

Principal Component Analysis (PCA) - An Efficient Way of Variable Reduction Method and its Applications in Oncology Trials

Author: Pradeep Acharya

Stream: Analytics and Statistics (AS)

Presentation #AS09

PHUSE EU Connect 2024, Strasbourg Convention Centre, 10-13 November

- Introduction to Principal Component Analysis (PCA)
- Steps to Compute Principal Components (PCs) and its Model
- Applications of PCA
- Methodology, Source Data and its Background
- Statistical Analysis and Results
- Advantages and Drawbacks
- Conclusion
- References

Introduction to PCA

Statistical technique for Variable Reduction in data

Easy to compute PCs

PCs are linear combination of actual variables

Uncorrelated to each other, has zero mean and constant variance

First few PCs explain maximum variance in the data

Steps to Compute PCs

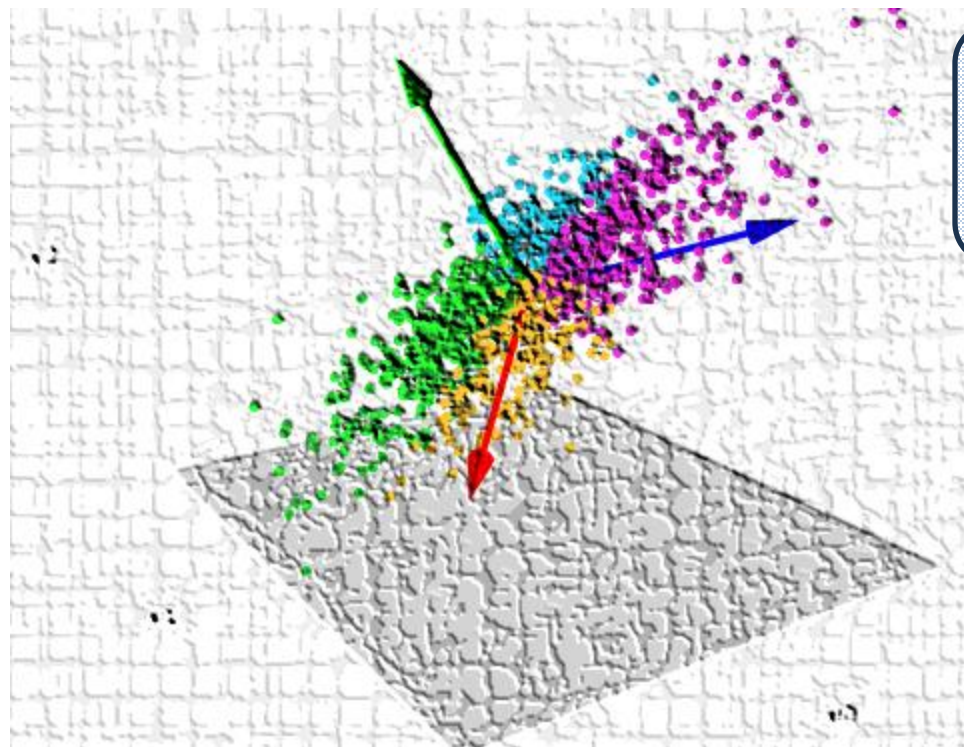
Standardization

Covariance Matrix Computation

Eigen Values and Eigen Vectors

Create Feature Vector

Transform the Data to New Coordinate System



General Model to Compute PCs

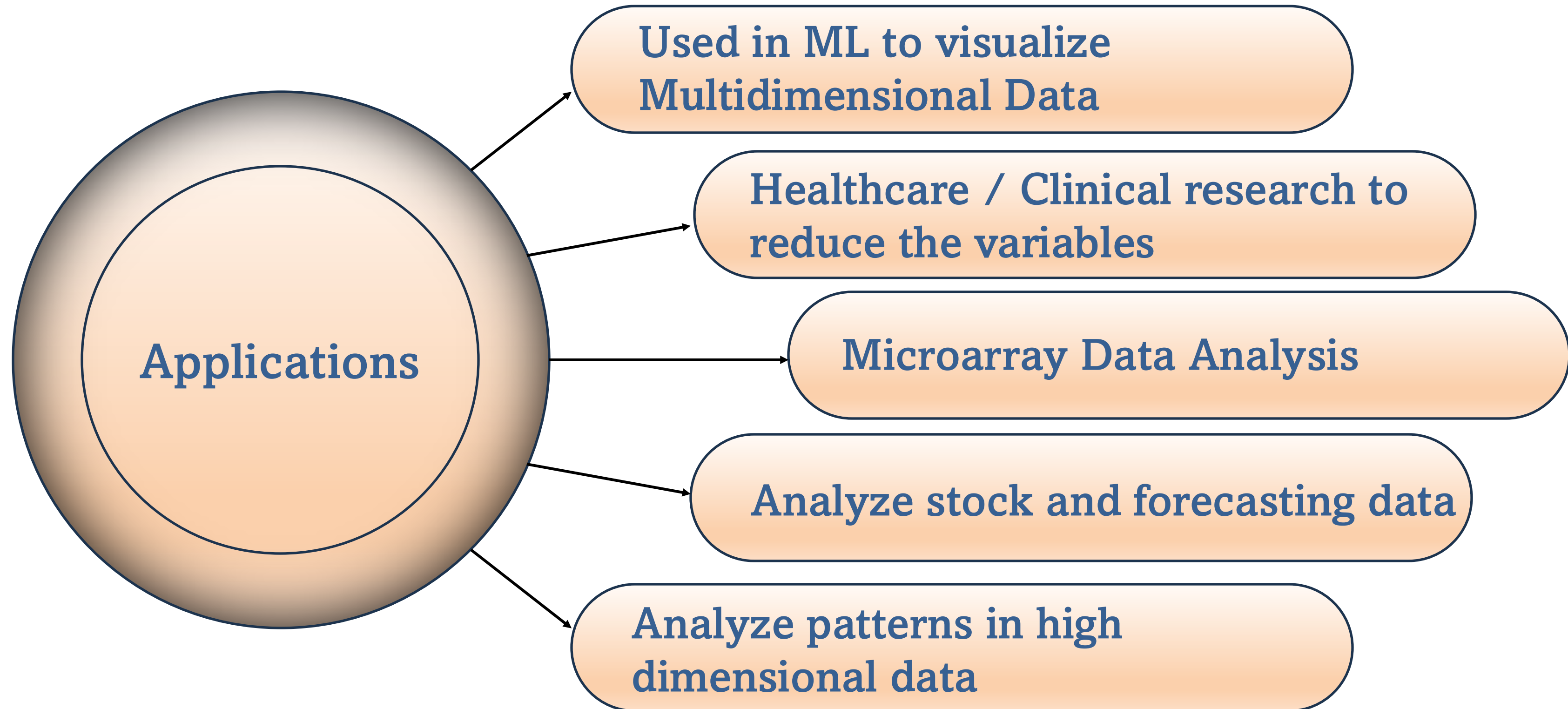
$$C_1 = b_{11}(X_1) + b_{12}(X_2) + \dots + b_{1p}(X_p)$$

C_1 = Subject's score on principal component 1 (the first component extracted)

b_{1p} = Regression coefficient (or weight) for observed variable p , as used in creating principal component 1

X_p = Subject's score on observed variable p

Applications of PCA



Methodology, Source Data and its Background

Hypothetical Data Structure

usubjid	trt01pn	age	param	paramcd	aval	cnsr	NUMCIGD	SMKDUR	INSMKDUR	EXPRADN	EXPASB
01-701-1015	1	63	Overall Survival	OS	2	0	16	15.2	11.5	7.7	9.5
01-701-1023	1	64	Overall Survival	OS	3	0	14	2.3	4	6.9	11.6
01-701-1028	0	71	Overall Survival	OS	3	0	16	15	11.4	6.5	11.9
01-701-1033	0	74	Overall Survival	OS	28	1	12	1.2	3.4	7.2	10.9
01-701-1034	0	77	Overall Survival	OS	58	0	9	15.3	12.5	7.7	9.9
01-701-1047	1	85	Overall Survival	OS	46	1	8	2.2	1.3	6.8	11.6
01-701-1097	0	68	Overall Survival	OS	3	0	18	15.8	7.6	7.6	10.4
01-701-1111	0	81	Overall Survival	OS	11	1	22	0.8	2.6	8.2	9.1
01-701-1115	0	84	Overall Survival	OS	3	0	12	4.6	3.9	8.6	9.1
01-701-1118	1	52	Overall Survival	OS	82	1	14	15.2	11.5	9.1	8.3
01-701-1130	1	84	Overall Survival	OS	97	0	12	15.3	11.5	7.5	10.5
01-701-1133	0	81	Overall Survival	OS	61	0	10	15.3	12.6	7.4	10.7
01-701-1146	0	75	Overall Survival	OS	13	0	16	3.2	0.3	7.8	9.5
01-701-1148	0	57	Overall Survival	OS	3	0	15	15.2	12.5	7.4	10.8
01-701-1153	1	79	Overall Survival	OS	91	1	6	14.6	6.3	8.2	9
01-701-1180	0	56	Overall Survival	OS	36	0	16	2.9	1.4	7.5	10.5

Influence of explanatory variables on Outcome

```
proc corr data=sltte;
  var aval numcigd smkdur insmkdur expasb bmibl;
run;
```

- All explanatory variables are significantly correlated with survival time
- It is also observed that, few independent variables are inter related within them
Ex: SMKDUR vs INSMKDUR
EXPASB vs BMIBL

Pearson Correlation Coefficients, N = 216 Prob > r under H0: Rho=0						
	aval	NUMCIGD	SMKDUR	INSMKDUR	EXPASB	bmibl
aval	1.00000	-0.19022 0.0050	0.53308 <.0001	0.44824 <.0001	-0.14803 0.0296	-0.14930 0.0282
NUMCIGD	-0.19022 0.0050	1.00000	-0.06486 0.3428	-0.02538 0.7107	-0.07502 0.2723	-0.07444 0.2761
SMKDUR	0.53308 <.0001	-0.06486 0.3428	1.00000	0.86267 <.0001	0.00888 0.8968	0.00856 0.9005
INSMKDUR	0.44824 <.0001	-0.02538 0.7107	0.86267 <.0001	1.00000	0.01343 0.8445	0.01338 0.8450
EXPASB	-0.14803 0.0296	-0.07502 0.2723	0.00888 0.8968	0.01343 0.8445	1.00000	0.99964 <.0001
bmibl	-0.14930 0.0282	-0.07444 0.2761	0.00856 0.9005	0.01338 0.8450	0.99964 <.0001	1.00000

Cox Proportional Hazards Model

```
proc phreg data=sltte;
  class trt01pn(ref="0");
  model aval*cnsr(1)=trt01pn numcigd smkdur insmkdur expasb bmibl/
    ties=breslow risklimits alpha=0.05;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.9068	6	0.1789

Analysis of Maximum Likelihood Estimates

Parameter		DF	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
trt01pn	1	1	1.7100	0.1910	0.699	0.409	1.196
NUMCIGD		1	1.2890	0.2562	1.035	0.976	1.097
SMKDUR		1	1.1561	0.2823	1.043	0.966	1.127
INSMKDUR		1	0.0784	0.7794	0.988	0.908	1.075
EXPASB		1	0.0497	0.8235	0.479	0.001	10.515
bmibl		1	0.0682	0.7940	1.390	0.117	16.481

Principal Component Analysis

```
proc princomp data=sltte out=sltte1 outstat=stat;  
    var numcigd smkdur insmkdur expasb bmibl;  
run;
```

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.01616243	0.15469869	0.4032	0.4032
2	1.86146374	0.87584263	0.3723	0.7755
3	0.98562111	0.84923061	0.1971	0.9726
4	0.13639050	0.13602828	0.0273	0.9999
5	0.00036222		0.0001	1.0000

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
NUMCIGD	-.113237	-.048797	0.991822	0.032928	0.000381
SMKDUR	0.132191	0.693725	0.025734	0.707537	-.000709
INSMKDUR	0.133253	0.691861	0.072688	-.705895	0.000591
EXPASB	0.689936	-.137207	0.071625	0.003729	0.707120
bmibl	0.689880	-.137356	0.072197	0.002448	-.707093

For each subject: PC1 = -0.113 (NUMCIGD) + 0.132 (SMKDUR) + 0.133 (INSMKDUR) + 0.69 (EXPASB) + 0.69 (BMIBL)

Statistical Analysis and Results Contd...

Principal Components created by PCA

usubjid	trt01pn	age	param	paramcd	aval	cnsr	NUMCIGD	SMKDUR	INSMKDUR	EXPRADN	EXPASB	Prin1	Prin2	Prin3	Prin4	Prin5
01-701-1015	1	63	Overall Survival	OS	2	0	16	15.2	11.5	7.7	9.5	0.3747015425	1.1427247578	1.0232703775	0.1240136499	-0.025145999
01-701-1023	1	64	Overall Survival	OS	3	0	14	2.3	4	6.9	11.6	2.5183452079	-1.915542857	0.5417854887	-0.377918539	0.0246649079
01-701-1028	0	71	Overall Survival	OS	3	0	16	15	11.4	6.5	11.9	3.3624495108	0.5091123945	1.3330977132	0.1274422089	-0.027936522
01-701-1033	0	74	Overall Survival	OS	28	1	12	1.2	3.4	7.2	10.9	1.7268832246	-1.938412452	-0.139734144	-0.448732417	-0.032592121
01-701-1034	0	77	Overall Survival	OS	58	0	9	15.3	12.5	7.7	9.9	1.123938024	1.2973583417	-0.950509781	-0.071616607	-0.014097534
01-701-1047	1	85	Overall Survival	OS	46	1	8	2.2	1.3	6.8	11.6	2.6432060652	-2.217712516	-1.24646049	-0.064103305	0.023684168
01-701-1097	0	68	Overall Survival	OS	3	0	18	15.8	7.6	7.6	10.4	1.3014840944	0.425866029	1.664576658	0.7754413083	0.0131142704
01-701-1111	0	81	Overall Survival	OS	11	1	22	0.8	2.6	8.2	9.1	-0.924151015	-1.785468336	2.5220870943	-0.29682662	0.0132548893
01-701-1115	0	84	Overall Survival	OS	3	0	12	4.6	3.9	8.6	9.1	-0.470691054	-1.010160042	-0.355937381	-0.11869536	0.0118299201
01-701-1118	1	52	Overall Survival	OS	82	1	14	15.2	11.5	9.1	8.3	-1.068136591	1.471739661	0.2831833061	0.0979220597	-0.011776278
01-701-1130	1	84	Overall Survival	OS	97	0	12	15.3	11.5	7.5	10.5	1.7281882159	0.9695378636	-0.0148429	0.1028973858	0.0037066685
01-701-1133	0	81	Overall Survival	OS	61	0	10	15.3	12.6	7.4	10.7	2.0678826434	1.1030471206	-0.556099752	-0.071779836	0.0094762491
01-701-1146	0	75	Overall Survival	OS	13	0	16	3.2	0.3	7.8	9.5	-0.244780041	-1.832290992	0.801482863	0.2647548653	0.0237437608
01-701-1148	0	57	Overall Survival	OS	3	0	15	15.2	12.5	7.4	10.8	2.006524012	0.9836979659	0.9103378602	-0.020783997	0.0251594211
01-701-1153	1	79	Overall Survival	OS	91	1	6	14.6	6.3	8.2	9	-0.066757748	0.6137056075	-2.034021234	0.6913750236	-0.029275467
01-701-1180	0	56	Overall Survival	OS	36	0	16	2.9	1.4	7.5	10.5	1.0437079382	-1.96642623	0.9482092406	0.0776825638	0.0044549962
01-701-1181	0	79	Overall Survival	OS	8	1	15	0.4	4.8	7.4	10.6	1.2216577403	-1.796627141	0.7070686712	-0.717440491	0.0201741014
01-701-1192	0	80	Overall Survival	OS	17	0	12	15.3	11.9	7.3	10.6	1.8488465547	1.003418417	0.0024654995	0.0465209498	0.0188776476
01-701-1203	1	81	Overall Survival	OS	83	1	13	15.3	13.5	7.8	9.8	0.9077944266	1.4012648072	0.2181951369	-0.175633637	-0.028654215
01-701-1234	1	69	Overall Survival	OS	77	1	18	14.8	8.1	8.7	8.9	-0.570621399	0.7476269784	1.4733243563	0.5750097537	0.0059026659
01-701-1275	0	61	Overall Survival	OS	18	0	12	9.5	7.1	8.2	9.3	0.0171815671	-0.04085842	-0.257146222	0.0203266919	-0.03128906
01-701-1287	0	56	Overall Survival	OS	29	0	14	15.3	11	8.2	8.9	-0.324612527	1.2637407887	0.3549670453	0.1844953085	-0.018644686

Cox Proportional Hazards Model with Principal Components

```
proc phreg data=slttel;
  class trt01pn(ref="0");
  model aval*cnsr(1)=trt01pn prin1 prin2 prin3/ties=breslow risklimits alpha=0.05;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.7808	4	0.0006

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
trt01pn	1	1	5.2503	0.0219	0.324	0.281	0.808
Prin1		1	12.2126	0.0005	1.227	1.084	1.389
Prin2		1	2.7872	0.0950	1.116	0.959	1.298
Prin3		1	0.9653	0.3259	1.118	0.915	1.367

Advantages & Drawbacks of PCA

Advantages

- Efficient technique for variable reduction
- Explains maximum variance in the data
- Highly efficient on large volume of data

Drawbacks

- Difficult to interpret PCs based on relevance
- Difficult to compute and interpret Covariances and Covariance matrices

Conclusion

- One of the powerful methods to reduce the dimension of the data and PCs obtained explains maximum variance in the data
- Recommended to do the Principal Component Analysis (PCA) when there are vast number of variables which are exceedingly difficult to analyze and interpret

References

- A Ghosha, S Barmanb. Prediction of Prostate Cancer cells Based on Principal Component Analysis Technique. Procedia Technology 10 (2013) 37-44
- Alexe G, Dalgin G S, Ganesan S, DeLisi C and Bhanot G. Analysis of breast cancer progression using principal component analysis and clustering; J. Biosci. 32 (2007) 1027–1039
- Ying-Lin Hsu, Po-Yu Huang, Dung-Tsa Chen. Sparse principal component analysis in cancer research. Transl. Cancer Res. 2014;3(3):182-190
- T Bankole-Oye, I El-Thalji and J Zec. Combined principal component analysis and proportional hazard model for optimizing condition-based maintenance. IOP Conf. Series: Materials Science and Engineering 1201 (2021)
- I.T. Jolliffe. Principal Component Analysis, Second Edition. 2002, Springer Verlag, New York, Berlin, Heidelberg.



ISO 9001 & 27001 CERTIFIED COMPANY



THANK YOU