(Keynote presentation)

A UNIFIED PERSPECTIVE ON ADVANCES OF INDEPENDENT SUBSPACES: BASIC, TEMPORAL, AND LOCAL STRUCTURES

LEI XU

Dept. of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong
E-MAIL: lxu@cse.cuhk.edu.hk

Abstract:

A general framework of independent subspaces is presented, based on which a number of unsupervised learning topics have been summarized from a unified perspective, featured by different combinations of three basic ingredients. Moreover, advances on these topics are overviewed in three streams, with roadmaps sketched. One consists of studies on the second order independence featured principal component analysis (PCA) and factor analysis (FA), in adaptive and robust implementations as well as with duality and temporal extensions. The other consists of studies on the higher order independence featured independent component analysis (ICA), binary FA, and nonGaussian FA. The third is called mixture based learning that combines the above individual tasks, proportionally or competitively to fulfill a complicated task.

Keywords:

Independence; Subspaces; PCA; MCA; Hebbian learning; Factor analysis (FA); ICA; Temporal FA; Binary FA; nonGaussian FA; Local subspaces; Local FA; Finite mixtures

1. Independent subspaces in a general framework

As shown in Fig.1(a), a sample x is projected to $\hat{x}$ on a subspace with an error $e=x-\hat{x}$. It is nature to minimize the error of using $\hat{x}$ to represent x, with the error e measured by its square length $d=\|e\|^2$ or the Euclidean distance between $\hat{x}$ and x. The minimum is reached when e is orthogonal to the subspace. Moreover, the subspace can be represented by a linear coordinate system, i.e., spanned by three linear independent unit basis vectors a_1, a_2, a_3, $\hat{x}$ can be further represented by its projection on each of the basis vectors, resulting in the coordinates $y=[y^{(1)},y^{(2)},y^{(3)}]^T$, i.e., $\hat{x}=\sum_j^3 y^{(1)} a_j$ or generally

$$x=\hat{x}+e=Ay+e \qquad (1)$$

where $\min E\|e\|^2$ is featured by the following natures:

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

b) e from $G(e|0,\sigma_e^2 I)$, $\sigma_e^2=\min E\|e\|^2$,

c) the coordinates in y is reached by an orthogonal transform $y=Wx$.

However, $\min E\|e\|^2$ does not lead to a unique A. Instead, it can consist of any linear independent unit basis vectors. To reduce this indeterminacy, we impose that

d) a_1, a_2, a_3 are orthonormal basis vectors (i.e., $A^TA=I$), which implies that $Eyy^T=\Lambda$ is a diagonal matrix, i.e., $y^{(1)}$, $y^{(2)}$, $y^{(3)}$ are mutually not correlated or independent in a 2nd order statistics sense.

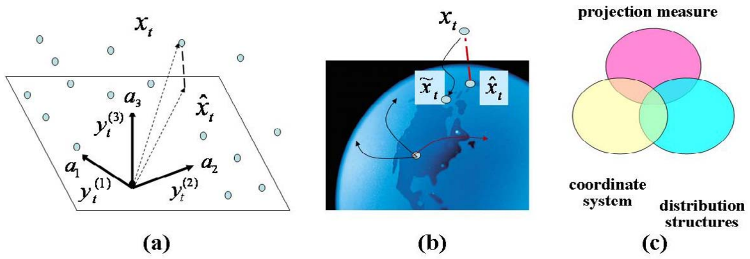


Fig.1 A general framework of independent subspaces

We further move to a general case shown in Fig.1(b). For a meaningful projection $\hat{x}$ on a manifold. there are at least three basic ingredients to be specified. First, the error $e=x-\hat{x}$ needs a measure, based on which a minimum error projection can be implemented. Actually, different measures define different projections. One choice is $d=\|e\|^2$ for a homogeneous medium between $\hat{x}$ and x. For inhomogeneous mediums, other choices can be used. One choice is $d=\|e\|_B^2=e^T\Sigma_e^{-1}e$ with $\Sigma_e^{-1}=B^TB$, as if e is

1-4244-0973-X/07/$25.00 ©2007 IEEE

mapped to a homogeneous medium by a linear mapping B and measured by Euclidean distance. In Fig.2 this measure is considered, with its special case $d=\|e\|^2$ (i.e., $B=I$) and the degenerated case e=0.

To further represent $\hat{x}$ within the manifold, we need the second ingredient, i.e., a coordinate system on the manifold, via either linear vectors in Fig.1(a) or a set of curves in Fig.1(b). In Fig.2, we still start at a linear coordinate system in Fig.1(a). However, minimizing $\|e\|_B^2$ no longer implies the nature a). Instead, it should be explicitly imposed as a condition. Also, the natures b) & c) are modified into

b) e from $G(e|0,\Sigma_e)$,

c) the coordinates y is reached by a linear $y=Wx$, but W is no longer orthogonal.

Moreover, impeding to reach the minimum $\|e\|_B^2$, the condition of $A^TA=I$ has to be replaced by a weaker condition:

d) $Eyy^T=\Lambda$ is a diagonal matrix, i.e., components of y becomes uncorrelated.

A re-scaling on components of y will not affect the natures a) & d). That is, there is an indeterminacy of a unknown diagonal matrix D. Also, there is an indeterminacy of a unknown rotation matrix Φ, and an indeterminacy on a specific allocation to its two additive terms in $Exx^T=A\Lambda A^T+\Sigma_e$.

$p(y_t^{(j)}\mid\mu_t^{(j)})$ projection measure		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 $\mu_t^{(j)\,y_t^{(j)}}(1-\mu_t^{(j)})^{1-y_t^{(j)}}$ $0\le\mu_t^{(j)}\le 1$
$d=0$, i.e., $e=0, x=Ay$		Decorrelation e.g., whitening $Exx^T=L^TL$ $y=L^{-1}x$	ICA, $y=Wx$, $WA=D\Pi$ D is diagonal Π is permutation	ICA, $y^{(j)}=f(z^{(j)})$, e.g., $f(r)=(1+e^{-r})^{-1}$ $z=Wx$, $WA=D\Pi$
$d=\|e\|^2$ e from $G(e\mid 0,\sigma_e^2 I)$	$x=Ay+e$	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		
$d=e^TB^TBe$ $x=Ay+e$ $e\sim G(x\mid 0,B^TB)$		Factor analysis (FA)	NonGaussian Factor analysis (NFA)	Binary Factor analysis (NFA)

Tab.1 Typical Examples of One Subspace for i.i.d. Samples

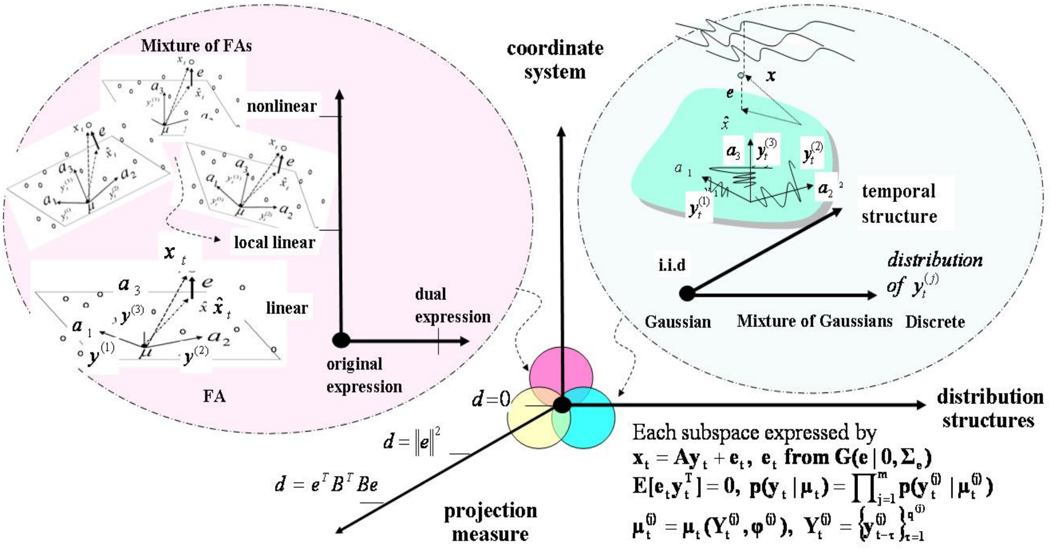

Mixture of FAs
nonlinear
local linear
linear
FA
x_t
e
$\hat{x}_t$
a_1
a_2
a_3
$y^{(1)}$
$y^{(2)}$
$y^{(3)}$
μ
coordinate system
dual expression
original expression
$d=0$
$d=\|e\|^2$
$d=e^TB^TBe$
projection measure
distribution structures
x
e
$\hat{x}$
$y_t^{(1)}$
$y_t^{(2)}$
$y_t^{(3)}$
temporal structure
i.i.d
distribution of $y_t^{(j)}$
Gaussian
Mixture of Gaussians
Discrete
Each subspace expressed by
$\mathbf{x}_t=\mathbf{A}\mathbf{y}_t+\mathbf{e}_t$, $\mathbf{e}_t$ from $\mathbf{G}(\mathbf{e}\mid 0,\Sigma_e)$
$\mathbf{E}[\mathbf{e}_t\mathbf{y}_t^T]=0$, $\mathbf{p}(\mathbf{y}_t\mid\boldsymbol{\mu}_t)=\prod_{j=1}^m \mathbf{p}(\mathbf{y}_t^{(j)}\mid\boldsymbol{\mu}_t^{(j)})$
$\boldsymbol{\mu}_t^{(j)}=\boldsymbol{\mu}_t(\mathbf{Y}_t^{(j)},\boldsymbol{\varphi}^{(j)})$, $\mathbf{Y}_t^{(j)}=\{\mathbf{y}_{t-\tau}^{(j)}\}_{\tau=1}^{q^{(j)}}$

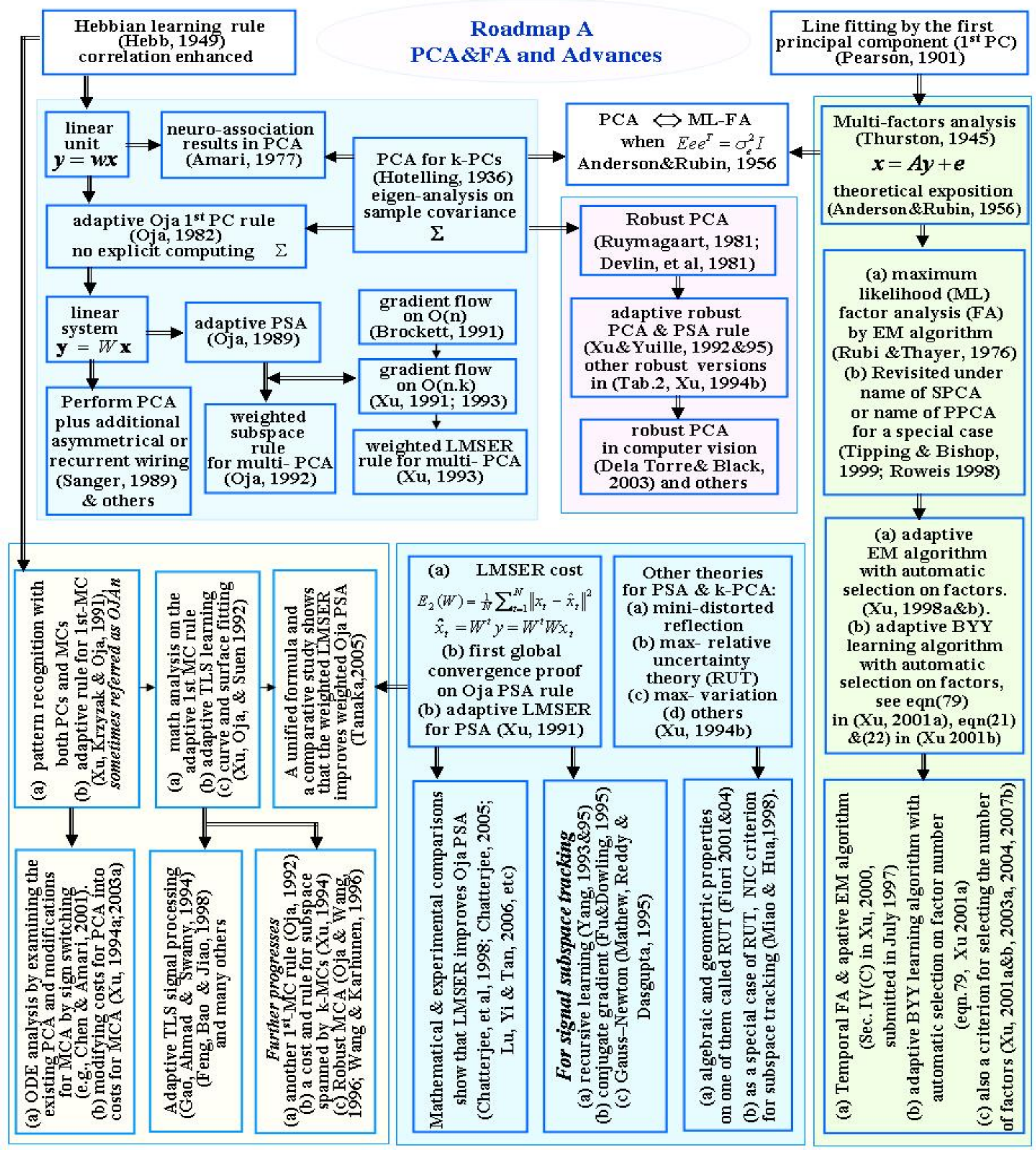
Fig.2 Typical variants of three ingredients

In fact, the above nature d) represents an example of the second ingredient, i.e., how y is distributed within an coordinate system. In Fig.2 we further extend it to an independence of any order statistics. The third ingredient varies from one linear coordinate system to multiple linear coordinate systems at different locations. Alternatively, each subspace can also be represented by another linear coordinate system for its complementary orthogonal subspace. Therefore, different specific types of independent subspaces can be summarized from a unified perspective, featured by different combinations of the three ingredients, as shown in Fig.2. Also, from this perspective, an overview can be made on the past studies.

2. Independent subspaces of 2nd order independence

We start at a subspace for samples of independently and identically distributed (i.i.d.), as shown in Tab.1 and the two bottom dimensions in Fig.2. Thus, there is no temporal structure, we have $\mathbf{Y}_t^{(j)}$ empty and $\boldsymbol{\mu}_t^{(j)}=\boldsymbol{\mu}^{(j)}$.

For a Gaussian $\mathbf{p}(\mathbf{y}_t^{(j)}\mid\boldsymbol{\mu}^{(j)})$, we encounter factor analysis (FA) at the general case $d=\|e\|_B^2=e^TB^TBe$ and PCA at *B*=*I*. Particularly, at the degenerated case ***e*=0** we have that ***y*=*Wx*** de-correlates components of ***y***. The pre-whitening in signal processing is such an example. Another is considering that the orthogonal complementary subspace spanned by the row vectors of ***U***, which leads to MCA. With these insights, we summarize the related existing studies on the Roadmap A. Instead of providing a complete review. It aims at a sketch with attentions put on


Roadmap A
PCA&FA and Advances
Hebbian learning rule (Hebb, 1949) correlation enhanced
Line fitting by the first principal component (1st PC) (Pearson, 1901)
linear unit $y = wx$
neuro-association results in PCA (Amari, 1977)
PCA for k-PCs (Hotelling, 1936) eigen-analysis on sample covariance Σ
PCA ⟺ ML-FA when $Eee^T=\sigma_e^2 I$ Anderson&Rubin, 1956
Multi-factors analysis (Thurston, 1945) $x = Ay + e$ theoretical exposition (Anderson&Rubin, 1956)
adaptive Oja 1st PC rule (Oja, 1982) no explicit computing Σ
Robust PCA (Ruymagaart, 1981; Devlin, et al, 1981)
linear system $y = Wx$
adaptive PSA (Oja, 1989)
gradient flow on O(n) (Brockett, 1991)
gradient flow on O(n.k) (Xu, 1991; 1993)
adaptive robust PCA & PSA rule (Xu&Yuille, 1992&95) other robust versions in (Tab.2, Xu, 1994b)
(a) maximum likelihood (ML) factor analysis (FA) by EM algorithm (Rubi &Thayer, 1976) (b) Revisited under name of SPCA or name of PPCA for a special case (Tipping & Bishop, 1999; Roweis 1998)
Perform PCA plus additional asymmetrical or recurrent wiring (Sanger, 1989) & others
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
(a) pattern recognition with both PCs and MCs (b) adaptive rule for 1st-MC (Xu, Krzyzak & Oja, 1991), 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 PSA (Tanaka,2005)
(a) LMSER cost $E_2(W)=\frac{1}{N}\sum_{t=1}^{N}\|x_t-\hat{x}_t\|^2$ $\hat{x}_t=W^ty=W^tWx_t$ (b) first global convergence proof on Oja PSA rule (b) adaptive LMSER for PSA (Xu, 1991)
Other theories for PSA & k-PCA: (a) mini-distorted reflection (b) max- relative uncertainty theory (RUT) (c) max- variation (d) others (Xu, 1994b)
(a) adaptive EM algorithm with automatic selection on factors. (Xu, 1998a&b). (b) adaptive BYY learning algorithm with automatic selection on factors, see eqn(79) in (Xu, 2001a), eqn(21) &(22) in (Xu 2001b)
(a) ODE analysis by examining the existing PCA and modifications for MCA by sign switching (e.g., Chen & Amari, 2001). (b) modifying costs for PCA into costs for MCA (Xu, 1994a;2003a)
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 k-MCs (Xu, 1994) (c) Robust MCA (Oja & Wang, 1996; Wang & Karhunen, 1996)
Mathematical & experimental comparisons show that LMSER improves Oja PSA (Chatterjee, et al, 1998; Chatterjee, 2005; Lu, Yi & Tan, 2006, etc)
For signal subspace tracking (a) recursive learning (Yang, 1993&95) (b) conjugate gradient (Fu&Dowling, 1995) (c) Gauss–Newton (Mathew, Reddy & Dasgupta, 1995)
(a) algebraic and geometric properties on one of them called RUT (Fiori 2001&04) (b) as a special case of RUT, NIC criterion for subspace tracking (Miao & Hua,1998).
(a) Temporal FA & apative EM algorithm (Sec. IV(C) in Xu, 2000, submitted in July 1997) (b) adaptive BYY learning algorithm with automatic selection on factor number (eqn.79, Xu 2001a) (c) also a criterion for selecting the number of factors (Xu, 2001a&b, 2003a, 2004, 2007b)

links among studies and on topics missing in the existing surveys and textbooks, to the author's best knowledge.

As shown on Roadmap A, this stream originated from over 100 years ago. The first adaptive learning one is Oja 1st-PC rule [66] that finds the first principal component (PC) without explicitly estimating the sample covariance Σ . Extended to find multi-PCs, one way is featured by an asymmetrical or a sequential implementation of one 1st-PC rule, but suffering error-accumulation. Readers are referred to [5,6,67,76,96] for overview. The other way is updating the weights of W symmetrically, e.g., Oja subspace rule [65]. Further studies are made in the following branches:

MCA, dual subspace, and TLS fitting Advocating to use not only a multi-PCs based subspace, but also its complementary part, i.e., minor components (MCs) that correspond the smallest eigenvalues of Σ, Xu, Krzyzak & Oja in 1991 suggests a dual pattern recognition with the first adaptive 1st-MC learning rule proposed via eqn.(11a) in [119]. Also, Minor component analysis (MCA) was firstly named by Xu, Oja & Suen in [116] and used for a total least square (TLS) curve fitting implemented by the above eqn.(11a) that finds the 1st-MC. Not only further progresses have been made on finding multi-MCs [62,63], but also this topic has been brought to the signal processing literature by Gao, Ahmad & Swamy [32] that was motivated by a visit of Gao to Xu's office where Xu introduced him the result of [116]. Thereafter, adaptive MCA learning for TLS filtering becomes a popular topic in the signal processing literature, e.g., see [24,30,58,60]. Also, efforts are made on performing either of PCA and MCA by simply switching the updating sign with a normalization as originally suggested in [119]. Since a PCA learning that converges correctly may become unstable or diverging after sign switching, studies have been made to examine the existing PCA rules on whether they remain stable after sign switching, while attempting to avoid its division computing for normalization The jobs are quite tedious and need heavy mathematical analyses of ODE stability (e.g., Chen & Amari, [16]). The other line is turning an optimization of a PCA cost into a stable optimization of an induced cost for MCA. One example that turns the LMSER cost into one for a subspace spanned by MCs is given in [111]. Generally, for a cost $\min_W J(W)$ for PCA by its gradient descending $\Delta W = -\eta\nabla J(W)$, after switching sign we have the updating $\Delta W = \eta\nabla J(W)$ in one of the following three situations:

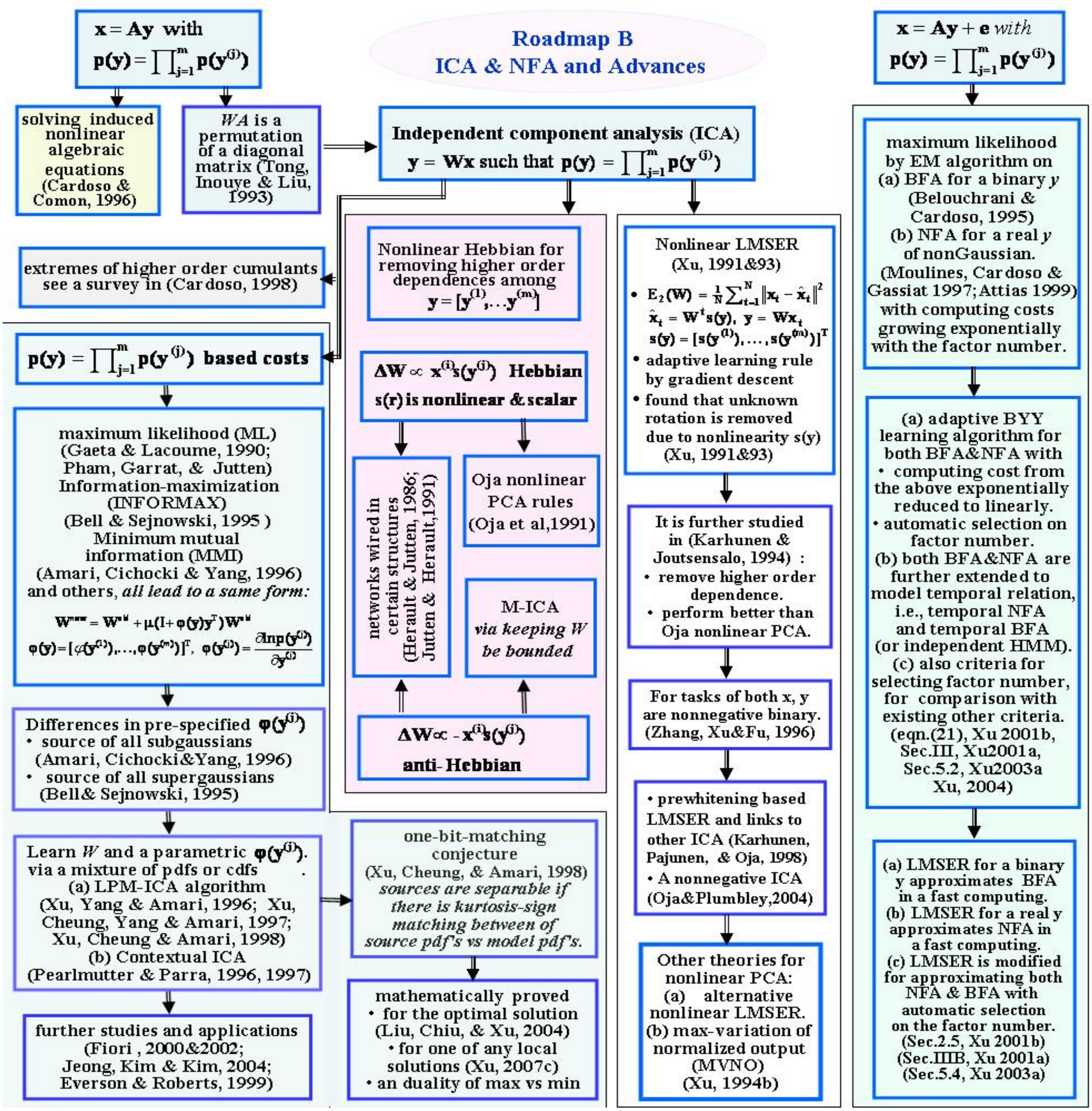
- Becoming divergence if $J(W)$ has no upper bound;
- Get a wrong solution if $J(W)$ is upper bounded but a maximum is reached at an non-orthogonal W ;
- Get the MCA solution if $J(W)$ is upper bounded and a maximum is reached at the MCA solution.

LMSER learning and Subspace tracking In Xu (1991)[118], a new adaptive PCA rule is derived from the gradient $\nabla E_2(W)$ for a least mean square error reconstruction (LMSER). In [118], the first proof on global convergence of Oja subspace rule was provided, which was previously regarded as difficult. Further comparative studies were made on Oja rule and the LMSER rule, e.g., in [14,15,47,48,54,71,72], and shown both mathematically and experimentally that LMSER improves Oja rule in both performance and converging speed. Two years after [118], Yang (1993) uses this $E_2(W)$ via a recursive least square method for signal subspace tracking [120], then followed by others in the signal processing literature [30,55]. Alternatively, Xu in 1994 also pointed out that PCA and subspace analysis can also be performed by several other theories or cost functions [111,112]. Recently in [25,28], Fiori analyzed the algebraic and geometric properties of one among them, called relative uncertainty theory (RUT). Moreover, the NIC criterion for subspace tracking [58] is actually a special case of RUT, which can be observed by comparing eqn.(20) in [58] with equation of ρ_e at the end of Sec.III.B in [111].

Principal subspace vs multi-PCs Oja subspace rule reaches a principal subspace but not truly the multi-PCs due to a unknown rotation, while it is experimentally demonstrated by Xu in 1991 that the converged rows of W approximate the multi-PCs well by adding a sigmoid function $s(r)$ [118]. Worked at Harvard by the late summer 1991, Xu got aware of Brockett (1991)[11] and 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 from modifying the LMSER rule and Oja subspace rule accordingly. The two rules were included as eqns (13)&(14) in Xu [115] that was submitted in 1991, independently and differently from that of Oja [63]. In [83], Tanaka unifies these rules into an expression controlled by one parameter and makes a comparative study on them as well as the rules in [16], with eqn(14) in [115] shown to be the most promising one. In addition, the multi-PCs were also shown to be adaptively learned by several other theories or costs (Xu, 1994b) [112].

Adaptive robust PCA In the robust statistics literature, robust PCA was proposed to resist outliers via a robust estimator on Σ [78,22]. First in 1992 [117] and then given in [117], Xu & Yuille generalized the rules of Oja, LMSER, and MCA into the robust adaptive learning by statistical physics, related to the M-estimators [40]. Also, PCA costs in [111] are extended to robust versions in Tab.2 of [112]. Thereafter, efforts have been further made, including its use in computer vision, e.g., [9,21,45,52].

On Roadmap A, another branch consists of advances on FA that is equivalent to PCA in the special case $\Sigma_e = \sigma_e^2 I$ (shown firstly [3] in 1956, revisted in [85,102]). In the past decade, there is a renewed interest on FA, not only the EM algorithm for FA [74] in is brought to PCA [75], but also adaptive EM algorithm and other advances are developed in help of the BYY harmony learning [96, 99,100,102,104], as listed on Roadmap A.



$\mathbf{x} = \mathbf{A}\mathbf{y}$ with $p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$
Roadmap B
ICA & NFA and Advances
$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}$ with $p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$
solving induced nonlinear algebraic equations (Cardoso & Comon, 1996)
WA is a permutation of a diagonal matrix (Tong, Inouye & Liu, 1993)
Independent component analysis (ICA) $\mathbf{y} = \mathbf{W}\mathbf{x}$ such that $p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$
extremes of higher order cumulants see a survey in (Cardoso, 1998)
$p(\mathbf{y}) = \prod_{j=1}^{m} p(y^{(j)})$ based costs
maximum likelihood (ML) (Gaeta & Lacoume, 1990; Pham, Garrat, & Jutten) Information-maximization (INFORMAX) (Bell & Sejnowski, 1995) Minimum mutual information (MMI) (Amari, Cichocki & Yang, 1996) and others, all lead to a same form:
$W^{new} = W^{old} + \mu(I + \varphi(y)y^T)W^{old}$
$\varphi(y) = [\varphi(y^{(1)}), \dots, \varphi(y^{(m)})]^T,\ \varphi(y^{(j)}) = \frac{\partial \ln p(y^{(j)})}{\partial y^{(j)}}$
Differences in pre-specified $\varphi(y^{(j)})$
• source of all subgaussians (Amari, Cichocki&Yang, 1996)
• source of all supergaussians (Bell& Sejnowski, 1995)
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) (b) Contextual ICA (Pearlmutter & Parra, 1996, 1997)
further studies and applications (Fiori , 2000&2002; Jeong, Kim & Kim, 2004; Everson & Roberts, 1999)
Nonlinear Hebbian for removing higher order dependences among $\mathbf{y} = [y^{(1)}, \dots y^{(m)}]$
$\Delta W \propto x^{(i)} s(y^{(j)})$ Hebbian
s(r) is nonlinear & scalar
networks wired in certain structures (Herault & Jutten, 1986; Jutten & Herault,1991)
Oja nonlinear PCA rules (Oja et al,1991)
M-ICA via keeping W be bounded
$\Delta W \propto -x^{(i)} s(y^{(j)})$
anti- Hebbian
one-bit-matching conjecture (Xu, Cheung, & Amari, 1998) sources are separable if there is kurtosis-sign matching between of source pdf's vs model pdf's.
mathematically proved
• for the optimal solution (Liu, Chiu, & Xu, 2004)
• for one of any local solutions (Xu, 2007c)
• an duality of max vs min
Nonlinear LMSER (Xu, 1991&93)
• $E_2(W) = \frac{1}{N}\sum_{t=1}^{N} \|x_t - \hat{x}_t\|^2$
$\hat{x}_t = W^t s(y),\ y = Wx_t$
$s(y) = [s(y^{(1)}), \dots, s(y^{(m)})]^T$
• 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) :
• remove higher order dependence.
• perform better than Oja nonlinear PCA.
For tasks of both x, y are nonnegative binary. (Zhang, Xu&Fu, 1996)
• prewhitening based LMSER and links to other ICA (Karhunen, Pajunen, & Oja, 1998)
• A nonnegative ICA (Oja&Plumbley,2004)
Other theories for nonlinear PCA: (a) alternative nonlinear LMSER. (b) max-variation of normalized output (MVNO) (Xu, 1994b)
maximum likelihood by EM algorithm on (a) BFA for a binary y (Belouchrani & Cardoso, 1995) (b) NFA for a real y of nonGaussian. (Moulines, Cardoso & Gassiat 1997; Attias 1999) with computing costs growing exponentially with the factor number.
(a) adaptive BYY learning algorithm for both BFA&NFA with
• computing cost from the above exponentially reduced to linearly.
• automatic selection on factor number.
(b) both BFA&NFA are further extended to model temporal relation, i.e., temporal NFA and temporal BFA (or independent HMM).
(c) also criteria for selecting factor number, for comparison with existing other criteria. (eqn.(21), Xu 2001b, Sec.III, Xu2001a, Sec.5.2, Xu2003a Xu, 2004)
(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 both NFA & BFA with automatic selection on the factor number. (Sec.2.5, Xu 2001b) (Sec.IIIB, Xu 2001a) (Sec.5.4, Xu 2003a)

3. Independent subspaces of higher order independence

In Fig.2 we further extend $\mathbf{p(y_t^{(j)}|\mu^{(j)})}$ to nonGaussian, with more constraint than a diagonal $Eyy^T = \Lambda$ imposed by $\prod_{j=1}^{m} \mathbf{p(y_t^{(j)} | \mu_t^{(j)})}$. We start at the degenerated case ***e=0***, shown at the left-upper corner on Roadmap B. The problem is solving ***x=Ay*** from samples of ***x*** and a constraint on higher order independence among components of ***y.*** One way to handle is solving induced nonlinear algebraic equations, with details referred to a survey paper [12]. From ***x=Ay**** with at most one component of ***y**** being Gaussian, it has been showed in [87] that "***y=Wx*** becomes component-wise independent" means "***y=Wx*** recovers ***y****up to scales and a permutation of components". Thus, the problem can be turned into independent component analysis (ICA), which was further tackled in four branches, as shown on Roadmap B.

One is seeking extremes of the higher order cumulants of *y* [13]. The other branch is featured by nonlinear extension of Hebbian learning. One example combines Hebbian and anti-Hebbian rules in a network with nonlinear neurons [37,44]. Another example is Oja nonlinear PCA rules by adding nonlinearity into Hebbian learning [64].

The third branch optimizes a cost or principle that directly aims at the independence in a product of component densities. Interestingly, different aspects lead to a same updating equation shown on Roadmap B, with difference coming from pre-specifying the nonlinearity of $\phi(y^{(j)})$. As a result, one works well when the source components of ***y**** are all subgaussians [2] while the other works well when the source components of ***y**** are all supergaussians [7]. The problem is solved by learning jointly *W* and $\phi(y^{(j)})$ via a parametric model [68, 106, 107, 108]. Also, it has been found that a rough estimate of each source pdf is enough for source separation, which leads to one-bit-matching conjecture that is mathematically proved to be true first in a weak sense that the global optimization is reached [51] and then in a strong sense that anyone of local optimal solutions is reached [94], also applicable to partial separation by a partial matching and a duality of maximization and minimization.

The last branch for ICA is featured by Nonlinear LMSER (Xu, 1991&93)[116,118], with details on Roadmap B. Additionally, clarifications are here made on two confusions. One relates to an omission of the origin of Nonlinear LMSER in [48], which has already been clarified in [41,47,71,72] that clearly spell out that both the nonlinear $E_2(W)$ and its adaptive gradient rule were firstly proposed in (Xu, 1991)[118]. The second comes from that ICA is usually regarded as a counterpart of PCA [44]. As already stated in [96,100], it is inappropriate and yields confusion. It follows from the first row of Tab.1 that ICA by ***y=Wx*** is an extension of decorrelation analysis, including any combinations of PCs and MCs, while it follows from the second row of Tab.1 that the counterpart of MCA is minor ICA (M-ICA) that is performed via nonlinear anti-Hebbian, minimizing cumulants, minimizing MVNO (See Roadmap B); while the counterpart of PCA is principal ICA (P-ICA) that is reached by nonlinear Hebbian, maximizing cumulants, and nonlinear LMSER.

Actually, the concept `principal' involves $e_t = x_t - Ay \neq 0$, i.e., eq.(1) or the model at the right-upper corner on Roadmap B. It extends FA to the cases beyond that *y* is from a Gaussian, e.g., binary FA (BFA) if *y* is binary, and nonGaussian FA (NFA) if *y* is real but not from a Gaussian, as shown by the third row in Tab.1. Details are referred to the leftmost column on Roadmap B. Particularly, similar to that FA becomes equivalent to PCA in the special case $\Sigma_e = \sigma_e^2 I$, both BFA and NFA perform P-ICA at this special case too. As discussed in [96], by using a nonlinear map $y_t^{(j)} = s(z_t^{(j)})$ and ***z=Wx*** to avoid expensive computing cost for *y*, the nonlinear LMSER implements either an approximate BFA with a Bernoulli $p(y^{(j)})$ in probability $p_j = \frac{1}{N}\sum_{t=1}^{N} s(z_t^{(j)})$ or a real factor NFA with $p(y^{(j)})$ being a pseudo uniform distribution on $[0,+\infty)$ or $(-\infty,+\infty)$ in help of BYY harmony learning [96], from which further results have been obtained too. Moreover, when $p(y^{(j)})$ is a Bernoulli or on $[0,+\infty)$, nonlinear LMSER relates to nonnegative ICA [71,72].

In the cases with $e_t = x_t - Ay \neq 0$, another critical point is that it is impossible for a linear ***y=Wx*** to lead to the minimum error $E\|e\|_B^2$ or the maximum likelihood. Instead, a nonlinear map ***y=f(x)*** is needed, which incurs open theoretical issues that will be addressed in the last section. From the aspect of algorithms, seeking an appropriate ***y=f(x)*** incurs expensive computing costs too. Further structures should be imposed on either or both of ***f(x)*** and $\mathbf{p(y_t^{(j)}|\mu^{(j)})}$. Nonlinear LMSER provides an approximate implementation with a post-sigmoid structure for ***f(x)***, which saves computing costs considerably and is experimentally shown to work well [48,71,72]. However, no quantitative analysis has been made on how this approximation affects performance yet. The EM algorithm in [4,61] implements the maximum likelihood learning by considering $\mathbf{p(y_t^{(j)}|\mu^{(j)})}$ via a Gaussian mixture, but suffering a computing cost that increases exponentially with the dimension *m* of $\boldsymbol{y}$. Also, the BYY harmony learning [99,100] considers $\mathbf{p(y_t^{(j)}|\mu^{(j)})}$ in a Gaussian mixture or other parametric models, and a preliminary experimental study showed that it gets a performance similar to the above EM algorithm based maximum likelihood learning but with a computing cost linearly increasing with *m* [95]. Further comparative studies on the three approaches are needed both theoretically and experimentally.

4. Extensions to temporal and localized structures

We further move to observe the first row of Tab.2 that extends the third row in Tab.1 after considering temporal structures. A typical one is embedded in $\mathbf{p(y_t^{(j)} \mid \mu_t^{(j)})}$ with $\mathbf{\mu_t^{(j)} = \mu^{(j)}(Y_t^{(j)}, \varphi_j)}$, $\mathbf{Y_t^{(j)} = \{y_{t-\tau}^{(j)}\}_{\tau=1}^{q^{(j)}}}$, e.g., a linear regression $\mathbf{\mu_t^{(j)}} = \sum_{\tau=1}^{q^{(j)}} \beta_\tau^{(j)} \mathbf{y_{t-\tau}^{(j)}}$. Information can be carried over time in two ways. One is via computing $\mathbf{\mu_t^{(j)}}$ by the regression, with learning on $\varphi_\mathbf{j}$ made in help of the gradient with respect to $\varphi_\mathbf{j}$ by the chain rule through the regression. The second way is via computing the integral $\int \mathbf{p(y_t^{(j)} \mid \mu_t^{(j)}) p(Y_t^{(j)}) dY_t^{(j)}}$ and getting the gradient with respect to $\varphi_\mathbf{j}$ by the chain rule through the integral. Details are referred to [95,96,99].

$p(y_t^{(j)} \mid \mu_t^{(j)})$ $d=e^TB^TBe$ $x=Ay+e$	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 ${\mu_t^{(j)}}^{y_t^{(j)}}(1-\mu_t^{(j)})^{1-y_t^{(j)}}$ $0 \le \mu_t^{(j)} \le 1$
From i.i.d. samples to temporal samples $\mu_t^{(j)} = \mu^{(j)}(y_t^{(j)}, \varphi_j)$ $y_t^{(j)} = \{y_{t-\tau}^{(j)}\}_{\tau=1}^{q^{(j)}}$	Temporal factor analysis (TFA)	Temporal nonGaussian Factor analysis (T-NFA) Temporal LMSER (T-LMSER) $y^{(j)} = f(z^{(j)}),\ z = Wx$	Temporal binary factor analysis (T-BFA) or independent HMM binary T-LMSER
From one subspace to multiple subspaces	Local FA (LFA) Local PCA/PSA	Local NFA LOCAL LMSER	Local BFA Local LMSER
Both from i.i.d. to temporal from one to multiple	Local TFA	Local T-NFA Local T-LMSER	Local T-BFA Local independent HMM Local T-LMSER

Tab.2 Typical examples with temporal and localized structures

Shown in the second row of Tab.2 are extensions of third row in Tab.1 to multiple subspaces at different locations and thus under the name of local sth. Studies on this stream are summarized on Roadmap C, where a key point is how to allocate a sample x_t to different coordinate systems. There are two typical ways. One is made during implementing the EM algorithm for a maximum likelihood (ML) learning or a Bayesian learning. The other is made via competition, including the classic competitive learning, and the rival penalized competitive learning (RPCL) and the one in the Bayesian Ying Yang (BYY) harmony learning [92,93].

Roadmap C Several Types of Mixtures of Subspaces

models / learning		*p(y)* is Gaussian		*p(y)* is nonGaussian	
		PCA/FA/TFA	MCA/Surface fitting	ICA	NFA/BFA/LMSER and temporal extensions
Competitive Learning	Classic	Local PCA (Sec.3.2, Xu 1995) (Kambhatla & Leen, 1997) PCA competitive learning (Lopez-Rubio, et al 2004)	Local MCA by MML (Sec.4.1, Xu 1995)	Competitive ICA (eqn.37, Xu 2001b) Competitive Temporal ICA (eqn. 88, Xu 2001a)	the degenerated cases from the ones below
	RPCL & BYY	Local PCA, Local FA, Local LMSER (eqns.14&16, Xu 1998b) (Sec.4.3, Xu 2001b) Local FA(Sec.3.3.2, Xu, 2007a) Local TFA (Xu, 2004)	Local MCA by MML (eqn.15, Xu, 1998b) (Sec.4.2, Xu, 2001b)	Improved competitive ICA (Sec.4, Xu 2002)	Local BFA, NFA, LMSER (Sec.4, Xu 2002) (eqns.43&44, Xu 2001b) temporal extensions (Xu, 2004)
		with automatic selection on the number of subspaces or both of the number and dimensions of subspaces.			
EM Algorithm	ML	Local PCA by a simplified EM (Sec.V(B)(D), Xu 1994c) Mixtures of FA (Ghahramani & Hinton, 1996) Mixtures of probabilistic PCA (Tipping & Bishop, 1999)	Local MCA by a simplified EM (Sec.V(C)(D), Xu 1994c) MCA Co-integration (Xu & Leung, 1998) Probabilistic MCA (Williams & Agakov, 2002)	ICA mixture (Lee, Lewicki, & Sejnowski, 2000)	One possible way is getting extension from (Moulines, Cardoso & Gassiat 1997; Attias 1999) but with much expensive computing costs
	Bayesian	Variational Mixture (Ghahramani & Beal, 2000; Utsugi & Kumagai, 2001)		Variational Mixture (Choudrey & Roberts 2003)	

5. Subspace number and subspace dimension

Another important problem is how to determine the number *k* of subspaces and the dimension m_ℓ of each subspace. It is usually referred under the name model selection. A classic way is implementing a two-phase procedure. First, a number of candidates are learned by enumerating *k* and m_ℓ. Second, a criterion is used to select the best among the candidates. An example is using AIC for one coordinate system based FA [10]. Other choices of criteria include BIC or equivalently MDL, the cross validation, etc. However, such a two-phase implementation is computationally very extensive and thus impractical. Moreover, when the sample size is finite and k, m_ℓ are not too small, the values of criterion are difficult to be estimated accurately, which makes the performance degenerate considerably.

Alternative solutions have been sought. One type includes incremental approaches, e.g., as m_ℓ increases to $m_\ell+1$, learning is made incrementally with the parts already learned kept or partially adjusted such that redundant computing can be saved. However, it usually leads to a suboptimal performance because not only newly added parameters have to be learned, but also the old parameter set have to be relearned. Oppositely, we can decrease from m_ℓ to $m_\ell-1$. Still we can not simply discard those extra parameters, i.e., all the parameters in the case $m_\ell-1$ have to be re-learned.

Another direction is seeking model selection made automatically during learning. That is, in a model with k and m_ℓ initially set to be large enough such that the correct one is included, learning will not only determine parameters but also automatically shrink k and m_ℓ down to appropriate ones. One such an effort is RPCL [114], which can make k determined automatically during learning. The other and better choice is the BYY harmony learning that is a general approach applicable to various statistical learning tasks with automatic model selection. Readers are referred to [92,93] for an adaptive algorithm to implement local FA with automatic determination of k and m_ℓ, and for a systematic comparative experimental study on a large number of simulated data sets and several real data sets from UCI repository, in comparison with AIC, CAIC, BIC/MDL, CV as well as two incremental approaches [34,80]. Both performances and computing times are compared, and it can be observed that the BYY harmony learning algorithm outperforms the counterparts considerably. Readers are referred to [92,93,95,97,98, 99,100] for the BYY harmony learning based criteria and algorithms for the rest cases in Tabs 1 &2.

4. Concluding remarks

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

There are already extensive studies on the cases with independence in a sense of second order statistics. Also, there are extensive studies on ICA with noise free (i.e, ***e=0***). Trends move towards the cases with components of ***y*** being mutually independent in higher order statistics and with noise $e\neq0$. Though a few algorithms are available, further comparative studies on them are needed both theoretically and experimentally. Even more importantly, there are still a number of open issues, some of which are listed below:

- Which part of unknown parameters in $x=Ay+e$ can be determined uniquely ? which part is indeterminable and how it can be improved ?
- Under what conditions, the independence of $\prod_{j=1}^{m} \mathbf{p}(\mathbf{y}_t^{(j)} \mid \boldsymbol{\mu}_t^{(j)})$ can be ensured conceptually ? How can it be further achieved by a learning algorithm ?
- When and under what condition, $\hat{x}=Ay$ can be said to be the best reconstruction of x ?
- Under what condition, both ensuring $\prod_{j=1}^{m} \mathbf{p}(\mathbf{y}_t^{(j)} \mid \boldsymbol{\mu}_t^{(j)})$ and the best reconstruction of x by $\hat{x}=Ay$ can be achieved jointly ? If not, how to trade off the two ?
- What is the best nonlinear map ***y=f(x)*** in term of $\mathbf{p}(\mathbf{y}_t^{(j)} \mid \boldsymbol{\mu}^{(j)})$ and ***e*** ? Can the best be obtained analytically or via an effective computing ?

Acknowledgements

The work was fully supported by a grant from the Research Grant Council of HKSAR (project No: CUHK 4177/07E).

References

[1] Amari S Biol. Cybernetics*,* 26, 175–185, 1977.
[2] Amari S, Cichocki A & Yang H, Advances in NIPS 8, MIT Press, 757-763, 1996.
[3] Anderson TW & Rubin H, Proc. Berkeley Symp. Math. Statist. Prob. 3rd, UC Berkeley, 111-150, 1956.
[4] Attias H, Neural Computation, 11, 803-851, 1999.
[5] Baldi P & Hornik K, In: Backpropagation: Theory, Architectures and Applications, Chauvin & Rumelhart, Eds., Hillsdale, NJ: Erlbaum Associates, 1991.
[6] Baldi P & Hornik K, Neural Networks 6(4), 837-858, 1995.
[7] Bell A & Sejnowski T, Neural Computation 17, 1129-1159, 1995.
[8] Belouchrani A & Cardoso JF, Proc. NOLTA95, 49-53, 1995.
[9] Black MJ & Jepson AD, Proc. ECCV96, 329–342, 1996.
[10] Bozdogan H & Ramirez DE, Psychometrika 53(3), 407-415, 1988.

[11] Brockett RW, Linear Algebra and Its Applications 146, 79-91, 1991.

[12] Cardoso JF & Comon P, Proc. IEEE ISCAS96, Vol2, 93-96, 1996.

[13] Cardoso JF, Proc. of IEEE. 86(10), 2009-25, 1998.

[14] Chatterjee C, Roychowdhury VP & Chong EKP, IEEE Trans. Neural Networks 9, 319–329, 1998.

[15] Chatterjee C, Neural Networks 18, 145–159, 2005.

[16] Chen T & Amari S, Neural Networks, 14(10), 1377–1387, 2001.

[17] Choudrey RA & Roberts SJ, Neural Computation 15(1), 213-252, 2003.

[18] Cirrincione G, Cirrincione M, Hérault J & Huffel SV, IEEE Trans. Neural Networks 13, 160-187, 2002.

[19] Comon P, Signal Processing 36, 287-314, 1994.

[20] Dayan P & Zemel RS, Neural Computation 7, 565-579, 1999.

[21] Dela TF & Black MJ, International Journal of Computer Vision 54(1-3), 117–142, 2003.

[22] Devlin SJ, Gnanadesikan R & Ketternrig JR, J.Am. Stat. Assoc. 76, 354-362, 1981.

[23] Everson R & Roberts S, Neural Computation 11, pp1957-1983, 1999,

[24] Feng DZ, Bao Z & Jiao LC, IEEE Trans. Signal Processing 46, 2122–2130, 1998.

[25] Fiori S, Intl. J Neural Systems 14(5), 293-311, 2004.

[26] Fiori S, Neural Networks 16, 453–467, 2003.

[27] Fiori S, Neural Networks, 15(1), 85–94, 2002.

[28] Fiori S, Neural Computation 13, 1625–1647, 2001.

[29] Fiori S, Neural Networks, 13(6), 597-611, 2000.

[30] Fu Z & Dowling EM, IEEE Trans. Signal Processing, 43, 1151–1160, 1995.

[31] Gaeta M & Lacoume JL, Proc. EUSIPO 90, 621-624, 1990.

[32] Gao K, Ahmad MO & Swamy MN, Electron. Lett., 28(4), pp430–432, 1992.

[33] Gao K, Ahmad MO & Swamy MN, IEEE Trans. Circuits Syst. Part II 41, 718–729, 1994.

[34] Ghahramani Z & Beal M, Advances in NIPS 12, Cambridge, MA: MIT Press, pp. 449–455, 2000.

[35] Ghahramani Z & Hinton GE,Neural Computation 12, 831–864, 2000.

[36] Hebb, O, The Organization of Behavior, Wiely, 1949.

[37] Herault J & Jutten C, In J. S. Denker (Ed.), Neural networks for computing: Proceedings of AIP Conf, American Institute of Physics, 206-211, 1986.

[38] Hinton GE & Salakhutdinov RR, Science 313 (5786), 504 – 507, 2006.

[39] Hotelling H, Psychometrika 1, 27-35, 1936.

[40] Huber PJ, Robust Statistics, John Wiley, NY, 1981.

[41] Hyvarinen A, Karhunen J & Oja A, Independent component analysis, John Wiley., New York, 2001.

[42] Jeong JW, Kim, TS & Kim SH, Intl J of Imaging System and technology 14, 170–180, 2004.

[43] Jutten C & Herault J, Signal Processing 24, 1-20, 1991.

[44] Jutten C & Herault J, Proc. EUSIPCO88, 643-646, 1988.

[45] Kamiya H & Eguchi S, Journal of Multivariate Analysis 77, 239-269, 2001 .

[46] Kambhatla N & Leen TK, Neural Computation 9(7), 1493–1516, 1997.

[47] Karhunen J, Pajunen P & Oja E, Neurocomputing 22, 5-20, 1998.

[48] Karhunen J & Joutsensalo J, Neural Networks 7(1), 113-127, 1994.

[49] Lee TW, Girolami M & Sejnowski TJ, Neural Computation 11, 417-441, 1999.

[50] Lee T, Lewicki M, & Sejnowski T, IEEE Trans. on Pattern Recognition and Machine Intelligence, 22(10), 1–12, 2000.

[51] Liu ZY, Chiu KC & Xu L, Neural Computation 16, 383-399, 2004.

[52] Li YM, Pattern Recognition 37, 1509 – 1518, 2004.

[53] Lopez-Rubio E, et al, Neural Computation 16(11), 2459 – 2481, 2004.

[54] Lu JC, Yi Z & Tan KK, Neurocomputing 70, 362–372, 2006.

[55] Mathew G, Reddy VU & Dasgupta S, IEEE Trans. Signal Processing, 43, 401–411, 1995.

[56] Matsuoka K, Ohya M & Kawamoto M, Neural Networks 8(3), 411-419, 1995.

[57] McLachlan GJ & Krishnan T The EM Algorithm and Extensions, John Wiley, 1997.

[58] Miao YF & Hua YB, IEEE Trans. Signal Processing, 46, pp1967–79, 1998.

[59] Molgedey L & Schuster H, Physical Review Letters, 72(23), 3634-37, 1994.

[60] Möller R, Intl J Neural Systems, 14(1), 1-8, 2004.

[61] Moulines E, Cardoso J & Gassiat E, Proc. ICASSP97, 3617-20, 1997.

[62] Oja E & Wang LY, Neural Networks 9, 435-444, 1996.

[63] Oja E, Neural Networks 5, 927-935, 1992.

[64] Oja E, Ogawa H & Wangviwattana J, Proc. ICANN'91, 385-390, 1991.

[65] Oja E, Int. J. Neural Systems 1, 61–68, 1989.

[66] Oja E, J. Math. Biol., 15(3), 267-273, 1982.

[67] Palmieri F, Zhu J, & Chang C, IEEE Trans. Neural Networks 4, 748–761, 1993.

[68] Pearlmutter BA & Parra LC, In M. Mozer, M. Jordan & T. Petsche (Eds.), Advances in NIPS 9, Cambridge, MA: MIT Press, 613–619, 1997.

[69] Pearson K, Phil. Mag. 2, 559-572, 1901.

[70] Pham DT, Garrat P, & Jutten C, Proc. EUSIPCO92, 771-774, 1992.
[71] Plumbley MD, IEEE Trans. Neural Networks 14, 534–543, May 2003.
[72] Plumbley MD & Oja E, IEEE Trans on Neural Networks 15(1), 66-76, 2004.
[73] Redner RA & Walker HF, SIAM Review 26, 195-239, 1984.
[74] Rubi D & Thayer D, Psychometrika 57, 69-76, 1976.
[75] Roweis ST, In M. Kearns, M. Jordan, and S. Solla (Eds.), Advances in NIPS 10, Cambridge, MA: MIT Press, 626–632, 1998.
[76] Roweis S & Ghahramani Z, Neural Computation 11, 305–345, 1999.
[77] Roweis ST & Saul LK, Science 290 (5500), 2323 – 2326, 2000
[78] Ruymagaart FH, Journal of Multivariate Analysis 11, 485–497, 1981.
[79] Sablatnig R & Kampel M, Computer Vision and Image Understanding 87, 90–103, 2002.
[80] Salah, A & Alpaydin, E, Proc.17th Intl Conf. on Pattern Recognition, vol.1, 276-279, 2004.
[81] Sanger TD, Neural Networks 2, 459–473, 1989.
[82] Sato M, Neural Computation 13(7), 1649-1681, 2001.
[83] Tanaka T, IEEE Trans. Signal Processing 53(4), 1243-1253, 2005.
[84] Tipping M & Bishop CM, J. of the Royal Statistical Society B61, 611–622, 1999.
[85] Tipping M & Bishop C, Neural Computation 11(2), 443–482, 1999.
[86] Thurston, L, Multiple factor analysis, Univ. of Chicago Press, Chicago, Illinois, 1945.
[87] Tong L, Inouye Y, & Liu R, IEEE Trans. on Signal Processing 41, 2461-2470, 1993.
[88] Tong L & Perreau S, Proc. of IEEE 86(10), 1951-1068, 1998.
[89] Turk M & Pentland A, Journal Cognitive Neuroscience 3(1), 71–86, 1991.
[90] Utsugi A & Kumagai T, Neural Computation 13(5), 993-1002, 2001.
[91] Williams CKI & Agakov FV, Neural Computation 14, 1169–1182, 2002.
[92] Xu L (2007a), Pattern Recognition 40, 2129-2153, 2007.
[93] Xu L (2007b), Bayesian Ying Yang Learning, In Scholarpedia, no10469, http://scholarpedia.org, 2007.
[94] Xu L (2007c), Neural Computation 19, 546-569, 2007.
[95] Xu L, IEEE Trans on Neural Networks 15(5), 1276-1295, 2004.
[96] Xu L (2003a), Neural Information Processing Letters and Reviews 1(1), 1-52, 2003.
[97] Xu L (2003b), Neural Networks 15(5-6), 817-825, 2003.
[98] Xu L, Neural Networks 15, 1125–1151, 2002.
[99] Xu L (2001a), IEEE Trans. Neural Networks 12, 822–849, 2001.
[100] Xu L (2001b), in: N. Allison et al. (Eds.), Advances in Self-Organizing Maps, Springer, 181–210, 2001.
[101] Xu L, IEEE Trans. on Signal Processing 48, 2132–2144, 2000.
[102] Xu L (1998a), Neurocomputing 22, 81-112, 1998.
[103] Xu L (1998b), in: Proceedings of IEEE-INNS IJCNN98, Anchorage, Alaska, vol II, 2525–30, 1998.
[104] Xu L (1998c), J. Computational Intelligence in Finance 6(5), 6-18, 1988.
[105] Xu L & Leung, WM, Proc. of IEEE/IAFE 1998 Intl. Conf. Computational Intelligence for Financial Engineering (CIFEr98-NYC)}, 157-160, 1998.
[106] Xu L, Cheung CC & Amari SI, Neurocomputing 22 (1-3), 69-80, 1998.
[107] Xu L, Cheung CC & Yang HH & Amari SI, Proc. IJCNN97 III, 1821-1826, 1997.
[108] Xu L, Yang HH, & Amari S, Presentation at RIKEN Frontier Forum, Japan, April, 1996.
[109] Xu L, & Yuille, AL, IEEE Trans. Neural Networks 6, 131–143, 1995.
[110] Xu L, Proc. WCNN95, Vol 1, pp35-42, 1995.
[111] Xu L (1994a), Proc. IEEE ICNN94, Vol.I, 315-320, 1994.
[112] Xu L (1994b), Proc. ICONIP94, vol.2, 943-949, 1994.
[113] Xu L (1994c), Proc. IEEE ICNN94, Vol.II, 1252-1257, 1994.
[114] Xu L Krzyzak A & Oja E, IEEE Trans Neural Networks 4, 636-649, 1993.
[115] Xu L Neural Networks 6, 627–648, 1993.
[116] Xu L, Oja E & Suen CY, Neural Networks 5, 393-407, 1992.
[117] Xu L & Yuille AL, Proc. of IJCNN92, Baltimore, MA, Vol. I, .812-817, 1992.
[118] Xu L Proc. of 1991 Intl Joint Conf on Neural Networks (IJCNN91-Singapore), vol.3, .2363-73.
[119] Xu L Krzyzak A & Oja E, Intl J. of Neural Systems 2(3), 169-184, 1991.
[120] Yang B, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. IV, 145–148, 1993.
[121] Yang B, IEEE Trans. Signal Processing, 43(1), 95–107, 1995.
[122] Yang X, Sarkar TK & Arvas E, IEEE Tr. Acoust., Speech, Signal Processing 37, 1550–1556, 1989.
[123] Zhang BL, Xu L & Fu MY, Intl J. Neural Systems 7(3), 223-236, 1996.