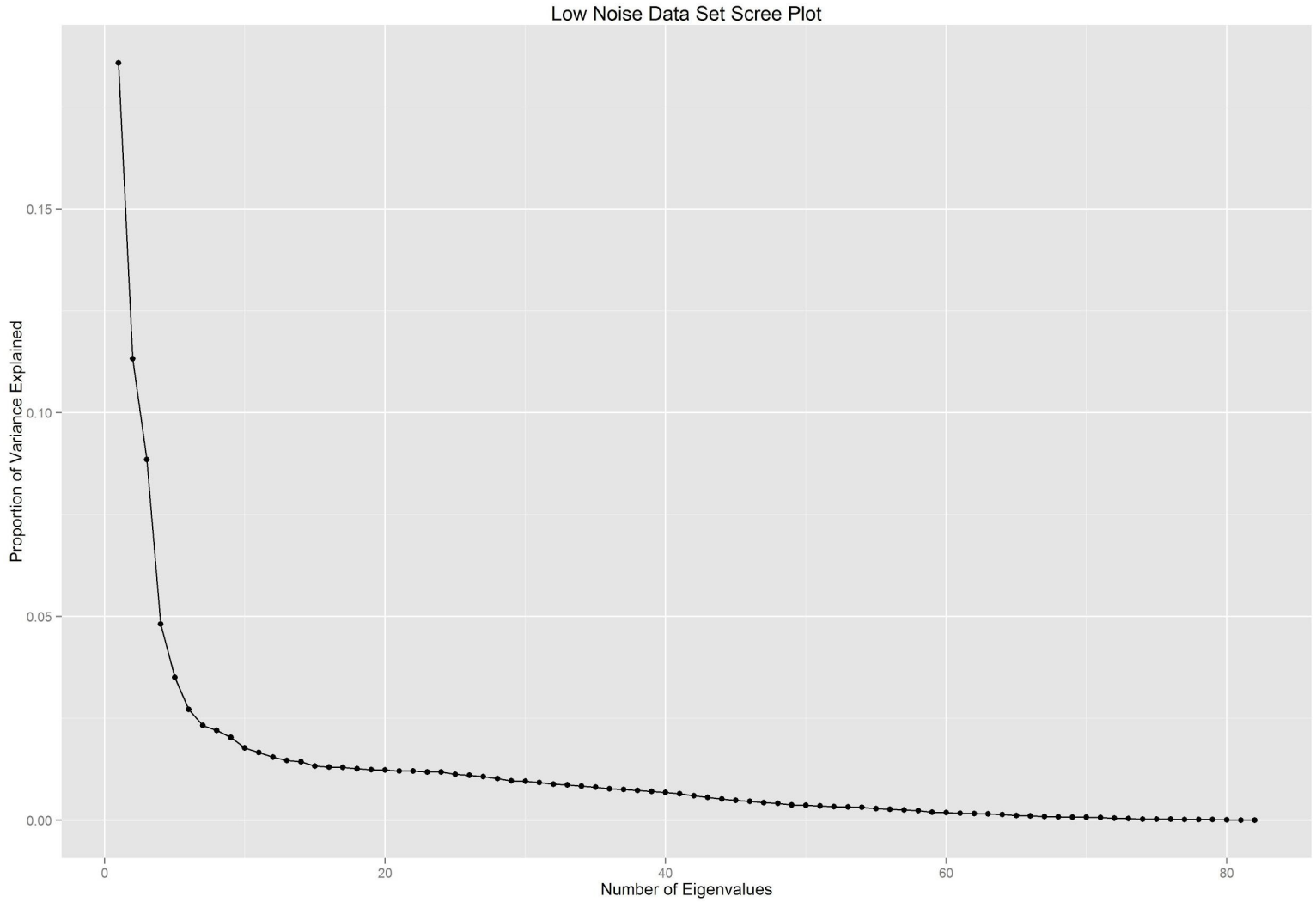
Clustering Astronomy Data

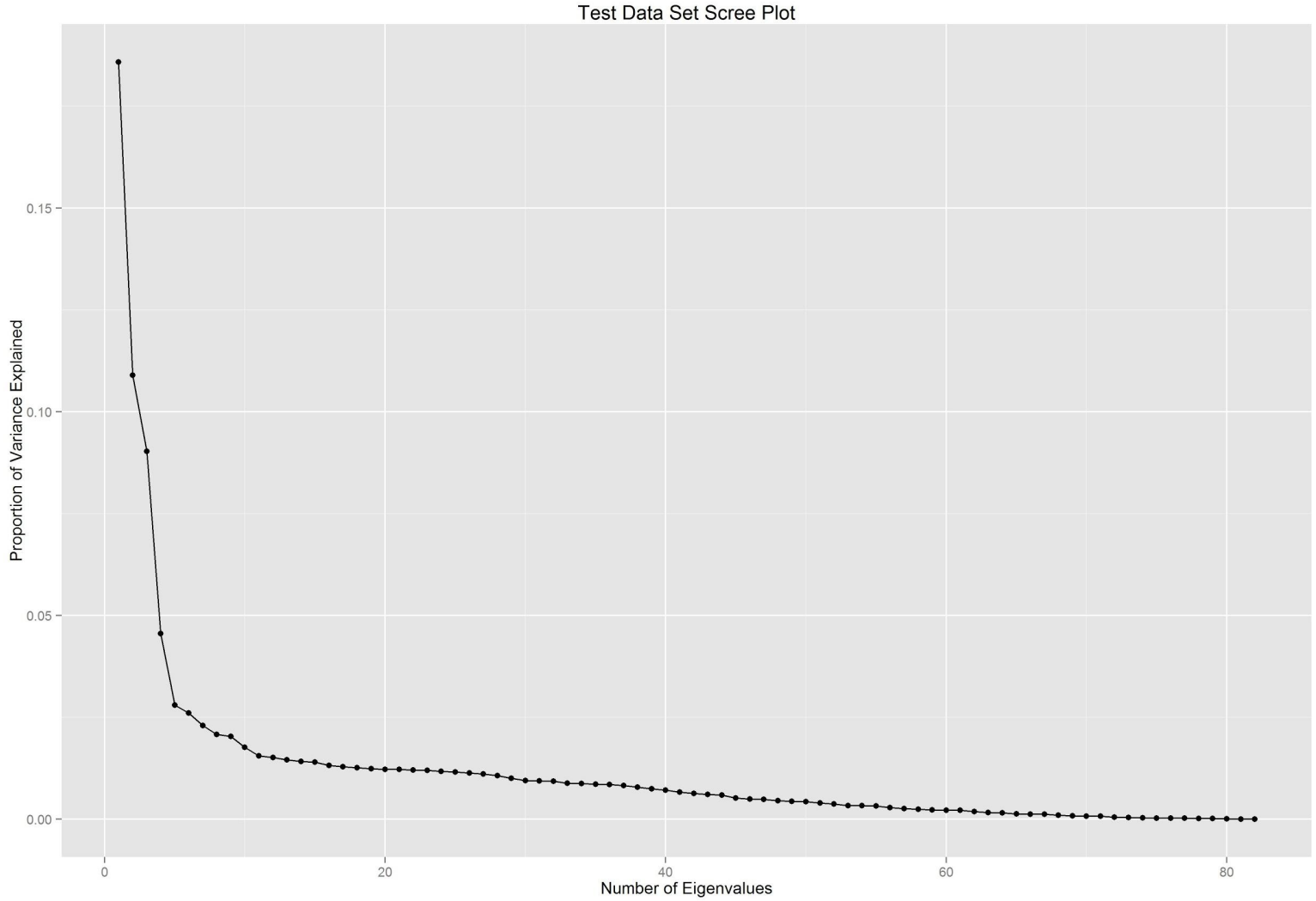
Katherine Eng, Jie Hu, Vanessa Lepe, Kalbi Zongo

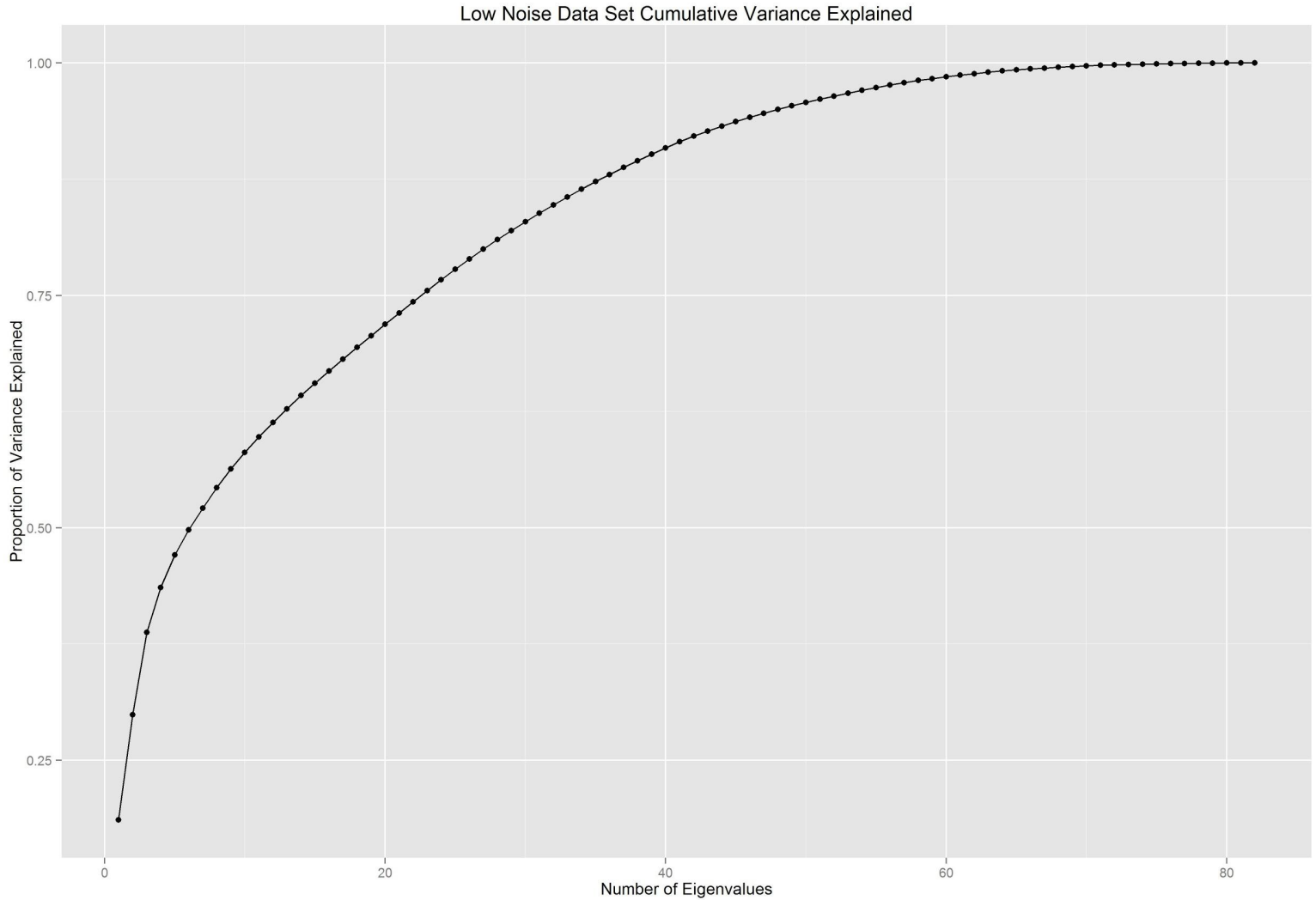
QUESTIONS OF INTEREST

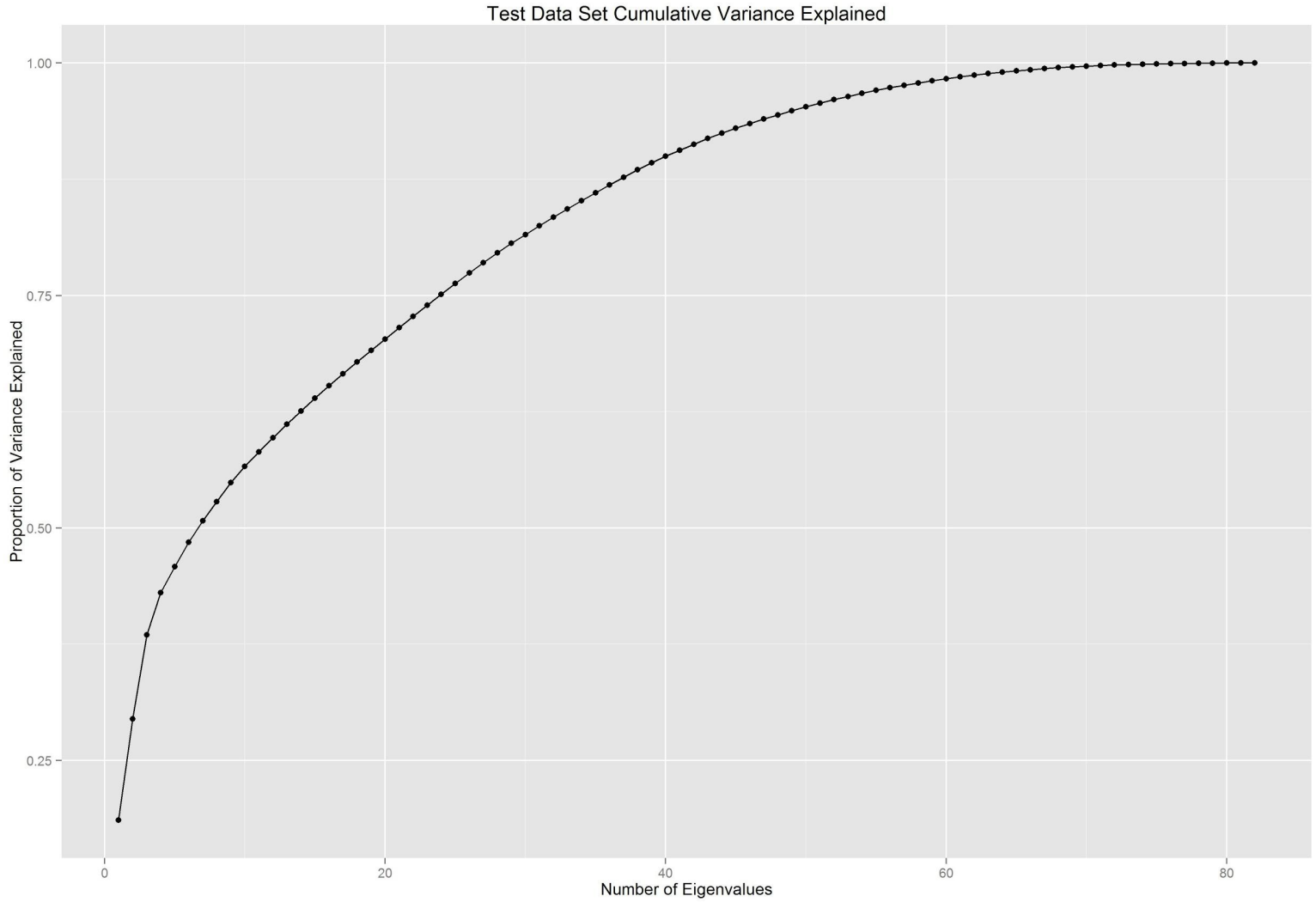
1. How does the principal component analysis (PCA) for low noise data compare to the PCA for the full test data set?
2. How do clustering methods (k-means vs hierarchical) differ in the low noise data set?
3. How do clustering methods (k-means vs hierarchical) differ in the test data set?

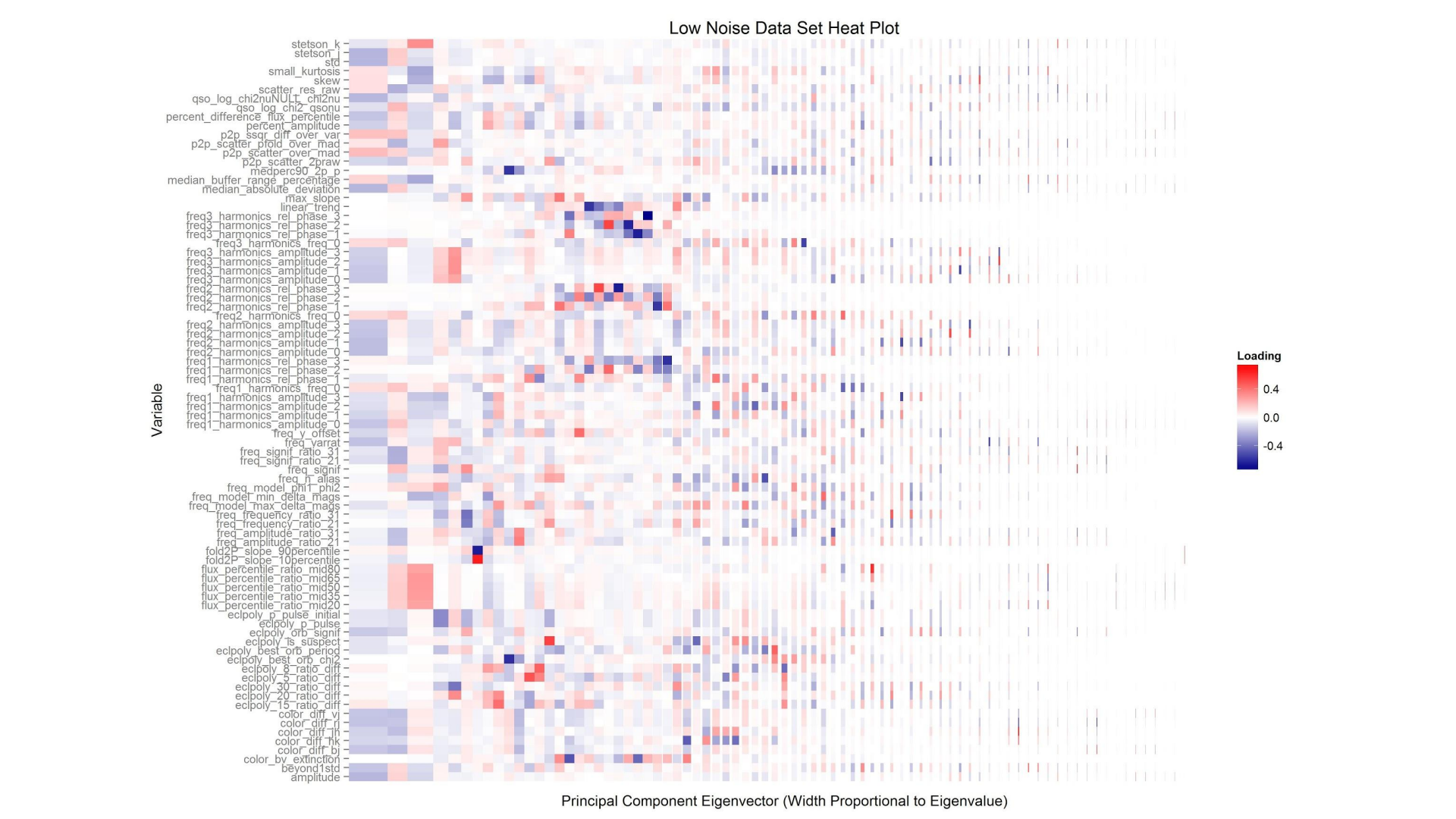
RESULTS


Low Noise Data Set Scree Plot
Proportion of Variance Explained
0.15
0.10
0.05
0.00
0
20
40
60
80
Number of Eigenvalues


Test Data Set Scree Plot
Proportion of Variance Explained
0.15
0.10
0.05
0.00
0
20
40
60
80
Number of Eigenvalues


Low Noise Data Set Cumulative Variance Explained
Proportion of Variance Explained
1.00
0.75
0.50
0.25
0
20
40
60
80
Number of Eigenvalues


Test Data Set Cumulative Variance Explained
Proportion of Variance Explained
1.00
0.75
0.50
0.25
0
20
40
60
80
Number of Eigenvalues

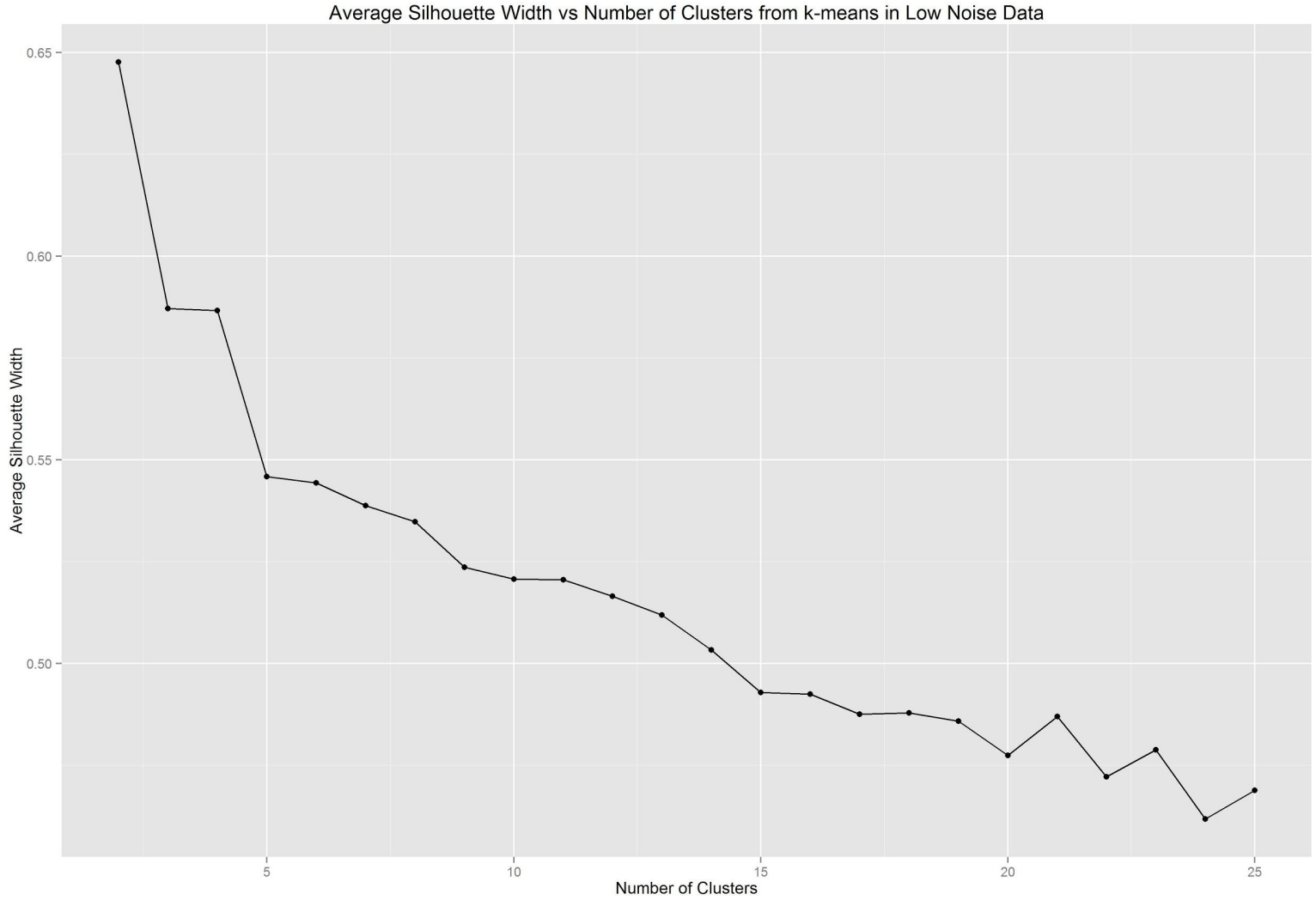

Low Noise Data Set Heat Plot
stetson_k
stetson_j
std
small_kurtosis
skew
scatter_res_raw
qso_log_chi2nuNULL_chi2nu
qso_log_chi2_qsonu
percent_difference_flux_percentile
percent_amplitude
p2p_ssqr_diff_over_var
p2p_scatter_pfold_over_mad
p2p_scatter_over_mad
p2p_scatter_2praw
medperc90_2p_p
median_buffer_range_percentage
median_absolute_deviation
max_slope
linear_trend
freq3_harmonics_rel_phase_3
freq3_harmonics_rel_phase_2
freq3_harmonics_rel_phase_1
freq3_harmonics_freq_0
freq3_harmonics_amplitude_3
freq3_harmonics_amplitude_2
freq3_harmonics_amplitude_1
freq3_harmonics_amplitude_0
freq2_harmonics_rel_phase_3
freq2_harmonics_rel_phase_2
freq2_harmonics_rel_phase_1
freq2_harmonics_freq_0
freq2_harmonics_amplitude_3
freq2_harmonics_amplitude_2
freq2_harmonics_amplitude_1
freq2_harmonics_amplitude_0
freq1_harmonics_rel_phase_3
freq1_harmonics_rel_phase_2
freq1_harmonics_rel_phase_1
freq1_harmonics_freq_0
freq1_harmonics_amplitude_3
freq1_harmonics_amplitude_2
freq1_harmonics_amplitude_1
freq1_harmonics_amplitude_0
freq_y_offset
freq_varrat
freq_signif_ratio_31
freq_signif_ratio_21
freq_signif
freq_n_alias
freq_model_phi1_phi2
freq_model_min_delta_mags
freq_model_max_delta_mags
freq_frequency_ratio_31
freq_frequency_ratio_21
freq_amplitude_ratio_31
freq_amplitude_ratio_21
fold2P_slope_90percentile
fold2P_slope_10percentile
flux_percentile_ratio_mid80
flux_percentile_ratio_mid65
flux_percentile_ratio_mid50
flux_percentile_ratio_mid35
flux_percentile_ratio_mid20
eclpoly_p_pulse_initial
eclpoly_p_pulse
eclpoly_orb_signif
eclpoly_ls_suspect
eclpoly_best_orb_period
eclpoly_best_orb_chi2
eclpoly_8_ratio_diff
eclpoly_5_ratio_diff
eclpoly_30_ratio_diff
eclpoly_20_ratio_diff
eclpoly_15_ratio_diff
color_diff_vj
color_diff_rj
color_diff_ijh
color_diff_hk
color_diff_bj
color_by_extinction
beyond1std
amplitude
Variable
Loading
0.4
0.0
-0.4
Principal Component Eigenvector (Width Proportional to Eigenvalue)

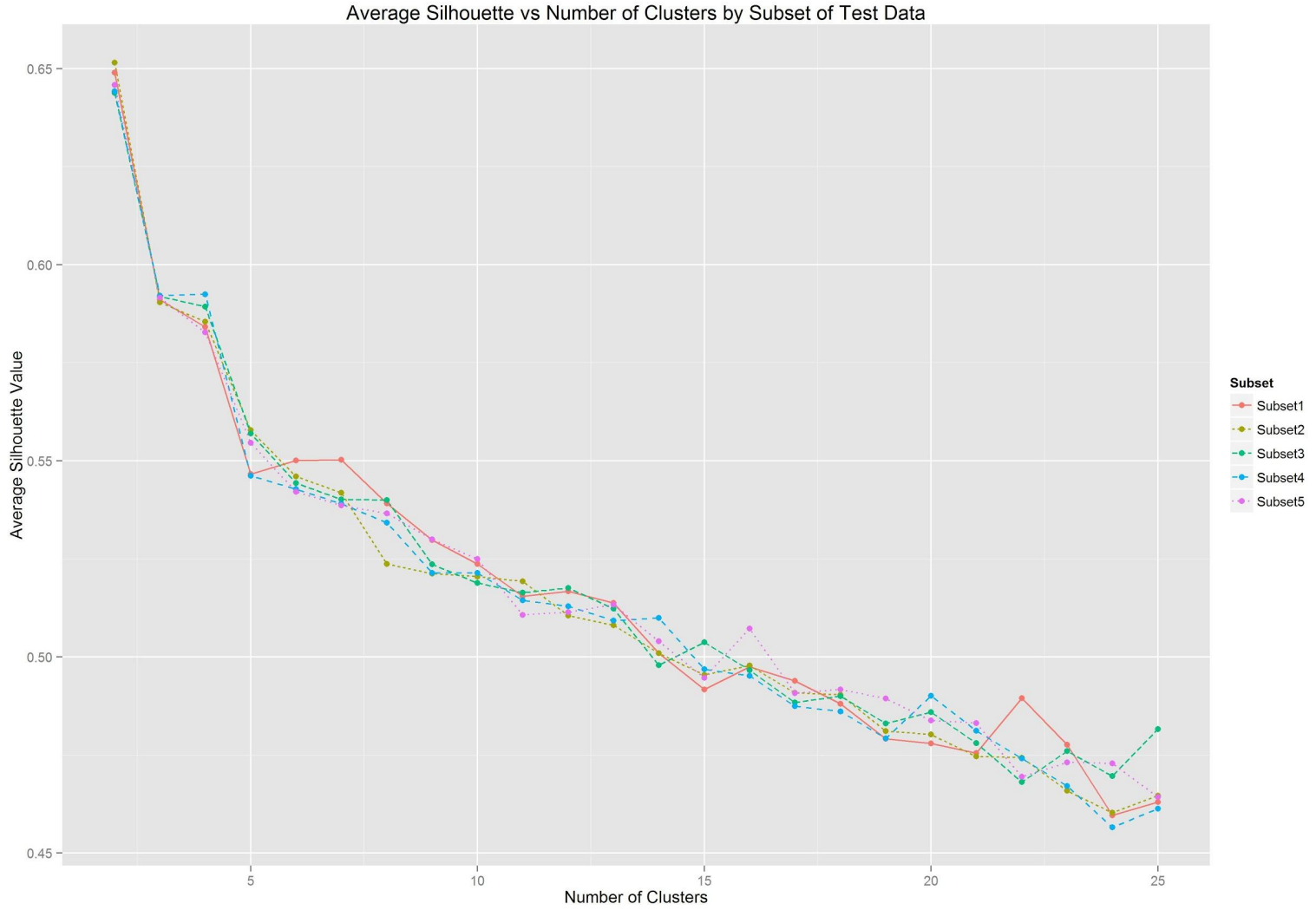

Test Data Set Heat Plot
Variable
stetson_k
stetson_j
std
small_kurtosis
skew
scatter_res_raw
qso_log_chi2nuNULL_chi2nu
qso_log_chi2_qsonu
percent_difference_flux_percentile
percent_amplitude
p2p_ssqr_diff_over_var
p2p_scatter_pfold_over_mad
p2p_scatter_over_mad
p2p_scatter_2praw
medperc90_2p_p
median_buffer_range_percentage
median_absolute_deviation
max_slope
linear_trend
freq3_harmonics_rel_phase_3
freq3_harmonics_rel_phase_2
freq3_harmonics_rel_phase_1
freq3_harmonics_freq_0
freq3_harmonics_amplitude_3
freq3_harmonics_amplitude_2
freq3_harmonics_amplitude_1
freq3_harmonics_amplitude_0
freq2_harmonics_rel_phase_3
freq2_harmonics_rel_phase_2
freq2_harmonics_rel_phase_1
freq2_harmonics_freq_0
freq2_harmonics_amplitude_3
freq2_harmonics_amplitude_2
freq2_harmonics_amplitude_1
freq2_harmonics_amplitude_0
freq1_harmonics_rel_phase_3
freq1_harmonics_rel_phase_2
freq1_harmonics_rel_phase_1
freq1_harmonics_freq_0
freq1_harmonics_amplitude_3
freq1_harmonics_amplitude_2
freq1_harmonics_amplitude_1
freq1_harmonics_amplitude_0
freq_y_offset
freq_varrat
freq_signif_ratio_31
freq_signif_ratio_21
freq_signif
freq_n_alias
freq_model_phi1_phi2
freq_model_min_delta_mags
freq_model_max_delta_mags
freq_frequency_ratio_31
freq_frequency_ratio_21
freq_amplitude_ratio_31
freq_amplitude_ratio_21
fold2P_slope_90percentile
fold2P_slope_10percentile
flux_percentile_ratio_mid80
flux_percentile_ratio_mid65
flux_percentile_ratio_mid50
flux_percentile_ratio_mid35
flux_percentile_ratio_mid20
eclpoly_p_pulse_initial
eclpoly_p_pulse
eclpoly_orb_signif
eclpoly_ls_suspect
eclpoly_best_orb_period
eclpoly_best_orb_chi2
eclpoly_8_ratio_diff
eclpoly_5_ratio_diff
eclpoly_30_ratio_diff
eclpoly_20_ratio_diff
eclpoly_15_ratio_diff
color_diff_vj
color_diff_rj
color_diff_jh
color_diff_hk
color_diff_bj
color_by_extinction
beyond1std
amplitude
Loading
0.4
0.0
-0.4
Principal Component Eigenvector (Width Proportional to Eigenvalue)

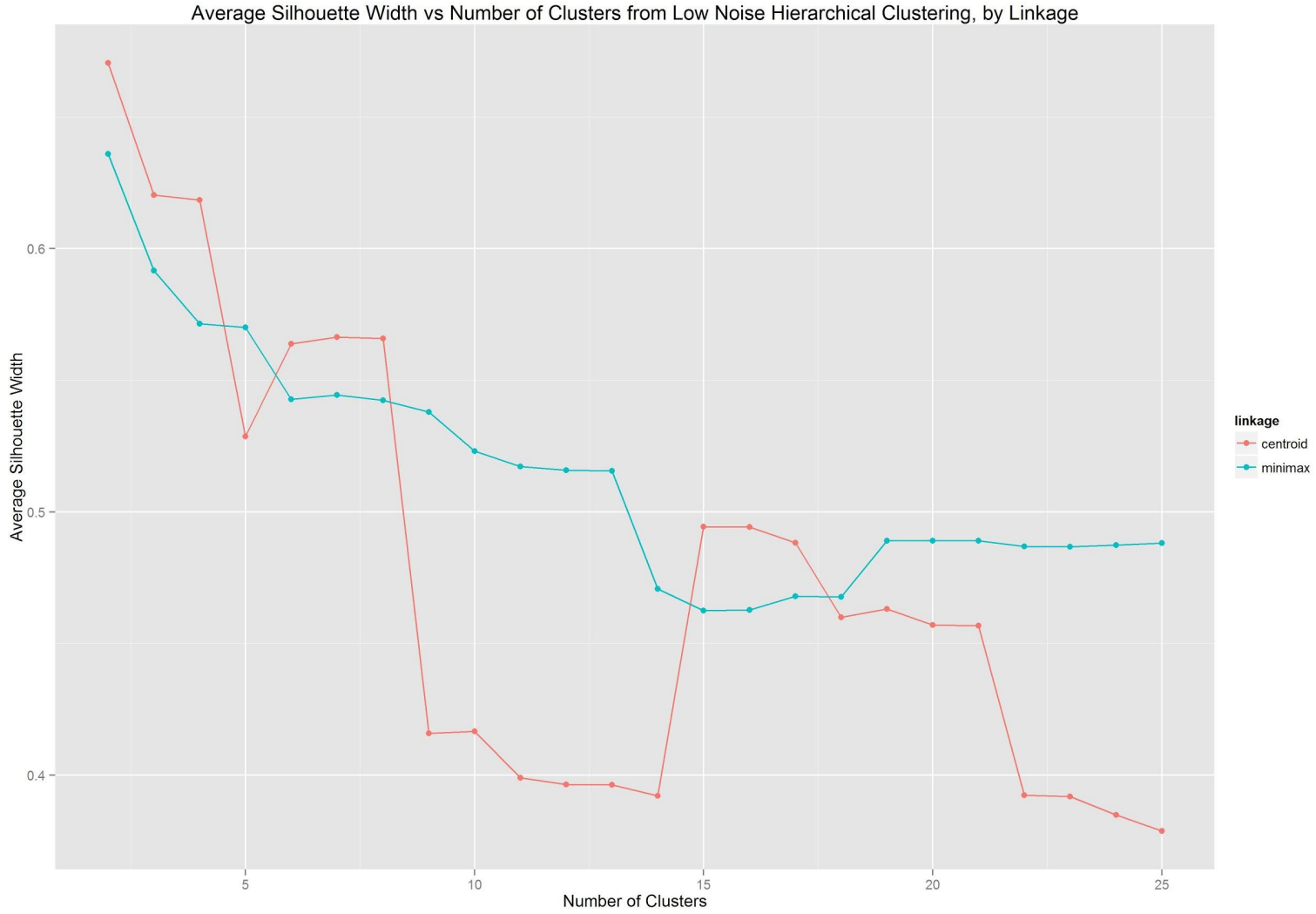
DATA FOR CLUSTERING

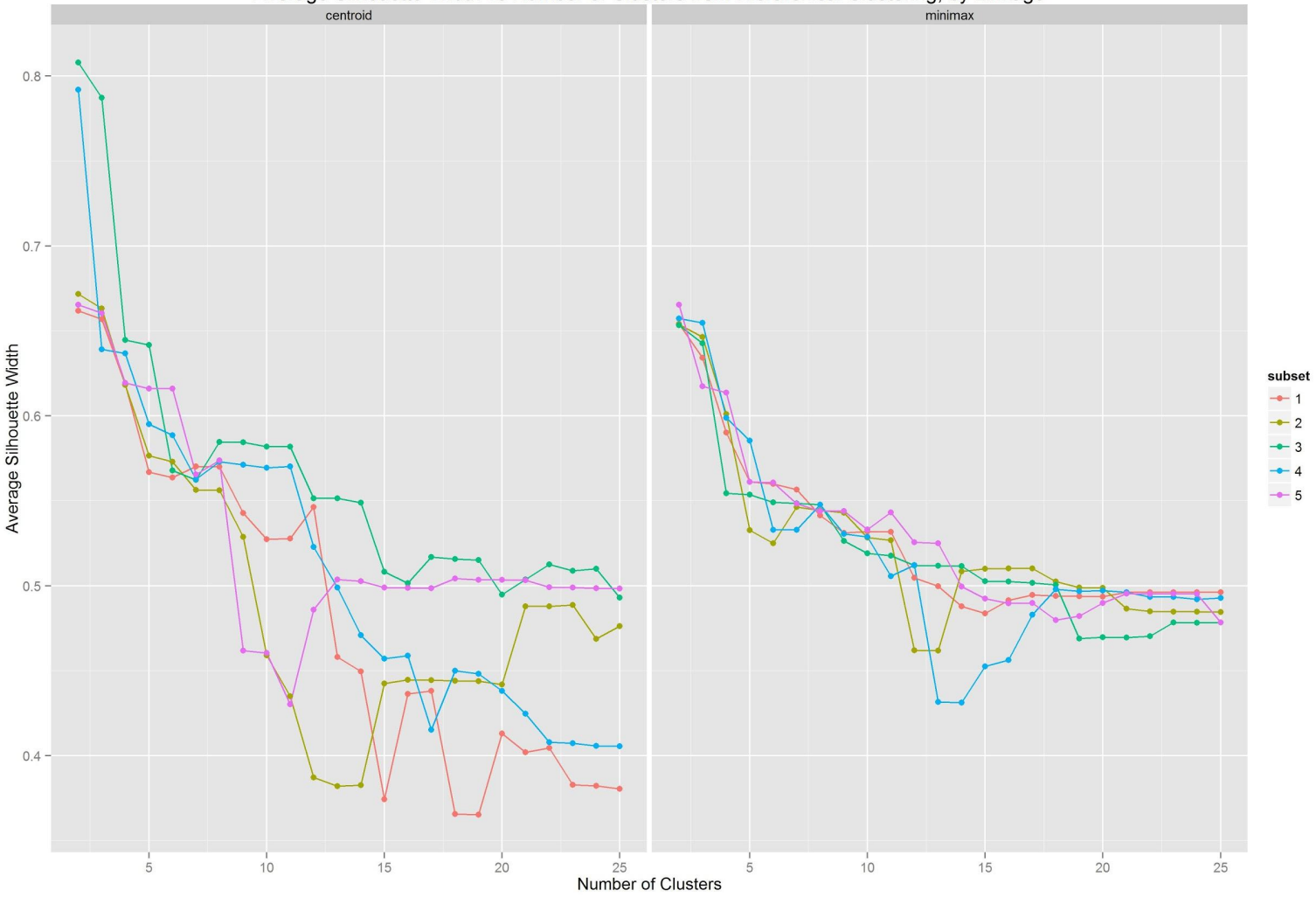
Arbitrarily chose 28 features (both low noise and test data) to reduce dimensionality.

Test data set divided into 5 random subsets (approximately equal size) for ease of computation.


Average Silhouette Width vs Number of Clusters from k-means in Low Noise Data
Average Silhouette Width
0.65
0.60
0.55
0.50
5
10
15
20
25
Number of Clusters


Average Silhouette vs Number of Clusters by Subset of Test Data
Average Silhouette Value
0.65
0.60
0.55
0.50
0.45
5
10
15
20
25
Number of Clusters
Subset
Subset1
Subset2
Subset3
Subset4
Subset5


Average Silhouette Width vs Number of Clusters from Low Noise Hierarchical Clustering, by Linkage
Average Silhouette Width
0.6
0.5
0.4
5
10
15
20
25
Number of Clusters
linkage
centroid
minimax


Average Silhouette Width vs Number of Clusters from Hierarchical Clustering, by Linkage
centroid
minimax
Average Silhouette Width
0.8
0.7
0.6
0.5
0.4
5
10
15
20
25
Number of Clusters
subset
1
2
3
4
5

OBSTACLES

1) Many observations had at least one missing feature

2) Some variables were constant across all observations

3) Clustering on all observations and features was computationally expensive

4) Visualizing clusters difficult (PCA, dendrograms)

FUTURE QUESTIONS/WORK

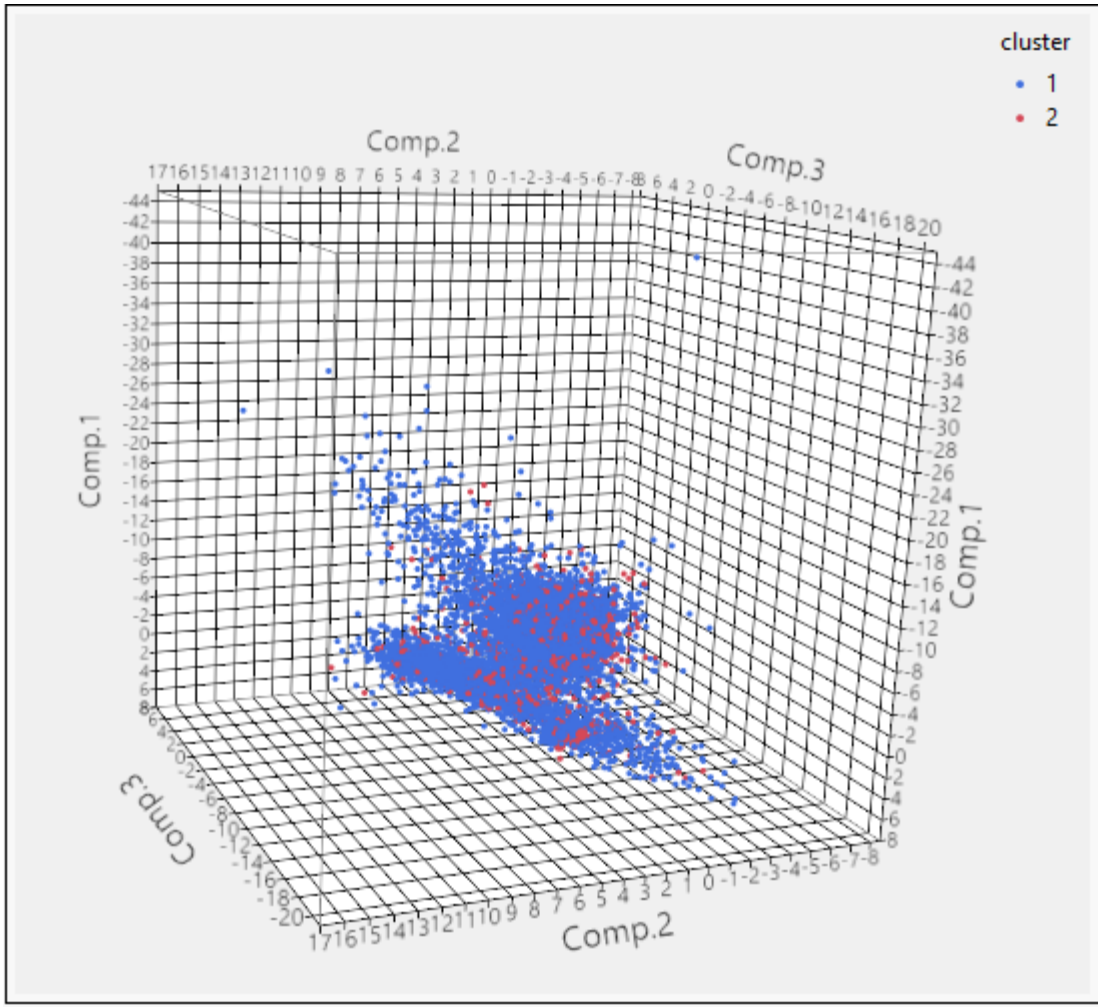
1) Other clustering methods (e.g., model-based)

2) Impute missing values

3) Find better ways to visualize data (e.g., use 3-d plots to display clustering forms)

4) Use COS Computing Cluster

End

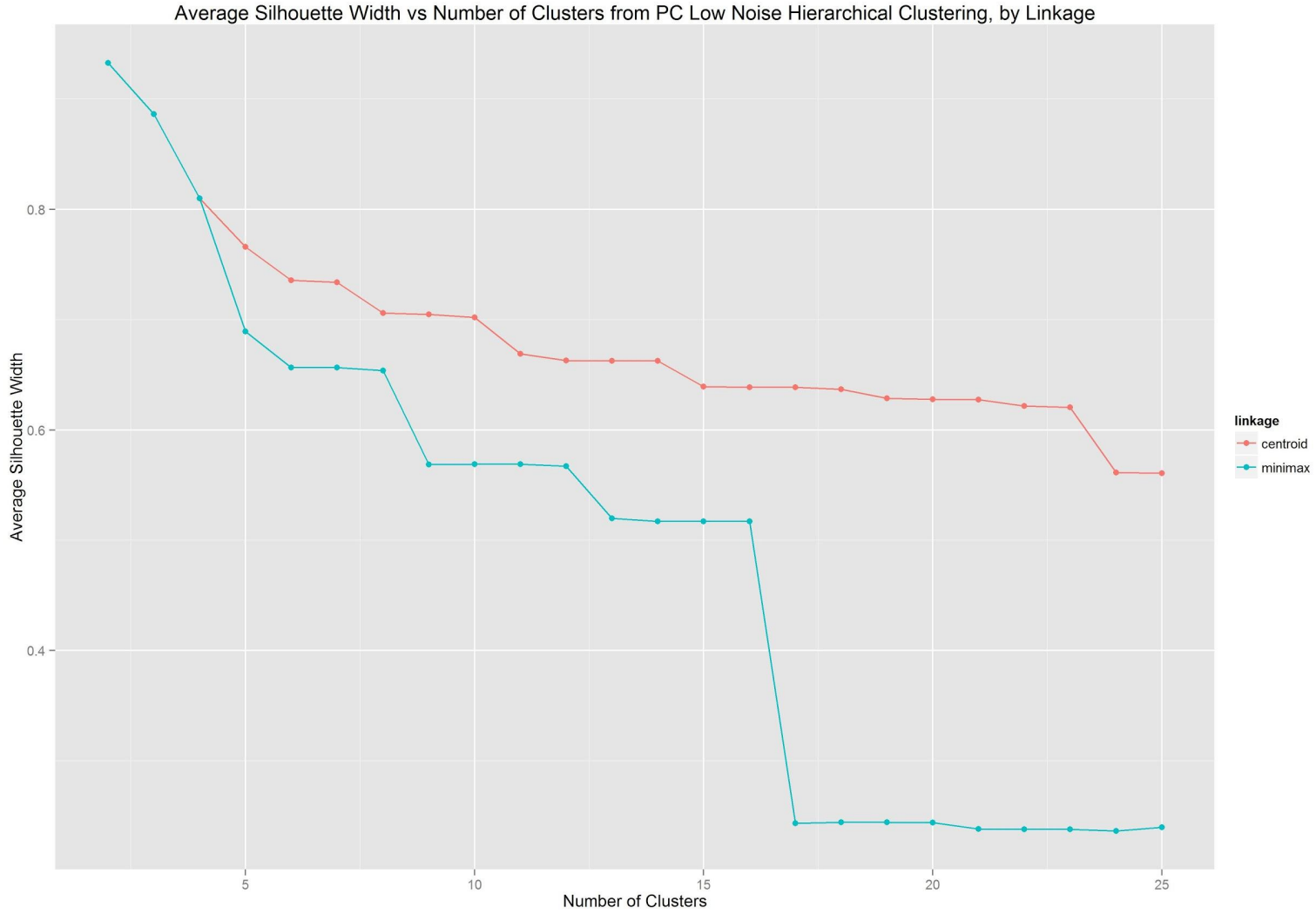


Low Noise data set

First three principal components

k = 2
cluster assignment from hierarchical clustering (28 features), minimax linkage

Cluster	1	2
Count	8759	1108


Average Silhouette Width vs Number of Clusters from PC Low Noise Hierarchical Clustering, by Linkage
Average Silhouette Width
0.8
0.6
0.4
5
10
15
20
25
Number of Clusters
linkage
centroid
minimax

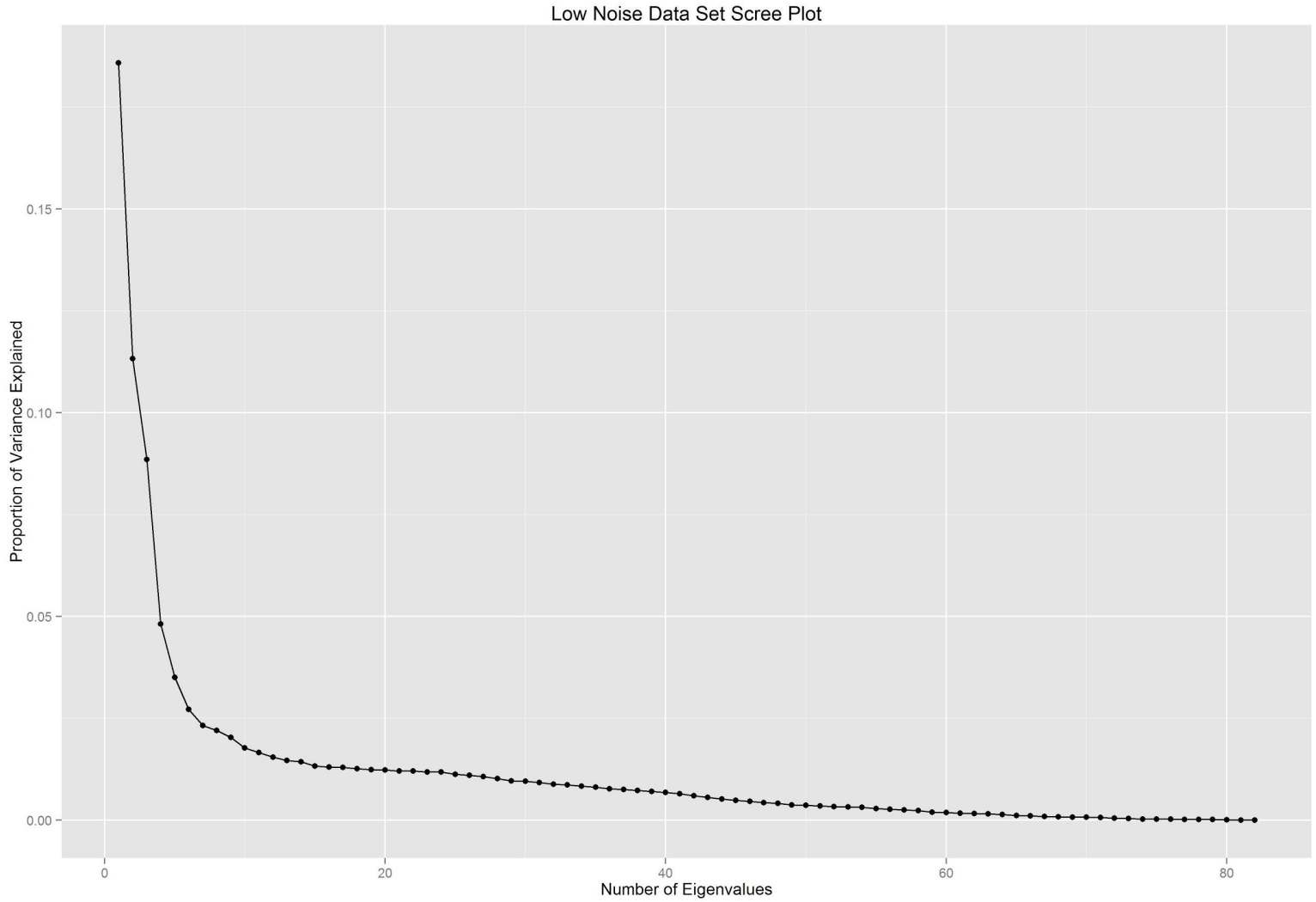
Clustering Astronomy Data

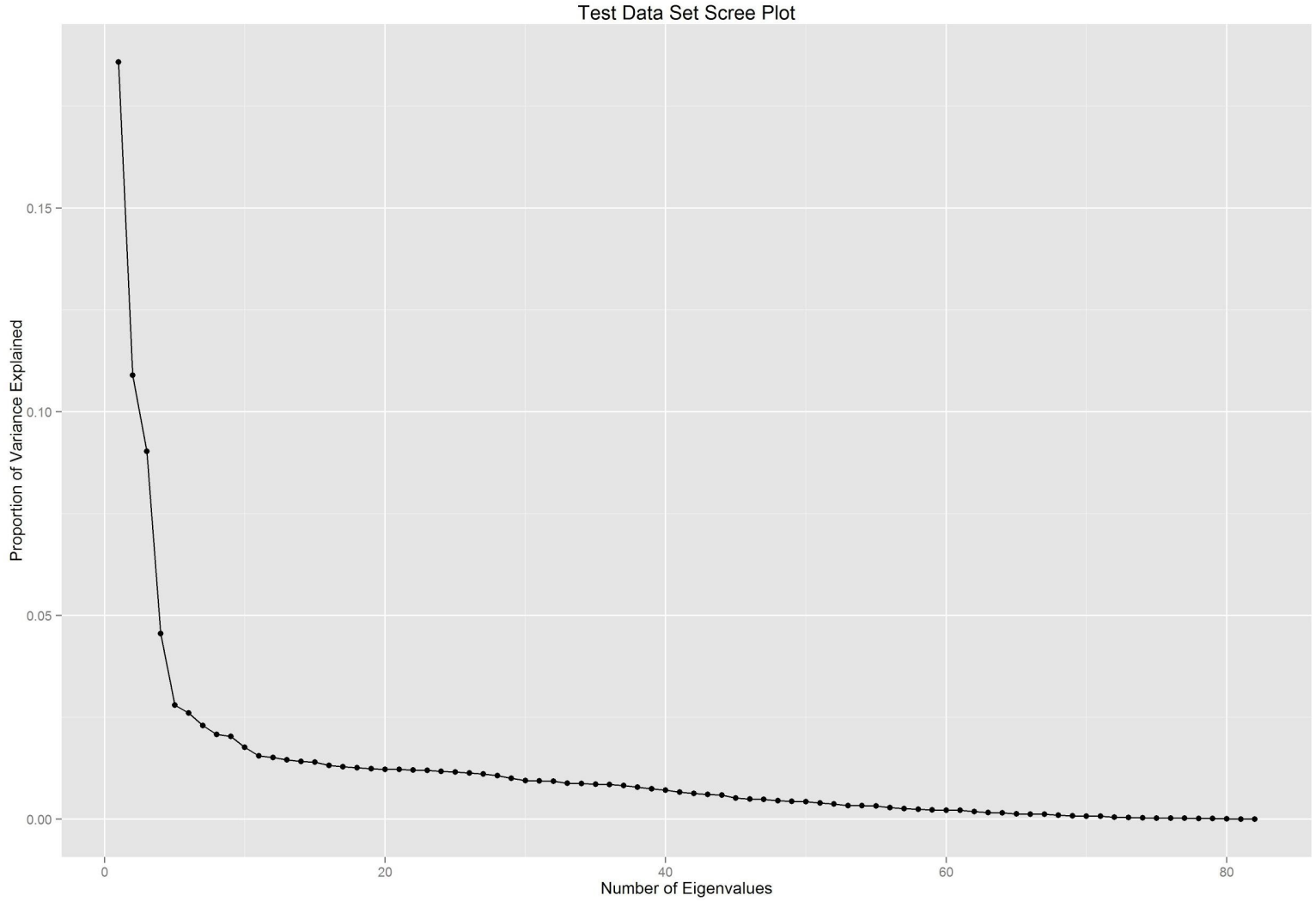
Katherine Eng, Jie Hu, Vanessa Lepe, Kalbi Zongo

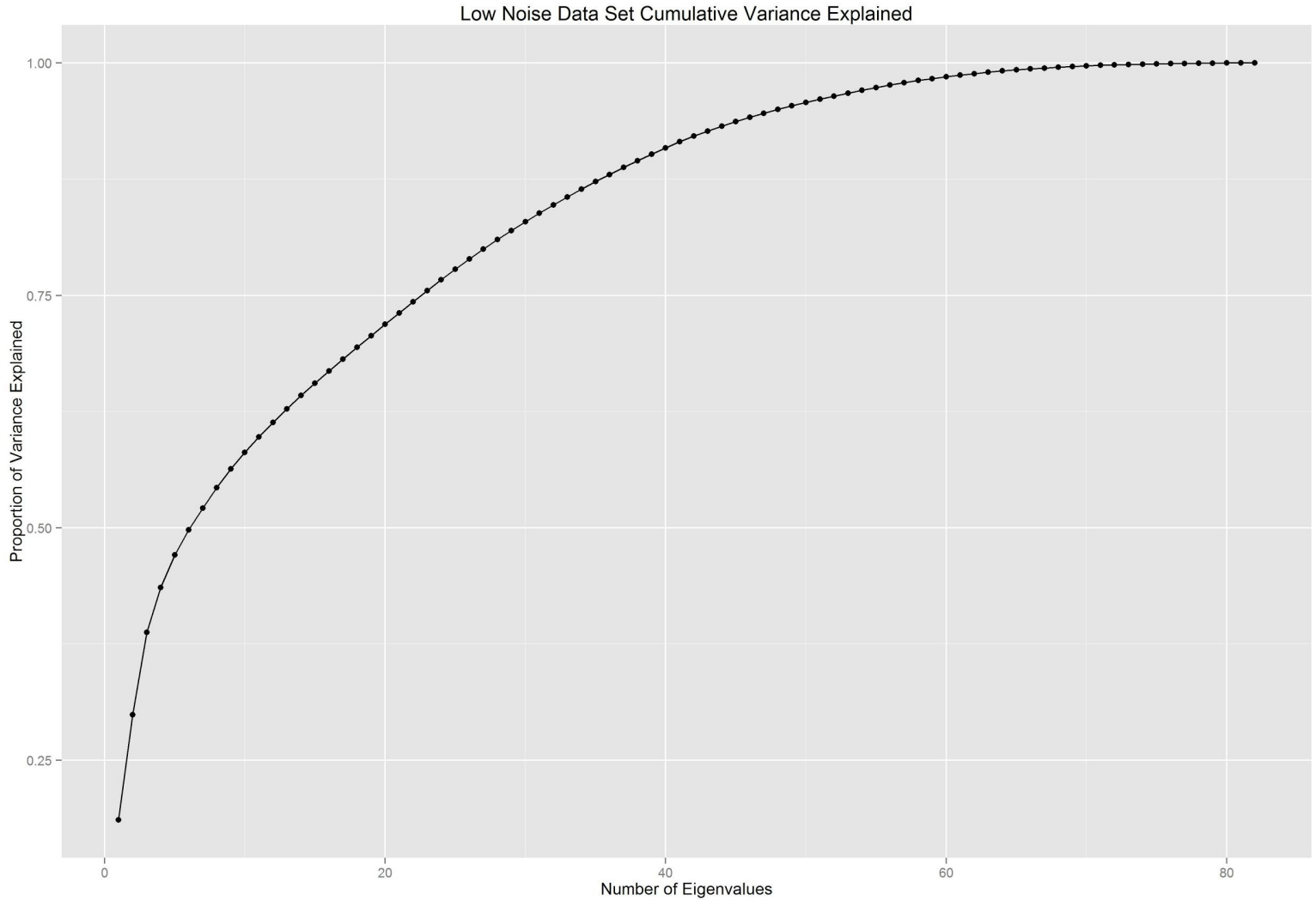
QUESTIONS OF INTEREST

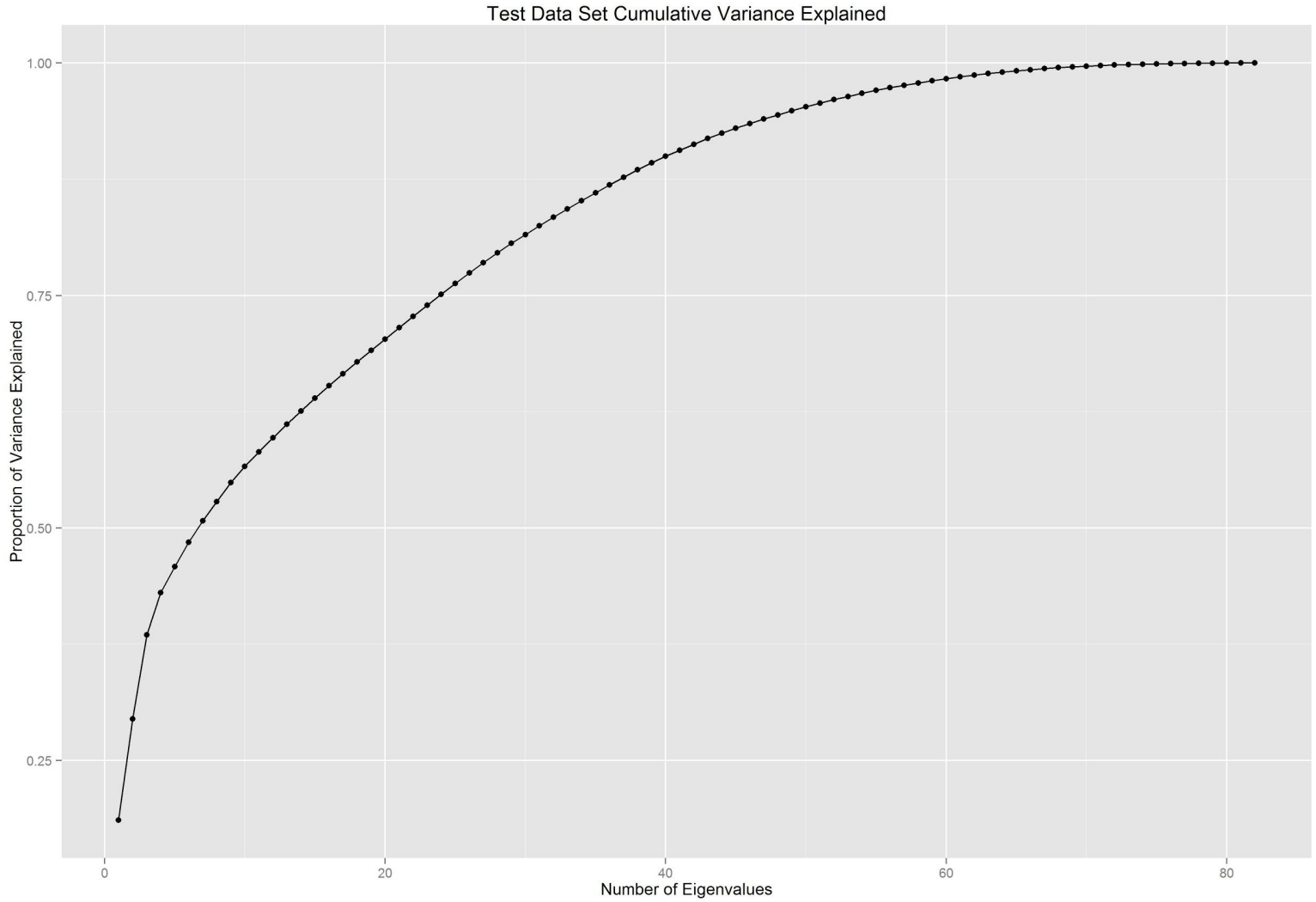
1. How does the principal component analysis (PCA) for low noise data compare to the PCA for the full test data set?
2. How do clustering methods (k-means vs hierarchical) differ in the low noise data set?
3. How do clustering methods (k-means vs hierarchical) differ in the test data set?

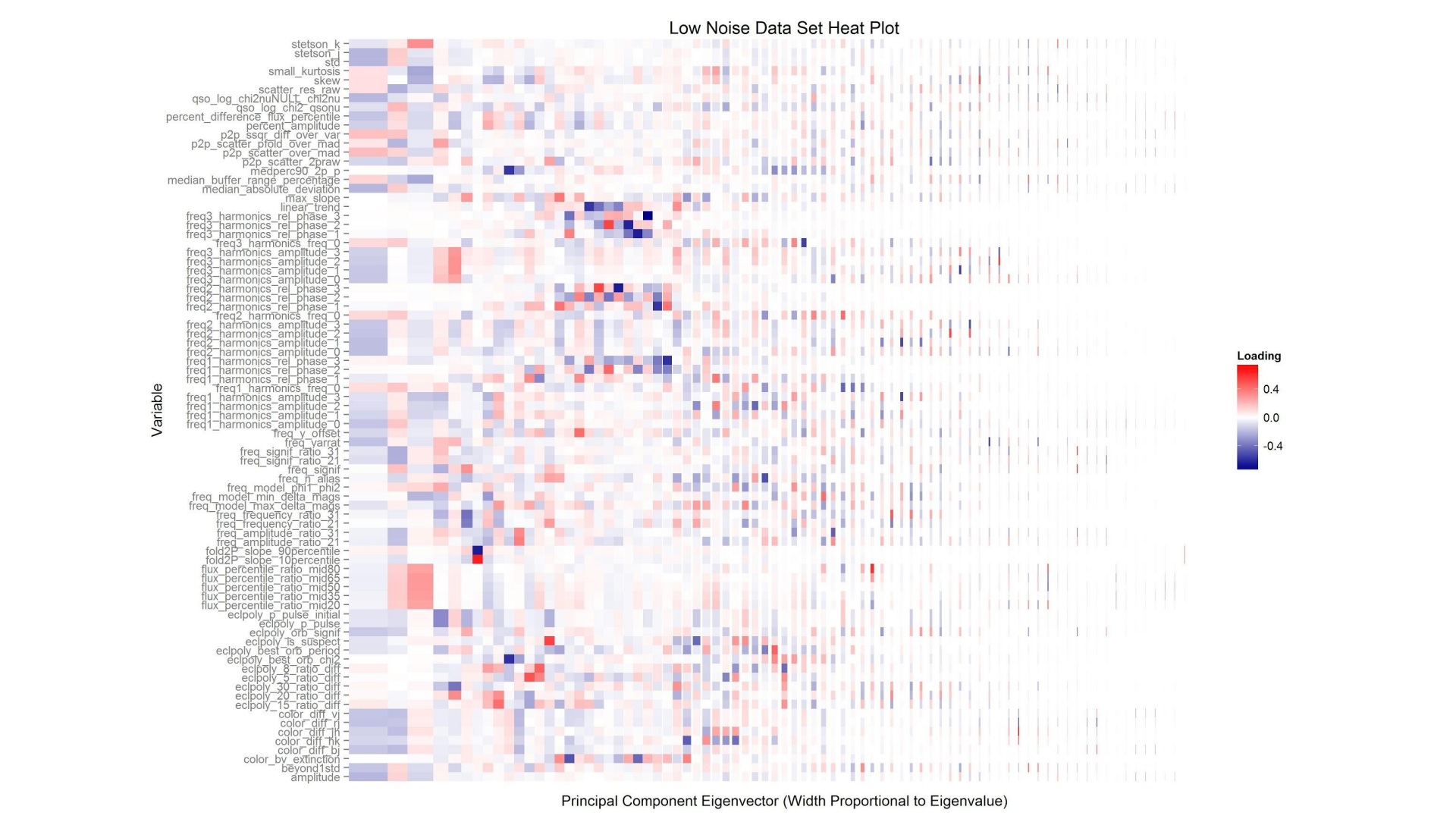
RESULTS


Low Noise Data Set Scree Plot
Proportion of Variance Explained
0.15
0.10
0.05
0.00
0
20
40
60
80
Number of Eigenvalues


Test Data Set Scree Plot
Proportion of Variance Explained
0.15
0.10
0.05
0.00
0
20
40
60
80
Number of Eigenvalues


Low Noise Data Set Cumulative Variance Explained
Proportion of Variance Explained
1.00
0.75
0.50
0.25
0
20
40
60
80
Number of Eigenvalues


Test Data Set Cumulative Variance Explained
Proportion of Variance Explained
1.00
0.75
0.50
0.25
0
20
40
60
80
Number of Eigenvalues

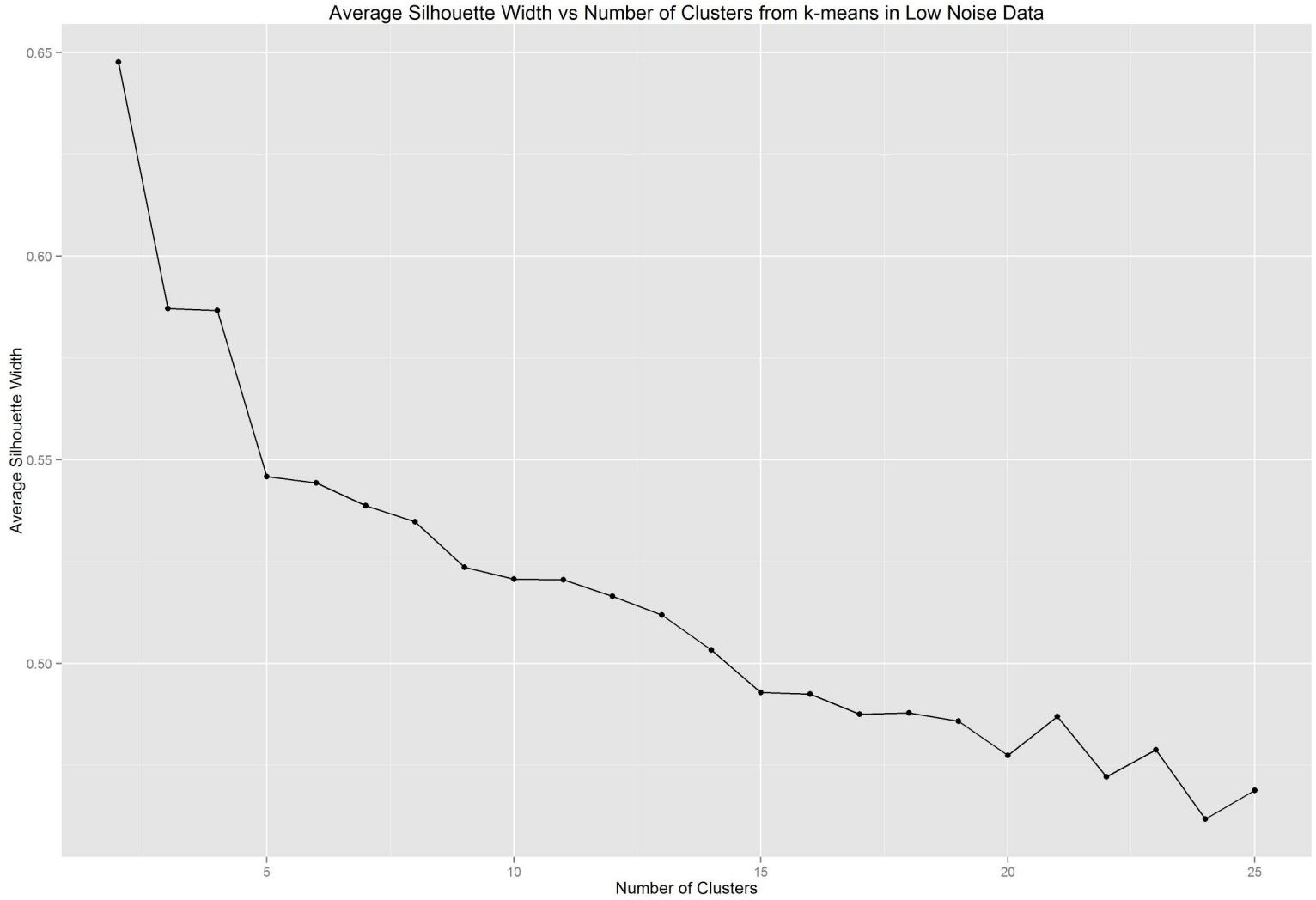

Low Noise Data Set Heat Plot
stetson_k
stetson_j
std
small_kurtosis
skew
scatter_res_raw
qso_log_chi2nuNULL_chi2nu
qso_log_chi2_qsonu
percent_difference_flux_percentile
percent_amplitude
p2p_ssqr_diff_over_var
p2p_scatter_pfold_over_mad
p2p_scatter_over_mad
p2p_scatter_2praw
medperc90_2p_p
median_buffer_range_percentage
median_absolute_deviation
max_slope
linear_trend
freq3_harmonics_rel_phase_3
freq3_harmonics_rel_phase_2
freq3_harmonics_rel_phase_1
freq3_harmonics_freq_0
freq3_harmonics_amplitude_3
freq3_harmonics_amplitude_2
freq3_harmonics_amplitude_1
freq3_harmonics_amplitude_0
freq2_harmonics_rel_phase_3
freq2_harmonics_rel_phase_2
freq2_harmonics_rel_phase_1
freq2_harmonics_freq_0
freq2_harmonics_amplitude_3
freq2_harmonics_amplitude_2
freq2_harmonics_amplitude_1
freq2_harmonics_amplitude_0
freq1_harmonics_rel_phase_3
freq1_harmonics_rel_phase_2
freq1_harmonics_rel_phase_1
freq1_harmonics_freq_0
freq1_harmonics_amplitude_3
freq1_harmonics_amplitude_2
freq1_harmonics_amplitude_1
freq1_harmonics_amplitude_0
freq_y_offset
freq_varrat
freq_signif_ratio_31
freq_signif_ratio_21
freq_signif
freq_n_alias
freq_model_phi1_phi2
freq_model_min_delta_mags
freq_model_max_delta_mags
freq_frequency_ratio_31
freq_frequency_ratio_21
freq_amplitude_ratio_31
freq_amplitude_ratio_21
fold2P_slope_90percentile
fold2P_slope_10percentile
flux_percentile_ratio_mid80
flux_percentile_ratio_mid65
flux_percentile_ratio_mid50
flux_percentile_ratio_mid35
flux_percentile_ratio_mid20
eclpoly_p_pulse_initial
eclpoly_p_pulse
eclpoly_orb_signif
eclpoly_ls_suspect
eclpoly_best_orb_period
eclpoly_best_orb_chi2
eclpoly_8_ratio_diff
eclpoly_5_ratio_diff
eclpoly_30_ratio_diff
eclpoly_20_ratio_diff
eclpoly_15_ratio_diff
color_diff_vj
color_diff_rj
color_diff_ijh
color_diff_hk
color_diff_bj
color_by_extinction
beyond1std
amplitude
Variable
Loading
0.4
0.0
-0.4
Principal Component Eigenvector (Width Proportional to Eigenvalue)

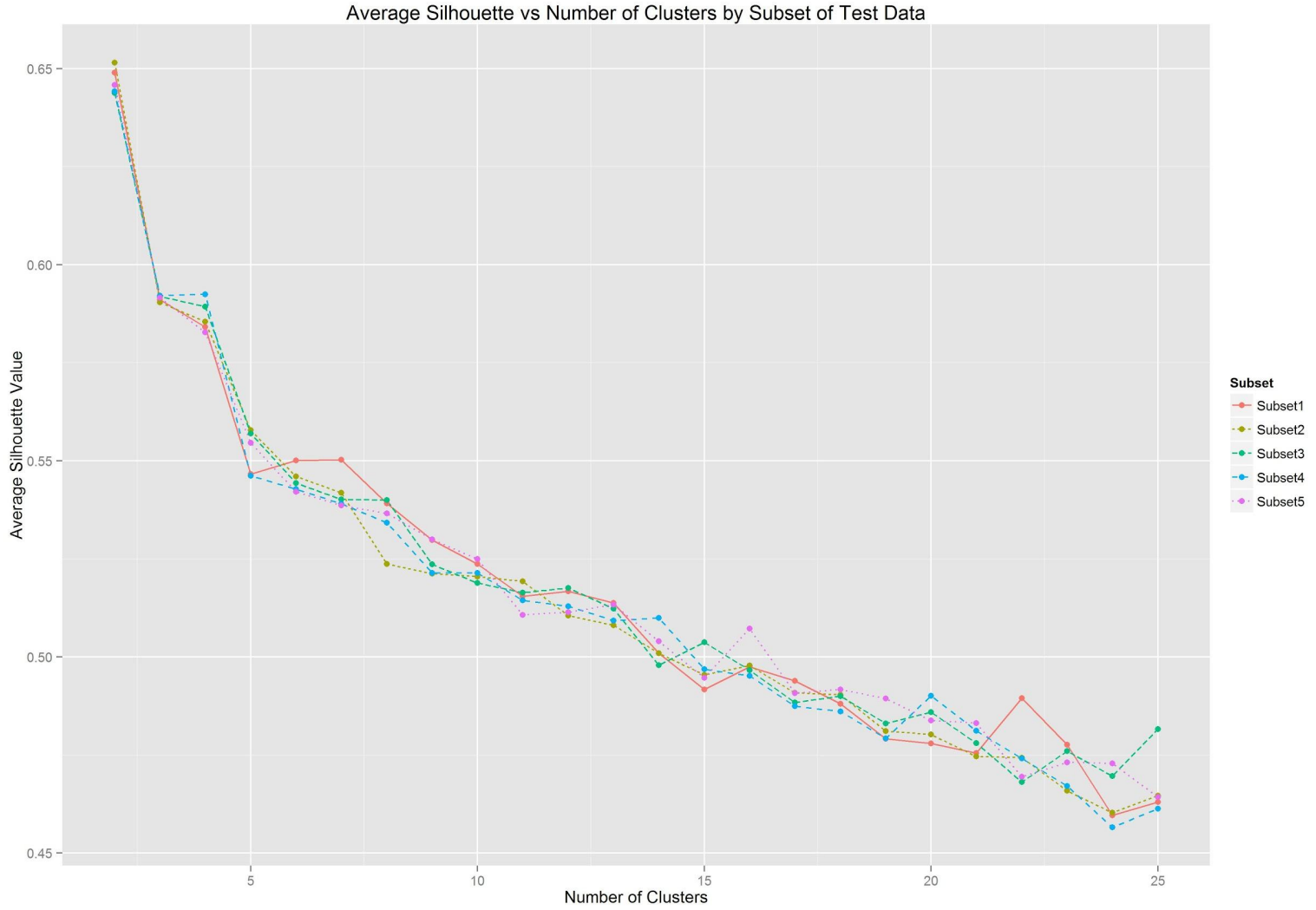

Test Data Set Heat Plot
Variable
stetson_k
stetson_j
std
small_kurtosis
skew
scatter_res_raw
qso_log_chi2nuNULL_chi2nu
qso_log_chi2_qsonu
percent_difference_flux_percentile
percent_amplitude
p2p_ssqr_diff_over_var
p2p_scatter_pfold_over_mad
p2p_scatter_over_mad
p2p_scatter_2praw
medperc90_2p_p
median_buffer_range_percentage
median_absolute_deviation
max_slope
linear_trend
freq3_harmonics_rel_phase_3
freq3_harmonics_rel_phase_2
freq3_harmonics_rel_phase_1
freq3_harmonics_freq_0
freq3_harmonics_amplitude_3
freq3_harmonics_amplitude_2
freq3_harmonics_amplitude_1
freq3_harmonics_amplitude_0
freq2_harmonics_rel_phase_3
freq2_harmonics_rel_phase_2
freq2_harmonics_rel_phase_1
freq2_harmonics_freq_0
freq2_harmonics_amplitude_3
freq2_harmonics_amplitude_2
freq2_harmonics_amplitude_1
freq2_harmonics_amplitude_0
freq1_harmonics_rel_phase_3
freq1_harmonics_rel_phase_2
freq1_harmonics_rel_phase_1
freq1_harmonics_freq_0
freq1_harmonics_amplitude_3
freq1_harmonics_amplitude_2
freq1_harmonics_amplitude_1
freq1_harmonics_amplitude_0
freq_y_offset
freq_varrat
freq_signif_ratio_31
freq_signif_ratio_21
freq_signif
freq_n_alias
freq_model_phi1_phi2
freq_model_min_delta_mags
freq_model_max_delta_mags
freq_frequency_ratio_31
freq_frequency_ratio_21
freq_amplitude_ratio_31
freq_amplitude_ratio_21
fold2P_slope_90percentile
fold2P_slope_10percentile
flux_percentile_ratio_mid80
flux_percentile_ratio_mid65
flux_percentile_ratio_mid50
flux_percentile_ratio_mid35
flux_percentile_ratio_mid20
eclpoly_p_pulse_initial
eclpoly_p_pulse
eclpoly_orb_signif
eclpoly_ls_suspect
eclpoly_best_orb_period
eclpoly_best_orb_chi2
eclpoly_8_ratio_diff
eclpoly_5_ratio_diff
eclpoly_30_ratio_diff
eclpoly_20_ratio_diff
eclpoly_15_ratio_diff
color_diff_vj
color_diff_rj
color_diff_jh
color_diff_hk
color_diff_bj
color_by_extinction
beyond1std
amplitude
Loading
0.4
0.0
-0.4
Principal Component Eigenvector (Width Proportional to Eigenvalue)

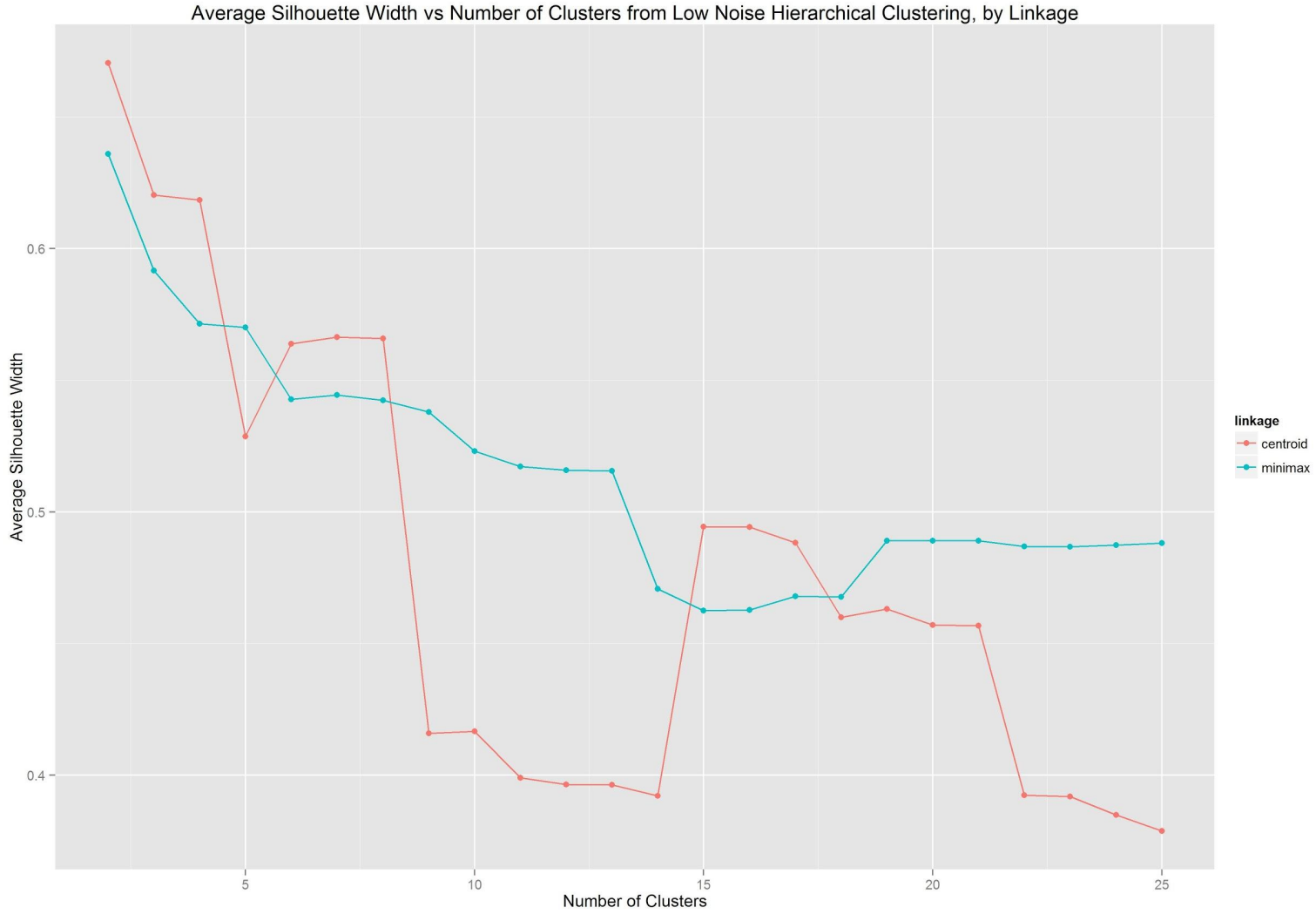
DATA FOR CLUSTERING

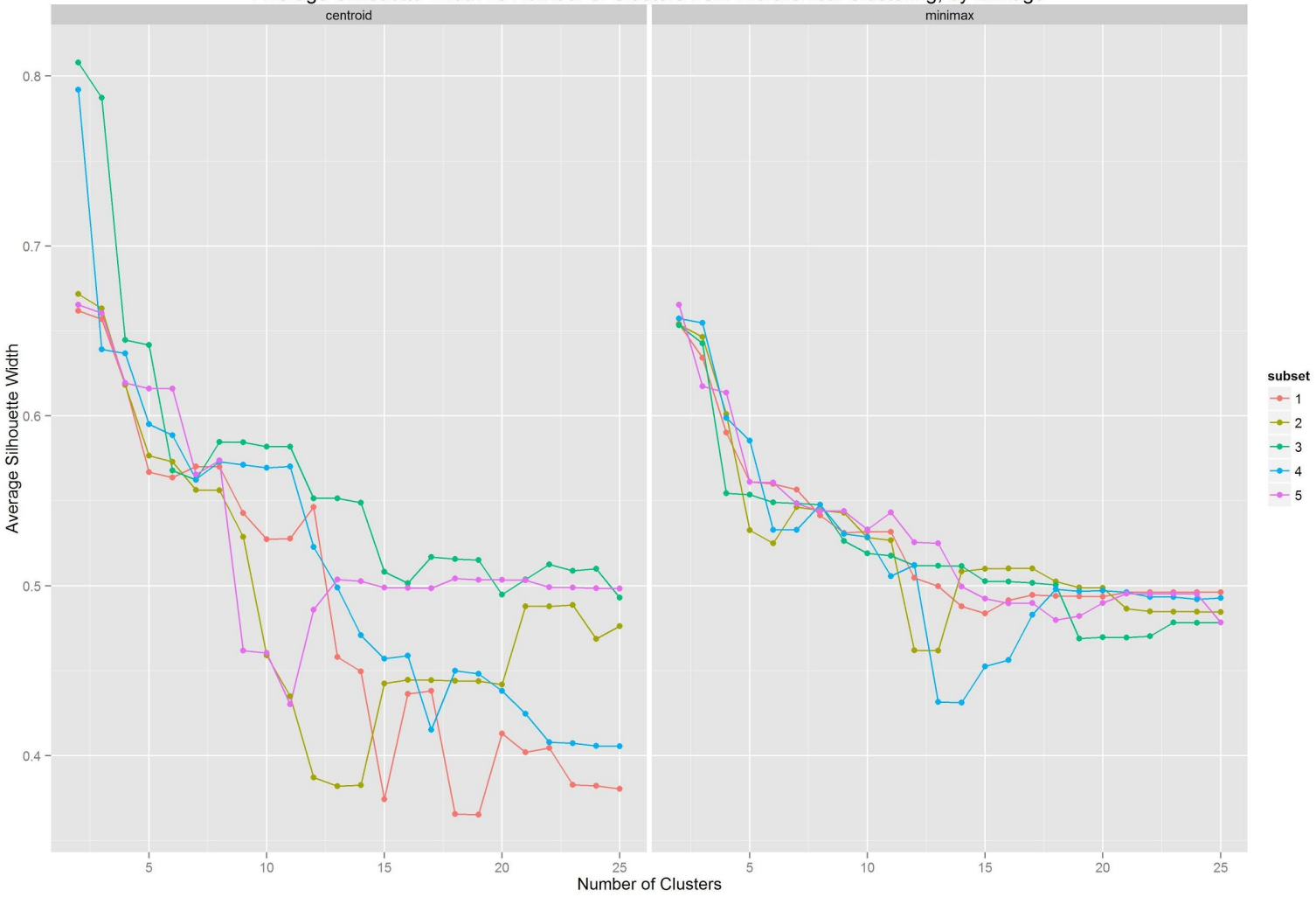
Arbitrarily chose 28 features (both low noise and test data) to reduce dimensionality.

Test data set divided into 5 random subsets (approximately equal size) for ease of computation.


Average Silhouette Width vs Number of Clusters from k-means in Low Noise Data
Average Silhouette Width
0.65
0.60
0.55
0.50
5
10
15
20
25
Number of Clusters


Average Silhouette vs Number of Clusters by Subset of Test Data
Average Silhouette Value
0.65
0.60
0.55
0.50
0.45
5
10
15
20
25
Number of Clusters
Subset
Subset1
Subset2
Subset3
Subset4
Subset5


Average Silhouette Width vs Number of Clusters from Low Noise Hierarchical Clustering, by Linkage
Average Silhouette Width
0.6
0.5
0.4
5
10
15
20
25
Number of Clusters
linkage
centroid
minimax


Average Silhouette Width vs Number of Clusters from Hierarchical Clustering, by Linkage
centroid
minimax
Average Silhouette Width
0.8
0.7
0.6
0.5
0.4
5
10
15
20
25
Number of Clusters
subset
1
2
3
4
5

OBSTACLES

1) Many observations had at least one missing feature

2) Some variables were constant across all observations

3) Clustering on all observations and features was computationally expensive

4) Visualizing clusters difficult (PCA, dendrograms)

FUTURE QUESTIONS/WORK

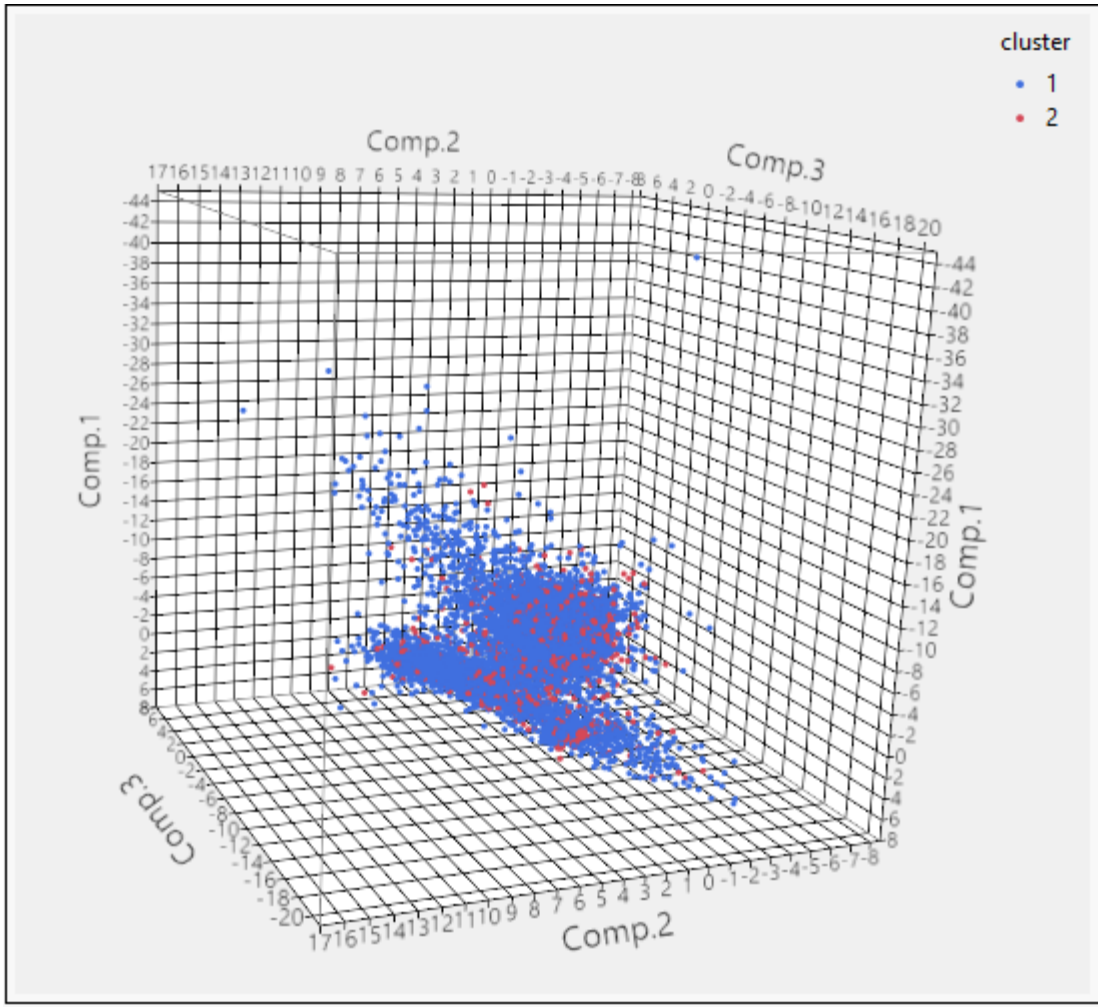
1) Other clustering methods (e.g., model-based)

2) Impute missing values

3) Find better ways to visualize data (e.g., use 3-d plots to display clustering forms)

4) Use COS Computing Cluster

End

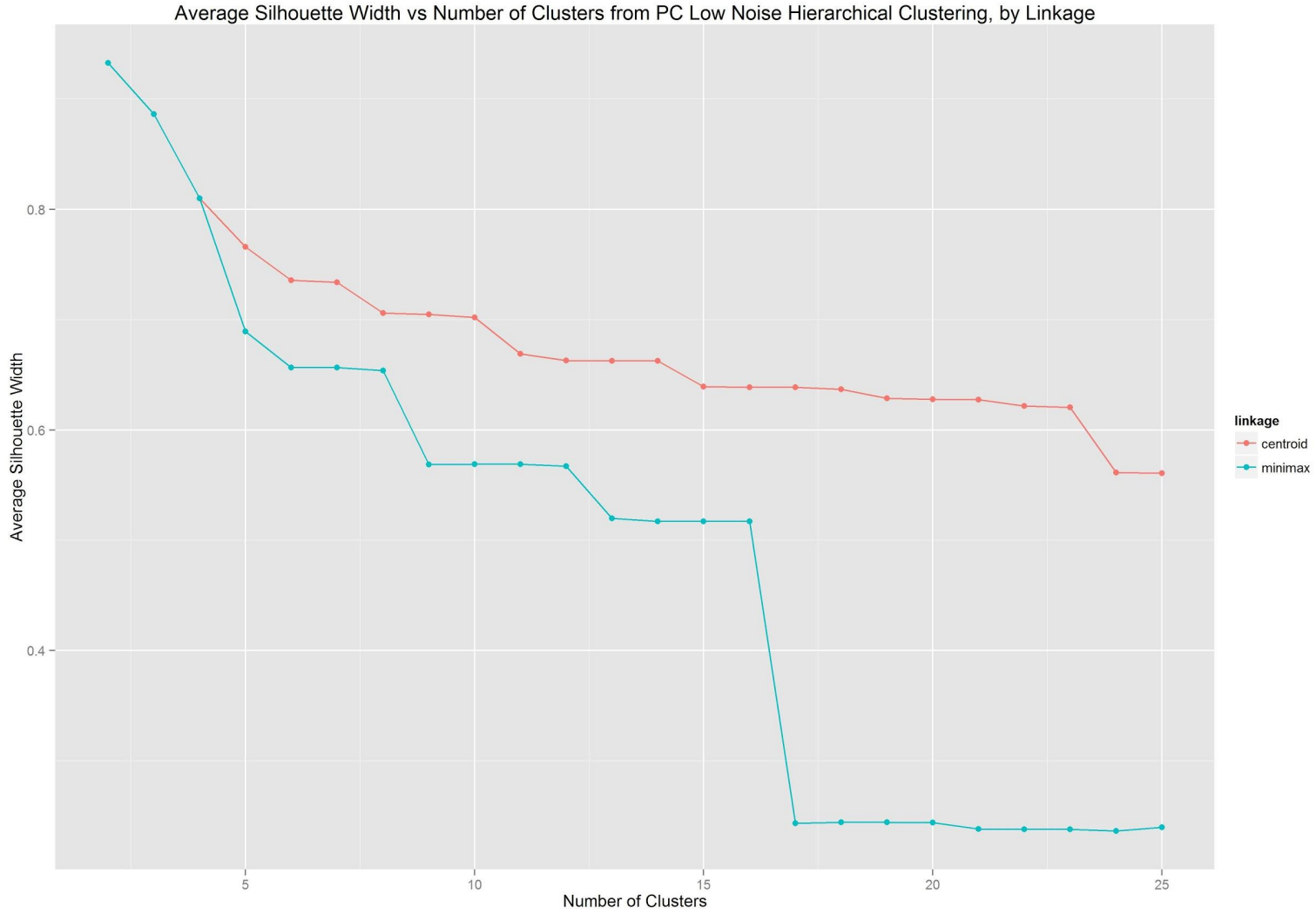


Low Noise data set

First three principal components

k = 2
cluster assignment from hierarchical clustering (28 features), minimax linkage

Cluster	1	2
Count	8759	1108


Average Silhouette Width vs Number of Clusters from PC Low Noise Hierarchical Clustering, by Linkage
Average Silhouette Width
0.8
0.6
0.4
5
10
15
20
25
Number of Clusters
linkage
centroid
minimax