

A canonical moment approach to the analysis of binomial mixtures¹

Un'analisi di modelli mistura binomiali basata sui momenti canonici

Luca Tardella²

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Università “La Sapienza”

luca.tardella@uniroma1.it

Riassunto: I modelli mistura di binomiali sono generalmente considerati modelli flessibili per il trattamento di dati di conteggio in grado di tener conto di eventuali aspetti di eterogeneità riscontrabili nelle osservazioni, ma anche un valido strumento per il trattamento di dati binari in presenza di scambiabilità. Per dare un'idea dell'ampiezza di applicazione si consideri che loro impiego spazia dalla teoria dei test psicologici ai contesti di affidabilità industriale nonché alla sperimentazione tossicologica. In questo lavoro vengono presentate diverse parametrizzazioni al fine di delineare un opportuno contesto di sviluppo per un'inferenza Bayesiana di riferimento in assenza di informazioni specifiche sulla forma funzionale della distribuzione misturante. Si mostra l'utilità dell'approccio proposto in problemi di stima di caratteristiche incognite con un'applicazione a dati relativi a disfunzioni polmonari. Viene accennato un primo tentativo di utilizzo di procedure bayesiane di riferimento nell'ambito dei problemi di scelta del modello in particolare per il problema di determinazione del numero di componenti.

Keywords: Mixture model, Binomial mixture, Canonical moments, Ordinary moments, Bayesian inference, Default prior.

1 Introduction

Often researchers have to deal with the following statistical context: there are n observations independently drawn from a binomial distribution, in symbols $X_i \sim \text{Bin}(N_i, \theta_i)$, for $i = 1, \dots, n$. There are circumstances where $N_i = N$, all the θ_i can be thought of as approximately equal hence the inferential problem reduces to making inference on a standard binomial set of data. However, sometimes the data suggest that a single binomial distribution does not fit the data well for the presence of overdispersion indicating that there might be substantial heterogeneity in the θ_i . Usually one way of dealing with this kind of heterogeneity is to set up the following hierarchical structure where θ_i are thought of as coming from a second-stage distribution F

$$\begin{cases} \theta_i \sim F \\ X_i \sim \text{Bin}(N_i, \theta_i) \end{cases} \quad i = 1, \dots, n. \quad (1)$$

¹Financial support from MIUR (Italian Ministry for University and Research) - Project Cofin 2002 “Costruzione ed analisi di misure di probabilità aleatorie e loro applicazione nell'ambito dell'inferenza bayesiana non parametrica”

²Mailing address: Dipartimento di Statistica, Probabilità e Statistiche Applicate - Università “La Sapienza” - Piazzale Aldo Moro, 5 - 00185 Roma - Italy

Usual ways of conceiving F are:

- *parametric* - the functional form of F is known up to the knowledge of few parameters; for instance in the beta-binomial configuration the mixing F is a $\text{Beta}(\alpha, \beta)$ and the inferential interest is basically focussed on the unknown values of those two parameters;
- *nonparametric* or *general mixture* - the functional form of F is left completely unspecified;
- *finite mixture* - the mixing distribution F is conceived as a distribution supported only on a finite number of points; this turns out to be a flexible strategy intermediate between the previous two items.

All these models have been studied under different frameworks: classical using maximum likelihood Mislevy (1984); Lindsay (1985); Brooks *et al.* (1997); Wood (1999); Böhning (1999); empirical Bayes Martz and Lian (1974); Albert (1984); and fully Bayesian Rolph (1968); Berry and Christensen (1979); Newton (2002) with some difficulties arising in the more general case of mixture due to non-identifiability of F . A much broader list of references can be gathered from classical monographs on mixture model such as Everitt and Hand (1981), Titterington *et al.* (1985), Lindsay (1995), Böhning (1999), McLachlan and Peel (2000). In fact mixture of binomial models represent a useful tool also for dealing with so-called clustered (exchangeable) binary data (Brooks *et al.*, 1997) and we will make use of this kind of approach for a specific data set in Section 6. In the next section we explore different possibilities of writing the likelihood function for the general model in terms of different parameterizations.

2 Parameterizations and likelihood

We begin with looking at model (1) in its greatest generality under a completely unspecified form of F . Our goal consists of making inference on F and, consequently, on the unobserved θ_i . If we denote with $\mathbf{x} = (x_1, \dots, x_i, \dots, x_n)$ the observed counts in n independent trials, one can immediately argue that not all features of F are identified in the likelihood function since the probability distribution of the observed data is

$$p(\mathbf{x}|F) = \prod_{i=1}^n \int_{[0,1]} \binom{N_i}{x_i} \theta^{x_i} (1 - \theta)^{N_i - x_i} dF(\theta) \quad (2)$$

and by using a simple polynomial expansion of $(1 - \theta)^{N_i - x_i}$ and integrating F out, one easily realizes that only the first $S = \max_{i=1, \dots, n} N_i$ moments of F , denoted with $(m_1^F, \dots, m_r^F, \dots, m_S^F)$, are involved in (2) so that the likelihood for F can be actually expressed as $L(F) = p(\mathbf{x}|F) = L(m_1^F, \dots, m_S^F | \mathbf{x}) = \prod_{i=1}^n \rho(m_1^F, \dots, m_S^F; N_i, x_i)$; the function $\rho(\cdot; N_i, x_i)$ can be shown to be a linear function of the first N_i moments of F involved in the polynomial expansion form of (2). In order to keep notation simple we expand in more details only the simpler case where $N_i = S$ for all $i = 1, \dots, n$. The most general case can be derived similarly. When $N_i = S$ for each $i = 1, \dots, n$ we get

$$\int_{[0,1]} \binom{S}{x_i} \theta^{x_i} (1 - \theta)^{S - x_i} dF(\theta) = \binom{S}{x_i} \sum_{r=x_i}^S (-1)^{r-x_i} \binom{S-x_i}{r-x_i} m_r^F$$

and hence for any $k = 0, 1, \dots, S$ denoting with $P_{k,F} = \sum_{r=k}^S (-1)^{r-k} \frac{S!}{(S-r)!(r-k)!} m_r^F$ and with f_k the number of observed counts x_i such that $x_i = k$, the likelihood of F can be written equivalently

$$L(F) = L(\mathbf{P}^F) = \prod_{i=1}^n P_{x_i, F} = \prod_{k=0}^S P_{k, F}^{f_k} \quad (3)$$

$$= L(\mathbf{m}_{[S]}^F) = \prod_{k=0}^S \left(\sum_{r=0}^S \gamma_{k,r} m_r^F \right)^{f_k} \quad (4)$$

where $\mathbf{m}_{[S]}^F = (m_0^F, m_1^F, \dots, m_r^F, \dots, m_S^F)$ and $\gamma_k^T = (\gamma_{k,0}, \dots, \gamma_{k,r}, \dots, \gamma_{k,S})$ with components

$$\gamma_{k,r} = \begin{cases} 0 & r < k \\ (-1)^{r-k} \frac{S!}{(S-r)!(r-k)!} & r = k, k+1, \dots, S. \end{cases} \quad (5)$$

Note also that in any case the likelihood depends on data $\mathbf{x}$ only through the vector of sufficient statistics $\mathbf{f} = (f_0, f_1, \dots, f_k, \dots, f_S)$. The above reformulations of the likelihood amount to say that the most general inferential problem of mixture of binomials in a so-called non-parametric setting reduces in fact to a finite-dimensional inference. So far we have explicitly seen that the problem can be parameterized either in terms of *marginal binomial mixture probabilities* $\mathbf{P} \in \mathcal{P}_{BinMix}$ where the range of the binomial mixture probabilities is a convex subset of the S -dimensional simplex Δ_S namely

$$\mathcal{P}_{BinMix} = \left\{ (P_0, \dots, P_k, \dots, P_S) \in \Delta_S : P_k = \int_0^1 \binom{S}{k} \theta^k (1-\theta)^{S-k} dF(\theta) \right. \\ \left. \text{for some distribution } F \text{ on } [0, 1] \right\} \quad (6)$$

or in terms of the first S *central moments* of the mixing distribution F on $[0, 1]$ whose range is a convex body and is usually referred to as the S -truncated moment space, in symbols

$$\mathcal{M}_S = \left\{ (m_1, \dots, m_r, \dots, m_S) : m_r = \int_0^1 \theta^r dF(\theta), r = 1, \dots, S \right. \\ \left. \text{for some distribution } F \text{ on } [0, 1] \right\}$$

As far as the ranges of $\mathbf{P}$ and $\mathbf{m}$ are concerned note that they are constrained S -dimensional subsets namely convex bodies which are one-to-one linearly related through $P_{k,F} = \sum_{r=k}^S (-1)^{r-k} \frac{S!}{(S-r)!(r-k)!} m_r^F$. Hence the space of the binomial mixture probabilities (6) can be reinterpreted as an image through an affine map of the convex body of the S -truncated moment space $\mathcal{M}_S$.

In this paper we will keep on dwelling further on another equivalent parametrization which prove itself very useful for an effective Bayesian analysis of binomial mixture. In fact the parameterization advocated in this paper stems from the moment parameterization as originally pioneered in Rolph (1968). However, there were **two main issues** that prevented that work to be broadly applicable and as a consequence induced later on other researchers to explore solutions based on modelling F through unidentifiable infinite-dimensional parameterizations like the ones used in Berry and Christensen (1979) and Newton (2002) through the Dirichlet process prior on the whole F .

- a) how to address the problem of **prior elicitation** and possibly suggesting some default choice;
- b) how to make **computation** in nontrivial examples.

Computational aspects would be a rather easy obstacle to overcome with Monte Carlo Markov Chain (MCMC) techniques if the n_{max} -truncated moment space involved in the parametrization (4) of the likelihood function were an easy-to-handle unrestricted space, which – unfortunately – is not, as already detailed earlier. In fact, there is another less known equivalent saturated parametrization of the first S moments which turns out to map either $\mathcal{P}_{BinMix}$ or $\mathcal{M}_S$ onto a convenient compact product space. That parametrization is expressed in terms of the so-called *canonical moments* (Skibinsky, 1967) which are quantities in the unit interval, otherwise unconstrained, which can be mapped back and forth from ordinary moments in terms of a simple recursive algorithm involving only summations and multiplications. Hence this can candidate itself to act as an appropriate parametrization to implement an MCMC scheme to produce approximations for any posterior output of interest in terms of canonical moments simulated from the Markov chain. An instance of actual implementation of this approach is basically already detailed in Tardella (2002) for a particular case of binomial mixture problems related to the estimation of a capture-recapture model. Another example of use of canonical moments within a Bayesian framework is sketched in Robert and Casella (1999, page 356) regarding geometric mixtures. We defer to the next section a formal description of canonical moments. For now we just highlight that the one-to-one mapping $\mathbf{m} = \psi(\mathbf{c})$ between canonical moments $\mathbf{c} = (c_1, \dots, c_S) \in [0, 1]^S$ and ordinary moments $\mathbf{m}$ allows to enrich the approaches to the binomial mixture model (1) with a third – less known – equivalent parametrization. Summing up we have the following three equivalent likelihood formulae. The first one as in (3).

$$L^1(\mathbf{P}) = \prod_{k=0}^S P_k^{f_k} \quad \mathbf{P} \in \mathcal{P}_{BinMix} \quad (7)$$

The second one as in (4)

$$L^2(\mathbf{m}) = \prod_{k=0}^S (\gamma_k^T \mathbf{m}_{[S]})^{f_k} \quad \mathbf{m} \in \mathcal{M}_S \quad (8)$$

Finally, in terms of canonical moments

$$L^3(\mathbf{c}) \propto L^2(\psi(\mathbf{c})) \quad \mathbf{c} \in [0, 1]^S. \quad (9)$$

where the likelihood is not expressed explicitly but rather as a composite function, through the use of $\psi(\cdot)$. In fact, as it will be explained later, some care is needed when considering points in the boundaries of the corresponding parameter spaces. We stress here again the importance of having a likelihood function like (9) whose range is a nice product space rather than an involved constrained subset.

3 Canonical moments

The basic features of canonical moments needed for our purposes are mentioned below, while further details and references can be found in Dette and Studden (1997). Consider a fixed value $\mathbf{m} \in \text{int}(\mathcal{M}_S)$ and $j \in \{0, 1, \dots, S-1\}$. Denote by $\mathbf{m}_{(j)}$ the vector of the first j components of $\mathbf{m}$, hence define the corresponding j -truncated moment

class of distributions $\mathcal{F}_{m_{(j)}} = \left\{ F \in \mathcal{F} : \int_{[0,1]} p^r dF(p) = m_r, r = 1, \dots, j \right\}$ and for this class we define the following quantities $m_{j+1}^-(m_{(j)}) = \inf_{F \in \mathcal{F}_{m_{(j)}}} \int_{[0,1]} p^{j+1} dF(p)$ and $m_{j+1}^+(m_{(j)}) = \sup_{F \in \mathcal{F}_{m_{(j)}}} \int_{[0,1]} p^{j+1} dF(p)$. Consider the mapping $\psi^{-1} : \text{int}(\mathcal{M}_S) \mapsto (0, 1)^S$, implicitly determined by the following relationship: $\mathbf{c} = (c_1, \dots, c_S) = \psi(\mathbf{m})$ is such that, for any $j = 0, \dots, S-1$,

$$c_{j+1} = \frac{m_j - m_{j+1}^-(m_{(j)})}{m_{j+1}^+(m_{(j)}) - m_{j+1}^-(m_{(j)})}, \quad j = 0, \dots, S-1. \quad (10)$$

To help understanding this relationship consider that the inverse mapping $\psi : (0, 1)^S \mapsto \text{int}(\mathcal{M}_S)$ can be determined by the mirror recursive relationship

$$m_{j+1} = (1 - c_{j+1}) m_{j+1}^-(m_{(j)}) + c_{j+1} m_{j+1}^+(m_{(j)}).$$

The quantities c_j are usually referred to as canonical moments. Simple modifications are needed to deal with moments on the boundary of the truncated moment space. Dette and Studden (1997, Theorem 1.5.1, page 27; Theorem 2.4.4, page 68) give algorithms for transforming ordinary moments into canonical moments and vice versa. Although the required computations grow with S as $O(S^2)$, the algorithm only involves iterative finite summations and multiplications.

A predictive interpretation of canonical moments can provide a useful insight for the elicitation of informative priors when exchangeable binary sequences are dealt with through binomial mixtures. It is based on the following representation: suppose that $\mathbf{Z}_i = (Z_1^i, \dots, Z_r^i, \dots, Z_S^i)$ is a finite exchangeable binary sequence so that $X_i \stackrel{d}{=} \sum_{r=1}^S Z_r^i$ can be regarded as an exhaustive summary. Conditionally on $\theta_i \sim F$ the Z_r^i sequence can be modelled as generated from Bernoulli experiments with success probability θ_i independently and with identical distribution for $r = 1, \dots, S$. This can be conceived when one deems appropriate to restrict the finitely exchangeable binary sequence to have non-negative pairwise correlation. Then for the first two canonical moments one easily get

$$c_1 = P\{Z_i = 1\}$$

$$c_2 = \frac{P\{Z_i = 1 | Z_r = 1\} - \max P\{Z_i = 1 | Z_r = 1\}}{\max P\{Z_i = 1 | Z_r = 1\} - \min P\{Z_i = 1 | Z_r = 1\}}$$

while Lau (1992), in order to generalize the interpretation of canonical moments as coefficients of expected agreement similarly to the so-called Cohen's Kappa coefficient (here corresponding to c_2) often used in psychological testing contexts, get a more general expression holding for any k and $r < k$

$$c_k = \frac{P\{Z_k = 1 | \sum_{t=1}^{k-1} Z_t = r\} - \min P\{Z_k = 1 | \sum_{t=1}^{k-1} Z_t = r\}}{\max P\{Z_k = 1 | \sum_{t=1}^{k-1} Z_t = r\} - \min P\{Z_k = 1 | \sum_{t=1}^{k-1} Z_t = r\}}$$

where the minimum and the maximum related to the generic c_j are taken over the space of mixing distributions with fixed values of the constrained predictions c_r (canonical moments) up to the order $j-1$.

Two further simple and meaningful interpretations of c_2 are to be mentioned: the first one is in terms of the correlation of any two binary outcomes $c_2 = \text{Corr}[Z_r, Z_s]$; the second one is related to (10) from which c_2 can be understood as an indicator of dispersion allowed (in terms of variance) for mixing distributions F in which the mean has already been fixed.

4 Default Prior

As far as prior specification for a Bayesian analysis is concerned we review in this section a range of possible choices for a prior distribution for the appropriate features of the mixing F in the framework of binomial mixture. Some naively intuitive choices could be either the uniform distribution on the truncated ordinary moment space $\mathcal{M}_S$ or the uniform distribution on the canonical moment space $[0, 1]^S$. However, if we are looking for prior specification which is theoretically well-grounded possibly being invariant under reparametrization one could choose a Jeffreys prior which has the additional appealing feature of usually leading to good frequentist properties of the corresponding posterior analysis. The derivation of a Jeffreys prior can be started from looking at the model in terms of the binomial mixture probabilities (7). In fact that likelihood turns out to be proportional to a multinomial likelihood constrained on the restricted space $\mathcal{P}_{BinMix}$ for which a Jeffreys prior is known and in this particular case coincides with the reference prior of Bernardo (1979). In fact the constrained space $\mathcal{P}_{BinMix}$ is a subset with nonempty interior of the original simplex hence the functional form of the Jeffreys prior in terms of $\mathbf{P}$ stays unchanged namely

$$\pi_J(\mathbf{P}) \propto \left(\prod_{k=0}^S P_k \right)^{-\frac{1}{2}} \quad \mathbf{P} \in \mathcal{P}_{BinMix}. \quad (11)$$

Also (11) can be rewritten in terms of the other parameterizations provided that the appropriate Jacobians are included. In fact, the Jacobian of the mapping $(P_1, \dots, P_S) \rightarrow (m_1, \dots, m_S)$ can be easily proved to be a constant term since they are linearly related while the Jacobian of the ψ mapping is derived in Chang *et al.* (1993) hence

$$\pi_J(\mathbf{c}) \propto L^3(\mathbf{c}) \prod_{j=1}^S \{c_j(1 - c_j)\}^{S-j}, \quad \mathbf{c} \in [0, 1]^S. \quad (12)$$

Another variation on this default prior theme could be the ordered reference prior approach of Berger and Bernardo (1992) for a multinomial likelihood (see also Sono (1983)) where, for an appropriately selected order of $(P_0, \dots, P_S)$ – say $(P_{\sigma(0)}, \dots, P_{\sigma(S)})$ –, the ordered reference prior of the constrained multinomial probabilities is written as

$$\pi_{OR}(\mathbf{P}) \propto \left\{ \prod_{k=0}^S P_{\sigma(k)} \left[\prod_{k=1}^{S-1} \left(\sum_{j=0}^k P_{\sigma(j)} \right) \right] \right\}^{-\frac{1}{2}}. \quad (13)$$

5 Finite mixture representation and canonical moments

We describe in this section how canonical moments can be useful also for dealing with mixtures of binomials where the mixing distribution F is supported on a finite number of points as follows

$$F(A) = \sum_{j=1}^k \pi_j \delta_{\theta_j}(A), \quad A \in \mathcal{B}_{[0,1]}. \quad (14)$$

Any mixture model with such a mixing distribution is usually referred to as *finite mixture* and its popularity is due to two main reasons:

- the possibility of representing a clustered structure of the observed population with the interpretation of the generic θ_j component as the parameter identifying a homogeneous subgroup (cluster).
- the possibility of having an increasingly flexible parametric structure where the flexibility grows with the number of components

It can be shown (Dette and Studden, 1997; Skibinsky, 1986) that in the case of mixing distributions on the unit interval one can basically represent the subspace $\mathcal{F}_k$ of discrete mixing distributions as in (14) through an equivalent canonical moment parametrization as a subset of $[0, 1]^S$ in the following way:

$$\mathcal{F}_k^{CM} = \{(c_1, \dots, c_r, \dots, c_{2k-1}, 0, \dots, 0), \quad c_r \in (0, 1)\} \quad (15)$$

Note that the nested model structure of the finite mixture representations for different values of k is now greatly simplified and also additional label-switching identification problems are avoided (Stephens, 2000; Celeux *et al.*, 2000). It is also well understood the dimensional match between the $2k - 1$ canonical moments and the $(2k - 1)$ -dimensional parametrization given through locations and weights of the mixture. In the above representation (15) we have actually omitted a couple of facts whose details are fully provided in Dette and Studden (1997) and Skibinsky (1986): on one hand we have in fact represented only mixing distributions which do not have either 0 or 1 (or both) in their support; on the other hand we have slightly extended the canonical moment representation of discrete distributions, whose ordinary truncated moment representation lies in fact on the boundary of the truncated moment space whereas the whole moment sequence is in fact one-to-one identified just with the first $2k$ canonical moments namely $(c_1, c_2, \dots, c_r, \dots, c_{2k-1}, 0)$.

Another by-product of the reference prior approach developed earlier is related to the possibility of considering that reference prior restricted to the subspace spanned by the finite mixing distributions supported on k points. In this case we can extend the use of the default prior derived in Section 4 on the restricted space of binomial mixture distributions with finitely supported mixing, which can be rewritten in terms of canonical moments; for instance (13) can be evaluated as follows

$$\pi_D^{\mathcal{F}_k}(c_1, \dots, c_{2k-1}) \propto \pi_{OR}(\mathbf{P}(\mathbf{m}(c_1, \dots, c_{2k-1}))) \left| \frac{\partial \mathbf{m}(c_1, \dots, c_{2k-1})}{\partial (c_1, \dots, c_{2k-1})} \right|. \quad (16)$$

The canonical moment parametrization provides an alternative to the well-recognized difficult task of deriving default prior distributions for finite mixtures. Previous general attempts can be found in Roeder and Wasserman (1997) Pérez and Berger (2002) and Moreno and Liseo (2003). On the other hand the finite mixture representation is often preferred due to its greater simplicity and easier understanding of what the mixing distribution looks like. Another important point in favor of such parametrization is that from a computational viewpoint it becomes much lighter to handle especially when the number of components of the mixture are few and the N_i parameters are high, say $N_i > 20$. In this last case the computing burden of the ψ mapping together with the stability of computation suggests to avoid the evaluation of the likelihood through $L^3(c)$ as in (9). As a marginal aside we note that formula (9) can be used for directly computing the maximum likelihood value (and the corresponding class of mixing distributions attaining that maximum) when F varies in $\mathcal{F}_k$ by using standard tools for optimization in several dimensions. In our limited experience for a reasonable range of S values we got a reliable behaviour possibly slower but sometimes a little bit more precise than the standard *EM* or *VEM* algorithm suggested in Böhning (1999).

Table 1: *COPD Data*

SIBLING SIZE	FREQUENCY OF CASES OF IMPAIRED PULMONARY FUNCTION WITHIN THE FAMILY							TOTAL
	0	1	2	3	4	5	6	
1	36	12	–	–	–	–	–	48
2	15	7	1	–	–	–	–	23
3	5	7	3	2	–	–	–	17
4	3	3	1	0	0	–	–	7
6	1	0	1	1	1	0	1	5

6 Application to clustered binary data

In this example we are interested in data originally described and analyzed in Cohen (1980) and later re-analyzed in Liang *et al.* (1992) and Stefanescu and Turnbull (2003). They concern Chronic Obstructive Pulmonary Disease (COPD) for which genetic epidemiologists recorded for each of 100 families, whose sizes varied from 1 to 6, the number of siblings with impaired pulmonary function. These data have been modelled in Stefanescu and Turnbull (2003) by means of a finitely exchangeable saturated parametrization for clustered binary data where each binary sequence $(Z_1, \dots, Z_{n_i})$ corresponds to Bernoulli outcomes in the set of n_i siblings of the same family and is assumed to be exchangeable. In fact, the data as analyzed in Stefanescu and Turnbull (2003) and summarized through the sufficient statistics displayed in Table 1, showed no evidence of negative correlation hence we restricted the attention to binary subsequences extracted from infinitely exchangeable binary outcomes for which only non negative correlation is allowed. In fact, this is equivalent to represent the summary statistics of a generic set of siblings $X_i = \sum_{s=1}^{n_i} Z_s^i$ as a mixture of binomial outcomes. Intra-class correlation $\rho = \text{Corr}[Z_r, Z_s]$ is one of the main objectives of researchers (Stefanescu and Turnbull, 2003, Section 5.1) and – as also stressed in Section 3 – that quantity corresponds to the second canonical moment $c_2 = \rho$. A second quantity used in that paper to get evidence of genetic association namely the odds ratio $\Psi = \frac{m_2(1-2m_1+m_2)}{(m_1-m_2)^2}$, corresponding to the ratio of expected probabilities of pairs of concordant binary outcomes against discordant ones, whose expression can be also translated in terms of the first two canonical moments.

The canonical moment Bayesian approach advocated in this paper together with the possible nested restrictions allowed for by setting to 0 the last canonical moments, corresponding in the particular case described in the previous section to varying dimensional finite mixture representation of F , delineates a flexible modelling framework ranging from the independent case of homogeneous binomial counts to increasingly complex (positive) correlation structures of the binary counts within each family. In fact such flexibility is not present in the analysis of Stefanescu and Turnbull (2003) where they skipped from a two-parameter Beta-binomial model directly to a six-parameter saturated parametrization. Recent trans-dimensional *MCMC* simulations techniques (Petrakis and Tardella, 2003) provided an automatic computational strategy for making inference with a full likelihood approach joint with flexible model averaging on the presence of genetic association. In Table 2 we display the posterior summaries resulting from posterior model averaging of

Table 2: *Posterior summaries of the COPD data analysis derived from model averaging of six nested submodels allowing for increasingly complex dependence structure among binary outcomes.*

	NESTED SUBMODEL					
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$
POSTERIOR PROBABILITY	0.01	0.03	0.13	0.25	0.29	0.29
	MEAN			STANDARD DEVIATION		
ρ	0.1981			0.0868		
Ψ	2.7456			1.2387		

the first 6 canonical moments over $\bigcup_{k=1}^6 \mathcal{C}_k$ where

$$\mathcal{C}_k = \left\{ (c_1, c_2, \dots, c_r, \dots, c_k, 0, \dots, 0) \in [0, 1]^6 : 0 < c_r < 1, r = 1, \dots, k \right\}.$$

We note that our results are in general agreement with the conclusions derived from the saturated parametrization model entertained in Stefanescu and Turnbull (2003) though their model selection discussion was in fact favoring a more parsimonious beta-binomial model which resulted in larger association figures together with larger variability of the estimates.

7 Determining the order of a mixture model: homogeneity versus heterogeneity

When the finite mixture representation is used as a mean to understand clustering structure of the data the number of clusters is likely to be not known in advance and represents actually one of the main inferential goal. A preliminary narrower question in determining the number of clusters is to see whether data have been generated by a homogeneous set of θ_i (i.e. $F \in \mathcal{F}_k$ with $k = 1$) or rather there is substantial heterogeneity so that at least two clusters can be detected (i.e. $F \in \mathcal{F}_k$ with $k > 1$). This problem is either faced by means of so-called homogeneity tests or can be regarded as a model selection instance with two alternative hypotheses:

$$\begin{cases} H_0 : & F \in \mathcal{F}_1 \text{ i.e. } F(\cdot) = \delta_{\{\theta_0\}}(\cdot), \theta_0 \in [0, 1] \\ H_1 : & F \in \mathcal{F}_2 \text{ i.e. } F(\cdot) = \pi_1 \delta_{\{\theta_1\}}(\cdot) + (1 - \pi_1) \delta_{\{\theta_2\}}(\cdot) \quad \theta_i \in [0, 1], \pi_1 \in [0, 1] \end{cases}$$

This problem is relatively simple and admits a long list both of general solutions such as $C(\alpha)$ tests (Neyman and Scott, 1965) and likelihood ratio tests and also some specific solutions for the case of mixture of Binomials (Potthoff and Whittinghill, 1966). However, there has been recent renewed interest in this topic mainly due to advances in overcoming technical difficulties in implementing the non-standard likelihood asymptotics (Chen *et al.*, 2001), deriving alternative powerful test statistics (Charnigo and Sun, 2004) and extending the results to more general cases of arbitrary selection between two alternative number of components (Chen *et al.*, 2004; Susko, 2003). In a Bayesian framework there are several ways to address the problem of determining whether there is homogeneity in binomial count data. We will not give an exhaustive account here but we just mention some of them which have been entertained recently: a model choice determination based on

Table 3: *Comparative performance in terms of empirical significance and power of CM2-Canonical Moment test, $C(\alpha)$ -Homogeneity test of Neyman and Scott (1965), CCK-penalized LRT of Chen et al. (2001)] ($\alpha = 0.04$); datasets have been simulated from the following mixing distributions: D1) $0.5\delta_{0.4} + 0.5\delta_{0.6}$ D2) $0.5\delta_{0.2} + 0.5\delta_{0.4}$ D3) $0.8\delta_{0.2} + 0.2\delta_{0.4}$ D4) $0.9\delta_{0.1} + 0.1\delta_{0.5}$ D5) $\delta_{0.5}$ D6) $\delta_{0.3}$ D7) $\delta_{0.24}$ D8) $\delta_{0.14}$*

SIMULATED DATASETS	SAMPLE SIZE $n = 30$			SAMPLE SIZE $n = 50$			SAMPLE SIZE $n = 100$		
	CM2	$C(\alpha)$	CCK	CM2	$C(\alpha)$	CCK	CM2	$C(\alpha)$	CCK
D1	0.27	0.19	0.21	0.25	0.18	0.29	0.42	0.30	0.36
D2	0.22	0.17	0.16	0.31	0.20	0.27	0.52	0.37	0.45
D3	0.22	0.16	0.16	0.23	0.19	0.17	0.41	0.30	0.34
D4	0.63	0.58	0.55	0.68	0.69	0.65	0.90	0.89	0.89
D5	0.04	0.01	0.02	0.13	0.04	0.03	0.06	0.03	0.03
D6	0.10	0.09	0.8	0.07	0.02	0.02	0.06	0.02	0.03
D7	0.07	0.04	0.04	0.11	0.11	0.08	0.05	0.06	0.04
D8	0.04	0.04	0.03	0.07	0.06	0.03	0.04	0.02	0.01

- posterior probabilities/Bayes factors for the two competing models either using trans-dimensional simulation or using recursive exact evaluation (Moreno and Liseo, 2003)
- posterior analysis of distances between parameters on the larger model space and appropriate projections of them in the smaller model space Sahu and Cheng (2003) following the original idea of Mengersen and Robert (1996)

However, we note that a thorough comparative analysis based on a common benchmark of the performance of the different solutions is not available. A more extensive computational comparison will be pursued in Liseo *et al.* (2004) for the special case addressed here of Binomial mixtures using some frequentist measure of performance as quantitative benchmark. Here we want to sketch another Bayesian solution whose frequentist properties are likely to be competitive with the most recent classical tools developed based on penalized likelihood distance, or L^2 distance. The idea is the following: one can use the larger model space - in this particular case represented by $\mathcal{F}_2^{CM} \leftrightarrow \{(c_1, c_2, c_3, 0), c_i \in (0, 1), i = 1, 2, 3\}$ together with the constrained reference prior π_D as described in (16) to get a posterior distribution over the first three canonical moments of the mixing distribution F . From the previous section it is natural to use the information of the proximity of the second canonical moment c_2 to 0 as an evidence that the data might have been drawn from a degenerate mixing. To understand the comparative performance of this proposal we have used the posterior mean $E[c_2|\mathbf{x}]$ as a test statistic and determined a classical critical region by resampling the distribution of $E[c_2|\mathbf{x}^r]$ with artificially fabricated data $\mathbf{x}^r$ from a single component mixture estimated through the original data. In fact in a series of simulations limited to the case $S = 5$ and presented in Table 3 we experienced a promising competitive frequentist performance of a test statistics derived from the posterior distribution of c_2 under the larger model. Of course this idea, though deserving a more extensive simulation and a more formal theoretical insight, confirms the importance of deriving formal reference priors for this particular model.

References

- Albert J.H. (1984) Empirical Bayes estimation of a set of binomial probabilities, *Journal of Statistical Computation and Simulation*, 20, 129–144.
- Berger J.O. and Bernardo J.M. (1992) Ordered group reference priors with application to the multinomial problem, *Biometrika*, 79, 1, 25–37.
- Bernardo J.M. (1979) Reference posterior distributions for Bayesian inference, *J. Roy. Statist. Soc. Ser. B*, 41, 2, 113–147, with discussion.
- Berry D.A. and Christensen R. (1979) Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes, *The Annals of Statistics*, 7, 558–568.
- Böhning D. (1999) *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping, and Others*, CRC Press.
- Brooks S.P., Morgan B.J.T., Ridout M.S. and Pack S.E. (1997) Finite mixture models for proportions, *Biometrics*, 53, 1097–1115.
- Celeux G., Hurn M. and Robert C. (2000) Computational and inferential difficulties with mixture posterior distributions, *J. Amer. Statist. Assoc.*, 95, 957–970.
- Chang F.C., Kemperman J.H.B. and Studden W.J. (1993) A normal limit theorem for moment sequences, *Ann. Probab.*, 21, 3, 1295–1309.
- Charnigo R. and Sun J. (2004) Testing for homogeneity in mixture distributions via the l^2 distance between competing models, *Journal of the American Statistical Association*, to appear.
- Chen H., Chen J. and Kalbfleisch J.D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models, *Journal of the Royal Statistical Society, Series B, Methodological*, 63, 1, 19–29.
- Chen H., Chen J. and Kalbfleisch J.D. (2004) Testing for a finite mixture model with two components, *Journal of the Royal Statistical Society, Series B, Methodological*, 66, 1, 95–115.
- Cohen B.H. (1980) Chronic obstructive pulmonary disease: a challenge in genetic epidemiology, *American Journal of Epidemiology*, 112, 2, 274–288.
- Detle H. and Studden W.J. (1997) *The theory of canonical moments with applications in statistics, probability, and analysis*, John Wiley & Sons Inc., New York, a Wiley-Interscience Publication.
- Everitt B.S. and Hand D.J. (1981) *Finite mixture distributions*, Chapman & Hall, London, monographs on Applied Probability and Statistics.
- Lau T.S. (1992) The reliability of exchangeable binary systems, *Statistics & Probability Letters*, 13, 153–158.
- Liang K.Y., Zeger S.L. and Qaqish B. (1992) Multivariate regression analyses for categorical data, *J. Roy. Statist. Soc. Ser. B*, 54, 1, 3–40, with discussion and a reply by the authors.
- Lindsay B.G. (1985) Errors in inspection: Integer parameter maximum likelihood in a finite population, *Journal of the American Statistical Association*, 80, 879–885.
- Lindsay B.G. (1995) *Mixture models: theory geometry and applications*, IMS, Hayward, CA, NSF-CBMS Regional Conference Series in Probability and Statistics.
- Liseo B., Petris G. and Tardella L. (2004) Alternative approaches for bayesian binomial mixtures: a computational comparative analysis, Technical report, Manuscript.
- Martz H.F. and Lian M.G. (1974) Empirical Bayes estimation of the binomial parameter, *Biometrika*, 61, 517–523.
- McLachlan G. and Peel D. (2000) *Finite mixture models*, Wiley Series in Probability and

- Statistics: Applied Probability and Statistics, Wiley-Interscience, New York.
- Mengersen K.L. and Robert C.P. (1996) Testing for mixtures: a Bayesian entropic approach, in: *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Univ. Press, New York, Oxford Sci. Publ., 255–276.
- Mislevy R.J. (1984) Estimating latent distributions, *Psychometrika*, 49, 359–381.
- Moreno E. and Liseo B. (2003) A default Bayesian test for the number of components in a mixture, *J. Statist. Plann. Inference*, 111, 1-2, 129–142, model selection, model diagnostics, empirical Bayes and hierarchical Bayes. Special issue, I (Lincoln, NE, 1999).
- Newton M.A. (2002) On a nonparametric recursive estimator of the mixing distribution, *Sankhya, Series A, Indian Journal of Statistics*, 64, 2, 306–322.
- Neyman J. and Scott E. (1965) On the use of $C(\alpha)$ optimal tests of composite hypotheses. (With discussion.), *Bull. Inst. Internat. Statist.*, 41, 477–497.
- Pérez J.M. and Berger J.O. (2002) Expected-posterior prior distributions for model selection, *Biometrika*, 89, 3, 491–511.
- Petris G. and Tardella L. (2003) A geometric approach to transdimensional markov chain monte carlo, *Canad. J. Statist.*, 31, 4, 469–482.
- Potthoff R.F. and Whittinghill M. (1966) Testing for homogeneity. I. The binomial and multinomial distributions, *Biometrika*, 53, 167–182.
- Robert C.P. and Casella G. (1999) *Monte Carlo statistical methods*, Springer-Verlag Inc.
- Roeder K. and Wasserman L. (1997) Practical Bayesian density estimation using mixtures of normals, *J. Amer. Statist. Assoc.*, 92, 439, 894–902.
- Rolph J.E. (1968) Bayes estimation of mixing distributions, *The Annals of Mathematical Statistics*, 39, 1289–1302.
- Sahu S.K. and Cheng R.C.H. (2003) A fast distance-based approach for determining the number of components in mixtures, *The Canadian Journal of Statistics*, 31, 3–22.
- Skibinsky M. (1967) The range of the $(n + 1)$ -th moment for distributions on $[0, 1]$, *J. Appl. Probability*, 4, 543–552.
- Skibinsky M. (1986) Principal representations and canonical moment sequences for distributions on an interval, *J. Math. Anal. Appl.*, 120, 1, 95–118.
- Sono S. (1983) On a noninformative prior distribution for Bayesian inference of multinomial distribution's parameters, *Ann. Inst. Statist. Math.*, 35, 2, 167–174.
- Stefanescu C. and Turnbull B.W. (2003) Likelihood inference for exchangeable binary data with varying cluster sizes, *Biometrics*, 59, 1, 18–24.
- Stephens M. (2000) Dealing with label switching in mixture models, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62, 4, 795–809.
- Susko E. (2003) Weighted tests of homogeneity for testing the number of components in a mixture, *Comput. Statist. Data Anal.*, 41, 3-4, 367–378, recent developments in mixture models (Hamburg, 2001).
- Tardella L. (2002) A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity, *Biometrika*, 89, 4, 807–817.
- Titterton D.M., Smith A.F.M. and Makov U.E. (1985) *Statistical analysis of finite mixture distributions*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Ltd., Chichester.
- Wood G.R. (1999) Binomial mixtures: Geometric estimation of the mixing distribution, *The Annals of Statistics*, 27, 1706–1721.