Review of Statistics and Information Theory Basics

COMP / ELEC / STAT 502, Rice University

Let ξ, η be continuous scalar random variables (r.v.-s).

(Cumulative) probability distribution function (CDF)

$$F_\xi(x) = Pr(\xi \leq x) = \int_{-\infty}^{x} f_\xi(u)du$$
$$\text{and} \tag{1}$$
$$F_\eta(y) = Pr(\eta \leq y) = \int_{-\infty}^{y} f_\eta(v)dv$$

where

$$F'_\xi(x) = f_\xi(x) \qquad \text{and} \qquad F'_\eta(y) = f_\eta(y) \tag{2}$$

and exist almost everywhere for continuous r.v.-s. by definition.

$F_\xi(x)$ is monotoniously non-decreasing; right-continuous, and
$F_\xi(-\infty) = 0$, $F_\xi(\infty) = 1$.

Probability density function (*pdf*):

$f_\xi(x)$ and $f_\eta(y)$ above are the *pdf*s of ξ and η, respectively. Then,

$$0 \leq f_\xi(x) \qquad \text{and} \qquad 0 \leq f_\eta(y)$$
$$\int_{-\infty}^{\infty} f_\xi(x)dx = 1 \qquad \text{and} \qquad \int_{-\infty}^{\infty} f_\xi(x)dx = 1 \tag{3}$$

and $f_\xi(x)$ and $f_\eta(y)$ are continuous almost everywhere.

Interpretation of $f_\xi(x)$ as "probability":

$$f_\xi(x) = \lim_{h\to 0} \frac{F_\xi(x + h/2) - F_\xi(x - h/2)}{h}$$
$$= \lim_{h\to 0} \frac{P_\xi[\xi \leq x + h/2] - P_\xi[\xi \leq x - h/2]}{h} = \lim_{h\to 0} \frac{P_\xi[x - h/2) < \xi \leq x + h/2]}{h} \tag{4}$$

The larger $f_\xi(x)$ the larger the probability that ξ falls close to x.

See more in Andrew Moore's tutorial under "Additional Links" at the course web site.

Moments of random variables

The kth moment $M_\xi^{(k)}$ of ξ is defined as the expected value of the kth power of the variable:

$$M_\xi^{(k)} = E[\xi^k] = \int_{-\infty}^{\infty} x^k dF_\xi(x) = \int_{-\infty}^{\infty} x^k f_\xi(x) dx \tag{5}$$

The kth *central* moment, $m_\xi^{(k)}$ is defined similarly except for a shift to obtain zero mean ($m_\xi^{(1)}$).

$$m_\xi^{(k)} = \int_{-\infty}^{\infty} (x - E[\xi])^k dF_\xi(x) = \int_{-\infty}^{\infty} (x - E[\xi])^k f_\xi(x) dx \tag{6}$$

Quantities of interest related to the first four moments, as estimated from a sample, are

$$Mean = M_\xi^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \xi_i \tag{7}$$

$$\begin{aligned} Variance &= m_\xi^{(2)} \\ &= \frac{1}{n-1}\sum_{i=1}^{n} (\xi_i - M_\xi^{(1)})^2 \end{aligned} \tag{8}$$

$$Standard\ \ Deviation \quad (STD) = \sqrt{Variance} \tag{9}$$

$$\begin{aligned} Skewness &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\xi_i - M_\xi^{(1)}}{STD} \right)^3 \\ &= {m_\xi^{(2)}}^{-3/2} \cdot m_\xi^{(3)} \end{aligned} \tag{10}$$

$$\begin{aligned} (Normalized)\ \ Kurtosis &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\xi_i - M_\xi^{(1)}}{STD} \right)^4 - 3 \\ &= {m_\xi^{(2)}}^{-2} \cdot m_\xi^{(4)} - 3 \end{aligned} \tag{11}$$

Kurtosis is often used as a quantitative measure of non-Gaussianity. If $kurt(\xi)$ denotes the Kurtosis of ξ,

$$kurt(\xi) = \begin{cases} < 0 & \text{for subgaussian or } \textit{platykurtic} \text{ distribution} \\ = 0 & \text{for gaussian or } \textit{mesokurtic} \text{ distribution} \\ > 0 & \text{for supergaussian or } \textit{leptokurtic} \text{ distribution} \end{cases} \tag{12}$$

The *covariance* of random variables ξ and η is defined as

$$c_{\xi\eta} = E[(\xi - M_\xi^{(1)})(\eta - M_\eta^{(1)})] \tag{13}$$

or in general, for the elements x_i of random vector $\mathbf{x}$

$$c_{ij} = E[(x_i - M_{x_i}^{(1)})(x_j - M_{x_j}^{(1)})] \quad \text{or} \tag{14}$$

$$\mathbf{C_x} = E[(\mathbf{x} - M_{\mathbf{x}}^{(1)})(\mathbf{x} - M_{\mathbf{x}}^{(1)})^T] \tag{15}$$

Similarly, the *cross-covariance* of random vectors $\mathbf{x}, \mathbf{y}$ is

$$\mathbf{C_{xy}} = E[(\mathbf{x} - M_{\mathbf{x}}^{(1)})(\mathbf{y} - M_{\mathbf{y}}^{(1)})^T] \tag{16}$$

Joint CDF and *pdf* of *jointly continuous* random variables ξ and η

$$F_{\xi,\eta}(x,y) = Pr(\xi \le x \wedge \eta \le y) \tag{17}$$

$$F_{\xi,\eta}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{\xi,\eta}(u,v)dudv \tag{18}$$

where

$$f_{\xi,\eta}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{\xi,\eta}(x,y) \tag{19}$$

exists almost everywhere, and it is the joint ***pdf*** of ξ and η. Properties of $F_{\xi,\eta}(x,y)$ and $f_{\xi,\eta}(x,y)$ are analogous to eqs (1)–(3).

Marginal distributions

$$F_\xi(x) \;\; = F_{\xi,\eta}(x,\infty) \;\; = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{\xi,\eta}(u,v)dudv \tag{20}$$

$$F_\eta(y) \;\; = F_{\xi,\eta}(\infty,y) \;\; = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f_{\xi,\eta}(u,v)dudv \tag{21}$$

A pair of joint random variables ξ and η are ***statistically independent*** iff their joint CDF factorizes into the product of the ***marginal CDFs***:

$$F_{\xi,\eta}(x,y) = F_\xi(x) \cdot F_\eta(y) \tag{22}$$

Equivalently, two jointly continuous random variables ξ and η are ***statistically independent*** iff their joint ***pdf*** factorizes into the product of the ***marginal density functions***:

$$f_{\xi,\eta}(x,y) = f_\xi(x) \cdot f_\eta(y) \tag{23}$$

$$\begin{aligned} F_{\xi,\eta}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{\xi,\eta}(u,v)dudv \\ &= \int_{-\infty}^{x} f_{\xi}(u)du \int_{-\infty}^{y} f_{\eta}(v)dv \\ &= F_{\xi}(x)\cdot F_{\eta}(y) \end{aligned} \tag{24}$$

(Prove the other direction at home.)

Conditional probability

$$F_{\xi,\eta}(x|y) = F_{\xi,\eta}(x,y)/F_{\eta}(y) \tag{25}$$
$$f_{\xi,\eta}(x|y) = f_{\xi,\eta}(x,y)/f_{\eta}(y) \tag{26}$$

Obviously, $F_{\xi,\eta}(x|y) = F_{\xi}(x)$ iff ξ and η are independent.
Equivalently, $f_{\xi,\eta}(x|y) = f_{\xi}(x)$ iff ξ and η are independent.

Information and entropy

Let X be a discrete random variable, taking its values from $\mathcal{H} = \{x_1, x_2, \ldots, x_n\}$. Let p_i be the probability of $x_i, p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. The amount of information in observing the event $X = x_i$ was defined by Shannon as

$$I(x_i) = -log_a(p_i) \tag{27}$$

$$\text{If } a = \begin{cases} 2 & \text{then } I \text{ is given in } bits, \\ e & \text{then } I \text{ is given in } nats, \\ 10 & \text{then } I \text{ is given in } digits. \end{cases} \tag{28}$$

The expected value of $I(x_i)$ over the entire set of events $\mathcal{H}$ is the *entropy* of X:

$$H(X) = -\sum_{i=1}^{n} p_i log(p_i) \tag{29}$$

(Note that X in $H(X)$ is not an argument of a function but the label of the random variable.) The convention $0 \cdot log0 = 0$ is used. The entropy is the average amount of information gained by observing one event from all possible events for the random variable X (the average information per message), the average length of the shortest description of a random variable, the amount of uncertainty, a measure of the randomness of X.

Some basic properties of $H(X)$:

$$H(X) = 0 \quad \text{iff } p_i = 1 \text{ for some } i \text{ and } p_k = 0 \text{ for } k \neq i. \tag{30}$$

(If an event occurs with probability 1 then there is no uncertainty.)

$$0 \leq H(X) \leq log|\mathcal{H}| \text{ with equality iff } p_i = 1/|\mathcal{H}| = 1/n \text{ for all } i. \tag{31}$$

(Uniform distribution carries maximum uncertainty.)

The *differential entropy* of a continuous random variable.

A quantity analogous to the entropy of a discrete random variable can be defined for continuous random variables.

Definition: The random variable ξ is ***continuous*** if its distribution function, $F_\xi(x)$ is continuous.

Definition: Let $f_\xi(x) = F'_\xi(x)$ be the pdf of ξ. The set $S : \{x|f_\xi(x) > 0\}$ is called the ***support set*** of ξ. The ***differential entropy*** of ξ is defined as

$$h(\xi) = -\int_S f_\xi(u) log f_\xi(u) du \qquad = -E[log f_\xi(x)] \tag{32}$$

where S is the support set of ξ. The lower case letter h is used to distinguish differential entropy from the discrete (or absolute) entropy H.

Relationship between discrete (absolute) and differential entropy

The differential entropy of a continuous variable ξ can be derived as a limiting case for the discrete entropy of the quantized version of ξ. (See, for example, Cover and Thomas, p. 228.) If $\Delta = 2^{-n}$ is the bin size of the quatization, and the quantized version of ξ is denoted by ξ^Δ then

$$H(\xi^\Delta) + log\Delta \longrightarrow h(\xi) \text{ as } n \longrightarrow \infty \quad (\Delta \longrightarrow 0) \tag{33}$$

That is, the absolute entropy of an ***n-bit*** quantization of a continuous random variable ξ is approximately $h(\xi) - log2^{-n} = h(\xi) + n$.

It follows that the absolute entropy of a continuous (non-quantized) variable is ∞.

Entropy from a continuous distribution — the Maximum Entropy Principle

Consider the following: suppose we have a stochastic system with a set of known states but unknown probabilities, and we also know some constraints on the distribution of the states. How can we choose the probability model that is optimum in some sense among the possibly infinite number of models that may satisfy the constraints? The Maximum Entropy principle (due to Jaynes, 1957) states that "When an inference is made on the

basis of incomplete information, it should be drawn from the probability ditribution that maximizes the entropy, subject to constraints on the distribution."

Solving this problem, in general, involves the maximization of the differential entropy

$$h(\xi) = -\int_{-\infty}^{\infty} f_\xi(u) log f_\xi(u) du \tag{34}$$

over all density functions $f_\xi(x)$ of the random variable ξ, subject to the following constraints:

$$f_\xi(x) \geq 0 \tag{35}$$

$$\int_{-\infty}^{\infty} f_\xi(x) dx = 1 \tag{36}$$

$$\int_{-\infty}^{\infty} f_\xi(x) g_i(x) dx = \alpha_i \quad \text{for } i = 2, \ldots, m \tag{37}$$

where $f_\xi(x) = 0$ holds outside of the support set of ξ, and α_i are the prior knowledge about ξ.

Using the Lagrange multipliers $\lambda_i \quad (i = 1, 2, \ldots, m)$ to formulate the objective function

$$J(f) = \int_{-\infty}^{\infty} \left[-f_\xi(x) log f_\xi(x) + \lambda_1 f_\xi(x) + \sum_{i=2}^{m} \lambda_i g_i(x) f_\xi(x) \right] dx \tag{38}$$

Differentiation of the integrand with respect to f_ξ and setting the results to 0 produces

$$-1 - log f_\xi(x) + \lambda_1 + \sum_{i=2}^{m} \lambda_i g_i(x) = 0 \tag{39}$$

Solving this for $f_\xi(x)$ yields the maximum entropy distribution for the given constraints:

$$f_\xi(x) = exp\left(-1 + \lambda_1 + \sum_{i=2}^{m} \lambda_i g_i(x) \right) \tag{40}$$

Example: If the prior knowledge consists of the mean μ and the variance σ^2 of ξ then

$$\int_{-\infty}^{\infty} f_\xi(x)(x - \mu)^2 dx = \sigma^2 \tag{41}$$

and from the constraints it follows that $g_2(x) = (x - \mu)^2$ and $\alpha_2 = \sigma^2$. The Maximum Entropy distribution for this case is the Gaussian distribution as shown on page 28 of Colin Fyfe 2 (CF2), or Haykin, Chapter 10, page 491.

Joint differential entropy (C-T 230)

$$h(\xi,\eta) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(u,v) log f_{\xi,\eta}(u,v) du dv \tag{42}$$

Conditional differential entropy

In CF 2/2, you have seen the formula of the *conditional entropy* of discrete random variable X given Y. It represents the amount of uncertainty remaining about the system input X after the system output Y has been observed. Analogous to that is the *conditional differential entropy* for continuous random variables:

$$h(\xi|\eta) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y) log f_{\xi,\eta}(x|y) dx dy \tag{43}$$

Since $f_{\xi,\eta}(x|y) = f_{\xi,\eta}(x,y)/f_\eta(y)$

$$h(\xi|\eta) = h(\xi,\eta) - h(\eta). \tag{44}$$

Relative entropy (Kullback-Leibler distance)

The relative entropy is an entropy based measure of the difference between two pdf's $f_\xi(x)$ and $g_\xi(x)$, which is defined as

$$D(f_\xi \parallel g_\xi) = \int_{-\infty}^{\infty} f_\xi(x) log \frac{f_\xi(x)}{g_\xi(x)} dx \tag{45}$$

Note that $D(f \parallel g)$ is finite only if the support set of f is contained in the support set of g. The K-L distance is a measure of the penalty (error) for assuming $g_\xi(x)$ when in reality the density function of ξ is $f_\xi(x)$. The convention $0 \cdot log\frac{0}{0} = 0$ is used.

It follows (can be proven) that $D(f_\xi \parallel g_\xi) \geq 0$ with equality iff $f_\xi(x) = g_\xi(x)$ almost everywhere. The relative entropy or K-L distance is also called the K-L divergence.

Mutual information

Mutual information is the measure of the amount of information one random variable contains about another.

$$I(\xi;\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y) log \frac{f_{\xi,\eta}(x,y)}{f_\xi(x) f_\eta(y)} dx dy \tag{46}$$

The concept of mutual information is of profound importance in the design of self-organizing systems, where the objective is to develop an algorithm that can learn an input-output relationship of interest on the basis of the input patterns alone. The *Maximum mutual*

information principle of Linsker (1988) states that the synaptic connections of a multilayered ANN develop such as to maximize the amount of information that is preserved when signals are transformed at each processing stage of the network, subject to certain constraints. This is based on earlier ideas that came from the recognition that encoding of data from a scene for the purpose of redundancy reduction is related to the identification of specific features in the scene (by the biological perceptual machinery).
Properties of mutual information:

$$\begin{aligned} I(\xi;\eta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y) log \frac{f_{\xi,\eta}(x,y)}{f_\xi(x) f_\eta(y)} dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y) log \frac{f_{\xi,\eta}(x,y)}{f_\eta(y)} dxdy \\ &- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y) log f_\xi(x) dxdy \\ &= -h(\xi|\eta) - \int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y)dy) log f_\xi(x) dx \\ &= -h(\xi|\eta) - \int_{-\infty}^{\infty} f_\xi(x) log f_\xi(x) dx \\ &= h(\xi) - h(\xi|\eta) \\ &= h(\eta) - h(\eta|\xi) \\ &= h(\xi) + h(\eta) - h(\xi,\eta) \end{aligned} \tag{47}$$

$$I(\xi;\eta) = I(\eta;\xi) \tag{48}$$

$$I(\xi;\eta) \geq 0 \tag{49}$$

$$I(\xi;\eta) = D(f_{\xi,\eta}(x,y) \parallel f_\xi(x) f_\eta(y)) \tag{50}$$

The last equation expresses mutual information as a special case of the K-L distance.

$$I(\xi;\eta) = D(\underbrace{f_{\xi,\eta}(x,y)}_{density\ \#1} \parallel \underbrace{f_\xi(x) f_\eta(y)}_{density\ \#2})$$

The larger the K-L distane the less independent ξ and η are, because $f_{\xi,\eta}(x,y)$ and $f_\xi(x) f_\eta(y)$ are farther apart $\Longrightarrow I$ is large.
See also diagram in CF 2/2 or in Haykin p.494 (fig. 10.1) showing insightful relations among mutual information, entropy, conditional and joint entropy.

Some additional properties

Statistically independent random variables satisfy the following

$$E[g(\xi)h(\eta)] = E[g(\xi)]E[h(\eta)] \tag{51}$$

where $g(\xi)$ and $h(\eta)$ are any absolutely integrable functions of ξ and η, respectively. This follows from

$$\begin{aligned} E[g(\xi)h(\eta)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)h(v)f_{\xi,\eta}(u,v)dudv \\ &= \int_{-\infty}^{\infty} g(u)f_{\xi}(u)du \int_{-\infty}^{\infty} h(v)f_{\eta}(v)dv \\ &= E[g(\xi)]E[h(\eta)] \end{aligned} \tag{52}$$

Statistical independence is a stronger property than *uncorrelatedness*.
ξ and η are uncorrelated iff

$$E[\xi\eta] = E[\xi]E[\eta] \tag{53}$$

In general two random vectors $\mathbf{x}, \mathbf{y}$ are uncorrelated iff

$$\mathbf{C_{xy}} = E[(\mathbf{x} - M_{\mathbf{x}}^{(1)})(\mathbf{y} - M_{\mathbf{y}}^{(1)})^T] = \mathbf{0} \tag{54}$$

or equivalently,

$$\mathbf{R_{xy}} = E[\mathbf{x}\mathbf{y}^T] = E[\mathbf{x}]E[\mathbf{y}^T] = M_{\mathbf{x}}^{(1)} {M_{\mathbf{y}}^{(1)}}^T \tag{55}$$

where $\mathbf{R_{xy}}$ is the *correlation* matrix of $\mathbf{x}, \mathbf{y}$. For zero mean the correlation matrix is the same as the covariance matrix.

As a special case, the components of a random vector $\mathbf{x}$ are mutually uncorrelated iff

$$\begin{aligned} \mathbf{C_x} &= E[(\mathbf{x} - M_{\mathbf{x}}^{(1)})(\mathbf{x} - M_{\mathbf{x}}^{(1)})^T] \\ &= \mathbf{D} = Diag(c_{11}, c_{22}, \ldots, c_{nn}) \\ &= Diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) \end{aligned} \tag{56}$$

Whiteness

Random vectors with

$$M_{\mathbf{x}}^{(1)} = 0 \text{ and } \mathbf{C_x} = \mathbf{R_x} = \mathbf{I} \tag{57}$$

i.e. having zero mean and unit covariance (and hence unit correlation) matrix are called *white*. Whiteness is a stronger property than uncorrelatedness but weaker than statistical independence.

Density functions of frequently used distributions

Uniform:

$$f_\xi(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & elsewhere \end{cases} \tag{58}$$

Gaussian:

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{59}$$

where μ and σ^2 denote the mean and variance, respectively.

Laplacian:

$$f_\xi(x) = \frac{\lambda}{2} e^{-\lambda|x-\mu|} \quad \lambda > 0 \tag{60}$$

These density functions are special cases of the ***generalized gaussian or exponential power family*** that has the general formula (for zero mean):

$$f_\xi(x) = C \cdot exp\left(-\frac{|x|^\nu}{\nu E[|x|^\nu]} \right) \tag{61}$$

As it is easy to verify, $\nu = 2$ yields the Gaussian, $\nu = 1$ the Laplacian, and $\nu \longrightarrow \infty$ the uniform density function. The values of $\nu < 2$ give rise to supergaussian densities, $\nu > 2$ to subgaussian ones.

Example of uncorrelated but not independent random variables:

Assume that ξ and η are discrete valued and their joint probabilities are 0.25 for any of (0,1), (0,-1), (1,0) and (-1,0). Then it is easy to calculate that ξ and η are uncorrelated. However, they are not independent since

$$E[\xi^2\eta^2] = 0 \neq 0.25 = E[\xi^2]E[\eta^2] \tag{62}$$

Principal Components Analysis (PCA) (also known as Karhunen-Loeve transform) seeks to find a linear transformation of an n-dimensional random variable $\mathbf{x}$, $\mathbf{y} = \mathbf{x}^T\mathbf{Q}$, such that the components of $\mathbf{y}$ are mutually uncorrelated, the column vectors of $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$ form an orthonormal basis ($\mathbf{q}_i\mathbf{q}_j = \delta_{ij}$), and $\mathbf{q}_1$ points in the direction of the largest variance, $\mathbf{q}_1$ points in the direction of the second largest variance (under the given constraints), and so on.

$$\mathbf{C_y} = \mathbf{Q}^T\mathbf{C_x}\mathbf{Q} = \mathbf{D} = Diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) \tag{63}$$

Since $\mathbf{Q}^T = \mathbf{Q}^{-1}$ for an orthonormal matrix

$$\mathbf{C_x}\mathbf{Q} = \mathbf{D}\mathbf{Q} \tag{64}$$

This is the familiar eigenvalue problem to which the solution is an orthogonal matrix $\mathbf{Q}$ whose column vectors are the eigenvectors of $\mathbf{C_x}$, and the eigenvalues of $\mathbf{C_x}$ are equal to the ${\sigma_i}^2$ in the diagonal of $\mathbf{D}$ (which are the variances of the components of $\mathbf{y} = \mathbf{x}^T\mathbf{Q}$). The details of this calculation are given in many standard linear algebra books (Haykin p. 396, for example). Since $\mathbf{C_x}$ is a real, symmetric matrix all of its eigenvalues are real. Let the eigenvalues be arranged in decreasing order: $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_n^2$. Now we can take the first m most significant eigenvalues and use the truncated projection operator $\mathbf{Q'} = (\mathbf{q}_1, \ldots, \mathbf{q}_m)$, where $m < n$ to transform $\mathbf{x}$ to an m-dimensional subspace such that the mean-square error between $\mathbf{x}$ and its projection on the subspace is minimal. (This is also detailed in standard textbooks.)

If we have K samples of $\mathbf{x}$ (an n-dimensional data cloud), with PCA we can compress the data to $m \leq n$ dimensional feature vectors such that the important characteristics of the data (based on second-order statistics) are preserved as much as possible in the MSE sense.