Chapter 1

Introduction

Blind Source Separation (BSS) is an area of research started in 1980’s, popularized in 1990’s and still expanding. It aims retrieval of unobserved sources from the observed mixtures in the absence of any prior information about the sources or the mixing system. Originated from a neuro-biological problem, now it finds applications in the areas of feature extraction, classification, telecommunications, brain signal processing, audio and speech processing, denoising and so on. The research has been nurtured and advanced by the research communities mainly working in the areas of neural net, signal processing and statistics. The detailed history can be found in (35, 65, 76).

1.1 The BSS Problem

The famous Cocktail party problem is a good example to explain the Source Separation problem. Let there be going on a cocktail party with multiple speakers, background music, breaking of a glass, continuous murmuring of a mass etc. as audio sources. Though there are multiple audio sources; for a human being present in that party, it is possible to focus on a particular audio source of choice. Whether a machine could have similar ability to separate the sources of choice from their mixtures? Elaborating more, let the recording devices are placed at some locations. The audio sources combine together to generate a mixture signal at each recording device. Let there be an assumption that the audio sources are linearly and instantaneously combined; where, the coefficients of linearity depend on their mutual geometrical locations, distances, characteristics of
the environment and others. Mathematically,
\[
\begin{align*}
    x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + \ldots + a_{1n}s_n(t) \\
    & \vdots \\
    x_m(t) &= a_{m1}s_1(t) + a_{m2}s_2(t) + \ldots + a_{mn}s_n(t)
\end{align*}
\]
where, \( m \) is the number of mixture signals; \( n \) is the number of sources; \( t \) is the time indice; \( x_i(t), i = 1 : m \) are the mixture signals; \( s_i(t), i = 1 : m \) are the source signals and \( a_{ij}, i = 1 : m, j = 1 : n \) are the mixing coefficients. Overall, the generation mechanism of the mixture signals can be explained, using matrix notations, as under:
\[
    x(t)_{m \times 1} = A_{m \times n}s(t)_{n \times 1} \tag{1.1}
\]
where, \( x(t) = (x_1(t), x_2(t), \ldots, x_m(t))^T \in \mathbb{R}^m \) is an observed mixture random vector;
\( s(t) = (s_1(t), s_2(t), \ldots, s_n(t))^T \in \mathbb{R}^n \) is a source random vector and \( A \) is a mixing transformation.

Formally, the **BSS model** explains generation of an observed random vector \( x(t) \) as an unknown transformation \( \mathcal{F} \) to an unobserved source vector \( s(t) \). Mathematically,
\[
x(t) = \mathcal{F}(s(t)) \tag{1.2}
\]
where, \( \mathcal{F} \) is an \( m \)-component invertible mixing transformation from \( \mathbb{R}^n \) to \( \mathbb{R}^m \).

Formally, the **BSS problem** is to estimate the unknown \( s(t) \) based on some *generic assumption* on the sources. The word *blind* implies that there is no other information available about the sources or recording devices; e.g. their mutual location geometry or source distributions or other.

---

*more generalized, in contrast to application specific assumptions in semi-BSS problem.*
sources or the mixing system, except the applicability of the generic assumption. If \( y(t) \) is the estimated source vector and \( \mathcal{G} \) is the estimated inverse of the mixing function \( \mathcal{F} \) then

\[
y(t) = \mathcal{G}(x(t)) = \mathcal{G}(\mathcal{F}(s(t))) = \mathcal{H}(s(t)) \tag{1.3}
\]

where, \( \mathcal{H} \) is an \( n \)-component transformation from \( \mathbb{R}^n \) to \( \mathbb{R}^n \) and each component \( h_i, i = 1, 2, \ldots, n \) is a left indeterminacy transformation giving one-one mapping between the \( i^{th} \) estimated source and the \( i^{th} \) actual source; i.e. \( y_i(t) = h_i(s_i(t)) \).

The mixing transformation \( \mathcal{F} \) can be linear or non-linear, instantaneous (without memory) or dynamic (with memory), stationary or time-varying. Similarly, the transformation with condition \( m > n \) (i.e., number of mixtures more than the number of sources) implies an \textit{overdetermined} system; with condition \( m < n \) (i.e., number of mixtures less than the number of sources) implies an \textit{underdetermined} system; and that with condition \( m = n \) (i.e., number of mixtures are same as the number of sources) implies a \textit{determined} system. The thesis, if not specified, assumes linear, instantaneous and \textit{determined} system.

### 1.1.1 Linear Instantaneous BSS Problem

Assuming the mixing transformation \( \mathcal{F} \) to be linear, instantaneous and invertible; similar to those in \textit{cocktail party problem} example; and extending the model for \( N \) sample instances, the BSS generation mechanism is explained as under:

\[
X(t)_{n \times N} = A_{n \times n} S(t)_{n \times N} \tag{1.4}
\]

Accordingly, the linear instantaneous \textit{BSS problem} is to estimate the source random vector \( S(t)_{n \times N} \) back from the available mixture random vector \( X(t)_{n \times N} \) without using any information about \( A \) or \( S(t) \). Mathematically, it is represented as in Equation (1.5) below:

\[
y(t)_{n \times N} = W_{n \times n} X(t)_{n \times N} = WAS(t) = H_{n \times n} S(t)_{n \times N} \tag{1.5}
\]

where, \( Y(t) = \hat{S}(t) \) denotes estimated source random vector, \( W = \hat{A}^{-1} \) denotes estimated inverse of the mixing matrix and \( H \) is the left indeterminacy transformation.

### 1.2 The BSS Solution

Towards the BSS solution, there are mainly four different approaches based on four different generic assumptions. The assumptions are:
1. Introduction

1. Second order uncorrelatedness among the sources

2. Statistically independence and identical distribution among the sources, which leads to Independent Component Analysis (ICA)

3. Sparsity of the sources, which leads to Sparse Component Analysis (SCA)

4. Nonnegativity of the sources, which leads to Non-negative Matrix Factorization (NMF)

Instead of using mere uncorrelatedness depending upon second order statistics, statistical independence considering higher order statistics among the sources is a more stronger property. Sparsity and nonnegativity are relatively new assumptions valid in specific applications. In general, BSS using the generic assumption of independence among the sources has a wider perspective to cover more applications and the topic under the scope of this thesis.

1.2.1 Can Independence Assumption Solve the Linear, Instantaneous BSS Problem?

The fact that independence assumption can lead to BSS solution is proved by Darmois-Skitovich Theorem. The Theorem is independently proved by both Darmois (38) and Skitovitch (117), in the context of factor analysis.

**Theorem 1.1 (The Darmois-Skitovich Theorem).** Let \( y_1 \) and \( y_2 \) be random variables defined as under:

\[
\begin{align*}
    y_1 &= g_{11}s_1 + g_{12}s_2 + \ldots + g_{1n}s_n \\
    y_2 &= g_{21}s_1 + g_{22}s_2 + \ldots + g_{2n}s_n
\end{align*}
\]

where, \( s_1, s_2, \ldots, s_n \) are independent random variables. Then, if \( y_1 \) and \( y_2 \) are independent, all variables \( s_k \) for which \( g_{1k}g_{2k} \neq 0 \) are Gaussian.

The theorem states that non-Gaussian independent random variables can not be mixed linearly and instantaneously to have independent mixture outcomes. Consequently, with linear and instantaneous (memoryless) transformations both \( G \) and \( F \) in the above equation (1.3), if \( s(t) \) and \( y(t) \) both are independent then \( G(F(\cdot)) \) must correspond to a product of scale and permutation transformations. This proves that independence assumption can lead to separation of sources if not more than one source is Gaussian.
1.2.2 Linear, Instantaneous ICA for BSS

The equation (1.4) reminds us factorization of a data matrix in Component Analysis (CA). The goal of CA is to remove redundancy in the data matrix by change of basis. The conventional Principal Component Analysis (PCA) achieves this goal of redundancy removal by finding directions with maximum variance and mutually uncorrelated. As based on second order statistics, PCA is used to separate Gaussian or wide sense stationary (WSS) random processes. The general framework to solve BSS, based on the assumption of statistical independence among the sources, is inspired by PCA and has been identified as Independent Component Analysis (ICA). It achieves redundancy removal by finding the components, which are statistically the most independent; in a sense that the information in a component direction can not be known by knowing the other components. Other than a BSS technique, as a CA tool, ICA has been reported for many applications as dimensionality reduction, pattern classification, pattern recognition, feature extraction, data compression and others (42). The formal definition of ICA, given by P. Comon in (33) for linear transformations, is as under:

**Definition 1.2.** The ICA of a random vector \( \mathbf{x} = (x_1, x_2, \ldots, x_m)^T \) with finite covariance \( \mathbf{C}_x \) is a pair \( \{ \mathbf{B}, \Delta^2 \} \) of matrices such that

i. the covariance factorizes into \( \mathbf{C}_x = \mathbf{B} \Delta^2 \mathbf{B}^* \), where \( \Delta \) is diagonal real positive and \( \mathbf{B} \) is full column rank matrix, \( ^* \) indicates complex conjugate;

ii. the observations can be written as \( \mathbf{x} = \mathbf{B} \mathbf{y} \), where \( \mathbf{y} \) is an \( n \times 1 \) random vector with covariance \( \Delta^2 \) and whose components are 'the most independent possible', in the sense of the maximization of a given 'contrast function'\(^†\), as defined in Chapter 2.

The first condition, similar to the PCA definition, identifies \( \mathbf{B} \) as a set of eigenvectors. In case of PCA, the second condition on \( \mathbf{B} \) is orthogonality. For ICA, the second condition restricts \( \mathbf{B} \) to give random variables \( y_i \) as independent components (ICs).

As can be noted, the ICA definition assumes \( \mathbf{x} \) and \( \mathbf{y} \) as random vectors without any time indices. Also, the instantaneous model, if extended for \( N \) number of samples, assumes identical distribution of the sources. Over all, ICA model assumes 'independent and identical distribution' (i.i.d.) assumption on the sources. Given the assumption is satisfied by a source random process, ICA can be used for BSS. Conventionally, linear ICA is considered equivalent to BSS. More details on the ICA solution and discussion on its' use for BSS is provided in Chapter 4.

\(^†\)Roughly speaking, contrast function is a maximization function based on independence measure, satisfying specific conditions to bring the ICs uniquely.
As a conclusion, the BSS solution for linear, instantaneous mixing system can be obtained by maximizing the independence among $y_i(t)$s with respect to the separation matrix $W$, as:

$$y^*(t) = \arg\max_W \Phi(y(t))$$  \hfill (1.6)

where, $\Phi(y(t))$ is the contrast function. The solution demands discussion on the suitable contrast functions as optimization criteria and suitable optimization technique corresponding to that contrast function.

1.3 The Linear ICA Problem and Solution

The Independent Component Analysis (ICA) model explains generation of an observed random vector $x$, as a linear transformation to another latent (hidden) random vector $s$. Mathematically, $x = As$, where $x = [x_1; x_2; \ldots; x_m]$; $s = [s_1; s_2; \ldots; s_n]$; $x_i, s_i$ are random variables with values in $R$; $m = n \geq 2$ and $A$ is full rank. Let there be available $N$umber of samples of each observed random variable. Assuming an identical distribution, the instantaneous model can be extended for $N$ realizations. Let $X = [x_1; x_2; \ldots; x_m]$ be the $m \times N$ data or observation matrix and $S$ be the $n \times N$ component or source matrix. Then,

$$X = AS$$  \hfill (1.7)

The problem of ICA is to estimate both the unknowns $A$ and $S$, with the only assumption of $s_i$ being mutually the most independent possible (m.i.p.) random variables with respect to a given contrast. If $W$ is the estimated inverse of the mixing matrix $A$ then the estimated source or component matrix $Y$ is:

$$Y = A^{-1}X = WX = WAS$$  \hfill (1.8)

The above ICA solution has the following inherent limitations or indeterminacy as discussed in (25, 33, 46):

- As $X = AS$, scaling to any source $s_i$ can be canceled by dividing corresponding column $a_i$ of $A$. So, both being unknown, the scaling or the variances and the signs of the ICs can not be estimated. Similarly, if $P$ is the permutation matrix, $X = (AP^{-1})(PS)$ i.e. the order of the estimated components can change with the change in the estimated mixing matrix. In short, the estimated sources $Y$ can be obtained as a scaled and permuted version of actual sources $S$. To have a unique solution, ICA assumes all components to be equivariant or univariant (33, Section 1.5).
Gaussian distribution is symmetric. It can not be modified by the mixing vector or any mixing vector produces the same distribution. Accordingly, if there are more than one components with Gaussian distribution, actual mixing vector can not be known from the mixture.

### 1.3.1 An Orthogonal Approach to ICA Solution

The goal is to obtain ICs \( (y_s) \) from the correlated mixtures \( (x_s) \) of them.

- By definition, statistical independence implies uncorrelatedness (the opposite is true only for Gaussian variable). The uncorrelated components with zero mean imply orthogonality. So, the ICs with zero mean are also mutually orthogonal.

- The zero mean uncorrelated components of the data matrix are mutually orthogonal and zero mean ICs of the data matrix are also mutually orthogonal. There are techniques available to get zero mean uncorrelated components of the data matrix. If we think of a transformation from the former to the latter, then it must be an orthogonal transformation. Because, only an orthogonal transformation can keep the orthogonal components, still orthogonal. To bring uniqueness of the estimated components, the uncorrelated components should be univariant or equivariant. Let a zero mean observed mixture data matrix \( X \), be linearly transformed through a whitening matrix \( V \), to give a zero mean, univariant, whiten (uncorrelated and equivariant) data matrix \( Z \).

\[
Z = VX
\]  
\[
\Rightarrow Z = VAS
\]

The goal is to obtain the estimated sources \( y_s \) as zero mean, univariant, ICs. Let us assume \( R \) is the linear transformation for that. Then,

\[
Y = RZ
\]

\[
\Rightarrow E\{yy^T\} = E\{RZZ^TR^T\}
\]

\[
\Rightarrow I = E\{RIR^T\}
\]

where, \( I \) is the identity matrix. It proves that \( R \) needs to be orthogonal. Accordingly,

\[
Y = RZ = RVAS = WAS
\]

where, \( W = RV \) is the estimated unmixing matrix.
• Orthogonal matrices with determinant +1 are rotation matrices and those with determinant -1 are reflection matrices. Given a reflection orthogonal matrix with determinant -1, by negating an odd number of columns, a new rotation orthogonal matrix with determinant +1 can be derived. As the estimated sources are allowed to be scaled or reflected version of the actual sources, the orthogonal transformation matrix $R$ can be a rotational matrix.

• Concluding above - a specific rotation matrix, from the infinite set of all $n$-dimensional rotation matrices, will be able to transform a set of zero mean whiten or eigen (not univariant but orthogonal) components to ICs. So, the ICA problem reduces to estimating a rotation matrix $R$ giving $y_i$.

$$Y^* = \arg \max_{W^*} \Phi(WZ) \quad (1.14)$$

where, $\Phi(y)$ or $\Phi(Y)$ is the contrast function based on the dependence or independence measure of random vector $Y$.

• The $n \times n$ rotation matrix $R$ has $d = \frac{n(n-1)}{2}$ entries to be estimated. Other way, $d = \frac{n(n-1)}{2}$ number of 2-d rotations are required to have $R_{n \times n}$. Overall, the linear ICA or linear BSS problem to separate $n$ number of sources reduces to $d$-dimensional optimization problem.

### 1.4 The Large Scale BSS (LSBSS) Problem

It is known and also proved in the Section 1.3.1 that the linear BSS problem with $n$ number of unknown sources is an $d = \frac{n(n-1)}{2}$ dimensional optimization problem. Accordingly, with number of sources $n > 14$ in BSS, the optimization problem has to deal with dimensions $d > 100$. Similarly, $n > 45$ corresponds to $d > 1000$ and $n > 141$ corresponds to $d > 10000$. The optimization research community refers a problem with dimensions $d \in [100, 10000)$ as the Large Scale Global Optimization (LSGO) problem and a problem with dimensions $d > 10000$ as the Big Scale Global Optimization problem. So, the thesis refers ‘Large Scale’ in LSBSS problem as the BSS in higher dimensions with number of sources ranging from more than 14 to less than 141, i.e. $n \in [15, 140] \Rightarrow d \in [100, 10000]$. The optimization techniques already face the problem of ‘curse of dimensionality’. With the linear increase in number of dimensions (d), the solution space increases exponentially ($a^d$, some $a$) and so does the difficulty in optimization. So, the LSGO for real world applications is still challenging and an identified problem (139). It has been a part of competitions at many conferences; such as, IEEE Congress on Evolutionary Computation (IEEE CEC) 2008, 2010, 2013, 2015. Overall, the solution of LSBSS demands multi-disciplinary approach.
1. Introduction

1.5 The Large Scale near-Independent BSS (LSnIBSS) Problem

The literary meaning of near-independence is - ‘not exact independence’. But, the ICA model allows the components, being separated, as mutually the ‘most independent possible’ (m.i.p.)† with respect to a given contrast function. So, the thesis defines ‘near-independent’ sources as the sources not being m.i.p. with respect to the used contrast function.

The m.i.p. sources correspond to the global optima of the optimization landscape. The near-independence among the sources may be exhibited in the following three ways in the optimization landscape of the used contrast:

i. The actual sources correspond to the solution near global optima i.e. there is a shift of the global optima such that the optimal solution do not correspond to the actual sources.

ii. There exists an added one or more local optima, which do not correspond to the actual sources.

iii. There is simultaneously an added local optima, as well as, shifted global optima.

The added local optima makes the optimization landscape difficult to be optimized but the shifted global makes either almost impossible to find the optima corresponding to the actual sources without any additional information or only an approximate solution can be obtained based on the amount of shift. At lower dimensions, a slight shift in global optima may allow atleast an approximate solution. With increasing dimension, cumulative slight shifts in pairwise optima, may cause the actual solution much far than the global optimal. Overall, the LSBSS problem demands special focus on the study of the circumstances causing these adverse optimization landscape and their consequences on separability. The thesis identifies the study area as ‘near-Independent’ BSS (nIBSS). The sources producing either shift of global optima or addition of spurious local optima or both with respect to the used contrast qualify to be near-independent for that contrast. It is to be noted that the near-independence is not a characteristic of sources alone, but it is the characteristic of sources exhibited in the presence of a specific contrast.

Though the term near-independence is new, there already exists local minima and extrema/optima analysis of different contrast functions with respect to various types of sources. Recently, there has been found situations that affect the optimization landscape in case of BSS of real world sources. There exists spurious local optima of information theoretic independence measures for multimodal source distributions. The empirical observations are supported by theoretical extrema analysis in (17, 90, 91, 92, 132, 134). On the other hand, it has been proved that lack of

†whether mention or not, m.i.p. implies most independent possible with respect to the given contrast function.
number of samples may bring overlearning phenomena in ICA for kurtosis like independence measures (83, 108, 109). The overlearning results into a shift of global optima. The near-independence terminology makes it possible to study two differently looking problems, under the same roof.

The near-independence condition is not same as the non-independent or non-\textit{i.i.d.} conditions stated in the BSS literature. Usually, non-independent sources imply time dependencies and non-identicle conditions imply non-stationarity. There exists BSS model extensions for non-\textit{i.i.d.} i.e. temporally dependent sources and/or non-stationary sources (41, 70, 73). More precisely, the non-\textit{i.i.d.} is the property of the sources only, while the near independence implies the source model violation with respect to the used contrast function only. Also, non-independence is more stronger than the near-independence, in terms of violating of ICA source model assumption.

Overall, the LSBSS of the more difficult real world near-independent sources give birth to the Large Scale near-Independent BSS (LSnIBSS) problem. The performance degradation in LSnIBSS is either due to the failure of an optimization techniques in the presence of local minima] or due to the shift of global (i.e. an optimization technique is successful in finding the optimal but the optimality does not assure separability) or due to both the former reasons.

\section*{1.6 Current State of the Art}

The linear BSS algorithms differ based on the used optimization criteria and optimization method.

Conventionally, independence interpretations in terms of minimization of mutual information, maximization of non-gaussianity and their approximations using higher order statistics (cumulants and moments) have served as the major guiding principles to derive the BSS contrasts. There exists ICA techniques using adaptive learning through neural net (11, 63, 64, 71, 77, 129). The nonlinearity used for learning has to be a function of probability density function (PDF) of the components to be estimated. In the absence of this knowledge, family of densities (e.g. super-Gaussian or sub-Gaussian) is used as an approximation to select the nonlinearity. This requires some prior knowledge of densities to be estimated and so violates the blind assumption. There exists algebraic techniques (22, 23, 27, 28, 33) trying to obtain uncorrelatedness of third or fourth order statistics, inspired by the diagonalization techniques for PCA through second order uncorrelatedness. With approximate independence measures, they offer a less precise solution (10). There are also likelihood (12, 24) based signal processing techniques for ICA.

Most of the ICA algorithms use gradient based optimization techniques though exhaustive search based optimization techniques have also been explored (75, 112). The gradient based optimization techniques lack global convergence, required specifically in near-independent BSS with local minima. The exhaustive search methods used for optimization are computationally demanding, specifically in large scale.
Overall, a BSS algorithm using contrast that allows blind estimation, offers precision in separation quality and computationally efficient is still in demand. Further, the large scale and near-independent scenario demands the same BSS algorithm using a global and computationally efficient optimization technique.

The kernel based nonparametric estimation methods (10, 18, 87) are both quite precise and blind but require high computational cost. In search of a BSS contrast with computation reduction, there has been explored alternative definitions of Entropy other than Shanon’s. So, the latest trend is to develop an ICA algorithms using kernel estimation of an independence measure that is based on alternative definitions of Entropy, specifically, the quadratic measures of independence offering low computational cost. The Information Theoretic Learning (ITL) - a new research area (96) and article (1) provide many such alternative definitions of independence and quadratic independence measures; for example, generalized $\beta$-Class Entropy, Renyi’s Entropy, Cross Information Potential (CIP), Euclidean distance ($D_{ED}$) based and Cauchy-Schwartz distance ($D_{CS}$) based Quadrature Mutual Information (QMI) (58, 112) and others. It should also be noted that Pham (93) proved that there are risks using Renyi’s entropy definition for BSS.

Concluding above, in the midst of existing many other algorithms, there is still a requirement for a linear BSS algorithm that is - blind in nature, based on computationally efficient kernel based nonparametric estimation of contrast and using optimization technique with good global convergence - even in small or medium scale BSS.

The performance degradation of existing ICA algorithms with increase in dimensions is a known fact (10, 75). The large scale in BSS, using independence assumption, has been addressed only in article (18) as per the knowledge of the author. There it is claimed that the ICA technique (NPICA) based on the nonparametric estimation of marginal entropies can seamlessly handle large scale. But, the empirical results reported in this thesis in the Chapter 4.8, show failure of NPICA in two dimensions, as well, in higher dimensions for near independent sources. There are efforts to solve LSBSS problem using other than independence assumption by (30) and (19). In general, the LSnIBSS, through conventional independence assumption, is still an unsolved problem. The performance degradation in LSnIBSS is either in terms of the failure of an optimization techniques to converge to an optimal or in terms of highly increased computation. Atleast the brain signal processing area, for EEG (Electro-Encephalo-Graph) and MEG (Megneto-Encephalo-Graph) data analysis like applications, currently demands LSBSS. Conventionally, they are derived through sequentially executing Blind Signal Extraction (BSE) algorithm to extract one or few important signals, instead of BSS (37, 133). It is anticipated that the LSnIBSS solution will find applications in brain signal processing, feature extraction and other data analysis problems.

The LSnIBSS solution demands contrast providing optimization landscape without any spurious local optima and shift of global optima. It demands optimization algorithm that is com-
putationally efficient and has good global convergence. The solution also demands study on near-independence scenario.

Overall, the requirements of LSnIBSS have been identified as the research problem for the thesis. This also justifies the title of the thesis.

1.7 The Motivation Summary and Work Directions

The thesis addresses the LSnIBSS problem in three directions:

1. Towards optimization criteria: As concluded in the previous Section 1.6, contrast that sticks to the blind assumption through kernel based nonparametric estimation, offers precision in separation quality, computationally efficient, without any local minima and using a 'prior' that does not violate the blind assumption is in demand. The work towards this direction is briefed in Chapter 2 and Chapter 3.

2. Towards optimization landscape: The near independent sources scenario needs to analyze situations affecting the optimization landscape, their consequences on separation quality and possible remedies. The related work is reported in Chapter 4.

3. Towards optimization technique: The Large Scale Global Optimization (LSGO) is reported to have linear time complexity ($O(d \ln d)$, where $d$ is the dimension of search) for a specific type of problems and an exponential time complexity ($O(d^d)$ or $O(\exp(d \ln d))$) for another type of problems (107). The BSS contrasts belong to which group of optimization functions for LSGO that need be identified first. Then, a suitable LSGO technique, either existing or newly defined, need be used for LSnIBSS. The work towards this direction, is reported in chapters 4, 5 and 6.

1.8 The Thesis Organization and Detailed Outline

The next Chapter 2 derives new contrasts for linear BSS. For differentiable multivariate functions with equal hyper volumes (region bounded by hyper surfaces) some results are proved relating equality of derivatives to equality of the functions. The results are applied to the independence definition stating equality of joint PDF and product of the marginal PDFs of a random vector. This avails new independence measures and BSS contrasts. The Chapter defines difference between the joint PDF and the product of the marginal PDFs as a Function Difference (FD) of a random vector. Similar to the Score Function Difference (SFD) definition in (7, 8), the gradient of FD (GFD) and the Hessian of FD (HFD) are defined. It is proved that FD, GFD, HFD all are zero everywhere
when the corresponding random variables are independent. The results lead to derive minimization of $L^p$-Norm of FD, GFD and HFD as contrasts for BSS.

The estimation method should be computationally efficient to match the requirement specifically for LSBSS. The contrasts depending upon joint PDF and marginal PDFs both are usually computationally demanding though more accurate (88). Instead of a conventional two stage estimation approach for a quantity like FD, a direct single stage estimation is more accurate. This is concluded and applied for ‘least squares’ based density difference estimation in (123). Also, the kernel theory identifies the fact that it is computationally more efficient to estimate the integration of square of PDF than the estimation of actual PDF. The ITL theory has given significance to this fact by defining integration of the square of PDF as an Information Potential of a random variable. The analogy, with the existing potential theory in Physics, also has given other concepts related to the information field; like, information forces, information particle interactions and others (96, 140, 141). The Section 2.7 targets both the efficient estimation of the proposed contrasts and extension of the potential theory for an information field. The potential theory has a concept of reference potential and it is used to derive closed form expression for the relative analysis of potential field. Analogous to it, the Section 2.8 introduces the concepts of Reference Information Potential (RIP) and Cross Reference Information Potential (CRIP) based on the potential due to kernel function placed at selected sample points as basis in kernel methods. The quantities are used to derive closed form expressions for information field analysis using least squares. The expressions are derived through multiplicative kernel basis in two ways: (a) basis placed at the selected paired sample points (b) basis placed at the selected paired or un-paired sample points. The expressions are used to estimate the required contrast functions. They are used to estimate $L^2$-Norm of FD and $L^2$-Norm of GFD based contrasts.

The performance of a kernel method depends upon efficient bandwidth parameter selection. The most popular and simple Silverman’s Rule-of-Thumb (ROT) (116) does not give precise bandwidth parameter and the precise solve-the-equation based plug-in methods are computationally too demanding. So, deriving data dependent bandwidth selection method for Kernel Density Estimation (KDE) that balances accuracy and computation is the focus of the next Chapter 3. It achieves this goal by deriving a novel Extended rule-of-thumb (ExROT). The ROT optimizes Asymptotic Mean Integrated Square Error (AMISE) with the assumption that the density being estimated is Gaussian. The ExROT uses infinite series expansion as an approximation to the unknown PDF. As an example here, the ExROT uses an extended assumption that the density being estimated is near Gaussian. The assumption makes it possible to use the Gram-Charlier (GC) A-series expansion of near Gaussian PDF with the same AMISE criteria for bandwidth optimization. There exist many other infinite series expansions of PDF based on which other variants of the ExROT could be derived. The multivariate ExROT is derived using the multivariate GC A-series. For that, the
multivariate GC A-series is derived by generalizing a specific derivation in (13) for the univariate Generalized Gram-Charlier (GGC) series expansion to multivariate using Kronecker algebra. The ExROT is also derived for gradient of multivariate density estimate. The empirical results on the standard test set for univariate density show the superiority of ExROT over ROT in all unimodal density estimation cases - skewed or kurtotic or with outliers and some of the multimodal cases. Thus, ExROT is a better option to ROT with comparable cost. The Chapter ends with the application of ExROT for previously derived estimators of FD and GFD based contrasts.

The Chapter 4 is devoted to the near-Independent BSS. It provides both theoretical and empirical local minima analysis of selected BSS contrasts for various source distributions. Then, it derives ICA algorithm using, Genetic Algorithm (GA) like, search based global optimization technique to allow BSS of near-independent source. It verifies the newly derived $L^2$-Norm of FD and $L^2$-Norm of GFD as BSS contrasts. The contrasts are estimated using the RIP and CRIP based formal expressions. The bandwidth selection for estimation is provided through both ROT and ExROT. The SRICA facilitates comparison of separation quality due to various BSS contrasts against varying source distributions by allowing use of same optimization algorithm for all the contrasts. The experimental results show failure of most ICA algorithms, including SRICA, in BSS of two sources with specific distributions and in BSS of higher dimensions. All the experimental results put together, bring further understanding of the overlearning phenomena and a discussion on the equivalence of ICA and BSS even in linear case. The failure of SRICA also necessitates focus on the success of GA as an optimization technique in the Chapter 5 and misconvergence of GA in higher dimensions in the Chapter 6.

The success of GA is explained through a notion of Schema (a template of similarity among the search strings) and Schema Theorem. The next Chapter 5 extends this GA search theory (GA algebra). The notion of schema, has been further generalized to dependency relation based Extended Formae from the current generalization as an equivalence relation based Formae. There are derived some operators exploiting the Extended Formae (plural of Formae) based similarities. Over all, the generalization achieves theoretical maximum possible schemata (plural of schema) for both the string and non-string representation structures using maximal alphabet. This has an impact on the current discussion on whether small alphabet (bi-nary) or maximal alphabet (float) for GA representation. Taking inspiration from the nature, the work recommends use of either an intermediate level alphabet - balancing maximal alphabet to avail maximum schemata and minimal alphabet to overcome some of the disadvantages due to maximum schemata - or varying representations during various stages of search. The above representation and operators are empirically used to derive Mendelian Genetic Algorithm (MGA). MGA, with abundance of schema, is inferred to avoid this misconvergence at least partially.

The Chapter 6 collaborates LSGO and BSS. It starts with the state of the art large scale
global optimization methods and the reason for possible misconvergence. The discussion also figures out that the BSS contrasts, in simultaneous mode, are non-separable functions, a difficult class of functions for LSGO. Towards the partial success to overcome misconvergence in GA and to reduce the computation for a non-separable global function optimization, there are discussed various search strategies with GA. The search strategies, for example, are - varying representations (gradual search), spiral search, delta search, refine search, population reinitialization and others. The concepts are mingled with existing Cooperative Coevolution search technique (95) and random grouping (143) for LSGO of standard test bench functions. The LSGO solutions are also applied to LSnIBSS.

Finally, the thesis ends with conclusion and possible future extensions in Chapter 7.

1.9 Contribution and Novelty

The contribution and novelty of the work can be briefed as under.

- For $k^{th}$ order differentiable multivariate functions with equal hyper volumes (region bounded by hyper surfaces) and added condition of bounded support, it is proved that equality of $k^{th}$ order derivatives implies equality of the corresponding functions.

- The $L^p$-norm of FD, GFD and HFD are derived as contrasts for BSS.

- The closed form expressions in terms of RIP and CRIP are derived for information field analysis using least squares and applied to estimate the derived $L^2$-norm of FD and GFD contrasts.

- The near-Gaussian PDF assumption and the Gram-Charlier A-Series based an Extended Rule-of-thumb is derived. It is experimentally proved to be better compare to ROT atleast in all sorts of (skewed or kurtotic or with outliers) unimodal density estimations. The ExROT is also derived for multivariate density estimation and gradient of univariate or multivariate density estimation.

- A specific derivation for the univariate GGC is extended to multivariate using kronecker algebra.

- The Search for Rotation based ICA (SRICA) algorithm using, GA like, global search based optimization method is derived.

- The near-Independent analysis brings some new conditions providing local minima. Also, it provides discussion on the use of ICA as BSS in linear case.
• The schema definition is further generalized to dependency relation based *Extended Forma* for both string and non-string structures of representation. The definition achieves theoretical maximum possible schemata with diploid representation (pair of chromosomes per individual) and corresponding operators. The new insights also leads to a discussion on whether maximal alphabet or minimal alphabet for representation.

• The novel Mendelian Genetic Algorithm using *Extended Forma* based representation and operators is derived.

• A varying representation, while search in progress, is achieved using gradual search and spiral search concepts. The concepts are empirically proved to be better than conventional search of a group of variables in nonseparable function global optimization. The concepts are mingled with existing Cooperative Coevolution search technique (95) and random grouping (143) for LSGO of standard test bench functions and a LSnIBSS application.