High Dimensional Big Data and Pattern Analysis: A Tutorial

Choudur K. Lakshminarayan

HP Software R&D, USA
Choudur.Lakshminarayan@hp.com

Abstract. Sensors and actuators embedded in physical objects being linked through wired/wireless networks known as "*internet of things*" are churning out huge volumes of data (McKinsey Quarterly report, 2010). This phenomenon has led to the archiving of mammoth amounts of data from scientific simulations in the physical sciences and bioinformatics, to social media and a plethora of other areas. It is predicted that over 30 billion devices with 200 billion intermittent connections will be connected by 2020. The creation and archival of the massive amounts of data spawned a multitude of industries. Data management and up-stream analytics is aided by data compression and dimensionality reduction. This review paper will focus on some foundational methods of dimensionality reduction by examining in extensive detail some of the main algorithms, and points the reader to emerging next generation methods that seek to identify structure within high dimensional data not captured by 2nd order statistics.

Keywords: Multivariate Analysis, Dimensionality Reduction**,** Projections**,** Principal Component Analysis, Factor Analysis, Canonical correlation Analysis, Independent Component Analysis, Exploratory Projection Pursuit.

1 Introduction

Needless to say, "Big Data" is the next frontier in scientific exploration and advancement. A recent report by a National Academy of Sciences commissioned study examines the challenges and opportunities [1]. The range of problems includes back-end systems for the ingestion and storing of large volumes of structured, semi-structured, and unstructured data (static and time-aware) in various formats and sources, to query engines, and *analytics* operations in the front-end. The volume to be processed at the speed of business requires parallel computing by distributed processing, a common set of analytical methods for repeatable analyses and rewriting existing algorithms to adapt to scale.

As the number of variables which purportedly describe a phenomenon, as well as frequency of sampling keeps increasing, it has become a challenge to tease out that subset of variables which indeed capture the dynamics and structure of the underlying phenomenon. Towards that end, data reduction techniques have become the *mainstay* of statistical data pre-processing. So we provide a tutorial review of some of the

V. Bhatnagar and S. Srinivasa (Eds.): BDA 2013, LNCS 8302, pp. 68–85, 2013.

foundational methods in dimensionality reduction in detail and point the reader to the next generation of algorithms. The field of dimensionality reduction is vast, and so we limit the scope of the paper to popular dimensionality reduction techniques such as *principal component analysis* (PCA), *Factor Analysis* (FA), *Canonical Correlation Analysis* (CCA) *Independent Component Analysis* (ICA), and *Exploratory projection Pursuit* (PP). We have chosen these methods because the vast majority of practitioners utilize them in daily applications. In this tutorial, we will study PCA, FA, CCA, ICA, and PP and relationships among them in some detail.

Since dimension reduction is not only desirable, but paramount, how should we go about it? Perhaps, a simple approach is to find a lower-dimensional embedding in which the data truly resides, while eliminating extraneous variance. This can be achieved by projecting the data by linear transformations into lower dimensional subspaces by maximizing a suitable objective function. This *genre* of algorithms is known as *projective methods* [13]. In the transformed domain the data is more interpretable as non-informative sources of variation can be eliminated, while retaining principal directions of variance. The other approach to dimensionality reduction is to exploit polynomial moments to unravel the hidden structure in the data. Well known projective methods are based on the covariance (2^{nd} order cross moments) which only capture the linear structure in the data. Methods that go beyond 2^{nd} order moments are exploratory projection pursuit, independent component analysis, and principal curves and surfaces [7,8]. A class of methods known as *manifold learning*, extract low-dimensional structure from high dimensional data in an unsupervised manner. These techniques typically try to unfold the underlying manifold ($\mathcal{M}$) into a lower dimensional space so that Euclidean distance in the new space is a meaningful measure of distance between pairs of points [16]. These methods have implications in making clustering methods more effective in the transformed space. In the following sub-sections, we will briefly introduce the techniques presented in this paper. The contents of the paper assume that the reader is familiar with elementary linear algebra, elementary probability theory, mathematical statistics, and multivariate analysis.

1.1 Principal Component Analysis

Principal components analysis (PCA) is one member of a family of methods for dimensionality reduction. It is a technique that involves transformations of set of variables into a smaller set of uncorrelated variables, while retaining intrinsic information in the original data set by exploiting correlations among the variables [2,3,15]. PCA is merely a linear projection of a set of observed variables on to basis vectors which turn out to be Eigen vectors under the average mean square error (MSE) objective function. PCA is one of the simplest and most common ways of doing dimensionality reduction. It is also one of the oldest, and has been variously alluded to in many fields as the Karhunen-Loève transformation (KLT) in communications, the Hotelling transformation, and latent semantic indexing (LSI) in text mining. But the *moniker* principal component analysis is the most popular.

1.2 Factor Analysis

Factor Analysis (FA) is a technique to find relationships between a set of observed variables and set of *latent* factors. The Factor analytic model is based entirely on the covariance matrix of the observed variables like the PCA models we studied in an earlier section. The key idea behind factor analysis is that multiple observed variables have similar patterns of responses because of their association with an underlying set of latent variables; the factors, which cannot be easily measured. For example, responses to questions about occupation, education, and income, are all associated with the latent variable socioeconomic stratum. In a factor analysis model, the number of factors always equal to the number of variables. Each factor contributes to a certain amount of the overall variance in the observed variables. The factors are then arranged in the decreasing order of variance explained. In a factor analytic model, each observed variable $\{X_i\}_{i=1}^{p}$is expressed as a sum of latent factors$\{F_i\}_{i=1}^{p}$, known as *common* factors, and an error term $\{\epsilon_i\}_{i=1}^{p}$, known as *specific* variance. The specific variance accounts for the unexplained variance in the observed variable. Mathematically, the observed variables are projected onto a set of basis vectors $\{\ell_{ij}\}_{i,j=1}^{p}$known as *loadings* in the FA literature. Under some assumptions on the latent factors, the loadings are the Eigen vectors obtained by decomposing the covariance matrix Σ. Typically spectral decomposition [2] is applied to Σ to obtain Eigen value, Eigen vector pairs$\{\lambda_i, \vec{l_i}\}_{i=1}^{p}$. The Eigen value is a measure of how much of the variance in the observed variables a factor explains. Any factor with an Eigen value ≥1 explains more variance than any single observed variable. In the exploratory mode, FA can be used to subset similar variables by examining the factor loadings on the original observed variables.

1.3 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) proposed by Professor Harold Hotelling in 1936 is a method for exploring linear relationships between two sets of multivariate variables (vectors), measured on the same object/entity. It finds two bases, one for each variable set such that the correlation between the inner products (linear projections) between the two bases and the two variable sets is maximized. The dimensionality of these new bases is less than or equal to the smallest dimensionality of the two variables. Succinctly, CCA reduces pairs of high dimensional variables into a smaller set of linear combinations which are more amenable for interpretation.

1.4 Independent Component Analysis (ICA)

All the methods introduced are based on 2nd order statistics (correlation structure). Independent component analysis (ICA) [9] in contrast attempts to reduce dependencies among higher order moments thereby increasing statistical independence among the original variables. It is a technique for identifying an underlying set of hidden factors from a multivariate set of random variables. It is among one of the popular techniques used in blind source separation [9]. An observed set of observations are assumed to be linear mixtures of hidden latent factors, and the mixing coefficients are unknown. ICA departs from previously known projective methods in that it assumes that it exploits higher order moments beyond the 2^{nd} order for identifying unknown factors (components) of the hidden mixture, besides assuming non-Gaussian distributions generating the data. In one version of the ICA, the latent sources are assumed to be non-Gaussian and independent. The objective function to estimate the unknown coefficients is parametrized by a likelihood function. The non-linear log-likelihood is then used to estimate the unknown coefficients by any of the stochastic gradient methods [11].

1.5 Projection Pursuit (PP)

Projection pursuit seeks to identify hidden structure in high dimensional data by using projections in lower dimensions that capture interesting features. The *interestingness* is determined by a numerical index known as the projection index. Techniques such as PCA, FA, and CCA depend on rotation, and scaling, to obtain linear projections. If the data vector $X \in \mathbb{R}^p$ observes a certain probability law, their sum $\vec{x}'\vec{w}$ would follow a Gaussian distribution by the central limit theorem [5]. And it is well known that the Gaussian is fully specified when the first two moments (mean, and covariance) are known. So, these methods capture only the linear structure in the data. Projection pursuit seeks to unravel the non-linear hidden structure by leveraging polynomial moments, and it is in this sense that PP departs from other projective methods.

2 Projective Methods and Dimensionality Reduction

2.1 Principal Component Analysis

Principal Component Analysis involves linear combinations of the p features x_1, x_2,....,x_p of an input pattern vector that are mean-centered. Geometrically, the linear combinations are obtained by rotating the original system with features x_1, x_2,....,x_p as the coordinate axes, thereby resulting in a new rotated coordinate system. The axes of the rotated coordinate system represent the directions with maximum variability. This lets elimination of low-variability coordinate axes to reduce the dimensionality of the original data. Although p principal components are required to account for the total system variability, majority of the variation is captured by a smaller number m. The m principal components can then replace the original p features. Thus, the original data set consisting of p features with n measurements each is replaced by p principal components with n measurements each. Thus principal components are vectors that span a lower m dimensional subspace. Material for this section has been adapted from [2,3,6,15] and the reader is encouraged to refer to these references.

Consider the random vector $X = (x_1, x_2, \cdots, x_p)$ with covariance matrix Σ whose Eigen values are $\{\lambda_i\}_{i=1}^{p}$, where each λ_i is ≥ 0. Let $\{z_i\}_{i=1}^{p}$, be a set of vectors obtained by composing linear combinations of the original features. Mathematically they are given as:

$$z_1 = w_{11}x_{11} + w_{21}x_{12} + \cdots + w_{p1}x_{1p}\,, w_{11}x_{21} + w_{21}x_{22} + \cdots + w_{p1}x_{2p}\,, \cdots w_{11}x_{n1} + w_{21}x_{n2} + \cdots + w_{p1}x_{np}$$

$$z_2 = w_{12}x_{11} + w_{22}x_{12} + \cdots + w_{p2}x_{1p}\,, w_{12}x_{21} + w_{22}x_{22} + \cdots + w_{p2}x_{2p}\,, \cdots, w_{12}x_{n1} + w_{22}x_{n2} + \cdots + w_{p2}x_{np}$$

$$\vdots$$

$$z_p = w_{1p}x_{11} + w_{2p}x_{12} + \cdots + w_{pp}x_{1p}, w_{1p}x_{21} + w_{2p}x_{22} + \cdots + w_{pp}x_{2p}, \cdots w_{1p}x_{n1} + w_{2p}x_{n2} + \cdots + w_{pp}x_{np}$$

Where $z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})$ is the i^{th} linear combination.

What we notice above is that each feature vector, say$\vec{x}_1 = (x_{11}, x_{12}, \cdots, x_{1p})$ is projected onto a vector $\vec{w} = (w_{11}, w_{21}, \cdots, w_{p1})$ given by; $\vec{x}_1{}'\vec{w}$ which is a simple inner product. The vector $\vec{w}$ is such that $\vec{w}'w = 1$. It is clear that the mean of the vectors $\{z_i\}_{i=1}^{p}$is 0 since the x's are mean-centered. Consider for example, the vector z_i. The mean $\bar{z} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n} w_{j1}x_{ij}}{n} = \sum_{j=1}^{p} w_{j1} \frac{\sum_{i=1}^{n} x_{ij}}{n}$ =0 as the x's are mean-centered.

In matrix form, the linear combinations $\{z_i\}_{i=1}^{p}$can be written as $Z = [XW]'$

The variance of Z can be expressed in matrix form as:

$$\sigma_z^2 = \frac{1}{n}[XW]'[XW] \xrightarrow{yields} W'\frac{X'X}{n}W \tag{1}$$

To derive principal components from linear combinations (projections), we invoke the notion of average *mean square error* (MSE). That is; we are searching for those projections that have the smallest mean square distance between the original feature vectors and their projections. Mathematically, we want to choose the vector $\vec{w}$ such that the variance σ_z^2 is minimized. To find the $\vec{w}$ that maximizes the variance (σ_z^2), we utilize constrained optimization by Lagrange multipliers [17]. Maximize σ_z^2 subject to $\vec{w}'\vec{w} = 1$

$$\mathcal{L}(\vec{w}, \lambda) = \sigma_z^2 - \lambda(\vec{w}'\vec{w} - 1) \tag{2}$$

$$\frac{\partial \mathcal{L}(\vec{w}, \lambda)}{\partial \vec{w}} = 2\frac{X'X}{n}\vec{w} - 2\lambda\vec{w} \tag{3}$$

$$\frac{\partial \mathcal{L}(\vec{w}, \lambda)}{\partial \lambda} = \vec{w}'\vec{w} - 1 \tag{4}$$

Let $\frac{X'X}{n} = S$ and setting $\frac{\partial\mathcal{L}(\vec{w},\lambda)}{\partial\vec{w}} = 0$ implies $S\vec{w}= \lambda\vec{w}$, the characteristic equation that links Eigen values and Eigen vectors. Therefore, the desired vector ($\vec{w}$) is the Eigen vector of the covariance matrix (S). These maximizing Eigen vectors will be associated with the largest Eigen values(λ). Since S is a covariance matrix, it is symmetric and positive definite. A matrix is said to be positive definite, if $\vec{x}'S\vec{x} > 0$ for any $\vec{x}$. It is well known that a symmetric, positive definite matrix has positive Eigen values and the corresponding Eigen vectors are orthogonal. The first principal component is the axis along which the data has the most variance, and corresponds to a projection on the Eigen vector with the largest Eigen value. Similarly, the 2nd principal component is the axis with the 2nd largest variance and is associated with the with the Eigen vector with 2nd largest Eigen value, and so on. And we obtain p principal components as the covariance matrix is of order (pxp). Since the Eigen vectors are orthogonal, the projections (principal components) are all uncorrelated with each other. As each principal component captures proportion of variance in the data along its axis; those components corresponding to low variance may be dropped. Thus a set of $q \ll p$ fewer components may be chosen, resulting in dimensionality reduction. In a practical setting, the Eigen values obtained by solving, $S\vec{w}= \lambda\vec{w}$, are given by; $\{\lambda_i\}_{i=1}^{p}$are ordered. The ordered set of Eigen values, from the smallest to the largest are$\left(\lambda_{(1)}, \lambda_{(2)}, \cdots, \lambda_{(p)}\right)$. The variance explained by each successive principal component is obtained by calculating the ratio;

$$\psi = \frac{\sum_{j=1}^{i}\lambda_j}{\sum_{j=1}^{p}\lambda_j}, j = 1,2,\cdots,p. \qquad (5)$$

When ψ exceeds 0.8 (say), then the number of principal components is equal to i for which 0.8 is attained. The number 0.8 corresponds to 80% of variance in the data. The experimenter is at liberty to choose the cut-off value appropriate for the application$(x\%)$. Many times, a graph known as the *scree* plot is drawn to select the appropriate number of principal components. The scree plot is merely a graph, where the X-axis represents the numbers $(1,2,\cdots,p)$-which are the indexes of the Eigen values and the Y-axis is the cumulative variance(ψ). The number of principal components is chosen by locating the elbow in the curve beyond which the additional variance is negligible. Fig. 1 is an illustration of a scree plot. The vertical arrow marks the point when the cumulative variance stabilizes (flattens).

In conclusion, while PCA is a useful data reduction technique, care should be exercised in extracting meaning out of the components, which are simply linear projections of the original data. If the end goal is to classify high dimensional objects/entities to one of a several classes (as in a classification problem), using PCA for data reduction is *fair game* and perhaps required.

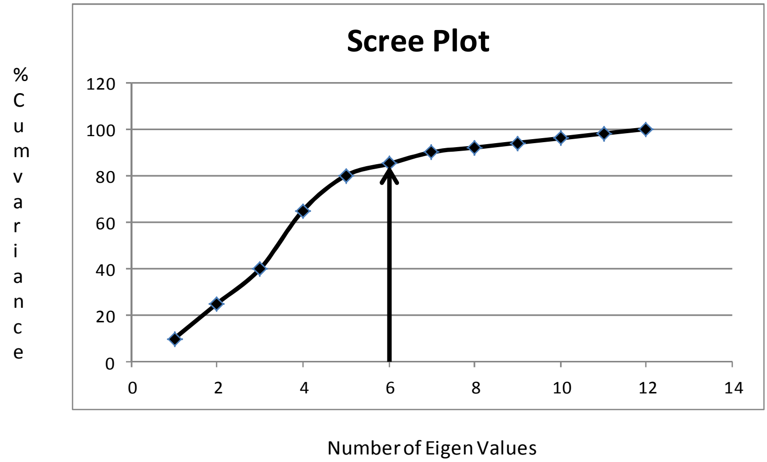


Fig. 1. Graph of Cumulative Variance versus Number of Eigen Values

2.2 Canonical Component Analysis

Canonical Correlation Analysis (CCA) proposed by Professor Harold Hotelling in 1936 is a method to correlate two different set of variables by projective transformations [4, 14]. It seeks to reduce pairs of high dimensional vectors into a few pairs of highly-correlated linear combinations of vectors known as *canonical* variables. Thus CCA can be construed as a feature reduction technique, while its origins was in being able to find relationships between manifestly different sets of variables, such as those related to government policies and economic impact. Operationally, CCA involves projecting each set of multi-dimensional variables $(\vec{x}, \vec{y})$onto basis vectors $(\vec{w}_x, \vec{w}_y)$ such that the correlation measure (ρ) between the projected vectors is maximized. The projections are given by $\vec{x}'\vec{w}_x$ and $\vec{y}'\vec{w}_y$. The idea is to find pairs of linear projections that are maximally correlated. The next iteration, we find those projections that are maximally correlated, but uncorrelated with the first pair, and the procedure continues until we find correlated projections that are uncorrelated with the predecessor pairs. Fig. 2 is a pictorial representation of Canonical Correlation Analysis. Two

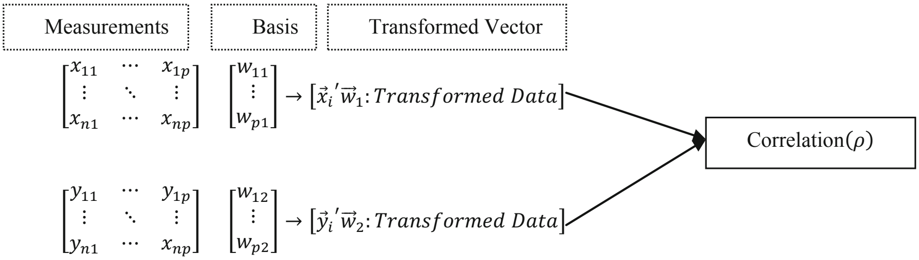

Measurements
Basis
Transformed Vector
$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$
$\begin{bmatrix} w_{11} \\ \vdots \\ w_{p1} \end{bmatrix} \rightarrow [\vec{x}_i{}'\vec{w}_1 : Transformed\ Data]$
Correlation(ρ)
$\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & x_{np} \end{bmatrix}$
$\begin{bmatrix} w_{12} \\ \vdots \\ w_{p2} \end{bmatrix} \rightarrow [\vec{y}_i{}'\vec{w}_2 : Transformed\ Data]$

Fig. 2. A Pictorial representation of Canonical Correlation Analysis

sets of variables (X, Y) are projected on to two basis vectors $\left(W_x, W_y\right)$ to yield linear projections, $\left(X'W_x, Y'W_y\right)$. We then determine the optimal values of the basis vectors; $\left(W_x, W_y\right)$ that maximizes the correlation (ρ) between the projections.

Operationally, we are seeking vectors $\vec{w}_x$ and $\vec{w}_y$ such that:

$$\rho = \max_{\vec{w}_x, \vec{w}_y} \frac{E\left(\vec{x}'\vec{w}_x\,\vec{y}'\vec{w}_y\right)}{\sqrt{E(\vec{x}'\vec{w}_x)^2 E\left(\vec{y}'\vec{w}_y\right)^2}} \tag{6}$$

The above expression can be further simplified for simplicity of expression by simple algebraic manipulations.

$$\rho = \max_{\vec{w}_x, \vec{w}_y} \frac{E\left(\,\vec{w}_x{}'\vec{x}\,\vec{y}'\vec{w}_y\right)}{\sqrt{E\left(\vec{w}_x{}'\vec{x}\vec{x}'\vec{w}_x\right)E\left(\overrightarrow{\vec{y}'\vec{w}_y\vec{y}'\vec{w}_y}\right)}} \xrightarrow{yields} \max_{\vec{w}_x, \vec{w}_y} \frac{\vec{w}_x{}'E(\vec{x}\,\vec{y}')\vec{w}_y}{\sqrt{\vec{w}_x{}'E(\vec{x}\vec{x}')\vec{w}_x E\left(\vec{w}_y{}'E(\vec{y}\vec{y}')\vec{w}_y\right)}} \tag{7}$$

$$\rho = \max_{\vec{w}_x, \vec{w}_y} \frac{\vec{w}_x{}'\Sigma_{xy}\vec{w}_y}{\sqrt{\vec{w}_x{}'\Sigma_{xx}\vec{w}_x\vec{w}_y{}'\Sigma_{yy}\vec{w}_y}} \tag{8}$$

where $\Sigma_{xx}, \Sigma_{yy}, \Sigma_{xy}$ are respectively, the variances and covariance between the random variables .

The maximum of ρ is the canonical correlation obtained by maximizing over $\left(\vec{w}_x, \vec{w}_y\right)$. We note that the canonical correlation is invariant to scaling the basis by a constant (c). This can easily seen by re-scaling to $c\vec{w}_x, c\vec{w}_y$ and substituting in (8). Thus we maximize the canonical correlation subject to the constraints $\vec{w}_x{}'\Sigma_{xx}\vec{w}_x = 1$ and $\overrightarrow{\vec{y}'\vec{w}_y y'}\vec{w}_y = 1$. Since we are seeking an optimization solution under constraints above, the Lagrangian formulation is as follows:

$$\mathcal{L}\left(w_x, w_y, \lambda\right) = \vec{w}_x{}'\Sigma_{xy}\vec{w}_y - \frac{\lambda_x}{2}\left(\vec{w}_x{}'\Sigma_{xx}\vec{w}_x - 1\right) - \frac{\lambda_y}{2}\left(\vec{w}_y{}'\Sigma_{yy}\vec{w}_y - 1\right) \tag{9}$$

Finding the derivatives of $\mathcal{L}(\cdot)$with respect to $\vec{w}_x, \vec{w}_y$and setting them equal to zero yields;

$$\frac{\partial\mathcal{L}\left(w_x, w_y, \lambda_x, \lambda_y\right)}{\partial w_x} = \Sigma_{xy}\vec{w}_y - \lambda_x\Sigma_{xx}\vec{w}_x = 0 \tag{10}$$

$$\frac{\partial\mathcal{L}\left(w_x, w_y, \lambda_x, \lambda_y\right)}{\partial w_y} = \Sigma_{xy}\vec{w}_x - \lambda_y\Sigma_{yy}\vec{w}_y = 0 \tag{11}$$

To solve this system of linear equations, we multiply, (10) by $\vec{w}_x{}'$ and the (11) by $\vec{w}_y{}'$ yielding,

$$\vec{w}_x{}'\Sigma_{xy}\vec{w}_y - \lambda_x\vec{w}_x{}'\Sigma_{xx}\vec{w}_x = 0 \tag{12}$$

$$\vec{w}_y{}'\Sigma_{xy}\vec{w}_x - \lambda_y\vec{w}_y{}'\Sigma_{yy}\vec{w}_y \ = 0 \tag{13}$$

Subtracting, the (13) from the (12), gives:

$$(\lambda_y - \lambda_x)(\vec{w}_y{}'\Sigma_{yy}\vec{w}_y - \vec{w}_x{}'\Sigma_{yy}\vec{w}_x) = 0 \tag{14}$$

Applying the constraints, $\vec{w}_x{}'\Sigma_{xx}\vec{w}_x = 1$ and $\vec{w}_y{}'\Sigma_{xx}\vec{w}_y = 1$, we obtain, $\lambda_y = \lambda_x$.

Also from (10) we get $\vec{w}_x = \frac{\Sigma_{xx}{}^{-1}\Sigma_{xy}\vec{w}_y}{\lambda_x}$. Substituting this value of $\vec{w}_x$ in in (11) we have:

$$\frac{\Sigma_{xy}{}'\Sigma_{xx}{}^{-1}\Sigma_{xy}}{\lambda_x}\vec{w}_y = \lambda_y\Sigma_{yy}\vec{w}_y \xrightarrow{yields}$$

$$(\Sigma_{xy}{}'\Sigma_{xx}{}^{-1}\Sigma_{xy})\vec{w}_y{=}\lambda_x\lambda_y\Sigma_{yy}\vec{w}_y \xrightarrow{yields} \lambda^2\Sigma_{yy}\vec{w}_y \tag{15}$$

since $\lambda_y = \lambda_x = \lambda$. Note that I use $\Sigma_{xy'}$ instead of Σ_{xy} in (15) as it is a symmetric matrix.

$(\Sigma_{xy}{}'\Sigma_{xx}{}^{-1}\Sigma_{xy})\vec{w}_y{=}\ \lambda^2\Sigma_{yy}\vec{w}_y$ in (15) is reminiscent of an Eigen equation. It is known as a generalized Eigen equation. This can be reduced to the form $Ay = \lambda y$, by noting that the matrix Σ_{yy} is symmetric and positive definite and can be expressed as the product $L_{yy}L_{yy}{}'$ (Cholesky Decomposition) [11]. Also, let $\vec{u}_y {=} L_{yy}{}'\vec{w}_y$ and re-writing (15), we have:

$$\Sigma_{xy}{}'\Sigma_{xx}{}^{-1}\Sigma_{xy}(L_{yy}^{-1})'\vec{u}_y = \lambda^2 L_{yy}{}'\vec{w}_y = \lambda^2\vec{u}_y \tag{16}$$

Clearly, this of the form $Ay = \lambda y$ is the Eigen equation seen in standard linear algebra! The Eigen equation can be used to find the $(\vec{w}_y, \vec{w}_x)$ to find the co-ordinate system that optimizes the correlation between the linear combinations of the two sets of vectors $(\vec{x}, \vec{y})$. To apply the theory developed above in a practical setting, the unknown population quantities, Σ_{xx}, Σ_{yy}, and $\Sigma_{xy}$are replaced by their sample counterparts, S_{xx}, S_{yy}, and S_{xy} respectively. Alternatively, one may use the correlation matrices, R_{xx}, R_{yy}, and R_{xy} as the roots/solutions derived from the application of the two representationss is the same.

In conclusion, CCA can be applied to large data sets where the correlations among the linear combinations of two sets of variables may reveal latent structures in the data that may not be captured by pair-wise correlations among the original variables. An added advantage of CCA is that it can identify relationships among observed variables and underlying latent factors/motivations. For example, it is applied in marketing analytics to understand the relationship between pricing (observed variable) and attributes such as form factor, ease of using, appeal, and other features (latent factors) attractive to a target segment of the market.

2.3 Factor Analysis

Consider a random vector X with p components with mean vector (μ) and covariance matrix(Σ). The factor analysis model is given by

$$\begin{gathered} X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\ \vdots \\ X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \end{gathered} \tag{17}$$

Where $\{X_i, \mu_i, F_i, \epsilon_i\}_{i=1}^{p}$are the p observed variables, unknown means of the observed variables, the hidden latent factors called the common factors and $\{\epsilon_i\}_{i=1}^{p}$are the specific variances respectively. And $\{l_{ij}\}_{i,j}^{m}$are the factor loadings. Fig. 3. is a graphical illustration of the FA model, where subsets of the observed variables are captured by the factors, and the specific variances (ϵ_i) are associated with the individual variables (X_i). Clearly from the set of equations (17), the observed variable resolves into a factor component(F_i) and a specific variance component (error) shown in Fig. 4.

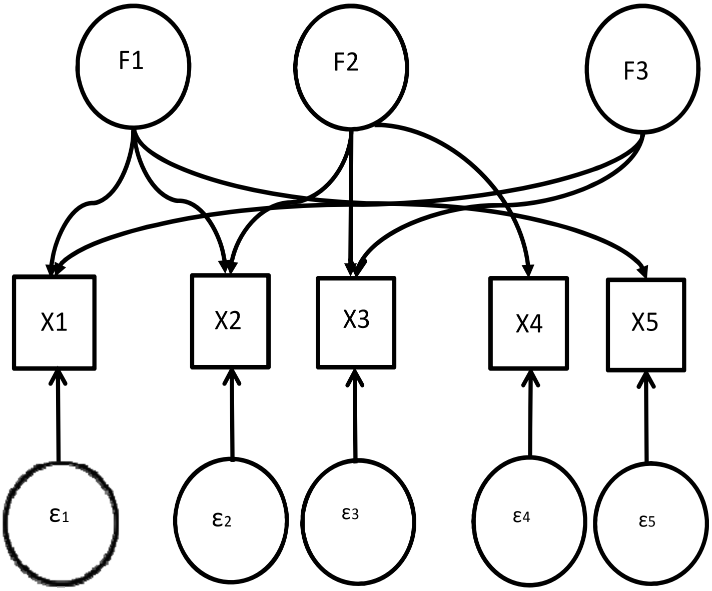


Fig. 3. A graphical representation of the factor analysis model

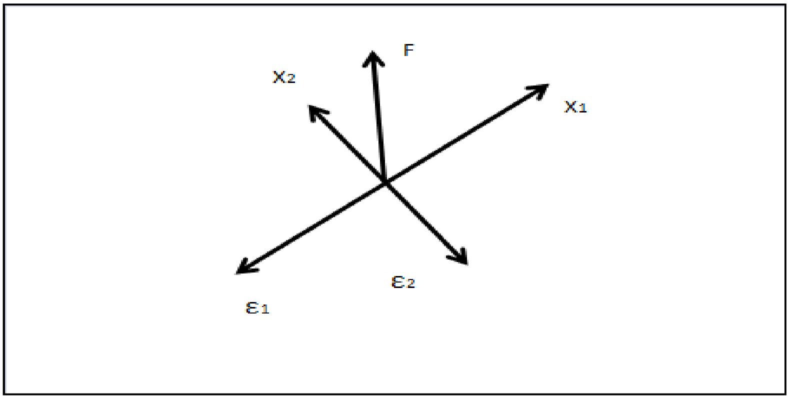


Fig. 4. Resolution of an observation vector (x), into a common factor (F) and error components $\varepsilon_1, \varepsilon_2$

The FA model postulates that a vector X is decomposable into a set of common factors $F_i, (i = 1,2,\cdots,m)$ and specific factors $\{\epsilon_i\}_{i=1}^{p}$. In matrix terms, the FA model in (17) can be written as;

$$X - \mu = LF + \varepsilon \tag{ 18}$$

We make the following assumptions. $E(F) = 0, \Sigma_F = E(FF') = I,, E(\varepsilon) = 0, E(\varepsilon\varepsilon') = \Psi$. The matrix Ψ is given as:

$$\Psi = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

Also, it is assumed that $E(\varepsilon F') = 0$. Note that $E(FF') = I$ means that the covariance of F is an identity matrix. The Factor model can be more elaborately expressed as:

$$X_{px1} = \mu_i + L_{pxm}F_{mx1} + \varepsilon_{px1} \tag{19}$$

The covariance matrix of X is given by;
$E\{(X-\mu)(X-\mu)'\} = E\{(LF+\varepsilon)(LF+\varepsilon)' = LE(FF')L' + E(\varepsilon\varepsilon')\} \xrightarrow{yields} LL' + \Psi$. The cross-products vanish due to our assumptions following (17)

$$\text{Thus, } \Sigma = LL' + \Psi \tag{20}$$

Simple algebraic calculations yield the following identities;

$$Var(X_i) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i \tag{21}$$

and

$$cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{im}l_{km} \tag{22}$$

and $cov(X_i, F_j) = l_{ij}$.

It is clear from the model formulation that factor analysis attempts to reproduce the p + p(p+1)/2 variances and covariance using pm factor loadings and specific variances. So the choice of the number of factors (m) is mighty important.

Heretofore, our discussion focused on what is known as the population model in the statistics literature. Since $\Sigma = LL' + \Psi$ is unknown, it is estimated by the sample covariance (S). We use the sample covariance matrix (S) to estimate the factor loadings. The loadings estimated from the sample are called sample loadings. And the specific variances may be construed as the unexplained sample variance.

$$S = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{p1} \\ \vdots & \ddots & \vdots \\ l_{1m} & \cdots & l_{pm} \end{bmatrix} + \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_{\mathrm{p}} \end{bmatrix} \tag{23}$$

Therefore,

$$s_{ii} = \underbrace{\textstyle\sum_{j=1}^{m} l_{ij}^2}_{communality} + \underbrace{\psi_i}_{specific\ variance} \tag{24}$$

Communality is the sum of the squared factor loadings on the m factors for a given variable. It is the variance in that variable accounted for, by the m factors. Another way to understand *communality* is that is a measure of percent of variance in a given variable explained by the m factors jointly and may be interpreted as the reliability of the indicator (latent factors).

In order to obtain a sample based solution, we use PCA. This is achieved by applying spectral decomposition [3] to the sample covariance matrix (S). The PCA approach decomposes S in terms of Eigen values and Eigen vector pairs. Mathematically, spectral decomposition of a pxp symmetric matrix is given as:

$S_{pxp} = \lambda_1 e_{1(px1)} e'_{1(1xp)} + \lambda_2 e_{2(px1)} e'_{2(1xp)} + \cdots + \lambda_2 e_{p(px1)} e'_{p(1xp)}$, where $\lambda_i, (i = 1,2,\cdots,p)$ are the Eigen values and e_i is the i^{th} Eigen vector. This representation of the sample covariance matrix is known as the famous spectral decomposition. For the FA model to be useful, only the top m << p Eigen vectors are retained, and the specific variances (ψ_i) are assumed to be negligible. In some cases the specific variances are assumed to be non-negligible as well. Furthermore, the spectral decomposition of S can be written as:

$$S_{pxp} = \left[\sqrt{\lambda_1} e_1 | \sqrt{\lambda_2} e_2 \cdots | \sqrt{\lambda_p} e_p\right] \begin{bmatrix} \sqrt{\lambda_1} e_1 \\ \vdots \\ \sqrt{\lambda_p} e_p \end{bmatrix} \tag{25}$$

Let us assume that the specific variances (ψ_i) are non-negligible. Then the FA model is given by

$$S_{pxp} = \left[\sqrt{\lambda_1}e_1 | \sqrt{\lambda_2}e_2 \cdots | \sqrt{\lambda_p}e_p\right] \begin{bmatrix} \sqrt{\lambda_1}e_1 \\ \vdots \\ \sqrt{\lambda_p}e_p \end{bmatrix} + \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \psi_\mathrm{p} \end{bmatrix} \tag{26}$$

If we assume the 2nd matrix (Ψ) to be negligible,

$$S_{pxp} \cong \left[\sqrt{\lambda_1}e_1 | \sqrt{\lambda_2}e_2 \cdots | \sqrt{\lambda_p}e_p\right] \begin{bmatrix} \sqrt{\lambda_1}e_1 \\ \vdots \\ \sqrt{\lambda_p}e_p \end{bmatrix} = L'L \tag{27}$$

Examining the decomposition equation for the sample covariance matrix (S), the loadings l_{ij} is the solution to the equation, $S_{pxp} = L'L$. The loadings appearing as coefficients in the observed variables (X_i) are used to impute meaning to factors and also identify sub groupings of observed variables.

The FA model, while useful for identifying for sub-groupings of original variables, is beset with some ambiguities. The loadings, L are only unique up to rotation. Consider an orthogonal matrix (R) such that $R'R = RR' = I$. The matrix R is a rotation matrix [17]. We saw that the FA model $X - \mu = LF + \varepsilon$ yields the covariance matrix, $\Sigma = LL' + \Psi$. Which can be rewritten as:

$$\Sigma = L\,RR'L' + \Psi \xrightarrow{\text{yields}} L_R L_R' + \Psi \tag{28}$$

where $L_R = LR$. So, we notice that both L andL_R yield the same covariance matrix Σ. So a rotated version of L leads to a set of loadings with the same covariance leading to an obvious ambiguity. So, in applying FA models, if the initial loadings do not yield a satisfactory solution in identifying a reduced set variable sub-groupings or meaningful interpretations, the experimenter can apply rotations such that the loadings L can be redistributed among the factors to possibly obtain more meaningful results. This task is often accomplished by such procedures as *varimax*, and *promax* rotations. Commercial and open source tools such as SAS, MATLAB, and open source program R provide this feature. The material for this section has largely been adapted from [3]. The reader is encouraged to consult it for a detailed exposition.

2.4 Independent Component Analysis

Independent component analysis (ICA) is a versatile technique that can be used for data reduction. ICA is a tool for discovering underlying latent factors that are statistically independent and do not observe the Gaussian law of errors to paraphrase [9]. While PCA and FA depend on the covariance matrix(Σ), ICA seeks projective directions that are statistically independent based on the probability distribution of the data and its higher order moments. Graphically, the ICA model given in Fig. 5, depicts the latent factors linearly combining to produce an output X_j at the j^{th} node with the edge weight equal to l_{ij} connecting the i^{th} latent source(F_i). ICA determines the optimal weights $\hat{l}_{ij}$. The weights may be construed as the correlation between the latent factor (F_i) and the observed output$\left(X_j\right)$. More formally, let, X be a $p \times n$ matrix consisting of n observed samples of a p-dimensional vector $\vec{x}_i$, F is also a $p \times n$ matrix of consisting of n samples of a p-dimensional latent source vector $\vec{f}_i$. And L is a $p \times p$ matrix of unknown weights to be determined. The ICA model in the matrix form is given below.

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} = \begin{pmatrix} l_{11} & \cdots & l_{1p} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pp} \end{pmatrix} \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pn} \end{pmatrix} \tag{29}$$

Concisely, it is given by; $X_{(p\times n)} = L_{(p\times p)}F_{(p\times n)}$. In this equation, X, L, and F are the observed data, the unknown weights, and the unknown latent factors respectively. The objective is to estimate the unknown weights and factors optimally. Notice that unlike the FA model, the ICA does not explicitly consider the specific variances (ϵ). In other words, we are trying to seek, $F = WX$, where $W = L^{-1}$. So we can recover the latent sources(F_i), where $F_i = WX_i$ [18].

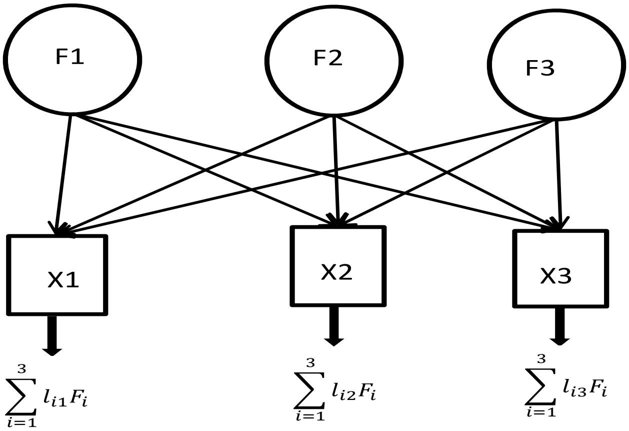

F1
F2
F3
X1
X2
X3
$\sum_{i=1}^{3} l_{i1}F_i$
$\sum_{i=1}^{3} l_{i2}F_i$
$\sum_{i=1}^{3} l_{i3}F_i$

Fig. 5. Graphical illustration of an ICA model. The latent sources F_i are linearly combined to produce an output $\sum_{i=1}^{3} l_{ij}F_i$, where l_{ij} is the weight connecting the i^{th} latent factor F_i to the i^{th} output X_i.

Let the probability distribution of each source F_i be $g_F(\cdot)$. The joint probability distribution of the p independent latent sources and n independent samples (28) is

$$g(F) = \prod_{i=1}^{n}\prod_{j=1}^{p} g\left(f_{ij}\right) = \prod_{i=1}^{n}\prod_{j=1}^{p} g\left(\vec{l}_j{}'\vec{x}_i\right)\left|L^{-1}\right| \tag{30}$$

The above joint probability density of the latent sources in (30) is based on a well-known theorem on transformation of random variables from mathematical statistics [5]. The quantity $|L^{-1}|$ is known as the Jacobian of transformation which is the 2nd term in (30). It corresponds to the 2nd term in (31) of theorem 2.4.1. Note that the vector $(\vec{x}_i)$ is the i^{th} observation and $\vec{l}_j$ is the corresponding weight vector in (29).

Theorem 2.4.1 If a random variable $X \sim f_X(x)$, then the transformed random variable,

$$Y = h(X) \sim f_X(Y)\left|\frac{\partial h^{-1}(y)}{\partial y}\right| \tag{31}$$

Given the observed data $\vec{x}_i, (i = 1,2,\cdots,n)$, the log-likelihood function (ℓ) relative to the joint density $g(F)$ is denoted and written as:

$$\ell(l) = \sum_{i=1}^{n}\sum_{j=1}^{p} log\left[g'\left(\vec{l}_j{}'\vec{x}_i\right)\right] + log|L^{-1}| \tag{32}$$

You will notice that the log-likelihood function is written assuming that the observed samples, and therefore the unknown latent sources are independent. In order to determine the weight vector $\left(\vec{l}_j, j = 1,2,\cdots,p\right)$, we invoke the stochastic gradient methods to maximize the log-likelihood function, and the iterative sequence is given by;

$$L \leftarrow L + \alpha\left\{\begin{bmatrix} 1 - 2g\left(\vec{l}_1{}'\vec{x}_i\right) \\ \vdots \\ 1 - 2g\left(\vec{l}_p{}'\vec{x}_i\right)\end{bmatrix}\vec{x}_i\right\} + (L)^{-1} \tag{33}$$

In the application ICA to obtain the best results, the probability density function $g(\cdot)$ is assumed to be non-Gaussian. The cumulative distribution function (CDF) parametrized by the sigmoid function, $\frac{1}{1-exp^{(-f)}}$ is a candidate CDF. It is well known from mathematical statistics that the probability density function (PDF) is simply the derivative of the CDF [5]. It can easily be checked that the derivative of the sigmoid does not result in the PDF of a Gaussian, which in its general form is;$\frac{1}{\sqrt{2\pi}}exp^{-\frac{1}{2\sigma^2}(f-\mu)^2}$, where$(\mu,\sigma^2)$are the mean and variance respectively. Equation (33) is derived based on the assumption of the sigmoid function.

In the updating equation above, the parameter (α) is the learning rate. The 2^{nd} term, $(L)^{-1}$ is obtained by finding the derivative of $log|L^{-1}|$. On finding the optimal values of L, the latent factors can be constructed from $F_i = WX_i$, where $W = L^{-1}$. It is noted *en passant* that another typical application of ICA is for identifying latent sources (these correspond to mixed signals of voice samples captured by microphones placed in room-the famous cocktail party problem), our perspective and purpose here is different. We assume that there is a model consisting of a $p \times n$ matrix (*X*) of observed data which is a linear combination of latent sources. The idea is to identify a smaller set containing linear combinations of latent sources which explain the variance in the observed data.

In deriving the ICA model, we used a sigmoid function parametrization. However, if the application suggested a certain parametric form for the PDF, one should incorporate it into the joint density function, and derive the updating equations accordingly.

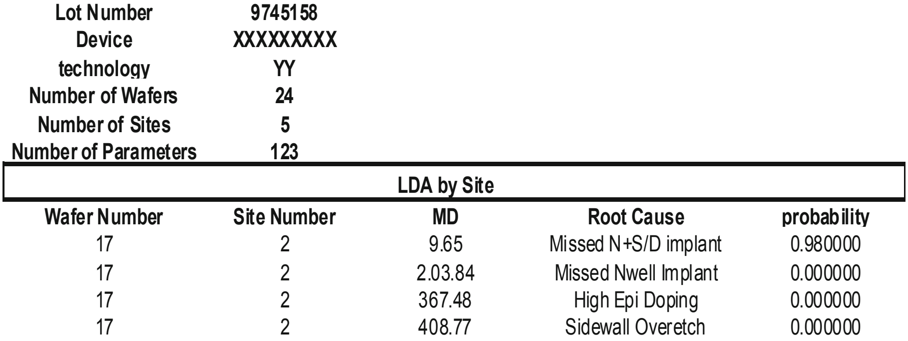
2.5 Projection Pursuit

Pursuant to our effort to discover hidden structure not captured by the covariance matrix(Σ), we introduce a popular technique called projection pursuit (PP) which is somewhat computationally intensive. Projection pursuit is a technique to reduce high-dimensional data by projecting on to a lower-dimensional space to reveal latent (hidden) structure in the higher dimensions [7]. Projection pursuit was first invented by Krushkal to discover *interesting* lower-dimensional projections. The notion of "interestingness" is parametrized by an index given by $I(w)$. If we recall, PCA, the goal there was to find axes of an ellipsoid (assuming the data is Gaussian) that corresponded to largest variation. So, the index $I(w)$ is the projection $(\vec{x}'\vec{w})$ of the data vector on to an Eigen vector$(\vec{w})$ subject to $\vec{w}'\vec{w} = 1$. The "interestingness" in the data is of course the linear projections which are the principal axes parametrized by the Eigen vectors. For comprehensive detail and a beautiful exposition of PP, the reader is referred to [7]. We alluded to *hidden structure* in high dimensions. The Gaussian distribution being rotationally symmetric does not produce interesting projections. Because a linear projection $(\vec{w}'X)$ where the random vector $X \in \mathbb{R}^p$, being a sum of random variables will again observe the Gaussian law by the central limit theorem. Therefore, a preponderance of linear projections do not reveal structure beyond the 2nd order moments. The projection index we seek is based on polynomial moments. The idea is to transform a projection$(\vec{w}'X)$ to $P = 2\Phi(\vec{w}'X) - 1$ where$\Phi(\cdot)$ denotes a Gaussian CDF. The transformation results in a Uniform distribution. The transformed projection is then compared against a Uniform random variable(U). A departure from the uniform distribution measured by $f_P(p) - g_U(u)$ is an indication of non-Gaussian structure. The symbols(f, g) are the distributions of the two random variables, (P, U)respectively. Operationally, it is the integral squared distance between the densities of P, U that is calculated. The integral square error statistic serves as a projection index. The reader is again encouraged to refer to [7].

3 Applications

In this section, we picked PCA to illustrate the importance of dimensionality reduction in a semiconductor manufacturing application. Signature analysis (SA) in semiconductor manufacturing is a statistical pattern recognition program designed to assign failed parts to one of several pre-determined root cause categories [10]. Engineers invest lots of time tracing back-end electrical parameter test failures to probable on-line root causes. It is desired to have an automated program based on sound statistical theory that enables the classification of a failing signature to a root cause category such that the probability of misclassification is minimized. Linear discriminant analysis (LDA) is an established parametric procedure that minimizes the probability of misclassification and allows the failure analysis engineer to state "The probability that a failing chip with a specific signature belongs to the k^{th} root cause category is *p %*." But prior to applying LDA, a database of signatures is created. A signature is merely a feature vector of measurements obtained from a chip. Associated with each signature is a *label* which indexes the failing chip with an associated root cause. In many semiconductor manufacturing settings, the size of the signature vector is in excess of 400 features due to the number of tests conducted to ensure the reliability of the finished product. A majority of these tests are electrical measurements that are correlated to one another. So applying of PCA not only reduces the dimensionality of the signature vector, but also eliminates the collinearity (correlations) among the features since the principal components are orthogonal to one another. In the example below, chips are manufactured using the LinBiCMOS technology. LinBiCMOS is a CMOS technology with bipolar components (see Wikipedia for details about the semiconductor technologies). The chips were tested at 5 locations on a wafer. A wafer is an array of chips laid out as a matrix on a circular disc. The wafers are processed in batches of 20 are known as *lots*. The five locations known as test structures are at the top, center, bottom, left, and right (T,C,B,L,R) locations on the wafer. The electrical test measurements were approximately 125. Many of the measured features were correlated and thus redundant. We applied PCA to reduce the feature set to 34 principal components, which is a reduction of ~75%. An example using LDA to determine root cause of failures is shown in Table 1.

Table 1. Classification by Linear Discriminant Analysis



Lot Number	**9745158**
Device	**XXXXXXXXX**
technology	**YY**
Number of Wafers	**24**
Number of Sites	**5**
Number of Parameters	**123**

LDA by Site

Wafer Number	Site Number	MD	Root Cause	probability
17	2	9.65	Missed N+S/D implant	0.980000
17	2	2.03.84	Missed Nwell Implant	0.000000
17	2	367.48	High Epi Doping	0.000000
17	2	408.77	Sidewall Overetch	0.000000

Two lots failing due to missing N+S/D implant were submitted to the automated signature analysis program for root cause identification (see the column headed, "probability" in Table 1). A signature of length 34 is applied to the program for pattern classification. The signature is from a certain device XXXXXXXXX belonging to technology YY. Table 1 shows the results of this analysis. The number of electrical test parameters measured for this technology is 123, but a signature of dimension 34 is applied for classification using PCA. Wafer 17 which failed some tests was applied to the SA program. The measurements were obtained from site 2 which corresponds to one of the locations (T,C,B,L,R). Clearly, LDA identified the correct root cause, and the dimensionality reduction by PCA captured sufficient information to draw the correct inference!

References

1. Committee on the Analysis of Massive Data, Frontiers in Massive Data Analysis. National Academies Press (2013)
2. Shalizi, C.R.: Advanced Data Analysis from an Elementary Point of View (2013), `http://www.stat.cmu.edu/~cshalizi`
3. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 3rd edn. Prentice Hall, Englewood Cliffs (1992)
4. Hotelling, H.: Relations Between Two Sets of Variables. Biometrika 28, 321–377 (1936)
5. Mood, A.M., Graybill, F.A., Boes, D.C.: Introduction to the Theory of Statistics, 3rd edn. McGraw-Hill (1974)
6. Hardle, W.K., Simar, L.: Applied Multivariate Statistical Analysis, 3rd edn. Springer (2011)
7. Friedman, J.H.: Exploratory Projection Pursuit. Journal of the American Statistical Association 82(397), 249–266 (1987)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer (2001)
9. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, Inter-Science (2001)
10. Lakshminarayan, C.K., Baron, M.I.: Pattern Recognition in Large-Scale Data Sets: Application in Integrated Circuit Manufacturing. In: Bhatnagar, V. (ed.) BDA 2013. Springer, Heidelberg (2013)
11. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C, The Art of Scientific Computing. Cambridge University Press (1990)
12. Strang, G.: Linear Algebra and its Applications, 4th edn. Brooks/Cole Publishing Company (2005)
13. Burgess, C.J.C.: Dimension Reduction: A guided Tour. Foundation and Trends in Machine Learning 2(4), 275–365 (2010)
14. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis; An overview with application to learning methods, technical report, CSD-TR-03-02, Dept. of Computer Science, Royal Holloway, University of London (2003)
15. Timm, N.H.: Applied Multivariate Analysis. Springer (2002)
16. Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer (2007)
17. Strang, G.: Introduction to Applied Mathematics. Wellesley-Cambridge Press (1986)
18. Ng, A.: Independent Component Analysis, CS229, Lecture Notes. Stanford University