3. Arbitrage Pricing Theory

- Capital Asset Pricing Model vs. Arbitrage Pricing Theory
- Temporal Factor Analysis (TFA) and APT
- TFA based APT for Prediction
- TFA based APT for Portfolio Management

Capital Asset Pricing Model

•Portfolio A is preferred to portfolio B if

(i) $E_A(R) \geq E_B(R)$ and

(ii) $\mathrm{var}_A(R) \leq \mathrm{var}_B(R)$ or $SD_A(R) \leq SD_B(R)$

•Portfolios that satisfy this known as the set of *efficient portfolios.*

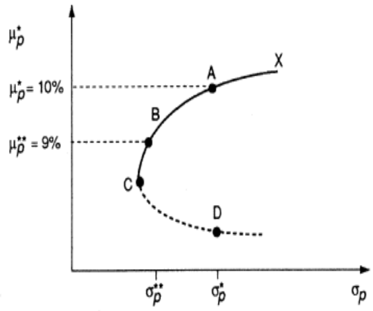
$$R_p = x_1 R_1 + x_2 R_2$$

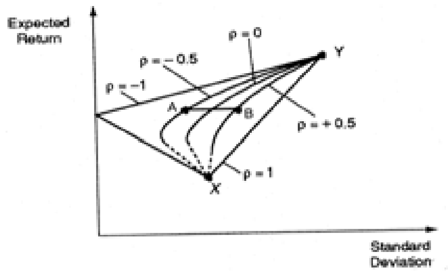
$$ER_p = \mu_p = \left(x_1 ER_1 + x_2 ER_2\right) = x_1 \mu_1 + x_2 \mu_2$$

$$\mu_p = \sum_{i=1}^{n} x_i \mu_i$$

$$\sigma_p^2 = \sum x_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} x_i x_j \sigma_{ij}$$

The *efficient frontier* shows all the combinations of (μ_p, σ_p) which *minimizes* risk σ_p for a *given level* of μ_p.





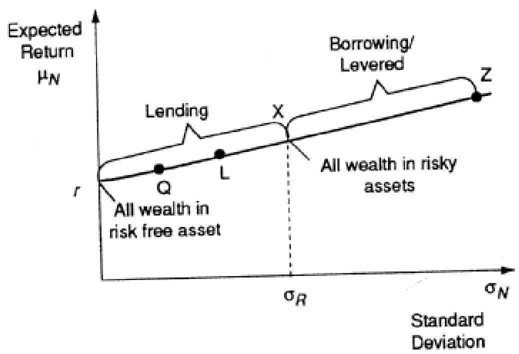
Efficient Frontier and Correlation.

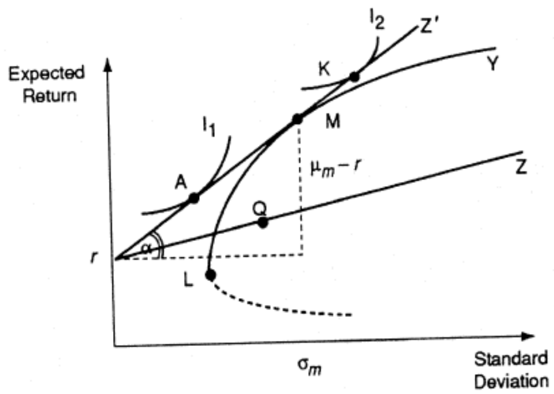
(the *set* of efficient portfolios forms the *efficient frontier*.)

Each *point* on the efficient frontier corresponds to a different set of *optimal* proportions $x_1^*, x_2^*, x_3^*, \cdots$ $\quad \sum x_i^* = 1$

The Optimal Portfolio

$$\mu_N = r + \left[\frac{\mu_R - r}{\sigma_R}\right]\sigma_N = \delta_0 + \delta_1\sigma_N$$





An investor can be anywhere along rZ', but M is always a fixed bundle of stocks (or fixed proportions of stocks) held by *all* investors.

•Hence point M is known as the *market portfolio* and rZ' is known as the *capital market line* (CML).

$$(ER_i - r) = \beta_i (ER^m - r) \Rightarrow ER_i = r + \beta_i (ER^m - r)$$

$$\beta_i = \text{cov}(R_i, R^m) / \text{var}(R^m)$$

ER^m is the expected return on the market portfolio that is the 'average' expected return from holding *all* assets in the optimal proportions x_i^*

Expected return = $\mu_i = ER_i$
Variance of returns = $\sigma_i^2 = \text{var}(R_i)$
Covariance of returns = $\sigma_{i,j} = \text{cov}(R_i, R_j)$

$$R_t = \overline{R} + By_t + e_t$$

actual return; mean return; factors; noise component

fundamental factor models

assume the B as given and estimate the y_t

macroeconomic factor models

assume the y_t as given and estimate the B

- e.g. changes in inflation, industrial production, investor confidence and interest rates

statistical models (factor analysis)

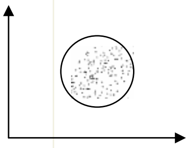
simultaneously estimate B and y_t

Rotation indeterminacy $y_t' = \Phi y$

Rotation indeterminacy

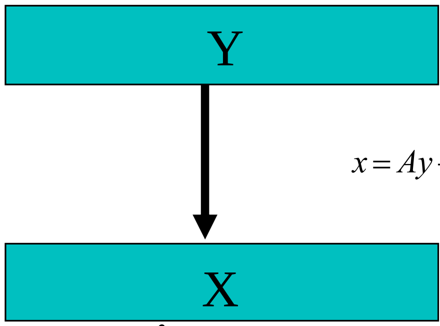
- Gaussian $q(y) = G(y|0,\Lambda) = \prod_{j=1}^{m} G(y_j|0,\lambda_j)$

Rotation Indeterminacy
factor analysis



$$\Sigma_x = A^T \Sigma_y A + \Sigma_e$$

- Nongaussian: *y* from nongaussian



$x = Ay + e$

$q(x) = \int q(x - Ay) q(y) dy$

IDENTIFYING THE FACTORS

Several researchers have investigated stock returns and estimated that there are anywhere from three to five factors. Subsequently, various people attempted to identify these factors.

By Chen, Roll, and Ross, the following factors were identified:

1. Growth rate in industrial production,
2. Rate of inflation (both expected and unexpected).
3. Spread between long-term and short-term interest rates,
4. Spread between low-grade and high-grade bonds.

Traditional Approach ONE

- Maximum Likelihood Factor Analysis
  - Likelihood Ratio (LR) test on the residuals to ascertain minimum factor number
- Limitations
  - ***k*** increases progressively with # of securities ***p*** used

    $\Rightarrow$ tends to bias towards more factors
  - Rotational indeterminacies

Traditional Approach TWO

[Chamberlain & Rothschild 1983]

- Eigenvalue Analysis Approach
  - ***k*** eigenvalues of Σ increases without bound as ***p*** increases

    $\Rightarrow$ eigenvectors can be used as factor loadings.
- Limitation
  - Assumption of infinite assets is strong and unrealistic [Shukla and Trzcinka 1990]

    $\Rightarrow$ tends to bias towards too few factors [Brown 1989]

3. Arbitrage Pricing Theory

- Capital Asset Pricing Model vs. Arbitrage Pricing Theory
- NonGaussian factor analysis (NFA), Temporal Factor Analysis (TFA), and APT
- TFA based APT for Prediction
- TFA based APT for Portfolio Management

Two Major Problems in APT Analysis

- Rotation indeterminacy (inherent in conventional maximum likelihood factor analysis)
- Determination of the appropriate number of priced factors ***k***

The problems can be solved by either of NonGaussian factor analysis (NFA) and Temporal Factor Analysis (TFA).

Non-Gaussian Factor Analysis (NFA)

$$p(y_t) = \prod_{j=1}^{k} p(y_t^{(j)})$$

Non-Gaussian

$$\begin{cases} y_t = \varepsilon_t \\ x_t = Ay_t + e_t \end{cases} \qquad t = 1, 2, \cdots, N$$

Gaussian

- Independence Constraint

Xu, L, ``BYY harmony learning, independent state space and generalized APT financial analyses ", IEEE Tr. on Neural Networks, 12 (4), 2001, 822-849.

Relationship between APT and NFA

- To analyze APT using NFA, the APT model may simply be rewritten in the following form:

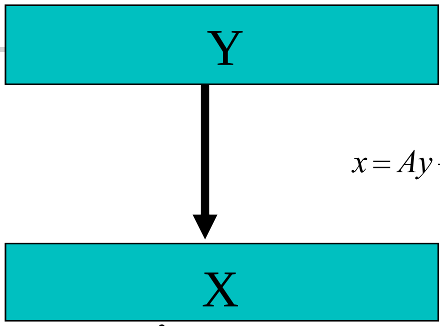
$$R_t - \overline{R} = Af_t + e_t$$

- If we let $x_t = R_t - \overline{R}$ and $y_t = f_t$, we get exactly the NFA model

$$x_t = Ay_t + e_t$$

Independent factor models

- Nongaussian: *y* from nongaussian



$$x = Ay + e$$

$$q(x) = \int q(x - Ay) q(y) dy$$

Moulines, Cardoso, & Gassiat, 1997, Attias, 1999

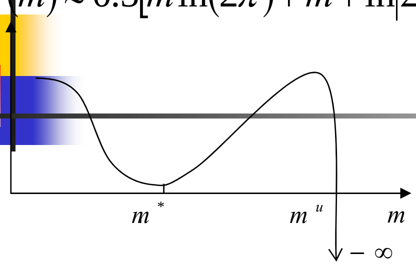
$$q\left(y^{(j)}\right) = \sum_i \beta_{ji} G(y^{(j)} \mid m_{ji}, \sigma_{ji}^2) \text{ subject to}$$
$$\int y^{(j)2} \; q\left(y^{(j)}\right) dy^{(j)} = 1, \quad \int y^{(j)} \; q\left(y^{(j)}\right) dy^{(j)} = 0$$

The EM algorithm: integral can be avoided

but with the computing complexity increasing with *m*.

NFA by Harmony Learning

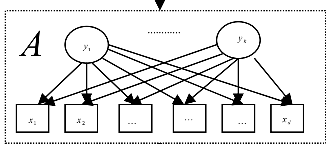
$$J(m) \approx 0.5\left[m \ln(2\pi) + m + \ln|\Sigma|\right]$$



$$q\left(y^{(j)}\right) = \sum_i \beta_{ji} G(y^{(j)} \mid m_{ji}, \sigma_{ji}^2) \text{ subject to}$$
$$\int y^{(j)2} \; q\left(y^{(j)}\right) dy^{(j)} = 1, \quad \int y^{(j)} \; q\left(y^{(j)}\right) dy^{(j)} = 0$$

$$q(y) = \prod_{j=1}^{m} q\left(y^{(j)}\right) \quad \text{is nongaussian}$$

$p(y|x)$ is free



ε

$$p_0(x) = \frac{1}{N} \sum_{t=1}^{N} \delta(x - x_t) \qquad x = Ay + \varepsilon$$
$$q(x) = \int G(x|Ay, \Sigma) p(y) dy$$

Harmony learning
$$y_t = \arg\max_y \; [q(x_t|y) q(y)]$$
from $\max_{p(y|x)} H(p\|q)$

versus

ML learning
$$p(y \mid x) = \frac{q(x \mid y) q(y)}{q(x)}$$
from $\min_{p(y|x)} KL(\theta)$

Xu, 1998, Neurocomputing, V.22, 81-112,1998

IEEE Trans. Neural Networks, Vol.12. July, 2001

NFA with automatic model selection

Automatic selection on *m*

$$q(y) = |\Lambda|^{-0.5}\, q_0(V\Lambda^{-0.5} y),\quad q_0(y) = \prod q(y_j)$$

$q(y_j)$ is a mixture of Gaussians or from a mixture of sigmoid, *Subject to*

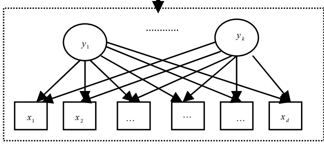
$y=[y_1 \ldots y_m]$

$$y_t = \arg\max_y \,[q(x_t|y)q(y)]$$

$$\Uparrow$$

$$\max H(p\|q) = \int p(y|x)p(x)\ln[q(x|y)q(y)]dxdy$$

$p(y|x)$ is free



$x=[x_1 \ldots x_d]$ $\quad x = Uy + \varepsilon$ $\quad +$ $\quad \varepsilon$

$$\int y_j^2\ q_0(y_j^2)dy_j = 1$$

$$\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_m]$$

$$U^T U = I, V^T V = I$$

$$\mathrm{A} = \mathrm{U\Lambda V^T}$$

$\delta U,\ \delta V$ are updated in the Stiefel manifold *via* updating $\delta\lambda_j$

Xu, L (2004a), in press, IEEE Trans on Neural Networks

Xu L, Neural Information Processing - Letters and Reviews, Vol.1, No.1, pp1-52, 2003.

Benefits of NFA for APT Analysis

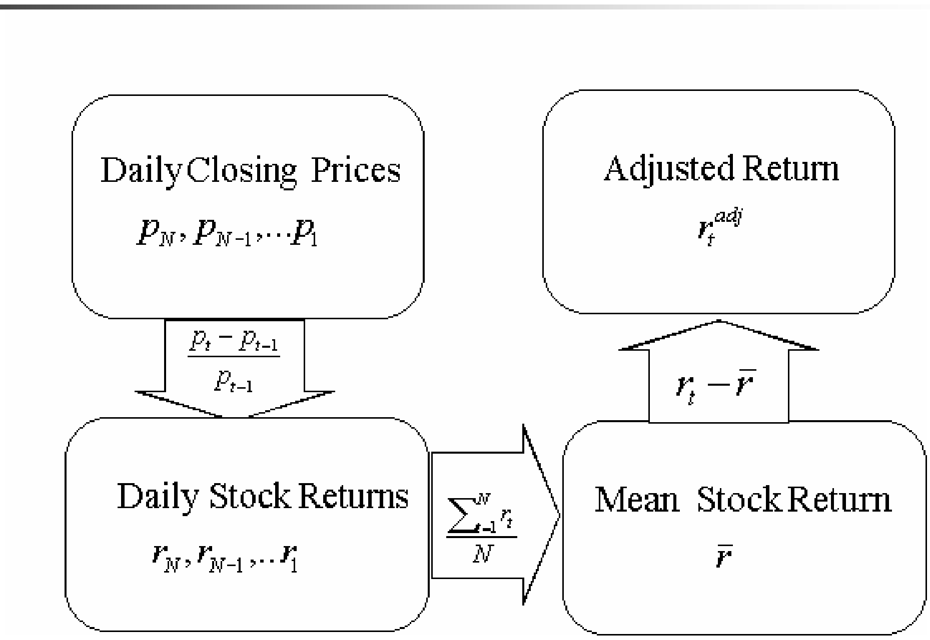
- Factors are independent
- Overcome rotation indeterminacies [Xu 2000]
- Factor determination via a simple cost function *J(k)* [Xu 2001]

Data Consideration

- Source: Hong Kong Stock Market
- Period: Jan 1, 1998 – Dec 31, 1999
- # of trading days: 522
- Total number of securities: 86
  - 30 Hang Seng Index (HSI)
  - 32 Hang Seng China-Affiliated Corporations Index (HSCCI)
  - 24 Hang Seng China Enterprises Index (HSCEI)

Kai-Chun Chiu, and **Lei Xu** (2003), ``NFA for Factor Number Determination in APT", *International Journal of Theoretical and Applied Finance*, pp 253-267, 2004.

Data Preprocessing



Test Methodology

- ML Factor Analysis
  - LR Statistics [Lawley & Maxwell 1963]

$$LR = (N - \frac{2p+4k+11}{6})\{(\ln|AA'+\Sigma| - \ln|S|) + (\mathrm{tr}[(AA'+\Sigma)^{-1}S] - p)\}$$

  - Follows χ^2 distribution with $[(p-k)^2 - (p+k)]/2$ degrees of freedom
  - Level of significance = 5%

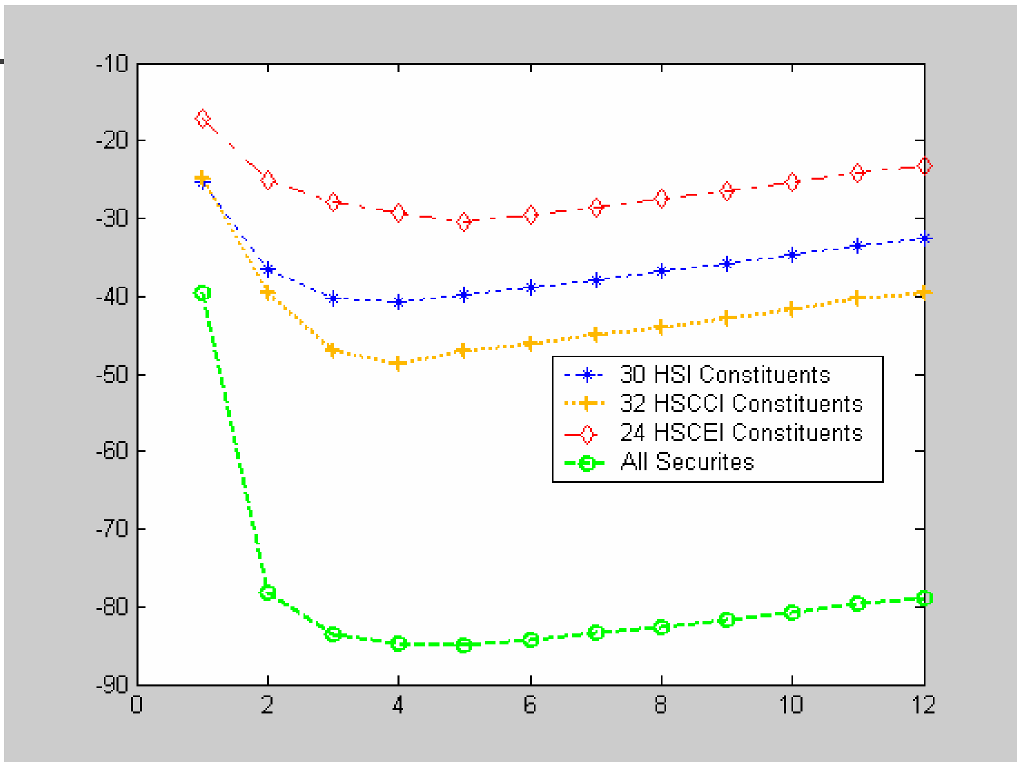
- Eigenvalues Analysis
  - Choose the number of eigenvalues that are significantly larger than the rest of the others.
- NFA
  - Model selection via the cost function *J(k)* [Xu 2001]

$$\min_k J(k) = \frac{1}{2}\left[\ln|\Sigma| - \frac{1}{N}\sum_{t=1}^{N} \ln q(\hat{y}_t \mid \hat{y}_{t-1}, \theta_y)\right]$$

Summarized Results

Stock Index	Total # of Securities	MLFA	Eigen-value	*J(k)*
HIS	30	11	1	4
HSCCI	32	12	1	4
HSCEI	24	9	1	5
All	86	33	1	5

NFA: Plot of *J(k)* for factor number determination



Result Interpretation and Analysis

- Implication by MLFA
  - factor # needed to explain cross-sectional security returns generation increases as more securities are added
- Implication by Eigenvalue Analysis
  - basically only one factor is needed to account for all returns (Conclusion in line with CAPM)
- Implication by NFA
  - Factor # is 4 or 5 (Consistent with the conjecture by Roll & Ross [1980])

Two Intuitive Question

- Q: Should factor # increases as more securities are added?
  - Probably not. So MLFA tends to bias towards more factors.
- Q: Is it likely that only one factor is enough?
  - Not quite so since the multi-factor APT is a generalization of the single-factor CAPM. So eigenvalue analysis tends to bias towards fewer factors

Temporal Factor Analysis

Xu, L (2001), ``BYY harmony learning, independent state space and generalized APT financial analyses '', IEEE Tr. on Neural Networks, 12 (4), 822-849.

Xu, L (2000), ``Temporal BYY learning for state space approach, hidden Markov model and blind source separation'', IEEE Tr. on Signal Processing 48, 2132-2144.

A Temporal Extension of APT

$$\begin{cases} y_t = By_{t-1} + \varepsilon_t \\ x_t = Ay_t + e_t \end{cases} \qquad t = 1,2,3\ldots n$$

y_t is independen t among its components

Adaptive Portfolio Management Algorithm

- **The way to find the hidden factors:**

Step 1 Fix A, B and Σ and estimate the hidden factors y_t by

$$y_t = \left[I + A^T \Sigma^{-1} A\right]^{-1} \left(A^T \Sigma^{-1} \overline{x}_t + By_{t-1}\right),$$

$$\varepsilon_t = y_t - By_{t-1},$$

$$e_t = \overline{x}_t - Ay_t.$$

Step 2 Fix y_t, update A, B and Σ_e by the gradient ascent approach

$$B^{new} = B^{old} + \eta diag\left[\varepsilon_t y_{t-1}\right],$$

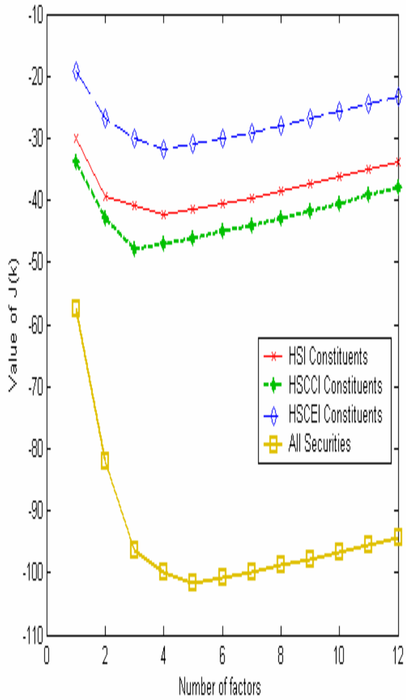
$$A^{new} = A^{old} + \eta e_t y_{t-1}{}^T,$$

$$\Sigma^{new} = (1-\eta)\Sigma^{old} + \eta e_t e_t{}^T.$$

Kai Chun Chiu and Lei Xu, ``A comparative study of Gaussian TFA learning and statistical tests for determination of factor number in APT'', Proceedings of International Joint Conference on Neural Networks 2002 (IJCNN '02), Honolulu, Hawaii, USA, May 12-17, 2002, pp2243-2248.

APT extensions

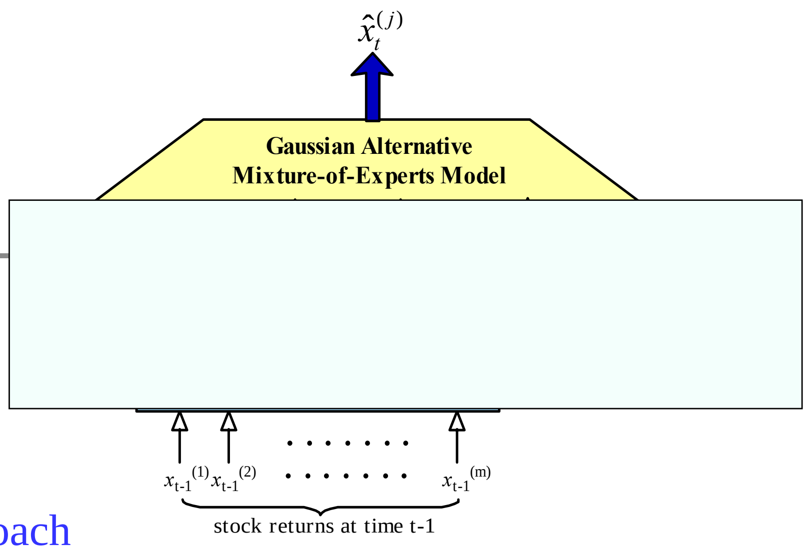
Stock Index	Total # of Securities	MLFA	Eigen-value	J(k)
HIS	30	11	1	4
HS CCI	32	12	1	3
HS CEI	24	9	1	4
All	86	33	1	5



3. Arbitrage Pricing Theory

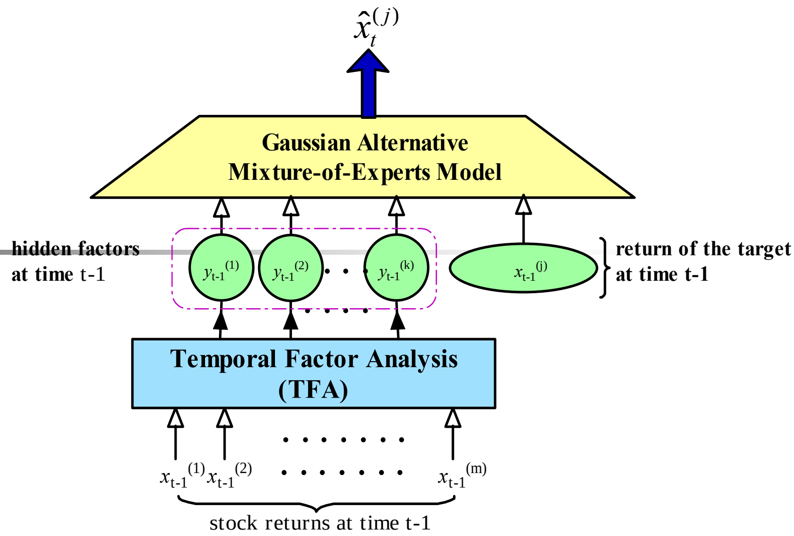
- Capital Asset Pricing Model vs. Arbitrage Pricing Theory
- Temporal Factor Analysis (TFA) and APT
- TFA based APT for Prediction
- TFA based APT for Portfolio Management

Kai Chun Chiu, and Lei Xu, (2002) "Stock price and index forecasting by arbitrage pricing theory-based gaussian TFA learning", in H. Yin et al., eds., Lecture Notes in Computer Sciences, Vol.2412, pp366-371, Springer Verlag.

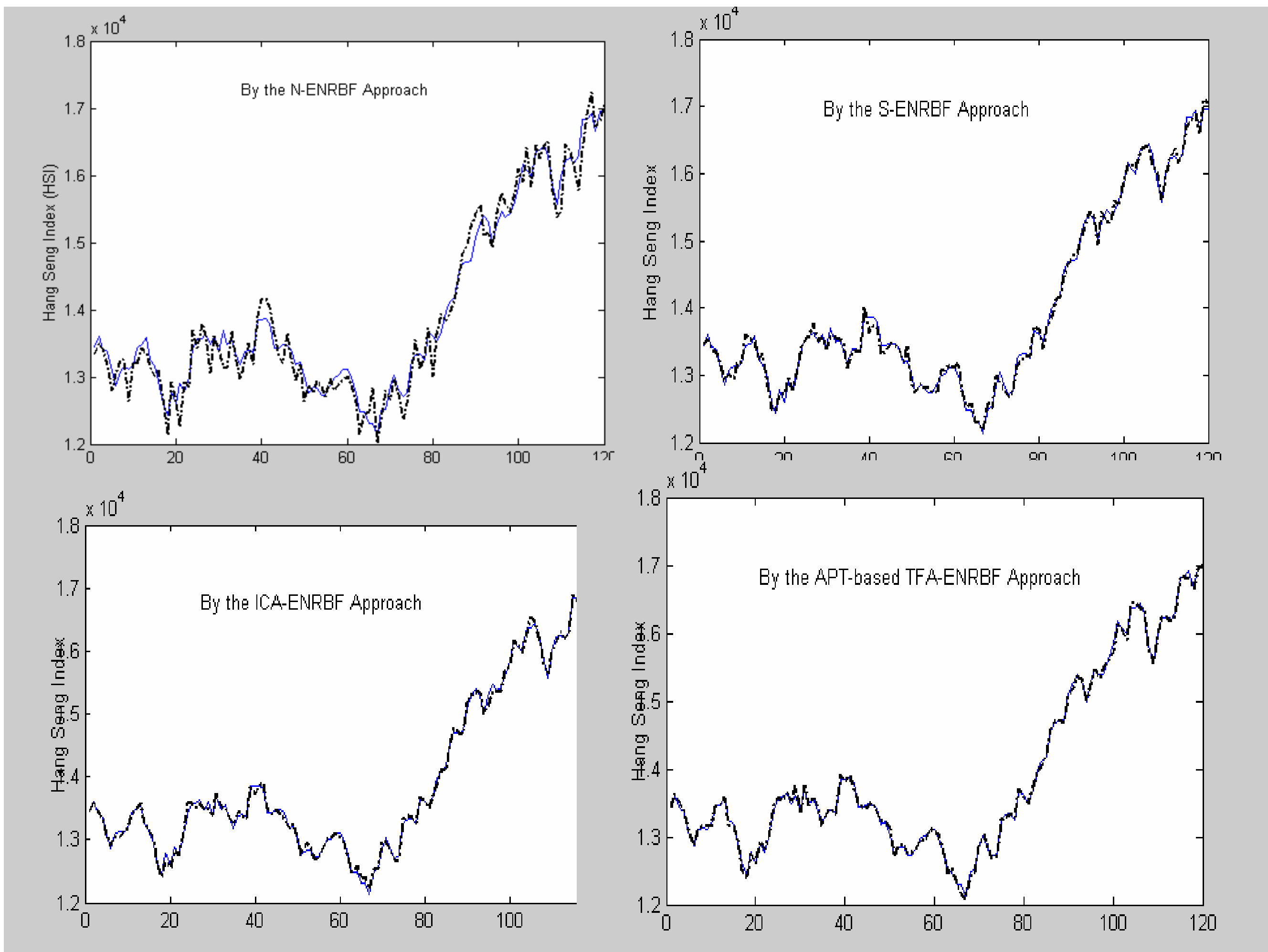


- N-ENRBF Approach
  - The adaptive ENRBF algorithm in [Xu, 1998] is used. The input vector consists of nonstationary raw index prices and is set as $x_t = [p_{t-1}, p_{t-2}, p_{t-3}]^T$ at time *t*.

- S-ENRBF Approach
  - Quite similar to the previous approach, the adaptive ENRBF algorithm is adopted. The input vector at time t is stationary returns $x_t = [\tilde{R}_{t-1}, \tilde{R}_{t-2}, \tilde{R}_{t-3}]^T$



- ICA-ENRBF Approach
  - Step 1: the inverse mapping $y_t = Wx_t$ is effected on the stock price of index constituents via the technique called Independent Component Analysis (ICA) for higher order dependence reduction;
  - Step 2: Then, the adaptive ENRBF algorithm is adopted for establishing the relationship between $y_{t-1}, x_{t-1}^{(j)}$ and $x_t^{(j)}$
- APT-Based TFA-ENRBF Approach
  - Step 1: the Gaussian TFA algorithm instead of the LPM-ICA algorithm is used to recover independent hidden factors;
  - Step 2: Same as the previous approach.

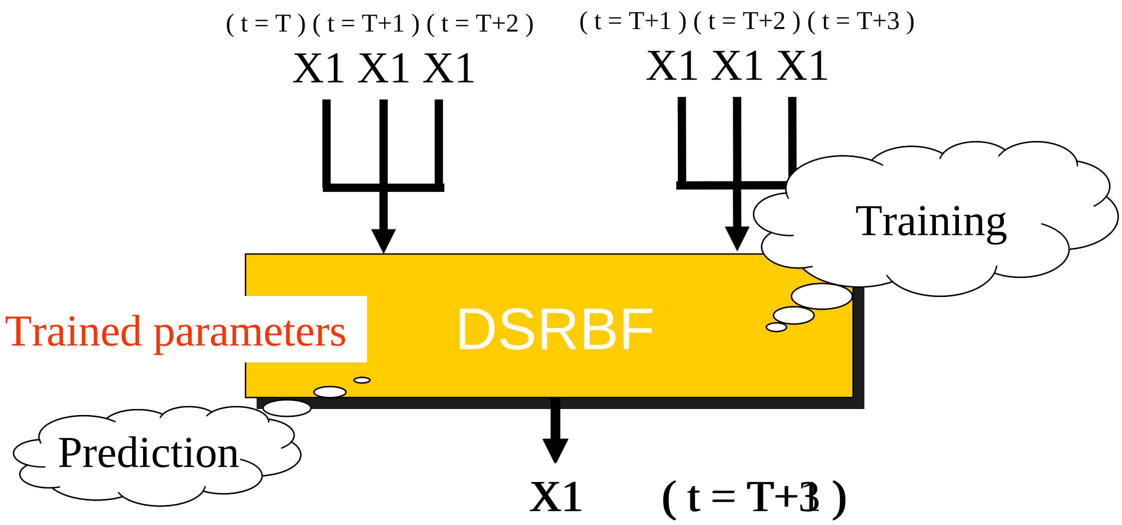


Experimental Results (RMSE)

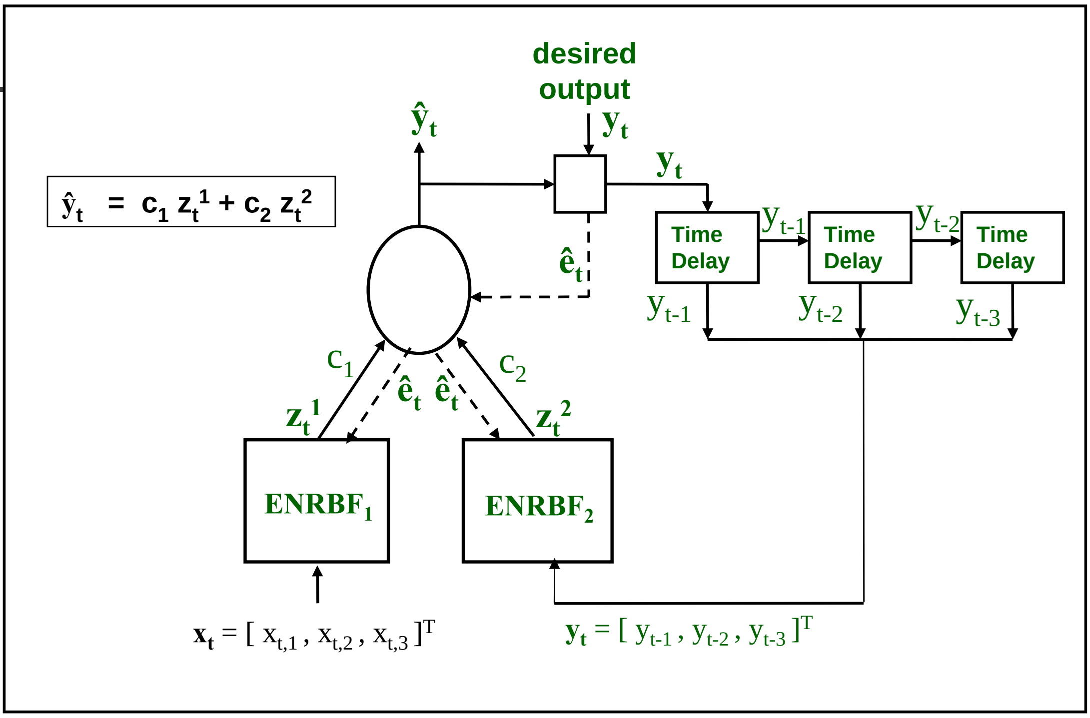
Approach	HSI	HSCCI	HSCEI	HSBC
N-Adaptive ENRBF	232.9625	25.8021	9.9819	0.7957
S-Adaptive ENRBF	80.8164	8.7290	4.2516	0.4347
ICA-ENRBF	63.9681	6.0765	3.4340	0.3147
APT-based TFA-ENRBF	47.6031	4.5202	2.2187	0.2346

Implementation of TFA

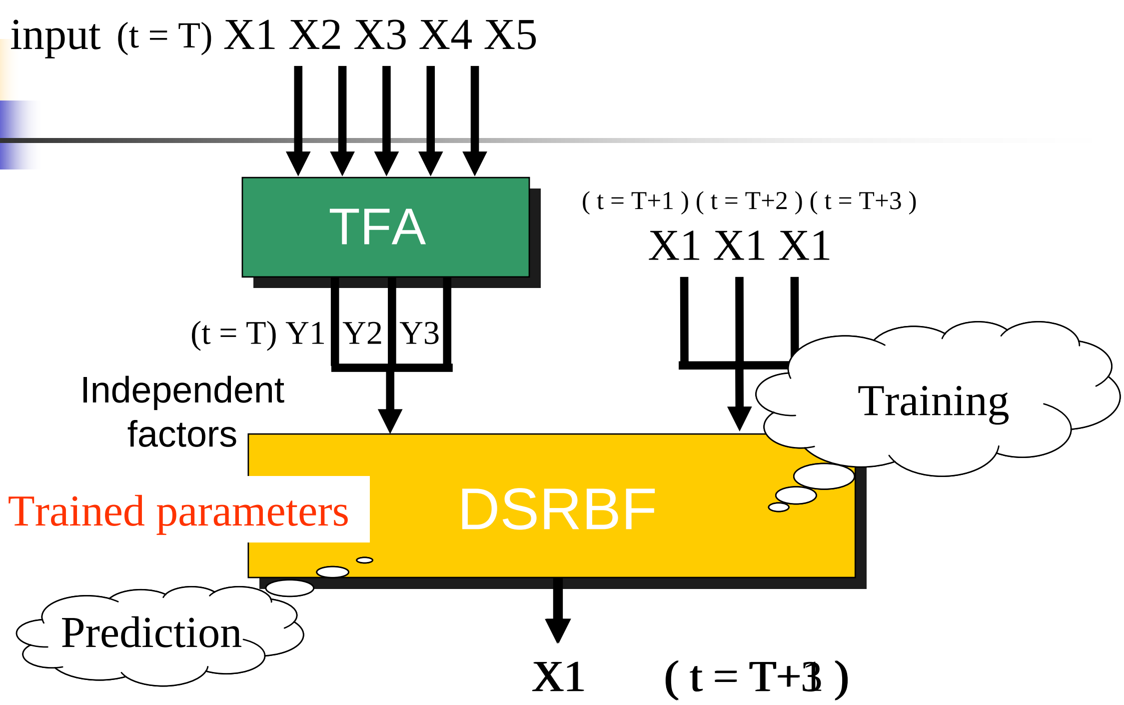
input



DSRBF

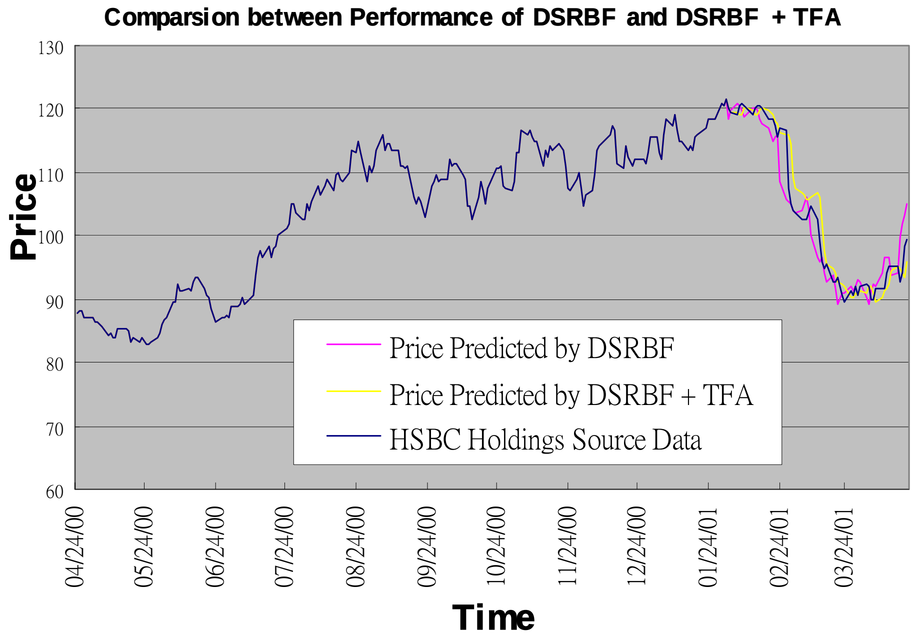


Implementation of TFA



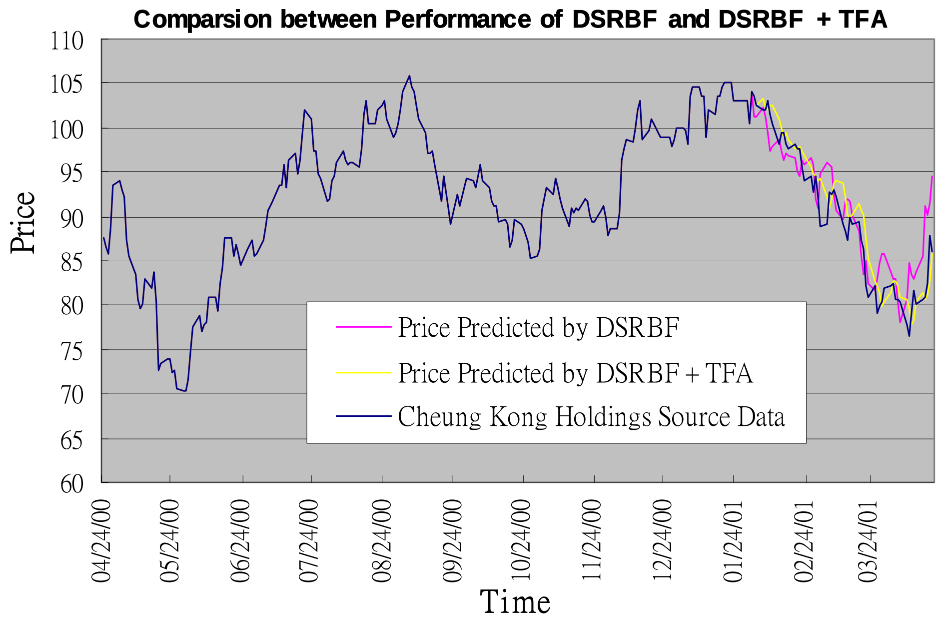
HSBC Holdings

Neural Network	*Mean Square Error of Testing Data*
DSRBF	*10.40589*
DSRBF + TFA	**8.268526**



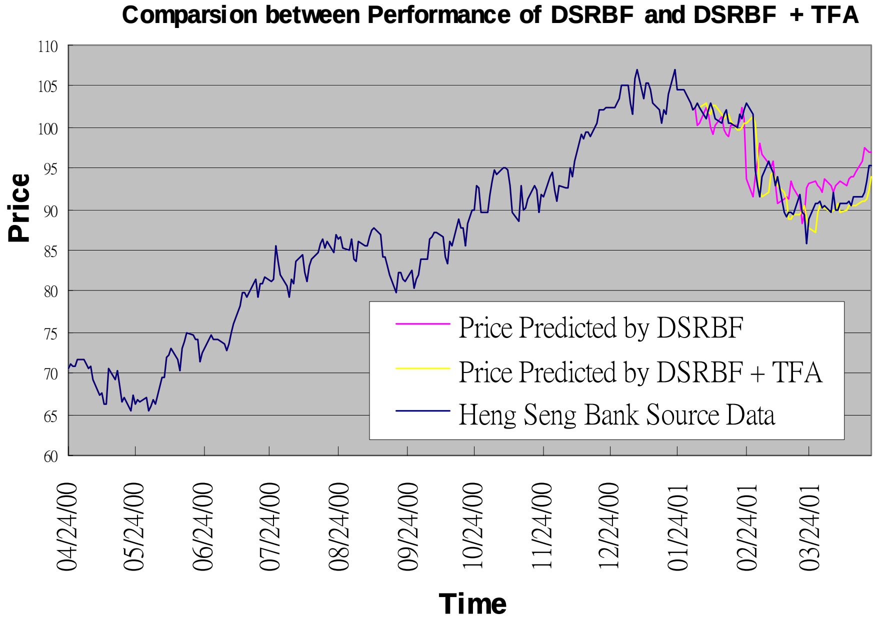
Cheung Kong Holdings

Neural Network	*Mean Square Error of Testing Data*
DSRBF	*13.48662*
DSRBF + TFA	**5.100805**



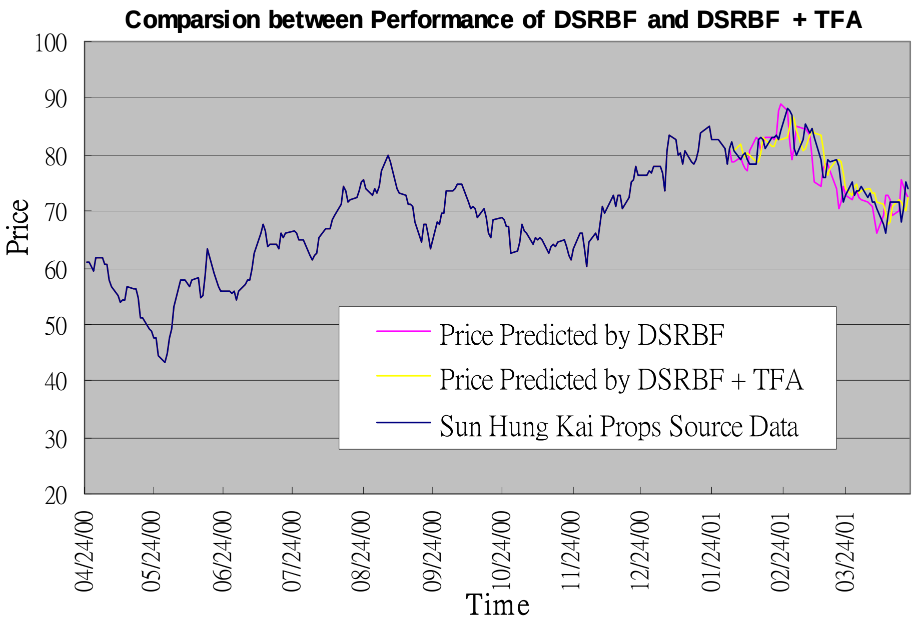
Heng Seng Bank

Neural Network	*Mean Square Error of Testing Data*
DSRBF	*10.46414*
DSRBF + TFA	**2.95054**



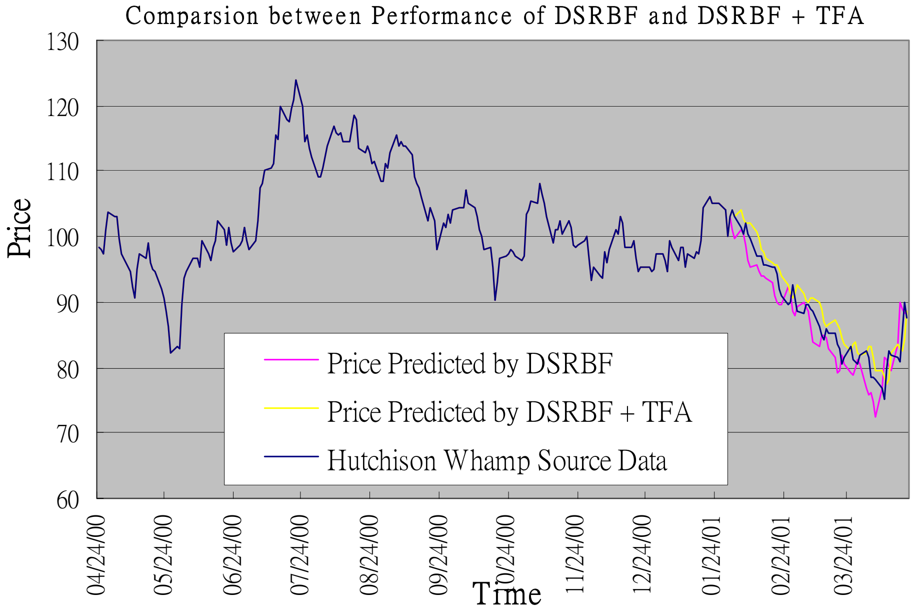
Sun Hung Kai Props

Neural Network	*Mean Square Error of Testing Data*
DSRBF	*11.38626*
DSRBF + TFA	**5.948012**



Hutchison Whamp

Neural Network	*Mean Square Error of Testing Data*
DSRBF	*10.03561*
DSRBF + TFA	**5.945724**



3. Arbitrage Pricing Theory

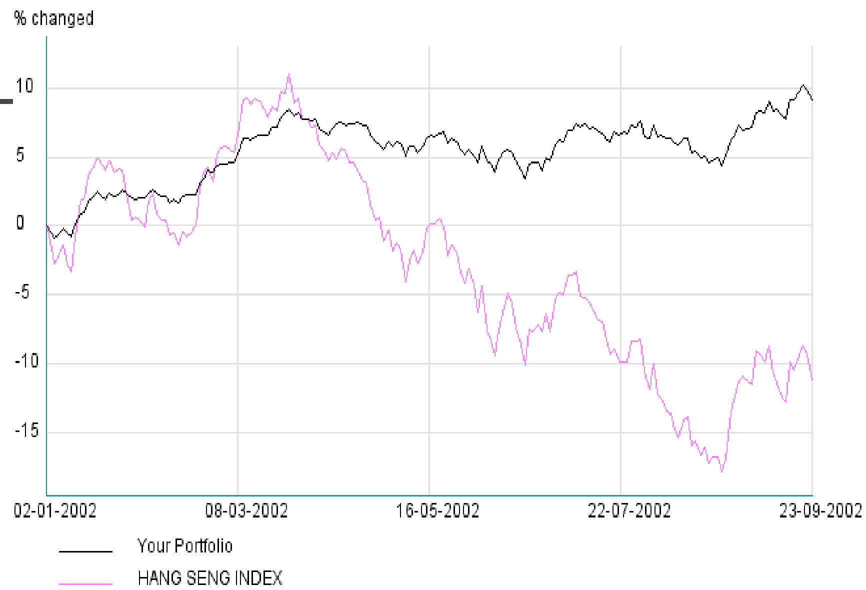
- Capital Asset Pricing Model vs. Arbitrage Pricing Theory
- Temporal Factor Analysis (TFA) and APT
- TFA based APT for Prediction
- TFA based APT for Portfolio Management

Kai-Chun Chiu and Lei Xu, (2004) ``Arbitrage Pricing Theory Based Gaussian Temporal Factor Analysis for Adaptive Portfolio Management'', Decision Support Systems 37, pp 485- 500, 2004.

Observations Based

	Mean	Standard Deviation	Max	Min	Sharpe Ratio
The Portfolio	*1.0249*	*0.0260*	*1.0896*	*0.9520*	*39.4192*
Hang Seng Index	0.9771	0.0727	1.1099	0.8211	13.4402

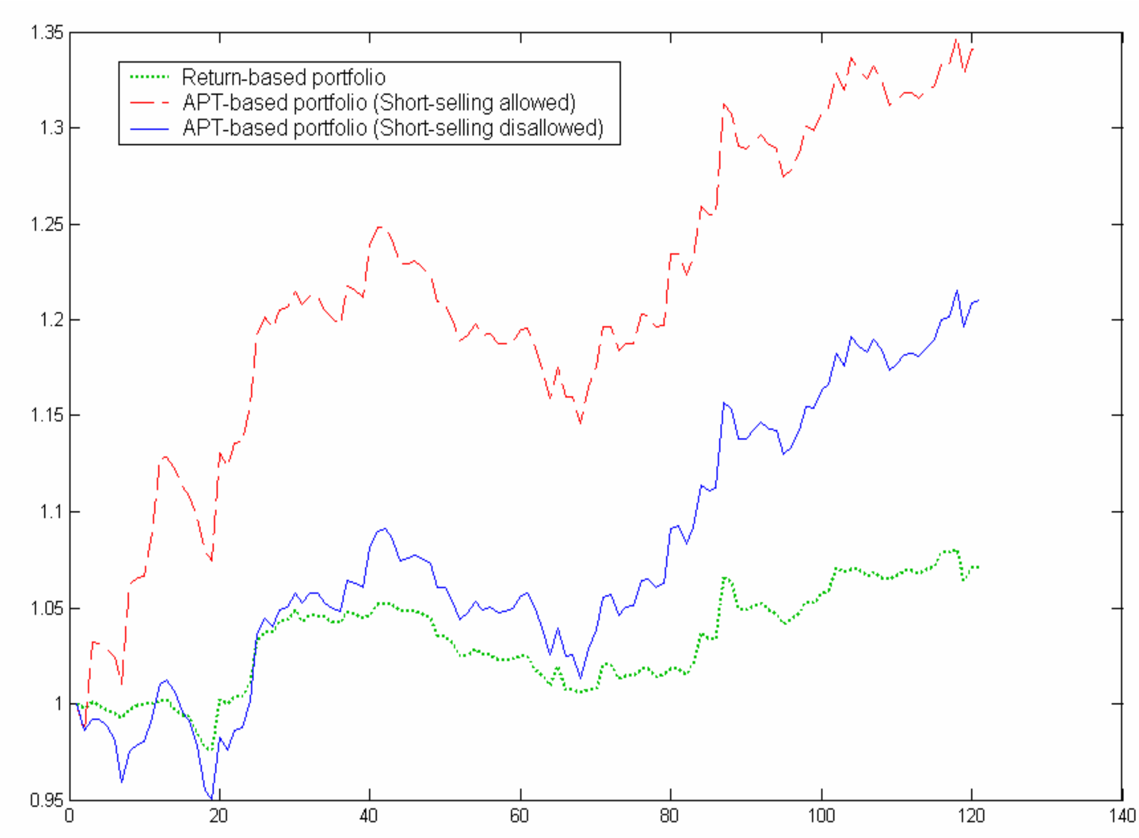
Hidden Factors Based



	Mean	Standard Deviation	Max	Min	Sharpe Ratio
The Portfolio	*1.0541*	*0.0235*	*1.1020*	*0.9910*	*44.8553*
Hang Seng Index	0.9771	0.0727	1.1099	0.8211	13.4402

Attributes	Mean	Standard Deviation	Maximum	Minimum	Sharpe Ratio
Change	*+2.8491%*	*-9.6154%*	*+1.1380%*	*+4.0966%*	*+13.7905%*

- hidden factors based
  - It generated a better return
  - Lower risk
  - Sharpe ratio increased by more than 13%



略

Risk-Return Statistics

Component Name	Expected Return	Risk	S_p
	(mean)	(std. dev)	
Risk-free Security	0.00148%	0.0018%	
HSI	0.18%	1.48%	
HSCCI	0.03%	2.51%	
HSCEI	-0.20%	2.55%	
Return-based Portfolio (short selling disallowed)	0.08%	0.61%	0.13
APT-based Portfolio (short selling disallowed)	0.19%	1.04%	0.18
APT-based Portfolio (short sell allowed)	0.33%	1.62%	0.20

4. Challenges and Advances of Statistical Learning

- Two types of Intelligent Ability: Learning from Samples
- Key Ingredients of Statistical Learning
- Two Key Challenges and Advances on Seeking Solutions
- A Unified Theory: Bayesian Ying-Yang Harmony Learning

Fundamentals, Challenges, and Advances of Statistical Learning for Knowledge Discovery and Problem Solving: *A BYY Harmony Perspective*

面向知识发现和问题求解的統計學習：
基本问题、主要挑战、和統一理论

Lei Xu

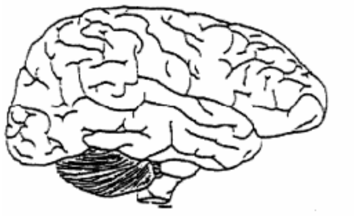
http://www.cse.cuhk.edu.hk/~lxu/

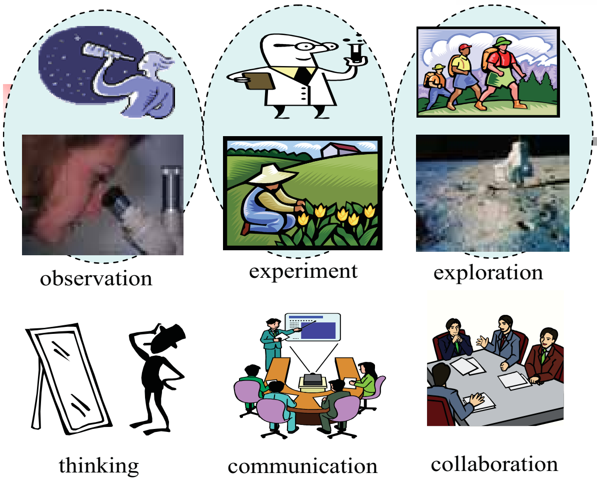
Department of Computer Science and Engineering,
The Chinese University of Hong Kong

Outlines

- **Two types of Intelligent Ability: Learning from Samples**
  发现知识和求解问题是体现智能的两个基本能力--通过学习获得
- **Key Ingredients of Statistical Learning**
  从有限个样本中学习--统计学习的三个基本要素
- **Two Key Challenges and Advances on Seeking Solutions**
  两个主要挑战--几十年来应对挑战的发展轮廓
- **A Unified Theory: Bayesian Ying-Yang Harmony Learning**
  一个统计学习之统一理论体系

Two types of Intelligent Ability





TYPE I

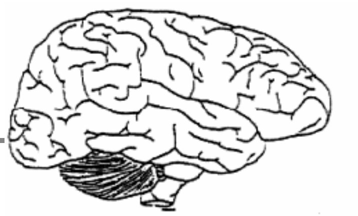
Knowledge about the world it survives

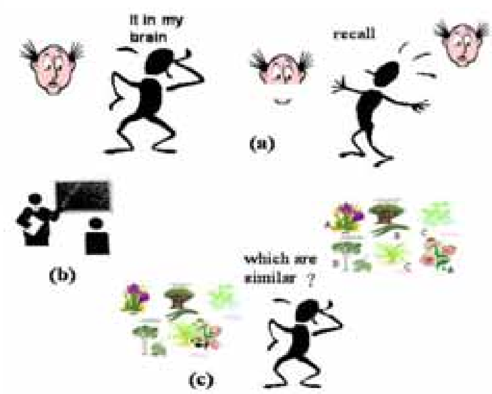
Why ?

(interpret what are observed)

The World

Two types of Intelligent Ability





TYPE II

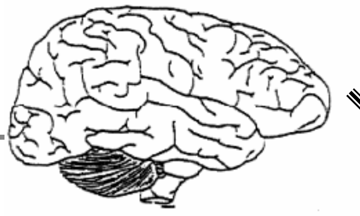
skill of handling each issue encountered in the world

How to do ?

(problem solving)

The World

How to get the abilities



TYPE II

Obtain Skills of Problem Solving via

• reasoning , inference, optimization

•learning from samples (fast implementation)

TYPE I

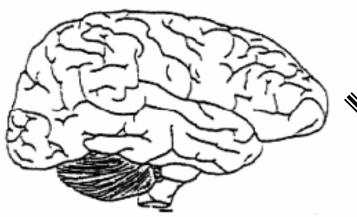
Obtain World Knowledge via

• loading from authorized sources (e.g., textbooks)

•learning from samples (pieces of uncertain evidences)

The World

Two Types of Learning from Samples



TYPE II

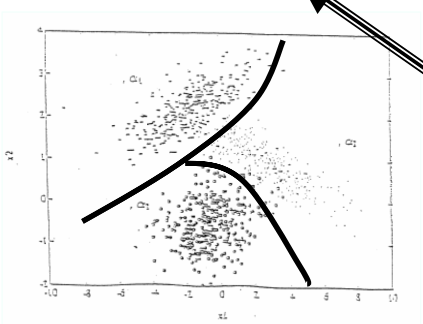
Training Skills of Problem Solving

via building up input-response type dependence per sample

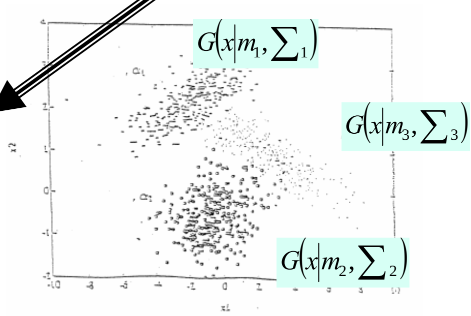
TYPE I

Discovering World Knowledge

via mining invariant dependence underlying a set of all samples

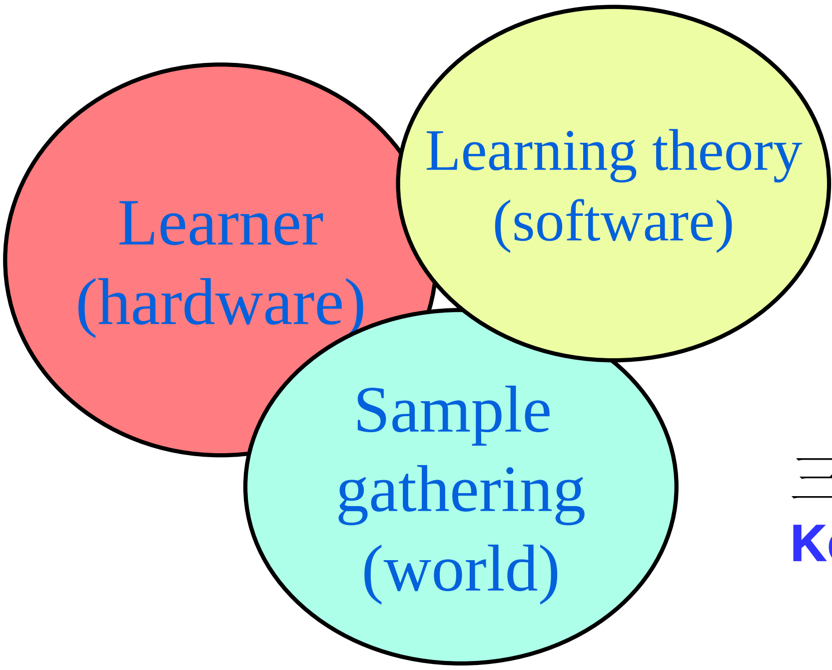


The World



Statistical Learning

Using statistical approach for removing uncertainties from **Sampling and observation noises**



三个基本要素

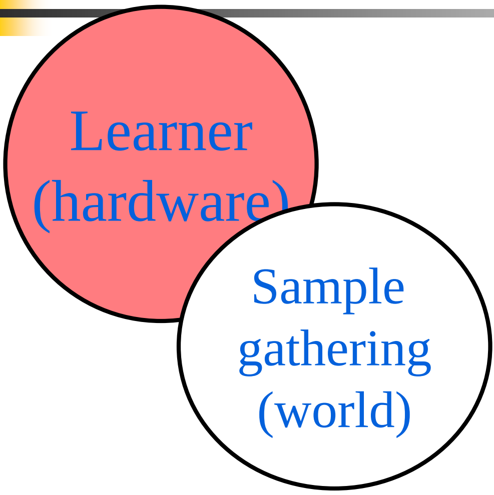
Key Ingredients

4. Challenges and Advances of Statistical Learning

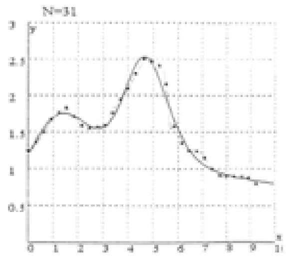
- Two types of Intelligent Ability: Learning from Samples
- Key Ingredients of Statistical Learning
- Two Key Challenges and Advances on Seeking Solutions
- A Unified Theory: Bayesian Ying-Yang Harmony Learning

Key Challenge 主要挑战 I

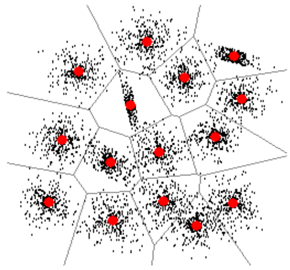
Learner's hardware appropriately represents dependences among data
(matching structures of underlying world)



Regression



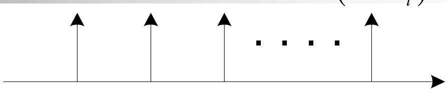
Multi-point



Memory based: individual 逐个记忆

Empirical density

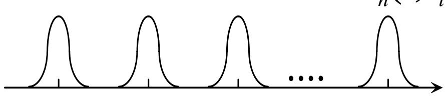
$$p_0(x)=\frac{1}{N}\sum_{t=1}^{N}\delta(x-x_t)$$



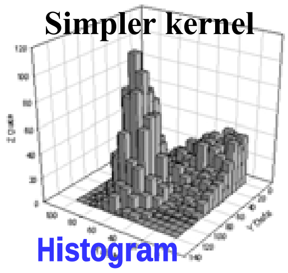
Parzen window density

Blurred memory 记忆与适当模糊

$$p_h(x)=\frac{1}{N}\sum_{t=1}^{N}K_h(x,x_t)$$

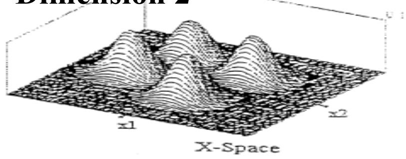


Simpler kernel



Histogram

Dimension 2

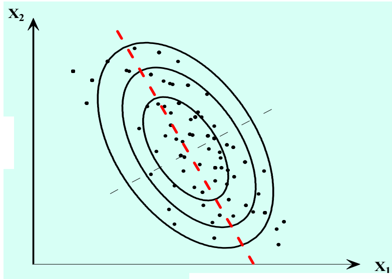


Curse of dimension !

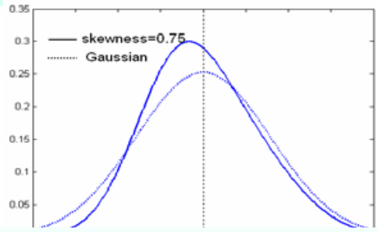
Ensemble **Feature based:** 总体特征

$$\mu_i = \frac{1}{N}\sum_{t=1}^{N} x_i^t \qquad \mu_i = E(x_i)$$

- Mean and covariance matrix

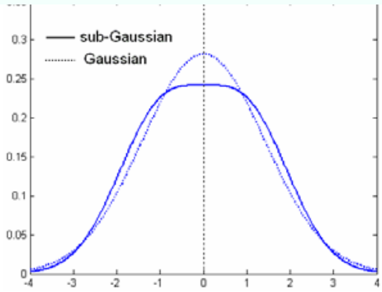


$$\sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j)$$



- higher order statistics
  - third-order: skewness
  - fourth-order: kurtosis
  - ...

$$\rho_{ii\cdot\cdot i} = E(x_i - \mu_i)(x_i - \mu_i)\cdot\cdot(x_i - \mu_i)$$



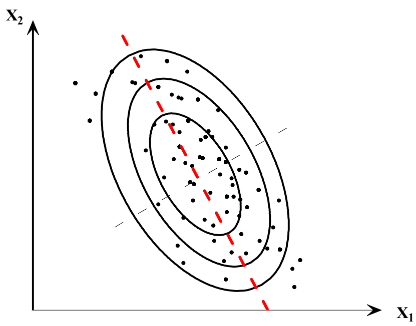
$$\rho_{ij\cdot\cdot m} = E(x_i - \mu_i)(x_j - \mu_j)\cdot\cdot(x_m - \mu_m)$$

The number increases exponentially !

Specific purpose: **Parametric family**

专用目的: 参数族

- Gaussian $G(x\,|\,m,\Sigma)$



Domain specific densities

e.g., exponential family

Case by case: too narrow for a general purpose !

Best: Seeking Structures that indirectly specify distribution families

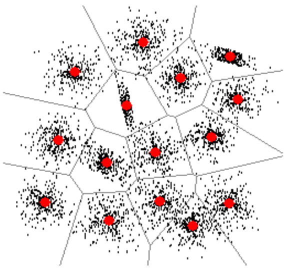
通过结构间接表示分布族

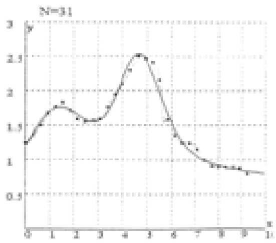
- Start at typical structures 典型结构

Multi-point

Regression

Two Previous examples





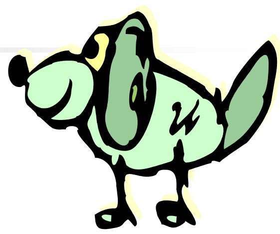
+ more

(to be introduced later)

- Aim at a general framework 通用框架 to integrate
    - existing studies
    - investigating new structures

One-body world

Dependence structures among samples from one-body world

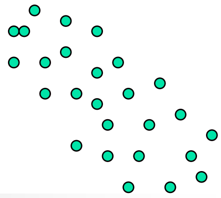


VS

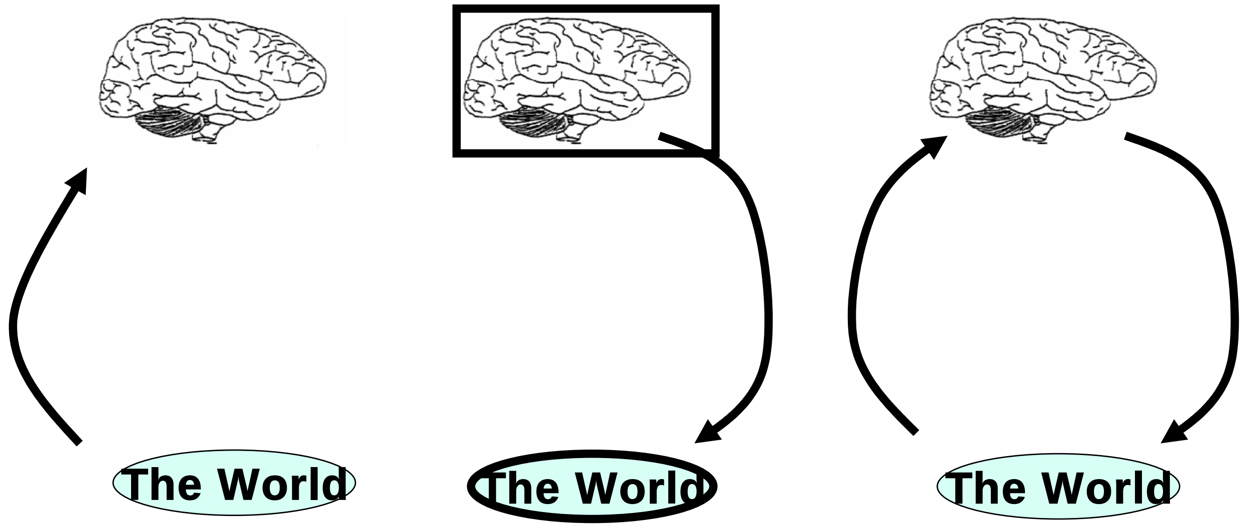
Multi-bodies world

Dependence structures among samples from multi-body world

Dependence structures among samples from one-body world



Three Architectures 三种构筑

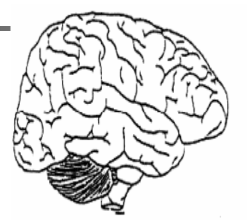


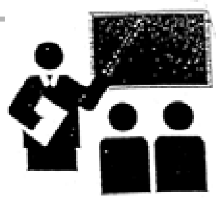
to be introduced one by one

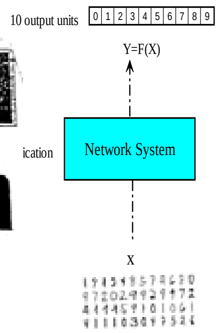
Forward Architecture

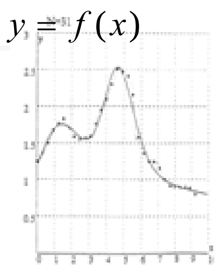
Pair-wise structures

$y = f(x)$









TYPE II

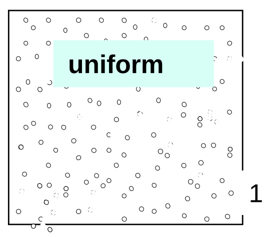
Training Skills of Problem Solving

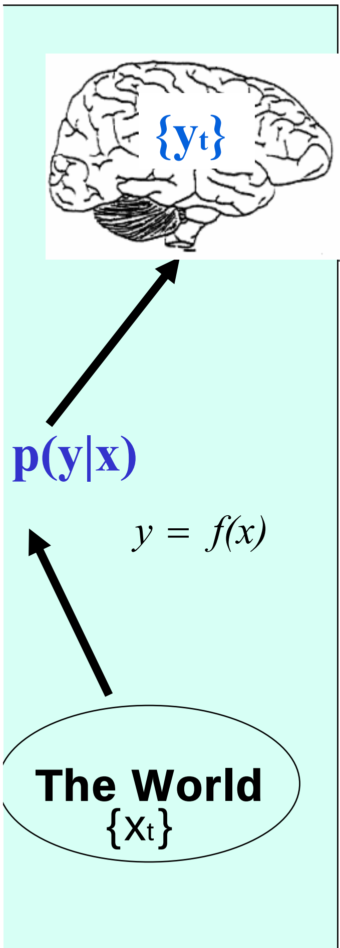
via building up input-response type dependence per sample

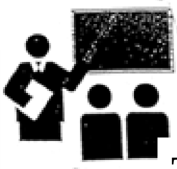
The World

Redundancy reduction structures

Redundancy's role for understanding perception (Attneave, 1954), sensory pathways (Barlow 1959, 1989), and pattern recognition (Watanabe, 1960)

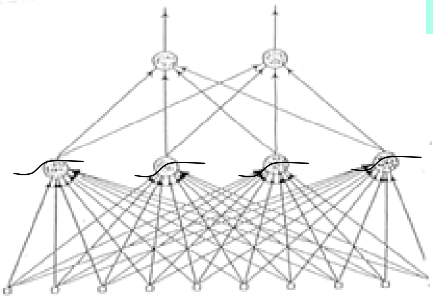




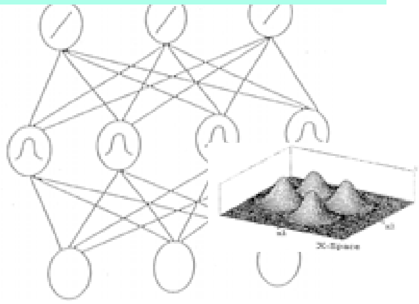


Pair-wise structures

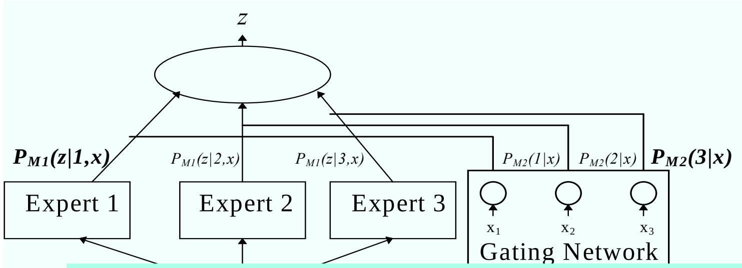
Three layer net



RBF & Kernel regression (Xu & Yuille, 92&94)
Adaptive EM algorithm (98)

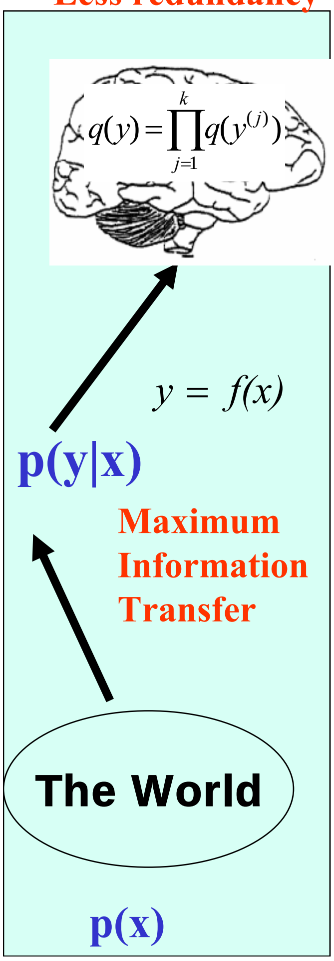


Mixture of Experts (ME)

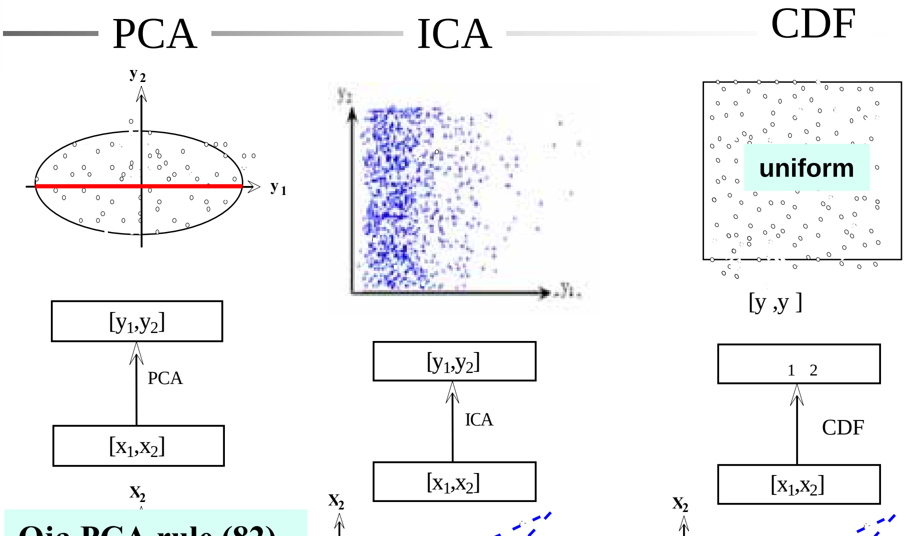


Classifier combination (Xu, Krzyzak, & Suen, (90&92)
EM convergence study (Jordan & Xu, 95)
Alternative ME model (Xu, Jordan, &Hinton, 94& 95)

Less redundancy



Redundancy reduction structures

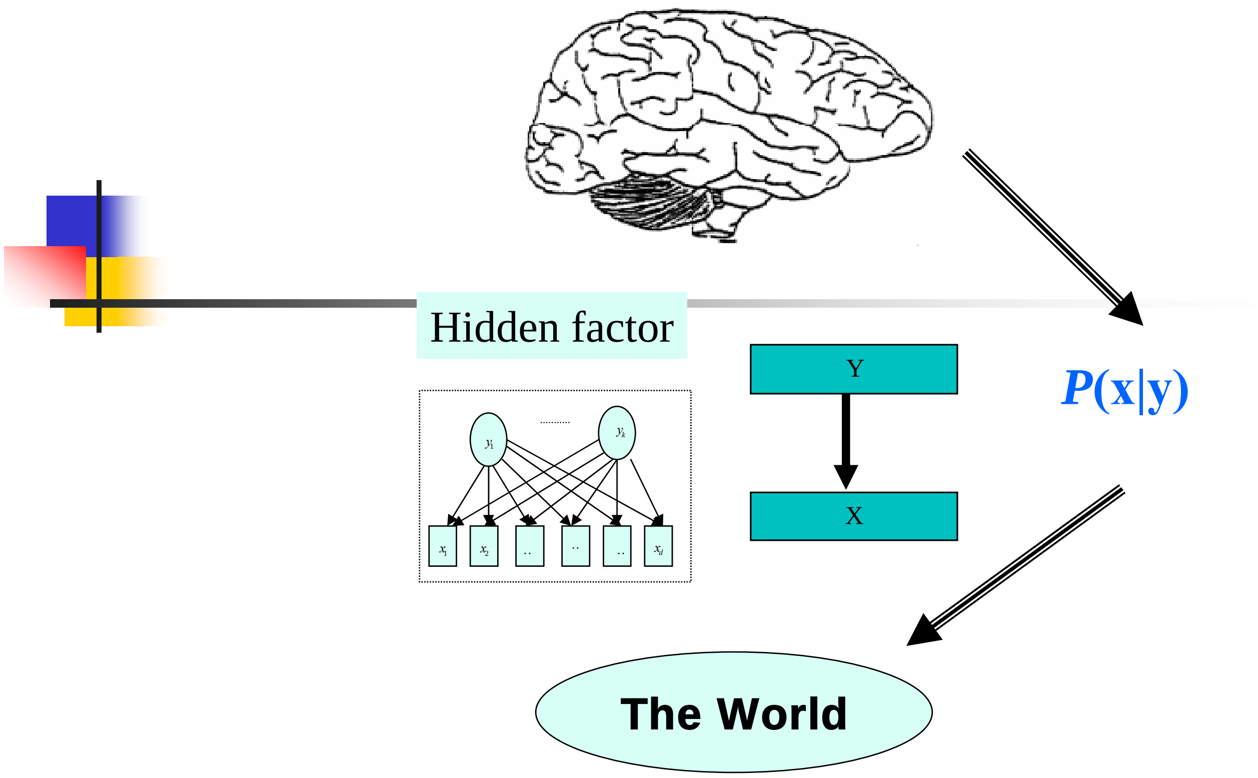


Oja PCA rule (82)
Oja PSA rule (89)
Xu (91) PSA global convergence and PSA → k-PCA (symmetry broken)

Informax (Bell, 94) for superGaussian
MMI (Amari, 95) for subGaussian
FastICA (Oja, 96) for both (sequential)
LPM-ICA(Xu, 96) for both (parallel)
One-bit-conjecture & proof (Xu, 98 &03)

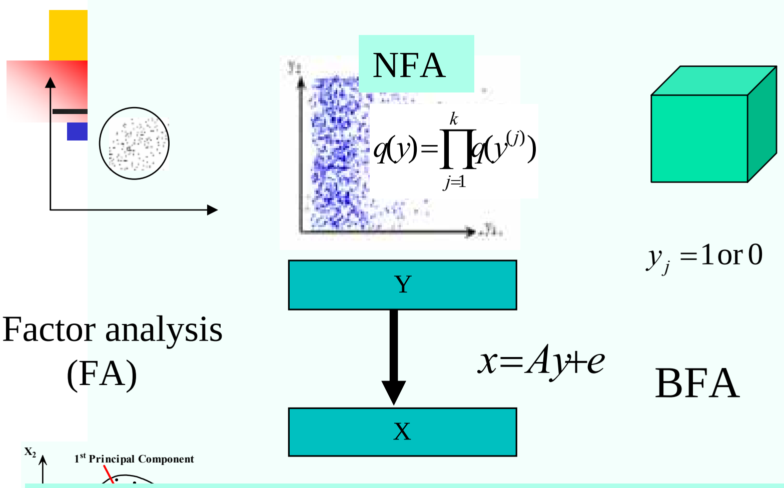
Backward Architecture

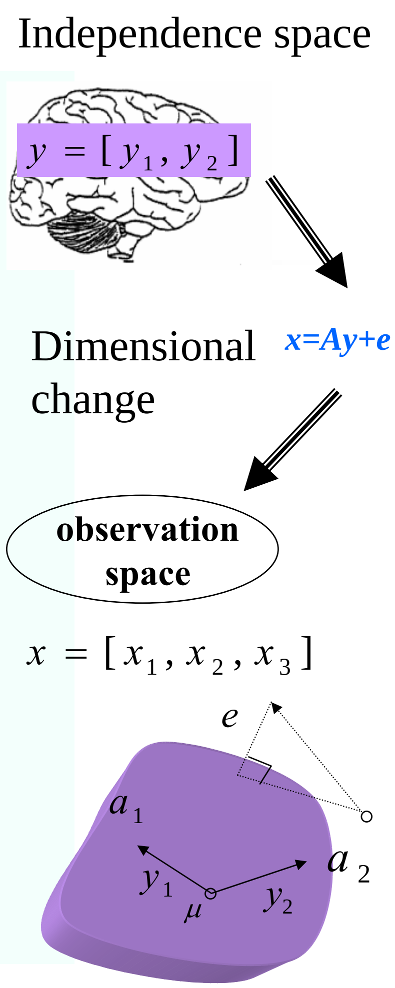
How observations generated



Independence subspace

(Linear Latent structures)

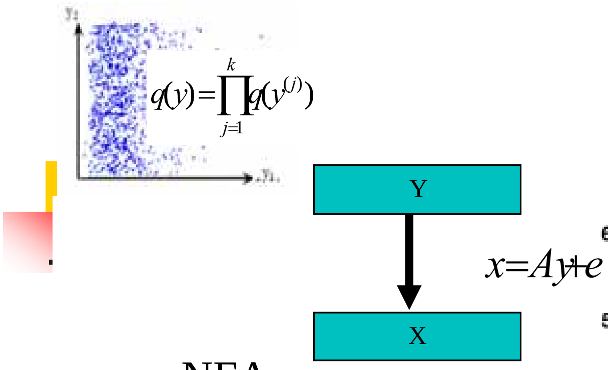




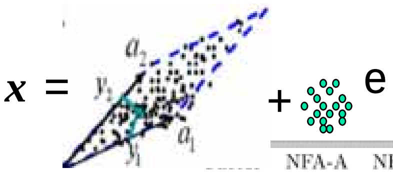
for FA & BFA：**adaptive algorithm & J(k) curve (Xu, 98)**

For NFA: **LMSER (Xu, 91&93), approximately**
EM algorithm (France,96), much exactly
BYY learning (fast !) and J(k) curve (Xu, 01&02)

For all the three: **adaptive BYY learning algorithm with k selected automatically during learning (Xu, 03&03).**

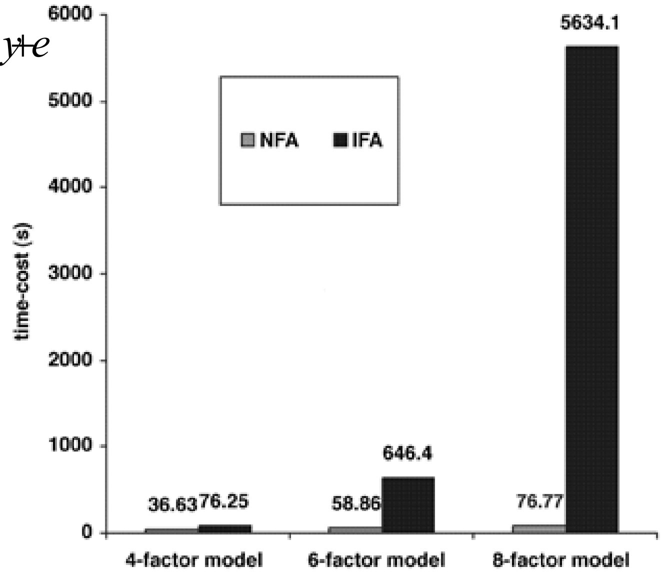


NFA



	NFA-A	NFA-O	IFA
1	.0369	.0379	.0426
2	.0401	.0299	.0292
3	.0446	.0336	.0400
4	.0189	.0245	.0344
Mean	.0351	.0315	.0366

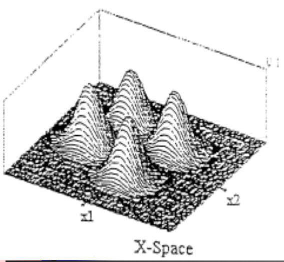
(a)

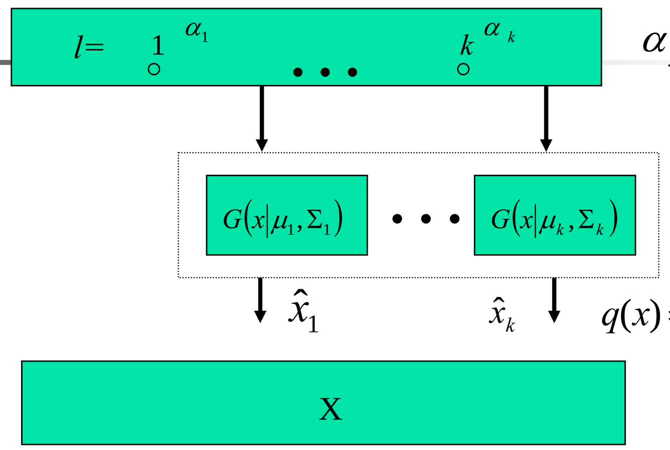


(b)

Fig. 7. Comparisons between NFA and IFA. (a) On the MSEs between the recovered factors and the original factors. (b) On time complexity.

Gaussian Mixture

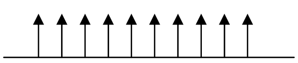


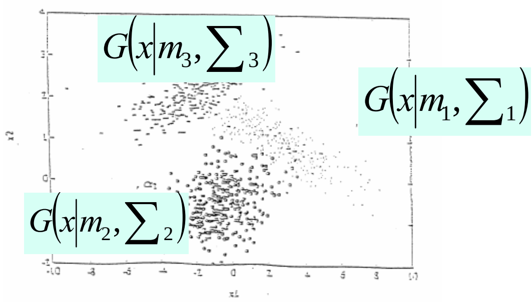


$$\alpha_j \geq 0, \quad \sum_{j=1}^{K} \alpha_j = 1$$

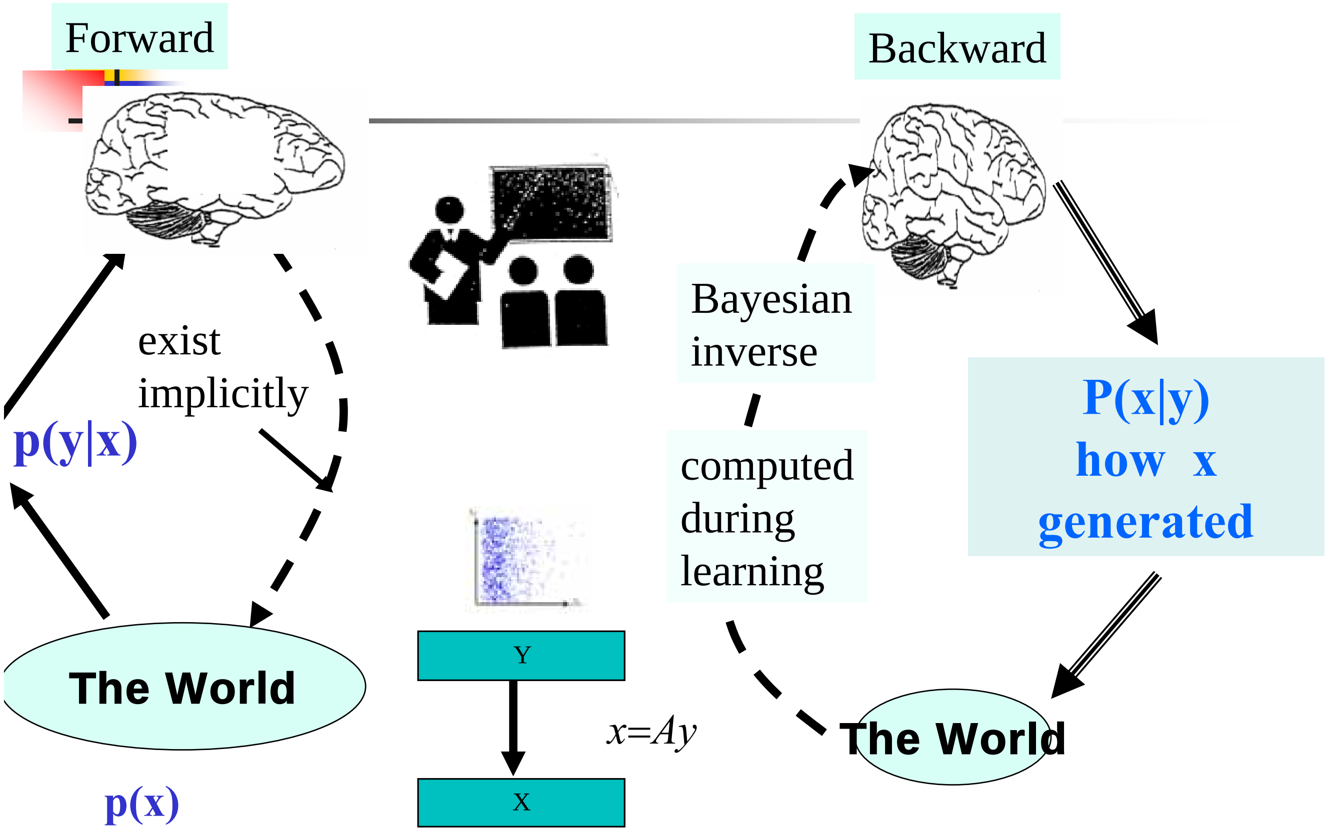
$$q(x) = \sum_{r=1}^{k} \alpha_r G(x|\mu_r, \Sigma_r)$$

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)$$

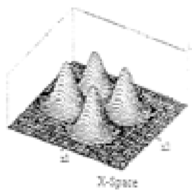


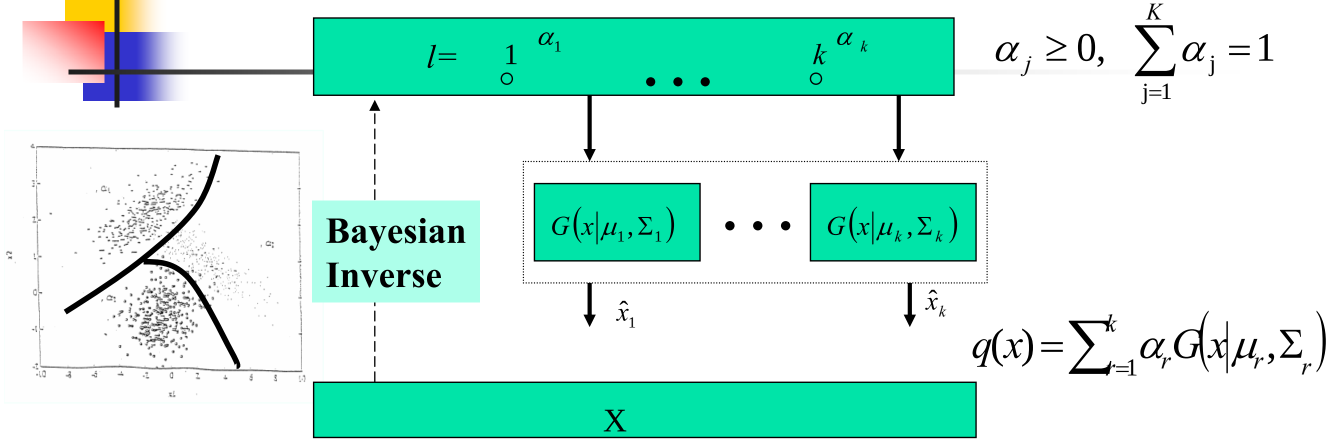


A bi-directional perspective



Gaussian Mixture



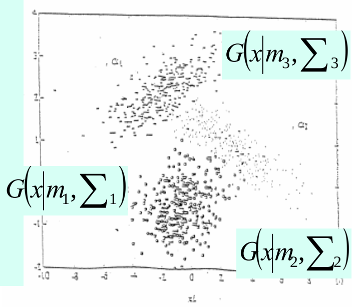


EM convergence and three advantages (Xu & Jordan, 92)

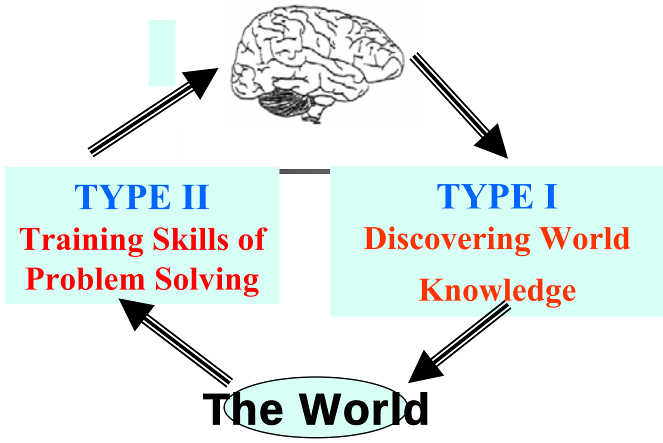
Hard-cut EM with automatic selection on k (Xu, 95&96)

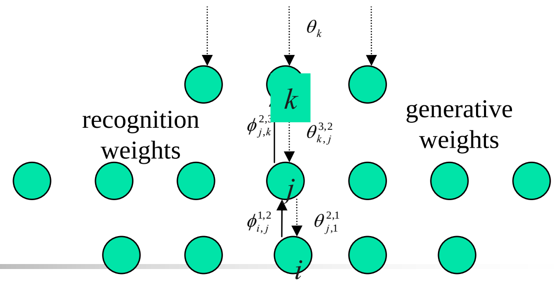
J(k) curve for k (Xu, 96 &97)

RPCL with automatic selection on k (Xu, Krzyzak, Oja, 91&93)

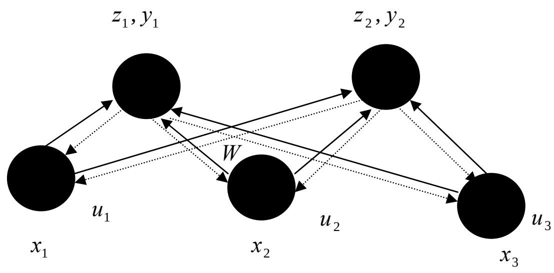


Bi-directional structures

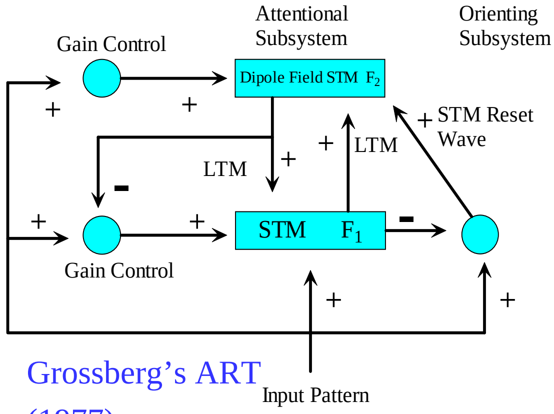




Helmholtz Machine (1995)



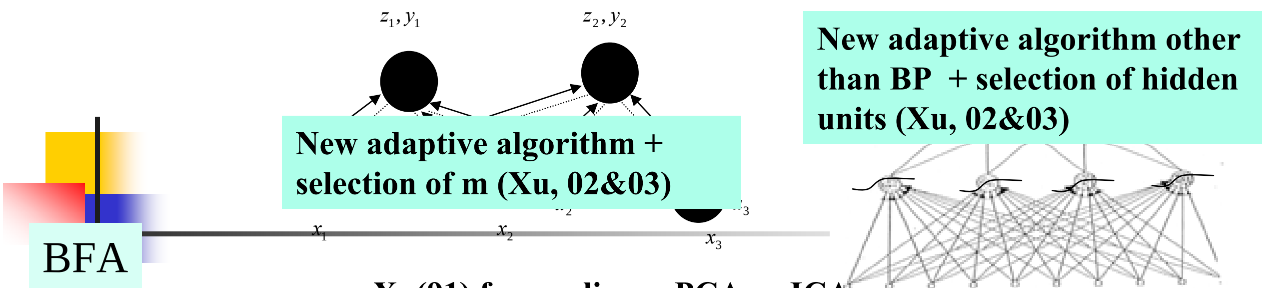
LMSER (Xu, 1991)



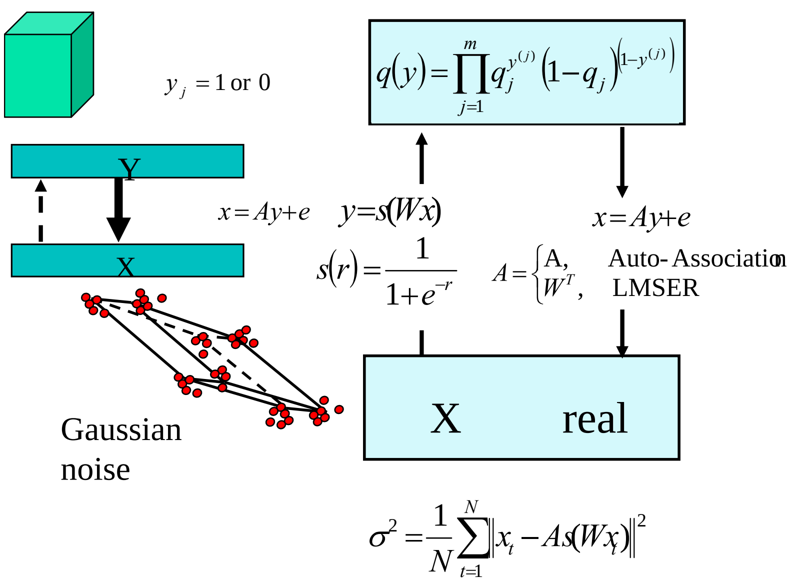
Grossberg's ART (1977)

Others

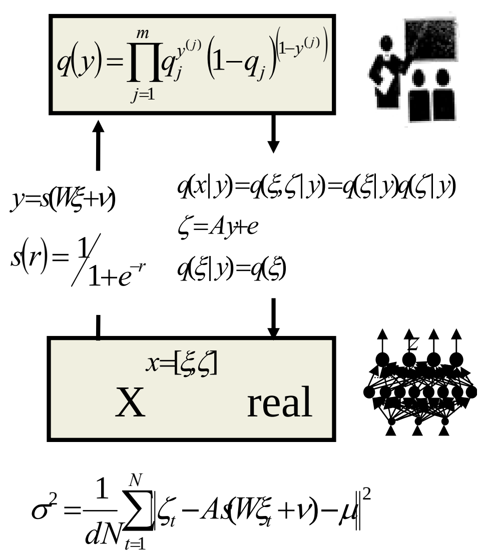
- **Kawato et al's Forward-inverse optics model**
- **Pattern Theory (Mumford, Grenander)**



Xu(91) for nonlinear PCA or ICA

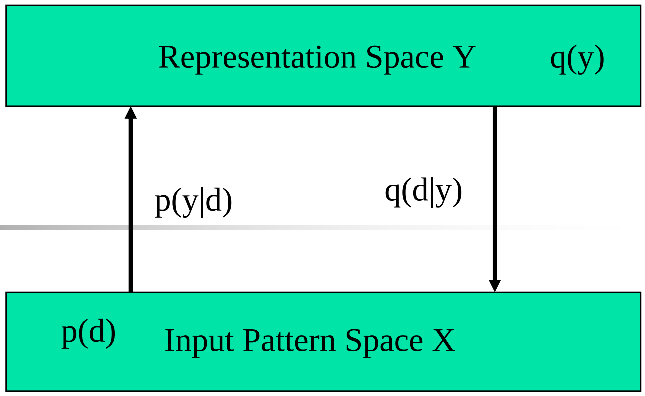


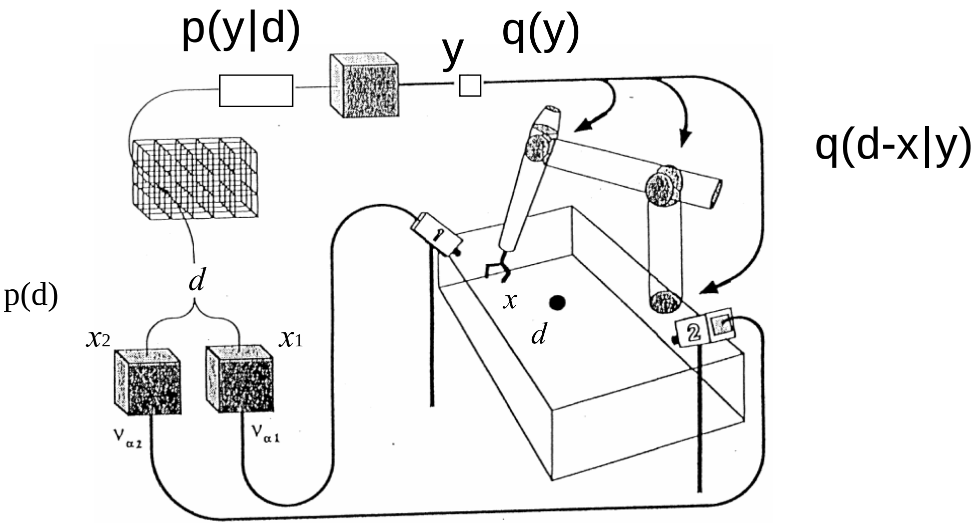
(a)



(b)

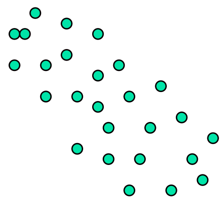
Motor Control





one-body world

(see Proceedings for details)

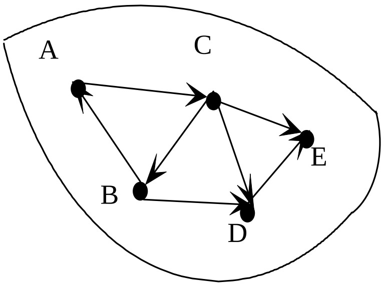


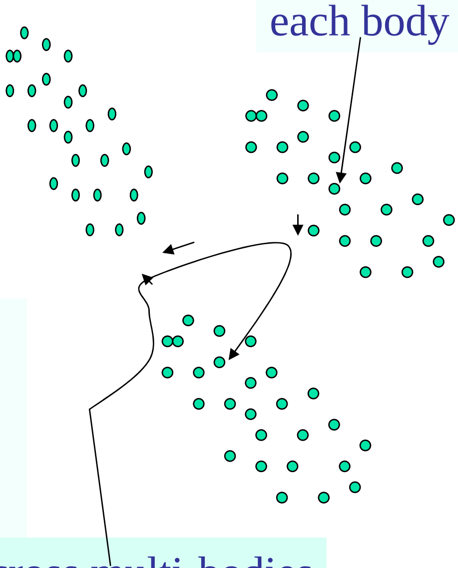
Dependence structures among samples from one-body world

- **pair-wise architectures**
  - Regression Function fitting
  - Co-occurrence association
  - Boundary structure
- **transform architectures**
  - Principal component analysis (PCA)
  - Independent component analysis (ICA)
  - CDF transform
- **factor architectures**
  - Gaussian factor
  - Binary factor
  - Non-Gaussian factor
  - Positive factor
- **bi-directional architectures**
  - ART
  - LMSER
  - three layer net
  - Helmholtz machine

Multi-body world

- *qualitatively* by the topology

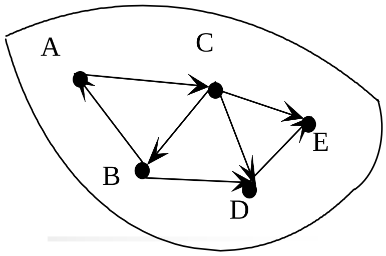




quantitatively by the dependence structures among variables within and across objects

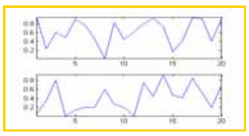
---

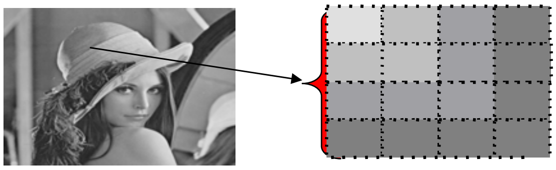
Difficult to learn topology from samples



three special cases studied in literature:

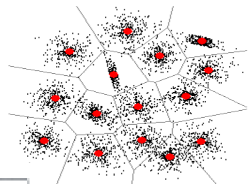
- Given topology: *learn quantitative dependence among variables;*



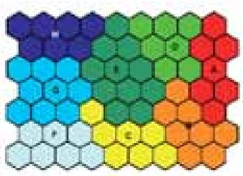


(c) 2-D lattice

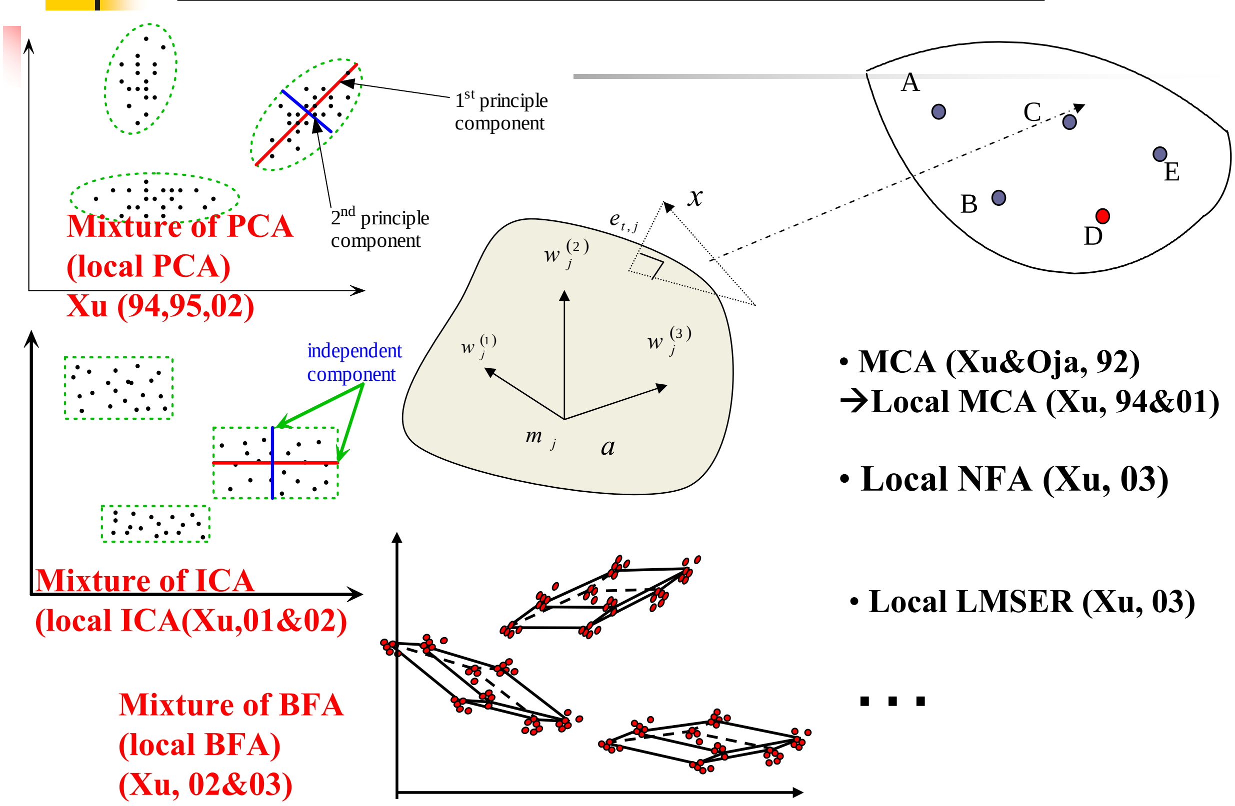
- Null topology: *learn ID and quantitative dependence among variables;*



- Allocation on lattice topology.



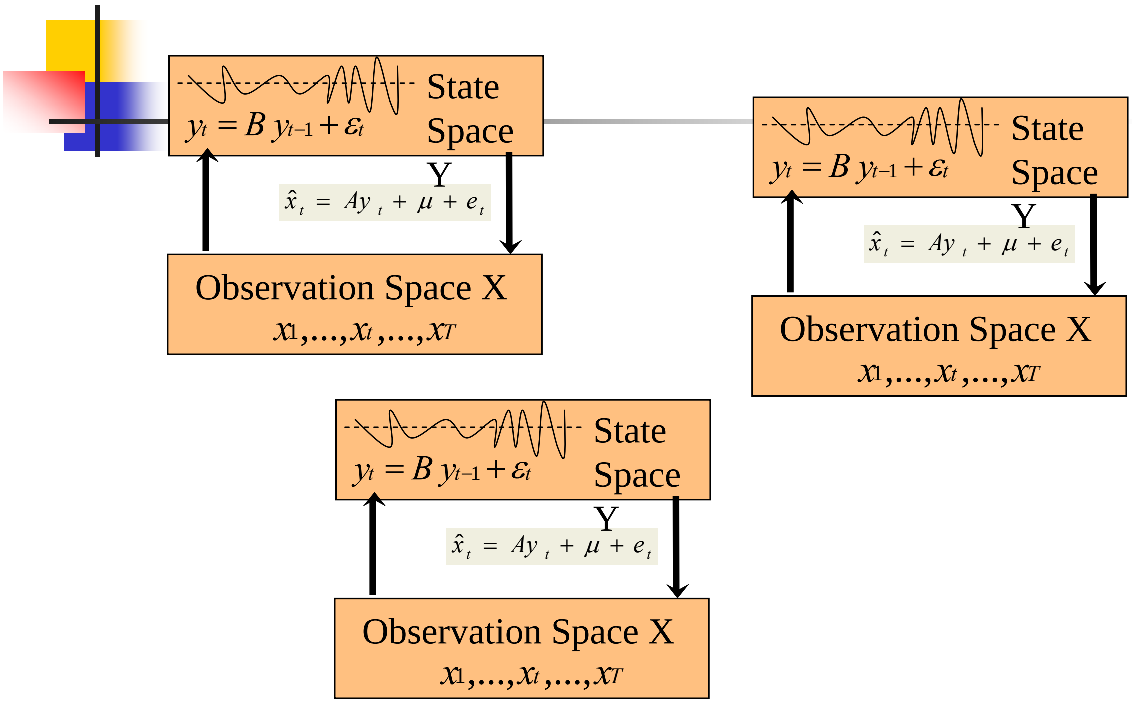
Beyond point structures



- MCA (Xu&Oja, 92) →Local MCA (Xu, 94&01)
- Local NFA (Xu, 03)
- Local LMSER (Xu, 03)

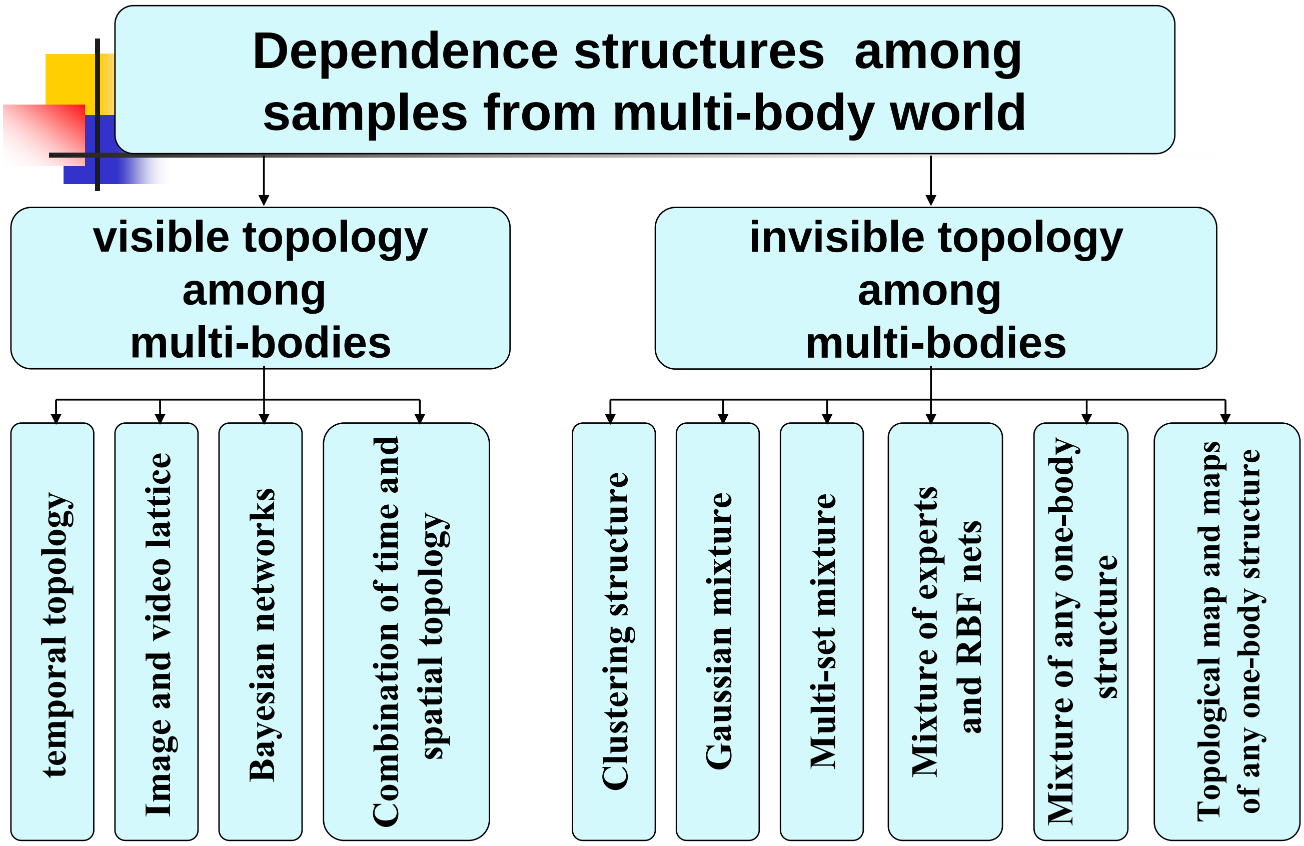
. . .

Mixture of independent state spaces

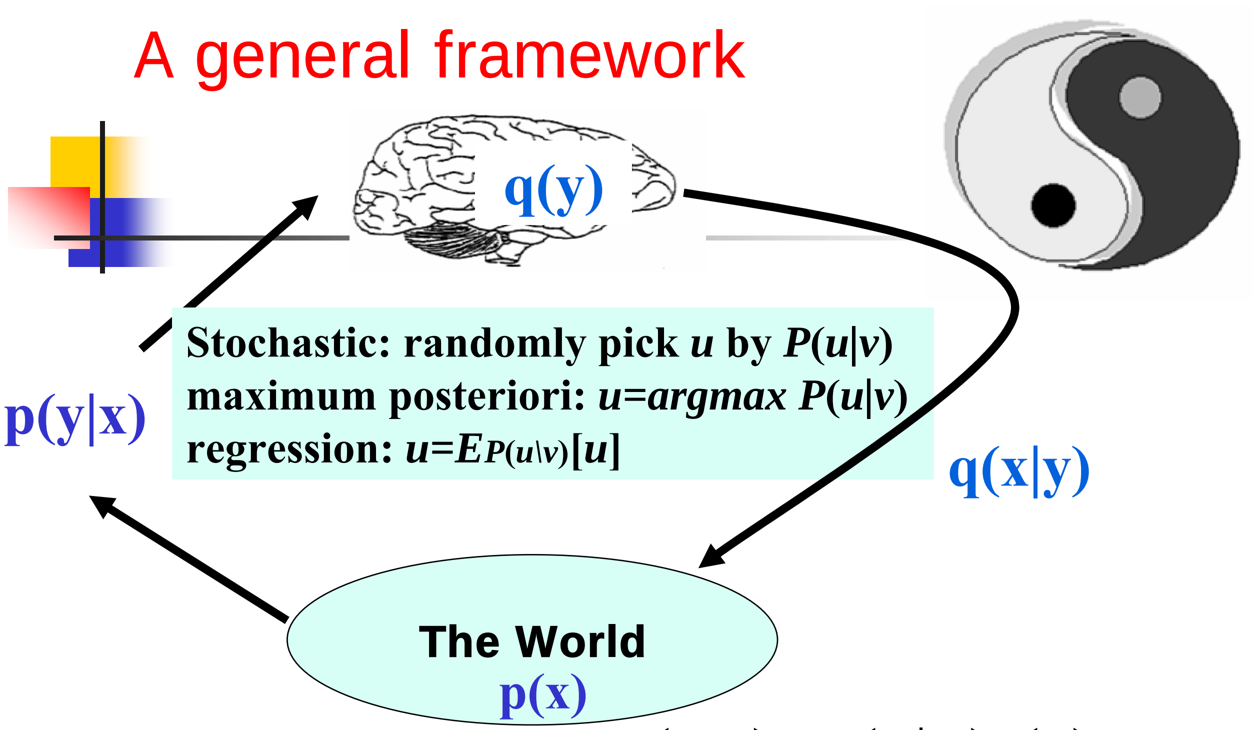


Xu, L. ``Temporal BYY Encoding, Markovian State Spaces, and Space Dimension Determination'', IEEE Tr. Neural Networks, Vol. 15, No. 5, pp1276-1295.

Multi-body world



A general framework

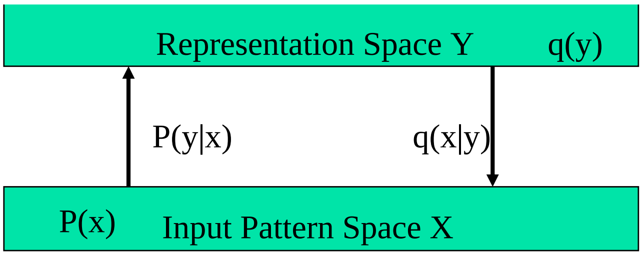


$$P(x,y) = P(y \mid x)P(x) \qquad q(x,y) = q(x \mid y)q(y)$$

YANG representation (Machine) YING representation (Machine)

Bayesian Ying Yang System

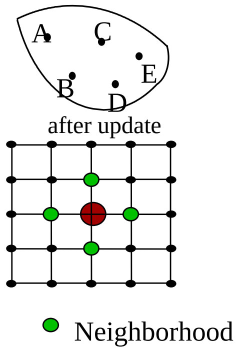
Integrated structures

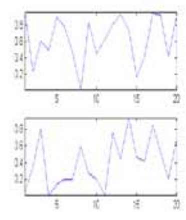


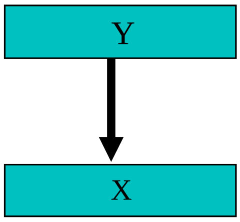
$$(\text{topology}) \times (\text{time}) \times (\text{inner-coding}) \times (\text{observation}) \times (\text{architecture})$$

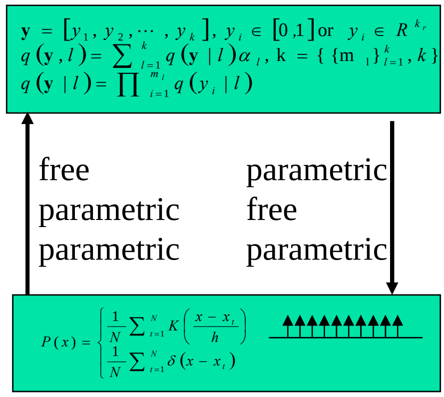
x

one learner	*no*	*no*	*full vector*	*forward*
multi-learner	*yes*	*Gaussian*	*vector in two pairs*	*backward*
learners in lattice		*real nonGaussian*	*multi-parts*	*bi-directional*
learner in tree		*binary*		





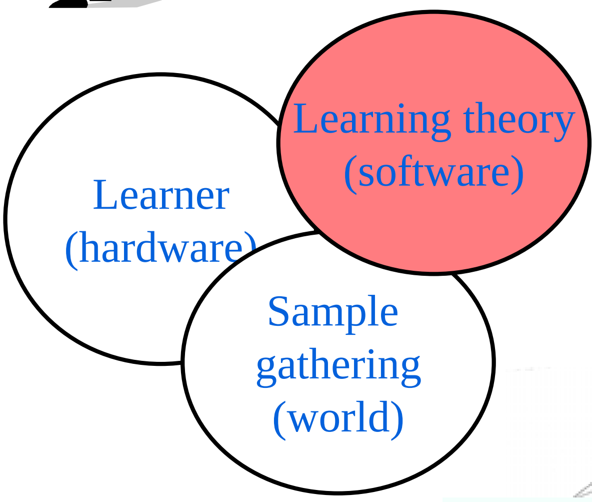


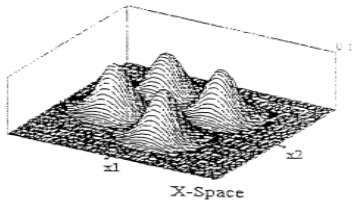


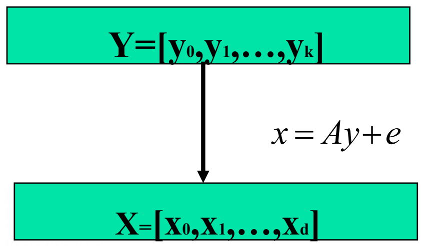
Key Challenge 主要挑战 II

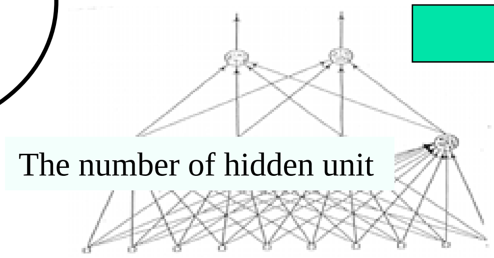
Complexity of Learner's structure →matching the size of samples

(reliable structures of underlying world)









One piece of evidence, take it by 100%

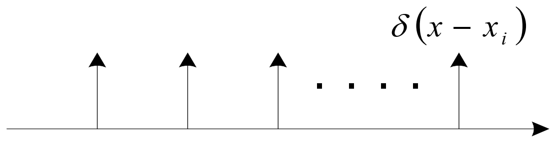
Two pieces of evidence, take each by 50%

…

More pieces of evidence

逐个记忆

The large number law

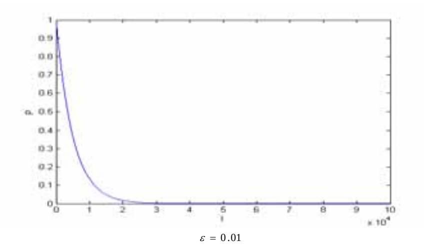


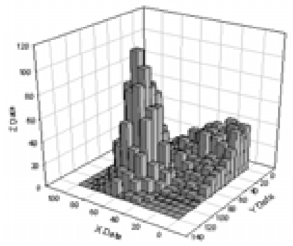
$$q_0(x) = \frac{1}{N}\sum_{t=1}^{N} \delta(x - x_t)$$

$$P\left\{ \sup_{\alpha \in \Lambda} [\, p_*(x \mid \theta^*) - q_0(x)] > \varepsilon \right\} =$$

$$\exp\{-2\varepsilon^2 N\} - 2\sum_{t=2}^{\infty} (-1)^t \exp\{-2\varepsilon^2 t^2 N\}$$

(Kolmogorov & Smirnov, 1930)





Chance of a failed retrieval of a memorized item or getting a wrong memorized item increases exponentially with dimension

Best parametric model matching 参数模型最佳匹配

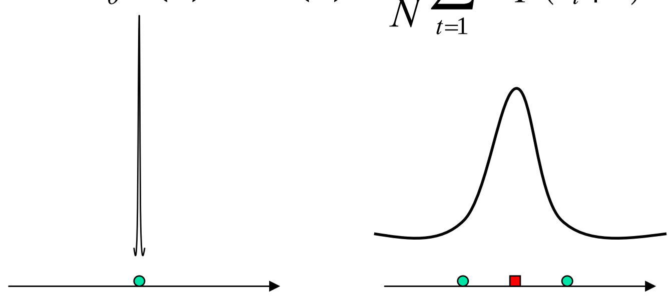
optimizing a matching cost

$$F(p(x \mid \theta), \mathrm{X}), \quad X = \{x_t\}_{t=1}^{N}$$

$$p(x \mid \hat{\theta}(X))$$

e.g., Maximum Likelihood (ML) 最大似然

$$\max_\theta L(\theta), \quad L(\theta) = \frac{1}{N}\sum_{t=1}^{N} \ln p(x_t \mid \theta)$$



One piece of evidence, take it by 100%

Two pieces of evidence take each by 50% subject to the template

More pieces of evidence

… …

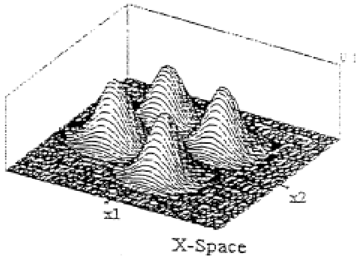
The large number law

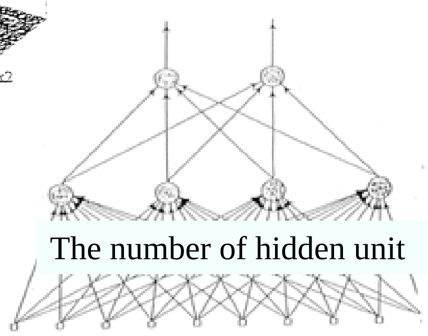
$$p_*(x \mid \theta) \to p_*(x \mid \theta_0) \qquad \hat{\theta}(\mathrm{X}) \to \theta_0 \qquad as \;\; N \to \infty$$

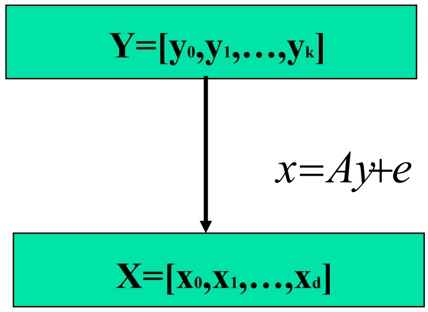
We do not known structure of $p_*(x \mid \cdot)$

a family with same structure but in different scales

$$p(x \mid \theta(k)),\ k = 1,2,\ldots.\infty$$

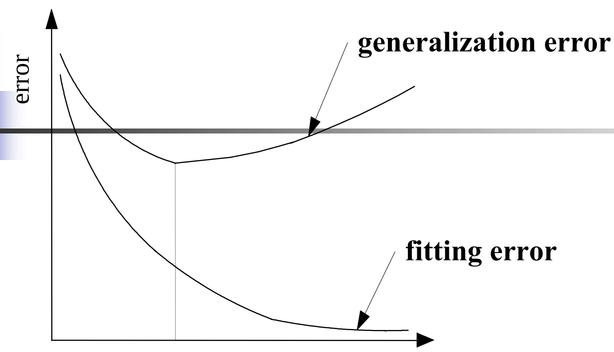


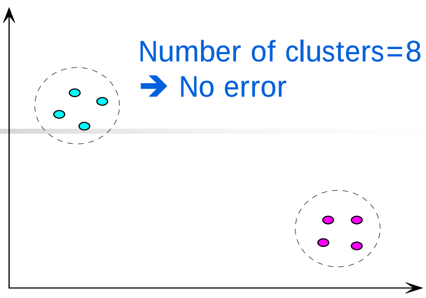


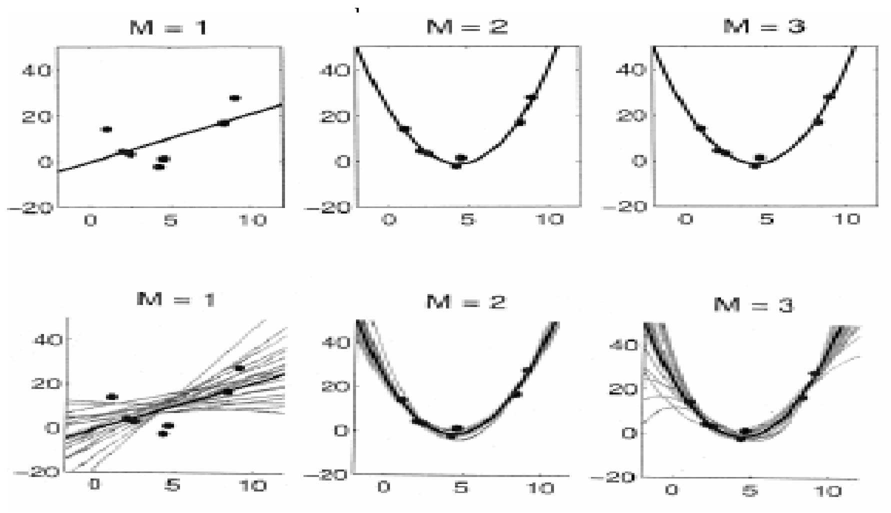


Provide that there is a k^* and $\theta^*(k^*)$ such that $p(x \mid \theta^*(k^*))$ is equal or close to the true $p_*(x \mid \theta_0)$

We do not have $N \to \infty$

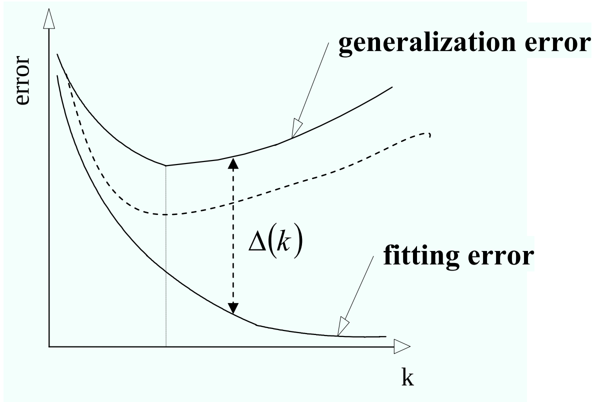






Existing Efforts

- VC Dimension based SRM
- AIC
- BIC, SIC
- Cross Validation
- MML/MDL
- Bayesian Approach



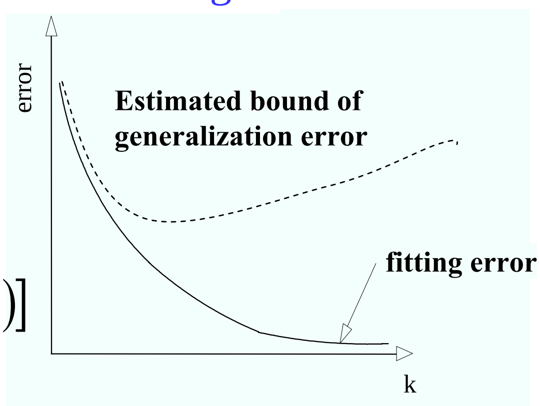
$$k^* = \arg\min_k \left[\Delta(k) + F\left(p(x \mid \theta(k)), \mathrm{X}\right)\right]$$

The existing efforts usually lead to a rough estimate $\Delta(k)$

Two Steps of Solving

Step 1 Enumerate k for a set of candidate values, fixed at each candidate, make learning

$$\theta^*(k) = \arg\min_\theta F\left(p(x \mid \theta), \mathrm{X}\right)$$



Step 2 Select the best one k^* by

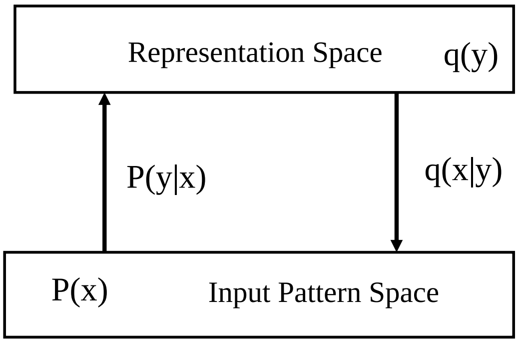
$$k^* = \arg\min_k \left[\Delta(k) + F\left(p(x \mid \theta(k)), \mathrm{X}\right)\right]$$

Very computational extensive !!!

3. Challenges and Advances of Statistical Learning

- Two types of Intelligent Ability: Learning from Samples
- Key Ingredients of Statistical Learning
- Two Key Challenges and Advances on Seeking Solutions
- A Unified Theory: Bayesian Ying-Yang Harmony Learning

Bayesian Ying-Yang Harmony Learning



Basic Learning Principle: Ying-Yang Harmony

(a) Best matching

$p(x,y)=p(y|x)p(x)$ ⟶ Best matching (Least difference) ⟵ $q(x,y)=q(x|y)q(y)$

$$\min KL(p\|q)=\int p(y|x)p(x)\ln\frac{p(y|x)p(x)}{q(x|y)q(y)}dxdy$$ Half job only

(b) The simplest one in complexity or most firm.

$$Max\ H(\theta,k)=H(p\|q)=\int p(y|x)p(x)\ln[q(x|y)q(y)]dxdy-\ln z_q$$

$$H(\theta,k)=H(p\|q)=\int p(y|x)p(x)\ln[q(x|y)q(y)]dxdy-\ln z_q$$

– $p(x)$ is fixed from $\{x_t\}_{t=1}^{N}$ but ***p(x|y)*** is at least not totally fixed.

Least complexity nature $\quad$ $\text{fix } q, \max_p H(p\|q) \Rightarrow p(y\,|\,x)=\delta(y-y_t),\ y_t=f(x_t)$

– pushes ***p(y|x)*** in the least complexity.

Matching nature $\quad$ $\text{fix } p, \max_q H(p\|q) \Rightarrow q_t = p_t$

– ***pushes q(x|y), q(y)*** in the least complexity also.

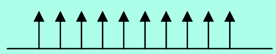
• *Therefore, we have* $\max_{\theta,k} H(\theta,k) \Rightarrow \begin{cases} \textit{parameter learning} \\ \textit{model selection} \end{cases}$

Parameter Learning with Automated Model Selection

$$q(\mathbf{y},l)=\sum_{l=1}^{k} q(\mathbf{y}\,|\,l)\alpha_l,\ \mathrm{k}=\{\{\mathrm{m}_l\}_{l=1}^{k},k\}$$
$$q(\mathbf{y}\,|\,l)=\prod_{i=1}^{m_l} q(y_i\,|\,l)$$

p(y, l |x) $\qquad$ q(x|y,l)

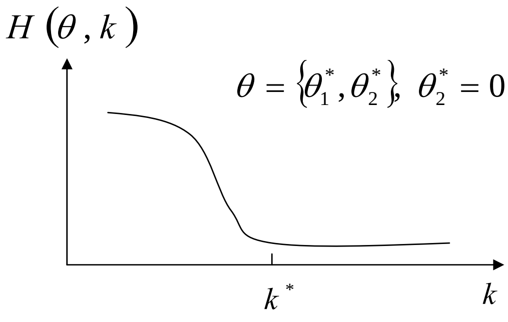
$$P(x)=\begin{cases}\frac{1}{N}\sum_{t=1}^{N}K\left(\frac{x-x_t}{h}\right)\\ \frac{1}{N}\sum_{t=1}^{N}\delta(x-x_t)\end{cases}$$



- Set some $\alpha_l = 0$ is equivalent to reduce $k \Rightarrow k$-1
- Set the variance of $q(y_i\,|\,l)$ to be 0 is equivalent to reduce $m_l \Rightarrow m_l$-1

k fixed at large value.

$$\underset{\theta}{Max}\ H(p\|q)$$

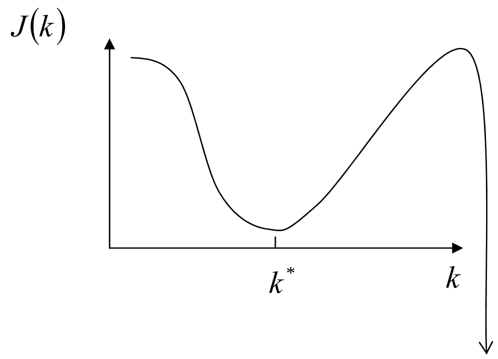


Parameter Learning Followed By Model Selection

Step 1 Enumerate **k**, at each **k**, make parameter learning

$$\max_{\theta} H(\theta, k),$$

Step 2 $k^* = \arg\min_k J(k),\quad J(k) = -H(\theta^*, k)$



Alternatively, parameter learning can also be made via $\min_{\theta} KL(p\|q)$

$$KL(p\|q) = \int_{x,y} p(y|x)p(x)\ln\frac{p(y|x)p(x)}{q(x|y)q(y)}dxdy$$

Also act as a general scheme that integrates:

Parameter Learning

Model selection

Regularization

Better performances in the cases of a small size of samples

Ying Yang Alternative Minimization

$\underset{\theta}{Max}\, H(p\|q)$ can be further implemented alternatively by

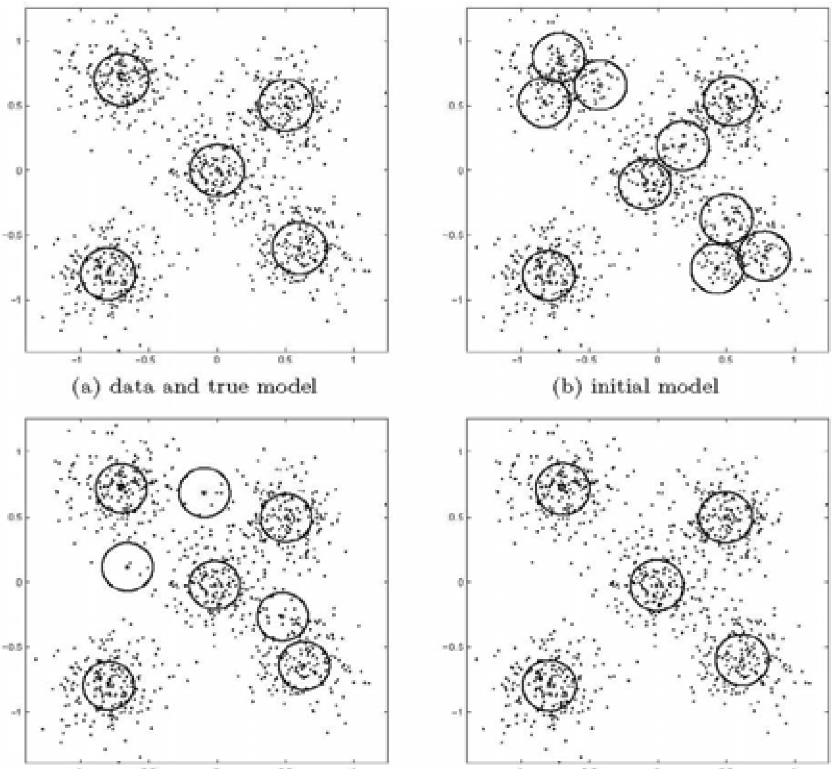
- $\text{Yang Step: Fix } q(x,y)$
  $\text{get } p(x,y) = \arg\max_{p(x,y)} H(p\|q)$

- $\text{Ying Step: Fix } p(x,y)$
  $\text{get } q(x,y) = \arg\max_{q(x,y)} H(p\|q)$

It will converge to a local maximum of ***H(p||q)***

It also works when ***H(p||q)*** is replaced by ***KL(p||q)***

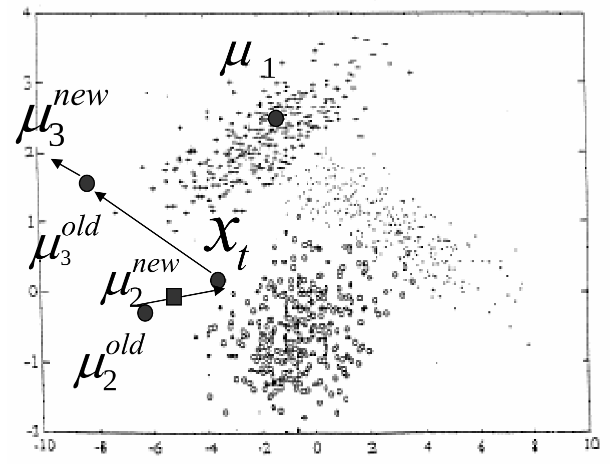
The well know Expectation-Maximization (EM) is its special case

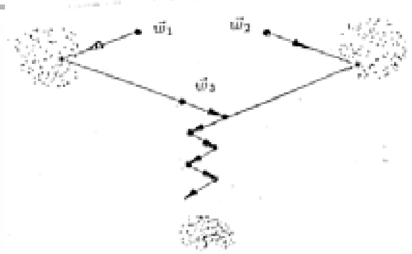


(a) data and true model (b) initial model (c) result after 50 iterations (d) final result

(a) there are five classes with each class consisting of 200 Gaussian samples.
(b) the initialize value of k is 10.
(c) after 50 iterations of implementing the best harmony learning
(d) a correct k=5 is determined after learning has converged.

Ying-Yang in a local alternative perspective:
Rival Penalized Competition





$$
\mu_j^{new} = \begin{cases} \mu_j^{old} + \eta_w (x_t - \mu_j^{old}), & \text{if } j = l_w \\ \mu_j^{old} - \eta_r (x_t - \mu_j^{old}), & \text{if } j = l_r \\ \mu_j^{old}, & \text{otherwise} \end{cases}
$$

$$0 < \eta_r << \eta_w$$

Rival Penalized Competitive Learning

(Xu, Krzyak, Oja, 91&93)

Listed in the following table are the results of 100 experiments on a Gaussian mixture with k=5, in comparison with three typical model selection criteria AIC, CAIC, BIC/MDL. Experiments were made by considering the ball shape 2x2 covariance matrix , the elliptic 2x2 covariance matrix , and the ball shape 10x10 covariance matrix , in different sizes n of samples. In this table, S denotes the rate of successes, O denotes over-estimated values of k, and U denotes under-estimated values of k. It can be clearly observed that the above J(k) (i.e., BYY-HDS) outperforms others considerably.

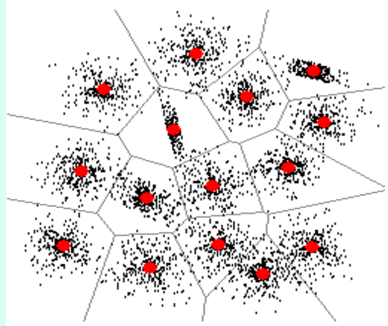
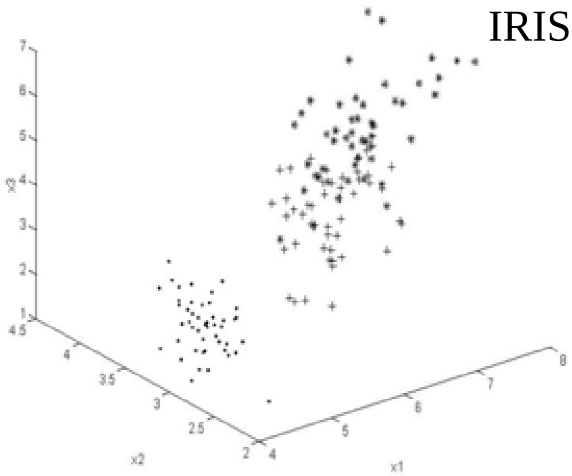
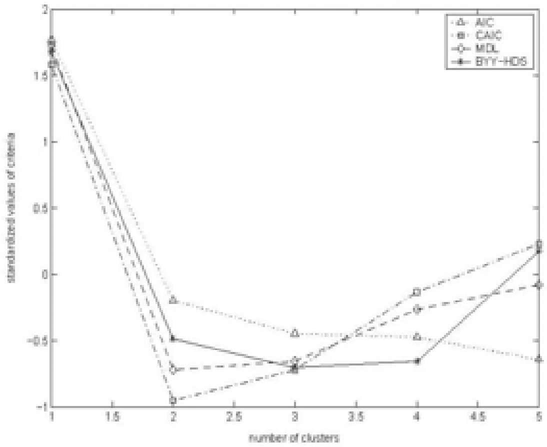


Table 1. Rates of underestimating (U), success (S), and overestimating (O) by each criteria on the simulation data sets in 100 replications

Example	Sample size	AIC			CAIC			MDL			BYY-HDS		
		U	S	O	U	S	O	U	S	O	U	S	O
Spherical	80	0	26	74	69	31	0	48	52	0	11	76	13
	200	0	48	52	16	79	5	12	85	3	6	84	10
	400	0	43	57	12	87	1	8	90	2	5	88	7
Elliptic	100	0	21	79	87	13	0	82	18	0	16	61	23
	250	0	34	66	69	31	0	57	43	0	14	59	27
	500	0	23	77	41	59	0	37	62	1	12	69	19
High Dimensional	100	0	27	73	39	48	13	25	51	24	23	55	22
	500	0	45	55	32	57	11	27	60	13	17	71	12
	1000	0	47	53	10	76	14	8	81	11	8	84	8
Average		**0**	**34.9**	**65.1**	**41.7**	**53.4**	**4.9**	**33.8**	**60.2**	**6.0**	**12.4**	**71.9**	**15.7**

IRIS Data





on the well known IRIS real data set of 150 samples from three classes (i.e., iris species setosa, versicolor, virginica) and each sample having four dimensions (i.e., sepal length, sepal width, petal length, and petal width). Again, the top figure shows the projections of the data set on the first three dimensions. The bottom figure gives the results of selection. It can be observed that only the above J(k) (i.e., BYY-HDS) has successfully determined the correct k=5, while AIC output a higher on k=5 but CAIC and both MDL output the wrong one k=2.

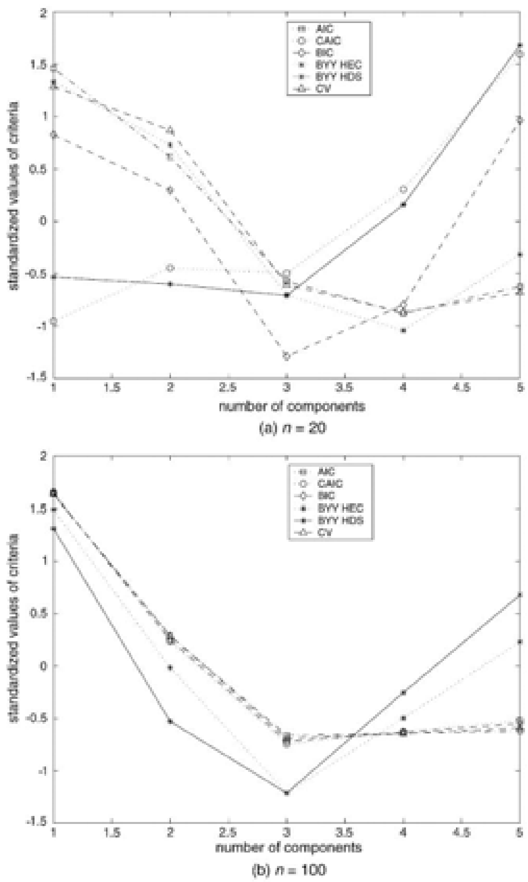
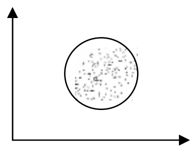
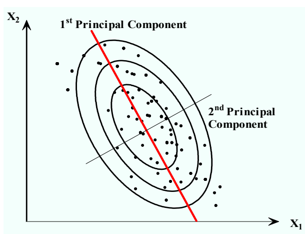


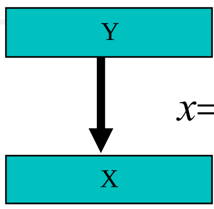
Fig. 1. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the data sets of a 10-dimensional x ($d = 10$) generated from a 3-dimensional y ($k = 3$) with different sample sizes. (a) $n = 20$ and (b) $n = 100$.

Factor analysis (FA)



k=3





$x = Ay + e$

d=10

Table 1
Rates of underestimating (U), success (S), and overestimating (O) by each criteria on simulation data sets with different sample sizes in 100 experiments

Criteria	$n = 20$			$n = 40$			$n = 100$		
	U	S	O	U	S	O	U	S	O
AIC	2	68	30	0	81	19	0	85	15
CAIC	26	73	1	2	98	0	0	100	0
BIC	10	84	6	1	99	0	0	100	0
BYY-HEC	6	74	20	0	98	2	0	100	0
BYY-HDS	11	86	3	1	99	0	0	100	0
10-Fold CV	3	71	26	0	87	13	0	92	8

Summary

- BYY system as a general framework that integrates typical structures for statistical learning
- BYY system + Kullback divergence KL(p||q) a unified perspective for maximum likelihood learning on various structures
- BYY system + Best harmony H(p||q) a new theory with a new mechanism for automatic model selection during parameter learning no need on two stage implementation
- BYY system + Best harmony + regularization further improve performances in the cases of a small size of samples.
- A natural perspective of alternative minimization algorithms

---

- Firstly proposed in 1995

Xu, L (1996), Advances in NIPS 8, 444-450 (1996). A part of its preliminary version on Proc. ICONIP95-Peking, 977-988(1995).

- Developed in past years (see recent papers below)

Xu, L. (2004), ``Temporal BYY Encoding, Markovian State Spaces, and Space Dimension Determination", IEEE Tr. Neural Networks, Vol. 15, No. 5, pp1276-1295.

Xu, L (2004), ``Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor auto-determination", IEEE Tr. Neural Networks, Vol. 15, No. 5, pp885-902 .

Xu, L (2003), ``Data smoothing regularization, multi-sets-learning, and problem solving strategies", Neural Networks, Vol. 15, Nos. 5-6, 817-825.

Xu, L (2003), `` BYY learning, regularized implementation, and model selection on modular networks with one hidden layer of binary units", *Neurocomputing* Vol. 51, 277-301.

Xu, L (2002), ``BYY harmony learning, structural RPCL, and topological self-organizing on mixture models", *Neural Networks*, Vol.15, Nos. 8-9, 1125-1151.

Xu, L (2001), ``BYY harmony learning, independent state space and generalized APT financial analyses ", IEEE Tr. on Neural Networks, 12 (4), 822-849.

Xu, L (2001), ``Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models", Intl J. of Neural Systems, 11(1), 43-69.

Xu, L (2000), ``Temporal BYY learning for state space approach, hidden Markov model and blind source separation", IEEE Tr. on Signal Processing 48, 2132-2144.

Relations to and Key differences from approaches below

- Maximum likelihood
- Information geometry
- Helmholtz machines
- Variational approximation
- Minimum description length (MDL)
- Bit-back based MDL
- Bayesian approach
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

Xu, L (2004), ``Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor auto-determination", IEEE Tr. Neural Networks, Vol. 15, No. 5, pp885-902 .。

For more details, see: **http://www.cse.cuhk.edu.hk/~lxu/**

其它工作

二十餘年來在模式識別、人工智能、信號處理、統計學習及統一理論等多個重要研究方向，不僅在理論方法方面且在技術應用方面都做出了若干開創性工作。

•發表學術期刊論文近百篇 (國際學術期刊上70餘篇，《中國科學》和《科學通报》上4篇)，還在主要國際出版社的編輯書中貢獻20餘篇,并發表了大量國際會議論文。

According to SCI-EXPANDED, his papers got over 1400 citations, and his 10 most cited papers scored near 850. Among them, one single his paper scored 275, each of the other nine papers are scored between 43—96.

According to Google Scholar, his papers scored over 1800 citations. The 10 most cited papers scored near 1200. Among them, one single paper scored 416, each of other nine papers are scored between 55—131.

By CiteSeer, ranked at the 2061-th among 10,000 most cited authors of 773109authors.

•還被國外30餘本學術專著或教科書中引用。

•應邀在國際主要學術大會做大會報告/特邀報告/學術講座40餘次。