MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Markov basis and Gröbner basis of Segre-Veronese configuration for testing independence in group-wise selections

Satoshi AOKI, Takayuki HIBI, Hidefumi OHSUGI and
Akimichi TAKEMURA

METR 2007–21 April 2007

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Markov basis and Gröbner basis of Segre-Veronese configuration for testing independence in group-wise selections

Satoshi Aoki
Department of Mathematics and Computer Science
Kagoshima University

Takayuki Hibi
Graduate School of Information Science and Technology
Osaka University

Hidefumi Ohsugi
Department of Mathematics
Rikkyo University
and
Akimichi Takemura
Graduate School of Information Science and Technology
University of Tokyo

April, 2007

Keywords: contingency table, diplotype, exact tests, haplotype, Hardy-Weinberg model, Markov chain Monte Carlo, National Center Test, structural zero

Abstract

We consider testing independence in group-wise selections with some restrictions on combinations of choices. We present models for frequency data of selections for which it is easy to perform conditional tests by Markov chain Monte Carlo (MCMC) methods. When the restrictions on the combinations can be described in terms of a Segre-Veronese configuration, an explicit form of a Gröbner basis consisting of moves of degree two is readily available for performing a Markov chain. We illustrate our setting with the National Center Test for university entrance examinations in Japan. We also apply our method to testing independence hypotheses involving genotypes at more than one locus or haplotypes of alleles on the same chromosome.

1 Introduction

Suppose that people are asked to select items which are classified into categories or groups and there are some restrictions on combinations of choices. For example, when a consumer buys a car, he or she can choose various options, such as a color, a grade of air conditioning, a brand of audio equipment, etc. Due to space restrictions for example, some combinations of options may not be available. The problem we consider in this paper is testing independence of people's preferences in group-wise selections in the presence of restrictions. We assume that observations are the counts of people choosing various combinations in group-wise selections, i.e., the data are given in a form of a multiway contingency table with some structural zeros corresponding to the restrictions.

If there are m groups of items and a consumer freely chooses just one item from each group, then the combination of choices is simply a cell of an m-way contingency table. Then the hypothesis of independence reduces to the complete independence model of an m-way contingency table. The problem becomes harder if there are some additional conditions in a group-wise selection. A consumer may be asked to choose up to two items from a group or there may be a restriction on the total number of items. Groups may be nested, so that there are further restrictions on the number of items from subgroups. Some restrictions may concern several groups or subgroups. Therefore the restrictions on combinations may be complicated.

As a concrete example we consider restrictions on choosing subjects in the *National Center Test* (NCT hereafter) for university entrance examinations in Japan (Section 2). Due to time constraints of the schedule of the test, the pattern of restrictions is rather complicated. However we will show that restrictions of NCT can be described in terms of a Segre-Veronese configuration.

Another important application of this paper is a generalization of the Hardy-Weinberg model in population genetics. We are interested in testing various hypotheses of independence involving genotypes at more than one locus and haplotypes of combination of alleles on the same chromosome. Although this problem seems to be different from the above introductory motivation on consumer choices, we can imagine that each offspring is required to choose two alleles for each gene (locus) from a pool of alleles for the gene. He or she can choose the same allele twice (homozygote) or different alleles (heterozygote). In the Hardy-Weinberg model two choices are assumed to be independently and identically distributed. A natural generalization of the Hardy-Weinberg model for a single locus is to consider independence of genotypes of more than one locus. In many epidemiological studies, the primary interest is the correlation between a certain disease and the genotype of a single gene (or the genotypes at more than one locus, or the haplotypes involving alleles on the same chromosome). Further complication might arise if certain homozygotes are fatal and can not be observed, thus becoming a structural zero.

In this paper we consider conditional tests of independence hypotheses in the above two important problems from the viewpoint of Markov bases and Gröbner bases. Evaluation of P-values by MCMC using Markov bases and Gröbner bases was initiated by Diaconis and Sturmfels [8]. See also [20]. Since then, this approach attracted much attention from

statisticians as well as algebraists. Contributions of the present authors are found, for example, in [1], [3], [15], [16], [17] and [21]. Methods of algebraic statistics are currently actively applied to problems in computational biology [18]. In algebraic statistics, results in commutative algebra may find somewhat unexpected applications in statistics. At the same time statistical problems may present new problems to commutative algebra. A recent example is a conjunctive Bayesian network proposed in [4], where a result of Hibi [10] is successfully used. In this paper we present application of results on Segre-Veronese configuration to testing independence in NCT and Hardy-Weinberg models. In fact, these statistical considerations have prompted further theoretical developments of Gröbner bases for Segre-Veronese type configurations and we will present these theoretical results in our subsequent paper.

Even in two-way tables, if the positions of the structural zeros are arbitrary, then Markov bases may contain moves of high degrees ([1]). However if the restrictions on the combinations can be described in terms of a Segre-Veronese configuration, then an explicit form of a Gröbner basis consisting of moves of degree two with a squarefree initial term is readily available for running a Markov chain for performing conditional tests of various hypotheses of independence. Therefore models which can be described by a Segre-Veronese configuration are very useful for statistical analysis.

The organization of this paper is as follows. In Section 2 we take a close look at patterns of selections of subjects in NCT and in Section 3 we consider various hypotheses of independence for NCT data and their conditional tests. In Section 4 we study generalizations of the Hardy-Weinberg model. In Section 5 we give a brief review of MCMC approach to conditional tests based on Markov basis and in Section 6 we define Segre-Veronese configuration. We give an explicit expression of a reduced Gröbner basis for the configuration and describe a simple procedure for running MCMC using the basis for conditional tests. We end the paper by some discussions in Section 8.

2 The case of National Center Test in Japan

One important example of group-wise selection is the entrance examination for universities in Japan. In Japan, as the common first-stage screening process, most students applying for universities take the National Center Test for university entrance examinations administered by National Center for University Entrance Examinations (NCUEE). Basic information in English on NCT in 2006 is available from the booklet published by NCUEE ([12] in the references). After obtaining the score of NCT, students apply to departments of individual universities and take second-stage examinations administered by the universities. Due to time constraints of the schedule of NCT, there are rather complicated restrictions on possible combination of subjects. Furthermore each department of each university can impose different additional requirement on the combinations of subjects of NCT to students applying to the department.

In NCT examinees can choose subjects in Mathematics, Social Studies and Science. These three major subjects are divided into subcategories. For example Mathematics is

divided into Mathematics 1 and Mathematics 2 and these are then composed of individual subjects. In the test carried out in 2006, examinees could select two mathematics subjects, two social studies subjects and three science subjects at most as shown below. The details of the subjects can be found in web pages and publications of NCUEE. In parentheses we show our abbreviations for the subjects in this paper.

- Mathematics:
  - Mathematics 1: One subject from {MathI, MathIA}
  - Mathematics 2: One subject from {MathII, MathIIB, Basics in Mathematics and Science for Industry (BMSI), Bookkeeping and Accounting (BKA), Basics in Information Processing (Info)}
- Social Studies:
  - Geography and History: One subject from {World History A (WHA), World History B (WHB), Japanese History A (JHA), Japanese History B (JHB), Geography A (GeoA), Geography B (GeoB)}
  - Civics: One subject from {Contemporary Society (ContSoc), Ethics, Politics and Economics (P&E)}
- Science:
  - Science 1: One subject from {Comprehensive Science B (CSciB), Biology I (BioI), Integrated Science (IntegS), Biology IA (BioIA)}
  - Science 2: One subject from {Comprehensive Science A (CSciA), Chemistry I (ChemI), Chemistry IA (ChemIA)}
  - Science 3: One subject from {Physics I (PhysI), Earth Science I (EarthI), Physics IA (PhysIA), Earth Science IA (EarthIA)}

Frequencies of the examinees selecting each combination of subjects in 2006 are given in the website of NCUEE. We reproduce part of them in Tables 8–14 at the end of the paper. As seen in these tables, examinees may select or not select these subjects. For example, one examinee may select two subjects from Mathematics, two subjects from Social Studies and three subjects from Science, while another examinee may select only one subject from Mathematics, one subject from Science and none from Social Studies. Hence each examinee is categorized into one of the $(2+1)\times\cdots\times(4+1) = 50400$ combinations of individual subjects. Here 1 is added for not choosing from the subcategory. As mentioned above, individual departments of universities impose different additional requirements on the choices of subjects of NCT. For example, many science or engineering departments of national universities ask the students to take two subjects from Science and one subject from Social Studies.

Let us observe some tendencies of the selections by the examinees to illustrate what kind of statistical questions one might ask concerning the data in Tables 8–14.

(i) The most frequent triple of Science subjects is {BioI, ChemI, PhysI} in Table 14, which seems to be consistent with Table 12 since these three subjects are the most

frequently selected subjects in Science 1, Science 2 and Science 3, respectively. However in Table 13, while the pairs {BioI, ChemI} and {ChemI, PhysI} are the most frequently selected pairs in {Science 1, Science2} and {Science 2, Science 3}, respectively, the pair {BioI, PhysI} is not the first choice in {Science 1, Science 3}. This fact indicates differences in the selection of Science subjects between the examinees selecting two subjects and those selecting three subjects.

(ii) In Table 11 the most frequent pair is {GeoB, ContSoc}. However the most frequent single subject from Geography and History is JHB both in Table 10 and 11. This result indicates the interaction effect in selecting pairs of Social Studies.

These observations lead to many interesting statistical questions. However Tables 8–14 only give frequencies of choices separately for Mathematics, Social Studies and Science, i.e., they are the marginal tables for these three major subjects. In this paper we are interested in independence across these three major subjects, such as "are the selections on Social Studies and Science related or not?" Unfortunately NCUEE currently do not provide cross tabulations of frequencies of choices across the major subjects. Although appropriate data are not available at present, in the next section we consider hypotheses of independence across the major subjects, since the conditional null distribution can be evaluated from the marginal tables.

3 Formulations of independence hypotheses for NCT and their conditional tests

In this section we formulate data types and their statistical models in view of NCT. Suppose that there are J different groups (or categories) and m_j different subgroups in group j for $j = 1, \ldots, J$. There are m_{jk} different *items* in subgroup k of group j ($k = 1, \ldots, m_j$, $j = 1, \ldots, J$). In NCT $J = 3$ and $m_1 = |\{\text{Mathematics 1, Mathematics 2}\}| = 2$ and similarly $m_2 = 2, m_3 = 3$. The sizes of subgroups are $m_{11} = |\{\text{MathI, MathIA}\}| = 2$ and similarly $m_{12} = 5$, $m_{21} = 6$, $m_{22} = 3$, $m_{31} = 4$, $m_{32} = 3$, $m_{33} = 4$.

Each individual selects c_{jk} items from the subgroup j of group k. We assume that the total number τ of items chosen is fixed and common for all individuals. In NCT c_{jk} is either 0 or 1. For example if an examinee is required to take two Science subjects in NCT, then (c_{31}, c_{32}, c_{33}) is $(1, 1, 0)$, $(1, 0, 1)$ or $(0, 1, 1)$. For the analysis of genotypes in Section 4, $c_{jk} \equiv 2$ although there is no nesting of subgroups, and the same item (allele) can be selected more than once (selection "with replacement").

We now set up our notation for indexing a combination of choices somewhat carefully. In NCT, if an examinee chooses WHA from "Geography and History" of Social Studies and PhysI from Science 3 of Science, we denote the combination of these two choices as (211)(331). In this notation, the selection of c_{jk} items from the subgroup k of group j are indexed as

$$\mathbf{i}_{jk} = (jkl_1)(jkl_2)\ldots(jkl_{c_{jk}}), \qquad 1 \le l_1 \le \cdots \le l_{c_{jk}} \le m_{jk}.$$

Here $\mathbf{i}_{jk}$ is regarded as a string. If nothing is selected from the subgroup, we define $\mathbf{i}_{jk}$ to be an empty string. Now by concatenation of strings, the set $\mathcal{I}$ of combinations is written as

$$\mathcal{I} = \{\mathbf{i} = \mathbf{i}_1 \dots \mathbf{i}_J\}, \qquad \mathbf{i}_j = \mathbf{i}_{j1} \dots \mathbf{i}_{jm_j}, \quad j = 1, \dots, J.$$

For example the choice of (MathIA, MathIIB, P&E, BioI, ChemI) in NCT is denoted by $\mathbf{i} = (112)(122)(223)(312)(322)$. Following the terminology of contingency tables, each $\mathbf{i} \in \mathcal{I}$ is called as a *cell*. We denote the number of possible combinations by $\nu = |\mathcal{I}|$.

We write $\mathbf{j} \subset \mathbf{i}$ to denote that a string $\mathbf{j}$ appears as a substring of $\mathbf{i}$. For $1 \le l \le m_{jk}$, we denote the number of times l appears in $\{l_1, \dots, l_{c_{jk}}\}$ by

$$\#(l; \mathbf{i}_{jk}) = \sum_{t=1}^{c_{jk}} 1_{\{l_t = l\}}.$$

If nothing is selected from the subgroup, $\#(l; \mathbf{i}_{jk}) = 0$, $1 \le l \le m_{jk}$. For NCT $\#(l; \mathbf{i}_{jk})$ is either 0 or 1.

Let $p(\mathbf{i})$ denote the probability of selecting the combination $\mathbf{i}$ (or the probability of cell $\mathbf{i}$) and write $\mathbf{p} = \{p(\mathbf{i})\}_{\mathbf{i} \in \mathcal{I}}$. When there are structural zeros or some other restrictions on the possible cells, it often becomes difficult to determine the normalizing constant $c = \sum_{\mathbf{i} \in \mathcal{I}} p(\mathbf{i})$ for a given unnormalized functional specification of $p(\cdot)$. Denote the result of the selections by n individuals as $\mathbf{x} = \{x(\mathbf{i})\}_{\mathbf{i} \in \mathcal{I}}$, where $x(\mathbf{i})$ is the frequency of the cell $\mathbf{i}$. We call $\mathbf{x}$ a frequency vector. Under the usual multinomial model, the joint probability of frequencies is given by

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}),\ \mathbf{i} \in \mathcal{I}) = \frac{n!}{\prod_{\mathbf{i} \in \mathcal{I}} x(\mathbf{i})!} \prod_{\mathbf{i} \in \mathcal{I}} p(\mathbf{i})^{x(\mathbf{i})}. \tag{1}$$

Now we consider some statistical models for $\mathbf{p}$. For NCT data, we consider three simple statistical models, namely, *complete independence model*, *subgroup-wise independence model* and *group-wise independence model*. The complete independence model is defined as

$$p(\mathbf{i}) = \prod_{j=1}^{J} \prod_{\substack{k=1 \\ \mathbf{i}_{jk} \subset \mathbf{i}}}^{m_j} \prod_{t=1}^{c_{jk}} q_{jk}(l_t) \tag{2}$$

for some parameters $q_{jk}(l),\ j = 1, \dots, J;\ k = 1, \dots, m_j;\ l = 1, \dots, m_{jk}$. Note that if $c_{jk} > 1$ we need a multinomial coefficient in (2) as in (10) below. The complete independence model means that each $p(\mathbf{i})$, the inclination of the combination $\mathbf{i}$, is explained by the set of inclinations $q_{jk}(l)$ of each item. Here $q_{jk}(l)$ corresponds to the marginal probability of the item (jkl). However we do not necessarily normalize them as $1 = \sum_{l=1}^{c_{jk}} q_{jk}(l)$, because the normalization for $\mathbf{p}$ is not trivial anyway. The same comment applies to other models below.

The subgroup-wise independence model is defined as

$$p(\mathbf{i}) = \prod_{j=1}^{J} \prod_{\substack{k=1 \\ \mathbf{i}_{jk} \subset \mathbf{i}}}^{m_j} q_{jk}(\mathbf{i}_{jk}) \tag{3}$$

for some parameters $q_{jk}(\cdot)$. This model means that each $p(\mathbf{i})$ is explained by the set of $q_{jk}(\cdot)$, the inclinations of each combination of items for each subgroup and there is no further structure in the specification of $q_{jk}(\cdot)$. Finally, the group-wise independence model is defined as

$$p(\mathbf{i}) = \prod_{j=1}^{J} q_j(\mathbf{i}_j) \tag{4}$$

for some parameters $q_j(\cdot)$. In this paper, we treat these models as the *null models* and give testing procedures to assess their fitting to observed data. Since the arguments for three models are almost similar, we first give a description for the complete independence model. We then consider the subgroup-wise and the group-wise independence model briefly.

Under the complete independence model (2), the joint probability function (1) is written as

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}),\ \mathbf{i} \in \mathcal{I}\ ) = \frac{n!}{\prod_{\mathbf{i}\in\mathcal{I}} x(\mathbf{i})!} \prod_{j=1}^{J} \prod_{k=1}^{m_j} \prod_{l=1}^{m_{jk}} q_{jk}(l)^{t_{jk}(l)}, \tag{5}$$

where

$$t_{jk}(l) = \sum_{\mathbf{i}\in\mathcal{I}} x(\mathbf{i}) \#(l; \mathbf{i}_{jk}) \tag{6}$$

is the frequency that the item (jkl) is selected. We see that $\mathbf{t} = \{t_{jk}(l)\}$ is the sufficient statistic for the parameter under the complete independence model. Note that (6) can be written in a matrix form. Regard $\mathbf{x} = \{x(\mathbf{i})\}_{\mathbf{i}\in\mathcal{I}}$ as a column vector with dimension $\nu = |\mathcal{I}|$ and regard $\mathbf{t}$ as a column vector with dimension $d = \sum_{j=1}^{J} \sum_{k=1}^{m_j} m_{jk}$. Then the relation (6) is written as

$$\mathbf{t} = A\mathbf{x}, \tag{7}$$

where A is a $d \times \nu$ matrix with $\#(l; \mathbf{i}_{jk})$ as the $((jkl), \mathbf{i})$ element. We call A a *configuration* in connection with the theory of toric ideals in Section 6. Once the sufficient statistic $\mathbf{t}$ is written in the form (7), the theory of Markov basis can be used to perform MCMC for conditional test of the complete independence model.

As for the group-wise and category-wise independence model, the sufficient statistics are simpler. Under the subgroup-wise independence model (3) the joint probability is written as

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}),\ \mathbf{i} \in \mathcal{I}) = \frac{n!}{\prod_{\mathbf{i}\in\mathcal{I}} x(\mathbf{i})!} \prod_{j=1}^{J} \prod_{k=1}^{m_j} q_{jk}(\mathbf{i}_{jk})^{t_{jk}(\mathbf{i}_{jk})},$$

where $t_{jk}(\mathbf{i}_{jk}) = \sum_{\mathbf{i}\in\mathcal{I}, \mathbf{i}_{jk}\subset\mathbf{i}} x(\mathbf{i})$ is the frequency of the combination of items $\mathbf{i}_{jk}$. Let $\mathbf{t} = \{t_{jk}(\mathbf{i}_{jk})\}$ be the column vector of the sufficient statistic. When $c_{jk} \equiv 1$ its dimension d is given as $d = \sum_{j=1}^{J} \prod_{k=1}^{m_j} m_{jk}$. Then as in (7) we can write $\mathbf{t} = A\mathbf{x}$. Here A is a $d \times \nu$ matrix of 0's and 1's such that $(\mathbf{i}_{jk}, \mathbf{i})$ element of A equals 1 if and only if $\mathbf{i}_{jk} \subset \mathbf{i}$. Similarly under the group-wise independence model (4) the joint probability is written as

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}),\ \mathbf{i} \in \mathcal{I}) = \frac{n!}{\prod_{\mathbf{i}\in\mathcal{I}} x(\mathbf{i})!} \prod_{j=1}^{J} q_j(\mathbf{i}_j)^{t_j(\mathbf{i}_j)},$$

where $t_j(\mathbf{i}_j) = \sum_{\mathbf{i}\in\mathcal{I},\mathbf{i}_j\subset\mathbf{i}} x(\mathbf{i})$ is the frequency of the combination of items $\mathbf{i}_j$. Again we can write the relation between the cell frequencies $\mathbf{x}$ and the sufficient statistic $\mathbf{t}$ as $\mathbf{t} = A\mathbf{x}$ with an appropriate A.

4 Hardy-Weinberg models

In this section, we consider problems of population genetics from the viewpoint of group-wise selections. Because of practical importance of this application, we present our framework adapted to this application from scratch without reference to the previous section. In addition the availability of haplotype data or diplotype data requires a separate treatment.

The allele frequency data are usually given as the genotype frequency. For multi-allele locus with alleles $A_1, A_2, \ldots, A_m$, the probability of the genotype A_iA_j in an individual from a random breeding population is q_i^2 $(i = j)$ or $2q_iq_j$ $(i \neq j)$, where q_i is the proportion of the allele A_i. These are known as the Hardy-Weinberg equilibrium probabilities. Since the Hardy-Weinberg law plays an important role in the field of population genetics and often serves as a basis for genetic inference, much attention has been paid to tests of the hypothesis that a population being sampled is in the Hardy-Weinberg equilibrium against the hypothesis that disturbing forces cause some deviation from the Hardy-Weinberg ratio. See [6] for example. [21] considers conditional tests of Hardy-Weinberg model by using Markov basis technique.

Due to the rapid progress of sequencing technology, more and more information is available on the combination of alleles on the same chromosome. A combination of alleles at more than one locus on the same chromosome is called a haplotype and data on haplotype counts are called haplotype frequency data. The haplotype analysis has gained an increasing attention in the mapping of complex-disease genes, because of the limited power of conventional single-locus analyses. Haplotype data may come with or without pairing information on homologous chromosomes. It is technically more difficult to determine pairs of haplotypes of the corresponding loci on a pair of homologous chromosomes. A pair of haplotypes on homologous chromosomes is called a diplotype. In this paper we are interested in diplotype frequency data, because haplotype frequency data on individual chromosomes without pairing information are standard contingency table data and can be analyzed by statistical methods for usual contingency tables. For the diplotype frequency data, the null model we want to consider is the independence model that the probability for each diplotype is expressed by the product of probabilities for each genotype.

In this section, first we consider the models for genotype frequency data in Section 4.1 and then consider the models for diplotype frequency data in Section 4.2.

4.1 Models for the genotype frequency data

We assume that there are J distinct loci. In the locus j, there are m_j distinct alleles, $A_{j1}, \ldots, A_{jm_j}$. In this case, we can imagine that each individual selects two alleles for each locus *with replacement*. Therefore the set of the combinations is written as

$$\mathcal{I} = \{\mathbf{i} = (i_{11}i_{12})(i_{21}i_{22})\ldots(i_{J1}i_{J2}) \mid 1 \le i_{j1} \le i_{j2} \le m_j,\ j = 1, \ldots, J\}.$$

Let $p(\mathbf{i})$ denote the probability of the combination $\mathbf{i}$ in the population and write $\mathbf{p} = \{p(\mathbf{i})\}_{\mathbf{i}\in\mathcal{I}}$. Write the genotype frequency by n individuals as $\mathbf{x} = \{x(\mathbf{i})\}_{\mathbf{i}\in\mathcal{I}}$. We consider conditional tests of various hypotheses of independence based on $\mathbf{x}$.

For the genotype frequency data, we consider two models of hierarchical structure, namely, *genotype-wise independence model*

$$p(\mathbf{i}) = \prod_{j=1}^{J} q_j(i_{j1}i_{j2}) \tag{8}$$

and the Hardy-Weinberg model

$$p(\mathbf{i}) = \prod_{j=1}^{J} \tilde{q}_j(i_{j1}i_{j2}), \tag{9}$$

where

$$\tilde{q}_j(i_{j1}i_{j2}) = \begin{cases} q_j(i_{j1})^2 & \text{if } i_{j1} = i_{j2}, \\ 2q_j(i_{j1})q_j(i_{j2}) & \text{if } i_{j1} \neq i_{j2}. \end{cases} \tag{10}$$

For the genotype-wise independence model the sufficient statistic is given by the set of frequencies $\mathbf{t} = \{t_j(i_{j1}i_{j2})\}$ of the genotypes, where

$$t_j(i_{j1}i_{j2}) = \sum_{\substack{\mathbf{i}'\in\mathcal{I} \\ (i'_{j1}i'_{j2})=(i_{j1}i_{j2})}} x(\mathbf{i}').$$

For the Hardy-Weinberg model the sufficient statistic is given by the set of frequencies $\mathbf{t} = \{t_j(i_j)\}$ of individual alleles, where

$$t_j(i_j) = 2\sum_{(i'_{j1}i'_{j2})=(i_ji_j)} x(\mathbf{i}') + \sum_{i_j=i'_{j1}<i'_{j2}} x(\mathbf{i}') + \sum_{i'_{j1}<i'_{j2}=i_j} x(\mathbf{i}').$$

Note that for both cases the sufficient statistic $\mathbf{t}$ can be written as $\mathbf{t} = A\mathbf{x}$ for appropriate matrix A as shown in Section 7.2. Given these sufficient statistics conditional tests of these models can be performed by Markov chain Monte Carlo methodology presented in Section 5.

4.2 Models for the diplotype frequency data

In order to illustrate the difference between genotype data and diplotype data, consider a simple case of $J = 2, m_1 = m_2 = 2$ and suppose that genotypes of $n = 4$ individuals are given as

$$\{A_{11}A_{11}, A_{21}A_{21}\},\ \{A_{11}A_{11}, A_{21}A_{22}\},\ \{A_{11}A_{12}, A_{21}A_{21}\},\ \{A_{11}A_{12}, A_{21}A_{22}\}.$$

In this genotype data, for an individual who has homozygote genotype on at least one loci, the diplotypes are uniquely determined. However, for the fourth individual who has the genotype $\{A_{11}A_{12}, A_{21}A_{22}\}$, there are two possible diplotypes as $\{(A_{11}, A_{21}), (A_{12}, A_{22})\}$ and $\{(A_{11}, A_{22}), (A_{12}, A_{21})\}$.

Now suppose that information on diplotypes are available. The set of combinations for the diplotype data is given as

$$\mathcal{I} = \{\mathbf{i} = \mathbf{i}_1\mathbf{i}_2 = (i_{11} \cdots i_{J1})(i_{12} \cdots i_{J2}) \mid 1 \le i_{j1}, i_{j2} \le m_j,\ j = 1, \ldots, J\}.$$

In order to determine the order of $\mathbf{i}_1 = (i_{11} \ldots i_{r1})$ and $\mathbf{i}_2 = (i_{12} \ldots i_{r2})$ uniquely, we assume that these two are lexicographically ordered, i.e., there exists some j such that

$$i_{11} = i_{12}, \ldots, i_{j-1,1} = i_{j-1,2},\ i_{j1} < i_{j2}$$

unless $\mathbf{i}_1 = \mathbf{i}_2$.

For the parameter $\mathbf{p} = \{p(\mathbf{i})\}$ where $p(\mathbf{i})$ is the probability for the diplotype $\mathbf{i}$, we can consider the same models as for the genotype case. Corresponding to the null hypothesis that diplotype data do not contain more information than the genotype data, we can consider the genotype-wise independence model (8) and the Hardy-Weinberg model (9). The sufficient statistics for these models are the same as in the previous subsection.

If these models are rejected, we can further test independence in diplotype data. For example we can consider a haplotype-wise Hardy-Weinberg model.

$$p(\mathbf{i}) = p(\mathbf{i}_1\mathbf{i}_2) = \begin{cases} q(\mathbf{i}_1)^2 & \text{if } \mathbf{i}_1 = \mathbf{i}_2, \\ 2q(\mathbf{i}_1)q(\mathbf{i}_2) & \text{if } \mathbf{i}_1 \neq \mathbf{i}_2. \end{cases}$$

The sufficient statistic for this model is given by the set of frequencies of each haplotype and the conditional test can be performed as in the case of Hardy-Weinberg model for a single gene by formally identifying each haplotype as an allele.

5 Markov chain Monte Carlo methods and Markov bases

In this section we give a brief review on performing MCMC for conducting conditional tests based on the theory of Markov basis. Markov basis was introduced by [8] and there are now many references on the use of Markov basis (e.g. [2]).

In the models considered in this paper the cell probability $p(\mathbf{i})$ is written as some product of functions, which correspond to various marginal probabilities. It should be noted that unlike the case of standard multiway contingency tables, our index set $\mathcal{I}$ can not be written as a direct product in general. Let $\mathcal{J}$ denote the index set of for the marginals. Then our models presented so far can be written as

$$p(\mathbf{i}) = h(\mathbf{i}) \prod_{\mathbf{j} \in \mathcal{J}} q(\mathbf{j})^{a_{\mathbf{ji}}}, \tag{11}$$

where $h(\mathbf{i})$ is a known function and $q(\mathbf{j})$'s are the parameters. Traditionally a model of the form (11) is called a log-linear model in statistics, but recently it is also called a toric model in algebraic statistics ([18, Chap.1]). Under the usual multinomial sampling, the joint probability of frequencies is written as

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}),\ \mathbf{i} \in \mathcal{I}) = \frac{n! \prod_{\mathbf{i} \in \mathcal{I}} h(\mathbf{i})^{x(\mathbf{i})}}{\prod_{\mathbf{i} \in \mathcal{I}} x(\mathbf{i})!} \prod_{\mathbf{j} \in \mathcal{J}} q(\mathbf{j})^{\sum_{\mathbf{i} \in \mathcal{I}} a_{\mathbf{ji}} x(\mathbf{i})}.$$

Therefore the sufficient statistic $\mathbf{t} = \{t(\mathbf{j}), \mathbf{j} \in \mathcal{J}\}$ is written in a matrix form as

$$\mathbf{t} = A\mathbf{x}, \qquad A = (a_{\mathbf{ji}})_{\mathbf{j} \in \mathcal{J}, \mathbf{i} \in \mathcal{I}},$$

where A is $d \times \nu$ matrix of non-negative integers and $d = |\mathcal{J}|$, $\nu = |\mathcal{I}|$.

By the standard theory of conditional tests (e.g. [11]), we can perform conditional test of the model (11) based on the conditional distribution given the sufficient statistic $\mathbf{t}$:

$$\Pr(X(\mathbf{i}) = x(\mathbf{i}), \mathbf{i} \in \mathcal{I} \mid \mathbf{t}) = c \prod_{\mathbf{i} \in \mathcal{I}} \frac{h(\mathbf{i})^{x(\mathbf{i})}}{x(\mathbf{i})!}, \quad c = \left(\sum_{\mathbf{x} \in \mathcal{F}_{\mathbf{t}}} \prod_{\mathbf{i}' \in \mathcal{I}} \frac{h(\mathbf{i}')^{x(\mathbf{i}')}}{x(\mathbf{i}')!} \right)^{-1}. \tag{12}$$

The conditional sample space given $\mathbf{t}$, called the $\mathbf{t}$-fiber, is

$$\mathcal{F}_{\mathbf{t}} = \{\mathbf{x} \in \mathbb{N}^{\nu} \mid \mathbf{t} = A\mathbf{x}\},$$

where $\mathbb{N} = \{0, 1, \dots\}$. If we can sample from the conditional distribution over $\mathcal{F}_{\mathbf{t}}$, we can evaluate P-values of any test statistic. One of the advantages of MCMC method of sampling is that it can be run without evaluating the normalizing constant c. Also once a connected Markov chain over the conditional sample space is constructed, then the chain can be modified to give a connected and aperiodic Markov chain with the stationary distribution by the Metropolis-Hastings procedure (e.g. [9]). Therefore it is essential to construct a connected chain and the solution to this problem is given by the notion of *Markov basis* ([8]).

Let $\mathbb{Z}$ denote the set of integers and $\mathcal{M}_A \subset \mathbb{Z}^{\nu}$ be the set of integer vectors in the kernel of A, i.e.,

$$\mathcal{M}_A = \{\mathbf{z} \mid A\mathbf{z} = \mathbf{0}\}.$$

We call an elements in $\mathcal{M}_A$ a *move* for A. Note that adding $\mathbf{z} \in \mathcal{M}_A$ to any frequency vector $\mathbf{x} \in \mathbb{N}^\nu$ does not change the sufficient statistics, i.e.,

$$A(\mathbf{x} + \mathbf{z}) = A\mathbf{x}. \tag{13}$$

$\mathbf{x} + \mathbf{z}$ above might contain a negative element. However, if $\mathbf{x} + \mathbf{z} \in \mathcal{F}_{A\mathbf{x}}$, we see that $\mathbf{x}$ is *moved* to $\mathbf{x} + \mathbf{z} \in \mathcal{F}_{A\mathbf{z}}$ by $\mathbf{z}$, which is why we call $\mathbf{z} \in \mathcal{M}^p$ a move. Now the definition of the Markov basis is as follows.

Definition 5.1. *A Markov basis for A is a set of moves $\mathcal{B} = \{\mathbf{z}_1, \ldots, \mathbf{z}_L\}, \mathbf{z}_j \in \mathcal{M}_A, j = 1, \ldots, L$, such that, for any $\mathbf{y}, \mathbf{y}^* \in \mathcal{F}_{\mathbf{t}^o}, \mathbf{t}^o = A\mathbf{y}^o$, there exist $S > 0$, $(\varepsilon_1, \mathbf{z}_{j_1}), \ldots, (\varepsilon_S, \mathbf{z}_{j_S})$ with $\varepsilon_s \in \{-1, +1\}$, $\mathbf{z}_{j_s} \in \mathcal{B}$, $s = 1, \ldots, S$, satisfying*

$$\mathbf{y} = \mathbf{y}^* + \sum_{s=1}^{S} \varepsilon_s \mathbf{z}_{j_s} \text{ and } \mathbf{y}^* + \sum_{s=1}^{r} \varepsilon_s \mathbf{z}_{j_s} \in \mathcal{F}_{\mathbf{t}^o} \text{ for } r = 1, \ldots, S.$$

By definition, a Markov basis enables us to construct a connected chain over the conditional sample space $\mathcal{F}_{\mathbf{t}}$ for any observed frequency vector $\mathbf{x}^o$. The fundamental contribution of [8] is to show that a Markov basis is a generator of the well-specified polynomial ideal (toric ideal) and it can be give as a Gröbner basis. In the next section, we show that our problem corresponds to a well-known toric ideal and give an explicit form of the reduced Gröbner basis.

6 Gröbner basis for Segre-Veronese configuration

In this section, we introduce toric ideals of algebras of Segre-Veronese type ([14]) with a generalization to fit statistical applications in the present paper.

First we define toric ideals. A *configuration* in $\mathbb{R}^d$ is a finite set $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_\nu\} \subset \mathbb{N}^d$. A can be regarded as a $d \times \nu$ matrix and corresponds to the matrix connecting the frequency vector to the sufficient statistic as in (7). Let K be a field and $K[\mathbf{q}] = K[q_1, \ldots, q_d]$ the polynomial ring in d variables over K. We associate a configuration $A \subset \mathbb{Z}^d$ with the semigroup ring $K[A] = K[\mathbf{q}^{\mathbf{a}_1}, \ldots, \mathbf{q}^{\mathbf{a}_\nu}]$ where $\mathbf{q}^{\mathbf{a}} = q_1^{a_1} \cdots q_d^{a_d}$ if $\mathbf{a} = (a_1, \ldots, a_d)$. Note that $d = |\mathcal{J}|$ and $\mathbf{q}^{\mathbf{a}_i}$ corresponds to to the term $\prod_{\mathbf{j} \in \mathcal{J}} q(\mathbf{j})^{a_{\mathbf{ji}}}$ on the right-hand side of (11). Let $K[Y] = K[y_1, \ldots, y_\nu]$ be the polynomial ring in ν variables over K. Here $\nu = |\mathcal{I}|$ and the variables $y_1, \ldots, y_\nu$ correspond to the cells of $\mathcal{I}$. The *toric ideal* I_A of A is the kernel of the surjective homomorphism $\pi : K[Y] \to K[A]$ defined by setting $\pi(y_i) = \mathbf{q}^{\mathbf{a}_i}$ for all $1 \leq i \leq \nu$. It is known that the toric ideal I_A is generated by the binomials $u - v$, where u and v are monomials of $K[Y]$, with $\pi(u) = \pi(v)$. More precisely,

Proposition 6.1. *Work with the same notation as above. Then*

$$I_A = \left\langle Y^{\mathbf{z}^+} - Y^{\mathbf{z}^-} \;\middle|\; \mathbf{z} \in \mathbb{Z}^\nu,\ A\mathbf{z} = \mathbf{0} \right\rangle,$$

where $\mathbf{z} = \mathbf{z}^+ - \mathbf{z}^-$ with $\mathbf{z}^+, \mathbf{z}^- \in \mathbb{N}^\nu$.

Toric ideals have a good property for a *Gröbner basis* (a set of generators satisfying a certain condition). Let $\mathfrak{M}$ denote the set of monomials belonging to $K[Y]$. Fix a *monomial order* $<$ on $K[Y]$, that is, $<$ is a total order on $\mathfrak{M}$ such that (i) $1 < u$ if $1 \neq u \in \mathfrak{M}$ and (ii) for $u, v, w \in \mathfrak{M}$, if $u < v$ then $uw < vw$. The *initial monomial* $in_<(f)$ of $0 \neq f \in K[Y]$ with respect to $<$ is the biggest monomial appearing in f with respect to $<$. The *initial ideal* of I_A with respect to $<$ is the ideal $in_<(I_A)$ of $K[Y]$ generated by all initial monomials $in_<(f)$ with $0 \neq f \in I_A$. Let $\mathcal{G}$ be a finite subset of I_A and write $in_<(\mathcal{G})$ for the ideal $\langle in_<(g) \mid g \in \mathcal{G} \rangle$ of $K[Y]$. A finite set $\mathcal{G}$ of I_A is called a *Gröbner basis* of I_A with respect to $<$ if $in_<(\mathcal{G}) = in_<(I_A)$. A Gröbner basis $\mathcal{G}$ is called *reduced* if, for each $g \in \mathcal{G}$, none of the monomials in g is divided by $in_<(g')$ for some $g \neq g' \in \mathcal{G}$. It is known that a Gröbner basis of I_A with respect to $<$ always exists. Moreover if $\mathcal{G}$ is a Gröbner basis of I_A, then I_A is generated by $\mathcal{G}$. It is known for Gröbner bases of toric ideals that

Proposition 6.2. *The reduced Gröbner basis of I_A is a finite subset of the set $\{Y^{\mathbf{z}^+} - Y^{\mathbf{z}^-} \mid \mathbf{z} \in \mathbb{Z}^\nu,\ A\mathbf{z} = \mathbf{0}\}$.*

The following proposition is shown by Diaconis–Sturmfels and associates Markov bases with toric ideals.

Proposition 6.3 (Diaconis–Sturmfels, [8]). *A set of moves $\mathcal{B} = \{\mathbf{z}_1, \ldots, \mathbf{z}_L\} \subset \mathcal{M}_A$ is Markov basis if and only if I_A is generated by binomials $Y^{\mathbf{z}_1^+} - Y^{\mathbf{z}_1^-}$, ..., $Y^{\mathbf{z}_L^+} - Y^{\mathbf{z}_L^-}$.*

Second, we introduce the notion of algebras of Segre-Veronese type. Fix integers $\tau \geq 2$, $M \geq 1$ and sets of integers $\mathbf{a} = \{a_1, \ldots, a_M\}$, $\mathbf{b} = \{b_1, \ldots, b_M\}$, $\mathbf{r} = \{r_1, \ldots, r_M\}$ and $\mathbf{s} = \{s_1, \ldots, s_M\}$ such that

(i) $0 \leq b_i \leq a_i$ for all $1 \leq i \leq M$;

(ii) $1 \leq s_i \leq r_i \leq d$ for all $1 \leq i \leq M$.

Let $A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}} \subset \mathbb{N}^d$ denote the configuration consisting of all nonnegative integer vectors $(f_1, f_2, \ldots, f_d) \in \mathbb{N}^d$ such that

(i) $\sum_{j=1}^d f_j = \tau$.

(ii) $b_i \leq \sum_{j=s_i}^{r_i} f_j \leq a_i$ for all $1 \leq i \leq M$.

Then the affine semigroup ring $K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}]$ is generated by all monomials $\prod_{j=1}^d {q_j}^{f_j}$ over K and called an *algebra of Segre-Veronese type*. Note that the present definition generalizes the definition in [14].

Several popular classes of semigroup rings are algebras of Segre-Veronese type. If $M = 2$, $\tau = 2$, $a_1 = a_2 = b_1 = b_2 = 1$, $s_1 = 1$, $s_2 = r_1 + 1$ and $r_2 = d$, then the affine semigroup ring $K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}]$ is the Segre product of polynomial rings $K[q_1, \ldots, q_{r_1}]$ and $K[q_{r_1+1}, \ldots, q_d]$. On the other hand, if $M = d$, $s_i = r_i = i$, $a_i = \tau$ and $b_i = 0$ for all $1 \leq i \leq M$, then the affine semigroup ring $K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}]$ is the classical τth Veronese subring of the polynomial ring $K[q_1, \ldots, q_d]$. Moreover, if $M = d$, $s_i = r_i = i$, $a_i = 1$ and

$b_i = 0$ for all $1 \leq i \leq M$, then the affine semigroup ring $K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}]$ is the τth squarefree Veronese subring of the polynomial ring $K[q_1, \ldots, q_d]$. In addition, algebras of Veronese type (i.e., $M = d$, $s_i = r_i = i$ and $b_i = 0$ for all $1 \leq i \leq M$) are studied in [7] and [20].

Let $K[Y]$ denote the polynomial ring with the set of variables

$$\left\{ y_{j_1 j_2 \cdots j_\tau} \;\middle|\; 1 \leq j_1 \leq j_2 \leq \cdots \leq j_\tau \leq d,\ \prod_{k=1}^{\tau} q_{j_k} \in \{\mathbf{q}^{\mathbf{a}_1}, \ldots, \mathbf{q}^{\mathbf{a}_\nu}\} \right\}.$$

where $K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}] = K[\mathbf{q}^{\mathbf{a}_1}, \ldots, \mathbf{q}^{\mathbf{a}_\nu}]$. The toric ideal $I_{A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}}$ is the kernel of the surjective homomorphism $\pi : K[Y] \longrightarrow K[A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}]$ defined by $\pi(y_{j_1 j_2 \cdots j_\tau}) = \prod_{k=1}^{\tau} q_{j_k}$.

A monomial $y_{\alpha_1 \alpha_2 \cdots \alpha_\tau} y_{\beta_1 \beta_2 \cdots \beta_\tau} \cdots y_{\gamma_1 \gamma_2 \cdots \gamma_\tau}$ is called *sorted* if

$$\alpha_1 \leq \beta_1 \leq \cdots \leq \gamma_1 \leq \alpha_2 \leq \beta_2 \leq \cdots \leq \gamma_2 \leq \cdots \leq \alpha_\tau \leq \beta_\tau \leq \cdots \leq \gamma_\tau.$$

Let $\mathrm{sort}(\cdot)$ denote the operator which takes any string over the alphabet $\{1, 2, \ldots, d\}$ and sorts it into weakly increasing order. Then the quadratic Gröbner basis of toric ideal $I_{A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}}$ is given as follows.

Theorem 6.1. *Work with the same notation as above. Then there exists a monomial order on $K[Y]$ such that the set of binomials*

$$\{y_{\alpha_1 \alpha_2 \cdots \alpha_\tau} y_{\beta_1 \beta_2 \cdots \beta_\tau} - y_{\gamma_1 \gamma_3 \cdots \gamma_{2\tau-1}} y_{\gamma_2 \gamma_4 \cdots \gamma_{2\tau}} \mid \mathrm{sort}(\alpha_1 \beta_1 \alpha_2 \beta_2 \cdots \alpha_\tau \beta_\tau) = \gamma_1 \gamma_2 \cdots \gamma_{2\tau}\} \quad (14)$$

is the reduced Gröbner basis of the toric ideal $I_{A_{\tau,\mathbf{a},\mathbf{b},\mathbf{r},\mathbf{s}}}$. The initial ideal is generated by squarefree quadratic (nonsorted) monomials.

In particular, the set of all integer vectors corresponding to the above binomials is a Markov basis. Furthermore the set is minimal as a Markov basis.

We omit a proof of this theorem. This theorem can be proved along the lines of [20, Theorem 14.2] and the proof in [16]. We will give a generalization of this theorem in our subsequent paper.

Finally we describe how to run a Markov chain using the Gröbner basis given in Theorem 6.1. First, given a configuration A in (7), we check that (with appropriate reordering of rows) that A is indeed a configuration of Segre-Veronese type. It is easy to check that our models in Sections 3 and 4 are of Segre-Veronese type, because the restrictions on choices are imposed separately for each group or each subgroup. Recall that each column of A consists of non-negative integers whose sum τ is common.

We now associate to each column $\mathbf{a}_i$ of A a set of indices indicating the rows with positive elements $a_{\mathbf{ji}} > 0$ and a particular index $\mathbf{j}$ is repeated $a_{\mathbf{ji}}$ times. For example if $d = 4, \tau = 3$ and $\mathbf{a}_i = (1, 0, 2, 0)'$, then row 1 appears once and row 3 appears twice in $\mathbf{a}_i$. Therefore we associate the index $(1, 3, 3)$ to $\mathbf{a}_i$. We can consider the set of indices as $\tau \times \nu$ matrix $\tilde{A}$. Note that $\tilde{A}$ and A carry the same information.

Given $\tilde{A}$, we can choose a random element of the reduced Gröbner basis of Theorem 6.1 as follows. Choose two columns (i.e. choose two cells from $\mathcal{I}$) of $\tilde{A}$ and sort $2 \times \tau$ elements of these two columns. From the sorted elements, pick alternate elements and

form two new sets of indices. For example if $\tau = 3$ and two chosen columns of $\tilde{A}$ are $(1, 3, 3)$ and $(1, 2, 4)$, then sorting these 6 elements we obtain $(1, 1, 2, 3, 3, 4)$. Picking alternate elements produces $(1, 2, 3)$ and $(1, 3, 4)$. These new sets of indices correspond to (a possibly overlapping) two columns of $\tilde{A}$, hence to two cells of $\mathcal{I}$. Now the difference of the two original columns and the two sorted columns of $\tilde{A}$ correspond to a random binomial in (14), hence to a move in (13). It should be noted that when the sorted columns coincide with the original columns, then we discard these columns and choose other two columns. The rest of the procedure for running a Markov chain is described in [8]. See also [2].

7 Numerical examples

In this section we present numerical experiments on NCT data and a diplotype frequency data.

7.1 The analysis of NCT data

First we consider the analysis of NCT data concerning selections in Social Studies and Science. We omit Mathematics for simplicity. Because NCUEE currently do not provide cross tabulations of frequencies of choices across the major subjects, we can not evaluate the P-value of the actual data. However for the models in Section 3, the sufficient statistics (the marginal frequencies) can be obtained from Tables 10–14. Therefore in this section we evaluate the conditional null distribution of the Pearson's χ^2 statistic by MCMC and compare it to the asymptotic χ^2 distribution.

In Section 3, we consider three models, complete independence model, subgroup-wise independence model and group-wise independence model, for the setting of group-wise selection problems. Note that, however, the subgroup-wise independence model coincides with the group-wise independence model for NCT data, since $c_{jk} \leq 1$ for all j and k. Therefore we consider fitting of the complete independence model and the group-wise independence model for NCT data.

As we have seen in Section 2, there are many kinds of choices for each examinee. However, it may be natural to treat some similar subjects as one subject. For example, WHA and WHB may well be treated as WH, ChemI and Chem IA may well be treated as Chem, and so on. As a result, we consider the following aggregation of subjects.

- In Social Studies: WH = {WHA,WHB}, JH = {JHA,JHB}, Geo = {GeoA,GeoB}
- In Science: CSiB = {CSiB, ISci}, Bio = {BioI, BioIA}, Chem = {ChemI, ChemIA}, Phys = {PhysI, PhysIA}, Earth = {EarthI, EarthIA}

In our analysis, we take a look at examinees selecting two subjects for Social Studies and two subjects for Science. Therefore

$$J = 2, m_1 = 2, m_2 = 3, m_{11} = m_{12} = 3, m_{21} = m_{22} = m_{23} = 2,$$
$$c_{11} = c_{12} = 1, (c_{21}, c_{22}, c_{23}) = (1, 1, 0) \text{ or } (1, 0, 1) \text{ or } (0, 1, 1).$$

The number of possible combination is then $\nu = |\mathcal{I}| = 3 \cdot 3 \times 3 \cdot 2^2 = 108$. Accordingly our sample size n is $n = 195094$, which is the number of examinees selecting two subjects on Science from Table 12. Our data set is shown in Table 1.

Table 1: The data set of number of the examinees in NCT in 2006 ($n = 195094$)

	ContS	Ethics	P&E
WH	32352	8839	8338
JH	51573	8684	14499
Geo	59588	4046	7175

	CSiA	Chem	Phys	Earth
CSiB	1648	1572	169	4012
Bio	21392	55583	1416	1845
Phys	3286	102856	—	—
Earth	522	793	—	—

From Table 1, we can calculate the maximum likelihood estimates of the numbers of the examinees selecting each combination of subjects. The sufficient statistics under the complete independence model are the numbers of the examinees selecting each subject, whereas the sufficient statistics under the group-wise independence model are the numbers of the examinees selecting each combination of subjects in the same group. The maximum likelihood estimates calculated from the sufficient statistics are shown in Table 2. For the complete independence model the maximum likelihood estimates can be calculated as in Section 5.2 of [5].

The configuration A for the complete independence model is written as

$$A = \begin{bmatrix} E_3 \otimes \mathbf{1}_3' & \otimes & \mathbf{1}_{12}' \\ \mathbf{1}_3' \otimes E_3 & \otimes & \mathbf{1}_{12}' \\ \mathbf{1}_9' & \otimes & B \end{bmatrix}$$

and the configuration A for the group-wise independence model is written as

$$A = \begin{bmatrix} E_9 \otimes \mathbf{1}_{12}' \\ \mathbf{1}_9' \otimes E_{12}' \end{bmatrix},$$

where E_n is the $n \times n$ identity matrix, $\mathbf{1}_n = (1, \ldots, 1)'$ is the $n \times 1$ column vector of 1's, $\otimes$ denotes the Kronecker product and

$$B = \begin{bmatrix} 111100000000 \\ 000011110000 \\ 100010001100 \\ 010001000011 \\ 001000101010 \\ 000100010101 \end{bmatrix}.$$

Note that the configuration B is the vertex-edge incidence matrix of the $(2, 2, 2)$ complete multipartite graph. Quadratic Gröbner bases of toric ideals arising from complete multipartite graphs are studied in [14].

Given these configurations we can easily run a Markov chain as discussed at the end of Section 6. After $5,000,000$ burn-in steps, we construct $10,000$ Monte Carlo samples.

Table 2: MLE of the number of the examinees selecting each combination of subjects under the complete independence model (upper) and the group-wise independence model (lower).

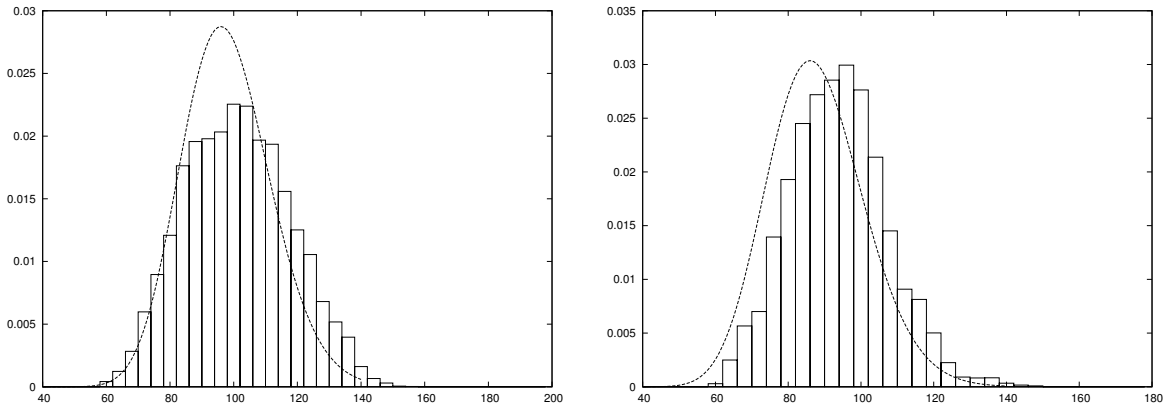
	WH			JH			Geo		
	ContS	Ethics	P&E	ContS	Ethics	P&E	ContS	Ethics	P&E
CSiB,CSiA	180.96	27.20	37.84	273.12	41.05	57.12	258.70	38.88	54.10
	273.28	74.66	70.43	435.65	73.36	122.48	503.35	34.18	60.61
CSiB,Chem	1083.82	162.89	226.65	1635.85	245.86	342.10	1549.48	232.88	324.03
	260.68	71.22	67.18	415.56	69.97	116.83	480.14	32.60	57.81
CSiB,Phys	110.04	16.54	23.01	166.09	24.96	34.73	157.32	23.64	32.90
	28.02	7.66	7.22	44.68	7.52	12.56	51.62	3.50	6.22
CSiB,Earth	7.33	1.10	1.53	11.06	1.66	2.31	10.47	1.57	2.19
	665.30	181.77	171.47	1060.57	178.58	298.16	1225.39	83.20	147.55
Bio,CSiA	1961.78	294.84	410.26	2960.99	445.02	619.21	2804.66	421.52	586.52
	3547.39	969.19	914.26	5654.96	952.20	1589.81	6533.81	443.64	786.74
Bio,Chem	11749.94	1765.93	2457.19	17734.63	2665.39	3708.74	16798.27	2524.66	3512.92
	9217.20	2518.26	2375.53	14693.34	2474.10	4130.82	16976.84	1152.72	2044.18
Bio,Phys	1193.01	179.30	249.49	1800.65	270.63	376.56	1705.58	256.34	356.68
	234.81	64.15	60.52	374.32	63.03	105.23	432.49	29.37	52.08
Bio,Earth	79.43	11.94	16.61	119.88	18.02	25.07	113.55	17.07	23.75
	305.95	83.59	78.85	487.72	82.12	137.12	563.52	38.26	67.85
CSiA,Phys	2691.94	404.58	562.95	4063.04	610.65	849.68	3848.52	578.41	804.82
	544.91	148.88	140.44	868.65	146.27	244.21	1003.65	68.15	120.85
CSiA,Earth	179.22	26.94	37.48	270.50	40.65	56.57	256.22	38.51	53.58
	86.56	23.65	22.31	137.99	23.24	38.79	159.44	10.83	19.20
Bio,Phys	16123.14	2423.20	3371.73	24335.27	3657.42	5089.09	23050.40	3464.31	4820.39
	17056.38	4660.03	4395.90	27189.93	4578.31	7644.05	31415.54	2133.10	3782.75
Bio,Earth	1073.41	161.33	224.48	1620.14	243.50	338.81	1534.60	230.64	320.92
	131.50	35.93	33.89	209.63	35.30	58.93	242.21	16.45	29.16

Figure 1 show histograms of the Monte Carlo sampling generated from the exact conditional distribution of the Pearson goodness-of-fit χ^2 statistics for the NCT data under the complete independence model and the group-wise independence model, respectively, along with the corresponding asymptotic distributions χ^2_{98} and χ^2_{88}.

7.2 The analysis of PTGDR (prostanoid DP receptor) diplotype frequencies data

Next we give a numerical example of genome data. Table 3 shows diplotype frequencies on the three loci, T-549C (locus 1), C-441T (locus 2) and T-197C (locus 3) in the human genome 14q22.1, which is given in [13]. Though the data is used for the genetic association studies in [13], we simply consider fitting our models. As an example, we only consider the diplotype data of patients in the population of blacks ($n = 79$).

First we consider the analysis of genotype frequency data. Though Table 3 is diplotype frequency data, here we ignore the information on the haplotypes and simply treat



Complete independence model ($df = 98$) Group-wise independence model ($df = 88$)

Figure 1: Asymptotic and Monte Carlo sampling distributions of NCT data

Table 3: PTGDR diplotype frequencies among patients and controls in each population. (The order of the SNPs in the haplotype is T-549C, C-441T and T-197C.)

Diplotype	Whites		Blacks	
	Controls	Patients	Controls	Patients
CCT/CCT	16	78	7	10
CCT/TTT	27	106	12	27
CCT/TCT	48	93	4	12
CCT/CCC	17	45	3	9
TTT/TTT	9	43	2	7
TTT/TCT	34	60	8	6
TTT/CCC	4	28	1	6
TCT/TCT	11	20	7	0
TCT/CCC	6	35	1	2
CCC/CCC	1	8	0	0

it as a genotype frequency data. Since $J = 3$ and $m_1 = m_2 = m_3 = 2$ holds, there are $3^3 = 27$ distinct set of genotypes, i.e., $|\mathcal{I}| = 27$, while only 8 distinct haplotypes appear in Table 3. Table 4 is the set of genotype frequencies of patients in the population of blacks. Under the genotype-wise independence model (8), the sufficient statistic is the genotype frequency data for each locus. On the other hand, under the Hardy-Weinberg model (9), the sufficient statistic is the allele frequency data for each locus, and the genotype frequencies for each locus are estimated by the Hardy-Weinberg law. Accordingly, the maximum likelihood estimates for the combination of the genotype frequencies are calculated as Table 5. The configuration A for the Hardy-Weinberg model is written as

Table 4: The genotype frequencies for patients among blacks of PTGDR data

locus 3		CC			CT			TT		
locus 2		CC	CT	TT	CC	CT	TT	CC	CT	TT
locus 1	CC	0	0	0	9	0	0	10	0	0
	CT	0	0	0	2	6	0	12	27	0
	TT	0	0	0	0	0	0	0	6	7

Table 5: MLE for PTGDR genotype frequencies of patients among blacks under the Hardy-Weinberg model (upper) and genotype-wise independence model (lower)

locus 3		CC			CT			TT		
locus 2		CC	CT	TT	CC	CT	TT	CC	CT	TT
locus 1	CC	0.1169	0.1180	0.0298	1.939	1.958	0.4941	8.042	8.118	2.049
		0	0	0	1.708	2.018	0.3623	6.229	7.361	1.321
	CT	0.2008	0.2027	0.0512	3.331	3.362	0.8486	13.81	13.94	3.519
		0	0	0	4.225	4.993	0.8962	15.41	18.21	3.268
	TT	0.0862	0.0870	0.0220	1.430	1.444	0.3644	5.931	5.988	1.511
		0	0	0	1.169	1.381	0.2479	4.262	5.037	0.9040

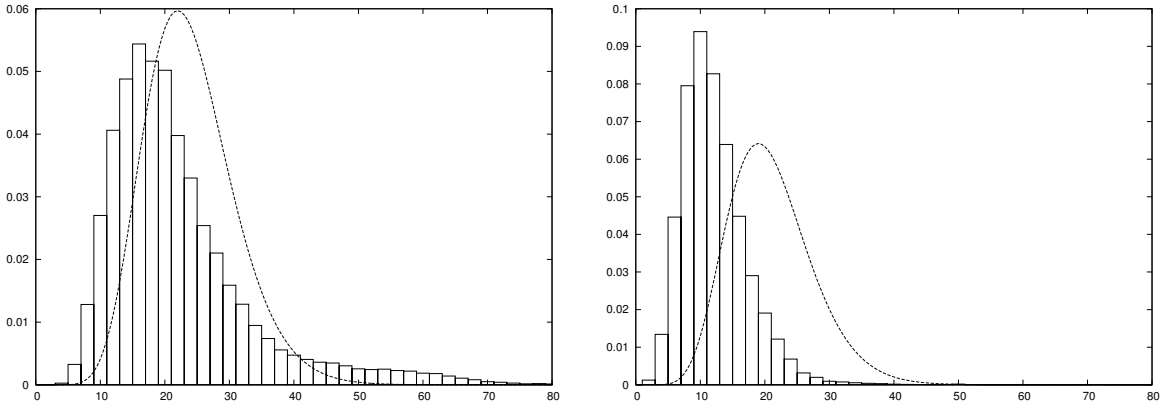
$$A = \begin{bmatrix} 222222222 & 111111111 & 000000000 \\ 000000000 & 111111111 & 222222222 \\ 222111000 & 222111000 & 222111000 \\ 000111222 & 000111222 & 000111222 \\ 210210210 & 210210210 & 210210210 \\ 012012012 & 012012012 & 012012012 \end{bmatrix}$$

and the configuration A for the genotype-wise independence model is written as

$$A = \begin{bmatrix} E_3 \otimes \mathbf{1}_3' \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes E_3 \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes \mathbf{1}_3' \otimes E_3' \end{bmatrix}.$$

Since these two configurations are of the Segre-Veronese type, again we can easily perform MCMC sampling as discussed in Section 6. After $100,000$ burn-in steps, we construct $10,000$ Monte Carlo samples. Figure 2 show histograms of the Monte Carlo sampling generated from the exact conditional distribution of the Pearson goodness-of-fit χ^2 statistics for the PTGDR genotype frequency data under the Hardy-Weinberg model and the genotype-wise independence model, respectively, along with the corresponding asymptotic distributions χ^2_{24} and χ^2_{21}.

From the Monte Carlo samples, we can also estimate the P-values for each null model. The values of the Pearson goodness-of-fit χ^2 for the PTGDR genotype frequency data of Table 4 are $\chi^2 = 88.26$ under the Hardy-Weinberg models, whereas $\chi^2 = 103.37$ under the genotype-wise independence model. These values are highly significant ($p < 0.01$ for both models), which implies the susceptibility of the particular haplotypes.



Hardy-Weinberg model ($df = 24$) Genotype-wise independence model ($df = 21$)

Figure 2: Asymptotic and Monte Carlo sampling distributions of PTGDR genotype frequency data

Next we consider the analysis of the diplotype frequency data. In this case of $J = 3$ and $m_1 = m_2 = m_3 = 2$, there are $2^3 = 8$ distinct haplotypes, and there are

$$|\mathcal{I}| = 8 + \binom{8}{2} = 36$$

distinct diplotypes, while there are only 4 haplotypes and 10 diplotypes appear in Table 3. The numbers of each haplotype are calculated as the second column of Table 6. Under the Hardy-Weinberg model, the haplotype frequencies are estimated proportionally to the allele frequencies, which is shown as the third column of Table 6. The maximum likelihood

Table 6: Observed frequency and MLE under the Hardy-Weinberg model for PTGDR haplotype frequencies of patients among blacks.

Haplotype	observed	MLE under HW	Haplotype	observed	MLE under HW
CCC	17	6.078	TCC	0	5.220
CCT	68	50.410	TCT	20	43.293
CTC	0	3.068	TTC	0	2.635
CTT	0	25.445	TTT	53	21.853

estimates of the diplotype frequencies under the Hardy-Weinberg model are calculated from the maximum likelihood estimates for each haplotype. These values coincide with appropriate fractions of the values for the corresponding combination of the genotypes in Table 5. For example, the MLE for the diplotype CCT/CCT coincides with the MLE for the combination of the genotypes (CC,CC,TT) in Table 5, whereas the MLE's for the diplotype CCC/TTT, CCT/TTC, CTC/TCT, CTT/TCC coincide with the $\frac{1}{4}$ fraction of the MLE for the combination of the genotypes (CT,CT,CT), and so on. Since we know that the Hardy-Weinberg model is highly statistically rejected, it is natural to consider the haplotype-wise Hardy-Weinberg model given in Section 4.2. Table 7 shows the maximum

likelihood estimates under the haplotype-wise Hardy-Weinberg model. It should be noted that the MLE for the other diplotypes are all zeros. We perform the Markov chain Monte

Table 7: MLE for PTGDR diplotype frequencies of patients among blacks under the haplotype-wise Hardy-Weinberg model.

Diplotype	observed	MLE	Diplotype	observed	MLE
CCT/CCT	10	14.6329	TTT/TCT	6	6.7089
CCT/TTT	27	22.8101	TTT/CCC	6	5.7025
CCT/TCT	12	8.6076	TCT/TCT	0	1.2658
CCT/CCC	9	7.3165	TCT/CCC	2	2.1519
TTT/TTT	7	8.8892	CCC/CCC	0	0.9146

Carlo sampling for the haplotype-wise Hardy-Weinberg model. The configuration A for this model is written as

$$A = \begin{bmatrix} 20000000111111100000000000000000000000 \\ 02000000100000011111100000000000000000 \\ 00200000010000010000011111000000000000 \\ 00020000001000001000010000111100000000 \\ 00002000000100000100001000100011100000 \\ 00000200000010000010000100010001100110 \\ 00000020000001000001000010001001010101 \\ 00000002000000100000100001000100101011 \end{bmatrix},$$

which is obviously of the Segre-Veronese type. We give a histogram of the Monte Carlo sampling generated from the exact conditional distribution of the Pearson goodness-of-fit χ^2 statistics for the PTGDR diplotype frequency data under the haplotype-wise Hardy-Weinberg model, along with the corresponding asymptotic distributions χ^2_9 in Figure 3.

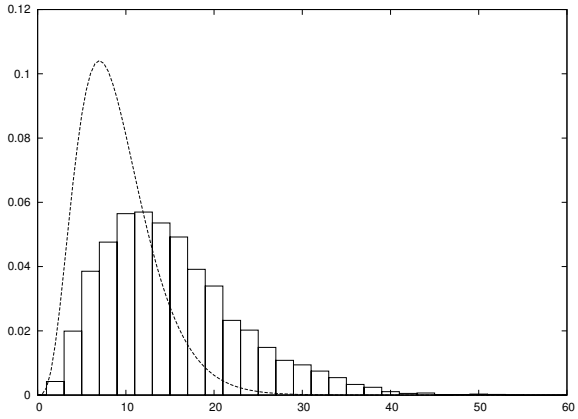


Figure 3: Asymptotic and Monte Carlo sampling distributions of PTGDR diplotype frequency data under the haplotype-wise Hardy-Weinberg model ($df = 9$).

The P-value for this model is estimated as 0.8927 with the estimated standard deviation 0.0029 (We also discard the first $100,000$ samples, and use a batching method to

obtain an estimate of variance, see [9] and [19]).

8 Some discussions

In this paper we considered independence models in group-wise selections, which can be described in terms of a Segre-Veronese configuration. We have shown that our framework can be applied to two important examples in educational statistics and biostatistics. We expect that the methodology of the present paper finds applications in many other fields.

In the NCT example, we assumed that the examinees choose the same number τ of subjects. We also assumed for simplicity that the examinees choose either nothing or one subject from a subgroup. This restricts our analysis to some subset of the examinees of NCT. Actually the examinees make decisions on how many subjects to take and modeling this decision making is clearly of statistical interest. Further complication arises from the fact that the examinees can choose which scores to submit to universities after taking NCT. For example after obtaining scores of three subjects on Science, an examinee can choose the best two scores for submitting to a university.

It seems that the simplicity of the reduced Gröbner basis for the Segre-Veronese configuration comes from the fact that the index set $\mathcal{J}$ of the rows of A can be ordered and the restriction on the counts can be expressed in terms of one-dimensional intervals. From statistical viewpoint, ordering of the elements of the sufficient statistic in group-wise selection seems to be somewhat artificial. It is of interest to look for other statistical models, where ordering of the elements of the sufficient statistic is more natural and the Segre-Veronese configuration can be applied.

References

[1] Aoki, S. and Takemura, A. (2005). Markov chain Monte Carlo exact tests for incomplete two-way contingency tables. *Journal of Statistical Computation and Simulation*. **75**, 787–812.

[2] Aoki, S. and Takemura, A. (2006). Markov chain Monte Carlo tests for designed experiments. `arXiv:math/0611463v1`. Submitted for publication.

[3] Aoki, S. and Takemura, A. (2007). Minimal invariant Markov basis for sampling contingency tables with fixed marginals. *Annals of the Institute of Statistical Mathematics*, To appear.

[4] Beerenwinkel, N., Eriksson, N. and Sturmfels, B. (2006). Conjunctive Bayesian Networks. `arXiv:math/0608417v2`.

[5] Bishop, Y. M. M., Fienberg, S. E. and Holland P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, Massachusetts.

[6] Crow, J. E. (1988). Eighty years ago: The beginnings of population genetics. *Genetics*, **119**, 473–476.

[7] De Negri, E. and Hibi, T. (1997). Gorenstein algebras of Veronese type, *J. Algebra*, **193**, no. 2, 629–639.

[8] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, **26**, 363–397.

[9] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[10] Hibi, T. (1987). Distributive lattices, affine semigroup rings and algebras with straightening laws. *Adv. Stud. Pure Math.*, **11**, 93–109.

[11] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York.

[12] National Center for University Entrance Examinations. (2006). Booklet on NCUEE. Available from `http://www.dnc.ac.jp/dnc/gaiyou/pdf/youran_english_H18_HP.pdf`

[13] Oguma, T., Palmer, L. J., Birben, E., Sonna, L. A. Asano, K. and Lilly, C. M. (2004). Role of prostanoid DP receptor variants in susceptibility to asthma, *The New England Journal of Medicine*, **351**, 1752–1763.

[14] Ohsugi, H. and Hibi, T. (2000). Compressed polytopes, initial ideals and complete multipartite graphs, *Illinois J. Math.* **44**, 391–406.

[15] Ohsugi, H. and Hibi, T. (2005). Indispensable binomials of finite graphs. *Journal of Algebra and Its Applications*, **4**, 421–434.

[16] Ohsugi, H. and Hibi, T. (2006). Quadratic Gröbner bases arising from combinatorics, preprint.

[17] Ohsugi, H. and Hibi, T. (2007). Toric ideals arising from contingency tables. In *Proceedings of the Ramanujan Mathematical Society's Lecture Notes Series*. To appear.

[18] Pachter, L. and Sturmfels, B. (2005). *Algebraic Statistics for Computational Biology.* Cambridge University Press, Cambridge.

[19] Ripley, B. D. (1987). *Stochastic Simulation*, Wiley, New York.

[20] Sturmfels, B. (1995). *Gröbner Bases and Convex Polytopes*. American Mathematical Society, Providence, RI.

[21] Takemura, A. and Aoki, S. (2004). Some characterizations of minimal Markov basis for sampling from discrete conditional distributions, *Annals of the Institute of Statistical Mathematics*, **56**, 1–17.

A Tables of numbers of examinees in NCT in 2006

Table 8: Number of examinees who takes subjects on Mathematics

	Mathematics 1		Mathematics 2					# total examinees	# actual examinees
	MathI	MathIA	MathII	MathIIB	BMSI	BKA	Info		
1 subject	6,454	33,381	113	363	4	480	93	40,888	40,888
2 subjects	7,553	322,758	12,076	317,099	83	591	462	660,622	330,311
Total	14,007	356,139	12,189	317,462	87	1,071	555	701,510	371,199

Table 9: Number of examinees who selects two subjects on Mathematics

	Mathematics 2					
Mathematics 1	MathII	MathIIB	BMSI	BKA	Info	Total
MathI	5,065	2,159	19	217	93	7,553
MathIA	7,011	314,940	64	374	369	322,758
Total	12,076	317,099	83	591	462	330,311

Table 10: Number of examinees who takes subjects on Social Studies

	Geography and History						Civics			# total examinees	# actual examinees
	WHA	WHB	JHA	JHB	GeoA	GeoB	ContS	Ethics	P&E		
1 subject	496	29,108	1,456	54,577	1,347	27,152	40,677	16,607	25,321	196,741	196,741
2 subjects	1,028	61,132	3,386	90,427	5,039	83,828	180,108	27,064	37,668	489,680	244,840
Total	1,524	90,240	4,842	145,004	6,386	110,980	220,785	43,671	62,989	686,421	441,581

Table 11: Number of examinees who selects two subjects on Social Studies

	Geography and History						
Civics	WHA	WHB	JHA	JHB	GeoA	GeoB	Total
ContSoc	687	39,913	2,277	62,448	3,817	70,966	180,108
Ethics	130	10,966	409	10.482	405	4,672	27,064
P&E	211	10253	700	17,497	817	8,190	37,668
Total	1,028	61,132	3,386	90,427	5,039	83,838	244,840

Table 12: Number of examinees who takes subjects on Science

	Science 1				Science 2			Science 3				# total	#actual
	CSciB	BioI	ISci	BioIA	CSciA	ChemI	ChemIA	PhysI	EarthI	PhysIA	EarthIA	examinees	examinees
1 subject	2,558	80,385	511	1,314	1,569	19,616	717	14,397	10,788	289	236	132,380	132,380
2 subjects	6,878	79,041	523	1,195	26,848	158,027	2,777	106,822	6,913	905	259	390,188	195,094
3 subjects	7,942	18,519	728	490	6,838	20,404	437	18,451	8,423	361	444	83,037	27,679
Total	17,378	177,945	1,762	2,999	35,255	198,047	3,931	139,670	26,124	1,555	939	605,605	355,153

Table 13: Number of examinees who selects two subjects on Science

		Science 2			Science 3			
		CSciA	ChemI	ChemIA	PhysI	EarthI	PhysIA	EarthIA
Science 1	CSciB	1,501	1,334	23	120	3,855	1	44
	BioI	21,264	54,412	244	1,366	1,698	5	52
	ISci	147	165	50	43	92	5	21
	BioIA	128	212	715	16	33	29	62
Science 3	Physics	3,243	101,100	934	—	—	—	—
	EarthI	485	730	20	—	—	—	—
	PhysIA	43	54	768	—	—	—	—
	EarthIA	37	20	23	—	—	—	—

Table 14: Number of examinees who selects three subjects on Science

Science 3		PhysI			EarthI			Physics IA			Earth science IA		
Science 2		CSciA	ChemI	ChemIA	CSciA	ChemI	ChemIA	CSciA	ChemI	ChemIA	CSciA	ChemI	ChemIA
Science 1	CSciB	1,155	5,152	17	1,201	317	7	16	5	16	48	5	3
	BioI	553	10,901	31	3,386	3,342	16	30	35	19	130	56	20
	ISci	80	380	23	62	34	4	32	13	27	48	14	11
	BioIA	6	114	39	22	22	10	12	6	150	57	8	44