Minimum Mean-Square Error Estimation

Now we will take our first look at estimating variables that are themsevles random, subject to a known probability law.

We start our discussion with a very basic problem. Suppose Y is a scalar random variable with a known pdf $f_Y(y)$. Here is a fun game: you guess what Y is going to be, then I draw a realization of Y corresponding to its probability law, then we see how close you were with your guess.

What is your best guess?

Well, that of course depends on what exactly we mean by "best", i.e. what price I pay for being a certain amount off. But if we penalize the *mean-squared error*, we know exactly how to minimize it.

Let g be your guess. The error in your guess is of course random (since the realization of Y is random), and so is the squared-error $(Y-g)^2$. We want to choose g so that the mean of the squared error is as small as possible:

$$\underset{g}{\text{minimize}} \;\; \mathrm{E}[(Y-g)^2].$$

Expanding the squared error makes it clear how to do this:

$$\mathrm{E}[(Y-g)^2] = \mathrm{E}[Y^2] - 2g\,\mathrm{E}[Y] + g^2.$$

No matter what the first moment $\mathrm{E}[Y]$ and second moment $\mathrm{E}[Y^2]$ are (as long as they are finite), the expression above is a convex quadratic function in g, and hence is minimized when its first derivative (w.r.t. g) is zero, i.e. when

$$-2\,\mathrm{E}[Y] + 2g = 0 \quad \Rightarrow \hat{g} = \mathrm{E}[Y].$$

The squared error for this choice $\hat{g}$ is of course exactly the variance of Y.

The story gets more interesting (and relevant) when we have multiple random variables, some of which we observe, some of which we do not. Suppose that two random variables (Y, Z) have joint pdf $f_{Y,Z}(y, z)$. Suppose that a realization of (Y, Z) is drawn, and I get to observe Z. What have I learned about Y?

If Y and Z are independent, then the answer is of course nothing. But if they are not independent, then the marginal distribution of Y changes. In particular, before the random variables were drawn, the (marginal) pdf for Y was

$$f_Y(y) = \int f_{Y,Z}(y, z) \ \mathrm{d}z.$$

After we observe $Z = z$, we have

$$f_Y(y|Z = z) = \frac{f_{Y,Z}(y, z)}{f_Z(z)} = \frac{f_{Y,Z}(y, z)}{\int f_{Y,Z}(y, z) \ \mathrm{d}y}.$$

Y is still a random variable, but its distribution depends on the value z that was observed for Z.

Now, given that I have observed $Z = z$, what is the best guess for Y? If by "best" we mean that which minimizes the mean squared error, it is the conditional mean. That is, the minimizer of

$$\underset{g}{\text{minimize}} \ \ \mathrm{E}[(Y - g)^2|Z = z]$$

is

$$g = \mathrm{E}[Y|Z = z].$$

Notice that unlike before, g is not pre-determined, it depends on the outcome $Z = z$. We might denote

$$g(z) = \mathrm{E}[Y|Z = z].$$

In fact, since Z is a random variable, g is a priori also random, we might say

$$g(Z) = \mathrm{E}[Y|Z].$$

It is then fair to ask: what is the mean of $g(Z)$? We have[1]

$$\begin{aligned}\mathrm{E}[g(Z)] &= \mathrm{E}[\mathrm{E}[Y|Z]] \\ &= \int \mathrm{E}[Y|Z = z] f_Z(z) \ \mathrm{d}z \\ &= \mathrm{E}[Y]\end{aligned}$$

So on average, we are doing the same thing as if we didn't observe Z at all, but our MSE will in general be much better.

Multivariate Gaussian

We say that a random variable $X \in \mathbb{R}^D$ is a **Gaussian random vector** if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a symmetric positive definite matrix $\boldsymbol{R}$ such that its density can be written as

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\boldsymbol{R})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

The vector $\boldsymbol{\mu}$ is the mean of this distribution, and $\boldsymbol{R}$ is the covariance:

$$\boldsymbol{\mu} = \mathrm{E}[X], \quad \boldsymbol{R} = \mathrm{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^{\mathrm{T}}].$$

[1] That $\mathrm{E}[\mathrm{E}[Y|Z]] = \mathrm{E}[Y]$ is known as the law of *iterated expectation*. The inside E above is over Y while the outside one is over Z.

We will denote this as

$$X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R}).$$

The geometry of the density reflects the eigenstructure of $\boldsymbol{R}$ — the level-surfaces of the density are ellipsoids with the eigenvectors of $\boldsymbol{R}$ as axes, and radii are proportional to the inverses of the eigenvalues. Example:
[INSERT]

It is a classical fact that $R[i,j] = 0$ if and only if entries X_i and X_j are independent from one another. It is also a fact that if $\boldsymbol{A}$ is a $D \times D$ invertible matrix, then $\boldsymbol{A}X$ is also a Gaussian random vector with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^{\mathrm{T}}$. There is a classic transformation that makes the components of X independent from one another. If $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$ is the eigenvalue decomposition of $\boldsymbol{R}$, then

$$X' = \boldsymbol{V}^{\mathrm{T}}X$$

is a Gaussian random vector that has independent entries. This follows from the fact that $\boldsymbol{V}$ diagonalizes $\boldsymbol{R}$:

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V} = \boldsymbol{\Lambda}.$$

This transformation is known as the *Karhunen-Loeve* transform.

Gaussian Estimation

What does observing part of a Gaussian random vector tell us about the part that we do not observe? That is, suppose

$$X \sim \text{Normal}(\mathbf{0}, \boldsymbol{R}),$$

and then we observe the first $1, \ldots, p$ entries of X while entries $p+1, \ldots, D$ stay hidden. We divide X into

$$X = \begin{bmatrix} X_o \\ X_h \end{bmatrix}, \quad \text{then observe } \; X_o = \boldsymbol{x}_o.$$

What is the conditional density for $X_h | X_o = \boldsymbol{x}_o$?

It turns out that the conditional density is also Gaussian, just with a different mean and different covariance. To see this, we partition the covariance matrix into 4 parts:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_o & \boldsymbol{R}_{oh} \\ \boldsymbol{R}_{oh}^{\mathrm{T}} & \boldsymbol{R}_h \end{bmatrix}.$$

The upper left corner contains the $p \times p$ covariance matrix for the random variables that end up being observed, the lower right corner contains the $D - p \times D - p$ covariance matrix for the unobserved random variables, and $\boldsymbol{R}_{oh}$ is the *cross-correlation* matrix, that captures the dependencies between the observed and unobserved random variables.

We can also partition the inverse covariance

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D} \end{bmatrix}.$$

Using the Schur complement (see the Technical Details section below), we can write out these blocks of the inverse as

$$\begin{aligned}\boldsymbol{D} &= (\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh})^{-1} \\ \boldsymbol{C} &= -\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}\boldsymbol{D} \\ \boldsymbol{B} &= (\text{something})\end{aligned}$$

We could write down the expression for $\boldsymbol{B}$ if we really wanted to, but it is long and complicated and it ends up that we don't use it. We will use these expressions later, but to ease the notation below, we will stick with $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$.

We can now compute the conditional density using

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = \frac{f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h)}{f_{X_o}(\boldsymbol{x}_o)}.$$

The numerator is proportional to

$$\begin{aligned}f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_o^{\mathrm{T}} & \boldsymbol{x}_h^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_o \\ \boldsymbol{x}_h\end{bmatrix}\right) \\ &= \exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o + \boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right).\end{aligned}$$

The first term above does not depend on $\boldsymbol{x}_h$, so we can write the conditional density as

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = g(\boldsymbol{x}_0)\,\exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right),$$

where $g(\boldsymbol{x}_0)$ is a function that incorporates $1/f_{X_o}(\boldsymbol{x}_o)$ and $\exp(-\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o/2)$ along with some constants. We are not too worried about what g actually is, just that it does not depend on $\boldsymbol{x}_h$.

To show that $X_h|X_o$ is a Gaussian random vector, we need a density that looks like

$$(\text{stuff with no } \boldsymbol{x}_h) \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_h - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{x}_h - \boldsymbol{\mu})\right).$$

To get our conditional density in this form, we *complete the square* in the exponent. You can easily check that the following relation holds:

$$\begin{aligned}\boldsymbol{x}_o^{\mathrm{T}} \boldsymbol{C} \boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}} \boldsymbol{D} \boldsymbol{x}_h = \\ (\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}} \boldsymbol{D} (\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o) - \boldsymbol{x}_o^{\mathrm{T}} \boldsymbol{C} \boldsymbol{D}^{-1} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{x}_o.\end{aligned}$$

Again, the last term above does not depend on $\boldsymbol{x}_h$. Thus we can write

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = h(\boldsymbol{x}_o) \exp\left(-\frac{1}{2}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}} \boldsymbol{D} (\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)\right),$$

where $h(\boldsymbol{x}_o) = g(\boldsymbol{x}_o) \exp(\boldsymbol{x}_o^{\mathrm{T}} \boldsymbol{C} \boldsymbol{D}^{-1} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{x}_o/2)$. Plugging in the expressions for $\boldsymbol{C}$ and $\boldsymbol{D}$ above, we see that $X_h|X_o = \boldsymbol{x}_o$ is a Gaussian random vector

$$X_h|X_o = \boldsymbol{x}_o \ \sim \ \text{Normal}(\boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_o^{-1} \boldsymbol{x}_o, \boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_o^{-1} \boldsymbol{R}_{oh}).$$

So given the observations $X_o = \boldsymbol{x}_o$, our best (MMSE) guess for $\boldsymbol{h}_h$ is the conditional mean:

$$\hat{\boldsymbol{x}}_h = \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_o^{-1} \boldsymbol{x}_o.$$

The MSE we will incur with this choice is

$$\mathrm{E}[\|\hat{\boldsymbol{x}}_h - X_h\|_2^2 | X_o = \boldsymbol{x}_o] = \mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_o^{-1} \boldsymbol{R}_{oh}).$$

It is a fact that $\mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_o^{-1} \boldsymbol{R}_{oh}) \leq \mathrm{trace}(\boldsymbol{R}_h)$ (why?), so the observing $X_o = \boldsymbol{x}_o$ also reduces the mean-squared error associated with our best guess.

Notice that for zero-mean Gaussian random variables, $R[i,j] = 0$ if and only if X_i and X_j are independent. Above, this means that if X_o and X_k are independent, we will have $\boldsymbol{R}_{oh} = \mathbf{0}$, and the conditional distribution for X_h is no different from its original marginal (exactly as we would expect).

Conditional independence

As we just mentioned above, It is easy to interpret a zero-valued entry in the covariance matrix: $R[i,j] = 0$ means $X[i]$ and $X[j]$ are independent. But the entries of the *inverse covariance* matrix also carry interesting information about the dependency structure in X. Let

$$\boldsymbol{S} = \boldsymbol{R}^{-1}.$$

What does it mean if $S[i,j] = 0$? The answer is that $X[i]$ and $X[j]$ are independent given observations of all of the other entries $\{X[k],\ k \neq i,j\}$ in X. To see this, suppose that we partition $\boldsymbol{S}$ the same way we partitioned $\boldsymbol{R}$:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_o & \boldsymbol{S}_{oh} \\ \boldsymbol{S}_{oh}^{\mathrm{T}} & \boldsymbol{S}_h \end{bmatrix}$$

We have already seen we can use the *Schur complement* to get an expression for $\boldsymbol{S}_h$ in terms of the blocks in $\boldsymbol{R}$:

$$\boldsymbol{S}_h = (\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_0^{-1} \boldsymbol{R}_{oh})^{-1}.$$

Notice that this is exactly the inverse covariance of the conditional random vector $X_h | X_o$. Then, consider the particular case when there are two "hidden" entries in X_h, say $X_h = \{X[i], X[j]\}$) and the other $d-2$ are in X_o. The off-diagonal terms in $\boldsymbol{S}_h$ above correspond to

$S[i,j]$ and $S[j,i]$, if these are zero, then $\boldsymbol{S}_h$ is diagonal, and so is $\boldsymbol{S}_h^{-1} = \boldsymbol{R}_{h|o}$. This means that $X[i]$ and $X[j]$ are **conditionally independent** given observations of $\{X[k],\ k \neq i,j\}$.

Independence and conditional independence

Let $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, and set $\boldsymbol{S} = \boldsymbol{R}^{-1}$. Then

$$R[i,j] = 0 \quad \Leftrightarrow \quad X[i] \text{ and } X[j] \text{ are independent,}$$

and with $X_{\overline{(i,j)}} = \{X[k], k \neq i,j\}$,

$$S[i,j] = 0 \quad \Leftrightarrow \quad X[i]|X_{\overline{(i,j)}} \text{ and } X[j]|X_{\overline{(i,j)}} \text{ are independent.}$$

Gaussian Graphical Models (a first look)

The non-zero patterns in the inverse covariance $\boldsymbol{S}$ often times more descriptive of the dependency structure in X than the non-zero entries in $\boldsymbol{R}$. For example, let $Z[1], \ldots, Z[d]$ be independent Gaussian random variables with $Z[k] \sim \text{Normal}(0,1)$, and $0 < a < 1$, and set

$$\begin{aligned} X[1] &= \sigma\, Z[1], \\ X[2] &= aX[1] + Z[2] \\ X[3] &= aX[2] + Z[3] \\ &\vdots \\ X[d] &= aX[d-1] + Z[d] \end{aligned}$$

where $\sigma = (1-a^2)^{-1/2}$ is chosen so that all the $X[k]$ have the same variance of $\sigma^2 = (1-a^2)^{-1}$. This is a standard auto-regressive process

— the next point in the vector is computed by taking the previous point, multiplying it by a fixed number, then adding an independent perturbation.

The "flow" of this process is very naturally described with this graph: [INSERT]

The covariance matrix for the vector described above is non-zero everywhere:

$$\boldsymbol{R} = \begin{bmatrix} \sigma^2 & a\sigma^2 & a^2\sigma^2 & \cdots & a^{d-1}\sigma^2 \\ a\sigma^2 & \sigma^2 & a\sigma^2 & \cdots & a^{d-2}\sigma^2 \\ \vdots & & \ddots & & \vdots \\ a^{d-1}\sigma^2 & a^{d-2}\sigma^2 & \cdots & & \sigma^2 \end{bmatrix},$$

but the inverse covariance immediately reveals the structure of the equations that generate X:

$$\boldsymbol{S} = \boldsymbol{R}^{-1} = \begin{bmatrix} 1 & -a & 0 & \cdots & & 0 \\ -a & (1+a^2) & -a & \cdots & & 0 \\ 0 & -a & (1+a^2) & -a & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & 0 & -a & (1+a^2) & -a \\ 0 & \cdots & & 0 & -a & 1 \end{bmatrix}$$

Technical Details: The Schur Complement

Suppose that $\boldsymbol{M}$ is an invertible matrix broken into four blocks:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{bmatrix}.$$

If $\boldsymbol{M}_{22}$ is invertible, then the inverse of $\boldsymbol{M}$ can be expressed in terms of these blocks using the *Schur complement* of $\boldsymbol{M}$ in $\boldsymbol{M}_{22}$:

$$\boldsymbol{S} = \boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}.$$

Then,

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \\ -\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1} & \boldsymbol{M}_{22}^{-1} + \boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \end{bmatrix}$$

Similarly, if $\boldsymbol{M}_{11}$ is invertible, we can do something similar with the Schur complement of $\boldsymbol{M}$ in $\boldsymbol{M}_{11}$:

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{M}_{11}^{-1} + \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & -\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & \boldsymbol{S}^{-1} \end{bmatrix},$$

where now

$$\boldsymbol{S} = \boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}.$$

These formulas can be checked simply by multiplying out $\boldsymbol{M}^{-1}\boldsymbol{M}$ and seeing that it is $\mathbf{I}$.