THE APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO MATERIALS SCIENCE DATA

Changwon Suh, Arun Rajagopalan, Xiang Li and Krishna Rajan*

Department of Materials Science and Engineering and Faculty of Information Technology
Rensselaer Polytechnic Institute, Troy NY 12180-3590 USA
** Corresponding author- Email: rajank@rpi.edu*

ABSTRACT

The relationship between apparently disparate sets of data is a critical component of interpreting materials' behavior, especially in terms of assessing the impact of the microscopic characteristics of materials on their macroscopic or engineering behavior. In this paper we demonstrate the value of principal component analysis of property data associated with high temperature superconductivity to examine the statistical impact of the materials' intrinsic characteristics on high temperature superconducting behavior.

Keywords: Combinatorial response maps, Data mining, High temperature superconductors, Materials science, Principal component analysis (PCA)

1 INTRODUCTION

The use of principal component analysis techniques is well established in many fields such as pharmacology, climatology, numerous aspects of the life sciences, economics, (Jolliffe, 1986; Faloutsos, Korn, Labrinidis, Kotidis, Kaplunuovich, & Perkovic, 1997; Preisendorfer, 1988; Shum, Ikeuchi, & Reddy, 1997) and even religious studies! See for example Wilker (2001) who has provided a very illustrative and imaginative use of this statistical methodology.

Principal Component Analysis (PCA) is a technique to reduce the information dimensionality that is often needed from the vast arrays of data obtained from a combinatorial experiment, in a way that minimises the loss of information. It relies on the fact that most of the descriptors are intercorrelated and these correlations in some instances are high (Bajorath, 2001). From a set of N correlated descriptors, we can derive a set of N uncorrelated descriptors (the principal components). Each principal component (PC) is a suitable linear combination of all the original descriptors. The first principal component accounts for the maximum variance (eigenvalue) in the original dataset. The second principal component is orthogonal (uncorrelated) to the first and accounts for most of the 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 calculated using eigenvalue/eigenvector matrix operations, only PCs with variances above a critical level are retained. The M-dimensional principal component space has retained most of the information from the initial N-dimensional descriptor space, by projecting it onto orthogonal axes of high variance. The complex tasks of prediction or classification are made easier in this compressed space.

In the materials sciences, it's use is not widespread and there may be numerous reasons for this. In research disciplines where the observance of patterns of behavior in nature such as weather patterns, migratory patterns of animals or the patterns in the efficacy of molecules in serving as building blocks for drugs, the primary question is whether we can determine what parameter or combination of parameters and to what extent do they appear to influence the macroscopic pattern. In this way, one can discern, in a statistical sense, the relative importance of these factors. In materials science the tendency is to begin with a paradigm of what we already think is important and use that as a basis for synthesizing or processing materials. This approach has certainly been utilized literally for centuries and in the process a "database" of knowledge has

been built through phenomenological associations, theory, computation and experiments. The process of materials discovery, however, is still a process governed by empiricism and accidental discoveries (high temperature ceramic superconductors, carbon fullerenes, and conducting polymers to mention some recent examples).

While incremental progress is made in specific technological areas of interest, we need to have a means of exploring vast combinations of structure-property relationships. If significant new advances in materials science are to be made, we need to have search tools that can accelerate the discovery process. The challenge in linking length scales in materials science is that we do not necessarily have theories linking every aspect of materials' characteristics in a unified manner. Much of materials design is based on phenomenological paradigms that provide guidelines for materials selection. The challenge that we are addressing in this proposal is how to integrate data at different length scales in such a way as to detect patterns of behavior (using statistical techniques) that could lead to (or suggest) new data or information (validated by experiments and theoretical formulations). Data mining is envisaged as a tool to exploit the masses of available data to accelerate the discovery of these relationships and possible new associations. For us the "association" is between structure and property of materials. Data mining acts as a descriptive tool for hypothesizing relationships between structures and materials that are interpretable by the material scientist. Materials Science offers a unique challenge in data mining due to the variety of data types, and their complex interconnections. During the material discovery process, there is a need to integrate multiple, heterogeneous databases to reach new and even unexpected conclusions as well as to use databases actively to design new processing strategies. This complex coupling of data models, data analysis methods and physical methods offer a unique computing challenge that has not yet been addressed sufficiently in information technology research. In this paper we provide an illustration of one of the important data mining techniques, namely principal component analysis, and how it helps us to manage information complexity in materials behavior.

2 COMBINATORIAL RESPONSE MAPS

As an example of the value of such data mining tools, we have conducted an approach to reduce the dimensionality of the multivariate problem in developing descriptors for high temperature compound superconductors. For the purposes of this discussion, we shall focus on situations where there exists a vast array of variables associated with a single set of compounds or chemistry. The recent discovery of MgB_2 a well known compound was accidentally to have found to possess superconducting characteristics (Nagamatsu, Nakagawa, Muranaka, Zenitini, & Akimitsu, 2001). We explored a wide array of descriptors based on "legacy" data of other inorganic compounds (intermetallic and ceramic systems) possessing high temperature superconducting behavior including: average number of valence electrons, electronegativity difference, radii difference, elemental concentration/mole fraction stoichiometry, cohesive energy and ionization energy. The choice of these 'descriptors' was initially based on earlier studies that had attempted to search for correlations between crystal chemistry and crystal structure and high Tc properties (Villiars & Phillips, 1988). However these studies attempted to look for correlations only between "raw" data and as will be shown here, much information can be lost in that manner. While the details of the calculations and methodologies of assessing these parameters will be discussed in a later publication, suffice it to say that the data was based on numerous sources from the archived published literature (Philips, 1989; Poole, Datta, & Farach, 1988; Poole, Farach, & Creswick, 1995; Satta, Profeta,, Bernardini, Continenza, & Massidda, 2001; Medvedeva, Ivanovskii, Medvedeva, & Freeman, 2001; Imai & Hirano, 1997). As no digital library on superconducting compounds exists, a detailed and exhaustive survey of the literature was conducted where information on the presence of high temperature superconductivity in compounds along with the information on all the relevant descriptors for each compound was available. This process of data warehousing was synthesized into multiple scatter plots and is shown below. We like to refer to this format of data representation as a "Combinatorial Response Map" as it maps out the vast array of combination of materials responses to a variety of descriptors. This also serves to graphically represent the challenge and need for techniques that can condense this information in a statistical manner to find which combinations of descriptors appear to have the most influence on the response function of interest (in this case, high

temperature superconductivity). For the purposes of this study we dealt with materials classification as a binary response problem (superconducting or non-superconducting).

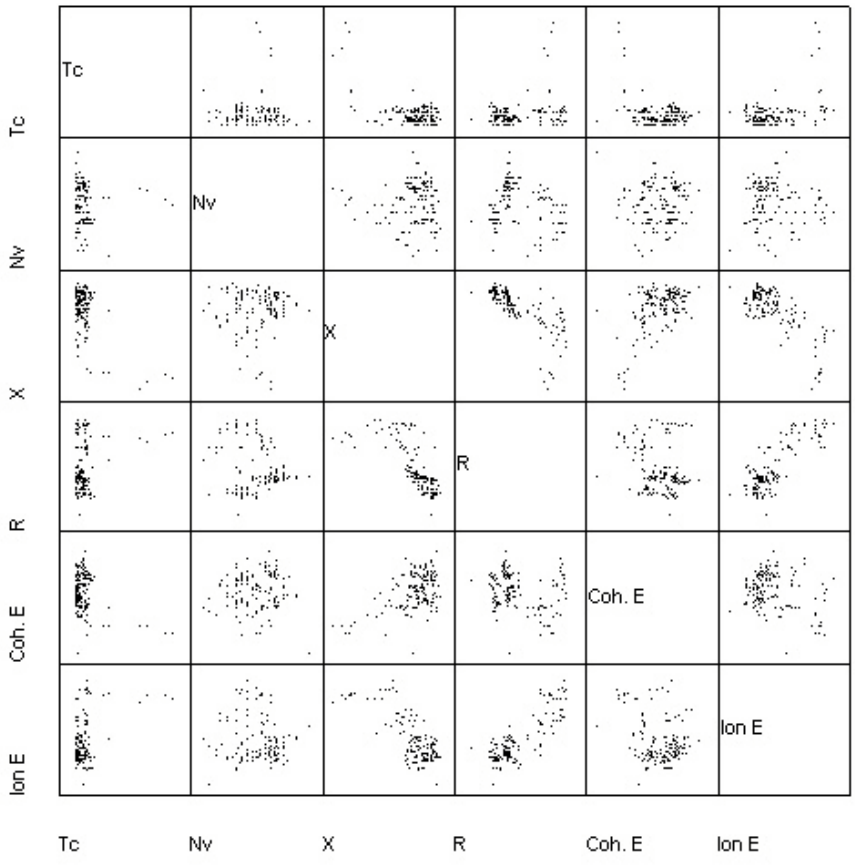


Figure 1. Combinatorial response map for high temperature superconductor descriptors.

3 RESULTS AND DISCUSSION

For the purposes of this discussion, we shall focus on situations where there exists a vast array of variables associated with a single set of compounds or chemistry. A statistical evaluation to search for each descriptor is computationally expensive and most probably ineffective. Principal Component Analysis (PCA) is a technique to reduce the information dimensionality that is often needed from the vast arrays of data obtained from a combinatorial experiment, in a way that minimizes the loss of information (Figure 3). It relies on the fact that most of the descriptors are intercorrelated and these correlations in some instances are high (Bajorath, 2001). From a set of N correlated descriptors, we can derive a set of N uncorrelated descriptors (the principal components). Each principal component (PC) is a suitable linear combination of all the original descriptors. The first principal component accounts for the maximum variance (eigenvalue) in the original dataset. The second principal component is orthogonal (uncorrelated) to the first and accounts for most of the 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 calculated using eigenvalue/eigenvector matrix operations, only PCs with variances above a critical level are retained. The M-dimensional principal component space has retained most of the information from the initial N-dimensional descriptor space, by

projecting it into orthogonal axes of high variance. Following the treatment of Wichern & Johnson (2002), we can describe this mathematically as follows:

Consider p random variables $X_1, X_2, \ldots, X_p$. The original system can be rotated and a new coordinate system obtained with the new axes representing directions with maximum variability. The new axes, which are linear combinations of the original axes, are the principal components.

Let $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X'} = [X_1, X_2, \ldots, X_p]$. The corresponding eigenvalue-eigenvector pairs are $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Then the ith principal component is given by

$$Y_i = \mathbf{e'}_i\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p, \quad i = 1,2, \ldots, p. \tag{1}$$

Then,

$$\text{Var}(Y_i) = \mathbf{e'}_i\mathbf{\Sigma}\,\mathbf{e}_i = \lambda_i \quad i = 1,2, \ldots, p \tag{2}$$

$$\text{and Cov}(Y_i, Y_k) = \mathbf{e'}_i\mathbf{\Sigma}\,\mathbf{e}_k = 0 \quad i \neq k(3) \tag{3}$$

Thus the principal components are uncorrelated and have variances equal to the eigenvalues of Σ. Another property of the principal components is

$$\sigma_{11} + \ldots + \sigma_{pp} = \text{Var}(X_1) + \ldots + \text{Var}(X_p) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \text{Var}(Y_1) + \text{Var}(Y_2) + \ldots + \text{Var}(Y_p) \tag{4}$$

Then the proportion of total population variance due to the kth principal component

$$= \lambda_k / (\lambda_1 + \lambda_2 + \ldots + \lambda_p) \qquad k = 1,2,\ldots, p. \tag{5}$$

Consequently, if most of the total population variance, for large p, can be attributed to the first two or three components, these can replace the original variables with a minimal loss of information.

When the variables have different ranges and are measured on different scales (as is the case with most materials problems), they are standardized

$$\begin{aligned} Z_1 &= (X_1 - \mu_1) / \sqrt{\sigma_{11}} \\ Z_2 &= (X_2 - \mu_2) / \sqrt{\sigma_{22}} \\ &\vdots \\ Z_p &= (X_p - \mu_p) / \sqrt{\sigma_{pp}} \end{aligned} \tag{6}$$

Then Cov $(\mathbf{Z}) = \boldsymbol{\rho}$ and Var $(Z_i) = 1$ and the principal components of $\mathbf{Z}$ are obtained from the eigenvectors of the correlation matrix $\boldsymbol{\rho}$ of $\mathbf{X}$. The corresponding eigenvalue-eigenvector pairs for ρ are $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Then the ith principal component is given by

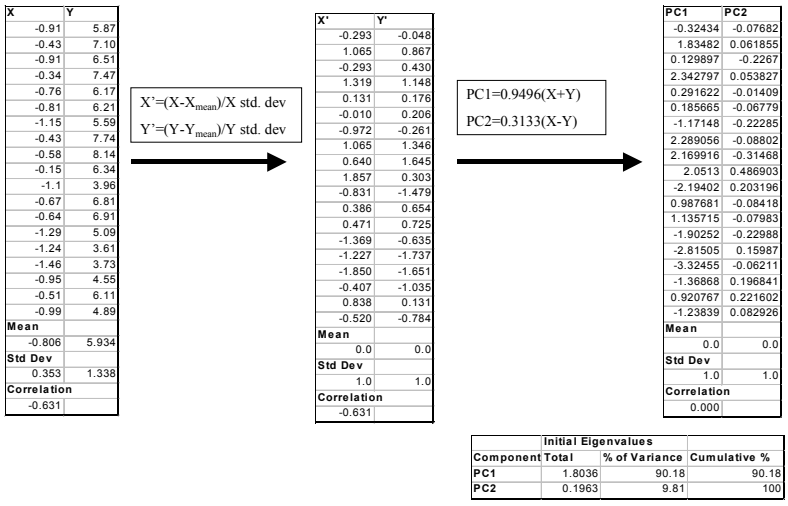
$$Y_i = \mathbf{e'}_i\mathbf{Z} = e_{i1}Z_1 + e_{i2}Z_2 + \ldots + e_{ip}Z_p, \quad i = 1,2, \ldots, p. \tag{7}$$

Then the proportion of total population variance due to the kth principal component $= \lambda_k / p$ where $k = 1,2,\ldots, p$ and λ_k's are the eigenvalues of $\boldsymbol{\rho}$.

The complex tasks of prediction and classification are made easier in this compressed space. PCA reduces the redundancy contained within the data by creating a new series of components in which the axes of the new coordinate systems point in the direction of decreasing variance. The resulting components are often more interpretable than the original data set (see for example the complexity of the combinatorial response map). For the purposes of this discussion, we shall focus on situations where a vast array of variables associated with single set of compounds or chemistry exists.

The information shown in the combinatorial response map provides the input for the PCA analysis (a typical example of the data calculations is shown below in Figure 2). An SPSS version 11 statistical

analysis package was used to preprocess, normalize and calculate the principal components on all the materials data.



X	Y
-0.91	5.87
-0.43	7.10
-0.91	6.51
-0.34	7.47
-0.76	6.17
-0.81	6.21
-1.15	5.59
-0.43	7.74
-0.58	8.14
-0.15	6.34
-1.1	3.96
-0.67	6.81
-0.64	6.91
-1.29	5.09
-1.24	3.61
-1.46	3.73
-0.95	4.55
-0.51	6.11
-0.99	4.89
Mean	
-0.806	5.934
Std Dev	
0.353	1.338
Correlation	
-0.631	

X'	Y'
-0.293	-0.048
1.065	0.867
-0.293	0.430
1.319	1.148
0.131	0.176
-0.010	0.206
-0.972	-0.261
1.065	1.346
0.640	1.645
1.857	0.303
-0.831	-1.479
0.386	0.654
0.471	0.725
-1.369	-0.635
-1.227	-1.737
-1.850	-1.651
-0.407	-1.035
0.838	0.131
-0.520	-0.784
Mean	
0.0	0.0
Std Dev	
1.0	1.0
Correlation	
-0.631	

PC1	PC2
-0.32434	-0.07682
1.83482	0.061855
0.129897	-0.2267
2.342797	0.053827
0.291622	-0.01409
0.185665	-0.06779
-1.17148	-0.22285
2.289056	-0.08802
2.169916	-0.31468
2.0513	0.486903
-2.19402	0.203196
0.987681	-0.08418
1.135715	-0.07983
-1.90252	-0.22988
-2.81505	0.15987
-3.32455	-0.06211
-1.36868	0.196841
0.920767	0.221602
-1.23839	0.082926
Mean	
0.0	0.0
Std Dev	
1.0	1.0
Correlation	
0.000	

	Initial Eigenvalues		
Component	**Total**	**% of Variance**	**Cumulative %**
PC1	1.8036	90.18	90.18
PC2	0.1963	9.81	100

Figure 2. An example of the data set calculations for the Principal Component Analysis.

Figures 3(a) & 3(b) shows the graphical three dimensional representation of the PCA analysis for one set of compounds, namely the intermetallic systems, which shows, particularly in Figure 3(b), the presence of a strong eigencomponent in the data set. This information for all sets of compound chemistries explored in the data set is summarized in Figure 4 and projected in two dimensions in this case. Figure 4(a) shows the data distribution of the PCA based on the temperature level of the superconducting transition. Figure 4(b) shows a similar type of plot, with the behavior classified according to crystal structure type. The new PCA scatter plots show that MgB_2 appears to be clustered around the response behavior space near other known superconductors.

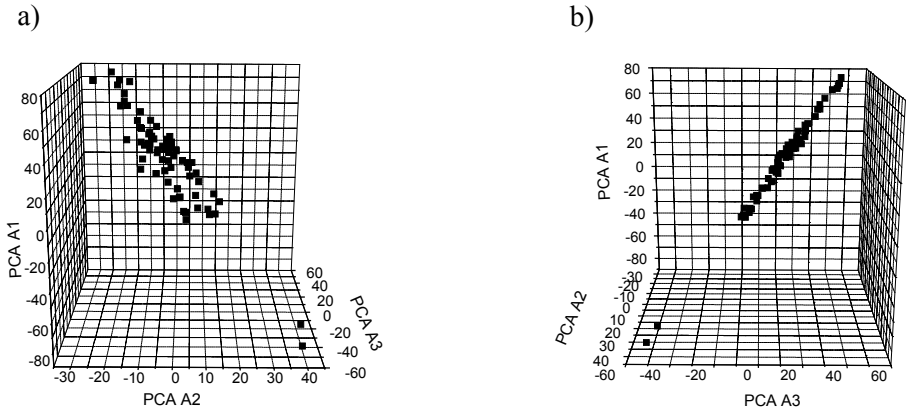


Figure 3. A15 compounds – showing strong eigencomponents in a 3-d PCA plot.

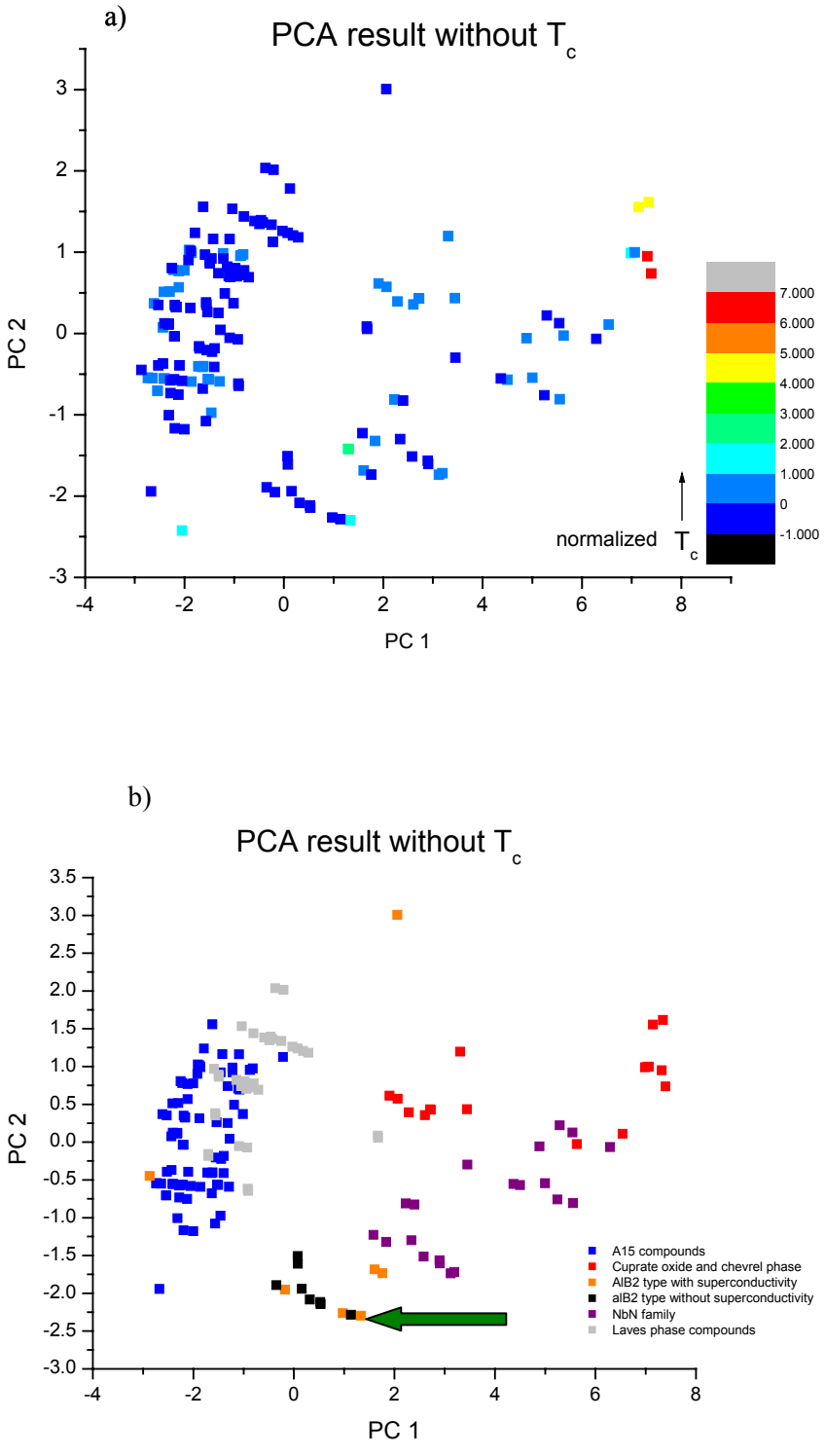


Figure 4. Two dimensional PCA plots showing clustering patterns of MgB_2 (arrowed) data compared with other superconductivity data sets. Note that the T_c data is deliberately not incorporated in this analysis to test the "discovery" capability of the PCA methodology to help one assess the potential of MgB_2 as a high temperature superconductor.

4 CONCLUSION

A key aspect of developing an "informatics" approach to materials discovery is the need to establish a critical array of descriptors of materials' attributes that may subsequently be input into a database. Having physically meaningful descriptors is key to developing and searching for associations between apparently disparate or disjointed datasets. Our initial work on this rapidly evolving field, has already provided potential descriptors that may indicate not only what materials may be worthwhile investigating further, but also strategies of how new chemistries must influence the structure if superconductivity is to be promoted. This in turn of course provides possible insights into the mechanisms that govern high temperature superconductivity in these new classes of materials. However we would also like data mining tools with high predictive accuracy, in order to identify materials likely to possess desirable properties from massive combinatorial libraries of materials. Thus material science presents a challenging testbed for the development of new algorithmic and mathematical foundations for integrating discovery and prediction in data mining.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the Air Force Office of Scientific Research, Contract No. F49620-01-1-0409.

6 REFERENCES

Bajorath, J. (2001) Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis and Virtual Screening. *J. Chem. Inf. Comput. Sci.* 41(2), 233-245.

Faloutsos, C., Korn, F., Labrinidis, A., Kotidis, Y., Kaplunuovich, A., & Perkovic, D. (1997) *Quantifiable Data Mining Using Principal Component Analysis*. Retrieved August 19, 2001 from the University of Maryland, Institute for Systems Research Website: http://www.isr.umd.edu/TechReports/ISR/1997/TR_97-25/TR_97-25.phtml

Imai, M. & Hirano,T. (1997) *In situ* measurements of the orthorhombic-to-trigonal transition in $BaSi_2$ under high-pressure and high-temperature conditions. *Phys.Rev.B* 55(1), 132-135.

Jolliffe, I.T. (1986) *Principal Component Analysis*, Berlin: Springer Verlag

Medvedeva, N.I., Ivanovskii,A.L., Medvedeva, J.E., & Freeman, A.J.(2001) Electronic structure of superconducting MgB_2 and related binary and ternary borides Website: http://arXiv.org/abs/cond-mat/0103157

Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitini, T., & Akimitsu, J. (2001) Superconductivity at 39K in magnesium boride. *Nature* 410, 63-64

Phillips, J.C. (1989) *Physics of High Tc superconductors* NY: Academic Press

Poole Jr, C.P., Datta, T., & Farach, H.A. (1988) *Copper Oxide Superconductors* NY: John Wiley & Sons

Poole Jr, C.P., Farach, H.A., & Creswick, R.J. (1995) *Superconductivity* NY: Academic Press

Preisendorfer, R.W. (1988) *Principal Component Analysis in Meterology and Oceanography,* Amsterdam: Elsevier

Satta, G., Profeta, G., Bernardini, F., Continenza, A., Massidda, S. (2001) Electronic and structural properties of superconducting diborides and calcium disilicide in the AlB_2 structure. Retrieved April 23, 2002 from arXiv.org e-Print archive: http://arXiv.org/abs/cond-mat/0102358

Shum H., Ikeuchi, K., & Reddy, P. (1995) *Principal Component Analysis with Missing Data and its Application to Polyhedral Object Modeling*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(9), 854-867.

Villars, P., & Phillips J.C. (1988) Quantum structural diagrams and high Tc superconductivity. *Phys. Rev. B* 37(4), 2345-2348.

Wichern, D. W., & Johnson, R.A., (2002) *Applied Multivariate Statistical Analysis*, 5th Edn., New Jersey: Prentice Hall

Wilker, W. (2001) *Principal Component Analysis of Manuscripts of the Gospel of John*. Retrieved August 15, 2001 from the Universität Bremen, Institut für Organische Chemie Website: http://www-user.uni-bremen.de/~wie/pub/Analysis-PCA.html