches have developed:

- The natural Bayesian approach is to use relatively uninformative prior distributions for the hyperparameters $\phi$ and then perform a fully Bayesian analysis.

- An alternative is to estimate the hyperparameters using classical statistical methods.

This second method is the so-called *empirical Bayes* approach which we shall explore below.
The empirical Bayesian approach

Suppose that we have a model $X|\theta \sim f(\cdot|\theta)$ with a hierarchical prior $p(\theta|\phi)$ where $\phi$ is a hyperparameter. Then conditional on $\phi$, we have from Bayes theorem that
\[
p(\theta|x) \propto f(x|\theta)p(\theta|\phi).
\]
Now suppose that we do not wish to specify a hyperprior distribution for $\phi$. Then, the empirical Bayes (EB) approach is to use the data to estimate $\phi$ (via e.g. maximum likelihood, the method of moments or some alternative approach). Then the analysis proceeds as if $\phi$ were known so that given an estimate, $\hat{\phi}$, of the hyperparameter, then we approximate the posterior distribution of $\theta$ by
\[
p(\theta|x) \propto f(x|\theta)p(\theta|\hat{\phi}).
\]
Multivariate normal example

Example 76
Suppose that

\[ X_i | \theta_i \sim \mathcal{N}(\theta_i, 1) \]
\[ \theta_i | \tau \sim \mathcal{N}(0, \tau^2) \]

for \( i = 1, \ldots, n \).

Then, it is easy to see that a posteriori, \( \theta_i | x_i, \tau \sim \mathcal{N}((1 - B)x_i, 1 - B) \) where \( B = \frac{1}{1 + \tau^2} \) with posterior mean \( E[\theta_i | x_i, \tau] = (1 - B)x_i \).

When \( \tau \) is unknown, this posterior mean estimate is unsatisfactory as it depends on \( \tau \). One possibility is thus to estimate \( \tau \) from the data.
In order to do this, note first that

\[ \mathbf{X} | \tau \sim \mathcal{N} \left( \mathbf{0}, (1 + \tau^2) \mathbf{I}_n \right) \]

and consider \( S = \sum_{i=1}^{n} X_i^2 \). From a classical viewpoint, we have \( S \frac{1}{B} \chi_n^2 \) so that if we define

\[ \hat{B} = \frac{n - 2}{S} \]

it is easy to show that \( E[\hat{B}] = B \) so that \( \hat{B} \) is an unbiased estimator of \( B \).

Substituting \( \hat{B} \) for \( B \), we now proceed as if \( B \) were known. Therefore, we have

\[ \theta_i | \mathbf{x}, (B = \hat{B}) \sim \mathcal{N} \left( \left( 1 - \hat{B} \right) x_i, \left( 1 - \hat{B} \right) \right) \]

and therefore

\[ E[\theta_i | \mathbf{x}] = \left( 1 - \frac{n - 2}{\sum_{i=1}^{n} x_i^2} \right) x_i \]

is an EB estimate of the posterior mean.
Stein’s paradox and the James Stein estimator

From a classical viewpoint, as $X \sim \mathcal{N}(\theta, I_n)$ it would appear that a natural estimator of $\theta$ would be $X$ itself.

However Stein (1955) showed that $X$ is not even an *admissable* estimator of $\theta$ when $n \geq 3$. James and Stein (1960) showed that an estimator which *dominates* $X$ is

$$
\left(1 - \frac{n - 2}{\sum_{i=1}^{n} X_i^2}\right)X
$$

which is exactly the EB estimator we have just derived.

For a fuller discussion of Stein’s paradox, see Yung (1999).
Example 77

The batting average of a baseball player is the number of hits $S_i$ divided by the number of times at bat. Supposing that $n$ players have each gone out to bat $k$ times, then (assuming exchangeability of hits for each player) the batting average of the $i$'th player is $S_i/k \sim B(k, p_i)$ where $p_i$ is the probability that they hit the ball on a given time at bat.

Then, using an arc sin transformation,

$$X_i = 2\sqrt{k} \sin^{-1} \sqrt{\frac{S_i}{k}}$$

$$\theta_i = 2\sqrt{k} \sin^{-1} \sqrt{p_i}$$

we have that approximately, $X_i|\theta_i \sim N(\theta_i, 1)$. Now assuming exchangeability of players, it is reasonable to assume that $\theta_i|\mu, \sigma \sim N(\mu, \sigma^2)$. 

Bayesian Statistics
The empirical Bayes approach is now to estimate $\mu$ and $\sigma$ from the data. Noting that the marginal distribution of $X_i$ given the hyperparameters is $X_i|\mu, \sigma \sim \mathcal{N}(\mu, 1 + \sigma^2)$ then,

$$E[\bar{X}|\mu, \sigma] = \mu \quad E\left[\frac{(n-3)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \frac{1}{1 + \sigma^2}$$

so we have method of moment estimates $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-3} - 1$.

Now, given $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, we have $\theta_i|x \sim \mathcal{N}\left((1 + \frac{1}{\hat{\sigma}^2})^{-1}\left(x_i + \frac{1}{\hat{\sigma}^2}\bar{x}\right), (1 + \frac{1}{\hat{\sigma}^2})^{-1}\right)$.

Thus, the EB estimator for $\theta_i$ takes the form $\hat{\theta}^{EB}_i = (1 + \frac{1}{\hat{\sigma}^2})^{-1}\left(x_i + \frac{1}{\hat{\sigma}^2}\bar{x}\right)$ and inverting of the arc sin law transformation, we have $\hat{p}^{EB}_i = \sin^2\left(\frac{\hat{\theta}^{EB}_i}{2\sqrt{n}}\right)$.

Efron and Morris (1975) analyzed the batting averages of $n = 18$ baseball players over their first 45 at bats and over the remainder of the season. A table of the data follow.
<table>
<thead>
<tr>
<th>Name</th>
<th>( \hat{p}_i(1st\ 45) )</th>
<th>( p_i(\text{Rest}) )</th>
<th>( \hat{p}_i^{EB} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clemente</td>
<td>0.400</td>
<td>0.346</td>
<td>0.290</td>
</tr>
<tr>
<td>F.Robinson</td>
<td>0.378</td>
<td>0.298</td>
<td>0.286</td>
</tr>
<tr>
<td>F.Howard</td>
<td>0.356</td>
<td>0.276</td>
<td>0.282</td>
</tr>
<tr>
<td>Johnstone</td>
<td>0.333</td>
<td>0.222</td>
<td>0.277</td>
</tr>
<tr>
<td>Berry</td>
<td>0.311</td>
<td>0.273</td>
<td>0.273</td>
</tr>
<tr>
<td>Spencer</td>
<td>0.311</td>
<td>0.270</td>
<td>0.273</td>
</tr>
<tr>
<td>Kessinger</td>
<td>0.289</td>
<td>0.263</td>
<td>0.268</td>
</tr>
<tr>
<td>Alvarado</td>
<td>0.267</td>
<td>0.210</td>
<td>0.264</td>
</tr>
<tr>
<td>Santo</td>
<td>0.244</td>
<td>0.269</td>
<td>0.259</td>
</tr>
<tr>
<td>Swoboda</td>
<td>0.244</td>
<td>0.230</td>
<td>0.259</td>
</tr>
<tr>
<td>Unser</td>
<td>0.222</td>
<td>0.264</td>
<td>0.254</td>
</tr>
<tr>
<td>Williams</td>
<td>0.222</td>
<td>0.256</td>
<td>0.254</td>
</tr>
<tr>
<td>Scott</td>
<td>0.222</td>
<td>0.303</td>
<td>0.254</td>
</tr>
<tr>
<td>Petrocelli</td>
<td>0.222</td>
<td>0.264</td>
<td>0.254</td>
</tr>
<tr>
<td>Rodriguez</td>
<td>0.222</td>
<td>0.226</td>
<td>0.254</td>
</tr>
<tr>
<td>Campaneris</td>
<td>0.200</td>
<td>0.285</td>
<td>0.249</td>
</tr>
<tr>
<td>Munson</td>
<td>0.178</td>
<td>0.316</td>
<td>0.244</td>
</tr>
<tr>
<td>Alvis</td>
<td>0.156</td>
<td>0.200</td>
<td>0.239</td>
</tr>
</tbody>
</table>

It can be seen immediately that the EB estimates correspond more closely to the true batting averages over the season than do the raw estimates.
Criticisms and characteristics of the empirical Bayes approach

In general, the EB approach leads to compromise (shrinkage) posterior estimators between the individual \((X_i)\) and group \((\bar{X})\) estimators.

One problem with the EB approach is that it clearly ignores the uncertainty in \(\phi\). Another problem with this approach is how to choose the estimators of the hyperparameters. Many options are possible, e.g. maximum likelihood, method of moments, unbiased estimators etc. and all will lead to slightly different solutions in general.

This is in contrast to the fully Bayesian approach which requires the definition of a hyperprior but avoids the necessity of selecting a given value of the hyperprior.
Fully hierarchical modeling

In order to implement a fully hierarchical model, we need to specify a hyperprior distribution \( p(\phi) \). Then, we have

\[
p(\theta|x) \propto \int f(x|\theta)p(\theta|\phi)p(\phi) \, d\phi
\]

\[
p(\phi|x) \propto \int f(x|\theta)p(\theta|\phi)p(\phi) \, d\theta.
\]

In many cases, these integrals cannot be evaluated analytically. However, often a Gibbs sampling approach can be implemented by sampling from the conditional posterior distributions

\[
p(\theta|x, \phi) \propto f(x|\theta)p(\theta|\phi)
\]

\[
p(\phi|x, \theta) = p(\phi|\theta) \propto p(\theta|\phi)p(\phi)
\]

which often do have conjugate forms.
Example 78
Consider Example 74 and suppose initially that the values of $\phi$ and $\psi$ are known and that we use a uniform distribution for $\mu$. Then:

$$p(\mu, \theta|y) \propto \exp \left( -\frac{\phi}{2} \sum_i (y_i - \theta_i)^2 - \frac{\psi}{2} \sum_i (\theta_i - \mu)^2 \right).$$

Integrating with respect to $\mu$, we have $\theta|y \sim \mathcal{N}(m, W)$ where

$$W^{-1} = \frac{1}{\phi + \psi} I + \frac{\psi}{n\phi(\phi + \psi)} J \quad \text{and} \quad Wm = \phi y$$

and the posterior mean of $\theta$ is given by

$$E[\theta|y] = \frac{\phi}{\phi + \psi} y + \frac{\psi}{\psi + \phi} \bar{y}$$

which is a weighted average of the MLE and the global mean.
Suppose now that both $\phi$ and $\psi$ are unknown and that we use the usual improper priors $p(\phi) \propto \frac{1}{\phi}$ and $p(\psi) \propto \frac{1}{\psi}$. Then it is easy to show that

$$\theta | y, \mu, \phi, \psi \sim \mathcal{N} \left( \frac{\phi y + \psi \mu}{\phi + \psi}, \frac{1}{\phi + \psi} \mathbf{I} \right)$$

$$\mu | y, \theta, \phi, \psi \sim \mathcal{N} \left( \bar{\theta}, \frac{1}{n\psi} \right)$$

$$\phi | y, \theta, \mu, \psi \sim \mathcal{G} \left( \frac{n}{2}, \frac{\sum_i (y_i - \theta_i)^2}{2} \right)$$

$$\psi | y, \theta, \mu, \phi \sim \mathcal{G} \left( \frac{n}{2}, \frac{\sum_i (\mu - \theta_i)^2}{2} \right)$$

and a Gibbs sampling algorithm could be set up. However, it is possible to demonstrate that the joint posterior distribution is *improper*.

It is important to check the propriety of the posterior distribution when improper hyperprior distributions are used. An alternative (as in for example *Winbugs*) is to use proper but high variance hyperprior distributions.
Directed acyclic graph representations

Hierarchical models are often well represented by directed acyclic graphs or DAGs as used in Winbugs.

**Example 79**

A DAG representing the model and prior structure of Example 75 is as below.
Here we can see that $X_i$ depends directly upon its rate $\lambda_i$ which depends on $t_i$ and $\theta_i$ through a logical relation ($\lambda_i = t_i \theta_i$).

In Winbugs, this can be converted into code for a Gibbs sampler.

```winbugs
model {
  for (i in 1 : N) {
    theta[i] ~ dgamma(alpha, beta)
    lambda[i] <- theta[i] * t[i]
    x[i] ~ dpois(lambda[i])
  }
  alpha ~ dexp(1)
  beta ~ dgamma(0.1, 1.0)
}
```
George et al (1993) give the following table of pump failure data.

<table>
<thead>
<tr>
<th>Pump</th>
<th>$t_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>15.7</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>62.9</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>126</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>5.24</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>31.4</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>1.05</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1.05</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2.1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>10.5</td>
<td>22</td>
</tr>
</tbody>
</table>

George et al (1993) assume a hierarchical prior distribution

\[
\alpha \sim \mathcal{E}(1) \\
\beta \sim \mathcal{G}(0.1, 1)
\]
The following table gives the posterior means and variances of the different parameters estimated via Winbugs.

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>0.7001</td>
<td>0.2699</td>
<td>0.004706</td>
<td>0.2851</td>
<td>0.6634</td>
<td>1.338</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>beta</td>
<td>0.929</td>
<td>0.5325</td>
<td>0.00978</td>
<td>0.1938</td>
<td>0.8315</td>
<td>2.205</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[1]</td>
<td>0.0598</td>
<td>0.02542</td>
<td>2.68E-4</td>
<td>0.02128</td>
<td>0.05627</td>
<td>0.1195</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[2]</td>
<td>0.1008</td>
<td>0.07855</td>
<td>8.177E-4</td>
<td>0.00838</td>
<td>0.08181</td>
<td>0.3023</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[3]</td>
<td>0.08927</td>
<td>0.03759</td>
<td>3.702E-4</td>
<td>0.0316</td>
<td>0.08469</td>
<td>0.1762</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[4]</td>
<td>0.116</td>
<td>0.03048</td>
<td>3.17E-4</td>
<td>0.06363</td>
<td>0.1132</td>
<td>0.1825</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[5]</td>
<td>0.6056</td>
<td>0.315</td>
<td>0.003087</td>
<td>0.1529</td>
<td>0.6529</td>
<td>1.359</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[6]</td>
<td>0.6105</td>
<td>0.1393</td>
<td>0.0014</td>
<td>0.3688</td>
<td>0.5996</td>
<td>0.9096</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[7]</td>
<td>0.9025</td>
<td>0.7252</td>
<td>0.007937</td>
<td>0.07569</td>
<td>0.7167</td>
<td>2.751</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[8]</td>
<td>0.8964</td>
<td>0.725</td>
<td>0.008262</td>
<td>0.07614</td>
<td>0.7096</td>
<td>2.785</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[9]</td>
<td>1.68</td>
<td>0.7767</td>
<td>0.009004</td>
<td>0.4828</td>
<td>1.452</td>
<td>3.452</td>
<td>1001</td>
<td>10000</td>
</tr>
<tr>
<td>theta[10]</td>
<td>1.983</td>
<td>0.4251</td>
<td>0.004915</td>
<td>1.264</td>
<td>1.968</td>
<td>2.916</td>
<td>1001</td>
<td>10000</td>
</tr>
</tbody>
</table>
The last diagram shows a kernel density estimate of the posterior density of \( \theta_1 \).