Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Coding invariance in factorial linear models and a new tool for assessing combinatorial equivalence of factorial designs

Ulrike Grömping

Beuth University of Applied Sciences, Luxemburger Str. 10, 13353 Berlin, Germany

ARTICLE INFO

Article history:
Received 24 March 2017
Received in revised form 21 July 2017
Accepted 22 July 2017
Available online xxxx

Keywords:
Experimental design
Qualitative factors
Combinatorial equivalence
Mean aberrations
Generalized word length pattern

ABSTRACT

This paper provides new insights into coding invariance for linear models with qualitative factors, including a coding invariant way of denoting the model coefficients. On this basis, "interaction contributions" (ICs) are proposed for decomposing generalized word counts for factorial designs into contributions that neither depend on level allocation nor on the coding of factors. Combinatorially equivalent designs yield the same ICs, so that ICs are suitable for classifying factorial designs with qualitative factors. ICs are based on singular value decomposition and have an interpretation in terms of bias contributions of an interaction on the estimation of the overall mean. The paper introduces ICs and their tabulations in interaction contribution frequency tables and illustrates their behavior in various examples. ICs are compared to several other tools for assessing combinatorial equivalence of general factorial designs, and they are found to provide a useful complement to existing methods.

1. Introduction

Two designs are called isomorphic, if they can be obtained from each other by swaps of columns and/or rows and/or *appropriate* relabelings of factor levels. Isomorphism has to be judged differently for designs with qualitative or quantitative factors: for designs with quantitative factors, isomorphism is sometimes called "geometric isomorphism" (see e.g. Cheng and Ye, 2004); here, changes in level orderings can lead to non-isomorphic designs. For qualitative factors, on the other hand, each relabeling of factor levels leads to an isomorphic design; this type of isomorphism will be called "combinatorial equivalence" in this paper; the expression "non-isomorphic" is used as a short form for "not combinatorially equivalent", and "equivalence screening" is used as a short form for "checking whether necessary conditions for equivalence are violated". When searching for appropriate designs, designs isomorphic to ones that have already been investigated need not be considered. Therefore, the ability to decide whether or not designs are isomorphic is important for efficiently handling resources. This is not only the case when searching for good or optimal designs, but can also be relevant for basic activities regarding experimental practice, for example when trying to adapt data from an existing experiment to a software tool that provides a factorial structure; the effort of obtaining an appropriate map between two isomorphic designs can be quite large and should only be undertaken if the designs in question are isomorphic.

Criteria for assessing combinatorial equivalence have to be coding invariant in two ways: they must not depend on swapping some levels in any design column, and for a given set of levels, they must not depend on a particular coding

E-mail address: pgroemping@beuth-hochschule.de.

Abbreviations: IC, Interaction Contribution; *ICFT*, Interaction Contribution Frequency Table; *DEFT*, Distance Enumeration Frequency Table; *ODFM*, Ordered Distance Frequency Matrices; *PFT*, Projection Frequency Table; *PMFT*, Power Moment Frequency Table; *MAFT*, Mean Aberration Frequency Table; *SCFT*, Squared Canonical correlation Frequency Table; *ARFT*, Average Frequency Table

http://dx.doi.org/10.1016/j.jspi.2017.07.004
0378-3758/

of the model matrix of a factorial linear model. In their seminal paper on generalized minimum aberration (GMA), Xu and Wu (2001) introduced the so-called normalized orthogonal coding (see Definition 2 in Section 2), which ensures that all coefficient estimators in a factorial linear model for a full factorial design are uncorrelated and have the same variance. This paper interprets coding invariance as invariance against the choice of a *normalized orthogonal* coding. Throughout the paper, the expression "coding invariant" will always be used in this sense, which comprises both level allocation and effect coding. It will be shown that outer products of effect model matrices are coding invariant in this sense. Furthermore, the paper will provide a coding invariant way of specifying effect coefficient vectors as a linear combination of right singular vectors of the effect model matrix. This allows, for example, to create simulation scenarios involving effect sizes for qualitative factors in a coding invariant way. The results on coding invariance will also serve as the basis for the development of the "interaction contributions" (ICs) that will be introduced in this paper as a tool for assessing combinatorial equivalence.

Clark and Dean (2001) and Katsaounis and Dean (2008) introduced necessary and sufficient conditions for combinatorial equivalence; checking these can be painfully slow, so that various faster tools for equivalence screening have been proposed, and Katsaounis (2012) proposed to use these also for screening designs with qualitative and quantitative factors. Section 2.2 will present a collection of existing tools for equivalence screening, including squared canonical correlation frequency tables (*SCFT*s) by Grömping (2017a) and mean aberration frequency tables (*MAFT*s) by Fontana et al. (2016), among several others. The ICs proposed in this paper focus on general factorial designs with at least some factors at more than two levels (since the toolbox for 2-level designs is already quite powerful); they will have to compete with the existing tools. This article considers screening tools only, and the Examples section contains a non-isomorphic set of designs that cannot be distinguished by any of the screening tools considered here (i.e. none of them provides sufficient conditions for equivalence).

ICs, like several of the other tools for equivalence screening, are based on generalized word counts: Xu and Wu (2001) introduced the generalized word length pattern (*GWLP*) which is now widely used as the basis of GMA. For a design with n factors, the *GWLP* can be written as $(A_0, A_1, \ldots, A_n)$, where $A_0 = 1$ generally holds. For $j > 0$, the generalized count A_j of words of length j can be written as a sum of generalized word counts $a_j(S)$ from all sets S of j factors. These $a_j(S)$ are called projected a_j values in this paper, and they will be defined and explained in Section 2.2. In many applications, the *GWLP* is applied to orthogonal array designs, which implies that $A_1 = A_2 = 0$, so that the first interesting entry is A_3. In this paper, $A_1 = 0$ is assumed (i.e., level balance of all factors), and the number R with $A_1 = \cdots = A_{R-1} = 0$ and $A_R > 0$ is called the resolution of the design; this is in line with the conventional understanding of resolution (e.g. in Hedayat et al., 1999 p. 280) for $R \geq 3$ and extends the concept to $R = 2$, e.g. for supersaturated designs. The majority of the tools for equivalence screening used in this paper is related to the projected a_j values, with particular focus on projected a_R values; the other tools are based on Hamming distances between design rows (see Section 2.2).

The ICs to be developed in this article provide a new coding invariant decomposition of the projected a_j values. They work for pure or mixed level designs with factors at arbitrary numbers of levels. ICs are based on singular value decomposition (SVD); ambiguities arising from singular values with multiplicity larger than one are resolved in two different ways, which leads to two types of interaction contribution frequency tables (*ICFT*s). Like the projected a_j values themselves, ICs have a statistical interpretation in terms of bias contributions of the interaction to estimation of the overall mean. It is proposed to use *ICFT*s for equivalence screening of general factorial designs, and the examples will demonstrate that they complement the existing tools for this purpose. Note that, in spite of also using singular values, the approach of the present paper is quite different from the proposal by Katsaounis et al. (2013) of using singular values for checking design equivalence for 2-level designs.

Section 2 will introduce notation and basic concepts, including a detailed introduction to the existing tools for equivalence screening. Section 3 will provide two fundamental theorems on coding invariance in factorial linear models. Section 4 will introduce the interaction contributions and their properties and will provide the aforementioned two types of *ICFT*s. Section 5 will provide several examples that exemplify the details of ICs as well as their performance in equivalence screening in comparison to other tools. The final section will discuss connections to further related work and reasonable future steps.

2. Notation and basic concepts

We consider factorial designs with n factors in N runs, with s_i levels for the ith factor. The designs are level-balanced, i.e. each factor has each level the same number of times, which implies at least resolution II. Subsets of the factors are denoted by $S \subseteq \{1, \ldots, n\}$, and the cardinality of a set S is denoted by $|S|$. For $j \in \{1, \ldots, n\}$, $\mathcal{S}_j = \{S \subseteq \{1, \ldots, n\} : |S| = j\}$ denotes the set of all j factor sets. The restriction of a design to the factors from a set $S \in \mathcal{S}_j$ is called a j factor projection and is for simplicity identified with the set S.

2.1. Matrix tools and factorial linear models

Before discussing factorial linear models, some matrix products are defined and rules for them established. In the following, the superscript $\top$ denotes transposition. $\mathbf{1}_N$ and $\mathbf{0}_N$ denote column vectors of N ones or zeros, respectively, $\mathbf{e}_i$ denotes a unit vector with the value "1" in position i and zeros everywhere else, and $\mathbf{I}_N$ denotes an N-dimensional identity matrix. An orthogonal matrix $\mathbf{Q}$ is an $r \times r$ matrix with $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_r$. Note that multiplication with an orthogonal matrix applies rotation and/or reflection operations only. This paper will use the term "rotation" for multiplication with any orthogonal matrix $\mathbf{Q}$, regardless whether $\mathbf{Q}$ involves only proper rotation ($det(\mathbf{Q}) = 1$) or not.