

# Advanced Image Processing

---

## Part IV: Image Modeling

S. Voloshynovskiy



# Course Outline

---

- Recall of Linear Algebra.
- Introduction. Human Visual System.
- Image Representation: pyramids and wavelets.
- Random Signals.
- Image Modeling.
  - Image Sensor Models. Noise Models.
  - Image Denoising.
  - Image Restoration.
  - Image Compression.
  - Video Modeling and Compression.
  - Digital Data Hiding.

# Recommended books

---

- A. K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, 1989.
- A. Bovik, Handbook of Image & Video Processing, Academic Press, 2000.

# Roadmap:

---

1. Model Validation
2. Prediction and Autoregressive Models
3. Markov Processes
4. Markov Random Fields and Smoothness
5. Doubly Stochastic Processes
6. Karhunen-Loève Expansions
7. Image Modeling: Fundamental Links
8. Transform Domain Models

# 1. Image modeling

---

Some difficulties involved in image modeling:

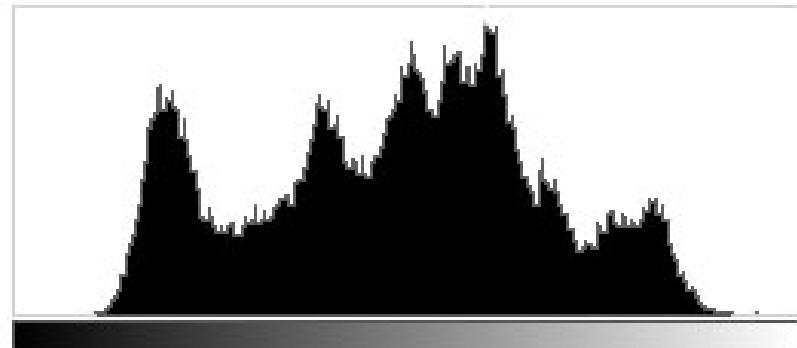
1. Define appropriate sampling space.
2. Very high dimensionality of images.
3. Model should be tractable (computationally).
4. Model should be robust (applies to many images).
5. Model Validation: Given two competing models P1 and P2, which one is better?

# 1. Model Validation

---

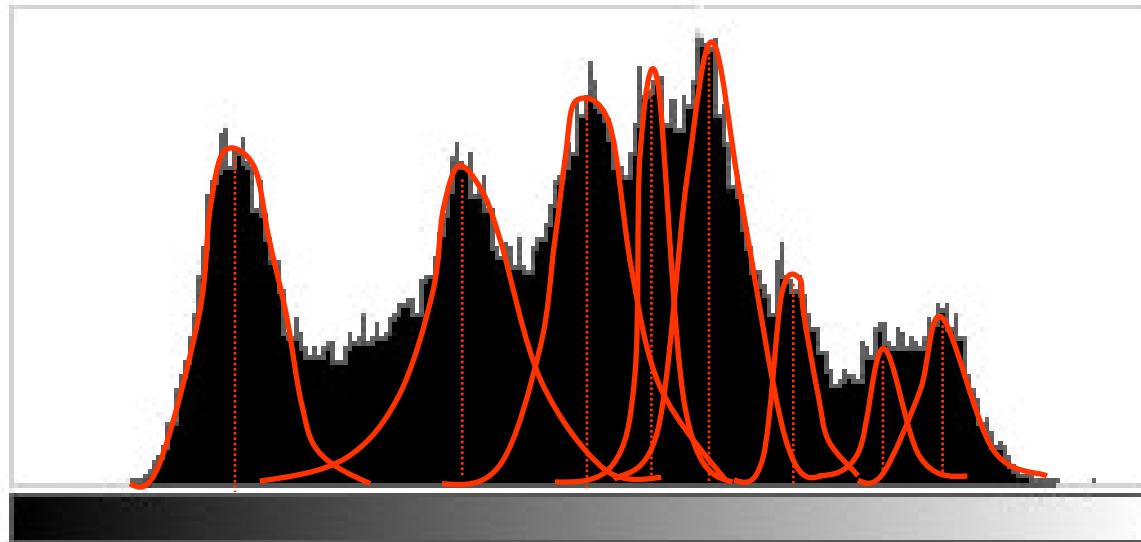


Histogram



# 1. Is image pdf Gaussian?

No, but can be approximated by mixture of Gaussians.



$m_{x_1}$

$m_{x_2}$

$m_{x_3}$

$m_{x_4}$

$m_{x_5}$

$m_{x_6}$

$m_{x_7}$

$m_{x_8}$

Number of coefficients  
in ith model

$$p_X(x) = \sum_{k=1}^M \alpha_i N(m_{x_i}, R_{x_i})$$

$$\alpha_i = \frac{N_i}{\sum_{j=1}^M N_j}$$

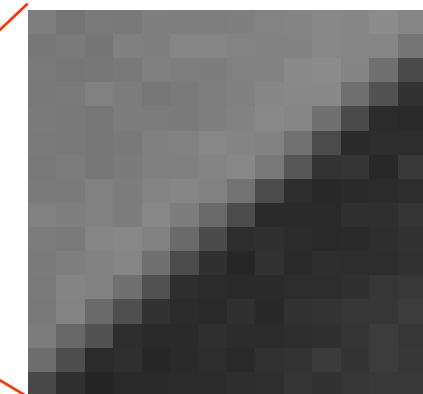
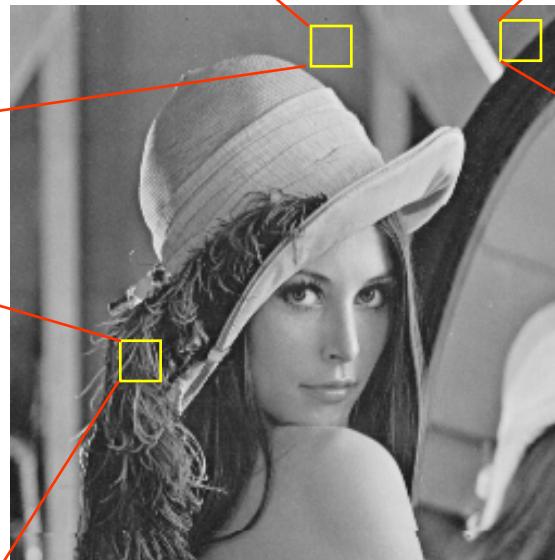
- Probability of model  
Total number of coefficients  
in all models

# 1. Multimode mechanisms: edges and textures

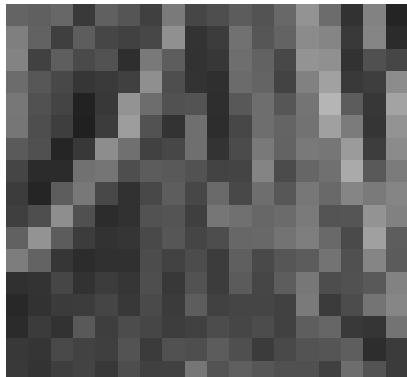
---



Flat region



Edges

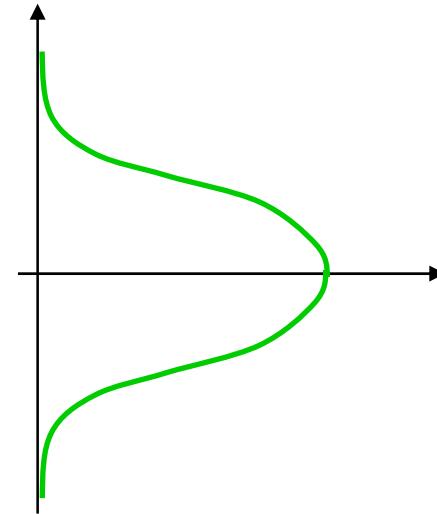
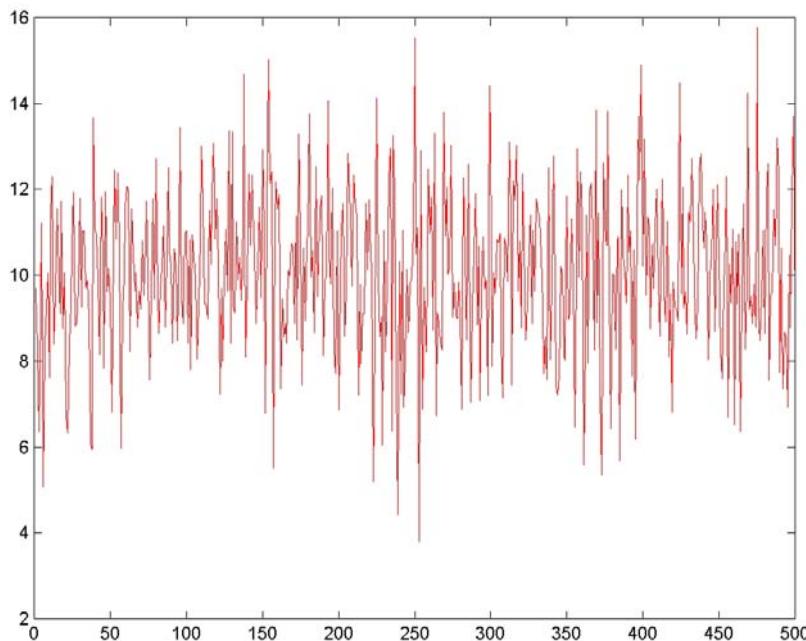


Textures

# 1. Multimode mechanisms: edges and textures

---

Flat region: 1D model:  $X \sim N(m_x, \sigma_x^2) = N(10, 4)$

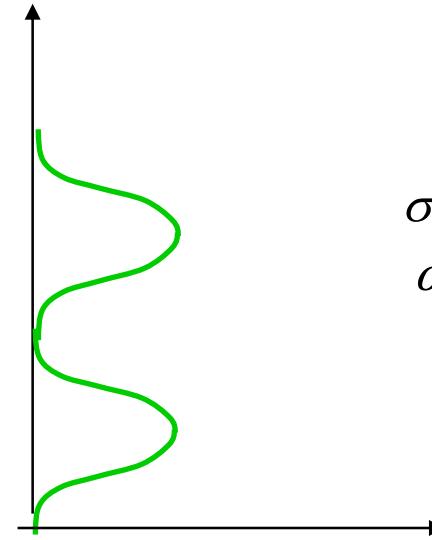
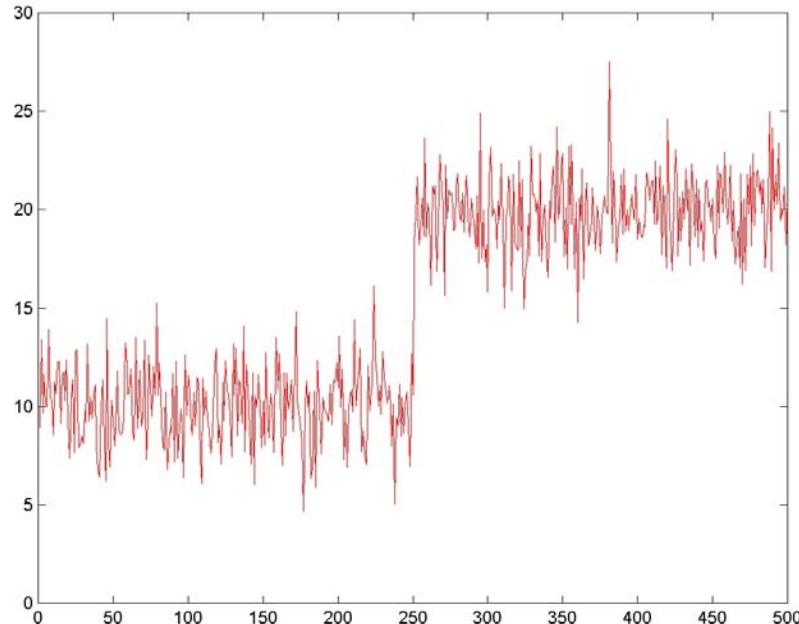


# 1. Multimode mechanisms: edges and textures

---

Edge: 1D model:

$$X \propto (1 - \alpha)N(m_{x_1}, \sigma_{x_1}^2) + \alpha N(m_{x_2}, \sigma_{x_2}^2) = (1 - 0.5)N(10, 4) + 0.5N(20, 4)$$



$$\begin{aligned}\sigma_{x_2}^2 &= \sigma_{x_1}^2 = 4 \\ \alpha &= 0.5\end{aligned}$$

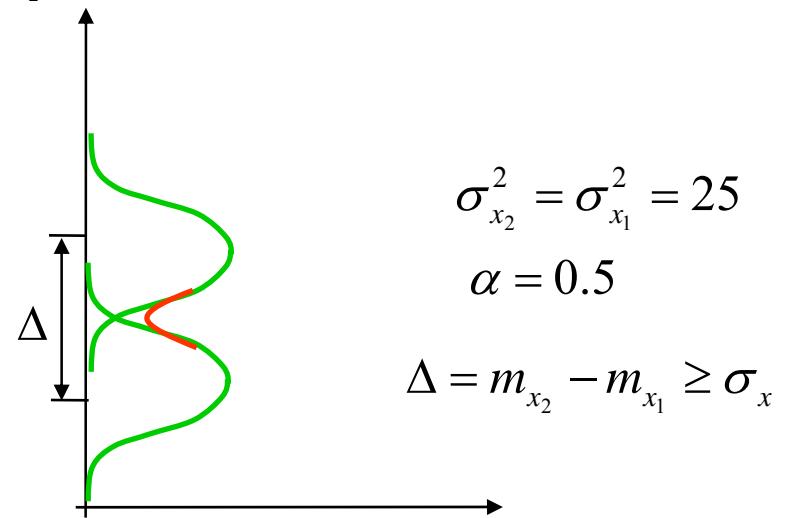
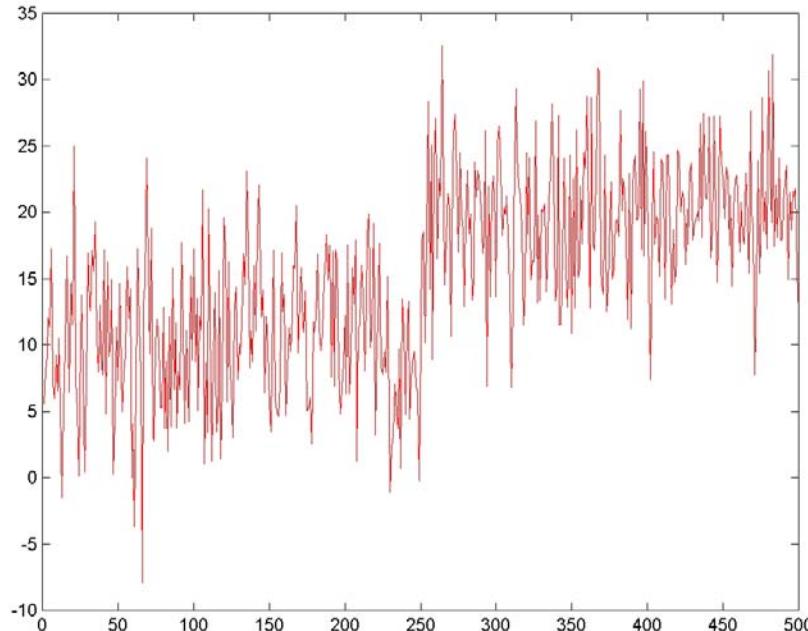
... increase the variance.

# 1. Multimode mechanisms: edges and textures

---

Edge or texture?: 1D model:

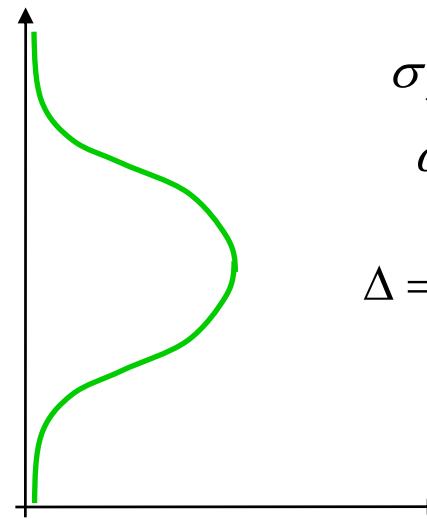
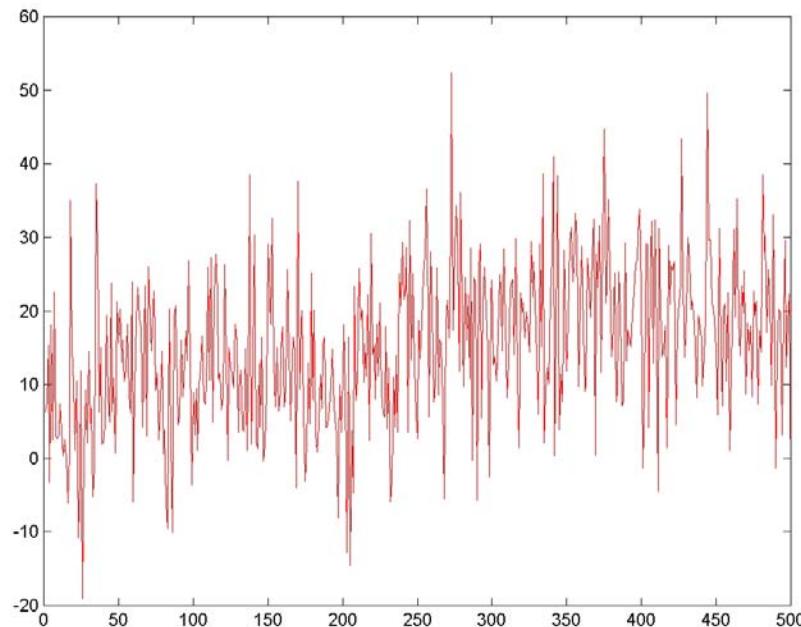
$$X \propto (1 - \alpha)N(m_{x_1}, \sigma_{x_1}^2) + \alpha N(m_{x_2}, \sigma_{x_2}^2) = (1 - 0.5)N(10, 25) + 0.5N(20, 25)$$



# 1. Multimode mechanisms: edges and textures

Edge is transformed into flat region: 1D model:

$$X \propto (1 - \alpha)N(m_{x_1}, \sigma_{x_1}^2) + \alpha N(m_{x_2}, \sigma_{x_2}^2) = (1 - 0.5)N(10, 100) + 0.5N(20, 100)$$



$$\sigma_{x_2}^2 = \sigma_{x_1}^2 = 100$$
$$\alpha = 0.5$$

$$\Delta = m_{x_2} - m_{x_1} \leq \sigma_x$$

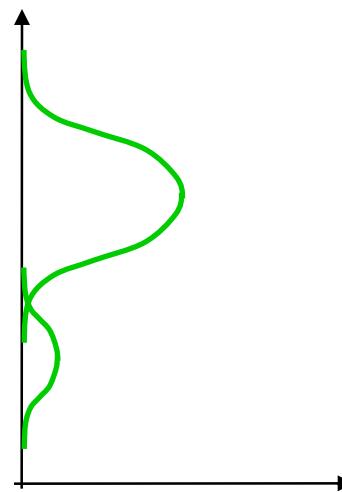
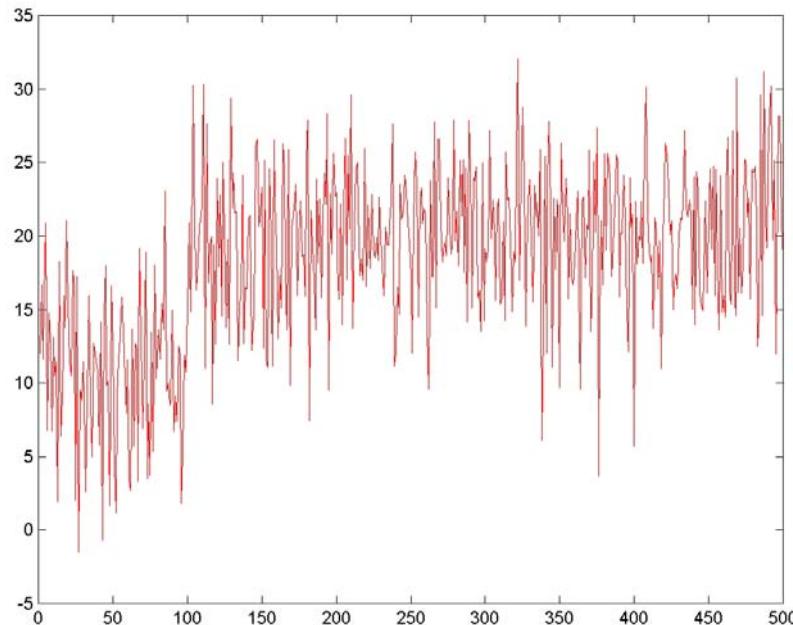
Mathematically: two processes with different means: observation: unimodal

# 1. Multimode mechanisms: edges and textures

---

Edge: 1D model:

$$X \sim (1 - \alpha)N(m_{x_1}, \sigma_{x_1}^2) + \alpha N(m_{x_2}, \sigma_{x_2}^2) = (1 - 0.8)N(10, 25) + 0.8N(20, 25)$$



$$\sigma_{x_2}^2 = \sigma_{x_1}^2 = 25$$
$$\alpha = 0.8$$

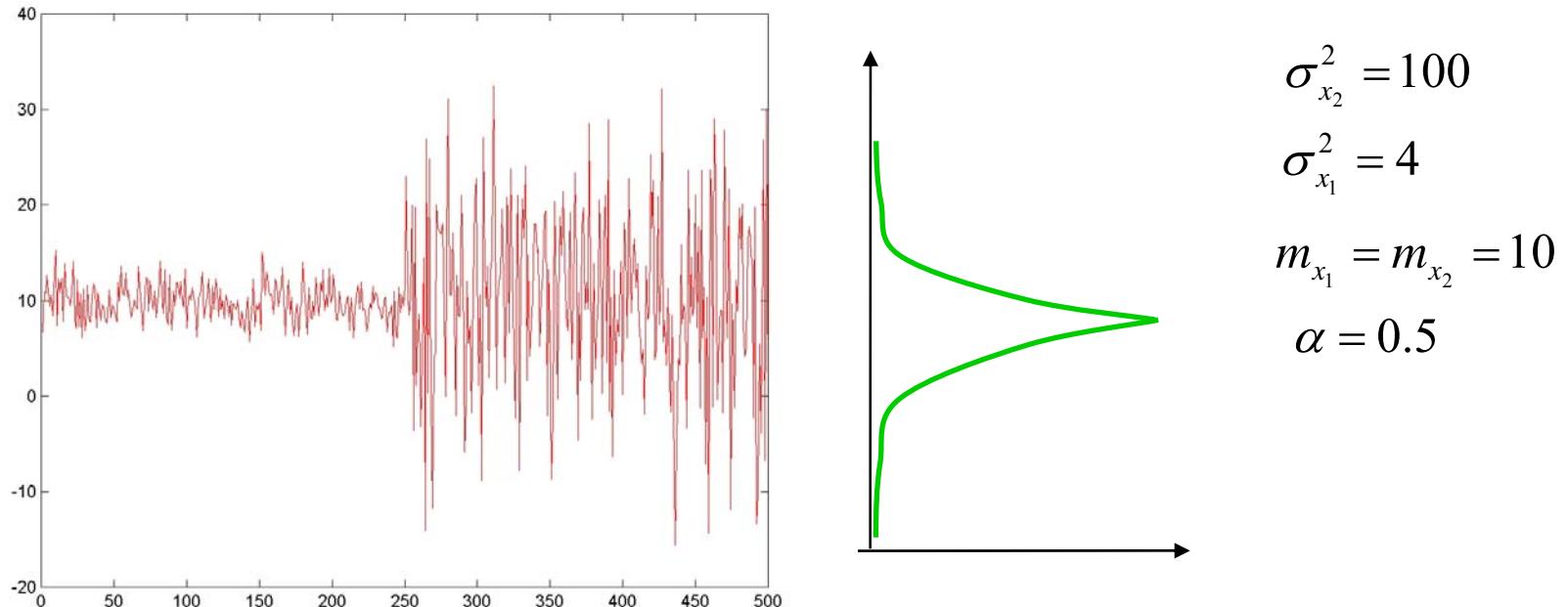
Asymmetric pdf.

# 1. Multimode mechanisms: edges and textures

---

Texture: 1D model:

$$X \propto (1 - \alpha)N(m_{x_1}, \sigma_{x_1}^2) + \alpha N(m_{x_2}, \sigma_{x_2}^2) = (1 - 0.5)N(10, 4) + 0.5N(10, 100)$$



Process is created by two Gaussians.

Does it look like Gaussian?

# 1. Multimode mechanisms: edges and textures

---

Mixture Gaussian model is quite simple.

We considered only a simple edge model and....

how many parameters should be estimated:

Variances:

$$\sigma_{x_1}^2$$

$$\sigma_{x_2}^2$$

Means:

$$m_{x_1}$$

$$m_{x_2}$$

Probability of model appearance:

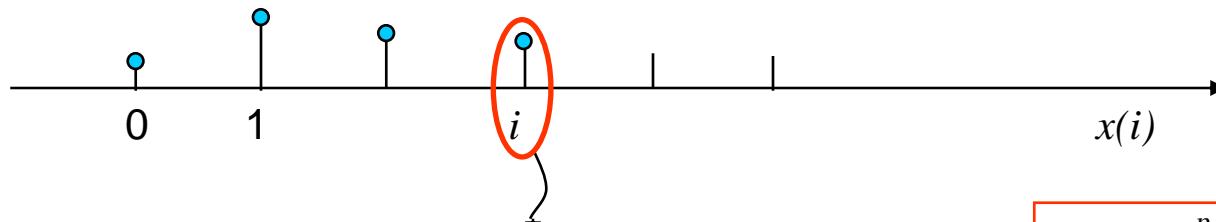
$$\alpha$$

..more complex compound images require more parameters to be estimated.

## 2. Prediction and Autoregressive Models (Jain, Ch.6)

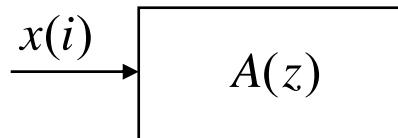
- Simple and tractable image models.
- Intuitive concept due to apparent high correlation of neighboring pixel values.

**Predictive models** (1D) [will also use for video]



① Predict value of  $x(i)$  using

$$\hat{x}(i) = \sum_{k=1}^p a_k x(i-k)$$

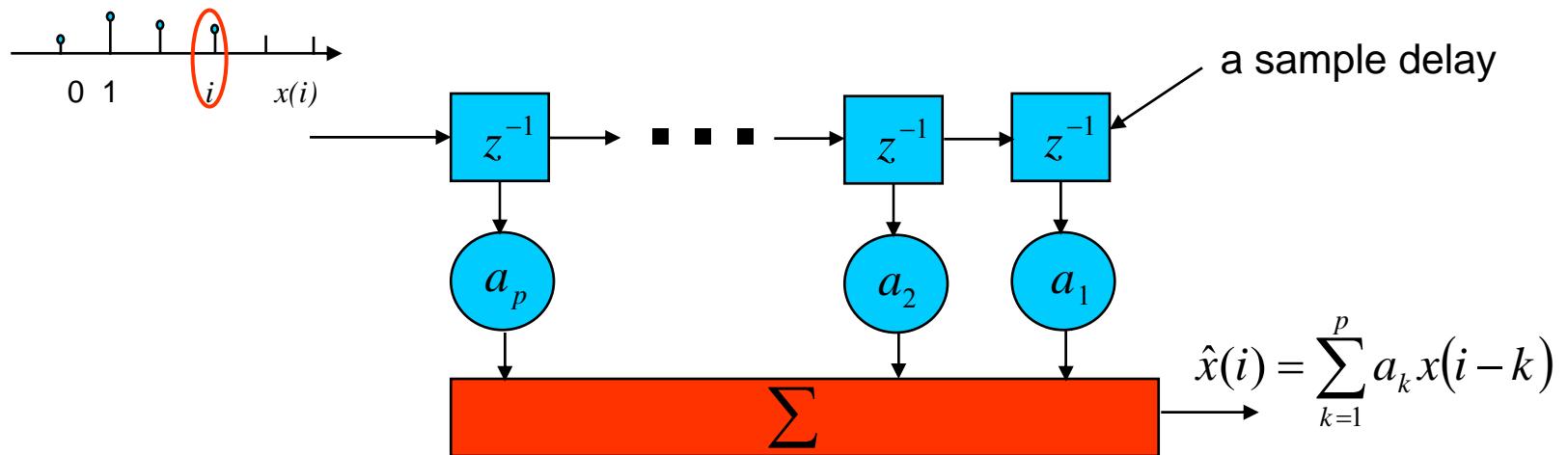


Causal prediction filter:

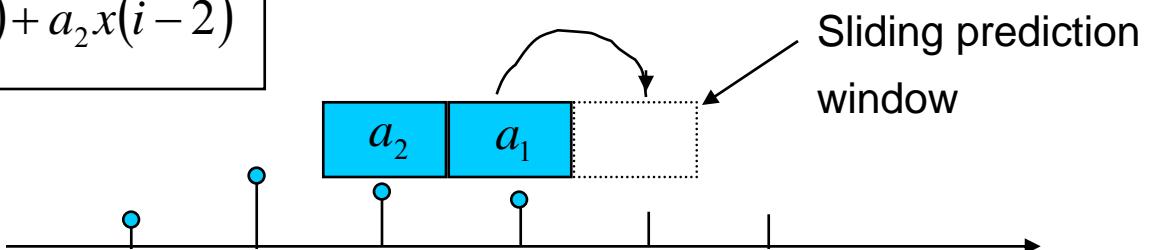
$$A(z) = \sum_{k=1}^p a_k Z^{-k}$$

## 2. Prediction and Autoregressive Models

---



$$p = 2 : \quad \hat{x}(i) = a_1 x(i-1) + a_2 x(i-2)$$

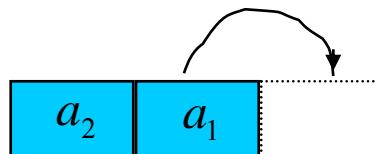


## 2. Prediction and Autoregressive Models

---

- ② Need model for causal filter coefficients  $\{a_k\}$

- local correlation between samples
- smoothness



Intuition: since  $a_1$  is closer to  $x(i)$  it is more likely to be of the same or close pixel value.

Therefore,  $a_1 \rightarrow 1$ .

The higher index  $k$  of filter coefficients  $\{a_k\}$ , the lower correlation with the value of pixel to be predicted.

(intuition: consider  $A(z)$  as a lowpass filer )

## 2. Prediction and Autoregressive Models

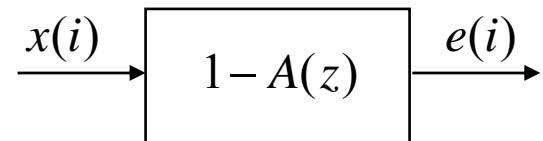
- ③ Need model for residual error of prediction:

$$e(i) = x(i) - \hat{x}(i)$$

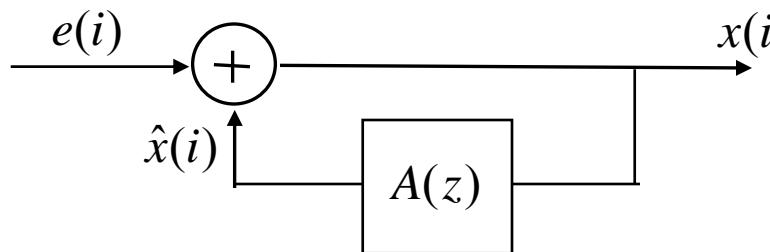
The goal of stochastic image modeling is to receive  
a very simple model for the residual error (**white noise as optimal case**).



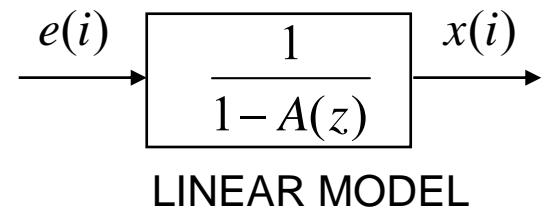
Assume  $\{e(i)\}$  = white noise sequence.



AR Model of order  $p$ :



Excitation sequence



## 2. Prediction and Autoregressive Models

---

AR Model of order  $p$ :

$$x(i) = \sum_{k=1}^p a_k x(i-k) + e(i)$$

Parameters of AR model:

① Transfer function:  $\frac{1}{1 - A(z)}$

Assume poles are in unit circle

② White noise excitation  $\{e(i)\}$

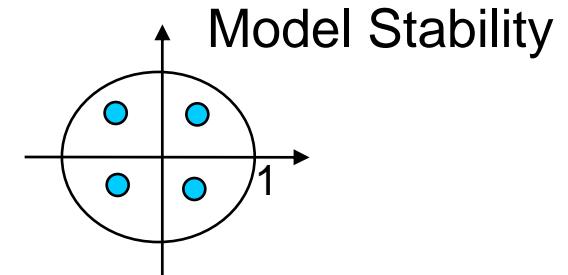
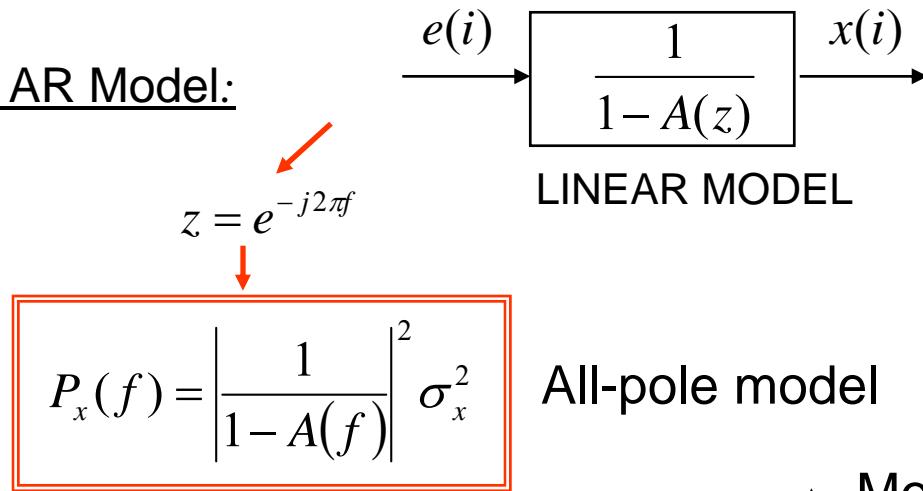
$$\begin{cases} E[e(i)] = 0 \\ E[e(i)e(n)] = \sigma_x^2 \delta_{in} \end{cases}$$

$1 - A(z)$  whitening filter

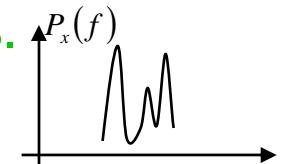
$$x(i) = \bar{x}(i) + e(i) \quad (\text{reminder: pyramid and wavelet image presentation})$$

## 2. Prediction and Autoregressive Models

Spectral density of AR Model:

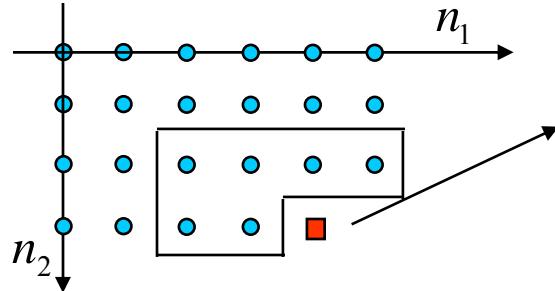


Model is well suited to model peaks, but without flat valleys.



## 2. 2D AR-Models

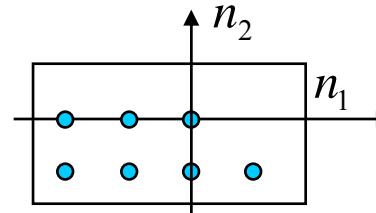
---



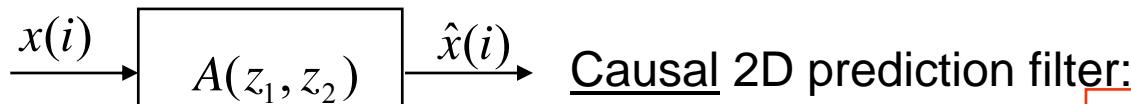
Predictor: causal filter in 2D

$$x(i) = \sum_{k \in M} a_k x(i - k)$$

$M$  is a mask



- Linear!
- Space invariant!

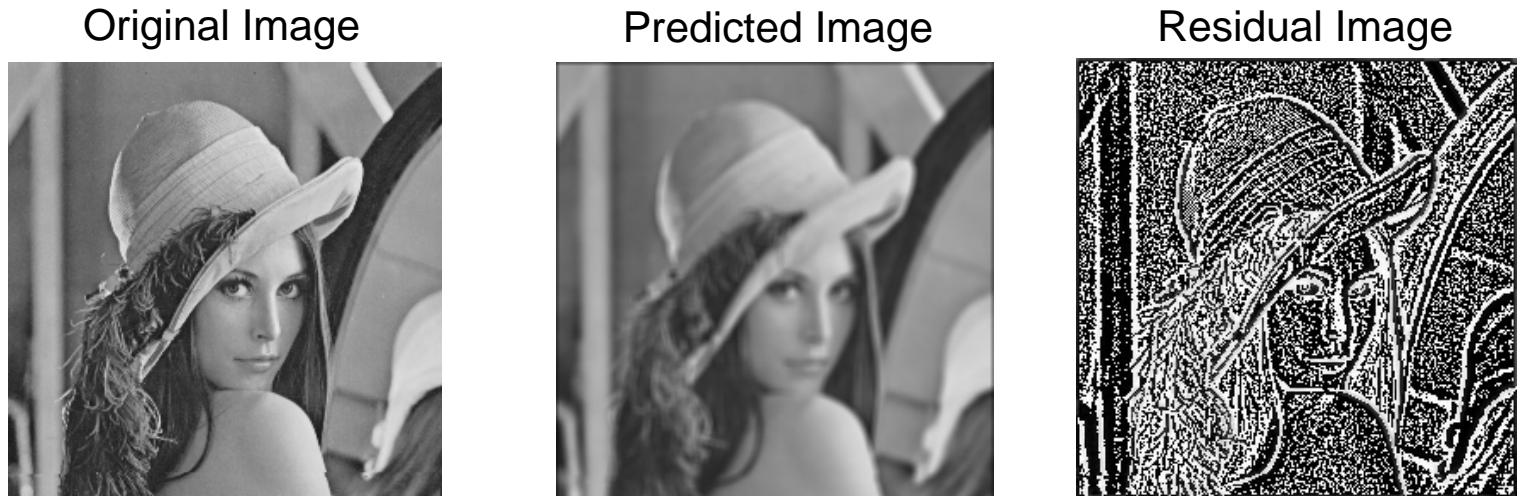


Causal 2D prediction filter:

$$A(z_1, z_2) = \sum_{k_1} \sum_{k_2} a_{k_1 k_2} Z_1^{-k_1} Z_2^{-k_2}$$

## 2. Image Modeling based on AR-Model

---



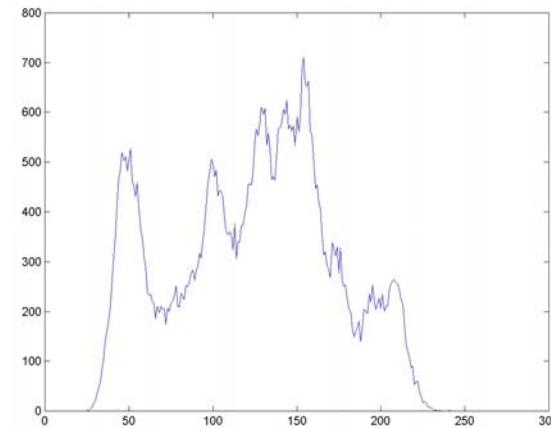
$$x(i) = \bar{x}(i) + e(i)$$

denotes prediction based on causal window.

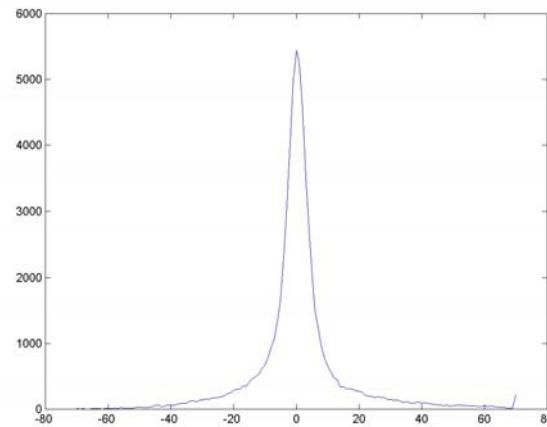
## 2. Image Modeling based on AR-Model

---

Original Image



Residual Image



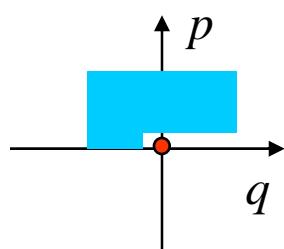
## 2. Image Modeling based on AR-Model: support

---

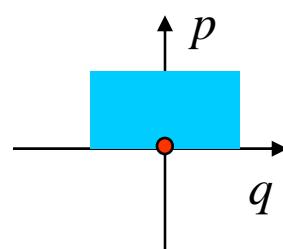
Jain

$$x(i, j) = \sum_{p,q \in M} a_{p,q} x(i-p, j-q) + e(i, j)$$

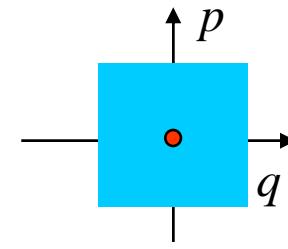
$$M = \begin{cases} \{(p, q) : (p \geq 0, q < 0) \cup (p > 0, q \geq 0)\}, & \text{causal} \\ \{(p, q) : (p > 0, \forall q) \cup (p = 0, q \neq 0)\}, & \text{semicausal} \\ \{(p, q) : (p, q) \neq 0\}, & \text{noncausal} \end{cases}$$



Causal



Semicausal



Noncausal

## 2. Image Modeling based on AR-Model

---

$$x(i) = \bar{x}(i) + e(i)$$

AR-process

Since AR-process is completely characterized by autocovariance function (or power spectrum), the question is how to model these quantities?

Consider typical models for autocovariance function.

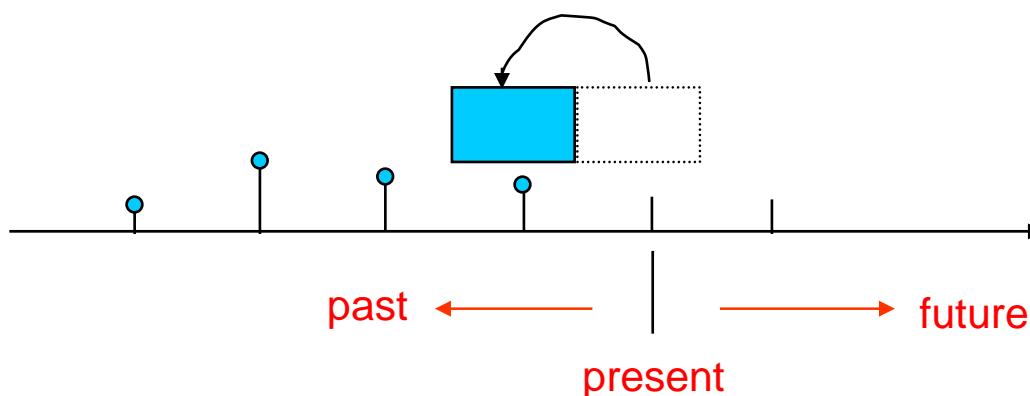
### 3. Image Modeling: Markov Processes

A Markov process  $\{x(i)\}$  is a random process whose past has no influence on the future, if its present is specified.

Definition:  $\{x(n)\}$  is a first-order Markov process, if

$$\mathbf{P}[x(n)|x(n-1), x(n-2), \dots] = \mathbf{P}[x(n)|x(n-1)]$$

“present” depends only on one sample in “past”



### 3. Markov Process: joint pdf (or marginal pdf)

---

Joint pdf:  $P[x^n] = P[x(n), x(n-1), x(n-2), \dots, x(0)] =$   
 $= P[x(n)|x(n-1), \dots, x(0)]P[x(n-1), \dots, x(0)]$

First order Markov

$$P[x(n)|x(n-1)]$$

$$P[x^{n-1}] = P[x(n-1), x(n-2), \dots, x(0)] =$$
  
 $= P[x(n-1)|x(n-2), \dots, x(0)]P[x(n-2), \dots, x(0)]$

First order Markov

$$P[x(n-1)|x(n-2)]$$

So on...

$$P[x^n] = \prod_{i=1}^n P(x(i)|x(i-1))P(x(0))$$

### 3. Markov Process: joint pdf

---

$$P[x^n] = \prod_{i=1}^n P(x(i)|x(i-1))P(x(0))$$

Terminology:

- $x(i)$  state
- $P(x(i)|x(i-1))$  transition probabilities

Model parameters:

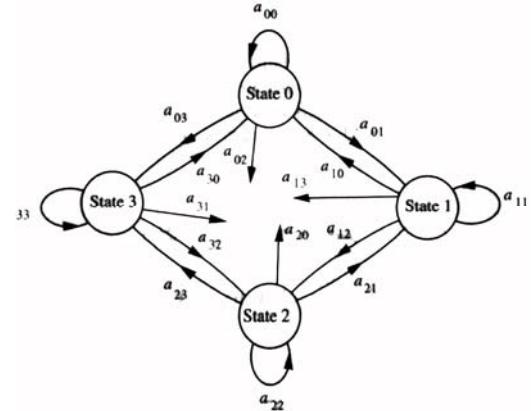
- Specify  $P(x(i)|x(i-1))$
- Use arbitrary  $P(x(0))$

Advantages: specifying only conditional probability between two samples, one can model very long sequences!

### 3. Markov Chains

---

- Terminology:
- If  $x(i)$  is in a finite set of L elements, then  $x(i)$  is said to be a Markov chain.
  - and  $P(x(i)|x(i-1))$  is LxL transition matrix for the Markov chain.



More details on Markov Chains will be considered when studying Hidden Markov Models (HMM).

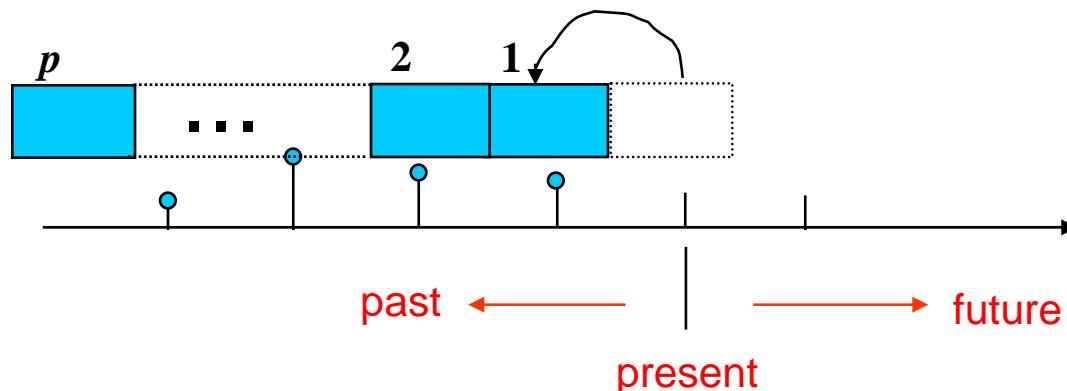
### 3. Image Modeling: $p$ th-order Markov Processes

Generalization:

Definition:  $\{x(n)\}$  is a  $p$ th-order Markov process, if

$$\mathbf{P}[x(n)|x(n-1), x(n-2), \dots] = \mathbf{P}[x(n)|x(n-1), \dots, x(n-p)]$$

“present” depends already on “ $p$ ” samples in “past”



### 3. First-order Markov Processes: a basic tool in IP

---

Example: AR(1) process is the first -order Markov process:

$$x(n) = \rho x(n-1) + e(n)$$

Since,  $e(n)$  is i.i.d. random variables with Gaussian pdf,  
the above process is called a Gauss-Markov process.

But Markov processes are:

$$x(i) = x(i-1)^2 + e(i)$$

$$x(i) = x(i-1)^2 e(i)$$

$$x(i) = x(i-1) \cos(e(i))$$

Depend on past, but are not AR(1)! (nonlinear)

Example: AR(p) process is the pth-order Markov process:

$$x(n) = \sum_{i=1}^p \rho(i)x(n-i) + e(n)$$

### 3. First-order Markov Processes: covariance

---

Covariance matrix of a zero-mean WSS sequence  $\{x(n)\}$  is determined as:

$$R_x = E[(X - M_x)^T (X - M_x)] = E[X^T X]$$

$$R_x = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N-1} \\ c_{21} & c_{22} & \cdots & c_{2N-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N-11} & c_{N-12} & \cdots & c_{N-1N-1} \end{bmatrix}$$

Elements:  $c_{ij} = E[(x(i) - m(i))(x(j) - m(j))] = E[x(i)x(j)]$

### 3. First-order Markov Processes: covariance

$$x(n) = \rho x(n-1) + e(n) \quad \xrightarrow{\text{Recursion}} \quad x(n) = \sum_{i=0}^{\infty} \rho^i e(n-i)$$

- ① Determine the variance of the process  $x(n)$ , assuming

$$m_e = E[e(n)] = 0 \rightarrow m_x = E[x(n)] = 0$$

$$\text{and } r_{ee}(m) = E[e(n)e(n+m)] = \sigma_e^2 \delta_{m0}$$

$$|\rho| < 1$$

$$\sigma^2 = E[x(n)^2] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \rho^i \rho^j E[e(n-i)e(n-j)] = \sigma_e^2 \sum_{i=0}^{\infty} \rho^{2i} = \frac{\sigma_e^2}{1-\rho^2}$$

### 3. First-order Markov Processes: covariance

---

- ② Determine the autocorrelation sequence:  $r_{xx}(m)$

$$\begin{aligned} r_{xx}(m) &= E[x(n)x(n+m)] = \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \rho^i \rho^j E[e(n-i)e(n-j+m)] = \sigma_e^2 \rho^{|m|} \sum_{i=0}^{\infty} \rho^{2i} = \frac{\sigma_e^2}{1-\rho^2} \rho^{|m|} \end{aligned}$$

### 3. First-order Markov Processes: covariance

---

The covariance matrix:

- ① for all diagonal elements (variances)

$$c_{nn} = E[x(n)x(n)] = E[x(n)^2] = \sigma^2$$

- ② for off-diagonal elements

$$c_{nm} = E[x(n)x(m)] = \sigma^2 \rho^{|m|}$$

### 3. First-order Markov Processes: covariance

---

Covariance matrix of a zero-mean WSS Markov sequence:

$$R_x = \begin{bmatrix} \sigma^2 & \sigma^2 \rho & \sigma^2 \rho^2 \dots & \sigma^2 \rho^{N-1} \\ \sigma^2 \rho & \sigma^2 & \dots & \sigma^2 \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2 \rho^{N-1} & \sigma^2 \rho^{N-2} & \dots & \sigma^2 \end{bmatrix}$$

This matrix is Toeplitz. It has a huge importance and influence on basic postulates of many image processing algorithms.

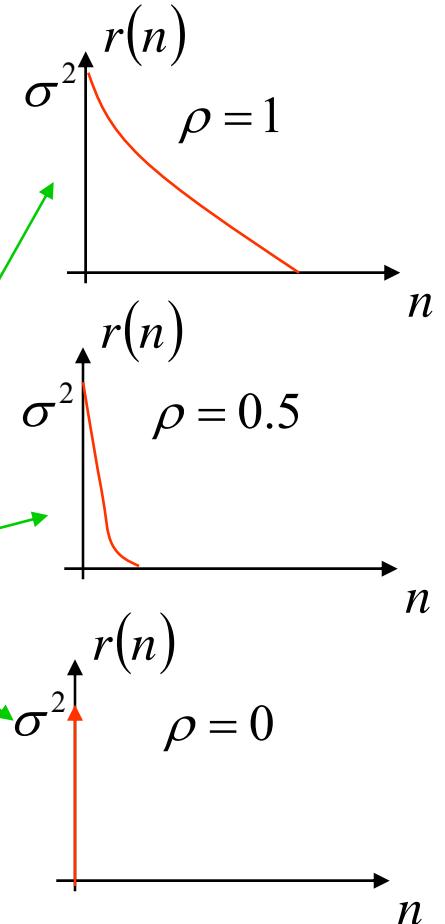
### 3. First-order Markov Processes: covariance

---

If  $\rho \rightarrow 0$ , then process will be uncorrelated.

$$R_x = \begin{bmatrix} \sigma^2 & & \cdots & 0 \\ \vdots & \sigma^2 & \cdots & \\ 0 & \vdots & \ddots & \vdots \\ & & \cdots & \sigma^2 \end{bmatrix}$$

General case:  $r(n) = \sigma^2 \rho^{|n|}, |\rho| < 1, \forall (n > 1)$



Conclusion:  $\rho$  is a correlation coefficient.

### 3. First-order Markov Processes: covariance

---

Covariance matrix of a zero-mean, unit variance WSS Markov sequence:

$$\sigma^2 = 1$$

$$R_x = \begin{bmatrix} 1 & \rho & \rho^2 \cdots & \rho^{N-1} \\ \rho & 1 & \ddots & \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \ddots & 1 \end{bmatrix}$$

$$r(n) = \rho^{|n|}, |\rho| < 1, \forall n$$

### 3. Discrete Random Fields: 2D formulation

---

#### Typical models for covariance matrix

The covariance function of a random field is called **separable** when it can be expressed as a product of covariance functions of one-dimensional sequences, i.e., if:

- ①  $r(m, n; m', n') = r_1(m, m')r_2(n, n')$  Nonstationary case
- ②  $r(m, n) = r_1(m)r_2(n)$  Stationary case

### 3. Image Modeling: Markov Processes

---

Separable stationary (Markov) covariance function is often used for IP:

$$r(m, n) = \sigma^2 \rho_1^{|m|} \rho_2^{|n|}, |\rho_1| < 1, |\rho_2| < 1$$

For  $\rho_1 = \rho_2$        $r(m, n) = \sigma^2 \rho^{|m|+|n|}$        $\rho_1, \rho_2 \approx 0.95$

More realistic model is non-separable exponential function:

$$r(m, n) = \sigma^2 e^{-\sqrt{m+n}}$$

Also called **isotropic** or **circularly symmetric**.

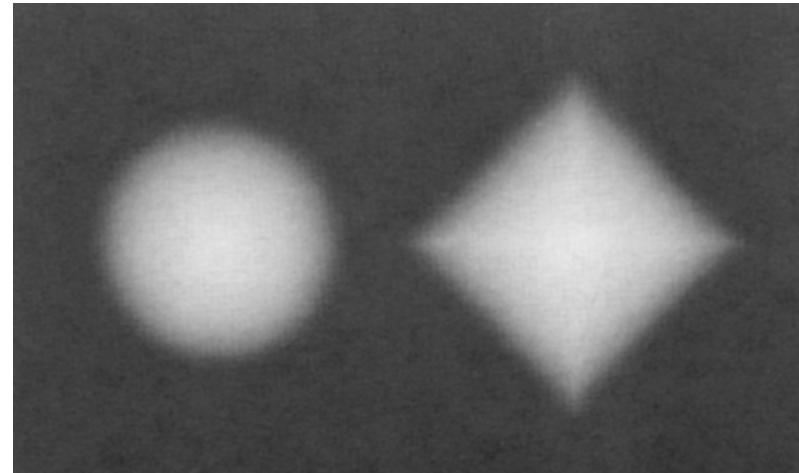
### 3. Image Modeling based on AR-Model

---

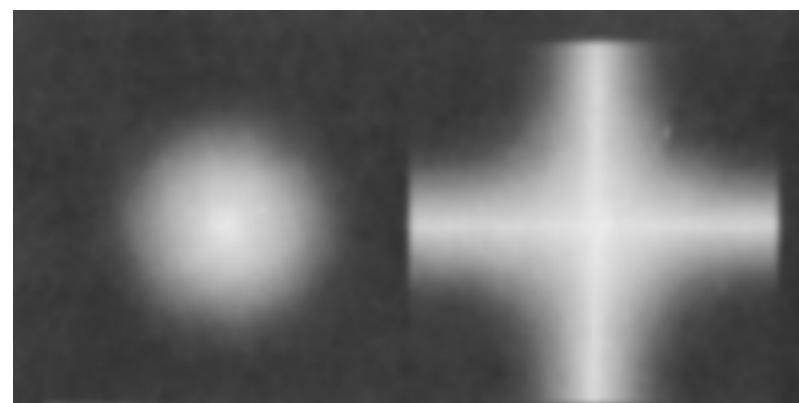
Isotropic

Separable

Covariance



Spectral Density



### 3. Image Modeling based on AR-Model

---

Problem of AR-models:

The above models fail to capture inhomogeneities in images.

Why to use:

- simple image presentation;
- basic models to capture correlation between image pixels;
- basic tool to understand optimal image processing in transform domain (see later FFT, DCT, wavelets and KLT in general).

## 4. Markov Random Fields (MRF)

- Noncausal model
- Advantages of MRF's:
  - Isotropic behavior
  - Only local dependencies
- Disadvantages of MRF's:
  - Computing joint probability is difficult
  - Parameter estimation is difficult
- Key theoretical result: Hammersley-Clifford theorem

## 4. Markov Random Fields (MRF)

---

The extension of the Markov property from 1-D to 2-D retains the concept of neighborhood structure in images.

But unlike in 1-D, **causality is unnatural** in 2-D and leads to **anisotropic models**.

### Main contributions:

- equivalence between MRFs and Gibbs distributions (Hammersley&Clifford, 1971): practical usage of MRFs;
- Further developed by Besag (1974);
- image analysis in the Bayesian framework (Geman&Geman, 1984);
- Generalized MRF theory (Bouman, 90s).

## 4. Markov Random Fields (MRF)

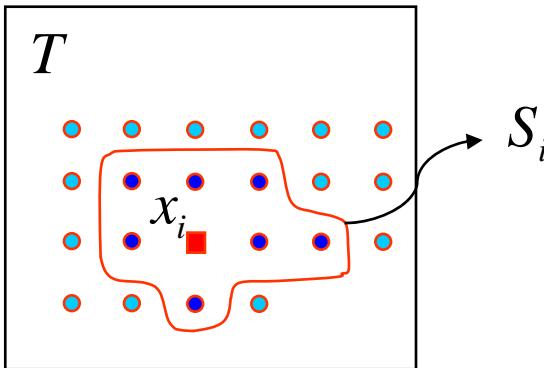
---

MRF theory tells us how to model the *a priori* probability of contextual dependent patterns, such as a class of textures and the arrangements of object features.

A particular MRF model favors its own class of patterns by associating them with larger probabilities than other pattern classes.

## 4. MRFs: Neighborhood System and Cliques

---



- Define neighborhood system

$$S = \{S_i \mid \forall i \in T\}$$

where  $S_i$  is the set of sites neighboring  $i$ ,  
 $T$  is a finite subset of the lattice  $\mathbb{Z}^2$ .

- Neighborhood system satisfies the following two conditions:
  - a site is not neighboring to itself:  $i \notin S_i$
  - the neighboring relationship is mutual: if  $i \in S_j \Leftrightarrow j \in S_i$

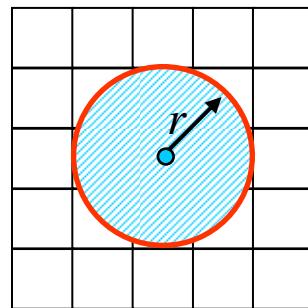
## 4. MRFs: Neighborhood System and Cliques

---

- For a regular lattice  $T$ , the neighbor set of  $i$  is defined as the set of nearby sites within a radius of  $r$ .

$$S_i = \left\{ i' \in T \mid \text{distance}[\text{pixel}_{i'}, \text{pixel}_i]^2 \leq r, i' \neq i \right\}$$

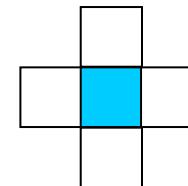
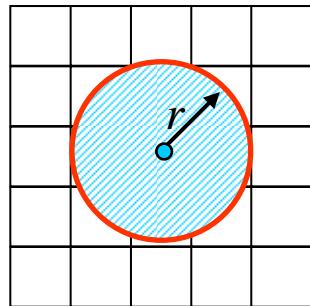
where  $\text{distance}[A, B]$  denotes the Euclidian distance between A and B and  $r$  takes the integer value.



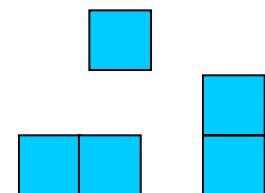
## 4. MRFs: Neighborhood System and Cliques

---

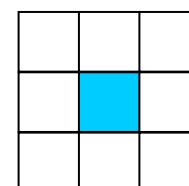
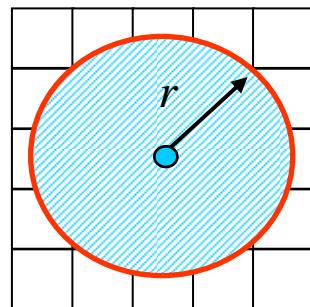
- First order neighborhood system (or 4-neighbor system)



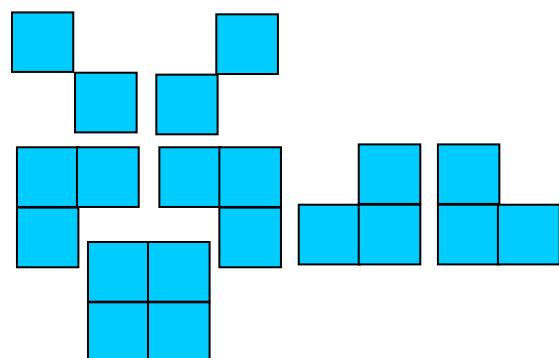
Possible cliques



- Second order neighborhood system (or 8-neighbor system)



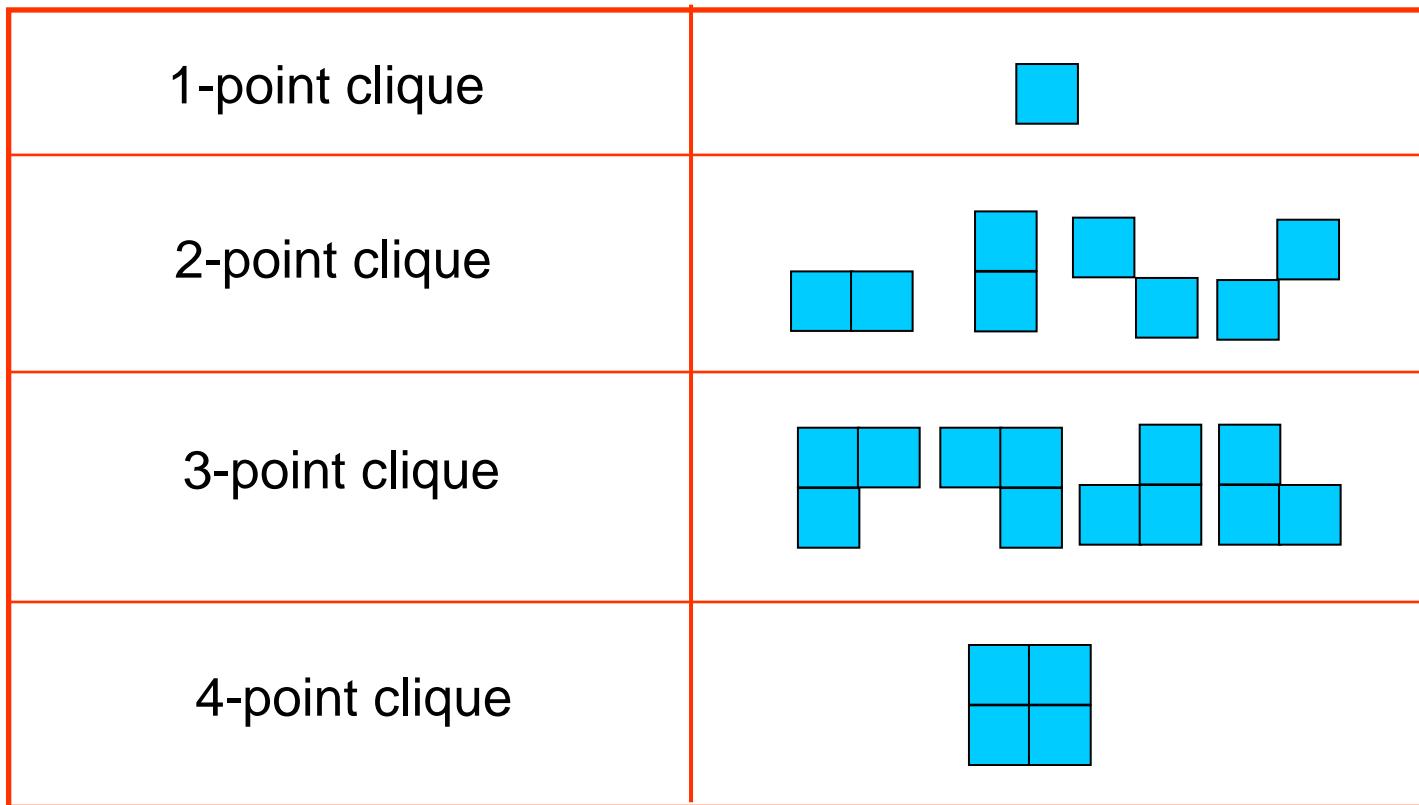
Additional Possible cliques



## 4. MRFs: Cliques on regular grid

---

The type of clique is determined by its **size**, **shape** and **orientation**.



## 4. MRFs: Cliques on regular grid

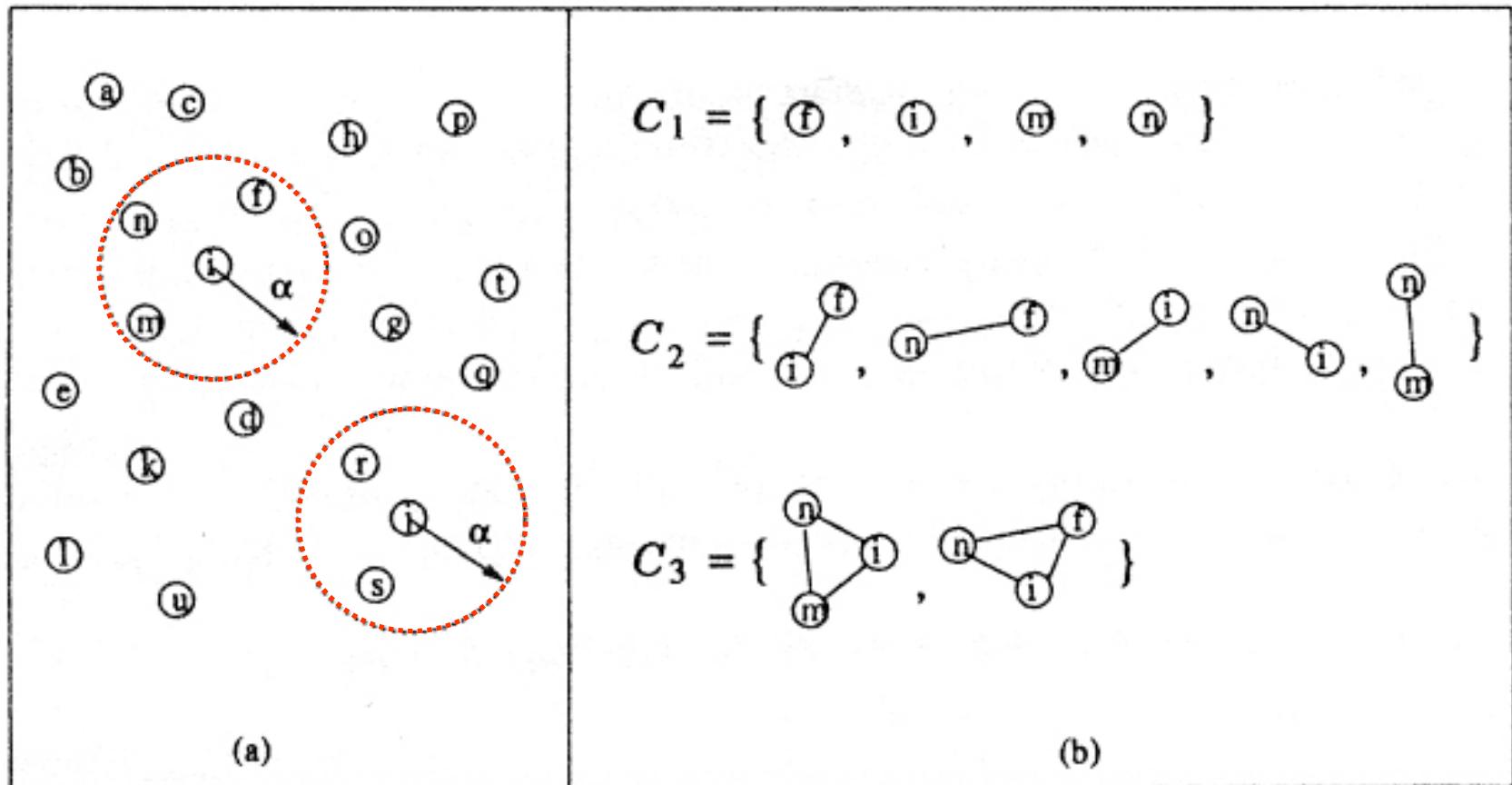
---

The order of the neighborhood system (and MRF respectively) is determined by the maximum order of clique.

5	4	3	4	5
4	2	1	2	4
3	1		1	3
4	2	1	2	4
5	4	3	4	5

A black oval highlights a 3x3 square centered at the cell containing the value 1. This highlights the neighborhood of the central cell, which includes the central cell itself and its four horizontal and vertical neighbors, forming a clique of size 3.

## 4. MRFs: General case: Cliques on irregular grid



## 4. MRFs: Definition

---

Assume finite **state space**  $x(i) \in \Lambda = \{0, \dots, L-1\}$   
# of gray levels

Definition:  $\{x(i), i \in T\}$  is a **MRF** with respect to (wrt) neighborhood system  $S$  if the following two conditions are satisfied:

- ①  $P(x) > 0 \quad \forall x \in \Lambda^{|T|}$  (positivity condition)  
(“technical” condition)
- ②  $P(x_i | x_j, i \neq j) = P(x_i | S_i)$  (Markovianity condition)  
(“local” condition: main feature of MRFs)  
 $P(x_i | \text{all points on lattice except } i) = P(x_i | \text{neighbors of } i)$
- ③ \* Stationarity (Homogeneity): Statistics are invariant to spatial shifts  
(do not depend on  $i$ ).

(This condition is often used but it is not central in definition of MRFs)

## 4. MRFs: Definition and further notes

---

Advantages of MRFs:

- ① No loss in generality for definition of neighborhood system and condition of positivity;
- ② Small neighborhoods still have long-range dependencies (for images!)

Problems of MRFs:

- ① No chain rule (like in Markov processes) to compute joint pdf  $P(x)$  of whole image in terms of  $P(x_i|S_i)$ .

$$P[x^n] = \prod_{i=1}^n P(x(i)|x(i-1))P(x(0))$$

Only for causal

- ② Does there exist some  $P(x)$  which is consistent with  $P(x_i|S_i)$ ?  
BAD NEWS: In general, no!

## 4. MRFs: Central question

---

How to compute the joint pdf for the whole image given the local interaction models based on the definition of local neighborhood sets and local correlations based on specified geometry of cliques?

## 4. Gibbs distributions (from statistical physics)

---

Definition: A Gibbs distribution relative to  $S$  is a probability measure  $P$  with the following representation:

$$P(x) = \frac{1}{Z} e^{-\frac{1}{T}U(x)}$$

Energy function  
Temperature  
Partition function

$$Z = \sum_{x \in \Lambda} e^{-\frac{1}{T}U(x)} \quad \text{Or normalizing constant to have integral probability } = 1$$

$$U(x) = \sum_{c \in C} V_c(x)$$

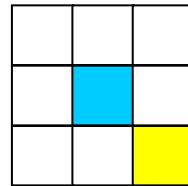
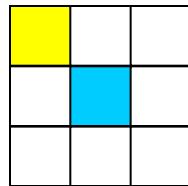
Potential function  
Set of cliques (summation is over all possible cliques)

Caussian distribution is a special member of Gibbs distribution family!

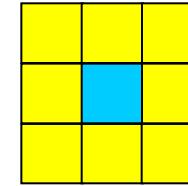
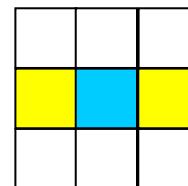
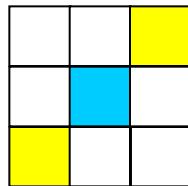
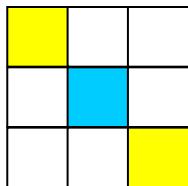
## 4. Gibbs distributions: properties

---

Definition: A Gibbs RF (GRF) is **homogeneous (strictly stationary)**, if  $V_c(x)$  is independent of the relative position of the clique  $c$  in  $S$ .



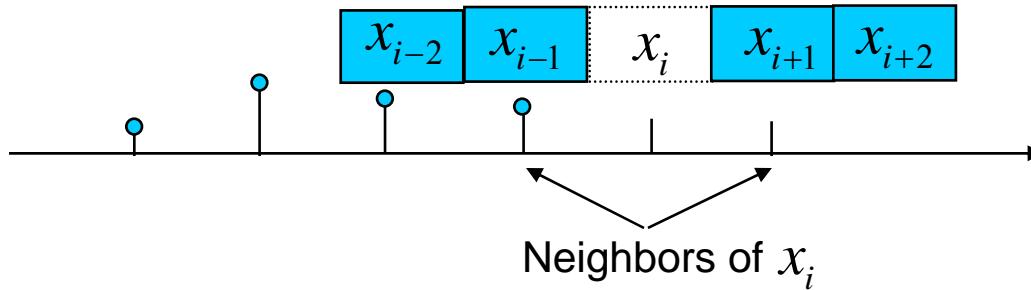
A Gibbs RF is **isotropic (rotationally invariant)**, if  $V_c(x)$  is independent of the orientation of  $c$ .



These assumptions are used to simplify mathematical description of GRF.

## 4. Example: Markov Processes are MRFs

---



- ① Neighborhood set:  $S_i = \{i-1, i+1\}$
  - ② Cliques have the form (for MP due to causality):  $c = \{i-1, i\}$
  - ③ Density has the form:  $p(x^n) = p(x(0)) \prod_{i=1}^n p(x(i)|x(i-1))$
- $z = e^{\ln z}$
- $$= p(x(0)) e^{\sum_{i=1}^n \underbrace{\ln p(x(i)|x(i-1))}_{V_c(x_i, x_{i-1})}} \quad \begin{aligned} \ln(ab) &= \ln a + \ln b \\ \ln(\prod) &= \sum \ln \end{aligned}$$
- Canonical form

## 4. Example: Markov Processes are MRFs

---

General form of Markov Process representation via Gibbs distribution  
(in exponential form)

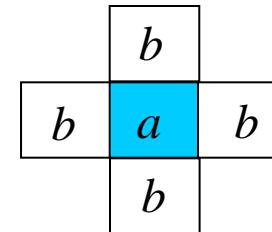
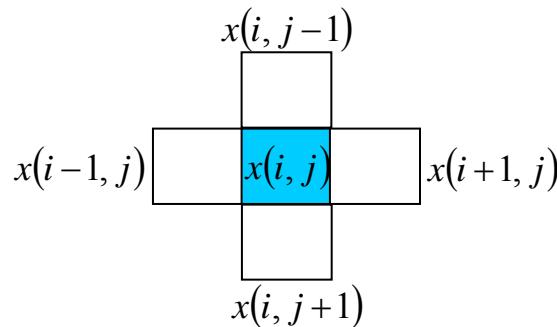
$$p(x^n) = p(x(0)) \exp \left\{ \sum_{i=1}^n V(x_i, x_{i-1}) \right\}$$

## 4. 2-D Example: Ising model (1925) (ferromagnetism)

---

Binary random field with 1-st order clique system

$$x_i \in \{0,1\}$$



$$U(x) = \sum_{c \in C} V_c(x) = \sum_{i \in S} V_1(x_i) + \sum_{i \in S} \sum_{i' \in S_i} V_2(x_i, x_{i'})$$

$$U(x(i, j)) = ax(i, j) + bx(i, j)(x(i-1, j) + x(i+1, j) + x(i, j-1) + x(i, j+1))$$

Ising model is homogeneous and isotropic.

## 4. Hammersley-Clifford Theorem (HCT)

---

$\{x(i)\}$  is a MRF with respect to  $S$   
if and only if  $P(x)$  is a Gibbs distribution with respect to  $S$ .

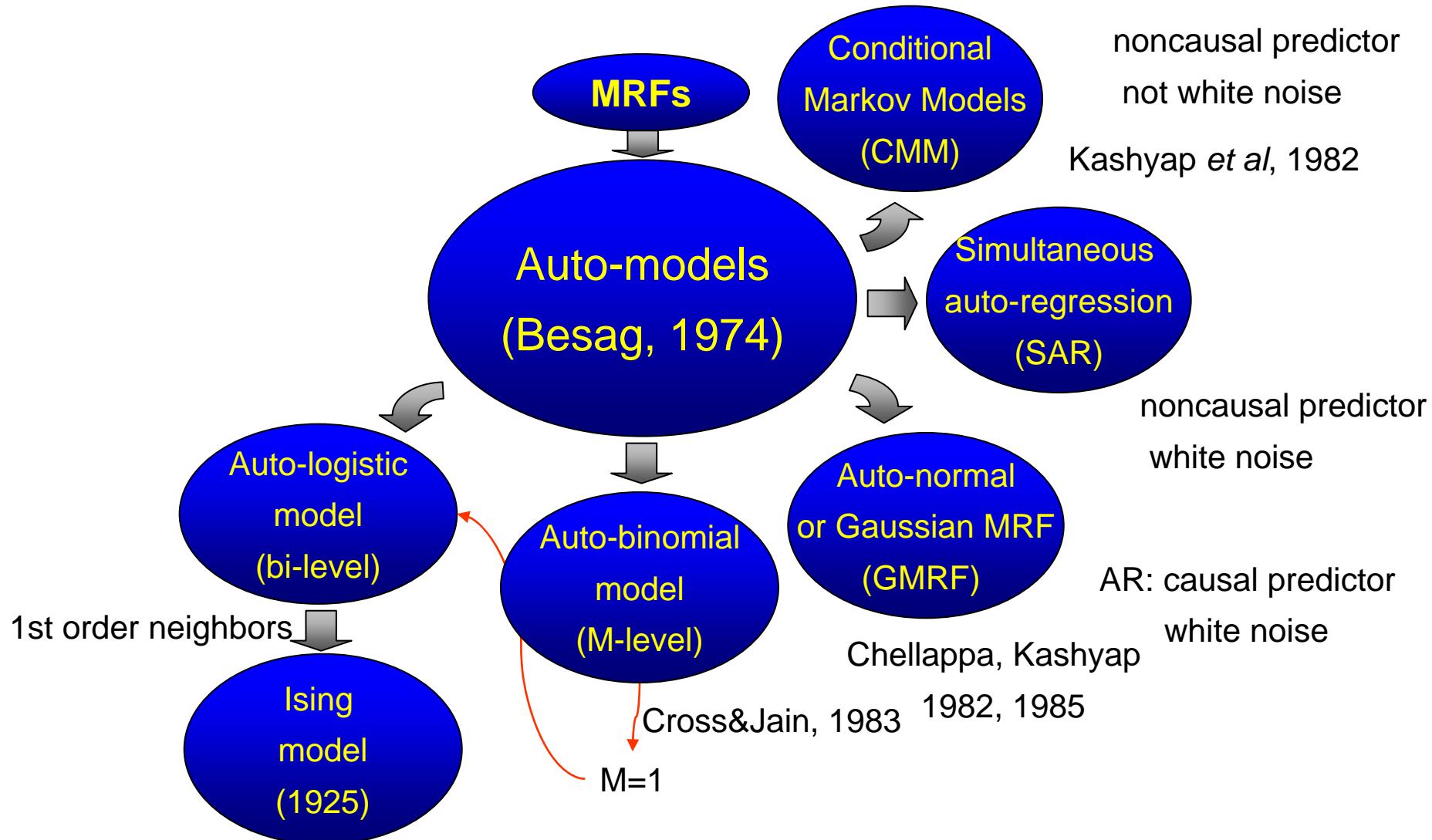
HCT provides a connection between the purely graph-theoretic relationships on a lattice with algebraic form of the density function.

It leads to the possibility to specify MRF via potentials  $\{V_c(x)\}$ .

There exist explicit formulas for computing the energy  $V_c(x)$  from the local characteristics of the MRFs.

Conversely, one can compute local characteristics from  $V_c(x)$ .

## 4. Auto-Models: Classification



## 4. Auto-models

---

Usage: especially useful for simulation of textures.

Main questions:

- conditional pdf (for a pixel in a given neighborhood system);
- joint pdf;
- corresponding energy functions.

## 4. Auto-models

---

We constrain our analysis only by pair-site neighborhood system.

The general form of energy function (cliques up to two):

$$U(x) = \sum_{c \in C} V_c(x) = \sum_{i \in S} V_1(x_i) + \sum_{i \in S} \sum_{i' \in S_i} V_2(x_i, x_{i'})$$

Or equivalently:

$$U(x) = \sum_{\{i\} \in C_1} V_1(x_i) + \sum_{\{i, i'\} \in C_2} V_2(x_i, x_{i'})$$

We can receive the different classes of MRFs depending on the proper selection of potential functions.

## 4. Auto-models: Definition

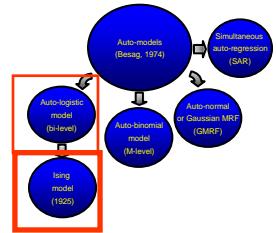
---

When  $\underline{V_1(x_i) = x_i G_i(x_i)}$  and  $\underline{V_2(x_i, x_{i'}) = \beta_{i,i'} x_i x_{i'}}$

where  $G_i(\cdot)$  are arbitrary functions and  $\beta_{i,i'}$  are constants reflecting the pair-site interaction between  $i$  and  $i'$ , then:

$$U(x) = \sum_{\{i\} \in C_1} x_i G_i(x_i) + \sum_{\{i, i'\} \in C_2} \beta_{i,i'} x_i x_{i'}$$

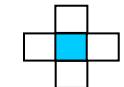
The above is called **auto-models** (Bessag, 1974).

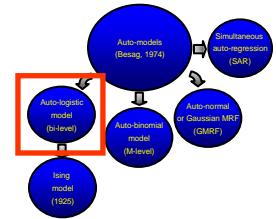


## 4. Auto-models: auto-logistic model

Bi-level model:  $x(i) \in \{0,1\}$  or  $x(i) \in \{-1,1\}$

$$U(x) = \sum_{\{i\} \in C_1} \alpha_i x_i + \sum_{\{i,i'\} \in C_2} \beta_{i,i'} x_i x_{i'}$$

When S (neighbors) is 4 nearest neighbors   
then auto-logistic model is reduced to Ising model.

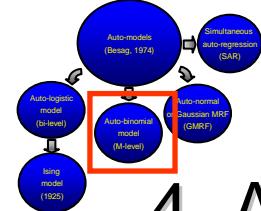


## 4. Auto-models: auto-logistic model

The conditional probability for auto-logistic model with  $x(i) \in \{0;1\}$

$$p(x_i | x_{S_i}) = \frac{e^{\alpha_i x_i + \sum_{i' \in S_i} \beta_{i,i'} x_i x_{i'}}}{\sum_{x_i \in \{0,1\}} e^{\alpha_i x_i + \sum_{i' \in S_i} \beta_{i,i'} x_i x_{i'}}} = \frac{e^{\alpha_i x_i + \sum_{i' \in S_i} \beta_{i,i'} x_i x_{i'}}}{1 + e^{\alpha_i + \sum_{i' \in S_i} \beta_{i,i'} x_{i'}}}$$

When the distribution is **homogeneous**, we have  $\alpha_i = \alpha$  and  $\beta_{i,i'} = \beta$  regardless of  $i$  and  $i'$ .



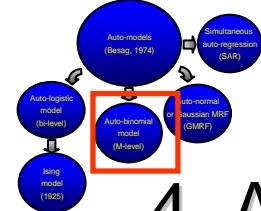
## 4. Auto-models: auto-binomial model Cross&Jain, 1983

M-level model:  $x(i) \in \{0, 1, \dots, M - 1\}$

and every  $x(i)$  has a conditional **binomial distribution** of  $M$  trials  
and probability of success  $q$

$$p(x_i | x_{S_i}) = \binom{M-1}{x_i} q^{x_i} (1-q)^{M-1-x_i}$$

$$q = \frac{e^{\alpha_i + \sum_{i' \in S_i} \beta_{i,i'} x_{i'}}}{1 + e^{\alpha_i + \sum_{i' \in S_i} \beta_{i,i'} x_{i'}}}$$

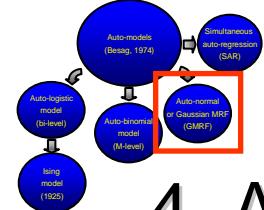


## 4. Auto-models: auto-binomial model Cross&Jain, 1983

Energy function for auto-binomial model:

$$U(x) = - \sum_{\{i\} \in C_1} \ln \left( \frac{M-1}{x_i} \right) - \sum_{\{i\} \in C_1} \alpha_i x_i + \sum_{\{i, i'\} \in C_2} \beta_{i, i'} x_i x_{i'}$$

It reduces to the auto-logistic model when  $M=1$ .



## 4. Auto-models: auto-normal (Gaussian MRF)

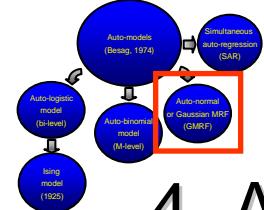
---

Continuous models:  $x(i) \in \Re$

The joint distribution is **multivariate normal**.

The conditional pdf is **normal distribution**:

$$p(x_i | x_{S_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \left[ x_i - m_{x_i} - \sum_{i' \in S_i} \beta_{i,i'} (x_{i'} - m_{x_{i'}}) \right]^2}$$



## 4. Auto-models: auto-normal (Gaussian MRF)

---

The conditional mean:

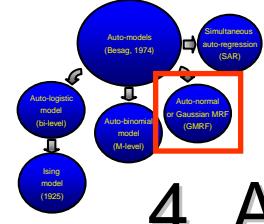
$$E(x_i | x_{S_i}) = m_{x_i} - \sum_{i' \in S_i} \beta_{i,i'} (x_{i'} - m_{x_{i'}})$$

The conditional variance:

$$\text{var}(x_i | x_{S_i}) = \sigma^2$$

The joint pdf:

$$p(x^n) = \frac{\sqrt{\det(B)}}{(2\pi\sigma^2)^N} e^{-\frac{1}{2\sigma^2} (x - m_x)^T B (x - m_x)}$$



## 4. Auto-models: auto-normal (Gaussian MRF)

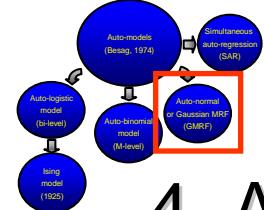
Compare:

$$p(x^n) = \frac{\sqrt{\det(B)}}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}(x-m_x)^T B (x-m_x)}$$

$$p(x^n) = \frac{1}{(2\pi)^{\frac{N}{2}} \sqrt{\det(R)}} e^{-\frac{1}{2}(x-m_x)^T R^{-1} (x-m_x)}$$

Multivariate Gaussian

$$R^{-1} = \frac{1}{\sigma^2} B$$



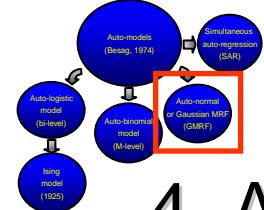
## 4. Auto-models: auto-normal (Gaussian MRF)

---

$B = [b_{i,i'}]$  is the NxN interaction matrix whose elements are unity and off-diagonal elements at  $i, i'$  are  $-\beta_{i,i'}$ .

$$b_{i,i'} = \delta_{i,i'} - \beta_{i,i'} \text{ with } \beta_{i,i} = 0.$$

See Jain, p. 153 to compare with the inverse covariance matrix of Gaussian process.



## 4. Auto-models: auto-normal (Gaussian MRF)

The single-site and pair-site potential functions for the auto-normal models are:

$$V_1(x_i) = (x_i - m_{x_i})^2 / 2\sigma^2$$

$$V_2(x_i, x_{i'}) = \beta_{i,i'}(x_i - m_{x_i})(x_{i'} - m_{x_{i'}}) / 2\sigma^2$$

Note: independent Gaussian model is a special case of GMRF whose Gibbs energy consists of only single-site clique potentials ( $V_2(x_i, x_{i'}) = 0$ ) .

Because all higher order clique potentials are zero, there is no contextual interaction in the above model (diagonal  $B$ ).

## 4. MRFs: simulation (generation)

---

- How to generate the random field with a Gibbs distribution?

$$p(x) = \frac{1}{Z} e^{-\frac{1}{T}U(x)}$$

- Generally, this problem is difficult.
- Markov Chains can be generated sequentially.
- Non-causal structure of MRFs makes simulation difficult.

## 4. MRFs: simulation (generation)

---

MRFs generators:

- Metropolis sampler
- Gibbs sampler
- Generalized Metropolis sampler

## 4. Metropolis sampler

---

Main references:

- N. Metropolis et al., Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087-1091, 1953.
- R. Kindermann and J. Snell. *Markov Random fields and their Applications*. American Mathematical Society, Providence, 1980.

Basic idea:

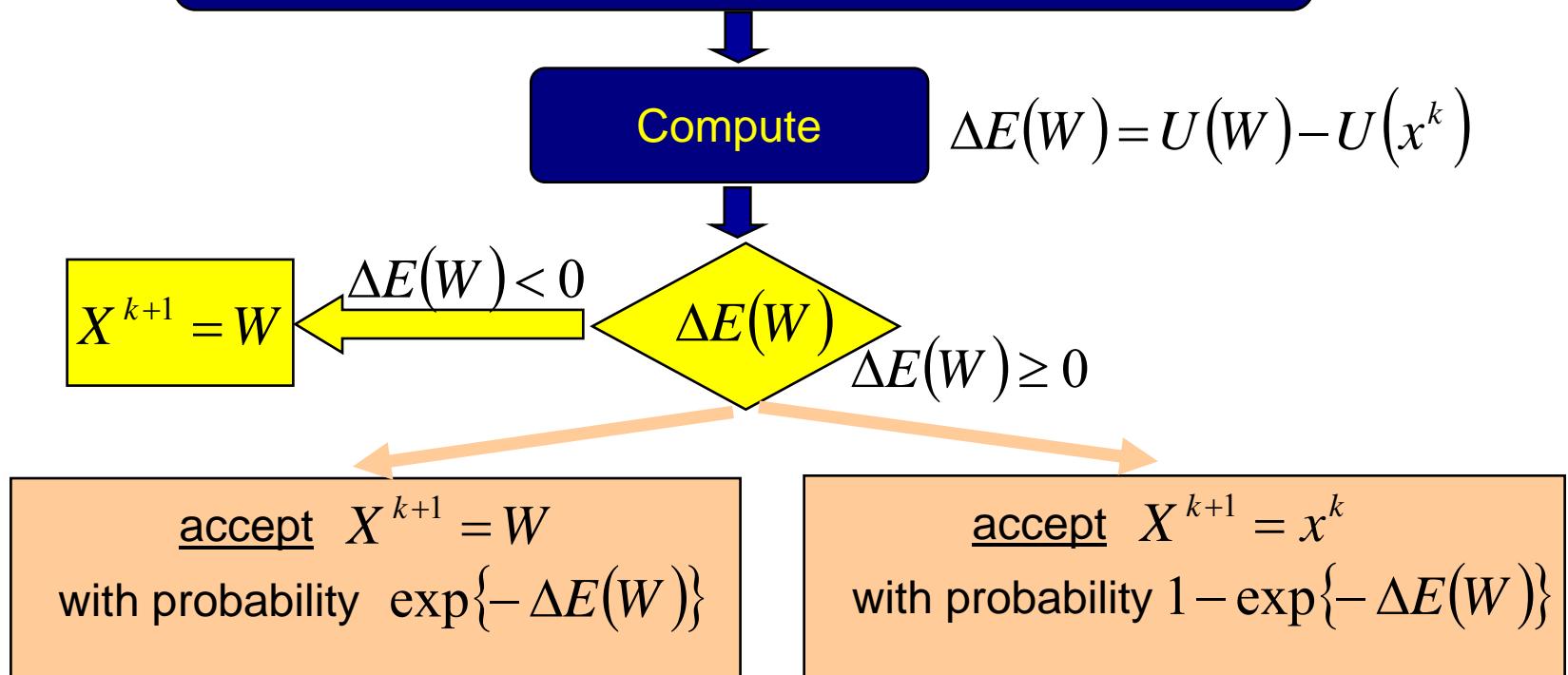
To construct a Markov chain in time which converges to an equilibrium distribution which should correspond to the desired MRF pdf.

## 4. Metropolis sampler

$$p(x) = \frac{1}{Z} e^{-\frac{1}{T} U(x)}$$

Start with the sample  $x^k$ , and generate a new sample

$W$  with symmetric probability  $q(w/x^k) = q(x^k/w)$ .



This algorithm has an asymptotic convergence (but very slow!)

## 4. Metropolis sampler: examples [Bouman, p.34]

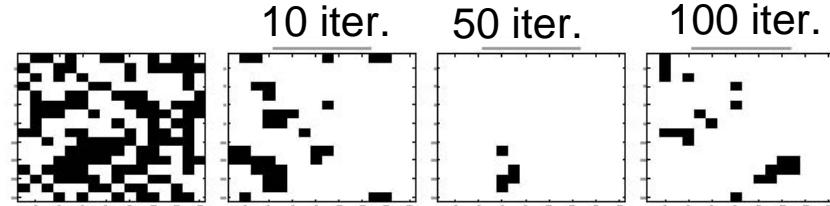
---

Ising MRF

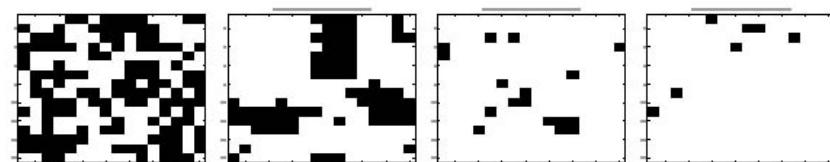
a=0

b=1

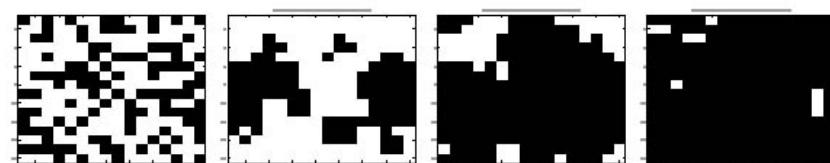
- Test 1



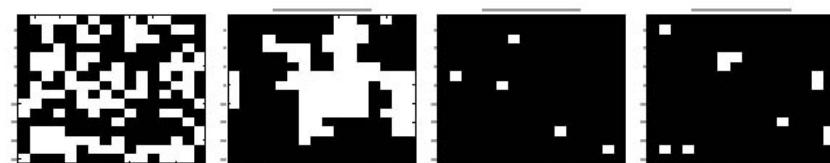
- Test 2



- Test 3



- Test 3



## 4. Metropolis sampler: properties

---

### Advantages:

- has guaranteed geometric convergence;
- can be easily implemented.

### Disadvantages:

- can be slow if there are many rejections;
- is constrained to use a symmetric transition function  $q(x^{k+1}/x^k)$ .  
(this constraint is solved in Generalized Metropolis Sampler)

## 4. Gibbs sampler [Geman&Geman]

---

Difference with Metropolis sampler:

- instead of choosing the state completely randomly,  
the state is chosen from the conditional density of the pixel given  
the value of his neighbors;
- scan through all the pixels in the image.

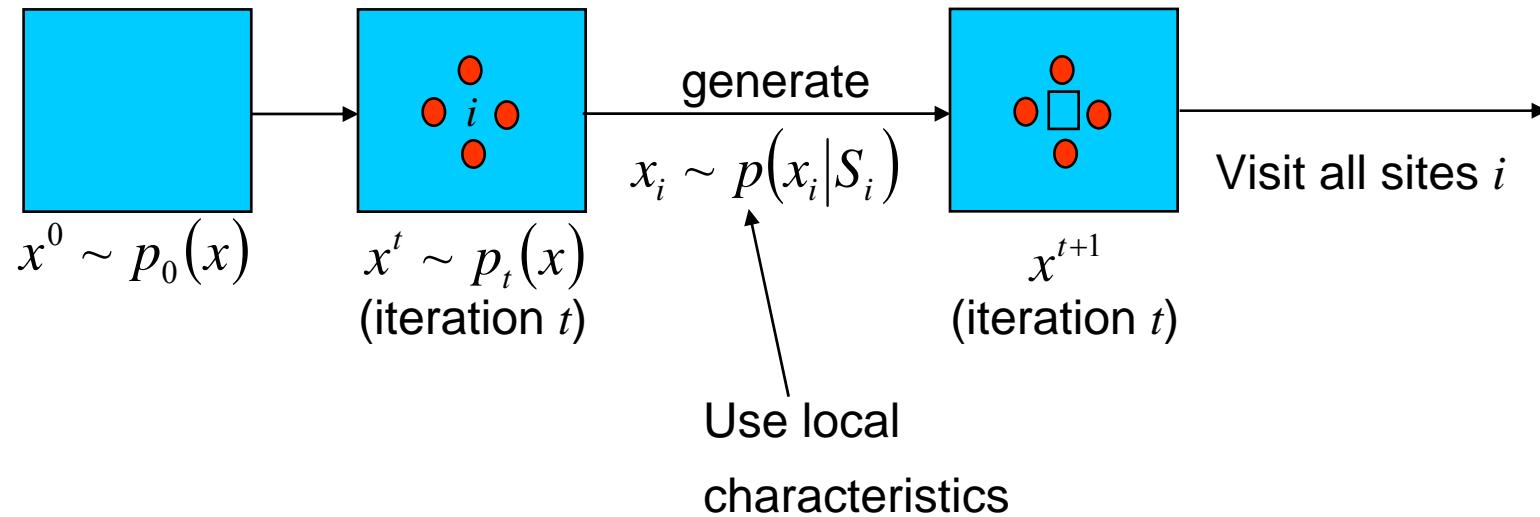
Advantage:

Eliminates need for rejections → faster convergence.

## 4. Gibbs sampler

---

- generate initial  $x^0$  from some arbitrary distribution  $p_0(x)$
- iterate as follows:



## 4. Texture Modeling (Cross&Jain, 1983): choice of potentials

---

Goal: provide realistic texture models

What is a texture: (6 features)

(Tamura et al (IEEE Trans. Syst, Man, Cybern., 1978), pp.460-473)

- coarseness, contrast, directionality, line-likeness, regularity, and roughness.

Applications:

- segmentation (looking for the objects: road, fields models....);
- classification (given 4 different textures: classify them);
- computer graphics (modeling of textures analytically);
- compression (field can be presented using only model parameters);
- restoration (prior model of image).

## 4. Texture Modeling

---

Probability of a point  $x(i, j)$  having gray level  $L$  is binomial, with parameter determined by its neighbors.

Besag formulated **autobinomial** model with energy function  $U(x)$ .

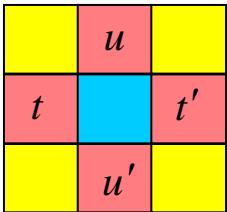
For binary fields (auto-logistic model):  $x(i, j) \in \{0,1\}$

$$p(x|neighbors) = \frac{e^{U(x)}}{\underbrace{e^{U(x=0)} + e^{U(x=1)}}_{1 + e^{U(x=1)}}} = \frac{e^{U(x)}}{1 + e^{U(x=1)}}$$

Partition function sum  $p(x)=1$

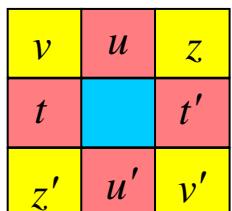
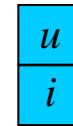
## 4. Texture Modeling

---



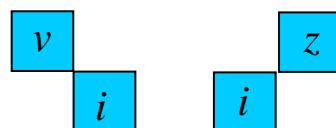
First order model:

$$U(x_i) = x_i(a + b_H(t + t') + b_V(u + u'))$$



Second order model (also diagonal):

$$U(x_i) = x_i(a + b_H(t + t') + b_V(u + u') + b_{D_L}(v + v') + b_{D_R}(z + z'))$$



## 4. Texture Modeling

---

Model parameters:  $(a; b_H; b_V; b_{D_L}; b_{D_R})$

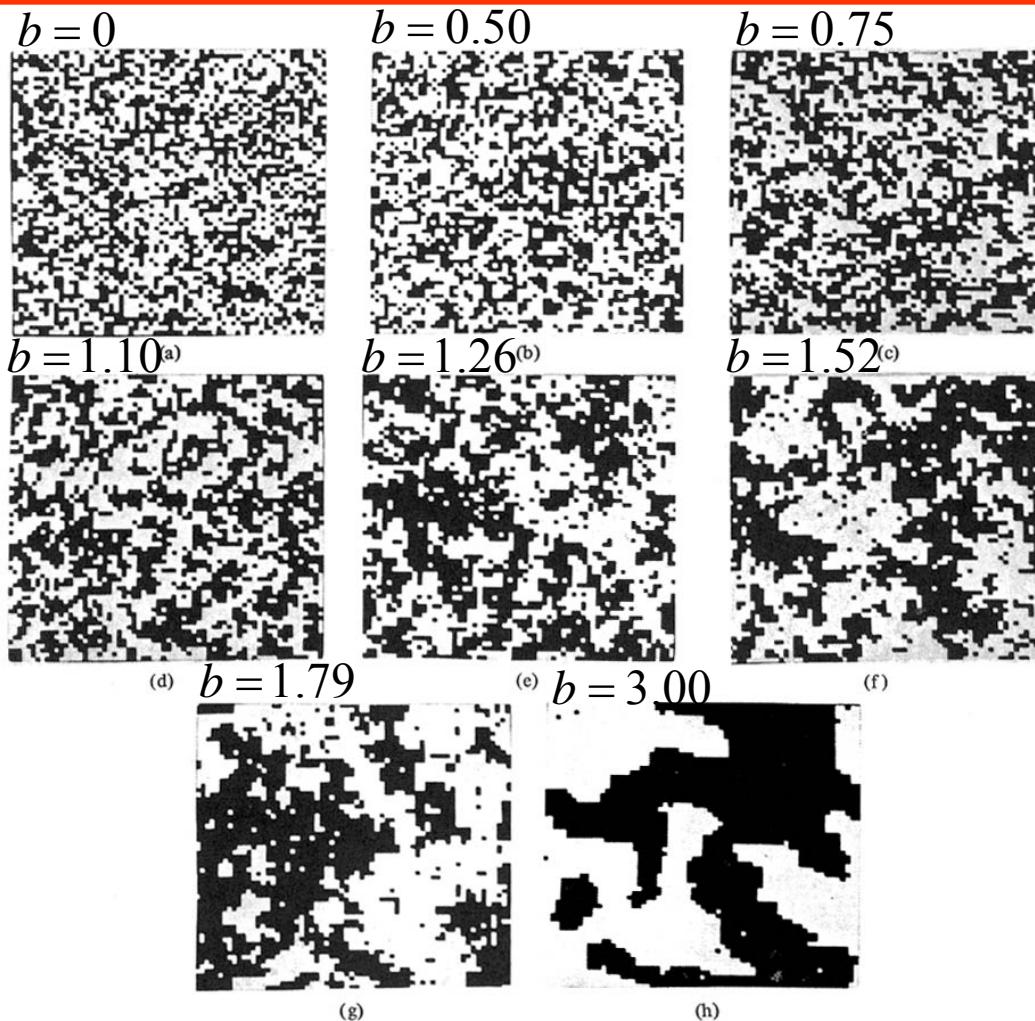
Positive values favor clustering  
in horizontal and vertical directions

Positive values favor clustering  
in left and right diagonal directions

If  $b_H = b_V$  or  $b_{D_L} = b_{D_R}$  model will have isotropic behavior.

Using Metropolis sampler (random field generation procedure)  
Cross&Jain have shown the possibility to generate different textures.

## 4. Texture generation: Clustering effects (Cross&Jain,p49)



Isotropic 1st order model

$$a = -2b$$

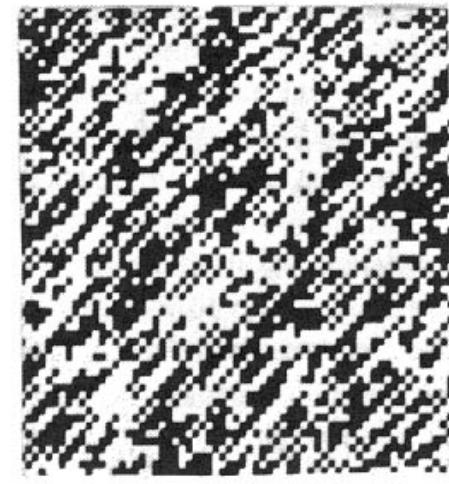
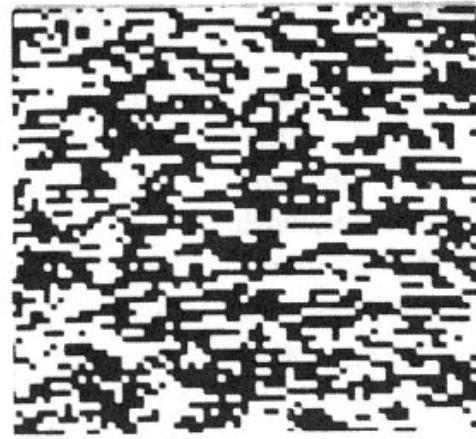
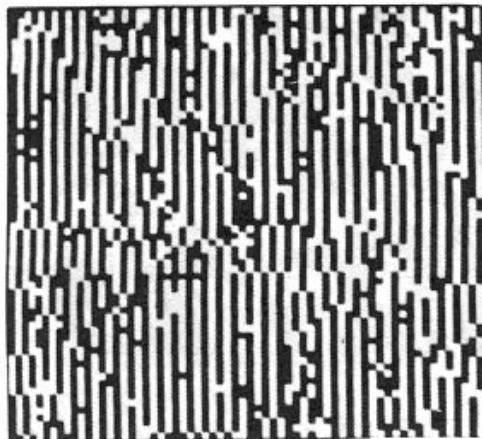
$$b = b_H = b_V$$

The increase of  $b$   
leads to the increase  
of clustering.

## 4. Texture generation: Anisotropic effects

---

Anisotropic 2nd order model



	$a$	$b_H$	$b_V$	$b_{D_R}$	$b_{D_L}$
A: vertical	-0.26	-2	2.1	0.13	0.015
B: horizontal	-2.04	1.93	0.16	0.07	0.02
C: Right diagonal	-1.90	-0.1	0.1	1.9	-0.07

## 4. Texture generation: Ordered patterns

---



## 4. Texture regeneration: Inverse problem

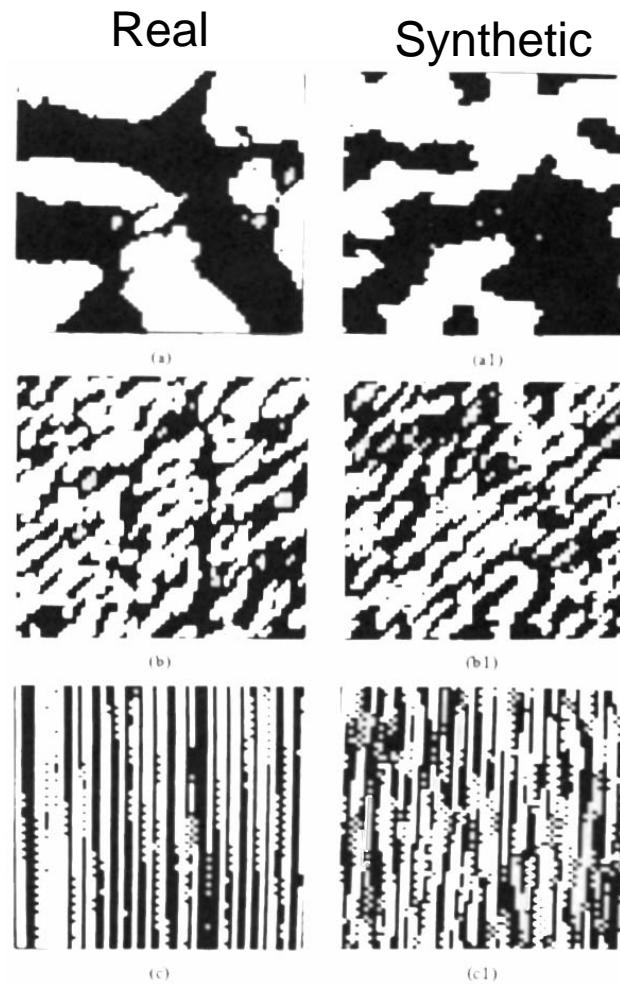
---

### Problem:

- given real texture pattern;
- regenerate the same pattern analytically.

### Solution:

- estimate parameters from real pattern;
- use sampling procedure considered above to generate new synthetic pattern.



## 4. Non-Gaussian MRF

---

Topics to be covered:

- Quadratic functions
- Non-convex functions
- Convex functions

## 4. Non-Gaussian MRFs

---

- Gaussian MRFs do not model transitions (edges) well  
(strongly penalize images with large changes ➔ favors homogeneity)
- In applications such as image restoration and tomography,  
Gaussian MRFs either
  - blur edges (ringing artifacts)
  - leave excessive amount of noise

## 4. Non-Gaussian MRFs: Line Process (LP) model

---

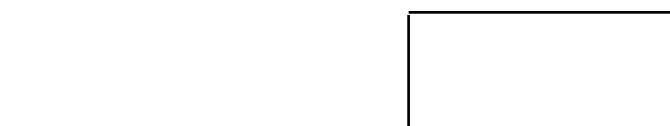
- One of the first attempts to avoid edge blurring due to Gaussian (quadratic) energy function was done by Geman and Geman (1984) that had a lot of influence on popularization of MRFs.
- The LP models assumes **piecewise smoothness** whereby the smoothness constraint is switched off at the points where the magnitude of the signal derivative exceeds certain threshold.

## 4. Non-Gaussian MRFs: Line Process (LP) model

---

### Line Process definition:

Signal  $x(i)$

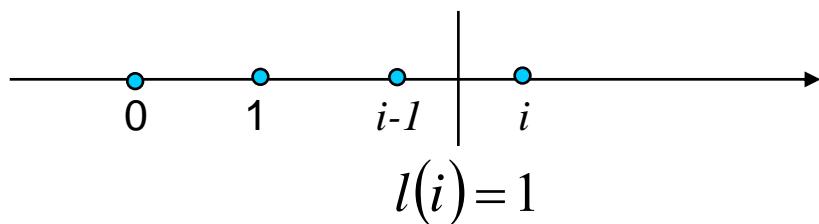


Derivative:  $[x(i) - x(i-1)]^2$



$l(i) = 1$  - discontinuity is detected

$l(i) = 0$  - signal is smoothed



## 4. Non-Gaussian MRFs: Line Process (LP) model

---

The LP energy function:

$$U(x, l) = \sum_{i=2}^m [x(i) - x(i-1)]^2 [1 - l(i)] + \alpha \sum_{i=2}^m l(i)$$

Model parameters:

$$l(i) = \begin{cases} 0, & \text{if } [x(i) - x(i-1)]^2 < \alpha \\ 1, & \text{otherwise.} \end{cases}$$

$\alpha$  - threshold, which determines edge detection

Model properties::

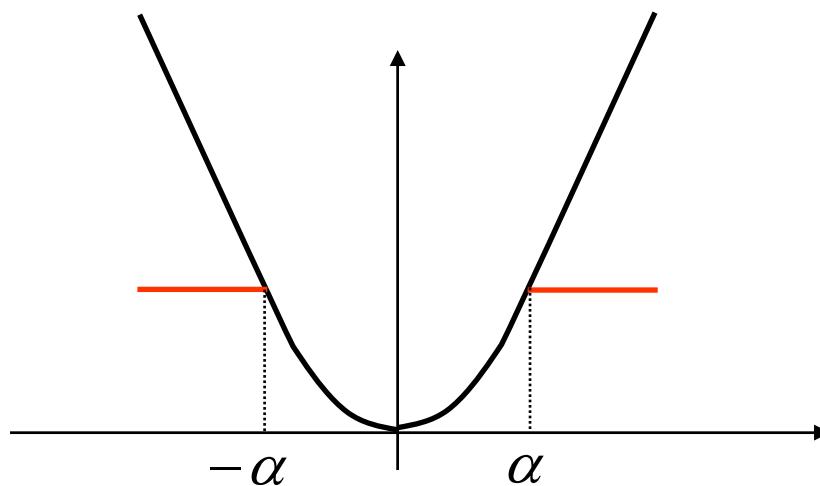
$$U(x, l) = \sum_{i=2}^m [x(i) - x(i-1)]^2 \quad \text{Gaussian MRF if } l(i) = 0$$

## 4. Non-Gaussian MRFs: Line Process (LP) model

---

Blake& Zisserman(1987) eliminated line process model parameter and converted the LP (Geman&Geman) into equivalent energy function:

$$U(x, \alpha) = \min \{ [x(i) - x(i-1)]^2, \alpha \}$$



## 4. Gaussian MRF: extension to Non-Gaussian MRFs

---

Gaussian MRF's have density functions with the form:

$$p(x) = \frac{1}{Z} \exp \left\{ - \sum_i a(i) x(i)^2 - \sum_{i,i'} \beta_{i,i'} |x(i) - x(i')|^2 \right\}$$

Let  $x(i) \in \{0,1,\dots,M-1\}$  and  $a(i) = 0$

$|x(i) - x(i')|$  - is change in gray level

$|x(i) - x(i')|^2$  - penalizes rapid changes in gray levels

- approximation of first order derivative  $|x(i) - x(i-1)|$
- approximation of second order derivative  $|x(i-1) - 2x(i) + x(i+1)|$

## 4. Non-Gaussian MRFs based on Pair-Wise Cliques

---

We will consider MRF's with pair-wise cliques:

$$p(x) = \frac{1}{Z} \exp \left\{ - \sum_{i,i'} \rho \left( \frac{x(i) - x(i')}{\sigma} \right) \right\}$$

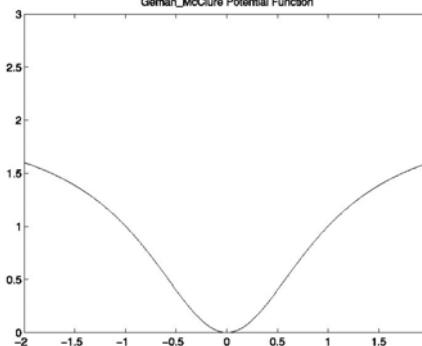
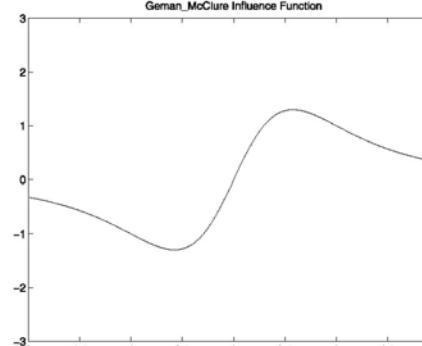
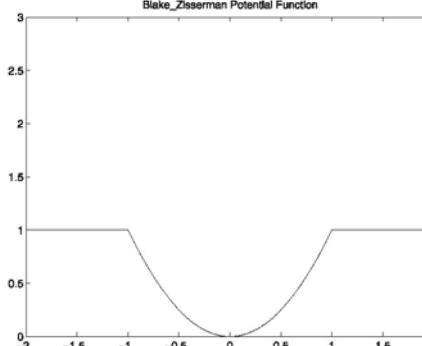
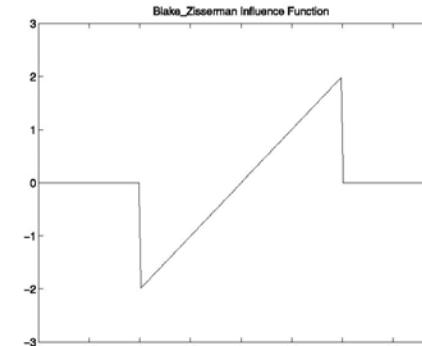
$\sigma$  - controls the gray level variation or scale.

$\rho(\Delta)$  is the **potential function**:

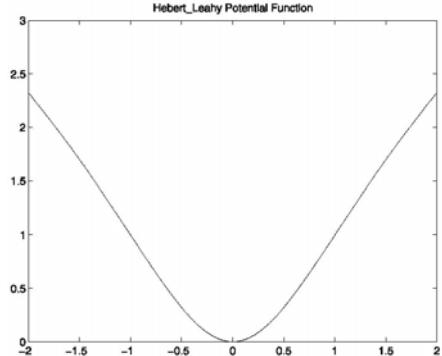
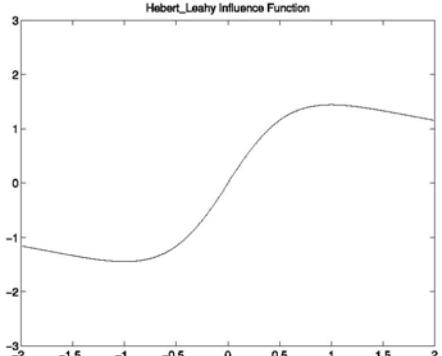
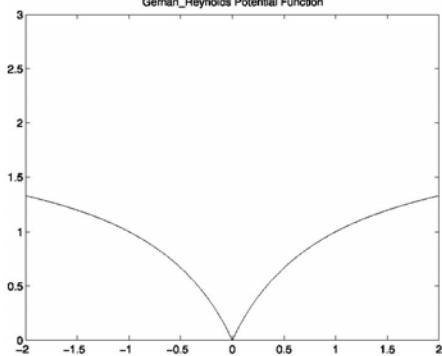
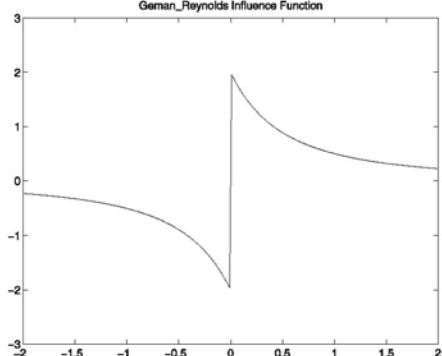
- determines the cost of abrupt changes on gray level;
- $\rho(\Delta) = |\Delta|^2$  is the Gaussian model.

$\rho'(\Delta) = \frac{d\rho(\Delta)}{d\Delta}$  is the **influence function** (from M-estimators).

## 4. Non-Gaussian MRFs: non-convex

Authors	$\rho(\Delta)$	Potential func.	Influence func.
Geman&McClure	$\frac{\Delta^2}{1 + \Delta^2}$	 <p>German_McClure Potential Function</p>	 <p>German_McClure Influence Function</p>
Blake&Zisserman (Line Process G&G, Talvar function )	$\min\{\Delta^2, 1\}$	 <p>Blake_Zisserman Potential Function</p>	 <p>Blake_Zisserman Influence Function</p>

## 4. Non-Gaussian MRFs: non-convex (cont.)

Authors	$\rho(\Delta)$	Potential func.	Influence func.
Hebert&Leahy	$\log(1 + \Delta^2)$	 A plot titled "Hebert_Leahy Potential Function" showing a convex U-shaped curve. The x-axis ranges from -2 to 2 with ticks at -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2. The y-axis ranges from 0 to 3 with ticks at 0, 0.5, 1, 1.5, 2, 2.5, 3. The curve starts at approximately (0, 2.3), dips to a minimum of about 0.1 at x=0, and rises back to approximately (0, 2.3).	 A plot titled "Hebert_Leahy Influence Function" showing a symmetric, bell-shaped curve. The x-axis ranges from -2 to 2 with ticks at -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2. The y-axis ranges from -3 to 3 with ticks at -3, -2, -1, 0, 1, 2, 3. The curve passes through (-2, -1.1), reaches a minimum of about -1.4 at x=-0.5, crosses the x-axis at x=0, reaches a maximum of about 1.4 at x=0.5, and returns to (-2, -1.1).
Geman&Reynolds	$\frac{ \Delta }{1 +  \Delta }$	 A plot titled "Geman_Reynolds Potential Function" showing a convex U-shaped curve. The x-axis ranges from -2 to 2 with ticks at -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2. The y-axis ranges from 0 to 3 with ticks at 0, 0.5, 1, 1.5, 2, 2.5, 3. The curve starts at approximately (0, 1.4), dips to a minimum of about 0.1 at x=0, and rises back to approximately (0, 1.4).	 A plot titled "Geman_Reynolds Influence Function" showing an asymmetric curve. The x-axis ranges from -2 to 2 with ticks at -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2. The y-axis ranges from -3 to 3 with ticks at -3, -2, -1, 0, 1, 2, 3. The curve is zero for  x  > 1. At x=1, it has a sharp peak reaching about 2. For x < -1, it decreases from 0 towards -3. For -1 < x < 1, it decreases from 0 towards -2. For x > 1, it decreases from 0 towards -0.5.

## 4. Properties of Non-Convex Potential Functions

---

### Advantages:

- Preserve edges
- Very general class of potential functions

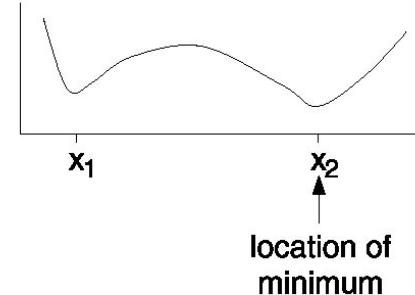
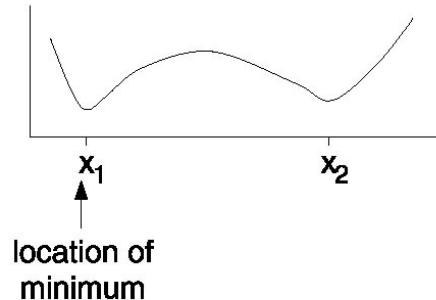
### Disadvantages:

- Difficult to compute some estimation using the above MRF as prior model
- Usually requires the choice of an edge threshold
- Unstable estimations and maxima can change abruptly.

## 4. Properties of Non-Convex Potential Functions

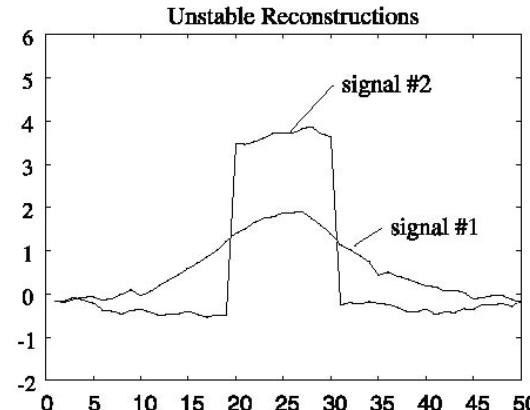
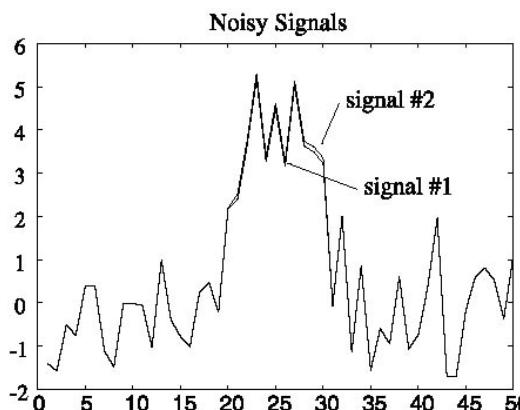
---

- Minimum of non-convex function can change abruptly.

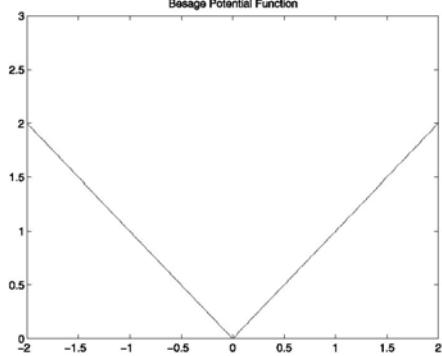
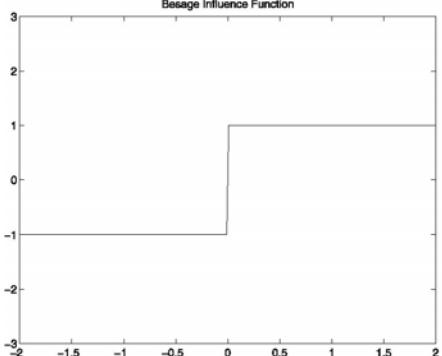
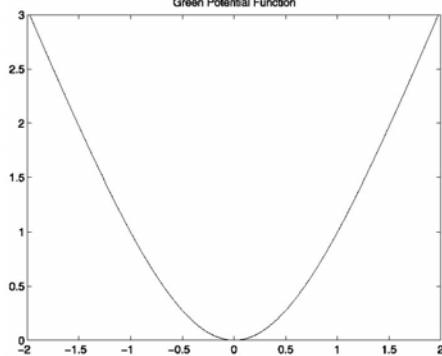
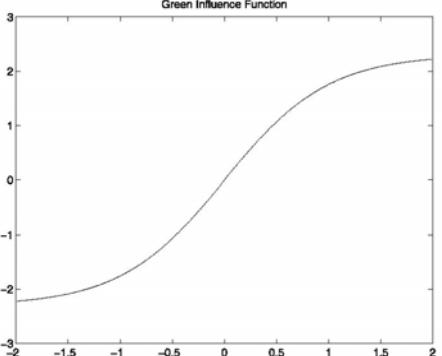


Noise can easily  
change the position  
of minima.

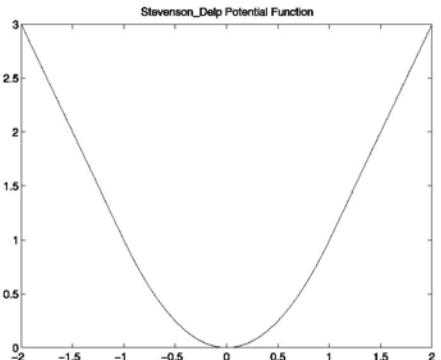
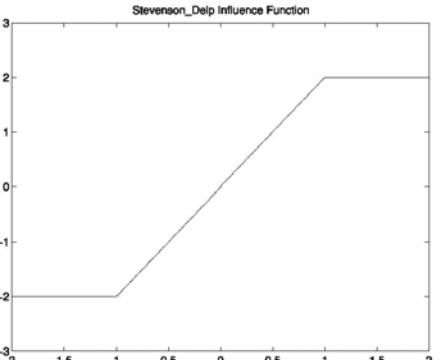
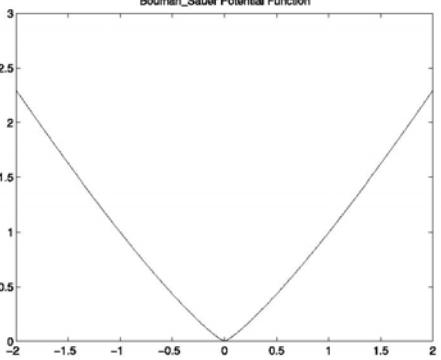
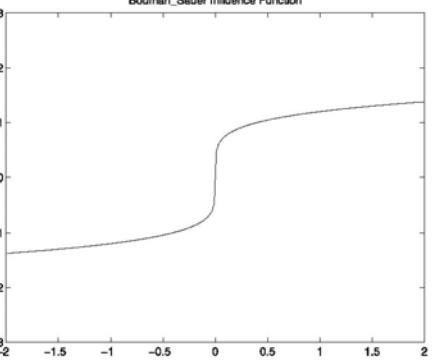
- Example: Blake and Zisserman potential function.



## 4. Non-Gaussian MRFs: convex potential functions

Authors	$\rho(\Delta)$	Potential func.	Influence func.
Besag (Laplacian)	$ \Delta $		
Green	$\log \cosh \Delta$		

## 4. Non-Gaussian MRFs: convex potential functions

Authors	$\rho(\Delta)$	Potential func.	Influence func.
Stevenson&Delp  (Huber, mixture model, epsilon-contaminated)	$\min\left\{\Delta^2, 2 \Delta  - 1\right\}$	 A plot titled "Stevenson_Delp Potential Function" showing a convex curve. The x-axis ranges from -2 to 2, and the y-axis ranges from 0 to 3. The curve is zero at x=0, reaches a minimum of approximately 0.33 at x ≈ ±0.57, and increases quadratically for  x  > 0.57.   A plot titled "Stevenson_Delp Influence Function" showing a piecewise linear function. The x-axis ranges from -2 to 2, and the y-axis ranges from -3 to 3. The function is constant at -2 for x < -1, increases linearly to 0 at x = -1, increases linearly to 2 at x = 1, and is constant at 2 for x > 1.  	
Bouman&Sauer  (Generalized Gaussian MRF, Generalized Gaussian)	$ \Delta ^p$	 A plot titled "Bouman_Sauer Potential Function" showing a convex curve. The x-axis ranges from -2 to 2, and the y-axis ranges from 0 to 3. The curve is zero at x=0, reaches a minimum of 0 at x=0, and increases with a decreasing slope for  x  > 0.   A plot titled "Bouman_Sauer Influence Function" showing a piecewise linear function. The x-axis ranges from -2 to 2, and the y-axis ranges from -3 to 3. The function is constant at -1 for x < 0, increases linearly to 0 at x = 0, and increases with a decreasing slope for x > 0.  	

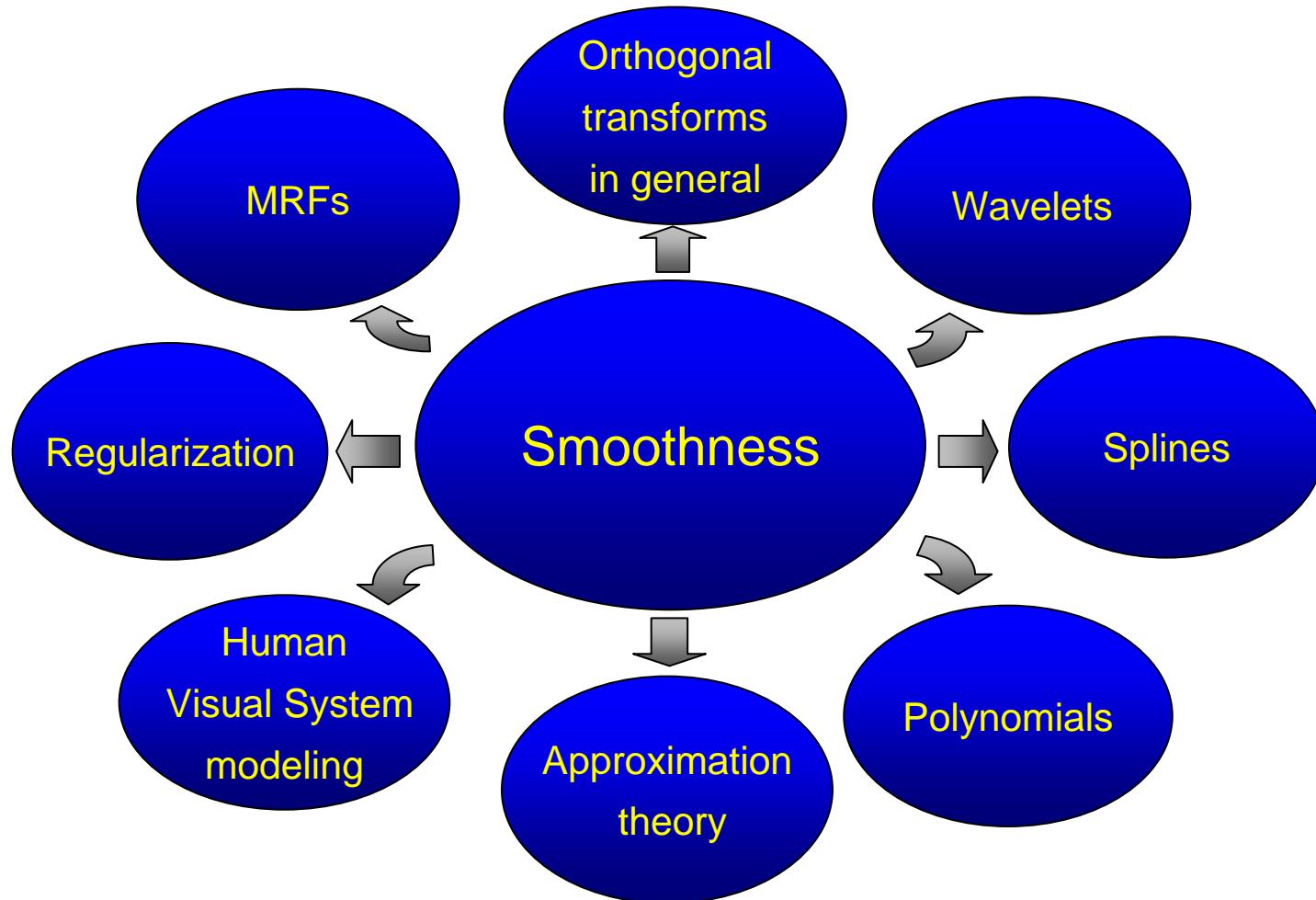
## 4. Properties of Convex Potential Functions

---

- Both  $\log \cosh \Delta$  and Huber functions:
  - Quadratic for  $|\Delta| \ll 1$
  - Linear for  $|\Delta| \gg 1$
  - Transition from quadratic to linear determines edge threshold.
- Generalized Gaussian MRF (GGMRF) functions:
  - Include  $|\Delta|$  function
  - Do not require an edge threshold parameter
  - Convex and differentiable for  $p > 1$ .

## 4. The Smoothness: unified approach

---



## 4. The Smoothness Models: MRFs and regularization

---

The generic contextual constraint on this world is the **smoothness**.

It assumes that physical properties in a neighborhood set present some correlation or generally do not change abruptly.

For spatially continuous MRFs, the smoothness often involves **derivatives**. (it is a case for **analytical regularization** that we will consider in restoration applications).

There exists a strong connection between the theory of MRFs and the regularization in terms of definition of local correlations within the neighborhood system and the order of derivative and its directivity.

## 4. The Smoothness Models: way to regularization

---

### Motivation: (intuitive)

derivatives detect the discontinuities or express the level of smoothness:

- for flat surface the derivative will be equal to zero;
- for edges or transitions, depending on order of derivative and type of transition, the derivative will be different from zero (can change the sign and the shape).

## 4. The Smoothness Models: link to Gaussian MRFs

---

The  $n$ th-order derivative (notations):  $[x^{(n)}(t)]^2$  ← Quadratic, i.e. Gaussian

The order of derivative  $n$  determines the number of sites in the involved cliques.

Example:  $n = 1$   $[x'(t)]^2$  corresponds to a pair-site smoothness potential.

Different orders imply different classes of smoothness.

## 4. The Smoothness Models: link to MRFs

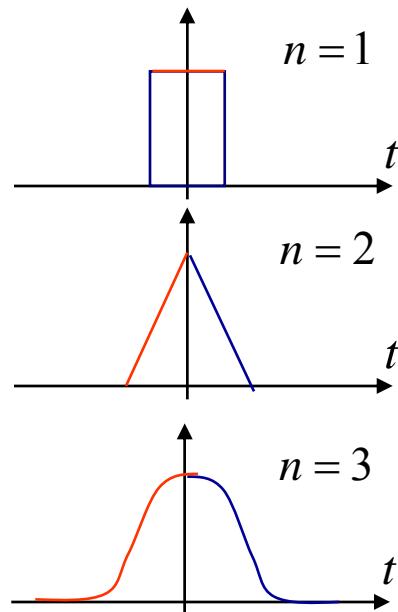
---

The Gibbs distribution  $p(x)$ , or equivalently the energy  $U(x)$ , depends on the type of the surface  $x$  we expect to model.

Consider different models of surface  $x(t)$ ,  $t \in [a, b]$

Generalized surface model

$$x(t) = a_0 + a_1 t + a_2 t^2$$

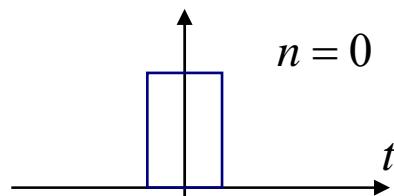


Isn't it very similar to splines?

## 4. The Smoothness Models: flat surface

The flat surface

$$\underline{x(t) = a_0}$$



The first-order derivative equals zero.

$$x'(t) = 0.$$

Energy function

$$U(x) = \int [x'(t)]^2 dt$$

string

The surface that minimizes the energy function has a constant gray level.

## 4. Flat surface: discrete model and link with MRFs

The flat surface

$$U(x) = \int [x'(t)]^2 dt$$

Discrete model

$$U(x) = \sum_i [x_i - x_{i-1}]^2$$



MRF

$$\begin{aligned} U(x) &= \sum_{c \in C} V_c(x) \\ &= \sum_{i \in S} \sum_{i' \in S_i} V_2(x_i, x_{i'}) \end{aligned}$$

$$C = \{(1,2), (2,3), \dots\}$$

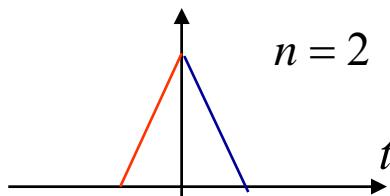
$$V_c(x) = V_2(x_i, x_{i'}) = \frac{1}{2}(x_i - x_{i'})^2$$

2-D model: membrane

## 4. The Smoothness Models: planar surface

The planar surface

$$\underline{x(t) = a_0 + a_1 t}$$



The second-order derivative equals zero.

$$x''(t) = 0.$$

Energy function

$$U(x) = \int [x''(t)]^2 dt$$

rod

The surface that minimizes the energy function has a constant gradient.

## 4. Planar surface: discrete model and link with MRFs

The planar surface

$$U(x) = \int [x''(t)]^2 dt$$

Discrete model

$$U(x) = \sum_i [x_{i+1} - 2x_i + x_{i-1}]^2$$



MRF

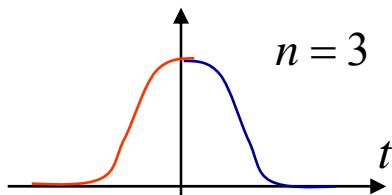
$$\begin{aligned} U(x) &= \sum_{c \in C} V_c(x) \\ &= \sum_{i \in S} \sum_{i' \in S_i} V_2(x_i, x_{i'}) \end{aligned}$$

2-D model: plate

## 4. The Smoothness Models: quadratic surface

The quadratic surface

$$x(t) = a_0 + a_1 t + a_2 t^2$$



The third-order derivative equals zero.

$$x'''(t) = 0.$$

Energy function

$$U(x) = \int [x'''(t)]^2 dt$$

The surface that minimizes the energy function has a constant curvature.

## 4. Quadratic surface: discrete model and link with MRFs

The quadratic surface

$$U(x) = \int [x'''(t)]^2 dt$$

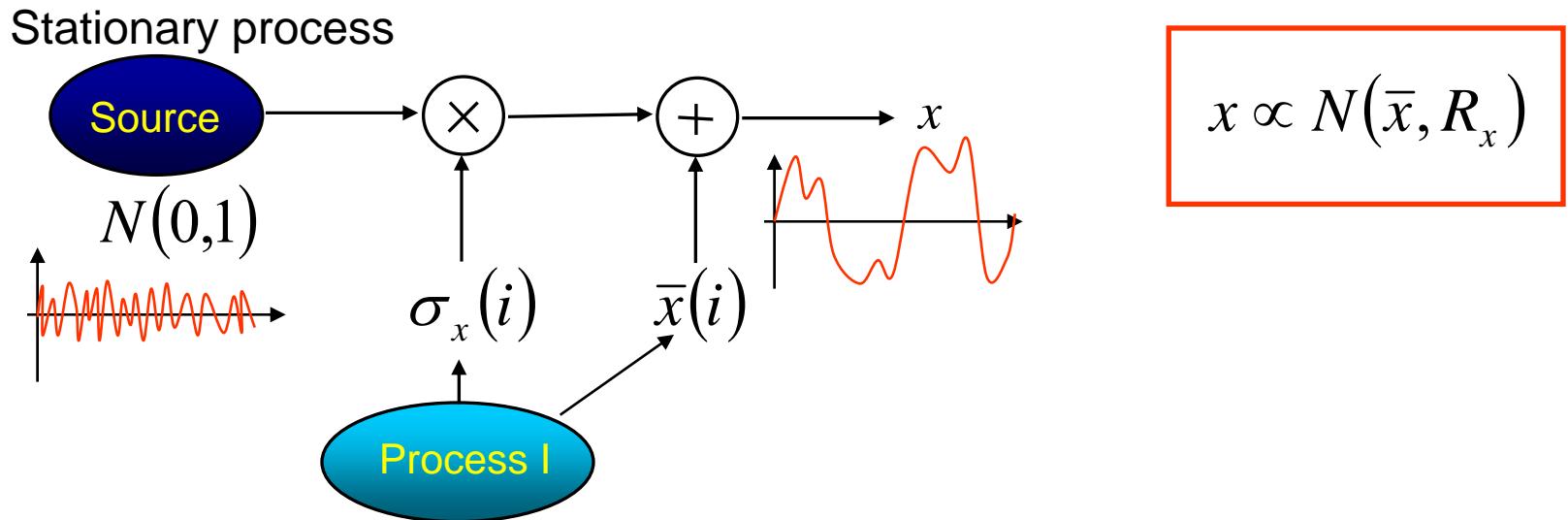
Discrete model

$$U(x) = \sum_i [x_{i+1} - 3x_i + 3x_{i-1} - x_{i-2}]^2$$

1	-3	3	-1
---	----	---	----

2-D model: squared Laplacian

## 5. Double Stochastic Processes: local models



Doubly stochastic process: gamma, exponential, Jeffrey  
 $\sigma_x(i)$  and  $\bar{x}(i)$  are slowly varying.

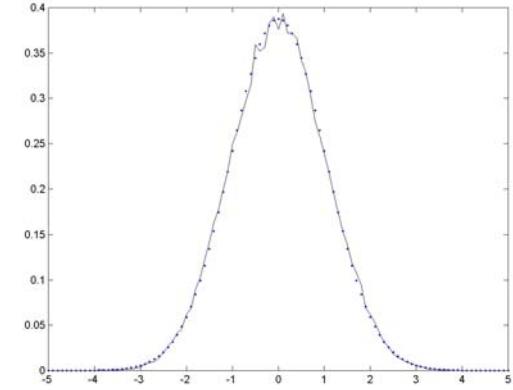
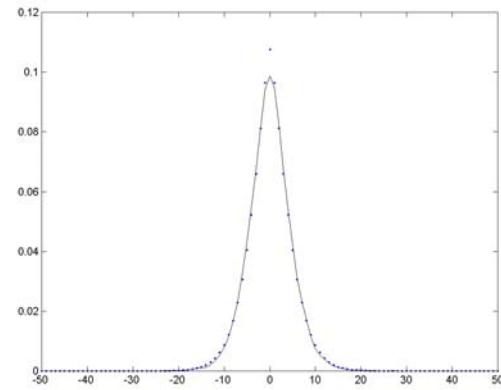
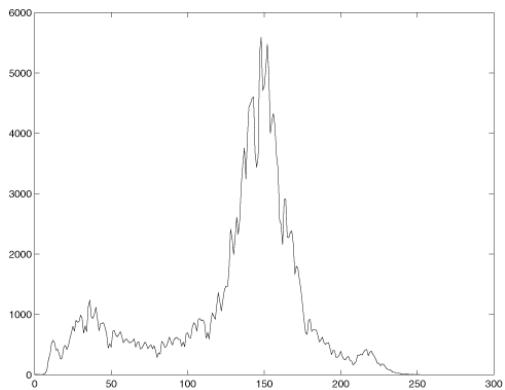
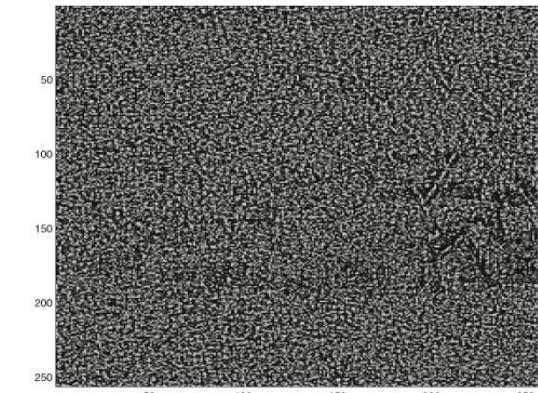
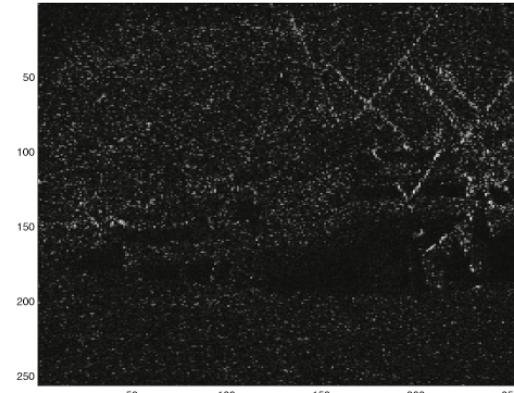
Main problem: estimation of  $\sigma_x^2$  and  $\bar{x}$ .

# 5. Non-stationary Gaussian

Original image



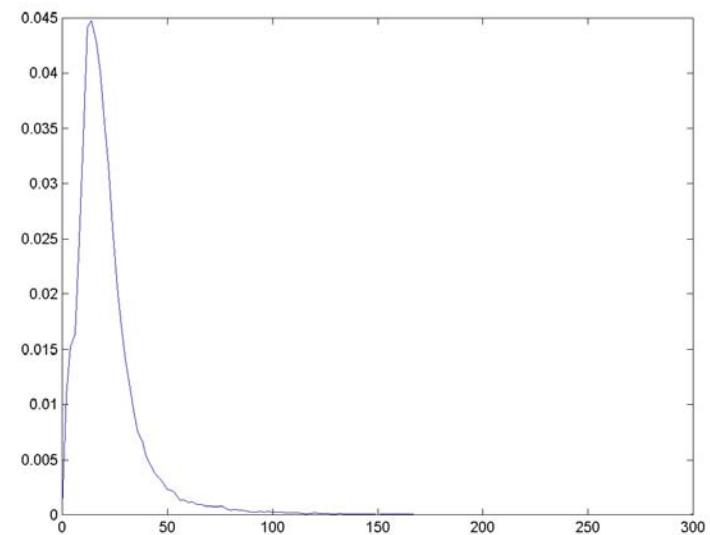
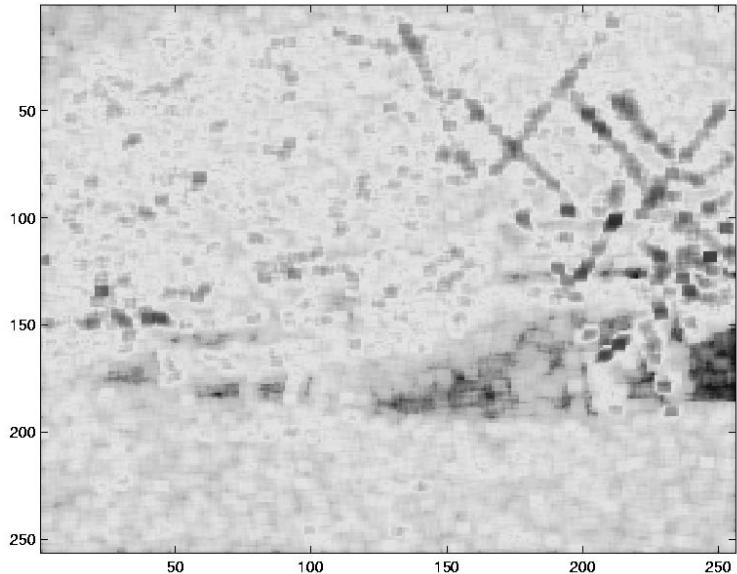
High-pass decorrelated image



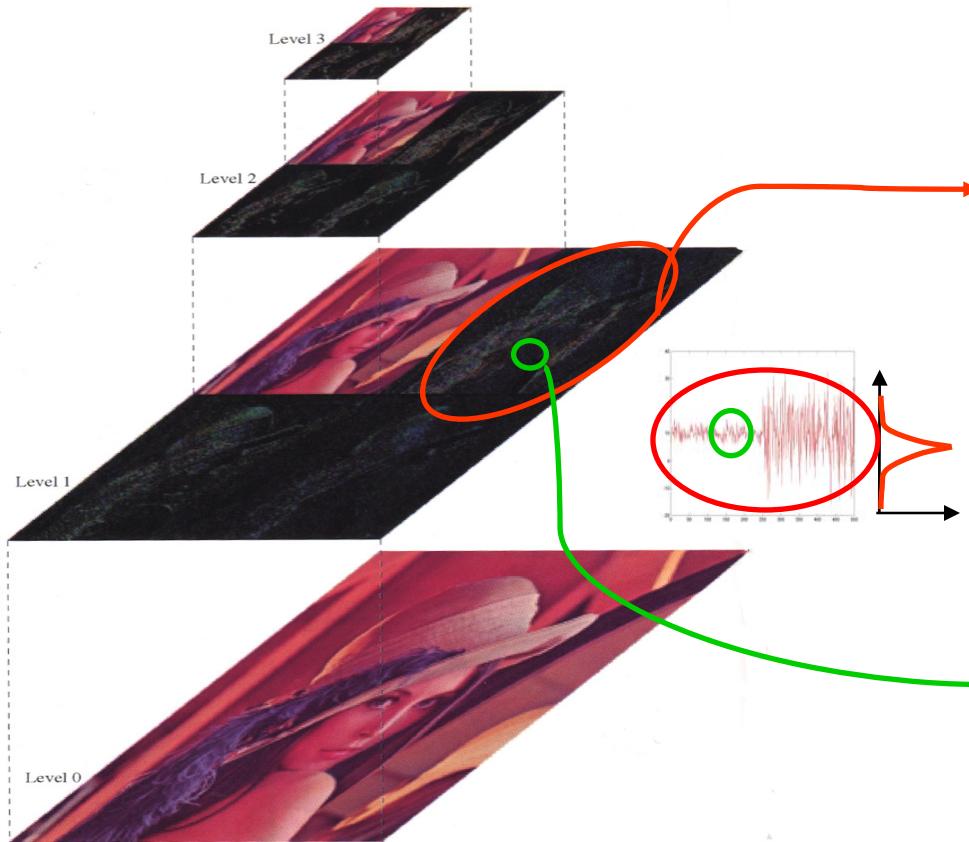
# 5. Non-stationary Gaussian

---

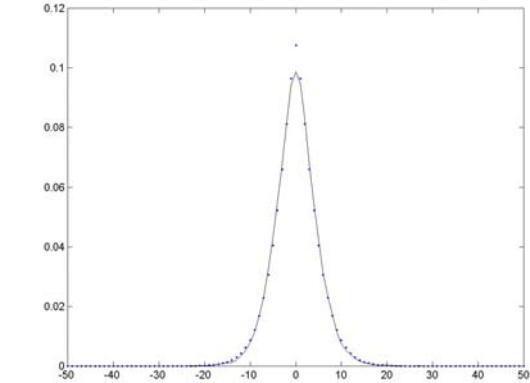
Variance model:



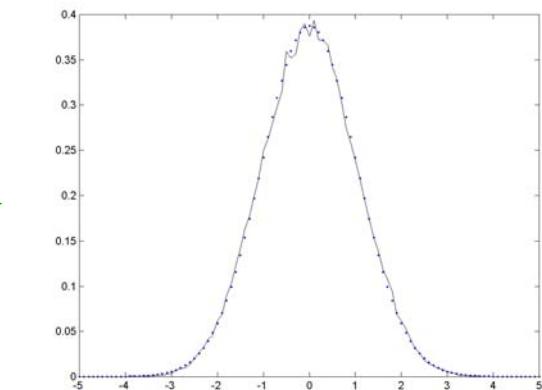
# 5. Non-stationary Gaussian: wavelet domain



Global i.i.d. non-Gaussian model



Local i.i.d. Gaussian model



# 5. Conclusions

---

We have seen that signals are modeled or analyzed using some assumptions about the local signal structure.

This local structure is expressed in terms of:

- signal shape (wavelets, pyramids, smoothness, derivatives and regularization) in scope of deterministic approximation theory;
- local correlations (AR- and MRFs-models) in scope of stochastic signal modeling.

In both cases signal is considered in some “transform” domain where the above local properties are captured.

# 5. Conclusions

---

The goal of the transformation is to decorrelate the signal, i.e. to represent it in more compact and convenient for analysis form.

## The question:

does there exist some generalized framework that makes possible to establish the fundamental link between the properties of the signal (both deterministic and stochastic) and the optimal representation of the signal in the maximum decorrelated form:

- minimization of energy function of MRFs,
- vanishing moments of wavelets,
- order of derivative in regularization.

# 6. The Principal Component Transform

---

The **Principal Component Transform** is also called:

- Karhunen-Loève Transform (KLT)
- Hotelling Transform
- Eigenvector Transform

## 6. KLT: Multivariate Random Signals (Jain, Ch. 5.11)

---

A real time signal  $x(t)$  can be considered as a random process and its samples  $x_m$  ( $m = 0, \dots, N - 1$ ) a random vector:

$$X = [x_0, \dots, x_{N-1}]^T$$

The *mean vector* of  $X$  is

$$M_X \triangleq E(X) = [E(x_0), \dots, E(x_{N-1})]^T = [\mu_{x_0}, \dots, \mu_{x_{N-1}}]^T$$

The *covariance matrix* of  $X$  is

$$R_x = E[(X - M_X)(X - M_X)^T] = E[XX^T] - M_X M_X^T = \begin{bmatrix} \dots & \dots & \dots \\ \dots & \sigma_{ij}^2 & \dots \\ \dots & \dots & \dots \end{bmatrix}$$

where  $\sigma_{ij}^2 \triangleq E(x_i x_j) - \mu_{x_i} \mu_{x_j}$  is the covariance of two random variables  $x_i$  and  $x_j$ . When  $i=j$ ,  $\sigma_{ij}^2$  becomes the variance of  $x_i$ ,  $\sigma_i^2 \triangleq E(x_i^2) - \mu_{x_i}^2$ .

---

## 6. KLT: Multivariate Random Signals

---

The *correlation matrix* of  $X$  is

$$K_x = E[XX^T] = \begin{bmatrix} \dots & \dots & \dots \\ \dots & k_{ij} & \dots \\ \dots & \dots & \dots \end{bmatrix}$$

where  $r_{ij} = \sigma_{ij}^2 + \mu_{x_i}\mu_{x_j}$ .

Both  $R_X$  and  $K_X$  are symmetric matrices (Hermitian if  $X$  is complex).

A signal vector  $X$  can always be easily converted into a zero-mean vector  $\bar{X}=X-M_X$  with all of its information (or dynamic energy) conserved. In the following, without loss of generality, we will assume  $M_X=0$  and therefore  $R_X = K_X$ .

## 6. The Principal Component Transform

---

Let  $\phi_i$  and  $\lambda_i$  be the  $i$ th eigenvector and eigenvalue of the covariance matrix  $R_X$ :

$$R_x \phi_i = \lambda_i \phi_i, \quad (i = 0, \dots, N-1)$$

We can construct an  $N \times N$  matrix  $\Phi$

$$\Phi = (\phi_0 \ \phi_1 \ \dots \phi_{N-1}) = \begin{pmatrix} \phi_{00} & \phi_{10} & \cdots & \phi_{N-10} \\ \phi_{01} & \phi_{11} & \cdots & \phi_{N-11} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{0N-1} & \phi_{1N-1} & \cdots & \phi_{N-1N-1} \end{pmatrix}$$

Since the columns of  $\Phi$  are the eigenvectors of a symmetric (Hermitian if  $X$  is complex) matrix  $R_X$ ,  $\Phi$  is orthogonal (unitary):  $\Phi^T \Phi = I$  i.e.,

$$\Phi^{-1} = \Phi^T$$

## 6. The Principal Component Transform

---

Therefore:

$$R_X \Phi = \Phi \Lambda$$

$$\Lambda = \begin{pmatrix} \lambda_0 & & & \boldsymbol{0} \\ & \lambda_1 & & \\ & & \ddots & \\ \boldsymbol{0} & & & \lambda_{N-1} \end{pmatrix}$$

where  $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{N-1})$ . Or, we have

$$R_Y = \Phi^{-1} R_X \Phi = \Phi^T R_X \Phi = \Lambda$$

We can now define the orthogonal (unitary if  $X$  is complex) *Principal Component Transform* of  $X$  by

$$\begin{cases} Y = \Phi^T X, \\ X = \Phi Y \end{cases}$$

## 6. The Principal Component Transform

---

The  $i$ th component of the forward transform  $Y = \Phi^T X$  is the projection of  $X$  on  $\phi_i$ :

$$y_i = (\phi_i, X) = \phi_i^T X$$

and the inverse transform  $X = \Phi^T Y$  represents  $X$  in the N-dimensional space spanned by  $\phi_i$  ( $i = 0, 1, \dots, N - 1$ )

$$X = \sum_{j=0}^{N-1} y_j \phi_j(i)$$

## 6. KLT Completely Decorrelates the Signal

---

KLT is the optimal orthogonal transform in the following sense:

- KLT completely decorrelates the signal
- KLT optimally compacts the energy (information) contained in the signal.

The first property is simply due to the definition of KLT, and the second property is due to the fact that KLT redistributes the energy among the  $N$  components in such a way that most of the energy is contained in a small number of components of  $Y = \Phi^T X$ .

## 6. KLT Completely Decorrelates the Signal

---

To see the first property, consider the covariance matrix  $R_Y$  of  $Y$ :

$$R_Y = E[YY^T] = E[\Phi^T X (\Phi^T X)^T] = E[\Phi^T X X^T \Phi] = \Phi^T E[X X^T] \Phi = \Phi^T R_X \Phi$$

$$R_Y = \Phi^T R_X \Phi = \Lambda$$

Or in matrix form, we have

$$R_Y = \Phi^T R_X \Phi = \Lambda = \begin{pmatrix} \lambda_0 & & & & \theta \\ & \lambda_1 & & & \\ & & \ddots & & \\ \theta & & & \lambda_{N-1} & \end{pmatrix} = \begin{pmatrix} \sigma_{y_0}^2 & & & & \theta \\ & \sigma_{y_1}^2 & & & \\ & & \ddots & & \\ \theta & & & \sigma_{y_{N-1}}^2 & \end{pmatrix}$$

## 6. KLT Completely Decorrelates the Signal

---

Conclusion:

We see that after KLT, the correlation matrix of the signal is diagonalized, i.e.  $\sigma_{ij} = 0$  , the correlation between any two components  $x_i$  and  $x_j$  is always zero.

In other words, the signal is completely decorrelated.

## 6. KLT Optimally Compacts the Energy

---

Consider a general orthogonal transform pair defined as

$$\begin{cases} Y = A^T X, \\ X = AY \end{cases} \quad A \text{ is not necessarily the KLT.}$$

where  $X$  and  $Y$  are  $N$  by 1 vectors and  $A$  is an arbitrary  $N$  by  $N$  orthogonal matrix  $A^{-1}=A^T$ .

## 6. KLT Optimally Compacts the Energy

---

Now the  $i$ th component of  $Y$  can be written as

$$y_i = (a_i, X) = a_i^T X$$

As we assume the mean vector of  $X$  is zero  $M_X=0$  (and obviously we also have  $M_Y=A^T M_X=0$ ), we have  $R_X = K_X$ , and the variance of the  $i$ th element in both  $X$  and  $Y$  are

$$\sigma_{x_i}^2 = E(x_i^2) \triangleq E(e_{x_i})$$

and

$$\sigma_{y_i}^2 = E(y_i^2) \triangleq E(e_{y_i})$$

where  $e_{x_i} \triangleq x_i^2$  and  $e_{y_i} \triangleq y_i^2$  represent the energy contained in the  $i$ th component of  $X$  and  $Y$ , respectively.

## 6. KLT Optimally Compacts the Energy

---

In other words, the trace of  $R_X$  (the sum of all the diagonal elements of the matrix) represents the expectation of the total amount of energy contained in the signal  $X$

$$\text{Total energy contained in } X = \text{tr } R_X = \sum_{i=0}^{N-1} \sigma_{x_i}^2 = \sum_{i=0}^{N-1} E(x_i^2) = E\left(\sum_{i=0}^{N-1} e_{x_i}\right)$$

Since an orthogonal transform  $A$  does not change the length of a vector  $X$ , i.e.  $\|Y\| = \|AX\| = \|X\|$ , where

$$\|X\| \triangleq \sqrt{\sum_{i=0}^{N-1} x_i^2} = \sqrt{\sum_{i=0}^{N-1} e_{x_i}}$$

the total energy contained in the signal vector  $X$  is conserved after the orthogonal transform.

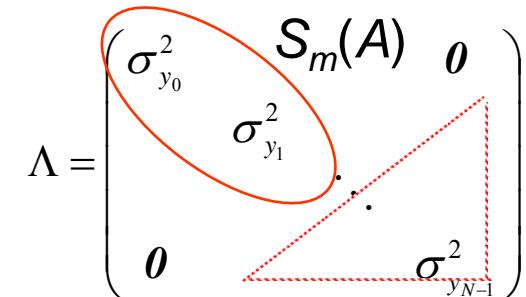
## 6. KLT Optimally Compacts the Energy

We next define

$$S_m(A) \triangleq \sum_{i=0}^{m-1} E(y_i^2) = \sum_{i=0}^{m-1} \sigma_{y_i}^2 = \sum_{i=0}^{m-1} E(e_{y_i})$$

where  $m \leq N$ .

$S_m(A)$  is a function of the transform matrix  $A$  and represents the amount of energy contained in the first  $m$  components of  $Y=AX$ .



Since the total energy is conserved,  $S_m(A)$  also represents the percentage of energy contained in the first  $m$  components.

In the following we will show that  $S_m(A)$  is maximized if and only if the transform  $A$  is the KLT:

$$S_m(\Phi) \geq S_m(A)$$

i.e., KLT optimally compacts energy into a few components of the signal.

## 6. Optimality of KLT

---

Example: Establish that only the KLT has maximum decorrelation and energy compaction properties in comparison with any orthogonal transform.

Consider the first-order Markov model:

$$R_x = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

KLT:  $\Phi$                           Any orthogonal transform:  $A$

$$R_Y = \begin{pmatrix} \lambda_0 & & \theta \\ & \lambda_1 & \\ & & \ddots \\ \theta & & \lambda_{N-1} \end{pmatrix}$$
$$R_Y = \begin{pmatrix} \lambda_0 & & \beta \\ & \lambda_1 & \\ & & \ddots \\ \beta & & \lambda_{N-1} \end{pmatrix}$$

Residual correlation

## 6. Optimality of KLT

---

$$R_x = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

KLT:  $\Phi$

Any orthogonal transform:



$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$A = \frac{1}{2} \begin{pmatrix} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{pmatrix}$$

Eigenvectors are found as solution of:

$$R_x \phi_i = \lambda_i \phi_i$$

Eigenvalues are found as all roots of:

$$\det|R_x - \lambda_i I| = 0$$

Find

$$Y = \Phi X$$
$$R_Y = \Phi^T R_X \Phi$$
$$Y = AX$$
$$R_Y = A^T R_X A$$


## 6. Optimality of KLT

KLT:  $\Phi$

$$R_Y = \Phi^T R_X \Phi = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \boxed{\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$R_Y = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

Any orthogonal transform:

$$R_Y = A^T R_X A = \frac{1}{4} \begin{pmatrix} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{pmatrix} \boxed{\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}} \begin{pmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{pmatrix}$$

$$R_Y = \begin{pmatrix} 1 + \frac{\sqrt{3}}{2}\rho & \frac{\rho}{2} \\ \frac{\rho}{2} & 1 - \frac{\sqrt{3}}{2}\rho \end{pmatrix}$$

## 6. Optimality of KLT: main properties

KLT: $\Phi$	Any orthogonal transform:
$R_Y = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$	$R_Y = \begin{pmatrix} 1+\frac{\sqrt{3}}{2}\rho & \frac{\rho}{2} \\ \frac{\rho}{2} & 1-\frac{\sqrt{3}}{2}\rho \end{pmatrix}$

1. Energy before and after both transforms are preserved (assume  $\rho \rightarrow 1$ ).

$Energy_x = \text{tr}R_X \sum_{i=0}^{N-1} \sigma_{x_i}^2$	$\sigma_{x_o}^2 = \rho = 1$	$Energy_x = 1+1=2$
$\sigma_{y_o}^2 = 1+\rho = 2$ $\sigma_{y_1}^2 = 1-\rho = 0$	$Energy_y = 2+0=2$	$\sigma_{y_o}^2 = 1+\frac{\sqrt{3}}{2}\rho = 1.866$ $\sigma_{y_1}^2 = 1-\frac{\sqrt{3}}{2}\rho = 0.134$ $Energy_y = 1.87 + 0.13 = 2$

## 6. Optimality of KLT: main properties

KLT:  $\Phi$

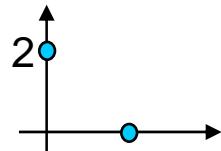
$$R_Y = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

Any orthogonal transform:

$$R_Y = \begin{pmatrix} 1+\frac{\sqrt{3}}{2}\rho & \frac{\rho}{2} \\ \frac{\rho}{2} & 1-\frac{\sqrt{3}}{2}\rho \end{pmatrix}$$

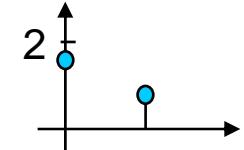
2. Energy compaction in the KLT case is higher (assume  $\rho \rightarrow 1$  ).

$$\begin{aligned}\sigma_{x_o}^2 &= \rho = 1 \\ \sigma_{x_l}^2 &= \rho = 1 \\ \sigma_{y_o}^2 &= 1 + \rho = 2 \\ \sigma_{y_l}^2 &= 1 - \rho = 0\end{aligned}$$



$$\begin{aligned}\sigma_{x_o}^2 &= \rho = 1 \\ \sigma_{x_l}^2 &= \rho = 1\end{aligned}$$

$$\begin{aligned}\sigma_{y_o}^2 &= 1 + \frac{\sqrt{3}}{2}\rho = 1.866 \\ \sigma_{y_l}^2 &= 1 - \frac{\sqrt{3}}{2}\rho = 0.134\end{aligned}$$



## 6. Optimality of KLT: main properties

---

KLT:  $\Phi$

$$R_Y = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

Any orthogonal transform:

$$R_Y = \begin{pmatrix} 1 + \frac{\sqrt{3}}{2}\rho & \frac{\rho}{2} \\ \frac{\rho}{2} & 1 - \frac{\sqrt{3}}{2}\rho \end{pmatrix}$$

3. Complete decorrelation in the KLT case.

---

$$R_Y = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

$$R_Y = \begin{pmatrix} 1 + \frac{\sqrt{3}}{2}\rho & \frac{\rho}{2} \\ \frac{\rho}{2} & 1 - \frac{\sqrt{3}}{2}\rho \end{pmatrix}$$

## 6. Comparison with Other Orthogonal Transforms

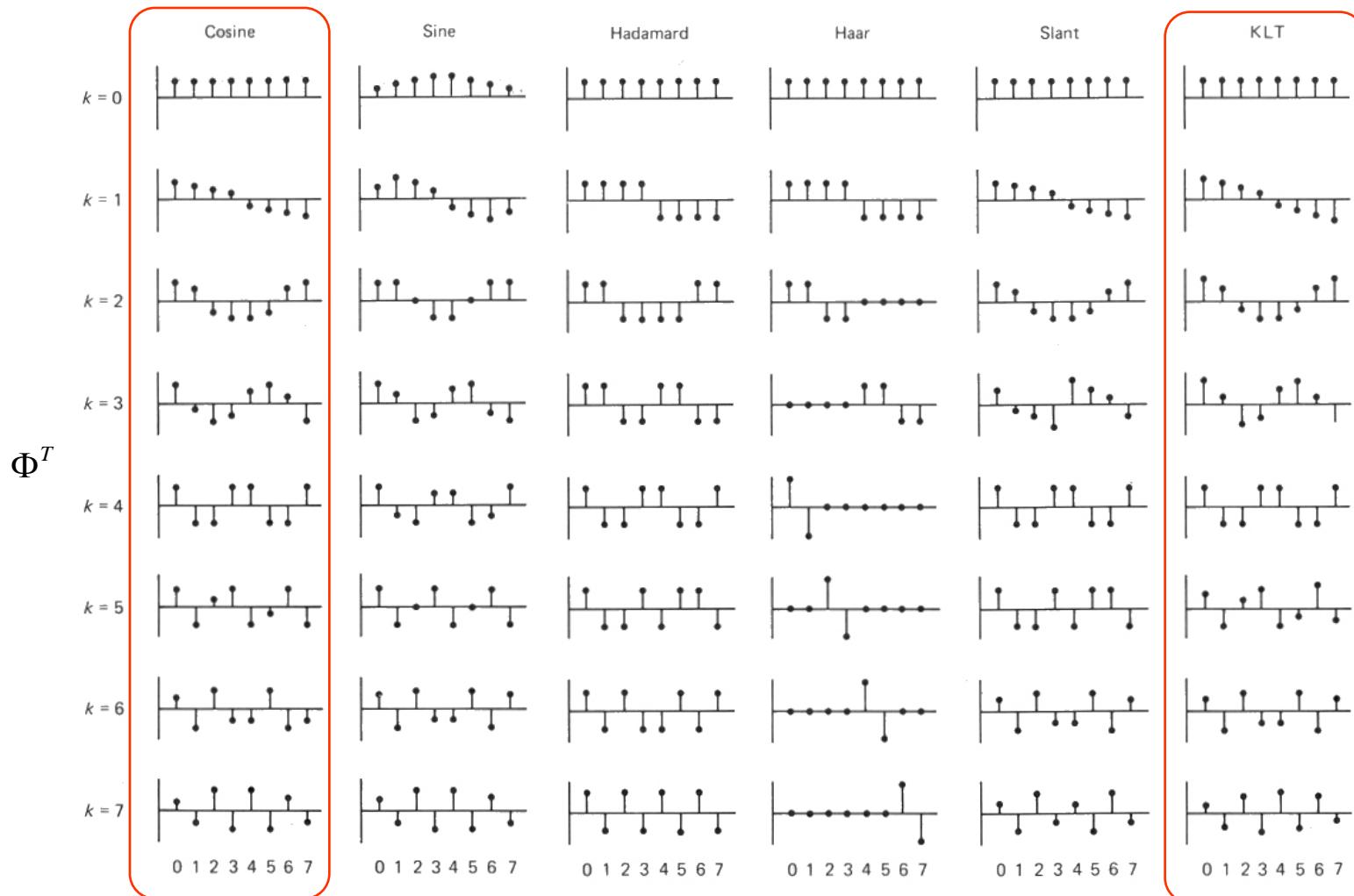
---

Other orthogonal transforms used in image processing:

- Discrete Cosine Transform (DCT);
- Discrete Sine Transform (DST);
- Discrete Fourier Transform (DFT);
- Haar Transform;
- Hadamard Transform;
- Slant Transform;
- Wavelet Transform.

Jain, Ch.5

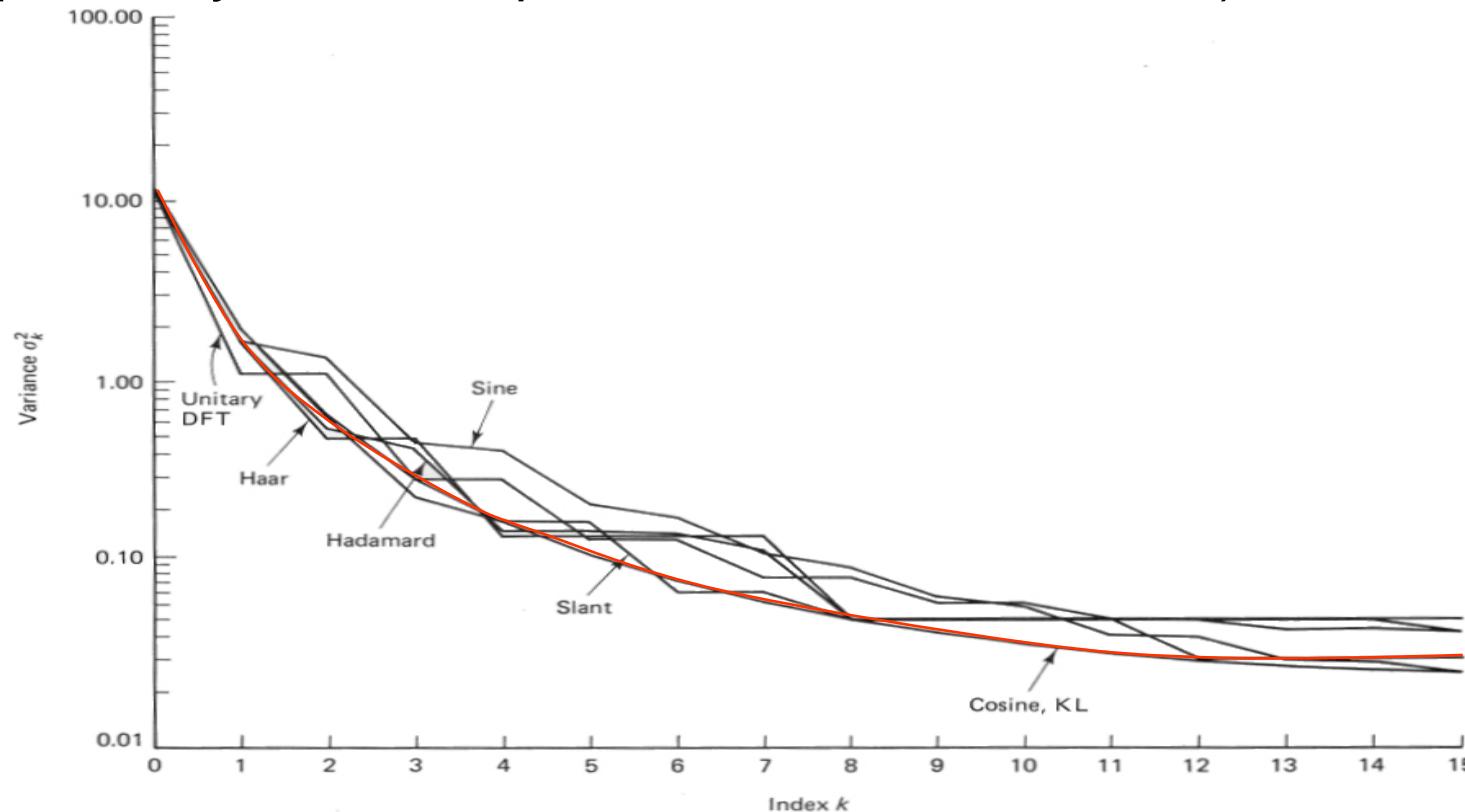
# 6. Basis Functions of Orthogonal Transforms



## 6. Comparison with Other Orthogonal Transforms

Here we compare KLT with other orthogonal transforms.

Different orthogonal transforms  $Y = A^T X$  are applied to the signal  $X$  (*stationary Markov sequence with  $N=16$  and  $\rho = 0.95$* ).



Jain, Ch.5

## 6. Geometric Interpretation of KLT

---

Assume the  $N$  random variables in  $X = [x_0, \dots, x_{N-1}]^T$  have a normal joint probability density function:

$$p(x_0, \dots, x_{N-1}) = N(X, M_X, \Sigma_X) = \frac{1}{(2\pi)^{N/2} |\Sigma_X|^{1/2}} \exp\left[-\frac{1}{2}(X - M_X)^T \Sigma_X^{-1} (X - M_X)\right]$$

where  $M_X$  and  $\Sigma_X$  are the mean vector and covariance matrix of  $X$ , respectively. When  $N=1$ ,  $\Sigma_X$  and  $M_X$  become  $\sigma_x$  and  $\mu_x$ , respectively, and the density function becomes single variable normal distribution.

The shape of this normal distribution in the  $N$ -dimensional space can be found by considering the iso-value hyper-surface in the space determined by equation

$$N(X, M_X, \Sigma_X) = c_0$$

where  $c_0$  is a constant.

## 6. Geometric Interpretation of KLT

---

Equivalently, this equation can be written as

$$(X - M_X)^T \Sigma_X^{-1} (X - M_X) = c_1$$

where  $c_1$  is another constant related to  $c_0$ ,  $M_X$  and  $\Sigma_X$ . In particular, with  $N=2$  variables  $x_0$  and  $x_1$ , we have

$$\begin{aligned}(X - M_X)^T \Sigma_X^{-1} (X - M_X) &= [x_0 - \mu_{x_0}, x_1 - \mu_{x_1}] \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x_0 - \mu_{x_0} \\ x_1 - \mu_{x_1} \end{bmatrix} \\ a(x_0 - \mu_{x_0})^2 + b(x_0 - \mu_{x_0})(x_1 - \mu_{x_1}) + c(x_1 - \mu_{x_1})^2 &= c_1\end{aligned}$$

Here we have assumed

$$\Sigma_X^{-1} = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \quad N(X, M_X, \Sigma_X) = c_0$$

The above quadratic equation represents an ellipse.

## 6. Geometric Interpretation of KLT

---

When  $X = [x_0, \dots, x_{N-1}]^T$  is completely decorrelated by KLT:

$$Y = \Phi^T X$$

the covariance matrix becomes diagonalized:

$$\Sigma_Y = \Lambda = \begin{bmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{N-1} \end{bmatrix} = \begin{bmatrix} \sigma_{y_0}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y_1}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{y_{N-1}}^2 \end{bmatrix}$$

and equation  $N(X, M_X, \Sigma_X) = c_0$  becomes  $N(Y, M_Y, \Sigma_Y) = c_0$ , or

$$(Y - M_Y)^T \Sigma_Y^{-1} (Y - M_Y) = \sum_{i=0}^{N-1} \frac{(y_i - \mu_{y_i})^2}{\lambda_i} = \sum_{i=0}^{N-1} \frac{(y_i - \mu_{y_i})^2}{\sigma_{y_i}^2} = c_1$$

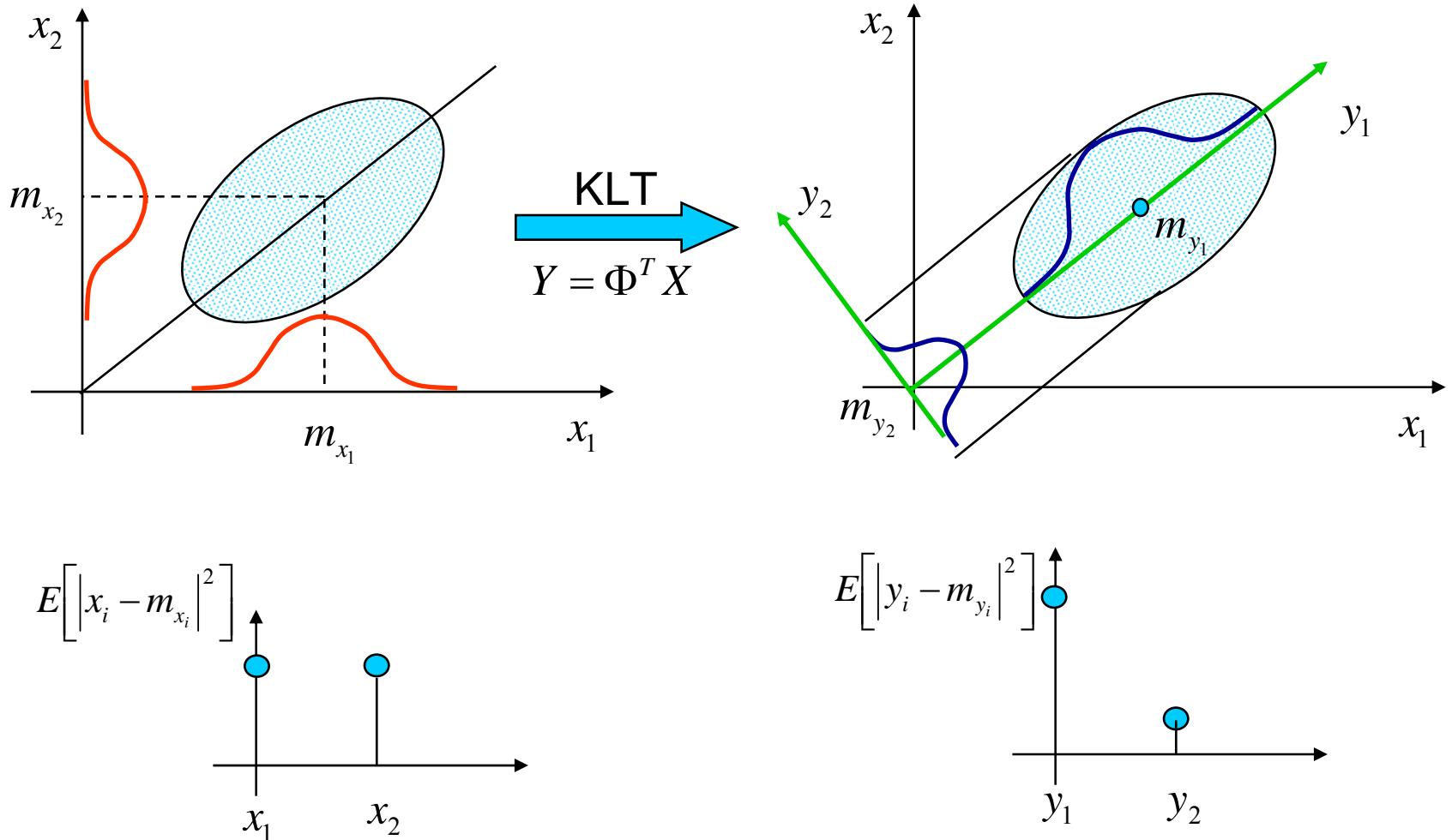
## 6. Geometric Interpretation of KLT

---

This equation represents a standard ellipsoid in the N-dimensional space. In other words, KLT  $Y = \Phi^T X$  rotates the coordinate system so that the ellipsoid associated with the normal distribution of  $X$  becomes a standardized ellipsoid associated with the normal distribution of  $Y$ , whose axes are parallel to  $\phi_i$  ( $i = 0, \dots, N - 1$ ), the axes of the new coordinate system, with the corresponding semi axes equal to  $\sqrt{\lambda_i} = \sigma_{y_i}$ .

The standardization of the ellipsoid is the essential reason why KLT has the two desirable properties: (a) decorrelation and (b) compaction of energy, as illustrated in the figure.

## 6. Geometric Interpretation of KLT



## 6. Extension to 2-D

---

If an  $N \times N$  image  $x(m, n)$  is represented by a random field whose autocorrelation function is given by

$$E[x(m, n)x(m', n')] = r(m, n; m', n') \quad 0 \leq m, m', n, n' \leq N - 1$$

then the basis images of the KLT transform are the orthonormalized eigenfunctions  $\psi_{k,l}(m, n)$  obtained by solving

$$\sum_{m'=0}^{N-1} \sum_{n'=0}^{N-1} r(m, n; m', n') \psi_{k,l}(m', n') = \lambda_{k,l} \psi_{k,l}(m, n) \quad 0 \leq m, n, k, l \leq N - 1$$

In matrix form:

$$\mathcal{R}\psi_i = \lambda_i \psi_i \quad \mathcal{R} = E[XX^T]$$

where  $\psi_i$  is an  $N^2 \times 1$  vector representation of  $\psi_{k,l}(m, n)$

$\mathcal{R}$  is  $N^2 \times N^2$  autocorrelation matrix of the image ( $x(m, n) \rightarrow \text{vector } x$ )

## 6. Extension to 2-D

---

If  $\mathfrak{R}$  is separable, then the  $N^2 \times N^2$  matrix  $\Psi$  whose columns are  $\{\psi_i\}$  becomes separable.

$$r(m, n; m', n') = r_1(m, m')r_2(n, n')$$

$$\psi_{k,l}(m, n) = \phi_1(m, k)\phi_2(n, l)$$

In matrix form:  $\mathfrak{R} = R_1 \otimes R_2$ ;  $\Psi = \Phi_1 \otimes \Phi_2$

$$\Phi_j R_j \Phi^T j = \Lambda_j, j = 1, 2$$

The KLT of  $x(m, n)$  is:

$$y = \Psi^T x$$

## 6. Extension to 2-D

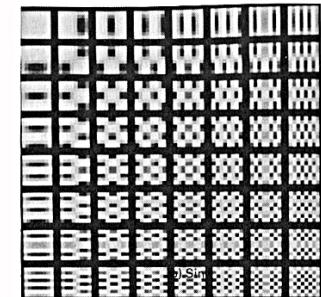
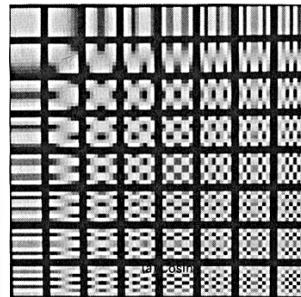
---

$$r(m, n; m', n') = \rho^{|m-m'|} \rho^{|n-n'|}$$

$$\rho = 0.95$$

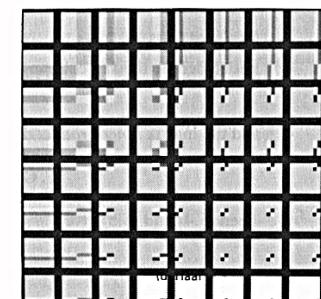
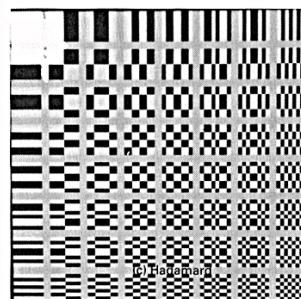
8x8

Cosine



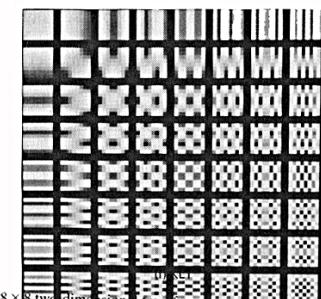
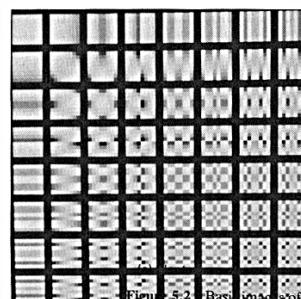
Sine

Hadamard



Haar

Slant



KLT

# 6. Karhunen-Loève Transform: Conclusions

---

## Conclusions:

The KLT is decorrelating transform for **Wide Sense Stationary** process when the second order statistics (correlation matrix) are known.

## Drawbacks of the KLT

Need:

- The correlation matrix (optimal basis) for every image.
- Side information ( $R_x$ ) for inverse transform.
- Local assumption about WSS (not true for edges and textures): trade-off window size.
- KLT is not separable for image blocks.
- Transformed matrix cannot be factored into sparse matrices.

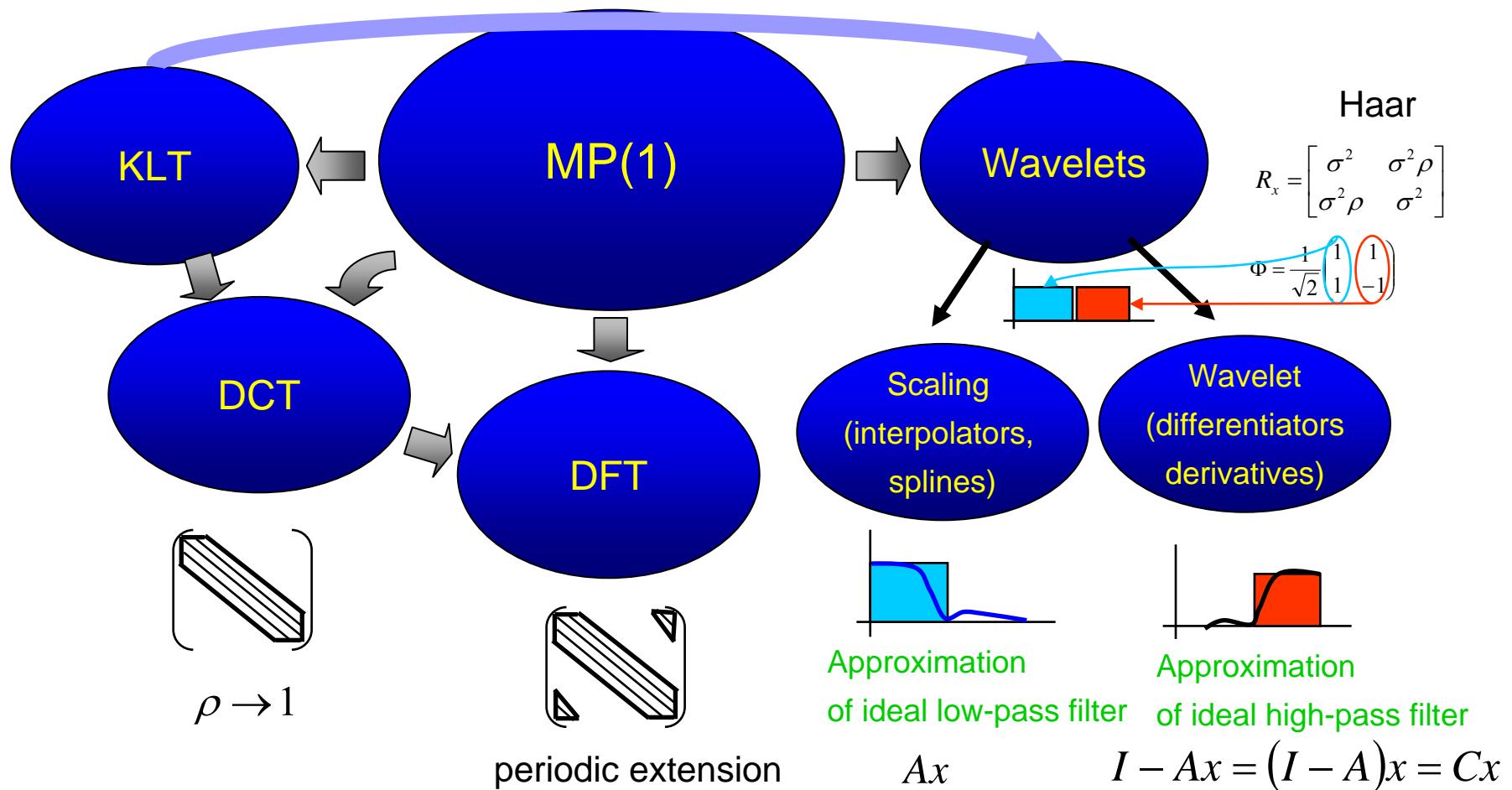
## 6. Karhunen-Loève Transform: Conclusions

---

### Advantages of the KLT:

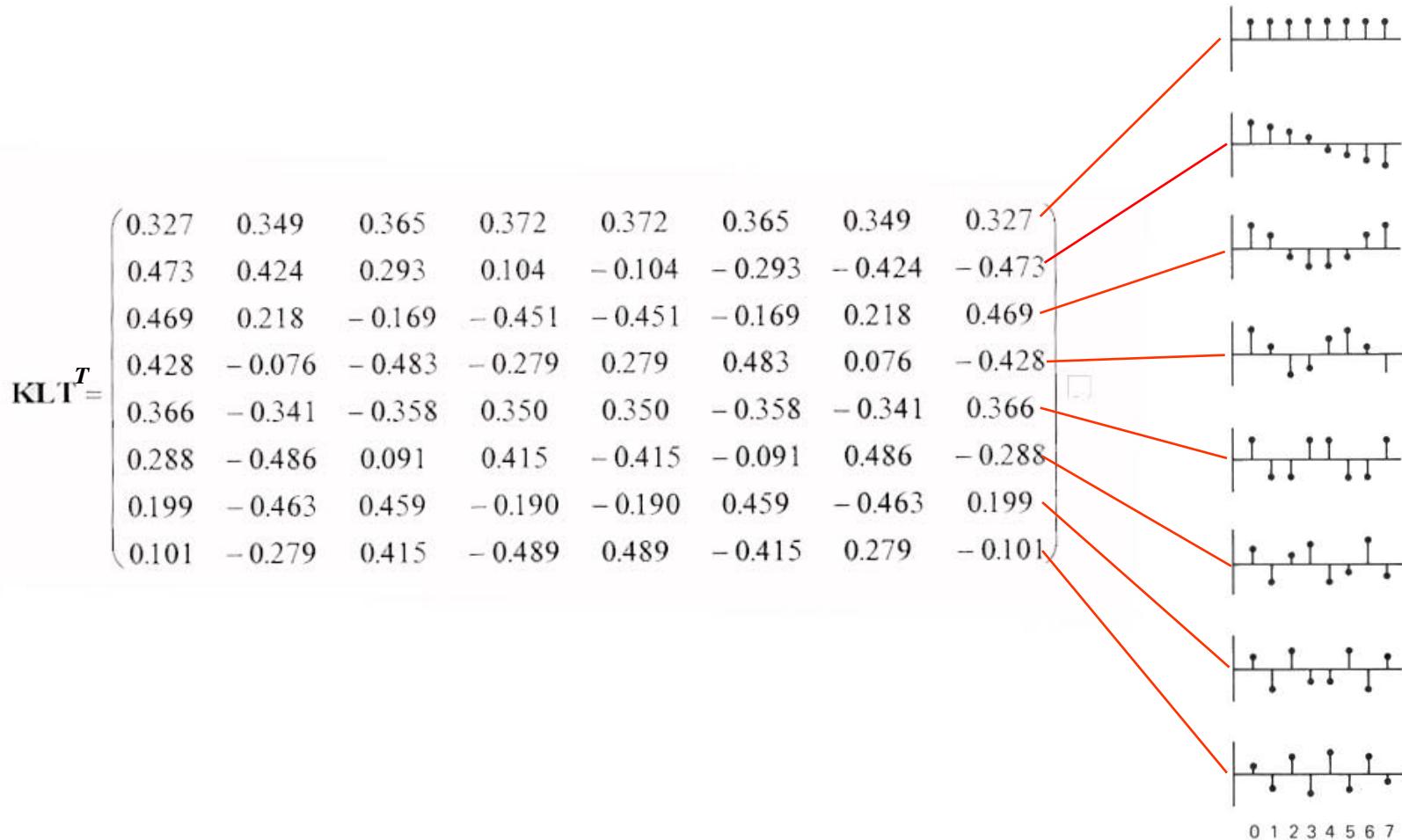
- Completely decorrelates signal (removes statistical dependencies between pixels).
- Optimally compacts the energy.
- Easy to invert: use separable 2D transforms.
- Leads to generic orthogonal transforms like DFT, DCT and wavelets.

# 7. Image Modeling: Fundamental Links



# 7. Image Modeling: Fundamental Links

---



# 7. Image Modeling: Fundamental Links

Compare:

$T$ $KLT =$	0.327	0.349	0.365	0.372	0.372	0.365	0.349	0.327
	0.473	0.424	0.293	0.104	-0.104	-0.293	-0.424	-0.473
	0.469	0.218	-0.169	-0.451	-0.451	-0.169	0.218	0.469
	0.428	-0.076	-0.483	-0.279	0.279	0.483	0.076	-0.428
	0.366	-0.341	-0.358	0.350	0.350	-0.358	-0.341	0.366
	0.288	-0.486	0.091	0.415	-0.415	-0.091	0.486	-0.288
	0.199	-0.463	0.459	-0.190	-0.190	0.459	-0.463	0.199
	0.101	-0.279	0.415	-0.489	0.489	-0.415	0.279	-0.101
	0.354	0.354	0.354	0.354	0.354	0.354	0.354	0.354
$T$ $C(8) =$	0.490	0.416	0.278	0.098	-0.098	-0.278	-0.416	-0.490
	0.462	0.191	-0.191	-0.462	-0.462	-0.191	0.191	0.462
	0.416	-0.098	-0.490	-0.278	0.278	0.490	0.098	-0.416
	0.354	-0.354	-0.354	0.354	0.354	-0.354	-0.354	0.354
	0.278	-0.490	0.098	0.416	-0.416	-0.098	0.490	-0.278
	0.191	-0.462	0.462	-0.191	-0.191	0.462	-0.462	0.191
	0.098	-0.278	0.416	-0.490	0.490	-0.416	0.278	-0.098
	0.354	0.354	0.354	0.354	0.354	0.354	0.354	0.354
	0.490	0.416	0.278	0.098	-0.098	-0.278	-0.416	-0.490

## 7. Image Modeling: Fundamental Links

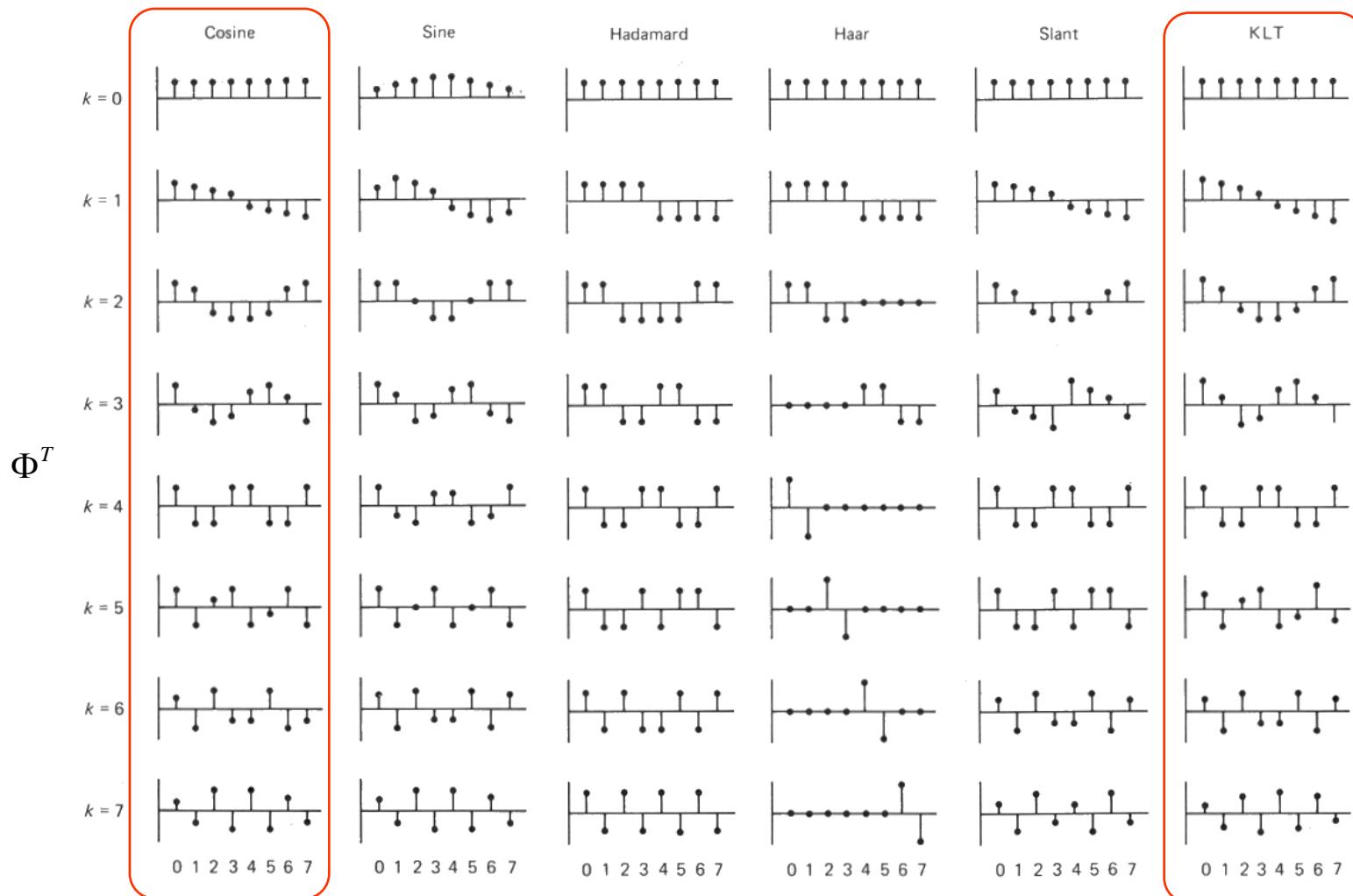
---

- For DFT of order  $N$ , define  $W_N = \exp\left\{\frac{-j2\pi}{N}\right\}$
- Transform matrix

$$\mathbf{DFT}_N^T = \frac{1}{\sqrt{N}} \begin{pmatrix} W_N^0 & W_N^0 & W_N^0 & \dots & W_N^0 \\ W_N^0 & W_N^1 & W_N^2 & \dots & W_N^{N-1} \\ W_N^0 & W_N^2 & W_N^4 & \dots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_N^0 & W_N^{N-1} & W_N^{2(N-1)} & \dots & W_N^{(N-1)(N-1)} \end{pmatrix}$$

- Definition for general  $N$ , fast algorithms for  $N=2^m$
- DFT coefficients are complex (even if input image is real)
- Inverse transform  $\mathbf{DFT}_N^{-1} = \mathbf{DFT}^H = \mathbf{DFT}^*$

# 7. Image Modeling: Fundamental Links



## 7. A Sinusoidal Family of Unitary Transforms

---

- Covariance matrix  $R_x$  and its inverse have an identical set of eigenvectors. Therefore, the KLT associated with  $R_x$  can be determined from the eigenvectors of inverse matrix.
- Consider the generalized inverse matrix:

$$J = J(k_1, k_2, k_3) = \begin{bmatrix} 1 - k_1\alpha & -\alpha & & k_3\alpha \\ -\alpha & 1 & 0 & -\alpha \\ & 1 & -\alpha & \\ k_3\alpha & 0 & -\alpha & 1 - k_2\alpha \end{bmatrix}$$

## 7. A Sinusoidal Family of Unitary Transforms

---

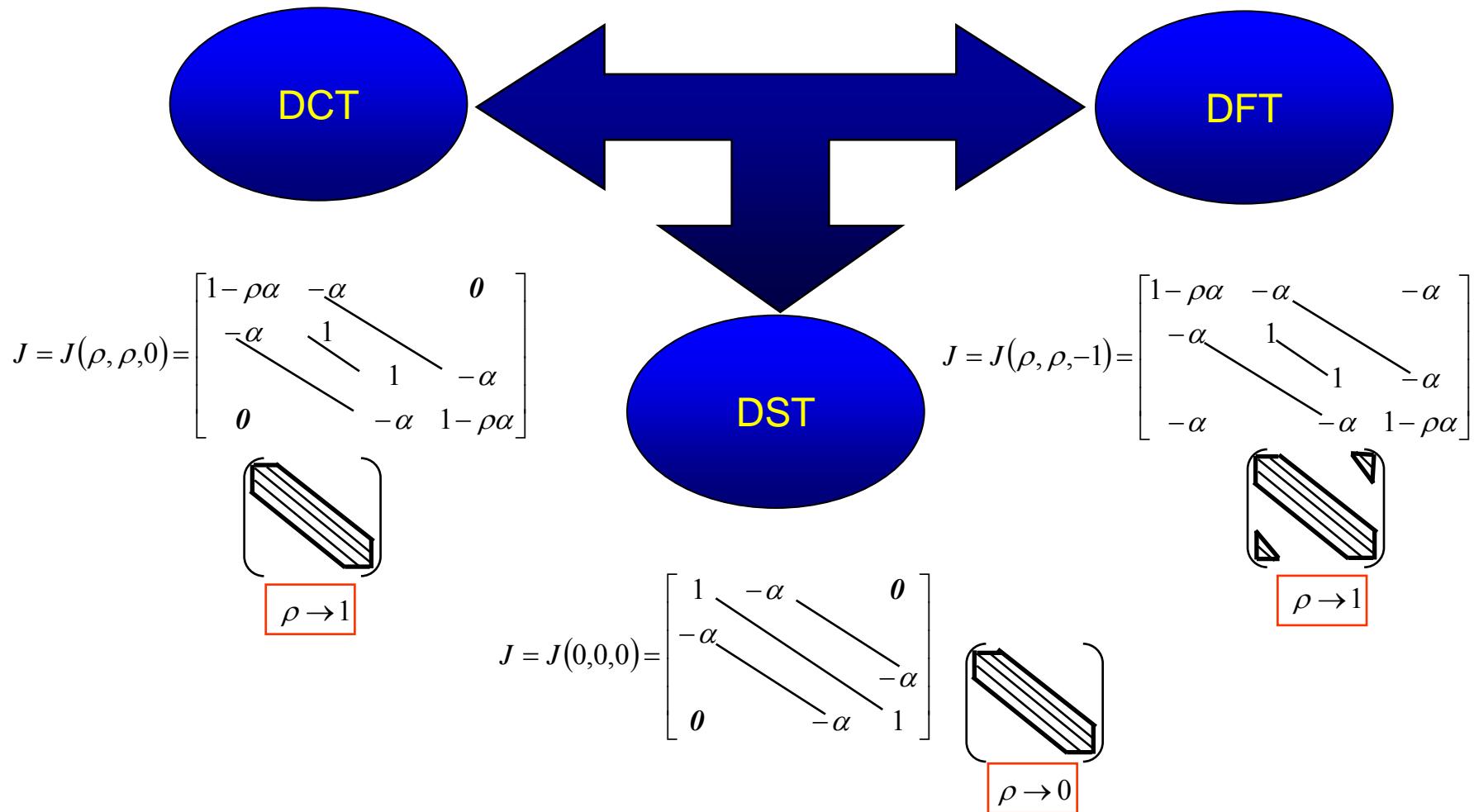
The first-order stationary Markov sequence:

$$R_x = \begin{bmatrix} 1 & \rho & \rho^2 \cdots & \rho^{N-1} \\ \rho & 1 & \cdots & \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \cdots & 1 \end{bmatrix}$$

Inverse matrix of the first-order stationary Markov sequence.

$$\beta^2 R^{-1} = \begin{bmatrix} 1 - \rho\alpha & -\alpha & & \theta \\ -\alpha & 1 & & \\ & & 1 & -\alpha \\ 0 & & -\alpha & 1 - \rho\alpha \end{bmatrix}$$
$$\beta^2 = \frac{(1 - \rho^2)}{(1 + \rho^2)} \quad \alpha = \frac{\rho}{(1 + \rho^2)}$$

# 7. A Sinusoidal Family of Unitary Transforms

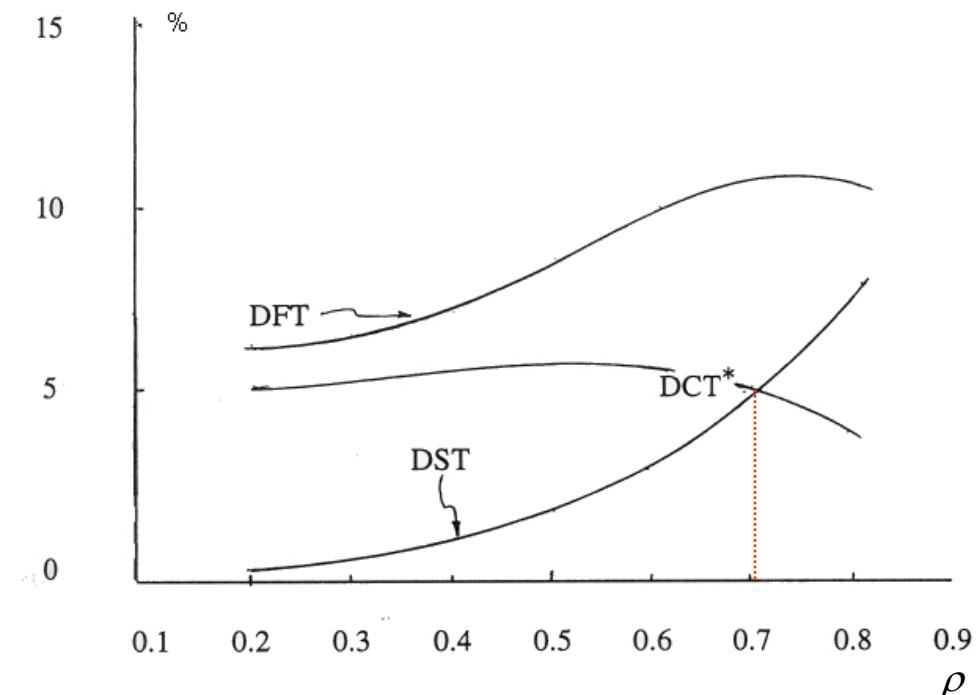


# 7. A Sinusoidal Family of Unitary Transforms

Variance Distributions for  $N = 16, \rho = 0.9$ .

$i$	DCT*	DST	DFT
0	9.835	9.218	9.835
1	2.933	2.640	1.834
2	1.211	1.468	1.834
3	0.581	0.709	0.519
4	0.348	0.531	0.519
5	0.231	0.314	0.250
6	0.166	0.263	0.250
7	0.129	0.174	0.155
8	0.105	0.153	0.155
9	0.088	0.110	0.113
10	0.076	0.099	0.113
11	0.068	0.078	0.091
12	0.062	0.071	0.091
13	0.057	0.061	0.081
14	0.055	0.057	0.081
15	0.053	0.054	0.078

Residual correlation



\*DCT is DCT-II here.

# 7. Eigenfaces: Principal Component Transform

---

Kirby&Sirovich, IEEE-PAMI, Jan, 1990

Eigenface recognition derives its name from the German prefix "eigen", meaning "own" or "individual".

The eigenface recognition approach was developed by Turk and Pentland (1991), both from MIT.

- Training data:  $\{x^i\}_{i=1}^{K=593}$  are 46x46 images. They are used to estimate image statistics.

Vector size  $N=46 \times 46 = 2116$

The training images differ in  $\left\{ \begin{array}{l} \text{facial features} \\ \text{orientation, expression, lighting} \end{array} \right\}$

# 7. Eigenfaces: Principal Component Transform

---

- Compute:

$$\hat{m}_x = \frac{1}{K} \sum_{I=1}^K x^i, \text{ let } \bar{x}^i = x^i - \hat{m}_x$$

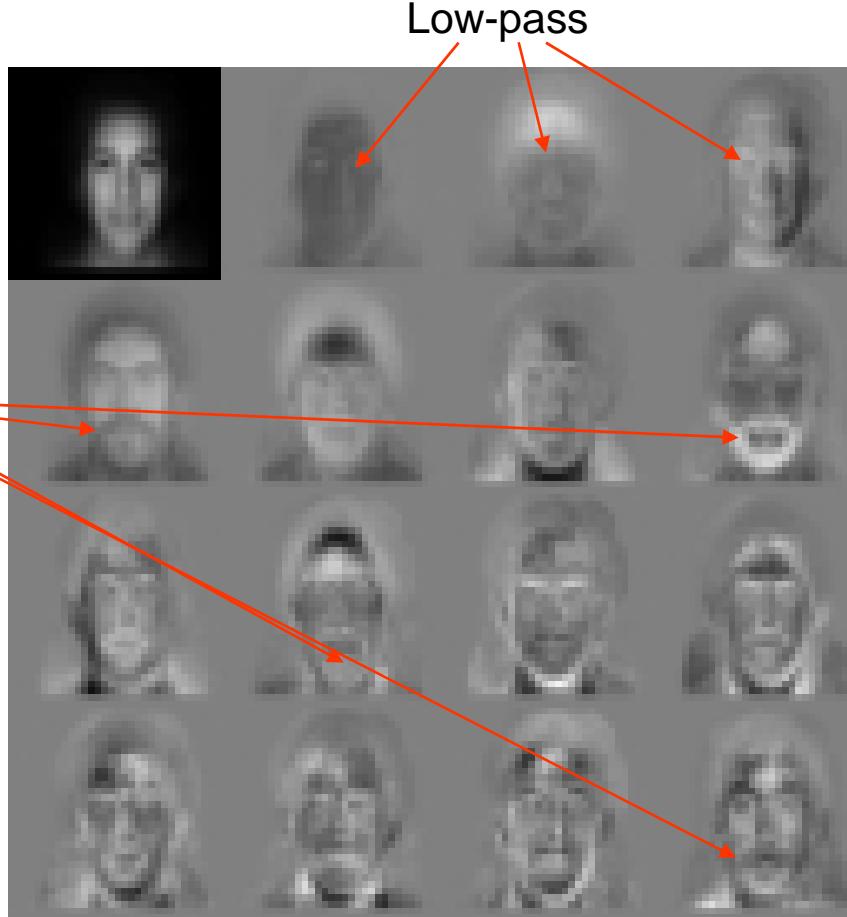
$$\hat{K}_x = \frac{1}{K} \sum_{I=1}^K \bar{x}^i \bar{x}^{i^T}$$

- Do eigenvector decomposition of  $\hat{K}_x$ .

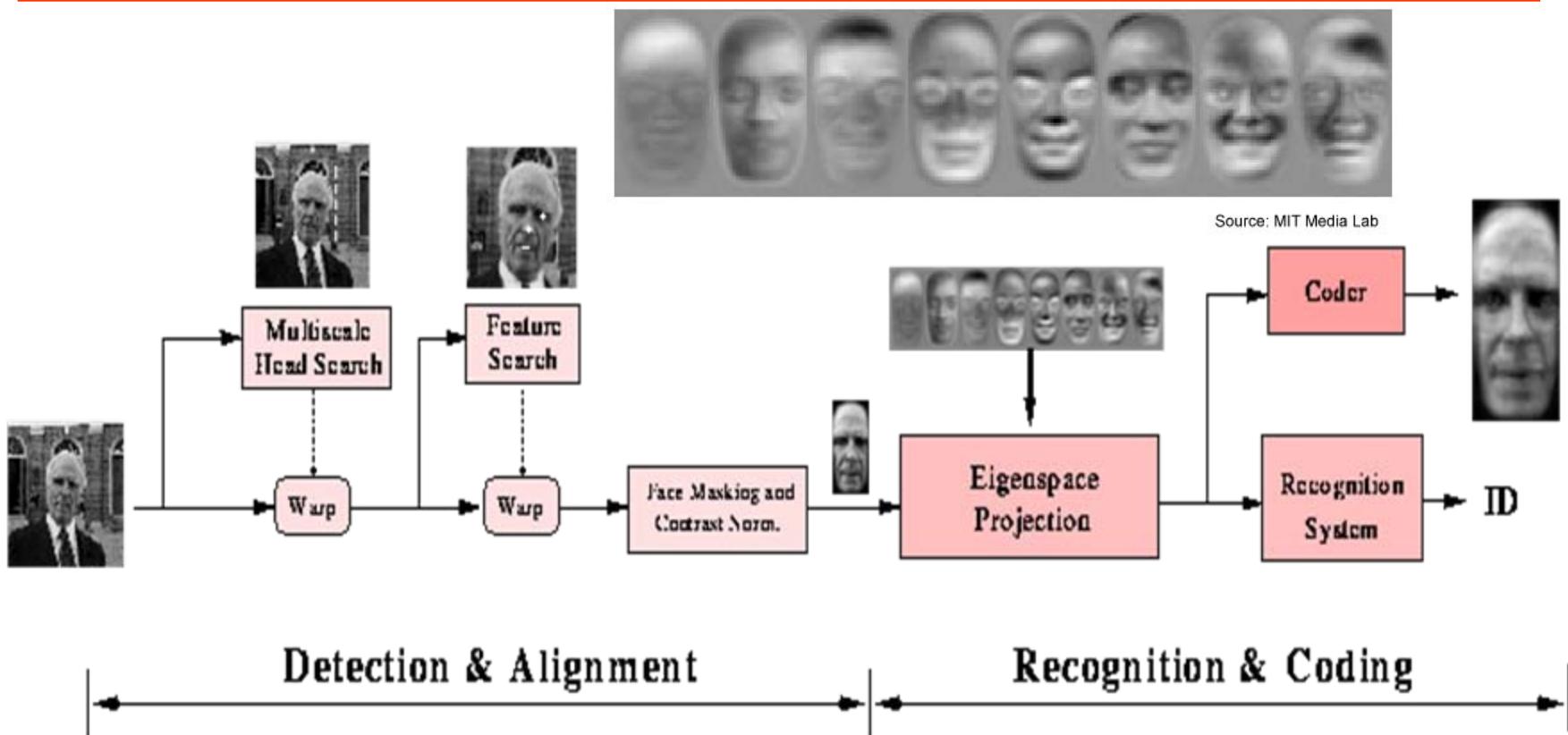
## 7. Eigenfaces: main features

---

More specific  
details are in “high  
frequencies”



## 7. Eigenfaces: Principal Component Transform



# 7. Eigenfaces: Principal Component Transform

---

- Applications:

- Compression: represent image using only smallest indexes in eigenfaces space.
- Recognition.
- Authentication.

For more info see:

<http://www.cs.rug.nl/~peterkr/FACE/frhp.html>

# Advanced Image Processing

---

## Part IV: Image Modeling

S. Voloshynovskiy

May 10, 2001



# Course Outline

---

- Recall of Linear Algebra.
- Introduction. Human Visual System.
- Image Representation: pyramids and wavelets.
- Random Signals.
- Image Modeling (the fourth part)
- Image Sensor Models. Noise Models.
- Image Denoising.
- Image Restoration.
- Image Compression.
- Video Modeling and Compression.
- Digital Data Hiding.

## 8. Transform domain Models

---

KLT:

$$y = Tx$$

Linear transform that decorrelates  $x$  depends on  $R_x$ .

The dependency on  $R_x$  implies some serious difficulties for estimating unknown  $R_x$  (recall eigenfaces application) or in coding, for communicating  $R_x$  to the decoder.

- Often want fixed transforms that nearly decorrelates  $x$  and are suitable for formulating image models in transform domain.

In this case,  $R_y$  is nearly diagonal.

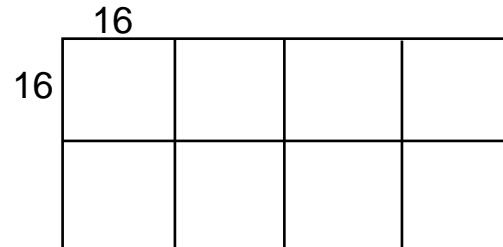
- Ex.1. DFT nearly diagonalizes  $R_x$  for WSS processes, N is large enough  
(  $R_x$  ~circulant approximation )  
But the DFT produces complex-valued coefficients y.

## 8. Discrete Cosine Transform (DCT)

---

Want a real-valued, orthonormal, DFT-like transform.

- There exist fast FFT-like algorithm for computing the DCT.
- The DCT basis functions are asymptotically equal to the KLT basis functions for AR(1) processes, as the correlation coefficient  $\rho \rightarrow 1$ .
- This has motivated the following image model: Images are “locally stationary”. Small (say 16x16) blocks obey an AR(1) model with  $\rho \rightarrow 0.95$ . Under this model, block DCT coefficients are nearly decorrelated. This property motivates independent coding of block-DCT coefficients in compression applications.



## 8. Discrete Cosine Transform (DCT)

---

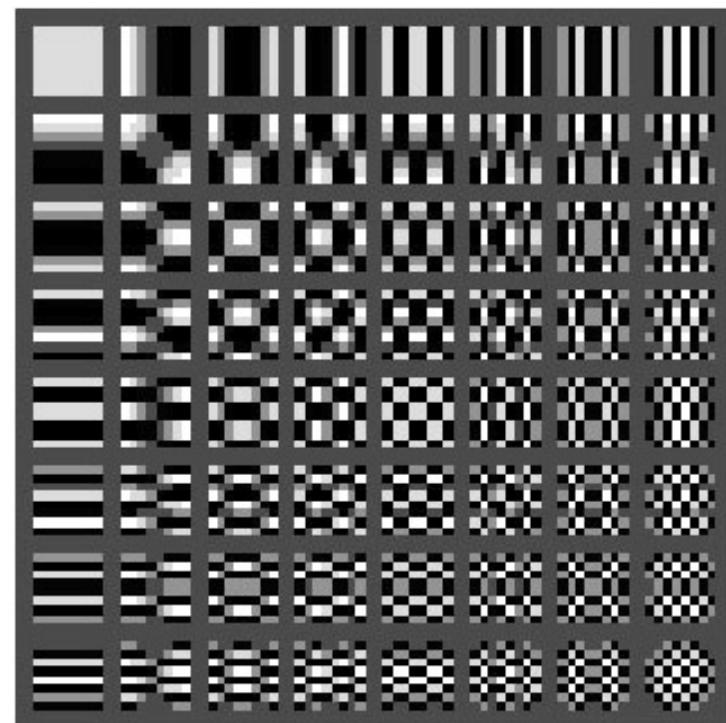
- Type II-DCT of blocksize  $M \times M$  is defined by transform matrix  $A$  containing elements
- 2D basis functions of the DCT:

$$a_{ik} = \alpha_i \cos \frac{\pi(2k+1)i}{2M}$$

for  $i, k = 0, \dots, M-1$

$$\text{with } \alpha_0 = \sqrt{\frac{1}{M}}$$

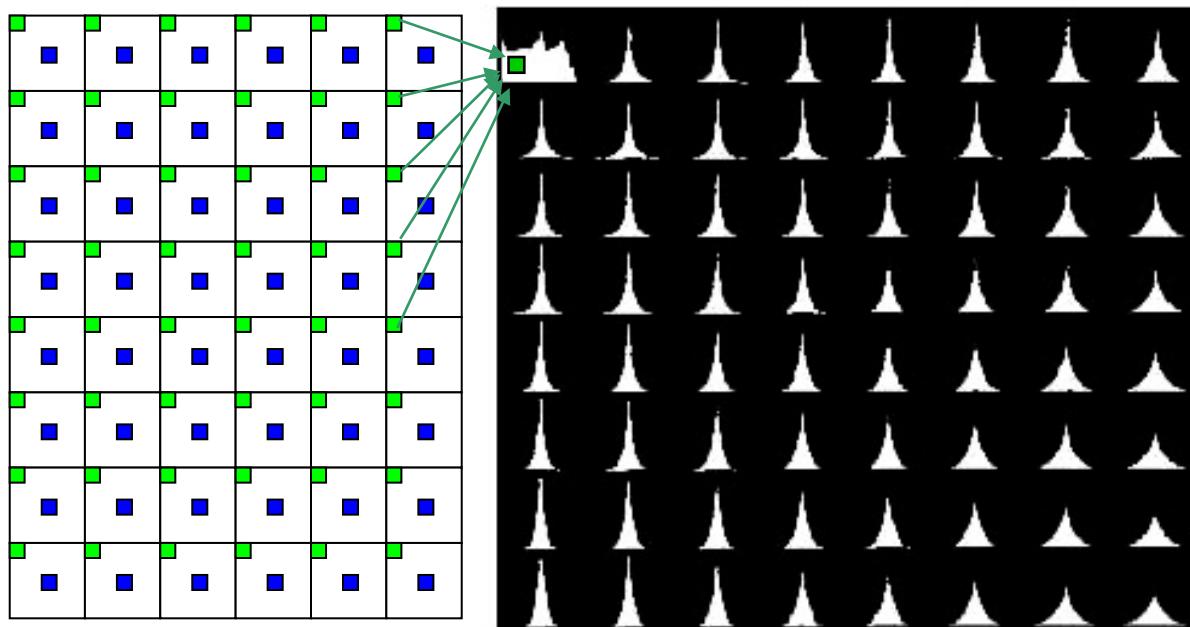
$$\alpha_i = \sqrt{\frac{2}{M}} \quad \forall i \neq 0$$



## 8. Discrete Cosine Transform (DCT)

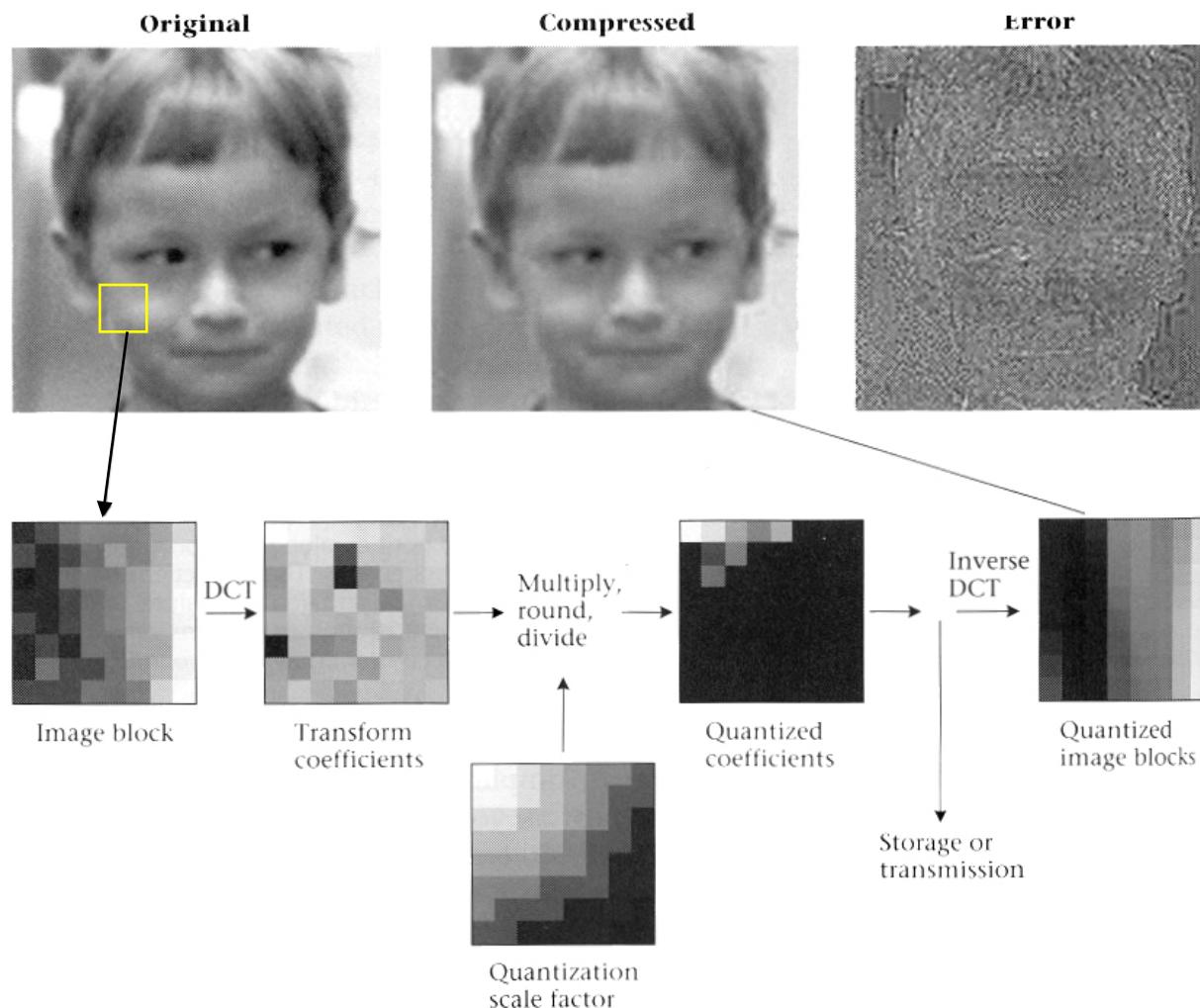
---

- Histograms for 8x8 DCT coefficient amplitudes measured for natural images (from Mauersberger):



- DC coefficient is typically uniformly distributed.

# 8. Discrete Cosine Transform: Compression



## 8. Discrete Cosine Transform (DCT)

---

The model above is of limited value due to the limitations of the AR(1) model:

- Also, just like KLT, the model is based on second-order statistics only.
- Finally, this model does not capture inter-block statistical dependencies.

## 8. Wavelets

---

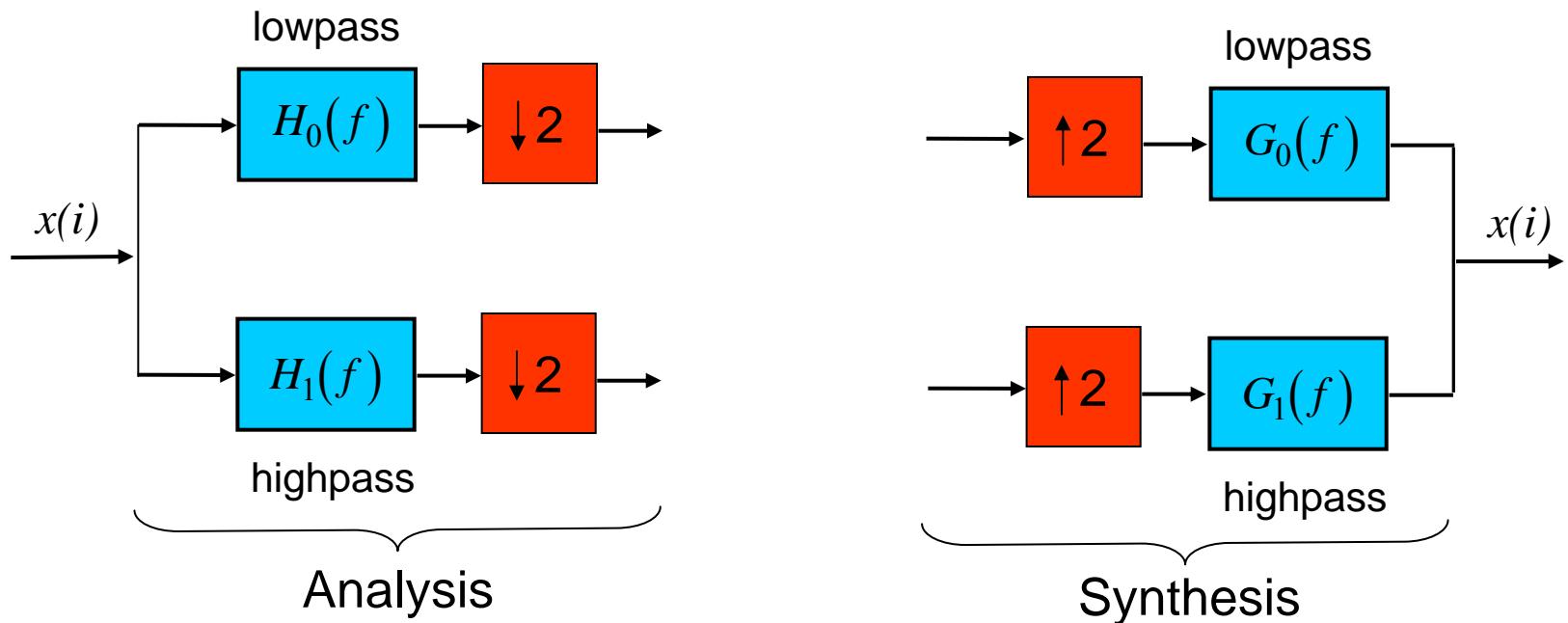
Wavelet transforms are another type of linear, orthogonal transform.

Advantages:

- existence of fast algorithms;
- no concept of “blocks”;
- decomposition of mage (signal) into coarse component and a sequence of detail images;
- Naturalness of multiresolution representation of visual information:
  - intuitive concept;
  - decomposition of images into frequency-selective channels in visual cortex.
- Wavelet coefficients are nearly decorrelated (0.1-0.3).

# 8. Wavelets

---



Wavelet decomposes  $x(i)$  into coarse component and detail signal.

---

## 8. Wavelets and KLT

---

Consider  $2 \times 1$  vector  $x = [x_1, x_2]^T$ , which represents the first-order Markov sequence with the covariance matrix:

$$R_x = \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix}$$

Find eigenvalues:  $\det[R_x - \lambda I] = 0 \rightarrow \lambda_1 = (1 + \rho)\sigma^2$   
 $\lambda_2 = (1 - \rho)\sigma^2$

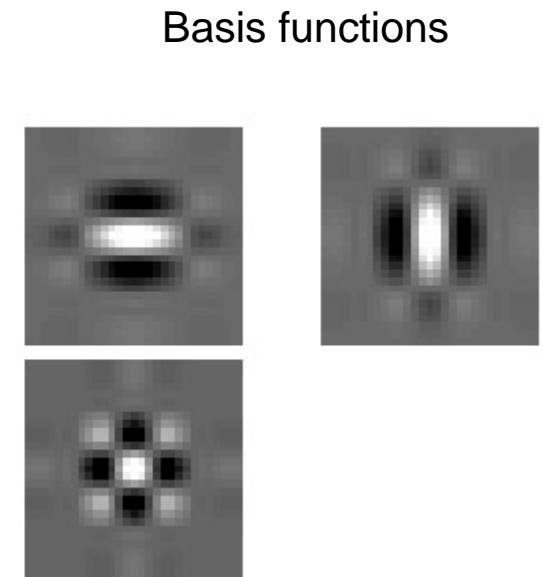
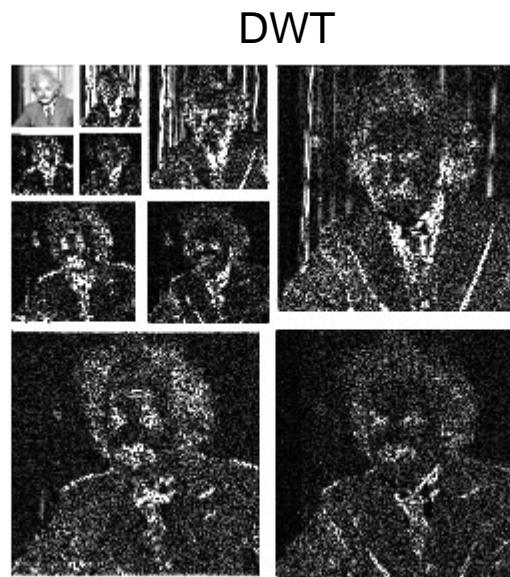
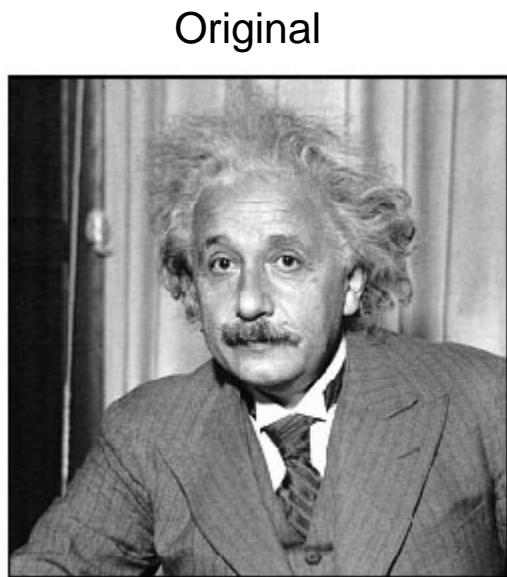
Then eigenvectors:  $\phi_1 = \begin{pmatrix} \phi_{11} \\ \phi_{12} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$   $\phi_2 = \begin{pmatrix} \phi_{21} \\ \phi_{22} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Haar  
first-order derivative approximation  
I-A

## 8. Wavelets and KLT

---



M J Wainwright and E P Simoncelli, Scale Mixtures of Gaussians and the Statistics of Natural Images. M J Wainwright and E P Simoncelli. *Adv. Neural Information Processing Systems*. v12, May 2000.

## 8. Wavelets

---

- Joint pdf for wavelet coefficients is non-Gaussian.
- There are 3 basic stochastic models that capture main properties of the wavelet coefficients:
  - those exploiting intrascale dependencies;  
(within one subband)
  - those exploiting interscale dependencies;  
(between subbands)
  - those exploiting both dependencies - hybrid models.

## 8. Wavelets: Intrascale Models

---

The group of intrascale models is the largest one and can be further divided on:

- Models that treat wavelet coefficients as i.i.d.

(due to nearly decorrelating properties of the KLT-wavelet);

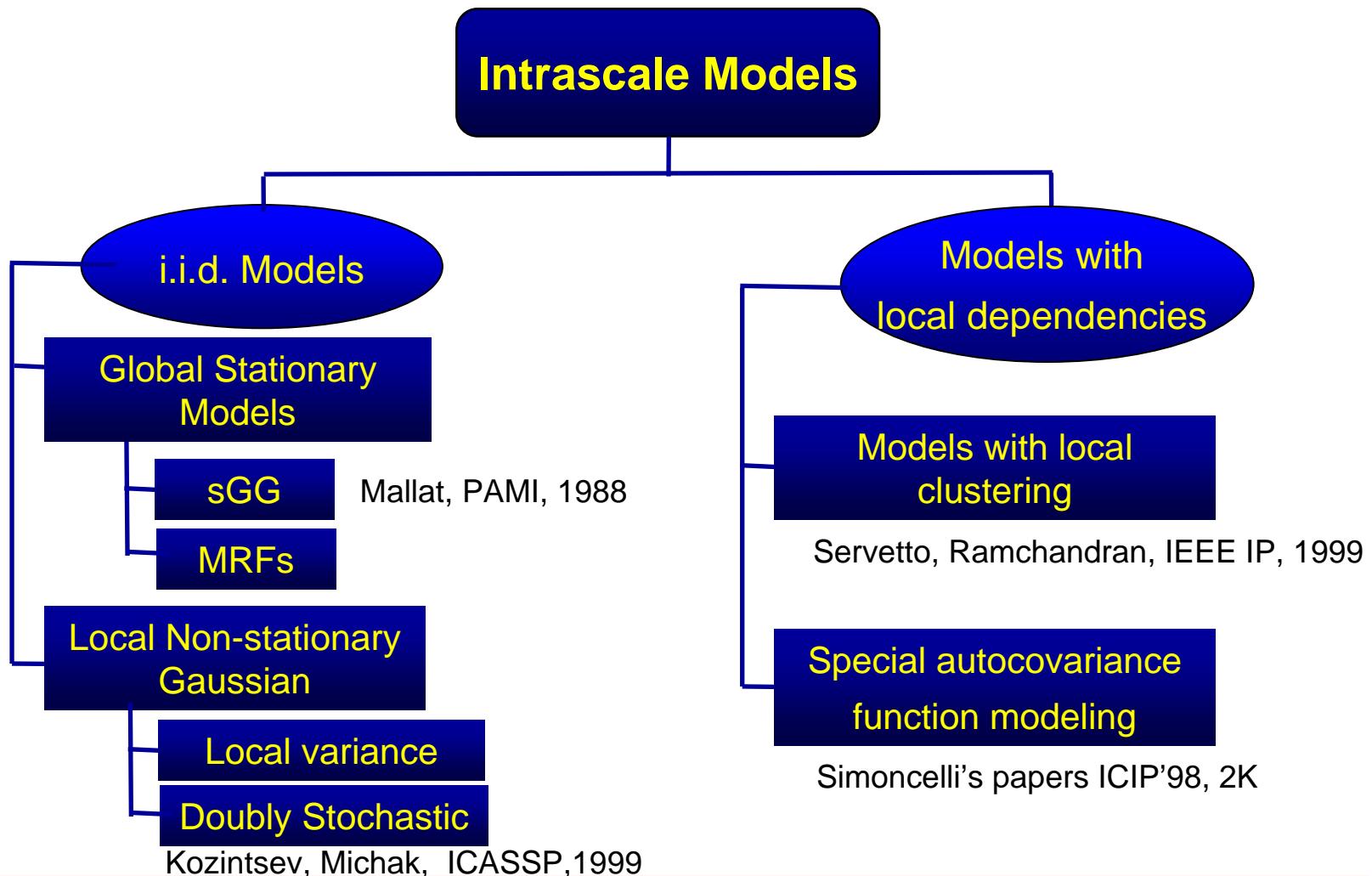
NOTE: decorrelation properties of the KLT implies only independence in the case of Gaussian distribution.

In the general case there exists Independent Component Analysis (ICA) contrary to Principal Component Analysis (PCA or KLT).

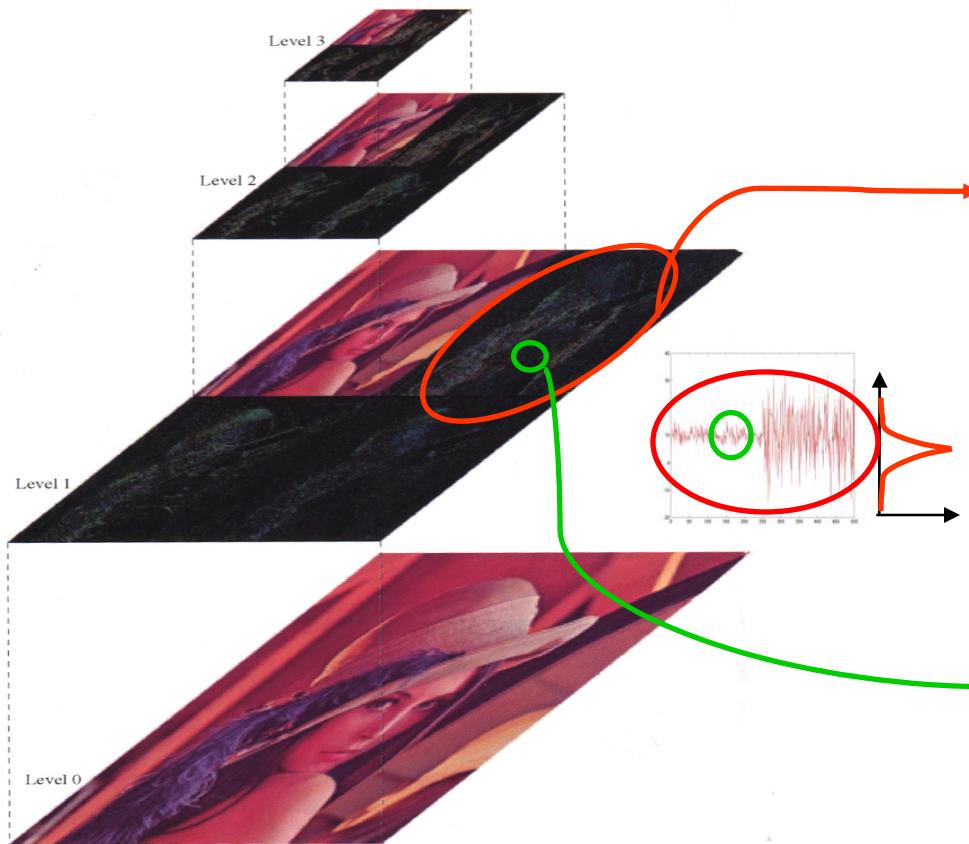
- Models that take into account the remaining correlation between the wavelet coefficients via local dependencies.

# 8. Wavelets: Intrascale Models

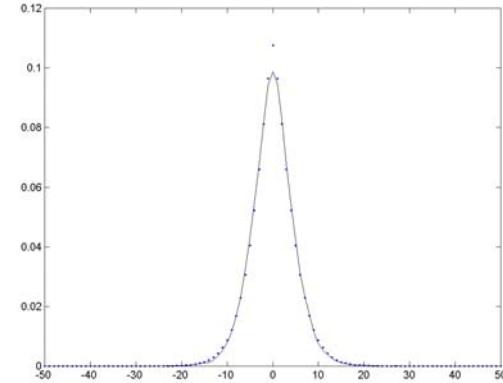
---



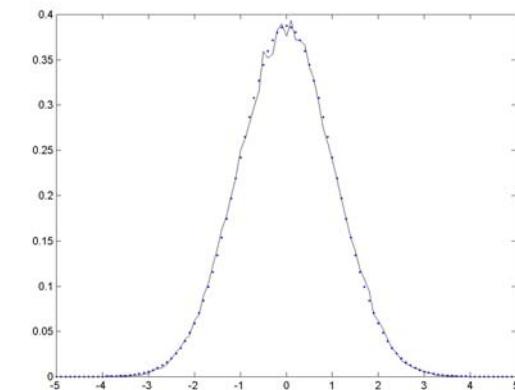
## 8. Wavelets: Intrascale i.i.d. Models



Global i.i.d. non-Gaussian model



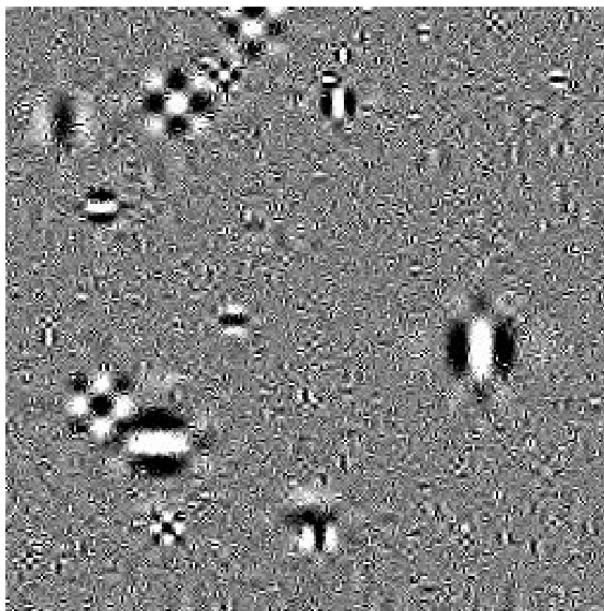
Local i.i.d. Gaussian model



## 8. Wavelets: Inrascale i.i.d. Models

---

Lets try to regenerate the image using the inrascale sGG model with the parameters estimated from Einstein image.



M J Wainwright and E P Simoncelli, Scale Mixtures of Gaussians and the Statistics of Natural Images. M J Wainwright and E P Simoncelli. *Adv. Neural Information Processing Systems*. v12, May 2000.

## 8. Wavelets: Intrascale i.i.d. Models

---

The problems with i.i.d. intrascale models:

- These models are used due to their simplicity and tractability of the obtained results.
- The wavelet coefficient are fairly decorrelated, but from another side they are not independent.
- Moreover, the large scale coefficients tend to occur near each other within subbands, and also (see it later) they occur at the same relative spatial locations in subbands at adjacent scale, and orientations.

Therefore, one has to capture these dependencies for correct stochastic image modeling.

# 8. Wavelets: Intrinsic Models with Local Dependencies

---

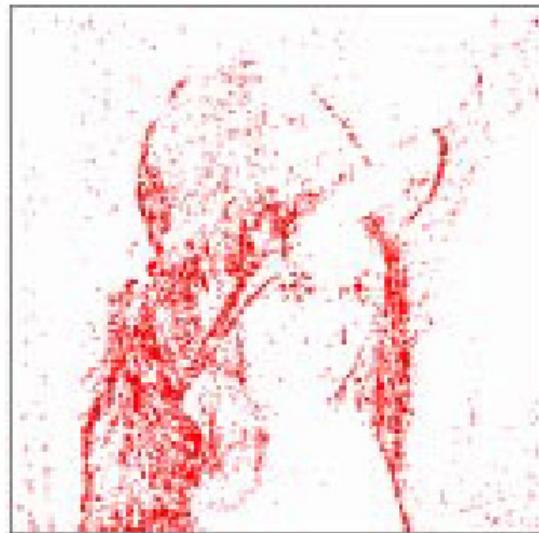
S. D. Servetto, K. Ramchandran, and M. T. Orchard.

Image Coding Based on a Morphological Representation of Wavelet Data.

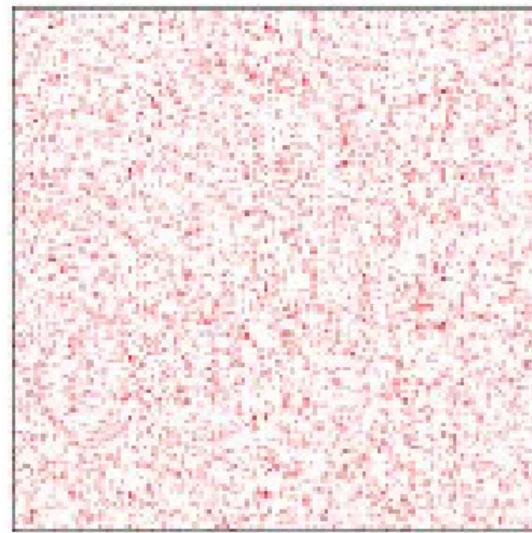
IEEE Transactions on Image Processing, 8(9):1161-1174, 1999.

Effect of wavelet coefficient clustering

(one more evidence of inconsistency of i.i.d. models )



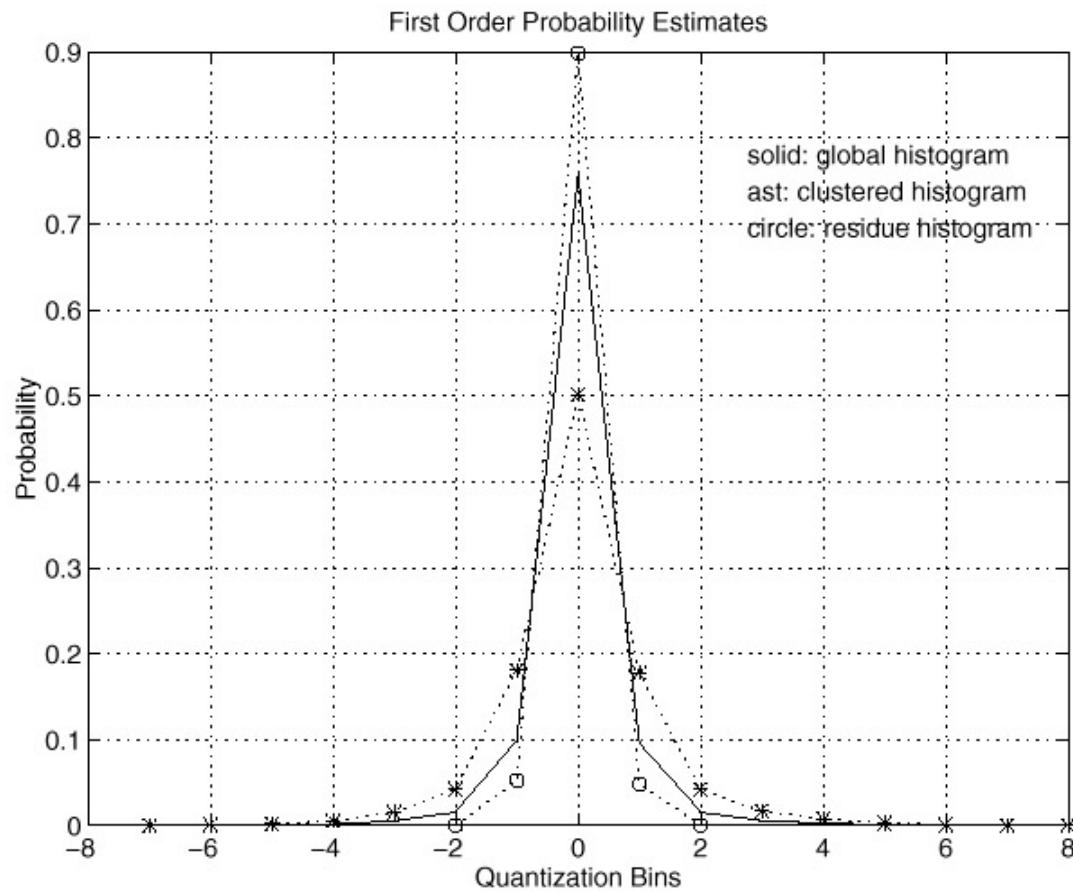
1st vertical subband  
(all coefficients above  $T$ )



Uniformly driven data with  
the same amount of coefficients

# 8. Wavelets: Intrascale Models with Local Dependencies

---



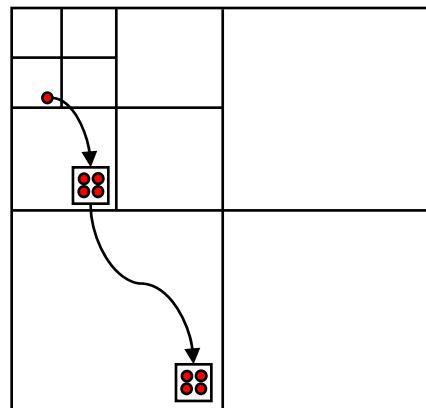
## 8. Wavelets: Interscale Models

---

There exists the strong clustering of wavelet coefficients not only in location with the subband, but also in scale (between different subbands).

E. Simoncelli, Statistical Models for Images: Compression, Restoration and Synthesis, 31st Asilomar Conf., Nov. 1997. (and ICIP'98)

In order to model the interscale dependencies it is useful to define parent/children dependencies.



## 8. Wavelets: Interscale Models

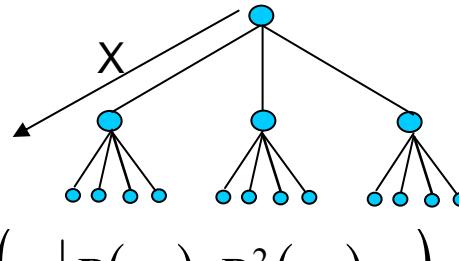
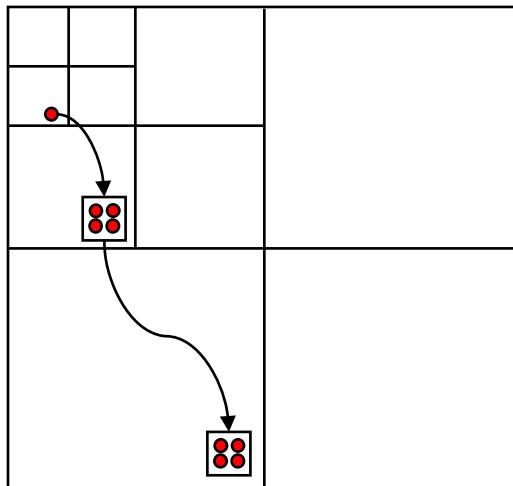
Markov model in scale has been proposed to capture interscale dependencies.

- A. Wilsky's group

IEEE SP, Dec'93, IEEE IT, Mar' 92

- Bouman, 1993

- recently Baranuk (Rice University)



$$\text{where } p(x_n | P(x_n), P^2(x_n), \dots) = p(x_n | P(x_n))$$

parent      grandparent

If the root of some subtree is zero, then all nodes in the subtree are zero.

The resulting subtree is called a zerotree. [EZT - Shapiro, IEEE SP, Dec'93

SPIHT - Said&Pearlman, IEEE CS, June' 96]

## 8. Wavelets: Interscale Models based on HMM

---

The hierarchical structure of interscale dependencies makes it possible to use Hidden Markov Models (HMM) to characterize these dependencies between parents and children in the tree.

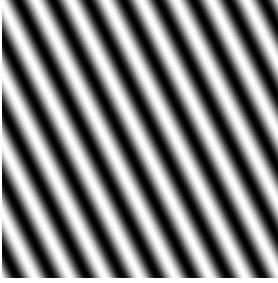
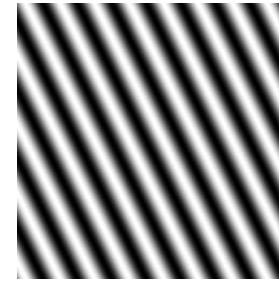
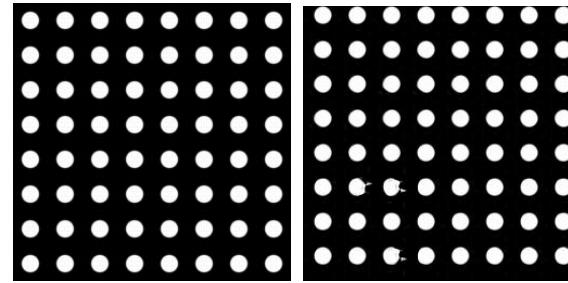
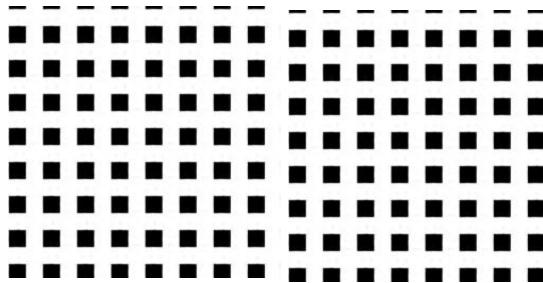
- M.S. Crouse, R. D. Nowak and R. Baraniuk, Wavelet-based statistical signal processing using Hidden Markov models, IEEE Trans. on Signal Processing, 46, 4, April 1998, pp. 886-902.
- J. K. Romberg, H. Choi and R. Baraniuk, Bayesian tree-structured image modeling using wavelet-domain Hidden Markov model, in Proc. of SPIE, Denver, USA, vol. 3816, July 1999, pp. 31-44.
- G. Fan. and X. Xia, Wavelet-based image denoising using Hidden Markov models, In IEEE ICIP, Vancouver Canada, October, 2000.

# 8. Wavelets: Intercalate Models with Local Dependencies

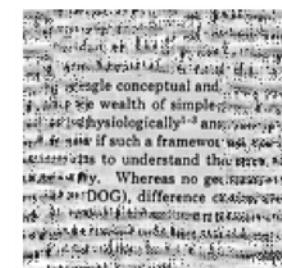
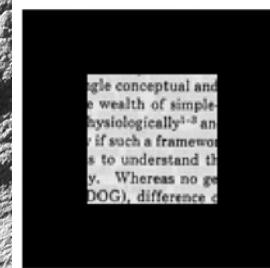
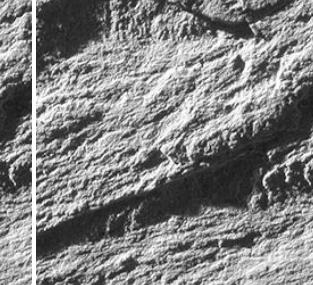
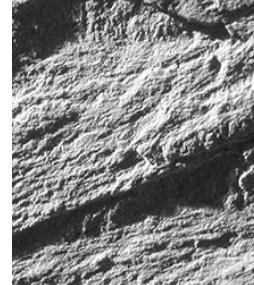
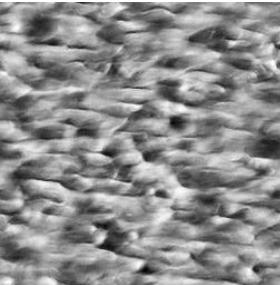
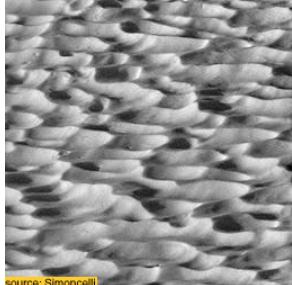
---

Texture synthesis: (A. Simoncelli: <http://www.cns.nyu.edu/~lcv/texture/>)

Artificial Periodic Textures



Photographic/random



source: Simoncelli

# 8. Wavelets: Intrinsic Models with Local Dependencies

---

Texture synthesis: (A. Simoncelli: <http://www.cns.nyu.edu/~lcv/texture/>)

Inhomogeneous (non-textures):



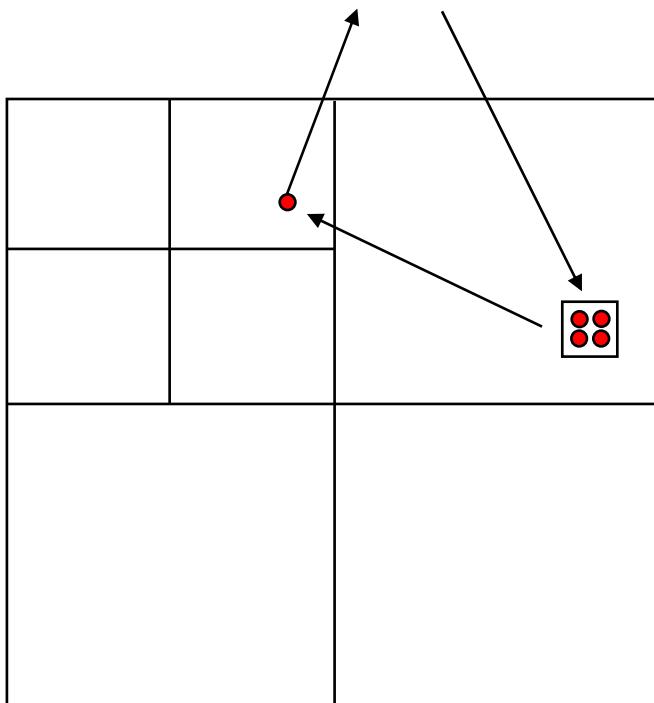
Color:



## 8. Wavelets: Hybrid Models

---

Compare with  
some threshold T



If  $>T$ , put x to Gsig  
else, put x to Ginsig

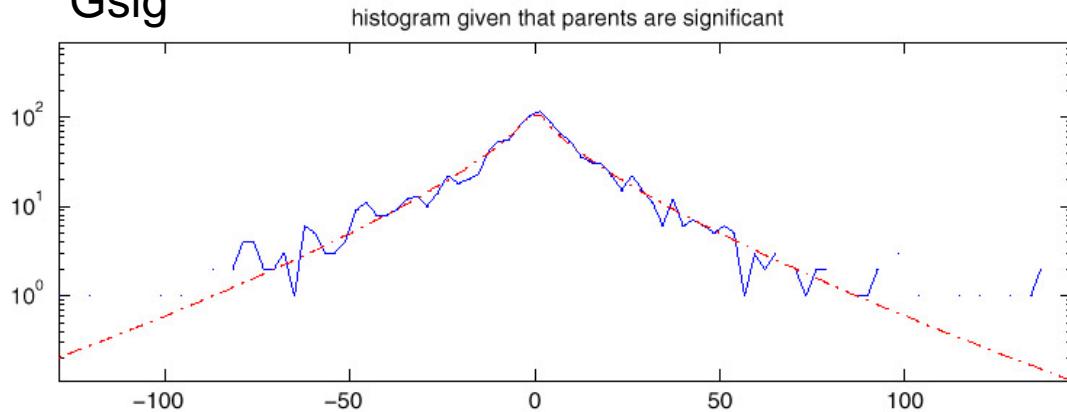
This concept is highly inspired by EZT:  
if the coefficient is significant ( $>T$ ), all  
his children will be significant (Gsig).

It creates clustering on two groups  
within each subband.

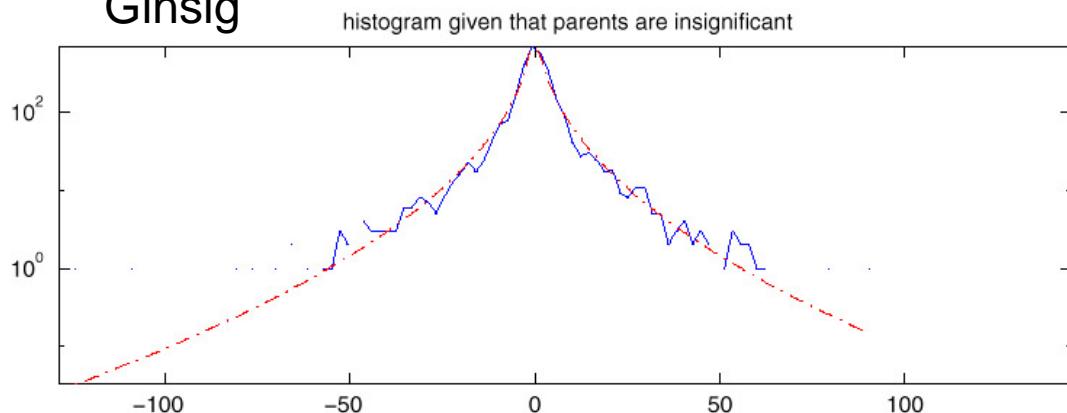
## 8. Wavelets: Hybrid Models

---

Gsig



Ginsig



i.i.d. sGG approximation  
of the clustered coefficients.

Variance and shape for  
the set of Gsig is higher  
than for Ginsig.

Ginsig: flat regions

Gsig: textures and edges