Business Analytics

PLS Regression

Instructor: Hrant Davtyan
Course: Business Analytics, Fall, MSSM, 2018

Content

1. Curse of dimensionality
2. Partial Least Squares (PLS)
3. Advantages and applications of PLS
4. Example

Curse of dimensionality

Problems of high dimensional data

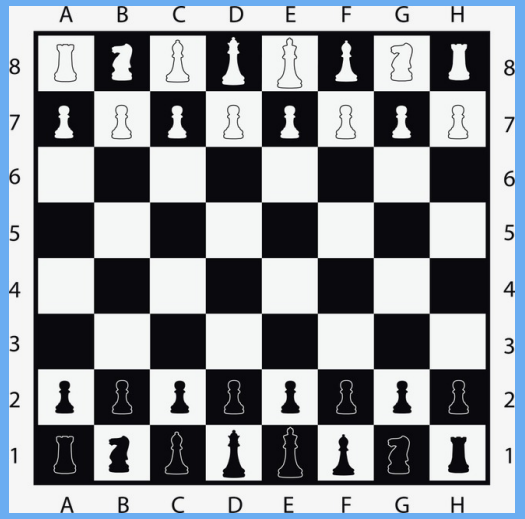
1. **High dimensional data** - data with large number of features, attributes or characteristics
2. **Dimensionality of a model** - the number of independent or input variables used by the model
3. **Problems:**
    - Many highly correlated input variables (in linear regression this will cause *multicollinearity*)
    - Number of variables may be bigger than number of observations
    - Traditional analysis methods (e.g. linear regression) may fail or become computationally intractable

Problems of high dimensional data

Curse of dimensionality- The difficulties posed by adding a variable increase exponentially with the inclusion of each variable.

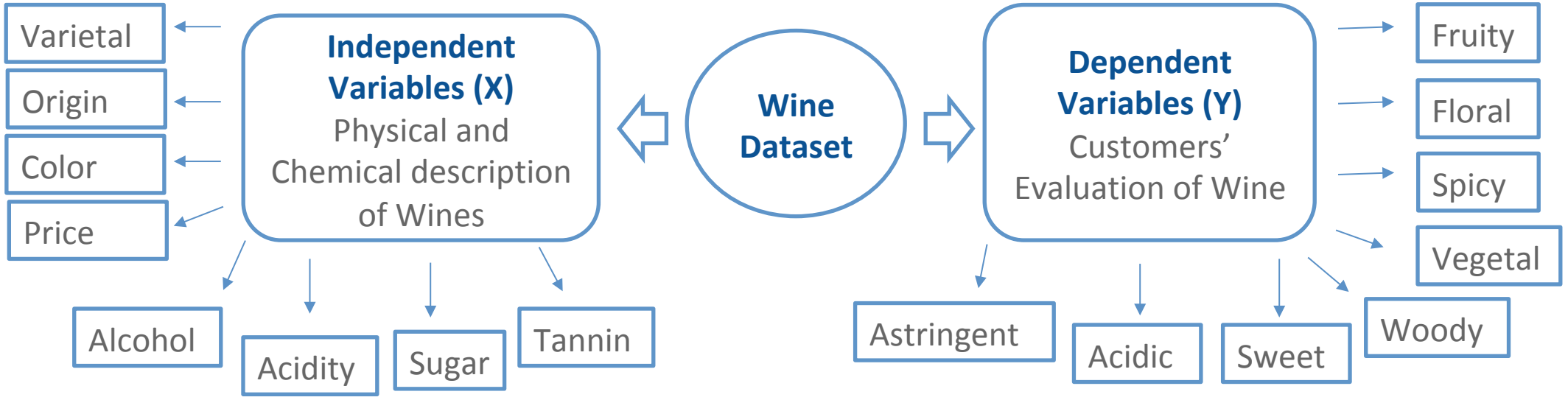
One way to think of this intuitively

If we expand the chessboard to a cube, we increase the dimensions by 50%-from 2 dimensions to 3 dimensions. However, the location options increase by 800%, to 512 (8x8x8).



Example

1. One more problem – there may be not only many dependent variables but also more than one independent variable
2. Example of data where dependent and independent variables are explained by many factors:



What to do?

Use models that will reduce the dimension of data!

Partial Least Squares (PLS)

What is PLS Regression?

1. The goal of partial least squares is to predict Y from X and to describe the common structure underlying the two variables
2. PLS **maximizes the covariance** between the target variables (Y) and the predictive variables (X).
3. Basically we want to do linear regression between two tables (1,2)

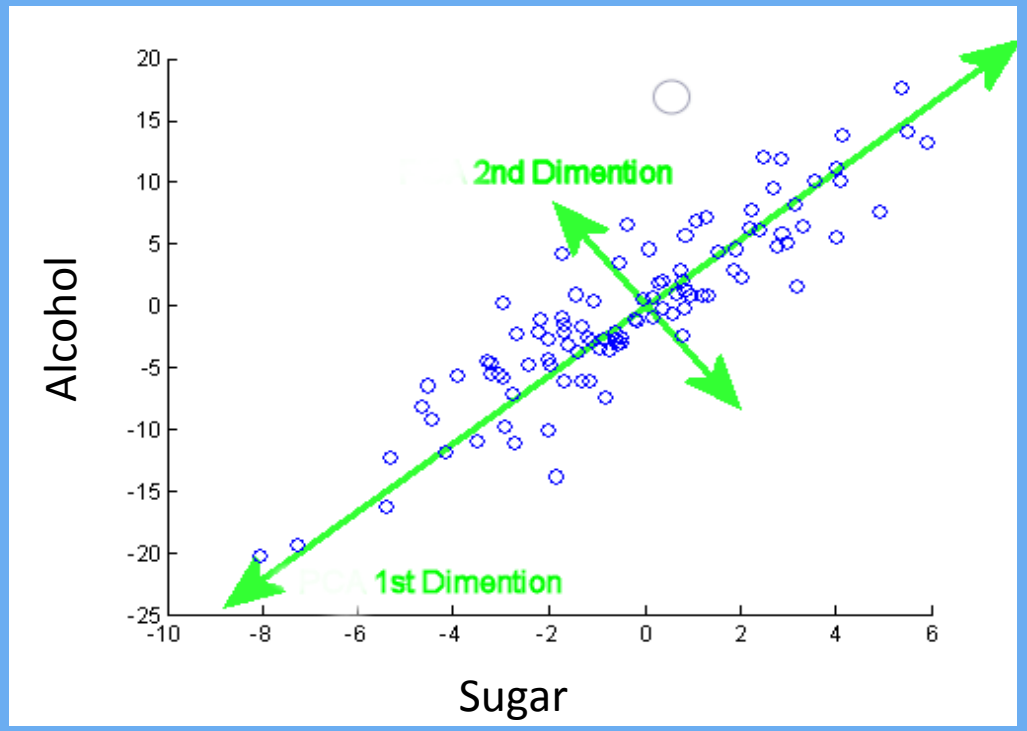
Table 1: X variables									Table 2: Y variables								
Wine Name	Varietal	Origin	Color	Price	Total acidity	Alcohol	Sugar	Tannin	Fruity	Floral	Vegetal	Spicy	Woody	Sweet	Astringent	Acidic	Hedonic
1	Merlot	Chile	Red	– 0.046	– 0.137	0.120	– 0.030	0.252	– 0.041	– 0.162	– 0.185	0.154	0.211	– 0.062	0.272	0.044	– 0.235
2	Cabernet	Chile	Red	– 0.185	– 0.165	0.140	– 0.066	0.335	– 0.175	– 0.052	– 0.030	0.041	0.101	– 0.212	0.385	– 0.115	– 0.235
3	Shiraz	Chile	Red	– 0.116	– 0.162	0.219	– 0.088	0.176	0.093	– 0.271	– 0.030	0.380	0.211	– 0.062	0.160	– 0.275	– 0.235
4	Pinot	Chile	Red	0.093	– 0.278	0.061	– 0.003	0.098	– 0.175	– 0.052	– 0.030	– 0.072	0.101	– 0.361	0.047	0.044	– 0.105
5	Chardonnay	Chile	White	0.023	– 0.283	0.022	– 0.045	– 0.124	– 0.175	0.058	– 0.185	0.041	0.101	– 0.212	– 0.178	0.044	0.025
6	Sauvignon	Chile	White	– 0.116	0.049	0.022	0.015	– 0.118	0.093	0.168	0.590	– 0.185	– 0.229	0.087	– 0.178	0.204	0.155
7	Riesling	Chile	White	– 0.081	0.210	– 0.175	– 0.093	– 0.127	– 0.041	0.387	– 0.030	– 0.072	– 0.119	– 0.062	– 0.178	0.364	0.220

PLS Regression

1. PLS Regression finds **latent variables (non-observable)** that explain X and are also the best for explaining Y.
2. Or in other words, PLS is a regression method that allows for the identification of underlying factors, which are a linear combination of the explanatory variables (X) (also known as latent variables) which best model the response or Y variables
3. PLS is preferred as **a predictive technique** rather than interpretive
4. Two types of PLS regression
    - **PLS1**- data includes only one dependent variable (Y)
    - **PLS2**- there are more than one dependent variables (Y)

Example of dimension reduction

An example of how **latent variables** can be extracted from 2 variables:



- From the graph, we see that Alcohol and Sugar are highly correlated.
- 1st Dimension explains the highest part of the variance of the data.
- 2nd Dimension explains the remaining part of the variance.
- So we got 2 new variables which are not correlated anymore.

Process

1. 1 set of latent variables is extracted for set of independent variables (X)
2. 1 set of latent variables is extracted **simultaneously** of set of dependent variables (Y)
3. The x-scores of independent latent variables are used to predict y-scores or dependent latent variables
4. Predicted y-scores are used to predict **observable** response variables
5. The x and y scores are selected by partial least squares so that the relationship of successive pairs of x and y scores is as strong as possible **(maximum covariance)**

Steps to run PLS

1. Standardize the data
2. Define the number of latent variables (components) we want to keep in our PLS regression.
3. Fit a set of components to X
4. Fit a set of components to Y
5. Reconcile the two sets of components so as to maximize covariance of X and Y

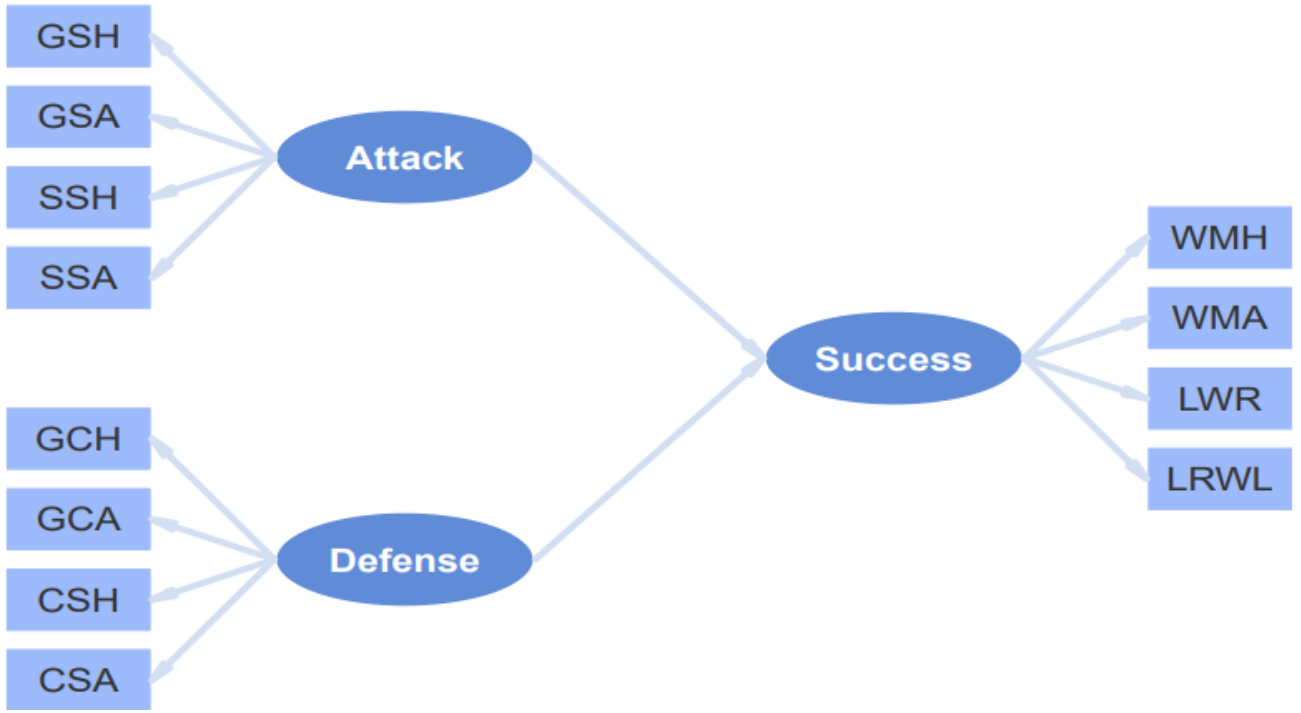
Advantages and applications of PLS

Advantages

1. PLS is a **soft modeling** technique (no hard assumptions as e.g. LINE)
2. It can model both **multiple dependent and multiple independent** variables
3. Can handle multicollinearity
4. Stronger predictions as when creating scores of X takes into account the correlation with Y to be as higher as possible
5. Can be applied to small samples
6. It can handle range of variables: nominal, ordinal, continuous
7. Partial least squares has the added benefit of providing a graphical representation of the relationships between the variables – **path diagrams**

Path Diagrams Representations

1. **Observable variables** are represented in a rectangular form
2. **Latent variables** are represented in an oval form
3. **Relationships between variables** are represented with straight arrows



Applications

1. Widely used in chemometric
2. Marketing (e.g. to find out factors that influence customer satisfaction which may be measured by many variables)
3. Industrial applications (e.g. to improve product quality through excellence in operation)
4. Economics (e.g. to model growth rate based on many economic and non-economic variables)

Example

Using Partial Least Squares Regression to Model Vehicle Sales

1. The first factor explains 20.9% of the variance in the predictors and 40.3% of the variance in the dependent variable.
2. The second factor explains 55.0% of the variance in the predictors and 2.9% of the variance in the dependent.
3. Together, the first three factors explain 81.3% of the variance in the predictors and 47.4% of the variance in the dependent.

Latent Factors	Statistics				
	X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
1	.209	.209	.403	.403	.395
2	.550	.760	.029	.431	.420
3	.053	.813	.043	.474	.460
4	.089	.902	.009	.483	.465
5	.026	.927	.002	.485	.464

Using Partial Least Squares Regression to Model Vehicle Sales

1. The parameters table shows the estimated regression coefficients for each independent variable for predicting the dependent variable.
2. Instead of the typical tests (e.g. t-test) of model effects, look to the variable importance in each latent variable.

Parameters

Independent Variables	Dependent Variables
	Insales
(Constant)	-2.107
[type=Automobile]	-.944
price	-.044
engine_s	.356
horsepow	-.002
wheelbas	.041
width	-.030
length	.018
curb_wgt	.068
fuel_cap	-.056
mpg	.079

	Latent Factors				
Variables	1	2	3	4	5
[type=Automobile]	1.037	1.053	1.011	1.057	1.055
price	2.088	2.028	1.965	1.949	1.946
engine_s	.512	.618	.900	.934	.932
horsepow	1.472	1.424	1.386	1.375	1.372
wheelbas	1.104	1.145	1.093	1.085	1.084
width	.139	.298	.301	.360	.370
length	.815	.882	.856	.861	.859
curb_wgt	.155	.293	.295	.300	.319
fuel_cap	.059	.175	.472	.475	.474
mpg	.457	.460	.546	.544	.561

Cumulative Variable Importance

Thank you