Title:

Classification and dimension reduction for high dimensional data

Brief introduction:

High-dimensional statistics refers to statistical inference when the number of unknown parameters $p$ is much larger than the sample size $n$. This includes regression and supervised classification models, when the number of covariates is of much larger order than $n$, and unsupervised settings, such as clustering, with more variables than observations. Image processing, information retrieval in text documents, food authentication studies are only a few examples of the applications in which such problems arise. In those contexts, standard statistical methods cannot be applied, as the involved matrices are in general not full rank and cannot be inverted. A solution to this problem, which has attracted large attention in the statistical literature, involves imposing a sparse structure on the estimated vector parameters through the introduction of an L1 penalty on their norm. Lasso (Tibshirani 1996) based approaches to regression, classification and dimension reduction methods have been populating the statistical literature since Tibshirani’s seminal paper. See Buhlmann, van de Geer (2011) and Hastie, Tibshirani, Wainwright (2015) for detailed references.

A different approach that is recently finding increasing interest is based on the recourse to random projections. This will be the approach dealt with in the present research proposal.

Background and statement of the problem:

Random projections have recently emerged as a powerful method for dimensionality reduction. Theoretical results indicate that the method preserves distances quite nicely. The original $p$-dimensional data is projected to a $k$-dimensional ($k \ll p$) subspace through the origin, using a random $k \times p$ matrix $R$ whose columns have been generated according to Haar measure (so that they are unit length and orthogonal). Using matrix notation where $X$ ($p \times n$) is the original set of $n$ $p$-dimensional observations, $X_{RP}$ ($k \times n$) = $R(k \times p)X(d \times n)$ is the projection of the data onto a lower $k$-dimensional subspace. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. This means that when $p$ is large compared with $\log n$, we may project the data at random into a lower dimensional space and run the statistical procedure on the projected data, potentially making great computational savings, while achieving comparable or even improved statistical performance.

The idea of combining random projections with ensemble methods has given very nice results in the supervised classification context (Cannings, Samworth 2017) where the task consists in assigning an object (or a number of objects) to one of two or more groups, on the basis of a sample of labelled training data. In the high dimensional context, popular methods such as Fisher’s linear discriminant rule or quadratic discriminant rule cannot be applied, as the sample covariance matrix cannot be inverted.

Cannings and Samworth introduce a very general method for high dimensional classification, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower dimensional space. The random projections are divided into disjoint groups, and within each group the projection yielding the smallest estimate of the test error are selected. The random-projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment. Theoretical results elucidate the effect on performance of increasing the number of projections.

The goal of this research proposal is to analyze the performances of the random projection ensemble classifier in depth and propose solutions for ensemble post pruning and variable selection. We also aim to extend the approach to the solution of one-class and unsupervised classification problems.
Research question or hypothesis, aim, objectives and deliveries:

We aim to address five research lines, which are indeed interconnected from a methodological perspective. Their most relevant aspects are detailed below:

1) Improvements on Cannings and Samworth solution. a) It is well known that the performance of ensemble classifier methods is strongly driven by the degree of the dependence between the classifiers in the ensemble (Brieman 2001). Including in the ensemble negatively dependent classifiers improves the performance while positively correlated classifiers may make the ensemble classifier redundant and may worsen its performance. Many researches have proved that ensemble post-pruning is a relevant strategy for the identification of the ensemble minimizing the misclassification rate. Based on some results we are obtaining regarding the possibility to use the multiplicative binomial distribution for modelling the performance of an ensemble, we would like to explore a strategy for post pruning that involves the parameters of this distribution. b) Ensemble methods are known to have good predictive performances, but because of the ensemble, they become a sort of black box and no longer allow detecting the most relevant variables for classification purposes. We would like to exploit the characteristics of random projections to propose a method that uses the coefficients of the linear combinations that identify the best discriminant projections in order to assess variable importance in classification.

2) Extension to one-class classification. One-class classification is a special classification problem where one class is fully known and the information on the other class are vague (Tax 2001). One typical example is given by food authentication issues where the characteristics of “good” food (target class) are known while the characteristics of “counterfeit” food may arise in many and almost unpredictable ways. Misclassification rate is no longer meaningful in this context. The goal is instead of finding a boundary around the target class so that the probability of wrongly labelling a unit form the target class as “counterfeit” is minimized. We would like to propose a measure of typicality aimed at identifying the boundary and to use the idea underlying random projection ensemble to allow the application of the method to high dimensional data.

3) Extension to clustering. Given the good performances of methods based on random projections in the context of supervised classification, we would like to study extensions of the method to the unsupervised context. The extension is not trivial, as it requires choosing among the existing simple clustering methods, identifying a criterion for choosing the “best random” projection and suggesting a sensible way for producing the ensemble. The last two expects are particularly tricky as they require a reflection of the characteristics one expects for a good clustering of the data and on clustering consensus. See Hennig (2015). The consistency of the recovered clusters will also be study.

4) Can random projections approach benefit from sparsity? This research line aims at merging the sparsity based approach to high dimensional inference with the random projection one with the goal of understanding whether the two approaches may reciprocally benefit. The underlying idea is that sparse random projection may be useful for the identification of relevant variables in both the supervised and the unsupervised classification context.

5) Comparison with latent variable models. Random projection is a dimension reduction method just like latent variable models. The two approaches profoundly differ from the methodological point of view but both may be useful in the high dimensional context. A comparison of the results obtained according to these different approaches will complete the research
The Post doc researcher is supposed to learn in detail the relevant aspects related to clustering, supervised classification and ensemble methods, to find original solutions to the research questions previously described and implement them in R, to analyze the data provided by our partners in the neuroimaging and molecular biology field and to tune the methods accordingly. The results will be presented at conferences and submitted for publication to relevant journals in the multivariate analysis context. Collaboration with our international partners will be required.

Participants in the study and the role they play:

The research will be partially funded by a Grant delivered by American Airforce. Besides the supervisor, it involves prof. Jeremy Jordan of the Air force institute of technology, who is an expert in big data issues, prof. Christian Hennig from UCL, who is an expert in clustering methods, prof Angela Monatanri and Dr Laura Anderlucci who have a wide experience in multivariate classification methods and clustering. It will also benefit from the contributions of researchers in Unibo who deal with images and with molecular biology and who have provided the data that have motivated this research.

References

L. Brieman (2001) Random forests, Machine learning, 45, 5-32


C. Hennig (2015) What are the true clusters?, Pattern Recognition Letters 64, 53-62
