AN ANALYTICAL PERSPECTIVE ON CHALLENGES AND FUTURE TRENDS IN GENOMIC DATA ANALYSIS

A. Zollanvari
Department of Electrical and Computer Engineering, Nazarbayev University

(Astana, Kazakhstan)

amin.zollanvari@nu.edu.kz

Predictive modeling of patient risk of a disease using "big" genomic data have great potential to improve healthcare. Big data can be big either in terms of sample size, number of variables, or both. Although a large sample size introduces various issues in terms of storage and computational needs, another subtle problem is raised once we face a large number of variables, namely, how to learn from a large number of variables (high-dimensional observations) and a *relatively* small sample size?

Classical statistical learning techniques have been fashioned for situations in which the sample size *(n)* is much larger than the number of variables *(p)*. This is in large part due to the classical notion of statistical consistency, which guarantees the performance of a statistical technique in situations where the number of measurements unboundedly increases *(n* — oo) for a fixed dimensionality of observation (fixed *p)*. In a finite sample operating regime, this implies that in order to expect an acceptable performance from a statistical technique, we need to have many more sample points (subjects) than variables (genes or SNPs) - a scenario that is exactly the opposite to what we currently face in genomics.

Two mathematical-statistical *machineries* that are potentially capable of constructing techniques for analyzing high-dimensional observations are based on: (1) shrinkage and sparsity assumption; and (2) high-dimensional asymptotics *(n* — oo, p — oo,- — / > 0). Despite remarkable

progress in these areas, many practitioners still utilize classical methods for analyzing high-dimensional datasets. This state of affairs can be attributed to: (1) a lack of knowledge about existing methods developed using these machineries; (2) the ready-to-use computational and statistical software packages that are well developed for classical techniques; and (3) the number of existing methods developed using these machineries is comparably much less than classical large sample techniques. The third issue introduces various research opportunities to develop statistical and signal processing techniques suitable for high-dimensional data analysis. As a simple example to judge the current state of affairs in statistical learning consider the fact that we do not know yet the estimator of the mean vector with minimum quadratic risk for a multivariate Gaussian distribution when the number of variables is as small as three, let alone thousands of variables!