Hands-On 1: Appendix

Konstantinos Chatzilygeroudis

29 October 2026

Read these appendices after the session. They give the full derivations and further analysis behind the six sections of the handout. The optional notebook contains numerical exercises on the same topics.

Appendices E–G use the true fault labels of the operational run: probe 7 is faulty from snapshot 200 onwards. We use them there only to explain what happened in this worked example. The main notebook never uses these labels for the diagnosis; it reveals them only in the final validation (Section 6 of the handout).

Contents

A Calibration as least squares (Lectures 1 and 5) 3
A.1 One probe . . . 3
A.2 Reading the residuals . . . 4
A.3 All probes at once . . . 5
A.4 From counts to temperature . . . 5

B Ill-conditioning, ridge and priors (Lectures 2 and 6) 6
B.1 Two reference points that are almost the same . . . 6
B.2 Plain ridge: stable, but where? . . . 7
B.3 Least squares is ML; ridge is MAP . . . 8
B.4 A physical prior: the rest of the rack . . . 8
B.5 Bias and variance over repeated short recalibrations . . . 9

C The covariance across probes (Lectures 4 and 5) 10
C.1 Estimating it from residuals . . . 10
C.2 Eigendecomposition and principal coordinates . . . 10
C.3 Estimated covariances overstate distances . . . 11

D The χ^2 distribution 12
D.1 Definition and first properties . . . 12
D.2 Why the threshold must grow with the dimension . . . 13

E Ellipses, whitening and detection (Lecture 4) 13
E.1 Covariance ellipses . . . 14
E.2 Whitening, and how to check it honestly . . . 14
E.3 The Mahalanobis distance and its distribution . . . 14
E.4 Why the Mahalanobis distance sees what a z-score misses . . . 15
E.5 Which probe? Leave one out . . . 17

F One temperature with an honest error bar (Lecture 5) **17**
F.1 The best linear combination 17
F.2 Why averaging does not beat a shared error 18
F.3 Honest and dishonest error bars 18

G Without a reference thermometer **19**

Notation.

$k = 1, \dots, N$; $j = 1, \dots, C$	snapshot index; probe index ($N = 400$, $C = 10$)
T_k	reference temperature of snapshot k [°C]
y_{kj}; $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$	count of probe j in snapshot k; all counts, one column per probe
a_j, b_j; $\boldsymbol{\theta}_j = [a_j,\, b_j]^\top$	offset [counts] and gain [counts/°C] of probe j
$\boldsymbol{X} \in \mathbb{R}^{N \times 2}$, $p = 2$	design matrix with rows $[1,\ T_k]$; parameters per probe
$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_C]$	all coefficients, $2 \times C$
$\boldsymbol{n}_k \in \mathbb{R}^C$; $\boldsymbol{\Sigma}$	noise of snapshot k across the probes; its covariance in °C^2
$\boldsymbol{E} \in \mathbb{R}^{N \times C}$	calibration residuals converted to °C
$\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$	the projection onto the column space of $\boldsymbol{X}$ (Lecture 1)
χ^2_ν; $\chi^2_\nu(q)$	chi-squared distribution with ν degrees of freedom; its q-quantile

The noise model. Each probe's error, measured in °C, is the sum of three parts: one that every probe shares (the ADC reference and supply rail), one that varies along the rack (the fan), and the probe's own noise. Errors are independent from one snapshot to the next, but *not* from one probe to the next within a snapshot:

$$\boldsymbol{n}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \ \text{ independently over } k, \qquad \boldsymbol{\Sigma} \text{ not diagonal.} \tag{0.1}$$

In counts, probe j's noise is b_j times its noise in °C.

A Calibration as least squares (Lectures 1 and 5)

Calibration means finding, for each probe, the line that turns its counts into degrees. The calibration run gives 400 snapshots at known temperatures, which is far more data than two unknowns per probe need, so we fit the line by least squares.

A.1 One probe

The LM35-style model of Lecture 1, written with counts as output, is

$$y_{kj} = a_j + b_j\, T_k + n_{kj} \quad \Longleftrightarrow \quad \boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\theta}_j + \boldsymbol{n}_j, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & T_1 \\ \vdots & \vdots \\ 1 & T_N \end{bmatrix}. \tag{A.1}$$

Here a_j is the count the probe would give at 0 °C, and b_j is the number of counts per degree.

Remark A.1 (Which variable goes on which side)**.** Least squares assumes that the noise is in $\boldsymbol{y}$ and that $\boldsymbol{X}$ is known exactly. The reference thermometer is far more accurate than the probes, so the reference temperatures go in $\boldsymbol{X}$ and the counts go in $\boldsymbol{y}$. We fit counts against temperature and invert the line afterwards (Section A.4). Fitting temperature against counts would put the noisy quantity inside $\boldsymbol{X}$, and the slope would be biased towards zero. The key word for that problem is *errors-in-variables*.

Least squares gives $\hat{\boldsymbol{\theta}}_j = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}_j$ (the normal equations, Lecture 1). With $S_{TT} = \sum_k (T_k - \bar{T})^2$, the 2×2 inverse can be written out:

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{1}{S_{TT}} \begin{bmatrix} S_{TT}/N + \bar{T}^2 & -\bar{T} \\ -\bar{T} & 1 \end{bmatrix}. \tag{A.2}$$

To check it, multiply by $\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} N & N\bar{T} \\ N\bar{T} & \sum_k T_k^2 \end{bmatrix}$ and use $\sum_k T_k^2 = S_{TT} + N\bar{T}^2$.

Proposition A.2 (Error bars of an affine calibration)**.** *If the noise of probe j is white with variance σ_j^2 (in counts), then* $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_j) = \sigma_j^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ *(Lecture 5). In particular,*

$$\mathrm{Var}(\hat{b}_j) = \frac{\sigma_j^2}{S_{TT}}, \qquad \mathrm{Var}(\hat{a}_j) = \sigma_j^2\Big(\frac{1}{N} + \frac{\bar{T}^2}{S_{TT}}\Big), \tag{A.3}$$

and the fitted line at temperature T, $\boldsymbol{x}_T^\top\hat{\boldsymbol{\theta}}_j$ with $\boldsymbol{x}_T = [1,\, T]^\top$, has

$$\mathrm{Var}\left(\boldsymbol{x}_T^\top\hat{\boldsymbol{\theta}}_j\right) = \sigma_j^2\, h(T), \qquad h(T) = \boldsymbol{x}_T^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_T = \frac{1}{N} + \frac{(T-\bar{T})^2}{S_{TT}}. \tag{A.4}$$

The proof is (A.2) read entry by entry, together with $\mathrm{Var}(\boldsymbol{x}_T^\top\hat{\boldsymbol{\theta}}) = \boldsymbol{x}_T^\top\,\mathrm{Cov}(\hat{\boldsymbol{\theta}})\boldsymbol{x}_T$. The number $h(T)$ is called the *leverage* of T. It is smallest at the centre of the calibration range and grows quadratically away from it.

Reading the formulas. The gain is determined by how *spread out* the reference temperatures are. The offset is the value of the line at $T = 0$, which lies outside the calibration range, so it inherits the leverage of $T = 0$. For the rack ($N = 400$, baths at 10, 20, 30 and 40 °C, $\bar{T} = 25\,°\mathrm{C}$, $S_{TT} = 400 \cdot 125 = 50\,000\,°\mathrm{C}^2$) and probe 3's noise $s_3 = 8.2$ counts, this gives ± 0.037 counts/°C and ± 1.0 count.

A.2 Reading the residuals

Proposition A.3 (Properties of the residuals)**.** *Let $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}$ with $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{n}$, and*

$$\mathbb{E}[\boldsymbol{r}] = \boldsymbol{0}, \qquad \mathrm{Cov}(\boldsymbol{r}) = \sigma^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}), \qquad \mathbb{E}\|\boldsymbol{r}\|^2 = (N-p)\,\sigma^2.$$

So $s^2 = \|\boldsymbol{r}\|^2/(N-p)$ is an unbiased estimate of σ^2 (Lecture 5), and residual k has variance $\sigma^2(1 - h(T_k))$, slightly less than σ^2.

Optional: proof

$\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ and $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}$, so $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{n}) = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{n}$. Its mean is zero. Its covariance is $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\,\sigma^2\boldsymbol{I}\,(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})^\top = \sigma^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})$, because $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$ is a symmetric projection, so $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})^2 = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$. The diagonal entries of $\boldsymbol{P}_{\boldsymbol{X}}$ are the leverages $h(T_k)$. Finally, $\mathbb{E}\|\boldsymbol{r}\|^2 = \mathrm{tr}\,\mathrm{Cov}(\boldsymbol{r}) = \sigma^2\,\mathrm{tr}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})$, and $\mathrm{tr}\,\boldsymbol{P}_{\boldsymbol{X}} = \mathrm{tr}\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right) = \mathrm{tr}\,\boldsymbol{I}_p = p$.

Proposition A.3 says what healthy residuals look like: no systematic pattern, and a spread of about σ. The notebook reads them in three ways.

1. **Against T.** A curve would show as the mean residual of each bath moving away from zero. Small bath-mean residuals support the affine model on these data.
2. **Their distribution.** If the noise is Gaussian, the standardised residuals r_{kj}/s_j are approximately $\mathcal{N}(0,1)$, and 4.6% of them should exceed 2 in absolute value. The rack gives 4.4%. This matters because "least squares is maximum likelihood" (Section B) relies on Gaussian noise.
3. **Across probes.** The residuals of probes 1 and 2 have correlation 0.95. That is the cross-probe structure of (0.1), studied in Section C.

A.3 All probes at once

All probes share the same $\boldsymbol{X}$, because they see the same baths. So the ten problems are the columns of one matrix equation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{\mathcal{N}}, \qquad \hat{\boldsymbol{\Theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}. \tag{A.5}$$

Stacking the columns of $\boldsymbol{Y}$ into one long vector $\operatorname{vec}\boldsymbol{Y} \in \mathbb{R}^{CN}$ gives one big least-squares problem. Its design matrix is block diagonal,

$$\operatorname{vec}\boldsymbol{Y} = (\boldsymbol{I}_C \otimes \boldsymbol{X})\operatorname{vec}\boldsymbol{\Theta} + \operatorname{vec}\boldsymbol{\mathcal{N}}, \qquad \boldsymbol{I}_C \otimes \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X} & & \\ & \ddots & \\ & & \boldsymbol{X} \end{bmatrix}, \tag{A.6}$$

where $\otimes$ is the Kronecker product: $\boldsymbol{A}\otimes\boldsymbol{B}$ replaces each entry A_{ij} by the block $A_{ij}\boldsymbol{B}$. Its normal matrix is block diagonal too, so the big problem splits into the C small ones.

The noise of different probes is correlated. In counts, $\operatorname{Cov}(\boldsymbol{n}_k^{\text{counts}}) = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B} =: \boldsymbol{\Sigma}_y$ with $\boldsymbol{B} = \operatorname{diag}(b_j)$. Should we then use weighted least squares (Lecture 5) with the full covariance $\boldsymbol{\Sigma}_y \otimes \boldsymbol{I}_N$ of $\operatorname{vec}\boldsymbol{\mathcal{N}}$? It turns out to make no difference.

Proposition A.4 (Shared design matrix: WLS equals per-probe LS)**.** *If every probe has the same design matrix $\boldsymbol{X}$ and $\operatorname{Cov}(\operatorname{vec}\boldsymbol{\mathcal{N}}) = \boldsymbol{\Sigma}_y \otimes \boldsymbol{I}_N$ with $\boldsymbol{\Sigma}_y \succ 0$, then the weighted least-squares estimate of $\operatorname{vec}\boldsymbol{\Theta}$ equals the column-by-column least-squares estimate* (A.5)*.*

Optional: proof of Proposition A.4

Use the Kronecker rules $(\boldsymbol{A}\otimes\boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top$, $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = \boldsymbol{AC}\otimes\boldsymbol{BD}$ and $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$. With $\boldsymbol{\mathcal{X}} = \boldsymbol{I}_C \otimes \boldsymbol{X}$ and $\boldsymbol{W} = (\boldsymbol{\Sigma}_y \otimes \boldsymbol{I}_N)^{-1} = \boldsymbol{\Sigma}_y^{-1} \otimes \boldsymbol{I}_N$,

$$\boldsymbol{\mathcal{X}}^\top \boldsymbol{W}\boldsymbol{\mathcal{X}} = (\boldsymbol{I}_C \otimes \boldsymbol{X}^\top)(\boldsymbol{\Sigma}_y^{-1} \otimes \boldsymbol{I}_N)(\boldsymbol{I}_C \otimes \boldsymbol{X}) = \boldsymbol{\Sigma}_y^{-1} \otimes \boldsymbol{X}^\top\boldsymbol{X}, \qquad \boldsymbol{\mathcal{X}}^\top\boldsymbol{W} = \boldsymbol{\Sigma}_y^{-1}\otimes\boldsymbol{X}^\top.$$

The WLS estimate is therefore

$$(\boldsymbol{\Sigma}_y \otimes (\boldsymbol{X}^\top\boldsymbol{X})^{-1})(\boldsymbol{\Sigma}_y^{-1}\otimes\boldsymbol{X}^\top)\operatorname{vec}\boldsymbol{Y} = (\boldsymbol{I}_C \otimes (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\operatorname{vec}\boldsymbol{Y},$$

which applies $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ to each column of $\boldsymbol{Y}$ separately. $\boldsymbol{\Sigma}_y$ has cancelled.

The same algebra gives the covariance of all the coefficients:

$$\operatorname{Cov}(\operatorname{vec}\hat{\boldsymbol{\Theta}}) = \boldsymbol{\Sigma}_y \otimes (\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{A.7}$$

The coefficients of *different* probes are correlated, exactly as their noises are. We return to this in Remark F.6.

Key idea

The correlation between probes does not change the calibration coefficients. It changes everything we do *with* the calibrated readings: how to combine them, how uncertain they are, and what counts as an outlier.

A.4 From counts to temperature

Inverting the calibration line gives each probe's temperature reading,

$$\hat{T}_j = \frac{y_j - \hat{a}_j}{\hat{b}_j}. \tag{A.8}$$

If the coefficients were exact, $\hat{T}_j = T + n_j$, with n_j the noise in °C. The calibration residuals in °C are therefore $e_{kj} = r_{kj}/\hat{b}_j$, the matrix $\boldsymbol{E}$ of Section C.

To see what calibration buys, compare it with a datasheet line (offset 200 counts and gain 40 counts/°C for every probe). In the calibration data that line leaves each probe with a mean error between −1.4 and +1.4 °C. After calibration the mean errors are zero, and what is left is the noise, about 0.2 °C per probe.

B Ill-conditioning, ridge and priors (Lectures 2 and 6)

A full calibration needs several bath temperatures, and each one takes time to settle. It is therefore tempting to check a serviced probe at one convenient temperature. This section asks what such a check can tell us, and how prior knowledge from the rest of the rack can supply what it cannot.

B.1 Two reference points that are almost the same

Probe 3 receives only a short recalibration: $n = 10$ readings at 25.0 °C and $n = 10$ at 25.2 °C. *With one bath only*, both columns of $\boldsymbol{X}$ are multiples of $\mathbf{1}$, so $\operatorname{rank}\boldsymbol{X} = 1$, and $\boldsymbol{X}^\top\boldsymbol{X}$ has a zero eigenvalue. Only the combination $a + 25\,b$ is identifiable.

With both baths, $\operatorname{rank}\boldsymbol{X} = 2$, but $S_{TT} = 20 \cdot 0.1^2 = 0.2\,°\mathrm{C}^2$ instead of 50 000. By (A.3),

$$\operatorname{sd}(\hat{b}) = \frac{8.2}{\sqrt{0.2}} = 18.3\ \tfrac{\text{counts}}{°\text{C}}, \qquad \operatorname{sd}(\hat{a}) = 8.2\sqrt{\tfrac{1}{20} + \tfrac{25.1^2}{0.2}} = 460 \text{ counts}. \tag{B.1}$$

These are 500 and 460 times larger than for the rack calibration. The coefficients explode: $\hat{a} = -578$ and $\hat{b} = 69.5$, against 162 and 40.0 from the full recalibration. Yet the fitted line still passes through the data.

The lever arm. Equation (A.4), divided by the gain, gives the uncertainty of the calibration line in °C:

$$\operatorname{sd}\big(\text{line at } T\big) = \frac{\sigma}{b}\sqrt{\frac{1}{n} + \frac{(T - \bar{T})^2}{S_{TT}}}. \tag{B.2}$$

At $\bar{T} = 25.1$ °C only the first term counts, which gives ±0.06 °C: the short recalibration pins the line down *where it was measured*. At 10 or 40 °C the second term dominates, $15/\sqrt{0.2} \approx 34$, and the line is uncertain by about ±7 °C. The slope is a lever, and a short lever magnifies every error at its end (Figure 1).

The sampling distribution. Proposition A.2 is a statement about repeated experiments. The notebook simulates 5000 short recalibrations with probe 3's rack-calibration coefficients as the truth. The fitted gains are centred correctly, because least squares is unbiased, but they spread with standard deviation about 18 counts/°C. In 1.7% of the checks the fitted gain is negative.

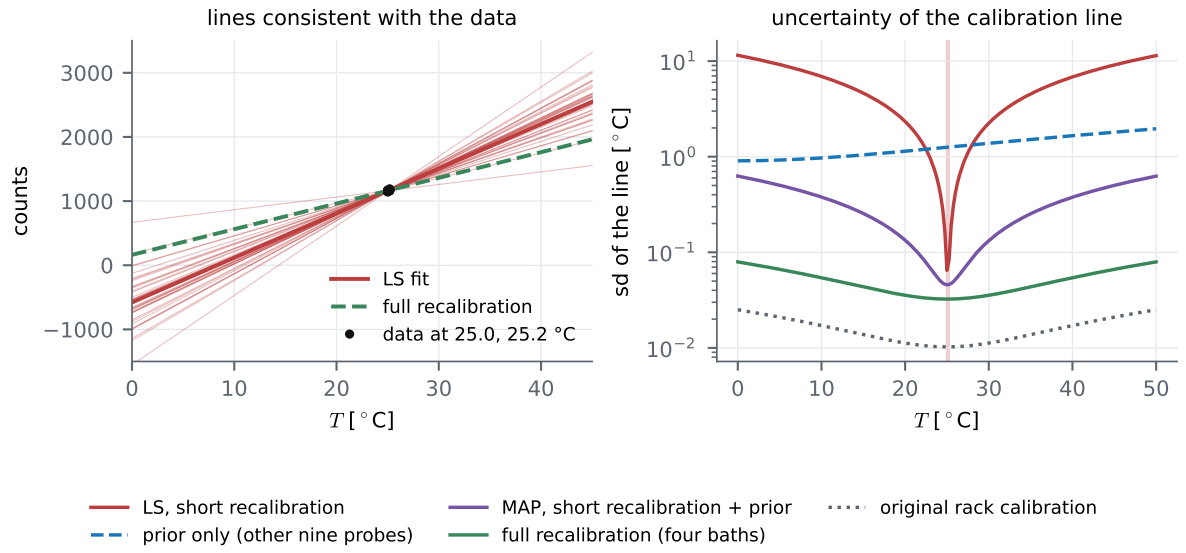


Figure 1: The short recalibration of probe 3. Left: 40 lines drawn from $\mathcal{N}(\hat{\boldsymbol{\theta}}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$. All of them fit the data at 25 °C, but they fan out across the range where the probe will be used. Right: the standard deviation (B.2) of the calibration line for each estimator of this section (log scale). The red band marks the two short-recalibration baths.

The SVD view. The singular values of $\boldsymbol{X}$ are $s_1 = 112.3$ and $s_2 = 0.018$, so $\kappa(\boldsymbol{X}) = s_1/s_2 \approx 6300$ and $\kappa(\boldsymbol{X}^\top\boldsymbol{X}) = \kappa(\boldsymbol{X})^2 \approx 4 \times 10^7$ (Lecture 2). The right singular vector of s_2 is

$$\boldsymbol{v}_2 \approx \pm\frac{1}{\sqrt{626}}\begin{bmatrix}25\\-1\end{bmatrix}.$$

Moving $\boldsymbol{\theta}$ along $\boldsymbol{v}_2$ raises the gain and lowers the offset so that predictions near 25 °C do not change. The data cannot see this direction, and noise is amplified along it by the factor $1/s_2$.

B.2 Plain ridge: stable, but where?

Ridge regression solves $(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\top\boldsymbol{y}$ (Lecture 2).

Proposition B.1 (Ridge through the SVD)**.** *With the thin SVD* $\boldsymbol{X} = \sum_i s_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$,

$$\hat{\boldsymbol{\theta}}_\lambda = \sum_i f_i\,\frac{\boldsymbol{u}_i^\top\boldsymbol{y}}{s_i}\,\boldsymbol{v}_i, \qquad f_i = \frac{s_i^2}{s_i^2+\lambda}. \tag{B.3}$$

Optional: proof

In the basis $\boldsymbol{v}_1, \boldsymbol{v}_2$ we have $\boldsymbol{X}^\top\boldsymbol{X} = \sum_i s_i^2\boldsymbol{v}_i\boldsymbol{v}_i^\top$ and $\boldsymbol{X}^\top\boldsymbol{y} = \sum_i s_i(\boldsymbol{u}_i^\top\boldsymbol{y})\boldsymbol{v}_i$. The ridge matrix becomes $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I} = \sum_i(s_i^2+\lambda)\boldsymbol{v}_i\boldsymbol{v}_i^\top$, whose inverse is $\sum_i(s_i^2+\lambda)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top$. Multiply by $\boldsymbol{X}^\top\boldsymbol{y}$.

Least squares has every filter factor $f_i = 1$. Ridge damps each direction by f_i. For $s_2^2 \ll \lambda \ll s_1^2$, roughly $10^{-2} \lesssim \lambda \lesssim 10^2$ here, the weak direction is switched off ($f_2 \approx 0$) and the strong one is untouched ($f_1 \approx 1$). Then $\hat{\boldsymbol{\theta}}_\lambda \approx (\boldsymbol{u}_1^\top\boldsymbol{y}/s_1)\boldsymbol{v}_1$, which is exactly the minimum-norm solution of the one-bath problem. This is the plateau of the ridge path in the notebook, at $a \approx 2$ and $b \approx 46$. Ridge has picked the shortest $\boldsymbol{\theta}$ that fits $a + 25b \approx 1160$. It is stable, and it is wrong by about 2.4 °C at 10 and 40 °C.

B.3 Least squares is ML; ridge is MAP

Proposition B.2 (Lectures 5 and 6)**.** *Let* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{n}$ *with* $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

1. *The maximum-likelihood estimate is the least-squares estimate.*
2. *With the prior* $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1})$*, the posterior is Gaussian with precision and mean*

$$\boldsymbol{P} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{\Lambda}_0, \qquad \hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \boldsymbol{P}^{-1}\Big(\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0\Big), \tag{B.4}$$

   and $\mathrm{Cov}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \boldsymbol{P}^{-1}$.

3. *Plain ridge with parameter* λ *is the case* $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\boldsymbol{\Lambda}_0 = \boldsymbol{I}/\sigma_0^2$, *with* $\lambda = \sigma^2/\sigma_0^2$.

Optional: proof

$-2 \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2/\sigma^2 + \text{const}$, so maximising the likelihood minimises the squared residual: this is part 1. For part 2, add $-2 \log p(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0 (\boldsymbol{\theta} - \boldsymbol{\mu}_0) + \text{const}$. The sum is a quadratic in $\boldsymbol{\theta}$ with Hessian $2\boldsymbol{P}$. Setting its gradient to zero gives (B.4), and completing the square shows that the posterior is Gaussian with precision $\boldsymbol{P}$ (Lecture 6). For part 3, multiply (B.4) by σ^2: with $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\boldsymbol{\Lambda}_0 = \boldsymbol{I}/\sigma_0^2$ it becomes $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\top \boldsymbol{y}$.

Part 3 says what plain ridge *claimed*: before seeing any data, that the offset and the gain are both near zero, with the same spread, even though they are measured in different units. For a probe whose counts are in the hundreds, that claim is false, and the answer inherits it.

B.4 A physical prior: the rest of the rack

The other nine probes are the same model from the same batch. Their fitted coefficients are nine draws of "what (a, b) looks like for this kind of probe". Their sample mean and standard deviations give

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 201 \\ 39.87 \end{bmatrix}, \qquad \boldsymbol{\Lambda}_0 = \mathrm{diag}\Big(\frac{1}{36^2}, \ \frac{1}{1.39^2}\Big).$$

This prior is a physical claim. In the language of part 3 it has one λ per parameter, $\lambda_a = \sigma^2/\sigma_{0,a}^2 = 0.051$ and $\lambda_b = \sigma^2/\sigma_{0,b}^2 = 34.8$. They differ by a factor of almost 700, so no single λ could have been right. This is Lecture 6's warning that $\lambda\|\boldsymbol{\theta}\|^2$ adds quantities in different units. With $\sigma = s_3 = 8.2$ counts, (B.4) gives

$$\hat{a}_{\mathrm{MAP}} = 180 \pm 25 \text{ counts}, \qquad \hat{b}_{\mathrm{MAP}} = 39.3 \pm 1.0 \text{ counts}/^\circ\mathrm{C},$$

within one posterior standard deviation of the full recalibration (162, 40.0). The calibration line is now within 0.3 °C of the full recalibration everywhere from 10 to 40 °C. Its own error bar, from $\boldsymbol{P}^{-1}$ in (A.4), is ±0.38 °C at the ends of the range (Figure 1, right).

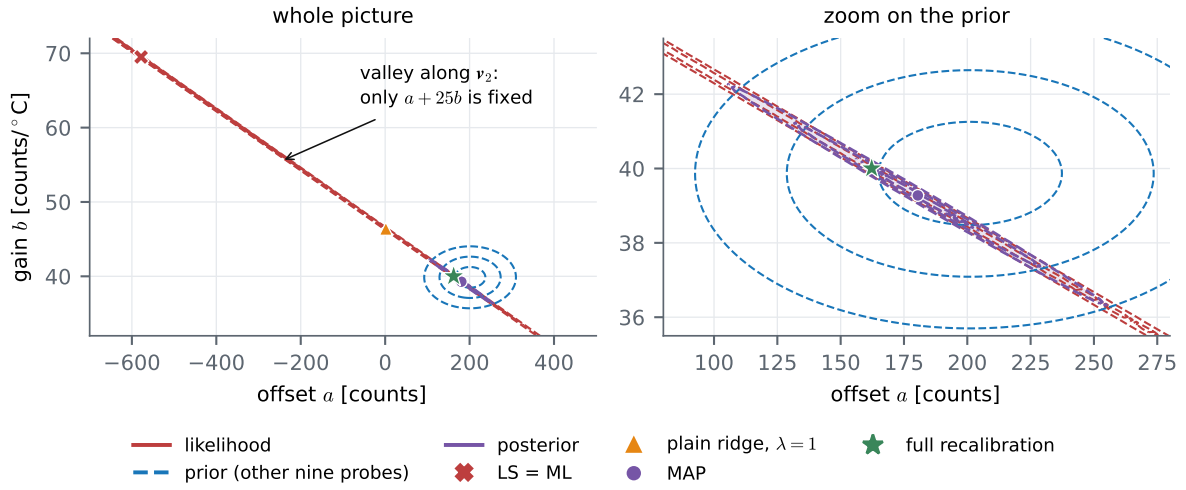


Figure 2: The short recalibration of probe 3 in the (a, b) plane. The likelihood (red) is a long valley along the weak direction $\boldsymbol{v}_2$. The prior from the other nine probes (blue) is wide in a different direction. The posterior (purple, shaded) is the small region where both agree. Contours at $\Delta \log = -0.5, -2, -4.5$, i.e. 1, 2 and 3 standard deviations.

Remark B.3 (Why the posterior is so much smaller than either factor)**.** Precisions add, as in (B.4). The likelihood's precision is huge along $\boldsymbol{v}_1$ and almost zero along $\boldsymbol{v}_2$. The prior's precision is moderate in every direction. Their sum is large in every direction. This is the matrix form of Lecture 6's "the prior can fill the data's blind spots". Here the posterior standard deviation of b (1.0) is set mostly by the prior (1.39), while $a + 25b$ is set by the data.

B.5 Bias and variance over repeated short recalibrations

Proposition B.4 (Mean and covariance of the MAP estimate)**.** *For a fixed true* $\boldsymbol{\theta}$ *and noise* $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$,

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}] = \boldsymbol{\theta} + \boldsymbol{P}^{-1}\boldsymbol{\Lambda}_0(\boldsymbol{\mu}_0 - \boldsymbol{\theta}), \qquad \mathrm{Cov}(\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}) = \boldsymbol{P}^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}\boldsymbol{P}^{-1}.$$

Optional: proof

Substitute $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{n}$ into (B.4): $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \boldsymbol{P}^{-1}\big(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}\boldsymbol{\theta} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\big) + \boldsymbol{P}^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{n}}{\sigma^2}$. Write $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} = \boldsymbol{P} - \boldsymbol{\Lambda}_0$ in the first term. It becomes $\boldsymbol{\theta} + \boldsymbol{P}^{-1}\boldsymbol{\Lambda}_0(\boldsymbol{\mu}_0 - \boldsymbol{\theta})$, which is the mean. The second term has covariance $\boldsymbol{P}^{-1}\frac{\boldsymbol{X}^\top(\sigma^2\boldsymbol{I})\boldsymbol{X}}{\sigma^4}\boldsymbol{P}^{-1}$.

The bias is the prior's pull, $\boldsymbol{P}^{-1}\boldsymbol{\Lambda}_0$, applied to how wrong the prior mean is, $\boldsymbol{\mu}_0 - \boldsymbol{\theta}$. A prior centred on the truth gives no bias. A prior centred far away gives a large one. The notebook's simulation measures the error of the calibration line at 40 °C:

estimator	bias [°C]	sd [°C]	rms error [°C]
least squares	+0.12	6.90	6.90
plain ridge, $\lambda = 1$	+2.40	0.07	2.40
MAP, rack prior	−0.28	0.06	0.29

Keep this distinction in mind

The posterior standard deviation at 40 °C is 0.38 °C, but the MAP estimate varies over repeated checks by only 0.06 °C. These answer different questions. The sampling spread holds the prior fixed and asks how much the estimate moves with the noise. The posterior also admits that we do not know how far probe 3 is from the average probe. That is exactly where the 0.28 °C bias comes from, and the posterior's 0.38 covers it.

Remark B.5 (Key words)**.** Building the prior from the other members of a population is called *empirical Bayes*. Its fully Bayesian version, which also puts a prior on $\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_0$, is a *hierarchical model*.

C The covariance across probes (Lectures 4 and 5)

The calibration gives each probe's line, and Section A showed that the correlation between probes does not change it. Everything that follows does depend on that correlation: combining the ten readings into one temperature, putting an error bar on it, and telling a faulty probe from ordinary shared noise. So we need the covariance $\boldsymbol{\Sigma}$ of the probe errors. The calibration residuals provide 400 samples of it.

C.1 Estimating it from residuals

Lecture 4 estimated a covariance from centred repeats with divisor $N-1$. The calibration residuals are centred by a fit with p parameters per probe, and the right divisor becomes $N-p$.

Proposition C.1 (Unbiased covariance from regression residuals)**.** *Let $\boldsymbol{R} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Theta}}$, and suppose the rows of the noise matrix are independent with covariance $\boldsymbol{\Sigma}_y$. Then $\mathbb{E}[\boldsymbol{R}^\top \boldsymbol{R}] = (N-p)\,\boldsymbol{\Sigma}_y$, so*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-p}\boldsymbol{E}^\top \boldsymbol{E} \tag{C.1}$$

estimates $\boldsymbol{\Sigma}$ without bias, up to the tiny error from using $\hat{b}_j$ instead of b_j.

Optional: proof

As in Proposition A.3, $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\mathcal{N}}$, so $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{\mathcal{N}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\mathcal{N}}$. Its (i,j) entry is $\sum_{k,l}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})_{kl}\,\mathcal{N}_{ki}\mathcal{N}_{lj}$. Rows are independent with mean zero, so $\mathbb{E}[\mathcal{N}_{ki}\mathcal{N}_{lj}] = \delta_{kl}(\boldsymbol{\Sigma}_y)_{ij}$, and the expectation is $\operatorname{tr}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\,(\boldsymbol{\Sigma}_y)_{ij} = (N-p)(\boldsymbol{\Sigma}_y)_{ij}$. Dividing column j by b_j converts counts to °C.

For the rack, every correlation in $\hat{\boldsymbol{\Sigma}}$ lies between 0.56 and 0.95. It is far from diagonal.

C.2 Eigendecomposition and principal coordinates

Write $\hat{\boldsymbol{\Sigma}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_C$ and orthonormal eigenvectors $\boldsymbol{u}_i$ (Lecture 4).

Proposition C.2 (Principal coordinates)**.** *If $\operatorname{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, the coordinates $\boldsymbol{p} = \boldsymbol{U}^\top\boldsymbol{e}$ (that is, $p_i = \boldsymbol{u}_i^\top\boldsymbol{e}$) satisfy $\operatorname{Cov}(\boldsymbol{p}) = \boldsymbol{\Lambda}$: they are uncorrelated, and $\operatorname{Var}(p_i) = \lambda_i$. Moreover $\sum_i \lambda_i = \operatorname{tr}\boldsymbol{\Sigma} = \sum_j \operatorname{Var}(e_j)$.*

The proof is $\boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{\Lambda}$, and the trace is unchanged by the rotation. So $\lambda_i / \sum_l \lambda_l$ is the *share of the total variance* along pattern $\boldsymbol{u}_i$. For the rack,

$$\lambda = (0.373,\ 0.042,\ 0.0050,\ \ldots,\ 0.0019)\ ^\circ\mathrm{C}^2,$$

and the first two patterns carry 85% and 9% of the variance (Figure 3).

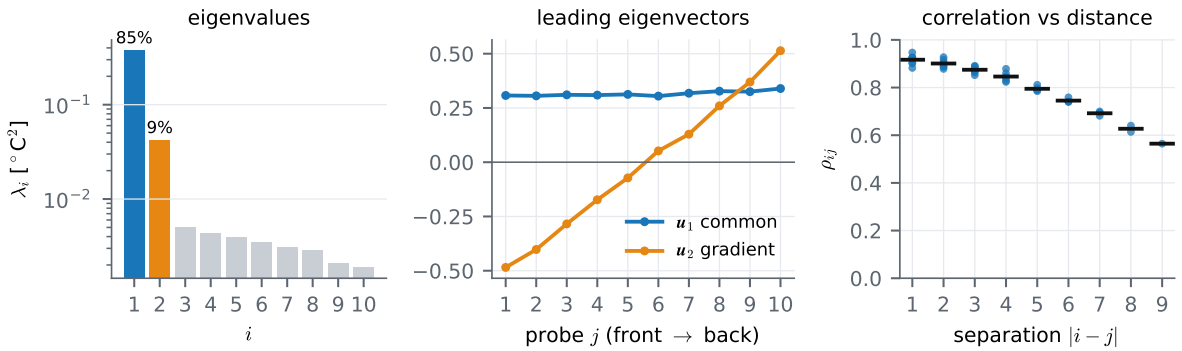


Figure 3: Left: eigenvalues of $\hat{\boldsymbol{\Sigma}}$ (log scale), with the share of the total variance for the two leading ones. Centre: the two leading eigenvectors, a common mode $\boldsymbol{u}_1$ and a front-to-back gradient $\boldsymbol{u}_2$. Right: correlation of each pair of probes against their distance along the rack; bars are means.

A model that explains the picture. Lecture 4's model of a shared disturbance is $\boldsymbol{n} = \boldsymbol{L}\boldsymbol{w}$ with $\mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{I}$, so $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. Take

$$\boldsymbol{n} = \sigma_c\,\mathbf{1}\,w_c + \sigma_g\,\boldsymbol{g}\,w_g + \boldsymbol{\epsilon}, \qquad \boldsymbol{\Sigma} = \sigma_c^2\mathbf{1}\mathbf{1}^\top + \sigma_g^2\boldsymbol{g}\boldsymbol{g}^\top + \boldsymbol{D}, \tag{C.2}$$

with a common mode, a gradient pattern $\boldsymbol{g}$ that runs linearly from -1 at the front to $+1$ at the back, and independent noise with $\boldsymbol{D}$ diagonal. Since $\mathbf{1}^\top\boldsymbol{g} = 0$, when $\boldsymbol{D}$ is small the leading eigenvectors are close to $\mathbf{1}/\sqrt{C}$ and $\boldsymbol{g}/\|\boldsymbol{g}\|$, with

$$\lambda_1 \approx C\sigma_c^2, \qquad \lambda_2 \approx \sigma_g^2\|\boldsymbol{g}\|^2.$$

From the data, $\sigma_c \approx 0.19\,°\mathrm{C}$ and $\sigma_g \approx 0.10\,°\mathrm{C}$. The remaining eigenvalues lie between 0.002 and $0.005\,°\mathrm{C}^2$. They are the independent noise, roughly $0.06\,°\mathrm{C}$ per probe. The model also predicts the correlations:

$$\rho_{ij} = \frac{\sigma_c^2 + \sigma_g^2 g_i g_j}{\sqrt{(\sigma_c^2 + \sigma_g^2 g_i^2 + D_{ii})(\sigma_c^2 + \sigma_g^2 g_j^2 + D_{jj})}}.$$

Neighbours have $g_i g_j > 0$, so the gradient adds to their correlation. The two ends have $g_1 g_{10} = -1$, so the gradient *subtracts*: $\rho_{1,10} \approx (0.037 - 0.010)/(0.037 + 0.010 + 0.003) \approx 0.53$, against 0.56 measured. This is why the correlation fades with distance but never drops below about 0.5 (Figure 3, right).

C.3 Estimated covariances overstate distances

Remark C.3. If $\hat{\boldsymbol{\Sigma}}$ is estimated from ν degrees of freedom and $\boldsymbol{e}$ is a new, independent draw from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}\big[\boldsymbol{e}^\top\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{e}\big] = C\,\frac{\nu}{\nu - C - 1}.$$

We quote this result without proof (key words: *Wishart distribution*, *inverse Wishart*). The reason is that the smallest eigenvalues of a sample covariance tend to come out too small, and $\hat{\boldsymbol{\Sigma}}^{-1}$ magnifies exactly those directions. Here $\nu = N - p = 398$ and $C = 10$, so the inflation is only about 3% on average. A particular $\hat{\boldsymbol{\Sigma}}$ can be off by more, and the notebook's average held-out variance of 1.10 combines this with sampling scatter. With only 100 calibration snapshots ($\nu = 98$) the average inflation would already be 13%.

D The χ^2 distribution

The rest of the session reduces a vector of errors to one number, "how far is this snapshot from normal?", and then has to decide *how large is too large*. For a single Gaussian variable we already know the answer.

The z-score. If a reading x has mean μ and standard deviation σ, its *z-score* is

$$z = \frac{x - \mu}{\sigma},$$

the number of standard deviations by which x differs from its mean. If x is Gaussian, then $z \sim \mathcal{N}(0, 1)$, whatever μ and σ are, so one table serves every variable: $|z| > 1.96$ happens 5% of the time and $|z| > 2.576$ happens 1% of the time. The standardised residuals r_{kj}/s_j of Section A are z-scores. For a probe error, the mean is zero when the probe is healthy, so its z-score is e_j/σ_j. A 1% test flags $|z| > 2.576$, and a healthy probe is flagged 1% of the time. This section extends that answer to many variables at once.

D.1 Definition and first properties

Definition D.1 (Chi-squared distribution)**.** If $z_1, \dots, z_\nu$ are independent standard normal variables, the distribution of

$$d^2 = z_1^2 + \cdots + z_\nu^2 = \|\boldsymbol{z}\|^2$$

is the *chi-squared distribution with ν degrees of freedom*, χ^2_ν. Its q-quantile, the value exceeded with probability $1 - q$, is written $\chi^2_\nu(q)$.

So d^2 is the squared distance from the origin of a point drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_\nu)$, a perfectly round Gaussian cloud. For $\nu = 1$ it is just z^2, and the 1% test $|z| > 2.576$ is the test $z^2 > 2.576^2 = 6.63 = \chi^2_1(0.99)$.

Proposition D.2 (Mean and variance)**.** *If $d^2 \sim \chi^2_\nu$, then $\mathbb{E}[d^2] = \nu$ and $\mathrm{Var}(d^2) = 2\nu$.*

Optional: proof

For a standard normal z, $\mathbb{E}[z^2] = 1$. Integrating by parts with $\varphi'(z) = -z\varphi(z)$, where φ is the standard normal density, gives $\mathbb{E}[z^4] = \int z^3 \cdot z\varphi(z)\,dz = \int 3z^2\varphi(z)\,dz = 3$. So $\mathrm{Var}(z^2) = 3 - 1 = 2$. The ν terms are independent, so their means add and their variances add.

A χ^2_ν value is therefore typically $\nu \pm \sqrt{2\nu}$; for $\nu = 10$ that is 10 ± 4.5. Each coordinate contributes about 1 on average, so the total grows with the number of coordinates.

Proposition D.3 (Two degrees of freedom)**.** *If $d^2 \sim \chi^2_2$, then $P(d^2 > x) = e^{-x/2}$ for $x \geq 0$. Hence $\chi^2_2(q) = -2\ln(1 - q)$; for example $\chi^2_2(0.95) = 5.99$.*

Optional: proof

The density of (z_1, z_2) is $\frac{1}{2\pi}e^{-(z_1^2+z_2^2)/2}$. In polar coordinates, with $\rho^2 = z_1^2 + z_2^2$ and $dz_1\,dz_2 = \rho\,d\rho\,d\phi$,

$$P(\rho^2 > x) = \int_0^{2\pi}\int_{\sqrt{x}}^{\infty} \frac{1}{2\pi} e^{-\rho^2/2}\rho\,d\rho\,d\phi = \Big[-e^{-\rho^2/2}\Big]_{\sqrt{x}}^{\infty} = e^{-x/2}.$$

For other ν there is no such simple formula, and we use tables or software. In Python, `scipy.stats.chi2.ppf(q, nu)` gives $\chi^2_\nu(q)$ and `scipy.stats.chi2.sf(x, nu)` gives $P(d^2 > x)$.

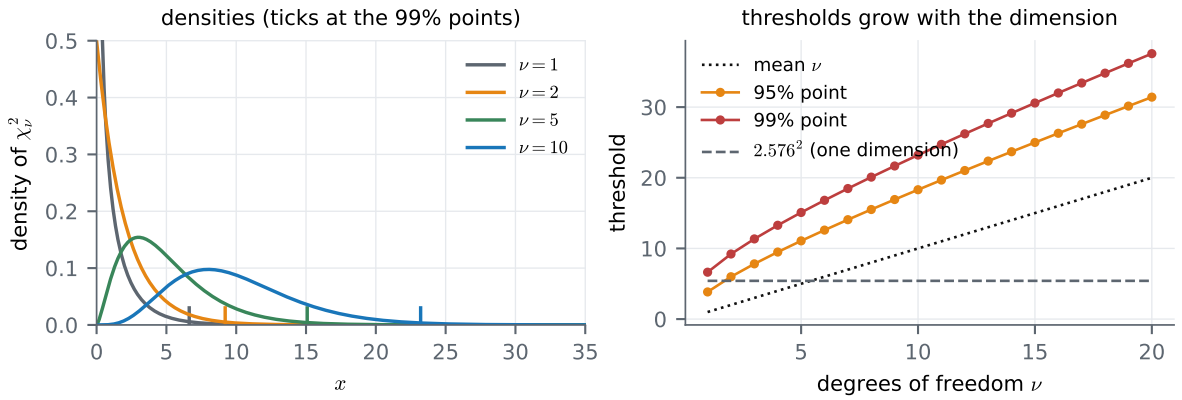


Figure 4: Left: densities of χ^2_ν; the short ticks mark the 99% points. As ν grows, the distribution moves to the right and widens. Right: the 95% and 99% points against ν, compared with the mean ν and with the one-dimensional 99% point 2.576^2.

D.2 Why the threshold must grow with the dimension

Figure 4 (right) shows the 99% point growing from 6.63 at $\nu = 1$ to 23.21 at $\nu = 10$. Reusing the one-dimensional threshold in ten dimensions would raise far too many alarms. The same effect appears if we test each coordinate separately. Even for ten *independent* probes, each tested at 1%, the probability that at least one raises an alarm when nothing is wrong is

$$1 - 0.99^{10} \approx 9.6\%.$$

A single χ^2_ν test looks at all coordinates together and keeps the overall false-alarm rate at exactly the level we choose.

Key idea

The χ^2_ν recipe applies to a sum of ν squared variables that are *independent* and *standard normal*. Raw probe errors are neither: they are correlated and have different variances. Whitening (Section E) is exactly the step that turns them into such variables.

Quantiles used in the session.

	$\chi^2_1(0.99)$	$\chi^2_2(0.95)$	$\chi^2_8(0.99)$	$\chi^2_9(0.99)$	$\chi^2_{10}(0.99)$
value	6.63	5.99	20.09	21.67	23.21
use	one variable	ellipses	no reference, LOO	no reference	with reference

LOO is leave one out (Section E.5); “no reference” is Section G.

E Ellipses, whitening and detection (Lecture 4)

We now turn to the validation run and ask, for every snapshot, whether the ten readings are consistent with a healthy rack. The answer must account for the shared noise: when all probes move together, that is normal, but when one probe moves against the others, something is wrong.

E.1 Covariance ellipses

For two probes, the block $\boldsymbol{\Sigma}_2$ of $\boldsymbol{\Sigma}$ defines the ellipses $\boldsymbol{e}^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{e} = c^2$. Their axes point along the eigenvectors of $\boldsymbol{\Sigma}_2$, with half-lengths $c\sqrt{\lambda_i}$ (Lecture 4). By Proposition E.2 below, applied in two dimensions, $\boldsymbol{e}^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{e} \sim \chi^2_2$. By Proposition D.3, the ellipse then holds a fraction $1 - e^{-c^2/2}$ of the draws:

c	1	2	3	$\sqrt{5.99} = 2.45$
fraction inside	39%	86%	99%	95%

The "one-sigma" ellipse holds only 39%, not the 68% of a one-sigma interval in one dimension. The notebook counts the calibration residuals inside each ellipse and finds 40%, 88% and 98% for probes 1 and 2.

E.2 Whitening, and how to check it honestly

With $\hat{\boldsymbol{\Sigma}} = \boldsymbol{L}\boldsymbol{L}^\top$ (Cholesky), the whitened vector $\boldsymbol{q}$ solves $\boldsymbol{L}\boldsymbol{q} = \boldsymbol{e}$, so $\mathrm{Cov}(\boldsymbol{q}) = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}\boldsymbol{L}^{-\top} = \boldsymbol{I}$ (Lecture 4).

Proposition E.1 (Two whiteners, one length)**.** *Let $\boldsymbol{W}$ be any matrix with $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{I}$, for example $\boldsymbol{L}^{-1}$ or $\boldsymbol{W}_{\mathrm{sym}} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top$. Then $\|\boldsymbol{W}\boldsymbol{e}\|^2 = \boldsymbol{e}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}$ for every $\boldsymbol{e}$. Two such whiteners differ by an orthogonal matrix.*

Optional: proof

From $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{I}$, the matrix $\boldsymbol{W}$ is invertible and $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1}\boldsymbol{W}^{-\top}$, so $\boldsymbol{\Sigma}^{-1} = \boldsymbol{W}^\top\boldsymbol{W}$ and $\|\boldsymbol{W}\boldsymbol{e}\|^2 = \boldsymbol{e}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{e} = \boldsymbol{e}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}$. If $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ both whiten, $\boldsymbol{O} = \boldsymbol{W}_2\boldsymbol{W}_1^{-1}$ satisfies $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{W}_2\boldsymbol{\Sigma}\boldsymbol{W}_2^\top = \boldsymbol{I}$.

The coordinates of $\boldsymbol{q}$ depend on the whitener, but its length does not. Only the length is used below.

Keep this distinction in mind

Whitening the *same* residuals that produced $\hat{\boldsymbol{\Sigma}}$ returns exactly $\boldsymbol{I}$, for any data. With $\boldsymbol{Q} = \boldsymbol{E}\boldsymbol{L}^{-\top}$,

$$\frac{\boldsymbol{Q}^\top\boldsymbol{Q}}{N-p} = \boldsymbol{L}^{-1}\frac{\boldsymbol{E}^\top\boldsymbol{E}}{N-p}\boldsymbol{L}^{-\top} = \boldsymbol{L}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^{-\top} = \boldsymbol{I}.$$

This checks the code, not the model. The model is checked on data that $\hat{\boldsymbol{\Sigma}}$ has never seen.

On held-out data, $\mathrm{Cov}(\boldsymbol{q})$ will not be exactly $\boldsymbol{I}$, for two reasons. The first is ordinary sampling scatter: with M draws, about $\pm 1/\sqrt{M}$ for off-diagonal entries and $\pm\sqrt{2/M}$ for diagonal ones. The second is that $\hat{\boldsymbol{\Sigma}}$ is an estimate (Remark C.3). On the first 200 snapshots of the validation run, the diagonal of $\mathrm{Cov}(\boldsymbol{q})$ averages 1.10.

E.3 The Mahalanobis distance and its distribution

In the validation run, the reference thermometer gives T, so each snapshot provides the error vector $\boldsymbol{e} = \hat{\boldsymbol{t}} - T\boldsymbol{1} \in \mathbb{R}^C$. Its Mahalanobis distance is

$$d^2 = \boldsymbol{e}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e} = \|\boldsymbol{q}\|^2, \qquad \boldsymbol{L}\boldsymbol{q} = \boldsymbol{e}. \tag{E.1}$$

Proposition E.2. *If the rack is healthy, $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_C)$ and $d^2 \sim \chi^2_C$.*

Proof. $\boldsymbol{q} = \boldsymbol{L}^{-1}\boldsymbol{e}$ is a linear map of a Gaussian vector, so it is Gaussian, with covariance $\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\boldsymbol{L}^{-\top} = \boldsymbol{I}$. Uncorrelated jointly Gaussian variables are independent, so $d^2 = \sum_i q_i^2$ is a sum of C independent squared standard normals. □

Proposition E.2 gives a test with a known false-alarm rate: flag a snapshot when $d^2 > \chi^2_C(0.99) = 23.21$, and a healthy rack is flagged 1% of the time. Over the first 200 snapshots, the notebook measures a mean d^2 of 11.0, against $\nu = 10$, and an alarm rate of 1.5%. Both are slightly high, which is consistent with Remark C.3.

Remark E.3 (Many snapshots, many tests)**.** A 1% test applied to 300 healthy snapshots raises about three alarms. A single alarm is therefore weak evidence. What signals a fault is a *rate* far above 1%, or alarms that persist. That is how the notebook reads the results.

E.4 Why the Mahalanobis distance sees what a z-score misses

The obvious detector tests each probe on its own: compute its z-score $z_j = e_j/\sqrt{\Sigma_{jj}}$ and flag $|z_j| > 2.576$. Suppose probe j develops an offset δ, so $\boldsymbol{e} \sim \mathcal{N}(\delta\,\boldsymbol{\epsilon}_j, \boldsymbol{\Sigma})$, where $\boldsymbol{\epsilon}_j$ is the j-th unit vector.

Proposition E.4 (What a single-probe offset does)**.** *1. The probe's own z-score shifts by $\delta/\sqrt{\Sigma_{jj}}$, measured using its* total *variance Σ_{jj}.*

2. The mean Mahalanobis distance is

$$\mathbb{E}[d^2] = C + \delta^2(\boldsymbol{\Sigma}^{-1})_{jj}.$$

Moreover,

$$\frac{1}{(\boldsymbol{\Sigma}^{-1})_{jj}} = \operatorname{Var}\left(e_j \mid e_i,\ i \neq j\right), \tag{E.2}$$

the variance of probe j that remains once the other probes are known.

Optional: proof of part 2

Write $\boldsymbol{e} = \delta\boldsymbol{\epsilon}_j + \boldsymbol{e}_0$ with $\boldsymbol{e}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then $\mathbb{E}[d^2] = \mathbb{E}[\boldsymbol{e}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_0] + \delta^2\boldsymbol{\epsilon}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}_j = C + \delta^2(\boldsymbol{\Sigma}^{-1})_{jj}$, since the cross term has mean zero. For (E.2), order the probes so that j comes last and partition $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{s} \\ \boldsymbol{s}^\top & \Sigma_{jj} \end{bmatrix}$. The block-inverse formula gives the last diagonal entry of $\boldsymbol{\Sigma}^{-1}$ as $1/(\Sigma_{jj} - \boldsymbol{s}^\top\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{s})$. The denominator is the variance of the error of the best linear prediction of e_j from the others, which for a Gaussian is the conditional variance.

Definition E.5 (Noncentral χ^2)**.** If $\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{I}_\nu)$, the distribution of $\|\boldsymbol{q}\|^2$ is the *noncentral chi-squared distribution* with ν degrees of freedom and noncentrality $\|\boldsymbol{m}\|^2$. In Python it is `scipy.stats.ncx2`.

After whitening, an offset δ on probe j shifts the mean of $\boldsymbol{q}$ by $\boldsymbol{L}^{-1}\delta\boldsymbol{\epsilon}_j$, whose squared length is $\delta^2(\boldsymbol{\Sigma}^{-1})_{jj}$. So d^2 is noncentral χ^2_C with that noncentrality, and the probability of an alarm can be computed for any δ.

For probe 7, the total standard deviation is 0.21 °C, but the conditional one is only 0.07 °C: the other nine probes predict most of its noise. An offset $\delta = 0.3$ °C is 1.5 total standard deviations but 4.3 conditional ones. The z-test at 1% therefore fires with probability 13%, while d^2 has noncentrality $4.3^2 \approx 18.4$ and exceeds 23.21 with probability 68%. The notebook observes 14% and 70%. To catch half of the faults, the z-test needs $\delta \approx 0.53$ °C, and the Mahalanobis test needs $\delta \approx 0.26$ °C (Figure 6, left).

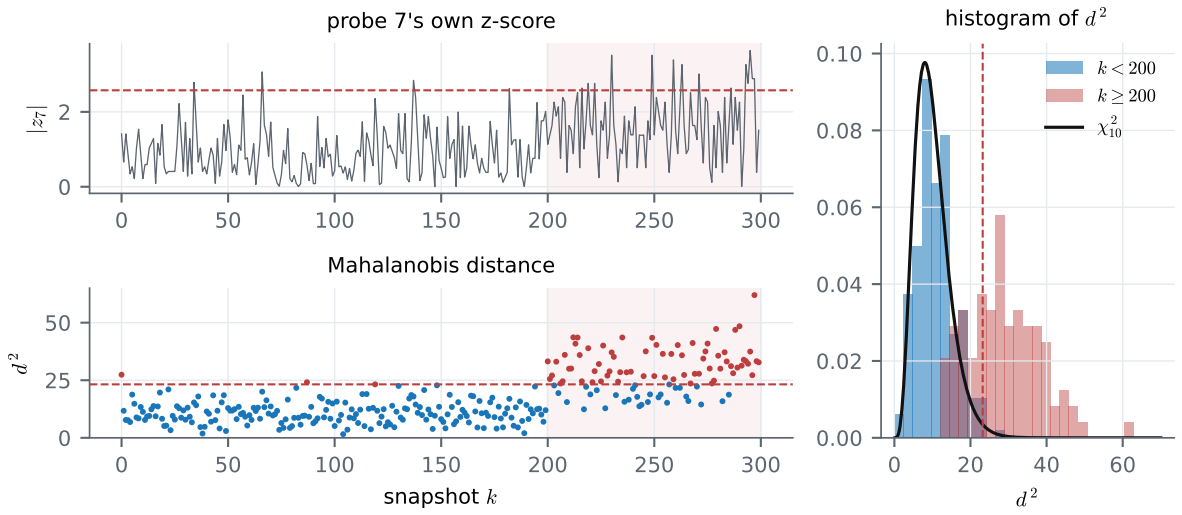


Figure 5: Validation run. Left: probe 7's own z-score barely reacts to its 0.3 °C fault from snapshot 200 (shaded). The Mahalanobis distance crosses $\chi^2_{10}(0.99)$ (dashed) in 1.5% of the healthy snapshots and in 70% of the faulty ones. Right: before the fault, the histogram of d^2 follows the χ^2_{10} density; after it, the histogram moves to the right.

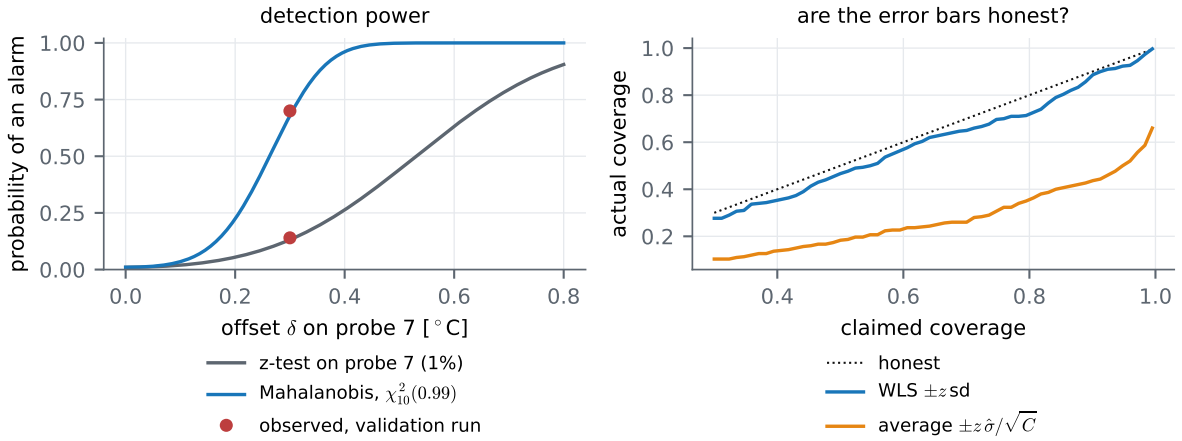


Figure 6: Left: probability of an alarm against the size of an offset on probe 7, for the per-probe z-test and for the Mahalanobis test, from Proposition E.4; dots are the rates observed after snapshot 200. Right: claimed against actual coverage of $\pm z \cdot$ sd intervals for the reported temperature (Section F).

Key idea

A z-score compares a probe with its own spread. The Mahalanobis distance compares it with *what the other probes say it should read*. When noise is shared, the second spread is much smaller, so faults stand out.

E.5 Which probe? Leave one out

Proposition E.6. *If $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{e}_{-j}$ removes component j, then $\boldsymbol{e}_{-j} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{-j})$, where $\boldsymbol{\Sigma}_{-j}$ removes row and column j. Hence $\boldsymbol{e}_{-j}^\top \boldsymbol{\Sigma}_{-j}^{-1} \boldsymbol{e}_{-j} \sim \chi^2_{C-1}$.*

The first claim holds because $\boldsymbol{e}_{-j} = \boldsymbol{S}\boldsymbol{e}$ for a selection matrix $\boldsymbol{S}$, and $\mathrm{Cov}(\boldsymbol{S}\boldsymbol{e}) = \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}^\top$ picks out the block. The second claim is Proposition E.2 in $C-1$ dimensions. If probe j is the faulty one, removing it leaves nine healthy probes, and the alarms drop to the 1% level. Removing any other probe leaves the fault in, and the alarms remain. In the notebook, removing probe 7 clears 97% of the alarms after snapshot 200, and removing any other probe clears at most 7%.

The fault, found

Probe 7 picks up an offset of about $+0.3\,^\circ$C at **snapshot 200**: its mean error is $-0.04\,^\circ$C before and $+0.32\,^\circ$C after. That is 1.5 of its own standard deviations, so its own z-test catches only 14% of the faulty snapshots. The Mahalanobis test catches 70%, because given the other nine probes the fault is 4.3 standard deviations. Leave-one-out then names the probe: removing probe 7 clears 97% of the alarms, and removing any other probe clears at most 7%. Section G finds the same probe without the reference thermometer.

F One temperature with an honest error bar (Lecture 5)

Ten readings of the same temperature are redundant. The user of the rack wants one number per snapshot, with an error bar that can be trusted.

F.1 The best linear combination

Each healthy probe reads $\hat{T}_j = T + e_j$ with $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. This is a linear model with a single parameter T and design matrix $\mathbf{1}$, and Lecture 5's weighted least squares applies directly.

Proposition F.1 (Combining correlated probes)**.** *Among all estimates $\hat{T} = \boldsymbol{w}^\top \hat{\boldsymbol{t}}$ that are unbiased ($\boldsymbol{w}^\top \mathbf{1} = 1$), the variance $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$ is smallest for*

$$\boldsymbol{w} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}}, \qquad \mathrm{Var}(\hat{T}) = \frac{1}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}}. \tag{F.1}$$

This is the WLS estimate $(\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{t}}$.

Optional: proof

Let $\boldsymbol{w}_\star$ be the weights in (F.1), and let $\boldsymbol{w}$ be any other unbiased weights. Write $\boldsymbol{w} = \boldsymbol{w}_\star + \boldsymbol{d}$ with $\mathbf{1}^\top \boldsymbol{d} = 0$. Then $\boldsymbol{d}^\top \boldsymbol{\Sigma} \boldsymbol{w}_\star = \boldsymbol{d}^\top \mathbf{1}/(\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}) = 0$, so $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} = \boldsymbol{w}_\star^\top \boldsymbol{\Sigma} \boldsymbol{w}_\star + \boldsymbol{d}^\top \boldsymbol{\Sigma} \boldsymbol{d} \geq \boldsymbol{w}_\star^\top \boldsymbol{\Sigma} \boldsymbol{w}_\star$. Finally, $\boldsymbol{w}_\star^\top \boldsymbol{\Sigma} \boldsymbol{w}_\star = \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1})^2 = 1/(\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1})$.

Example F.2 (Two probes and a negative weight)**.** Take two probes with standard deviations $\sigma_1 = 0.1\,^\circ$C and $\sigma_2 = 0.2\,^\circ$C and correlation $\rho = 0.8$. For two probes, (F.1) reduces to

$$w_1 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \qquad w_2 = 1 - w_1, \qquad \mathrm{Var}(\hat{T}) = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Here the denominator is $0.01 + 0.04 - 0.032 = 0.018$, so $w_1 = 0.024/0.018 = 4/3$ and $w_2 = -1/3$. The combination has standard deviation $\sqrt{0.0004 \cdot 0.36/0.018} = 0.089\,^\circ$C, better than the better probe alone. The noisy probe gets a negative weight: most of its error is shared with probe 1, so subtracting a third of it removes part of probe 1's error. In general $w_2 < 0$ exactly when $\rho > \sigma_1/\sigma_2$.

The rack behaves the same way. With probe 7 removed, three probes get negative weights (1, 8 and 10), and probe 6 gets almost half of the weight. Probes 1 and 10 sit at the two ends of the fan gradient, and their best use is as references for subtracting it.

Proposition F.3 (When the plain average is optimal)**.** *If* $\mathbf{1}$ *is an eigenvector of* $\boldsymbol{\Sigma}$*, then* $\boldsymbol{w} = \mathbf{1}/C$*: the plain average is the best combination.*

Proof. If $\boldsymbol{\Sigma}\mathbf{1} = \mu\mathbf{1}$, then $\boldsymbol{\Sigma}^{-1}\mathbf{1} = \mathbf{1}/\mu$, and (F.1) gives $\boldsymbol{w} = \mathbf{1}/C$. □

In the model (C.2), $\mathbf{1}$ is an eigenvector when every probe has the same independent noise and the gradient is balanced ($\mathbf{1}^\top \boldsymbol{g} = 0$). The real rack breaks both conditions: the probes' own noise levels differ, and removing probe 7 unbalances the gradient. That is why the weights are uneven.

F.2 Why averaging does not beat a shared error

Proposition F.4 (The common-mode floor)**.** *If* $\boldsymbol{\Sigma} = \sigma_c^2 \mathbf{1}\mathbf{1}^\top + \boldsymbol{\Sigma}_r$ *with* $\boldsymbol{\Sigma}_r \succeq 0$*, then every unbiased combination has* $\mathrm{Var}(\hat{T}) \geq \sigma_c^2$*, however many probes are used.*

Proof. For $\boldsymbol{w}^\top \mathbf{1} = 1$, $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} = \sigma_c^2 (\boldsymbol{w}^\top \mathbf{1})^2 + \boldsymbol{w}^\top \boldsymbol{\Sigma}_r \boldsymbol{w} = \sigma_c^2 + \boldsymbol{w}^\top \boldsymbol{\Sigma}_r \boldsymbol{w} \geq \sigma_c^2$. □

Here $\sigma_c \approx 0.19\,^\circ\mathrm{C}$, and (F.1) gives $\mathrm{sd}(\hat{T}) = 0.190\,^\circ\mathrm{C}$. The WLS estimate is right at the floor. The plain average is barely worse, with standard deviation $\sqrt{\mathbf{1}^\top \boldsymbol{\Sigma} \mathbf{1}}/C = 0.193\,^\circ\mathrm{C}$.

F.3 Honest and dishonest error bars

Keep this distinction in mind

The naive error bar $\hat{\sigma}/\sqrt{C} = 0.070\,^\circ\mathrm{C}$ treats the probes as independent votes. The true variance of the plain average is $\mathbf{1}^\top \boldsymbol{\Sigma} \mathbf{1}/C^2$, which includes every covariance.

An error bar can be tested whenever the truth is known. If it is honest, the interval $\pm z \cdot \mathrm{sd}$ with $z = \Phi^{-1}((1+q)/2)$ covers the truth in a fraction q of the cases, for every q. Against the reference thermometer (Figure 6, right), the naive bar claims 95% and covers 50%, and claims 68% and covers 26%. The WLS bar covers 92% and 64%. It is slightly below the diagonal because $\hat{\boldsymbol{\Sigma}}$ is slightly too small (Remark C.3).

Remark F.5 (What the fault does to the reported temperature)**.** If probe 7 is *not* removed, its offset shifts the ten-probe WLS estimate by $w_7 \delta = 0.169 \times 0.3 \approx 0.05\,^\circ\mathrm{C}$. That is a quarter of the error bar, too small to notice in the readings. So the fault is far more visible to the outlier test than in the reported temperature. A slowly growing fault would eventually bias every reading, which is why it pays to catch it early.

Remark F.6 (Uncertainty in the calibration itself)**.** The reported standard deviation treats $\hat{a}_j, \hat{b}_j$ as exact. To first order, a coefficient error adds $-[(\hat{a}_j - a_j) + (\hat{b}_j - b_j)T]/b_j = -\boldsymbol{x}_T^\top (\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)/b_j$ to $\hat{T}_j$. By (A.7), the covariance of these additions across the probes is $h(T)\,\boldsymbol{\Sigma}$, where $h(T)$ is the leverage (A.4). The full covariance is therefore $(1+h(T))\boldsymbol{\Sigma}$. For the rack, $h = 0.0025$ at 25 °C and 0.007 at 10 or 40 °C, so including it would widen the error bar by less than 0.4%. That is negligible here, but it would not be after a short recalibration like the one in Section B. This error is also *fixed* once the calibration is done: it does not average out over snapshots.

G Without a reference thermometer

Scope. The calculations below concern snapshot alarms. The main session adds a batch alarm-count check and a unique-candidate check before excluding a probe; isolated alarms are insufficient.

The fault detector of Section E compares the readings with the reference thermometer. In normal operation there is no reference thermometer: T is unknown, so $\boldsymbol{e}$ is not available. The natural substitute is the residual about the WLS estimate,

$$\boldsymbol{r} = \hat{\boldsymbol{t}} - \hat{T}\,\mathbf{1} = (\boldsymbol{I} - \mathbf{1}\boldsymbol{w}^\top)\hat{\boldsymbol{t}} = (\boldsymbol{I} - \mathbf{1}\boldsymbol{w}^\top)\boldsymbol{e}, \tag{G.1}$$

where the last step uses $(\boldsymbol{I} - \mathbf{1}\boldsymbol{w}^\top)\mathbf{1} = \mathbf{0}$: the unknown T cancels.

Proposition G.1 (One degree of freedom for the estimated temperature)**.** *If* $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{w}$ *is given by* (F.1)*, then* $\boldsymbol{r}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r} \sim \chi^2_{C-1}$.

Optional: proof

Whiten: $\boldsymbol{q} = \boldsymbol{L}^{-1}\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{g} = \boldsymbol{L}^{-1}\mathbf{1}$. Since $\boldsymbol{w}^\top \boldsymbol{e} = \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e} / (\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}) = \boldsymbol{g}^\top \boldsymbol{q} / (\boldsymbol{g}^\top \boldsymbol{g})$,

$$\boldsymbol{L}^{-1}\boldsymbol{r} = \boldsymbol{q} - \boldsymbol{g}\,\frac{\boldsymbol{g}^\top \boldsymbol{q}}{\boldsymbol{g}^\top \boldsymbol{g}} = (\boldsymbol{I} - \boldsymbol{\Pi}_{\boldsymbol{g}})\boldsymbol{q},$$

where $\boldsymbol{\Pi}_{\boldsymbol{g}}$ is the orthogonal projection onto $\boldsymbol{g}$ (Lecture 1). So $\boldsymbol{r}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r} = \|(\boldsymbol{I} - \boldsymbol{\Pi}_{\boldsymbol{g}})\boldsymbol{q}\|^2$. Choose an orthonormal basis whose first vector is $\boldsymbol{g}/\|\boldsymbol{g}\|$. In it, $\boldsymbol{q}$ still has independent standard normal coordinates, because a rotation of $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is still $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The projection removes exactly the first coordinate. What remains is a sum of $C - 1$ squared standard normals.

This is the whitened version of Lecture 5's "each fitted parameter absorbs about one σ^2": fitting T uses up one of the C dimensions. The operational test flags a snapshot when $\boldsymbol{r}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r} > \chi^2_{C-1}(0.99) = 21.67$, and leave-one-out then compares with $\chi^2_{C-2}(0.99) = 20.09$. On the validation run this test, which uses no operational reference temperature, flags 1.5% of the healthy snapshots and 70% of the faulty ones, the same rates as the test with the reference.