Computational Statistics and Data Analysis 66 (2013) 1–7

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Two algorithms for fitting constrained marginal models

R.J. Evans [a,*], A. Forcina [b]

[a] Statistical Laboratory, Wilberforce Road, Cambridge, CB3 0WB, UK
[b] Dipartimento di Economia, Finanza e Statistica, University of Perugia, Italy

ARTICLE INFO

Article history:
Received 13 October 2011
Received in revised form 1 February 2013
Accepted 1 February 2013
Available online 16 February 2013

Keywords:
Categorical data
L_1-penalty
Marginal log-linear model
Maximum likelihood
Non-linear constraint

ABSTRACT

The two main algorithms that have been considered for fitting constrained marginal models to discrete data, one based on Lagrange multipliers and the other on a regression model, are studied in detail. It is shown that the updates produced by the two methods are identical, but that the Lagrangian method is more efficient in the case of identically distributed observations. A generalization is given of the regression algorithm for modelling the effect of exogenous individual-level covariates, a context in which the use of the Lagrangian algorithm would be infeasible for even moderate sample sizes. An extension of the method to likelihood-based estimation under L_1-penalties is also considered.

1. Introduction

The application of marginal constraints to multi-way contingency tables has been much investigated in the last 20 years; see, for example, McCullagh and Nelder (1989), Liang et al. (1992), Lang and Agresti (1994), Glonek and McCullagh (1995), Agresti (2002), and Bergsma et al. (2009). Bergsma and Rudas (2002) introduced marginal log-linear parameters (MLLPs), which generalize other discrete parameterizations including ordinary log-linear parameters and Glonek and McCullagh's multivariate logistic parameters. The flexibility of this family of parameterizations enables their application to many popular classes of conditional independence models, and especially to graphical models (Forcina et al., 2010; Rudas et al., 2010; Evans and Richardson, in press). Bergsma and Rudas (2002) show that, under certain conditions, models defined by linear constraints on MLLPs are curved exponential families. However, naïve algorithms for maximum likelihood estimation with MLLPs face several challenges: in general, there are no closed form equations for computing raw probabilities from MLLPs, so direct evaluation of the log-likelihood can be time consuming; in addition, MLLPs are not necessarily variation independent and, as noted by Bartolucci et al. (2007), ordinary Newton–Raphson or Fisher scoring methods may become stuck by producing updated estimates which are incompatible.

Lang (1996) and Bergsma (1997), amongst others, have tried to adapt a general algorithm introduced by Aitchison and Silvey (1958) for constrained maximum likelihood estimation to the context of marginal models. In this paper, we provide an explicit formulation of Aitchison and Silvey's algorithm, and show that an alternative method due to Colombi and Forcina (2001) is equivalent; we term this second approach the *regression algorithm*. Though the regression algorithm is less efficient, we show that it can be extended to deal with individual-level covariates, a context in which Aitchison and Silvey's approach is infeasible, unless the sample size is very small. A variation of these algorithms, which can be used to fit marginal log-linear models under L_1-penalties, and therefore perform automatic model selection, is also given.

Section 2 reviews the marginal log-linear models and their basic properties, while in Section 3 we formulate the two algorithms, show that they are equivalent and discuss their properties. In Section 4 we derive an extension of the regression

* Corresponding author. Tel.: +44 0 1223 337952; fax: +44 0 1223 337956.
E-mail address: rje42@cam.ac.uk (R.J. Evans).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2013.02.001

algorithm which can incorporate the effect of individual-level covariates. Finally Section 5 considers similar methods for L_1-constrained estimation.

2. Notations and preliminary results

Let $X_j, j = 1, \dots, d$ be categorical random variables taking values in $\{1, \dots, c_j\}$. The joint distribution of $X_1, \dots, X_d$ is determined by the vector of joint probabilities $\boldsymbol{\pi}$ of dimension $t = \prod_1^d c_j$, whose entries correspond to cell probabilities, and are assumed to be strictly positive; we take the entries of $\boldsymbol{\pi}$ to be in lexicographic order. Further, let $\boldsymbol{y}$ denote the vector of cell frequencies with entries arranged in the same order as $\boldsymbol{\pi}$. We write the multinomial log-likelihood in terms of the canonical parameters as

$$l(\boldsymbol{\theta}) = \boldsymbol{y}'\boldsymbol{G}\boldsymbol{\theta} - n \log[\mathbf{1}_t{}' \exp(\boldsymbol{G}\boldsymbol{\theta})]$$

(see, for example, Bartolucci et al., 2007, p. 699); here n is the sample size, $\mathbf{1}_t$ a vector of length t whose entries are all 1, and $\boldsymbol{G}$ a $t \times (t-1)$ full rank design matrix which determines the log-linear parameterization. The mapping between the canonical parameters and the joint probabilities may be expressed as

$$\log(\boldsymbol{\pi}) = \boldsymbol{G}\boldsymbol{\theta} - \mathbf{1}_t \log[\mathbf{1}_t{}' \exp(\boldsymbol{G}\boldsymbol{\theta})] \;\Leftrightarrow\; \boldsymbol{\theta} = \boldsymbol{L}\log(\boldsymbol{\pi}),$$

where $\boldsymbol{L}$ is a $(t-1) \times t$ matrix of row contrasts and $\boldsymbol{L}\boldsymbol{G} = \boldsymbol{I}_{t-1}$.

The score vector, $\boldsymbol{s}$, and the expected information matrix, $\boldsymbol{F}$, with respect to $\boldsymbol{\theta}$ take the form

$$\boldsymbol{s} = \boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi}) \quad \text{and} \quad \boldsymbol{F} = n\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G};$$

here $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$.

2.1. Marginal log-linear parameters

Marginal log-linear parameters (MLLPs) enable the simultaneous modelling of several marginal distributions (see, for example, Bergsma et al., 2009, Chapters 2 and 4) and the specification of suitable conditional independences within marginal distributions of interest (see Evans and Richardson, in press). In the following, let $\boldsymbol{\eta}$ denote an arbitrary vector of MLLPs; it is well known that this can be written as

$$\boldsymbol{\eta} = \boldsymbol{C}\log(\boldsymbol{M}\boldsymbol{\pi}),$$

where $\boldsymbol{C}$ is a suitable matrix of row contrasts, and $\boldsymbol{M}$ a matrix of 0's and 1's producing the appropriate margins (see, for example, Bergsma et al. (2009, Section 2.3.4)).

Bergsma and Rudas (2002) have shown that, if a vector of MLLPs $\boldsymbol{\eta}$ is *complete* and *hierarchical* (two properties defined below), models determined by linear restrictions on $\boldsymbol{\eta}$ are curved exponential families, and thus smooth. Like ordinary log-linear parameters, MLLPs may be grouped into interaction terms involving a particular subset of variables; each interaction term must be defined within a margin of which, it is a subset.

Definition 1. A vector of MLLPs $\boldsymbol{\eta}$ is called *complete* if every possible interaction is defined in precisely one margin.

Definition 2. A vector of MLLPs $\boldsymbol{\eta}$ is called *hierarchical* if there is a non-decreasing ordering of the margins of interest $M_1, \dots, M_s$ such that, for each $j = 1, \dots s$, no interaction term which is a subset of M_j is defined within a later margin.

3. Two algorithms for fitting marginal log-linear models

Here we describe the two main algorithms used for fitting models of the kind described above.

3.1. An adaptation of Aitchison and Silvey's algorithm

Aitchison and Silvey (1958) study maximum likelihood estimation under non-linear constraints in a very general context, showing that, under certain conditions, the maximum likelihood estimates exist and are asymptotically normal; they also outline an algorithm for computing those estimates. Suppose, we wish to maximize $l(\boldsymbol{\theta})$ subject to $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$, a set of r non-linear constraints, under the assumption that the second derivative of $\boldsymbol{h}(\boldsymbol{\theta})$ exists and is bounded. Aitchison and Silvey consider the stationary points of the function $l(\boldsymbol{\theta}) + \boldsymbol{h}(\boldsymbol{\theta})'\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers; this leads to the system of equations

$$\begin{aligned} &\boldsymbol{s}(\hat{\boldsymbol{\theta}}) + \boldsymbol{H}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\lambda}} = \mathbf{0} \\ &\boldsymbol{h}(\hat{\boldsymbol{\theta}}) = \mathbf{0}, \end{aligned} \qquad (1)$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimate and $\boldsymbol{H}$ the derivative of $\boldsymbol{h}'$ with respect to $\boldsymbol{\theta}$. Since these are non-linear equations, they suggest an iterative algorithm which proceeds as follows: suppose that at the current iteration we have $\boldsymbol{\theta}_0$, a value reasonably close to $\hat{\boldsymbol{\theta}}$.