Tumor Phylogenetic Lineage Separation by Medoidshift Clustering with Non-Positive Kernel

Lu Xie
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lxie1@andrew.cmu.edu

Committee
Dr. Russell Schwartz, Dr. Kathryn Roeder, Dr. Roy Maxion

Abstract

Background. Computational approaches have been widely used for dissecting tumor subtypes through mixture models. One successful approach involves formulating this as a problem equivalent to identifying simplices in a high-dimensional space of genomic point clouds, where the vertices of a simplex are assumed to represent unique cell clones such as healthy stem cells, tumor progenitors and tumor subtypes. The highly heterogeneous nature and the complex evolutionary trajectories of tumors pose significant challenges to conventional methods for unmixing tumor composition, which can be simplified by using clustering methods to separate tumor evolution lineages.

Aim. The goal of this project is to design a clustering method that addresses the specific challenges posed by the properties of tumor genomic data: 1) the number of cluster is unknown; 2) data points are clustered according to their belonging simplices; 3) the data clouds from two clusters may overlap; 4) there is no existing probabilistic model yet developed to describe the distribution of the data points inside a cluster; 5) the edges of two simplices may form sharp angles.

Methods. The new clustering method is based on standard medoidshift but coupled with specialized non-positive kernel functions. It converges on vertices of the simplices instead of density maxima, and classify data points with respect to the distance to the converging vertices. The convergence is guaranteed when the shadow kernel function is non-increasing and concave.

Data. Synthetic datasets are produced from 5 hypothetical tumor progression scenarios. The method was also applied on two real breast cancer datasets and one real lung cancer dataset.

Results. The new method shows consistently better clustering than stand medoidshift in all test scenarios, and improvements over k-medoids in certain applications. It also captures some interesting structural characteristics of the real cancer dataset that could not be detected by prior automated approaches to tumor mixture separation.

Conclusions. The medoidshift-based method described in this work better handles the structural constraints of subdividing simplicial complex by finding cluster "centers" at vertices of the simplices, whereas standard medoidshift is heavily influenced by the density on shared margins and may fail to distinguish adjacent simplices. Comparing to the previously used k-medoids method, the new method removed the often difficult model selection question of choosing k, which requires *a priori* knowledge of the number of oncogenetic lineages.

1 Background

Genomics research has dramatically improved our understanding of the nature of tumor progression and the means of possible treatment. Tumor progression, or tumorigenesis, is an evolutionary process by which tumor cells accumulate successive mutations that lead to decreased growth control, increased invasiveness and eventually metastasis. There has been much interest in reconstructing this process of evolution because of its relevance to identifying the driving factors of mutation and predicting drug response and future prognosis. Our understanding of tumor progression has been radically reshaped by the application of new technologies for probing the genome, gene and protein expression profiles of tumors, which have made it possible to identify key sub-types of tumors that may be clinically indistinguishable yet have very distinct prognoses and responses to treatments [1, 2, 3, 4]. Uncovering these tumor sub-types has helped drive the development of novel therapeutics, known as "targeted therapeutics", that are more specifically targeted to the particular genetic defects that cause each cancer [5, 6, 7]. Despite the profound impact molecular genetics had on cancer research, however, we are only starting to embrace the full complexity of tumor evolution. As a result of that, some recognized sub-types remain poorly defined and many patients do not fall into any currently recognized sub-type [8]. Nevertheless, clinical treatment of cancer could receive considerable benefit from better techniques to identify sub-types missed by the prevailing expression clustering approaches, their diagnostic signatures, and the genes essential to the pathogenicity of particular sub-types [8].

More sophisticated computational models have stemmed from the field of phylogenetics, where tumors are not considered as merely random collections of mutated cells, but rather evolving populations. The pioneering work of Desper et al. inferred tumor pylogenies, or oncogenetic trees, by treating observed tumors as leaf nodes in a species tree, estimating evolutionary distances from genomic and gene expression profiles, and applying a variety of methods to obtain reasonable models of the major progression pathways by which tumors evolve across a patient population. [9, 10, 11]. Maximum likelihood estimation has also been used to work with common measurements of tumor state [12].

An alternative to the above tumor-by-tumor phylogenetics, the cell-by-cell approach relies instead on heterogeneity between individual cells within single tumors to identify likely pathways of progression [13, 14, 15, 16, 17]. The Pennington model forms tumor evolution as a Steiner tree problem within individual patients, and uses pooled data from many patients to build a global consensus network describing common evolutionary pathways across a patient population [13, 14]. It is based on the assumption that tumors preserve remnants of their earlier cell populations as they develop, and any given tumor sample will therefore consist of a heterogeneous combination of cells at different stages of progression along a common pathway, as well as possibly contamination by healthy cells of various kinds. This assumption is supported by numerous studies using recent techniques such as fluourescence in situ hybridization (FISH) and single cell sequencing. [15, 16, 17, 18].

Each of the above two methods has its shortcomings. The tumor-by-tumor approach overlooks the intratumor heterogeneity that can provide valuable clues to tumor progression, and the cell-by-cell approach considers the intratumor heterogeneity information but at the cost of limiting the number of probes per cell and therefore the resolution of measure. The more recent single-cell sequencing technique increased the resolution of a single cell measurement to the genomic level, but so far the data is affected by noise and the sample size is limited by the high cost [19, 20, 21, 22, 23].

2 Related works

Schwartz et al. proposed a gap-bridging method that computationally infers cell population from tissue-wide measurements by using "unmixing", a mathematical formalism of the problem of separating fundamental components of mixed samples in which each observation is presumed to be an unknown convex combination of several hidden components. [8]. In the context of tumor phylogeny, each component corresponds to a cell clone ("node") on the underlying oncogenetic tree. Their specific approach views components as vertices of a multi-dimensional simplex that encloses the observed points, which makes unmixing essentially the problem of inferring the vertices and boundaries of the simplex from a survey of the points inside it [24]. Tolliver et al. revised the

unmixing model and derived a "soft" geometric unmixing algorithm that robustly handles noisy observations [25].

Instead of focusing on sub-structural information of the data clouds, other popular mixture inference methods are centered around the probability density estimation that best fits the observed data [26]. Both optimized for data acquired from single-nucleotide polymorphism (SNP) array, ABSOLUTE fits optimal CNV model and uses a karyotype likelihood model as a "prior" [27], and ASCAT uses additional information such as variant allele frequency (VAF) based upon the availability of the raw sequencing data [28]. Some methods also relies on sequencing and VAF data, such as PyClone that uses hierarchical Bayesian clustering [29] and SciClone that uses Bayersian mixture modeling [26]. CNAnorm is a another method designed for sequencing data and infers heterogeneity and copy numbers separately, but it relies on the assumption that tumor is largely monoclonal [30]. Assuming that the tumor sub-populations can be distinguished by CNV information, the THetA algorithm poses heterogeneity inference as a maximum likelihood mixture decomposition problem [31].

The above numerous strategies towards deconvolution of tumor genomic data can typically resolve up to approximately 10 distinct cell types, which limits their ability to resolve finer details of cellular heterogeneity within tumors [32]. To address this issue, Roman et al. proposed a methodological improvement on genomic mixture modeling by exploiting the fact that mixed genomic data from cells evolving according to an evolutionary tree model would be expected to have finer mathematical substructures than a single uniform simplex assumed by prior work. In particular, point clouds produced by representing tumors as points in a genomic space (e.g., by gene expression or gene copy numbers) would be expected to yield simplicial complexes: conjunctions of low-dimensional sub-simplices, corresponding to distinct tumor subtypes, joined to one another via lower-dimensional surfaces corresponding to shared ancestral cell populations including healthy cells [32]. Reconstructing these simplicial complexes can be considered a special case of the technique of manifold learning [33]. In the work of Roman et al., k-medoids clustering is employed to separate point clouds from different sub-simplices, but it requires the pre-selection of cluster number k that corresponds to the number of lineages in the unknown oncogenetic tree.

In this project, the structural-based clustering approach is extended with a medoidshift-based method that is designed to better identify clusters of points on distinct low-dimensional subspaces of the full manifold and remove the often difficult task of model selection regarding the choice of k. The standard medoidshift method automatically seeks the density centers of the point clouds known as "modes", but it may break down in the context of separating individual simplices from simplical complex due to the possibility that density centers may locate on the conjunction surface. The general methodology of medoidshift provides a good platform for designing problem-specific classifier for tumor genomic data, as it is free from any pre-assumption on data distribution or cluster number, and guaranteed to converge with only one-time computation of pairwise distances [34],

3 Problem description

The project is aimed to design a problem-specific algorithm for clustering tumor samples with respect to their belonging simplices in a simplicial complex embedded in high-dimensional feature-space. The vertices of each simplex are assumed to represent the cell clones that map to the nodes on the oncogenetic tree of the particular type of tumor, and the number of shared nodes by two lineages determines the number of vertices shared by two simplices. The input dataset is organized as a matrix whose rows and columns are associated with samples and measurements of certain features, respectively.

4 Method

4.1 Algorithms and implementation

Besides the development of medoidshift with non-positive kernel (NPK), this project also involves standard medoidshfit and k-medoids as competing algorithms, and adjusted rand index as a measure of clustering results. NPK-medoidshift, standard medoidshift and ARI are scripted in MATLAB R2014b, and k-medoids is implemented as a built-in function of MATLAB R2014b with k-means++

seeding. The built-in function *graphallshortestpaths* of MATLAB R2014b is used whenever the shortest L2-squared path is needed as distance metric.

4.2 Standard medoidshift revisited

The derivation of the medoidshift algorithm mainly follows the work of Sheikh et al. [34] and Comaniciu et al. [35]. Similar to meanshift, the medoidshift algorithm is designed to find the modes of estimated kernel density, but only at data points rather than in the whole continuous parameter space:

$$f(x_i) = \frac{1}{c}\sum_{j}^{n}\Phi(z_{ij}) \tag{1}$$

Here $\Phi(\cdot)$ is the kernel function using profile notation, and x_i, x_j are data points such that $1 \leq i \leq n, 1 \leq j \leq n$. The factors c and h are the normalizer and kernel bandwidth, respectively. The distance measure z_{ij} can be any non-negative distance metric between data points x_i and x_j, for example, the scaled L2-squared distance:

$$z_{ij} = \|\frac{x_i - x_j}{h}\|^2 \geq 0 \tag{2}$$

To formulate a medoidshift task, first the original kernel, or "shadow kernel" $\Phi(\cdot)$ must be translated into a medoidshift kernel by:

$$\varphi(\cdot) = -\Phi'(\cdot) \tag{3}$$

where the iterative update rule is:

$$y_{p+1} = \underset{y\in\{x_1\ldots x_n\}}{\arg\min}\sum_{i}^{n}\Big(\|x_i - y\|^2\varphi\big(\|\frac{x_i - y_p}{h}\|^2\big)\Big) \tag{4}$$

Here y_p is the medoid at p^{th} step. In fact, because the sequence of y belongs to the set of data points, one iteration of the above minimization will provide enough information to trace out the convergence trajectory for all data points. The trace terminates when $y_{k+1} = y_k$.

Sheikh et al. proved that medoidshift is guaranteed to converge when the shadow kernel function $\Phi(\cdot)$ is convex. The resulting sequence y has the property that $f(y_{p+q}) > f(y_p)$ for all $q > 0$ (Theorem 2.1 in [34]). Equivalently, a convex shadow kernel means $\varphi'(\cdot) \leq 0$, or $\varphi(x)$ is a monotonically non-increasing function on $x \geq 0$ because of the symmetry of $\Phi(\cdot)$ and $\varphi(\cdot)$.

4.3 Medoidshift with non-positive kernel (NPK)

As discussed in the background section, the standard medoidshift clustering may not work well for separating simplices from a simplicial complex due to the possibility that the simplices, despite occupying distinct subspaces, may yield point clouds so overlapping that will result in a super cluster that includes all adjacent simplices. By the structural characteristics of a simplicial complex, a better strategy is to find the class centers located at density minima, often one vertex that is farthest from the density center for each simplex, and group the points in the containing simplex into one cluster. Intuitively, this can be achieved by reversing the sign of the kernel function in Equation 3:

$$\phi(\cdot) = \Phi'(\cdot) \tag{5}$$

The convergence of medoidshifht with $\phi(\cdot)$ is guaranteed with a concave shadow kernel $\Phi(\cdot)$ and the same update rule in Equation 4 with $\phi(\cdot)$ instead of $\varphi(\cdot)$. To prove that, it is sufficient to show that the resulting sequence y has the property that $f(y_{p+q}) > f(y_p)$ for all $q > 0$. From Equation 1,

$$f(y_{p+1}) - f(y_p) = \frac{1}{c}\sum_i^n \Big(\Phi\big(\|\frac{x_i - y_{p+1}}{h}\|^2\big) - \Phi\big(\|\frac{x_i - y_p}{h}\|^2\big)\Big) \tag{6}$$

and the concavity of $\Phi(\cdot)$,

$$f(y_{p+1}) - f(y_p) \leq \frac{1}{h^2 c}\sum_i^n \Big(\phi\big(\|\frac{x_i - y_p}{h}\|^2\big)\big(\|x_i - y_{p+1}\|^2 - \|x_i - y_p\|^2\big)\Big) \tag{7}$$

By the update rule in Equation 4 and the termination at $y_{p+1} = y_p$,

$$\sum_i^n \Big(\|x_i - y_{p+1}\|^2 \phi\big(\|\frac{x_i - y_p}{h}\|^2\big)\Big) < \sum_i^n \Big(\|x_i - y_p\|^2 \phi\big(\|\frac{x_i - y_p}{h}\|^2\big)\Big) \tag{8}$$

Combining the two above inequalities, we conclude that $f(y_{p+1}) < f(p)$ for all p, thus the sequence $f(y)$ is strictly decreasing. This guarantee of convergence will however break down if a hard cut-off is applied. Hard cut-offs, such as $\Phi(x) = 0$ if $\|x\| > 1$, are common for other kernels but they will break the concavity of $\Phi(x)$. Although the convergence only requires concavity, non-decreasing $\Phi(\cdot)$ such as $\log(z_{ij} + 1)$ will result in an estimated kernel density that is inversely proportional to data density, making a decreasing sequence of $f(\cdot)$ converge at the density maxima. A non-increasing $\Phi(\cdot)$ implies a non-positive first derivative, i.e. $\phi(\cdot) \leq 0$, and the concavity of $\Phi(\cdot)$ implies $\phi'(\cdot) \leq 0$, i.e. $\phi(\cdot)$ is non-increasing.

4.4 2-stage medoidshift

To suppress the noise in data clouds, a standard medoidshift is used as a generic denoiser before using NPK medoidshift. Without knowing the underlying noise distribution, we collapse the data clouds by stepping one iteration forward with standard medoidshift, and then use NPK medoidshift to cluster the remaining points. Labeling is preserved when tracking the trajectory of medoids across the two stages.

4.5 Assessment of clustering results

It is a non-trivial work to find a good measure that tells the quality of clustering, as the values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are hard to define when the number of clusters given by medoidshift is not always the same as ground truth. One promising measure is the adjusted Rand index (ARI), which defines TP, TN, FP and FN according to the pairwise relations among data points [36]:

- TP: a pair of points that belong to the same true cluster are also in the same inferred cluster.
- TN: a pair of points that belong to different true clusters are not in the same inferred cluster.
- FP: a pair of points that belong to different true clusters are in the same inferred cluster.
- FN: a pair of points that belong to the same true cluster are not in the same inferred cluster.

Assume there are v true clusters denoted by $V_i, 1 \leq i \leq v$, w inferred clusters denoted by $W_j, 1 \leq j \leq w$, and $m_{ij} = |V_i \cap W_j|$, then the value of ARI is computed as [37]:

$$ARI = \frac{\sum_i^v \sum_j^w \binom{m_{ij}}{2} - t_3}{t_1 + t_2 - t_3} \tag{9}$$

where

$$t_1 = \frac{1}{2}\sum_i^v \binom{|V_i|}{2}; t_2 = \frac{1}{2}\sum_j^w \binom{|W_j|}{2}; t_3 = \frac{8t_1t_2}{n(n-1)} \tag{10}$$

One advantage of using ARI is its stringent penalty for random clustering and trivial clustering. As mentioned above, standard medoidshift often creates a trivial cluster that includes all data points, which will be rated as low as 0 by ARI but still as much as 0.5 by Jaccard index, another popular measure of clustering results [38, 39].

4.6 Kernel functions and bandwidth selection

In this project, the following shadow kernel and kernel are used with the standard medoidshift method and the first stage of 2-stage medoidshift:

$$\begin{aligned} \Phi_0(z_{ij}) &= \begin{cases} \frac{1}{2}(1-z_{ij})^2 & z_{ij} \leq 1 \\ 0 & z_{ij} > 1 \end{cases} \\ \varphi_0(z_{ij}) &= \begin{cases} 1 - z_{ij} & z_{ij} \leq 1 \\ 0 & z_{ij} > 1 \end{cases} \end{aligned} \tag{11}$$

and the following shadow kernel and kernel are used with NPK medoidshift and the second stage of 2-stage medoidshift:

$$\begin{aligned} \Phi_1(z_{ij}) &= z_{ij} - \exp(z_{ij}) \\ \phi_1(z_{ij}) &= 1 - \exp(z_{ij}) \end{aligned} \tag{12}$$

To provide a relatively fairer comparison across datasets with varying sparsity and distance metric, the kernel bandwidth was not set to a fixed value, but rather decided by the mean of pairwise distance and the choice of a bandwidth factor h_c. The following is an example for L2-squared distance:

$$h^2 = \frac{h_c}{n^2}\sum_i^n\sum_j^n \|x_i - x_j\|^2 \tag{13}$$

and the distance metric is therefore scaled as:

$$z_{ij} = \frac{n^2\|x_i - x_j\|^2}{h_c\sum_i^n\sum_j^n \|x_i - x_j\|^2} \tag{14}$$

5 Dataset

5.1 Synthetic scenarios

Evaluation of the effectiveness of the methods is complicated by the lack of ground truth mixture compositions from real heterogeneous tumor data. Synthetic datasets generated from simulated scenarios with known mixture components, mixture fractions, and simplicial structures therefore provide a relatively fair testbed to quantify effectiveness of the methods. Five tumor evolution scenarios are considered here, each involving a model of tumors evolving into two or three subtypes:

- Scenario T2E consists of a model in which a healthy state evolves into an early precancerous state that subsequently branches into two subtypes of late state (Figure 1 A). This model results in a structure of two triangles sharing an edge (Figure 1 F). One triangle has 50 data points and the other has 100.

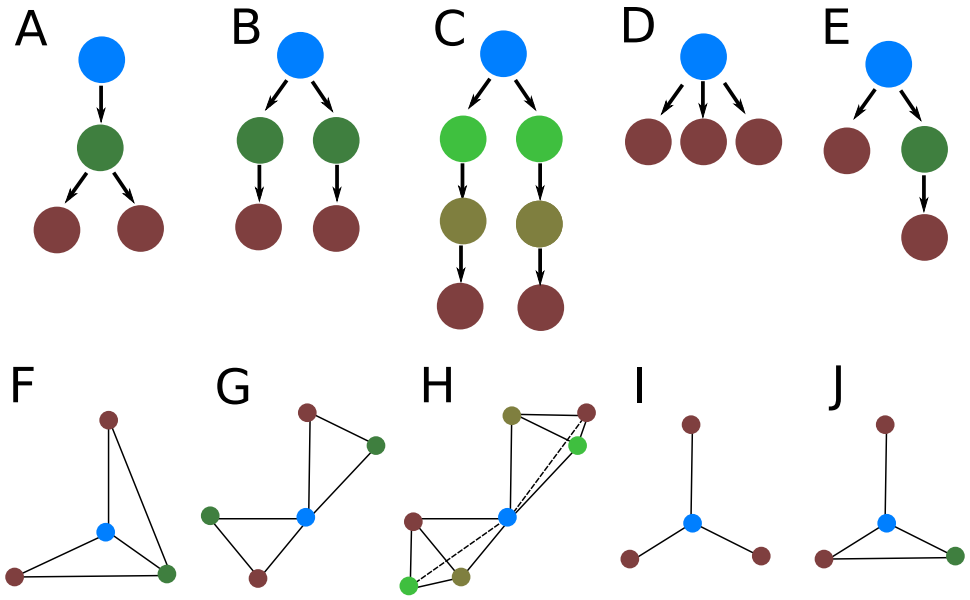

Figure 1: The simulated tumor progression scenarios for producing synthetic datasets. Subfigures A, B, C, D, E show five oncogenetic trees of the synthetic model (T2E, T2V, Q2V, L3V, TLV, respectively), each with healthy cells at root, tumor progenitor cells as internal nodes, and tumor subtypes at leaves. Their corresponding simplicial complex structures are shown in subfigures F, G, H, I, J, respectively, and synthetic data points are randomly drawn from the enclosed simplices. Note that the structures are embedded in 10-Dimensional space while illustrated in 3-D for diagrammatic view.

- Scenario T2V consists of two subtypes each with its own early and a late progression stages (Figure 1 B), yielding a structure of two triangles joined at a vertex (Figure 1 G). One triangle has 50 data points and the other has 100.
- Scenario Q2V consists of a more complex two-subtype model, in which each subtype is defined by early, intermediate, and late progression stages (Figure 1 C). The resulting structure is a pair of tetrathedra joined at a vertex (Figure 1 H). One tetrahedron has 50 data points and the other has 100.
- Scenario L3V assumes an ancestral healthy cell subpopulation that diverges into three discrete subtypes each with a single progression state (Figure 1 D). The result is a simplicial complex structure consisting of three lines joined at a vertex (Figure 1 I). Each of the lines have 25, 50, 75 data points, respectively.
- Scenario TLV consists of two subtypes with unequal number of intermediate stages: one is the direct descendant of healthy cell, the other has one precancerous stage (Figure 1 E). The resulting structure is a line and a triangle sharing a vertex (Figure 1 J). The line has 25 data points and the triangle has 100 data points.

The structures are embedded in 10-Dimensional space by placing the vertex that represents healthy cells at the origin, and each of the other vertices on a unique base vector with unit distance away from the origin. Under noiseless condition, data points are uniformly randomly placed inside the enclosed simplex; and under noisy conditions, Gaussian noises with standard deviation $\sigma = 0.05, 0.1, 0.2$ are added to each dimension of each data point.

5.2 Real tumor data

Using array comparative genomic hybridization (aCGH), a molecular cytogenetic technique for analysing copy number variations (CNVs) relative to ploidy level in the DNA of a test sample compared to a reference sample, Navin et al. assayed CNVs of 83,055 genes from 87 breast tumor samples, forming an 87-by-83,055 matrix [40]. Principle component analysis (PCA) reduced the

dimensionality down to 86, with the projected space consisting of the first 10 principle components (PC) capturing approximately 90% of the data covariance.

Jones et al. measured the expression levels of 40368 genes in 91 lung cancer cells by complementary DNA (cDNA) microarray [41]. The pre-process of the dataset follows previous work in Schwartz et al. [8], where the data points were translated from log space to linear space and the upper and lower limits are set to 2^5 and 2^{-5}, respectively. The values outside the limits were set to the closer limit, and two outliers were discard from the dataset. The dimensionality was reduced to 88 by PCA.

Retrieved from the cancer genome atlas (TCGA), the third dataset consists of 1100 breast tumor samples profiled at 20,243 expression levels with RNASeq [42], a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time. Three outliers were discarded, and the remaining data points were normalized by dimension-wise standard deviation. The dimensionality was reduced to 1096 by PCA.

For all three datasets, the data points projected in the PC space were linearly scaled to fit in a range of [0, 1] in each dimension.

6 Results

The standard medoidshift, NPK medoidshift and 2-stage medoidshift methods were applied to all synthetic scenarios (Figure 2), where data points are randomly generated for 100 replicates per scenario per choice of σ. For the first 2 methods, h_c takes the values from 0.2 to 2 with a step size of 0.2, and for 2-stage medoidshift, h_c is fixed at 1 for the first stage and ranges from 0.2 to 2 with a step size of 0.2 for the second stage. The results of the three methods obtained at $h_c = 1$ are then compared against k-medoids with the data generated under the same scenarios and levels of noise (Figure 3).

By applying NPK medoidshift (with $h_c = 1$) on the first 3-PC space of the real aCGH data, one can find 3 line-shaped clusters joined at one vertex, and the cluster "centers" located on the disjoined endpoints of the lines (Figure 4 B). The spatial distribution of the data points projected in the first 3-PC space may inspire people using 3-medoids clustering, which yields indeed quite similar clusters for aCGH data, but with cluster centers located at the density centers (Figure 4 A).

Applied to the first 5-PC space of the aCGH cancer data, NPK medoidshift (with $h_c = 1$) finds one more cluster that resides near the origin in the first 3-PC space (Figure 5 A). Although indistinguishable in the first 3-PC space, the new cluster shows considerable divergence in the additional PC spaces (Figure 5 B, C) from the other clusters.

The distribution of cDNA data points in the first 3-PC space shows a similar "3-arm" shape to the aCGH case, with the difference being arched "arms" and sharper angles between each pair of "arms" (Figure 4 C, D). Shortest L2 squared path is used as distance measure between points in the aCGH data, and NPK medoidshift (with $h_c = 1$) classified each "arm" as its own cluster with "centers" located again on the disjoined endpoints of the "arms" (Figure 4 D). The application of 3-medoids failed to give desirable clusters that could separate the "arms" (Figure 4 C).

The data clouds in TCGA dataset have a much noisier appearance which makes the separation of potential clusters a much harder task than in the previous two real datasets (Figure 6). Coupled with shortest L2-path distance metric and kernel bandwidth factor $h_c = 1$, the application of 2-stage medoidshift on TCGA dataset discovered 3 clusters in the first 4-PC space (Figure 6). Despite the noisiness, the red cluster shapes like a tetrahedron in the first 3-PC space, and a flat disk in the 2-3-4-PC space; and the blue and black clusters appear to be a triangle and a rod respectively in both 1-2-3-PC and 2-3-4-PC spaces (Figure 6).

7 Discussion

Comparing across the five scenarios, finding the right clusters for Q2V and T2V is much easier than for the others, because under these scenarios the two simplices share only one vertex while having most of the data clouds away from the joining vertex. Scenario T2E is a much harder case due to its density center being placed on the joining edge. In theory, TLV is an easier case than L3V, which is

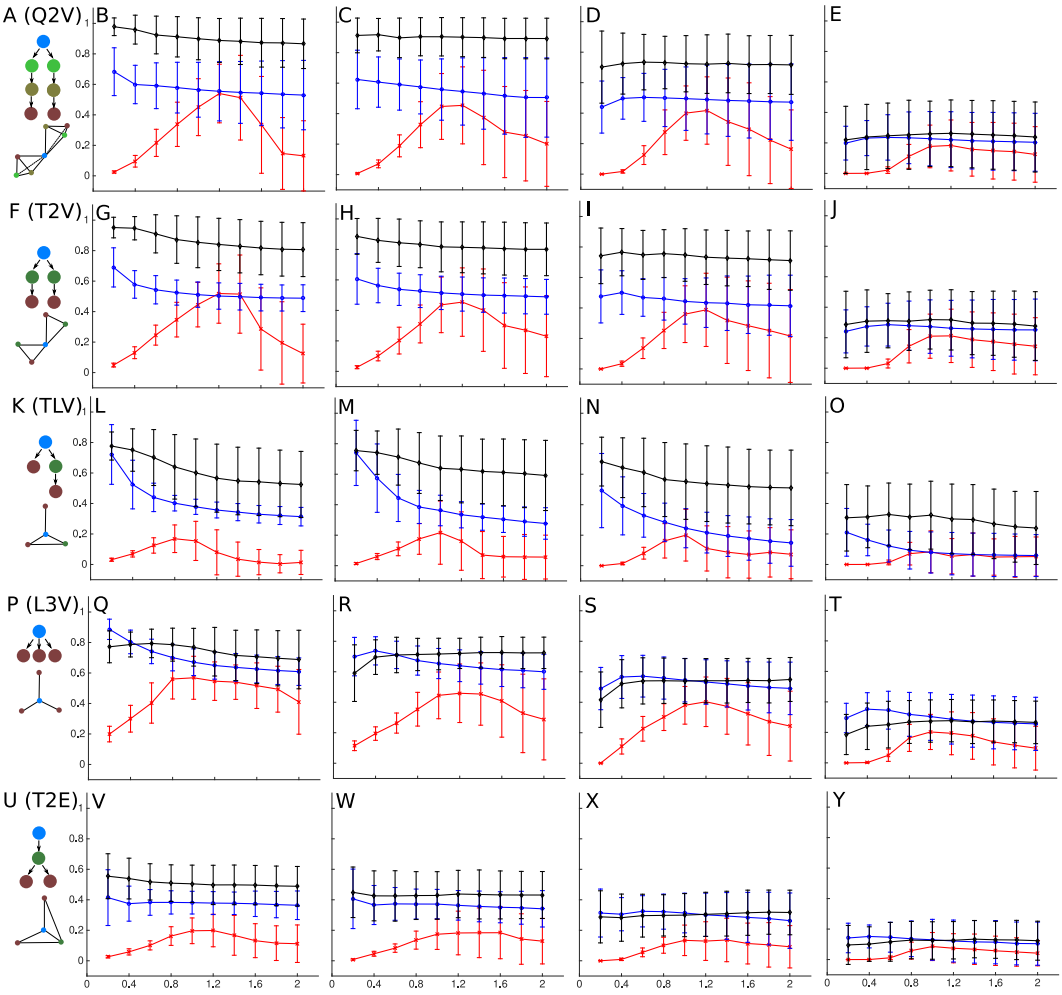

Figure 2: Comparing clustering results of synthetic Scenarios Q2V (B, C, D, E), T2V (G, H, I, J), TLV (L, M, N, O), L3V (Q, R, S, T) and T2E (V, W, X, Y) by standard medoidshift (red curves), NPK medoidshift (blue curves) and 2-stage medoidshift (black curves). Gaussian noise is added to the datasets with $\sigma = 0$ (B, G, L, Q, V), 0.05 (C, H, M, R, W), 0.1 (D, I, N, S, X), and 0.2 (E, J, O, T, Y). For each subfigure of errorbars, the vertical axis indicates the value of ARI for every subfigure, the horizontal axis indicates the kernel bandwidth factor, and the errorbars show one standard deviation above and below the mean ARI value. The structural illustrations of the scenarios are shown in Subfigures A, F, K, P, U, respectively. In general, 2-stage medoidshift (black curves) gives the highest ARI values, while standard medoidshift gives the lowest and most bandwidth-sensitive ARI.

proved by k-medoids, but NPK medoidshift has a slightly lower success rate in clustering TLV than L3V, which is caused by occasionally mistaking the two vertices in the triangle as two endpoints in different simplices. All methods perform poorly at noise level 0.2 under all scenarios.

Comparing across the first three methods, the standard medoidshift yields low and h_c-sensitive ARIs for all scenarios, while 2-stage medoidshift gives overall the highest and most stable ARIs. Scenario TLV is relatively sensitive to kernel bandwidth with NPK medoidshift (Figure 2 L, M, N, P), which prefers the choice of smaller bandwidth value under all noise conditions. The performance of NPK medoidshift is close to 2-stage medoidshift in term of ARI value under Scenario L3V (Figure 2 Q, R, S, T). Adding k-medoids to the comparison, it is obvious that the k-medoids method has slight advantages over 2-stage medoidshift under Scenarios Q2V (Figure 3 B, C, D, E), T2V (Figure 3 G, H, I, J) and TLV (Figure 3 L, M, N), but slight disadvantages under the other two harder scenarios (Figure 3 Q, R, S, V, W, X).

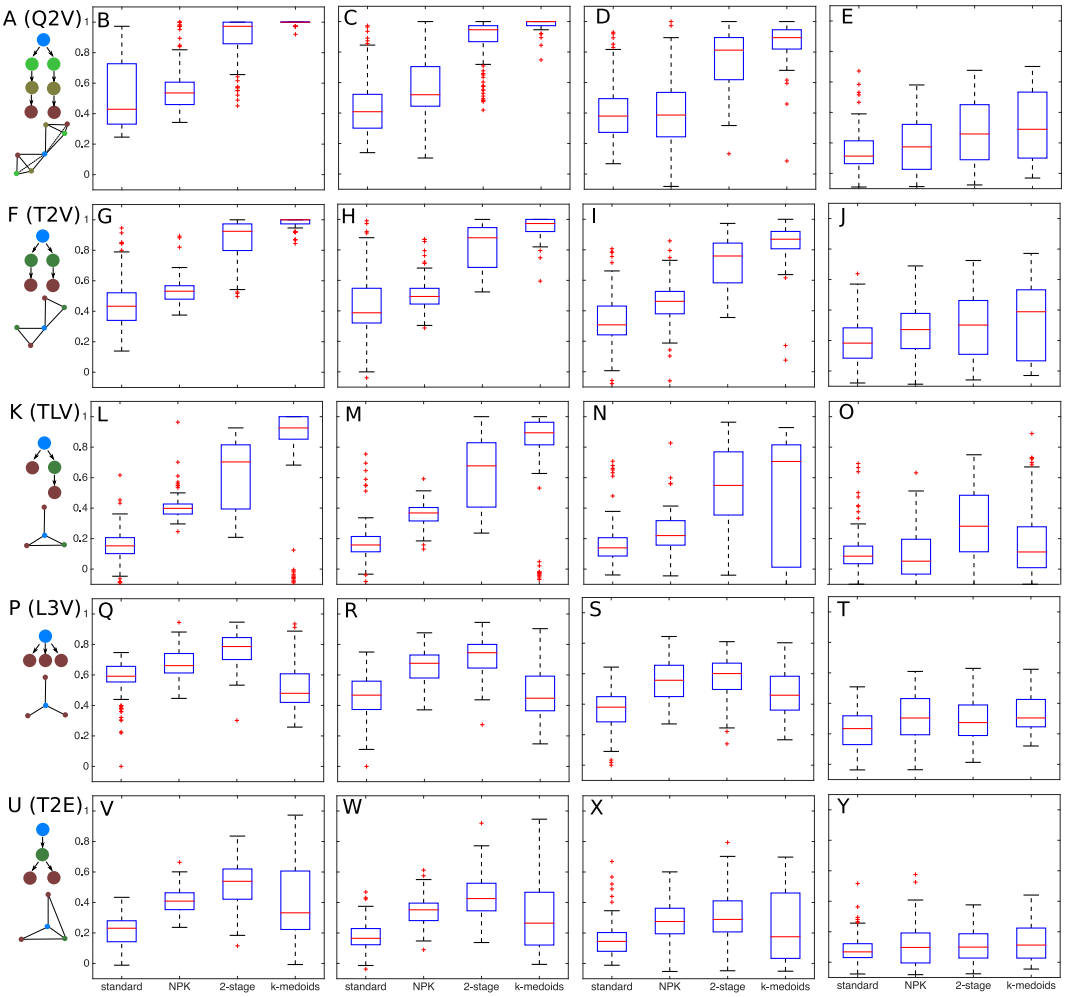

Figure 3: Comparing clustering results of synthetic Scenarios Q2V (B, C, D, E), T2V (G, H, I, J), TLV (L, M, N, O), L3V (Q, R, S, T) and T2E (V, W, X, Y) by standard medoidshift, NPK medoidshift, 2-stage medoidshift and k-medoids with the correct k. Gaussian noise is added to the datasets with $\sigma = 0$ (B, G, L, Q, V), 0.05 (C, H, M, R, W), 0.1 (D, I, N, S, X), and 0.2 (E, J, O, T, Y). For each subfigure of boxplots, the vertical axis indicates the value of ARI, the horizontal axis corresponds to the list of methods, the red bars indicate the median, the boxes include the first to the third quarter quantile, the whiskers reach to extrema, and the red pluses mark the outliers. The structural illustrations of the scenarios are shown in Subfigures A, F, K, P, U, respectively. K-medoids gives the highest ARI under scenarios Q2V, T2V and TLV, while 2-stage medoidshift and NPK medoidshift give higher ARI under scenarios L3V and T2E.

Applications of NPK medoidshift and 2-stage medoidshift on real cancer datasets show reasonable clustering results which one would expect from their sub-structural appearances. In the cases of aCGH and cDNA datasets, where data clouds spread in 3 distinct arms in the first 3-PC space, the NPK medoidshift performs superior than standard medoidshift even k-medoids, which is in accordance with the similar synthetic scenario L3V. In the cases of cDNA and TCGA datasets where inter-cluster distances may be shorter than intra-cluster distances for some data points measured in Euclidean space, using shortest path as distance metric might be able to create the correct clustering as in other cases of manifold learning [33].

In comparison, standard medoidshift failed to produce any desirable clustering with a range of bandwidth when being applied to the first 3-PC space of real aCGH and cDNA data(Figure 9). Affected

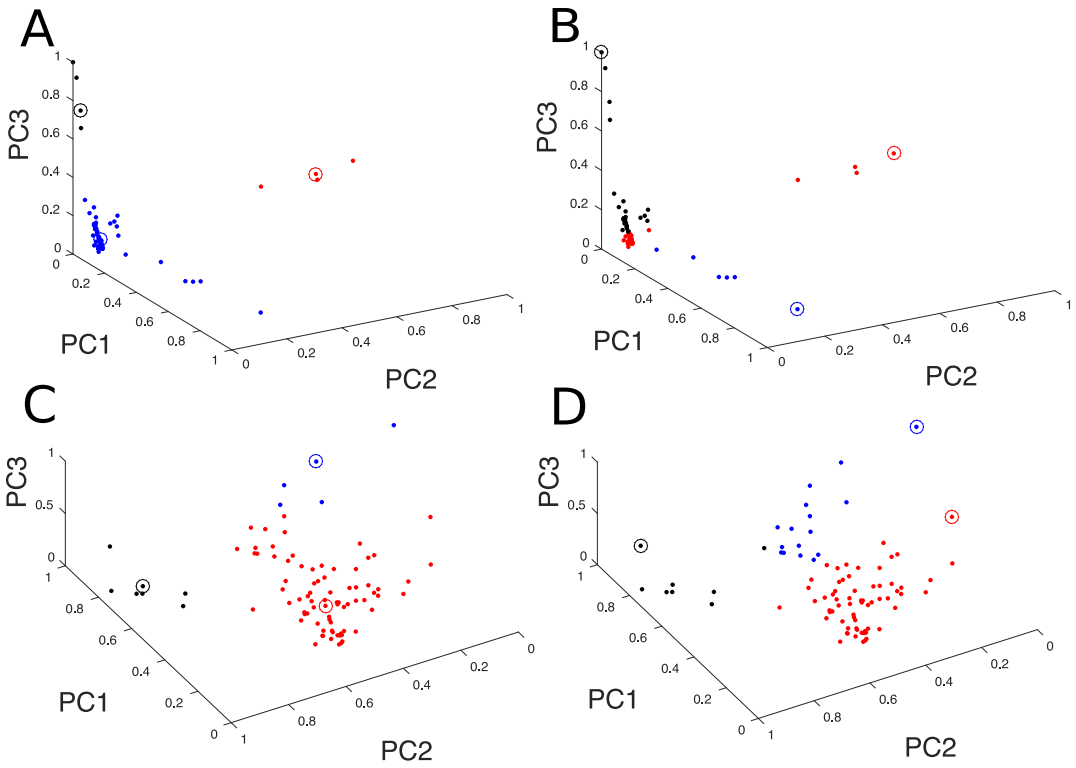

Figure 4: Comparing clustering results of the aCGH (A, B) and cDNA (C, D) data projected in the space consisting of the first 3 PC by 3-medoids (A, C) and NPK medoidshift (B, D). Data points are colored with respect to their belonging clusters, and the circles indicate the corresponding cluster centers. The 3-medoids method found the cluster center located at local density maxima, while NPK medoidshift found them located at vertices farthest away from the conjoint point.

by the fluctuation in the spatial density of the data cloud, a small bandwidth may lead to too many smaller clusters, and a large bandwidth may include two or more clusters in a super cluster.

K-medoids works reasonably well with the correct number of clusters. In the applications on aCGH and cDNA datasets, it seems obvious that k should be set to 3 while clustering the first 3-PC space with k-medoids method in both cases, but adding more PCs to the space hinders the determination of cluster numbers by visually checking the distribution of data points, and the use of k-medoids method therefore requires the inference of k by model selection under these conditions. Medoidshift-based methods do not rely on predefined cluster numbers; instead the number of inferred clusters is influenced by kernel bandwidth, and such influence is less prominent with NPK medoidshift.

8 Limitations

When developing the synthetic scenarios, the vertices of the simplices are assumed to occupy distinct subspaces in the synthetic scenarios, resulting in perpendicular edges joined at the origin. This assumption follows as a consequence of embedding a low-dimensional manifold in a much higher dimension, where it is less likely to have more than one vertex in one dimension. Another simplification in the scenarios is that the vertices have equal distance from the original, which follows the sole criterion for classifying data points with respect to the found cluster centers being minimizing the distance from each data point to the center of its belonging cluster. As a consequence of that, when a vertex is chosen to be the "center" of a simplex by NPK medoidshift, the correctness of clustering relies on the distance from that vertex to the decision boundary. Taking scenario L3V for an example, the resulting clusters might be skewed if the lines have different length measured in

certain distance metric. A possible remedy for more complicated situation in manifold learning is to develop more sophisticated metrics.

Lots of factors in the real world may affect the accuracy of clustering, with an important one being the noisy observations. In the synthetic scenarios, Gaussian noise is added with standard deviation up to $\sigma = 0.2$, and the k-medoids and 2-stage medoidshift can produce meaningful clusters up to $\sigma = 0.1$ for some scenarios. It is non-trivial to estimate the noise induced by experimental measures with different genomic profiling techniques; Su et al. estimated σ to range from 0.05 to 0.38 for RNA-seq [43]. When these datasets usually have much higher dimensionality than the number of data points, PCA may serve as a denoiser but complicates the quantification of the remaining noise after PCA. Two other factors - the surface shared by two or more simplices and the number of data points in each simplex - are inherent to the structure of the tumor phylogenetic tree and the source of the samples. The results from clustering the synthetic datasets show the tendency of a less accurate clustering if a higher proportion of the precancerous states is shared among the lineages. The synthetic scenarios are created deliberately with unequal cluster size, which accounts for the uncertainty in the possible real applications to some extent. The accuracy of every method increases substantially when data points are evenly distributed across the simplices (results not shown).

The clustering results are influenced by more PCs adding to the space (Figure 7, 8), but whether the fluctuation in clustering is a reflection of actual oncogenetic meaning or merely driven by noise requires further oncogenetic analysis.

9 Conclusion

The work in this project fits in the framework of tumor phylogenetic study by facilitating unmixing tumor heterogeneity with the separation of possible phylogenetic lineages. The new-introduced NPK medoidshift method works as a desirable classifier that handles the special challenges in clustering generic tumor genomic data, and is less dependent on the *a priori* knowledge and experimental source of the data. Non-conventional kernels such as $\Phi_1(\cdot)$ and $\phi_1(\cdot)$ may have little meaning in most statistical applications that involve kernel density estimation, but they serve the purpose in finding data points located at geometric extrema of the simplices. Application on the synthetic scenarios proves that 2-stage medoidshift can produce clusters of similar quality to k-medoids with known cluster numbers, and outperforms standard medoidshift method in general. NPK medoidshift and 2-stage medoidshift captured previously undiscovered sub-spacial characteristics of the point clouds from real breast cancer data, but the biological meaning of the discovered clusters in real dataset requires verification from unmixing and further oncogenetic studies.

10 Future works

The results suggest further work may be needed to build more realistic synthetic scenarios that can better reflect the nature of tumor genomic data. Some directions seem promising in making improvements in synthetic data generation, for example, deriving a better data generating method that approximates the noise in real measurements, and designing new scenarios without the assumptions of vertices occupying individual dimension and equal distance from vertices to the origin.

In case of clustering data with more complicated structures, one might consider finding a problem-specific distance metric, but might also benefit from a problem-specific kernel function. Unlike standard kernel density estimation where the choice of kernel functions plays a much less important role than that of the kernel bandwidth, coupling medoidshift with different NPKs may lead to unexpected clustering results. For example, by using a flat kernel $\phi_1(\cdot) = -1$ medoidshift will simply create a single trivial cluster centered at the point that has the largest total distance to the other points:

$$y = \underset{y \in \{x_1 \dots x_n\}}{\arg\min} \sum_{i}^{n} -\|x_i - y\|^2 \tag{15}$$

The use of $\phi_1(z_{ij}) = -z_{ij}^r$ will result in undesirable insensitivity to the choice of bandwidth:

$$\begin{aligned} y_{p+1} &= \underset{y\in\{x_1\ldots x_n\}}{\arg\min} \sum_i^n \|x_i - y\|^2 \big(- \frac{\|x_i - y_p\|^{2r}}{h^{2r}} \big) \\ &= \underset{y\in\{x_1\ldots x_n\}}{\arg\min} \frac{1}{h^{2r}} \sum_i^n -\|x_i - y\|^2 \|x_i - y_p\|^{2r} \\ &= \underset{y\in\{x_1\ldots x_n\}}{\arg\min} \sum_i^n -\|x_i - y\|^2 \|x_i - y_p\|^{2r} \end{aligned} \tag{16}$$

The NPK $\phi_1(z_{ij}) = exp(-z_{ij}) - 1$ has some interesting features. When h is too small, $exp(-z_{ij}) - 1$ approximates -1 and creates trivial clustering. When h is too large, $exp(-z_{ij}) - 1$ approximates $-z_{ij}$ and becomes insensitive to h.

Nevertheless, the results from clustering tumor genomic data, even the candidates from unmixing each cluster, are inconclusive of any biological meaning until further oncogenetic validations have been performed.

Acknowledgment

The author appreciates the advisory of Dr. Russell Schwartz and the effort of the committee members. Special thanks to Theodore Roman for answering countless questions, and to Dr. Roy Maxion for formulating the layout of this report. This work is funded by NIH Award R01CA140214 and the Department of Computational Biology, School of Computer Science, Carnegie Mellon University.

References

[1] Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**: 531-537.

[2] Perou C, Sorlie T, Eisen M, Rijn van der M, Rees S, et al. (2000) Molecular portraits of human breast tumors. *Nature*, **406**: 747-752.

[3] Sorlie T, Perou C, Tibshirani R, Aas T, Geisler S, et al. (2001) Gene expression profiles of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of National Academy of Science*, **98**: 10869-10864.

[4] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, et al. (2003) Repeated observation of breast tumor subtypes in indepednent gene expression data sets. *Proceedings of National Academy of Science*, **100**: 8418-8423.

[5] Pegram M, Konecny G, Slamon D. (2000) The molecular and cellular biology of HER2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer. *Cancer Treatment Research*, **103**: 57-75.

[6] Atkins J, Gershell L. (2002) From the analyst's couch: Selective anticancer drugs. *Nataure Reviews Cancer*, **2**: 645-646.

[7] Bild A, Potti A, Nevins J. (2006) Opinion: Linking oncogenic pathways with therapeutic opportunities. *Nature Reviews Cancer*, **6**: 735-741.

[8] Schwartz R, Shackney S. (2010) Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, **11**(42): 1-20.

[9] Desper R, Jiang F, Kallioniemi O, Moch H, Papadimitriou C, et al. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, **6**: 37-51.

[10] Desper R, Jiang F, Kallioniemi O, Moch H, Papadimitriou C, et al. (2000) Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, **7**: 789-803.

[11] Desper R, Khan J, Schaffer A. (2004) Tumor classification using phylogenetic methods on expression data. *Journal of Theoretical Biology*, **228**: 477-496.

[12] von Heydebreck A, Gunawan B, Fzesi L. (2004) Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, **5**: 545-556.

[13] Pennington G, Smith C, Shackney S, Schwartz R. (2006) Expectation-maximization method for the reconstruction of tumor phylogenies from single-cell data. *Computational Systems Bioinformatics Conference*, 371-380.

[14] Pennington G, Smith C, Shackney S, Schwartz R. (2007) Reconstructing tumor phylogenies from single-cell data. *Journal of Bioinformatics and Computational Biology*, **5**: 407-427.

[15] Smith C, Pollice A, Gu L, Brown K, Singh S, et al. (2000) Correlations among p53, Her-2/neu and ras overexpression and aneuoploidy by multiparameter flow cytometry in human breast cancer: evidence fro a common phenotypic evolutionary pattern in infiltrating ductal carcinomas. *Clinical Cancer Research*, **6**: 112-126.

[16] Janocko L, Brown K, Smith C, Gu L, Pollice A, et al. (2001) Distinctive patterns of Her-2/neu, c-myc and cyclin D1 gene amplification by fluorescence in situ hybridization in primary human breast cancers. *Cytometry*, **46**: 136-149.

[17] Shackney S, Smith C, Pollice A, Brown K, Day R, et al. (2004) Intracellular patterns of Her-2/neu, ras, and ploidy abnormalities in primary human breast cancers predict postoperative clinical disease-free survival. *Clinical Cancer Research*, **10**: 3042-3052.

[18] Shah S, Morin R, Khattra J, Prentice L, Pugh T, et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**: 809-813.

[19] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**(7341): 90-94.

[20] Wang D, Bodovitz S. (2010) Single cell analysis: the new frontier in 'omics'. *Trends in biotechnology*, **28**(6): 281-290.

[21] Tao Y, Ruan J, Yeh S, Lu X, Wang Y, et al. (2011) Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proceedings of the National Academy of Sciences*, **108**(29): 12042-12047.

[22] Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, **148**(5): 873-885.

[23] Xu X, Hou Y, Yin X, Bao L, Tang A, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**(5): 886-895.

[24] Ehrlich R, Full W. (1987) Sorting out geology - unmixing mixtures. *Use and Abuse of Statistical Methods in the Earth Sciences*, Oxford University Press, 33-46.

[25] Tolliver D, Tsourakakis C, Subramanian A, Shackney S, Schwartz R. (2010) Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics*, **26**(ISMB): i106-i114.

[26] Ding L, Wendl M, McMichael J, Raphael B. (2014) Expanding the computational toolbox for mining cancer genomes. *Nature Reviews*, **15**: 556-570.

[27] Carter S, Cibulskis K, Helman E, McKenna A, Shen H, et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, **30**: 412-421.

[28] van Loo P, Nordgard S, Lingjærde O, Russnes H, Rye I, et al. (2010) Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Science*, **107**(39): 16910-16915.

[29] Shah S, Roth A, Goya R, Oloumi A, Ha G, et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**: 395-399.

[30] Gusnanto A, Wood H, Pawitan Y, Rabbitts P, Berri S. (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**(1): 40-47.

[31] Oesper L, Mahmoody A, Raphael B. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, **14**: R80.

[32] Roman T, Nayyeri A, Fasy B, Schwartz R. (2014) A simplicial complex-based approach to unmixing tumor progression data. *Submitted*

[33] Roweis S, Saul L. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500): 2323-2326.

[34] Sheikh Y, Khan E, Kanade T. (2007) Mode-seeking by medoidshift. *IEEE 11th International Conference on Computer Vision (ICCV)*.

[35] Comaniciu D, Meer P. (2002) Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5): 603-619.

[36] Hubert L, Arabie P. (1985) Comparing partitions. *Journal of Classification*, **2**: 193-218.

[37] Kuncheva L, Hadjitodorov S. (2004) Using Diversity in Cluster Ensembles. *IEEE SMC International Conference on Systems, Man and Cybernetics*.

[38] Wagner S, Wagner D. (2007) Comparing Clusterings - An Overview. *http://i11www.iti.uni-karlsruhe.de/extra/publications/ww-cco-06.pdf*

[39] Denœud L, Garreta H, Guénoche A. (2006) Comparison of distance indices between partitions. it Studies in Classification, Data Analysis, and Knowledge Organization, 21-28.

[40] Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. *Genome research*, **20**(1): 68-80.

[41] Jones M, Virtanen C, Honjoh D, Miyoshi C, Satoh Y, et al. (2004) Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*, **363**: 775-781.

[42] Weinstein J, Collisson E, Mills G, Shaw K, Ozenberger B, et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**(10): 1113-1120.

[43] Su Z, Łabaj P, Li S, Thierry-Mieg J, Thierry-Mieg D, et al. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature Biotechnology*, **32**(9): 903-914.

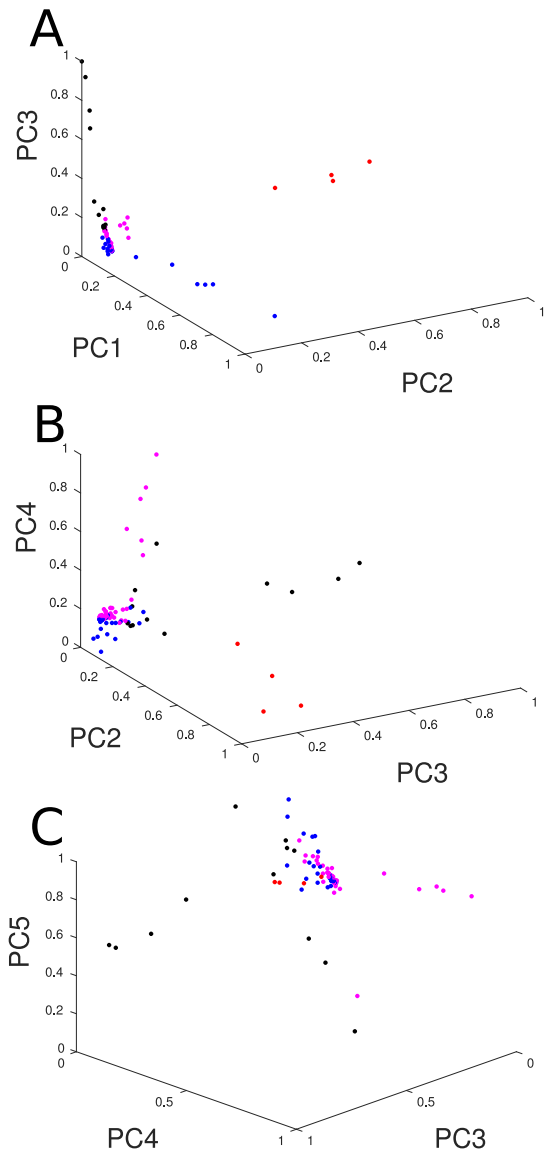

Figure 5: Resulting clusters of the aCGH data projected in the space consisting of the first 5 PC by NPK medoidshift. Data points are colored with respect to their belonging clusters. The data points are shown in 3D spaces consisting of the 1st-2nd-3rd (A), 2nd-3rd-4th (B), and 3rd-4th-5th (C) PC. Despite indistinguishable in Subfigure A, the magenta cluster diverse from the other clusters in Subfigures B and C.

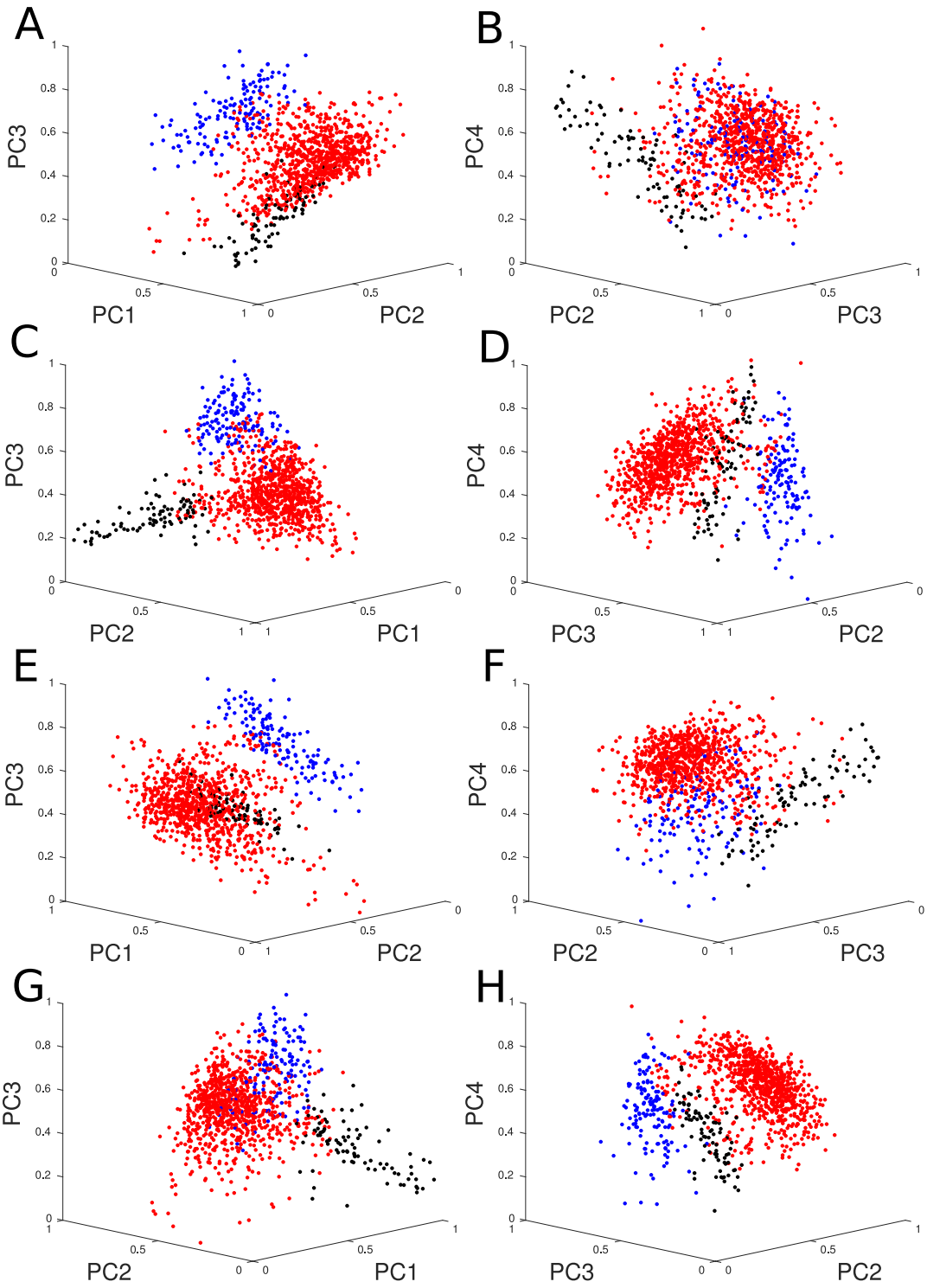

Figure 6: Resulting clusters of TCGA data projected in the space consisting of the first 4 PC by 2-stage medoidshift. Data points are colored with respect to their belonging clusters. The data points are shown in different view angles in 3D spaces consisting of the 1st-2nd-3rd (A, C, E, G) and 2nd-3rd-4th (B, D, F, H) PC. The clusters appear to have distinct shapes and spacial occupation.

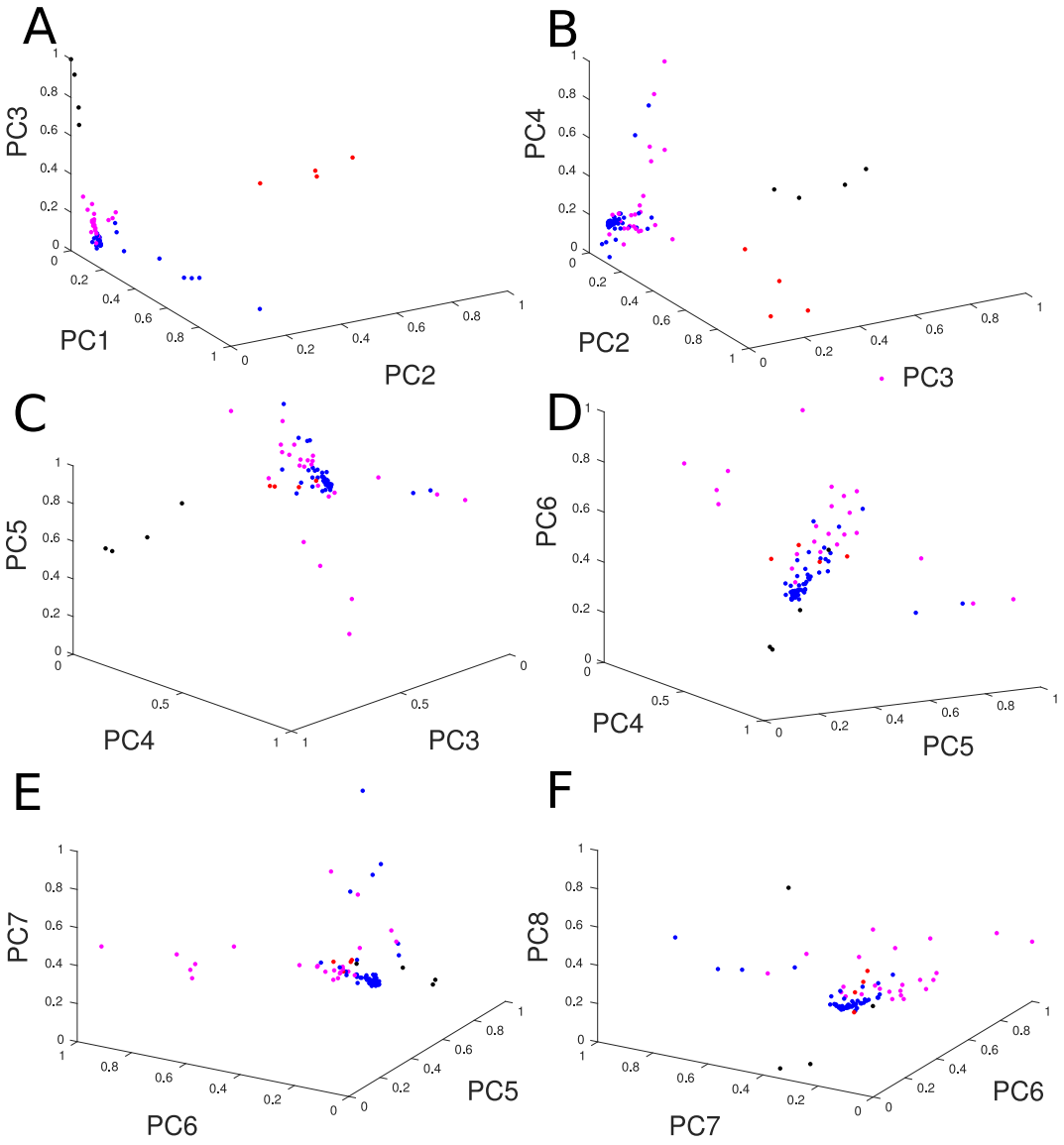

Figure 7: Resulting clusters of aCGH data projected in the space consisting of the first 8 PC by medoidshift with non-positive kernel. Data points are colored with respect to their belonging clusters. The data points are shown in 3D spaces consisting of the 1st-2nd-3rd (A), 2nd-3rd-4th (B), 3rd-4th-5th (C), 4th-5th-6th (D), 5th-6th-7th (E), and 6th-7th-8th (F) PC. The increased dimensionality influences the clustering result with the risk of fitting noise.

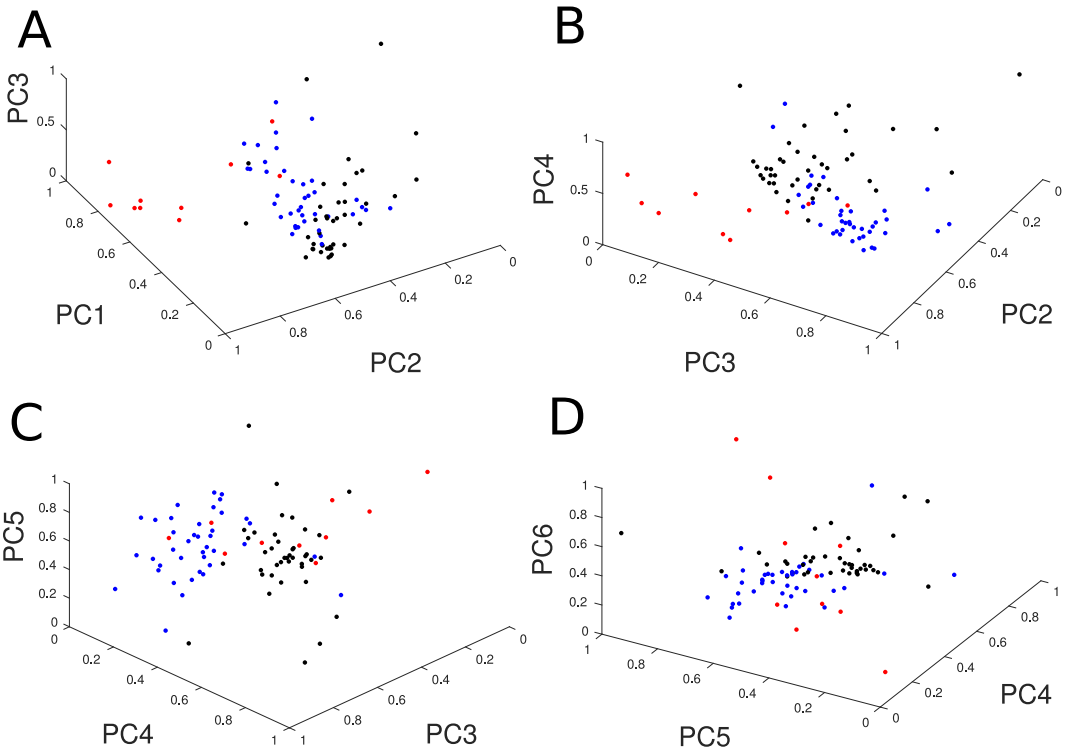

Figure 8: Resulting clusters of cDNA data projected in the space consisting of the first 6 PC by medoidshift with non-positive kernel. Data points are colored with respect to their belonging clusters. The data points are shown in 3D spaces consisting of the 1st-2nd-3rd (A), 2nd-3rd-4th (B), 3rd-4th-5th (C), and 4th-5th-6th (D) PC. The increased dimensionality influences the clustering result with the risk of fitting noise.

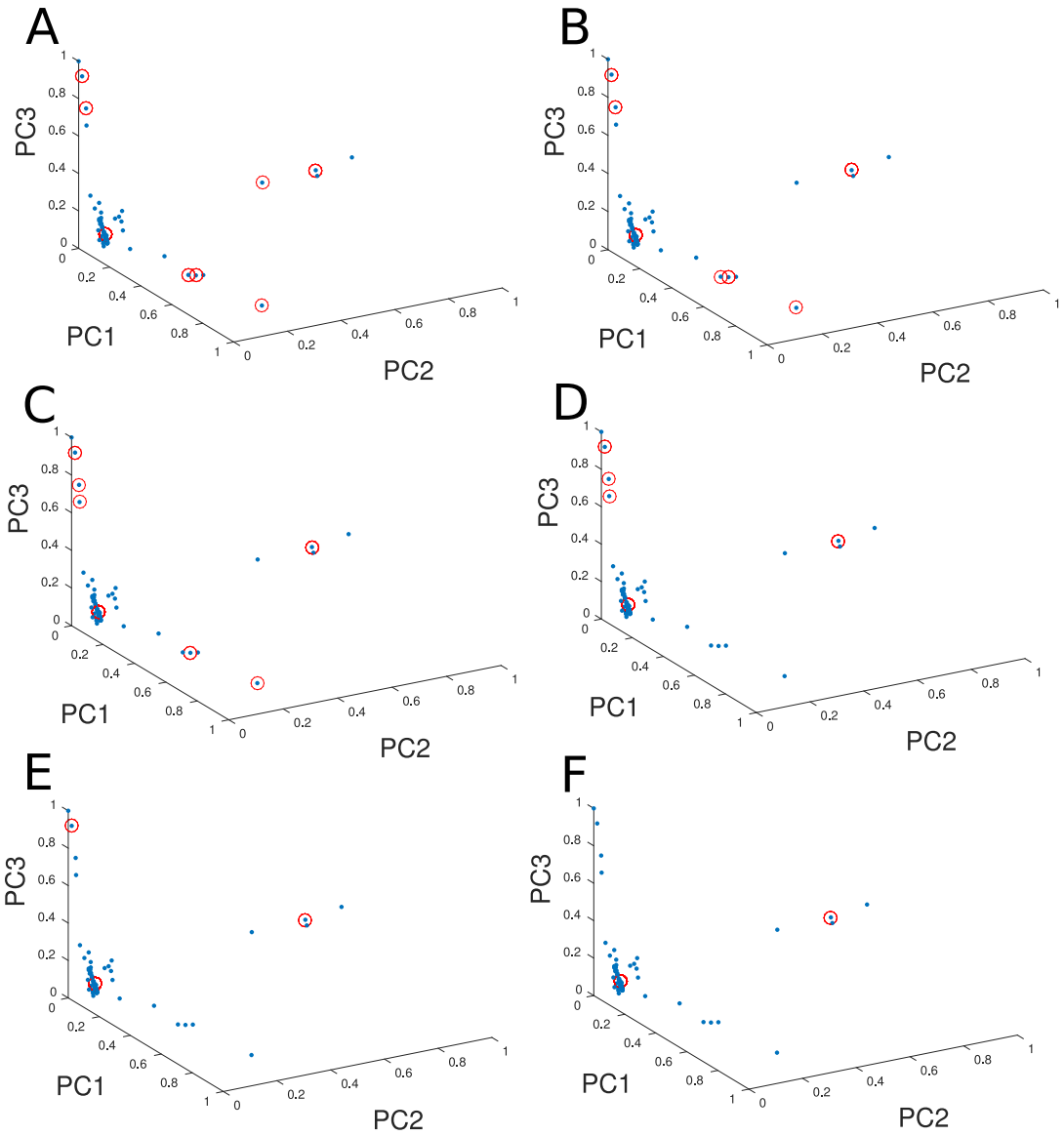

Figure 9: Comparing clustering results of aCGH data projected in the space consisting of the first 3 PC by standard medoidshift with kernel bandwidth factors of 0.5 (A), 1.0 (B), 1.5 (C), 2.0 (D), 2.5 (E), and 3.0 (F). Cluster centers are marked with red circles. The clustering is sensitive to the choice of bandwidth factor, and it fails to give desirable clustering in all cases.

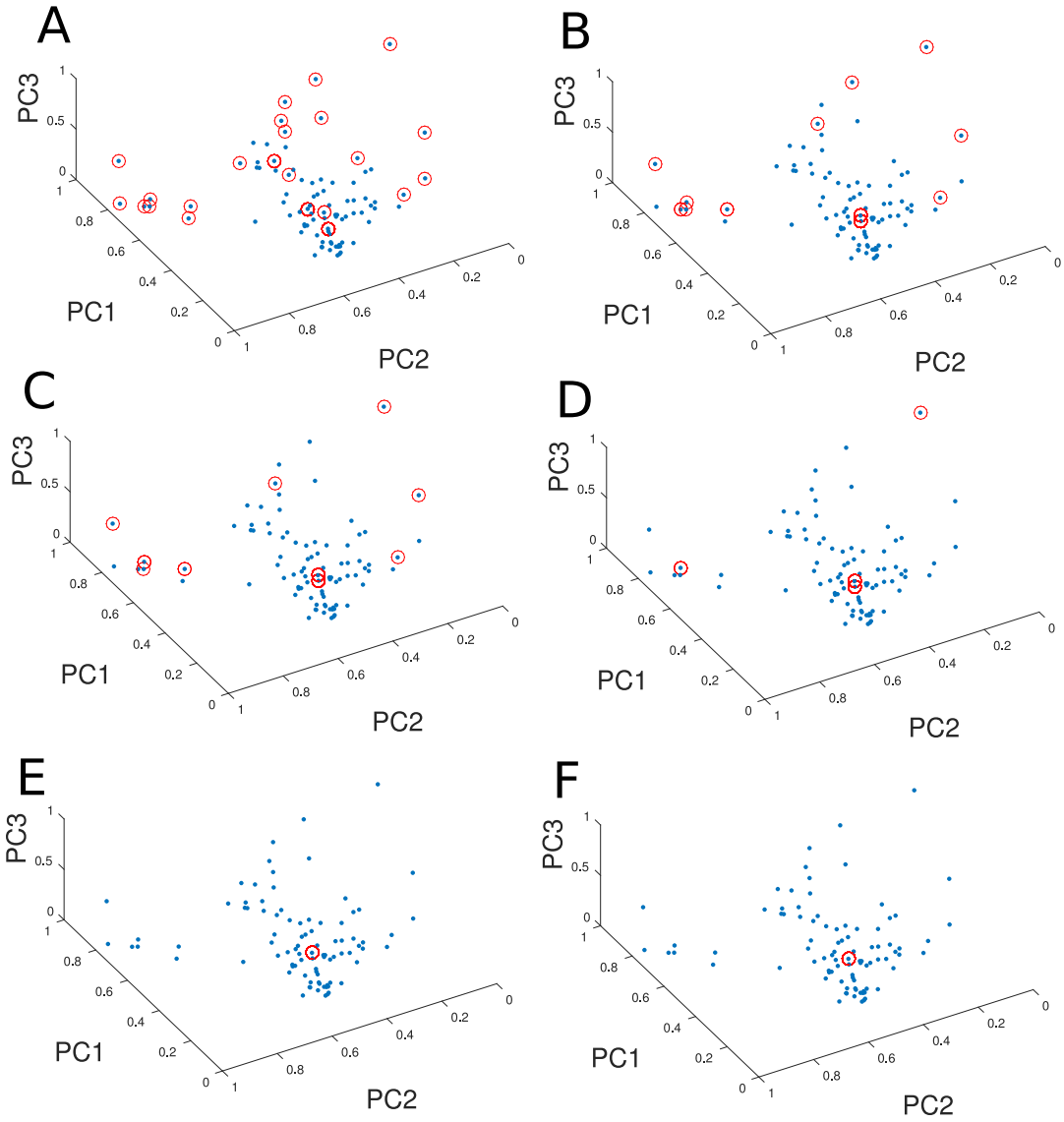

Figure 10: Comparing clustering results of cDNA data projected in the space consisting of the first 3 PC by standard medoidshift with kernel bandwidth factors of 0.5 (A), 1.0 (B), 1.5 (C), 2.0 (D), 2.5 (E), and 3.0 (F). Cluster centers are marked with red circles. The clustering is sensitive to the choice of bandwidth factor, and it fails to give desirable clustering in all cases.