Statistica Sinica **31** (2021), 1125-1143
doi:https://doi.org/10.5705/ss.202018.0100

PARAMETER REDUNDANCY AND THE EXISTENCE OF MAXIMUM LIKELIHOOD ESTIMATES IN LOG-LINEAR MODELS

Serveh Sharifi Far[1], Michail Papathomas[2] and Ruth King[1]

[1] *University of Edinburgh and* [2] *University of St Andrews*

Abstract: Log-linear models are typically fitted to contingency table data to describe and identify the relationships between categorical variables. However, these data may include observed zero cell entries, which can have an adverse effect on the estimability of the parameters, owing to parameter redundancy. We describe a general approach to determining whether a given log-linear model is parameter-redundant for a pattern of observed zeros in the table, prior to fitting the model to the data. We derive the estimable parameters or the functions of the parameters, and explain how to reduce the unidentifiable model to an identifiable model. Parameter-redundant models have a flat ridge in their likelihood function. We explain when this ridge imposes additional parameter constraints on the model, which can lead to unique maximum likelihood estimates for parameters that otherwise would not have been estimable. In contrast to other frameworks, the proposed approach informs on those constraints, elucidating the model being fitted.

Key words and phrases: Contingency table, extended maximum likelihood estimate, identifiability, parameter redundancy, sampling zero.

1. Introduction

Observations from multiple categorical random variables can be cross-classified according to combinations of the variables' levels. This type of data is often displayed in a contingency table, where each cell count is the number of subjects with a given cross-classification. Log-linear models are typically fitted to such tables; examples of applications are given in Agresti (2002), Bishop, Fienberg and Holland (1975), and McCullagh and Nelder (1989).

Zero cell counts can have an adverse effect on the estimability of log-linear model parameters. Zero entries are of two main types: structural zeros, and sampling zeros. If the expectation and variance of a cell count are zero, then the entry is a structural zero. A sampling zero is an observed zero entry in a cell with a positive expectation. In this study, we examine how zero cell entries

Corresponding author: Serveh Sharifi Far, School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK. E-mail: serveh.sharifi@ed.ac.uk.

influence the estimability of log-linear model parameters, addressed with respect to parameter redundancy.

A model is not identifiable if two different sets of parameter values generate the same model for the data, which often happens when a model is overparametrized. This cause of non-identifiability is termed parameter redundancy (Catchpole and Morgan (1997)). A parameter-redundant model can be rearranged as a function of a smaller set of parameters, which are themselves functions of the initial parameters. Parameter-redundant models have a flat ridge in their likelihood surface, which precludes unique maximum likelihood (ML) estimates for some of the parameters (Catchpole and Morgan (1997)). For a log-linear parameter-redundant model, undefined or large standard errors for non-estimable parameters are often reported using numerical optimization methods. An overview of identifiability and parameter redundancy is given by Catchpole and Morgan (1997) and Catchpole, Morgan and Freeman (1998). Cole, Morgan and Titterington (2010) provide several ecological examples on this topic. Identifiability is crucial when exploring complex associations between factors, because interaction terms quickly become nonestimable in the presence of zero cell counts. Therefore, we require methods that identify the highest level of interaction complexity that can be explored for a given data set.

We develop a method for detecting parameter redundancy in log-linear models in the presence of sampling zero observations. We derive the estimable parameters and combinations of parameters, and show how a parameter-redundant model can be reduced to a nonredundant one that is also identifiable. We refer to the proposed method as the "parameter redundancy" approach. In the presence of structural zeros, the corresponding cells are omitted from the modeling and analysis because they are associated with cross-classifications that cannot be observed.

A comprehensive study of log-linear models for contingency tables was performed by Haberman (1973), who proves that ML estimates of the model parameters are unique when they exist, and provides a necessary and sufficient condition for the existence of cell mean estimates in the presence of zero cell entries. This was studied further by Brown and Fuchs (1983), who compared iterative methods, and by Lauritzen (1996), who examined a polyhedral and graphical model framework. A polyhedral version of Haberman's condition for the existence of the maximum likelihood estimator (MLE) is provided by Eriksson et al. (2006). The estimability of parameters under a nonexistent MLE, within the extended exponential families, is studied by Fienberg and Rinaldo (2012a), and is developed for higher-dimensional problems by Wang, Rauhy and Massam (2019). We refer to these developments collectively as the "existence of the maximum likelihood estimator" (EMLE) framework. The method demonstrates that some of the parameters cannot be estimated when the MLE does not exist. However, an extended estimator, in which some of the elements of the estimated cell mean vector are zero, always exists (Eriksson et al. (2006)). In this case, it is possible to reduce the model and estimate a subset of the initial parameters.

We compare the proposed parameter redundancy approach with the EMLE method. The reduced models obtained by the two methods may differ in terms of their parametrization, but the parameter redundancy approach provides a reparametrization that retains the original interpretation of the parameters. This is because this method provides estimable parameters and linear combinations of parameters instead of just an estimable subset of the model's initial parameters. The parameter redundancy approach also reveals additional constraints imposed by the likelihood function on some parameter-redundant models. Standard statistical software packages report parameter estimates for such a model without informing on the additional implied constraints.

Section 1.1 introduces the necessary notation. Section 2 describes the parameter-redundant model and the proposed adaptation to log-linear models, including examples and a study on saturated log-linear models. We also show when additional constraints enable us to determine unique ML estimates for the additional parameters, thus specifying the model that is, in fact, fitted to the sparse table. In Section 3, the EMLE framework is reviewed, and in Section 4, the two approaches are compared using illustrative examples. Section 5 concludes the paper.

1.1. Log-linear models for contingency tables

Adopting the notation in Overstall and King (2014), let $V = \{V_1, \ldots, V_m\}$ denote a set of m categorical variables, where the jth variable has l_j levels. The corresponding contingency table has $n = \prod_{j=1}^{m} l_j$ cells. Let $\mathbf{y}$ denote an $n \times 1$ vector corresponding to the observed cell counts. Each element of $\mathbf{y}$ is denoted by $y_{\mathbf{i}}$, for $\mathbf{i} = (i_1 \ldots i_m)$, such that $0 \leqslant i_j \leqslant l_j - 1$ and $j = 1, \ldots, m$. Here, $\mathbf{i}$ identifies the combinations of variable levels that cross-classify the given cell. We define L as the set of all n cross-classifications, such that $L = \otimes_{j=1}^{m} [l_j]$, where $[l_j] = \{0, 1, \ldots, l_j - 1\}$. Then, $N = \sum_{\mathbf{i} \in L} y_{\mathbf{i}}$ denotes the sum of all cell counts. The $y_{\mathbf{i}}$ are assumed to be observations from independent Poisson random variables, $Y_{\mathbf{i}}$, such that $\mu_{\mathbf{i}} = E(Y_{\mathbf{i}})$. Let $\mathcal{E}$ denote a set of subsets of V. By adapting the notation of Johndrow, Bhattacharya and Dunson (2017), the log-linear model assumes the form

$$m_{\mathbf{i}} = \log \mu_{\mathbf{i}} = \sum_{e \in \mathcal{E}} \theta^e(\mathbf{i}), \tag{1.1}$$

where $\theta^e(\mathbf{i}) \in \mathcal{R}$ denotes the main effects or the interactions between the variables in e corresponding to the levels in $\mathbf{i}$. The summation is over all members of $\mathcal{E}$, which could be the set of all subsets of the variables (for a saturated model) or a set of desirable subsets (for a smaller model). By convention, θ corresponds to $e = \emptyset$, such that when the set $\mathcal{E}$ contains $e = \emptyset$, there is an intercept θ in the model. To allow for the existence of unique parameter estimates, corner point constraints are applied; as such, parameters that incorporate the lowest level of a variable are set to zero. To clarify the notation, consider this minimal example. Assume two categorical variables, $V = \{X, Y\}$, with $l_1 = l_2 = 2$ levels. Then, the number of cells in the $l_1 \times l_2$ table is four and $L = \{00, 10, 01, 11\}$. The set of subsets of V, $\mathcal{E} = \{\emptyset, \{X\}, \{Y\}\}$ constructs the following independence log-linear model, shown as model (X, Y),

$$\begin{aligned} m_{00} &= \log \mu_{00} = \theta, & m_{10} &= \log \mu_{10} = \theta + \theta_1^X, \\ m_{01} &= \log \mu_{01} = \theta + \theta_1^Y, & m_{11} &= \log \mu_{11} = \theta + \theta_1^X + \theta_1^Y. \end{aligned}$$

Alternatively to (1.1), for p parameters, we can write $\mathbf{m}_{n\times 1} = \log \boldsymbol{\mu}_{n\times 1} = A_{n\times p}\boldsymbol{\theta}_{p\times 1}$, where A is a full-rank design matrix with elements $\{0, 1\}$. Therefore, this model can be written as below; note that the subscript indices of the parameters are removed because there are only two possible variable levels:

$$\begin{bmatrix} \log \mu_{00} \\ \log \mu_{10} \\ \log \mu_{01} \\ \log \mu_{11} \end{bmatrix} = \begin{bmatrix} 1\,0\,0 \\ 1\,1\,0 \\ 1\,0\,1 \\ 1\,1\,1 \end{bmatrix} \begin{bmatrix} \theta \\ \theta^X \\ \theta^Y \end{bmatrix}.$$

For a model fitted to an l^m table (with m variables, each classified in l levels), an alternative notation can be used to denote the cell counts in (1.1) that sets a one-to-one correspondence between the elements of L and the integers, $i = 1, \ldots, l^m$, as follows:

$$\mathbf{i} = (i_1 \ldots i_m) = i_1 l^0 + i_2 l^1 + \cdots + i_{m-1} l^{m-2} + i_m l^{m-1} + 1. \tag{1.2}$$

Thus, in the aforementioned example, the elements in $L = \{00, 10, 01, 11\}$ correspond to $\{1, 2, 3, 4\}$, respectively.

2. The Parameter Redundancy Approach

2.1. The derivative method

Goodman (1974) first used a derivative approach to detect identifiability in latent structure models and m-way contingency tables. The generic approach for the exponential family of distributions that we summarize here was presented by Catchpole and Morgan (1997) and Catchpole, Morgan and Freeman (1998), but was also developed independently by Chappell and Gunn (1998) and Evans and Chappell (2000) for compartmental models.

The mean vector $\boldsymbol{\mu} = E(\mathbf{Y})$ of observations from a distribution that belongs to the exponential family of distributions is expressible as a function of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The derivative matrix $D(\boldsymbol{\theta})$, which describes the relationship between $\boldsymbol{\mu}$ (or a monotonic function of it) and $\boldsymbol{\theta}$, has elements

$$D_{si}(\boldsymbol{\theta}) = \frac{\partial \mu_i}{\partial \theta_s}, \quad s = 1, \ldots, p, \quad i = 1, \ldots, n. \tag{2.1}$$

Theorem 1 of Catchpole and Morgan (1997) states that the model that relates $\boldsymbol{\mu}$ to $\boldsymbol{\theta}$ is parameter-redundant if and only if the derivative matrix is symbolically rank deficient; that is, there exists a nonzero vector $\boldsymbol{\alpha}(\boldsymbol{\theta})$, such that for all $\boldsymbol{\theta}$,

$$\boldsymbol{\alpha}(\boldsymbol{\theta})^{\mathsf{T}} D(\boldsymbol{\theta}) = \mathbf{0}. \tag{2.2}$$

As an alternative, Cole, Morgan and Titterington (2010) construct a derivative matrix by differentiating an "exhaustive summary" of the model. An exhaustive summary is a vector of parameter combinations that uniquely defines the model.

The rank of the derivative matrix, r, is the number of estimable parameters and combinations of parameters. The model deficiency is defined as $d = p - r$, which is the number of linearly independent $\boldsymbol{\alpha}(\boldsymbol{\theta})$ vectors, labeled as $\boldsymbol{\alpha}_j(\boldsymbol{\theta})$, for $j = 1, \ldots, d$. Any elements of these vectors that are zero for all j correspond to directly estimable parameters (Catchpole, Morgan and Freeman (1998)). To find the estimable combinations of parameters, we must solve the auxiliary equations of the following system of linear first-order partial differential equations:

$$\sum_{s=1}^{p} \alpha_{sj} \frac{\partial f}{\partial \theta_s} = 0, \quad j = 1, \ldots, d \tag{2.3}$$

(Catchpole, Morgan and Freeman (1998)). The solution can be obtained using software such as `Maple`, which allows symbolic computations.

2.2. Parameter redundancy for log-linear models

Parameter redundancy can be the result of the model structure or a lack of data (Catchpole and Morgan (2001); Cole, Morgan and Titterington (2010)), with the latter type referred to as "extrinsic" parameter redundancy (Gimenez et al. (2004)). Model (1.1) is constructed so that it is not over-parametrized owing to its structure. To detect extrinsic parameter redundancy for a log-linear model, we adjust the derivative matrix elements (2.1) using $y_i \log \mu_i$ as a monotonic function of μ_i, such that

$$D_{si} = \frac{\partial y_i \log \mu_i}{\partial \theta_s}, \quad s = 1, \ldots, p, \quad i = 1, \ldots, n. \tag{2.4}$$

In effect, each sampling zero turns a column of the derivative matrix to zero, and may decrease the rank of the derivative matrix.

If the rank of the derivative matrix is smaller than p, the model is parameter redundant. Finding all estimable parameters and estimable combinations of parameters identifies which cell means are estimable. The vector of estimable quantities ($\boldsymbol{\theta}'$) and the vector of estimable cell means ($\boldsymbol{\mu}'$) specify a reduced model via a smaller design matrix (A'). The reduced model is full rank with rank r, and its degrees of freedom is the number of estimable cells means minus r.

To clarify the notation, consider the independence log-linear model (X, Y) for a 2×2 table. The derivative matrix (2.4) for observations $\mathbf{y}^\mathsf{T} = (y_1, y_2, y_3, y_4) = (y_{00}, y_{10}, y_{01}, y_{11})$ and parameters $\boldsymbol{\theta}^\mathsf{T} = (\theta, \theta^X, \theta^Y)$ is,

$$D = \left[\frac{\partial y_i \log \mu_i}{\partial \theta_s}\right] = \left[\begin{array}{c|cccc} & \mu_{00} & \mu_{10} & \mu_{01} & \mu_{11} \\ \hline \theta & y_1 & y_2 & y_3 & y_4 \\ \theta^X & 0 & y_2 & 0 & y_4 \\ \theta^Y & 0 & 0 & y_3 & y_4 \end{array}\right], \quad s = 1, 2, 3, \quad i = 1, 2, 3, 4.$$

Now, for example, assume that $y_1 = y_2 = 0$. Then, $r = 2$, $d = 1$, and $\boldsymbol{\alpha}^\mathsf{T} = (1, 0, -1)$. Equation (2.3) is $\partial f / \partial \theta - \partial f / \partial \theta^Y = 0$, and solving it gives the estimable parameters $\boldsymbol{\theta}'^\mathsf{T} = (\theta^X, \theta + \theta^Y)$. Thus only $\boldsymbol{\mu}'^\mathsf{T} = (\mu_{01}, \mu_{11})$ are estimable. Therefore, the reduced design matrix A' is 2×2 with two rows $[(0, 1), (1, 1)]$.

Alternative approaches for investigating identifiability are not suitable in the context of Poisson log-linear models for contingency tables. Specifically, using the log-likelihood function elements as exhaustive summaries is a common option when creating the derivative matrix (Cole, Morgan and Titterington (2010)). Similarly, Catchpole and Morgan (2001) use the score vector of a multinomial log-linear model to assess the effect of missing data on the model redundancy.

Table 1. Observations in a 3^3 contingency table.

0	y_4	y_7	y_{10}	y_{13}	y_{16}	0	y_{22}	0^*
0	y_5	y_8	y_{11}	y_{14}	0^*	0	y_{23}	y_{26}
y_3	y_6	y_9	y_{12}	0	0	y_{21}	y_{24}	y_{27}

In addition, using the information matrix instead of a derivative matrix is an alternative method for detecting non-identifiability (Rothenberg (1971)). However, these approaches do not necessarily show the rank deficiency caused by the zero cell counts for a Poisson log-linear model. The next two examples further illustrate the use of the parameter redundancy method.

Example 1. The data pattern in Table 1, taken from Fienberg and Rinaldo (2012a), describes cell counts for variables X (rows), Y (columns), and Z (layers), with three levels $(0, 1, 2)$ for each. Eight cell counts are observed as sampling zeros. All other cell counts are positive Poisson observations, numbered according to (1.2). We fit the hierarchical model (XY, XZ, YZ), which can be shown as $\log \boldsymbol{\mu}_{27\times 1} = A_{27\times 19}\boldsymbol{\theta}_{19\times 1}$, with parameters,

$$\boldsymbol{\theta}^{\mathsf{T}} = (\theta, \theta_1^X, \theta_2^X, \theta_1^Y, \theta_2^Y, \theta_1^Z, \theta_2^Z, \theta_{11}^{XY}, \theta_{21}^{XY}, \theta_{12}^{XY}, \theta_{22}^{XY}, \theta_{11}^{YZ}, \theta_{21}^{YZ}, \theta_{12}^{YZ}, \theta_{22}^{YZ}, \theta_{11}^{XZ}, \theta_{21}^{XZ}, \theta_{12}^{XZ}, \theta_{22}^{XZ}).$$

The matrix form of this model is given in the Supplementary Material.

The rank of the derivative matrix in accordance with (2.4) is 18; that is, there are only 18 estimable parameters or combinations of them. Therefore, $d = 19 - 18 = 1$, and the $\boldsymbol{\alpha}$ that satisfies (2.2) is $\boldsymbol{\alpha}^{\mathsf{T}} = (1, 0, -1, -1, -1, -1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0)$. Solving (2.3) gives the estimable quantities as

$$\boldsymbol{\theta}'^{\mathsf{T}} = (\theta_1^X, \theta + \theta_2^X, \theta + \theta_1^Y, \theta + \theta_2^Y, \theta + \theta_1^Z, \theta_2^Z, \theta_{11}^{XY}, -\theta + \theta_{21}^{XY}, \theta_{12}^{XY}, -\theta + \theta_{22}^{XY}, -\theta + \theta_{11}^{YZ}, -\theta + \theta_{21}^{YZ}, \theta_{12}^{YZ}, \theta_{22}^{YZ}, \theta_{11}^{XZ}, -\theta + \theta_{21}^{XZ}, \theta_{12}^{XZ}, \theta_{22}^{XZ}).$$

The elements of $\boldsymbol{\theta}'$ determine that 21 out of 27 cell means are estimable, including cells 17 and 25, indicated in Table 1 with asterisks. Therefore, for this model and this specified pattern of zeros, the cell means $1, 2, 15, 18, 19, 20$ are not estimable; thus, we remove the corresponding cells from the model. This is equivalent to assuming that these observations are structural zeros. Considering $\boldsymbol{\theta}'$ and the 21 estimable cell means, the reduced model with three degrees of freedom is $\log \boldsymbol{\mu}'_{21\times 1} = A'_{21\times 18}\boldsymbol{\theta}'_{18\times 1}$, given in the Supplementary Material.

Example 2. Hung et al. (2008) performed a genome-wide association study of lung cancer by studying 500 single nucleotide polymorphisms (SNP). Each SNP is categorized at levels 0, 1, and 2 to identify the number of minor alleles. Papathomas et al. (2012) selected 50 of these SNPs by applying a profile regression. We further select five SNPs (as representatives of uncorrelated groups of SNPs): `rs7748167_C` (A), `rs4975616_G` (B), `rs6803988_T` (C), `rs11128775_G` (D), and `rs9306859_A` (E).

A crucial variable in this study describes the presence or absence of cancer in each of the individuals. Adding this variable (F) creates a $3^5 \times 2^1$ contingency table with 486 cells. We consider fitting a log-linear model with main effects and first-order interactions. This table has 298 zero cell counts, and the derivative matrix has rank 59, with $d = 62 - 59 = 3$. After solving the partial differential equations for the three $\boldsymbol{\alpha}$ vectors, the 59 estimable parameters are obtained; see in the Supplementary Material.

Only three parameters ($\theta_{22}^{AD}, \theta_{22}^{AE}$, and θ_{22}^{DE}) are not estimable. The estimable parameters make 360 out of 486 cell means estimable, and the reduced model is $\log \boldsymbol{\mu}'_{360\times1} = A'_{360\times59}\boldsymbol{\theta}'_{59\times1}$, with degrees of freedom $360 - 59 = 301$. In this model, the presence of cancer has a significant positive interaction with level 1 of variables A and D, and a significant negative interaction with level 1 of C and E and with level 2 of B, C, and E.

2.3. Parameter redundancy for a saturated log-linear model

We provide general results on parameter redundancy for a saturated log-linear model fitted to an l^m contingency table, and determine which parameters become nonestimable after observing a zero cell count. Example S1 in the Supplementary Material illustrates the proposed approach, and shows that a saturated log-linear model is always full rank when all the cell counts are positive.

Definition 1. For a saturated log-linear model, we define *the parameter corresponding to the cell with count* y_i, for $i = 1, \ldots, n$ (according to (1.2)), as that with the maximum number of variables in its superscript, within the set of all parameters in $\log \mu_i = A_{(i)}\boldsymbol{\theta}$, where $A_{(i)}$ is the ith row of A.

For example, for a 3^3 contingency table with variables $\{X, Y, Z\}$, the parameter corresponding to observation y_{201} (or y_{12}, according to the ordering given by (1.2)) is θ_{21}^{XZ}.

Definition 2. For a given log-linear model parameter, *the parameters associated with a higher-order interaction* are all those specified by including additional variables in the given parameter's superscript.

For example, for the same 3^3 table, the parameters associated with a higher order interaction, given θ_{21}^{XZ}, are θ_{211}^{XYZ} and θ_{221}^{XYZ}.

The following theorem determines which model parameters become nonestimable as a result of a given zero observation.

Theorem 1. *Assume a saturated Poisson log-linear model fitted to an l^m table with a single zero cell count. If $\exists \boldsymbol{i}, \boldsymbol{i} \in L$, such that $y_{\boldsymbol{i}} = 0$, then the parameter that corresponds to that cell, and all other parameters associated with a higher-order interaction, given that parameter, are nonestimable.*

The proof by induction and examples are given in the Supplementary Material. Note that additional zero cells in the table cannot make previously nonestimable parameters estimable, because the amount of information is further reduced. Then, the set of nonestimable parameters is at least as large as the union of the nonestimable parameters per zero cell. The estimable parameters and linear combinations of them can be derived by solving (2.3).

2.4. The esoteric constraints

The likelihood function of a parameter-redundant model has a flat ridge that is occasionally orthogonal to the axes of some parameters; as such, these associated parameters still have unique ML estimates (Catchpole, Morgan and Freeman (1998)). This is when, in all $\boldsymbol{\alpha}(\boldsymbol{\theta})$, the elements corresponding to these parameters are zero. In addition, for some log-linear parameter-redundant models, maximizing the likelihood function imposes one or more extra constraints on the model parameters, owing to the placement of the likelihood ridge in the parameter space. The extra constraints can make more parameters uniquely estimable compared with those specified by solving the partial differential equations in (2.3). We refer to these extra constraints as "esoteric constraints". Standard statistical software packages do not provide information on these constraints when maximizing the likelihood function; thus, informing on them reveals the log-linear model that is, in fact, being fitted. After detecting a parameter-redundant model, we can check the existence of such constraints, as explained below.

The log-likelihood function of model (1.1) is $l(\boldsymbol{\theta}) = \sum_{\mathbf{i}}(y_{\mathbf{i}} \log \mu_{\mathbf{i}}(\boldsymbol{\theta}) - \mu_{\mathbf{i}}(\boldsymbol{\theta}))$. The corresponding score vector is $\mathbf{U}(\boldsymbol{\theta}) = (\partial l/\partial\theta_1, \ldots, \partial l/\partial\theta_p)^{\mathsf{T}}$, where the partial derivatives for $s = 1, \ldots, p$, are,

$$\frac{\partial l}{\partial \theta_s} = \sum_{\mathbf{i}} \left(\frac{y_{\mathbf{i}}}{\mu_{\mathbf{i}}(\boldsymbol{\theta})} - 1 \right) \frac{\partial \mu_{\mathbf{i}}(\boldsymbol{\theta})}{\partial \theta_s} = \sum_{\mathbf{i}} (y_{\mathbf{i}} - \mu_{\mathbf{i}}(\boldsymbol{\theta})) \frac{\partial \mu_{\mathbf{i}}(\boldsymbol{\theta})}{\partial \theta_s} \frac{1}{\mu_{\mathbf{i}}(\boldsymbol{\theta})}.$$

Therefore, $\mathbf{U}(\boldsymbol{\theta}) = A^{\mathsf{T}}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$. When a model is parameter redundant, there exists at least one $\boldsymbol{\alpha}(\boldsymbol{\theta})$, such that $\boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{\theta})D(\boldsymbol{\theta}) = \mathbf{0}$. If the observations are from a multinomial distribution, it follows that $\boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = 0$, which means the likelihood surface has a completely flat ridge (Theorem 2 of Catchpole and Morgan (1997)). Note that $\boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = 0$ implies that the directional derivative is zero; therefore, the likelihood function is constant in the direction of $\boldsymbol{\alpha}(\boldsymbol{\theta})$. This makes a ridge in the likelihood surface, which is along the curve generated by the direction field $\boldsymbol{\alpha}(\boldsymbol{\theta})$ through any point at which the likelihood is maximized.

For a Poisson log-linear model determined to be parameter-redundant by the derivative matrix in (2.4), we set $\boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = 0$. The constraints that hold this equality for finite values of the model parameters are the esoteric constraints. These extra constraints, along with the estimable quantities in $\boldsymbol{\theta}'$, may make more parameters estimable, and permit one to obtain unique ML estimates for parameters that otherwise would not have been estimable. In addition, reducing the parameter space according to the esoteric constraints and, therefore, removing the flat ridge, can make it possible to uniquely maximize the likelihood. If $\boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})$ cannot be zero with finite θ, then the esoteric constraints do not exist, and some of the θ tend to negative infinity. These constraints do not exist for models described in Theorem 1 and in Examples 1 and 2. A model with an esoteric constraint is given in Example 4.

3. The EMLE for Log-Linear Models

The methods summarized in this section are referred to as the EMLE approach, and will be used in Examples 3 and 4 in Section 4. Refer to Fienberg and Rinaldo (2006, 2012a,b) for further background and details.

Decomposable log-linear models (Agresti (2002)) have an explicit formula for $\hat{\mu}_{\mathbf{i}}$. For these models, the positivity of the minimal sufficient statistics is a necessary and sufficient condition for the existence of the MLE of $\boldsymbol{\mu}$ (Agresti (2002)). For non-decomposable models, $\hat{\mu}_{\mathbf{i}}$ does not have a closed form, and is calculated using iterative methods. In this case, the positivity of the sufficient table marginals is still necessary for the existence of the estimator, but is no longer a sufficient condition.

A condition for the existence of the MLE of $\mathbf{m}$ in a hierarchical log-linear model, regardless of the presence of positive or zero table marginals, is provided by Haberman (1973). Assume $\mathcal{M}$ is a p-dimensional linear manifold contained in $\mathcal{R}^{|L|}$, and

$$\mathcal{M}^{\perp} = \left\{ \mathbf{x} \in \mathcal{R}^{|L|} : (\mathbf{x}, \mathbf{m}) = \mathbf{x}^{\mathsf{T}}\mathbf{m} = 0, \forall \mathbf{m} \in \mathcal{M} \right\}. \tag{3.1}$$

Then, Theorem 3.2 of Haberman (1973) states that a necessary and sufficient condition such that the MLE $\hat{\mathbf{m}}$ of $\mathbf{m}$ exists is that there is a $\boldsymbol{\delta} \in \mathcal{M}^{\perp}$, such that $y_{\mathbf{i}} + \delta_{\mathbf{i}} > 0$, for every $\mathbf{i} \in L$. Here, $\boldsymbol{\mu}$ in $\mathbf{m} = \log \boldsymbol{\mu}$ is assumed to be positive.

The theorem specifies, for any pattern of zeros in the table, whether the MLE of the cell means exists. In the extended ML estimate case, a cell mean estimate could be $\hat{\mu}_{\mathbf{i}} = 0$, but its log transformation is not defined; then, estimates of some corresponding θ parameters tend to infinity (Haberman (1974)).

A polyhedral version of Haberman's necessary and sufficient condition states that under any sampling design, the MLE of $\mathbf{m}$ exists if and only if the vector of observed marginals, $\mathbf{t} = A^{\mathsf{T}}\mathbf{y}$, lies in the relative interior of the marginal of the polyhedral cone (Eriksson et al. (2006)). The polyhedral cone, generated by spanning columns of A with rank p, is defined as,

$$C_A = \left\{ \mathbf{t} : \mathbf{t} = A^{\mathsf{T}}\mathbf{y}, \mathbf{y} \in \mathcal{R}^{|L|}_{\geqslant 0} \right\}. \tag{3.2}$$

The MLE does not exist if and only if the vector of marginals lies on a facet or a facial set of the marginal cone (Fienberg and Rinaldo (2006)). In other words, the estimator does not exist if and only if the vector of marginals belongs to the relative interior of some proper face, F, of the marginal cone. A face of the marginal cone is defined as a set, $F = \{\mathbf{t} \in C_A : (\mathbf{t}, \boldsymbol{\zeta}) = 0\}$, for some $\boldsymbol{\zeta} \in \mathcal{R}^p$, such that $(\mathbf{t}, \boldsymbol{\zeta}) \geqslant 0$, for all $\mathbf{t} \in C_A$, with $(\mathbf{t}, \boldsymbol{\zeta})$ representing the inner product. The facial set $\mathcal{F}$ is a set of cell indices of the rows of A, the conic hull of which is precisely F. For any design matrix A for $\mathcal{M}$, $\mathcal{F} \subseteq L$ is a facial set of F if there exists some $\boldsymbol{\zeta} \in \mathcal{R}^p$, such that

$$\begin{aligned} (A_{(i)}, \boldsymbol{\zeta}) &= 0, \quad \text{if} \quad i \in \mathcal{F}, \\ (A_{(i)}, \boldsymbol{\zeta}) &> 0, \quad \text{if} \quad i \in \mathcal{F}^c, \end{aligned} \tag{3.3}$$

where $\mathcal{F}^c = L - \mathcal{F}$ is the co-facial set of F (Fienberg and Rinaldo (2012a)). If such $\boldsymbol{\zeta}$ and $\mathcal{F}$ exist, the MLE does not exist, and only the cell means corresponding to members of $\mathcal{F}$ are estimable. The nonestimable cells in $\mathcal{F}^c$ are treated as structural zeros, and are omitted from the model. An estimable subset of model parameters could be determined by finding $A_{\mathcal{F}}$, the matrix containing rows from A with coordinates in $\mathcal{F}$. Then $A_{\mathcal{F}}$, which is an $|\mathcal{F}| \times p$ design matrix with rank p_F, is reduced to full rank $A^*_{\mathcal{F}}$, with dimensions $|\mathcal{F}| \times p_F$. By implementing this reduced design matrix, the log-likelihood function is strictly concave with a unique maximizer. Then, the extended MLE is

$$\hat{\boldsymbol{\theta}}^e = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathcal{R}^{p_F}} l_{\mathcal{F}}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathcal{R}^{p_F}} \mathbf{t}_F^{\mathsf{T}}\boldsymbol{\theta} - \mathbf{1}^{\mathsf{T}} \exp(A^*_{\mathcal{F}}\boldsymbol{\theta}),$$

where $\mathbf{t}_F = (A^*_{\mathcal{F}})^{\mathsf{T}}\mathbf{y}_{\mathcal{F}}$, and the extended MLE of the cell mean vector is $\hat{\mathbf{m}}^e = \exp(A^*_{\mathcal{F}}\hat{\boldsymbol{\theta}}^e)$ (Fienberg and Rinaldo (2012b)).

Another way to define the facial set is by considering the sub-matrices A_+ and A_0, obtained from A. These are formed by the rows of A, indexed by $L_+ = \{i : y_i \neq 0\}$ and $L_0 = \{i : y_i = 0\}$, respectively. The vector of marginals belongs to the relative interior of some proper face of the marginal cone if and only if $\mathcal{F}^c \subseteq L_0$. This is equivalent to the existence of a vector $\boldsymbol{\zeta}$ satisfying the following three conditions (Fienberg and Rinaldo (2012b)):

$$a.\ A_+\boldsymbol{\zeta} = \mathbf{0}, \tag{3.4}$$

$$b.\ A_0\boldsymbol{\zeta} \gneq \mathbf{0},$$

c. The set $\{i : (A\boldsymbol{\zeta})_{(i)} \neq 0\}$ has maximal cardinality among all sets of $\{i : (A\mathbf{x})_{(i)} \neq 0\}$, with $A\mathbf{x} \gneq 0$, for $\mathbf{x}$ that satisfies the first two conditions.

In (3.3) and (3.4), the inequality signs can be changed to less than zero without loss of generality. Using $\gneq 0$, we describe a nonnegative vector with at least one element greater than zero. In conclusion, if $\text{rank}(A_+) = \text{rank}(A)$, then the MLE exists, because no vector $\boldsymbol{\zeta}$ exists and $\mathcal{F}^c = \emptyset$. If $\text{rank}(A_+) < \text{rank}(A)$, the MLE may still exist; in this case, we should search for a facial set.

The degrees of freedom for the reduced model is $|\mathcal{F}| - \text{rank}(A^*_{\mathcal{F}})$, which is the number of estimable cell means minus the number of estimable model parameters (Fienberg and Rinaldo (2012a)). Fienberg and Rinaldo (2012b) provide computational algorithms for detecting the existence of the MLE and deriving the co-facial set by converting these methods into linear and nonlinear optimization problems. However, those algorithms are inefficient for a model with a large number of variables (Wang, Rauhy and Massam (2019)). The `R` packages `eMLEloglin` and `SparseMSE` use the EMLE approach to fit log-linear models (Chan, Silverman and Vincent (2020); Friedlander (2016)).

4. Comparison of the EMLE and Parameter Redundancy Approaches

The two approaches described in Sections 2 and 3 can be used to check the identifiability of a log-linear model fitted to a sparse table. We compare them, and summarize the comparison in the following three possible cases:

i. Within the EMLE framework, the MLE exists when the co-facial set, as defined in (3.3), is null. This is equivalent to the parameter redundancy outcome in which the model is not parameter-redundant.

ii. When there are facial and co-facial sets, as defined in (3.3), the MLE of $\boldsymbol{\mu}$ does not exist, and some zero cells are treated as structural zeros. In the parameter redundancy approach, this is equivalent to having $\boldsymbol{\alpha}^\mathsf{T} D = \mathbf{0}$

and no esoteric constraints determined by $\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{U}(\boldsymbol{\theta}) = 0$. In practice, for such a model, the determinant of the information matrix and at least one of its eigenvalues are very close to zero, considering numerical approximations and rounding errors.

iii. If there is no co-facial set, as described in (3.3), then the MLE exists. This is equivalent to the parameter redundancy outcome in which the model is parameter-redundant, with at least one esoteric constraint that allows one to uniquely estimate the model parameters.

The next theorem explains a link between the EMLE method and the parameter redundancy approach through the score vector $\mathbf{U}(\boldsymbol{\theta})$.

Theorem 2. *For a parameter-redundant model, the MLE of $\boldsymbol{\mu}$ does not exist if and only if one or more $\boldsymbol{\alpha}_j$ vectors, for $j = 1, \ldots, d$, do not satisfy $\boldsymbol{\alpha}_j^{\mathsf{T}}(\boldsymbol{\theta})\, \boldsymbol{U}(\boldsymbol{\theta}) = 0$ for finite elements of $\boldsymbol{\theta}$.*

The proof is given in the Appendix.

We use two examples to illustrate the similarities and differences between the two approaches. Example 3 shows a parameter-redundant model without any possible additional esoteric constraints (comparison case ii). The two reduced models found by the respective approaches have different reparametrizations of $\boldsymbol{\theta}$, although the ML estimates of the estimable cell means are identical. The parameters in the reduced model obtained using parameter redundancy have the same interpretation as those in the initial model, in terms of variable interactions. Example 4 presents a model that is parameter-redundant and its MLE does exist (comparison case iii). This model has an esoteric constraint, extracted using the parameter redundancy approach, that makes all parameters estimable. This approach allows us to consider two possible ways to address the model's redundancy. First, we can reduce the model to a smaller, saturated, nonredundant model. Second we can adopt the esoteric constraint and estimate all parameters, which is equivalent to using numerical methods, such as the "iteratively reweighted least squares" method to maximize the likelihood.

Example 3. We fit model (4.1), which can be shown as (XY, XZ, YZ), to the contingency table in Table 2(a).

$$\log \mu_{ijk} = \theta + \theta_i^X + \theta_j^Y + \theta_k^Z + \theta_{ij}^{XY} + \theta_{ik}^{XZ} + \theta_{jk}^{YZ}, \quad i, j, k = \{0, 1\}^2. \tag{4.1}$$

According to (1.2), the vector of cell counts is $\mathbf{y}^{\mathsf{T}} = (y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8) = (y_{000}, y_{100}, y_{010}, y_{110}, y_{001}, y_{101}, y_{011}, y_{111})$. The nonzero cell counts in the table are

Table 2. Observations in two 2^3 contingency tables.

(a)

	$Z=0$		$Z=1$	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
$X=0$	0	y_3	y_5	y_7
$X=1$	y_2	y_4	y_6	0

(b)

	$Z=0$		$Z=1$	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
$X=0$	0	y_3	y_5	y_7
$X=1$	y_2	0	y_6	y_8

assumed to be positive. The parameter vector is shown as $\boldsymbol{\theta}^\mathsf{T} = (\theta, \theta^X, \theta^Y, \theta^{XY}, \theta^Z, \theta^{XZ}, \theta^{YZ})$ because subscripts are superfluous. The model in the form $\log \boldsymbol{\mu}_{8\times1} = A_{8\times7}\boldsymbol{\theta}_{7\times1}$ is given in the Supplementary Material.

We apply the parameter redundancy approach first. The derivative matrix formed using formula (2.4) is given in the Supplementary Material; it has rank six, indicating that $d = 1$. From (2.2), $\boldsymbol{\alpha}^\mathsf{T} = (1, -1, -1, 1, -1, 1, 1)$, and solving (2.3) yields the estimable parameters,

$$\boldsymbol{\theta}'^\mathsf{T} = (\theta + \theta^X, \theta + \theta^Y, -\theta + \theta^{XY}, \theta + \theta^Z, -\theta + \theta^{XZ}, -\theta + \theta^{YZ}).$$

Therefore, all cell means other than μ_{000} (for which $\log \mu_{000} = \theta$) and μ_{111} (for which $\log \mu_{111} = \theta + \theta^X + \theta^Y + \theta^{XY} + \theta^Z + \theta^{XZ} + \theta^{YZ}$) are estimable. No esoteric constraint exists because,

$$\boldsymbol{\alpha}^\mathsf{T}\mathbf{U}(\boldsymbol{\theta}) = y_{000} + y_{111} - e^{\theta} - e^{\theta+\theta^X+\theta^Y+\theta^{XY}+\theta^Z+\theta^{XZ}+\theta^{YZ}} \neq 0,$$

for finite θ. We treat y_{000} and y_{111} as structural zeros, and remove them from the model. Then, we reduce the model to a saturated one with a design matrix of rank six in accordance with the estimable parameters $\boldsymbol{\theta}'$. The reduced nonredundant model is

$$\begin{bmatrix} \log \mu_{100} \\ \log \mu_{010} \\ \log \mu_{110} \\ \log \mu_{001} \\ \log \mu_{101} \\ \log \mu_{011} \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 1\,0\,0\,1\,1\,0 \\ 0\,1\,0\,1\,0\,1 \end{bmatrix} \begin{bmatrix} \theta + \theta^X \\ \theta + \theta^Y \\ -\theta + \theta^{XY} \\ \theta + \theta^Z \\ -\theta + \theta^{XZ} \\ -\theta + \theta^{YZ} \end{bmatrix}.$$

Now, we consider the EMLE method. Model (4.1) has no zero sufficient marginals. However, positive estimates for all the cell means do not exist, according to Haberman's sufficient and necessary condition and the polyhedral condition. To reduce this to an identifiable model, from the polyhedral method and (3.3), we obtain, $\mathcal{F} = \{100, 010, 110, 001, 101, 011\}$, $\mathcal{F}^c = \{000, 111\}$, and $\boldsymbol{\zeta} = (1, -1, -1, 1, -1, 1, 1)$. The design matrix for the reduced model is $A^*_{\mathcal{F}}$, which is an $|\mathcal{F}| \times p_F = 6 \times 6$ matrix, and is found using Proposition 5.1 in

Fienberg and Rinaldo (2012b). The final model is

$$\begin{bmatrix} \log\mu_{100} \\ \log\mu_{010} \\ \log\mu_{110} \\ \log\mu_{001} \\ \log\mu_{101} \\ \log\mu_{011} \end{bmatrix} = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&0&1&0&0&0 \\ 1&1&1&1&0&0 \\ 1&0&0&0&1&0 \\ 1&1&0&0&1&1 \\ 1&0&1&0&1&0 \end{bmatrix} \begin{bmatrix} \theta \\ \theta^X \\ \theta^Y \\ \theta^{XY} \\ \theta^Z \\ \theta^{XZ} \end{bmatrix}.$$

The estimable cell means are the same as those derived using the parameter redundancy approach (as must be the case). However, θ^{YZ} is dropped from the model, reducing it to (XY, XZ).

In a numerical example, the ML estimates for the six estimable cell means are identical under the two methods, and the log-linear model parameter estimates are consistent. Although both methods reduce the model to six parameters, the parameter interpretations differ. The parameters derived using the parameter redundancy approach are linear combinations of those in the initial model. However, for instance, the estimate of θ in the second reduced model is not the intercept estimate for the initial model.

Example 4. Consider fitting model (4.1) to the pattern of zeros in Table 2(b). For the parameter redundancy approach, the derivative matrix is given in the Supplementary Material, and has rank six, thus, $d = 1$. Then, $\boldsymbol{\alpha}^{\mathsf{T}} = (1, -1, -1, 0, -1, 1, 1)$ indicates the estimable parameters as

$$\boldsymbol{\theta}'^{\mathsf{T}} = (\theta + \theta^X, \theta + \theta^Y, \theta^{XY}, \theta + \theta^Z, -\theta + \theta^{XZ}, -\theta + \theta^{YZ}).$$

Therefore, $\log\mu_{000}$ and $\log\mu_{110}$ are not estimable. The initial model is reduced to one with a design matrix of rank six:

$$\begin{bmatrix} \log\mu_{100} \\ \log\mu_{010} \\ \log\mu_{001} \\ \log\mu_{101} \\ \log\mu_{011} \\ \log\mu_{111} \end{bmatrix} = \begin{bmatrix} 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&1&0&0 \\ 1&0&0&1&1&0 \\ 0&1&0&1&0&1 \\ 1&1&1&1&1&1 \end{bmatrix} \begin{bmatrix} \theta + \theta^X \\ \theta + \theta^Y \\ \theta^{XY} \\ \theta + \theta^Z \\ -\theta + \theta^{XZ} \\ -\theta + \theta^{YZ} \end{bmatrix}.$$

However, an esoteric constraint exists, and is derived by considering

$$\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{U}(\boldsymbol{\theta}) = y_{000} - y_{110} - e^{\theta} + e^{\theta+\theta^X+\theta^Y+\theta^{XY}} = 0.$$

This translates to $\theta^X + \theta^Y + \theta^{XY} = 0$ or $\log\mu_{000} = \log\mu_{110}$. Adding this constraint to model (4.1) makes all parameters estimable.

In accordance with the EMLE approach for model (4.1), we identify a $\boldsymbol{\delta}$

that satisfies (3.1), such that $y_{\mathbf{i}} + \delta_{\mathbf{i}} > 0, \forall \mathbf{i} \in L$. Let $0 < \delta < 1$. Then, $\boldsymbol{\delta} = (+\delta, -\delta, -\delta, +\delta, -\delta, +\delta, +\delta, -\delta)$ holds the necessary and sufficient condition for the existence of the estimator of $\boldsymbol{\mu}$. This is confirmed by the polyhedral condition, because the observed marginals lie in the relative interior of the marginal of the polyhedral cone, owing to vector $\boldsymbol{y} = (y_1 + \delta, y_2 - \delta, y_3 - \delta, y_4 + \delta, y_5 - \delta, y_6 + \delta, y_7 + \delta, y_8 - \delta)$ satisfying (3.2). In other words, no $\boldsymbol{\zeta}$ or $\mathcal{F}$ can satisfy (3.3) or (3.4). Thus, we can maximize the likelihood function using numerical methods, and obtain estimates for all parameters of model (4.1). This is possible because of the esoteric constraint, which is not reported by this method, but is explicit in the parameter redundancy approach.

5. Conclusion

We have proposed a parameter redundancy approach for evaluating the effect of zero cell counts on the estimability of log-linear model parameters. For a parameter-redundant model, we obtain the estimable parameters and reduce the model to be identifiable.

We compare the parameter redundancy approach with a method that focuses on the existence of the MLE for the expected cell counts of a hierarchical model. Models with a nonexistent MLE are parameter redundant; some log-linear models are parameter-redundant despite an existent MLE. The latter occurs when maximizing the likelihood function that has a flat ridge imposes hidden extra constraints on the model to make a unique MLE possible.

The EMLE method is reported by Wang, Rauhy and Massam (2019) to be inefficient in finding the co-facial sets when the number of variables in the model is larger than 16. The authors propose using an approximation for the cone's face to make the method work for more variables. In the parameter redundancy approach, the symbolic algebra package `Maple` can be used to simultaneously solve a number of corresponding partial differential equations. However, as `Maple` runs out of memory, problems arise in the calculations when the model deficiency increases and becomes as large as 40. This limitation depends on the fitted model and the pattern of zeros in the table. For example, it may be more apparent in applications such as large cohort studies in which observations are concentrated in a small subspace of the entire sample space.

Future research could further explore parameter-redundant models with an existent MLE. This includes investigating the properties of the esoteric constraints and the goodness of fit of the model they imply.

Supplementary Material

The online Supplementary Material provides additional details for some of the examples, as well as Example S1 and the proof of Theorem 1 by induction.

Acknowledgments

We would like to thank the referees and the journal editor for their helpful comments and suggestions. The work of first author was supported by EPSRC PhD grants EP/J500549/1, EP/K503162/1 and EP/L505079/1.

Appendix

Proof of Theorem 2. Assume the MLE does not exist for a parameter-redundant model. We prove by contradiction that at least one $\boldsymbol{\alpha}_j$ vector does not satisfy $\boldsymbol{\alpha}_j^\top(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = 0$ for finite elements of $\boldsymbol{\theta}$. Suppose that all $\boldsymbol{\alpha}_j$ vectors, for $j = 1, \ldots, d$, satisfy $\boldsymbol{\alpha}_j^\top(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = 0$ for finite elements of $\boldsymbol{\theta}$. We know $\mathbf{U}(\boldsymbol{\theta}) = A^\top(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$. Then,

$$\begin{aligned} \boldsymbol{\alpha}_j^\top(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) &= 0 \\ \boldsymbol{\alpha}_j^\top A^\top(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) &= 0, \\ \boldsymbol{\alpha}_j^\top A_+^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_+ + \boldsymbol{\alpha}_j^\top A_0^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_0 &= 0, \end{aligned}$$

where $(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_+$ denotes a vector with the elements of $(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$ that correspond to the rows in A_+, and $(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_0$ denotes a vector with the elements of $(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$ that correspond to the rows in A_0. Now, $\boldsymbol{\alpha}_j^\top A_+^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_+ = 0$, because $\boldsymbol{\alpha}_j^\top A_+^\top = \mathbf{0}$, owing to $\boldsymbol{\alpha}_j^\top D = \mathbf{0}$. This implies $\boldsymbol{\alpha}_j^\top A_0^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_0 = 0$, or equivalently that $\boldsymbol{\alpha}_j^\top A_0^\top(-\boldsymbol{\mu}(\boldsymbol{\theta}))_0 = 0$. As the MLE does not exist, from (3.4), a $\boldsymbol{\zeta}$ vector exists such that $A_0\boldsymbol{\zeta} \gneq \mathbf{0}$. However, $\boldsymbol{\zeta}$ is also an $\boldsymbol{\alpha}$ vector, because $A_+\boldsymbol{\zeta} = \mathbf{0}$. Now suppose, without any loss of generality, that $\boldsymbol{\alpha}_{j'} = \boldsymbol{\zeta}$, for $1 \leqslant j' \leqslant d$. Then,

$$A_0\boldsymbol{\alpha}_{j'} \gneq \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\alpha}_{j'}^\top A_0^\top(-\boldsymbol{\mu}(\boldsymbol{\theta}))_0 < 0,$$

because all elements of $(-\boldsymbol{\mu}(\boldsymbol{\theta}))_0$ are nonzero and negative. Thus, this contradicts $\boldsymbol{\alpha}_j^\top A_0^\top(-\boldsymbol{\mu}(\boldsymbol{\theta}))_0 = 0$.

To prove the converse, assume an $\boldsymbol{\alpha}_j$ vector exists, for $1 \leqslant j \leqslant d$, such that $\boldsymbol{\alpha}_j^\top(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) < 0$ and cannot be zero for finite $\boldsymbol{\theta}$. This implies that

$$\boldsymbol{\alpha}_j^\top A_+^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_+ + \boldsymbol{\alpha}_j^\top A_0^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))_0 < 0,$$

$$\boldsymbol{\alpha}_j^\top A_0^\top(-\boldsymbol{\mu}(\boldsymbol{\theta}))_0 < 0,$$

since $\boldsymbol{\alpha}_j^\top D = \mathbf{0}$ means $\boldsymbol{\alpha}_j^\top A_+^\top = \mathbf{0}$. Thus, $\boldsymbol{\alpha}_j^\top A_0^\top \gneq \mathbf{0}$. From all $\boldsymbol{\alpha}_j$ such that $\boldsymbol{\alpha}_j^\top A_0^\top \gneq \mathbf{0}$, we choose the $\boldsymbol{\alpha}_{j'}$ that corresponds to the set $\{i : (A\mathbf{x})_{(i)} \neq 0\}$ with maximal cardinality. Then, $\boldsymbol{\alpha}_{j'}$ satisfies the three conditions in (3.4), and the MLE does not exist. This completes the proof of Theorem 2.

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. Wiley, New York.

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

Brown, M. B. and Fuchs, C. (1983). On Maximum likelihood estimation in sparse contingency tables. *Computational Statistics and Data Analysis* **1**, 3–15.

Catchpole, E. A. and Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika* **84**, 187–196.

Catchpole, E. A., Morgan, B. J. T. and Freeman, S. N. (1998). Estimation in parameter redundant models. *Biometrika* **85**, 462–468.

Catchpole, E. A. and Morgan, B. J. T. (2001). Deficiency of parameter redundant models. *Biometrika* **88**, 593–598.

Chan, L., Silverman, B. and Vincent, K. (2020). Multiple systems estimation for sparse capture sata: Inferential challenges when there are non-overlapping Lists. *Journal of the American Statistical Association*. `DOI: 10.1080/01621459.2019.1708748.`

Chappell, M. J. and Gunn, R. N. (1998). A procedure for generating locally identifiable reparameterisations of unidentifiable non-linear systems by the similarity transformation approach. *Mathematical Biosciences* **148**, 21–41.

Cole, D. J., Morgan, B. J. T. and Titterington, D. M. (2010). Detecting the parametric structure of models. *Mathematical Biosciences* **228**, 16–30.

Eriksson, N., Fienberg, S. E., Rinaldo, A. and Sullivant, S. (2006). Polyderal conditions for the nonexistence of the MLE for hierarchical log-linear models. *Journal of Symbolic Computation* **41**, 222–233.

Evans, N. D. and Chappell, M. J. (2000). Extensions to a procedure for generating locally identifiable reparameterisations of unidentifiable systems. *Mathematical Biosciences* **168**, 137–159.

Fienberg, S. E. and Rinaldo, A. (2006). Computing maximum likelihood estimation in log-linear models. Carnegie Mellon University. `http://www.stat.cmu.edu/tr/tr835/tr835.pdf`.

Fienberg, S. E. and Rinaldo, A. (2012a). Maximum likelihood estimation in log-linear models. *The Annals of Statistics* **40**, 996–1023.

Fienberg, S. E. and Rinaldo, A. (2012b). *Maximum Likelihood Estimation in Log-Linear Models.* Supplementary material: Algorithms. `http://www.stat.cmu.edu/~arinaldo/Fienberg_Rinaldo_Supplementary_Material.pdf`.

Friedlander, M. (2016). Fitting log-linear models in sparse contingency tables using the eMLEloglin R package. *arXiv:1611.07505*.

Gimenez, O., Viallefont, A., Catchpole, E. A., Choquet, R. and Morgan, B. J. T. (2004). Methods for investigating parameter redundancy. *Animal Biodiversity and Conservation*

27, 1–12.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics* **1**, 617–632.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago press, Chicago.

Hung, R. J., McKay, J. D., Gaborieau, V. et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637.

Johndrow, J. E., Bhattacharya, A. and Dunson, D. (2017). Tensor decompositions and sparse log-linear models. *The Annals of Statistics* **45**, 1–38.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition. Chapman and Hall, London.

Overstall, A. M. and King, R. (2014). conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software* **58**, 1–26.

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D. and Richardson, S. (2012). Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: Application to searching for gene × gene patterns. *Genetic Epidemiology* **36**, 663–674.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica* **39**, 577–591.

Wang, N., Rauhy, J. and Massam, H. (2019). Approximating faces of marginal polytopes in discrete hierarchical models. *The Annals of Statistics* **47**, 1203–1233.

Serveh Sharifi Far

School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK.

E-mail: serveh.sharifi@ed.ac.uk

Michail Papathomas

Department of Statistics, School of Mathematics and Statistics, University of St Andrews, KY16 9LZ, UK.

E-mail: m.papathomas@st-andrews.ac.uk

Ruth King

School of Mathematics, University of Edinburgh, EH9 3FD, UK.

E-mail: ruth.king@ed.ac.uk

(Received March 2018; accepted September 2019)