Supplement to "A k-mean-directions Algorithm for Fast Clustering of Data on the Sphere"

Ranjan Maitra and Ivan P. Ramler

S-1. ADDITIONAL EXPERIMENTAL EVALUATIONS

The k-mean-directions algorithm developed in the paper consists of the algorithm itself, initialization and determining the number of clusters (K). In Section 3 of the paper, we evaluated all three aspects together. Here we focus on the algorithm with initializer for K known as well as each portion of the initialization procedure. The experimental suite is the same as described in the paper: 25 repetitions for each of $(p, n, c) = \{2, 6, 10\} \times \{5000, 10000, 20000\} \times \{1, 2, 4\}$ with $K = \{3, 6, 12\}$ for $p = 2$ and 6 and $K = \{3, 8, 20\}$ for $p = 10$.

S-1.1 Evaluating k-mean-directions with the initializer and K known

Table S-1 summarizes the performance of k-mean-directions with its initializer when K is known. Performance is very good throughout and the algorithm classifies the observations nearly perfectly in many situations, in particular for moderate to high levels of separation (c = 2 and 4). Additionally, for small number of clusters, the derived groupings are excellent even for the poor separation scenario.

S-1.2 Evaluating the each portion of the initialization algorithm

The initialization method described in Section 2.2 of the paper consisted of two major parts, the modification to the deterministic initializer of Maitra (2007) and the best of multiple random starts. Here we separately evaluate each portion of the initialization for known K using classifications obtained from the initial centers to calculate the adjusted Rand measure ($\mathcal{R}_a$) relative to the true group identities for each dataset.

Table S-2 summarizes the performance of the deterministic initializer for the twenty-five replicates of each scenario of the experiment. Overall, performance is good for moderate to high sep-

Table S-1: Summary of k-mean-directions with number of clusters given as known. Each cell contains the median adjusted Rand value (top) and interquartile of the adjusted Rand values (bottom).

	$p = 2$				$p = 6$				$p = 10$			
		c				c				c		
	K	1.0	2.0	4.0	K	1.0	2.0	4.0	K	1.0	2.0	4.0
$n = 5,000$	3	0.954	0.997	0.999	3	0.947	0.992	0.998	3	0.928	0.990	0.998
		0.027	0.003	0.000		0.027	0.005	0.002		0.054	0.006	0.003
	6	0.949	0.997	1.000	6	0.801	0.993	0.998	8	0.790	0.909	0.998
		0.037	0.002	0.001		0.284	0.005	0.002		0.139	0.214	0.002
	12	0.861	0.926	0.984	12	0.865	0.988	0.999	20	0.726	0.979	0.998
		0.145	0.047	0.090		0.235	0.006	0.002		0.129	0.088	0.002
$n = 10,000$	3	0.953	0.998	0.999	3	0.952	0.994	0.999	3	0.930	0.994	0.998
		0.030	0.001	0.001		0.025	0.003	0.001		0.059	0.004	0.001
	6	0.960	0.998	1.000	6	0.835	0.995	0.999	8	0.797	0.988	0.998
		0.025	0.001	0.000		0.245	0.003	0.001		0.146	0.198	0.002
	12	0.861	0.955	0.964	12	0.918	0.993	0.998	20	0.739	0.990	0.998
		0.142	0.060	0.038		0.181	0.002	0.001		0.135	0.006	0.003
$n = 20,000$	3	0.950	0.998	1.000	3	0.961	0.995	0.999	3	0.921	0.994	0.999
		0.027	0.002	0.001		0.048	0.004	0.001		0.086	0.004	0.001
	6	0.959	0.998	1.000	6	0.809	0.996	0.999	8	0.803	0.994	0.999
		0.038	0.002	0.000		0.189	0.002	0.001		0.197	0.006	0.001
	12	0.958	0.966	0.981	12	0.935	0.996	0.999	20	0.740	0.992	0.998
		0.142	0.143	0.019		0.152	0.003	0.001		0.075	0.004	0.002

Table S-2: Summary of the deterministic portion of the initialization method with number of clusters given as known. Each cell contains the median adjusted Rand value (top) and interquartile of the adjusted Rand values (bottom).

	$p = 2$				$p = 6$				$p = 10$			
		c				c				c		
	K	1.0	2.0	4.0	K	1.0	2.0	4.0	K	1.0	2.0	4.0
$n = 5,000$	3	0.556 0.169	0.990 0.005	0.995 0.002	3	0.432 0.398	0.979 0.426	0.996 0.004	3	0.496 0.405	0.873 0.025	0.996 0.004
	6	0.928 0.199	0.992 0.007	0.995 0.099	6	0.650 0.241	0.988 0.011	0.998 0.002	8	0.542 0.102	0.870 0.137	0.997 0.003
	12	0.846 0.201	0.799 0.109	0.879 0.060	12	0.807 0.190	0.975 0.015	0.999 0.001	20	0.625 0.120	0.975 0.052	0.999 0.002
$n = 10,000$	3	0.565 0.083	0.999 0.002	0.999 0.001	3	0.474 0.343	0.990 0.001	0.997 0.003	3	0.530 0.301	0.872 0.437	0.992 0.003
	6	0.945 0.045	0.995 0.001	0.997 0.030	6	0.767 0.271	0.992 0.001	0.999 0.001	8	0.500 0.162	0.978 0.134	0.998 0.002
	12	0.846 0.169	0.767 0.167	0.845 0.220	12	0.901 0.121	0.994 0.005	0.996 0.003	20	0.587 0.144	0.982 0.017	0.994 0.003
$n = 20,000$	3	0.563 0.098	0.997 0.001	0.999 0.001	3	0.511 0.357	0.995 0.001	0.999 0.001	3	0.541 0.309	0.975 0.428	0.999 0.001
	6	0.939 0.192	0.998 0.001	1.000 0.001	6	0.778 0.248	0.992 0.001	0.999 0.001	8	0.589 0.117	0.994 0.128	0.999 0.001
	12	0.908 0.137	0.902 0.119	0.909 0.306	12	0.903 0.105	0.996 0.003	0.998 0.002	20	0.625 0.083	0.992 0.005	0.997 0.003

aration between clusters. However, for low separation and low number of clusters, the initializer doesn't perform as well. For low separation and moderate to high number of clusters, performance is mixed. In some cases (eg., $p = 6$, $K = 12$) performance is very good, in others not so, implying there is some variability in the quality of initial centers.

The performance of the best of 250 random starts is summarized in Table S-3. Once again, overall performance is quite good. The results stay fairly consistent across n and show a upward trend as cluster separation increases. The most notable trend is that as the number of clusters increases, performance decreases. This may imply that as K increases, more random starts should be evaluated to increase the chance that the initial centers will have good representatives from each of the clusters.

In summary, while both the deterministic initializer and best of multiple random starts exhibit areas of good performance, it appears that for low separation and small number of clusters, the random starts are better while the deterministic portion becomes better as the number of clusters

Table S-3: Summary of the initialization method using the best of 250 random starts with number of clusters given as known. Each cell contains the median adjusted Rand value (top) and interquartile of the adjusted Rand values (bottom).

	$p = 2$				$p = 6$				$p = 10$			
		c				c				c		
	K	1.0	2.0	4.0	K	1.0	2.0	4.0	K	1.0	2.0	4.0
$n = 5,000$	3	0.936	0.997	0.999	3	0.849	0.990	0.995	3	0.706	0.969	0.996
		0.054	0.003	0.000		0.055	0.014	0.001		0.116	0.012	0.001
	6	0.869	0.997	1.000	6	0.774	0.977	0.998	8	0.633	0.909	0.998
		0.122	0.004	0.001		0.184	0.029	0.002		0.163	0.156	0.154
	12	0.818	0.812	0.878	12	0.744	0.832	0.893	20	0.593	0.792	0.802
		0.084	0.097	0.085		0.091	0.089	0.054		0.108	0.058	0.034
$n = 10,000$	3	0.953	0.998	0.999	3	0.845	0.987	0.999	3	0.729	0.990	0.994
		0.036	0.002	0.001		0.088	0.009	0.000		0.116	0.003	0.001
	6	0.901	0.990	1.000	6	0.759	0.985	0.999	8	0.625	0.986	0.997
		0.066	0.008	0.000		0.124	0.012	0.000		0.160	0.161	0.157
	12	0.801	0.847	0.817	12	0.738	0.859	0.886	20	0.560	0.782	0.812
		0.048	0.094	0.093		0.107	0.093	0.072		0.102	0.037	0.057
$n = 20,000$	3	0.941	0.997	1.000	3	0.842	0.985	0.997	3	0.706	0.990	0.999
		0.046	0.002	0.001		0.098	0.011	0.003		0.083	0.002	0.001
	6	0.898	0.992	1.000	6	0.732	0.984	0.999	8	0.627	0.904	0.899
		0.094	0.007	0.000		0.123	0.021	0.001		0.191	0.158	0.164
	12	0.834	0.841	0.820	12	0.757	0.819	0.876	20	0.565	0.768	0.796
		0.108	0.089	0.096		0.076	0.094	0.063		0.097	0.082	0.054

increases. Thus combining the methods is suitable to rely on the strength of each. This is apparent as the performance for k-mean-directions (as seen in Table S-1) is quite a bit higher than that seen by each initializer separately. In light of these results, a few suggestions not implemented here, that may improve initialization, would be to run k-mean directions completely through based on the initial centers for both methods choosing the results that have the lowest value of the objective function (Equation 1 in Section 2 of the paper). Additionally, as mentioned earlier, for larger K the number of random starts evaluated should be increased to find good initial centers. This then may imply that to save computing time when K is large, for small K fewer random starts would be sufficient.

S-1.3 Comparison to k-means

To assess the overall usefulness of spherically constrained cluster algorithms we evaluate the performance of the k-means algorithm of Hartigan and Wong (1979). Table S-4 shows the performance of k-means on the experimental datasets. It is apparent that for a small number of clusters, the k-means algorithm does excellent. In fact, performance is often better for k-means than it was for k-mean-directions. However, a disturbing trend is the very large interquartile ranges. This implies that when k-means does not correctly identify the clusters, it seems to be very inaccurate. Further, performance deteriorates rapidly when the number of clusters increases. Surprisingly, this is more apparent for higher separation as the Adjusted Rand values are much lower than those for low separation. One possible explanation for this is that by not constraining the cluster centers appropriately, the k-means algorithm does not update itself well and may converge to locally optimal areas. Another explanation may be that if the initial centers are not located in each of the actual clusters, k-means converges to locally optimal areas without properly updating.

S-1.4 Comparative running times between k-mean-directions and $spkmeans$

To assess the efficiency of the k-mean-directions algorithm compared to $spkmeans$, we ran both algorithms on each dataset initialized using the proposed methodology (Section 2.2) in the main paper. The experiments were completed on a Dell Optiplex 755 with 1.9 GiB memory and Intel(R) Core2 Duo 2.66 Ghz Processor with the Linux 2.6.27.24-170.2.68 kernel and Fedora 10 as the operating system. For each dataset, we considered the ratio of the time running $spkmeans$ to k-mean-directions as our measure of comparison. Thus, a ratio greater than one indicates k-mean-directions was the faster algorithm. Figure S-1 illustrates the connection between the amount of

Table S-4: Summary of the performance of the standard k-means algorithm with number of clusters given as known. Each cell contains the median adjusted Rand value (top) and interquartile of the adjusted Rand values (bottom).

	$p = 2$				$p = 6$				$p = 10$			
		c				c				c		
	K	1.0	2.0	4.0	K	1.0	2.0	4.0	K	1.0	2.0	4.0
$n = 5,000$	3	0.986	1.000	1.000	3	0.995	1.000	1.000	3	0.994	1.000	1.000
		0.014	0.555	0.000		0.005	0.001	0.000		0.004	0.000	0.554
	6	0.987	0.774	0.779	6	0.996	0.769	0.775	8	0.987	0.836	0.721
		0.264	0.231	0.229		0.231	0.233	0.005		0.170	0.015	0.142
	12	0.773	0.795	0.729	12	0.850	0.884	0.797	20	0.868	0.821	0.758
		0.128	0.179	0.115		0.109	0.083	0.165		0.064	0.109	0.095
$n = 10,000$	3	0.991	1.000	1.000	3	0.997	1.000	1.000	3	0.995	1.000	1.000
		0.010	0.000	0.000		0.002	0.001	0.555		0.002	0.000	0.554
	6	0.982	0.772	0.773	6	0.995	0.780	0.773	8	0.989	0.828	0.830
		0.243	0.011	0.010		0.231	0.230	0.009		0.168	0.119	0.140
	12	0.746	0.731	0.731	12	0.854	0.799	0.798	20	0.897	0.826	0.771
		0.165	0.141	0.147		0.117	0.160	0.184		0.048	0.062	0.065
$n = 20,000$	3	0.988	1.000	1.000	3	0.998	1.000	1.000	3	0.997	1.000	1.000
		0.009	0.000	0.000		0.002	0.000	0.555		0.002	0.000	0.554
	6	0.743	0.770	0.774	6	0.997	0.773	0.775	8	0.995	0.837	0.835
		0.287	0.234	0.007		0.009	0.232	0.008		0.176	0.169	0.286
	12	0.786	0.792	0.729	12	0.870	0.796	0.789	20	0.919	0.821	0.779
		0.133	0.109	0.104		0.079	0.168	0.145		0.056	0.081	0.085

running time and the type of dataset clustered. It is apparent that regardless of the number of clusters, k-mean-directions becomes the more efficient algorithm as the number of dimensions increases. One reason for this could be that since k-mean-directions only updates as necessary, for higher dimensions (and to a lesser extent higher number of clusters) there are less unnecessary calculations being performed. However, Figure S-1 also indicates that for low dimensions, $spkmeans$ is the more efficient algorithm. This may imply that for smaller datasets the time saved in updating only as necessary does not make up for the initial time spent determining the closest and second closest cluster centers (step 1 from Section 2.1). However, as the datasets are smaller to begin with, the total amount of time should still be quite small. Finally, it should be noted that regardless of the size of the dataset, both algorithms still perform fairly fast as both were able to complete most datasets in under one-tenth of a second. Thus, while k-mean-directions may be the most time efficient, $spkmeans$ may still be a viable options for at least moderately sized datasets and either algorithm may be a better option than other potentially slower algorithms such as those based on the EM algorithm.

S-1.5 Illustrative Summary of Large-sample Simulation Experiments

Section 3.2 of the main paper describes the performance of the k-mean-directions with K considered unknown and displays the results in Table 2. Here, Figure S-2 include boxplots of the distributions of the Adjusted Rand values for each experiment, grouped in the same fashion as Table 2 from the main paper. It is again clear that as performance is often very good, especially for moderate separation ($c = 2$). However, as indicated by the outliers in Figure S-2, there are the rare occasions when performance is quite poor. This occurs when the number of clusters is not accurately estimated (in particular when K is grossly underestimated) and is quite rare as, for $c = 2$, only four of the 675 datasets (about 0.6%) resulted in extremely low Adjusted Rand values.

S-2. 2008 JOINT STATISTICAL MEETINGS ABSTRACTS

The top seven words along with brief descriptions and cardinality of each of the 55 clusters found in the 2,107 abstracts of talks presented at the 2008 Joint Statistical Meetings are provided in Tables S-5 and S-6. The dataset is more thoroughly examined in Section 4.2 of the paper.

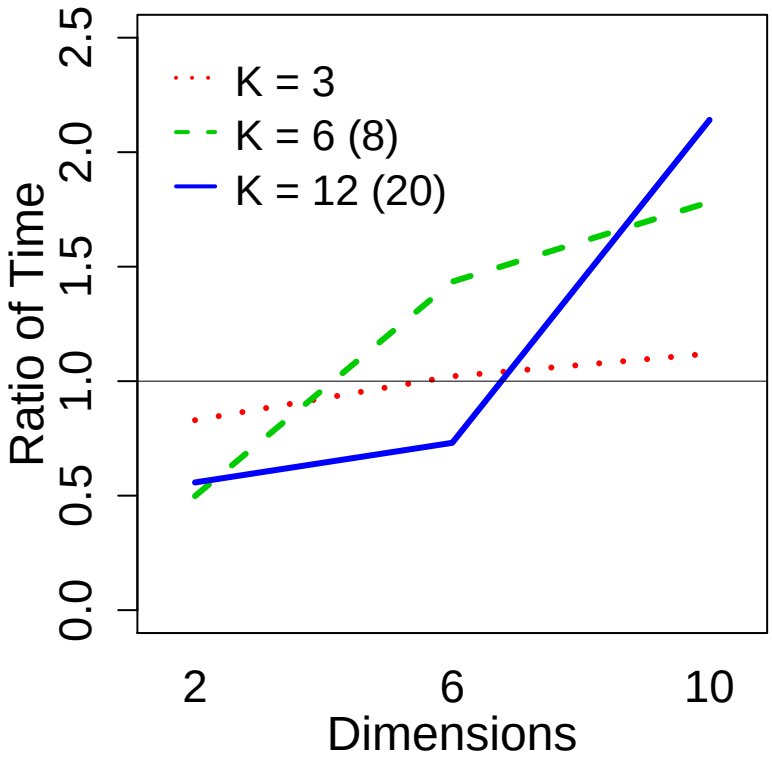

Figure S-1: Ratio of running time for $spkmeans$ to k-mean-directions across dimensions (p) and number of clusters (K). The lines are the ratios averaged over sample sizes, cluster (c) separation and the 25 repetitions. The numbers in parentheses indicate the number of clusters for p=10 dimensions.

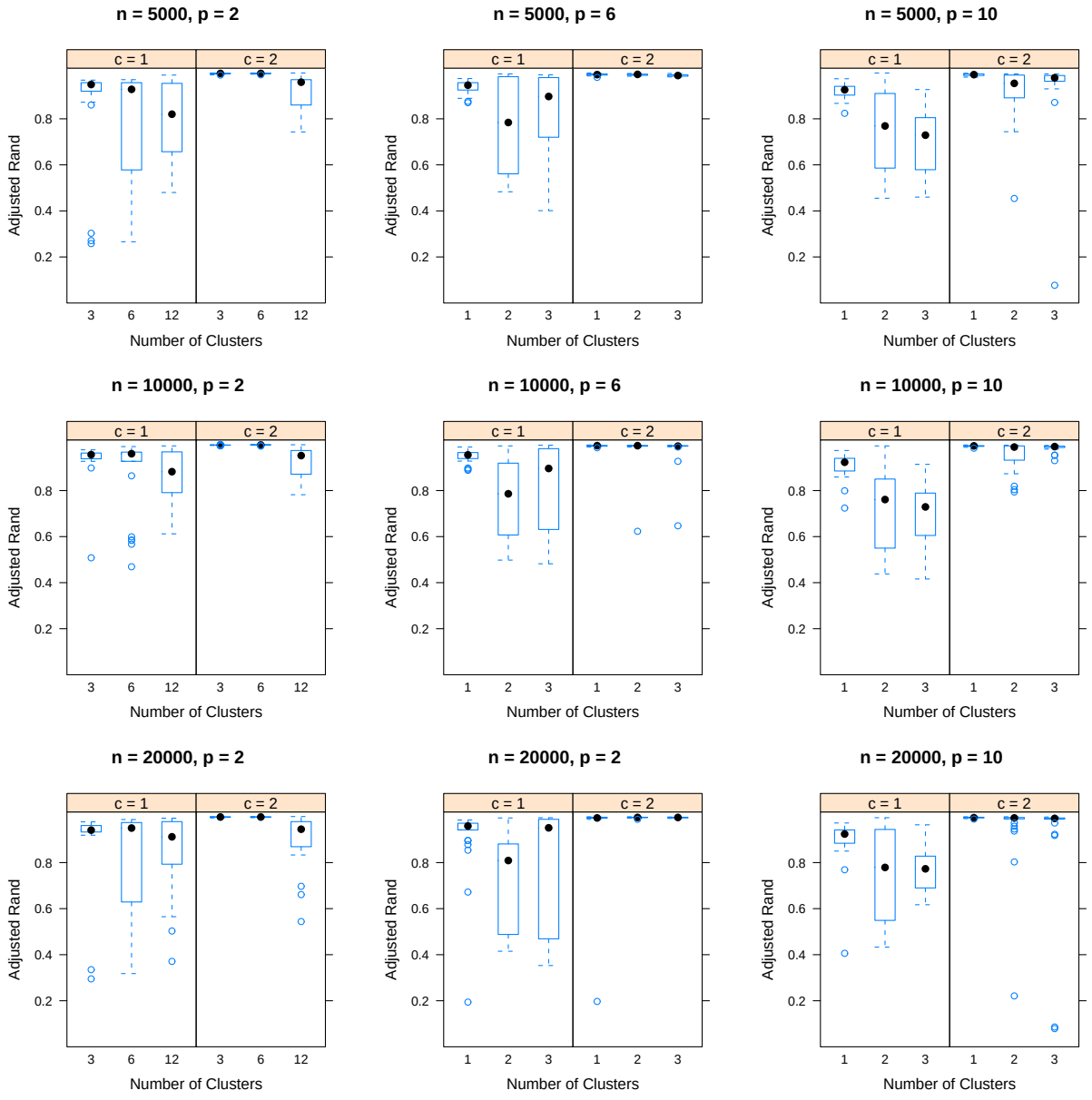

Figure S-2: Distributions of $\mathcal{R}_a$ of the twenty-five repitions for each design scenario when using k-mean-directions with estimated number of clusters.

Table S-5: Top seven words for the first twenty-eight of the fifty-five clusters in the 2008 Joint Statistical Meetings abstracts. The first column provides a rough interpretation of the cluster subject with the cardinality of the cluster given in parentheses.

Analysis of Clinical Trials (56)	treatment	effect	trial	intervention	clinical	randomized	placebo
Variable Selection (51)	selection	variable	variables	regression	procedure	predictors	lasso
Hypothesis Testing (50)	tests	test	null	hypothesis	hypotheses	testing	distribution
Introductory Statistical Education (49)	statistics	students	courses	statisticians	teaching	learning	online
Missing Data (49)	missing	imputation	longitudinal	information	covariates	values	multiple
Health Surveys - NHANES (48)	survey	surveys	nonresponse	bias	nhanes	health	response
SNPs Data (46)	snps	genetic	association	snp	genome	markers	disease
Bootstrap (46)	estimators	estimator	bootstrap	robust	regression	estimation	parameters
Experimental Design (45)	designs	design	criterion	criteria	choice	optimal	dose
Spatial Statistics (45)	spatial	clustering	temporal	dependence	spatio	climate	kernel
Cancer Studies (43)	risk	risks	recurrent	exposure	metrics	cancer	service
Designing Clinical Trials (43)	trials	sequential	power	clinical	size	interim	adaptive
Gene Expressions (43)	gene	expression	association	genes	genetic	interaction	interactions
Linear Models (42)	nonparametric	function	parametric	regression	test	anova	likelihood
Markov Chain Monte Carlo (42)	sampling	markov	hiv	chain	population	algorithm	scheme
Cancer Studies (42)	patients	cure	cancer	survival	groups	misclassification	treatment
Generalized Estimating Equations (41)	covariance	matrix	interval	gee	variance	symbolic	estimator
Online Education (41)	students	student	learning	literacy	program	statistics	education
Reliability & Stat. Quality Control (40)	distributions	distribution	discrete	weibull	moments	parameters	random
Pharmaceutical Clinical Trials (40)	risk	clinical	cox	trial	safety	drug	event
National Household Survey (40)	survey	frame	weights	sampling	surveys	households	nonresponse
Climate Statistics (39)	error	climate	type	research	measurement	issues	classical
Statistical Consulting (39)	consulting	training	center	service	collaborative	collaboration	graduate
Regression Analysis (39)	regression	estimate	polynomial	quantile	local	kernel	asymptotic
Health Surveys - NHIS (39)	health	panel	nhis	survey	meps	income	national
Regression Methodology (38)	coefficient	varying	response	protein	predictor	predictors	estimation
Bayesian Anaysis (38)	effects	random	prior	inference	bayesian	fixed	distribution
Nonparametric Statistics (38)	nonlinear	parametric	semiparametric	estimator	response	linear	responses

Table S-6: Top seven words for the last twenty-seven of the fifty-five clusters in the 2008 Joint Statistical Meetings abstracts. The first column provides a rough interpretation of the cluster subject with the cardinality of the cluster given in parentheses.

Time Series (38)	series	stationary	noise	spectral	domain	innovations	matrix
High Dimensions (38)	dimensional	problems	chain	independence	high	distribution	markov
Long-memory Processes (38)	memory	order	ensemble	sizes	assimilation	asymptotic	confidence
Estimation in Survival Analysis (38)	limits	prediction	intervals	confidence	tolerance	distribution	survival
Business Surveys (38)	respondents	users	survey	mode	statistics	internet	business
Undergraduate Research (37)	students	projects	statistics	project	research	statistician	topics
Engineering Statistics (37)	curve	calibration	roc	curves	predictions	censored	randomization
Census Studies (36)	census	acs	form	bureau	county	unit	estimates
Microarrays (35)	genes	microarray	expressed	fdr	differentially	gene	expression
Demographic Surveys (35)	estimates	direct	insurance	census	survey	level	population
Dimension Reduction (35)	reduction	dimension	weight	loss	plots	subspace	space
Imaging Biomarkers (34)	biomarker	imaging	biomarkers	search	clinical	drug	specificity
Mixture Distributions (34)	mixture	mixtures	gamma	chemical	gaussian	distributions	poisson
Spatial Processes & Applications (34)	patterns	point	geographic	network	social	process	change
Multiple Testing (33)	adjustment	testing	multiplicity	superiority	text	procedure	test
Discrete Data (33)	count	poisson	logit	counts	fit	binary	inflated
Biopharmaceutical Statistics (33)	charts	recovery	control	contamination	drug	monitoring	blood
Secondary Education (32)	school	dropout	students	student	schools	graduation	university
Polling & Voting (31)	system	election	polls	voters	candidate	elections	audit
Statistics & Demography (30)	ratio	national	matter	health	segments	node	aspect
Baysian Spatial Statistics (30)	spatial	mcmc	traffic	glm	parameter	dispersion	regularity
Sports Applications (29)	players	series	games	game	forecasting	round	placement
Bayesian Additive Regression Trees (29)	subgroup	home	bart	specific	haplotype	analyses	multiple
Applied Stochastic Processes (29)	experiments	actual	outcome	measurements	vaccination	scientific	categories
Financial Statistics (27)	frequencies	visual	extreme	economic	frequency	arima	bank
Economics Applications (26)	monthly	gdp	sales	prices	algorithm	quarterly	volume
Mortality Studies (26)	mortality	deaths	influenza	epidemic	death	infant	disease

REFERENCES

Hartigan, J. A., and Wong, M. A. (1979), "A K-means clustering algorithm," *Applied Statistics*, 28, 100–108.

Maitra, R. (2007), "Initializing Partition-Optimization Algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, preprint, 14 Aug 2007. doi:10.1109/TCBB.2007.70244.