Biostatistics (2007), **8**, 1, *pp.* 9–31
doi:10.1093/biostatistics/kxj029
Advance Access publication on April 13, 2006

Are clusters found in one dataset present in another dataset?

AMY V. KAPP*

Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA
akapp@stanford.edu

ROBERT TIBSHIRANI

Department of Health Research and Policy and Department of Statistics,
Stanford University, Stanford, CA, USA

SUMMARY

In many microarray studies, a cluster defined on one dataset is sought in an independent dataset. If the cluster is found in the new dataset, the cluster is said to be "reproducible" and may be biologically significant. Classifying a new datum to a previously defined cluster can be seen as predicting which of the previously defined clusters is most similar to the new datum. If the new data classified to a cluster are similar, molecularly or clinically, to the data already present in the cluster, then the cluster is reproducible and the corresponding prediction accuracy is high. Here, we take advantage of the connection between reproducibility and prediction accuracy to develop a validation procedure for clusters found in datasets independent of the one in which they were characterized. We define a cluster quality measure called the "in-group proportion" (IGP) and introduce a general procedure for individually validating clusters. Using simulations and real breast cancer datasets, the IGP is compared to four other popular cluster quality measures (homogeneity score, separation score, silhouette width, and weighted average discrepant pairs score). Moreover, simulations and the real breast cancer datasets are used to compare the four versions of the validation procedure which all use the IGP, but differ in the way in which the null distributions are generated. We find that the IGP is the best measure of prediction accuracy, and one version of the validation procedure is the more widely applicable than the other three. An implementation of this algorithm is in a package called "clusterRepro" available through The Comprehensive R Archive Network (http://cran.r-project.org).

Keywords: Breast cancer subtypes; Cluster validation; In-group proportion; Prediction accuracy.

1. INTRODUCTION

As the name suggests, cluster validation is concerned with "assessing the validity of classifications that have been obtained from the application of clustering procedure" (Gordon, 1999). In general, cluster validation procedures define a cluster quality measure (e.g. a measure of isolation or a measure of cohesion)

*To whom correspondence should be addressed.

and determine how likely given values of that measure are to occur under a null model of no structure. Either graph theory or Monte Carlo simulations can be used to find the null distribution of the cluster quality values.

Interest in cluster validation has been re-ignited by the need for gauging the significance of gene and array clusters in microarray studies. The majority of the literature has centered on determining which clustering procedure to use and on determining how many clusters are present in a microarray dataset (Datta and Datta, 2003; Chen *and others*, 2002; Kerr and Churchill, 2001; Yeung *and others*, 2001; Levine and Domany, 2001).

Although many of these papers used a cluster quality measure based on within-cluster and/or between-cluster variance, three papers (Dudoit and Fridlyand, 2002; Dudoit *and others*, 2002; Tibshirani and Walther, 2005) used prediction error to evaluate the quality of clusters. When the true classifications of the test dataset were known, as in Dudoit *and others* (2002), the estimate of prediction error was the proportion of correct classifications in the test dataset. When the true classifications are unknown, as in Tibshirani and Walther (2005), cluster quality can be estimated by how well training centroids predict test set co-memberships, i.e. pairs of observations classified to the same cluster. Instead of concentrating on a single measure of prediction accuracy, Dudoit and Fridlyand (2002) compared a variety of indices to measure the agreement between the training set partition and the test set partition.

Despite their differences, all three papers argued that the use of a measure of test set clusters defined by a classifier made from the training data is the most appropriate approach to cluster validation when the aim of analyzing the microarray data is to identify reproducible clusters of genes or arrays with similar expression profiles. The genes or samples of a microarray dataset are partitioned into clusters and used to build a classifier which is applied to new data. If the new data classified to a cluster are like the samples already present in the cluster (molecularly or clinically), then the cluster is validated because it is reproducible and may be biologically significant.

In other words, a classifier built using previously defined clusters is used to predict which new data have certain molecular or clinical characteristics in common with the other members of the cluster. A cluster is validated if enough predictions are correct because accurate predictions mean the cluster is present in the new data. Therefore, when the goal of a study is the identification of reproducible clusters, validation is related to prediction accuracy which is defined to be the proportion of data whose predicted classifications are identical to the true classifications.

This paper extends the idea of using prediction accuracy (or strength) from validating the number of clusters or the choice of clustering method to validating individual clusters found in a new dataset. First, a new cluster quality measure is proposed. The "in-group proportion" (IGP) is similar to the measure of co-memberships in Tibshirani and Walther (2005). It is defined to be the proportion of observations classified to a cluster whose nearest neighbor is also classified to the same cluster.

The IGP also resembles a cluster quality measure proposed by Bailey and Dubes in 1982. Their "measure of cohesion" was defined for a random graph with m edges and n vertices: $W_C(m) = \#\{(a, b) | a, b \in C, (a, b) \in S_m\}$, where S_m was the set of (a, b) edges in the graph. If C is made up of observations in the same cluster and S_m is the set of edges that connect observations in C to their nearest neighbors, then in certain situations W_C is equal to the product of the IGP and the total number of observations in the cluster (m). For example, consider three points on the real line: 0, 1, and $\frac{3}{2}$. If $C = \{0, 1\}$, S_m is as defined above, and Euclidean distance is used, then $W_C(2) = 1$ and the IGP is $\frac{1}{2} = \frac{W_C(2)}{m}$.

In the subsequent sections, a more explicit description of the IGP and four other cluster quality measures are presented, after which a cluster validation procedure is proposed. Four different versions of the cluster validation procedure are described. In all versions, the IGPs for all the clusters in a new dataset identified by centroids built on a previous dataset are computed and then compared to an appropriate null distribution to obtain p-values. The null distributions of IGPs, however, are generated differently in each version of the cluster validation procedure. Finally, simulations and real datasets are used to compare the

IGP with four other cluster quality measures and to compare the four versions of the cluster validation procedure.

2. Methods

Before describing the cluster quality measures and validation procedure, some basic definitions must be established. We let A be an $m \times n$ matrix of microarray data where m is the number of features (genes) and n is the number of samples (arrays). We assume that a subset of the samples of A have been partitioned into p groups (labeled $1, 2, \ldots, p$) and C is the $m \times p$ matrix of the centroids. The uth column of C is made by averaging over the features (rows) of the samples in A classified to group u. If X is an $m \times q$ matrix of microarray data independent of A, then all the samples (columns) of X can be classified to one of the p groups or to a "below-cutoff group" using C and a cutoff (c). The function $d(x, y)$ is defined to be the Pearson's (centered) correlation for vectors x and y and $\text{Class}_X(j)$ is the class label for sample j of X:

$$\text{Class}_X(j) = \begin{cases} 0, & \text{if } \max_{1 \leqslant u \leqslant p} d(X[,j], C[,u]) < c, \\ \operatorname{argmax}_{1 \leqslant u \leqslant p} d(X[,j], C[,u]), & \text{if } \max_{1 \leqslant u \leqslant p} d(X[,j], C[,u]) \geqslant c. \end{cases} \tag{2.1}$$

Since $d(x, y)$ is a measure of correlation not distance, $0 \leqslant d(x, y) \leqslant 1$ and a $d(x, y)$ near 1 means x and y are close together. Thus, every sample of X is classified to the the group whose centroid with which it most highly correlates. If the maximum correlation for a sample and any of the centroids is less than c, the sample receives the class label 0. The below-cutoff group is composed of all the samples i of X for which $\text{Class}_X(i) = 0$. A cluster quality measure is subsequently computed for each group to which at least one array of X is classified.

2.1 *The IGP*

We propose a new cluster quality measure based on the idea of prediction accuracy. The IGP (Figure 1) is defined to be the proportion of samples in a group whose nearest neighbors are also in the same group. In other words, the IGP quantifies how often points near each other are predicted to belong to the same group. Define $j^N = \operatorname{argmax}_{k \neq j} d(X[,j], X[,k])$ for each columnn of X, and let u be the class label for all the samples whose $\text{Class}_X = u$, then

$$\text{IGP}(u, X) = \frac{\#\{j | \text{Class}_X(j) = \text{Class}_X(j^N) = u\}}{\#\{j | \text{Class}_X(j) = u\}}. \tag{2.2}$$

For the jth sample of X, j^N is j's nearest neighbor and so $\text{IGP}(u, X)$ is the proportion of samples in class u whose nearest neighbor is also in class u.

If a distance function is used instead of Pearson's (centered) correlation coefficient, then the above definitions still hold with min replacing max, argmin replacing argmax, $\leqslant$ replacing $\geqslant$, and $>$ replacing $<$.

2.2 *Other cluster quality measures*

The IGP is not the only measure of cluster quality. Chen *and others* (2002) described several others. Since Chen *and others* (2002) defined the "homogeneity score" (HS), separation score (SS), "silhouette width" (SW), and "weighted average discrepant pairs" (WADP) for entire clusterings as opposed to

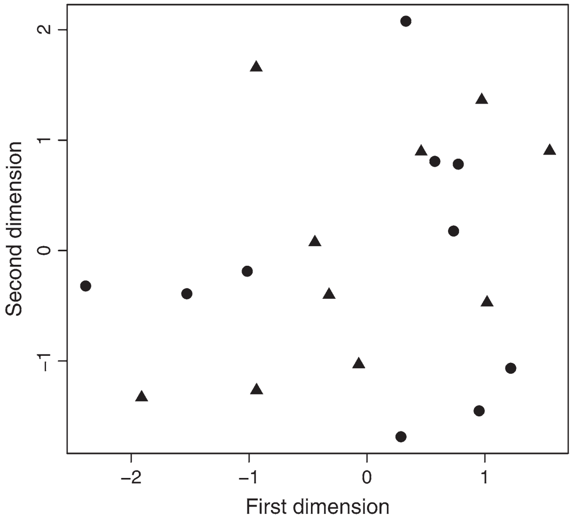

Fig. 1. Using Euclidean distance, the IGP for the circles is 0.8 and that for the triangles is 0.6.

individual clusters, we slightly modified the scores to apply to an individual cluster. First, the HS is defined to be the average correlation between a cluster's centroid and the members of the cluster. If $\text{Set}_u = \{j|\text{Class}_X(j) = u\}$ and N_u is the number of elements in Set_u, then the HS for cluster u is: $\text{HS}_u = \frac{1}{N_u}\sum_{j\in\text{Set}_u} d(X[,j], C[,u])$.

The SS for cluster u is the weighted average of the correlation between the uth cluster's centroid and every other centroid: $\text{SS}_u = \frac{1}{\sum_{v\neq u} N_v}\sum_{v\neq u} N_v d(C[,v], C[,u])$.

Next, we assume that $j \in \text{Set}_u$ and let $a(j)$ be the average dissimilarity (or distance) between sample j and the other samples in Set_u and $b(j)$ be the average dissimilarity (or distance) between sample j and the samples not in Set_u. The SW for cluster u is thus defined: $\text{SW}_u = \frac{1}{N_u}\sum_{j\in\text{Set}_u} \frac{b(j)-a(j)}{\max\{a(j),b(j)\}}$.

Since Pearson's (centered) correlation coefficient is a measure of similarity, 1 − |Pearson's (centered) correlation| was used as the measure of dissimilarity to compute the SWs and every time a measure of distance was required. Therefore, for each member of a cluster u, the discrepancy between the average value of 1 − |Pearson's (centered) correlation| between that member and the other members of the cluster and the average 1 − |Pearson's (centered) correlation| between the member and the members outside of the cluster is calculated and then divided by the maximum of those two quantities. The SW for cluster u is the average of these quotients over all the members of cluster u.

Finally, the WADP score measures the consistency of a classifier when the samples are subject to small perturbations. We generate an $m \times q$ matrix of Gaussian random variables: $R = [r_{i,j}]$ where $r_{i,j} \overset{iid}{\sim} N(0, \sigma^2_{\text{WADP}})$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant q$. R is added to X and the samples of X are reclassified. For each cluster, we calculate the number of sample pairs that were in the same cluster in the original classification, but not in the same cluster after reclassification. That quantity is divided by the total number of sample pairs originally in the cluster. This process is repeated many times and the WADP score for cluster u is the average of these ratios taken over the perturbations of X. The value of σ^2_{WADP} is specified by the user and has a large impact on the WADP score. If σ^2_{WADP} is too small, the WADP score is always 0; if σ^2_{WADP} is too large, the WADP score is close to 1. In this paper, σ^2_{WADP} was always chosen to be large enough for the WADP scores to vary between groups.

2.3 *Null distribution generation*

To validate the groups found in X, the IGPs of those groups are compared to a null distribution of IGPs. Four different versions of the same procedure are used in this paper to generate null distributions of IGPs. The basic null distribution generation procedure repeatedly generates an $m \times p$ centroid matrix (C^*), computes Class_X^* in the way described above with C^* replacing C, and calculates the IGPs for the groups in Class_X^*. Each version of the null distribution generation procedure generates C^* differently.

Version 1 permutes each row of C to get C^*.

Version 2 permutes the rows of A, hierarchically clusters the columns (average linkage), automatically cuts the dendogram to make p groups, and averages over the rows of the arrays with the same group labels to get C^*.

Version 3 transforms C to get C^* (transformation described below).

Version 4 transforms A (transformation described below), hierarchically clusters the columns (average linkage), automatically cuts the dendogram to make p groups, and averages over the rows of the arrays with the same group labels to get C^*.

The first two versions assume independence of the genes in the centroids or raw data. As many microarray studies have demonstrated, however, genes are not completely independent. Therefore, the centroids produced by Versions 1 and 2 may not be near the data. To remedy this problem, the third and fourth versions permute the samples within the box aligned with their principal components. This transformation increases the chance that the centroids are near the data without being too similar to the actual data or actual centroids which would bias the p-values towards 1.

1. Let $W = UDV^T$ be the singular value decomposition of W. (W can be either C or A.)
2. Define $W' = WV$.
3. Permute the columns of W' to obtain Z'.
4. Let $Z = Z'V^T$.
5. Substitute Z for W.

The null distribution generation methods were designed to produce centroids that are placed randomly in the data. As a consequence, the groups defined by the centroids most likely are not high-quality clusters. Thus, the null distributions are composed of IGPs that come from groups of data that are not high-quality clusters. Since a cluster of high-quality will have an IGP close to 1, the p-value of a group is the fraction of the null distribution IGPs that are as close or closer to 1 than the group's actual IGP. In other words, the null hypothesis is that a group is not a high-quality cluster and it is rejected if the actual IGP of the group is high enough (i.e. close enough to 1).

A group of data with a significant p-value is a high-quality cluster. In addition, that group of data corresponds to a cluster in an independent dataset (the one in which the original centroids were formed). Therefore, a significant p-value means a high-quality cluster (as opposed to a group of data near each other) corresponding to the original cluster was found in an independent dataset. Hence, the cluster is reproducible and thus validated.

Since the IGPs depend on the size of the group, $\text{IGP}(u, X)$ is compared only to the IGPs from the null distribution generation procedure that come from groups of the same size. When a cutoff is used, the below-cutoff group is compared to the IGPs of all the below-cutoff groups obtained from the null distribution procedure because the sizes of the generated below-cutoff groups so rarely match the size of the actual below-cutoff group.

Nothing about the null distribution generation procedure is specific to the IGP. Therefore, the overall cluster validation method and its four versions could be used with any of the cluster quality measures described in Section 2.2. In light of the results presented in Section 3.1, however, we only compared the null distribution generation versions for the IGP.

3. SIMULATIONS

Results from five simulations are presented: two (Simulation 1 and Simulation 2) were done to compare the five cluster quality measures described in Section 2.2 and two (Simulation 3, Simulation 4, and Simulation 5) were done to compare the different versions of the null distribution generation procedure described in Section 2.3. All the simulations used Pearson's (centered) correlation coefficient and datasets of 300 observations. The details and results of the simulations are described in Sections 3.1 and 3.2.

3.1 *Comparison of cluster quality measures*

The datasets for Simulation 1 and Simulation 2 were generated in the same fashion. First, a single vector of length 500 was defined: $P = (p_1, p_2, \ldots, p_{500})$ such that $p_i \overset{iid}{\sim} N(0, 50)$. Then, the S_u $(u = 1, 2, 3, 4)$ were defined to be random samples of size 50 drawn without replacement from the set $\{1, 2, \ldots, 500\}$. Using P and the S_u's, a 500×4 matrix (Q), which can be thought of as the matrix of true centroids, was defined:

$$Q[i, u] = \begin{cases} p_i + y_{i,u}, & \text{if } i \in S_u, \\ p_i, & \text{if } i \notin S_u. \end{cases} \tag{3.1}$$

The $y_{i,u}$ were independent identically distributed $N(0, \sigma_u^2)$. To produce the data matrix of observations, the variable T_j was defined to be a random sample of size 100 drawn without replacement from the set $\{1, 2, \ldots, 500\}$ for $j = 1, 2, \ldots, 300$. Thus, the data matrix (R) was defined:

$$R[i, j] = \begin{cases} Q[i, 1] + z_{i,j,1}, & \text{if } j \leqslant 50 \text{ and } i \in T_j, \\ Q[i, 1], & \text{if } j \leqslant 50 \text{ and } i \notin T_j, \\ Q[i, 2] + z_{i,j,2}, & \text{if } 51 \leqslant j \leqslant 100 \text{ and } i \in T_j, \\ Q[i, 2], & \text{if } 51 \leqslant j \leqslant 100 \text{ and } i \notin T_j, \\ Q[i, 3] + z_{i,j,3}, & \text{if } 101 \leqslant j \leqslant 200 \text{ and } i \in T_j, \\ Q[i, 3], & \text{if } 101 \leqslant j \leqslant 200 \text{ and } i \notin T_j, \\ Q[i, 4] + z_{i,j,4}, & \text{if } 201 \leqslant j \leqslant 300 \text{ and } i \in T_j, \\ Q[i, 4], & \text{if } 201 \leqslant j \leqslant 300 \text{ and } i \notin T_j. \end{cases} \tag{3.2}$$

The $z_{i,j,u}$ were independent identically distributed $N(0, \eta_u^2)$.

R is like a matrix of microarray data where the 500 rows are genes (features) and the 300 columns are arrays (samples). Each column of R was classified to one of four groups: columns 1–50 were the first group, columns 51–100 were the second group, columns 101–200 were the third group, and columns 201–300 were the fourth group. The 500×4 matrix, $\overline{Q}$, was found by averaging over the columns of R which were generated from the same column of Q: $\overline{Q}[i, u] = \frac{1}{n_u} \sum_{k=a_u}^{b_u} R[i, j]$, where $n_1 = n_2 = \frac{1}{2} n_3 = \frac{1}{2} n_4 = 50$, $(a_1, a_2, a_3, a_4) = (1, 51, 101, 201)$, and $(b_1, b_2, b_3, b_4) = (50, 100, 200, 300)$.

In Simulation 1, $\eta_u^2 = 100$ for all u, but $\sigma_1^2 = \sigma_3^2 = 2\sigma_2^2 = 2\sigma_4^2$ and $\sigma_1^2 \in \{2, 4, 6, \ldots, 40\}$. In Simulation 2, $\sigma_u^2 = 25$ for all j, but $\eta_1^2 = \eta_3^2 = \frac{1}{2}\eta_2^2 = \frac{1}{2}\eta_4^2$ and $\eta_1^2 \in \{10, 30, 50, \ldots, 250\}$. In other words, if Q is thought of as the centroid matrix, then in Simulation 1, as σ_1^2 increased the centroids moved further apart from one another, but the correlations between the data and the centroids remained constant. Furthermore, in Simulation 2, as η_1^2 increased the data moved further away from their centroids, but the between-centroid correlations remained constant.

For each value of σ_1^2 in Simulation 1 and η_1^2 in Simulation 2, 100 datasets (R) were generated in the manner described above. For each of the datasets, the IGPs, HSs, SSs, SWs, and WADP scores ($\sigma_{\text{WADP}} = 10$) were computed using two different classifications. One was the true classification:

$$\text{True classification of } R[, j] = \begin{cases} 1, & \text{if } j \leqslant 50, \\ 2, & \text{if } 51 \leqslant j \leqslant 100, \\ 3, & \text{if } 101 \leqslant j \leqslant 200, \\ 4, & \text{if } 201 \leqslant j \leqslant 300. \end{cases} \tag{3.3}$$

The other was the estimated classification: Estimated classification of $R[, j] = \text{argmax}_{1 \leqslant u \leqslant 4}\, d(\overline{Q}[, u], R[, j])$.

For both the true classifications and the estimated classifications, the average values for all five cluster quality measures are presented in Figures 2 and 3.

A cluster is said to be “isolated” if the members of the cluster are very different from the members of other clusters and a cluster is said to be “cohesive” if the members of the cluster are very similar to each other (Gordon, 1999). In contrast, a measure of prediction accuracy needs to quantify how likely a point classified to one cluster is to have been classified to another cluster because prediction accuracy is the proportion of data whose predicted classifications are identical to the true classifications. If a cluster is both isolated and cohesive, its members are unlikely to be classified to another cluster. Therefore, a measure of prediction accuracy will be sensitive to both isolation and cohesion.

In the context of the above simulations, this implies an appropriate cluster quality measure should consistently increase (or decrease) as σ_1^2 increased in Simulation 1 (causing the between-column correlation of Q to decrease) and as η_1^2 increased in Simulation 2 (causing the correlation between the columns of R and their associated columns of Q to decrease). In Figure 2, the true classification curves of all the cluster quality measures consistently increased or decreased with σ_1^2: the IGP, HS, and SW increased while the SS and WADP score decreased. Although the HS’s true classification curves increased, the scale of the increase over the range of σ_1^2 was so small (0.78–0.79) that the HS was basically unchanged. In other words, as the groups become more isolated, the HS was constant. Therefore, the HS is not an appropriate measure of isolation.

In Figure 3, all the true classification curves for the IGP, HS, and WADP score decreased with the increase in η_1^2. In contrast, the true classification curves for the SS were constant and two of the SW true classification curves changed direction, first decreasing and then increasing. Therefore, the SS and SW are not appropriate measures of cohesion.

Based upon the true classification curves, only the IGP and WADP score were appropriate cluster quality measures. The WADP score had two major drawbacks as a cluster quality measure, however. First, the estimated WADP scores differed greatly from the true WADP scores in Figure 2 for $\sigma_1^2 < 10$. Second, the WADP score required us to choose the value of σ_{WADP}^2. The true IGPs were different from the estimated IGPs for $\sigma_1^2 < 10$, but the difference was not as great. In addition, the IGP did not require values for parameters to be chosen before calculating the score. Thus, these simulations demonstrate that the IGP was a better cluster quality measure than the HS, SS, SW, and WADP score.

These simulations also show a cluster’s IGP depended upon the cluster’s size, average correlation between members, and average correlation between each member and the centroids. In Figure 2 IGP graph, the two groups of 50 observations traced out similar IGP true classification curves (black and green) that differed from those of the two groups of 100 observations (red and blue). In Figure 3 IGP graph, however, all four groups traced out different IGP true classification curves.

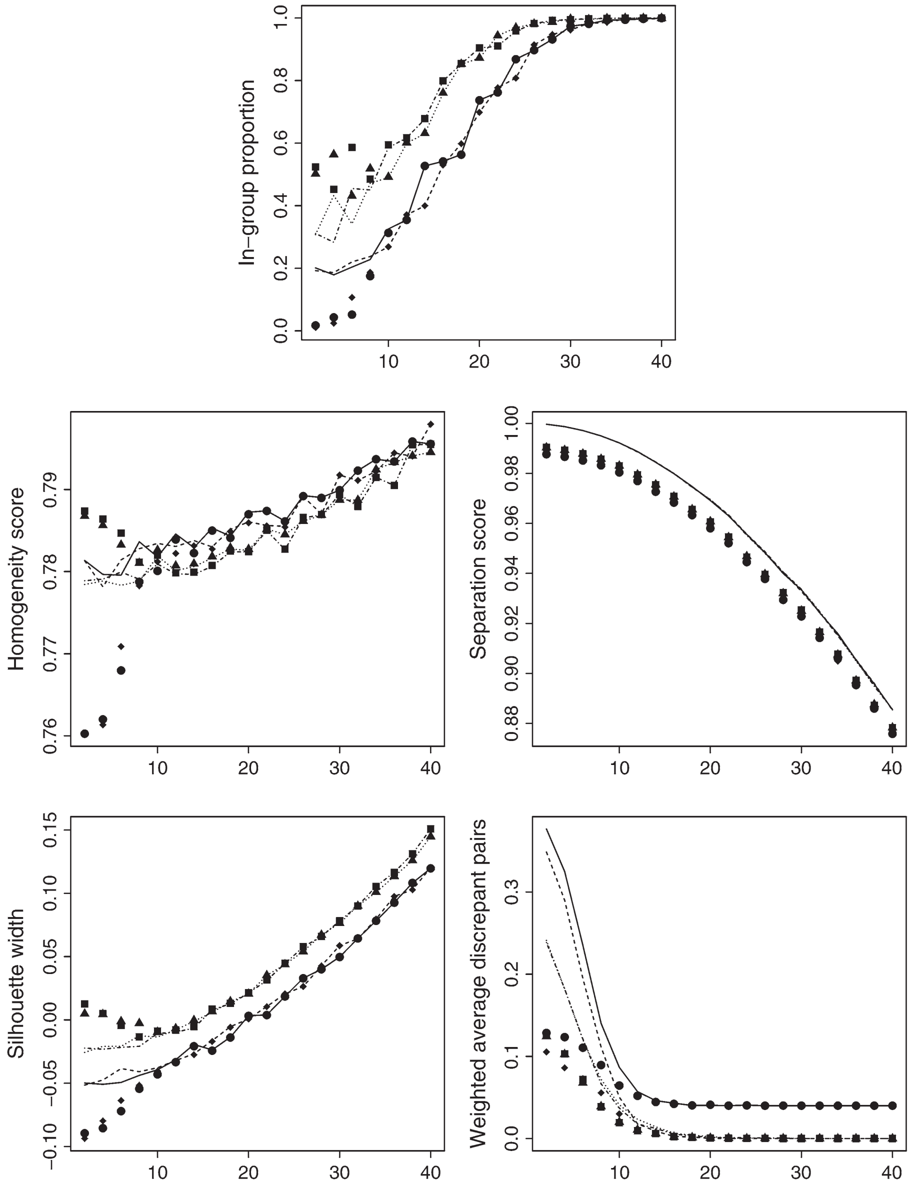

Fig. 2. These graphs show the results of Simulation 1. The horizontal axis on each graph is the value of σ_1^2 used to generate each Q matrix. The vertical axis is the cluster quality measure. The lines trace out the average cluster quality measure when the true classifications were used; the solid points are the average cluster quality measure values when the estimated classifications were used. The averages from observations generated using σ_1^2, σ_2^2, σ_3^2, and σ_4^2 are circles (solid line), diamonds (dashed line), triangles (dotted line), and squares (dashed and dotted line), respectively.

Finally, the IGP of a cluster depended upon the composition of the entire population. When the third and fourth groups were removed from Simulation 1 and Simulation 2, the IGPs for the first and second groups increased when the clusters were not very cohesive or not very isolated (Figure 4). (In each of the 100 repetitions, the third and fourth columns of Q were removed, columns 101–300 were removed from

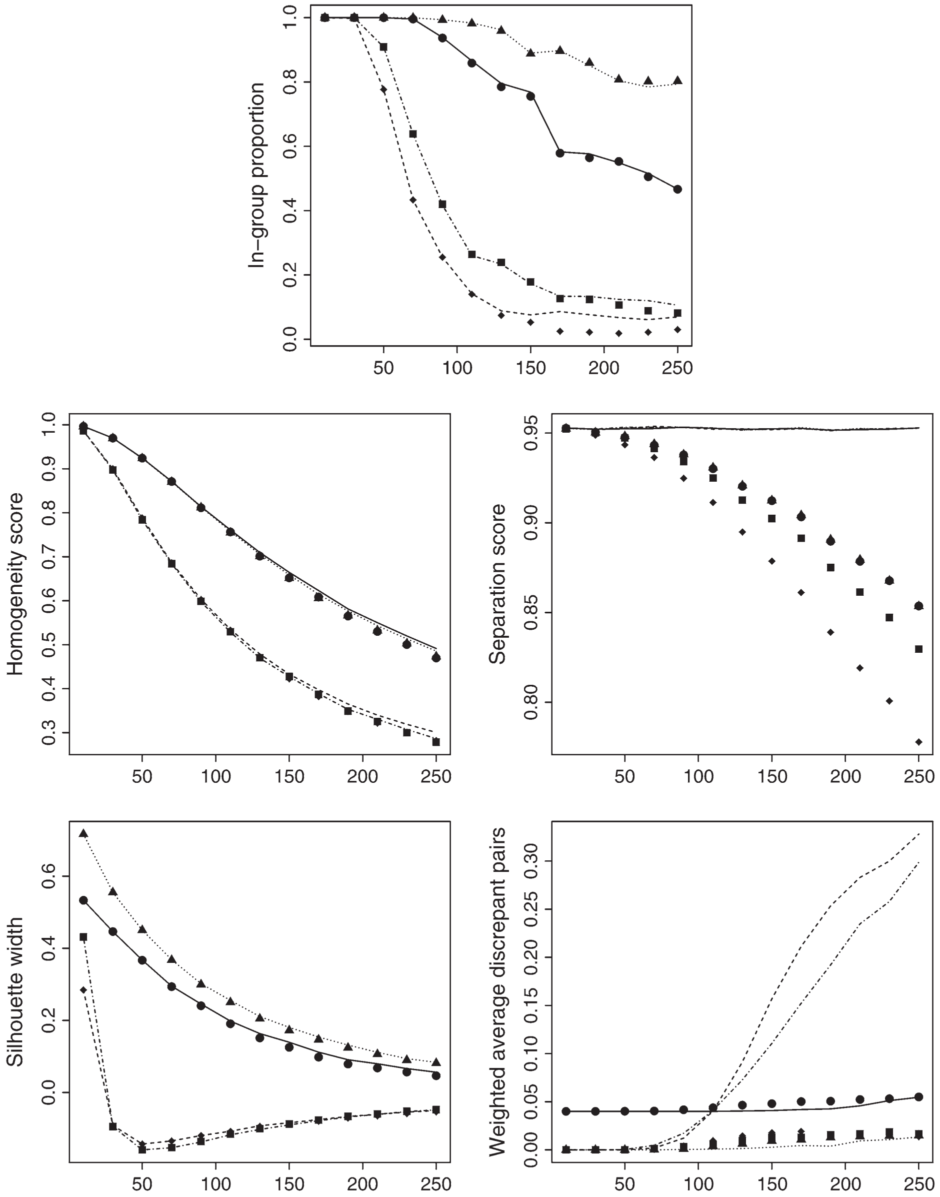

Fig. 3. These graphs show the results of Simulation 2. The horizontal axis on each graph is the value of η_1^2 used to generate each data matrix (R) of 300 observations. The vertical axis is the cluster quality measure. The lines trace out the cluster quality measure when the true classifications were used; the solid points are the average cluster quality measure values when the estimated classifications were used. The averages for the observations generated using η_1^2, η_2^2, η_3^2, and η_4^2 are circles (solid line), diamonds (dashed line), triangles (dotted line), and squares (dashed and dotted line), respectively.

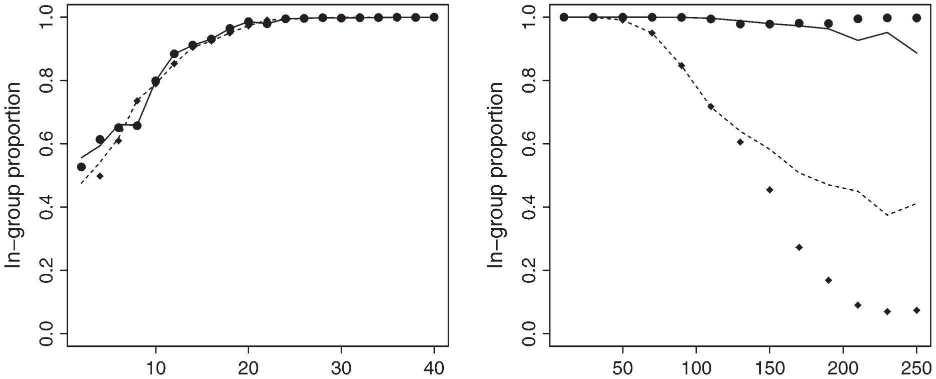

Fig. 4. The IGP averages for the first and second groups in the absence of the third and fourth groups for Simulation 1 (left) and Simulation 2 (right). The lines trace out the average IGPs for the true classifications; the solid points are the average IGPs for the estimated classifications. (Left) The IGP averages of observations generated by σ_1^2 and σ_2^2 are circles (solid line) and diamonds (dashed line), respectively. (Right) The IGP averages of observations generated by η_1^2 and η_2^2 are circles (solid line) and diamonds (dashed line), respectively.

R, the true classifications for the first 50 columns of R were 1, and the true classifications for the second 50 columns of R were 2.) Therefore, the IGP of a cluster that is not very isolated or cohesive will decrease in the presence of other clusters.

3.2 *Comparison of null distribution generation versions*

To apply the null distribution generation method proposed in Section 2.3, two independent datasets are required. The first is the one in which the clusters are initially identified and upon which the centroids are formed, and the second is the one whose columns are classified using these centroids. Hence, pairs of independent R matrices were made repeatedly for Simulations 3–5 which compare the null distribution generations versions. In Simulation 3, both R matrices were made identically to the R matrix in Simulation 1. In Simulation 4, both R matrices were made identically to the R matrix in Simulation 2. In Simulation 5, both R matrices were made like the R matrix in Simulation 1 with one important difference. For $u = 2, 4$, not all $Q[i, u]$ and $Q[l, u]$ were independent. After the Q matrix was generated and before the R matrix was generated, the following transformation was performed: $Q[i, u] = Q[250 + i, u]$ for $1 \leqslant i \leqslant 250$ and $u = 2, 4$.

In Simulations 3–5, the true classifications of both R matrices were the same. The true classifications were used to make $\overline{R}$ from one R only by averaging over the rows of columns with the same classifications. These centroids were then applied to the other R matrix that was not used to make the centroids.

Unlike Simulations 1 and 2, only 20 datasets were generated for each standard deviation value (either σ_1^2 or η_1^2). Although 20 times is not as many as one would like, it was large enough to see differences between each of the null distribution generation methods and complete the simulations within a realistic time frame.

In all three simulations, all four null distribution methods were applied and p-values were computed as described in Section 2.3. Each null distribution of IGPs was made by generating 500 centroid matrices (C^*). For each standard deviation value (either σ_1^2 or η_1^2) and column of Q, the average and standard error of the p-values of the 20 repetitions were computed. They are presented in Figures 5–7.

Version 3 and Version 4 consistently produced p-values while Version 2 did not produce any p-values and Version 1 did not always produce p-values (the second and third panels in the first column of Figure 6).

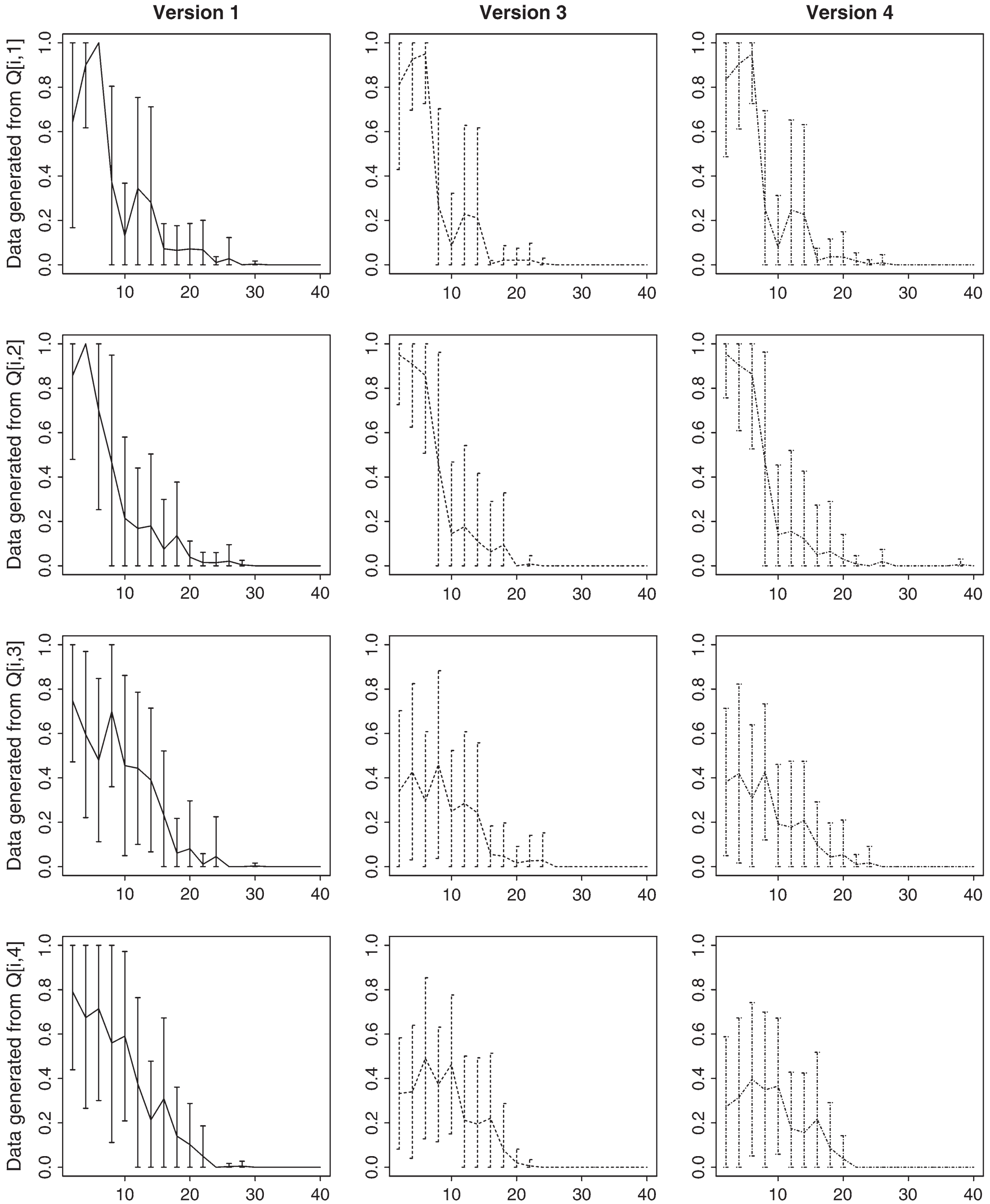

Fig. 5. The results of Simulation 3 are shown in these graphs. The horizontal axis of every graph is the value of σ_1^2 used to generate the elements of Q; the vertical axis of every graph is the p-value. The average p-values are plotted with their corresponding standard error bars. Results from Version 1 are represented by solid lines; results from Version 3, by dashed lines; and results from Version 4, by dotted and dashed lines. Results for data generated from $Q[i, 1]$ are in the first row; results for data generated from $Q[i, 2]$ are in the second row; results for data generated from $Q[i, 3]$ are in the third row; and results for data generated from $Q[i, 4]$ are in the fourth row ($1 \leqslant i \leqslant 500$).

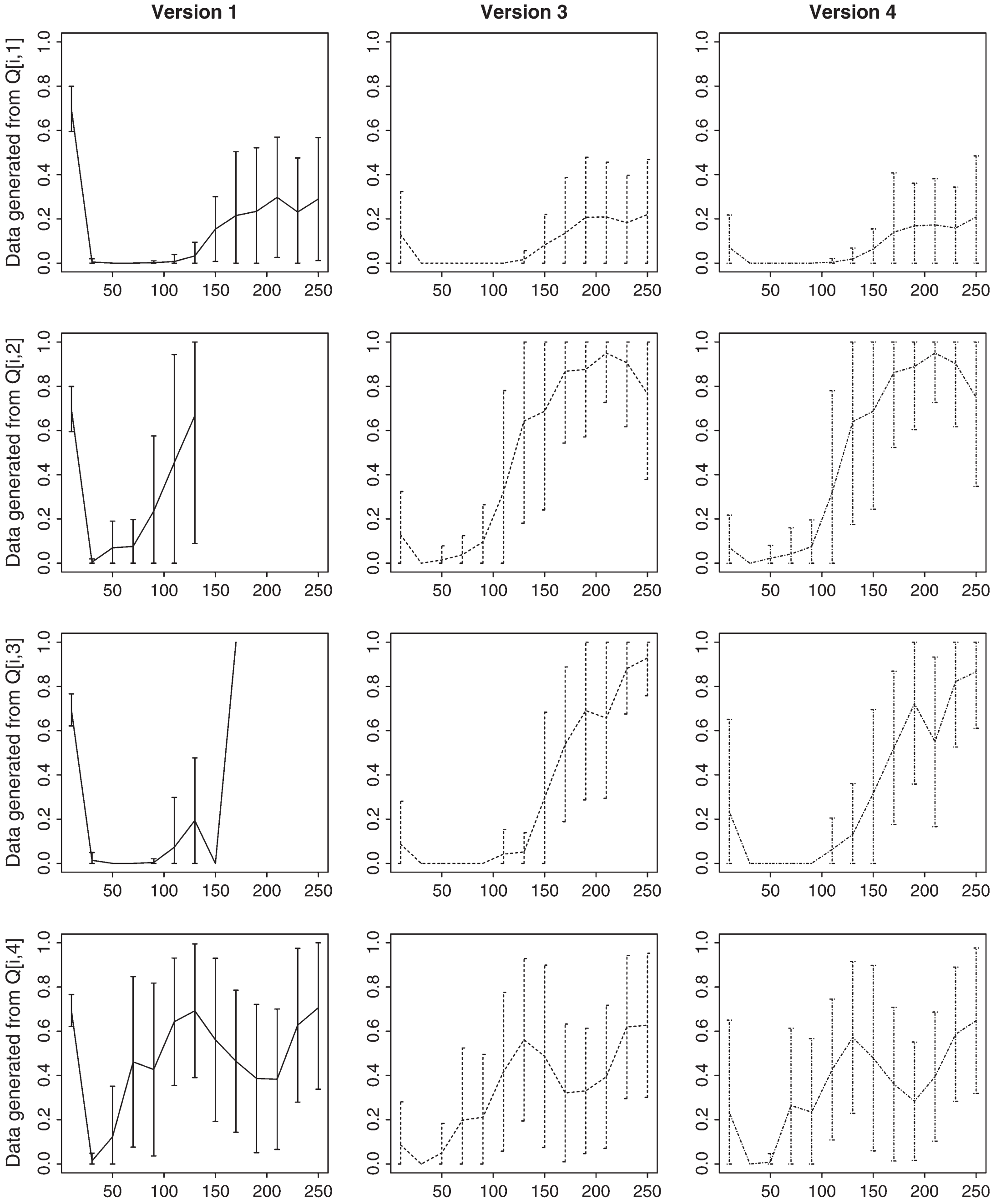

Fig. 6. The results of Simulation 4 are shown in these graphs. The horizontal axis of every graph is the value of η_1^2 used to generate the elements of R; the vertical axis of every graph is the p-value. The average p-values are plotted with their corresponding standard error bars. Results from Version 1 are represented by solid lines; results from Version 3, by dashed lines; and results from Version 4, by dotted and dashed lines. Results for data generated from $Q[i, 1]$ are in the first row; results for data generated from $Q[i, 2]$ are in the second row; results for data generated from $Q[i, 3]$ are in the third row; and results for data generated from $Q[i, 4]$ are in the fourth row ($1 \leqslant i \leqslant 500$).

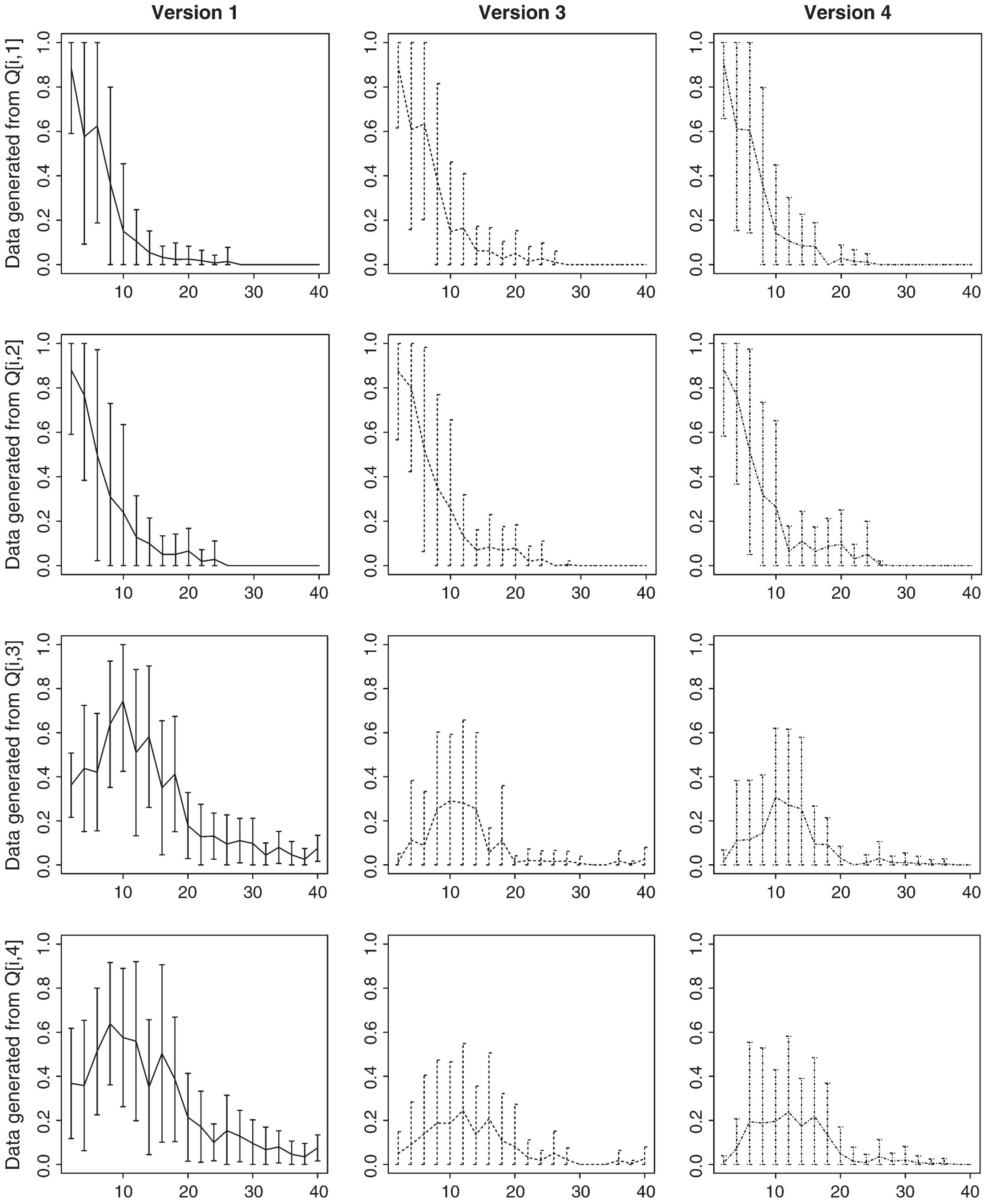

Fig. 7. The results of Simulation 5 are shown in these graphs. The horizontal axis of every graph is the value of σ_1^2 used to generate the elements of Q; the vertical axis of every graph is the p-value. The average p-values are plotted with their corresponding standard error bars. Results from Version 1 are represented by solid lines; results from Version 3, by dashed lines; and results from Version 4, by dotted and dashed lines. Results for data generated from $Q[i, 1]$ are in the first row; results for data generated from $Q[i, 2]$ are in the second row; results for data generated from $Q[i, 3]$ are in the third row; and results for data generated from $Q[i, 4]$ are in the fourth row ($1 \leqslant i \leqslant 500$).

Version 2 did not produce any *p*-values because none of the groups induced by any C^* was the same size as any of the actual groups. This occurs when not all columns of C^* are near the data, e.g. $\text{Class}^*_V(j) = u$ for all j and a single $u \in \{1, 2, 3, 4\}$. Therefore, Version 3 and Version 4 (transform versions) were applicable to more situations than Version 1 and Version 2 (no transform versions). Moreover, Version 3 and Version 4 produced very similar *p*-values. Finally, as expected, in all three null distribution generation versions, more isolated (as σ_1^2 increased in Figures 5 and 7) or more cohesive (as η_1^2 decreased in Figure 6) clusters had lower *p*-values.

As in Simulation 1 IGP results (top panel of Figure 2), Simulation 3 *p*-values for data groups of the same size were similar to each other, but different for data groups of a different size. The first and second rows of Figure 5 are similar to each other but different from the third and fourth rows, and the third and fourth rows of Figure 5 are similar to each other but different from the first and second rows. Specifically, the groups of size 50 had higher *p*-values than the groups of size 100 over $0 < \sigma_1^2 < 10$. This difference in *p*-values between groups of different sizes corresponded to a difference between the IGPs of the smaller groups and the larger groups (Figure 2, top graph). Furthermore, the dependence within the elements of $Q[i, 2]$ and $Q[i, 4]$ for $1 \leqslant i \leqslant 500$ had the most impact when the centroids were not isolated ($0 < \sigma_1^2 < 10$). For $\sigma \geqslant 10$, the curves in Figure 7 greatly resembled those for Figure 5.

The results of Simulation 4 (Figure 6) were similar to the IGP results for Simulation 2 (Figure 3): the curves for the four groups all differed. Even though the relationship between the groups' IGP curves is not the same as the relationship between the groups' *p*-value curves, higher IGPs tend to correspond to lower *p*-values. Therefore, if the third and fourth groups were absent from Simulations 3 and 4, the *p*-values of the first and second groups would probably be lower over the regions where the first and second groups are not very isolated or cohesive. When a group is not very isolated or cohesive, the presence of additional groups will lower the *p*-value of the group.

In these simulations, both R matrices were identically generated which means the number of columns of both R were the same (300 columns). In real situations, however, the sample sizes of two independent datasets are not always the same. To determine the effect sample size had upon *p*-values a final simulation, whose results are not shown, was conducted. Pairs of R matrices were generated where $\sigma_i^2 = 25$ for all i and $\eta_1^2 = 2\eta_2^2 = \eta_3^2 = 2\eta_4^2 = 100$. While the number of columns of the R matrix used to define the centroids remained constant, the number of columns of R to which the centroids were applied varied. The proportions the classes within the latter R matrix remained constant, however.

The number of columns of the second R ranged from 300 to 30. For all four groups and all three null distribution generation versions, the *p*-values fluctuated over this range for number of columns of R. Sometimes the *p*-values were lower for smaller sample sizes; sometimes the *p*-values were higher for smaller sample sizes. Therefore, the relationship between sample size and *p*-values is not easily summarized.

4. APPLICATION TO BREAST CANCER DATA

Although breast cancer is the most common form of cancer to affect women, men are also affected, although in smaller numbers. Like other cancers, the seven stages of breast cancer are based upon the development of the disease (National Cancer Institute, 2004). Stage 0 is the least severe stage and is characterized by the appearance of non-invasive tumors in the breast. In contrast, Stage IV is the most severe and this advanced stage is characterized by breast tumors which have spread to the underarm, internal lymph nodes and beyond (Breastcancer.org, 2004).

In practice, breast cancer stages currently are based entirely upon clinical parameters. The invention of microarrays, however, created the possibility of identifying subtypes using gene expression profiles instead of tumor characteristics. Previous studies (Sørlie *and others*, 2001, 2003; Perou *and others*, 2000)

Table 1. *These are the summary statistics for the sample groups in the van't Veer and others and West and others datasets formed by classification using the Norway/Stanford centroids. The size of each group, average Pearson's (centered) correlation coefficients between the members of each group, and the standard error of the Pearson's (centered) correlation coefficients between the members of each group both with and without the* 0.1 *cutoff are shown*

Sample group	van't Veer *and others* (no cutoff)			van't Veer *and others* (0.1 cutoff)		
	Size	Average	Standard error	Size	Average	Standard error
Normal-like	10	0.3967592	0.08673876	8	0.3964162	0.07539638
ERBB2$^+$	7	0.345514	0.1014053	5	0.3447617	0.09556526
Luminal A	38	0.2997569	0.121031	26	0.2907002	0.1272165
Luminal B	27	0.2972715	0.1013435	18	0.2900479	0.1024578
Basal	35	0.5966506	0.1105372	35	0.5966506	0.1105372
	West *and others* (no cutoff)			West *and others* (0.1 cutoff)		
ERBB2$^+$	3	0.3068818	0.1285781	2	0.4453106	NA
Luminal A	20	0.2301113	0.1243890	18	0.2462426	0.1261459
Luminal B	8	0.3238005	0.128713	8	0.3238005	0.128713
Basal	18	0.3872948	0.1097535	18	0.3872948	0.1097535

have done just that. Sørlie *and others* (2003) analyzed 122 microarrays (115 of which were from malignant breast cancer tissues, seven of which were from non-malignant tissues) and identified five subtypes: two luminal-like (luminal A and luminal B), one ERBB2-overexpressing (ERBB2$^+$), one basal-like (basal), and one normal breast tissue-like (normal-like). The five subtypes were identified in a semi-supervised way: the samples were hierarchically clustered on 534 "intrinsic" genes but the five groups are not identified by cutting the tree at a certain height or specifying the number of groups. As in Sørlie *and others* (2003), we will refer to this dataset as the Norway/Stanford dataset. The Norway/Stanford data are available from the Stanford Microarray Database (http://genome-www.stanford.edu/MicroArray/). A link to them is provided on the webpage: http://genome-www.stanford.edu/breast_cancer/.

Each of the five subtypes was characterized by a centroid made by averaging the gene expression values across all the samples belonging to a subtype. These centroids were then used to classify breast cancer microarrays from independent datasets: van't Veer *and others* (2002) (shared 461 of the 534 intrinsic genes) and West *and others* (2001) (shared 222 of the 534 intrinsic genes). The Pearson's (centered) correlation coefficient was computed for each sample and each centroid. The sample was classified to the group whose centroid had the maximum correlation with the sample. In addition, a correlation cutoff of 0.1 was used, so samples which did not have a correlation of at least 0.1 with any of the centroids were not classified to any subtype (labeled "below-cutoff") (Table 1).

In Sections 4.1 and 4.2, we go a few steps further with the Norway/Stanford, van't Veer *and others* and West *and others* datasets. They are all used to compare the five cluster quality measures and to compare the four cluster validation versions. In addition, these comparisons lead us to validate three of the five subtypes.

4.1 *Comparison of cluster quality measures*

As in Sørlie *and others* (2003), the Norway/Stanford centroids were applied to the van't Veer *and others* and West *and others* datasets to classify the samples. After classifying all the samples in the van't Veer *and others* dataset (no cutoff was used), the five cluster quality measures were computed for each subtype

Table 2. *These are the values of the five cluster quality measures of groups defined by application of Norway/Stanford centroids to van't Veer and others and West and others datasets. The WADP score was computed using* 500 *perturbation matrices whose entries are from a normal distribution (mean was* 0 *and standard deviation was* 1*)*

Dataset	Norway/ Stanford centroid	IGP	HS	SS	SW	WADP score ($\sigma_{\text{WADP}} = 1, n = 500$)
van't Veer *and others*	Normal-like	0.3000000	0.1679855	−0.05489459	0.06801421	0.7398667
	ERBB2$^+$	0.4285714	0.1948864	−0.08013015	−0.07170683	0.7863810
	Luminal A	0.6578947	0.1612159	−0.37187887	0.12308451	0.6734595
	Luminal B	0.4814815	0.1510008	−0.10731030	0.05778799	0.7287350
	Basal	0.9714286	0.5648812	−0.19752388	0.49865359	0.1508403
West *and others*	ERBB2$^+$	0.6666667	−0.06782225	−0.13109415	−0.31311813	0.248000000
	Luminal A	0.7500000	−0.14169784	−0.49751854	0.06339894	0.188600000
	Luminal B	0.8750000	0.06045730	−0.03817666	0.03127546	0.410357143
	Basal	1.0000000	0.07931599	−0.36440829	0.24752699	0.001555556

which contained at least one sample ($\sigma_{\text{WADP}} = 1$ and data perturbed 500 times). The procedure was repeated for the West *and others* dataset. The results for both datasets are presented in Table 2.

For the IGP, HS, and SW, positive value is directly related to cluster quality. For the WADP score, clusters whose scores are closer to zero are of higher quality. For the SS, clusters whose scores are closer to -1 are of higher quality. Using this information to rank the subtypes from highest quality to lowest quality, we saw that the SW and WADP score ranked the West *and others* groups in the same order. In every other case, however, the subtypes were ranked differently by each of the cluster quality measures. Nevertheless, in four of the five cases, the basal subtype had the best score. For the SS, the Luminal A centroid is the most isolated, but the basal-like centroid is the second-most isolated. Therefore, while none of the measures was equivalent, the cohesive and isolated basal-like cluster stood out when using any of the five cluster quality measures, especially when the IGP, HS, SW, or WADP score was used.

4.2 *Comparison of null distribution generation versions*

All four versions of the null distribution generation procedure were applied to the van't Veer *and others* and West *and others* datasets twice, first without a cutoff and then with a cutoff of 0.1 (Tables 7–10). As in Section 2, the group of samples for which $\text{Class}_V(j) = 0$ was called the below-cutoff group.

In other words, the Norway/Stanford microarray data were A and the five centroids made from the dataset comprised C. First, C was used to classify the samples of the van't Veer *and others* dataset ($n = 461$ and $q = 117$), then C was used to classify the West *and others* dataset ($n = 222$ and $q = 49$). In both cases, no cutoff was used ($c = 0$) and a 0.1 cutoff was used ($c = 0.1$).

The minimum number of permutations used to generate the null distributions was 2500; the maximum was 250 000. The number of permutations was chosen so that at least 100 permutations would be used to compose the null distributions for each group. (NB recall that the null distributions depend upon the size of the group.) In only one case were fewer than 100 permutations used: the basal group when the van't Veer *and others* raw data were permuted (Version 2) with 0.1 cutoff. Out of 50 500 permutations, a group of size 35 only occurred three times. At this rate, over one million permutations would have been necessary to get 100 IGPs for groups of size 35.

Table 3. *Each entry is the number of samples classified to the row subtype and whose nearest neighbors were classified to the column subtype. The right-hand most column is the proportion of samples in the row subtype whose nearest neighbors were also classified to the same subtype*

van't Veer *and others* (no cutoff)		Nearest-neighbor classification					IGPs
		Normal-like	ERBB2$^+$	Luminal A	Luminal B	Basal	
Sample classification	Normal-like	3	5	2	0	0	0.300
	ERBB2$^+$	1	3	0	3	0	0.429
	Luminal A	6	2	25	5	0	0.658
	Luminal B	2	4	7	13	1	0.875
	Basal	0	1	0	0	34	0.971

Table 4. *Each entry is the number of samples assigned to the row group and whose nearest neighbors were assigned to the column group. The right-hand most column is the proportion of samples in the row subtype whose nearest neighbors were also classified to the same subtype*

van't Veer *and others* (0.1 cutoff)		Nearest-neighbor classification						IGPs
		Normal-like	ERBB2$^+$	Luminal A	Luminal B	Basal	Below-cutoff	
Sample classification	Normal-like	2	0	1	0	0	5	0.250
	ERBB2$^+$	0	1	0	3	0	1	0.200
	Luminal A	2	0	18	1	0	5	0.692
	Luminal B	1	2	0	6	1	8	0.333
	Basal	0	1	0	0	34	0	0.971
	Below-cutoff	3	0	0	1	0	21	0.840

When the Norway/Stanford centroids were applied to the van't Veer *and others* and West *and others* datasets both with and without a 0.1 cutoff, the basal-like subtype had the highest IGPs. In the West *and others* dataset, no samples were classified to the normal breast tissue-like subtype. In the van't Veer *and others* dataset, this subtype had the lowest IGPs. The IGPs for the ERBB2$^+$, luminal A, and luminal B subtypes varied. For these three subtypes, however, at least one IGP was above 0.75.

The IGPs for the samples not classified to any group when a cutoff was used were 0 and 0.84 for the West *and others* and van't Veer *and others* datasets, respectively. More than 20 samples were not classified in the van't Veer *and others* dataset when the cutoff was used.

The groups were much more cohesive in the West *and others* dataset than in the van't Veer *and others* dataset. In the West *and others* dataset, the nearest neighbor of a sample classified to one subtype was classified to only two subtypes if it was classified at all. In contrast, in the van't Veer *and others* dataset, the nearest neighbor could have been classified to any subtype. The one exception was the normal breast tissue-like subtype. No normal breast tissue-like samples had a nearest neighbor classified to the basal-like subtype. The reverse also held true (Tables 3–6).

In every case except for that in which the null distribution was generated by permuting the centroids with a cutoff of 0.1, the estimated p-value for the basal-like subtype was less than 0.05. The estimated p-values for the luminal B subtype were all below 0.05 in the West *and others* dataset. In addition, the ERBB2$^+$ subtype's p-values were below 0.05 in every case using the West *and others* data and the cutoff. When a cutoff was not used when the Norway/Stanford centroids were applied to the West *and others* data, the ERBB2$^+$ p-values were below 0.05 in two cases and below 0.10 in the other two cases. Only the

Table 5. *Each entry is the number of samples classified to the row subtype and whose nearest neighbors were classified to the column subtype. The right-hand most column is the proportion of samples in the row subtype whose nearest neighbors were also classified to the same subtype*

West *and others* (no cutoff)		Nearest-neighbor classification					IGPs
		Normal-like	ERBB2$^+$	Luminal A	Luminal B	Basal	
Sample classification	Normal-like	0	0	0	0	0	∞
	ERBB2$^+$	0	2	1	0	0	0.667
	Luminal A	0	0	15	4	1	0.750
	Luminal B	0	0	1	7	0	0.875
	Basal	0	0	0	0	18	1.00

Table 6. *Each entry is the number of samples assigned to the row group and whose nearest neighbors were assigned to the column group. The right-hand most column is the proportion of samples in the row subtype whose nearest neighbors were also classified to the same subtype*

West *and others* (0.1 cutoff)		Nearest-neighbor classification						IGPs
		Normal -like	ERBB2$^+$	Luminal A	Luminal B	Basal	Below -cutoff	
Sample classification	Normal-like	0	0	0	0	0	0	∞
	ERBB2$^+$	0	2	0	0	0	0	1.00
	Luminal A	0	0	13	4	0	1	0.722
	Luminal B	0	0	1	7	0	0	0.875
	Basal	0	0	0	0	18	0	1.00
	Below-cutoff	0	0	2	0	1	0	0

basal-like subtype was validated at the $\alpha = 0.05$ level by any version of the null distribution generation procedure applied to the van't Veer *and others* data.

Although the estimated p-values varied widely across the datasets and across versions of the null distribution generation procedure, three trends were evident. First, the p-values were higher when the 0.1 cutoff was used because the van't Veer *and others* and West *and others* samples were not close to the centroids. When compared to the null distributions made without a cutoff, the null distributions made with the cutoff were skewed toward 1.0 (Figure 8). Second, the versions which applied the transformation before permuting yielded smaller p-values than the versions that did not apply the transformation and just permuted. Third, the p-values obtained from a centroid version were very similar to the p-values obtained from the equivalent version using the raw data.

5. DISCUSSION

Although the IGP, HS, SS, SW, and WADP score all measure cluster quality, they are not equivalent. In Simulation 1 (Figure 2), the average HS for the true classifications was unaffected by the change in correlation between the columns of Q. Therefore, the HS does not capture any information about the isolation of the centroids. On the other hand, the SS only captured information about the correlation between centroids. In Simulation 2 (Figure 3), the average SS for the true classifications was unaffected by the change in correlation between members of a groups and their centroids. In addition, in Simulation 2, two of the SW true classification curves increased while two of the SW true classification curves decreased between $50 < \eta_1^2 < 250$. Therefore, the HS, SS, and SW are poor choices for a cluster quality measure.

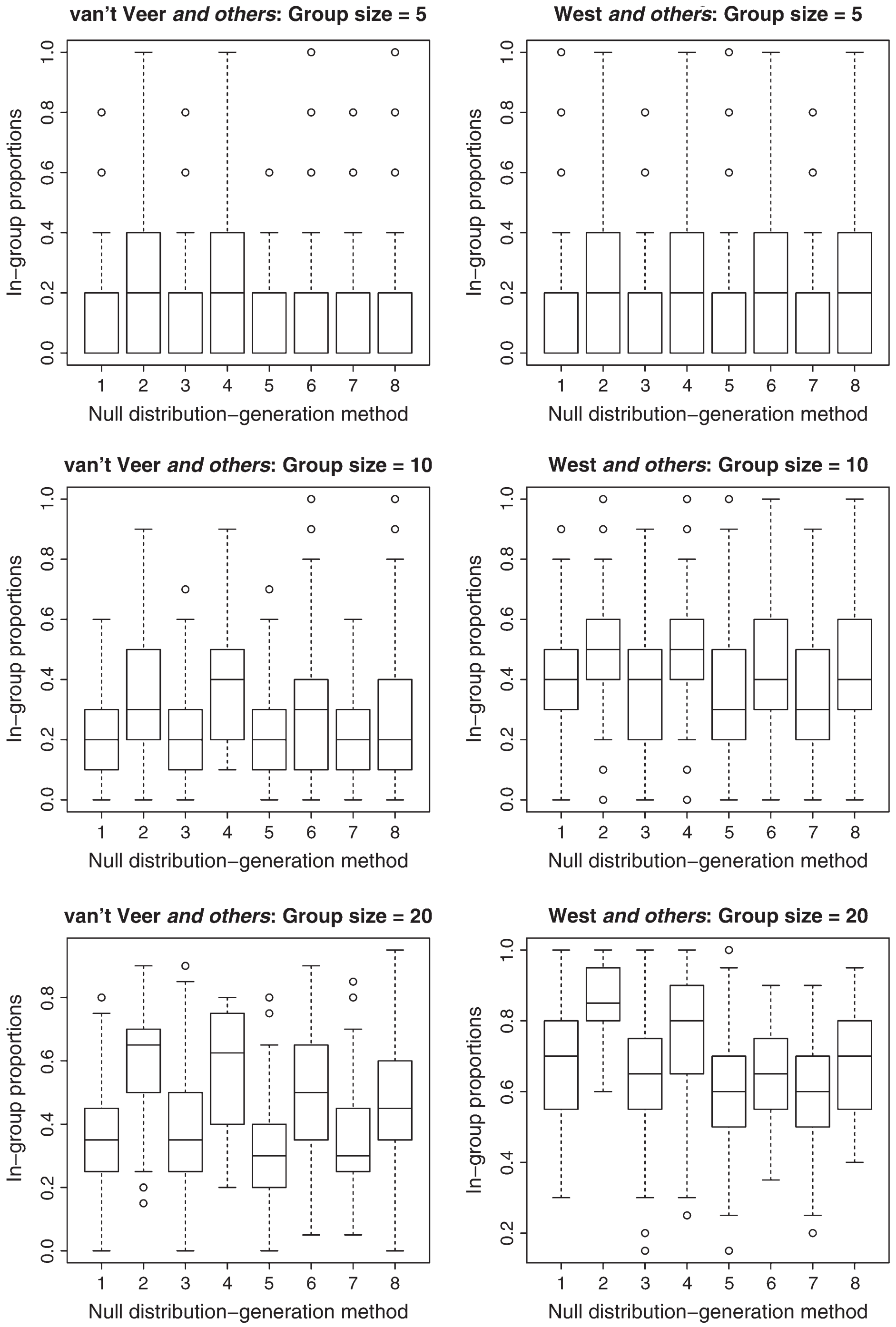
van't Veer *and others*: Group size = 5
West *and others*: Group size = 5
van't Veer *and others*: Group size = 10
West *and others*: Group size = 10
van't Veer *and others*: Group size = 20
West *and others*: Group size = 20
In−group proportions
Null distribution−generation method
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8

Table 7. *Estimated p-values for van't Veer and others data by null distribution generation version (no cutoff). The size of the null distribution is also given (n)*

van't Veer *and others* (no cutoff)	IGPs	Version 1: permute centroids		Version 2: permute raw data		Version 3: transform centroids		Version 4: transform raw data	
		p-value	*n*	*p*-value	*n*	*p*-value	*n*	*p*-value	*n*
Normal-like	0.300	0.3313	335	0.3860	329	0.3232	362	0.2779	367
ERBB2$^+$	0.429	0.1108	352	0.1534	365	0.0811	333	0.1095	338
Luminal A	0.658	0.3648	159	0.3519	108	0.2500	132	0.2340	141
Luminal B	0.481	0.5258	202	0.4684	190	0.3824	238	0.3333	207
Basal	0.971	0.0221	181	0.0070	142	0.0000	157	0.0000	155

Table 8. *Estimated p-values for van't Veer and others data by null distribution generation version (*0.1 *cutoff). The size of the null distribution is also given (n)*

van't Veer *and others* (0.1 cutoff)	IGPs	Version 1: permute centroids		Version 2: permute raw data		Version 3: transform centroids		Version 4: transform raw data	
		p-value	*n*	*p*-value	*n*	*p*-value	*n*	*p*-value	*n*
Normal-like	0.250	0.6379	486	0.6804	1527	0.5151	13137	0.5029	3426
ERBB2$^+$	0.200	0.5196	766	0.5219	3545	0.4508	32089	0.4583	6118
Luminal A	0.692	0.7288	177	0.7843	102	0.3705	502	0.2776	425
Luminal B	0.333	0.9264	231	0.9010	293	0.7814	1830	0.7296	1028
Basal	0.971	0.4122	131	0.0000	3	0.0265	113	0.0099	203
Below-cutoff	0.840	0.9638	4000	0.9844	50500	0.9137	250000	0.4680	25000

In both Simulation 1 and Simulation 2, the WADP score true classification curves differed greatly from the estimated classification curves for some values of σ_1^2 or η_1^2. This in addition to the requirement that we choose a value for σ_{WADP} prevents the WADP score from being a good cluster quality measure.

Consequently, the IGP is the best choice for a cluster quality measure. First, the true classification curves for all groups increased as σ_1^2 increased and η_1^2 decreased. Second, the true classification curves are close to the estimated classification curves. Third, the IGP did not require us to choose a parameter value.

Nevertheless, when all five cluster quality measures were applied to the breast cancer datasets, the basal subtype was given the best score by four of the five measures (Table 2). In both datasets, the basal subtype had the highest IGP, HS, and SW and had the lowest WADP score. The differences between these four scores appeared when a cluster was not very cohesive or isolated. Thus, if all the clusters in the datasets are of high quality, which of the four is used in the validation procedure may not matter.

Fig. 8. These are boxplots of the IGPs for eight different null distribution generation methods. Each version of the null distribution generation procedure was twice applied (with and without cutoff) to both datasets for three different group sizes (5, 10, and 20). Each even-numbered method and its odd-numbered neighbor to the left are identical except that the even-numbered method uses the 0.1 cutoff and the odd-numbered method does not use a cutoff. All the even-numbered plots are skewed toward 1.0 when compared to its left-hand neighbor. Null distribution generation method labels: (1) permute centroids without cutoff, (2) permute centroids with 0.1 cutoff, (3) permute raw data without cutoff, (4) permute raw data with 0.1 cutoff, (5) transform centroids without cutoff, (6) transform centroids with 0.1 cutoff, (7) transform raw data without cutoff, and (8) transform raw data with 0.1 cutoff.

Table 9. *Estimated p-values for West and others data by null distribution generation version (no cutoff). The size of the null distribution is also given (n)*

van’t Veer *and others* (no cutoff)	IGPs	Version 1: permute centroids		Version 2: permute raw data		Version 3: transform centroids		Version 4: transform raw data	
		p-value	*n*	*p*-value	*n*	*p*-value	*n*	*p*-value	*n*
ERBB2^+	0.667	0.0670	806	0.0426	798	0.0435	620	0.0527	683
Luminal A	0.750	0.4185	227	0.2775	227	0.2020	203	0.2000	180
Luminal B	0.875	0.0058	685	0.0085	704	0.0047	852	0.0045	892
Basal	1.000	0.0162	309	0.0034	292	0.0000	276	0.0000	271

Table 10. *Estimated p-values for West and others data by null distribution generation version (*0.1 *cutoff). The size of the null distribution is also given (n)*

van’t Veer *and others* (no cutoff)	IGPs	Version 1: permute centroids		Version 2: permute raw data		Version 3: transform centroids		Version 4: transform raw data	
		p-value	*n*	*p*-value	*n*	*p*-value	*n*	*p*-value	*n*
ERBB2^+	1.000	0.0181	1380	0.0105	17209	0.0317	19711	0.0202	3166
Luminal A	0.722	0.9191	136	0.6806	144	0.4259	108	0.4514	144
Luminal B	0.875	0.0240	458	0.0248	2297	0.0207	3000	0.0229	1003
Basal	1.000	0.1324	136	0.0139	144	0.0093	108	0.0278	144
Below-cutoff	0.000	1.0000	2500	1.0000	25000	1.0000	25000	1.0000	5000

Since the quality of the clusters may not be known beforehand, the IGP was used to compare the four versions of the null distribution generation procedure. Simulations 3–5 (Figures 5–7) showed when a null distribution generation version produced *p*-values, they were lower for more isolated or more cohesive clusters. In all three simulations, Version 2 did not yield any *p*-values and Version 1 tended to produce more conservative *p*-values than Versions 3 and 4. Also, the *p*-values from Version 3 and Version 4 were very similar. In addition, dependence between the rows of two of the columns of the centroid matrix (C) had the most impact when the clusters were not isolated ($0 < \sigma_1^2 < 10$). When the rows of C were not completely independent and the clusters were not isolated, the average *p*-value for a cluster was lower than the average *p*-value for the same cluster when all the rows of C were independent, even for the columns of C that were generated identically in Simulation 3 and Simulation 5. The difference in the *p*-values between Simulation 3 and Simulation 5 was most dramatic for the clusters of larger size.

Similar conclusions were seen when the null distribution generation versions were compared using real data (Tables 7–10). Not only were the *p*-values produced by Version 3 and Version 4 very similar for the breast cancer data but also were the *p*-values produced by Version 1 and Version 2. In other words, a *p*-value from a raw data version was close to the *p*-value from the corresponding centroid version. Furthermore, as was seen when Version 1 and Version 3 *p*-values were compared in the simulations, a *p*-value from a version that did not use a transformation was more conservative than the *p*-value from the corresponding transformation version. Moreover, the *p*-values from the versions using a 0.1 cutoff were more conservative than the versions not using a cutoff (Figure 8). Most likely, this is due to the IGP of each group without a cutoff being as high or higher than the IGP of the group with the 0.1 cutoff in all but one case (West *and others* ERBB2^+). Based upon these results, a cutoff should not be used because it may reduce the quality of a cluster when quality is measured by the IGP.

Of the cluster quality measures considered, the IGP was the best at quantifying how likely a point was to be assigned to a different cluster. Of the null distribution generation versions, Versions 3 and 4 most reliably generated p-values. Although the two versions generated very similar p-values, Version 3 is superior to Version 4. Version 3 takes less time to implement, does not use raw data which may be unavailable, and does not require the user to make a choice about which hierarchical clustering method to use.

Therefore, using the IGP with the transform centroids version of the null distribution generation procedure without a cutoff (Version 3, $c = 0$, and $\alpha = 0.05$), Section 4 shows that only the ERBB2^{+1}, luminal B, and basal-like groups are reproducible and potentially biologically significant.

An implementation of null distribution generation Version 3 without a cutoff is available on-line through CRAN (http://cran.r-project.org) in the clusterRepro package.

As this breast cancer application demonstrates, the cluster validation method proposed here has the potential to be very useful. For a cluster found in datasets independent of the one in which it was defined, we believe this method (using the IGP and null distribution generation Version 3 without a cutoff) reliabily and efficiently gauges the significance of the cluster's reproducibility.

ACKNOWLEDGMENTS

Robert Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

REFERENCES

BAILEY, T. A. AND DUBES, R. (1982). Cluster validity profiles. *Pattern Recognition* **15**, 61–83.

BREASTCANCER.ORG. (2004). *Stages of Breast Cancer*. http://www.breastcancer.org/cmn_sta_idx.html.

CHEN, G., JARADAT, S. A., BANERJEE, N., TANAKA, T. S., KO, M. S. H. AND ZHANG, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica* **12**, 241–62.

DATTA, S. AND DATTA, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**, 459–66.

DUDOIT, S. AND FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**, research0036.1–21.

DUDOIT, S., FRIDLYAND, J. AND SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

GORDON, A. D. (1999). *Classification*. Boca Raton, FL: Chapman & Hall.

KERR, M. K. AND CHURCHILL, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the Unites States of America* **98**, 8961–5.

LEVINE, E. AND DOMANY, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation* **13**, 2573–93.

NATIONAL CANCER INSTITUTE (2004). *Staging: Questions and Answers*. http://www.cancer.gov/cancertopics/factsheet/Detection/staging.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A. *and others* (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747–52.

[1] ERBB2^{+} was validated on the West *and others* dataset and only two or three samples were classified to this subtype so this result is somewhat suspect.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S. *and others* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–74.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S. *and others* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8418–23.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*. **14**, 511–28.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. *and others* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–6.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olsen, Jr, J. A., Marks, J. R. and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11462–7.

Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics* **17**, 309–18.

[*Received July 28, 2005; revised February 28, 2006; accepted for publication March 8, 2006*]