

D3-3D-LSA for QuantNet 2.0 and GitHub

Lukas Borke

Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E. – Center for Applied Statistics
and Economics

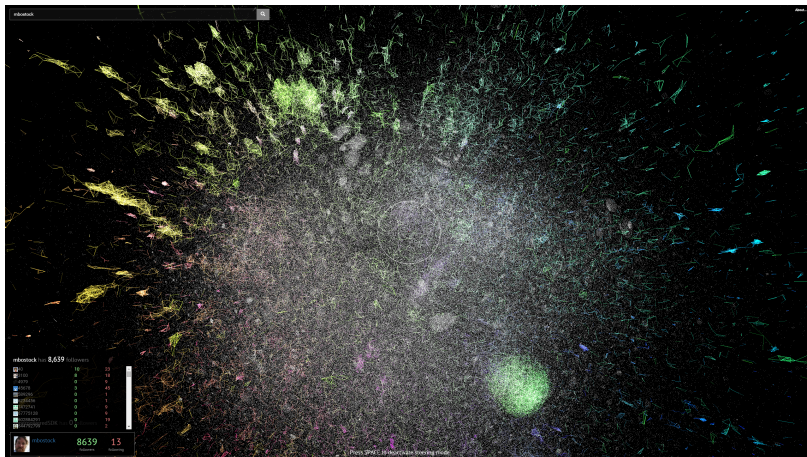
Humboldt-Universität zu Berlin

<http://lvb.wiwi.hu-berlin.de>

<http://www.case.hu-berlin.de>



Seek and ye shall find



Modern Scientific Practice

Modern scientific practice:

- Transparency
- Reproducibility
- Collaborative Reproducible Research

Also: Want to publicize new technologies!

Problem: Need and want to publish our technologies and data!



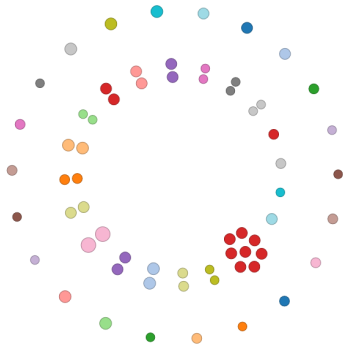
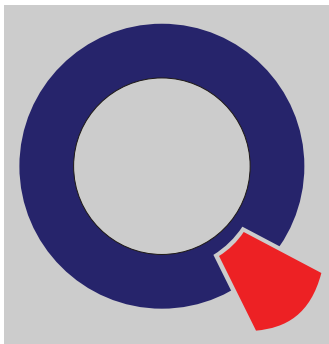
Outline

1. Motivation ✓
2. QuantNet 2.0 and GitHub
3. Challenges
4. Vector Space Model (VSM)
5. Empirical results
6. Interactive Structure
7. Conclusion



The Solution

QuantNet 2.0



The Solution

QuantNet 2.0 - The Next Generation

- ▣ $\approx$ 2000 Quantlets
- ▣ Technology to easily share data and programs
- ▣ Searchable technology
- ▣ Enabled collaboration via seamless GitHub integration
- ▣ Connections between technologies

Boosting transparent and reproducible science



What is GitHub?



- A distributed version control system (Git)
- A collaboration platform (Hub)



Advantages of QuantNet 2.0

- Fully integrated with GitHub
- Proprietary GitHub-R-API developed from core package (Arizona State University)
- Text Mining Pipeline via R packages providing D3 and 3D Visualizations and Document clustering
- Tuned and integrated Search engine within the main D3 Visu based on validated meta information in Quantlets
- Ease of discovery and use of your technology
- Audit of your technology



Objectives

- D3: D3.js – Data-Driven Documents
 - ▶ Knowledge discovery via information visualization
 - ▶ [visit on GitHub](#)

- 3D: Three.js – Next logical step
 - ▶ cross-browser JavaScript library/API
 - ▶ animate 3D computer graphics in a web browser
 - ▶ [visit on GitHub](#)

- LSA: Latent Semantic Analysis
 - ▶ Semantic Embedding
 - ▶ [visit on GitHub](#)



Statistical Challenges

□ Text Mining

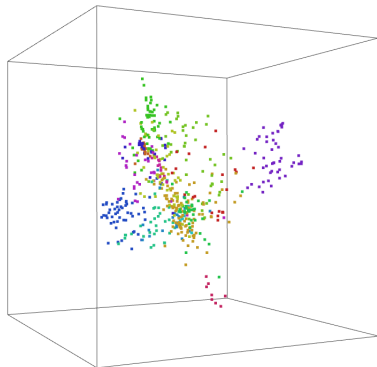
- ▶ Model calibration
- ▶ Dimension reduction
- ▶ Semantic based Information Retrieval
- ▶ Cluster validation for Document Clustering

□ Visualization

- ▶ Projection techniques
- ▶ 2D, 3D
- ▶ Geometry



Vector Space Model (VSM)



□ Model calibration

- ▶ Text to Vector: Weighting scheme, Similarity, Distance
- ▶ Generalized VSM (GVSM)
Latent Semantic Analysis



Text to Vector

- $Q = \{d_1, \dots, d_n\}$ **set of documents** (Quantlets/Gestalten).
- $T = \{t_1, \dots, t_m\}$ **dictionary** (set of all terms).
- $tf(d, t)$ **absolute frequency** of term $t \in T$ in $d \in Q$.
- $idf(t) \stackrel{\text{def}}{=} \log(|Q|/n_t)$ **inverse document frequency**, with $n_t = |\{d \in Q | t \in d\}|$.
- $w(d) = \{w(d, t_1), \dots, w(d, t_m)\}^\top \in \mathbb{R}^m, d \in Q$, **document as vector**.
- $w(d, t_i)$ calculated by a **weighting scheme**.
- $D = [w(d_1), \dots, w(d_n)] \in \mathbb{R}^{m \times n}$, **term by document matrix (TDM)**.



Weighting scheme, Similarity, Distance

- Salton et al. (1994): the **tf-idf** – weighting scheme

$$w(d, t) = \frac{tf(d, t)idf(t)}{\sqrt{\sum_{j=1}^m tf(d, t_j)^2 idf(t_j)^2}}, m = |T|$$

- (normalized tf-idf) Similarity S of two documents d_1 and d_2

$$S(d_1, d_2) = \sum_{k=1}^m w(d_1, t_k) \cdot w(d_2, t_k) = w(d_1)^T w(d_2)$$

- Euclidian distance measure:

$$dist_d(d_1, d_2) \stackrel{\text{def}}{=} \sqrt{\sum_{k=1}^m \{w(d_1, t_k) - w(d_2, t_k)\}^2}$$



Generalized VSM (GVSM)

Generalize similarity S with a linear mapping P :

$$S(d_1, d_2) = (Pd_1)^\top (Pd_2) = d_1^\top P^\top P d_2$$

Every P defines another VSM:

$$M_S^{(P)} = D^\top (P^\top P) D$$

M_S : similarity matrix, D : term by document matrix



GVSM

□ Basic VSM (BVSM): ▶ BVSM

- ▶ $P = I_m$ and $w(d) = \{tf(d, t_1), \dots, tf(d, t_m)\}^\top$
- ▶ classical tf-similarity: $M_S^{tf} = D^\top D$

□ Term-Term correlations: ▶ GVSM(TT)

- ▶ $P = D^\top$, $M_S^{TT} = D^\top (DD^\top) D$
- ▶ $DD^\top$: term by term correlation matrix

□ Latent Semantic Analysis ▶ LSA

- ▶ $D = U\Sigma V^\top$: singular value decomposition (SVD)
- ▶ $P = U_k^\top = I_k U^\top$: projection onto the first k dimensions
- ▶ $M_S^{LSA} = D^\top (U I_k U^\top) D$
- ▶ The k dimensions as the main semantic components and $U_k U_k^\top = U I_k U^\top$ their correlation.



Power of LSA

- Highest-performing variants of LSA-based search algorithms perform as well as PageRank-based Google search engine (Miller et al., 2009)
- In half of the studies with 30 sets LSA performance equal to or better than that of humans (Bradford, 2009)
- Positive correlation of LSA comparable with the more sophisticated WordNet based methods and also human ratings ($r = 0.88$), (Mohamed et al., 2014)



M^3 : 3 Models, 3 Methods, 3 Measures

- Dataset: the whole Quantnet
- Documents: 1170 Gestalten (from 1826 individual Quantlets)
- 3 Models: BVSM, GVSM(TT) and GVSM(LSA)
 - ▶ 3 configurations in LSA with dimension parameter k equal to 300, 171 (50% of the weight of all singular values) and 50
- 3 Clustering methods: ▶ *k*-Means, ▶ *k*-Medoids, ▶ HCA
- 3 Cluster validation measures:
 - ▶ Connectivity ▶ Connectivity
 - ▶ Silhouette width ▶ Silhouette
 - ▶ Dunn Index ▶ Dunn



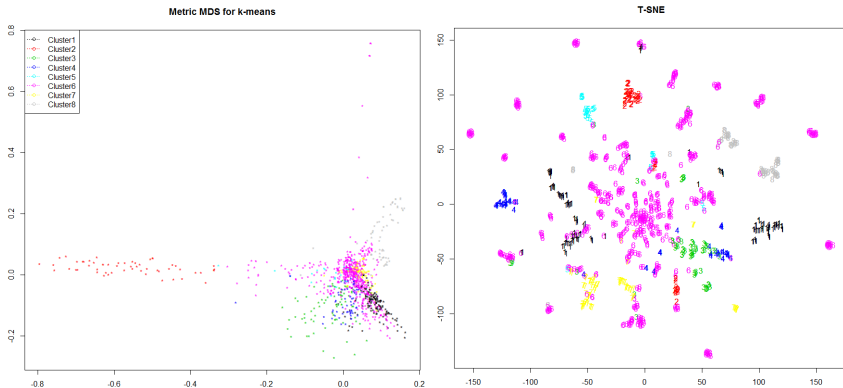
M^3 evaluation results

Measure	Model	Method
Connectivity	LSA50	hca
Silhouette	LSA50	hca
Dunn	BVSM/LSA	hca

- Hierarchical Clustering(hca) better or comparable to other methods in all measure aspects and in all models
- LSA50 superior wrt. Connectivity and Silhouette
- BVSM/LSA slightly better than LSA50 wrt. Dunn, but still comparable (small range of values in all models)
- Conclusion: hca under LSA/LSA50 is the optimal method

► More results





2D-Geometry via MDS (left) and t-SNE (right) in LSA:50

8 *k*-Means-Clusters:

- 1: distribut copula normal gumbel pdf;
- 2: call option blackschol put price;
- 3: return timeseri dax stock financi;
- 4: portfolio var pareto return risk;
- 5: interestr filter likelihood cir term;
- 6: visual dsfm requir kernel test;
- 7: regress nonparametr linear logit lasso;
- 8: cluster analysi pca principalcompon dendrogram

[▶ More about t-SNE](#) [▶ Geometry via 3D/Three.js](#)



Dendrogram: subset from SFE, SFS, IBT

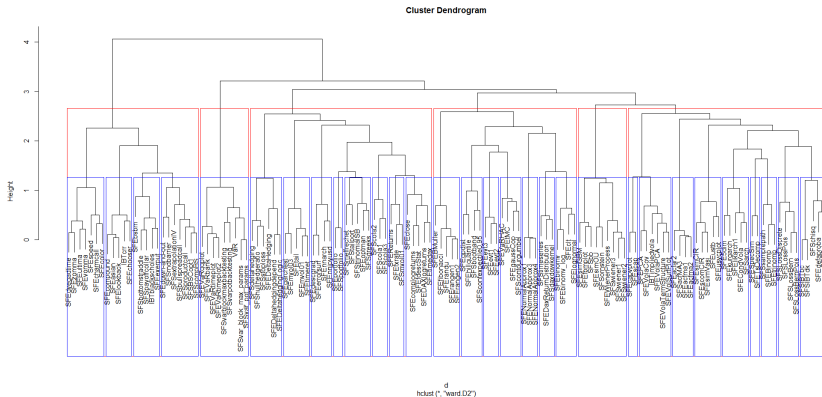


Figure 1: Created by hierarchical clustering (ward-method) in LSA model, cut in 6 clusters and 30 subclusters, 137 Gestalten

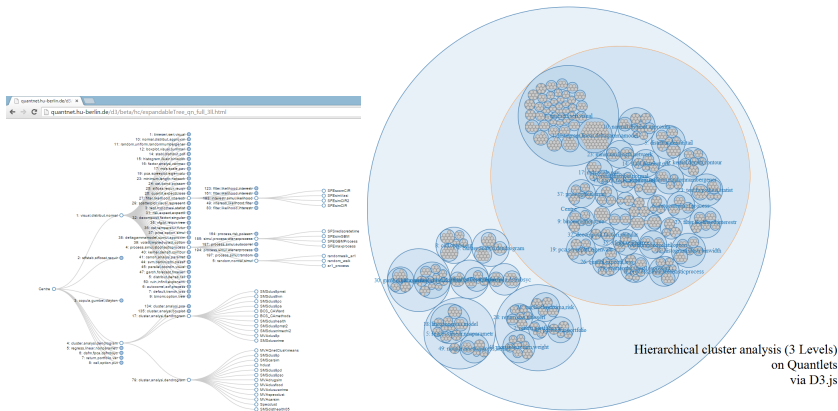
► D3 Scheme 1

► D3 Scheme 2

► D3 Scheme 3



Figure 2: Come in and Quant out under quantnet.wiwi.hu-berlin.de/d3/beta/



Search Options: Select the language: Animation on/off Color palette: Palette 1

Search: pca extended ☒ R ☒ Matlab ☒ SAS ☒

Found jets: 31

SPAdemoH1Japan
FPCAAscan
FPCAago
FPCAindividual
FPCAmultibic
FPCAsimulate_input
FPCAsimulation
FPCAvariance
LETFactorFuncs
LETFVSurfPlot
LETFVTrueMonsc
LETFSelectLoads
MVAcpair
MVAncpauco
MVAncpauco2
MVAncpauus
MVAncpauus2
MVAclusbank
MVAclusbsh
MVAclusfod
MVAncpabank
MVAncpabank2
MVAncpabank3
MVAncpabank4
MVAncpabank5
MVAncpabank6
MVAncpabank7
MVAncpabank8
MVAncpabank9
MVAncpabank10
MVAncpabank11
MVAncpabank12
MVAncpabank13
MVAncpabank14
MVAncpabank15
MVAncpabank16
MVAncpabank17
MVAncpabank18
MVAncpabank19
MVAncpabank20
MVAncpabank21
MVAncpabank22
MVAncpabank23
MVAncpabank24
MVAncpabank25
MVAncpabank26
MVAncpabank27
MVAncpabank28
MVAncpabank29
MVAncpabank30
MVAncpabank31
MVAncpabank32
MVAncpabank33
MVAncpabank34
MVAncpabank35
MVAncpabank36
MVAncpabank37
MVAncpabank38
MVAncpabank39
MVAncpabank40
MVAncpabank41
MVAncpabank42
MVAncpabank43
MVAncpabank44
MVAncpabank45
MVAncpabank46
MVAncpabank47
MVAncpabank48
MVAncpabank49
MVAncpabank50
MVAncpabank51
MVAncpabank52
MVAncpabank53
MVAncpabank54
MVAncpabank55
MVAncpabank56
MVAncpabank57
MVAncpabank58
MVAncpabank59
MVAncpabank60
MVAncpabank61
MVAncpabank62
MVAncpabank63
MVAncpabank64
MVAncpabank65
MVAncpabank66
MVAncpabank67
MVAncpabank68
MVAncpabank69
MVAncpabank70
MVAncpabank71
MVAncpabank72
MVAncpabank73
MVAncpabank74
MVAncpabank75
MVAncpabank76
MVAncpabank77
MVAncpabank78
MVAncpabank79
MVAncpabank80
MVAncpabank81
MVAncpabank82
MVAncpabank83
MVAncpabank84
MVAncpabank85
MVAncpabank86
MVAncpabank87
MVAncpabank88
MVAncpabank89
MVAncpabank90
MVAncpabank91
MVAncpabank92
MVAncpabank93
MVAncpabank94
MVAncpabank95
MVAncpabank96
MVAncpabank97
MVAncpabank98
MVAncpabank99
MVAncpabank100
MVAncpabank101
MVAncpabank102
MVAncpabank103
MVAncpabank104
MVAncpabank105
MVAncpabank106
MVAncpabank107
MVAncpabank108
MVAncpabank109
MVAncpabank110
MVAncpabank111
MVAncpabank112
MVAncpabank113
MVAncpabank114
MVAncpabank115
MVAncpabank116
MVAncpabank117
MVAncpabank118
MVAncpabank119
MVAncpabank120
MVAncpabank121
MVAncpabank122
MVAncpabank123
MVAncpabank124
MVAncpabank125
MVAncpabank126
MVAncpabank127
MVAncpabank128
MVAncpabank129
MVAncpabank130
MVAncpabank131
MVAncpabank132
MVAncpabank133
MVAncpabank134
MVAncpabank135
MVAncpabank136
MVAncpabank137
MVAncpabank138
MVAncpabank139
MVAncpabank140
MVAncpabank141
MVAncpabank142
MVAncpabank143
MVAncpabank144
MVAncpabank145
MVAncpabank146
MVAncpabank147
MVAncpabank148
MVAncpabank149
MVAncpabank150
MVAncpabank151
MVAncpabank152
MVAncpabank153
MVAncpabank154
MVAncpabank155
MVAncpabank156
MVAncpabank157
MVAncpabank158
MVAncpabank159
MVAncpabank160
MVAncpabank161
MVAncpabank162
MVAncpabank163
MVAncpabank164
MVAncpabank165
MVAncpabank166
MVAncpabank167
MVAncpabank168
MVAncpabank169
MVAncpabank170
MVAncpabank171
MVAncpabank172
MVAncpabank173
MVAncpabank174
MVAncpabank175
MVAncpabank176
MVAncpabank177
MVAncpabank178
MVAncpabank179
MVAncpabank180
MVAncpabank181
MVAncpabank182
MVAncpabank183
MVAncpabank184
MVAncpabank185
MVAncpabank186
MVAncpabank187
MVAncpabank188
MVAncpabank189
MVAncpabank190
MVAncpabank191
MVAncpabank192
MVAncpabank193
MVAncpabank194
MVAncpabank195
MVAncpabank196
MVAncpabank197
MVAncpabank198
MVAncpabank199
MVAncpabank200
MVAncpabank201
MVAncpabank202
MVAncpabank203
MVAncpabank204
MVAncpabank205
MVAncpabank206
MVAncpabank207
MVAncpabank208
MVAncpabank209
MVAncpabank210
MVAncpabank211
MVAncpabank212
MVAncpabank213
MVAncpabank214
MVAncpabank215
MVAncpabank216
MVAncpabank217
MVAncpabank218
MVAncpabank219
MVAncpabank220
MVAncpabank221
MVAncpabank222
MVAncpabank223
MVAncpabank224
MVAncpabank225
MVAncpabank226
MVAncpabank227
MVAncpabank228
MVAncpabank229
MVAncpabank230
MVAncpabank231
MVAncpabank232
MVAncpabank233
MVAncpabank234
MVAncpabank235
MVAncpabank236
MVAncpabank237
MVAncpabank238
MVAncpabank239
MVAncpabank240
MVAncpabank241
MVAncpabank242
MVAncpabank243
MVAncpabank244
MVAncpabank245
MVAncpabank246
MVAncpabank247
MVAncpabank248
MVAncpabank249
MVAncpabank250
MVAncpabank251
MVAncpabank252
MVAncpabank253
MVAncpabank254
MVAncpabank255
MVAncpabank256
MVAncpabank257
MVAncpabank258
MVAncpabank259
MVAncpabank260
MVAncpabank261
MVAncpabank262
MVAncpabank263
MVAncpabank264
MVAncpabank265
MVAncpabank266
MVAncpabank267
MVAncpabank268
MVAncpabank269
MVAncpabank270
MVAncpabank271
MVAncpabank272
MVAncpabank273
MVAncpabank274
MVAncpabank275
MVAncpabank276
MVAncpabank277
MVAncpabank278
MVAncpabank279
MVAncpabank280
MVAncpabank281
MVAncpabank282
MVAncpabank283
MVAncpabank284
MVAncpabank285
MVAncpabank286
MVAncpabank287
MVAncpabank288
MVAncpabank289
MVAncpabank290
MVAncpabank291
MVAncpabank292
MVAncpabank293
MVAncpabank294
MVAncpabank295
MVAncpabank296
MVAncpabank297
MVAncpabank298
MVAncpabank299
MVAncpabank300
MVAncpabank301
MVAncpabank302
MVAncpabank303
MVAncpabank304
MVAncpabank305
MVAncpabank306
MVAncpabank307
MVAncpabank308
MVAncpabank309
MVAncpabank310
MVAncpabank311
MVAncpabank312
MVAncpabank313
MVAncpabank314
MVAncpabank315
MVAncpabank316
MVAncpabank317
MVAncpabank318
MVAncpabank319
MVAncpabank320
MVAncpabank321
MVAncpabank322
MVAncpabank323
MVAncpabank324
MVAncpabank325
MVAncpabank326
MVAncpabank327
MVAncpabank328
MVAncpabank329
MVAncpabank330
MVAncpabank331
MVAncpabank332
MVAncpabank333
MVAncpabank334
MVAncpabank335
MVAncpabank336
MVAncpabank337
MVAncpabank338
MVAncpabank339
MVAncpabank340
MVAncpabank341
MVAncpabank342
MVAncpabank343
MVAncpabank344
MVAncpabank345
MVAncpabank346
MVAncpabank347
MVAncpabank348
MVAncpabank349
MVAncpabank350
MVAncpabank351
MVAncpabank352
MVAncpabank353
MVAncpabank354
MVAncpabank355
MVAncpabank356
MVAncpabank357
MVAncpabank358
MVAncpabank359
MVAncpabank360

The resulting 31 objects are concentrated on 3 clusters with the topics:
"pca, eigenvalue, standard", "regress, model, estimation" and
"volatility, option, implied" [► D3 Visu](#) [► 3D Visu](#)

D3-3D-LSA for QuantNet 2.0 and GitHub



Collaboration Timeline via GitHub-API

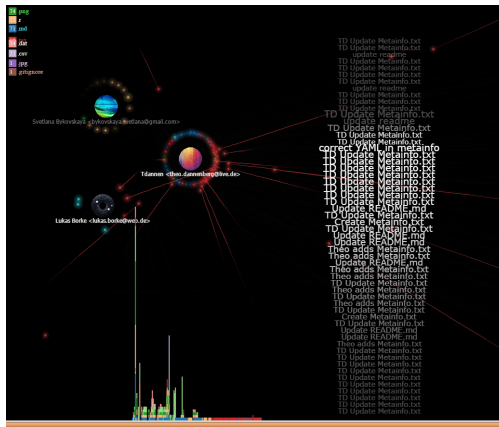
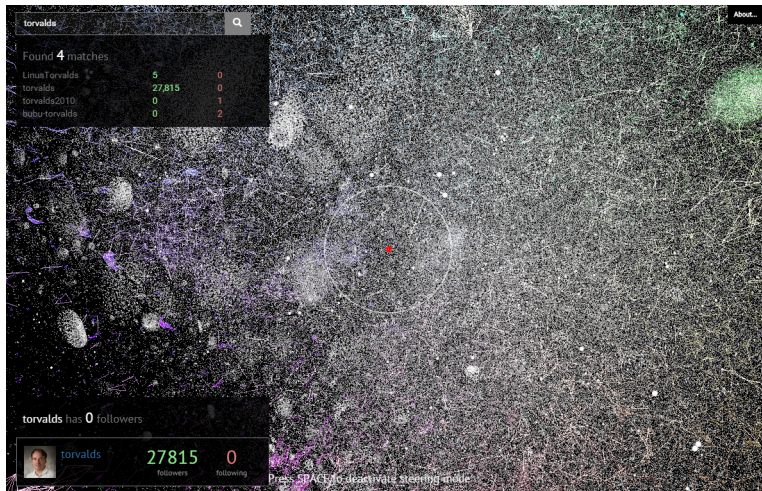


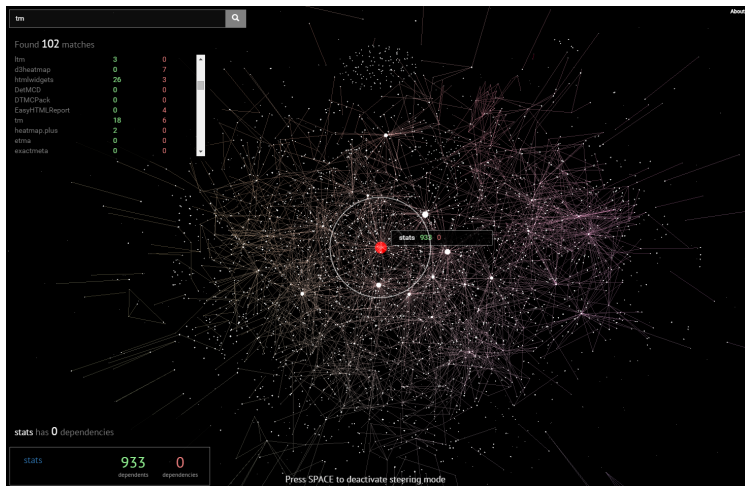
Figure 4: Snapshot of the development of the MVA repository
More examples of collaboration projects



3D GitHub Network Graph



3D CRAN Network Graph - R Language



Conclusion

- Different model configurations allow adapted Similarity based Document Clustering and Knowledge Discovery
- LSA/LSA50 and HCA optimal under M^3 evaluation
- Incorporating **term-term Correlations** and **Semantics**:
 - ▶ Sparsity reduction
 - ▶ higher clustering performance and better semantic topics
 - ▶ more recall/precision (IR)



Future Perspectives

- More clustering methods and validation measures for performance validation: from M^3 to M^k
- Optimization of cluster labels for easier human readability
- Implementation of „upgrades“ into QuantNet 2.0 via D3-3D
- Extension of D3-3D-LSA to further parts of GitHub
 - ▶ from BigData to SmartData



D3-3D-LSA for QuantNet 2.0 and GitHub

Lukas Borke

Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E. – Center for Applied Statistics

and Economics

Humboldt-Universität zu Berlin



<http://lvb.wiwi.hu-berlin.de>

<http://www.case.hu-berlin.de>



References



Borgelt, C. and Nürnberger, A.

Experiments in Term Weighting and Keyword Extraction in Document Clustering

LWA, pp. 123-130, Humboldt-Universität Berlin, 2004



Bostock, M., Heer, J., Ogievetsky, V. et al.

D3: Data-Driven Documents

available on d3js.org, 2014



Bradford, R.

Comparability of LSI and Human Judgment in Text Analysis Tasks

Proceeding MMACTEE'09, pp. 359-366, 2009

available on dl.acm.org



References



Chen, C., Härdle, W. and Unwin, A.

Handbook of Data Visualization

Springer, 2008



Elsayed, T., Lin, J. and Oard, D. W.

Pairwise Document Similarity in Large Collections with MapReduce

Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL), pp. 265-268, 2008



Feldman, R. and Dagan, I.

Mining Text Using Keyword Distributions

Journal of Intelligent Information Systems, 10(3), pp. 281-300,
DOI: 10.1023/A:1008623632443, 1998



References

-  Franke, J., Härdle, W. and Hafner, C.
Statistics of Financial Markets
Springer, 4th ed., 2015
-  Gentle, J. E., Härdle, W. and Mori, Y.
Handbook of Computational Statistics
Springer, 2nd ed., 2012
-  Hansen, K., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F. and Sarkar, D.
Rgraphviz: Provides plotting capabilities for R graph objects, R package version 2.8.1, available on www.bioconductor.org



References

-  Hastie, T., Tibshirani, R. and Friedman, J.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Springer, 2nd ed., 2009
-  Härdle, W. and Simar, L.
Applied Multivariate Statistical Analysis
Springer, 4th ed., 2015
-  Hotho, A., Nürnberger, A. and Paass, G.
A Brief Survey of Text Mining
LDV Forum, 20(1), pp 19-62, available on www.jlcl.org, 2005



References



Miller, T., Klein, B. and Wolf, E.

Exploiting Latent Semantic Relations in Highly Linked Hypertext for Information Retrieval in Wikis

Proceedings of the International Conference RANLP-2009, pp. 241-245, 2009, available on aclweb.org



Mohamed, M. and Oussalah, M.

A Comparative Study of Conversion Aided Methods for WordNet Sentence Textual Similarity

Proceedings of the First AHA!-Workshop on Information Discovery in Text, pp. 37-42, 2014, available on aclweb.org



References



Salton, G., Allan, J., Buckley, C. and Singhal, A.
Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts
Science, 264(5164), pp. 1421-1426,
DOI: 10.1126/science.264.5164.1421, 1994



Witten, I., Paynter, G., Frank, E., Gutwin, C. and Nevill-Manning, C.
KEA: Practical Automatic Keyphrase Extraction
DL '99 Proceedings of the fourth ACM conference on Digital libraries, pp. 254-255, DOI: 10.1145/313238.313437, 1999



Zhang, J., Chen, C., Härdle, W. and Bommers, E.
Distillation of News Flow into Analysis of Stock Reactions
SFB 649 Discussion Paper 2015-005, 2015



Text to Vector

- ▣ $Q = \{d_1, \dots, d_n\}$ **set of documents** (Quantlets/Gestalten).
- ▣ $T = \{t_1, \dots, t_m\}$ **dictionary** (set of all terms).
- ▣ $tf(d, t)$ **absolute frequency** of term $t \in T$ in $d \in Q$.

	terms	Non-/sparse entries
all terms (after preprocessing)	2223	17878/2583032
discarding $tf = 1$	1416	17071/1639649
discarding $tf \leq 2$	1039	16317/1199313
discarding $tf \leq 3$	846	15738/974082

Table 1: Total number of documents in QNet: 1170 Gestalten/1826 Quantlets; term sparsity: 98% – 99%



Example: NASDAQ Text Data

Let $Q = \{d_1, d_2, d_3\}$ be the set of NASDAQ news.

The *TDM* is a 1022×3 - matrix.

Document 1: Apple text 1 (total word number: 1729)

Document 2: J. P. Morgan (total word number: 584)

Document 3: Apple text 2 (total word number: 1012)

- ▣ [NASDAQ articles source](#)
- ▣ Data available at [RDC](#)
- ▣ [Sentiment Index](#) (Distillation of News Flow into Analysis of Stock Reactions, Zhang, J., Chen, C., Härdle, W. and Bommes, E., 2015)





D3-3D-LSA for QuantNet 2.0 and GitHub

Similarity matrix M_S and Distance matrix M_D for:

all 1022 terms (in normalized TF-form):

$$M_S = \begin{pmatrix} 1 & 0.28 & 0.17 \\ 0.28 & 1 & 0.11 \\ 0.17 & 0.11 & 1 \end{pmatrix} \quad M_D = \begin{pmatrix} 0 & 1.20 & 1.29 \\ 1.20 & 0 & 1.34 \\ 1.29 & 1.34 & 0 \end{pmatrix}$$

229 special terms ($tf > 1$, in normalized TF-form):

$$M_S = \begin{pmatrix} 1 & 0.51 & 0.28 \\ 0.51 & 1 & 0.15 \\ 0.28 & 0.15 & 1 \end{pmatrix} \quad M_D = \begin{pmatrix} 0 & 0.99 & 1.20 \\ 0.99 & 0 & 1.30 \\ 1.20 & 1.30 & 0 \end{pmatrix}$$

41 special terms ($tf > 2$, in normalized TF-form):

$$M_S = \begin{pmatrix} 1 & 0.52 & 0.53 \\ 0.52 & 1 & 0.69 \\ 0.53 & 0.69 & 1 \end{pmatrix} \quad M_D = \begin{pmatrix} 0 & 0.98 & 0.96 \\ 0.98 & 0 & 0.79 \\ 0.96 & 0.79 & 0 \end{pmatrix}$$



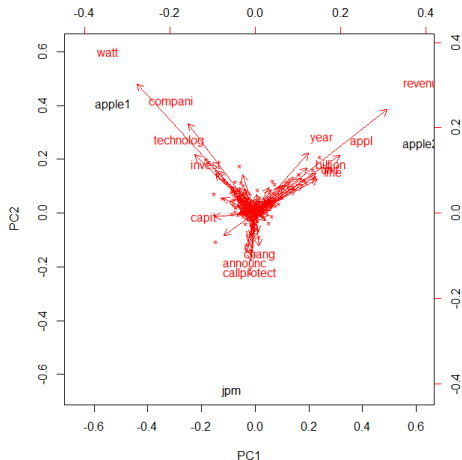


Figure 6: PCA projection of NASDAQ Texts on PC1 and PC2 (all terms)

PC1 (top 5 words): revenu, appl, line, billion, fiscal

PC2 (top 5 words): watt, revenu, compani, year, technolog

The Apple texts are well separated from J.P.M. by PC2 with words like watt, company and technology.



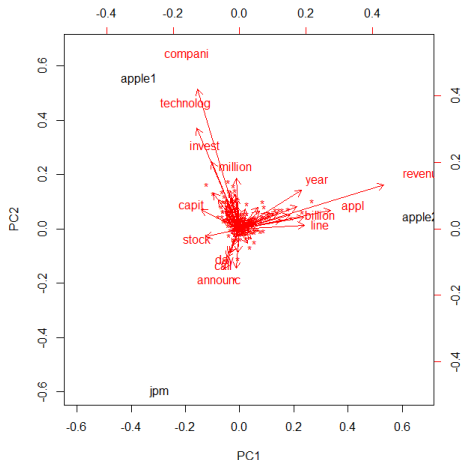


Figure 7: PCA projection of NASDAQ Texts on PC1 and PC2 (229 terms)

PC1 (top 5 words): revenu, appl, line, billion, year

PC2 (top 5 words): compani, technolog, invest, million, revenu

The Apple texts are well separated from J.P.M. by PC2 with words like company, technology and invest.



Partitional Clustering methods

- k -Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- k -Medoids clustering is related to the k -means. Both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k -means, k -medoids chooses datapoints as centers (medoids) and works with an arbitrary matrix of distances.

▶ [Back to \$M^3\$ -evaluation](#)



Hierarchical Clustering + Silhouette width

- Hierarchical cluster analysis (HCA) is a method which seeks to build a hierarchy of clusters using a set of dissimilarities for the n objects being clustered. It uses agglomeration methods like "ward.D", "ward.D2", "single", "complete", "average".
- The silhouette of a datum is a measure of how closely it is matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster, i.e. the cluster whose average distance from the datum is lowest. A silhouette close to 1 implies the datum is in an appropriate cluster, while a silhouette close to -1 implies the datum is in the wrong cluster.

► Back to M^3 -evaluation



Cluster validation measures

- The connectivity indicates the degree of connectedness of the clusters, as determined by the k -nearest neighbors. The connectedness considers to what extent observations are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between zero and ∞ and should be minimized.
- The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn Index has a value between zero and ∞ , and should be maximized.

► [Back to \$M^3\$ -evaluation](#)



t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm for nonlinear dimensionality reduction. It comprises two main stages:

1. Construct a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked, whilst dissimilar points have an infinitesimal probability of being picked.
2. Define a similar probability distribution over the points in the low-dimensional map, and minimize the Kullback-Leibler divergence between the two distributions with respect to the locations of the points in the map.

► [Back to LSA geometry](#)



Data Mining: DM

DM is the computational process of discovering/representing patterns in large data sets involving methods at the intersection of **artificial intelligence, machine learning, statistics, and database systems.**

1. Numerical DM
2. Visual DM
3. Text Mining
(applied on considerably weaker structured text data)



Text Mining

Text Mining or **Knowledge Discovery from Text (KDT)** deals with the machine supported analysis of text (Feldman et al., 1995).

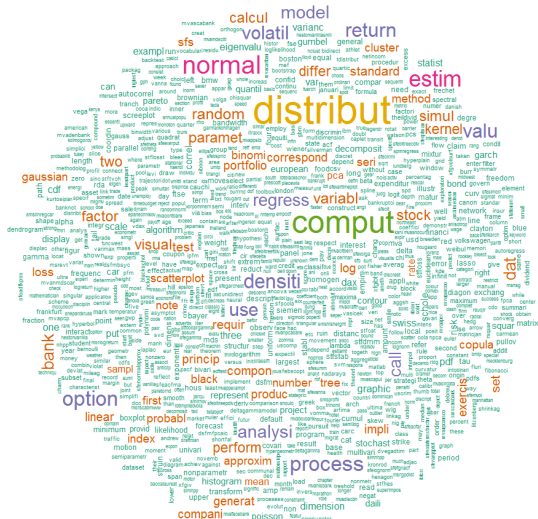
It uses techniques from:

- ▣ Information Retrieval (IR)
- ▣ Information extraction
- ▣ Natural Language Processing (NLP)

and connects them with the methods of DM.



Wordcloud of the words/terms in QNet



Most frequent words/terms in QNet

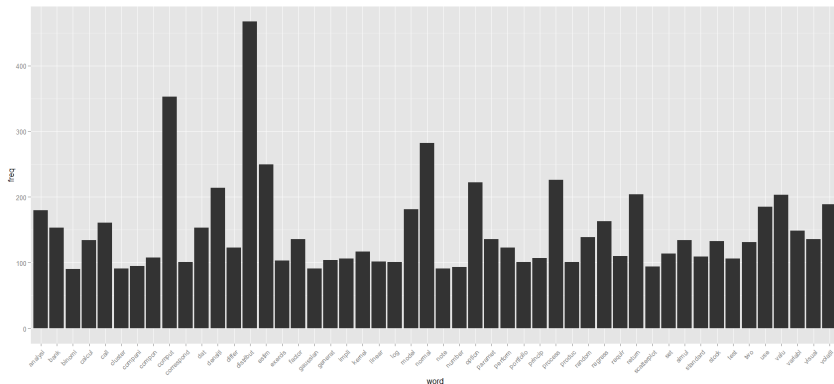


Figure 8: Words with more then 90 occurrences

Correlation graph of the QNet terms

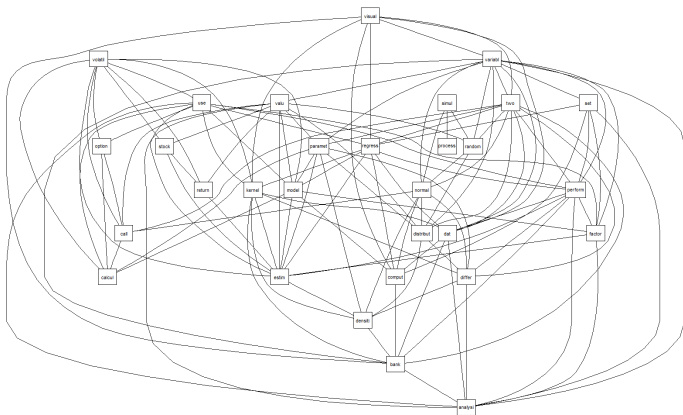


Figure 9: 30 most frequent terms with threshold = 0.05



Distance measure

A frequently used distance measure is the **Euclidian distance**:

$$dist_d(d_1, d_2) \stackrel{\text{def}}{=} dist\{w(d_1), w(d_2)\} \stackrel{\text{def}}{=} \sqrt{\sum_{k=1}^m \{w(d_1, t_k) - w(d_2, t_k)\}^2}$$

It holds for tf-idf:

$$\cos \phi = \frac{x^\top y}{|x| \cdot |y|} = 1 - \frac{1}{2} dist^2 \left(\frac{x}{|x|}, \frac{y}{|y|} \right),$$

where $\frac{x}{|x|}$ means $w(d_1)$, $\frac{y}{|y|}$ means $w(d_2)$ and $\cos \phi$ is the angle between x and y .



Drawbacks of the classical tf-idf approach

- Uncorrelated/orthogonal terms in the feature space
- Documents must have common terms to be similar
- Sparsity of document vectors and similarity matrices

Solution

- Using statistical information about term-term correlations
- Incorporating information about semantics (Semantic smoothing)



GVSM – Basic VSM (BVSM)

- $P = I_m$ and $w(d) = \{tf(d, t_1), \dots, tf(d, t_m)\}^\top$
classical tf-similarity: $M_S^{tf} = D^\top D$
- diagonal $P(i, i)^{idf} = idf(t_i)$ and
 $w(d) = \{tf(d, t_1), \dots, tf(d, t_m)\}^\top$
classical tf-idf-similarity: $M_S^{tfidf} = D^\top (P^{idf})^\top P^{idf} D$
- starting with
 $w(d) = \{tf(d, t_1)idf(t_1), \dots, tf(d, t_m)idf(t_m)\}^\top$
and letting $P = I_m$:
 $M_S^{tfidf} = D^\top I_m D = D^\top D$

► Back to BVSM



GVSM – term-term correlations

- $P = D^T$
- $S(d_1, d_2) = (D^T d_1)^T (D^T d_2) = d_1^T D D^T d_2$
- $M_S^{TT} = D^T (D D^T) D$
- $D D^T$ – term by term matrix, having a nonzero ij entry if and only if there is a document containing both the i -th and the j -th terms
- terms become semantically related if co-occurring often in the same documents
- also known as a dual space method (Sheridan and Ballerini, 1996)
- when there are less documents than terms – dimensionality reduction

► [Back to GVSM\(TT\)](#)



GVSM – Latent Semantic Analysis (LSA)

- LSA measures semantic information through co-occurrence analysis (Deerwester et al., 1990)
- Technique – singular value decomposition (SVD) of the matrix $D = U\Sigma V^T$
- $P = U_k^T = I_k U^T$ – projection operator onto the first k dimensions
- $M_S = D^T(U I_k U^T)D$ – similarity matrix
- It can be shown: $M_S = V\Lambda_k V^T$, with $D^T D = V\Sigma^T U^T U \Sigma V^T = V\Lambda V^T$ and $\Lambda_{ii} = \lambda_i$ eigenvalues of V ; Λ_k consisting of the first k eigenvalues and zero-values else.

► [Back to GVSM\(LSA\)](#)



Latent Semantic Space

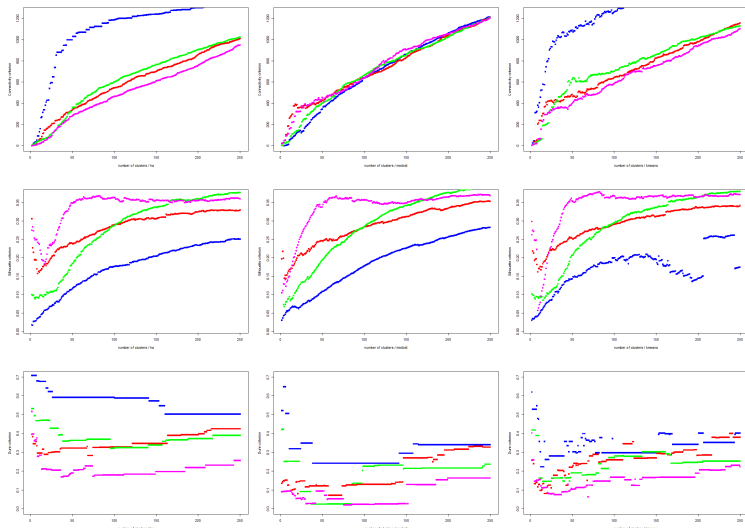
1. Create directly by using the quantlets, matrix D = the set of quantlets
2. First train by domain-specific and generic background documents
 - ▶ Fold in Quantlets into the semantic space after the previous SVD process
 - ▶ Gain of higher retrieval performance (bigger vocabulary set, more semantic structure)
 - ▶ Chapters or sub-chapters from our e-books well suited



Generalized VSM – Semantic smoothing

- More natural method of incorporating semantics is by directly using a semantic network
- (Miller et al., 1993) used the semantic network WordNet
- Term distance in the hierarchical tree provided by WordNet gives an estimation of their semantic proximity
- (Siolas and d'Alche-Buc, 2000) have included the semantics into the similarity matrix by handcrafting the VSM matrix P
- $M_S = D^T(P^T P)D = D^T P^2 D$ – similarity matrix





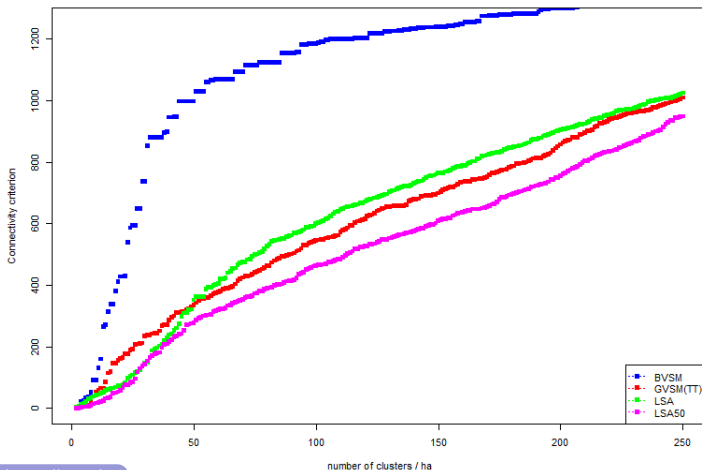
Rows: Connectivity, Silhouette, Dunn; Columns: hca, k -medoids, k -means

Colors: BVSM, GVSM(TT), LSA, LSA50; [► more details](#) [► Back to Slides](#)

D3-3D-LSA for QuantNet 2.0 and GitHub



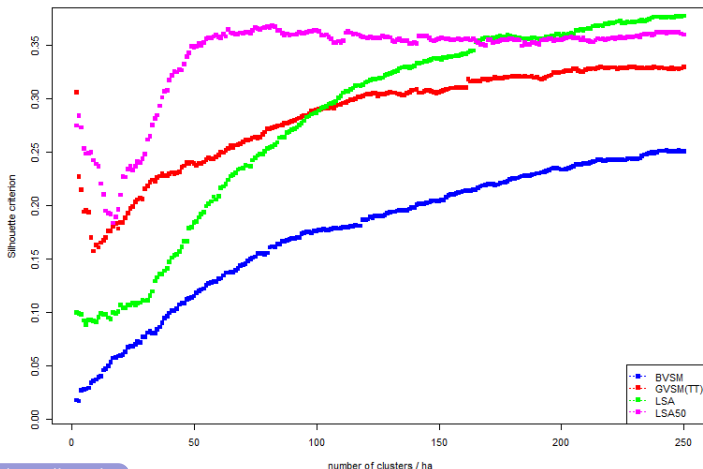
Connectivity performance - hca



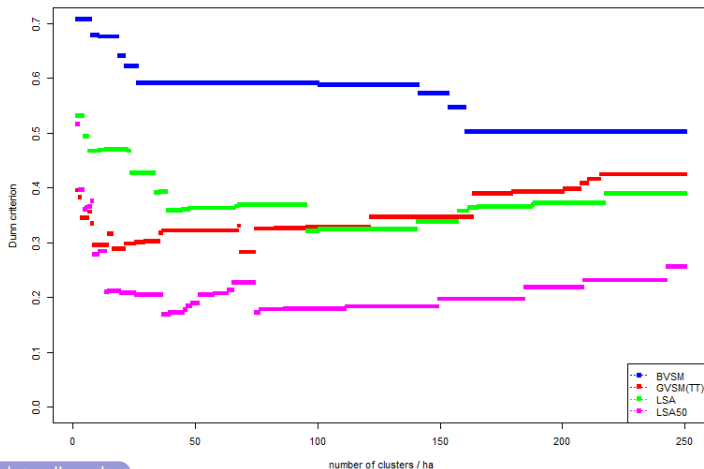
[Back to all results](#)



Silhouette performance - hca

[▶ Back to all results](#)

Dunn performance - hca

[Back to all results](#)

LSA - A first insight into the interpretation

Coincidence of the terms
in the semantic space principal components
and in the labels of the dendrogram clusters

PC1 (5.6): visual return option call distribut

PC2 (4.9): call option blackschol put price

PC3 (4.5): dsfm fpca dsfmbysc dsfmfpcaic cluster

PC4 (4.4): dsfm copula distribut densiti gumbel

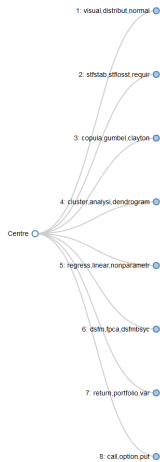
PC5 (4.3): return visual portfolio timeseri dax

PC6 (4.1): regress kernel nonparametr linear estim

PC7 (4.0): copula regress gumbel nonparametr var

PC8 (3.9): copula visual gumbel scatterplot clayton

PC number	cluster number
1	1
2	8
3	6
4	3
5	7
6	5
7	3, 5
8	3



Sparsity results

	BVSM	TT	LSA:300	LSA:171(50%)	LSA:50
TDM	0.99	0.65	0.51	0.51	0.47
M_S	0.65	0.07	0.35	0.36	0.35

Table 2: Model Performance regarding the sparsity of the term by document matrix TDM and the similarity matrix M_S in the appropriate models (weighting scheme: tf-idf normed).

Sparsity: the ratio of the number of zero entries to the total number of entries of a matrix. In general: the lower the sparsity, the better.

More details about sparsity and similarity structure in

▶ BVSM

▶ GVSM(TT)

▶ GVSM(LSA:300)

▶ GVSM(LSA:155(50%))

▶ GVSM(LSA:50)



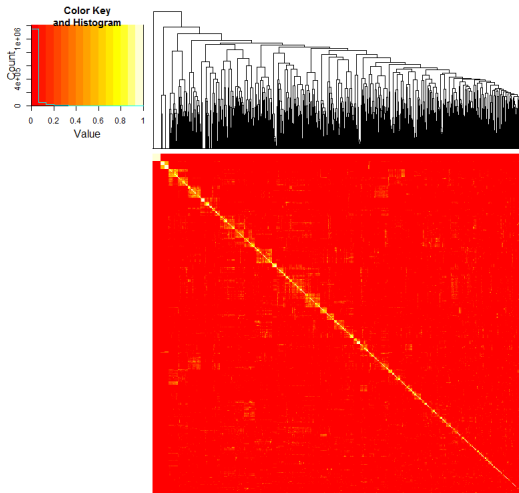


Figure 10: Heat map with Dendrogram - BVSM SimMatrix

[▶ Back to sparsity results](#)



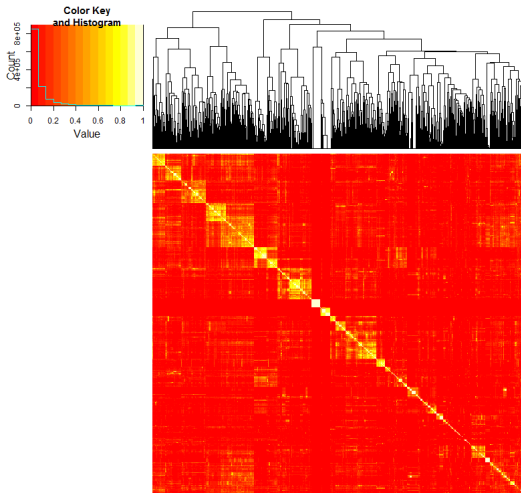


Figure 11: Heat map with Dendrogram - GVSM(TT) SimMatrix

[▶ Back to sparsity results](#)



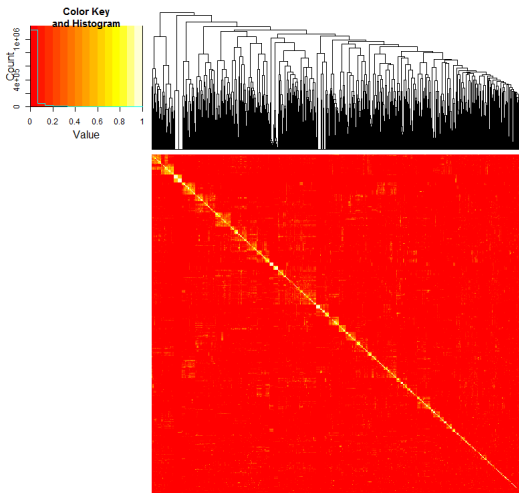


Figure 12: Heat map with Dendrogram - LSA:300 SimMatrix

[▶ Back to sparsity results](#)



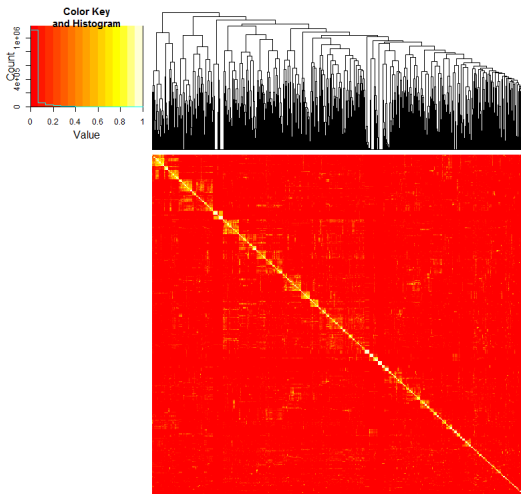


Figure 13: Heat map with Dendrogram - LSA:155(50%) SimMatrix

[▶ Back to sparsity results](#)



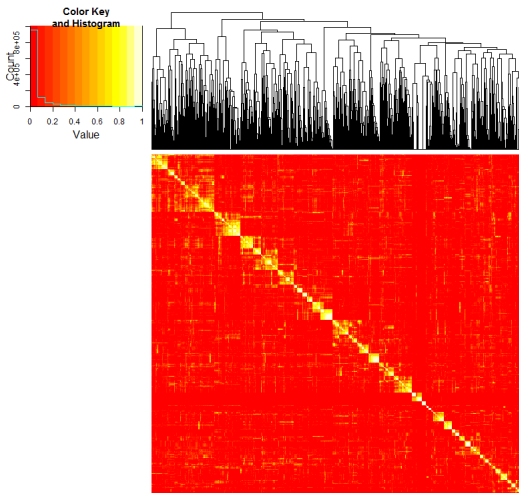


Figure 14: Heat map with Dendrogram - LSA:50 SimMatrix

[▶ Back to sparsity results](#)

