# STATISTICAL INFERENCE VIA EMPIRICAL BAYES APPROACH FOR STATIONARY AND DYNAMIC CONTINGENCY TABLES

Farzad Eskandari

Supervisor
Professor Mohammad R. Meshkani

A dissertation submitted to the Department of Statistics, Shahid Beheshti University in partial fulfillment of requirements for the degree of Doctor of Philosophy in STATISTICS

March 2002

## ACKNOWLEDGEMENT

First of all, I would like to thank my supervisor professor Meshkani. It is a fact that I could not have done this research without his patience, advice and encouragement.

Further my thanks go to Professor Shafie who have been my consultant professor.

Also, I am thankful to Professor Parsian, Arghami, Zokaie and Shahlaie for carefully reading this dissertation and attending my defence session. Furthermore, I appreciate professor Noorbaloochi and Mr. Azadeh who have supported me during my studies period at this department.

Many thanks go to my wife, without her help, support, patience and encouragement I could not continue my studies. I also owe a special debt to my sons for my failure in fathering.

Finally, many thanks are due to my father, mother, brothers and sisters who their helps and encouragements have been very valuable for me

# Contents

# Statistical Inference via Empirical Bayes Approach for Stationary and Dynamic Contingency Tables

Farzad Eskandari
Supervisor: Mohammad R.Meshkani

Department of Statistics, Shahid Beheshti University
2002

## ABSTRACT

The literature on the theory of Markov chain models for a sequence of contingency tables began by Bush and Mosteller (1955) and Anderson (1957). The Bayesian model selection for Markov chain contingency tables and estimation of transition probabililies matrix were presented by Mashkani and Billard (1992). One of the main results presented in this dissertation is the Bayesian model selection method, using empirical Bayes approach for finite stationary Markov chains when in each state one has a log-linear model.

This dissertation consists of six chapters. The first chapter provides introductory material on Empirical Bayes Method (EBM) and it includes four case studies from political, acturial, medical and behavioral sciences as real ex-

amples where our results can be applied. While these exmples have been analyzed either by Bayesian or classical approaches, they could have been analyzed by empirical Bayes method, which enjoys both Bayesian strenght and frequentist as well, obhectivity. The second chapter gives a general overview and discussion of model selection techniques including stepwise and Bayesian techniques. The empirical Bayes analysis of log-linear and logistic models in stationary contingency tables are presented in Chapters 3 and 4. In Chapter 5, we present the model selection problem using the empirical Bayes method in dynamic contingency tables. In Chapter 6, three real-world examples illustrate the implementation of the empirical Bayes method developed in Chapters 4 and 5. Finally, we conclude by comparing a number of models via techniques presented in chapter 2 and 3.

# Chapter 1

# Historical Background

## 1.1 Introduction

One of the most important issues in statistical science is the construction of probabilistic models which represent, or sufficiently approximate, the true generating mechanism of a random phenomenon. The non-judicious choice of such models may possibly lead to misleading results concerning the description of the phenomenon under study. Model building, on the other hand, is the procedure that decides which probabilistic structure we should finally select as an appropriate model from a specified set of models. The identification of a statistical procedure for the selection of *good* models is still problematic even in the simple case of covariate selection, that is, selection of variables which influence a response variable $Y$ under study. Particularly, in variable selection procedures the most broadly employed methods are the stepwise procedures which consist of a sequential application of single significance test. The simplest arguement against the stepwise methods is that the exact distribution of the estimand parameters may not be known. Moreover, significance tests in some cases such as contingency tables cannot discriminate between non-nested models and therefore between models with different distributional structures. A variety of alternative criterias have been presented in the statistical literature. These criterias select the model which maximizes a quantity usually expressed as the log-likelihood minus a penalty function which depends on the dimension of the model. One of the most important methods is the likelihood ratio test. This alternative procedure, like a stepwise type method, may not trace the models which maximizes the criterion used since, in collinear cases between covariates, some *good* models will not be visited at all.

The goal of statistical inference, in general, is to extract and to report all available information about an unknown state of nature ( parameter of interest) $\theta$. Bayesian inference about $\theta$ is performed by combining the prior information( or uncertainty) about $\theta$ and the information provided by the experimental data via Baye's theorem. The statistician's prior uncertainty(before experimentation) about the parameter $\theta$ may be described by a probability density function of the parameter $\theta$, namely $\pi(\theta|\eta)$ which depends on some superparameter $\eta$. In order to learn more about $\theta$, the statistician observes the outcome of an experiment, $Y$ (a vector of observations or a single observation), which is related to $\theta$. This relationship is represented by a family of probability distributions of $Y$, conditional on $\theta$ *i.e.* $f(y|\theta)$. When regarded as a function of $\theta$, $f(y|\theta)$ is called the likelihood of the data $y$ given the parameter $\theta$. The likelihood is a function that describes the experiment and the relationship between $Y$ and $\theta$. The posterior distribution of $\theta$ given the data $y$, *i.e.* $\pi(\theta|y,\eta)$, is obtained from $\pi(\theta|\eta)$ and $f(y|\theta)$ according to Baye s theorem:

$$\pi(\theta|y,\eta) = \frac{f(y|\theta)\pi(\theta|\eta)}{\int_\Theta f(y|\theta)\pi(\theta|\eta)d\theta} = \frac{Likelihood \times Prior\ distribution}{Marginal\ ditribution\ of\ y} \quad (1-1)$$

The main idea in Bayesian inference is that the posterior ditribution contains all the available information about $\theta$. Consequently, any inference about $\theta$ should be deriven from this distribution and some Bayesians state that inference about $\theta$ is complete whenever the entire posterior $\pi(\theta|y)$ is reported.

The statistical models are a collection of joint probability distributions for the data $Y$, given $\theta$. The frequentist and Bayesian models are extreme cases in this paradigm.

In the words of Neyman (1977):

*"Frequentist models are extreme in the general framework because their prior families consist of all one−point distributions, that is, the parameter* $\theta$ *is a fixed but unknown constant".*

and according to Hill(1990) in Bayesian approach:

*"Bayesian models [as defined in* (1-1)] *form the other extreme because they each have only one element".*

Morris(1986) Commented that:

*"The practical statisticians encounter a variety of problems, and the frequency, objective Bayes and subjective Bayes methods provide a range of possible responses. There can be no clear victory for any approach for all applications, rather, we should train the statisticians for a frequency-Bayes compromise so that they can more flexibly respond to the new situations."*

Thus perhaps the best choice is the Empirical Bayes Methods. The adjective *empirical* in the empirical Bayes method arises from the fact that we are using the data to help determine the prior through the estimation of the superparameter $\eta$.

The empirical Bayes method does not belong to either of these extremes, but it does belong to the general paradigm. The Empirical Bayes Method(EBM) has a long, and sometimes philosophically confused past, a vibrant present, and an uncertain future. We consider these aspects in turn.

**past**

Somewhat ironically, the history of empirical Bayes method is not particularly Bayesian and certainly has little in common with the traditional, subjectivist Bayesian viewpoint. As noted above, *essentially there were attempts by frequentist decision theorists to use Bayesian tools to produce decision rules having good frequentist (not Bayesian) properties* (Robbins 1955). Stein(1955) showed that in the case where $y_i|\theta_i \sim N(\theta_i, \sigma^2)$, $i = 1, 2, \cdots, n$ and $\sigma^2$ is assumed known, the maximum likelihood estimator $\hat{\theta}_i(y) = y_i$ is inadmissible as an estimator of $\theta_i$. That is, under average squared error loss there must exist another estimator with frequentist risk no larger than $\sigma^2$ for every possible $\theta$ value. This dominating estimator was obtained by James and Stein(1961) as

$$\hat{\theta}_i^{JS}(\mathbf{y}) = \left(1 - \frac{(n-2)\sigma^2}{\| \mathbf{y} \|^2}\right) y_i.$$

The connection to EB was provided later in a celebrated series of papers by Efron and Morris(1971,1972,1973,1975,1977). Amongst many other things, these authors showed that $\hat{\theta}^{JS}$ is exactly the EB point estimator obtained under the assumption that $\theta_i|\tau^2 \sim N(0, \tau^2)$ where $\tau^2 = \left(\frac{1-B}{B}\right)\sigma^2$ and $B = \frac{\sigma^2}{\sigma^2+\tau^2}$ is estimated by $\hat{B} = (n-2)\sigma^2 / \| \mathbf{y} \|^2$.

**Present**

The cumulative impact of empirical Bayes methods on statistical applications continues to be enormous. Statisticians and users of statistics, many of whom were trained to distrust Bayesian methods as overly subjective and theoretically mysterious, can nonetheless often appreciate the value of borrowing strengh from similar but independent experiments. The empirical Bayes methods have, for example, enjoyed broad applications in the analysis of longitudinal, survival, and spatially correlated data (Laird and Ware,1982; Clayton and Kaldor, 1987). Efron (1996) develops an empirical Bayes (EB) approach to combining likelihoods for similar but independent parameters $\theta_i$. Meshkani and Billard(1992) and Billard and Meshkani(1995) using empirical Bayes method to estimate the transition probability matrix for Markov chains. Other important EB works include Rohanathan(1993) on spatial statistics and Altman and Casella(1995) on nonparametric growth curves.

**Future**

Returning to the theme of our opening sentence, one's view of the future of empirical Bayes method is indelibly tied to his/her view of its past and present, as well as one's own upbringing. With the widespread availability of Monte Carlo Markov Chain (MCMC) tools such as the BUGS (Baeysian inference using Gibbs sampling) software, Spiegelhalter (1995), this produces a much more pessimistic outlook for empirical Bayes method, since the need for such approximations has more or less vanished.

For categorical data, statistical methodology has only recently reached the level of sophistication achieved early in 20th century by methodology for continuous data. In contingency tables, log-linear models and logistic models have become increasingly popular tools for the analysis of data via frequentist and Bayesian methods. The empirical Bayes analysis has not as fully developed for inference about categorical data as in many other areas of statistics. An empirical Bayesian approach in estimating cell probabilities lead to estimates which usually are a combination of sample proportions and estimated moments of the prior. Leonard(1975) and Laird(1978) gave an alternative Bayesian approach, focusing on parameters of the saturated log-linear model. For two-way tables, Laird(1978) suggested an empirical Bayesian analysis, estimating parameters by finding the value that maximizes an approximation to the marginal distribution of the cell counts, evaluted at the observed data.

Here we present four case studies from political, acturial, medical, and behavoral sciences in order to illustrate how we can exploit the possibilities for contingency tables, panel data and state space modelling of cross-classified time series of counts via integration of the frequentist and Bayesian methods.

## 1.2 Some examples

**Example 1-1:**(Nazaret, 1987; Bayesian approach using posterior mode)
Nazaret using the data from Upton (1977) as an example refers to the voting transitions between 1964 and 1970 of a subset of members of a panel who

remained, throughout that period, in a constituency contested by the Conservative, Labour and Liberal parties alone. The corresponding $4 \times 4 \times 4$ array is given in Table 1-1, where $A$ corresponds to the states ( Conservative, Labour, Liberal, Abstention) in 1964, $B$ corresponds to the states in 1966, and $C$ corresponds to the states in 1970.

**Table 1-1: Observed counts for a panel of voters between 1964-1970**

**1964 (Conservative)**

| 1966 \ 1970 | *Con.* | *Lab.* | *Lib.* | *Ab.* |
|---|---|---|---|---|
| *Con.* | 57 | 0 | 1 | 5 |
| *Lab.* | 4 | 2 | 0 | 0 |
| *Lib.* | 1 | 0 | 0 | 1 |
| *Ab.* | 5 | 0 | 0 | 0 |

**1964(Labour)**

| 1966 \ 1970 | *Con.* | *Lab.* | *Lib.* | *Ab.* |
|---|---|---|---|---|
| *Con.* | 2 | 1 | 0 | 0 |
| *Lab.* | 7 | 52 | 3 | 10 |
| *Lib.* | 1 | 0 | 1 | 1 |
| *Ab.* | 1 | 0 | 0 | 3 |

**1964 (Liberal)**

| 1966 \ 1970 | *Con.* | *Lab.* | *Lib.* | *Ab.* |
|---|---|---|---|---|
| *Con.* | 8 | 0 | 0 | 0 |
| *Lab.* | 1 | 5 | 2 | 0 |
| *Lib.* | 10 | 3 | 12 | 0 |
| *Ab.* | 0 | 1 | 3 | 1 |

**1964(Abstention)**

| 1966 \ 1970 | *Con.* | *Lab.* | *Lib.* | *Ab.* |
|---|---|---|---|---|
| *Con.* | 1 | 0 | 0 | 2 |
| *Lab.* | 0 | 1 | 0 | 0 |
| *Lib.* | 0 | 0 | 0 | 0 |
| *Ab.* | 2 | 1 | 0 | 2 |

In this example, we can study, whether given the vote at the begining of a period, the vote at the end of it is independent of the transition period, and in this case we wish to explain voters behavior via Bayesian approach in time. Let $\{P_{ijk}\}$ be the transition probabilities for each cell. We use, the posterior mode method for estimating voters transition probabilities, $P_{ijk}$. This approach consists of three stages:

**1.** Under the saturated log-linear model with parameters $\gamma_{ijk}$, for $i = 1, 2, \cdots, a$ $j = 1, 2, \cdots, b$ $k = 1, 2, \cdots, c$, *i.e.*

$$\gamma_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \cdots + \lambda_{ijk}^{ABC},$$

the likelihood function is given by

$$exp\Big\{\sum_{ijk} \gamma_{ijk} y_{ijk} - N\mathcal{D}(\gamma)\Big\},$$

where $\{y_{ijk}\}$ is the table of frequencies and $N = \sum_{ijk} y_{ijk}$ is the sample size. $\mathcal{D}(\gamma) = Ln\Big(\sum_{ijk} exp(\gamma_{ijk})\Big)$ plays the role of a normalizing constant.

**2.** To specify our prior distribution we argue as follows: in many cases the researcher is in a position of weak knowledge about the effects defined in saturated log-linear model. Here, we assume that the prior knowledge about the effects can be described independently by means of exchangeable distributions. Hence, we suppose the prior distribution of $\lambda_i^A$, as the first main effect, given the variance, $\sigma_1^2$, is

$$\pi(\lambda_i^A|\sigma_1^2) \sim exp\Big\{-\frac{1}{2\sigma_1^2}\sum_i(\lambda_i^A - \lambda_0^A)\Big\}$$

where $\lambda_0^A$ is the average of the $\lambda_i^A$'s. Accordingly, the density of $\lambda_i^A$, $\cdots$ and $\lambda_{jk}^{BC}$, conditional on $\sigma_1^2$, $\sigma_2^2$, $\cdots$, $\sigma_{123}^2$ is proportional to the product of multivariate normal densities.

**3.** Hence,

$$\pi(\lambda_i^A, \cdots, \lambda_{ijk}^{ABC}|\Sigma, y),$$

is the posterior distribution, where $\Sigma = (\sigma_1^2, \cdots, \sigma_{123}^2)$. Now, we will find those values which maximize the posterior distribution. These are called posterior modes. However, in this approach, $\hat{P}_{ijk}$ are called expected voters transition probabilities given a log-linear model.

**Example 1-2:**(Ntzoufras, 2000; Bayesian approach for contingency table) Insurance companies often do not pay the outstanding claims as soon as they occur. Instead, the claims are settled with a time delay which may be years or, in some extreme cases, decades. Reserving for outstanding claims is of central interest in actuarial practice and has attracted the attention of many researchers because of the challenging stochastic uncertainties involved. Mathematically, the problem can be formulated as follows. There exist data with a structure given by Table 1-2 where $A_i, i = 1, 2, \cdots, r$ denote the accident years and $B_j, j = 1, 2, \cdots, r$ denote the year that the claim was settled. In Table 1-2, $n_{ij}$ represents claim counts which were paid by the insurance company with a delay of $j-1$ years for accidents originated at year $i$ and $T_i$ denotes the total number of accidents. Finally the inflation factor for each cell is $f_{ij}$, which is used to deflate the claim amounts, is also assumed to be known.

**Table 1-2: Structure of the outstanding claim counts data**

| | | B | | | | |
|---|---|---|---|---|---|---|
| | | $B_1$ | $B_2$ | $\cdots$ | $B_r$ | *total* |
| | $A_1$ | $n_{11}$ | $n_{12}$ | | $n_{1r}$ | $T_1$ |
| A | $A_2$ | $n_{21}$ | $n_{22}$ | | | $T2$ |
| | $\vdots$ | $\vdots$ | | | | |
| | $A_r$ | $n_{r1}$ | | | | $T_r$ |

Many models and techniques have been proposed for the prediction of the lower triangle of Table 1-2. Two of most broadly used models are the log-linear and the log-normal ANOVA-type models; see Renshaw and Verrall (1994). Markov et.al. (1996) provides a short review of the Bayesian methods and description of its application. The following data came from a major Greek insurance company. Tables 1-3 and 1-4 give the claim counts, the total counts for car accidents and the inflation factor.

**Table 1-3: The outstanding claim counts for a Greek Insurance Company**

| *Year* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Total* |
|---|---|---|---|---|---|---|---|---|
| 1989 | 6622 | 1943 | 489 | 138 | 61 | 223 | 66 | 9542 |
| 1990 | 6943 | 2133 | 632 | 154 | 162 | 390 | | 10496 |
| 1991 | 8610 | 2216 | 736 | 651 | 256 | | | 12601 |
| 1992 | 9791 | 3167 | 1570 | 624 | | | | 15565 |
| 1993 | 11722 | 3192 | 1773 | | | | | 17735 |
| 1994 | 13684 | 3664 | | | | | | 19746 |
| 1995 | 13068 | | | | | | | 18600 |

**Table 1-4: Inflation factor for Greece**

| year | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 |
|---|---|---|---|---|---|---|---|---|
| Inflation (%) | 100.0 | 120.4 | 143.9 | 166.6 | 190.6 | 214.2 | 235.6 | 257.0 |

**Example 1-3:**(Singh & Roberts, 1992; Bayesian approach for State Space time series)

The problem of modeling and projecting cross-classified categorical time series is quite common in the area of planning and policy decisions. The data are generaly in the form of a fairly long-series of tables of counts based on a large number of observations at each point in time. For instance, the Canadian cancer mortality data series consists of annual counts for each province cross-classified by cancer site, age and sex. Following Cox(1981) the time series approach to non-normal and non-linear data, can be classified into two types, namely, *observation driven* and *parameter driven*. In this example, the proposed model is termed a State Space Generalized Linear Model (SSGLM) in which the technique of recursive algorithm is modified to suit the non-normal and non-linear modeling. The SSGLM can be defined in terms of the following two equations:

i) Cross-sectional behavior: For each $t = 1, 2, \cdots, T$

$$\begin{cases} Y_t = \mu_t + e_t \\ \eta_t = g(\mu_t) = F_t \boldsymbol{\theta}_t \end{cases} \qquad (1-4)$$

where $e_t|\mu_t \sim N[0, V_t(\mu_t)]$, $Cov(e_t, e_s) = 0$ for $s < t$ , $\boldsymbol{\theta}$ is the vector of parameters in the model, $g(.)$ is a monotone differentiable link function, $F_t$ is a known

$m \times r$ matrix and the functional form of $V_t(\mu_t)$ is assumed to be known.

**ii)** Longitudinal behaviour: For $t = 2, 3, \cdots, T$

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \nu_t$$

where $G_t$ is a known $r \times r$ transition matrix, and the errors $\nu_t$ are specified by

$$\nu_t \sim N[0, W_t], \qquad Cov(\nu_t, \nu_s) = 0, \qquad s < t$$

$$Cov(\nu_t, e_s) = 0, \qquad Cov(\nu_t, \boldsymbol{\theta}_s) = 0. \qquad s < t$$

The covariance matrix $W_t$ is also assumed to be known. For fitting SSGLM, the following algorithm has been proposed by whom for estimation of the model parameters.

**Estimation Algorithm**

The Filtered and Iterative Weighted Least Squares(FIWLS) algorithm for estimating $\boldsymbol{\theta}_T$ consists of two stages, each requiring a series of iterative steps.

**Stage I: Linearization for state space formulation**

First transform $Y_t$ (response) to $\mathbf{Z}_t$ for each $t = 1, 2, \cdots, T$ as

$$\mathbf{Z}_t^{(i-1)} = \eta_t^{(i-1)} + (d\eta_t | d\mu_t)(\mathbf{Y}_t - \mu_t)\Big|_{\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^{(i-1)}}$$

where

$$E(\mathbf{Z}_t^{(i-1)}) = \eta_t^{(i-1)} = F_t \boldsymbol{\theta}_t^{(i-1)} \qquad (1-5)$$