GENERALIZED BILINEAR MODELS

VARTAN CHOULAKIAN
UNIVERSITE DE MONCTON, MONCTON, N.B.

Generalized bilinear models are presented for the statistical analysis of two-way arrays. These models combine bilinear models and generalized linear modeling, and yield a family of models that includes many existing models, as well as suggest other potentially useful ones. This approach both unifies and extends models for two-way arrays, including the ability to treat response and explanatory variables differently in the models, and the incorporation of external information about the variables directly into the analysis. A unifying framework for the generalized bilinear models is provided by considering four particular cases which have been proposed and used in the existing statistical literature. A three-step procedure is proposed to analyze data sets by generalized bilinear models. Two data sets of different nature are analyzed.

Key words: exponential family, quasi-likelihood, Newton's method, Fisher scoring, logit, response variable, explanatory variable, external information.

1. Introduction

This paper uses the basic bilinear model proposed by Takane and Shibayama (1991), coupled with the setup of generalized linear models for the analysis of data arranged in two-way arrays. The resulting models are named generalized bilinear models (GBM). GBM generalize both bilinear models, discussed quite in detail by Takane and Shibayama (1991) and generalized linear models, discussed in detail in the monograph by McCullagh and Nelder (1989). Such generalizations have been suggested in the statistical literature, see for instance Gower (1989). In fact, some particular GBM have been developed in the statistical literature, such as, among others, Goodman's RC(M), U + RC(M), R + RC(M), C + RC(M) and R + C + RC(M) models, refer to Goodman (1979, 1985, 1986, 1991), and "one degree of freedom for nonadditivity" model, see Pettitt (1989). The approach proposed both unifies and extends models for two-way arrays, including the ability to treat response and explanatory variables differently in the models. Often, external information is also available about the variables; for example a qualitative variable can be ordinal. In GBM, external information is directly incorporated into the analysis. Usually the structural part of GBM is composed of two parts: linear and bilinear. The linear part represents the portion of the data which is explained by the external information, while the bilinear part explains the remaining hidden structure. If we delete the bilinear terms in GBM, we get generalized linear models. It is evident from the above discussion that models with mixed additive and multiplicative terms are named bilinear.

This paper is organized as follows: Section 2 introduces GBM using a basic general bilinear model proposed by, among others, Nishisato and Lawrence (1989) and Takane and Shibayama (1991), and presents a three-step procedure for the analysis of a data set. Section 3 discusses four particular cases of GBM which have been proposed and used in the statistical literature. In section 4 two data sets are analyzed.
2. The Framework of Generalized Bilinear Models

In this section we briefly present the basic bilinear model proposed by Takane and Shibayama (1991), see also Nishisato and Lawrence (1989); the reader should refer to the original papers for further details and applications. Let $X = (x_{ij})$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ be a two-way array. We assume that there are an $I$ by $p$ ($p \leq I$) row information matrix, $G$, and a $J$ by $q$ ($q \leq J$) column information matrix, $H$. We consider the following bilinear model

$$X(M) = GSH' + BH' + GC' + UDV', \quad (1)$$

where $S$ ($p \times q$), $B$ ($I \times q$), $C$ ($J \times p$), $U$ ($I \times M$), a diagonal matrix $D$ ($M \times M$), and $V$ ($J \times M$) are matrices of coefficients to be estimated under the following constraints

$$G'mB = 0, \quad (2)$$

$$CnH = 0, \quad (3)$$

$$G'mU = 0, \quad (4)$$

$$V'nH = 0, \quad (5)$$

$$U'mU = I_u, \quad (6)$$

$$V'nV = I_M; \quad (7)$$

in which $m$ is the row metric matrix and $n$ is the column metric matrix, both known in advance; and $I_M$ is the identity matrix of order $M$, where $M = \min (I - p, J - q)$ and it represents the number of bilinear terms in (1). The diagonal elements of the matrix $D$ are arranged in a descending order of magnitude. The four terms in (1) explain some distinct parts of the data matrix $X$: The first term represents the part which can be explained by both $G$ and $H$, the second and third terms represent the parts which can be explained only by $H$ and $G$, respectively; and the last term represents the part which can be explained by neither $H$ nor $G$. Takane and Shibayama (1991) discuss in detail how the parameters are estimated according to the least squares criterion. We generalize the framework of the problem, and adopt the modeling approach associated with the generalized linear models or quasi-likelihood, see McCullagh and Nelder (1989), by assuming the following structure:

1. there is a bilinear predictor, $\eta$, that replaces $X$ in (1), written as

$$\eta(M) = GSH' + BH' + GC' + UDV', \quad (8)$$

where the parameters are estimated under the constraints (2) through (7);

2. $g$ is the link function, and $\eta_{ij} = g(\mu_{ij})$, where $\mu_{ij} = E(x_{ij})$ is the expected value of the $(i, j)$-th element of $X$;

3. $\var(x_{ij}) = \phi(v(\mu_{ij}))$ is the variance function of $x_{ij}$, where $\phi$ is a scale parameter and $v(\mu_{ij})$ is a function of $\mu_{ij}$.

To this three-part structure corresponds a three-step procedure to analyze a data set by GBM. We briefly explain how to handle the three-step procedure in a particular problem. The structural bilinear part (8) describes the nature and the relationship existing between the row and column variables; for example if both these variables are nominal and explanatory, then (8) should be symmetric in the indices $i$ and $j$. The three most used link functions are the canonical, identical and logarithmic link functions, and these will be used in section 4 of this paper. For ease of interpretation, the first of these