CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2010
The University of Texas at Austin

Unit 6: Multivariate Normal Distributions and Chi Square

(Let me explain where we're going here…)

- Building up prerequisites to do a fairly sophisticated treatment of model fitting
  - Bayes parameter estimation ✓
  - p-value tail tests ✓
  - really understand multivariate normal and covariance
  - really understand chi-square
- Then, we get to appreciate the actual model fitting stuff
  - fitted parameters
  - their uncertainty expressed in several different ways
  - goodness-of-fit
- And it will in turn be a nice "platform" for learning some other things
  - bootstrap resampling

Multivariate Normal Distributions

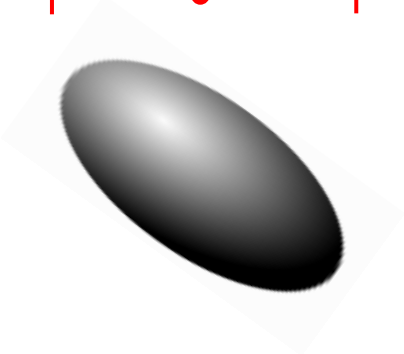
Generalizes Normal (Gaussian) to M-dimensions
Like 1-d Gaussian, completely defined by its mean and (co-)variance
Mean is a M-vector, covariance is a M x M matrix

$$N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

The mean and covariance of r.v.'s from this distribution are*

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \qquad \boldsymbol{\Sigma} = \left\langle (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right\rangle$$

In the one-dimensional case σ is the standard deviation, which can be visualized as "error bars" around the mean.



In more than one dimension Σ can be visualized as an error ellipsoid around the mean in a similar way.

$$1 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

*really?

Because mean and covariance are easy to estimate from a data set, it is easy – perhaps too easy – to fit a multivariate normal distribution to data.

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \approx \frac{1}{N} \sum_i \mathbf{x}_i \qquad \boldsymbol{\Sigma} = \left\langle (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \right\rangle \approx \frac{1}{N} \sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$$

I.e., estimate by sample averages.

But back to "really?" The mean follows from the symmetry argument

$$0 = \int \cdots \int (\mathbf{x} - \boldsymbol{\mu}) \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})] \, d^M \mathbf{x}$$

It's *not* obvious that the covariance in fact obtains from the definition of the multivariate Normal. One has to do the multidimensional (and tensor) integral:

$$\mathbf{M}_2 = \int \cdots \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})] \, d^M \mathbf{x}$$

The only way I know how to do this integral is by trickery involving the Cholesky decomposition ("square root of a positive definite matrix"):

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T \text{ (Cholesky)}, \quad \mathbf{\Sigma}^{-1} = (\mathbf{L}^T)^{-1}\mathbf{L}^{-1}, \quad \mathbf{L}\mathbf{y} \equiv \mathbf{x}$$

we're setting μ to 0 for convenience

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right|$$

Jacobian determinant. The transformation law for multivariate probability distributions.

$$= \frac{\det(\mathbf{L})}{(2\pi)^{N/2}\det(\mathbf{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{y}^T\mathbf{L}^T)(\mathbf{L}^{T-1}\mathbf{L}^{-1})(\mathbf{L}\mathbf{y})]$$

$$= \prod_i (2\pi)^{-1/2} \exp(-\tfrac{1}{2}y_i^2)$$

This is the distribution of N independent univariate Normals N(0,1)!

$$\left\langle \mathbf{x}\mathbf{x}^T \right\rangle = \left\langle \mathbf{L}\mathbf{y}\mathbf{y}^T\mathbf{L}^T \right\rangle = \mathbf{L}\left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{L}^T = \mathbf{L}\mathbf{L}^T = \mathbf{\Sigma}$$

Ha!

(I don't know an elementary proof, i.e., without some matrix decomposition. Can you find one?)

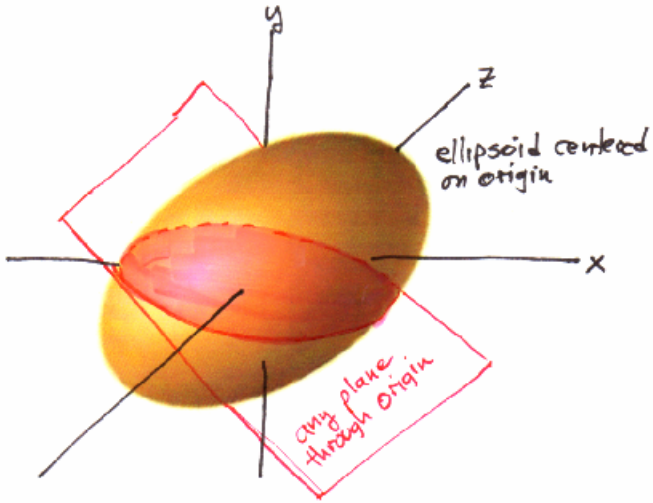
Reduced dimension properties of multivariate normal

$$N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

1. Any slice through a m.v.n. is a m.v.n ("constraint" or "conditioning")

2. Any projection of a m.v.n. is a m.v.n ("marginalization")

You can prove both assertions by "completing the square" in the exponential, producing an exponential in (only) the reduced dimension times an exponential in (only) the lost dimensions. Then the second exponential is either constant (slice case) or can be integrated over (projection case).



How to generate multivariate normal deviates N(μ,Σ):

Cholesky: $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$

Fill $\mathbf{y}$ with independent Normals: $\mathbf{y} = \{y_i\} \sim \mathrm{N}(0,1)$

Transform: $\mathbf{x} = \mathbf{L}\mathbf{y} + \boldsymbol{\mu}$ That's it! x is the desired m.v.n.

Proof: $\left\langle \mathbf{y}\mathbf{y}^T \right\rangle = \mathbf{1}$

$$\begin{aligned} \left\langle (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right\rangle &= \left\langle (\mathbf{L}\mathbf{y})(\mathbf{L}\mathbf{y})^T \right\rangle \\ &= \left\langle \mathbf{L}(\mathbf{y}\mathbf{y}^T)\mathbf{L}^T \right\rangle = \mathbf{L}\left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{L}^T \\ &= \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma} \end{aligned}$$

Even easier: MATLAB has a built-in function mvnrnd(MU,SIGMA). But be sure you get a bunch of m.v.n.'s all in one call, because it (probably) re-does the Cholesky decomposition on each call!

Notice that the proof never used Normality. You can fill **y** with anything with zero mean and variance one, and you'll reproduce Σ. But the result won't be Normal!

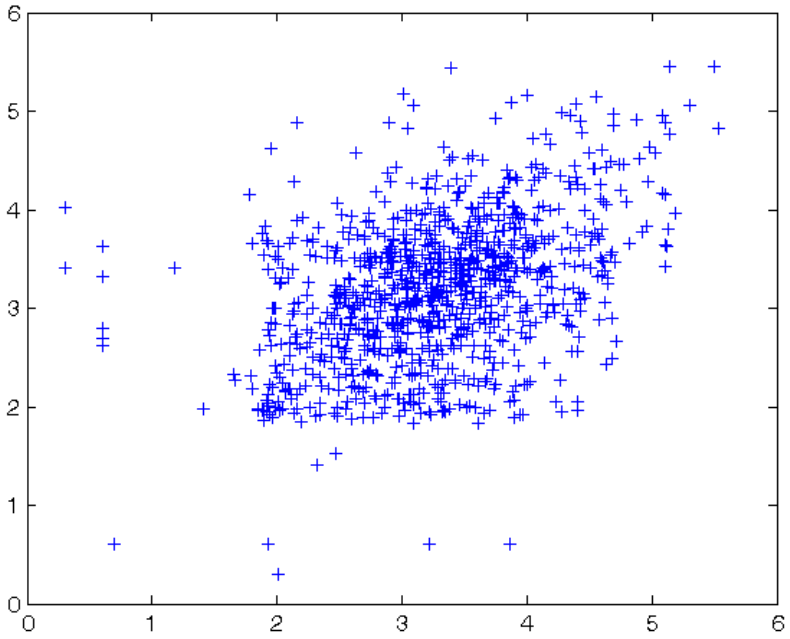
So, easy operations are:

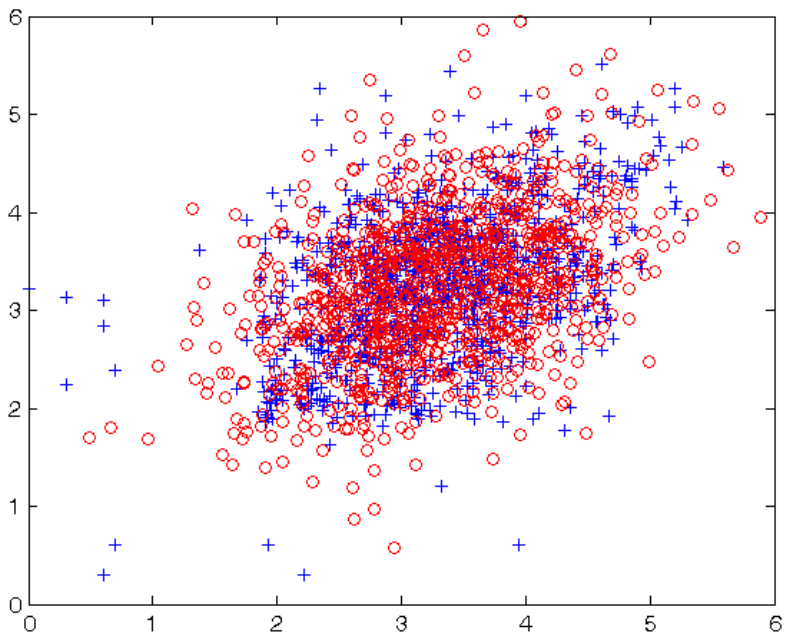
1. Fitting a multivariate normal to a set of points (just compute the sample mean and covariance!)
2. Sampling from the fitted m.v.n.

```
mu = mean([len1 len2])
sig = cov(len1,len2)
mu =
    3.2844
    3.2483
sig =
    0.6125    0.2476
    0.2476    0.5458
rsamp = mvnrnd(mu,sig,1000);
```

In MATLAB, for example, these are one-line operations.

Example:





A related, useful, Cholesky trick is to draw error ellipses (ellipsoids, …)

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$$

So, locus of points at 1 standard deviation is

$$1 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \quad \Rightarrow \quad \left|\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right| = 1$$

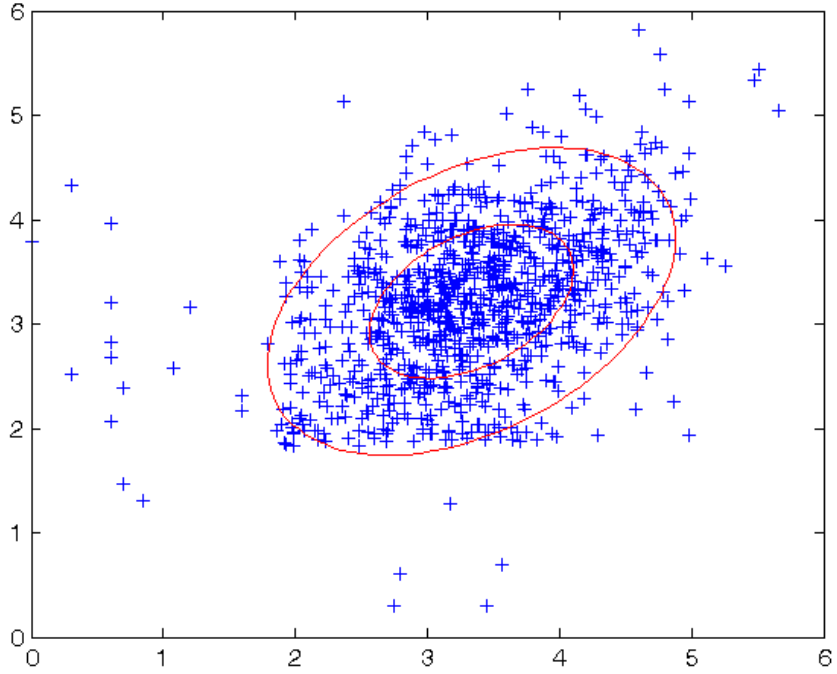
So, if z is on the unit circle (sphere, …) then

$$\mathbf{x} = \mathbf{L}\mathbf{z} + \boldsymbol{\mu}$$

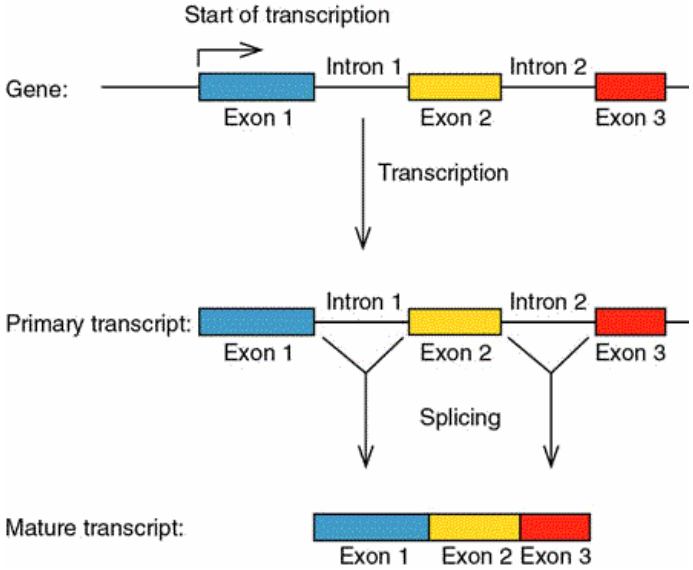
will be on the error ellipse.

my coding of this idea looks like this

```
function [x y] = errorellipse(mu,sigma,stdev,n)
L = chol(sigma,'lower');
circle =
 [cos(2*pi*(0:n)/n); sin(2*pi*(0:n)/n)].*stdev;
ellipse = L*circle + repmat(mu,[1,n+1]);
x = ellipse(1,:);
y = ellipse(2,:);
```



The distribution we have been looking at has some interesting biology in it!



file genestats.dat (on course web site) contains 20694 lines like this:

number of exons N

N exon lengths

N-1 intron lengths

```
ENST00000341866    17470     3262 0.00002 4 1290 349 1412 211 169 678 13361 <EOL>
ENST00000314348    22078     1834 0.00001 7 100 166 113 178 165 262 850 5475
385 3273 1149 2070 7892
ENST00000313081    13858     1160 0.00001 6 496 150 107 85 151 171 2068 76 2063
674 7817
ENST00000298622    80000     6487 0.00001 24 135 498 216 120 147 132 36 60 129
129 84 63 99 99 54 66 69 78 204 66 73 1081 397 2452 12133 15737 1513 769 942
103 829 2272 1340 3058 327 2371 1361 471 2922 735 85 9218 1257 2247 897 822
12104
```

gene name

total length

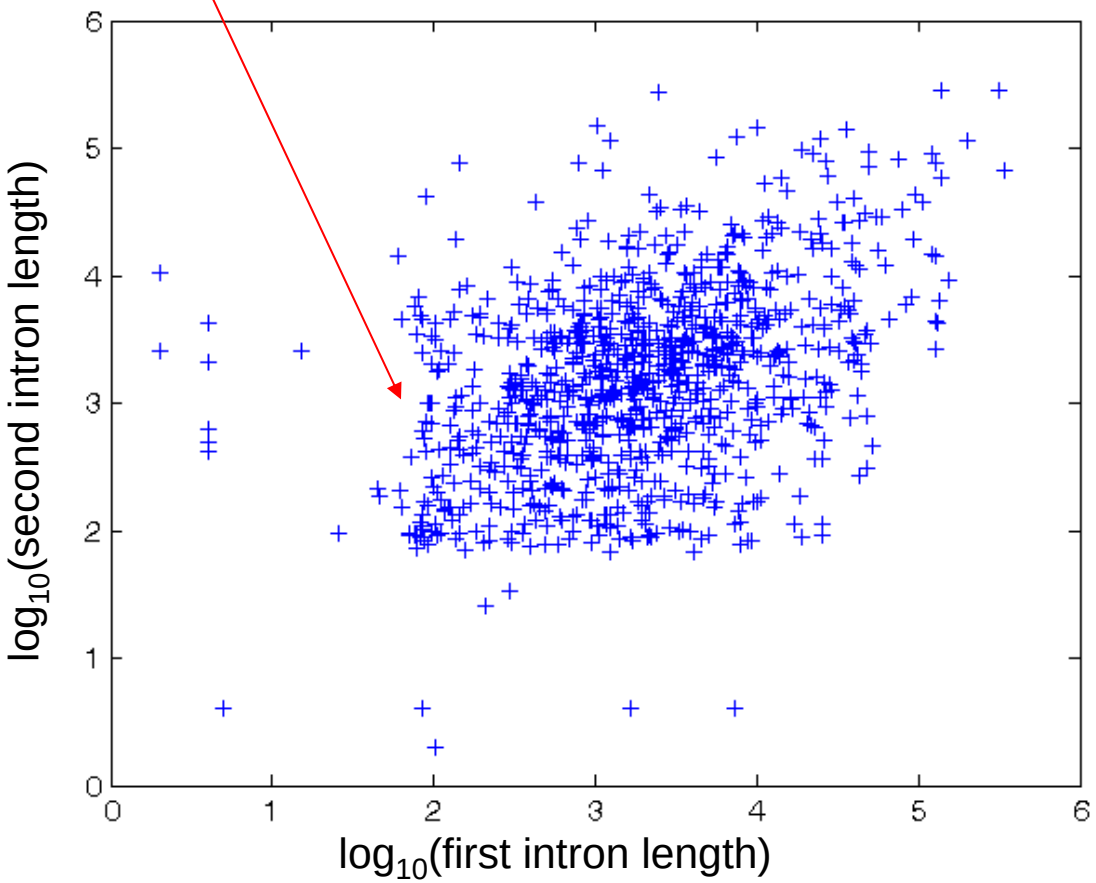
total length of exons

ignore for now

Log_{10} of size of 1st and 2nd introns for 1000 genes:

This is kind of fun, because it's not just the usual featureless scatter plot

notice the "hard edges"
this is biology!



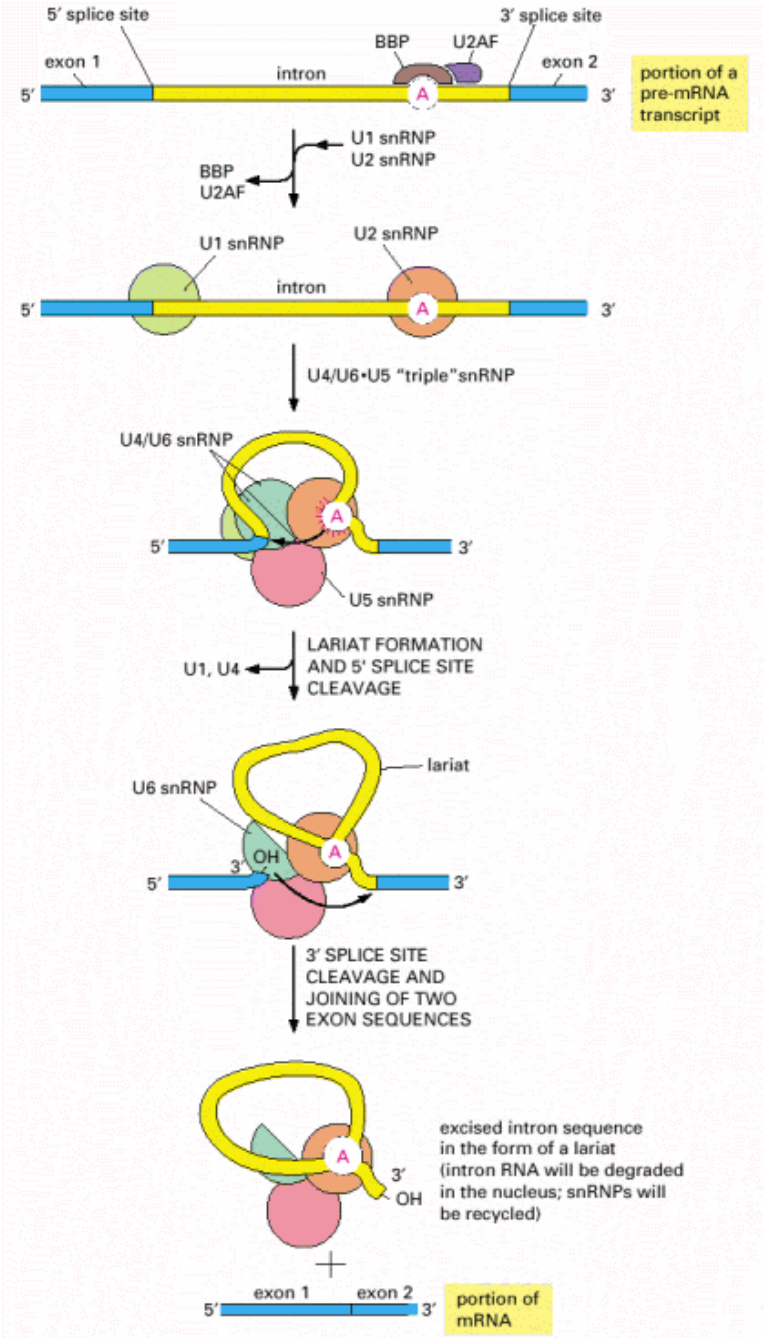
Is there a significant correlation here? If the first intron is long, does the second one also tend to be? Or is our eye being fooled by the non-Gaussian shape?

Biology:

The hard lower bounds on intron length are because the intron has to fit around the “big” spliceosome machinery!

It’s all carefully arranged to allow exons of any length, even quite small.

Why? Could the spliceosome have evolved to require a minimum exon length, too? Are we seeing chance early history, or selection?



credit: Alberts et al.
Molecular Biology of the Cell

The covariance matrix is a more general idea than just for multivariate Normal.
You can compute the covariances of any set of random variables.
It's the generalizaton to M-dimensions of the (centered) second moment Var.

$$\mathrm{Cov}\,(x,y) = \langle (x-\overline{x})(y-\overline{y}) \rangle$$

For multiple r.v.'s, all the possible covariances form a (symmetric) matrix:

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\,(x_i,x_j) = \langle (x_i-\overline{x_i})(x_j-\overline{x_j}) \rangle$$

Notice that the diagonal elements are the variances of the individual variables.

The variance of any linear combination of r.v.'s is a quadratic form in $\mathbf{C}$:

$$\begin{aligned}
\mathrm{Var}\,\left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i-\overline{x_i}) \sum_j \alpha_j (x_j-\overline{x_j}) \right\rangle \\
&= \sum_{ij} \alpha_i \left\langle (x_i-\overline{x_i})(x_j-\overline{x_j}) \right\rangle \alpha_j \\
&= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha}
\end{aligned}$$

This also shows that $\mathbf{C}$ is positive definite, so it can still be visualized as an ellipsoid in the space of the r.v.'s., where the directions are the different linear combinations.

The covariance matrix is closely related to the linear correlation matrix.

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

more often seen written out as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

When the null hypothesis is that X and Y are independent r.v.'s, then r is useful as a p-value statistic ("test for correlation"), because

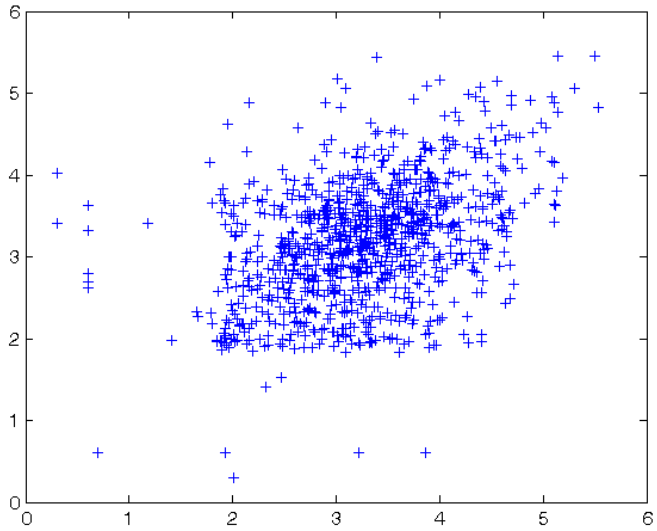
1. For large numbers of data points *N*, it is normally distributed,

   $r \sim \mathrm{N}(0, N^{-1/2})$

   so $r\sqrt{N}$ is a normal t-value

2. Even with small numbers of data points, if the underlying distribution is multivariate normal, there is a simple form for the p-value (comes from a Student t distribution).

For the exon length data, we can easily now show that the correlation is highly significant.




```
r = sig ./ sqrt(diag(sig) * diag(sig)')
tval = sqrt(numel(len1))*r
r =
    1.0000    0.3843
    0.3843    1.0000
tval =
   31.6228   12.1511
   12.1511   31.6228
```

statistical significance of the correlation in standard deviations (but note: uses CLT)

```
[rr p] = corrcoef(i1llen,i2llen)
rr =
    1.0000    0.3843
    0.3843    1.0000
p =
    1.0000    0.0000
    0.0000    1.0000
```

Matlab has built-ins

not clear why Matlab reports 1 on the diagonals. I'd call it 0!

Let's talk more about **chi-square**.
Recall that a t-value is (by definition) a deviate from $\mathrm{N}(0,1)$

χ^2 is a "statistic" defined as the sum of the squares of n independent t-values.

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

$\mathrm{Chisquare}(\nu)$ is a distribution (special case of Gamma), defined as

$$\chi^2 \sim \mathrm{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)} (\chi^2)^{\frac{1}{2}\nu - 1} \exp\left(-\tfrac{1}{2}\chi^2\right) d\chi^2, \qquad \chi^2 > 0$$

The important theorem is that $\chi 2$ is in fact distributed as Chisquare.

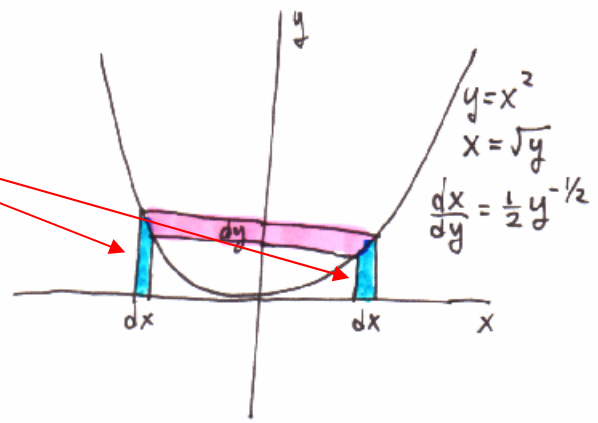
Let's prove it.

Prove first the case of ν=1:

Suppose $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \Rightarrow \quad x \sim \mathrm{N}(0,1)$

and $y = x^2$

$p_Y(y)\,dy = 2p_X(x)\,dx$



$$\text{So, } p_Y(y) = y^{-1/2} p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$$

$$\sim \mathrm{Chisquare}(1)$$

To prove the general case for integer ν, compute the characteristic function

$$\chi^2 \sim \text{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\tfrac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

```
In[9]:= pchi2 = (1 / (2^(nu / 2) Gamma[nu / 2])) y^(nu / 2 - 1) Exp[-y / 2]
```

Out[9]= $\dfrac{2^{-\text{nu}/2}\, e^{-y/2}\, y^{-1+\frac{\text{nu}}{2}}}{\text{Gamma}\left[\frac{\text{nu}}{2}\right]}$

```
In[10]:= Integrate[pchi2, {y, 0, Infinity}, GenerateConditions → False]
```

```
Out[10]=
1
```

```
In[11]:= Integrate[pchi2 Exp[I t y], {y, 0, Infinity},
 GenerateConditions → False]
```

```
Out[11]=
(1 - 2 i t)^(-nu/2)
```

Since we already proved that ν=1 is the distribution of a single t^2-value, this proves that the general ν case is the sum of ν t^2-values.

Question: What is the generalization of

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

to the case where the x_i's are normal, **but not independent**?
I.e., **x** comes from a multivariate Normal distribution?

Answer:

$$\chi^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \qquad \mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Proof is one of those Cholesky things,

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T, \quad \mathbf{L}\mathbf{y} = \mathbf{x} - \boldsymbol{\mu},$$

show that $\mathbf{y}$ is product of independent N(0,1)'s, as we did before, and that

$$\chi^2 = \mathbf{y}^T \mathbf{y} = \sum y_i^2$$