Encyclopedia of Artificial Intelligence (3 Volumes)
Edited By: Juan Ramón Rabuñal Dopico; Julian Dorado; Alejandro Pazos,
IGI Global (IGI) publishing company ISBN: 978-1-59904-849-9

Independent Subspaces

Lei Xu
Chinese University of Hong Kong, Hong Kong, & Peking University, Beijing, China

INTRODUCTION

Several unsupervised learning topics have been extensively studied with wide applications for decades in the literatures of statistics, signal processing, and machine learning. The topics are mutually related and certain connections have been discussed partly, but still in need of a systematical overview. The article provides a unified perspective via a general framework of independent subspaces, with different topics featured by differences in choosing and combining three ingredients. Moreover, an overview is made via three streams of studies. One consists of those on the widely studied principal component analysis (PCA) and factor analysis (FA), featured by the second order independence. The second consists of studies on a higher order independence featured independent component analysis (ICA), binary FA, and nonGaussian FA. The third is called mixture based learning that combines individual jobs to fulfill a complicated task. Extensive literatures make it impossible to provide a complete review. Instead, we aim at sketching a roadmap for each stream with attentions on those topics missing in the existing surveys and textbooks, and limited to the authors' knowledge.

A GENERAL FRAMEWORK OF INDEPENDENT SUBSPACES

A number of unsupervised learning topics are featured by its handling on a fundamental task. As shown in Fig.1(b), every sample $\boldsymbol{x}$ is projected into $\hat{x}$ on a manifold and the error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ of using $\hat{x}$ to represent $\boldsymbol{x}$ is minimized collectively on a set of samples. One widely studied situation is that a manifold is a subspace represented by linear coordinates, e.g., spanned by three linear independent basis vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ as shown in Fig.1(a). So, $\hat{x}$ can be represented by its projection $y^{(j)}$ on each basis vector, i.e.,

$$\hat{\boldsymbol{x}} = \sum_j^3 \boldsymbol{y}^{(1)} \boldsymbol{a}_j$$

or

$$x = \hat{x} + e = Ay + e, \ [y = \ y^{(1)}, y^{(2)}, y^{(3)}]^T. \qquad (1)$$

Typically, the error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ is measured by the square norm, which is minimized when $\boldsymbol{e}$ is orthogonal to $\hat{x}$. Collectively, the minimization of the average error $\|\boldsymbol{e}\|^2$ on a set of samples or its expectation $\boldsymbol{E}\|\boldsymbol{e}\|^2$ is featured by those natures given at the bottom of Fig.1(a).

Generally, the task consists of three ingredients, as shown in Fig.2. First, how the error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ is measured. Different measures define different projections. The square norm $\boldsymbol{d} = \|\boldsymbol{e}\|^2$ applies to a homogeneous medium between $\boldsymbol{x}$ and $\hat{x}$. Other measures are needed for inhomogeneous mediums. In Fig.1(c), a non-orthogonal but still linear projection is considered via $\boldsymbol{d} = \|\boldsymbol{e}\|_{\boldsymbol{B}}^2 = \boldsymbol{e}^{\boldsymbol{T}} \Sigma_e^{-1} \boldsymbol{e}$ with $\Sigma_e^{-1} = \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{B}$, as if $\boldsymbol{e}$ is first mapped to a homogeneous medium by a linear mapping $\boldsymbol{e}$ and then measured by the square norm. Shown at the bottom of Fig.1(c) are the natures of this $\boldsymbol{Min}\|\boldsymbol{e}\|_{\boldsymbol{B}}^2$. Being considerably different from those of $\boldsymbol{Min}\|\boldsymbol{e}\|^2$, more assumptions have to be imposed externally.

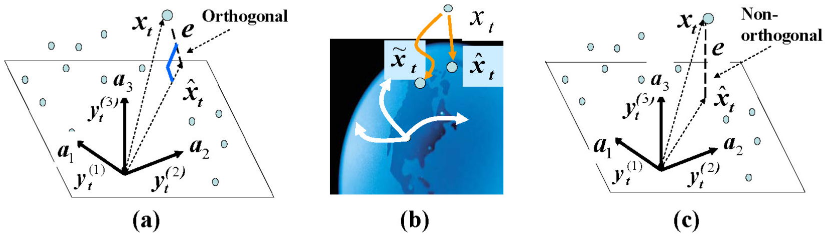
The second ingredient is a coordinate system, via either linear vectors in Fig.1(a)&(c) or a set of curves on a nonlinear manifold in Fig.1(b). Moreover, there is the third ingredient that imposes certain structure to further constrict how $\boldsymbol{y}$ is distributed within the coordinates, e.g., by the nature d).

The differences in choosing and combining the three ingredients lead to different approaches. We use the name "independent subspaces" to denote those structures with the components of $\boldsymbol{y}$ being mutually independent, and get a general framework for accommodating several unsupervised learning topics.

Subsequently, we summarize them via three streams of studies by considering

- $\boldsymbol{d} = \|\boldsymbol{e}\|_{\boldsymbol{B}}^2 = \boldsymbol{e}^{\boldsymbol{T}} \Sigma_e^{-1} \boldsymbol{e}$ and two special cases,
- three types of independence structure, and whether there is temporal structure among samples,
- varying from one linear coordinate system to multiple linear coordinate systems at different locations, as shown in Fig.2.

Figure 1



For $\min E\|e\|^2$, we have

Consequences

a) $\boldsymbol{e}$ from $\boldsymbol{G}(e \mid 0, \sigma_e^2 \boldsymbol{I})$, $\sigma_e^2 = \min \boldsymbol{E}\|e\|^2$,

b) the coordinates in $\boldsymbol{y}$ is reachable by an orthogonal transform $y = Wx$

c) $Eey^T = 0$ (i.e., not correlated)

Assumptions

d) for a unique A, it needs to impose that $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are orthonormal $A^T A = I$, which implies the nature: $Eyy^T = \Lambda$ *is diagonal, i.e.,* $y^{(1)}$, $y^{(2)}$, $y^{(3)}$ *are of the* 2nd *order independence.*

Indeterminacy

e) any rotation of the coordinate system leads to the same subspace, i.e., there is an indeterminacy of a rotation matrix Φ with $A' = \Phi A$: $A'^T A' = A^T \Phi^T \Phi A = A^T A = I$.

For $\min E\|e\|_B^2$, $\|e\|_B^2 = e^T \Sigma_e^{-1} e$, we have

Consequences

a) $\boldsymbol{e}$ from $\boldsymbol{G}(e \mid 0, \Sigma_e)$,

b) $\boldsymbol{y}$ is reached by a $y = Wx$, but $\boldsymbol{W}$ is non-orthogonal.

Assumptions

c) $Eey^T = 0$ (i.e., not correlated), here it is no longer a consequence.

d) $Eyy^T = \Lambda$ is diagonal, *while* $A^T A = I$ *is removed because it impedes* $E\|e\|_B^2$ *to reach its minimum.*

Indeterminacy

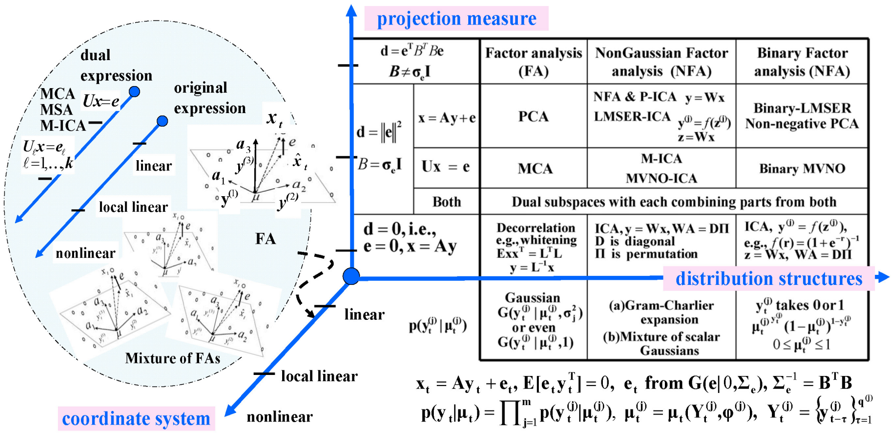
e) still a rotation matrix Φ since $A' = \Phi A$ spans the same subspace.

f) a diagonal D with $A' = AD, y' = D^{-1} y$: $Ay + e = A'y' + e$ & $Ey'y'^T$ is still diagonal.

g) a unknown allocation between the two additive terms in $Exx^T = A \Lambda A^T + \Sigma_e$.

Fig.1 A General Framework of Independent Subspaces

Figure 2



$d = e^T B^T B e$ $B \neq \sigma_e I$		Factor analysis (FA)	NonGaussian Factor analysis (NFA)	Binary Factor analysis (NFA)
$d = \|e\|^2$ $B = \sigma_e I$	$x = Ay + e$	PCA	NFA & P-ICA $y = Wx$ LMSER-ICA $y^{(j)} = f(z^{(j)})$ $z = Wx$	Binary-LMSER Non-negative PCA
	$Ux = e$	MCA	M-ICA MVNO-ICA	Binary MVNO
	Both	Dual subspaces with each combining parts from both		
$d = 0$, i.e., $e = 0, x = Ay$		Decorrelation e.g.,whitening $Exx^T = L^T L$ $y = L^{-1} x$	ICA, $y = Wx, WA = D\Pi$ D is diagonal Π is permutation	ICA, $y^{(j)} = f(z^{(j)})$, e.g., $f(r) = (1+e^{-r})^{-1}$ $z = Wx$, $WA = D\Pi$
$p(y_t^{(j)} \mid \mu_t^{(j)})$		Gaussian $G(y_t^{(j)} \mid \mu_t^{(j)}, \sigma_j^2)$ or even $G(y_t^{(j)} \mid \mu_t^{(j)}, 1)$	(a)Gram-Charlier expansion (b)Mixture of scalar Gaussians	$y_t^{(j)}$ takes 0 or 1 $\mu_t^{(j)\,y_t^{(j)}} (1-\mu_t^{(j)})^{1-y_t^{(j)}}$ $0 \le \mu_t^{(j)} \le 1$

$$x_t = Ay_t + e_t,\ E[e_t y_t^T] = 0,\ e_t \text{ from } G(e \mid 0, \Sigma_e),\ \Sigma_e^{-1} = B^T B$$

$$p(y_t \mid \mu_t) = \prod_{j=1}^{m} p(y_t^{(j)} \mid \mu_t^{(j)}),\ \mu_t^{(j)} = \mu_t(Y_t^{(j)}, \varphi^{(j)}),\ Y_t^{(j)} = \{y_{t-\tau}^{(j)}\}_{\tau=1}^{q^{(j)}}$$

Fig.2 Three gradients and their typical choices

STUDIES FEATURED BY SECOND ORDER INDEPENDENCE

We start at considering samples of independently and identically distributed (i.i.d.) by linear coordinates and an independent structure of a Gaussian $\mathbf{p(y_t^{(j)} \,|\, \text{ì}^{(j)})}$, with the projection measure varying as illustrated within the first column of the table in Fig.2. We encounter factor analysis (FA) in the general case $d = \|e\|_B^2 = e^T B^T Be$. At the special case $B = \sigma_e I$, the linear coordinates span a principal subspace of data. Further imposing $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$ and requiring the columns of $\boldsymbol{A}$ given by the first $\boldsymbol{m}$ principal components (PCs), i.e., eigenvectors that correspond the largest eigenvalues of $\Sigma = (\boldsymbol{B}^T \boldsymbol{B})^{-1}$. It becomes equivalent to PCA. Moreover, at the degenerated case $e = 0, y = xW$ de-correlates components of y, e.g., performing a pre-whitening as encountered in signal processing.

We summarize studies on the Roadmap A. The first stream originated from 100 years ago. The first adaptive learning one is Oja rule that finds the 1st-PC (i.e., the eigenvector that corresponds the largest eigenvalue of Σ), without explicitly estimating Σ. Extended to find multi-PCs, one way is featured by either an asymmetrical or a sequential implementation of the 1st-PC rule, but suffering error-accumulation. Details are referred to Refs.5,6,67,76,96 in (Xu, 2007a). The other way is finding multi-PCs symmetrically, e.g., Oja subspace rule. Further studies are summarized into the following branches:

MCA, Dual Subspace, and TLS Fitting

In (Xu, Krzyzak&Oja, 1991), a dual pattern recognition is suggested by considering both the principal subspace and its complementary subspace, as well as both the multiple PCs and its complementary counterparts--the components that correspond the smallest eigenvalues of Σ (i.e., the row vectors of U in Fig.2). Moreover, the first adaptive rule is proposed by eqn.(11a) in (Xu, Krzyzak&Oja, 1991) to get the component that corresponds the smallest eigenvalue of Σ, under the name Minor component analysis (MCA) firstly coined by Xu, Oja&Suen (1992), and it is also used for implementing a total least square (TLS) curve fitting. Subsequently, this topic has been brought to the signal processing literature by Gao, Ahmad & Swamy (1992) that was motivated by a visit of Gao to Xu's office where Xu introduced him the result of Xu,Oja&Suen (1992). Thereafter, adaptive MCA learning for TLS filtering becomes a popular topic of signal processing, see (Feng,Bao&Jiao,1998) and Refs.24,30,58,60 in (Xu,2007a).

It was also suggested in (Xu,Krzyzak&Oja,1992) that an implementation of PCA or MCA is made by switching the updating sign in the above eqn.(11a). Efforts were subsequently made to examine the existing PCA rules on whether they remain stable after such a sign switching. These jobs usually need tedious mathematical analyses of ODE stability, e.g., Chen & Amari (2001). An alternative way is turning an optimization of a PCA cost into a stable optimization of an induced cost for MCA, e.g., the LMSER cost is turned into one for subspace spanned by multiple MCs (Xu, 1994, see Ref.111, Xu2007a). A general method is further given by eqns(24-26) in (Xu, 2003) and then discussed in (Xu, 2007a).

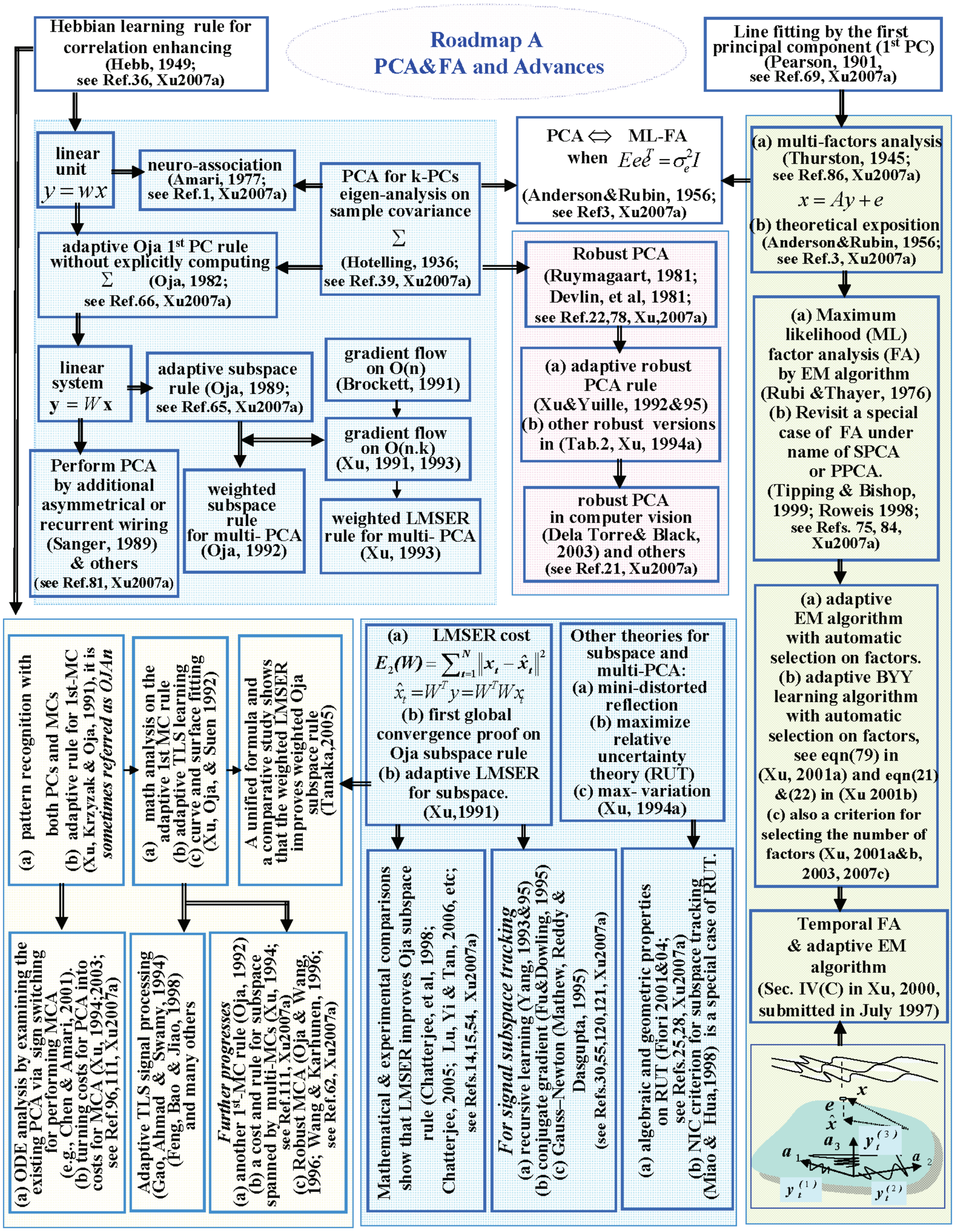
LMSER Learning and Subspace Tracking

A new adaptive PCA rule is derived from the gradient $\nabla E_2(W)$ for a least mean square error reconstruction (LMSER) (Xu,1991), with the first proof proposed on global convergence of Oja subspace rule--a task that was previously regarded as difficult. It was shown mathematically and experimentally that LMSER improves Oja rule by further comparative studies, e.g, see (Karhunen,Pajunen&Oja,1998) and see (Refs14,15,48,54,71,72, Xu2007a). Two years after (Xu,1991), this $E_2(W)$ is used for signal subspace tracking via a recursive least square technique (Yang,1993), then followed by others in the signal processing literature (Refs.33&55, Xu2007a). Also, PCA and subspace analysis can be performed by other theories or costs (Xu, 1994a&b). The algebraic and geometric properties were further analyzed on one of them, namely relative uncertainty theory (RUT), by Fiori (2000&04, see Refs.25,29, Xu2007a). Moreover, the NIC criterion for subspace tracking is actually a special case of this RUT, which can be observed by comparing eqn.(20) in (Miao& Hua,1998) with the equation of ρ_e at the end of Sec.III.B in (Xu,1994a).

Principal Subspace vs. Multi-PCs

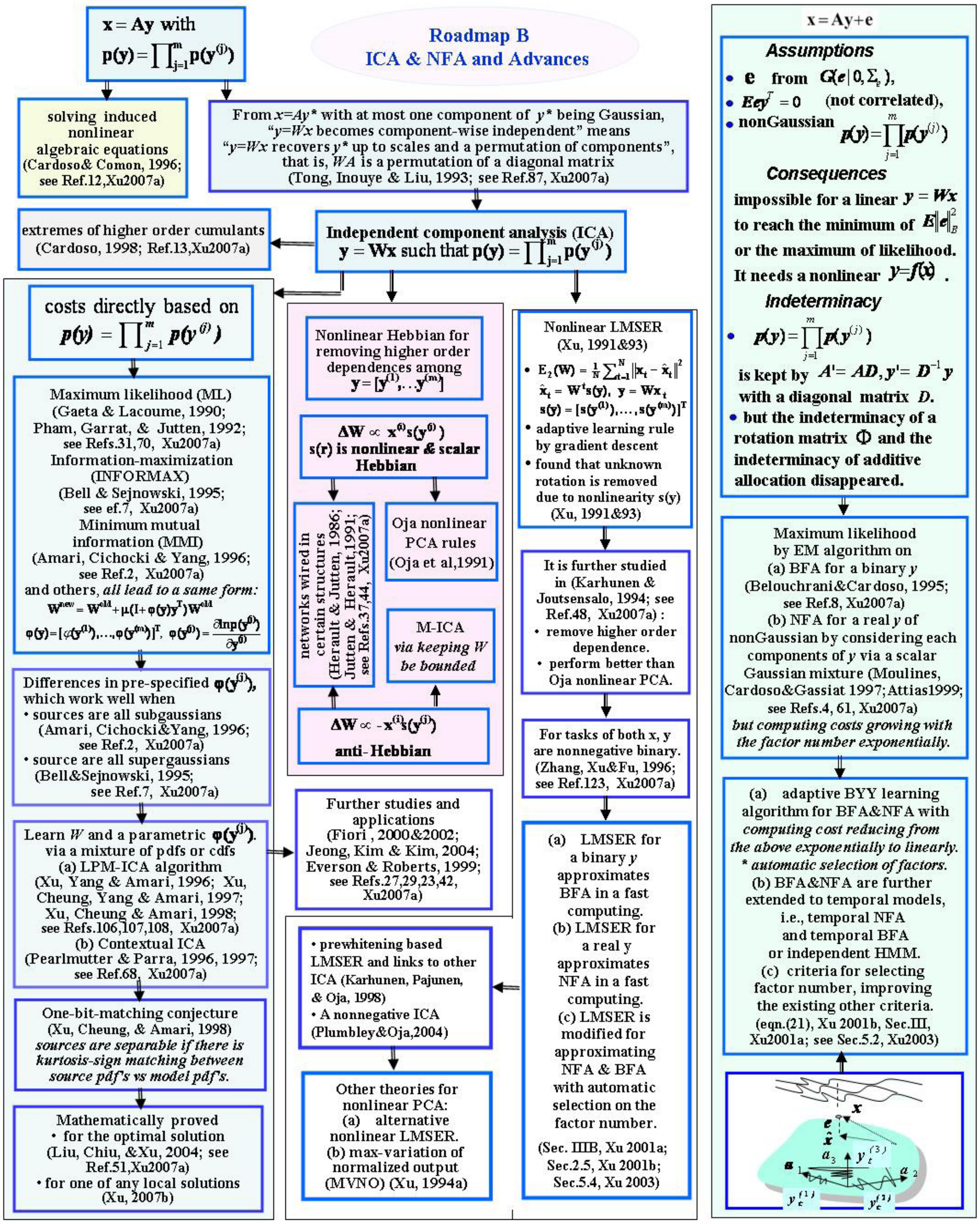
Oja subspace rule does not truly find the multi-PCs due to a rotation indeterminacy. Interestingly, it is demonstrated experimentally that adding a sigmoid function makes LMSER approximate the multi-PCs

Figure 3


Roadmap A
PCA&FA and Advances
Hebbian learning rule for correlation enhancing (Hebb, 1949; see Ref.36, Xu2007a)
Line fitting by the first principal component (1st PC) (Pearson, 1901, see Ref.69, Xu2007a)
linear unit $y = wx$
neuro-association (Amari, 1977; see Ref.1, Xu2007a)
PCA for k-PCs eigen-analysis on sample covariance Σ (Hotelling, 1936; see Ref.39, Xu2007a)
PCA ⇔ ML-FA when $Eee^T = \sigma_e^2 I$ (Anderson&Rubin, 1956; see Ref3, Xu2007a)
(a) multi-factors analysis (Thurston, 1945; see Ref.86, Xu2007a) $x = Ay + e$ (b) theoretical exposition (Anderson&Rubin, 1956; see Ref.3, Xu2007a)
adaptive Oja 1st PC rule without explicitly computing Σ (Oja, 1982; see Ref.66, Xu2007a)
Robust PCA (Ruymagaart, 1981; Devlin, et al, 1981; see Ref.22,78, Xu,2007a)
(a) Maximum likelihood (ML) factor analysis (FA) by EM algorithm (Rubi &Thayer, 1976) (b) Revisit a special case of FA under name of SPCA or PPCA. (Tipping & Bishop, 1999; Roweis 1998; see Refs. 75, 84, Xu2007a)
linear system $y = Wx$
adaptive subspace rule (Oja, 1989; see Ref.65, Xu2007a)
gradient flow on O(n) (Brockett, 1991)
gradient flow on O(n.k) (Xu, 1991, 1993)
(a) adaptive robust PCA rule (Xu&Yuille, 1992&95) (b) other robust versions in (Tab.2, Xu, 1994a)
Perform PCA by additional asymmetrical or recurrent wiring (Sanger, 1989) & others (see Ref.81, Xu2007a)
weighted subspace rule for multi- PCA (Oja, 1992)
weighted LMSER rule for multi- PCA (Xu, 1993)
robust PCA in computer vision (Dela Torre& Black, 2003) and others (see Ref.21, Xu2007a)
(a) adaptive EM algorithm with automatic selection on factors. (b) adaptive BYY learning algorithm with automatic selection on factors, see eqn(79) in (Xu, 2001a) and eqn(21) &(22) in (Xu 2001b) (c) also a criterion for selecting the number of factors (Xu, 2001a&b, 2003, 2007c)
(a) pattern recognition with both PCs and MCs (b) adaptive rule for 1st-MC (Xu, Krzyzak & Oja, 1991), it is *sometimes referred as OJAn*
(a) math analysis on the adaptive 1st MC rule (b) adaptive TLS learning (c) curve and surface fitting (Xu, Oja, & Suen 1992)
A unified formula and a comparative study shows that the weighted LMSER improves weighted Oja subspace rule (Tanaka,2005)
(a) LMSER cost $E_2(W) = \sum_{t=1}^{N} \|x_t - \hat{x}_t\|^2$ $\hat{x}_t = W^T y = W^T W x_t$ (b) first global convergence proof on Oja subspace rule (b) adaptive LMSER for subspace. (Xu,1991)
Other theories for subspace and multi-PCA: (a) mini-distorted reflection (b) maximize relative uncertainty theory (RUT) (c) max- variation (Xu, 1994a)
Temporal FA & adaptive EM algorithm (Sec. IV(C) in Xu, 2000, submitted in July 1997)
(a) ODE analysis by examining the existing PCA via sign switching for performing MCA (e.g., Chen & Amari, 2001). (b) turning costs for PCA into costs for MCA (Xu, 1994;2003; see Ref.96,111, Xu2007a)
Adaptive TLS signal processing (Gao, Ahmad & Swamy, 1994) (Feng, Bao & Jiao, 1998) and many others
Further progresses (a) another 1st-MC rule (Oja, 1992) (b) a cost and rule for subspace spanned by multi-MCs (Xu, 1994; see Ref.111, Xu2007a) (c) Robust MCA (Oja & Wang, 1996; Wang & Karhunen, 1996; see Ref.62, Xu2007a)
Mathematical & experimental comparisons show that LMSER improves Oja subspace rule (Chatterjee, et al, 1998; Chatterjee, 2005; Lu, Yi & Tan, 2006, etc; see Refs.14,15,54, Xu2007a)
For signal subspace tracking (a) recursive learning (Yang, 1993&95) (b) conjugate gradient (Fu&Dowling, 1995) (c) Gauss–Newton (Mathew, Reddy & Dasgupta, 1995) (see Refs.30,55,120,121, Xu2007a)
(a) algebraic and geometric properties on RUT (Fiori 2001&04; see Refs.25,28, Xu2007a) (b) NIC criterion for subspace tracking (Miao & Hua,1998) is a special case of RUT.
x
e
x̂
a3
a1
a2
y(1)t
y(2)t
y(3)t

Note: due to a limited space, it is impossible to put all the references into the reference list of this article. The problem is solved in help of (Xu, 2007a) via citing papers in its reference list where there are 123 entries. E.g., "see Refs.22,78, Xu2007a" means "see the entries [22][78] in the reference list of (Xu, 2007a).

Figure 4



Roadmap B
ICA & NFA and Advances

$\mathbf{x} = \mathbf{A}\mathbf{y}$ with
$p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$

solving induced nonlinear algebraic equations (Cardoso& Comon, 1996; see Ref.12,Xu2007a)

From $x=Ay^*$ with at most one component of y^* being Gaussian, "$y=Wx$ becomes component-wise independent" means "$y=Wx$ recovers y^* up to scales and a permutation of components", that is, WA is a permutation of a diagonal matrix (Tong, Inouye & Liu, 1993; see Ref.87, Xu2007a)

extremes of higher order cumulants (Cardoso, 1998; Ref.13,Xu2007a)

Independent component analysis (ICA)
$\mathbf{y} = \mathbf{W}\mathbf{x}$ such that $p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$

costs directly based on
$p(y) = \prod_{j=1}^{m} p(y^{(j)})$

Maximum likelihood (ML) (Gaeta & Lacoume, 1990; Pham, Garrat, & Jutten, 1992; see Refs.31,70, Xu2007a)
Information-maximization (INFORMAX) (Bell & Sejnowski, 1995; see ef.7, Xu2007a)
Minimum mutual information (MMI) (Amari, Cichocki & Yang, 1996; see Ref.2, Xu2007a)
and others, *all lead to a same form:*

$$\mathbf{W}^{new} = \mathbf{W}^{old} + \mu(I + \varphi(\mathbf{y})\mathbf{y}^T)\mathbf{W}^{old}$$

$$\varphi(\mathbf{y}) = [\varphi(y^{(1)}), \ldots, \varphi(y^{(m)})]^T, \ \varphi(y^{(j)}) = \frac{\partial \ln p(y^{(j)})}{\partial y^{(j)}}$$

Differences in pre-specified $\varphi(y^{(j)})$, which work well when
- sources are all subgaussians (Amari, Cichocki&Yang, 1996; see Ref.2, Xu2007a)
- source are all supergaussians (Bell&Sejnowski, 1995; see Ref.7, Xu2007a)

Learn W and a parametric $\varphi(y^{(j)})$. via a mixture of pdfs or cdfs
(a) LPM-ICA algorithm (Xu, Yang & Amari, 1996; Xu, Cheung, Yang & Amari, 1997; Xu, Cheung & Amari, 1998; see Refs.106,107,108, Xu2007a)
(b) Contextual ICA (Pearlmutter & Parra, 1996, 1997; see Ref.68, Xu2007a)

One-bit-matching conjecture (Xu, Cheung, & Amari, 1998) *sources are separable if there is kurtosis-sign matching between source pdf's vs model pdf's.*

Mathematically proved
- for the optimal solution (Liu, Chiu, &Xu, 2004; see Ref.51,Xu2007a)
- for one of any local solutions (Xu, 2007b)

Nonlinear Hebbian for removing higher order dependences among $\mathbf{y} = [y^{(1)}, \ldots y^{(m)}]$

$\Delta \mathbf{W} \propto \mathbf{x}^{(i)} s(y^{(j)})$
s(r) is nonlinear & scalar Hebbian

networks wired in certain structures (Herault & Jutten, 1986; Jutten & Herault,1991; see Refs.37,44, Xu2007a)

Oja nonlinear PCA rules (Oja et al,1991)

M-ICA *via keeping W be bounded*

$\Delta \mathbf{W} \propto -\mathbf{x}^{(i)} s(y^{(j)})$
anti-Hebbian

Further studies and applications (Fiori , 2000&2002; Jeong, Kim & Kim, 2004; Everson & Roberts, 1999; see Refs.27,29,23,42, Xu2007a)

- prewhitening based LMSER and links to other ICA (Karhunen, Pajunen, & Oja, 1998)
- A nonnegative ICA (Plumbley&Oja,2004)

Other theories for nonlinear PCA:
(a) alternative nonlinear LMSER.
(b) max-variation of normalized output (MVNO) (Xu, 1994a)

Nonlinear LMSER (Xu, 1991&93)
- $E_2(\mathbf{W}) = \frac{1}{N}\sum_{t=1}^{N} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2$
  $\hat{\mathbf{x}}_t = \mathbf{W}^t s(\mathbf{y}), \ \mathbf{y} = \mathbf{W}\mathbf{x}_t$
  $s(\mathbf{y}) = [s(y^{(1)}), \ldots, s(y^{(m)})]^T$
- adaptive learning rule by gradient descent
- found that unknown rotation is removed due to nonlinearity s(y) (Xu, 1991&93)

It is further studied in (Karhunen & Joutsensalo, 1994; see Ref.48, Xu2007a) :
- remove higher order dependence.
- perform better than Oja nonlinear PCA.

For tasks of both x, y are nonnegative binary. (Zhang, Xu&Fu, 1996; see Ref.123, Xu2007a)

(a) LMSER for a binary y approximates BFA in a fast computing.
(b) LMSER for a real y approximates NFA in a fast computing.
(c) LMSER is modified for approximating NFA & BFA with automatic selection on the factor number.
(Sec. IIIB, Xu 2001a; Sec.2.5, Xu 2001b; Sec.5.4, Xu 2003)

$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}$

Assumptions
- $\mathbf{e}$ from $G(e|0, \Sigma_e)$,
- $Eey^T = 0$ (not correlated),
- nonGaussian $p(y) = \prod_{j=1}^{m} p(y^{(j)})$

Consequences
impossible for a linear $y = Wx$ to reach the minimum of $E\|e\|_B^2$ or the maximum of likelihood. It needs a nonlinear $y = f(x)$.

Indeterminacy
- $p(y) = \prod_{j=1}^{m} p(y^{(j)})$ is kept by $A' = AD, y' = D^{-1} y$ with a diagonal matrix D.
- but the indeterminacy of a rotation matrix Φ and the indeterminacy of additive allocation disappeared.

Maximum likelihood by EM algorithm on
(a) BFA for a binary y (Belouchrani&Cardoso, 1995; see Ref.8, Xu2007a)
(b) NFA for a real y of nonGaussian by considering each components of y via a scalar Gaussian mixture (Moulines, Cardoso&Gassiat 1997;Attias1999; see Refs.4, 61, Xu2007a)
but computing costs growing with the factor number exponentially.

(a) adaptive BYY learning algorithm for BFA&NFA with *computing cost reducing from the above exponentially to linearly.*
* *automatic selection of factors.*
(b) BFA&NFA are further extended to temporal models, i.e., temporal NFA and temporal BFA or independent HMM.
(c) criteria for selecting factor number, improving the existing other criteria. (eqn.(21), Xu 2001b, Sec.III, Xu2001a; see Sec.5.2, Xu2003)

well (Xu,1991). Working at Harvard in the late summer 1991, Xu got aware of Brockett (1991) and thus extended the Brockett flow of $n \times n$ orthogonal matrices to that of $n \times n_1$ orthogonal matrices with $n > n_1$, from which two learning rules for truly the multi-PCs are obtained through modifying the LMSER rule and Oja subspace rule. The two rules were included as eqns (13)&(14) in Xu (1993) that was submitted in 1991, which are independent and also different from Oja (1992). Recently, Tanaka (2005) unifies these rules into one expression controlled by one parameter, and a comparative study was made to show that eqn(14) in (Xu,1993) turned out to be the most promising one.

Adaptive Robust PCA

In the statistics literature, robust PCA was proposed to resist outliers via a robust estimator on Σ. Xu&Yuille (1992&95) generalized the rules of Oja, LMSER, and MCA into robust adaptive learning by statistical physics, related to the Huber M-estimators. Also, the PCA costs in (Xu,1994b) are extended to robust versions in Tab.2 of (Xu, 1994a). Thereafter, efforts have been further made, including its use in computer vision, e.g., see (Refs9,21,45,52, Xu2007a).

On Roadmap A, another branch consists of advances on FA, which includes PCA as its special case at $\Sigma_e = \sigma_e^2 I$. In the past decade, there is a renewed interest on FA, not only the EM algorithm for FA is brought to implementing PCA, but also adaptive EM algorithm and other advances are developed in help of the Bayesian Ying Yang (BYY) harmony learning.

SUBSPACES OF HIGHER ORDER INDEPENDENCE

Noticing the table in Fig.2, we proceed as $\mathbf{p(y_t^{(j)} \mid \mu^{(j)})}$ becomes nonGaussian ones in the last two columns. Shown at the left-upper corner on Roadmap B, the degenerated case $e = 0$ leads to the problem of solving $x = Ay$ from samples of x and an independence constraint

$$\boldsymbol{p(y)} = \prod_{j=1}^{\boldsymbol{m}} \boldsymbol{p(y}^{(j)})$$

.

One way is solving induced nonlinear algebraic equations. Another way is called independent component analysis (ICA), tackled in the following four branches:

- Seeking extremes of the higher order cumulants of y.
- Using nonlinear Hebbian learning for removing higher order dependences among components of y, actually from which ICA studies originate.
- Optimizing a cost that bases on

  $$\boldsymbol{p(y)} = \prod_{j=1}^{\boldsymbol{m}} \boldsymbol{p(y}^{(j)})$$

  directly. As shown on Roadmap B, a same updating equation is reached from several aspects, with actual differences coming from pre-specifying the nonlinearity of $\phi(y^{(j)})$. One works when the source components of y^* are all subgaussians while the other works when the components of $y^*$are all supergaussians. This problem is solved by learning jointly $W$and $\phi(y^{(j)})$ via a parametric model. It is further found that a rough estimate of each source is already enough, which motivates the so called one-bit-matching conjecture that is recently proved to be true mathematically (Xu, 2007b).
- Implementing nonlinear LMSER (Xu, 1991&93). Details are referred to Roadmap B. Here, we add clarifications on two previous confusions. One relates to an omission of the origin of nonlinear LMSER. This has already been clarified in (Karhunen,Pajunen, &Oja,1998; Hyvarinen, Karhunen, & Oja, 2001;Plumbley &Oja,2004), clearly spelling out that the nonlinear $E_2(W)$ and its adaptive gradient rule were both proposed firstly in (Xu, 1991&93). The second confusion is about that ICA is usually regarded as a counterpart of PCA. As stated in (Xu,2001b&03) and observed from the Table in Fig.2, ICA by $y = xW$ is actually an extension of de-correlation analysis, in any combinations of PCs and MCs. The counterpart of MCA is minor ICA (M-ICA) while the counterpart of PCA is principal ICA (P-ICA).

In fact, the concept `principal' emerges from $e_t = x_t - Ay \neq 0$. As shown within the table in Fig.2 and on the rightmost column on Roadmap B, as $\mathbf{p(y_t^{(j)} \mid \mu^{(j)})}$

becomes nonGaussian ones, FA is extended to a binary FA (BFA) if y is binary, and a nonGaussian FA (NFA) if y is real but nonGaussian. Similar to FA performing PCA at $\Sigma_e = \sigma_e^2 I$, both BFA and NFA become to perform a P-ICA at $\Sigma_e = \sigma_e^2 I$.

Observing the first box in this column, for $e_t = x_t - Ay \neq 0$ we need to seek an appropriate nonlinear map $y = f(x)$. It usually has no analytical solution but needs an expensive computation to approximate. As discussed in (Xu, 2003), nonlinear LMSER uses a sigmoid nonlinearity $y_t^{(j)} = s(z_t^{(j)}), z = xW$ to avoid computing costs and approximately implements a BFA for a Bernoulli $p(y^{(j)})$ with a probability $p_j = \frac{1}{N}\sum_{t=1}^{N} s(z_t^{(j)})$ and a NFA for $p(y^{(j)})$ with a pseudo uniform distribution on $(-\infty, +\infty)$, as well as a nonnegative ICA (Plumbley&Oja,2004) when $p(y^{(j)})$ is on $[0, +\infty)$. However, further quantitative analysis is needed for this approximation.

Without approximation, the EM algorithm is developed for maximum likelihood learning since 1997, still suffering expensive computing costs. Favorably, further improvements have also been achieved by the BYY harmony learning. Details are referred to the rightmost column on Roadmap B.

TEMPORAL AND LOCALIZED EXTENSIONS

We further consider temporal samples shown at the bottom of the rightmost column on both Roadmap A and Roadmap B, via embedding a temporal structure in $\mathbf{p(y_t^{(j)} | \mu_t^{(j)})}$. A typical one is using

$$\boldsymbol{\mu}_\mathbf{t}^\mathbf{(j)} = \boldsymbol{\mu}^\mathbf{(j)}(\mathbf{Y}_\mathbf{t}^\mathbf{(j)}, \phi_\mathbf{j}), \quad \mathbf{Y}_\mathbf{t}^\mathbf{(j)} = \left\{\mathbf{y}_{\mathbf{t}-\tau}^\mathbf{(j)}\right\}_{\tau=1,}^{\mathbf{q^{(j)}}}$$

e.g., a linear regression

$$\boldsymbol{\mu}_\mathbf{t}^\mathbf{(j)} = \sum_{\tau=1}^{q^{(j)}} \beta_\tau^{(j)} \mathbf{y}_{\mathbf{t}-\tau,}^\mathbf{(j)}$$

to turn a model (e.g., one in the table of Fig.2) into temporal extensions. Information is carried over time in two ways. One is computing $\boldsymbol{\mu}_\mathbf{t}^\mathbf{(j)}$ by the regression, with learning on $\boldsymbol{\mu}_\mathbf{t}^\mathbf{(j)}$ made through the gradient with respect to $\varphi_\mathbf{j}$ by a chain rule. The second is computing $\int \mathbf{p(y_t^{(j)} | \mu_t^{(j)}) p(Y_t^{(j)})\, dY_t^{(j)}}$ and getting the gradient with respect to $\varphi_\mathbf{j}$. Details are referred to Xu (2000&01a&03).

Next, we move to multiple subspaces at different locations as shown in Fig.2. Studies are summarized on Roadmap C, categorized according to one key point, i.e., a scheme $\mathrm{p}_{\ell,\mathrm{t}}$ that allocates a sample $\boldsymbol{x_t}$ to different subspaces. This $\mathrm{p}_{\ell,\mathrm{t}}$ bases on two issues.

One is a local measure on how the ℓ -th subspace is suitable for representing $\boldsymbol{x_t}$. The other is a mechanism that summarizes the local measures of subspaces to yield $\mathrm{p}_{\ell,\mathrm{t}}$. One typical mechanism is that emerges in the EM algorithm for the maximum likelihood or Bayesian learning, where $\boldsymbol{x_t}$ is fractionally allocated among subspaces proportional to their local measures. Another typical mechanism is that $\boldsymbol{x_t}$ is nonlinearly located to one or more winners via a competition based on the local measures, e.g,, as in the classic competitive learning and the rival penalized competitive learning (RPCL).

Also, a scheme $\mathrm{p}_{\ell,\mathrm{t}}$ may come from blending both types of mechanisms, as that from the BYY harmony learning. Details are referred to (Xu,2007c) and its two http-sites.

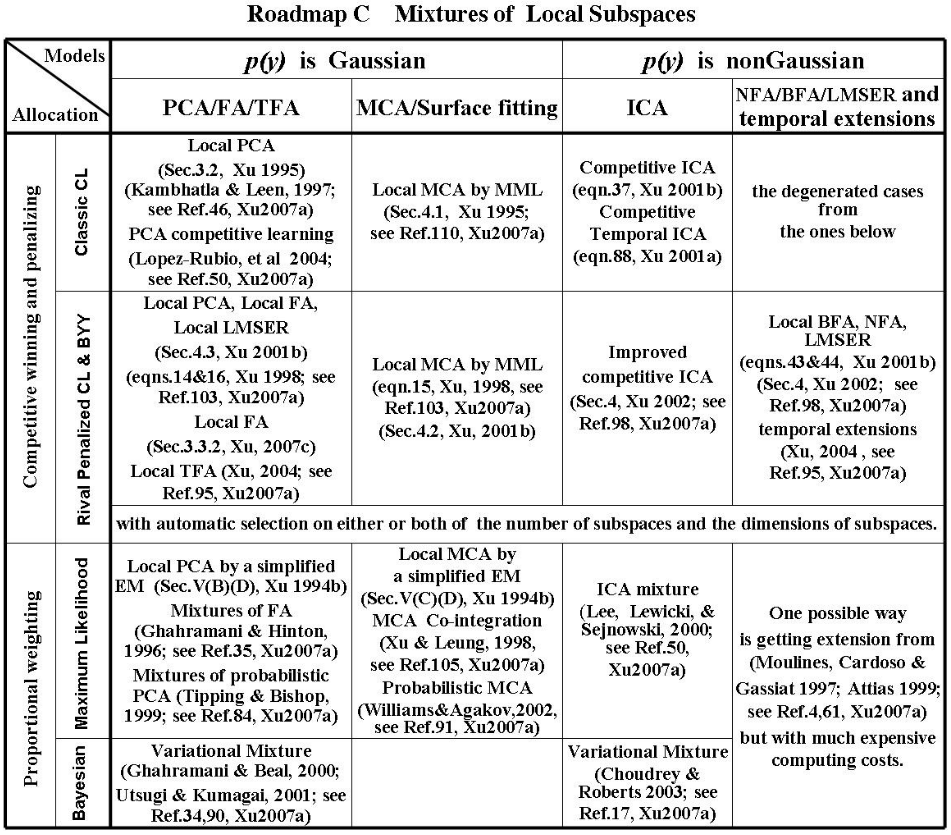
FUTURE TRENDS

Another important task is how to determine the number $\boldsymbol{k}$ of subspaces and the dimension $\boldsymbol{m}_\ell$ of each subspace. It is called model selection, usually implemented in two phases. First, a set of candidates are considered by enumerating $\boldsymbol{k}$ and $\boldsymbol{m}_\ell$, with unknown parameters estimated by the maximum likelihood learning. Second, the best among the candidates is selected by one of criteria, such as AIC, CAIC, SIC/BIC/MDL, Cross Validation, etc. However, this two-phase implementation is computationally very extensive. Moreover, the performance will degenerate considerably when the sample size is finite while $\boldsymbol{k}$ and $\boldsymbol{m}_\ell$ are not too small.

One trend is letting model selection to be made automatically during learning, i.e., on a candidate with $\boldsymbol{k}$ and $\boldsymbol{m}_\ell$ initially being large enough, learning not only determines unknown parameters but also automatically shrinks $\boldsymbol{k}$ and $\boldsymbol{m}_\ell$ to appropriate ones. Two such efforts are RPCL and the BYY harmony learning. Details are referred to (Xu,2007c) and its two http-sites.

Also, there are open issues on $x = Ay + e$, $e \neq 0$, with components of y mutually independent in higher order statistics. Some are listed below:

Figure 5



Roadmap C Mixtures of Local Subspaces

Models / Allocation		*p(y)* is Gaussian		*p(y)* is nonGaussian	
		PCA/FA/TFA	MCA/Surface fitting	ICA	NFA/BFA/LMSER and temporal extensions
Competitive winning and penalizing	Classic CL	Local PCA (Sec.3.2, Xu 1995) (Kambhatla & Leen, 1997; see Ref.46, Xu2007a) PCA competitive learning (Lopez-Rubio, et al 2004; see Ref.50, Xu2007a)	Local MCA by MML (Sec.4.1, Xu 1995; see Ref.110, Xu2007a)	Competitive ICA (eqn.37, Xu 2001b) Competitive Temporal ICA (eqn.88, Xu 2001a)	the degenerated cases from the ones below
	Rival Penalized CL & BYY	Local PCA, Local FA, Local LMSER (Sec.4.3, Xu 2001b) (eqns.14&16, Xu 1998; see Ref.103, Xu2007a) Local FA (Sec.3.3.2, Xu, 2007c) Local TFA (Xu, 2004; see Ref.95, Xu2007a)	Local MCA by MML (eqn.15, Xu, 1998, see Ref.103, Xu2007a) (Sec.4.2, Xu, 2001b)	Improved competitive ICA (Sec.4, Xu 2002; see Ref.98, Xu2007a)	Local BFA, NFA, LMSER (eqns.43&44, Xu 2001b) (Sec.4, Xu 2002; see Ref.98, Xu2007a) temporal extensions (Xu, 2004 , see Ref.95, Xu2007a)
		with automatic selection on either or both of the number of subspaces and the dimensions of subspaces.			
Proportional weighting	Maximum Likelihood	Local PCA by a simplified EM (Sec.V(B)(D), Xu 1994b) Mixtures of FA (Ghahramani & Hinton, 1996; see Ref.35, Xu2007a) Mixtures of probabilistic PCA (Tipping & Bishop, 1999; see Ref.84, Xu2007a)	Local MCA by a simplified EM (Sec.V(C)(D), Xu 1994b) MCA Co-integration (Xu & Leung, 1998, see Ref.105, Xu2007a) Probabilistic MCA (Williams&Agakov,2002, see Ref.91, Xu2007a)	ICA mixture (Lee, Lewicki, & Sejnowski, 2000; see Ref.50, Xu2007a)	One possible way is getting extension from (Moulines, Cardoso & Gassiat 1997; Attias 1999; see Ref.4,61, Xu2007a) but with much expensive computing costs.
	Bayesian	Variational Mixture (Ghahramani & Beal, 2000; Utsugi & Kumagai, 2001; see Ref.34,90, Xu2007a)		Variational Mixture (Choudrey & Roberts 2003; see Ref.17, Xu2007a)	

- Which part of unknown parameters in $x = Ay + e$ can be determined uniquely ?
- Under which conditions, the independence

$$p(y) = \prod_{j=1}^{m} p(y^{(j)})$$

can be ensured in concept? Can it be further achieved by a learning algorithm?

- In what a sense, both ensuring

$$p(y) = \prod_{j=1}^{m} p(y^{(j)})$$

and the best reconstruction of x by $\hat{x} = Ay$ can be achieved simultaneously? If not, what is the best nonlinear $y = f(x)$ in term of both

$$p(y) = \prod_{j=1}^{m} p(y^{(j)})$$

and $e \neq 0$?

- Can such a best be obtained analytically or via an effective computing?

CONCLUSION

Studies of three closely related unsupervised learning streams have been overviewed in an extensive scope and from a systematic perspective. A general framework of independent subspaces is presented, from which a number of learning topics are summarized via different features of choosing and combining the three basic ingredients.

ACKNOWLEDGMENT

The work is supported by Chang Jiang Scholars Program by Chinese Ministry of Education for Chang Jiang Chair Professorship in Peking University.

REFERENCES

Brockett, R.W., (1991), Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems, Linear Algebra and Its Applications 146,79-91.

Chen, T., & Amari, S., (2001), Unified stabilization approach to principal and minor components extraction algorithms, Neural Networks 14(10),1377–1387.

Feng, D.Z., Bao, Z., & Jiao, L.C., (1998), Total least mean squares algorithm, IEEE Transactions Signal Processing 46,2122–2130.

Gao, K., Ahmad, M.O., & Swamy, M.N., (1992), Learning algorithm for total least-squares adaptive signal processing, Electronic Letters 28(4),430–432.

Hyvarinen, A., Karhunen, J., & Oja, E., (2001), Independent component analysis, John Wiley, NY, 2001.

Karhunen, J., Pajunen, P. & Oja , E., (1998), The nonlinear PCA criterion in blind source separation: relations with other approaches, Neurocomputing 22,5-20.

Miao, Y.F., & Hua, Y.B., (1998), Fast subspace tracking and neural network learning by a novel information criterion, IEEE Transactions Signal Processing 46,1967–79.

Oja, E., (1992), Principal components, minor components, and linear neural networks, Neural Networks 5,927-935.

Oja, E., Ogawa, H., & Wangviwattana, J., (1991), Learning in nonlinear constrained Hebbian networks, Proc.ICANN’91, 385-390.

Plumbley, M.D., & Oja, E., (2004), A “nonnegative PCA” algorithm for independent component analysis, IEEE Transactions Neural Networks 15(1),66-76.

Tanaka, T., (2005), Generalized weighted rules for principal components tracking, IEEE Transactions Signal Processing 53(4),1243- 1253.

Xu, L., (2007a), A unified perspective on advances of independent subspaces: basic, temporal, and local structures, Proc.6th.Intel.Conf.Machine Learning and Cybernetics, Hong Kong, 19-22 Aug.2007, 767-776.

Xu, L., (2007b), One-bit-matching ICA theorem, convex-concave programming, and distribution approximation for combinatorics, Neural Computation 19,546-569.

Xu, L., (2007c), A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving, Pattern Recognition 40,2129-2153. Also see http://www.scholarpedia.org/article/Rival_Penalized_Competitive_Learning http://www.scholarpedia.org/article/Bayesian_Ying_Yang_Learning.

Xu, L., (2003), Independent component analysis and extensions with noise and time: A Bayesian Ying-Yang learning perspective, Neural Information Processing Letters and Reviews 1(1),1-52.

Xu, L., (2001a), BYY harmony learning independent state space and generalized APT financial analyses, IEEE Transactions Neural Networks 12,822–849.

Xu, L., (2001b), An Overview on Unsupervised Learning from Data Mining Perspective, Advances in Self-Organizing Maps, Allison et al, Eds., Springer, 2001,181–210.

Xu, L., (2000), Temporal BYY learning for state space approach, hidden Markov model and blind source separation, IEEE Transactions Signal Processing 48,2132–2144.

Xu, L., Cheung, C.C., & Amari, S., (1998), Learned parametric mixture based ICA algorithm, Neurocomputing 22,69-80.

Xu, L., (1994a), Beyond PCA learning: from linear to nonlinear and from global representation to local representation, Proc.ICONIP94, Vol.2,943-949.

Xu, L., (1994b), Theories for unsupervised learning: PCA and its nonlinear extensions, Proc.IEEE ICNN94, Vol.II,1252-1257.

Xu, L., (1993), Least mean square error reconstruction principle for self-organizing neural-nets, Neural Networks 6,627–648.

Xu, L., Oja, E., & Suen, C.Y., (1992), Modified Hebbian learning for curve and surface fitting, Neural Networks 5,393-407.

Xu, L., & Yuille, A.L., (1992&95), Robust PCA learning rules based on statistical physics approach, Proc.IJCNN92-Baltimore, Vol.I:812-817. An extended version on IEEE Transactions Neural Networks 6,131–143.

Xu, L., (1991), Least MSE reconstruction for self-organization, Proc.IJCNN91-Singapore, Vol.3,2363-73.

Xu, L., Krzyzak, A., & Oja, E., (1991), A neural net for dual subspace pattern recognition methods, International Journal Neural Systems 2(3),169-184.

Yang, B., (1993), Subspace tracking based on the projection approach and the recursive least squares method, Proc.IEEE ICASSP93, Vol.IV,145–148.

KEY TERMS

BYY Harmony Learning: It is a statistical learning theory for a two pathway featured intelligent system via two complementary Bayesian representations of the joint distribution on the external observation and its inner representation, with both parameter learning and model selection determined by a principle that two Bayesian representations become best harmony. See http://www.scholarpedia.org/article/Bayesian_Ying_Yang_Learning.

Factor Analysis: A set of samples $\{\mathbf{x}_t\}_{t=1}^{N}$ is described by a linear model $x = Ay + \mu + e$, where μ is a constant, y and e are both from Gaussian and mutually uncorrelated, and components of y are called factors and mutually uncorrelated. Typically, the model is estimated by the maximum likelihood principle.

Independence Subspaces: It refers to a family of models, each of which consists of one or several subspaces. Each subspace is spanned by linear independent basis vectors and the corresponding coordinates are mutually independent.

Least Mean Square Error Reconstruction (LMSER): For an orthogonal projection x_t onto a subspace spanned by the column vectors of a matrix W, maximizing $\frac{1}{N}\sum_{t=1}^{N}(w^{T}\mathbf{x}_t)^2$ subject to $W^{t}W = I$ is equivalent to minimizing the mean square error $\frac{1}{N}\sum_{t=1}^{N}\left\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\right\|^2$ by using the projection $\hat{\mathbf{x}}_t = \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}_t$ as reconstruction of x_t, which is reached when W spans the same subspace spanned by the PCs.

Minor Component (MC): Being orthogonal complementary to the PC, the solution of $\mathbf{min}_{(\mathbf{w}^{t}\mathbf{w}=1\}}\mathbf{J}(\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{w}^{r}\mathbf{x}_t)^2 = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\,\mathbf{w}$ is the MC, while the m-MCs are referred to the columns of W that minimizes $\mathbf{J}(W) = \frac{1}{N}\sum_{t=1}^{N}\| W^{r}\mathbf{x}_t \|^2 = Tr[W^{\mathrm{T}}\boldsymbol{\Sigma}\,W]$ subject to $W^{t}W = I$.

Principal Component (PC): For samples $\{\mathbf{x}_t\}_{t=1}^{N}$ with a zero mean, its PC is a unit vector $\boldsymbol{w}$ originated at zero with a direction along which the average of the orthogonal projection by every sample is maximized, i.e., $\mathbf{max}_{(\mathbf{w}^{t}\mathbf{w}=1\}}\mathbf{J}(\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{w}^{T}\mathbf{x}_t)^2 = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\,\mathbf{w}$, the solution is the eigenvector of the sample covariance matrix $\boldsymbol{\Sigma} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}$, corresponding to the largest eigen-value. Generally, the m-PCs are referred to the m orthonormal vectors as the columns of W that maximizes $\mathbf{J}(W) = \frac{1}{N}\sum_{t=1}^{N}\| W^{r}\mathbf{x}_t \|^2 = Tr[W^{\mathrm{T}}\boldsymbol{\Sigma}W]$.

Rival Penalized Competitive Learning: It is a development of competitive learning in help of an appropriate balance between participating and leaving mechanisms, such that an appropriate number of agents or learners will be allocated to learn multiple structures underlying observations. See http://www.scholarpedia.org/article/Rival_Penalized_Competitive_Learning.

Total Least Square (TLS) Fitting: Given samples $\{\mathbf{z}_t\}_{t=1}^{N}$, $\mathbf{z}_t = [\mathbf{y}_t, \mathbf{x}_t^{\mathrm{T}}]^{\mathrm{T}}$, instead of finding a vector $\boldsymbol{w}$ to minimize the error $\frac{1}{N}\sum_{t=1}^{N}\left\|y_t - \boldsymbol{w}^{T}\boldsymbol{x}_t\right\|^2$, the TLS fitting is finding an augmented vector $\tilde{w} = [w^{T}, c]^{T}$ such that the error $\frac{1}{N}\sum_{t=1}^{N}\left\|\tilde{\boldsymbol{w}}^{T}\boldsymbol{z}_t\right\|^2$ is minimized subject to $\tilde{w}^{T}\tilde{w} = 1$, the solution is the MC of $\{\mathbf{z}_t\}_{t=1}^{N}$.