

Extra Exercises
for
Modeling and Analysis Principles
for
Chemical and Biological Engineers
2nd Edition

Michael D. Graham

Department of Chemical and Biological Engineering
University of Wisconsin-Madison
Madison, Wisconsin

James B. Rawlings

Department of Chemical Engineering
University of California, Santa Barbara
Santa Barbara, California

December 29, 2025

The logo for Nob Hill Publishing features the words "Nob" and "Hill" in a serif font, positioned above a stylized, wavy line that represents a hill. The word "Publishing" is written in a sans-serif font to the right of the wavy line.

Madison, Wisconsin

The extra exercises were set in Lucida using L^AT_EX.

Copyright © 2026 by Nob Hill Publishing, LLC

All rights reserved.

Nob Hill Publishing, LLC
Cheryl M. Rawlings, publisher
Santa Barbara, CA 93101
orders@nobhillpublishing.com
<http://www.nobhillpublishing.com>

The extra exercises are intended for use by students and course instructors.

This document has been posted electronically on the website: www.chemengr.ucsb.edu/~jbraw/principles.

1

Linear Algebra

Exercise 1.79: Using the SVD to solve engineering problems

Consider the system depicted in Figure 1.12 in which we can manipulate an input $u \in \mathbb{R}^3$ to cancel the effect of a disturbance $d \in \mathbb{R}^2$ on an output $y \in \mathbb{R}^2$ of interest. The steady-state relationship between the variables is modeled as a linear relationship

$$y = Gu + d$$

and y, u, d are in deviation variables from the steady state at which the system was linearized. Experimental tests on the system have produced the following model parameters

$$G = [U] \begin{bmatrix} \Sigma & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

$$U = \begin{bmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1.21 & 0.00 \\ 0.00 & 0.08 \end{bmatrix} \quad V_1 = \begin{bmatrix} -0.98 & 0.11 \\ 0.20 & 0.84 \\ -0.098 & 0.54 \end{bmatrix}$$

If we have measurements of the disturbance d available, we would like to find the input u that exactly cancels d 's effect on y , and we would like to know ahead of time what is the worst-case disturbance that can hit the system.

- Can you use u to exactly cancel the effect of d on y for *all* d ? Why or why not?
- In terms of U, Σ, V_1, V_2 , and d , what are all the inputs u that minimize the effect of d on y ?
- What is the smallest input u that minimizes the effect of d on y ?

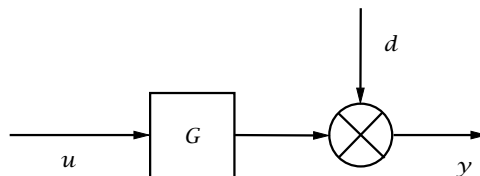


Figure 1.12: Manipulated input u and disturbance d combine to affect output y .

- (d) What is the worst-case disturbance d for this process, i.e., what d of unit norm requires the *largest* response in u ? What is the response u to this worst d ?
- (e) What is the best-case disturbance d for this process, i.e., what disturbance d of unit norm requires the *smallest* response u . What is the response u for this best d ?

Exercise 1.80: Functions of matrices

I recall that two solutions to the linear, scalar differential equation $\ddot{x} = x$ are $\cosh t$ and $\sinh t$. Say the boundary conditions for this differential equation are the values of x and its first derivative at $t = 0$, $x(0) = x_0$ and $\dot{x}(0) = \dot{x}_0$.

- (a) What is the full solution (in terms of \sinh and \cosh) to the scalar differential equation

$$\ddot{x} = mx$$

with these initial conditions and parameter $m > 0$?

- (b) Therefore, what is the solution to the set of n coupled differential equations

$$\ddot{x} = Mx \quad x(0) = x_0 \quad \dot{x}(0) = \dot{x}_0$$

in which $x, x_0, \dot{x}_0 \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times n}$ and matrix M has positive eigenvalues. Hint: make sure that all matrix-vector multiplications are defined.

- (c) Explain how you would evaluate this solution numerically. You may assume M has distinct eigenvalues.

Exercise 1.81: Concepts related to eigenvalues, SVD, and the fundamental theorem

Consider a general complex-valued matrix $A \in \mathbb{C}^{m \times n}$.

- (a) Establish the relationship between the eigenvalues of A^*A and AA^* . For example: (i) how many eigenvalues does each matrix have? (ii) are these eigenvalues equal to each other? (iii) all of them? etc. Do not use the SVD factorization of A in your derivation because this result is used to establish the SVD factorization.
- (b) Show that the following two null spaces are the same: $N(A^*A) = N(A)$. Hint: first show the simple result that $x \in N(A)$ implies $x \in N(A^*A)$ so $N(A)$ is contained in $N(A^*A)$. Next show that $x \in N(A^*A)$ implies that $x \in N(A)$ to complete the derivation. You may want to use the fundamental theorem of linear algebra for this second step.

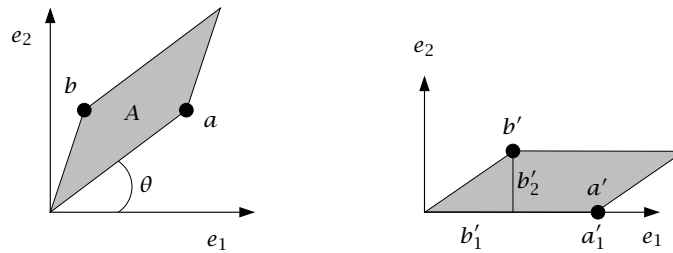


Figure 1.13: Parallelogram and rotation to align with x_1 axis.

Exercise 1.82: Determinant of a matrix, area, and volume

- (a) Given a parallelogram with sides $a, b \in \mathbb{R}^2$ depicted in Figure 1.13, show that its area is given by the formula

$$A = |\det C| \quad C = \begin{bmatrix} a & b \end{bmatrix}$$

In other words, form a partitioned matrix by placing the column vectors representing the two sides next to each other, and take the absolute value of the determinant of that matrix to calculate the area. Note that the absolute value is required so that area is positive.

Hint: to reduce the algebra, first define the rotation matrix that rotates the parallelogram so that the a side is parallel to the x_1 -axis; then use the fact that the area of a parallelogram is the base times the height.

- (b) Show that the result generalizes to finding the volume of a parallelepiped with sides given by $a, b, c \in \mathbb{R}^3$

$$V = |\det C| \quad C = \begin{bmatrix} a & b & c \end{bmatrix}$$

- (c) How does the result generalize to n dimensions.

Note that the determinant without absolute value is then called the *signed* area (volume). Signed area is positive when rotation of side a into side b is counter-clockwise as shown in Figure 1.13, and negative when the rotation is clockwise (right-hand rule, same sign as cross product).

Exercise 1.83: Pseudoinverse

We have already seen two different forms of the pseudoinverse arise when solving in the least-squares sense

$$Ax = b$$

with the singular value decomposition in Section 1.4.7. The different forms motivate a more abstract definition of pseudoinverse that is general enough to cover all these cases. Consider the following definition

Definition 1.1 (Pseudoinverse). Let $A \in \mathbb{C}^{m \times n}$. A matrix $A^\dagger \in \mathbb{C}^{n \times m}$ satisfying

$$(1) AA^\dagger A = A \quad (2) A^\dagger AA^\dagger = A^\dagger \quad (3) (AA^\dagger)^* = AA^\dagger \quad (4) (A^\dagger A)^* = A^\dagger A$$

is called a pseudoinverse (or Moore-Penrose pseudoinverse) of A .

- (a) Show that $(A^*)^\dagger = (A^\dagger)^*$ by direct substitution into the four properties of the definition.
- (b) Show that (if it exists) A^\dagger is unique for all A . Hint: let both B and C satisfy the properties of the pseudoinverse. First show that $AB = AC$ and that $BA = CA$. Given these, show that $B = C$.

Obtaining uniqueness of the pseudoinverse is the motivation for properties 3 and 4 in the definition.

- (c) Let $A \neq 0$ and consider the SVD, $A = U\Sigma V^*$. Show that

$$A^\dagger = V_1 \Sigma_r^{-1} U_1^* \quad A \neq 0 \quad (1.42)$$

in which $r \geq 1$ is the rank of A , satisfies the definition of the pseudoinverse.

- (d) From the definition and inspection, what is a pseudoinverse for $A = 0$? Therefore we have shown that the pseudoinverse exists and is unique for all A , and we have a means to calculate it using the SVD.
- (e) Show that the pseudoinverse reduces to $A^\dagger = (A^*A)^{-1}A^*$ when A has linearly independent columns, and that it reduces to $A^\dagger = A^*(AA^*)^{-1}$ when A has linearly independent rows.

Exercise 1.84: Minimum-norm, least-squares solution to $Ax = b$

- (a) Show that the solution

$$x^0 = A^\dagger b$$

with the pseudoinverse defined in Exercise 1.83 is the unique minimum-norm least-squares solution to $Ax = b$ for arbitrary A, b . This result covers *all* versions of the least-squares problem.

- (b) When A has neither linearly independent columns nor linearly independent rows, find the set of *all* least-squares solutions.

Exercise 1.85: Pseudoinverse and Kronecker product

Establish the formula

$$(A \otimes B)^\dagger = A^\dagger \otimes B^\dagger$$

which generalizes (1.28) in the text.

Exercise 1.86: Solution to the general matrix equation $AXB = C$

Show that the unique minimum-norm X that minimizes $\|AXB - C\|_F$ is given by

$$X^0 = A^\dagger CB^\dagger$$

Here we are using the square root of the sum of the squares of the elements in the matrix as the matrix norm. This norm is called the Frobenius norm (see also Exercise 1.90). This result covers *all* versions of the matrix least-squares problem.

Exercise 1.87: Some linear algebra questions and the singular value decomposition (SVD)

I have an $A \in \mathbb{R}^{12 \times 12}$ and I call Octave's or MATLAB's $[U, S, V] = \text{svd}(A)$ function and obtain the return arguments shown on the next page.

Answer the following questions.

- (a) What is the rank of A ? What is $\|A\|_2$?
- (b) What are the smallest dimension matrices that you can multiply together to obtain A ? Clearly mark and label these matrices as parts of the U, S, V , matrices printed on the next page.
- (c) Is there a solution x to $Ax = b$ for every $b \in \mathbb{R}^{12}$? How do you know?
- (d) Say I happen to choose a b such that $Ax = b$ has a solution. Is this solution unique? How do you know?
- (e) Does the solution to the optimization problem $\min_x \|Ax - b\|_2^2$ exist for every right-hand side b ? Is this solution unique? How do you know?
- (f) What is A^\dagger in terms of the three matrices you identified part (b)? What do you call this matrix?
- (g) How can you calculate the smallest solution to $\min_x \|Ax - b\|_2^2$ given a b ? Is this solution unique?

U =

-0.4328	-0.0409	-0.3260	0.2898	-0.0375	0.2669	-0.0471	0.0107	-0.5901	-0.4301	0.0769	0.0815
-0.2593	0.5397	-0.2572	-0.0451	-0.0347	0.0161	0.0883	0.0042	0.3733	-0.2694	-0.5661	-0.1779
-0.3013	0.1062	0.1712	0.5850	0.3381	-0.0406	0.4231	0.0140	0.3349	0.0650	0.3365	0.0395
-0.2345	0.0860	-0.1324	-0.2227	0.4816	-0.2773	0.0806	0.1276	-0.3850	0.3865	-0.0405	-0.4895
-0.2267	0.0518	0.4199	-0.4300	0.4074	0.3383	0.1524	-0.2094	-0.1005	-0.0433	-0.1801	0.4391
-0.3732	-0.4287	-0.4275	-0.3879	-0.0369	-0.3253	0.1659	0.0834	0.3058	-0.0743	0.1558	0.2805
-0.2056	0.4478	-0.1227	-0.0921	-0.2727	-0.1315	-0.1419	-0.6237	-0.0594	0.3056	0.3442	0.1320
-0.3213	0.0349	0.4643	-0.3101	-0.4018	0.0897	0.1180	0.1767	0.0393	-0.2405	0.3237	-0.4527
-0.3199	-0.3116	0.1391	0.1287	0.2369	0.0435	-0.7090	-0.2495	0.2899	-0.0421	-0.0673	-0.2268
-0.3137	-0.3351	0.1398	0.2182	-0.4261	0.0718	0.2340	-0.1007	-0.0910	0.4807	-0.4809	-0.0023
-0.1626	0.2053	-0.1698	-0.0586	-0.0508	0.4846	-0.2808	0.5557	0.1496	0.4404	0.1626	0.1773
-0.1905	0.2204	0.3511	0.1205	-0.0882	-0.6021	-0.2889	0.3622	-0.1693	-0.0583	-0.1288	0.3791

S =

```

2.6e+01  0  0  0  0  0  0  0  0  0  0  0
0  2.4e+00  0  0  0  0  0  0  0  0  0  0
0  0  1.9e+00  0  0  0  0  0  0  0  0  0
0  0  0  9.8e-01  0  0  0  0  0  0  0  0
0  0  0  0  9.0e-01  0  0  0  0  0  0  0
0  0  0  0  0  4.2e-01  0  0  0  0  0  0
0  0  0  0  0  0  3.6e-01  0  0  0  0  0
0  0  0  0  0  0  0  2.1e-01  0  0  0  0
0  0  0  0  0  0  0  0  4.8e-02  0  0  0
0  0  0  0  0  0  0  0  0  1.0e-15  0  0
0  0  0  0  0  0  0  0  0  0  3.5e-16  0
0  0  0  0  0  0  0  0  0  0  0  8.0e-17

```

V =

```

-0.4328 -0.0409 -0.3260  0.2898 -0.0375  0.2669 -0.0471  0.0107 -0.5901  0.0736  0.1188  0.4220
-0.2593  0.5397 -0.2572 -0.0451 -0.0347  0.0161  0.0883  0.0042  0.3733  0.6505 -0.0258  0.0309
-0.3013  0.1062  0.1712  0.5850  0.3381 -0.0406  0.4231  0.0140  0.3349 -0.3280  0.1022  0.0311
-0.2345  0.0860 -0.1324 -0.2227  0.4816 -0.2773  0.0806  0.1276 -0.3850  0.0460  0.2453 -0.5730
-0.2267  0.0518  0.4199 -0.4300  0.4074  0.3383  0.1524 -0.2094 -0.1005  0.0276 -0.4263  0.2114
-0.3732 -0.4287 -0.4275 -0.3879 -0.0369 -0.3253  0.1659  0.0834  0.3058 -0.1975 -0.1298  0.2293
-0.2056  0.4478 -0.1227 -0.0921 -0.2727 -0.1315 -0.1419 -0.6237 -0.0594 -0.4583 -0.0529 -0.1284
-0.3213  0.0349  0.4643 -0.3101 -0.4018  0.0897  0.1180  0.1767  0.0393 -0.0370  0.6028  0.0534
-0.3199 -0.3116  0.1391  0.1287  0.2369  0.0435 -0.7090 -0.2495  0.2899  0.1488  0.1678 -0.0863
-0.3137 -0.3351  0.1398  0.2182 -0.4261  0.0718  0.2340 -0.1007 -0.0910  0.2288 -0.3892 -0.5084
-0.1626  0.2053 -0.1698 -0.0586 -0.0508  0.4846 -0.2808  0.5557  0.1496 -0.3689 -0.2213 -0.2584
-0.1905  0.2204  0.3511  0.1205 -0.0882 -0.6021 -0.2889  0.3622 -0.1693  0.0090 -0.3483  0.2057

```

Exercise 1.88: Solutions of linear differential equations with repeated eigenvalues

We want to solve a set of linear differential equations

$$\frac{d}{dt}x = Ax \quad x(0) = x_0$$

with $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$

- (a) What is the solution to this differential equation? Give the Taylor series for evaluating the exponential of a matrix e^C .

To help us express the solution, we performed an eigenvalue decomposition and found $Q, \Lambda \in \mathbb{C}^{n \times n}$ such that

$$A = Q\Lambda Q^{-1}$$

- (b) Substitute this expression for A into the Taylor series of your solution and express the solution in terms of t , Λ , Q , and Q^{-1} .

Unfortunately, it appears that our A is not diagonalizable because Λ has the form

$$\Lambda = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

- (c) Perform the indicated multiplications on Λ and write out the following: Λ^2 , Λ^3 , Λ^4 .
- (d) How does this result generalize for Λ^k , $k \geq 1$?
- (e) Substitute these results into the Taylor series applied to $e^{\Lambda t}$ and obtain a series expansion in terms of λ . What function of λ and t appear on the diagonal in $e^{\Lambda t}$?

- (f) Simplify the sum of terms in your series expansion and show what functions of λ and t appear above the diagonal in $e^{\Lambda t}$?
- (g) OK, based on your work so far, let's make a conjecture. What functions will appear in the matrix $e^{\Lambda t}$ when you have p repeated eigenvalues λ in $\Lambda_p \in \mathbb{C}^{p \times p}$

$$\Lambda_p = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda \end{bmatrix} \quad e^{\Lambda_p t} = ?$$

(Note that missing entries denote zeroes.)

Exercise 1.89: Changing basis for a linear space

Let V be a linear space of dimension n with a set of basis vectors $\{b_i\}_{i=1}^n$. Consider an arbitrary set of n -linearly independent vectors $\{a_i\}_{i=1}^n$. Show that the set $\{a_i\}_{i=1}^n$ is also a basis for V .

Exercise 1.90: The Frobenius norm of a matrix, the trace, and the vec operator

- (a) Let $A, B \in \mathbb{C}^{m \times n}$. Show that

$$\text{tr}(A^* B) = \text{vec}(A)^* \text{vec}(B)$$

- (b) Let $X \in \mathbb{C}^{m \times n}$. Show that

$$\|X\|_F^2 = \text{tr}(X^* X) = (\text{vec} X)^* \text{vec} X = \|\text{vec} X\|_2^2$$

Exercise 1.91: Useful properties of the pseudoinverse in least squares¹

Let $X \in \mathbb{R}^{m \times n}$ have independent columns and $E \in \mathbb{R}^{n \times n}$ be symmetric and full rank. We wish to show that

$$M = (X^T (XEX^T)^\dagger X)^{-1} X^T (XEX^T)^\dagger$$

is the pseudoinverse of X for all such E .

- (a) First show that the indicated inverse exists. Hint: show that $X^\dagger X = I_n$ if X has linearly independent columns using the result in Exercise 1.83(e). Therefore $E^{-1} X^\dagger X E = I_n$. Use that result to show that $X^T (XEX^T)^\dagger X = E^{-1}$, so that $(X^T (XEX^T)^\dagger X)^{-1} = E$.

¹JBR would like to thank Travis Arnold of UW for helpful discussion of this exercise.

- (b) Next note that $MX = I_n$ so M is a left-inverse of X . Then show that M is the pseudoinverse of X by verifying the four properties of the pseudoinverse (see Exercise 1.83).

We can also conclude that since the pseudoinverse is unique (see Exercise 1.83), M is independent of E .

Exercise 1.92: More least squares²

Let $X \in \mathbb{R}^{m \times n}$ have independent columns, $E \in \mathbb{R}^{n \times n}$ be symmetric and full rank, and $V \in \mathbb{R}^{m \times m}$ be symmetric and full rank. We wish to show that

$$M = (X^T(V + XEX^T)^\dagger X)^{-1} X^T(V + XEX^T)^\dagger$$

is independent of E . Note that M is a left-inverse of X as in the previous exercise, but it is no longer the pseudoinverse for $V \neq 0$.

- (a) From the matrix inversion lemma, we know that for full rank A and C (Rawlings, Mayne, and Diehl, 2020, Exercise 1.12)

$$\begin{aligned} (A + BCD)^{-1} &= A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1} \\ (A + BCD)^{-1}BC &= A^{-1}B(DA^{-1}B + C^{-1})^{-1} \end{aligned}$$

Use this result to show that

$$(V + XEX^T)^{-1} = V^{-1} - V^{-1}X(X^T V^{-1}X + E^{-1})^{-1}X^T V^{-1} \quad (1.43)$$

$$(V + XEX^T)^{-1}XE = V^{-1}X(X^T V^{-1}X + E^{-1})^{-1} \quad (1.44)$$

- (b) Use (1.44) to show that

$$(X^T(V + XEX^T)^\dagger X)^{-1} = E + (X^T V^{-1}X)^{-1}$$

- (c) Use this result and both (1.43) and the transpose of (1.44) to show that

$$M = (X^T V^{-1}X)^{-1} X^T V^{-1}$$

Note that E does not appear in this expression.

Exercise 1.93: A useful limit for matrix inverse and pseudoinverse

Let $U \in \mathbb{R}^{m \times r}$, and let $M \in \mathbb{R}^{m \times m}$ be positive semidefinite.

- (a) Show that for positive scalar α

$$\lim_{\alpha \rightarrow 0} (U^T M U + \alpha I_r)^{-1} U^T M = (U^T M U)^\dagger U^T M$$

²JBR would like to thank Travis Arnold of UW for helpful discussion of this exercise.

(b) Given this result show that for a positive definite matrix $D \in \mathbb{R}^{r \times r}$

$$\lim_{\alpha \rightarrow 0} U(U^T M U + \alpha D)^{-1} U^T M = \tilde{U}(\tilde{U}^T M \tilde{U})^\dagger \tilde{U}^T M$$

with $\tilde{U} = U\sqrt{D}^{-1}$. Without loss of generality, we take square root to be positive.

Exercise 1.94: Least squares with zero E and rank deficient V

In Exercise 1.92 we showed that for $V > 0$

$$M = (X^T V^\dagger X)^{-1} X^T V^\dagger$$

because we can set $E = 0$. We wish to extend this result to the case $V \geq 0$. Note that we cannot use the expression above for M because $X^T V^\dagger X$ can be singular for $V \geq 0$.

So let scalar $\alpha > 0$ and note that $V + \alpha I_m > 0$ so we have that

$$M_\alpha = (X^T (V + \alpha I_m)^{-1} X)^{-1} X^T (V + \alpha I_m)^{-1}$$

Derive a formula for M valid for $V \geq 0$ by setting $M = \lim_{\alpha \rightarrow 0} M_\alpha$ and taking this limit.

Exercise 1.95: Least squares with full rank E and rank deficient V ³

Let $X \in \mathbb{R}^{m \times n}$ have independent columns, $E \in \mathbb{R}^{n \times n}$ be positive definite, and $V \in \mathbb{R}^{m \times m}$ be semidefinite. We wish to show that

$$M = (X^T (V + X E X^T)^\dagger X)^{-1} X^T (V + X E X^T)^\dagger$$

is independent of E . This exercise subsumes Exercises 1.91 and 1.92, in which E was full rank, and V was either zero or full rank, respectively.

(a) First show that the indicated inverse exists and is given by

$$(X^T (V + X E X^T)^\dagger X)^{-1} = E + X^\dagger (V - V U_2 (U_2^T V U_2)^\dagger U_2^T V) (X^\dagger)^T$$

where the SVD of X is given by $X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} W^T$.

(b) Using this result, next show that

$$M = X^\dagger - X^\dagger V U_2 (U_2^T V U_2)^\dagger U_2^T$$

Note that M is independent of E for all $E > 0$. Finally, notice that this formula can be extended to $E \geq 0$ if we define the result for semidefinite E by taking the limit $E \rightarrow 0^+$.

³JBR would like to thank Travis Arnold of UW for helpful discussion of this exercise.

Exercise 1.96: Show candidate solutions are the same

In Exercises 1.94 and 1.95 we have two candidate solutions for the case $E = 0$ and $V \geq 0$. One is based on setting $E = 0$ with $V > 0$ and taking the limit as $V \rightarrow 0^+$. The other is based on setting $V \geq 0$ with $E > 0$ and taking the limit as $E \rightarrow 0^+$.

Show that these two candidate solutions are identical.

Hint: first show that for any matrix A , $(AA^T)^\dagger A = A(A^T A)^\dagger$.

Exercise 1.97: Pseudoinverse as limit

Show that for $A \in \mathbb{R}^{m \times n}$

$$\lim_{\alpha \rightarrow 0^+} (A^T A + \alpha I)^{-1} A^T = A^\dagger$$

Hint: show that for $A \neq 0$

$$\left\| (A^T A + \alpha I)^{-1} A^T - A^\dagger \right\|_2 = \frac{\alpha}{\underline{\sigma}(\alpha + \underline{\sigma}^2)}$$

in which $\underline{\sigma} > 0$ is the smallest singular value of A (Golub and Van Loan, 1996, Exercise 5.5.2).

Note that by taking the transpose of this result, we have also established that

$$\lim_{\alpha \rightarrow 0^+} A^T (AA^T + \alpha I)^{-1} = A^\dagger$$

Exercise 1.98: Concave quadratic optimization problems and eigenvalues/eigenvectors

In Section 1.5.2 we discussed that minimizing the strictly convex quadratic function gives a unique solution. Here we consider the opposite problem, minimizing a strictly *concave* quadratic function of $x \in \mathbb{R}^n$

$$\min_x f(x) = -x^T Q x$$

where $Q > 0$ so that $-Q < 0$. The solution to the unconstrained problem is not interesting because it is *unbounded*. Note that $f(x) \rightarrow -\infty$ as $x \rightarrow \infty$. But if we constrain the optimization, we can define an interesting problem. Consider the constrained problem

$$\min_x f(x) = -x^T Q x \quad \text{s.t. } x^T x \leq 1$$

This problem has a compelling geometric interpretation: minimize the concave quadratic function while searching over all x inside the unit sphere. It is easy to show by scaling arguments that the solution must lie *on* the sphere. So an equivalent problem is

$$\min_x f(x) = -x^T Q x \quad \text{s.t. } x^T x = 1 \quad (1.45)$$

where we have replaced the inequality constraint with an equality constraint.

- (a) Define the Lagrangian $L(x, \lambda)$ and unconstrained minmax problem that is equivalent to (1.45).

- (b) Write out the necessary conditions for a solution to the minmax problem, namely $L_x(x, \lambda) = 0, L_\lambda(x, \lambda) = 0$.
- (c) Solve the necessary conditions and show that the solution to (1.45) is

$$x^0 = \pm \bar{v} \quad \lambda^0 = \bar{\sigma} \quad f^0 = -\bar{\sigma}$$

where $(\bar{\sigma}, \bar{v})$ is the eigenvalue/eigenvector pair corresponding to the largest eigenvalue of Q , i.e., $\bar{\sigma} = \max(\text{eig}(Q))$ and $Q\bar{v} = \bar{\sigma}\bar{v}$.

- (d) Note that the necessary conditions are also satisfied by the eigenpair $(\underline{\sigma}, \underline{v})$ where $\underline{\sigma}$ is the *smallest* eigenvalue of Q . State the optimization problem solved by the eigenpair $(\underline{\sigma}, \underline{v})$.
- (e) Draw a sketch of your results for a two-dimensional problem, $x \in \mathbb{R}^2$.

Exercise 1.99: Generalization of concave quadratic optimization⁴

Consider the general version of the previous exercise

$$\min_x f(x) = -(x - b)^T B(x - b) \quad \text{s.t. } (x - c)^T C(x - c) \leq e$$

with $B > 0, C > 0, e > 0$. As in the previous exercise (where $b = c = 0$), when the value $x = b$ satisfies the constraint, we consider the companion problem

$$\min_x f(x) = (x - b)^T B(x - b) \quad \text{s.t. } (x - c)^T C(x - c) \geq e$$

But if $x = b$ does not satisfy the constraint, we consider the modified companion problem

$$\min_x f(x) = (x - b)^T B(x - b) \quad \text{s.t. } (x - c)^T C(x - c) \leq e$$

which is the only convex problem we have considered.

- (a) Show that these three problems can be expressed with fewer parameters as

$$\min_z -(z - d)^T D(z - d) \quad \text{s.t. } z^T z \leq 1 \quad (1.46)$$

$$\min_z (z - d)^T D(z - d) \quad \text{s.t. } z^T z \geq 1 \quad (1.47)$$

$$\min_z (z - d)^T D(z - d) \quad \text{s.t. } z^T z \leq 1 \quad (1.48)$$

with $D > 0$. Provide the transformation from z to x and expressions for D, d in terms of B, b, C, c, e so that the solutions of (1.46)-(1.47) can be used to solve the general problems.

⁴JBR would like to thank Robin Straesser of the University of Stuttgart for helpful discussion of this exercise.

- (b) Following the procedure of the previous exercise show that the necessary conditions for an optimal solution to all three problems, (1.46), (1.47), and (1.48) are

$$(D - \lambda I)z = Dd \quad z^T z = 1$$

Show that λ satisfying these necessary conditions can be found as the real eigenvalues (σ) of the following augmented eigenvalue problem

$$\begin{bmatrix} D & -I_n \\ -Ddd^T D^T & D \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \sigma \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (1.49)$$

- (c) Consider the following two examples.

$$D_1 = \begin{bmatrix} 1.2 & 0.5 \\ 0.5 & 0.75 \end{bmatrix} \quad d_1 = \begin{bmatrix} 0.25 \\ 0.5 \end{bmatrix} \quad D_2 = D_1 \quad d_2 = 3d_1$$

Find the solution to optimization problems (1.46) and (1.47) for the first example, and (1.46) and (1.48) for the second example. Plot the elliptical contours displaying the optimal solutions. Note the essential difference that d_1 is *inside* the unit circle (so $z = d_1$ satisfies the constraint), but d_2 is *outside* the unit circle (so $z = d_2$ does not satisfy the constraint).

Exercise 1.100: Maximum real eigenvalues of a matrix

The previous constrained optimization problems have established the following result on eigenvalues of a matrix.

Proposition 1.2 (Maximum real eigenvalues of a matrix). *Let $D \in \mathbb{R}^{n \times n} \geq 0$ have 2-norm (maximum singular value) $\|D\|$ and let $d \in \mathbb{R}^n$. Let λ_P denote the largest real eigenvalue of matrix P*

$$P = \begin{bmatrix} D & I_n \\ dd^T & D \end{bmatrix}$$

Then

$$(a) \lambda_P \geq \|D\| \quad \text{for all } d$$

$$(b) \lambda_P = \|D\| \quad \text{if and only if } d \in R(D - \|D\| I_n) \text{ and } \|(D - \|D\| I)d\| \leq 1$$

Prove this proposition based on your previous optimization results. Can you think of a simpler algebraic proof based only on eigenvalue/eigenvector properties?

Exercise 1.101: The ℓ_0 pseudonorm

In many optimization problems, a “sparse” solution, one with only a few nonzero elements, may be desired. The ℓ_0 “pseudonorm” of a vector x , denoted $\|x\|_0$, is the number of nonzero elements of x (in a given basis). For example, if $x = [0.6, 0, 3, 0]^T$, then $\|x\|_0 = 2$. Why is this quantity not a norm?

Exercise 1.102: Fredholm Alternative Theorem

The matrix version of the Fredholm Alternative Theorem reads as follows. For real matrix $A \in \mathbb{R}^{m \times n}$ and vectors x, y , and b : Either

- a) $Ax = b$ has at least one solution, or
- b) $A^T y = 0$ has a solution y that satisfies $y^T b \neq 0$.

Prove this theorem, using the results you have learned regarding the range and null space of a matrix. Start by letting y be a vector in the null space of A^T and taking the inner product between y and the equation $Ax = b$ to yield

$$y^T Ax = y^T b.$$

Either this equation can be satisfied or it cannot. Take it from here.

Exercise 1.103: Eigenvalues of AB and BA

Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$.

- (a) Show that the *nonzero* eigenvalues of AB and BA are the same.
- (b) Let $m = n$. Show that the eigenvalues of AB and BA are the same.

Exercise 1.104: QR decomposition for solving the normal equation

The standard method for solving least squares problems turns out not to be direct solution of the normal equation. A more computationally efficient approach is to use the QR decomposition of A .

- (a) If A is full rank (i.e. has rank = number of LI columns = n), then Q and R will also be full rank (each with n LI columns). Why? Start with the fact that, since A is full rank, the only x for which $Ax = 0$ is $x = 0$. Recall that Q is $m \times n$ and R is $n \times n$.
- (b) Given the normal equation, the QR decomposition of A , and the fact that Q and R have rank n , show that x is the solution to a triangular system of equations, (and thus trivial to solve).

Exercise 1.105: The Levi-Civita symbol and determinants

The Levi-Civita symbol will be introduced in Chapter 3 when defining the cross product. But we can also make good use it in linear algebra when analyzing determinants. Consider n integers $1, 2, \dots, n$ for some positive n . Given n indices, i_1, i_2, \dots, i_n , taking values $1, 2, \dots, n$, we define the Levi-Civita symbol $\epsilon_{i_1, i_2, \dots, i_n}$ as follows

$$\epsilon_{i_1 i_2 \dots i_n} = \begin{cases} 1 & \text{for } i_1 i_2 \dots i_n = 12 \dots n \\ 0 & \text{if any indices repeat} \\ (-1)^p & \end{cases}$$

where p is the number of exchanges required to put the indices i_1, i_2, \dots, i_n into the order $1, 2, \dots, n$.

This definition is quite useful in describing the matrix determinant. Let $n = 3$ and consider the $n \times n$ matrix A . Expansion by cofactors verifies the following result

$$\det A = \epsilon_{ijk} A_{i1} A_{j2} A_{k3}$$

We sum all the products of three elements of the matrix A taken from *different* rows and columns, and choose the sign of these terms exactly according to the sign rule for the Levi-Civita symbol. Moreover, there is no reason that we must associate the 123 column index of A with the ijk row index; we could equivalently express the determinant as

$$\det A = \epsilon_{ijk} A_{i3} A_{j1} A_{k2} = \epsilon_{ijk} A_{i2} A_{j3} A_{k1}$$

Therefore, we can generalize the expression for this 3×3 matrix to be

$$\det A \epsilon_{rst} = \epsilon_{ijk} A_{ir} A_{js} A_{kt}$$

Finally, for an $n \times n$ matrix A we have that

$$\det A \epsilon_{i_1 i_2 \dots i_n} = \epsilon_{j_1 j_2 \dots j_n} A_{j_1 i_1} A_{j_2 i_2} \cdots A_{j_n i_n} \quad (1.50)$$

Using (1.50) as the expression for the determinant, prove that for $A, B \in \mathbb{R}^{n \times n}$

$$\det AB = \det A \det B$$

as stated in the chapter.

Exercise 1.106: Singularity of $I + AB$ and $I + BA$

Let matrices A and B have dimensions so that AB is square. Show that $I + AB$ being (non)singular is equivalent to $I + BA$ being (non)singular.

Hint: Assume that $I + AB$ is singular, note that $B(I + AB) = (I + BA)B$, and show that $I + BA$ is singular.

Exercise 1.107: Least-squares solutions

Setting derivatives to zero to find solutions to least-squares problems as done in the text is useful to get started, but since this is only a necessary condition for optimality, it does not establish that the proposed solution is actually the minimizer, and it is silent about the uniqueness of the solution. So let's establish that these proposed solutions do in fact minimize the various least-squares objective functions, and that they are unique.

(a) To get started, consider the quadratic function

$$e(x) = (1/2)x^T A x$$

with $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n} > 0$. Show that the solution to the optimization problem $\min_x e(x)$ is

$$x^0 = 0 \quad e^0 = 0$$

and that this solution is unique.

(b) Consider next the function

$$f(x) = (1/2)x^T A x + b^T x + c$$

Show that the solution to the optimization problem $\min_x f(x)$ is

$$x^0 = -A^{-1}b \quad f^0 = -(1/2)b^T A^{-1}b + c$$

and that this solution is unique.

Hint: first show that $f(x)$ can be written as the bottom row in Table 1.1 and then use the result of the previous part.

(c) Consider next the function

$$g(x) = (1/2)(Ax - b)^T (Ax - b)$$

with $A \in \mathbb{R}^{m \times n}$ and assume that A has linearly independent columns. Show that the solution to the optimization problem $\min_x g(x)$ is

$$x^0 = (A^T A)^{-1} A^T b \quad g^0 = (1/2)b^T (I - A(A^T A)^{-1} A^T) b$$

and that this solution is unique.

Hint: multiply out the terms in $g(x)$ and use the result of the previous part.

(d) Finally consider the function

$$h(x) = (1/2)(Ax - b)^T R^{-1} (Ax - b)$$

with

with $A \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{m \times m}$ and assume that $R > 0$ and A has linearly independent columns. Show that the solution to the optimization problem $\min_x h(x)$ is

$$x^0 = (A^T R^{-1} A)^{-1} A^T R^{-1} b \quad h^0 = (1/2)b^T R^{-1} (R - A(A^T R^{-1} A)^{-1} A^T) R^{-1} b$$

and that this solution is unique.

Hint: Show that R has a square root $R = DD$ and redefine A and b in $h(x)$ to reduce this problem to the previous case.

Exercise 1.108: Spectroscopy and linear algebra—ideal mixture

The basics of spectroscopy are to study molecules by exposing them to various frequencies of electromagnetic radiation and observing the emission or absorption of the sample. Here we look at the basic mathematical problem of inferring the composition of a sample based on its measured spectrum and the spectrum of a standard comprising samples of known composition.

We consider a ternary mixture of three different molecular species, denoted A, B and C. Let the mixture's molar concentrations be denoted c_A, c_B, c_C . Denote the absorption/emission signal strength as a function of radiation frequency (or wavenumber) by $s_A, s_B, s_C \in \mathbb{R}^f$, where f is the number of frequencies at which we collect the spectrum. Figure 1.14 shows the absorption spectrum of the pure samples of A, B and C.

We stack these pure component spectra as columns in a matrix S , and the corresponding compositions of these samples, C , are then

$$S_0 = \begin{bmatrix} s_A & s_B & s_C \end{bmatrix} \quad C_0 = \begin{bmatrix} \begin{bmatrix} c_A \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ c_B \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ c_C \end{bmatrix} \end{bmatrix}$$

which are the three column vectors corresponding to the molar concentrations of the three pure components. Here $S_0 \in \mathbb{R}^{f \times 3}$, $C_0 \in \mathbb{R}^{3 \times 3}$. Typically we measure spectra at hundreds of frequencies so S_0 is quite a tall matrix.

Ideal mixture. We start the analysis with a simple model for the mixture, the ideal mixture assumption. For an ideal mixture with composition $c_m = (c_A, c_B, c_C)$ we assume the spectrum of the mixture s_m is given by the linear combination of the pure

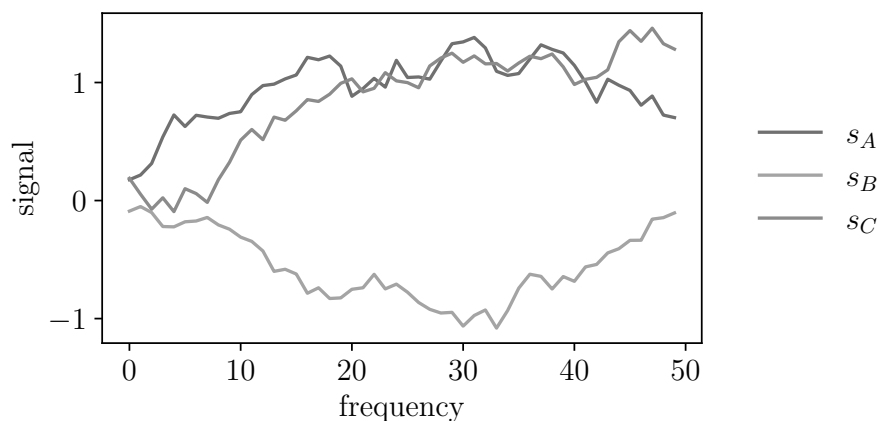


Figure 1.14: Measured absorption/emission spectra of pure components A, B, and C.

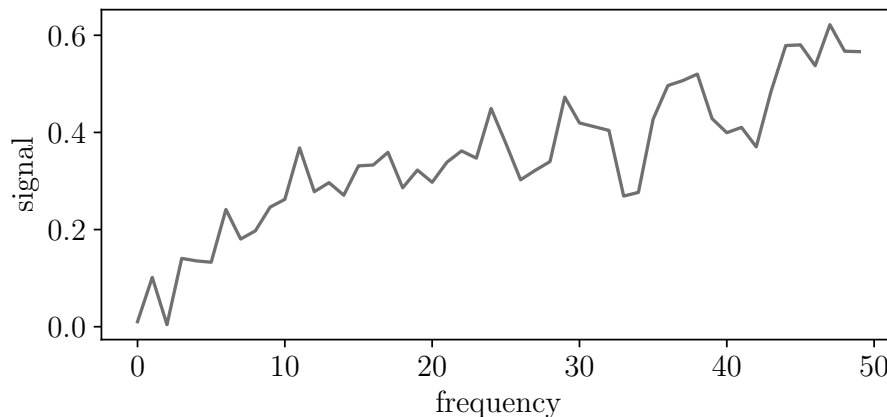


Figure 1.15: A measured spectrum of an (A,B,C) mixture of unknown composition.

component spectra weighted by the corresponding mole fraction in the mixture

$$s_m = S_0 c_m \quad (1.51)$$

So, given a measured spectrum $s_m \in \mathbb{R}^f$ of a mixture with unknown composition, we wish to estimate the composition $c_m \in \mathbb{R}^3$ with the known spectrum of the standard given by the $S_0 \in \mathbb{R}^{f \times 3}$ matrix. As (1.51) indicates, this estimation will involve some form of “solving” a linear system.

- (a) Create your own s_m spectrum for an equimolar mixture. Next “solve” (1.51) and recover c_m . Make sure that you obtain the value $c_m = (c/3)(1, 1, 1)$ where c is the total molar concentration of the mixture that you used to create the measured spectrum. Note that you are solving an overdetermined system with $f = 50$ equations and only 3 unknowns, the mixture composition.

Plot also the fit of your created and fitted spectra.

- (b) Now take the spectrum shown in Figure 1.15, which I created as you did in the previous part, but I chose a different composition and added measurement noise to s_m .

Give your estimated c_m .

Also plot the measured spectrum and your best fit of the spectrum. Comment on the quality of the fit to the measurement.

Exercise 1.109: Spectroscopy and linear algebra—nonideal mixture

The analysis of spectroscopic measurements would be a short discussion if all mixtures were ideal. Unfortunately, this is seldom the case for real mixtures of interest. So next we address how to estimate the composition of a nonideal mixture. The basic idea is to measure a more complete set of mixture compositions rather than only the three pure components. Consider the ten composition samples of mole fractions given in the following table

mixture	1	2	3	4	5	6	7	8	9	10
x_A	0	0	0	0	0.33	0.33	0.33	0.67	0.67	1.00
x_B	0	0.33	0.67	1.00	0	0.33	0.67	0	0.33	0
x_C	1.00	0.67	0.33	0	0.67	0.33	0	0.33	0	0

Here we have taken all combinations of compositions with a mole fraction increment of $1/3$. Given the *four* values chosen in the interval, $(0, (1/3), (2/3), 1)$, we have created a total of $(4)(3)/2 = 6$ mixtures. Note in the ideal case of the previous exercise, we required only the two end points of $(0,1)$ and had only $(2)(3)/2 = 3$ samples, i.e, the three pure component samples. The idea here is to measure the spectrum at all of these samples, so that we can account for the nonideality (nonlinearity) in the mixture when we solve the linear algebra problem for the composition.

So let $n_s \geq 3$ denote the total number of mixture compositions that we have chosen.⁵ We then measure the spectra of these n_s samples and place those in the S matrix as in the previous exercise. So now $S \in \mathbb{R}^{n_f \times n_s}$. The more mixture samples we choose, the more accurate our estimation method will be, but the more work we will have to do in the laboratory to prepare all the samples and measure their spectra.

Nonideal mixture. So now we create the spectral data for a nonideal mixture. We choose the following model for illustrative purposes. The spectrum of an ideal mixture with composition $c = (c_A, c_B, c_C)$ satisfies

$$s_{id}(c) = c_A s_A + c_B s_B + c_C s_C \quad (1.52)$$

as in the previous exercise. Here we add nonlinear binary interaction terms between each component in the mixture as follows

$$s(c) = s_{id}(c) + \gamma_{AB} c_A c_B s_{AB} + \gamma_{AC} c_A c_C s_{AC} + \gamma_{BC} c_B c_C s_{BC} \quad (1.53)$$

where scalars $\gamma_{AB}, \gamma_{AC}, \gamma_{BC}$ represent the strength of the interactions. Setting $\gamma_{AB} = \gamma_{AC} = \gamma_{BC} = 0$ recovers the ideal solution. The spectrum chosen for the three interaction terms are shown in Figure 1.16. These are not supposed to have physical significance; they are chosen here simply to create a nonideal solution to test our composition estimation scheme.

⁵Note that the mixture samples do not need to be evenly spaced as we have shown above.

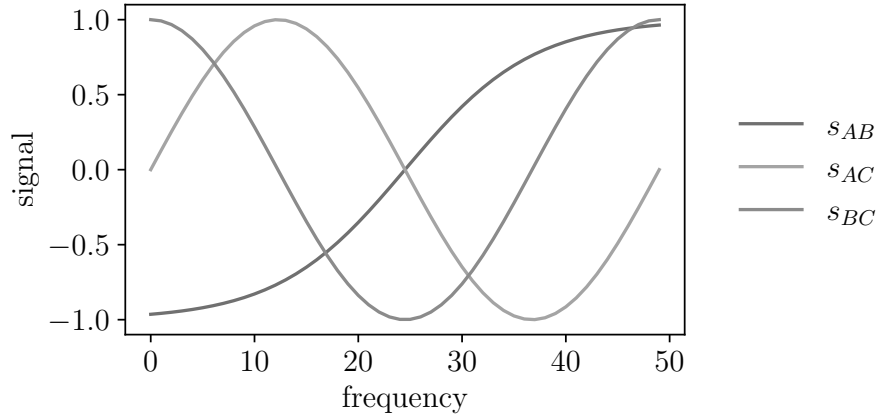


Figure 1.16: The interaction spectra used to create the nonideal mixture data.

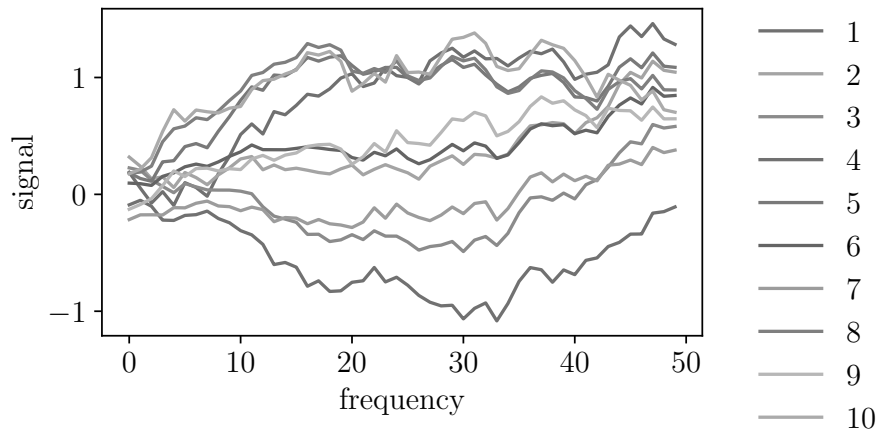


Figure 1.17: Measured absorption/emission spectra of the 10 mixtures in the standard.

Given the nonideal mixture model in (1.53) and spectral interactions given in Figure 1.16, we then create the spectral dataset for the ten mixture samples given in matrix C . These are shown in Figure 1.17 for the parameter values $\gamma_{AB} = \gamma_{AC} = \gamma_{BC} = 1$. So in an application, Figure 1.17 would be provided by actual spectroscopic measurements of the sample mixtures.

So now we are prepared for the estimation step, i.e., given these spectra of the

n_s samples in our standard, how do we estimate the unknown composition of a new mixture from its measured spectrum? We do this in two stages. First we estimate the linear combination of the n_s spectra in the standard that best fit the new measured spectrum. In mathematical terms, we now wish to solve for vector $\alpha \in \mathbb{R}^{n_s}$ so that

$$s_m = S\alpha$$

Since n_f is typically much larger than n_s , we again have an overdetermined system to estimate α . Denote the estimate by

$$\hat{\alpha} = S^+ s_m \tag{1.54}$$

where S^+ denotes the pseudoinverse of S that we have discussed in class. Once we have estimated α , we then compute the estimated mixture composition by

$$c_m = C\hat{\alpha} = CS^+ s_m = R s_m \tag{1.55}$$

where $C \in \mathbb{R}^{3 \times n_s}$ as shown in the table above is our composition matrix, and $R = CS^+$. Since we know C and S and can therefore compute S^+ and R offline, obtaining c_m from s_m is simply a matrix multiply by R and is a *fast* operation suitable for an *online* composition “measurement.”⁶

- (a) Give the dimension of S^+ and R . What is the rank of C ? What is the dimension of the null space of C ?
- (b) Show that $\text{rank}(S) = 3$ for all $n_s \geq 3$ if the mixture is ideal. So there is no reason to use more than $n_s = 3$ samples in the standard for an ideal mixture. What is the rank of S for the data given in this problem?

Note that by how much the rank of S exceeds the ideal case of 3 provides a rough measure of how nonideal the mixture is.

Download the data provided on the class website. What is the rank of S for the data shown in Figure 1.17?

Calculate the SVD of the downloaded S . How many singular values do you consider to be above the noise level in the data? By how much does this number exceed the value of 3 for an ideal solution. Can you provide a physical explanation for this result?

- (c) Consider the measured spectrum displayed in Figure 1.18. Inverting only the singular values considered significant in the previous part, give the estimate $\hat{\alpha}$,

⁶Note that, as in most applications, composition is *not* measured. What is measured is an emission/absorption spectrum. After multiplication by R , we have a composition *estimate*. But it is conventional shorthand to blur this distinction and call this a composition *measurement* when, e.g., designing a feedback controller for a chemical reactor equipped with an online spectroscopic sensor.

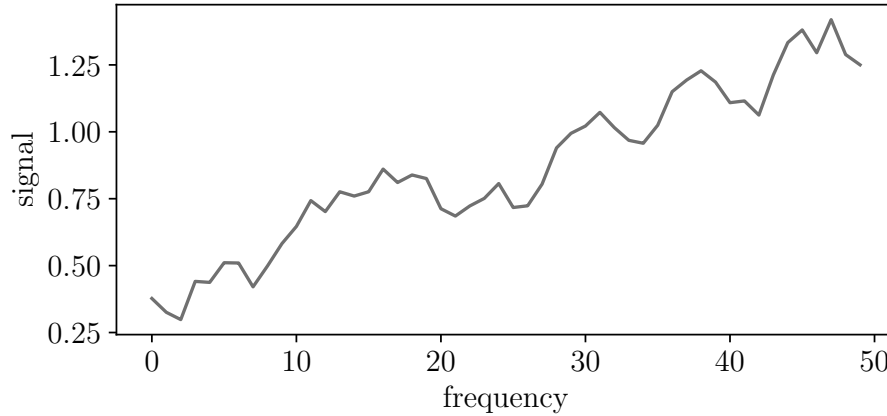


Figure 1.18: A measured spectrum of an (A,B,C) mixture of unknown composition.

and the estimated composition c_m in (1.54)-(1.55). Plot the spectral data in Figure 1.18 and the corresponding fit based on your estimated c_m . How well do you fit the data?

- (d) Now assume that you consider *all* the singular values of S to be significant. Recalculate $\hat{\alpha}$ and c_m under this assumption. Again, plot the spectral data in Figure 1.18 and the corresponding fit based on this estimated c_m . How well do you fit the data?

Which estimate of c_m do you think is better and why?

Exercise 1.110: Derive the singular value decomposition

We wish to show that for any $A \in \mathbb{C}^{m \times n}$, there exists a singular value decomposition (SVD) $A = USV^*$ in which $U \in \mathbb{C}^{m \times m}$, $V \in \mathbb{C}^{n \times n}$ are unitary, and $S \in \mathbb{R}^{m \times n}$ has the structure

$$S = \begin{matrix} & r & n-r \\ \begin{matrix} r \\ m-r \end{matrix} & \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \end{matrix}$$

in which r is the rank of the A matrix. The matrix Σ is diagonal and real

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$$

Finally, we wish to show that the SVD matrices U and V satisfy the properties shown in Figure 1.4, i.e., the columns of V_1, V_2, U_1 , and U_2 are unitary bases for the subspaces $R(A^*), N(A), R(A)$, and $N(A^*)$, respectively.

- (a) Begin by assuming $A \in \mathbb{R}^{m \times n}$ and $A \neq 0$. Show that for $A \in \mathbb{R}^{m \times n}$ of rank r , $N(A^T A) = N(A)$ and $R(A^T A) = R(A^T)$. Show that $\text{rank}(A^T A) = r$.
- (b) Consider the (ordered) symmetric Schur decomposition (Theorem 1.17) of the symmetric $A^T A$ matrix

$$A^T A = V S_V V^T$$

where $V \in \mathbb{R}^{n \times n}$ is orthonormal and S_V is diagonal. As shown in Section 1.4.7, the diagonal elements of S_V are nonnegative real numbers. Since $A^T A$ has rank r , there are r positive eigenvalues and $n - r$ zero eigenvalues of $A^T A$. So $A^T A$ can be partitioned as

$$A^T A = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = V_1 \Sigma^2 V_1^T$$

Show that the columns of V_1 are an orthonormal basis for $R(A^T)$ and the columns of V_2 are an orthonormal basis for $N(A)$.

- (c) Define $U_1 = AV_1 \Sigma^{-1}$ and show that $U_1 \in \mathbb{R}^{m \times r}$ and $U_1^T U_1 = I_r$. Show that the columns of U_1 are an orthonormal basis for $R(A)$.
- (d) Define U_2 so that its columns are an orthonormal basis for $N(A^T)$. Show that $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{m \times m}$ is an orthonormal matrix.
- (e) Show that $A^T U = V S^T$ and therefore

$$A^T = V S^T U^T$$

Taking the transpose gives the singular value decomposition of A with the properties depicted in Figure 1.4.

- (f) How do you then treat the special case $A = 0$, and how do you generalize the approach to treat $A \in \mathbb{C}^{m \times n}$.

Exercise 1.111: Existence of solutions to games for quadratic functions

Consider a quadratic $V(x, y)$ whose contours are shown Figure 1.19. Note that the two solid straight lines intersecting at the origin are the $V = 0$ contours, and that V is both positive and negative in quadrants I and III, and negative only in quadrants II and IV.

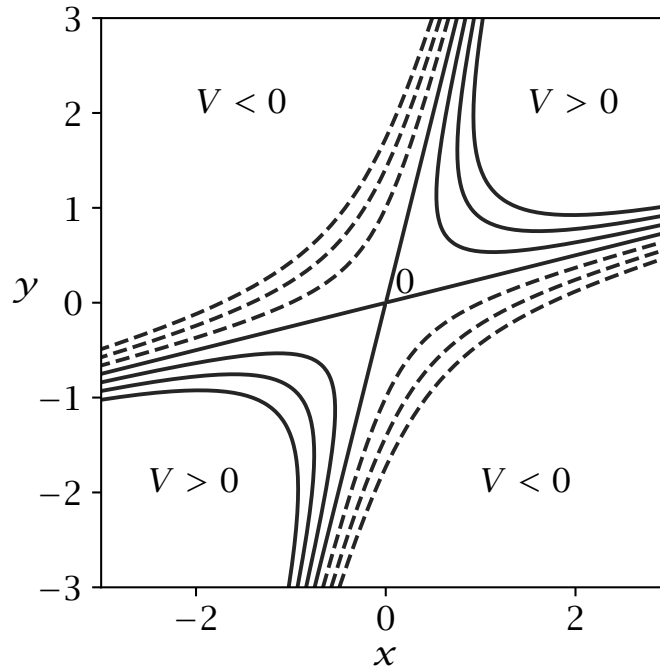


Figure 1.19: Contour lines for quadratic function $V(x, y) = \text{constant}$. Solid lines are positive V contours and dashed lines are negative V contours.

Denote the optimizers of the single variable optimizations as follows:

$$\begin{aligned} \arg \min_y V(x, y) &= y_0(x) & \arg \max_x V(x, y) &= x^0(y) \\ \arg \max_y V(x, y) &= y^0(x) & \arg \min_x V(x, y) &= x_0(x) \end{aligned}$$

Denote the four different games by:

$$\begin{aligned} \max_x \min_y V(x, y) &= \max_x V(x, y_0(x)) & \min_y \max_x V(x, y) &= \min_y V(x^0(y), y) \\ \min_x \max_y V(x, y) &= \min_x V(x, y^0(x)) & \max_y \min_x V(x, y) &= \max_y V(x_0(x), y) \end{aligned}$$

Without knowing exactly the numerical values of the contours, answer the following questions from inspecting the geometry of the contours of $V(x, y)$.

- (a) Does $V(x, y)$ have a stationary point, i.e., a value (x, y) such that $\partial V / \partial(x, y) = 0$?
If so, where is it located?

Hint: yes, there is a single stationary point.

- (b) Which of the four single optimization problems defined above have solutions? For what values of the independent variable do these solutions exist?

Hint: only two of the single optimizations have solutions, and these solutions are defined for all values of the independent variable.

- (c) Given your answer to the previous problem, which of the four games have solutions?

Hint: only two of the four games have solutions.

- (d) Do any of these problems satisfy strong duality, i.e., the order of the optimizations in the game can be reversed without changing the solution?

Hint: no.

- (e) For the games that have solutions, what are their optimal $x, y, V(x, y)$ values?

Exercise 1.112: Existence of solutions to games for quadratic functions

Now consider the modified quadratic $V(x, y)$ whose contours are shown Figure 1.20. Note that Figure 1.20 is simply a rotation of the V function in Figure 1.19 by about $+20^\circ$. Now V has both positive and negative contours in all four quadrants.

- (a) Which of the four single optimization problems defined in Exercise 1.111 have solutions? For what values of the independent variable do these solutions exist?

Hint: only two of the single optimizations have solutions, and these solutions are defined for all values of the independent variable.

- (b) Given your answer to the previous problem, which of the four games have solutions?

Hint: only two of the four games have solutions.

- (c) Do any of these problems satisfy strong duality, i.e., the order of the optimizations in the game can be reversed without changing the solution?

Hint: Yes! The origin is therefore a saddle point for the two games that have solutions.

- (d) Contrast the function depicted in Figure 1.20 with the one in Figure 1.19 of Exercise 1.111. Both functions seem to have a similar saddle-point geometry. What feature in the geometry causes one of them to have a saddle point and the other not to have a saddle point?

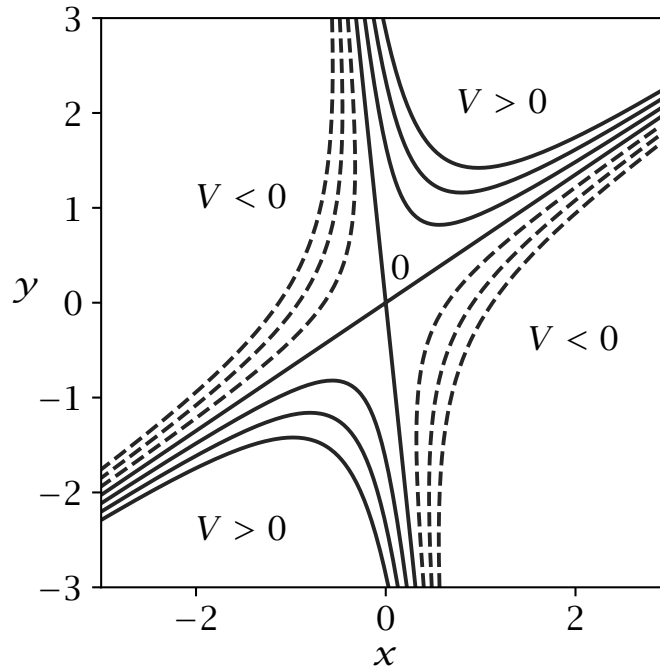


Figure 1.20: Contour lines for quadratic function $V(x, y) = \text{constant}$. Solid lines are positive V contours and dashed lines are negative V contours.

- (e) Find matrices A in $V(x, y) = (x, y)^T A(x, y)$ corresponding to Figure 1.19 and Figure 1.20. What are their eigenvalues?

Hint: The slopes of the straight lines for the $V = 0$ contour in Figure 1.19 were chosen as 4 and $1/4$. Start by building a quadratic function that vanishes on those two lines, and calculate A for that function. Rotate the axes by -20° to find A for Figure 1.20. See also Example 1.5 for discussion of rotating vectors with matrix multiplication.

Exercise 1.113: Some null space relationships

Given symmetric $D \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{m \times n}$, let the SVD of $A = USV = U_1 \Sigma V_1^T$, with $A^+ = V_1 \Sigma^{-1} U_1^T$ and the columns of V_2 serving as an orthonormal basis for $N(A)$, the nullspace of A . Consider the following partitioned matrices

$$M_1 = \begin{bmatrix} D & -A^T \\ -A & 0 \end{bmatrix} \quad M_2 = \begin{bmatrix} V_2^T D \\ -A \end{bmatrix}$$

Define the following sets in \mathbb{R}^n

$$\mathbb{N}_1 = \{u \mid \text{there exists } \lambda \in \mathbb{R}^m \text{ with } \begin{bmatrix} u \\ \lambda \end{bmatrix} \in N(M_1)\}$$

$$\mathbb{N}_2 = N(M_2)$$

$$\mathbb{N}_3 = V_2 N(V_2^T D V_2)$$

Show that these three sets are equal.

Hint: Show $\mathbb{N}_1 = \mathbb{N}_2$ and $\mathbb{N}_2 = \mathbb{N}_3$.

Exercise 1.114: Some range space relationships

Given symmetric $D \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{m \times n}$, let the SVD of $A = USV = U_1 \Sigma V_1^T$, with $A^+ = V_1 \Sigma^{-1} U_1^T$ and the columns of V_2 serving as an orthonormal basis for $N(A)$, the nullspace of A . Consider the following partitioned matrices

$$M_1 = \begin{bmatrix} D & -A^T \\ -A & 0 \end{bmatrix} \quad M_2 = \begin{bmatrix} V_2^T D \\ -A \end{bmatrix} \quad M_3 = \begin{bmatrix} V_2^T & V_2^T D A^+ \\ 0 & I \end{bmatrix}$$

Define the following sets in \mathbb{R}^n

$$\mathbb{R}_1 = R(M_1)$$

$$\mathbb{R}_2 = \{(d, b) \mid \begin{bmatrix} V_2^T d \\ b \end{bmatrix} \in R(M_2)\}$$

$$\mathbb{R}_3 = \{(d, b) \mid M_3 \begin{bmatrix} d \\ b \end{bmatrix} \in R\left(\begin{bmatrix} V_2^T D V_2 & 0 \\ 0 & A \end{bmatrix}\right)\}$$

Show that these three sets are equal.

Hint: Show $\mathbb{R}_1 = \mathbb{R}_2$ and $\mathbb{R}_2 = \mathbb{R}_3$.

Exercise 1.115: And finally, the partitioned matrix pseudoinversion formula

Given symmetric $D \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{m \times n}$, let the SVD of $A = USV = U_1 \Sigma V_1^T$, with $A^+ = V_1 \Sigma^{-1} U_1^T$ and the columns of V_2 serving as an orthonormal basis for $N(A)$, the nullspace of A . Assume that both V_1 and V_2 are nonzero and that $V_2^T D V_2 \neq 0$.⁷

Consider the following partitioned matrices

$$M = \begin{bmatrix} D & -A^T \\ -A & 0 \end{bmatrix} \quad X = \begin{bmatrix} W & (WD - I)A^+ \\ (A^T)^+(DW - I) & (A^T)^+(DWD - D)A^+ \end{bmatrix}$$

⁷See the next exercise for relaxing these assumptions.

where

$$W = V_2(V_2^T D V_2)^+ V_2^T$$

Show that $M^+ = X$.

Hint: It is useful to first show that $V_2^T(DWD - D) = 0$.

Exercise 1.116: Trapping the edge cases

Given symmetric $D \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{m \times n}$, let the SVD of $A = USV = U_1 \Sigma V_1^T$, with $A^+ = V_1 \Sigma^{-1} U_1^T$ and the columns of V_2 serving as an orthonormal basis for $N(A)$, the nullspace of A . Consider the following partitioned matrices

$$M = \begin{bmatrix} D & -A^T \\ -A & 0 \end{bmatrix} \quad X = \begin{bmatrix} W & (WD - I)A^+ \\ (A^T)^+(DW - I) & (A^T)^+(DWD - D)A^+ \end{bmatrix}$$

where

$$W = V_2(V_2^T D V_2)^+ V_2^T$$

We wish to show when $X = M^+$, the unique Moore-Penrose pseudoinverse of M , and when X is a reflexive generalized inverse of M , i.e. X satisfies only $MXM = M$ and $XXM = X$, but *not* $MX = (MX)^T$ and *not* $XM = (XM)^T$. Thus, X is a reflexive generalized inverse when it satisfies conditions (1) and (2) of Definition 1.1, but not conditions (3) and (4).

We shall break the problem into several edge cases to streamline the development.

- (a) First consider $A = 0$. Note that there is no V_1 matrix in the SVD of A in this case. Show that

$$X = \begin{bmatrix} D^+ & 0 \\ 0 & 0 \end{bmatrix}$$

and verify that $X = M^+$ by checking the four conditions of Definition 1.1.

- (b) Next let A have independent columns. Note that there is no V_2 matrix in the SVD of A in this case. For this case, interpret the W formula to mean $W = 0 \in \mathbb{R}^{n \times n}$. Again, verify that $X = M^+$ for this case.

For the final three cases, we assume $A \neq 0$ and A 's columns are not linearly independent. Hence, there are nonzero V_1 and V_2 matrices.

- (c) Next consider the case $V_2^T D V_2 \neq 0$. Verify that $X = M^+$ for this case. Some would consider this case to be the main case of interest, and the other cases to be the edge cases.
- (d) Next consider the case with $V_2^T D V_2 = 0$ and $D V_2 = 0$. Verify that $X = M^+$ for this case.

- (e) Finally, consider the possibility that $V_2^T D V_2 = 0$ and $V_2 V_2^T D A^+ \neq 0$. Verify that X is *not* the pseudoinverse of M but is a reflexive generalized inverse of M for this case.

Show also that if D is positive (or negative) semidefinite, $V_2^T D V_2 = 0$ if and only if $D V_2 = 0$, which implies $V_2 V_2^T D A^+ = 0$. Show that $D V_2 = 0$ is sufficient for $V_2^T D V_2 = 0$, but it is *not* necessary when D is symmetric but *indefinite*. Therefore, just because $V_2^T D V_2 = 0$, we cannot rule out that $V_2 V_2^T D A^+ \neq 0$ when D is indefinite.

To settle the matter, provide a numerical example of $D \in \mathbb{R}^{2 \times 2}$ and $A \in \mathbb{R}^{1 \times 2}$ that shows this last case is in fact possible. Calculate M^+ for this example and show that it is not equal to X , and that MX and XM are not symmetric matrices.

2

Ordinary Differential Equations

Exercise 2.80: Changing variables of integration

(a) For the two-dimensional integral, derive the formula

$$\int_{\mathbb{A}_x} g(x_1, x_2) dx_1 dx_2 = \int_{\mathbb{A}_y} \bar{g}(y_1, y_2) |\det(\partial x / \partial y)| dy_1 dy_2$$

in which the variable transformation is $y = f(x)$, which is assumed invertible, $\bar{g}(y) = \bar{g}(y(x)) = g(x)$, $\partial x / \partial y$ is the Jacobian matrix of the transformation, and the area of integration in the y -variables is given by

$$\mathbb{A}_y = \{(y_1, y_2) \mid (y_1, y_2) = f(x_1, x_2), (x_1, x_2) \in \mathbb{A}_x\}$$

Hint: consider the rectangular element of integration shown in Figure 2.35 with area $dx_1 dx_2$, and find its area under the transformation f ; then use the properties of the determinant derived in Exercise 1.82.

(b) How does this result generalize to three-dimensional integrals? n -dimensional integrals?

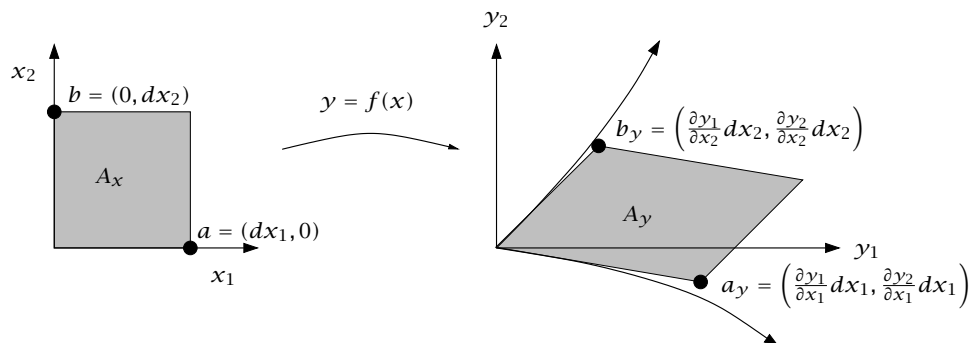


Figure 2.35: Change of area element due to coordinate transformation

Exercise 2.81: Existence and uniqueness of a nonhomogeneous boundary-value problem

Consider the second-order nonhomogeneous boundary-value problem

$$Lu = f \quad B_1 u = 0 \quad B_2 u = 0 \quad (2.104)$$

for unknown function $u(x)$, $x \in [0, 1]$, in which $f(x)$ is given. The differential operator and two boundary functionals are defined by

$$\begin{aligned} Lu(x) &= \frac{d^2}{dx^2} u(x) + k^2 u(x) \\ B_1 u(x) &= u(0) \\ B_2 u(x) &= u(1) \end{aligned}$$

- (a) Find the adjoint operator and boundary functionals L^* , B_1^* , B_2^* so that $(Lu, v) = (u, L^*v)$ for all $u(x)$, $v(x)$ satisfying $B_1 u = 0$, $B_2 u = 0$, $B_1^* v = 0$, and $B_2^* v = 0$.

Is the boundary-value problem self adjoint?

- (b) Find the null space of L , $N(L)$, for this boundary-value problem. Find also $N(L^*)$.
- (c) Given your results, what does the alternative theorem say about the existence and uniqueness of the *nonhomogeneous* BVP 2.104? In particular, does the existence of the solution depend on the form of $f(x)$?

Exercise 2.82: More forcing on the boundary

We may wonder how general is the technique to move nonhomogeneity from the boundary conditions to the differential equation using impulses. To explore this issue, consider the fully nonhomogeneous second-order BVP for $u(x)$, with $x \in [0, 1]$

$$Lu = f \quad B_1 u = \gamma_1 \quad B_2 u = \gamma_2$$

with the general form

$$\begin{aligned} Lu &= u_{xx} + \rho_1 u_x + \rho_0 u \\ \begin{bmatrix} B_1 u \\ B_2 u \end{bmatrix} &= \begin{bmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ u_x(0) \\ u_x(1) \end{bmatrix} \end{aligned}$$

and we assume that the rows of the B matrix are linearly independent so the boundary conditions are well posed for a second-order BVP.

- (a) The first task is to define the adjoint problem. Find operator L^* and $J(u, v)$ so that

$$\langle Lu, v \rangle = \langle u, L^*v \rangle + J(u, v) \Big|_0^1$$

(b) Define the following vectors to contain the boundary information

$$\mathbf{u} = [u(0) \quad u(1) \quad u_x(0) \quad u_x(1)]^T \quad \mathbf{v} = [v(0) \quad v(1) \quad v_x(0) \quad v_x(1)]^T$$

and find matrix M such that

$$J(u, v)|_0^1 = \mathbf{v}^T M \mathbf{u}$$

What is the rank of matrix M ?

(c) We define the adjoint boundary conditions so that $J(u, v)|_0^1 = 0$ for all \mathbf{u} satisfying $B\mathbf{u} = 0$ and \mathbf{v} satisfying $\tilde{B}\mathbf{v} = 0$.¹

Equivalently, we are saying that $\mathbf{v}^T M \mathbf{u} = 0$ for all $\mathbf{u} \in N(B)$ and $\mathbf{v} \in N(\tilde{B})$. Let the columns of matrix N_B be a basis for $N(B)$, the null space of B , and let the columns of matrix $N_{\tilde{B}}$ be a basis for $N(\tilde{B})$, the null space of \tilde{B} .

What is the rank and dimension of matrix N_B ? Justify your answer.

(d) We then have that $\mathbf{u} \in N(B)$ and $\mathbf{v} \in N(\tilde{B})$ are of the form $N_B\alpha$ and $N_{\tilde{B}}\beta$, respectively, for arbitrary vectors α, β . So we have the condition

$$\beta^T (N_{\tilde{B}}^T M N_B) \alpha = 0 \quad \text{for all } \alpha, \beta$$

Show that this requirement implies

$$N_{\tilde{B}}^T M N_B = 0 \tag{2.105}$$

(e) So we can compute the adjoint BC $N_{\tilde{B}}$ matrix by defining its columns to be a basis for $N((MN_B)^T)$. In MATLAB or Octave,

$$N_B_tilde = \text{null}((M*N_B)')$$

What is the rank and dimension of matrix $N_{\tilde{B}}$ given the ranks of M and N_B . Justify your answer.

(f) With the linear algebra preliminaries out of the way, we are ready to move the nonhomogeneity into the differential equation. We consider adding a term to the ODE of the form

$$p(x) = [p_1 \quad p_2 \quad p_3 \quad p_4] \begin{bmatrix} \delta(x) \\ \delta(x-1) \\ -\delta_x(x) \\ -\delta_x(x-1) \end{bmatrix}$$

¹We use matrix \tilde{B} to denote the adjoint's boundary functional matrix in place of B^* to avoid conflicting with the notation for the adjoint (conjugate, transpose) of matrix B .

where we consider adding singlets and doublets to both ends of the interval $[0, 1]$. Note the minus sign on the last two elements.

Show that this $p(x)$ produces the following integral

$$\langle p, v \rangle = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix} \begin{bmatrix} v(0) \\ v(1) \\ v_x(0) \\ v_x(1) \end{bmatrix}$$

or

$$\langle p, v \rangle = \mathbf{p}^T \mathbf{v} = \mathbf{v}^T \mathbf{p}$$

- (g) The solvability condition is $\langle f, v_k \rangle = J(u, v_k)|_0^1 = \mathbf{v}_k^T M \mathbf{u}$ where $v_k(x)$ is any solution to the fully homogeneous adjoint problem, and $u(x)$ is the solution to the nonhomogeneous problem. To make $\hat{f} = f + p$ orthogonal to v_k , show that we require that

$$\mathbf{v}_k^T (M \mathbf{u} + \mathbf{p}) = 0$$

- (h) Since $u(x)$ satisfies the nonhomogeneous problem, $B \mathbf{u} = \gamma$. Show that all \mathbf{u} satisfying this restriction are given by $\mathbf{u} = B^\dagger \gamma + N_B \alpha$ with α arbitrary. Since $v_k(x)$ satisfies the homogeneous adjoint problem, it satisfies $\mathbf{v}_k = N_{\tilde{B}} \beta$ with β arbitrary. Therefore, show that \mathbf{p} satisfying

$$\mathbf{p} = -M B^\dagger \gamma$$

satisfies the orthogonality equation of the previous part. Is this perturbation unique? Justify your answer.

We have therefore found a perturbation $p(x)$ using impulses at the boundaries for any second-order nonhomogeneous problem. The same procedure works for any n th-order nonhomogeneous BVP with n linearly independent nonhomogeneous boundary conditions.

Exercise 2.83: Forcing on the boundary without tears

For problems arising in applications, we usually do not face the completely general nonhomogeneous BVP of the last exercise. For most of the common BVPs we can deduce the forcing on the boundary by inspection. To see this consider boundary conditions of the form

$$B \mathbf{u} = \gamma \quad B = \begin{bmatrix} B_{11} & B_{12} & B_{13} & 0 \\ 0 & B_{22} & B_{23} & B_{24} \end{bmatrix}$$

If we ever see two columns with zeros in different rows, we can create the forcing term by inspection. First move the fourth column to the second column so that we have

$$B = \begin{bmatrix} B_{11} & 0 & B_{12} & B_{13} \\ 0 & B_{24} & B_{22} & B_{23} \end{bmatrix}$$

Note that we have now also switched the boundary data to

$$u = [u(0) \quad u_x(1) \quad u(1) \quad u_x(0)]$$

in the boundary condition $Bu = \gamma$. From the first row of B , when we change to $Bu = 0$, we require $B_{11}u(0^+) = \gamma_1$, a jump in $u(0)$ of size γ_1/B_{11} . From the second row of B , we require a jump in $u_x(1^+)$ of size $-\gamma_2/B_{24}$. Since these jumps handle the nonhomogeneity in the BCs, we can set $u(1) = u_x(0) = 0$ for convenience. Next we examine what the selected jumps do to the differential operator L

$$\begin{aligned} u_{xx} &= (\gamma_1/B_{11})\delta_x(x) - (\gamma_2/B_{24})\delta(x-1) \\ u_x &= (\gamma_1/B_{11})\delta(x) \end{aligned}$$

So the total change to the operator $Lu = u_{xx} + \rho_1 u_x + \rho_0 u$ is

$$Lu = f + \frac{\gamma_1}{B_{11}}(\delta_x(x) + \rho_1 \delta(x)) - \frac{\gamma_2}{B_{24}}\delta(x-1)$$

Check that the solvability condition for the original nonhomogeneous problem is equivalent to the orthogonality relation $\langle \hat{f}, v_k \rangle = 0$ with modified $\hat{f} = f + p$ with p defined above.

Exercise 2.84: Amaze your friends; homogenous BVPs by inspection

Given the general second-order nonhomogeneous, $u_{xx} + \rho_1 u_x + \rho_0 u = f$, $B_1 u = \gamma_1$, $B_2 u = \gamma_2$, consider the following special cases that commonly arise in applications. Give ODE perturbations $p(x)$ for $Lu = f + p$ that provide orthogonal solvability conditions $\langle f + p, v_k \rangle = 0$.

- (a) $\rho_1 = 0, B_1 u = u(0), B_2 u = u(1)$.
- (b) $\rho_1 = 0, B_1 u = u_x(0), B_2 u = u_x(1)$.
- (c) $\rho_1 = 0, B_1 u = u(0) - u(1), B_2 u = u_x(0) + u_x(1)$.

3

Vector Calculus and Partial Differential Equations

Exercise 3.60: Brief transport discussion

- (a) In polar coordinates (r, θ) , calculate $\nabla \cdot \mathbf{x}$ in which \mathbf{x} is the position vector, $\mathbf{x} = r\mathbf{e}_r$.
- (b) Show that you obtain the same result for $\nabla \cdot \mathbf{x}$ if you operate in Cartesian coordinates (x, y) with $\mathbf{x} = x\mathbf{e}_x + y\mathbf{e}_y$. How do you suppose this result generalizes to an n -dimensional space?
- (c) Starting from a statement of conservation of moles in a volume element, derive the continuity equation

$$\frac{\partial}{\partial t} c_A = -\nabla \cdot c_A \mathbf{v}_A + R_A$$

in which c_A is the molar concentration of species A, $c_A \mathbf{v}_A$ is the molar flux vector of species A, and R_A is the molar rate/volume of generation of species A. What is a common model for the flux of species A in a multicomponent fluid with average velocity \mathbf{v} ?

Exercise 3.61: A sphere with prescribed heat flux at its surface

A sphere of radius R , initially at uniform temperature T_0 , is placed in a uniform radiation field that delivers constant heat flux F_0 normal to the surface of the sphere. We wish to find the transient temperature response of the sphere.

- (a) Starting with the general energy balance

$$\rho \hat{C}_p \frac{\partial}{\partial t} T = \nabla \cdot k \nabla T$$

reduce the model as far as possible in spherical coordinates for the conditions given in the problem statement. State appropriate boundary and initial conditions.

- (b) Using the following dimensionless variables

$$\tau = \frac{k}{\rho \hat{C}_p R^2} t \quad \xi = \frac{r}{R} \quad \Theta = \frac{T - T_0}{T_0}$$

nondimensionalize the PDE and IC and BCs. How many dimensionless parameters are there in this problem?

- (c) Take the Laplace transform of your PDE, IC, and BCs and show that the transform satisfies the second-order ODE

$$\frac{1}{\xi^2} \frac{d}{d\xi} \left(\xi^2 \frac{d\bar{\Theta}}{d\xi} \right) - s\bar{\Theta} = 0$$

Choose two linearly independent solutions to this equation that allow you to eliminate one of the solutions easily using the BCs.

- (d) Solve the ODE, apply the two BCs, and show that

$$\bar{\Theta}(\xi, s) = \frac{\beta}{\xi} \frac{\sinh \sqrt{s}\xi}{s(\sqrt{s} \cosh \sqrt{s} - \sinh \sqrt{s})}$$

in which β is a dimensionless heat flux.

- (e) To invert the transform, we need to find and classify the order of the singularities of $\bar{\Theta}(s)$. Obviously there is a singularity at $s = 0$ because of the s term in the denominator.

Now for the other term. Using the substitution $\sqrt{s} = i\alpha$, for what α is $\sqrt{s} \cosh \sqrt{s} - \sinh \sqrt{s} = 0$? Draw a sketch of these $\alpha_n, n = 1, 2, 3, \dots$ roots. You should find an infinite number of them. These are all simple zeros.

But notice that $\alpha = 0$, which implies $s = 0$, is a zero of this term as well, so the zero at $s = 0$ is *second order*.

- (f) So the structure of the inverse of the transform is

$$\Theta(\xi, \tau) = a_{00}\tau + a_{01} + \sum_{n=1}^{\infty} a_n e^{-\alpha_n^2 \tau}$$

where the first two terms come from the double zero at $s = 0$. Find the $a_n, n \geq 1$ using the Heaviside theorem since these are simple zeros.

- (g) Find a_{00} and a_{01} using the Heaviside theorem for repeated roots for the root at $s = 0$.
- (h) Given the appearance of this linear τ term, what is the final value of the temperature? Discuss the physical significance of this result, and critique the model we are using.

Exercise 3.62: Steady-state heat conduction in two dimensions.

Consider a square thin plate of side length ℓ . We are interested in finding the steady-state temperature profile $T(x, y)$ in response to an arbitrary heat generation rate $f(x, y)$ within the body. The four sides of the plate are maintained at steady temperature T_0 .

- (a) Write down the steady-state energy balance considering heat conduction and heat generation in the solid. Note that this will be a partial differential equation. State the boundary conditions for this PDE.
- (b) Nondimensionalize your problem and choose variables so that the boundary conditions are homogeneous. Denote the nondimensional T, x, y variables as Θ, ξ, η , respectively.

State the nondimensional PDE and BCs for this problem. How many nondimensional parameters appear in the problem?

- (c) Take the Laplace transform in the ξ variable. Show that the ordinary differential equation that arises for the transform is

$$\begin{aligned} \bar{T}_{\eta\eta} + s^2\bar{T} &= \bar{g}(s, \eta) \\ \bar{T}(s, 0) = 0 \quad \bar{T}(s, 1) &= 0 \end{aligned} \tag{3.109}$$

What is $\bar{g}(s, \eta)$ for this problem?

- (d) Solve the ODE for the transform, and show that the solution is

$$\bar{T}(s, \eta) = \frac{1}{s} \int_0^1 \bar{G}(s, \eta, \eta') \bar{g}(s, \eta') d\eta'$$

with

$$\bar{G}(s, \eta, \eta') = \begin{cases} -\frac{\sin(s\eta') \sin s(1-\eta)}{\sin s}, & \eta' < \eta \\ -\frac{\sin(s\eta) \sin s(1-\eta')}{\sin s}, & \eta < \eta' \end{cases}$$

Note that you may find it helpful to review Example 2.15 where it is shown that (in our variable names)

$$\sinh s(\eta - \eta') - \frac{\sinh(s\eta) \sinh s(1 - \eta')}{\sinh s} = -\frac{\sinh(s\eta') \sinh s(1 - \eta)}{\sinh s}$$

Without rederiving anything, how do you know that the result above holds also for sin replacing sinh?

- (e) Using the inverse transform of Exercise 3.63, show that the inverse of the Green's function transform is

$$G(\xi, \eta, \eta') = 2 \sum_{n=1}^{\infty} \sin(n\pi\eta') \sin(n\pi\eta) \cosh(n\pi\xi)$$

- (f) Invert $\bar{T}(s, \eta)$ and show that

$$\begin{aligned} \Theta(\xi, \eta) &= \int_0^1 \int_0^\xi G(\xi', \eta, \eta') \Theta_\xi(0, \eta') d\xi' d\eta' + \\ &\quad \int_0^1 \int_0^\xi \int_0^\xi G(\xi' - \xi'', \eta, \eta') F(\xi'', \eta') d\xi'' d\xi' d\eta' \end{aligned}$$

Substitute the above expression for G , switch the order of ξ' and ξ'' integrals in the second term, and then perform the ξ' integrals in both terms to obtain

$$\Theta(\xi, \eta) = 2 \sum_{n=1}^{\infty} \frac{1}{n\pi} \left[a_n \sinh(n\pi\xi) + \int_0^1 \int_0^{\xi} \sin(n\pi\eta') \sinh(n\pi(\xi - \xi')) F(\xi', \eta') d\xi' d\eta' \right] \sin(n\pi\eta) \quad (3.110)$$

with the Fourier coefficients of the flux defined as

$$a_n = \int_0^1 \Theta_{\xi}(0, \eta') \sin(n\pi\eta') d\eta'$$

(g) Evaluate at $\xi = 1$ and use the boundary condition to show that

$$0 = \sum_{n=1}^{\infty} \frac{1}{n\pi} \left[\sinh(n\pi) a_n + \int_0^1 \int_0^1 \sin(n\pi\eta') \sinh(n\pi(1 - \xi')) F(\xi', \eta') d\eta' d\xi' \right] \sin(n\pi\eta)$$

Since the right-hand side of this equation is the Fourier series of the zero function for $\eta \in [0, 1]$, its coefficients must all be zero giving

$$a_n = \frac{-1}{\sinh(n\pi)} \int_0^1 \int_0^1 \sin(n\pi\eta') \sinh(n\pi(1 - \xi')) F(\xi', \eta') d\eta' d\xi'$$

(h) Substitute this result for a_n into (3.110) and show that the solution can be expressed as

$$\Theta(\xi, \eta) = 2 \sum_{n=1}^{\infty} \frac{1}{n\pi} \int_0^1 \int_0^1 G_n^{\xi}(\xi, \xi') G_n^{\eta}(\eta, \eta') F(\xi', \eta') d\eta' d\xi'$$

with

$$G_n^{\eta}(\eta, \eta') = \sin(n\pi\eta) \sin(n\pi\eta')$$

$$G_n^{\xi}(\xi, \xi') = \begin{cases} -\frac{\sinh(n\pi\xi') \sinh(n\pi(1-\xi))}{\sinh n\pi}, & \xi' < \xi \\ -\frac{\sinh(n\pi\xi) \sinh(n\pi(1-\xi'))}{\sinh n\pi}, & \xi < \xi' \end{cases}$$

Notice that both G_n^{η} and G_n^{ξ} are symmetric functions.

Exercise 3.63: Laplace transform inverse.

Invert the following transform

$$\bar{f}(s) = \frac{\sin(as) \sin(bs)}{\sin s}$$

Hint: use the Heaviside expansion theorem. Note that the Heaviside expansion theorem does not strictly apply to this $\bar{f}(s)$. Why not? See Exercise 3.64 for a way around this technical difficulty.

Exercise 3.64: The fine print.

We are stretching the Laplace transform pretty hard to cover Exercise 3.62, so let's check that the approach is providing a valid solution.

- (a) First of all, an infinite sum of increasing exponentials does not appear to converge, so let's check that the $G_n(\xi, \xi')$ is well defined in the limit $n \rightarrow \infty$.

$$G_n^\xi(\xi, \xi') = \begin{cases} -\frac{\sinh(n\pi\xi') \sinh(n\pi(1-\xi))}{\sinh n\pi}, & \xi' < \xi \\ -\frac{\sinh(n\pi\xi) \sinh(n\pi(1-\xi'))}{\sinh n\pi}, & \xi < \xi' \end{cases}$$

What is $\lim_{n \rightarrow \infty} G_n^\xi(\xi, \xi')$?

- (b) Since $G_n^\xi(\xi, \xi')$ is bounded as $n \rightarrow \infty$, we can probably justify the use of the Laplace transform. The first serious problem is that we blithely applied the Heaviside expansion theorem to

$$\bar{f}(s) = \frac{\sin(as) \sin(bs)}{\sin s}$$

even though $1/\sin(s)$ has poles at $s = \pm n\pi, n = 0, 1, 2, \dots$, and we cannot choose a positive constant c such that all the poles of $1/\sin(s)$ are to the left of the line $\text{Re}(s) = c$. And we see the problem when we try to invert anyway and obtain

$$f(t) = 2 \sum_{n=1}^{\infty} (-1)^n \sin(n\pi a) \sin(n\pi b) \cosh(n\pi t)$$

which is not (absolutely) convergent for any t , even *finite* t , because of the increasing orders of the cosh terms!

So we have to fix this problem. First consider the Mittag-Leffler expansion of $1/\sin(s)$

$$\frac{1}{\sin s} = \sum_{n=-\infty}^{\infty} \frac{(-1)^n}{s - n\pi}$$

and truncate this series for $|n| > M$ and define

$$\frac{1}{\sin_M(s)} = \sum_{n=-M}^M \frac{(-1)^n}{s - n\pi}$$

and we have that $\lim_{M \rightarrow \infty} 1/\sin_M(s) = 1/\sin(s)$ at all $s \neq n\pi, n = 0, \pm 1, \pm 2, \dots$. Next replace $1/\sin(s)$ with $1/\sin_M(s)$, resolve Exercise 3.63, and show that

$$\bar{f}_M(s) = \frac{\sin(as) \sin(bs)}{\sin_M(s)} \quad f_M(t) = 2 \sum_{n=1}^M (-1)^n \sin(n\pi a) \sin(n\pi b) \cosh(n\pi t)$$

which is now well defined for all $t \in [0, \infty)$ and finite integer $M > 0$.

- (c) Notice what changes when you resolve the rest of Exercise 3.62 using $1/\sin_M(s)$ in place of $1/\sin(s)$. Show that

$$\Theta_M(\xi, \eta) = 2 \sum_{n=1}^M \frac{1}{n\pi} \int_0^1 \int_0^1 G_n^\eta(\eta, \eta') G_n^\xi(\xi, \xi') F(\xi', \eta') d\xi' d\eta'$$

Finally take the limit as $M \rightarrow \infty$ to solve the original problem and obtain the result stated in Exercise 3.62.

Exercise 3.65: Two forms of the steady-state heat conduction solution

Review the eigenfunction expansion approach to the steady-state heat conduction problem presented in Example 3.7.

- (a) Show that another form of the solution to Exercise 3.62 is

$$\Theta(\xi, \eta) = -\frac{4}{\pi^2} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{n^2 + m^2} \left[\int_0^1 \int_0^1 F(\xi', \eta') \sin(m\pi\xi') \sin(n\pi\eta') d\xi' d\eta' \right] \sin(m\pi\xi) \sin(n\pi\eta)$$

Notice that this solution has a pleasing symmetry in ξ and η that is missing from the previous solution. On the other hand, this solution requires a double summation, which is computationally more expensive than the previous solution.

Let's establish that these two solutions are equivalent.

- (b) Equate the two forms of the solution and show that they are the same if and only if

$$\frac{2}{n\pi} G_n^\xi(\xi, \xi') = -\frac{4}{\pi^2} \sum_{m=1}^{\infty} \frac{1}{n^2 + m^2} \sin(m\pi\xi') \sin(m\pi\xi)$$

which must hold for all $n \geq 1$ and $\xi, \xi' \in [0, 1]$.

- (c) Consider ξ' a constant parameter and ξ a variable. Show that for the right-hand side to be a Fourier series representation of the left-hand side, the coefficients must satisfy

$$-\frac{n}{\pi(n^2 + m^2)} \sin(m\pi\xi') = \int_0^1 G_n^\xi(\xi, \xi') \sin(m\pi\xi) d\xi \quad (3.111)$$

which must hold for all $n, m \geq 1$ and all $\xi' \in [0, 1]$.

(d) Given $G_n(\xi, \xi')$ from the previous exercise

$$G_n^\xi(\xi, \xi') = \begin{cases} -\frac{\sinh(n\pi\xi') \sinh(n\pi(1-\xi))}{\sinh n\pi}, & \xi' < \xi \\ -\frac{\sinh(n\pi\xi) \sinh(n\pi(1-\xi'))}{\sinh n\pi}, & \xi < \xi' \end{cases}$$

perform the indicated integral and verify that the coefficients satisfy 3.111.

Hints: Before jumping in, it pays to have a few integrals available. Verify the following pair of integrals

$$\int_0^a \sinh(n\pi x) \sin(m\pi x) dx = \frac{1}{\pi(n^2 + m^2)} [-m \cos(m\pi a) \sinh(n\pi a) + n \sin(m\pi a) \cosh(n\pi a)]$$

$$\int_a^1 \sinh(n\pi(1-x)) \sin(m\pi x) dx = \frac{(-1)^m}{\pi(n^2 + m^2)} [m \cos(m\pi(1-a)) \sinh(n\pi(1-a)) - n \sin(m\pi(1-a)) \cosh(n\pi(1-a))]$$

The first can be derived by noting that

$$\sinh(n\pi x) \sin(m\pi x) = -\text{Im}(\cos((n + im)\pi x))$$

and the second by noting that

$$\sinh(n\pi x) \sin(m\pi(1-x)) = -\text{Im}((-1)^n \cos((-n + im)\pi(1-x)))$$

(e) Choose a few values of n and ξ' and plot $G_n^\xi(\xi, \xi')$ versus ξ and its Fourier series. How many terms in the series are required for an accurate reconstruction of G_n^ξ ? Which of the two solutions do you prefer and why?

Exercise 3.66: Transient heat conduction

In Exercise 2.81, you showed that the following steady-state heat conduction problem with forcing term (heat addition rate $-f(x)$) has a unique solution provided $k \neq n\pi, n = 1, 2, \dots$

$$\frac{d}{dx^2} u(x) + k^2 u(x) = f(x) \quad u(0) = u(1) = 0$$

One of your colleagues is wondering how we can show analytically whether the steady-state solution is stable?

So consider the transient problem with no external heating, $f(x) = 0$, for which we know that a steady state exists for *all* k ,

$$\frac{\partial}{\partial t} u(x, t) = \frac{\partial^2}{\partial x^2} u(x, t) + k^2 u(x, t)$$

The initial and boundary conditions are

$$u(x, t) = \begin{cases} u_0(x), & t = 0, x \in (0, 1) \\ 0, & t > 0, x = 0 \\ 0, & t > 0, x = 1 \end{cases}$$

with arbitrary initial condition function $u_0(x)$.

- (a) Take the Laplace transform of the PDE and BCs and show that the transform satisfies

$$\begin{aligned} \frac{d^2}{dx^2} \bar{u}(x, s) + (k^2 - s) \bar{u}(x, s) &= -u_0(x) \\ \bar{u}(0, s) = \bar{u}(1, s) &= 0 \end{aligned} \quad (3.112)$$

- (b) Given the arbitrary $u_0(x)$ initial condition, let's solve this differential equation using Fourier series. Given the form of the boundary conditions, we choose the complete, orthogonal set $\{\sin n\pi x\}$, $n = 1, 2, \dots$ since these satisfy the boundary conditions at $x = 0, 1$. Denote the Fourier coefficients of the solution as $a_n(s)$ and the initial condition as t_n , so that

$$\bar{u}(x, s) = \sum_{n=1}^{\infty} a_n(s) \sin n\pi x \quad u_0(x) = \sum_{n=1}^{\infty} t_n \sin n\pi x$$

Provide a formula for calculating the t_n , $n = 1, 2, \dots$

- (c) Double check that $\bar{u}(x, s)$ satisfies the boundary conditions for all choices of $a_n(s)$. Substitute these expansions for $\bar{u}(x, s)$ and $u_0(x)$ into (3.112) and solve for $a_n(s)$.
- (d) With $\bar{u}(x, s)$ determined, where are its singularities? Invert the transform to obtain $u(x, t)$.
- (e) For what k values is the steady state stable? What is the steady-state solution?

Exercise 3.67: Transient diffusion with source term

Consider again the transient diffusion problem in an unbounded domain with an arbitrary forcing term

$$u_t = D u_{xx} + f(x, t)$$

with $x \in (-\infty, \infty)$ and initial condition $u(x, 0) = u_0(x)$.

(a) Take the Fourier *and* Laplace transform of this problem and show

$$\tilde{u}(k, s) = \frac{\hat{u}_0(k)}{s + Dk^2} + \frac{\tilde{f}(k, s)}{s + Dk^2}$$

where the double transform is defined as

$$\tilde{u}(k, s) = \int_0^\infty \int_{-\infty}^\infty e^{-st} e^{-ikx} u(x, t) dx dt$$

(b) Invert the Laplace transform to obtain

$$\hat{u}(k, t) = \hat{u}_0(k) e^{-Dk^2 t} + \int_0^t \hat{f}(k, \tau) e^{-Dk^2(t-\tau)} d\tau$$

(c) Invert the Fourier transform to obtain the solution

$$u(x, t) = \int_{-\infty}^\infty u_0(\xi) g(x - \xi, t) d\xi + \int_0^t \int_{-\infty}^\infty f(\xi, \tau) g(x - \xi, t - \tau) d\xi d\tau \quad (3.113)$$

in which

$$g(x, t) = \frac{1}{2\sqrt{\pi Dt}} e^{-x^2/(4Dt)}$$

(d) Check that the solution satisfies the initial condition and PDE.

(e) Try inverting in the reverse order. First invert the Fourier transform and show that

$$\bar{u}(x, s) = \int_{-\infty}^\infty \left(u_0(\xi) + \bar{f}(\xi, s) \right) \frac{e^{-\sqrt{s} \frac{|x-\xi|}{\sqrt{D}}}}{2\sqrt{Ds}} d\xi$$

(f) Now invert the Laplace transform. Do you obtain (3.113)? Which order of inversion do you prefer and why?

(g) Now extend the problem to three dimensions as in Example 3.13 and solve

$$u_t = D(u_{xx} + u_{yy} + u_{zz}) + f(x, y, z, t)$$

with initial condition $u(x, y, z, t) = u_0(x, y, z)$ for $t = 0$. Compare the solution to (3.113).

Exercise 3.68: Reaction and diffusion in a membrane

Revisit Example 3.15 and show that the solution given in the text

$$c(z, \tau) = \frac{\sinh \sqrt{k}(1-z)}{\sinh \sqrt{k}} - 2 \sum_{n=1}^\infty \frac{(-1)^{n+1} n\pi}{n^2 \pi^2 + k} \sin(n\pi(1-z)) e^{-(n^2 \pi^2 + k)\tau}$$

satisfies the PDE

$$\frac{\partial c}{\partial \tau} = \frac{\partial^2 c}{\partial z^2} - kc$$

and boundary and initial conditions.

$$c(z, \tau) = 1 \quad z = 0, \quad \tau > 0$$

$$c(z, \tau) = 0 \quad z = 1, \quad \tau > 0$$

$$c(z, \tau) = 0 \quad 0 < z < 1, \quad \tau = 0$$

4

Probability, Random Variables, and Estimation

Exercise 4.60: The marginal intervals for the unknown measurement variance case and the t -statistic

Consider again the maximum likelihood estimation problem presented in Section 4.7.2 for the linear model with scalar measurement y , and unknown measurement variance σ^2 .

- (a) Show that the marginal box for this case is given by

$$\hat{\theta} = \theta_0 \pm m$$
$$m_i = \left(F_F^{-1}(\alpha; 1, n - n_p) s^2 (X^T X)_{ii}^{-1} \right)^{1/2}$$

- (b) Compare your formula for m_i above to c_i given in the text for the bounding box interval. Which one is larger and why?
- (c) Next use the approach in Exercise 4.34 to show that the marginal box can equivalently be expressed with a t -statistic

$$m_i = F_t^{-1} \left(\frac{1 + \alpha}{2}; n - n_p \right) (s^2 (X^T X)_{ii}^{-1})^{1/2}$$

in which F_t^{-1} is the inverse of the cumulative t -distribution, i.e., $\int_{-\infty}^{F_t^{-1}(\alpha; n)} p_t(z; n) dz = \alpha$ for all $n \geq 1$ and $\alpha \in [0, 1]$. Therefore, comparing the two formulas for m_i , we have also established the following relationship between the t -statistic and the F -statistic

$$F_t^{-1} \left(\frac{1 + \alpha}{2}; n \right) = \sqrt{F_F^{-1}(\alpha; 1, n)} \quad n \geq 1, \quad \alpha \in [0, 1]$$

Exercise 4.61: PLS regression versus minimum norm regression

The text lists the PLS regression formula as

$$B_{\text{PLS}} = R T^T Y$$

and mentions that this is a (nonunique) least-squares solution to the regression problem after replacing X with the low rank (rank ℓ) approximation $X_\ell = T P^T$

$$Y = X_\ell B_{\text{PLS}} \tag{4.97}$$

Consider the SVD of X_ℓ written as

$$X_\ell = \tilde{U}_\ell \tilde{\Sigma}_\ell \tilde{V}_\ell^T$$

- Work out the dimensions of matrices \tilde{U}_ℓ , $\tilde{\Sigma}_\ell$, and \tilde{V}_ℓ .
- What is the minimum-norm least-squares solution B_{MLS} of (4.97) in terms of the SVD matrices? Is this solution unique?
- For the data in Example 4.23, compare the estimates B_{PLS} and B_{MLS} for all $\ell = 1, 2, \dots, q$. Check that both are least-squares solutions. Check that indeed the norm of B_{PLS} is larger than B_{MLS} for all ℓ . Is it much larger?

Exercise 4.62: Expectation of squared norm of sample mean estimate error

Consider random variable $\xi \in \mathbb{R}^p$ with mean m and variance matrix P . Let \hat{x}_n be an estimator of the mean of ξ based on n samples of ξ , x_1, x_2, \dots, x_n . Say we have worked out the mean and variance of the particular estimator, and they are

$$\mathcal{E}(\hat{x}_n) = m' \quad \text{var}(\hat{x}_n) = P'$$

- What is the bias of this estimator?
- Calculate the expectation of the square of the 2-norm of the estimate error $\mathcal{E}(\|\hat{x}_n - m\|^2)$. State this result in terms of the bias and variance of the estimator. This metric of estimate error allows us, for example, to quantify the tradeoff between bias and variance in estimation.
Hint: start by expressing $\|\hat{x}_n - m\|^2 = (\hat{x}_n - m)^T(\hat{x}_n - m) = ((\hat{x}_n - m') - (m - m'))^T((\hat{x}_n - m') - (m - m'))$. Expand the quadratic and then take expectation. Also recall that $x^T x = \text{tr}(x x^T)$.
- For the sample-average estimator, what are m' and P' ? What is the expectation of the square of the norm of the estimate error using the sample-average estimator?
- Consider taking just one of the samples, x_i , as the estimator of the mean. What is the expectation of the square of the norm of the estimate error using the single-sample estimator?

Exercise 4.63: The limiting value of $x^T(x x^T)^{-1}x$

Consider a nonzero vector $x \in \mathbb{R}^p$. We would like to find the limiting value of $x^T(x x^T)^{-1}x$.

- Show that $(x x^T)^{-1}$ does not exist (for $p \geq 2$) by showing that $\text{rank}(x x^T) = 1$.

- (b) Therefore consider the full rank matrix $xx^T + \rho I$ and take the limit as $\rho \rightarrow 0$ to show that

$$\lim_{\rho \rightarrow 0} x^T (xx^T + \rho I)^{-1} x = 1$$

Hint: Consider the SVDs of x and xx^T .

- (c) Next consider a collection of $n \geq 1$ x_i vectors, not all zero, and show that

$$\text{rank}\left(\sum_{i=1}^n x_i x_i^T\right) = r$$

where r is the number of linearly independent x_i vectors. Then show that

$$\lim_{\rho \rightarrow 0} \sum_{j=1}^n x_j^T \left(\left(\sum_{i=1}^n x_i x_i^T \right) + \rho I \right)^{-1} x_j = r$$

Hint: Place the vectors in matrix $X = [x_1 \ x_2 \ \cdots \ x_n]$ and consider the SVDs of X and the product XX^T .

Exercise 4.64: Maximum likelihood estimate of mean and variance of a normal

Let random variable $x \in \mathbb{R}^p$ be distributed normally $x \sim N(m, P)$. Let $x_i, i = 1, 2, \dots, n$ denote $n \geq 1$ samples of x .

- Compute the maximum likelihood estimate of m and P , denoted \hat{m}, \hat{P} . Compare your result to the one stated in Theorem 4.22.
- How large does n need to be for the maximum likelihood problem to have a solution? What happens, for example, for a single sample, $n = 1$?
Hint: consider the result derived in Exercise 4.63.
- Assume that the mean is known to be zero, i.e., $x \sim N(0, P)$. How large does n need to be for the maximum likelihood problem for estimating variance (only) to have a solution?
- Take the expectations of the estimates \hat{m} and \hat{P} . Are these unbiased? Explain why or why not.

Exercise 4.65: A maximum likelihood estimation problem

Consider a model

$$y = X\theta + e$$

with $y \in \mathbb{R}^p$ a vector of measured responses that are linear in parameter vector $\theta \in \mathbb{R}^{n_p}$ with measurement error e modeled as a random variable distributed as $e \sim N(0, R)$ with known positive definite variance R .

- (a) Given this density of e , what is the density of measurement y as a function of model parameter (not RV) θ .
- (b) One of your colleagues recalls from a class he took as a graduate student at Wisconsin, that if you maximize the probability of the measurements, $p_y(y; \theta)$, over parameter θ , you are also solving a weighted least-squares problem. What is the equivalent weighted least-squares problem that you are solving here?
- (c) Solve this least-squares problem and find the maximum likelihood estimate of parameter θ , denoted $\hat{\theta}$. What have you assumed about matrix X when solving this problem.
- (d) Given your result above, find the residual of the model fit $r = y - X\hat{\theta}$, assuming that the measurements come from the linear model with some true model parameter value θ_0 , i.e., $y = X\theta_0 + e$.
- (e) Show that the following matrix B is a projection operator

$$B = I - X(X^T R^{-1} X)^{-1} X^T R^{-1}$$

- (f) Finally compute the expectation of the square of the norm of the residual for the maximum likelihood estimate of θ .

$$\mathcal{E}(r^T r)$$

Given the density of e , what is $\mathcal{E}(e^T e)$? Which is smaller, $\mathcal{E}(r^T r)$ or $\mathcal{E}(e^T e)$? What does your result reduce to if $R = \sigma^2 I_p$, which is the first estimation problem solved in the text in Section 4.7.1 with the number of measurements p replacing the number of samples n of the scalar measurement.

Exercise 4.66: Least squares with deficient model matrix

Complications arise when solving least-squares problems with rank deficient matrices. Consider the following problem

$$\min_{\theta} (y - X\theta)^T H (y - X\theta) \quad H > 0 \quad \text{any } X$$

Here we keep a positive definite Hessian, $H \in \mathbb{R}^{p \times p} > 0$, but allow any $X \in \mathbb{R}^{p \times n_p}$, and do not assume that the columns (or even rows) are linearly independent as in the standard least-squares problem.

- (a) Show that the set of all solutions to this least-squares problem is

$$\hat{\theta} = \left(V_1 \Sigma_r^{-1} (U_1^T H U_1)^{-1} U_1^T H \right) y + V_2 \alpha_2$$

with free parameter $\alpha_2 \in \mathbb{R}^{n_p-r}$, $r = \text{rank}(X)$, and U and V are defined by the partitioned SVD of X as follows

$$X = USV^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_r & \\ & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

- (b) What does this solution reduce to when X has linearly independent columns, $r = n_p$?
- (c) What is the minimum norm solution, $\hat{\theta}_{\text{mn}}$?

Exercise 4.67: Least squares with deficient variance matrix

Complications also arise in least-squares estimation for semidefinite variance of the measurement error. Consider the following problem

$$\min_{\theta} (\mathcal{y} - X\theta)^T R^{-1} (\mathcal{y} - X\theta) \quad R \geq 0 \quad \text{any } X$$

with semidefinite matrix, $R \in \mathbb{R}^{p \times p} \geq 0$, and arbitrary $X \in \mathbb{R}^{p \times n_p}$.

- (a) Show that the solution exists if and only if the data \mathcal{y} satisfy the condition

$$\tilde{\mathcal{y}}_2 \in R(\tilde{X}_2)$$

with $\tilde{\mathcal{y}}_2 = U_2^T \mathcal{y}$, $\tilde{X}_2 = U_2^T X$, and the U matrix is defined in the partitioned (symmetric) SVD of matrix R

$$R = USU^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_r & \\ & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}$$

and r is the rank of R .

How would you numerically check that the data satisfy this existence condition?

- (b) Given that \mathcal{y} satisfies the existence condition, show that the original problem can be converted into the following equality-constrained least-squares problem with a positive definite Hessian

$$\min_{\theta} (\tilde{\mathcal{y}}_1 - \tilde{X}_1 \theta)^T \Sigma_r^{-1} (\tilde{\mathcal{y}}_1 - \tilde{X}_1 \theta) \quad \text{subject to } \tilde{X}_2 \theta = \tilde{\mathcal{y}}_2$$

and $\tilde{\mathcal{y}}_1 = U_1^T \mathcal{y}$, $\tilde{X}_1 = U_1^T X$.

Exercise 4.68: Least squares with equality constraint

Consider the equality-constrained least-squares problem that appears in the previous exercise

$$\min_{\theta} (\mathbf{y} - X\theta)^T H (\mathbf{y} - X\theta) \quad \text{subject to } A\theta = \mathbf{b}$$

with $\mathbf{b} \in R(A)$ and $H > 0$.

- (a) Define the SVD of A as $A = USV$ and change the coordinate system so that $\theta = V\alpha$. Show that the constraint $A\theta = \mathbf{b}$ can be solved in the transformed coordinates $\alpha = (\alpha_1, \alpha_2)$ to obtain

$$\alpha_1 = \Sigma_r^{-1} U_1^T \mathbf{b} \quad \alpha_2 \text{ arbitrary}$$

with the solvability condition $\mathbf{b} \in R(U_1)$. Show that this condition is equivalent to the original solvability condition $\mathbf{b} \in R(A)$.

- (b) With α_1 determined by the constraint, show that the remaining unconstrained optimization problem is

$$\min_{\alpha_2} (\tilde{\mathbf{y}} - M_2 \alpha_2)^T H (\tilde{\mathbf{y}} - M_2 \alpha_2)$$

with $M_1 = XV_1$, $M_2 = XV_2$, and $\tilde{\mathbf{y}} = \mathbf{y} - M_1 \Sigma_r^{-1} U_1^T \mathbf{b}$. Since $H > 0$, this problem can be solved as in Exercise 4.66.

Exercise 4.69: The whole enchilada—Generalized maximum likelihood estimation for the linear model

With the preliminaries of the last three exercises, we are ready to solve the problem of interest. Consider again the linear model with normally distributed measurement error

$$\mathbf{y} = X\theta + e \quad e \sim N(0, R)$$

with $\mathbf{y}, e \in \mathbb{R}^p$, $\theta \in \mathbb{R}^{n_p}$, and $X \in \mathbb{R}^{p \times n_p}$. We are interested in the case where X may not have independent columns, and $R \geq 0$ may not be positive definite. Denote the SVDs of X and R by

$$X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T & V_2^T \end{bmatrix} = U_1 \Sigma_r V_1^T$$

$$R = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} S_s & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1^T & Z_2^T \end{bmatrix} = Z_1 S_s Z_1^T = \tilde{Z}_1 \tilde{Z}_1^T$$

where we define $\tilde{Z}_1 = Z_1 \sqrt{S_s}$. We denote the rank of X as r , satisfying $1 \leq r \leq \min(n_p, p)$ and the rank of R as s , satisfying $1 \leq s \leq p$. We handle $X = 0$ and/or $R = 0$ as special cases below.

- (a) To treat the singular normal, $R \geq 0$, we consider the nonsingular perturbed problem $R_\rho = R + \rho I_p$ with scalar $\rho > 0$ so $R_\rho > 0$, find the maximum likelihood estimate by solving the optimization problem

$$\min_{\theta} (\mathbf{y} - X\theta)^T R_\rho^{-1} (\mathbf{y} - X\theta)$$

and take the limit as $\rho \rightarrow 0^+$. Show that the solution exists for \mathbf{y} satisfying the solvability condition $Z_2^T \mathbf{y} \in R(Z_2^T X)$, and the set of all solutions is given by

$$\hat{\theta} = X^\dagger (I_p - \tilde{Z}_1 (U_2^T \tilde{Z}_1)^\dagger U_2^T) \mathbf{y} + V_2 \alpha_2 \quad Z_2^T \mathbf{y} \in R(Z_2^T X) \quad (4.98)$$

where $\alpha_2 \in \mathbb{R}^{n_p - r}$ is an arbitrary vector, and $R(A)$ denotes the range of matrix A . Specialize this general result to the following cases.

- (b) $R > 0$ and X has linearly independent columns. Show that the solution exists for all \mathbf{y} and is unique, and is equivalent to the solution derived in the chapter

$$\hat{\theta}_2 = (X^T R^{-1} X)^{-1} X^T R^{-1} \mathbf{y}$$

- (c) $R > 0$. Show that the solution exists for all \mathbf{y} . Show that the set of all solutions is equivalent to the form given in Exercise 4.67

$$\hat{\theta}_3 = X^\dagger U_1 (U_1^T R^{-1} U_1)^{-1} U_1^T R^{-1} \mathbf{y} + V_2 \alpha_2$$

- (d) The columns of X are independent. Show that the solution is unique for \mathbf{y} satisfying the solvability condition, and is given by

$$\hat{\theta}_4 = X^\dagger (I_p - \tilde{Z}_1 (U_2^T \tilde{Z}_1)^\dagger U_2^T) \mathbf{y} \quad Z_2^T \mathbf{y} \in R(Z_2^T X)$$

- (e) The rows of X are independent. Show that a solution exists for all \mathbf{y} , and the set of all solutions is given by

$$\hat{\theta}_5 = X^\dagger \mathbf{y} + V_2 \alpha_2$$

- (f) $R = 0$. The set of solutions and solvability condition are

$$\hat{\theta}_6 = X^\dagger \mathbf{y} + V_2 \alpha_2 \quad \mathbf{y} \in R(X)$$

- (g) $X = 0$. The set of solutions and solvability condition are

$$\hat{\theta} \in \mathbb{R}^{n_p} \quad Z_2^T \mathbf{y} = 0$$

- (h) $X = 0$ and $R = 0$. The set of solutions and solvability condition are

$$\hat{\theta} \in \mathbb{R}^{n_p} \quad \mathbf{y} = 0$$

Exercise 4.70: The whole enchilada—Lagrangian approach

Consider again the problem statement of Exercise 4.69.

- (a) Show that the necessary conditions of the Lagrangian formulation reduce the problem to solving the linear algebra problem

$$L \begin{bmatrix} \theta \\ \lambda \end{bmatrix} = \begin{bmatrix} X^T R^\dagger \\ Z_2^T \end{bmatrix} \mathbf{y} \quad L = \begin{bmatrix} X^T R^\dagger X & X^T Z_2 \\ Z_2^T X & 0 \end{bmatrix}$$

- (b) Let the SVD of $L \geq 0$ be given by

$$L = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} \Sigma_l & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W_1' \\ W_2' \end{bmatrix} = W_1 \Sigma_l W_1'$$

where the rank of L is denoted l . Show that the solution to the Lagrangian's linear algebra problem is

$$\begin{bmatrix} \hat{\theta} \\ \lambda^0 \end{bmatrix} = L^\dagger \begin{bmatrix} X^T R^\dagger \\ Z_2^T \end{bmatrix} \mathbf{y} + W_2 w_2 \quad \begin{bmatrix} X^T R^\dagger \\ Z_2^T \end{bmatrix} \mathbf{y} \in R(L) \quad (4.99)$$

where $w_2 \in \mathbb{R}^{n_p+p-l}$ is an arbitrary vector.

- (c) Show that this solution is equivalent to (4.98) by showing that

$$\begin{bmatrix} I_{n_p} & 0 \end{bmatrix} L^\dagger \begin{bmatrix} X^T R^\dagger \\ Z_2^T \end{bmatrix} = X^\dagger (I_p - \tilde{Z}_1 (U_2^T \tilde{Z}_1)^\dagger U_2^T)$$

and

$$\begin{bmatrix} X^T R^\dagger \\ Z_2^T \end{bmatrix} \mathbf{y} \in R(L) \text{ if and only if } Z_2^T \mathbf{y} \in R(Z_2^T X)$$

and

$$R(V_2) = R\left(\begin{bmatrix} I_{n_p} & 0 \end{bmatrix} W_2\right)$$

Which form of the solution, (4.98) or (4.99), do you prefer and why?

Exercise 4.71: Expectation and covariance under linear transformations

Consider random variable $x \in \mathbb{R}^n$ with density p_x and mean and covariance

$$\mathcal{E}(x) = m_x \quad \text{cov}(x) = P_x$$

Consider the random variable $y \in \mathbb{R}^p$ defined by the linear transformation

$$y = Cx$$

- (a) Show that the mean and covariance for \mathbf{y} are given by

$$\mathcal{E}(\mathbf{y}) = C\mathbf{m}_x \quad \text{cov}(\mathbf{y}) = CP_xC^T$$

Does this result hold for all C ? If yes, prove it; if no, provide a counterexample.

- (b) Apply this result to solve Exercise 4.9.

Exercise 4.72: Complete and incomplete gamma function

The (complete) gamma function is defined by the integral

$$\Gamma(n) = \int_0^{\infty} e^{-t} t^{n-1} dt$$

- (a) Perform the integral for $n = 1$ and show that $\Gamma(1) = 1$.
 (b) Perform integration by parts and show that $\Gamma(n + 1)$ satisfies the recursion

$$\Gamma(n + 1) = n\Gamma(n)$$

for all $n > 0$. Therefore for integer-valued $n \geq 1$, show that

$$\Gamma(n) = (n - 1)!$$

- (c) Plot $\Gamma(n)$ for real-valued $0 < n \leq 5$. Notice that the integral formula generalizes the factorial to noninteger values.
 (d) The incomplete gamma function is defined by the integral

$$\gamma(n, x) = \int_0^x e^{-t} t^{n-1} dt$$

so that $\lim_{x \rightarrow \infty} \gamma(n, x) = \Gamma(n)$ for all n . Show that for all $n > 0$

$$\gamma(n + 1, x) = n\gamma(n, x) - x^n e^{-x}$$

Exercise 4.73: The density of a noninvertible transformation¹

Given random variable $\xi \in \mathbb{R}^n$ with density p_ξ and $\eta \in \mathbb{R}^k$ defined by the transformation $\eta = f(\xi)$ we wish to find p_η in terms of p_ξ . In the text, the *distribution* for η (rather than the density) is derived and stated in (4.24)

$$F_\eta(\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} p_\xi(\mathbf{x}) d\mathbf{x}$$

¹JBR would like to thank Scott Shell and Jordan Finzel of UCSB for helpful discussion of this and the next two exercises.

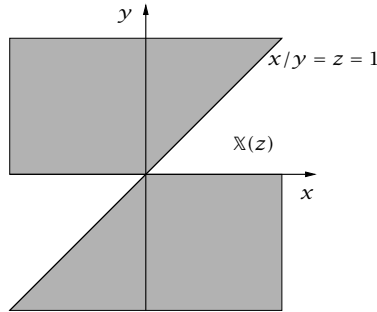


Figure 4.19: The region $\mathbb{X}(z) = \{(x, y) \mid x/y \leq z\}$.

where region $\mathbb{X}(y)$ is defined as $\mathbb{X}(y) = \{x \mid f(x) \leq y\}$.

For the special case $\eta \in \mathbb{R}^1$ and $\xi \in \mathbb{R}^2$, find the density $p_\eta(y)$ by differentiating $F_\eta(y)$ with respect to y , and show that

$$p_\eta(y) = \iint_{-\infty}^{\infty} \delta(y - f(x_1, x_2)) p_\xi(x_1, x_2) dx_1 dx_2 \quad (4.100)$$

Hint: use the indicator function for $(x_1, x_2) \in \mathbb{X}(y)$ in the integral for $F_\eta(y)$.

Exercise 4.74: Delta function formula for transformed densities

Derive the following fact of integration involving the delta function. In particular, explain why $|a|$ appears in the formula.

$$\int_{-\infty}^{\infty} g(x) \delta(ax - b) dx = \frac{1}{|a|} g(b/a), \quad a \neq 0 \quad (4.101)$$

Exercise 4.75: Density of the ratio of two normals

Let scalar random variables be independent and identically distributed as $X, Y \sim N(0, 1)$. Define $Z = X/Y$. We wish to derive Z 's density and show that

$$p_Z(z) = \frac{1}{\pi(z^2 + 1)}$$

- Sketch the region $\mathbb{X}(z) = \{(x, y) \mid x/y \leq z\}$, and use (4.24) to compute $F_Z(z)$ in terms of the density p_{XY} . Differentiate with respect to z to obtain p_Z .
- Use the formulas (4.100) and (4.101) and compute p_Z . Which approach do you prefer and why?

Exercise 4.76: Revisit the maximum of two random variables

Use the delta function approach given by (4.100) and (4.101) to resolve Example 4.6 where we showed that if $\eta = \max(\xi_1, \xi_2)$ then its density is given by

$$p_\eta(y) = p_{\xi_1}(y) \int_{-\infty}^y p_{\xi_2}(x) dx + p_{\xi_2}(y) \int_{-\infty}^y p_{\xi_1}(x) dx$$

Which approach do you prefer and why?

Exercise 4.77: Partitioned semidefinite matrices

(a) Show that a positive semidefinite matrix $P \in \mathbb{R}^{n \times n}$ can be expressed as

$$P = \begin{bmatrix} U^T U & U^T V \\ V^T U & V^T V \end{bmatrix}$$

for matrices $U \in \mathbb{R}^{n \times p}$ and $V \in \mathbb{R}^{n \times n-p}$ for any $0 \leq p \leq n$.

Hint: consider the partitioned eigenvalue decomposition of matrix P .

(b) Given this result, consider the partitioned positive semidefinite matrix

$$P = \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}$$

and show that if $P_y = 0$, then $P_{xy} = P_{yx}^T = 0$ as well.

Exercise 4.78: Expressing degenerate normal with pseudoinverse

In Exercise 4.23, we expressed the singular normal density of random variable $\xi \sim N(m, P)$ with $P \geq 0$ by

$$p_\xi(x) = \frac{1}{(2\pi)^{r/2} (\det \Lambda)^{1/2}} \exp \left[-\frac{1}{2} (x - m)^T Q_1 \Lambda^{-1} Q_1^T (x - m) \right] \delta(Q_2^T (x - m))$$

in which diagonal matrix $\Lambda \in \mathbb{R}^{r \times r}$ and orthonormal $Q \in \mathbb{R}^{n \times n}$ are obtained from the eigenvalue (singular value) decomposition of P

$$P = Q S Q^T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$$

with $\text{rank}(P) = r$.

(a) Show that we can express this density in pseudoinverse form as

$$p_\xi(x) = \frac{1}{(2\pi)^{r/2} (\det_+ P)^{1/2}} \exp \left[-\frac{1}{2} (x - m)^T P^\dagger (x - m) \right], \quad x - m \in R(P)$$

in which P^\dagger is the pseudoinverse of P , and $\det_+ P$ denotes the pseudodeterminant of P , defined as²

$$\det_+ P = \begin{cases} \lambda_1 \lambda_2 \cdots \lambda_r & r > 0 \\ 1 & r = 0 \end{cases}$$

Note that in this notation, the probability is defined on the *restricted* set $x - m \in R(P)$ where $R(P)$ is the range of the P matrix. By not defining the density on all of \mathbb{R}^n , we do not require the δ function in the density as in the previous notation.

(b) What does this density reduce to for $P = 0$?

Exercise 4.79: Conditional density for the degenerate normal

In Example 4.19 it was shown for ξ and η normally distributed as $(\xi, \eta) \sim N(m, P)$ with

$$m = \begin{bmatrix} m_x \\ m_y \end{bmatrix} \quad P = \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}$$

and $P > 0$, that the conditional density of ξ given η is also normal

$$p_{\xi|\eta}(x|y) = n(x, m, P)$$

in which the mean and covariance are

$$m = m_x + P_{xy}P_y^{-1}(y - m_y) \quad P = P_x - P_{xy}P_y^{-1}P_{yx}$$

Extend this result to the degenerate normal with $P \geq 0$ and show for this case

$$m = m_x + P_{xy}P_y^\dagger(y - m_y), \quad y - m_y \in R(P_y) \\ P = P_x - P_{xy}P_y^\dagger P_{yx}$$

where the pseudoinverse replaces the inverse, and we have a restriction on the values of y .

Hint: Try perturbing P_y to make it positive definite and then take the limit as the perturbation tends to zero. You may find the results in Exercises 4.77 and 4.78 useful.

Exercise 4.80: Establishing optimality in maximum likelihood problems

Two useful intermediate results for solving maximum likelihood problems are the following.

Lemma 4.1. *Given positive semidefinite $H \in \mathbb{R}^{p \times p}$, scalar $n > 0$, and optimization problem*

$$\min_{H \geq 0} f(H) = -n \ln \det H + \text{tr} H$$

the optimizer and optimal value are

$$H^0 = nI_p \quad f^0 = pn(1 - \ln n)$$

²Note that the definition implies that the pseudodeterminant of a zero matrix is unity.

Corollary 4.2 (Anderson (2003, Lemma 3.2.2)). *Given positive definite $R, G \in \mathbb{R}^{p \times p}$, scalar $n > 0$, and optimization problem*

$$\min_{R > 0} g(R) = n \ln \det R + \text{tr}(R^{-1}G)$$

the optimizer and optimal value are

$$R^0 = (1/n)G \quad g^0 = n \ln \det G + pn(1 - \ln n)$$

- Consider the scalar version ($p = 1$) of the lemma and show $H^0 = n$. How do you know that this is the unique, global minimizer?
- Then using the scalar result, establish that in the matrix case, all eigenvalues of H^0 are equal to n . Clearly $H = nI_p$ has this property. Establish that this is the *only* solution. Then evaluate the objective function with this H^0 and verify the result for f^0 .
- Using the result of the lemma, establish the corollary by considering the substitutions $G = UU$ and $H = UR^{-1}U$.

Exercise 4.81: Maximizing likelihood

Consider the maximum likelihood problems in Sections 4.7.3 and 4.7.4. The negative of the log-likelihood function is given in Section 4.7.3 as

$$-L(\Theta, R) = \frac{np}{2} \ln 2\pi + \frac{n}{2} \ln \det R + \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \Theta \mathbf{x}_i)^T R^{-1} (\mathbf{y}_i - \Theta \mathbf{x}_i)$$

Rather than setting derivatives to zero for *necessary* conditions for optimality, let's establish the optimality directly.

- Assume that the matrix $\sum_i \mathbf{x}_i \mathbf{x}_i^T \in \mathbb{R}^{q \times q}$ has full rank q . Maximize the likelihood (minimize $-L$) over Θ and show that the unique optimizer is

$$\hat{\Theta} = \left(\sum_i \mathbf{y}_i \mathbf{x}_i^T \right) \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1}$$

Notice that this estimate does not depend on the covariance R .

Hint: Stack all the samples as in Exercise 4.40 and express the likelihood with data matrices Y and X . Apply the vec operator and maximize the likelihood over vector $\text{vec} \Theta$ using the results of Exercise 1.107.

- Maximize the likelihood over the remaining variable R and show that the unique optimizer and optimal value are

$$\hat{R} = \frac{1}{n} \sum_i (\mathbf{y}_i - \hat{\Theta} \mathbf{x}_i) (\mathbf{y}_i - \hat{\Theta} \mathbf{x}_i)^T$$

$$-L^0 = \frac{np}{2} (1 + \ln 2\pi) + \frac{n(n-p)}{2} \ln n + \frac{n}{2} \ln \det \hat{R}$$

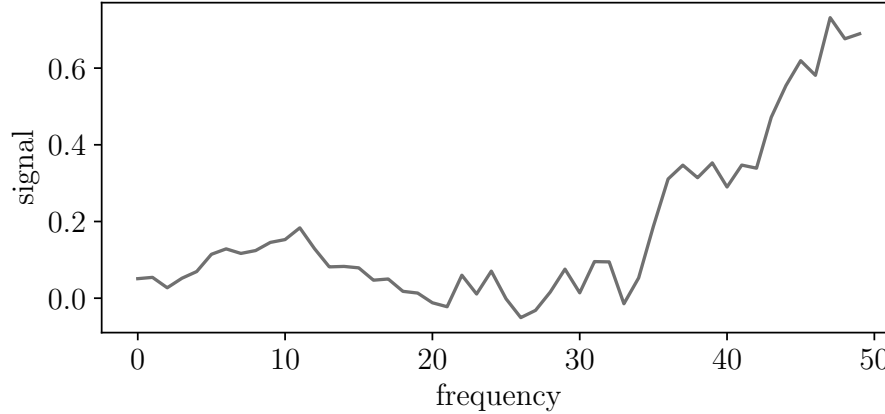


Figure 4.20: The exact spectrum for the $c = (1/4, 1/2, 1/4)$ mixture.

Hint: Corollary 4.2 is useful here.

Exercise 4.82: Spectroscopy, measurement uncertainty, and composition estimation

Given this chapter's new methods for describing measurement uncertainty, we revisit estimating mixture composition from spectroscopic measurements, Exercises 1.108 and 1.109.

- (a) Download the data for the ideal spectra of the three pure components in Figure 1.14, the three nonlinear interaction spectra in Figure 1.16, and the sample mixture matrix $S \in \mathbb{R}^{f \times 10}$ in Figure 1.17, where we have chosen 10 compositions for the standard.

Choose a mixture composition of $c = (1/4, 1/2, 1/4)$, and create the zero-noise nonideal spectrum for this mixture, using (1.52)–(1.53) with $\gamma_{AB} = \gamma_{AC} = \gamma_{BC} = 1$ as in Exercise 1.109

$$s_e = c_A s_A + c_B s_B + c_C s_C + \gamma_{AB} c_A c_B s_{AB} + \gamma_{AC} c_A c_C s_{AC} + \gamma_{BC} c_B c_C s_{BC}$$

We denote this spectrum as s_e . Make a plot of s_e versus frequency. Compare to Figure 4.20.

- (b) Now add normally distributed independent random noise with variance $\sigma^2 = 10^{-4}$ to the signal s_e . We denote that as a *sample*, s_{m1} , of the spectroscopic measurement of the chosen $(1/4, 1/2, 1/4)$ mixture. Plot the signals s_e and s_{m1} versus frequency and make sure that your sampling code is working properly, i.e., the two signals are close but not identical.

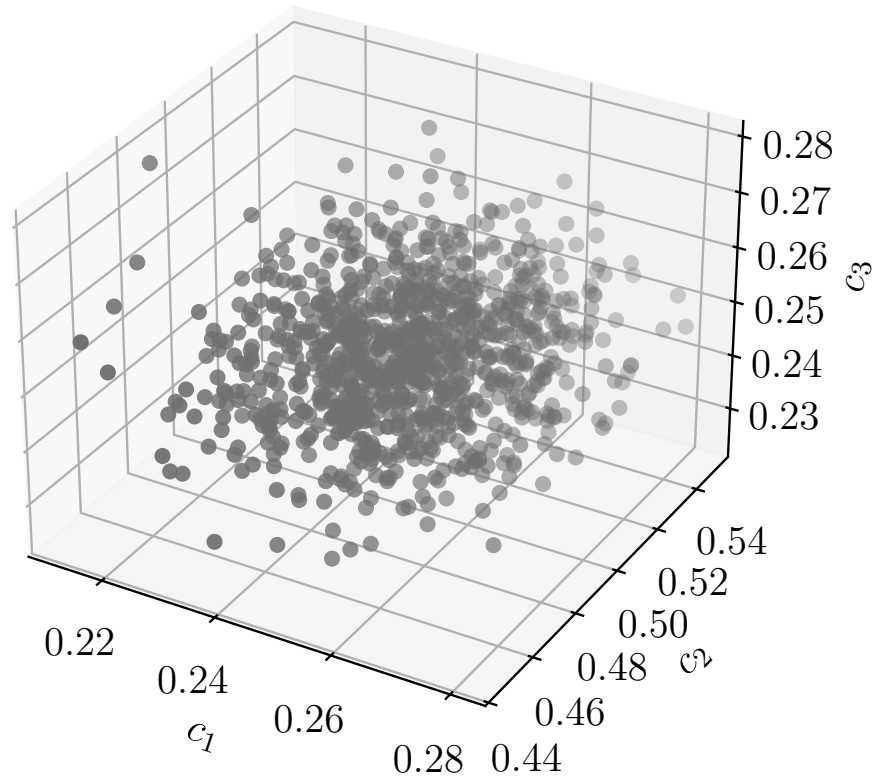


Figure 4.21: Scatter plot of the concentration estimates using the 6 largest singular values for S^\dagger . Note that this estimator has small bias and small variance.

Next generate $n_s = 1000$ samples of $s_{mi}, i = 1, 2, \dots, 1000$ and stack these as column vectors in a large matrix $S_m \in \mathbb{R}^{f \times n_s}$.

- (c) Now, as in part (c) of Exercise 1.109, estimate the composition for each of these n_s measurement samples using only the singular values in the S matrix that you deem reliable and above the noise level. Plot these composition estimates in a 3-d

scatter plot. Your results should look something like Figure 4.21.

Compute the sample mean and sample covariance of the composition estimates.³ Note that the sample covariance is a 3×3 matrix since the random variable is a concentration vector with 3 components. Is the sample mean close to $(1/4, 1/2, 1/4)$ using $n_s = 1000$ samples?

- (d) Next, as in part (d) of Exercise 1.109, estimate the composition for each of these n_s samples using *all* of the nonzero singular values in the S matrix. Plot these composition estimates in a 3-d scatter plot. Compute the sample mean and sample covariance of these composition estimates. Compare the sample mean and sample covariance of this estimator with those of the previous part.
- (e) Finally, compute the norm of the estimate error for the samples, $\|\hat{c}_i - c_0\|$ for $i = 1, 2, \dots, n_s$. On the same plots, show two histograms of the norms of the estimate error for the two approaches.

What do you conclude about these two composition estimators? Which one would you use and why?

³See commands `numpy.mean` and `numpy.cov` in Python.

5

Stochastic Models and Processes

Exercise 5.28: Optimizing a constrained, quadratic matrix function

We are familiar with the vector version of the problem

$$\min_x (1/2)x^T Qx \quad \text{s.t. } Ax = b$$

with $x \in \mathbb{R}^n$, and $Q > 0$, and $A \in \mathbb{R}^{p \times n}$ having linearly independent rows. The solution is

$$x^0 = Q^{-1}A^T(AQ^{-1}A^T)^{-1}b \quad (5.87)$$

which is readily derived using the method of Lagrange multipliers. The solution exists for all $b \in \mathbb{R}^p$.

We would like to extend this result to the matrix version of the problem

$$\min_X (1/2)\text{tr}(X^T QX) \quad \text{s.t. } AX = B$$

with $X \in \mathbb{R}^{n \times n}$, $Q > 0$, and $A \in \mathbb{R}^{p \times n}$ having linearly independent rows. See also Humpherys, Redd, and West (2012).

To solve this constrained problem we consider a matrix of Lagrange multipliers, $\Lambda \in \mathbb{R}^{p \times n}$, and express the equality constraints as the vector equation $\text{vec}(AX - B) = 0$. Next we augment the objective function in the usual way to form the Lagrangian

$$L(X, \Lambda) = (1/2)\text{tr}(X^T QX) - (\text{vec}\Lambda)^T \text{vec}(AX - B)$$

The constrained problem is then equivalent to the following unconstrained minmax problem

$$\min_X \max_\Lambda L(X, \Lambda)$$

(see Section 1.5.2 of Chapter 1 for a brief review of minmax games.) The necessary and sufficient conditions for a solution to the minmax problem are the matrix equations

$$\frac{dL}{dX} = 0 \quad \frac{dL}{d\Lambda} = 0$$

(a) Show that for any matrices A and B with identical dimensions

$$(\text{vec}A)^T (\text{vec}B) = \text{tr}(A^T B)$$

Use this result to convert the Lagrangian to

$$L(X, \Lambda) = (1/2)\text{tr}\left(X^T QX - 2\Lambda^T (AX - B)\right)$$

- (b) Compute the required derivatives (Table A.3 may be useful for this purpose) and show that the optimal solution satisfies the linear equations

$$\begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} X^0 \\ \Lambda^0 \end{bmatrix} = \begin{bmatrix} 0 \\ B \end{bmatrix}$$

- (c) Solve these equations to obtain

$$X^0 = Q^{-1}A^T(AQ^{-1}A^T)^{-1}B \quad (5.88)$$

(and $\Lambda^0 = (AQ^{-1}A^T)^{-1}B$). Compare (5.88) to (5.87).

Exercise 5.29: Optimal linear controller

Consider the linear discrete time system

$$x^+ = Ax + Bu$$

with infinite horizon control objective function

$$V(x, \mathbf{u}) = \sum_{k=0}^{\infty} L(x(k), u(k))$$

with input sequence defined as $\mathbf{u} = \{u(0), u(1), \dots\}$, and stage cost given as

$$L(x, u) = (1/2)(x^T Qx + u^T Ru)$$

From dynamic programming we have already found that the optimal controller is a linear feedback control $u^0(x) = Kx$, and the optimal cost is $V^0(x) = (1/2)x^T \Pi x$. The matrix Π was shown to satisfy the following discrete algebraic Riccati equation with the optimal gain K given by

$$\Pi = Q + A^T \Pi A - A^T \Pi B (B^T \Pi B + R)^{-1} B^T \Pi A \quad (5.89)$$

$$K = -(B^T \Pi B + R)^{-1} B^T \Pi A \quad (5.90)$$

Now we would like to find a shortcut method to derive these two results. Postulate that the optimal control is a linear control law $u = Kx$, and substitute that control law into the model, assume $A + BK$ is stable, compute $V(x)$ for this control law, and verify that $V(x) = (1/2)x^T \Pi x$ with Π satisfying

$$\Pi = Q_K + A_K^T Q_K A_K + (A_K^T)^2 Q_K A_K^2 + \dots$$

with $Q_K = Q + K^T R K$ and $A_K = A + BK$. Multiply both sides from the left by A_K^T and right by A_K and subtract from Π to show that Π satisfies

$$\Pi - (A + BK)^T \Pi (A + BK) = Q + K^T R K \quad (5.91)$$

which is a Lyapunov equation for Π given a fixed value of K .

So the optimization problem of interest is

$$\min_K (1/2)x^T \Pi(K)x \quad (5.92)$$

where $\Pi(K)$ denotes the solution to (5.91). For a controller gain K_0 to be an optimal gain, we require that for all $x \in \mathbb{R}^n$ and all K

$$x^T \Pi(K_0)x \leq x^T \Pi(K)x$$

or $\Pi(K_0) \leq \Pi(K)$.

- (a) Assume K_0 is the unique optimal gain and denote the solution to (5.91) as $\Pi_0 = \Pi(K_0)$. Let P_K be an arbitrary perturbation and consider $K = K_0 + P_K$. Substitute this K into (5.91), and denoting the solution as $\Pi(K) = \Pi(K_0) + P_\Pi$, derive an equation for the perturbation P_Π . It should be a Lyapunov equation.
- (b) Using the result of Exercise 2.63, find a condition for K_0 such that $P_\Pi \geq 0$ for all P_K . Show that this condition is (5.90). Thus K_0 is optimal if it satisfies (5.90). Since the optimal solution is assumed unique, this K_0 is the only optimal solution.
- (c) Substitute the optimal gain into (5.91) and simplify to show that the optimal Π is given by the solution to (5.89).

Dynamic programming goes further and establishes that *linear* feedback is indeed optimal over all feedback policies, including nonlinear policies, and that the optimal gain is unique.

Exercise 5.30: Rank ordering positive matrices and their inverses

We would like to treat the optimal linear estimator in an analogous fashion as the optimal linear controller, but we first require a useful relationship on ordering positive definite matrices and their inverses.¹ Let $A, B > 0, \in \mathbb{R}^{n \times n}$. Then the following holds

$$A > B \text{ if and only if } B^{-1} > A^{-1}$$

We establish this result in a few steps (Horn and Johnson, 1985, pp. 465, 471)

- (a) First we show that for $A, B > 0$, there exists a nonsingular $M \in \mathbb{R}^{n \times n}$ such that

$$A = MDM^T \quad B = MEM^T \quad D, E \text{ are diagonal matrices}$$

In other words, there exists a M matrix that can simultaneously diagonalize A and B .

¹The authors would like to thank Steven Kuntz of UCSB and Moritz Diehl of the University of Freiburg for helpful discussion of this and the next exercises.

To establish this result, let $P = A + B$. Note that P is also positive definite and can be expressed as $P = CC^T$ for a nonsingular C .²

Next note that $C^{-1}A(C^T)^{-1}$ is positive semidefinite for any nonsingular C , and let its singular value decomposition be denoted USU^T with U unitary and S diagonal and nonnegative. Then show that A and B are both diagonalized by choosing $M = CU$.

- (b) Show that for the special case of *diagonal* positive-definite matrices D and E , $D > E$ if and only if $E^{-1} > D^{-1}$. Using this fact, and the simultaneous diagonalization result above, establish that the same ordering result holds for all positive definite matrices.

Exercise 5.31: Optimal linear state estimator

Given the linear stochastic system

$$x^+ = Ax + w \quad y = Cx + v \quad w \sim N(0, Q) \quad v \sim N(0, R)$$

we showed that the optimal steady-state estimator is given by

$$\hat{x}^+ = A\hat{x} + L(y - C\hat{x}) \quad (5.93)$$

in which the steady-state filter variance and gain are given by

$$P = Q + APA^T - APC^T(CPC^T + R)^{-1}CPA^T \quad (5.94)$$

$$L = APC^T(CPC^T + R)^{-1} \quad (5.95)$$

Note that these are the optimal state estimate and variance *before* measurement, i.e., $\hat{x} = \hat{x}^-$, $P = P^-$.

As in the previous exercise with regulation, we would like to derive the optimal estimator formulas with a shortcut method. First consider a linear estimator of the form given in (5.93), but with an arbitrary L .

- (a) Show that for this estimator, the estimate error $\tilde{x} = x - \hat{x}$ satisfies

$$\tilde{x}^+ = (A - LC)\tilde{x} + w - Lv$$

Assuming that $A - LC$ is stable, show that, at steady state, the estimate error is zero mean with variance that satisfies

$$P = (A - LC)P(A - LC)^T + Q + LRL^T \quad (5.96)$$

Therefore, at steady state $\tilde{x} \sim N(0, P)$.

²Hint: you can use the singular value decomposition of P to establish this result.

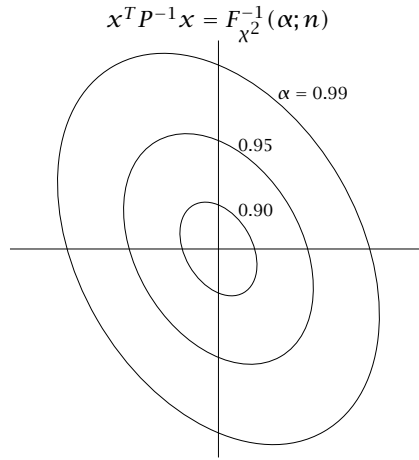


Figure 5.20: The α -level confidence intervals for a normal distribution. Maximizing the value of $x^T P^{-1} x$ over L gives the *smallest* confidence intervals for the estimate error.

- (b) As derived in Chapter 4, the α -level confidence interval for a normally distributed random variable is the ellipsoid

$$x^T P^{-1} x \leq F_{\chi^2}^{-1}(\alpha; n)$$

in which $F_{\chi^2}^{-1}(\alpha; n)$ is the inverse cumulative distribution function for the chi-squared distribution. We would like to derive the optimal L so that the confidence interval for estimate error is as small as possible. As shown in Figure 5.20, this means we would like to solve the equivalent optimization³

$$\max_L (1/2) x^T P^{-1} x$$

and we would like this L to be optimal for all $x \in \mathbb{R}^n$.

- (c) Use Exercise 5.30 to show this optimization problem is equivalent to

$$\min_L (1/2) x^T P x \tag{5.97}$$

Note the similarity to problem (5.92).

- (d) This strong similarity allows us to solve (5.97) by renaming variables and using the solution to Exercise 5.29. Consider the following transformation of the variables in Exercise 5.29

$$\Pi \rightarrow P \quad Q \rightarrow Q \quad R \rightarrow R \quad B \rightarrow C^T \quad A \rightarrow A^T \quad K \rightarrow -L^T$$

³We introduce the factor (1/2) for convenience as we will see shortly.

Show that under this variable transformation, the solution derived in Exercise 5.29 gives exactly the results (5.94)–(5.95).

Recall that we assumed in Exercise 5.29 that the optimal gain K_0 gave a stable $(A + BK_0)$ matrix and that this gain was unique. So we have made the same assumptions here about the optimal estimator gain L_0 , and the stability of $(A - L_0C)$. The derivation of the statistically optimal filter in Section 5.4, on the other hand, *established* that the linear estimator is optimal over all estimators, including nonlinear estimators, and that the optimal estimator gain is unique.

Exercise 5.32: Diffusion and random walk with n species

One of your experimental colleagues has measured the multicomponent diffusivity matrix D corresponding to the flux model

$$N_i = -D_{ij}\nabla c_j \quad i, j = 1, 2, \dots, n$$

i.e., a gradient in *any* concentration results in a nonzero flux of *all* species. The matrix D is positive semidefinite (symmetric, with nonnegative eigenvalues).

You are considering the following random process to model this multicomponent diffusion in a single spatial dimension

$$x(k+1) = x(k) + Gw(k)$$

with $x \in \mathbb{R}^n$, $k \in \mathbb{I}_{\geq 0}$, and matrix $G \in \mathbb{R}^{n \times n}$, and $w(k) \sim N(0, I_n)$ distributed as a vector of independent, zero-mean, unit-variance normals.

- Solve this difference equation for $x(k)$, $k = 0, 1, \dots$, as a function of the random term $w(k)$. What is the mean and variance of $x(k)$?
- If we let $t = k\Delta t$ and $GG^T = 2\Delta t D$, what is the mean and variance of the resulting continuous time process $x(t)$. Write down the probability density $p(x, t)$ for $x(t)$.
- If the simulation for $x(t)$ is considered to be a sample of some random process, provide the evolution equation of its probability density $p(x, t)$? In other words, what is the corresponding diffusion equation for this n -species system. You may assume that the diffusivity matrix D does not depend on x .
- How do you calculate a matrix G so that it corresponds to the measured diffusivity matrix D . Hint: think about the SVD (or, equivalently in this case, the eigenvalue decomposition) of D .
- We now want to make *sure* that we have the correct random process for the stated diffusion equation. To verify the diffusion equation, we need to take one t derivative and two x derivatives of $p(x, t)$. First establish the following helpful fact

$$D: \nabla \nabla \exp(a x^T D^{-1} x) = 2a \exp(a x^T D^{-1} x) (n + 2a x^T D^{-1} x)$$

or in component form

$$D_{ij} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} \exp(a x_k D_{kl}^{-1} x_l) = 2a \exp(a x_k D_{kl}^{-1} x_l) (n + 2a x_k D_{kl}^{-1} x_l)$$

for a an arbitrary constant scalar.

Even if this part proves elusive, you can still do the next part.

- (f) Using the result above, show that the $p(x, t)$ from your random walk satisfies your stated diffusion equation.

Exercise 5.33: Mean and variance of a controller cost function

When controlling the state x of a system to the origin subject to random disturbances, the best that a controller can do is usually obtain $x \sim N(0, P)$ where the variance P depends on the controller and the variance of the random disturbances. Given a quadratic cost function for the controller, $\ell = x^T Q x$, show that (Zagrobelny, Ji, and Rawlings, 2013)

$$\mathcal{E}(\ell) = \text{tr}(QP) \quad \text{var}(\ell) = 2\text{tr}(QPQP)$$

Note that ℓ is distributed as a generalized chi-squared distribution. Hint: the result in Exercise 4.32 may prove useful.

Bibliography

- T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, third edition, 2003.
- G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- J. Humpherys, P. Redd, and J. West. A fresh look at the Kalman filter. *SIAM Rev.*, 54(4):801–823, 2012.
- J. B. Rawlings, D. Q. Mayne, and M. M. Diehl. *Model Predictive Control: Theory, Design, and Computation*. Nob Hill Publishing, Santa Barbara, CA, 2nd, paperback edition, 2020. 770 pages, ISBN 978-0-9759377-5-4.
- M. A. Zagrobelny, L. Ji, and J. B. Rawlings. Quis custodiet ipsos custodes? *Annual Rev. Control*, 37(2):260–270, 2013.