THE UNIVERSITY OF CHICAGO

INFINITE EXCHANGEABILITY AND PARTITIONS
AND
PERMANENT PROCESS AND CLASSIFICATION MODELS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
JIE YANG

CHICAGO, ILLINOIS
AUGUST 2006

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The dissertation consists of two parts.

In Part I: *Infinite Exchangeability and Partitions*, we develop a partition model with applications to multiple comparisons and cluster analysis. Unlike an ordinary Bayesian setup, we construct an infinitely exchangeable variety process and assign positive probability to each partition of the varieties. Using this process as a prior in a Gaussian model, we obtain inferences in the form of a posterior distribution on partitions. For typical multiple comparison applications, we suggest the Ewens family as a class of prior distributions on partitions with parameter in the range roughly 1-4. We also give inference for variety contrasts from the partition model, which allows positive probabilities for the events that two or more varieties are equal. For application to cluster analysis, we develop MCMC algorithms to estimate summary statistics, especially the similarity matrix.

In Part II: *Permanent Process and Classification Models*, we develop a classification model based on the permanent process. In the model, there are only 2-3 estimable parameters, regardless of the number of classes or the dimension of the feature space. The model works well even if the classes occupy non-convex regions or disconnected regions in the feature space. Under the model, we express the conditional distribution of the class of a subsequent unit given the training data in terms of ratios of weighted

permanents. We propose an analytic approximation for the weighted permanent ratios based on the cycle expansion of the weighted permanent. Our experience is that the approximation usually has acceptable error for typical classification problems.

ACKNOWLEDGEMENTS

I am grateful to Professor Peter McCullagh, for all the advice and support I need throughout the years. It is fortunate for me to work with him, to learn statistics and other matters from him, and to complete this thesis with him.

I want to thank the faculty, students and staff in the Department of Statistics at the University of Chicago for such a wonderful environment. Particularly, I am thankful to Professors Steven Lalley, Wei-Biao Wu and Zhiyi Chi for helping me during some discussions.

I feel blessed to have Liping and share the life with her. Without her encouragement and love, this work could hardly be done. I must also thank my parents for supporting me, and my brother for always being there for me.

Finally, I acknowledge the support provided by NSF Grant DMS-0305009.

PART I: INFINITE EXCHANGEABILITY AND PARTITIONS

CHAPTER 0

PRELIMINARIES

In this chapter, we shortly review several concepts including set partition, integer partition, etc. We will revisit them in the later chapters.

0.1 Set Partition

0.1.1 Partition of Set

A *partition* of a non-empty set S is a collection of disjoint non-empty subsets, called *blocks*, whose union is S ([4], [63]). In most cases, we are interested in the partitions of a finite set such as

$$[n] = \{1, 2, \ldots, n\} \ .$$

For example, there are totally 5 different partitions of $[3] = \{1, 2, 3\}$:

$$\{\{1\}, \{2\}, \{3\}\}, \quad \{\{1, 2\}, \{3\}\}, \quad \{\{1, 3\}, \{2\}\}, \quad \{\{1\}, \{2, 3\}\}, \text{ and } \{\{1, 2, 3\}\}. \quad (1)$$

Note that the order of the subsets or blocks in a partition does not matter.

A more convenient notation for a set partition is to use vertical bars to indicate the partitioning ([37]). Then the 5 partitions above can be rewritten as

$$1|2|3, \quad 12|3, \quad 13|2, \quad 1|23, \quad 123. \tag{2}$$

Any partition of $[n]$ naturally implies an *equivalence relation* among $[n]$ by

$$i \sim j \qquad \text{if and only if} \qquad i, j \text{ belong to the same block.}$$

We denote sometimes a set partition by its equivalence relation matrix (see Section 3.2.2) . For example, the matrix form of the partition $13|2|45$ is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{3}$$

As usual, the entry in the ith row and the jth column is 1 if and only if $i \sim j$. The matrix is always positive semi-definite. Its rank is equal to the number of blocks in the corresponding partition.

0.1.2 Partition Lattice

Denote by $\mathcal{E}_n$ the set of all partitions of $[n]$. There is a natural partial order among $\mathcal{E}_n$, called *sub-partition* ([37]). Indeed, given two set partitions E^1 and E^2 of $[n]$, we say E^1 is a *sub-partition* of E^2, denoted by $E^1 \leq E^2$, if each block of E^1 is a subset

of some block of E^2. In matrix form,

$$E^1 \leq E^2 \iff E^1_{ij} = 1 \text{ always implies } E^2_{ij} = 1. \tag{4}$$

If both $E^1 \leq E^2$ and $E^2 \leq E^1$ are true, E^1 and E^2 must be the same partition.

The partial order "sub-partition" makes $\mathcal{E}_n$ a *complete lattice*, which means each nonempty subset of $\mathcal{E}_n$ has both a *least upper bound* and a *greatest lower bound* in $\mathcal{E}_n$ (see [13] for a good introduction on general lattices). For example, the least upper bound of $E^1 = 1|23|4$ and $E^2 = 1|24|3$, denoted by $E^1 \bigvee E^2$, is $1|234$; the greatest lower bound of E^1 and E^2, denoted by $E^1 \bigwedge E^2$, is $1|2|3|4$.

Note that there exist partial order sets which are not lattices because the "least" upper bound or the "greatest" lower bound of two arbitrary elements may not exist (see [37] for such an example).

0.1.3 Bell Number

A fundamental question on set partitions is how many of them there are. In the literature, the number of partitions of $[n]$ is called a *Bell number*, denoted by B_n ([57]).

There is no explicit formula for B_n. Nevertheless, Bell numbers can be calculated conveniently via the number of partitions of $[n]$ including exactly k blocks, denoted

by $S(n,k)$ and known as the *Stirling number of the second kind* ([1], [57]). Indeed,

$$B_n = \sum_{k=1}^{n} S(n,k).$$

The Stirling numbers of the second kind can be obtained by the recursive relation

$$S(n,k) = S(n-1,k-1) + kS(n-1,k), \tag{5}$$

with initial values $S(n,1) \equiv S(n,n) \equiv 1$. Indeed, there are only two ways to generate a k-block partition E_n of $[n]$ by adding the item n into a partition E_{n-1} of $[n-1]$. One way is to attach n as a single-item-block to E_{n-1} which includes exactly $k-1$ blocks; the other is to insert n into one of the existing blocks of E_{n-1} containing exactly k blocks. Thus $S(n,k)$ is the sum of $S(n-1,k-1)$ and k times of $S(n-1,k)$. In practice, it is fairly convenient to establish by hand the triangle of $S(n,k)$ as follows:

$$\begin{array}{cccccccc}
1 & & & & & & & \\
1 & 1 & & & & & & \\
1 & 3 & 1 & & & & & \\
1 & 7 & 6 & 1 & & & & \\
1 & 15 & 25 & 10 & 1 & & & \\
1 & 31 & 90 & 65 & 15 & 1 & & \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots &
\end{array}$$

Here the entry in the nth row and kth column is $S(n,k)$. The sum of the nth row provides the value of B_n. To see how fast B_n increases with n, we list the first 20 Bell numbers in Table 1 ([37]).

Table 1: First 20 Values of Bell Numbers

n	B_n	n	B_n
1	1	11	678,570
2	2	12	4,213,597
3	5	13	27,644,437
4	15	14	190,899,322
5	52	15	1,382,958,545
6	203	16	10,480,142,147
7	877	17	82,864,869,804
8	4,140	18	682,076,806,159
9	21,147	19	5,832,742,205,057
10	115,975	20	51,724,158,235,372

Lovász ([33]) showed the following asymptotic property of the Bell numbers

$$B_n \sim n^{-\frac{1}{2}}[\lambda(n)]^{n+\frac{1}{2}}e^{\lambda(n)-n-1},$$

where $\lambda(n)$ is defined by

$$\lambda(n)\ln[\lambda(n)] = n.$$

Thus, B_n increases slower than $n!$ and faster than e^n ([27]).

The *Bell polynomial* $B_n(\lambda)$ ([58]) is the coefficient of $t^n/n!$ in the Taylor expansion of $\exp\{\lambda(e^t-1)\}$, which is the moment generating function of the Poisson distribution with mean λ. In other words, $B_n(\lambda) = E(X^n)$ if X is a Poisson random variable with mean λ. Note that all the cumulants of X are equal to λ. The relation between moments and cumulants implies

$$B_n(\lambda) = \sum_{E\in\mathcal{E}_n} \lambda^{\#E},$$

where $\#E$ is the number of blocks of partition E. Particularly, the Bell number

$$B_n = B_n(1) = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!},$$

which is known as Dobinski's Formula ([59]).

0.1.4 Partition Path

In many cases, we use a partition of $[n]$ to describe the homogeneous relationship among a finite number of objects (see Section 3.2.2) . If additional objects need to be considered, we also need to use partitions of $[n+1]$, $[n+2]$, and so on. In other words, we need to consider a sequence of partitions.

To make the sequence of partitions consistent, we insist that the partition of $[n]$ in the sequence can be embedded into the partition of $[n+1]$ and so on. That is, the homogeneous relationship among the first n objects remains the same after adding the $(n+1)$th object and so on.

In general, we call a sequence of partitions $\{E_n\}_{n=1,2,\ldots}$ a *partition path*, or a *partition-valued path*, if

(i) E_n is a partition of $[n] = \{1, 2, \ldots, n\}$, and

(ii) E_n is the restriction of E_{n+1} to $[n]$,

for each n. The "restriction of E_{n+1} to $[n]$" indicates deleting the item $n+1$ from E_{n+1} in the original form (1) or (2) of partitions. For example, the restriction of

partition $13|2$ to $[2]$ is $1|2$. If we denote by $\pi_{[n]}$ the restriction operation from $\mathcal{E}_{n+1}$ to $\mathcal{E}_n$, then $\pi_{[2]}(13|2) = 1|2$. In the matrix form (3), the restriction indicates deleting both the $(n+1)$th row and the $(n+1)$th column. An example of a partition path is

$$\begin{aligned} E_1 &= 1, \\ E_2 &= 1|2, \\ E_3 &= 13|2, \\ E_4 &= 13|2|4, \\ E_5 &= 13|2|45, \quad \ldots \end{aligned}$$

Indeed, a partition path $E = \{E_n\}_{n=1,2,\ldots}$ can also be regarded as a partition of the set of natural numbers $\mathcal{N} = \{1, 2, 3, \ldots, n, \ldots\}$. The restriction of E to $[n]$ is E_n. From this point of view, a partition path E is not only a sequence of partitions $\{E_n\}_n$, but also a set of partitions $\{E_S\}_S$, where S runs all finite subset of $\mathcal{N}$ and E_S is the restriction of E to S.

0.2 Integer Partition

0.2.1 Partition of a Natural Number

A *partition of a natural number* n is a way of writing n as a sum of positive integers and without regard to their order. By convention, a partition of n is normally written from the largest to the smallest addend ([4], [60]).

For example, for $n = 4$, there are totally 5 different integer partitions:

$$4, \quad 3+1, \quad 2+2, \quad 2+1+1, \quad 1+1+1+1, \tag{6}$$

which can also be written in *frequency representation* ([4], [60]) as

$$1^0 2^0 3^0 4^1, \quad 1^1 2^0 3^1 4^0, \quad 1^0 2^2 3^0 4^0, \quad 1^2 2^1 3^0 4^0, \quad 1^4 2^0 3^0 4^0.$$

Here the indices of 1, 2, 3, or 4 indicate the corresponding frequencies in integer partitions.

In general, a partition of n can be written as

$$1^{\alpha_1} 2^{\alpha_2} \cdots n^{\alpha_n},$$

with α_j indicating exactly α_j occurrences of j's in the integer partition. So

$$\sum_{j=1}^{n} j \ \alpha_j = n.$$

0.2.2 Partition Function

A natural question is how many integer partitions there are for each n. In the mathematical literature, the number of partitions of n is called *partition function*, denoted by $P(n)$ or $p(n)$ ([61]).

Just like computing Bell numbers, there is no simple formula to calculate $P(n)$ either. Instead, we first calculate $P(n,k)$, which is the number of partitions of n containing exactly k terms. Then

$$P(n) = \sum_{k=1}^{n} P(n,k).$$

To compute $P(n,k)$, we may use the recursive relation ([50], [61])

$$P(n,k) = P(n-1,k-1) + P(n-k,k), \tag{7}$$

with $P(n,k) = 0$ for $k > n$, $P(n,n) = 1$, and $P(n,0) = 0$. Following a similar argument as in (5), the equation (7) can be derived by classifying the k-term partitions into two groups according to their smallest addends in the original form (6). Those k-term partitions containing addend 1 form group one. The other k-term partitions form group two. There is a one-to-one correspondence between group one and the $(k-1)$-term partitions of $n-1$ if one addend 1 is deleted from the k-term partition. Similarly, a one-to-one correspondence between the partitions in group two and the k-term partitions of $n-k$ is established if each addend in the former is reduced by 1. So $P(n,k)$ is the sum of $P(n-1,k-1)$ and $P(n-k,k)$. Based on (7), it is convenient and practically useful to set up the triangle of $P(n,k)$ by hand ([61])

$$
\begin{array}{ccccccc}
1 & & & & & & \\
1 & 1 & & & & & \\
1 & 1 & 1 & & & & \\
1 & 2 & 1 & 1 & & & \\
1 & 2 & 2 & 1 & 1 & & \\
1 & 3 & 3 & 2 & 1 & 1 & \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots
\end{array}
$$

Here the entry in the nth row and the kth column is $P(n,k)$. The sum of the entries in the nth row is $P(n)$. To see how fast $P(n)$ increases with n, we list the first 20 values of $P(n)$ in Table 2 ([37]). More precisely ([26], [61]),

$$P(n) \sim \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{2n/3}}, \text{ as } n \to \infty.$$

Table 2: First 20 Values of Partition Function $P(n)$

n	1	2	3	4	5	6	7	8	9	10
$P(n)$	1	2	3	5	7	11	15	22	30	42
n	11	12	13	14	15	16	17	18	19	20
$P(n)$	56	77	101	135	176	231	297	385	490	627

0.2.3 From Set Partition to Integer Partition

Given a set partition of $[n]$, there is a corresponding integer partition of n if we ignore the difference among the n subjects and count the block sizes only. For example, given the set partition $1|23$, the corresponding integer partition is $2+1$ in its original form where $\{2,1\}$ are the blocks sizes of $1|23$. Denote by π_I the mapping from set partitions to the corresponding integer partitions via block sizes of the former. Then $\pi_I(1|23) = 2+1$ or $1^1 2^1 3^0$.

As in Section 0.1.2, denote by $\mathcal{E}_n$ the set of partitions of $[n]$. Furthermore, denote by ϖ_n the set of partitions of n ([31]). Evidently, the mapping π_I from $\mathcal{E}_n$ to ϖ_n is onto. That is, for each $e_n \in \varpi_n$, there exists an $E_n \in \mathcal{E}_n$ such that $\pi_I(E_n) = e_n$. Indeed, given the frequency representation $e_n = 1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}$, the number of E_n's satisfying $\pi_I(E_n) = e_n$ is ([44])

$$N(\alpha_1, \alpha_2, \ldots, \alpha_n) = \frac{n!}{\prod_{j=1}^{n}(j!)^{\alpha_j}\ \alpha_j!}.$$

CHAPTER 1

PARTITION DISTRIBUTION

In this chapter, we first review two classes of probability distributions on set partitions, the exchangeable type and the product type. To construct a partition-valued process, we need a non-inference assumption or Kolmogorov consistency. Then we reveal that self-similarity is the characteristic property of a partition-valued process of the product type. Finally, our discussion leads to the Ewens family, the partition-valued processes that are both exchangeable and self-similar.

1.1 Exchangeable Partition Distribution

As in Section 0.1.2, we denote by $\mathcal{E}_n$ the set of all partitions of $[n]$. If a probability distribution on $\mathcal{E}_n$ serves as a prior for uncertain homogeneous relationships among n subjects, the principle of egalitarianism or symmetry requires that the partition distribution should remain invariant under permutations of $[n]$. Such a distribution on partitions is called *finitely exchangeable*, or *exchangeable* in short.

For example, given an exchangeable partition distribution P on $\mathcal{E}_3$, then $P(1|23) = P(2|13) = P(3|12)$, because $\sigma_{12}(1|23) = 2|13$, $\sigma_{23}(2|13) = 3|12$, where $\sigma_{12} = \binom{1\ 2\ 3}{2\ 1\ 3}$, $\sigma_{23} = \binom{1\ 2\ 3}{1\ 3\ 2}$ are permutations of [3]. Notice that the set partitions 1|23, 2|13 and

$3|12$ have the same set of block sizes $\{1, 2\}$. Following the notation in Section 0.2.3, $\pi_I(1|23) = \pi_I(2|13) = \pi_I(3|12)$, where π_I is the mapping from set partitions to integer partitions.

In general, given any two partitions E_1 and E_2 of $[n]$, there exists a permutation σ of $[n]$ such that $\sigma(E_1) = E_2$ if and only if E_1 and E_2 have the same set of block sizes. Therefore the set of block sizes or $\pi_I(\cdot)$ is the maximal invariant under permutations given only P is exchangeable.

Proposition 1.1 *A partition distribution P on $\mathcal{E}_n$ is exchangeable if and only if there exists a function g defined on the set of integer partitions ϖ_n such that*

$$P(E) \propto g \circ \pi_I(E). \tag{1.1}$$

Denote by $B_1, B_2, \ldots, B_k$ the blocks of partition E of $[n]$. Let $|B_1|, |B_2|, \ldots, |B_k|$ be the corresponding block sizes. Then the function g in (1.1) is indeed a function of the set of block sizes $\{|B_1|, |B_2|, \ldots, |B_k|\}$ ([46]).

Example 1.1 : Gibbs partition distribution *A probability distribution P on $\mathcal{E}_n$ is called a Gibbs partition distribution ([46]) if there exist two sequences of non-negative real numbers $\{v_i\}_{i=1,\ldots,n}$ and $\{w_i\}_{i=1,\ldots,n}$ such that*

$$P(E = B_1|B_2|\ldots|B_k) \propto v_k \prod_{i=1}^{k} w_{|B_i|} \, . \tag{1.2}$$

The Gibbs partition distribution is exchangeable based on Proposition 1.1 .

1.2 Product Partition Distribution

To combine data from different sources, Hartigan ([27]) developed a method for constructing probability models by means of random partitions, known as *product partition model*. The model comes with partition distribution P of the product type:

$$P(E = B_1|B_2|\dots|B_k) \propto \prod_{i=1}^{k} c(B_i), \tag{1.3}$$

where $B_1, B_2, \dots, B_k$ are blocks of the partition E, $c(\cdot)$ is a *cohesion* which attaches a non-negative real number $c(B)$ to each subset B of $[n]$. Such a partition distribution is called a *product partition distribution*. Note that different cohesions in (1.3) may lead to the same partition distribution. Indeed, for any sequence of positive real numbers $\{a_i\}_{i=1,\dots,n}$, the cohesions $c(\cdot)$ and $c(\cdot)\prod_{i\in\cdot} a_i$ determine the same P by (1.3). Thus we may always assume $c(\{i\}) = 1$ for $i = 1, \dots, n$.

A product partition distribution is not necessary exchangeable. For example, if

$$c(B) = \begin{cases} 2, & \text{if } B = \{2, 3\}; \\ 1, & \text{otherwise,} \end{cases}$$

then the corresponding P on $\mathcal{E}_3$ satisfies $P(1|23) = 1/3$ while $P(2|13) = 1/6$.

It is interesting to see what product partition distributions are exchangeable. Indeed, if P with cohesion c is exchangeable, then $c(B_1) = c(B_2)$ whenever $|B_1| = |B_2|$. In other words, the cohesion c is determined by a sequence of non-negative real

numbers $\{w_i\}_{i=1,2,\ldots}$. Thus,

Proposition 1.2 *Given an exchangeable partition distribution P on $\mathcal{E}_n$, P is of product type if and only if there exists a sequence of non-negative real numbers $\{w_i\}_{i=1,2,\ldots}$ such that*

$$P(E = B_1|B_2|\ldots|B_k) \propto \prod_{i=1}^{k} w_{|B_i|}. \tag{1.4}$$

Note that a partition distribution satisfying (1.4) is a special case of the Gibbs partition distribution. Letting $w_i \equiv \lambda$ for some $\lambda > 0$, (1.4) leads to a distribution on partitions as follows

Example 1.2 : Exponential family on partitions *A set of probability distributions $\{p_n(\cdot;\lambda),\ \lambda > 0\}$ on $\mathcal{E}_n$ is called the exponential family on partitions generated from the uniform distribution with canonical parameter $\theta = \log\lambda$ and canonical statistic equal to the number of blocks if*

$$p_n(E;\lambda) = \lambda^{\#E}/B_n(\lambda), \tag{1.5}$$

where E is a partition of $[n]$, $\#E$ is the number of blocks of E, $B_n(\lambda)$ is a Bell polynomial (see Section 0.1.3) .

Each partition distribution in the exponential family is of the product type and exchangeable. Note that it is uniform on partitions if $\lambda = 1$. The exponential family on $\mathcal{E}_n$ can be generated from the uniform distribution by exponential weighting with canonical statistic $\#E$.

1.3 Kolmogorov Consistency

1.3.1 Partition Process

Consider a sequence of random partitions $\{E_n\}_{n=1,2,\ldots}$, where E_n is a random partition of $[n]$ with distribution P_n on $\mathcal{E}_n$. Typically, we need the sequence of partition distributions $\{P_n\}_{n=1,2,\ldots}$ to be *consistent*. That is, we require that the restriction of P_{n+1} to $[n]$ be identical to P_n for each n. Following the notation in Section 0.1.4,

$$P_n(\cdot) = \sum_{x \in \pi_{[n]}^{-1}(\cdot)} P_{n+1}(x), \quad \text{for each } n. \tag{1.6}$$

For example, the restriction of P_3 to $[2]$ is identical to P_2 if and only if

$$\begin{aligned} P_2(12) &= P_3(123) + P_3(12|3), \\ P_2(1|2) &= P_3(13|2) + P_3(1|23) + P_3(1|2|3). \end{aligned}$$

If there exists a random partition path $E = \{E'_n\}$ (see Section 0.1.4) such that E'_n has the distribution P_n for each n, then (1.6) must be true. On the other hand, given a sequence of partition distributions $\{P_n\}$ satisfying (1.6), such a random partition path $E = \{E'_n\}$ can always be constructed. Indeed, if we denote the set of partitions of $\mathcal{N}$ by $\mathcal{E}_\infty$, then each partition of $[n]$ can be regarded as a subset of $\mathcal{E}_\infty$ specifying only the relationship among $[n]$. Therefore any subset of $\mathcal{E}_n$ can be embedded into $\mathcal{E}_\infty$. The union $\mathcal{A} = \cup_{k=1}^{\infty} 2^{\mathcal{E}_k}$ containing every subset of $\mathcal{E}_n$ for each n is a *field* which

is closed under complement and finite union. The existence of a probability measure P on $(\mathcal{E}_\infty, \mathcal{A})$ is ensured by the consistency condition (1.6). By the extension theorem ([9], Theorem 3.1) in measure theory, P can be extended uniquely to the σ-field $\mathcal{F}$ generated by $\mathcal{A}$. Thus,

Proposition 1.3 *Let $\{P_n\}_{n=1,2,...}$ be a sequence of partition distributions, where P_n is a distribution on $\mathcal{E}_n$ for each n. Then $\{P_n\}$ satisfies the consistency condition (1.6) if and only if there exists a random partition path $E = \{E'_n\}$ such that the random partition E'_n has distribution P_n for each n.*

In other words, the consistency condition (1.6) guarantees that a sequence of random partitions $\{E_n\}$ has an equivalent-in-probability modification $\{E'_n\}$ which also serves as a partition-valued process, or a partition process in short. It is similar in spirit to the Kolmogorov consistency for the set of finite-dimensional distributions for real-valued processes. So the condition (1.6) is called *Kolmogorov consistency* too.

Note that Proposition 1.3 is only a special case of a more general result as follows. As long as there is a sequence of onto mappings $\pi_n : \mathcal{E}_{n+1} \to \mathcal{E}_n$ such that $P_n = P_{n+1}\pi_n^{-1}$, we can construct a partition-valued Markov chain $\{E'_n\}$ such that the random partition E'_n of $[n]$ has the the distribution P_n.

For statistical problems involving random partitions, Kolmogrov consistency is often an essential assumption. It indicates that the same probabilistic statements hold for an extended trial including additional subjects, known as *non-interference*.

Example 1.2 : Exponential family on partitions*(Continued) The Bell polynomials* $B_2(\lambda) = \lambda^2 + \lambda$, $B_3(\lambda) = \lambda^3 + 3\lambda^2 + \lambda$. *Then*

$$\begin{aligned} p_2(12;\lambda) &= \frac{\lambda}{B_2(\lambda)} = \frac{1}{\lambda+1}\,, \\ p_3(123;\lambda) &= \frac{\lambda}{B_3(\lambda)} = \frac{1}{\lambda^2+3\lambda+1}\,, \\ p_3(12|3;\lambda) &= \frac{\lambda^2}{B_3(\lambda)} = \frac{\lambda}{\lambda^2+3\lambda+1}\,. \end{aligned}$$

Note that for each $\lambda > 0$, the exponential family $\{p_n(\cdot;\lambda)\}_n$ is not Kolmogorov consistent. Particularly, the set of uniform distributions $\{p_n(\cdot;\lambda=1)\}_n$ on partitions doesn't determine a process. In fact, if $\{p_n(\cdot;\lambda)\}_n$ satisfies the Kolmogorov consistency, then

$$p_2(12;\lambda) = p_3(123;\lambda) + p_3(12|3;\lambda)\,,$$

which implies $\lambda = 0$. It can't be true because $p_n(E;\lambda) = \lambda^{\#E}/B_n(\lambda)$ is a probability distribution.

Following the discussion in Section 0.1.4, any partition path $\{E'_n(\omega)\}_n$ also indicates a set of partitions $\{E'_S(\omega)\}_S$. Therefore, a partition process $\{E'_n\}_n$ indexed by n induces a partition process $\{E'_S\}_S$ indexed by nonempty finite subsets S of $\mathcal{N}$, since $\{E'_S\}_S$ satisfies the Kolmogorov consistency too. That is, E'_S has the same distribution as the restriction of $E'_{S'}$ to S given any pair of finite subsets $S \subset S'$.

If $\{P_n\}$ or $\{P_S\}_S$ satisfies both Kolmogorov consistency and exchangeability, it is called *infinitely exchangeable* ([46]).

1.3.2 Permutation Process

A permutation of $[n]$ is a one-to-one correspondence from $[n]$ to itself. It can be represented uniquely as a product of permutation cycles up to the order of cycles ([62]). For example, the permutation $\sigma_6 = \binom{1\ 2\ 3\ 4\ 5\ 6}{1\ 5\ 4\ 3\ 6\ 2}$ can be rewritten as (1)(256)(34).

Let $\mathcal{S}_n$ be the set of permutations of $[n]$. Define a mapping ϕ_n^* from $\mathcal{S}_{n+1}$ to $\mathcal{S}_n$ such that, for each permutation $\sigma \in \mathcal{S}_{n+1}$ and $i = 1, \ldots, n$,

$$(\phi_n^* \sigma)(i) = \begin{cases} \sigma^2(i), & \text{if } \sigma(i) = n+1; \\ \sigma(i), & \text{otherwise.} \end{cases}$$

For example, $\phi_5^*((1)(256)(34)) = (1)(25)(34)$. So $\phi_n^*(\sigma)$ is the permutation of $[n]$ derived by removing $n+1$ from σ in its cyclic form. Furthermore, $(\phi_{n-1}^* \circ \phi_n^*)(\sigma) = \phi_{n-1}^*(\phi_n^*(\sigma))$ is the permutation of $[n-1]$ derived by removing $n, n+1$ from σ.

Similar to the definition of partition path (see Section 0.1.4), a sequence of permutations $\{\sigma_n\}_n$ is called a *permutation path* if $\phi_n^*(\sigma_{n+1}) = \sigma_n$ for each $n \geq 1$, where σ_n is a permutation of $[n]$. Let $\{P_n\}_n$ be a set of probability distributions on permutations, where P_n is defined on $\mathcal{S}_n$. Then $\{P_n\}_n$ is called *Kolmogorov consistent* if for each n,

$$P_n(\cdot) = \sum_{\sigma \in (\phi_n^*)^{-1}(\cdot)} P_{n+1}(\sigma)\,. \tag{1.7}$$

For example, the consistency of $\{P_n\}_n$ implies

$$P_3((13)(2)) = P_4((143)(2)) + P_4((134)(2)) + P_4((13)(24)) + P_4((13)(2)(4)).$$

Following a similar argument as in Proposition 1.3, we conclude that

Proposition 1.4 *Let $\{P_n\}_{n=1,2,...}$ be a sequence of permutation distributions, where P_n is defined on $\mathcal{S}_n$ for each n. Then $\{P_n\}_n$ satisfies the Kolmogorov consistency (1.7) if and only if there exists a random permutation path $\{\Sigma_n\}_n$ such that the random permutation Σ_n has the distribution P_n.*

Let $\{\Sigma_n\}_n$ be a sequence of random permutations with distribution functions $\{P_n\}_n$. Based on Proposition 1.4, if $\{P_n\}_n$ satisfies the Kolmogorov consistency, then there exists a sequence of random permutations $\{\Sigma'_n\}_n$ defined on a common probability space $(\Omega, \mathcal{F}, P)$ such that $\{\Sigma'_n(\omega)\}_n$ is a permutation path for each $\omega \in \Omega$ and Σ'_n has the distribution function P_n for each n.

In general, the Kolmogorov consistency may be defined for the set of permutation distributions $\{P_B\}_B$, where B runs through all possible finite subsets of an index set $\mathcal{I}$. For example, $\mathcal{I} = \mathcal{N}$. For any two finite subsets B and B' such that $B \subseteq B'$, the insertion $\phi : B \to B'$ induces a mapping ϕ^* from $\mathcal{S}_{B'}$ to $\mathcal{S}_B$ similar to the mapping ϕ_n^* from $\mathcal{S}_{n+1}$ to $\mathcal{S}_n$. So the Kolmogorov consistency requires

$$P_B(\cdot) = \sum_{\sigma \in (\phi^*)^{-1}(\cdot)} P_{B'}(\sigma).$$

Fortunately, the subsequence $\{P_{[n]}\}_n$ is enough to determine $\{P_B\}_B$ uniquely as long as the Kolmogorov consistency is satisfied. Therefore, we only need to consider the sequence $\{P_n\}_n = \{P_{[n]}\}_n$ in the case $\mathcal{I} = \mathcal{N}$.

Example 1.3 : Exponential families on permutations *([38]) The set of probability distributions*

$$p_n(\sigma;\lambda) = \lambda^{\#\sigma}\Gamma(\lambda)/\Gamma(n+\lambda),\ \lambda > 0 \tag{1.8}$$

on $\mathcal{S}_n$ is called the exponential family on permutations generated from the uniform distribution with canonical parameter $\theta = \log\lambda$ and canonical statistic equal to the number of cycles, where σ is a permutation of $[n]$, and $\#\sigma$ is the number of cycles of σ .

For each $\lambda > 0$, the set of permutation distributions $\{p_n(\cdot;\lambda)\}_n$ is Kolmogorov consistent. In fact, for each permutation σ of $[n]$,

$$\begin{aligned}
\sum_{\sigma_{n+1}\in(\phi_n^*)^{-1}(\sigma)} p_n(\sigma_{n+1};\lambda) &= n\cdot\lambda^{\#\sigma}\frac{\Gamma(\lambda)}{\Gamma(n+1+\lambda)} \ +\ \lambda^{\#\sigma+1}\frac{\Gamma(\lambda)}{\Gamma(n+1+\lambda)} \\
&= \lambda^{\#\sigma}\cdot(n+\lambda)\cdot\frac{\Gamma(\lambda)}{(n+\lambda)\Gamma(n+\lambda)} \\
&= \lambda^{\#\sigma}\frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \\
&= p_n(\sigma;\lambda)\,.
\end{aligned}$$

If $n = 1$, (1.8) is a probability distribution on $\mathcal{S}_1$. For $n > 1$, the Kolmogorov consistency (1.7) ensures that (1.8) is a probability distribution on $\mathcal{S}_n$ too.

Note that the cycles of a permutation of $[n]$ determine a partition of $[n]$. Let π_n be the mapping from $\mathcal{S}_n$ to $\mathcal{E}_n$ changing cycles into blocks. For example, $\pi_6((1)(256)(34))$ $= 1|256|34$. The mappings are onto. Then a random permutation induces a random partition, and a permutation process induces a partition process. For example, the partition process induced by (1.8) is called the *Ewens process* (see Section 1.5).

Similar to the exchangeability for partitions, we introduce the exchangeability for permutations. A permutation distribution P_n on $\mathcal{S}_n$ is called *exchangeable* if, for each permutation σ of $[n]$,

$$P_n(\sigma_1) = P_n(\sigma\sigma_1\sigma^{-1}), \text{ for each } \sigma_1 \in \mathcal{S}_n.$$

In the cyclic form of permutations, the conjugate mapping $g_\sigma : \cdot \to \sigma \cdot \sigma^{-1}$ keeps the cycle structure invariant and permutes the orders of items only. For example, if $\sigma = (1234)(5)(6)$, $\sigma_1 = (1235)(46)$, then

$$g_\sigma(\sigma_1) = \sigma\sigma_1\sigma^{-1} = (\sigma(1)\sigma(2)\sigma(3)\sigma(5))(\sigma(4)\sigma(6)) = (2345)(16).$$

Note that the permutations also have matrix representations. Indeed, a permutation $\sigma \in \mathcal{S}_n$ can be represented as a *permutation matrix* $\Sigma = (\Sigma_{ij})_{n\times n}$, such that $\Sigma_{ij} = 1$ if $j = \sigma(i)$ and 0 otherwise. A permutation matrix is always orthogonal ([6]). Given two permutations $\sigma, \sigma_1 \in \mathcal{S}_n$ with corresponding permutation matrices Σ, Σ_1, the composition $\sigma\sigma_1$ is equivalent to the matrix multiplication $\Sigma_1\Sigma$ in reverse order.

So $g_\sigma(\sigma_1)$ indicates permuting the rows and columns of Σ_1 simultaneously according to σ. In the matrix form, a random permutation is exchangeable if and only if the corresponding random permutation matrix remains invariant under simultaneous row and column permutations. It is similar to the case of exchangeable random partitions in matrix forms. However, to construct a permutation process in matrix form, the mapping ϕ_n^* from $\mathcal{S}_{n+1}$ to $\mathcal{S}_n$ is not simply deleting the $(n+1)$th row and column of the permutation matrices. Instead, $(\phi_n^*\Sigma)_{ij} = \max\{\Sigma_{ij}, \Sigma_{i,n+1}\Sigma_{n+1,j}\},\ i,j = 1,\ldots,n$.

An exchangeable random permutation induces an exchangeable random partition, and an exchangeable permutation process induces an exchangeable partition process. For example, the exponential families on permutations are exchangeable. So the induced partition distribution *Ewens sampling distribution* (see Section 1.5) is exchangeable too.

On the other hand, given an exchangeable partition process, it is simple to generate an exchangeable permutation process by assigning the probability of each partition E uniformly on the set of permutations $\pi_n^{-1}(E)$. For example, if the probability of a partition $E = B_1|B_2|\cdots|B_k$ of $[n]$ is p, then any permutation σ of n satisfying $\pi_n(\sigma) = E$ is assigned the probability $p/\prod_{i=1}^k(|B_i|-1)!$.

1.4 Self-Similarity

For each finite subset S of $\mathcal{N}$, denote by $\mathcal{E}_S$ the set of partitions of S. Let $\{P_S\}_S$ be a set of partition distributions, where S runs all nonempty finite subsets of $\mathcal{N}$ and P_S

is a distribution on $\mathcal{E}_S$. Then $\{P_S\}_S$ is called *self-similar* if, for all non-overlapping, nonempty, finite subsets S, S' of $\mathcal{N}$,

$$P_{S\cup S'}(E|E') = \pi(S|S')P_S(E)P_{S'}(E'), \tag{1.9}$$

where E and E' are partitions of S and S' respectively, $E|E'$ is the partition of $S\cup S'$ collecting simply all the blocks of E and E', $\pi(\cdot)$ is the cumulative probability function in the sense of partition lattice (see Section 0.1.2) . Specifically,

$$\pi(S|S') = \sum_{E''\leq S|S'} P_{S\cup S'}(E'') = \sum_{E\leq S,\ E'\leq S'} P_{S\cup S'}(E|E'). \tag{1.10}$$

If $\{P_S\}_S$ is self-similar, then for any non-overlapping S_1, S_2, S_3,

$$P_{S_1\cup S_2\cup S_3}(E_1|E_2|E_3) \ \propto\ P_{S_1}(E_1)P_{S_2\cup S_3}(E_2|E_3) \ \propto\ P_{S_1}(E_1)P_{S_2}(E_2)P_{S_3}(E_3),$$

where E_1, E_2, E_3 are partitions of S_1, S_2, S_3 respectively. So

$$P_{S_1\cup S_2\cup S_3}(E_1|E_2|E_3) = \pi(S_1|S_2|S_3)P_{S_1}(E_1)P_{S_2}(E_2)P_{S_3}(E_3),$$

which implies $\pi(S_1|S_2|S_3) = \pi(S_1|S_2\cup S_3)\pi(S_2|S_3)$. By induction,

Proposition 1.5 *If the set of partition distributions $\{P_S\}_S$ is self-similar, then for*

any positive integer k and any non-overlapping nonempty finite subsets $S_1, S_2, \ldots, S_k$,

$$P_{S_1 \cup \cdots \cup S_k}(E_1|\cdots|E_k) = \pi(S_1|\cdots|S_k) P_{S_1}(E_1) \cdots P_{S_k}(E_k), \tag{1.11}$$

where $E_1, \ldots, E_k$ are partitions of $S_1, \ldots, S_k$ respectively.

Self-similarity on partitions is a property similar in spirit as the lack-of-memory property of the exponential distribution on the real line. For any fixed subset S' and any fixed partition E' of S', self-similarity ensures that the conditional partition process below $S|E'$ behaves as the original process. In other words, once a subset S is identified as isolated from its coset, this property guarantees that no adjustment is needed after the coset is removed. In the literature, it is also known as *subset deletion* ([31], [3]).

Proposition 1.6 *Let $\{P_S\}_S$ be a set of partition distributions, where S runs all nonempty finite subsets of $\mathcal{N}$. Then $\{P_S\}_S$ is self-similar if and only if there exists a cohesion c associating a nonnegative real number to each nonempty finite subset of $\mathcal{N}$, such that P_S is of product type determined by the common c for each S.*

Proof **(1)** If there exists such a c independent of S, then $P_{S \cup S'}$, P_S and $P_{S'}$ are all of product type, where S and S' are two non-overlapping subsets of $\mathcal{N}$. For any partitions $E = B_1|B_2|\cdots|B_s$ of S, $E' = B'_1|B'_2|\cdots|B'_t$ of S',

$$P_{S \cup S'}(E|E') \;\propto\; c(B_1)c(B_2)\cdots c(B_s) \cdot c(B'_1)c(B'_2)\cdots c(B'_t),$$

$$P_S(E) \propto c(B_1)c(B_2)\cdots c(B_s),$$

$$P_{S'}(E') \propto c(B'_1)c(B'_2)\cdots c(B'_t),$$

Therefore, there exists a constant $K(S, S')$ determined only by S and S' such that

$$P_{S\cup S'}(E|E') = K(S, S') \cdot P_S(E)P_{S'}(E'). \tag{1.12}$$

Based on (1.10) and (1.12),

$$\pi(S|S') = \sum_{E,E'} P_{S\cup S'}(E|E') = K(S, S') \sum_{E} P_S(E) \sum_{E'} P_{S'}(E') = K(S, S').$$

So (1.9) is true and $\{P_S\}_S$ is self-similar.

(2) If $\{P_S\}_S$ is self-similar, a cohesion c can be constructed as follows:

(i) $c(\{i\}) = 1$, for each $i \in \mathcal{N}$;

(ii) $c(B) = P_B(B)/P_B(\|B\|)$, for each finite subset $B = \{i_1, \ldots, i_k\}$ of $\mathcal{N}$, where $\|B\|$ denotes the smallest partition $\{i_1\}|\{i_2\}|\cdots|\{i_k\}$ of B, and B itself can be regarded as the single-block partition of B.

Let $\{P^*_S\}_S$ be the set of partition distributions determined by c. The only thing left is to verify $P^*_S = P_S$ for each S. Indeed, for any partition $E = B_1|\cdots|B_k$ of S, $\|S\|$ is a sub-partition of E. By self-similarity and (1.11),

$$P_S(\|S\|) = \pi(B_1|\cdots|B_k)P_{B_1}(\|B_1\|)\cdots P_{B_k}(\|B_k\|),$$

$$\begin{aligned} P_S^*(E) &\propto \prod_{i=1}^{k} c(B_i) \\ &= \prod_{i=1}^{k} P_{B_i}^{-1}(\|B_i\|) P_{B_i}(B_i) \\ &= P_S^{-1}(\|S\|) \cdot \pi(B_1|\cdots|B_k) \prod_{i=1}^{k} P_{B_i}(B_i) \\ &\propto P_S(E) \end{aligned}$$

Note that both P_S^* and P_S are distributions on $\mathcal{E}_S$. Therefore $P_S^* = P_S$. #

Note that a self-similar set of partition distributions $\{P_S\}_S$ need not be Kolmogorov consistent or exchangeable. If each P_S is exchangeable, then by Proposition 1.2,

Corollary 1.1 *Let $\{P_S\}_S$ be a set of exchangeable partition distributions. Then $\{P_S\}_S$ is self-similar if and only if there exists a sequence of non-negative real numbers $\{w_n\}_{n=1,2,\ldots}$ such that for each S,*

$$P_S(B_1|B_2|\cdots|B_k) \propto \prod_{i=1}^{k} w_{|B_i|}, \tag{1.13}$$

where $B_1|B_2|\cdots|B_k$ is a partition of S.

1.5 Ewens Sampling Distribution

An interesting question is, what partition distributions $\{P_S\}_S$ satisfy Kolmogorov consistency, self-similarity and exchangeability at the same time?

Theorem 1.1 *([31]) Let $\{P_S\}_S$ be a set of exchangeable partition distributions. Suppose $0 < P_{\{12\}}(1|2) < 1$. Then $\{P_S\}_S$ is both self-similar and Kolmogorov consistent if and only if there exists a positive real number λ such that for each S,*

$$P_S(B_1|\cdots|B_k) = \frac{\Gamma(\lambda)\lambda^k}{\Gamma(|S|+\lambda)} \prod_{i=1}^{k} (|B_i|-1)!\,, \tag{1.14}$$

where $B_1|\cdots|B_k$ is a partition of S, and $\Gamma(\cdot)$ is the gamma function.

Proof **(1)** If $\{P_S\}$ is self-similar, then by Corollary 1.1, there exists a sequence of non-negative real numbers $\{w_n\}_n$ such that (1.13) is true.

Without any loss of generality, we assume $w_1 = 1$. Denote

$$\lambda = \frac{P_{\{12\}}(1|2)}{1 - P^{-1}_{\{12\}}(1|2)} = \frac{P_{\{12\}}(1|2)}{P_{\{12\}}(12)} = w_2^{-1}.$$

Since $0 < P_{\{12\}}(1|2) < 1$, λ is a positive real number. We claim that

$$w_n = \lambda^{1-n}(n-1)! \tag{1.15}$$

if $\{P_S\}_S$ satisfies Kolmogorov consistency. Evidently, (1.15) is true for $n = 1$ and 2. For $k \geq 2$,

$$P_{[k]}(12\cdots(k-1)k)/P_{[k]}(12\cdots(k-1)|k) = w_k/w_{k-1}.$$

Since $P_{[k]}(12\cdots(k-1)k) = P_{[k+1]}(12\cdots k(k+1)) + P_{[k+1]}(12\cdots k|(k+1))$,

$$P_{[k+1]}(12\cdots k(k+1))/P_{[k+1]}(12\cdots k|(k+1)) = w_{k+1}/w_k$$

implies $P_{[k+1]}(12\cdots k|(k+1)) = P_{[k]}(12\cdots(k-1)k)\cdot w_k/(w_{k+1}+w_k)$. On the other hand, $P_{[k+1]}(12\cdots k|(k+1)) = P_{[k]}(12\cdots(k-1)|k)\cdot w_k/(w_k+w_{k-1}w_2+w_{k-1})$ since

$$\begin{aligned} & P_{[k]}(1\cdots(k-1)|k) \\ = \; & P_{[k+1]}(1\cdots(k-1)(k+1)|k) \; + \; P_{[k+1]}(1\cdots(k-1)|k(k+1)) \\ & + P_{[k+1]}(1\cdots(k-1)|k|(k+1)). \end{aligned}$$

Thus, the Kolmogorov consistency implies

$$1 = w_k/w_{k-1}\cdot(w_k + w_{k-1}w_2 + w_{k-1})/(w_{k+1}+w_k).$$

If (1.15) is true for $k-1$ and k, then $w_{k+1} = \lambda^{-k}k!$, which indicates (1.15) is true for $k+1$ too. By induction, (1.15) is true for each n.

Based on (1.13), (1.15) and the exchangeability, there exists a sequence of real numbers $\{a_n\}_n$ such that

$$P_S(B_1|B_2|\cdots|B_k) = a_{|S|}\cdot\lambda^{k-|S|}\cdot\prod_{i=1}^{k}(|B_i|-1)!\,.$$

To prove (1.14), the only thing left is to verify that

$$a_n = \lambda^n \Gamma(\lambda)/\Gamma(n+\lambda). \tag{1.16}$$

It is straightforward to derive $a_1 = 1$, $a_2 = 1/(1+w_2) = \lambda/(1+\lambda)$. Since $\Gamma(1+x) \equiv x\Gamma(x)$ for $x > 0$, (1.16) is true for $n = 1$ and 2. If (1.16) is true for $n = k \geq 2$, then the Kolmogorov consistency implies $a_k w_k = a_{k+1} w_{k+1} + a_{k+1} w_k$ and

$$a_{k+1} = a_k \cdot \lambda/(k+\lambda) = \lambda^{k+1}\Gamma(\lambda)/\Gamma(k+1+\lambda).$$

By induction, (1.16) is true for each n.

(2) If $\{P_S\}_S$ is defined by (1.14), in other words, if it is determined by the sequence $w_n = \lambda^{1-n}(n-1)!$ as in (1.13), then $\{P_S\}_S$ is self-similar by Corollary 1.1.

To verify that $\{P_S\}_S$ is Kolmogorov consistent, it is sufficient to prove that for each S, each partition $B_1|\cdots|B_k$ of S and each u which does not belong to S,

$$P_S(B_1|\cdots|B_k) = P_{S\cup\{u\}}(B_1|\cdots|B_k|\{u\}) + \sum_{i=1}^{k} P_{S\cup\{u\}}(B_1|\cdots|B_i\cup\{u\}|\cdots|B_k),$$

which is straightforward based on (1.14).

Therefore Theorem 1.1 is proved. #

Given the set of partition distributions $\{P_S\}_S$ satisfying Kolmogorov consistency, it is sufficient to consider a subset $\{P_n\}_n$ instead, where $P_n = P_{[n]}$ for each n. In

those cases, $\{P_S\}_S$ or $\{P_n\}_n$ is called a partition process (see Section 1.3) . For example, the partition process defined by (1.14) can be represented by a sequence of partition distributions as follows

$$p_n(E;\ \lambda) = \frac{\Gamma(\lambda)\lambda^{\#E}}{\Gamma(n+\lambda)} \prod_{\text{blocks}} (b-1)!\ , \tag{1.17}$$

where $\lambda > 0$, $\#E$ is the number of blocks of E, the product runs over the blocks of E, b is a block size.

The partition distribution (1.17) is called the *Ewens sampling distribution*, named after Warren J. Ewens ([19], [20]). It was first proposed for population genetics purposes. Consider a sample of n gametes taken from a population and classified into blocks according to the alleles at a certain locus. Under suitable conditions ([31]) including moderate and non-recurrent mutations and negligible selection effect at the locus, etc, the random partition E based on genotype follows the Ewens sampling distribution.

The Ewens sampling distribution is also related to the urn models widely used in physics. For example, Costantini ([16]) considered the case where balls are sequentially put into m urns as follows. If n balls have been placed into the urns and the jth urn contains n_j balls, then the $(n+1)$th ball is put into the jth urn with probability $(n_j+\delta)/(n+m\delta)$. The distribution (1.17) can be derived as a limit if $\delta \to 0$, $m \to \infty$ and $m\delta = \lambda$ ([29], Chapter 41). On the other hand, the classical Maxwell-Boltzmann,

Bose-Einstein and Fermi-Dirac partition formulae correspond to $\delta \to \infty$, $\delta = 1$ and $\delta = -1$ respectively.

The Ewens sampling distribution is of the full exponential-family type with canonical statistic $\#E$. The canonical parameter is $\theta = \log \lambda$, and the cumulant function is $\log \Gamma(n + \lambda) - \log \Gamma(\lambda)$. Writing $\psi(\cdot) = (\log \Gamma(\cdot))'$, the expected number of blocks is $E(\#E) = \lambda(\psi(n + \lambda) - \psi(\lambda))$, which behaves asymptotically as $\lambda \log n$ for large n. The variance is a little smaller than the mean, and the number of blocks is roughly Poisson for large n. For $\lambda = 1$, the probability of one block is n^{-1}, and the probability of n blocks is $1/n!$. Basically, greater λ indicates that the random partition tends to have more blocks. For population genetics purpose, the parameter λ indicates the level of mutation rate up to a constant depending on the reproductive mechanism.

Consider extreme cases of Theorem 1.1: $P_{\{12\}}(1|2) = 0$ or 1. If $P_{\{12\}}(1|2)$ goes to 0, then $\lambda = P_{\{12\}}(1|2)/P_{\{12\}}(12)$ goes to 0. The limit distribution P_S based on (1.14) satisfies $P_S(S) = 1$. In other words, the partition distribution P_S puts mass 1 on the maximal partition with only 1 block. Similarly, if $P_{\{12\}}(1|2) = 1$ which indicates $\lambda = \infty$, then the limit distribution P_S puts mass 1 on the minimal partition $\|S\|$.

Corollary 1.2 *Let $\{P_S\}_S$ be a set of exchangeable partition distributions.*

(i) If $P_{\{12\}}(1|2) = 0$, then $\{P_S\}_S$ is both self-similar and Kolmogorov consistent if and only if $P_S(S) = 1$ for each S.

(ii) If $P_{\{12\}}(1|2) = 1$, then $\{P_S\}_S$ is both self-similar and Kolmogorov consistent if and only if $P_S(\|S\|) = 1$ for each S.

The cases (i) and (ii) in Corollary 1.2 can be regarded as two extreme cases of the Ewens sampling distribution corresponding to $\lambda = 0$ and $\lambda = \infty$ respectively. The only exchangeable and self-similar partition process is the extended Ewens family with $0 \leq \lambda \leq \infty$. Kingman ([31]) derived the same conclusion under the terminology *partition structure* and *subset deletion* (see also [3]).

CHAPTER 2

SAMPLING FROM A PARTITION DISTRIBUTION

In this chapter, we first review the Chinese Restaurant process which generates independent, identically distributed samples from the Ewens sampling distribution. Then we describe the so-called cocktail process on partitions with unique stationary distribution that belongs to the exponential family. As a variant of the cocktail process, we propose the Ewens-cocktail process that converges to the Ewens sampling distribution. It could be used for Markov chain monte carlo if combined with the Metropolis-Hastings algorithm. Because the number of partitions of $[n]$ increases even faster than e^n, it's practically important to generate a random sample of partitions conveniently and efficiently.

2.1 Sequential Construction: Chinese Restaurant Process

Assume a restaurant has infinitely many tables numbered from 1 to ∞. Each table is capable of seating infinitely many customers. Suppose customers arrive one by one and are seated according to the following rule:

1) The 1st customer sits at table 1;

2) After the first n customers have occupied the first k tables with n_i customers sitting at table i, $i = 1, 2, \ldots, k$, the $(n+1)$th customer randomly chooses one of the first $k+1$ tables with probabilities proportional to

$$n_1 \ : \ n_2 \ : \ \cdots \ : \ n_k \ : \ \lambda \, , \tag{2.1}$$

where $\lambda > 0$ is a parameter.

If the labels of those tables are ignored, the determined partition process of customers with blocks indicated by tables is known as the *Chinese restaurant process*. It was devised by Dubins and Pitman and published first on page 92 in [2]. The name came from some Chinese restaurants in San Francisco which seem to be capable of seating infinitely many customers ([10]).

It is straightforward to verify by induction that the Chinese restaurant process follows the Ewens sampling distribution with parameter λ ([43]). This fact provides an alternative way to derive some important properties of the Ewens family. For example, the Chinese restaurant process shows clearly that the probability that the first two customers sit both at table 1 is $1/(1+\lambda)$. By exchangeability, the probability that any two customers sit at the same table is also $1/(1+\lambda)$, no matter where the other customers sit. Besides, the Chinese restaurant process suggests an easy way to generate independent, identical distributed random partitions from the Ewens family.

If the proportions in (2.1) are changed into

$$(n_1 - \delta) \ : \ (n_2 - \delta) \ : \ \cdots \ : \ (n_k - \delta) \ : \ (\lambda + k\delta)\,,$$

where $0 \leq \delta < 1$ and $\lambda > -\delta$, then the determined partition process is called the *two-parameter generalization of the Ewens sampling distribution* ([43]). The corresponding partition distribution function defined on $\mathcal{E}_n$ is

$$p_n(E; \lambda, \delta) = \frac{\Gamma(\lambda + 1 - \delta) \prod_{j=0}^{\#E-1} (\lambda + j\delta)}{\Gamma(\lambda + n - \delta)\Gamma(1 - \delta)^{\#E}[\lambda + (\#E - 1)\delta]} \prod_{\text{blocks}} \Gamma(b - \delta), \qquad (2.2)$$

which yields (1.17) as a special case if $\delta = 0$. Evidently, the partition process determined by (2.2) is exchangeable too.

If the cyclic order of customers sitting at the same table is taken into accounts, the Chinese restaurant process generates an exchangeable permutation process which belongs to the exponential family on permutations. Here the exchangeability of a permutation process indicates that the distributions remain invariant under conjugate symmetric group operations (see Section 1.3.2 for more details). Suppose there is always an empty seat between any two occupied ones. The $(n + 1)$th customer in the Chinese restaurant process chooses a seat at random between any two adjacent customers or at a table occupied by a single customer with probability $1/(n + \lambda)$, and sits alone at the next unoccupied table with probability $\lambda/(n + \lambda)$. Since each occupied table with its cyclic order of customers indicates a cycle of permutation,

the seating plan of the first m customers generates a random permutation. It is straightforward to verify that the random permutation belongs to the exponential family. The distribution is

$$p_n(\sigma;\lambda) = \lambda^{\#\sigma}\Gamma(\lambda)/\Gamma(n+\lambda),$$

where $\#\sigma$ indicates the number of cycles of σ.

2.2 Markov Chain Monte Carlo: Cocktail Process

In this section, we describe a partition-valued Markov chain called the *cocktail process*, which has the unique stationary distribution belonging to the exponential family (see Example 1.2).

> Guests at a cocktail party numbered from 1 to n arrange themselves in conversational blocks, which determine a partition of the n individuals present. At regular time intervals, a transition occurs from the partition E to another partition E' as follows. An individual u chosen uniformly at random splits off from the block to which he or she belongs. Suppose there are k blocks left formed by the other $n-1$ guests. The individual u joins one of the k blocks with probability $1/(\lambda+k)$ for each block, or strikes out on his or her own and forms a new block with probability $\lambda/(\lambda+k)$, where λ is a positive real number.

Proposition 2.1 *The cocktail process has the unique stationary distribution*

$$p_n(E;\lambda) = \lambda^{\#E}/B_n(\lambda)\,,$$

which satisfies the detailed balance condition

$$p_n(E;\lambda)P(E,E') = p_n(E';\lambda)P(E',E)\,, \tag{2.3}$$

where E, E' are partitions of $[n]$, $B_n(\lambda) = \sum_{k=1}^{n} S(n,k)\lambda^k$ is known as the Bell polynomial, $S(n,k)$ is the Stirling number of the second kind, and $P(E,E')$ is the transition probability from E to E'.

Proof Evidently, the cocktail process is both aperiodic and irreducible (see [55] or [15] for a good review on Markov chains). Since it has only finitely many states, the cocktail process has unique stationary distribution.

The only thing left is to prove (2.3), which implies $p_n(\cdot;\lambda)$ is the unique stationary distribution. Note $P(E,E') > 0$ only if there exist two integers $0 \le i, i' < n$ and one partition $E_\wedge$ formed by the $n-1$ guests other than u, such that,

1) the individual u comes from a size-i block in E;

2) the individual u belongs to a size-i' block in E'.

The transition probabilities are

$$P(E,E') = \begin{cases} \frac{1}{\lambda+\#E_\wedge}, & \text{if } i' > 0; \\ \frac{\lambda}{\lambda+\#E_\wedge}, & \text{if } i' = 0. \end{cases} \qquad P(E',E) = \begin{cases} \frac{1}{\lambda+\#E_\wedge}, & \text{if } i > 0; \\ \frac{\lambda}{\lambda+\#E_\wedge}, & \text{if } i = 0. \end{cases}$$

With these notations, it is straightforward to verify that (2.3) is true. #

If we ignore the labels of the guests and count only the block sizes, then the cocktail process induces a Markov chain on integer partitions. The exchangeability and the detailed balance condition of the cocktail process imply that the induced integer-partition-valued Markov chain also satisfies the detailed balance condition.

In general, let $\{E^{(k)}\}_k$ be a homogenous Markov chain taking values in partitions of $[n]$. We call $\{E^{(k)}\}_k$ *exchangeable* if its one-step transition probability $P(\cdot,\cdot)$ remains invariant under permutations of $[n]$. That is,

$$P(E,E') = P(E_\sigma, E'_\sigma), \text{ for each } \sigma \in \mathcal{S}_n, \tag{2.4}$$

where E_σ is the partition defined by $E_\sigma(i,j) = E(\sigma(i),\sigma(j))$, $i,j = 1,\ldots,n$. For example, if $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{smallmatrix}\right)$ and $E = 13|2$, then $E_\sigma = \sigma^{-1}(1)\sigma^{-1}(3)|\sigma^{-1}(2) = 32|1 = 1|23$. We call the transition probability $P(\cdot,\cdot)$ *exchangeable* if it satisfies (2.4). In that case, it is straightforward to verify that the k-step transition probability $P^k(\cdot,\cdot)$ is exchangeable too. If $\{E^{(k)}\}_k$ is both irreducible and aperiodic, then the unique stationary distribution on partitions is exchangeable as the limit of $P^k(E,\cdot)$ which is

independent of E.

Let $\{e^{(k)}\}_k$ be the induced Markov chain by $\{E^{(k)}\}_k$ taking values in integer partitions of n. If $\{E^{(k)}\}_k$ is exchangeable, then $\{e^{(k)}\}_k$ is a homogenous Markov chain with one-step transition probability

$$P^{(I)}(e, e') = \sum_{E' \in \pi_I^{-1}(e')} P(E, E'), \tag{2.5}$$

where π_I is the mapping from set partitions to integer partitions, and E is an arbitrary partition of $[n]$ satisfying $\pi_I(E) = e$.

Proposition 2.2 *Let $\{E^{(k)}\}_k$ be a homogenous Markov chain taking values in partitions of $[n]$. Assume it is exchangeable, irreducible and aperiodic. Denote by p_n the unique stationary distribution of $\{E^{(k)}\}_k$. Let $\{e^{(k)}\}_k$ be the induced Markov chain taking values in integer partitions of n. Then $\{e^{(k)}\}_k$ has unique stationary distribution $p_n^{(I)}$ such that*

$$p_n^{(I)}(e) = N(\alpha_1, \ldots, \alpha_n) \cdot p_n(E),$$

where $e = 1^{\alpha_1} \cdots n^{\alpha_n}$ is an integer partition of n, $N(\alpha_1, \ldots, \alpha_n)$ is the combination number $n!/[\prod_{j=1}^{n} (j!)^{\alpha_j}\ \alpha_j!]$, E is an arbitrary partition satisfying $\pi_I(E) = e$. Furthermore, if $\{E^{(k)}\}_k$ satisfies the detailed balance condition, so does $\{e^{(k)}\}_k$.

Proof Evidently, $\{e^{(k)}\}_k$ has unique stationary distribution. Let $P(\cdot,\cdot)$ be the transition probability of $\{E^{(k)}\}_k$. Then

$$p_n(E) = \sum_{E' \in \mathcal{E}_n} p_n(E')P(E', E) = \sum_{e' \in \mathcal{S}_n} \sum_{E' \in \pi_I^{-1}(e')} p_n(E')P(E', E).$$

Note that p_n is exchangeable. Write $e = \pi_I(E)$. Then

$$\begin{aligned}
p_n^{(I)}(e) &= \sum_{E \in \pi_I^{-1}(e)} p_n(E) \\
&= \sum_{E \in \pi_I^{-1}(e)} \sum_{e' \in \mathcal{S}_n} \sum_{E' \in \pi_I^{-1}(e')} p_n(E')P(E', E) \\
&= \sum_{e' \in \mathcal{S}_n} \sum_{E' \in \pi_I^{-1}(e')} p_n(E') \cdot \sum_{E \in \pi_I^{-1}(e)} P(E', E) \\
&= \sum_{e' \in \mathcal{S}_n} \sum_{E' \in \pi_I^{-1}(e')} p_n(E')P^{(I)}(e', e) \\
&= \sum_{e' \in \mathcal{S}_n} p_n^{(I)}(e')P^{(I)}(e', e).
\end{aligned}$$

Therefore, $p_n^{(I)}$ is the stationary distribution of $\{e^{(k)}\}_k$. Using the same trick, it is straightforward to verify that

$$p^{(I)}(e)P^{(I)}(e, e') = p^{(I)}(e')P^{(I)}(e', e)$$

if $\{E^{(k)}\}_k$ satisfies the detailed balance condition. #

Corollary 2.1 *The integer-partition-valued process induced by the cocktail process*

has the unique stationary distribution

$$p_n^{(I)}(e_n;\lambda) = N(\alpha_1,\ldots,\alpha_n)\cdot\lambda^{\sum_j \alpha_j}/B_n(\lambda),$$

which satisfies the detailed balance condition, where $e_n = 1^{\alpha_1}\cdots n^{\alpha_n}$ *is an integer partition of* n*, and* $B_n(\lambda)$ *is the Bell polynomial.*

2.3 Ewens-Cocktail Process

If we modify the cocktail process as follows, the new Markov chain yields the Ewens sampling distribution as its stationary distribution on partitions.

> As in the original cocktail process, an individual u chosen uniformly at random splits off from the block to which he or she belongs. Suppose there are k blocks left formed by the other $n-1$ guests. The individual u joins one of the k blocks with probability *block size*$/(\lambda+n-1)$, or strike out on his or her own and form a new block with probability $\lambda/(\lambda+n-1)$.

We call the modified process the *Ewens-cocktail process.* Following a similar argument as in Proposition 2.1, we conclude that

Proposition 2.3 *The Ewens-cocktail process has the unique stationary distribution*

$$p_n(E;\lambda) = \frac{\Gamma(\lambda)\lambda^{\#E}}{\Gamma(n+\lambda)}\prod_{\text{blocks}}(b-1)!\ ,$$

which satisfies the detailed balance condition.

By Proposition 2.2, the induced process on integer partitions satisfies the detailed balance condition too.

Corollary 2.2 *The integer-partition-valued process induced by the Ewens-cocktail process has the unique stationary distribution*

$$p_n^{(I)}(e_n = 1^{\alpha_1} \cdots n^{\alpha_n}; \lambda) = N(\alpha_1, \ldots, \alpha_n) \cdot \frac{\Gamma(\lambda)\lambda^{\alpha_\bullet}}{\Gamma(n+\lambda)} \prod_{j=1}^{n} [(j-1)!]^{\alpha_j} \ ,$$

which satisfies the detailed balance condition, where $\alpha_\bullet = \sum_{j=1}^{n} \alpha_j$.

The Ewens-cocktail process provides an alternative way to simulate the Ewens sampling distribution. Unlike the Chinese restaurant process generating i.i.d. samples, the Ewens-cocktail process generates a reversible Markov chain $\{E^{(k)}\}_k$ with stationary distribution $p_n(\cdot; \lambda)$. By the ergodic theorem (for example, Theorem 3 in [52]), for any real-valued function f defined on $\mathcal{E}_n$,

$$\frac{1}{m} \sum_{k=1}^{m} f(E^{(k)}) \xrightarrow{\text{a.s.}} \sum_{E_n \in \mathcal{E}_n} f(E_n) p_n(E_n; \lambda), \text{ as } m \to \infty.$$

Composed with the Metropolis-Hastings algorithm (see [12], [14] for a good review), it may be used conveniently to sample a distribution close to the Ewens'.

2.4 Poisson-Ewens-Cocktail Process

If the number of guests in the cocktail process is no longer fixed, a Markov process on integer partitions can be introduced as follows.

> Suppose the arrival process is Poisson with constant rate λ. Once a new guest arrives, he or she follows the same rule as the Ewens-cocktail process to join the party. Suppose the departure process is Poisson with rate $N_t\delta$ independent of the arrivals, where N_t is the number of guests in the party at time t. All the current guests at time t have the equal chance to leave the party, independently of each other.

Note that we do not arrange the arrivals in any order, since there are infinite candidates. Nevertheless, if we focus only on the related integer-partition-valued process, called Poisson-Ewens-Cocktail process, we get the stationary distribution as follows.

Proposition 2.4 *The Poisson-Ewens-cocktail process is time reversible with respect to the stationary distribution on integer partitions as follows*

$$P(n, e_n = 1^{\alpha_1} \cdots n^{\alpha_n}; \lambda, \delta) = p_{\lambda/\delta}(n) \times N(\alpha_1, \ldots, \alpha_n) \cdot \frac{\Gamma(\lambda)\lambda^{\alpha_\bullet}}{\Gamma(n+\lambda)} \prod_{j=1}^{n} [(j-1)!]^{\alpha_j},$$

where $\lambda > 0$, $\delta > 0$, $\alpha_\bullet = \sum_j \alpha_j$, *and* $p_{\lambda/\delta}(n)$ *is the Poisson mass function with rate* λ/δ.

In other words, the stationary distribution is separable. The marginal distribution of the number of guests is still Poisson. Fixing the number of guests in the party, the conditional distribution on integer partitions is the induced distribution by the Ewens'.

CHAPTER 3

A PARTITION MODEL FOR BAYESIAN MULTIPLE COMPARISONS

In this chapter, we propose a partition model for Bayesian multiple comparisons. Instead of making inference based on pairwise comparisons, the partition model provides inference in the form of set partitions. We suggest the marginal prior on partitions be infinitely exchangeable and the posterior distribution on partitions be location-scale invariant.

3.1 Remarks on the Literature

The purpose of multiple comparisons or simultaneous inference is to summarize a set of statistical statements. The statements can be confidence intervals, significance tests, etc. A typical problem might be how to compare the means of k populations. Two different approaches are used in the literature to solve this problem.

The classical approach is based on multiple pairwise comparisons. For example, the standard t-procedure may be used for each of the $k(k-1)/2$ pairwise comparisons. To control the chance of false discoveries, for example, the *Fisher protected least significant difference (LSD) test* uses an initial F-test to check the hypothesis

that all means are equal. Pairwise comparisons are meaningful only when the initial hypothesis is rejected. Some other methods, such as *Bonferroni intervals*, *Sheffé intervals* or *Tukey's Q-method*, provide wider confidence intervals than the t-type intervals to reduce false discoveries. Typically, the lengths of adjusted intervals depend on the number k (see [34] for a good review). In other words, the inference about whether or not two populations a and b have equal means might depend on some other population c.

The more recent approach is a Bayesian one. Berry ([7]) suggested considering every possible combination of equality and inequality among the k means. A mixture of Dirichlet processes ([21], [5]) might be used to model the population means. It allows a statement of the type "population a and b have the same mean" has positive probability. Since then, Bayesian analysis using the Dirichlet prior ([24]) or the uniform prior ([41]) has been applied to the multiple comparisons problem.

We propose a Bayesian model for multiple comparisons using infinitely exchangeable priors on the partitions of the k populations. The partitions are identified by the equality and inequality relationships among the population means. The uniform prior on partitions however is not infinitely exchangeable. Unlike Gopalan and Berry ([24]), we use residual statistics to make the inference on partitions be location-scale invariant under data transformations.

3.2 A Bayesian Model Permitting Multiple Comparisons

3.2.1 Infinite Exchangeability

Consider finitely many varieties. From a Bayesian point of view, the principle of egalitarianism or symmetry requires that the prior on the variety effects be exchangeable. For example, if the statement "the mean for variety a exceeds that for b by more than 10%" has prior probability 0.15, then the same is true with varieties c and d.

Infinite exchangeability is just finite exchangeability plus non-interference, which means that the same probabilistic statements hold for an extended trial that includes additional varieties. As a consequence, if a is found to be superior to b in a trial involving ten additional varieties, a may be said to be superior to b without qualification. The notion is a natural one, that these additional varieties do not interfere with the values for the pair being compared.

Infinite exchangeability in this sense is a model assumption. It is not a statement of personal belief, nor is it a statement of agricultural, chemical, or educational fact. The non-interference component is especially critical. Without it, we cannot deduce by mathematical argument alone that the observed contrast for $a - b$ might not be reversed if variety c had been included in the trial.

3.2.2 The Variety Process

Perhaps the simplest model for the observation process is such that, conditional on the variety effects $\tau_1, \ldots, \tau_k$, the observations are generated by $Y_{ij} = \tau_i + \epsilon_{ij}$, where the ϵs are independent $N(0, \sigma^2)$. The variety effects are distributed marginally as $N(\mu, \sigma_\tau^2)$, but for each finite set of indices $\{i_1, \ldots, i_r\}$ the probability is positive that $\tau_{i_1} = \cdots = \tau_{i_r}$. In other words, two or more varieties having different labels may be genetically identical or otherwise equivalent in terms of yield.

To construct such a process, the distribution for $(\tau_1, \ldots, \tau_k)$ on $\mathcal{R}^k$ is obtained in two steps as follows.

$$\mathbf{E} \sim p_k \quad \text{on } \mathcal{E}_k,$$
$$\tau \sim N_k(\mu\mathbf{1}, \sigma_\tau^2\mathbf{E}) \quad \text{on } \mathcal{R}^k.$$

In these expressions $\mathcal{E}_k$ is the set of partitions of $[k] = \{1, \ldots, k\}$, so $\mathbf{E} \in \mathcal{E}_k$ is an equivalence relation on the set. $\mathbf{1}$ is the constant vector. For example if $k = 4$, the partition denoted by $\mathbf{E} = 3|1|24$ may also be represented by the equivalence relation

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix},$$

a positive semi-definite matrix of rank equal to the number of blocks; $\text{rank}(\mathbf{E}) = \#\mathbf{E}$. For two levels such that $\mathbf{E}_{ij} = 1$, i.e. varieties i and j in the same block, the distribution $N_k(\mu\mathbf{1}, \mathbf{E})$ implies that $\tau_i = \tau_j$. Operationally, therefore, a random

partition $\mathbf{E}$ is chosen with distribution p_k on $\mathcal{E}_k$, and independent random variables are generated, one for each block. If $\mathbf{E} = 1|24|3$, three independent random variables are generated, and the sequence of length 4 is such that $\tau_4 = \tau_2$. The Gaussian assumption is not essential for exchangeability, but it does simplify computations.

In order that the marginal distribution of τ on $\mathcal{R}^k$ be finitely exchangeable, it is necessary and sufficient that each distribution p_k be invariant under permutation of elements. For $k = 4$, the 15 partitions of $\{a, b, c, d\}$ are conventionally listed in the form

$$abcd\ [1], \quad abc|d\ [4], \quad ab|cd\ [3], \quad ab|c|d\ [6], \quad a|b|c|d\ [1]$$

showing an orbit representative followed by the orbit size. A distribution is finitely exchangeable, or symmetric, if and only if it is uniform on each of the orbits. There are as many group orbits in $\mathcal{E}_k$ as there are partitions of the number k, i.e. three orbits for $k = 3$, five for $k = 4$, seven for $k = 5$ and so on. In order that the construction determine a process, the Kolmogorov consistency condition must be satisfied, so p_k is necessarily the marginal distribution of p_{k+1} under elementwise deletion. In particular, this means that

$$p_3(123) = p_4(1234) + p_4(123|4)$$

$$p_3(1|23) = p_4(14|23) + p_4(1|234) + p_4(1|4|23)$$

$$p_3(1|2|3) = 3p_4(14|2|3) + p_4(1|2|3|4).$$

Thus $p_3(123) = 0$ implies that every partition of $[k]$ having a block of size three or more has zero probability. Conversely $p_3(1|2|3) = 0$ implies that only binary partitions have positive probability.

The marginal distribution on $\mathcal{E}_3$ of the uniform distribution on $\mathcal{E}_4$ is not uniform, so the uniform distributions on $\mathcal{E}_k$ do not determine an infinitely exchangeable partition. An exchangeable prior on partitions is necessarily non-uniform. For the purposes of this section, we restrict attention to processes such that p_k is strictly positive on $\mathcal{E}_k$, so all partitions have positive prior probability. Specific examples of such exchangeable partition processes are given below.

3.2.3 Exchangeable Partition Processes

The following are instances of infinitely exchangeable partition processes, most of which are not suitable for multiple comparisons.

(i) For each finite set, p_k puts mass one on the minimal partition with k blocks in $\mathcal{E}_k$.

(ii) p_k puts mass one on the maximal partition with only 1 block in $\mathcal{E}_k$.

(iii) Each infinitely exchangeable random sequence $X = (X_1, X_2, \ldots)$ determines an infinitely exchangeable random partition as follows. For each set $[k]$, define the random partition $\mathbf{E}$ by $\mathbf{E}_{ij} = 1$ if $X_i = X_j$, and zero otherwise. Let $p_k(\mathbf{E})$ be the probability induced by X.

(iv) A binomial random partition is obtained by the construction (iii) in which X is an i.i.d. Bernoulli sequence with parameter θ. Each partition $\mathbf{E}$ of $[k]$ containing two blocks of sizes $r, k-r$, has probability

$$p_k(\mathbf{E};\theta) = \theta^r(1-\theta)^{k-r} + (1-\theta)^r\theta^{k-r}.$$

Non-binary partitions having three or more blocks have zero probability.

(v) If, in the preceding construction, the parameter θ is a symmetric beta random variable with parameter α, the probability is

$$p_k(\mathbf{E};\alpha) = \frac{2\,\Gamma(r+\alpha)\,\Gamma(k-r+\alpha)\,\Gamma(2\alpha)}{\Gamma(k+2\alpha)\,\Gamma^2(\alpha)}.$$

(vi) If, in (iv) and (v), the Bernoulli is replaced by the multinomial with n levels, and the beta by the symmetric Dirichlet, the limiting probability as $\alpha \to 0$ and $n \to \infty$ such that $n\alpha = \lambda > 0$ is fixed, is

$$p_k(\mathbf{E};\lambda) = \frac{\Gamma(\lambda)\,\lambda^{\#\mathbf{E}}}{\Gamma(k+\lambda)} \prod_{\text{blocks}} (b-1)!, \tag{3.1}$$

where $\#\mathbf{E}$ is the number of blocks, b is a block size, and the product runs over the blocks ([30], [32]). Note that the distribution (3.1) is just the Ewens sampling distribution (see Section 1.5) . The limit distribution of (3.1) as $\lambda \to 0$ (or ∞), is (ii) (or (i)).

(vii) If, in (iv) and (v), the Bernoulli is replaced by random discrete distribution $\{P_i\}_{i=1,2,\ldots}$, and the beta is replaced by a sequence of independent random variables $W_j \sim \text{beta}(1-\delta, \lambda+(j-1)\delta)$ with parameters $\lambda > 0$ and $0 \leq \delta < 1$, such that $P_i = (1-W_1)\ldots(1-W_{i-1})W_i$, then each partition $\mathbf{E}$ of $[k]$ has probability

$$p_k(\mathbf{E};\lambda,\delta) = \frac{\Gamma(\lambda+1-\delta)\prod_{j=0}^{\#\mathbf{E}-1}(\lambda+j\delta)}{\Gamma(\lambda+k-\delta)\Gamma(1-\delta)^{\#\mathbf{E}}[\lambda+(\#\mathbf{E}-1)\delta]}\prod_{\text{blocks}}\Gamma(b-\delta),$$

which yields (vi) as a special case if $\delta = 0$ ([43], [45]). It is the two-parameter generalization of the Ewens' (see Section 2.1) .

Although it is more usually defined on number partitions, the process constructed in (vi) with distribution functions $\{p_k(\cdot;\lambda)\}_k$ is called the Ewens partition process (see Section 1.5) . It has the essential property for present purposes that all partitions have positive probability. More than that, in all exchangeable partition processes, the extended Ewens family, which allows λ taking values in $[0,\infty]$, is the only one satisfying the self-similarity. It makes the Ewens family the preferred prior when the self-similarity serves as an assumption.

3.2.4 Likelihood Function

We consider in this section what is meant by the likelihood function based on a finite-dimensional observation $\mathbf{y} \in \mathcal{R}^n$. If $\mathbf{X}$ is the indicator matrix for varieties, the

simplest model has three principal components as follows:

$$\begin{aligned} \mathbf{Y} &\sim N_n(\mathbf{X}\tau, \sigma^2\mathbf{I_n}) \quad \text{on } \mathcal{R}^n \\ \tau &\sim N_k(\mu\mathbf{1}, \sigma_\tau^2\mathbf{E}) \quad \text{on } \mathcal{R}^k \\ \mathbf{E} &\sim p_k(\cdot) \quad \text{on } \mathcal{E}_k. \end{aligned} \tag{3.2}$$

Our focus initially is on the determination of the varieties that are equivalent, so we aim first to obtain a posterior distribution on variety partitions in $\mathcal{E}_k$ rather than variety effects in $\mathcal{R}^k$ or contrasts in the quotient space $\mathcal{R}^k/\mathbf{1}$. Accordingly, the likelihood is based on the first two components, and the third component is a part of the prior. To the extent that it is necessary, a prior distribution may be required for the three remaining parameters $(\mu, \sigma^2, \sigma_\tau^2)$. The first two components of the model imply that, given the parameter values $\mathbf{E} \in \mathcal{E}_k$, plus $(\mu, \sigma^2, \sigma_\tau^2)$,

$$\mathbf{Y} \sim N(\mu\mathbf{1},\ \sigma^2\mathbf{I_n} + \sigma_\tau^2\mathbf{X}\mathbf{E}\mathbf{X}^T). \tag{3.3}$$

It is important here to understand that $\mathbf{X}$ is the matrix whose columns are the indicator vectors for the k varieties. The basis matters, and $\mathbf{X}$ is not the matrix produced by the model formula $y \sim$ *variety* in R or S.

To simplify the exercise, we base the likelihood function on a particular function of $\mathbf{Y}$ rather than $\mathbf{Y}$ itself. Let $\bar{Y}, s^2$ be the sample mean and variance. The standardized residual vector $\check{\mathbf{Y}} = (\mathbf{Y} - \bar{Y}\mathbf{1})/s$ evidently has a distribution depending only on

$(\sigma_\tau^2/\sigma^2, \mathbf{E})$. The marginal log likelihood ([54]) for the pair $(\theta = \sigma_\tau^2/\sigma^2, \mathbf{E})$ is

$$\check{l}(\theta, \mathbf{E}; \check{\mathbf{y}}) = -\frac{n-1}{2}\log(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{Q}\mathbf{y}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}|, \tag{3.4}$$

where $\boldsymbol{\Sigma} = \mathbf{I_n} + \theta\mathbf{X}\mathbf{E}\mathbf{X}^T$ and $\mathbf{Q} = \mathbf{I_n} - \mathbf{1}(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{\Sigma}^{-1}$. The choice of sample mean and sample standard deviation for standardization purposes is of no consequence: any location and scale statistic such as the sample median and inter-quartile range yields exactly the same marginal likelihood. The use of the standardized residual vector for likelihood calculations implies that the conclusions regarding $(\theta, \mathbf{E})$ are unaffected by component-wise affine transformation $y \mapsto a + by$.

Most of the terms occurring in the marginal log likelihood can be expressed in terms of block sizes and block averages as follows. The $k \times k$ matrix $\mathbf{E}$ is a partition of the varieties into $\#\mathbf{E}$ blocks, and $\mathbf{X}\mathbf{E}\mathbf{X}^T$ is the corresponding $n \times n$ matrix that partitions the n units into $\#\mathbf{E}$ blocks. Let $b_1, b_2, \ldots, b_{\#\mathbf{E}}$ be the block sizes of $\mathbf{X}\mathbf{E}\mathbf{X}^T$, and let $\mathbf{J}_b$ be the $b \times b$ matrix whose elements are all one. It is convenient to reorder the components of $\mathbf{y}$ such that

$$\boldsymbol{\Sigma} = \mathbf{I_n} + \theta \cdot \mathrm{diag}\{\mathbf{J}_{b_1}, \ldots, \mathbf{J}_{b_{\#\mathbf{E}}}\}.$$

in which case the inverse is

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I_n} - \theta \cdot \mathrm{diag}\left\{\frac{\mathbf{J}_{b_1}}{1 + b_1\theta}, \ldots, \frac{\mathbf{J}_{b_{\#\mathbf{E}}}}{1 + b_{\#\mathbf{E}}\theta}\right\}.$$

If we set $w_i = b_i/(1+\theta b_i)$, the various terms in the likelihood function reduce to

$$\begin{aligned} |\mathbf{\Sigma}| &= \prod_{i=1}^{\#\mathbf{E}} (1+\theta b_i) \\ \mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1} &= \sum w_i \\ \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{Q}\mathbf{y} &= S_w^2 + \sum w_i(\bar{y}_i - \bar{y}_w)^2 \end{aligned}$$

where S_w^2 is the within-blocks sum of squares, $\bar{y}_i$ is the mean of $\mathbf{y}$ in the ith block, and $\bar{y}_w = \sum w_i\bar{y}_i / \sum w_i$ is the weighted average of the block means.

3.2.5 Posterior Distribution on Partitions

Following the analysis in the preceding section, the marginal likelihood function at $(\theta, \mathbf{E})$ is

$$L(\theta, \mathbf{E}; \mathbf{y}) \propto (\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{Q}\mathbf{y})^{-\frac{n-1}{2}} \, |\mathbf{\Sigma}|^{-\frac{1}{2}} \, (\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-\frac{1}{2}}.$$

For a formal Bayesian analysis we must choose a proper prior distribution for the variance ratio, and the symmetric F-family

$$\pi(\theta) \propto \frac{\theta^{\alpha-1}}{(1+\theta)^{2\alpha}}$$

with $\alpha > 0$ offers a range of reasonable choices. It is implicit in the construction that the variance ratio is independent of the partition, so the marginal likelihood function

at $\mathbf{E} \in \mathcal{E}_k$ is

$$L(\mathbf{E}; \mathbf{y}) = \int_0^\infty L(\theta, \mathbf{E}; \mathbf{y})\, \pi(\theta)\, d\theta.$$

The posterior distribution at $\mathbf{E} \in \mathcal{E}_k$ is then

$$p(\mathbf{E}|\mathbf{y}) \propto p_k(\mathbf{E}) \times L(\mathbf{E}; \mathbf{y}). \tag{3.5}$$

In all calculations in the examples, $p_k(\mathbf{E}) = p_k(\mathbf{E}; \lambda)$ is the Ewens distribution with parameter λ.

The posterior mean partition

$$\bar{\mathbf{E}} = \sum_{\mathbf{E} \in \mathcal{E}_k} \mathbf{E} \cdot p(\mathbf{E} \,|\, \mathbf{y})$$

is a convex combination of symmetric positive semi-definite binary matrices and is not itself a partition. Nevertheless the elements

$$\bar{\mathbf{E}}_{rs} = \Pr(r \sim s \,|\, \mathbf{y})$$

have a simple straightforward interpretation, so this is a useful partial summary of the posterior. The values, which can also be obtained from the cumulative posterior distribution $P(\mathbf{E}' \,|\, \mathbf{y}) = \sum_{\mathbf{E}} \Pr(\mathbf{E}' \leq \mathbf{E} \,|\, \mathbf{y})$, are shown for two examples in Tables 3.3 and 3.6 . The posterior distribution on the number of blocks is another useful summary statistic.

3.3 Examples

3.3.1 Example 1: Fat Absorbed by Doughnuts

The first example taken from from Snedecor & Cochran ([51], p. 259), is a balanced one-way design with $k = 4$ and $n = 24$. Six batches of doughnuts were cooked using each of four types of fat, making a total of 24 batches. The response is the amount of fat in grams absorbed by each batch.

The marginal posterior distribution on partitions is shown in Table 3.2 for a range of nine prior distributions of the form

$$\frac{\theta^{\alpha-1}}{(1+\theta)^{2\alpha}} \times p_4(\mathbf{E};\lambda)$$

with $\lambda = 1, 2, 10$ and $\alpha = 0.5, 1, 2$. The partitions are arranged in descending order at $\lambda = 1$. For this range of prior distributions, the conclusions depend to a moderate extent on the choice of λ but are relatively insensitive to α, which determines the prior on the variance ratio. For $\lambda = 1$, the prior distribution on the number of blocks is $(0.25, 0.46, 0.25, 0.04)$, and the posterior distribution is $(0.06, 0.44, 0.41, 0.09)$ for $\alpha = 1$. For $\lambda = 2$, the distributions are $(0.10, 0.37, 0.40, 0.13)$ and $(0.02, 0.27, 0.49, 0.22)$. Since the prior expected number of blocks, is monotone increasing in λ, a similar trend is evident in the posterior, larger values favoring more fragmented partitions. The prior probability that all four fats are equivalent is $6/((\lambda+1)(\lambda+2)(\lambda+3)$, or $1/4, 1/10, 1/286$ for the values in Table 3.1 . At the other extreme, the completely

fragmented partition into four blocks has prior probability $0.04, 0.13, 0.58$. These values make it clear that the range of λ-values considered here is fairly extreme.

Table 3.1: Grams of Fat Absorbed Per Batch of Doughnuts

Type of Fat	a	b	c	d
	64	78	75	55
	72	91	93	66
	68	97	78	49
	77	82	71	64
	56	85	63	70
	95	77	76	68
$\bar{Y}_j$	72	85	76	62
s_j	13.3	7.8	9.9	8.2
Pooled s=10.0				

The posterior probability in Table 3.2 is uniformly small for each partition such that b, d occur in the same block. At the other extreme, most partitions in which a, c are in the same block have appreciable probability. These conclusions are consistent with the pattern of observed sample means as shown in Table 3.1, where b has the highest mean, d the lowest, and a, c are the closest pair. For $(\lambda, \alpha) = (1, 1)$ the marginal posterior probability that a, c are equivalent is 55%, while the posterior probability that b, d are equivalent is only 9%. The posterior probabilities for all six pairwise comparisons are shown in Table 3.3 . Although the numerical values depend to a great extent on the prior, the ordering of pairwise contrasts is strongly consistent across the entire range of priors. For all nine priors in this example, the ordering

$$P(a \sim c|\mathbf{y}) > P(b \sim c|\mathbf{y}) > P(a \sim d|\mathbf{y}) > P(a \sim b|\mathbf{y}) > P(c \sim d|\mathbf{y}) > P(b \sim d|\mathbf{y})$$

Table 3.2: Marginal Posterior Probabilities for the 15 Partitions, $p(\mathbf{E}|\mathbf{y}, \lambda, \alpha) \times 100\%$

	Prior parameter values (λ, α)								
E	$(1, 0.5)$	$(1, 1)$	$(1, 2)$	$(2, 0.5)$	$(2, 1)$	$(2, 2)$	$(10, 0.5)$	$(10, 1)$	$(10, 2)$
$abc\|d$	18	19	19	12	11	11	2	2	1
$ac\|b\|d$	16	17	17	19	20	21	12	12	12
$acd\|b$	13	13	13	8	8	8	1	1	1
$a\|bc\|d$	9	10	10	12	12	12	7	7	7
$ad\|bc$	9	9	9	6	5	5	1	1	1
$a\|b\|c\|d$	8	9	9	21	22	23	66	66	67
$ad\|b\|c$	8	8	8	9	10	10	6	6	6
$abcd$	8	6	5	3	2	2	0	0	0
$ab\|c\|d$	3	3	4	4	4	4	3	3	3
$a\|b\|cd$	3	3	3	3	3	3	2	2	2
$ab\|cd$	2	1	1	1	1	1	0	0	0
$abd\|c$	1	1	1	1	1	0	0	0	0
$a\|bcd$	1	1	1	1	1	0	0	0	0
$ac\|bd$	1	0	0	0	0	0	0	0	0
$a\|bd\|c$	0	0	0	0	0	0	0	0	0

is consistent with the ascending order of the differences among sample means. However, the relationship between posterior probabilities and sample mean differences is not invariably monotone.

Table 3.3: Posterior Probabilities That Two Types Belong to the Same Block, $P(\cdot \sim \cdot|\mathbf{y}, \lambda, \alpha) \times 100\%$

(λ, α)	$(1, 0.5)$	$(1, 1)$	$(1, 2)$	$(2, 0.5)$	$(2, 1)$	$(2, 2)$	$(10, 0.5)$	$(10, 1)$	$(10, 2)$
$a \sim c$	56	55	55	42	42	41	15	15	15
$b \sim c$	46	44	43	33	31	30	10	9	9
$a \sim d$	39	37	36	27	26	25	8	8	7
$a \sim b$	33	31	29	20	19	18	4	4	4
$c \sim d$	27	25	23	16	15	14	3	3	3
$b \sim d$	11	9	7	5	3	3	1	0	0

Snedecor & Cochran ([51], pp. 272-273) give the results of several classical significance tests, all based on comparisons of differences between sample means. The least

significant difference (LSD) test declares significant at the 5% level any pair whose sample means differ by more than 12.1 grams. By this criterion, the differences between $a\&b$, $c\&d$ and $b\&d$ are large enough that they are deemed unlikely to be due to random variation alone. More conservatively, Tukey's Q-method requires the difference in sample means to be at least 16.2 grams for significance to be declared at the 5% level. Only the mean difference of $b\&d$ is significantly nonzero by this criterion.

Each of these classical tests computes the probability under the null hypothesis of equal means of obtaining a value of the statistic as extreme as the value observed. The smaller this probability, the less likely it is that the observed difference is due to random variation, and accordingly the greater the evidence against the hypothesis. The Bayesian analysis aims instead to find a posterior distribution on partitions. The posterior values reported in Tables 3.2 and 3.3 are thus not directly comparable with p-values in significance tests. Nevertheless, they are likely to be interpreted in a similar manner, and we aim here to make such a comparison.

Roughly speaking, at the 5% level, the LSD test is consistent with the posterior for $\lambda \simeq 10$ in Table 3.3 . The set of partitions such that $a \sim b$ has posterior probability 4%, while $c \sim d$ has probability 3% and $b \sim d$ less than 1%. These are also the contrasts declared significantly different from zero by the LSD test at the 5% level. The more conservative Q-method is consistent with $\lambda \simeq 2$ in that only the pair (b, d) has posterior probability or tail probability less than 5%. In other words, the LSD test and the Q-method are consistent with Bayesian calculations based on very different

prior distributions. From Table 3.3, it appears the more conservative Q-method corresponds to a moderately uniform prior on partitions, while the LSD method corresponds to a prior that puts the greater part of its mass on highly fragmented partitions. The prior probability for each pair $a \sim b$ is $1/(\lambda + 1)$, so the Q-method has an implicit prior of 1/3 for each contrast, while the LSD method has a prior of 1/11.

3.3.2 Example 2: Dyestuff Data

Example 2 is a comparison of six dyestuffs taken from Davies ([18], p. 105). These data have been analyzed by Box and Tiao ([11], p. 246, 370) to illustrate the differences between fixed-effects models and random effects models for estimation and prediction. Five observations are available on each of six dyestuffs. The six sample means and sample standard deviations are shown in Table 3.4 .

Table 3.4: Sample Means and Sample Standard Deviations for the Dyestuff Data

Dyestuff	1	2	3	4	5	6
$\bar{Y}_j$	105	128	164	98	200	70
s_j	63	33	38	69	50	31

With $\alpha = 1$ in the prior for the variance ratio, the posterior probability distribution on partitions was computed for various values of λ. The posterior distribution of the number of blocks shown in Table 3.5 indicates that the varieties are not all equivalent. The posterior probability of six blocks is small, but it is greater than the prior probability. For $\lambda = 1$ it was found that the modal partition is 1246|35 with

posterior probability 15%, followed by 1246|3|5 with probability 9%, and 12346|5 with probability 7%. The unpartitioned set has probability 3%, and these values are fairly constant for $0.5 < \lambda < 2$.

Table 3.5: Posterior Distribution on the Number of Blocks for the Dyestuff Data

Number of blocks	1	2	3	4	5	6
λ=1	0.03	0.33	0.41	0.19	0.04	0.00
λ=2	0.01	0.14	0.37	0.33	0.13	0.02

The posterior probabilities for pairwise comparisons are shown in Table 3.6 for $\alpha = 1$ and for a range of λ-values. The p-values for the standard t-test and Tukey's Q test are also shown for comparison. The prior probability that any specific pair is equivalent is $1/(\lambda + 1)$, and the posterior probabilities also decrease with λ. For $\lambda = 2$, the posterior probability that dyestuffs 1 and 4 belong to the same block, or are equivalent, is 51%. At the other extreme, the posterior probability that 5 and 6 are equivalent is only 2%. At the 5% level, the conclusions from Tukey's Q test correspond roughly to $\lambda = 2$, while the standard LSD test corresponds to $\lambda = 9$.

Table 3.6 illustrates another curious feature of the Bayesian model, namely that the order of the posterior probabilities for pairs does not correspond to the order of the sample mean differences. Several instances of such inversions may be observed. For example, we find that

$$\bar{Y}_3 - \bar{Y}_2 \quad : \quad 164 - 128 = 36$$

$$\bar{Y}_5 - \bar{Y}_3 \quad : \quad 200 - 164 = 36$$

All standard comparisons via pairwise differences inevitably conclude that $\mu_3 = \mu_2$ is just as likely as $\mu_5 = \mu_3$ because the observed differences are the same. The p-values in the final two columns based on pairwise standardized differences reflect this view, 26% for the LSD test and 86% for the Q test. However, in the Bayesian comparison with $\lambda = 2$, the posterior probability that $\mu_3 = \mu_5$ is 40%, while the posterior probability that $\mu_2 = \mu_3$ is only 26%.

Table 3.6: Posterior Probabilities $\times 100$ for Pairwise Comparisons As a Function of λ

					λ						p-value	
Pair	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	LSD	Q
1,4	66	51	42	35	30	27	24	21	20	18	83	100
4,6	64	49	40	33	29	25	22	20	18	17	38	94
1,6	61	45	36	30	25	22	20	17	16	14	28	87
1,2	56	42	33	28	24	21	18	16	15	14	47	98
2,4	54	40	31	26	22	19	17	15	14	12	35	93
3,5	53	40	33	28	25	22	20	18	16	15	26	86
2,6	45	30	22	17	14	12	10	9	8	7	8	45
2,3	35	26	21	18	16	14	12	11	10	9	26	86
1,3	23	16	12	10	8	7	6	6	5	5	7	43
3,4	21	14	10	8	7	6	5	4	4	4	5	31
2,5	20	13	10	8	6	6	5	4	4	4	3	23
3,6	15	8	5	4	3	2	2	2	1	1	0.6	6
1,5	10	5	4	3	2	2	2	1	1	1	0.6	6
4,5	8	4	3	2	2	1	1	1	1	1	0.3	3
5,6	5	2	1	1	1	0	0	0	0	0	0.0	0.0

The explanation for these inversions is fairly simple. In order for a difference of 36 units to occur well separated from the body of the data, either a single large outlier in $\tau_1, \ldots, \tau_k$ has occurred and the two groups are equivalent, or two large outliers have occurred and the two groups are not equivalent. The latter event is considerably less likely than the former, increasing the likelihood that the two groups are homogeneous.

A similar difference in the main body of the data could equally well be generated by one group or by two since no outliers are required.

3.4 Inference for Variety Contrasts

If we have six observations on each of four varieties $\{a, b, c, d\}$, the analysis in Section 3.2.5 yields a posterior distribution on the 15 partitions of $\{a, b, c, d\}$. Suppose, for example, that the partition $ac|bd$ has probability 0.75 suggesting that varieties a, c are similar, as are b, d, but all four varieties are unlikely to be equivalent. This piece of information may be useful, but we would ordinarily like to know more, for example which pair of varieties has the higher yield and by how much. Quantitative information of this sort is not available directly from the posterior distribution on partitions, but it may be obtained indirectly as follows. Let i^* be a new unit or plot, and let $x(i^*) = r$ be the variety. Given the parameter values $\mathbf{E}, \mu, \sigma^2, \sigma^2_\tau$, the conditional distribution of $Y(i^*)$ given the observed data is Gaussian with mean and variance

$$\frac{\sigma^2\mu + n_r\sigma^2_\tau\bar{y}_r}{\sigma^2 + n_r\sigma^2_\tau} \qquad \sigma^2 + \frac{\sigma^2_\tau}{1 + n_r\sigma^2_\tau/\sigma^2}.$$

where $n_r = \#\{i \mid \mathbf{E}_{x(i),x(i^*)} = 1\}$ is the number of sample units whose variety is equivalent to $x(i^*)$, and $\bar{y}_r$ is the average of the response on these units. The predictive distribution can now be obtained by posterior averaging.

The more conventional approach leading ultimately to the same conclusion is to

modify the model by the inclusion of variety effects as a component of the parameter space. For this purpose, the second component in (3.2) is a part of the prior, not a part of the parametric model in the conventional sense. This distinction is critical in the definition of likelihood and the interpretation of the likelihood principle. From a Bayesian point of view, the joint distribution function including variety effects is

$$f(\mathbf{y}, \tau, \mu, \sigma^2, \sigma_\tau^2, \mathbf{E}) \;\propto\; p(\mu, \sigma^2, \sigma_\tau^2, \mathbf{E}) f(\tau|\mu, \sigma_\tau^2, \mathbf{E}) f(\mathbf{y}|\tau, \sigma^2).$$

The posterior distribution on variety contrasts depends on the data, but it depends on the scale of y and is not a function of the residual configuration. Nevertheless, from the posterior distribution on variety contrasts it is possible to calculate probabilities such as $\Pr(\tau_a = \tau_c \,|\, \mathbf{y})$ from which the marginal distribution on partitions can be determined. If contradictory conclusions are to be avoided, we must arrange matters so that the marginal distribution on partitions coincides with the calculations in Section 3.2.5 . This consistency condition implies that the marginal posterior distribution on partitions depends only on the residual configuration statistic.

3.4.1 Compatible Prior on the Original Parameter Space

In order to avoid contradictory conclusions on the variety effects, we aim to construct a prior p on the original parameter space $\Theta = \{(\mu, \sigma^2, \sigma_\tau^2, \mathbf{E})\}$ such that the following diagram is commutative:

$$\begin{array}{ccc} \mathbf{y} \in \mathcal{R}^n & \xrightarrow{\text{Model, prior}} & \mathcal{P}(\{(\mu, \sigma^2, \sigma_\tau^2, \mathbf{E})\}) \\ \Big\downarrow T_S & & \Big\downarrow T_\Theta \\ \check{\mathbf{y}} \in \check{\mathcal{R}}^n & \xrightarrow{\text{Model}', \text{prior}'} & \mathcal{P}(\{(\mathbf{E}, \theta)\}) \end{array}$$

Here $T_S(\mathbf{y}) = (\mathbf{y} - \bar{y}\mathbf{1})/s$, $T_\Theta((\mu, \sigma^2, \sigma_\tau^2, \mathbf{E})) = (\mathbf{E}, \sigma_\tau^2/\sigma^2)$, and $\mathcal{P}(\cdot)$ indicates the set of all possible probability measures defined on the given parameter space.

Proposition 3.1 *If the (improper) prior p on Θ takes the form of*

$$p(\mu, \sigma^2, \sigma_\tau^2, \mathbf{E})d\mu\ d\sigma^2 d\sigma_\tau^2 = c \cdot p_k(\mathbf{E}) \cdot (\sigma^2)^{-2} \cdot p_1(\frac{\sigma_\tau^2}{\sigma^2})d\mu\ d\sigma^2 d\sigma_\tau^2, \tag{3.6}$$

where $c > 0$, $p_1(\cdot)$ is a positive and integrable function defined on $(0, \infty)$, then the above diagram is commutative with the prior p' on $\Theta' = \{(\mathbf{E}, \theta)\}$ defined by

$$p'(\mathbf{E}, \theta)d\theta = c \cdot p_k(\mathbf{E})p_1(\theta)d\theta. \tag{3.7}$$

Proof Write $\vartheta = (\mu, \sigma^2, \sigma_\tau^2, \mathbf{E})$, $d\vartheta = d\mu d\sigma^2 d\sigma_\tau^2$, and $\mathbf{\Sigma}_0 = \sigma^2\mathbf{I_n} + \sigma_\tau^2\mathbf{XEX}^T$. Let $\mathbf{\Sigma} = \mathbf{I_n} + \theta \cdot \mathbf{XEX}^T$, $\mathbf{Q} = \mathbf{I_n} - \mathbf{1}(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{\Sigma}^{-1}$ as in Section 3.2.4 . Then

$$\begin{aligned} & \int_{\vartheta \in T_\Theta^{-1}((\mathbf{E}, \theta))} f(\mathbf{y}|\vartheta)p(\vartheta)d\vartheta \\ = & \int_{\mu \in \mathcal{R}, \frac{\sigma_\tau^2}{\sigma^2} = \theta} (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_0|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu\mathbf{1})^T {\mathbf{\Sigma}_0}^{-1}(\mathbf{y} - \mu\mathbf{1})\} \\ & \cdot c \cdot p_k(\mathbf{E}) \cdot (\sigma^2)^{-2} \cdot p_1(\frac{\sigma_\tau^2}{\sigma^2})d\mu\ d\sigma^2 d\sigma_\tau^2 \end{aligned}$$

$$
\begin{aligned}
\overset{u=\sigma^2,\theta=\frac{\sigma_\tau^2}{\sigma^2}}{=} & \int_{\mu\in\mathcal{R},u\in\mathcal{R}_+}(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}}u^{-\frac{n}{2}}\exp\{-\frac{1}{2}(\mathbf{y}-\mu\mathbf{1})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\mu\mathbf{1})\frac{1}{u}\}\\
& \cdot c\cdot p_k(\mathbf{E})\cdot u^{-2}\cdot p_1(\theta)\cdot u\ d\mu\ du\cdot d\theta\\
= & \ c\ p_k(\mathbf{E})(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}}p_1(\theta)\int_{u\in\mathcal{R}_+}u^{-\frac{n+2}{2}}\exp\{-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{Q}\mathbf{y}\frac{1}{u}\}du\\
& \cdot\int_{\mu\in\mathcal{R}}\exp\{-\frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{2u}\cdot(-\mu+(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{y})^2\}d\mu\cdot d\theta\\
= & \ c\ p_k(\mathbf{E})(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}}p_1(\theta)\int_{u\in\mathcal{R}_+}u^{-\frac{n+2}{2}}\exp\{-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{Q}\mathbf{y}\cdot\frac{1}{u}\}\\
& \cdot(2\pi)^{\frac{1}{2}}u^{\frac{1}{2}}(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-\frac{1}{2}}du\cdot d\theta\\
= & \ \Gamma(\frac{n-1}{2})\pi^{-\frac{n-1}{2}}\cdot|\mathbf{\Sigma}|^{-\frac{1}{2}}(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-\frac{1}{2}}(T_S(\mathbf{y})^T\mathbf{\Sigma}^{-1}\mathbf{Q}T_S(\mathbf{y}))^{-\frac{n-1}{2}}\\
& \cdot c\ p_k(\mathbf{E})p_1(\theta)d\theta\cdot\|\mathbf{y}-\bar{y}\mathbf{1}\|^{1-n}\\
\propto & \ f(T_S(\mathbf{y})|(\mathbf{E},\theta))\cdot p'(\mathbf{E},\theta)d\theta.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\int_{\vartheta\in T_\Theta^{-1}((\mathbf{E},\theta))}f(\vartheta|\mathbf{y})d\vartheta &= \frac{\int_{\vartheta\in T_\Theta^{-1}((\mathbf{E},\theta))}f(\mathbf{y}|\vartheta)p(\vartheta)d\vartheta}{\int_{\vartheta_1\in\Theta}f(\mathbf{y}|\vartheta_1)p(\vartheta_1)\nu(d\vartheta_1)}\\
&= \frac{\int_{\vartheta\in T_\Theta^{-1}((\mathbf{E},\theta))}f(\mathbf{y}|\vartheta)p(\vartheta)d\vartheta}{\sum_{\mathbf{E}_1}\int_{\theta_1\in\mathcal{R}_+}\left[\int_{\vartheta\in T_\Theta^{-1}((\theta_1,\mathbf{E}_1))}f(\mathbf{y}|\vartheta)p(\vartheta)d\vartheta\right]}\\
&= \frac{f(T_S(\mathbf{y})|(\mathbf{E},\theta))p'(\mathbf{E},\theta)d\theta}{\sum_{\mathbf{E}_1}\int_{\theta_1\in\mathcal{R}_+}f(T_S(\mathbf{y})|(\mathbf{E}_1,\theta_1))p'(\mathbf{E}_1,\theta_1)d\theta_1}\\
&= f((\mathbf{E},\theta)|T_S(\mathbf{y}))d\theta.
\end{aligned}
$$

So the diagram is commutative. #

Based on Proposition 3.1, the prior (3.6) defined on $\{(\mu,\sigma^2,\sigma_\tau^2,\mathbf{E})\}$ for the original

model (3.2) is compatible with the corresponding prior (3.7) for the reduced model depending only on the residual configuration statistic. Specifically, if we choose

$$p_1(\theta) \propto \theta^{\alpha-1}/(1+\theta)^{2\alpha}$$

with $\alpha > 0$, the Bayesian inference on variety effects is consistent with the marginal posterior distribution on partitions calculated in Section 3.2.5 .

3.4.2 Posterior Distribution on Variety Effects

Fixing any partition $\mathbf{E}$, let b_i, $\bar{y}_i$ be the number of observations and the sample mean in the ith block respectively as in Section 3.2.4 . Let $i(r)$ be the block in which variety r occurs. Given the sample $\mathbf{y}$, the marginal posterior distribution of τ conditional on $(\mathbf{E}, \theta)$ is the multivariate t distribution ([28]) defined on $\mathcal{R}^k$ with location vector $\hat{\tau}^{\mathbf{E}} = (\hat{\tau}_1^{\mathbf{E}}, \ldots, \hat{\tau}_k^{\mathbf{E}})^T$, scale matrix $\mathbf{\Lambda}$ and $n-1$ degrees of freedom, where

$$\hat{\tau}_r^{\mathbf{E}} = \bar{y}_{i(r)} - \frac{1}{1+\theta b_{i(r)}}(\bar{y}_{i(r)} - \bar{y}_w),$$

$$\Lambda_{rs} = \frac{1}{n-1}(\mathbf{y}^T \mathbf{\Sigma}^{-1} \mathbf{Q} \mathbf{y}) \left[\frac{\theta \mathbf{E}_{rs}}{1+\theta b_{i(r)}} + \frac{1/\mathbf{\Sigma} w_i}{(1+\theta b_{i(r)})(1+\theta b_{i(s)})} \right].$$

Furthermore, if the varieties r and s are not in the same block of $\mathbf{E}$, the marginal posterior distribution of the variety contrast $\tau_r - \tau_s$ conditional on $(\mathbf{E}, \theta)$ satisfies

$$\frac{(\tau_r - \tau_s) - (\hat{\tau}_r^{\mathbf{E}} - \hat{\tau}_s^{\mathbf{E}})}{(\mathbf{D}_{rs}^T \mathbf{\Lambda} \mathbf{D}_{rs})^{\frac{1}{2}}} \sim t(n-1),$$

where the elements of $\mathbf{D}_{rs} \in \mathcal{R}^k$ are all zero except that the rth one is 1 and the sth one is -1. Otherwise, $\tau_r - \tau_s \equiv 0$.

After averaging over θ and $\mathbf{E}$, the marginal posterior distribution of the variety contrast $\tau_r - \tau_s$ is a mixture, which has an atom at 0 with probability $\Pr(r \sim s \,|\, \mathbf{y}) = \sum_{\mathbf{E}:\mathbf{E}_{rs}=1} p(\mathbf{E}|\mathbf{y})$ and a continuous component with density function

$$\sum_{\mathbf{E}:\mathbf{E}_{rs}=0} p(\mathbf{E}|\mathbf{y}) \int_{\theta>0} f(\tau_r - \tau_s|\theta, \mathbf{E}, \mathbf{y}) f(\theta|\mathbf{E}, \mathbf{y}) d\theta. \tag{3.8}$$

Both the marginal posterior distribution of θ and the marginal posterior distribution of $\mathbf{E}$ have the forms given in Section 3.2.5 .

3.4.3 Example 2: Dyestuff Data (Continued)

With $\alpha = 1$ and $\lambda = 1$ in the prior for $(\mathbf{E}, \theta)$, Figure 3.1 shows the continuous component of the marginal posterior distribution of the greatest observed difference $\tau_5 - \tau_6$, as well as the marginal posterior distributions from the classical random and fixed effect models in Box and Tiao ([11], sec. 7.2). Indeed, the distribution from the random effect model can also be derived from the partition model by fixing

$\mathbf{E} = 1|2|3|4|5|6$ and setting $p_1(\theta) = 1/(1 + k\theta)$, which is the Box and Tiao prior.

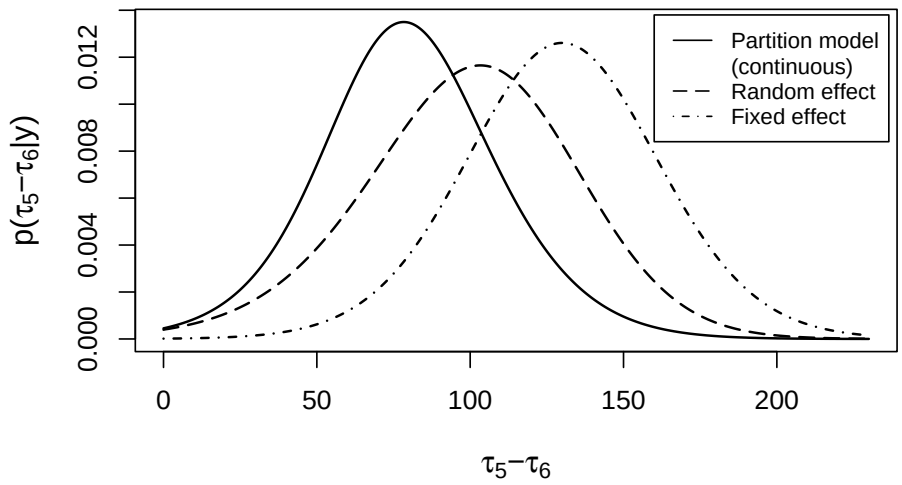


Figure 3.1: Marginal Posterior Distributions of $\tau_5 - \tau_6$ for the Dyestuff Data

Based on (3.8), the continuous component of $p(\tau_5 - \tau_6|\mathbf{y})$ from the partition model is indeed a linear combination of $p(\tau_5 - \tau_6|\mathbf{E}, \mathbf{y})$'s such that $\mathbf{E}_{56} = 0$. The shrinkage towards zero is more severe than the conventional random effect model. Besides that, unlike the random or fixed effect models, it allows a positive probability for $p(\tau_5 = \tau_6|\mathbf{y})$.

Table 3.7: Posterior Probabilities on Variety Contrast $\tau_5 - \tau_6$ $\times 100\%$

Models	$\tau_5 < \tau_6$	$\tau_5 = \tau_6$	$\tau_5 > \tau_6$
Fixed effect model	0.02	0	99.98
Random effect model	0.32	0	99.67
Model with partitions	0.42	4.97	94.61

CHAPTER 4

APPLICATION TO CLUSTER ANALYSIS

The aim of cluster analysis or unsupervised classification is to cluster the subjects into homogeneous groups based their observed features (see [25] for a good review). A set of non-overlapping clusters naturally forms a partition of the subjects. From a Bayesian point of view, the purpose of cluster analysis is not only to identify the modal partition, but to make inference on the marginal posterior distribution on partitions. In this chapter, we apply the partition model described in the previous chapter to cluster analysis. Since the number of partitions B_n increases rapidly with n, the marginal posterior probabilities of partitions are not practical to be calculated explicitly for large n. Instead, the Markov Chain Monte Carlo (MCMC) methods are used to estimate the posterior probabilities that two subjects belong to the same block, as well as the posterior distribution on the number of blocks. We describe two different Metropolis-Hastings algorithms for this purpose.

4.1 A Partition Model for Cluster Analysis

Consider a special case of the model (3.2). There are n varieties numbered from 1 to n. Each variety has only one observation. Denote by y_i the observed value

belonging to the variety i, $i = 1, \ldots, n$. Throughout this chapter, we consider $y_i \in \mathcal{R}$ only. For applications to higher-dimensional cases, please see [39]. Denote by $\mathbf{E}$ the homogeneous relationship among the n varieties, which is not observed. The model is as follows

$$\begin{aligned} \mathbf{Y} \mid \mu, \sigma^2, \sigma_\tau^2, \mathbf{E} &\sim N_n(\mu \mathbf{1}, \sigma^2 \mathbf{I_n} + \sigma_\tau^2 \mathbf{E}) \quad \text{on } \mathcal{R}^n, \\ \mathbf{E} &\sim p_n(\cdot) \quad \text{on } \mathcal{E}_n. \end{aligned}$$

Evidently, if additional variety $n+1$ with observation y_{n+1} is considered, the joint density of $(\mathbf{E}, \mathbf{y})$ satisfies the Kolmogorov consistency condition ([38]):

$$p_n(\mathbf{E}, y_1, \ldots, y_n) = \sum_{\mathbf{E}' \in \pi_{[n]}^{-1}(\mathbf{E})} \int_0^\infty p_{n+1}(\mathbf{E}', y_1, \ldots, y_n, y_{n+1}) dy_{n+1},$$

where $\pi_{[n]} : \mathcal{E}_{n+1} \to \mathcal{E}_n$ is the deletion operator that removes the $(n+1)$th row and column.

In all calculations in this chapter, $p_n(\cdot) = p_n(\cdot; \lambda)$ is the Ewens sampling distribution with parameter λ. The corresponding joint density of $(\mathbf{E}, \mathbf{y})$ belongs to the *Gauss-Ewens cluster process* ([38]), which is infinitely exchangeable.

For the same reason as in Section 3.4.1, we choose the prior on $(\mu, \sigma^2, \sigma_\tau^2)$

$$p(\mu, \sigma, \sigma_\tau^2) \propto (\sigma^2)^{-2} (\sigma_\tau^2/\sigma^2)^{\alpha-1} / (1 + \sigma_\tau^2/\sigma^2)^{2\alpha}$$

for $\alpha > 0$. Following a similar argument as in Section 3.2.5, the posterior distribution at $\mathbf{E} \in \mathcal{E}_n$ is

$$p(\mathbf{E}|\mathbf{y}) = c \cdot p_n(\mathbf{E}) \times L(\mathbf{E}; \mathbf{y}).$$

The normalized constant c is just the reciprocal of $\sum_{\mathbf{E}} p_n(\mathbf{E}) \times L(\mathbf{E}; \mathbf{y})$. It is not practical to calculate c explicitly if n is large. For example, $n \geq 20$.

4.2 Simple Metropolis-Hastings Algorithm

Fortunately, we can still use the Metropolis-Hastings algorithms to generate Markov chains with stationary distribution $p(\mathbf{E}|\mathbf{y})$ without knowing c. Then we can estimate summary statistics such as $p(\#\mathbf{E} = k|\mathbf{y})$, $E(\#\mathbf{E}|\mathbf{y})$ or $\mathrm{pr}(r \sim s|\mathbf{y})$ by Markov Chain Monte Carlo methods.

Perhaps one of the simplest Metropolis-Hastings algorithms is as follows:

0° Simulate the initial partition $\mathbf{E}_1$ of the index set $[n] = \{1, \ldots, n\}$ following the Ewens sampling distribution (see Section 2.1) .

1° Choose unit i from $[n]$ randomly. Denote by

$$B_1,\ B_2,\ \ldots,\ B_k$$

the blocks formed by the left $n-1$ elements. Denote by $i^{\#} = 1, 2, \ldots, k$ or $k+1$ the index of block where i comes from. Here $i^{\#} = k+1$ indicates that i comes

from a single-element block.

2° Randomly choose $j^{\#} = 1, 2, \ldots, k$ or $k+1$ such that $j^{\#} \neq i^{\#}$. Put the unit i into the block $B_{j^{\#}}$ and form a new partition $\mathbf{E}_2$. Again, $j^{\#} = k+1$ indicates letting i form a single-element block in $\mathbf{E}_2$.

3° Accept $\mathbf{E}_2$ with probability

$$r = \min\left\{1,\ \frac{p(\mathbf{E}_2|\mathbf{y})}{p(\mathbf{E}_1|\mathbf{y})}\right\}.$$

Otherwise keep $\mathbf{E}_1$ for the next iteration.

4° Repeat steps from 1° to 3° until the target statistics converge.

Basically, the transition distribution from $\mathbf{E}_1$ to $\mathbf{E}_2$ is just uniform. It won't reflect any information about the target distribution $p(\mathbf{E}|y)$. In a typical simulation, the average acceptance ratio is only 5% or lower. So the MCMC based on this simple algorithm is not quite efficient.

4.3 Proposed Metropolis-Hastings Algorithm

We propose the following Metropolis-Hastings algorithm. In each iteration, we either "split" a single block into two smaller ones or "combine" two blocks into a bigger one.

0° Simulate the initial partition $\mathbf{E}_1$ of $[n]$ following the Ewens sampling distribution.

1° Determine to do the "split" step or the"combine" step.

In detail, if $\#\mathbf{E}_1 = 1$, go to Step 2 ("split"); if $\#\mathbf{E}_1 = k$, go to Step 3 ("combine"); if $1 < \#\mathbf{E}_1 < k$, go to Step 2 with probability $1 - p_c$, or go to Step 3 with probability p_c. Here, p_c is pre-determined. Typically, $p_c = 0.5$.

2° If $\#\mathbf{E}_1 < k$, choose one block of $\mathbf{E}_1$ containing more than two elements, then "split" it into two nonempty blocks.

In detail, given $\mathbf{E}_1 = \{B_1, B_2, \ldots, B_L\} \in \mathcal{E}_k$ with nonzero block sizes b_1, b_2, $\ldots$, b_L, do the following steps:

a) Choose block B_l with probability proportional to the sample variance of B_l (0, if $b_l = 1$).

b) Choose $m \in \{1, 2, \ldots, b_l - 1\}$ randomly to split B_l into two nonempty blocks with block sizes m, $b_l - m$.

c) Choose $0 \leq r \leq R = \min\{m, b_l - m\}$ with probability proportional to δ^r $(0 < \delta < 1)$. For example, $\delta = 0.5$.

d) Rewrite $B_l = \{a_1, a_2, \ldots, a_{b_l}\}$ in ascending order, randomly choose r elements from $\{a_1, \ldots, a_m\}$, then exchange them with r randomly chosen elements from $\{a_{m+1}, \ldots, a_{b_l}\}$. Denote the two subsets after the exchange by B_{l1} and B_{l2}.

The new partition candidate after "split" is

$$\mathbf{E}_2 = \{B_1, B_2, \ldots, B_{l-1}, B_{l1}, B_{l2}, B_{l+1}, \ldots, B_L\} \ .$$

3° If $\#\mathbf{E}_1 > 1$, choose two blocks of $\mathbf{E}_1$ amd "combine" them into one block.

In detail, given $\mathbf{E}_1 = \{B_1, B_2, \ldots, B_L\} \in \mathcal{E}_k$ with block means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_L$, choose block B_i randomly, then choose block $B_j \neq B_i$ with probability proportional to $[(|\bar{y}_j - \bar{y}_i| + \epsilon_0]^{-1}$, where $0 < \epsilon_0 \ll 1$. For example, $\epsilon_0 = 0.0001$. Combine B_i & B_j into a new block B_{ij}. The new partition candidate $\mathbf{E}_2$ after "combine" consists of B_{ij} and the blocks of $\mathbf{E}_1$ other than B_i, B_j.

4° Calculate the acceptance ratio

$$\gamma = \min\left\{1, \frac{p(\mathbf{E}_2|y)p(\mathbf{E}_1|\mathbf{E}_2)}{p(\mathbf{E}_1|y)p(\mathbf{E}_2|\mathbf{E}_1)}\right\} = \min\left\{1, \frac{L(\mathbf{E}_2; y)}{L(\mathbf{E}_1; y)} \cdot \frac{p_k(\mathbf{E}_2)}{p_k(\mathbf{E}_1)} \cdot \frac{p(\mathbf{E}_1|\mathbf{E}_2)}{p(\mathbf{E}_2|\mathbf{E}_1)}\right\}.$$

Accept $\mathbf{E}_2$ with probability γ, otherwise keep $\mathbf{E}_1$.

In detail, if $\mathbf{E}_2$ is generated after the "combine" step,

$$p(\mathbf{E}_2|\mathbf{E}_1) = \frac{p_c^{1\{\#\mathbf{E}_1 < k\}}}{\#\mathbf{E}_1}\left[\frac{(|\bar{y}_i - \bar{y}_j| + \epsilon_0)^{-1}}{\sum_{l\neq i}(|\bar{y}_l - \bar{y}_i| + \epsilon_0)^{-1}} + \frac{(|\bar{y}_i - \bar{y}_j| + \epsilon_0)^{-1}}{\sum_{l\neq j}(|\bar{y}_l - \bar{y}_j| + \epsilon_0)^{-1}}\right].$$

Otherwise, $\mathbf{E}_2$ is generated by the "split" step. Then

$$p(\mathbf{E}_2|\mathbf{E}_1) = \frac{(1-p_c)^{1\{\#\mathbf{E}_1>1\}}\, s^2(B_l)}{(b_l-1)\sum_{i=1}^{\#\mathbf{E}_1} s^2(B_i)} \cdot \left[\frac{\delta^{r_{12}} / \sum_{i=0}^{R} \delta^i}{\binom{m}{r_{12}}\binom{b_l-m}{r_{12}}} + \frac{\delta^{r_{21}} / \sum_{i=0}^{R} \delta^i}{\binom{m}{r_{21}}\binom{b_l-m}{r_{21}}} \right],$$

where r_{12} is the number of elements needed to be exchanged to get B_{l1} & B_{l2} if B_{l1} contains the first m elements before exchange, r_{21} is the number of elements needed to be exchanged if B_{l2} contains the first $b_l - m$ elements before exchange, and $s^2(B_i)$ is the sample variance of block B_i.

5° Repeat Step 1 $\sim$ Step 4 until the target statistics got from MCMC converge.

The algorithm described above is much more complicated than the simple one in the previous section. It reflects $p(\mathbf{E}|y)$ better by splitting a block according roughly to the increasing order of the observations in it or combining two blocks if their sample means are close to each other. Besides, in each iteration, the proposed algorithm has much greater chance to change the entire partition structure such as the number of blocks. In a typical simulation, the average acceptance ratio is between 0.35 and 0.40 .

4.4 Simulation Study

4.4.1 Comparison of Two Algorithms

To compare the proposed algorithm in Section 4.3 with the simple one in Section 4.2, we simulate a data set of size 20 from

$$Y_1, ..., Y_{10} \quad \text{i.i.d.} \sim N(-2, 1);$$

$$Y_{11}, ..., Y_{20} \quad \text{i.i.d.} \sim N(2, 1).$$

For the same data set, we run 9 independent Markov chains for each algorithm to estimate $E(\#\text{block}|\mathbf{y})$ and the average of $P(i \sim j|\mathbf{y})$. Each chain starts from a randomly chosen partition E following the Ewens sampling distribution with $\lambda = 1$. For each Markov chain, $100,000$ iterations are counted after 500 burn-in iterations. It takes an Intel Pentium 2.4GHz desktop with 512 Mb of RAM about 11 minutes to complete each Markov chain.

Table 4.1 shows that the proposed algorithm converges much faster than the simple one when estimating $E(\#\text{block}|\mathbf{y})$ and $p(i \sim j|\mathbf{y})$. Indeed, the standard deviations of the estimates by the proposed algorithm are as small as 1/6 of the ones by the simple algorithm.

Table 4.1: Estimated $E(\#\text{block}|\mathbf{y})$ and $p(i \sim j|\mathbf{y})$ by Metropolis-Hastings Algorithms

	$E(\#\text{block}|\mathbf{y})$		Average of $p(i \sim j|\mathbf{y})$	
Algorithm	Simple	Proposed	Simple	Proposed
Chain 1	3.748	3.621	0.455	0.475
Chain 2	3.492	3.625	0.499	0.479
Chain 3	3.573	3.637	0.478	0.472
Chain 4	3.592	3.627	0.463	0.473
Chain 5	3.614	3.644	0.471	0.469
Chain 6	3.802	3.657	0.457	0.469
Chain 7	3.690	3.658	0.465	0.469
Chain 8	3.884	3.631	0.430	0.471
Chain 9	3.649	3.621	0.462	0.476
Mean	3.671	3.636	0.465	0.472
Std	0.123	0.015	0.019	0.003

4.4.2 Cluster Analysis

To see how much information about the clusters we can get from data, we simulate four data sets in $\mathcal{R}^1$ as follows. First, the number of clusters $k = 1, 2, 3$ or 4 is specified. Secondly, the cluster centers are simulated from $N(0, k^2)$ to make the clusters well-separated. Thirdly, the cluster labels of points are simulated from the discrete uniform distribution on $\{1/k, \ldots, 1/k\}$. Finally, the observations $y_1, \ldots, y_{20}$ are just the cluster centers plus random noises simulated i.i.d. from $N(0, 1)$. The realized block sizes with cluster centers in increasing order are $\{8, 12\}$ for $k = 2$, $\{10, 5, 5\}$ for $k = 3$, and $\{5, 3, 3, 9\}$ for $k = 4$. The four data sets are shown in the same plot (see Figure 4.1) .

We assume that the prior on the partitions is the Ewens sampling distribution with $\lambda = 1$ and the prior on the variance ratio $\theta = \sigma_\tau^2/\sigma^2$ has density $1/(1+\theta)^2$. To make

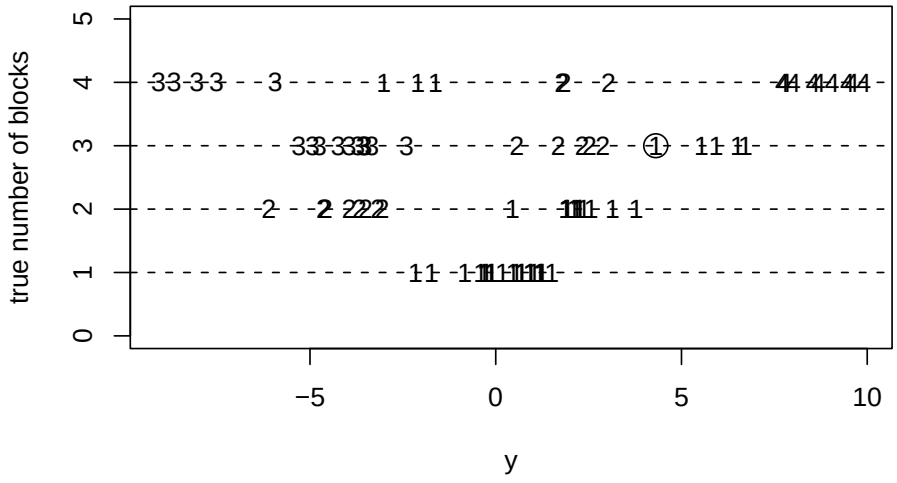


Figure 4.1: Four Simulated Data Sets Given Number of Clusters 1, 2, 3, or 4

inference on the partition $\mathbf{E}$, we use the proposed algorithm to estimate $P(\#\mathbf{E} = i|\mathbf{y})$ and $P(i \sim j|\mathbf{y}) = E(\mathbf{E}(i,j)|\mathbf{y})$, $i, j = 1, \ldots, n$. The latter $P(i \sim j|\mathbf{y})$ indicates how likely the ith and jth points belong to the same cluster. The average estimates based on 5 independent Markov chains are summarized in Table 4.2, Table 4.3 and Table 4.4 . Chains as long as 200,000 after 500 burn-in iterations are used to make sure the target statistics converge. For illustrative purpose, only part of $P(i \sim j|\mathbf{y})$'s are list in Table 4.3 and Table 4.4 .

In Table 4.2, $P(\#\mathbf{E} = \cdot|\mathbf{y})$ for Case 1 is fairly close to the Ewens' prior. Roughly speaking, the smallest i such that $P(\#\mathbf{E} = i|\mathbf{y})$ is greater than the prior or at least comparable with the prior might indicate the number of observed clusters. Bigger probability shows stronger evidence. For example, the first significantly large item

Table 4.2: Posterior Distribution of Number of Blocks, $P(\#\mathbf{E}|\mathbf{y}) \times 1000$

$\#\mathbf{E}$	1	2	3	4	5	6	7	8	9	10	11	$E(\#\mathbf{E}\|y)$
Case 1	44	167	273	257	160	69	23	6	1	0	0	3.66
Case 2	0	196	344	274	131	43	10	2	0	0	0	3.52
Case 3	7	50	222	325	243	110	34	7	1	0	0	4.26
Case 4	5	23	59	261	344	211	75	17	3	0	0	4.95
Ewens	50	177	275	251	153	66	21	5	1	0	0	3.60

for Case 2 is $P(\#\mathbf{E} = 2|\mathbf{y}) = 0.196$ while $P(\#\mathbf{E} = 2)$ is 0.177 in the prior. Table 4.2 also shows that $P(\#\mathbf{E} = 1|\mathbf{y})$ is almost 0. Indeed, the 5 Markov chains of length $200,000$ visit the single-block partition 14 times on average. The evidence for $\#\mathbf{E} = 2$ in Case 2 is fairly strong.

Table 4.3: Posterior Probabilities for Case 2, $P(i \sim j|\mathbf{y}) \times 100$

i, j	1	3	5	7	8	9	11	13	15	17	19	20
1	100	65	58	55	54	0	0	0	0	0	0	0
3	65	100	78	75	74	0	0	0	0	0	0	0
5	58	78	100	80	80	0	0	0	0	0	0	0
7	55	75	80	100	80	0	0	0	0	0	0	0
8	54	74	80	80	100	0	0	0	0	0	0	0
9	0	0	0	0	0	100	70	69	69	68	66	64
11	0	0	0	0	0	70	100	87	87	87	85	82
13	0	0	0	0	0	69	87	100	87	88	86	82
15	0	0	0	0	0	69	87	87	100	88	86	83
17	0	0	0	0	0	68	87	88	88	100	86	84
19	0	0	0	0	0	66	85	86	86	86	100	85
20	0	0	0	0	0	64	82	82	83	84	85	100

The posterior probabilities that two units belong to the same block in Table 4.3 also support the two-block conclusion for Case 2. Clearly, the matrix $P(i \sim j|\mathbf{y})$ identifies the true partition for Case 2, where the indices $1, \ldots, 20$ are corresponding to the reordered points in increasing order. Because the two clusters in Case 2 are

well separated (see Figure 4.1) .

Table 4.4: Posterior Probabilities for Case 3, $P(i \sim j|\mathbf{y}) \times 100$

i, j	1	3	6	9	10	11	13	15	16	17	19	20
1	100	85	83	82	76	11	6	6	5	5	4	4
3	85	100	84	83	77	11	6	6	5	5	5	4
6	83	84	100	83	79	11	6	6	5	5	5	5
9	82	83	83	100	79	12	7	6	5	5	5	5
10	76	77	79	79	100	13	7	6	6	5	5	5
11	11	11	11	12	13	100	54	48	24	14	13	13
13	6	6	6	7	7	54	100	65	38	24	22	21
15	6	6	6	6	6	48	65	100	43	28	25	25
16	5	5	5	5	6	24	38	43	100	58	55	54
17	5	5	5	5	5	14	24	28	58	100	77	76
19	4	5	5	5	5	13	22	25	55	77	100	81
20	4	4	5	5	5	13	21	25	54	76	81	100

Case 3 shows an example that the clusters in the data set are not separated very well. Note that $P(\#\mathbf{E} = 2|\mathbf{y}) = 0.050$ for Case 3 (see Table 4.2), which is a little less than 1/3 of the corresponding prior probability 0.177. If we denote by $\mathbf{E}'$ the partition generated by combining the clusters 1 and 2 in the true partition $\mathbf{E}^0$ to form a two-block partition, then $P(\mathbf{E}'|\mathbf{y})/P(\mathbf{E}^0|\mathbf{y}) = 0.25$. If we check $P(i \sim j|\mathbf{y})$'s in Table 4.4, the smallest item is $P(11 \sim 20|\mathbf{y}) = 0.13$ if both i and j belong to the clusters 1 and 2. Besides, the clustering information based on the matrix $P(i \sim j|\mathbf{y})$ is not so clear as in Case 2. The 16th point, which is the circled "1" in Figure 4.1, also has chance to be clustered into the cluster 2. Indeed, the probability ratio $P(\mathbf{E}''|\mathbf{y})/P(\mathbf{E}^0|\mathbf{y}) = 0.46$, where $\mathbf{E}''$ indicates the partition clustering 16 into the block $\{11, 12, \ldots, 15\}$ and keeping the other relationships as in $\mathbf{E}^0$.

Typically, $P(\#\mathbf{E} = i|\mathbf{y})$ is still large if i is bigger than the true number of clusters

by 1, 2 or even 3. For example, $P(\#\mathbf{E} = i|y) = 0.325$, 0.243 or 0.110 for $i = 4, 5$ or 6 in Case 3. It suggests $E(\#\mathbf{E}|y)$ may not be a good estimate for the number of observed clusters unless the clusters are separated very well. Fortunately, the matrix $P(i \sim j|\mathbf{y})$ is fairly informative for clustering.

PART II: PERMANENT PROCESS AND CLASSIFICATION MODELS

CHAPTER 5

PERMANENT PROCESS

In this chapter, we review the permanent process described by McCullagh and Møller ([36]). We develop an algorithm for searching the maximum likelihood estimate for the parameters of the permanent process.

5.1 Permanent Polynomial

For each n by n matrix $K = (K_{ij})$, there is a polynomial of degree n

$$\mathrm{per}_\alpha(K) = \sum_\sigma \alpha^{\#\sigma} K_{1\sigma(1)} \cdots K_{n\sigma(n)} \,,$$

where the sum runs over all permutations of $\{1, 2, \ldots, n\}$ and $\#\sigma$ indicates the number of cycles of the permutation σ. For example,

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ \\ 1 & 3 & 2 \end{pmatrix} \implies \sigma = (1)(23) \implies \#\sigma = 2.$$

In the mathematical literature, $\mathrm{per}_\alpha(K)$ is called the *α-permanent* of K ([56]). Particularly, $\mathrm{per}_1(K)$ is called the *permanent* of K (see [40] for more details), and

$\mathrm{per}_{-1}(-K)$ is just the determinant of K. The quantity $\det_\alpha(A) = \alpha^n \mathrm{per}_{1/\alpha}(A)$ is called the α-*determinant* ([49]).

The coefficient of α in the permanent polynomial $\mathrm{per}_\alpha(K)$ is the sum of cyclic products ([36])

$$\mathrm{cyp}(K) = \lim_{\alpha \to 0} \alpha^{-1} \mathrm{per}_\alpha(K) = \sum_{\sigma : \#\sigma = 1} K_{1\sigma(1)} \cdots K_{n\sigma(n)} \,.$$

5.2 Gaussian Moments

The permanent polynomial arises naturally in statistical works associated with the Cox process as follows ([36]). Let Z be a Gaussian random field on the feature space $\mathcal{X}$ with mean 0 and covariance function $C/2$. In other words, $Z(x) \sim N(0, C(x,x)/2)$ and $\mathrm{Cov}(Z(x), Z(x')) = C(x,x')/2$ for any $x, x' \in \mathcal{X}$. Then the joint cumulant and the joint moment of $Z(x_1)^2, \ldots, Z(x_n)^2$ ([36], [49]) are

$$\begin{aligned} E\left(Z(x_1)^2 Z(x_2)^2 \cdots Z(x_n)^2\right) &= \mathrm{per}_{1/2}[C](X), \\ \mathrm{cum}\left(Z(x_1)^2, Z(x_2)^2, \ldots, Z(x_n)^2\right) &= \mathrm{cyp}[C](X)/2, \end{aligned}$$

where $X = (x_1, x_2, \ldots, x_n)^T$, and $[C](X)$ is the n by n matrix $(C(x_i, x_j))_{ij}$.

Proposition 5.1 *([56]) The cumulant generating function of* $(Z(x_1)^2, \ldots, Z(x_n)^2)$ *is*

$$C_X(t_1, t_2, \ldots, t_n) = -\frac{1}{2} \log \left| I_n - \mathrm{diag}(t_1, t_2, \ldots, t_n) \cdot [C](X) \right| . \tag{5.1}$$

Proof Recall three facts in matrix theory (for example, see [6]). Let A be any n by n matrix, and let $\|\cdot\|$ be a normalized submultiplicative norm (that is, $\|I_n\| = 1$, $\|AB\| \le \|A\| \cdot \|B\|$), then

1) $\log|e^A| = \mathrm{tr}(A)$, where $e^A = \sum_{k=0}^{\infty} A^k/k!$;

2) $\log(I_n - A) = -\sum_{i=1}^{\infty} A^i/i$ is well defined, if $\|A\| < 1$;

3) $e^{\log(I_n - A)} = I_n - A$, if $\|A\| < 1$.

Thus, $\log|I_n - A| = \log|e^{\log(I_n - A)}| = \mathrm{tr}(\log(I_n - A)) = -\sum_{k=1}^{\infty} \mathrm{tr}(A^k)/k$, if $\|A\| < 1$.

Let $A = \mathrm{diag}(t_1, t_2, \ldots, t_n) \cdot [C](X) = (t_i C_{ij})$, where $C_{ij} = C(x_i, x_j)$. If $\max_i |t_i|$ is small enough such that $\|\mathrm{diag}(t_1, \ldots, t_n)\| < \|[C](X)\|^{-1}$, then $\|A\| < 1$. The right hand of (5.1) is

$$-\frac{1}{2}\log|I_n - \mathrm{diag}(t_1, t_2, \ldots, t_n) \cdot [C](X)| = -\frac{1}{2}\log|I_n - A| = \frac{1}{2}\sum_{k=1}^{\infty} \mathrm{tr}(A^k)/k,$$

where $\mathrm{tr}(A^k) = \sum_{i_1,\ldots,i_k} t_{i_1} \cdots t_{i_k} C_{i_1,i_2} C_{i_2,i_3} \cdots C_{i_k,i_1}$. The left hand of (5.1) is

$$\begin{aligned}
& C_X(t_1, t_2, \ldots, t_n) \\
= & \sum_{k=1}^{\infty} \frac{1}{k!} \sum_{i_1,\ldots,i_k} t_{i_1} \cdots t_{i_k} \mathrm{cum}(Z(x_{i_1})^2, \ldots, Z(x_{i_k})^2) \\
= & \frac{1}{2}\sum_{k=1}^{\infty} \frac{1}{k!} \sum_{i_1,\ldots,i_k} t_{i_1} \cdots t_{i_k} \mathrm{cyp}[C](x_{i_1}, \ldots, x_{i_k}) \\
= & \frac{1}{2}\sum_{k=1}^{\infty} \frac{1}{k!} \sum_{i_1,\ldots,i_k} t_{i_1} \cdots t_{i_k} \sum_{\sigma \in S_k, \#\sigma = 1} C_{i_1, i_{\sigma(1)}} C_{i_{\sigma(1)}, i_{\sigma^2(1)}} \cdots C_{i_{\sigma^{k-1}(1)}, i_1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k!}\sum_{\sigma\in S_k,\#\sigma=1}\sum_{i_1,\ldots,i_k} t_{i_1}\cdots t_{i_k}C_{i_1,i_{\sigma(1)}}C_{i_{\sigma(1)},i_{\sigma^2(1)}}\cdots C_{i_{\sigma^{k-1}(1)},i_1}\\
&= \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k!}\cdot(k-1)!\sum_{i_1,\ldots,i_k} t_{i_1}\cdots t_{i_k}C_{i_1,i_2}\cdots C_{i_k,i_1}\\
&= \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k}\mathrm{tr}(A^k)
\end{aligned}
$$

Thus, (5.1) is true, if $\|\mathrm{diag}(t_1,\ldots,t_n)\| < \|[C](X)\|^{-1}$. #

Corollary 5.1 *The moment generating function of* $(Z(x_1)^2,\ldots,Z(x_n)^2)$ *is*

$$M_X(t_1,t_2,\ldots,t_n) = |I_n - \mathrm{diag}(t_1,t_2,\ldots,t_n)\cdot[C](X)|^{-\frac{1}{2}}, \tag{5.2}$$

if $\|\mathrm{diag}(t_1,\ldots,t_n)\| < \|[C](X)\|^{-1}$.

Note that Corollary 5.1 is a special case of the theorem proved by D. Vere-Jones ([56]).

5.3 Density Function

The *permanent process* on $\mathcal{X}$ is a Cox process with the intensity function

$$\Lambda(x) = \sum_{r=1}^{k} Z_r^2(x),$$

where $Z_1,\ldots,Z_k$ are independent and identically distributed Gaussian random fields on $\mathcal{X}$ with zero mean and covariance function $C/2$. So C is symmetric and positive

definite. For many applications, $\mathcal{X} = \mathcal{R}^d$.

Typically, a spatial point pattern $\{x_1, \ldots, x_n\}$ is observed within a compact subset S, or a bounded window, in $\mathcal{X}$. If C is also bounded on S, it has the spectral representation ([23]):

$$C(x, x') = \sum_{r=0}^{\infty} \lambda_r e_r(x) e_r(x'),\ \forall x, x' \in S,$$

where $\{\lambda_r\}_r$ and $\{e_r\}_r$ are the eigenvalues and the normalized eigenfunctions of C respectively. In other words,

$$\int_S C(x, x') e_r(x) dx = \lambda_r e_r(x'),\ \forall x' \in S,$$

$$\int_S e_r(x) e_s(x) dx = \delta_{rs}.$$

Due to the symmetry of C, $\{e_r\}_r$ forms an orthonormal basis of $\mathcal{L}_2(S)$. Because C is positive definite, $\lambda_r \geq 0$. If we write $\tilde{\lambda}_r = \lambda_r/(1+\lambda_r)$, we can define a new covariance function

$$K(x, x') = \sum_{r=0}^{\infty} \tilde{\lambda}_r e_r(x) e_r(x').$$

McCullagh and Møller ([36]) obtained the marginal density of the permanent process

$$f(\mathbf{X}) = e^{|S| - \alpha D} \mathrm{per}_{\alpha}[K](\mathbf{X}),$$

where $\mathbf{X} = (x_1, x_2, \ldots, x_n)^T$, $\alpha = k/2$, $D = \sum_{r=0}^{\infty} \log(1+\lambda_r) = -\sum_{r=0}^{\infty} \log(1-\tilde{\lambda}_r)$.

Unlike general Cox processes, the permanent process has its density function in explicit form. The flexibility in choosing k (or α) and C makes the permanent process potentially useful.

5.4 Numerical Computation

Consider a specific case. Let $S = [0,T] \times [0,T] \subseteq \mathcal{R}^2$,

$$C(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\{-\|\mathbf{x} - \mathbf{x}'\|^2/\tau^2\} = C(x, x') \cdot C(y, y'), \tag{5.3}$$

where $\mathbf{x} = (x,y)^T$, $\mathbf{x}' = (x',y')^T$, $C(x,x') = \sigma \exp\{-(x-x')^2/\tau^2\}$. Given the observations $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \subseteq S$, where $\mathbf{x}_i = (x_i, y_i)^T, i = 1, 2, \ldots, n$, we want to calculate the log likelihood

$$l(\alpha, \sigma, \tau) = \log \mathrm{per}_\alpha[K](\mathbf{X}) + |S| - \alpha D.$$

5.4.1 Proposed Algorithm

We propose an algorithm to calculate $l(\alpha, \sigma, \tau)$ as follows:

1° Find the eigenvalues and eigenfunctions of C on $[0,T]$.

Fix a large enough integer m. Divide $[0,T]$ equally into m subintervals. Let $a_i = (i - 1/2)T/m,\ i = 1, 2, \ldots, m$ be the centers of the subintervals.

If λ_r and e_r are the eigenvalue and eigenfunction of C respectively, then

$$\sum_{j=1}^{m} C(a_i, a_j) e_r(a_j) \doteq \frac{m}{T} \lambda_r \cdot e_r(a_i).$$

Calculate the eigenvalue $\lambda_r m/T$ and the eigenvector $e_r(\mathbf{a})$ of $C(\mathbf{a}, \mathbf{a}^T)$, where $\mathbf{a} = (a_1, \ldots, a_m)^T$. Standardize $e_r(\mathbf{a})$.

2° Estimate $e_r(x_i), e_r(y_i), i = 1, 2, \ldots, n$ based on linear interpolation or extrapolation.

3° Estimate D using

$$D \doteq \sum_{r,s=1}^{m} \log(1 + \lambda_r \lambda_s).$$

4° Estimate K using

$$K(\mathbf{x}_i, \mathbf{x}_j) \doteq \sum_{r,s=1}^{m} \lambda_r \lambda_s \cdot e_r(x_i) e_s(y_i) \cdot e_r(x_j) e_s(y_j).$$

5° Estimate the log likelihood using

$$l(\alpha, \sigma, \tau) = \log(\mathrm{per}_\alpha[K](\mathbf{X})) + |S| - \alpha D.$$

5.4.2 Numerical Illustration

In this section, we use a numerical example to illustrate how to calculate the log likelihood given the observations. Furthermore, we can get the maximum likelihood estimate (MLE) for the parameters.

Let $S = [0, 2\pi] \times [0, 2\pi]$. Thus $|S| = 4\pi^2$. We simulate $n = 20$ data points uniformly in S, which can be found in Figure 5.1(a). We assume the permanent model with covariance function described in (5.3). For illustrative purpose, we assume $\alpha = 1$ to simplify the calculation. Then the purpose is to calculate the log likelihood $l(\sigma, \tau)$ and the MLE for (σ, τ).

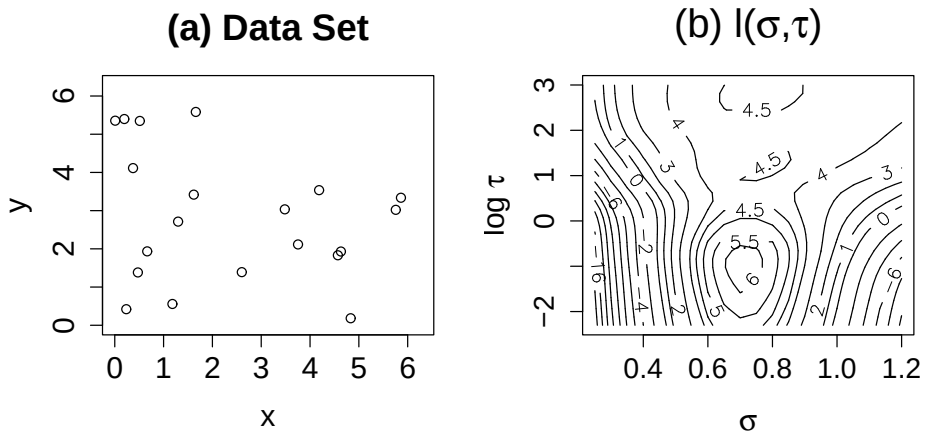


Figure 5.1: Simulated Data Set and Contour Plot for Log-Likelihood

Based on the algorithm described in Section 5.4.1, we calculate the log likelihood $l(\sigma, \tau)$ for different combinations of the parameters. For this particular case, $m = 300$ is used as the number of subintervals. Figure 5.1(b) shows the contour plot for $l(\sigma, \tau)$. Note that there are more than one local maximum. After further exploration, we get

the MLE

$$\hat{\sigma} = 0.71, \ \hat{\tau} = 0.38 \ .$$

CHAPTER 6

CLASSIFICATION MODELS

In this chapter, an exchangeable cluster process based on the permanent process is constructed for classification problems. In the corresponding classification model, only 2-3 parameters need to be estimated, regardless of the number of classes or the dimension of the feature space.

6.1 Remarks on the Literature

In the literature, there are two kinds of classification problems, supervised or unsupervised (see [25], [42] for a good review). Given observations with information on measured features and class labels, the aim of supervised classification is to classify a new unit u on the basis of its measured features $x(u) \in \mathcal{X}$, the feature space. In this chapter, we focus on supervised classification problems.

If the inference based on the classification model provides a probability distribution on the set of class labels, the model is called *stochastic classification model* ([38]). Besides, if the set of classes is pre-determined and fixed, the classification model is called *closed*. Otherwise, if the model permits a new unit to be assigned to a class that has not been observed, it is called *open classification model*.

The modern theory on supervised classification problems begins with Fisher's discriminant model ([22], [47]). Logistic regression model was proposed first in 1960s (see [48] for a good review). Most stochastic classification models used in the literature assume that the class labels Y_i's are independent. Many efforts are made to estimate the functions f_r such that

$$\log(\text{pr}(Y(u) = r|X)) = f_r(X(u)).$$

To get good estimate for f_r, we need either stronger assumption to narrow the candidate set, or large sample size to gather enough information. As the number of classes or the dimension of feature space increases drastically, the mission becomes hopeless.

The approach proposed in this chapter is quite different. The goal is to construct a classification model with no more than 4-5 parameters regardless of the number of classes or the dimension of the feature space. Besides, we assume exchangeable components instead of independent ones.

6.2 A Marked Point Process

Let $\mathbf{X}$ be a general Poisson process in the feature space $\mathcal{X}$ with intensity function $\lambda(x)$ with respect to some baseline measure μ on $\mathcal{X}$. Then for each measurable set $A \subset \mathcal{X}$, $X(A)$, which is the number of events occurred in A, follows the usual Poisson distribution with parameter $\int_A \lambda(x)\mu(dx)$. In addition, for any two non-overlapping

subsets A and A' of $\mathcal{X}$, the event counts $X(A)$ and $X(A')$ are independent. More details about Poisson process could be found in [32].

If the Poisson process $\mathbf{X}$ is driven by a random intensity function $\Lambda(x)$, it is known as a *Cox process*, or a doubly stochastic Poisson process ([32], [17]). As mentioned in Section 5.3, the permanent process is a special kind of the Cox process, whose marginal density function can be written in explicit form.

Let μ be a non-random baseline measure in $\mathcal{X}$. Given a random non-negative intensity function $\Lambda(x)$, the expected number of events following the associated Cox process in an infinitesimal ball dx centered at x is $E(\Lambda(x))\mu(dx)$. Furthermore, the expected number of ordered pairs of distinct events in $dx\ dx'$ is $E(\Lambda(x)\Lambda(x')\mu(dx)\mu(dx')$. In general, for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we call

$$m^{(n)}(\mathbf{x}) = E(\Lambda(x_1) \cdots \Lambda(x_n)) \tag{6.1}$$

the nth order product density at x. It indicates how likely we can get the observations $\{x_1, \ldots, x_n\}$.

Suppose $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are k independent Cox processes on $\mathcal{X}$ driven by independent random intensity functions $\Lambda_1(x), \ldots, \Lambda_k(x)$ with respect to the same baseline measure μ. The marked process can be presented by $(\mathbf{X}, y)$. Here, $\mathbf{X} = \cup \mathbf{X}^{(r)}$ is the superposition process, which is still a Cox process with intensity function $\Lambda_\bullet(x) = \sum_r \Lambda_r(x)$, and y is the list of labels indicating which Cox process the obser-

vation belongs to. Then the conditional distribution of y given the observed feature vectors $\mathbf{x} \in \mathcal{X}^n$ is

$$p_n(y|\mathbf{x}) = \frac{\prod_r m_r(\mathbf{x}^{(r)})}{m_\bullet(\mathbf{x})}, \tag{6.2}$$

where $\mathbf{x}^{(r)}$ are observations labelled r, m_r and $m_\bullet$ are corresponding to the product density functions (6.1) for the rth Cox process and the superposition Cox process respectively. For the empty set, we define $m_r(\emptyset) = 1$, which permits a class having not been observed yet. In other words, the model is an open classification model. Besides, the conditional distribution remains invariant if we change the order of the paired observations $(x_1, y_1), \ldots, (x_n, y_n)$. Here x_i and y_i are observed feature vector and class label respectively. So we only assume the observations are exchangeable.

For prediction purpose, we are interested in classifying a new unit u' based on its observed feature variables x'. The the conditional probability assigning u to class r can be derived directly from (6.2). That is,

$$p_{n+1}(y(u') = r|\text{data}) \propto m_r(\mathbf{x}^{(r)} \cup \{x'\})/m_r(\mathbf{x}^{(r)}). \tag{6.3}$$

Specifically, the probability that u' is assigned to a class r which has not been observed is proportional to $E(\Lambda(x'))$.

6.3 Permanent Cluster Process

Suppose each component process $\mathbf{X}^{(r)}$ in Section 6.2 is a permanent process with parameter α_r and covariance function $C/2$. In other words, $\mathbf{X}^{(r)}$ is driven by the random intensity function

$$\Lambda_r(x) = Z_1^{(r)}(x)^2 + \cdots Z_{2\alpha_r}^{(r)}(x)^2,$$

where $Z_1^{(r)}, \ldots, Z_{2\alpha_r}^{(r)}$ are i.i.d. copies of the Gaussian random field Z with mean 0 and covariance function $C/2$.

McCullagh and Møller ([36], Theorem 1) proved

$$E(\Lambda_r(x_1) \cdots \Lambda_r(x_n)) = \mathrm{per}_{\alpha_r}[C](x_1, \ldots, x_n),$$

for any $x_1, \ldots, x_n \in \mathcal{X}$. In other words, the product density for process r is $m_r(\mathbf{x}) = \mathrm{per}_{\alpha_r}[K](\mathbf{x})$. Besides, the superposition process $\mathbf{X}$ of the k independent permanent processes is still a permanent process ([36]). The parameters of $\mathbf{X}$ are $\alpha_\bullet = \sum_r \alpha_r$ and $C/2$. So the product density for $\mathbf{X}$ is $m_\bullet(\mathbf{x}) = \mathrm{per}_{\alpha_\bullet}[C](\mathbf{x})$. By (6.2), the conditional distribution of the labels y given feature observations $\mathbf{x}$ is

$$p_n(y|\mathbf{x}) = \frac{\mathrm{per}_{\alpha_1}[C](\mathbf{x}^{(1)}) \cdots \mathrm{per}_{\alpha_k}[C](\mathbf{x}^{(k)})}{\mathrm{per}_{\alpha_\bullet}[C](\mathbf{x})}. \tag{6.4}$$

For a new unit u' with observed features x', the conditional probability of class r

following (6.3) is proportional to the permanent ratio

$$p_{n+1}(y(u') = r|\text{data}) \ \propto \ \text{per}_{\alpha_r}[C](\mathbf{x}^{(r)}, x')/\text{per}_{\alpha_r}[C](\mathbf{x}^{(r)}). \tag{6.5}$$

We define $\text{per}_{\alpha}[C](\emptyset) \equiv 1$. For a class r which has not been observed, $\mathbf{x}^{(r)}$ is empty, then (6.5) yields $p_{n+1}(y(u') = r|\text{data}) \ \propto \ \alpha_r C(x', x')$.

Same as in Section 6.2, both (6.4) and (6.5) keep invariant if we change the order of observations $(x_i, y_i)_{i=1,\ldots,n}$. So we only need to assume the observations are exchangeable. Besides, the classification model based on the permanent cluster process is open. Because the model allows a new unit to be assigned to a class r which has not been observed yet.

CHAPTER 7

PERMANENT RATIO APPROXIMATION

In this chapter, we propose analytic approximations for the ratio of two α-permanents. It's valid for large α, but also reasonably accurate for α as small as 1.

7.1 Cyclic Approximations for Permanent Ratio

To make the permanent model applicable, we need to calculate the permanent ratio

$$R(t;\mathbf{x}) = \frac{\mathrm{per}_\alpha[K](\{t\}\cup\mathbf{x})}{\mathrm{per}_\alpha[K](\mathbf{x})}\,, \tag{7.1}$$

where K is the given covariance function. Unfortunately, exact computation of permanents is a NP-hard problem. The best of approximation algorithms ([8]) runs at an unappealing rate $O(n^7\log^4 n)$. For these reasons we propose an analytic approximation as follows.

The α-permanent of the matrix $[K](t,x_1\ldots,x_n)$ is a sum over $(n+1)!$ terms. In a subset consisting of $n!$ terms, the index t occurs in a cycle of length 1, giving rise to the partial sum

$$\alpha K(t,t)\mathrm{per}_\alpha[K](\mathbf{x})\,.$$

The index t may also occur in a cycle of length two such as (t, x_1) or (t, x_2) and so on. There are $n!$ permutations in which t occurs in a 2-cycle, giving rise to the additional sum

$$\sum_{i=1}^{n} \alpha K(t, x_i) K(x_i, t) \mathrm{per}_\alpha[K](\mathbf{x}_{-i}) \, ,$$

where $\mathbf{x}_{-i}$ is the set of $n-1$ points with the ith element removed. Similarly, the index t may occur in a 3-cycle such as (t, x_i, x_j) or (t, x_j, x_i), giving rise to the sum

$$\sum_{i \neq j} \alpha K(t, x_i) K(x_i, x_j) K(x_j, t) \mathrm{per}_\alpha[K](\mathbf{x}_{-i-j}) \, .$$

In the cycle expansion of the permanent of order $n+1$, there are $n!$ terms in which t occurs in a 1-cycle, $n!$ terms in which t occurs in a 2-cycle, $n!$ terms in which t occurs in a 3-cycle, and so on up to cycles of length $n+1$. Therefore, we obtain the following finite expansion by cycles for (7.1):

$$\begin{aligned} R_n(t; \mathbf{x}) \;=\; & \alpha K(t,t) + \alpha \sum_i \frac{1}{R_{n-1}(x_i; \mathbf{x}_{-i})} \Big(|K(t, x_i)|^2 + \\ & \sum_{j \neq i} \frac{1}{R_{n-2}(x_j; \mathbf{x}_{-i-j})} \Big(K(t, x_i) K(x_i, x_j) K(x_j, t) + \\ & \sum_{k \neq i,j} \frac{1}{R_{n-3}(x_k; \mathbf{x}_{-i-j-k})} \Big(K(t, x_i) K(x_i, x_j) K(x_j, x_k) K(x_k, t) + \cdots \Big) \Big) \Big) \end{aligned}$$

This cycle expansion suggests a recursive approximation in which $R_n^0(t; \mathbf{x}) = \alpha K(t, t)$

is the uni-cycle approximation;

$$\begin{aligned} R_n^1(t;\mathbf{x}) &= \alpha K(t,t) + \alpha \sum_i |K(t,x_i)|^2 / R_{n-1}^0(x_i;\mathbf{x}_{-i}) \\ &= \alpha K(t,t) + \sum_i |K(t,x_i)|^2 / K(x_i,x_i) \end{aligned}$$

is the two-cycle approximation,

$$R_n^2(t;\mathbf{x}) = \alpha K(t,t) + \alpha \sum_i \frac{1}{R_{n-1}^1(x_i;\mathbf{x}_{-i})} \Big(|K(t,x_i)|^2 + \sum_{j\neq i} \frac{K(t,x_i)K(x_i,x_j)K(x_j,t)}{R_{n-2}^0(x_j;\mathbf{x}_{-i-j})} \Big)$$

is the three-cycle approximation, and so on. The four-cycle approximation is

$$\begin{aligned} &\alpha K(t,t) + \alpha \sum_i \frac{1}{R_{n-1}^2(x_i;\mathbf{x}_{-i})} \Big(|K(t,x_i)|^2 + \sum_{j\neq i} \frac{1}{R_{n-2}^1(x_j;\mathbf{x}_{-i-j})} \times \\ &\Big(K(t,x_i)K(x_i,x_j)K(x_j,t) + \sum_{k\neq i,j} \frac{K(t,x_i)K(x_i,x_j)K(x_j,x_k)K(x_k,t)}{R_{n-3}^0(x_k;\mathbf{x}_{-i-j-k})} \Big) \Big). \end{aligned}$$

Up to order four, this sequence is easy to compute, even for fairly large values of n. It is an asymptotic approximation for large α, so the accuracy improves as α increases.

Some specific cases are list as follows:

(i) $K(x,y) = \delta_{xy} f(x)$.

Here f is some positive non-random function on $\mathcal{X}$, and $\delta_{xy} = 1$ if $x = y$ and 0 otherwise. If $t, x_1, \ldots, x_n$ are pairwise different, then

$$R(t;\mathbf{x}) = R^0(t;\mathbf{x}) = R^1(t;\mathbf{x}) = \cdots \equiv \alpha f(t).$$

(ii) $K(x,y) \equiv c$ for some constant c. Then

$$R^0(t;\mathbf{x}) \equiv c\alpha,$$

$$R(t;\mathbf{x}) = R^1(t;\mathbf{x}) = R^2(t;\mathbf{x}) = \cdots \equiv c(\alpha+n).$$

(iii) K is a projection of rank ν on $\mathcal{X}$. That is,

$$\begin{aligned} \int_{\mathcal{X}} K(x,x)\mu(dx) &= \nu, \\ \int_{\mathcal{X}} K(x,t)K(t,y)\mu(dt) &= K(x,y). \end{aligned}$$

Then the two-cycle approximation determines a probability density in the sense that it is non-negative and integrates to one:

$$\begin{aligned} (n+\alpha\nu)^{-1}\int_{\mathcal{X}} R_n^1(t;x)\,\mu(dt) &= (n+\alpha\nu)^{-1}\Big(\alpha\nu + \sum_i \int \frac{|K(t,x_i)|^2}{K(x_i,x_i)}\,\mu(dt)\Big) \\ &= (n+\alpha\nu)^{-1}(\alpha\nu + \sum_i \frac{K(x_i,x_i)}{K(x_i,x_i)}) \\ &= 1. \end{aligned}$$

A similar argument shows that the three-cycle and four-cycle approximations also integrate to one, but it is not clear whether they are non-negative.

7.2 Accuracy of the Cyclic Approximations

For $n < 20$, the accuracy of the approximation can be checked directly by comparison with the exact computation. Our experience is that the 3-cycle approximation is adequate in this range, and the four-cycle approximation usually has negligible error. For larger values, say $n > 50$, the accuracy can be checked by examining special cases in which the permanent can be calculated exactly in reasonable time. Examples include tri-diagonal and similar banded matrices, and the constant matrix, and in each of these cases the approximation is essentially exact. For more general matrices, the accuracy can be gauged to some extent from an examination of the sequence of approximations.

Figure 7.1: Approximations of $\mathrm{per}_\alpha[K](t, \mathbf{x})/\mathrm{per}_\alpha[K](\mathbf{x})$

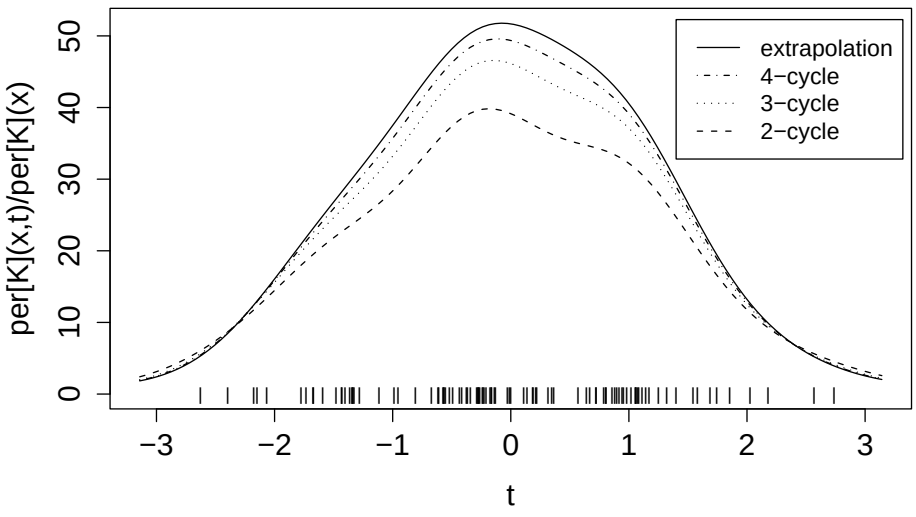


Figure 7.1 shows the approximated values of (7.1) for a sample of 100 x-values in

$(-\pi, \pi)$, plotted as a function of t in the same range. For this example, $\alpha = 1$, and $K = \exp(-(x - x')^2/\tau^2)$ with $\tau = 1$. In the central peak, the lowest curve is the first-order approximation, and the next three curves are successive approximations up to order 4. The highest curve is an extrapolation based on the third and fourth approximations. The shape of these relative intensity functions depends fairly strongly on the value of τ, but only slightly on α. In all cases, the difference between the second and third-order approximations is considerably smaller than the difference between the first and second. For $\alpha = \tau = 1$, the third-order approximation is approximately 6% larger than the second in the central peak, while the second-order approximation is approximately 18% larger than the first. The points for this example were generated from the symmetric triangular distribution on $(-\pi, \pi)$.

CHAPTER 8

SIMULATION STUDY

In this chapter, two examples are simulated to illustrate how the permanent classification model works. Only 2-3 parameters need to be estimated for those cases. The model performs well even if the classes occupy non-convex regions or disconnected regions in the feature space.

8.1 Chequerboard Pattern

The first artificial example has two classes in a 3×3 chequer-board layout labelling as follows

1	2	1
2	1	2
1	2	1

We assume the two-class permanent model (6.4) with $\alpha_1 = \alpha_2$, $K(x, x') = \exp(-\|x - x'\|/\tau^2)$. The training data consists of 90 units, with 10 feature values uniformly distributed in each small square as shown in Figure 8.1(a). We choose parameters (α, τ) to maximize

$$\sum_{i=1}^{n} \log p(Y_i = y_i | \mathbf{x}, \mathbf{y}_{-i}; \alpha, \tau),$$

where $\mathbf{y}_{-i}$ is the class list with unit i removed. The permanent ratios are calculated based on the four-cycle approximation.

Figure 8.1: Chequerboard Pattern – Part I

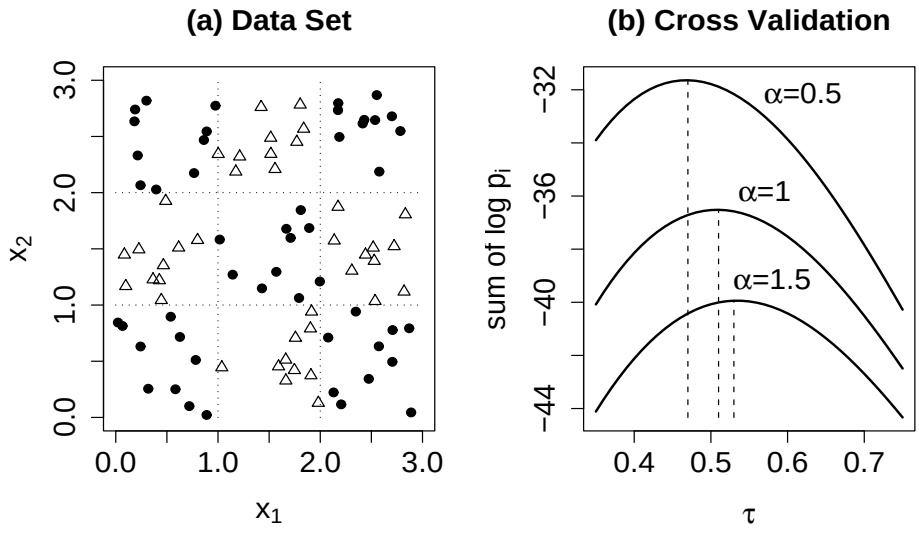
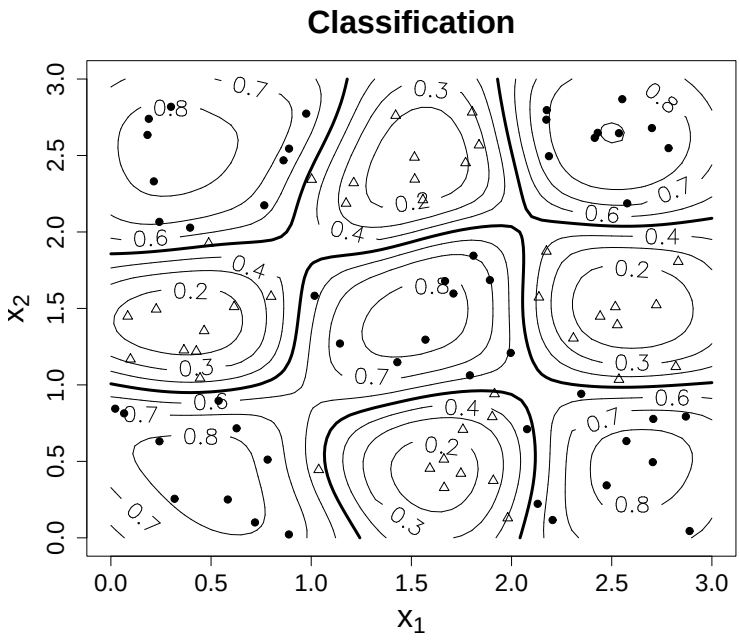


Figure 8.1(b) suggests $\hat{\alpha} = 0.5$ and $\hat{\tau} = 0.47$. Nevertheless, there is little difference if we choose $\alpha = 1$ or $\alpha = 1.5$ with corresponding τ. In other words, the classification inference is not sensitive to α.

Figure 8.2 provides the contour plot based on the probability that a new point is assigned to class 1. For the parameter values chosen, the range of predictive probabilities depends to a moderate extent on the configuration of x-values in the training sample, but the extremes are seldom below 0.1 or above 0.9 for a configuration of 90 points with 10 in each small square. The range of predictive probabilities decreases as τ increases, but the 50% contour line (the solid line in Figure 8.2) is little affected, so the classification is fairly stable. Given that the correct classification is

Figure 8.2: Chequerboard Pattern – Part II



determined by the chequerboard rule, the error rate for the permanent model using this particular training configuration is about 9%. This error rate is a little misleading because most of those errors occur near an internal boundary where the predictive probability is close to 0.5 .

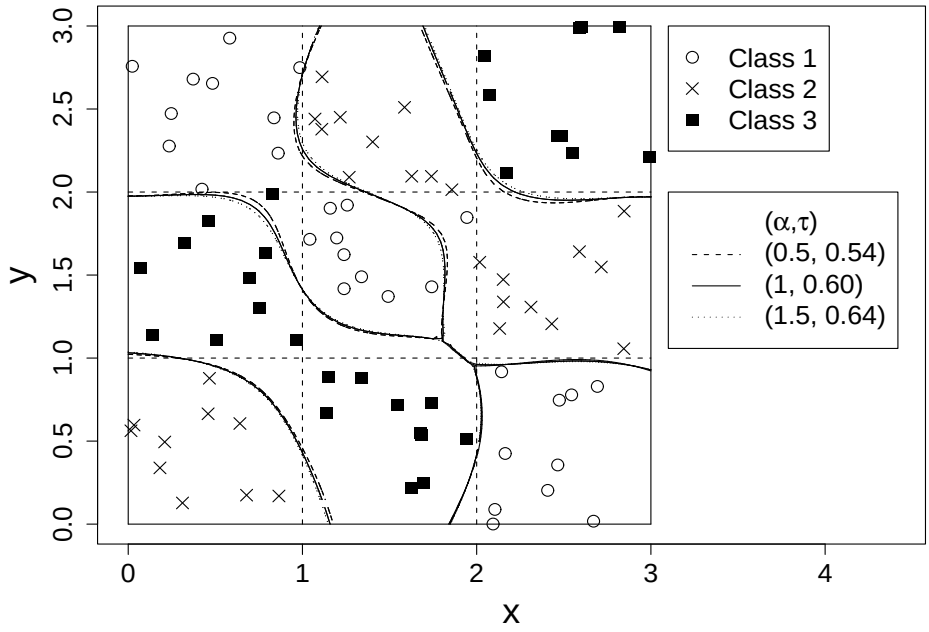
8.2 Latin Square Pattern

The second example has three classes in a 3×3 Latin-squre layout labelling as follows

1	2	3
3	1	2
2	3	1

As in Section 8.1, the training data consists of 90 units, with 10 feature values uniformly distributed in each small square as shown in Figure 8.3 . We assume the

Figure 8.3: Latin Square Pattern



three-class permanent model (6.4) with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$ and $K(x, x') = \exp(-\|x - x'\|/\tau^2)$. Different combinations of parameters (α, τ) are chosen to illustrate how the classification inference varies with α. The permanent ratios are calculated based on the four-cycle approximation. As we can see from Figure 8.3, the boundaries separating different class regions are fairly stable.

REFERENCES

[1] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*, New York: Dover.

[2] Aldous, D. J. (1985). Exchangeability and Related Topics. In *École d'Été de Probabilités de Saint-Flour XIII*, New York: Springer.

[3] Aldous, D. (1996). Probability Distributions on Cladograms. In *Random Discrete Structures*, eds. D. Aldous and R. Premantle, New York: Springer-Verlag, 1-18.

[4] Andrews, G. E. (1976). *The Theory of Partitions*. Reading, MA: Addison-Wesley.

[5] Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparamatric Problems. *The Annals of Statistics*, 2, 1152-1174.

[6] Bernstein, D. S. (2005). *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*, Princeton University Press.

[7] Berry, D. A. (1988). Multiple Comparisons, Multiple Tests, and Data Dredging: A Bayesian Perspective. In *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford University Press, 79-94.

[8] Bezáková, I., Štefankovič, D., Vazirani, V. V. and Vigoda, E. (2006). Accelerating Simulated Annealing Algorithm for the Permanent and Combinatorial Counting Problems. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA'06)*, 900-907.

[9] Billingsley, P. (1995). *Probability and Measure*, 3rd edition, New York: John Wiley & Sons.

[10] Blei, D., Griffiths, T., Jordan, M. and Tenenbaum, J. (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems 16*, Cambridge, MA: MIT Press.

[11] Box, G. E. P. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*, New York: John Wiley & Sons.

[12] Brooks, S. P. (1998). Markov Chain Monte Carlo Method and Its Application. *The Statistician*, 47, 69-100.

[13] Burris, S. and Sankappanavar, H. P. (1982). *A Course in Universal Algebra*, New York: Springer-Verlag.

[14] Cappe, O. and Robert, C. P. (2000). Markov Chain Monte Carlo: 10 Years and Still Running! *Journal of the American Statistical Association*, 95, 1282-1286.

[15] Ching, W. K. and Ng, M. K. (2006). *Markov Chains: Models, Algorithms and Applications*, New York: Springer Science + Business Media.

[16] Costantini, D. (1987). Symmetry and the Indistinguishability of Classical Particles. *Physics Letters A*, 123, 433-436.

[17] Daley D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd edition, New York: Springer.

[18] Davies, O. L. (ed.) (1967). *Statistical Methods in Research and Production*, 3rd edition, London: Oliver and Boyd.

[19] Ewens, W. J. (1972). The Sampling Theory of Selectively Neutral Alleles. *Theoretical Population Biology*, 3, 87-112.

[20] Ewens, W. J. and Tavaré, S. (1995). The Ewens Sampling Formula. In *Multivariate Discrete Distributions*, eds. N. S. Johnson, S. Kotz and N. Balakrishnan, New York: Wiley.

[21] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1, 209-230.

[22] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179-188.

[23] Ghanem, R. G. and Spanos, P. D. (1991). *Stochastic Finite Elements: A Spectral Approach*, revised edition, New York: Springer-Verlag.

[24] Gopalan, R. and Berry, D. A. (1998). Bayesian Multiple Comparisons Using Dirichlet Process Priors. *Journal of the American Statistical Association*, 93, 1130-1139.

[25] Gordon, A. D. (1999). *Classification*, 2nd edition, New York: Chapman & Hall/CRC.

[26] Hardy, G. H. and Ramanujan, S. (1918). Asymptotic Formulae in Combinatory Analysis. *Proceedings of the London Mathematical Society*, 17, 75-115.

[27] Hartigan, J. A. (1990). Partition Models. *Communications in Statistics, Part A - Theory and Methods*, 19, 2745-2756.

[28] Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*, New York: J. Wiley & Sons, 132-157.

[29] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: Wiley-Interscience.

[30] Kingman, J. F. C. (1975). Random Discrete Distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 37, 1-22.

[31] Kingman, J. F. C. (1978). Random Partitions in Population Genetics. *Proceedings of the Royal Society of London: Series A (Mathematical and Physical Sciences)*, 361, 1-20.

[32] Kingman, J. F. C. (1993). *Poisson Processes*, New York: Oxford University Press, 90-99.

[33] Lovász, L. (1979). *Combinatorial Problems and Exercises*, Amsterdam: North-Holland, 144-145.

[34] Miller, R. (1985). Multiple Comparisons. In *Encyclopedia of Statistical Sciences* (Vol. 5), eds. S. Kotz and N. L. Johnson, New York: John Wiley & Sons, 679-689.

[35] McCullagh, P. (1987). *Tensor Methods in Statistics*, London: Chapman and Hall.

[36] McCullagh, P. and Møller, J. (2005). The Permanent Process. Available via http://www.stat.uchicago.edu/∼pmcc/permanent.pdf .

[37] McCullagh, P. and Wilks, A. R. (1988). Complementary Set Partitions. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 415, 347-362.

[38] McCullagh, P. and Yang, J. (2006). Stochastic Classification Models. To appear in the *Proceedings of the International Congress of Mathematicians 2006*.

[39] McCullagh, P. and Yang, J. (2006). How Many Clusters? Available via http://www.stat.uchicago.edu/∼pmcc/reports/clusters.pdf .

[40] Minc, H. (1978). *Permanents*, London: Addison-Wesley.

[41] Neath, A. A. and Cavanaugh, J. E. (2006). A Bayesian Approach to the Multile Comparisons Problem. *Journal of Data Science*, 4, 131-146.

[42] Pal, S. K. and Pal, A. (eds.) (2001). *Pattern Recognition: From Classical to Modern Approaches*, New Jersey: World Scientific.

[43] Pitman, J. (1995). Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, 102, 145-158.

[44] Pitman, J. (1992). The Two-Parameter Generalization of Ewens' Random Partition Structure. Technical Report No. 345, Department of Statistics, U.C. Berkeley.

[45] Pitman, J. (1996). Random Discrete Distributions Invariant under Size-Biased Permutation. *Advances in Applied Probability*, 28, 525-539.

[46] Pitman, J. (2005). *Combinatorial Stochastic Processes*, New York: Springer.

[47] Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society, Series B*, 10, 159-203.

[48] Ripley, B. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

[49] Shirai, T. and Takahashi, Y. (2003). Random Point Fields Associated with Certain Fredholm Determinants I: Fermion, Poisson and Boson Point Processes. *Journal of Functional Analysis*, 205, 414-463.

[50] Skiena, S. (1990). *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Reading, MA: Addison-Wesley.

[51] Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th edition, Ames: Iowa State University Press.

[52] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22, 1701-1728.

[53] Tukey, J. W. (1953). The Problem of Multiple Comparisons. In *The Collected Works of John W. Tukey* (Vol. VIII), ed. H. Braun, New York: Chapman and Hall, 1-300, 1994.

[54] Tunnicliffe Wilson, G. (1989). On the Use of Marginal Likelihood in Time Series Model Estimation. *Journal of the Royal Statistical Society, Series B*, 51, 15-27.

[55] Tweedie, R. L. (2001). Markov Chains: Structure and Applications. In *Handbook of Statistics* (Vol. 19), eds. D. N. Shanbhag and C. R. Rao, Neitherlands: Elsevier Science, 817-851.

[56] Vere-Jones, D. (1988). A Generalization of Permanents and Determinants. *Linear Algebra and Its Applications*, 111, 119-124.

[57] Weisstein, E. W. Bell Number. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/BellNumber.html .

[58] Weisstein, E. W. Bell Polynomial. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/BellPolynomial.html .

[59] Weisstein, E. W. Dobinski's Formula. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/DobinskisFormula.html .

[60] Weisstein, E. W. Partition. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/Partition.html .

[61] Weisstein, E. W. Partition Function P. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/PartitionFunctionP.html .

[62] Weisstein, E. W. Permutation. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/Permutation.html .

[63] Weisstein, E. W. Set Partition. From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/SetPartition.html .