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

- Bernardo, J.M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, **4**, 111–121. Available from

  http://www.uv.es/~bernardo/Exchangeability.pdf

- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, **39**, 83–87.

- Yung, K.H. (1999). Explaining the Stein paradox. Available from

  http://www.cs.toronto.edu/~roweis/csc2515/readings/stein_paradox.pdf

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

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

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 their 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 θ.
- θ 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

$$f(Y_n = y_n) = \int_0^1 \left(\begin{array}{c} n \\ y_n \end{array} \right) \theta^{y_n} (1-\theta)^{n-y_n} P(\theta) d\theta.$$

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

$$F(x_1, \ldots, x_n) = \int_{\mathcal{G}} \prod_{i=1}^{n} G(x_i) dP(G)$$

where $\mathcal{G}$ 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 $\boldsymbol{\theta}$ such that

$$F(\mathbf{x}) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{n} F(x_i|\boldsymbol{\theta}) dP(\boldsymbol{\theta}).$$

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 $\mathbf{X}_n$ is the same as the distribution of $\mathbf{A}\mathbf{X}_n$ for any orthogonal matrix $\mathbf{A}$, 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 $\mathbf{x}$ with density $f(\mathbf{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Often we may make extra assumptions about the structural relationships between the elements of $\boldsymbol{\theta}$.

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

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

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 θ_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 μ, representing the mean true IQ in the population, and ψ.

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

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

where θ_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 α and β are the unknown hyperparameters.

Fitting hierarchical models

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

Thus, two main approaches have developed:

- The natural Bayesian approach is to use relatively uninformative prior distributions for the hyperparameters ϕ 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

Robbins

This approach is originally due to Robbins (1955). For a full review see

http://en.wikipedia.org/wiki/Empirical_Bayes_method

Suppose that we have a model $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ with a hierarchical prior $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ where $\boldsymbol{\phi}$ is a hyperparameter. Then conditional on $\boldsymbol{\phi}$, we have from Bayes theorem that

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi}).$$

Now suppose that we do not wish to specify a hyperprior distribution for $\boldsymbol{\phi}$. Then, the *empirical Bayes* (EB) approach is to use the data to estimate $\boldsymbol{\phi}$ (via e.g. maximum likelihood, the method of moments or some alternative approach). Then the analysis proceeds as if $\boldsymbol{\phi}$ were known so that given an estimate, $\hat{\boldsymbol{\phi}}$, of the hyperparameter, then we approximate the posterior distribution of $\boldsymbol{\theta}$ by

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\hat{\boldsymbol{\phi}}).$$

Multivariate normal example

Example 76

Suppose that

$$\begin{aligned} X_i|\theta_i &\sim \mathcal{N}(\theta_i, 1) \\ \theta_i|\tau &\sim \mathcal{N}\left(0, \tau^2\right) \end{aligned}$$

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 τ is unknown, this posterior mean estimate is unsatisfactory as it depends on τ. One possibility is thus to estimate τ from the data.

In order to do this, note first that

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

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 $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ it would appear that a natural estimator of $\boldsymbol{\theta}$ would be $\mathbf{X}$ itself.

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

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

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 $S_i/k \sim \mathcal{B}(k, p_i)$ where p_i is the probability that they hit the ball on a given time at bat.

Then, using an arc sin transformation,

$$\begin{aligned} X_i &= 2\sqrt{k}\sin^{-1}\sqrt{\frac{S_i}{k}} \\ \theta_i &= 2\sqrt{k}\sin^{-1}\sqrt{p_i} \end{aligned}$$

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

The empirical Bayes approach is now to estimate μ and σ from the data. Noting that the marginal distribution of X_i given the hyperparameters is $X_i|\mu,\sigma \sim \mathcal{N}\left(\mu, 1+\sigma^2\right)$ 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 $\theta_i|\mathbf{x} \sim \mathcal{N}\left(\left(1+\frac{1}{\hat{\sigma}^2}\right)^{-1}\left(x_i+\frac{1}{\hat{\sigma}^2}\bar{x}\right), \left(1+\frac{1}{\hat{\sigma}^2}\right)^{-1}\right)$.

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

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.

Name	$\hat{p}_i$(1st 45)	p_i(Rest)	$\hat{p}_i^{EB}$
Clemente	0.400	0.346	0.290
F.Robinson	0.378	0.298	0.286
F.Howard	0.356	0.276	0.282
Johnstone	0.333	0.222	0.277
Berry	0.311	0.273	0.273
Spencer	0.311	0.270	0.273
Kessinger	0.289	0.263	0.268
Alvarado	0.267	0.210	0.264
Santo	0.244	0.269	0.259
Swoboda	0.244	0.230	0.259
Unser	0.222	0.264	0.254
Williams	0.222	0.256	0.254
Scott	0.222	0.303	0.254
Petrocelli	0.222	0.264	0.254
Rodriguez	0.222	0.226	0.254
Campaneris	0.200	0.285	0.249
Munson	0.178	0.316	0.244
Alvis	0.156	0.200	0.239

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 ϕ. 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 $p(\boldsymbol{\phi})$. Then, we have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})\, d\phi \\ p(\boldsymbol{\phi}|\mathbf{x}) &\propto \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})\, d\boldsymbol{\theta}. \end{aligned}$$

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

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) &\propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi}) \\ p(\boldsymbol{\phi}|\mathbf{x}, \boldsymbol{\theta}) &= p(\boldsymbol{\phi}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}) \end{aligned}$$

which often do have conjugate forms.

Example 78

Consider Example 74 and suppose initially that the values of ϕ and ψ are known and that we use a uniform distribution for μ. Then:

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

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

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

and the posterior mean of $\boldsymbol{\theta}$ is given by

$$E[\boldsymbol{\theta}|\mathbf{y}] = \frac{\phi}{\phi + \psi}\mathbf{y} + \frac{\psi}{\psi + \phi}\mathbf{1}\bar{y}$$

which is a weighted average of the MLE and the global mean.

Suppose now that both ϕ and ψ 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

$$\begin{aligned}
\boldsymbol{\theta}|\mathbf{y}, \mu, \phi, \psi &\sim \mathcal{N}\left(\frac{\phi\mathbf{y} + \psi\mu\mathbf{1}}{\phi + \psi}, \frac{1}{\phi + \psi}\mathbf{I}\right) \\
\mu|\mathbf{y}, \boldsymbol{\theta}, \phi, \psi &\sim \mathcal{N}\left(\bar{\theta}, \frac{1}{n\psi}\right) \\
\phi|\mathbf{y}, \boldsymbol{\theta}, \mu, \psi &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_i (y_i - \theta_i)^2}{2}\right) \\
\psi|\mathbf{y}, \boldsymbol{\theta}, \mu, \phi &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_i (\mu - \theta_i)^2}{2}\right)
\end{aligned}$$

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

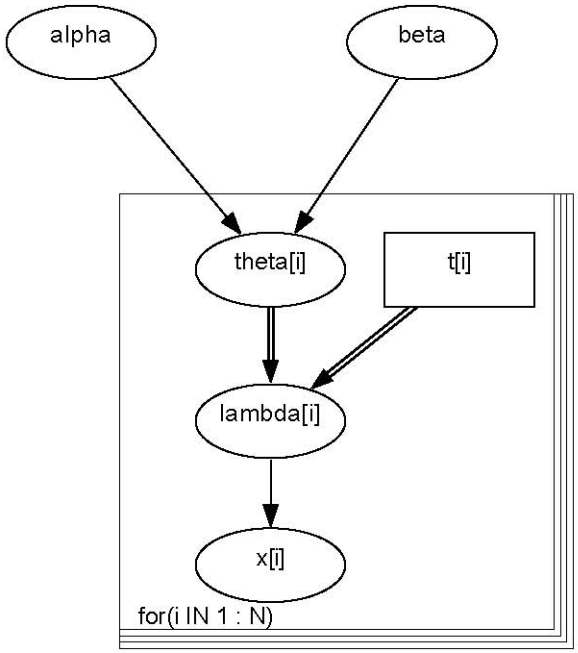
It is important to check the propriety of the posterior distribution when improper hyperprior distributions are used. An alternative (as in for example `Winbugs`) is to use proper but high variance hyperprior distributions.

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 λ_i which depends on t_i and θ_i through a logical relation ($\lambda_i = t_i\theta_i$).

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

```
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.

Pump	t_i	x_i
1	94.5	5
2	15.7	1
3	62.9	5
4	126	14
5	5.24	3
6	31.4	19
7	1.05	1
8	1.05	1
9	2.1	4
10	10.5	22

George et al (1993) assume a hierarchical prior distribution

$$\begin{aligned} \alpha &\sim \mathcal{E}(1) \\ \beta &\sim \mathcal{G}(0.1, 1) \end{aligned}$$

The following table gives the posterior means and variances of the different parameters estimated via `Winbugs`

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	0.7001	0.2699	0.004706	0.2851	0.6634	1.338	1001	10000
beta	0.929	0.5325	0.00978	0.1938	0.8315	2.205	1001	10000
theta[1]	0.0598	0.02542	2.68E-4	0.02128	0.05627	0.1195	1001	10000
theta[2]	0.1008	0.07855	8.177E-4	0.00838	0.08181	0.3023	1001	10000
theta[3]	0.08927	0.03759	3.702E-4	0.0316	0.08469	0.1762	1001	10000
theta[4]	0.116	0.03048	3.17E-4	0.06363	0.1132	0.1825	1001	10000
theta[5]	0.6056	0.315	0.003087	0.1529	0.5529	1.359	1001	10000
theta[6]	0.6105	0.1393	0.0014	0.3668	0.5996	0.9096	1001	10000
theta[7]	0.9025	0.7252	0.007937	0.07559	0.7167	2.751	1001	10000
theta[8]	0.8964	0.725	0.008262	0.07614	0.7098	2.785	1001	10000
theta[9]	1.59	0.7767	0.009004	0.4828	1.452	3.452	1001	10000
theta[10]	1.993	0.4251	0.004915	1.264	1.958	2.916	1001	10000

The last diagram shows a kernel density estimate of the posterior density of θ_1.