Testing R-Port of Q5 Activity Logs

Testers: Ranjani Ramakrishan (ramakris@ohsu.edu) and Ted Laderas (laderast@ohsu.edu)

Test ID R1

Objective: Compare outputs of Matlab version of Q5 to R version of Q5 algorithm under similar conditions.

Procedure:

1) Run Matlab version of Q5 with one hundred replicates on synthetic dataset
2) Run R version of Q5 with one hundred replicates on same dataset.
3) Report average and standard deviations of statistics.

Results: FAIL (version 2.0). PASS (version 3.0)

This comparison between Matlab Q5 and the R-port of Q5 was done on two separate problems: group 1 versus group 4 (table R1.1) and group 3 versus group 4 (Table R1.3) of the synthetic data set. The data was interpolated using spline-based interpolation and any negative values were winsorized to zero.

The Matlab version performs consistently better across the different probability classification thresholds (PCT) than the R Port Q5 version 2.0. As information about the PCT is not available for the R Port, we compare the results for PCT = 0.5, for illustrative purposes (Table R1.2 and R1.4.). After these results were made available to Dartmouth, an implementation of Q5 (3.0) closer to the original Matlab implementation was provided. This version (3.0) was again tested and performance was comparable to the Matlab implementation and that of a Support Vector Machine classifier.

Complete results (Tables R1.2 and R1.4) are included. Some differences that we noted between the R and Matlab implementations are:

1) The R Port (version 2.0) does not exclude any eigenvectors after the PCA step, whereas the number of eigenvectors dropped is n-k for the Matlab version (where n is the number of samples and k is the number of classes). Version 3.0 has changed the PCA implementation to closely mimic the Matlab implementation.
2) The Matlab version provides functionality for three class classification whereas this is not implemented in the R port. The RPort could be generalized for any number of classes, which would increase its utility.
3) The Matlab implementation classifies using a prespecified probability threshold; the R Port does not.

Table R1.1: Comparison of group 1 versus group 4

Group	P1	C1	C2	C3	Sample Size
1	Absent	Absent	0.5x	1x	49
4	1x	2x	4x	1x	51

Table R1.2 Classification metrics for Matlab and R Port of Q5 for group 1 versus group 4 (at PCT = 0.5 and 50:50 test :train split)

Interpolated Data, negative values winsorized to zero				
	PPV	**%CC**	**Sens**	**Spec**
Matlab Q5	0.9843 (0.0297)	0.9915 (0.0160)	1.0000 (0.0000)	0.9831 (0.0320)
R Q5 v2.0	0.7197 (0.1330)	0.7200 (0.1366)	0.7300 (0.1815)	0.7100 (0.1569)
R Q5 v3.0	0.9764 (0.0322)	0.9873 (0.0174)	1.0000 (0.0000)	0.9746 (0.0347)
SVM	0.9749 (0.0386)	0.9504 (0.0395)	0.9262 (0.0699)	0.9746 (0.0395)

Table R1.3: Comparison of Group 3 versus Group 4.

Group	P1	C1	C2	C3	Sample Size
3	1x	1x	2x	1x	50
4	1x	2x	4x	1x	51

Table R1.4 Classification metrics for Matlab and R Port of Q5 for group 3 versus group 4 (at PCT = 0.5 and 50:50 test :train split)

Group 3 versus Group 4 (hardest problem)				
Interpolated Data, negative values not winsorized to zero				
	PPV	**%CC**	**Sens**	**Spec**
Matlab Q5	0.9836 (0.0302)	0.9912 (0.0163)	1.0000 (0.0000)	0.9696 (0.0264)
R Q5 v2.0	0.7286 (0.1091)	0.7127 (0.0872)	0.7100 (0.1320)	0.7154 (0.1525)
R Q5 v3.0	0.9807 (0.0319)	0.9896 (0.0172)	1.0000 (0.0000)	0.9792 (0.0343)
SVM	0.8818 (0.0875)	0.8677 (0.0684)	0.8585 (0.0945)	0.8769 (0.0972)

Test ID R2

Objective: Compare effectiveness of Q5 algorithm to differentiate between two groups of different effect sizes of synthetic data.

Procedure:

1) Load dataset into memory (using groups with different effect sizes – group 3 and group 4)
2) Produce 100 random partitions of data for three sets of train/test splits: 50%/50%, 75%/25%, 90%/10%.
3) Run Q5 algorithm on 100 random partitions produced above.
4) Report average and standard deviations of statistics.

Results: PASS.

These results are for Q5 version 3.0.

Even at the 50/50 train/test split, Q5 classifies the data with a high positive predictive value. Full results are presented in Table R2.1

Table R2.1 Results of Classification of two groups with different effect sizes. Standard deviations are given in parentheses.

Train/Test Ratio	PPV	% Correctly Classified	Sensitivity	Specificity
0.5	0.9792 (0.0193)	0.9892 (0.0098)	0.9996 (0.0040)	0.9792 (0.0193)
0.75	0.9800 (0.0338)	0.9892 (0.0182)	1.0000 (0.0000)	0.9785 (0.0364)
0.9	0.9783 (0.0563)	0.9882 (0.0307)	1.0000 (0.0000)	0.9783 (0.0563

Test ID R3

Objective: Compare effectiveness of Q5 algorithm to differentiate between two groups of that contain different sets of protein mixtures.

Procedure:

1) Load dataset into memory (using groups that contain different protein mixtures)
2) Produce 100 random partitions of data for three sets of train/test splits: 50%/50%, 75%/25%, 90%/10%.
3) Run Q5 algorithm on random partitions produced above step 2.
4) Report average and standard deviations of statistics.

Results: PASS.

These results are for Q5 version 3.0.

Even at the 50/50 train/test split, Q5 classifies the data with high sensitivity, high specificity, and high positive predictive value. Full results are presented in table R3.1.

Table R3.1 Results of Classification of samples with different sets of protein mixtures

Train/Test Ratio	PPV	% Correctly Classified	Sensitivity	Specificity
0.5	0.9773 (0.0190)	0.9884 (0.0097)	1.0000 (0.0000)	0.9773 (0.0190)
0.75	0.9764 (0.0322)	0.9873 (0.0174)	1.0000 (0.0000)	0.9746 (0.0347)
0.9	0.9850 (0.0479)	0.9918 (0.0261)	1.0000 (0.0000)	0.9850 (0.0479)