Computational Statistics and Data Analysis 56 (2012) 2317–2333

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Joint adaptive mean–variance regularization and variance stabilization of high dimensional data

Jean-Eudes Dazard [a,*], J. Sunil Rao [b]

[a] Division of Bioinformatics, Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA

[b] Division of Biostatistics, Department of Epidemiology and Public Health, The University of Miami, Miami, FL 33136, USA

ARTICLE INFO

Article history:
Received 8 October 2010
Received in revised form 8 January 2012
Accepted 10 January 2012
Available online 25 January 2012

Keywords:
Bioinformatics
Inadmissibility
Regularization
Shrinkage estimators
Normalization
Variance stabilization

ABSTRACT

The paper addresses a common problem in the analysis of high-dimensional high-throughput "omics" data, which is parameter estimation across multiple variables in a set of data where the number of variables is much larger than the sample size. Among the problems posed by this type of data are that variable-specific estimators of variances are not reliable and variable-wise tests statistics have low power, both due to a lack of degrees of freedom. In addition, it has been observed in this type of data that the variance increases as a function of the mean. We introduce a non-parametric adaptive regularization procedure that is innovative in that (i) it employs a novel "similarity statistic"-based clustering technique to generate *local*-pooled or *regularized* shrinkage estimators of population parameters, (ii) the regularization is done *jointly* on population moments, benefiting from C. Stein's result on *inadmissibility*, which implies that usual sample variance estimator is improved by a shrinkage estimator using information contained in the sample mean. From these *joint regularized* shrinkage estimators, we derived regularized *t*-like statistics and show in simulation studies that they offer more statistical power in hypothesis testing than their standard sample counterparts, or regular common value-shrinkage estimators, or when the information contained in the sample mean is simply ignored. Finally, we show that these estimators feature interesting properties of variance stabilization and normalization that can be used for preprocessing high-dimensional multivariate data. The method is available as an R package, called 'MVR' ('Mean–Variance Regularization'), downloadable from the CRAN website.

1. Introduction: estimation of population parameters

1.1. Scope—motivation

We introduce a regularization and variance stabilization method for parameter estimation, normalization and inference of data with many variables. In a typical setting, this method applies to high-dimensional high-throughput 'omics'-type data, where the number of variable measurements or input variables (gene, peptide, protein, etc. …) hugely dominates the number of samples (so called $p \gg n$ paradigm). The data may be any kind of continuous covariates.

It is common to deal in high-dimensional setting with the following issues:

- A severe lack of degrees of freedom, generally due to tiny sample sizes ($n \ll 1$), where usual variable-wise estimators lack of statistical power (Storey et al., 2004; Smyth, 2004; Tong and Wang, 2007; Wang et al., 2009) and lead to false positives (Efron et al., 2001; Tusher et al., 2001).

* Corresponding author. Tel.: +1 216 368 3157; fax: +1 216 368 6846.
E-mail addresses: jxd101@case.edu (J.-E. Dazard), JRao@med.miami.edu (J. Sunil Rao).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.01.012

- Spurious correlation and collinearity between a large number of variables ($p \gg 1$) in part due to the nature of the data, but most of which due to an artifact of the dimensionality (see Cai and Lv, 2007 and Fan and Lv, 2008 for a detailed discussion). In addition, False Detection Rates (FDR) get high in part because of the regression-to-the-mean effect induced by correlated parameter estimates (Ishwaran and Rao, 2005).
- Variables in high-dimensional data recurrently exhibit a complex mean–variance dependency with standard deviations severely increasing with the means (Rocke and Durbin, 2001; Huber et al., 2002; Durbin et al., 2002), while statistical procedures usually assume their independence.

In general, statistical inference procedures rely on a set of assumptions about the ideal form of the data such as normality of the measurements or errors, sample group homoscedasticity, and i.i.d variables. These issues make usual assumptions unrealistic, usual moment estimators unreliable (generally biased and inconsistent), and inferences inaccurate. The goal of this method is to get lower estimation errors of mean and variance population parameters and more accurate inferences in high-dimensional data.

1.2. Estimation in high-dimensional setting

A large majority of authors have used *regularization* techniques for estimating population parameters in high dimensional data. The premise is that because many variables are measured simultaneously, it is likely that most of them will behave similarly and share similar parameters. The idea is to take advantage of the parallel nature of the data by borrowing information (pooling) across *similar* variables to overcome the problem of lack of degrees of freedom.

Non-parametric regularization techniques for variance estimation have shown that shrinkage estimators can significantly improve the accuracy of inferences. Jain et al. (2003), proposed a local-pooled error estimation procedure, which borrows strength from variables in local intensity regions to estimate variability. Shrinkage estimation was used by Wright and Simon (2003), Cui et al. (2005) and Ji and Wong (2005). Similarly to Jain et al., Papana and Ishwaran (2006) proposed a strategy to generate an equal variance model. This is a form of variance stabilization that is achieved by *quantile regularization* of sample standard deviations by means of a recursive partitioning (CART-like) algorithm, which was initially used in Bayesian model selection (Ishwaran and Rao, 2005). Tong and Wang proposed a family of optimal shrinkage estimators for variances raised to a fixed power (Tong and Wang, 2007) by borrowing information across variables. The idea of borrowing strength across variables was also recently exploited by Efron in gene sets enrichment analyses (Efron and Tibshirani, 2007), and by Storey's Optimal Discovery Procedure (ODP) to control for compound error rates in multiple-hypothesis testing (Storey, 2007).

Shrinkage estimators have also been successfully combined with empirical Bayes approaches, where posterior estimators have been shown to follow distributions with augmented degrees of freedom, greater statistical power, and far more stable inferences in the presence of few samples (Lonnstedt and Speed, 2002; Smyth, 2004). Following this approach, Baldi and Long estimated population variances by a weighted mixture of the individual variable sample variance and an overall inflation factor selected using all variables (Baldi and Long, 2001). Lonnstedt and Speed (2002) and later Smyth (2004) proposed an empirical Bayes approach that combines information across variables. Kendziorski et al. extended the empirical Bayes method using hierarchical gamma–gamma and log-normal–normal models (Kendziorski et al., 2003).

In a similar vein, shrinkage estimation was also used to generate (Bayesian- or not) "moderated" statistics. There, variable-specific variance estimators are inflated by using an overall offset. Efron et al. derived a t-test that estimates the offset by using a percentile of the distribution of sample standard deviations (Efron et al., 2001). Tusher et al. (2001) and Storey and Tibshirani (2003) added a small constant to the variable-specific variance estimators in their t-test to stabilize the small variances (SAM). Smyth and Cui et al. proposed regularized t-tests and F-tests by replacing the usual variance estimator with respectively a Bayesian-adjusted denominator (Smyth, 2004) or a James–Stein-based shrinkage estimator (Cui et al., 2005).

A commonality to all previous method is that (i) they focus on variance estimation alone, (ii) they involve shrinkage of the sample variance towards a *global* value, which is used for *all* variables. First, regularization of the variance is still a problem if the variance depends on the mean and this dependency is ignored. For instance, denoting by $y_{i,j}$ the individual response (expression level, signal, intensity, ...) of variable $j \in \{1, \ldots, p\}$ (gene, peptide, protein, ...) in sample $i \in \{1, \ldots, n\}$, and the usual population mean estimates by $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}$ and standard deviation estimates by $\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{i,j} - \hat{\mu}_j)^2$, clearly the assumption that a variance estimator can be used in common to all variables (i.e., an equal variance model where $\sigma_j^2 = \sigma_0^2$ for all $j \in \{1, \ldots, p\}$) is unrealistic because of the mean-dependency issue, and because we still expect sampling variability at play even if an homoscedastic model was true. Exploiting the observation that the variance is an unknown function of the mean (Rocke and Durbin, 2001; Huber et al., 2002; Durbin et al., 2002) and Stein's *inadmissibility* result on variance estimators (Stein, 1964), it is clear that shrinkage variance estimates should improve if information contained in the sample mean is known or estimated. In line, Wang recently proposed to use a constant coefficient of variation model and a quadratic variance–mean model for variance estimation as a function of an unknown mean (Wang et al., 2009).

Second, a model that has a variable-specific variance estimator will lack power due to the aforementioned lack-of-degrees-of-freedom issue in high-dimensional data. Using for instance a variable-by-variable z-score transformation such as $y_{i,j}^* = \frac{y_{i,j} - \hat{\mu}_j}{\hat{\sigma}_j}$ for $j \in \{1, \ldots, p\}$, using regular sample mean and standard deviation estimates $\hat{\mu}_j$ and $\hat{\sigma}_j$ of variable j, will generate corresponding variable-specific mean and standard deviation estimates $\hat{\mu}_j^* = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}^*$, and $\hat{\sigma}_j^{*2} =$