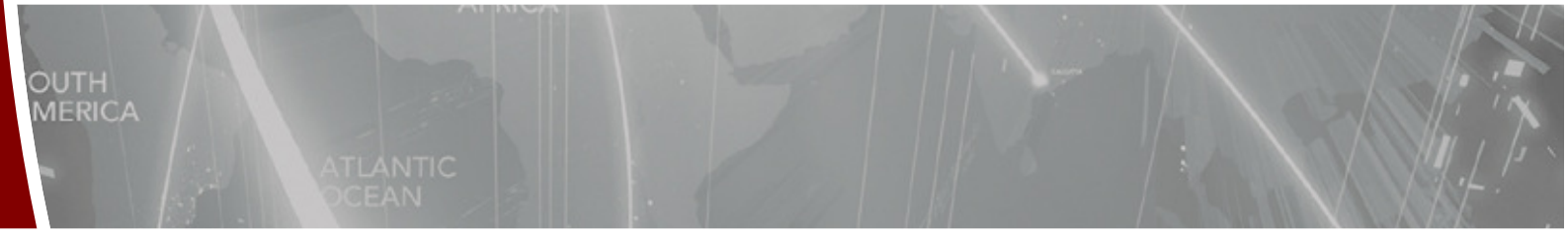
Techniques For Dimension Reduction

David Otto, FCAS, MAAA

EMB America LLC



EMB

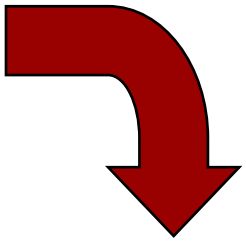
What Is Dimension Reduction

- Definition
  - Reducing the dimensionality of a data set by extracting a number of underlying factors, dimensions, clusters, etc., that can account for the variability in the data set
- Given a table of data:
  - Columns represent both the dimensions and facts of the data
  - Rows represent the observation
- Dimension reduction focuses on reducing both the number of columns (associations among variables) and the number of rows (associations among observations)

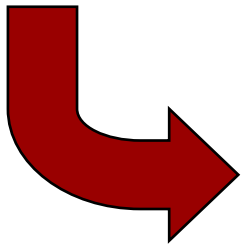
Dimensionality Revisited

- Understanding the design matrix

Obs	Age	Gender	Response
1	Youthful	Male	
2	Adult	Male	
3	**Mature**	**Male**	
4	Senior	Male	
5	Youthful	Female	
6	Adult	Female	
7	Mature	Female	
8	Senior	Female	



Data as understood in traditional tables, marts, and warehouses



	Age				Gender		
Obs	Youthful	Adult	Mature	Senior	Male	Female	Response
1	1	0	0	0	1	0	
2	0	1	0	0	1	0	
3	0	0	1	0	1	0	
4	0	0	0	1	1	0	
5	1	0	0	0	0	1	
6	0	1	0	0	0	1	
7	0	0	1	0	0	1	
8	0	0	0	1	0	1	

Data as translated in a statistical design matrix

EMB

Dimensionality Revisited

- Base class selection

	Age				Gender		
Obs	Youthful	Adult	Mature	Senior	Male	Female	Response
1	1	0	0	0	1	0	
2	0	1	0	0	1	0	
3	0	0	1	0	1	0	
4	0	0	0	1	1	0	
5	1	0	0	0	0	1	
6	0	1	0	0	0	1	
7	0	0	1	0	0	1	
8	0	0	0	1	0	1	

Data as translated in a statistical design matrix

	Base	Age			Gender	
Obs		Youthful	Adult	Senior	Female	Response
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	
4	1	0	0	1	0	
5	1	1	0	0	1	
6	1	0	1	0	1	
7	1	0	0	0	1	
8	1	0	0	1	1	

Data as translated in a statistical design matrix incorporating a base class

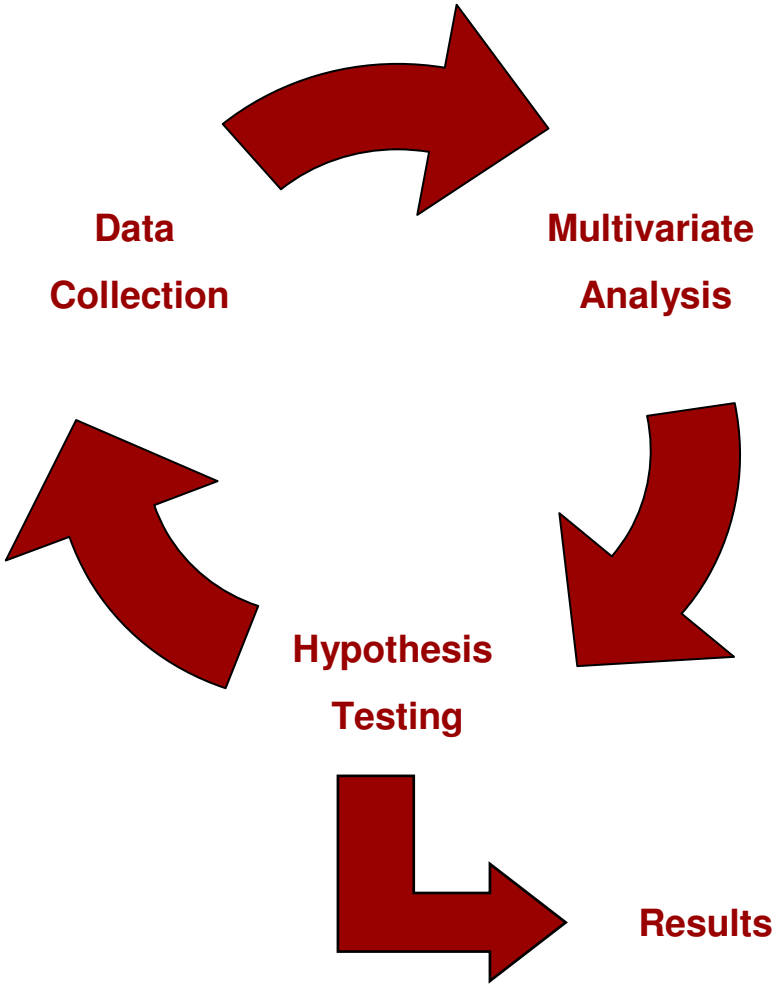
EMB

Dimension Reduction

- Dimension reduction in the modeling process:

Pre Modeling

- Principal Components
- Clustering



During Modeling

- Stepwise Regression
- Eliminating
- Grouping
- Curve Fitting
- Scoring
- Proxy Categorization Analysis

EMB

Stepwise Regression

Forward Selection

- Build a model with no factors and add based on prespecified criteria regarding improvement in model fit:

Model	Variables	Deviance	Degrees of Freedom	Chi Squared Compare to Base
Base	Mean	12,380.23	18,596	
1	Mean + Gender	12,377.02	18,594	20.1%
2	Mean + Policyholder Age	12,214.88	18,570	0.0%
3	Mean + Rating Area	12,365.50	18,581	47.1%
4	Mean + Vehicle Age	9,997.75	18,576	0.3%
	.			
	.			
17	Mean + MTA Indicator	12,370.30	18,595	0.2%
18	Mean + Time	12,371.45	18,594	0.1%

- Add the factor that performed the best on the Chi Square test (Policyholder Age)
- Iterate process with the new base model until no further factors indicated for selection

Stepwise Regression

Backward Elimination

- Build a model with all variables and delete based on prespecified criteria regarding improvement in model fit:

Model	Variables	Deviance	Degrees of Freedom	Chi Squared Compare to Base
Base	All	8,906.44	18,469	
1	All excl Gender	8,907.09	18,471	65.2%
2	All excl Policyholder Age	8,959.74	18,495	0.1%
3	All excl Rating Area	8,951.61	18,484	0.0%
4	All excl Vehicle Age	10,824.07	18,489	0.0%
	.			
	.			
17	All excl MTA Indicator	8,906.45	18,470	92.2%
18	All excl Time	8,982.06	18,471	0.0%

- Remove factor that performed the worst on the Chi Square test (MTA Indicator)
- Iterate process with the new base model until no further factors indicated for elimination

EMB

Stepwise Regression

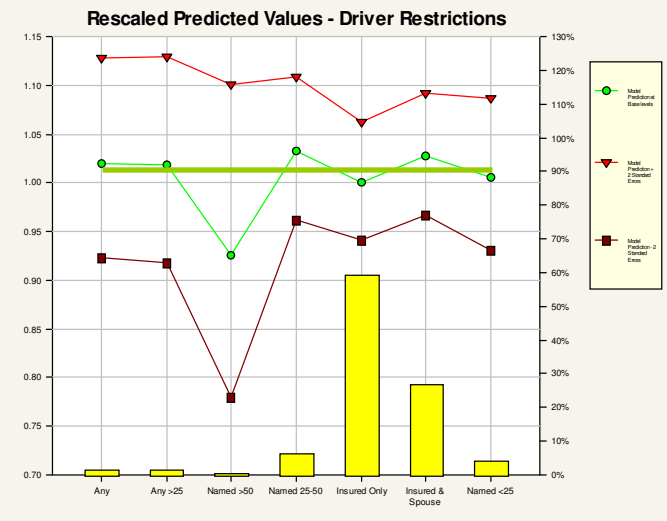
- Drawbacks:
  - Tendency to overfit the data
  - Problems in the presence of collinearity
  - Short cuts the exploratory process through which the researcher gains an intuitive feel for the data
- Advantages
  - Forward selection is a good way to develop an initial model
  - Automated process
  - Decisions made are within the multivariate framework
  - Indication of relative variable importance

Elimination

- Excluding factors entirely is the easiest and most straightforward way to simplify a model
- Things to look for:
  - Parameter estimates
    - All parameter estimates are small
    - All parameter estimates are within two standard errors of zero (i.e., the standard error percentages are all > 50%)
    - Sensible Patterns
  - Consistency over time
  - Hypothesis testing
    - Chi Square tests

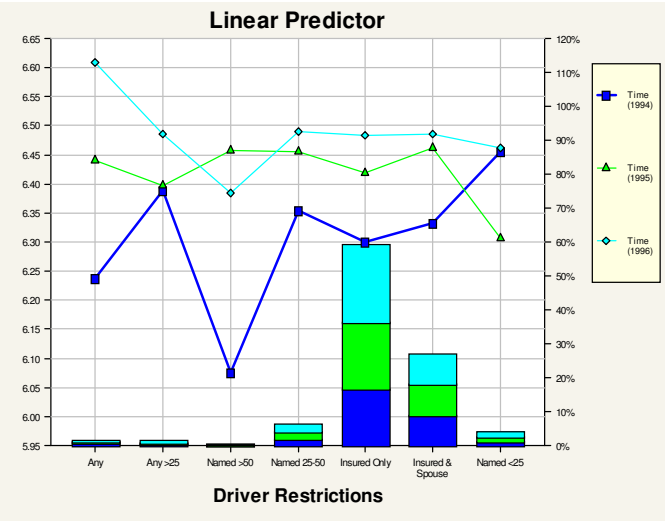
Elimination Example

Parameter estimates



High standard errors indicate flat relativities

Consistency over time



No consistent pattern over time

Hypothesis testing

Model	With	Without
Deviance	8,906.4414	8,909.6226
Degrees of Freedom	18,469	18,475
Scale Parameter	0.4822	0.4823
Chi Square Test		78.6%

Chi square test indicates models are statistically similar

EMB

Elimination Example

- Elimination reduction effect on the design matrix

		Age			Gender	
Obs	Base	Youthful	Adult	Senior	Female	Response
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	
4	1	0	0	1	0	
5	1	1	0	0	1	
6	1	0	1	0	1	
7	1	0	0	0	1	
8	1	0	0	1	1	

Data as understood in a statistical design matrix incorporating a base class

		Age			
Obs	Base	Youthful	Adult	Senior	Response
1	1	1	0	0	
2	1	0	1	0	
3	1	0	0	0	
4	1	0	0	1	
5	1	1	0	0	
6	1	0	1	0	
7	1	0	0	0	
8	1	0	0	1	

Elimination occurs by removing the gender dimensions

EMB

Grouping

- While a factor might be significant, it may be possible to band certain levels within a factor to create a more parsimonious model
- Things to look for:
  - Parameter estimates
    - Parameter estimates that are not significantly different from each other
    - Levels where there is low exposure
    - Sensible Patterns
  - Consistency over time
  - Models with and without the factor are not significantly different
    - Chi Square tests

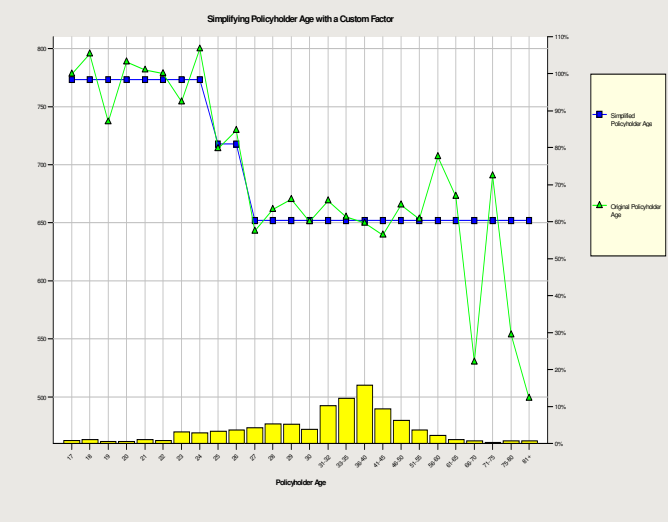
Grouping Example

Similarity in parameter estimates

	Policyholder Age (lt 17)	Policyholder Age (17)	Policyholder Age (18)	Policyholder Age (19)	Policyholder Age (20)	Policyholder Age (21)	Policyholder Age (22)	Policyholder Age (23)	Policyholder Age (24)	Policyholder Age (25)	Policyholder Age (26)	Policyholder Age (27)
Policyholder Age (lt 17)												
Policyholder Age (17)	92.3											
Policyholder Age (18)	87.3	306.0										
Policyholder Age (19)	110.0	132.4	91.0									
Policyholder Age (20)	94.1	1,414.6	277.0	154.2								
Policyholder Age (21)	97.7	333.2	153.0	196.1	468.5							
Policyholder Age (22)	97.7	357.7	162.6	200.2	512.2	10,113.7						
Policyholder Age (23)	104.4	130.8	79.6	498.2	158.5	205.1	213.3					
Policyholder Age (24)	90.2	912.9	378.7	101.2	530.3	188.0	204.9	68.4				
Policyholder Age (25)	108.2	104.4	67.8	4,227.8	123.3	141.4	148.0	307.4	56.6			
Policyholder Age (26)	101.6	161.6	90.9	293.4	203.2	322.2	330.4	388.3	82.2	161.7		
Policyholder Age (27)	147.1	41.4	31.8	77.9	46.1	41.5	44.2	35.7	23.5	38.5	29.4	
Policyholder Age (28)	134.7	48.0	35.9	103.7	54.0	49.4	52.7	44.1	26.4	49.4	35.3	132.5
Policyholder Age (29)	129.7	52.4	38.7	123.8	59.1	55.0	58.7	51.2	28.9	59.2	40.6	91.8
Policyholder Age (30)	147.2	41.6	32.0	78.2	46.3	41.8	44.6	36.6	24.1	39.8	30.7	38,134.8
Policyholder Age (31-32)	132.0	46.8	36.0	110.9	55.2	50.3	54.0	43.8	25.6	49.8	34.7	97.8
Policyholder Age (33-35)	142.4	41.8	31.5	82.5	46.9	41.7	44.8	34.2	22.0	37.1	27.6	345.5
Policyholder Age (36-40)	147.6	39.1	29.6	73.9	43.9	38.6	41.5	30.7	20.5	33.0	25.1	2,196.5
Policyholder Age (41-45)	156.5	36.5	28.1	64.9	40.7	35.7	38.3	28.8	19.9	30.5	24.0	192.1
Policyholder Age (46-50)	135.3	47.6	35.7	102.1	53.6	49.2	52.5	44.2	26.6	49.9	36.1	145.0
Policyholder Age (51-55)	147.1	42.0	32.5	79.0	46.8	42.6	45.2	37.9	24.9	41.6	32.3	43,108.5
Policyholder Age (56-60)	114.0	85.6	59.8	481.2	98.7	105.4	110.3	150.3	52.2	254.0	106.9	55.1
Policyholder Age (61-65)	138.1	55.1	43.3	111.9	60.8	59.7	62.1	63.5	38.6	73.0	55.0	288.4
Policyholder Age (66-70)	652.1	23.6	20.7	30.6	25.1	23.2	24.0	21.6	18.3	22.3	20.4	31.5
Policyholder Age (71-75)	127.5	95.4	75.5	243.1	103.9	114.7	116.3	153.2	78.0	194.2	129.6	191.1
Policyholder Age (75-80)	431.4	25.4	22.1	33.8	27.1	25.2	26.0	23.5	19.6	24.4	22.2	36.9
Policyholder Age (81+)	1,822.1	19.7	17.5	24.6	20.8	19.3	19.9	17.8	15.6	18.3	17.0	23.8

Standard error of parameter differences

Sensible patterns



Grouping levels that have sensible patterns

Hypothesis testing

Model	Ungrouped	Grouped
Deviance	8,906.4414	8,934.1620
Degrees of Freedom	18,469	18,493
Scale Parameter	0.4822	0.4823
Chi Square Test		27.2%

Chi square test indicates models are statistically similar

EMB

Grouping Example

- Grouping reduction effect on the design matrix

Obs	Base	Age			Gender	Response
		Youthful	Adult	Senior	Female	
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	
4	1	0	0	1	0	
5	1	1	0	0	1	
6	1	0	1	0	1	
7	1	0	0	0	1	
8	1	0	0	1	1	

Data as understood in a statistical design matrix incorporating a base class

Obs	Base	Age		Gender	Response
		Youthful/Adult	Senior	Female	
1	1	1	0	0	
2	1	1	0	0	
3	1	0	0	0	
4	1	0	1	0	
5	1	1	0	1	
6	1	1	0	1	
7	1	0	0	1	
8	1	0	1	1	

Grouping occurs by combining the youthful and adult dimensions

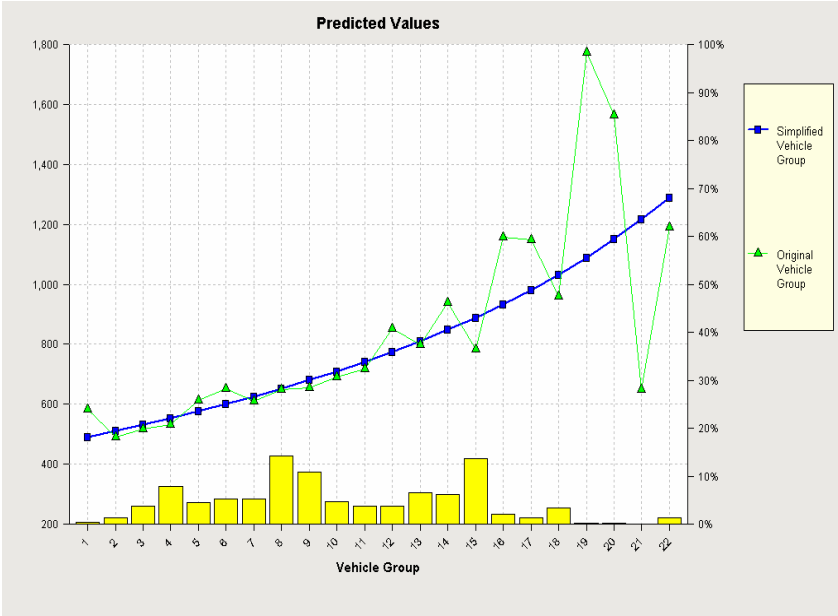
EMB

Curve Fitting

- While a factor might be significant, it may be desirable to smooth adjacent levels to create a more parsimonious model
- Things to look for:
  - Factors which have a natural x-axis that can be converted to a continuous scale
  - Factors with a sufficient number of levels to justify curve fitting
  - Factors with a definite trend or progression
  - Models with and without the factor are not significantly different
    - Chi Square tests

Fitting Curves

- Simplify trends in rating factors in order to remove random noise, by fitting an nth degree curve…



- Additional Curve Fitting Options
  - Degree of the polynomial
  - Multiple curves across the same variable
  - Splines

EMB

Curve Fitting Example

- Curve fitting reduction effect on the design matrix

Obs	Base	Age			Gender	Response
		Youthful	Adult	Senior	Female	
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	
4	1	0	0	1	0	
5	1	1	0	0	1	
6	1	0	1	0	1	
7	1	0	0	0	1	
8	1	0	0	1	1	

Data as understood in a statistical design matrix incorporating a base class

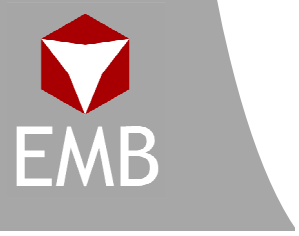
Obs	Base	Age		Gender	Response
		Age^1	Age^2	Female	
1	1	20.0	400.0	0	
2	1	35.0	1,225.0	0	
3	1	55.5	3,080.3	0	
4	1	82.5	6,806.3	0	
5	1	20.0	400.0	1	
6	1	35.0	1,225.0	1	
7	1	55.5	3,080.3	1	
8	1	82.5	6,806.3	1	

Curve fitting transforms the age columns into continuous vectors

EMB

Scoring

- Predictors can be combined into an overall score
  - Need multivariate estimators for each rating factor and some type of systematic way to assign a load for each point in the score (i.e. measure of sensitivity)
- The idea is to decompose rating variables in linear combinations of latent traits
  - Scores are the location of the original observations in the reduced factor space
- Once you have merged the rating variables into the score you need to come up with a risk metric for each score total
  - Done by evaluating in the multivariate framework
    - Take unused rating factors out of the pure premium model
    - Replace with underwriting score point values

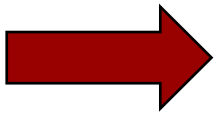


Scoring Example

- Example:

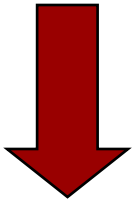
Loss Control			Company Size	
Yes	0		Small	0
No	7		Medium	9
			Large	6
Territiory				
A	0			
B	6		**Years Renewed**	
C	9		New	12
D	13		1	7
E	16		2	4
			3	0
AND OTHERS...				

Scoring rules are coming from the multivariate estimators



Score
<120
121-125
126-30
131+

Cumulate rules to a total score for each observation



Score	Score Factor
<120	0.90
121-125	1.00
126-30	1.05
131+	1.20

Derive score factors using standard multivariate modeling techniques

EMB

Scoring Example

- Scoring reduction effect on the design matrix

Obs	Base	Age			Gender	Response
		Youthful	Adult	Senior	Female	
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	
4	1	0	0	1	0	
5	1	1	0	0	1	
6	1	0	1	0	1	
7	1	0	0	0	1	
8	1	0	0	1	1	

Data as understood in a statistical design matrix incorporating a base class

Obs	Base	Score			Response
		Group1	Group2	Group3	
1	1	1	0	0	
2	1	0	1	0	
3	1	0	0	0	
4	1	0	0	1	
5	1	1	0	0	
6	1	0	1	0	
7	1	0	0	0	
8	1	0	0	1	

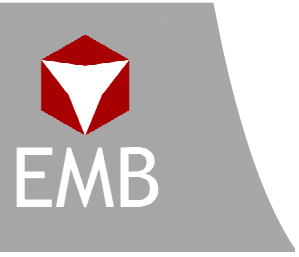
Scoring translates continuous and categorical concepts into a single concept

EMB

Proxy Categorization Analysis

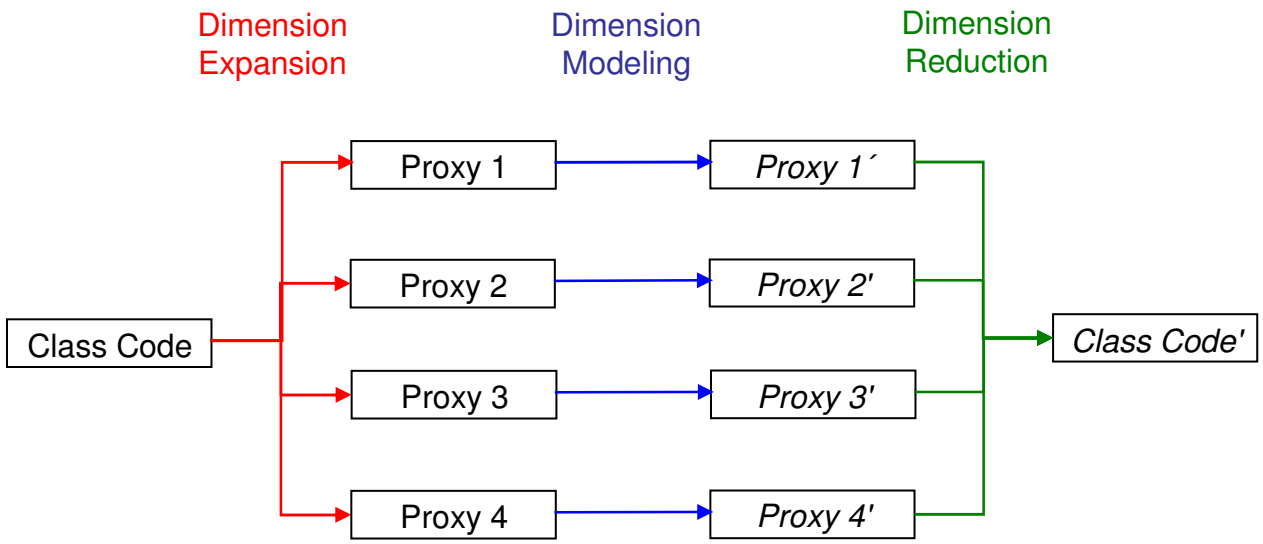
Used for handling high dimensional categorical rating variables:

- What is a high dimensional categorical rating variable?
  - Rating variable with a large number of levels.
  - Levels within the factor do not have a natural x-axis that can be converted to a continuous scale.
- Examples:
  - WC Class Code
  - Zip code
  - VINs



Proxy Categorization Analysis

- Issues with dimension reduction:
  - Single dimension with a large number of levels
  - Basic techniques do not perform well:
    - Grouping becomes difficult to evaluate with a large number of levels .
    - Cannot curve-fit because no natural x-axis
- Solution is a three step process:



EMB

Step 1: Dimension Expansion

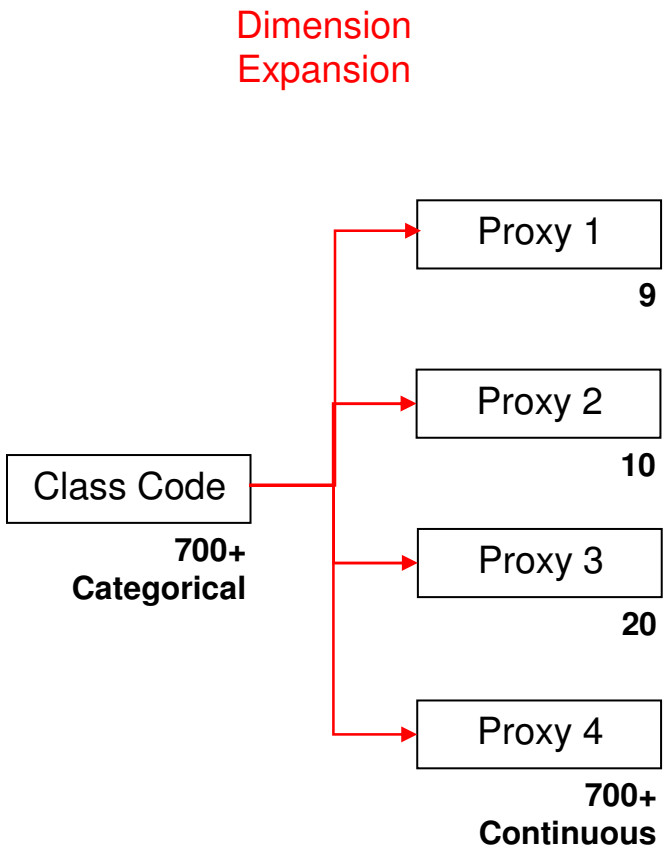
- Introduce multiple dimensions into the dataset using proxies 1 through n.
- Desired characteristics of proxies:
  - Each level of class code should have a direct relationship to a given level within the proxy rating factor.
  - Enables separation of the class code signal from class code noise in the response variable.
  - Performs well with basic dimension reduction techniques:
    - Fewer number of levels.
    - Factors which have a natural x-axis that can be converted to a continuous scale.

Step 1: Dimension Expansion

- Examples of proxies to use for class code:
  - Hazard Group (4,9 groups).
  - Best's Hazard Index (10 categories).
  - NAICS - North American Industry Classification System (20 sectors, numerous subsectors).
- Examples of proxies to use for zip code:
  - Population density
  - Median home value
  - Percent of population using public transportation

Step 1: Dimension Expansion

- Merge proxies into the dataset.



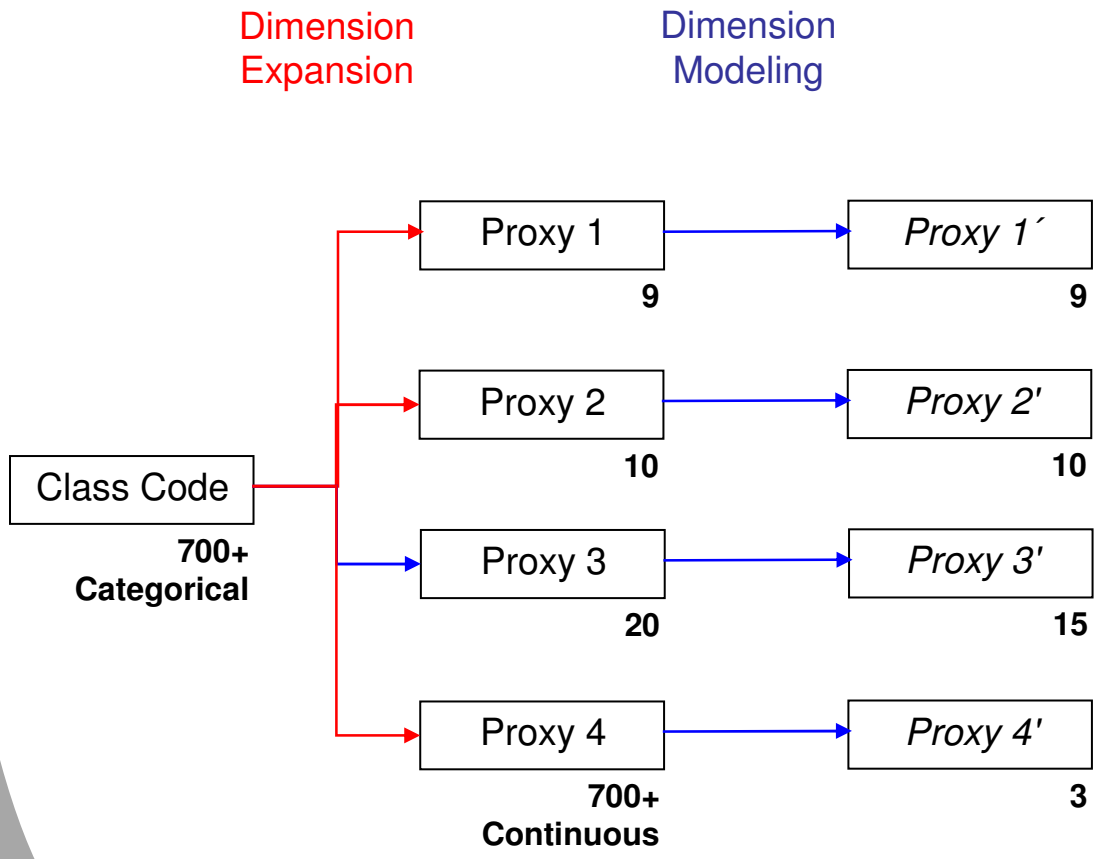
Step 2: Dimension Modeling

- Now we are ready to begin modeling at the component level:
  - Frequency - Severity
  - Coverage
  - Cause of Loss
- For each component: Incorporate proxies and other rating factors that exhibit predictive power into the model.
  - Perform basic dimension reduction techniques:
    - Elimination
    - Grouping
    - Curve fitting
- Do not use original variable as a rating factor at the component level.

EMB

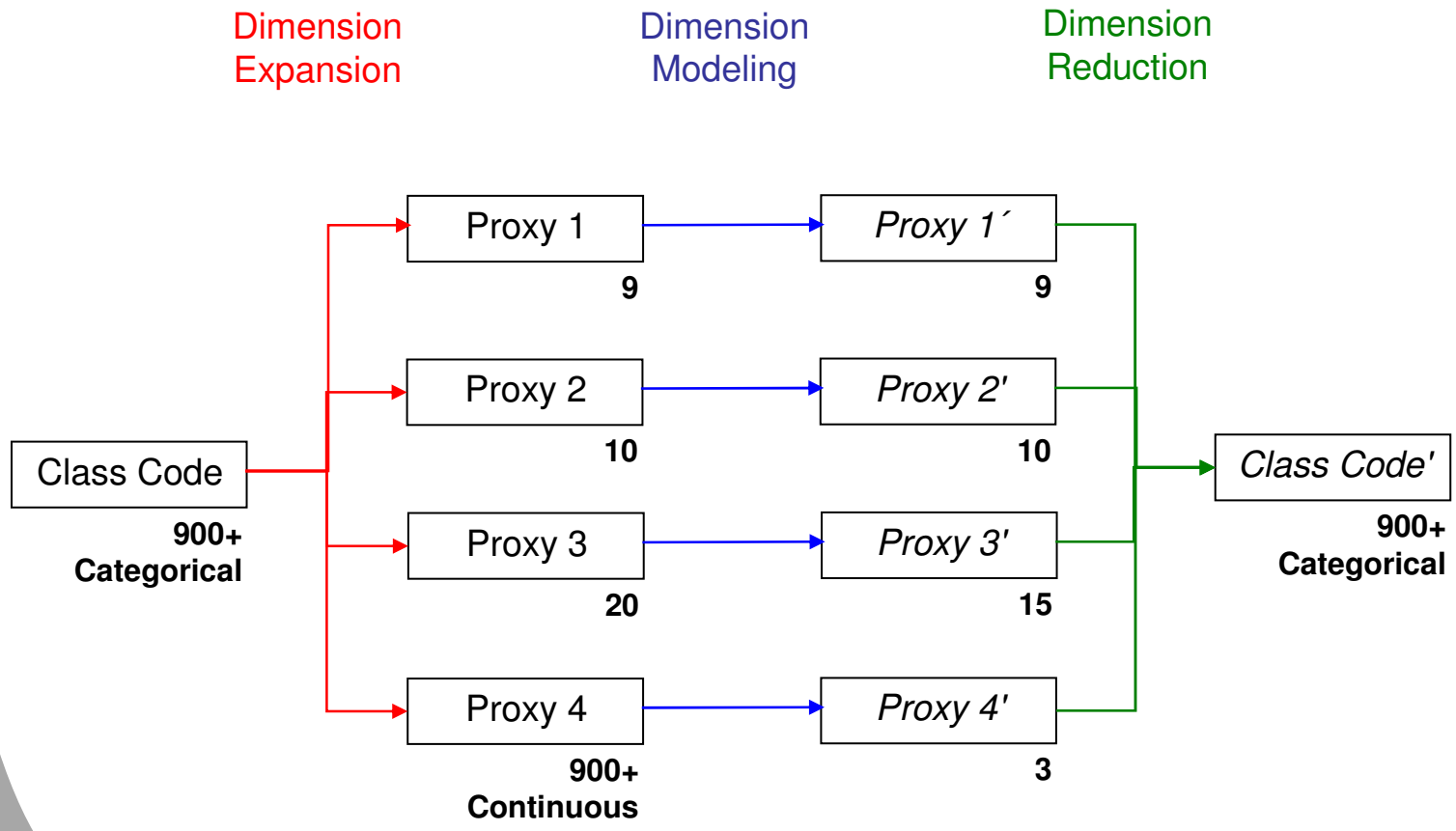
Step 2: Dimension Modeling

- Perform modeling at the component level using basic dimension reduction technique.



Step 3: Dimension Reduction

- Combine completed component models to produce a pure premium model.
- Remove the proxy rating factors and replace with original rating factor.



Dimension Reduction - Summary

- Modern data warehouses contain more data than ever before so modeling techniques need to be able to handle the added complexities
- Objective is to use techniques to identify which factors are predictive thus identifying the signal from the noise
- Data decisions should occur within a multivariate framework
  - Selecting the appropriate dimensions
  - Transforming dimensions
  - Grouping observations

EMB America LLC

Mission:

EMB America seeks to help our clients solve complex problems and identify opportunities by providing the appropriate blend of value-added consulting and state-of-the-art software. In so doing, EMB America strives to develop long-term relationships with clients and be the consulting firm of choice for the business community.

For information:

- Phone: 858.793.1425
- Email: info@embamerica.com
- Website: www.embamerica.com