ARTICLE IN PRESS

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

Sample size determination for high dimensional parameter estimation with application to biomarker identification

Binyan Jiang [a], Jialiang Li [b,*]

[a] Department of Applied Mathematics, Hong Kong Polytechnic University, Hum Hung, Kowloon, Hong Kong
[b] Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

ARTICLE INFO

Article history:
Received 26 October 2016
Received in revised form 30 April 2017
Accepted 14 August 2017
Available online xxxx

Keywords:
Bernstein inequality
Bonferroni inequality
IDI
NRI
Sample size calculation
Training sample

ABSTRACT

We consider sample size calculation to obtain sufficient estimation precision and control the length of confidence intervals under high dimensional assumptions. In particular, we intend to provide more general results for sample size determination when a large number of parameter values need to be computed for a fixed sample. We consider three design approaches: normal approximation, inequality method and regression method. These approaches are applied to sample size calculation in estimating the Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI) for a diagnostic or screening study. Two medical examples are also provided as illustration. Our results suggest the regression method in general can yield a much smaller sample size than other methods.

1. Introduction

Diagnostic or screening tests are used to detect the patient disease status in medical practice. The accuracy of these tests may be assessed by all kinds of traditional statistical methods such as sensitivity and specificity (Zhou et al., 2011). In recent population studies it becomes more and more imperative to evaluate the accuracy gain when new information such as new biomarker or new model structure has been added into the existing diagnostic procedure. In practice in order to study the general diagnostic accuracy performance of statistical methods, we must obtain an appropriate data set with a reasonable sample size. A study with inadequate sample size may not have sufficient statistical efficiency to achieve a meaningful finding. On the other hand, it may be wasteful and unethical to conduct a study with too large a sample size. There are abundant sample size calculation approaches for various statistical problems; see for example Chow et al. (2007). Within the diagnostic medicine literature, one can find a comprehensive review for sample size calculation in chapter 6 of Zhou et al. (2011). Additionally, Obuchowski and Zhou (2002) considered sample size calculation for diseased and non-diseased subjects required for attaining a prespecified conditional power to test hypotheses regarding diagnostic accuracy measures. Li and Fine (2004) extended earlier sample size formula for case-control studies to prospective cohort studies and provided a justification for the commonly used prevalence inflation method. Steinberg et al. (2009) investigated sample size methods for positive and negative predictive values which may depend on the disease prevalence.

In general, sample size calculation is performed to meet certain optimality criteria, controlling either the Type I/II errors in a hypothesis test problem or the length and confidence level of a confidence interval in an estimation problem (Zhou et al., 2011). Earlier authors (Pencina et al., 2012; Leening et al., 2014) usually prefer the confidence interval approach to make inference in lieu of the hypothesis test approach. In this paper we aim at designing the sample size to attain sufficient

* Corresponding author.
E-mail address: stalj@nus.edu.sg (J. Li).

http://dx.doi.org/10.1016/j.csda.2017.08.010
0167-9473/

Please cite this article in press as: Jiang B., Li J., Sample size determination for high dimensional parameter estimation with application to biomarker identification. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.08.010.

estimation precision and controlling the length of confidence intervals, and we will focus on sample size methods for the interval estimation.

Accuracy measures are frequently reported for biomarker studies where a large number of tests are evaluated simultaneously using the same data set. See Li and Fine (2008) and Li et al. (2013) for examples. Sample size calculation thus needs to acknowledge the high-dimension feature of the data set. A common approach is to use the asymptotic distribution of the estimator. However, if the study does not admit very large number of subjects, asymptotic approximation may be questionable. Without assuming the asymptotic distribution, van der Laan and Bryan (2001) proposed an inequality approach to calculate sample size for the mean estimation using Bernstein's inequality. This method provides a bound for the sample size required for a fixed significance level and only requires the existence of second order moments. We will adopt a similar method of using probability inequalities to calculate the sample size needed to obtain certain estimation accuracy. We further propose a regression approach when a training set is available. This approach may be less conservative than the normal approximation and the inequality approach. A regression calibration method has been used recently in Dobbin and Song (2013) for the estimation of regression coefficients in proportional hazards models. However, the authors considered a deterministic sample size computation under a very complicated calibration model. In this paper, we propose three sample size calculation approaches. The first approach is based on the normal approximation while the second approach is based on the probability inequalities. These methods may lead to very large sample size requirement. A third approach based on regression calibration is also proposed and may provide more realistic sample sizes in practice.

Pencina et al. (2008) proposed two quantitative criteria based on reclassification to directly evaluate the extent to which a new predictor improves classification performance: the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). These new statistics received wide acceptance in health science research. Uno et al. (2013) and Li et al. (2013) extended the formulation of NRI and IDI to failure time outcomes and multicategory outcomes, respectively. Because the NRI and the IDI yield lucid probability assess on diagnostic accuracy improvement, they have both been widely reported and discussed in medical literature since their creation (Pencina et al., 2008). Recently Steyerberg et al. (2010) assessed the performance of prediction models using a variety of methods and metrics and suggested that the NRI and the IDI can gain insight into the value of adding a novel predictor to an established model; Pencina et al. (2012) compared the NRI, the IDI and the ROC curve under nested models and recommended to report these three measures together to characterize the performance of the final model as these three measures offered complementary information. Some authors suggest combining these reclassification statistics with various calibration measures and decision analytic measures to avoid spurious claims of improved prediction and erroneous clinical inference (Pencina et al., 2011, 2012; Leening et al., 2014; Kerr et al., 2014). However, very little research work is available on the design of an epidemiological study for the estimation of the NRI and the IDI. As an application of our approaches we obtain explicit sample size calculation for studies aiming to evaluate the NRI and the IDI.

In the rest of this paper we will first introduce three approaches for sample size calculation, followed by extensive simulations results and two medical examples when these approaches are applied to evaluate the NRI and IDI estimation. Some remarks will also be provided at the end of this paper.

2. Methods

Suppose we are interested in estimating a parameter $\theta = (\theta_1, \ldots, \theta_p)^T \in R^p$ with a sample of size n where $n \ll p$. This is the so-called large-p-small-n setting. We usually construct an estimator $\hat{\theta} = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{np})^T$ from the sample which may have nice asymptotic properties. The research question of this article is to design a sample size n such that the estimation errors of all the covariates are bounded by ϵ with high probability $1 - \alpha$.

2.1. Method 1: normal approximation

In this section we assume that the distribution or asymptotic distributions of $\sqrt{n}\left(\hat{\theta}_{nj} - \theta_j\right)$ is $N\left(0, \sigma_j^2\right)$ for $k = 1, \ldots, p$. This is achievable for many parameter estimation problems. We may use such asymptotic results to compute the sample size. For large n, we have

$$P\left(\frac{\sqrt{n}|\hat{\theta}_{nj} - \theta_j|}{\sigma_j} > z_{\alpha/2}\right) < \alpha, \tag{1}$$

where z_α is the upper α quantile of the standard normal distribution. Let $\epsilon = |\hat{\theta}_{nj} - \theta_j|$ be the anticipated error margin. We obtain the following sample size formula

$$n^* = \frac{z_{\alpha/2}^2 \sigma_j^2}{\epsilon^2}. \tag{2}$$

Using this formula ensures that the estimation error for θ_k is bounded by ϵ with probability $1 - \alpha$. However, this formula is appropriate if we only study one parameter ($p = 1$).

Please cite this article in press as: Jiang B., Li J., Sample size determination for high dimensional parameter estimation with application to biomarker identification. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.08.010.