Parameter redundancy and the existence of the maximum likelihood estimates in log-linear models

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Statistica Sinica

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
PARAMETER REDUNDANCY AND THE EXISTENCE
OF MAXIMUM LIKELIHOOD ESTIMATES
IN LOG-LINEAR MODELS

Serveh Sharifi Far, Michail Papathomas* and Ruth King

University of Edinburgh, *University of St Andrews

Abstract: In fitting log-linear models to contingency table data, the presence of zero cell entries can have an adverse effect on the estimability of parameters, due to parameter redundancy. We describe a general approach for determining whether a given log-linear model is parameter redundant for a pattern of observed zeros in the table. We derive the estimable parameters or functions of parameters and show how to reduce the unidentifiable model to that of an identifiable one. For some parameter redundant models, the likelihood function has a flat ridge orthogonal to the axes of some parameters, that imposes additional parameter constraints which lead to obtaining unique maximum likelihood estimates for more parameters. In contrast to a mentioned alternative framework, the proposed approach informs on the existence of those constraints and determines them, elucidating the model that is, in fact, being fitted.

Key words and phrases: Contingency table, Extended maximum likelihood estimate, Identifiability, Parameter redundancy, Sampling zero.
1. Introduction

Categorical variables cross-classify subjects over combinations of their 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 and are popular due to their ability to model different interactions among the categorical variables. Log-linear models have been well studied and some examples of their applications in social, medical, and biological sciences are given by Agresti (2002), Bishop et al. (1975) and McCullagh & Nelder (1989).

Zero cell counts can have an adverse effect on the estimability of log-linear model parameters and expected cell counts. Zero entries are of two main types; structural and sampling zeros. If the expectation and variance for a cell count are zero, then it is a structural zero. A sampling zero is an observed zero entry to a cell with positive expectation. Sampling zeros contribute to the estimation of parameters and cell means through likelihood calculations.

In this manuscript, we examine how zero cell entries influence the estimability of log-linear model parameters and this is addressed with respect to identifiability and parameter redundancy. A model is parameter redundant when statistical methods are not able to estimate all parameters. For
a log-linear model this inability is displayed by large standard errors for some parameter estimates provided by optimization numerical methods. The concept of parameter redundancy is related to that of identifiability. 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 over-parametrized (Catchpole & 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. An overview of identifiability and parameter redundancy is given by Catchpole & Morgan (1997) and Catchpole et al. (1998). Identifiability is crucial when exploring the association between factors and it is relevant to studies which observe a large number of categorical variables, such as cohort studies. For example, it is known that most diseases are multi-factorial, and that different risk factors combine and interact to affect the risk of disease (Vineis, et al. 2008). Observing increased numbers of categorical variables leads to sparse contingency tables that include zero cell counts. The development of methods that identify the highest level of interaction complexity that can be explored is therefore important.

We develop a method for the detection of parameter redundancy for log-linear models in the presence of sampling zero observations. The es-
timable parameters and combinations of parameters are derived, and it is shown how a parameter redundant model can be reduced such that the new reduced model is identifiable. We refer to the proposed method as the “parameter redundancy” approach. The focus here is on the log-linear model parameters, rather than cell means. This is because, typically, the interest of practitioners lies on the significance and magnitude of main effects and interactions, rather than the mean of specific combinations of variable levels. In the presence of structural zeros, the corresponding cells are omitted from the modelling and analysis, since they are associated with cross-classifications that cannot be observed, with the resulting model checked for parameter redundancy.

A comprehensive study of log-linear models for contingency tables was developed by Haberman (1973), who proved that maximum likelihood estimates of model parameters are unique when they exist, and provided a necessary and sufficient condition for the existence of cell mean estimates in presence of zero cell entries. Brown & Fuchs (1983) investigated the effect of sampling zero observations on the existence of the maximum likelihood estimates for mean of cell counts by considering and comparing iterative methods. This effect was further studied in a polyhedral and graphical model framework defined by Lauritzen (1996). A polyhedral version of
Haberman’s necessary and sufficient condition for existence of the maximum likelihood estimator is provided by Eriksson et al. (2006). Estimability of parameters under non-existent maximum likelihood estimator, within the extended exponential families, is studied by Fienberg & Rinaldo (2012), and is developed to higher dimensional problems by Wang et al. (2016). We refer to these developments collectively as “the existence of the maximum likelihood estimator” or EMLE framework. The method demonstrates that when the maximum likelihood estimator does not exist, some parameters cannot be estimated. However, an extended MLE always exists so 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, while parameter redundancy provides a re-parametrization that retains the original interpretation of the parameters. It is because this method provides estimable parameters and estimable linear combinations of parameters instead of just a subset of the model’s initial parameters. Importantly, parameter redundancy also reveals additional constraints imposed by the likelihood function on some parameter redundant models for which the maximum likelihood estimator exist. This is particularly interesting, as standard statistical software reports pa-
parameter estimates without informing on the additional implied constraints. Consequently, one may not be aware of the exact log-linear model that is being fitted.

Section 1.1 introduces the necessary notation. Section 2 describes the determination of a parameter redundant model and its adaptation to log-linear models. The idea is illustrated by examples including a real-data example with a large sparse contingency table in a genome-wide association study of lung cancer, formed by some selected factors observed within a genome. We consider the saturated log-linear model (with \( m \) variables and \( l \) levels each) and determine which parameters become inestimable after observing a zero cell count. We also show when additional constraints, implied by the shape of the likelihood function, allow to determine unique maximum likelihood estimates for more parameters of a parameter redundant model. We refer to these constraints, whose presence is not reported by numerical optimization methods, as esoteric constraints. The proposed parameter redundancy approach determines them, thus revealing the model that is in fact fitted to the sparse table. In section 3, we review the idea of the existence of the maximum likelihood estimates for log-linear models and then compare the two approaches using illustrative examples. Section 4 concludes with a discussion of results.
1. INTRODUCTION

1.1 Log-linear models for contingency tables

Adopting the notation in Overstall & King (2014), let $V = \{V_1, ..., V_m\}$ denote a set of $m$ categorical variables, where the $j$th variable has $l_j$ levels. The corresponding contingency table has $n = \prod_{j=1}^{m} l_j$ cells. Let $y$ denote an $n \times 1$ vector corresponding to the observed cell counts. Each element of $y$ is denoted by $y_i$, $i = (i_1, \ldots, i_m)$ such that $0 < i_j < l_j - 1$ and $j = 1, \ldots, m$. Here, $i$, identifies the combination of variable levels that cross-classify the given cell. We define $L$ as the set of all $n$ cross-classifications, so that $L = \bigotimes_{j=1}^{m} [l_j]$, in which $[l_j] = \{0, 1, \ldots, l_j - 1\}$. Then, $N = \sum_{i \in L} y_i$ denotes the sum of all cell counts. The $y_i$'s are observations from independent Poisson random variables $Y_1$, such that, $\mu_i = E(Y_1)$. Let $\mathcal{E}$ denote a set of subsets of $V$. By adapting the notation of Johndrow et al. (2014), the log-linear model assumes the form,

$$m_i = \log \mu_i = \sum_{e \in \mathcal{E}} \theta^e(i), \quad (1.1)$$

where $\theta^e(i) \in \mathbb{R}$ denotes the main effect or interaction among the variables in $e$ corresponding to the levels in $i$. The summation is over all members of $\mathcal{E}$, which could be the set of all subsets of variables (for a saturated model) or a set of desirable subsets (for a smaller model). As a convention, $\theta$ corresponds to $e = \emptyset$, which guarantees the existence of an intercept. To
allow for the existence of unique parameter estimates, corner point con-
straints are applied, so that parameters that incorporate the lowest level
of a variable are set to zero. Alternatively, for $p$ parameters, we write,

$$\mathbf{m}_{n \times 1} = \log \mu_{n \times 1} = A_{n \times p} \theta_{p \times 1},$$

where $A$ is a full rank design matrix. For a model fitted to an $l^m$ table, an alternative way to identify cell counts in

$$\text{(1.1)}$$
is to set a one-to-one correspondence between the set $\mathbf{i} = (i_1, ..., i_m)$
and integer numbers, $i = 1, ..., l^m$, as

$$\mathbf{i} = (i_1, ..., i_m) = i_1l^0 + i_2l^1 + \cdots + i_{m-1}l^{m-2} + i_m l^{m-1} + 1. \quad \text{(1.2)}$$

2. The Parameter Redundancy approach

2.1 The derivative method

The generic approach introduced in [Catchpole & Morgan (1997)] and [Catch-
pole et al. (1998)] is summarized here to identify a parameter redundant
model and to determine the set of estimable parameters. The mean vector
of observations $E(Y) = \mu$ for a distribution from the exponential family of
distributions, is expressible as a function of parameters $\theta = (\theta_1, ..., \theta_p) \in \Omega$.
The derivative matrix $D(\theta)$, describes the relationship between $\mu$ (or a
monotonic function of it) and $\theta$,

$$D_{si}(\theta) = \frac{\partial \mu_i}{\partial \theta_s}, \quad i = 1, \ldots, n, \quad s = 1, \ldots, p. \quad \text{(2.3)}$$
Theorem 1 of Catchpole & Morgan (1997) states that the model which relates \( \mu \) to \( \theta \) is parameter redundant, if and only if the matrix is symbolically rank deficient, i.e. if there exists a non-zero vector, \( \alpha(\theta) \), such that for all \( \theta \),

\[
\alpha(\theta)^T D(\theta) = 0.
\] (2.4)

In the likelihood surface of parameter redundant models, there is a flat ridge which causes non-existent MLE for some parameters (Catchpole & Morgan, 1997). Parameter redundancy could occur due to model structure or because of lack of data (Catchpole & Morgan, 2001; Cole et al., 2010), the latter type is sometimes referred to as “extrinsic” parameter redundancy (Gimenez et al., 2004). Non-parameter redundant models are referred to as full rank models.

The rank of the derivative matrix, \( r \), is the number of estimable parameters and estimable combinations of parameters. The model deficiency is defined as \( d = r - p \), which is the number of possible \( \alpha(\theta) \) vectors. The positions where all \( \alpha(\theta) \) vectors have zero elements correspond to the parameters that are directly estimable. To find estimable combinations of parameters, the auxiliary equations of the following system of linear first
order partial differential equations are solved,
\[
\sum_{s=1}^{p} \alpha_{sj} \frac{\partial f}{\partial \theta_s} = 0, \quad j = 1, \ldots, d. 
\] (2.5)

After detecting the set of estimable parameters and estimable combina-

tions of them, the initial model can be reduced to a full rank model

including only estimable parameters and combinations of them.

2.2 Parameter redundancy for log-linear models

The generic approach cannot be applied to log-linear models by utilizing

(2.3), as this will not readily determine a zero cell count as a missing obser-

vation. We adjust the derivative matrix (2.3) first, so that it is a function

of the data. Hence, \( y_i \log \mu_i \) is used as the monotonic function of \( \mu_i \), such

that,
\[
D_{si} = \frac{\partial y_i \log \mu_i}{\partial \theta_s}, \quad i = 1, \ldots, n, \quad s = 1, \ldots, p. 
\] (2.6)

The presence of cell counts in the monotonic function of cell means allows

us to investigate the effect of zero observations in the model (which does

not need to be hierarchical). Each sampling zero turns a column of the

matrix to zero and may decrease the rank of the derivative matrix, thus
detecting an extrinsic parameter redundant model.

The model is full rank if the rank of the derivative matrix is not smaller

than the number of parameters \( p \), otherwise, it is parameter redundant.
Finding all estimable parameters and estimable combinations of parameters further identifies which cell means are estimable. The vector of estimable quantities ($\theta'$), and the vector of estimable cell means ($\mu'$), specify a reduced model through a smaller design matrix ($A'$). The reduced model is full rank with rank $r$ and its degree of freedom is the number of estimable cell means minus the number of estimable quantities.

Using the log-likelihood function elements is a common option to consider in forming the derivative matrix (Cole et al., 2010). However, when using those elements for the Poisson log-linear model, setting cell counts to zero does not decrease the rank of the derivative matrix. Catchpole & Morgan (2001) mention the role of the score vector for a multinomial log-linear model to assess the effect of missing data on the redundancy of the model. Utilising the information matrix instead of a derivative matrix is an alternative for detecting parameter redundancy (Rothenberg, 1971), which again does not show the rank deficiency caused by zero cell counts here.

The process of detecting parameter redundancy and reducing the model is illustrated in the following examples.

**Example 1.** The data pattern in Table [1] taken from Fienberg & Rinaldo (2012), 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
2. THE PARAMETER REDUNDANCY APPROACH

Table 1: Observations in a $3^3$ contingency table

<table>
<thead>
<tr>
<th></th>
<th>$y_1$</th>
<th>$y_7$</th>
<th>$y_{10}$</th>
<th>$y_{13}$</th>
<th>$y_{16}$</th>
<th>$y_{22}$</th>
<th>$0^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>$y_5$</td>
<td>$y_8$</td>
<td>$y_{11}$</td>
<td>$y_{14}$</td>
<td>0*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_3$</td>
<td>$y_6$</td>
<td>$y_9$</td>
<td>$y_{12}$</td>
<td>0</td>
<td>0</td>
<td>$y_{21}$</td>
<td>$y_{24}$</td>
</tr>
</tbody>
</table>

as sampling zeros. All other cell counts are positive Poisson observations, named according to (1.2). The fitted hierarchical model is $(XY, XZ, YZ)$ for which, $\log \mu_{27 \times 1} = A_{27 \times 19} \theta_{19 \times 1}$ with parameters,

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

The rank of the derivative matrix is 18, i.e. there are only 18 estimable parameters or combinations of them. So, $d = 19 - 18 = 1$, and the only $\alpha$ in (2.4) is, $\alpha^T = (1, 0, -1, -1, -1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0)$. After solving (2.5), the estimable quantities are,

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

Vector $\theta^T$ determines that 21 out of 27 cell means are estimable, including cells 17 and 25, indicated in Table 1 with zero asterisks. Cell means
1, 2, 15, 18, 19, 20 are not estimable. We treat those cells as structural zeros and remove them from the model. Considering $\theta'$ and the 21 estimable cell means, the reduced model with three degrees of freedom is,

$$\log \mu'_{21 \times 1} = A'_{21 \times 18} \theta'_{18 \times 1}.$$

**Example 2.** Hung et al. (1973) studied a genome-wide association of lung cancer by genotyping for 317,139 single nucleotide polymorphisms (SNP) in chromosomes 6 and 15. Each SNP is categorized at three levels of 0, 1 and 2 to identify the number of their minor allele. The complete data sample consists of 3841 individuals from six European countries. 500 SNPs with the highest p-value in the test for association with lung cancer are chosen to be considered. Papathomas et al. (2012) decreased this number of SNPs to 50 via applying profile regression. We further select the following five SNPs (as representatives of uncorrelated groups of SNPs) which form a $3^5$ contingency table; rs7748167_C ($A$), rs4975616_G ($B$), rs6803988_T ($C$), rs11128775_G ($D$), rs9306859_A ($E$).

We choose to fit a log-linear model with main effects, first-order, and second-order interactions of the variables to the selected data, so the model has 243 cell counts and 131 parameters. The corresponding contingency table includes 132 sampling zero cells which cause large standard errors for many of the parameter estimates in naively fitting the log-linear model. By
applying the parameter redundancy method, we can immediately eliminate the inestimable parameters and construct a smaller identifiable model to proceed with. The rank of the derivative matrix formed according to (2.6) is 95, indicating only 95 parameters or linear combinations of parameters are estimable in this model. The deficiency of the model is \( d = p - r = 131 - 95 = 36 \). Using the 36 \( \alpha \) vectors and solving the corresponding differential equations the 95 estimable parameters are specified and given in the Appendix.

These estimable parameters make 12 cell means estimable of those 132 cells with zero cells entries. Thus, 95 quantities and 123 \((111 + 12)\) cell means of the model are estimable. This leads to reducing the model to a smaller one with the corresponding design matrix \( \mathcal{A}' \), shown as,

\[
\log \mu'_{123 \times 1} = \mathcal{A}'_{123 \times 95} \theta'_{95 \times 1},
\]

with degree of freedom of \( d.f = 123 - 95 = 28 \). After forming the new design matrix and fitting the model to the data for 123 cells, the \( \theta' \) parameter estimates are provided without repeating the previous large standard errors.

A crucial variable in this study describes the presence or absence of cancer in each of the individuals and by adding this variable \((F)\) to the table, the \(3^5 \times 2^1\) contingency table has 486 cells. To study the interactions between the five SNPs and the outcome variable, we only consider the main
effects of variables and the first-order interactions between them which make 62 parameters. Then the contingency table has 298 zero cell counts and the corresponding derivative matrix has the rank 59 and $d = 62 - 59 = 3$. It indicates that there are 59 estimable parameters in the model fitted to this sparse table. After finding the three $\alpha$ vectors and solving the corresponding partial differential equations, the 59 estimable parameters are obtained and given in the Appendix.

Thus, only three parameters $\theta_{22}^{AD}, \theta_{22}^{AE}, \theta_{22}^{DE}$ are not estimable due to the sparseness of the table. The estimable parameters make 360 out of 486 cell means estimable, so 62 cells with observed zero counts have estimable cell means. The reduced model including only the estimable parameters and the estimable cell means is shown as,

$$\log \mu_{360 \times 1}' = \Lambda_{360 \times 59}' \theta_{59 \times 1}'$$

with degree of freedom of $d.f = 360 - 59 = 301$. After forming the new design matrix and fitting the model to the data for 360 cells, the parameter estimates with no large standard errors are obtained by fitting a log-linear model. All the main effect coefficients are significant at the 0.05 level. The presence of cancer has a significant positive interaction with the level 1 of variables $A$ and $D$, and a significant negative interaction with the level 1 of variables $C, E$ and the level 2 of variables $B, C, E$. 
2. PARAMETER REDUNDANCY APPROACH

2.3 Parameter redundancy for a saturated log-linear model

We provide some general results on parameter redundancy for a saturated log-linear model fitted to an $l^m$ contingency table ($m$ variables, each classified in $l$ levels). The next example illustrates the proposed approach by showing that for all positive cell counts a saturated log-linear model is always full rank.

**Example 3.** It is known that for a log-linear model fitted to a contingency table with all positive $y_i$, the log-likelihood function is strictly concave and the maximum likelihood estimates exist for all model parameters \cite{Haberman1973}. Consider fitting a saturated Poisson log-linear model to an $l^m$ ($m \geq 1, l \geq 2$) contingency table. The derivative matrices for a $2 \times 1$ and a $2 \times 2$ table, according to (2.6), are,

$$
D_1 = \begin{bmatrix}
\mu_0 & \mu_1 \\
\theta & y_1 & y_2 \\
\theta^X & 0 & y_2
\end{bmatrix}, \quad D_2 = \begin{bmatrix}
\mu_{00} & \mu_{10} & \mu_{01} & \mu_{11} \\
\theta & y_1 & y_2 & y_3 & y_4 \\
\theta^X & 0 & y_2 & 0 & y_4 \\
\theta^Y & 0 & 0 & y_3 & y_4 \\
\theta^{XY} & 0 & 0 & 0 & y_4
\end{bmatrix}.
$$

We can always consider an ordering of cell means and corresponding parameters that produces the D matrix in the form of an upper triangular matrix with main diagonal elements equal to the cell counts. So when $y_i > 0, \forall i \in L$, the D matrix is always full rank, as expected, and all of the model parameters are estimable.
Now we assume that the contingency table contains one zero cell count and the following theorem determines exactly which model parameters become inestimable as a result of one zero observation. Some definitions are required first.

**Definition 1.** For a saturated log-linear model, we define the parameter corresponding to the cell with count \( y_{i}, i = 1, \ldots, n \) (according to (1.2)), as the one with the maximum number of variables in its superscript, within the set of all parameters in \( \log \mu_{i} = A_{(i)} \theta \).

For example, for a \( 3 \times 3 \) 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 \( \theta_{XZ}^{21} \).

**Definition 2.** For a given log-linear model parameter, 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 \( 3 \times 3 \) table, the parameters associated with a higher order interaction given \( \theta_{XZ}^{21} \), are \( \theta_{XYZ}^{211} \) and \( \theta_{XYZ}^{221} \).

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

The proof by induction is given in the online supplementary material.

Additional zero cells in the table cannot make previously inestimable parameters estimable, as the amount of information is further reduced. Thus, the set of the model’s inestimable parameters is at least as large as the union of the inestimable parameters per zero cell. Estimable parameters and linear combinations of them are derived by solving (2.5).

2.4 The esoteric constraints

The likelihood function of parameter redundant models has a flat ridge which is occasionally orthogonal to the axes of some parameters, so that the associated parameters still have unique maximum likelihood estimates (Catchpole et al., 1998). In this case, the likelihood function essentially imposes one or more extra constraints on the model parameters. We refer to these constraints as “esoteric constraints”. Knowing these implied constraints reveals the model that is actually fitted, as standard statistical software does not provide any information on them when maximizing the likelihood function. After detecting a parameter redundant model, we can check the existence of such constraints for the model.
3. Maximum likelihood estimation and parameter redundancy

The log-likelihood function for (1.1) is, 
\[ l(\theta) = \sum_i (y_i \log \mu_i(\theta) - \mu_i(\theta)). \]

When a model is parameter redundant, then \( \alpha_j^T(\theta)D(\theta) = 0 \) for \( j = 1, \ldots, d \), and it follows that \( \alpha_j^T(\theta)U(\theta) = 0 \) for \( j = 1, \ldots, d \), in which \( U(\theta) = (\partial l/\partial \theta_1, \ldots, \partial l/\partial \theta_p)^T \) (Catchpole & Morgan, 1997). The partial derivatives are,
\[
\frac{\partial l}{\partial \theta_s} = \sum_i \left( \frac{y_i}{\mu_i(\theta)} - 1 \right) \frac{\partial \mu_i(\theta)}{\partial \theta_s} = \sum_i (y_i - \mu_i(\theta)) \frac{\partial \mu_i(\theta)}{\partial \theta_s} \frac{1}{\mu_i(\theta)}.
\]

If \( \alpha_j^T(\theta)U(\theta) \) is impossible to be zero with finite \( \theta_s \), then the esoteric constraints do not exist. If imposing one or more constraints can make it zero, then those are the esoteric constraints. Such constraints do not exist for models in Theorem 1, Example 1, and Example 2, and the only way to deal with those models is to reduce them to an identifiable one. A model with an esoteric constraint is given in Example 5 of Section 3.2.

3. Maximum likelihood estimation and parameter redundancy

3.1 The existence of the maximum likelihood estimator for log-linear models

For decomposable log-linear models with an explicit formula for the maximum likelihood estimator or \( \hat{\mu}_i \), positivity of minimal sufficient statistics is a necessary and sufficient condition for the existence of the maximum
likelihood estimator (EMLE) of $\mu$ (Agresti, 2002). For non-decomposable models, $\hat{\mu}_i$ is calculated by iterative methods. In this case, positivity of sufficient table marginals is still necessary for the existence of the estimator but it is no longer a sufficient condition.

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

$$\mathcal{M}^\perp = \{ x \in \mathcal{R}^{\mid L\mid} : (x, m) = x^T m = 0, \forall m \in \mathcal{M} \} .$$ (3.7)

Then, theorem 3.2 of Haberman (1973) states that a necessary and sufficient condition that the maximum likelihood estimate $\hat{m}$ of $m$ exists is that there exist $\delta \in \mathcal{M}^\perp$ such that $y_i + \delta_i > 0$ for every $i \in L$. Here, $\mu$ in $m = \log \mu$ is assumed to be positive. The theorem specifies, for any pattern of zeros in the table, whether the MLE of the cell means exists or not. In the extended maximum likelihood estimate case, a cell mean estimate could be $\hat{\mu}_i = 0$, but its log transformation is not defined and then estimates of some corresponding $\theta$ parameters tend to infinity (Haberman, 1974).

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

\[
C_A = \{ t : t = A^T y, y \in \mathbb{R}^{|L|}_+ \}. \tag{3.8}
\]

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 \( \text{(McCullagh & Nelder, 1989)} \). 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, \( F = \{ t \in C_A : (t, \zeta) = 0 \} \), for some \( \zeta \in \mathbb{R}^p \), such that \( (t, \zeta) \geq 0 \) for all \( t \in C_A \), with \( (t, \zeta) \) representing the inner product. The facial set \( \mathcal{F} \) is a set of cell indexes of the rows of \( A \) whose conic hull is precisely \( F \). \( \mathcal{F} \subseteq L \) is a facial set of \( F \) for any design matrix \( A \) for \( \mathcal{M} \), if there exists some \( \zeta \in \mathbb{R}^p \) such that,

\[
(A_{(i)}, \zeta) = 0, \quad \text{if} \quad i \in \mathcal{F}, \tag{3.9}
\]

\[
(A_{(i)}, \zeta) > 0, \quad \text{if} \quad i \in \mathcal{F}^c,
\]

where \( A_{(i)} \) is the \( i \)th row of \( A \), and \( \mathcal{F}^c = L - \mathcal{F} \) is the co-facial set of \( F \) \( \text{(Fienberg & Rinaldo, 2012)} \). If such \( \zeta \) and \( \mathcal{F} \) exist, the MLE does not exist and only cell means corresponding to members of \( \mathcal{F} \) are estimable. The inestimable cells in \( \mathcal{F}^c \) are treated as structural zeros and are omitted from
3. MAXIMUM LIKELIHOOD ESTIMATION AND PARAMETER REDUNDANCY

the model. An estimable subset of model parameters could be determined by finding $A_F$, the matrix whose rows are the ones from $A$ with coordinates in $\mathcal{F}$. $A_F$ which is a $|\mathcal{F}| \times p$ design matrix with rank $p_F$, is then reduced to $A^*_F$ of rank $p_F$ and identical column range to have a minimal representation. By implementing this reduced design matrix, the log-likelihood function is strictly concave with a unique maximizer. Then the extended MLE is,

$$\hat{\theta}^e = \operatorname{argmax}_{\theta \in \mathbb{R}^{p_F}} l_F(\theta) = \operatorname{argmax}_{\theta \in \mathbb{R}^{p_F}} t_F^T \theta - 1^T \exp(A^*_F \theta),$$

in which $t_F = (A^*_F)^T y_F$ and the extended MLE of the cell mean vector is $\hat{m}^e = \exp(A^*_F \hat{\theta}^e)$ (Fienberg & Rinaldo, 2012b).

Another way to define the facial set is by denoting sub-matrices obtained from $A$, named $A_+$ and $A_0$. They are made of rows 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 $\zeta$ satisfying the following three conditions:
3. MAXIMUM LIKELIHOOD ESTIMATION AND PARAMETER REDUNDANCY

1. \( A_+ \zeta = 0 \). \hspace{1cm} (3.10)

2. \( A_0 \zeta \succeq 0 \).

3. The set \( \{ i : (A \zeta)_i \neq 0 \} \) has maximal cardinality among all sets of the form \( \{ i : (A x)_i \neq 0 \} \) with \( A x \succeq 0 \), for \( x \) that satisfy the first two conditions. (Fienberg & Rinaldo, 2012b).

In (3.9) and (3.10) the inequality signs could be switched to less than zero without loss of generality. With \( \succeq \) we describe a vector with at least one element greater than zero. In conclusion, if \( \text{rank}(A_+) = \text{rank}(A) \), the MLE exists, since no vector \( \zeta \) exists and \( F^c = \emptyset \). If \( \text{rank}(A_+) < \text{rank}(A) \), the MLE may still exist, so we should search for a facial set.

The degree of freedom for the reduced model is \( d.f = |F| - \text{rank}(A_F^*) \) which is the number of cell means that are estimable minus the number of estimable log-linear model parameters (Fienberg & Rinaldo, 2012). Computational algorithms for detecting the existence of the MLE and deriving the co-facial set, by converting these methods into linear and non-linear optimization problems, are described in (Fienberg & Rinaldo, 2012b). However, those algorithms are inefficient for a model with a large number of variables (Wang et al., 2016).
3. MAXIMUM LIKELIHOOD ESTIMATION AND PARAMETER REDUNDANCY

3.2 The EMLE framework versus the parameter redundancy approach

Within the EMLE framework, when there are facial and co-facial sets as defined in (3.9), the maximum likelihood estimator of $\mu$ does not exist and some zero cells are treated as structural zeros. In the parameter redundancy approach, this is equivalent to $\alpha^T D = 0$ and no esoteric constraints determined by $\alpha^T U(\theta) = 0$. When the co-facial set is either null or there is no co-facial set as described in (3.9), then the maximum likelihood estimates exist. This is equivalent to the parameter redundancy outcomes in which the model is either full rank, or the necessary esoteric constraints exist. The next theorem explains a link between the EMLE method and the parameter redundancy approach through the score vector $U(\theta)$.

**Theorem 2.** For a parameter redundant model, the maximum likelihood estimator of $\mu$ does not exist if and only if one or more $\alpha_j$ vectors, $j = 1, \ldots, d$, do not satisfy $\alpha_j^T(\theta) U(\theta) = 0$ for finite elements of $\theta$.

The proof is given in the Appendix.

Two examples are utilized here to illustrate similarities and differences between the two approaches. For a parameter redundant model without any possible additional esoteric constraints, as in Example 4 below, the
only way to proceed with the initially considered model is to reduce it. The two reduced models found by the two approaches have a different re-parametrization of $\theta$, although the maximum likelihood estimates of the estimable cell means are identical. The parameters in the reduced model obtained by parameter redundancy, have the same interpretations as they had in the initial model, in terms of interactions of variable levels. When the model is parameter redundant and the maximum likelihood estimator does exist, there are more options to consider. For the model in Example 5, there is an esoteric constraint, extracted by the parameter redundancy approach, that makes all parameters estimable. Two ways can be considered to deal with this type of model. Reduce the model to a smaller, saturated and identifiable one, or adopt the esoteric constraint which is equivalent to using numerical methods to maximize the likelihood.

**Example 4.** We consider fitting log-linear model (3.11) to the contingency table (a) in Table 2 with three variables, each categorized in two levels. Cell counts denoted by $y_{i}$, $i = 2, \ldots, 7$ are positive.

$$
\log \mu_{ijk} = \theta + \theta_i^X + \theta_j^Y + \theta_k^Z + \theta_{ij}^{XY} + \theta_{ik}^{XZ} + \theta_{jk}^{YZ}, \quad i, j, k = \{0, 1\}^2. \quad (3.11)
$$

The method described in Section 2.2 is applied first. The rank of the derivative matrix, formed using (2.6) with columns one and eight set to zero
Table 2: Observations in two $2^3$ contingency tables

<table>
<thead>
<tr>
<th></th>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
<th></th>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y = 0$</td>
<td>$Y = 1$</td>
<td>$Y = 0$</td>
<td>$Y = 1$</td>
<td></td>
</tr>
<tr>
<td>$X = 0$</td>
<td>0</td>
<td></td>
<td>$y_3$</td>
<td></td>
<td>$y_5$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$y_2$</td>
<td>$y_4$</td>
<td></td>
<td>$y_6$</td>
<td>0</td>
</tr>
</tbody>
</table>

and the parameter vector $\theta^T = (\theta, \theta^X, \theta^Y, \theta^{XY}, \theta^Z, \theta^{XZ}, \theta^{YZ})$, is 6 indicating that $d = 1$. From (2.4), we have $\alpha^T = (1, -1, -1, 1, -1, 1, 1)$. Solving (2.5) gives the estimable parameters,

$$\theta^T = (\theta + \theta^X, \theta + \theta^Y, -\theta + \theta^{XY}, \theta + \theta^Z, -\theta + \theta^{XZ}, -\theta + \theta^{YZ}).$$

Therefore, all cell means but the first (log $\mu_{000} = \theta$) and the last (log $\mu_{111} = \theta + \theta^X + \theta^Y + \theta^{XY} + \theta^Z + \theta^{XZ} + \theta^{YZ}$) are estimable. No esoteric constraint exists as,

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

for finite $\theta$s. We treat $y_{000}$ and $y_{111}$ as structural zeros and remove them from the model. Then, reduce the model to a saturated one with a design matrix of rank 6, in accordance with the estimable parameters and linear combinations of them. The reduced model, for which the maximum
In accordance with the EMLE method, model (3.11) has no zero sufficient marginals, and positive estimates for all the cell means do not exist according to the Haberman’s sufficiency and necessary condition and also the polyhedral condition. For reducing this model to an identifiable one, according to the polyhedral method, we obtain,

\[ \mathcal{F} = \{100, 010, 110, 001, 101, 011\}, \quad \mathcal{F}^c = \{000, 111\}, \quad \mathbf{\zeta} = (1, -1, -1, 1, -1, 1, 1). \]

The design matrix for the reduced model is \( A^*_f \), which is a \( |\mathcal{F}| \times p_F = 6 \times 6 \) matrix and is found by using the suggested proposition 5.1 in Fienberg & Rinaldo (2012b). The final model is,

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

The estimable cell means are the same as derived by the parameter redundancy approach (as must be the case). However, \( \theta^{YZ} \) is dropped from the
model now, i.e. the model is reduced to \((XY, XZ)\).

In a numerical example, the maximum likelihood estimates for the six estimable cell means are identical under the two methods and log-linear model parameter estimates are also consistent. Although both methods reduce the model to one with six parameters, their parameter interpretations are different. The parameters derived by the parameter redundancy approach are the ones in the initial model. However, for instance, the estimate of \(\theta\) in the second reduced model is not the intercept estimate for the initial model.

**Example 5.** Consider fitting model (3.11) to the pattern of zeros in contingency table (b) in Table 2. For the parameter redundancy approach, the rank of the derivative matrix, with columns one and four set to zero, is 6 and \(d = 1\). Then, \(\mathbf{\alpha}^T = (1, -1, -1, 0, -1, 1, 1)\) which indicates the estimable parameters as,

\[
\mathbf{\theta}^T = (\theta + \theta^X, \theta + \theta^Y, \theta^{XY}, \theta + \theta^Z, -\theta + \theta^{XZ}, -\theta + \theta^{YZ}).
\]

Therefore, \(\log \mu_{000}\) and \(\log \mu_{110}\) are not estimable. We reduce the initial model to a saturated one with a design matrix of rank 6 as,
3. MAXIMUM LIKELIHOOD ESTIMATION AND PARAMETER REDUNDANCY

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

An esoteric constraint exists and is derived by setting,

\[\alpha^T U(\theta) = y_{000} - y_{110} - e^\theta + e^{\theta^X + \theta^Y + \theta^{XY}} = 0.\]

This translates to \(\theta^X + \theta^Y + \theta^{XY} = 0\) or \(\log \mu_{000} = \log \mu_{110}\). Imposing this constraint on model (3.11) makes all parameters estimable, although the model is parameter redundant and has a flat ridge in the likelihood surface.

In accordance with the EMLE approach for model (3.11), we identify a \(\delta\) which satisfies (3.7), such that \(y_i + \delta_i > 0, \forall i \in L\). Let \(0 < \delta < 1\), then \(\delta = (+\delta, -\delta, -\delta, +\delta, -\delta, +\delta, +\delta, -\delta)\) holds the necessary and sufficient condition for the existence of the estimator of \(\mu\). This is also confirmed by the polyhedral condition since the observed marginals lie in the relative interior of the marginal of the polyhedral cone, as vector \(y = (y_1 + \delta, y_2 - \delta, y_3 - \delta, y_4 + \delta, y_5 - \delta, y_6 + \delta, y_7 + \delta, y_8 - \delta)\) satisfies (3.8). In other words, there exist no \(\zeta\) or \(F\) such that satisfy (3.9). Thus, we are able to maximize the likelihood function by numerical methods and obtain the estimates for all parameters of model (3.11). This is possible because of the esoteric
constraint, which is not reported by this method but is explicit in the parameter redundancy approach.

4. Discussion

We proposed a general approach for evaluating the effect of zero cell counts on the estimability of log-linear model parameters as a parameter redundancy problem. For a parameter redundant model, we obtained the estimable parameters and reduced the model to an identifiable one. As a special case, we considered a saturated model with only one zero cell count and determined which model parameters are not directly estimable because of that zero cell.

The parameter redundancy approach was compared with a different method that focuses on the existence of the MLE of the mean counts. Models with non-existent MLE are parameter redundant, whilst some log-linear models are parameter redundant despite their existent MLE. The latter happens because the likelihood function has a flat ridge, orthogonal on some parameters axes, which imposes hidden extra constraints in the model to make a unique MLE possible. These constraints are derivable by the parameter redundancy method.

The EMLE method is reported in Wang et al. (2016) to be inefficient in
finding the co-facial sets when the number of variables in the hierarchical model is larger than 16. The authors propose 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 could be used to simultaneously solve a number of corresponding partial differential equations. However, problems rise in the calculations for solving a large number of partial differential equations when the model deficiency increases and becomes as large as 40.

Investigating and comparing possible ways to handle parameter redundant models with existent maximum likelihood estimator is the subject of ongoing work. Our aim is to further explore properties of the esoteric constraints, as well as alternative ways to derive identifiable models that may fit the data better than the model implied by the esoteric constraints.

Supplementary Materials

The online supplementary material contains the proof of Theorem 1 by induction.

Appendix

Example 2. The vector of 95 estimable parameters obtained by parameter
redundancy for the $3^5$ contingency table.

\[ \theta^T = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}, \theta_{18}, \theta_{19}, \theta_{20}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{26}, \theta_{27}, \theta_{28}, \theta_{29}, \theta_{30}, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34}, \theta_{35}) \]

\[ \theta_{11}^{AC}, \theta_{12}^{AC}, \theta_{13}^{AC}, \theta_{14}^{AC}, \theta_{15}^{AC}, \theta_{16}^{AC}, \theta_{17}^{AC}, \theta_{18}^{AC}, \theta_{19}^{AC}, \theta_{20}^{AC}, \theta_{21}^{AC}, \theta_{22}^{AC}, \theta_{23}^{AC}, \theta_{24}^{AC}, \theta_{25}^{AC}, \theta_{26}^{AC}, \theta_{27}^{AC}, \theta_{28}^{AC}, \theta_{29}^{AC}, \theta_{30}^{AC}, \theta_{31}^{AC}, \theta_{32}^{AC}, \theta_{33}^{AC}, \theta_{34}^{AC}, \theta_{35}^{AC} \]

The vector of 59 estimable parameters obtained by parameter redundancy for the $3^5 \times 2^1$ contingency table.

\[ \theta^T = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}, \theta_{18}, \theta_{19}, \theta_{20}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{26}, \theta_{27}, \theta_{28}, \theta_{29}, \theta_{30}, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34}, \theta_{35}) \]

\[ \theta_{11}^{AC}, \theta_{12}^{AC}, \theta_{13}^{AC}, \theta_{14}^{AC}, \theta_{15}^{AC}, \theta_{16}^{AC}, \theta_{17}^{AC}, \theta_{18}^{AC}, \theta_{19}^{AC}, \theta_{20}^{AC}, \theta_{21}^{AC}, \theta_{22}^{AC}, \theta_{23}^{AC}, \theta_{24}^{AC}, \theta_{25}^{AC}, \theta_{26}^{AC}, \theta_{27}^{AC}, \theta_{28}^{AC}, \theta_{29}^{AC}, \theta_{30}^{AC}, \theta_{31}^{AC}, \theta_{32}^{AC}, \theta_{33}^{AC}, \theta_{34}^{AC}, \theta_{35}^{AC} \]

\textbf{Proof of Theorem 2}. Assume the MLE does not exist for a parameter redundant model. Suppose that all $\alpha_j$ vectors, $j = 1, \ldots, d$, satisfy
\[ \alpha_j^T(\theta) U(\theta) = 0 \] for finite elements of \( \theta \). For simplicity, consider having 0 and 1’s in the derivative matrix instead of 0 and \( y_i \)'s and set zero columns corresponding to zero observations, which gives \( U(\theta) = A^T(y - \mu(\theta)) \).

Then,
\[ \alpha_j^T(\theta) U(\theta) = 0 \]
\[ \alpha_j^T A^T(y - \mu(\theta)) = 0, \]
\[ \alpha_j^T A_+^T(y - \mu(\theta))_+ + \alpha_j^T A_0^T(y - \mu(\theta))_0 = 0, \]

where \((y - \mu(\theta))_+\) denotes a vector with the elements of \((y - \mu(\theta))\) that correspond to the rows in \(A_+\), and \((y - \mu(\theta))_0\) denotes a vector with the elements of \((y - \mu(\theta))\) that correspond to the rows in \(A_0\). Now, \(\alpha_j^T A_+^T(y - \mu(\theta))_+ = 0\), because \(\alpha_j^T A_+^T = 0\), since \(\alpha_j^T D = 0\). This implies that \(\alpha_j^T A_0^T(y - \mu(\theta))_0 = 0\), or equivalently that \(\alpha_j^T A_0^T(-\mu(\theta))_0 = 0\).

As the MLE does not exist, from (3.10), a \(\zeta\) vector exists so that \(A_0\zeta \succeq 0\). However, \(\zeta\) is also an \(\alpha\) vector, as \(A_+\zeta = 0\). Assume, without any loss of generality, that \(\alpha_{j'} = \zeta\), \(1 \leq j' \leq d\). Then,
\[ A_0\alpha_{j'} \succeq 0 \quad \Rightarrow \quad \alpha_{j'}^T A_0^T(-\mu(\theta))_0 < 0, \]
as all elements of \((-\mu(\theta))_0\) are non-zero and negative. This proves, by contradiction, that if the the MLE does not exist, then at least one \(\alpha_j\) vector does not satisfy \(\alpha_j^T(\theta) U(\theta) = 0\), for finite elements of \(\theta\).
To prove the converse, assume an $\alpha_j$, $1 \leq j \leq d$ vector exists, so that $\alpha_j^T(\theta)U(\theta) > 0$ for finite $\theta$. This implies that,

$$
\alpha_j^T A^T_+ (y - \mu(\theta))_+ + \alpha_j^T A^T_0 (y - \mu(\theta))_0 > 0,
$$

$$
\alpha_j^T A^T_0 (\mu(\theta))_0 > 0.
$$

Here, $\alpha_j^T A^T_0 \preceq 0$. This is because $\alpha_j^T A^T_0 = 0$ implies that $\alpha_j^T A^T_0 (\mu(\theta))_0 = 0$, which is not true from (4.12). Now, from all $\alpha_j$’s so that $\alpha_j^T A^T_0 \preceq 0$, we choose the $\alpha_j'$ that corresponds to the set $\{i : (Ax)_i \neq 0\}$ with maximal cardinality. Then, $\alpha_j'$ satisfies the three conditions in (3.10), and the maximum likelihood estimator does not exist. This completes the proof of Theorem 2. 

\[ \square \]

References


