Published in final edited form as:

J Am Stat Assoc. 2011 ; 106(496): 1394–1404. doi:10.1198/jasa.2011.tm10322.

Order restricted inference for multivariate binary data with application to toxicology

Ori Davidov

Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905 Israel,
davidov@stat.haifa.ac.il

Shyamal Peddada

Biostatistics Branch, National Institute of Environmental Health Sciences, Alexander Drive, RTP,
NC 27709, peddada@niehs.nih.gov

Abstract

In many applications researchers collect multivariate binary response data under two or more, naturally ordered, experimental conditions. In such situations one is often interested in using all binary outcomes simultaneously to detect an ordering among the experimental conditions. To make such comparisons we develop a general methodology for testing for the multivariate stochastic order between $K \geq 2$ multivariate binary distributions. The proposed test uses order restricted estimators which, according to our simulation study, are more efficient than the unrestricted estimators in terms of mean squared error. The power of the proposed test was compared with several alternative tests. These included procedures which combine individual univariate tests for order, such as union intersection tests and a Bonferroni based test. We also compared the proposed test with unrestricted Hotelling's T^2 type test. Our simulations suggest that the proposed method competes well with these alternatives. The gain in power is often substantial. The proposed methodology is illustrated by applying it to a two-year rodent cancer bioassay data obtained from the US National Toxicology Program (NTP). Supplemental materials are available online.

Keywords

dose response studies; multivariate binary data; order restricted statistical inference; stochastic order relations

1. Introduction

In this article we develop a methodology for analyzing multivariate binary (MVB) response data when the explanatory variable is an ordered categorical variable such as the dose of a drug or toxin, the level of a treatment, or the stage of a cancer.

The random vector $\mathbf{X} = (X_1, \dots, X_m)$ is said to follow a MVB distribution if its components X_i , $i = 1, \dots, m$ are binary random variables (RVs). Various probability models for MVB data have been proposed (cf., Joe 1997). MVB response data arise in diverse applications such as epidemiology (Ananth and Kantor 2004), dentistry (Burnside et al. 2007), genetics and animal breeding (Hoschele et al. 1986), psychology and economics (Maydeu-Olivares et al. 2006), clinical trials (Pocock et al. 1987) and quality of life studies (Ribaud and Thompson 2002). Our investigation is motivated by applications in toxicology.

A variety of statistical methods are available for analyzing unordered MVB response data depending on the scientific question of interest. One popular methodology is to model the

multivariate binary outcomes using a latent multivariate normal random vector and a set of fixed thresholds that induce binary random variables (Hoschele et al., 1986, Maydeu-Olivares et al. 2006). This approach is appealing because it is easy to interpret. Moreover the data can be analyzed by means of well studied hierarchical models. However, a problem with such a methodology is the arbitrary choice of various thresholds and the suitability of the latent variable model in the given application. Another family of methods is based on the generalized estimating equations (GEE) methodology of Liang and Zeger (1986) which fit marginal models to the data. One concern with this approach is the potential loss of power (cf. Fitzmaurice et al. 1996). Several modifications to the GEE approach have been proposed in the literature by considering multivariate logistic regression where, in addition to the marginal probabilities, marginal odds-ratios are modeled (e.g., Glonke and McCullach 1995 and Fitzmaurice and Laird 1993). Another reason for not considering GEE in the present article is that in our application, dealing with animal studies, one cannot be certain about the cause of death. The use of logistic regression models for modeling tumor incidence in terms of life time of an animal implicitly assumes that the tumor is non-lethal which may not be appropriate. Finally, resampling methods provide another strategy for testing for equality among $K - 2$ MVBs (Westfall and Young 1989, and Troendle 2005). One nice feature of resampling methods is that they naturally account for dependence without explicitly modeling it. Many of these methods and others are described in greater detail in the recent book by Song (2007).

Industries and various agencies, such as the Environmental Protection Agency (EPA), the Organization for Economic Cooperation and Development (OECD), the National Toxicology Program (NTP) etc., routinely conduct dose response studies in rodents to investigate the toxicity and/or carcinogenicity of various chemicals (Keenan et al. 2009). Toxicologists have long believed that some tumors may be associated or co-exist. Therefore while making decisions regarding the effect of a chemical on a particular tissue, the NTP may also consider information regarding an associated tumor. For example tumors of the pituitary gland and the mammary gland are believed to be associated through the Prolactin pathway (cf. McComb et al. 1984, NTP TR 527, 2005). It follows that a dose-related trend in both tumors may provide stronger evidence of carcinogenic activity. Often the NTP draws conclusions informally by examining the two tumors individually because formal methods for combining inferences from different tumors sites do not exist. Similar issues arise in clinical trials and drug safety studies. For example in a clinical trial one treatment may be preferred over another if it reduces the probability of observing a collection of pre-specified adverse events.

Researchers are often interested in demonstrating that the joint probability of events is increasing with dose. The above methods may not be directly applicable for making such comparisons. A simple approach to the problem, which is often used, is to conduct univariate analysis on each binary variable separately and then perform a Bonferroni correction to the resulting p -values. Since Bonferroni corrections do not require any information on the underlying dependence structure, the resulting methodology does not require any complicated modeling of the association among the response variables. A potential drawback of such a methodology is that it may lose power as the number of binary variables increases. Rather than testing for the equality of each marginal binomial probability separately, Agresti and Klingenberg (2005) and Klingenberg and Agresti (2006) introduced binary analogs to the classical Hotelling T^2 statistic. Although the methodology developed in these papers is ideal for comparing two or more experimental conditions in terms of all marginal probabilities simultaneously, they are not designed to test for an ordering among the experimental conditions. To the best of our knowledge the only methodology in the literature that addresses the problem considered in this paper is based on combining p -values from marginal tests in a permutation framework, cf. Finos et al. (2007)

and Pesarin (2001). It is important to emphasize that although the permutation framework honors the dependence among the binary RVs, the underlying test statistics ignores it. Moreover the framework described by Finos et al. (2007) does not use the order restrictions. For further discussion of this approach see Basso et al. (2009). Consequently, we conclude that there does not exist any procedure in the literature that addresses the problem considered in this paper in a principled way.

The main contributions of this paper are: (i) formulating the problem of comparing $K - 2$ MVBs across ordered experimental conditions as a problem of testing for the multivariate stochastic order among groups; and (ii) developing the methodology to carry out such tests when the data are high dimensional. The new methodology does not require any modelling of the dependence structure. Unlike the existing methodologies, the proposed formulation allows us not only to draw inferences on the joint probabilities, but more importantly it is possible to test for an ordering among the experimental conditions. Precise formulation of the problem is provided in Section 2. In Section 3 we develop estimation and testing procedures that explicitly account for multivariate stochastic order constraints and show, by simulations, that the proposed methodology improves the precision of the estimators and increases power for testing hypotheses. In Section 4 the proposed methodology is modified so that it can be applied to rodent cancer bioassay data where tumor incidence data are frequently incomplete. In that section we also provide results of a small simulation study which demonstrates that the proposed test maintains the Type I error rates. In Section 5 we apply the methodology to a data set obtained from the National Toxicology Program (NTP), where female rats were exposed to four different doses of Malachite Green Chloride, an antifungal agent used in the fish industry. We shall formally demonstrate that the joint incidence of mammary gland and pituitary gland tumors increases in a dose-related manner. NTP noted this outcome informally. A brief discussion and concluding remarks are provided in Section 6. Proofs are provided in an Appendix.

2. Formulation

Let $\mathbf{X}_i$ and $\mathbf{X}_j$ be m dimensional RVs. The RV $\mathbf{X}_i$ is said to be smaller than $\mathbf{X}_j$ in multivariate stochastic order, denoted $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$, if $\mathbb{P}(\mathbf{X}_i \in U) \leq \mathbb{P}(\mathbf{X}_j \in U)$ for all upper sets $U \in \mathbb{R}^m$ (Shaked and Shanthikumar, 2007). A set $U \in \mathbb{R}^m$ is called an upper set if $\mathbf{u} \in U$ implies that $\mathbf{v} \in U$ whenever $u_i \leq v_i, i = 1, \dots, m$. It follows that if $\mathbf{X}_i$ and $\mathbf{X}_j$ are MVBs then $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ if and only if for all upper sets $U \in \mathbb{R}^m$

$$\sum_{\mathbf{t} \in U} p_i(\mathbf{t}) \leq \sum_{\mathbf{t} \in U} p_j(\mathbf{t}), \quad (2.1)$$

where $p_i(\mathbf{t}) = \mathbb{P}(X_{i1}=t_1, \dots, X_{im}=t_m)$ for $t_l \in \{0, 1\}, l = 1, \dots, m$ and $p_j(\mathbf{t})$ is similarly defined. Furthermore, if there exists $U \in \mathbb{R}^m$ for which the inequality in (2.1) is strict then $\mathbf{X}_i$ is said to be strictly smaller than $\mathbf{X}_j$ in the multivariate stochastic order. This relationship is

denoted by $\mathbf{X}_i <_{st} \mathbf{X}_j$. Thus the event $\{\mathbf{X}_i \leq_{-st} \mathbf{X}_j\}$ is equivalent to $\{\mathbf{X}_i =_{st} \mathbf{X}_j\} \cup \{\mathbf{X}_i <_{st} \mathbf{X}_j\}$. This notion of multivariate stochastic order generalizes the univariate one and simply indicates that one random vector is more likely than another to take on larger values. In toxicology this statement implies that there is equal or greater chance of observing tumors in the high dose group (group j) than in the low dose group (group i).

Equation (2.1) implies that two MVBs are ordered provided a set of linear inequalities on their probabilities hold. It is easily verified that an upper set U generates a non-trivial inequality if the probabilities $\sum_{\mathbf{t} \in U} p_i(\mathbf{t})$ and $\sum_{\mathbf{t} \in U} p_j(\mathbf{t})$ are in $(0, 1)$. For example if $m = 1$

then X_i and X_j are binary RVs and $X_i \leq_{-st} X_j$ if and only if $p_i(1) \leq p_j(1)$. If $m = 2$, i.e., X_i and X_j are bivariate binary RVs, then it can be verified that $X_i \leq_{-st} X_j$ if and only if the following inequalities hold

$$\begin{aligned} p_i(1, 1) &\leq p_j(1, 1), \\ p_i(1, 0) + p_i(1, 1) &\leq p_j(1, 0) + p_j(1, 1), \\ p_i(0, 1) + p_i(1, 1) &\leq p_j(0, 1) + p_j(1, 1), \\ p_i(0, 1) + p_i(1, 0) + p_i(1, 1) &\leq p_j(0, 1) + p_j(1, 0) + p_j(1, 1). \end{aligned} \quad (2.2)$$

In principle one could enumerate all upper sets for any given m . Unfortunately, and as explained in Remark 2.1 below, the number of upper sets grows at a super-exponential rate.

Remark 2.1 Let N_m be the number of upper sets for which the inequality (2.1) is non-trivial. Let $S_m = \{(t_1, \dots, t_m) : t_i \in \{0, 1\}\}$ denote the support of an m dimensional MVB and let $\mathcal{U}(S_m)$ be the family of upper sets on S_m . Note that the sets $\emptyset$ and S_m are the only members of $\mathcal{U}(S_m)$ which guarantee a trivial inequality in (2.1) (i.e., $0 \leq 0$ and $1 \leq 1$). It is well known that there is a one to one correspondence between the elements of $\mathcal{U}(S_m)$ and the antichains in S_m . The antichains are those subsets of S_m in which no two elements are ordered (cf. Davey and Priestley, 2002). Furthermore the number of antichains is given by the so-called Dedekind numbers, D_m , $m \geq 0$. It follows that $N_m = D_{m+1} - 2$ where $m \geq 1$. The first six elements of this sequence are $N_1 = 1$, $N_2 = 4$, $N_3 = 18$, $N_4 = 166$, $N_5 = 7579$, $N_6 = 7828352$. The values of N_7 and N_8 are also known but N_9 is not (lower bounds and approximations are available for all m). For more detail see the ON-LINE ENCYCLOPEDIA OF INTEGER SEQUENCES at <http://www.research.att.com/~njas/sequences/A000372>.

In many applications m is large. For instance, in neurotoxicology a researcher may be interested in studying the effect of a toxin on the nervous system of an animal by measuring as many as 20 to 30 binary response variables, known as functional observational battery. Examples include, response to touch, tail pinch, and so forth (cf., Moser, 2000, Han et al., 2004). In such cases the enumeration of all upper sets is practically impossible. Thus, in order to proceed we must first reduce the dimension of the problem. This can be done with the help of Theorem 2.1 below.

Theorem 2.1 Suppose that $X_i \leq_{-st} X_j$ then $X_i \leq_{st} X_j$ if and only if $X_{i_l} \leq_{st} X_{j_l}$ for some $1 \leq l \leq m$.

Theorem 2.1 shows that if X_i is equal or smaller stochastically than X_j then a strict ordering between X_i and X_j holds provided at least one of the marginal distributions is strictly ordered. Note that in many dose-response studies, such as those conducted by toxicologists at NTP, the investigators believe apriori that the chemicals studied are harmful and that their effects get worse as the dose increases (without such prior belief agencies such as the NTP would not spend millions of dollars to conduct the two-year bioassay). Thus the assumption that $X_i \leq_{-st} X_j$, whenever $i \leq j$, is a reasonable one in our application.

Although the focus of this paper is binary outcomes Theorem 2.1 is valid for discrete and continuous variables and combinations thereof. Consequently it can be used to develop methods for a much broader collection of applications than described here. For example, in early phase clinical trials one expects the number and severity of adverse effects to have a monotone trend in dose. In such studies researchers collect continuous outcomes (e.g., blood chemistry) ordinal outcomes (e.g., quality of life measures) and binary outcomes (e.g., existence of specified adverse effects).

Remark 2.2 The relationship between the ordering of multivariate distributions and their marginals was noted by Finos et al. (2007, 2008), who quoted Baccelli and Makowski (1989). Giancristofaro and Bonnini (2008) showed that if the marginal distributions are supported on $\{0, \dots, s\}$ then they can be ordered by ordering their first $s - 1$ moments. Klingenberg et al. (2009) formulated a theorem similar to Theorem 2.1 for the case of multivariate discrete distributions with finite support. Their proof establishes the orthant orders. More specifically they show that $\mathbb{P}(\mathbf{X}_i \leq x) \geq \mathbb{P}(\mathbf{X}_j \leq x)$ and $\mathbb{P}(\mathbf{X}_i > x) \leq \mathbb{P}(\mathbf{X}_j > x)$ for all $\mathbf{x}$, with a strict inequality for some $\mathbf{x}$. The first inequality is equivalent to the lower orthant order denoted $\mathbf{X}_i \leq_{-lo} \mathbf{X}_j$ and the second to the upper orthant order denoted by $\mathbf{X}_i \leq_{-uo} \mathbf{X}_j$. It can be shown, see the Supplementary Materials A, that if $\mathbf{X}_i$ and $\mathbf{X}_j$ are MVBs then $\mathbf{X}_i \leq_{-uo} \mathbf{X}_j$ and $\mathbf{X}_i \leq_{-lo} \mathbf{X}_j$ are jointly equivalent to $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ if and only if $m = 2$. Theorem 2.1, however, establishes a strict multivariate stochastic order, which is stronger than the orthant orders, moreover it does so for arbitrary RVs. For further comparison of the multivariate stochastic order and the orthant orders see Shaked and Shanthikumar (2007).

Remark 2.3 Note that, in general, the univariate stochastic ordering of marginal distributions does not imply the multivariate stochastic ordering of the joint distribution. Consider, for example, two bivariate binary distributions. The Bahadur (1961) representation of their probability function is

$$p_i(t_1, t_2) = \prod_{j=1}^2 p_{ij}^{t_j} (1 - p_{ij})^{1-t_j} \left\{ 1 + \rho_i \frac{t_1 - p_{i1}}{\sqrt{p_{i1}(1 - p_{i1})}} \frac{t_2 - p_{i2}}{\sqrt{p_{i2}(1 - p_{i2})}} \right\}$$

where $i = 1, 2$ and $(t_1; t_2) \in \{0, 1\}^2$. In our notation $p_{i1} = p_i(1, 1) + p_i(1, 0)$ is the probability of observing the first outcome in the i^{th} group. All other probabilities are similarly defined. The quantity ρ_i is the Pearson correlation coefficient in the i^{th} group. Suppose that $(p_{11}, p_{12}, \rho_1) = (.1, .1, .4)$ and $(p_{21}, p_{22}, \rho_2) = (.15, .15, 0)$. Clearly $p_{1j} < p_{2j}$ for $j = 1, 2$ so the marginal distributions are ordered. However $p_1(1, 1) = .0460 > .0225 = p_2(1, 1)$ so the joint distributions are not ordered by the multivariate stochastic order (see the first inequality in (2.2)). A simple calculation shows that $p_1(1, 1) > p_2(1, 1)$ if $\rho_2 > 0.185$ in which case they will be ordered.

Theorem 2.1 is particularly useful in the context of MVBs because $\mathbf{X}_{il} \leq_{st} \mathbf{X}_{jl}$ means that the marginal probabilities, i.e., the probabilities of observing specific tumors are ordered. Thus

$$p_{il} = \sum_{t \in T_l} p_i(t) \leq \sum_{t \in T_l} p_j(t) = p_{jl} \quad (2.3)$$

for all $1 \leq l \leq m$, where $T_l = \{t : t_l = 1\}$. Hence only m inequalities rather than N_m need be considered (see Remark 2.1). For example in the bivariate case we need to consider only the second and third inequalities in (2.2). For large m the reduction in the number of constraints to be considered is extraordinary and therefore ordered inference on MVBs of any dimension is practical. This reduction in dimensionality will be exploited in this paper. It is particularly useful in the context of testing problems where the hypothesis

$$H_0: \mathbf{X}_i \leq_{st} \mathbf{X}_j \text{ versus } H_1: \mathbf{X}_i <_{st} \mathbf{X}_j \quad (2.4)$$

can be reformulated using Theorem 2.1 as

$$H'_0: \cap_{l=1}^m [X_{il} =_{st} X_{jl}] \text{ versus } H'_1: \cup_{l=1}^m [X_{il} <_{st} X_{jl}]. \quad (2.5)$$

The null in (2.5) was first referred to as simultaneous marginal homogeneity (SMH) by Agresti and Klingenber (2005).

The above formulation can be extended from 2 to $K > 2$ groups as follows. Let $\mathbf{P} = (\mathbf{P}_1, \dots, \mathbf{P}_K)^T$ where $\mathbf{P}_j = (P_{j1}, \dots, P_{jm})^T$ are the marginal probabilities for the j^{th} group as defined by (2.3). With this notation and using (2.5) we formulate our hypothesis as

$$H_0: \mathbf{P} \in \mathcal{P}_0 \text{ versus } H_1: \mathbf{P} \in \mathcal{P}_1 \setminus \mathcal{P}_0 \quad (2.6)$$

where $\mathcal{P}_0 = \{\mathbf{P}: \mathbf{P}_1 = \dots = \mathbf{P}_K\}$, $\mathcal{P}_1 = \{\mathbf{P}: \mathbf{P}_1 \leq \dots \leq \mathbf{P}_K\}$, and the inequalities defining $\mathcal{P}_1$ are interpreted component wise. Note that $\mathcal{P}_0$ is a linear space and $\mathcal{P}_1$ is a convex polyhedral cone, often called the simple order cone.

Remark 2.4 It is important to note that the null hypothesis in (2.6) is valid provided the marginal distributions are equal, i.e., if $X_{j1} =_{st} \dots =_{st} X_{Kj}$ for $j = 1, \dots, m$. Thus (2.6) can be tested without assuming that the multivariate distributions of $\mathbf{X}_1, \dots, \mathbf{X}_K$ are all equal. If we are willing to assume apriori (as we do here) that $\mathbf{X}_1 \leq_{st} \dots \leq_{st} \mathbf{X}_K$ then rejecting the null in (2.6) we may conclude that the K MVBs are (strictly) ordered by the multivariate stochastic order. Otherwise rejecting the null in (2.6) leads to the conclusion that some univariate distributions are strictly ordered by the univariate stochastic order.

For completeness we note that Sampson and Whitaker (1989) as well as Lucas and Wright (1991) studied stochastically ordered distributions. They focused on bivariate discrete distributions, i.e., on $I \times J$ cross classified data. Sampson and Whitaker (1989) developed *min* – *max* formulas for estimation whereas Lucas and Wright (1991) extended Robertson and Wright's (1981) projection formulas. Testing procedures with chi-bar limits were developed. These authors did not formulate their procedure in terms of the parameters $\mathbf{P}_1, \dots, \mathbf{P}_K$ as in (2.6) but rather in terms of $\mathbf{p}_1, \dots, \mathbf{p}_K$ where $\mathbf{p}_j$ is the $2^m - 1$ dimensional probability mass function of $\mathbf{X}_j$. It is not surprising, in light of Remark 2.1, that they noted the difficulty of extending their methodology to high dimensional data.

Following Pesarin (2001) in a series of papers Finos et al. (2007, 2008) developed union intersection type tests that do not explicitly incorporate any ordering. Their tests are based on combining p-values using prespecified p-value combination functions from pairwise one-sided tests and are implemented using permutation methods. Klingenber et al. (2009) proposed a two sample test for multivariate ordinal data. Their test is equivalent to the testing of homogeneity in a cumulative logit model and is implemented by assigning scores to the ordinal outcomes, p-values were computed by permutation methods. The above body of work is summarized by Basso et al. (2009).

In contrast with these approaches our proposed methodology explicitly incorporates: (1) the multivariate stochastic order constraints, i.e., that $\mathbf{P} \in \mathcal{P}_0$ under the null and that $\mathbf{P} \in \mathcal{P}_1 \setminus \mathcal{P}_0$ under the alternative; and (2) the dependence among all binary RVs. Thus our methodology addresses the ordering in (2.6) in a very natural way. In addition our methodology is completely model free and does not require the specification of any extraneous quantities such as scores or p-value combination functions. It is also important to note that unlike the above mentioned methods our methodology can be applied when the data are not exchangeable and resampling methods fail. Animal studies subject to censoring by survival are an important example of such a situation (cf. Peddada et al. 2007).

3. Inference for ordered multivariate binary data

3.1. The proposed methodology

Recall that $\mathbf{P}_i = (P_{i1}, \dots, P_{im})^T$ is the vector of marginal probabilities for the i^{th} group. For $i = 1, \dots, K$ and $j = 1, \dots, m$ we may estimate P_{ij} by its MLE $\hat{P}_{ij} = n_i^{-1} \sum_{t \in T_j} N_i(t)$ where T_j is defined in (2.3), $N_i(t)$ is the random number of subjects in the i^{th} group with outcome t , and $n_i = \sum_t N_i(t)$ is the size of the i^{th} group. It follows that $\text{Var}(\hat{\mathbf{P}}_i) = n_i^{-1} \Sigma_i$ where the variance matrix Σ_i is a function of $\mathbf{p}_i$ the probability mass function of $\mathbf{X}_i$.

We also note that under the null in (2.6) $\mathbf{P}$ is estimated by $\bar{\mathbf{P}}_1 = (\bar{P}_1, \dots, \bar{P}_K)$ where

$$\bar{P}_1 = \dots = \bar{P}_K = \frac{\sum_{i=1}^K n_i \hat{P}_i}{\sum_{i=1}^K n_i}.$$

We propose to estimate $\mathbf{P}$ under the alternative in (2.6) by

$$\tilde{\mathbf{P}} = \Pi_{\hat{\Sigma}}(\hat{\mathbf{P}} | \mathcal{D}_1) = \arg \min \left\{ (\hat{\mathbf{P}} - \mathbf{P})^T \hat{\Sigma}^{-1} (\hat{\mathbf{P}} - \mathbf{P}) : \mathbf{P} \in \mathcal{D}_1 \right\}. \quad (3.1)$$

Thus $\tilde{\mathbf{P}}$ is the projection of $\hat{\mathbf{P}}$ onto the convex cone $\mathcal{D}_1$ with respect to the matrix $\hat{\Sigma}$ defined as

$$\hat{\Sigma} = \widehat{\text{Var}} \left(\sqrt{n} (\hat{\mathbf{P}} - \mathbf{P}) \right) = \text{BlockDiag} \left((n/n_1) \hat{\Sigma}_1, \dots, (n/n_K) \hat{\Sigma}_K \right) \quad (3.2)$$

where $\hat{\Sigma}_i$ are the usual plug in estimators, i.e., the (j, j') element of $\hat{\Sigma}_i$ is given by $n_i^{-1} \sum_{a=1}^{n_i} (X_{ij}(a) - N_{ij}/n_i) (X_{ij'}(a) - N_{ij'}/n_i)$ where $X_{ij}(a) = 1$ if the a^{th} individual in the i^{th} group experienced outcome j and $X_{ij}(a) = 0$ otherwise, and N_{ij} is the number of individuals with the j^{th} outcome in the i^{th} group.

Note that the projection (3.1) accounts for the dependence among the components of $\hat{\mathbf{P}}$. Projection estimators of the type (3.1) have been proposed for a wide variety of constrained inference problems, see Silvapulle and Sen (2005) for a recent summary. The value of $\tilde{\mathbf{P}}$ can be readily obtained by quadratic programming.

Proposition 3.1. Let $\mathbf{P} \in \mathcal{D}_1$. If $\min(n_1, \dots, n_K) \rightarrow \infty$ then $\tilde{\mathbf{P}}$ is consistent for $\mathbf{P}$.

Note that while $\hat{\mathbf{P}}$ is unbiased the constrained estimator $\tilde{\mathbf{P}}$ is biased. This is typical of constrained estimators. However the bias is negligible.

We propose testing (2.6) using the statistic D_n given by

$$D_n = n \left\{ \|\hat{\mathbf{P}} - \Pi_{\hat{\Sigma}_0}(\hat{\mathbf{P}} | \mathcal{D}_0)\|_{\hat{\Sigma}_0}^2 - \|\hat{\mathbf{P}} - \Pi_{\hat{\Sigma}_0}(\hat{\mathbf{P}} | \mathcal{D}_1)\|_{\hat{\Sigma}_0}^2 \right\} \quad (3.3)$$

where $\|x\|_V \equiv (x^T V^{-1} x)^{1/2}$ and $\hat{\Sigma}_0$ estimates the scaled variance of $\hat{\mathbf{P}}$ under the null. The matrix Σ_0 may be estimated by $\hat{\Sigma}_0 = \text{BlockDiag} \left((n/n_1) \bar{\Sigma}, \dots, (n/n_K) \bar{\Sigma} \right)$ where

$\bar{\Sigma} = \text{Var}(\bar{P}_i) = n^{-1} \sum_{i=1}^K n_i \hat{\Sigma}_i$. This estimator is appropriate if we further assume that the data are exchangeable under the null. If only marginal homogeneity is assumed then we use (3.2) instead of $\hat{\Sigma}_0$. Further observe that (3.3) simplifies to $n \left\{ (\hat{P} - \bar{P})^T \hat{\Sigma}_0^{-1} (\hat{P} - \bar{P}) - (\hat{P} - \bar{P})^T \hat{\Sigma}_0^{-1} (\hat{P} - \bar{P}) \right\}$ and has the form of a likelihood ratio statistic (LRT) for an asymptotically normal random vector (Silvapulle and Sen, 2005). In fact, D_n may be viewed as test based on a likelihood derived from a normal approximation to the joint distribution of the marginal means. We now provide the limiting distribution of the proposed test statistic under the null hypothesis.

Theorem 3.1. If $\in \mathcal{P}_0$ and $n_i \rightarrow \infty$ with $n_i/n \rightarrow \lambda_i > 0$ then we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_n \geq x) = \sum_{k=1}^{m(K-1)} w_k(\Sigma_0, R) \mathbb{P}(\chi_k^2 \geq x) \quad (3.4)$$

for some weights $w_k(\Sigma_0, R)$, $k = 1, \dots, m(K-1)$ discussed below.

Thus (3.3) follows a chi-bar distribution. The weights $w_k(\Sigma_0, R)$, $k = 1, \dots, m(K-1)$ are non negative and sum to unity. Note that the weights depend on the unknown, but estimable, variance of the unconstrained estimator under the null and the restriction matrix R defined in the Appendix. In general the weights can not be computed in closed form whenever $m(K-1) > 5$. However the weights can be estimated by Monte-Carlo methods. For a thorough discussion on calculation of the weights and the simulation of chi-bar distributions see Section 3.5 in Silvapulle and Sen (2005) or Davidov et al. (2010). To circumvent these problems we test (2.4) using standard bootstrap methods which we have found to be more powerful than permutation based methods. We have also found that the (approximate) parametric bootstrap works well. The latter methodology compares the observed value of D_n to simulated values from the (approximate) large sample distribution of D_n given in the proof of Theorem 3.1. One advantage of the (approximate) parametric bootstrap is that it can be applied in situations where the data is not exchangeable and standard resampling methods do not hold. One such situation is described in the next section.

When dealing with binary variables with small “success” probabilities, it is inevitable to observe small or even zero number of “successes” in the sample. This would potentially lead to singular or near singular variance matrices. To deal with this problem, as done in Teoh et al. (2008), we applied the classical ridge-regression based methodology by adding a small positive constant δ to the diagonal elements of $\hat{\Sigma}$ and/or $\hat{\Sigma}_0$. This modification guarantees their invertibility and is a strategy that is widely used in a variety of contexts, for examples see Montgomery et al. (2001) or Tusher et al. (2001). Although there exists considerable amount of literature on the choice of delta, there is no consensus. Fortunately, in our specific situation, the exact value of the ridge parameter is not critical. Firstly, our simulation studies suggest that the impact on the estimators is minimal for a reasonably wide range of values experimented in our simulation study. Secondly, and more interestingly, it can be shown that when the ridge parameter goes to zero, we obtain the restricted estimator and as it goes to infinity we actually recover the pooled adjacent violator algorithm (PAVA).

Finally we remark that in the special case of $m = 1$ the multivariate methodology developed in this paper generalizes the well-known procedures for the univariate case. The constrained estimator (3.1) reduces to the isotonic regression problem

$\bar{P} = \arg \min \left\{ \sum_{i=1}^K \hat{w}_i (\hat{p}_i - p_i)^2 : p_1 \leq \dots \leq p_K \right\}$ where $\hat{w}_i = n_i / \text{Var}(\hat{p}_i)$. It is well known (e.g., Silvapulle and Sen 2005 page 53) that the MLE in this case is

$\mathbf{P}^* = \arg \min \left\{ \sum_{i=1}^K n_i (\hat{p}_i - p_i)^2 : p_1 \leq \dots \leq p_K \right\}$. Thus $\bar{\mathbf{P}}$ may be different from the MLE in some cases. Nevertheless it can be shown that $\mathbf{P}^* = \bar{\mathbf{P}} + o_p(1)$. Moreover, if we recompute the weights, w_j , when samples are merged in the pool adjacent violator algorithm (PAVA) then $\mathbf{P}^* = \bar{\mathbf{P}}$. Further note that when $m = 1$ the limiting distribution of D_n is again a chi-bar square. In this case the limit (3.4) simplifies considerably when $n_1 = \dots = n_K$ where the weight are given in closed form by $w_k = K!/(2^K k!(K-k)!)$ where $k = 1, \dots, K$.

3.2. A simulation study

We conducted a simulation study to evaluate the performance of the proposed methodology. The design of our simulation study is as follows. We consider four test groups: a control and three “dose” groups (low, medium and high) and sample sizes of $n = 50$ subjects per group as in our application. MVB data was generated using the method of Emrich and Piedmonte (1991), see also Qaqish (2003). Various patterns for the marginal probabilities and correlation coefficients and the dimension of the problem were investigated. A subset of these patterns, reported in this paper, are summarized in Table 1. Here we report only on trivariate MVBs, i.e., $m = 3$. Results for $m = 5$ are included in our supplementary materials.

We considered three different patterns of correlation structures among the three binary outcomes considered in our simulation study. These include; (i) Independent structure, (ii) Serial correlation structure where the correlation coefficient between components 1 and 2 and between 2 and 3 were each 0.5 while the correlation coefficient between components 1 and 3 was 0.25, and (iii) Intraclass correlation structure with a correlation coefficient of 0.5 between any pair of components. All simulations are based on 1000 simulation runs and the null distribution of the test statistics were derived empirically using 1000 bootstrap samples.

Estimates of the total mean squared error (TMSE) for the unrestricted and the order restricted estimators are provided in Table 2. Note that TMSE is defined as the sum of the MSEs of all components.

The TMSE is computed using $\hat{\Sigma}_0$, the pooled variance estimator, under the null and using the unpooled estimator $\hat{\Sigma}$ under the alternative. In most cases we observe that the proposed estimator achieves a substantial reduction in the TMSE compared to the unconstrained estimator. In some cases almost 40% reduction was noted.

We compared the estimated Type I error rate and power of the proposed test statistic (D_n) with six competitors. These are: (1) The unconstrained Hotelling T^2 type procedure (U_n)

based on the statistic $n(\hat{\mathbf{P}} - \bar{\mathbf{P}})^T \hat{\Sigma}_0^{-1} (\hat{\mathbf{P}} - \bar{\mathbf{P}})$ which is equivalent to the test of Agresti and Klingenberg (2005). Note that U_n is not designed for ordered alternatives. It is important to note that Theorem 2.1 can be used to develop tests for (2.4) which combine m marginal tests for order in various ways. These include Union–Intersection type tests (UIT) (see Sen (2007) for a review of UIT methodology for ordered alternatives) and a Bonferroni test. In particular in our simulation study we consider: (2) the max–test (M_n); (3) the sum–test (S_n); (4) a Bonferroni test (B_n); (5) a Williams type test (W_n); and (6) the test of Finos et al. (2007) (F_n). All of the above mentioned tests, except for F_n , are new. The max–test is based on the statistic $M_n = \max_{1 \leq j \leq m} D_n^{[j]}$ where $D_n^{[j]}$ is the statistic (3.3) based only on the j^{th} tumor as explained at the end of the previous section. Note that $D_n^{[j]}$ is the usual univariate likelihood ratio test for order. The sum–statistic is given by $S_n = \sum_{j=1}^m D_n^{[j]}$. The Bonferroni test (B_n) is based on the marginal p–values of the tests $D_n^{[j]}$, $j = 1, \dots, m$ while adjusting for

the number of tests. The Williams (1971) type test is based on the statistic

$W_n = \max_{1 \leq j \leq m} \sqrt{n} (P_{kj}^* - P_{1j}^*) / \sqrt{V_j^*}$ where $P_{kj}^* - P_{1j}^*$ is the difference between the constrained MLEs for the first and last groups and V_j^* is the estimated variance. Due to their simplicity Williams type tests are common in order restricted inference. The F_n test is somewhat more involved and is based on the combination of p-values. The test statistic is given by $F_n = \Psi(Q_1, \dots, Q_m)$ where Q_j is the p-value associated with the j^{th} marginal test and Ψ is some p-value combination function. In our simulations we used Fisher's product method. For more details see Finos et al. (2007).

Results of Type I error rates and powers of the above test statistics are provided in Table 3.

We observe that all test procedures attain the nominal error rate of 0.05, although U_n , the Hotelling's type T^2 statistic was slightly liberal. In most cases considered in our simulation study, the proposed test D_n enjoys higher power than the alternative procedures when the data are correlated. In particular, D_n has higher power than F_n (Finos et al., 2007). The gain in power in the settings we examined is usually substantial. For example, in some cases D_n has double the power of F_n . As one would expect, the test statistic F_n and the union intersection type tests M_n , S_n , W_n are often comparable, and at times slightly superior, to D_n in the independent case. Even then the gain in power by the best test (which varies by pattern) relative to D_n is minimal. We note that the Bonferroni test B_n never performed better than D_n .

To summarize D_n is comparable to the best test under independence and much more powerful when the binary outcomes are dependent. We note that dependent outcomes are common in a variety of applications such as the one considered here. Thus we conclude that the overall performance of D_n in practical settings is likely to be superior to any of the alternatives considered.

4. Application to analysis of rodent cancer bioassay

4.1. Modification of the proposed methodology

Toxicologists routinely conduct dose response studies to evaluate the carcinogenicity of a chemical by examining multiple organs for tumors. For instance, a typical NTP bioassay consists of exposing male and female rats and mice to one of several doses of a chemical of interest. Fifty animals are randomly assigned to each of the four dose groups, one of which is a control, and followed for two years. At the end of the study all animals are necropsied and various organs are evaluated for tumors. The NTP performs a statistical test to detect any dose-related trend in tumor incidence for each tumor type. Based on the statistical significance of the trend test and other biological evidence the NTP declares if a chemical is a carcinogen or not. This weight of evidence approach for declaring the carcinogenicity of a chemical includes, among several considerations, informal use of information on multiple tumor sites (cf. Keenan, et al., 2009). A formal procedure that jointly evaluates a dose-related trend in two or more associated tumors would be of great value to toxicologists. Thus, the methodology proposed in this paper will be also useful in the analysis of rodent cancer bioassay data.

An important feature of rodent cancer bioassay data is that not all animals survive until the end of the study. Consequently, animals that die without developing the tumor of interest produce a censored observation. A further complication is that survival rates may vary by dose group and thus observations at different dose levels are not exchangeable even under the null. This means that the missingness is not ignorable, i.e., the data is not missing at random and standard methods would fail. To deal with such censored data, toxicologists use

a widely accepted sample size correction known as the Poly-3 correction (Bailer and Portier, 1988). Recall first that if the data are complete, i.e., it is not subject to censoring by survival,

the unconstrained estimator of P_{ij} equals N_{ij}/n_i where $N_{ij} = \sum_{t \in T_j} N_i(t)$ is the number of animal with the j^{th} tumor type in the i^{th} group and n_i is the total number of animals in the i^{th} group. Unfortunately, and as noted above, we do not observe these quantities. Instead P_{ij} is commonly estimated by

$$\widehat{P}_{ij}^o = \frac{O_{ij}}{W_{ij}} \quad (4.1)$$

where O_{ij} is the observed number of animals in the i^{th} group with tumor j and W_{ij} is the effective sample size in the i^{th} group with respect to the j^{th} tumor. The effective sample size is given by $W_{ij} = \sum_{a=1}^{n_i} w_{ij}(a)$ where $w_{ij}(a)$ is a weight associated with the a^{th} animal.

Bailer and Portier (1988) suggested using $w_{ij}(a) = 1$ if the a^{th} animal in the i^{th} dose group survived to the end of the study or died with the j^{th} tumor. Otherwise $w_{ij}(a) = (t_i(a)/T)^3$ where $t_i(a)$ is the time of death for the a^{th} animal in the i^{th} group and T is the duration of the study. Their correction, derived assuming a Weibull distribution for tumor onset and survival, is known as the Poly-3 correction and is widely accepted by toxicologists (Peddada et al., 2005). It is important to realize that the key idea underlying Bailer and Portier approach is that with properly chosen weights (4.1) consistently estimates the true value of P_{ij} . We will assume so henceforth.

We now proceed as in Section 3 but instead of using $\widehat{P}$ we base our inferences on $\widehat{P}^o$. First observe that the consistency of (4.1) together with the multivariate central limit theorem imply that $\sqrt{n_i}(\widehat{P}_i^o - P_i) \Rightarrow N_m(0, \Sigma_i^o)$ for $i = 1, \dots, K$ where $\Sigma_i^o = n_i \text{Var}(\widehat{P}_i^o)$. Using a standard series approximations we find that

$$\text{Cov}(\widehat{P}_{ij}^o, \widehat{P}_{i'j'}^o) = \frac{\mathbb{E}(O_{ij})\mathbb{E}(O_{i'j'})}{\mathbb{E}(W_{ij})\mathbb{E}(W_{i'j'})} \times \left\{ \frac{\text{Cov}(O_{ij}, O_{i'j'})}{\mathbb{E}(O_{ij})\mathbb{E}(O_{i'j'})} - \frac{\text{Cov}(O_{ij}, W_{i'j'})}{\mathbb{E}(O_{ij})\mathbb{E}(W_{i'j'})} - \frac{\text{Cov}(O_{i'j'}, W_{ij})}{\mathbb{E}(O_{i'j'})\mathbb{E}(W_{ij})} + \frac{\text{Cov}(W_{ij}, W_{i'j'})}{\mathbb{E}(W_{ij})\mathbb{E}(W_{i'j'})} \right\} + O(n_i^{-1}) \quad (4.2)$$

Note that if the data are fully observed then $W_{ij} = n_i$ for all j and (4.2) reduces to the variance in the fully observed case. Clearly the variance matrix Σ_i^o may be consistently estimated from the data. The expectations $\mathbb{E}(O_{ij})$ and $\mathbb{E}(W_{ij})$ are estimated by O_{ij} and W_{ij} . The covariances terms appearing in (4.2) may be estimated empirically. For example, $\widehat{\text{Cov}}(O_{ij}, W_{i'j'}) = n_i \sum_{a=1}^{n_i} (o_{ij}(a) - O_{ij}/n_i)(w_{i'j'}(a) - W_{i'j'}/n_i)$ where $o_{ij}(a)$ is a binary RV indicating whether the a^{th} animal in the i^{th} dose group developed the j^{th} tumor type prior to its death or sacrifice. The other covariances are estimated similarly.

Following the developments in Section 3 define

$$\overline{P}_i^o = \frac{\sum_{i=1}^K n_i \widehat{P}_i^o}{\sum_{i=1}^K n_i} \quad \text{for } i=1, \dots, K$$

$$\hat{P}^o = \Pi_{\hat{\Sigma}^o}(\hat{P}^o | \mathcal{D}^1) = \arg \min \left\{ (\hat{P}^o - P)^T (\hat{\Sigma}^o)^{-1} (\hat{P}^o - P) : P \in \mathcal{D}_1 \right\}$$

to be the estimators of P under the null and alternative respectively. Here

$\hat{\Sigma}^o = \text{BlockDiag}((n/n_1)\hat{\Sigma}_1, \dots, (n/n_K)\hat{\Sigma}_K)$ is the estimated variance of $\hat{P}^o$. Further let

$$D_n^o = n \left\{ \|\hat{P}^o - \Pi_{\hat{\Sigma}_0^o}(\hat{P}^o | \mathcal{D}_0)\|_{\hat{\Sigma}_0^o}^2 - \|\hat{P}^o - \Pi_{\hat{\Sigma}_0^o}(\hat{P}^o | \mathcal{D}_1)\|_{\hat{\Sigma}_0^o}^2 \right\}$$

be the test statistic for (2.6).

It can be shown (by repeating the analysis in Proposition 3.1 and Theorem 3.1), that $\hat{P}^o$ is consistent and that D_n^o follows a chi-bar distribution with weights $w_k(\hat{\Sigma}_0^o, R)$, $k = 1, \dots, m(K - 1)$. The test D_n^o may be applied using the (approximate) parametric bootstrap as explained earlier.

Note that if survival rates differ among groups then $\hat{\Sigma}_i^o$ will be generally different than $\hat{\Sigma}_j^o$. Nevertheless our numerical calculations show that it is often better to estimate Σ^o by

$\text{BlockDiag}((n/n_1)\bar{\Sigma}^o, \dots, (n/n_K)\bar{\Sigma}^o)$ where $\bar{\Sigma}^o = n^{-1} \sum_{i=1}^K n_i \hat{\Sigma}_i^o$ is the pooled variance estimator.

4.2. A simulation study

We conducted a simulation study to evaluate the performance of the proposed modification of the methodology. The design of our simulation study mimics the NTP's 2-year cancer bioassay and is similar to those conducted previously (e.g. Peddada et al., 2005). We consider four test groups: a control and three dose groups (low, medium and high) and sample sizes of 50 subjects per group.

Our simulation study honors the prevailing latent Weibull structure and the Poly-3 adjustment (cf. Bailer and Portier 1988) assumed in rodent cancer bioassay studies. Thus, corresponding to each subject we generated an m variate normal random vector with mean 0 and covariance matrix $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = 1$ if $i = j$ and $\sigma_{ij} = \rho$ for $i \neq j$. Let $Q_{\lambda_{ij}; \kappa_{ij}}(\cdot)$ denote the quantile function of a Weibull random variable with survival function given by $\exp(-\lambda_{ij} t^{\kappa_{ij}})$. Note that λ_{ij} is the scale parameter and κ_{ij} is the shape parameter. Then $T_{ij}(a) = Q_{\lambda_{ij}; \kappa_{ij}}(\Phi(Z(a)))$ where $a = 1, \dots, n_i$ denotes the latent time to onset of the j^{th} tumor for the a^{th} subject in the i^{th} group. By the above construction the components of $T_i(a) = (T_{i1}(a), \dots, T_{im}(a))$ are correlated Weibull random variables. For each animal we generated an additional independent Weibull random variable $\Delta_j(a)$ which represents the time to death. The scale and shape parameters for Δ_j are denoted by β_j and γ_j , respectively. Let S be the duration of the study, which in the present case is taken to be 2 years then the a^{th} animal in the i^{th} group develops the j^{th} tumor if $T_{ij}(a) \leq \min(\Delta_j(a), S)$. For $i = 1, \dots, K; j = 1, \dots, m$ and $a = 1, \dots, n_i$ define the binary RVs $X_{ij}(a) = \mathbb{I}(T_{ij}(a) \leq \min(\Delta_j(a), S))$. Thus for each animal we generate a vector of binary RVs and a time of death $\min(\Delta_i(a), S)$. This is exactly the information analyzed by the NTP.

We evaluated the performance of the test procedure D_n^o for detecting an ordering among m correlated tumors in a 2-year cancer bioassay in terms of the Type I error. All our

simulations are based on 1000 random realizations and in each case the critical values for D_n° were generated using 1000 parametric bootstrap samples.

Although a broad range of patterns were considered, in this paper we summarize a small subset of results we obtained. Various patterns of animal survival and tumor rates are provided in Table 4.

The patterns of survival rates considered in this simulation study ranged from no survival difference among the dose groups to somewhat large survival differences between the control and the high dose group. Similarly, we considered a wide range of tumor rate patterns, ranging from rare tumors (about 1% rate) to somewhat common tumors (40%). We considered three correlated tumors with correlation coefficient ρ of patterns of 0.1 and 0.9. Corresponding to each of the above types patterns, we considered three different patterns of shape parameters for tumor incidence curve, namely, 1.5, 3 and 6 as commonly done in this area (e.g. Peddada et al., 2005). Results of the simulations are summarized in Table 5.

When the data conforms to the Poly-3 model, i.e., the shape parameter governing the Weibull distribution for time to tumor is equal to 3, the proposed test maintains the nominal level. The nominal level is also maintained when the data generating model is misspecified to be Poly-6 model. However, it can be liberal if the data are generated by (the misspecified) Poly-1.5 model and the survival differences among groups are large.

5. Illustration: an analysis of NTP's Malachite Green Chloride data

As part of their toxicity and carcinogenicity evaluation of chemicals that humans are exposed to, the NTP recently conducted a 2-year cancer bioassay on Malachite Green Chloride an antifungal agent used in the Fish industry (NTP TR-527, 2005). There were three dose groups of the chemical (100, 300 or 600 parts per million (ppm)) and a control group (0 ppm), with 48 female rats randomly assigned to each of the groups. The animals were followed for two years. The incidence of mammary gland adenomas and pituitary gland adeno-carcinomas in female rats are reported in Table 6. Neither of these tumor types displayed a significant dose-related trend according to the NTP's Poly-3 trend test (with p-values 0.113 and 0.162, respectively). However, increased incidence of mammary gland carcinoma in the highest dose group, was one of the reasons for NTP to conclude that there was an equivocal evidence of carcinogenic activity in female rats. It is interesting to note that the NTP highlighted the fact that there was a significant increase in the incidence rates of pituitary gland adeno-carcinoma in female rats from control group to 100 ppm group (p-value=0.014). They also noted that the incidences in the higher dose groups, namely, 300 ppm and 600 ppm, were well above the historical control range. Note that in the case of the mammary gland adenoma the observed rate at the 300 ppm dose group violates the hypothesized monotone order, while in the case of pituitary gland adeno-carcinoma the violation may be in the 100 ppm dose group.

It is well documented in the literature that the pituitary gland tumors may be associated with mammary gland tumors via the prolactin pathway (cf. McComb, et al. 1984, and NTP TR 527, 2005). In view of this known correlation between the two tumors, we applied the proposed methodology to this data. The order-restricted estimates under stochastically increasing order for the two tumor types are provided in Table 6.

Applying our methodology we find that the p-value of the test for the monotonic bivariate stochastic order based on 20,000 bootstrap samples is 0.0253. Thus we may conclude that the two tumors exhibit jointly a dose-related trend and that tumor incidence is ordered by the multivariate stochastic order as a function of dose.

We performed a simulation study to evaluate the Type I error rate of our procedure, under the null hypothesis that two correlated tumors had incidence rates of 0.06 and 0.73 for each of the four dose groups. These incidence rates are obtained by averaging the observed incidence rates of mammary gland and pituitary tumors in the current example. We used the same survival rates in our simulation study as in the NTP study. As in the simulation section, we generated tumor incidence and survival data using independent Weibull random variables. We considered three different patterns of shape parameters for tumor incidence, namely, 1.5, 3 and 6 and two patterns of correlation coefficients, namely, 0.1 and 0.9. We estimated the Type I errors using 1000 bootstraps and 1000 simulation runs. In every case the Type I error was close to the nominal level of 0.05. Thus we believe that the significant result obtained in the above example is not a false positive.

6. Discussion

Comparison of two or more experimental conditions on the basis of multivariate binary response data is an important problem. Much of the existing methodology is designed to compare experimental conditions that are not ordered. However, in many instances the experimenter is interested in comparing treatment conditions which are ordered as in dose-response studies, time-course experiments, tumor stages in oncology and so forth. In such situations the existing methodologies are not helpful in answering questions of practical interest such as Do a set of responses change (increase or decrease) monotonically with dose of a chemical?

In this article we extend the existing literature in two important directions. First, we formulate the scientific question as a multivariate stochastic ordering problem among multiple groups. By doing so, we provide a framework for comparing two or more experimental conditions in terms of all m binary response variables jointly. In dose-response studies, such as those conducted by toxicologists, a fundamental assumption is that the response is monotonic in dose. We demonstrate that this assumption reduces the number of order restrictions substantially and allows the development of practical estimation and testing procedures for MVB data of any dimension. As a consequence, true multivariate inference on the hypothesized ordering of the experimental conditions is enabled. Furthermore, our methodology does not require the specification or verification of any complicated modeling assumptions regarding the dependence among the binary variables. Our simulations suggest that the constrained estimators proposed in this article enjoy a substantial increase in efficiency (i.e., reduction in total mean squared error) relative to the standard unconstrained estimators. Similarly, the proposed test for the multivariate stochastic ordering is more powerful than the unrestricted Hotelling type T^2 test (cf. Agresti and Klingenberg, 2005) and the Bonferroni correction based methodology in the cases we investigated. When the binary outcomes are dependent the proposed test is more powerful than a host of UIT type tests. In particular it improves considerably over Finos et al. (2007), currently the only available test in the literature. The gain in power is substantial in many cases. When the binary outcomes are independent we found that there usually exists a test (among the tests M_n , S_n , W_n or F_n) which has comparable power to D_n . However the best test among the above tests varies according to the underlying pattern.

The proposed methodology is modified to handle the competing risks problem commonly encountered in animal cancer bioassays. Essentially, in long term animal cancer studies an animal may die before the end of the study without developing tumors. In such cases, the estimates of tumor incidence rate have to be corrected for the length of time animals were on the study. As routinely done by toxicologists in the case of univariate binary response, we use the Poly-3 methodology to estimate tumor incidence rates. As expected, the proposed methodology performs accurately (i.e., achieves the nominal Type I error rate) when there

are no survival differences among dose groups. It also performs accurately (or conservatively) if the shape parameter in the Weibull model is 3 or more. But it may be liberal when the true shape parameter is 1.5, the tumors are somewhat common and there are survival differences among groups. We note that in this situation the Poly-3 correction commonly used by toxicologists is incorrect. Hence it seems that inflated Type I error rates are intrinsic to the methodology used by toxicologists when the model is misspecified. If it is of paramount importance, the Type I error can be controlled by using larger values of the ridging constant δ . Note that the constrained estimator is consistent for all values of $\delta > 0$. We have experimented with a variety of values for δ and found that by choosing $\delta = 1$ the Type I error rates are maintained over a broad range of patterns. However a price in terms of power will be paid. Perhaps, borrowing ideas from the well established ridge regression theory, one may empirically derive the ridge constant δ using cross-validation or empirical Bayes methodology.

Finally the framework provided in this paper may be generalized in several directions. Our focus had been on multivariate binary distributions but the theory can be extended to multivariate ordinal and even multivariate continuous data. The method can also be extended to situations where a scientist is interested in comparing experimental conditions after adjusting for confounders or covariates. Such extensions may be feasible by couching the proposed methodology in generalized multivariate linear models, although this may not be trivial. In some applications, such as reproductive toxicology, one may observe multivariate binary responses over time or generations. Such experimental designs induce correlations not only among the binary responses at given time (generation) but also over time.

7. Supplementary materials

We have included two appendices as supplementary materials. In the first, Appendix A, the relationship between the orthant orders and the multivariate stochastic order, discussed in Remark 2.2, is investigated and clarified. In the second, Appendix B, additional simulations for higher dimensional data, i.e., $m = 5$, are reported on along with simulations results for some situations where the ordering assumptions are violated.

8. Proofs

Proof of Theorem 2.1:

Proof. Theorem 2.1 states that: (i) If $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ and $X_{il} \leq_{st} X_{jl}$ for some $1 \leq l \leq m$ then $\mathbf{X}_i \leq_{st} \mathbf{X}_j$; and (ii) If $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ and $\mathbf{X}_i \leq_{st} \mathbf{X}_j$ then $X_{il} \leq_{st} X_{jl}$ for some $1 \leq l \leq m$. We start with (i). Taken together the relationships $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ and $X_{il} \leq_{st} X_{jl}$ for some $1 \leq l \leq m$ imply that $\mathbb{E}[\phi(\mathbf{X}_i)] \leq \mathbb{E}[\phi(\mathbf{X}_j)]$ for all increasing functions $\phi: \mathbb{R}^m \rightarrow \mathbb{R}$ with a strict inequality for some functions ϕ of the form $\phi(x_1, \dots, x_m) = \phi(x_l)$. It follows by definition that $\mathbf{X}_i \leq_{st} \mathbf{X}_j$ establishing (i). Now note that Theorem 6.B.16 in Shaked and Shanthikumar (2007) implies that statement in Theorem 2.1 is invariant to monotone transformations. Therefore we may assume that the RVs $\mathbf{X}_i$ and $\mathbf{X}_j$ are nonnegative. Suppose by contradiction that (ii) does not hold. The relationship $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ implies that $X_{il} \leq_{-st} X_{jl}$ for all $1 \leq l \leq m$. Since $X_{il} \leq_{st} X_{jl}$ does not hold then we must have $X_{il} =_{st} X_{jl}$ for all $1 \leq l \leq m$. Let $\theta \in \mathbb{R}^m$ be nonnegative, i.e., $\theta_i \geq 0$ for all $1 \leq i \leq m$. Since $\mathbf{X}_i \leq_{-st} \mathbf{X}_j$ we have $\theta^T \mathbf{X}_i \leq_{-st} \theta^T \mathbf{X}_j$ but the equalities $X_{il} =_{st} X_{jl}$ imply that $\theta^T \mathbf{X}_i =_{st} \theta^T \mathbf{X}_j$. The positivity of $\mathbf{X}_i$ and $\mathbf{X}_j$ guarantees the existence of their Laplace transforms, so $L_{\mathbf{X}_i}(\theta) = \mathbb{E}[\exp(-\theta^T \mathbf{X}_i)] = \mathbb{E}[\exp(-\theta^T \mathbf{X}_j)] = L_{\mathbf{X}_j}(\theta)$. The equality of the

Laplace transforms implies that $X_i =_{st} X_j$ contradicting the fact that $X_i <_{st} X_j$ and proving (ii).

Proof of Proposition 3.1:

Proof. Let $P_0 \in \mathcal{P}_1$ be the true value of P . Recall that $\bar{P} = \Pi_{\Sigma_0}(\hat{P}|\mathcal{P}_1)$ is the $\widehat{\Sigma}_0$ -projection of the unconstrained MLE onto $\mathcal{P}_1$ where Σ_0 denotes the true scaled variance of the MLE. By definition

$$0 \leq (\tilde{P} - P_0)^T \widehat{\Sigma}_0^{-1} (\tilde{P} - P_0) \leq (\tilde{P} - P_0)^T \widehat{\Sigma}_0^{-1} (\tilde{P} - P_0) = o_p(1)$$

since $n_i \rightarrow \infty$ implies the consistency of $\hat{P}$. Thus $\hat{P}$ and $\bar{P}$ have the same limit P_0 as required.

Proof of Theorem 3.1:

Proof. Let $P_0 \in \mathcal{P}_0$ be the true value of P . Note that (2.6) may be rewritten as $H_0: \mathbf{R}P = 0$ versus $H_1: \mathbf{R}P \neq 0$ where $\mathbf{R} = (r_{ij})$ is a $m(K-1) \times mK$ matrix with elements $r_{ij} = -1$ if $j = i$, $r_{ij} = 1$ if $j = i + m$ and $r_{ij} = 0$ otherwise. By the multivariate central limit theorem

$$\sqrt{n}(\hat{P} - P_0) \Rightarrow N(0, \Sigma_0) \quad (8.1)$$

where $n = \sum_{i=1}^K n_i$, the symbol $\Rightarrow$ denotes convergence in distribution, and $Var(\hat{P}) = \Sigma_0/n$ at P_0 . It follows that

$$\begin{aligned} D_n &= n \left\{ \|\hat{P} - \Pi_{\Sigma_0}(\hat{P}|\mathcal{P}_0)\|_{\Sigma_0}^2 - \|\hat{P} - \Pi_{\Sigma_0}(\hat{P}|\mathcal{P}_1)\|_{\Sigma_0}^2 \right\} \\ &= \min_{P \in \mathcal{P}_0} n(\hat{P} - P)^T \Sigma_0^{-1} (\hat{P} - P) - \min_{P \in \mathcal{P}_1} n(\hat{P} - P)^T \Sigma_0^{-1} (\hat{P} - P) + o_p(1) \\ &\Rightarrow \min_{P \in \mathcal{P}_0} (Z - P_0)^T \Sigma_0^{-1} (Z - P_0) - \min_{P \in \mathcal{P}_1} (Z - P_0)^T \Sigma_0^{-1} (Z - P_0) \\ &= (Z - P_0)^T \Sigma_0^{-1} (Z - P_0) - \min_{P \in \mathcal{P}_1 \cap \mathcal{P}_0^\perp} (Z - P_0)^T \Sigma_0^{-1} (Z - P_0). \end{aligned}$$

The second equality above follows by the consistency of $\widehat{\Sigma}_0$. Convergence follows from (8.1) where Z follows a $\mathcal{N}_{m(K-1)}(P_0, \Sigma_0)$ distribution; and the third equality follows from Theorem 3.7.1 in Silvapulle and Sen (2005) where $\mathcal{P}_1 \cap \mathcal{P}_0^\perp$ is a convex cone ($\mathcal{P}_0^\perp$ is the linear space orthogonal to $\mathcal{P}_0$). Thus in Silvapulle's and Sen's (2005) notation

$D_n \Rightarrow \bar{\chi}^2(\Sigma_0, \mathcal{P}_1 \cap \mathcal{P}_0^\perp)$ is a chi-bar square distributed RV. Since Σ_0 is of full rank and $\mathbf{R}$ is a restriction matrix (a matrix whose rows are permutations of $(-1, 1, 0, \dots, 0)$) it now follows from their Corollary 3.7.2 that

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_n \geq x) = \sum_{k=0}^{m(K-1)} w_k(\Sigma_0, R) \mathbb{P}(\chi_k^2 \geq x)$$

establishing (3.4).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research of Ori Davidov was partially supported by the Israeli Science Foundation Grant No 875/09 and was conducted when visiting Shyamal Das Peddada at the NIEHS. This research [in part] was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04). We would like to thank Professor Doron Zeilberger (Rutgers University) for pointing out the connection between the number of upper sets and the Dedekind numbers and Drs. Grace Kissling (NIEHS) and Wenge Guo (New Jersey Institute of Technology) for several useful comments that improved the presentation. We would also like to thank the editors and reviewers for constructive comments that helped improve many aspects of the paper.

References

- [1]. Agresti A, Klingenberg B. Multivariate tests comparing binomial probabilities with application to safety studies for drugs. *Applied Statistics*. 2005; 54:691–706.
- [2]. Ananth C, Kantor M. Modeling multivariate binary responses with multiple levels of nesting based on alternating logistic regressions: an application to caries aggregation. *Journal of Dental Research*. 2004; 83:776–781. [PubMed: 15381718]
- [3]. Baccelli F, Makowski AM. Multidimensional stochastic ordering and associated random variables. *Operations Research*. 1989; 37:478–487.
- [4]. Bahadur, RR. A representation of the joint distribution of responses to n dichotomous items. In: Solomon, H., editor. *Studies in Item Analysis and Prediction*. Stanford Mathematical Studies in Social Sciences VI. Stanford University Press; Stanford CA: 1961. p. 158–168.
- [5]. Bailer JA, Portier CJ. Effects of treatment induced mortality and tumor induced mortality on tests for carcinogenicity in small samples. *Biometrics*. 1988; 44:417–431. [PubMed: 3390507]
- [6]. Basso, D.; Salmaso, L.; Pesarin, F.; Solari, A. *Permutation Tests for Stochastic Ordering and ANOVA*. Springer; 2009.
- [7]. Burnside G, Pine C, Williamson P. The application of multilevel modeling to dental caries data. *Statistics in Medicine*. 2007; 26:4139–4149. [PubMed: 17340596]
- [8]. Davidov O, Fokianos K, Iliopoulos G. Order restricted semiparametric inference for the power bias model. *Biometrics*. 2010; 66:549–557. [PubMed: 19522874]
- [9]. Davey, BA.; Priestley, HA. *Introduction to Lattices and Order*. Cambridge University Press; 2002.
- [10]. Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. *The American Statistician*. 1991; 54:302–304.
- [11]. Finos L, Salmaso L, Solari A. Conditional inference under simultaneous stochastic ordering constraints. *Journal of Statistical Planning and Inference*. 2007; 137:2633–2641.
- [12]. Finos L, Pesarin F, Salmaso L, Solari A. Exact inference for multivariate ordered alternatives. *Statistical Methods and Applications*. 2008; 17:195–208.
- [13]. Fitzmaurice G, Laird N. A likelihood based method for analyzing longitudinal binary responses. *Biometrika*. 1993; 80:141–151.
- [14]. Fitzmaurice G, Laird N, Zahner G. Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*. 1996; 91:99–108.
- [15]. Giancristofaro RA, Bonnini S. Moment based multivariate permutation tests for ordinal categorical data. *Journal of Nonparametric Statistics*. 2008; 20:383–393.
- [16]. Glonek GFV, McCullagh P. Multivariate logistic models. *Journal of the Royal Statistical Society*. 1995; 57:533–546. Series B
- [17]. Han KE, Catalano PJ, Senechaudhuri P, Mehta C. Exact analysis of dose response for multiple correlated binary outcomes. *Biometrics*. 2004; 60:216–224. [PubMed: 15032792]
- [18]. Hoschele I, Foulley J, Colleau J, Gianola D. Genetic evaluation for multiple binary responses. *Genet. Sel. Evol.* 1986; 18:299–320. [PubMed: 22879249]
- [19]. Joe, H. *Multivariate Models and Dependence Concepts*. Chapman and Hall; 1997.

- [20]. Keenan C, Elmore S, Franck-Carroll S, Kemp R, Kerlin R, Ped-dada SD, Pletcher J, Rinke M, Schmidt S, Taylor I, Wolf D. Best Practices for Use of Historical Control Data of Proliferative Rodent Lesions. *Toxicologic Pathology*. 2009; 37:679–693. [PubMed: 19454599]
- [21]. Klingenberg B, Agresti A. Multivariate extensions of Mc-Nemar's test. *Biometrics*. 2006; 62:921–928. [PubMed: 16984337]
- [22]. Klingenberg B, Solari A, Salmaso L, Pesarin F. Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*. 2009; 65:452–462. [PubMed: 18510649]
- [23]. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- [24]. Lucas LA, Wright FT. Testing for and against stochastic ordering between multivariate Multinomial populations. *Journal of Multivariate Analysis*. 1991; 38:167–186.
- [25]. McComb DJ, Kovacs K, Beri J, Zak F. Pituitary adenomas in old Sprague-Dawley rats: A histologic, ultrastructural, and immunocytochemical study. *Journal of the National Cancer Institute*. 1984; 73:1143–1166. [PubMed: 6092768]
- [26]. Montgomery, DC.; Peck, EA.; Vining, GG. *Introduction to Linear Regression Analysis*. 3rd Edition. Wiley & Sons; 2001.
- [27]. Moser V. Observational batteries in neurotoxicity testing. *International Journal of Toxicology*. 2000; 19:407–411.
- [28]. Maydeu-Olivares A, Hernandez A, McDonald R. A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research*. 2006; 41:445–471.
- [29]. NTP TR-527. The National Toxicology Program. NIEHS, RTP; NC: 2005. NTP Technical Report on the Toxicology and Carcinogenesis Studies of Malachite Green Chloride and Leucomalachite Green (CAS Nos. 569-64-2 and 129-73-7) in F344/N Rats and B6C3F1 Mice (Feed Studies).
- [30]. Peddada SD, Dinse G, Haseman J. A survival-adjusted quantal response test for comparing tumor incidence rates. *Journal of the Royal Statistical Society*. 2005; 54:51–61. Series C
- [31]. Peddada SD, Dinse GE, Kissling GE. Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association*. 2007; 102:1212–1220. [PubMed: 20396669]
- [32]. Pesarin, F. *Multivariate Permutation Tests : With Applications in Biostatistics*. Wiley & Sons; 2001.
- [33]. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987; 43:487–498. [PubMed: 3663814]
- [34]. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*. 2003; 90:455–463.
- [35]. Ribaudo HJ, Thompson SG. The analysis of repeated multivariate binary quality of life data: A hierarchical model approach. *Statistics Methods in Medical Research*. 2002; 11:69–83.
- [36]. Sampson AR, Whitaker LR. Estimation of multivariate distributions under stochastic ordering. *Journal of the American Statistical Association*. 1989; 84:541–548.
- [37]. Sen PK. Union–intersection principle and constrained statistical inference. *Journal of Statistical Planning and Inference*. 2007; 137:3741–3752.
- [38]. Shaked, M.; Shanthikumar, JG. *Stochastic Orders*. Springer; 2007.
- [39]. Silvapulle, MJ.; Sen, PK. *Constrained Statistical Inference*. Wiley & Sons; 2005.
- [40]. Song, PXX. *Correlated Data Analysis*. Springer; 2007.
- [41]. Teoh, E.; Nyska, A.; Wormser, U.; Peddada, SD. Statistical inference under order restrictions on both rows and columns of a matrix, with an application in toxicology. In: Balakrishnan, N.; Pena, E.; Silvapulle, MJ., editors. *Beyond Parametrics in Interdisciplinary Research*, Festschrift to P.K. Sen. IMS Collections; 2008. p. 62–77.
- [42]. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci*. 2001; 98:5116–5121. [PubMed: 11309499]
- [43]. Troendle JT. Multiple comparisons between two groups on multiple Bernoulli outcomes while accounting for covariates. *Statistics in Medicine*. 2005; 24:3581–3591. [PubMed: 15977268]

- [44]. Westfall PH, Young SS. P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*. 1989; 84:780–786.
- [45]. Williams D. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*. 1971; 27:103–117. [PubMed: 5547548]