EECS 225A Statistical Signal Processing Lecture 3 - 1/28/20

Lecture 3: Joint Gaussian Distribution and WSS Stochastic Processes

Lecturer: Jiantao Jiao *Scribe: Tiffany Chien*

In a previous lecture, we derived the optimal *linear* estimator $\hat{X}(Y)$. A natural question to ask is when this linear estimator is also the actual optimal estimator, i.e. when there is no non-linear estimator better than it. It turns out to be the case when (X, Y) is jointly Gaussian. However, it is not the only case where the optimal linear estimator is also the optimal non-linear estimator. Indeed, if Y only takes values in a set with cardinality two, then any deterministic function of Y can be written as a linear function of Y.

1 Joint Gaussian distribution and Gaussian random vectors

We first review the definition and properties of joint Gaussian distribution and Gaussian random vectors. For a detailed exposition, the readers are referred to [1, Section 3.4].

We say that a random variable X is Gaussian with mean μ and variance $\sigma^2 > 0$ if X has probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

with notation

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

In the degenerate case $\sigma^2 = 0$, we say the random variable X is Gaussian with mean μ and variance 0 if $\mathbb{P}(X = \mu) = 1$. It turns out that the general way to describe (multivariate) Gaussian distribution is via the characteristic function. For $X \sim \mathcal{N}(\mu, \sigma^2)$, the characteristic function $\Phi_X(u)$ is given by

$$\Phi_X(u) \triangleq \mathbb{E}[e^{juX}] = \exp\left(-\frac{u^2\sigma^2}{2} + j\mu u\right).$$

We say $X \in \mathbb{R}^d$ is a Gaussian random vector if every finite linear combination of the coordinates of X is a Gaussian random variable. We write $X \sim \mathcal{N}(\mu, \Sigma)$ if X is a Gaussian random vector with mean vector μ and covariance matrix Σ. It has the following properties:

- The characteristic function of an $\mathcal{N}(\mu, \Sigma)$ Gaussian random vector is given by

$$\Phi_X(u) \triangleq \mathbb{E}[e^{ju^T X}] = \exp(ju^T\mu - \frac{1}{2}u^T\Sigma u)$$

- An $\mathcal{N}(\mu, \Sigma)$ random vector $X \in \mathbb{R}^d$ such that Σ is non-singular has a probability density function given by

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right).$$

  Any random vector such that its covariance matrix is singular does not have a pdf. Here $|\Sigma|$ denotes the determinant of the matrix Σ.

- If X and Y are jointly Gaussian vectors, then they are independent if and only if $\Sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] = 0$.

- Affine transformation: if $X \sim \mathcal{N}(\mu, \Sigma)$, then

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T).$$

The next theorem characterizes the conditional distribution for joint Gaussian distributions.

Theorem 1. *Suppose real-valued random vectors X, Y are jointly Gaussian*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} \right)$$

Then, there exists (one version) of the regular probability distribution function for $X|Y$ which is jointly Gaussian:

$$X|Y \sim \mathcal{N}\Big(\mu_X + \Sigma_{XY}\Sigma_Y^{-1}(Y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Big)$$

It implies that

$$\begin{aligned} \mathbb{E}[X|Y] &= \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(Y - \mu_Y) \\ \mathbb{E}[(X - \mathbb{E}[X|Y])(X - \mathbb{E}[X|Y])^T] &= \Sigma_{XX} - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} = \Sigma_e. \end{aligned}$$

If Σ_Y is singular then we may replace Σ_Y^{-1} by its pseudoinverse.

A few remarks are in order. This result shows that to derive the optimal linear estimator, there are "two" possible approaches: one is to use the orthogonality principle, and the other is to assume everything is jointly Gaussian and compute the conditional expectation. Both approaches may be easy to execute for simple problems, but for problems with complicated structures usually one approach will stand out and the idea of using both approaches is usually called the *Gaussian trick* in the literature.

This result also shows that for given mean and covariance structures of (X, Y), jointly Gaussian distributions attain the *worst* estimation error. Indeed, for any P_{XY}, and the optimal linear estimator $\hat{X}_{\text{Lin}}$ we have

$$\min_{g(y)} \mathbb{E}[\|X - g(Y)]\|^2 \leq \mathbb{E}[\|X - \hat{X}_{\text{Lin}}\|_2^2] = \mathsf{Tr}(\Sigma_e) \tag{1}$$

But in the Gaussian case, we find that this inequality is tight.

Proof To demonstrate the power of the "Gaussian trick" we prove the theorem without using the concrete expressions for the pdf of (X, Y). In particular, (X, Y) does not have a pdf if its joint covariance matrix is singular.

Define the optimal linear (in fact, affine) estimator $\hat{X}_L(Y)$ and let $e = X - \hat{X}_L$. By the orthogonality principle, we know $\mathbb{E}[e] = 0$, $\mathsf{Cov}(e, Y) = 0$. Since (Y, e) is obtained from (X, Y) by affine transformations, they are jointly Gaussian. Since $\mathsf{Cov}(e, Y) = 0$, e and Y are independent.

Since $X = e + \hat{X}_L(Y)$, $\hat{X}_L$ is a function of Y, e is independent of Y with covariance Σ_e, we know conditioned on Y, $e \sim \mathcal{N}(0, \Sigma_e)$. Hence, conditioned on Y, X is nothing but the sum of a deterministic vector $\hat{X}_L(Y)$ and a Gaussian random vector $\mathcal{N}(0, \Sigma_e)$, which is distributed as

$$\mathcal{N}(\hat{X}_L(Y), \Sigma_e).$$

□

2 Discrete-time random processes

Definition 2. *A discrete-time random process is a countably infinite collection of random variables on the same probability space:* $\{X(n) : n \in \mathbb{Z} \text{ or } \mathbb{N}\}$*, with mean function* $\mu_n = \mathbb{E}[X(n)]$ *and auto-correlation function* $R_X(n_1, n_2) = \mathbb{E}[X(n_1)X(n_2)^*]$*, where* $*$ *represents the conjugate transpose operation.*

Definition 3. *A wide sense stationary process (WSS) is a discrete-time random process that satisfies the following conditions:*

1. $\mu_n = \mu$*: mean is time invariant*
2. $R_X(n_1, n_2) = R_X(n_1 - n_2)$*: autocorrelation function is a function of only the difference* $n_1 - n_2$
3. $\mathbb{E}[|X(n)|^2] < \infty, \ \forall n$

Because of condition (2), we redefine $R_X(n_1, n_2)$ as $R_X(k)$ where $k = n_1 - n_2$, and call it the auto-correlation function of $X(n)$.

We list some properties of $R_X(k)$ below.

1. $R_X(k) = R_X^*(-k)$
2. $R_X(0) \geq |R_X(k)| \ \forall k$. Indeed,

$$\begin{aligned} |R_X(k)| &= |\mathbb{E}[X_n X_{n-k}^*]| \\ &\leq \sqrt{\mathbb{E}[X_n X_n^*]}\sqrt{\mathbb{E}[X_{n-k} X_{n-k}^*]} \\ &= \sqrt{R_X(0)R_X(0)} \\ &= R_X(0). \end{aligned}$$

3. Let d be a positive integer. The following three conditions are equivalent:
   - $R_X(d) = R_X(0)$
   - $\mathbb{P}(X[n+d] = X[n]) = 1$ for all $n \in \mathbb{Z}$
   - $R_X(d+n) = R_X(n)$ for all $n \in \mathbb{Z}$ (i.e., periodic with period d)

   Proof Suppose the first statement is true. Since $R_X(0)$ is real-valued, so is $R_X(d)$, yielding

$$\begin{aligned} \mathbb{E}[|X_{n+d} - X_n|^2] &= \mathbb{E}[X_{n+d}X_{n+d}^* - X_{n+d}X_n^* - X_n X_{n+d}^* + X_n X_n^*] \\ &= R_X(0) - R_X(d) - R_X^*(d) + R_X(0) \\ &= 0, \end{aligned}$$

   which implies the second statement. Since two random variables that are equal with probability one have the same expectation, the second statement implies that for any $n \in \mathbb{Z}$,

$$R_X(n+d) = \mathbb{E}[X_{n+d}X_0^*] = \mathbb{E}[X_n X_0^*] = R_X(n).$$

□

3 Harmonic process

We define the DTFT of the autocorrelation function $R_X(k)$ as the power spectral density (PSD) $S_X(\omega)$ of WSS process X:

$$R_X(k) \longleftrightarrow S_X(\omega)$$

Why is it meaningful to look at its DTFT (and why do we call it the power spectral density)? We look at the example of a random harmonic function for some evidence, and we will make the connection precise in the next lecture.

Define stochastic process $X(n) = \sum_{m=1}^{N} A_m \cos(\omega_m n + \Phi_m)$, where $\omega_m, m \in [N]$ are deterministic frequencies, $\Phi_m \sim U[-\pi, \pi]$ are random phases such that they are mutually independent across m, and $A_m, m \in [N]$ are orthogonal zero mean real-valued random variables independent of all the random phases Φ_m. Concretely, $\mathbb{E}[A_m] = 0$ and $\mathbb{E}[A_m A_{m'}] = \sigma_m^2 \delta_{mm'}$ ($\delta_{mm'} = 1$ if $m = m'$, else 0).

In other words, $X(n)$ is a linear combination of random sinusoids.

We will show X is indeed WSS, and confirm that our intuition that the power of X is only at frequencies ω_m is reflected in $S_X(\omega)$, thus justifying it as the 'power spectral density'.

$$\mu_n = \mathbb{E}[X(n)] = 0$$

$$R_X(k) = \mathbb{E}[X_{n+k} X_n] = \sum_{m=1}^{N} \sum_{l=1}^{N} \mathbb{E}[A_m A_l \cos(\omega_m (n+k) + \Phi_m) \cos(\omega_l n + \Phi_l)] \quad \text{(double sum for all cross terms)}$$

condition on $\{\Phi_m\}_{m=1}^N \rightarrow$ only terms where $m = l$ are nonzero because $\{A_m\}$ are uncorrelated

$$= \sum_{m=1}^{N} \mathbb{E}[A_m^2 \cos(\omega_m (n+k) + \Phi_m) \cos(\omega_m n + \Phi_m)]$$

Recall: $\cos\alpha \cos\beta = \frac{1}{2}(\cos(\alpha+\beta) + \cos(\alpha-\beta))$.

Here, the $\alpha + \beta$ terms are all zero because Φ_m is uniform on $[-\pi, \pi]$, so cosine is zero in expectation.

Considering only the $\alpha - \beta$ terms (Φ_m cancels):

$$R_X(k) = \sum_{m=1}^{N} \sigma_m^2 \frac{1}{2} \cos(\omega_m k) \text{ not depending on } n$$

Taking the DTFT, $S_X(\omega) = \frac{1}{2} \sum_{m=1}^{N} \sigma_m^2 \pi (\delta(\omega - \omega_m) + \delta(\omega + \omega_m))$

And we see that it is only nonzero at ω_m and $-\omega_m$, corresponding to the signal only having power at those frequencies.

Interested readers may have observed that the harmonic process is in fact a *deterministic* process in the sense that once observing a finite interval of X, one can determine the unique random variables $\{A_m\}_{m=1}^N, \{\Phi_m\}_{m=1}^N$, thus determining all the future values of X. We will later show in this course that not only any WSS process with line spectrum (a countable collection of delta functions) is *deterministic*, but also any bandlimited WSS process. Here by bandlimited we mean that $S_X(\omega) = 0$ for a set of positive Lebesgue measure.

References

[1] B. Hajek, *Random processes for engineers.* Cambridge university press, 2015.