



DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale

Department of Computer Science and
Engineering

DATA ANALYTICS

Unit 5: Sparse data processing, Sparse PCA

Swati Pratap Jagdale

Department of Computer Science and Engineering

- **Sparse principal component analysis (sparse PCA)** is a specialised technique used in statistical analysis and, in particular, in the analysis of multivariate data sets.
- It extends the classic method of principal component analysis (PCA) for the reduction of dimensionality of data by introducing sparsity structures to the input variables.
- A particular disadvantage of ordinary PCA is that the principal components are usually linear combinations of all input variables.
- Sparse PCA overcomes this disadvantage by finding linear combinations that contain just a few input variables.
- Contemporary datasets often have the number of input variables comparable with or even much larger than the number of samples.

- **Mathematical Formulation**
- Consider a data matrix X , where each of the p columns represent an input variable and each of the n rows represents an independent sample from data population.
- One assumes each column of X has mean zero, otherwise one can subtract column-wise mean from each element of X .
- Let $\Sigma = \frac{1}{n-1} X^T X$ be the empirical covariance matrix of X , which has

dimensions $p \times p$. Given integer k , $1 \leq k \leq p$, the sparse PCA problem can be formulated as maximizing the variance along a direction represented by vector v , while constraining its cardinality.

$$\begin{aligned} & \max \quad v^T \Sigma v \\ & \text{subject to} \quad \|v\|_2 = 1 \\ & \quad \quad \quad \|v\|_0 \leq k. \end{aligned}$$

- First constraint specifies that v is a unit vector, In the second constraint, $||v||_0$ represents the L0 norm of v , which is defined as the number of its non-zero components.
- So the second constraint specifies that the number of non-zero components in v is less than or equal to k , which is typically an integer that is much smaller than dimension p .
- The optimal value of above Eq is known as the k -sparse largest eigenvalue.
- If one takes $k=p$, the problem reduces to the ordinary PCA, and the optimal value becomes the largest eigenvalue of covariance matrix Σ .
- After finding the optimal solution v , one deflates Σ to obtain a new matrix.

$$\Sigma_1 = \Sigma - (v^T \Sigma v) v v^T,$$

- Iterate this process to obtain further principal components.
- However, unlike PCA, sparse PCA cannot guarantee that different principal components are orthogonal. In order to achieve orthogonality, additional constraints must be enforced.
- The following equivalent definition is in matrix form.
- Let, V be a $p \times p$ symmetric matrix, one can rewrite the sparse PCA problem as:

$$\begin{aligned} & \max \quad \text{Tr}(\Sigma V) \\ & \text{subject to} \quad \text{Tr}(V) = 1 \\ & \quad \quad \quad \|V\|_0 \leq k^2 \\ & \quad \quad \quad \text{Rank}(V) = 1, V \succeq 0. \end{aligned}$$

- Tr is the matrix trace, and $\|V\|_0$ represents the non-zero elements in matrix V . The last line specifies that V has matrix rank one and is positive semidefinite. The last line means that one has $V = \mathbf{v}\mathbf{v}^T$

- **Applications**
- **Financial Data Analysis**
- Suppose ordinary PCA is applied to a dataset where each input variable represents a different asset, it may generate principal components that are weighted combination of all the assets
- In contrast, sparse PCA would produce principal components that are weighted combination of only a few input assets, so one can easily interpret its meaning.
- Furthermore, if one uses a trading strategy based on these principal components, fewer assets imply less transaction costs.

- **Applications**
- **Biology**
- Consider a dataset where each input variable corresponds to a specific gene. Sparse PCA can produce a principal component that involves only a few genes, so researchers can focus on these specific genes for further analysis.
- **High-dimensional Hypothesis Testing**
- Contemporary datasets often have the number of input variables (p) comparable with or even much larger than the number of samples (n)
- It has been shown that if p/n does not converge to zero, the classical PCA is not consistent. But sparse PCA can retain consistency even if $p \gg n$

Sparse PCA & rare event modelling



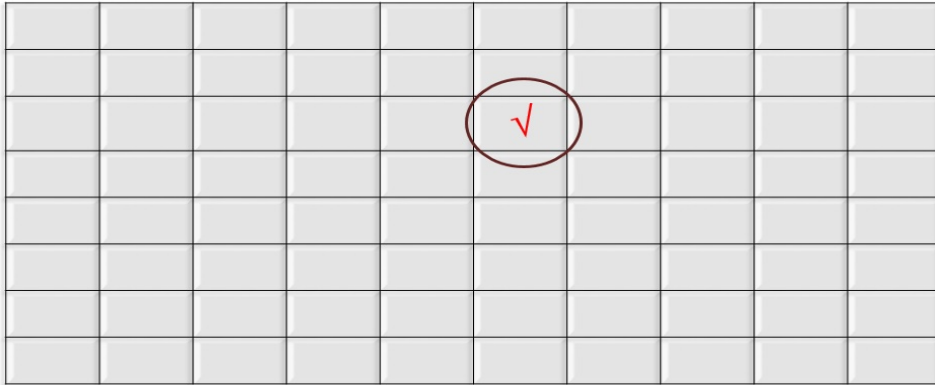
“Only 531 out of a population of 50,431 customer closed their saving account in a year, but the dollar value lost because of such closures was more than \$ 5 Million. “

The best way to arrest these attrition was by predicting the propensity of attrition for individual customer and then pitch retention offers to these identified customers. This was a typical case of modeling in a rare event population. This kind of problems are also very common in Health care analytics.

In such analysis, there are two challenges :

1. Accurate prediction is difficult because of small sample bias.
2. The accuracy of prediction need to be extremely high to make an implementable strategy. This is because high number of false positive, unnecessarily burdens the retention budgets.

step by step process to make a logistic regression model in a rare event population:



					✓				

How to model rare events?

The problem basically is that maximum likelihood estimation of the logistic model is well-known to suffer from small-sample bias. And the degree of bias is strongly dependent on the number of cases in the less frequent of the two categories.

Sparse PCA & rare event modelling



estimating the degree of bias in each of the following samples ??

A. 20 events in a sample size of 1000 (Response rate : 2%)

B. 180 events in a sample size of 10000 (Response rate : 1.8%)

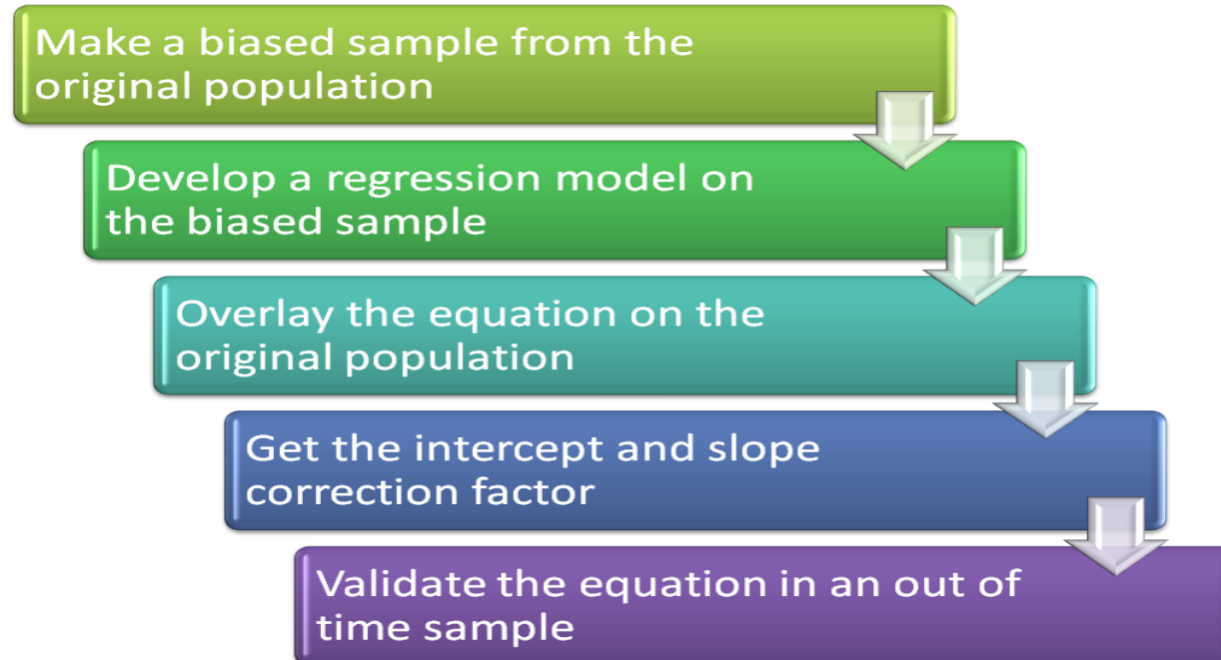
C. 990 events in a sample size of 1000 (Response rate : 99%)

$C > A > B$. C will suffer with the problem of small-sample bias most

Sparse PCA & rare event modelling

The solution in such problems is slightly longer than a normal logistic regression model. In such cases, we make a biased sample to increase the proportion of events.

Now, we run logistic regression on the sample created. Once we have the final Logit equation, we transform the equation to fit the entire population.



Sparse PCA & rare event modelling



consider the case in hand and walk through the step by step process. We have a population of 50,431 customers out of which 531 attrite in 12 months. We need to predict the probability of attrition, minimizing the false positives.

Step 1 :Select a biased sample

Total number of non attritors in the population is 49,900.

We plan to take a sample of 1000 customers.

As a thumb rule, we select 25% of the sample size as the responders.

Hence, we select 250 customers out of the 531 attriting customers. And rest 750 come from the 49,900 base.

This sample of 1000 customers is a biased sample we will consider for our analysis.

Step 2 : Develop the regression model

We now build a logistic regression model on the biased sample selected. We make sure that all the assumptions of the logistic regression are met and we get a reasonable lift because the lift tends to decrease after the transformations.

Step 3 : Overlay equation on the population

Using the equation found in step 2, get the number of attritors in each decile of the overall population. In the table below, -Log odds (Predicted) directly comes from the regression equation. Using this function, one can find the Predicted attrition for each decile. (Refer to the table- next slide)

Sparse PCA & rare event modelling

Decile	Population	Actual attrition	Predicted Attrition	-LOG Odds (Actual)	-LOG Odds (Predicted)
1	5043	100	2500	1.69	0.01
2	5043	90	2200	1.74	0.11
3	5043	74	1500	1.83	0.37
4	5043	52	1300	1.98	0.46
5	5043	47	1100	2.03	0.55
6	5043	42	950	2.08	0.63
7	5043	40	800	2.10	0.72
8	5043	35	650	2.16	0.83
9	5043	30	600	2.22	0.87
10	5044	21	400	2.38	1.06
Total	50431	531	12000		

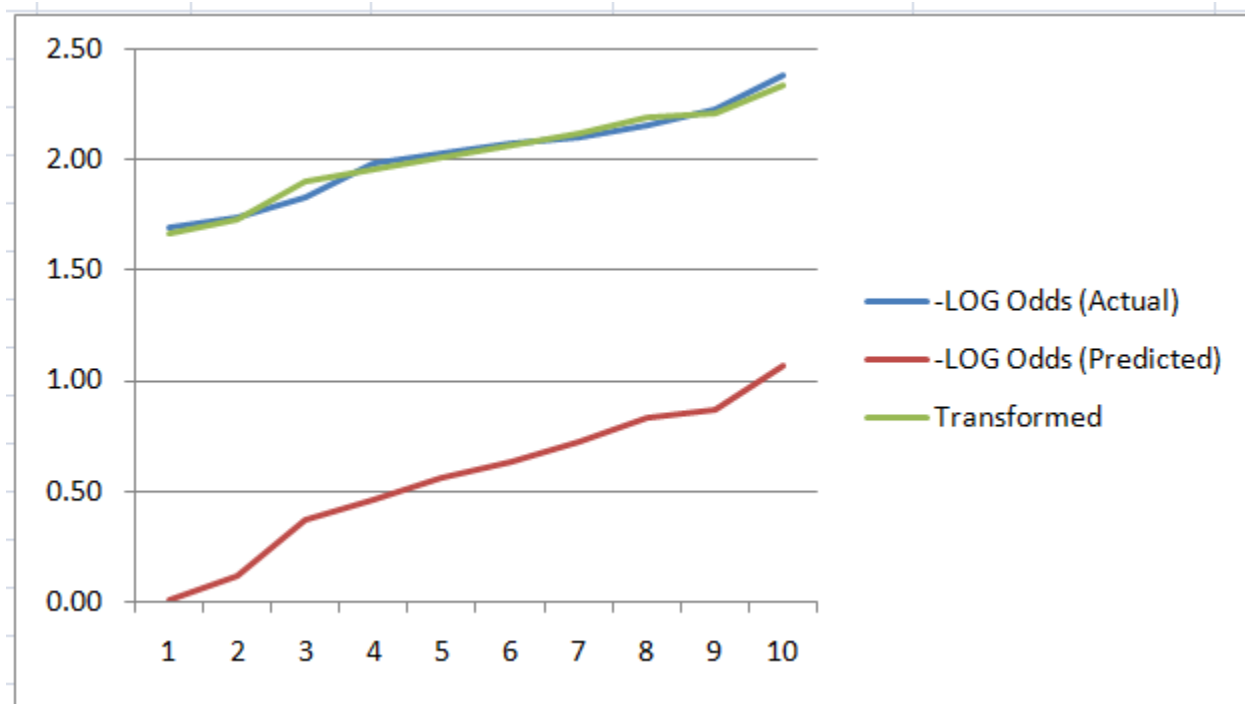
Step 4: Solve for intercept and slope transformation

Using the actual and the predicted decile value of the log odds, we find the slope and the intercept required to transform the equation of the sample to the equation of the population. This equation is given by,

$$\{-\text{Log odds (actual)}\} = \text{slope} * \{-\text{Log odds(predicted)}\} + \text{Intercept}$$

Find the slope and intercept using the 10 data-points, each corresponding to each decile. In this case slope is 0.63 and the intercept is 1.66 .

Decile	Population	Actual attrition	Predicted Attrition	-LOG Odds (Actual)	-LOG Odds (Predicted)	Transformed
1	5043	100	2500	1.69	0.01	1.67
2	5043	90	2200	1.74	0.11	1.73
3	5043	74	1500	1.83	0.37	1.90
4	5043	52	1300	1.98	0.46	1.95
5	5043	47	1100	2.03	0.55	2.01
6	5043	42	950	2.08	0.63	2.07
7	5043	40	800	2.10	0.72	2.12
8	5043	35	650	2.16	0.83	2.19
9	5043	30	600	2.22	0.87	2.22
10	5044	21	400	2.38	1.06	2.34
Total	50431	531	12000			



As seen from the above figure, the actual and the transformed logit curve for each decile is much closer compared to the predicted curve.

Step 5: Validate the equation on out of time sample

Once we reach a final equation of the logit function, we now validate the same on out of time samples. For the case in hand, we take a different cohort and compile the lift chart. If the model holds on out of time as well, we are good to go.

References

<https://www.analyticsvidhya.com/blog/2014/01/logistic-regression-rare-event/>

https://en.wikipedia.org/wiki/Sparse_PCA#Financial_Data_Analysis



THANK YOU

Swati Pratap Jagdale

Department of Computer Science

swatigambhire@pes.edu