CHAPTER 3

BAYES, MINIMAX AND EMPIRICAL BAYES
ESTIMATION OF PARAMETERS IN MIGRATION
PROBLEM WITH COMPUTATIONAL METHODS

3.1 Introduction and summary

Bayes and Empirical Bayes

In recent years, a progressive new trend has been growing in applied
statistics: It is becoming even more popular to build application-specific
models that are designed to account for the hierarchical and latent struc­
tures inherent in any particular data generation mechanism. Such multilevel
models have long been advocated on theoretical grounds, but the develop­
ment of methodological and computational tools for statistical analysis has
now begun to bring such model fitting into routine practice.

Empirical Bayes (EB) estimation is a type of Bayesian estimation involves
estimating parameter $p$ of the distribution (which is multinomial in this case)
without knowing or assessing the parameters of the prior (i.e., parameter of
the dirichlet prior in this case) involves in the distribution of $p$. The term em­
pirical Bayes was coined by Robbins (1955). In the most frequent adopted
version of empirical Bayes estimation, the structure is used with a preas­
signed family of prior distributions to obtain the Bayes estimator. But the
parameters of the prior distribution are estimated from the current dataset
by using the marginal distribution of the data, given the hyperparameters. They are estimated by maximum likelihood method or method of moments. The data-based estimator of the hyperparameters are then substituted for the hyperparameters in the prior distribution, as well as in the Bayesian estimator, to obtain an approximate Bayesian estimator, an empirical Bayesian estimator.

Here empirical Bayes estimator assuming dirichlet prior is carried out by maximum likelihood method, and this estimation involves iterative algorithm.

**EM Algorithm**

It is to be noted that the dirichlet prior is a conjugate prior, in this case the expectation-maximisation (EM) algorithm is an important technique for finding maximum likelihood estimates from multinomial data. Each iteration of EM consists of an E-step (expectation step) and an M step (maximisation step). For detail exposition on EM algorithm one can look at McLachlan and Krishnan (1997).

The EM algorithm is a broadly applicable algorithm that provides an iterative procedure of computing MLE's in situations where, useful in incomplete data problems. Hence in this context, the observed data vector \( n \) is viewed as being incomplete and as regarded as an observable function of the so-called complete data. The notion of 'incomplete data' includes the conventional sense of missing data, but it also applies to situations where the complete data represent what would be available from some hypothetical experiment. In the latter case the complete data may contain some variables that are never observable in a data sense. Within this framework, let \( x \) de-
note the vector containing the augmented or so-called complete data, and let \( \mathbf{p} \) denote the vector containing the additional data, referred to as the unobservable or missing data.

As it is evident when a problem does not at first appear to be an incomplete-data one, computation of MLE is often greatly facilitated by artificially formulating it to be as such. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data.

In Section 3.2 Bayes estimation of the parameters under Dirichlet prior is given. Section 3.3 includes Bayes estimation where the states or places seem to be equal (i.e., they seem to be equally important).

There is a way to find out minimax estimator through Bayes estimation e.g., Equaliser Bayes is Minimax. This way, in Section 3.4 Minimax estimator through Bayes procedure is obtained. Also it is compared with MLE.

In Section 3.5 the empirical Bayes procedures carried out for the two cases considered in Section 3.2 and Section 3.3. It is to be remembered that estimation of hyperparameters is difficult as these involve complicated marginal. That is why in literature several computational methods like Markov Chain Monte Carlo (MCMC) and others are used. Here Quasi-Newton accelerated EM algorithm is used. Method is fitted accordingly to these cases. Following such algorithm one can obtain the empirical Bayes estimates.

In Section 3.5.1 computational method for EB is given when hyperparameters are different and Section 3.5.2 computational method for EB is given where hyperparameters are same for \( k \)-places. Section 3.5.3 is meant to study the robustness of EB numerically.

We develope the relevant programs for evaluating the estimates by using
S-Plus platform and those are attached in a section Chapter 6 which is pre
meant for numerical works.

3.2 Bayes Estimation under Conjugate Prior

In this sub-section we would like to evaluate the Bayes’ estimator for the
multinomial parameter \( p_i = (p_{i1}, p_{i2}, \ldots, p_{ik}) \), for all \( i = 1, 2, \ldots, k \) under
conjugate prior distribution. Bayes estimator under Dirichlet prior, since it
is conjugate to the p.m.f. of \( n_i \), for all \( i = 1, 2, \ldots, k \). Bayes estimator is
easy to find out. W3e know if there are sufficiently large sample then such
assumption on prior is not inappropriate.

**Result:** The Bayes estimator of \( p_{ij} \)'s under Dirichlet prior with constant
vector \( \alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}) \), for all \( i = 1, 2, \ldots, k \) is
\( \delta(n_{ij}) = \frac{n_{ij} + \alpha_{ij}}{n + \alpha_i} \), for all \( i, j = 1, 2, \ldots, k \).

**Proof.** A natural conjugate prior for the parameter \( p_i \), for all \( i = 1, 2, \ldots, k \) is given by the Dirichlet distribution density with parameter \( \alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik}) \), for all \( i = 1, 2, \ldots, k \). Thus the joint prior density is given by

\[
\pi(p) = \prod_{i=1}^{k} \pi(p_i) = \prod_{i=1}^{k} \left( \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right) \prod_{j=1}^{k} \Gamma(\alpha_{ij}) \prod_{j=1}^{k} p_{ij}^{\alpha_{ij}-1} }{\prod_{j=1}^{k} \Gamma(\alpha_{ij}) \prod_{j=1}^{k} p_{ij}^{\alpha_{ij}-1} } \right)
\]

The joint distribution of \( n = (n_1^T, n_2^T, \ldots, n_k^T)^T \) and \( p = (p_1^T, p_2^T, \ldots, p_k^T)^T \) is
given by

\[
f(n|\mathbf{p}) \pi(\mathbf{p}) = \prod_{i=1}^{k} f(n_i|p_i) \pi(p_i) = \prod_{i=1}^{k} \left( \frac{n_i!}{\prod_{j=1}^{k} n_{ij}! \prod_{j=1}^{k} \Gamma(\alpha_{ij})} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} p_{ij}^{n_{ij}+\alpha_{ij}-1} \right)
\]

The marginal pmf of \( n \) is

\[
g(n) = \int_{S_1} \cdots \int_{S_k} f(n|\mathbf{p}) \pi(\mathbf{p}) \prod_{i=1}^{k} \prod_{j=1}^{k} dp_{ij}
\]

where the integration is carried out over the region

\[
S_i = \left\{ p_{ij} : p_{ij} \geq 0 \text{ and } \sum_{j=1}^{k} p_{ij} = 1 \right\}, \text{ for all } i = 1, 2, \ldots, k
\]

After integration we have the marginal pmf of \( n \) is

\[
g(n) = \prod_{i=1}^{k} \left( \frac{n_i!}{\prod_{j=1}^{k} n_{ij}! \prod_{j=1}^{k} \Gamma(\alpha_{ij})} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} \frac{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)}{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})} \right)
\]

which is the product of \( k \) independent Multinomial-Dirichlet distribution.

Then the posterior distribution of \( \mathbf{p} \) is given by

\[
\pi(\mathbf{p}|n) = \frac{f(n|\mathbf{p}) \pi(\mathbf{p})}{g(n)} = \prod_{i=1}^{k} \frac{f(n_i|p_i) \pi(p_i)}{g(n_i)}
\]

\[
= \prod_{i=1}^{k} \left( \frac{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)}{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})} \frac{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)}{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})} p_{ij}^{n_{ij}+\alpha_{ij}-1} \right)
\]
which is the product of $k$ independent Dirichlet distribution with parameter $\alpha_i', i = 1, 2, \cdots, k$

where $\alpha_i' = (\alpha_{i1}', \cdots, \alpha_{ik}')^T = (n_{i1} + \alpha_{i1}, \cdots, n_{ik} + \alpha_{ik})^T$, for all $i = 1, 2, \cdots, k$.

Now we are interested to find out the posterior mean of $p_{ij}$'s

Let in particular $r$ be the specified value of $i$, then from (3.2.1) we have,

$$\pi(p_r | n_r) = \frac{\Gamma \left( \sum_{j=1}^{k} (n_{rj} + \alpha_{rj}) \right)}{\prod_{j=1}^{k} \Gamma(n_{rj} + \alpha_{rj})} \frac{1}{\prod_{j=1}^{k} p_{rj}^{n_{rj} + \alpha_{rj} - 1}}$$

Let $j_o$ be the specified value of $j$, then

$$E(p_{rj_o} | n_r) = \frac{\Gamma \left( \sum_{j=1}^{k} (n_{rj} + \alpha_{rj}) \right) \prod_{j=1 \atop j \neq j_o}^{k} p_{rj_o}^{n_{rj_o} + \alpha_{rj_o} - 1}}{\prod_{j=1}^{k} \Gamma(n_{rj} + \alpha_{rj})} \prod_{j=1}^{k} \prod_{j \neq j_o}^{k} dp_{rj}$$

$$= \frac{\Gamma \left( \sum_{j=1}^{k} (n_{rj} + \alpha_{rj}) \right) \prod_{j=1 \atop j \neq j_o}^{k} \Gamma(n_{rj_o} + \alpha_{rj_o} + 1) \prod_{j=1 \atop j \neq j_o}^{k} \Gamma(n_{rj} + \alpha_{rj})}{\prod_{j=1}^{k} \Gamma(n_{rj} + \alpha_{rj})} \frac{\Gamma \left( \sum_{j=1}^{k} (n_{rj} + \alpha_{rj}) + 1 \right)}{\prod_{j=1 \atop j \neq j_o}^{k} (\sum_{j=1 \atop j \neq j_o}^{k} \alpha_{rj} + n_{rj})}$$

$$= \frac{\alpha_{rj_o} + n_{rj_o}}{\sum_{j=1}^{k} \alpha_{rj} + n_r}$$

Since $r$ and $j_o$ are specified value of $i$ and $j$ respectfully, hence we have,

$$E(p_{ij}) = \frac{\alpha_{ij} + n_{ij}}{\sum_{j=1}^{k} \alpha_{ij} + n_i}, \text{ for all } i, j = 1, 2, \cdots, k$$
Therefore the Bayes estimator of $p_{ij}$ for sum of squared error loss is

$$\delta(n_{ij}) = \frac{\alpha_{ij} + n_{ij}}{\sum_{j=1}^{k} \alpha_{ij} + n_{i}}$$

where

$$\alpha_{i} = \sum_{j=1}^{k} \alpha_{ij}, \text{ for all } i = 1, 2, \cdots, k$$

also note that

$$\sum_{j=1}^{k} \delta(n_{ij}) = 1$$

3.3 Bayes Estimation where the States Seem to be Equal

In this sub-section we would like to evaluate the Bayes estimator under the constraint that all the states are equally important i.e., $p_{11} = p_{22} = \cdots = p_{kk} = p$, when the value of $p$ is unknown.

We have parameters $p = (p_{11}, p_{12}, \cdots, p_{1k}; p_{21}, p_{22}, \cdots, p_{2k}; \cdots; p_{k1}, p_{k2}, \cdots, p_{kk})$ and our interest is in $(p_{11}, p_{22}, \cdots, p_{kk})$, then others are nuisance parameter.

Let us denote $p_{-i} = (p_{i1}, \cdots, p_{i,i-1}, p_{i,i+1}, \cdots, p_{ik})$, for all $i = 1, 2, \cdots, k$ and also $p_{-} = (p_{T-1}, \cdots, p_{T-k})^T$.

Then we have under the above the joint p.m.f. of $n_{ij}$'s, for all $i, j = 1, 2, \ldots, k$ is

$$f(n|p, p_{-}) = \prod_{i=1}^{k} f(n_{i}|p, p_{-i}) = \prod_{i=1}^{k} \left( \frac{n_{i_{1}}! \cdots n_{i_{k}}!}{n_{i_{1}}! n_{i_{2}}! \cdots n_{i_{k}}!} p_{i_{1}}^{n_{i_{1}}} p_{i_{2}}^{n_{i_{2}}} \cdots p_{i_{i-1}}^{n_{i_{i-1}}} p_{i_{i+1}}^{n_{i_{i+1}}} p_{i_{k}}^{n_{i_{k}}} \right)$$

$$= \prod_{i=1}^{k} \left( \frac{n_{i_{i}}!}{n_{i_{1}}! n_{i_{2}}! \cdots n_{i_{k}}!} \right) \times \sum_{p} \prod_{i=1}^{k} \prod_{j \neq i}^{k} p_{ij}^{n_{ij}} \quad (3.3.1)$$
Under \( p_{11} = p_{22} = \cdots = p_{kk} = p(\text{unspecified}) \), the prior distribution can be taken as

\[
\pi(p, p_-) = \prod_{i=1}^{k} \left[ \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_{i1})\Gamma(\alpha_{i2}) \cdots \Gamma(\alpha_{ik})} \times p_{i1}^{\alpha_{i1} - 1} \cdots p_{i,i-1}^{\alpha_{i,i-1} - 1} p_{i,i}^{\alpha_{i,i} - 1} p_{i,i+1}^{\alpha_{i,i+1} - 1} \cdots p_{ik}^{\alpha_{ik} - 1} \right]
\]

where \( \alpha_i = \sum_{j=1}^{k} \alpha_{ij} \), for all \( i = 1, \ldots, k \)

\[
= \prod_{i=1}^{k} \left[ \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_{i1})\Gamma(\alpha_{i2}) \cdots \Gamma(\alpha_{ik})} \right] \sum_{p_{i1} = 1}^{k} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{\alpha_{ij} - 1} (3.3.2)
\]

**Theorem 3:** Bayes’ estimator of \( p_{ij} \)'s when \( p_{11} = p_{22} = \cdots = p_{kk} = p(\text{unspecified}) \) is

\[
\hat{\theta}(n_{ij}) = \frac{\sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - (k - 1)}{n + \alpha - 2(k - 1)} \quad \text{; for all } i = 1, 2, \ldots, k
\]

and

\[
\hat{\theta}(n_{ij}) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) - (k - 1)}{\sum_{j=1}^{k} \sum_{j \neq i} (n_{ij} + \alpha_{ij})} \quad \text{; for all } i, j(j \neq i) = 1, 2, \ldots, k.
\]

**Proof.**

We have from (3.3.1) and (3.3.2) the joint distribution of \( n_{ij} \)'s and \( p, p_{ij} \)'s

\( i \neq j \) is

\[
f(n|p, p_-) = \prod_{i=1}^{k} \left[ \frac{n_{ii}!}{n_{i1}! \cdots n_{ik}!} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_{i1}) \cdots \Gamma(\alpha_{ik})} \right] \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{n_{ij} + \alpha_{ij} - 1}.
\]
Then the posterior distribution is of the form

\[
\pi(p, p_i | n) = \frac{1}{C(n, \alpha)} \times p^{\sum_{i=1}^{k} (n_{ii} + \alpha_{ii})} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{n_{ij} + \alpha_{ij} - 1}
\]  

(3.3.3)

where,

\[
C(n, \alpha) = \int_{0}^{1} \left( \int_{S_i} \cdots \int_{S_k} p^{\sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - k} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{n_{ij} + \alpha_{ij} - 1} \prod_{i=1}^{k} \prod_{j=1}^{k} d(p_{ij}) \right) dp
\]

where the integration is carried out over the region

\[
S_i = \left\{ p_{ij} : p_{ij} \geq 0, p \geq 0, \text{ and } \sum_{j=1}^{k} p_{ij} = 1 - p \right\} ; \text{ for all } i = 1, 2, \cdots, k
\]

Now, for a particular \(i\),

\[
\int_{S_i} p_{ii}^{n_{ii} + \alpha_{ii} - 1} \cdots p_{ii+1}^{n_{ii+1} + \alpha_{ii+1} - 1} \cdots p_{ik}^{n_{ik} + \alpha_{ik} - 1} \prod_{j=1}^{k} d(p_{ij})
\]

\[
= \prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij}) \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) - 1
\]

\[
= \frac{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})}{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)} (1 - p)^{j \neq i}
\]

(3.3.4)

for all \(i = 1, 2, \cdots, k\)
So after integrating out $p_{ij}$'s ($i \neq j$) from (3.3.3) we have,

\[
C(n, \alpha) = \frac{\prod_{i=1}^{k} \prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})}{\prod_{i=1}^{k} \Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)} \int_0^1 \frac{\prod_{i=1}^{k} \Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) - (k - 1) \right)}{\prod_{i=1}^{k} \Gamma \left( n_{ii} + \alpha_{ii} \right)} \left( 1 - p \right)^{(k-1)} \left( 1 - p \right)^{-(k-1)} \ dp
\]

\[
= \frac{\prod_{i=1}^{k} \prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij}) \Gamma \left( \sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - (k - 1) \right) \Gamma \left( \sum_{i=1}^{k} \sum_{j \neq i}^{k} (n_{ij} + \alpha_{ij}) - (k - 1) \right)}{\prod_{i=1}^{k} \Gamma \left( n_{ii} + \alpha_{ii} \right) \Gamma(n + \alpha - 2(k - 1))} \Gamma(n + \alpha - 2(k - 1))
\]

where $n = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}$ and $\alpha = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_{ij}$

Thus we get,

\[
\frac{1}{C(n, \alpha)} = \frac{\prod_{i=1}^{k} \Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) - (k - 1) \right) \Gamma \left( \sum_{i=1}^{k} \sum_{j \neq i}^{k} (n_{ij} + \alpha_{ij}) - (k - 1) \right)}{\prod_{i=1}^{k} \prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})}
\]

(3.3.5)

From (3.3.3) we see that the marginal distribution of $p$ follows Beta-distribution with parameters $\sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - (k - 1)$ and $\sum_{i=1}^{k} \sum_{j \neq i}^{k} (n_{ij} + \alpha_{ij}) - (k - 1)$
Therefore we have,

\[
E(p) = \frac{\sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - (k - 1)}{n + \alpha - 2(k - 1)}
\]

Thus we get restricted Bayes estimator of \( p_{ii} \) is

\[
\hat{\delta}(n_{ii}) = \frac{\sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - (k - 1)}{n + \alpha - 2(k - 1)}; \text{ for all } i = 1, 2, \ldots, k
\]

Now we want to find out the posterior mean of \( p_{ij} \)'s for \( i \neq j \)

In particular,

\[
E(p_{12}) = \frac{1}{C(n, \alpha)} \int_{0}^{1} p \sum_{i=1}^{k} (n_{ii} + \alpha_{ii}) - k \left( \int_{S_1}^{S_k} p_{12}^n_{12} + \alpha_{12} - 1 \cdots p_{1k}^n_{1k} + \alpha_{1k} - 1 \prod_{j=2}^{k} dp_{1j} \right) \times \left( \int_{S_{k+1}}^{S_{k+2}} p_{21}^n_{21} + \alpha_{21} - 1 \cdots p_{2k}^n_{2k} + \alpha_{2k} - 1 \prod_{j=1, j \neq 2}^{k} dp_{2j} \right) \times \cdots
\]

\[
\cdots \times \left( \int_{S_{k-j+1}}^{S_{k-j+2}} p_{k1}^n_{k1} + \alpha_{k1} - 1 \cdots p_{k,k-1}^n_{k,k-1} + \alpha_{k,k-1} - 1 \prod_{j=1, j \neq k}^{k} dp_{kj} \right) dp \tag{3.3.6}
\]

Now,

\[
\int_{S_{k-j+1}}^{S_{k-j+2}} p_{12}^n_{12} + \alpha_{12} - 1 \cdots p_{1k}^n_{1k} + \alpha_{1k} - 1 \prod_{j=2}^{k} dp_{1j}
\]

\[
\Gamma(n_{12} + \alpha_{12} + 1) \prod_{j=3}^{k} \Gamma(n_{1j} + \alpha_{1j}) \sum_{j=2}^{k} (n_{1j} + \alpha_{1j})
\]

\[
= \frac{\Gamma\left( \sum_{j=2}^{k} (n_{1j} + \alpha_{1j}) + 1 \right)}{(1 - p)^{k-2}} \tag{3.3.7}
\]

51
Thus we have, from (3.3.6) using (3.3.7) and (3.3.4) for $i = 2, 3, \ldots, k$

$$E(p_{12}) = C'(n, \alpha) \int_0^1 \sum_{p_i=1}^k (n_{ii} + \alpha_{ii}) - k \sum_{i=1}^k \sum_{j=1, j \neq i}^k (n_{ij} + \alpha_{ij}) - (k - 1) dp$$

where

$$C'(n, \alpha) = \frac{1}{C(n, \alpha)} \frac{\Gamma(n_{12} + \alpha_{12} + 1) \prod_{j=3}^k \Gamma(n_{1j} + \alpha_{1j}) \prod_{i=2}^k \Gamma(n_{ij} + \alpha_{ij})}{\Gamma\left(\sum_{j=2}^k (n_{ii} + \alpha_{ii}) + 1\right) \prod_{i=2}^k \Gamma\left(\sum_{j=1, j \neq i}^k (n_{ij} + \alpha_{ij})\right)}$$

$$= \frac{n_{12} + \alpha_{12}}{\sum_{j=2}^k (n_{1j} + \alpha_{1j}) \Gamma\left(\sum_{i=1}^k (n_{ii} + \alpha_{ii}) - (k - 1)\right) \Gamma\left(\sum_{i=1}^k \sum_{j=1, j \neq i}^k (n_{ij} + \alpha_{ij}) - (k - 1)\right)}$$

[using (3.3.5)]

(3.3.8)

Therefore

$$E(p_{12}) = C'(n, \alpha) \frac{\sum_{i=1}^k \sum_{j=1, j \neq i}^k (n_{ij} + \alpha_{ij}) - (k - 1)}{\Gamma(n + \alpha - 2(k - 1))}$$

$$= \frac{n_{12} + \alpha_{12}}{\sum_{j=2}^k (n_{1j} + \alpha_{1j})} \frac{\sum_{i=1}^k \sum_{j=1, j \neq i}^k (n_{ij} + \alpha_{ij}) - (k - 1)}{n + \alpha - 2(k - 1)}$$

[using (3.3.8)]

52
Therefore we get, the restricted Bayes estimator of \( p_{ij}, (j \neq i) \) is
\[
\hat{\delta}(n_{ij}) = \frac{n_{ij} + \alpha_{ij} - \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) - (k - 1)}{\frac{\alpha_{ij} + n_{ij}}{n + 2(k - 1)}}; \text{for all } i, j(j \neq i) = 1, 2, \ldots, k.
\]
that completes the proof.

**3.4 Minimax Estimator**

In this sub-section, we develop minimax estimator for \( p_{ij} \)'s, for all \( i, j \) with respect to sum of squared error loss. Here we use the result that a Bayes estimator with constant risk is minimax. Now the risk function for \( i \)th State is
\[
E_{p_{ij}} \left[ \sum_{j=1}^{k} \{ \delta(n_{ij}) - p_{ij} \}^2 \right] = \sum_{j=1}^{k} \text{Bias}(\delta(n_{ij}))^2 + \sum_{j=1}^{k} \text{Var}(\delta(n_{ij}))
\]
\[
= \sum_{j=1}^{k} \left[ p_{ij} - \frac{1}{\alpha_i + n_i} \right]^2 + \frac{n_{ij} p_{ij} (1 - p_{ij})}{\left[ \frac{1}{\alpha_i + n_i} \right]^2} + \sum_{j=1}^{k} \left[ \frac{n_{ij} p_{ij} (1 - p_{ij})}{\left[ \frac{1}{\alpha_i + n_i} \right]^2} \right]
\]
\[
= \sum_{j=1}^{k} (p_{ij} \alpha_i - \alpha_{ij})^2 + \sum_{j=1}^{k} n_{ij} p_{ij} (1 - p_{ij})
\]
\[
= \frac{(\alpha_i - n_i) \sum_{j=1}^{k} p_{ij}^2 + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} (n_{ij} - 2 \alpha_i \alpha_{ij}) p_{ij}}{\left( \alpha_i + n_i \right)^2}
\]
53
\[
\begin{align*}
&= \frac{1}{(\alpha_i + n_i)^2} \left[ (\alpha_i^2 - n_i) \left( \sum_{j=1}^{k-1} p_{ij}^2 + (1 - \sum_{j=1}^{k-1} p_{ij})^2 \right) + \sum_{j=1}^{k} \alpha_{ij}^2 \\
&\quad + \sum_{j=1}^{k-1} ((n_i - 2\alpha_i \alpha_{ij}) p_{ij} + (n_i - 2\alpha_i \alpha_{ik})(1 - \sum_{j=1}^{k-1} p_{ij}) \right]
\end{align*}
\]

\[
= \frac{1}{(\alpha_i + n_i)^2} \left[ (\alpha_i^2 - n_i) \left( \sum_{j=1}^{k-1} p_{ij}^2 + (1 - \sum_{j=1}^{k-1} p_{ij})^2 \right) + \sum_{j=1}^{k} \alpha_{ij}^2 \\
&\quad + \sum_{j=1}^{k-1} ((n_i - 2\alpha_i \alpha_{ij}) - (n_i - 2\alpha_i \alpha_{ik}) p_{ij} + (n_i - 2\alpha_i \alpha_{ik}) \right] \tag{3.4.1}
\]

For this risk function to be constant over all \( p_{ij} \), we need coefficient of \( p_{ij} \) and that of \( \alpha_{ij} \) is zero. Which implies

\[
\alpha_i^2 - n_i = 0, \text{ for all } i = 1, 2, \ldots, k \text{, and,}
\]

\[
(n_i - 2\alpha_i \alpha_{ij}) - (n_i - 2\alpha_i \alpha_{ik}) = 0, \text{ for all } j = 1, 2, \ldots, k - 1.
\]

Which gives,

\[
\alpha_i = \sqrt{n_i}, \text{ for all } i = 1, 2, \ldots, k \text{, and,}
\]

\[
\alpha_{i1} = \alpha_{i1} = \cdots = \alpha_{ik} = \frac{1}{k} \sqrt{n_i}, \text{ for all } i = 1, 2, \ldots, k.
\]

So that minimax estimator of \( p_{ij} \), using sum of squared error loss, is

\[
\delta^*(n_{ij}) = \frac{n_{ij} + \frac{1}{k} \sqrt{n_i}}{n_i + \sqrt{n_i}}, \text{ for all } i, j = 1, 2, \ldots, k
\]

From (3.4.1) we have the minimax risk

\[
r_{\delta^*} = \frac{k-1}{k} \frac{1}{(1 + \sqrt{n_i})^2}, \text{ for all } i = 1, 2, \ldots, k
\]

54
and risk of ML estimator

\[
r_{\hat{p}_i} = E \left[ \sum_{j=1}^{k} (p_{ij} - \hat{p}_{ij})^2 \right] = \sum_{j=1}^{k} \frac{p_{ij}(1 - p_{ij})}{n_i} = \frac{1 - \sum_{i=1}^{k} p_{ij}^2}{n_i}, \text{ for all } i = 1, 2, \ldots, k
\]

This shows that minimax risk is less than the maximum possible risk of m.l.e, for all \( i = 1, 2, \ldots, k \). Which holds when in particular

\[p_{11} = p_{22} = \cdots = p_{ik} = \frac{1}{k}, \text{ for all } i = 1, 2, \ldots, k\]

Then risk of MLE

\[
r_{\hat{p}_i} = \frac{(k - 1) \frac{1}{k}}{n_i}, \text{ for all } i = 1, 2, \ldots, k.
\]

Which is larger than minimax risk. Thus minimax estimator is better than maximum likelihood estimator in this particular case also.

### 3.5 Empirical Bayes Estimation for the cases mentioned in Section 3.2 and Section 3.3

In this section we want to find out the estimator of the unknown parameters involved in the prior distribution.

It can be seen that the Bayes estimator \( \delta(n_{ij}) \) of \( p_{ij} \) depends on prior distribution \( \pi(p|\alpha) \). When \( \alpha_{ij} \)'s are unknown, it is not possible to implement \( \delta(\cdot) \). However, according to the model described previously, the \( n_i \), for \( i = 1, 2, \ldots, k \), are independent. The empirical Bayes approach is employed to combine information from observations \( n_i \), \( i = 1, 2, \ldots, k \).

If the prior distribution of \( p \) is not completely known the posterior distribution can not be used to make probability statements concerning the
unknown variables. But in some situations it may be possible to compute the posterior distribution of \( p \) on \( n \) without the complete knowledge of the distribution function of \( p \). But the estimate may involve unknown parameters which are characteristics of the marginal distribution of \( n \) alone. If past data on \( n \) are available, the marginal distribution of \( n \) can be estimated empirically.

### Case 1: Priors have different hyperparameters

Now let \( N \) be the random vector corresponding to the observed data \( n \) having Multinomial distribution with parameter \( p \) (unknown), where

\[
 n_i \sim \text{Multinomial}(n_{i1}, n_{i2}, \ldots, n_{ik}), \quad \text{for all } i = 1, 2, \ldots, k
\]

such that

\[
 n_i = \sum_{j=1}^{k} n_{ij}, \quad \sum_{j=1}^{k} p_{ij} = 1, \quad \text{for all } i = 1, 2, \ldots, k
\]

and \( p_i \) have Dirichlet prior distribution i.e.,

\[
p_i \sim \text{Dirichlet}(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}), \quad \text{for all } i = 1, 2, \ldots, k
\]

Then the Bayes estimator of \( p_{ij} \) under squared error loss is

\[
 \delta(n_{ij}) = \frac{n_{ij} + \alpha_{ij}}{n_i + \alpha_i}, \quad \text{for all } i, j = 1, 2, \ldots, k
\]

### 3.5.1 Computational Method when hyperparameters are different

Consider the case where \( \alpha_{ij} \)'s are unknown parameter, for \( i, j = 1, 2, \ldots, k \).

We take

\[
 \alpha = (\alpha_1^T, \alpha_2^T, \ldots, \alpha_k^T)^T = (\alpha_{11}, \ldots, \alpha_{1k}; \alpha_{21}, \ldots, \alpha_{2k}; \ldots; \alpha_{k1}, \ldots, \alpha_{kk})^T
\]

56
We are interested to find out the estimate of the Dirichlet parameter $\alpha$ from observed data $n$. But it is not possible to estimate $\alpha$ from observed data analytically. In this case it may be possible to compute iteratively the MLE of $\alpha$ from observed data by using Quasi-Newton accelerated EM algorithm.

We estimate $\alpha_i$ individually, since estimates of $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})$ depends only on $n_i = (n_{i1}, \ldots, n_{ik})$ and independent of $n_j = (n_{j1}, \ldots, n_{jk})$, for $(j \neq i)$.

Thus we iterate each $\alpha_i$ individually and take the limiting value of $\alpha_i$ as $\alpha_i^*$, for all $i = 1, 2, \ldots, k$.

Finally we have

$$\alpha^* = (\alpha_1^*, \ldots, \alpha_k^*)^T$$

By $x$ we denote the vector containing the augmented or so called complete data, and by $p$ we denote the vector containing the additional data, referred to as the unobservable or missing data. Then the complete data vector $x$ taken to be

$$x = (n^T, p^T)^T$$

The missing variables $(p_{i1}, p_{i2}, \ldots, p_{ik})$ are defined so that

$$f(n_{i1}, n_{i2}, \ldots, n_{ik} | p_{i1}, p_{i2}, \ldots, p_{ik}) = \frac{n_{i!}^{n_{i1}} p_{i1}^{n_{i2}} p_{i2}^{n_{i3}} \cdots p_{ik}^{n_{iik}}}{n_{i1}! n_{i2}! \cdots n_{ik}!} \quad (3.5.1)$$

for all $i = 1, 2, \ldots, k$ and

$$\pi(p_{i1}, p_{i2}, \ldots, p_{ik}) = \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} p_{i1}^{\alpha_{i1}-1} p_{i2}^{\alpha_{i2}-1} \cdots p_{ik}^{\alpha_{ik}-1}, \text{ for all } i = 1, 2, \ldots, k \quad (3.5.2)$$
After integrating out \((p_{i1}, p_{i2}, \cdots, p_{ik})\) from joint density function of \((n_{i1}, n_{i2}, \cdots, n_{ik})\) and \((p_{i1}, p_{i2}, \cdots, p_{ik})\) that can be formed from (3.5.1) and (3.5.2), the density function of \((n_{i1}, n_{i2}, \cdots, n_{ik})\), for all \(i = 1, 2, \cdots, k\) is given by

\[
\begin{align*}
 f(n_{i1}, n_{i2}, \cdots, n_{ik} | \alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik}) = & \frac{n_{i!}}{\prod_{j=1}^{k} n_{ij!}} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} \int_{S_i} \prod_{j=1}^{k} p_{ij}^{n_{ij} + \alpha_{ij} - 1} \prod_{j=1}^{k} dp_{ij} \\
& \text{where the integration is carried out over the region} \\
& S_i = \left\{ p_{ij} : p_{ij} \geq 0 \text{ and } \sum_{j=1}^{k} p_{ij} = 1 \right\}, \text{ for all } i = 1, 2, \cdots, k \\
\end{align*}
\]

Thus we get marginal density of \((n_{i1}, n_{i2}, \cdots, n_{ik})\), for all \(i = 1, 2, \cdots, k\) is

\[
\begin{align*}
 f(n_{i1}, n_{i2}, \cdots, n_{ik} | \alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik}) = & \frac{n_{i!}}{\prod_{j=1}^{k} n_{ij!}} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} \frac{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})}{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)} \\
& \text{which is Multinomial-Dirichlet distribution.} \\
\end{align*}
\]

By \((n_{i1}, n_{i2}, \cdots, n_{ik})\) we denote an observed random sample from Multinomial-Dirichlet distribution with parameters \((\alpha_{i1}, \cdots, \alpha_{ik})\) for all \(i = 1, 2, \cdots, k\). Then for incomplete-data log likelihood is of the form [ after taking logarithm of (3.5.3) and summing over all \(i = 1, 2, \cdots, k\) ]

\[
\log L(\alpha) = \sum_{i=1}^{k} \log L(\alpha_i) = \sum_{i=1}^{k} \log(n_{i!}) - \sum_{i=1}^{k} \sum_{j=1}^{k} \log(n_{ij}!) + \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(\alpha_i) - \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(\alpha_{ij})
\]

58
\[ + \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(n_{ij} + \alpha_{ij}) - \sum_{i=1}^{k} \log \Gamma(n_{i} + \alpha_{i}) \]

(3.5.4)

where \( n_{i} = \sum_{j=1}^{k} n_{ij} \), and \( \alpha_{i} = \sum_{j=1}^{k} \alpha_{ij} \), for all \( i = 1, 2, \ldots, k \).

Because of the conditional structure of the complete-data model specified by (3.5.1) \& (3.5.2), the complete-data likelihood can be factored into product of the conditional density of \((n_{i1}, n_{i2}, \ldots, n_{ik})\) given \((p_{i1}, p_{i2}, \ldots, p_{ik})\) and the marginal density of \((p_{i1}, p_{i2}, \ldots, p_{ik})\) for all \( i = 1, 2, \ldots, k \). Accordingly, the complete-data log likelihood be written as

\[
\log L_{c}(\alpha) = \sum_{i=1}^{k} \log L_{c}(\alpha_{i}) = \sum_{i=1}^{k} \log(n_{i}) - \sum_{i=1}^{k} \sum_{j=1}^{k} \log(n_{ij}) + \sum_{i=1}^{k} \log \Gamma(\alpha_{i}) - \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(\alpha_{ij})
\]

\[ + \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + \alpha_{ij} - 1) \log p_{ij} \]

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation (3.5.4) indirectly by proceeding iteratively in terms of the complete-data log likelihood function \( \log L_{c}(\alpha) \). As it is unobservable, this obstacle is overcome by averaging the complete-data likelihood over its conditional distribution given the observed data \( n \). But in order to calculate this conditional expectation, we have to specify a value for \( \alpha \).

Let \( \alpha_{i}^{(0)} \) denotes the starting value of \( \alpha_{i} \) and \( \alpha_{i}^{(r)} \) the value of \( \alpha_{i} \) on the \( r \)th subsequent iteration of the EM algorithm.

Then on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of \( \log L_{c}(\alpha) \) given \( n \), using \( \alpha_{i}^{(0)} \) for \( \alpha_{i} \), which can be written as

\[
Q(\alpha_{i}, \alpha_{i}^{(0)}) = E_{\alpha_{i}^{(0)}}[\log L_{c}(\alpha_{i})|n_{i}] \]
where $Q$-function is used to denote the conditional expectation of the complete-data log likelihood function, $\log L_c(\alpha)$, given the observed data $n$, using the current fit for $\alpha$. We have on the $(r+1)$th iteration of the E-step

$$Q(\alpha_i, \alpha^{(r)}_i) = E_{\alpha^{(r)}_i}\{\log L_c(\alpha_i)|n_i\}$$

where the expectation operator $E$ has the subscript $\alpha^{(r)}_i$ to explicitly convey that this (conditional) expectation is being affected using $\alpha^{(r)}_i$ for $\alpha_i$. Therefore $Q(\alpha_i, \alpha^{(r)}_i)$ can be written as

$$Q(\alpha_i, \alpha^{(r)}_i) = \log(n_{i1}) - \sum_{j=1}^{k} \log(n_{ij})! + \log \Gamma(\alpha_i) - \sum_{j=1}^{k} \log \Gamma(\alpha_{ij})$$

$$+ \sum_{j=1}^{k} (n_{ij} + \alpha_{ij} - 1) E_{\alpha^{(r)}_i}\{\log p_{ij}|n_{i1}, \cdots, n_{ik}\} \quad (3.5.5)$$

It is seen from (3.5.5) that, in order to carry out M-step, we need to calculate the term

$$E_{\alpha^{(r)}_i}\{\log p_{ij}|n_{i1}, \cdots, n_{ik}\}$$

The calculation of the above term can be avoided if one make uses of the identity [ Lange (1995b) ],

$$S_i(n_i; \alpha^{(r)}_i) = [\delta Q(\alpha_i, \alpha^{(r)}_i)/\delta \alpha_i]_{\alpha_i=\alpha^{(r)}_i}, \text{ for all } i = 1, 2, \cdots, k$$

where $S_i(n_i; \alpha^{(r)}_i)$ is the score statistics given by

$$S_i(n_i; \alpha^{(r)}_i) = \delta \log L(\alpha^{(r)}_i)/\delta \alpha^{(r)}_i, \text{ for all } i = 1, 2, \cdots, k$$

On evaluating $S_{ij}(n_i; \alpha^{(r)}_i)$, the derivative of (3.5.4) with respect to $\alpha_{ij}$ at the point $\alpha_i = \alpha^{(r)}_i$, we have that
\[
S_{ij}(n_i; \alpha_i^{(r)}) = \delta \log L(\alpha_i^{(r)}/\delta \alpha_{ij}^{(r)} = \delta \log \Gamma(\alpha_i^{(r)})/\delta \alpha_{ij}^{(r)} - \delta \log \Gamma(n_i + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)}
\]
\[
+ \delta \log \Gamma(n_{ij} + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)} - \delta \log \Gamma(n_i + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)}, \text{ for all } i, j = 1, 2, \ldots, k
\]
(3.5.6)

On equating \( S_{ij}(n_i; \alpha_i^{(r)}) \) equal to the derivative of (3.5.5) with respect to \( \alpha_{ij} \) at the point \( \alpha_i = \alpha_i^{(r)} \), we obtain
\[
E_{\alpha_i^{(r)}} \log(p_{ij}|n_{i1}, \ldots, n_{ik}) = \delta \log \Gamma(n_{ij} + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)} - \delta \log \Gamma(n_i + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)}
\]
\[
= \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_i + \alpha_{ij}^{(r)})
\]
where
\[
\psi(s) = \delta \log \Gamma(s)/\delta s
\]
is the digamma function of \( s \).

Therefore from (3.5.5) we have
\[
Q(\alpha_i, \alpha_i^{(r)}) = \log(n_i) - \sum_{j=1}^k \log(n_{ij}) + \log \Gamma(\alpha_i) - \log \Gamma(\alpha_{ij})
\]
\[
+ \sum_{j=1}^k (n_{ij} + \alpha_{ij} - 1) \left[ \delta \log \Gamma(n_{ij} + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)} - \delta \log \Gamma(n_i + \alpha_{ij}^{(r)})/\delta \alpha_{ij}^{(r)} \right]
\]
(3.5.7)

**M-step:** On the M-step at the \((r + 1)\)th iteration of EM algorithm is to maximise \( Q(\alpha_i, \alpha_i^{(r)}) \), i.e., \( \alpha_{ij}^{(r+1)} \) obtained after solving
\[
\delta Q(\alpha_i, \alpha_i^{(r)})/\delta \alpha_{ij} = 0, \text{ for all } i = 1, \ldots, k
\]

it follows that \( \alpha_{ij}^{(r+1)} \) is a solution of the equation
\[
\psi(\alpha_i) - \psi(\alpha_{ij}) + \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_i + \alpha_{ij}^{(r)}) = 0
\]
The E-step and M-step are alternated repeatedly until the difference 
\( L(\alpha_i^{(r+1)}) - L(\alpha_i^{(r)}) \) changes by an arbitrary small amount in the case of 
convergence of the sequence of likelihood values \( \{L(\alpha_i^{(r)})\} \).

Concerning now the M-step, it can be seen that presence of the terms like 
\( \log \Gamma(\alpha_{ij}) \) in (3.5.7) prevents a closed-form solution for \( \alpha_i^{(r+1)} \).

The quasi-Newton acceleration procedure defines \( \alpha_i^{(r+1)} \) to be [Lange(1995b)]

\[
\alpha_i^{(r+1)} = \alpha_i^{(r)} + [I_c(\alpha_i^{(r)}; n_i) + B_i^{(r)}]^{-1} S_i(n_i; \alpha_i^{(r)}), \text{ for all } i = 1, \ldots, k \quad (3.5.8)
\]

The components of the score statistic \( S_i(n_i; \alpha_i^{(r)}) \) are available from (3.5.6),
and the information matrix \( I_c(\alpha_i^{(r)}; n_i) \) is given by

\[
I_c(\alpha_i^{(r)}; n_i) = -[\delta^2 Q(\alpha_i, \alpha_i^{(r)}) / \delta \alpha_i \delta \alpha_i^T]_{\alpha_i=\alpha_i^{(r)}} = D_i - C_i \mathbf{1} \mathbf{1}^T, \text{ for all } i = 1, 2, \ldots, k \quad (3.5.9)
\]

where each \( D_i \) is the diagonal matrix with \( j \)th diagonal entry

\[
d_{ij} = \psi'(\alpha_{ij}^{(r)}) - \psi'(n_{ij} + \alpha_{ij}^{(r)}), \text{ for all } j = 1, 2, \ldots, k
\]

\( C_i \) is the constant \( \psi'(\alpha_i^{(r)}) - \psi'(n_i + \alpha_i^{(r)}) \), and \( \mathbf{1} \) is a column vector of all 1's.

where \( \psi'(s) \) is the trigamma function \( \delta^2 \log \Gamma(s) / \delta s^2 \). Because the trigamma function is decreasing, \( d_{ij} > 0 \) when \( n_{ij} > 0 \). For the same reason, \( C_i > 0 \).

Since the presentation (3.5.9) is preserved under finite sums, it holds, in fact, for the entire sample.

The observed information matrix (3.5.9) is the sum of a diagonal matrix, 
which is trivial to invert, plus a symmetric, rank-one perturbation.

Thus the presence of the term \( B_i^{(r)} \) in (3.5.8) can be viewed as an attempt 
to approximate the Hessian of \( H(\alpha; \alpha_i^{(r)}) \) at the point \( \alpha = \alpha_i^{(r)} \).

62
Taking $B_i^{(r)}$ to be based on Davidon’s (1959) symmetric, rank-one update defined by

$$B_i^{(r)} = B_i^{(r-1)} + c_i^{(r)} u_i^{(r)} u_i^{(r)T}, \text{ for all } i = 1, \ldots, k$$

and where the constant $c_i^{(r)}$ and the vector $u_i^{(r)}$ are specified as

$$c_i^{(r)} = 1/(v_i^{(r)T} v_i^{(r)})$$

$$v_i^{(r)} = h_i^{(r)} + B_i^{(r-1)} e_i^{(r)}$$

here

$$e_i^{(r)} = \alpha_i^{(r)} - \alpha_i^{(r-1)}$$

is the error term, and

$$h_i^{(r)} = [\delta H(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i]_{\alpha_i=\alpha_i^{(r-1)}}$$

Taking $B_i^{(0)} = I_d$, which corresponds to initially performing an EM gradient step. When the matrix

$$I_c(\alpha_i^{(r)}; n_i) + B_i^{(r)}$$

fails to be positive definite, then replacing it by

$$I_c(\alpha_i^{(r)}; n_i) + (1/2)^m B_i^{(r)}$$

$m$ is the smallest positive integer such that the above matrix is positive definite.

In view of the identities in $\alpha$ and $\alpha^{(r)}$, namely that

$$\delta H(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i + \delta \log L(\alpha_i)/\delta \alpha_i = \delta Q(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i$$

63
and that

\[ [\delta H(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i]_{\alpha_i=\alpha_i^{(r)}} = 0 \]

we can express \( h_i^{(r)} \) as

\[
h_i^{(r)} = [\delta Q(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i]_{\alpha_i=\alpha_i^{(r-1)}} - \delta \log L(\alpha_i^{(r-1)})/\delta \alpha_i
\]

\[
= [\delta Q(\alpha_i; \alpha_i^{(r)})/\delta \alpha_i]_{\alpha_i=\alpha_i^{(r-1)}} - [\delta Q(\alpha_i; \alpha_i^{(r-1)})/\delta \alpha_i]_{\alpha_i=\alpha_i^{(r-1)}}
\]

\[
= [\psi(\alpha_i) - \psi(\alpha_{ij}) + \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_i + \alpha_i^{(r)})]_{\psi = \psi^{(r-1)}}
\]

\[
- [\psi(\alpha_i) - \psi(\alpha_{ij}) + \psi(n_{ij} + \alpha_{ij}^{(r-1)}) - \psi(n_i + \alpha_i^{(r-1)})]_{\alpha_i=\alpha_i^{(r-1)}}
\]

\[
= \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_i + \alpha_i^{(r)}) - \psi(n_{ij} + \alpha_{ij}^{(r-1)}) + \psi(n_i + \alpha_i^{(r-1)})
\]

Finally we get

\[ \alpha^* = (\alpha_1^* T, \alpha_2^* T, \ldots, \alpha_k^* T)^T \]

Where \( \alpha_i^* \) the limiting value of \( \alpha_i \), for all \( i = 1, 2, \ldots, k \).

Thus we have the empirical Bayes estimator of \( p_{ij} \) is

\[ \delta^*(n_{ij}) = \frac{n_{ij} + \alpha_{ij}^*}{n_i + \alpha_i^*}, \text{ for all } i, j = 1, 2, \ldots, k \]

where \( \alpha_{ij}^* \) is the limiting value of \( \alpha_{ij} \) for all \( i, j = 1, \ldots, k \), and

\[ \alpha_i^* = \sum_{j=1}^k \alpha_{ij}^*, \text{ for all } i = 1, \ldots, k \]
Case 2: Priors have same hyperparameters

Suppose

\[ n_i = (n_{i1}, n_{i2}, \ldots, n_{ik}) \sim \text{Multinomial}(n_i; p_{i1}, p_{i2}, \ldots, p_{ik}) \]

and

\[ p_i = (p_{i1}, p_{i2}, \ldots, p_{ik}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k), \text{ for all } i = 1, 2, \ldots, k \]

i.e., the \( k \) groups are tied together by the common prior distribution. But the Dirichlet parameters are unknown. We are interested to find out ML estimate of \((\alpha_1, \alpha_2, \ldots, \alpha_k)\) based on available observations \(n_i\), for all \( i = 1, 2, \ldots, k \).

Here \( n_i = \sum_{j=1}^{k} n_{ij} \), for all \( i = 1, 2, \ldots, k \).

Given \((n_{i1}, n_{i2}, \ldots, n_{ik})\), the maximum likelihood estimators of \((p_{i1}, p_{i2}, \ldots, p_{ik})\) are the relative frequencies [for details, J. Medhi (1987)]

\[ \hat{p}_{ij} = \frac{n_{ij}}{k} = \frac{n_{ij}}{n_i}, \text{ for all } i, j = 1, 2, \ldots, k \]

If we write

\[ y_{ij} = \frac{n_{ij}}{n_i}, \text{ for all } i, j = 1, 2, \ldots, k \text{ and } y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})^T, \text{ for all } i = 1, 2, \ldots, k \]

Then the vector \( y_i \) approximately follow Dirichlet distribution [Johnson and Kotz (1969)].

Without loss of generality let \( Y_1, Y_2, \ldots, Y_k \) be \( k \) sets of random vector from Dirichlet \((\alpha_1, \alpha_2, \ldots, \alpha_k)\).

Then the likelihood function is given by
\[ L(\alpha) = \left( \frac{\Gamma(\alpha)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \right)^k \prod_{i=1}^{k} (y_{i1}^{\alpha_1-1}y_{i2}^{\alpha_2-1} \cdots y_{ik}^{\alpha_k-1}) \]  

(3.5.10)

where \( \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)^T \) and \( \alpha = \sum_{j=1}^{k} \alpha_j \)

\[
\frac{\delta \log L(\alpha)}{\delta \alpha_j} = k\psi(\alpha) - k\psi(\alpha_j) + \sum_{i=1}^{k} \log y_{ij}
\]

where \( \psi(t) = \frac{\delta}{\delta t} \log \Gamma(t) \) is a digamma function.

Then the maximum-likelihood equations for estimators of \((\alpha_1, \alpha_2, \cdots, \alpha_k)\) respectively are given by [Johnson and Kotz (1994)].

\[
\psi(\hat{\alpha}_j) - \psi(\hat{\alpha}) = \frac{1}{k} \sum_{i=1}^{k} \log(y_{ij}), \text{ for all } j = 1, 2, \cdots, k
\]

(3.5.11)

as the likelihood equation, which clearly does not yield an explicit solution for \( \alpha_j \). Thus equations (3.5.11) must be solved by trial and error method.

### 3.5.2 Computational Method when hyperparameters are same for \( k \)-places

If we use the approximation

\[
\psi(t) \equiv \log(t - \frac{1}{2})
\]

then from (3.5.11) we have,

\[
\log(\alpha_j - \frac{1}{2}) - \log(\alpha - \frac{1}{2}) = \frac{1}{k} \sum_{i=1}^{k} \log(y_{ij}), \text{ for all } j = 1, 2, \cdots, k
\]

\[
\Rightarrow \log \left( \frac{\alpha_j - \frac{1}{2}}{\alpha - \frac{1}{2}} \right) = \log \left( \prod_{i=1}^{k} y_{ij}^{\frac{1}{k}} \right), \text{ for all } j = 1, 2, \cdots, k
\]
Thus the approximate values of \((\hat{\alpha}_j - \frac{1}{2})/(\hat{\alpha} - \frac{1}{2})\), for all \(j = 1, 2, \ldots, k\) is given by

\[
\frac{\alpha_j - \frac{1}{2}}{\alpha - \frac{1}{2}} = \prod_{i=1}^{k} y_{ij}^{\frac{1}{2}}, \text{ for all } j = 1, 2, \ldots, k \tag{3.5.12}
\]

summing over all \(j\) we have,

\[
\Rightarrow 1 - \frac{k-1}{\alpha - \frac{1}{2}} = \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{\frac{1}{2}}
\]

\[
\Rightarrow \alpha - \frac{1}{2} = \frac{k-1}{1 - \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{\frac{1}{2}}}
\]

Thus we have from (3.5.12), the estimator of \(\alpha_j\) which follow, as first approximation to \(\hat{\alpha}_j\) is given by

\[
\hat{\alpha}_j = \frac{1}{2} + \frac{k-1}{2} \prod_{i=1}^{k} y_{ij}^{\frac{1}{2}}, \text{ for all } j = 1, 2, \ldots, k \tag{3.5.13}
\]

Starting from these values of \(\hat{\alpha}_j\) from (3.5.13), solutions of (3.5.11) can be obtained by an iterative process.

Let us denote \(\alpha^{(r)}\) the value of \(\alpha\) at the \(r\)th subsequent iteration of Newton-Raphson method, then the value of \(\alpha\) for the next iteration is given by

\[
\alpha^{(r+1)} = \alpha^{(r)} + I(\alpha^{(r)})^{-1} S(\alpha^{(r)})
\]

67
where $I(\alpha) = -\frac{\partial^2}{\partial \alpha \partial \alpha'} \log L(\alpha)$ is the observed information matrix and $S(\alpha) = \frac{\partial}{\partial \alpha} \log L(\alpha)$ is the score vector.

From the likelihood function (3.5.10), we have the score has entries

$$\frac{\delta}{\delta \alpha_j} \log L(\alpha) = k \psi(\alpha_j) - k \psi(\alpha_j) + \sum_{i=1}^{k} \log y_{ij}$$

The observed information has entries

$$-\frac{\delta^2}{\delta \alpha_j \delta \alpha_{j'}} \log L(\alpha) = k \{ 1_{(j=j')} \psi'(\alpha_j) - \psi'(\alpha_j) \}$$

where $1_{(j=j')}$ is the indicator function of the event $\{j = j'\}$, and $\psi'(t)$ is the trigamma function $\frac{d^2}{dt^2} \log \Gamma(t)$.

The observed information can be summarised in matrix form by

$$-\frac{\delta^2}{\delta \alpha \delta \alpha'} \log L(\alpha) = I(\alpha) = k(D - c11^T)$$

where $D$ is a diagonal matrix with $j$th diagonal entry

$$d_j = \psi'(\alpha_j), \quad j = 1, 2, \ldots, k$$

c is the constant $\psi'(\alpha_j)$, and $1$ is a column vector of all 1's.

Thus the limiting value of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ as $\alpha^* = (\alpha^*_1, \alpha^*_2, \ldots, \alpha^*_k)$ gives the empirical Bayes estimate when all prior distributions have same set of parameters.

### 3.5.3 Robustness of Empirical Bayes Estimate

Empirical Bayes estimation, falls outside of the formal Bayesian paradigm. However it has been proven to be an effective technique of constructing estimators that perform well under both Bayesian and frequentist criteria. One
reason for this, as we see, is that empirical bayes estimators tend to be more robust against misspecification of the prior distribution.

Bayes risk performance of the empirical Bayes estimator is often robust; i.e., its Bayes risk is reasonably close to that of the Bayes estimator, no matter what values the hypeparameters attain.

We have Bayes risk for unbiased estimator \( \left( \frac{n_1}{n_i}, \frac{n_2}{n_i}, \cdots, \frac{n_k}{n_i} \right) \), for the ith state for all \( i = 1, 2, \cdots, k \)

\[
r(\pi, n_i|n_i) = E_{p_i} \left[ \sum_{j=1}^{k} \frac{p_{ij}(1 - p_{ij})}{n_i} \right] = \frac{1}{n_i} E_{p_i} \left[ \sum_{j=1}^{k} p_{ij}(1 - p_{ij}) \right]
\]

\[
= \frac{1}{n_i} E_{p_i} \left[ \sum_{j=1}^{k} p_{ij} \right] - \frac{1}{n_i} E_{p_i} \left[ \sum_{j=1}^{k} p_{ij}^2 \right] = \frac{1}{n_i} \left[ \sum_{j=1}^{k} \alpha_{ij} \right] - \frac{1}{n_i} \left[ \sum_{j=1}^{k} \alpha_{ij}(\alpha_{ij} + 1) \right]
\]

\[
= \frac{\alpha_i^2 - \sum_{j=1}^{k} \alpha_i^2}{n_i \alpha_i(\alpha_i + 1)}
\]

We have risk function for Bayes estimator for the ith state

\[
R(\delta, p_i) = E_{n_i|p_i} \left[ \sum_{j=1}^{k} \{ \delta(n_{ij}) - p_{ij} \}^2 \right] = \sum_{j=1}^{k} \text{Var}[\delta(n_{ij})] \]

\[
= \sum_{j=1}^{k} \left[ p_{ij} - E \left( \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right) \right]^2 + \sum_{j=1}^{k} \text{Var} \left[ \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right]
\]

\[
= \sum_{j=1}^{k} \left[ p_{ij} - \left( \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right) \right]^2 + \sum_{j=1}^{k} \left[ n_i p_{ij}(1 - p_{ij}) \right] \frac{(\alpha_i + n_i)^2}{(\alpha_i + n_i)^2}
\]

\[
= \frac{\sum_{j=1}^{k} [p_{ij}\alpha_i - \alpha_i] + \sum_{j=1}^{k} [n_i p_{ij}(1 - p_{ij})]}{(\alpha_i + n_i)^2}
\]


69
\[
\frac{(\alpha_i^2 - n_i) \sum_{j=1}^{k} p_{ij}^2 + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} (n_i - 2\alpha_i \alpha_{ij})p_{ij}}{(\alpha_i + n_i)^2}, \text{ for all } i = 1, 2, \ldots, k
\]

Bayes risk for Bayes estimator

\[
r(\pi, \delta) = E[R(\delta, p_i)] = E_{p_i}E_{n_i|p_i} \left[ \sum_{j=1}^{k} \{\delta(n_{ij}) - p_{ij}\}^2 \right]
\]

\[
= E_{p_i} \left[ \frac{(\alpha_i^2 - n_i) \sum_{j=1}^{k} p_{ij}^2 + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} (n_i - 2\alpha_i \alpha_{ij})p_{ij}}{(\alpha_i + n_i)^2} \right]
\]

\[
= \frac{1}{(\alpha_i + n_i)^2} \left[ (\alpha_i^2 - n_i) \sum_{j=1}^{k} \frac{\alpha_{ij}(\alpha_{ij} + 1)}{\alpha_i(\alpha_i + 1)} + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} (n_i - 2\alpha_i \alpha_{ij})\frac{\alpha_{ij}}{\alpha_i} \right]
\]

\[
= \frac{1}{(\alpha_i + n_i)^2} \left[ \left( \sum_{j=1}^{k} \alpha_{ij}^2 + \alpha_i \right) + \sum_{j=1}^{k} \alpha_{ij}^2 + n_i - 2\sum_{j=1}^{k} \alpha_{ij}^2 \right]
\]

\[
= \frac{\alpha_i^2 - \sum_{j=1}^{k} \alpha_{ij}^2}{(\alpha_i + n_i)\alpha_i(\alpha_i + 1)}, \text{ for all } i = 1, 2, \ldots, k
\]

Bayes risk for empirical Bayes estimator

\[
r(\hat{\pi}, \delta) = E_{\alpha_i=\hat{\alpha_i}}[R(\delta, p_i)] = \frac{\hat{\alpha}_i^2 - \sum_{j=1}^{k} \hat{\alpha}_{ij}^2}{(\hat{\alpha}_i + n_i)\hat{\alpha}_i(\hat{\alpha}_i + 1)}, \text{ for all } i = 1, 2, \ldots, k
\]

The Bayes risk of empirical Bayes estimator is only slightly higher than that of the Bayes estimator and is given in Table 1. For comparison, we also
Table 1: Bayes Risk of Bayes, empirical Bayes and unbiased estimator, for $k=3$, and sample size = 30

<table>
<thead>
<tr>
<th>Prior Parameters</th>
<th>Bayes Risk</th>
<th>EB (Δk)</th>
<th>Unbiased (n_{ij}/n_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(1,1,1)$</td>
<td>0.0151515</td>
<td>0.0157934</td>
<td>0.0166667</td>
</tr>
<tr>
<td>$(1,5,5)$</td>
<td>0.0129342</td>
<td>0.0131800</td>
<td>0.0176767</td>
</tr>
<tr>
<td>$(5,5,10)$</td>
<td>0.01190476</td>
<td>0.0119227</td>
<td>0.019841</td>
</tr>
<tr>
<td>$(1,5,1)$</td>
<td>0.0106178</td>
<td>0.0119566</td>
<td>0.013095</td>
</tr>
<tr>
<td>$(1,5,10)$</td>
<td>0.0103900</td>
<td>0.0106867</td>
<td>0.0159314</td>
</tr>
<tr>
<td>$(1,10,10)$</td>
<td>0.0101860</td>
<td>0.0103957</td>
<td>0.0173760</td>
</tr>
</tbody>
</table>

include the Bayes risk of the unbiased estimator $n_{ij}/n_i$, where $n_i = \sum_{j=1}^{k} n_{ij}$ for a particular $i$. The risk of the empirical Bayes estimator is between that of the Bayes estimator and that of unbiased estimator.

Program for Maximum Likelihood Estimation of Dirichlet parameters $\alpha_{ij}$'s by iterative procedure (Quasi-Newton accelerated EM algorithm) is given in Chapter 6, which includes different programs and numerical applications relevant to the methods used in this dissertation.

### 3.6 Concluding Remarks

In this chapter robustness (in Section 3.5.3) of EB in this case has been demonstrated, but this should be done elaborately. More importantly we
have worked out in Section 3.2, 3.3 and 3.5 for a class of priors. But from practical point of view if there are other guesses regarding transitions, then that kind of prior should be chosen by imposing conditions on hyperparameters of the prior model.

******************