Multivariate Analysis for Chemistry-Property Relationships in Molten Salts

Changwon Suh[a], Slobodan Gadzuric[b,c], Marcelle Gaune-Escard[b], and Krishna Rajan[a]

[a] Combinatorial Sciences and Materials Informatics Collaboratory (CoSMIC), NSF International Materials Institute, Department of Materials Science and Engineering, Iowa State University, Ames, IA 50011, USA
[b] Ecole Polytechnique, IUSTI CNRS 6595, Technopôle de Château-Gombert, 5 rue Enrico Fermi, 13453 Marseille cedex 13, France
[c] Faculty of Science, Department of Chemistry, University of Novi Sad, Trg. D. Obradovica 3, 21000 Novi Sad, Serbia

Reprint requests to Prof. K. R.; Fax: 515-294-5444; E-mail: krajan@iastate.edu

Z. Naturforsch. **64a,** 467 – 476 (2009); received Dezember 4, 2006 / revised June 27, 2007

Presented at the EUCHEM Conference on Molten Salts and Ionic Liquids, Hammamet, Tunisia, September 16 – 22, 2006.

We systematically analyze the molten salt database of Janz to gain a better understanding of the relationship between molten salts and their properties. Due to the multivariate nature of the database, the intercorrelations amongst the molten salts and their properties are often hidden and defining them is challenging. Using principal component analysis (PCA), a data dimensionality reduction technique, we have effectively identified chemistry-property relationships. From the various patterns in the PCA maps, it has been demonstrated that information extracted with PCA not only contains chemistry-property relationships of molten salts, but also allows us to understand bonding characteristics and mechanisms of transport and melting, which are difficult to otherwise detect.

Key words: Molten Salts; Multivariate Analysis; Data Mining; Principal Component Analysis (PCA).

1. Introduction

Molten salts have many unique characteristics advantagous in industrial applications, such as improving the processing of metals where thermodynamic or kinetic constraints exist. Molten salts can be used in the electrodeposition of metals and composites, for better waste processing and recycling, and in enhancing a wide range of energy applications. The ever-growing field constituted by low-temperature multi-component molten salts [1] as well as room-temperature ionic liquids [2] should be stressed in view of the even larger possible applications related to the organic nature of cations in the latter. In the past, a huge demand to collect and publish all known properties of molten salts existed. In the present paper, we will use data mining tools to systematically analyze this data to extract new knowledge that will permit a better understanding of molten salts, related to aspects such as chemistry, processing and properties.

In the study of molten salts for any given chemistry, there exist corresponding structural, chemical, physical, and thermodynamic attributes (Table 1). The analysis of pre-existing empirical and theoretical data as well as the virtual design of new materials is a multivariate problem, which requires the use of data mining tools to find new information regarding the properties, both microscopic and macroscopic. Following the series of books published in the pioneering activity of G. J. Janz, an early version of a molten salt database was released, which is still the most comprehensive compilation of property data on molten salts available today. Some twenty years later, this numerical information was converted into a relational database, with Web-access capability [3 – 5]. In this paper, PCA is applied to this database, originally designed as a static compilation of materials data, to search for trends that can be useful in guiding future work to advance the molten salts field.

The data sets used in the present study are composed of seven variables for 1658 samples [6]. Since data of viscosity is limited in this database, two different cases are shown to illustrate applications of the PCA. While the data matrix having 473 samples for all seven vari-

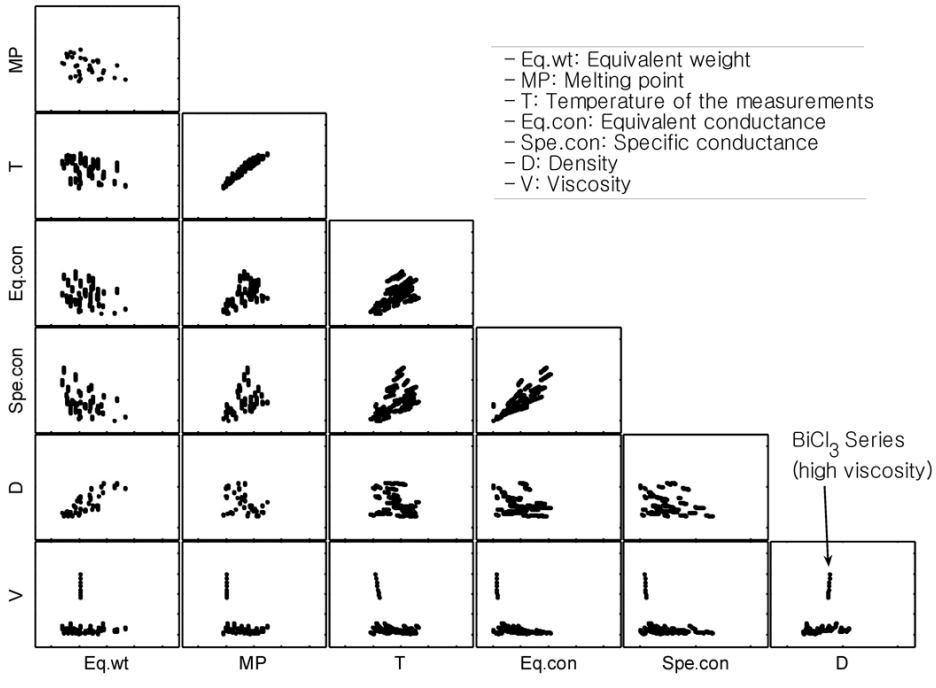


Fig. 1. Data of molten salts shown in multivariate arrays. This scatter plot is used to identify mutual relationships between two variables in each cell, and strong outliers can be detected (ex. $BiCl_3$ – high viscosity). Because of the effect of strong outliers, the correlations between viscosity and other variables are not clearly identified. Relative behaviours (e. g. trends with respect to all attributes) between variables and samples cannot be easily detected due to the numerous combinations in the scatter plot.

Table 1. Descriptors for molten salts.

- Structural data
  - Atomic/ionic size
  - Coordination number
  - Cation/cation distance
  - Anion/anion distance
- Viscosity
- Heat capacity
- Conductivity
- Surface tension
- Refractive index
- Density
- Compressibility
- Electrochemical potential
- Phase equilibria
- Cryoscopic behaviour
- Heat conductance
- Solubility
- Melting point
- Transport numbers
- Raman spectra
- Neutron scattering
- X-ray scattering
- NMR/EPR data

etc.

ables including viscosity was used as the first example, the second case studied does not include viscosity and is comprised of 1377 samples and six variables. As an example of traditional data visualization, all of the data in Fig. 1 is shown in the form of a scatter plot to see in a bivariate manner the relationships existing between variables. While some correlations can be seen, it is difficult to describe the relationship among multiple parameters and extract this information so as to guide future work. For this reason, we are analyzing this data through a systematic data mining methodology. Future work will look to apply the statistical techniques developed here to more recent databases of molten salts [7 – 11].

2. Principal Component Analysis: Informatics Tool for Multivariate Data

The multivariate data analysis method used in this paper is PCA, which is a useful projection tool in qualitatively guiding the interpretation of huge amounts of multivariate data with interrelated variables. By reducing the information dimensionality in a way that minimizes the loss of information, PCA constructs uncorrelated axes leading to the transformation (i. e. rotation) of the original coordinate system. The constructed PCA axis is referred to as a "latent variable" (LV) or "principal component" (PC). LVs which are independent (i. e. orthogonal) of each other are the linear combinations of original variables. By using just a few LVs, the dimensionality of the original multivariate data sets are reduced and visualized by their projections in the 2-dimensional (2D) or 3-dimensional (3D) space with a minimal loss of information. Therefore, PCA allows dimensionally reduced mapping of multivariate data sets. PCA is used in this paper because of the multiple physical parameters in the molten salts data.

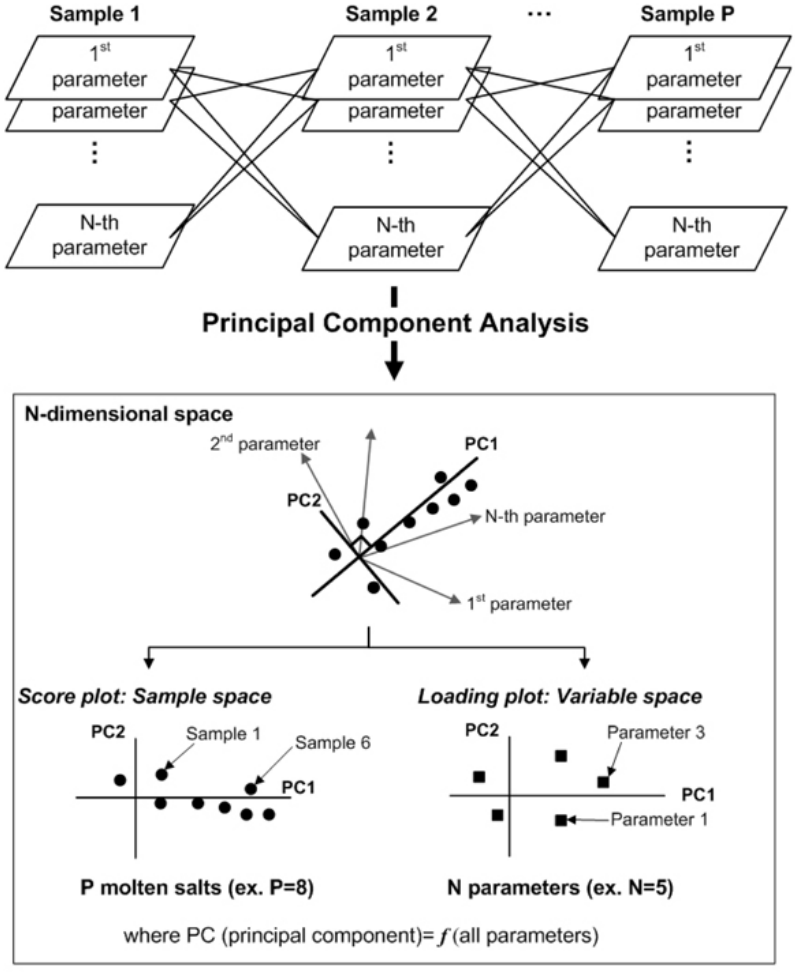


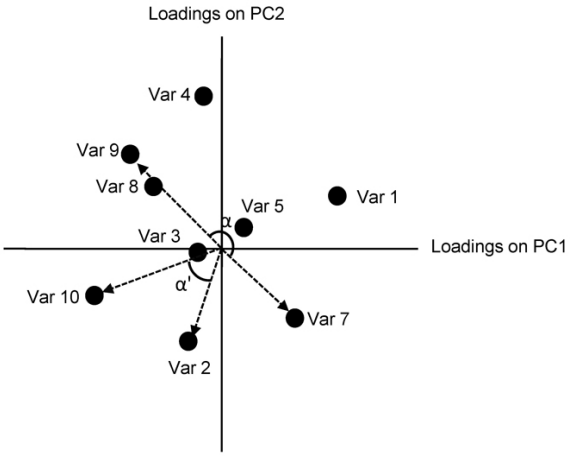
Fig. 2. A graphical schematic of the PCA approach from given N parameters. Variable mappings are done in hyperspace (i. e. N-dimensional space). The direction of maximum variance in the samples in the hyperspace becomes PC1. PC2 explains the second largest variance of samples. PC1 and PC2 are orthogonal to each other. Score and loading plots from PCA represent correlation maps with respect to reduced dimensionality. This approach provides a guide to identify interdependency of parameters (i. e. in a loading plot) and correlations between molten salts (i. e. in a score plot) in the 2D space.

PCA relies on the fact that most of the variables are intercorrelated. As shown in Fig. 2, we can derive a set of N uncorrelated variables (the principal components) from a set of N correlated variables. Each selected parameter is then combined to make latent variables in the form of linear combinations. Therefore, each PC is a suitable linear combination of all the original variables. The first principal component (PC1) accounts for maximum variance (information in data) in the original dataset. The second principal component (PC2) is orthogonal (uncorrelated) to the first one and accounts for the largest amount of remaining variance. Thus, the m-th PC is orthogonal to all others and has the m-th largest variance in the set of PCs. Once the N PCs have been populated using eigenvalue/eigenvector matrix operations, only PCs with variances above a critical level determined from a scree plot are retained. By exploiting the low dimensionality of data sets formed by a few PCs, molten salts are easily classified by the effects of variables, and variables are clustered with their statistical similarities as well. Through the process of eigenvector decomposition in PCA, the original data is decomposed by two matrices, loadings and scores.

The loadings are the weights of each original variable while scores contain information of original samples in a rotated coordinate system. Thus, PCA loading plots are mapping the correlation between variables, and score plots are mapping the correlation between samples (Fig. 3). Interpretations of the loading plot are twofold:

- Degree of correlation between variables: angle between variable-origin-variable.
- Relative impact of variables on the PCA model: distance from the origin to variable.

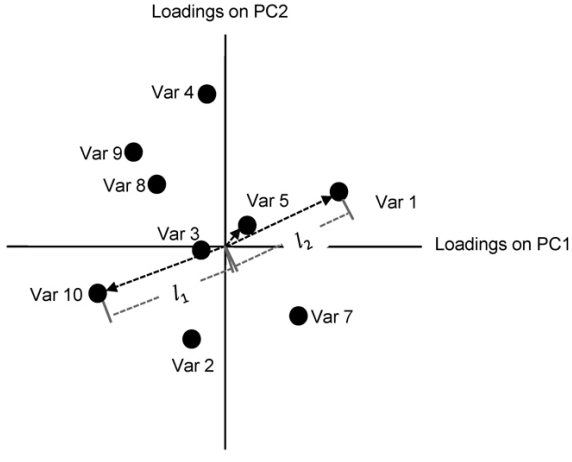
Degree of correlation between variables:
Angle between arrows (loading position)



1) Similar properties (correlated) are grouped together.
2) Inversely correlated variables sit on opposite (or diagonal) sides.
3) These relations can be explained using cosine angle between arrows from origin to variables.

Examples:
- Var 8 and 9 are highly correlated [$\cos(0) \sim 1$].
- Var 8 and 9 are also inversely correlated with var 7 [$\cos(\alpha) = \pi \rightarrow 1$].
- Var 10 and 2 are somewhat correlated [$\cos(\alpha') < \pi/2$].
- Similarly, there is almost no correlation between var 9 and var 5.

Relative impact of variables on PCA model:
Length of arrows (distance from the origin)



1) A longer distance indicates a stronger impact, while a shorter distance corresponds to a weaker impact.
2) Similar distances represent similar impacts of variables.

Examples:
- Var 10 has a strong impact on PC1 while PC2 is a strong function of var 4.
- Var 1 and 10 have similar impacts on the PC1-PC2 model ($l_1 \approx l_2$).
- Impact of var 3 or var 5 on PC1-PC2 model is quite small.

Fig. 3. A schematic illustration of PCA interpretation to track correlations between variables (or properties). The PCA loading plot simply shows how an N-dimensional correlation plot appears, while still retaining many of the qualitative features shown in the bivariate case such as Figure 1. Each point represents loadings of each variable (Var). Thus, ten variables are shown in this example (i. e. $N = 10$).

The degree of correlation between variables is determined by the angles (cosine) between them. If the angle between two variables at the origin is θ, then $\theta = 0^\circ$ for highly positively correlated variables, $\theta = 180^\circ$ for highly inversely correlated variables, and $\theta = 90^\circ$ if no correlation exists. A PCA loading plot captures all the possible bivariate correlations within a multivariate data set in the 2D space. Since PCA reduces the dimensionality of variables with a minimum loss of information, it should be noted that correlations on the PCA map depend on the variance captured by the confined dimensions, which means that they could be different from bivariate correlation coefficients (e. g. Pearson's). On the other hand, the relative impact of each variable can be identified by measuring the distance from the origin. For instance in Fig. 3, the PC1 axis mainly measures variable 1 and 10 while the PC2 axis captures variables 2 and 4. Therefore, information contained by the first two PCs is highly related to these four variables.

The same logic can be applied to the score plot. Samples having similar properties sit closely in the score plot and samples of different behaviours sit separately. Outliers generally exist at a long distance from the origin. The key point in loading and score plots is that all the variables and samples are simultaneously explained by their relative behaviours in terms of correlations in reduced dimensions. More detailed mathematical description of PCA can be found in literatures [12, 13].

3. Results and Discussion

3.1. Case Study 1: (473 × 7) Data Matrix including Viscosity

In this section, we provide examples of our analysis using the PCA technique. The impact of a multivari-

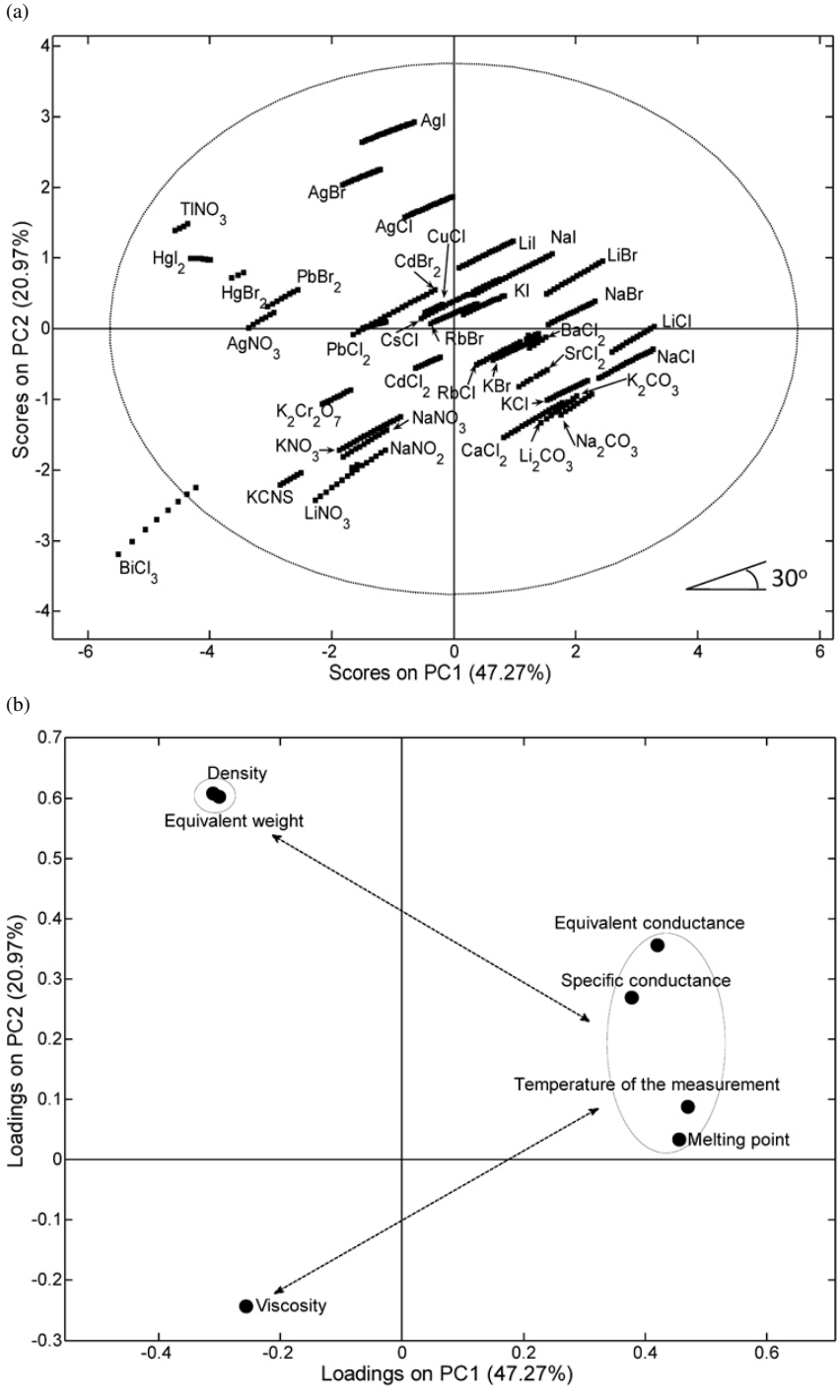


Fig. 4. (a) The PC1-PC2 score plot for the complete data using principal components with 95% confidence limits (ellipse). The use of two PCs to concisely describe a variable space spanned by seven descriptors is a good example of this method's advantages. The outliers can be seen outside the 95% confidence limit envelope ($BiCl_3$ series). (b) The loading plot displays the relationships between the variables. The odd behaviour of $BiCl_3$ in the score plot is due to the high viscosity of this sample as shown in the loading plot.

ate analysis of seven descriptors was explored on conductivity and density behaviour. These are presented in Fig. 4 and contain all seven descriptors as a function of the temperature of the measurement. The first two PCs explain 68.24% (PC1: 47.27%, PC2: 20.97%) of the variance (information) in the data. These PCs are

linear combinations of seven variables according to:

$$\begin{aligned}\text{score on PC1} = &-0.301Eq.wt + 0.455MP \\ &+ 0.469T + 0.420Eq.con + 0.378Spe.con \\ &- 0.311D - 0.257V,\end{aligned} \quad (1)$$

$$\begin{aligned}\text{score on PC2} = &\ 0.602Eq.wt + 0.033MP \\ &+ 0.088T + 0.356Eq.con + 0.270Spe.con \\ &+ 0.608D - 0.243V,\end{aligned} \quad (2)$$

where *Eq.wt* is the equivalent weight, *MP* the melting point, *T* the temperature of the measurements, *Eq.con* the equivalent conductance, *Spe.con* the specific conductance, *D* the density, and *V* the viscosity. Each coefficient in the linear combination is defined by the loading value and is used to create loading plots such as in Figure 4b.

In the score plot, Fig. 4a, most of the trajectories for temperature appear to lie at around 30° to the chemistry trajectory due to the combined effect of viscosity and temperature, as seen in Figure 4b. $BiCl_3$, an outlier in the PC space of Fig. 4a, also appears as an outlier in the bivariate plot of Fig. 1, confirming our results. In the loading plot, Fig. 4b, melting point, temperature of the measurement, and equivalent/specific conductance are strongly correlated. Most molten salts are classical examples of this behaviour since they have high density but low viscosity. We see this relationship from the loading points through the PC2 axis of loadings, as well as the relationships between viscosity and other variables with the PCA plot of Fig. 4b. Viscosity has negative correlations with melting point, temperature of the measurement, and equivalent/specific conductance, according to their PC values in diagonal quadrants. Correlations between density (or equivalent weight) and conductance are not strong because they have positive PC2 values, although they have different signs in PC1 values.

To screen the relationships between samples and variables, we should explore the scores and loadings simultaneously (i. e. Figs. 4a and b). For instance, samples in the third quadrant have high viscosity. Similarly, samples in the second quadrant have relatively high density and equivalent weight. Consequently, the PCA plot serves as a correlation map of samples, variables, and samples-variables for qualitative and quantitative interpretations of physical behaviours of molten salt systems. These correlations can be found by plotting parameters against each other, as shown in Fig. 1, but that requires many plots and is inefficient.

3.2. Case study 2: (1377 × 6) Data Matrix without Viscosity

In the second example, we perform PCA using all the single-salt records without any viscosity information in the Janz data set. The size of this data matrix is 1377×6. In this case, we use PC1, PC2, and PC3 because PC3 accounts for over 19% of the variance. In Fig. 5, all the Janz data without viscosity is compressed and visualized in the 3D space in a way that minimizes the loss of information. From Fig. 5, we choose two interesting projections, PC1-PC2 and PC1-PC3 for interpretation (Figs. 6 and 7, respectively). As shown in Fig. 6a and Fig. 7a, changes in bonding characteristics along the PC1 axis are observed. The compounds on the right-hand side in the score plot have typical ionic characteristics while more covalent compounds sit on the left-hand side of the plot.

In Fig. 7, scores and loadings spanned by PC1 and PC3 are shown. Two described compounds on the PC1-PC3 projection in Fig. 7a are not in the 95% confidence limit. These melts are LiCl and LiBr, which have the most ionic bonds compared to the other compounds presented here. Our procedure and the descriptors mentioned above describe the melts well, with the exception of those with the most extreme bonding (high ionicity). For possible improvement, we could integrate new descriptors that are characteristic of ionic compounds.

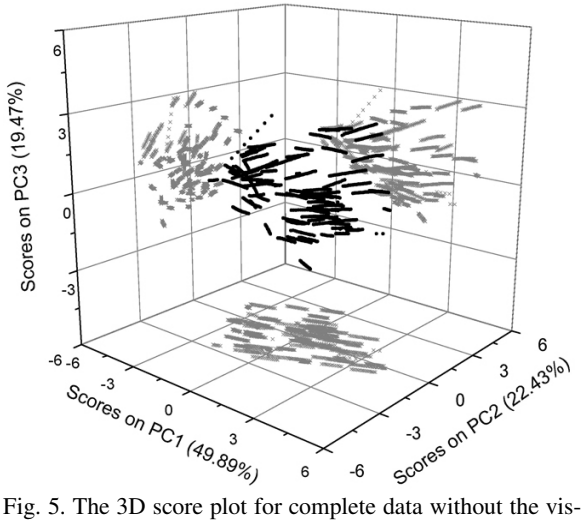


Fig. 5. The 3D score plot for complete data without the viscosity information. This dimensionally compressed plot contains most of the sample trends with respect to six variables, with a minimal loss of information. Black, for raw data; gray, for PC1-PC2, PC1-PC3, and PC2-PC3 projection as appropriate.

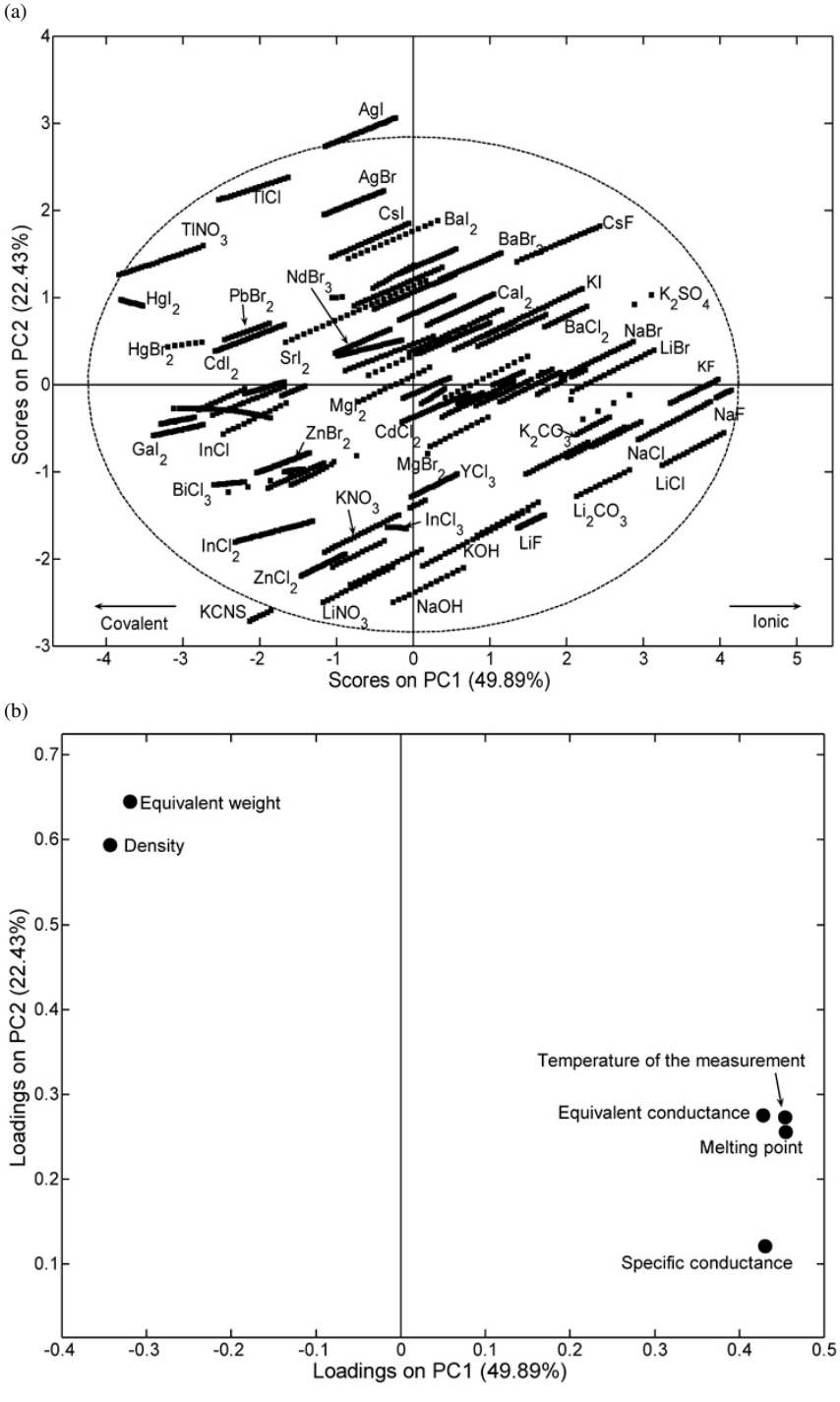
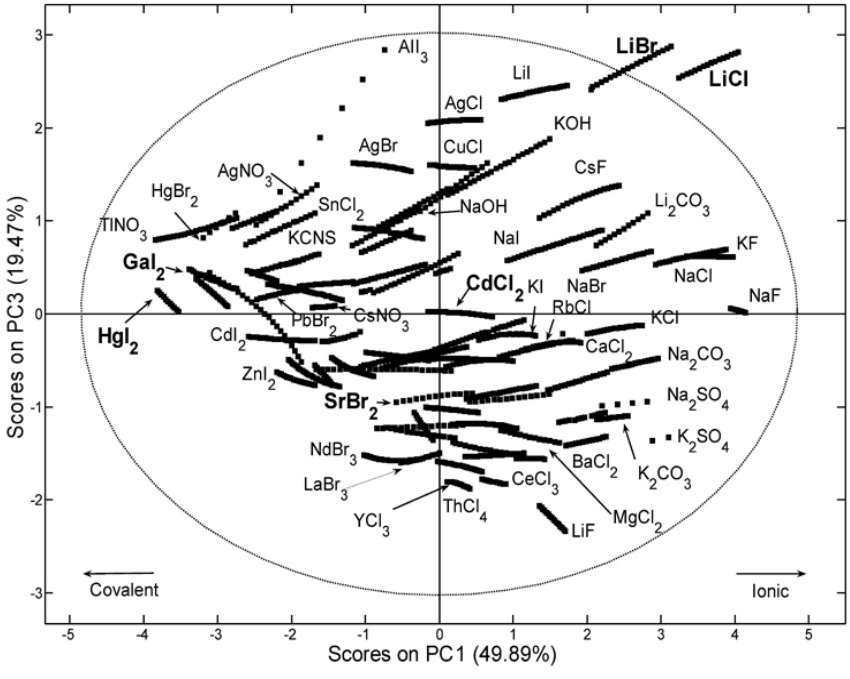


Fig. 6. (a) The score plot of PC1-PC2 for complete data without the viscosity data using principal components with 95% confidence limits. (b) The loading plot of PC1-PC2.

As shown in Fig. 7b, the loading values of temperature of the measurement, melting point, and conductance are similar in PC1 while the loading values are different in PC3. Therefore, PC3 captures the independent characteristics of conductance, temperature of the measurement and melting point. This effect is depicted with marked arrows in Figure 7b.

(a)



(b)

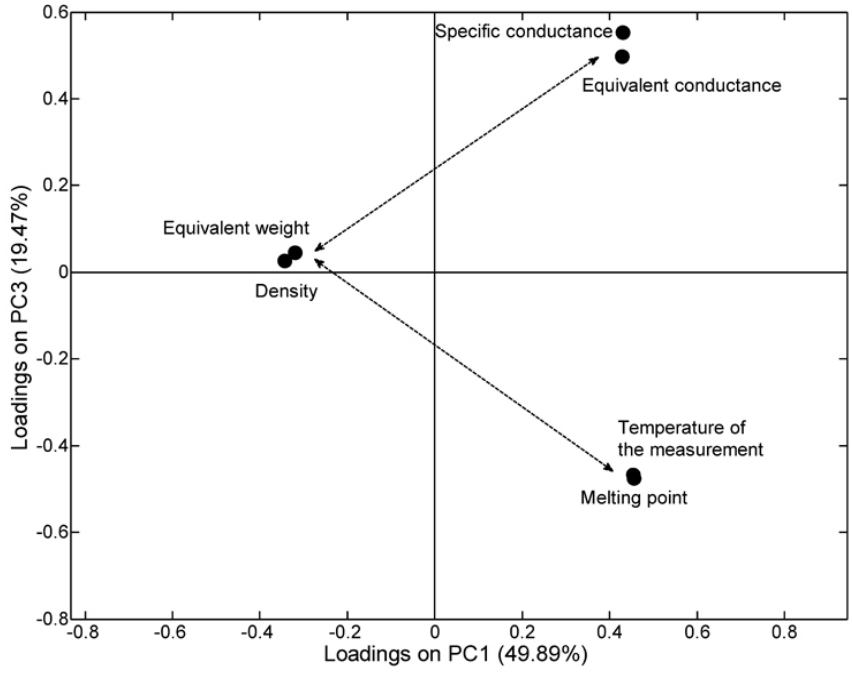


Fig. 7. (a) The score plot of PC1-PC3 for complete data without the viscosity column using principal components with 95% confidence limits. The two emphasized melts, $CdCl_2$ and $SrBr_2$, in the centre of the score plot have nearly temperature-independent conductivity and mixed bonding characteristics (ionic and covalent). The other two emphasized melts (GaI_2 and HgI_2) on the left of the plot have low transport coefficients. (b) The loading plot of PC1-PC3. Note that PC3 captures effects due to conductance, temperature and melting point while density and equivalent weight are not explained fully [arrows on (b)]. Converging trends are shown on the left side of the score plot which show conducting mechanisms as a function of temperature.

Different slopes for each molten salt in the score plots, Fig. 6a and Fig. 7a, explain different behaviours of molten salts. For example, the trajectory for each sample describes the conduction mechanism and transport coefficients. Differences in slopes for each sample are more clearly seen on the PC1-PC3 projection,

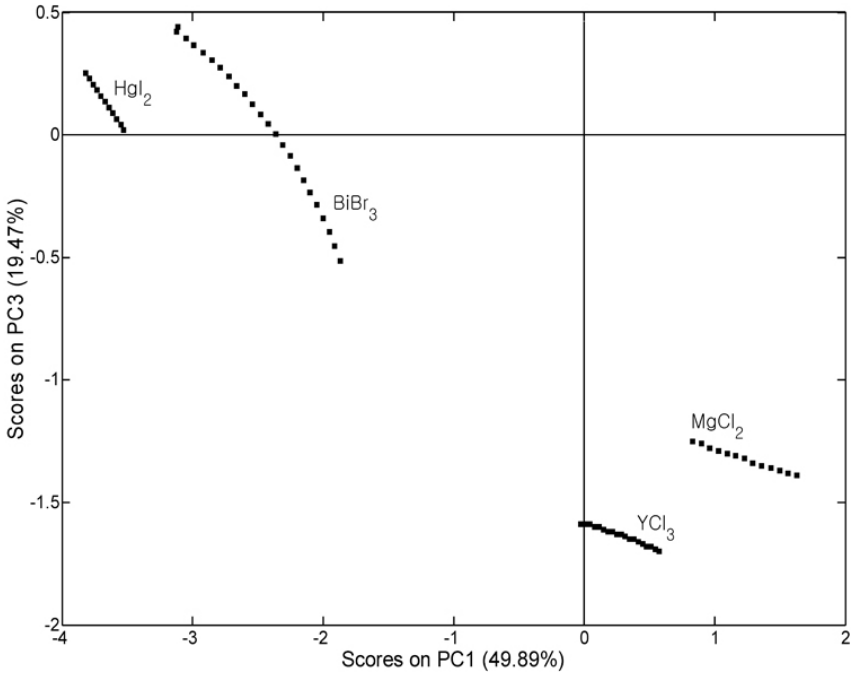


Fig. 8. The magnified score plot of Figure 7a. For clarity, only four interesting melts are shown.

Figure 7a. Since bonding characteristics of the melts such as $CdCl_2$ and $SrBr_2$ are neither covalent nor ionic based on their score values (PC1 $\rightarrow$ 0), their conductivity change with temperature is minor, as seen by a trajectory parallel to the PC1 axis. While samples having a positive slope of trajectory (i. e. increasing conductance against temperature) are on the positive side of the PC3 axis, samples with a negative slope of trajectory are on the negative side of the PC3 axis. While transport properties are very high in ionic melts, they are low in molecular liquids (ex. HgI_2 and GaI_2) and in the network type of melts, similar to the case of polymers (ex. halides of zinc). Since the thermal and specific conductivity are directly temperature-dependent properties, the same trends in conductivity are observed on the loading plots (Fig. 6b and Fig. 7b).

Comparing the effect of ionicity-covalency on melting points (PC1 axis) for all compounds, we observed that increased covalency is consistent with a decrease in melting point. The highest melting points pertain to ionic materials (for instance, LiF, NaF, CsF, $BaCl_2$ with melting points in the range 1000 – 1200 K) and covalent melts have lower melting points (HgI_2, $TlNO_3$, and $SnCl_2$ with melting points in the range 450 – 600 K). Materials in the middle range of the plot (score values of PC1 nearly zero) were identified as corresponding to the melting range 600 – 1000 K.

The well-known linear relationship between entropy of melting (ΔS_m) and estimated volume change ($\Delta V_m/V$) on melting can be used to describe different behaviours of the systems in this study. For the systems with a special melting mechanism, i. e. from disordered solid or a network-forming liquid into a molecular liquid, some exceptions from linearity are observed [14, 15]. The same melts also deviate from the others as shown in Fig. 8, a magnified version of Figure 7a. In this figure, a steeper slope is observed when the melting system goes through a transition from ionic crystal to molecular liquid (ex. HgI_2 and $BiBr_3$). The melting mechanisms of $MgCl_2$ and YCl_3, which have relatively gentle slopes, can be viewed as a transition from an ionic crystal to an ionic liquid without any drastic difference in behaviour during the melting, as compared with the other molten salts. Even in the absence of those values, these observations confirm our descriptions of selected molten salts through PCA.

Most patterns of samples converge to a single type of melt on the left-hand side in the PC1-PC3 score plots of Figure 7. We also assume that melts around this converging region should be molecular liquids without any change of the molecular structure during the melting process. These melts would also have a high molar mass, limited conductivity, high density

and low melting point and might even be liquid at room temperature.

4. Conclusions

We have provided examples of statistical approaches for identifying chemistry-property relationships in a classic materials database (molten salts). By using PCA, we described multiple physical parameters easily, helping us to identify outliers, find global and local patterns in the samples, and study the correlations between the variables. The results showed that extracted information from PCA captures bonding characteristics, transport mechanisms, and the melting of molten salts. Consequently, it has been demonstrated that the classical Janz's molten salts database is a good template for applying data mining for design and analysis of molten salts. It is desirable to include structural data into Janz's database or incorporate it into other databases for further study.

Acknowledgements

The authors gratefully acknowledge support from the National Science Foundation International Materials Institute Program on Combinatorial Sciences and Materials Informatics Collaboratory (CoSMIC-IMI) (Grant # DMR: 0603644). S. G. wishes to thank Department of Materials Science and Engineering, Iowa State University, Ames, USA, Ministry of Science of Republic of Serbia, and Ecole Polytechnique de Marseille for hospitality and support during this work.

[1] B. Mishra, I. Maroef, and D. J. Hebditch, International Conference on Molten Salts and Fluxes 2000, Sweden-Finland, CD-ROM Pub., June 2000, p. 1.

[2] V. Kamavaram and R. G. Reddy, in: Proceedings of the Recycling and Waste Treatment in Mineral and Metal Processing: Technical and Economic Aspects (Eds. B. Bjrokman, C. Samelsson, and J.-O. Wikstrom), Vol. 2, LUT, Lulea, Sweden 2002, p. 517.

[3] M. Gaune-Escard, The Web-based molten salt database project, Contract NIST (2000 – 2002).

[4] M. Gaune-Escard and J. Fuller, High Temp. Materials Processes **20**, 309 (2001).

[5] J. Fuller and M. Gaune-Escard, in: NATO Science Series "Green Industrial Applications of Ionic Liquids", Vol. 92 (Eds. R. D. Rogers, K. R. Seddon, and S. Volkov), Kluwer Academic Publishers, Dordrecht, 2001, p. 275.

[6] G. J. Janz, A. T. Ward, and R. D. Reeves, Electrical Conductance, Density, and Viscosity, Technical bulletin series, Rensselaer Polytechnic Institute, 1964.

[7] G. J. Janz, G. M. Dijkhius, G. R. Lakshminarayanan, R. P. Tomkins, and J. Wong, Molten Salts: Vol. 2, Section 1, Electrochemistry on Molten Salts, Gibbs Free Energies and Excess Free Energies from Equilibrium-Type Cells, Section 2, Surface Tension Data, NSRDS-NBS 28, 1968.

[8] G. J. Janz, C. B. Allen, J. R. Downey Jr., and R. P. Tomkins, Physical Properties Data Compilations Relevant to Energy Storage. I. Molten Salts: Eutectic Data, NSRDS-NBS 61, Part I, 1978.

[9] G. J. Janz, C. B. Allen, N. P. Bansal, R. M. Murphy and R. P. Tomkins, Physical Properties Data Compilations Relevant to Energy Storage. II. Molten Salts: Data on Single and Multi-Component Systems, NSRDS-NBS 61, Part II, 1979.

[10] G. J. Janz and R. P. Tomkins, Physical Properties Data Compilations Relevant to Energy Storage. IV. Molten Salts: Data on Additional Single and Multi-Component Salt Systems, NSRDS-NBS 61, Part IV, 1981.

[11] G. J. Janz, J. Phys. Chem. Ref. Data **17**, Supplement No. 2, 1988.

[12] A. Rajagopalan, C. Suh, X. Li, and K. Rajan, Appl. Catal. A-Gen. **254**, 147 (2003).

[13] I. T. Joliffe, Principle Component Analysis, 2nd ed., Springer Series in Statistics, Springer, New York 2002.

[14] N. H. March and M. P. Tosi, Introduction to Liquid State Physics, World Scientific Publishing, London 2002.

[15] Z. Akdeniz and M. P. Tosi, Proc. R. Soc. Lond. A **437**, 85 (1992).