# Estimating Cell Probabilities in Contingency Tables with Constraints on Marginals/Conditionals by Geometric Programming with Applications

Xinlei WANG, Johan LIM, Seung-Jean KIM, Kyu S. HAHN *

**Abstract**

Contingency tables are often used to display the multivariate frequency distribution of variables of interest. Under the common multinomial assumption, the first step of contingency table analysis is to estimate cell probabilities. It is well known that the unconstrained Maximum Likelihood Estimator (MLE) is given by cell counts divided by the total number of observations. However, in the presence of (complex) constraints on the unknown cell probabilities or their functions, the MLE or other types of estimators may often have no closed form and have to be obtained numerically. In this paper, we focus on finding the MLE of cell probabilities in contingency tables under two common types of constraints: known marginals and ordered marginals/conditionals, and propose a novel approach based on geometric programming. We present two important applications that illustrate the usefulness of our approach via comparison with existing methods. Further, we show that our GP-based approach is flexible, readily implementable, effort-saving and can provide a unified framework for various types of constrained estimation of cell probabilities in contingency tables.

*Keywords*: known marginals; ordered conditionals; ordered marginals; judgement post-stratification; Matlab; monomial; multinomial distribution; posynomial; stochastic ordering.

*Xinlei Wang is Associate Professor, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275 (swang@smu.edu). Johan Lim is Professor, Department of Statistics, Seoul National University, Seoul, Korea, johanlim@snu.ac.kr. Seung-Jean Kim is Vice President, Citi Capital Advisors, New York City, NY. Kyu S. Hahn is Assistant Professor, Department of Communication, Seoul National University, Seoul, Korea.

# 1 Introduction

Let $g(x)$ be a *monomial* function defined in the form of

$$g(x) = cx_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}$$

where $c > 0$, $a_i \in R$ and $x_i \in \mathbf{R}^+$ for $1 \le i \le n$. Let $f(x)$ be a *posynomial* function defined as the sum of one or more monomials

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}.$$

Then a special type of convex optimization problem, called a *geometric program* (GP), can be described as follows:

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 1 \quad i = 1, \cdots m, \\ & g_j(x) = 1 \quad j = 1, \cdots, p, \end{aligned} \tag{1.1}$$

where $x = (x_1, \cdots, x_n)$ is the row vector containing all (positive) optimization variables, $f_i$'s are posynomial functions and $g_j$'s are monomial functions for all $i = 0, \cdots, m$ and $j = 1, \cdots, p$.

*Geometric programming* is a common technique for solving the above optimization problem. It was first introduced by Duffin et al. (1967). More recently, new methods have been developed to solve large-scale GPs very efficiently and reliably (Nesterov and Nemirovsky 1994, Boyd et al. 2005). Since then, useful applications such as large digital/analog circuit designing, floor planning, and wireless power control have emerged. As pointed out by Boyd et al. (2005), the new methods have several advantages: (1) fast computing speed; (2) global optimal solution guaranteed for any feasible GP; and (3) no extra human effort required (such as parameter tuning, starting point selection, or initial guessing). These technical advances, combined with the wide availability of software for implementation (e.g., a freeware named GGPLAB, Mosek, YALMIP), make geometric programming even more useful for statisticians' toolkit. Once GP modeling, i.e., formulating problems into the GP form (1.1), is done, which hardly requires any knowledge of technical details and is conceptually simple, one has an effective and reliable way for solving a practical problem. For detail about geometric programming, see Boyd et al. (2007) and Lim et al. (2009a).

Although it is less well known among statisticians than it should be, the idea of using geometric programming for constrained estimation has been adopted to solve several statistical problems in the literature. For example, Alldredge and Armstrong (1974) considered GP-based estimation of overlap sizes created by interlocking sampling schemes; Mazumdar and Jefferson (1983) used the GP method for estimating gene frequencies or success probabilities when sums of independent Bernoulli random variables are observed; Lim et al. (2009a) conducted GP modeling for estimating survival functions subject to stochastic orderings, and then solved the problem by GP; and Lim and Won (2012) proposed a GP-based method for estimating a concave ROC curve. However, little work has been done by other researchers under the context of contingency tables, which play an important role in summarizing and displaying categorical data. The only work that we are aware of is Bricker et al. (1997), where GP-based approaches were proposed for estimating cell probabilities in a two-way contingency table given positive/negative association (i.e., certain inequality constraints on local odd ratios need to be posed) or in a one-way table with upper bounds on ratios of cell probabilities. Clearly, this earlier application of GP to contingency tables is restricted to specific and small problems, involving low-dimensional tables and monomial inequality constraints only.

In this paper, we consider the use of geometric programming with contingency tables, to address two important types of problems that occur frequently in practice. The first is estimating cell probabilities with known marginals; and the second is estimating cell probabilities under order constraints on marginals/conditionals. We show that under the multinomial assumption, the problem of finding the constrained maximum likelihood estimates can be solved via GP in various situations. As opposed to most existing work that handles one or two-way tables only, we thoroughly examine distinctive situations in three-way contingency tables; and for higher dimensional tables, we provide sufficient conditions, which guarantee that the problem can be handled using GP. In addition, the problems we consider are more general and harder in nature than those in Bricker et al. (1997), with useful applications from various fields, as will be discussed later. Through simulation and examples, we show that for sparse tables (i.e., a large number of cells have zero or small counts), the GP method can provide higher estimation efficiency than Iterative Proportional Fitting (IPF), which is known as raking in the context of contingency tables, the most popular algorithm for solving the first problem. Also, for the second problem, the existing algorithms, Pooled Adjacent Violator or its variants (e.g., Barlow et al. 1972, Jewell and Kalbfleisch 2004), cannot be used with tables of three ways or higher. When compared with the Newton-Raphson method (NR, a well known general-purpose optimization method implemented via quadratic programming in this work), it can run much faster besides providing improved estimation for sparse tables. Even in cases where both methods can work well, the GP method is more attractive because of ease of implementation; and there is no need for initial guessing and it can always guarantee a global optimal solution to any feasible GP problem. By contrast, implementing NR may require effort and advanced knowledge in optimization; and the performance of NR may rely on specification of starting points. In some cases, even after many trials, NR can be still slow in convergence.

Our work is partly motivated by two scientific studies, which differ significantly in application areas. The first is to estimate the population mean from judgment post-stratification (JP-S) samples with multiple rankers (e.g., MacEachern et al. 2004, Wang et al. 2006, Stokes et al. 2007, Frey et al. 2007, Frey and Ozturk 2011), where estimation of cell probabilities needs to be done under known one-dimensional marginal probabilities; and the second is to examine whether there exists gender-based inequality in returns to education through analysis of a three-way table, where order constraints on conditional probabilities arise naturally in this context. Both applications show the usefulness of our GP-based approach. The Matlab code and data used in this paper are available at the URL http://stat.snu.ac.kr/johanlim/gp_contingency_examples.html.

The paper is organized as follows. In Sections 2 and 3, we discuss our estimation of cell probabilities in contingency tables with known marginals and with ordered marginals/conditionals, respectively. Sections 4 and 5 present the two applications and numerical results. Section 6 discusses the general use of GP in contingency tables, including relevant previous work. Section 7 concludes the paper with discussion.

## 2 Estimation under known marginal probabilities

The problem of estimating cell probabilities in contingency tables with known marginals has a rich history in the literature. For example, Deming and Stephan (1940) mentioned that, in census applications, one may often want to estimate probabilities from contingency tables when the marginal probabilities are known from an established theory or a much larger sample. Imposing such equality constraints can reduce estimation bias arising from non-responses and sampling variability. Estimation methods using different objective functions had been proposed, such as Maximum Likelihood, Minimum Discrimination Information, and Quasi-Bayes, mostly with focus on two-dimensional tables; see Pelz and Good (1986) for a summary. Optimizing those objective functions with equality constraints generally has no closed-form solution and so needs to be solved numerically. Deming and Stephan (1940) and Stephan (1942) proposed the well known IPF procedure, also called "raking", for estimating two or three-way tables, which iteratively adjusts column and row proportions to satisfy known marginals. This procedure is simple and fast, as opposed to other existing numerical procedures (e.g., NR, steepest ascent/descent, Southwell relaxation). Little and Wu (1991) compared several estimation methods in the analysis of two-way contingency tables and concluded that raking was one of the top two winners. Bishop and Fienberg (1969) considered extension of raking for two-way contingency tables with zero counts. Also, raking has been proved to converge to the Maximum Likelihood Estimator (MLE) when no empty cells exist (Thompson, 1981). Moreover, it is possible to extend raking to higher dimensional tables (e.g.,Ireland and Kullback 1968).

In this section, we propose a novel approach via geometric programming to find the MLE, whose performance will be compared with that of raking through simulation and an empirical study in Section 4. Although our discussion of GP modeling is based on the MLE, the approach can be extended to other estimators, as will be discussed in Section 6. In what follows, we first analyze all distinctive situations in three-way contingency tables, and then generalize the results to $m-$way tables ($m > 3$) given certain sufficient conditions.

### 2.1 Three-way contingency tables

Consider a $r \times s \times t$ contingency table, where the three dimensions are associated with discrete variables $X_1$, $X_2$ and $X_3$, respectively. Let $i$ index the $i$th value of $X_1$, $j$ index the $j$th value of $X_2$, and $k$ index the $k$th value of $X_3$. Let $p_{ijk}$ denote the probability that an observation falls in the $(i, j, k)$th cell, whose observed frequency is denoted by $n_{ijk}$ in a sample of size $n$ (i.e, $\sum_{i,j,k} n_{ijk} = n$). Assuming these frequencies follow a multinomial distribution with parameters $n$ and $p_{ijk}$s, the likelihood function is then proportional to

$$\mathcal{L}\left(\{p_{ijk}\}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{n_{ijk}}, \tag{2.1}$$

which is a monomial function of $p_{ijk}$s.

For three-way contingency tables, there are two types of marginal probabilities that can be known: one dimensional such as $p_{i++}$ (i.e., some univariate information about $X_1$ is known ) and two-dimensional such as $p_{ij+}$ (i.e., some bivariate information about $X_1$ and $X_2$ is known). In what follows, we show under three distinctive situations, the MLE of $p_{ijk}$s can be computed via geometric programming. Instead of maximizing the likelihood function $\mathcal{L}$, we minimize the inverse of $\mathcal{L}$ equivalently.

(i) **All three sets of one-dimensional marginal probabilities are known.** Here, the optimization problem can be described as

$$\begin{array}{lll} \text{minimize} & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}}, & \\ \text{subject to} & \sum_{j,k} p_{ijk} = p_{i++}, & \text{for } i = 1, \cdots, r, \\ & \sum_{i,k} p_{ijk} = p_{+j+}, & \text{for } j = 1, \cdots, s, \\ & \sum_{i,j} p_{ijk} = p_{++k}, & \text{for } k = 1, \cdots, t, \end{array} \tag{2.2}$$

where $p_{i++}$, $p_{+j+}$ and $p_{++k}$ are positive constants satisfying $\sum_i p_{i++} = \sum_j p_{+j+} = \sum_k p_{++k} = 1$.

(ii) **One set of one-dimensional marginal probabilities and the remaining two-dimensional marginal probabilities are known.** Without loss of generality, we can describe the optimization problem as

$$\begin{array}{lll} \text{minimize} & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} & \\ \text{subject to} & \sum_{j,k} p_{ijk} = p_{i++}, & \text{for } i = 1, \cdots, r, \\ & \sum_{i} p_{ijk} = p_{+jk}, & \text{for } j = 1, \cdots, s,\ k = 1, \cdots, t, \end{array} \tag{2.3}$$

where $p_{i..}$ and $p_{.jk}$ are all positive constants satisfying $\sum_i p_{i++} = \sum_{j,k} p_{+jk} = 1$.

(iii) **Any two sets of two-dimensional marginal probabilities are known.** Without loss of generality, we can describe the problem as

$$\begin{array}{lll} \text{minimize} & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} & \\ \text{subject to} & \sum_{k} p_{ijk} = p_{ij+}, & \text{for } i = 1, \cdots, r,\ j = 1, \cdots, s, \\ & \sum_{j} p_{ijk} = p_{i+k}, & \text{for } i = 1, \cdots, r,\ k = 1, \cdots, t, \end{array} \tag{2.4}$$

where $p_{ij+}$ and $p_{i+k}$ are all positive constants satisfying $\sum_j p_{ij+} = \sum_k p_{i+k}$ for $1, \cdots, r$ and $\sum_{ij} p_{ij+} = \sum_{ik} p_{i+k} = 1$.

**Theorem 1.** ***Changing all "$=$"s to "$\leq$"s in the constraints, the optimization problem in any of the three situations (i)-(iii) turns into a GP; and the GP is equivalent to the corresponding original optimization problem in estimating the probabilities of all nonempty cells.***

*Proof.* The GP relaxation (i.e., relaxing all "=" constraints s to the "$\leq$"s) is based on the existence of a pivotal cell when constraints in marginal probabilities are violated (i.e., some "<" instead of "=" holds). Here, a pivotal cell is a cell whose probability can be adjusted to remove the violation to the constraints. See Appendix A for a detailed proof that shows the existence of the pivotal cell. □

**Remarks:**

1. There are redundant "$=$" constraints in (2.2)-(2.4). However, once each of the problems has been converted to the corresponding GP, they are no longer redundant and must be kept when solving the GP.

2. If all the observed counts are non-zero (i.e., no empty cells exist) in a contingency table, then the optimal solution from the GP is also optimal for the original optimization problem we try to solve. If there exist some empty cells, then we can find the optimal solution(s) to the original

optimization problem in two steps: (i) solve the corresponding GP to get the estimates for non-empty cells; (ii) plug in the estimates of non-empty cells into the equality constraints in the original optimization problem and then solve the linear equations for the estimates of empty cells. Note that the solution to the probabilities of empty cells may not be unique; and unlike the raking method, the solution may not be zero for empty cells.

In practical situations, it is often the case that only some subset of the constraints is available in each of the three situations discussed above. For example, in the first situation, only one or two sets of one-dimensional marginal probabilities may be known.Or in the second and third situations, only one set of two-dimensional marginal probabilities may be known. Or in all three situations, an incomplete set of marginal probabilities may be known (e.g., $p_{i++}$s may be known for some $i$, not all $i$; $p_{ij+}$s may be known for some $i$ and $j$ instead of all possible $(i,j)$ combinations). Note that, when no sets are complete, the constraint $\sum_{i,j,k} p_{ijk} = 1$ must be present in the original optimization problem. In such situations, GP modeling can be easily done in the same spirit; and the proof of equivalence in estimating the probabilities of all nonempty cells is essentially the same as before since the existence of the pivotal cell would not be affected. For example, for situation (i), if the one-dimensional marginal probabilities are known for only some $i$'s, some $j$'s and some $k$'s (say $i \in \mathcal{A}$, $j \in \mathcal{B}$ and $k \in \mathcal{C}$, where $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are true subsets of the complete sets $\{1, \cdots, r\}$, $\{1, \cdots, s\}$ and $\{1, \cdots, t\}$, respectively). Then the original optimization problem is

$$\begin{array}{lll} \text{minimize} & \prod_{i=1}^{r} \prod_{j=1}^{s} \prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} & \\ \text{subject to} & \sum_{j,k} p_{ijk} = p_{i++}, & \text{for } i \in \mathcal{A}, \\ & \sum_{i,k} p_{ijk} = p_{+j+}, & \text{for } j \in \mathcal{B}, \\ & \sum_{i,j} p_{ijk} = p_{++k}, & \text{for } k \in \mathcal{C}, \\ & \sum_{i,j,k} p_{ijk} = 1. & \end{array}$$

Note that $\sum_{i,j,k} p_{ijk} = 1$ is equivalent to $\sum_{i \in \mathcal{A}^c} \sum_{j,k} p_{ijk} = 1 - \sum_{i \in \mathcal{A}} p_{i++}$, or $\sum_{j \in \mathcal{B}^c} \sum_{i,k} p_{ijk} = 1 - \sum_{j \in \mathcal{B}} p_{+j+}$ or $\sum_{k \in \mathcal{C}^c} \sum_{i,j} p_{ijk} = 1 - \sum_{j \in \mathcal{C}} p_{++k}$. Then we can easily recognize the GP that is equivalent to the problem above in estimating the probabilities of all nonempty cells:

$$
\begin{array}{lll}
\text{minimize} & \prod_{i=1}^{r} \prod_{j=1}^{s} \prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} & \\
\text{subject to} & \sum_{j,k} p_{ijk} \le p_{i++}, & \text{for } i \in \mathcal{A}, \\
& \sum_{i,k} p_{ijk} \le p_{+j+}, & \text{for } j \in \mathcal{B}, \\
& \sum_{i,j} p_{ijk} \le p_{++k}, & \text{for } k \in \mathcal{C}, \\
& \sum_{i \in \mathcal{A}^c} \sum_{j,k} p_{ijk} \le 1 - \sum_{i \in \mathcal{A}} p_{i++}; & \\
& \sum_{j \in \mathcal{B}^c} \sum_{i,k} p_{ijk} \le 1 - \sum_{j \in \mathcal{B}} p_{+j+}; & \\
& \sum_{k \in \mathcal{C}^c} \sum_{i,j} p_{ijk} \le 1 - \sum_{j \in \mathcal{C}} p_{++k}. &
\end{array}
$$

So far, we have shown that for three-way contingency tables with known marginal information, nearly all cases can be converted to GPs, which can be solved very efficiently and reliably. There is only one open case, in which all three sets of two-dimensional marginal probabilities are known; that is,

$$
\begin{aligned}
&\sum_k p_{ijk} = p_{ij+}, \text{ for } i = 1, \cdots, r, \ j = 1, \cdots, s, \\
&\sum_j p_{ijk} = p_{i+k}, \text{ for } i = 1, \cdots, r, \ k = 1, \cdots, t, \\
&\sum_i p_{ijk} = p_{+jk}, \text{ for } j = 1, \cdots, s, \ k = 1, \cdots, t,
\end{aligned} \tag{2.5}
$$

where $p_{ij+}$, $p_{i+k}$ and $p_{+jk}$ are all positive constants satisfying $\sum_{ij} p_{ij+} = \sum_{ik} p_{i+k} = \sum_{jk} p_{+jk} = 1$, $\sum_j p_{ij+} = \sum_k p_{i+k}$, $\sum_i p_{ij+} = \sum_k p_{+jk}$ and $\sum_i p_{i+k} = \sum_j p_{+jk}$ (i.e., all the constraints are feasible). In this case, relaxing "=" to "$\le$" leads to a GP, which, as in the other cases discussed above, was conjectured to be equivalent to the original optimization problem. Although much efforts have been made, neither a proof to the equivalence nor a counterexample has been found. So questions remain about whether the conjecture is true or not. However, this case has very limited practical importance because the complete bivariate information for all the three variables is rarely available in real situations. Still, one handy strategy is to solve the corresponding GP (by simply changing all "=" to "$\le$ " constraints), and then examine whether (2.5) is satisfied; If yes, we have solved the problem.

## 2.2 Extension to multi-way contingency tables

In an $m-$way contingency table, there are $m-1$ types of marginal probabilities, 1-dimensional to $m-1$ dimensional, resulting in $2^m - 2$ different marginal probabilities in total. Thus, the space for the possible combinations of known marginals can be enormous, which makes finding a general approach that is workable in the entire space very difficult. In practice, low-dimensional marginal probabilities are often easier to obtain and more comfortable to use than higher-dimensional ones, especially for the one-dimensional ones that require only univariate information. Even when all the known marginals are one-dimensional, there is no existing solution method that is generally acceptable. As mentioned before, the popular raking method is fast but does not converge to the true MLE for tables with empty cells (a problem widely present in tables of three dimensions or higher); and the NR method can be extremely slow in convergence and requires human effort per dataset. Below we show that the GP method can provide an effortless, reliable and efficient solution for multi-way contingency tables when one-dimensional marginals are known.

Let $i_k$ index the $i_k$th value of $X_k$, where $i_k \in \{1, \ldots, I_k\}$ and $k = 1, \ldots, m$. When all $m$ complete sets of one-dimensional marginal probabilities are known, the MLE of $\{p_{i_1 \ldots i_m}\}$ is the solution to the problem

$$\begin{array}{ll} \text{minimize} & \prod_{i_1,\ldots,i_m} p_{i_1 \ldots i_m}^{-n_{i_1 \ldots i_m}}, \\ \text{subject to} & \sum_{i_2,\ldots,i_m} p_{i_1 \ldots i_m} = p_{i_1 + \cdots +}, i_1 = 1, \ldots I_1, \\ & \ldots \\ & \sum_{i_1,\ldots,i_{m-1}} p_{i_1 \ldots i_m} = p_{+ \cdots + i_m}, i_m = 1, \ldots I_m. \end{array} \tag{2.6}$$

**Theorem 2.** ***Suppose all the marginal constraints in (2.6) are feasible. Then changing all "="s to "$\leq$"s in the constraints, (2.6) turns into a GP that is equivalent to (2.6) in estimating the probabilities of all nonempty cells.***

*Proof.* It is a straightforward extension of the proof to Situation (i) in Section 2.1, in which we can show the existence of a pivotal cell when any of the "$\leq$" constraints in the GP is not tight. □

In practice, it is rare that the one-dimensional marginal probabilities of all the corresponding variables $X_1, \ldots, X_m$ are known. However, even if one or more of them are missing from (2.6), the existence of the pivotal cell would not be affected. Hence, when any subset of the one-dimensional marginal probabilities is available, the MLE can be obtained via GP.

More generally, the above results are not restricted to one-dimensional marginal probabilities.

**Theorem 3.** ***Let $A_1, \ldots, A_a$ be disjoint subsets of $\{1, \ldots, m\}$, and $i(A_j^c) \equiv \{'i_k', k \notin A_j\}$ be a set of index labels. If any subset of the marginal probabilities $\sum_{i(A_1^c)} p_{i_1 \ldots i_m}$, ..., $\sum_{i(A_a^c)} p_{i_1 \ldots i_m}$ is known and these constraints are***

***feasible, then the problem of finding the MLE of cell probabilities is equivalent to the GP in estimating the probabilities of all nonempty cells, which is obtained by changing all "=" constraints to "$\leq$" constraints in the original optimization problem (with each equality constraint expressed in the least possible number of optimization variables).***

Take a simple example. If $A_1 = \{1,3\}$, $A_2 = \{2,4\}$, and $A_3 = \{5\}$ in a five-way contingency table, then $i(A_1^c) = \{'i_2',' i_4',' i_5'\}$, $i(A_2^c) = \{'i_1',' i_3',' i_5'\}$ and $i(A_3^c) = \{i_1',' i_2', i_3',' i_4'\}$. Theorem 3 tells that if any subset of $\{p_{i_1+i_3++}, p_{+i_2+i_4+}, p_{++++i_5}, \forall i_1, i_2, i_3, i_4, i_5\}$' is available, the MLE can be obtained via GP because the existence of the pivotal cell is guaranteed by the disjointness of $A_1$, $A_2$ and $A_3$.

We shall mention that cases of disjoint index sets, as described in Theorem 3, cover important applications, of which a recent example is given in Section 4 (i.e., application to a relatively new method of data collection, judgment post-stratification with multiple rankers). Note that raking is possible in more general cases. However, its application to high-dimensional tables is greatly limited by the fact that it cannot properly handle sparse tables while the problem of empty cells becomes prevalent as the dimension increases. By contrast, the GP-based approach does not have this problem and can greatly enhance the efficiency of estimation in the presence of empty cells, as indicated by the results reported in Table 1 (see Section 4.2).

# 3 Estimation under ordered marginal/conditional probabilities

In contingency tables, variables and their marginal or conditional probabilities can be naturally ordered (a motivating example is given in Section 5). Several authors developed algorithms for estimating multinomial parameters under order constraints on the probabilities of $Y$ or on the conditional probabilities of $Y$ given $X$ (e.g., Barlow et al. 1972, Jewell and Kalbfleisch 2004). However, the work is limited to one or two-way tables only, and these existing algorithms cannot be used with higher-way tables. As in Section 2, we begin with GP-based estimation in three-way tables, and then discuss the extension to higher-dimensional tables.

### 3.1 Three-way contingency tables

Suppose that the marginal probabilities $p_{i++}$s or conditional probabilities $p_{jk|i}$s are ordered in accordance with the values of $X_1$ indexed by $i$, $i \in \mathcal{I} = \{1, \ldots, r\}$. That is,

$$p_{1++} \leq p_{2++} \leq \cdots \leq p_{r++}, \tag{3.1}$$

or

$$p_{jk|1} \leq p_{jk|2} \leq \cdots \leq p_{jk|r} \quad \text{for } (j,k) \in \mathcal{C}. \tag{3.2}$$

In (3.2), $\mathcal{C}$ is assumed to be a true subset of $\mathcal{J} \times \mathcal{K}$ (i.e., $C \subset \mathcal{J} \times \mathcal{K}$), where $\mathcal{J} = \{1, \ldots, s\}$ and $\mathcal{K} = \{1, \ldots, t\}$ are index sets for $X_2$ and $X_3$. Note that if $\mathcal{C} = \mathcal{J} \times \mathcal{K}$, then (3.2) implies $p_{jk|1} = p_{jk|2} = \cdots = p_{jk|r}$ for all $j$ and $k$ because $\sum_{jk} p_{jk|i} = 1$ for all $i$, representing a trivial case. We further assume some very mild regularity conditions under each order constraint. That is, under (3.1), we assume that for each $i$, $\exists (j,k)$ s.t. $n_{ijk} > 0$; under (3.2), we assume that for each $i$, $\exists (j,k) \notin \mathcal{C}$. s.t., $n_{ijk} > 0$. Thus, under these regularity conditions, the likelihood function to be maximized here is given by (2.1), subject to the order constraints (3.1) or (3.2), plus the equality constraint $\sum_{ijk} p_{ijk} = 1$.

To do GP modeling, we re-parameterize the likelihood function (2.1) with $p_{jk|i}$ and $p_{i++}$, namely

$$\mathcal{L}(\{p_{ijk}\}) = \left( \prod_{i=1}^{r} \prod_{j=1}^{s} \prod_{k=1}^{t} p_{jk|i}^{n_{ijk}} \right) \left( \prod_{i=1}^{r} p_{i++}^{n_{i++}} \right). \tag{3.3}$$

The equality constraint $\sum_{ijk} p_{ijk} = 1$ is then equivalent to

$$\sum_{jk} p_{jk|i} = 1 \quad \text{and} \quad \sum_{i} p_{i++} = 1, \quad \text{for } i = 1, 2, \ldots, r. \tag{3.4}$$

**Theorem 4.** ***Under (3.1), the optimization problem is equivalent to the following GP with optimization variables*** $\{p_{jk|i}, p_{i++}\}$**,**

$$\begin{array}{lll} \textit{minimize} & \left(\prod_{i=1}^{r} \prod_{j=1}^{s} \prod_{k=1}^{t} p_{jk|i}^{-n_{ijk}}\right) \left(\prod_{i=1}^{r} p_{i++}^{-n_{i++}}\right) & \\ \textit{subject to} & \sum_{jk} p_{jk|i} \leq 1, & \textit{for } i = 1, \cdots, r, \\ & \sum_{i} p_{i++} \leq 1, & \\ & p_{1++} \leq p_{2++} \leq \cdots \leq p_{r++} & \end{array} \tag{3.5}$$

*Proof.* See Appendix B. □

**Theorem 5.** ***Under (3.2), the optimization problem is equivalent to the following GP with optimization variables*** $\{p_{jk|i}, p_{i++}\}$,

$$\begin{array}{lll} \textit{minimize} & \left(\prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{jk|i}^{-n_{ijk}}\right)\left(\prod_{i=1}^{r} p_{i++}^{-n_{i++}}\right) & \\ \textit{subject to} & \sum_{jk} p_{jk|i} \le 1, & \textit{for } i = 1, \cdots, r, \\ & \sum_i p_{i++} \le 1, & \\ & p_{jk|1} \le p_{jk|2} \le \cdots \le p_{jk|r}. & \textit{for } (j,k) \in \mathcal{C} \end{array} \tag{3.6}$$

*where* $\mathcal{C} \subset \mathcal{J} \times \mathcal{K}$.

*Proof.* See Appendix C. □

## 3.2 Extension to multi-way contingency tables

Suppose a subset of the $r-$dimensional marginal probabilities indexed by $i_1, \ldots, i_r$, say $p_{i_1\cdots i_r+\cdots+} = \sum_{i_{r+1},\ldots,i_m} p_{i_1\ldots i_m}$, $r < m$, is ordered in some way. Then, the MLE of $\{p_{i_1\ldots i_m}\}$ is the solution to

$$\begin{array}{ll} \text{minimize} & \prod_{i_1,\ldots,i_m} p_{i_1\ldots i_m}^{-n_{i_1\ldots i_m}}, \\ \text{subject to} & \sum_{i_1,\ldots,i_m} p_{i_1\ldots i_m} = 1, \\ & \text{and order constraints on } p_{i_1\cdots i_r+\cdots+}\text{s}. \end{array} \tag{3.7}$$

As in Section 3.1, we re-parameterize (3.7) using the marginal probabilities $p_{i_1\cdots i_r+\cdots+}$ and the corresponding conditional probabilities $p_{i_{r+1}\cdots i_m|i_1\cdots i_r} = p_{i_1\ldots i_m}/p_{i_1\cdots i_r+\cdots+}$, so that the optimization problem becomes

$$\begin{array}{ll} \text{minimize} & \prod_{i_1,\ldots,i_m} \left[p_{i_{r+1}\cdots i_m|i_1\cdots i_r} \cdot p_{i_1\cdots i_r+\cdots+}\right]^{-n_{i_1\ldots i_m}} \\ \text{subject to} & \sum_{i_{r+1},\ldots,i_m} p_{i_{r+1}\cdots i_m|i_1\cdots i_r} = 1, \\ & \sum_{i_1,\ldots,i_r} p_{i_1\cdots i_r+\cdots+} = 1, \text{ for every } (i_1, \ldots, i_r), \\ & \text{and order constraints on } p_{i_1\cdots i_r+\cdots+}\text{s}. \end{array} \tag{3.8}$$

Similarly, if a subset of the $r-$dimensional conditional probabilities $p_{i_{r+1}\cdots i_m|i_1\cdots i_r}$ is ordered according to the values of $X_1, \ldots, X_r$ in some way. Then, the MLE is the solution to

$$
\begin{array}{ll}
\text{minimize} & \prod_{i_1,\ldots,i_m}\left[p_{i_{r+1}\cdots i_m|i_1\cdots i_r}\cdot p_{i_1\cdots i_r+\cdots+}\right]^{-n_{i_1\ldots i_m}} \\
\text{subject to} & \sum_{i_{r+1},\ldots,i_m} p_{i_{r+1}\cdots i_m|i_1\cdots i_r}=1, \\
& \sum_{i_1,\ldots,i_r} p_{i_1\cdots i_r+\cdots+}=1,\ \text{for every } (i_1,\ldots,i_r), \\
& \text{and order constraints on } p_{i_{r+1}\cdots i_m|i_1\cdots i_r}\text{s} \\
& \text{for } (i_{r+1},\ldots,i_m)\in\mathcal{C}\ ,\ \mathcal{C}\subset\mathcal{I}_{r+1}\times\cdots\times\mathcal{I}_m\ .
\end{array}
\tag{3.9}
$$

Under very mild regularity conditions (similar to those described in Section 3.1), we introduce the following theorem.

**Theorem 6.** ***Changing all "="s to "$\le$"s in the constraints, (3.8)/(3.9) turns into a GP that is equivalent to (3.8)/(3.9).***

We rely on the same arguments made for Theorem 4/5 to relax the equality constraints in (3.8)/(3.9). The proof for Theorem 6 is omitted for brevity.

Finally, we note that the above extension is only for *one set* of $r-$dimensional marginal order constraints. It might not be applicable to multiple sets of multi-dimensional marginal order constraints. The difficulty in obtaining the MLE with $K$ sets of ordered marginal probabilities arises from expressing the likelihood function as a product of different sets of the marginal probabilities.

# 4 Application to JP-S with multiple rankers

## 4.1 Background

Judgment post-stratification (JP-S) sampling was introduced by MacEachern et al. (2004) as an alternative to ranked set sampling (Chen et al. 2006; McIntyre 2005) for studying the characteristics (such as mean, variance and distribution, etc) of some variable, say $Y$, which is expensive to quantify, but relatively cheap to rank by judgement. To obtain a JP-S sample, a simple random sample (SRS) of $n$ units is first drawn from a population and the value of $Y$ is recorded for each, denoted $y_i$, $1 \le i \le n$. For each measured unit $i$, an additional sample of size $H-1$ is chosen at random and the rank of $y_i$ among the $H$ units is assessed by one or more rankers using some inexpensive ranking method not requiring measurement of the $H-1$ units. MacEachern et al.

(2004), Stokes et al. (2007) and Wang et al. (2012) developed nonparametric mean estimators that make use of imprecise ranking information from multiple rankers.

Suppose there are $m$ rankers. For each measured unit $i$, the vector of ranks is denoted by $\mathbf{R}_i = (R_{i1}, \cdots, R_{im})$, where $R_{ij} \in \{1, \ldots, H\}$ is the rank assigned to $y_i$ by Ranker $j$. There are thus $H^m$ post-strata jointly grouped by the ranks $\mathbf{R} = (\mathbf{R}_i)_{i=1}^n$, forming an $m-$way contingency table. Let $\mathbf{r} = (r_1, \ldots, r_m)$. Let $\mathrm{PS}_\mathbf{r}$ denote the post-stratum with $\mathbf{R}_i = \mathbf{r}$; that is, the $\mathbf{r}$th cell in the contingency table that contains measured units whose ranks are given by $R_{i1} = r_1$,..., $R_{im} = r_m$. Further, let $\pi_\mathbf{r}$, $n_\mathbf{r}$, and $\bar{Y}_{[\mathbf{r}]}$ denote the probability, number and sample mean of observations falling in $\mathrm{PS}_\mathbf{r}$. Stokes et al. (2007) considered the stratified estimator of the mean $\mu$, $\hat{\mu} = \sum_\mathbf{r} \hat{\pi}_\mathbf{r}(\mathbf{n})\bar{Y}_{[\mathbf{r}]}$, where $\mathbf{n}$ is the random vector containing the counts of $Y$ in the $H^m$ post-strata, and $\hat{\pi}_\mathbf{r}(\cdot)$ is the estimate of the cell probability in $\mathrm{PS}_\mathbf{r}$. Here, the summation is over all $H^m$ realizations of the rank vector. However, empty cells frequently arise in JP-S samples with multiple rankers, whose sample means are not observable. For such cells, we proceed as if they do not exist. That is,

$$\hat{\mu} = \sum_{\mathbf{r} \notin \mathcal{E}} \hat{\pi}_\mathbf{r}(\mathbf{n})\bar{Y}_{[\mathbf{r}]}, \tag{4.1}$$

where $\mathcal{E}$ denotes the set of empty cells. Even under this naive treatment, the efficiency of $\hat{\mu}$ can be much higher than the corresponding SRS mean estimator $\bar{Y}$ (see Table 1), given the cell probabilities are estimated appropriately.

In practice, as long as appropriate blinding is employed in the ranking process, it is guaranteed that each of the $H$ ranks is equally likely for each ranker. Thus, it is reasonable to assume that for each ranker $j$ and a randomly selected unit $i$, $P(R_{ij} = 1) = \cdots = P(R_{ij} = H) = 1/H$. This induces the constraints that all one-dimensional marginal probabilities are uniform on $\{1, \ldots, H\}$. Stokes et al. (2007) used the raking method (Deming and Stephan 1940) to estimate $\pi_\mathbf{r}$s. Though easy to implement, the estimates from raking do not necessarily converge to the true MLE when empty cells exist, as mentioned in Section 2.

### 4.2 Simulation

We conducted a simulation study, where we simulated JP-S samples and applied the GP method to obtain $\hat{\pi}_{\mathbf{r}}$ in (4.1), to examine its performance in mean estimation.

We generated JP-S data with three rankers in the form of $(Y_i, R_{i1}, R_{i2}, R_{i3})_{i=1}^{n}$, where the rankers can be either equally effective or not. Here, we assume that Ranker $j$, $j = 1, 2, 3$, behaves as if he assesses the rank of $Y_i$ by assigning it the true rank that some ranking variable $X_{ij}$ has among its comparison group of size $H$. Further, we assume $X_{ij} = Y_i + e_{ij}$, where $Y_i \sim N(0, 1)$, $e_{ij} \sim N(0, \sigma_j^2)$ and $e_{ij}$ is independent of $Y_i$. If $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, all three rankers are equally effective. We set $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ to be $(1/2, 1/2, 1/2)$ or $(1/4, 1/2, 1/2)$, and $(H, n)$ to (3,12), (3,18), (5,20), and (5,30). Here, we chose $\sigma_j^2$ to be 1/2 or 1/4 so that each ranker has a reasonable accuracy in assessing ranks. Note that the ranking quality can be measured by $\rho$, the correlation between $X_{ij}$ and $Y_i$. When $\sigma_j^2 = 1/2$ or $1/4$, $\rho \approx 0.8$ or 0.9 (approximately 64% or 81% of the variability in $Y$ can be explained by latent $X_j$). Also, the number of rank strata $H$ needs to be small in practical situations because ranking a larger number of units cheaply by judgment with reasonable accuracy is difficult. That's why we chose $H = 3$ or 5. In addition, the total sample size $n$ should not be large because JP-S is typically applied to achieve cost efficiency when precisely measuring $Y$ is expensive. We set $n = \bar{n} \times H$ so that for each rank stratum, the average sample size $\bar{n}$ is very small (4 or 6). For each parameter combination, we generated 1000 data sets and computed the mean estimator (4.1), using the cell probabilities estimated from the raking and GP methods. When using the GP method, $\hat{\pi}_{stu}$s in (4.1) are obtained by maximizing the likelihood $\mathcal{L}\left(\{\pi_{stu}\}\right) = \prod_{s=1}^{H}\prod_{t=1}^{H}\prod_{u=1}^{H}\pi_{stu}^{n_{stu}}$ subject to the marginal constraints $\pi_{1++} = \ldots = \pi_{H++} = \pi_{+1+} = \ldots = \pi_{+H+} = \pi_{++1} = \ldots = \pi_{++H} = 1/H$, where $n_{stu}$ is the cell count in the post-stratum with $\mathbf{R} = (s, t, u)$.

Table 1 reports the approximate relative efficiency (RE) of the JP-S estimator $\hat{\mu}$ versus the SRS mean estimator $\bar{Y}$, as well as two types of measure ($U$ and $L_2$) for assessing estimation errors of cell probabilities. Here, RE is defined as the ratio of mean square errors (MSEs), namely

$$\mathrm{RE} \equiv \frac{\text{MSE}\left(\bar{Y}\right)}{\text{MSE}\left(\hat{\mu}\right)} = \frac{\frac{\sigma_Y^2}{n}}{\mathrm{E}(\hat{\mu} - \mu)^2},$$

where $\mathrm{E}(\hat{\mu}-\mu)^2$ is approximated by the average of $(\hat{\mu}-\mu)^2$ over the 1000 data sets. If RE $> 1$, $\hat{\mu}$ is more efficient than $\bar{Y}$ for estimating the population mean $\mu$ in terms of MSE. Also, the higher the RE value is, the more gain in efficiency is achieved from using $\hat{\mu}$ over $\bar{Y}$. We further define the uniform error $U \equiv \max_{s,t,u} \left|\pi_{stu} - \hat{\pi}_{stu}\right|$ and the $L_2$ error $L_2 \equiv \sum_{s,t,u}(\pi_{stu} - \hat{\pi}_{stu})^2$, to assess the performance in estimating cell probabilities. Table 1 shows that $\hat{\mu}$ based on GP provides much better mean estimation than $\bar{Y}$ in all cases we considered while raking does not. Also, it shows that GP has consistently higher relative efficiency than raking. This is because GP provides much better estimates of cell probabilities than raking, as shown by both $U$ and $L_2$ values in the table. The improvement from using GP is substantial, especially when $n/H$ is small.

Table 1: Simulated relative efficiencies of the JP-S estimator $\hat{\mu}$ versus the SRS mean estimator $\bar{Y}$, along with two types of estimation error $U$ and $L_2$ for estimating cell probabilities, are reported for the raking and GP methods, respectively.

| | | equally effective | | unequally effective | |
|---|---|---|---|---|---|
| ($H$,$n$) | | raking | GP | raking | GP |
| (3,12) | RE | 0.857 | 1.926 | 1.000 | 2.003 |
| | ($U$, $L_2$) | (0.256, 0.133) | (0.121, 0.047) | (0.258, 0.132) | (0.118, 0.043) |
| (3,18) | RE | 1.02 4 | 1.736 | 1.110 | 1.913 |
| | ($U$, $L_2$) | (0.217, 0.097) | (0.099, 0.034) | (0.224, 0.100) | (0.096, 0.032) |
| (5,20) | RE | 1.173 | 2.587 | 1.206 | 2.618 |
| | ($U$, $L_2$) | (0.159, 0.073) | (0.092, 0.033) | (0.162, 0.072) | (0.092, 0.037) |
| (5,30) | RE | 1.246 | 2.363 | 1.147 | 2.246 |
| | ($U$, $L_2$) | (0.130,0.053) | (0.076, 0.029) | (0.164, 0.072) | (0.091, 0.036) |

### 4.3 An Empirical Study: Tree Height Data

We consider a data set containing heights of 399 conifer trees (in feet) given in Chen et al. (2006). For illustrative purposes, we set our target parameter to be the mean height of the 399 trees. To generate a JP-S sample (with replacement) of size $n$ with $H$ strata, we repeated the following procedure $n$ times: first, randomly select a group of $H$ trees from the entire data set; Among the $H$ selected trees, ranking is done by three "perceived" rankers based on ranking variables $X_1$, $X_2$ and $X_3$, and then one of the $H$ trees is randomly selected to enter the JP-S sample. Again, for $j = 1, 2, 3$, $X_j = Y + \epsilon_j$, where $Y$ is the tree height, $\epsilon_j \overset{iid}{\sim} N(0, \sigma_j^2)$, and $Y$ and $\epsilon_j$ are independent. Here, we set $H = 5$, $n = 40$, $\sigma_1^2 = \sigma_2^2 = 5^2$ and $\sigma_3^2 = 10^2$ and generated 1000 JP-S samples.

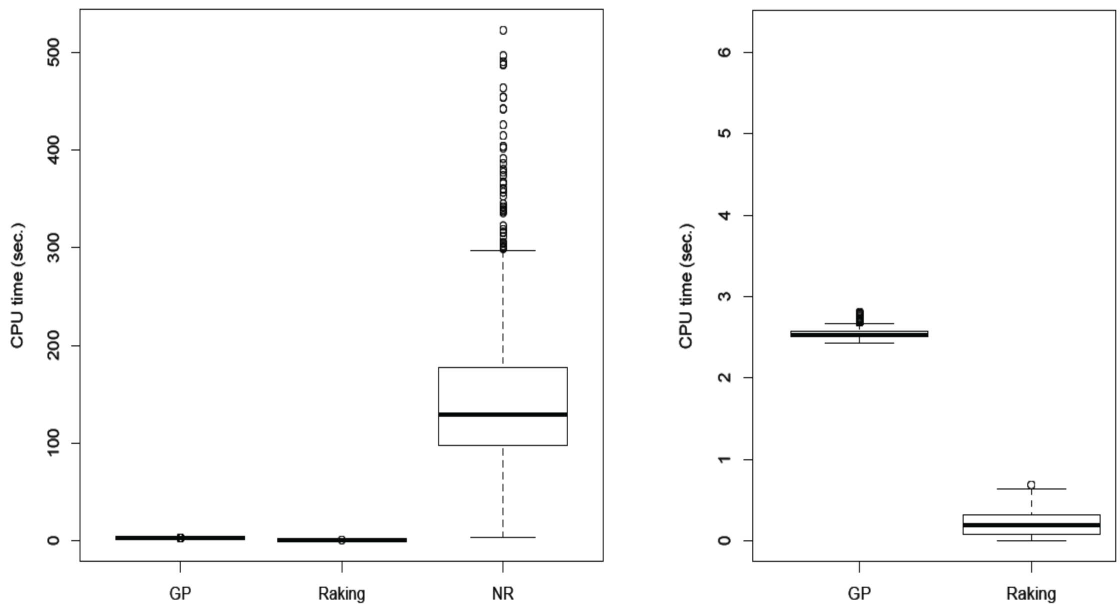


Figure 4.1: Tree height data example: the left panel reports the computing time (in seconds) for three methods, GP, raking and NR, using box plots based on 1000 JP-S samples; the right panel provides an enlarged view for comparing GP and Raking easily. Here, NR is implemented by the "quadprog" procedure in MATLAB.

The left panel of Figure 4.1 compares the computing time (in seconds) of three methods using box plots based on the 1000 JP-S samples, namely GP, raking, and NR. Both GP and raking are much faster than NR. Raking is actually slightly faster than GP on average; however, the small difference shown by the enlarged view in the right panel of Figure 4.1 may not have much practical impact, especially when compared with the huge difference from NR. The performance of raking in estimation is worse than GP: 1.93 vs. 2.24 in terms of relative efficiency. NR is very slow due to the extreme sparsity of the tables involved (i.e., lots of empty cells and small counts). For some of the 1000 JP-S samples, it does not seem to converge even after reaching the maximum 1000 iterations. The RE of NR is very poor and much lower than one, indicating it is even much worse than the SRS mean estimator $\bar{Y}$ in estimating the population mean.

# 5 Application to Income Data

In this section, we examine the existence of gender-based inequality in returns to education. Schooling is conceived as a process by which people acquire credentials. Traditional theories of stratification assume that education contributes to inequality by endowing people with different amounts of credentials (Blau and Duncan 1967). Others maintain that education is instrumental in determining one's life chances because it functions as a criterion which employers use to screen job applicants (e.g., Arrow 1973). Against this backdrop, many scholars have argued that returns for female education are low, because of lower earnings, lower labor force participation, and shorter working hours of women, compared with men (e.g., Moreh 1971; Belman and Heywood 1991; Orazem and Vodopivec 1995; Haider 2001). This suggests that education might not be a vehicle for accomplishing economic equality between men and women.

Our data are from the 2004 American National Election Studies (ANES) Survey, publicly accessible at the ANES website (http://www.electionstudies.org/). The sample consists of a cross-section of respondents that yielded 1,212 face-to-face interviews in the pre-election study prior to the presidential election, conducted September 7 through November 1, 2004. We grouped income data into 23 levels, and turned education into a dichotomous variable (i.e., 2 for "college graduates and above" and 1 for "others"). Three categorical variables, gender $(X_1)$, income $(X_2)$, and education $(X_3)$, were used in our analysis, so the observed table is $2 \times 23 \times 2$.

Given the total sample size $n$, the cell frequencies $\mathbf{n}_{ijk}$s are assumed to follow a multinomial distribution with probabilities $\{p_{ijk},\ i = '\text{m}'$ for men and $'\text{f}'$ for women, $j = 1, \ldots, 23,\ k = 1, 2\}$. Our main purpose is to test that at each education level, as the income level increases, the proportion of men increases so that the proportion of women decreases. More specifically, we test (i) $\mathcal{H}_{11}$, (ii) $\mathcal{H}_{21}$, and (iii) $\mathcal{H}_{11} \cap \mathcal{H}_{21}$, where $\mathcal{H}_{k1} : p_{\mathrm{m}|1k} \le p_{\mathrm{m}|2k} \le \cdots \le p_{\mathrm{m}|23,k}$ for $k = 1, 2$ and $p_{i|jk} = p_{ijk}/p_{+jk}$. The null hypothesis $\mathcal{H}_0$ for all three cases is that the $p_{m|jk}$'s are equal for all income levels.

Let $\mathcal{L}(\mathbf{p}) = \prod_{i=\mathrm{m,f}} \prod_{j=1}^{23} \prod_{k=1}^{2} p_{ijk}^{n_{ijk}}$. Under the order constraints above, the MLE of $p_{ijk}$s is the solution to

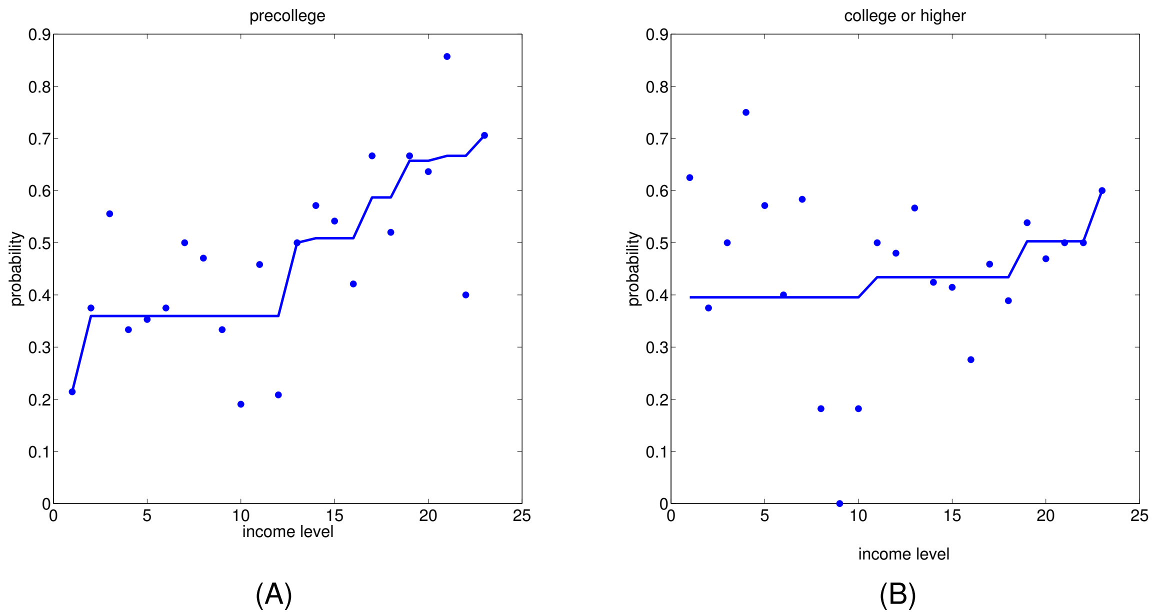


Figure 5.1: Income data example: the estimated proportions of men for education level (A) not college graduates $(k = 1)$ and (B) college or higher $(k = 2)$. The circle and the solid line represent the unconstrained and constrained MLE, respectively.

$$
\begin{aligned}
\text{minimize} \quad & \mathcal{L}(\mathbf{p}) = \prod_{j=1}^{23} \prod_{k=1}^{2} \left[ \left( \prod_{i=\mathrm{m,f}} p_{i|jk}^{-n_{ijk}} \right) p_{+jk}^{-n_{+jk}} \right] \\
\text{subject to} \quad & p_{\mathrm{m}|jk} + p_{f|jk} = 1 \quad \text{for all } j, k; \\
& \sum_{jk} p_{+jk} = 1; \\
& \text{under } \mathcal{H}_{11} : \quad p_{\mathrm{m}|11} \le p_{\mathrm{m}|21} \le \cdots \le p_{\mathrm{m}|23,1}; \\
& \text{under } \mathcal{H}_{21} : \quad p_{\mathrm{m}|12} \le p_{\mathrm{m}|22} \le \cdots \le p_{\mathrm{m}|23,2}; \\
& \text{under } \mathcal{H}_{11} \cap \mathcal{H}_{21} : p_{\mathrm{m}|1k} \le p_{\mathrm{m}|2k} \le \cdots \le p_{\mathrm{m}|23,k}, \quad \text{for } k = 1, 2.
\end{aligned}
$$

where $n_{+jk} = n_{\mathrm{m}jk} + n_{\mathrm{f}jk}$. In Figure 5.1, we plot the constrained and unconstrained MLEs of $p_{\mathrm{m}|jk}$ for $j = 1, \ldots, 23$ and $k = 1, 2$. The unconstrained MLE of $p_{m|jk}$ is simply $n_{mjk}/(n_{\mathrm{m}jk} + n_{\mathrm{f}jk})$, the corresponding sample proportion of men. The constrained MLE is obtained by our GP method where all "=" constraints are changed to "$\le$" in the optimization problems above.

To test the hypotheses above, we consider a likelihood ratio test (LRT), $\mathbf{T} = \max_{\mathbf{p} \in \Omega_0} \mathcal{L}(\mathbf{p}) /$

$\max_{\mathbf{p}\in\Omega} \mathcal{L}(\mathbf{p})$, where $\Omega_0$ and $\Omega$ are the parameter space under the null hypotheses $\mathcal{H}_0$s and the entire parameter space, respectively. Since the reference distribution of $T$ in each test is unknown, we computed the $p$-value using a permutation procedure where we permuted the levels of incomes. Let $\pi = (\pi(1), \pi(2), \ldots, \pi(23))$ be a permutation of $\{1, 2, \ldots, 23\}$, and $\Pi$ be the set of all possible $\pi$s. Let $\mathbf{T}(\pi)$ be the LRT statistic using the permuted sample $\mathbf{n}(\pi)$, where $\mathbf{n}(\pi) \equiv \left\{n_{i\pi(j)k},\ i = \mathrm{m}, \mathrm{f},\ j = 1, 2, \ldots, 23,\ k = 1, 2\right\}$. Then the $p$-value can be approximated by $\sum_{\pi\in\Pi} \mathrm{I}\left(\mathbf{T}_{\mathrm{obs}} > \mathbf{T}(\pi)\right)/|\Pi|$, where $|\Pi|$ is the total number of permutations and $\mathbf{T}_{\mathrm{obs}}$ is the observed test statistic. To avoid heavy computation, we computed the $p$-value using 1000 random permutations, not based on the enumeration $\Pi$. Here, one important step is to find the MLEs of the cell probabilities with/without the order constraints for each of the permutations, where the proposed GP method is applied.

The test results are reported in Table 2. Among non-college graduates, the proportion of men increases significantly as the income scale moves up. In contrast, this pattern was substantially less pronounced among college graduates, suggesting a much weaker level of gender-based income inequality. This suggests that education may help eliminate the income inequality between men and women.

Finally, we compare the GP method with the NR method in this example, where the raking method (i.e., the IPF algorithm) is not applicable. Both GP and NR gave essentially the same estimates in each permutation, which led to the same testing results. In terms of speed, NR was comparable or slightly faster than GP. This is perhaps due to the fact that the table does not contain any empty cell so that NR converged quickly. The computing time needed for any of the permutations is less than 0.2 second for either method, indicating no noticeable difference between the two. Here, we prefer using GP, because it is simple to implement and does not require us to provide (and test) starting points. Implementing the general-purpose NR method is not as easy and requires great effort as well as a good understanding about how NR works.

Table 2: The LRT test results for the inequality in income between men and women.

| Hypothesis | not college graduates ($\mathcal{H}_{11}$) | college and higher ($\mathcal{H}_{21}$) | both ($\mathcal{H}_{11} \cap \mathcal{H}_{21}$) |
|---|---|---|---|
| p-value | 0.0000 | 0.0816 | 0.0020 |

## 6 General use of GP in contingency tables

Although we focus on the constraints discussed in Sections 2 and 3, we show that our approach can be used to solve a wide variety of problems for contingency tables. To do so, we summarize below all the problems of contingency tables that can be handled by GP, including the relevant previous work.

P1 A discrete random variable $Y$ follows $\mathrm{Multinomial}(n, p_1, \cdots p_r)$, subject to constraints of the form $p_i \leq \alpha_{ij} p_j$ for $i \neq j$ and constants $\alpha_{ij} > 0$. This problem fits in constrained estimation of one-way contingency tables. An important special case, $p_{i+1} \leq p_i$ for $i = 1, \ldots, r-1$, had been solved by Robertson et al. (1988) through an iterative method based upon Fenchel duality. Bricker et al. (1997) considered the problem using a GP-based approach.

P2 Suppose $X$ and $Y$ index the rows and columns of a $r \times s$ contingency table. Let $p_{ij}$ be the probability that observations fall in the $(i,j)^{th}$ cell. To specify a positive (negative) association between $X$ and $Y$, the constraints in the local odds ratios $\theta_{ij}$ are given by

$$\theta_{ij} = \frac{p_{ij} p_{i+1,j+1}}{p_{i+1,j} p_{i,j+1}} \geq 1 \ \ (\leq 1)$$

for all $i$ and $j$. To estimate cell probabilities given positive/negative association, Bricker et al. (1997) reformulated the problem as a GP and then utilized GP software for the computation.

P3 Suppose that data are available on a discrete variable $Y$ at $r$ different values of a discrete explanatory variable $X$. Let $U_i = (Y | X = x_i)$, $1 \leq i \leq r$, each taking values in the same set of outcomes $o_1, \cdots, o_s$ and following $U_i \sim$ Multinomial$(n_i, p_{1|i}, \cdots, p_{s|i})$. For outcome $j$, $1 \leq j \leq s-1$, there exist order constraints $p_{j|1} \leq \cdots \leq p_{j|r}$. Jewell and Kalbfleisch (2004) proposed a modified pooled-adjacent violator (m-PAV) algorithm to solve this problem. Lim et al. (2009b) proposed a GP-based approach that is much faster than the m-PAV algorithm.

P4 In a multi-way contingency table, certain marginal probabilities are known. As discussed before, such problems were largely solved by the popular raking method. We have proposed a GP-based approach in Section 2. Though a recent application described in Section 4, i.e., estimating the population mean from JPS samples with multiple rankers, we have shown the superiority of the proposed method over the raking method.

P5 In a multi-way contingency table, marginal/conditional probabilities are ordered in accordance with one (or one set) of the discrete random variables. Note that P1 and P3 are both special cases of P5. We have discussed GP modeling for this type of problem in Section 3 and applied it to hypothesis testing using income data in Section 5.

Note that the earlier GP-based work (P1-P3) was restricted to low-dimension tables ($m = 1$ or 2) and the problems were solved on a case-by-case basis. In P4-P5, by contrast, we mainly focus on three-way tables, which are as common as two-way tables in applications. It is trivial to verify that GP modeling in P4-P5 can be done in all two-way tables. Also, we provide sufficient conditions for GP modeling for higher dimensional tables. Thus, by adding P4 and P5 into the pool, which are more general and harder in nature, the scope of GP applications in contingency tables has been greatly widened.

We note that a mixture of any two GP-feasible constraints is also GP feasible. Thus, the GP-based approach has the potential to handle a large variety of constraints beyond the five cases; for example, estimating cell probabilities in a three-way contingency table with ordered marginals on $X$ and positive association between $Y$ and $Z$. Also, it is possible to combine the two types of constraints discussed in Sections 2 and 3, and the resulting constrained optimization problem can be still solved via GP. For example, as long as the set of cell probabilities that appear in the known marginal constraints does not overlap with the set of cell probabilities that appear in the left hand sides of "$\leq$" constraints arising from order constraints, GP modeling can be done similarly. In addition, it can deal with any objective function in the form of monomials, including the likelihood function but not limited to it. For example, we could consider other types of estimators for P4-P5, such as the one minimizing discrimination information (Kullback, 1959) given by

$$\prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t}\left(\frac{n_{ijk}}{np_{ijk}}\right)^{\frac{n_{ijk}}{n}},$$

or from quasi-Bayes estimation (Good, 1965, 1967; Bishop et al., 1975) by maximizing

$$\prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{\frac{n_{ijk}+k}{n+tk}}.$$

## 7 Discussion

We have proposed a GP-based approach to obtain the MLEs of cell probabilities under two important types of constraints on marginals/conditionals, as discussed in Sections 2 and 3. Below we comment on how the GP method behaves under sparse contingency tables, which often pose great difficulty for the existing methods. In fact, the optimization problems in this paper are convex because the negative of the log likelihood functions are convex and the constraints involved are all linear in the optimization variables. As a result, the MLEs exist unless the constraints are inconsistent (i.e., no feasible values of the optimization variables satisfy all the constraints). Recall that the GP algorithm always finds the (true, globally) optimal solution for any feasible GP. Due to this attractive feature, the sparsity of contingency tables is not a concern for the GP method as long as the constraints are consistent. For example, our first application in Section 4 involves sparse contingency tables when the sample size $n$ is not large and/or the set size $H$ is not small. Specifically, in our first data example with $n = 40$, $H = 5$ and $m = 3$ (i.e., three rankers) in Section 4.3, the 40 observations in one JP-S sample fall into $5^3 = 125$ cells, and so the involved contingency tables are extremely sparse. The GP method worked well in this example while the other two methods, raking and NR, did not.

Geometric programming is able to provide a unified approach for various types of constrained estimation problems in contingency tables as summarized in Section 6. However, there are still open cases, for which we cannot guarantee solutions provided by the GP-based approach. They include three-way contingency tables with all three sets of two-dimensional marginals known, high-dimensional tables with known marginals where the index sets generally overlap, and tables with more than one set of $r-$dimensional marginal order constraints. For such cases, we could try to extend raking, which can be difficult, especially when empty cells exist, requiring future research. Or we could seek feasible algorithms on a per-table basis; for example, given an observed table, we conduct the GP relaxation, solve the corresponding GP, and then examine whether it yields a feasible solution before trying other numerical algorithms.

Although the main focus of this paper is on point estimation, we note that the covariance matrix of the constrained MLE, say Var$(\hat{\mathbf{p}})$, can be generally obtained through bootstrapping, no matter which type of constraints is involved. This entails (i) taking a sample of size $q$ ($q \leq n$) with replacement from the ungrouped data where each individual observation is recorded as one data line; (ii) calculating $\hat{\mathbf{p}}$ based on the generated sample using our proposed GP method; (iii) repeat the above steps $Q$ times. Then we calculate the sample covariance matrix based on $\hat{\mathbf{p}}^{(1)}, \cdots, \hat{\mathbf{p}}^{(Q)}$, which can be rescaled to approximate Var$(\hat{\mathbf{p}})$. To estimate the variance of the JP-S mean estimator in Section 4, we further calculate $\hat{\mu}^{(1)}, \cdots \hat{\mu}^{(Q)}$ using (4.1) based on the $Q$ samples, and then compute the sample variance to estimate Var$(\hat{\mu})$ in a similar manner. Also, we can construct confidence intervals for each cell probability or $\hat{\mu}$ from the corresponding quantiles of the $Q$ estimates. There are some other methods for variance estimation include random grouping, jackknife, etc; and we refer readers to Lohr (1999) for details.

## Appendix A: Proof of Theorem 1

Theorem 1 deals with three situations in three-way contingency tables, as described in Section 2.1.

**Proof for situation (i):**

The relaxed problem is a GP because (i) the objective is a monomial function that is a special case of posynomial functions; and (ii) all "$\leq$" constraints are posynomial functions of $p_{ijk}$s. Let $\{\bar{p}_{ijk}\}$ denote one optimal solution to the GP.

1. If $\{\bar{p}_{ijk}\}$ satisfies all the equality constraints in (2.2) , then $\{\bar{p}_{ijk}\}$ provides an optimal solution to (2.2) , including both empty and nonempty cells. The proof is done in this case.

2. Suppose there exists at least one constraint with "$<$" held at $\{\bar{p}_{ijk}\}$ so that $\{\bar{p}_{ijk}\}$ cannot provide a solution to the original optimization problem in (2.2). Without loss of generality, assume $\sum_{j,k} \bar{p}_{i^1jk} < p_{i^1++}$. Then $\sum_i \sum_{j,k} \bar{p}_{ijk} < \sum_i p_{i++} = 1$, which gives $\sum_j \sum_{i,k} \bar{p}_{ijk} < \sum_j p_{+j+}$ and $\sum_k \sum_{i,j} \bar{p}_{ijk} < \sum_k p_{++k}$. So there exist $j^1$ and $k^1$ s.t. $\sum_{i,k} \bar{p}_{ij^1k} < p_{+j^1+}$ and

$\sum_{i,j} \bar{p}_{ijk^1} < p_{++k^1}$. Then the cell $(i^1, j^1, k^1)$ must be empty, i.e. $n_{i^1j^1k^1} = 0$. Otherwise, we can construct $\{\tilde{p}_{ijk}\}$ by increasing $\bar{p}_{i^1j^1k^1}$ by a very small amount while keeping all the other $\bar{p}_{ijk}$s unchanged, so that the related three "$\leq$" constraints still hold. Doing so would increase the value of $\mathcal{L}$ if the cell $(i^1, j^1, k^1)$ is nonempty, which contradicts that $\{\bar{p}_{ijk}\}$ is an optimal solution to the GP.

3. Since the cell $(i^1, j^1, k^1)$ is empty, we can increase $\bar{p}_{i^1j^1k^1}$, while keeping the probabilities of all the other cells unchanged (without increasing the value of $\mathcal{L}$), until at least one of the three equality constraints $\sum_{j,k} p_{i^1jk} = p_{i^1++}$, $\sum_{i,k} p_{ij^1k} = p_{+j^1+}$ and $\sum_{i,j} p_{ijk^1} = p_{++k^1}$ holds (i.e., the one with the least slack). Denote the increased value by $\bar{p}^*_{i^1j^1k^1}$. Now update $\{\bar{p}_{ijk}\}$ by letting $\bar{p}_{i^1j^1k^1} = \bar{p}^*_{i^1j^1k^1}$. Note that the updated $\{\bar{p}_{ijk}\}$ not only provides an alternative optimal solution to the GP, but also changes at least one inequality constraint from the strict "<" sign to the "=" sign.

4. Repeat steps 1 through 3 until all the "<" constraints are changed to the corresponding "=" constraints.

After the above steps, the adjusted $\{\bar{p}_{ijk}\}$ provides an optimal solution to both the GP and the original optimization problem in (2.2). Note that all the adjustments are done to empty cells. Thus, the GP provides an optimal solution to all non-empty cells in (2.2).

**Proof for situation (ii):**

This proof is omitted for brevity because it is similar to the proof in situation (i) with very slight modifications.

**Proof for situation (iii):**

Let $\{\bar{p}_{ijk}\}$ denote one optimal solution to the GP.

1. If $\{\bar{p}_{ijk}\}$ satisfies all the equality constraints in (2.4) , then $\{\bar{p}_{ijk}\}$ provides an optimal solution to (2.4), including both empty and nonempty cells. This concludes the proof in this case.

2. Suppose there exists at least one constraint in which "<" holds at $\{\bar{p}_{ijk}\}$ so that $\{\bar{p}_{ijk}\}$ cannot provide a solution to the original optimization problem in (2.4). Without loss of generality, assume $\sum_k \bar{p}_{i^*j^*k} < p_{i^*j^*+}$. Then $\sum_j \sum_k \bar{p}_{i^*jk} < \sum_j p_{i^*j+} = \sum_k p_{i^*+k}$, which gives $\sum_k \sum_j \bar{p}_{i^*jk} < \sum_k p_{i^*+k}$. So there exists $k^*$ s.t. $\sum_j \bar{p}_{i^*jk^*} < p_{i^*+k^*}$. Then the cell $(i^*, j^*, k^*)$ must be empty, i.e. $n_{i^*j^*k^*} = 0$. Otherwise, we can construct $\{\tilde{p}_{ijk}\}$ by increasing $\bar{p}_{i^*j^*k^*}$ by a very small amount while keeping all the other $\bar{p}_{ijk}$s unchanged, so that the related three "$\leq$" constraints still hold. Doing so would increase the value of $\mathcal{L}$ if the cell $(i^*, j^*, k^*)$ is nonempty, which contradicts that $\{\bar{p}_{ijk}\}$ is an optimal solution to the GP.

3. Since the cell $(i^*, j^*, k^*)$ is empty, we can increase $\bar{p}_{i^*j^*k^*}$, while keeping the probabilities of all the other cells unchanged (without increasing the value of $\mathcal{L}$), until at least one of the three equality constraints $\sum_{j,k} p_{i^*jk} = p_{i^*++}$, $\sum_{i,k} p_{ij^*k} = p_{+j^*+}$ and $\sum_{i,j} p_{ijk^*} = p_{++k^*}$ holds (i.e., the one with the least slack). Denote the increased value as $\bar{p}^*_{i^*j^*k^*}$. Now update $\{\bar{p}_{ijk}\}$ by letting $\bar{p}_{i^*j^*k^*} = \bar{p}^*_{i^*j^*k^*}$. Note that the updated $\{\bar{p}_{ijk}\}$ not only provides an alternative optimal solution to the GP, but also changes at least one inequality constraint from the strict "<" sign to the "=" sign.

4. Repeat steps 1 through 3 until all the "<" constraints are changed to the corresponding "=" constraints.

After the above steps, the adjusted $\{\bar{p}_{ijk}\}$ provides an optimal solution to both the GP and the original optimization problem in (2.4). Note that all the adjustments are done to empty cells. Thus, the GP provides an optimal solution to all non-empty cells in (2.4).

## Appendix B: Proof of Theorem 4

We need to show that the GP (3.5) achieves the optimal value only when (3.4) is satisfied. Let $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ denote the optimal solution to the GP. Suppose there exists at least one of the terms among $\sum_{jk} p_{jk|i}$ and $\sum_i p_{i++}$ with "<1" held at $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ .

If for some $i$, say $i^*$, $\sum_{jk} \bar{p}_{jk|i^*} < 1$, then we can obtain $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ by setting $\tilde{p}_{j^*k^*|i^*} = \bar{p}_{j^*k^*|i^*} + 1 - \sum_{jk} \bar{p}_{jk|i^*}$, $\tilde{p}_{jk|i} = \bar{p}_{jk|i}$ for $i \neq i^*$ or $j \neq j^*$ or $k \neq k^*$, and $\tilde{p}_{i++} = \bar{p}_{i++}$ for all $i$, where $(j^*, k^*)$ satisfies $n_{i^*j^*k^*} > 0$. Note that $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ satisfies all the inequality constraints in (3.5) with $\sum_{jk} \tilde{p}_{jk|i^*} = 1$. Since $\tilde{p}_{j^*k^*|i^*} > \bar{p}_{j^*k^*|i^*}$, the objective function is smaller at $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$, indicating $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ is not optimal. Hence, $\sum_{jk} \bar{p}_{jk|i} = 1$ must hold for all $i$.

If $\sum_i \bar{p}_{i++} < 1$, then we can obtain $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ by setting $\tilde{p}_{r++} = \bar{p}_{r++} + 1 - \sum_i \bar{p}_{i++}$, $\tilde{p}_{i++} = \bar{p}_{i++}$ for $i \neq r$, and $\tilde{p}_{jk|i} = \bar{p}_{jk|i}$ for all $i, j, k$. Note that

$$\bar{p}_{1++} = \tilde{p}_{1++} \leq \bar{p}_{2++} = \tilde{p}_{2++} \leq \cdots \leq \bar{p}_{r++} < \tilde{p}_{r++}$$

and $\sum_i \tilde{p}_{i++} = 1$ so that $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ satisfies all the inequality constraints in (3.6). Since $\tilde{p}_{r++} > \bar{p}_{r++}$, the objective function is smaller at $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$, indicating $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ is not optimal. Hence, $\sum_i \bar{p}_{i++} = 1$ must hold.

## Appendix C: Proof of Theorem 5

Again, we need to show that the GP (3.6) achieves the optimal value only when (3.4) is satisfied. Since $\mathcal{C} \subset \mathcal{J} \times \mathcal{K}$, the complement $\bar{\mathcal{C}}$ is nonempty. Let $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ denote the optimal solution to the GP. Suppose there exists at least one of the terms among $\sum_{jk} p_{jk|i}$ and $\sum_i p_{i++}$ with "<1" held at $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ .

If for some $i$, say $i^*$, $\sum_{jk} \bar{p}_{jk|i^*} < 1$, then we can obtain $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ by setting $\tilde{p}_{j^*k^*|i^*} = \bar{p}_{j^*k^*|i^*} + 1 - \sum_{jk} \bar{p}_{jk|i^*}$, $\tilde{p}_{jk|i} = \bar{p}_{jk|i}$ for $i \neq i^*$ or $j \neq j^*$ or $k \neq k^*$, where $(j^*, k^*) \in \bar{\mathcal{C}}$ and $n_{i^*j^*k^*} > 0$, and $\tilde{p}_{i++} = \bar{p}_{i++}$ for all $i$. Note that $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ satisfies all the inequality constraints in (3.6) with $\sum_{jk} \tilde{p}_{jk|i^*} = 1$. Since $\tilde{p}_{j^*k^*|i^*} > \bar{p}_{j^*k^*|i^*}$, the objective function is smaller at $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$, indicating $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$ is not optimal. Hence, $\sum_{jk} \bar{p}_{jk|i} = 1$ must hold for all $i$.

If $\sum_i \bar{p}_{i++} < 1$, then a similar argument can be made as in the proof of Theorem 4 to show that $\sum_i \bar{p}_{i++} = 1$ must hold; that is, we can obtain $\{\tilde{p}_{jk|i}, \tilde{p}_{i++}\}$ by setting $\tilde{p}_{1++} = \bar{p}_{1++} + 1 - \sum_i \bar{p}_{i++}$, $\tilde{p}_{i++} = \bar{p}_{i++}$ for $i \neq 1$, and $\tilde{p}_{jk|i} = \bar{p}_{jk|i}$ for all $i, j, k$, which gives a smaller value of the objective function than $\{\bar{p}_{jk|i}, \bar{p}_{i++}\}$.

## References

Alldredge, J. R. and Armstrong, D. W. (1974). Maximum likelihood estimation for multinomial distribution using geometric programming. *Technometrics*, 16(4):585–587.

Arrow, K. J. (1973). *Information and Economic Behavior.* Stockholm: Federation of Swedish Industries.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

Belman, D. and Heywood, J. S. (1991). Sheepskin effects in the returns to education: An examination of women and minorities. *The Review of Economics and Statistics*, 73(4):720–724.

Bishop, J., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge.

Bishop, Y. M. and Fienberg, S. E. (1969). Incomplete two-dimensional contingency tables. *Biometrics*, 25(1):119–28.

Blau, P. M. and Duncan, O. D. (1967). *The American Occupational Structure.* New York: Wiley.

Boyd, S. P., Kim, S.-J., Patil, D. D., and Horowitz, M. A. (2005). Digital circuit optimization via geometric programming. *Operations Research*, 53(6):899–932.

Boyd, S. P., Kim, S.-J., Vandenberghe, L., and Hassibi, A. (2007). A tutorial on geometric programming. *Optimization and Engineering*, 8:67–127.

Bricker, D. L., Kortanek, K. O., and Xu, L. (1997). Maximum likelihood estimates with order restrictions on probabilities and odds ratios: a geometric programming approach. *Journal of Applied Mathematics and Decision Sciences*, 1(1):53–65.

Chen, Z., Bai, Z., and Sinha, B. K. (2006). *Ranked Set Sampling: Theory and Applications*. Springer, New York.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Duffin, R. J., Peterson, E. L., and Zener, C. (1967). *Geometric Programming - Theory and Applications*. John Wiley & Sons, New York.

Frey, J. and Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 63:769–789.

Frey, J., Ozturk, O., and Deshpande, J. V. (2007). Nonparametric tests for perfect judgment rankings. *Journal of the American Statistical Association*, 102:708–717.

Good, I. (1965). *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*. MIT Press, Cambridge.

Good, I. (1967). A bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society, Series B*, 29:399–431.

Haider, S. J. (2001). Earnings instability and earnings inequality of males in the united states: 1967-1991. *Journal of Labor Economics*, 19:799–836.

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55:179–188.

Jewell, N. P. and Kalbfleisch, J. D. (2004). Maximum likelihood estimation of ordered multinomial parameters. *Biostatistics*, 5(2):291–306.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lim, J., Kim, S. J., and Wang, X. (2009a). Estimation of stochastically ordered survival functions by geometric programming. *Journal of Computational and Graphical Statistics*, 18:978–994.

Lim, J., Wang, X., and Choi, W. (2009b). Maximum likelihood estimation of ordered multinomial probabilities by geometric programming. *Computational Statistics and Data Analysis*, 53:889–893.

Lim, J. and Won, J.-H. (2012). Estimation of convex ROC curve: ROC convex hull, conditional, and unconditional maximum likelihood estimators. *Machine Learning*, 88:433–444.

Little, R. J. A. and Wu, M. M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86:87–95.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. New York: Duxbury.

MacEachern, S. N., Stasny, E. A., and Wolfe, D. A. (2004). Judgment post-stratification with imprecise rankings. *Biometrics*, 60:207–215.

Mazumdar, M. and Jefferson, T. R. (1983). Maximum likelihood estimates for multinomial probabilities via geometric programming. *Biometrika*, 70(1):257–261.

McIntyre, G. A. (2005). A method for unbiased selective sampling, using ranked sets. *The American Statistician*, 59:230–232.

Moreh, J. (1971). Human capital and economic growth: United kingdom, 1951-1961. *Economic and Social Review*, October.

Nesterov, Y. and Nemirovsky, A. (1994). Interior-point polynomial methods in convex programming. In *Studies in Applied Mathematics*, volume 13, Philadelphia, PA. SIAM.

Orazem, P. F. and Vodopivec, M. (1995). Winners and losers in transition: Returns to education, experience, and gender in slovenia. *The World Bank Economic Review*, 9(2):201–230.

Pelz, W. and Good, I. J. (1986). Estimating probabilities from contingency tables when the marginal probabilities are known, by using additive objective functions. *The Statistician*, 35(1):45–50.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order-Restricted Inferences*. New York: Wiley.

Stephan, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13:166–178.

Stokes, S. L., Wang, X., and Chen, M. (2007). Judgment post-stratification with multiple rankers. *Journal of Statistical Theory and Applications*, 6:344–359.

Thompson, J. H. (1981). Convergence properties of the iterative 1980 census estimator. In *Proceedings of the Survey Research Methods Section*, pages 182–185, Washington, DC. The American Statistical Association.

Wang, X., Stokes, L., Lim, J., and Chen, M. (2006). Concomitants of multivariate order statistics with application to judgment poststratification. *Journal of the American Statistical Association*, 101(476):1693–1704.

Wang, X., Wang, K., and Lim, J. (2012). Isotonized CDF estimation from judgment post-stratification data with empty strata. *Biometrics*, 68:194–202.