Data, Models, and Uncertainty in the Natural Sciences

GEO422

© Frederik J Simons
Princeton University
November 24, 2020

Contents

Part I

Inferential Statistics

1

Basic concepts

1.1 Data, models, and uncertainty

Literal quote from Hand.

Read, in the book by Tarantola [1], the *Preface* and *Chapter 1*, sections 1.2.1 through 1.2.4, for some choice remarks. Read Bendat & Piersol [2] Chapter 3 and 4.

1.2 Location and scale

Location (center) and scale (spread, dispersion, ...). See Figure 1.1.

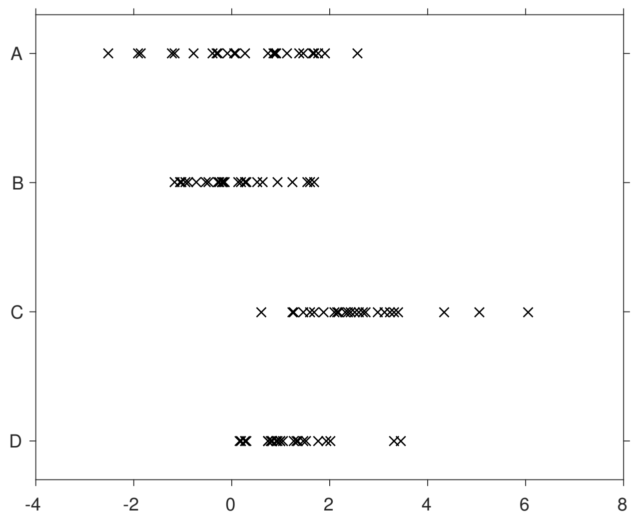


Fig. 1.1. Guess the distributions and the parameters of these distributions.

1.3 The histogram

1.3.1 Discrete data

For a set of possible *outcomes*, $m = 1, \ldots, M$, we denote by f_m the number of times each outcome is observed in a series of N experiments. If this number is expressed relative to the number of tries, we obtain the *(relative) frequencies* f_m/N of the *histogram*. The frequencies satisfy

$$\sum_{m=1}^{M} f_m = N \qquad \text{and} \qquad \sum_{m=1}^{M} \frac{f_m}{N} = 1. \tag{1.1}$$

We define the *distribution function* as the relative number of times that the observed outcome is smaller than one of the outcomes m', i.e.

$$F_N(m') = \frac{1}{N} \sum_{m=1}^{m'} f_m. \tag{1.2}$$

1.3.2 Continuous data

The continuous observable x is *binned* into a set of bins labeled $m = 1, \ldots, M$. As above, the absolute number of times that the outcome of an experiment consisting of N trials is observed to fall in a certain bin m is denoted f_m, the relative number is then f_m/N, and eq. (1.2) still holds of course. Thus, f_m/N is again the *(relative) frequency*, but now of the observations that lie in the bin around x_m of width Δ.

Read Tarantola Chapter 1... Finiteness and continuity, smoothness, discretization, not the same thing.

1.3.3 Samples and populations

All of this is *empirical*, i.e. based on the finite data *sample*. Let us now rescale the relative frequency by the size of the interval, i.e.

$$\frac{f_m}{N\Delta} \to p(x) \tag{1.3}$$

such that the *area* of the bar,

$$\Delta \frac{f_m}{N\Delta} \tag{1.4}$$

equals the relative frequency, then let $N \to \infty$ and $\Delta \to 0$ and we get information on the *population*: and $p\,(x)$ is the *probability density function*:

$$\lim_{N\to\infty} \lim_{\Delta\to 0} \frac{f_m}{N\Delta} = p(x). \tag{1.5}$$

1.4 The probability density function — small p and big P

We define $p(x)$, the probability density function, with the following properties. All values are hit, some value had to have occurred, thus the probability density function is normalized

$$\int_{-\infty}^{+\infty} p(x)\ dx = 1. \tag{1.6}$$

The probability density function is positive or zero:

$$p(x) \geq 0. \tag{1.7}$$

And the probability of finding x in the interval $[x, x+dx]$ is given by

$$P(x') = \int_{-\infty}^{x'} p(x)\ dx = \text{Prob}\,[x \leq x'] \tag{1.8}$$

This defines $P(x)$, the *probability distribution function*,

$$\frac{dP(x)}{dx} = p(x) \tag{1.9}$$

which has the following properties:

$$P(-\infty) = 0, P(\infty) = 1 \qquad \text{and} \qquad P(a) \leq P(b)\ \text{ if } a \leq b \tag{1.10}$$

1.5 Two random variables — $P(x)$, $P(y)$ and $P(x, y)$

Let $P(x)$ and $P(y)$ be two different pdfs, then we introduce the *joint probability distribution function* $P(x, y)$, with properties as follows:

$$P(x', y') = \text{Prob}[x \leq x' \text{ and } y \leq y'] \tag{1.11}$$

$$P(-\infty, y) = P(x, -\infty) = 0 \tag{1.12}$$

$$P(\infty, \infty) = 1 \tag{1.13}$$

$$p(x, y) \geq 0 \tag{1.14}$$

Normalization

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x, y)\ \ dx\, dy = 1 \tag{1.15}$$

Positivity

$$p(x, y) \geq 0 \tag{1.16}$$

The relation to the distribution function

$$P\left(x', y'\right) = \int\limits_{-\infty}^{x'} \int\limits_{-\infty}^{y'} p\left(x, y\right)\, dx\, dy \tag{1.17}$$

$$\frac{\partial}{\partial y}\left[\frac{\partial P\left(x, y\right)}{\partial x}\right] = p\left(x, y\right) \tag{1.18}$$

These are the *marginal probabilities*:

$$p\left(x\right) = \int\limits_{-\infty}^{+\infty} p\left(x, y\right)\, dy \qquad \text{and} \qquad p\left(y\right) = \int\limits_{-\infty}^{+\infty} p\left(x, y\right)\, dx. \tag{1.19}$$

Maybe here Bayes from Tarantola's first edition.

1.6 Statistical independence

Let's define a notation for an "event" happening,

$$P(A) = P(x \in A) = \int_A p(x)\, dx \tag{1.20}$$

What is the *conditional probability*, i.e. the probability that event A happens *given that* event B happens:

$$P(A|B) \tag{1.21}$$

What is the *joint probability* of two *events*?

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{1.22}$$

The above is a statement of *Bayes' theorem*, often written as:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}.} \tag{1.23}$$

See this in words - e.g. posterior equals likelihood times prior. Young and Smith (2005). Or the probability of the hypothesis given the data is the probability of the data given the hypothesis times the probability of the hypothesis. Sivia ().

Independence of both events means

$$P\left(A|B\right) = P\left(A\right) \tag{1.24}$$

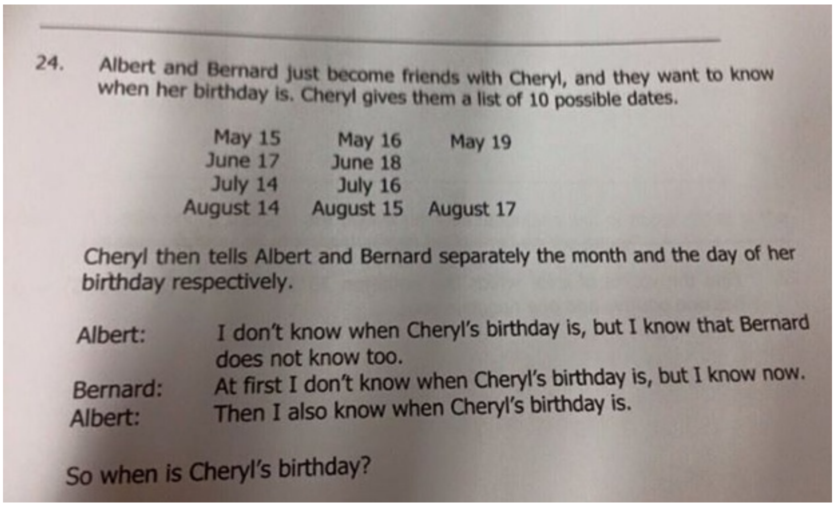
$$P\left(A \cap B\right) = P\left(A\right) P\left(B\right) \tag{1.25}$$

The equivalent for the probability density and distribution functions for independently distributed variables:

$$p\,(x, y) = p\,(x)\,p\,(y) \tag{1.26}$$

$$P\,(x, y) = P\,(x)\,P\,(y) \tag{1.27}$$

Illustrate using Strogatz' thing. Distinguish "events-type" Bayes theorem from "pdf-type Bayes theorem"? Tarantola between editions changes his mind on the presentation.



24. Albert and Bernard just become friends with Cheryl, and they want to know when her birthday is. Cheryl gives them a list of 10 possible dates.

May 15	May 16	May 19
June 17	June 18	
July 14	July 16	
August 14	August 15	August 17

Cheryl then tells Albert and Bernard separately the month and the day of her birthday respectively.

Albert: I don't know when Cheryl's birthday is, but I know that Bernard does not know too.

Bernard: At first I don't know when Cheryl's birthday is, but I know now.

Albert: Then I also know when Cheryl's birthday is.

So when is Cheryl's birthday?

Fig. 1.2. As read online: "Albert, Bernard and Cheryls threesome sets the web aflutter".

1.7 Expectation

I am going to define the expectation via an intermediary which is most certainly NOT the expectation, purely for didactical reasons.

Now, let x_i be the $i = 1, \ldots, N$ observations in some experiment, and define the **arithmetic mean** of the data as given by

$$\bar{x} = \frac{1}{N} \sum_i x_i. \tag{1.28}$$

Suppose x was not a continuous variable but rather could take on only the discrete values $m = 1, \ldots, M$, then we would formulate an alternative whereby for every such outcome m we add up the relative number of times that it has occurred in our experiment:

$$\bar{x} = \frac{1}{N} \sum_m m f_m, \tag{1.29}$$

and if the data were binned with bin width Δ we would define

$$\bar{x} = \frac{1}{N} \sum_m m \frac{f_m}{\Delta} \Delta. \tag{1.30}$$

Make next thing intuitive by saying that in the limit that $N \to \infty$ and $\Delta \to 0$ you **now** have the *population mean* or *expected* or *average value*:

$$E\{x\} = \langle x \rangle = \int_{-\infty}^{+\infty} x \, p\,(x) \; dx = \mu_x \tag{1.31}$$

What is the expected value of a function of x?

$$E\{g(x)\} = \langle g(x) \rangle = \int_{-\infty}^{+\infty} g(x) \, p(x) \, dx \tag{1.32}$$

Further properties for two random variables:

$$E\{x + y\} = E\{x\} + E\{y\} \tag{1.33}$$

which we derive using the joint and the marginal distributions.

Small aside.

$$E\{x + y\} = \int \int (x + y) p(x, y) \, dx \, dy \tag{1.34}$$

$$= \int \int x p(x, y) \, dx \, dy + \int \int y p(x, y) \, dx \, dy \tag{1.35}$$

$$= \int x \underbrace{\left[\int p(x, y) \, dy \right]}_{p(x)} dx + \int y \underbrace{\left[\int p(x, y) \, dx \right]}_{p(y)} dy \tag{1.36}$$

$$= E\{x\} + E\{y\} \tag{1.37}$$

In the above I talk about *linearity* and this being a *moment*.

1.8 Variance

Again, I start by writing something that it isn't, but which makes sense. What is the mean-squared value of x?

$$E\{x^2\} = \left\langle x^2 \right\rangle = \int_{-\infty}^{+\infty} x^2\, p(x)\, dx. \tag{1.38}$$

Now define the **variance**, the mean-squared variation of x about the mean?

$$E\{(x - E\{x\})^2\} = \left\langle\, (x - \langle x \rangle)^2 \,\right\rangle = \int_{-\infty}^{+\infty} (x - \mu_x)^2\; p(x)\, dx = \sigma_x^2 \tag{1.39}$$

This is the *variance*, the square of the *standard deviation*. Now of course

$$\left\langle\, (x - \langle x \rangle)^2 \,\right\rangle = \left\langle x^2 + \langle x \rangle^2 - 2x \langle x \rangle \,\right\rangle \tag{1.40}$$

$$= \left\langle x^2 \right\rangle + \langle x \rangle^2 - 2 \langle x \rangle \langle x \rangle \tag{1.41}$$

$$= \left\langle x^2 \right\rangle - \langle x \rangle^2 \tag{1.42}$$

And thus

$$\mathrm{var}\{x\} = E\{x^2\} - E^2\{x\} \tag{1.43}$$

Now: properties of the mean and variance:

$$E\{a + g(x) + h(x)\} = a + E\{g(x)\} + E\{h(x)\} \tag{1.44}$$

$$E\{a\, g(x)\} = a\, E\{g(x)\} \tag{1.45}$$

$$\mathrm{var}\{ax + b\} = a^2 \mathrm{var}\{x\} \tag{1.46}$$

The last thing defines a *standardized variable* of mean zero and variance one.

$$E\{x - E\{x\}\} = 0 \tag{1.47}$$

$$\mathrm{var}\left\{ \frac{x - E\{x\}}{\sqrt{\mathrm{var}\{x\}}} \right\} = 1 \tag{1.48}$$

And THAT is when I draw a picture of why everyone, all the time, should refer all things to their "location" and scale by their "spread"... not distinguishing for now population vs sample, just driving the point home that the units should be shifted and scaled to start noticing things. I draw some crosses on an axis, and then redraw the axis once after shifting, once after shifting and scaling. This is not captured in my Blackboard picture, which I erased before taking a picture.

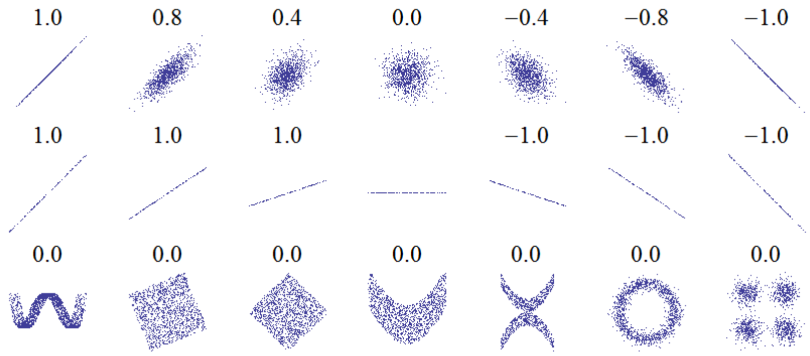


Fig. 1.3. How you can be wrong about correlation and dependence. From the Wikipedia: *The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.*

1.9 Covariance

$$\begin{aligned} E\{(x - E\{x\})\,(Y - E\{y\})\} &= \langle (x - \langle x\rangle)\,(y - \langle y\rangle)\rangle \\ &= \langle xy - x\,\langle y\rangle - \langle x\rangle\, y + \langle x\rangle\,\langle y\rangle\rangle \\ &= \langle xy\rangle - \langle x\rangle\,\langle y\rangle \qquad (1.49) \\ &= \mathrm{cov}\{x, y\} \qquad (1.50) \end{aligned}$$

thus for independent variables:

$$\mathrm{cov}\{x, y\} = \iint x\,y\,p\,(x, y)\;dx\,dy - \int x\,p(x)\,dx \int y\,p\,(y)\;dy = 0 \quad (1.51)$$

Independence implies zero covariance, though the reverse is not true!

Population *correlation coefficient*:

$$\rho\,(x, y) = \frac{\mathrm{cov}\{x, y\}}{\sqrt{\mathrm{var}\,(x)\;\mathrm{var}\,(y)}}. \qquad (1.52)$$

It of course follows that

$$-1 \leq \rho(x, y) \leq 1. \qquad (1.53)$$

Perhaps here Tarantola first edition grey boxes?

I need to introduce **iid** some time very soon.

1.10 A suitable notation

Need to talk about the dot product. And the transpose of a dot product. And the dyad, but really should not be writing the transpose.

The variance of a linear combination of variables is:

$$\mathrm{var}\Big\{ \sum_i w_i x_i \Big\} = \sum_i \sum_j w_i \,\mathrm{cov}\{x_i, x_j\}\, w_j, \tag{1.54}$$

and

$$\mathrm{cov}\{x, x\} = \mathrm{var}\{x\} \tag{1.55}$$

Note: if x_i and x_j are independent, scrap cov of cross terms and leave var terms only.

It just gets tiresome inventing new letters for different random variables such as x, y, and so on. For convenience, let's call them all x_i, $i = 1, ..., N$, however many we need. A bit of linear algebra here would be good. At the minimum level, go with

$$\mathrm{cov}\{\mathbf{x}\} = \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle \quad \text{if} \quad \langle \mathbf{x} \rangle = \mathbf{0} \tag{1.56}$$

and then do an arbitrary linear transform of which eqs and are special cases.

$$\begin{aligned} \mathrm{cov}\{\mathbf{A}\cdot\mathbf{x}\} &= \langle (\mathbf{A}\cdot\mathbf{x}) \left(\mathbf{x}^{\mathrm{T}} \cdot \mathbf{A}^{\mathrm{T}}\right) \rangle & (1.57) \\ &= \mathbf{A} \cdot \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle \cdot \mathbf{A}^{\mathrm{T}} & (1.58) \\ &= \mathbf{A} \cdot \mathrm{cov}\{\mathbf{x}\} \cdot \mathbf{A}^{\mathrm{T}} & (1.59) \end{aligned}$$

Some simple index rules. The elements of $\tilde{\mathbf{x}} = \mathbf{A}\cdot\mathbf{x}$ are given in index notation by $\tilde{x}_i = \sum_j A_{ij} x_j$ and the elements of the transformed covariance matrix $\tilde{\mathbf{C}} = \mathbf{A}\cdot\mathbf{C}\cdot\mathbf{A}^{\mathrm{T}}$ are given by $\tilde{C}_{ij} = \sum_k \sum_l A_{ik} C_{kl} A^T_{lj}$ as per the common rules of tensor transformation. And if A_{ij} is simply $w_i \delta_{ij}$ then get eq. (1.54) back.

Put a simulation picture here, e.g. BMI.

Order of things was 1. moments, 2. eq. (1.43), 3. eq. (1.47).

1.11 Estimation properties

Truth versus estimate. Constant versus random. Unbiased, consistent, efficient in words.

Let's say we have a property, s, and we estimate it, $\hat{s}$. How good is the estimate? Brackets denote repeated measurements. Properties: (s is the only truth, $\hat{s}$ is an estimate)

BIAS OF THE ESTIMATOR ("accuracy")

$$\boxed{b = \langle \hat{s} \rangle - s}$$

Note that $b^2 = \langle \hat{s} \rangle^2 + s^2 - 2s\langle \hat{s} \rangle$.

VARIANCE OF THE ESTIMATOR ("precision")

$$\boxed{v = \left\langle \, (\hat{s} - \langle \hat{s} \rangle)^2 \, \right\rangle}$$

Note that

$$\begin{aligned} v = \left\langle \, (\hat{s} - \langle \hat{s} \rangle)^2 \, \right\rangle &= \left\langle \hat{s}^2 + \langle \hat{s} \rangle^2 - 2\hat{s}\langle \hat{s} \rangle \right\rangle \\ &= \langle \hat{s}^2 \rangle + \langle \hat{s} \rangle^2 - 2\langle \hat{s} \rangle^2 \\ &= \langle \hat{s}^2 \rangle - \langle \hat{s} \rangle^2 \end{aligned} \tag{1.60}$$

and that we'll continue to use the notation $\mathrm{var}\{\}$ when we have something concrete to stick within the squiggly brackets.

ERROR OF THE ESTIMATE

$$\boxed{\epsilon = \hat{s} - s}$$

MEAN-SQUARED ERROR

$$\begin{aligned} \langle \epsilon^2 \rangle = \left\langle \, (\hat{s} - s)^2 \, \right\rangle &= \langle \hat{s}^2 + s^2 - 2\hat{s}s \rangle \\ &= \langle \hat{s}^2 \rangle + s^2 - 2\langle \hat{s} \rangle s \\ &= v + b^2 \end{aligned} \tag{1.61}$$

Thus in conclusion, the mean squared error is the sum of the variance plus the square of the bias.

$$\boxed{\langle \epsilon^2 \rangle = v + b^2.}$$

Now launch into a philosophical discussion of how minimum-bias or minimum-variance are only part of the problem, rather one minimizes the mse — which often means a slight bias to get variance reduction.

Unbiasedness, consistency (if the sequence of estimators converges in probability to the truth), efficiency (if there is some best possible state reached, usually the minimal mean-squared error).

1.12 The sample mean

So, we have a *population*, and calculate for a sample of **uncorrelated** observations x_i the arithmetic mean:

$$\boxed{\hat{\mu}_x = \frac{1}{N} \sum_{i=1}^{N} x_i.}$$

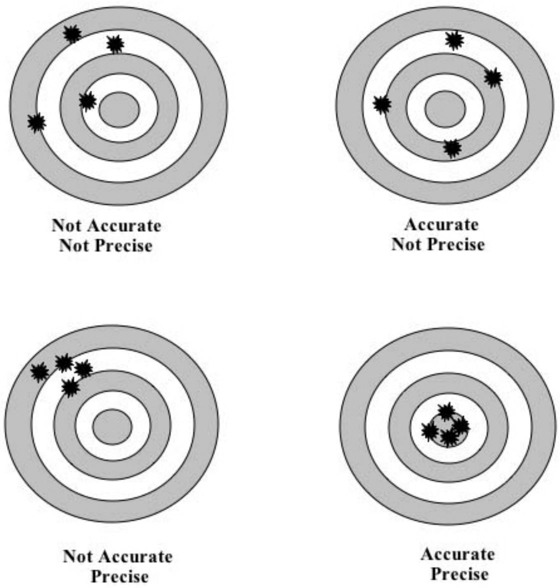


Fig. 1.4. Accuracy, precision. Bias, variance.

How good of an estimate is this for the mean of the population $\mu_x = \langle x \rangle$? We consider each of the possible observations as random variables themselves and wonder about their properties should we have access to more than one set of them.

Bias of the sample mean

$$\langle \hat{\mu}_x \rangle - \mu_x = \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle - \mu_x \tag{1.62}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \mu_x - \mu_x \tag{1.63}$$

$$= 0. \tag{1.64}$$

So the arithmetic mean is an *unbiased* estimator for the population mean:

$$\boxed{b\{\hat{\mu}\} = 0}$$

Variance of the sample mean

The long way round.

$$\begin{aligned}
\langle(\hat{\mu}_x - \mu_x)^2\rangle &= \Big\langle \left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu_x\right)^2 \Big\rangle &(1.65)\\
&= \frac{1}{N^2}\Big\langle \left(\sum_{i=1}^{N} x_i - N\mu_x\right)^2 \Big\rangle \\
&= \frac{1}{N^2}\Big\langle \left(\sum_{i=1}^{N} (x_i - \mu_x)\right)^2 \Big\rangle &(1.66)\\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle (x_i - \mu_x)(x_j - \mu_x)\rangle &(1.67)\\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sigma_x^2 &(1.68)\\
&= \frac{\sigma_x^2}{N} &(1.69)
\end{aligned}$$

since x_i and x_j are drawn from the same parent distribution with common mean and variance — and no covariance between the samples, $\langle x_i, x_j\rangle = 0$.

Key "notational" tricks are (search for the right place to put this!):

$$\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \hat{\mu}_x\right)^2 = \left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i - N\hat{\mu}_x\right)\right]^2 \tag{1.70}$$

and

$$\left(\sum_{i=1}^{N} x_i - N\hat{\mu}_x\right) = \sum_{i=1}^{N}(x_i - \hat{\mu}_x) \tag{1.71}$$

So the arithmetic mean is a **consistent** estimator for the population mean:

$$\boxed{v\{\hat{\mu}\} = \frac{\sigma_x^2}{N}}$$

We could have also done this straight from the result of unbiasedness and the rule eq. (1.54)

$$\langle(\hat{\mu}_x - \mu_x)^2\rangle = \mathrm{var}\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\right\} \tag{1.72}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{var}\{x_i\} \tag{1.73}$$

$$= \frac{\sigma_x^2}{N} \tag{1.74}$$

since — or rather, if! — there are no covariances in this case. Variance and mean squared error coincide.

Conclusion: the arithmetic mean $\hat{\mu}_x$ is an unbiased estimate for the population mean μ_x, with variance and mean-squared error both given by σ_x^2/N, with σ_x^2 the population variance.

Nothing here has implied the estimators are any "good"... no considerations of efficiency. Efficiency is another matter – is this the best we can do? Is this a mininum in any sense (yes!)

Perhaps **here** is where I show Tarantola's grey boxes?

1.13 The sample variance

Maybe do what I do with Olhede in "methodology", which is much shorter.

Going in knowing the answer, define this as the estimate:

$$\boxed{\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2}$$

based on the sample mean. Note: Press et al say "If the difference between N and $N-1$ matters to you, you are probably up to no good anyways."

Bias of the sample variance:

What is the bias due to this? Start with the expected value with respect to the sample mean. We know that

$$\mathrm{var}\{x_i\} = \langle x_i^2\rangle - \langle x_i\rangle^2 = \sigma_x^2 \tag{1.75}$$

We know from eqs (1.62)–(1.64) that $\langle\hat{\mu}_x\rangle = \mu_x$. We know from eqs (1.72)–(1.74) that $\mathrm{var}\{\hat{\mu}_x\} = \langle\hat{\mu}_x^2\rangle - \langle\hat{\mu}_x\rangle^2 = \frac{\sigma_x^2}{N}$.

We're dealing with independent samples, $\langle x_i x_j \rangle = 0$. So now we are ready for this calculation:

$$\langle \hat{\sigma}_x^2 \rangle = \Big\langle \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_x)^2 \Big\rangle \tag{1.76}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \langle (x_i - \hat{\mu}_x)^2 \rangle \tag{1.77}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left(\langle x_i^2 \rangle + \langle \hat{\mu}_x^2 \rangle - 2 \langle x_i \hat{\mu}_x \rangle \right) \tag{1.78}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \langle x_i^2 \rangle + \frac{N}{N-1} \langle \hat{\mu}_x^2 \rangle - \frac{2}{N-1} \sum_{i=1}^{N} \langle x_i \hat{\mu}_x \rangle \tag{1.79}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \langle x_i^2 \rangle - \frac{N}{N-1} \langle \hat{\mu}_x^2 \rangle \tag{1.80}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left(\sigma_x^2 + \hat{\mu}_x^2 \right) - \frac{N}{N-1} \left(\frac{\sigma_x^2}{N} + \mu_x^2 \right) \tag{1.81}$$

$$= \frac{N}{N-1} \sigma_x^2 + \frac{N}{N-1} \mu_x^2 - \frac{\sigma_x^2}{N-1} - \frac{N}{N-1} \mu_x^2 \tag{1.82}$$

$$= \sigma_x^2 \tag{1.83}$$

It's unbiased:

$$\boxed{b\{\hat{\sigma}_x^2\} = 0}$$

Note the trick:

$$\frac{2}{N-1} \sum_{i=1}^{N} \langle x_i \hat{\mu}_x \rangle = \frac{2}{N-1} \Big\langle \sum_{i=1}^{N} x_i \hat{\mu}_x \Big\rangle \tag{1.84}$$

$$= \frac{2}{N-1} \langle N \hat{\mu}_x \hat{\mu}_x \rangle \tag{1.85}$$

$$= \frac{2N}{N-1} \langle \mu_x^2 \rangle \tag{1.86}$$

Variance of the sample variance:

Need a proper definition of moments, raw, centered, just a moment, higher orders.

$$\mu_n = \int_{-\infty}^{+\infty} (x - \mu_x)^2 \, p\,(x) \; dx \tag{1.87}$$

When you do this – and it takes a while, you end up with:

$$\mathrm{var}\{\hat{\sigma}_x^2\} = \frac{1}{N} \left(\mu_4 - \frac{N-3}{N-1} \mu_2^2 \right) \tag{1.88}$$

which considerably simplifies for the normal distribution since the *central moments* (to be distinguished from the *crude* or *raw* moments around zero instead of around the expected value) are well known: $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$ thus the variance of the sample variance for normally distributed variables is given by

$$\mathrm{var}\{\hat{\sigma}_x^2\} = \frac{2\sigma^4}{N-1}. \tag{1.89}$$

We have mentioned the normal, only now, but wait till later.

NOW link this with the spatial statistics, RB X p 21, and compare how for correlated fields the variance loses degrees of freedom, whereas for Whittle likelihood, it doesn't!

2
Distributions

We have the notion of a pdf. We need to know what happens to a pdf when we add variables, when we differentiate variables, when we add MANY variables together. This quite naturally introduces the normal and the chi-squared distribution, and then we continue to motivate the central-limit theorem and the general notion of "fitting" data as "modeling residuals". Through hypothesis testing as a check whether all the assumptions are fullfilled, and then, yes, a certain p-value, if you must.

2.1 Adding variables

Refer to "a suitable notation" since we know something about the moments already. But now we need to learn about the distributions.

Let x and y be two random variables with a joint probability density distribution $p\,(x, y)$. What is the probability density function of the random variable which is the sum $z = x + y$? *For each fixed value of x, the corresponding* $y = z - x$, and thus the joint pdf is given by

$$p\,(x, y) = p\,(x, z - x)\,. \tag{2.1}$$

For each fixed value of z, the variable x can range from $-\infty$ to $+\infty$, and so the marginal distribution of z will be

$$p\,(z) = \int_{-\infty}^{+\infty} p\,(x, z - x)\;dx. \tag{2.2}$$

If x and y are independent, and their probability density functions are $p_1(x)$ and $p_2(y)$, respectively, then we obtain the general result that the pdf of the sum variable is the *convolution* of the individual ones:

$$\boxed{p\,(z) = \int\limits_{-\infty}^{+\infty} p_1(x)\, p_2(z-x)\, dx.} \tag{2.3}$$

Example

Let x and y be random variables with uniform distributions

$$p_1(x) = \frac{1}{a} \qquad \text{for} \qquad 0 \le x \le a, \tag{2.4}$$

$$p_2(y) = \frac{1}{a} \qquad \text{for} \qquad 0 \le y \le a. \tag{2.5}$$

What is the distribution of $z = x + y$? As we know

$$p_2(y) = p_2(z-x), \tag{2.6}$$

which applies in the range

$$0 \le z - x \le a, \tag{2.7}$$

$$0 \ge -z + x \ge -a, \tag{2.8}$$

$$z \ge x \ge z - a. \tag{2.9}$$

In other words, the distribution of the sum is the restricted convolution integral

$$p(z) = \int_{z-a}^{z} p_1(x)\, p_2(z-x)\, dx. \tag{2.10}$$

There are two regimes for this, since $0 \le z \le 2a$ but $0 \le x \le a$ and x cannot be bigger than z. Write down eq. (2.10) but then point to the need to truncate the upper, and the lower limit at 0 and at a. We have

$$p(z) = \begin{cases} \int\limits_0^z \left(\frac{1}{a}\right)^2 dx = \left(\frac{1}{a}\right)^2 z & \text{for } 0 \le z \le a, \\ \int\limits_{z-a}^{a} \left(\frac{1}{a}\right)^2 dx = \dfrac{2a-z}{a^2} & \text{for } a \le z \le 2a, \end{cases} \tag{2.11}$$

and 0 otherwise, which is a triangular function. Properly normalized. Picture here: Uniform for y on 0 to a. Uniform for x on 0 to a. Triangle for z on a to 2a.

An alternative derivation of the convolution rule

Working with the distribution function:

$$P(x+y \le z) = P(z) = \iint_{x+y\le z} p(x,y)\,dx\,dy, \tag{2.12}$$

$$= \iint_{x+y\le z} p_1(x)\,p_2(y)\,dx\,dy, \tag{2.13}$$

$$= \int_{-\infty}^{+\infty} p_1(x) \left[\int_{-\infty}^{z-x} p_2(y)\,dy \right] dx, \tag{2.14}$$

$$= \int_{-\infty}^{+\infty} p_1(x)\,P_2(z-x)\,dx. \tag{2.15}$$

On to the probability density function:

$$p(z) = \frac{dP}{dz} = \int_{-\infty}^{+\infty} p_1(x) \frac{dP_2}{dz}(z-x)\,dx, \tag{2.16}$$

$$= \int_{-\infty}^{+\infty} p_1(x)\,p_2(z-x)\,dx, \tag{2.17}$$

which is what we had before. We are of course using the chain rule of differentiation, by which $(dy/dz = 1)$

$$\frac{dP_2}{dz} = \frac{dP_2}{dy}\frac{dy}{dz} = \frac{dP_2}{dy} = p_2. \tag{2.18}$$

Well, keep doing this and get the following. For *any* underlying distribution by CENTRAL LIMIT THEOREM. In fact, *any* distribution with finite variance would do it = CLT.

2.2 The Gaussian distribution

The standard form of the *Gaussian* or *normal* probability density function is (`normpdf`):

$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right].} \tag{2.19}$$

The distribution function is (`normcdf`):

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left[-\frac{(x'-\mu)^2}{2\sigma^2}\right] dx'. \tag{2.20}$$

This is not called the *error function* (`erf`), but it is close. Variables X that are distributed normally with expectation μ and variance σ^2 are denoted

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right). \tag{2.21}$$

Expectation of the Gaussian distribution

$$\begin{aligned}
\langle x \rangle &= \int\limits_{-\infty}^{+\infty} x\, p(x)\, dx && (2.22)\\
&= \int\limits_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx && (2.23)\\
&= \int\limits_{-\infty}^{+\infty} (x+\mu) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx && (2.24)\\
&= \mu \int\limits_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\
&\qquad + \int\limits_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx && (2.25)\\
&= \mu. && (2.26)
\end{aligned}$$

The first term equals 1, because it is just simply the normalization of the probability density function, $\int p(x)\, dx$. The second term is the integral of an odd function over a symmetric interval – it thus vanishes. Matlab example of how to do this integral?

Variance of the Gaussian distribution

$$\langle (x - E\{x\})^2 \rangle = \int\limits_{-\infty}^{+\infty} (x-\mu)^2 p(x)\, dx = \sigma^2, \tag{2.27}$$

which we prove using the well-known integral that, depending on the text book, bears the names of Gauss, Poisson and Euler, namely:

$$\int_{-\infty}^{+\infty} e^{-x^2}\, dx = \sqrt{\pi}. \tag{2.28}$$

Usingr Euler-Poisson formula that gives us $\sqrt{\pi} = \Gamma(1/2)$. Trick is to square the integral to a double, then switch to polar coordinates. See “Gaussian integral” on Wikipedia. Matlab example of how to do this integral?

Percentiles and Intervals

Quantile-quantile plots are a good thing to talk about here. `qqplot` Invented by Wilkes! Remember Tukey. Not that long ago.

Scaling

If $x \sim \mathcal{N}(\mu, \sigma^2)$, how is $z = (x - \mu)/\sigma$ distributed? We already talked about how this works in terms of the variance and the mean.

Matlab example of how it’s just an axis scaling.

2.3 Changing variables

If $x \sim \mathcal{N}(\mu, \sigma^2)$, how is $(x - \mu)/\sigma$ distributed? Well, it’s just a *change of variables*? How’s that done? Talk about x and then the functional mapping $g(x) = y$. Knowing $p(x)$, what is $p(y)$? From the graph, we can see that

$$p(y)\, dy = p(x)\, dx. \tag{2.29}$$

But how does dy relate to dx? Let’s do a first-order **Taylor series** for the function $g(x) = y$ and try to find the value of $g(x + dx) = y + dy$:

$$g(x + dx) = g(x) + \frac{dg}{dx} dx = y + dy. \tag{2.30}$$

From this we conclude that, to first order:

$$dy = \frac{dg}{dx} dx = \frac{dy}{dx} dx, \tag{2.31}$$

of course, and thus

$$p(x)\, dx = p(y)\, \frac{dy}{dx}\, dx, \tag{2.32}$$

which we formalize to ensure positivity by writing:

$$\boxed{p(y) = p(x)\left|\frac{dy}{dx}\right|^{-1}.} \tag{2.33}$$

Jacobian, my friend. BUT: need to ensure normalization.

Picture. Make infinitesimal or else the argument won't work well. Deform a p(x) to a p(g(x)). Shade areas of equal probability. Relate dx and dy.

First example: The standard normal distribution

We return to the question at hand, whereby $x \sim \mathcal{N}(\mu, \sigma^2)$ is a Gaussian variable with expectation μ and variance σ^2, and we are finding out what the distribution is of the variable obtained by subtracting the mean from x and dividing the result by the standard deviation of x. The transformation is:

$$y = \frac{x-\mu}{\sigma} \qquad \text{and} \qquad \frac{dy}{dx} = \frac{1}{\sigma} \qquad \text{and} \qquad \left|\frac{dy}{dx}\right|^{-1} = \sigma, \tag{2.34}$$

and thus the distribution of the resulting variable is:

$$p(y) = \frac{\sigma}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.35}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right), \tag{2.36}$$

which is nothing else but $p(x)$ when $\mu = 0$ and $\sigma = 1$, and which is therefore called *standard-normally* distributed. We will denote important result this as:

$$\boxed{\text{If} \quad x \sim \mathcal{N}\left(\mu, \sigma^2\right) \quad \text{then} \quad \left(\frac{x-\mu}{\sigma}\right) \sim \mathcal{N}(0,1).} \tag{2.37}$$

See something about the distribution of the maxima of the sample as a function of N, which I heard by Bouchard (?) at PCTS and reminded me of Donoho and $\sqrt{(2\ln N)}$ in the thresholding paper.

Alternative derivation of the standard normal

$$z = \frac{x-\mu}{\sigma} \quad \text{when } X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{2.38}$$

$$\begin{aligned}
P_z(z') &= \text{Prob}(z \leq z') && (2.39)\\
&= \text{Prob}\left(\frac{x-\mu}{\sigma} \leq z'\right) && (2.40)\\
&= \text{Prob}(x-\mu \leq \sigma z') && (2.41)\\
&= \text{Prob}(x \leq \mu + \sigma z') && (2.42)\\
&= P_{\text{Gaussian}}(\mu + \sigma z') && (2.43)\\
&= P_x(\mu + \sigma z') && (2.44)
\end{aligned}$$

but

$$x \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{2.45}$$

so we know it already

$$P_z(z) = P_x(\mu + \sigma z) \tag{2.46}$$

$$p(z) = \frac{dP}{dz} = \sigma p_x(\mu + \sigma z) \tag{2.47}$$

by the chain rule, therefore, by substitution,

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \sim \mathcal{N}(0,1) \tag{2.48}$$

This is the easy way for a linear transformation. See BP Chap 3-4, Trauth, and Aster.

Second example: The chi-squared distribution

Now let x be a standard normal variable,

$$x \sim \mathcal{N}(0,1) \tag{2.49}$$

and let's ask ourselves the question what the distribution is of its square:

$$y = x^2. \tag{2.50}$$

Evidently

$$\frac{dy}{dx} = 2x = 2\sqrt{y} \qquad \text{and} \qquad \left|\frac{dy}{dx}\right|^{-1}, \tag{2.51}$$

leading to:

$$\boxed{p(x^2) = p(y) = \frac{1}{\sqrt{2\pi}} \frac{e^{-y/2}}{\sqrt{y}}.} \tag{2.52}$$

This is the so-called χ^2-distribution with one degree of freedom, χ_1^2. We remind ourselves of some properties of the gamma function

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \tag{2.53}$$

$$\Gamma(1) = 1, \tag{2.54}$$

$$\Gamma(N-1) = N\Gamma(N). \tag{2.55}$$

$$\Gamma(N) = (N-1)!. \tag{2.56}$$

In Matlab, see `chi2pdf` and `chi2cdf`. Make the connection with `factorial` and `gamma`. Don't get bogged down, though having the formula for the Gamma function would be nice.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt \tag{2.57}$$

2.4 The chi-squared distribution

Because in standard form the χ^2-distribution with n degrees of freedom is going to be the result of adding multiple sums of squares of normal variables together as

$$y = z_1^2 + z_2^2 + \cdots + z_n^2 \tag{2.58}$$

which results in

$$\boxed{p(y) = \frac{1}{2^{n/2}\Gamma(n/2)} \exp\left(-\frac{y}{2}\right)(y)^{n/2-1}.} \tag{2.59}$$

We will be using the notation with N "degrees of freedom" that:

$$\boxed{\text{If} \quad x_i \sim \mathcal{N}(0,1) \quad \text{then} \quad \sum_i^N x_i^2 \sim \chi_N^2.} \tag{2.60}$$

The χ^2 distribution is a one-parameter distribution, whose expectation and variance are determined solely by the number of its degrees of freedom, n,

$$E\left(\chi_n^2\right) = \mu_{\chi^2} = n \tag{2.61}$$

$$E\left[(\chi_n^2 - \mu_{\chi^2})^2\right] = \sigma_{\chi^2}^2 = 2n \tag{2.62}$$

2.5 General statements about fitting data

Here we make the point that yall need to know about how "sums of squares" are distributed. I reiterated variance and bias and rmse with a hypothetical example of how to fit a third-order polynomial with a first, then second, etc. degree fit. Each time calculating the metrics.

Here I had some 3 pictures in Lecture 5 which worked well. Whatever goes to χ^2. Are the data normal? Are the errors? Residuals? What is a qq-plot? χ^2 again, somewhat self-referentially. .

It's all about **sums of squares**.

Polynomial fitting. Need to explain this better and hands-on. Good Blackboard Oct 3, 2017.

2.6 Convolution

See Papoulis book on Systems and Optics. See PW p 161. Start from the general transform with reproducing-type kernels, then specify to LTI which implies convolution! Simple, elegant.

convolution.png

commutativity

Mention the Hilbert transform AS a convolution with 1/pi, so they've all heard of it! As a way of first bringing out the Hilbert transform.

What is convolution? We will need it often.

$$y(t) = \int_{-\infty}^{+\infty} h(\tau)\, x(t-\tau)\, d\tau, \tag{2.63}$$

whereby t is an arbitrary input. $h(t)$ is the impulse response. Discrete version – see Strang. The value of the *output* $y(t)$ is given as a weighted, linear, infinite sum, over the entire history of the *input* $x(t)$. Th weighting function $h(t)$ is called the *impulse* response, the value of the output is due to a unit impulse input at a time τ before.

$$h(t) = \int_{-\infty}^{+\infty} h(\tau)\delta(t-\tau)\, d\tau \tag{2.64}$$

Say what the δ function is.

For a causal (i.e. *physically* realizable) system, we cannot have any output before we received an input, in other words

$$h(\tau) = 0 \qquad \text{for} \qquad \tau \leq 0 \tag{2.65}$$

How about in the discrete case? This is probably how we can best understand the procedure by which functional (also stochastic, statistical) inputs get mapped via deterministic "filter" functions into outputs:

$$y(n) = \sum_{k=0}^{N} h(k)x(n-k) \tag{2.66}$$

Note: Green functions, is a smoothing kernel. This is not a cross-correlation, just so you know. Cross-correlation has POSITIVE sign inside, that's all. Distro of X+Y is convo, of X-Y is cross-co!

So construct this sequence and then read it from the bottom up, you'll see eq. (2.66) appear before your eyes for the simple case when $N = 1$ thus the filter is of length $N + 1 = 2$:

$$h = \begin{bmatrix} h(0) & h(1) \end{bmatrix} \tag{2.67}$$

Every input gets multiplied by however long the response takes, and this for every sample of the inputs. The results are added. Whatever lingers.

Need some kind of arrows up on the top that say, at time one, at time two, etc.

$$
\begin{array}{llll}
\text{time 1} & \text{time 2} & \text{time 3} & \\
\times\left\{\begin{array}{l} x(0) \\ h(0) \end{array}\right. & \begin{array}{l} x(0) \\ h(1) \end{array} & & \\
 & \times\left\{\begin{array}{l} x(1) \\ h(0) \end{array}\right. & \begin{array}{l} x(1) \\ h(1) \end{array} & \\
 & & \times\left\{\begin{array}{l} x(2) \\ h(0) \end{array}\right. & \begin{array}{l} x(2) \\ h(1) \end{array} \\
\hline
\end{array}
$$

$$
\begin{array}{l}
y(0) = h(0)x(0) \\
\qquad\qquad y(1) = h(0)x(1) + h(1)x(0) \\
\qquad\qquad\qquad\qquad y(2) = h(0)x(2) + h(1)x(1)
\end{array}
$$

Read the results and write them out from the bottom up.

Now take a good look at we just did:

$$y(0) = h(0)\,x(0) \tag{2.68}$$
$$y(1) = h(0)\,x(1) + h(1)\,x(0) \tag{2.69}$$
$$y(2) = h(0)\,x(2) + h(1)\,x(1) \tag{2.70}$$
$$\vdots \qquad \vdots \qquad + \qquad \vdots \tag{2.71}$$
$$y(n) = h(0)\,x(n) + h(1)\,x(n-1) \tag{2.72}$$

and the general form follows. Picture. See Strang and Nguyen, first page. Should you want to represent this as a matrix operation, you could, too! You'd get a Toeplitz matrix.

discrete impulse $x(0) = 1$

$$y(0) = h(0) \tag{2.73}$$
$$y(1) = h(1) \tag{2.74}$$

H is a "filter" causal, with finite impulse response.

First Example: Continuous convolution

The impulse response:

$$h(\tau) = e^{-\tau}, \quad \tau \geq 0 \tag{2.75}$$

The signal:

$$x(t) = 1 \quad \tau \geq 0 \tag{2.76}$$

To calculate the convolution,

$$[h * x](t) = \int_{-\infty}^{+\infty} x(\tau)\, h(t-\tau)\, d\tau \tag{2.77}$$

Now, flip the signs and shift

$$h(-\tau) = e^{\tau}, \quad \tau \leq 0 \tag{2.78}$$

$$h(t-\tau) = e^{t-\tau} \quad \tau \leq t \tag{2.79}$$

No restriction on the sign of t: it can be anything. Multiply, then integrate:

$$(e^{\tau} * 1)[t] = \int_{-\infty}^{+\infty} e^{-(t-\tau)}\, d\tau = \int_{0}^{t} e^{\tau-t}\, d\tau = \left. e^{\tau-t} \right|_0^t = 1 - e^{-t}. \tag{2.80}$$

Now $t - \tau \geq 0$ and $\tau \geq 0$ thus $0 \leq \tau \leq t$.

Second example: Discrete convolution

Now the signal is given by

$$s = \begin{bmatrix} -1 & 2 & 4 & -6 & 5 & 2 & 0 & 1 \end{bmatrix}, \tag{2.81}$$

and we consider the filters

$$h_1 = \begin{bmatrix} 1 & 1 \end{bmatrix} \quad \text{and} \quad h_2 = \begin{bmatrix} 1 & -1 \end{bmatrix}. \tag{2.82}$$

Think about what could be going on with h_1. It's some kind of a moving average, or smoothing, or lowpass.

$$\begin{array}{ccccccccc}
-1 & -1 & & & & & & & \\
 & 2 & 2 & & & & & & \\
 & & 4 & 4 & & & & & \\
 & & & -6 & -6 & & & & \\
 & & & & 5 & 5 & & & \\
 & & & & & 2 & 2 & & \\
 & & & & & & 0 & 0 & \\
 & & & & & & & 1 & 1 \\
\hline
\end{array} \tag{2.83}$$

$$\begin{bmatrix} -1 & 1 & 6 & -2 & -1 & 7 & 2 & 1 & 1 \end{bmatrix} \tag{2.84}$$

Think about what could be going on with h_1. It's some kind of a moving difference, or roughening, or highpass:

$$\begin{array}{ccccccccc}
-1 & 1 & & & & & & & \\
 & 2 & -2 & & & & & & \\
 & & 4 & -4 & & & & & \\
 & & & -6 & 6 & & & & \\
 & & & & 5 & -5 & & & \\
 & & & & & 2 & -2 & & \\
 & & & & & & 0 & 0 & \\
 & & & & & & & 1 & -1 \\
\hline
\end{array} \tag{2.85}$$

$$\begin{bmatrix} -1 & 3 & 2 & -10 & 11 & -3 & -2 & 1 & -1 \end{bmatrix} \tag{2.86}$$

Also do example of Strang and Nguyen p. 6.

2.7 Confidence Intervals

Generalities. We've so far contented ourselves with describing what happens to the mean and variances of certain types of variables. Clearly, we want the whole distribution to be able to make quantitative inference of any kind.

2.7.1 How good is the sample mean?

Let μ be the population mean. let $\hat{\mu}$ be the estimate based on N samples.

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.87}$$

i.e. the arithmetic mean.

Sure, $\langle \hat{\mu} \rangle = \mu$ as we have proved. This is unbiased. also,

$$\mathrm{var}\{\hat{\mu}\} = \langle \hat{\mu}^2 \rangle - \langle \hat{\mu} \rangle^2 = \frac{\sigma^2}{N} \tag{2.88}$$

as we have seen. Note: we did it via $\mathrm{mse} = v + b^2$. Note: known variance

Wikipedia: The standard error (SE) is the standard deviation of the sampling distribution of a statistic,[1] most commonly of the mean (e.g. standard error of the mean). So compared to the above, you may want to do

$$\frac{s^2}{N} \tag{2.89}$$

if s is your estimate of σ.

But really, we would like a *confidence interval*, i.e. to be able to say that the true μ falls in some interval around the estimated $\hat{\mu}$ with some degree of confidence. We need the whole pdf of the estimator in eq. (2.87). Where there are N independent observations of the random variable, x.

The data are normally distributed and their variance is known

Let's say $x \sim \mathcal{N}(\mu, \sigma^2)$, what is the pdf of

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.90}$$

for $n \gg 1$?

Even if the x weren't normal the argument would still hold...

(i) its *normal*, too (we know this from the CLT on $n \gg 1$)
(ii) its mean is μ

(iii) its standard deviation is $\sqrt{\sigma^2/N}$

In other words,

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \tag{2.91}$$

Therefore,

$$\boxed{\frac{\hat{\mu}-\mu}{\sigma}\sqrt{N} \sim \mathcal{N}(0,1)} \tag{2.92}$$

is standard-normally distributed. We've done this too.

What does this mean for us? If we did a hundred tests of $N = 100$, and computed each time the sample mean according to eq. (2.87), we quantify the probability of

$$\text{Prob}\left(z_{\alpha/2} \leq \frac{\hat{\mu}-\mu}{\sigma}\sqrt{N} \leq z_{1-\alpha/2}\right) = 1-\alpha \tag{2.93}$$

from this we derive a *confidence interval*.

$$\text{Prob}\left(\frac{\sigma}{\sqrt{N}}z_{\alpha/2} + \mu \leq \hat{\mu} \leq \frac{\sigma}{\sqrt{N}}z_{1-\alpha/2} + \mu\right) = 1-a \tag{2.94}$$

We define z_α as the value for which the distribution function of the standard normal reaches the probability value of $\alpha \times 100\%$.

If μ is the mean, then, with a probability $1-\alpha$ will we find $\hat{\mu}$ in our trials:

$$\boxed{\mu = \hat{\mu} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}} \tag{2.95}$$

Note: $z_{1-\alpha/2} = -z_{\alpha/2}$ symmetric distributions, such as the standard normal z. Mention **twosided**. Picture here of a Gaussian, let's say, with $\alpha/2$ on the left and $1-\alpha/2$ on the right.

Goodman2016, Baker2016, Nuzzo2014.

Note — once we're done, we have found a single $\hat{\mu}$, no longer a random variable! So technically, either $\hat{\mu}$ falls into the interval or it doesn't. The previous equation is the interpretation as the basis for hypothesis testing.

This is a it of a step... (BR p63 "slight logical step") What BP 4.44 says is equivalent to treating μ like the unknown and $\hat{\mu}$ like the given. In the same number cases does μ fall within the confidence interval based on $\hat{\mu}$.

Link it back to Bayes and likelihood and Sivia.

So, if we know the *mean* and *variance* of the data, we can find the distribution of the sample mean $\hat{\mu}$ – we can use this to *test hypotheses*:

If mean is μ and variance is σ^2 and sample size is N, then what are the chances of observing $\hat{\mu}$ if this *null hypothesis* is true?

High? Accept. Low? Reject. Pick a threshold, that would b ethe (1-a)100 percent confidence interval.

The data are normally distributed but their variance is unknown

Clearly somewhere in here we'll need a test for normality to begin with!

We'll have to put in the sample variance for the lack of anything better. Thus we first figure out what the distribution is of the sample variance!

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{2.96}$$

where x is normal with μ and σ^2. We already know much about it:

(i) it's a sum of squares, so χ^2 will come in
(ii) its expected value is the variance, σ^2.
(iii) eq. (1.88) gave us an inkling as to its variance, remember $2\sigma^4/(N-1)$.

Let us start from the answer and work backwards. It has been derived (Papoulis p260) that the sum of squares of standard-normally distributed variables

$$\boxed{\sum_{i=1}^{N}\frac{(x_i - \hat{\mu})^2}{\sigma^2} \sim \chi^2_{N-1},} \tag{2.97}$$

We are penalized by one degree of freedom for having first computed the sample mean. One of the terms in the equation is fixed when we have all but one of the terms and their mean, quite intuitively.

To verify the expectation of the variable in eq. (2.97), take eq. (2.96) and multiply it by

$$\left\langle \sum_{i=1}^{N}\frac{(x_i - \hat{\mu})^2}{\sigma^2} \right\rangle = \left\langle \frac{(N-1)}{\sigma^2}\,\hat{\sigma}^2 \right\rangle = \frac{(N-1)}{\sigma^2}\langle\hat{\sigma}^2\rangle = N-1 \tag{2.98}$$

which follows from eq. (2.96), $\langle\hat{\sigma}^2\rangle = \sigma^2$. So eq. (2.98) follows by linearity of the expectation.
The variance of the variable in eq. (2.97) is $2(N-1)$. So,

$$\mathrm{var}\left(\frac{\sigma^2}{N-1}\frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{\sigma^2}\right) = \frac{2(N-1)}{(N-1)^2}\sigma^4 = \frac{2\sigma^4}{N-1} \tag{2.99}$$

by which we have found again the variance of the sample variance for normally distributed variables with variance σ^2, $\mathrm{var}\{\hat{\sigma}^2\}$ (Kenny and Keeping p164).

Note that this is the short form of an equation which, to say it most generally, would have a couple of other terms in it. We've encountered this once, it is worth reverifying what the results are. See the Wolfram article on this (Sample Variance Distribution), quite helpful.

So the distribution of eq. (2.97), simply rewritten as

$$\frac{(N-1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{N-1}, \tag{2.100}$$

which gives us all the information we need about the distribution of eq. (2.96). The distribution of the sample variance of size N is proportional to a χ^2 distribution with $N-1$ degrees of freedom.

We also know the distribution of the mean from eq. (2.92). We'd like to have an equation of the same form. Note: which we know to be true by the CLT and the scaling argument. But we don't know the true variance of the sample. Let's try something finding the distribution of

$$\boxed{\frac{\hat{\mu}-\mu}{\hat{\sigma}}\sqrt{N} \sim \text{what?}} \tag{2.101}$$

Point out the difference with eq. (2.92).

We have a Gaussian variable $\hat{\mu}$ distributed as eq. (2.91), and a χ^2 variable $\hat{\sigma}^2$ distributed as eq. (2.100). That which we want to know, eq. (2.101) is in other words the ratio of a standard normally distributed variable, the numerator of eq. (2.101),

$$z = \frac{\hat{\mu}-\mu}{\sigma}\sqrt{N} \sim \mathcal{N}(0,1) \tag{2.102}$$

and another variable, the denominator of eq. (2.101),

$$\frac{\hat{\sigma}}{\sigma} = \sqrt{\frac{y}{N-1}} \tag{2.103}$$

whereby y is the variable whose distribution we know already also:

$$y = (N-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-1} \tag{2.104}$$

The distribution of this ratio of variables was worked out a long time ago by Gosset,

$$\boxed{\frac{\hat{\mu}-\mu}{\hat{\sigma}}\sqrt{N} \sim t_{N-1}} \tag{2.105}$$

whereby t_n has a pdf

$$p(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\sqrt{n}\,\Gamma(\frac{n}{2})}\left[1+\frac{t^2}{n}\right]^{-\frac{n+1}{2}} \tag{2.106}$$

The student t distribution $\overset{n\longrightarrow\infty}{\longrightarrow}$ Gaussian. (Papoulis p207, Biometrica p908) asymptotically.

Because we *can* calculate the student t distribution, and we can go through the same argument as before to construct confidence intervals on the *population* mean, based on the *sample mean* and the *sample variance*

$$\boxed{\mu = \hat{\mu} \pm t_{N-1;1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}} \tag{2.107}$$

All of these were symmetric — even when the distributions are asymmetric we can do this.

2.7.2 How good is the sample variance?

We've just worked out

$$(N-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-1}. \tag{2.108}$$

and from this we learn

$$\text{Prob}\left(\chi^2_{N-1;\alpha/2} \leq (N-1)\frac{\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{N-1;1-\alpha/2}\right) = 1-\alpha \tag{2.109}$$

Now translate this to a statement on σ^2 given $\hat{\sigma}^2$.

$$\text{Prob}\left(\frac{\sigma^2}{N-1}\chi^2_{N-1;\alpha/2} \leq \hat{\sigma}^2 \leq \chi^2_{N-1;1-\alpha/2}\frac{\sigma^2}{N-1}\right) = 1-\alpha \tag{2.110}$$

rephrase this to a $(1-\alpha)\times 100\%$ confidence interval.

Or else start from before and do 1/over it,

$$\frac{1}{\chi^2_{N-1;1-\alpha/2}} \geq \frac{\sigma^2}{(N-1)\,\hat{\sigma}^2} \geq \frac{1}{\chi^2_{N-1;\alpha/2}} \tag{2.111}$$

or a confidence interval

$$\boxed{\frac{\hat{\sigma}^2(N-1)}{\chi^2_{N-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(N-1)}{\chi^2_{N-1;\alpha/2}}} \tag{2.112}$$

Now we can ask the question: how big does N need to be to get a certain level of confidence interval?

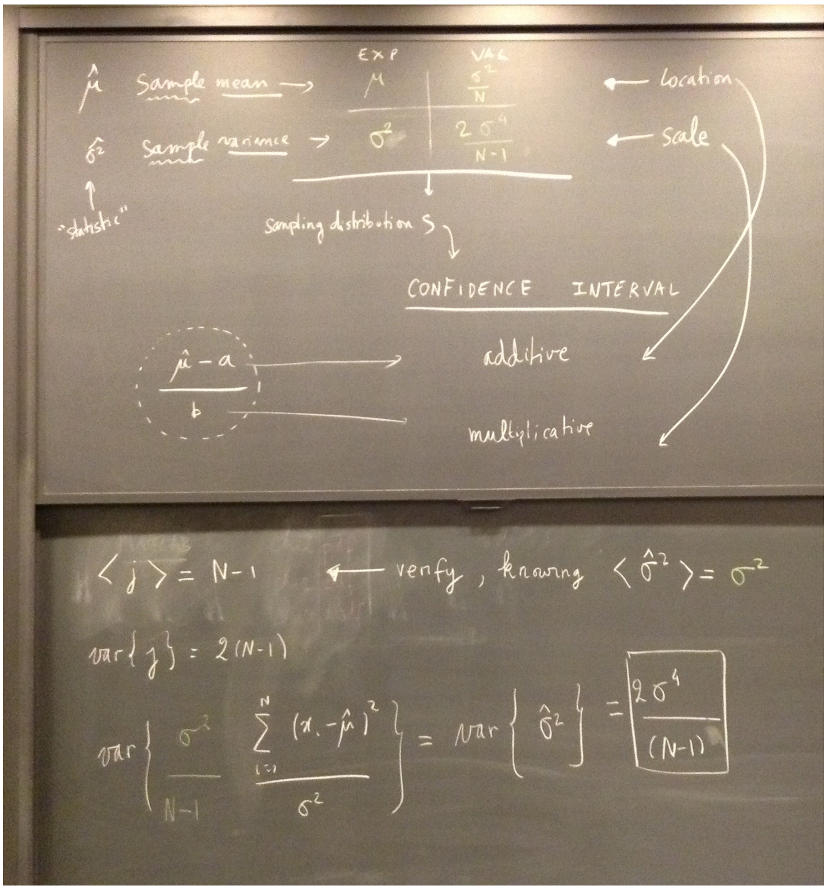


Fig. 2.1. Pieces of lecture. Right.

2.7.3 How good is the sample correlation coefficient?

Here the formula from BP. Never, ever, quote an “r-value” without quoting the significance level. Look at Snedecor, VandeCar+90.

HERE MAYBE NEED TO WORK OUT THE DISTRIBUTION OF THE RATIO OF TWO CHI-SQUAREDS? TO GET AT LEAST TO F.

THEN POINT OUT THAT WE HAD ADDING VARIABLES, CHANGING VARIABLES, SHOULD HAVE A SECTION ON FINDING THE PDFS OF MULTIPLIED VARIABLES.

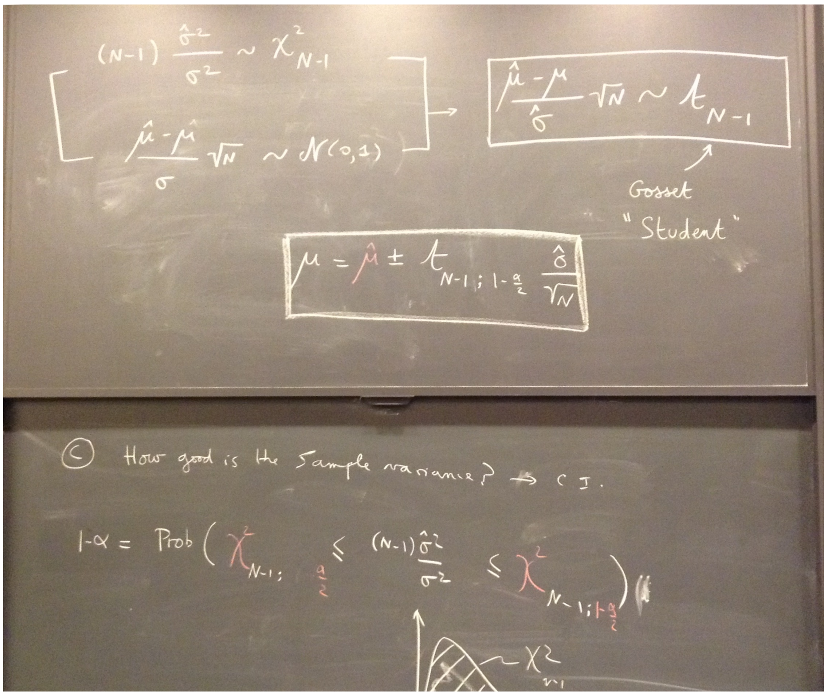


Fig. 2.2. Pieces of lecture. Left. A was How good is the sample mean – known variance. B was unknown variance.

2.8 Hypothesis testing

As to the lab, a good thing would be to run the test for 1:N each time M times, and plot the means and stds of the p values versus N with the thresholding.

Let's say you're after s, and you estimate $\hat{s}$ from its samples,

- We can figure out *bias*, *variance* and *mse*
- We can construct a confidence interval
- We can query the data under various hypotheses

E.g. the *mean*; the *sample* mean; its distribution
Null Hypothesis: H_0

- The truth is $\mu = \mu_0$
- We observe $\hat{\mu}$ until the shown probability
- What is the *chance* that the observed $\hat{\mu}$ falls as far apart from μ_0 as it does? or even further?

[graphic] big chance $\geq \alpha$ if this is $\geq \alpha$, accept H_0

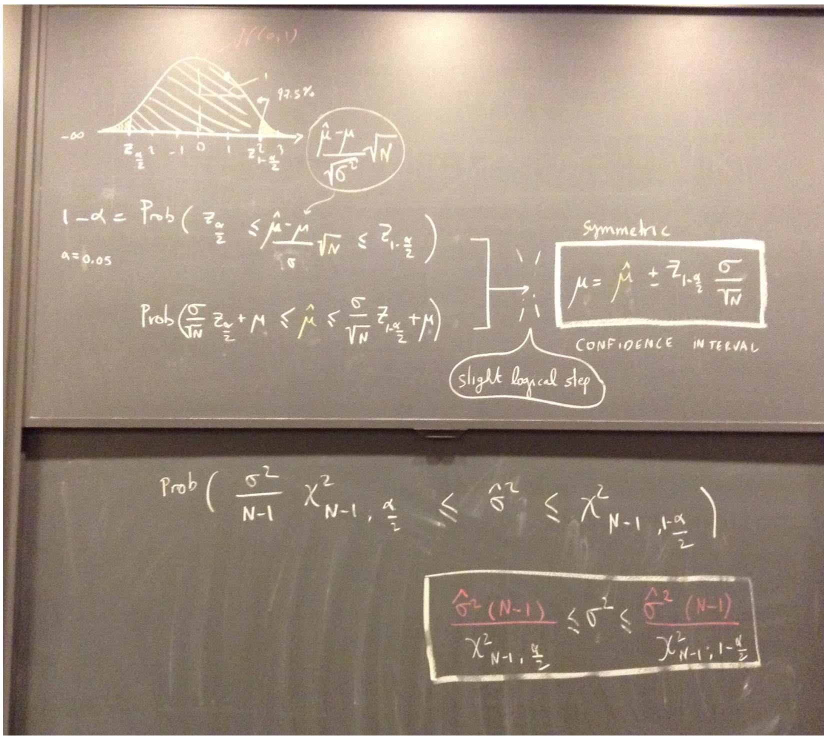


Fig. 2.3. Pieces of lecture. Middle.

[graphic] small chance $<$ a, it'd be extremely unlikely for H_0 to be true. Nevertheless, the chance exists, it's the Type I error.

Accept H_O if the change of seeing the result $< \alpha$. α is a type 1 error, you can *live* with.

Wrong hypothesis H_O

- the truth is $\mu \neq \mu_o$ but $\mu = \mu_1$
- we observe $\hat{\mu}$, with chance $\geq \alpha$ for H_0
- we decide in favor of $\mu = \mu_0$, wrongly

[graphic] this is the chance that you accept $\hat{\mu}$ for H_O, but wrongly

β is a type II error. This is the chance that the wrong hypothesis is accepted.

Reduce α, get better *significance*, but the probability β is increased.

$1 - \beta$ is called the *power* of the fit. Only increasing N can reduce α and β at the same time, by making the pdf's more peaked and thus better separated.

Example 4.2 is nice.

population	sample	“statistic”	distribution	expectation	variance
μ	mean	$\hat{\mu}$	$\mathcal{N}$	$\langle\hat{\mu}\rangle = \mu$	$\frac{\sigma^2}{N}$
σ^2	variance	$\hat{\sigma}^2$	$\frac{\sigma^2}{N-1}\chi^2_{N-1}$	$\langle\hat{\sigma}^2\rangle = \sigma^2$	$\frac{2\sigma^4}{N-1}$

Do lab 1! Give nice Matlab illustrations of this, with the tests. Also plot the chi2 that these X2 should look like, etc. 1 Graph for changing N, the X2 data, and the accepts/rejects 1 Graph for changing N, the p value, and the accepts/rejects. All for many loops and cases. Show that the chi2 is actually chi2 by its histogram!

Part II

Linear Inverse Theory

3
Best-fit type approaches

Let us be vague, in terms of notation, about the notions **model** (that which we want to know), **noise** (that which corrupts) and the **data** (that which we have), using the simple mnemonics m, n and d, and let us describe the **mapping** between model space and data space by some operator $\mathscr{G}$, as follows:

$$\mathscr{G}(m) + n = d. \tag{3.1}$$

3.1 The forward problem

A *forward model* is an operation that turns a set of *model parameters* into a set of observable *data*. Speaking quite generally and neglecting the influence of noise anywhere in this system, we write

$$G(m) = d. \tag{3.2}$$

We shall focus quite exclusively on *linear problems*, i.e. those for which, for some scalar values a and b, the following expression holds:

$$G(a\,m_1 + b\,m_2) = a\,G(m_1) + b\,G(m_2). \tag{3.3}$$

3.1.1 Continuous problems

Some general model

$$\int g(x,y)\,m(y)\,dy = d(x) \tag{3.4}$$

De/convolution

$$\int g(x-y)\,m(y)\,dy = d(x) \tag{3.5}$$

When is it nice to think about something in continuous, functional form? I last gave an example of how knowing an analytical inverse helps, or how, in the forward sense, to do an integral of something whose primitive function you know. Hansen, barcode reading thing. Maybe an midocean ridge for us?

Deconvolution/Toeplitz form - inversion... how it works for [1 1] and [1 2 1].

Example 1 (without a solution)

First an example of how a function is smoothed, running average, low-pass filtering. Draw a picture. We anticipate that going back is going to be hard. This is convolution like we've had in the continuous sense. But we also write the same equation out again in discrete form for continuity with the below and with the above on convolution that we'd already done.

3.1.2 Discrete problems

There is always a matrix

$$\boxed{\mathbf{G}\cdot\mathbf{m}=\mathbf{d}} \tag{3.6}$$

But this is abstract so we do it by example.

Example 2 (without a solution)

Second example: fitting a parabola of a throw in a gravity field. Draw a nice picture of some data, some parabolas through it, the distance to the data points, the mean squared error of the result.

$$d(t)=a+bt+ct^2 \tag{3.7}$$

write this as a matrix operation as follows:

$$N\text{ data}\left\{\begin{pmatrix} d_1\\ d_2\\ \vdots\\ d_N\end{pmatrix}\right. = \underbrace{\begin{pmatrix} 1 & t_1 & t_1^2\\ 1 & t_2 & t_2^2\\ \vdots & \vdots & \vdots\\ 1 & t_N & t_N^2\end{pmatrix}}_{N\times M\text{ matrix}} \left.\begin{pmatrix} a\\ b\\ c\end{pmatrix}\right\} M\text{ unknowns} \tag{3.8}$$

3.2 Solving linear inverse problems

Forget math class where you could get a *solution* by inverting something. Example: a scalar inverse, a 2 by 2 doable inverse. See examples.

Example 1: Estimating the Earth's average density

From either the mass or the moment of inertia. Earth mass, M, and moment of inertia, I, for $a = 6371 \times 10^3$ m the mean radius. Both of these are observables, call them d_1 and d_2:

$$M = 5.974 \times 10^{24} \text{ kg} = d_1, \tag{3.9}$$

$$I/a^2 = 1.975 \times 10^{24} \text{ kg} = d_2. \tag{3.10}$$

Two equations of the same kind: looking for a scalar from a single observation. The unknown is the average density, $\bar{\rho}$, let's call this m_1. Let's write how the mass and the moment of inertia depend on the density of a uniform solid sphere with this average density. Incidentally, just like we did with pdf's, these are moments of the density distribution of the Earth! In general, this would be the mass, M which is the zeroth moment of the "location" variable of the "density distribution", the center of mass, which would be the first, and the moment-of-inertia which would be the second. If we remember that the differential element of mass is given by

$$dM(r) = \rho(r)\, dV = \rho(r)\, r^2 \sin\theta\, dr\, d\theta\, d\varphi \tag{3.11}$$

we quickly arrive at the expressions for the total mass and moment-of-inertia with respect to the rotation axis which is the north pole of the coordinate system when $\theta = 0$, and let's integrate out the azimuthal variable already:

$$M = 2\pi \int_0^a \int_0^\pi \rho(r)\, r^2 \sin\theta\, dr\, d\theta, \tag{3.12}$$

$$I = 2\pi \int_0^a \int_0^\pi (r\sin\theta)^2\, \rho(r)\, r^2 \sin\theta\, dr\, d\theta. \tag{3.13}$$

Check that the center of mass is in the center and that it does not get the perpendicular projection that the moi requires, as it is with respect to the center not to an axis! Though it gets the cosine which makes it vanish. Think about what the average radius is compared to the center of mass. Rather, isn't the center of mass that location about which the variance of the density distribution is minimized. But no, we don't want the average radius of all of them, we want the average position on each of the three axes. So we've got to still project on the axes to get x, y, and z to ultimately be zero.

These are the forward models, and both are linear, so let's just call them G_{11} and G_{22} for now

$$M = \frac{4\pi}{3} a^3\, \bar{\rho} = G_{11}\, \bar{\rho} \tag{3.14}$$

$$I/a^2 = \frac{8\pi}{15} a^3\, \bar{\rho} = G_{22}\, \bar{\rho}. \tag{3.15}$$

and so from either of these equations taken individually, we can estimate the mean density to be. The solutions are obtained by simple linear scalar inversion. So we won't write the hat but we'll identify them by a superscript which says which linear model and which observation they are derived from:

$$\bar{\rho}_1 = 5.515 \times 10^3 \text{ kg m}^{-3} \qquad \text{from eq (3.14)} \tag{3.16}$$

$$\bar{\rho}_2 = 4.558 \times 10^3 \text{ kg m}^{-3} \qquad \text{from eq. (3.15).} \tag{3.17}$$

Note that another way to put it would be to say that we did this by simple inversion of the system of independently considered equations

$$\begin{bmatrix} M \\ I/a^2 \end{bmatrix} = \begin{bmatrix} \frac{4\pi}{3}a^3 & 0 \\ 0 & \frac{8\pi}{15}a^3 \end{bmatrix} \begin{bmatrix} \bar{\rho}_1 \\ \bar{\rho}_2 \end{bmatrix} \tag{3.18}$$

Example 2: Estimating the Earth's two-layer density

$a = 6371$ km, two-layer earth, $c = 3485$ From either the mass or the moment of inertia. Slightly more complicated, but completely solvable through matrix inversion.

Exact, unique solution.

$$M = \frac{4\pi}{3}c^3\,\rho_c + \frac{4\pi}{3}(a^3 - c^3)\,\rho_m \tag{3.19}$$

$$= G_{11}\,\rho_c + G_{12}\,\rho_m \tag{3.20}$$

$$I/a^2 = \frac{8\pi}{15}\frac{c^5}{a^2}\,\rho_c + \frac{8\pi}{15}\left(a^3 - \frac{c^5}{a^2}\right)\,\rho_m \tag{3.21}$$

$$= G_{21}\,\rho_c + G_{22}\,\rho_m. \tag{3.22}$$

When you're done

$$\rho_c = 12.492 \times 10^3 \text{ kg m}^{-3} \tag{3.23}$$

$$\rho_m = 4.150 \times 10^3 \text{ kg m}^{-3}. \tag{3.24}$$

We've solved

$$\begin{bmatrix} M \\ I/a^2 \end{bmatrix} = \frac{4\pi}{3} \begin{bmatrix} c^3 & \left(a^3 - c^3\right) \\ \frac{2}{5}\frac{c^5}{a^2} & \frac{2}{5}\left(a^3 - \frac{c^5}{a^2}\right) \end{bmatrix} \begin{bmatrix} \bar{\rho}_1 \\ \bar{\rho}_2 \end{bmatrix}. \tag{3.25}$$

Example 3: Estimating the Earth's average density, again

Another estimate would take both constraints on the single parameter. This is left for later. What if we try to satisfy both of these data *at the same time*?

We'd have:

$$\begin{bmatrix} M \\ I/a^2 \end{bmatrix} = \begin{bmatrix} \frac{4\pi}{3}a^3 \\ \frac{8\pi}{15}a^3 \end{bmatrix} \bar{\rho} \tag{3.26}$$

Counterexample: a rectangular matrix, an undoable inverse. Look again at the discrete form of the convolution matrix.

There is always noise, and perhaps the system has no solution. Talk about *signal* and *noise*. We now want an estimate *estimate* of the model, $\hat{\mathbf{m}}$ such that the solution

$$\mathbf{G} \cdot \mathbf{m} = \mathbf{d} = \mathbf{s} + \mathbf{n} \tag{3.27}$$

which is by definition to the *prediction*

$$\mathbf{G} \cdot \hat{\mathbf{m}} = \hat{\mathbf{d}} \approx \mathbf{d} \tag{3.28}$$

How close is close? There is the error or residual of course.

$$\mathbf{e} = \mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}} \tag{3.29}$$

We need to go back to our notions of *bias*, *variance* and *mean squared error*, which here is, up to a scaling, given by the norm or length of the vector with residuals.

3.3 The overdetermined problem

Do the full derivation here. Start again with a noisy version of

$$\mathbf{d} = \mathbf{G} \cdot \mathbf{m} + \mathbf{n} \qquad \text{and} \qquad \langle \mathbf{n} \rangle = \mathbf{0} \qquad \text{and} \qquad \langle \mathbf{s}\,\mathbf{n} \rangle = \mathbf{0} \tag{3.30}$$

Goal is to find an estimator $\hat{\mathbf{m}}$ that predicts this data using the linear model. The prediction error is

$$\mathbf{e} = \mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}. \tag{3.31}$$

We want to keep it small (to within where we assume it could be just noise, so back to the notions of significance and chi-squared statistics etc.) Define this the mean-squared error:

$$\phi = ||\mathbf{e}||_2^2 = \mathbf{e} \cdot \mathbf{e} = (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \tag{3.32}$$

Or, in index notation with(out) the summation convention:

$$\phi = \sum_{i=1}^{N} \left(d_i - \sum_{j=1}^{M} G_{ij} \hat{m}_j \right)^2 = (d_i - G_{ij}\hat{m}_j)\,(d_i - G_{in}\hat{m}_n) \tag{3.33}$$

And now take the derivative with respect to a generic model parameter

$$\frac{\partial \phi}{\partial \hat{m}_k} = (d_i - G_{ij}\hat{m}_j)(-G_{ik}) - G_{ik}(d_i - G_{in}\hat{m}_n) \tag{3.34}$$

$$= -2G_{ik}(d_i - G_{ij}\hat{m}_j) = 0. \tag{3.35}$$

Which is, if you will, in vector notation:

$$\boldsymbol{\nabla}_{\hat{\mathbf{m}}}\, \Phi = -2\mathbf{G}^{\mathrm{T}} \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) = \mathbf{0} \tag{3.36}$$

and thus clearly the minimzer of ϕ is when

$$\mathbf{G}^{\mathrm{T}} \cdot \mathbf{d} = \mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} \cdot \hat{\mathbf{m}}, \tag{3.37}$$

from which we derive the best-fitting least-squares estimate as use argmin here!

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{d}.} \tag{3.38}$$

In conclusion, this here is the "left-inverse" of the design matrix, etc.

$$\mathbf{G}^{-\mathrm{g}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}}. \tag{3.39}$$

Because of this, plugging in we see there is no bias.

$$\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot (\mathbf{G} \cdot \mathbf{m} + \mathbf{n}) \qquad \text{and} \qquad \langle \hat{\mathbf{m}} \rangle = \mathbf{m}. \tag{3.40}$$

How does the estimated model vector relate to the true model vector? The **model resolution matrix** in this case is the identity

$$\mathbf{R} = (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G})^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} = \mathbf{I}. \tag{3.41}$$

Example 3: Fitting a curve through data

This we can't solve yet, so we need a new tool. How do we characterize the solution? Still only have bias and variance to play with. Let's decide on a good measure. L2. Write this solution out for the linear regression with two parameters and you'll see the familiar sum of squares of *Menke's book*.

Example 4: Finding the mean of several measurements

Consider the problem of finding the value of a single parameter, m, from N repeated measurements. Written in matrix form, this becomes

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} m = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \tag{3.42}$$

And thus the solution is

$$\hat{m} = \left(\begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \tag{3.43}$$

which is the same as saying

$$\hat{m} = N^{-1} \sum_{i}^{N} d_i \tag{3.44}$$

which we recognize as the arithmetic mean.

Woohoo. Now change the norm, comment on the outlier behavior, go do an ℓ_1 problem to come up with the median. Refer to [3].

Example 5: Incorporating data uncertainty

Do this derivation without going into detail of the statistics just yet. However, we are once again solving

$$\mathbf{d} = \mathbf{G} \cdot \mathbf{m} + \mathbf{n} \tag{3.45}$$

by finding some linear operator that works on the data to return the model

$$\hat{\mathbf{m}} = \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d} \tag{3.46}$$

Start by assuming that the data and the model parameters have zero mean. Motivate this more properly but anyway

$$\langle \mathbf{d} \rangle = \mathbf{0} \qquad \text{and} \qquad \langle \mathbf{m} \rangle = \mathbf{0} \tag{3.47}$$

and defining the data covariance matrix to be the dyad or the "squared" pair

$$\mathbf{C}_{\mathrm{d}} = \langle \mathbf{d}\mathbf{d}^{\mathrm{T}} \rangle \tag{3.48}$$

We're looking for another $\mathbf{G}^{-\mathrm{g}}$ that works, that we can apply to the data to construct a model. We'd love to keep the model unbiased as before, but now we're talking variance, we can talk about model covariance as well. Whatever the linear model, the

$$\mathbf{C}_{\mathrm{m}} = \langle (\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d}) \cdot (\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d})^{\mathrm{T}} \rangle = \mathbf{G}^{-\mathrm{g}} \cdot \langle \mathbf{d}\mathbf{d}^{\mathrm{T}} \rangle \cdot \mathbf{G}^{-\mathrm{g}^{\mathrm{T}}} \tag{3.49}$$

$$= \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{C}_{\mathrm{d}} \cdot \mathbf{G}^{-\mathrm{g}^{\mathrm{T}}} \tag{3.50}$$

so it's just a quadratic form as expected.

This is the key to **error propagation for linear models** and students need to see it.

A new penalty function

Let's make a new penalty function, one that weights the prediction error of the data somehow, e.g., and minimize

$$\phi = (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W} \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \tag{3.51}$$

$$= (\mathbf{W} \cdot \mathbf{d} - \mathbf{W} \cdot \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot (\mathbf{W} \cdot \mathbf{d} - \mathbf{W} \cdot \mathbf{G} \cdot \hat{\mathbf{m}}) \tag{3.52}$$

$$= (\tilde{\mathbf{d}} - \tilde{\mathbf{G}}\hat{\mathbf{m}}) \cdot (\tilde{\mathbf{d}} - \tilde{\mathbf{G}}\hat{\mathbf{m}}) \tag{3.53}$$

Now it looks again like the first time around except for where

$$\tilde{\mathbf{G}} = \mathbf{W} \cdot \mathbf{G} \qquad \text{and} \qquad \tilde{\mathbf{d}} = \mathbf{W} \cdot \mathbf{d} \tag{3.54}$$

Solution once again is

$$\hat{\mathbf{m}} = (\tilde{\mathbf{G}}^{\mathrm{T}} \cdot \tilde{\mathbf{G}})^{-1} \cdot \tilde{\mathbf{G}}^{\mathrm{T}} \cdot \tilde{\mathbf{d}} \tag{3.55}$$

which, in other words, is

$$\boxed{\hat{\mathbf{m}} = (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{G})^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{d}} \tag{3.56}$$

See [4] p. 103 re IRLS.

At this point we should have already written the likelihoods from the next chapter. But maybe not. Now let's propose to use the inverse covariance for the interior weighting, i.e.

$$\mathbf{C}_{\mathrm{d}}^{-1} = \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W} \tag{3.57}$$

thereby minimizing

$$\phi = (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \tag{3.58}$$

forming the

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{d}} \tag{3.59}$$

Probably too much detail

In other words can show that this is the minimum-variance unbiased estimate without writing the likelihoods yet.

$$\mathbf{G}^{-\mathrm{g}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \tag{3.60}$$

What is the bias of this thing? Plug in, average, observe the cancellation. It's unbiased. What is the variance of this thing? This is called *BLUE*. [5, p. 138]. Minimizer of also minimizes the *variance matrix* of the estimate. Minimal *model covariance*? Previously we assumed we'd demeaned but now let's return for the full form. It doesn't matter. Well, the data covariance, $\mathbf{C}_{\mathrm{d}}$, is given by the dyad

$$\mathbf{C}_{\mathrm{d}} = \langle(\mathbf{d} - \mathbf{G} \cdot \mathbf{m})(\mathbf{d} - \mathbf{G} \cdot \mathbf{m})^{\mathrm{T}}\rangle \tag{3.61}$$

The model covariance, $\mathbf{C}_{\mathrm{m}}$, is given by the dyad

$$\begin{aligned}
\mathbf{C}_{\mathrm{m}} &= \langle(\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^{\mathrm{T}}\rangle && (3.62)\\
&= \langle(\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d} - \mathbf{m})(\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d} - \mathbf{m})^{\mathrm{T}}\rangle \\
&= \langle\mathbf{G}^{-\mathrm{g}} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m})(\mathbf{d} - \mathbf{G} \cdot \mathbf{m})^{\mathrm{T}} \cdot \mathbf{G}^{-\mathrm{g}^{\mathrm{T}}}\rangle \\
&= \mathbf{G}^{-\mathrm{g}} \cdot \langle(\mathbf{d} - \mathbf{G} \cdot \mathbf{m})(\mathbf{d} - \mathbf{G} \cdot \mathbf{m})^{\mathrm{T}}\rangle \cdot \mathbf{G}^{-\mathrm{g}^{\mathrm{T}}} \\
&= \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{C}_{\mathrm{d}} \cdot \mathbf{G}^{-\mathrm{g}^{\mathrm{T}}} && (3.63)
\end{aligned}$$

which is just like we had it. Variance is the about the average variation about the mean. Taking the mean out before you do it or leaving it in makes no damn difference. Make a note of the symmetry to transposition which is vital for all covariances and which we need later.

Doing is properly requires us to figure out what "minimum" means in a matrix sense, we'd have to have derivatives, etc. – Hilbert-Schmidt. We could notice it's equal to the ML solutions and involve Fisher, Cramér, Rao. We could notice it is minimum mse but NO – that is not enough... it's not true, we are minimizing $(\mathbf{d}-\mathbf{G}\mathbf{m})\cdot\mathbf{C}_\mathrm{d}\cdot(\mathbf{d}-\mathbf{G}\mathbf{m})$.

Definitely too much detail

[5, p. 144]. So now we have the model covariance. Let's just say we want to "minimize" the "matrix" in eq. (3.63) subject to the inverse operator $\mathbf{G}^{-\mathrm{g}}$ being given by the expression eq. (3.60) for purely overdetermined systems, thus $\mathbf{G}^{-\mathrm{g}}\cdot\mathbf{G}=\mathbf{I}$ but let's say there actually is *another* operator $\mathbf{D}^{-g}$ that does the job, by which we mean that both

$$\mathbf{G}^{-\mathrm{g}}\cdot\mathbf{G}=\mathbf{I} \qquad \text{and} \qquad \mathbf{D}^{-\mathrm{g}}\cdot\mathbf{G}=\mathbf{I} \tag{3.64}$$

Let's say there is another $\mathbf{D}^{-\mathrm{g}}$ that does the job, and express

$$\mathbf{D}^{-\mathrm{g}}=\mathbf{G}^{-\mathrm{g}}+(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}}) \tag{3.65}$$

The resultant

$$\begin{aligned}\mathbf{C}_\mathrm{m} = {} & \mathbf{D}^{-\mathrm{g}}\cdot\mathbf{C}_\mathrm{d}\cdot{\mathbf{D}^{-g}}^{\mathrm{T}} \\ = {} & [\mathbf{G}^{-\mathrm{g}}+(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})]\cdot\mathbf{C}_\mathrm{d}\cdot[\mathbf{G}^{-\mathrm{g}}+(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})]^{\mathrm{T}} \\ = {} & \mathbf{G}^{-\mathrm{g}}\cdot\mathbf{C}_\mathrm{d}\cdot\mathbf{G}^{-\mathrm{g}}+(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{C}_\mathrm{d}\cdot(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})^{\mathrm{T}} \\ & +2(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{C}_\mathrm{d}\cdot{\mathbf{G}^{-\mathrm{g}}}^{\mathrm{T}}\end{aligned} \tag{3.66}$$

Note that we've used the transpose symmetry of $\mathbf{C}_\mathrm{d}$. The last term is zero which we see by substitution of the expression $\mathbf{G}^{-\mathrm{g}}$ so

$$\begin{aligned}&(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{C}_\mathrm{d}\cdot\mathbf{C}_\mathrm{d}^{-1}\cdot\mathbf{G}\cdot(\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_\mathrm{d}^{-1}\mathbf{G})^{-1} \\ &\quad = (\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{G}\cdot(\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_\mathrm{d}^{-1}\cdot\mathbf{G})^{-1}\end{aligned} \tag{3.67}$$

Note that we've used the transpose symmetry of $\mathbf{C}_\mathrm{d}^{-1}$. But since we supposed that both $\mathbf{D}^{-\mathrm{g}}\cdot\mathbf{G}=\mathbf{I}$ and $\mathbf{G}^{-\mathrm{g}}\cdot\mathbf{G}=\mathbf{I}$ were solutions, we have

$$(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{G}=0 \tag{3.68}$$

and thus eq. (3.67) is zero. What's left is the two first terms

$$\mathbf{C}_\mathrm{m}=\mathbf{G}^{-\mathrm{g}}\cdot\mathbf{C}_\mathrm{d}\cdot{\mathbf{G}^{-\mathrm{g}}}^{\mathrm{T}}+(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})\cdot\mathbf{C}_\mathrm{d}\cdot(\mathbf{D}^{-\mathrm{g}}-\mathbf{G}^{-\mathrm{g}})^{\mathrm{T}} \tag{3.69}$$

this is positive definite always so smallest and equal to

$$\mathbf{C}_\mathrm{m} = \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{C}_\mathrm{d} \cdot \mathbf{G}^{-\mathrm{g}\mathrm{T}} \tag{3.70}$$

when

$$\mathbf{D}^{-\mathrm{g}} = \mathbf{G}^{-\mathrm{g}} \tag{3.71}$$

So our estimate eq. (3.60) is BLUE, and unique. The fact that we have a minimum doesn't for that matter mean that it's a "good" and "significant" estimate — for that we'll still need the statistical viewpoint of testing.

3.4 The underdetermined problem: first cut

What if neither eq. (3.39) nor eq. (3.60) are good enough? Reasons could be simply zero determinant, low eigenvalues, numerically zero eigenvalues, and the lot. Why is that so bad? Getting it a little wrong on the model or in the data leads to huges effects in the model. Bad!

Example 6: Asking the wrong questions

A dumb example: let $\mathbf{C}_\mathrm{d} = \mathbf{I}$, and try fitting a line through a single point. Clearly, there is no "best" solution in the usual sense, since any "solution" has zero prediction error. Written in matrix form, this becomes

$$\begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} d_1 \end{bmatrix} \tag{3.72}$$

And thus the solution is:

$$\hat{\mathbf{m}} = \left(\begin{bmatrix} 1 \\ x_1 \end{bmatrix} \begin{bmatrix} 1 & x_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \begin{bmatrix} d_1 \end{bmatrix} \tag{3.73}$$

and we need to invert this – whether you remember or not, inverses of a matrix are universally proportional to their determinant

$$(\mathbf{G}^\mathrm{T} \cdot \mathbf{G})^{-1} \propto \frac{1}{\begin{vmatrix} 1 & x_1 \\ x_1 & x_1^2 \end{vmatrix}} = \frac{1}{0} = \infty \tag{3.74}$$

This matrix is "singular", least-squares regression fails. M data, N unknowns. There isn't enough information here – it is an ill-conditioned, ill-posed, *underdetermined* problem.

Example 6: A quick fix

A priori information and the lot. Making sure both parameters are equal (*smoothing*). Making sure both parameters are small (*damping*). Fixing one value. Making sure one is positive. Let's pick the example of damping, and amend the problem as:

$$\begin{bmatrix} 1 & x_1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} d_1 \\ 0 \\ 0 \end{bmatrix} \tag{3.75}$$

More generally, introducing a single scalar multiplier λ,

$$\begin{bmatrix} \mathbf{G} \\ \lambda\mathbf{I} \end{bmatrix} \cdot \mathbf{m} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \tag{3.76}$$

minimize error

$$\left(\begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{G} \\ \lambda\mathbf{I} \end{bmatrix} \cdot \mathbf{m} \right)^2 = \begin{bmatrix} \mathbf{d} - \mathbf{G} \cdot \mathbf{m} \\ \mathbf{0} - \lambda\mathbf{m} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{d} - \mathbf{G} \cdot \mathbf{m} \\ \mathbf{0} - \lambda\mathbf{m} \end{bmatrix} \tag{3.77}$$

$$\phi = (\mathbf{d} - \mathbf{G}\mathbf{m}) \cdot (\mathbf{d} - \mathbf{G}\mathbf{m}) + \lambda^2 \mathbf{m} \cdot \mathbf{m} \tag{3.78}$$

solution is by direct minimization of the gradient

$$\boldsymbol{\nabla}\Phi = -2\mathbf{G}^{\mathrm{T}} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) + 2\lambda^2 \mathbf{m} = \mathbf{0}, \tag{3.79}$$

which leads to

$$\mathbf{0} = -\mathbf{G}^{\mathrm{T}} \cdot \mathbf{d} + \mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} \cdot \mathbf{m} + \lambda^2 \mathbf{m} \tag{3.80}$$

$$= -\mathbf{G}^{\mathrm{T}} \cdot \mathbf{d} + (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} + \lambda^2 \mathbf{I}) \cdot \mathbf{m} \tag{3.81}$$

leading to the general *damped least-squares inverse*

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} + \lambda^2 \mathbf{I}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{d}} \tag{3.82}$$

See Gubbins 115.

Some details that are welcome

Alternative derivation has us work straight from the definition of the inverse of the overdetermined problem:

$$\begin{bmatrix} \mathbf{G} \\ \lambda\mathbf{I} \end{bmatrix} \cdot \mathbf{m} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \tag{3.83}$$

which leads straight to the formulation

$$\mathbf{m} = \left(\begin{bmatrix} \mathbf{G}^{\mathrm{T}} & \lambda \mathbf{I} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{G} \\ \lambda \mathbf{I} \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} \mathbf{G}^{\mathrm{T}} & \lambda \mathbf{I} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \tag{3.84}$$

and thus, when we also once again stick the data error covariance in there, we obtain the solution

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G} + \lambda^2 \mathbf{I} \right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{d}} \tag{3.85}$$

What's a good choice?

Clearly this is a very special form of extra constraints, let's see this in its general form using some weighted model norm as

$$\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G} + \lambda^2 \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W} \right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{d} \tag{3.86}$$

as in [4]. And then it's a small step to "propose" to use the inverse model covariance which must be "a priori" for this also. So this time

$$\phi = (\mathbf{d} - \mathbf{G}\mathbf{m}) \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G}\mathbf{m}) + \mathbf{m} \cdot \mathbf{C}_{\mathrm{m}}^{-1} \cdot \mathbf{m} \tag{3.87}$$

definitely makes sense after identifying

$$\mathbf{C}_{\mathrm{m}}^{-1} = \lambda^2 \mathbf{W}^{\mathrm{T}} \cdot \mathbf{W}, \tag{3.88}$$

and when you're done

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G} + \mathbf{C}_{\mathrm{m}}^{-1} \right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{d}} \tag{3.89}$$

We have to do this — watch the effect on the model uncertainty. If we are using eq. (3.85) the model covariance matrix is an ugly thing that returns *zero* when the data contribute nothing to the solution — when $\mathbf{G} = \mathbf{0}$ if that should be the case. But if we do eq. (3.89) we properly get model uncertainty even if the data don't contribute. Do the full derivation but not here, see [4] p. 119.

Link up with the notion of "variance reduction" (unexplained total variance) and with [3] eq. 2.20

$$\sum_{i=1}^{N} \frac{(\mathrm{obs} - \mathrm{pred})^2}{\mathrm{var}(\mathrm{obs})} \sim \chi^2_{\mathbf{N}-\mathbf{m}} \tag{3.90}$$

The (i, j)th element of the *data covariance matrix*, a dyad,

$$\langle \mathbf{d}\mathbf{d}^{\mathrm{T}} \rangle \tag{3.91}$$

that is, for data of zero mean, is

$$\langle d_i d_j \rangle \tag{3.92}$$

Now let A be

$$(\mathbf{G'}^{\mathrm{T}}\mathbf{G'})^{-1}\mathbf{G'}^{\mathrm{T}} \tag{3.93}$$

in weighted form

then

$$\mathbf{C}_{\mathrm{m}} = \langle \mathbf{m}\mathbf{m}^{\mathrm{T}} \rangle = (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G'})^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}} \cdot \mathbf{G'} \cdot (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G'})^{\mathrm{T}} \tag{3.94}$$

Note: but $\mathbf{C}_{\mathrm{d}}$ of weighted data $= \mathbf{1}$

but $\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G}$ is symmetric if $\mathbf{C}_{\mathrm{d}} = \mathbf{1}$, so

$$\mathbf{C}_{\mathrm{m}} = (\mathbf{G'}^{\mathrm{T}}\mathbf{G'})^{-1}(\mathbf{G'}^{\mathrm{T}}\mathbf{G'})(\mathbf{G'}^{\mathrm{T}}\mathbf{G'})^{-1} \tag{3.95}$$

$$\mathbf{C}_{\mathrm{m}} = (\mathbf{G'}^{\mathsf{T}'}\mathbf{G})^{-1} \tag{3.96}$$

Bring the σ^2 back, for uncorrelated data

$$\mathbf{C}_{\mathrm{m}} = \sigma^2(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G})^{-1} \tag{3.97}$$

typically not diagonal, thus: may construct confidence interval. Perhaps diagonalize. Perhaps perform constrained minimization and thus eliminate model parameters.

Trade-off. Illustration level?

[Example why χ^2 shouldn't be zero? Can always make it zero by fitting an $N - 1$ polynomial through data (Lagrange's theorem)].

Once you have a fit, you can interpret χ^2. Having a minimum doesn't mean it is *good*.

For this we must go to the Bayesian form. But first we do another bit.

3.5 Constrained minimization

Make pictures of fitting the ping-pong ball going forcing through a certain point. Or to have the average of some parameters be zero. Or some other such motivational thing. An example could be the constrained fitting of an otherwise overdetermined problem.

General case, graphical interpretation

Let us consider the minimization of a certain scalar misfit function Φ, subject to the constraint that a certain *other* scalar Γ or vector $\mathbf{\Gamma}$ function of the unknown

Fig. 3.1. Make the picture now! Contours of Φ, contours of Γ. Make something real.

parameter set $\mathbf{m}$ is also satisfied. We want to find the solution $\hat{\mathbf{m}}$ that achieves a *minimum* $\Phi(\hat{\mathbf{m}})$ while satisfying $\Gamma(\hat{\mathbf{m}})$ or $\mathbf{\Gamma}(\hat{\mathbf{m}})$ *exactly*, in the sense

$$\min \Phi(\hat{\mathbf{m}}) \quad \text{subject to} \quad \Gamma(\hat{\mathbf{m}}) = 0, \tag{3.98}$$

$$\min \Phi(\hat{\mathbf{m}}) \quad \text{subject to} \quad \mathbf{\Gamma}(\hat{\mathbf{m}}) = \mathbf{0}. \tag{3.99}$$

This all different from inequality-constrained minimization. Or from the minimization of two quadratics, etc. But still.

The intuitive explanation and picture (which is easiest to draw graphically for eq. 3.98) is that where this happens, the gradients of both functions with respect to the parameters estimated, $\hat{\mathbf{m}}$, are *anti-parallel*,

$$\boldsymbol{\nabla}_{\hat{\mathbf{m}}} \Phi = -\lambda \boldsymbol{\nabla}_{\hat{\mathbf{m}}} \Gamma, \tag{3.100}$$

Go as far as you can in minimizing Φ while staying on the contour of $\Gamma = v$. If these constraints are satisfied, the minimum of $\phi + \lambda\Gamma$ is the minimum of ϕ. So you add a new equation to the old one, but you impose that the addition is zero. What is the connection? See [4]: may as well pick a λ as [6] on p 56. Or, think of a valley of minimum of ϕ and pick on it the one that satisfies the constraints.

Another way to understand this has us consider the need to minimize "new" penalty functions,

$$\Phi + \lambda\Gamma \qquad \text{and} \qquad \Phi + \boldsymbol{\lambda} \cdot \mathbf{\Gamma}, \tag{3.101}$$

with the "lambda(s)" one or more **Lagrange multiplier**(s). Thus we now require, in the case of eq. (3.98), that

$$\boldsymbol{\nabla}_{\hat{\mathbf{m}}} \Phi + \lambda \boldsymbol{\nabla}_{\hat{\mathbf{m}}} \Gamma = \mathbf{0}, \tag{3.102}$$

where λ is a scalar Lagrange multiplier, and in the case of eq. (3.99),

$$\boldsymbol{\nabla}_{\hat{\mathbf{m}}} \Phi + \boldsymbol{\nabla}_{\hat{\mathbf{m}}} \mathbf{\Gamma} \cdot \boldsymbol{\lambda} = \mathbf{0}, \tag{3.103}$$

with $\boldsymbol{\lambda}$ a vector Lagrange multiplier. In the latter case, the Lagrange vector as many entries as you want to impose additional constraints. Write out in index notation and think about gamma also being a linear operator (see my own notes in [4], p. 233). From now on we drop the subscript from $\boldsymbol{\nabla}$ as we shall never get confused.

Special case, linear constraint equalities

This is a **quadratic** minimization plus a **linear** constraint. Should be able to formalize this in matrix form. Also see Hansen about double quadratics. That's where the "quick fix" comes in, presumably. I do this in class with an eye on eq. (3.78) which should have made intuitive sense by that time.

A (linear) *scalar* constraint on the model parameters is of the general form

$$\Gamma(\hat{\mathbf{m}}) = \mathbf{f} \cdot \hat{\mathbf{m}} - v = 0, \tag{3.104}$$

for a certain vector $\mathbf{f}$ and some value v, while a (linear) *vector* constraint, for some matrix $\mathbf{F}$ and a certain vector $\mathbf{v}$, is of the general form,

$$\boldsymbol{\Gamma}(\hat{\mathbf{m}}) = \mathbf{F} \cdot \hat{\mathbf{m}} - \mathbf{v} = \mathbf{0}. \tag{3.105}$$

Perhaps do the following two examples twice, once explicitly and once for the Lagrange matrix equations if we can make it look nice.

Top left: taking the gradient of the quadratic, which is going to be some form of $\mathbf{m} \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \mathbf{m}$ thus $\mathbf{W}^T \cdot \mathbf{W}$. Bottom right is zero. Remainder is the constraints, i.e. the vectors themselves.

Let us now turn our attention to the problem of finding the minimizer $\hat{\mathbf{m}}$ of the quadratic penalty function Φ given by

$$\min \Phi = (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}), \tag{3.106}$$

supplemented by constraints of the form (3.104) or (3.105). Note that this situation is different from the purely underdetermined case, we are still saving this for later. There, we found that we were able to obtain a perfect data fit, $\Phi = 0$, and we attempted to select from a large class of models, all of which explained the data without error, by minimizing some other quantity, e.g. the Euclidean length of the model vector.

Here, we solve

$$\mathbf{0} = \boldsymbol{\nabla}_{\hat{\mathbf{m}}} \left[(\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \right] + \lambda \boldsymbol{\nabla}_{\hat{\mathbf{m}}} (\mathbf{f} \cdot \hat{\mathbf{m}}) \tag{3.107}$$

$$= -2\,\mathbf{G}^{\mathrm{T}} \cdot (\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) + \lambda \mathbf{f}. \tag{3.108}$$

In other words, satisfying eq. (3.108)

$$\mathbf{0} = -2\,\mathbf{G}^{\mathrm{T}} \cdot \mathbf{d} + 2\,\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} \cdot \hat{\mathbf{m}} + \lambda \mathbf{f} \tag{3.109}$$

simultaneously with the initial constraint (3.104)

$$v = \mathbf{f} \cdot \hat{\mathbf{m}} \tag{3.110}$$

can be written in matrix form (the transposes on vectors start to matter!) as

$$\begin{bmatrix} \mathbf{G}^{\mathrm{T}} \cdot \mathbf{d} \\ v \end{bmatrix} = \begin{pmatrix} \mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} & \mathbf{f}^{\mathrm{T}} \\ \mathbf{f} & 0 \end{pmatrix} \begin{bmatrix} \hat{\mathbf{m}} \\ \lambda/2 \end{bmatrix}. \tag{3.111}$$

Top left is the equivalent of the original problem that you minimize, the bottom right is what you try to fix.

Example 7: Constrained fitting of a straight line

$$\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_M \end{bmatrix} \qquad \text{and} \qquad \mathbf{m} = \begin{bmatrix} b \\ a \end{bmatrix} \tag{3.112}$$

and let the constraint be that the line go through the point (x^*, y^*), as in

$$\begin{bmatrix} 1 & x^* \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = y^* \tag{3.113}$$

Lagrange equations from eq. (3.111)

$$\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G} = \begin{pmatrix} M & \sum_{i=1}^{M} x_i \\ \sum_{i=1}^{M} x_i & \sum_{i=1}^{M} x_i^2 \end{pmatrix} \tag{3.114}$$

is now square and can be inverted [conditions?] who knows invert this, instead.

3.6 The underdetermined problem: the pure form

Let's do a different example — suppose we can fit the data exactly, but we need to choose a particular solution. We'd have to satisfy eq. (3.99) where

$$\min \Phi = \hat{\mathbf{m}} \cdot \hat{\mathbf{m}} \quad \text{subject to} \quad \boldsymbol{\Gamma} = \mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}} = \mathbf{0}. \tag{3.115}$$

What we did before we knew constraints

Thus we solve

$$\mathbf{0} = \boldsymbol{\nabla}(\hat{\mathbf{m}} \cdot \hat{\mathbf{m}}) + \boldsymbol{\nabla}(\mathbf{d} - \mathbf{G} \cdot \hat{\mathbf{m}}) \cdot \boldsymbol{\lambda} \tag{3.116}$$

$$= 2\hat{\mathbf{m}} - \mathbf{G}^{\mathrm{T}} \cdot \boldsymbol{\lambda}, \tag{3.117}$$

by in other words having

$$\hat{\mathbf{m}} = -\tfrac{1}{2}\, \mathbf{G}^{\mathrm{T}} \cdot \boldsymbol{\lambda} \tag{3.118}$$

but of course we know from eq. (3.115) that $\mathbf{G} \cdot \hat{\mathbf{m}} = \mathbf{d}$ and thus,

$$\mathbf{d} = -\tfrac{1}{2}\, \mathbf{G} \cdot \mathbf{G}^{\mathrm{T}} \cdot \boldsymbol{\lambda} \tag{3.119}$$

and thus we can determine the vector of multipliers as

$$\boldsymbol{\lambda} = -2\left(\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}}\right)^{-1}\cdot\mathbf{d}, \tag{3.120}$$

which we plug back into eq. (3.118) to yield

$$\boxed{\hat{\mathbf{m}} = \mathbf{G}^{\mathrm{T}}\cdot(\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}})^{-1}\cdot\mathbf{d}} \tag{3.121}$$

By construction, eq. (3.121) is the *minimum-length* solution of all those that satisfy the data *exactly*. Now say something about the null-space interpretation?

$$\mathbf{G}^{-\mathrm{g}} = \mathbf{G}^{\mathrm{T}}\cdot(\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}})^{-1} \tag{3.122}$$

is a “right-inverse” of the design matrix. Correct. See SV’s thesis on page 13 for the SVD version of this.

In the formalism of minimization with constraints

Purely underdetermined problem – you can fit te data exactly but you need to minimize the solution length. We’ve already done this, we’re going to do this again in the slighly more general framework.

$$\min\Phi = \hat{\mathbf{m}}\cdot\hat{\mathbf{m}} \quad \text{subject to} \quad \boldsymbol{\Gamma} = \mathbf{d}-\mathbf{G}\cdot\hat{\mathbf{m}} = \mathbf{0}. \tag{3.123}$$

Indeed

$$\hat{\mathbf{m}}\cdot\hat{\mathbf{m}} = (\mathbf{I}\cdot\hat{\mathbf{m}})\cdot(\mathbf{I}\cdot\hat{\mathbf{m}}) \tag{3.124}$$

minimize

$$\hat{\mathbf{m}}\cdot\hat{\mathbf{m}} = \hat{\mathbf{m}}^2 \rightarrow^{\text{equiv}} \mathbf{I}\cdot\mathbf{m} = \mathbf{0} \tag{3.125}$$

least squares subject to

$$\mathbf{G}\cdot\hat{\mathbf{m}} = \mathbf{d} \rightarrow^{\text{equiv}} \mathbf{F}\cdot\mathbf{m} = \mathbf{d} \tag{3.126}$$

See how the roles are switched! The equations of Lagrange are

$$\begin{pmatrix} \mathbf{I} & \mathbf{G}^{\mathrm{T}} \\ \mathbf{G} & \mathbf{0} \end{pmatrix}\begin{bmatrix} \hat{\mathbf{m}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{d} \end{bmatrix} \tag{3.127}$$

And then eq. (3.121) is the least-squares solution to the problem which is just a square inversion leading to

$$\begin{bmatrix} \hat{\mathbf{m}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{G}^{\mathrm{T}}\cdot(\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}})^{-1} \\ (\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}})^{-1}\cdot\mathbf{G} & -(\mathbf{G}\cdot\mathbf{G}^{\mathrm{T}})^{-1} \end{pmatrix}\begin{bmatrix} \mathbf{0} \\ \mathbf{d} \end{bmatrix} \tag{3.128}$$

as you can check by substituting.

Purely overdetermined problem—you cannot find an exact solution but you can minimize the residuals. Minimize $(\mathbf{G} \cdot \hat{\mathbf{m}} - \mathbf{d})^2$ subject to what? The norm being a certain value? Need to work this out in more general detail. The board lecture was good. Maybe ask Kyle for his notes, he seems to have decent handwriting.

WRAPUP: how we went from data norm to model norm to mixtures thereof.

More examples of ad hoc stuff

The tomographic inverse problem

slowness s_i, d, G, m, $m_2 =?$, $m_1 =?$

underdetermined: too little information, supple a priori constraints

overdetermined: to much information, look for best solution

[6] p. 51: average is overdetermined, individual is underdetermined, same length. A constraint is just another (set of) equation(s) satisfied by the model parameters that have nothing to do with the actual data.

[6] p. 55: i.e. the mean of the model parameters has a value known a priori, "value v"

$$\frac{\mathbf{d}}{v} = \frac{\mathbf{G}}{1/N1/N1/N1/N1/N1/N1/N} \quad m \tag{3.129}$$

or a specific data point is known

$$\frac{\mathbf{d}}{v} = \frac{\mathbf{G}}{000100} \quad m \tag{3.130}$$

or reducing the roughness of the solution

$$\begin{array}{c} \underline{\mathbf{d}} \\ 0 \\ \downarrow \end{array} = \begin{array}{c} \mathbf{G} \\ \hline \begin{array}{cccccc} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & & \searrow & \end{array} \end{array} \quad m \tag{3.131}$$

It is clear that choices have to be made as to how one has to weight the information contained in the data versus the information supplied *a priori*.

So we need to talk about a trade-off curve.

3.7 Model and data resolution matrices

$$\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{G} = \mathbf{I} \tag{3.132}$$

purely overdetermined, the *left inverse*

$$(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G})^{-1} \cdot \mathbf{G}^{\mathrm{T}} = \mathbf{G}^{-\mathrm{g}} \tag{3.133}$$

$$\mathbf{G} \cdot \mathbf{G}^{-\mathrm{g}} = \mathbf{I} \tag{3.134}$$

purely underdetermined, the *right inverse*

$$\mathbf{G}^{\mathrm{T}} \cdot (\mathbf{G} \cdot \mathbf{G}^{\mathrm{T}})^{-1} = \mathbf{G}^{-\mathrm{g}} \tag{3.135}$$

in practice, the problem is neither purely under or overdetermined, and $\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{G}$ is the model resolution matrix of whatever you pick to solve by

$$\hat{\mathbf{m}} = \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d} \tag{3.136}$$

that necessitates the discussion

$\mathbf{G}^{-\mathrm{g}} \cdot \mathbf{G}$ is the **model resolution matrix**. It only equals the identity for the purely overdetermined case. Similarly,

$$\hat{\mathbf{d}} = \mathbf{G} \cdot \hat{\mathbf{m}} = \mathbf{G} \cdot \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{d} \tag{3.137}$$

$\mathbf{G} \cdot \mathbf{G}^{-\mathrm{g}}$ is the **data resolution matrix**, see [6] = I for purely underdetermined. Maybe the model $\mathbf{G}$ matrix has a character in between over/underdetermined cases? $\mathbf{G}$ embodies model and experimental geometry.

SEE NOTES 10/23/2008

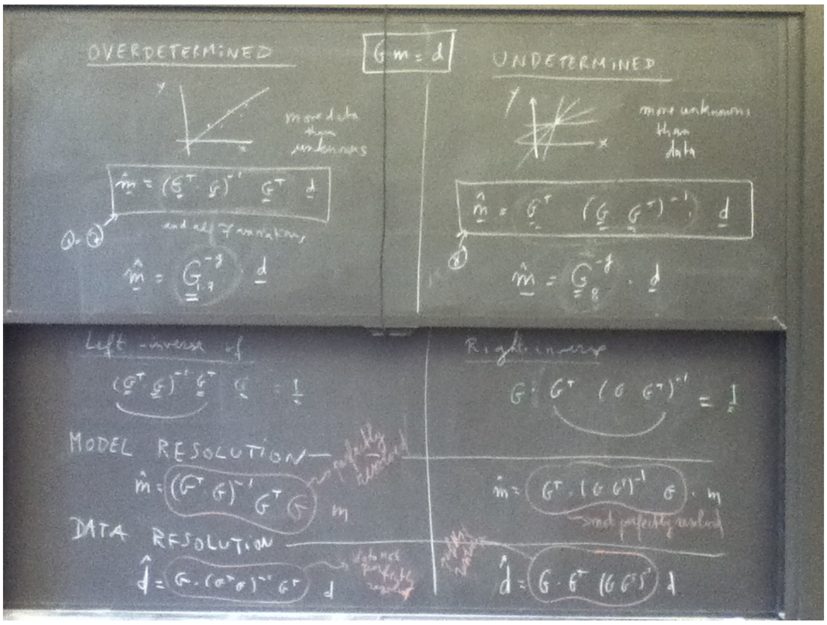

OVERDETERMINED
Gm = d
UNDETERMINED
more data than unknowns
more unknowns than data
Left inverse of
MODEL RESOLUTION
DATA RESOLUTION

Fig. 3.2. Wuut

3.8 Mixed-determined systems

What if it's a mixture — this is when eigenvalue expansions and the SVD should be discussed. Lay and Wallace discussion? Diagrams a la Myres?

Before or after next should talk about Landweber method. Malta conference: Landweber is supposed to converge to the Moore-Penrose Inverse, illustrate this with some really simple examples, and look up the key paper positing this.

3.9 Inversion of a nonlinear function / Geiger's method

$$d = T(m_i, z_j) \tag{3.138}$$

$$T(m^{k+1}) = T(m^k) + \left.\frac{\partial T}{\partial m_i}\right|_{m^k} \cdot (m_i^{k+1} - m_i^k) \tag{3.139}$$

$$d^{obs} - d^{pred} = \left.\frac{\partial T}{\partial m_i}\right|_{m^{pred}} \cdot \Delta m \tag{3.140}$$

$$\Delta d = \boldsymbol{\nabla} T \cdot \Delta m \tag{3.141}$$

same linear problem as before - guess; invert for update until convergence, i.e until Δd, the misfit, is 0, or Δm, the model update, is 0 as well.

Earthquake location problem.

Exercise: earthquake location and double difference?

method 1

method 2 - Geiger's method for earthquake location - an example of non-linear inversion by first-order approximation

get course notes- lay and wallace, cahier, any other notes?

Method 1: good discussion on p22 of Lay and Wallace.

(i) guess a solution x,yz,t
(ii) calculate arival times
(iii) calculate distances, azimuths
(iv) calculate time/distance residual
(v) plot residual distance vs azimuth
(vi) shift time by average residual
(vii) shift location along azimuth of maximum by distance difference
(viii) repeat

error: when residuals are small: say Δt

What method 1 is, is a series of forward modeling exercises with adjustments that supposedly go in the direction of the gradient of the misfit function. Is this really true?

$$t_{\text{pred}} = f(x, y, z, v) \tag{3.142}$$

f is nonlinear. v is velocity

$$t_i = t + \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}}{v} \tag{3.143}$$

can't do it? Linearize!

$$\mathbf{d} = T(\mathbf{m}) \tag{3.144}$$

guess $\mathbf{m}_0$, then $\mathbf{d}_0 = T(\mathbf{m}_0)$

Taylor series to first order:

$$T(\mathbf{m}) = T(\mathbf{m}_0) + \left.\frac{\partial T}{\partial \mathbf{m}}\right|_{\mathbf{m}_0} \cdot \partial \mathbf{m} \tag{3.145}$$

$$= T(\mathbf{m}_0) + \frac{\partial T}{\partial x}dx + \frac{\partial T}{\partial y}dy + \frac{\partial T}{\partial z}dy + \frac{\partial T}{\partial t}dt \tag{3.146}$$

In summary,

$$(\mathbf{d} - \mathbf{d}_0) = \boldsymbol{\nabla}\mathbf{T} \cdot \Delta\mathbf{m}, \tag{3.147}$$

whereby $\boldsymbol{\nabla}\mathbf{T}$ is the Jacobian, each row is with respect to a particular rotation. Iterate. Don't confuse derivative of linear t which is 1 and derivative of nonlinear t_i which we call t-prime.

So now: write out the derivatives of the function T_i that takes an earthquake location x, y, z, t and calculates its travel time to station i.

$$(t_i - t)v = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} \tag{3.148}$$

$$\left.\frac{\partial t_i}{\partial t}\right|_{x0,y0,z0,t0} = 1 \tag{3.149}$$

$$\left.\frac{\partial t_i}{\partial x}\right|_{x0,y0,z0,t0} = -\frac{(x_i - x)}{v^2(t_i - t)} \tag{3.150}$$

$$\left.\frac{\partial t_i}{\partial y}\right|_{x0,y0,z0,t0} = -\frac{(y_i - y)}{v^2(t_i - t)} \tag{3.151}$$

$$\left.\frac{\partial t_i}{\partial z}\right|_{x0,y0,z0,t0} = -\frac{(z_i - z)}{v^2(t_i - t)} \tag{3.152}$$

so now these are ready to be stuck into a matrix for inversion.

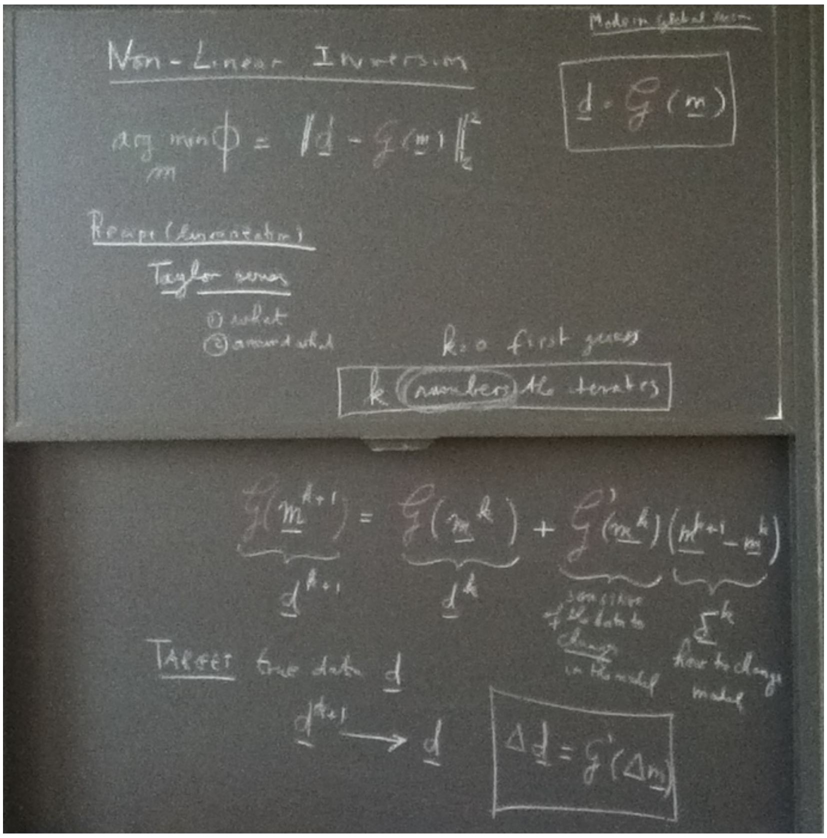


Fig. 3.3. Geiger 1

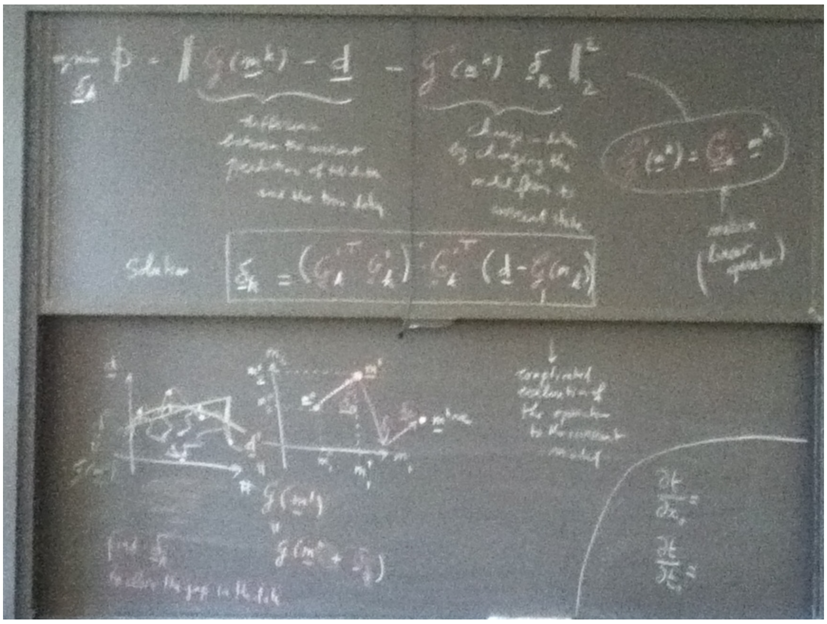


Fig. 3.4. Geiger 2

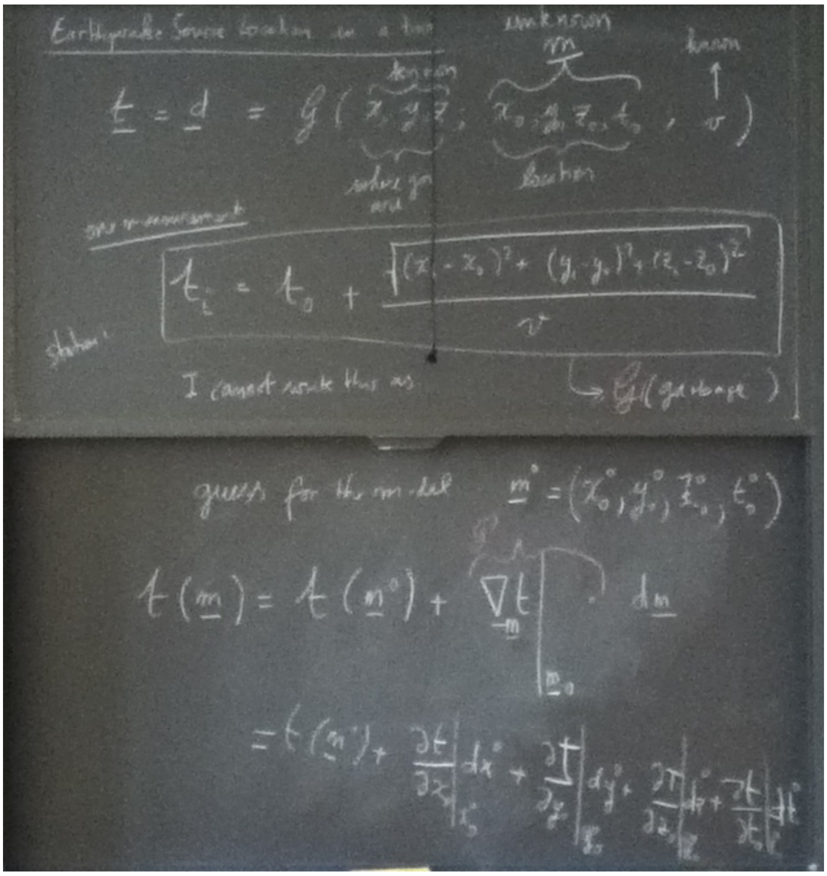


Fig. 3.5. Geiger 3

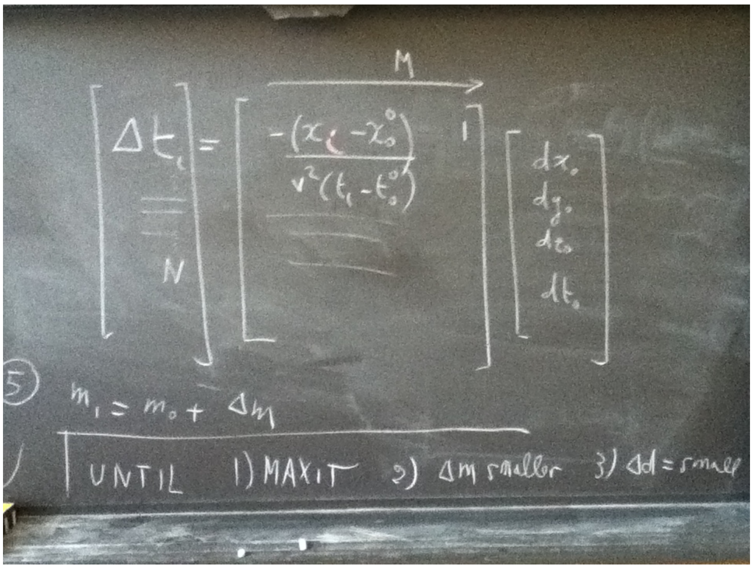


Fig. 3.6. Geiger 3

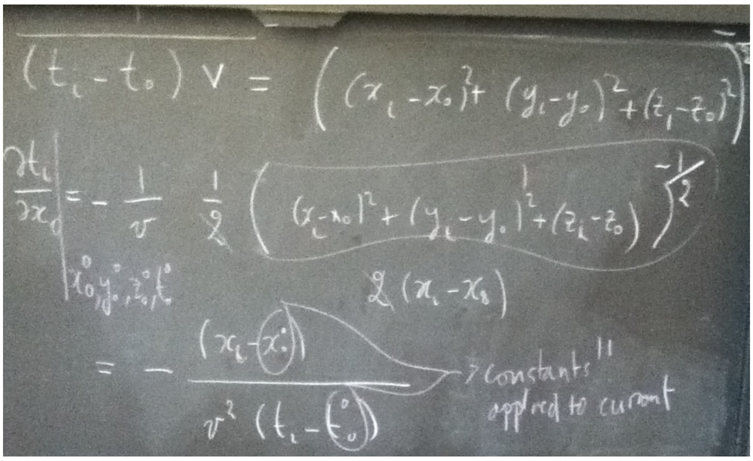


Fig. 3.7. Geiger 4

4
Distributional approach

In this section we want to begin with Chapter X and use Chapter Z to find Chapter Y again, in the same order. Number the solutions with subscripts, 1, 2, ..., 8 and end up with a little table perhaps.

4.1 Likelihoods

See Sivia in the very beginning for some notation etc. Begin with Bayes.

Independent identically distributed data:

$$p(\mathbf{d}|\mathbf{m}) = p(d_1|\mathbf{m}) \cdots p(d_N|\mathbf{m}) \tag{4.1}$$

In practice, we *have the data*, we want the *model* (another "leap of faith")

$$\mathcal{L}(\mathbf{m}|\mathbf{d}) = p(\mathbf{d}|\mathbf{m}) \tag{4.2}$$

Let's assume an individual pdf looks like

$$p(d_i|\mathbf{m}) = \frac{1}{\sigma_i\sqrt{2\pi}}\, e^{-\frac{\left(d_i - \sum_j^M G_{ij}m_j\right)^2}{2\sigma_i^2}} = \frac{1}{\sigma_i\sqrt{2\pi}}\, e^{-\frac{(d_i - \mathbf{G}\cdot\mathbf{m})^2}{2\sigma_i^2}} \tag{4.3}$$

Why? You hope that $\sum_j G_{ij}m_j$ explains the data in the mean, in the true sense of the word, it is what we *expect*!

$$\mathcal{L}(\mathbf{m}|\mathbf{d}) = p(\mathbf{d}|\mathbf{m}) = \frac{1}{(2\pi)^{N/2}}\, e^{-\sum_i^N \frac{\left(d_i - \sum_j^M G_{ij}m_j\right)^2}{2\sigma_i^2}} \tag{4.4}$$

not the value, but the argument

$$\text{argmax}(\log\, \mathcal{L}) = \text{argmin} \sum_{i=1}^{N} \frac{\left(d_i - \sum_j G_{ij}m_j\right)^2}{2\sigma_i^2} \tag{4.5}$$

Now for data that have covariance the likelihood is

$$\mathcal{L}(\mathbf{m}|\mathbf{d}) \sim \exp\left[-(\mathbf{d}-\mathbf{G}\cdot\mathbf{m})\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot(\mathbf{d}-\mathbf{G}\cdot\mathbf{m})\right] \tag{4.6}$$

is just a multivariate Gaussian. Data can have correlated error, i.e. non-diagonal $\mathbf{C}_{\mathrm{d}}$.

See variance reduction etc before.

4.2 Two approaches to solving inverse problems

$$\mathbf{G}\cdot\mathbf{m} = \mathbf{d} \tag{4.7}$$

minimize the residual in the mean squared sense:

$$\phi = ||\mathbf{d}-\mathbf{G}\cdot\hat{\mathbf{m}}||_2^2 = \mathbf{e}\cdot\mathbf{e} \tag{4.8}$$

where the model prediction error, or the residual is given by.

$$\mathbf{e} = \mathbf{d}-\mathbf{G}\cdot\hat{\mathbf{m}} \tag{4.9}$$

Minimization is the easiest in the index notation. The repeated j speaks for itself, the i is repeated due to the square.

$$\phi = (d_i - G_{ij}\hat{m}_j)^2 \tag{4.10}$$

Note: Two sums implied! Over i and over j. Take the derivative with respect to a generic model parameter m_k

$$\frac{\partial\phi}{\partial m_k} = 2(d_i - G_{ij}\hat{m}_j)(-G_{ik}) = 0 \tag{4.11}$$

Convert back to vector/matrix notation:

$$\boldsymbol{\nabla}\Phi\mathbf{G}^{\mathrm{T}}\cdot(\mathbf{d}-\mathbf{G}\cdot\hat{\mathbf{m}}) = \mathbf{0} \tag{4.12}$$

$$\mathbf{G}^{\mathrm{T}}\cdot\mathbf{d} = \mathbf{G}^{\mathrm{T}}\cdot\mathbf{G}\cdot\hat{\mathbf{m}} \tag{4.13}$$

and the *least-squares* estimate is again:

$$\boxed{\hat{\mathbf{m}} = (\mathbf{G}^{\mathrm{T}}\cdot\mathbf{G})^{-1}\cdot\mathbf{G}^{\mathrm{T}}\cdot\mathbf{d}} \tag{4.14}$$

as eq. (**??**).

This estimate is identical to the *maximum likelihood estimate* of the data (and the forward operator write it) *rescaled to unit variance*. In other words the estimated solution $\hat{\mathbf{m}}$ maximizes

$$\mathcal{L}(\mathbf{m}|\mathbf{d}) \sim \exp\left(-[\mathbf{d}-\mathbf{G}\cdot\mathbf{m}]\cdot[\mathbf{d}-\mathbf{G}\cdot\mathbf{m}]\right) \tag{4.15}$$

which is appropriate for independent data. Now with a real data covariance,

$$\langle \mathbf{d}\,\mathbf{d}^{\mathrm{T}} \rangle = \mathbf{C}_{\mathrm{d}} \tag{4.16}$$

Maximize $\mathcal{L}$, equal to minimize

$$(\mathbf{d} - \mathbf{G}\cdot\mathbf{m})\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot(\mathbf{d} - \mathbf{G}\cdot\mathbf{m}) \tag{4.17}$$

We have *no* prior info on the model, we leave that for later.
SVD must bring up *total least squares?*
differentiate $(\mathbf{d} - \mathbf{G}\cdot\mathbf{m})\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot(\mathbf{d} - \mathbf{G}\cdot\mathbf{m})$ convert to m and we get

$$\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot(\mathbf{d} - \mathbf{G}\cdot\mathbf{m}) \tag{4.18}$$

Why the transpose?

$$\frac{\partial(G_{ij}m_j)}{\partial m_k} = G_{ik} + (\mathbf{d} - \mathbf{G}\cdot\mathbf{m})\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{G} = 0 \tag{4.19}$$

do this for every k but keep the sum over i which therefore requires the transpose. Thus

$$\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot(\mathbf{d} - \mathbf{G}\cdot\mathbf{m}) = \mathbf{0} \tag{4.20}$$

also, clearly $\mathbf{C}_{\mathrm{d}} = \mathbf{C}_{\mathrm{d}}^{\mathrm{T}}$ and for vectors this doesn't matter. Or else

$$\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{d} = \mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{G}\cdot\mathbf{m} \tag{4.21}$$

and thus the solution is again given by

$$\boxed{\hat{\mathbf{m}} = \left(\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{G}\right)^{-1}\cdot\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{d}} \tag{4.22}$$

as eq. (**??**), where we presumed the linearity of the forward model.

Think about noise a bit more formally

$$\mathbf{d} = \mathbf{G}\cdot\mathbf{m} + \mathbf{n} \tag{4.23}$$

General case ABT 2.101. What's the bias of this selection?

$$\langle \hat{\mathbf{m}} \rangle = (\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{G})^{-1}\cdot(\mathbf{G}^{\mathrm{T}}\cdot\mathbf{C}_{\mathrm{d}}^{-1}\cdot\mathbf{G})\cdot\mathbf{m} \tag{4.24}$$

$\langle \hat{\mathbf{m}} \rangle = \mathbf{m}$ immediately falls out, but let's also do the case where uncorrelated data

$$\mathbf{C}_{\mathrm{d}} = \sigma^2\mathbf{I} \qquad \text{and} \qquad \mathbf{C}_{\mathrm{d}}^{-1} = \sigma^{-2}\mathbf{I} \tag{4.25}$$

In other words,

$$\langle \hat{\mathbf{m}} \rangle = \sigma^2 (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G})^{-1} \cdot (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{G}) \cdot \mathbf{m}\, \sigma^{-2} \tag{4.26}$$

$$= \mathbf{m}. \tag{4.27}$$

Unbiased since it was minimum RMSE already RMSE = var + bias^2

What is the variance?

$$\mathbf{C}_{\mathrm{m}} = \mathbf{G}^{-\mathrm{g}} \cdot \mathbf{C}_{\mathrm{d}} \cdot (\mathbf{G}^{-\mathrm{g}})^{\mathrm{T}} \tag{4.28}$$

so it's also minimum variance in the sense that

$$\langle (\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^{\mathrm{T}} \rangle \tag{4.29}$$

is minimized. Refer to above whole section.

4.3 Full-blown Bayesian form

Up until now, we've provided "mock" model constraints. I suppose it makes a lot more sense to specify a-priori model covariance matrix... Let's not anymore try to maximize the likelihood of observing the data given the postulated model parameters. We postulate an *a priori* probability density function on the model parameters of the Gaussian kind: we have

$$p(\mathbf{m}) \sim \exp\left[-(\mathbf{m} - \mathbf{m}_0) \cdot \mathbf{C}_{\mathrm{m}}^{-1} \cdot (\mathbf{m} - \mathbf{m}_0)\right] \tag{4.30}$$

we had for the conditional probability on the data the Gaussian quadratic form

$$P(\mathbf{d}|\mathbf{m}) \sim \exp\left[-(\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m})\right] \tag{4.31}$$

We use Bayes' theorem and we optimize the probability of the, now stochastic, model given the data.

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})} \tag{4.32}$$

which is proportional to—let the initial model be $\mathbf{m}_0 = \mathbf{0}$:

$$p(\mathbf{m}|\mathbf{d}) \sim \exp\left[(\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) - \mathbf{m} \cdot \mathbf{C}_{\mathrm{m}}^{-1} \cdot \mathbf{m}\right] \tag{4.33}$$

and the maximum-likelihood solution now given by minimizing the exponent. Same procedure as ever: maximize the likelihood function is equivalent to minimizing the penalty function

$$\phi = (\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) + \mathbf{m} \cdot \mathbf{C}_{\mathrm{m}}^{-1} \cdot \mathbf{m} \tag{4.34}$$

So take the derivative with respect to $\hat{\mathbf{m}}$. In index notation, with respect to a generic model parameter

$$\frac{\partial \phi}{\partial m_k} = \frac{\partial}{\partial m_k}\left[(d_i - G_{ij}m_j)(C_d^{-1})_{ie}(d_e - G_{en}m_n) - m_i C_d m_j\right] \tag{4.35}$$

$$0 = (d_i - G_{ij}m_j)(C_d)_{ie} - (G_{lk} - m_i C_{mik}^{-1}) \tag{4.36}$$

or back in vector form:

$$\mathbf{0} = \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{d} - \mathbf{G} \cdot \mathbf{m}) + \mathbf{C}_{\mathrm{m}}^{-1} \cdot \mathbf{m} \tag{4.37}$$

and the solution is once again

$$\boxed{\hat{\mathbf{m}} = (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G} + \mathbf{C}_{\mathrm{m}}^{-1})^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{d}} \tag{4.38}$$

which is the solution to

$$\begin{bmatrix} \mathbf{C}_{\mathrm{d}}^{-1/2} \cdot \mathbf{G} \\ \mathbf{C}_{\mathrm{m}}^{-1/2} \end{bmatrix} \cdot \mathbf{m} = \begin{bmatrix} \mathbf{C}_{\mathrm{d}}^{-1/2} \cdot \mathbf{d} \\ \mathbf{0} \end{bmatrix}. \tag{4.39}$$

We should try to write all of our solution in this familiar form (like we had for the regularized and constrained cases already) so that we can use the standard algorithms to solve them. Sometimes one will find this relation instead

$$\boxed{\hat{\mathbf{m}} = \mathbf{C}_{\mathrm{m}} \cdot \mathbf{G}^{\mathrm{T}} \cdot (\mathbf{C}_{\mathrm{d}} + \mathbf{G} \cdot \mathbf{C}_{\mathrm{m}} \cdot \mathbf{G}^{\mathrm{T}})^{-1} \cdot \mathbf{d}} \tag{4.40}$$

which is identical; the equivalence is through

$$\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot (\mathbf{G} \cdot \mathbf{C}_{\mathrm{m}} \cdot \mathbf{G}^{\mathrm{T}} + \mathbf{C}_{\mathrm{d}}) = (\mathbf{G}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot \mathbf{G} + \mathbf{C}_{\mathrm{m}}^{-1}) \cdot \mathbf{C}_{\mathrm{m}} \cdot \mathbf{G}^{\mathrm{T}} \tag{4.41}$$

Bayesian approach vs frequentist/Tikhonov approach. Now must say something about the covariance of the result. Gubbins on the curvature of the likelihood function.

Next - maybe Gauss-Newton methods? Maximum likelihood? Nonlinear methods using linearization. Lab: Earthquake location. Landweber iteration.

Board scheme: optimization to testing to error bars is the classic way. This was the new way. Evaluation is now in terms of "resolution" or "information added". So resolution is back and the choice of prior needs to be discussed.

4.4 Tarantolia

Forget inversion, rather do the mapping of probabilities. Apply to the quake location problem, see the SIAM book. Basically, it's about evaluating

$$p(\mathbf{m}|\mathbf{d}) \sim p(\mathbf{d}|\mathbf{m})\, p(\mathbf{m}) \tag{4.42}$$

a possibly nonlinear function of the model parameters. E.g. for the Gaussian case this would be using the notation $\mathcal{G}$ for some functional, now here is the likelihood function

$$p(\mathbf{d}|\mathbf{m}) \sim \exp\left([\mathbf{d} - \mathcal{G}(\mathbf{m})] \cdot \mathbf{C}_{\mathrm{d}}^{-1} \cdot [\mathbf{d} - \mathcal{G}(\mathbf{m})]\right) \tag{4.43}$$

Now, Metropolis Hastings is a way of not having to do this evaluation analytically, but rather to sample the distribution directly. See the last chapter in the Tarantola web book and [7].

Part III

Time-series Analysis

5
Fourier Analysis

What's a time series? Never mind "time": it's the "series" part that is important. What is meant is that, in contrast to a bunch of numbers drawn at random out of a hat (the domain of most of what preceded this chapter), from now on, the *order* in which they are drawn will *matters*. More than that: the ordering will *mean* something. Whether the axis on which you order represents space, or time, or anything else, we will talk about **time series** very generally. They are strings of data, "sequences" of values that, by their values *and* by their ordering, carry *information*—which we want to unearth.

It requires hardly any imagination to consider the daily weather as an archtypical time series, and the earth being rotund and rotating, *sine* and *cosine* functions as the natural functions with which to describe the temporal behavior of fundamental meterological variables such as, e.g., the outside temperature, will need no special introduction.

5.1 Pure harmonics

A pure harmonic, e.g., the real-valued $\sin(2\pi\phi)$ and $\cos(2\pi\phi)$ or the complex-valued $\exp(i2\pi\phi)$, is an oscillatory, periodic, function. As I write it, a pure harmonic associates a dimensionless ratio ϕ with a number between -1 and $+1$.

Any dimensionless ratio, depending on your application, will fit in this framework. Should the independent variable under study be *time*, t, we would define the fractional variable $\phi = t/T$, in terms of a **period**, T. If, on the other hand, the variable of interest is a *position*: a dimension, x, y or z, of Euclidean *space*, we could define $\phi = x/L$, measured in terms of some generic applicable length scale, L. While it introduces extra baggage in the terminology, unnecessary at this point, a common alternative (symbol) choice would be $\phi = x/\lambda$, which would relate position to a specific **wavelength**, λ.

But we do not need *waves* (with prescribed relations between space and

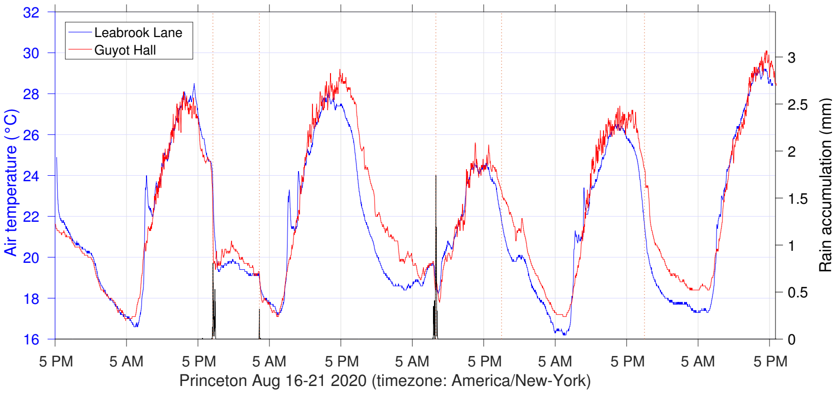


Fig. 5.1. Weather cycles in Princeton, New Jersey, late Summer 2020. Air temperature and rain accumulation measured by a Kestrel Drop D3 Fire on Leabrook Ln and a Vaisala WXT-520 instrument at Guyot Hall, WGS84 latitude $40.3458°$, longitude $-74.6547°$, elevation 46.692 m.

time) to proceed: *oscillations*, quantities that rise and fall around an equilibrium state (for example, a long-term average), will do.

Since $\sin$, $\cos$ and complex exponentials are functions that are 2π-periodic in their argument, as the fraction ϕ ranges from 0 to 1, the argument $2\pi\phi$ of $\sin(2\pi\phi)$ will range from 0 to 2π, and the function values from 0 over $+1$ back down to 0 and then to -1 and then back to 0 in one complete oscillation. Similarly, the values of $\cos(2\pi\phi$ will go from $+1$ down to 0 and then to -1 and 0 before rising back up to $+1$.

As to the *complex exponentials*, Euler's magic formula stipulates that

$$e^{i2\pi\phi} = \cos(2\pi\phi) + i\sin(2\pi\phi), \tag{5.1}$$

and, hence,

$$\cos(2\pi\phi) = \frac{1}{2}\left(e^{i2\pi\phi} + e^{-i2\pi\phi}\right), \tag{5.2}$$

$$\sin(2\pi\phi) = \frac{1}{2i}\left(e^{i2\pi\phi} - e^{-i2\pi\phi}\right). \tag{5.3}$$

The repeating nature of "periodic" signals is manifest in the realization that, for any integer $n \in \mathbb{Z}^+$, adding n times 2π to the argument of $\sin$, $\cos$ and $\exp$ leaves the function value unchanged. In the case of a time series with period T,

$$\sin\left(2\pi\frac{t}{T}\right) = \sin\left(2\pi\frac{t}{T} + 2\pi n\right) = \sin\left(\frac{2\pi}{T}\left[t + nT\right]\right), \tag{5.4}$$

from which we learn indeed that T is the stretch of time that passes until the oscillation repeats itself.

If one oscillation takes T units of time t to complete, then $1/T = f$, the *temporal* **frequency**, is the number of oscillations that can be completed in precisely *one* unit of t. If that frequency is measured in radians, $\omega = 2\pi f$ is the **angular frequency**. So we should be able to rewrite eq. (5.4), or, rather, reread it with new eyes, as

$$\sin\left(2\pi\frac{t}{T}\right) = \sin(2\pi f t) = \sin(\omega t), \tag{5.5}$$

where $ft = \phi$ is again the dimensionless quantity that counts the number of times a complete cycle (however long in physical units), is being completed. For a period T in seconds (s), the unit of frequency $f = 1/T$ is in Hertz (Hz), and the unit of angular frequency $\omega = 2\pi/T$ is in radians per second.

As to spatial oscillations, taking $\lambda = L$, the equivalent to eq (5.4) is

$$\sin\left(2\pi\frac{x}{\lambda}\right) = \sin\left(2\pi\frac{x}{\lambda} + 2\pi n\right) = \sin\left(\frac{2\pi}{\lambda}\left[x + n\lambda\right]\right). \tag{5.6}$$

I find no reason not to reuse the symbol $f = 1/\lambda$ for the *spatial* **frequency**, and I stick with convention to use $k = 2\pi/\lambda$ for the **wavenumber**.

$$\sin\left(2\pi\frac{x}{\lambda}\right) = \sin(2\pi f x) = \sin(kx). \tag{5.7}$$

Again, I caution that there needn't be any *waves* yet. It's just symbols, for different domains of applications, time or space, and no *dispersion* relation between k and ω is implied: there is no propagation *speed*.

5.2 Orthogonality

What happens upon integration over a complete cycle? Nothing much:

$$\sqrt{2}\int_0^1 \cos(2\pi\phi)\, d\phi = \left.\frac{\sin(2\pi\phi)}{\sqrt{2}\pi}\right|_0^1 = 0, \tag{5.8}$$

$$\sqrt{2}\int_0^1 \sin(2\pi\phi)\, d\phi = \left.\frac{-\cos(2\pi\phi)}{\sqrt{2}\pi}\right|_0^1 = \frac{-1}{\sqrt{2}\pi} + \frac{1}{\sqrt{2}\pi} = 0. \tag{5.9}$$

As to the complex exponentials, we will note the indefinite integral

$$\int e^{i2\pi\phi}\, d\phi = \frac{e^{i2\pi\phi}}{i2\pi}. \tag{5.10}$$

What about the integral of the *product* of two periodic functions? Let us imagine ourselves on an interval of size T or L, with a function $\sin(2\pi f t)$ or

$\sin(2\pi f x)$ whose *fundamental* **frequency** f is either $1/T$ or $1/L$, i.e., precisely matched to the *duration* or *length* in question, cycling once within that interval, and repeating identically outside of it interval, *ad infinitum*.

Of course, functions oscillating faster, at **integer multiples** of that fundamental frequency would equally well "fit" within the interval under consideration. For the case of proper "time series", we thus consider the integrals of products of $\sin(2\pi\, m\phi)$ and $\sin(2\pi\, n\phi)$, with $m, n \in \mathbb{Z}_0^+$ integers, over the scaled interval (i.e. in terms of the fraction ϕ), to formulate the relationships:

$$2\int_0^1 \left\{ \begin{matrix} \cos \\ \sin \end{matrix} \right\} (2\pi\, m\phi) \left\{ \begin{matrix} \cos \\ \sin \end{matrix} \right\} (2\pi\, n\phi)\, d\phi = \delta_{mn}, \quad \text{for } m, n \neq 0, \tag{5.11}$$

where the **Kronecker delta** means that the integral vanishes unless $m = n$, when it evaluates to 1. The curly braces in eq. (5.11) are meant to convey that you can pick either $\cos$ or $\sin$ and have the equation hold—but remember that you are *not* allowed to pick $\sin$ twice and $m = n = 0$, since in that case, the integral returns 0, and not 1. Also, you are not allow to pick $\cos$ twice and $m = n = 0$, since in that case, you'd get 2, not 1. All the other choices remain valid, and we'll simply ban $\sin(0)$ out of our heads for now, and remember that $\cos(0)$ will acquire a prefactor of 1 rather than the $\sqrt{2}$ that we endowed the other terms with in eqs (5.8), which helped us arrive at eq. (5.11).

In contrast, in terms of complex exponentials, the equivalent to eq. (5.11) is universally true for all integers m and n, as is easily shown:

$$\int_0^1 e^{+i2\pi m\phi} e^{-i2\pi n\phi}\, d\phi = \frac{1}{i2\pi} \frac{e^{i2\pi(m-n)\phi}}{(m-n)} \bigg|_0^1 = \delta_{mn}, \tag{5.12}$$

though you will need de l'Hôpital's rule to prove it for $m = n$.

5.3 Orthonormal basis expansions of square-integrable functions

What do we take away from eq. (5.11)? That the set

$$\left\{ \; 1 \quad \sqrt{2}\cos(2\pi\, m\phi) \quad \sqrt{2}\sin(2\pi\, m\phi) \; \right\} \qquad \text{for} \qquad m \in \mathbb{Z}^+, \tag{5.13}$$

is *orthonormal*. Even more conveniently, in the space of functions that are *square-integrable* over the interval $[0, 1]$, e.g., for all real-valued $s(\phi)$ whose energy is finite:

$$\int_0^1 |s(\phi)|^2\, d\phi < \infty, \tag{5.14}$$

the set (5.13) supplies a **complete orthonormal basis**. What that means is that any such square-integrable function $s(\phi)$ can be decomposed as a linear combination of these sinusoidal building blocks at the perfectly matched frequencies. The more elements we take from the set, the better the approximation.

One caveat is that the convergence is to be understood in the *mean-squared*, not in a *pointwise* sense. The mean-squared "misfit" over the interval can be brought arbitrarily close to zero if we bring enough "parameters" $m = 0, ..., M-1$ into the game:

$$\lim_{M\to\infty} \int_0^1 \left| \underbrace{a_0 + \sqrt{2} \sum_{m=1}^{M-1} \left[a_m \cos(2\pi\, m\phi) + b_m \sin(2\pi\, m\phi) \right]}_{\text{the } M\text{-term approximation to } s(\phi)} - s(\phi) \right|^2 d\phi = 0. \tag{5.15}$$

At which *rate* the misfit (a norm, or mean-squared error!) decays to zero, we cannot say at this point. Nor are we ready to list all the ways by which we can determine what the proper coefficients are.

Except for this one: orthogonality! If we are comfortable writing

$$s(\phi) \stackrel{ms}{=} a_0 + \sqrt{2} \sum_{m=1}^{\infty} \left[a_m \cos(2\pi\, m\phi) + b_m \sin(2\pi\, m\phi) \right], \tag{5.16}$$

with the augmented "equals" sign to be interpreted in the "mean-squared" sense, we should be able to multiply both sides by a certain element of the set (5.13), e.g., $\sqrt{2}\cos(2\pi\, n\phi)$ or $\sqrt{2}\sin(2\pi\, n\phi)$, integrate over the interval, and apply eq. (5.11) to recover exactly the coefficients a_n, and b_n, and so, one at a time, for all candidates n, corresponding ultimately to all $m = 1, ..., M-1$.

For certain simple functions you can find closed-form expressions analytically by hand. (Like boxcars, triangles, polynomials?)

Right away we understand why, in electrical engineering, a_0 is known as the direct-current or "DC" term, the *average* over the interval, since indeed

$$a_0 = \int_0^1 s(\phi) \cos(0)\, d\phi = \int_0^1 s(\phi)\, d\phi. \tag{5.17}$$

More generally, we have

$$a_m = \int_0^1 s(\phi) \cos(m\phi)\, d\phi, \tag{5.18}$$

$$b_m = \int_0^1 s(\phi) \sin(m\phi)\, d\phi. \tag{5.19}$$

Note that eqs (5.18)–(5.19) embody a **correlation** of the function $s(\phi)$ with the sinusoidal basis functions: a *similarity* measure, a *projection* in some sense. Various synonymous interpretations come to mind.

Maybe here we list some popular by-hand expansions (square, triangle, etc), but most importantly, their coefficient decay which would start building the intuition. So here we mention the *Gibbs effect* as PW 93, at points of discontinuity.

Picture here of interval approximation. Maybe some random thing, numerically approximated, in addition to the by-hand things. Make them more interesting than usual. Smiley face, semicircle, see Mathworld. Think about the "half" frequency. Think of Scott's heat equation examples?

Mention this is just one orthogonal expansion. With integrals for coefficients. Polynomials connect to what we have done in earlier chapters. Taylor series would be expansions with differentiated terms. Bessel, Chebyshev, etc.

5.4 The Fourier series (CTDF)

Big picture. Continuous time on the interval, discrete frequencies. Give up interval, turn the frequencies to continuous. Discretize the time domain, periodize the frequencies. Notice the equivalence, describe aliasing.

But how do we find the required coefficients? We may as well rewrite eq. (5.16) as an expansion into *complex* exponentials and absorb all constants into the unknown $\tilde{s}_m$:

$$\boxed{s(\phi) \overset{ms}{=} \sum_{m=-\infty}^{+\infty} \tilde{s}_m \, e^{+i2\pi m\phi} \qquad \text{for} \qquad 0 \le \phi < 1} \tag{5.20}$$

and then... repeat [8] (p 267). Now from eq. (5.12) the orthonormal basis is

$$\left\{1, e^{\pm i2\pi m\phi}\right\} \qquad \text{for} \qquad m \in \mathbb{Z}, \tag{5.21}$$

and the sign can be chosen freely. The expansion coefficients

$$\boxed{\tilde{s}_m = \int_0^1 s(\phi) \, e^{-i2\pi m\phi} \, d\phi.} \tag{5.22}$$

CTDF Eqs (5.20) and (5.22) form a *transform pair*: the continuous-time discrete-frequency (CTDF) **Fourier series**.

More detail

Let us minimize the sum of the squared error over the unit interval, defining the misfit function (watch the complex conjugation!), let's do this for real $s(\phi)$ only. Let us define as usual the mean-squared misfit function as the quantity to be minimized

$$\Phi = \int_0^1 \left| s(\phi) - \sum_{m=-\infty}^{\infty} \tilde{s}_m \, e^{+i2\pi m\phi} \right|^2 d\phi, \tag{5.23}$$

which leads to the condition that its derivative over the coefficients must vanish at the condition $\tilde{s}_k$, namely

$$\begin{aligned}
&\frac{\partial}{\partial \tilde{s}_k} \left[\int_0^1 \left(s(\phi) - \sum_{m=-\infty}^{\infty} \tilde{s}_m^* \, e^{-i2\pi m\phi} \right) \left(s(\phi) - \sum_{n=-\infty}^{\infty} \tilde{s}_n \, e^{+i2\pi n\phi} \right) d\phi \right] \\
&\quad = \int_0^1 \left[\left(s(\phi) - \sum_{m=-\infty}^{\infty} \tilde{s}_m^* \, e^{-i2\pi m\phi} \right) \left(- e^{+i2\pi k\phi} \right) \right. \\
&\qquad\qquad \left. + \left(-e^{-i2\pi k\phi} \right) \left(s(\phi) - \sum_{n=-\infty}^{\infty} \tilde{s}_n \, e^{+i2\pi n\phi} \right) \right] d\phi \\
&\quad = \int_0^1 \left[\left(s(\phi) \left(- e^{+i2\pi k\phi} \right) - \sum_{m=-\infty}^{\infty} \tilde{s}_m^* \, e^{-i2\pi m\phi} \left(- e^{+i2\pi k\phi} \right) \right) \right. \\
&\qquad\qquad \left. + \left(s(\phi) \left(-e^{-i2\pi k\phi} \right) - \sum_{n=-\infty}^{\infty} \tilde{s}_n \, e^{+i2\pi n\phi} \left(-e^{-i2\pi k\phi} \right) \right) \right] d\phi
\end{aligned}$$

Using the orthogonality eq. (5.12), this requires

$$\begin{aligned}
0 &= \int_0^1 \left[-s(\phi) \, e^{+i2\pi k\phi} + \tilde{s}_k^* - s(\phi) e^{-i2\pi k\phi} + \tilde{s}_k \right] d\phi \\
&= \int_0^1 \left[-s(\phi) \, e^{+i2\pi k\phi} - s(\phi) \, e^{-i2\pi k\phi} \right] d\phi + \left[\tilde{s}_k^* + \tilde{s}_k \right] \int_0^1 d\phi,
\end{aligned}$$

In other words, we require

$$\left[\tilde{s}_k^* + \tilde{s}_k \right] = \int_0^1 \left[s(\phi) e^{+i2\pi k\phi} + s(\phi) e^{-i2\pi k\phi} \right] d\phi,$$

from which we find eq. (5.22) again.

5.4.1 Out of the unit interval

Rewrite eqs (5.20)–(5.22) by moving to an interval of length T, and make $s(t)$ repeat periodically, but respecting the mean-squared integrability $\int_{-\infty}^{+\infty} |s(t)|^2 \, dt < \infty$, as follows:

$$s(t) = \sum_{m=-\infty}^{+\infty} \tilde{s}_m \, e^{+i2\pi m \frac{t}{T}} \quad \text{for all} \quad t \in \mathbb{R}, \tag{5.24}$$

$$\tilde{s}_m = \frac{1}{T} \int_{-T/2}^{T/2} s(t) \, e^{-i2\pi m \frac{t}{T}} \, dt, \tag{5.25}$$

and then define $f_m = m/T$, the **discrete frequencies**, to arrive at:

$$\boxed{s(t) = \sum_{m=-\infty}^{+\infty} \tilde{s}_m \, e^{+i2\pi f_m \, t},} \tag{5.26}$$

$$\boxed{s_m = \frac{1}{T} \int_{-T/2}^{T/2} s(t) \, e^{-i2\pi f_m \, t} \, dt,} \tag{5.27}$$

then let $T \to \infty$. The **Fourier coefficients** are $\tilde{s}_m$. Spacing of the frequencies is $\Delta f = 1/T$.

5.5 The Fourier integral transform (CTCF)

Should put a subscript for "periodicity". Argument is to replicate a general function over the period T and then let the interval $[-T/2, T/2]$ over which the function is assumed to be periodic stretch all the way from $-\infty$ to $+\infty$, thereby spacing the frequencies continuously, $\Delta f \to df$ and allowing for effectively *non-periodic* functions $s(t)$ to be represented by the continuous-time
CTCF continuous-frequency (CTCF) **Fourier integral-transform** pair

$$\boxed{s(t) = \int_{-\infty}^{+\infty} \tilde{s}(f) \, e^{+i2\pi f \, t} \, df,} \tag{5.28}$$

$$\boxed{\tilde{s}(f) = \int\limits_{-\infty}^{+\infty} s(t)\, e^{-i2\pi f\, t}\, dt.} \tag{5.29}$$

5.6 Discrete-time continuous-frequency (DTCF)

In the real world, however, t is **sampled**, so it is **discrete**. Now t is merely an integer *index* and Δt is a *spacing* or a *sampling interval*:

$$s(t) = s(t\Delta t), \qquad \text{for} \qquad t = 0, \pm 1, \pm 2, ... \tag{5.30}$$

In analogy with eq. (5.29) we *define*

$$\boxed{\tilde{s}_p(f) = \sum_{t=-\infty}^{\infty} s(t\Delta t)\, e^{-i2\pi f\, t\Delta t} \Delta t,} \tag{5.31}$$

essentially a Riemann sum or rectangular *approximation* to the integral in eq. (5.29).

We used a new symbol, $\tilde{s}_p(f)$, for the following reason: for $n = 0, \pm 1, \pm 2, ...$, we notice that

$$\begin{aligned} \tilde{s}_p\left(f + \frac{n}{\Delta t}\right) &= \sum_{t=-\infty}^{\infty} s(t\Delta t)\, e^{-i2\pi f\, t\Delta t} \Delta t \underbrace{\left(e^{-i2\pi\, nt\, \Delta t/\Delta t}\right)}_{1} \\ &= \tilde{s}_p(f), \end{aligned} \tag{5.32}$$

i.e. since both t and n are integers so nt is *also* an integer, the last factor vanishes... hence turns out that the $\tilde{f}$ is *periodic* with period $1/\Delta t$. It's discrete-time, continuous frequency. So we wrote $\tilde{s}_p$ as a mnemonic, to remind us of the periodicity.

We now turn the argument around and note that the *discrete* inverse Fourier transform of eq. (5.31) is given by simple *evaluation* of eq. (5.28), namely

$$\boxed{s(t\Delta t) = \int\limits_{-1/(2\Delta t)}^{1/(2\Delta t)} \tilde{s}_p(f)\, e^{i2\pi f t\Delta t}\, df,} \tag{5.33}$$

because in this case the orthogonality property is

$$\int\limits_{-1/(2\Delta t)}^{1/(2\Delta t)} e^{-i2\pi f t\Delta t} e^{+i2\pi f\, t\Delta t} \Delta t\, df = \Delta t \left[\frac{1}{2\Delta t} + \frac{1}{2\Delta t}\right] = 1. \tag{5.34}$$

DTCF

Eqs (5.31)–(5.33) give us the discrete-time continuous-frequency (DTCF) relation.

5.7 Sampling, aliasing, etc.

What now is the connection between **continuous-continuous** transform pair of eqs (5.29) and (5.28), which we rewrite here

$$\tilde{s}(f) = \int_{-\infty}^{+\infty} s(t)e^{-i2\pi ft}\, dt \tag{5.35a}$$

$$s(t) = \int_{-\infty}^{+\infty} \tilde{s}(f)e^{+i2\pi ft}\, df \tag{5.35b}$$

and the **discrete-continuous** version of eqs (5.31) and (5.33) which we rewrite here (note that $s_p \neq s$) as

$$\tilde{s}_p(f) = \sum_{t=-\infty}^{\infty} s(t\Delta t)\, e^{-i2\pi f\, t\Delta t}\Delta t \tag{5.36a}$$

$$s(t\Delta t) = \int_{-1/(2\Delta t)}^{1/(2\Delta t)} \tilde{s}_p(f)\, e^{i2\pi f\, t\Delta t}\, df \tag{5.36b}$$

Something appears to be missing: what was formerly an interval between $-\infty$ and ∞ now seems to contain the complete spectral information between the frequencies $-1/(2\Delta t)$ and $1/(2\Delta t)$. But as we can see by evaluating eq. (5.35b) at $t\Delta t$, using a trick to split the infinite integration interval into a infinite sum of finite integration intervals of width $1/\Delta t$, with the central one at $n = 0$ being the one that appears in eq. (5.36b), and then making a change of variables from $f' \to (f + n/\Delta t)$, we obtain

$$s(t\Delta t) = \int_{-\infty}^{+\infty} \tilde{s}(f')\, e^{i2\pi f' t\Delta t}\, df' \tag{5.37}$$

$$= \sum_{n=-\infty}^{\infty} \int_{\frac{2n-1}{2\Delta t}}^{\frac{2n+1}{2\Delta t}} \tilde{s}(f')\, e^{i2\pi f' t\Delta t} df' \tag{5.38}$$

$$= \sum_{n=-\infty}^{\infty} \int_{-1/(2\Delta t)}^{1/(2\Delta t)} \tilde{s}\left(f + \frac{n}{\Delta t}\right)\, e^{i2\pi (f+n/\Delta t)t\Delta t}\, df. \tag{5.39}$$

We thus conclude that

$$s(t\Delta t) = \int\limits_{-1/(2\Delta t)}^{1/(2\Delta t)} \underbrace{\left[\sum_{n=-\infty}^{\infty} \tilde{s}\left(f + \frac{n}{\Delta t}\right)\right]}_{\tilde{s}_p(f)} e^{i2\pi f\, t\Delta t}\, df, \tag{5.40}$$

where we identify the underbrace from eq. (5.36b) to notice that the periodic $\tilde{s}_p(f)$ which we defined in eq. (5.31) simply *copies* the portion of the frequencies of $\tilde{s}(f)$ as it appeared in eq. (5.28) contained inside and *wraps* them into the central interval:

$$\boxed{\tilde{s}_p(f) = \sum_{n=-\infty}^{\infty} \tilde{s}\left(f + \frac{n}{\Delta t}\right) \qquad \text{for} \qquad |f| < \frac{1}{2\Delta t}.} \tag{5.41}$$

So here is how $\tilde{s}$ relates to $\tilde{s}_p$: by *sampling* in the time domain we get copies of the *amplitude spectrum* (need to define before) in the frequency domain. But what was originally outside of $\pm 1/(2\Delta t)$ appears now within. What we get out at f for $\tilde{s}_p(f)$ depends not just on $\tilde{s}(f)$ but on all the possible values at frequencies of $n/\Delta t$ away from the target f! This effect is called **aliasing**.

The highest unaliased frequency is called the **Nyquist frequency**.

Mallat: wheels in films. Moiré effect. Wunsch says: "*Those who do not understand it are condemned to foolish results*".

5.8 Bandlimited signals and the sampling theorem

What, pray tell, is the relation between the original *continuous* signal and its *sampled* equivalent?

If, however, the original signal is bandlimited, which is as much as being able to restrict the infinite integral in eq. (5.28) to

$$s(t() = \int_{-W}^{+W} \tilde{s}(f)\, e^{+i2\pi f t}\, df \tag{5.42}$$

or indeed when the bandwidth is *precisely* limited to the interval bounded by the **Nyquist frequencies**,

$$\tilde{s}(f) = 0 \quad \text{when} \quad |f| > W = \frac{1}{2\Delta t} \tag{5.43}$$

in other words

$$\tilde{s}_p(f) = \tilde{s}(f) \quad \text{for} \quad |f| \leq \frac{1}{2\Delta t} \tag{5.44}$$

then we have the special lucky case that we will be use eq. (5.31) or eq. (5.36a)

$$s(t) = \int_{-1/2\Delta t}^{1/2\Delta t} \tilde{s}(f)\, e^{i2\pi f t}\, df \tag{5.45}$$

$$= \int_{-1/2\Delta t}^{1/2\Delta t} \left[\sum_{t'=-\infty}^{+\infty} s(t'\Delta t)\, e^{-i2\pi f t'\Delta t} \Delta t \right] e^{i2\pi f t}\, df \tag{5.46}$$

$$= \sum_{t'=-\infty}^{+\infty} s(t'\Delta t)\Delta t \left[\int_{-1/2\Delta t}^{1/2\Delta t} e^{i2\pi f(t-t'\Delta t)}\, df \right] \tag{5.47}$$

$$= \sum_{t'=-\infty}^{+\infty} s(t'\Delta t)\Delta t \left[\frac{\sin \frac{\pi(t-t'\Delta t)}{\Delta t}}{\pi(t-t'\Delta t)} \right] \tag{5.48}$$

The last factor is something we recognized from the formulas in the first section... Summarizing we obtain: the **Whittaker-Shannon** sampling theorem:

$$\boxed{s(t) = \sum_{x'=-\infty}^{+\infty} s(t'\Delta t)\, \mathrm{sinc}\left(\frac{t - t'\Delta t}{\Delta t} \right).} \tag{5.49}$$

Ours is the engineering or signal-processing definition of the "cardinal sine" or "sinc" function, $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. A bandlimited signal can be recovered from its samples by interpolation if it is sampled at the **Nyquist rate**, with the samples spaced Δt apart, with the **Nyquist frequencies** $\pm 1/(2\Delta t)$.

More detail

Was it obvious?

$$\frac{\Delta t}{2i\pi(t - t'\Delta t)} \int_{-1/2\Delta t}^{1/2\Delta t} e^{i2\pi k(t-t'\Delta t)}\, df\; (i2\pi k(t - t'\Delta t))$$

$$= \frac{\Delta t}{2i\pi(t - t'\Delta t)}\; e^{i2\pi f(t-t'\Delta t)} \Big|_{k=-1/2\Delta t}^{f=1/2\Delta t}$$

$$= \frac{\Delta t}{2i\pi(t - t'\Delta t)} \left[e^{\frac{i2\pi(t-t'\Delta t)}{2\Delta t}} - e^{\frac{-i2\pi(t-t'\Delta t)}{2\Delta t}} \right]$$

$$= \frac{\sin \frac{\pi(t-t'\Delta t)}{\Delta t}}{\frac{\pi(t-t'\Delta t)}{\Delta t}} = \mathrm{sinc}\left(\frac{t - t'\Delta t}{\Delta t} \right). \tag{5.50}$$

5.9 Sampling, filtering, and interpolation

Later, think of this as a filtering of the samples with an ideal boxcar. Remind ourselves of the continuous convolution (a commutative operation),

$$[g * h](t) = \int_{-\infty}^{+\infty} g(t')\, h(t - t')dt' \qquad , \tag{5.51}$$

which is making a discrete comeback. Let us have a discrete signal g be convolved with a continuous function h and evaluated at the continuous variable t:

$$[g * h](t) = \sum_{t'=-\infty}^{+\infty} g(t'\Delta t)\, h(t - t'\Delta t)\Delta t, \tag{5.52}$$

So the sampling theorem (5.49) looks like a filtering. The continuous signal at t is a filtered/interpolated version of the sampled one. Why should this be obvious? Because of the convolution theorem, which is up next.

5.10 The convolution-Fourier duality

Let us consider two signals sampled with the same sampling interval, Δt, and convolve them. Let g be a discrete-time function, i.e. $g = g(t\Delta t)$. Let h be a discrete-time function i.e., $h = h(t\Delta t)$. Define

$$[g * h](t\Delta t) = \sum_{t'=-\infty}^{+\infty} g(t'\Delta t)\, h(t\Delta t - t'\Delta t)\Delta t. \tag{5.53}$$

Using eq. (5.31), let the Fourier transform $\tilde{f}$ be (let us not bother with any subscripts: in reality we are talking bandlimited functions here, a property which is preserved upon convolution, as will be obvious from the outcome):

$$\tilde{g}(f) = \sum_{t=-\infty}^{+\infty} g(t\Delta t)\, e^{-2i\pi f\, t\Delta t}\Delta t. \tag{5.54}$$

Need to refer to eq. (5.31). Let $\tilde{h}$ be

$$\tilde{h}(f) = \sum_{t=-\infty}^{+\infty} h(t\Delta t)\, e^{-2i\pi f\, t\Delta t}\Delta t. \tag{5.55}$$

And the argument should be that you can add a shift to this in the time domain variable and you don't get a difference at all. The convolution of g and h is given be the discrete sum as in eq. (5.53) and let's invent a new symbol for it:

$$\mathcal{F}(t\Delta t) = [g * h](t\Delta t). \tag{5.56}$$

The continuous-frequency Fourier transform of the outcome is given by applying eq. (5.31), and then splitting the exponential terms, to return

$$\tilde{\mathcal{F}}(f) = \sum_{x=-\infty}^{+\infty} \left[\sum_{t'=-\infty}^{+\infty} g(t'\Delta t)\, h(t\Delta t - t'\Delta t)\Delta t \right] e^{-i2\pi f t \Delta x} \Delta t \tag{5.57}$$

$$= \left(\sum_{t'=-\infty}^{+\infty} g(t'\Delta t')\, e^{-i2\pi f t' \Delta t'} \Delta t \right)$$
$$\times \left(\sum_{t=-\infty}^{+\infty} h(t\Delta t - t'\Delta t)\, e^{-i2\pi f (t\Delta t - t'\Delta t)} \Delta t \right) \tag{5.58}$$

The last factor can be evaluated first to yield the Fourier transform $\tilde{h}$ and the first factor is the Fourier transform $\tilde{g}$. So the final result is

$$\tilde{\mathcal{F}}(f) = \tilde{g}(f)\, \tilde{h}(f). \tag{5.59}$$

Convolution in the time domain equals multiplication in the frequency domain. The Fourier transform of the convolution of two signals is the multiplication of their Fourier transforms.

5.11 Sincs and boxcars

We did the above for discrete-discrete as this how you'll be doing this in practice. However, we need to revisit one more thing and ask ourselves of which continuous function the $sinc$ is the continuous Fourier transform... Now back to the interpretation of the Whitaker-Shannon theorem. The sinc and the boxcar are each other's transform pairs. This is simple: we use the CTCF formalism. We've of course already done this and should be able to notice this straight from the shifted transform right before the Whitaker-Shannon theorem, yet we'll be very explicit here...

$$\begin{aligned}
\tilde{h}(f) = \int_{-1/2\Delta t}^{1/2\Delta t} e^{i2\pi f t}\, df &= \frac{1}{i2\pi t} \left[e^{i2\pi f t} \right]_{-1/2\Delta t}^{+1/2\Delta t} \\
&= \frac{1}{i2\pi t} \left[e^{i\pi t/\Delta t} - e^{-i\pi x/\Delta t} \right] \\
&= \frac{\sin \pi t/\Delta t}{\pi t} \\
&= \frac{1}{\Delta t} \operatorname{sinc}\left(\frac{t}{\Delta t} \right).
\end{aligned} \tag{5.60}$$

And this finally puts eq. (5.49) in the proper context as the product of the Fourier transforms. The continuous $h(x)$ is indeed the convolution as per eq. (5.52) with the Fourier transform of the frequency-domain signal that bandlimits the continuous input to the band between. After sampling, we can never expect to gain anything better than that.

Picture Notes 56.

Now we can rewrite this for W and T generic bandwidth and period, and move on to the next section.

5.12 Heisenberg, Plancherel, Parseval

Let $T \to 0$ and $1/T \to \infty$, the function becomes narrower and narrower, and $\mathrm{sinc}(kT)$ becomes broader and broader. This is not a proof, but an illustration of **Heisenberg's uncertainty principle**: the broader a function is in the time domain, the narrower in the frequency domain.

The behavior is nicely illustrated by considering the Fourier transform of the Gaussian density.

Energy in the time domain equals that in the frequency domain. One such equation for every form of our transforms. Here I pick one

$$\Delta t \sum_{t=0}^{N-1} |f_t|^2 = \frac{\Delta t}{(N\Delta t)^2} \sum_{n=0}^{N-1} \sum_{n'=0}^{N-1} \tilde{f}_n \tilde{f}^*_{n'} \underbrace{\sum_{t=0}^{N-1} e^{i2\pi \frac{(n-n')t}{N}}}_{N\delta_{nn'}} \tag{5.61}$$

$$= \frac{1}{N\Delta t} \sum_{n=0}^{N-1} \left|\tilde{f}_n\right|^2 \tag{5.62}$$

The underbrace is as before. Here too it would have been easier to symmetrize the transform, wouldn't it. The boys trade credit for deriving the various transforms and whether you're using one or more functions in the relation. Energy is conserved.

5.13 Discrete-time discrete-frequency (DTDF)

Since our sequence—imagine we have it—contains only a finite number, N, of time-sampled values, with sampling interval Δt, it seems logical to look for a representation of it that only involves a finite number—the same—of frequencies.

Define a grid of frequencies,

$$f_n = \frac{n}{N\Delta t}, \tag{5.63}$$

where $N\Delta t$ is the total signal length and $n = 0, 1, ..., N-1$. And let us define the **discrete Fourier transform** as:

$$\boxed{\tilde{s_n} = \Delta t \sum_{t=0}^{N-1} s_t \, e^{-i2\pi \dfrac{nt\Delta t}{N\Delta t}},} \tag{5.64}$$

completely inspired by eq. (5.31) except truncated to a finite set of frequencies.

Now... get the inverse transform

$$\boxed{\tilde{s}_t = \frac{1}{N\Delta t} \sum_{n=0}^{N-1} \tilde{s}_n \, e^{i2\pi \dfrac{nt\Delta t}{N\Delta t}}.} \tag{5.65}$$

What we needed for this is (see my RB p 42, it uses geometric series and basic identities))

$$\sum_{n=0}^{N-1} e^{i2\pi \frac{n}{N} z} = N \tag{5.66}$$

of $z = mN$ for integer m or 0 otherwise. When is $(t-t')$ an integer multiple of N, remember they are never more than $N-1$ apart, hence only when $t'-t = 0$. This completes the pair verification.

More detail

My notes on the last page of PW. Graphical argument is intuitive.

Not surprisingly if we use eq 5.64 to *define* $\tilde{f}_n$, even for all n the we notice again that $\tilde{f}_n$ is periodic with period N as expected.

$$e^{-i2\pi \frac{n+N}{N} x} = e^{-i2\pi \frac{n}{N} x} \underbrace{e^{-i2\pi x}}_{1} \tag{5.67}$$

since x is an integer. But also $\tilde{f}_n e^{i2\pi \frac{n}{N} x}$ is periodic with period N so we may pick an integer m and add it to the sum as

$$\tilde{f}_x = \frac{1}{N\Delta x} \sum_{n=m}^{N+m-1} \tilde{f}_n \, e^{i2\pi \frac{n}{N} x}. \tag{5.68}$$

The frequencies used to be $k_0 \to k_{N-1}$ but now they are $k_m \to k_{N+m-1}$. It used to be the interval $0, \frac{N-1}{N\Delta x}$ but that falls shy of the Nyquist frequency, so why don't we simply take

$$m = -\left[\frac{(N+1)}{N\Delta x}\right] + 1 \tag{5.69}$$

and construct a frequency axis with that.

Frequencies NOTE chance to floor

$$f = \frac{-[\frac{N+1}{2}] + 1 + n}{N\Delta t} \qquad \text{where} \qquad n = 0, \ldots, N-1 \tag{5.70}$$

When N is even the smalles frequency is

$$\frac{-N/2+1}{N\Delta x} > \frac{-1}{2\Delta x} \tag{5.71}$$

but the biggest frequency

$$\frac{N/2}{N\Delta x} = \frac{1}{2\Delta x} \tag{5.72}$$

exactly. Where is the “zero” frequency? at NOTE floor

$$n = [\frac{N+1}{2}] - 1 \tag{5.73}$$

that is the longest wavelength that you can extract:

$$f = \frac{1}{N\Delta t} \tag{5.74}$$

one over the signal length (“Rayleigh”) frequency, and the shortest wavelenegth you can extract is

$$f = \frac{1}{2\Delta t} \tag{5.75}$$

the Nyquist frequency.

So this axes has

$$-f_N < f_n \le f_N \tag{5.76}$$

but always through zero. Take a look again at eqs (5.64)-(5.65) and attempt to write it as a matrix equation as does Strang p290.

$$W = e^{-i2\pi x/N} \tag{5.77}$$

then the matrix is filled with powers of w. FFT writes and solves this in a clever way. w is unitary. In the column dimensions the powers come about due to the samples, in the row dimensions they come out due to the frequencies. LOOK AT THE TILDES

$$\begin{bmatrix} \tilde{f}_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & e^{-i2\pi\frac{x}{N}} & x = 0, \ldots, N-1 & & \\ 1 & e^{-i2\pi\frac{2x}{N}} & & & \\ \vdots & & & & \end{bmatrix} \begin{bmatrix} \tilde{f}_x \end{bmatrix} \tag{5.78}$$

5.14 Windowing, etcetera

We have already seen that convolution in the time domain corresponds to multiplication in the frequency domain. Now, with the DTDF of course the reverse is also true: convolution in the frequency domain corrresponds to multiplication in the time domain.

Picture. For the sake of the argument, let it be Nyquist bandlimited. First you've blurred it with a sinc function and then you've periodized it. Clearly you've messed it up.

PW119b

$$(g \star h)_x = \Delta x \sum_{y=0}^{N-1} g_y h_{x-y} \tag{5.79}$$

$$\Delta x \sum_{x=0}^{N-1} (g \star h)_x e^{-i2\pi \frac{n}{N} x} = \tilde{g}_n \tilde{h}_n \tag{5.80}$$

Use cyclically, $h_s = h_{\mod (s,N)}$ for s outside the range $[0, N-1]$.

$$\tilde{g}_n = \Delta x \sum_{x=0}^{N-1} g_x e^{-i2\pi \frac{n}{N} x} \tag{5.81}$$

Need padding to make cyclic. It's best to use FFT even for convolution, N^2 vs $3N \log N$.

6
Spectral Analysis

Need a spectral variance-based representation, just like we had a Fourier representation. The key reference is [9].

6.1 Parseval/Plancherel again?

Let's write Parseval/Plancherel again for DTCF processes. PW89. For that case we have

$$\Delta x \sum_{x=-\infty}^{\infty} |f_x|^2 = \int_{-1/2\Delta x}^{1/2\Delta x} |\tilde{s}_p(k)|^2 \, df \tag{6.1}$$

This is about one realization. Let's now interpret the "process" as a random variable, i.e. we have samples or realizations of $\{f_x\}$ at different x. Let's take $E\{f_x\} = 0$ for simplicity without trouble. Let's take f_x be real. The left hand side of eq. (6.1) is the total energy of the signal, some measure of the variance, and apparently $|\tilde{s}_p(k)|^2$ is a measure of the energy spectral density: $|\tilde{s}_p(k)|^2 \, df$ is the amount of energy to to sinusoids/exponentials of the frequencies in the interval k to $k + df$, i.e. a measure of how much energy there is in a small frequency interval. Now $\tilde{s}_p(p)$ are random variables.

This sort of interpretation would have worked equally well in other domain, obviously. E.g. PW66.

$$\int_{-\infty}^{\infty} |g(t)|^2 \, dt = \int_{-\infty}^{\infty} |G(f)|^2 \, df. \tag{6.2}$$

6.2 The power spectral density

Take $\Delta t = 1$ and let us rewrite eq. (5.33) as

$$s_t = \int_{-1/2}^{1/2} e^{i2\pi ft} \underbrace{\tilde{s}_p(f)\,df}_{dZ(f)}. \tag{6.3}$$

This is known as **Cramér**'s spectral representation, whereby $dZ(f)$ is an **orthogonal increment spectral process**. Let us have

$$E\{dZ(f)\} = 0, \tag{6.4}$$

and let us also have

$$\text{cov}\{dZ(f), dZ(f')\} = S(f)\,df\,\delta_{ff'}, \tag{6.5}$$

which applies under **stationarity**. We define the true **energy spectrum** to be

$$\text{cov}\{dZ(f), dZ(f)\} = E\{|dZ(f)|^2\} = S(f)\,df. \tag{6.6}$$

The question is: what can we learn from a finite set of observations about the true $S(f)$ that we are interested in? Let's use the (D)FT of what we got, i.e. the finite sequence to form an estimate, a *direct* one, without averaging, and see what we get in expectation. Sums only range over the samples, not infinity. Use the square of eq (5.31) to define $\hat{S}(f)$, and find its expectation:

$$E\{\hat{S}(f)\} = E\left\{\left|\sum_t s_t\, e^{-i2\pi ft} a_t\right|^2\right\} \tag{6.7}$$

$$= E\left\{\sum_t \sum_{t'} s_t s_{t'}^*\, e^{i2\pi k(t'-t)} a_t a_{t'}^*\right\} \tag{6.8}$$

$$= \sum_t \sum_{t'} E\{s_t s_{t'}^*\} e^{i2\pi f(t'-t)} a_t a_{t'}^*. \tag{6.9}$$

the only stochastic variable being s_t, and for some window a_t.

Now let us use the spectral representation eq. (6.3) to evaluate the autocovariance at different times, perhaps call this some other symbol.

$$E\{s_t s_{t'}^*\} = \int_{-1/2}^{1/2} e^{i2\pi f'(t-t')} S(f')\,df', \tag{6.10}$$

the Wiener/Wold/Khintchine theorem. The autocovariance at lag $(t - t') = \tau$ is the Fourier transform of the spectrum. White spectrum you end up with $\delta_{tt'}$.

And therefore the expected value of the estimate constructed from a windowed discrete Fourier transform is given by

$$E\{\hat{S}(f)\} = \sum_t \sum_{t'} a_t a_{t'}^* \, e^{i2\pi f(t'-t)} \int_{-1/2}^{1/2} e^{i2\pi f'(t-t')} S(f') \, df' \tag{6.11}$$

$$= \int_{-1/2}^{1/2} S(f') \sum_t \sum_{t'} a_t a_{t'}^* e^{i2\pi t(f-f')} e^{-i2\pi t'(f'-f)} \, df' \tag{6.12}$$

$$= \int_{-1/2}^{1/2} S(f') \left| \sum_t a_t e^{-i2\pi t(f-f')} \right|^2 df' \tag{6.13}$$

$$= \int_{-1/2}^{1/2} S(f') \left| \tilde{a}(f-f') \right|^2 df', \tag{6.14}$$

where we notice the convolution with the power spectrum of the applied window $\tilde{a}(f)$. Hence we arrive at the conclusion

$$E\{\hat{S}(f)\} = S * A, \tag{6.15}$$

where A is the **spectral window**. And clearly, a "good" estimate is going to have $A \approx 1$. Windowing introduces bias, keep it low.

If we look at *white* spectra, we can keep the result unbiased if we keep the power of the spectral window to unity, since then

$$E\{\hat{S}(f)\} = S \underbrace{\int_{-1/2}^{1/2} \left| \tilde{a}(f-f') \right|^2 df}_{1} . \tag{6.16}$$

At least then we're good for white spectra. And of course, using Parseval's identity, we require that

$$\int_{-1/2}^{1/2} \left| \tilde{a}(t-t') \right|^2 dt = \sum_{t=0}^{N-1} a_t^2 = 1. \tag{6.17}$$

So the search is on for properly normalized windows that have good, low, sidelobes and tall, narrow main lobes. Peaky picture, that's what you want!

What is the variance of the estimate? Windowed or not, it's a sum of squares, and you can show that for large N, This is an *inconsistent* estimator: it never gets better!

$$\mathrm{var}\{\hat{S}(f)\} \approx \begin{cases} S^2(f) & \text{for} \quad 0 < f < \frac{1}{2\Delta t} \\ 2S^2(f) & \text{for} \quad f = 0 \quad \text{and} \quad f = \frac{1}{2\Delta t} \end{cases} \tag{6.18}$$

If windowed, you get a "smoothed" grid of uncorrelated frequencies. Windowing introduces bias and covariance. The below needs work.

$$\mathrm{cov}\{\hat{S}(f), \hat{S}(k')\} = 0 \qquad \text{for} \qquad 0 \le f', f \le \frac{1}{2\Delta t} \tag{6.19}$$

We won't be able to do better on bias, but we can try to lower the variance and make the estimate more consistent.

6.3 Multitaper spectral estimation

$$\hat{S}^{\mathrm{MT}}(f) = \frac{1}{K} \sum_{k'=0}^{K-1} \left| a_{t,k} s_t e^{-i2\pi f t} \right|^2, \tag{6.20}$$

the average of many tapered estimates which we should call $\hat{S}_k(f)$. PW333.

For every one of these we have

$$E\{\hat{S}_{k'}(f)\} = \int_{-1/2}^{1/2} S(f') \left| \tilde{a}_{k'}(f - f') \right|^2 \, df', \tag{6.21}$$

and for the combined estimate

$$E\{\hat{S}_{k'}^{\mathrm{MT}}(f)\} = \int_{-1/2}^{1/2} S(f') \mathcal{A}(f - f') \, df', \tag{6.22}$$

where the average spectral window is

$$\mathcal{A}(f) = \frac{1}{K} \sum_{k'=0}^{K-1} \left| \tilde{a}_{k'}(f) \right|^2. \tag{6.23}$$

Variance properties are more complicated to derive (depend on assumptions, and **Isserlis theorem**... but for *orthogonal* tapers, the variance *between* tapered estimates are just about *uncorrelated*. **Slepian** functions/sequences are one particular such choice.

6.4 A unifying framework

Fourier being such a tool. But only a tool.

6.5 Alternative approaches, and heuristics

Other ways. Looking. Correlation. Direct inversion. Stacking.

Maybe completely close the loop with periodfit and harmonic processes a la Pete and PW.

Bibliography

A. Tarantola, *Inverse Problem Theory*, Elsevier, Amsterdam, Neth., 1987.

J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, John Wiley, New York, 3 ed., 2000.

R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, vol. 90 of *International Geophysics Series*, Elsevier Academic Press, San Diego, Calif., 2005.

D. Gubbins, *Time Series Analysis and Inverse Theory for Geophysicists*, Cambridge Univ. Press, Cambridge, UK, 2004.

G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge, Wellesley, Mass., 1986.

W. Menke, *Geophysical Data Analysis: Discrete Inverse Theory*, vol. 45 of *International Geophysics Series*, Academic Press, San Diego, Calif., Rev. ed., 1989.

K. Mosegaard and A. Tarantola, "Monte Carlo sampling of solutions to inverse problems," *J. Geophys. Res.* **100**(B7), pp. 12431–12447, 1995.

G. Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, San Diego, Calif., 3 ed., 1998.

D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications, Multitaper and Conventional Univariate Techniques*, Cambridge Univ. Press, New York, 1993.