LASSO Penalised Likelihood in High-Dimensional Contingency Tables

Student Oral Presentation

Susana Conde\textsuperscript{1,2} and Gilbert MacKenzie\textsuperscript{1,3}

\textsuperscript{1} Centre of Biostatistics, Department of Mathematics & Statistics, The University of Limerick, Ireland, \texttt{susana.conde@ul.ie} & \texttt{gilbert.mackenzie@ul.ie}
\textsuperscript{2} School of Mathematical Sciences, Room 1.65, Western Gateway Complex, University College Cork, Ireland
\textsuperscript{3} ENSAI, Rennes, France

Abstract: We consider several least absolute shrinkage and selection operator (LASSO) penalized likelihood approaches in high dimensional contingency tables and with hierarchical log-linear models. These include the proposal of a parametric, analytic, convex, approximation to the LASSO. We compare them with “classical” stepwise search algorithms. The results show that both backwards elimination and forward selection algorithms select more parsimonious (i.e. sparser) models which are always hierarchical, unlike the competing LASSO techniques.

Keywords: high dimensional contingency tables; LASSO; model selection; penalized likelihood; stepwise search algorithms.

1 Introduction

Conde and MacKenzie (2008) have recently developed new stepwise search algorithms for binary variables, in the context of high dimensional contingency tables and hierarchical log-linear models. They introduced the idea of measuring dependence between binary comorbidities using interactions in a hierarchical log-linear setting. The algorithms can work with any number of binary variables, and constitute one approach to the problem of model selection in this context.

In this paper we consider a different approach which has been more recently developed in the literature, based on Penalized Likelihood. The idea is to attach a penalty to the usual likelihood function. Different penalties may be adopted to achieve various desirable properties: e.g., sparsity (Friedman, 2008) or smoothness of solutions (Eilers and Marx, 1996), etc. Here we are primarily interested in encouraging sparse solutions in order to identify a more parsimonious model in high-dimensional contingency tables.

We compare our methods with the penalized likelihood approach given by Dahinden et al. (2007), who provided just such an extension for contingency tables using the least absolute shrinkage and selection operator.
 Penalized Likelihood Inference

Given $p$ binary variables, let consider the $p$-dimensional contingency table with $q = 2^p$ cells. If we define $\mu_i = E(Y_i)$, the expected value in the $i$th cell, $i = 1, \ldots, q$ let consider a log-linear regression model with $k$ parameters (with $k \leq q$):

$$\ln(\mu_i) = \sum_{j=1}^{k} a_{ij} \theta_j.$$  \hspace{1cm} (1)

where $A = (a_{ij})$ is a $(q \times k)$ design matrix, $k$ the number of linearly independent parameters; and $\theta$, the vector of unknown parameters measuring the influence of the constant, main effects and interactions on the response. The dimension of $\theta$ is that from the vector space spanned by the columns of $A$. Thus, $A$ has full rank = $k$.

For inference, we consider that the penalized negative log-likelihood is:

$$-\ell^p(\theta, \lambda) = -\ell_{\text{mult}}(\theta) + \text{pen}_{\lambda}$$
where $\ell_{\text{mult}}(\theta)$ is the log-likelihood of a multinomial random variable, and $\text{pen}_\lambda$, the penalty term, is, for $\lambda > 0$

$$(a) : \quad \lambda \sum_{j=2}^{k} |\theta_j|, \quad (b) : \quad \lambda \sum_{j=2+p}^{k} |\theta_j|,$$

and where for the case of the smooth approximation, we have that $\sum_j |\theta_j| \approx \sum_j Q_\omega(\theta_j)$ with $Q_\omega(\theta_j) = \omega \ln \left( \cosh \left( \frac{\theta_j}{\omega} \right) \right)$ for a certain constant $\omega$ that regulates the approximation of the function to that of the absolute value, see Figure 1, whence the penalty term is

$$(c) : \quad \lambda \sum_{j=l}^{k} \omega \ln \left[ \cosh \left( \frac{\theta_j}{\omega} \right) \right]$$

where $l = 2$ or $p + 2$. Note that $Q_\omega(\theta_j) \in C^\infty$, the set of functions that are infinitely differentiable, and is convex. We define then the maximum penalised likelihood estimates (MPLEs), according to the terminology of Green and Silverman (1994) as

$$\hat{\theta}_P(\lambda) := \arg \min_{\theta \in \Theta} \{-\ell_{\text{mult}}(\theta) + \text{pen}_\lambda\}.$$

For a large $\lambda$, all the estimates have go to 0; and for $\lambda = 0$, there is no constraint whence $\hat{\theta}_P(0) \equiv \hat{\theta}$, the MLEs. We estimated the regularization parameter using cross-validation with different folds (5-, 10-, 20-) as required.

We also note that the LASSO penalty is a non-differentiable function, which can complicate optimization. Muggeo (2010) proposes a penalty which is a smoother parametric approximation to the LASSO; nevertheless, that approximation is only once differentiable (Conde, 2011) and standard Newton algorithms require that the function is at least twice differentiable.

### 2.1 HLL Models

We note too that the models derived from penalties encouraging sparsity are not necessarily hierarchical: the penalty and the estimation procedure do not take the hierarchical rules into account. Overall, this is a major disadvantage of the methodology since non-hierarchical models are not invariant to the choice of design matrix (Conde, 2011) and, accordingly, are of no scientific interest. This remark does not apply to simple main effects analysis.

### 2.2 Computation

To compute the solutions of (2), we used the logilasso package, contributed by Dahinden (2007). The package fits log-linear models in sparse
contingency tables using penalized likelihoods. The penalties supported are
the LASSO, the group-L$_1$, and the L$_2$. The \texttt{logilasso} procedure calculates
the estimates of the parameters along a path of $\lambda$s, estimating $\lambda$ by cross-
validation. The functions in the package use a path following algorithm
(Dahinden \textit{et al.}, 2007). They re-scale $\lambda^* = 0$ to 1 so that the latter value
corresponds to $\lambda = +\infty$. The algorithm starts with $\lambda^* = 1$, for which all
the parameters are 0. Then, in each step, it tries to add, to the active set,
which is the set of non-zero parameters, the parameters for which the con-
dition for a minimum in the previous inactive set, has been violated. The
estimates of the parameters are calculated using a Newton formula with
the current $\lambda$ and the previous estimate. Our penalty (b) is not included
in the \texttt{logilasso} package and we used \texttt{nlm} in R and as (c) is an analytic
penalty we again used \texttt{nlm}, obtaining the standard errors directly.

3 Results

In this paper, merely as an illustrative example of the methodology, we
analyze a simulated contingency table, corresponding to $p = 5$, $n = 2000$,
sampled from the model with all two-way interactions present. We used a
design matrix up to and including all the 3-ways. Then, the model has the
constant, 5 main effects, 10 2-way interactions, and 10 3-way interactions,
a total of 25 parameters (without including the constant). The table is in
vector format and Fortran standard order:

$$Y^* = (39, 23, 21, 27, 42, 7, 37, 21, 75, 70, 21, 56, 50, 21, 14, 28,$$
$$87, 55, 9, 21, 46, 13, 12, 4, 325, 520, 28, 129, 103, 61, 10, 25)^T;$$

In Figure 2 we present the graph showing the MPLEs corresponding to the
10 3-way interactions along the path of $\lambda$s, with a LASSO penalty: panel
(a) using the \texttt{logilasso} package (Dahinden, 2007). Irrespectively of the
fold of the cross-validation, the final model found is the same (i.e. 6 three-
ways went to zero); panel (b) using the \texttt{nlm} function in R with the LASSO
penalty; panels (c) and (d), using the \texttt{nlm} function in R with the smooth
LASSO for $w = 1$. The paths of the 3-ways are not stabilized until $\lambda$ is
very large ( panel (c)) and they go to 0 much slower than in the previous
cases. Note that the range of $\lambda$s in panel (d) corresponds to those in panels
(a) and (b). Finally, panels (a) and (b) compare the use of the \texttt{logilasso}
package with \texttt{nlm}.

The final models found using each penalty are: for the LASSO, six three-
way interactions out of the ten went to zero; for the LASSO only in the
interactions, five three-ways went to zero; for the smooth LASSO, if we use
the 95\% confidence interval includes 0 as a cut-off criterion, the method
found the correct model, i.e. the all two-way interactions model. Our step-
wise search algorithms (Conde and MacKenzie, 2008; Conde, 2011) found
either the true model (the backwards elimination algorithms), or a model
FIGURE 2. MPLEs of the 3-way interactions, with a LASSO penalty. (a) Using the package `logilasso`, $\lambda^* \in [0, 1]$; (b) Using the `nlm` function; (c) and (d) Using the smooth approximation for $\omega = 1$. The range of $\lambda$ in (a), (b), and (d) coincide.

with nine two-ways (the Forward Selection algorithm), i.e., in all cases a more parsimonious (and hierarchical) model.

For this example table, the LASSO penalized likelihood method found the least parsimonious or least sparse model, in “regularization” terminology. This is just one table, but it illustrates a direct contradiction to the view that penalized likelihood methods produce sparse(st) solutions. We have many more examples including simulated and real data, and including cases from the $q > p$ scenario, with similar conclusions (Conde, 2011).
4 Final Remarks

As far as we know, this is the first time that penalized likelihood approaches have been compared with “classical” stepwise search algorithms in contingency tables. In the light of the results, we recommend the use of the stepwise algorithms which outperform the LASSO penalized likelihood approaches. We will present more detailed finding at the workshop.

Acknowledgments: The work for this paper was supported Glaxosmithkline (GSK) who funded the first author from 2006-2010. It was also supported by Science Foundation Ireland (SFI, www.sfi.ie) under their Mathematics Initiative, II, via the BIO-SI (www.ul.ie/bio-si) research programme in the Centre of Biostatistics, University of Limerick, Ireland (grant number 07/MI/012). The second author is the PI of BIO-SI project.

References


