matrix identities

sam roweis

(revised June 1999)

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

0.1 basic formulae

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

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T+\mathbf{B}^T \tag{1b}$$

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{1c}$$

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

$$(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1} \tag{1e}$$

0.2 trace, determinant and rank

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \tag{2a}$$

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \tag{2b}$$

$$|\mathbf{A}| = \prod \text{evals} \tag{2c}$$

$$\text{Tr}\,[\mathbf{A}] = \sum \text{evals} \tag{2d}$$

if the cyclic products are well defined,

$$\text{Tr}\,[\mathbf{ABC}\ldots] = \text{Tr}\,[\mathbf{BC}\ldots\mathbf{A}] = \text{Tr}\,[\mathbf{C}\ldots\mathbf{AB}] = \ldots \tag{2e}$$

$$\text{rank}\,[\mathbf{A}] = \text{rank}\left[\mathbf{A}^T\mathbf{A}\right] = \text{rank}\left[\mathbf{AA}^T\right] \tag{2f}$$

$$\text{condition number} = \gamma = \sqrt{\frac{\text{biggest eval}}{\text{smallest eval}}} \tag{2g}$$

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

0.3 derivatives of traces

$$\frac{\partial \mathrm{Tr}\left[\mathbf{X}\right]}{\partial \mathbf{X}} = \mathbf{I} \tag{3a}$$

$$\frac{\partial \mathrm{Tr}\left[\mathbf{XA}\right]}{\partial \mathbf{X}} = \frac{\partial \mathrm{Tr}\left[\mathbf{AX}\right]}{\partial \mathbf{X}} = \mathbf{A}^T \tag{3b}$$

$$\frac{\partial \mathrm{Tr}\left[\mathbf{X}^T\mathbf{A}\right]}{\partial \mathbf{X}} = \frac{\partial \mathrm{Tr}\left[\mathbf{AX}^T\right]}{\partial \mathbf{X}} = \mathbf{A} \tag{3c}$$

$$\frac{\partial \mathrm{Tr}\left[\mathbf{X}^T\mathbf{AX}\right]}{\partial \mathbf{X}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{X} \tag{3d}$$

$$\frac{\partial \mathrm{Tr}\left[\mathbf{X}^{-1}\mathbf{A}\right]}{\partial \mathbf{X}} = -\mathbf{X}^{-1}\mathbf{A}^T\mathbf{X}^{-1} \tag{3e}$$

0.4 derivatives of determinants

$$\frac{\partial |\mathbf{AXB}|}{\partial \mathbf{X}} = |\mathbf{AXB}|(\mathbf{X}^{-1})^T = |\mathbf{AXB}|(\mathbf{X}^T)^{-1} \tag{4a}$$

$$\frac{\partial \ln|\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1} \tag{4b}$$

$$\frac{\partial \ln|\mathbf{X}(z)|}{\partial z} = \mathrm{Tr}\left[\mathbf{X}^{-1}\frac{\partial \mathbf{X}}{\partial z}\right] \tag{4c}$$

$$\text{for real, square } \mathbf{A} \quad \frac{\partial |\mathbf{X}^T\mathbf{AX}|}{\partial \mathbf{X}} = |\mathbf{X}^T\mathbf{AX}|(\mathbf{A} + \mathbf{A}^T)\mathbf{X}(\mathbf{X}^T\mathbf{AX})^{-1} \tag{4d}$$

0.5 derivatives of scalar forms

$$\frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a} \tag{5a}$$

$$\frac{\partial(\mathbf{x}^T\mathbf{Ax})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \tag{5b}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{Xb})}{\partial \mathbf{X}} = \mathbf{ab}^T \tag{5c}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{X}^T\mathbf{b})}{\partial \mathbf{X}} = \mathbf{ba}^T \tag{5d}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{Xa})}{\partial \mathbf{X}} = \frac{\partial(\mathbf{a}^T\mathbf{X}^T\mathbf{a})}{\partial \mathbf{X}} = \mathbf{aa}^T \tag{5e}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{X}^T\mathbf{CXb})}{\partial \mathbf{X}} = \mathbf{C}^T\mathbf{Xab}^T + \mathbf{CXba}^T \tag{5f}$$

$$\frac{\partial\left((\mathbf{Xa} + \mathbf{b})^T\mathbf{C}(\mathbf{Xa} + \mathbf{b})\right)}{\partial \mathbf{X}} = (\mathbf{C} + \mathbf{C}^T)(\mathbf{Xa} + \mathbf{b})\mathbf{a}^T \tag{5g}$$

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

$$\frac{\partial(\mathbf{X}^{-1})}{\partial z} = -\mathbf{X}^{-1}\frac{\partial \mathbf{X}}{\partial z}\mathbf{X}^{-1} \tag{6a}$$

$$\frac{\partial(\mathbf{A}\mathbf{x})}{\partial z} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial z} \tag{6b}$$

$$\frac{\partial(\mathbf{X}\mathbf{Y})}{\partial z} = \mathbf{X}\frac{\partial \mathbf{Y}}{\partial z} + \frac{\partial \mathbf{X}}{\partial z}\mathbf{Y} \tag{6c}$$

$$\frac{\partial(\mathbf{A}\mathbf{X}\mathbf{B})}{\partial z} = \mathbf{A}\frac{\partial \mathbf{X}}{\partial z}\mathbf{B} \tag{6d}$$

$$\frac{\partial(\mathbf{x}^T\mathbf{A})}{\partial \mathbf{x}} = \mathbf{A} \tag{6e}$$

$$\frac{\partial(\mathbf{x}^T)}{\partial \mathbf{x}} = \mathbf{I} \tag{6f}$$

$$\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x}\mathbf{x}^T)}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}\mathbf{x}^T + \mathbf{x}^T\mathbf{A}\mathbf{x}\mathbf{I} \tag{6g}$$

0.7 constrained maximization

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

$$\boldsymbol{\mu}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} \tag{7a}$$

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

$$\mathbf{A}\boldsymbol{\mu} + \mathbf{A}\mathbf{C}\boldsymbol{\Lambda}, \qquad \boldsymbol{\Lambda} = -4(\mathbf{C}^T\mathbf{A}\mathbf{C})\mathbf{C}^T\mathbf{A}\boldsymbol{\mu} \tag{7b}$$

where the jth 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:

$$\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^T \tag{8}$$

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:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}| \cdot |\mathbf{F}_{11}| = |\mathbf{A}_{11}| \cdot |\mathbf{F}_{22}| \tag{9a}$$

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{F}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_{22}^{-1} \\ -\mathbf{F}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_{22}^{-1} \end{bmatrix} \tag{9b}$$

$$= \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{F}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{F}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{F}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \tag{9c}$$

where

$$\mathbf{F}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \tag{9d}$$

$$\mathbf{F}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \tag{9e}$$

for block *diagonal* matrices things are much easier:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}||\mathbf{A}_{22}| \tag{9f}$$

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix} \tag{9g}$$

0.10 matrix inversion lemma

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

$$(\mathbf{A} + \mathbf{X}\mathbf{B}\mathbf{X}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{B}^{-1} + \mathbf{X}^T\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1} \tag{10}$$

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

gaussian identities

sam roweis

(revised July 1999)

0.1 multidimensional gaussian

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

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

it has entropy:

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

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

0.2 linear functions of a normal vector

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

$$\mathrm{E}[\mathbf{A}\mathbf{x}+\mathbf{y}] = \mathbf{A}(\mathrm{E}[\mathbf{x}]) + \mathbf{y} \tag{3a}$$
$$\mathrm{Covar}[\mathbf{A}\mathbf{x}+\mathbf{y}] = \mathbf{A}(\mathrm{Covar}[\mathbf{x}])\mathbf{A}^T \tag{3b}$$

in particular this means that for normal distributed quantities:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{A}\mathbf{x}+\mathbf{y}) \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{y}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right) \tag{4a}$$
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4b}$$
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_n \tag{4c}$$

0.3 marginal and conditional distributions

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

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right) \tag{5a}$$

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:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{A}) \tag{5b}$$
$$\mathbf{y} \sim \mathcal{N}(\mathbf{b}, \mathbf{B}) \tag{5c}$$

and the conditional distributions are:

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T\right) \tag{5d}$$
$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{b} + \mathbf{C}^T\mathbf{A}^{-1}(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right) \tag{5e}$$

0.4 multiplication

the multiplication of two gaussian functions is another gaussian function (although no longer normalized). in particular,

$$\mathcal{N}(\mathbf{a}, \mathbf{A}) \cdot \mathcal{N}(\mathbf{b}, \mathbf{B}) \propto \mathcal{N}(\mathbf{c}, \mathbf{C}) \tag{6a}$$

where

$$\mathbf{C} = \left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1} \tag{6b}$$
$$\mathbf{c} = \mathbf{C}\mathbf{A}^{-1}\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}\mathbf{b} \tag{6c}$$

amazingly, the normalization constant z_c is Gaussian in either $\mathbf{a}$ or $\mathbf{b}$:

$$z_c = (2\pi)^{-d/2}|\mathbf{C}|^{+1/2}|\mathbf{A}|^{-1/2}|\mathbf{B}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{C}^{-1}\mathbf{c})\right] \tag{6d}$$

$$z_c(\mathbf{a}) \sim \mathcal{N}\left((\mathbf{A}^{-1}\mathbf{C}\mathbf{A}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{C}\mathbf{B}^{-1})\mathbf{b}, (\mathbf{A}^{-1}\mathbf{C}\mathbf{A}^{-1})^{-1}\right) \tag{6e}$$
$$z_c(\mathbf{b}) \sim \mathcal{N}\left((\mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{C}\mathbf{A}^{-1})\mathbf{a}, (\mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-1})^{-1}\right) \tag{6f}$$

0.5 quadratic forms

the expectation of a quadratic form under a gaussian is another quadratic form (plus an annoying constant). in particular, if $\mathbf{x}$ is gaussian distributed with mean $\mathbf{m}$ and variance $\mathbf{S}$ then,

$$\int_{\mathbf{x}} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \mathcal{N}(\mathbf{m},\mathbf{S})\, d\mathbf{x} = (\boldsymbol{\mu}-\mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\mathbf{m}) + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \quad (7a)$$

if the original quadratic form has a linear function of $\mathbf{x}$ the result is still simple:

$$\int_{\mathbf{x}} (\mathbf{W}\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{W}\mathbf{x}-\boldsymbol{\mu}) \mathcal{N}(\mathbf{m},\mathbf{S})\, d\mathbf{x} = (\boldsymbol{\mu}-\mathbf{W}\mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\mathbf{W}\mathbf{m}) + \mathrm{Tr}\left[\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{S}\right] \quad (7b)$$

0.6 convolution

the convolution of two gaussian functions is another gaussian function (although no longer normalized). in particular,

$$\mathcal{N}(\mathbf{a},\mathbf{A}) * \mathcal{N}(\mathbf{b},\mathbf{B}) \propto \mathcal{N}(\mathbf{a}+\mathbf{b},\mathbf{A}+\mathbf{B}) \quad (8)$$

this is a direct consequence of the fact that the Fourier transform of a gaussian is another gaussian and that the multiplication of two gaussians is still gaussian.

0.7 Fourier transform

the (inverse)Fourier transform of a gaussian function is another gaussian function (although no longer normalized). in particular,

$$\mathcal{F}\left[\mathcal{N}(\mathbf{a},\mathbf{A})\right] \propto \mathcal{N}\left(j\mathbf{A}^{-1}\mathbf{a},\mathbf{A}^{-1}\right) \quad (9a)$$

$$\mathcal{F}^{-1}\left[\mathcal{N}(\mathbf{b},\mathbf{B})\right] \propto \mathcal{N}\left(-j\mathbf{B}^{-1}\mathbf{b},\mathbf{B}^{-1}\right) \quad (9b)$$

where $j = \sqrt{-1}$

0.8 constrained maximization

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

$$\boldsymbol{\mu}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} \tag{10a}$$

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

$$\mathbf{A}\boldsymbol{\mu} + \mathbf{AC}\boldsymbol{\Lambda}, \qquad \boldsymbol{\Lambda} = -4(\mathbf{C}^T\mathbf{AC})\mathbf{C}^T\mathbf{A}\boldsymbol{\mu} \tag{10b}$$

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