Procedia Computer Science

Volume 51, 2015, Pages 2417–2425

ICCS 2015 International Conference On Computational Science

Unsupervised Feature Extraction Using Singular Value Decomposition

Kourosh Modarresi
Adobe Inc., San Jose, U.S.
Stanford University, Stanford, U.S.
kouroshm@alumni.stanford.edu

Abstract

Though modern data often provides a massive amount of information, much of the insight might be redundant or useless (noise). Thus, it is significant to recognize the most informative features of data. This will help the analysis of the data by removing the consequences of high dimensionality, in addition of obtaining other advantages of lower dimensional data such as lower computational cost and a less complex model. Modern data has high dimension, sparsity and correlation besides its characteristics of being unstructured, distorted, corrupt, deformed, and massive. Feature extraction has always been a major toll in machine learning applications. Due to these extraordinary features of modern data, feature extraction and feature reduction models and techniques have even more significance in analyzing and understanding the data.

Keywords: Modern Data, Feature Reduction, Singular Value Decomposition, Regularization, principal Component Analysis

1 Introduction

Feature reduction and al its related topics of feature selection, feature extraction or dimension reduction have a long history in statistical analysis [22]. The main purpose for the application of feature reduction has been to reduce the amount of data and its complexity, to save time and cost and to make analysis more effective and simpler.

There have been a variety of algorithms and models used for feature reductions. Some of these algorithms include the ones using clustering, CUR, PCA/SVD [22].

In this work, a data matrix X of m by n dimensions is defined with its m rows called users (objects, observations, customers, Items, measurements, replications Records). The data matrix n columns are called features (variables, covariates, predictors, dimensions, attributes, factors, regressors, inputs, fields, and so on).

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.424

2 Feature Extraction and Dimensional Reduction for Modern Data

The enormity of the observed and collected modern data is of great advantage in machine learning, data mining, data analytics and all related fields. Apart from its massive size, modern data has some other unique features such as high dimensionality, high sparsity, non-Gaussian distribution, high correlation, unstructured format, and high frequency (data stream). Unfortunately, the commonly used phrase of “big data” cannot possibly describe the characteristics of “modern data”. Modern data possesses distinct features that forcing the data science and all its related areas to come up with innovative solution in dealing with today data.

These features of modern data have brought fundamental challenges for data science. Some of these difficulties are related to “data storage”, “data management”, “data platform architecture” and “data collection techniques”. Analytical and modeling challenges are perhaps the most serious ones in dealing with modern data. They include high computational cost and scaling related adjustments. The curse of dimensionality poses even more significant challenge in analyzing modern data. Some examples of these are;

1. Because of Curse of dimensionality, you may get patterns by chance
2. Learning from data is accompanied by overfitting
3. Random and noisy but false model validation

The modern data properties of high level of sparsity and high dimensionality introduce another major difficulty for data science since the definition of distance and hence similarities/dissimilarities are not well-defined for high dimensional and sparse data. Due to the fact the data is lying in high dimension and is also very sparse, data points tend to be places at the edges of the high dimensional space and thus the distances are very large. As a result, the usual distance metrics would compute data points distances to be maximum and all objects are rendered to be dissimilar.

Since Gaussian noise distribution is not a valid assumption for the modern data, many available models that are based on Gaussian distribution assumption, cannot be readily applied to the analysis and modeling of modern data.

Though, modern data has some advantageous properties that could help us in its modeling and analysis. Some of these helpful features are:

1. Concentration of measure
2. Existence of structure
3. Massive size
4. High Correlation
5. Rand deficiency of the data matrix X, i.e.,

   rank(X)<< min(m,n), indicating X is severely ill- conditioned.

2.1 Feature Selection as a Means for Data Analysis

The main task of machine learning is to make actionable and useful insights from the massive amount of modern data. The data has very large dimension representing much knowledge and information. In the past, the amount of data collected was limited and mostly gathered through controlled experimentations and it was carefully done so basically a targeted selected data would be observed or measured. The data gathered this way was mostly very relevant data to the analysis that the data was collected for. Today, though, we have access to huge volume of the data, in part since the cost of collecting the data is negligible.

Though, as a consequence we collect much of irrelevant (noise) data. But how do we know which part of data has hidden information knowledge about the metrics (phenomena) we are interested in and which part of the data does not. This is very crucial to recognize the portion of the data having impact on our metrics. This, in no small part, is important because of the curse of dimensionality which makes any analysis of data impossible. Other good reasons for this is to save storage, making analysis faster and more stable and to provide more intuitive understanding of the data.

As an example of using feature extraction in machine learning is to find out which of the users or visit sessions' attributes have the most significant with respect to the general areas of online targeting and campaign more efficient. By recognizing these important features, we could stress more resources and focus on them to make our online campaign more effective.

There are many methods for feature selections in machine learning, data mining and data analytics fields. These methods are divided into two classes of supervised and unsupervised feature selections. For unsupervised feature selection approaches, unlike the supervised ones, the data is not labeled in the sense that there is no specific output or particular application of data in mind when using feature selection models. Another division of feature selection methods is based on whether we select a new, though smaller number, of features that are combination of the old features or we select a subsection of the original features. The first methods are called "feature selection" and the latter types of approach is called "feature extraction". Feature extraction has the advantage of recognizing a smaller group of the original features and thus related directly to the original data. Feature selection methods use some combination of the original features, thus making interpretation of the new features non trivial. Both "feature reduction" approaches of "feature selection" and "feature extractions" are examples of data "dimension reduction".

In general, and for both types of feature reduction models, we have two criteria of correlation and variation of data in mind. In the sense that the reduced features must represents the correlation and or variation of the original data well. In choosing the best model and algorithm for feature reduction, there is often a tradeoff between variation- and –correlation representation of the original data in the new space. In other words the projection of the data onto a reduced dimension space considers the best projection or approximation of the data by having the variation and correlation of the data as the objective function to be optimized.

3 Description of the Feature Extraction Model Using SVD

In this section, we describe the feature extraction model that is based on principal component analysis of our data matrix X. PCA uses singular value decomposition of the centered X and thus is equivalent to SVD for the purposes of this work.

3.1 Singular Value Decomposition (SVD)

For any matrix X (m by n), SVD exists and is unique up to the signs. The singular value decomposition for the data matrix X is;

$$X = UDV^t$$

Where: U, the left singular vectors, is m×n orthogonal matrix,

$$UU^t = U^tU = I$$

V, the right singular vectors, is n×n orthogonal matrix

$$VV^t = V^tV = I$$

and D = diag $(d_1, d_2, \dots, d_n)$ with the singular vectors;

$$d_1 \geq d_2 \geq \cdots . \geq d_n \geq 0$$

Using a threshold (often it is 80%-90%) for the amount of original data variation explained by the new features leads to the selection of a small number (k) of the new features. These new features contain a weighted combination of all of the original features. Thus, SVD cannot be directly used for feature extraction because the new features (columns of U) combine the original features (columns of X). In this work, we impose rank constraints on the singular value decomposition to have only a selected limited number of nonzero factors in each new principal coordinates (columns of U). This will lead to extraction of original features with their significance. The model follows the following steps;

Step (1) Matrix completion: All missing values of matrix X is computed at this step using an iterative svd algorithm [60]. The algorithm has the following steps:

For a centered X ;

- Step (1.1) compute
  $\underbrace{min}_{U_q, V_q, D_q} \|X - U_q D_q V_q\|$ to obtain U_q , D_q and V_q
  For q= numerical rank of the matrix.
- Step(1.2) compute the rank –q of X;

$$X_q = U_q D_q V_q$$

using newly computed X_q, we have new values for the missing entries.

Step (1.3) Iterate steps (1.1) and (1.2) till convergence ;

$$\|X_{q(i+1)} - X_{q(i)}\| / \|X_{q(i)}\| \leq \delta$$

for small δ.

Step (2) Computing rank constrained SVD;

- Using Rank-1 approximation to our data matrix X;

$$\text{argmin}_{(u,v,\sigma)} \|X - \sigma u v^t\|_2^2 \quad s.t. \qquad \|v\|_2^2 = \|u\|_2^2 = 1$$

With the rank constraints;

$$min\ \|v\|_0 \text{ and } min\ \|u\|_0$$

- Since norm-zero computation is NP hard problem and thus not feasible, we use a surrogate constraint (second norm), or equivalently

$$\text{argmin}_{(u,v,\sigma)} \|X - \sigma u v^t\|_2^2 \;\; s.t. \;\; \|v\|_2^2 = \|u\|_2^2 = 1 \text{ with the constraints of;}$$

$$\|v\|_0 \le \delta \text{ and } \|u\|_0 \le \eta$$

Equivalently, Using Minimum Reconstruction Error [32] in Approximating X

$$\text{argmin}_{(u,v,\sigma)} \|X - Xvu^t\|_2^2 \;\; s.t. \;\; \|v\|_2^2 = \|u\|_2^2 = 1 \text{ also } min\, \|v\|_0 \text{ and } min\, \|u\|_0$$

- Which is equivalent to

$$\text{argmin}_{(u,v,\sigma)} \|X - Xvu^t\|_2^2 \;\; s.t. \;\; \|v\|_2^2 = \|u\|_2^2 = 1 \text{ also } \|v\|_0 \le \delta \text{ and } \|u\|_0 \le \eta$$

Similarly, since norm-zero computation is not tractable, a surrogate constraints of norm one is used. [32,33,60]

4 Results

The model in section 3 has been applied to two different data sets. The first example is a 2722×122 matrix containing the length of time spent by users on different sites (variables).

Figure 1 shows the most significant features with their significance.

The second example is a data matrix is of 75715×12 dimension containing the conversion of different ad campaigns for a variety of regions (variables). Figure 2 shows the most significant features with their significance for the second data set.

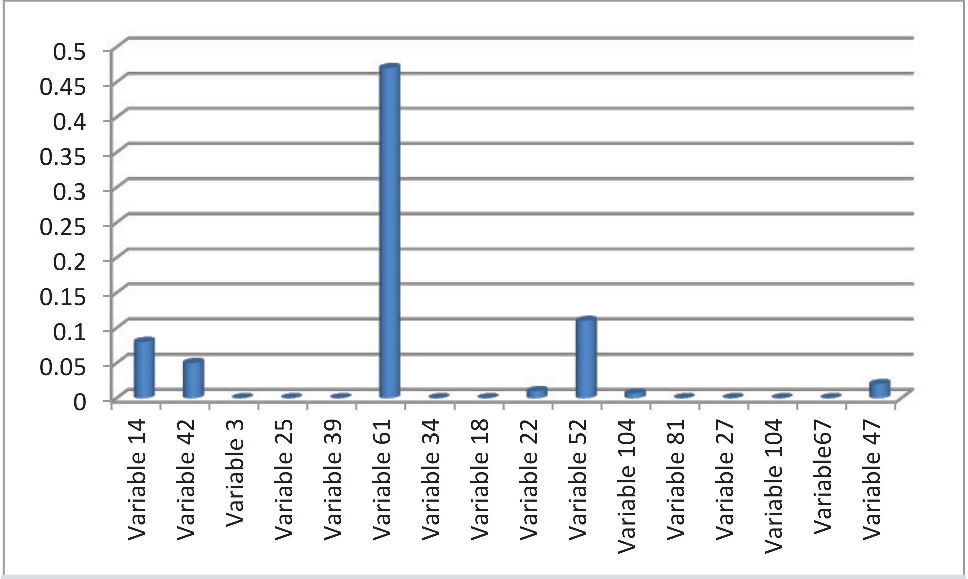

Figure 1. the most significant features with their significance for the data matrix in Example1.

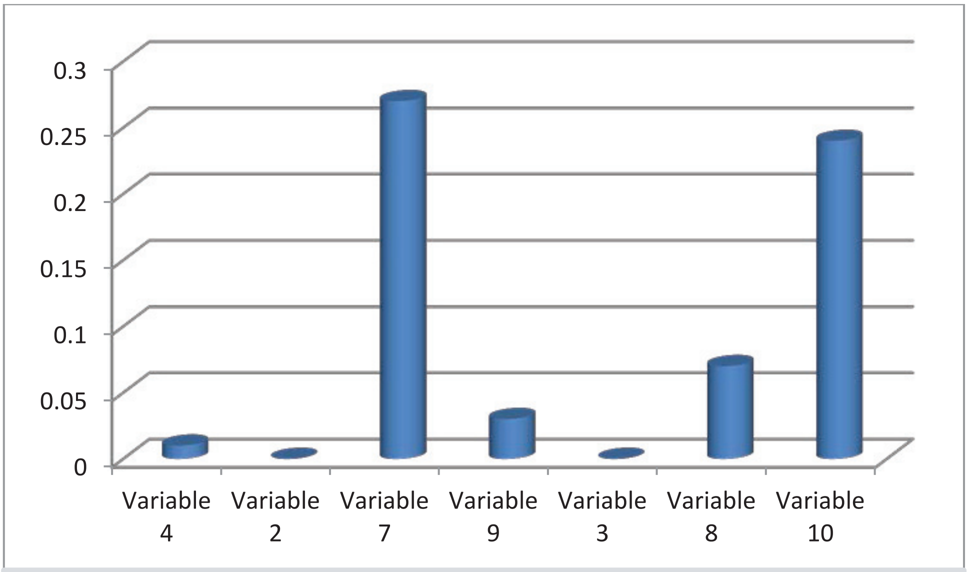

Figure 2. the most significant features with their significance for the data matrix in Example2.

The test is done by using the extracted features for SVD (PCA) computation. We see the difference between k-rank svd for both cases in terms of the relative error based of Euclidean difference between the new coordinates in both cases.

References

[1] A. Bjorck, "Numerical Methods for Least Squares Problems" ,SIAM, Philadelphia,1996.

[2] S. Boyd and L. Vandenberghe, "Convex Optimization", Cambridge University Press, 2004.

[3] P. A. Businger, G. H. Golub, "Singular value decomposition of a complex Matrix", Algorithm 358, Comm. Acm, No. 12, pp. 564-565, 1969.

[4] J. Cadima and I. T. Jolliffe, " Loadings and correlations in the interpretation of principal components", Journal of Applied Statistics, 22:203–214, 1995.

[5] E. J. Cand`es and T. Tao, "Decoding by linear programming", IEEE Transactions on Information Theory, 51(12):4203–4215, 2005.

[6] R. Courant and D. Hilbert, "Methods of Mathematical Physics", Vol. II, Interscience, New York, 1953.

[7] A. R. Davies and M. F. Hassan, "Optimality in the regularization of ill-posed inverse problems", in P. C. Sabatier (Ed.), Inverse Problems: An interdisciplinary study, Academic

Press, London, UK, 1987.

[8] B. DeMoor, G. H. Golub, "The restricted singular value decomposition: properties and applications", SIAM J. Matrix Anal. Appl., 12, No. 3, pp. 401-425, 1991.

[9] D. L. Donoho and J. Tanner, " Sparse nonnegative solutions of underdetermined linear equations by linear programming", Proc. of the National Academy of Sciences, 102(27):9446–9451, 2005.

[10] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., "Least Angle Regression," The Annals of Statistics, 32, 407–499, 2004.

[11] Lars Elden, "Algorithms for the Regularization of Ill-Conditioned Least Squares Problems", BIT 17, pp. 134-145, 1977.

[12] Lars Elden, "A Note on the Computation of the Generalized Cross-Validation Function for Ill-Conditioned Least Squares Problems", BIT 24, pp. 467-472, 1984.

[13] Heinz. W. Engl, M. Hanke, and A. Neubauer, "Regularization methods for the stable solution of inverse problems" , Surv. Math. Ind., No. 3, pp. 71-143, 1993.

[14] H. W. Engl, M. Hanke, and A. Neubauer, "Regularization of Inverse Problems", Kluwer, Dordrecht, 1996.

[15] H. W. Engl , C. W. Groetsch (Eds), "Inverse and Ill-Posed Problems", Academic Press, London, 1987.

[16] M. Fazel, H. Hindi, and S. Boyd. "A rank minimization heuristic with application to minimum order system approximation", Proceedings American Control Conference, 6:4734–4739, 2001.

[17] W. Gander, "On the linear least squares problem with a quadratic Constraint", Technical report STAN-CS-78-697, Stanford University, 1978.

[18] G. H. Golub, C. F. Van Loan, "Matrix Computations", 4th Ed., Computer Assisted Mechanics and Engineering Sciences, Johns Hopkins University Press, US, 2013.

[19] Gene H. Golub, Charles F. Van Loan, "An Analysis of the Total Least Squares Problem", Siam J. Numer. Anal., No. 17, pp. 883-893, 1980.

[20] Gene H. Golub, W. Kahan, "Calculating the Singular Values and Pseudo-Inverse of a Matrix", SIAM J. Numer. Anal. Ser. B 2, pp. 205-224, 1965.

[21] Gene H. Golub, Michael Heath, Grace Wahba, "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter", Technometrics 21, pp. 215-223, 1979.

[22] Hastie, T., Tibshirani, R., and Friedman, J. ," The Elements of Statistical Learning; Data mining, Inference and Prediction", New York: Springer Verlag, 2001.

[23] Hastie, T.J and Tibshirani, R. "Handwritten Digit Recognition via Deformable Prototypes", AT&T Bell Laboratories Technical Report, 1994.

[24] Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D., " 'Gene Shaving' as a Method for Identifying Distinct Sets of Genes With Similar Expression Patterns," Genome Biology, 1, 1–21, 2000.

[25] T. Hein, "Some analysis of Tikhonov regularization for the inverse problem of option pricing in the price-dependent case," ,SIAM Review, (21)No. 1, pp. 100-111, 1979.

[26] T. Hein and B. Hofmann, "On the nature of ill-posedness of an inverse problem in option pricing," ,Inverse Problems,(19), pp. 1319-11338, 2003.

[27] B. Hofmann, "Regularization for Applied Inverse and Ill-Posed problems ," Teubner, Stuttgart, Germany, 1986.

[28] B. Hofmann, "Regularization of nonlinear problems and the degree of illposedness," in G. Anger, R. Goreno, H. Jochmann, H. Moritz, and W Webers (Eds.), inverse Problems: principles and Applications in Geophysics,Technology, and Medicine, Akademic Verlag, Berlin, 1993.

[29] T. A. Hua and R. F. Gunst, "Generalized ridge regression: A note on negative ridge parameters," Comm. Statist. Theory Methods, 12, pp. 37-45, 1983.

[30] Jeffers, J., “Two Case Studies in the Application of Principal Component,” Applied Statistics, 16, 225–236, 1967.

[31] Jolliffe, I. , Principal Component Analysis, New York: Springer Verlag, 1986

[32] Jolliffe, Trendafilov, and Uddin (2003) ’A modified principal component technique based on the lasso’, Journal of Computational and Graphical Statistics 12 531-547.

[33] I. T. Jolliffe, N.T. Trendafilov, and M. Uddin, “A modified principal component technique based on the LASSO,” Journal of Computational and Graphical Statistics, 12:531–547, 2003.

[34] M. Journ´ee, Y. Nesterov, P. Richt´arik, and R. Sepulchre, “Generalized power method for sparse principal component analysis,” arXiv:0811.4724, 2008.

[35] Misha E. Kilmer and Dianne P. OLeary, “Choosing regularization parameters in iterative methods for ill-posed problems,” SIAM J. MATRIX ANAL. APPL., Vol. 22, No. 4, pp. 1204-1221. 2001.

[36] Andreas kirsch, “An Introduction to the Mathematical theory of Inverse problems ,” Springer Verlag, New York, 1996.

[37] Mardia, K., Kent, J., and Bibby, J., “Multivariate Analysis,” New York: Academic Press, 1979.

[38] McCabe, G., “Principal Variables,” Technometrics, 26, 137–144, 1984.

[39] Kourosh Modarresi and Gene H Golub, “An Adaptive Solution of Linear Inverse Problems”, Proceedings of Inverse Problems Design and Optimization Symposium (IPDO2007), April 16-18, Miami Beach, Florida, pp. 333-340, 2007.

[40] Kourosh Modarresi, “A Local Regularization Method Using Multiple Regularization Levels”, Stanford, CA, April 2007.

[41] Kourosh Modarresi and Gene H Golub, “An Efficient Algorithm for the Determination of Multiple Regularization Parameters,” Proceedings of Inverse Problems Design and Optimization Symposium (IPDO), April 16-18, 2007, Miami Beach, Florida, pp. 395-402, 2007.

[42] D. W. Marquardt, “Generalized inverses, ridge regression, biased linear estimation,” and nonlinear estimation, Technometrics, 12, pp. 591-612, 1970.

[43] K. Miller, “Least Squares Methods for Ill-Posed Problems with a prescribed bond,” SIAM J. Math. Anal., No. 1, pp. 52-74, 1970.

[44] B. Moghaddam, Y. Weiss, and S. Avidan, “Spectral bounds for sparse PCA: exact and greedy algorithms,” Advances in Neural Information Processing Systems, 18, 2006.

[45] V. A. Morozov, “On the solution of functional equations by the method of regularization”,Sov. Math. Dokl., 7, pp. 414-417, 1966.

[46] V. A. Morozov, “Methods for Solving Incorrectly Posed Problems, “ Springer-Verlag, New York, 1984.

[47] B. K. Natarajan, “Sparse approximate solutions to linear systems,” SIAM J. Comput., 24(2):227–234, 1995.

[48] R. L. Parker , “Understanding inverse theory,” Ann. Rev. Earth Planet. Sci., No. 5, pp. 35-64, 1977.

[49] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: transfer learning from unlabeled data. In 24th International Conference on Machine learning, 2007.

[50] T. Raus, “The principle of the residual in the solution of ill-posed problems with nonselfadjoint operator,” Uchen. Zap. Tartu Gos. Univ., 75, pp. 12-20, 1985.

[51] T. Reginska, “A Regularization Parameter in Discrete Ill-Posed Problems,” SIAM J. Sci. Comput., No. 17, pp. 740-749, 1996.

[52] A. Tarantola and B. Valette , “Generalized nonlinear inverse problems solved using the least squares criterion,” Reviews of Geophysics and Space Physics, No. 20, pp. 219-232 , 1993.

[53] Tibshirani, R., “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B, 58, 267–288, 1996.

[54] A. N. Tikhonov, “Solution of Incorectly Formulated Problems and the Regularization Method,” Soviet Math. Dokl., 4(1963), pp. 1035-1038; English translation of Dokl. Akad. Nauk. SSSR, 151(1963), pp. 501-504, 1963.

[55] A. N. Tikhonov, “Regularization of incorrectly posed problems,” Dokl. Akad. Nauk. SSSSR, 153, (1963), pp. 49-52= Soviet Math. Dokl., 4, 1963.

[56] A. N. Tikhonov, V. Y. Arsenin, “Solutions of Ill-Posed Problems,” Winston, Washington, D.C.(1977).

[57] A. N. Tikhonov, A. V. Goncharsky(Eds), “Ill-Posed Problems in the Natural Sciences,”,MIR, Moscow, 1987.

[58] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, A. G. Yagola, “Numerical Methods for the Solution of Ill-Posed Problems,” Kluwer, Dordrecht, the Netherlands, 1995.

[59] Trendafilov and Jolliffe (2006) ’Projected gradient approach to the numerical solution of the SCoTLASS’, Computational Statistics and Data Analysis 50 242-253.

[60] Witten, Tibshirani and Hastie (2009) “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”, Biostatistics (2009), 10, 3, pp. 515–534, 2009 .

[61] Z. Zhang, H. Zha, and H. Simon, “Low rank approximations with sparse factors I: basic algorithms and error analysis,” SIAM journal on matrix analysis and its applications, 23(3):706–727, 2002.

[62] H. Zou, T. Hastie, and R. Tibshirani, “Sparse Principal Component Analysis,” Journal of Computational & Graphical Statistics, 15(2):265–286, 2006.