Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 6 (2011) 352–357

Procedia Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

Correspondence analysis for symbolic contingency tables based on interval algebra

Ikufumi Takagi[a,*], Hiroshi Yadohisa[b]

[a]*Graduate School of Culture and Information Science, Doshisha University, Kyoto, 610-0394, Japan.*
[b]*Department of Culture and Information Science, Doshisha University, Kyoto, 610-0394, Japan.*

Abstract

In this paper, we propose interval algebraic correspondence analysis (IACA), a new correspondence analysis method for interval contingency tables based on interval algebra. The interval contingency table, which is made by counting up the observations measured by two multi-valued variables, is an extension of the classical contingency table.

Correspondence analysis for the interval contingency table has been proposed by Rodríguez[8] (SymCA); this analysis is based on the centers method in principal component analysis for the interval variables (Cazes, et al.,[2]). However, his method has the disadvantage that when computing statistical indices, the internal variation of intervals is lost. To overcome this problem, we propose a new correspondence analysis through which the internal variation of the interval is retained. A numerical example using IACA is discussed and the usefulness is shown.

Keywords: symbolic data analysis, large and complex data, contingency table analysis, multi-valued variables, interval contingency table

1. Introduction

Symbolic data analysis (SDA) is a method to analyze large and complex data (Bock and Diday[1], Diday and Noirhomme[4]). In SDA, we assume that a cell of the data table, which is often observed by n individuals and p variables or by $n \times m$ modalities of two variables, is not a single description but a complex description. The data table is called a symbolic data table. The main study objectives of SDA is to develop data analysis methods for symbolic data tables.

An interval contingency table is a type of symbolic data table. In classical data analysis, when we obtain observations measured by two categorical single variables, we make a contingency table whose cell is described as a single numerical value (classical contingency table). Then, we generally explain the relations between the modalities of the two variables using the results of a chi-square test or by performing correspondence analysis (CA), depending

*Corresponding author
Email addresses: dik0010@mail4.doshisha.ac.jp (Ikufumi Takagi), hyadohis@mail.doshisha.ac.jp (Hiroshi Yadohisa)

doi:10.1016/j.procs.2011.08.065

on the research objectives. However, when we obtain the observations measured by two multi-valued variables, we typically encounter difficulties in constructing the table. This is because we do not know the general way to treat such observations. To treat multi-valued variables, Rodríguez[8] has proposed an interval contingency table whose cell is represented as an interval.

A cell described as an interval makes it possible to represent the degree of the observations belonging to it. Besides, we can also treat the uncertainty and the measurement errors by using such cells.

Recently, the data analysis methods for an interval contingency table have been studied. One of the surveys for the methods is Symbolic Correspondence Analysis (SymCA), which is the extension of classical CA (Rodríguez[8]). The purpose of SymCA is the visualization and comprehension of the relations between the two modalities of the two variables in low-dimensional space, as well as classical CA. Unlike classical CA, SymCA reflects the internal variation of the observation because the result is described as an interval value. However, SymCA sums up an observation into a single value once when calculating the statistical indices. It is important to note that SymCA cannot reflect the internal variation of the observation completely.

In this paper, we propose the interval algebraic correspondence analysis (IACA), which is based on interval algebra. By calculating the statistical indices based on the interval algebra, we can obtain values that are described as interval values. Thus, the proposing the analysis methods based on interval algebra would overcomes the problems. Studies on the SDA method based on interval algebra have been reported in the past, e.g., Gioia and Lauro[5].

First, we introduce the concept of interval algebra. Next, we define an interval contingency table and show some situations for obtaining the table in detail. Finally, we describe the IACA and present a numerical example and interpretations of the results.

2. Interval algebra

We introduce the basic concept of interval algebra. The statistical indices based on this algebra make it possible to retain the internal variation for the interval. Interval algebra is discussed in the literature, Moore[6] and Neumaier[7].

Definition 2.1. *interval, interval matrix*
Let $x^L, x^U \in \mathbb{R}$ with $x^L \leq x^U$. First, an interval x^I is defined as follows:

$$x^I =: [x^L,\ x^U] = \{x |\ x^L \leq x \leq x^U\},$$

and let $\mathcal{I}$ be the set of all intervals. Second, an interval matrix $\boldsymbol{X}^I$ is defined as follows:

$$\boldsymbol{X}^I =: [\boldsymbol{X}^L,\ \boldsymbol{X}^U] = \{\boldsymbol{X} |\ \boldsymbol{X}^L \leq \boldsymbol{X} \leq \boldsymbol{X}^U\},$$

where $\boldsymbol{X}^L (= (x_{ij}^L))$ and $\boldsymbol{X}^U (= (x_{ij}^U))$ are $(n \times p)$ matrices, and let $\mathcal{M}_{np}$ be the set of all $(n \times p)$ interval matrices.

We will define the arithmetic operations on $\mathcal{I}$ by extending the arithmetic operations on $\mathbb{R}$.

Definition 2.2. *interval arithmetic*
Let $x^I,\ y^I \in \mathcal{I}$. An arithmetic operation $\bullet \in \{+,\ -,\ \times,\ /\}$ on $\mathcal{I}$ is defined as follows:

$$x^I \bullet y^I = \{x \bullet y |\ x \in x^I,\ y \in y^I\},$$

in case that $0 \in y^I$, we do not define x^I / y^I.

Proposition 2.1.
Let $x^I,\ y^I \in \mathcal{I}$. Then, Definition2.2 is represented as follows:

$$\begin{aligned}
x^I + y^I &= [x^L + y^L,\ x^U + y^U], \quad x^I - y^I = [x^L - y^U,\ x^U - y^L],\\
x^I \times y^I &= [\min(x^L y^L,\ x^L y^U,\ x^U y^L,\ x^U y^U),\ \max(x^L y^L,\ x^L y^U,\ x^U y^L,\ x^U y^U)],\\
x^I / y^I &= [x^L,\ x^U] \times [1/y^U,\ 1/y^L]. \quad (0 \notin y_I)
\end{aligned}$$

We define arithmetic operations on $\mathcal{M}_{np}$. This definition allows us to calculate some statistical indices for an interval matrix easily. The arithmetic is the basis of the interval algebraic data analysis.

Definition 2.3. *interval matrices arithmetic*
Let $\boldsymbol{X}^I = (x_{ij}^I) \in \mathcal{M}_{np}$, $\boldsymbol{Y}^I = (y_{ij}^I) \in \mathcal{M}_{np}$, $\boldsymbol{Z}^I = (z_{ij}^I) \in \mathcal{M}_{pq}$. *Then, the sum interval matrix, the difference interval matrix and the product interval matrix are defined as follows:*

$$\boldsymbol{X}^I + \boldsymbol{Y}^I = \left(x_{ij}^I + y_{ij}^I\right), \quad \boldsymbol{X}^I - \boldsymbol{Y}^I = \left(x_{ij}^I - y_{ij}^I\right), \quad \boldsymbol{X}^I\boldsymbol{Z}^I = \left(\sum_{k=1}^{p} x_{ik}^I z_{kj}^I\right).$$

In the classical CA, we solve the eigenvalue problem. Also, in our proposal method, we treat the the eigenvalue problem for the interval matrix. So, we consider the interval eigenvalue problem, which is an extension of the eigenvalue problem of the classical matrix.

Definition 2.4. *interval eigenvalue, interval eigenvector*
Let $\boldsymbol{X}^I \in \mathcal{M}_{np}$. *The interval eigenvalue problem for the interval matrix* $\boldsymbol{X}^I$ *is as follows:*

$$\boldsymbol{X}^I\boldsymbol{u}^I = \lambda^I\boldsymbol{u}^I,$$

where λ^I *are called interval eigenvalues of* $\boldsymbol{X}^I$ *and any interval vectors* $\boldsymbol{u}^I$ *satisfying the interval eigenvalue problem are called interval eigenvectors of* $\boldsymbol{X}^I$. *The interval eigenvalues* λ_α^I *and interval vectors* $\boldsymbol{u}_\alpha^I$ *are represented in more detail as follows:*

$$\lambda_\alpha^I = \left[\min_{\boldsymbol{X}\in\boldsymbol{X}^I} \lambda_\alpha(\boldsymbol{X}),\ \max_{\boldsymbol{X}\in\boldsymbol{X}^I} \lambda_\alpha(\boldsymbol{X})\right], \quad \boldsymbol{u}_\alpha^I = \left[\left(\min_{\boldsymbol{X}\in\boldsymbol{X}^I} u_{\alpha i}(\boldsymbol{X})\right),\ \left(\max_{\boldsymbol{X}\in\boldsymbol{X}^I} u_{\alpha i}(\boldsymbol{X})\right)\right],$$

where $(\lambda_\alpha(\boldsymbol{X}),\ \boldsymbol{u}_\alpha(\boldsymbol{X}))$ *is the* α*-th eigenpair of* $\boldsymbol{X} \in \boldsymbol{X}^I$ *and* $u_{\alpha i}(\boldsymbol{X})$ *is the elements of* $\boldsymbol{u}_\alpha(\boldsymbol{X})$. *The pair* $(\lambda_\alpha^I,\ \boldsymbol{u}_\alpha^I)$ *is the* α*-th interval eigenpair of* $\boldsymbol{X}^I$.

Previously, the exact bounds for λ_α^I were not determinable, but now there can be determined using Deif [3]'s idea. However, the idea has the problem that the results tend to be oversize interval by calculating unconsidered matrices, when we analyze an interval data (Gioia and Lauro[5]). The interval eigenvectors are also determined by solving the linear optimization problem as described in Seif et al.[9], though it comes to be the same problem for calculating the eigenvalues.

Therefore, in this paper, we estimate the interval eigenvalues and eigenvectors of $\boldsymbol{X}^I$ by Monte Carlo simulation from multivariate uniform distribution.

3. Interval contingency table

We define the interval contingency table and interval contingency table matrix, and then consider the table's application to actual situations.

Definition 3.1. *interval contingency table, interval contingency table matrix*
Let X and Y be two categorical multi-valued variables and each domain is $\Omega_X = \{x_i |\ i = 1,\ 2,\ \ldots,\ n\}$ *and* $\Omega_Y = \{y_j |\ j = 1,\ 2,\ \ldots,\ m\}$, *respectively. Then, a contingency table whose cell of* (x_i, y_j) *is described as the interval* $f_{ij}^I = [f_{ij}^L,\ f_{ij}^U]$ *is called the* $(n \times m)$ *interval contingency table for the variables X and Y, and we regard it as the matrix* $\boldsymbol{F}_{XY}^I$, *which is called an interval contingency table matrix,*

$$\boldsymbol{F}_{XY}^I = \left(f_{ij}^I\right) = \begin{pmatrix} f_{11}^I & f_{12}^I & \cdots & f_{1m}^I \\ f_{21}^I & f_{22}^I & \cdots & f_{2m}^I \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1}^I & f_{n2}^I & \cdots & f_{nm}^I \end{pmatrix} = \begin{pmatrix} [f_{11}^L,\ f_{11}^U] & [f_{12}^L,\ f_{12}^U] & \cdots & [f_{1m}^L,\ f_{1m}^U] \\ [f_{21}^L,\ f_{21}^U] & [f_{22}^L,\ f_{22}^U] & \cdots & [f_{2m}^L,\ f_{2m}^U] \\ \vdots & \vdots & \ddots & \vdots \\ [f_{n1}^L,\ f_{n1}^U] & [f_{n2}^L,\ f_{n2}^U] & \cdots & [f_{nm}^L,\ f_{nm}^U] \end{pmatrix}.$$

Then, $\boldsymbol{F}_{XY}\left(=(f_{ij})\in \boldsymbol{F}^I_{XY}\right)$ *is called a classical interval contingency table matrix. Since* f_{ij} *is the element of natural number in the classical contingency table, we note that* f^I_{ij} *is regarded as the subset of natural number.*

We also introduce an interval row sum $f^I_{i\cdot}$*, an interval column sum* $f^I_{\cdot j}$ *and an interval total sum* $f^I_{\cdot\cdot}$ *as follows,*

$$f^I_{i\cdot}=\sum_{j=1}^{m} f^I_{ij},\quad f^I_{\cdot j}=\sum_{i=1}^{n} f^I_{ij},\quad f^I_{\cdot\cdot}=\sum_{i=1}^{n}\sum_{j=1}^{m} f^I_{ij}.$$

Next, we describe the situations for construction of the interval contingency table. These situations are similar to those used to construct the classical contingency table, and there are two situations. Thus, one situation is to count up the observations measured by two categorical multi-valued variables, and the other situation is to measure the observations directly. Now, we discuss each situation in detail.

In the first situation, there are the problems how to get the observations which measured by multi-valued variables, and three possible ways to get the such observations are as follows:

- The observed objects are the groups (e.g. class, category and concept) which are constituted of the aggregation of individuals, and all of individuals are described by the categorical single value.
- The observation is described by multi-values for an answer of the multi-selection question directly.
- The observation is described by considering the measurement error and uncertainty.

From the above mentioned situation, we can build an interval contingency table by counting up the observations that are described by the categorical multi-valued variables . The way where constructs the table from the variables is described by Rodríguez[8] in details.

In the second situation, we could describe the observation by considering the uncertainty because the accuracy of the observation is not guaranteed, e.g., when each observed object is always moving. The table is also used for privacy reasons, because it can prevent the disclosure of confidential information to unauthorized peoples.

This is because there are many cases where we use the interval contingency tables in the real world. So, studies of the interval contingency table have been meaningful and the development of data analysis methods for the table has been as expected.

4. Correspondence analysis for interval contingency table

In this section, we explain the IACA for the $(n\times m)$ interval contingency table matrix $\boldsymbol{F}^I_{XY}$. The purpose of IACA is similar to that of the CA for the classical tables, thus, visualization and comprehension of the relations between the modalities of the two categorical multi-valued variables in low-dimensional space.

In IACA, we start with the interval contingency table matrix $\boldsymbol{F}^I_{XY}=\left(f^I_{ij}\right)$ $(i=1,\ 2,\ \ldots,\ n;\ j=1,\ 2,\ \ldots,\ m)$, which count up the observation described by the two categorical multi-valued variables $X,\ Y$. Next, we apply the interval principal component analysis (IPCA) proposed by Gioia and Lauro[5] for the interval matrices $\boldsymbol{S}^I$ described as follows:

$$\boldsymbol{S}^I=\left(s^I_{ij}\right)=\left(\boldsymbol{D}^I_X\right)^{-1/2}\boldsymbol{F}^I_{XY}\left(\boldsymbol{D}^I_Y\right)^{-1/2}$$

where the $\boldsymbol{D}^I_X$ is $\mathrm{diag}(f^I_{i\cdot})$ $(i=1,\ 2,\ \ldots,\ n)$ and the $\boldsymbol{D}^I_Y$ is the $\mathrm{diag}(f^I_{\cdot j})$ $(j=1,\ 2,\ \ldots,\ m)$.

Applying IPCA is equivalent to solving the following interval eigenvalues problems:

$$\boldsymbol{S}^{I\prime}\boldsymbol{S}^I\boldsymbol{u}^I=\lambda^I\boldsymbol{u}^I,\quad \boldsymbol{S}^I\boldsymbol{S}^{I\prime}\boldsymbol{v}^I=\lambda^I\boldsymbol{v}^I,$$

where $\boldsymbol{U}^I=\left(\boldsymbol{u}^I_{(1)},\ \boldsymbol{u}^I_{(2)},\ \ldots,\ \boldsymbol{u}^I_{(m)}\right)$, $\boldsymbol{u}^I_{(i)}=(u^I_{1i},\ u^I_{2i},\ \ldots,\ u^I_{mi})'$, $\boldsymbol{V}^I=\left(\boldsymbol{v}^I_{(1)},\ \boldsymbol{v}^I_{(2)},\ \ldots,\ \boldsymbol{v}^I_{(n)}\right)$ and $\boldsymbol{v}^I_{(i)}=(v^I_{1i},\ v^I_{2i},\ \ldots,\ v^I_{ni})'$.

If you permit the overestimation of the bounds of eigenvalues and eigenvectors, the bounds can be obtained easily for using $\boldsymbol{S}^{I\prime}\boldsymbol{S}^I$ and $\boldsymbol{S}^I\boldsymbol{S}^{I\prime}$ by Deif[3] and Seif, et al.[9]'s idea. Otherwise, you should estimate the bounds for running Monte Carlo simulations.

After solving each interval eigenvalues problem, we get the $\hat{\boldsymbol{U}}^I$, $\hat{\boldsymbol{V}}^I$ and the interval factorial score $\hat{\boldsymbol{Z}}^I$ and $\hat{\boldsymbol{W}}^I$ by calculating the following equations:

$$\hat{\boldsymbol{Z}}^I = \left(\boldsymbol{D}_X^I\right)^{-1/2} \boldsymbol{S}^I \hat{\boldsymbol{U}}^I, \quad \hat{\boldsymbol{W}}^I = \left(\boldsymbol{D}_Y^I\right)^{-1/2} \boldsymbol{S}^{I\prime} \hat{\boldsymbol{V}}^I.$$

Then, the two eigenvectors $\hat{\boldsymbol{u}}_{(\alpha)}^I (\in \hat{\boldsymbol{u}}_{(\alpha)}^I)$ and $\hat{\boldsymbol{v}}_{(\alpha)} (\in \hat{\boldsymbol{v}}_{(\alpha)}^I)$ of the α-th eigenvalue $\lambda_\alpha (\in \lambda_\alpha^I)$ have the following relations:

$$\hat{\boldsymbol{v}}_{(\alpha)} = \frac{1}{\sqrt{\lambda_\alpha}} \boldsymbol{S} \hat{\boldsymbol{u}}_{(\alpha)}, \quad \hat{\boldsymbol{u}}_{(\alpha)} = \frac{1}{\sqrt{\lambda_\alpha}} \boldsymbol{S}' \hat{\boldsymbol{v}}_{(\alpha)},$$

We can derive the one factorial score $\hat{\boldsymbol{z}}_{(\alpha)}$ from the other factorial score $\hat{\boldsymbol{w}}_{(\alpha)}$ or the one factorial score $\hat{\boldsymbol{w}}_{(\alpha)}$ from the other factorial score $\hat{\boldsymbol{z}}_{(\alpha)}$ by the above relations and the following equations:

$$\hat{\boldsymbol{z}}_{(\alpha)} = \frac{1}{\sqrt{\lambda_\alpha}} \left(\boldsymbol{D}_X\right)^{-1} \boldsymbol{F}_{XY} \hat{\boldsymbol{w}}_{(\alpha)}, \quad \hat{\boldsymbol{w}}_{(\alpha)} = \frac{1}{\sqrt{\lambda_\alpha}} \left(\boldsymbol{D}_Y\right)^{-1} \left(\boldsymbol{F}_{XY}\right)' \hat{\boldsymbol{z}}_{(\alpha)},$$

where the $\boldsymbol{D}_X$ is diag$(f_{i\cdot})$ $(i = 1,\ 2,\ \ldots,\ n)$ and the $\boldsymbol{D}_Y$ is diag$(f_{\cdot j})$ $(j = 1,\ 2,\ \ldots,\ m)$ of the contingency table matrix $\boldsymbol{F}_{XY}$.

From these relations, we constitute the interval factorial scores matrix $\hat{\boldsymbol{Z}}^{I\prime}$ (or $\hat{\boldsymbol{W}}^{I\prime}$) and represent the modalities of the two categorical multi-valued variables in low-dimensional space. As the result of the plot, we can recognize the relative dissimilarities among the modalities. The advantage of these methods, in contradiction to the classical CA, is the fact that the internal variations can be represented as boxes.

Finally, we introduce the interval contribution ratio γ_α^I, which measures the degree of the variation which the original interval data matrix has.

Definition 4.1. *interval contribution ratio*
Let the interval eigenvalues of the interval matrix $\boldsymbol{X}^I$ be $\lambda_\alpha^I = [\lambda_\alpha^L,\ \lambda_\alpha^U]$ $(\alpha = 2,\ 3,\ \ldots,\ \min(n,\ m))$. γ_α^I is defined as,

$$\gamma_\alpha^I = \left[\lambda_\alpha^L \Big/ \left(\lambda_\alpha^L + \sum_{k=2,\, k\neq\alpha}^{\min(n,\, m)} \lambda_k^U\right), \quad \lambda_\alpha^U \Big/ \left(\lambda_\alpha^U + \sum_{k=2,\, k\neq\alpha}^{\min(n,\, m)} \lambda_k^L\right)\right],$$

and is called the interval contribution ratio for the α-th factorial axis.

The reason why we expect λ_1^I is $\lambda_1^I = 1$ inevitably. So, we start with the second factorial axis, when we put the factorial scores on the plain.

5. Numerical example

We discuss a numerical example and the results of IACA as applied to the hair color and eye color data (Rodríguez[8] First, we show the hair color and eye color data (Table.1).

Table 1: hair color and eyes color data

		color of the hair			
		black-hair	brown-hair	red-hair	blond-hair
color	black-eye	[60, 60]	[119, 123]	[20, 28]	[4, 7]
of	brown-eye	[15, 15]	[50, 58]	[14, 20]	[5, 11]
the	green-eye	[5, 5]	[24, 26]	[10, 12]	[11, 12]
eyes	blue-eye	[20, 20]	[70, 84]	[16, 17]	[90, 100]

This contingency table is consists of two categorical multi-valued variables for “eye color” and “hair color”. Here, the variable “eye color” takes the four modalities; black, green, red and blue and the variable “hair color” takes the four modalities, black, brown, red and gold.

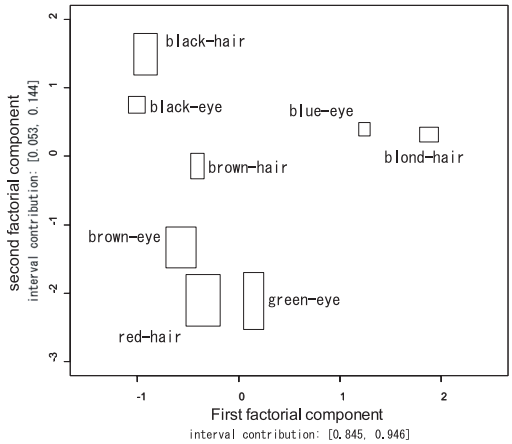

Figure 1: plot of interval factorial score

From this result (Fig.1), we recognize the relations between eye color and hair color. For example, people with red-hair tend to have green-eye or brown-eye, and people with blond-hair tend not to have black-hair. Besides the results of previous work, we recognize the internal variation of a modality as the box size. The box size is represented as the frequency of response under multi-selection. For example, people with red-hair have more modalities than people with blue-eye.

6. Conclusion

In this paper, we proposed the correspondence analysis for the interval contingency table based on interval algebra (IACA). The advantage of this method is that it can retain the internal variation of interval values, unlike the SymCA which loses the internal variations when calculating the statistical indices. However, this method needs some improvements. For example, the results of calculations based on interval algebra tend to produce interval data which is too oversized. Therefore, when we interpret the results, we should compare the results to the other symbolic correspondence analysis results. Finally, as a future work we want to propose another correspondence analysis for the interval contingency table.

Acknowledgements

The authors are grateful to the Editor and Referees for their useful suggestions and comments.

References

[1] Bock, H. -H. and Diday, E. (2000): *Analysis of Symbolic Data: Exploratory Methods for Extraction Statistical Information from Complex Data*, Springer, Berlin.
[2] Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997): Extensions de l' analyse en Composantes Principales a des Données de Type Intervalle, *Revue de Statistique Apploqué*, **24**, 5-24.
[3] Deif, A. S. (1991): The interval eigenvalue problem, *ZAMM*, **71**(1), 61-64.
[4] Diday, E. and Noirhomme, M. (2008): *Symbolic Data Analysis and the SODAS software*, John Wiley & Sons, Chichester.
[5] Gioia, F. and Lauro, C. N. (2006): Principal Component Analysis on Interval Data, *Computational Statistics*, **21**(2), 343-363.
[6] Moore, R. E. (1966): Interval Analysis, *Prentice Hall*, Englewood Cliffs, NJ.
[7] Neumaier, A. (1990): *Interval Methods for Systems of Equations*, Cambridge Univ. Press, Cambridge.
[8] Rodríguez, R. O. (2007): Correspondence analysis for symbolic multi-valued variables, "*http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/SymCA_CARME2007.229151706.pdf*".
[9] Seif, N. P., Hashem, S. and Deif, A. S. (1992): Bounding the Eigenvectors for Symmetric Interval Matrices, *ZAMM*, **72**, 233-236.