FORECASTING OF SUBJECTIVE COMFORT IN TRAM USING ORDINAL LOGISTIC REGRESSION AND MANIFOLD LEARNING

Jacek Pietraszek, Witold Grzegożek, Jarosław Szczygieł

Cracow University of Technology
Department of Mechanical Engineering
Jana Pawła II Av. 37, 31-864 Krakow, Poland
tel.: +48 12 6283580
e-mail: pmpietra@mech.pk.edu.pl
witek@mech.pk.edu.pl, jarek.szczygiel@gmail.com

Abstract

Comfort in a vehicle has a very important role to play as one of the most important dynamic performance characteristics of rail vehicles. It is the factor of ever-increasing importance, even creating a specialized branch of engineering associated with relation between human limitations and designing of machines: human–factors engineering. The vibration is known to be a major factor that affects and deteriorates ride comfort. For evaluating ride comfort in rail vehicles, there have been developed methods resulting in the creation of many standards and multiple criteria used and even standardized in different countries. One of the authors, J. Szczygieł designed and performed a passive experiment to collect data describing physical conditions of ride and associated subjective assessments of comfort. Panel of fourteen people during the tram ride made synchronous subjective assessments of comfort, assessing it on a discrete ordinal scale of 0 to 5, using electronic panels connected to the computer. At the same time computer through sensors recorded values of acceleration in three perpendicular axes. It made possible to correlate the fuzzy subjective evaluations with objective physical measurements. Because of the discrete type of fuzzy ratings of comfort, natural way of modelling is the ordinal logistic regression. The classic form of the ordinal logistic regression assumes that in the space of explanatory factors there are parallel activation hyper-planes slightly disturbed by unknown or uncontrolled noise factors. In fact, the assumption of linearity is a very strong idealization and leads to considerable misclassifications. The original space of explanatory factors is 11-dimensional with ten continuous dimensions and one discrete. Then the multivariate method, principal component analysis (PCA), was used to identify principal components, which are responsible most to the variability of the studied set. The scree plot was used to identify the number of significant PCA factors. The use of PCA revealed that the area occupied by the data set is approximately 6-dimensional. However, the dimensionality reduction of explanatory variables set did not lead to better forecasting accuracy. A more subtle analysis involving discretization techniques showed that activation hyper-planes are highly curved in the six-dimensional area identified by PCA but their dimensionality is much lower. The details of the procedure are described in the article. The article conclusion is that is necessary to introduce curvilinear coordinate system embedded into the shapes of activation hyper-planes to obtain better classification.

Keywords: *rail transport, vibrations, ordinal logistic regression, principal component analysis, manifold learning*

1. Introduction

Passenger comfort in a vehicle has very important role to play as one of the most significant dynamic performance characteristics of rail vehicles. The ride comfort is determined by many factors, inner (related to particular person) and outer (related to vehicle environment and ride dynamic). The factor recognized with particularly importance is vibration. The human response to vibration is highly variable and depends on magnitude, frequencies, direction and duration of vibrations.

ISO Standard 2631 [11] and British Standard 6841 [3] precisely define procedures for prediction of vibration discomfort basing on measured vibration at the seat pan, the seat back and the feet of seated person. The standards use RMS (root mean square) of measured acceleration of vibration.

Experimental studies [9] revealed that doubling vibration magnitude requires sixteen-fold reduction of duration to maintain equivalence comfort feeling. It led to introduction of RMQ (root mean quadruple) of measured acceleration of vibration. Both RMS and RMQ are averages of acceleration over time interval. It may lead to false prediction of discomfort if the vibration signal is not stationery. The remedy for this problem is Vibration Dose Value [2, 11], which is related to RMQ but not averaged over time. It means that VDV measures cumulative dose of vibration. Other approaches to comfort assessment base on completely three-dimensional vector of vibration acceleration like e.g. CEN ENV12299 [4] introducing NMV measure.

BS 6841 and ISO 2631 standards introduce discomfort scale defined as a set of overlapped classes of RMS named: not uncomfortable (less than 0.315), a little uncomfortable (0.315-0.63), fairly uncomfortable (0.5-1.0), uncomfortable (0.8-1.6), very uncomfortable (1.25-2.5), extremely uncomfortable (greater than 2.0). It clearly leads to fuzzy assessment. Analogically, CEN ENV 12299 defines its discomfort scale as a set of non-overlapped classes of its measure NMV named: very comfortable (less than 1.5), comfortable (1.5-2.5), medium (2.5-3.5), uncomfortable (3.5-4.5), very uncomfortable (greater than 4.5).

Regardless of the physical factors, objectively measurable, it remains to consider the subjective feeling of a ride comfort. The issue of subjective feeling of comfort came to be widely seen in recent years. Nausea and discomfort in high-speed tilting trains was studied by Forstberg since 1996 [6-8]. In addition to vibration measurements, Lee, Shin, Song, Han and Lee [14] studied biological parameters: heart rate and blood pressure, testing of volunteers on the specially constructed tilting train simulator. Recently Scherer [16] considered various subjective and objective reasons why the light train ride is more attractive than the bus. Long, Wei, Shi and Wang [15] have developed a method to evaluate a ride comfort in the context of the track alignment for high-speed railways. Cheng [5] applied a conceptual model based on the railway passenger service chain perspective to study passengers anxieties associated with train travel measured using a modern psychometric method: the Rasch's model. Um, Choi, Yang and Kim [17] studied the relationship between driving comfort and superimposed horizontal and vertical curves in the case of railway construction/renovation.

Grzegożek, Szczygieł and Król [10] began in 2009 the study on rail vehicles with particular emphasis on the impact of vibration on comfort. Some problems, which appeared during model's identification, are described later in this article.

2. Materials and methods

2.1. Experimental unit

A passive experiment was designed and performed in year 2011 by J. Szczygieł to collect data describing physical conditions of ride and associated subjective assessment of comfort. The tram model NGT6 produced by Bombardier was provided by MPK Krakow. Accelerations in three perpendicular axes X, Y, Z were measure by three accelerometers model B12/200 produced by Hottinger Baldwin Messtechnik (HBM). Sensors are attached to the floor under the passenger seats. Signals were received and a/d converted by a universal amplifier model Spider 8 produced by HBM. Measuring circuit is controlled by software CatMan. Push-button panels were designed and made by J. Szczygieł as voltage dividers.

Fourteen people were driving in a tram and synchronously evaluating ride comfort using push-button panels. The comfort was evaluated on the discrete ordinal scale of 0 to 5. At the same time on-board computer were recording through sensors values of acceleration in three perpendicular axes. The route was entirely located in Krakow and it led through the streets of Old Town as well as longer distances between the Old Town and the district of Nowa Huta. This route allowed performing measurements on tracks of different standards.

2.2. Mathematical model and methods

Collected multivariate measures form a set of point labelled with chosen discomfort level and associated count. The main goal is discrimination: use collected data to construct a classifier separating predefined classes. The main difficult is overlapping of class boundaries due to noise and individual preferences in comfort assessment.

Measured physical factors (directly accelerations, categorical type of track section and indirectly seven others calculated from accelerations in longer time window: RMS, RMQ, VDV etc.) are associated with human stress level. It is assumed that stress level SL is described as a scalar number and it is a function of stress factors x_i. If the stress level exceeds a certain critical value SL_i, the individual expresses this by getting higher scores y_i to discomfort:

$$y = y_i \Leftrightarrow \left(\left(SL_i \le SL \right) \wedge \left(SL < SL_{i+1} \right) \right). \tag{1}$$

Typically, it is assumed that the stress level is a simple linear function of stress factors:

$$SL\left(\mathbf{x} \right) = \boldsymbol{\beta}^T \mathbf{x} \,. \tag{2}$$

where stress factors x_j are consolidated into one vector $\mathbf{x}$ with respective number of components. That formulation of the problem belongs to the linear discrimination analysis (LDA) [12].

The main difficulty of the considered modelling is that measured subjective responses are discrete, particularly binary: y_i selected (1) or y_i not selected (0). A solution proposed by categorical data analysis [1] leads to the model returning the probability of particular response, not the response itself like in classical regression. It means that the model has the form:

$$\mathrm{P}\left(y_i \left| \mathbf{x} \right. \right) = \Phi\left(\alpha_i + \boldsymbol{\beta}^T \mathbf{x} \right), \tag{3}$$

where Φ is a link function associated with generalized linear model and α_i constant terms are associated with critical values SL_i. The particular form of link function depends on the random distribution associated with noise factors and individual characteristics of tested persons e.g. normal distribution leads to probit link function, binomial distribution leads to logit link function [12]:

$$\Phi\left(y_i \left| \mathbf{x} \right. \right) = \frac{e^{\alpha_i + \boldsymbol{\beta}^T \mathbf{x}}}{1 + e^{\alpha_i + \boldsymbol{\beta}^T \mathbf{x}}}. \tag{4}$$

The model (3) is known under the name of ordinal logistic regression.

In the case of model (3) the only difference between forms for difference scores are the constant terms α_i. It is required due to the lack of constraints associated with stress factors. In geometric approach, it means that hyper-planes associated with scores and located in stress factors space are parallel to each other. This corresponds to the condition that the order of scores must be maintained.

In the case of the local modelling, this condition need not be met. It means that the model (3) could be generalized to the form:

$$\mathrm{P}\left(y_i \left| \mathbf{x} \right. \right) = \Phi\left(\alpha_i + \boldsymbol{\beta}_i^T \mathbf{x} \right), \tag{4}$$

where each score could have individual coefficient vector $\boldsymbol{\beta}_i$. The only condition that must be satisfied is that the hyper-planes cannot intersect in the area of modelling. Additionally, authors propose to include in the model (4) two-way interactions between stress factors. This will increase the number of vector $\mathbf{x}$ components.

In general, the model (4) is an example of latent-variable model [1] because the stress level and modelled probability are not observable. The only measureable variables are stress factors and categorical scores. The constant terms a_i and coefficient vectors $\boldsymbol{\beta}_i$ have to be calculated based on

maximum likelihood estimation associated with assumed probability distribution and iterative methods.

High dimensional data usually contain highly correlated factors. It may introduce collinearity into discrimination problem and dimensionality reduction should be considered. In this problem high dimensionality appears: 10 continuous and 1 categorical variable of stress factors space. The principal component analysis (PCA) [13] as an unsupervised procedure for linear dimensionality reduction was applied.

Additionally, the analysis of discretization was performed to detect more subtle structure of support data. Discretization analysis is a simple method [12, 13], which bases on the detection of effective exponent at increasingly finer discretization of considered factors space. For every k-dimensional simple-connected object embedded into discretized n-dimensional space, ($k \leq n$) increasing discretization by factor p results in increasing occupied cells by factor p^k. Comparison of occupied cell numbers for different discretization (eq.5) reveals dimensionality of the object k, which may be observed at this level of details.

$$\frac{V_2}{V_1} = \left(\frac{p_2}{p_1}\right)^k \quad \Rightarrow \quad k = \ln\left(\frac{V_2}{V_1}\right) \Big/ \ln\left(\frac{p_2}{p_1}\right). \tag{5}$$

It should emphasized that such dimensionality may be fractional because different parts of data set may realize simple-connectedness at different levels of dimensionality e.g. one part as two-dimensional surface and different part as one-dimensional curve. As if this were not enough, sufficiently small discretization can reveal the holes in the data set structure and thus the area can no longer be simple-connected. Regardless of these difficulties of interpretation, this method is a valuable tool to detect the subtleties of the structure lie outside the scope of the PCA and other LDA methods.

3. Results

3.1. Principal component analysis

Belonging to LDA group principal component analysis [13] was carried out on all 10 continuous stress factors. It revealed that the problem might be practically treated as with 6 dimensions. The remaining 4 dimensions explain only 1.12% of variability. It is clearly showed in Fig.1.

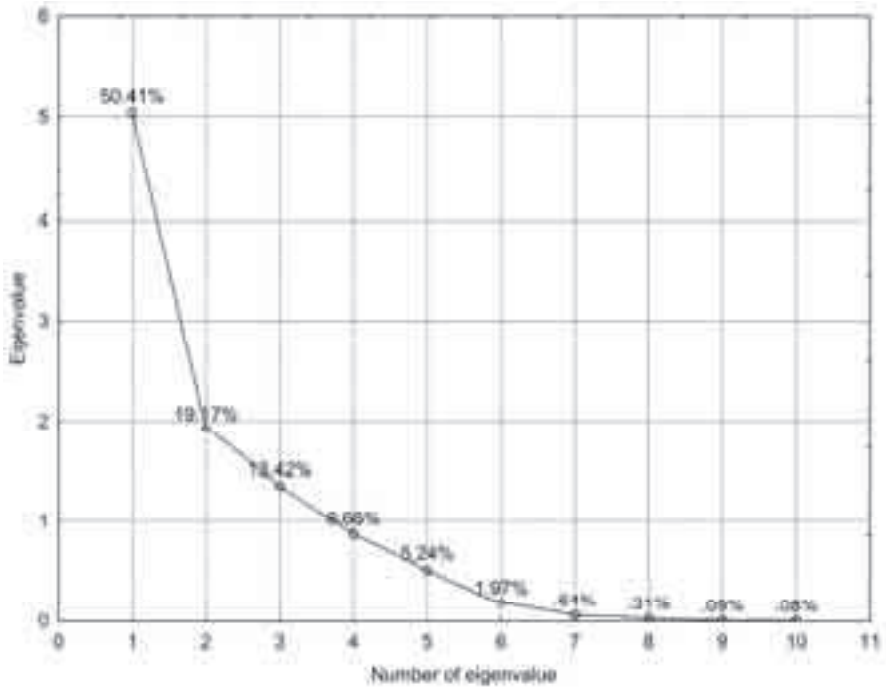

Fig. 1. Scree plot for PCA carried out on 10 continuous stress factors

Tab. 1. PCA coefficient for physical stress factors

Stress factors	PCA factors					
	1	2	3	4	5	6
A	-0.98545	-0.05950	-0.01234	-0.02955	0.00021	0.00326
B	-0.44192	0.04165	-0.11557	0.88702	0.01425	0.00270
C	-0.98949	-0.04997	-0.01371	-0.02044	-0.00007	-0.00035
D	-0.98583	-0.04437	-0.01816	0.01571	0.00010	-0.00205
E	-0.96387	-0.06632	0.00582	-0.20941	0.00061	0.00679
F	-0.97598	-0.06318	-0.00060	-0.15862	-0.00045	0.00622
G	0.01392	-0.49653	0.69463	0.08309	-0.51363	-0.00598
H	0.11958	-0.87157	-0.35258	0.00001	0.05425	-0.31433
I	0.03539	-0.51489	0.68779	0.05851	0.50273	0.06647
J	0.16365	-0.79220	-0.49821	-0.01067	-0.06379	0.30533

It may be interpreted that data set in stress factors space occupies 6-dimension sub-space and values of physical 10 continuous stress factors belonging to data set may be linearly expressed by 6 variables with no physical meaning. The coefficients of these expressions are shown in Tab. 1.

It should emphasize that this does not mean that data set is 6-dimensional itself. A simple explanation may be one-dimensional spiral located in two-dimensional plane embedded in three-dimensional space. In this example PCA detects only two-dimensional plane but does not detect more subtle one-dimensional spiral structure.

3.2. Discretization analysis

The discretization of the stress factors space was carried out for the range from 2 to 24 intervals on each of dimension range. The evaluated dimensionality (eq. 5) is shown in Tab. 2.

Tab. 2. Dimensionality of the data set support

No of intvls	Hyper-voxels	*k*-dim	No of intvls	Hyper-voxels	*k*-dim
2	130	7.02	14	3675	0.24
3	472	3.18	15	3808	0.52
4	953	2.44	16	3994	0.74
5	1403	1.73	17	4139	0.59
6	1770	1.27	18	4275	0.57
7	2198	1.40	19	4412	0.58
8	2495	0.95	20	4423	0.05
9	2797	0.97	21	4511	0.40
10	3021	0.73	22	4574	0.30
11	3248	0.76	23	4695	0.59
12	3434	0.64	24	4818	0.61
13	3611	0.63	–	–	–

3.3. Ordinal logistic regression

J. Szczygieł conducted the experiment and collected raw data. After pre-processing of raw data with various method including wavelets [10], he obtained 6341 data point located in 11-dimensional factors space: 10-dimensional sub-space of continuous factors and 1-dimensional categorical factor (type of track section).

First attempt to identification of logistic discrimination classifier with simple linear model (eq.3) has led to too large misclassification probability. The second attempt has used linear model with two-way interactions (eq.4) between stress factors. The calculations have shown large number of statistically insignificant terms of the model: they were sequentially eliminated during BE stepwise regression [12]. Also noteworthy is the large number of main effects eliminated. The identification was made separately for each of score categories and resulted formulas were quite different. It means that activation surfaces in stress factors space have different, very complicated shapes and – in opposition to typical LDA ordinal logistic classification – they are not parallel.

The second attempt has significantly reduced misclassification probability. Unfortunately, due to limited space of this article, it is not possible to present complete identified models. The initial model has 67 terms: constant term, linear main effects and two-way interactions.

4. Discussion

The whole stress factors space is formally 11-dimensional, but due to categorical character of type of track section, it should be decomposed into five separate 10-dimensional subspaces. Application of PCA to the data set revealed that the area occupied by the data points is a 6-dimensional subspace.

Analysing of discretization sequence provided very serious suspicions that the actual data set support has a lower dimension with a very complex shape, so folded that it occupies 6-dimensional subspace. The course of the *k*-dimensional (Tab.2) shows that the cloud of data points has a relatively dense core surrounded by a less dense halo of single points. A clear decrease in *k*-dimensional (Tab.2) at two values of intervals: 14 and 20 is clearly associated with detection of a large number of isolated groups of hyper-voxels with small number of data points whose dimension (at this level) is zero, i.e. like a point.

An attempt to create a classifier based on logistic regression analysis was successful, when the model was augmented by two-way interactions. The misclassification was lowered to acceptable level however still large. This fact, combined with information obtained from the PCA and discretization analysis can formulate the notion that the data set support has the highly complicated shape and it is not simply a trivial subset of the stress factors space.

To get higher accuracy of the classifier (e.g. logistic regression), it is desirable to introduce a curvilinear coordinate system (similar to the linear arrangement of the eigenvectors in PCA), consistent with the shape of the data set support. There is a high probability that the classifier built in such a coordinate system will have a higher accuracy i.e. lower number of misclassifications.

5. Conclusion

A passive experiment to collect data about subjective level of comfort was performed during a tram ride. The data set was analysed with statistical multivariate methods and the classifier based on logistic regression was built. The obtained information revealed that the shape of the data set support, originally identified by PCA as simple-connected 6-dimensional subspace of 11-dimensional stress factors space, is far more complicated: its dimensionality is lower but folding leads to high curvature. Due to this fact, the logistic regression classifier, even augmented with two-way interactions, still has a large number of misclassifications. To obtain better forecasting accuracy two possible approaches seems possible:

- parametric modelling with curvilinear coordinate system embedded into the shape of activation hyper-planes,
- non-parametric modelling based on manifold learning or, perhaps, empirical likelihood estimating.

References

[1] Agresti, A., *Categorical Data Analysis*, John Wiley & Sons, Hoboken, NJ 2002.

[2] BS6472-1, *Guide to evaluation of human exposure to vibration in buildings. Vibration sources other than blasting*, BSI, London 2008.

[3] BS6841, *Guide to measurement and evaluation of human exposure to whole-body mechanical vibration and repeated shoc*, BSI, London 1987.

[4] CEN_ENV12299, *Railway applications - Ride comfort for passengers - Measurement and evaluation*, CEN-CENELEC, Brussels 2009.

[5] Cheng, Y.H., *Exploring passenger anxiety associated with train travel*, Transportation, Vol. 37, 875-896, 2010.

[6] Forstberg, J., *Nausea and comfort in tilting trains: Possible regression models for nausea*, Taylor & Francis Ltd, London 2003.

[7] Forstberg, J., et al., *Influence of different conditions of tilt compensation on motion and motion-related discomfort in high speed trains*, Vehicle Syst Dyn, Vol. 29, 729-734, 1998.

[8] Forstberg, J., et al., *Influence of different compensation strategies on comfort in tilting high speed trains*, In: Claussen, C.F., et al. (eds.) Giddiness & Vestibulo-Spinal Investigations; Combined Audio-Vestibular Investigations; Experimental Neurootology, Vol. 1133, pp. 325-331, Elsevier Science Publ B V, Amsterdam 1996.

[9] Griffin, M. J., *Discomfort from feeling vehicle vibration*, Vehicle Syst Dyn, Vol. 45, 679-698, 2007.

[10] Grzegożek, W., et al., *An Attempt of an Employment of a Continuous Wavelet Transform for Evaluation of Temporary Comfort Distrubances*, Journal of KONES Powertrain and Transport, Vol. 16, 165-172, 2009.

[11] ISO2631, *Mechanical vibration and shock - Evaluation of human exposure to whole-body vibration* - Part 1: General requirements, ISO, Geneve 1997.

[12] Izenman, A. J., *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning.*, Springer Science+Business Media, LLC, New York 2008.

[13] Jolliffe, I. T., *Principal Component Analysis*, Springer-Verlag Inc., New York 2010.

[14] Lee, Y., et al., *Research of Ride Comfort for Tilting Train Simulator Using ECG*, In: VanderSloten, J., et al. (eds.) 4th European Conference of the International Federation for Medical and Biological Engineering, vol. 22, pp. 1906-1909, Springer New York 2009.

[15] Long, X. Y., et al., *Dynamic Analysis and Ride Comfort Evaluation of Track Alignment for High-Speed Railway*, China Railway Publishing House, Beijing 2010.

[16] Scherer, M., *Is Light Rail More Attractive to Users Than Bus Transit? Arguments Based on Cognition and Rational Choice*, Transp. Res. Record, Vol. 2144, 11-19, 2010.

[17] Um, J. H., et al., *Optimization of alignment considering ride comfort for superimposition of vertical and horizontal curves*, P I Mech Eng F-J Rai, Vol. 225, 649-662, 2011.