Part 1: Probability

Suppose that a random player's chance of winning the lottery is 1 in a million.
But you go to a psychic, whose predictions are 99% accurate. Really, 99% accurate.

a. The psychic tells you that you will win this week! What is the probability you will win?

b. The psychic tells you that you will lose this week. What is the probability you will win?

Hint: it may help to consider 100 million people going to psychics and playing the lottery, and tally them as integers in a contingency table (joint distribution).

c. Elucidate how your answer is an application of Bayes' theorem
http://en.wikipedia.org/wiki/Bayes%27_theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

---------------- background --------------
http://yudkowsky.net/rational/bayes
Your friends and colleagues are talking about something called "Bayes' Theorem" or "Bayes' Rule", or something called Bayesian reasoning. They sound really enthusiastic about it, too, so you google and find a webpage about Bayes' Theorem and...It's this equation. That's all. Just one equation. The page you found gives a definition of it, but it doesn't say what it is, or why it's useful, or why your friends would be interested in it. It looks like this random statistics thing.

So you came here. Maybe you don't understand what the equation says. Maybe you understand it in theory, but every time you try to apply it in practice you get mixed up trying to remember the difference between $p(a|x)$ and $p(x|a)$, and whether $p(a)*p(x|a)$ belongs in the numerator or the denominator. Maybe you see the theorem, and you understand the theorem, and you can use the theorem, but you can't understand why your friends and/or research colleagues seem to think it's the secret of the universe. Maybe your friends are all wearing Bayes' Theorem T-shirts, and you're feeling left out. Maybe you're a girl looking for a boyfriend, but the boy you're interested in refuses to date anyone who "isn't Bayesian". What matters is that Bayes is cool, and if you don't know Bayes, you aren't cool.

Why does a mathematical concept generate this strange enthusiasm in its students? What is the so-called Bayesian Revolution now sweeping through the sciences, which claims to subsume even the experimental method itself as a special case? What is the secret that the adherents of Bayes know? What is the light that they have seen?
Soon you will know. Soon you will be one of us.

Part II: Mean, deviations, mean square, mean squared deviation (variance), and RMS (standard deviation).

DEFINITIONS
Let an overbar indicate averaging over a discrete set y of data values y = $\{y_i\}$. Let y' indicate the same set of data values with the mean (which is a single, constant number) subtracted from each value.

$$y = \{y_i\}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$y' = \{(y_i - \bar{y})\} = \{y'_i\}$$

The variance var(y), and its square root the standard deviation or RMS, are:

$$\mathrm{var}(y) = \overline{y'^2} = \frac{1}{N}\sum_{i=1}^{N} (y'_i)^2$$

$$\sigma_y = sd(y) = RMS(y) = \mathrm{var}^{1/2}(y)$$

Show using the summation notation that $\overline{y'} = 0$
Show using the above that $\overline{y^2} = \bar{y}^2 + \overline{y'^2}$.

Write Matlab code using the above fact to calculate both the mean and standard deviation using only a single Matlab loop over all $y_{i.}$
Hint: define variables sum_of_y and sum_of_y2 (pronounce it “sum of y-squared”), and update them inside the loop. Then, after the loop, compute what you want.

```
for i=1:N




end
```

Now suppose we have 2-dimensional data arrays. It is clearest to bring in function-like notation f(x,t) rather than the indexed set $\{f_{ij}\}$ and summation notation used above to define the bar (averaging) and prime (deviation). For example, our homework example data of equatorial SST could be called SST(x,t).

ASIDE: Since every longitude grid cell is equal size, the area average on Earth is simple average over the x (longitude) values. But if one of the dimensions is latitude y, we have to weight all averages by the area of lat-lon grid cells, which is proportional to cos(lat), to get a true geographical average.

There are (at least) 3 kinds of average we can take:

$$\overline{\overline{SST}}^{x,t} = \frac{1}{NT \bullet NX}\sum_{i=1}^{NX}\sum_{j=1}^{NT} SST_{ij}$$

$$\overline{SST}^{x}(t) = \frac{1}{NX}\sum_{i=1}^{NX} SST_{ij}$$

$$\overline{SST}^{t}(x) = \frac{1}{NT}\sum_{j=1}^{NT} SST_{ij}$$

Notice how the function notation (x) and (t) helps keep the size or dimension of each of these mathematical objects clear. Note that the order of averaging doesn't matter: summation operators and multiplication by 1/NT vs. 1/NX commute.

Corresponding to the various "bar" averages, there are at least 3 kinds of primes or deviations or "perturbations" or "anomalies."

$$SST'^{\,x,t}(x,t) = SST(x,t) - \overline{\overline{SST}}^{x,t}$$

$$SST'^{\,x}(x,t) = SST(x,t) - \overline{SST}^{x}(t)$$

$$SST'^{\,t}(x,t) = SST(x,t) - \overline{SST}^{t}(x)$$

Simply using a prime for all 3 leads to all kinds of confusion, especially since all 3 kinds of deviation are mathematical objects (or computer arrays) of the same dimensions. There are spatial deviations (from the zonal mean), called "zonal eddy" SST; temporal deviations (from the time mean), called SST "transients"; and deviations from the grand mean (which might just be called perturbations or anomalies). You may be learning these terms in General Circulation class where at least the convention is to use [] and * for zonal average & deviations; and bar & prime for temporal average & deviations.

YOU at least need to know at all times what kind averages you have removed from whatever you are showing. For communicating with others, generous frequent use of the words "temporal", "spatial", "zonal" (or east-west), "meridional" (or latitudinal or north-south), "azimuthal", etc. can hardly be overdone. Later, when we further decompose variability into frequency bands, like a climatological annual cycle and deviations from that mean cycle, you should frequently use the words "year to year differences" or "deviations from monthly climatology," since that quantity includes high frequencies like month to month weather fluctuations, as well as truly interannual-timescale anomalies like El Nino. A reader who is following your statements carefully can flow with these technical words easily enough, while any reader who is confused or has forgotten is frequently reminded and kept on track.

What are the properties of these 3 kinds of primes (deviations)? Well, each one vanishes when subjected to its corresponding kind of bar (average). For example, it is easy to see that

$$\overline{\overline{SST^{\overset{x,t}{\prime}}}}^{x,t} \equiv \overline{\overline{(SST - \overline{\overline{SST}}^{x,t})}}^{x,t} = \overline{\overline{SST}}^{x,t} - \overline{\overline{\overline{\overline{SST}}^{x,t}}}^{x,t} = 0$$

because the second term is just an average over x,t of a constant, which is equal to that constant. In the same way,

$$\overline{SST^{\overset{x}{\prime}}}^{x}(t) \equiv \overline{(SST - \overline{SST}^{x})}^{x} = \overline{SST}^{x}(t) - \overline{\overline{SST}^{x}(t)}^{x} = 0$$

because the second term is the average over x of something that's not a function of x. Similarly the t-average of temporal perturbations is zero. But the *time* average of *spatial* perturbations is not zero in general:

$$\overline{SST^{\overset{x}{\prime}}}^{t}(x) = \overline{(SST - \overline{SST}^{x})}^{t} = \overline{SST}^{t}(x) - \overline{\overline{SST}^{x}(t)}^{t} \neq 0$$

What is the relationship among all these different kinds of perturbations and averages (temporal, spatial, and temporo-spatial)? Well, the total SST(x,t) can be decomposed into them in several ways, rearranging the above:

$$SST(x,t) = \overline{\overline{SST}}^{x,t} + SST^{\overset{x,t}{\prime}}(x,t)$$

$$SST(x,t) = \overline{SST}^{x}(t) + SST^{\overset{x}{\prime}}(x,t) = \overline{SST}^{x}(t) + \left(\overline{SST^{\overset{x}{\prime}}}^{t}(x) + SST^{\overset{x\,t}{\prime\,\prime}}(x,t) \right)$$

$$SST(x,t) = \overline{SST}^{t}(x) + SST^{\overset{t}{\prime}}(x,t) = \overline{SST}^{t}(x) + \left(\overline{SST^{\overset{t}{\prime}}}^{x}(t) + SST^{\overset{t\,x}{\prime\,\prime}}(x,t) \right)$$

These are all "orthogonal" decompositions: the terms on the right are orthogonal (the product of any two of them, integrated over all x&t, is 0). For this reason, total variance (or total squared SST) can be partitoned cleanly into the terms in the right, with each acting as a truly separate variance "bin" or category. It is helpful to measure all the terms on the right of all these decompositions by their summed-square or RMS.

The final terms are equal in the second and third equations above, but not the first:

$$SST^{\overset{x\,t}{\prime\,\prime}}(x,t) = SST^{\overset{t\,x}{\prime\,\prime}}(x,t) \neq SST^{\overset{x,t}{\prime}}(x,t)$$

To see all this, let's look at some data in a domain with two very different dimensions: atmospheric profiles of temperature T(p,t). Let's remove the time-mean pressure-mean value $\overline{\overline{T}}^{p,t}$ so we don't have the square of 273K cluttering up our budget numbers of squared temperature: In other words, define

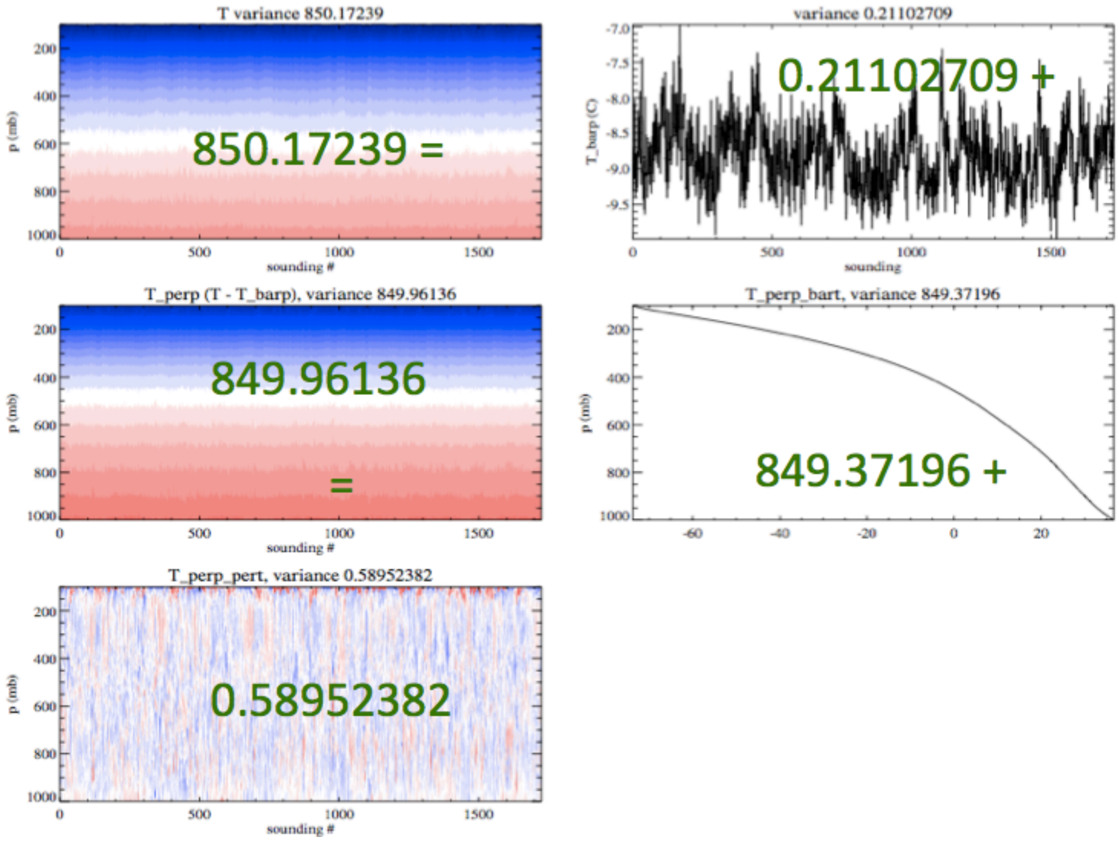
$$T(p,t) = T_{abs}{}^{\overset{p,t}{|}}(p,t)$$

A time series and a profile are very different: T(p) has very large variance, hundreds of square degrees, while a time series of temperature T(t) in the tropics fluctuates little, with standard deviations less than a degree. Nonetheless, the total variance can be budgeted into the various terms to full accuracy, so we can see the different contributions in different parts (the hundreds and the tenths of square degrees) of the total variance budget.

Decomposition 1: starting from vertical average

$$T(t,p) = \overline{T}^{p}(t) + T^{\overset{p}{|}}(t,p) = \overline{T}^{p}(t) + T^{\overline{\overset{p}{|}}^{t}}(p) + T^{\overset{p}{|}\,\overset{t}{|}}(t,p)$$

var: 850.17=0.21 + 849.96 =0.21 +849.37 + 0.59 ←numbers aligned w/ terms


T variance 850.17239

variance 0.21102709

T_perp (T - T_barp), variance 849.96136

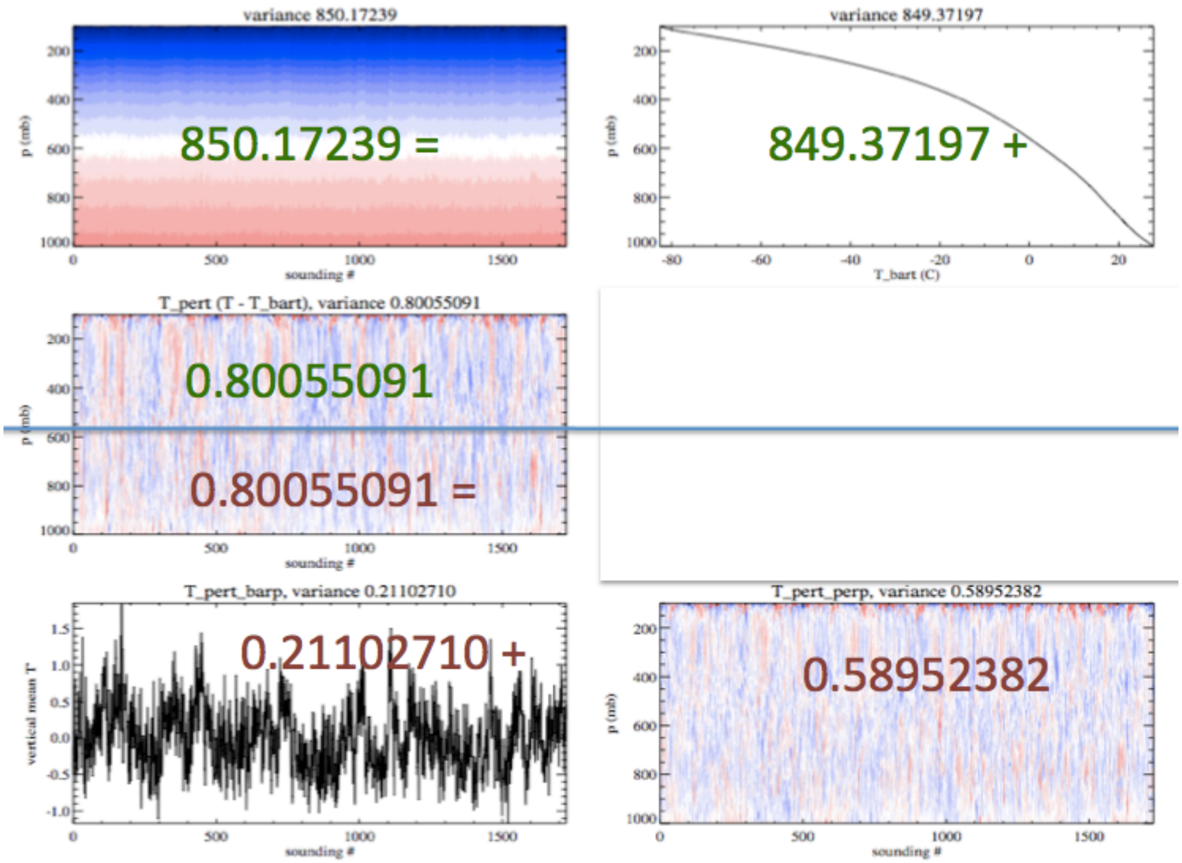
T_perp_bart, variance 849.37196

T_perp_pert, variance 0.58952382

Decomposition 2: starting from a time average

$$T(t,p) = \overline{T}^{t}(p) + T^{\prime t}(t,p) = \overline{T}^{t}(p) + \overline{T^{\prime t}}^{p}(t) + T^{\prime t \prime p}(t,p)$$

var: 850.17=849.37 +0.8 =849.37 +0.21 + 0.59 ←numbers aligned w/ terms



Note that the 0.21 values in the two decompositions are not identical, but the 0.59 values are. Show that panels = 0!

The total variance adds up this way to the total because the terms on the right are orthogonal.

Square both sides. All the cross terms vanish if the terms are orthogonal. Averages and deviations from an average are orthogonal. So are Fourier components in a Fourier decomposition using an integer number of periods of the sine and cosine components). Parseval's theorem tells us the total variance is the sum over Fourier components, and is equally true for any orthogonal decomposition.

Q: What is the variance of a trend?
Q: What is the variance of a sine or cosine curve over its whole period?
Q: What is the variance budget of the Black Rock Forest temperature series? You have the *amplitude* of trend, annual, semiannual, diurnal, etc. so you can answer: How much does each component contribute to the total variance?

Part III: Covariance, correlation coefficient, and regression coefficient.

A powerful result in the book is the matrix result for linear regression, but it helps to see it in a close example. But first let’s review univariate regression.

We try to “fit” data to a straight line, with equation y = mx + b, where m is slope and b is intercept. For clarity here, let’s remove the mean of both variables so that we just have y’ = mx’ + error. Fitting means minimizing the summed square of error or residual variance.

The covariance cov(x,y) has units of x times y, and is given by:

$$\mathrm{cov}(x,y) = \mathrm{cov}(y,x) = \overline{x'y'} = \overline{y'x'}$$

Notice that the covariance **matrix** C_{xy} as used in the book is a peculiar math object: it is a bundle of numbers with different units! So minimizing it, or explaining it in a budget sense, is a subtle idea that is hard to think straight about.

From cov(x,y) we can get the dimensionless correlation coefficient:

$$cor(x,y) = r_{xy} = \frac{\mathrm{cov}(x,y)}{sd(x)sd(y)} = \frac{\overline{x'y'}}{\sigma_x \sigma_y}$$

Note that both of the above quantites are symmetric in x and y: cor(x,y) = cor(y,x).

The optimal value of m in our linear fit y’=mx’ is the regression slope or regression coefficient reg(y,x):

$$m_{best} = reg(y,x) = r_{xy}\frac{\sigma_y}{\sigma_x} = \overline{x'y'}\Big/\sigma_x^2$$

Show that the $m_{best}x'$ term accounts for a fraction r_{xy}^2 of the variance of y.

Note that reg(y,x) is NOT symmetric in x and y: it has units of y divided by x. In fact this is how I remember its form: it is proportional to correlation of course, and has the right units from the only quantity we have: the standard deviation. This would be called “regressing y on x” (fitting a line y’=mx’). It is not the same as regressing x on y (fitting a line x’=wy’), because you are minimizing different squared errors or residuals, with different units, in the two cases. Thus w is not equal to 1/m!

If you standardize all variables at the outset, $y_{std} = (y-\bar{y})\Big/\sigma_{y,}$ $x_{std} = (x-\bar{x})\Big/\sigma_{x,}$

then w = 1/m, the regression slope (or regression coefficient) is just the correlation coefficient, and regression is symmetric just like correlation. But that assumes all percentage fluctuations are equally important, in all quantities with all units. When dealing with relationships among various quantities with different units, it is all we can do sometimes. But this is a consequential choice and needs to be done consciously: are you partitioning the *covariance matrix* or the *correlation matrix*?

Now on to *multiple* linear regression: it is simple, and analogous, but an example helps a lot to see the power of the method. Consider Lin and Mapes (2004 MWR).

Suppose we have two kinds of rainfall: Convective and Stratiform. Let the two rainrate time series (with the means removed) be C'(t) and S'(t). Let's postulate that the two have different vertical profiles of heating, vertical velocity, and thus wind divergence D, which we can call Ds(p) and Dc(p). We want to estimate these from data.

A radar can measure (or, strictly, we should say 'estimate') C'(t) and S'(t) as defined by the horizontal and vertical texture of radar echoes. If it is a Doppler radar, it can measure (or we can estimate from its measurements) the total divergence profile with time D(t,p) through the Velocity Azimuth Display technique.

How can we solve for estimates of Ds(p) and Dc(p)? Set it up as a matrix problem.

We define our take on the problem by postulating a decomposition of the data into what we desire, plus the remainder:

$$D(t,p) = C'(t)Dc(p) + S'(t)Ds(p) + resid(t,p)$$

Now we want to estimate Dc(p) and Ds(P) such that the residual is minimized. Specifically, we want to minimize the summed square or RMS of the residual. This is like a **d = Gm** example from the book. With **d** the data, **m** the parameters we want to estimate, and **G** the "data kernel" matrix. Here time (0,1,2,…Nt) runs down, and pressure (surface, midlevels,... Np) runs from left to right:

$$\begin{bmatrix} D(0,sfc) & D(0,mid)\ldots & D(0,Np) \\ D(1,sfc) & D(1,mid)\ldots & D(1,Np) \\ \ldots & \ldots & \ldots \\ D(Nt,sfc) & D(Nt,mid)\ldots & D(Nt,Np) \end{bmatrix} = \begin{bmatrix} C'(0) & S'(0) \\ C'(1) & S'(1) \\ \ldots & \ldots \\ C'(Nt) & S'(Nt) \end{bmatrix} \begin{bmatrix} Dc(sfc) & Dc(mid)\ldots & Dc(Np) \\ Ds(sfc) & Ds(mid)\ldots & Ds(Np) \end{bmatrix} + resic$$

d = **G** **m**

To solve the best estimate of **m** (minimizing the RMS of the residual), we multiply both sides by the transpose $\mathbf{G}^{\mathbf{T}}$, and then "divide" both sides by $\mathbf{G}^{\mathbf{T}}\mathbf{G}$ to get m. Of course for a matrix, division is really left-multiplication by $(\mathbf{G}^{\mathbf{T}}\mathbf{G})^{-1}$ so we have:

$$\mathbf{m}_{best} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d} = \begin{bmatrix} \overline{C'C'} & \overline{C'S'} \\ \overline{C'S'} & \overline{S'S'} \end{bmatrix}^{-1} \mathbf{G}^T\mathbf{d} = K \begin{bmatrix} \overline{S'S'} & -\overline{C'S'} \\ -\overline{C'S'} & \overline{C'C'} \end{bmatrix} \mathbf{G}^T\mathbf{d}$$

For 2x2 we have done the math of inversion manually, with K=1/det($\mathbf{G}^{\mathbf{T}}\mathbf{G}$). Finally **m**, which is just the bundle of the "pure" convective and stratiform divergence profiles we seek, can be expressed in terms of covariances and standard deviations:

$$\mathrm{m}_{best} = K\begin{bmatrix} \sigma_S^2 & -\overline{C'S} \\ -\overline{C'S} & \sigma_C^2 \end{bmatrix}\begin{bmatrix} \mathrm{cov}(D_{sfc}, C) & \mathrm{cov}(D_{mid}, C)\ldots & \mathrm{cov}(D_{Np}, C) \\ \mathrm{cov}(D_{sfc}, S) & \mathrm{cov}(D_{mid}, S)\ldots & \mathrm{cov}(D_{Np}, S) \end{bmatrix}$$

Or, unbundling it, our best estimates are the scalar equations (not matrices):

$$\mathrm{Dc(sfc)} = \mathrm{K}(\ \mathrm{cov}(\mathrm{D_{sfc}},\mathrm{C})\ \sigma_S^2\ - \mathrm{cov}(\mathrm{D_{sfc}},\mathrm{S})\mathrm{cov}(\mathrm{S},\mathrm{C})\ )$$
$$\mathrm{Ds(sfc)} = \mathrm{K}(\ \mathrm{cov}(\mathrm{D_{sfc}},\mathrm{S})\ \sigma_c^2\ - \mathrm{cov}(\mathrm{D_{sfc}},\mathrm{C})\mathrm{cov}(\mathrm{S},\mathrm{C})\ )$$

And similarly for other pressure levels above the surface.

I find this a little inscrutable in terms of all the weighting factors – why is it σ_S^2 in the first term in the D*c* equation? It came from the matrix inverse operation $(\mathbf{G}^T\mathbf{G})^{-1}$ which is just a magic formula to me at this moment. But I can see the key dependences at least, with an understandable SIGN:

Dc(sfc) = a term α cor(D_{sfc},C)   MINUS   a term α cor(D_{sfc},S) x cor(S,C)

If all the time series involved (C, S, and D at every altitude) are standardized, then K=1 and there are no units and the above is literally just the part I understand:

$$\mathrm{Dc_{std}(sfc)} = \mathrm{cor}(\mathrm{D_{sfc}},\mathrm{C}) - \mathrm{cor}(\mathrm{D_{sfc}},\mathrm{S})\mathrm{cor}(\mathrm{S},\mathrm{C})$$
$$\mathrm{Ds_{std}(sfc)} = \mathrm{cor}(\mathrm{D_{sfc}},\mathrm{S}) - \mathrm{cor}(\mathrm{D_{sfc}},\mathrm{C})\mathrm{cor}(\mathrm{S},\mathrm{C})$$

This equation makes clear that the correlation of the data with each of our predictors (the first term) is simply adjusted (in the second term) for the partial redundancy ("multi-collinearity") of the predictors. In this example, convective and stratiform rain tend to fluctuate together, so we don't want to assign all the D fluctuations that go with convective rain fluctuations purely to those C fluctuations. Some are attributable to the S fluctuations, which happen to be partially correlated with convection. But as long as C and S are independent – that is, as long as they aren't totally redundant (in which case det→0 so that K →1/0) – we can tease apart the two entangled signals.

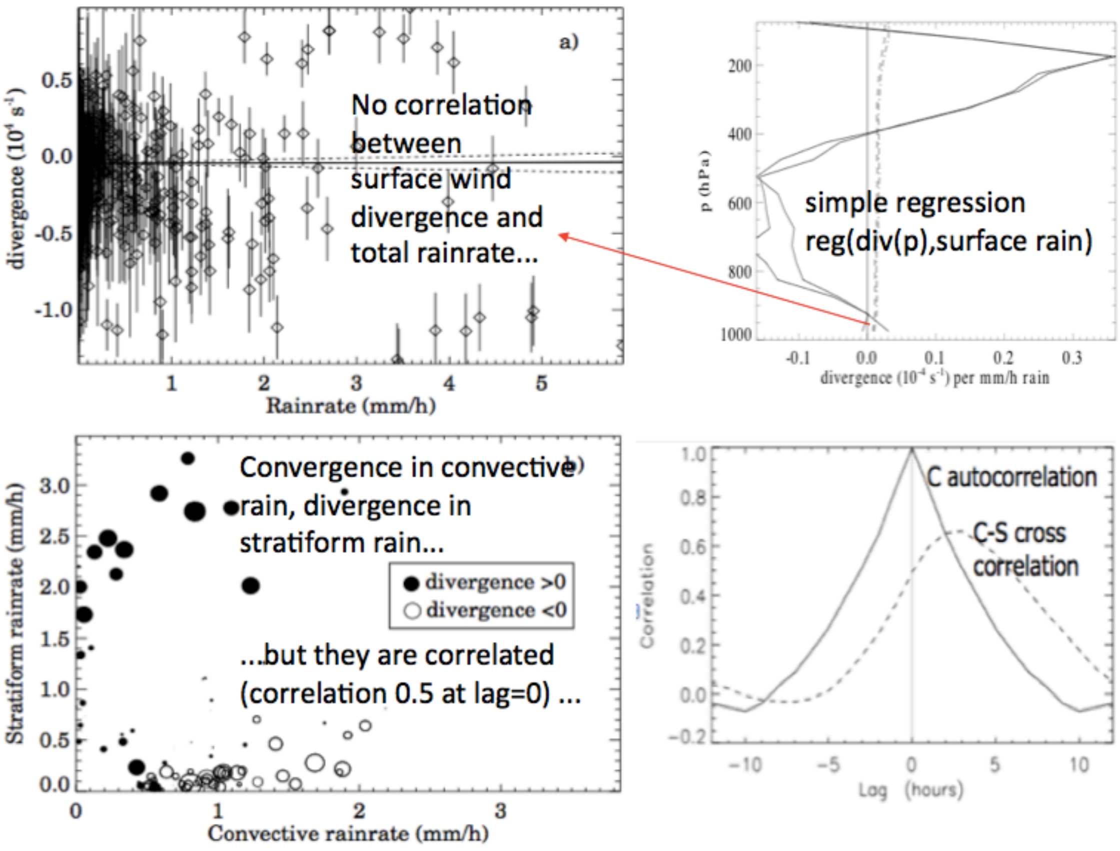
Here's a double check from a Web search.
http://luna.cas.usf.edu/~mbrannic/files/regression/Reg2IV.html
For the two variable case, fitting y = b1x1 + b2x2:

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

and

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Here are the relevant key graphs from Lin and Mapes 2004:



No correlation between surface wind divergence and total rainrate...

simple regression reg(div(p),surface rain)

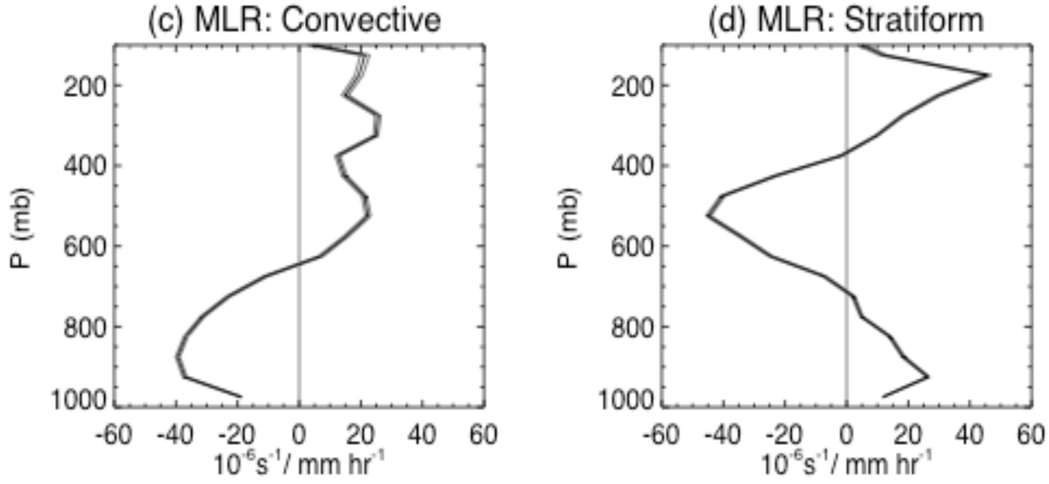
Convergence in convective rain, divergence in stratiform rain...

...but they are correlated (correlation 0.5 at lag=0) ...

C autocorrelation

C-S cross correlation

MLR teases it apart!



Part 4: Another way to use (and think of) matrices: as “filters” of time series.

$$\begin{bmatrix} g(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix} = \begin{bmatrix} 1 &&&&& \\ & 1 &&& 0 & \\ && 1 &&& \\ &&& 1 && \\ & 0 &&& \cdots & \\ &&&&& 1\end{bmatrix} \begin{bmatrix} f(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix}$$

No filter (the identity matrix)

$$\begin{bmatrix} g(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix} = \begin{bmatrix} ? &&&&& \\ -1 & 0 & 1 && 0 & \\ & -1 & 0 & 1 && \\ && -1 & 0 & 1 & \\ & 0 && -1 & 0 & 1 \\ &&&&& ?\end{bmatrix} \begin{bmatrix} f(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix}$$

Time derivative. Ends need a special rule.

$$\begin{bmatrix} g(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix} = 1/3\begin{bmatrix} 2 & 1 &&&& \\ 1 & 1 & 1 && 0 & \\ & 1 & 1 & 1 && \\ && 1 & 1 & 1 & \\ & 0 && 1 & 1 & 1 \\ &&&& 1 & 2\end{bmatrix} \begin{bmatrix} f(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix}$$

Centered smoother. Ends need a special rule.

$$\begin{bmatrix} g(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix} = 1/3\begin{bmatrix} 1 & 0 &&&& \\ 1/2 & 1 & 0 && 0 & \\ & 1/2 & 1 & 0 && \\ && 1/2 & 1 & 0 & \\ & 0 && 1/2 & 1 & 0 \\ &&&& 1/2 & 1\end{bmatrix} \begin{bmatrix} f(t)\\ \downarrow\\ \downarrow\\ \downarrow\\ \cdots\end{bmatrix}$$

Causal filter. Only the past affects g(t).

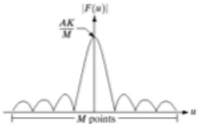
A spike in f would pick out a column of the filter and return it as the column vector result g(t). Results for arbitrary f can be built up as a sequence of such spikes.

I might give you a matrix, and a graph of a time series f(t), and ask you to sketch g(t).

Or I give graphs of f(t) and a filtered g(t), and ask you to sketch the matrix M.

Part V: Fourier analysis

Try to gain a deep understanding of how the Fourier transform maps variance (power) from the physical domain (time or space) to the corresponding spectral domain (frequency or wavenumber).
Also, a time series and its autocorrelation function have the SAME power spectrum (only the phase spectrum, i.e. the phase of each sinusoidal componenent, is different). For example, a time series of uncorrelated random values in time has a white power spectrum (equal amplitude of all frequencies, with random phase). Meanwhile the same power spectrum (white), if all the components are phase aligned at t=0, gives an infinitely tall spike (the delta function). The delta function is the *autocorrelation function* of random white noise: perfect correlation at lag 0, zero correlation at any lag other than zero. Autocorrelation is always a sum of all the Fourier components with their phase aligned at lag=0.

physical space (time t)	spectral space (frequency f)
sine wave cos(ft)	delta function (spike) at frequency f
localized spike $\delta(t)$	white spectrum (flat)
uncorrelated random values	same as above
wide or smoothed spike	red spectrum (reduced high frequencies compared to white noise)
Gaussian	Gaussian
boxcar	sinc() function: 

You should understand how a finite time interval appears in the power spectrum (as a lowest possible frequency, equal to the bandwidth = finest possible resolution in frequency).

You should understand how a finite sampling interval appears (as a highest resolvable frequency, the Nyquist frequency).

You should be alert to the existence of vertical error bars on spectral peaks in power spectra (from aliasing; or from having too few cycles in the sample so that it might involve the statistics of small numbers, or appear by happenstance of sampling). For example, if you have a decade of hourly data, and find a peak at 5-year period, it is only from 2 cycles even though it may be from thousands of datapoints. The information content and statistical significance is low, despite all the datapoints.

You should have opinions about power spectrum plotting conventions. Visibility is a virtue, but's nice to keep track of the variance budget by eye, so weird stretchings of one axis should be compensated on the other to keep the variance budget "honest".