# Model order selection for bio-molecular data clustering

Alberto Bertoni[1], Giorgio Valentini[*1]

[1]DSI, Dipartimento di Scienze dell' Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy

Email: Alberto Bertoni - bertoni@dsi.unimi.it; Giorgio Valentini[*]- valentini@dsi.unimi.it;

[*]Corresponding author

## Abstract

**Background:** Cluster analysis has been widely applied for investigating structure in bio-molecular data. A drawback of most clustering algorithms is that they cannot automatically detect the "natural" number of clusters underlying the data, and in many cases we have no enough "a priori" biological knowledge to evaluate both the number of clusters as well as their validity. Recently several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters, but despite their successful application to the analysis of complex bio-molecular data, the assessment of the statistical significance of the discovered clustering solutions and the detection of multiple structures simultaneously present in high-dimensional bio-molecular data are still major problems.

**Results:** We propose a stability method based on randomized maps that exploits the high-dimensionality and relatively low cardinality that characterize bio-molecular data, by selecting subsets of randomized linear combinations of the input variables, and by using stability indices based on the overall distribution of similarity measures between multiple pairs of clusterings performed on the randomly projected data. A $\chi^2$-based statistical test is proposed to assess the significance of the clustering solutions and to detect significant and if possible multi-level structures simultaneously present in the data (e.g. hierarchical structures).

**Conclusions:** The experimental results show that our model order selection methods are competitive with other state-of-the-art stability based algorithms and are able to detect multiple levels of structure underlying both synthetic and gene expression data.

## Background

Unsupervised clustering algorithms play a crucial role in the exploration and identification of structures underlying complex bio-molecular data, ranging from transcriptomics to proteomics and functional genomics [1–4].

Unfortunately, clustering algorithms may find structure in the data, even when no structure is present instead. Moreover, even if we choose an ap- propriate clustering algorithm for the given data, we need to assess the reliability of the discovered clusters, and to solve the model order selection problem, that is the proper selection of the "natural" number of clusters underlying the data [5, 6]. From a machine learning standpoint, this is an intrinsically "ill-posed" problem, since in unsupervised learning we lack an external objective criterion, that is we have not an equivalent of a priori known class label as in

supervised learning, and hence the evaluation of the reliability of the discovered classes becomes elusive and difficult. From a biological standpoint, in many cases we have no sufficient biological knowledge to "a priori" evaluate both the number of clusters (e.g. the number of biologically distinct tumor classes), as well as the validity of the discovered clusters (e.g. the reliability of new discovered tumor classes) [7].

To deal with these problems, several methods for assessing the validity of the discovered clusters and to test the existence of biologically meaningful clusters have been proposed (see [8] for a review).

Recently, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in complex bio-molecular data [9–11]. In this conceptual framework multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations.

Different procedures have been introduced to randomly perturb the data, ranging from bootstrapping techniques [9, 12, 13], to noise injection into the data [14] or random projections into lower dimensional subspaces [15, 16].

In particular, Smolkin and Gosh [17] applied an unsupervised version of the random subspace method [18] to estimate the stability of clustering solutions. By this approach, subsets of features are randomly selected multiple times, and clusterings obtained on the corresponding projected subspaces are compared with the clustering obtained in the original space to assess its stability. Even if this approach gives useful information about the reliability of high-dimensional clusterings, we showed that random subspace projections may induce large distortions in gene expression data, thus obscuring their real structure [15]. Moreover, a major problem with data perturbations obtained through random projections from a higher to a lower dimensional space is the choice of the dimension of the projected subspace.

In this paper we extend the Smolkin and Gosh approach to more general randomized maps from higher to lower-dimensional subspaces, in order to reduce the distortion induced by random projections. Moreover, we introduce a principled method based on the Johnson and Lindenstrauss lemma [19] to properly choose the dimension of the projected subspace. Our proposed stability indices are related to those proposed by Ben-Hur et al. [13]: their sta-

bility measures are obtained from the distribution of similarity measures across multiple pairs of clustered data perturbed through resampling techniques. In this work we propose stability indices that depend on the distribution of the similarity measures between pairs of clusterings, but data perturbation is realized through random projections to lower dimensional subspaces, in order to exploit the high-dimensionality of bio-molecular data.

Another major problem related to stability-based methods is to estimate the statistical significance of the structures discovered by clustering algorithms. To face this problem we propose a $\chi^2$-based statistical test that may be applied to any stability method based on the distribution of similarity measures between pairs of clusterings. We experimentally show that by this approach we may discover multiple structures simultaneously present in the data (e.g. hierarchical structures), associating a p-value to the clusterings selected by a given stability-based method for model order selection.

## Methods

In this section we present our approach to stability-based model order selection, considering randomized maps with bounded distortion to perturb the data, stability indices based on the distribution of the clustering similarity measures, and finally we present our $\chi^2$-based test for assessing the significance of the clustering solutions.

### Data perturbations using randomized maps with bounded distortions

A major requirement for clustering algorithms is the reproducibility of their solutions with other data sets drawn from the same source; this is particularly true with bio-molecular data, where the robustness of the solutions is of paramount importance in bio-medical applications. From this standpoint the reliability of a clustering solution is tied to its stability: we may consider reliable a cluster if it is stable, that is if it is maintained across multiple data sets drawn from the same source. In real cases, however, we may dispose only of limited data, and hence we need to introduce multiple "small" perturbations into the original data to simulate multiple "similar" samples from the same underlying unknown distribution. By applying appropriate indices based on similarity measures be-

tween clusterings we can then estimate the stability and hence the reliability of the clustering solutions.

We propose to perturb the original data using random projections $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ from high $d$-dimensional spaces to lower $d'$-dimensional subspaces. A related approach is presented in [17], where the authors proposed to perturb the data randomly choosing a subset of the original features (*random subspace* projection [18]); the authors did not propose any principled method to choose the dimension of the projected subspace, but a key problem consists in finding a $d'$ such that for every pair of data $p, q \in \mathbb{R}^d$, the distances between the projections $\mu(p)$ and $\mu(q)$ are approximately preserved with high probability. A natural measure of the approximation is the distortion $dist_\mu$:

$$dist_\mu(p, q) = \frac{||\mu(p) - \mu(q)||_2}{||p - q||_2} \tag{1}$$

If $dist_\mu(p, q) = 1$, the distances are preserved; if $1 - \epsilon \leq dist_\mu(p, q) \leq 1 + \epsilon$, we say that an $\epsilon$-*distortion* level is introduced.

In [15] we experimentally showed that random subspace projections used in [17] may introduce large distortions into gene expression data, thus introducing bias into stability indices based on this kind of random projections. For these reasons we propose to apply randomized maps with guaranteed low distortions, according to the *Johnson-Lindenstrauss (JL) lemma* [19], that we restate in the following way:

Given a $d$-dimensional data set $D = \{p_1, p_2, \ldots, p_n\} \subset \mathbb{R}^d$ and a distortion level $\epsilon$, randomly choosing a $d'$-dimensional subspace $S \subset \mathbb{R}^d$, with $d' = c \, \log n / \epsilon^2$, where $c$ is a suitable constant, with high probability (say $\geq 0.95$) the random projection $\mu : \mathbb{R}^d \rightarrow S$ verifies $1 - \epsilon \leq dist_\mu(p_i, p_j) \leq 1 + \epsilon$ for all $p_i \neq p_j$.

In practice, using randomized maps that obey the JL lemma, we may perturb the data introducing only bounded distortions, approximately preserving the metric structure of the original data [15]. Note that the dimension of the projected subspace depends only on the cardinality of the original data and the desired $\epsilon$-distortion, and not from the dimension $d$ of the original space.

The embedding exhibited in [19] consists in projections from $\mathbb{R}^d$ in random $d'$-dimensional subspaces. Similar results may be obtained by using simpler maps [20, 21], represented through random $d' \times d$ matrices $R = 1/\sqrt{d'}(r_{ij})$, where $r_{ij}$ are random

variables such that:

$$E[r_{ij}] = 0, \qquad Var[r_{ij}] = 1$$

Strictly speaking, these are not projections, but for sake of simplicity, we call random projections even this kind of embeddings. Examples of random projections are the following:

1. *Bernoulli* random projections: represented by $d' \times d$ matrices $R = 1/\sqrt{d'}(r_{ij})$, where $r_{ij}$ are uniformly chosen in $\{-1, 1\}$, such that $Prob(r_{ij} = 1) = Prob(r_{ij} = -1) = 1/2$ (that is the $r_{ij}$ are *Bernoulli* random variables). In this case the *JL lemma* holds with $c \simeq 4$.

2. *Achlioptas* random projections [20]: represented by $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where $r_{ij}$ are chosen in $\{-\sqrt{3}, 0, \sqrt{3}\}$, such that $Prob(r_{ij} = 0) = 2/3$, $Prob(r_{ij} = \sqrt{3}) = Prob(r_{ij} = -\sqrt{3}) = 1/6$. In this case also we have $E[r_{ij}] = 0$ and $Var[r_{ij}] = 1$ and the *JL lemma* holds.

3. *Normal* random projections [21, 22]: this *JL lemma* compliant randomized map is represented by a $d' \times d$ matrix $R = 1/\sqrt{d'}(r_{ij})$, where $r_{ij}$ are distributed according to a gaussian with 0 mean and unit variance.

4. *Random Subspace (RS)* [17, 18]: represented by $d' \times d$ matrices $R = \sqrt{d/d'}(r_{ij})$, where $r_{ij}$ are uniformly chosen with entries in $\{0, 1\}$, and with exactly one 1 per row and at most one 1 per column. Unfortunately, *RS* does not satisfy the *JL lemma*.

Using the above randomized maps (with the exception of *RS* projections), the *JL lemma* guarantees that, with high probability, the "compressed" examples of the data set represented by the matrix $D_R = RD$ have approximately the same distance (up to a $\epsilon$-distortion level) of the corresponding examples in the original space, represented by the columns of the matrix $D$, as long as $d' \geq c \, \log n / \epsilon^2$.

We propose a general *MOSRAM* (Model Order Selection by RAndomized Maps) algorithmic scheme, that implements the above ideas about random projection with bounded distortions to generate a set of similarity indices of clusterings obtained by pairs of randomly projected data. The main difference with respect to the method proposed in [13] is that by our approach we perturb the original data using a randomized mapping $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$:

3

**MOSRAM algorithm**:

```
Input:
    D : a dataset;
    k_max: max number of clusters;
    m : number of similarity measures;
    μ : a randomized map;
    C: a clustering algorithm;
    sim : a clustering similarity measure.
Output:
    M(i,k): a bidimensional list of similarity mea-
sures for each k (1 ≤ i ≤ m, 2 ≤ k ≤ k_max)
begin
    for k := 2 to k_max
        for i := 1 to m
        begin
            proj_a := μ(D)
            proj_b := μ(D)
            C_a := C(proj_a, k)
            C_b := C(proj_b, k)
            M(i,k) := sim(C_a, C_b)
        end
end.
```

The algorithm computes $m$ similarity measures for each desired number of clusters $k$. Every measure is achieved by applying $sim$ to the clustering $C_a$ and $C_b$, outputs of the clustering algorithm $C$, having as input $k$ and the projected data $proj_a$ and $proj_b$. These data are generated through randomized maps $\mu$, with a desired distortion level $\epsilon$. It is worth noting that we make no assumptions about the shape of the clusters, and in principle any clustering algorithm $C$, randomized map $\mu$, and clustering similarity measure $sim$ may be used (e.g. the Jaccard or the Fowlkes and Mallows coefficients [23]).

**Stability indices based on the distribution of the similarity measures**

Using the similarity measures obtained through the *MOSRAM* algorithm, we may compute stability indices to assess the reliability of clustering solutions.

More precisely, let $C$ be a clustering algorithm, $\rho$ a random perturbation procedure (e.g. a resampling or a random projection) and $sim$ a suitable similarity measure between two clusterings (e.g. the Fowlkes and Mallows similarity).

We may define the random variable $S_k$, $0 \leq S_k \leq 1$:

$$S_k = sim\left(C(D_1, k), C(D_2, k)\right) \qquad (2)$$

where $D_1 = \rho^{(1)}(D)$ and $D_2 = \rho^{(2)}(D)$ are obtained

through random and independent perturbations of the data set $D$; the intuitive idea is that if $S_k$ is concentrated close to 1, the corresponding clustering is stable with respect to a given controlled perturbation and hence it is reliable.

Let $f_k(s)$ be the density function of $S_k$ and $F_k(s)$ its cumulative distribution function. A parameter of concentration implicitly used in [13] is the integral $g(k)$ of the cumulative distribution:

$$g(k) = \int_0^1 F_k(s)ds \qquad (3)$$

Note that if $S_k$ is centered in 1, $g(k)$ is close to 0, and hence it can be used as a measure of stability.

Moreover, the following facts show that $g(k)$ is strictly related to both the expectation $E[S_k]$ and the variance $Var[S_k]$ of the random variable $S_k$:

**Fact 1**: $E[S_k] = 1 - g(k)$.
Indeed, integrating by parts:

$$E[S_k] = \int_0^1 s f_k(s)ds = \int_0^1 s F_k'(s)ds =$$
$$= 1 - \int_0^1 F_k(s)ds = 1 - g(k) \qquad (4)$$

**Fact 2**: $Var[S_k] \leq g(k)(1 - g(k))$.
Since $0 \leq S_k \leq 1$ it follows $S_k^2 \leq S_k$; therefore, using Fact 1:

$$Var[S_k] = E[S_k^2] - E[S_k]^2 \leq E[S_k] - E[S_k]^2 =$$
$$= g(k)(1 - g(k)) \quad (5)$$

In conclusion, $g(k) \simeq 0$ then $E[S_k] \simeq 1$ and $Var[S_k] = 0$, i.e. $S_k$ is centered close to 1. As a consequence, $E[S_k]$ can be used as an index of the reliability of the $k$-clustering: if $E[S_k] \simeq 1$, the clustering is stable, if $E[S_k] \ll 1$ the clustering can be considered less reliable.

We can estimate $E[S_k]$ by means of $m$ similarity measures $M(i,k)$ ($1 \leq i \leq m$) computed by the *MOSRAM* algorithm. In fact $E[S_k]$ may be estimated by the empirical mean $\xi_k$:

$$\xi_k = \sum_{i=1}^m \frac{M(i,k)}{m} \qquad (6)$$

**A $\chi^2$-based test for the assessment of the significance of the solutions**

In this section we propose a method for automatically finding the "optimal" number of clusters and to detect significant and possibly multi-level structures simultaneously present in the data. First of all,

let us consider the vector $(\xi_2, \xi_3, \ldots, \xi_{H+1})$ (eq. 6) computed by using the output of the *MOSRAM* algorithm. We may perform a sorting of this vector:

$$(\xi_2, \xi_3, \ldots, \xi_{H+1}) \overset{sort}{\rightarrow} (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)}) \quad (7)$$

where $p$ is the permutation index such that $\xi_{p(1)} \geq \xi_{p(2)} \geq \ldots \geq \xi_{p(H)}$. Roughly speaking, this ordering represents the "most reliable" $p(1)$-clustering down to the least reliable $p(H)$-clustering; exploiting this we would establish which are the significant clusterings (if any) discovered in the data.

To this end, for each $k \in \mathcal{K} = \{2, 3, \ldots, H+1\}$, let us consider the random variable $S_k$ defined in eq. 2, whose expectation is our proposed stability index. For all $k$ and for a fixed threshold $t^o \in [0, 1]$ consider the Bernoulli random variable $B_k = I(S_k > t^o)$, where $I$ is the indicator function: $I(P) = 1$ if $P$ is True, $I(P) = 0$ if $P$ is False. The sum $X_k = \sum_{j=1}^{m} B_k^j$ of i.i.d. copies of $B_k$ is distributed according to a binomial distribution with parameters $m$ and $\theta_k = Prob(I(S_k > t^o))$.

If we hypothesize that all the binomial populations are independently drawn from the same distribution (i.e. $\theta_k = \theta$, for all $k \in \mathcal{K}$), for sufficiently large values of $m$ the random variables $\frac{X_k - m\theta_k}{\sqrt{m\theta_k(1-\theta_k)}}$ are independent and approximately normally distributed. Consider now the random variable:

$$\sum_{k \in \mathcal{K}} \frac{(X_k - m\hat{\theta})^2}{m\theta(1-\theta)} \quad \text{with} \quad \hat{\theta} = \frac{\sum_{k \in \mathcal{K}} X_k}{|\mathcal{K}| \cdot m} \quad (8)$$

This variable is known to be distributed as a $\chi^2$ with $|\mathcal{K}| - 1$ degrees of freedom, informally because the constraint $\hat{\theta}$ between the random variables $X_k, k \in \mathcal{K}$ introduces a dependence between them, thus leading to a loss of one degree of freedom. By estimating the variance $m\theta(1-\theta)$ with the statistic $m\hat{\theta}(1-\hat{\theta})$, we conclude that the following statistic

$$Y = \sum_{k \in \mathcal{K}} \frac{(X_k - m\hat{\theta})^2}{m\hat{\theta}(1-\hat{\theta})} \sim \chi^2_{|\mathcal{K}|-1} \quad (9)$$

is approximately distributed according to $\chi^2_{|\mathcal{K}|-1}$ (see, e.g. [24] chapter 12, or [25] chapter 30 for more details).

A realization $x_k$ of the random variable $X_k$ (and the corresponding realization $y$ of $Y$) can be computed by using the output of the *MOSRAM* algorithm:

$$x_k = \sum_{i=1}^{m} I(M(i,k) > t^o) \quad (10)$$

Using $y$, we can test the following alternative hypotheses:

- Ho: all the $\theta_k$ are equal to $\theta$ (the considered set of k-clusterings are equally reliable)

- Ha: the $\theta_k$ are not all equal between them (the considered set of k-clusterings are not equally reliable)

If $y \geq \chi^2_{\alpha, |\mathcal{K}|-1}$ we may reject the null hypothesis at $\alpha$ significance level, that is we may conclude that with probability $1 - \alpha$ the considered proportions are different, and hence that at least one k-clustering significantly differs from the others.

Using the above statistical test, we propose an iterative procedure to detect the significant number(s) of clusterings:

1. Consider the ordered vector $\xi = (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)})$

2. Repeat the $\chi^2$-based test until no significant difference is detected or the only remaining clustering is $p(1)$ (the top-ranked one). At each iteration, if a significant difference is detected, remove the bottom-ranked clustering from $\xi$

The output of the proposed procedure is the set of the remaining (top sorted) k-clusterings that correspond to the set of the estimate stable number of clusters (at $\alpha$ significance level). Equivalently, following the sorting of $\xi$, we may compute the p-value (probability of committing an error if we reject the null hypothesis) for all the ordered groups of clusterings from the $p(1)...p(H)$ to the the $p(1), p(2)$ group, each time removing the bottom ranked clustering from the $\xi$ vector. Note that if the set of the remaining top-ranked clusterings contains more than one clustering, we may find multiple structures simultaneously present in the data (at $\alpha$ significance level).

## Results and Discussion

We present experiments with synthetic and gene expression data to show the effectiveness of our approach. At first, using synthetic data, we show that our proposed methods can detect not only the "correct" number of clusters, but also multiple structures underlying the data. Then we apply our *MOSRAM*

algorithm to discover the "natural" number of clusters in gene expression data, and we compare the results with other algorithms for model order selection. In our experiments we used the classical *k-means* [26] and *Prediction Around Medoid* (*PAM*) [27] clustering algorithms, and we applied the *Bernoulli, Achlioptas* and *Normal* random projections, but in this section we show only the results obtained with *Bernoulli* projections, since with the other randomized maps we achieved the same results without any significant difference. In all our experiments we set the threshold $t^o$ (see Section "A $\chi^2$-based test for the assessment of the significance of the solutions" to 0.9. Moreover we applied our proposed $\chi^2$-based procedure to individuate sets of significant $k$-clusterings into the data. The methods and algorithms described in this paper have been implemented in the *mosclust R* package, publicly available at [28].

## Detection of multiple levels of structure in synthetic data

To show the ability of our method to discover multiple structures simultaneously present in the data, we propose an experiment with a 1000-dimensional synthetic multivariate gaussian data set (*sample1*) with relatively low cardinality (60 examples), characterized by a two-level hierarchical structure, highlighted by the projection of the data into the two main principal components (Figure 1): indeed a two-level structure, with respectively 2 and 6 clusters is self-evident in the data.

Two clusterings (using the Prediction Around Medoid algorithm) are detected at $10^{-4}$ significance level by applying our *MOSRAM* algorithm and the proposed $\chi^2$-based statistical procedure. In particular we performed 100 pairs of *Bernoulli* projections with a distortion bounded to 1.2 ($\epsilon = 0.2$), yielding to random projections from 1000 to 479-dimensional subspaces. Indeed Table 1 reports the sorted means of the stability measures together with their variance and the corresponding p-values computed according to the proposed $\chi^2$-based statistical test, showing that 2 and 6-clusterings are the best scored, as well as the most significant $k$-clusterings discovered in the data. This situation is depicted in Figure 2, where the histograms of the similarity measures for $k = 2$ and $k = 6$ clusters are tightly concentrated near 1, showing that these clusterings are very stable, while for other values of $k$ the similarity measures are spread across multiple values. Note that the p-

values of Table 1 (as well as the p-values of Table 2 and 3) refer to the probability of committing an error if we reject the null hypothesis. The clusterings with $p \geq \alpha$ are considered equally reliable (in this case the null hypothesis cannot be rejected), while the clusterings for which $p < \alpha$ are considered less reliable (at $\alpha$ significance level).

## Experiments with DNA microarray data

To show the effectiveness of our methods with gene expression data we applied *MOSRAM* and the proposed statistical test to *Leukemia* [29] and *Lymphoma* [1] samples. These data sets have been analyzed with other model order selection algorithms previously proposed [10, 13, 30–32]: at the end of this section we compare the results obtained with the cited methods with our proposed *MOSRAM* algorithm.

### Leukemia

This well known data set [29] is composed by 72 leukemia samples analyzed with oligonucleotide Affymetrix microarrays. The *Leukemia* data set is composed by a group of 25 acute myeloid leukemia (AML) samples and another group of 47 acute lymphoblastic leukemia (ALL) samples, that can be subdivided into 38 B-Cell and 9 T-Cell subgroups, resulting in a two-level hierarchical structure.

We applied the same pre-processing steps performed by the authors of the *Leukemia* study [29], obtaining 3571 genes from the original 7129 gene expression values. We further selected the 100 genes with the highest variance across samples, since low variance genes are unlikely to be informative for the purpose of clustering [10, 31]. We analyzed both the 3571-dimensional data and the data restricted to the 100 genes with highest variance, using respectively *Bernoulli* projections with $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ and projections to 80-dimensional subspaces. In both cases the k-means clustering algorithm has been applied.

Figure 3 summarizes the results using gene expression levels of the genes with highest variance. Table 2 reports the sorted means of the stability measures together with their variance and the corresponding p-values computed according to the proposed $\chi^2$-based statistical test. By these results 2 and 3 clusters are correctly predicted at $\alpha = 10^{-5}$ significance level. Indeed the empirical means of

the stability measures (eq. 6) for 2 and 3 clusters are quite similar and the corresponding lines (black and red) of the empirical cumulative distributions (ecdfs) cross several times, while the other ecdfs are clearly apart from them (Figure 3). The same results are approximately obtained also using the 3571-dimensional data with random projections to 428-dimensional subspaces ($\epsilon = 0.2$), but 2 and 3 clusters are predicted at $\alpha = 10^{-13}$ significance level. Similar results are also achieved with $\epsilon = 0.1$ and $\epsilon = 0.3$, while with $\epsilon = 0.4$ the results are less reliable due to the relatively large distortion induced (data not shown). Also using the PAM [27] and hierarchical clustering algorithms with the Ward method [33] we obtained a two-level structure with 2 and 3 clusters at $\alpha = 10^{-5}$ significance level.

*Lymphoma*

Three different lymphoid malignancies are represented in the *Lymphoma* gene expression data set [1]: Diffuse Large B-Cell Lymphoma (DLBCL), Follicular Lymphoma (FL) and Chronic Lymphocytic Leukemia (CLL). The gene expression measurements are obtained with a cDNA microarray specialized for genes related to lymphoid diseases, the Lymphochip, which provides expression levels for 4026 genes [34]. The 62 available samples are subdivided in 42 DLBCL, 11 CLL and 9 FL. We performed pre-processing of the data according to [1], replacing missing values with 0 and then normalizing the data to zero mean and unit variance across genes. As a final step, according to [10], we further selected the 200 genes with highest variance across samples, obtaining a resulting data set with 62 samples and 200 genes. As in the previous experiment, we processed both the high-dimensional original data and the data with the reduced set of high-variance genes, using respectively *Bernoulli* projections with $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ and projections to 160-dimensional subspaces. The k-means clustering algorithm has been applied.

The results with highest variance genes are summarized in Figure 4 and Table 3. The statistical test identifies as significant only the 2-clustering. Indeed, looking at the ecdf of the stability index values (left), the 2-clustering (black) is clearly separated from the others. The 3 (red) and 4-clustering (green) graphs, are quite distinct from the others, as shown also by the corresponding empirical mean of the stability index values (Figure 4), but they are

also clearly separated from the 2-clustering curve. Accordingly, our proposed $\chi^2$-based test found a significant difference between the 2-clustering and all the others. Similar results are obtained also with hierarchical clustering and PAM algorithms. Using all the 4026 genes and *Bernoulli* random projections to 413-dimensional ($\epsilon = 0.2$) subspaces with the Ward's hierarchical clustering algorithm our method finds as significant the 3-clustering as well as the 2-clustering. In this case also similar results are obtained with $\epsilon = 0.1$ and $\epsilon = 0.3$. It is worth noting that the subdivision of *Lymphoma* samples in 3 classes (DLBCL, CLL and FL) have been defined on histopathological and morphological basis and it has been shown that this classification does not correspond to the bio-molecular characteristics and to clinical outcome classes of non Hodgkin lymphomas. In particular studies based on the gene expression signatures of the DLBCL patients [1] and on their supervised analysis [35], showed the existence of two subclasses of DLBCLs. Moreover Shipp et al. [36] highlighted that FL patients frequently evolve over time and acquire the clinical features of DLBCLs, and Lange et al. [10] found that a 3-clustering solution groups together FL, CLL and a subgroup of DLBCLs, while another subgroup of DLBCLs sets up another cluster, even if the overall stability of the clustering is lower with respect to the 2-clustering solution. The relationships between FL and subgroups of DLBCL patients are confirmed also by recent studies on the individual stability of the clusters in DLBCL and FL patients [15]. These considerations show also that the stability analysis of patients clusters in DNA microarray analysis are only the first step to discover significant subclasses of pathologies at bio-molecular level, while another necessary step is represented by the bio-medical validation.

*Comparison with other methods*

We compared the results obtained by the *MOSRAM* algorithm with other model order selection methods using the *Leukemia* and *Lymphoma* data sets analyzed in the previous section. In particular we focused our comparison with other state-of-the-art stability-based methods proposed in the literature.

The *Model Explorer* algorithm adopts subsampling techniques to perturb the data (data are randomly drawn without replacement) and applies stability measures based on the empirical distribution of the stability measures [13]. This approach is quite

similar to ours but we applied random projections to perturb the data and a statistical test to identify significant numbers of clusters, instead of simply qualitatively looking at the distributions of the stability indices. The *Figure of Merit* measure is based on a resampling approach too, but the stability of the solutions is assessed directly comparing the solution obtained on the full sample with that obtained on the subsamples [32]. We considered also stability-based methods that apply supervised algorithms to assess the quality of the discovered clusterings instead of comparing pairs of perturbed clusterings [10, 31]: the main differences between these last approaches are the choice of the supervised predictor and other parameters (no guidance is given in [31], while in [10] a more structured approach is proposed). Finally we considered also a non-stability-based method, the *Gap statistic*, that applies an estimates of the gap between the total sum within-class dissimilarities and a null reference distribution (the uniform distribution on the smallest hyper-rectangle that contains all the data) to assess the "optimal" number of clusters in the data.

Table 4 shows the number of clusters selected by the different methods, as well as their "true" number. The "true" number is estimated according to the a priori biological knowledge about the data [1, 29] (see Section *Experiments with DNA microarray data*). The best results achieved with the two gene expression data sets are highlighted with a bounding box. The *MOSRAM* algorithm achieves results competitive with the other state-of-the-art model order selection methods. Indeed *MOSRAM* correctly predicts the "true" number of clusters with the *Leukemia* data set and partially with the *Lymphoma* data set. Note that the 2-clustering prediction with *Lymphoma* may be considered reliable, as outlined in the corresponding experimental section.

These results show that our proposed methods based on randomized maps are well-suited to the characteristics of DNA microarray data: indeed the low cardinality of the examples, the very large number of features (genes) involved in microarray chips, the redundancy of information stored in the spots of microarrays are all characteristics in favour of our approach. On the contrary using bootstrapping techniques to obtain smaller samples from just small samples of patients should induce more randomness in the estimate of cluster stability. A resampling based approach appears to be better suited to evaluate the cluster stability of genes, since significantly

larger samples are available in this case [12]. The alternative based on noise injection into the data to obtain multiple instance of perturbed data poses difficult statistical problems for evaluating what kind and which magnitude of noise should be added to the data [17].

All the perturbation-based methods need to properly select a parameter to control the amount of perturbation of the data: resampled-based methods need to select the "optimal" fraction of the data to be subsampled; noise-injection-based methods needs to choice the amount of noise to be introduced; random subspace and random projections-based methods needs to select the proper dimension of the projected data. Anyway, our approach provides a theoretically motivated method to automatically find an "optimal" value for the perturbation parameter, and in our experiments we observed that values of $\epsilon \leq 0.2$ led to reliable results. Moreover our proposed approach provides also a statistical test that may be applied also with other stability-based methods to assess the significance of the discovered solutions.

Despite of the convincing experimental results obtained with stability-based methods there are some drawbacks and open problems associated with these techniques. Indeed, as shown by [8], a given clustering may converge to a suboptimal solution owing to the shape of the data manifold and not to the real structure of the data, thus introducing bias in the stability indices. Moreover in [37] it has been shown that stability based methods based on resampling techniques, when cost-based clustering algorithms are used, may fail to detect the correct number of clusters, if the data are not symmetric. However it is unclear if these results may be extended to other stability-based methods (e.g. to our proposed methods based on random projections) or to other more general classes of clustering algorithms.

## Conclusions

We proposed a stability-based method, based on random projections, for assessing the validity of clusterings discovered in high-dimensional post-genomic data. The reliability of the discovered k-clusterings may be estimated exploiting the distribution of the clustering pairwise similarities, and a $\chi^2$-based statistical test tailored to unsupervised model order selection. In the theoretical framework of randomized maps that satisfy the *JL lemma*, a principled ap-

proach to select the dimension of the projected data, and to approximately preserve the structure of the original data is given, thus yielding to the design of reliable stability indices for model order selection in bio-molecular data clusterings.

The $\chi^2$-based statistical test may be applied to any stability method that make use of the distribution of the similarity measures between pairs of clusterings.

Our experimental results with synthetic data and real gene expression data show that our proposed method is able to find significant structures, comprising multiple structures simultaneously present into bio-molecular data.

As an outgoing development, considering that the $\chi^2$-based test assumes that the random variables representing distributions for different number of clusters are normally distributed, we are developing a new distribution-independent approach based on the Bernstein inequality to assess the significance of the discovered k-clusterings.

## Authors contributions

The authors equally contributed to this paper.

## Acknowledgments

## References

1. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**:503–511.

2. Hoehe M, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd K, Berrettini W, Church G: **Sequence variability and candidate gene analyisis in complex disease: association of mu opioid receptor gene variation with substance dependence**. *Hum. Mol. Gen.* 2000, **9**:2895–2908.

3. Kaplan N, Friedlich M, Fromer M, Linial M: **A functional hierarchical organization of the protein sequence space**. *BMC Bioinformatics* 2004, **5**.

4. Bilu Y, Linial M: **The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classification**. *Journal of Computational Biology* 2002, **9**:193–210.

5. Bolshakova N, Azuaje F, Cunningham P: **An integrated tool for microarray data clustering and cluster validity assessment**. *Bioinformatics* 2005, **21**(4):451–455.

6. Datta S, S D: **Comparison and validation of statistical clustering techniques for microarray gene expression data**. *Bioinformatics* 2003, **19**:459–466.

7. Alizadeh A, Ross D, Perou C, van de Rijn M: **Towards a novel classification of human malignancies based on gene expression**. *J. Pathol.* 2001, **195**:41–52.

8. Handl J, Knowles J, Kell D: **Computational cluster validation in post-genomic data analysis**. *Bioinformatics* 2005, **21**(15):3201–3215.

9. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning* 2003, **52**:91–118.

10. Lange T, Roth V, Braun M, Buhmann J: **Stability-based Validation of Clustering Solutions**. *Neural Computation* 2004, **16**:1299–1323.

11. Garge N, Page G, Sprague A, Gorman B, Allison D: **Reproducible Clusters from Microarray Research: Whither?** *BMC Bioinformatics* 2005, **6(Suppl2):S10**.

12. Kerr M, Curchill G: **Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments**. *PNAS* 2001, **98**:8961–8965.

13. Ben-Hur A, Ellisseeff A, Guyon I: **A stability based method for discovering structure in clustered data**. In *Pacific Symposium on Biocomputing, Volume 7*. Edited by Altman R, Dunker A, Hunter L, Klein T, Lauderdale K, Lihue, Hawaii, USA: World Scientific 2002:6–17.

14. McShane L, Radmacher D, Freidlin B, Yu R, Li M, Simon R: **Method for assessing reproducibility of clustering patterns observed in analyses of microarray data**. *Bioinformatics* 2002, **18**(11):1462–1469.

15. Bertoni A, Valentini G: **Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses**. *Artificial Intelligence in Medicine* 2006, **37**(2):85–109.

16. Valentini G: **Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data**. *Bioinformatics* 2006, **22**(3):369–370.

17. Smolkin M, Gosh D: **Cluster stability scores for microarray data in cancer studies**. *BMC Bioinformatics* 2003, **36**(4).

18. Ho T: **The random subspace method for constructing decision forests**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20**(8):832–844.

19. Johnson W, Lindenstrauss J: **Extensions of Lipshitz mapping into Hilbert space**. In *Conference in modern analysis and probability, Volume 26 of* Contemporary Mathematics, Amer. Math. Soc. 1984:189–206.

20. Achlioptas D: **Database-friendly random projections.** In *Proc. ACM Symp. on the Principles of Database Systems*, Contemporary Mathematics. Edited by Buneman P, New York, NY, USA: ACM Press 2001:274–281.

21. Bingham E, Mannila H: **Random projection in dimensionality reduction: Applications to image and text data**. In *Proc. of KDD 01*, San Francisco, CA, USA: ACM 2001.

22. Fern X, Brodley C: **Random Projections for High Dimesnional Data Clustering: A Cluster Ensemble Approach**. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*. Edited by Fawcett T, Mishra N, Washington D.C., USA: AAAI Press 2003.

23. Jain A, Murty M, Flynn P: **Data Clustering: a Review**. *ACM Computing Surveys* 1999, **31**(3):264–323.

24. Freund J: *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall 1962.

25. Cramer H: *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press 1958.

26. McQueen J: **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the Fifth Berkeley Symposium of Mathematical Statistics and Probability*. Edited by LeCam L, Neyman J, University Of California Press 1967:281–297.

27. Kaufman L, Rousseeuw P: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley 1990.

28. Valentini, G: **Mosclust: a software library for discovering significant structures in bio-molecular data**. *Bioinformatics* 2007, **23**(3):387–389. Mosclust web-site: `http://homes.dsi.unimi.it/∼valenti/SW/mosclust`.

29. Golub T, et al.: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**. *Science* 1999, **286**:531–537.

30. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a dataset via the gap statistic**. *Journal of the Royal Statistical Society B* 2001, **63**(2):411–423.

31. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset**. *Genome Biology* 2002, **3**(7):1–21.

32. Levine E, Domany E: **Resampling method for unsupervised estimation of cluster validity**. *Neural Computation* 2001, **13**(11):2573–2593.

33. Ward J: **Hierarchical grouping to optimize an objective function**. *J. Am. Stat. Assoc.* 1963, **58**:236–244.

34. Alizadeh A, et al.: **The Lymphochip: a specialized cDNA microarray for genomic-scale analysis of gene expression in normal and malignant lymphocytes**. In *Cold Spring Harbor Symp. Quant. Biol.* 2001.

35. Valentini G: **Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles**. *Artificial Intelligence in Medicine* 2002, **26**(3):283–306.

36. Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, Gaasenbeek M, Angelo M, Reich M, Pinkus G, Ray T, Koval M, Last K, Norton A, Lister T, Mesirov J, Neuberg D, Lander E, Aster J, Golub T: **Diffuse large B-cell Lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nature Medicine* 2002, **8**:68–74.

37. Ben-David S, von Luxburg U, Pal D: **A Sober Look at Clustering Stability**. In *19th Annual Conference on Learning Theory, COLT 2006, Volume 4005 of* Lecture Notes in Computer Science, Springer 2006:5–19.
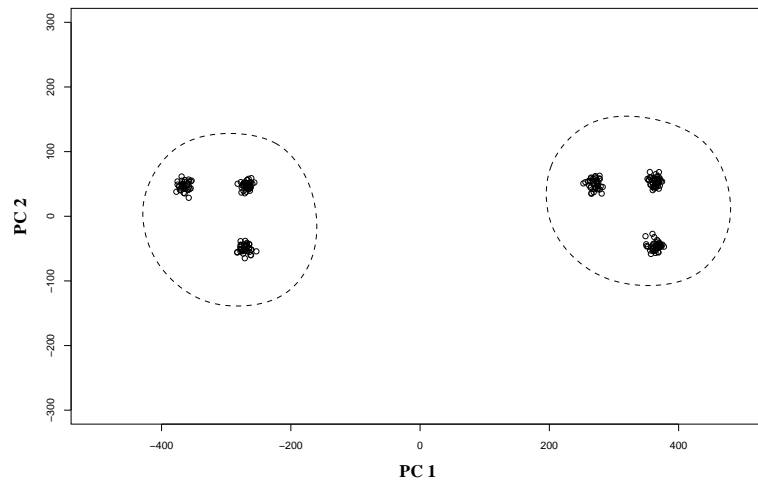
# Figures



Figure 1: A two-level hierarchical structure with 2 and 6 clusters is revealed by principal components analysis (data projected into the two components with highest variance).
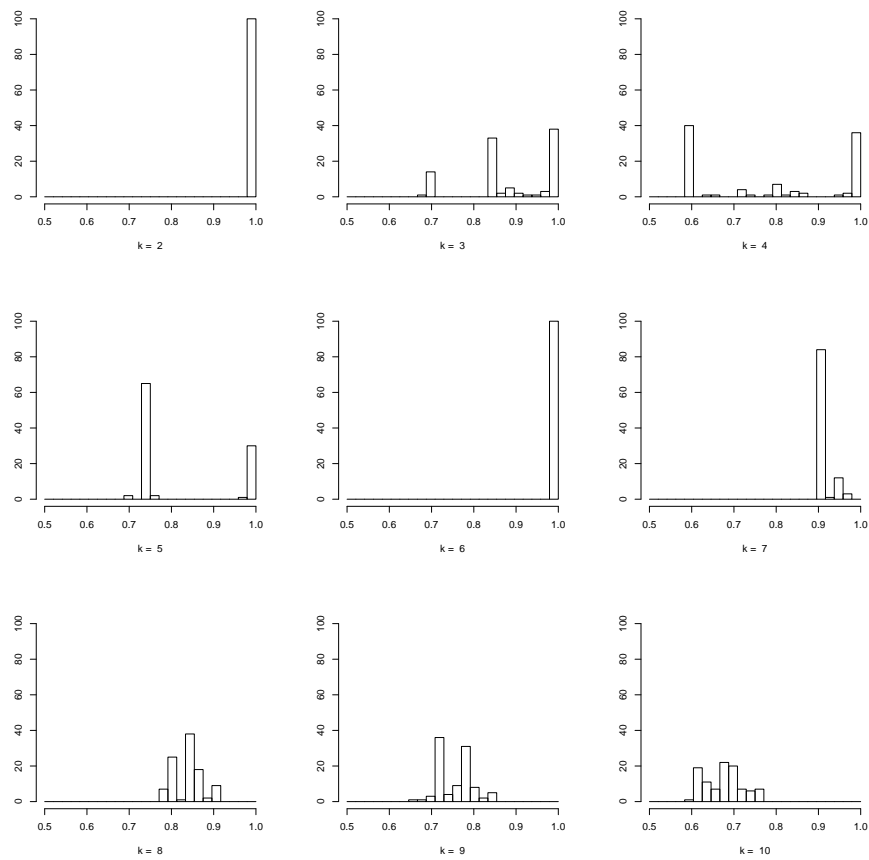
Figure 2: Histograms of the similarity measure distributions for different numbers of clusters.
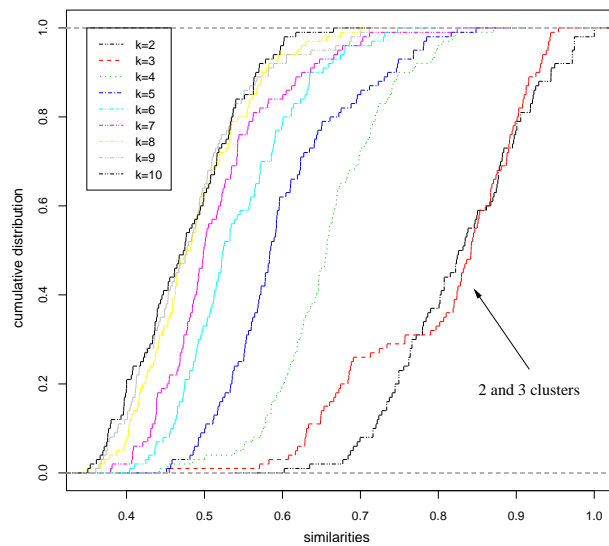


Figure 3: *Leukemia* data set: empirical cumulative distribution functions of the similarity measures for different number of clusters $k$.
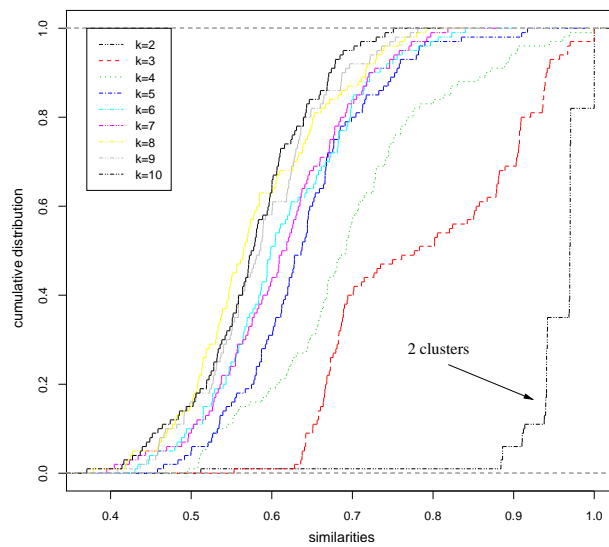
Figure 4: *Lymphoma* data set: empirical cumulative distribution functions of the similarity measures for different number of clusters $k$.

13

## Tables

### Table 1 - Sample1: similarity indices

Similarity indices for the synthetic *sample1* data set for different $k$-clusterings, sorted with respect to their mean values.

| k | mean | variance | p-value |
|---|------|----------|---------|
| 2 | 1.0000 | 0.0000 | 1.0000 |
| 6 | 1.0000 | 0.0000 | 1.0000 |
| 7 | 0.9217 | 0.0016 | 0.0000 |
| 8 | 0.8711 | 0.0033 | 0.0000 |
| 9 | 0.8132 | 0.0042 | 0.0000 |
| 5 | 0.8090 | 0.0104 | 0.0000 |
| 3 | 0.8072 | 0.0157 | 0.0000 |
| 10 | 0.7715 | 0.0056 | 0.0000 |
| 4 | 0.7642 | 0.0158 | 0.0000 |

.

### Table 2 -Leukemia data set

Stability indices for different $k$-clusterings sorted with respect to their mean values.

| k | mean | variance | p-value |
|---|------|----------|---------|
| 2 | 0.8285 | 0.0077 | 1.0000 |
| 3 | 0.8060 | 0.0124 | 0.7328 |
| 4 | 0.6589 | 0.0060 | 2.3279e-06 |
| 5 | 0.6012 | 0.0073 | 9.5199e-11 |
| 6 | 0.5424 | 0.0057 | 6.3282e-15 |
| 7 | 0.5160 | 0.0062 | 0.0000 |
| 8 | 0.4865 | 0.0050 | 0.0000 |
| 9 | 0.4819 | 0.0060 | 0.0000 |
| 10 | 0.4744 | 0.0049 | 0.0000 |

### Table 3 -Lymphoma data set.

Stability indices for different $k$-clusterings sorted with respect to their mean values.

| k | mean | variance | p-value |
|---|------|----------|---------|
| 2 | 0.9566 | 0.0028 | 1.0000 |
| 3 | 0.7900 | 0.0149 | 0.0000 |
| 4 | 0.6963 | 0.0128 | 0.0000 |
| 5 | 0.6387 | 0.0075 | 0.0000 |
| 6 | 0.6135 | 0.0082 | 0.0000 |
| 7 | 0.6129 | 0.0079 | 0.0000 |
| 9 | 0.5864 | 0.0063 | 0.0000 |
| 8 | 0.5792 | 0.0079 | 0.0000 |
| 10 | 0.5744 | 0.0058 | 0.0000 |

**Table 4 - Results comparison**

Comparison between different methods for model order selection in gene expression data analysis

| Methods | Class. risk (Lange et al. 2004) | Gap statistic (Tibshirani et al. 2001) | Clest (Dudoit and Fridlyand 2002) | Figure of Merit (Levine and Domany 2001) | Model Explorer (BenHur et al 2002) | MOSRAM | "True" number k |
|---------|------------|---------------|--------|------------|----------------|--------|-----------------|
| Data set | | | | | | | |
| *Leukemia* (Golub et al. 1999) | k=3 | k=10 | k=3 | k=2,8,19 | k=2 | k=2,3 | k=2,3 |
| *Lymphoma* (Alizadeh et al. 2000) | k=2 | k=4 | k=2 | k=2,9 | k=2 | k=2 | k=2,(3) |