Convex Clustering Shrinkage

K. Pelckmans, J. De Brabanter †, J.A.K. Suykens, and B. De Moor

K.U.Leuven ESAT-SCD/SISTA, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
†: Hogeschool KaHo Sint-Lieven (Associatie KULeuven), Departement Industrieel Ingenieur, 9000 Gent, Belgium
{kristiaan.pelckmans,johan.suykens}@esat.kuleuven.ac.be,
http://www.esat.kuleuven.ac.be/sista/lssvmlab

Abstract. This paper proposes a convex optimization view towards the task of clustering. Herefor, a shrinkage term is proposed resulting in sparseness amongst the differences between the centroids. Given a fixed trade-off term between clustering loss and this shrinkage term, clusters is obtained by solving a convex optimization problem. Varying the trade-off term yields an hierarchical clustering tree. An efficient algorithm for larger datasets is derived and the method is illustrated briefly.

1 Introduction

The term cluster analysis encompasses a number of different algorithms and methods (Tree Clustering, Block Clustering, k-Means Clustering and EM algorithms) for grouping objects of similar kind into respective categories. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Clustering techniques have been applied to a wide variety of research problems. Hartigan [5] provides an excellent summary of the many published studies reporting the results of cluster analyses. There are many books on clustering, including [4] , [3] and [9]. The classic k-Means algorithm was popularized and refined by [5]. The EM algorithm for clustering is described in detail e.g. in [15]. A recent overview of the literature on the subject is given in [16].

Shrinkage techniques, for regression and discriminant analysis, have been studied extensively since the seminal works by [12] and [8]. With a term borrowed from approximation theory, these methods are also called regularization methods [14]. Ridge regression, a fundamental shrinkage construction, was introduced by [7], while the Least Absolute Shrinkage Selection Operator (LASSO) was proposed by [13]. The solution results from solving a quadratic programming problem which can be accelerated by using dedicated decomposition methods (as SMO, [11]), sometimes all solutions corresponding with any regularization trade-off constants (the *solution path*) can be computed efficiently by exploiting the Karush-Kuhn-Tucker conditions for optimality as in the case of Least Angle Regression (LARS) for computing the solution path of the LASSO estimator [2] and in the related SVMs [6].

The following method realizes following objectives: (i) obtaining clustering by solving a convex optimization problem (ii) obtaining an implicit representation of clusters

through the occurence of equal centroids (iii) relating the solution path for varying regularization constants with hierarchical cluster techniques, and (iv) obtaining a method for constructing efficiently the solution path. Section 2 discusses (i) and (ii), while section 3 sketches the issued in (iii) and (iv). Section 4 reports briefly an example.

2 Clustering with Shrinkage

The considered datastructure consists of a centroid $M_i \in \mathbb{R}^D$ corresponding with any datapoints $x_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$.

Definition 1 *An implicit form of clustering using shrinkage can be obtained by solving the following convex programming problem with regularization constant $\gamma > 0$*

$$\min_{M_i} \mathscr{J}_\gamma^p(M_i) = \frac{1}{2}\sum_{i=1}^{N} \|x_i - M_i\|_2^2 + \gamma \sum_{i<j} \left\|M_i - M_j\right\|_p, \tag{1}$$

where $p \geq 1$ is a given constant.

When $\gamma = 0$, there are as many clusters as datapoints as $x_i = M_i$ for all $i = 1, \ldots, N$. When $\gamma \to \infty$, only one cluster remains as the sum of the differences between the centroids tends to zero. We consider the convex cases where $p = 1$ or $p = \infty$. Let $M = \left(M_1^T, \ldots, M_N^T\right)^T \in \mathbb{R}^{N \times D}$ denote a vector (matrix) containing all centroids and let M^d denote the dth column of M. The following result holds for all proper norms $\|\cdot\|_p$:

Proposition 1 *When $\gamma \to \infty$, all centroids become equal to the vector of the empirical mean.*

Proof. When $\gamma \to \infty$, the second part $\sum_{i<j} \|M_i - M_j\|_p \to 0$ which can only be achieved when $M_1 = \cdots = M_N$ if $\|\cdot\|_p$ is a proper norm. Eliminating the individual centroids M_i by using the global $\bar{M} = M_i$ for all $i = 1, \ldots, N$ results into the optimization problem $\min_{\bar{M}} \mathscr{J}(\bar{M}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(x_{i,d} - \bar{M}_d\right)^2$. The necessary and sufficient first order conditions for optimality with respect to the variables $\bar{M}^d$ for all $d = 1, \ldots, M$ become $1_N^T 1_N \bar{M}_d = 1_N^T X^d \Leftrightarrow \bar{M}_d = \frac{1}{N}\sum_{i=1}^{N} x_{i,d}$ for all $d = 1, \ldots, D$ where $x_{i,d}$ denotes the dth variables of the ith sample and $X^d = (x_{1,d}, \ldots, x_{N,d})^T \in \mathbb{R}^N$ is a vector. This concludes the proof. □

The choice of the p-norm influences the behavior of the solution when ranging the parameter γ between 0 and towards ∞. The use of the 1-norm and the ∞-norm results in a solution vector containing typically many zeros, which yields insight into the data in the form a small set of clusters. This result is related to the formulation and analysis of the LASSO estimator [13, 2] in the context of regression. The use of a 2-norm is computationally more eficient but lacks the interpretability of the result as a form of clustering.

2.1 Least absolute clustering shrinkage

In the case when $p = 1$, the shrinkage clustering problem can be written as a quadratic convex programming problem as follows

$$\min_{M_i, t_{ij,d}} \mathscr{J}_\gamma(M_i, t_{ij}) = \frac{1}{2}\sum_{i=1}^{N} \|x_i - M_i\|_2^2 + \gamma \sum_{d=1}^{D}\sum_{i<j} t_{ij,d} \quad \text{s.t.} \quad -t_{ij} \leq \Delta_{ij}^T M^d \leq t_{ij} \ \forall d = 1,\ldots,D, \tag{2}$$

where the scalars $t_{ij,d} \in \mathbb{R}^+$ are slack variables for all $i < j = 1,\ldots,N$ and $d = 1,\ldots,D$. Let $\Delta_{ij} \in \mathbb{R}^N$ be defined as a vector of zeros except for the ith place and the jth place which equals respectively 1 and -1 for all $1 \leq i < j \leq N$.

Proposition 2 *The optimization problem (2) with $DN(N+1)/2$ unknowns and $DN(N-1)$ inequalities can be solved as*

$$\min_{\xi} \frac{1}{2}\sum_{d=1}^{D} \xi^{d^T}\xi^d \ - \ \sum_{d=1}^{D} \xi^{d^T} X^d \quad s.t. \quad G\xi^d \ \leq \ \gamma g \qquad \forall d \ = \ 1,\ldots,D \tag{3}$$

where G and g are defined as

$$\left\{ \forall \xi \in \mathbb{R}^N \;\middle|\; G\xi \leq g \Leftrightarrow \exists a \in \mathbb{R}^{N'} : \ \xi = \Delta^T a \ \ \& \ \ -1_{N'} \leq a^d \leq 1_{N'} \right\}, \tag{4}$$

denoting the projected polytope of the box-constraints.

Proof. The proof is based on the derivation of the dual problem and reducing its complexity. Let $\alpha_{ij,d}, \beta_{ij,d} \in \mathbb{R}^+$ be a positive Lagrange multiplier for all $1 \leq i < j \leq N$ and $d = 1,\ldots,D$. The Lagrangian becomes $\mathscr{L}_\gamma(M_i, t_{ij,d}; \alpha, \beta) = \frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{N}\left(x_{i,d} - M_{i,d}\right)^2 + \gamma\sum_{d=1}^{D}\sum_{i<j} t_{ij,d} + \sum_{i<j}\sum_{d=1}^{D}\alpha_{ij,d}\left(-t_{ij,d} - \Delta_{ij}^T M^d\right) + \sum_{i<j}\sum_{d=1}^{D}\beta_{ij,d}\left(-t_{ij,d} + \Delta_{ij}^T M^d\right)$. Let $T^d \in \mathbb{R}^{N(N-1)/2}$ denote the vector containing all slack variables $t_{ij,d}$ for $1 \leq i < j \leq N$ for all $d = 1,\ldots,D$ and let $\Delta \in \mathbb{R}^{N(N-1)/2 \times N}$ contain all vectors Δ_{ij} for all $1 \leq i < j \leq N$. The Karush-Kuhn-Tucker conditions [1] characterize uniquely the global optimum:

$$\begin{cases} \dfrac{\partial \mathscr{L}_\gamma}{\partial M} = 0 \rightarrow & M^d - X^d = \Delta^T(\alpha^d - \beta^d) & d = 1,\ldots,D \\ \dfrac{\partial \mathscr{L}_\gamma}{\partial T^d} = 0 \rightarrow & \gamma = \alpha_{ij,d} + \beta_{ij,d} & 0 \leq i < j \leq N, \forall d \\ & -T^d \leq \Delta M^d \leq T^d & d = 1,\ldots,D \\ & \alpha_{ij,d}, \beta_{ij,d} \geq 0 & 0 \leq i < j \leq N, \forall d \\ & \alpha_{ij,d}\left(-t_{ij,d} - \Delta_{ij} M^d\right) = 0 & 0 \leq i < j \leq N, \forall d \\ & \beta_{ij,d}\left(-t_{ij,d} + \Delta_{ij} M^d\right) = 0 & 0 \leq i < j \leq N, \forall d, \end{cases} \tag{5}$$

with $\alpha^d = \left(\alpha_{12,d} \ldots, \alpha_{(n-1)n,d}\right)^T \in \mathbb{R}^{N(N-1)/2}$ and $\beta^d = \left(\beta_{12,d} \ldots, \beta_{(n-1)n,d}\right)^T \in \mathbb{R}^{N(N-1)/2}$ for all $d = 1,\ldots,D$. The dual problem becomes $\min_{\alpha,\beta} \frac{1}{2}\sum_{d=1}^{D}(\alpha^d - \beta^d)^T \Delta\Delta^T(\alpha^d - \beta^d) - \sum_{d=1}^{D}(\alpha^d - \beta^d)^T \Delta X^d$ such that $0 \leq \alpha_{ij,d}, \beta_{ij,d}$ and $(\alpha_{ij,d} + \beta_{ij,d}) = \gamma$ for all

$1 \leq i < j \leq N$ and $d = 1, \ldots, D$. This problem can be written equivalently by introducing the unknown vectors $\xi^d = \Delta^T \left(\alpha^d - \beta^d\right) \in \mathbb{R}^N$ and $a^d \in \mathbb{R}^{N(N-1)/2}$ for all $d = 1, \ldots, D$ as follows:

$$\min_{\xi^d, a^d} \frac{1}{2} \sum_{d=1}^{D} \xi^{dT} \xi^d - \sum_{d=1}^{D} \xi^{dT} X^d \quad \text{s.t.} \quad \xi^d = \Delta^T a^d, \quad -\gamma 1_{N'} \leq a^d \leq \gamma 1_{N'} \quad \forall d \tag{6}$$

where $a^d = \left(\alpha^d - \beta^d\right)$ should satisfy the constraints. If one were in the possibility to replace the box constraints in (6) by a set of inequalities on the subspace after projection by Δ^T - say $G \in \mathbb{R}^{K \times N}$ and $g \in \mathbb{R}^K$ - the dual problem can be written as in (3) where we use the property that a rescaled version of the box-constraints results into rescaled bounds $\gamma g \in \mathbb{R}^K$. The final estimate of the primal variables can be recovered from the relation

$$\hat{M}^d = X^d + \hat{\xi}^d \quad \forall d = 1, \ldots, D \tag{7}$$

where $\hat{\xi}^d$ are the solutions to the dual problem (3). □

Algorithm 1 *The following algorithm can be used to derive the matrices $G \in \mathbb{R}^{K \times N}$ and $g \in \mathbb{R}^K$.*

- *Compute C containing all corners $(\pm\gamma, \ldots, \pm\gamma) \in \mathbb{R}^{N(N-1)/2}$ of the box-constraints in $R^{N/(N-1)/2}$ for $\gamma = 1$,*
- *Compute $C' = \Delta^T C$*
- *Compute the convex hull of the projected corners C'*

This conceptual approach is motivated by the fact that any element of the convex hull of C' can be described as a convex combination of the projected corners denoted as $C'^T w$ with $1_N^T w = 1$ and $0_N \leq w \leq 1_N$. Then the vector $C^T w$ (for which $\Delta^T (C^T w) = C'^T w$) satisfies the box-constraints by convexity of the latter. More advanced versions as the Fourier-Motzkin method are described in detail in [17] and implemented in Matlab in the Multi-Parameteric Toolbox (MPT) [10].

3 Hierarchical clustering using shrinkage

This section illustrates how one can use the proposed convex optimization strategy to build a hierarchical cluster tree from the data. It can be prooven that the solutionvector follows a tree-like structure when varying γ from 0 to $+\infty$. It turns out that the computation of the hierarchical structure diagram as presented can be accelerated considerably with respect to the naive approach of computing for each regularization constant the solution explicitly.

Algorithm 2 *Given datapoints $X = \{x_i\}_{i=1}^N$, the hierarchical tree can be computed as*

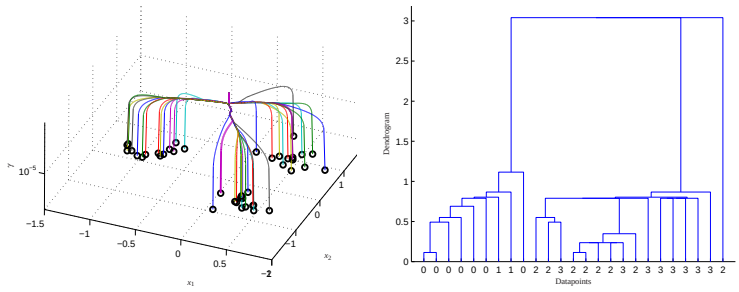
- *Compute G and g given the box constraints $\{-1 \leq a_{ij} \leq 1, \forall i < j\}$ and the matrix Δ^T;*
- *Compute the initial solution for $\gamma = 0$ as $\hat{M}_0 = X$ and $\xi_0 = 0_N$;*

– *Let $\gamma_t > \gamma_{t-1}$, compute the solution $\hat{\xi}_t$ and the corresponding $\hat{M}_t$ to (3) given the set of linear equations G and γg and startingvalue ξ_{t-1}.*

A weakness of this approach is that the critical points $\gamma_0 < \gamma_1 < \cdots < \gamma_t$ are not known explicitly and the grid need to be sufficiently fine to construct the tree.

4 Illustrative example

Two simple examples are elaborated to illustrate the method. The first dataset consists of a sample of 60 datapoints $x_i \in \mathbb{R}^2$ distributed around centra $(0.4, 0)$, $(-0.3, 0.6)$ and $(-0.3, -0.6)$. Figure 1.a displays the clustering tree obtained by varying the regularization constant. The plot shows that the method can discrimate clearly the three classes as can be derived from the three first branches of the tree. Additionally, the iris dataset of the UCI benchmark repository is taken to illustrate the method. Figure 1.b displays the resulting dendrogram using only a subset of the available samples.



(a) (b)

Fig. 1. (a) *The whole solution path for the clustering shrinkage technique for the toy example. The markers 'o' denote the positions of the samples.* **(b)** *Dendrogram for the iris dataset: given the datapoints with labels indicated on the X-axis, the dendrogram displays the hierarchical clustering tree resulting from the algorithm. Only a small subset of the points is shown for graphical convenience. One can see that the presented unsupervised learning strategy already explains the data already partially according to the unseen labels (*$0, 1, 2$ *or* 3*).*

5 Conclusions

This paper presented a convex optimization perspective towards the task of clustering. A technique of shrinkage results in the merging of datapoints into a small set of clusters. Varying the trade-off between shrinkage and clustering loss yields a hierarchical tree representation of the clusters. The technique is presented as a means to detect structure into the data similar as to the case of LASSO estimators. A major advantage of the optimization view is that the whole set of theoretical as well as practical results on

large scale solvers for convex programming problems can be applied to the clustering problem. Further work includes (i) a formal result confirming the tree structure of the solution path (ii) quantifying the effect of shrinkage for different p-norms (appart from sparsity), (iii) finding efficient means to learn the parameter γ from data, (iv) quantification of stability of the presented convex approach.

Acknowledgments. This research work was carried out at the ESAT laboratory of the KUL. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006), Program Sustainable Development PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

1. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
2. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
3. A. Gordon. *Classification*. Chapman and Hall, London, 1999.
4. J. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
5. J. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979. Witten and Frank (2001).
6. T. Hastie, S. Rosset, and R. Tibshirani. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.
7. A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–82, 1970.
8. W. James and C. Stein. Estimation with quadratic loss. In *in Proceedings of the Fourth Berkeley Symposium*, volume 1, pages 311–319, 1960.
9. L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.
10. M. Kvasnica, P. Grieder, and M. Baotić. Multi-Parametric Toolbox (MPT), 2004.
11. J. Platt. Fast training of support vector machines using sequential minimal optimization. *in* Advances in Kernel Methods - Support Vector Learning, *eds. B. Schölkopf and C. Burges and A. Smola*, pages 185–208, 1999. MIT Press.
12. C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *in Proceedings of the Third Berkeley Symposium*, volume 1, pages 197–206, Berkeley, 1955. University of California Press.
13. R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58:267–288, 1996.
14. N. Tikhonov. Regularization of ill-posed problems. *Doklady Akad. Nauk. SSSR*, 1963.
15. I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, New York, 2000.
16. R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
17. G.M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, 1994.