

# matrix and gaussian identities

sam roweis\*

---

**note** that  $\mathbf{a}, \mathbf{b}, \mathbf{c}$  and  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  do not depend on  $\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$  or  $z$

## 0.1 basic formulae

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \quad (1)$$

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \quad (2)$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T \quad (3)$$

$$\text{if individual inverses exist } (\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1} \quad (4)$$

$$(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1} \quad (5)$$

## 0.2 trace, determinant and rank

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \quad (6)$$

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \quad (7)$$

$$|\mathbf{A}| = \prod \text{evals} \quad (8)$$

$$\text{Tr}[\mathbf{A}] = \sum \text{evals} \quad (9)$$

if the cyclic products are well defined,

$$\text{Tr}[\mathbf{ABC} \dots] = \text{Tr}[\mathbf{BC} \dots \mathbf{A}] = \text{Tr}[\mathbf{C} \dots \mathbf{AB}] = \dots \quad (10)$$

$$\text{rank}[\mathbf{A}] = \text{rank}[\mathbf{A}^T \mathbf{A}] = \text{rank}[\mathbf{A} \mathbf{A}^T] \quad (11)$$

$$\text{condition number} = \gamma = \sqrt{\frac{\text{biggest eval}}{\text{smallest eval}}} \quad (12)$$

**derivatives** of scalar forms with respect to scalars, vectors, or matrices are indexed in the obvious way. similarly, the indexing for derivatives of vectors and matrices with respect to scalars is straightforward.

---

\*Thanks to Mike Brooks and Zoubin Ghahramani for contributing identities to these pages.

### 0.3 derivatives of traces

$$\frac{\partial \text{Tr}[\mathbf{X}]}{\partial \mathbf{X}} = \mathbf{I} \quad (13)$$

$$\frac{\partial \text{Tr}[\mathbf{X}\mathbf{A}]}{\partial \mathbf{X}} = \frac{\partial \text{Tr}[\mathbf{A}\mathbf{X}]}{\partial \mathbf{X}} = \mathbf{A}^T \quad (14)$$

$$\frac{\partial \text{Tr}[\mathbf{X}^T \mathbf{A}]}{\partial \mathbf{X}} = \frac{\partial \text{Tr}[\mathbf{A}\mathbf{X}^T]}{\partial \mathbf{X}} = \mathbf{A} \quad (15)$$

$$\frac{\partial \text{Tr}[\mathbf{X}^T \mathbf{A}\mathbf{X}]}{\partial \mathbf{X}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{X} \quad (16)$$

$$\frac{\partial \text{Tr}[\mathbf{X}^{-1}\mathbf{A}]}{\partial \mathbf{X}} = -\mathbf{X}^{-1}\mathbf{A}^T\mathbf{X}^{-1} \quad (17)$$

### 0.4 derivatives of determinants

$$\frac{\partial |\mathbf{A}\mathbf{X}\mathbf{B}|}{\partial \mathbf{X}} = |\mathbf{A}\mathbf{X}\mathbf{B}|(\mathbf{X}^{-1})^T = |\mathbf{A}\mathbf{X}\mathbf{B}|(\mathbf{X}^T)^{-1} \quad (18)$$

$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1} \quad (19)$$

$$\frac{\partial \ln |\mathbf{X}(z)|}{\partial z} = \text{Tr} \left[ (\mathbf{X}^{-1})^T \frac{\partial \mathbf{X}}{\partial z} \right] \quad (20)$$

for real, square  $\mathbf{A}$   $\frac{\partial |\mathbf{X}^T \mathbf{A}\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}^T \mathbf{A}\mathbf{X}|(\mathbf{A} + \mathbf{A}^T)\mathbf{X}(\mathbf{X}^T \mathbf{A}\mathbf{X})^{-1}$  (21)

### 0.5 derivatives of scalar forms

$$\frac{\partial (\mathbf{a}^T \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}^T \mathbf{a})}{\partial \mathbf{x}} = \mathbf{a} \quad (22)$$

$$\frac{\partial (\mathbf{x}^T \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \quad (23)$$

$$\frac{\partial (\mathbf{a}^T \mathbf{X}\mathbf{b})}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \quad (24)$$

$$\frac{\partial (\mathbf{a}^T \mathbf{X}^T \mathbf{b})}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \quad (25)$$

$$\frac{\partial (\mathbf{a}^T \mathbf{X}\mathbf{a})}{\partial \mathbf{X}} = \frac{\partial (\mathbf{a}^T \mathbf{X}^T \mathbf{a})}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \quad (26)$$

$$\frac{\partial (\mathbf{a}^T \mathbf{X}^T \mathbf{C}\mathbf{X}\mathbf{b})}{\partial \mathbf{X}} = \mathbf{C}^T \mathbf{X}\mathbf{a}\mathbf{b}^T + \mathbf{C}\mathbf{X}\mathbf{b}\mathbf{a}^T \quad (27)$$

$$\frac{\partial ((\mathbf{X}\mathbf{a} + \mathbf{b})^T \mathbf{C}(\mathbf{X}\mathbf{a} + \mathbf{b}))}{\partial \mathbf{X}} = (\mathbf{C} + \mathbf{C}^T)(\mathbf{X}\mathbf{a} + \mathbf{b})\mathbf{a}^T \quad (28)$$

the **derivative** of one vector  $\mathbf{y}$  with respect to another vector  $\mathbf{x}$  is a matrix whose  $(i, j)^{th}$  element is  $\partial y(j)/\partial x(i)$ . such a derivative should be written as  $\partial \mathbf{y}^T/\partial \mathbf{x}$  in which case it is the *Jacobian* matrix of  $\mathbf{y}$  wrt  $\mathbf{x}$ . its determinant represents the ratio of the hypervolume  $d\mathbf{y}$  to that of  $d\mathbf{x}$  so that  $\int f(\mathbf{y})d\mathbf{y} = \int f(\mathbf{y}(\mathbf{x}))|\partial \mathbf{y}^T/\partial \mathbf{x}|d\mathbf{x}$ . however, the sloppy forms  $\partial \mathbf{y}/\partial \mathbf{x}$ ,  $\partial \mathbf{y}^T/\partial \mathbf{x}^T$  and  $\partial \mathbf{y}/\partial \mathbf{x}^T$  are often used for this Jacobain matrix.

## 0.6 derivatives of vector/matrix forms

$$\frac{\partial(\mathbf{X}^{-1})}{\partial z} = -\mathbf{X}^{-1}\frac{\partial \mathbf{X}}{\partial z}\mathbf{X}^{-1} \quad (29)$$

$$\frac{\partial(\mathbf{A}\mathbf{x})}{\partial z} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial z} \quad (30)$$

$$\frac{\partial(\mathbf{X}\mathbf{Y})}{\partial z} = \mathbf{X}\frac{\partial \mathbf{Y}}{\partial z} + \frac{\partial \mathbf{X}}{\partial z}\mathbf{Y} \quad (31)$$

$$\frac{\partial(\mathbf{A}\mathbf{X}\mathbf{B})}{\partial z} = \mathbf{A}\frac{\partial \mathbf{X}}{\partial z}\mathbf{B} \quad (32)$$

$$\frac{\partial(\mathbf{x}^T \mathbf{A})}{\partial \mathbf{x}} = \mathbf{A} \quad (33)$$

$$\frac{\partial(\mathbf{x}^T)}{\partial \mathbf{x}} = \mathbf{I} \quad (34)$$

$$\frac{\partial(\mathbf{x}^T \mathbf{A} \mathbf{x} \mathbf{x}^T)}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \mathbf{x}^T + \mathbf{x}^T \mathbf{A} \mathbf{x} \mathbf{I} \quad (35)$$

## 0.7 constrained maximization

the maximum over  $\mathbf{x}$  of the quadratic form:

$$\boldsymbol{\mu}^T \mathbf{x} - \frac{1}{2} \mathbf{x}^T \mathbf{A}^{-1} \mathbf{x} \quad (36)$$

subject to the  $J$  conditions  $c_j(\mathbf{x}) = 0$  is given by:

$$\mathbf{A}\boldsymbol{\mu} + \mathbf{A}\mathbf{C}\boldsymbol{\Lambda}, \quad \boldsymbol{\Lambda} = -4(\mathbf{C}^T \mathbf{A} \mathbf{C})\mathbf{C}^T \mathbf{A}\boldsymbol{\mu} \quad (37)$$

where the  $j$ th column of  $\mathbf{C}$  is  $\partial c_j(\mathbf{x})/\partial \mathbf{x}$

## 0.8 symmetric matrices

have real eigenvalues, though perhaps not distinct and can always be diagonalized to the form:

$$\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^T \quad (38)$$

where the columns of  $\mathbf{C}$  are (orthonormal) eigenvectors (i.e.  $\mathbf{C}\mathbf{C}^T = \mathbf{I}$ ) and the diagonal of  $\mathbf{\Lambda}$  has the eigenvalues

## 0.9 block matrices

for conformably partitioned block matrices, addition and multiplication is performed by adding and multiplying blocks in exactly the same way as scalar elements of regular matrices

however, determinants and inverses of block matrices are very tricky; for 2 blocks by 2 blocks the results are:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}| \cdot |\mathbf{F}_{11}| = |\mathbf{A}_{11}| \cdot |\mathbf{F}_{22}| \quad (39)$$

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{F}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_{22}^{-1} \\ -\mathbf{F}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_{22}^{-1} \end{bmatrix} \quad (40)$$

$$= \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{F}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{F}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{F}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \quad (41)$$

where

$$\mathbf{F}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \quad (42)$$

$$\mathbf{F}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \quad (43)$$

for block *diagonal* matrices things are much easier:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| |\mathbf{A}_{22}| \quad (44)$$

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix} \quad (45)$$

## 0.10 matrix inversion lemma

using the above results for block matrices we can make some substitutions and get the following important result:

$$(\mathbf{A} + \mathbf{X}\mathbf{B}\mathbf{X}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{B}^{-1} + \mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1} \quad (46)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are *square* and *invertible* matrices but need not be of the same dimension. this lemma often allows a really hard inverse to be converted into an easy inverse. the most typical example of this is when  $\mathbf{A}$  is large but diagonal, and  $\mathbf{X}$  has many rows but few columns

### 0.11 multidimensional gaussian

a  $d$ -dimensional multidimensional gaussian (normal) density for  $\mathbf{x}$  is:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \quad (47)$$

it has entropy:

$$S = \frac{1}{2} \log_2 \left[ (2\pi e)^d |\boldsymbol{\Sigma}| \right] - \text{const} \quad \text{bits} \quad (48)$$

where  $\boldsymbol{\Sigma}$  is a symmetric positive semi-definite covariance matrix and the (unfortunate) constant is the log of the units in which  $\mathbf{x}$  is measured over the “natural units”

### 0.12 linear functions of a normal vector

no matter how  $\mathbf{x}$  is distributed,

$$\mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{y}] = \mathbf{A}(\mathbb{E}[\mathbf{x}]) + \mathbf{y} \quad (49)$$

$$\text{Covar}[\mathbf{A}\mathbf{x} + \mathbf{y}] = \mathbf{A}(\text{Covar}[\mathbf{x}])\mathbf{A}^T \quad (50)$$

in particular this means that for normal distributed quantities:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{A}\mathbf{x} + \mathbf{y}) \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{y}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) \quad (51)$$

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (52)$$

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_n^2 \quad (53)$$

### 0.13 marginal and conditional distributions

let the vector  $\mathbf{z} = [\mathbf{x}^T \mathbf{y}^T]^T$  be normally distributed according to:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right) \quad (54)$$

where  $\mathbf{C}$  is the (non-symmetric) cross-covariance matrix between  $\mathbf{x}$  and  $\mathbf{y}$  which has as many rows as the size of  $\mathbf{x}$  and as many columns as the size of  $\mathbf{y}$ . then the marginal distributions are:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{A}) \quad (55)$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{b}, \mathbf{B}) \quad (56)$$

and the conditional distributions are:

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T) \quad (57)$$

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{b} + \mathbf{C}^T\mathbf{A}^{-1}(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}) \quad (58)$$