

Matrix inverse and log-determinant derivatives

Ten problems on differentiating through a matrix inverse and a determinant: the differential of A^{-1} , the trace trick that reads a gradient off a differential, Jacobi's formula for $\det$ and $\log \det$, the chain rule through a parametrised matrix, backprop through a linear solve and through an inverse, the Sherman–Morrison and determinant lemmas for a rank-one update, $\log \det$ through a Cholesky factor, the Gaussian log-likelihood gradient with respect to that factor, and $\log \det$ through the eigenvalues, with worked solutions and the mistakes that put the factors in the wrong order.

BEFORE YOU START

A matrix inverse and a determinant appear wherever a Gaussian does: the log-likelihood of a covariance, the marginal likelihood of a Gaussian process, the entropy of a VAE's posterior, a Kalman update. Differentiating through them by entries is hopeless, but there is a method that makes every one of these a few lines: write the *differential* of the expression, move everything into a trace, and read the gradient off the trace. These ten problems build that method from the differential of A^{-1} and Jacobi's formula for $\det A$, then use it for what a framework actually has to do, which is backpropagate through a linear solve and an inverse, and for the parametrisations that appear in practice: a rank-one update, a Cholesky factor, and the eigenvalues. The five mistakes at the end each give an answer of the right shape: the inverse differentiated like a reciprocal, a transpose lost in the trace trick, $\det$ confused with $\log \det$, a Cholesky gradient that lands above the diagonal, and a chain rule through $LL^\top$ that counts L once.

- Gradients follow the conventions of the matrix-calculus page: for a scalar $f(X)$ with $X \in \mathbb{R}^{m \times n}$, $\nabla_X f$ has the shape of X and $(\nabla_X f)_{ij} = \partial f / \partial X_{ij}$.
- The Frobenius inner product of two matrices of the same shape is $\langle A, B \rangle = \text{tr}(A^\top B) = \sum_{ij} A_{ij} B_{ij}$. The trace is linear, $\text{tr}(A^\top) = \text{tr}(A)$, and it is cyclic: $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$ whenever the products are defined. Cyclic is not commutative: $\text{tr}(ABC) \neq \text{tr}(ACB)$ in general.
- The differential dX of a matrix variable is an arbitrary small change of the same shape. For a scalar f , the differential df is linear in dX , and if $df = \langle C, dX \rangle$ for every dX then $\nabla_X f = C$: this is the *trace trick*, and it is proved in Problem 2. Derivatives of products follow the product rule, $d(AB) = (dA)B + A dB$, in that order, because matrices do not commute.
- The derivative of f along a direction E is $\frac{d}{dt} f(A + tE)$ at $t = 0$; for a scalar f it equals $\langle \nabla_A f, E \rangle$.
- A is a square invertible matrix, $A^{-\top}$ is $(A^{-1})^\top = (A^\top)^{-1}$, and $\log \det A$ is taken where $\det A > 0$. I is the identity.

- L is a scalar loss, and the upstream gradient of an intermediate quantity is written with g or G : for $x = A^{-1}b$, $g = \nabla_x L$; for $Y = A^{-1}$, $G = \nabla_Y L$.
- A Cholesky factor L (the letter is reused, as it is everywhere) is lower triangular with positive diagonal and $\Sigma = LL^\top$; its free parameters are the $n(n+1)/2$ entries on and below the diagonal. $\text{tril}(M)$ keeps the entries of M on and below the diagonal and zeroes the rest.
- The multivariate-Gaussian page showed $\nabla_\Sigma \log \det \Sigma = \Sigma^{-1}$ for symmetric Σ and, for the Gaussian log-likelihood $\ell(\Sigma) = -\frac{N}{2} \log \det \Sigma - \frac{1}{2} \text{tr}(\Sigma^{-1}S) + \text{const}$ with scatter matrix S , that $\nabla_\Sigma \ell = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} S \Sigma^{-1}$. Problems 3 and 9 use these.

Builds on: [Matrix calculus conventions](#), [Jacobians and the chain rule](#)

Problems

1. Show that $d(A^{-1}) = -A^{-1} (dA) A^{-1}$, and write the derivative of A^{-1} along a direction E . difficulty ·

2. Prove the trace trick: if $df = \text{tr}(C^\top dX)$ for every dX , then $\nabla_X f = C$. difficulty ··
Then use it on $f(X) = a^\top X^{-1}b$ with X invertible and a, b fixed vectors.

3. Show that $\frac{d}{dt} \det(I + tM) \Big|_{t=0} = \text{tr } M$, and derive Jacobi's formula difficulty ··
 $d(\det A) = \det(A) \text{tr}(A^{-1}dA)$. Deduce $\nabla_A \det A$ and $\nabla_A \log \det A$.

4. Let $A(\theta)$ be an invertible matrix depending on a scalar θ , with difficulty ··
derivative $A'(\theta)$. Show that $\frac{d}{d\theta} \log \det A(\theta) = \text{tr}(A^{-1}A')$ and $\frac{d}{d\theta} A(\theta)^{-1} = -A^{-1}A'A^{-1}$.
Evaluate both for $A(\theta) = B + \theta I$ with B symmetric positive definite and eigenvalues λ_i , and $\theta > 0$.

5. A layer solves $x = A^{-1}b$ (as `solve(A, b)`, without forming A^{-1}). difficulty ···
With upstream gradient $g = \nabla_x L$, compute $\nabla_b L$ and $\nabla_A L$, and say what the backward pass costs.

6. A layer computes $Y = A^{-1}$ explicitly. With upstream gradient $G = \nabla_Y L$, compute $\nabla_A L$ in terms of Y and G . difficulty ··

7. Rank-one update. For invertible A and vectors u, v with $1 + v^\top A^{-1}u \neq 0$, show the Sherman–Morrison formula $(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}$ and the matrix determinant lemma $\det(A + uv^\top) = \det(A) (1 + v^\top A^{-1}u)$. Then, with $1 + v^\top A^{-1}u > 0$ and $\det A > 0$, compute $\nabla_u \log \det(A + uv^\top)$. difficulty ..

8. Let $\Sigma = LL^\top$ with L a Cholesky factor. Show that $\log \det \Sigma = 2 \sum_i \log L_{ii}$, and compute the gradient of $\log \det \Sigma$ with respect to the free entries of L , once directly and once through $\nabla_\Sigma \log \det \Sigma = \Sigma^{-1}$. difficulty ...

9. The Gaussian log-likelihood of Before you start, $\ell(\Sigma) = -\frac{N}{2} \log \det \Sigma - \frac{1}{2} \text{tr}(\Sigma^{-1}S) + \text{const}$, is optimised through $\Sigma = LL^\top$. Starting from $\nabla_\Sigma \ell$, compute the gradient with respect to the free entries of L , simplify it using $M = L^{-1}SL^{-\top}$, and show that it vanishes at $\Sigma = S/N$. difficulty ...

10. Let A be symmetric with distinct eigenvalues λ_i and orthonormal eigenvectors q_i , and let E be symmetric. Show that along $A + tE$ each eigenvalue has derivative $\frac{d\lambda_i}{dt} = q_i^\top E q_i$ at $t = 0$, and use it to recover $\frac{d}{dt} \log \det(A + tE) = \text{tr}(A^{-1}E)$ for positive definite A . difficulty ..

Answers

1. $d(A^{-1}) = -A^{-1} (dA) A^{-1}$; along E , $\frac{d}{dt}(A + tE)^{-1} \big|_{t=0} = -A^{-1}EA^{-1}$
2. $\nabla_X f = C$ whenever $df = \text{tr}(C^\top dX)$; $\nabla_X(a^\top X^{-1}b) = -X^{-\top}ab^\top X^{-\top}$
3. $\frac{d}{dt} \det(I + tM) \big|_0 = \text{tr } M$; $d(\det A) = \det(A) \text{tr}(A^{-1}dA)$; $\nabla_A \det A = \det(A)A^{-\top}$ and $\nabla_A \log \det A = A^{-\top}$
4. $\frac{d}{d\theta} \log \det A = \text{tr}(A^{-1}A')$, $\frac{d}{d\theta} A^{-1} = -A^{-1}A'A^{-1}$; for $B + \theta I$: $\sum_i 1/(\lambda_i + \theta)$ and $-(B + \theta I)^{-2}$
5. $\nabla_b L = y = A^{-\top}g$ and $\nabla_A L = -yx^\top = -(A^{-\top}g)x^\top$
6. $\nabla_A L = -A^{-\top}GA^{-\top} = -Y^\top GY^\top$
7. $(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}$; $\det(A + uv^\top) = \det(A)(1 + v^\top A^{-1}u)$; $\nabla_u \log \det(A + uv^\top) = \frac{A^{-\top}v}{1 + v^\top A^{-1}u}$
8. $\log \det \Sigma = 2 \sum_i \log L_{ii}$; the gradient over the free entries of L is $2/L_{ii}$ on the diagonal and 0 below it, which is $\text{tril}(2\Sigma^{-1}L) = \text{tril}(2L^{-\top})$
9. $\nabla_L \ell = \text{tril}(2 \nabla_\Sigma \ell L) = \text{tril}(L^{-\top}(M - NI))$ with $M = L^{-1}SL^{-\top}$; it is 0 at $\Sigma = S/N$
10. $\frac{d\lambda_i}{dt} = q_i^\top E q_i$, and $\frac{d}{dt} \log \det(A + tE) = \sum_i q_i^\top E q_i / \lambda_i = \text{tr}(A^{-1}E)$

Worked solutions

Problem 1

Show that $d(A^{-1}) = -A^{-1} (dA) A^{-1}$, and write the derivative of A^{-1} along a direction E .

$A^{-1}A = I$ for every invertible A .	The definition of the inverse, and the identity does not change, so its differential is 0.
$d(A^{-1})A + A^{-1}dA = 0$.	Product rule on the left side, factors kept in order; $dI = 0$ on the right.
$d(A^{-1})A = -A^{-1}dA$.	Move the second term across.
$d(A^{-1}) = -A^{-1}(dA)A^{-1}$.	Multiply both sides by A^{-1} on the right. It has to go on the right, where the A is; multiplying on the left would not cancel it.
$d(A^{-1}) = -A^{-1}(dA)A^{-1}$; along E , $\left. \frac{d}{dt}(A + tE)^{-1} \right _{t=0} = -A^{-1}EA^{-1}$	For a 1×1 matrix this is $d(1/a) = -da/a^2$. The matrix version keeps the two factors of A^{-1} apart with dA between them, and Problem 2 onwards never needs more than this line and the trace trick.

Problem 2

Prove the trace trick: if $df = \text{tr}(C^{\top}dX)$ for every dX , then $\nabla_X f = C$. Then use it on $f(X) = a^{\top}X^{-1}b$ with X invertible and a, b fixed vectors.

$\text{tr}(C^{\top}dX) = \sum_{ij} C_{ij} dX_{ij}$.	Entry (j, j) of $C^{\top}dX$ is $\sum_i C_{ij}dX_{ij}$; the trace sums over j .
$df = \sum_{ij} \frac{\partial f}{\partial X_{ij}} dX_{ij}$.	The differential of a scalar function of many variables is the sum of its partial derivatives times the changes in the variables.

Both are linear in the entries of dX and agree for every dX , so $\partial f / \partial X_{ij} = C_{ij}$.	Take dX with a single nonzero entry at (i, j) : each side reduces to that one coefficient. That is $\nabla_X f = C$.
$df = a^\top d(X^{-1}) b = -a^\top X^{-1} (dX) X^{-1} b.$	a and b are constant, and Problem 1 gives the differential of the inverse.
$= -\operatorname{tr} (a^\top X^{-1} (dX) X^{-1} b) =$ $-\operatorname{tr} (X^{-1} b a^\top X^{-1} dX).$	A scalar is its own trace, and the cyclic property moves the vector factor $X^{-1} b$ from the end to the front, keeping the order $X^{-1} b$, then $a^\top X^{-1}$, then dX .
$C^\top = -X^{-1} b a^\top X^{-1}$, so $C =$ $-(X^{-1} b a^\top X^{-1})^\top = -X^{-\top} a b^\top X^{-\top}.$	Match step 5 to the form $\operatorname{tr}(C^\top dX)$, then transpose: the order reverses, $(b a^\top)^\top = a b^\top$, and $(X^{-1})^\top = X^{-\top}$.
$\nabla_X f = C$ whenever $df = \operatorname{tr}(C^\top dX)$; $\nabla_X (a^\top X^{-1} b) = -X^{-\top} a b^\top X^{-\top}$	$n \times n$, the shape of X . The Gaussian page's $\nabla_\Sigma (v^\top \Sigma^{-1} v) = -\Sigma^{-1} v v^\top \Sigma^{-1}$ is the case $a = b = v$ with X symmetric, so that the transposes disappear.

Problem 3

Show that $\frac{d}{dt} \det(I + tM) \big|_{t=0} = \operatorname{tr} M$, and derive Jacobi's formula $d(\det A) = \det(A) \operatorname{tr}(A^{-1} dA)$. Deduce $\nabla_A \det A$ and $\nabla_A \log \det A$.

Let $\mu_1, \dots, \mu_n$ be the eigenvalues of M , with multiplicity and possibly complex. Then $I + tM$ has eigenvalues $1 + t\mu_i$.	If $Mw = \mu w$ then $(I + tM)w = (1 + t\mu)w$, and the characteristic polynomial of $I + tM$ is that of M shifted, so multiplicities carry over.
$\det(I + tM) = \prod_i (1 + t\mu_i) = 1 + t \sum_i \mu_i + O(t^2).$	The determinant is the product of the eigenvalues; expand and collect the terms with one factor of t .
$\sum_i \mu_i = \operatorname{tr} M$, so $\frac{d}{dt} \det(I + tM) \big _{t=0} = \operatorname{tr} M$.	The trace is the sum of the eigenvalues; the derivative at 0 is the coefficient of t .
$\det(A + tE) = \det(A(I + tA^{-1}E)) = \det(A) \det(I + tA^{-1}E).$	Factor A out on the left, then $\det(XY) = \det X \det Y$.

$\frac{d}{dt} \det(A + tE) _{t=0} = \det(A) \operatorname{tr}(A^{-1}E).$	Step 3 with $M = A^{-1}E$. Since E is any direction, this is $d(\det A) = \det(A) \operatorname{tr}(A^{-1}dA).$
$\det(A) \operatorname{tr}(A^{-1}dA) = \operatorname{tr}((\det(A)A^{-\top})^\top dA),$ so $\nabla_A \det A = \det(A) A^{-\top}.$	Match to $\operatorname{tr}(C^\top dA)$: $C^\top = \det(A)A^{-1}$, so $C = \det(A)A^{-\top}$ (Problem 2).
$d \log \det A = \frac{d(\det A)}{\det A} = \operatorname{tr}(A^{-1}dA),$ so $\nabla_A \log \det A = A^{-\top}.$	Chain rule through log; the $\det A$ cancels, which is why $\log \det$ is the quantity that gets differentiated in practice.
$\frac{d}{dt} \det(I + tM) _0 = \operatorname{tr} M; d(\det A) = \det(A) \operatorname{tr}(A^{-1}dA); \nabla_A \det A = \det(A)A^{-\top}$ and $\nabla_A \log \det A = A^{-\top}$	For symmetric Σ , $\Sigma^{-\top} = \Sigma^{-1}$, which is the Gaussian page's result. $\det(A)A^{-\top}$ is the transpose of the adjugate, whose (i, j) entry is the cofactor of A_{ij} : differentiating the cofactor expansion along row i gives the same thing entry by entry.

Problem 4

Let $A(\theta)$ be an invertible matrix depending on a scalar θ , with derivative $A'(\theta)$. Show that $\frac{d}{d\theta} \log \det A(\theta) = \operatorname{tr}(A^{-1}A')$ and $\frac{d}{d\theta} A(\theta)^{-1} = -A^{-1}A'A^{-1}$. Evaluate both for $A(\theta) = B + \theta I$ with B symmetric positive definite and eigenvalues λ_i , and $\theta > 0$.

$dA = A'(\theta) d\theta.$	A depends on the single variable θ , so its differential is its derivative times $d\theta$.
$d \log \det A = \operatorname{tr}(A^{-1}dA) = \operatorname{tr}(A^{-1}A') d\theta.$	Problem 3, step 7, with step 1 substituted; $d\theta$ is a scalar and comes out of the trace.
$d(A^{-1}) = -A^{-1}(dA)A^{-1} = -A^{-1}A'A^{-1} d\theta.$	Problem 1 with step 1 substituted.
For $A = B + \theta I$: $A' = I$, so $\frac{d}{d\theta} \log \det A = \operatorname{tr}(A^{-1})$ and $\frac{d}{d\theta} A^{-1} = -A^{-2}.$	$\frac{d}{d\theta}(B + \theta I) = I$, and $A^{-1}IA^{-1} = A^{-2}$. Here the factors do commute, because one of them is the identity.

$B = Q\Lambda Q^\top$ gives $A = Q(\Lambda + \theta I)Q^\top$ and $A^{-1} = Q(\Lambda + \theta I)^{-1}Q^\top$.	B is symmetric, so it has an orthonormal eigenbasis; $QQ^\top = I$ lets θI be written as $Q(\theta I)Q^\top$, and the inverse of $QDQ^\top$ with Q orthogonal is $QD^{-1}Q^\top$.
$\text{tr}(A^{-1}) = \text{tr}((\Lambda + \theta I)^{-1}) = \sum_i \frac{1}{\lambda_i + \theta}$.	Cyclic trace: $\text{tr}(QDQ^\top) = \text{tr}(DQ^\top Q) = \text{tr} D$, and D is diagonal with entries $1/(\lambda_i + \theta)$.
$\frac{d}{d\theta} \log \det A = \text{tr}(A^{-1}A')$, $\frac{d}{d\theta} A^{-1} = -A^{-1}A'A^{-1}$; for $B + \theta I$: $\sum_i 1/(\lambda_i + \theta)$ and $-(B + \theta I)^{-2}$	$B + \theta I$ is the ridge matrix $X^\top X + \lambda I$ with $\theta = \lambda$, and the marginal likelihood of a Gaussian process has exactly this log det with θ the noise variance (the Gaussian-process page). The sum is positive and decreasing in θ : log det grows with the ridge, fastest when some λ_i is small.

Problem 5

A layer solves $x = A^{-1}b$ (as `solve(A, b)`, without forming A^{-1}). With upstream gradient $g = \nabla_x L$, compute $\nabla_b L$ and $\nabla_A L$, and say what the backward pass costs.

$dx = d(A^{-1})b + A^{-1}db = -A^{-1}(dA)A^{-1}b + A^{-1}db$.	Product rule on $A^{-1}b$, then Problem 1 for the first term.
$= -A^{-1}(dA)x + A^{-1}db$.	$A^{-1}b = x$, the forward output, which is saved.
$dL = g^\top dx = -g^\top A^{-1}(dA)x + g^\top A^{-1}db$.	L depends on A and b only through x , and $dL = \sum_i (\partial L / \partial x_i) dx_i = g^\top dx$.
Let $y = A^{-\top}g$, so $g^\top A^{-1} = y^\top$.	$y^\top = (A^{-\top}g)^\top = g^\top A^{-1}$. y is one solve with $A^\top$: <code>solve(A.T, g)</code> .
$g^\top A^{-1}db = y^\top db$, so $\nabla_b L = y = A^{-\top}g$.	Matching $dL = \langle \nabla_b L, db \rangle$ with $dA = 0$.

$-g^\top A^{-1}(dA)x = -y^\top(dA)x =$ $-\text{tr}(y^\top dA x) = -\text{tr}(xy^\top dA).$	A scalar is its own trace; the cyclic property moves x to the front.
$\nabla_A L = -(xy^\top)^\top = -yx^\top.$	Trace trick with $C^\top = -xy^\top$.
$\nabla_b L = y = A^{-\top} g \textbf{ and } \nabla_A L = -yx^\top =$ $-(A^{-\top} g)x^\top$	$\nabla_A L$ is $n \times n$ and rank one: an outer product of the solve with $A^\top$ and the forward output. The backward pass costs one more solve (or one more triangular pair if the factorisation of A is kept) and an outer product; A^{-1} is never formed, forward or backward. Check against Problem 2: $L = a^\top x$ has $g = a$, and $-A^{-\top} a (A^{-1} b)^\top = -A^{-\top} a b^\top A^{-\top}$.

Problem 6

A layer computes $Y = A^{-1}$ explicitly. With upstream gradient $G = \nabla_Y L$, compute $\nabla_A L$ in terms of Y and G .

$dL = \langle G, dY \rangle = \text{tr}(G^\top dY).$	L depends on A only through Y, and the differential of a scalar function of a matrix is the Frobenius inner product of its gradient with the change.
$dY = -A^{-1}(dA)A^{-1} = -Y(dA)Y.$	Problem 1, with $A^{-1} = Y$ saved from the forward pass.
$dL = -\text{tr}(G^\top Y dA Y) = -\text{tr}(Y G^\top Y dA).$	Cyclic: the trailing Y moves to the front.
$\nabla_A L = -(Y G^\top Y)^\top = -Y^\top G Y^\top.$	Trace trick with $C^\top = -Y G^\top Y$; transposing reverses the order and $(G^\top)^\top = G$.

$$\nabla_A L = -A^{-\top} G A^{-\top} = -Y^{\top} G Y^{\top}$$

$n \times n$, the shape of A . For $L = \langle G, Y \rangle$ with $G = ab^{\top}$, $L = a^{\top} Y b = a^{\top} A^{-1} b$ and this gives $-A^{-\top} ab^{\top} A^{-\top}$, Problem 2 again. The two transposes are the whole content; when A is symmetric and G is not, $-YGY$ is still wrong, because G sits between them.

Problem 7

Rank-one update. For invertible A and vectors u, v with $1 + v^{\top} A^{-1} u \neq 0$, show the Sherman–Morrison formula $(A + uv^{\top})^{-1} = A^{-1} - \frac{A^{-1} uv^{\top} A^{-1}}{1 + v^{\top} A^{-1} u}$ and the matrix determinant lemma $\det(A + uv^{\top}) = \det(A) (1 + v^{\top} A^{-1} u)$. Then, with $1 + v^{\top} A^{-1} u > 0$ and $\det A > 0$, compute $\nabla_u \log \det(A + uv^{\top})$.

Write $s = 1 + v^{\top} A^{-1} u$, a scalar.

$$(A + uv^{\top}) \left(A^{-1} - \frac{A^{-1} uv^{\top} A^{-1}}{1 + v^{\top} A^{-1} u} \right) = I + uv^{\top} A^{-1} - \frac{uv^{\top} A^{-1}}{s} - \frac{u (v^{\top} A^{-1} u) v^{\top} A^{-1}}{s}.$$

Multiply out the four terms; $AA^{-1} = I$, and in the last term the scalar $v^{\top} A^{-1} u$ has been pulled out of the middle of $uv^{\top} A^{-1} uv^{\top} A^{-1}$.

$$= I + uv^{\top} A^{-1} \left(1 - \frac{1}{s} - \frac{v^{\top} A^{-1} u}{s} \right) = I + uv^{\top} A^{-1} \cdot \frac{s - 1 - v^{\top} A^{-1} u}{s} = I.$$

The last three terms share the factor $uv^{\top} A^{-1}$; the bracket's numerator is $s - 1 - (s - 1) = 0$. A right inverse of a square matrix is its inverse.

$$\det(I + uv^{\top}) = 1 + v^{\top} u \text{ for any vectors } u, v.$$

$I + uv^{\top}$ sends w to $(1 + v^{\top} u)w$, so w is an eigenvector with eigenvalue $1 + v^{\top} u$ (when $w \neq 0$); every z with $v^{\top} z = 0$, an $(n - 1)$ -dimensional space, is an eigenvector with eigenvalue 1. When $v^{\top} u \neq 0$ these n eigenvectors are independent and the determinant is the product of the eigenvalues; when $v^{\top} u = 0$ both sides equal 1 by continuity in w .

$$\det(A + uv^{\top}) = \det(A(I + A^{-1} uv^{\top})) = \det(A) \det(I + A^{-1} u v^{\top}) = \det(A) (1 + v^{\top} A^{-1} u).$$

Factor out A , split the determinant, and apply step 3 with $w = A^{-1} u$.

$\log \det(A + uv^\top) = \log \det A + \log(1 + v^\top A^{-1}u).$	Log of step 4; both factors are positive by assumption.
$\nabla_u(v^\top A^{-1}u) = (v^\top A^{-1})^\top = A^{-\top}v.$	$v^\top A^{-1}u = c^\top u$ with $c = (v^\top A^{-1})^\top$, and $\nabla_u(c^\top u) = c$ (the matrix-calculus page).
$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u};$ $\det(A + uv^\top) = \det(A)(1 + v^\top A^{-1}u);$ $\nabla_u \log \det(A + uv^\top) = \frac{A^{-\top}v}{1 + v^\top A^{-1}u}$	The first two turn an $O(n^3)$ inverse and determinant into $O(n^2)$ work once A^{-1} (or a factorisation) is known: adding one example to a kernel matrix, one observation to a Kalman covariance, or a rank-one step in a quasi-Newton method. The gradient agrees with Problem 3 through the chain rule: $\nabla_A \log \det$ at $A + uv^\top$ is $(A + uv^\top)^{-\top}$, and u enters through $uv^\top$.

Problem 8

Let $\Sigma = LL^\top$ with L a Cholesky factor. Show that $\log \det \Sigma = 2 \sum_i \log L_{ii}$, and compute the gradient of $\log \det \Sigma$ with respect to the free entries of L , once directly and once through $\nabla_\Sigma \log \det \Sigma = \Sigma^{-1}$.

$\det \Sigma = \det L \det L^\top = (\det L)^2 = \prod_i L_{ii}^2.$	$\det(XY) = \det X \det Y$, $\det L^\top = \det L$, and the determinant of a triangular matrix is the product of its diagonal.
$\log \det \Sigma = 2 \sum_i \log L_{ii}.$	Log of step 1; each $L_{ii} > 0$ so the logs are defined. This is how a library computes $\log \det$ of a covariance: no determinant is ever formed.
Directly: $\frac{\partial}{\partial L_{ij}} 2 \sum_k \log L_{kk} =$ $\begin{cases} 2/L_{ii} & i = j \\ 0 & i > j \end{cases}.$	Only the diagonal entries appear in the formula, so the below-diagonal entries have zero derivative, and $\frac{d}{dx} \log x = 1/x$.
Through Σ : $d\Sigma = (dL)L^\top + L(dL)^\top.$	Product rule on $LL^\top$, in order.

$d \log \det \Sigma = \text{tr}(\Sigma^{-1} d\Sigma) =$ $\text{tr}(\Sigma^{-1}(dL)L^\top) + \text{tr}(\Sigma^{-1}L(dL)^\top).$	Problem 3, step 7, with $d\Sigma$ from step 4 and the trace split over the sum.
$\text{tr}(\Sigma^{-1}(dL)L^\top) = \text{tr}(L^\top \Sigma^{-1} dL)$ and $\text{tr}(\Sigma^{-1}L(dL)^\top) = \text{tr}(dL L^\top \Sigma^{-1}) =$ $\text{tr}(L^\top \Sigma^{-1} dL).$	Cyclic trace for the first; for the second, transpose inside the trace (which leaves it unchanged, using $\Sigma^{-\top} = \Sigma^{-1}$) and then cycle. The two terms are equal.
$d \log \det \Sigma = 2 \text{tr}((\Sigma^{-1}L)^\top dL)$, so over all entries the gradient is $2\Sigma^{-1}L =$ $2L^{-\top}L^{-1}L = 2L^{-\top}.$	Trace trick with $C^\top = 2L^\top \Sigma^{-1}$, so $C =$ $2\Sigma^{-1}L$; then $\Sigma^{-1} = (LL^\top)^{-1} = L^{-\top}L^{-1}.$
$L^{-\top}$ is upper triangular with diagonal $1/L_{ii}$, so $\text{tril}(2L^{-\top}) =$ $\text{diag}(2/L_{11}, \dots, 2/L_{nn}).$	The inverse of a lower-triangular matrix is lower triangular with reciprocal diagonal, and its transpose is upper triangular. Only the entries where L has free parameters count: dL is zero above the diagonal, so those entries of C multiply nothing.
$\log \det \Sigma = 2 \sum_i \log L_{ii}$; the gradient over the free entries of L is $2/L_{ii}$ on the diagonal and 0 below it, which is $\text{tril}(2\Sigma^{-1}L) = \text{tril}(2L^{-\top})$	Both routes agree, and the second is the one a framework's autodiff takes: it computes the full $2\Sigma^{-1}L$ and the restriction to the parametrised entries discards the upper triangle. A gradient step on the off-diagonal entries of L that came from $\log \det$ alone would be a bug; those entries only move through the other terms of a likelihood (Problem 9).

Problem 9

The Gaussian log-likelihood of Before you start, $\ell(\Sigma) = -\frac{N}{2} \log \det \Sigma - \frac{1}{2} \text{tr}(\Sigma^{-1}S) + \text{const}$, is optimised through $\Sigma = LL^\top$. Starting from $\nabla_\Sigma \ell$, compute the gradient with respect to the free entries of L , simplify it using $M = L^{-1}SL^{-\top}$, and show that it vanishes at $\Sigma = S/N$.

Write $C = \nabla_\Sigma \ell = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}S\Sigma^{-1}$, which is symmetric because Σ and S are.

$d\ell = \langle C, d\Sigma \rangle = \langle C, (dL)L^\top \rangle + \langle C, L(dL)^\top \rangle.$	ℓ depends on L only through Σ , with $d\Sigma$ from Problem 8, step 4, and the inner product is linear.
$\langle C, (dL)L^\top \rangle = \text{tr} (C^\top (dL)L^\top) = \text{tr} (L^\top C^\top dL) = \langle CL, dL \rangle.$	Cyclic trace, then $L^\top C^\top = (CL)^\top$.
$\langle C, L(dL)^\top \rangle = \text{tr} (C^\top L(dL)^\top) = \text{tr} (dL L^\top C) = \text{tr} (L^\top C dL) = \langle C^\top L, dL \rangle.$	Transpose inside the trace, then cycle; $L^\top C = (C^\top L)^\top$.
$d\ell = \langle (C + C^\top)L, dL \rangle = \langle 2CL, dL \rangle$, so over the free entries $\nabla_L \ell = \text{tril}(2CL)$.	Trace trick; C is symmetric; and as in Problem 8 only the lower triangle is parametrised.
$2CL = -N\Sigma^{-1}L + \Sigma^{-1}S\Sigma^{-1}L = -NL^{-\top} + L^{-\top}L^{-1}SL^{-\top}.$	$\Sigma^{-1} = L^{-\top}L^{-1}$ and $L^{-1}L = I$, applied to each term: $\Sigma^{-1}L = L^{-\top}$, and $\Sigma^{-1}S\Sigma^{-1}L = L^{-\top}L^{-1}SL^{-\top}$.
$2CL = L^{-\top}(M - NI)$ with $M = L^{-1}SL^{-\top}$.	Factor $L^{-\top}$ out on the left of both terms.
At $\Sigma = S/N$: $LL^\top = S/N$, so $M = L^{-1}(NLL^\top)L^{-\top} = NI$ and $2CL = 0$.	$L^{-1}L = I$ and $L^\top L^{-\top} = I$. Then the lower triangle of 0 is 0.
$\nabla_L \ell = \text{tril} (2 \nabla_\Sigma \ell L) = \text{tril} (L^{-\top}(M - NI))$ with $M = L^{-1}SL^{-\top}$; it is 0 at $\Sigma = S/N$	M is the scatter matrix of the whitened data $L^{-1}(x_n - \mu)$, and the gradient says: move L until the whitened scatter is N times the identity, that is, until Σ explains the data's spread exactly. The maximum-likelihood covariance from the Gaussian page is unchanged by the reparametrisation, as it must be; what changes is the gradient the optimiser sees, which here needs two triangular solves and no inverse.

Problem 10

Let A be symmetric with distinct eigenvalues λ_i and orthonormal eigenvectors q_i , and let E be symmetric. Show that along $A + tE$ each eigenvalue has derivative $\frac{d\lambda_i}{dt} = q_i^\top E q_i$ at $t = 0$, and use it to recover $\frac{d}{dt} \log \det(A + tE) = \text{tr}(A^{-1}E)$ for positive definite A .

Take as known that for distinct eigenvalues, $\lambda_i(t)$ and a unit eigenvector $q_i(t)$ of $A + tE$ can be chosen to vary differentiably in t .

$(A + tE) q_i(t) = \lambda_i(t) q_i(t).$	The eigenvector equation at every t .
$E q_i + A q'_i = \lambda'_i q_i + \lambda_i q'_i$ at $t = 0$.	Differentiate both sides with respect to t by the product rule; $\frac{d}{dt}(A + tE) = E$. Primes are t -derivatives at 0.
$q_i^\top E q_i + q_i^\top A q'_i = \lambda'_i + \lambda_i q_i^\top q'_i.$	Multiply on the left by $q_i^\top$; $q_i^\top q_i = 1$.
$q_i^\top A q'_i = (A q_i)^\top q'_i = \lambda_i q_i^\top q'_i.$	A is symmetric, so $q_i^\top A = (A q_i)^\top$, and $A q_i = \lambda_i q_i$.
$\lambda'_i = q_i^\top E q_i.$	The two q'_i terms in step 3 are equal by step 4 and cancel. The eigenvector's derivative never needed to be found.
$\frac{d}{dt} \log \det(A + tE) = \sum_i \frac{\lambda'_i}{\lambda_i} = \sum_i \frac{q_i^\top E q_i}{\lambda_i}.$	$\log \det = \sum_i \log \lambda_i$ for positive eigenvalues, and $\frac{d}{dt} \log \lambda_i = \lambda'_i / \lambda_i$.
$\sum_i \frac{q_i^\top E q_i}{\lambda_i} = \sum_i \text{tr} \left(\frac{q_i q_i^\top}{\lambda_i} E \right) = \text{tr} \left(\left(\sum_i \frac{q_i q_i^\top}{\lambda_i} \right) E \right) = \text{tr}(A^{-1}E).$	$q_i^\top E q_i = \text{tr}(q_i q_i^\top E)$ by the cyclic property; the trace is linear; and $\sum_i q_i q_i^\top / \lambda_i = Q \Lambda^{-1} Q^\top = A^{-1}$ is the eigendecomposition of the inverse.
$\frac{d\lambda_i}{dt} = q_i^\top E q_i$, and $\frac{d}{dt} \log \det(A + tE) = \sum_i q_i^\top E q_i / \lambda_i = \text{tr}(A^{-1}E)$	Jacobi's formula from Problem 3, reached through the spectrum. The per-eigenvalue form says what the trace hides: a perturbation moves $\log \det$ most through the smallest eigenvalues, with weight $1/\lambda_i$ along q_i , which is why a nearly singular covariance makes its $\log \det$ gradient large along the nearly flat direction.

Where this goes wrong

1. Differentiating the inverse like a reciprocal

The scalar rule $d(1/a) = -da/a^2$ is well worn, and A^{-1} looks like $1/A$.

$$A^{-1}A = I$$

Right so far: Problem 1, step 1.

“ A^{-1} is $1/A$, so its derivative is $-1/A^2$ times dA .”

The analogy that causes the mistake: in the scalar rule the two factors of $1/a$ sit together because numbers commute; matrices do not.

$$d(A^{-1}) = -A^{-2} dA \quad \times$$

Problem 1 gives $-A^{-1}(dA)A^{-1}$, with dA between the two inverses. The two agree only when dA commutes with A , for instance $A = B + \theta I$ differentiated in θ (Problem 4), which is exactly the case in which the habit is formed. In Problem 5 the wrong form gives $\nabla_A L = -A^{-2\top} g b^\top$ in place of $-A^{-\top} g x^\top$: the right shape, and a gradient check fails on every entry.

2. Trace trick without the transpose

Once df is inside a trace, the matrix next to dX looks like the gradient.

$$df = -\text{tr}(X^{-1} b a^\top X^{-1} dX) \text{ for } f = a^\top X^{-1} b$$

Right so far: Problem 2, step 5.

“The gradient is whatever multiplies dX inside the trace.”

The shortcut that causes the mistake: the trace trick reads $\text{tr}(C^\top dX)$, so what multiplies dX is $C^\top$, not C .

$$\nabla_X(a^\top X^{-1}b) = -X^{-1}ba^\top X^{-1} \mathbf{x}$$

The gradient is the transpose, $-X^{-\top}ab^\top X^{-\top}$ (Problem 2). Both are $n \times n$, so no shape check objects, and for symmetric X with $a = b$ they coincide, which is the case the Gaussian page uses. Test the general case on one entry: with $X = I$, $\partial f / \partial X_{ij} = -a_i b_j$, the (i, j) entry of $-ab^\top$, not of $-ba^\top$.

3. The gradient of det taken to be the gradient of log det

$A^{-\top}$ is the answer everyone remembers, and the log is easy to drop or to add.

$$d(\det A) = \det(A) \operatorname{tr}(A^{-1}dA)$$

Right so far: Jacobi's formula, Problem 3.

“So the gradient of the determinant is the inverse transpose.”

The slip that causes the mistake: the factor $\det A$ in front of the trace is treated as part of the bookkeeping, when it is the difference between $\det$ and $\log \det$.

$$\nabla_A \det A = A^{-\top} \mathbf{x}$$

That is $\nabla_A \log \det A$; the determinant's gradient is $\det(A)A^{-\top}$ (Problem 3). With $A = 2I$ in $n = 4$ dimensions, $\det A = 16$ and $\partial \det A / \partial A_{11} = 8$ (the cofactor), while $A^{-\top}$ has entry $\frac{1}{2}$ there. The error is a factor of $\det A$ on every entry, which in a likelihood is the difference between an update that scales with the volume of the covariance and one that does not.

4. Cholesky gradient left above the diagonal

Autodiff produces $2\Sigma^{-1}L$ for the full matrix, and it is tempting to use all of it.

$$d \log \det \Sigma = \langle 2\Sigma^{-1}L, dL \rangle \text{ over all } n^2 \text{ entries of } L$$

Right so far: Problem 8, step 7.

“ L is a matrix parameter, so update every entry of L with this gradient.”

The shortcut that causes the mistake: $2\Sigma^{-1}L = 2L^{-\top}$ is upper triangular, and its nonzero off-diagonal entries sit exactly where L has no parameters.

$L \leftarrow L - \eta \cdot 2L^{-\top}$, so the entries above the diagonal of L become nonzero ✗

L is no longer lower triangular, $\log \det(LL^\top) = 2 \sum_i \log L_{ii}$ no longer holds, and the next Cholesky-based solve is wrong. The gradient over the free entries is $\text{tril}(2L^{-\top})$, which is only the diagonal (Problem 8); a parametrisation that stores the lower triangle as a vector, or masks the gradient with tril , never sees the problem.

5. Chain rule through $LL^\top$ counting L once

$\Sigma = LL^\top$ is a product, but it is a product of a matrix with itself.

$\nabla_\Sigma \ell = C$, symmetric

Right so far: the Gaussian page.

“ $\Sigma = LL^\top$ is linear in L with $L^\top$ fixed, so by the chain rule $\nabla_L \ell = CL$.”

The shortcut that causes the mistake: $d\Sigma = (dL)L^\top + L(dL)^\top$ has two terms, and the second is dropped because $L^\top$ is treated as a different variable.

$\nabla_L \ell = \text{tril}(CL)$ ✗

The two terms are equal (Problem 9, steps 2 and 3), so the gradient is $\text{tril}(2CL)$: the error is a factor of 2 on every entry. The stationary point is unchanged, which is why training can still reach the maximum-likelihood covariance, only with half the step the learning rate promises; the same slip in $\log \det \Sigma$ alone gives $1/L_{ii}$ in place of $2/L_{ii}$ (Problem 8).