Appendix B

Additional Figures & Tables

Factor	Values	`MixSim` argument
Cluster Overlap	0.01, 0.05, 0.10, 0.15, 0.20	`BarOmega`
Cluster sphericity	TRUE, FALSE	`sph`
Number of clusters	3, 5, 8	`K`
Number of observations	100, 600, 1000	`n`
Number of variables	8, 12, 16	`p`
Proportion of categorical variables	0.20, 0.50, 0.80	-
Cluster density	1, 0.01	`PiLow`

Table B.1: Summary of the factors included in benchmarking study and relevant `MixSim` arguments.

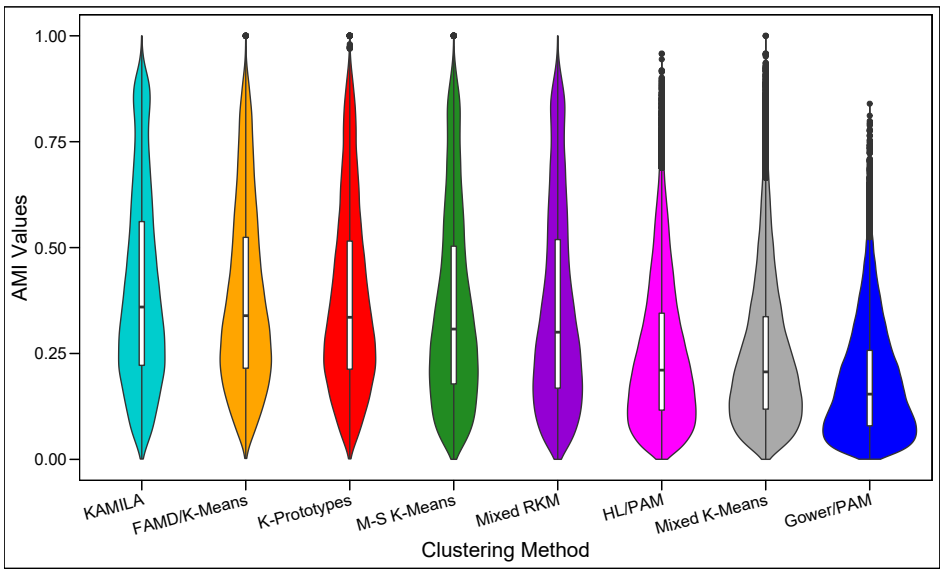

Figure B.1: Violin/box plots of Adjusted Mutual Information (AMI) values by method. Methods are sorted from left to right by decreasing mean AMI.

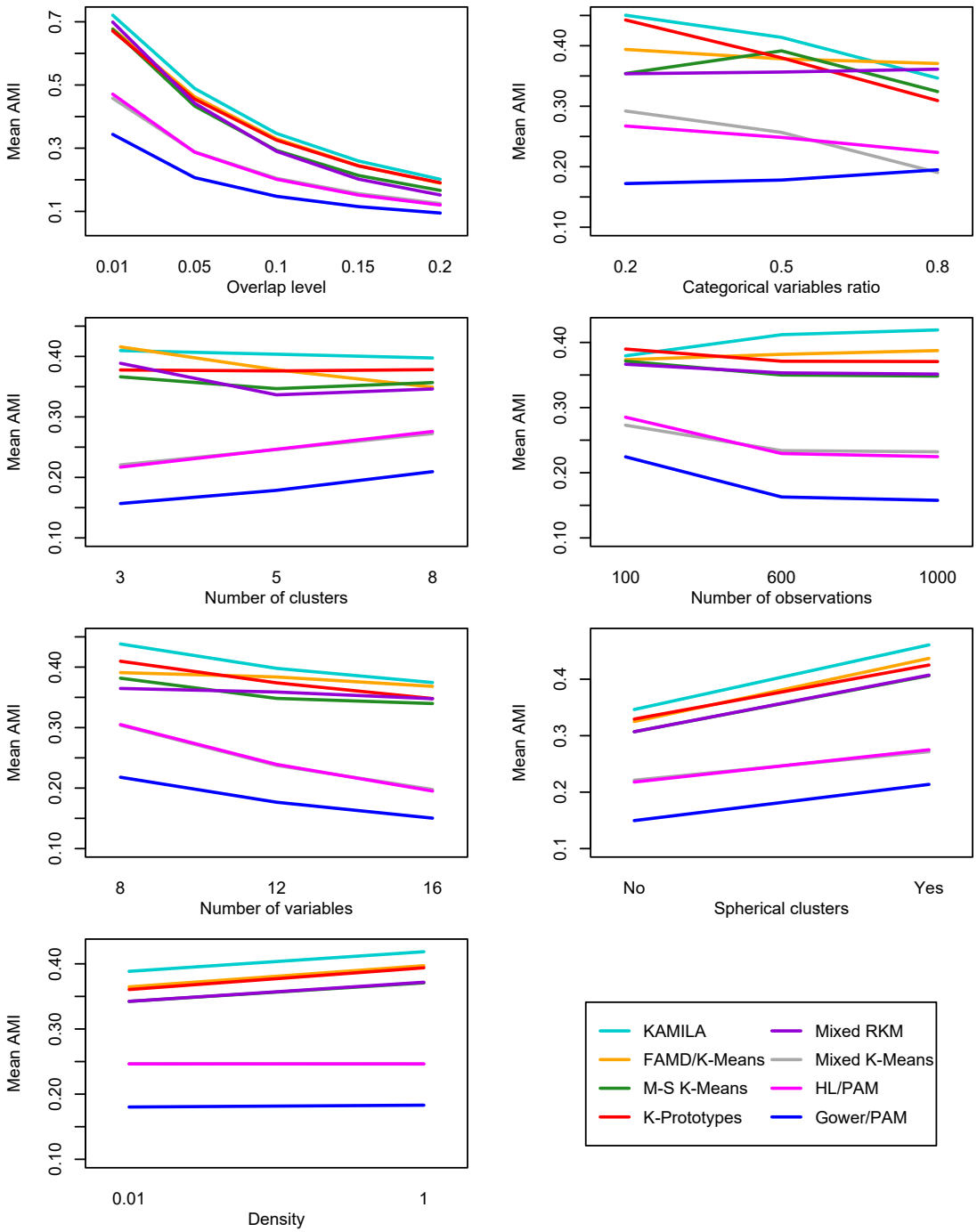

Figure B.2: Two-way interactions of method by overlap level, percentage of categorical variables, number of clusters, number of observations, number of variables, density, and cluster sphericity (mean AMI values). Subplots/factors are arranged, from left to right, by decreasing effect size η^2_p.

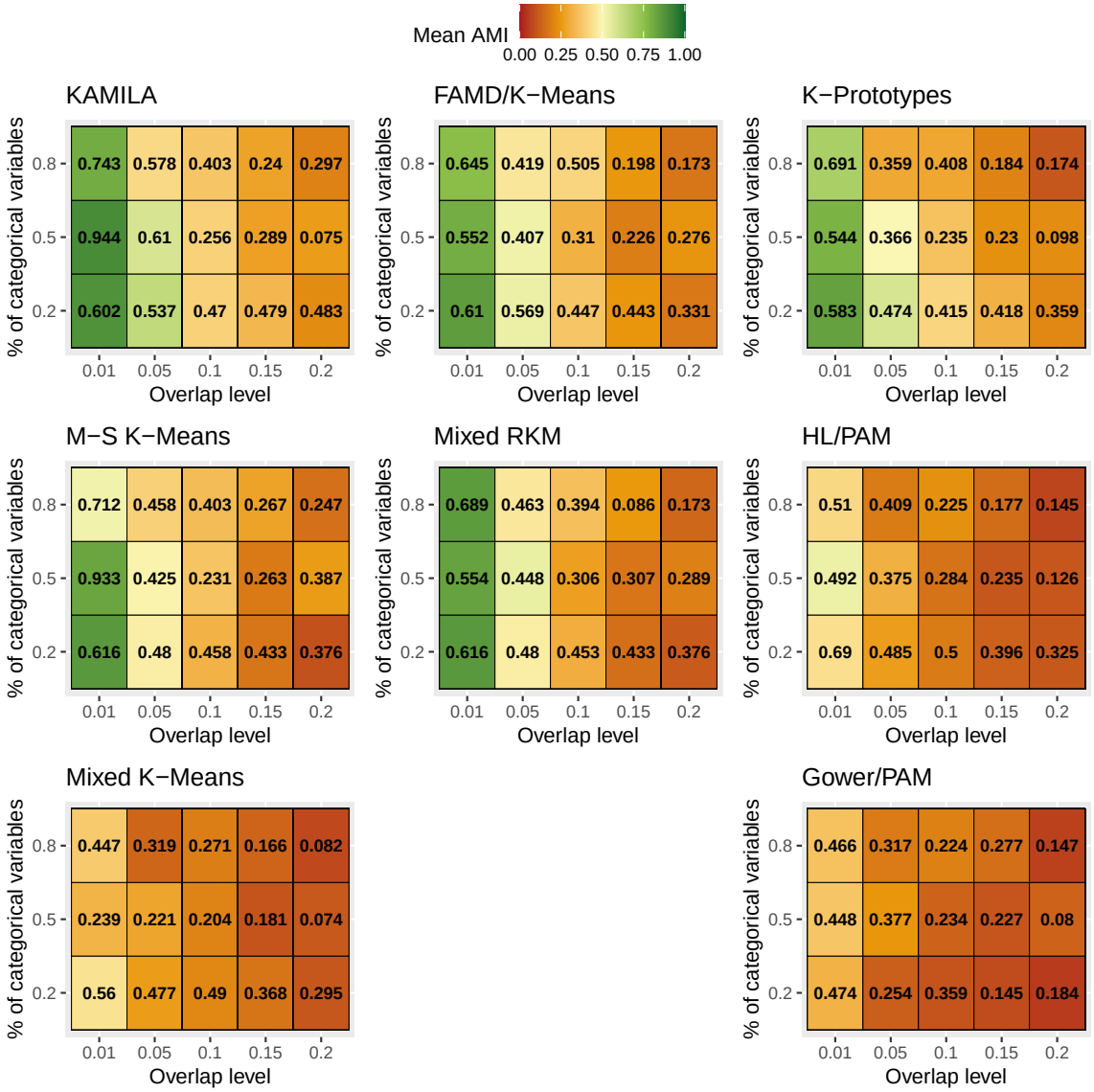

Figure B.3: Three-way interaction of method by overlap level and percentage of categorical variables. The numbers indicate the mean AMI for each combination of overlap level and percentage of categorical variables.

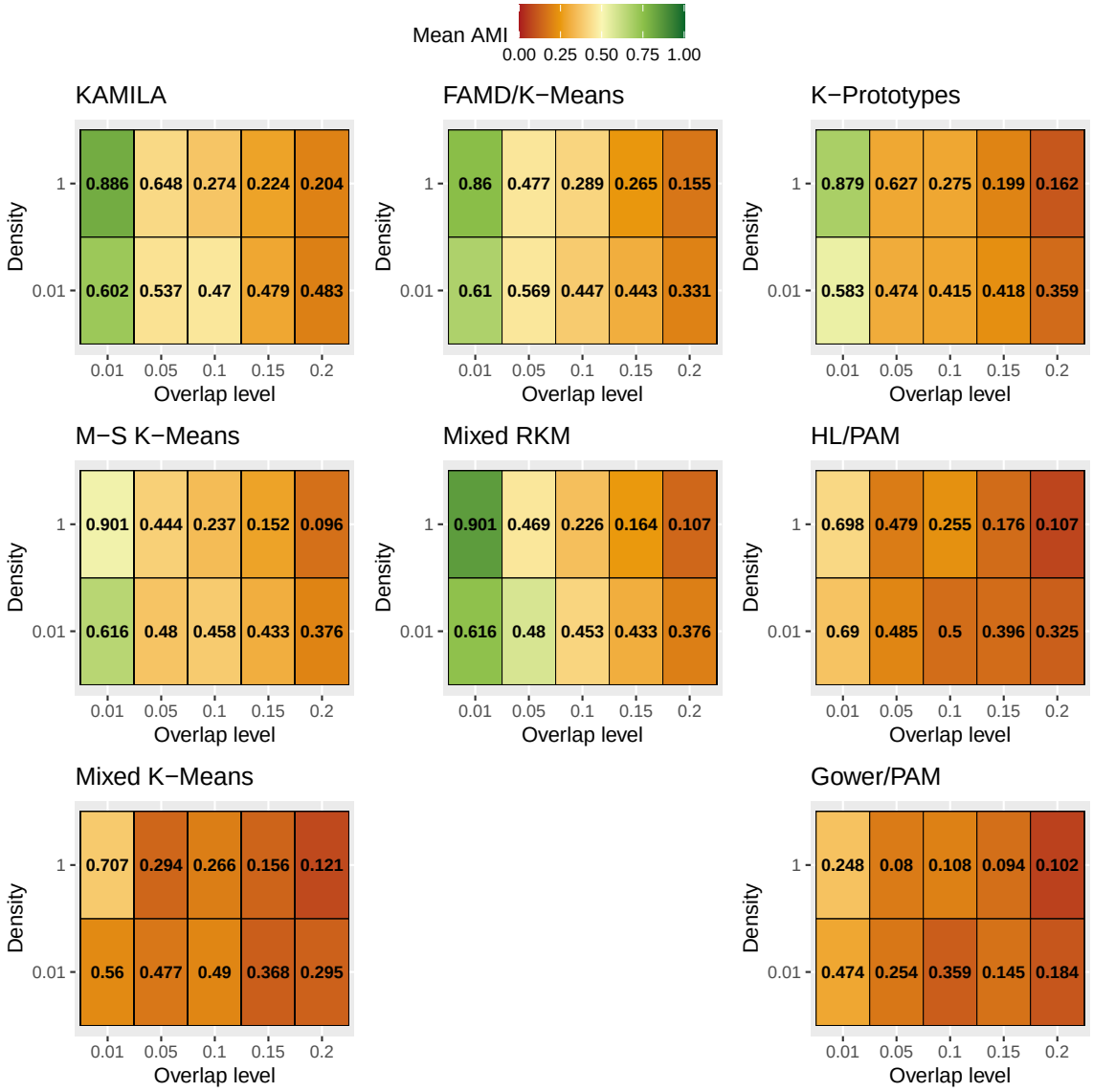

Figure B.4: Three-way interaction of method by overlap level and cluster density. The numbers indicate the mean AMI for each combination of overlap level and cluster density.

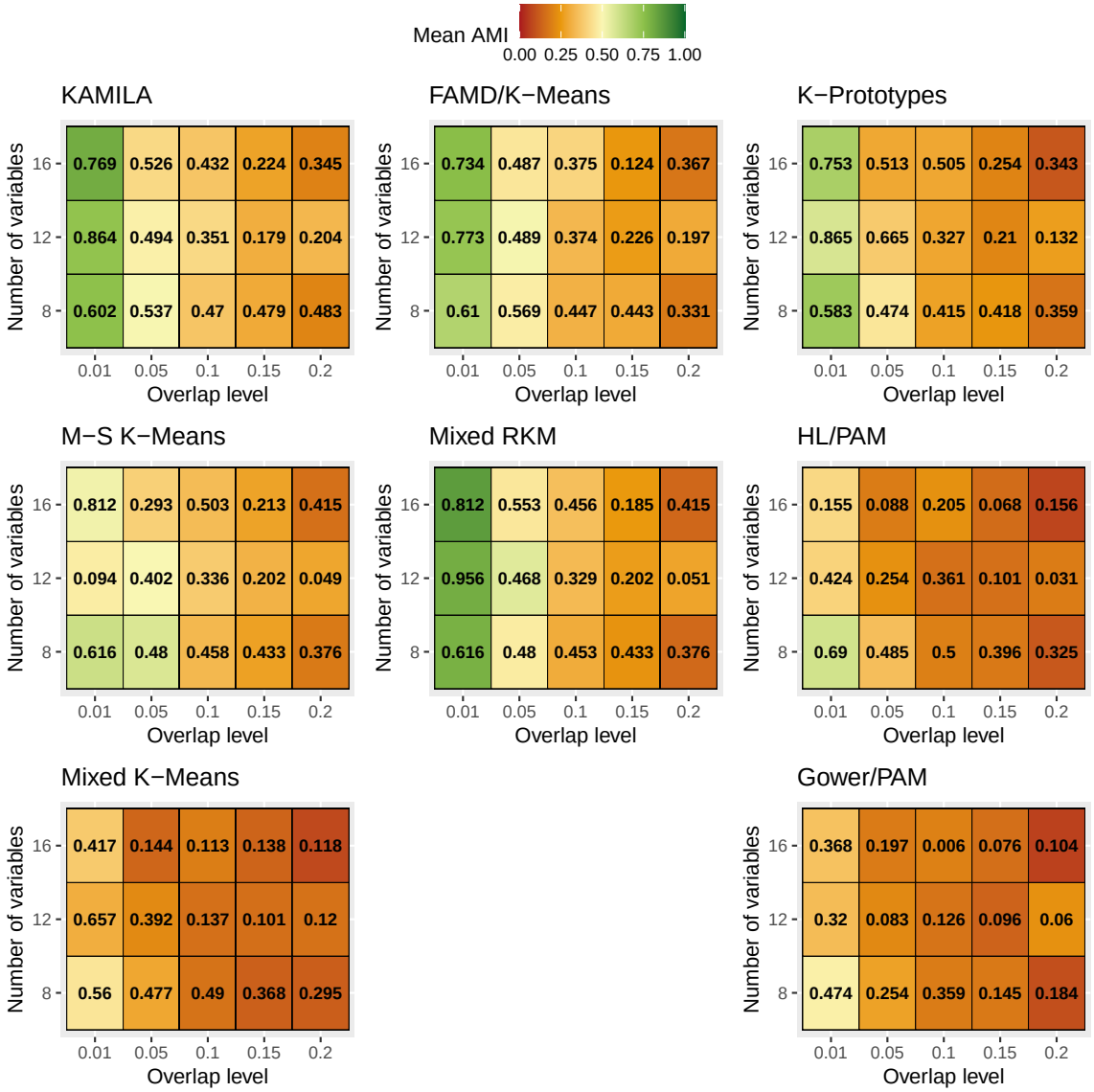

Figure B.5: Three-way interaction of method by overlap level and number of variables. The numbers indicate the mean AMI for each combination of overlap level and number of variables.

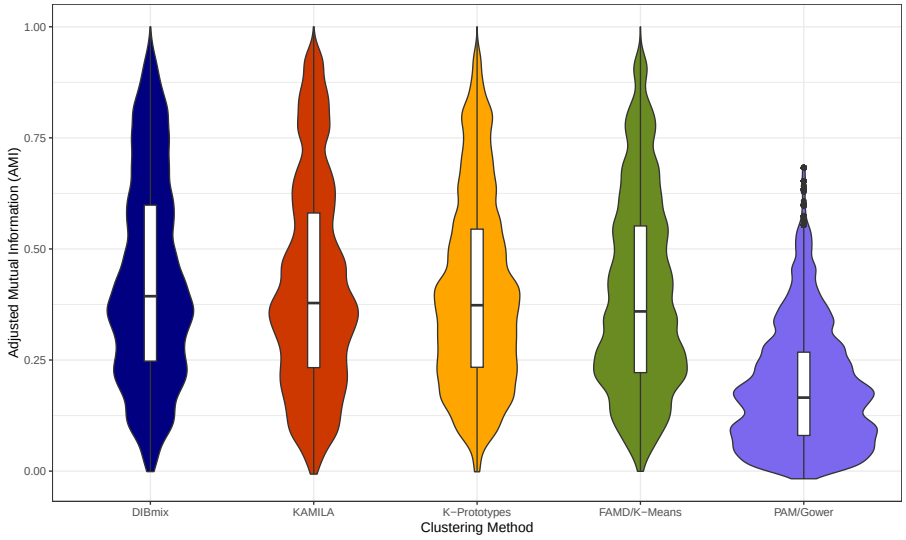

Figure B.6: Violin/box plots of Adjusted Mutual Information (AMI) values by method. Methods are sorted from left to right by decreasing mean AMI.

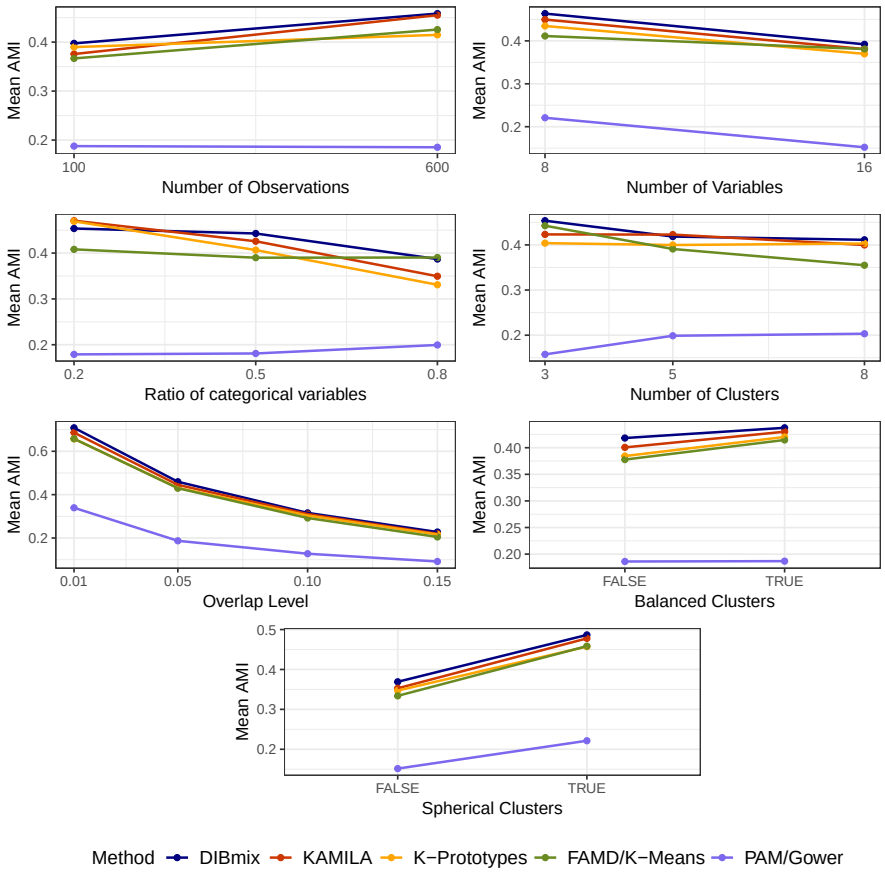

Figure B.7: Two-way interactions of method by overlap level, percentage of categorical variables, number of clusters, number of observations, number of variables, density, and cluster sphericity (mean AMI values).

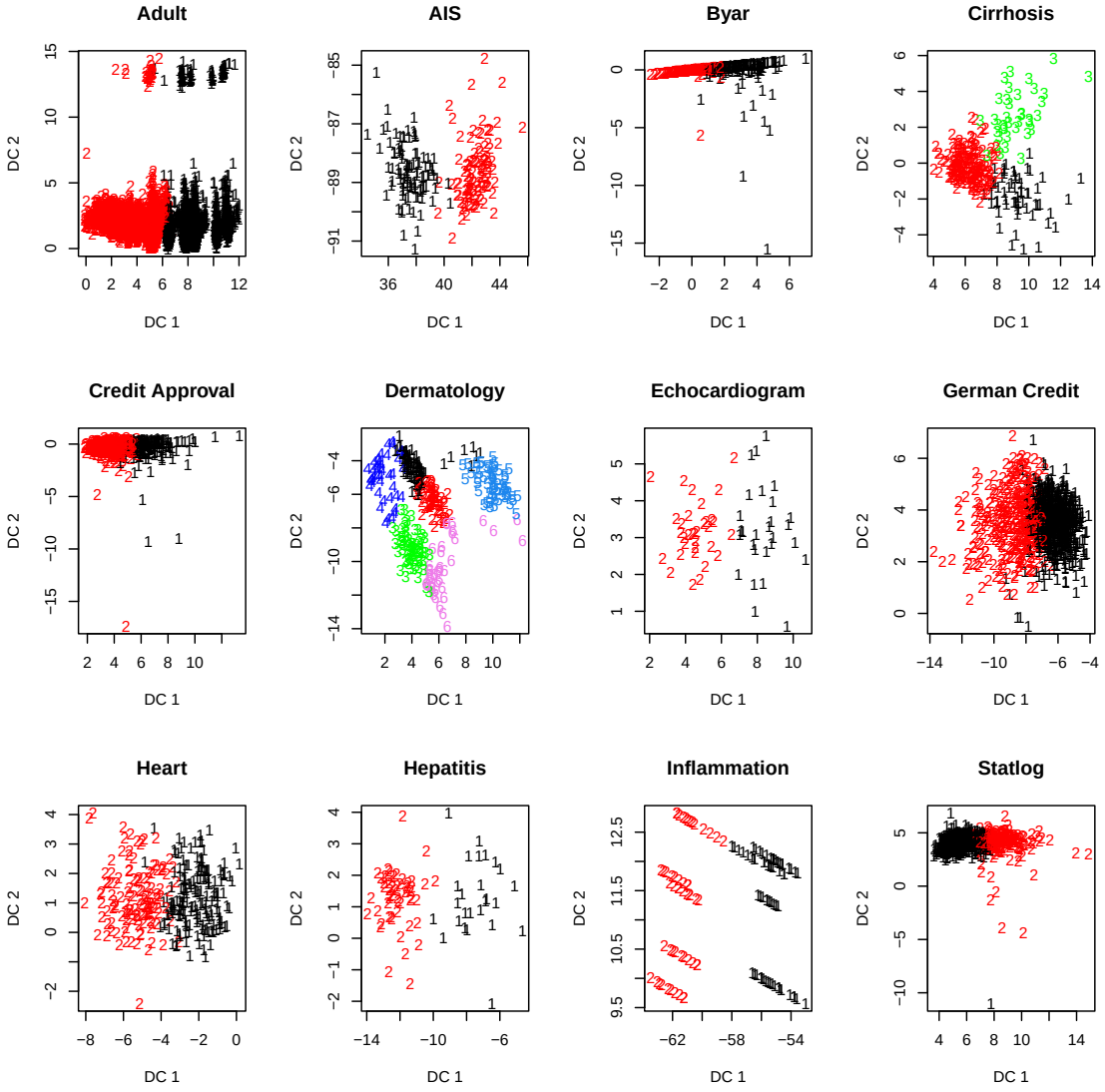

Figure B.8: Projection of cluster partitions obtained using KAMILA on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

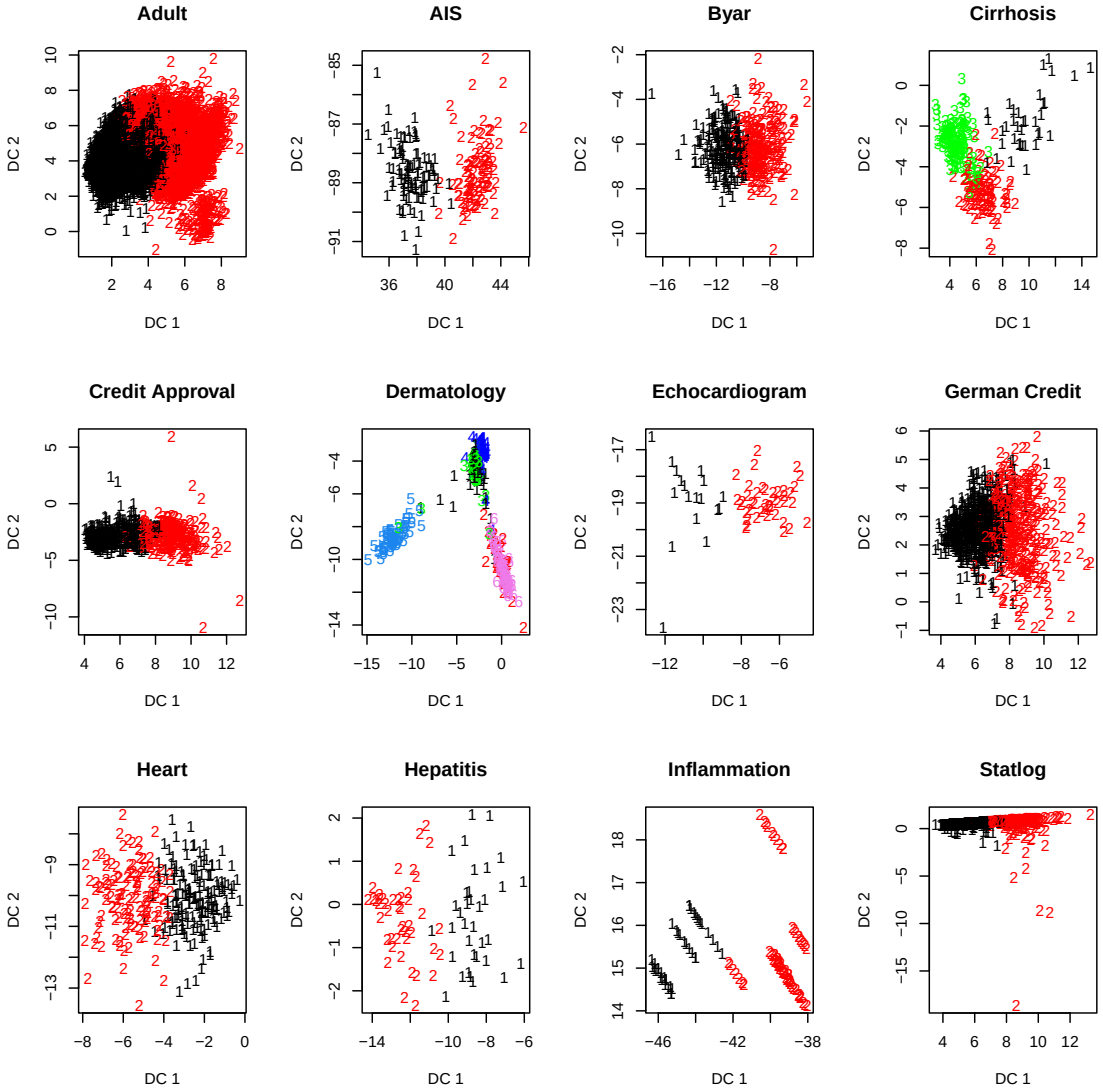

Figure B.9: Projection of cluster partitions obtained using K-Prototypes on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

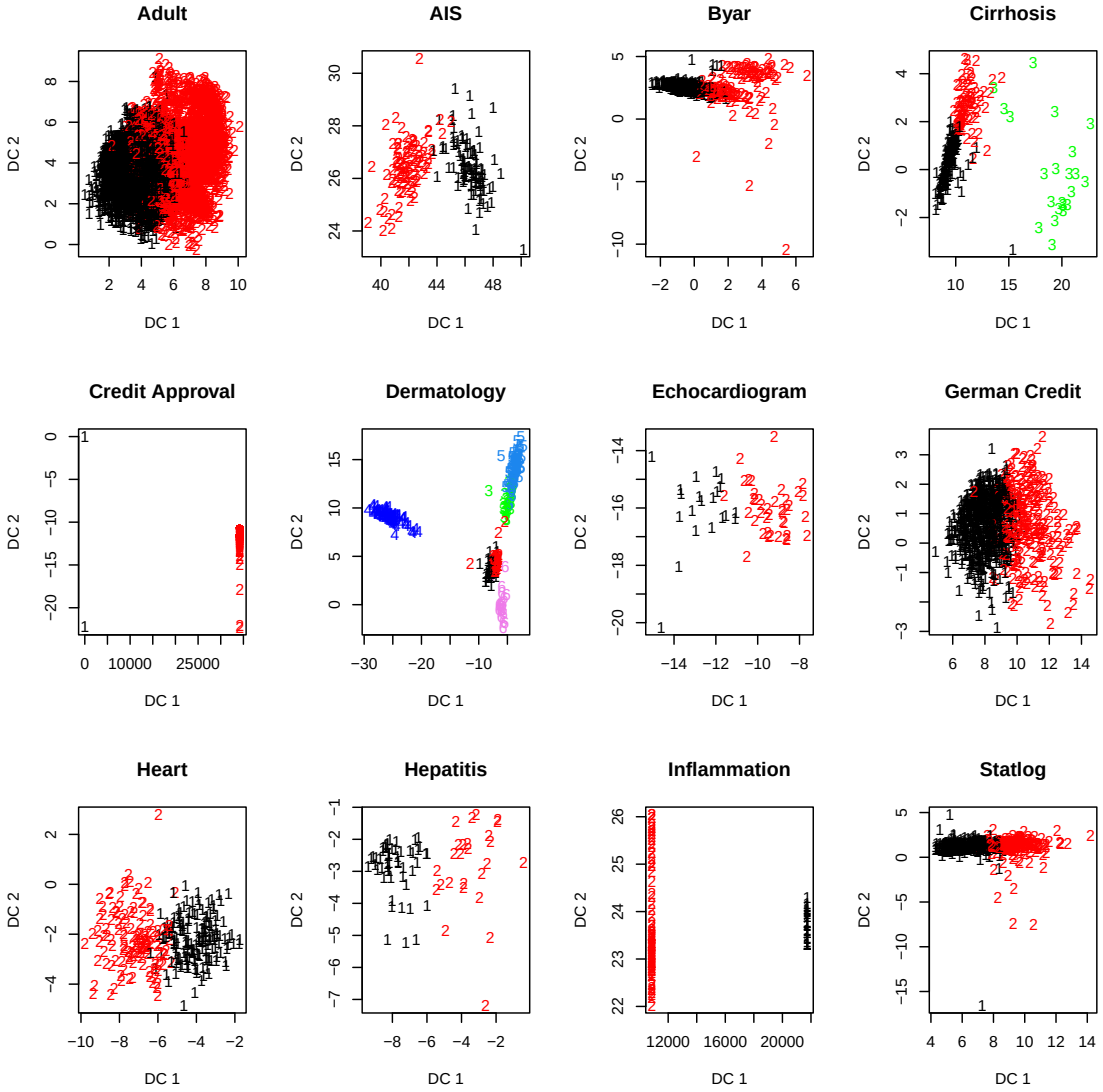

Figure B.10: Projection of cluster partitions obtained using FAMD/K-Means on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

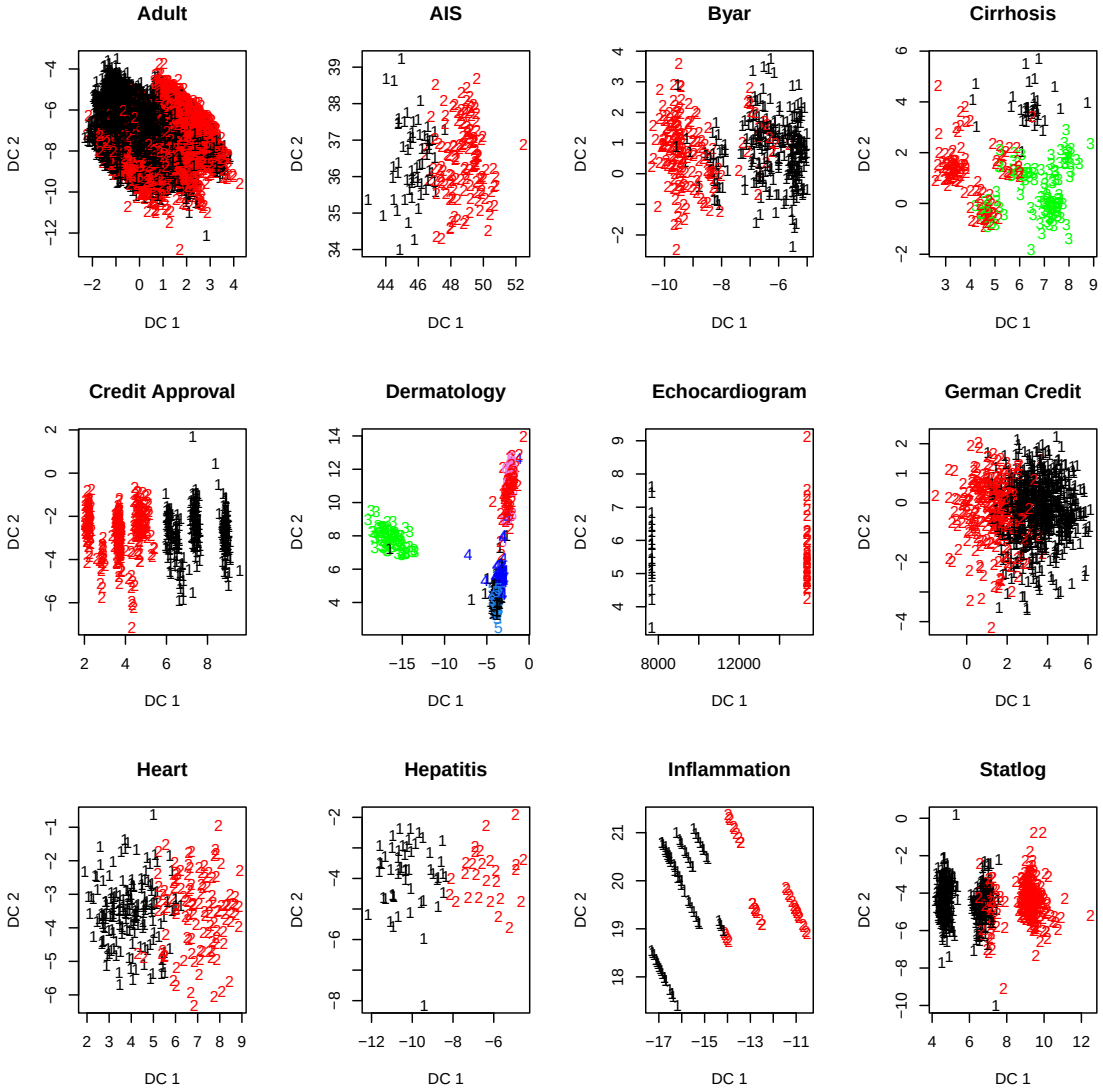

Figure B.11: Projection of cluster partitions obtained using Gower/PAM on the first two discriminant coordinates (DC1, DC2) for the twelve UCI data sets.

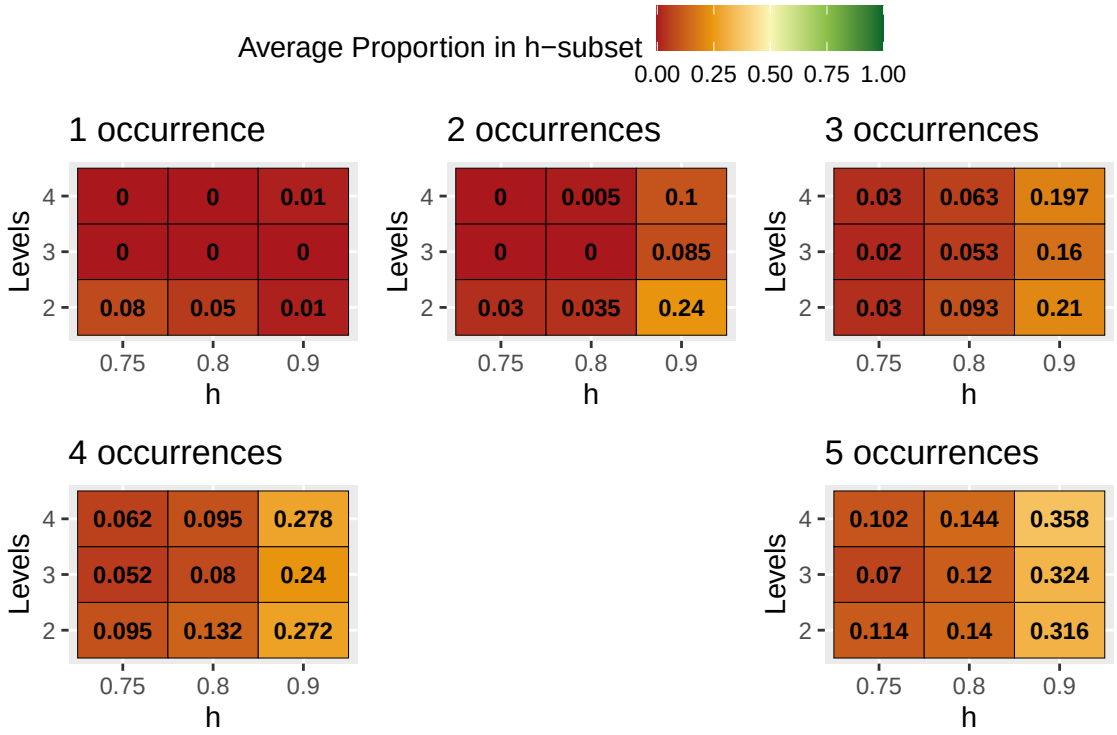

Figure B.12: Average proportion of observations possessing infrequent ordinal levels included in the h-subset for varying h, number of categorical levels, and number of occurrences.

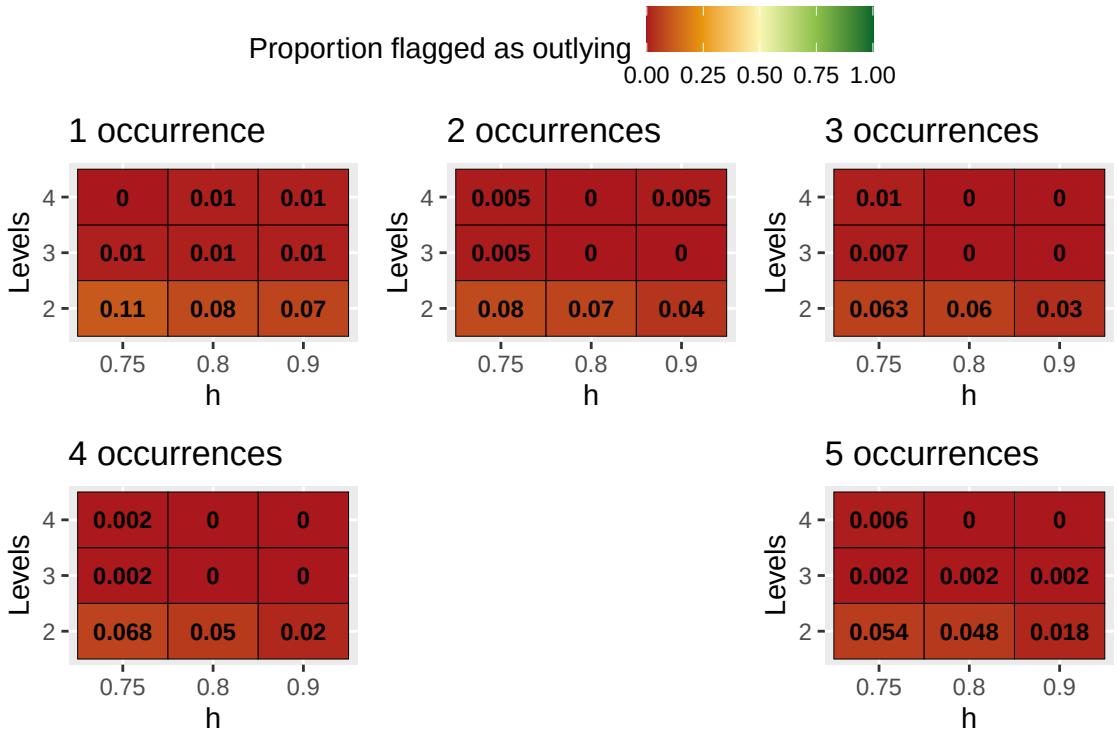

Figure B.13: Average proportion of observations possessing infrequent ordinal levels flagged as outlying for varying h, number of categorical levels, and number of occurrences.

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
`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.

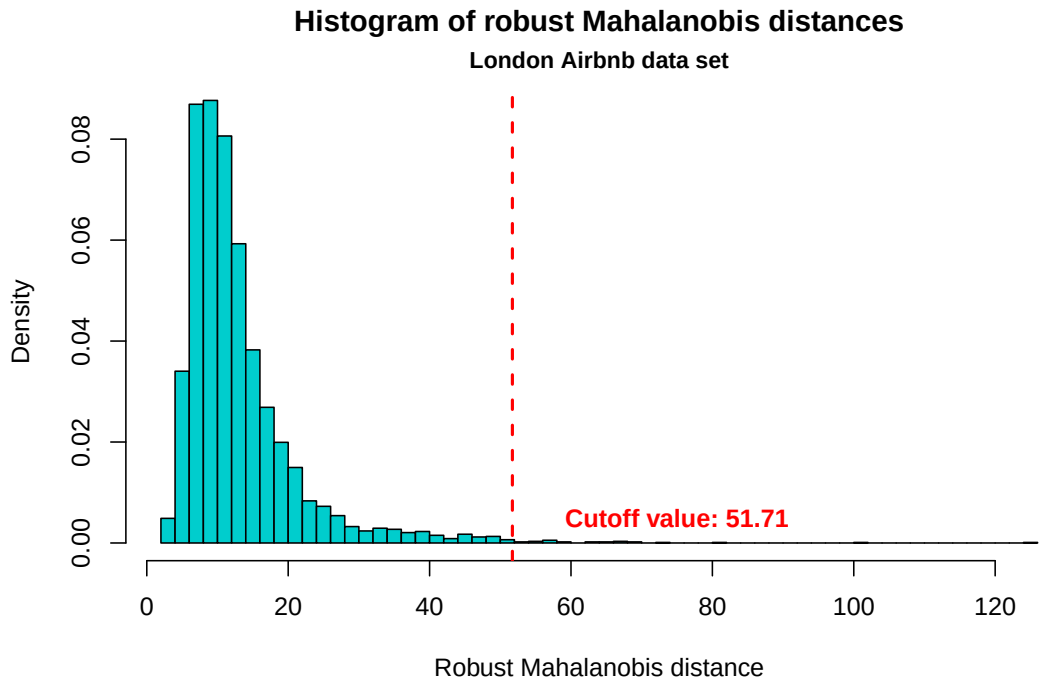

Figure B.14: Histogram of robust Mahalanobis distances on the Lonodon Airbnb data set. The red dashed line denotes the cuttoff value, based on a significance level of $\alpha = 0.05$, indicating outlying observations.

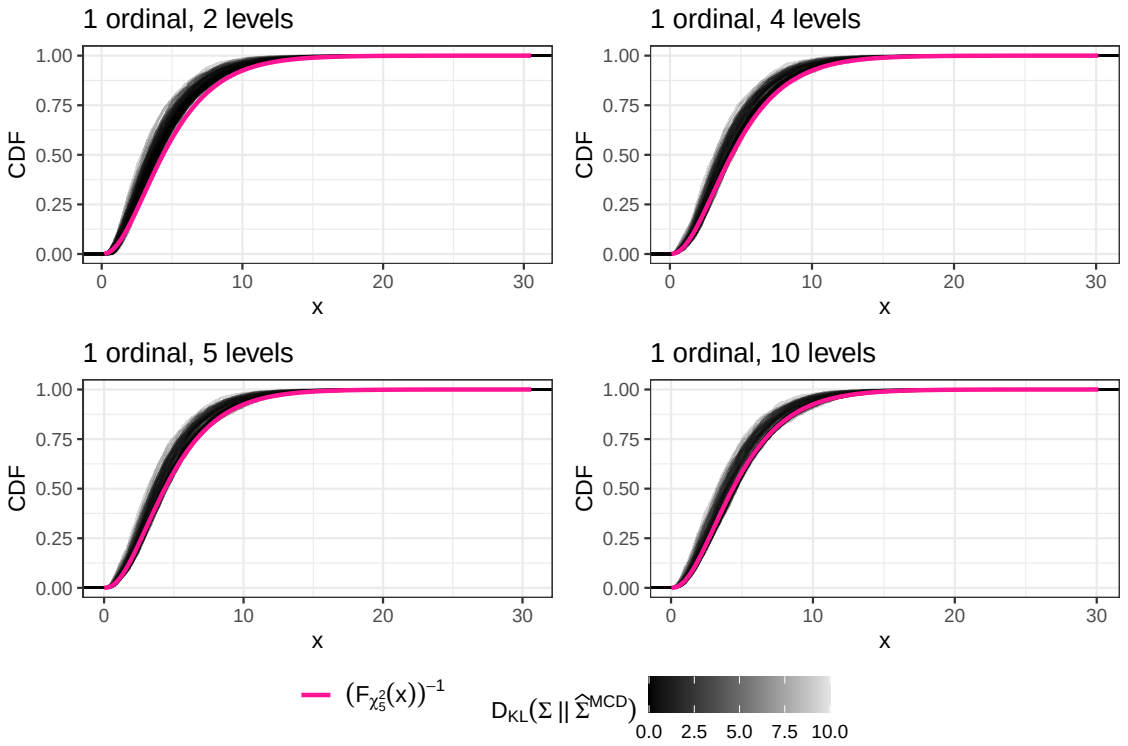

Figure B.15: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and one ordinal variable with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_5 distribution.

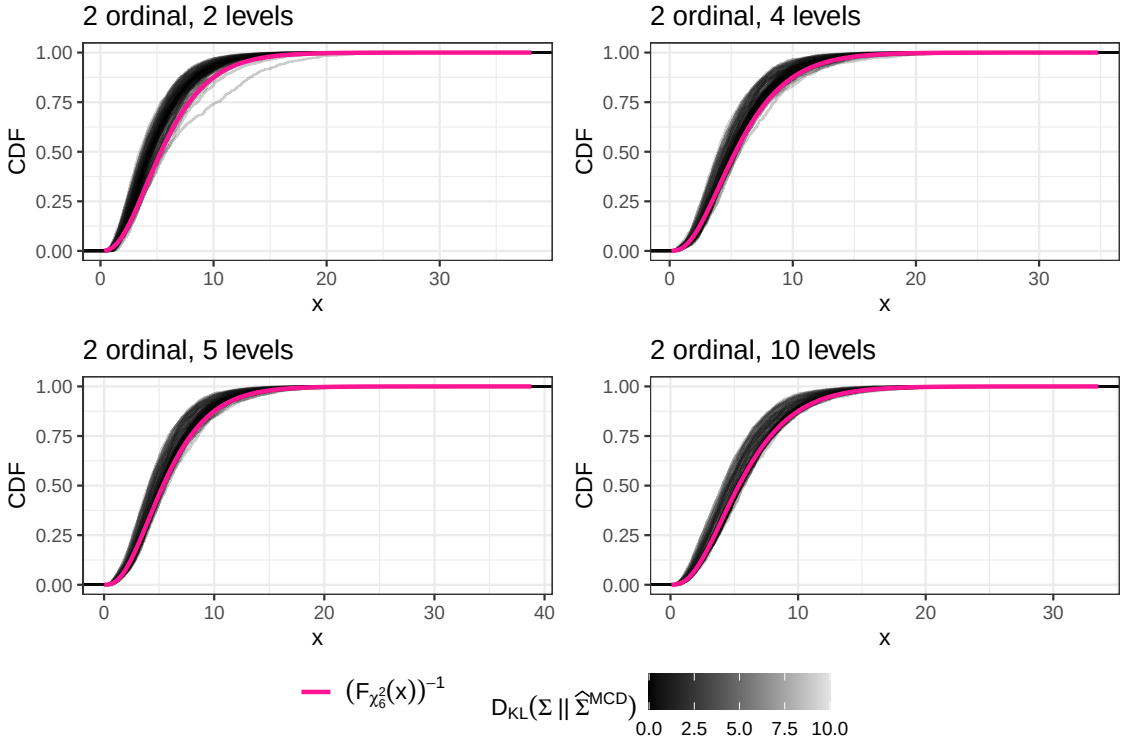

Figure B.16: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and two ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_6 distribution.

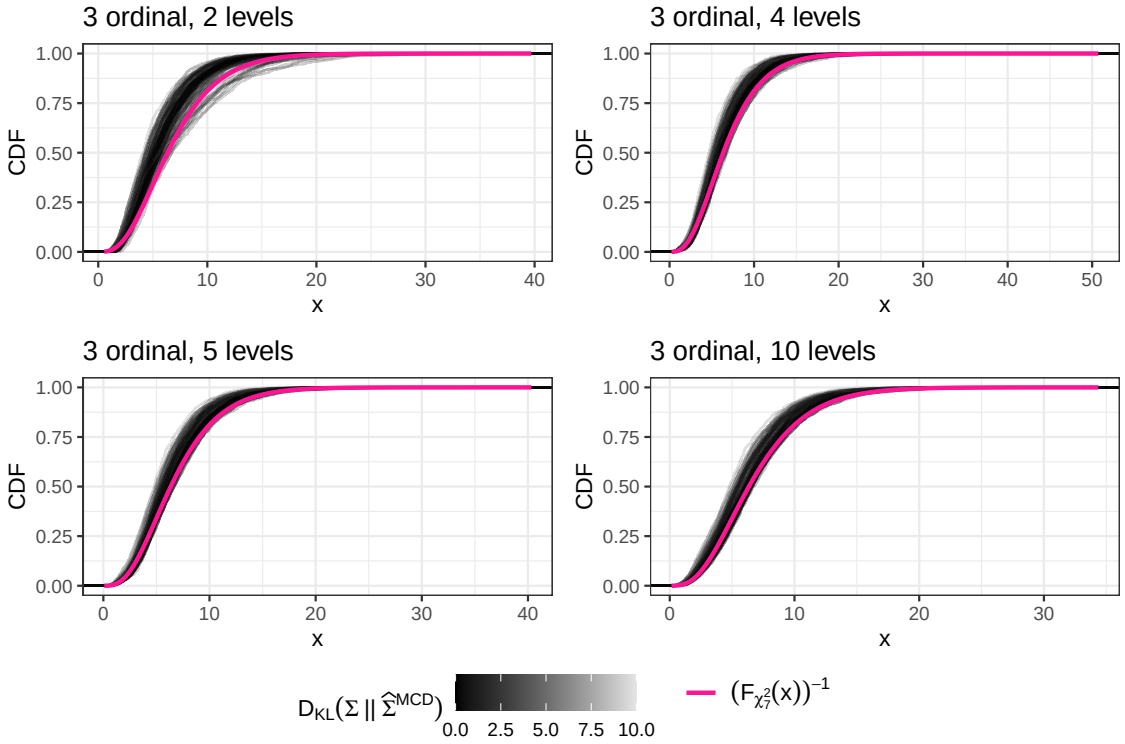

Figure B.17: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and three ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_7 distribution.

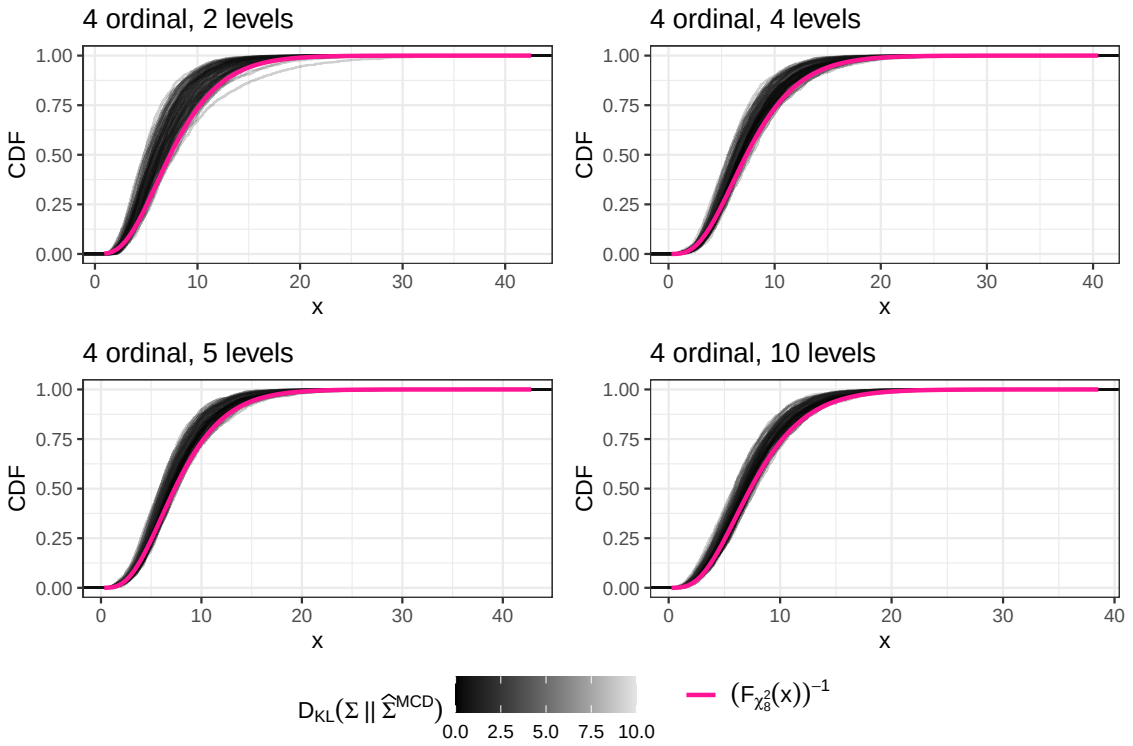

Figure B.18: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and four ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_8 distribution.

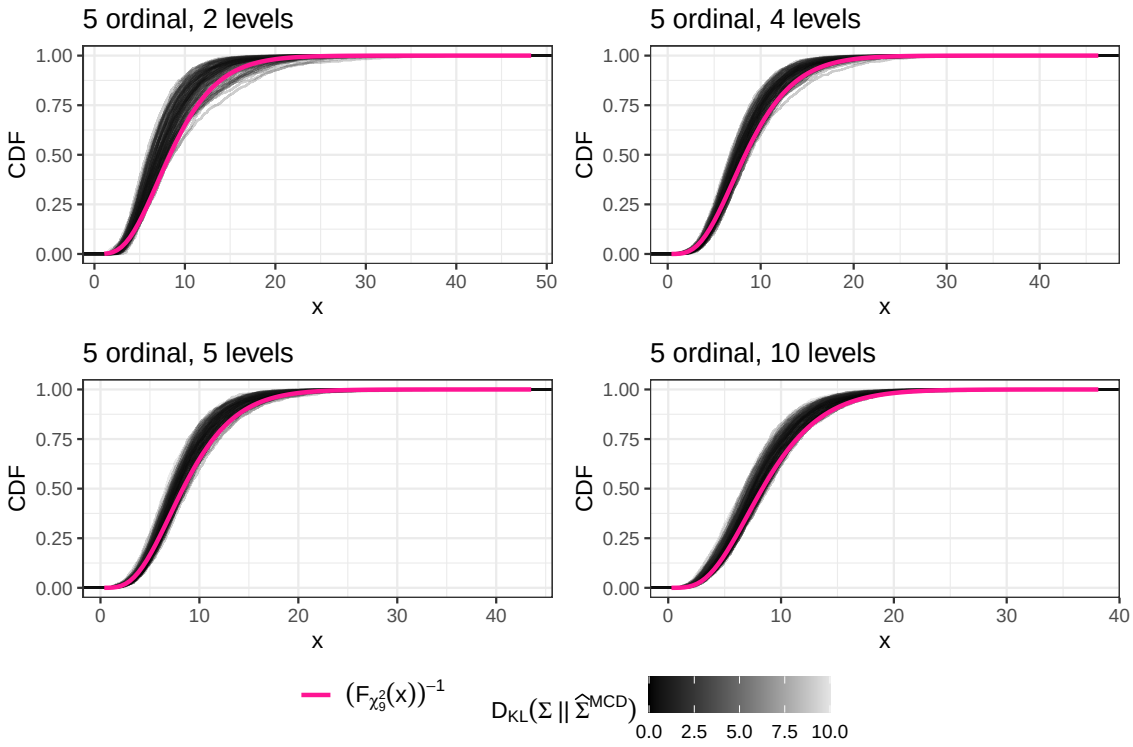

Figure B.19: Empirical cumulative distribution functions (CDFs) for robust distances on an outlier-free data set with 1000 observations, four continuous, and five ordinal variables with two, four, five, and ten levels. The pink line is the theoretical CDF of the χ^2_9 distribution.

	Quantile	2 levels	4 levels	5 levels	10 levels
1 ordinal ($\mathsf{p} = 5$)	0.950	1.31	0.96	0.89	**0.67**
	0.955	1.33	1.00	0.90	**0.69**
	0.960	1.34	1.02	0.93	**0.70**
	0.965	1.36	1.05	0.97	**0.73**
	0.970	1.39	1.06	0.98	**0.75**
	0.975	1.44	1.08	1.03	**0.78**
	0.980	1.49	1.14	1.07	**0.83**
	0.985	1.56	1.17	1.15	**0.92**
	0.990	1.61	1.28	1.18	**0.98**
	0.995	1.60	1.21	1.15	**0.95**
2 ordinal ($\mathsf{p} = 6$)	0.950	1.83	1.23	1.11	**0.91**
	0.955	1.87	1.25	1.15	**0.94**
	0.960	1.90	1.27	1.16	**0.95**
	0.965	1.93	1.27	1.18	**0.94**
	0.970	1.96	1.31	1.17	**0.97**
	0.975	1.99	1.33	1.22	**0.99**
	0.980	2.01	1.36	1.22	**1.00**
	0.985	2.08	1.43	1.27	**1.03**
	0.990	2.13	1.48	1.34	**1.04**
	0.995	2.04	1.39	1.29	**0.98**
3 ordinal ($\mathsf{p} = 7$)	0.950	1.86	1.35	1.26	**1.16**
	0.955	1.88	1.38	1.28	**1.18**
	0.960	1.92	1.40	1.31	**1.22**
	0.965	1.94	1.43	1.35	**1.24**
	0.970	1.99	1.46	1.40	**1.28**
	0.975	2.02	1.50	1.46	**1.36**
	0.980	2.08	1.56	1.50	**1.39**
	0.985	2.23	1.61	1.59	**1.41**
	0.990	2.24	1.65	1.60	**1.36**
	0.995	2.05	1.49	1.51	**1.34**
4 ordinal ($\mathsf{p} = 8$)	0.950	2.28	1.65	1.64	**1.61**
	0.955	2.30	1.66	1.68	**1.65**
	0.960	2.37	1.68	**1.66**	1.69
	0.965	2.44	1.70	**1.67**	1.74
	0.970	2.46	1.71	**1.70**	1.78
	0.975	2.54	1.76	**1.73**	1.81
	0.980	2.55	1.82	**1.76**	1.83
	0.985	2.57	1.93	**1.76**	1.93
	0.990	2.58	1.91	**1.79**	2.00
	0.995	2.35	**1.73**	1.74	1.93

Table B.3: Difference between large quantiles of the χ^2_{p} distribution ($\mathsf{p} = 5, 6, 7, 8$) and the mean 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 ($\mathsf{p} = 9$)	0.950	1.98	**1.63**	1.66	1.80
	0.955	1.99	**1.66**	1.69	1.83
	0.960	2.01	**1.66**	1.73	1.85
	0.965	2.01	**1.70**	1.73	1.88
	0.970	2.01	**1.72**	1.78	1.91
	0.975	2.02	**1.76**	1.83	1.97
	0.980	2.08	**1.80**	1.91	2.02
	0.985	2.08	**1.86**	1.96	2.15
	0.990	2.10	**1.78**	1.91	2.18
	0.995	1.70	**1.32**	1.72	1.97

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