I. Summary

Here are the key ideas and results:

• Theorem 1: If \((X, Y)\) are \(N(0, K)\) with \(|K| \neq 0\), then \(f_{X|Y}[y] = N(Ay, \Sigma)\) where \(A, \Sigma\) are given by (1).

• Theorem 2: Under same assumptions, \(E[X|Y] = (2)\).

• The square integrable RVs form a Hilbert space (Section IV).

II. Example - Continued

Recall our little example from the end of L4:

Assume that \((X, Y)^T = N(0, K)\) with

\[
K = \begin{bmatrix}
3 & 1 \\
1 & 1
\end{bmatrix}.
\]

We found that \(X - Y \perp Y\), so that \(X - Y\) and \(Y\) are independent. We used that to calculate \(E[X|Y] = E[X - Y + Y|Y] = Y\).

There is another useful consequence of the independence of \(X - Y\) and \(Y\): Given that \(Y = y\), we see that

\[
X = (X - Y) + Y |_{Y=y} = N(y, \sigma^2)
\]

where

\[
\sigma^2 = \text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y) = 3 + 1 - 2 = 2.
\]

That is,

\[
f_{X|Y}[y] = N(y, 2).
\]

Here are a few key observations:

• The mean of \(X\) given \(Y = y\) depends on \(y\) (it is \(E[X|Y = y] = y\)).

• However, the variance of \(X\) given \(Y = y\) does not depend on \(y\)! Again, this fact follows from the independence of \(X - Y\) and \(Y\). The ‘noise’ \(X - Y\) that is added to \(Y\) to get \(X\) does not depend on \(Y\).

• The variance of \(X\) given \(Y\) is smaller than the variance of \(X\). (Here, it is 2 instead of 3.)

We generalize these observations in the next section.

III. Conditional Distribution

*Theorem 1*: Conditional Distribution

Assume that \((X, Y)\) are \(N(0, K)\) with \(|K| \neq 0\). Then, given \(Y = y\), \(X = N(Ay, \Sigma)\) where

\[
A = K_{XY}K_Y^{-1} \quad \text{and} \quad \Sigma = K_X - K_{XY}K_Y^{-1}K_{YX}.
\]  

**(Proof):**

First note that \(|K_Y| \neq 0\). This follows from the fact that \(K\) is symmetric, so that it is positive semi-definite and is positive definite if and only if \(|K| \neq 0\). However, if \(|K_Y| = 0\), then there is some \(a \neq 0\) such that \(K_Ya = 0\) and, consequently, \([0, a^T]K[0, a^T]^T = 0\). Hence \(K\) cannot be positive definite if \(|K_Y| = 0\). (See [2], Section VIII.B). Second, observe that

\[
Z := X - AY \perp Y \quad \text{if} \quad K_{XY} = AK_Y, \quad \text{i.e.,} \quad A = K_{XY}K_Y^{-1}.
\]

Then, \(X = AY + Z\) where \(Z\) and \(Y\) are independent. Also,

\[
\Sigma := K_Z = E(X - AY)(X - AY)^T
= K_X - AK_{YX} - K_{XY}A^T + AK_YA^T
= K_X - K_{XY}K_Y^{-1}K_{YX}.
\]
One interesting observation is that the variance $\Sigma$ of $X$ given that $Y = y$ does not depend on the value of $y$. Another observation is the reduction of the variance $K_X$ due to the observation.

We can also derive the following consequence.

**Theorem 2:** Conditional Expectation

Under the same assumptions as Theorem 1,

$$E[X|Y] = K_{XY}K_Y^{-1}Y.$$  \hfill (2)

**Some Comments:** Note the following steps that lead to the above theorem;

* $X - AY \perp Y \Rightarrow \{X - AY, Y\}$ independent
* $\Rightarrow \{X - AY, g(Y)\}$ independent for all $g(\cdot)$
* $\Rightarrow X - AY \perp g(Y)$ for all $g(\cdot)$
* $\Rightarrow AY = E[X|Y].$

The first step is valid because $X, Y$ are jointly Gaussian. If they were not, all we could conclude would be that

$$X - AY \perp \mu + BY, \forall (\mu, B).$$

This would show that $AY$ is the linear function $g(Y)$ of $Y$ that minimizes $E(||X - g(Y)||^2).$ In other words, this would show that $AY$ is the linear least squares estimator of $X$ given $Y.$ To show that $AY$ is the function, possibly nonlinear, $g(Y)$ that minimizes $E(||X - g(Y)||^2),$ we needed the property that orthogonal implies independence. Generally, this is not true. It’s OK for JG. The simplification is remarkable....

**Exercise 1:** Extend the results of Theorems 1 and 2 to nonzero-mean random variables.

**IV. Vector Space of Random Variables**

In this section, we make a little digression to explore the geometry of random variables. We consider the collection of square integrable random variables $X$ on a given probability space $\{\Omega, F, P\}.$ That is, we look at $S := \{X \mid X$ is a RV and $E(X^2) < \infty\}.$

Note that $S$ is a linear space since any linear combination of elements of $S$ is also in $S.$ We equip this space with a scalar product as follows:

$$<X, Y> := E(XY).$$

The scalar product defines a norm:

$$||X|| := <X, X>^{1/2} = E(X^2)^{1/2}.$$ A nice result states that $S$ is a complete space. This means that, as is the case for real numbers, Cauchy sequences converge to a point in $S.$ More precisely, if $\sup_{m, k \geq n} ||X_m - X_k|| \to 0$ as $n \to \infty,$ with $X_n \in S,$ then $X_n \to X \in S.$

In particular, let us fix a random variable $Y$ and define $\mathcal{V}$ to be the subset of $S$ that consists of random variables that are functions of $Y.$ That is, $\mathcal{V} = \{Z \mid Z = g(Y) \text{ and } E(Z^2) < \infty\}.$ The linear space $\mathcal{V}$ is again complete. Given $X$ and $Y,$ we can find a sequence of random variables $Z_n \in \mathcal{V}$ such that $||X - Z_n|| \leq d + 1/n$ where $d = \inf\{||X - Z||, Z \in \mathcal{V}\}.$ One can show that $Z_n$ is Cauchy. It follows that $Z_n \to Z \in \mathcal{V}.$ That random variable $Z$ defines $E[X|Y].$

To make this construction of $\mathcal{S}$ a bit more concrete, think of a finite probability space $\Omega = \{1, \ldots, N\}$ with $P(\{i\}) = p_i$ for $i = 1, \ldots, N.$ In that case, one associates a random variable $X$ with a vector $x \in \mathbb{R}^N$ with components $x_i = X(i)\sqrt{p_i}.$ With this definition, if we have two random variables $X, Y$ with their associated vectors $x, y,$ then we find that

$$<X, Y> = E(XY) = \sum_{i=1}^{N} X(i)Y(i)p_i = x^TY.$$ That is, the abstract scalar product $<X, Y>$ is the ordinary scalar product in $\mathbb{R}^N.$ Thus, the geometry of the Hilbert space $S$ is the ordinary geometry of $\mathbb{R}^N.$ You can then view the general case as an extension to infinitely many dimensions. In that sense, Hilbert spaces look like $\mathbb{R}^N.$

**REFERENCES**
