

Remote Sounding 1: Estimation Theory

Myles Allen, Department of Physics, University of Oxford.

`myles.allen@physics.ox.ac.uk`

Estimation theory is crucial to many problems in atmospheric, oceanic and planetary physics (and much else besides, from other branches of physics to financial forecasting):

- Estimating the current or past state of the system from remote or indirect observations – the retrieval or inverse problem.
- Assessing the implications of direct or indirect observations in terms of underlying theory – the data analysis and interpretation problem.
- Direct use of observations as constraints on a prognostic model derived from underlying theory – the data assimilation problem.
- An example: dual-view sea-surface temperature (SST) retrieval from the Along-Track Scanning Radiometer.

A worked example: dual-view SST retrieval from the Along-Track Scanning Radiometer.

- ATSR looks down at the same point on the surface, first with a “forward view” at an angle of $\Theta = 55^\circ$ then, 100s later, with a “nadir view” from directly overhead.
- If we assume neither the SST nor overlying atmosphere changes in the intervening period, L_N and L_F are the radiances observed in the nadir and forward views respectively, while L_S and L_A are the contributions to L_N from the surface and overlying atmosphere respectively, we have

$$L_N = L_S + L_A \tag{1}$$

$$L_F = L_S + (\sec \Theta)L_A \tag{2}$$

because the atmospheric path length increases by $\sec \Theta$ in the forward view (ignoring various things like surface emissivity variations).

- If we linearise the Planck function about a single reference temperature T_R , then we can replace all radiances with “brightness temperature anomalies” relative to that reference temperature, giving

$$T_N = T_S + T_A \tag{3}$$

$$T_F = T_S + (\sec \Theta)T_A, \tag{4}$$

where T_S is the surface brightness temperature and T_A the “atmospheric correction”.

- Thus, given observations $\mathbf{y} = (T_N, T_F)$, we can solve a simple pair of simultaneous equations for the “state vector” $\mathbf{x} = (T_S, T_A)$, to obtain

$$T_N - (\cos \Theta)T_F = (1 - \cos \Theta)T_S. \tag{5}$$

So what is going on here? We can linearise the dependence of (T_N, T_A) on surface and atmospheric temperatures in terms of two “response-functions” or “weighting functions”, being the vectors $\mathbf{k}_S = (1, 1)$ and $\mathbf{k}_A = (1, \sec \Theta)$, thus

$$\begin{pmatrix} T_N \\ T_F \end{pmatrix} = T_S \mathbf{k}_S + T_A \mathbf{k}_A = T_S \begin{pmatrix} 1 \\ 1 \end{pmatrix} + T_A \begin{pmatrix} 1 \\ \sec \Theta \end{pmatrix} \quad (6)$$

If we write $(\mathbf{k}_S, \mathbf{k}_A)$ as a matrix, \mathbf{K} (known as the “measurement operator” or “linearised forward model”), where

$$K_{i,j} = \frac{\partial y_i}{\partial x_j}, \quad (7)$$

then the solution for (T_S, T_A) is equivalent to

$$\begin{pmatrix} T_S \\ T_A \end{pmatrix} = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \begin{pmatrix} T_N \\ T_F \end{pmatrix}. \quad (8)$$

Exercise: convince yourself that (5) is equivalent to (8).

Aside on notation: I will use \mathbf{K} to refer to the measurement operator, for consistency with the set text for this part of the course: Clive D. Rodgers, “Inverse Methods for Atmospheric Remote Sounding”, World Scientific Publishing, 2000. Almost everyone else uses \mathbf{H} to refer to the measurement operator, and \mathbf{K} to refer to the “Kalman Gain Matrix”, which in this simple example is just $(\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T$, although it gets more complicated as we bring in more information. Clive correctly observes that Kalman did not use \mathbf{K} to refer to the Kalman Gain Matrix, but this just means Kalman wasn’t incredibly arrogant. The rest of the community also uses \mathbf{R} to refer to the observation error covariance matrix, whereas we, following Clive, will use \mathbf{S}_ϵ .

Is that all there is to it? Of course not – this example assumed exactly the same number of observations as quantities to be estimated. In reality these two numbers are very different – overconstrained and underconstrained problems. We also assumed (perhaps without realising it) equal and uncorrelated levels of noise in T_N and T_F , plus no uncertainty or non-linearity in \mathbf{K} – unpacking these assumptions likes at the heart of:

Estimation theory

You’ve probably heard of “least squares” as a method of estimation. The aim of this lecture is to discuss *why* minimising squared residuals is a good idea in terms of basic probability theory. We will talk about:

- Minimum variance estimators: the Gauss-Markov Theorem in pictures.
 - Dealing with correlated observational noise: pre-whitening operators.
 - Maximum likelihood estimators and Bayes Theorem.
-

The basic linear model:

$$\mathbf{y} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}, \quad (9)$$

where

\mathbf{y} : the rank- m vector of observations.

\mathbf{x} : the rank- n vector of quantities we wish to estimate, describing the state of the system.

\mathbf{K} : the linear(ised) measurement operator or forward model, arranged in an $m \times n$ matrix. We can think of the columns of \mathbf{K} as n patterns \mathbf{k}_j contributing to \mathbf{y} (these are known as “independent variables” in standard linear regression) with unknown amplitudes x_j . For example, the k^{th} column of \mathbf{K} might be:

- $\mathbf{k}_j = [1, 1, 1, 1, 1, \dots]$, to estimate the mean.
- $\mathbf{k}_j =$ a straight line, to estimate the gradient.
- $\mathbf{k}_j =$ a model-predicted spatio-temporal pattern of anthropogenic climate change (my job).
- $\mathbf{k}_j =$ the j^{th} column of a linearised forward model predicting satellite radiances \mathbf{y} from the atmospheric state \mathbf{x} . That is, the change in \mathbf{y} resulting from a small perturbation on x_j , normalised by the size of the perturbation (Don’s job).

$\boldsymbol{\epsilon}$: the (unobserved) noise term, comprising all sources of observation/model discrepancy, including observation errors, un-modelled variability, etc.

Key assumptions of the basic linear model: We assume that the “expected value” of $\boldsymbol{\epsilon}$ (the mean value of $\boldsymbol{\epsilon}$ which would result if we were able to repeat the observation (experiment) an infinite number of times) is zero:

$$\langle \boldsymbol{\epsilon} \rangle = \mathbf{0}. \quad (10)$$

We also assume that the covariance of $\boldsymbol{\epsilon}$ is given by a well-defined $m \times m$ matrix \mathbf{S}_ϵ :

$$\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \rangle = \mathbf{V}(\boldsymbol{\epsilon}) = \mathbf{S}_\epsilon, \quad (11)$$

so $(S_\epsilon)_{i,j} = \langle \epsilon_i \epsilon_j \rangle$.

The “ordinary least squares” estimator for \mathbf{x} : In the absence of any information on \mathbf{x} other than that provided by \mathbf{y} , the standard OLS estimator is

$$\hat{\mathbf{x}} \equiv (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{y} \equiv \hat{\mathbf{G}} \mathbf{y}. \quad (12)$$

By substitution into (9), and using the properties of $\langle \boldsymbol{\epsilon} \rangle$ and $\mathbf{V}(\boldsymbol{\epsilon})$, we have:

$$\langle \hat{\mathbf{x}} \rangle = \mathbf{x} \quad \text{and} \quad (13)$$

$$\mathbf{V}(\hat{\mathbf{x}}) \equiv \langle (\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)(\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)^T \rangle \quad (14)$$

$$= \langle (\hat{\mathbf{G}}(\mathbf{K}\mathbf{x} - \boldsymbol{\epsilon}) - \mathbf{x})(\hat{\mathbf{G}}(\mathbf{K}\mathbf{x} - \boldsymbol{\epsilon}) - \mathbf{x})^T \rangle \quad (15)$$

$$= \hat{\mathbf{G}} \mathbf{S}_\epsilon \hat{\mathbf{G}}^T. \quad (16)$$

A simple example: \mathbf{K} consists of a single column of 1s, to estimate the mean.

$$\mathbf{K}^T \mathbf{K} = m \quad (17)$$

$$\hat{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^m y_i \quad (\text{reassuringly}) \quad (18)$$

$$V(\hat{\mathbf{x}}) = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m R_{ij} \quad (19)$$

Suppose $(S_\epsilon)_{i,j} = \sigma^2 \delta_{ij}$ (i.e., un-correlated “white noise”), then

$$V(\hat{\mathbf{x}}) = \frac{\sigma^2}{m}. \quad (20)$$

But why least squares? What exactly is going on in equation (12)? Suppose \mathbf{K} consists of two vectors, \mathbf{k}_1 and \mathbf{k}_2 and ϵ is unit-variance, uncorrelated, “white noise” ($\mathbf{S}_\epsilon = \mathbf{I}$). What can we say about \mathbf{G} , a linear operator that gives us an estimate of \mathbf{x} based on observations \mathbf{y} : $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$?

Case 1: \mathbf{k}_1 and \mathbf{k}_2 orthogonal. To work out how much of \mathbf{k}_1 is required to reproduce \mathbf{y} , we “project” \mathbf{y} onto \mathbf{k}_1 , or multiply \mathbf{y} by a vector \mathbf{g}_1 , aligned with \mathbf{k}_1 with length such that $\mathbf{g}_1^T \mathbf{k}_1 = 1$. Hence, in this case, $\hat{x}_1 = \mathbf{g}_1^T \mathbf{y}$.

Case 2: \mathbf{k}_1 and \mathbf{k}_2 not orthogonal. To work out the optimal combination of \mathbf{k}_1 and \mathbf{k}_2 to reproduce \mathbf{y} , we need to project onto directions orthogonal to \mathbf{k}_2 and \mathbf{k}_1 respectively. That is

$$\mathbf{g}_1^T \mathbf{k}_2 = 0 \quad \text{and} \quad (21)$$

$$\mathbf{g}_2^T \mathbf{k}_1 = 0. \quad (22)$$

We still need

$$\mathbf{g}_1^T \mathbf{k}_1 = 1 \quad \text{and} \quad (23)$$

$$\mathbf{g}_2^T \mathbf{k}_2 = 1, \quad (24)$$

which should be clear if you think about the case $\mathbf{y} = \mathbf{k}_1 + \mathbf{k}_2$. Noting that the \mathbf{g} are the *rows* of \mathbf{G} , we can just write

$$\mathbf{G}\mathbf{K} = \mathbf{I} \quad (25)$$

as a condition for $\hat{\mathbf{x}}$ being unbiased. Notice that this condition is satisfied by the least squares estimator $\hat{\mathbf{G}} = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T$, but it would also be satisfied by lots of other estimators: what is so special about least squares?

Some points to notice: If $\mathbf{S}_\epsilon = \mathbf{I}$, then ϵ has the same expected length in all directions, and the variance in \hat{x}_1 is simply the squared length of \mathbf{g}_1 , or more generally,

$$V(\hat{\mathbf{x}}) = \mathbf{G}\mathbf{G}^T. \quad (26)$$

Hence the larger the sum squared elements in \mathbf{g}_1 , the “worse” (less accurate) the estimate of \mathbf{x}_1 : what happens to \mathbf{g}_1 as \mathbf{k}_1 and \mathbf{k}_2 move closer together (become more correlated)?

And here's the punchline: So far, everything has been represented in the plane containing \mathbf{k}_1 and \mathbf{k}_2 , but of course \mathbf{y} can have components in other directions as well, making \mathbf{G} rectangular. We can add anything we like to \mathbf{g}_1 perpendicular to the plane of the white-board, and condition (25) is still satisfied. But doing so *must* increase the length of \mathbf{g}_1 , and hence the variance in \hat{x}_1 . Hence $\hat{\mathbf{x}}$ has the least variance of all possible linear unbiased estimators for \mathbf{x} provided $\mathbf{S}_\epsilon = \mathbf{I}$, or some multiple thereof. In the jargon, $\hat{\mathbf{x}}$ is the “Best Linear Unbiased Estimator”, or BLUE.

This is the Gauss-Markov Theorem: and what neat about it is that we've proved it without making any assumptions about the shape of the distribution of ϵ in any particular direction, aside from requiring it to be the same in all directions. Hence least-squares estimators are minimum-variance even when noise is non-Gaussian.

How do we apply the Gauss-Markov theorem with correlated noise? OLS estimators are based on minimising the sum squared residuals, giving equal weight to every datapoint and every pattern in \mathbf{y} . If $\mathbf{S}_\epsilon \neq \sigma^2\mathbf{I}$, then some datapoints or combinations of datapoints (i.e. patterns, such as low-frequency Fourier modes) will contain much more variance than others. The simplest way to deal with this is to introduce a coordinate transformation or “pre-whitening operator”, \mathbf{P} , defined such that

$$\epsilon' \equiv \mathbf{P}\epsilon \tag{27}$$

and

$$\mathbf{V}\epsilon' = \langle \mathbf{P}\epsilon\epsilon^T\mathbf{P}^T \rangle \tag{28}$$

$$= \mathbf{I}. \tag{29}$$

Since \mathbf{P} is constant, we can bring it out from under the expectation operator, giving

$$\mathbf{P}\mathbf{S}_\epsilon\mathbf{P}^T = \mathbf{I} \tag{30}$$

$$\mathbf{P}^T\mathbf{P}\mathbf{S}_\epsilon\mathbf{P}^T\mathbf{P} = \mathbf{P}^T\mathbf{P}. \tag{31}$$

This is satisfied if (but *not* only if) $\mathbf{P}^T\mathbf{P} = \mathbf{S}_\epsilon^{-1}$, so substituting “pre-whitened” for original variables in equation (12) we have

$$\tilde{\mathbf{x}} \equiv (\mathbf{K}^T\mathbf{P}^T\mathbf{P}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{P}^T\mathbf{P}\mathbf{y} \tag{32}$$

$$= (\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{y}, \tag{33}$$

which you will see again. So everything we have proved about OLS estimators carries over to appropriately pre-whitened correlated noise. In particular, the variance of $\tilde{\mathbf{x}}$ is given by

$$\mathbf{V}(\tilde{\mathbf{x}}) = \tilde{\mathbf{G}}^T\mathbf{S}_\epsilon\tilde{\mathbf{G}} = (\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}, \tag{34}$$

and $\tilde{\mathbf{x}}$ is BLUE.

Exercise: Prove the Gauss-Markov Theorem using algebra rather than pictures. Hints: introduce another linear unbiased estimator, $\mathbf{x}' = \tilde{\mathbf{x}} + \mathbf{A}\mathbf{y}$, where \mathbf{A} is a linear operator and $\langle \mathbf{x}' \rangle = \mathbf{x}$, and show that $V(\mathbf{x}') - V(\tilde{\mathbf{x}}) = V(\mathbf{x}' - \tilde{\mathbf{x}})$, a positive-definite matrix.

But what about prior information on \mathbf{x} ? Prior information can be considered as just another kind of “observation”, so the arguments in favour of least-squares solutions carry over. Explicitly including prior information in this geometrical, minimum-variance, treatment gets rather fiddly, so we move to finding maximum-likelihood estimators instead.

This requires Bayes Theorem which, in our notation, looks like:

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}, \quad (35)$$

or, since $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(\mathbf{y})$ if \mathbf{K} and \mathbf{S}_ϵ are given,

$$P(\mathbf{x}|\tilde{\mathbf{x}}) = \frac{P(\tilde{\mathbf{x}}|\mathbf{x})P(\mathbf{x})}{P(\tilde{\mathbf{x}})}. \quad (36)$$

Some definitions from Rodgers (2000)

- $P(\mathbf{x})$: the *a priori* p.d.f. of the state \mathbf{x} , describing what is known about the state before the measurement is made.
- $P(\mathbf{y})$: the *a priori* p.d.f. of the measurement, describing what is known about the measurement before it is made.
- $P(\mathbf{x}, \mathbf{y})$: the joint *a priori* p.d.f. of \mathbf{x} and \mathbf{y} .
- $P(\mathbf{y}|\mathbf{x})$: the conditional p.d.f. of the measurement given the value of the state vector. This depends on the p.d.f. of experimental error and on the forward function.
- $P(\mathbf{x}|\mathbf{y})$: the conditional p.d.f. of the state given the measurement. This is what is wanted to describe our knowledge of the state after the measurement.

Considering the illustration, it is clear that the conditional and joint probabilities are related by

$$P(\mathbf{x}, \mathbf{y}) = P(\mathbf{x}|\mathbf{y})P(\mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}), \quad (37)$$

and leading immediately to Bayes' theorem:

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}. \quad (38)$$

A schematic explanation of Bayes Theorem: _____

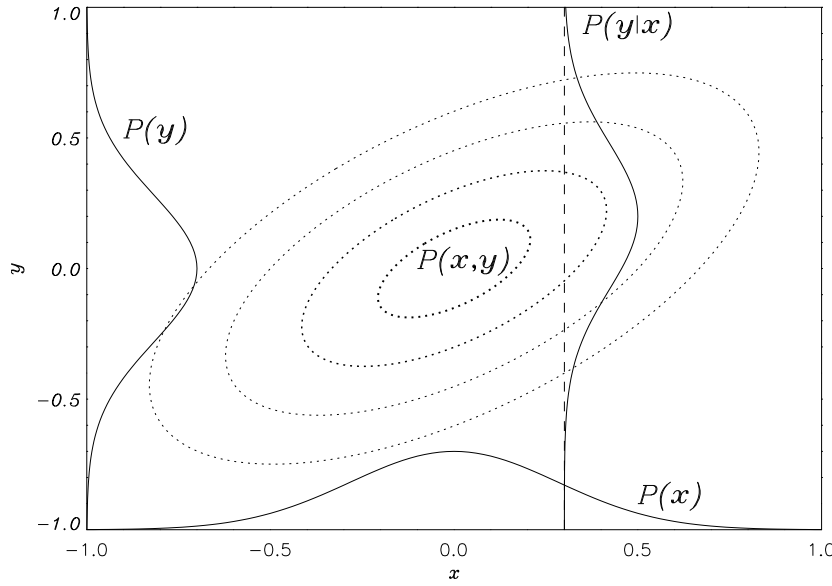


Figure 1: The dotted ellipses represent contours of the joint p.d.f. $P(x, y)$ of two scalar random variables x and y . The dashed line is a cut through $P(x, y)$ at a given value of x , and the curve labelled $P(y|x)$ represents the conditional p.d.f. of y given that value of x . The individual p.d.f.'s $P(x)$ and $P(y)$ are each obtained by integrating $P(x, y)$ over the other variable (from Rodgers, 2000).

Bayesian treatment of prior information: If all distributions are Gaussian, and the prior distributions are

$$\mathbf{x}_a \text{ priori} \sim N(\mathbf{x}_a, \mathbf{S}_a) \quad (39)$$

$$\mathbf{y}_a \text{ priori} \sim N(\mathbf{y}_a, \mathbf{V}(\mathbf{y}_a)) \quad (40)$$

$$\boldsymbol{\epsilon}_a \text{ priori} \sim N(\mathbf{0}, \mathbf{S}_\epsilon). \quad (41)$$

The distribution of, e.g. $\boldsymbol{\epsilon}$ is given by

$$P(\boldsymbol{\epsilon}) = \exp\left(-\frac{\boldsymbol{\epsilon}^T \mathbf{S}_\epsilon^{-1} \boldsymbol{\epsilon}}{2}\right) \quad (42)$$

giving an explicit log-likelihood (or “cost”) function

$$-2 \ln P(\mathbf{x}|\mathbf{y}) \equiv 2J = -2 [\ln P(\mathbf{y}|\mathbf{x}) + \ln P(\mathbf{x}) - \ln P(\mathbf{y})] \quad (43)$$

$$\begin{aligned} &= (\mathbf{y} - \mathbf{K}\mathbf{x})^T \mathbf{S}_\epsilon^{-1} (\mathbf{y} - \mathbf{K}\mathbf{x}) + \\ &\quad (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) - \\ &\quad (\mathbf{y} - \mathbf{y}_a)^T \mathbf{V}(\mathbf{y}_a)^{-1} (\mathbf{y} - \mathbf{y}_a) \end{aligned} \quad (44)$$

Differentiating once w.r.t. \mathbf{x} gives

$$\frac{\partial J}{\partial \mathbf{x}} = (\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1}) \mathbf{x} - (\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{y} + \mathbf{S}_a^{-1} \mathbf{x}_a). \quad (45)$$

J is minimized and $\ln P(\mathbf{x}|\mathbf{y})$ is maximized where $\partial J / \partial \mathbf{x} = \mathbf{0}$, or

$$\mathbf{x} = \tilde{\mathbf{x}} = (\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} (\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{y} + \mathbf{S}_a^{-1} \mathbf{x}_a). \quad (46)$$

Differentiating twice w.r.t. \mathbf{x} gives

$$\frac{\partial^2 J}{\partial \mathbf{x}^2} = \mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1}, \quad (47)$$

known as the “curvature” or “Hessian” matrix. IF (and it’s a big if), all our assumptions about Gaussian distributions are satisfied, then the variance in $\tilde{\mathbf{x}}$ is given by the inverse Hessian

$$\mathbf{V}(\tilde{\mathbf{x}}) = \left(\frac{\partial^2 J}{\partial \mathbf{x}^2} \right)^{-1} = \left(\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1} \right)^{-1}. \quad (48)$$

To see why, think about how the $P(\mathbf{x}|\mathbf{y})$ changes as we move away from the minimum: if it declines only slowly in a particular direction, the solution is poorly constrained and there is a lot of uncertainty in that direction.

Note how as $\mathbf{S}_a \rightarrow \infty$ (no prior information), $\mathbf{S}_a^{-1} \rightarrow \mathbf{0}$ and we recover equations (33) and (34).

Problem: In the ATSR problem, T_N and T_F are nadir and forward view brightness temperatures (linearizing $L(T)$ about a single reference temperature), T_S is surface brightness temperature and T_A the “atmospheric correction” term,

$$\begin{aligned} T_N &= T_S + T_A \\ T_F &= T_S + (\sec \Theta) T_A, \end{aligned}$$

where $\Theta = 55^\circ$.

1. If the standard deviations of T_N and T_F due to measurement noise are both 0.1K and the noise is uncorrelated, what is the optimal estimator for T_S and what is the standard error of this estimate?
2. If $T_N = 295\text{K}$ and $T_F = 287.5\text{K}$, what is the best estimate of T_S ? If $T_N = 279\text{K}$ and $T_F = 278.25\text{K}$, what is the best estimate of T_S ? What are the values for T_A in both cases? Comment on your result, where these observations might have been taken, and why the atmospheric corrections are very different.
3. If, because of thermal inertia for example, errors in T_N and T_S are found to be correlated with $\rho = 0.5$ (they aren’t, but for the sake of argument) but with s.d.s still 0.1K, what is the new optimal estimator for T_S and what is its standard error? Has it increased or decreased because of the correlation? Why?
4. Setting $\rho = 0$ again, suppose you have some additional information (from a weather forecast model, for example) that suggests that $T_A = -5\text{K}$ with a standard error of 0.5K. If $T_N = 295\text{K}$ and $T_F = 287.5\text{K}$, what is your new best estimate of T_S and what is its standard error?