Simple factor analysis of measured data

Ivica Kožar[*1], Danila Lozzi Kožar[2a] and Neira Torić Malić[1b]

[1]Faculty of Civil Engineering, University of Rijeka, Radmile Matejčić 3, 51000 Rijeka, Croatia
[2]Croatian Waters-Unit Rijeka, Đure Šporera Street 3, 51000 Rijeka, Croatia

Abstract.   Quite often we have a lot of measurement data and would like to find some relation between them. One common task is to see whether some measured data or a curve of known shape fit into the cumulative measured data. The problem can be visualized since data could generally be presented as curves or planes in Cartesian coordinates where each curve could be represented as a vector. In most cases we have measured the cumulative 'curve', we know shapes of other 'curves' and would like to determine unknown coefficients that multiply the known shapes in order to match the measured cumulative 'curve'. This problem could be presented in more complex variants, e.g., a constant could be added, some missing (unknown) data vector could be added to the measured summary vector, and instead of constant factors we could have polynomials, etc. All of them could be solved with slightly extended version of the procedure presented in the sequel. Solution procedure could be devised by reformulating the problem as a measurement problem and applying the generalized inverse of the measurement matrix. Measurement problem often has some errors involved in the measurement data but the least squares method that is comprised in the formulation quite successfully addresses the problem. Numerical examples illustrate the solution procedure.

Keywords: factor analysis; least squares method; measurement data; measurement error; measurement problem

1. Introduction

Engineering experiments produce a lot of measurement data that have to be somehow related to one another or to results of some known formulation. One often makes cumulative measurements in a process that comprises other simpler relations, e.g., we are measuring a cumulative mixture of some constituents or forces in a structure that includes some simpler mechanisms. Usually, factor analysis is applied to this type of problem (see e.g., Menke 2012). In Kožar *et al*. (2020) cumulative curve stochastically comprises experimental data and factor analysis is applied as $S = CF$ where S is the matrix of experimental results, F is the matrix of factors that comprise experimental results, C is the matrix of factor loadings. The matrix S that is known is decomposed using singular value decomposition into a product of matrices C and F that are to be determined. The resulting eigenvalues describe the presence of factors within the cumulative curve. However, the above decomposition is not unique and only the largest eigenvector has a physical meaning, it represents the mean of the contributing curves.

∗Corresponding author, Professor, E-mail: ivica.kozar@gradri.uniri.hr

Other approaches are possible as well and in this work we will treat the problem as a measurement problem and in vector notation (e.g., see Kožar 2016, Ibrahimbegović *et al*. 2009). It is well known (see e.g., Gibbs 2011) that any vector can be represented as a sum of vectors that span its vector space, $\boldsymbol{v} = \sum_{i=1}^{m} c_i \boldsymbol{h}_i$. This problem is further complicated with errors in measurement, represented as an error vector that can be additive or multiplicative; in the later case error is correlated with measurement since larger measurements have lager error. This is important in formulation of correction schemes like weighted least squares, etc. In this paper error correction is formulated as the weighted least squares procedure based on the calculation residual.

2. Formulation of the problem

A common task in engineering experimental analysis is to see whether some measured data fit into another measured data, see e.g., Kožar and Lozzi-Kožar (2017), Lozzi-Kožar and Kožar (2017) or Kožar *et al*. (2011). There are *n* independent experiments, each having *m* measured data values. The problem can be visualized since data could generally be represented as curves in the Cartesian coordinate plane (or in the space for 3D data). Each curve could be represented as a vector $\boldsymbol{q}$ with *m* components and the mathematical formulation of our problem is a problem of vector representation $\boldsymbol{Q} = \sum_{i=1}^{n} c_i \boldsymbol{q}_i$ where '*i*' is the number of measurement curves (or components), $\boldsymbol{q}_i$ is the result of a series of measurements belonging to the '*i*-th' experiment in the vector form comprising *m* data values, c_i is an unknown coefficient describing the content of the '*i*-th' component in the cumulative data and $\boldsymbol{Q}$ is the total result. In most cases we have measured $\boldsymbol{Q}$ and we (somehow) know $\boldsymbol{q}_i$ (e.g., from the theory of the problem we are analyzing) and would like to determine its contribution to the total result (determine unknown coefficients c_i). This problem could be presented in more complex variants, e.g., a constant could be added, some missing (unknown) data vector could be added to the measured summary vector, instead of constant factors c_i we could have polynomials, like $\boldsymbol{Q} = \sum_{i}^{n} (c_{i0} x^2 + c_{i1} x + c_{i2}) \boldsymbol{q}_i$, etc. All of them could be solved with slightly extended version of the procedure presented in the sequel. Our measurement problem often has some error involved in the measurement data, so the general formulation is

$$\boldsymbol{Q} = \sum_{i=1}^{n} c_i \boldsymbol{q}_i + \boldsymbol{q}_{err} \tag{1}$$

or

$$\boldsymbol{Q} = \sum_{i}^{n} (c_{i0} x^2 + c_{i1} x + c_{i2}) \boldsymbol{q}_i + \boldsymbol{q}_{err} \tag{2}$$

where $\boldsymbol{q}_{err}$ is the measurement error vector. We assume that terms in the error vector are *iid* (independent and identically distributed) with zero-mean distribution, i.e., $E[\boldsymbol{q}_{err}] = 0$, see JCGM (2008).

Solution procedure can be devised using the measurement matrix $\boldsymbol{H}$ (see e.g., Kožar 2019) and reformulating the problem into

$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{c} \tag{3}$$

where c is the vector of unknown coefficients and H is the measurement matrix constructed according to the problem, i.e., problem represented with Eq. (1) or Eq. (2), etc. For the problem described with Eq. (1) we have $H_{j,i} = (q_i)_j$. The matrix H has dimension $[m \times n], m \geq n$ and its rank should be equal to n, the number of unknown parameters. The case of rank deficiency of the measurement matrix H will not be discussed here since then the decomposition according to Eq. (1) is not possible although we could apply some regularization technique like e.g., Tikhonov regularization (see Lozzi-Kožar and Kožar 2017).

Solution procedure of the measurement problem given by Eqs. (1) or (2) could be obtained using the equation $c = H^{-1}Q$, where H^{-1} is the generalized inverse of the measurement matrix. That formulation is equivalent to application of the least squares method on the measurement problem. In the case $rank(H) = n$, H^{-1} is invertible and the vector of unknown coefficients c could be found.

It is possible to check whether the number of parameters in the model is appropriate by establishing a relationship between the number of data and parameters. We introduce the residual of the model $r = Q - Hc$. For the least squares measurement problem the expected value of the sum of squares of the residual $E(r^T r)$ is $E(r^T r) = \sigma_r^2 tr[I - H(H^T H)^{-1}H^T]$ where σ_r is variance of the residual. Applying SVD (singular value decomposition) we obtain the control relationship

$$tr[I - H(H^T H)^{-1}H^T] = m - n \tag{4}$$

In Eq. (4) m is number of data points and n is the number of parameters.

Other versions of the least squares method are possible, like Levenberg-Marquardt method (Kožar *et al*. 2018). One common way of treating a measurement error is use of the weighted least squares (Gibbs 2011) where we give different importance to different measured values. Minimization of the measurement problem described with Eq. (3) could be enhanced with introduction of the positive definite symmetric weighing matrix W into the optimization goal $J = (Q - Hc)^T W (Q - Hc)$. This changed optimization function J leads to somewhat different solution equations (derived from $\frac{\partial J}{\partial c} = 0$)

$$c = P^{-1}Q_W \qquad P = H^T W H \quad Q = H^T W Q \tag{5}$$

The matrix P^{-1} is the information or Fisher matrix and it represents the theoretical information available in the measurement since $P = E[cc^T]$ (the 'a posteriori' error covariance). Eq. (5) can outperform the Eq. (3) only with the proper formulation of the matrix W. General advice is to use the inverse covariant matrix of the measurement noise R, so $W = R^{-1}$. R is assumed to be diagonal (meaning error covariances are not correlated) since that greatly simplifies the computations but that is not a restriction. The matrix R can be formulated from the known covariance of measurement errors of given sensors. In our example errors are produced as a series of random numbers, completely uncorrelated with any function value so we do not have any 'a priori' information. Instead, we will use 'a posteriori' error analysis and correction, i.e.,

With weighted least squares check whether the number of parameters in the model is appropriate is somewhat more involved. Here we are looking at the weighted expected value of the sum of squares of the residual $E(r^T R^{-1} r)$ which is $E(r^T R^{-1} r) = tr[I - L^{-1}H(H^T L^{-T} L^{-1}H)^{-1}H^T L^{-T}]$ where L^{-1} comes from the factored covariance matrix $R = LL^T$ that is easily determined for diagonal R. Applying SVD (singular value decomposition) we obtain $L^{-1}H = USV^T$ and the control relationship is

$$tr[U_2 U_2^T] = m - n \tag{6}$$

In Eq. (6) $\boldsymbol{U} = [\boldsymbol{U}_1 \boldsymbol{U}_2]$ where $\boldsymbol{U}_1$ is $[m \times n]$ matrix and $\boldsymbol{U}_2$ is $[m \times (m - n)]$ matrix (belonging to zero eigenvalues).

In order to properly formulate the matrix $\boldsymbol{R}$ one should understand the nature of errors appearing in the measurement process.

Various numerical examples have been performed, with and without an error, and the effectiveness of the proposed method has been demonstrated. Also, we have some examples where missing data has been successfully identified.

3. Numerical examples

3.1 Least-squares examples

The first example is of the form $\boldsymbol{Q} = \sum_{i=1}^{n} c_i \boldsymbol{q}_i + \boldsymbol{q}_{err}$ with ten measurement points, m=10 and two unknown coefficients, c_1 and c_2, i.e., $n = 2$. We will apply the forward model to produce the data, add some error and then try to estimate the coefficients used in the forward model. We will make two examples with different curve shapes and the forward model data is

a) example
$\boldsymbol{q}_1 = [2.0, 3.0, 5.0, 7.0, 11.0, 13.0, 17.0, 19.0, 23.0, 29.0]$
$\boldsymbol{q}_2 = [1.841, 2.909, 3.141, 3.243, 4.041, 5.721, 7.657, 8.989, 9.412, 9.456]$
$\boldsymbol{q}_{err} = [0.042, -0.042, 0.012, 0.026, 0.029, 0.002, -0.015, -0.019, 0.015, 0.003]$
$E[\boldsymbol{q}_{err}] = 0.0053$
b) example
$\boldsymbol{q}_1 = [2.0, 1.5, 1.67, 1.75, 2.2, 2.17, 2.43, 2.38, 2.56, 2.9]$
$\boldsymbol{q}_2 = [-0.683, -1.319, 0.051, 1.764, 2.118, 0.725, -1.171, -1.854, -0.713, 1.188]$
$\boldsymbol{q}_{err} = [-0.154, -0.203, -0.372, 0.107, -0.278, -0.115, -0.386, 0.440, 0.334, 0.155]$
$E[\boldsymbol{q}_{err}] = -0.04718$

In the first example the first vector is formed from prime numbers, the second is sampled from the function $k + \sin(k); k = [1, \ldots, n]$ and the error vector is made from uniformly distributed random numbers with the zero mean; as it is evident, the mean is not exactly zero. In the second example the first vector is formed from prime numbers decided by their position, the second is sampled from the function $\frac{1}{k} + 2\sin(k); k = [1, \ldots, n]$ and the error vector is made from uniformly distributed random numbers with the zero mean but, as before, the mean is not exactly zero.

Note: The mean could be made equal to zero by subtracting it from all elements in a vector. However, it did not appeared to be significant for the examples. If there are zeros in data vectors, it is best to leave out that data since it reduces accuracy.

The cumulative vector (including an error) is now $\boldsymbol{Q} = c_1 \boldsymbol{q}_1 + c_2 \boldsymbol{q}_2 + \boldsymbol{q}_{err}$ and for $c_1 = 1.23$ and $c_2 = -0.56$ it follows (rounded to 3 decimal figures)

a) example
$\boldsymbol{Q} = [1.471, 2.019, 4.403, 6.820, 11.296, 12.788, 16.607, 18.317, 23.034, 30.378]$
b) example
$\boldsymbol{Q} = [2.689, 2.381, 1.654, 1.272, 1.242, 2.148, 3.258, 4.406, 3.882, 3.056]$

Graphical representation of data vectors is given in Fig.1 where $\boldsymbol{Q}0 = c_1 \boldsymbol{q}_1 + c_2 \boldsymbol{q}_2$ is the cumulative vector without an error and $\boldsymbol{Q} = \boldsymbol{Q}0 + \boldsymbol{q}_{err}$ is the cumulative vector with the error; one

could see the influence of data vectors q_1 and q_2 represented graphically. Fig. 1 also enables assessment of the magnitude of the error vector, i.e., comparison of vectors Q and $Q0$. It should be clear that recovery of the exact coefficients c_1 and c_2 is not possible when the error vector is too large (introduction of an error vector, i.e., replacement of the vector Q with $Q0$ changes the original problem).

<table>
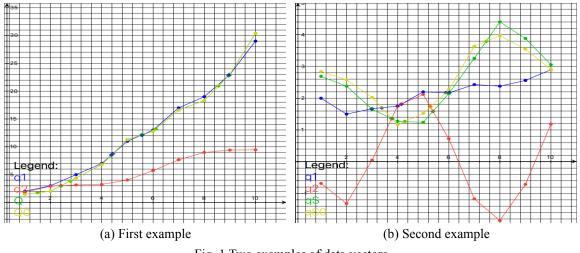<tr><td>(a) First example</td><td>(b) Second example</td></tr>
</table>

Fig. 1 Two examples of data vectors

"Unknown" coefficients $[c_1, c_2]$ are estimated using the measuring equation $c = H^{-1}Q$ where H is the measuring matrix $[q_1, q_2]^T$ and $(\cdot)^{-1}$ represents the generalized inverse of a matrix. Solution of the measuring equation gives the estimate (the exact value is $c = [c_1, c_2] = [1.23, -0.56]$)

a) example
$c = H^{-1}Q = [c_{e1} \quad c_{e2}] = [1.2319, -0.5643]$
b) example
$c = H^{-1}Q = [c_{e1} \quad c_{e2}] = [1.2200, -0.5919]$
Residual $r_{LS} = Q - c_{e1}q_1 - c_{e2}q_2$ resulting from this estimate is (in %)
a) example
$r_{LS} = [-3.13, 1.74, -0.36, -0.39, -0.23, -0.02, 0.08, 0.09, -0.05, 0.03]$
b) example
$r_{LS} = [5.78, 9.65, 21.36, -14.23, 15.17, 3.28, 12.27, -9.19, -8.67, -7.23]$
Residual in the second example is noticeably larger because the error vector much has larger mean, -0.04718 compared to 0.0053, which is almost 9 times larger in absolute value.
Note that residual is calculated in relation to the measurement with error and the residual related to the exact (unknown) measurement $Q_0 = c_1q_1 + c_2q_2$ is significantly smaller (in %)
a) example
$r0_{LS} = [-0.29, -0.09, -0.01, 0.03, -0.002, -0.006, -0.006, -0.02, 0.01, 0.04]$
b) example
$r0_{LS} = [0.064, 1.048, -0.903, -6.326, -5.885, -1.977, 0.359, 0.892, -0.079, -2.302]$

In practice this residual can not be calculated since the exact measurement Q_0 is unknown. This confirms the previous statement that we are searching for the projection of the measurement vector in the component space.

3.2 Weighted least-squares examples

We are solving the same examples as before with the same q_1, q_2, q_{err} only this time with a weighing matrix R following the weighted least squares procedure.

a) example

We have adapted the covariance matrix to the error vector, which we could not do in reality since q_{err} is unknown. Absolute error in percent relative to the measured value Q is

$q_{per} = [2.86, 2.08, 0.27, 0.38, 0.26, 0.02, 0.09, 0.1, 0.07, 0.01]$

$R = diag[q_{per}]$

$P^{-1} = (H^T W H)^{-1} = \begin{bmatrix} 5.1717E - 06 & -1.4505E - 05 \\ -1.4505E - 05 & 4.1427E - 05 \end{bmatrix}$

$Q = H^T W Q = [11128283.9 \quad 3882869.3]$

$c = P^{-1} Q_W = [c_{w1}, c_{w2}] = [1.2308, -0.5619]$

For this example with small error we have obtained only slightly better results then with the least squares procedure.

b) example

Here, the absolute error in percent relative to the measured value Q is much larger (over 20% for measuring points)

$q_{per} = [5.72, 8.53, 22.48, 8.44, 22.36, 5.35, 11.85, 10.0, 8.59, 5.06]$

$R = diag[q_{per}]$

$P^{-1} = (H^T W H)^{-1} = \begin{bmatrix} 0.001662 & -0.0001787 \\ -0.0001787 & 0.005731 \end{bmatrix}$

$Q = H^T W Q = [736.407 \quad -75.4468]$

$c = P^{-1} Q_W = [c_{w1}, c_{w2}] = [1.2375, -0.5640]$

For this example with large error we have obtained much better results then with the least squares

(a) First example   (b) Second example
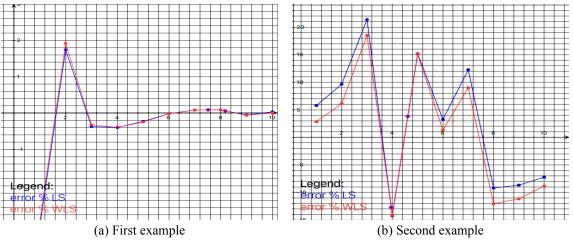
Fig. 2 Residuals for two examples

procedure. Residual (in %) $r_{LS} = Q - c_{w1}q_1 - c_{w2}q_2$ resulting from this estimate is much smaller than for pure least squares.

3.3 Polynomial example

Here the example is of the form $Q = \sum_i^m [(c_{11}x_i + c_{12})q_1 + c_{20}q_2]$ with fifteen measurement points, m=15 and three unknown coefficients, c_{11}, c_{12} and c_{20}, i.e., $n = 3$. We will apply the forward model to produce the data, add some error and then try to estimate the coefficients used in the forward model. Parameters for the forward model are

$c_{11} = -0.2222;$ $c_{12} = 3.2222;$ $c_{20} = -0.56$
$q_1 = [2.0, 1.5, 1.67, 1.75, 2.2, 2.17, 2.43, 2.38, 2.56, 2.9]$
$q_2 = [-0.683, -1.319, 0.051, 1.764, 2.118, 0.725, -1.171, -1.854, -0.713, 1.188]$
$q_{err} = [-0.154, -0.203, -0.372, 0.107, -0.278, -0.115, -0.386, 0.440, 0.334, 0.155]$
$E[q_{err}] = -0.3159$
$Q = [5.746, 6.319, 10.30, 13.74, 20.31, 20.67, 23.69, 22.12, 23.03, 23.36, 18.59, 13.16, 6.209,$
     $-3.510, -14.76]$

Graphical representation of data vectors is given in Fig. 3 where $Q0$ is the cumulative vector without an error and $Q = Q0 + q_{err}$ is the cumulative vector with the error; one could see the influence of data vectors q_1 and q_2 represented graphically and asses the error introduced with the error vector q_{err}.

Solution (the exact coefficients are $c = [c_{11}, c_{12}, c_{20}] = [-0.2222, 3.2222, -0.5600])$

$c = H^{-1}Q = [c_{e11} \quad c_{e12} \quad c_{e20}] = [-0.2198, 3.1784, -0.5468]$. The errors in percent are [-1.08%, -1.36%, -2.4%] respectively. Here, the measurement matrix H is $[15 \times 3]$ and comprises the Vandermonde matrix V where $V_{ij} = x_i^{j-1}$ and data vectors $H_{j,i} = [V_{j1}q_{j1}, (q_i)_j]$.
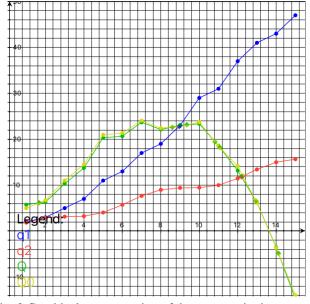

Fig. 3 Graphical representation of data vectors in the example

3. Discussion

Error vector is not known in advance and cannot be directly included into the solution procedure. Introduction of an error vector changes the original problem and the recovery of the exact value of the model coefficients is not possible when the error vector is too large. We have obtained very good results with weighted least squares procedure only because we have assumed some knowledge about the error vector. One could use various assumptions about error stochastic properties (e.g., the error is represented as a random filed with a stationary mean) to improve the procedure for recovery of the unknown coefficients. There are many possible procedures for error treatment: one could use weighted least squares (Gibbs 2011), some 'a priori' assumptions like in Ibrahimbegović *et al.* (2020), Sarfaraz *et al.* (2018) or some 'a posteriori' assumptions. However, improvements in error handling remain open for further investigation.

One should mention that nowadays neural networks could be used for the problems presented here as they are capable of computing any function (Nielsen 2015). However, solving for presence of known functions in another function is better suited for the approach based on the least squares method as given in the paper.

4. Conclusions

Data vectors are a common result of measurements on engineering structures. Often we old like to know the relationships between various data vectors, i.e., perform a factor analysis. We have demonstrated how one could perform factor analysis on data vectors corrupted with some sensor errors by transforming the formulation into a measurement problem. This greatly simplifies the problem and enables the use of least squares methods, like least squares or weighted least squares. The most important factor in weighted least squares is a proper determination of the error covariance matrix. Results obtained in practical examples are satisfactory for practical engineering applications.

Acknowledgments

References

Gibbs, B.P. (2011), *Advanced Kalman Filtering, Least-Squares and Modeling*, Wiley, Hoboken, New Jersey.
Ibrahimbegovic, A., Hervé, G. and Villon, P. (2009), "Nonlinear impact dynamics and field transfer suitable for parametric design studies", *Eng. Comput.*, **26**(1/2), 185-204. https://doi.org/10.1108/02644400910924861.
Ibrahimbegović, A., Mathies, H.G. and Karavelić, E. (2020), "Reduced model of macro-scale stochastic plasticity identification by Bayesian inference in application to quasi-brittle failure of concrete", *Comput. Meth. Appl. Mech. Eng.*, **372**, 113428. https://doi.org/10.1016/j.cma.2020.113428.
JCGM 100 (2008), Evaluation of Measurement Data-Guide to the Expression of Uncertainty in Measurement,

Joint Committee for Guides in Metrology.

Kožar, I. (2016), "Relating structure and model", *Computational Methods for Solids and Fluids, Multiscale Analysis, Probability Aspects and Model Reduction*, Ed. A. Ibrahimbegovic, Springer, New York, USA,

Kožar, I. and Lozzi-Kožar, D. (2017), "Flux determination using finite elements: global vs. local calculation", *Technical Gazette*, **24**, 247-252. https://doi.org/10.17559/TV-20160208123711.

Kožar, I., Lozzi Kožar, D. and Jeričević, Ž. (2011), "Limit kriging in finite element environmental modeling", *International Convention on Information, Communication and Electronic Technology* MIPRO, Opatija, Croatia.

Kožar, I., Torić Malić, N. and Rukavina, T. (2018), "Inverse model for pullout determination of steel fibers", *Couple. Syst. Mech.*, **7**, 197-209. https://doi.org/10.12989/csm.2018.7.2.197.

Kožar, I., Torić Malić, N., Simonetti, D. and Božić, Ž. (2020), "Stochastic properties of bond-slip parameters at fibre pull-out", *Eng. Fail. Anal.*, **111**, 104478. https://doi.org/10.1016/j.engfailanal.2020.104478.

Lozzi-Kožar, D. and Kožar, I. (2017), "Estimation of the eddy thermal conductivity for Lake Botonega", *Eng. Rev.*, **37**, 322-334.

Menke, W. (2012), *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press Elsevier, Cambridge, Massachusetts, USA.

Nielsen, M.A. (2015), *Neural Networks and Deep Learning*, Determination Press.

Sarfaraz, S.M., Rosic, B.V, Matthies, H.G. and Ibrahimbegovic, A. (2018), "Stochastic upscaling via linear bayesian updating", *Couple. Syst. Mech.*, **7**, 211-231. http://doi.org/10.12989/csm.2018.7.2.211.