Journal of Statistical Planning and Inference 209 (2020) 144–159

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Support point of locally optimal designs for multinomial logistic regression models

Shuai Hao [a], Min Yang [b,*]

[a] Abbvie Inc. North Chicago, IL, 60064, United States of America
[b] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago Chicago, IL 60607, United States of America

ARTICLE INFO

Article history:
Received 16 April 2019
Received in revised form 18 March 2020
Accepted 22 March 2020
Available online 25 April 2020

MSC:
62K05
62J12

Keywords:
Lowner ordering
Baseline link
Cumulative link
Adjacent link
Continuation-ratio link

ABSTRACT

While multinomial logistic models have been widely applied in practice, the research on design selection has not kept pace. The complication in studying optimal/efficient designs for multinomial logistic models is the complicated structure of information matrices due to the model complexity and existence of many variants. A critical step in deriving optimal/efficient designs is to determine the number of support points needed. In this paper, we systematically characterize the optimal designs through the complete class framework. The results hold for any optimal designs, regardless of optimality criterion chosen, parameters of interest, one-stage or multi-stage designs. It provides insight in the structure of optimal designs for multinomial logistic models from theoretical perspective and makes the follow-up derivation much easier.

1. Introduction

1.1. Multinomial logistic regression model

The logistic regression model is a common tool for analyzing binary responses. Under many circumstances, however, binary responses may not be sufficient for describing the results (Perevozskaya et al., 2003). For example, in a dose–response study, while binary response is used to estimate dose–response curve, such response usually does not contain information about the severity of toxicity. According to Schacter et al. (1997), a subject can suffer from five types of adverse effects ranging from self-limiting nausea to death in phase I cancer trial. In general, if responses from an experiment take values from a fixed set containing J (>2) categories, they are called polytomous responses, which usually follow the multinomial distribution. Polytomous data is often modeled by the multinomial logistic regression model (MLRM) (Agresti, 2013, chap 6), a special case of generalized linear models (GLM) (McCullagh and Nelder, 1989). In particular, given a polytomous response, say Y, it is modeled by the following,

$$G(E(Y)) = \eta. \tag{1}$$

Here $G(\cdot)$ is called *link function* that transforms observed responses to log-odds, E is the expectation operator, and η is the *linear component*.

* Corresponding author.
E-mail addresses: shao21@uic.edu (S. Hao), myang2@uic.edu (M. Yang).

https://doi.org/10.1016/j.jspi.2020.03.006
0378-3758/

MLRM is a broad class of models. Despite of its simple looking like (1), it could be arbitrarily complex in the following perspectives.

First, unlike the binary logistic regression where a single log-odds is modeled, one need to model $J-1$ log-odds simultaneously if the response is polytomous and has J response categories.

Second, judged from the relation among response categories, polytomous response can be categorized into three kinds: nominal, ordinal, and hierarchical. For nominal response, categories are considered as equally important. For example, blood types, car makes, and etc. On the contrary, there is a nature order among categories of ordinal response, such as beef quality grade, people's preference rating to a restaurant, and etc. Hierarchical response is different because some of response categories are nested in others. For example, in McCullagh and Nelder (1989), a study of mortality due to radiation consists of three stages. At first stage, outcomes are 'alive' and 'dead'; then at second stage, those who died are divided into 'due to cancer' and 'other cause'; at last, those who died from cancer are labeled either 'other cancer' or 'leukemia'.

Third, each kind of polytomous response requires a properly chosen link function. For nominal response, baseline link (3) is appropriate since the conclusion drawn from the fitted model with baseline link is still valid if the label of categories are permuted (McCullagh and Nelder, 1989). As to ordinal response, cumulative link (4) (McCullagh and Nelder, 1989), or adjacent link (5) (Liu and Agresti, 2005; Agresti, 2013), are preferred for this case because if the order of categories is reversed, the conclusions made from the fitted model with either of those link functions remain unchanged. If response is hierarchical, continuation ratio link (6) is recommended by Zocchi and Atkinson (1999).

Fourth, the complexity also lies on the linear component in (1). The linear components are usually summarized into three types of model assumptions: the proportional odds model (*po*), the non-proportional odds model (*npo*) and the partial proportional odds model (*ppo*) (Bu et al., 2019). Here the 'odds' refers to 'log-odds', which is discussed in detail in next section. For the proportional model, linear components across categories share the same set of parameters, whereas the non-proportional model assumes each category has its own set of parameters that distinguish themselves across categories. The members in the partial proportional model share parameters across categories while each of them possesses its own set of parameters. It is obvious that the partial proportional model is an amalgamation of po and npo models and therefore is the most general.

1.2. The present knowledge of optimal design for MLR models

While MLRMs are widely applied in practice and the methodology of analyzing such models is well established, the optimal design research for MLRMs is arguably in its infancy stage with little optimality result available. The available results, which are summarized in the following paragraphs, are scattered around and lack of systematical work.

As mentioned before there are at least twelve types (at least four types of link functions coupled with at least three types of model assumptions) of MLRM due to the variety of link functions and model assumptions. The information matrix, which is the key to the study of optimal design, has its own structure under each model. Therefore, one has to develop tools for optimal design case by case.

One major obstacle of studying optimal designs for MLRMs is that the information matrix depends on the unknown parameter $\boldsymbol{\theta}$ due to the nonlinearity. A common approach to solve this dilemma is to use locally optimal designs, which are based on one's best guess of the unknown parameters. While a good guess is not always guaranteed, this approach remains of value to obtain benchmarks for all designs (Ford et al., 1992). There are other ways to address this issue, for example, by using a Bayesian approach (Chaloner and Verdinelli, 1995).

The complicated structure of the information matrix makes it notoriously difficult to derive the corresponding optimal designs under MLRMs. There are, however, some nice attempts to attack this complexity problem. In Zocchi and Atkinson (1999), they considered Bayesian D-optimal design for a multinomial logistic model based on hierarchical responses collected from an experiment on emergence of houseflies. They used Markov Chain Monte Carlo to generate a sample of parameters in order to access to the objective function. Some properties of the information matrix of the proportional odds model with cumulative link were explored in Perevozskaya et al. (2003), and locally optimal designs under multiple optimal criteria were investigated through numerical construction therein. A model with cumulative link for ordinal data was studied by Yang et al. (2017) and they had shown the size of minimally supported design only depends on number of predictors. Locally D- and EW-D optimal designs were derived through algorithm approaches. Most recently, Bu et al. (2019) conducted a comprehensive study on all 12 variants of multinomial logistic regression models and provide general conclusions on the cardinality of minimally supported designs. Algorithm for D-optimal designs was also provided.

While these results explore some optimal designs, there is a lack of systematic understanding of their characterizations — arguably speaking, little is known about them. In this paper, we study the characterization of optimal designs for MLRMs through a complete class framework proposed by a series of papers (Yang and Stufken, 2009; Yang, 2010; Dette and Melas, 2011; Yang and Stufken, 2012; Dette and Schorning, 2013). The strategy is to find a subclass with simple format such that, for any design outside the complete class, say, ξ_1, there always exists a design in this subclass, say, ξ_2, and the information matrix of ξ_2 dominates that of ξ_1 in Lowner ordering. Notice that other strategies are also feasible. For example, the functional approach (Melas, 2006). The main idea of this approach is to express the support points (and sometimes also the weights) of optimal designs as implicit functions of some auxiliary parameters. In many cases these functions, which are real and analytic, can be expanded into Taylor series, for the coefficients of which recursive formulae

are available. While this approach is valuable in studying optimal designs, in this paper, we shall focus on the complete class strategy.

Utilizing this strategy, we obtain complete class results for a broad class of MLRMs. The results are significant for three reasons. First, it is the first time the characterizations of optimal designs under a varieties of MLRMs are derived. The results can help us understand the structure of optimal designs systematically. Second, the characterizations can significantly simplify the search of any specific optimal designs, both analytically and numerically, regardless of parameters of interest, optimality criteria, one-stage or multiple stage design. Third, a pressing research direction in big data analysis is the trade-off between computation complexity and statistical efficiency under the constraint of limited computing resources. The derived characterizations can guide us to develop efficient algorithms of selecting an informative subdata which can address the trade-off adequately (Wang et al., 2018).

The rest of this paper is organized as follows. Notations and model settings are given in Section 2. The main results are given in Section 3. Some applications are provided in Section 4. A brief discussion is given in Section 5. All proofs are included in the Appendix.

2. Notations and settings

2.1. Multinomial logistic regression model

Suppose in an experiment, we observe n polytomous responses with J possible response categories from m distinct experiment settings. Particularly, at ith experiment settings, n_i responses, say y_{ij} for $j = 1, \ldots, n_i$, are collected, where $\sum_{i=1}^{m} n_i = n$.

Typically, y_{ij}'s from the same experiment setting are summarized into a count vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ij})'$, where Y_{ik} means the counts of responses obtained at ith experiment setting that belong to the kth category, or equivalently, if we code response categories as integers from 1 to J, then $Y_{ik} = \sum_{j=1}^{n_i} \mathbb{1}(y_{ij} = k)$ where $\mathbb{1}(\cdot)$ is an indicator function. Let the $\pi_{ik} = Prob(y_{ij} = k)$ for $k = 1, 2, \ldots, J$, $\sum_{j=1}^{J} \pi_{ij} = 1$, the distribution of $\mathbf{Y}_i$ is multinomial, $\mathbf{Y}_i \sim Multinomial(n_i, \pi_{i1}, \ldots, \pi_{iJ})$ with Probability Mass Function being

$$Prob[\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ij})] = \binom{n_i}{Y_{i1}, \ldots, Y_{iJ}} \prod_{j=1}^{J} \pi_{ij}^{Y_{ij}}$$

In probability theory, the multinomial distribution is generalized from the binomial distribution and it belongs to the exponential family. Therefore, the multinomial logistic model, a generalized version of the logistic model, is appropriate to model the probabilities. For the ith experiment setting, say $s_i = (x_{i1}, \ldots, x_{ip})$, one needs to model all probabilities simultaneously in the following general form,

$$G(\boldsymbol{\pi}_i) = \eta_i = \mathbf{X}_i\boldsymbol{\theta}, \tag{2}$$

where the link function $G(\cdot)$ is a map $\mathbb{R}^J \mapsto \mathbb{R}^{J-1}$, $\eta_i = (\eta_1(\boldsymbol{\pi}_i), \ldots, \eta_{J-1}(\boldsymbol{\pi}_i))'$ is a $(J-1)\times 1$ vector, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})'$ is a $J \times 1$ vector, $\mathbf{X}_i = (f_1(s_i), \ldots, f_{J-1}(s_i))'$ is a design matrix of order $(J-1)\times \nu$. Here $f_s'(\mathbf{X}_i)$ stands for its sth row, f is a function on $\mathbb{R}^p \mapsto \mathbb{R}^\nu$, $\boldsymbol{\theta}$ is a vector of unknown parameters of length ν. The linear component is $\eta = \mathbf{X}_i\boldsymbol{\theta}$.

The link function $G(\cdot)$ transforms responses to log-odds. All four links functions can be summarized as follows.

$$baseline \quad \log\frac{\pi_{ij}}{\pi_{iJ}} \qquad \text{for} \quad j = 1, \ldots, J-1, \tag{3}$$

$$cumulative \quad \log\frac{\pi_{i1} + \cdots + \pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}} \qquad \text{for} \quad j = 1, \ldots, J-1, \tag{4}$$

$$adjacent \quad \log\frac{\pi_{ij}}{\pi_{i,j+1}} \qquad \text{for} \quad j = 1, \ldots, J-1, \tag{5}$$

$$continuation - ratio \quad \log\frac{\pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}} \qquad \text{for} \quad j = 1, \ldots, J-1. \tag{6}$$

For baseline link (3), one can arbitrarily choose a so called *reference category*, of which the probability will be fixed on the bottom. Here we put π_{iJ} on the denominator only for illustration purpose. Meanwhile, there are many possible pairs to be used in getting log-odds, but most of them are redundant. For example, with J response categories and baseline link, there are $J(J-1)/2$ possible pairs, however one only needs to model $J-1$ selected pairs and the rest of them can be obtained using the existing ones. In general, for all the link functions above, $J-1$ log-odds are sufficient.

Notice that there is no standard criterion of choosing the right type of link functions. For some cases, as illustrated in McCullagh and Nelder (1989, chap 6), both baseline link and cumulative link yield similar parameter estimates and conclusions.

The linear component of (2) depends on model assumptions. There are at least three model assumptions in literature (Bu et al., 2019): proportional odds (*po*), non-proportional odds (*npo*), and partial proportional odds (*ppo*). For $j = 1, \ldots, J-1$, let X_i^{rt} be the rtth entry of design matrix $\mathbf{X}_i$,

$$po \quad \eta_j(\boldsymbol{\pi}_i) = X_i^{j1}\theta_1 + \cdots + X_i^{j\nu}\theta_\nu, \tag{7}$$

$$npo \quad \eta_j(\boldsymbol{\pi}_i) = X_i^{j1}\theta_{j1} + \cdots + X_i^{jv}\theta_{jv}, \tag{8}$$

$$ppo \quad \eta_j(\boldsymbol{\pi}_i) = X_i^{j1}\theta_1 + \cdots + X_i^{j\tilde{v}}\theta_{\tilde{v}} + X_i^{j,\tilde{v}+1}\theta_{j,\tilde{v}+1} + \cdots + X_i^{jv}\theta_{jv}, \tag{9}$$

where $\tilde{v}$ is the number of parameters shared across categories in (9).

2.2. Unified model

In an effort to unify them, Glonek (see Glonek and McCullagh, 1995) proposed a transformation that covers a wide scope of link functions between the multinomial logistic model and the log-linear model. It is written as

$$\mathbf{C}\log(\mathbf{L}\boldsymbol{\pi}_i) = \begin{pmatrix}\eta_i \\ 0\end{pmatrix} = \mathbf{X}_i\boldsymbol{\theta} \quad \text{for} \quad i = 1, \ldots, m. \tag{10}$$

where η_i is defined in (2), $\mathbf{C}$ is a $J \times (2J-1)$ constant matrix, with $\mathbf{I}_{J-1}$ being the identity matrix of order $J-1$ and $\mathbf{0}_{J-1}$ is a vector of $(J-1)$ 0's,

$$\mathbf{C} = \begin{pmatrix}\mathbf{I}_{J-1} & -\mathbf{I}_{J-1} & \mathbf{0}_{J-1} \\ \mathbf{0}'_{J-1} & \mathbf{0}'_{J-1} & 1\end{pmatrix},$$

$\mathbf{L}$ is a $(2J-1) \times J$ matrix varies through link functions. For baseline, cumulative, adjacent, and continuation-ratio link functions, the concrete structure of $\mathbf{L}$ matrices can be found in the Appendix.

There is a major difference between (2) and (10). Because matrix $\mathbf{C}$ has J rows, which means (10) models simultaneously model J log-odds. A close look at $\mathbf{C}$ reveals that the last row is merely for imposing the constraint $\sum_{j=1}^{J}\pi_{ij} = 1$, and this is the reason that the last row of $\mathbf{L}$ is all 1's regardless of type link functions.

2.3. Information matrix

An important step for deriving the Fisher information matrix is to invert η_i for $\boldsymbol{\pi}_i$. Provided all the link functions we introduced, we can find the closed form of $\boldsymbol{\pi}_i$ in terms of $\mathbf{X}_i\boldsymbol{\theta}$. As an example, for baseline link, the π_{ij}'s could be calculated via

$$\pi_{ij} = \frac{exp\{\mathbf{X}_i\boldsymbol{\theta}\}}{1 + exp\{\mathbf{X}_i\boldsymbol{\theta}\}} \quad \text{for} \quad j = 1, \ldots, J-1.$$

Following Bu et al. (2019), the information matrix for $\boldsymbol{\theta}$ in (10) is

$$I_i(\boldsymbol{\theta}) = (\frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\theta}'})'diag\{\boldsymbol{\pi}_i\}^{-1}(\frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\theta}'}),$$

where $\partial\boldsymbol{\pi}_i/\partial\boldsymbol{\theta}' = (\mathbf{C}\mathbf{D}_i^{-1}\mathbf{L})^{-1}\mathbf{X}_i$ and $\mathbf{D}_i = diag\{\mathbf{L}\boldsymbol{\pi}_i\}$. Matrices $\mathbf{C}$ and $\mathbf{L}$ are defined in (10). $\mathbf{X}_i$ is the design matrix related to design point $s_i = (s_1, \ldots, s_p)$.

2.4. Optimal designs for MLRM

Let s_i be a *design point* (or *experiment setting*), which is a vector of regression variables. A collection of all possible design points is named *design space* and denoted by $\boldsymbol{\chi}$. Let d be an *exact design* with n runs and *support* $\mathcal{S}$, where $\mathcal{S}$ is a set of m distinct design points. It can be written as

$$d = \{(s_i, n_i), s_i \in \mathcal{S}, \sum_{i=1}^{m} n_i = n, n_i \in \mathcal{Z}^+\},$$

where n_i's are repetitions associated with s_i's and are restricted to be positive integers. The set of all positive integers is $\mathcal{Z}^+$. An *optimal exact design* is therefore a collection of (s_i, n_i) that collectively optimizes an objective function that defined in terms of the information matrix. Provided a design d, the information matrix for the unknown parameter can be represented by

$$I_d = \sum_{i=1}^{m} n_i I_i,$$

where I_i is the information matrix for the design point s_i. However, it is often an intractable issue to find optimal exact designs due to its restrictions on repetitions. In particular, the optimal exact design in closed form is frequently sought via combinatorial tools, but the solution only exists for certain combinations of experiment configurations, such as the number of total runs, levels of regression variables and etc. Moreover, because this discrete nature on repetitions, numerical algorithms that work with derivatives are not applicable either. Consequently, optimal designs are often studied in the context of *approximate designs* by relaxing the discrete repetitions to continuous *weights* (or *proportions* in some literature).

Formally, n_i are replaced by $w_i = n_i/n$ and the w_i's (weights) are assumed to be real numbers in the interval $[0, 1]$. An approximate design ξ as well as its information matrix, is written as follows,

$$\xi = \{(s_i, w_i), s_i \in S, \sum_{i=1}^{m} w_i = 1, w_i \in [0, 1]\},$$
$$I_\xi = \sum_{i=1}^{m} w_i I_i = \sum_{i=1}^{m} \frac{n_i}{n} I_i = \frac{1}{n} I_d.$$

An *optimal approximate design* is hence sought for. The consequence of relaxation on repetitions is profound since there are a variety of optimization tools and numerical algorithms that are available in literature. As a trade-off, one has to take extra effort to carefully round an approximate design to an exact design which is either optimal or efficient prior to the implementation. Throughout this paper, the term 'optimal design' refers to an optimal approximate design unless otherwise specified.

3. Main result

The strategy for developing those findings is inspired by the complete class framework developed by Yang and Stufken (2012).

3.1. Model under consideration

Our results are mainly on the baseline proportional odds model and some special cases for models with other links. In general, a multinomial logistic model with J ($\geq$3) response categories and p continuous regression variables is

$$\mathbf{C}' \log(\mathbf{L}\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\theta}, \tag{11}$$

where

$$\mathbf{C}' = \begin{pmatrix} \mathbf{I}_{J-1} & -\mathbf{I}_{J-1} & \mathbf{0}'_{J-1} \\ \mathbf{0}_{J-1} & \mathbf{0}_{J-1} & 1 \end{pmatrix}, \quad \mathbf{X}_i = \begin{pmatrix} 1 & & & x_{i1} & \cdots & x_{ip} \\ & \ddots & & x_{i1} & \cdots & x_{ip} \\ & & 1 & x_{i1} & \cdots & x_{ip} \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{12}$$

$\mathbf{L}$ is an $(2J-1)\times J$ constant matrix and depends on choice of link functions. $\mathbf{X}_i$ is the design matrix associated to design point $s_i = (x_{i1}, \ldots, x_{ip})$, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_{J-1}, \beta_1, \ldots, \beta_p)$ is the parameter vector of which α_j's are intercepts and β_i's are coefficients of regression variables. Here we only assume $x_{ij} \in [U_j, V_j]$ for $j = 1, \ldots, p-1$, where U_j, V_j are real numbers, and x_{ip} is unbounded.

3.2. Information matrix and its blocks

Since the analytical approach requires identification of the maximal set of linear independent non-constant functions, we first introduce the general structure of the information matrix.

Following Bu et al. (2019), given design point s_i, the information matrix for $\boldsymbol{\theta}$ is

$$\begin{aligned} I_i(\boldsymbol{\theta}) &= (\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\theta}'})' diag\{\boldsymbol{\pi}_i\}^{-1} (\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\theta}'}) \\ &= \mathbf{X}_i'[(\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}]' diag\{\boldsymbol{\pi}_i\}[(\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}]\mathbf{X}_i \end{aligned} \tag{13}$$

where $\partial\boldsymbol{\pi}_i/\partial\boldsymbol{\theta}' = (\mathbf{C}'\mathbf{D}_i^{-1}\mathbf{L})^{-1}\mathbf{X}_i$ and $\mathbf{D}_i = diag\{\mathbf{L}\boldsymbol{\pi}_i\}$.

We let $\mathbf{U}$ be the matrix in the middle except for the design matrix, then (13) can be written as $I_i = \mathbf{X}_i'\mathbf{U}\mathbf{X}_i$. Although the concrete expression of $\mathbf{U}$ varies case by case, according to Corollary 3.1 in Bu et al. (2019), it has a general structure

$$\mathbf{U} = \begin{pmatrix} \mathbf{M} & \mathbf{0}'_{J-1} \\ \mathbf{0}_{J-1} & 1 \end{pmatrix}, \tag{14}$$

where $\mathbf{M}$ is a $(J-1)\times(J-1)$ symmetric matrix. In addition, if the design matrix $\mathbf{X}_i$ is partitioned as follows for blockwise matrix multiplication

$$X_i = \begin{pmatrix} \mathbf{I}_{J-1} & \mathbf{S} \\ \mathbf{0}_{J-1} & \mathbf{0}_1 \end{pmatrix}, \tag{15}$$

where the submatrix $\mathbf{S}$ is a $(J-1)\times p$ matrix that holds values of regression variables, and $\mathbf{0}_{J-1}$ and $\mathbf{0}_1$ are vectors of 0's with appropriate orders. As a result, we reach to the following lemma for the structure of the information matrix.

Lemma 3.1. *Given matrix partitions in (14) and (15), the information matrix at $s_i = (x_{i1}, \ldots, x_{ip})$ can be presented by blocks.*

$$I_i(\theta) = \begin{pmatrix} \mathbf{I}_{J-1} & \mathbf{S} \\ \mathbf{0}_{J-1} & \mathbf{0}_1 \end{pmatrix}' \begin{pmatrix} \mathbf{M} & \mathbf{0}'_{J-1} \\ \mathbf{0}_{J-1} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I}_{J-1} & \mathbf{S} \\ \mathbf{0}_{J-1} & \mathbf{0}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{M} & \mathbf{MS} \\ \mathbf{S}'\mathbf{M} & \mathbf{S}'\mathbf{MS} \end{pmatrix} = \begin{pmatrix} B1 & B2' \\ B2 & B3 \end{pmatrix}.$$

We name those blocks by 'B' + numbers, where the letter 'B' is short for 'Block', and $B2'$ is B2 block transposed.

Furthermore, let $\mathbf{M} = \{M_{ij}\}$, and define $M_{\cdot j} = \sum_{i=1}^{J} M_{ij}$ and $M_{\cdot\cdot} = \sum_{i=1}^{J-1}\sum_{j=1}^{J-1} M_{ij}$,

$$B2 = \begin{pmatrix} x_1 M_{\cdot 1} & x_1 M_{\cdot 2} & \cdots & x_1 M_{\cdot p} \\ x_2 M_{\cdot 1} & x_2 M_{\cdot 2} & \cdots & x_2 M_{\cdot p} \\ \vdots & \vdots & \ddots & \vdots \\ x_p M_{\cdot 1} & x_p M_{\cdot 2} & \cdots & x_p M_{\cdot p} \end{pmatrix}, \qquad B3 = \begin{pmatrix} x_1^2 M_{\cdot 1} & x_1 x_2 M_{\cdot 2} & \cdots & x_1 x_p M_{\cdot p} \\ x_2 x_1 M_{\cdot 1} & x_2^2 M_{\cdot 2} & \cdots & x_2 x_p M_{\cdot p} \\ \vdots & \vdots & \ddots & \vdots \\ x_p x_1 M_{\cdot 1} & x_p x_2 M_{\cdot 2} & \cdots & x_p^2 M_{\cdot p} \end{pmatrix}.$$

The proof is merely matrix multiplications and is therefore omitted here. Lemma 3.1 plays an important role in following sections where those structures will be extensively exploited.

3.3. Proportional model with 3 response categories

For $J = 3$ and $p = 1$, the design point, $s_i = x_i$, reduces to a scalar and the design matrix as well as $\boldsymbol{\theta}$ become

$$X_i = \begin{pmatrix} 1 & 0 & x_i \\ 0 & 1 & x_i \\ 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}.$$

In the information matrix, the B2 block reduces to a row vector and the B3 is now a scalar. We investigated such a matrix and reach to the following theorem on a complete class.

Theorem 3.2. *For proportional odds model (11) with 1 continuous regression variable $x_i \in [U, V]$, where U, V are real numbers, and 3 response categories, the following results on complete class hold.*

1. *For baseline link, designs with at most 2 support points form a complete class.*
2. *For cumulative, continuation ratio or adjacent link, designs with at most 4 support points form a complete class.*

The proof is given in the Appendix. Theorem 3.2 provides upper bounds of the number of support points for MLRM with 1 continuous covariate and 3 response categories. In particular, optimal designs for such model with baseline link will have at most 2 support points. Meanwhile, the model with cumulative, continuation ratio and adjacent link will have at most 4 design points. According to Bu et al. (2019), the minimal number of support points for this case is 2 for baseline link and 3 for the rest type of links. Combined with Theorem 3.2, optimal designs for the baseline multinomial logistic regression model with 3 response categories are minimally supported.

3.4. Baseline proportional odds model with J categories

In this section, we generalized complete class result for the baseline proportional odds model to the one with $J \geq 3$ response categories. The model is

$$\log(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta x, \quad j = 1, \ldots, J-1.$$

If it is written in matrices like (11), the design matrix X_i and $\boldsymbol{\theta}$ now become

$$X_i = \begin{pmatrix} 1 & & & x_i \\ & \ddots & & \vdots \\ & & 1 & x_i \\ 0 & \ldots & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{J-1} \\ \beta \end{pmatrix}$$

In its information matrix, the B2 block is still a row vector and B3 is a scalar. But B1 block now is of order $J-1$ by $J-1$. We have the following complete class result for this case.

Theorem 3.3. *For the baseline proportional odds model (11) with $J \geq 3$ response categories and 1 continuous regression variable $x_i \in [U, V]$, where U, V are real numbers, designs with at most 2 support points form a complete class.*

Theorem 3.3 generalizes complete class result for the baseline proportional model to an arbitrary number of response categories. That is, the optimal design for the baseline proportional odds model consists at most 2 support points regardless of the number of response categories. It broadens the scope of its applications.

3.5. Baseline proportional odds model with J categories and p regression variables

We now consider arbitrary J and p. The baseline proportional odds model with $J \geq 3$ response categories and $p \geq 2$ continuous regression variables can be written as

$$\log(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta_1 x_1 + \cdots \beta_p x_p, \quad j = 1, \ldots, J-1,$$

where Jth response category is conventionally set to be a reference category, and x_i are the value of ith regression variable. Here we only assume $x_j \in [U_j, V_j]$ for $j = 1, \ldots, p-1$, where U_j, V_j are finite real numbers.

As introduced at the beginning of this section, the design matrices and parameter vector are exactly the same as (12). For example, when $J = 4$, $p = 2$, the design matrix is

$$X = \begin{pmatrix} 1 & 0 & 0 & x_1 & x_2 \\ 0 & 1 & 0 & x_1 & x_2 \\ 0 & 0 & 1 & x_1 & x_2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

In general, by Lemma 3.1, key components in the information matrix are as follows.

The B1 block is the matrix $\mathbf{M}$ of order $(J-1) \times (J-1)$ with

$$M_{ij} = \begin{cases} -\frac{e^{\alpha_i + \alpha_j + 2\sum_{t=1}^{p} \beta_t x_t}}{[1 + (\sum_{s=1}^{J-1} e^{\alpha_s}) e^{\sum_{t=1}^{p} \beta_t x_t}]^2}, & i \neq j \\ \frac{e^{\alpha_i + \sum_{t=1}^{p} \beta_t x_t} [1 + (\sum_{s=1, s\neq j}^{J-1} e^{\alpha_s}) e^{\sum_{t=1}^{p} \beta_t x_t}]}{[1 + (\sum_{s=1}^{J-1} e^{\alpha_s}) e^{\sum_{t=1}^{p} \beta_t x_t}]^2}, & i = j. \end{cases}$$

The B2 block is a $p \times (J-1)$ matrix with the ijth entry being

$$x_i M_{\cdot j} = x_i \frac{e^{\alpha_j + \sum_{t=1}^{p} \beta_t x_t}}{[1 + (\sum_{j=1}^{J-1} e^{\alpha_j}) e^{\sum_{t=1}^{p} \beta_t x_t}]^2}, \quad \text{for } j = 1, \ldots, J-1$$

and the B3 block is a $p \times p$ matrix with ijth entry being

$$x_i x_j M_{\cdot\cdot} = \frac{e^{\sum_{t=1}^{p} \beta_t x_t} \sum_{s=1}^{J-1} e^{\alpha_s}}{[1 + (\sum_{j=1}^{J-1} e^{\alpha_j}) e^{\sum_{t=1}^{p} \beta_t x_t}]^2}.$$

We focus on the transformed design points, with support points $s_i = (x_{i1}, \ldots, x_{i,p-1}, c_i)$, where $c_i = \sum_{t=1}^{p} \beta_t x_t$ and $\beta_t \neq 0$ for all possible t. Note that such transformation does not change the complete class result, because of the following factorization of the information matrix. For a design point x and its transformed design point s

$$I(s, \boldsymbol{\theta}) = \mathbf{X}'\mathbf{U}\mathbf{X} = \mathbf{Q}'\mathbf{F}'\mathbf{U}\mathbf{F}\mathbf{Q}, \tag{16}$$

where $\mathbf{X}$ is the design matrix for $x = (x_1, \ldots, x_p)$ and $\mathbf{F}$ is the design matrix for $s = (x_1, \ldots, c)$, and $\mathbf{F}\mathbf{Q} = X$,

$$A(\theta) = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ 0 & \cdots & \beta_1 & \cdots & \beta_p \end{pmatrix}^{-1}, \qquad F = \begin{pmatrix} 1 & & & x_1 & \cdots & c \\ & \ddots & & \vdots & \vdots & \vdots \\ & & 1 & x_1 & \cdots & c \end{pmatrix}.$$

Let $I(s, \boldsymbol{\theta})$ stand for the information matrix at $s = (x_1, \ldots, c)$, one can easily obtain $I(s, \boldsymbol{\theta})$ from $I(x, \boldsymbol{\theta})$ by (16). The structures of them are identical. Under this setting, we have the following theorem.

Theorem 3.4. *In the transformed design space, for an arbitrary design $\xi = \{(s_i, w_i), i = 1, \ldots, m; \sum_{i=1}^{m} w_i = 1\}$, there exists a design $\tilde{\xi}$ such that the following inequality for information matrices hold:*

$$I_{\xi}(\theta) \leq I_{\tilde{\xi}}(\theta),$$

where

$$\tilde{\xi} = \{(\tilde{F}_{\ell 1}, w_{\ell 1}) \quad \text{and} \quad (\tilde{F}_{\ell 2}, w_{\ell 2}), \ell = 1, \ldots, 2^{p-1}\}$$

and $\tilde{F}_{\ell 1} = (a_{\ell 1}, \ldots, a_{\ell, p-1}, \tilde{c}_1)$, $\tilde{F}_{\ell 2} = (a_{\ell 1}, \ldots, a_{\ell, p-1}, \tilde{c}_2)$. Here $a_{\ell, j} = U_j$ or V_j, and $(a_{\ell 1}, \ldots, a_{\ell, p-1})$ are all combinations of them for $\ell = 1, \ldots, 2^{p-1}$, and $\tilde{c}_1$ and $\tilde{c}_2$ are two numbers need to be solved.

The proof is deferred to the Appendix as well. Theorem 3.4 shows that the optimal designs for the baseline proportional model with p covariates are made of two equivalent classes of design points of which the value of its first $p-1$ covariates

Table 1
Locally optimal designs for the baseline proportional odds models.

$(\alpha_1, \alpha_2, \alpha_3, \beta)$	criterion	design space	N	design	# of points
(0.5, −0.6, 0.9, 2)	D	[−3, 3]	121	(−1.25, 0.3106) (0.35, 0.6894)	2
(0.5, −0.6, 0.9, 2)	A	[−3, 3]	121	(−1.4, 0.2198) (0.2, 0.7802)	2
(0.5, −0.6, 0.9, 2)	D	[0, 3]	61	(0.0, 0.4580) (1.1, 0.5420)	2
(0.5, −0.6, 0.9, 2)	A	[0, 3]	61	(0.0, 0.4509) (1.2, 0.5491)	2
(1, 2, 4, −0.3)	D	[−10, 30]	81	(6.5, 0.6877) (17.0, 0.3123)	2
(1, 2, 4, −0.3)	A	[−10, 30]	81	(5.0, 0.8536) (21.5, 0.1464)	2
(1, 2, 4, −0.3)	D	[0, 30]	61	(6.5, 0.6877) (17.0, 0.3123)	2
(1, 2, 4, −0.3)	A	[0, 30]	61	(6.0, 0.7751) (24.5, 0.2249)	2

are easily found. The significance is not only the optimal designs for such a general model are in a simple structure, also algorithms would benefit from it since it reduces the dimension of an optimization problem from p to 1.

Note that when $J = 2$, Theorem 3.4 reduces to Theorem 2 in Yang et al. (2011), where similar result for binary logistic regression is derived. Therefore, it generalized Yang's result to the baseline log-odds model.

4. Applications

All designs in this section are locally optimal. Therefore, one needs to provide initial values of parameters in order to derive a design aiming at estimating them. Initial values are not randomly chosen, on the contrary, it should be determined prudently by either consulting experts or look up historical experiment results. We have mentioned that the locally optimal design can serve as a benchmark for other designs. However, a set of badly chosen initial values which are far away from the 'truth' would result in a benchmark that has not too much practical meaning even though it is locally optimal.

We use the optimal weights exchange algorithm (OWEA) in Yang et al. (2013) for approximate designs. Although the scope of this paper is on models with continuous regression variables, we still need to discretize the design space in order to use the algorithm. A common practice is to use an equally spaced grid on the design space, as we did in following examples.

4.1. Examples

Example 1. Consider the following baseline proportional odds model, with $J = 4$

$$\log(\frac{\pi_1}{\pi_4}) = \alpha_1 + \beta x,$$
$$\log(\frac{\pi_2}{\pi_4}) = \alpha_2 + \beta x,$$
$$\log(\frac{\pi_3}{\pi_4}) = \alpha_3 + \beta x.$$

We select two sets of initial parameters $(0.5, -0.6, 0.9, 2)$ and $(1, 2, 4, -0.3)$, for the given design spaces, we use R program to find optimal approximate designs for both A- and D-optimal criteria. These approximate designs are summarized in Table 1. Here the N stands for number of grid points in design space, and entries on the column 'design' are written in the format of $(point, weight)$. Finally, we count the number of support points and add them to the last column.

It is noticeable that all those designs consist of two support points, which is consistent with our findings in Theorem 3.3. In fact, according to Bu et al. (2019), designs with 2 support points for this model are also minimally supported which is the minimum requirement for the information matrix being non-singular and hence parameter estimation being unbiased. Therefore, optimal designs could be both optimal and minimally supported. An interesting observation is, contrary to common case where the D-optimal design has equal weights, the minimally supported D-optimal design is not equally weighted. For example, for the first set of initial parameter values, the D-optimal design has two points $-1.25, 0.35$ with weights 0.3106 and 0.6894. Meanwhile, the A-optimal design has two weights being 0.2198 and 0.7802.

Table 2
Locally optimal designs for the cumulative proportional odds model.

$(\alpha_1, \alpha_2, \beta)$	Criterion	Design space	N	Design	# of points
$(1, 2.88, -0.5)$	D	$[-10, 15]$	501	$(0.55, 0.3970)$ $(3.85, 0.0214)$ $(3.90, 0.1853)$ $(7.20, 0.3963)$	4*
$(1, 2.88, -0.5)$	A	$[-10, 15]$	501	$(-1.00, 0.6539)$ $(3.45, 0.3461)$	2
$(1, 2.88, -0.5)$	D	$[0, 15]$	1501	$(0.53, 0.3949)$ $(3.86, 0.2066)$ $(7.18, 0.3985)$	3
$(1, 2.88, -0.5)$	A	$[0, 15]$	1501	$(0.00, 0.7456)$ $(3.95, 0.2209)$ $(9.27, 0.0335)$	3
$(-2, 1, 0.8)$	D	$[-10, 10]$	2001	$(-2.20, 0.3180)$ $(0.62, 0.3630)$ $(3.44, 0.3190)$	3
$(-2, 1, 0.8)$	A	$[-10, 10]$	2001	$(-2.10, 0.2800)$ $(0.85, 0.6918)$ $(3.68, 0.0282)$	3
$(-2, 1, 0.8)$	D	$[0, 9]$	901	$(0.00, 0.6361)$ $(3.94, 0.0454)$ $(4.00, 0.3185)$	3
$(-2, 1, 0.8)$	A	$[0, 9]$	901	$(0.00, 0.8630)$ $(4.07, 0.0117)$ $(4.50, 0.1253)$	3

Example 2. In Perevozskaya et al. (2003), an early pioneer paper that studies designs for MLRM, they provided the locally optimal designs for the following model, which is a cumulative link model with 4 response categories.

$$\log \frac{\gamma_j(x)}{1-\gamma_j(x)} = x - \alpha_j \quad \text{for } j = 1, 2, 3,$$

where $\gamma_j(x) = Prob(Y \le j|x) = \sum_{s=1}^{j} \pi_{is}$. Here those intercept terms are unknown parameters and they set the slope to be a constant. Such a model is used for dose–response study. Inspired by this paper, we reparameterize model in the fashion of the proportional odds model and assume the slope is also unknown. For simplicity, we only consider 3 response categories. The model is formulated as

$$\log \frac{\gamma_1(x)}{1-\gamma_1(x)} = \alpha_1 + \beta x,$$
$$\log \frac{\gamma_2(x)}{1-\gamma_2(x)} = \alpha_2 + \beta x.$$

Notice that there is a natural order that $\alpha_1 \ge \alpha_2$.

Similarly, two sets of initial parameter values are chosen upon which A- and D-optimal designs are derived for given design spaces. Table 2 summarizes key information of those designs.

Most of those designs in Table 2 have 3 support points, except those on the first two rows. In particular, the D-optimal design on the first row have two points, 3.85, 3.90, that can be combined as one since they are actually two adjacent grid points and 3.85 has very small weight. Such a design can be considered as the one with three support points, 0.55, 7.20 and a, where $a \in (3.85, 3.90)$. As Theorem 3.2 shows optimal designs have at most 4 support points, the abundance of 3-point designs and absence of 4-point designs might give a hint that our current result could be improved. Finally, it is worth mentioning that, according to Bu et al. (2019), those D-optimal designs are also minimally supported.

In practice, when there is no information on unknown parameters, an intuitive yet commonly used strategy is to implement uniform designs. Such a design puts equal weights on design points, and sometimes those points are equally spread in design space as well. However, they are known as lacking efficiency. For example, Yang et al. (2017) and Bu et al. (2019) proved uniform designs are less efficient for D-optimality under some variants of multinomial logistic regression. We have the same observation here. Consider the following two uniform designs in Table 3, of which puts equal weights to its support.

For simplicity, we compare optimal designs with uniform designs. Here in Table 3, 'A-eff' and 'D-eff' are shorts for the relative efficiency under A- and D-optimality respectively, and they are calculated by

$$eff = \frac{\Phi(\Sigma_{optimal})}{\Phi(\Sigma_{uniform})}. \tag{17}$$

Table 3
Uniform designs and efficiency to the optimal designs.

Design space	Design points	# of points	A-eff	D-eff
[−10, 15]	−10, −5, 0, 5, 10, 15	6	0.5339	0.2589
[0, 15]	0, 2, 4, 8, 15	5	0.6740	0.5862
[−10, 10]	−10, −6, −2, 0, 2, 6, 10	7	0.4822	0.1892
[0, 9]	0, 1, 2, 3, 4, 5, 6, 7, 8, 9	10	0.3563	0.2278

Table 4
Original design in a toxicity study.

Dose	0	62.5	125	250	500
Observations	297	242	312	299	285

When $eff > 1$, it indicates uniform design is more efficient, and vice versa. As shown, uniform designs are not as efficient as those optimal designs. On the contrary, the difference in efficiency is quite huge. For example, given the design space $[-10, 15]$, there are 2 points and 4 points for the A- and D-optimal designs, and they are almost 1 and 3 times more efficient than the uniform design with 5 points.

Example 3. In Agresti (2013, chap 6), a developmental toxicity study with pregnant mice was introduced. In this experiment, a certain chemical substance in distilled water of different concentrations (from 0 to 500 mg/kg per day) was given to pregnant mice in successive 10 days and their uterine contents were analyzed in order to examine the defects of fetuses. There are three outcomes for each fetus: nonlive, malformation, or normal. The outcome is ordinal with 'nonlive' being the most preferable. The original design has 5 levels of concentration, 0, 62.5, 125, 250, 500, where 0 is the level of control group. Those design points spread out in the design space, [0, 500]. Design points and the number of observations are summarized in Table 4.

A continuation-ratio proportional model is considered because the response is hierarchical. With 3 response categories, if we target at the following model,

$$\begin{aligned} \log \frac{\pi_1}{\pi_{i2} + \pi_{i,3}} &= \alpha_1 + \beta x, \\ \log \frac{\pi_2}{\pi_{i,3}} &= \alpha_2 + \beta x, \end{aligned} \tag{18}$$

where the x means the concentration. For this example, we set $(\alpha_1, \alpha_2, \beta) = (0.1, -0.5, 0.016)$, which is the initial estimate provided by Agresti (2013, chap 6). We use the OWEA algorithm to find the A- and D-optimal designs in the space [0, 500]. Both designs are summarized in Table 5. The numbers of support points, regardless of optimal criteria, are all equal to 2, which is less than the upper bound of 4 provided in Theorem 3.2. Also, the design points are not uniformly allocated.

In fact, for this toxicity study, Agresti (2013, chap 6) fitted a continuation-ratio non-proportional model, and give estimations for $\beta_1 = 0.0064$, $\beta_2 = 0.0174$.

$$\begin{aligned} \log \frac{\pi_1}{\pi_{i2} + \pi_{i,3}} &= \alpha_1 + \beta_1 x, \\ \log \frac{\pi_2}{\pi_{i,3}} &= \alpha_2 + \beta_2 x. \end{aligned} \tag{19}$$

Although we have not derived any complete class result for such model, optimal designs can still be derived numerically. In this case, we search locally optimal designs at $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (0.4, 1, 0.0064, 0.0174)$, which serves as initial 'guess' of unknown parameters.

As shown in Table 5, designs have 3 levels of concentrations for the D-optimality and only 2 for the A-optimality. Under both criteria, the control level 0 is always included. The significance is that an experimenter can reduce the levels of concentrations which saves labor and reduces the experimental cost. Lastly, we still observed that those designs have the number of design points that are less than the theoretical maximum as we derived in Theorem 3.2.

The last column in Table 5 is the relative efficiency that comparing the original design in Agresti (2013, chap 6) to the optimal design in Table 5. The formula is similar to (17). It is obvious that the original design is lacking the efficiency for both A- and D-optimality. For example, the D-optimal design based on 3 points for the npo model is almost 2 times as efficient as the original design. The take away message is optimal or efficient designs for models like (18) and (19), can be based on only a limited number of design points. Those observations are in line with the spirit of theorems derived in this paper.

Table 5
Locally optimal designs for continuation models.

$(\alpha_1, \alpha_2, \beta_1, \beta_2)$	Criterion	Design space	N	Design	# of points	Efficiency
Locally Optimal Designs for the Proportional Odds Model						
$(0.1, -0.5, 0.016)$	D	[0, 500]	501	$(0, 0.6323)$ $(121, 0.3677)$	2	0.2296
$(0.1, -0.5, 0.016)$	A	[0, 500]	501	$(0, 0.9926)$ $(122, 0.0074)$	2	0.2984
Locally Optimal Designs for the Non-proportional Odds Model						
$(0.4, 1, 0.0064, 0.0174)$	D	[0, 500]	501	$(0, 0.4653)$ $(117, 0.3741)$ $(365, 0.1606)$	3	0.5108
$(0.4, 1, 0.0064, 0.0174)$	A	[0, 500]	501	$(0, 0.9797)$ $(105, 0.0203)$	2	0.2533

5. Discussion

The multinomial logistic regression model plays an important role in statistical analysis. However, the research on optimal designs is still at its infancy stage. Deriving optimal designs for MLRM in general is difficult. As stated in introduction, there are two major obstacles. First, the MLRM consists of at least 12 types of variants and each has its own concrete expression of both the model and the information matrix. So far relevant optimal designs are generated case by case. Second, the information matrix depends on the unknown parameters, and mostly, the locally optimal designs are studied.

Although there are an increasing number of researches on related fields, the optimal design for MLRM is still under development and almost all of existing designs emerged in literature so far are constructed in a numerical manner. While these results are helpful in some sense, they are in fact merely computational and cannot provide further insights. In recent decades, some preliminary theoretical results have been established regarding the unified model representation, the information matrix and etc. There are, nevertheless, still no such studies for optimal designs from theoretical perspective.

In this paper, we accessed the optimal designs for MLRMs via an analytical approach. The main result is on the complete class of optimal designs for some prevalent models. In particular, we derived the upper bound for the number of support points of optimal designs. Such results provide evidence for the claim that optimal designs for MLRM usually do not have many support points. This is important because one can expect a simple design constructed by numerical algorithms.

Numerical examples are also explored. It is shown that the number of support points of those designs are in line with our theory. In particular, some examples have the number of supports that is exactly what indicated by our theorem. More interestingly, designs from some other examples have less support points than what we derived in theory. Since our theory holds regardless of initial values of parameters and optimal criteria, there might be some other cases that have exactly the maximum number of support points. Moreover, and even more exciting, it might be possible to improve our result in the future.

In addition, selecting initial values for parameters is tricky. For example, Agresti (2013, chap 6) argues that cumulative link indicates that the cumulative probability must be stochastically ordered, otherwise, the model will be poorly fitted. This is the general guidelines for choosing initial parameters. Some bad chosen sets not only result in inadequately fitted models, but also ill-organized designs. As to our experience, some of the choice of initial parameter would result in the singular information matrix, and one has to be prudent to exclude design points like this in the algorithm, since the framework of OWEA relies on the non-singular information matrices.

The study of designs for the multinomial logistic regression model is still under development. There are many interesting yet untouched topics in this field. For example, designs for MLRM with mixed type of regression variables, or when there are higher power terms or interactions in linear components, and etc. We hope our work can trigger more research in these topics.

CRediT authorship contribution statement

Shuai Hao: Methodology, Writing - original draft. **Min Yang:** Methodology, Writing - review & editing.

Acknowledgments

The authors are grateful for many insightful comments and suggestions from an anonymous referee, an associate editor, and editor, which helped to improve the article. Yang's research was supported by NSF grant DMS-1811291.

Appendix A. L matrices

$$\mathbf{L}_{baseline} = \begin{pmatrix} 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix} \quad \mathbf{L}_{cumulative} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & 1 & \dots & 1 & 0 \\ 0 & 1 & \dots & 1 & 1 \\ 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

$$\mathbf{L}_{continuation} = \begin{pmatrix} 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \\ 0 & 1 & \dots & \dots & 1 \\ 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix} \quad \mathbf{L}_{adjacent} = \begin{pmatrix} 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \\ 0 & 1 & & & \\ 0 & & 1 & & \\ \vdots & & & \ddots & \\ 0 & & & & 1 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

Appendix B. Proofs

Proof of Theorem 3.2. We only provide proofs for baseline link and cumulative link. For continuation ratio and adjacent link, the arguments are similar to that of cumulative link. Since the information matrix at $J = 3, p = 1$ is simple, we work on them directly.

The main task is to identify the complete class. For a complete class Ξ, define two designs $\xi \notin \Xi$ and $\tilde{\xi} \in \Xi$ on the design space χ,

$$\begin{aligned} \xi &= \{(c_i, w_i), c_i \in \chi, \sum_{i=1}^{m} w_i = 1\} \\ \tilde{\xi} &= \{(\tilde{c}_i, \tilde{w}_i), \tilde{c}_i \in \chi, \sum_{i=1}^{k} \tilde{w}_i = 1\} \end{aligned} \tag{B.1}$$

Part I. For the baseline link, the information matrix at design point x is

$$I = \begin{pmatrix} \frac{e^{\alpha_1+\beta x}(1+e^{\alpha_2+\beta x})}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & -\frac{e^{\alpha_1+\alpha_2+2\beta x}}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & \frac{e^{\alpha_1+\beta x}x}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} \\ -\frac{e^{\alpha_1+\alpha_2+2\beta x}}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & \frac{e^{\alpha_2+\beta x}(1+e^{\alpha_1+\beta x})}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & \frac{e^{\alpha_2+\beta x}x}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} \\ \frac{e^{\alpha_1+\beta x}x}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & \frac{e^{\alpha_2+\beta x}x}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} & \frac{e^{\beta x}(e^{\alpha_1}+e^{\alpha_2})x^2}{(1+e^{\alpha_1+\beta x}+e^{\alpha_2+\beta x})^2} \end{pmatrix}. \tag{B.2}$$

To prove the complete class result,

Step 1: (Selection) Let $c = \beta x$ (where $\beta \neq 0$), then there is a bijection between x and c, and $x = c/\beta$. Among the first two columns, select the following set as maximal linear independent nonconstant functions:

$$\begin{aligned} \Psi_1(c) &= \frac{e^{\alpha_1+c}(1+e^{\alpha_2+c})}{(1+e^{\alpha_1+c}+e^{\alpha_2+c})^2}, \\ \Psi_2(c) &= -\frac{e^{\alpha_1+\alpha_2+2c}}{(1+e^{\alpha_1+c}+e^{\alpha_2+c})^2}, \\ \Psi_3(c) &= \frac{e^{\alpha_1+c}c}{\beta(1+e^{\alpha_1+c}+e^{\alpha_2+c})^2}, \end{aligned}$$

and let

$$\Psi_4(c) = \frac{c^2 e^c(e^{\alpha_1}+e^{\alpha_2})}{\beta^2(1+e^{\alpha_1+c}+e^{\alpha_2+c})^2}.$$

Here let $g(c) = (1+e^{\alpha_1+c}+e^{\alpha_2+c})^2$, and inequality $g(c) > 0$ holds on its domain. Such an arrangement is due to the fact that the B1 block in (B.2) only has two linear independent functions in terms of c.

Step 2: (Simplification) The task is to show the following system for any two designs ξ and $\tilde{\xi}$ in (B.1),

$$\begin{aligned}
\sum_{i=1}^{m} w_i\Psi_1(c_i) &= \sum_{i=1}^{k} \tilde{w}_i\Psi_1(\tilde{c}_i),\\
\sum_{i=1}^{m} w_i\Psi_2(c_i) &= \sum_{i=1}^{k} \tilde{w}_i\Psi_2(\tilde{c}_i),\\
\sum_{i=1}^{m} w_i\Psi_3(c_i) &= \sum_{i=1}^{k} \tilde{w}_i\Psi_3(\tilde{c}_i),\\
\sum_{i=1}^{m} w_i\Psi_4(c_i) &\le \sum_{i=1}^{k} \tilde{w}_i\Psi_4(\tilde{c}_i),
\end{aligned} \tag{B.3}$$

and it is sufficient to show

$$\begin{aligned}
&\{1, \Psi_1, \Psi_2, \Psi_3\} \quad \text{and} \quad \{1, \Psi_1, \Psi_2, \Psi_3, \Psi_4\} \quad \text{are Chebyshev Systems,}\\
&\{1, \Psi_1, \Psi_2, \Psi_3\} \quad \text{and} \quad \{1, \Psi_1, \Psi_2, \Psi_3, -\Psi_4\} \quad \text{are Chebyshev Systems.}
\end{aligned} \tag{B.4}$$

Due to the existence of denominators in $\Psi(c)$, the recursive construction of $F(c)$ described in Theorem 2 in Yang and Stufken (2012) are expected to be cumbersome and the resultant function $F(c)$ can be rather complicated. Instead, we perform a series simplifications which preserve either the equality in (B.3) or the Chebyshev System in (B.4) but with more simple functions.

First, we omit the '−' sign in Ψ_2 and β in Ψ_3 which does not change the equality in (B.3). Then multiply all Ψ functions including the constant $\Psi_0 = 1$ by the denominator and conduct row or column operations that does not change the sign of matrix determinant. At last we get rid of positive constants like e^{α_1}, e^{α_2} and β^2 which preserve the Chebyshev System. Eventually, a set of functions Ψ is simplified to

$$\{1, e^c, e^{2c}, ce^c, c^2e^c\}.$$

To show (B.4) is equivalent to verifying either those following claims hold

$$\begin{aligned}
&\{1, e^c, e^{2c}, ce^c\} \quad \text{and} \quad \{1, e^c, e^{2c}, ce^c, c^2e^c\} \quad \text{are Chebyshev Systems,}\\
&\{1, e^c, e^{2c}, ce^c\} \quad \text{and} \quad \{1, e^c, e^{2c}, ce^c, -c^2e^c\} \quad \text{are Chebyshev Systems.}
\end{aligned}$$

Step 3: (Calculation) The sequence of $f_{\ell,\ell}$ functions can be easily calculated according to Theorem 3 of Yang and Stufken (2012). Here $f_{11} = e^c, f_{22} = 2e^c, f_{33} = -e^c/2, f_{44} = 2$, and $F(c) = \prod_{i=1}^{4} f_{ii}(c) = -2e^c < 0$. Then designs with at most 2 support points form a complete class is a direct consequence of the case (b) of Theorem 2 in Yang and Stufken (2012).

Part II. For cumulative link, the information matrix at the support point x is

$$I = \begin{pmatrix}
\frac{e^{\alpha_1+\alpha_2+\beta x}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_1+\beta x})^2} & -\frac{e^{\alpha_1+\alpha_2+\beta x}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_1+\beta x})(1+e^{\alpha_2+\beta x})} & \frac{xe^{\alpha_1+\alpha_2+2\beta x}}{(1+e^{\alpha_1+\beta x})^2(1+e^{\alpha_2+\beta x})}\\
-\frac{e^{\alpha_1+\alpha_2+\beta x}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_1+\beta x})(1+e^{\alpha_2+\beta x})} & \frac{e^{2\alpha_2+\beta x}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_2+\beta x})^2} & \frac{xe^{\alpha_2+\beta x}}{(1+e^{\alpha_1+\beta x})(1+e^{\alpha_2+\beta x})^2}\\
\frac{xe^{\alpha_1+\alpha_2+2\beta x}}{(1+e^{\alpha_1+\beta x})^2(1+e^{\alpha_2+\beta x})} & \frac{xe^{\alpha_2+\beta x}}{(1+e^{\alpha_1+\beta x})(1+e^{\alpha_2+\beta x})^2} & \frac{x^2e^{\alpha_2+\beta x}(1+2e^{\alpha_1+\beta x}+e^{\alpha_1+\alpha_2+2\beta x})}{(1+e^{\alpha_1+\beta x})^2(1+e^{\alpha_2+\beta x})^2}
\end{pmatrix}$$

Following the same steps:

Step 1: Let $c = \beta x$, we propose the assignments of functions:

$$\begin{aligned}
\Psi_1 &= \frac{e^{\alpha_1+\alpha_2+c}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_1+c})^2},\\
\Psi_2 &= -\frac{e^{\alpha_1+\alpha_2+c}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_1+c})(1+e^{\alpha_2+c})},\\
\Psi_3 &= \frac{e^{2\alpha_2+c}}{(e^{\alpha_2}-e^{\alpha_1})(1+e^{\alpha_2+c})^2},\\
\Psi_4 &= \frac{ce^{\alpha_1+\alpha_2+2c}}{\beta(1+e^{\alpha_1+c})^2(1+e^{\alpha_2+c})},\\
\Psi_5 &= \frac{ce^{\alpha_2+c}}{\beta(1+e^{\alpha_1+c})(1+e^{\alpha_2+c})^2},\\
\Psi_6 &= \frac{c^2e^{\alpha_2+c}(1+2e^{\alpha_1+c}+e^{\alpha_1+\alpha_2+2c})}{\beta^2(1+e^{\alpha_1+c})^2(1+e^{\alpha_2+c})^2}.
\end{aligned}$$

One can easily verify functions Ψ_1 to Ψ_5 form the set of maximal linear independent nonconstant functions among the first two columns $I(\boldsymbol{\theta})$.

Step 2. To show

$$\sum_{i=1}^{m} w_i\Psi_1(c_i) = \sum_{i=1}^{k} \tilde{w}_i\Psi_1(\tilde{c}_i),$$
$$\sum_{i=1}^{m} w_i\Psi_2(c_i) = \sum_{i=1}^{k} \tilde{w}_i\Psi_2(\tilde{c}_i),$$
$$\sum_{i=1}^{m} w_i\Psi_3(c_i) = \sum_{i=1}^{k} \tilde{w}_i\Psi_3(\tilde{c}_i),$$
$$\sum_{i=1}^{m} w_i\Psi_4(c_i) = \sum_{i=1}^{k} \tilde{w}_i\Psi_4(\tilde{c}_i),$$
$$\sum_{i=1}^{m} w_i\Psi_5(c_i) = \sum_{i=1}^{k} \tilde{w}_i\Psi_5(\tilde{c}_i),$$
$$\sum_{i=1}^{m} w_i\Psi_6(c_i) \le \sum_{i=1}^{k} \tilde{w}_i\Psi_6(\tilde{c}_i),$$

or its sufficient condition

$$\{1, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5\} \quad \text{and} \quad \{1, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, \Psi_6\} \quad \text{are Chebyshev Systems,}$$
$$\{1, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5\} \quad \text{and} \quad \{1, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, -\Psi_6\} \quad \text{are Chebyshev Systems.}$$

We did similar simplifications as introduced in part I and it turns out one needs to work with the following set of functions,

$$\{(1+e^{\alpha_1+c})^2(1+e^{\alpha_2+c})^2, e^c(1+e^{\alpha_2+c})^2, e^c(1+e^{\alpha_1+c})(1+e^{\alpha_2+c}), e^c(1+e^{\alpha_1+c})^2, ce^c \qquad (1+e^{\alpha_1+c}),$$
$$ze^{2c}(1+e^{\alpha_2+c}), c^2e^c(1+2e^{\alpha_1+c}+e^{\alpha_1+\alpha_2+2c})\}$$

However, the major difficulty is the resultant function $F(c)$ is still way too complicated from which one can draw conclusions regarding complete class. Instead, the best we can do so far is to investigate the Chebyshev System on an augmented set of linear independent functions:

$$\{1, e^c, e^{2c}, e^{3c}, e^{4c}, ce^c, ce^{2c}, ce^{3c}, c^2e^c(1+2e^{\alpha_1+c}+e^{\alpha_1+\alpha_2+2c})\}.$$

That is, we managed to reach to check the following,

$$\{1, e^c, e^{2c}, e^{3c}, e^{4c}, ce^c, ce^{2c}, ce^{3c}\} \quad \text{and}$$
$$\{1, e^c, e^{2c}, e^{3c}, e^{4c}, ce^c, ce^{2c}, ce^{3c}, c^2e^c(1+2e^{\alpha_1+c}+e^{\alpha_1+\alpha_2+2c})\} \quad \text{are Chebyshev Systems,}$$
or
$$\{1, e^c, e^{2c}, e^{3c}, e^{4c}, ce^c, ce^{2c}, ce^{3c}\} \quad \text{and}$$
$$\{1, e^c, e^{2c}, e^{3c}, e^{4c}, ce^c, ce^{2c}, ce^{3c}, -c^2e^c(1+2e^{\alpha_1+c}+e^{\alpha_1+\alpha_2+2c})\} \quad \text{are Chebyshev Systems.}$$

Step 3 Direct calculation shows $F(c) = \prod_{\ell=1}^{8} f_{\ell\ell} = -8e^z(3+2e^{\alpha_1+z}+3e^{\alpha_1+\alpha_2+2z}) < 0$. Then according to case (b) of Theorem 2 in Yang and Stufken (2012), designs with at most 4 points form a complete class. □

Proof of Theorem 3.3. For baseline link, by Lemma 3.1, the information matrix at the support point x is summarized blockwise.

The B1 block,

$$M_{ij} = \begin{cases} -\frac{e^{\alpha_i+\alpha_j+2\beta x}}{[1+(\sum_{s=1}^{J-1} e^{\alpha_s})e^{\beta x}]^2}, & i \neq j, \\ \frac{e^{\alpha_i+\beta x}[1+(\sum_{s=1,s\neq j}^{J-1} e^{\alpha_s})e^{\beta x}]}{[1+(\sum_{s=1}^{J-1} e^{\alpha_s})e^{\beta x}]^2}, & i = j. \end{cases}$$

The B2 block is a row vector and its jth entry is

$$xM_{\cdot j} = x\left\{\frac{e^{\alpha_j+\beta x}[1+(\sum_{s=1,s\neq j}^{J-1}e^{\alpha_s})e^{\beta x}]}{[1+(\sum_{s=1}^{J-1}e^{\alpha_s})e^{\beta x}]^2} - \sum_{i=1,i\neq j}^{J-1}\frac{e^{\alpha_i+\alpha_j+2\beta x}}{[1+(\sum_{s=1}^{J-1}e^{\alpha_s})e^{\beta x}]^2}\right\}$$
$$= \frac{xe^{\alpha_j+\beta x}}{[1+(\sum_{j=1}^{J-1}e^{\alpha_j})e^{\beta x}]^2}, \quad \text{for } j=1,\ldots,J-1,$$

and the B3 block is a scalar,

$$x^2M_{\cdot\cdot} = \frac{x^2e^{\beta x}\sum_{s=1}^{J-1}e^{\alpha_s}}{[1+(\sum_{j=1}^{J-1}e^{\alpha_j})e^{\beta x}]^2}.$$

In order to select a maximal set of linear independent nonconstant functions, we first introduce the following lemma that summaries the relevant property of the information matrix.

Lemma. *For the information matrix for the baseline proportional model with J categories, its B1 block only has two linear independent functions and its B2 block only has one linear independent function.*

The proof is evident in the calculations of the B1, B2, and B3 blocks.
Following standard steps.

Step 1: Let $c=\beta x$ (where $\beta\neq 0$), then there is a bijection between x and c, and $x=c/\beta$,

$$\Psi_1 = \frac{e^c}{[1+(\sum_{s=1}^{J-1}e^{\alpha_s})e^c]^2},$$
$$\Psi_2 = \frac{e^{2c}}{[1+(\sum_{s=1}^{J-1}e^{\alpha_s})e^c]^2},$$
$$\Psi_3 = \frac{ce^c}{\beta[1+(\sum_{j=1}^{J-1}e^{\alpha_j})e^c]^2},$$

and let

$$\Psi_4 = x^2M_{\cdot\cdot} = \frac{c^2e^c\sum_{s=1}^{J-1}e^{\alpha_s}}{\beta^2[1+(\sum_{j=1}^{J-1}e^{\alpha_j})e^c]^2}.$$

Let $g(c)=[1+(\sum_{j=1}^{J-1}e^{\alpha_j})e^c]^2$, and inequality $g(c)>0$ holds on all over its domain. Then one needs to verify either of those two claims hold.

$\{1,\Psi_1,\Psi_2,\Psi_3\}$ and $\{1,\Psi_1,\Psi_2,\Psi_3,\Psi_4\}$ are Chebyshev Systems,
$\{1,\Psi_1,\Psi_2,\Psi_3\}$ and $\{1,\Psi_1,\Psi_2,\Psi_3,-\Psi_4\}$ are Chebyshev Systems.

After simplification, it is equivalent to work on the following set of functions

$$\{1,e^c,e^{2c},ce^c,c^2e^c\}.$$

That is one need to verify those following claims:

$\{1,e^c,e^{2c},ce^c\}$ and $\{1,e^c,e^{2c},ce^c,c^2e^c\}$ are Chebyshev Systems,
$\{1,e^c,e^{2c},ce^c\}$ and $\{1,e^c,e^{2c},ce^c,-c^2e^c\}$ are Chebyshev Systems.

The result in Theorem 3.2 applies, and designs with at most 2 support points form a complete class. □

Proof of Theorem 3.4. The proof is inspired by Yang et al. (2011). For a given design ξ, the information matrix is

$$I_\xi(\theta) = n\sum_{i=1}^{m} w_i\mathbf{F}_i'\mathbf{U}_i\mathbf{F}_i.$$

First of all, define following weights, $r_j=\frac{V_j-x_{ij}}{V_j-U_j}$ such that

$$\begin{aligned} &r_jU_j+(1-r_j)V_j = x_{ij}\\ &r_jU_j^2+(1-r_j)V_j^2 \geq x_{ij}^2\\ &\textit{for}\quad j=1,\ldots,p-1. \end{aligned} \tag{B.5}$$

The first equality is easy to verify, and the second inequality is due to the fact that the function $f(x) = x^2$ is convex. Note that this is exactly the Lemma 1 appears in Yang et al. (2011).

For an arbitrary design point, say $s_i = (x_{i1}, \ldots, x_{i,p-1}, c_i)$, consider the following two design points, $\tilde{s}_{i1} = (U_1, x_{i2}, \ldots, x_{i,p-1}, c_i)$ and $\tilde{s}_{i2} = (V_1, x_{i2}, \ldots, x_{i,p-1}, c_i)$, and their design matrices are $\tilde{\mathbf{F}}_{i1}$ and $\tilde{\mathbf{F}}_{i2}$. Let $\tilde{w}_{i1} = r_1 w_i$ and $\tilde{w}_{i2} = w_i - \tilde{w}_{i1}$, then $w_i \mathbf{F}_i' \mathbf{U}_i \mathbf{F}_i$ and $\sum_{\ell=1}^{2} \tilde{w}_{i\ell} \tilde{\mathbf{F}}_{i\ell}' \tilde{\mathbf{U}}_{i\ell} \tilde{\mathbf{F}}_{i\ell}$ are exactly the same except the first diagonal element in their B3 blocks. Here $\tilde{\mathbf{U}}$ is the matrix $\mathbf{U}$ evaluated at $\tilde{s}_{i\ell}$.

This is true due to two facts. First the (B.5) holds. Second, entries in B1, B2 as well as off-diagonal ones in B3 are linear in x_{i1}, and only the first diagonal components in the B3 block are quadratic in x_{i1}. As a result,

$$w_i \mathbf{F}_i' \mathbf{U}_i \mathbf{F}_i \le \sum_{\ell=1}^{2} \tilde{w}_{i\ell} \tilde{\mathbf{F}}_{i\ell}' \tilde{\mathbf{U}}_{i\ell} \tilde{\mathbf{F}}_{i\ell}.$$

Repeat the procedures until $x_{i,p-1}$, and we have the following

$$w_i \mathbf{F}_i' \mathbf{U}_i \mathbf{F}_i \le \sum_{\ell=1}^{2^{p-1}} \tilde{w}_{i\ell} \tilde{\mathbf{F}}_{i\ell}' \tilde{\mathbf{U}}_{i\ell} \tilde{\mathbf{F}}_{i\ell}. \tag{B.6}$$

Note that the right hand side of (B.6) only depends on c_i, and they have the same set of linear independent nonconstant functions. Then following Theorem 3.3, there exist two points $\tilde{c}_{i1}$ and $\tilde{c}_{i2}$ such that

$$I_\xi(\boldsymbol{\theta}) \le \sum_{i=1}^{m} \sum_{\ell=1}^{2^{p-1}} \tilde{w}_{i\ell} \tilde{\mathbf{F}}_{i\ell}' \tilde{\mathbf{U}}_i \tilde{\mathbf{F}}_{i\ell} \le \sum_{i=1}^{2} \tilde{w}_i \sum_{\ell=1}^{2^{p-1}} \tilde{w}_{i\ell} \tilde{\mathbf{F}}_{i\ell}' \tilde{\mathbf{U}}_{i\ell} \tilde{\mathbf{F}}_{i\ell}$$

That is the complete class consists of two equivalent classes of 2^p design points in total. □

References

Agresti, A., 2013. Categorical Data Analysis, third ed. Wiley Series in Probability and Statistics, Wiley, URL https://books.google.com/books?id=UOrr47-2oisC.

Bu, X., Majumdar, D., Yang, J., 2019. D-optimal designs for multinomial logistic models. arXiv:1707.03063v3.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statist. Sci. 10 (3), 273–304. http://dx.doi.org/10.1214/ss/1177009939, URL http://projecteuclid.org/euclid.ss/1177009939.

Dette, H., Melas, V.B., 2011. A note on the de la Garza phenomenon for locally optimal designs. Ann. Statist. 39 (2), 1266–1281. http://dx.doi.org/10.1214/11-AOS875, URL http://projecteuclid.org/euclid.aos/1304947050.

Dette, H., Schorning, K., 2013. Complete classes of designs for nonlinear regression models and principal representations of moment spaces. Ann. Statist. 41 (3), 1260–1267. http://dx.doi.org/10.1214/13-AOS1108.

Ford, I., Torsney, B., Wu, C.-F.J., 1992. The use of a canonical form in the construction of locally optimal designs for nonlinear problems. J. R. Stat. Soc. Ser. B Stat. Methodol. 54 (2), 569–583, URL http://links.jstor.org/sici?sici=0035-9246(1992)54:2<569:TUOACF>2.0.CO;2-8&origin=MSN.

Glonek, G.F.V., McCullagh, P., 1995. Multivariate logistic-models. J. R. Stat. Soc. Ser. B 57 (3), 533–546.

Liu, I., Agresti, A., 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. Test 14 (1), 1–73. http://dx.doi.org/10.1007/BF02595397, URL http://link.springer.com/10.1007/BF02595397.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models, second ed. http://dx.doi.org/10.1007/978-1-4899-3242-6, arXiv:arXiv:1011.1669v3.

Melas, V., 2006. Functional Approach to Optimal Experimental Design, Vol. 184. Springer, http://dx.doi.org/10.1007/0-387-31610-8.

Perevozskaya, I., Rosenberger, W., Haines, L., 2003. Optimal design for the proportional odds model. Canad. J. Statist. 31 (2), 225–235. http://dx.doi.org/10.2307/3316068, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-0242595932&partnerID=40.

Schacter, L., Birkhofer, M., Carter, S., Canetta, R., Hellmann, S., Onetto, N., Weil, C.B., Rozencweig, M.R., 1997. Anticancer drugs. In: O'Grady, J., Joubert, P.H. (Eds.), Handbook of Phase I/II Clinical Drug Trials, first ed. CRC Press, Boca Raton, pp. 523–534.

Wang, H., Yang, M., Stufken, J., 2018. Information-based optimal subdata selection for big data linear regression. J. Amer. Statist. Assoc. 1–13. http://dx.doi.org/10.1080/01621459.2017.1408468.

Yang, M., 2010. On the de la Garza phenomenon. Ann. Statist. 38 (4), 2499–2524. http://dx.doi.org/10.1214/09-AOS787.

Yang, M., Biedermann, S., Tang, E., 2013. On optimal designs for nonlinear models: a general and efficient algorithm. J. Amer. Statist. Assoc. 108 (504), 1411–1420. http://dx.doi.org/10.1080/01621459.2013.806268.

Yang, M., Stufken, J., 2009. Support points of locally optimal designs for nonlinear models with two parameters. Ann. Statist. 37 (1), 518–541. http://dx.doi.org/10.1214/07-AOS560, arXiv:arXiv:0903.0728v1.

Yang, M., Stufken, J., 2012. Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. Ann. Statist. 40 (3), 1665–1681. http://dx.doi.org/10.1214/12-AOS992.

Yang, J., Tong, L., Mandal, A., 2017. D-optimal designs with ordered categorical data. Statist. Sinica 27, 1879–1902. http://dx.doi.org/10.5705/ss.202016.0210, arXiv:1502.05990.

Yang, M., Zhang, B., Huang, S., 2011. Optimal designs for generalized linear models with multiple design variables. Statist. Sinica 21 (3), 1415–1430. http://dx.doi.org/10.5705/ss.2009.115, URL http://www.jstor.org/stable/24309568, http://www3.stat.sinica.edu.tw/statistica/j21n3/J21N318/J21N318.html.

Zocchi, S., Atkinson, A., 1999. Optimum experimental designs for multinomial logistic models. Biometrics 55 (2), 437–444. http://dx.doi.org/10.2307/2533789.