

APPENDIX B

ADDITIONAL FIGURES & TABLES

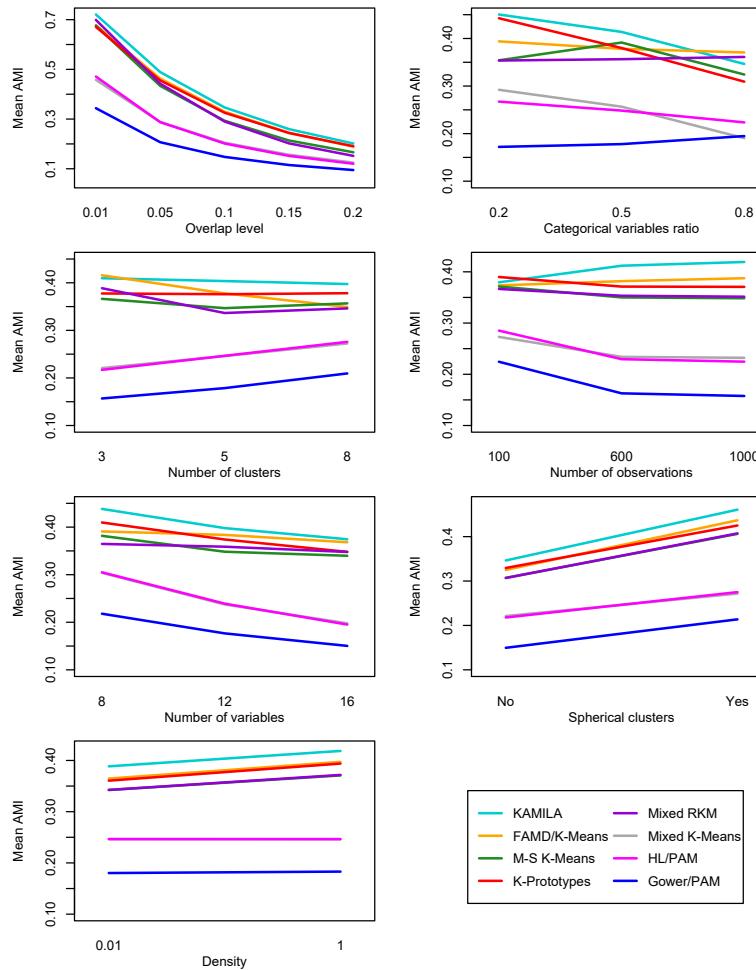


Figure B.1: Two-way interactions of method by overlap level, percentage of categorical variables, number of clusters, number of observations, number of variables, density, and cluster sphericity (mean AMI values). Subplots/factors are arranged, from left to right, by decreasing effect size η_p^2 .

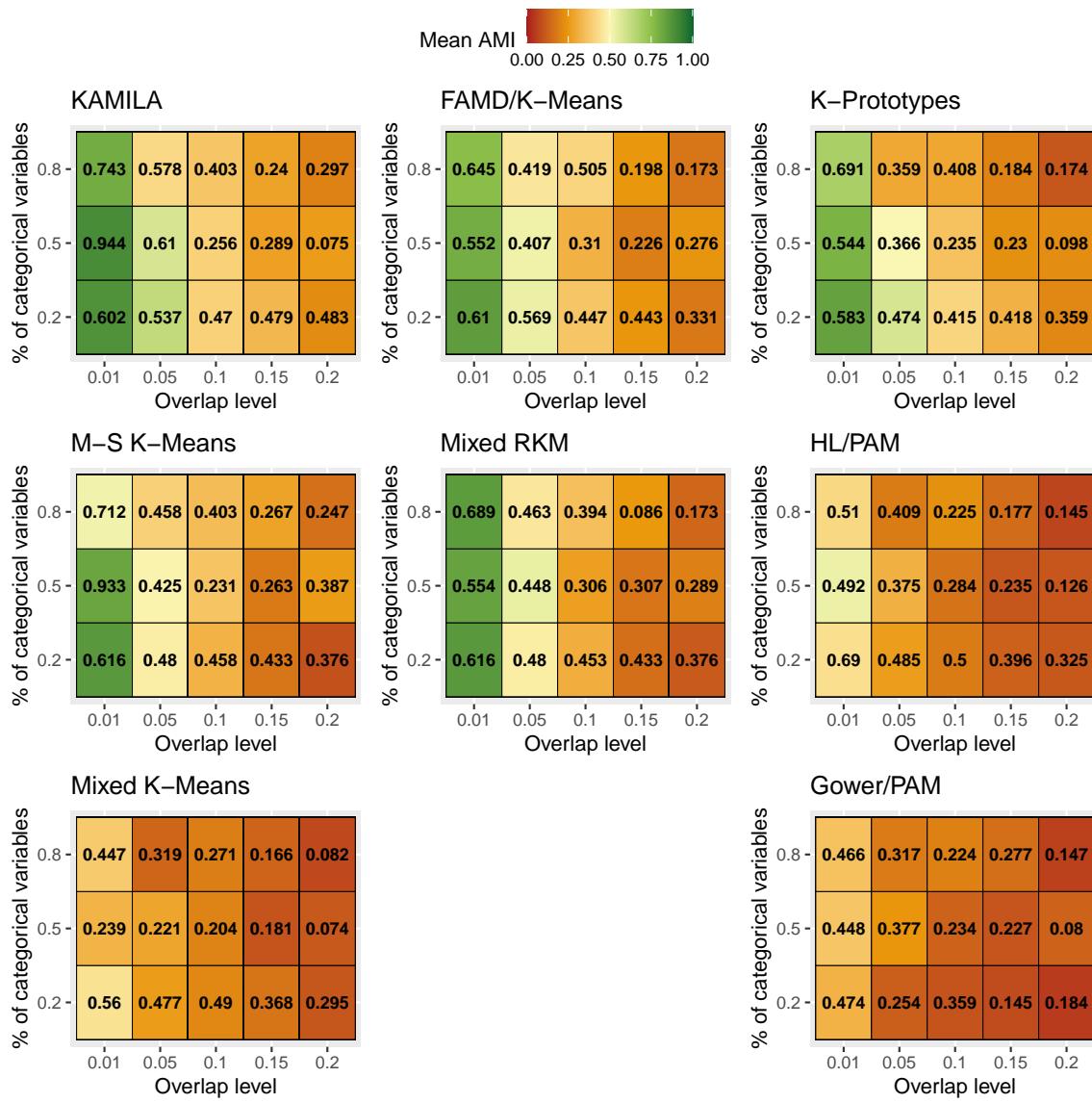


Figure B.2: Three-way interaction of method by overlap level and percentage of categorical variables. The numbers indicate the mean AMI for each combination of overlap level and percentage of categorical variables.

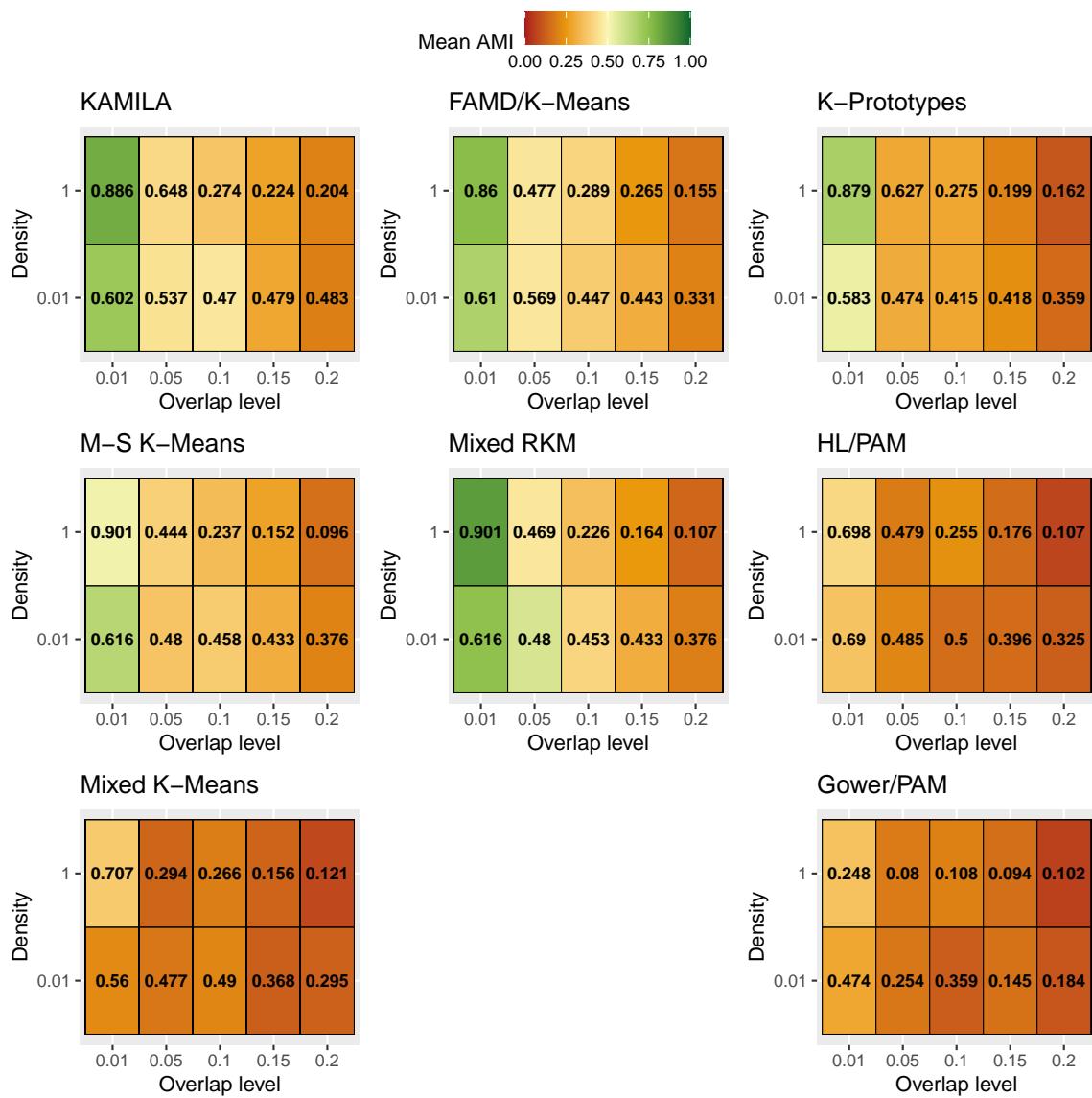


Figure B.3: Three-way interaction of method by overlap level and cluster density. The numbers indicate the mean AMI for each combination of overlap level and cluster density.

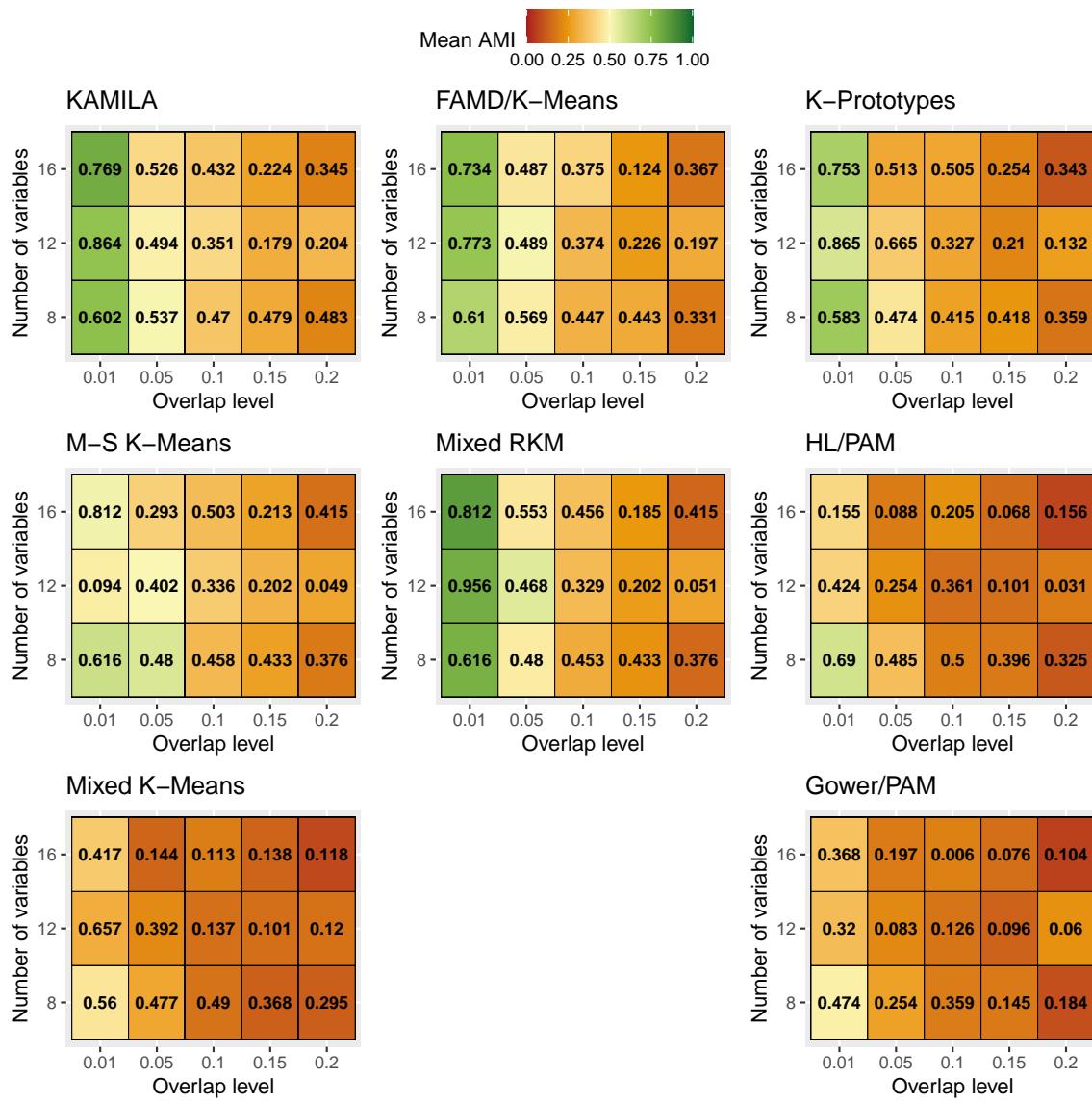


Figure B.4: Three-way interaction of method by overlap level and number of variables. The numbers indicate the mean AMI for each combination of overlap level and number of variables.

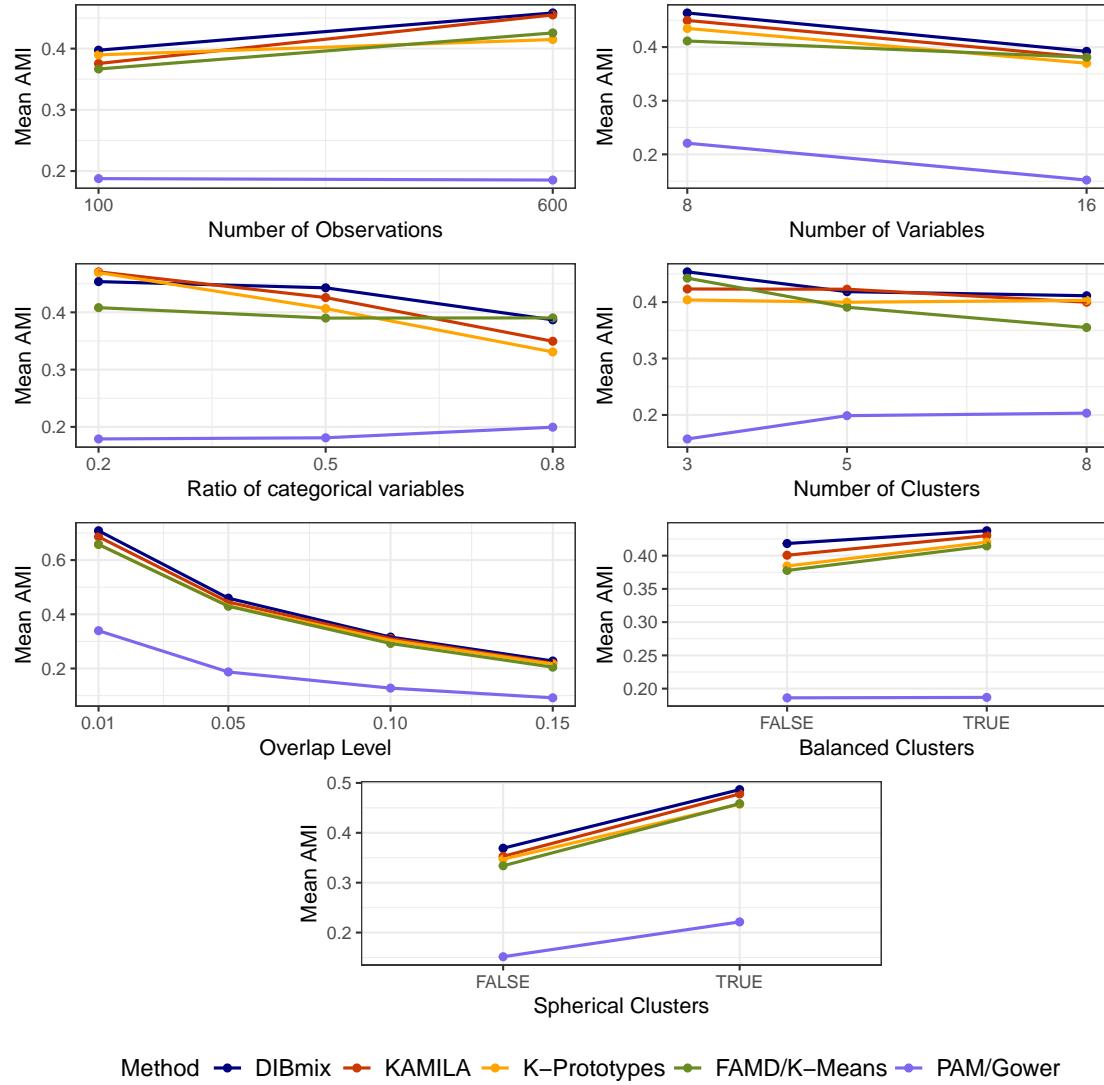


Figure B.5: Two-way interactions of method by overlap level, percentage of categorical variables, number of clusters, number of observations, number of variables, density, and cluster sphericity (mean AMI values).

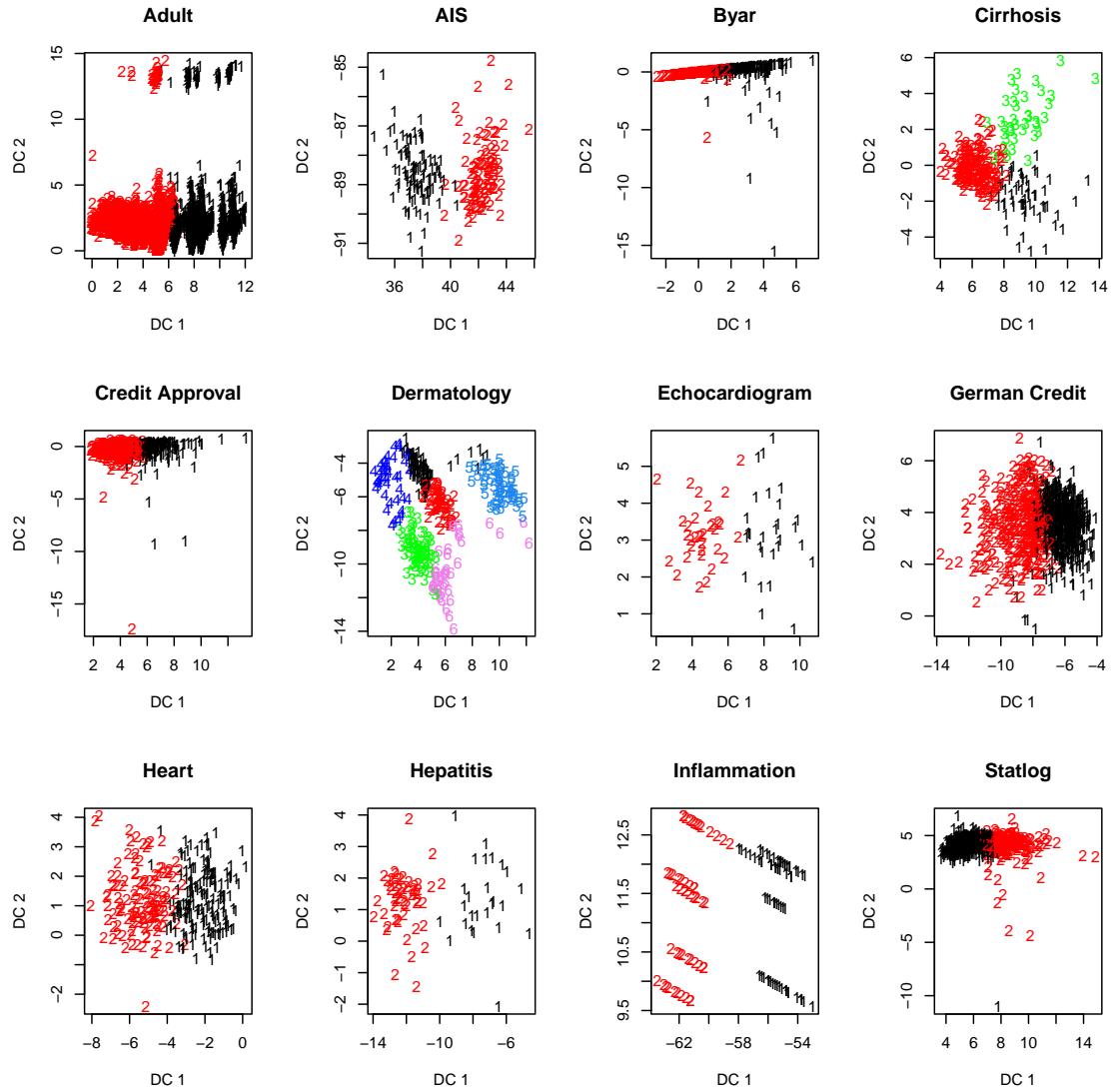


Figure B.6: Projection of cluster partitions obtained using KAMILA on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

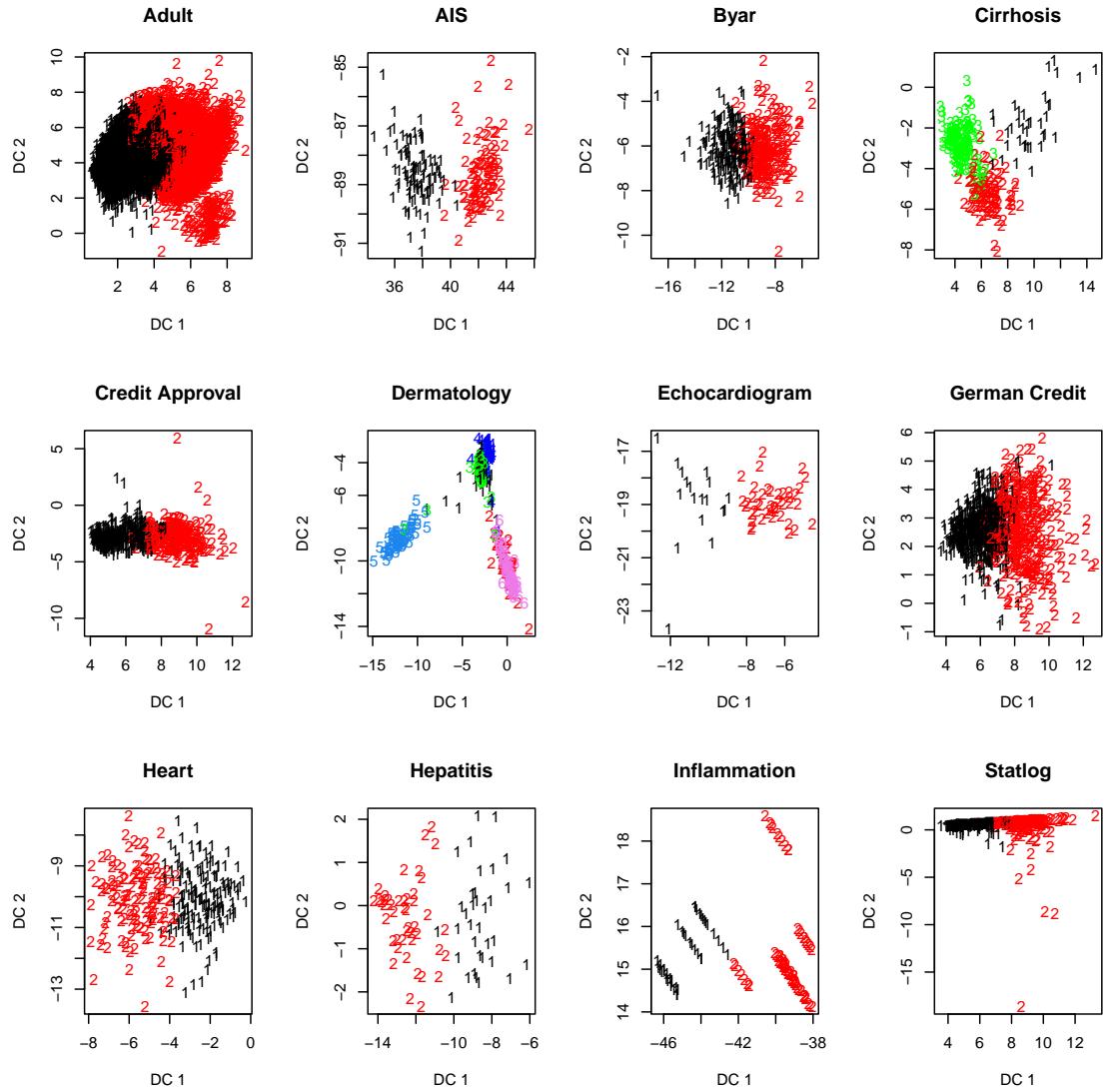


Figure B.7: Projection of cluster partitions obtained using K-Prototypes on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

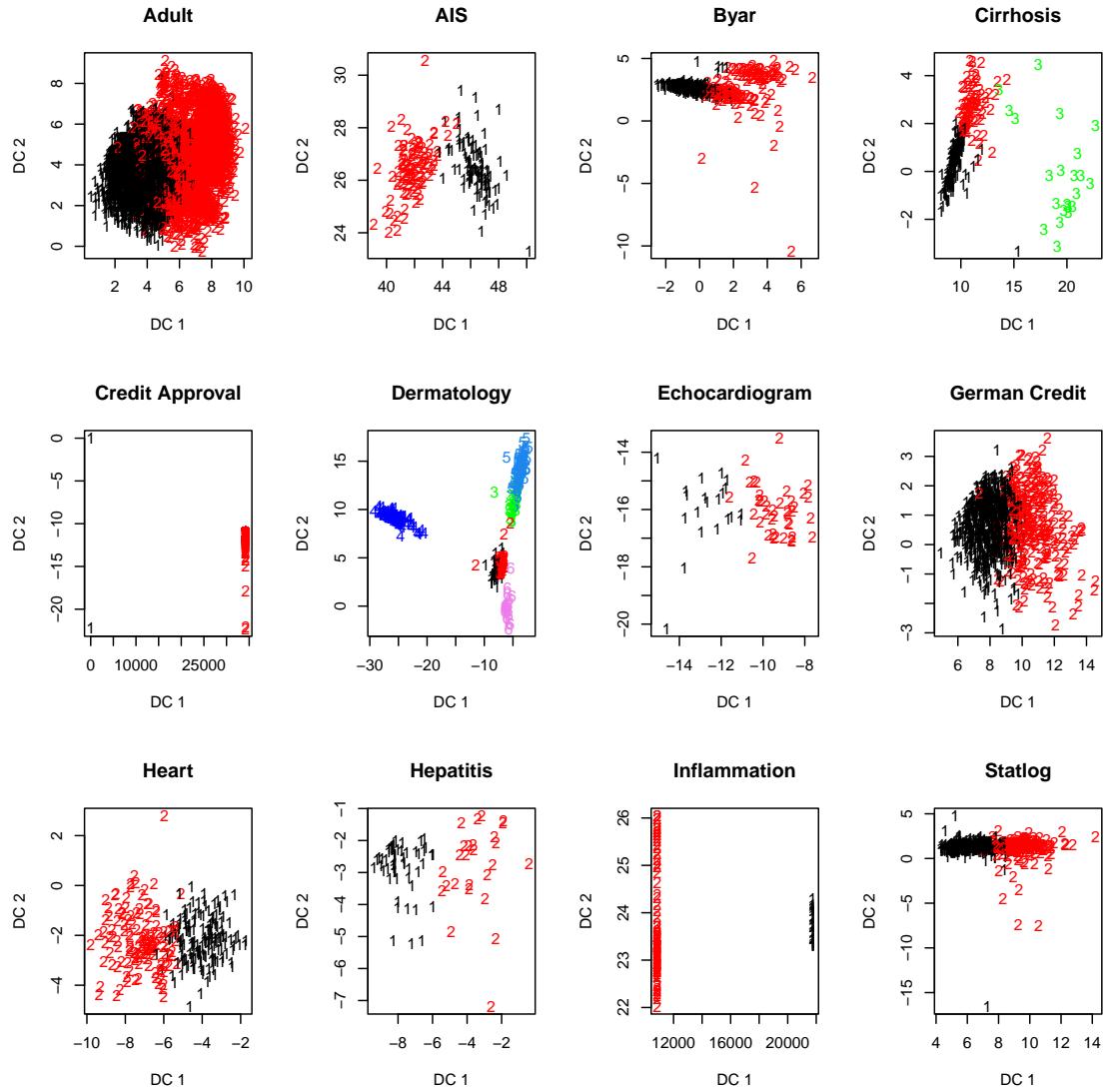


Figure B.8: Projection of cluster partitions obtained using FAMD/K-Means on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

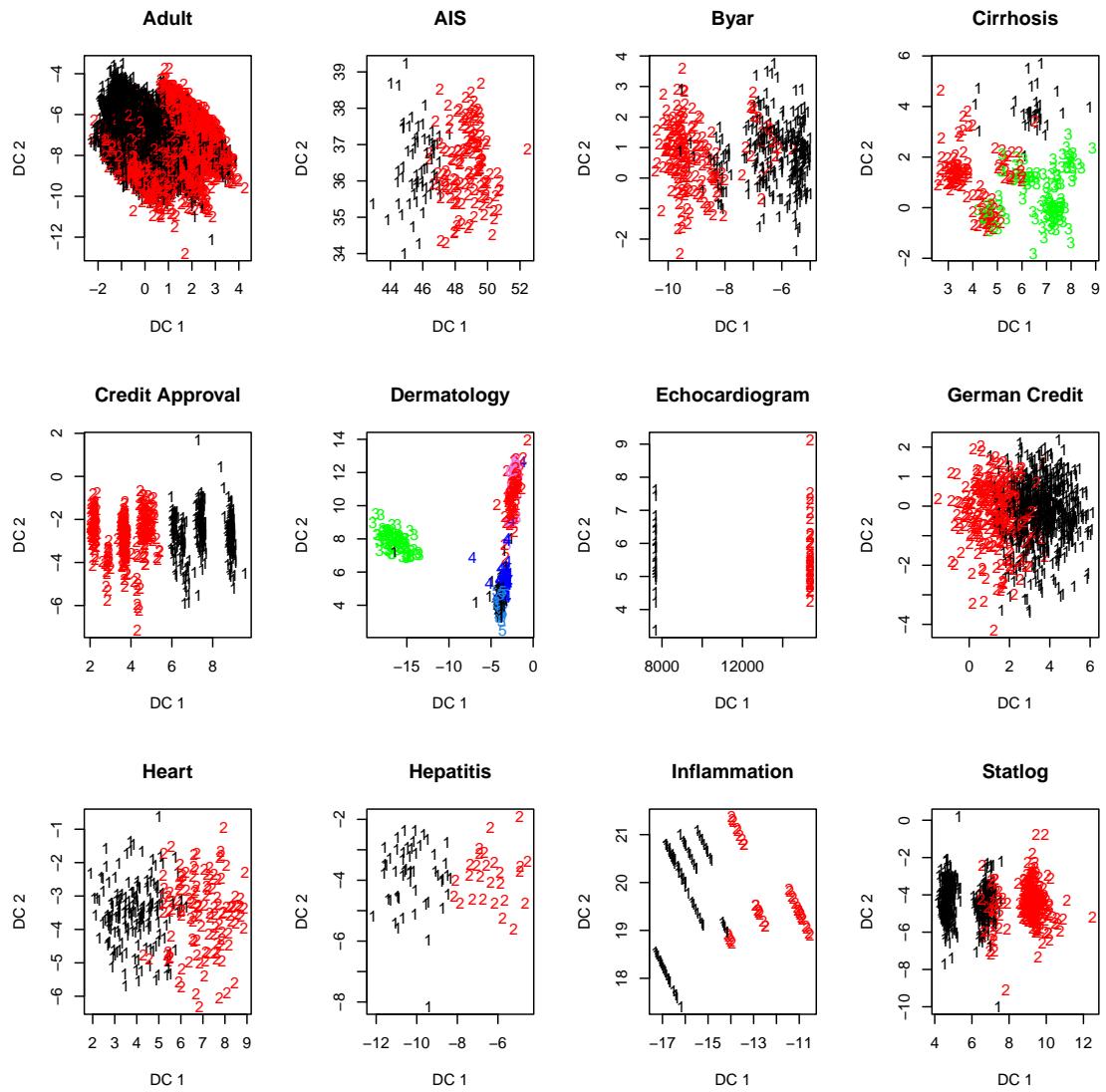


Figure B.9: Projection of cluster partitions obtained using Gower/PAM on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

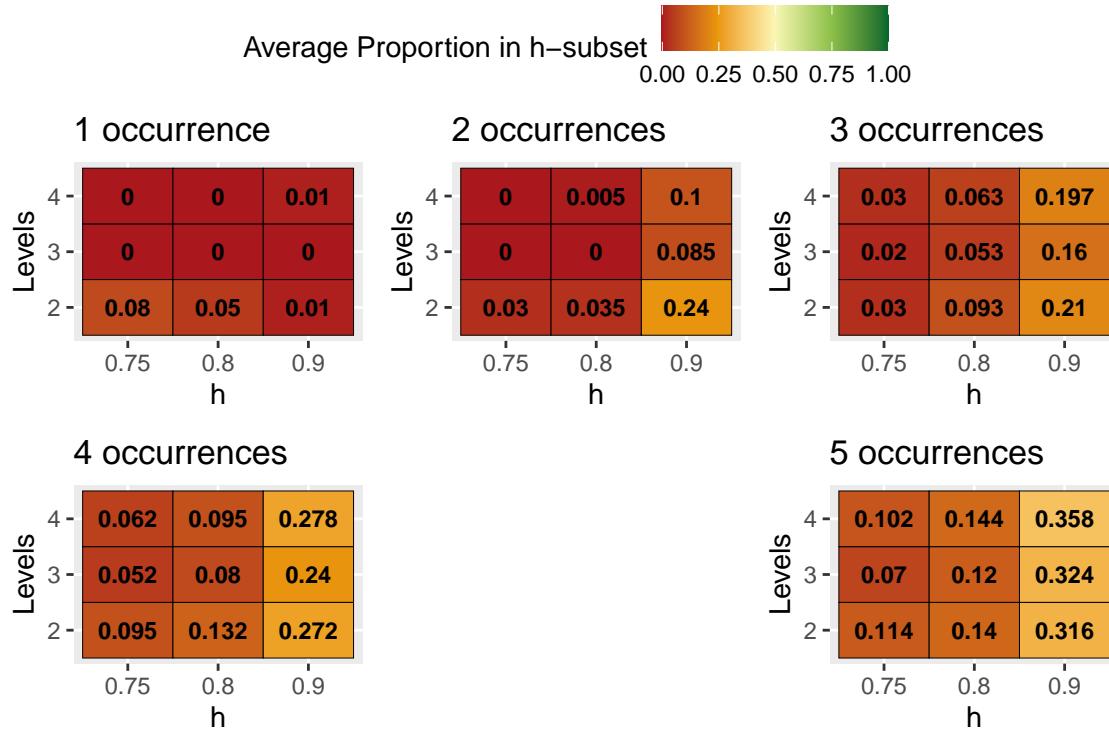


Figure B.10: Average proportion of observations possessing infrequent ordinal levels included in the h-subset for varying h, number of categorical levels, and number of occurrences.

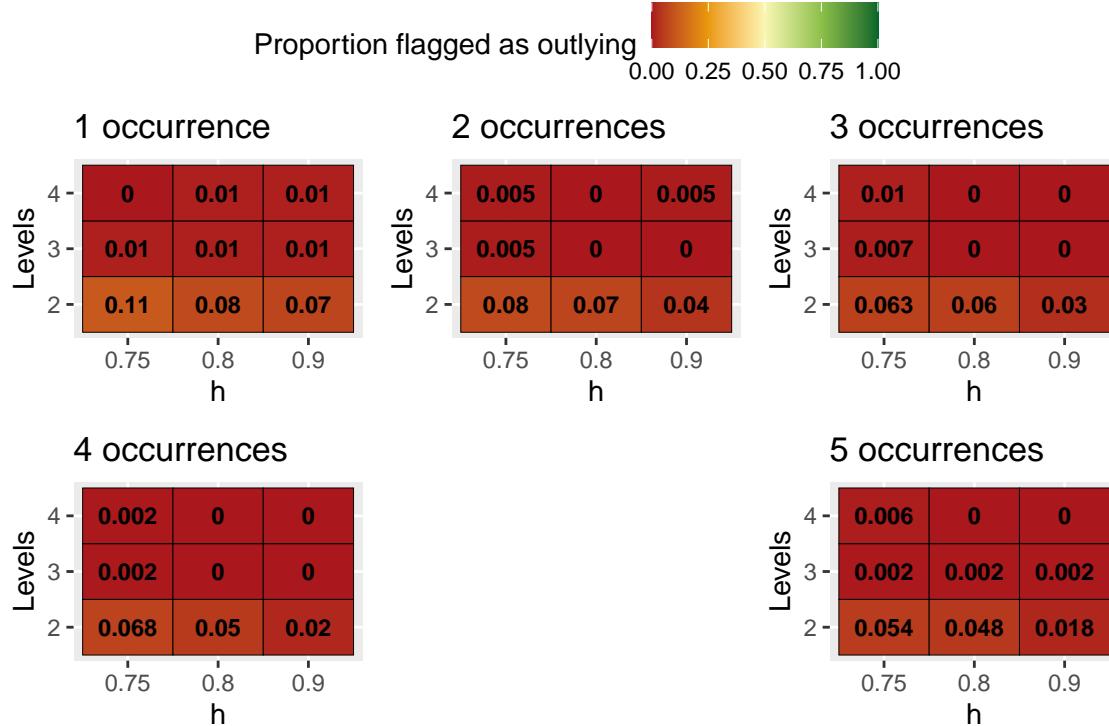


Figure B.11: Average proportion of observations possessing infrequent ordinal levels flagged as outlying for varying h, number of categorical levels, and number of occurrences.

	Shift Contamination			Cluster Contamination		
	$h = 3n/4$	$h = 4n/5$	$h = 9n/10$	$h = 3n/4$	$h = 4n/5$	$h = 9n/10$
$\epsilon = 0.05$ (25 outliers)	4.00	1.00	0.00	4.00	1.00	0.00
$\epsilon = 0.10$ (50 outliers)	15.00	7.00	1.00	15.00	7.00	1.00

Table B.1: Median number of inliers flagged as outlying for mixed continuous-ordinal data sets with shift and cluster contamination and varying values of ϵ and h . The lowest numbers of inliers flagged as outlying for each ϵ value appear in bold.

Variable	Number of levels	Levels
room_type	3	Shared room, Private room, Entire home/apartment
room_shared	2	True, False
room_private	2	True, False
person_capacity	5	2, 3, 4, 5, 6
host_is_superhost	2	True, False
multi	2	True, False
biz	2	True, False
cleanliness_rating	9	2, 3, 4, 5, 6, 7, 8, 9, 10 [0, 50), [50, 60), [60, 70), [70, 80), [80, 90), [90, 100]
guest_satisfaction_overall	6	[0, 50), [50, 60), [60, 70), [70, 80), [80, 90), [90, 100]
bedrooms	7	0, 1, 2, 3, 4, 5, 8

Table B.2: Description of ordinal variables in the London Airbnb data set.

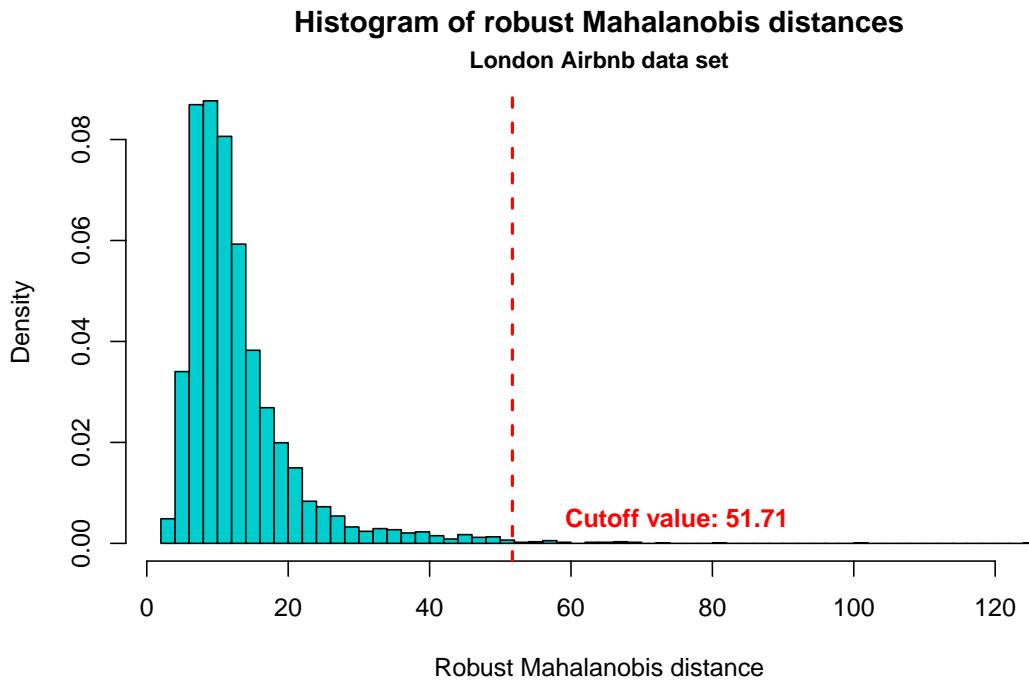


Figure B.12: Histogram of robust Mahalanobis distances on the Lonodon Airbnb data set. The red dashed line denotes the cutoff value, based on a significance level of $\alpha = 0.05$, indicating outlying observations.

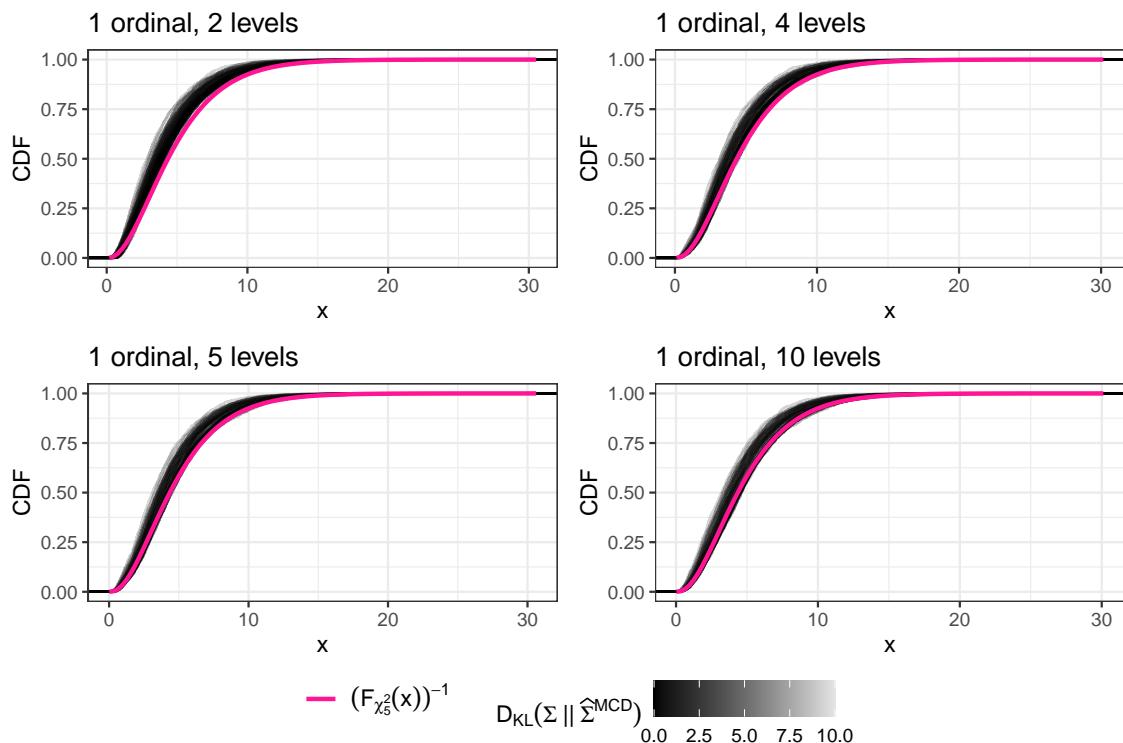


Figure B.13: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and one ordinal variable with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_5 distribution.

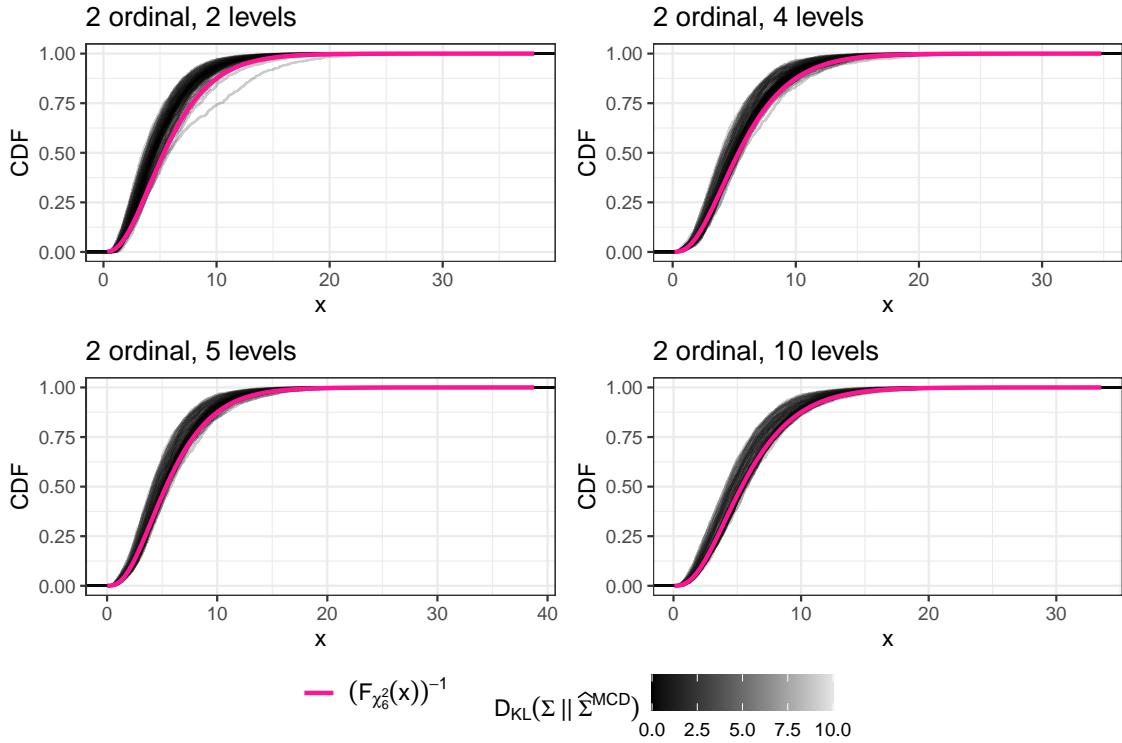


Figure B.14: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and two ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ_6^2 distribution.

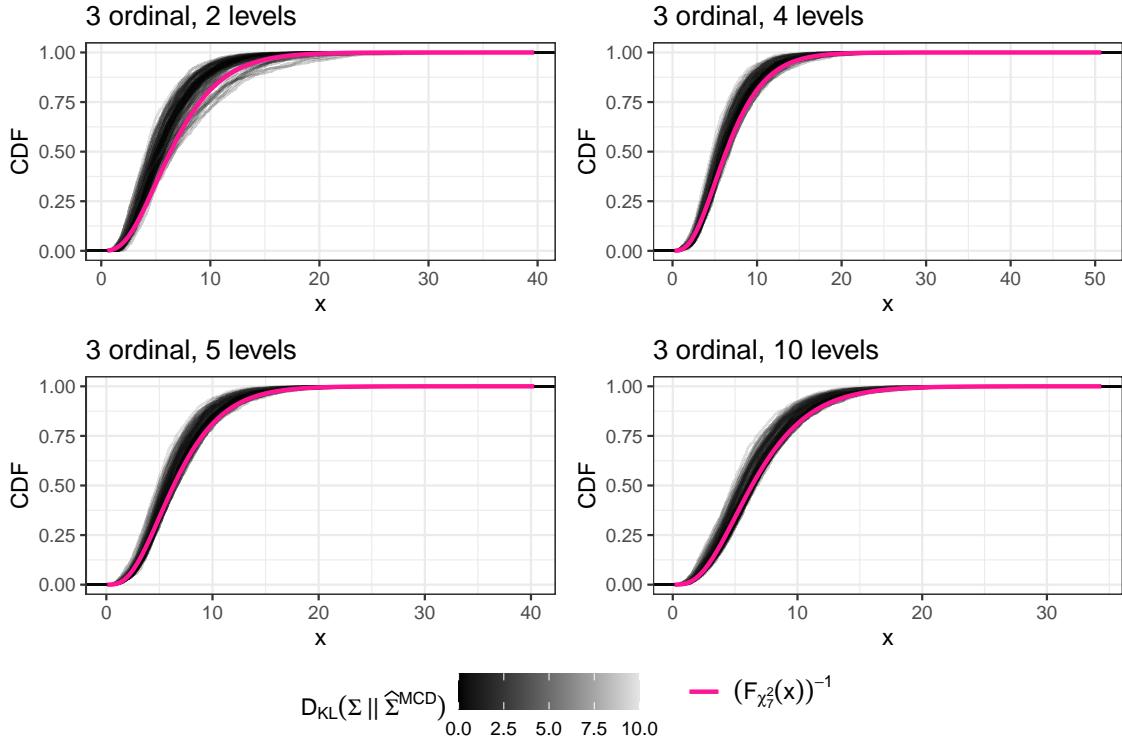


Figure B.15: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and three ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ_7^2 distribution.

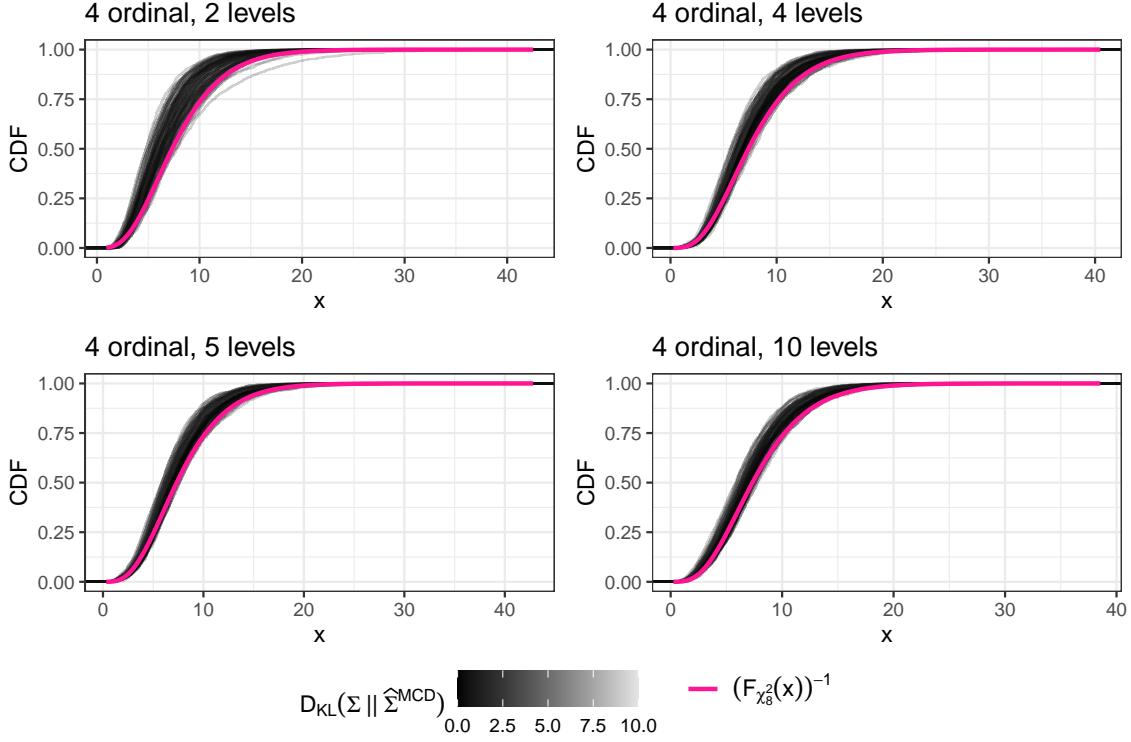


Figure B.16: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and four ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_8 distribution.

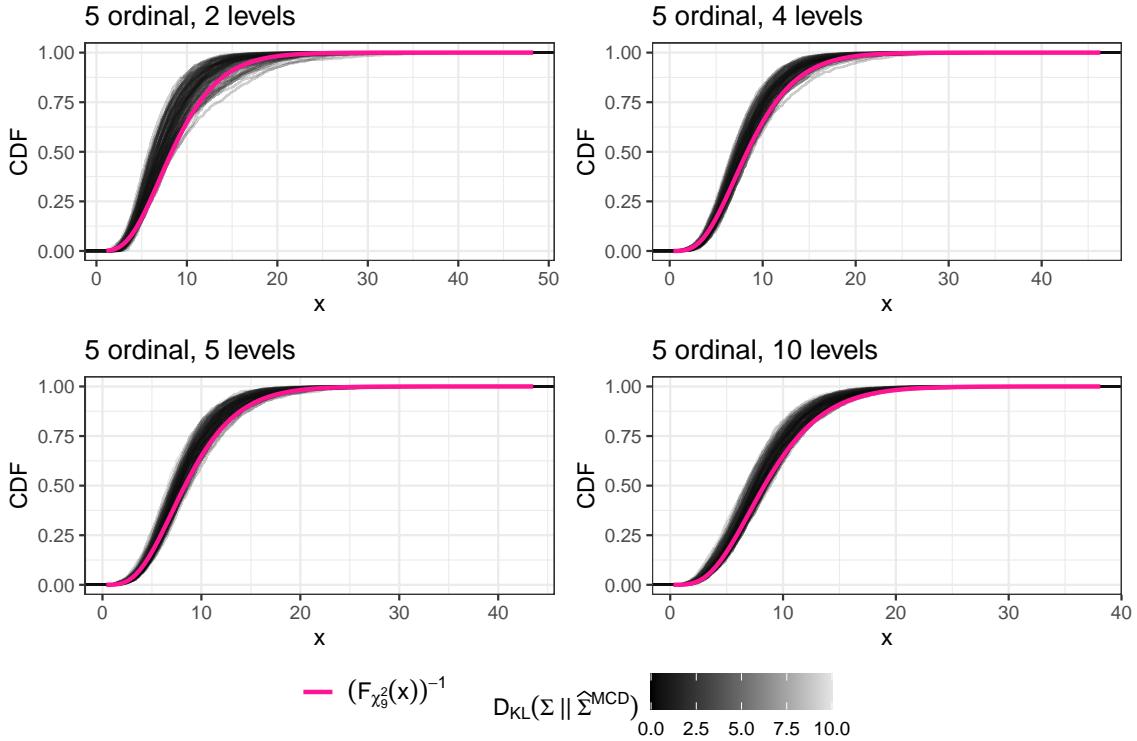


Figure B.17: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and five ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_9 distribution.

	Quantile	2 levels	4 levels	5 levels	10 levels
1 ordinal ($p = 5$)	0.950	1.11	0.74	0.68	0.38
	0.955	1.14	0.79	0.66	0.40
	0.960	1.18	0.78	0.71	0.42
	0.965	1.25	0.80	0.78	0.48
	0.970	1.23	0.92	0.80	0.48
	0.975	1.40	0.94	0.90	0.59
	0.980	1.36	0.98	0.93	0.70
	0.985	1.38	0.97	1.02	0.73
	0.990	1.55	1.12	1.02	0.86
	0.995	1.62	1.09	0.97	0.86
2 ordinal ($p = 6$)	0.950	1.94	1.23	1.06	0.70
	0.955	1.98	1.23	1.05	0.69
	0.960	2.03	1.24	1.07	0.75
	0.965	2.05	1.26	1.07	0.68
	0.970	2.12	1.28	1.11	0.77
	0.975	2.22	1.34	1.20	0.73
	0.980	2.19	1.40	1.18	0.76
	0.985	2.25	1.48	1.27	0.83
	0.990	2.39	1.69	1.33	0.94
	0.995	2.27	1.67	1.40	0.96
3 ordinal ($p = 7$)	0.950	2.20	1.26	1.07	1.01
	0.955	2.27	1.32	1.13	1.02
	0.960	2.30	1.36	1.17	1.05
	0.965	2.30	1.39	1.25	1.07
	0.970	2.30	1.43	1.28	1.06
	0.975	2.32	1.48	1.32	1.24
	0.980	2.46	1.53	1.43	1.26
	0.985	2.60	1.52	1.30	1.34
	0.990	2.62	1.61	1.65	1.26
	0.995	2.50	1.63	1.77	1.36
4 ordinal ($p = 8$)	0.950	2.60	1.62	1.52	1.60
	0.955	2.61	1.63	1.55	1.57
	0.960	2.70	1.63	1.62	1.59
	0.965	2.80	1.58	1.57	1.62
	0.970	2.81	1.68	1.52	1.67
	0.975	2.97	1.79	1.62	1.64
	0.980	2.83	1.88	1.79	1.76
	0.985	2.82	1.96	1.76	1.89
	0.990	2.78	1.95	1.76	2.01
	0.995	2.61	1.75	1.76	2.00

Table B.3: Difference between large quantiles of the χ_p^2 distribution ($p = 5, 6, 7, 8$) and the median empirical quantiles of robust distances based on a hundred data sets. Smallest difference for each quantile appear in bold.

Quantile	2 levels	4 levels	5 levels	10 levels
5 ordinal ($p = 9$)	0.950	2.51	1.77	1.73
	0.955	2.47	1.78	1.79
	0.960	2.59	1.86	1.84
	0.965	2.75	1.82	1.92
	0.970	2.64	1.86	1.95
	0.975	2.67	1.99	1.98
	0.980	2.65	2.05	1.90
	0.985	2.80	2.10	1.99
	0.990	2.94	1.86	1.98
	0.995	2.56	1.62	1.95

Table B.4: Difference between large quantiles of the χ_p^2 distribution ($p = 9$) and the median empirical quantiles of robust distances based on a hundred data sets. Smallest difference for each quantile appear in bold.