Principal Component Analysis for Detection of Globally Important Input Parameters in Nonlinear Finite Element Analysis

Sascha Ackermann[1]*, Lothar Gaul[1], Michael Hanss[1], Thomas Hambrecht[2]

[1] IAM - Institute of Applied and Experimental Mechanics, Stuttgart, Germany

[2] Aluminium- and Lightweight Centre, AUDI AG, Neckarsulm, Germany

Abstract

The presented approach utilizing the *Principal Component Analysis* (*PCA*) shows a multivariate analysis method to detect correlations of input parameters onto a complete structural Finite-Element-Model. The advantage of the approach is, that it permits to incorporate all nodes or elements in the analysis of an output parameter like total displacement or plastic strain. This avoids to miss important input parameters or to overestimate unimportant input parameters, because all information of the model is used and no previous knowledge like choice of a certain node for displacement is necessary. After a theoretical development of the approach and an historical application of the method, the feasibility of the presented approach for crash applications is examined on a comprehensible U-structure. It shows that the Principal Component Analysis reduces the amount of data to a manageable size and therefore offers the user a perfect alternative for further insights into the model.

Keywords: PCA, Crash Simulation, Coefficient of Correlation, PAMCRASH, Latin Hypercube Sampling

*Contact: Dipl.-Ing. Sascha Ackermann, IAM - Institute of Applied and Experimental Mechanics, University of Stuttgart, Pfaffenwaldring 9, D-70569 Stuttgart, Germany, E-Mail: ackermann@iam.uni-stuttgart.de

1 Introduction

In many engineering applications nonlinear Finite Element Analysis is used to calculate the time-dependent behaviour of dynamically loaded structures. The employed models often have dozens of input parameters to describe the physical properties of the investigated structure. Whereas some of these parameters do not show significant influence on the output of the model, others have great impact on the results. It is essentially to identify the most important input parameters. This procedure is often referred to as sensitivity analysis.
Usually in sensitivity analyses scalar input parameters are compared with scalar output parameters of the structural model. As output parameters often the deflection of a single node or the plastic strain of a single finite element are analysed. For simple structures and load cases the choice of the most significant node to observe i.e. for deflection is quite simple. The situation is more difficult for complex structures combined with complex loading conditions: In this case it is hard to completely guarantee that the chosen node is really significant.
An alternative would be to observe all nodes and to detect the correlations between the input parameters and the nodal deflections. However, this is no more feasible for large models. A solution can be found in the application of a data reduction technique such as the Principal Component Analysis (PCA), Marinell (1995). This method reduces the variation of the output values to a low number of few important correlation rates, which explain a major part of variation of the model output (i.e. more than 95% or more than 99%). In this way, the global behaviour of the structure can be described with a marginal number of parameters. For each input parameter a correlation rate for the reduced number of output parameters can be formed.
Whereas the first section of this text gives a short introduction and explains the motivation for the presented approach, the second section originally explains the method mathematically, continues with a simple example and ends with the proposed algorithm implementing a software tool for the automatic PCA of time-dependent output data from a nonlinear FEM solver. The third section shows some applications and their results. Conclusions are drawn in the last section.

2 Main Part

2.1 Numerical background of PCA

The PCA starts with base data described by a $n \times p$ matrix X,

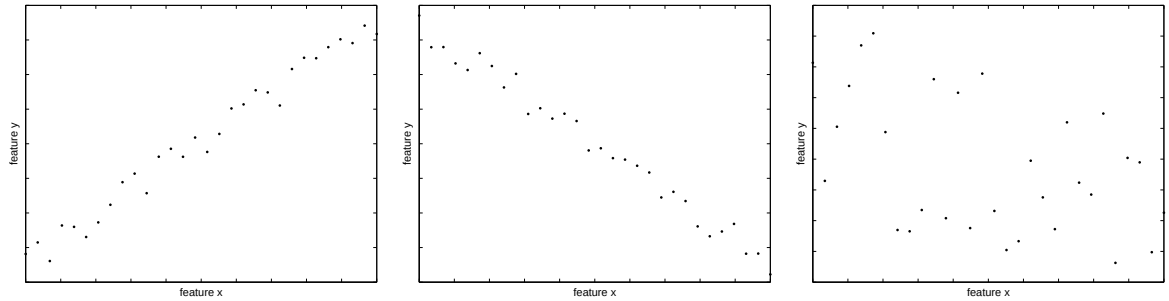
$$X = (x_{ij})_{1 \le i \le n,\ 1 \le j \le p} \tag{1}$$

where n is the number of samples and p the number of different measured features. So each line of the matrix stands for one sample with its measured properties and each row of the matrix gives the different measured values for one feature. In a more mathematical sense this matrix gives a discrete set of n data items in a p-dimensional space. Thus utilizing the Principal Component Analysis the data items are projected in a q-dimensional subspace R_q $(q < p)$ therefore the information loss becomes preferably low.
Technically the approach of the PCA is a principal axis transformation, which minimizes the correlation of (Gaussian-)multivariate variables by the conversion to a vector space with a new basis. The principal axis transformation can be specified by an orthogonal matrix, which is established by eigenvectors of the covariance matrix (respectively correlation matrix) of the

samples and each feature. For this reason the PCA is always problem-related, in the sense that already a single different sample or an additional sample will produce more or less different results for PCA.
Before performing the principal axis transformation initially the correlations of the different features have to be calculated using the *empirical Coefficient of Correlation* according to Eq. (2)

$$r = r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2}$$

This equation often also is referred to as *Bravais-Pearson Coefficient of Correlation* . The numerator of the equation describes the empirical covariance and the denominator stands for standardization. Further $\overline{x}$ and $\overline{y}$ express the arithmetical means of x and y respectively.
The Coefficient of Correlation gives information about the qualitative relationship of two parameters. Unlike to a full regression analysis it establishes for instance no quantitative relationship of the form $f(x) = y = ax + b$. For the empirical Coefficient of Correlation holds $r_{xy} \in [-1, 1]$. Thus a straight line with positive gradient results in $r_{xy} = 1$, whereas a negative gradient will produce $r_{xy} = -1$. For non-optimal relations of two parameters the Correlation Coefficient is somewhere between $r_{xy} \in]-1, 1[$. The restriction of this kind of measuring the



$r_{xy} \approx 1$ $r_{xy} \approx -1$ $r_{xy} \approx \pm 0$

Figure 1: different correlations of two features

relationship of two parameters is the lack of the Coefficient of Correlation to detect only links of linear kind. Additionally certain special linear relationships of parameters are not perceived, Fahrmeir et al. (2002).
The Coefficient of Correlation is composed for all combinations of features p of matrix X according to Eq. (1). This yields a correlation matrix K of dimension $dim(K) = p \times p$, which contains the normalized relations of all items.
For detecting the fundamental characteristics of the correlation matrix K a special eigenvalue problem, according to Eq. (3) is solved.

$$(K - \lambda E)x = 0 \tag{3}$$

Each resulting eigenvector forms a linear combination with its components. The larger an eigenvalue the more variance of the model is explained by the associated eigenvector. Thus starting with the highest eigenvalue, the respective eigenvector explains the maximum possible amount of the variance of the matrix values. With the second largest eigenvalue respectively its linear combination, the maximum possible amount of the datas remaining variance is explained. The same holds for all subsequent eigenvalues and eigenvectors. All linear combinations together

explain the complete variance of the matrix data.
The rate of overall variance σ^2, which comprises a linear combination of an eigenvector i is described by Eq. 4.

$$\sigma_i^2 = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{4}$$

For the overall variance, which is declared by eigenvalues $[\lambda_r, \lambda_s]$ and its associated eigenvectors, holds Eq. 5.

$$\sum_{i=r}^{s} \sigma_i^2 = \frac{\sum_{i=r}^{s} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{5}$$

To achieve data reduction, a restriction to the most important respectively highest eigenvalues and its linear combinations is necessary. Important is a reasonable rate of data reduction, to explain a suitable amount of variation. Consequently the number of eigenvectors to incorporate has to be determined. The chosen, most important eigenvectors are denoted as *Principal Components* (*PC*).
For the identification of the number of Principal Components there exist numerous approaches. The most prevalent one is the Eigenvalue Criterion, Marinell (1995) according to Eq. 6, which is also referred to as Kaiser-(Guttmann)-Criterion, Guttman (1954).

$$PC = \{\lambda_1, \lambda_2, ..., \lambda_k\} \;\; \forall \;\; \lambda_i > R \quad with \quad R = 1 \tag{6}$$

A modification of this is the Joliffe-Criterion, which sets $R = 0.7$. Background of Eq. 6 is to retain only linear combinations, which explain more variance than the original variables. This holds only for eigenvalues λ_i with $\lambda_i > 1$. Alternatively a kind of coefficient of determination according to Eq. 5 can be utilized to explain a defined rate of total variance, i.e. 99%, Abonyi and Feil (2007). Also the number of eigenvalues can be fixed, or the maximum number of Principal Components is capped by the number n of the samples. This corresponds with information theory and its applications in image processing. Furthermore the so-called *Elbow-Criterion* , often referred to as Scree-Plot, can be applied. Best is to combine several criteria.
Typically, before application of PCA a statistical test is necessary to ascertain, if a data reduction of the assessed correlation matrix is reasonable and even possible, see Backhaus et al. (2005), Marinell (1995). For meaningful applications in the area of uncertain model parameters in Finite-Element-Structural-Analysis this seems to be needless.
After realization of PCA the new variables for each sample are calculated from the linear combinations of the old, original variables. Again utilizing Eq. 2 the correlations of the considered variables are estimated. Hence the number of necessary correlations to examine reduces to $|Corr| = |PC| \cdot n$.

2.2 Out of ordinary topic, but original applications

Like many statistical methods the Principal Component Analysis was originally used primarily in the field of sociology and psychology until powerful and fast computer technology emerged. A very similar approach is the Factor Analysis, Backhaus et al. (2005). With this method the psychologist Charles Spearman showed in 1904 that the results of intelligence tests with different variables can be expressed by a single one-dimensional character item of a person. This established the general factor of intelligence.
The principal procedure of the PCA can be described pretty easily by means of an example of

sociology/market research, a modified version of the example is stated in Marinell (1995). 172 Children were examined regarding 8 emotional behavioural patterns ($x_i \; \forall \; i \; \in \; [1, ..., 8]$). With Eq. 2 the correlation matrix in tab. 1 yields.

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.0	0.59	0.35	0.34	0.63	0.4	0.28	0.20
X2		1.0	0.42	0.51	0.49	0.52	0.31	0.36
X3			1.0	0.38	0.19	0.36	0.73	0.24
X4				1.0	0.29	0.46	0.27	0.39
X5					1.0	0.34	0.17	0.23
X6						1.0	0.32	0.33
X7							1.0	0.24
x8								1.0

Table 1: Correlation matrix with 8 emotional behavioural patterns

From this data eight eigenvalues result:

$\lambda_1 = 3.625$	$\lambda_2 = 1.241$	$\lambda_3 = 0.953$	$\lambda_4 = 0.655$
$\lambda_5 = 0.536$	$\lambda_6 = 0.418$	$\lambda_7 = 0.325$	$\lambda_8 = 0.245$

Utilizing the Eigenvalue Criterion according to Eq. 6 produces 2 Principal Components. The corresponding eigenvectors show the following linear combinations:

$$e_1^T = 0.379x_1 + 0.422x_2 + 0.358x_3 + 0.358x_4 + 0.328x_5 + 0.369x_6 + 0.318x_7 + 0.277x_8 \quad (7)$$

$$e_2^T = -0.331x_1 - 0.189x_2 + 0.529x_3 - 0.007x_4 - 0.477x_5 - 0.048x_6 + 0.587x_7 + 0.022x_8 \quad (8)$$

With Eq. 5 the first two Principal Components explain (3.625 + 1.241)/8 accordingly to 60.82 % of the total variance. Often a low number of Principal Components explains much higher rates of the total variance. Thus 5 - 10 Principal Components define up to 90% of total variance of models described originally by hundreds of parameters.

2.3 Implementation

A software implementation of PCA for automatized analysis of result data of an explicit FEM solver like PAMCRASH 2G, ESI GROUP (2006) offers additional potentials. Thus the correlations of Principal Components at variable times of the simulated period can be checked. Also possible is a visualisation of the Principal Components in proportions of incorporated nodes or elements for items like displacement or plastic strain. Another interesting option in this context is the analysis of different structural subassemblies as well as an outlier detection for the samples, which correspond to the different solver runs. The course of the software implementation is as follows:

1. Definition of analysis items (i.e. node displacements)

2. Definition of output point in time to analyse the node respectively element output parameters

3. Definition of nodes respectively elements to analyse through individual selection or choice of complete subassemblies

4. Choice of further analysis sets (i.e. element strains at a different output time) back to 1.) , else continue

5. Determination of correlation matrix

6. Calculation of eigenvalues and eigenvectors of correlation matrix

7. Estimation of number of Principal Components

8. For more than one analysis set, repeat 5.) - 8.) for relevant times

9. Calculation of new variables with the help of linear combinations

10. Estimation of new correlations by Eq. 2 between all input parameters and all new Principal Components

11. Output of correlation matrix

12. Visual evaluation by user, meaning: scatter plots, visual presentation of linear combinations

13. Outlier detection by cluster analysis or adjustment with reference sample

3 Application

3.1 Academic U-structure

The feasibility of the presented approach was evaluated with a comprehensible, academic example: A U-structure modeled with shell elements (Fig. 2) . The pillars of the structure are fixed at the floor in all degrees of freedom and a mass is falling onto the crossbar with a given velocity. The shape of the falling mass consists of four rigid shell elements. All elements contribute to a total sum of nodes of 134.
The evaluation was performed for 24 input parameters and the variations of these parameters. For this, 30 calls of the FEM Solver PAMCRASH 2G were conducted with varying input parameters predetermined by an algorithm for Latin Hypercube Sampling. Hence the distribution of input parameters for this 30 solver runs was nearly uncorrelated by Eq. 2, Will (2007b, a). Additionally a solver run with nominal parameters was conducted.

The analysis was performed for the node parameters total displacement (length of summed vector of x,y,z-deflection) and displacement in y direction. Utilizing Eq. 6 for choice of number of Principal Components, each eigenvalue λ_i with $\lambda_i > 1$ and its associated eigenvector contributes for evaluation. In complete for the chosen example and the investigated parameters result 4 Principal Components for total displacement and 7 Principal Components for y-displacement. These linear combinations declared 98.5% respectively 98.8% of the total

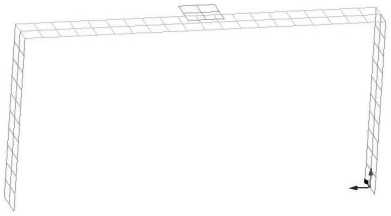


Figure 2: Original shape

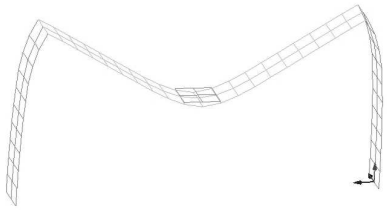


Figure 3: Deflected shape

$\sum_{i=1}^{s} \sigma_i^2$	98.5	98.8
λ_1	53.631	50.242
λ_2	36.205	33.579
λ_3	4.8791	6.0063
λ_4	3.8259	3.5128
λ_5	-	2.4038
λ_6	-	1.7545
λ_7	-	1.3721

Table 2: Defined total variance

variance of original model output values. As Table 2 shows, the largest share of total variance of both output items is actually already explained by the first two principal components.

A graphical depiction of the correlation of the 24 input parameters (x-axis) and the Principal Components (y-axis) shows the correlation matrix in Fig. 4. Though the last four lines correspond to the correlations of the input parameters and the Principal Components of the total node displacements, whereas the first seven lines show correlations of the input parameters and the Principal Components of y-axis node displacements. Accordingly to the definition of the Coefficient of Correlation, the values of the correlation matrix range in the interval of $[-1, 1]$. More blackened fields in Fig. 4 show high either negative or positive correlations between an input parameter and a Principal Component as output parameter.
Evaluating the matrix it is always important to have a look onto the participation of the particular Principal Component, which is described through its contribution to the declaration of total variance according to Eq. 4.

Thus in the example the sixth Principal Component of y-displacement has only share of 1.75% for the degree of declaration of total variance, however the correlation of the 24th input parameter shows a particular strong correlation with this Principal Component. Due to the low value for the degree of declaration the information about importance of the 24th input value on the whole model should be considered critical.
However the input values 1, 2 and 18 are very important. Input parameter 1 and 2 correspond to the thickness of the crossbar and the pillars, whereas input parameter 18 stands for the mass of the falling rigid 4-shell-plate. Furthermore a mid-strong correlation of the input variables 6, 11 and the already mentioned variable 24 seems to exist. Parameters 6 and 11 relate to yield strength of the materials and parameter 24 is the initial velocity of the falling mass.

Coming from engineering knowledge, all previously stated and as important labeled variables make sense. For all other fields of the correlation matrix, differing a little from the clear green color the following questions are recommended to be checked:

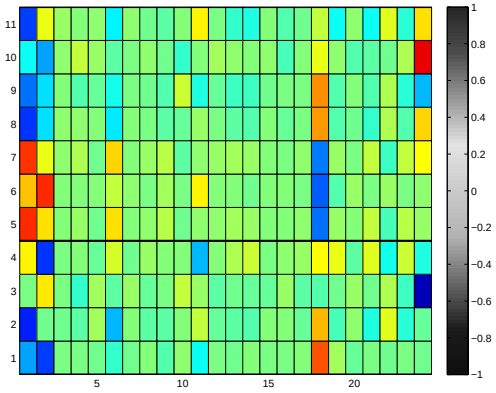


Figure 4: correlation matrix: 24 input parameters vs. 13 Principal Components

- What is the contribution of the particular Principal Component to the degree of declaration formed by Eq. 4 ?
- Which are the nodes contributing particularly strong to the respective Principal Component (= have large coefficients in the particular linear combination) and where are these nodes located ?
- Check the scatter plot of the respective Principal Component and the respective input variable, if any abnormalities (outlier, nonlinear relations) are to detect, so that this could have distorted the value for the Coefficient of Correlation according to Eq. 2.
- Check the variation range of the input parameters ? Is the variation too broad or too narrow ?

A comparison to the simple observation of a single node deflection is given in Figs. 5 and 6. The observed single item is one of the nodes at the exact middle of the horizontal bar in the structure depicted in Fig. (2). Whereas the left illustration shows a clear relation between an increase of the thickness of the horizontal beam, no significant relation between an increase of Poissons ratio of the beam and the deflection of the observed node is to detect.

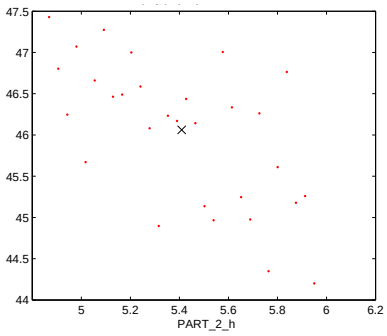


Figure 5: $r_{XY} \approx -0.63$

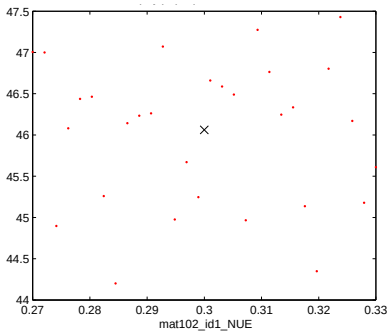


Figure 6: $r_{XY} \approx -0.03$

The determination of the most significant node for deflection observation of this comprehensive example is quite clear. The results for important correlations between the deflection of this node and an exemplary input parameter coincide with the results derived for the application

of the data reduction technique of Principal Component Analysis. For each other important input parameter determined by application of PCA the results are easy to follow by engineering knowledge.

input parameter	attribution	minimum	nominal	maximum
1	shell thickness pillars	4.5409	5.0455	5.55
2	shell thickness horizontal beam	4.8682	5.4091	5.95
3	shell thickness four rigid shells	4.5	5	5.5
4	material density pillars	2.43e-006	2.7e-006	2.97e-006
5	Youngs modulus material pillars	65.045	72.273	79.5
6	yield strength material pillars	0.16528	0.18364	0.202
7	Poissons ratio material pillars	0.27	0.3	0.33
8	material damping pillars	0.18	0.2	0.22
9	material density crossbeam	2.43e-006	2.7e-006	2.97e-006
10	Youngs modulus material crossbeam	65.045	72.273	79.5
11	yield strength material crossbeam	0.18	0.2	0.22
12	Poissons ratio material crossbeam	0.27	0.3	0.33
13	material damping crossbeam	0.18	0.2	0.22
14	contact thickness	0.36	0.4	0.44
15	nonlinear contact stiffness	0	0	200
16	contact friction	0	0	0.3
17	contact damping	0.09	0.1	0.11
18	mass of four rigid shells	0.98181	1.0909	1.2
19	1st deviatoric moment of rigid shells	0.9	1	1.1
20	2nd deviatoric moment of rigid shells	0.9	1	1.1
21	3rd deviatoric moment of rigid shells	0.9	1	1.1
22	initial velocity of rigid shells, 1st direction	-0.1	0	0.1
23	initial velocity of rigid shells, 2nd direction	-0.1	0	0.1
24	initial velocity of rigid shells, 3rd direction	-10.473	-11.636	-12.8

Table 3: Varied input parameters

4 Conclusions

The presented approach utilizing the *Principal Component Analysis* (*PCA*) shows a multivariate analysis method to detect correlations of input parameters onto a complete structural Finite-Element-Model. The advantage of the approach is, that it permits to incorporate all nodes or elements in the analysis of an output parameter like total displacement or plastic strain. This avoids to miss important input parameters or to overestimate unimportant input parameters, because all information of the model is used and no previous knowledge like choice of a certain node for displacement is necessary. The Principal Component Analysis reduces the amount of data to a manageable size and therefore offers the user a perfect alternative for further insights into the model.

The authors wish to express their appreciation and thanks to ESI GmbH in Eschborn, Germany for software and support. Finite Element simulation has been executed using PAM-CRASH 2G 2007. Result analysis and pre-, postprocessing was supported using the graphical simulation environment Visual-Environment v4.5.

References

J. Abonyi and B. Feil. *Cluster Analysis for Data Mining and System Identification*. Birkhaeuser Basel, 1 edition, 8 2007. ISBN 9783764379872.

K. Backhaus, B. Erichson, W. Plinke, and R. Weiber. *Multivariate Analysemethoden: Eine anwendungsorientierte Einfuehrung (Springer-Lehrbuch)*. Springer, 11., berarb. aufl. edition, 10 2005. ISBN 9783540278702.

ESI GROUP. *PAM-CRASH(TM) PAM-SAFE(TM), 2006, Solver Reference Manual*. ESI GROUP, 2006.

L. Fahrmeir, R. Kuenstler, I. Pigeot, and G. Tutz. *Statistik. Der Weg zur Datenanalyse*. Springer, Berlin, 9 2002. ISBN 9783540440000.

L. Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2): 149–161, 1954.

G. Marinell. *Multivariate Verfahren*. Oldenbourg Verlag Muenchen, Wien, 1995. ISBN 3486233025.

J. Will. optiSLang - the optimizing Structural Language, Manual. *www.dynardo.de*, 2007a.

J. Will. Introduction of robustness evaluation in CAE-based virtual prototyping processes of automotive applications. *EUROMECH Colloquium, London*, 2007b.