Supplement to "Clustering Gene Expression Data with Repeated Measurements"

Ka Yee Yeung, Mario Medvedovic, Roger E. Bumgarner

Results on the yeast galactose data (without imputing missing values)

In the paper, we reported results on the yeast galactose data after imputing missing values. Here, we show results on the yeast galactose data *without* imputing any missing values, i.e., we ignore the corresponding values when computing the similarity measures. Since the current implementations of the model-based approaches cannot deal with missing data values, their results are not available.

Cluster Accuracy: agreement with the functional categories

Each entry shows the adjusted Rand index of the corresponding algorithm with the functional categories. The maximum adjusted Rand index of each row is shown in bold. The algorithms (rows) are sorted in descending order of the maximum adjusted Rand in each row. DIANA and single-link produce the least accurate clusters. *CAST did not converge.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical	true mean
average-link	dist	0.843	0.858	**0.957**	0.869	0.159
CAST	corr	0.737	0.815	NA*	NA	**0.938**
complete-link	dist	0.695	**0.937**	0.928	0.633	0.438
average-link	corr	0.840	0.790	0.764	0.799	**0.924**
centroid-link	dist	0.861	0.868	**0.869**	0.861	0.850
k-means	dist	0.843	0.860	0.821	NA	**0.861**
centroid-link	corr	0.833	0.772	0.176	0.833	**0.841**
complete-link	corr	0.636	0.764	0.619	0.455	**0.792**
k-means	corr	0.693	0.678	0.568	NA	**0.728**
DIANA	dist	0.043	0.137	**0.162**	NA	-0.156
single-link	dist	0.017	0.038	0.132	**0.160**	**0.160**
single-link	corr	0.005	-0.009	0.005	**0.156**	0.017
DIANA	corr	-0.059	-0.143	**0.046**	NA	-0.173

Cluster Stability: with respect to synthetic re-measured data

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets. The maximum average adjusted Rand index for each row is shown in bold. The rows are sorted in descending order of the maximum average adjusted Rand index. DIANA and single-link produced very unstable clusters.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical
average-link	dist	0.543	**0.983**	0.828	0.680
centroid-link	dist	0.466	**0.934**	0.714	0.466
k-means	dist	0.768	**0.827**	0.754	NA
complete-link	dist	0.622	**0.813**	0.675	0.393
average-link	corr	0.746	**0.780**	0.700	0.765
CAST	corr	0.671	**0.695**	NA*	NA
k-means	corr	**0.677**	0.674	0.463	NA
centroid-link	corr	**0.665**	0.568	0.201	**0.665**
complete-link	corr	0.591	**0.620**	0.509	0.503
single-link	dist	0.348	0.506	0.175	**0.578**
single-link	corr	**0.252**	0.246	0.026	0.141
DIANA	dist	**-0.016**	-0.487	-0.318	NA
DIANA	corr	-0.174	-0.291	**-0.034**	NA

Results on the yeast galactose data (after imputing missing values)

Most of the results in the following tables are shown in the main paper. Here, we added the results from single-link and DIANA. As shown in the following two tables, both single-link and DIANA produced unstable and inaccurate clusters. Comparing the results before and after imputing missing values, it is clear that clusters tend to be more accurate and more stable after the missing value imputation step.

Cluster Accuracy: agreement with the functional categories

Each entry shows the adjusted Rand index of the corresponding algorithm with the functional categories. The maximum adjusted Rand index of each row is shown in bold. The algorithms (rows) are sorted in descending order of the maximum adjusted Rand in each row. DIANA and single-link produce the least accurate clusters. *CAST did not converge.

algorithm	sim measure/model	average	SD-weighted	CV-weighted	hierarchical	true mean	IMM
centroid-link	dist	**0.968**	0.849	0.802	**0.968**	0.159	NA
MCLUST	NA	**0.968**	NA	NA	**0.968**	0.806	NA
complete-link	dist	0.957	**0.968**	0.957	0.643	0.695	NA
complete-link	spherical	NA	NA	NA	NA	NA	**0.968**
complete-link	elliptical	NA	NA	NA	NA	NA	**0.968**
centroid-link	corr	**0.942**	0.807	0.753	**0.942**	**0.942**	NA
CAST	corr	0.881	0.682	NA*	NA	**0.898**	NA
k-means	corr	0.871	0.640	0.827	NA	**0.897**	NA
average-link	spherical	NA	NA	NA	NA	NA	0.897
average-link	elliptical	NA	NA	NA	NA	NA	0.897
average-link	dist	0.858	0.858	0.847	**0.869**	0.159	NA
average-link	corr	**0.866**	0.817	0.841	0.865	0.857	NA
k-means	dist	**0.857**	**0.857**	0.767	NA	0.159	NA
single-link	dist	**0.850**	0.144	0.194	0.160	0.017	NA
single-link	corr	**0.748**	-0.011	0.005	0.172	0.017	NA
complete-link	corr	0.677	0.724	0.730	0.503	**0.744**	NA
DIANA	dist	0.044	0.098	**0.388**	NA	-0.100	NA
DIANA	corr	0.142	**0.269**	-0.156	NA	-0.114	NA

Cluster Stability: with respect to synthetic re-measured data

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets. The maximum average adjusted Rand index for each row is shown in bold. The rows are sorted in descending order of the maximum average adjusted Rand index. DIANA and single-link produced very unstable clusters.

algorithm	sim measure/model	average	SD-weighted	CV-weighted	hierarchical	IMM
complete-link	elliptical	NA	NA	NA	NA	**0.998**
complete-link	spherical	NA	NA	NA	NA	**0.991**
average-link	dist	0.820	**0.985**	0.914	0.650	NA
MCLUST	NA	**0.963**	NA	NA	0.916	NA
complete-link	dist	0.927	**0.937**	0.830	0.441	NA
centroid-link	dist	0.893	**0.924**	0.841	0.893	NA
average-link	spherical	NA	NA	NA	NA	**0.923**
k-means	dist	**0.905**	0.867	0.798	NA	NA
average-link	elliptical	NA	NA	NA	NA	**0.895**
centroid-link	corr	**0.889**	0.758	0.644	**0.889**	NA
average-link	corr	0.842	0.842	**0.855**	0.828	NA
k-means	corr	**0.799**	0.709	0.781	NA	NA
CAST	corr	**0.756**	0.714	NA*	NA	NA
complete-link	corr	0.655	**0.700**	0.666	0.577	NA
single-link	dist	0.12	**0.606**	0.256	0.518	NA
single-link	corr	0.054	**0.307**	0.164	0.161	NA
DIANA	dist	**-0.016**	-0.612	-0.335	NA	NA
DIANA	corr	**-0.061**	-0.232	-0.292	NA	NA

Synthetic Data: randomly resampled yeast galactose data

We complement our empirical study with another set of synthetic data. We modified the randomly resampled approach in [1] to incorporate repeated measurements on the yeast galactose data.

1. For each experiment j (where j= 1, ..., E), estimate the true mean intensity levels (μ_{xij},μ_{yij}) for each gene i, and the gene-independent parameters ($\sigma_{\varepsilon xj}$, $\sigma_{\varepsilon yj}$, $\rho_{\varepsilon j}$, $\sigma_{\delta xj}$, $\sigma_{\delta yj}$, $\rho_{\delta j}$).
2. To generate the rth random repeated measurement for gene i in experiment j and class c, where i = 1, ..., 205, j=1, ..., 20, r= 1,..., 4:
    a. Randomly sample with replacement the estimated true mean intensity levels under the same experiment j in the same class c from the yeast galactose data. Let the randomly sampled estimated true mean intensities be (μ_{xij},μ_{yij}).
    b. Generate a random replicate observation using (μ_{xij},μ_{yij}) and the gene-independent parameters estimated in Step 1. Let the random observation be (x_{ij}^r, y_{ij}^r). The log ratio for this observation is $\log\frac{x_{ij}^r}{y_{ij}^r}$.

The size of each class in this synthetic data set is the same as in the real yeast galactose data. Due to independent random sampling from each experiment, any possible correlation between experiments is lost. Using this method, we generated synthetic data with R = 4, 10, or 100 repeated measurements from the yeast galactose data.

Results on randomly resampled synthetic yeast galactose data (without imputing missing values)

The results from the model-based approaches (MCLUST and IMM) are not available because the current implementations do not handle missing values.

Cluster Accuracy: agreement with the 4 classes

Each entry shows the adjusted Rand index of the corresponding algorithm with the 4 classes. The maximum adjusted Rand index of each row is shown in bold. The algorithms (rows) are sorted in descending order of the maximum adjusted Rand in each row. DIANA and single-link produced the least accurate clusters.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical
average-link	dist	0.788	**0.912**	0.858	0.859
centroid-link	dist	0.573	**0.908**	0.806	0.573
k-means	dist	0.838	**0.899**	0.895	NA
complete-link	dist	0.717	**0.890**	0.709	0.493
average-link	corr	0.807	0.765	0.736	**0.864**
centroid-link	corr	**0.817**	0.704	0.404	**0.817**
k-means	corr	**0.711**	0.674	0.649	NA
CAST	corr	**0.664**	0.595	0.627	NA
complete-link	corr	**0.624**	0.612	0.588	0.446
single-link	dist	0.159	**0.300**	0.123	**0.300**
DIANA	dist	0.028	0.097	**0.206**	NA
single-link	corr	**0.104**	-0.001	0.011	0.059
DIANA	corr	-0.117	-0.026	**-0.018**	NA

Cluster Stability: with respect to synthetic re-measured data

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets. The maximum average adjusted Rand index for each row is shown in bold. The rows are sorted in descending order of the maximum average adjusted Rand index. Entries marked with "-" means that we don't expect the results to be very interesting and didn't run the experiment.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical
centroid-link	dist	0.661	**0.945**	0.823	0.661
average-link	dist	0.702	**0.934**	0.907	0.892
k-means	dist	0.826	**0.892**	0.810	NA
single-link	dist	0.877	0.821	-	-
average-link	corr	0.783	0.758	0.712	**0.863**
complete-link	dist	0.664	**0.813**	0.682	0.302
centroid-link	corr	**0.712**	0.563	0.264	**0.712**
k-means	corr	**0.681**	0.668	0.542	NA
CAST	corr	0.645	0.632	-	NA
complete-link	corr	**0.598**	0.569	0.494	0.386
single-link	corr	0.113	0.079	-	-
DIANA	corr	-0.187	-0.209	-	NA
DIANA	dist	-0.247	-0.409	-	NA

Results on randomly resampled synthetic yeast galactose data (after imputed missing values)

Cluster Accuracy: agreement with the 4 classes

Each entry shows the adjusted Rand index of the corresponding algorithm with the 4 classes. The maximum adjusted Rand index of each row is shown in bold. The algorithms (rows) are sorted in descending order of the maximum adjusted Rand in each row. DIANA and single-link produced the least accurate clusters.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical
MCLUST	NA	0.968	NA	NA	**0.991**
complete-link	dist	0.917	**0.982**	0.902	0.524
average-link	dist	**0.971**	0.915	0.933	0.859
k-means	dist	**0.970**	0.911	0.903	NA
centroid-link	dist	0.899	**0.964**	0.828	0.899
centroid-link	corr	0.906	0.875	0.626	**0.906**
average-link	corr	0.878	0.836	0.809	**0.879**
k-means	corr	0.774	0.761	**0.847**	NA
complete-link	corr	**0.757**	0.664	0.699	0.584
single-link	corr	**0.460**	-0.010	0.008	0.042
single-link	dist	0.300	**0.440**	0.438	0.300
DIANA	dist	0.025	0.073	**0.130**	NA
DIANA	corr	-0.127	0.099	**0.123**	NA

Cluster Stability: with respect to synthetic re-measured data

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets. The maximum average adjusted Rand index for each row is shown in bold. The rows are sorted in descending order of the maximum average adjusted Rand index.

algorithm	sim measure	average	SD-weighted	CV-weighted	hierarchical
MCLUST	NA	**0.993**	NA	NA	0.982
average-link	dist	0.935	**0.953**	0.935	0.866
centroid-link	dist	**0.940**	0.930	**0.940**	**0.940**
k-means	dist	**0.932**	0.910	0.898	NA
complete-link	dist	0.882	**0.925**	0.904	0.367
centroid-link	corr	**0.915**	0.766	0.915	0.915
average-link	corr	**0.902**	0.801	0.821	0.893
k-means	corr	0.749	0.740	**0.786**	NA
complete-link	corr	**0.693**	0.630	0.589	0.504

Results on completely synthetic data

Cluster Accuracy: different numbers of simulated repeated measurements and different noise levels

Each entry shows the average adjusted Rand index of the corresponding algorithm with the 6 classes (average over 5 sets of completely synthetic data). For each number of simulated repeated measurements and noise level, the most accurate result is shown in bold. Since the modified SD-weighted approach only makes sense for distance, the results are not available (NA) for correlation.

algorithm	# rep	noise	sim measure or model	average	SD-weighted	modified SD-weighted dist	IMM
average-link	1	lo	corr	0.680	NA	NA	NA
average-link	1	lo	dist	0.789	NA	NA	NA
average-link	1	lo	spherical	NA	NA	NA	0.804
average-link	1	lo	elliptical	NA	NA	NA	0.804
complete-link	1	lo	corr	0.679	NA	NA	NA
complete-link	1	lo	dist	**0.831**	NA	NA	NA
complete-link	1	lo	spherical	NA	NA	NA	0.428
complete-link	1	lo	elliptical	NA	NA	NA	0.428
average-link	1	hi	corr	0.259	NA	NA	NA
average-link	1	hi	dist	0.000	NA	NA	NA
average-link	1	hi	spherical	NA	NA	NA	**0.395**
average-link	1	hi	elliptical	NA	NA	NA	**0.395**
complete-link	1	hi	corr	0.252	NA	NA	NA
complete-link	1	hi	dist	0.053	NA	NA	NA
complete-link	1	hi	spherical	NA	NA	NA	0.214
complete-link	1	hi	elliptical	NA	NA	NA	0.214
average-link	4	lo	corr	0.764	0.576	NA	NA
average-link	4	lo	dist	0.877	0.927	0.962	NA
average-link	4	lo	spherical	NA	NA	NA	0.926
average-link	4	lo	elliptical	NA	NA	NA	**0.957**
complete-link	4	lo	corr	0.755	0.584	NA	NA
complete-link	4	lo	dist	0.925	0.876	0.992	NA
complete-link	4	lo	spherical	NA	NA	NA	0.897
complete-link	4	lo	elliptical	NA	NA	NA	**0.957**
average-link	4	hi	corr	0.389	0.519	NA	NA
average-link	4	hi	dist	0.000	0.713	0.831	NA
average-link	4	hi	spherical	NA	NA	NA	0.589
average-link	4	hi	elliptical	NA	NA	NA	**0.911**
complete-link	4	hi	corr	0.450	0.518	NA	NA
complete-link	4	hi	dist	0.498	0.798	0.871	NA
complete-link	4	hi	spherical	NA	NA	NA	0.559
complete-link	4	hi	elliptical	NA	NA	NA	0.910
average-link	20	lo	corr	0.854	0.701	NA	NA
average-link	20	lo	dist	0.891	0.964	**1.000**	NA
average-link	20	lo	spherical	NA	NA	NA	0.962
average-link	20	lo	elliptical	NA	NA	NA	0.957
complete-link	20	lo	corr	0.809	0.675	NA	NA
complete-link	20	lo	dist	0.890	0.999	**1.000**	NA
complete-link	20	lo	spherical	NA	NA	NA	0.872
complete-link	20	lo	elliptical	NA	NA	NA	0.954
average-link	20	hi	corr	0.602	0.651	NA	NA
average-link	20	hi	dist	0.590	0.819	**0.961**	NA
average-link	20	hi	spherical	NA	NA	NA	0.688
average-link	20	hi	elliptical	NA	NA	NA	0.953
complete-link	20	hi	corr	0.608	0.690	NA	NA
complete-link	20	hi	dist	0.710	0.920	0.960	NA
complete-link	20	hi	spherical	NA	NA	NA	0.854
complete-link	20	hi	elliptical	NA	NA	NA	0.957

Cluster Stability: at low noise levels

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets.

alg	# rep	sim	avg over replicates	SD-weighted	CV-weighted
average-link	1	corr	0.977	NA	NA
average-link	1	dist	0.945	NA	NA
complete-link	1	corr	0.892	NA	NA
complete-link	1	dist	0.868	NA	NA
centroid-link	1	corr	0.980	NA	NA
centroid-link	1	dist	0.948	NA	NA
k-means	1	corr	0.979	NA	NA
k-means	1	dist	0.941	NA	NA
average-link	4	corr	**0.958**	0.832	0.854
average-link	4	dist	0.970	**0.998**	**0.998**
complete-link	4	corr	**0.947**	0.580	0.671
complete-link	4	dist	0.968	0.923	**0.981**
centroid-link	4	corr	**0.959**	0.861	0.866
centroid-link	4	dist	0.984	**0.998**	0.986
k-means	4	corr	**0.958**	0.693	0.634
k-means	4	dist	0.954	**0.998**	0.967
average-link	20	corr	**0.978**	0.962	0.947
average-link	20	dist	0.958	**0.999**	**0.999**
complete-link	20	corr	**0.972**	0.735	0.783
complete-link	20	dist	0.936	0.973	**0.999**
centroid-link	20	corr	**0.978**	0.975	0.903
centroid-link	20	dist	0.965	0.992	**0.999**
k-means	20	corr	**0.978**	0.870	0.706
k-means	20	dist	0.958	0.995	**0.996**

Lung cancer data

Bhattacharjee *et al*. [2] studied the variation of expression levels in lung tumor tissue samples over 12600 transcript sequences using Affymetrix oligonucleotide arrays. On Affymetrix arrays, 11-20 probe pairs of short sequences are used to interrogate each target gene. In other words, these probe pairs measure the expression levels of different portions of the same gene, and hence provide a form of repeated information. Many statistics have been proposed to summarize the expression levels of genes from measured intensities of probe pairs, for example, [3], [4], [5], [6]. Some of these statistics also provide estimated standard errors of expression levels. We treat these standard errors as variability estimates of expression levels. In particular, we used the AvLog(PM-BG) statistic and the corresponding standard errors from Irizarry *et al*. [6]. There are 254 experiments on five types of tissue samples, including lung adenocarcinomas, pulmonary carcinoids, squamous cell lung carcinomas, small-cell lung carcinomas, and normal lung specimens. Our goal is to cluster the experiments, and these five tissue types form our external knowledge. We used a subset of approximately 800 probe sets which have high reproducibility across the adenocarcinoma samples [2].

Results on the lung cancer data

Cluster Accuracy: agreement with the 5 tissue types

Table 4 shows selected results of the adjusted Rand indices of comparing clustering results on the lung cancer data to the five tissue types. We produced five clusters using various clustering algorithms with both correlation and distance. In the "AvLog(PM-BG)" column, we used the AvLog(PM-BG) statistic [6] in computing similarity measures. In the "SE-weighted" column, we used the AvLog(PM-BG) statistic and the associated standard error (SE) [6] in computing

variability-weighted similarity measures. The SE-weighted approach produced more accurate clusters in 4 out of 6 cases (except when average-link and centroid-link are applied with correlation).

algorithm	sim measure	AvLog(PM-BG)	SE-weighted AvLog(PM-BG)
average-link	corr	**0.414**	0.402
centroid-link	corr	**0.399**	0.368
centroid-link	dist	0.366	**0.376**
average-link	dist	0.275	**0.372**
complete-link	dist	0.098	**0.354**
complete-link	corr	0.232	**0.299**

Cluster Stability: varying numbers of replicates

Each entry shows the average adjusted Rand index of the clusters on the original data with clusters from the synthetic re-measured data sets with different numbers of replicates

alg	sim	4 rep avg over rep	4 rep SD-weighted	10 rep avg over rep	10 rep SD-weighted	20 rep avg over rep	20 rep SD-weighted
average-link	corr	**0.784**	0.071	**0.869**	0.854	0.875	**0.974**
average-link	dist	**0.946**	0.553	**0.948**	0.721	**1.000**	0.828
complete-link	corr	**0.750**	0.118	**0.588**	0.377	**0.720**	0.484
complete-link	dist	0.409	**0.464**	**0.553**	0.479	0.576	0.575
centroid-link	corr	**0.979**	0.028	**1.000**	0.082	**1.000**	0.770
centroid-link	dist	**0.960**	0.867	0.975	**0.981**	0.960	**0.967**
k-means	corr	0.969	NA	0.996	NA	0.996	NA
k-means	dist	0.944	NA	0.961	NA	0.993	NA

References

1. KY Yeung, C Fraley, A Murua, AE Raftery, WL Ruzzo: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**:977-987.
2. A Bhattacharjee, WG Richards, J Staunton, C Li, S Monti, P Vasa, C Ladd, J Beheshti, R Bueno, M Gillette, et al: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proceedings of the National Academy of Science USA* 2001, **98**:13790-13795.
3. Affymetrix: **Statistical Aglorithms Reference Guide**. 2001.
4. C Li, WH Wong: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection**. *Proceedings of the National Academy of Science USA* 2001, **98**:31-36.
5. WJ Lemon, JJT Palatini, R Krahe, FA Wright: **Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays**. 2001.
6. RA Irizarry, B Hobbs, F Collin, YD Beazer-Barclay, KJ Antonellis, U Scherf, TP Speed: **Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data**. 2002.