**86-631/42-631 Neural Data Analysis**
**Lecture 4: Review of probability III**

1) Covariance / Correlation
    a. Mean and variance of random vectors
    b. The multivariate normal distribution
    c. Example: Corr=0 does not imply independence
    d. Corr measures linear dependence
2) Expectation and variance of a sum of RVs
3) Conditional expectation
    a. Law of total expectation, law of total variance
4) The Poisson process
5) Signal correlation, noise correlation
6) Bayes Theorem

***7) Slides: Decoding with correlations (Averbeck et al. 2006)***

**Covariance and correlation**

What if two random variables are *not* independent? How do we characterize their dependence? One common method is to use the covariance:

Weighted average

$$Cov[X,Y] \equiv E[(X-\mu_X)(Y-\mu_Y)] = E[XY]-E[X]E[Y]$$

$$Cov[X,Y] \equiv \sum_{x\in X}\sum_{y\in Y}(x-\mu_X)(y-\mu_Y)p(x,y) \qquad Cov[X,Y] \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy$$

Note: Cov[X,Y] = Cov[Y,X]

Symmetric

$Cov[X,X] = Var[X]$

The covariance is analogous to the variance for single random variables. Now, the covariance depends not only on the joint variation of X and Y, but also on their individual variation. To see this, note that Cov[aX,Y]=aCov[X,Y]. This means that the covariance depends on the *scaling* of the random variables. For example, if X were the voltage signal measured off of one electrode, and Y were something else (say, concentration of neurotransmitter measured near a synapse), and if we were to suddenly start measuring X in mV instead of V, then the covariance would change! Why? Because it has units.

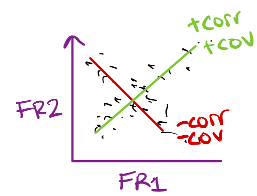
If instead we want a measure of association that is independent of scaling, we use the *correlation*:

rho

$$\rho = Corr[X,Y] \equiv Cov[X,Y]/\sigma_X\sigma_Y.$$

This is also called the *Pearson correlation*.

Now note that if we rescale X to aX, the correlation doesn't change. It can also be shown that the correlation must be between -1 and 1.




***Example:***

| | | X = 0 | X = 1 | X = 2 |
|---|---|---|---|---|
| Y | 2 | .05 | .05 | .1 |
| Y | 1 | .05 | .15 | .1 |
| Y | 0 | .3 | .15 | .05 |

$Corr(X)=0 \neq$ independent

P(X)=[.4,.35,.25]; P(Y)=[.5,.3,.2]

Going back to our spike count example, we can compute the covariance and correlation between X and Y.

$\mu_X$=0*.4+1*.35+2*.25=.85. $\mu_Y$=0*.5+1*.3+2*.2=.7.

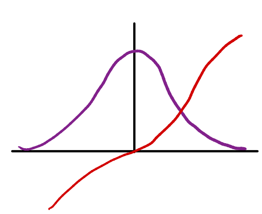
$\sigma_X=\text{sqrt}(.4(0-.85)^2+.35(1-.85)^2+.25(2-.85)^2)\approx.79$ $\sigma_Y=\text{sqrt}((.5(0-.7)^2+.3(1-.7)^2+.2(2-.7)^2)\approx.78$.

Cov[X,Y]≈.26, and $\rho\approx.41$

Although correlation is THE MOST common measure of association between two random variables, it really doesn't measure all kinds of dependence. Corr[X,Y]=0 DOES NOT IMPLY X and Y are independent!

$X \sim N(0,1)$
$Y = X^2$

$$Cov[X,Y] = E[XY] - E[X]E[Y] = E[XY] = E[X^3]$$



***Mean and variance of a random vector.***

The mean of a random vector is a vector. If

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad \text{then } \vec{\mu} = E[\vec{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

Similarly, the variance of a random vector is a matrix:

$$Var[\vec{X}] = \mathbf{\Sigma} = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] & \cdots & Cov[X_1, X_n] \\ Cov[X_2, X_1] & Var[X_2] & & \vdots \\ \vdots & & \ddots & Cov[X_{n-1}, X_n] \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_{n-1}] & Var[X_n] \end{bmatrix}$$

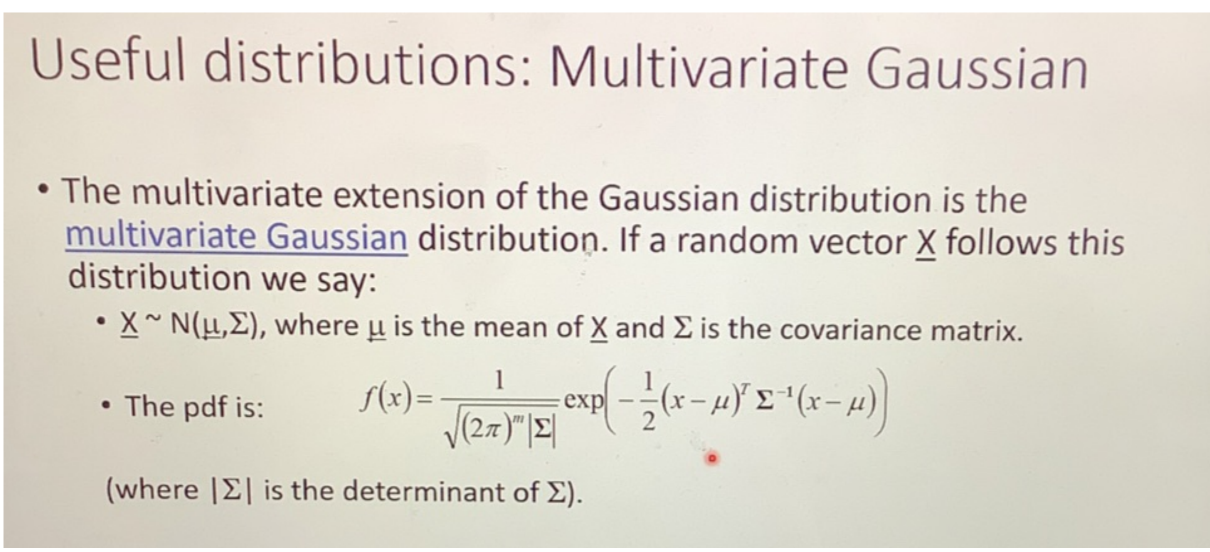
$$= \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \rho_{n-1n}\sigma_{n-1}\sigma_n \\ \rho_{1n}\sigma_1\sigma_n & \cdots & \rho_{n-1n}\sigma_{n-1}\sigma_n & \sigma_n^2 \end{bmatrix}$$

Note that since Corr[X,Y]=Corr[Y,X], $\rho_{ij}=\rho_{ji}$, so $\Sigma$ is a symmetric matrix. It is called the *covariance matrix*.

Let $w$ be an $n$-dimensional vector. It can be shown that

$$E[w^T X] = w^T\mu, \quad \text{and}$$

$$Var[w^T X] = w^T \Sigma w.$$



Normal = Gaussian

***Example: The multivariate normal distribution.***

The multivariate normal distribution. Let X be an m-dimensional multivariate normal having mean vector $\mu$ and covariance matrix $\Sigma$, then its pdf is given by:

$$f(x) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} \exp\left( -\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu) \right)$$

where $|\Sigma|$ is the determinant of $\Sigma$.

$X \sim N(\mu, \Sigma)$

$\mu$ = mean of $\underline{X}$

$\Sigma$ = cov matrix

***Illustration: dependent variables with zero correlation.***

Let X ~ N(0,1). E[X]=0. Also, $E[X^3]=0$. (Show by even odd argument.)

Now let $Y=X^2$. Obviously Y and X are not independent: if we know X=x, then we know $Y=x^2$.

What's the correlation between Y and X?

$$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]. \text{ But } \mu_X=0, \text{ so…}$$

$$Cov(X,Y) = E[X(Y-\mu_Y)] = E[X^3-X\mu_Y] = E[X^3]-E[X]\mu_Y= 0.$$

**Expectation and Variance of sums of rvs**

A very useful and important fact concerning two or more random variables is that their expectation is linear in the sense that the expectation of a linear combination of them is the correspondingly linear combination of their expectations.

For two RVs $X_1$ and $X_2$ we have

holds even when dealing w/ sums of RVs

$$E[aX_1+bX_2]=aE[X_1]+bE[X_2]$$

In general,

$$E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E[X_i]$$

Proof:

$$E[aX_1+bX_2] = \int_{A_2}^{B_2}\int_{A_1}^{B_1}(ax_1+bx_2)f_{12}(x_1,x_2)dx_1dx_2$$

$$E[aX_1+bX_2] = a\int_{A_2}^{B_2}\int_{A_1}^{B_1} x_1 f_{12}(x_1,x_2)dx_1dx_2 + b\int_{A_2}^{B_2}\int_{A_1}^{B_1} x_2 f_{12}(x_1,x_2)dx_1dx_2$$

$$E[aX_1+bX_2] = a\int_{A_1}^{B_1} x_1 \int_{A_2}^{B_2} f_{12}(x_1,x_2)dx_2dx_1 + b\int_{A_2}^{B_2} x_2 \int_{A_1}^{B_1} f_{12}(x_1,x_2)dx_1dx_2$$

$$E[aX_1+bX_2] = a\int_{A_1}^{B_1} x_1 f_1(x_1)dx_1 + b\int_{A_2}^{B_2} x_2 f_2(x_2)dx_2$$

$$E[aX_1+bX_2] = aE[X_1]+bE[X_2]$$

The variance of a sum of random variables is slightly more complicated:

$$Var[aX_1+bX_2]=a^2Var[X_1]+b^2Var[X_2]+2abCov[X_1,X_2].$$

Sketch of Proof: $Var[X_1+X_2] = Var[X_1] + Var[X_2] + 2Cov[X_1,X_2]$

$Var[aX_1 + bX_2] = E[(aX_1+b X_2-E[aX_1+bX_2])^2] = E[(aX_1-E[aX_1]+bX_2-E[bX_2])^2]$.

Let $A= aX_1-E[aX_1]$ and $B= bX_2-E[bX_2]$.   Then $Var[aX_1+bX_2]=E[(A+B)^2] = E[A^2+B^2+2AB]$.

So, $Var[aX_1 + bX_2] = E[A^2]+E[B^2]+2E[AB] = Var[aX_1] + Var[bX_2] + 2Cov[aX_1,bX_2]$.

**Bayes Theorem**

Bayes theorem tells us how to compute a conditional distribution from its opposite partner:

$$P(A|B)=P(B|A)P(A)/P(B).$$

It is fairly easy to prove. Recall $P(B|A)=P(B\cap A)/P(A)$, so $P(B\cap A)=P(B|A)P(A)$. Similarly, $P(A|B)=P(A\cap B)/P(B)$, so $P(A\cap B)=P(A|B)P(B)$. Since $P(A\cap B)=P(B\cap A)$, we have $P(A|B)P(B)=P(B|A)P(A)$, and so we have Bayes' theorem.

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X)P(X)=P(X|Y)P(Y)$$

$$P(Y,X)=P(X,Y)$$

Remember the *law of total probability*?

If $A_1, \ldots, A_n$ represent n mutually exclusive, exhaustive events, then:

$$p(B)=\sum_{i=1}^{n} p(B\cap A_i)=\sum_{i=1}^{n} p(B|A_i)p(A_i)$$

With this in mind, we can write an expanded version of Bayes' Theorem:

$$p(A|B)=\frac{p(B|A)p(A)}{\sum_{i=1}^{n} p(B|A_i)p(A_i)}$$

To get the conditional probability p(A|B), we don't ever have to get the joint probability. We don't even need the marginal p(B). The only two things we need are p(B|A) and p(A).

P(A) is often called the *prior*, and P(A|B) is called the *posterior.* (P(A) is considered the probability of A before you knew B, P(A|B) is the probability of A after you take B into account.)

Although this is a really simple concept, it is tremendously powerful in practice:

(1) It can be used for decoding. Imagine A is a stimulus, and B is the firing rate of a neuron. P(B|A) is the distribution of spikes we would record from a neuron given a particular stimulus – we can figure this out. P(A|B) is the probability that *this* stimulus was responsible for *these* spikes. This is the problem the brain must solve.

(2) There's evidence that we use Bayesian integration when we combine information from multiple sensory modalities.

Posterior

Prior

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X)=\frac{P(X|Y)P(Y)}{\sum_{i=1}^{n} P(X|Y_i)P(Y_i)}$$

***Some non-neural examples of Bayes' rule in practice.***

***Example, tests for rare diseases:***

Let's say we are testing for prostate cancer with a test that's fairly accurate: If you have the disease, you will test positive 90% of the time, if you don't have it, you will test positive 10% of the time. Let's assume that the disease strikes 1 out of 900 men (it doesn't – the actual rate is ~1/1500). Say you test positive for the disease. How worried should you be?

Let D = have disease, and ND = don't have disease. We are interested in P(D|P): the probability that you have the disease, given that you test positive for it. By Bayes:

$$p(D|P) = \frac{p(P|D)p(D)}{p(P|D)p(D) + p(P|ND)p(ND)}$$

$$p(D|P) = \frac{\left(\frac{9}{10}\right)\left(\frac{1}{900}\right)}{\left(\frac{9}{10}\right)\left(\frac{1}{900}\right) + \left(\frac{1}{10}\right)\left(1 - \frac{1}{900}\right)}$$

$$p(D|P) = \frac{\left(\frac{1}{1000}\right)}{\left(\frac{1}{1000}\right) + \left(\frac{899}{9000}\right)} = \frac{\left(\frac{1}{1000}\right)}{\left(\frac{908}{9000}\right)} = \frac{9}{908}$$

***Example: The prosecutor's fallacy.***

Assume that DNA is found at a crime scene that matches with somebody. Suppose that the odds of a match of DNA to a random person are 1 in a million ($10^{-6}$). Clearly, the person did it, right?

Not necessarily. If the subject were really guilty, say we definitely get a match: P(M|G)=1. And, from the "false positive" rate above, P(M|NG) = $10^{-6}$. We are interested in the probability of the subject being guilty, given a match: P(G|M).

From Bayes', we know P(G|M) = P(M|G)P(G) / (P(M|G)P(G)+P(M|NG)P(NG))

$$p(G|M) = \frac{p(M|G)p(G)}{p(M|G)p(G) + p(M|NG)p(NG)}$$

Divide this through by p(G):

$$p(G|M) = \frac{p(M|G)}{p(M|G) + p(M|NG)\frac{p(NG)}{p(G)}}$$

Using p(NG) = 1-p(G), and p(M|G)=1, we get

$$p(G|M) = \frac{1}{1 + 10^{-6}\frac{1 - p(G)}{p(G)}}$$

| P(G) | P(G∣M) |
|---|---|
| $10^{-9}$ | 0.001 |
| $10^{-8}$ | 0.01 |
| $10^{-7}$ | 0.09 |
| $10^{-6}$ | 0.5 |
| $10^{-5}$ | 0.9 |
| $10^{-4}$ | 0.99 |

***If the prior probability of guilt is low, the posterior probability of guilt is still low!***

# Problem # 1

**The Poisson Process**

A random process $\{N(t), t\geq 0\}$ is said to be a *counting process* if $N(t)$ represents the total number of 'events' that have occurred up to time t. Hence, a counting process $N(t)$ must satisfy:

(i) $N(t) \geq 0$
(ii) $N(t) \in$ integers
(iii) If $s<t$, then $N(s) \leq N(t)$
(iv) For $s<t$, $N(t)-N(s)$ equals the number of events that have occurred in the interval $(s,t]$.

A counting process is said to have *independent increments* if the numbers of events that occur in disjoint time intervals are independent. So, $N(t)$ must be independent of $N(t+s)-N(t)$.

A counting process is said to possess *stationary increments* if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval. In other words, $N(t_2+s)-N(t_1+s)$ must have the same distribution for all s.

Perhaps the most important counting process is the *Poisson process*. The counting process $\{N(t), t\geq 0\}$ is said to be a Poisson process having rate $\lambda$, $\lambda>0$, if:

(i) $N(0) = 0$.
(ii) The process has independent increments.
(iii) The number of events in any interval of length *t* is Poisson distributed with mean $\lambda t$. That is, for all $s, t \geq 0$,

$$P\{N(t+s)-N(s)=n\}=e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

Note that it follows from (iii) that the Poisson process has stationary increments and also $E[N(t)]=\lambda t$.

*Note*
The result that $N(t)$ has a Poisson distribution is a consequence of the Poisson approximation to the binomial distribution. Subdivide the interval t into k parts. As k goes to infinity, the probability of getting an event in any given interval goes to zero.

*The interarrival time distribution of a Poisson process follows an exponential distribution*
Let $X_i$ denote the time of the $i^{th}$ event. The sequence of $\{X_n, n\geq 1\}$ is called the sequence of interarrival times.

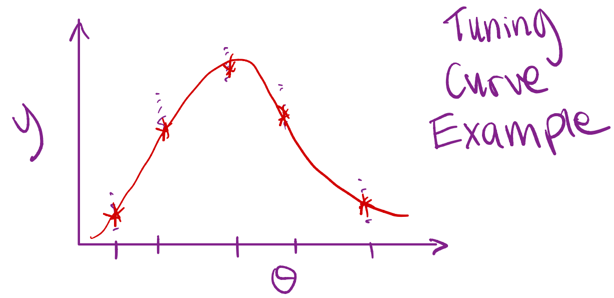
First, note that $X_1>t$ iff no events of the Poisson process occur in $[0,t]$. Thus, $P\{X_1>t\}=e^{-\lambda t}$. Hence, $X_1 \sim Exp(\lambda)$.

To get the distribution of $X_2$, condition on $X_1$. This gives:

$P\{X_2>t|X_1=s\}=P\{0 \text{ events in } (s,s+t]|S_1=s\} = P\{0 \text{ events in } (s,s+t]\}$ (by independent increments)

$=e^{-\lambda t}$ (by stationary increments)

Therefore, $X_n$, n=1,2,… are iid $Exp(\lambda)$ RVs.




**<u>Conditional Expectation</u>**

***Definition of a tuning curve:***

Let's say we measure the firing rate of a neuron when presented with several stimuli, $\theta_1$ through $\theta_n$. The tuning curve is the mean firing rate of the cell as a function of the stimulus. That is, if we can come up with some function that allows us to interpolate the non-sampled points, we write the tuning curve $\lambda(\theta)$ where $\lambda(\theta)$ is really $E[Y;\theta]$. (Draw example on board.)

Now, sometimes the stimulus is also a random variable, X. (From the point of view of the organism, the stimulus is always a random variable that can only be inferred by listening to the spikes!) In this case, each point on the tuning curve is $\lambda(x) = E[Y|X=x]$. This is an example of ***conditional expectation***.

Definition of conditional expectation:

$E[Y|X=x]$.

$$E[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,dy$$

Note that this is a *constant*. x here is *known*. This is perhaps easier to see in the discrete case:

$$E[Y|X=x] = \sum_{y \in Y} y \underbrace{p(y|x)}_{\text{Weighted avg of FR times } p(FR|stim)}$$

Now, because $\lambda$ is a function of x, $\lambda$ is a random variable in its own right! In this case, we say $\lambda(X) = E[Y|X]$. In this case, it becomes really clear why we have to be careful when we write $E[Y|X=x]$ vs $E[Y|X]$.

$\lambda(X) = \mathbb{E}[Y|X]$

*Law of total expectation*

E[E[Y|X]] = E[Y].

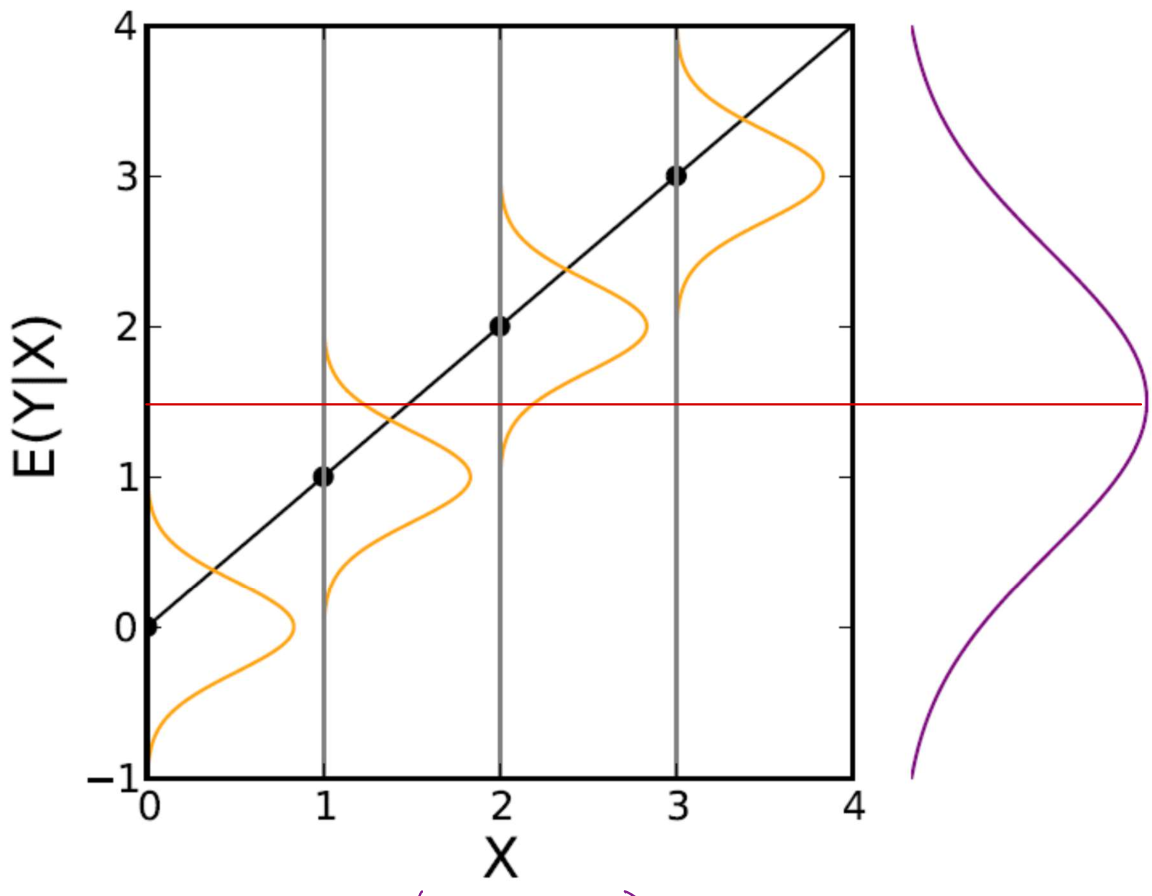
*Proof:*

$$E\left[E\left[Y|X=x\right]\right]=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y f_{Y|X}\left(y|x\right)dy\right)f_X\left(x\right)dx$$

$$E\left[E\left[Y|X=x\right]\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|X}\left(y|x\right)f_X\left(x\right)dxdy$$

$$E\left[E\left[Y|X=x\right]\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}\left(x,y\right)dxdy$$

$$E\left[E\left[Y|X=x\right]\right]=\int_{-\infty}^{\infty} y f_Y\left(y\right)dy=E\left[Y\right]$$




(ex: Contrast)

*Law of total variance*

$\text{Var}[Y] = \text{Var}[E[Y|X]] + E[\text{Var}[Y|X]]$

Total Variance    Signal Variance    Noise Variance

*Proof:*

$\text{Var}[Y]= E[Y^2]-E^2[Y]$.   By the law of total expectation, $E[Y^2]=E[E[Y^2|X]]$, $(E[Y])^2 = (E[E[Y|X]])^2$

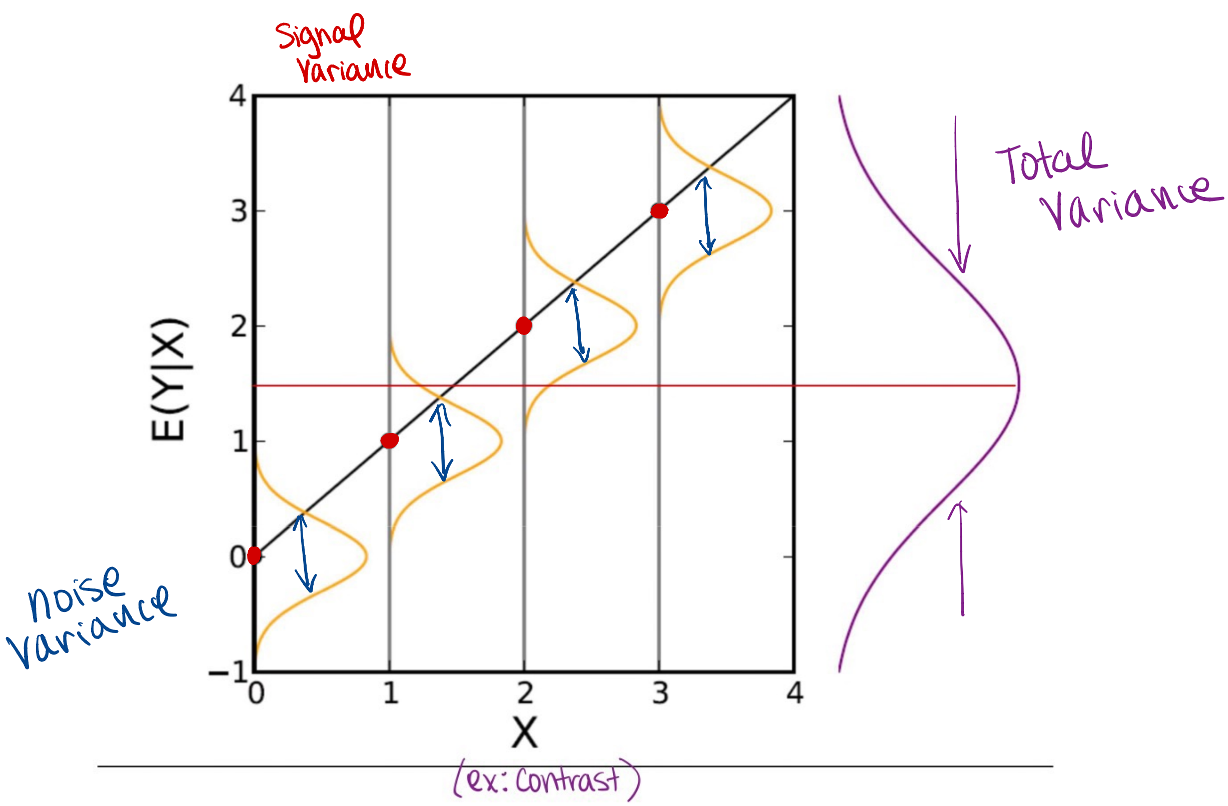
$\text{Var}[Y]=E[E[Y^2|X]]- (E[E[Y|X]])^2$

Now, $E[Y^2|X] = \text{Var}[Y|X] + (E[Y|X])^2$.

So, $\text{Var}[Y] =E[\ \text{Var}[Y|X]+(E[Y|X])^2\ ] - (E[E[Y|X]])^2$

$\text{Var}[Y] = E[\text{Var}[Y|X]] + (E[(E[Y|X])^2]-(E[E[Y|X]])^2) = E[\text{Var}[Y|X]] + \text{Var}[E[Y|X]]$.

Interpret on plot. E[Var]] is the average conditional variance, Var[E] is the variance of the tuning curve changes (the dynamic range).



## <u>Signal correlation and noise correlation</u>

You may have heard or read about signal correlation and noise correlation. *Signal correlation* is the correlation in the tuning curves of two neurons, and *noise correlation* is the correlation between the firing rates of the neurons when you control for the stimulus. Let's be a bit more exact in our definition.

Let the firing rate of neuron 1 be the random variable $Y_1$, and the firing rate of neuron 2 be the random variable $Y_2$. Also, suppose there is a random variable X representing the stimulus.

$E[Y_1|X]$ is the random variable that defines neuron 1's tuning to the stimulus. Similarly,

$E[Y_2|X]$ is the random variable that defines neuron 2's tuning to the stimulus.

Then the *signal covariance* is defined as:

$Cov[E[Y_1|X],E[Y_2|X]]$.

(The signal correlation would be $Cov[E[Y_1|X],E[Y_2|X]] / sqrt(Var[E[Y_1|X]]Var[E[Y_2|X]])$.)

The noise covariance is defined as:

$Cov[Y_1|X=x, Y_2|X=x]$. Note that this is a function of x.

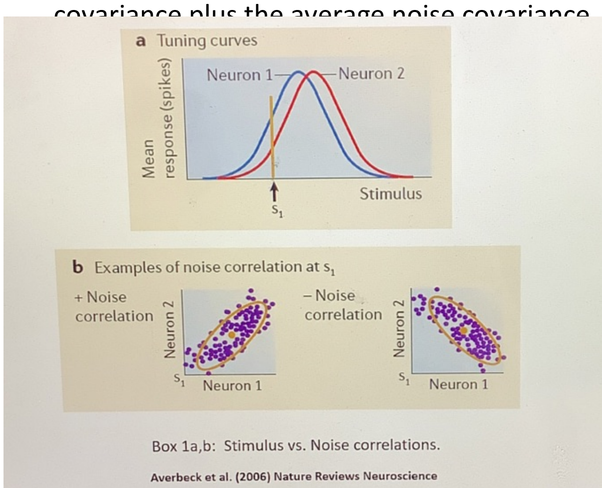
(The noise correlation would be $Cov[Y_1|X=x,Y_2|X=x] / sqrt(Var[Y_1|X=x]Var[Y_2|X=x])$.)

The average noise covariance would be given by $E[Cov[Y_1|X,Y_2|X]]$.

Finally, analogous to the Law of Total Variance is the Law of Total Covariance:

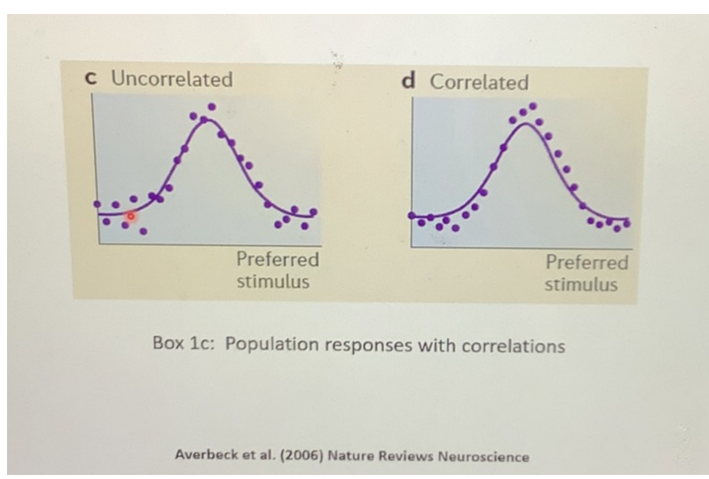
$Cov[Y_1,Y_2] = Cov[E[Y_1|X],E[Y_2|X]] + E[Cov[Y_1|X,Y_2|X]]$.

In words, the covariance of neuron 1's firing rate with neuron 2's firing rate is equal to the signal covariance plus the average noise covariance.



Box 1a,b: Stimulus vs. Noise correlations.

Averbeck et al. (2006) Nature Reviews Neuroscience



Box 1c: Population responses with correlations

Averbeck et al. (2006) Nature Reviews Neuroscience