A Mathematical Appendix

A.1 ℓ_p Norm on $\mathbb{C}^n$

Let $x \in \mathbb{C}^n$ be represented by its components $x = (x_1, x_2, \ldots, x_n)$ in the standard basis. We can also think of $(x_1, x_2, \ldots, x_n)$ as a sequence of n numbers in $\mathbb{C}$. The ℓ_p-norm of x is defined as

$$\|x\|_p := \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \max_i\{|x_i|\}, & \text{if } p = \infty \end{cases} \tag{A.1}$$

In this appendix, we will prove that this definition indeed satisfies all three properties of a norm outlined in Definition 1.10. The homogeneity and non-negativity conditions are easily checked and left as an exercise. To verify the triangle inequality, we need some preliminary work.

Lemma A.1 (Young's Inequality)**.** *Let p, q be real numbers satisfying $p, q > 1$ and*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then, for all real numbers $a, b \geq 0$,

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q.$$

Proof. If either a or b is 0, then $ab = 0$, while $\frac{1}{p}a^p + \frac{1}{q}b^q \geq 0$, proving the claim. If $a, b > 0$, then

$$\log(ab) = \frac{1}{p}\log a^p + \frac{1}{q}\log b^q \leq \log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right),$$

where we have used the fact the log is a concave function in the last inequality. Noting that log is also a strictly increasing function proves the claim. □

Lemma A.2 (Hölder's Inequality)**.** *Let $1 \leq p \leq \infty$ and $1 \leq q \leq \infty$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then, for any $x, y \in \mathbb{R}^n$, with components $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$,*

$$\sum_{i=1}^n |x_i|\,|y_i| \leq \|x\|_p \|y\|_q.$$

Proof. If either $x = 0$ or $y = 0$, then both sides of the inequality are 0, proving the claim. If $x, y \neq 0$ and $p, q > 1$, then

$$\frac{1}{\|x\|_p\|y\|_q}\sum_{i=1}^n |x_i|\,|y_i| = \sum_{i=1}^n \frac{|x_i|}{\|x\|_p}\frac{|y_i|}{\|y\|_q} \leq \sum_{i=1}^n \left[\frac{1}{p}\left(\frac{|x_i|}{\|x\|_p}\right)^p + \frac{1}{q}\left(\frac{|y_i|}{\|y\|_q}\right)^q\right] = 1$$

where we have used Young's inequality (Lemma A.1) and the fact that $\frac{1}{p} + \frac{1}{q} = 1$ after summing the individual terms. If $p = 1, q = \infty$, then

$$\sum_{i=1}^{n} |x_i|\,|y_i| \le \sum_{i=1}^{n} |x_i|\,\|y\|_\infty \le \|x\|_1 \|y\|_\infty.$$

A similar argument proves the claim for $p = \infty, q = 1$. $\square$

Theorem A.1. *For $1 \le p \le \infty$, $\|\cdot\|_p$ is a norm on $\mathbb{C}^n$.*

Proof. As previously described, the only remaining property to check is the triangle inequality. If $p = 1$, then for all $x, y \in \mathbb{C}^n$,

$$\|x+y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \le \sum_{i=1}^{n} (|x_i| + |y_i|) = \|x\|_1 + \|y\|_1.$$

If $p = \infty$, then for all $x, y \in \mathbb{C}^n$,

$$\|x+y\|_\infty = |x_j + y_j| \le \max_i\{|x_i|\} + \max_i\{|y_i|\} = \|x\|_\infty + \|y\|_\infty$$

where $j = \arg\max_i\{|x_i + y_i|\}$. Finally, if $1 < p < \infty$, then for all $x, y \in \mathbb{C}^n$, applying Hölder's inequality (Lemma A.2) yields

$$\begin{aligned}
\|x+y\|_p^p = \sum_{i=1}^{n} |x_i + y_i|^p &\le \sum_{i=1}^{n} (|x_i| + |y_i|)|x_i + y_i|^{p-1} \\
&= \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1} \\
&\le (\|x\|_p + \|y\|_p) \left(\sum_{i}^{n} |x_i + y_i|^{(p-1)q} \right)^{1/q},
\end{aligned}$$

where q satisfies the constraint $1/p + 1/q = 1$. From this constraint, we see that $1/q = (p-1)/p$ and $(p-1)q = p$. Hence,

$$\|x+y\|_p^p \le (\|x\|_p + \|y\|_p) \left(\sum_{i}^{n} |x_i + y_i|^p \right)^{(p-1)/p} = (\|x\|_p + \|y\|_p)\|x+y\|_p^{p-1}. \tag{A.2}$$

If $\|x+y\|_p = 0$, then the condition $\|x\|_p + \|y\|_p \ge 0$ implies the triangle inequality. If $\|x+y\|_p > 0$, then we can cancel $\|x+y\|_p^{p-1}$ from (A.2), thereby proving the theorem. $\square$

A.2 Determinant Formulae

Theorem A.2 (Cauchy-Binet)**.** *Let A be an $m \times n$ matrix and B an $n \times m$ matrix, where $m \le n$. Let S denote an m-element subset of $\{1, \ldots, n\}$; given S, let $A_{:S}$ denote the submatrix containing the m columns of A indexed by S, and $B_{S:}$ the submatrix containing* *the m rows of B indexed by S. Note that the columns of A_S and the rows of B_S have increasing indices. Then,*

$$\det(AB) = \sum_S \det(A_{:S}) \det(B_{S:}),$$

where the sum is over all $\binom{n}{m}$ possible choices of S.

Proof. By definition, we have

$$\begin{aligned}
\det(AB) &= \sum_{k_1=1,\cdots,k_m=1}^{m} \epsilon_{k_1\cdots k_m} (AB)_{1k_1} \cdots (AB)_{mk_m} \\
&= \sum_{k_1=1,\ldots,k_m=1}^{m} \sum_{j_1=1,\ldots,j_m=1}^{n} \epsilon_{k_1\cdots k_m} A_{1j_1} B_{j_1k_1} \cdots A_{mj_m} B_{j_mk_m}.
\end{aligned}$$

Exchanging the two multi-sums and rearranging terms, we get

$$\begin{aligned}
\det(AB) &= \sum_{j_1=1,\ldots,j_m=1}^{n} A_{1j_1} \cdots A_{mj_m} \sum_{k_1=1,\ldots,k_m=1}^{m} \epsilon_{k_1\cdots k_m} B_{j_1k_1} \cdots B_{j_mk_m} \\
&= \sum_{j_1=1,\ldots,j_m=1}^{n} A_{1j_1} \cdots A_{mj_m} \det \begin{pmatrix} B_{j_1\cdot} \\ \vdots \\ B_{j_m\cdot} \end{pmatrix} \\
&= \sum_{j_1=1,\ldots,j_m=1}^{n} A_{1j_1} \cdots A_{mj_m} \epsilon_{j_1\cdots j_m} \det(B_{\{j_1,\ldots,j_m\}\cdot}) \\
&= \sum_S \left[\sum_{j_1,\ldots,j_m \in S} A_{1j_1} \cdots A_{mj_m} \epsilon_{j_1\cdots j_m} \right] \det(B_{S:}) \\
&= \sum_S \det(A_{:S}) \det(B_{S:}).
\end{aligned}$$

□

REMARK A.1. *When $m > n$, AB is not full rank, and $\det(AB) \equiv 0$. We thus impose the condition $m \leq n$ in the theorem.*

Corollary A.1. *For any $m \times n$ real matrix A,*

$$\det(AA^t) \geq 0.$$

Proof. If $m > n$, then the rank of AA^t is strictly less than m, so $\det(AA^t) = 0$. If $m \leq n$, then by Theorem A.2, $\det(AA^t) = \sum_S \det(A_{:S}) \det(A^t_{S:}) = \sum_S \det(A_{:S})^2 \geq 0$. □

Theorem A.3. *Let M be a $(m+n)\times(m+n)$ block matrix of the form*

$$M = \begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ & \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{array}{c} m \\ n \end{array} \end{array}.$$

If A is invertible, then

$$\det(M) = \det(A)\det(M/A), \tag{A.3}$$

where $M/A \equiv D - CA^{-1}B$, called the Schur complement of A in M. If D is invertible, then

$$\det(M) = \det(D)\det(M/D), \tag{A.4}$$

where $M/D \equiv (A - BD^{-1}C)$, called the Schur complement of D in M.

Proof. If A is invertible, we can express M as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix}\begin{pmatrix} A & 0 \\ 0 & M/A \end{pmatrix}\begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix},$$

from which the theorem follows. Similarly, if D is invertible, then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} M/D & 0 \\ 0 & D \end{pmatrix}\begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}.$$

□

A.3 Condition Number of a Matrix

Many numerical calculations in real life require solving equations of the form

$$Ax = b$$

for $x \in \mathbb{R}^n$, where a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ and a vector $b \in \mathbb{R}^n$ are obtained from measurements. For example, x might represent the predictive power of each feature encoded in the columns of A, or it might be the hidden signal to be inferred in the system. Because most measurements in real life are subject to errors, we need to consider how perturbations in A and b might affect the desired solution x. If small changes in A and/or b result in large changes in x, then our confidence in what the obtained solution x implies in terms of real-world consequences would need to be moderated with caution.

As in Quantum Mechanics, let us choose a perturbation expansion parameter ϵ and consider solving the perturbed system

$$(A + \epsilon\,\delta A)\,x(\epsilon) = (b + \epsilon\,\delta b),$$

by expanding

$$x(\epsilon) = x_0 + \epsilon\, x_1 + \mathcal{O}(\epsilon^2),$$

where x_0 is the solution to the original unperturbed system, i.e. $Ax_0 = b$. Then, to linear order in ϵ, we must impose

$$(Ax_0 - b) + \epsilon(Ax_1 + \delta A\, x_0 - \delta b) + \mathcal{O}(\epsilon^2) = 0.$$

The zeroth term is automatically satisfied by the assumption that $Ax_0 = b$, while the first order term requires

$$Ax_1 = \delta b - \delta A\, x_0 \Rightarrow x_1 = A^{-1}(\delta b - \delta A\, x_0).$$

Thus, using any norm on $\mathbb{R}^n$ and the induced matrix norm, the relative change in $x(\epsilon)$ to leading order is

$$\frac{\|x(\epsilon) - x_0\|}{\|x_0\|} = |\epsilon|\, \frac{\|A^{-1}(\delta b - \delta A\, x_0)\|}{\|x_0\|} \leq |\epsilon|\, \|A^{-1}\| \frac{\|\delta b - \delta A, x_0\|}{\|x_0\|} \leq |\epsilon|\, \|A^{-1}\| \frac{\|\delta b\| + \|\delta A\| \|x_0\|}{\|x_0\|}.$$

We can rewrite the last fraction as

$$\frac{\|\delta b\| + \|\delta A\| \|x_0\|}{\|x_0\|} = \|A\| \left(\frac{\|\delta b\|}{\|A\| \|x_0\|} + \frac{\|\delta A\|}{\|A\|} \right).$$

Since $b = Ax_0$, we have $\|b\| \leq \|A\| \|x_0\|$ and, thus,

$$\frac{\|\delta b\|}{\|A\| \|x_0\|} \leq \frac{\|\delta b\|}{\|b\|}.$$

Combining these results, we finally get

$$\boxed{\frac{\|x(\epsilon) - x_0\|}{\|x_0\|} \leq \kappa(A)\, (\rho_b + \rho_A)}. \tag{A.5}$$

where

$$\kappa(A) \equiv \|A\| \|A^{-1}\|, \;\; \rho_b = \frac{|\epsilon| \|\delta b\|}{\|b\|}, \quad \text{and } \rho_A = \frac{|\epsilon| \|\delta A\|}{\|A\|}.$$

Definition A.1 (Condition Number of a Matrix)**.** *$\kappa(A) \equiv \|A\| \|A^{-1}\|$ is called the condition number of matrix A, with respect to the specified norm.*

REMARK A.2. *Note that ρ_b is the relative error in b, and ρ_A is the relative error in A. The relative error in x is thus bounded above by the sum of relative errors in A and b scaled by the condition number.*

REMARK A.3. *The precise value of the condition number depends on the chosen matrix norm. For the spectral norm, the condition number, denoted κ_2, is*

$$\kappa_2(A) \equiv \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

where $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the maximum and minimum singular values of A, respectively. This expression follows from the fact that

$$\sigma_{\min}(A) = \min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \min_{x \neq 0} \frac{\|x\|_2}{\|A^{-1}x\|_2} = \left(\max \frac{\|A^{-1}x\|_2}{\|x\|_2} \right)^{-1} = \frac{1}{\|A^{-1}\|_2}.$$

REMARK A.4. *If the condition number of a matrix is large, according to some notion of being large appropriate for the given problem, then the matrix is said to be ill-conditioned. A matrix with a small condition number is said to be well-conditioned. Note that orthogonal matrices have a condition number of 1 in the 2-norm and are thus well-conditioned.*

REMARK A.5. *The generalized condition number of any matrix A, not necessarily an invertible square matrix, is defined as $K(A) = \|A\|\|A^+\|$, where A^+ is the pseudo-inverse of A (see Theorem A.9). For an invertible square matrix A, $\kappa(A) = K(A)$. Similar to the condition number, $K(A)$ provides a bound on the relative error of a solution x to the problem $Ax = b$ when A and b are perturbed.*

REMARK A.6. *Note that (A.5) provides only a general upper bound, which might be an overestimate of the actual relative error for a specific perturbation. To get a better handle on the relative error, consider the singular value decomposition of A:*

$$A = \sum_{i=1}^{n} \sigma_i w_i v_i^t \;\Rightarrow\; A^{-1} = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i w_i^t.$$

Using the expansions $b = \sum_{i=1}^{n} b_i w_i$ and $\delta b = \sum_{i=1}^{n} \delta b_i w_i$ of the vectors b and δb in terms of the left singular vectors, we get

$$x = \sum_{i=1}^{n} \frac{b_i}{\sigma_i} v_i \quad \text{and} \quad \delta x = \sum_{i=1}^{n} \frac{\delta b_i}{\sigma_i} v_i$$

Hence, the relative error is

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\sum_{i=1}^{n} (\delta b_i)^2/\sigma_i^2}{\sum_{i=1}^{n} (b_i)^2/\sigma_i^2} = \frac{\sum_{i=1}^{n} (\delta b_i)^2 (\sigma_{\max}/\sigma_i)^2}{\sum_{i=1}^{n} (b_i)^2 (\sigma_{\max}/\sigma_i)^2},$$

which will be large for a large condition number $\kappa(A)$, if b_i are originally negligible along the smallest singular value directions, but non-negligible perturbations are introduced along these directions.

The condition number also appears in the bound of perturbation of A^{-1} when A itself is perturbed by infinitesimally small δA:

Theorem A.4.

$$\frac{\|(A + \delta A)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|} + \mathcal{O}(\|\delta A\|^2).$$

Sketch of proof. Rewriting $(A + \delta A)^{-1}$ as $(I + A^{-1}\delta A)^{-1} A^{-1}$, and using the series representation

$$(I + A^{-1}\delta A)^{-1} = \sum_{k=0}^{\infty}(-A^{-1}\delta A)^n,$$

we get

$$\|(A+\delta A)^{-1} - A^{-1}\| = \|A^{-1}\,\delta A\,A^{-1}\| + \mathcal{O}(\|\delta A\|^2) \leq \|A^{-1}\|\,\|\delta A\|\,\|A^{-1}\| + \mathcal{O}(\|\delta A\|^2).$$

□

A.4 Matrix Identities

Theorem A.5. *Let X be any $m \times n$ matrix. Let $I_{n\times n}$ and $I_{m\times m}$ denote the identity matrices of dimension n and m, respectively. Then, for all $\alpha > 0$, the following identify holds*

$$(X^tX + \alpha I_{n\times n})^{-1}X^t = X^t(XX^t + \alpha I_{m\times m})^{-1}.$$

Proof. Let $Z = (X^tX + \alpha I_{n\times n})^{-1}X^t - X^t(XX^t + \alpha I_{m\times m})^{-1}$. Then,

$$(X^tX + \alpha I_{n\times n})Z = X^t - (X^tX + \alpha I_{n\times n})X^t(XX^t + \alpha I_{m\times m})^{-1}.$$

But,

$$(X^tX + \alpha I_{n\times n})X^t = X^tXX^t + \alpha X^t = X^t(XX^t + \alpha I_{m\times m}).$$

Hence,

$$(X^tX + \alpha I_{n\times n})Z = 0.$$

Since $(X^tX + \alpha I_{n\times n})$ is invertible, we have $Z \equiv 0$. □

The same proof shows that

Theorem A.6. *Let X be any $m \times n$ matrix and Σ any $n \times n$ positive definite matrix. Let $I_{n\times n}$ and $I_{m\times m}$ denote the identity matrices of dimension n and m, respectively. Then, for all $\alpha > 0$, the following identify holds*

$$(X^tX\Sigma + \alpha I_{n\times n})^{-1}X^t = X^t(X\Sigma X^t + \alpha I_{m\times m})^{-1}.$$

REMARK A.7. *Note that $X^tX\Sigma + \alpha I_{n\times n}$ is invertible, because we can write it as*

$$\Sigma^{-1}(\Sigma X^tX\Sigma + \alpha\Sigma)$$

and $\Sigma X^tX\Sigma + \alpha\Sigma$ is positive definite.

Theorem A.7 (Woodbury, Sherman & Morrison Matrix Inversion Formula)**.** *Let Z, W, U, V be $n \times n$, $m \times m$, $n \times m$, $n \times m$ matrices, respectively. Then,*

$$(Z + UWV^t)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^tZ^{-1}U)^{-1}V^tZ^{-1}.$$

when the indicated inverses exist.

A.5 Orthogonal Matrices

Let $O(n)$ be the set of $n \times n$ matrices U such that the ℓ_2-norm of any vector $v \in \mathbb{R}^n$ is invariant under $v \mapsto Uv$. That is,

$$\forall v \in \mathbb{R}^n, \|Uv\|_2^2 = v^t U^t U v = v^t v \Rightarrow \forall v \in \mathbb{R}^n, v^t(U^t U - I_{n\times n})v = 0 \Rightarrow U^t U = I_{n\times n}.$$

Because $\det(U^t U) = \det(U^t)\det(U) = (det(U))^2 = \det(I) = 1$, we have

$$\det U = \pm 1.$$

Definition A.2 (Orthogonal Group). *The set $O(n)$ of $n \times n$ matrices U satisfying $U^t U = I$ is called the orthogonal group.*

REMARK A.8. *By the equality of left and right inverses, we also have $UU^t = I$.*

Definition A.3 (Special Orthogonal Group). *The special orthogonal group $SO(n)$ is the component of $O(n)$ connected to the identity, i.e. $SO(n) = \{U \in O(n) | \det U = 1\}$.*

REMARK A.9. *$O(n)$ is the set of all rotations and permutations in $\mathbb{R}^n$. This set is actually a Lie group, i.e. a smooth manifold with a group structure.*

EXERCISE A.1. *Show that the dimension of $O(n)$ and $SO(n)$ is $n(n-1)/2$.*

A.6 Courant-Fischer Theorem

Theorem A.8 (Courant-Fischer). *Let M be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \le \ldots \le \lambda_n$. Let $Gr(k,n)$ denote the set of all k-dimensional subspaces in $\mathbb{R}^n$. Then,*

$$\lambda_k = \min_{V \in Gr(k,n)} \left(\max_{v \in V\setminus\{0\}} \frac{v^t M v}{v^t v} \right) = \max_{V \in Gr(n-k+1,n)} \left(\min_{v \in V\setminus\{0\}} \frac{v^t M v}{v^t v} \right).$$

A.7 Moore-Penrose Pseudo-inverse

Theorem A.9 (Moore-Penrose). *Let A be a real $m \times n$ matrix. Then, there exists a unique $n \times m$ matrix A^+, called the pseudo-inverse, that satisfies*

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)^t = AA^+$
4. $(A^+A)^t = A^+A$

Proof. Condition 3 implies that $(A^{+^t} A^t)A = (AA^+)A = A$, where the last equality follows from Condition 1. Taking transpose of this equation yields, $(A^tA)A^+ = A^t$, which implies that $\ker(A^t) \supseteq \ker(A^+)$. Now, for any $v \in \ker(A^t)$, we have $(A^tA)A^+v = A^tv = 0$; i.e. $A^+v \in \ker(A^tA) = \ker(A)$. But, then, Condition 2 implies that $A^+v = A^+AA^+v = 0$, since $A^+v \in \ker(A)$; thus, $\ker(A^t) \subseteq \ker(A^+)$. Together, we have $\ker(A^t) = \ker(A^+)$. Similarly, Condition 4 can be used to show that $\ker(A) = \ker(A^{+t})$. Hence, if $\mathcal{R}$ is the range of A, then $A^+(\mathcal{R})$ is orthogonal to $\ker(A)$, and A^tA is thus invertible on $A^+(\mathcal{R}) = A^t(\mathcal{R})$. Thus, $A^+ = (A^tA)^{-1}A^t$ on $\mathcal{R}$. In terms of the SVD of A, we thus have

$$\boxed{A = \sum_{i=1}^{k} \sigma_i w_i v_i^t} \Longleftrightarrow \boxed{A^+ = \sum_{i=1}^{k} \sigma_i^{-1} v_i w_i^t}.$$

□

A.8 Matrix Tensor Products

Theorem A.10. *Let A and B be $I \times K$ and $J \times K$ matrices, respectively. The Khatri-Rao product $A \odot B$ satisfies the following properties:*

1. $(A \odot B)^t(A \odot B) = (A^tA) * (B^tB)$, *where $*$ denotes the Hadamard product;*
2. $(A \odot B)^+ \equiv ((A^tA) * (B^tB))^+(A \odot B)^t = [(A \odot B)^t(A \odot B)]^+(A \odot B)^t$ *is the Moore-Penrose pseudo-inverse of $A \odot B$.*

Proof. 1. By definition, the i-th row and j-th column of the left-hand side is

$$[(A \odot B)^t(A \odot B)]_{ij} = (A_{:,i} \otimes B_{:,i})^t(A_{:,j} \otimes B_{:,j}) = ((A_{:,i})^tA_{:,j})((B_{:,i})^tB_{:,j}) = (A^tA)_{ij}(B^tB)_{ij}.$$

2. We have $(A \odot B)(A \odot B)^+(A \odot B) = (A \odot B)((A^tA) * (B^tB))^+(A \odot B)^t(A \odot B)$. But, using Property 1, we get

$$(A \odot B)(A \odot B)^+(A \odot B) = (A \odot B)[(A^tA) * (B^tB)]^+[(A^tA) * (B^tB)].$$

Note that $P \equiv [(A^tA) * (B^tB)]^+[(A^tA) * (B^tB)]$ satisfies $P^2 = P$, i.e. it is a projection operator. Since $(A^tA) * (B^tB) = (A \odot B)^t(A \odot B)$, $\ker((A^tA) * (B^tB)) = \ker(A \odot B)$. Hence, P is a projection operator onto the domain of $A \odot B$, and $(A \odot B)P = A \odot B$. We have thus shown that

$$(A \odot B)(A \odot B)^+(A \odot B) = (A \odot B).$$

Similarly,

$$\begin{aligned}(A \odot B)^+(A \odot B)(A \odot B)^+ &= [(A^tA) * (B^tB)]^+[(A^tA) * (B^tB)](A \odot B)^+ \\ &= [(A^tA) * (B^tB)]^+[(A^tA) * (B^tB)][(A^tA) * (B^tB)]^+(A \odot B)^t \\ &= [(A^tA) * (B^tB)]^+(A \odot B)^t = (A \odot B)^+,\end{aligned}$$

where we have used the fact that $[(A^tA) * (B^tB)]^+$ is the Moore-Penrose pseudo-inverse of $(A^tA) * (B^tB)$.

To check the symmetry property, note that

$$(A \odot B)(A \odot B)^+ = (A \odot B)[(A^tA) * (B^tB)]^+(A \odot B)^t,$$

which is manifestly symmetric. Similarly,

$$(A \odot B)^+(A \odot B) = [(A^tA) * (B^tB)]^+[(A^tA) * (B^tB)],$$

which has to be symmetric since $[(A^tA) * (B^tB)]^+$ is the Moore-Penrose pseudo-inverse of $(A^tA) * (B^tB)$. $\square$

The following theorem now follows upon using mathematical induction:

Theorem A.11. *Let $A_1, \ldots, A_p$ be matrices with K columns. Then, the Khatri-Rao product $A_1 \odot \cdots \odot A_p$ satisfies the following properties:*

1. $(A_1 \odot \cdots \odot A_p)^t(A_1 \odot \cdots \odot A_p) = (A_1^tA_1) * \cdots * (A_p^tA_p)$;
2. $(A_1 \odot \cdots \odot A_p)^+ = ((A_1^tA_1) * \cdots * (A_p^tA_p))^+(A_1 \odot \cdots \odot A_p)^t$.

Theorem A.12. *Let A and B be matrices. Then,*

$$(A \otimes B)^t = (A^t \otimes B^t).$$

Proof. Exercise. $\square$

Using mathematical induction, we can also prove

Theorem A.13. *Let $A_1, \ldots, A_p$ be matrices. Then,*

$$(A_1 \otimes \ldots \otimes A_p)^t = (A_1^t \otimes \cdots \otimes A_p^t).$$

Theorem A.14. *Let $A \in O(m)$ and $B \in O(n)$ be orthogonal matrices. Then $A \otimes B$ is also orthogonal.*

Proof. Using Theorem A.13, we have

$$(A \otimes B)^t(A \otimes B) = (A^t \otimes B^t)(A \otimes B) = (A^tA) \otimes (B^tB) = I_{m\times m} \otimes I_{n\times n}.$$

$\square$

Again by mathematical induction, we can prove

Theorem A.15. *Let $A_1, \ldots, A_p$ be orthogonal matrices. Then, $A_1 \otimes \ldots \otimes A_p$ is an orthogonal matrix.*

A.9 Block Matrix Inversion

Let M be an invertible $(m + n) \times (m + n)$ matrix in a block form

$$M = \begin{array}{cc} \begin{array}{cc} m & n \end{array} & \\ \begin{pmatrix} A & B \\ C & D \end{pmatrix} & \begin{array}{c} m \\ n \end{array} \end{array},$$

where A, D, $A - BD^{-1}C$ and $D - CA^{-1}B$ are themselves invertible. Then,

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}. \tag{A.6}$$

A.10 Multivariate Normal Random Vectors

A random vector $X = (X_1, X_2, \ldots, X_n)$ is said to be a multivariate normal random variable if any linear combination $\sum_{i=1}^{n} \alpha_i X_i$, $\alpha_i \in \mathbb{R}$, is a univariate normal random variable. In particular, it implies that the marginal distribution of each X_i is normal. The covariance matrix Σ of X is defined by the matrix elements $\Sigma_{ij} = Cov[X_i, X_j]$. N.B. Note that Σ is symmetric and, thus, can be diagonalized. When Σ is invertible, we can define the joint density of $X_1, \ldots, X_n$ as

$$p(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}Q(x_1,\ldots,x_n)},$$

where $|\Sigma|$ is the determinant of Σ and the quadratic form Q is defined as

$$Q = (\boldsymbol{x} - \boldsymbol{\mu})^t\, \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{A.7}$$

Multivariate normal distributions satisfy interesting partition theorems. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a multivariate normal random vector. Let $\boldsymbol{X_a} = (X_1, \ldots, X_k)$ and $\boldsymbol{X_b} = (X_{k+1}, \ldots, X_n)$. The covariance matrix of $\boldsymbol{X}$ can be decomposed as

$$\Sigma = Var[\boldsymbol{X}] = \begin{pmatrix} Cov(\boldsymbol{X_a}, \boldsymbol{X_a}) & Cov(\boldsymbol{X_a}, \boldsymbol{X_b}) \\ Cov(\boldsymbol{X_b}, \boldsymbol{X_a}) & Cov(\boldsymbol{X_b}, \boldsymbol{X_b}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{aa}} & \boldsymbol{\Sigma_{ab}} \\ \boldsymbol{\Sigma_{ba}} & \boldsymbol{\Sigma_{bb}} \end{pmatrix}.$$

Its inverse matrix Λ, called the precision matrix, can be similarly decomposed as

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \boldsymbol{\Lambda_{aa}} & \boldsymbol{\Lambda_{ab}} \\ \boldsymbol{\Lambda_{ba}} & \boldsymbol{\Lambda_{bb}} \end{pmatrix}.$$

By the block matrix inversion formula (A.6), we see that

$$\boldsymbol{\Lambda_{aa}} = (\boldsymbol{\Sigma_{aa}} - \boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}}^{-1}\boldsymbol{\Sigma_{ba}})^{-1}$$

and

$$\boldsymbol{\Sigma_{aa}}^{-1} = (\boldsymbol{\Lambda_{aa}} - \boldsymbol{\Lambda_{ab}}\boldsymbol{\Lambda_{bb}}^{-1}\boldsymbol{\Lambda_{ba}}). \tag{A.8}$$

Theorem A.16 (Marginal Distribution)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a multivariate normal random vector with mean $\boldsymbol{\mu}$, and $\boldsymbol{X_a} = (X_1, \ldots, X_k)$ and $\boldsymbol{X_b} = (X_{k+1}, \ldots, X_n)$. Then, the marginal distribution of $\boldsymbol{X_a}$ is multivariate normal with mean $\boldsymbol{\mu_a}$ and covariance matrix $\Sigma' = Cov(\boldsymbol{X_a}, \boldsymbol{X_a})$.*

Proof. You will show in Problem Set that integrating out $\boldsymbol{X_b}$ from the full joint distribution yields the right-hand side of (A.8) in the exponentiated quadratic form. □

REMARK A.10. *We can apply (A.8), because we have assumed that the covariance matrix is invertible and is thus positive definite.*

Theorem A.17 (Conditional Distribution)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a multivariate normal random vector with mean $\boldsymbol{\mu}$, and $\boldsymbol{X_a} = (X_1, \ldots, X_k)$ and $\boldsymbol{X_b} = (X_{k+1}, \ldots, X_n)$. Then, the conditional distribution of $\boldsymbol{X_a}$ given $\boldsymbol{X_b} = \boldsymbol{x_b}$ is multivariate normal with mean*

$$E[\boldsymbol{X_a}|\boldsymbol{X_b} = \boldsymbol{x_b}] = \boldsymbol{\mu_a} + \boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b} - \boldsymbol{\mu_b})$$

and variance

$$Var[\boldsymbol{X_a}|\boldsymbol{x_b}] = \boldsymbol{\Sigma_{aa}} - \boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}} = \boldsymbol{\Lambda_{aa}^{-1}}.$$

Proof. Treating $\boldsymbol{X_b}$ as a constant vector, the quadratic form Q can be written as

$$Q = \boldsymbol{Y_a}^t\boldsymbol{\Lambda_{aa}}\boldsymbol{Y_a} + 2\,\boldsymbol{Y_a}^t\boldsymbol{\Lambda_{ab}}\boldsymbol{y_b} + \text{const}$$

where $\boldsymbol{Y_a} = \boldsymbol{X_a} - \boldsymbol{\mu_a}$ and $\boldsymbol{y_b} = \boldsymbol{x_b} - \boldsymbol{\mu_b}$. Completing the square, we get

$$Q = (\boldsymbol{Y_a} + \boldsymbol{\Lambda_{aa}^{-1}}\boldsymbol{\Lambda_{ab}}\boldsymbol{y_b})^t\boldsymbol{\Lambda_{aa}}(\boldsymbol{Y_a} + \boldsymbol{\Lambda_{aa}^{-1}}\boldsymbol{\Lambda_{ab}}\boldsymbol{y_b}) + \text{const}.$$

Hence,

$$Var[\boldsymbol{X_a}|\boldsymbol{x_b}] = \boldsymbol{\Lambda_{aa}^{-1}} \equiv \boldsymbol{\Sigma_{aa}} - \boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}},$$

and

$$E[\boldsymbol{X_a}|\boldsymbol{X_b} = \boldsymbol{x_b}] = \boldsymbol{\mu_a} - \boldsymbol{\Lambda_{aa}^{-1}}\boldsymbol{\Lambda_{ab}}(\boldsymbol{x_b} - \boldsymbol{\mu_b}).$$

Finally, it is left as an exercise to use (A.6) and show that

$$\boldsymbol{\Lambda_{aa}^{-1}}\boldsymbol{\Lambda_{ab}} = -\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}.$$

□

Example A.1. *For $n = 2$, X_1 and X_2 have a variance matrix given by*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

whose inverse is

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}.$$

The bivariate normal distribution can be thus written as

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}.$$

The marginal distributions of X_1 and X_2 are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$.

Example A.2 (Partial Correlation)**.** *Let $X = (X_1, X_2, \ldots, X_n)$ be multivariate normal as defined above. Let $\boldsymbol{X_a} = (X_1, X_2)$ and $\boldsymbol{X_b} = (X_3, \ldots, X_n)$. Then, the partial correlation of X_1 and X_2 given $\boldsymbol{X_b} = \boldsymbol{x_b}$ is defined as*

$$\rho_{12;\cdot} = \frac{Cov(X_1, X_2 | \boldsymbol{x_b})}{\sqrt{Var[X_1|\boldsymbol{x_b}]Var[X_2|\boldsymbol{x_b}]}}.$$

But, by Theorem A.17, the conditional covariance matrix $Var[\boldsymbol{X_a}|\boldsymbol{x_b}]$ is

$$Var[\boldsymbol{X_a}|\boldsymbol{x_b}] = \boldsymbol{\Lambda_{aa}^{-1}} = \frac{1}{\Lambda_{11}\Lambda_{22} - \Lambda_{12}\Lambda_{21}} \begin{pmatrix} \Lambda_{22} & -\Lambda_{12} \\ -\Lambda_{21} & \Lambda_{11} \end{pmatrix}.$$

Thus, we have

$$\rho_{12;\cdot} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}}.$$

A.11 Matrix Version of Completing the Squares

Let $x, y \in \mathbb{R}^n$, and let $M \in \mathbb{R}^{n \times n}$ be an invertible symmetric matrix.

$$x^t y + x^t M x = \left(x + \frac{M^{-1}y}{2}\right)^t M \left(x + \frac{M^{-1}y}{2}\right) - \frac{y^t M^{-1} y}{4} \tag{A.9}$$

A.12 Mean Value Theorem

Recall the following theorem from calculus.

Theorem A.18 (Mean Value Theorem)**.** *Let $[a, b] \subset \mathbb{R}$ be a closed interval, and let $f : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable in (a, b). Then, there exists $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c)(b - a).$$

Corollary A.2. *Let $g : [a, b] \to \mathbb{R}$ be in class $\mathcal{C}^{m-1}([a, b])$, i.e. it is continuous with continuous derivatives $g^{(1)}, \ldots, g^{(m-1)}$ up to order $m - 1$. If g has m-th derivative $g^{(m)}$ on (a, b) and satisfies*

$$g^{(k)}(a) = 0,\ 0 \le k < m, \quad \text{and} \quad g(b) = 0$$

then there exists a sequence of points c_i, $i = 1, \ldots, m$, such that

$$a < c_m < c_{m-1} < \cdots < c_1 < b$$

and

$$g^{(k)}(c_k) = 0,\ 1 \le k \le m.$$

Proof. We will prove this theorem by induction. For $m = 1$, note that since $g(a) = 0$ and $g(b) = 0$, Theorem A.18 implies that there exists $c_1 \in (a, b)$, such that $g^{(1)}(c_1) = 0$. Now, for some positive integer n, assume that the claim holds for all functions in $\mathcal{C}^{n-1}$ satisfying the above assumptions for $m = n$, and choose any $g \in \mathcal{C}^n$ satisfying the above assumptions for $m = n+1$. Then, by the induction hypothesis, there exist $a < c_n < c_{n-1} < \cdots < c_1 < b$ such that

$$g^{(k)}(c_k) = 0,\ 1 \leq k \leq n.$$

But, since $g^{(n)}(a) = 0$ and $g^{(n)}(c_n) = 0$, Theorem A.18 again implies that there exists $c_{n+1} \in (a, c_n)$, such that $g^{(n+1)}(c_{n+1}) = 0$. □

Theorem A.19 (Generalized Mean Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be in class $\mathcal{C}^{m-1}([a,b])$, and assume that its m-th derivative exists on (a,b). Then, there exists $c_m \in (a,b)$ such that*

$$f(b) = \sum_{k=0}^{m-1} \frac{f^{(k)}(a)}{k!}\,(b-a)^k + \frac{f^{(m)}(c_m)}{m!}\,(b-a)^m.$$

Proof. For $z \in [a,b]$, let

$$r(z) = f(z) - \sum_{k=0}^{m-1} \frac{f^{(k)}(a)}{k!}\,(z-a)^k,$$

which is the remainder of the Taylor expansion of f around a. Define

$$g(z) \equiv f(z) - \sum_{k=0}^{m-1} \frac{f^{(k)}(a)}{k!}\,(z-a)^k - r(b)\frac{(z-a)^m}{(b-a)^m},$$

which can be easily checked to satisfy

$$g^{(k)}(a) = 0,\ 0 \leq k < m, \quad \text{and} \quad g(b) = 0.$$

The condition $g(b) = 0$ is equivalent to

$$f(b) = \sum_{k=0}^{m-1} \frac{f^{(k)}(a)}{k!}\,(b-a)^k + r(b).$$

But, Corollary A.2 implies that there exists $c_m \in (a,b)$, such that

$$g^{(m)}(c_m) = 0 = f^{(m)}(c_m) - r(b)\,\frac{m!}{(b-a)^m} \Rightarrow r(b) = \frac{f^{(m)}(c_m)}{m!}\,(b-a)^m.$$

□

Corollary A.3. *Let f be a real-valued function twice differentiable on an open set in $\mathbb{R}^n$ containing the line segment $\{(1-\tau)x + \tau y \,|\, 0 \leq \tau \leq 1\}$ for some fixed $x, y \in \mathbb{R}^n$. Then,*

$$f(y) = f(x) + (y-x)^t \nabla f(x) + \frac{1}{2}(y-x)^t Q((1-\tau^*)x + \tau^* y)(y-x)$$

for some $0 < \tau^* < 1$, where $Q((1-\tau^*)x + \tau^* y)$ is the Hessian matrix of f at $(1-\tau^*)x + \tau^* y$.

Proof. Define $h : [0,1] \to \mathbb{R}$ by

$$h(\tau) = f((1-\tau)x + \tau y).$$

Then, Theorem A.19 implies that $\exists\, \tau^* \in (0,1)$ such that

$$h(1) = h(0) + h'(0) + \frac{1}{2} h''(\tau^*).$$

But, we have $h(1) = f(y)$, $h(0) = f(x)$ and

$$h'(0) = (y-x)^t \nabla f(x) \quad \text{and} \quad h''(\tau^*) = (y-x)^t Q((1-\tau^*)x + \tau^* y)(y-x).$$

$\square$