Cluster Analysis in Loss Development

Dave Clark
Victoria Jiang
Munich Reinsurance America Inc.

RPM Conference – March 2019

Munich RE

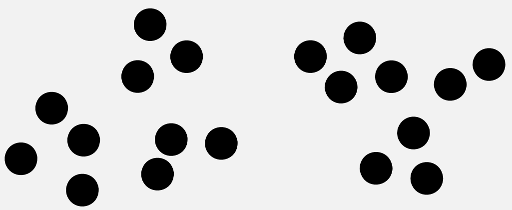
Agenda

Munich RE

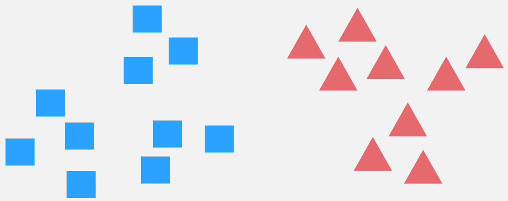
- Introduction
- How to find clusters:
  - Cluster analysis
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

Introduction
What is Clustering?

- Clustering is about finding groups in a set of objects
  - The objects in a group should be similar and groups should be different from each other
  - No need to define the groups in advance (i.e. unsupervised learning)
  - Essential to assess the usefulness and meaning of the identified groups



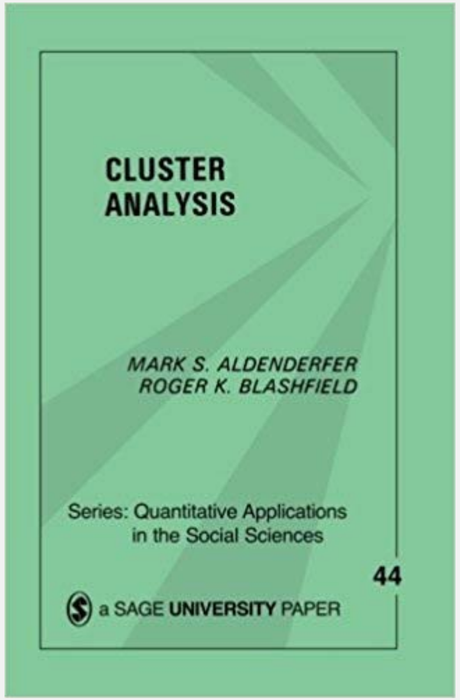
Original data



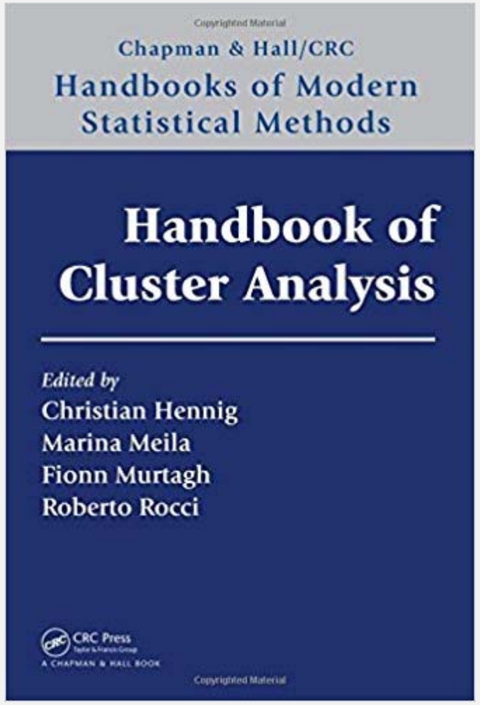
Two clusters

Introduction
Publications on Clustering

Cluster Analysis has grown rapidly, especially as computer software has become more readily available.



1984 - **88 pages**



2015 - **773 pages**

Introduction
Why Clustering?

- What questions could be answered with cluster analysis?
  - Test the data homogeneity
  - Find a benchmark
- What kind of data can be clustered?
  - Segments, contracts or claims
  - County or Region
  - Loss development patterns, loss ratios, severity, frequency…

Agenda

Munich RE

- Introduction
- How to find clusters:
  - Cluster analysis – Schedule P example
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

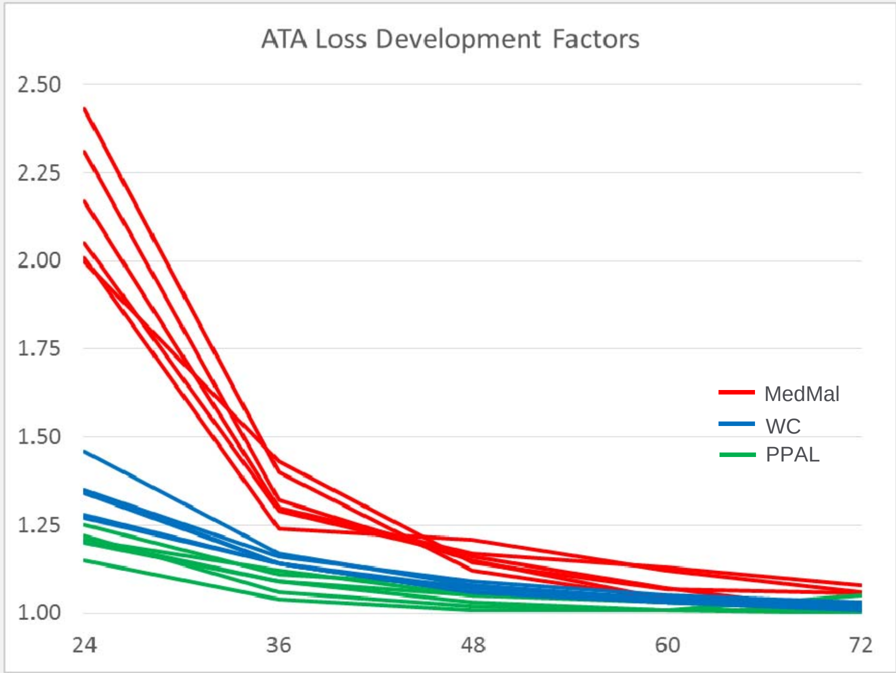
Cluster Analysis
Schedule P (Annual Statement) Example

Co.	Line	Ownership	Geographic	Distribution
1	MedMal	Mutual	Regional	Direct, Ind Agency
2	MedMal	Stock	National	Direct, Ind Agency
3	PPAL	Stock	National	MGA, Ind Agency
4	PPAL	Stock	Regional	Ind Agency
5	WC	Stock	National	MGA
6	WC	Mutual	Regional	Ind Agency

…

Co.	24	36	48	60	72
1	2.01	1.24	1.21	1.12	1.06
2	2.05	1.29	1.16	1.07	1.00
3	1.20	1.09	1.05	1.03	1.01
4	1.15	1.04	1.01	1.01	1.00
5	1.34	1.14	1.07	1.04	1.02
6	1.28	1.14	1.06	1.04	1.02

…



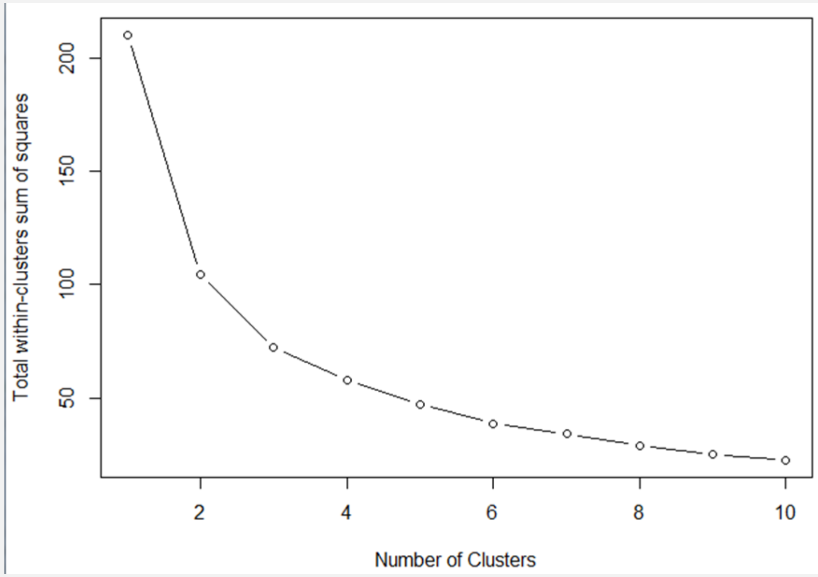
Cluster Analysis
Preparing Data

Explanatory Variables					Variables used for clustering, PCA, …				
Co.	**Line**	**Ownership**	**Geographic**	**Distribution**	**24**	**36**	**48**	**60**	**72**
1	MedMal	Mutual	Regional	Direct, Ind Agency	2.01	1.24	1.21	1.12	1.06
2	MedMal	Stock	National	Direct, Ind Agency	2.05	1.29	1.16	1.07	1.00
3	PPAL	Stock	National	MGA, Ind Agency	1.20	1.09	1.05	1.03	1.01
4	PPAL	Stock	Regional	Ind Agency	1.15	1.04	1.01	1.01	1.00
5	WC	Stock	National	MGA	1.34	1.14	1.07	1.04	1.02
6	WC	Mutual	Regional	Ind Agency	1.28	1.14	1.06	1.04	1.02
			…				…		

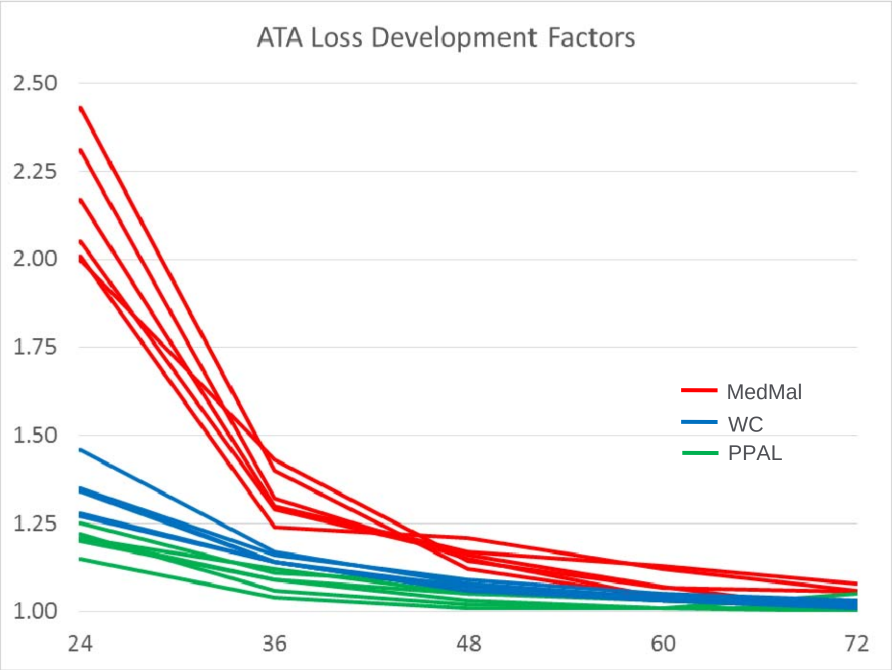
Cluster Analysis
K-means Algorithm

- K-means partitions the data in a user-specified number of clusters (K), in which each observation belongs to the cluster with the nearest mean

- No definitive answer for selecting K
  - Scree plot: locate the sharpest drop in within-cluster sum of squares



Cluster Analysis
K-means Clustering Results



LOB	K-means 2 clusters	K-means 3 clusters	K-medoids 3 clusters
MedMal	1	1	1
MedMal	1	1	1
MedMal	1	2	1
MedMal	1	1	1
MedMal	1	2	1
MedMal	1	2	1
PPAL	2	3	2
PPAL	2	3	2
PPAL	2	3	2
PPAL	2	3	2
PPAL	2	3	2
PPAL	2	3	2
WC	2	3	3
WC	2	3	3
WC	2	3	3
WC	2	3	3
WC	2	3	3
WC	2	3	3

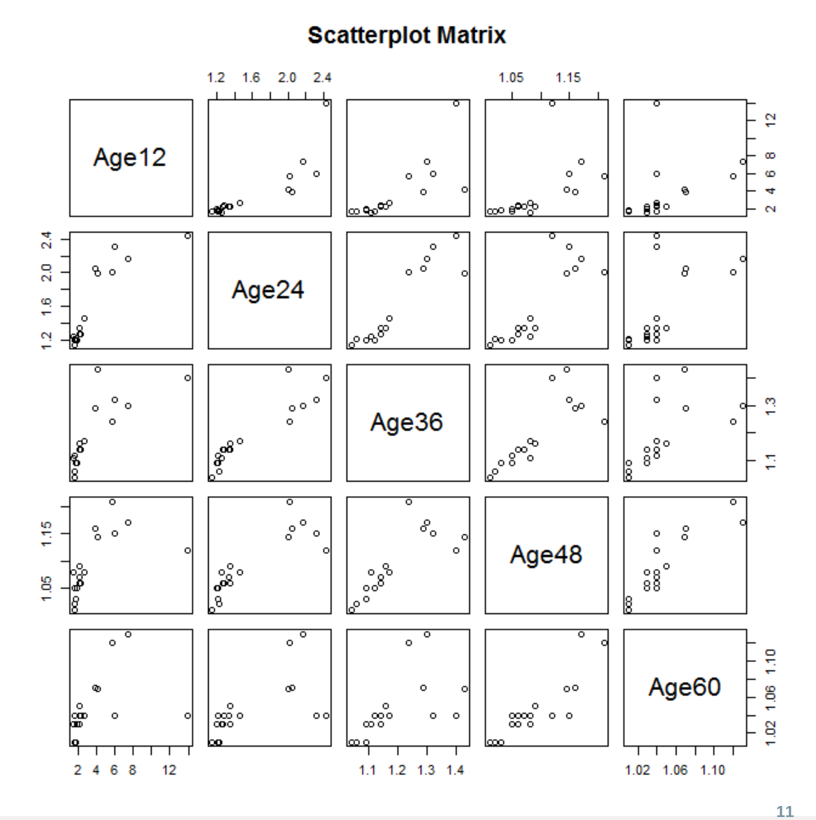
Cluster Analysis
Too Many Dimensions

Munich RE

- Difficulty visualizing more than two dimensions for validation purposes

12	24	36	48	60	72
5.70	2.01	1.24	1.21	1.12	1.06
3.86	2.05	1.29	1.16	1.07	1.00
1.92	1.20	1.09	1.05	1.03	1.01
1.64	1.15	1.04	1.01	1.01	1.00
2.19	1.34	1.14	1.07	1.04	1.02
2.33	1.28	1.14	1.06	1.04	1.02

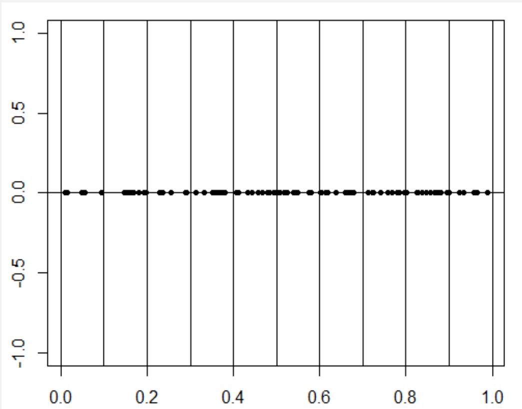
…

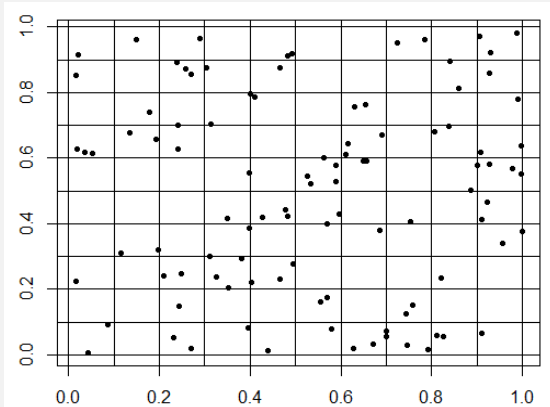


Cluster Analysis
Too Many Dimensions

- Data gets “lost in space”

Randomly generated 100 points in 1D and 2D





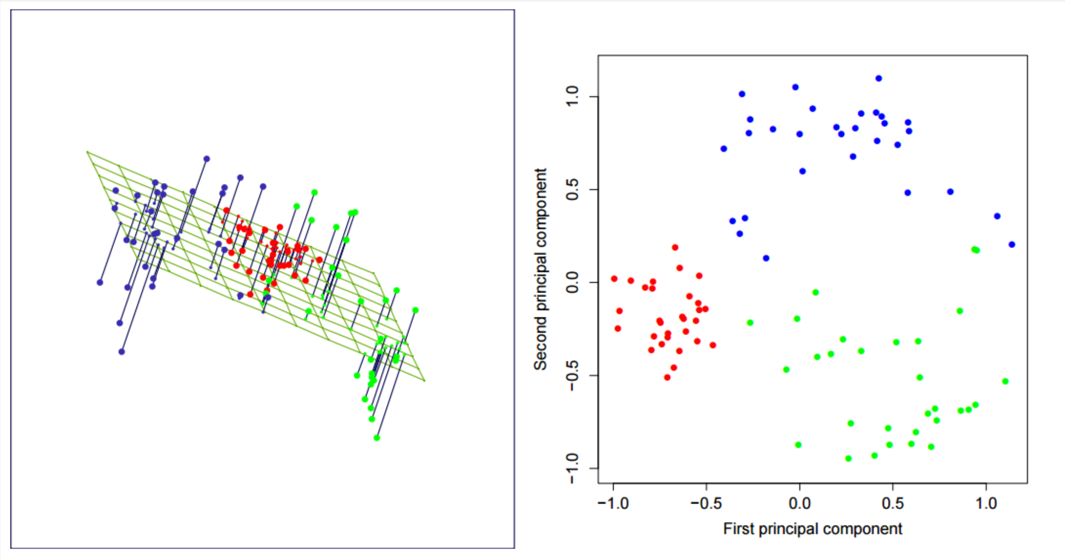
- The performance of clustering algorithms relying on L_1 (sum of absolute values) or L_2 (Euclidian) metrics in high dimensional data may be compromised

Agenda

- Introduction
- How to find clusters:
  - Cluster analysis – Schedule P example
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

PCA
Principal Component Analysis

- **PCA stretches and rotates data** with the goal to derive the best possible k-dimensional representation of the Euclidean distance among objects.



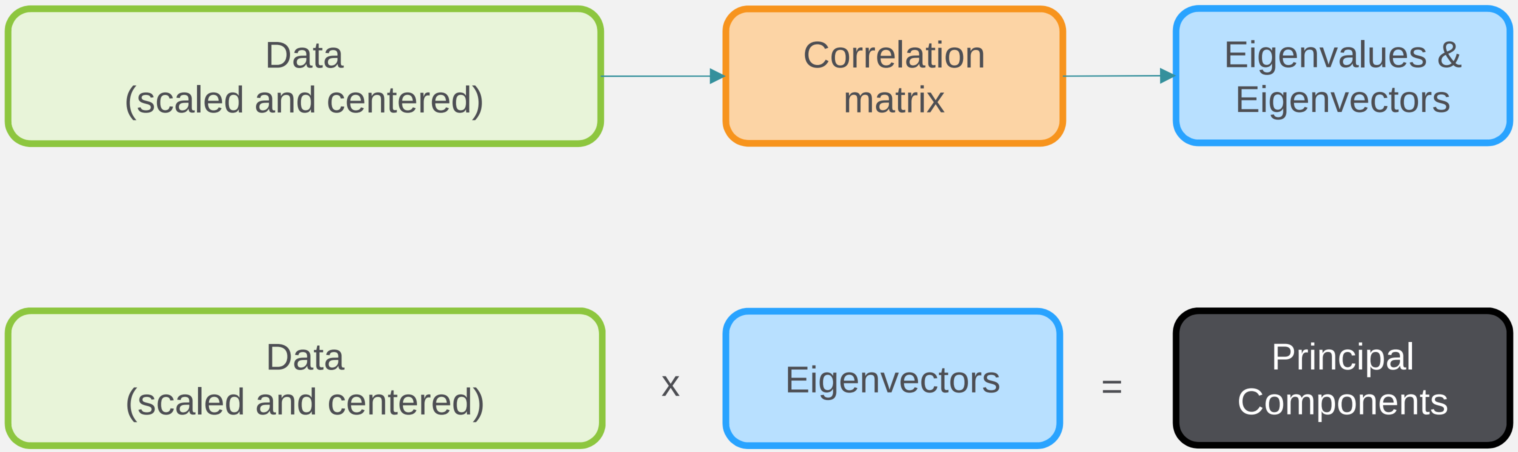
Source: The Elements of Statistical Learning

PCA
Principal Component Analysis

➢ Think about viewing a galaxy from “above” rather than the side: what angle do we want in order to get the most understanding of the “shape” of the galaxy?

Source: https://www.nasa.gov/feature/goddard/2017/a-new-angle-on-two-spiral-galaxies-for-hubbles-27th-birthday

PCA
How to perform a PCA?



Munich RE

- PCA provides an opportunity for interpretation
  - PC1 captures the mean loss development
  - PC2 indicates a change in the loss curve shape

Data (scaled and centered) x Eigenvectors = Principal Components

Co	24	36	48	60	72
1	0.99	0.46	2.07	2.18	1.43
2	1.08	0.89	1.17	0.66	-1.04
3	-0.83	-0.82	-0.75	-0.60	-0.78
4	-0.94	-1.30	-1.39	-1.14	-0.96
			…		

x

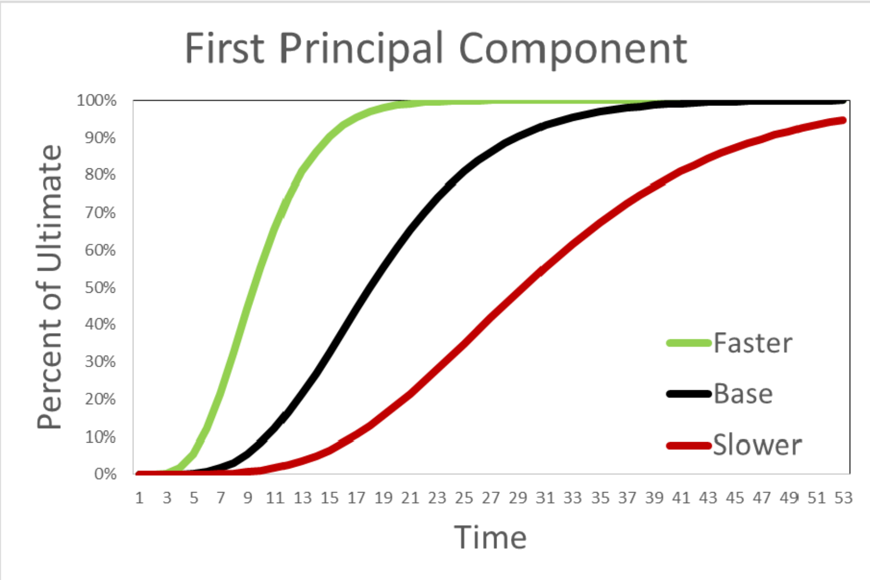
Dim	1	2
24	0.47	-0.39
36	0.46	-0.38
48	0.49	-0.12
60	0.46	0.36
72	0.34	0.75

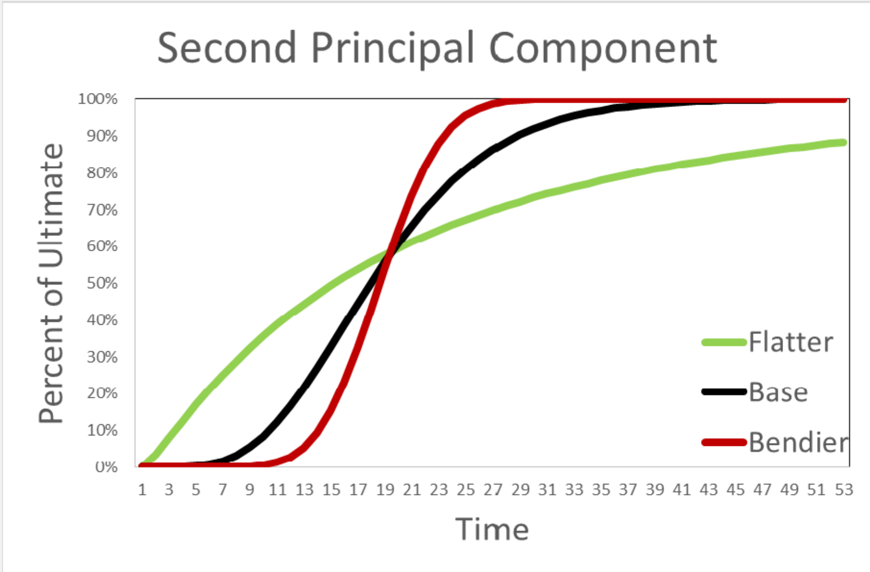
=

Co	PC1	PC2
1	3.18	1.04
2	1.45	-1.44
3	-1.67	-0.07
4	-2.57	-0.10
	…	

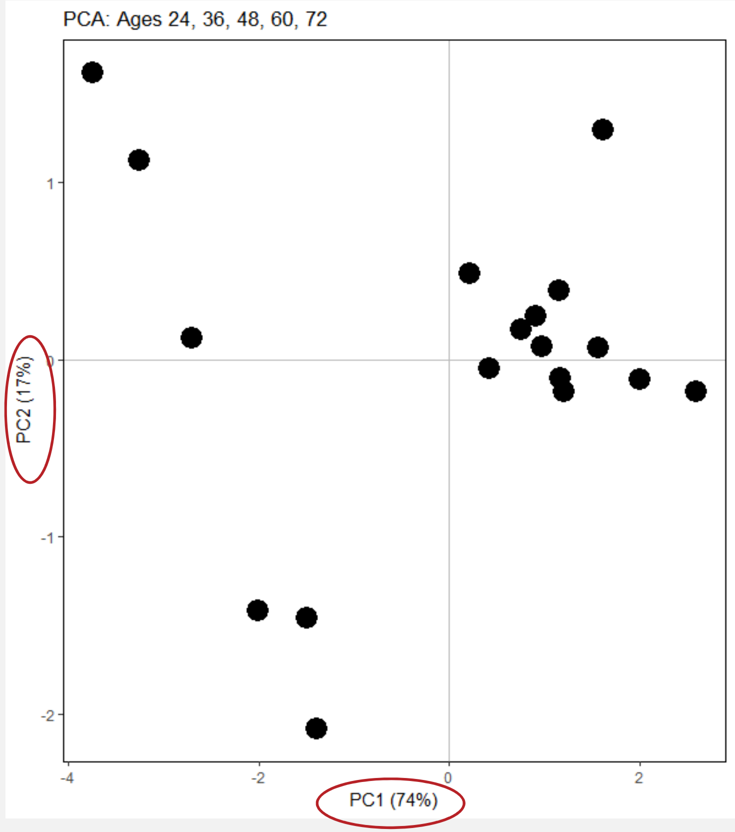
PCA

Interpretation





PCA
Schedule P example: Visualization



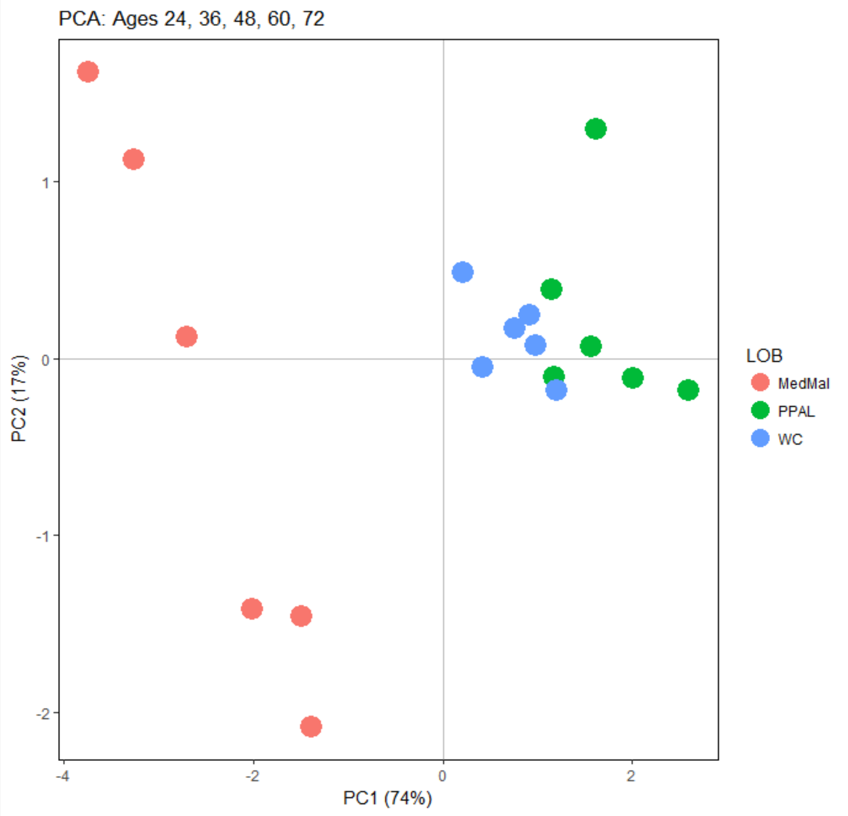
PCA
Explanatory Variables

Munich RE

Explanatory Variables

Co.	Line	Ownership	Geographic	Distribution	24	36	48	60	72
1	MedMal	Mutual	Regional	Direct, Ind Agency	2.01	1.24	1.21	1.12	1.06
2	MedMal	Stock	National	Direct, Ind Agency	2.05	1.29	1.16	1.07	1.00
3	PPAL	Stock	National	MGA, Ind Agency	1.20	1.09	1.05	1.03	1.01
4	PPAL	Stock	Regional	Ind Agency	1.15	1.04	1.01	1.01	1.00
5	WC	Stock	National	MGA	1.34	1.14	1.07	1.04	1.02
6	WC	Mutual	Regional	Ind Agency	1.28	1.14	1.06	1.04	1.02
			…				…		

PCA
Schedule P example: Visualization - LOB



Agenda

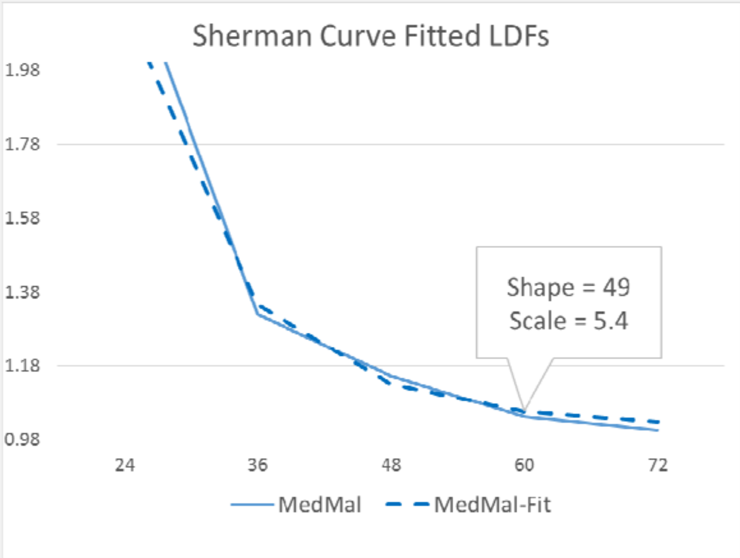
Munich RE

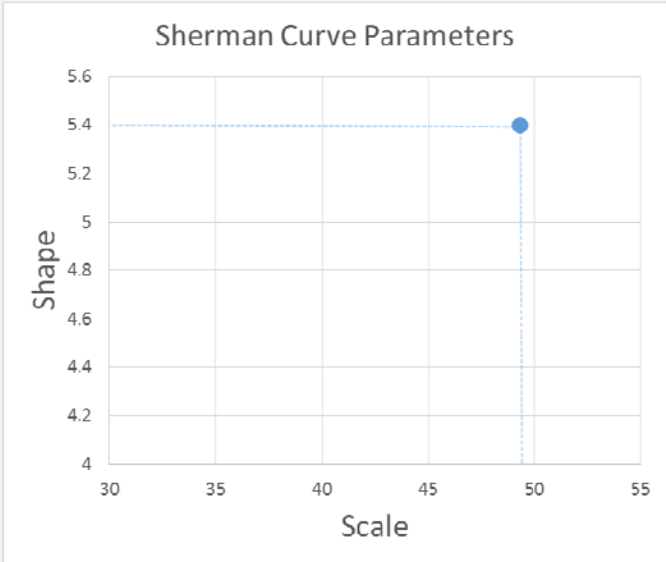
- Introduction
- How to find clusters:
  - Cluster analysis – Schedule P example
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

Data Transformation
Sherman Curve

- Sherman proposed a curve that fits to the typical LDF pattern

$$ATA_t = 1 + \left(\frac{Scale}{t + c}\right)^{Shape}$$





Data Transformation

How to estimate the parameters?

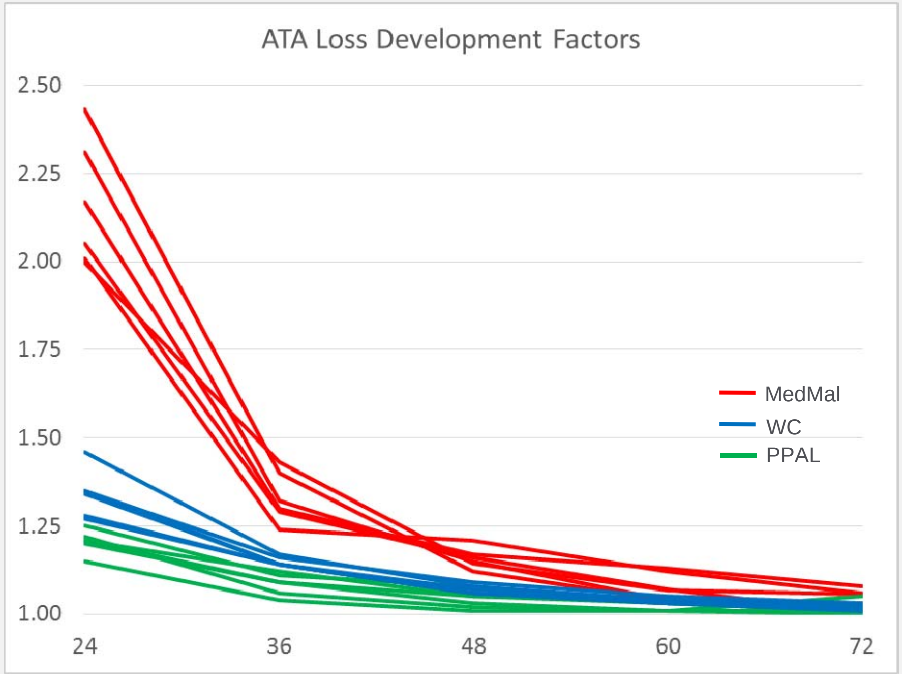
- Sherman recommends estimating the parameters by using log-linear regression
  - All actual age-to-age factors must be strictly greater than 1
  - Fitting a logged value rather than actual amounts
- GLM to the rescue!
  - Apply GLM with log-link on actual data

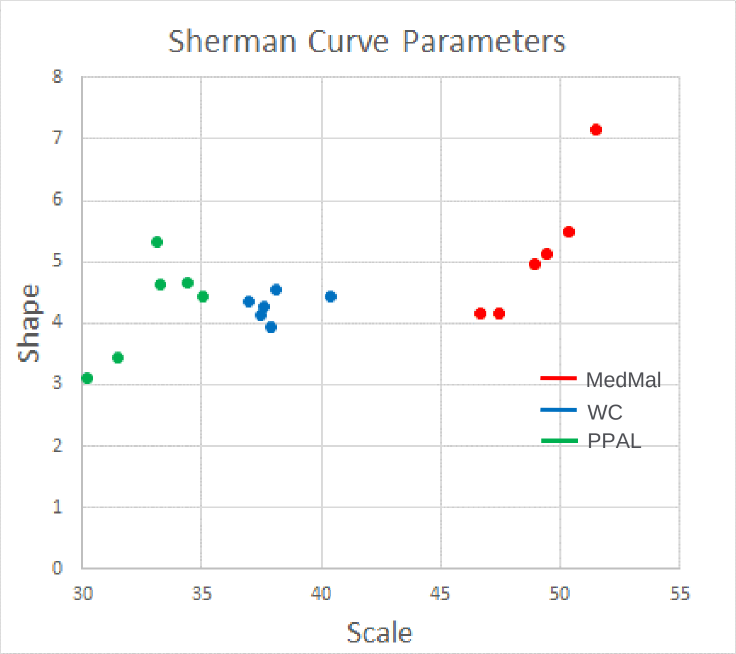
Data Transformation
Pros & Cons

Munich RE

- Allows comparison of loss development patterns of different sizes
- Does not work well for flat curves
- The focus is on the fit and not on maintaining the distances between points

Data Transformation
Schedule P example: Sherman curve





Agenda

Munich RE

- Introduction
- How to find clusters:
  - Cluster analysis – Schedule P example
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

Practical Considerations

Correlations between lines of business

- Compare the first principal component for two different lines, written by the same company

- Schedule P data for loss reserving posted on the CAS website
  - 54 companies with CAL and GL lines
  - 20 companies with WC and GL lines
  - Data is from 1988 to 1997

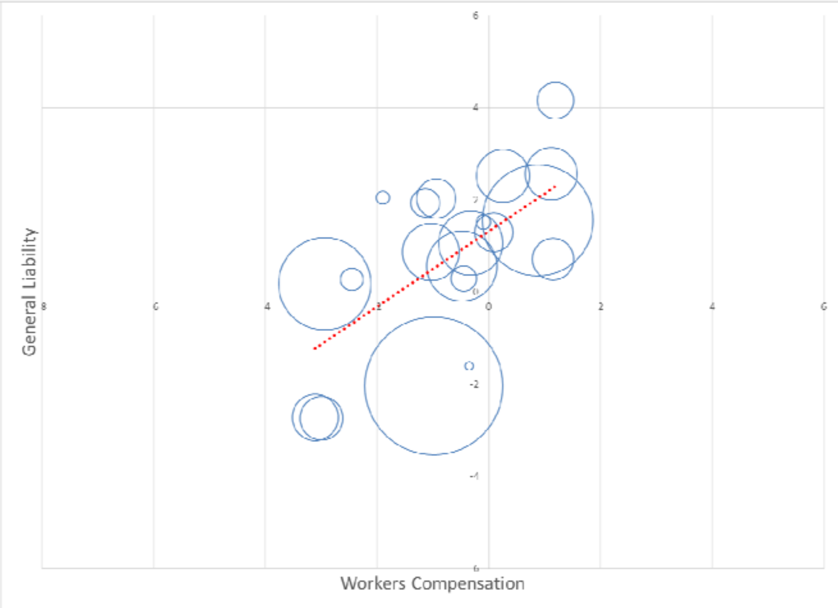
- Check if historical dependency is preserved in more recent years

Practical Considerations

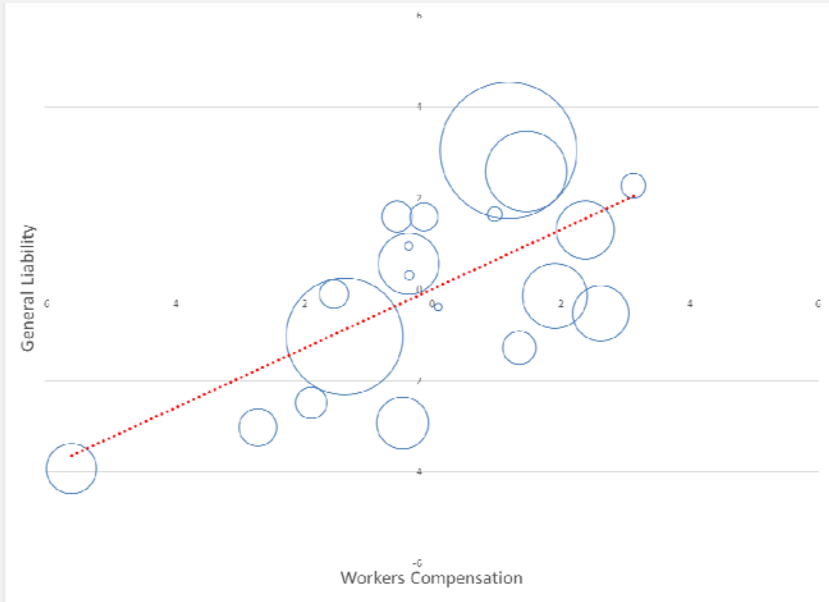
First principal component for WC/GL

- PCA on Reported loss

1988 - 1997



1998 - 2007



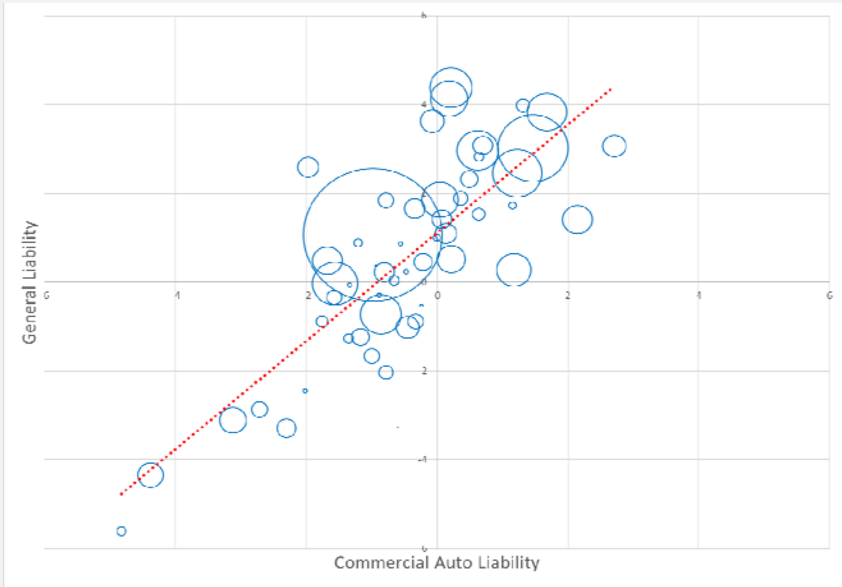
Note: bubble size corresponds to a company's average yearly premium volume

Practical Considerations
First principal component for CAL/GL

➤ PCA on Reported loss

1988 - 1997

1998 - 2007



Note: bubble size corresponds to a company's average yearly premium volume

Agenda

- Introduction
- How to find clusters:
  - Cluster analysis – Schedule P example
  - Principal Component Analysis (PCA)
  - Data transformation (curve fitting)
- Practical considerations
  - Correlations between LoB
  - Identifying drivers of loss development

Practical Considerations

What are the drivers of loss development?

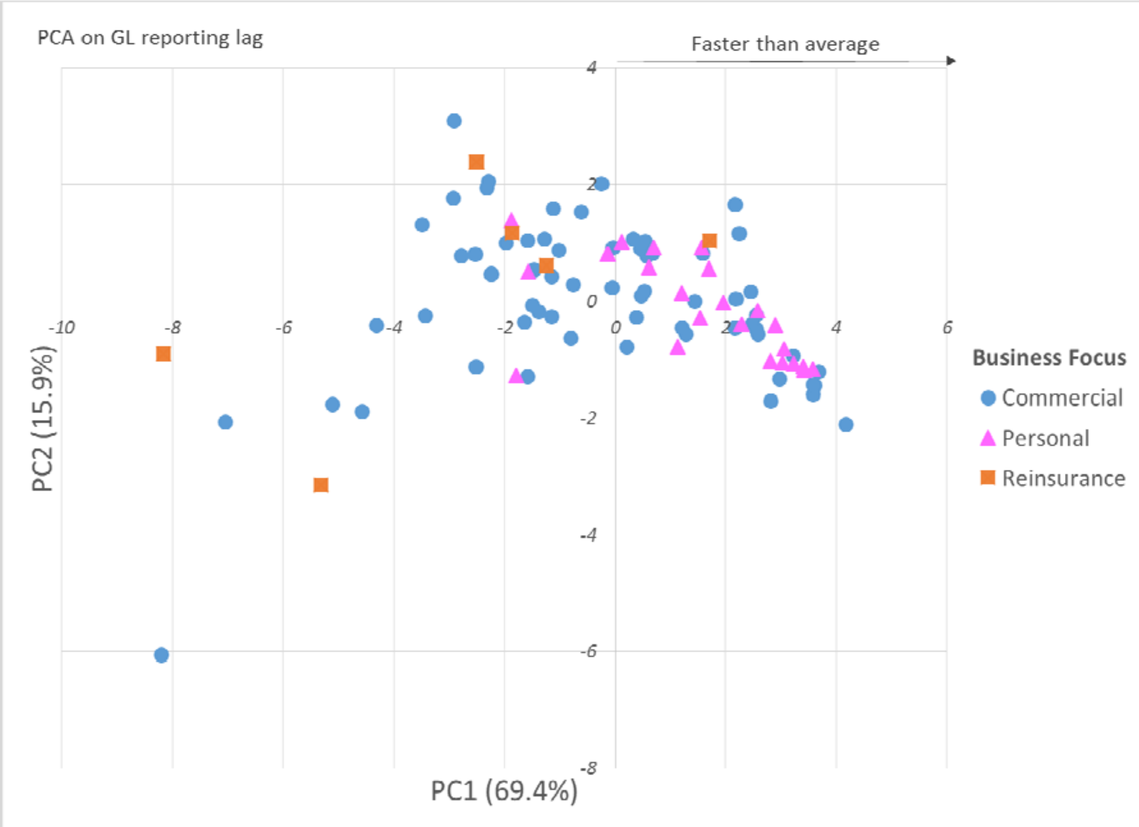
- Identify potential predictors
  - Business focus (Commercial, Personal, Reinsurance)
  - Ownership (Stock, Mutual, Other)
  - Distribution channel (Broker vs Non-Broker)
  - Geography (Regional vs National)

- Schedule P GL data & SNL company profile
  - Top 100 insurers by market share
  - Loss data is from 2008 to 2017

Practical Considerations

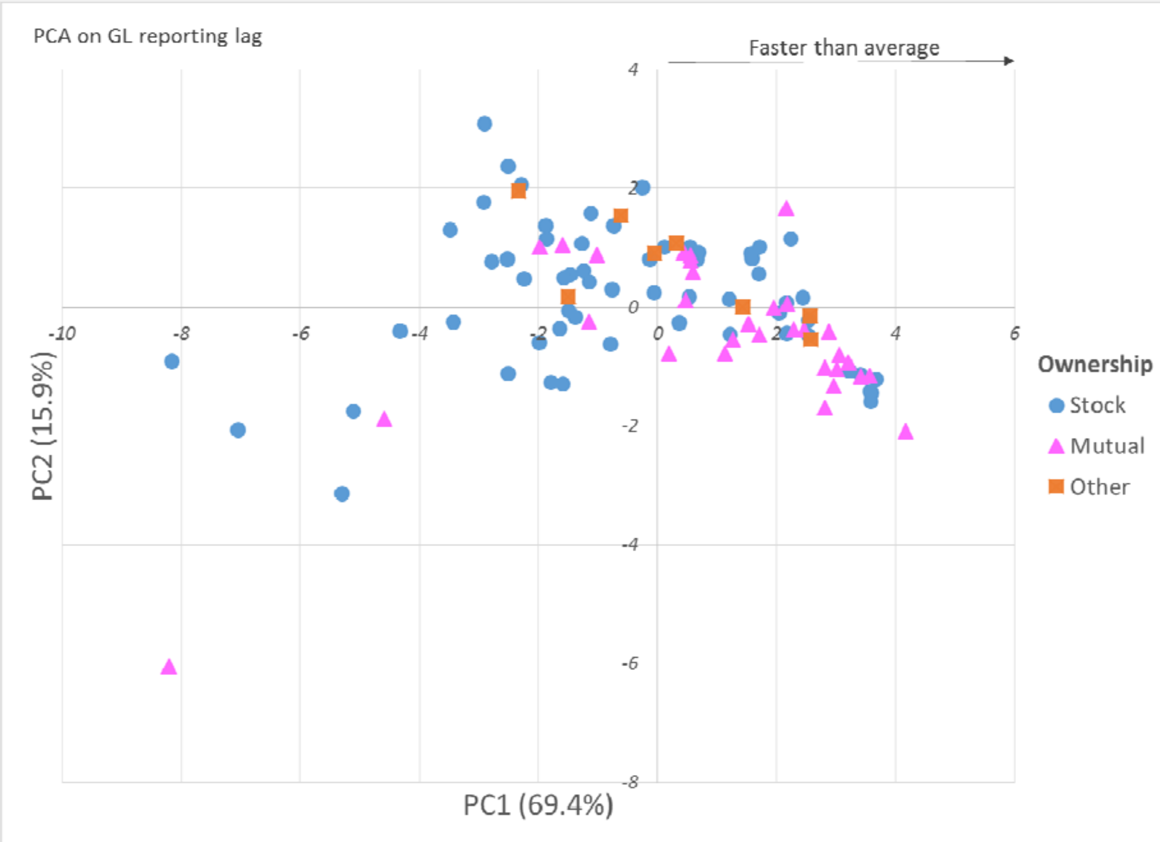
Visualization: are the explanatory variables logical?

Munich RE



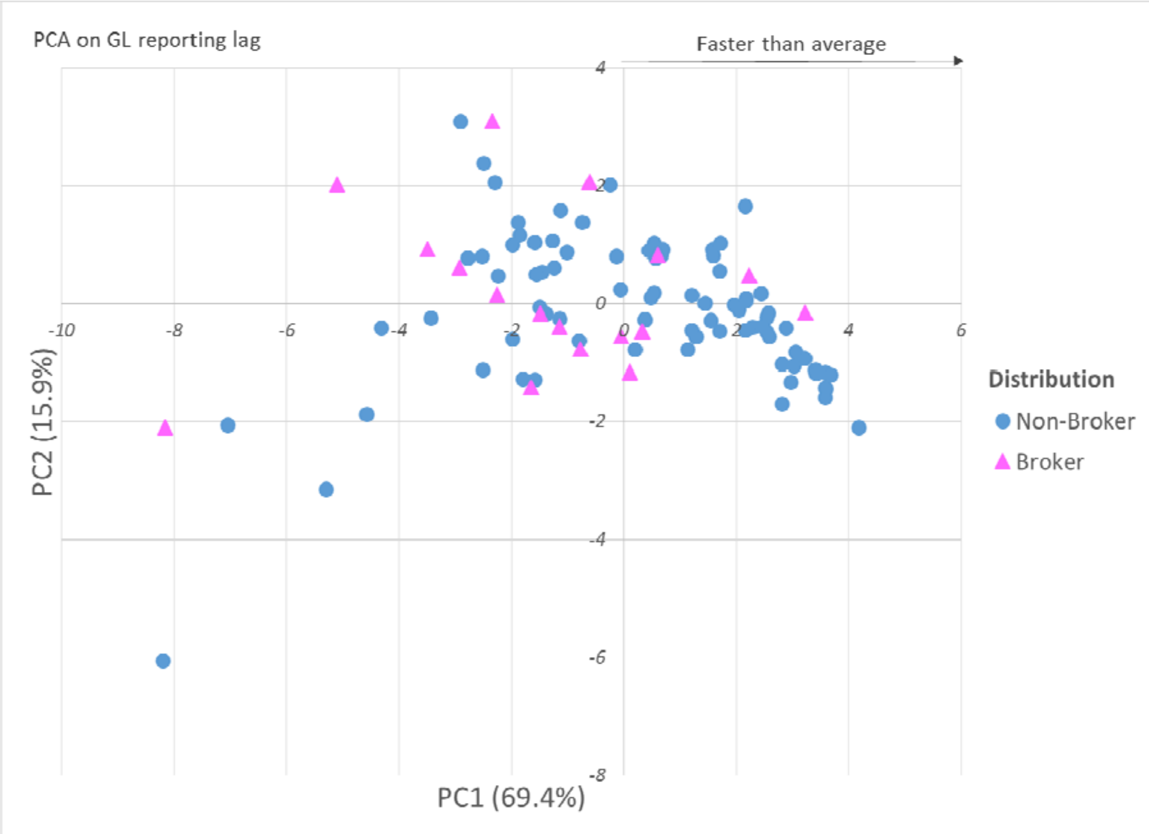
Practical Considerations
Visualization: are the explanatory variables logical?

Munich RE



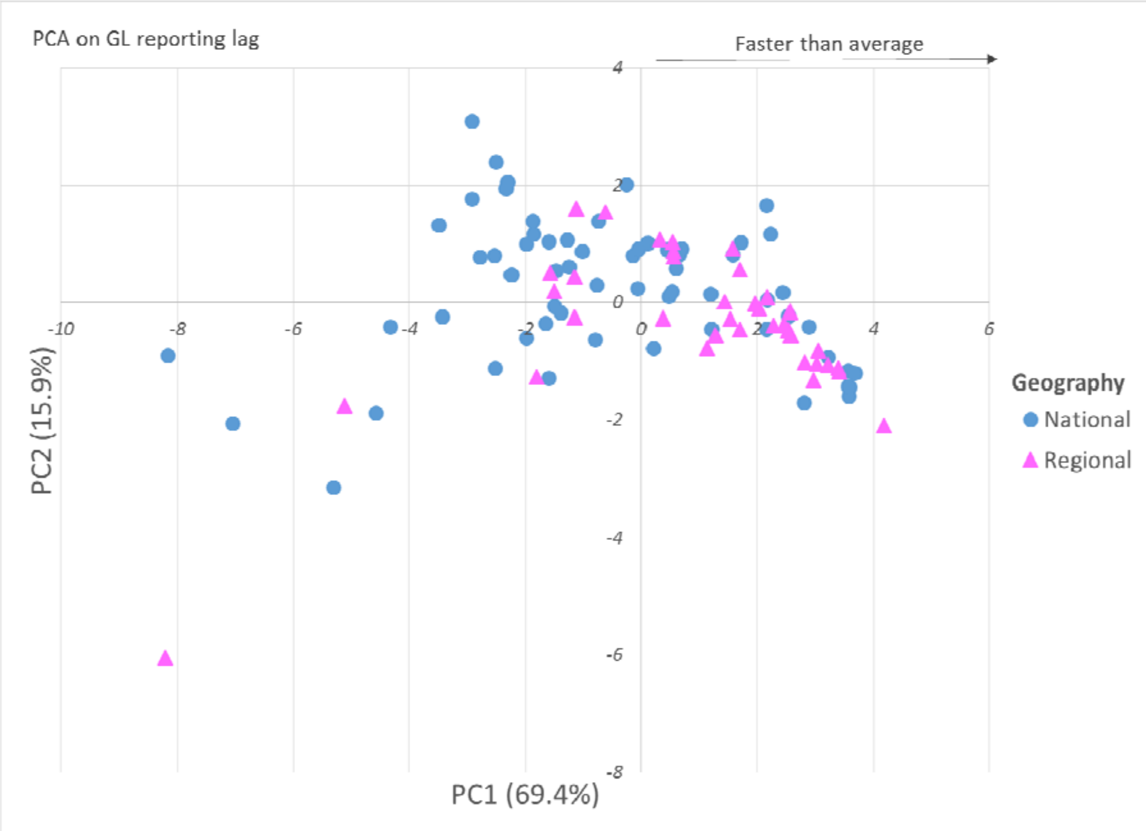
Practical Considerations

Visualization: are the explanatory variables logical?



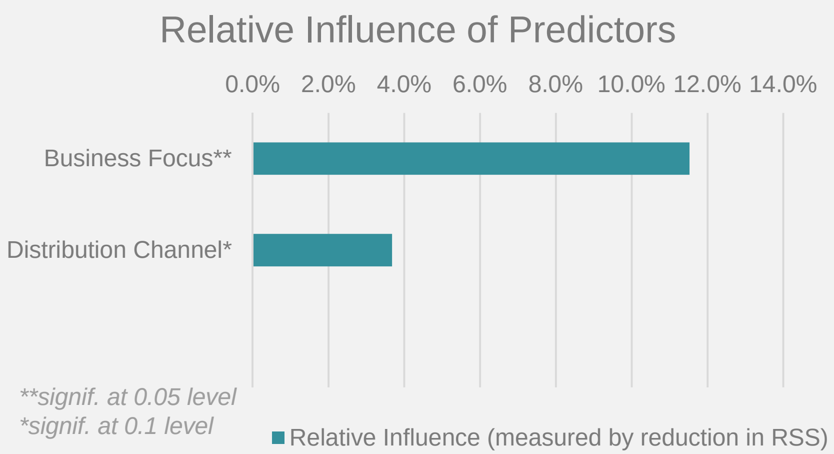
Practical Considerations

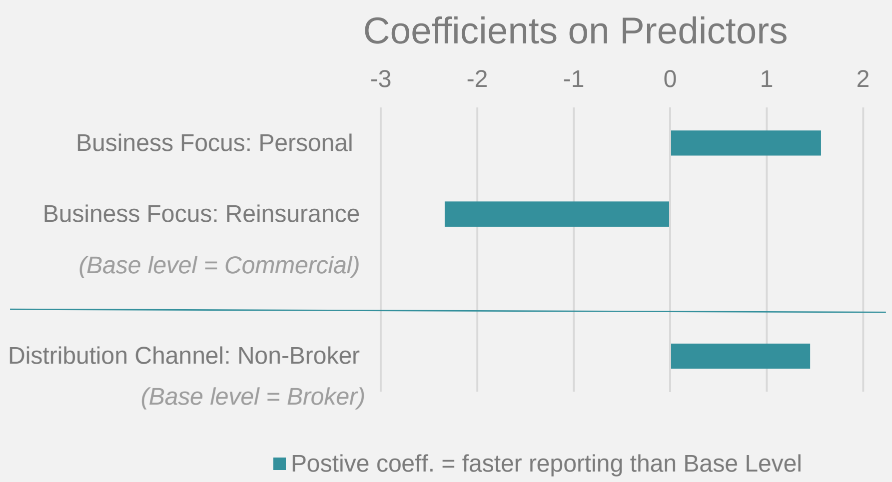
Visualization: are the explanatory variables logical?



Practical Considerations

Linear regression: are the explanatory variables logical?





Conclusion
Key Takeaways

- Clustering techniques help us obtain a better understanding of the loss development:
  - Explore the structure of data
  - Go beyond “just” practical grouping of data
  - Identify variables impacting the development

- Each method has strengths and weaknesses
  - Look for robustness between methods

Selected References

Munich RE

1. D. Clark (2017) **Estimation of Inverse Power Parameters via GLM**, *Actuarial Review, May-June 2017* , https://ar.casact.org/estimation-of-inverse-power-parameters-via-glm/

2. T. Hastie, R. Tibshirani, J. Friedman (2009) **The Elements of Statistical Learning - Data Mining, Inference, and Prediction**, Springer https://web.stanford.edu/~hastie/Papers/ESLII.pdf

3. C. Hennig (2015) **Clustering strategy and method selection,** In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.). *Handbook of Cluster Analysis*. Chapman and Hall/CRC, http://www.homepages.ucl.ac.uk/~ucakche/

4. C. Hennig, M.Meila, F. Murtagh, R.Rocci (2017) **Handbook of Cluster Analysis,** CRC Press

5. P. Tan, M. Steinbach, V. Kumar (2005) **Cluster Analysis: Basic Concepts and Algorithms,** In P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining,* Pearson Addison Wesley, http://www-users.cs.umn.edu/~kumar/dmbook/index.php

6. J. Shlens (2003) **A Tutorial on Principal Component Analysis: Derivation, Discussion and Singular Value Decomposition,** arXiv preprint arXiv:1404.1100, 2014, https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf

7. M. Steinbach, L. Ertoz, V. Kumar, **"The Challenges of Clustering High Dimensional Data"**, https://www-users.cs.umn.edu/~kumar001/papers/high_dim_clustering_19.pdf

8. J. VanderPlas, "**Python Data Science Handbook",** O'Reilly Media, http://shop.oreilly.com/product/0636920034919.do

9. CAS Schedule P data for Loss Reserving: http://www.casact.org/research/index.cfm?fa=loss_reserves_data

Thank you!

Munich RE

Appendix I: Soft Clusters and Mixed Models

Munich RE

- Soft (a.k.a. fuzzy) clustering allows each data point to belong to more than one cluster
- Membership grades are assigned to each data point
- In R, use *fanny(data, k=2,…)* from package "cluster" for fuzzy clustering
- Gaussian Mixed Models can also produce soft clusters

LOB	Fuzzy 1 (MedMal)	Fuzzy 2 (PPAL)	Fuzzy 3 (WC)
MedMal	45%	27%	28%
MedMal	54%	22%	24%
MedMal	66%	17%	18%
MedMal	46%	26%	28%
MedMal	65%	17%	18%
MedMal	66%	17%	18%
PPAL	6%	57%	38%
PPAL	12%	51%	37%
PPAL	16%	44%	40%
PPAL	8%	55%	37%
PPAL	5%	45%	49%
PPAL	6%	49%	44%
WC	5%	51%	44%
WC	5%	41%	54%
WC	9%	36%	56%
WC	5%	34%	61%
WC	5%	37%	58%
WC	13%	36%	51%

Appendix II: R Packages

Munich RE

Important R packages:

- Package "stats" *(kmeans, prcomp,…)* - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- Package "cluster" *(pam, fanny,…)* - https://cran.r-project.org/web/packages/cluster/cluster.pdf
- Package "factoextra" (get_eigenvalue, fviz_cluster,…) - https://cran.r-project.org/web/packages/factoextra/factoextra.pdf
- Package "ggplot2" - https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
- Package "mclust"(*mclust*) - https://cran.r-project.org/web/packages/mclust/mclust.pdf
- Package "Rmixmod" *(mixmodCluster)* - https://cran.r-project.org/web/packages/Rmixmod/Rmixmod.pdf

Comparison of packages for mixed model:

Package	Version	Clustering	Classification	Density estimation	Non-Gaussian components
mclust	5.2	✓	✓	✓	✗
Rmixmod	2.0.3	✓	✓	✗	✓
mixture	1.4	✓	✓	✗	✗
EMCluster	0.2–5	✓	✓	✗	✗
mixtools	1.0.4	✓	✗	✓	✓
bgmm	1.7	✓	✓	✗	✗
flexmix	2.3–13	✓	✗	✗	✓