Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Parsimonious and powerful composite likelihood testing for group difference and genotype–phenotype association

Zhendong Huang, Davide Ferrari *, Guoqi Qian

School of Mathematics and Statistics, University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 6 December 2016
Accepted 6 December 2016
Available online xxxx

Keywords:
Composite likelihood
Wald test
Forward selection
SNPs association test

ABSTRACT

Studying the association between a phenotype and a number of genetic variants from case-control data is an important goal in many genetic studies. Association analysis is often carried out by testing the null hypothesis that two groups of multi-dimensional data are generated by the same population. Testing based on genotype data is a challenging task as the full likelihood of the data is usually intractable. This difficulty may be tackled by composite likelihood (MCL) tests which do not entail the full likelihood. But currently available MCL tests are subject to severe power loss for involving non-informative or redundant sub-likelihoods. To reduce the power loss, a forward search and test method for simultaneous powerful group difference testing and informative sub-likelihoods composition is developed. The new method constructs a sequence of Wald-type test statistics by including only informative sub-likelihoods progressively so as to improve the test power under local sparsity alternatives. Numerical studies show it achieves considerable improvement over the available tests as the modeling complexity grows. The new method is illustrated through an analysis of genotype data from a case-control study on breast cancer.

1. Introduction

Testing population difference between two groups of multivariate data is common in many fields of statistical research. Due to the significant development of data acquisition technologies in recent years, more and more complex data – e.g. involving temporal or spatial dependence among the sample units – can now be readily collected for statistical analysis. However, this entails the use of tractable statistical models that are not easily available. In particular, it may be difficult or even impossible to specify the full likelihood function for testing the group difference. These challenges are common in analyzing case-control data in genotype–phenotype association studies, where for example we test associations between a binary breast cancer phenotype and various genotype variants known as the single nucleotide polymorphisms (SNPs). Note that testing genotype–phenotype association from case-control data can be formulated as a two-sample statistical test problem. But association testing for many genotype variants altogether entails a high-dimensional statistical model, and makes it difficult to formulate a computationally tractable full likelihood (Han and Pan, 2012).

These issues naturally suggest approximating the full likelihood function by a computationally tractable one for constructing the test statistics for association testing. A well-developed approximation is based on the maximum composite likelihood estimator (MCLE), obtained by maximizing the product of low-dimensional sub-likelihood objects instead of the full likelihood. Besag (1974) proposed composite likelihood estimation for spatial data while Lindsay (1988) developed composite likelihood estimation in its generality. Over the years, composite likelihood methods have proved useful in many

* Correspondence to: School of Mathematics and Statistics, Richard Berry Building, Parkville 3010, VIC, Australia.
E-mail address: dferrari@unimelb.edu.au (D. Ferrari).

http://dx.doi.org/10.1016/j.csda.2016.12.004
0167-9473/

applied fields, including geo-statistics, spatial extremes and statistical genetics. See Varin et al. (2011) for a comprehensive survey on methods and applications.

Like the familiar maximum likelihood estimator (MLE), the MCLE is asymptotically unbiased and normally distributed under regularity conditions. This feature, being beneficial for constructing Wald-type statistics for testing group differences (see Geys et al., 1999 and Molenberghs and Verbeke, 2005 among others), can also be used in MCLE based testing. The standard approach here is to form a statistic using all the available data-subsets (so that the MCLE is computed by combining all the feasible sub-likelihood components). Although the resulting Wald test has known null distribution in the limit due to the asymptotic normality of MCLE, it may exhibit unsatisfactory power when the number of parameters in the model is moderate or large relative to the sample size.

In our view, forming a test statistic using all the available sub-likelihoods is not always well-justified from either a statistical or computational perspective. Specifically, when the noise in the data is evident and the statistical model considered is very complex, inclusion of sub-likelihoods that do not explain group differences will mainly be adding noise to the Wald statistic. Clearly, this unwanted noise has the undesirable effect of deteriorating the overall test power. A better strategy would be to choose only informative sub-likelihoods relevant to group differences, while dropping noisy or redundant components as much as possible.

Prompted by the above discussion, we propose a new approach – referred to as the forward step-up composite likelihood (FS-CL) testing – for group difference testing. Given a set of candidate data subsets used for constructing the sub-likelihood objects, our FS-CL method carries out simultaneous testing and data noise reduction by selecting a best set of sub-likelihoods so as to improve the resulting test power. Differently from the existing approaches, we impose a sparsity requirement on our alternative hypothesis reflecting the notion that only a certain portion of data subsets fundamentally explains the difference between groups. While testing the null hypothesis of no difference between groups, our method makes efficient use of data by dropping noisy or redundant data subsets to the maximum extent. This procedure is implemented by a forward search algorithm which, similar to the well-established methods in variable selection, progressively includes one more sub-likelihood at each step until no significant improvement in terms of power is observed.

The new approach proposed can be extended to general linear hypothesis testing (cf. Chapter 7 of Lehmann and Romano, 2005) without fundamental difficulty, but will not be pursued in detail in this paper. The remainder of the paper is organized as follows. In Section 2, we describe the main framework for composite likelihood estimation and overview the existing Wald-type association tests. In Section 3, we describe the new FS-CL methodology and propose the forward search algorithm. In Section 4, we study the finite-sample properties of our method in terms of Type I error probability and power using simulated data. In Section 4.4, we apply our test to the case-control genotype data from the Australian Breast Cancer Family Study. In Section 5, we conclude the paper by providing some final remarks.

2. Composite likelihood inference

2.1. Sparse composite likelihood estimation

Consider a random sample of n observations on a d-dimensional random vector $Y = (Y_1, \ldots, Y_d)^T$ following a probability density function $f(y; \theta)$, with unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^q$ and $q = \dim(\Theta) \geq 1$. Let $\hat{\theta}(w)$ be the profiled maximum composite likelihood estimator (MCLE) of θ, obtained by maximizing the composite likelihood function

$$\ell_{cl}(\theta; w) = \left(\sum_{k=1}^{N_{cl}} w_k\right)^{-1} \sum_{k=1}^{N_{cl}} w_k \ell_k(\theta), \tag{1}$$

where N_{cl} is the total number of sub-likelihood objects being considered, $w = (w_1, \ldots, w_{N_{cl}})^T \in \Omega = \{0, 1\}^{N_{cl}}$ is a vector of binary weights referred to as composition rule, and $\ell_k(\theta) \propto \log f(S_k; \theta)$ is the sub-likelihood defined on the kth data subset S_k. The composite likelihood design is typically user-specified (Varin et al., 2011; Lindsay et al., 2011). For example, ℓ_k can be based on all marginal events ($S_k = \{y_k\}$, $k = 1, \ldots, d$), all pair-wise events ($S_k = \{y_j, y_l\}$, $1 \leq j < l \leq d$), or conditional events ($S_k = \{y_k | y_j,\ j \neq k\}$, $k = 1, \ldots, d$).

In our parsimonious composition framework, each sub-likelihood $\ell_k(\theta)$ is allowed to be selected or not, depending on whether w_k takes the value of 1 or 0, which results in an efficient use of the data. The total number of selected sub-likelihoods, $\|w\| = \sum_{k=1}^{N_{cl}} w_k$, can be much smaller than the total N_{cl} ones available. This is in contrast with the frequently used composite likelihood setting where all the N_{cl} sub-likelihoods are selected. Particularly, in the latter case $w = w_{\text{all}} = (1, \ldots, 1)^T$, and no data noise reduction is attained.

A complication related to notations in composite likelihood is that the parameter θ does not always have all its elements involved in each sub-likelihood $\ell_k(\theta)$. To facilitate presentation in the sequel, we rewrite $\ell_k(\theta)$ as $\ell_k(\theta_k)$ by using θ_k to represent the parameter involved in $\ell_k(\cdot)$. Thus the parameter θ is equivalently represented by $(\theta_1, \ldots, \theta_{N_{cl}})$ in composite likelihood. This necessarily means $(\theta_1, \ldots, \theta_{N_{cl}})$ may contain common elements or elements of known values. For example, if Y follows a d-variate normal distribution $Y \sim N_d(\mu, \sigma^2 I)$ with $\mu = (\mu_1, \ldots, \mu_{d-1}, 0)^T$ and I being the identity matrix, one may define sub-likelihoods using marginal normal distributions $N_1(\mu_k, \sigma_k^2)$, $k = 1, \ldots, d = N_{cl}$ and equate μ_d with 0 and all σ_k^2's with σ^2. In applying parsimonious likelihood composition a subset of $(\theta_1, \ldots, \theta_{N_{cl}})$ indexed by the composition rule

Please cite this article in press as: Huang, Z., et al., Parsimonious and powerful composite likelihood testing for group difference and genotype–phenotype association. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.12.004