Jointly Gaussian random variables, MMSE and linear MMSE estimation

Namrata Vaswani, Iowa State University

April 8, 2012

Most notes are based on Chapter IV-B and Chapter V of Poor's Introduction to Signal Detection and Estimation book [1].

1 Jointly Gaussian random variables

1. The $n \times 1$ random vector X is jointly Gaussian if and only if the scalar

$$u^T X$$

is Gaussian distributed for all $n \times 1$ vectors u

2. The random vector X is jointly Gaussian if and only if its characteristic function, $C_X(u) := \mathbb{E}[e^{iu^T X}]$ can be written as

$$C_X(u) = e^{iu^T \mu} e^{-u^T \Sigma u/2}$$

where $\mu = \mathbb{E}[X]$ and $\Sigma = cov(X)$.

- Proof: X is j G implies that $V = u^T X$ is G with mean $u^T \mu$ and variance $u^T \Sigma u$. Thus its characteristic function, $C_V(t) = e^{itu^T \mu} e^{-t^2 u^T \Sigma u/2}$. But $C_V(t) = \mathbb{E}[e^{itV}] = \mathbb{E}[e^{itu^T X}]$. If we set $t = 1$, then this is $\mathbb{E}[e^{iu^T X}]$ which is equal to $C_X(u)$. Thus, $C_X(u) = C_V(1) = e^{iu^T \mu} e^{-u^T \Sigma u/2}$.
- Proof (other side): we are given that the charac function of X, $C_X(u) = \mathbb{E}[e^{iu^T X}] = e^{iu^T \mu} e^{-u^T \Sigma u/2}$. Consider $V = u^T X$. Thus, $C_V(t) = \mathbb{E}[e^{itV}] = C_X(tu) = e^{iu^T \mu} e^{-t^2 u^T \Sigma u/2}$. Also, $\mathbb{E}[V] = u^T \mu$, $var(V) = u^T \Sigma u$. Thus V is G.

3. The random vector X is jointly Gaussian if and only if its joint pdf can be written as

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^n det(\Sigma)} e^{-(X-\mu)^T \Sigma^{-1} (X-\mu)/2} \tag{1}$$

- Proof: follows by computing the characteristic function from the pdf and vice versa

4. The random vector X is j G if and only if it can be written as an affine function of i.i.d. standard Gaussian r.v's.
   - Proof: if $X = AZ + a$ where $Z \sim \mathcal{N}(0, I)$, then easy to show that X has joint pdf given by (1) and thus it is j G.
   - Proof (other side): if X is j G, then it has the joint pdf given by (1). Then can show that $Z := \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I)$, i.e. it is i.i.d. standard G. Thus, $X = \Sigma^{1/2}Z + \mu$, i.e. it is an affine function of Z.

5. The random vector X is j G if and only if it can be written as an affine function of jointly Gaussian r.v's.
   - Proof: Suppose X is an affine function of a j G r.v. Y, i.e. $X = BY + b$. Since Y is j G, by 4, it can be written as $Y = AZ + a$ where $Z \sim \mathcal{N}(0, I)$ (i.i.d. standard Gaussian). Thus, $X = BAZ + (Ba + b)$, i.e. it is an affine function of Z, and thus, by 4, X is j G.
   - Proof (other side): X is j G. So by 4, it can be written as $X = BZ + b$. But $Z \sim \mathcal{N}(0, I)$ i.e. Z is a j G r.v.

Properties

1. If X_1, X_2 are j G, then the conditional distribution of X_1 given X_2 is also j G
2. If the elements of a j G r.v. X are pairwise uncorrelated (i.e. non-diagonal elements of their covariance matrix are zero), then they are also mutually independent.
3. Any subset of X is also j G.

2 Bayesian Minimum Mean Squared Error (MMSE) estimation

1. X is the unknown, Y is the observation. We assume that X itself is a random variable with a prior distribution that is known. We are also given the conditional distribution of Y given X.
2. Bias of a Bayesian estimator $\hat{X}(Y)$ is defined as

$$\mathbb{E}[\hat{X}(Y)] - E[X] \tag{2}$$

   where $\mathbb{E}[.]$ means we take expectation over all random variables (here X, Y).

3. Bayesian MSE of an estimator $\hat{X}(Y)$ is

$$\mathbb{E}[\|X - \hat{X}(Y)\|^2] \tag{3}$$

4. Claim: $\mathbb{E}[X|Y]$ is the minimum MSE (MMSE) estimator of X from Y. Proof:

   (a) We try to show that

   $$\mathbb{E}[\|X - \mathbb{E}[X|Y]\|^2] \leq \mathbb{E}[\|X - \hat{X}(Y)\|^2] \tag{4}$$

   (b) To do this, add and subtract $\mathbb{E}[X|Y]$ from RHS, expand and show that the cross term is zero. To show cross term is zero, use law of iterated expectations. Thus,

   $$\begin{aligned} \mathbb{E}[\|X - \hat{X}(Y)\|^2] &= \mathbb{E}[\|X - \mathbb{E}[X|Y] + \mathbb{E}[X|Y] - \hat{X}(Y)\|^2] \\ &= \mathbb{E}[\|X - \mathbb{E}[X|Y]\|^2] + \mathbb{E}[\|\mathbb{E}[X|Y] - \hat{X}(Y)\|^2] + 2\text{cross} \end{aligned} \tag{5}$$

   where

   $$\begin{aligned} \text{cross} &= \mathbb{E}[(\mathbb{E}[X|Y] - \hat{X}(Y))^T (X - \mathbb{E}[X|Y])] \\ &= \mathbb{E}_Y[\mathbb{E}[(\mathbb{E}[X|Y] - \hat{X}(Y))^T (X - \mathbb{E}[X|Y])|Y]] \\ &= \mathbb{E}_Y[(\mathbb{E}[X|Y] - \hat{X}(Y))^T \mathbb{E}[(X - \mathbb{E}[X|Y])|Y]] \\ &= \mathbb{E}_Y[(\mathbb{E}[X|Y] - \hat{X}(Y))^T [\mathbb{E}[X|Y] - \mathbb{E}[X|Y])]] = 0 \end{aligned} \tag{6}$$

   The second row uses law of iterated expectations, the third row follows because $\mathbb{E}[X|Y]$ and $\hat{X}(Y)$ are constants given Y. The last row follows because $\mathbb{E}[X|Y]$ is a constant given Y.

   (c) Using the above and since $\mathbb{E}[\|\mathbb{E}[X|Y] - \hat{X}(Y)\|^2] \geq 0$, the result follows.

5. Claim: Variance of the error of $\mathbb{E}[X|Y]$ is smallest in any direction, i.e. for any unit vector, c,

   $$c^T \mathbb{E}[(X - \mathbb{E}[X|Y])(.)^T]c \leq c^T \mathbb{E}[(X - \hat{X}(Y))(.)^T]c \tag{7}$$

   Proof:

   (a) Consider $Z := c^T X$. By the previous result, its MMSE estimator is $\mathbb{E}[Z|Y] = c^T \mathbb{E}[X|Y]$. Thus,

   $$\mathbb{E}[(c^T X - c^T \mathbb{E}[X|Y])^2] \leq \mathbb{E}[(Z - \hat{Z}(Y))^2] \tag{8}$$

   (b) Using $(c^T v)^2 = c^T v v^T c$ and using $Z = c^T X$, we get

   $$\mathbb{E}[c^T (X - \mathbb{E}[X|Y])(.)^T c] \leq \mathbb{E}[(c^T X - \hat{Z}(Y))^2] \tag{9}$$

(c) The above is true for all estimators of Z, $\hat{Z}(Y)$. In particular, it is true if we consider the class of estimators that can be written as $\hat{Z}(Y) = c^T \hat{X}(Y)$. Thus,

$$\mathbb{E}[c^T(X - \mathbb{E}[X|Y])(.)^T c] \leq \mathbb{E}[c^T(X - \hat{X}(Y))(.)^T c] \tag{10}$$

This finishes the proof.

6. By letting $c = e_i$ (e_i is a vector with a one at the i^{th} location and zero everywhere else), we see that $\mathbb{E}[X_i|Y]$ is the MMSE of X_i from Y.

7. Claim: $\mathbb{E}[X|Y]$ is unbiased, i.e. $\mathbb{E}[\mathbb{E}[X|Y]] - E[X] = 0$.

   (a) Proof: This follows because $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

8. Read Chapter IV-B of Poor's book.

3 Linear MMSE estimation

1. We call this linear MMSE estimation, but that is a misnomer, we actually look for the minimum MSE estimator among all affine functions of the observation, i.e. among all functions of the form $HY + c$.

2. Let the set of affine estimators of X from Y be

$$\mathcal{H} := \{\hat{X}(Y) : \hat{X}(Y) = HY + c\}$$

The linear MMSE estimator $\hat{X}_{LMMSE}(Y)$ is defined as the solution of

$$\min_{\hat{X}(Y) \in \mathcal{H}} \mathbb{E}[\|X - \hat{X}(Y)\|^2] \tag{11}$$

for a matrix H and a vector c.

3. Orthogonality Principle 1: $\hat{X}_L(Y) \in \mathcal{H}$ is the linear MMSE of X from Y if and only if

$$\mathbb{E}[(X - \hat{X}_L(Y))Z^T] = 0 \text{ for all } Z \in \mathcal{H} \tag{12}$$

Proof (one side):

   (a) Suppose $\hat{X}_L(Y) \in \mathcal{H}$ satisfies (12), but it is not the LMMSE, i.e. there exists an $\hat{X}_0(Y) \neq \hat{X}_L(Y)$ such that $\hat{X}_0(Y) \in \mathcal{H}$ and

$$\mathbb{E}[\|X - \hat{X}_0(Y)\|^2] \leq \mathbb{E}[\|X - \hat{X}_L(Y)\|^2] \tag{13}$$

   (b) We can write the LHS as $\mathbb{E}[\|X - \hat{X}_0(Y)\|^2] = \mathbb{E}[\|X - \hat{X}_L(Y) + \hat{X}_L(Y) - \hat{X}_0(Y)\|^2] = \mathbb{E}[\|X - \hat{X}_L(Y)\|^2] + \mathbb{E}[\|\hat{X}_L(Y) - \hat{X}_0(Y)\|^2] + 2\text{cross}$ where

$$\text{cross} = \mathbb{E}[(\hat{X}_L(Y) - \hat{X}_0(Y))^T(X - \hat{X}_L(Y))] \tag{14}$$

(c) Since $\hat{X}_L(Y) \in \mathcal{H}$ and $\hat{X}_0(Y) \in \mathcal{H}$, thus $(\hat{X}_L(Y) - \hat{X}_0(Y)) \in \mathcal{H}$. Thus by (12), $\mathbb{E}[(X - \hat{X}_L(Y))(\hat{X}_L(Y) - \hat{X}_0(Y))^T] = 0$.

(d) Using trace(AB) = trace(BA) and the fact that trace is a linear operator, we can see that for any two n dimensional vectors X_1, X_2,

$$\mathbb{E}[X_2^T X_1] = \mathbb{E}[\text{trace}(X_2^T X_1)] = \mathbb{E}[\text{trace}(X_1 X_2^T)] = \text{trace}(\mathbb{E}[X_1 X_2^T]) \qquad (15)$$

(e) Using (15), cross = trace$(\mathbb{E}[(X - \hat{X}_L(Y))(\hat{X}_L(Y) - \hat{X}_0(Y))^T])$, thus cross = 0.

(f) Thus, $\mathbb{E}[\|X - \hat{X}_0(Y)\|^2] = \mathbb{E}[\|X - \hat{X}_L(Y)\|^2] + \mathbb{E}[\|\hat{X}_L(Y) - \hat{X}_0(Y)\|^2] \geq \mathbb{E}[\|X - \hat{X}_L(Y)\|^2]$ and this is a contradiction to (13) unless $\hat{X}_0(Y) = \hat{X}_L(Y)$.

Proof (other side):

(a) Suppose $\hat{X}_L(Y)$ is the LMMSE but it does not satisfy (12), i.e. there exists a $Z_0 \in \mathcal{H}$ for which $\mathbb{E}[(X - \hat{X}_L(Y))Z_0^T] \neq 0$.

(b) Define another estimator, $\hat{X}_0 = \hat{X}_L + BZ_0$.

(c) Let us try to find B to minimize the MSE, $\mathbb{E}[\|X - \hat{X}_L - BZ_0\|^2]$. If we differentiate this and set to zero, we get $B_{\min} = \mathbb{E}[(X - \hat{X})Z_0^T]\mathbb{E}[Z_0 Z_0^T]^{-1}$. Thus, we consider the estimator $\hat{X}_0 = \hat{X}_L + B_{\min} Z_0$.

(d) Consider $\mathbb{E}[\|X - \hat{X}_0\|^2]$ and simplify it:

$$\begin{aligned}
\mathbb{E}[\|X - \hat{X}_0\|^2] &= \mathbb{E}[\|X - \hat{X}_L - B_{\min} Z_0\|^2] \\
&= \mathbb{E}[\|X - \hat{X}_L\|^2] + \mathbb{E}[Z_0^T B_{\min}^T B_{\min} Z_0] - 2\mathbb{E}[Z_0^T B_{\min}^T (X - \hat{X}_L)] \quad (16)
\end{aligned}$$

(e) Using (15), we can rewrite the second term of (16) as

$$\begin{aligned}
\mathbb{E}[Z_0^T B_{\min}^T B_{\min} Z_0] &= \text{trace}(\mathbb{E}[B_{\min} Z_0 Z_0^T B_{\min}^T]) \\
&= \text{trace}(B_{\min}\mathbb{E}[Z_0 Z_0^T]B_{\min}^T] \\
&= \text{trace}(\mathbb{E}[(X - \hat{X})Z_0^T]\mathbb{E}[Z_0 Z_0^T]^{-1}\mathbb{E}[(X - \hat{X})Z_0^T]^T) \quad (17)
\end{aligned}$$

(f) Using (15) we can also rewrite the third term of (16) as

$$\begin{aligned}
\mathbb{E}[Z_0^T B_{\min}^T (X - \hat{X}_L)] &= \text{trace}(\mathbb{E}[(X - \hat{X}_L) Z_0^T B_{\min}^T]) \\
&= \text{trace}(\mathbb{E}[(X - \hat{X}_L)Z_0^T]B_{\min}^T) \\
&= \text{trace}(\mathbb{E}[(X - \hat{X}_L)Z_0^T]\mathbb{E}[Z_0 Z_0^T]^{-1}\mathbb{E}[(X - \hat{X})Z_0^T]^T) \quad (18)
\end{aligned}$$

(g) Substituting the last two equations into (16),

$$\mathbb{E}[\|X - \hat{X}_0\|^2] = \mathbb{E}[\|X - \hat{X}_L\|^2] - \text{trace}(\mathbb{E}[(X - \hat{X}_L)Z_0^T]\mathbb{E}[Z_0 Z_0^T]^{-1}\mathbb{E}[(X - \hat{X})Z_0^T]^T) \quad (19)$$

The second term is the trace of a positive semi-definite matrix and hence it is non-negative. Thus, $\mathbb{E}[\|X - \hat{X}_0\|^2] \leq \mathbb{E}[\|X - \hat{X}_L\|^2]$, i.e. $\hat{X}_L$ is not the LMMSE. This is a contradiction.

4. Orthogonality Principle 2: $\hat{X}_L(Y) \in \mathcal{H}$ is the linear MMSE of X from Y if and only if

$$\mathbb{E}[(X - \hat{X}_L(Y))] = 0 \text{ and } \mathbb{E}[(X - \hat{X}_L(Y))Y^T] = 0 \tag{20}$$

   Proof (one side): follows easily from the first one.

   (a) Suppose $\hat{X}_L(Y)$ is the LMMSE. Then by orthogonality principle 1,

$$\mathbb{E}[(X - \hat{X}_L(Y))Z^T] = 0 \text{ for all} Z \in \mathcal{H}$$

   (b) If we set $H = 0$ in $\mathcal{H}$, then we get $\mathbb{E}[(X - \hat{X}_L(Y))c^T] = 0$. Since c is a constant, this means that $\mathbb{E}[(X - \hat{X}_L(Y))] = 0$.

   (c) If we set $H = I$, $c = 0$, in $\mathcal{H}$, then we get $\mathbb{E}[(X - \hat{X}_L(Y))Y^T] = 0$.

   Proof (other side): follows directly from first one

   (a) Suppose $\mathbb{E}[(X - \hat{X}_L(Y))] = 0$ and $\mathbb{E}[(X - \hat{X}_L(Y))Y^T] = 0$. Thus, $\mathbb{E}[(X - \hat{X}_L(Y))Y^T H^T] = 0$.

   (b) Using, $\mathbb{E}[(X - \hat{X}_L(Y))] = 0$ we get $\mathbb{E}[(X - \hat{X}_L(Y))c^T] = 0$.

   (c) Combining the above two, we get $\mathbb{E}[(X - \hat{X}_L(Y))(Y^T H^T + c^T)] = \mathbb{E}[(X - \hat{X}_L(Y))(HY + c)^T] = 0$.

   (d) Thus, $\mathbb{E}[(X - \hat{X}_L(Y))Z^T] = 0$ for all $Z \in \mathcal{H}$. By orthogonality principle 1, $\hat{X}_L(Y)$ is the linear MMSE.

5. Wiener-Hopf equations: using the orthogonality principle 2, we can derive the Weiner-Hopf equations to compute an LMMSE estimate.

   (a) The LMMSE estimate is of the form $\hat{X}_L = H_L Y + c_L$. Using the ortho principle, this satisfies

$$\begin{aligned} \mathbb{E}[(X - H_L Y - c_L)] = 0, \text{ and} \\ \mathbb{E}[(X - H_L Y - c_L)Y^T] = 0 \end{aligned} \tag{21}$$

   (b) Using the first equation of (21)

$$c_L = \mathbb{E}[(X - H_L Y)] = \mathbb{E}[X] - H_L \mathbb{E}[Y] \tag{22}$$

   Using the second equation of (21) and above,

$$\mathbb{E}[(X - H_L Y - c_L)Y^T] = \mathbb{E}[((X - \mathbb{E}[X]) - H_L(Y - \mathbb{E}[Y]))Y^T] = 0 \tag{23}$$

(c) Thus,

$$\mathbb{E}[(X - \mathbb{E}[X])Y^T] = H_L \mathbb{E}[(Y - \mathbb{E}[Y]))Y^T] \tag{24}$$

Since $cov(X,Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] = \mathbb{E}[(X - \mathbb{E}[X])Y^T]$, thus, we get

$$H_L = cov(X,Y)cov(Y,Y)^{-1} \tag{25}$$

and so

$$c_L = \mathbb{E}[X] - cov(X,Y)cov(Y,Y)^{-1}\mathbb{E}[Y] \tag{26}$$

6. Special cases:

(a) If the sequence $Y_1, Y_2, \ldots Y_n$ is wide sense stationary, then $cov(Y,Y)$ is a Toeplitz matrix. This allows for efficient matrix inversion: $O(n^2)$ cost compared to $O(n^3)$ for any general matrix.

(b) If $Y = [Y_1, Y_2, \ldots Y_t]$ and $X = Y_{t+1}$, then X, Y are jointly wide sense stationary. In this case, the Levinson algorithm can be used to find the solution efficiently.

(c) Non-causal Wiener filter: estimate X_t using $\{Y_\tau\}_{\tau=-\infty}^{\infty}$, when they are jointly WSS

- Due to joint WSS assumption, the problem can be converted into frequency domain, and one gets an expression for the squared magnitude of the filter's frequency response.
- Since the filter can be non-causal, one can just pick a zero phase filter.

(d) Causal Wiener: estimate X_t using $\{Y_\tau\}_{\tau=-\infty}^{t}$ when they are jointly WSS

- Can design a causal Wiener filter also in the frequency domain (see Chapter V of Poor's book or see DSP texts).
- If X_t's and Y_t's satisfy the linear dynamic model (model used by Kalman filter) and are jointly WSS, then the Kalman filter update exactly gives the causal Wiener solution.

References

[1] H. Vincent Poor, *An Introduction to Signal Detection and Estimation*, Springer, second edition.