A fast homotopy algorithm for a large class of weighted classification problems and application to phylogeny

Pierre Gutierrez[1,2], Julien Chiquet[2], and Guillem Rigaill[1]

[1] Unité de Recherche en Génomique Végétale INRA-CNRS-Universitée d'Évry Val d'Essonne, Évry, France
[2] Laboratoire Statistique et Génome UMR CNRS 8071-USC INRA-Universitée d'Évry Val d'Essonne, Évry, France

Abstract. We propose a fusion penalty to aggregate a large number of groups starting from multidimensional quantitative data. In the unidimensional case it boils down to solve the one-way ANOVA problem by collapsing the coefficients of K conditions. We introduce a large class of weights for which our homotopy algorithm is in $\mathcal{O}(K\log(K))$. These weights induce a balanced tree structure and simplify the interpretation of the results. Some of these weights also enjoy asymptotic oracle properties. As an example we consider phenotypic data: given one or several traits, we reconstruct a balanced tree structure and assess its agreement with the known phylogeny.

Background

With the advent of new high-throughput technologies, it is possible to compare features across a very large number, K, of conditions. Considering for instance the case of one single feature, one typically applies one-way ANOVA to test for any significant difference between conditions. Large K leads to multiple-testing and algorithmic problems since the number of pairwise tests is in $\mathcal{O}(K^2)$. Furthermore, each test is performed independently and the resulting structure between the conditions is not necessarily simple and easily interpretable.

In this work, we propose a ℓ_1-fusion penalty achieving these goals by constructing a hierarchical structure on the conditions at a low computational cost. Our penalty collapses the coefficients within the conditions in the same manner as the fused-Lasso [1]. We prove that for a large class of weights no split can occur along the path of solutions. These weights lead to a balanced tree structure. Besides adaptive versions of these weights enjoy asymptotically an oracle property. This guarantees selection of the true underlying structure for an appropriate choice of the tuning parameter λ which control the level of aggregation.

In the unidimensional settings, an analogous strategy called "Cas-ANOVA" has been investigated in [2] for multi-factor ANOVA. They propose some weights which enjoys similar asymptotic consistency. Still, these weights to do not lead to a tree as soon as the number of individual by condition is unbalanced. Moreover,

the optimization procedure of [2] is quadratic in K and only provides the solution for a given λ. We also experienced numerical instability using their weights.

In the multidimensional setting a similar penalty was proposed in [3]. When there is just one individual per condition and for fixed weights equal to one, they showed that no split can occur along the path of solutions and proposed an efficient algorithm. However these weights typically lead to unbalanced hierarchies. We extend their results to the case of several individuals per condition and to a larger class of weights that induces a balanced tree structure.

Fast homotopy algorithm for distance decaying weights

The optimization problem that we consider can be solved by the homotopy algorithm proposed in [4]. For unspecified weights, split events may occur in this algorithm. However, the absence of splits is highly desirable because if there is no split,

1. the order of the estimated means always matches the order of the empirical means of each condition;
2. the recovered structure is a tree which simplifies the interpretation;
3. the total number of iterations is guaranteed to be small and equal to K;
4. we avoid maximum flow problems whose resolution is computationally demanding.

We prove that for a large class of weights that induced a balanced tree structure there can be no split in the path of solutions. We implemented both the general and the without split version of the algorithm in C++. For the latter, the complexity of our implementation is $\mathcal{O}(K \log K)$. We also provide a fast cross validation (CV) procedure to select λ. The main idea behind this procedure is to take advantage of the DAG structure of the path of solutions along λ to avoid unnecessary computations.

Application to phylogenetic data

We applied our penalized approach to aggregate various species of bacteria based on one ore several features. We demonstrate the good agreement between our classification and the official phylogeny.

References

1. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B **67**(1) (2005) 91–108
2. Bondell, H., Reich, B.: Simultaneous factor selection and collapsing levels in anova. Biometrics **65**(1) (2008) 169–177
3. Hocking, T., Vert, J.-P.and Bach, F., Joulin, A.: Clusterpath: an algorithm for clustering using convex fusion penalties. In: Proceedings of the 28th ICML. (2011) 745–752
4. Hoefling, H.: A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics **19**(4) (2010) 984–1006