DETECTING THE POSITIONS OF NONCONSESUS AMINO ACIDS IN HIV PATIENTS BY MARGINAL LIKELIHOOD THRESHOLDING

Claudia Di Caterina [1]

[1] Department of Economics, University of Verona, (e-mail: claudia.dicaterina@univr.it)

ABSTRACT: We show how marginal likelihood thresholding can be applied in the context of multiple hypothesis testing, proposing a rule to select the tuning parameter involved. For detecting the positions of nonconsensus amino acids in patients suffering from two different HIV variants, we use a logistic regression framework and see that our results are in line with those from standard and advanced procedures controlling the false discovery rate, i.e. the proportion of incorrectly rejected null hypotheses.

KEYWORDS: composite marginal likelihood, logistic regression, multiple testing.

1 Setup and Methods

Let Y be a $p \times 1$ random vector with probability mass or density function $f(y;\theta)$ indexed by the parameter $\theta = (\theta_1,\dots,\theta_p)^\top$, which is sparse in the sense that a small number $p^* \ll p$ of its elements are different from zero. Suppose the full $f(y;\theta)$ is difficult to specify or compute, but we can identify the p conditional univariate marginal distributions of the single Y_js, $f_j(y|x;\theta_j)$ $(j = 1,\dots,p)$ where x is a k-vector of covariates. Specifically, we assume a generalized linear model $\mu_j = E(Y_j) = g^{-1}(\alpha_j + \theta_j x)$ with link function $g(\cdot)$ and dispersion parameter $\phi > 0$.

Given independent observations $(Y^{(i)}, x^{(i)})$ $(i = 1,\dots,n)$, the composite marginal likelihood (CML) estimator $\tilde{\theta}$ (Varin *et al.*, 2011) maximizes

$$\ell(\theta; Y^{(1)},\dots,Y^{(n)}) = \sum_{j=1}^{p} w_j \ell_j(\theta_j; Y^{(1)},\dots,Y^{(n)}), \tag{1}$$

where $\ell_j(\theta_j; Y^{(1)},\dots,Y^{(n)}) = \sum_{i=1}^{n} \log f_j(Y^{(i)}|x^{(i)};\theta_j)$ is the jth marginal log-likelihood and $w = (w_1,\dots,w_p)^\top$ is the design vector of weights that determines which margins are included in (1). Finally, we assume that p grows with the sample size n, but at a slower rate.

1.1 Marginal likelihood thresholding

We review here the method presented in Di Caterina & Ferrari, 2022, for the current setting. Since the marginal log-likelihoods depend on separate parameters, we have $\tilde{\theta}_j = \left\{\theta_j : \sum_{i=1}^n u_j(\theta_j; Y^{(i)}) = 0\right\}$ $(j = 1, \ldots, p)$ where $u_j(\theta_j; y) = \partial \ell_j(\theta_j; y)/\partial\theta_j$ denotes the jth marginal score. Sparsity in the final estimator $\hat{\theta}$ is induced via the marginal likelihood thresholding (MLT)

$$\hat{\theta}_j = \begin{cases} \tilde{\theta}_j & \text{if} \quad \hat{w}_j \neq 0 \\ 0 & \text{if} \quad \hat{w}_j = 0 \end{cases} \qquad (j = 1, \ldots, p),$$

where $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^\top$ is a sparse design vector, selected by minimizing for some $\lambda > 0$ the convex criterion that balances statistical efficiency and sparsity:

$$\hat{d}_\lambda(w) = \frac{1}{2} w^\top \hat{C} w - w^\top \mathrm{diag}(\hat{C}) + \frac{\lambda}{n} \sum_{j=1}^p \frac{|w_j|}{\tilde{\theta}_j^2}, \tag{2}$$

where $\hat{C}$ is the sample covariance matrix of the marginal scores and, if $g(\cdot)$ takes canonical form, has entries $\hat{C}_{jk} = \sum_{i=1}^n (Y_j^{(i)} - \tilde{\mu}_j^{(i)})(Y_k^{(i)} - \tilde{\mu}_k^{(i)})(x^{(i)})^2/(\phi^2 n)$ with $\tilde{\mu}_j^{(i)} = g^{-1}(\hat{\alpha}_j + \tilde{\theta}_j x^{(i)})$.

1.2 Selection of the tuning parameter

The tuning parameter λ is crucial in determining the proportion of nonzero elements in the final MLT estimator $\hat{\theta}$. From the Karush-Kuhn-Tucker (KKT) first-order conditions for the minimization of (2), we find that $\hat{\theta}_j$ is set to zero if the corresponding rescaled z-statistic is smaller than $\sqrt{\lambda}$. This condition is an acceptance region for the null hypothesis $\theta_j = 0$ and suggests that λ may be selected by some form of error control for multiple tests based on the family of hypotheses $\mathcal{H}_\lambda = \{H_0^j : \theta_j = 0 \text{ vs } H_a^j : \theta_j \neq 0, j \in \hat{\mathcal{A}}_\lambda\}$, where $\hat{\mathcal{A}}_\lambda = \{j : \hat{w}_j \neq 0\}$. Rejecting all the hypotheses in $\mathcal{H}_\lambda$ indicates that the selected parameters are probably useful and a larger model could be considered by decreasing λ.

By this rationale, using the asymptotic normality of the z-statistic for θ_j, Slutsky's Theorem and the KKT conditions, if the false discovery rate (FDR) is set equal to $\alpha \in (0,1)$ we obtain the following selection rule for λ:

$$\hat{\lambda} = \inf\left\{\lambda : \frac{\tilde{\theta}_j^2}{SE_j^2} > q_\alpha, \text{ for all } j \in \hat{\mathcal{A}}_\lambda\right\}, \tag{3}$$

where $SE_j = \phi\{\sum_{i=1}^n (Y_j^{(i)} - \tilde{\mu}_j) x^{(i)}\}^{-1}$ if $g(\cdot)$ is canonical and q_α is the upper α-quantile of the χ_1^2 distribution.

2 Analysis of HIV data

We analyze data from Gilbert, 2005, to investigate differences between two variants of HIV. The gag p24 amino acid sequence with $p = 118$ positions was obtained from $n = 146$ individuals, half infected with subtype C (group 1, $n_1 = 73$) and half infected with subtype B (group 2, $n_2 = 73$). For each jth position, the number of subjects with a nonconsesus amino acid was recorded in groups 1 and 2. Our aim is to detect the differentially polymorphic positions, where the probability of a nonconsensus amino acid differs in the two groups.

Both Gilbert, 2005, and Chen *et al.*, 2018, §5, assumed the counts per position were distributed as $Bin(\tau_{jg}, n_g)$ in the gth group $(g = 1, 2)$, computed Fisher's exact statistics to test the null hypotheses $H_0^j : \tau_{j1} = \tau_{j2}$ for $j = 1, \ldots 118$, and adjusted for multiple comparison. They discussed that the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995), which here finds 12 relevant positions controlling the FDR at level $\alpha = 5\%$, has less power and possibly yield unreliable results in discrete settings. Because the first 50 positions have Fisher's exact test statistics with p-values almost surely equal to 1, the BH procedure is expected to be extremely conservative here, meaning to have a FDR much lower than α.

Instead, we model the presence/absence of a nonconsensus amino acid in subject i on position j as $Y_j^{(i)} \sim Ber(\pi(i)_j)$ with $\pi(i)_j = \text{logit}^{-1}(\alpha_j + \theta_j x^{(i)})$, where $x(i)$ is a dummy variable encoding the ith subject's group $(i = 1, \ldots, n)$. We can then apply the MLT method to such logistic regression scenario using $p = 118$ univariate marginal likelihoods: a nonzero estimate of the logit coefficient θ_j will indicate to reject the hypothesis $H_0^j : \theta_j = 0$ and so will identify the jth position as differentially polymorphic.

Since quasi-complete separation occurs when fitting the logistic regression in some positions, it is convenient to set the marginal $\tilde{\theta}_j$s equal to the equally consistent bias-reduced estimates (Firth, 1993). If we choose $\hat{\lambda}$ as described in (3) with $\alpha = 5\%$, we select $\hat{p}^* = 15$ nonzero parameters corresponding to 15 differentially polymorphic positions. This is in line with what found by Gilbert, 2005, via their modified BH procedure. Chen *et al.*, 2018, §5, noticed that the classical BH method applied after excluding the first 50 positions also leads to the same conclusion. In terms of positions selected by MLT, Table 1 shows that 13 out of 15 were identified also by at least another multiple testing procedure conducted on this data set in Gilbert, 2005, and Chen *et al.*, 2018, §5, controlling the FDR at level $\alpha = 5\%$. Note that, when the complete data are analyzed, neither of the FDR-controlling procedures considered selects any

Table 1. *Number of positions selected via MLT classified by alternative FDR-controlling method. A tick indicates the corresponding method also selects those positions at level* $\alpha = 5\%$. *The* * *marks methods run after excluding the first 50 positions.*

# Selected			Modified	adaptive	adaptive
positions	BH	BH*	BH	BH*	BH-Heyse*
by MLT			(Gilbert, 2005)	(Chen *et al.*, 2018)	
7	✓	✓		✓	✓
3			✓		
1		✓		✓	✓
1				✓	✓
1					✓
2					
Tot: 15					

of the first 50 positions. It would then appear sensible that results did not change once those were discarded. Yet this sort of robustness holds only for the modified BH procedure and our proposal.

References

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Chen, X., Doerge, R. W., & Heyse, J. F. 2018. Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, **60**, 761–779.

Di Caterina, C., & Ferrari, D. 2022. Sparse composite likelihood selection. *Pages 423–426 of:* Torelli, N., Bellio, R., & Muggeo, V. (eds), *Proceedings of the 36th International Workshop on Statistical Modelling*, vol. 3.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Gilbert, P. B. 2005. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society C*, **54**, 143–158.

Varin, C., Reid, N., & Firth, D. 2011. An Overview of Composite Likelihood Methods. *Statist. Sinica*, **21**, 5–42.