Old and New Matrix Algebra Useful for Statistics

Thomas P. Minka

December 28, 2000

Contents

Warning: This paper contains a large number of matrix identities which cannot be absorbed by mere reading. The reader is encouraged to take time and check each equation by hand and work out the examples. This is advanced material; see Searle (1982) for basic results.

1 Derivatives

Maximum-likelihood problems almost always require derivatives. There are six kinds of derivatives that can be expressed as matrices:

	Scalar	Vector	Matrix
Scalar	$\frac{\mathrm{d}y}{\mathrm{d}x}$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}x} = \left[\frac{\partial y_i}{\partial x}\right]$	$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}x} = \left[\frac{\partial y_{ij}}{\partial x}\right]$
Vector	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y}{\partial x_j}\right]$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]$	
Matrix	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \left[\frac{\partial y}{\partial x_{ji}}\right]$		

The partials with respect to the numerator are laid out according to the shape of $\mathbf{Y}$ while the partials with respect to the denominator are laid out according to the transpose of $\mathbf{X}$. For example, $\mathrm{d}\mathbf{y}/\mathrm{d}x$ is a column vector while $\mathrm{d}y/\mathrm{d}\mathbf{x}$ is a row vector (assuming $\mathbf{x}$ and $\mathbf{y}$ are column vectors—otherwise it is flipped). Each of these derivatives can be tediously computed via partials, but this section shows how they instead can be computed with matrix manipulations. The material is based on Magnus and Neudecker (1988).

Define the differential $\mathrm{d}y(x)$ to be that part of $y(x+\mathrm{d}x) - y(x)$ which is linear in $\mathrm{d}x$. Unlike the classical definition in terms of limits, this definition applies even when x or y are not scalars.

For example, this equation:

$$\mathbf{y}(\mathbf{x}+\mathrm{d}\mathbf{x}) = \mathbf{y}(\mathbf{x}) + \mathbf{A}\mathrm{d}\mathbf{x} + \text{(higher order terms)} \tag{1}$$

is well-defined for any $\mathbf{y}$ satisfying certain continuity properties. The matrix $\mathbf{A}$ is the derivative, as you can check by setting all but one component of $\mathrm{d}\mathbf{x}$ to zero and making it small. The matrix $\mathbf{A}$ is also called the Jacobian matrix $\mathbf{J}_{x\to y}$. Its transpose is the gradient of $\mathbf{y}$, denoted $\nabla\mathbf{y}$. The Jacobian is useful in calculus while the gradient is useful in optimization.

Therefore, the derivative of any expression involving matrices can be computed in two steps:

1. compute the differential
2. massage the result into canonical form

after which the derivative is immediately read off as the coefficient of $\mathrm{d}x$, $\mathrm{d}\mathbf{x}$, or $\mathrm{d}\mathbf{X}$.

The differential of an expression can be computed by iteratively applying the following rules:

$$\begin{aligned}
\mathrm{d}\mathbf{A} &= 0 \quad \text{(for constant } \mathbf{A}\text{)} && (2)\\
\mathrm{d}(\alpha\mathbf{X}) &= \alpha\mathrm{d}\mathbf{X} && (3)\\
\mathrm{d}(\mathbf{X}+\mathbf{Y}) &= \mathrm{d}\mathbf{X}+\mathrm{d}\mathbf{Y} && (4)\\
\mathrm{d}(\mathrm{tr}(\mathbf{X})) &= \mathrm{tr}(\mathrm{d}\mathbf{X}) && (5)\\
\mathrm{d}(\mathbf{X}\mathbf{Y}) &= (\mathrm{d}\mathbf{X})\mathbf{Y}+\mathbf{X}\mathrm{d}\mathbf{Y} && (6)\\
\mathrm{d}(\mathbf{X}\otimes\mathbf{Y}) &= (\mathrm{d}\mathbf{X})\otimes\mathbf{Y}+\mathbf{X}\otimes\mathrm{d}\mathbf{Y} \quad \text{(see section 2)} && (7)\\
\mathrm{d}(\mathbf{X}\circ\mathbf{Y}) &= (\mathrm{d}\mathbf{X})\circ\mathbf{Y}+\mathbf{X}\circ\mathrm{d}\mathbf{Y} \quad \text{(see section 5)} && (8)\\
\mathrm{d}\mathbf{X}^{-1} &= -\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1} && (9)\\
\mathrm{d}\,|\mathbf{X}| &= |\mathbf{X}|\,\mathrm{tr}(\mathbf{X}^{-1}\mathrm{d}\mathbf{X}) && (10)\\
\mathrm{d}\log|\mathbf{X}| &= \mathrm{tr}(\mathbf{X}^{-1}\mathrm{d}\mathbf{X}) && (11)\\
\mathrm{d}\mathbf{X}^{\star} &= (\mathrm{d}\mathbf{X})^{\star} && (12)
\end{aligned}$$

where * is any operator that rearranges elements, e.g. transpose, vec, and vec-transpose (section 3). The rules can be iteratively applied because of the chain rule, e.g. $\mathrm{d}(\mathbf{AX}+\mathbf{Y}) = \mathrm{d}(\mathbf{AX}) + \mathrm{d}\mathbf{Y} = \mathbf{A}\mathrm{d}\mathbf{X} + (\mathrm{d}\mathbf{A})\mathbf{X} + \mathrm{d}\mathbf{Y} = \mathbf{A}\mathrm{d}\mathbf{X} + \mathrm{d}\mathbf{Y}$. Most of these rules can be derived by subtracting $\mathbf{F}(\mathbf{X}+\mathrm{d}\mathbf{X}) - \mathbf{F}(\mathbf{X})$ and taking the linear part. For example,

$$(\mathbf{X}+\mathrm{d}\mathbf{X})(\mathbf{Y}+\mathrm{d}\mathbf{Y}) = \mathbf{XY} + (\mathrm{d}\mathbf{X})\mathbf{Y} + \mathbf{X}\mathrm{d}\mathbf{Y} + (\mathrm{d}\mathbf{X})(\mathrm{d}\mathbf{Y})$$

from which (6) follows.

To derive $\mathrm{d}\mathbf{X}^{-1}$, note that

$$0 = \mathrm{d}\mathbf{I} = \mathrm{d}\mathbf{X}^{-1}\mathbf{X} = (\mathrm{d}\mathbf{X}^{-1})\mathbf{X} + \mathbf{X}^{-1}\mathrm{d}\mathbf{X}$$

from which (9) follows.

The next step is to massage the differential into one of the six canonical forms (assuming $\mathbf{x}$ and $\mathbf{y}$ are column vectors):

$\mathrm{d}y = a\mathrm{d}x$	$\mathrm{d}\mathbf{y} = \mathbf{a}\mathrm{d}x$	$\mathrm{d}\mathbf{Y} = \mathbf{A}\mathrm{d}x$
$\mathrm{d}y = \mathbf{a}\mathrm{d}\mathbf{x}$	$\mathrm{d}\mathbf{y} = \mathbf{A}\mathrm{d}\mathbf{x}$	
$\mathrm{d}y = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X})$		

This is where the operators and identities developed in the following sections are useful. For example, since the derivative of $\mathbf{Y}$ with respect to $\mathbf{X}$ cannot be represented by a matrix, it is customary to use $\mathrm{dvec}(\mathbf{Y})/\mathrm{dvec}(\mathbf{X})$ instead (vec is defined in section 2). If the purpose of differentiation is to equate the derivative to zero, then this transformation doesn't affect the result. So after expanding the differential, just take vec of both sides and use the identities in sections 2 and 3 to get it into canonical form.

One particularly helpful identity is:

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \tag{13}$$

Examples:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}\mathrm{tr}(\mathbf{AXB}) &= \mathbf{BA} \qquad (14)\\ \text{because } \mathrm{dtr}(\mathbf{AXB}) &= \mathrm{tr}(\mathbf{A}(\mathrm{d}\mathbf{X})\mathbf{B}) = \mathrm{tr}(\mathbf{BA}\mathrm{d}\mathbf{X}) \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}\mathrm{tr}(\mathbf{AX}'\mathbf{BXC}) &= \mathbf{CAX}'\mathbf{B} + \mathbf{A}'\mathbf{C}'\mathbf{X}'\mathbf{B}' \qquad (15)\\ \text{because } \mathrm{dtr}(\mathbf{AX}'\mathbf{BXC}) &= \mathrm{tr}(\mathbf{AX}'\mathbf{B}(\mathrm{d}\mathbf{X})\mathbf{C}) + \mathrm{tr}(\mathbf{A}(\mathrm{d}\mathbf{X})'\mathbf{BXC})\\ &= \mathrm{tr}((\mathbf{CAX}'\mathbf{B} + \mathbf{A}'\mathbf{C}'\mathbf{X}'\mathbf{B}')\mathrm{d}\mathbf{X}) \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}\mathrm{tr}(\mathbf{AX}^{-1}\mathbf{B}) &= -\mathbf{X}^{-1}\mathbf{BAX}^{-1} \qquad (16)\\ \text{because } \mathrm{dtr}(\mathbf{AX}^{-1}\mathbf{B}) &= -\mathrm{tr}(\mathbf{AX}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1}\mathbf{B})\\ &= -\mathrm{tr}(\mathbf{X}^{-1}\mathbf{BAX}^{-1}\mathrm{d}\mathbf{X}) \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}\mathrm{tr}(\mathbf{A}(\mathbf{X}\Sigma\mathbf{X}')^{-1}\mathbf{B}) &= -\Sigma\mathbf{X}'(\mathbf{X}\Sigma\mathbf{X}')^{-1}(\mathbf{BA} + \mathbf{A}'\mathbf{B}')(\mathbf{X}\Sigma\mathbf{X}')^{-1} \qquad (17)\\ & \text{(for symmetric } \Sigma)\\ \text{because } \mathrm{dtr}(\mathbf{A}(\mathbf{X}\Sigma\mathbf{X}')^{-1}\mathbf{B}) &= -\mathrm{tr}(\mathbf{A}(\mathbf{X}\Sigma\mathbf{X}')^{-1}((\mathrm{d}\mathbf{X})\Sigma\mathbf{X}' + \mathbf{X}\Sigma(\mathrm{d}\mathbf{X})')(\mathbf{X}\Sigma\mathbf{X}')^{-1}\mathbf{B})\\ &= -\mathrm{tr}(\Sigma\mathbf{X}'(\mathbf{X}\Sigma\mathbf{X}')^{-1}(\mathbf{BA} + \mathbf{A}'\mathbf{B}')(\mathbf{X}\Sigma\mathbf{X}')^{-1}\mathrm{d}\mathbf{X}) \end{aligned}$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}|\mathbf{X}| = |\mathbf{X}|\,\mathbf{X}^{-1} \tag{18}$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}|\mathbf{X}'\mathbf{X}| &= 2\,|\mathbf{X}'\mathbf{X}|\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' && (19)\\
\text{because } \mathrm{d}\,|\mathbf{X}'\mathbf{X}| &= |\mathbf{X}'\mathbf{X}|\,\mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathrm{d}(\mathbf{X}'\mathbf{X}))\\
&= |\mathbf{X}'\mathbf{X}|\,\mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathrm{d}\mathbf{X} + (\mathrm{d}\mathbf{X})'\mathbf{X}))\\
&= 2\,|\mathbf{X}'\mathbf{X}|\,\mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{d}\mathbf{X})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}f(\mathbf{X}\mathbf{z}) &= \mathbf{z}\,\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}f(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{X}\mathbf{z}} && (20)\\
\text{because } \mathrm{d}f(\mathbf{x}) &= (\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}f(\mathbf{x}))\mathrm{d}\mathbf{x} \quad \text{(by definition)}\\
\mathrm{d}f(\mathbf{X}\mathbf{z}) &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}f(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{X}\mathbf{z}}(\mathrm{d}\mathbf{X})\mathbf{z}\\
&= \mathrm{tr}(\mathbf{z}\,\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}f(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{X}\mathbf{z}}\mathrm{d}\mathbf{X})
\end{aligned}$$

Constraints Sometimes we want to take the derivative of a function whose argument must be symmetric. In this case, $\mathrm{d}\mathbf{X}$ must be symmetric, so we get

$$\mathrm{d}y(\mathbf{X}) = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X}) \Rightarrow \frac{\mathrm{d}y(\mathbf{X})}{\mathrm{d}\mathbf{X}} = (\mathbf{A}+\mathbf{A}') - (\mathbf{A}\circ\mathbf{I}) \tag{21}$$

where $\mathbf{A}\circ\mathbf{I}$ is simply $\mathbf{A}$ with off-diagonal elements set to zero. The reader can check this by expanding $\mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X})$ and merging identical elements of $\mathrm{d}\mathbf{X}$. An example of this rule is:

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma}\log|\Sigma| = 2\Sigma^{-1} - (\Sigma^{-1}\circ\mathbf{I}) \tag{22}$$

when Σ must be symmetric. This is usually easier than taking an unconstrained derivative and then using Lagrange multipliers to enforce symmetry.

Similarly, if $\mathbf{X}$ must be diagonal, then so must $\mathrm{d}\mathbf{X}$, and we get

$$\mathrm{d}y(\mathbf{X}) = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X}) \Rightarrow \frac{\mathrm{d}y(\mathbf{X})}{\mathrm{d}\mathbf{X}} = (\mathbf{A}\circ\mathbf{I}) \tag{23}$$

Example: Principal Component Analysis Suppose we want to represent the zero-mean random vector $\mathbf{x}$ as one random variable a times a constant unit vector $\mathbf{v}$. This is useful for compression or noise removal. Once we choose $\mathbf{v}$, the optimal choice for a is $\mathbf{v}'\mathbf{x}$, but what is the best $\mathbf{v}$? In other words, what $\mathbf{v}$ minimizes $\mathrm{E}[(\mathbf{x}-a\mathbf{v})'(\mathbf{x}-a\mathbf{v})]$, when a is chosen optimally for each $\mathbf{x}$?

Let $\Sigma = \mathrm{E}[\mathbf{x}\mathbf{x}']$. We want to maximize

$$f(\mathbf{v}) = \mathbf{v}'\Sigma\mathbf{v} - \lambda(\mathbf{v}'\mathbf{v} - 1)$$

where λ is a Lagrange multiplier. Taking derivatives gives

$$\nabla f(\mathbf{v}) = 2\Sigma\mathbf{v} - 2\lambda\mathbf{v}$$

so the gradient is zero at any eigenvector of Σ. (Recall that the gradient is the transpose of the derivative.) If $\mathbf{v}$ is an eigenvector then $f(\mathbf{v}) = \lambda$ so the maximum is attained when $\mathbf{v}$ has the largest eigenvalue.

Example: Blind source separation Suppose we have k microphones listening to k overlapped sound sources. Can we recover the individual sources? More generally, suppose we've observed data $\mathbf{x}$ generated by the function $\mathbf{x} = \mathbf{A}^{-1}\mathbf{u}$ where $\mathbf{u}$ is a set of independent hidden causes and $\mathbf{A}$ is an unknown square mixing matrix. Assume each u_i is distributed according to some density $f_i(w_i)$ with unknown parameter w_i. We want to find the mixing matrix which maximizes the likelihood of the data, so that we can then recover the hidden causes.

$$\mathbf{p}(\mathbf{u}|\mathbf{w}) = \prod_i f_i(u_i|w_i) \tag{24}$$

$$\mathbf{p}(\mathbf{x}|\mathbf{A},\mathbf{w}) = |\mathbf{A}| \prod_i f_i(u_i|w_i) \tag{25}$$

$$\log \mathbf{p}(\mathbf{x}|\mathbf{A},\mathbf{w}) = \log|\mathbf{A}| + \sum_i \log f_i(u_i|w_i) \tag{26}$$

$$\nabla_{\mathbf{A}} \log \mathbf{p}(\mathbf{x}|\mathbf{A},\mathbf{w}) = \mathbf{A}^{-\mathrm{T}} + \left[\frac{\nabla_{u_i} f_i(u_i|w_i)}{f_i(u_i|w_i)}\right]\mathbf{x}' \tag{27}$$

$$\nabla_{w_i} \log \mathbf{p}(\mathbf{x}|\mathbf{A},\mathbf{w}) = \frac{\nabla_{w_i} f_i(u_i|w_i)}{f_i(u_i|w_i)} \tag{28}$$

These equations can be used in a gradient-based optimization to find $\mathbf{A}$ and $\mathbf{w}$. This approach comes from Pearlmutter and Parra (1996).

Example: Gaussian covariance Suppose we've observed vectors $\mathbf{x}_i$ independently sampled from a zero-mean Gaussian distribution, i.e.

$$\mathbf{p}(\mathbf{x}|\Sigma) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x})$$

We want to determine the most likely covariance matrix Σ, keeping in mind that the solution must be symmetric. Maximizing the log-likelihood gives:

$$\sum_i \log \mathbf{p}(\mathbf{x}_i|\Sigma) = \sum_i(-(d/2)\log(2\pi) - (1/2)\log|\Sigma| - \frac{1}{2}\mathbf{x}_i{}'\Sigma^{-1}\mathbf{x}_i) \tag{29}$$

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma} = \sum_i(-\Sigma^{-1} + \Sigma^{-1}\mathbf{x}_i\mathbf{x}_i{}'\Sigma^{-1} - \frac{1}{2}((-\Sigma^{-1} + \Sigma^{-1}\mathbf{x}_i\mathbf{x}_i{}'\Sigma^{-1}) \circ \mathbf{I})) = 0 \tag{30}$$

$$\Sigma = (\sum_i \mathbf{x}_i\mathbf{x}_i{}')/(\sum_i 1) \tag{31}$$

2 Kronecker product and vec

The Kronecker product (Lancaster and Tismenetsky, 1985) (Horn and Johnson, 1991) is

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix} \tag{32}$$

which, like ordinary matrix product, is associative and distributive but not commutative.

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}' \tag{33}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \tag{34}$$

which implies $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

If $\mathbf{A}$ and $\mathbf{B}$ are square, then the eigenvectors and eigenvalues of $(\mathbf{A} \otimes \mathbf{B})$ are given by

$$\mathbf{A} \otimes \mathbf{B} = \mathbf{V}_A \Lambda_A {\mathbf{V}_A}^{-1} \otimes \mathbf{V}_B \Lambda_B {\mathbf{V}_B}^{-1} = (\mathbf{V}_A \otimes \mathbf{V}_B)(\Lambda_A \otimes \Lambda_B)(\mathbf{V}_A \otimes \mathbf{V}_B)^{-1} \tag{35}$$

which implies

$$\mathrm{rank}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{rank}(\mathbf{A})\mathrm{rank}(\mathbf{B}) \tag{36}$$

$$\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\Lambda_A \otimes \Lambda_B) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B}) \tag{37}$$

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^{\mathrm{rank}(\mathbf{B})} |\mathbf{B}|^{\mathrm{rank}(\mathbf{A})} \tag{38}$$

Define $\mathrm{vec}(\mathbf{A})$ to be the stacked columns of $\mathbf{A}$:

$$\mathrm{vec}(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix} \tag{39}$$

Then the main result is

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\mathrm{vec}(\mathbf{B}) \tag{40}$$

Example The Lyapunov equation is

$$\mathbf{AX} + \mathbf{XB} = \mathbf{C} \tag{41}$$

$$(\mathbf{I} \otimes \mathbf{A} + \mathbf{B}' \otimes \mathbf{I})\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{C}) \tag{42}$$

which can be solved for $\mathrm{vec}(\mathbf{X})$.

The other properties of vec will be presented in the context of vec-transpose.

3 Vec-transpose

Vec-transpose is a new operator that generalizes vec and transpose. It is essential for expressing derivatives of Kronecker products and is also useful for expressing multilinear forms. It was called "vector transposition" by Marimont and Wandell (1992).

Define

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \end{bmatrix}^{(2)} = \begin{bmatrix} a_{11} & a_{31} & a_{51} \\ a_{21} & a_{41} & a_{61} \\ a_{12} & a_{32} & a_{52} \\ a_{22} & a_{42} & a_{62} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \end{bmatrix}^{(3)} = \begin{bmatrix} a_{11} & a_{41} \\ a_{21} & a_{51} \\ a_{31} & a_{61} \\ a_{12} & a_{42} \\ a_{22} & a_{52} \\ a_{32} & a_{62} \end{bmatrix}$$

and similarly $\mathbf{A}^{(p)}$ for any integer p dividing rows($\mathbf{A}$).

The basic properties are:

$$\begin{aligned} \mathbf{A}^{(1)} &= \mathbf{A}' && (43) \\ \mathbf{A}^{(\mathrm{rows}(\mathbf{A}))} &= \mathrm{vec}(\mathbf{A}) && (44) \\ \mathrm{vec}(\mathbf{A})^{(r)} &= \mathrm{reshape}(\mathbf{A}, r, c) \ \ \text{(in Matlab notation)} && (45) \\ \mathbf{A}^{(p)(p)} &= \mathbf{A} && (46) \\ (\alpha\mathbf{A} + \beta\mathbf{B})^{(p)} &= \alpha\mathbf{A}^{(p)} + \beta\mathbf{B}^{(p)} && (47) \end{aligned}$$

We can freely apply vec-transpose inside of a trace expression:

$$\mathrm{tr}(\mathbf{A}'\mathbf{B}) = \mathrm{tr}((\mathbf{A}^{(p)})'\mathbf{B}^{(q)}) \tag{48}$$

assuming conformability. This generalizes $\mathrm{tr}(\mathbf{A}'\mathbf{B}) = \mathrm{vec}(\mathbf{A})'\mathrm{vec}(\mathbf{B})$ as well as $\mathrm{tr}(\mathbf{A}'\mathbf{B}) = \mathrm{tr}(\mathbf{A}\mathbf{B}')$. In fact,

$$\mathrm{tr}(\mathbf{A}'\mathbf{B}) = \mathrm{tr}((\mathbf{A}^{\star})'\mathbf{B}^{\star}) \tag{49}$$

for any operator * that rearranges elements.

We can generalize equations 33 and 40:

$$(\mathbf{A} \otimes \mathbf{B})^{(p)} = \mathbf{A}' \otimes \mathbf{B}^{(p)} \tag{50}$$

$$
\begin{aligned}
((\mathbf{D}' \otimes \mathbf{A})\mathbf{BC})^{(p)} &= (\mathbf{C}' \otimes \mathbf{A})\mathbf{B}^{(p)}\mathbf{D} \\
\text{where } p &= \operatorname{cols}(\mathbf{A})
\end{aligned} \tag{51}
$$

We now have the tools to express the derivative of a Kronecker product:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\operatorname{tr}(\mathbf{X}'(\mathbf{A} \otimes \mathbf{B})\mathbf{Y}) &= \tfrac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\operatorname{tr}((\mathbf{X}^{(p)})'((\mathbf{A} \otimes \mathbf{B})\mathbf{Y})^{(p)}) \\
&= \tfrac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\operatorname{tr}((\mathbf{X}^{(p)})'(\mathbf{I} \otimes \mathbf{B})\mathbf{Y}^{(p)}\mathbf{A}') \\
&= (\mathbf{Y}^{(p)})'(\mathbf{I} \otimes \mathbf{B}')\mathbf{X}^{(p)}
\end{aligned} \tag{52}
$$

where p is uniquely defined by conformability to be $\operatorname{cols}(\mathbf{B})$.

Equation 51 gives us the following rule for pulling a matrix out of nested vec-transposes:

$$
((\mathbf{AB})^{(p)}\mathbf{C})^{(p)} = (\mathbf{C}' \otimes \mathbf{I})\mathbf{AB} = (\mathbf{A}^{(p)}\mathbf{C})^{(p)}\mathbf{B} \tag{53}
$$

This formula is useful in fitting multilinear forms (see the next section). Unlike regular transpose, it is not true in general that $(\mathbf{B}^{(p)}\mathbf{C})^{(p)} = \mathbf{C}^{(p)}\mathbf{B}$, as we can see by setting $\mathbf{A} = \mathbf{I}$ in (53).

4 Multilinear forms

Multilinear statistical models are more expressive than linear models yet still easy to use. A multilinear form $f(\mathbf{x}, \mathbf{y}, ..., \mathbf{z})$ is linear in each component separately, i.e.

$$
f(..., \alpha\mathbf{y}_1 + \beta\mathbf{y}_2, ...) = \alpha f(..., \mathbf{y}_1, ...) + \beta f(..., \mathbf{y}_2, ...)
$$

For example, face images can be modeled as linear in identity and linear in lighting conditions, yielding a bilinear model (Tenenbaum and Freeman, 1997). Another example is colored objects under colored light.

Just as every bilinear form can be written as $\mathbf{x}'\mathbf{G}\mathbf{y} = (\mathbf{y}' \otimes \mathbf{x}')\operatorname{vec}(\mathbf{G})$, every multilinear form can be written as $(\mathbf{z}' \otimes ... \otimes \mathbf{y}' \otimes ... \otimes \mathbf{x}')\operatorname{vec}(\mathbf{G})$ (Magnus and Neudecker, 1988) (Prasolov, 1991) (Dodson and Poston, 1991). $\mathbf{G}$ is the tensor defining the form.

For example, the trilinear form

$$
\sum_{ijk} G_{ijk} x_i y_j z_k = (\mathbf{z}' \otimes \mathbf{y}' \otimes \mathbf{x}')\operatorname{vec}(\mathbf{G}) = (\mathbf{y}' \otimes \mathbf{x}')\mathbf{Gz} = \mathbf{x}'(\mathbf{Gz})^{(p)}\mathbf{y}
$$

Thus we can express a multilinear form either with tensor products or with vec-transpose. Matlab users have often used this kind of reshaping to manipulate higher-dimensional objects.

Just as $\mathbf{A}\mathbf{x} = [\,\mathbf{c}_1 \quad \cdots \quad \mathbf{c}_n\,]\,\mathbf{x} = \sum_i \mathbf{c}_i x_i$, a linear combination of vectors, we can write

$$(\mathbf{A}\mathbf{x})^{(p)} = ([\,\mathbf{C}_1{}^{(p)} \quad \cdots \quad \mathbf{C}_n{}^{(p)}\,]\,\mathbf{x})^{(p)} = (\sum_i \mathbf{C}_i{}^{(p)} x_i)^{(p)} = \sum_i \mathbf{C}_i x_i$$

which is a linear combination of matrices.

Therefore, we can think of $\mathbf{G}$ in the trilinear form as a three-dimensional stack of matrices. The formula $\mathbf{x}'(\mathbf{G}\mathbf{z})^{(p)}\mathbf{y}$ says to first combine the stack according to $\mathbf{z}$, then combine the columns according to $\mathbf{y}$, and finally to combine the elements according to $\mathbf{x}$.

The vector-valued bilinear form is:

$$(\mathbf{y}' \otimes \mathbf{x}' \otimes \mathbf{I})\mathrm{vec}(\mathbf{G}) = (\mathbf{G}\mathbf{y})^{(p)}\mathbf{x}$$

which is the same as the scalar-valued trilinear form, except that the three-dimensional tensor $\mathbf{G}$ is only being summed in two dimensions. This form was used in Marimont and Wandell (1992) and subsequently by Tenenbaum and Freeman (1997). To fit this model to data, note that by (53) we have

$$((\mathbf{G}\mathbf{y})^{(p)}\mathbf{x})^{(p)} = (\mathbf{G}^{(p)}\mathbf{x})^{(p)}\mathbf{y}$$

so given an observation and the value of $\mathbf{y}$ we can solve for $\mathbf{x}$ and vice-versa, by applying vec-transpose to the observation. Therefore we can iterate from an initial guess until we reach a fixed point. This method generalizes to any multilinear form. Compare this to the complex algorithm without vec-transpose given in Magnus and Neudecker (1988).

What if $\mathbf{G}$ must be learned as well as $\mathbf{x}$ and $\mathbf{y}$? In this case, we need an entire observation matrix $\mathbf{D} = (\mathbf{G}\mathbf{Y})^{(p)}\mathbf{X}$. Without loss of generality, we can assume that $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal matrices, since $\mathbf{G}$ can always be chosen to make this so. (This can be proven with a polar decomposition of $\mathbf{X}$ and $\mathbf{Y}$.) Therefore, if we know $\mathbf{Y}$, we can solve for $\mathbf{X}$ by singular-value decomposition of $\mathbf{D}$, and vice-versa with $\mathbf{D}^{(p)}$. Once $\mathbf{X}$ and $\mathbf{Y}$ have settled, it is easy to solve for vec($\mathbf{G}$). This algorithm comes from Marimont and Wandell (1992).

5 Hadamard product and diag

The Hadamard product is simply the product of corresponding elements:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \circ \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{bmatrix} \tag{54}$$

Schur's product theorem (Horn and Johnson, 1991) says

$$\mathbf{A} \geq 0, \mathbf{B} \geq 0 \Rightarrow \mathbf{A} \circ \mathbf{B} \geq 0 \tag{55}$$

This is also true for Kronecker product (by (35)), but not for regular matrix product. To prove it for Hadamard product, define a random vector $\mathbf{z} = \mathbf{x} \circ \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are independent random vectors with covariance Σ_x and Σ_y. Then the covariance of $\mathbf{z}$ can be shown to be $\Sigma_x \circ \Sigma_y$. Since every covariance matrix is nonnegative definite, the theorem follows.

Define

$$\operatorname{diag}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = \begin{bmatrix} x_1 & 0 \\ 0 & x_2 \end{bmatrix} \tag{56}$$

$$\operatorname{diag}^{-1}(\begin{bmatrix} x_1 & a \\ b & x_2 \end{bmatrix}) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{57}$$

diag^{-1} is a kind of pseudoinverse because

$$\operatorname{diag}^{-1}(\operatorname{diag}(\mathbf{x})) = \mathbf{x} \tag{58}$$

but

$$\operatorname{diag}(\operatorname{diag}^{-1}(\mathbf{D})) = \mathbf{D} \tag{59}$$

only for diagonal $\mathbf{D}$.

The basic properties are:

$$\begin{aligned}
\operatorname{diag}(\alpha\mathbf{x} + \beta\mathbf{y}) &= \alpha\operatorname{diag}(\mathbf{x}) + \beta\operatorname{diag}(\mathbf{y}) && (60)\\
\operatorname{diag}(\mathbf{x} \circ \mathbf{y}) &= \operatorname{diag}(\mathbf{x}) \circ \operatorname{diag}(\mathbf{y}) && (61)\\
\operatorname{diag}(\mathbf{x} \otimes \mathbf{y}) &= \operatorname{diag}(\mathbf{x}) \otimes \operatorname{diag}(\mathbf{y}) && (62)\\
\operatorname{diag}^{-1}(\mathbf{A} \circ \mathbf{B}) &= \operatorname{diag}^{-1}(\mathbf{A}) \circ \operatorname{diag}^{-1}(\mathbf{B}) && (63)\\
\operatorname{diag}^{-1}(\mathbf{A} \otimes \mathbf{B}) &= \operatorname{diag}^{-1}(\mathbf{A}) \otimes \operatorname{diag}^{-1}(\mathbf{B}) && (64)\\
\operatorname{diag}^{-1}((\mathbf{A} \circ \mathbf{B})\mathbf{C}') &= \operatorname{diag}^{-1}(\mathbf{A}(\mathbf{B} \circ \mathbf{C})') = \operatorname{diag}^{-1}(\mathbf{B}(\mathbf{A} \circ \mathbf{C})') && (65)\\
\operatorname{vec}(\mathbf{A} \circ \mathbf{B}) &= \operatorname{diag}(\operatorname{vec}(\mathbf{A}))\operatorname{vec}(\mathbf{B}) && (66)
\end{aligned}$$

Equation 66 can be used to remove all Hadamard products from an expression.

Another way to remove Hadamard products is with the diag^{-1} matrix, which is the unique matrix $\mathbf{R}_n$ satisfying

$$\mathbf{R}_n \text{vec}(\mathbf{A}) = \text{diag}^{-1}(\mathbf{A}) \tag{67}$$

where $\mathbf{A}$ is n by n. $\mathbf{R}_n$ is an $n \times n^2$ matrix with orthogonal rows (each row picks out one element of $\mathbf{A}$). Some useful properties are:

$$\begin{aligned}
\mathbf{R}^+ &= \mathbf{R}' \qquad \text{(pseudo-inverse is the transpose)} & (68)\\
\mathbf{R}'\mathbf{x} &= \text{vec}(\text{diag}(\mathbf{x})) & (69)\\
\mathbf{A} \circ \mathbf{B} &= \mathbf{R}(\mathbf{A} \otimes \mathbf{B})\mathbf{R}' & (70)
\end{aligned}$$

These properties cause Hadamard product and diag to have a kind of duality with Kronecker product and vec, as seen in the following table:

$$\begin{aligned}
\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\text{vec}(\mathbf{B}) \qquad & \text{diag}^{-1}(\mathbf{A}\text{diag}(\mathbf{x})\mathbf{C}) = (\mathbf{C}' \circ \mathbf{A})\mathbf{x} & (71)\\
& \text{diag}(\mathbf{x} \circ \mathbf{y} \circ \mathbf{z}) = (\mathbf{z}' \otimes \mathbf{x}) \circ \text{diag}(\mathbf{y}) & (72)
\end{aligned}$$

A useful special case of (71) is

$$\text{diag}^{-1}(\mathbf{AB}') = (\mathbf{A} \circ \mathbf{B})\mathbf{1} \tag{73}$$

If we factor $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$ then by (71),

$$\text{diag}^{-1}(\mathbf{A}) = (\mathbf{V} \circ \mathbf{V}^{-\mathrm{T}})\text{diag}^{-1}(\Lambda) \tag{74}$$

which relates the diagonal of a matrix to its eigenvalues. Many facts about the matrix $\mathbf{V} \circ \mathbf{V}^{-\mathrm{T}}$ can be found in Horn and Johnson (1991).

Many identities for $\text{diag}^{-1}(\mathbf{A})$ also apply to $\text{tr}(\mathbf{A})$, because

$$\text{tr}(\mathbf{A}) = \mathbf{1}'\text{diag}^{-1}(\mathbf{A}) \tag{75}$$

where $\mathbf{1}$ is a column vector of ones. For example, $\text{tr}(\mathbf{A}\text{diag}(\mathbf{x})\mathbf{C}) = \mathbf{1}'(\mathbf{C}' \circ \mathbf{A})\mathbf{x} = \text{diag}^{-1}(\mathbf{CA})\mathbf{x}$.

Similarly to (48) we have

$$\mathbf{x}'\mathbf{y} = \text{tr}(\text{diag}(\mathbf{x})'\text{diag}(\mathbf{y})) \tag{76}$$

which allows us to compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\mathbf{x}'(\mathbf{A} \circ \mathbf{B})\mathbf{y} = {} & \tfrac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\text{tr}(\text{diag}(\mathbf{x})'\mathbf{B}\text{diag}(\mathbf{y})\mathbf{A}') = \\
= {} & \text{diag}(\mathbf{y})\mathbf{B}'\text{diag}(\mathbf{x}) = \mathbf{B}' \circ \mathbf{y}\mathbf{x}' & (77)
\end{aligned}$$

cf (52).

6 Inverting partitioned matrices

If we partition $\mathbf{P}$ into $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ then the Schur complement of $\mathbf{A}$ in $\mathbf{P}$ (Prasolov, 1991) is

$$(\mathbf{P}|\mathbf{A}) = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \tag{78}$$

Similarly,

$$\begin{aligned}
(\mathbf{P}|\mathbf{D}) &= \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} && && (79) \\
-(\mathbf{P}|\mathbf{B}) &= \mathbf{C} - \mathbf{D}\mathbf{B}^{-1}\mathbf{A} && (\text{if } \mathbf{B}^{-1} \text{ exists}) && (80) \\
-(\mathbf{P}|\mathbf{C}) &= \mathbf{B} - \mathbf{A}\mathbf{C}^{-1}\mathbf{D} && (\text{if } \mathbf{C}^{-1} \text{ exists}) && (81)
\end{aligned}$$

Define

$$(\mathbf{A}|\mathbf{P}) = (\mathbf{P}|\mathbf{A})^{-1} \tag{82}$$

Then the main result is

$$\mathbf{P}^{-1} = \begin{bmatrix} (\mathbf{D}|\mathbf{P}) & -(\mathbf{B}|\mathbf{P}) \\ -(\mathbf{C}|\mathbf{P}) & (\mathbf{A}|\mathbf{P}) \end{bmatrix} \tag{83}$$

This formula still holds if all inverses are replaced by pseudo-inverses. The reader may want to check this formula when $\mathbf{P}$ is a 2×2 matrix.

Since $\mathbf{P}\mathbf{P}^{-1} = \mathbf{I}$ and $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$, we know

$$\begin{aligned}
\mathbf{A}(\mathbf{B}|\mathbf{P}) &= \mathbf{B}(\mathbf{A}|\mathbf{P}) && (84) \\
\mathbf{C}(\mathbf{D}|\mathbf{P}) &= \mathbf{D}(\mathbf{C}|\mathbf{P}) && (85) \\
(\mathbf{A}|\mathbf{P})\mathbf{C} &= (\mathbf{C}|\mathbf{P})\mathbf{A} && (86) \\
(\mathbf{B}|\mathbf{P})\mathbf{D} &= (\mathbf{D}|\mathbf{P})\mathbf{B} && (87)
\end{aligned}$$

These identities define $(\mathbf{B}|\mathbf{P})$ and $(\mathbf{C}|\mathbf{P})$ when $\mathbf{B}$ and $\mathbf{C}$ are singular.
Since $(\mathbf{A}|\mathbf{P})(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) = \mathbf{I}$, we know

$$\begin{aligned}
(\mathbf{A}|\mathbf{P}) &= \mathbf{D}^{-1} + (\mathbf{A}|\mathbf{P})\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\mathbf{D}^{-1} \\
&= \mathbf{D}^{-1} + (\mathbf{C}|\mathbf{P})\mathbf{B}\mathbf{D}^{-1} \\
&= \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{D}|\mathbf{P})\mathbf{B}\mathbf{D}^{-1} && (88) \\
\text{similarly } (\mathbf{D}|\mathbf{P}) &= \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}|\mathbf{P})\mathbf{C}\mathbf{A}^{-1} && (89)
\end{aligned}$$

which are handy for rewriting (83) solely in terms of $(\mathbf{D}|\mathbf{P})$ or $(\mathbf{A}|\mathbf{P})$.

The Schur complement has the flavor of a division. Equation 84, for example, tells us that $(\mathbf{B}|\mathbf{P})(\mathbf{P}|\mathbf{A}) = \mathbf{A}^{-1}\mathbf{B}$ which is a kind of cancellation of $\mathbf{P}$. The clearest example of this is the formula for the determinant of $\mathbf{P}$:

$$|\mathbf{P}| = |(\mathbf{P}|\mathbf{A})|\,|\mathbf{A}| = |(\mathbf{P}|\mathbf{D})|\,|\mathbf{D}| \tag{90}$$

When $\mathbf{A} = \mathbf{I}$ and $\mathbf{D} = \mathbf{I}$ (not necessarily the same size) in (90) we get the handy formula

$$|\mathbf{I} + \mathbf{CB}| = |\mathbf{I} + \mathbf{BC}| \tag{91}$$

for any matrices $\mathbf{B}$ and $\mathbf{C}$.

Example Magnus and Neudecker give a formula for $\begin{bmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{B}' & \mathbf{D} & 0 \\ \mathbf{C}' & 0 & \mathbf{E} \end{bmatrix}^{-1}$ but leave the proof to the reader. This is easily handled using Schur complements. Define $\mathbf{X} = [\mathbf{B} \quad \mathbf{C}]$ and $\mathbf{Y} = \begin{bmatrix} \mathbf{D} & 0 \\ 0 & \mathbf{E} \end{bmatrix}$. Then

$$\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{Y} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{Y}|\mathbf{P}) & -(\mathbf{X}|\mathbf{P}) \\ -(\mathbf{X}|\mathbf{P})' & (\mathbf{A}|\mathbf{P}) \end{bmatrix}$$

where

$$(\mathbf{Y}|\mathbf{P}) = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}' - \mathbf{C}\mathbf{E}^{-1}\mathbf{C}')^{-1} \tag{92}$$

$$(\mathbf{X}|\mathbf{P}) = (\mathbf{Y}|\mathbf{P})\mathbf{X}\mathbf{Y}^{-1} \tag{93}$$

$$= [\,(\mathbf{Y}|\mathbf{P})\mathbf{B}\mathbf{D}^{-1} \quad (\mathbf{Y}|\mathbf{P})\mathbf{C}\mathbf{E}^{-1}\,] \tag{94}$$

$$(\mathbf{A}|\mathbf{P}) = \mathbf{Y}^{-1} + \mathbf{Y}^{-1}\mathbf{X}'(\mathbf{Y}|\mathbf{P})^{-1}\mathbf{X}\mathbf{Y}^{-1} \tag{95}$$

$$= \begin{bmatrix} \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{B}'(\mathbf{Y}|\mathbf{P})\mathbf{B}\mathbf{D}^{-1} & \mathbf{D}^{-1}\mathbf{B}'(\mathbf{Y}|\mathbf{P})\mathbf{C}\mathbf{E}^{-1} \\ \mathbf{E}^{-1}\mathbf{C}'(\mathbf{Y}|\mathbf{P})\mathbf{B}\mathbf{D}^{-1} & \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{C}'(\mathbf{Y}|\mathbf{P})\mathbf{C}\mathbf{E}^{-1} \end{bmatrix} \tag{96}$$

Example: Conditioning a Gaussian density Suppose we partition a zero-mean Gaussian random vector as $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ with partitioned covariance matrix $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{yx} & \mathbf{K}_{yy} \end{bmatrix}$. The distribution of $\mathbf{x}$ conditioned on $\mathbf{y}$ is $\mathbf{p}(\mathbf{x}|\mathbf{y}) = \mathbf{p}(\mathbf{x},\mathbf{y})/\mathbf{p}(\mathbf{y})$ which is proportional to the joint distribution. Knowing that this conditional is also Gaussian, we can immediately derive its mean and variance by dropping the terms in the joint distribution that depend on $\mathbf{y}$:

$$\begin{aligned}
\mathbf{p}(\mathbf{x}|\mathbf{y}) &\propto \exp(-\frac{1}{2}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}' \begin{bmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{yx} & \mathbf{K}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}) \\
&\propto \exp(-\frac{1}{2}\mathbf{x}'(\mathbf{K}_{yy}|\mathbf{K})\mathbf{x} + \mathbf{x}'(\mathbf{K}_{xy}|\mathbf{K})\mathbf{y}) \\
&= \exp(-\frac{1}{2}\mathbf{x}'(\mathbf{K}_{yy}|\mathbf{K})\mathbf{x} + \mathbf{x}'(\mathbf{K}_{yy}|\mathbf{K})\mathbf{m}) \qquad \text{(defining } \mathbf{m}\text{)} \\
&\propto \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m})'(\mathbf{K}_{yy}|\mathbf{K})(\mathbf{x} - \mathbf{m}))
\end{aligned}$$

where the conditional mean $\mathbf{m} = (\mathbf{K}|\mathbf{K}_{yy})(\mathbf{K}_{xy}|\mathbf{K})\mathbf{y} = \mathbf{K}_{xy}\mathbf{K}_{yy}^{-1}\mathbf{y}$. The conditional variance is therefore $(\mathbf{K}|\mathbf{K}_{yy}) = \mathbf{K}_{xx} - \mathbf{K}_{xy}\mathbf{K}_{yy}^{-1}\mathbf{K}_{yx}$.

7 Polar decomposition

Suppose a set of points $\mathbf{B}$ has been subjected to an unknown rotation and then jittered by white Gaussian noise to give a new set of points $\mathbf{A}$. What is the most likely rotation? More generally, what unitary matrix minimizes $f(\mathbf{U}) = \mathrm{tr}((\mathbf{A}-\mathbf{UB})'(\mathbf{A}-\mathbf{UB}))$?

Expanding $f(\mathbf{U})$ gives

$$f(\mathbf{U}) = \mathrm{tr}(\mathbf{A}'\mathbf{A}) - 2\mathrm{tr}(\mathbf{A}'\mathbf{UB}) + \mathrm{tr}(\mathbf{B}'\mathbf{B})$$

so the problem reduces to maximizing $\mathrm{tr}(\mathbf{A}'\mathbf{UB}) = \mathrm{tr}(\mathbf{UBA}')$. Define unitary $\mathbf{V}$ and $\mathbf{W}$ and positive diagonal $\mathbf{S}$ so that $\mathbf{BA}' = \mathbf{VSW}'$. This is the singular value decomposition of $\mathbf{BA}'$. Then

$$\mathrm{tr}(\mathbf{UBA}') = \mathrm{tr}(\mathbf{UVSW}') = \mathrm{tr}(\mathbf{W}'\mathbf{UVS}) \stackrel{\mathrm{def}}{=} \mathrm{tr}(\mathbf{XS})$$

where $\mathbf{X}$ is also unitary. Since $\mathbf{S}$ is diagonal,

$$\mathrm{tr}(\mathbf{XS}) = \sum_i s_{ii}x_{ii}$$

which is maximum when $x_{ii} = 1$; that is, $\mathbf{X} = \mathbf{I}$. Therefore $\mathbf{U} = \mathbf{WV}'$ is the desired solution.

If $\mathbf{B} = \mathbf{I}$, this solution minimizes $(\mathbf{A}-\mathbf{U})$, i.e. it is the closest unitary matrix to an arbitrary matrix $\mathbf{A}$. This $\mathbf{U}$ has the property that there exists a positive definite $\mathbf{P}$ such that $\mathbf{A} = \mathbf{PU}$. These two matrices are called the polar decomposition of $\mathbf{A}$: $\mathbf{U}$ is the rotation and $\mathbf{P}$ is the magnitude, exactly analogous to the decomposition of a complex number.

Scott and Longuet-Higgins (1991) used this technique to match columns in $\mathbf{A}$ with those in $\mathbf{B}$. Matching assumes that $\mathbf{U}$ is a permutation matrix, but finding $\mathbf{U}$ in this case is difficult. So they first found the optimal unitary matrix and obtained a permutation from it.

8 Hessians

The Hessian matrix is a matrix of second derivatives. The Hessian of a scalar function with respect to a vector argument is

$$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'} = \left[\frac{\partial y}{\partial x_i \partial x_j}\right]$$

This section is based on Magnus and Neudecker (1988).

The Hessian is the derivative of the first derivative. The first derivative $\mathbf{a}(\mathbf{x}) = \left[\frac{\partial y}{\partial x_j}\right]$ is a row vector function of $\mathbf{x}$, and the derivative of this function with respect to $\mathbf{x}'$ is a matrix

$$\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}\mathbf{x}'} = \left[\frac{\partial a_j}{\partial x_i}\right] = \left[\frac{\partial y}{\partial x_i \partial x_j}\right]$$

The Hessian can also be defined by the Taylor expansion of y:

$$y(\mathbf{x} + \mathrm{d}\mathbf{x}) = y(\mathbf{x}) + \mathbf{a}'\mathrm{d}\mathbf{x} + \frac{1}{2}\mathrm{d}\mathbf{x}'\mathbf{H}\mathrm{d}\mathbf{x} + \text{(higher order terms)} \tag{97}$$

The matrix $\mathbf{H}$ is the Hessian, as you can check by setting all but two components of $\mathrm{d}\mathbf{x}$ to zero.

The Hessian can be computed in three steps:

1. Compute the first differential
2. Compute the differential of the first differential
3. Massage the result into canonical form

The only new differential rule we need is:

$$\mathrm{d}(\mathrm{d}\mathbf{x}) \quad = \quad 0 \tag{98}$$

because $\mathrm{d}\mathbf{x}$ is not a function of $\mathbf{x}$.

The second differential has three canonical forms:

$\mathrm{d}^2 y = h(\mathrm{d}x)^2$
$\mathrm{d}^2 y = \mathrm{d}\mathbf{x}'\mathbf{H}\mathrm{d}\mathbf{x}$
$\mathrm{d}^2 y = \mathrm{dvec}(\mathbf{X})'\mathbf{H}\mathrm{dvec}(\mathbf{X})$

where $\mathbf{H}$ must be symmetric.

Some of these forms require rewriting the differential in terms of $\mathrm{dvec}(\mathbf{X})$, which can be tricky. Equation 48 is particularly helpful for introducing vec into an expression. To eliminate terms like $\mathrm{vec}(\mathbf{X}')$, Magnus and Neudecker (1988) define the *commutation matrix* $\mathbf{K}_{nm}$, which is the permutation matrix satisfying

$$\mathbf{K}_{nm}\mathrm{vec}(\mathbf{X}') = \mathrm{vec}(\mathbf{X}) \tag{99}$$

where $\mathbf{X}$ is n by m. Also,

$$\mathbf{K}'_{nm} = \mathbf{K}^{-1}_{nm} = \mathbf{K}_{mn} \tag{100}$$

Examples:

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'}\mathrm{tr}(\mathbf{X}'\mathbf{X}) &= 2\mathbf{I}_{n^2} && (101)\\
\text{because } \mathrm{dtr}(\mathbf{X}'\mathbf{X}) &= 2\mathrm{tr}(\mathbf{X}'\mathrm{d}\mathbf{X})\\
\mathrm{d}^2\mathrm{tr}(\mathbf{X}'\mathbf{X}) &= 2\mathrm{tr}(\mathrm{d}\mathbf{X}'\mathrm{d}\mathbf{X}) = 2\mathrm{vec}(\mathrm{d}\mathbf{X})'\mathrm{vec}(\mathrm{d}\mathbf{X})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'}\mathrm{tr}(\mathbf{X}'\mathbf{A}\mathbf{X}) &= 2\mathbf{I}_n \otimes \mathbf{A} && (102)\\
\text{because } \mathrm{dtr}(\mathbf{X}'\mathbf{A}\mathbf{X}) &= 2\mathrm{tr}(\mathbf{X}'\mathbf{A}\mathrm{d}\mathbf{X})\\
\mathrm{d}^2\mathrm{tr}(\mathbf{X}'\mathbf{A}\mathbf{X}) &= 2\mathrm{tr}(\mathrm{d}\mathbf{X}'\mathbf{A}\mathrm{d}\mathbf{X}) && (103)\\
&= 2\mathrm{vec}(\mathrm{d}\mathbf{X})'\mathrm{vec}(\mathbf{A}\mathrm{d}\mathbf{X}) && (104)\\
&= 2\mathrm{vec}(\mathrm{d}\mathbf{X})'(\mathbf{I}\otimes\mathbf{A})\mathrm{vec}(\mathrm{d}\mathbf{X}) && (105)
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'}\mathrm{tr}(\mathbf{X}^2) &= 2\mathbf{K}_{nn} && (106)\\
\text{because } \mathrm{dtr}(\mathbf{X}^2) &= 2\mathrm{tr}(\mathbf{X}\mathrm{d}\mathbf{X})\\
\mathrm{d}^2\mathrm{tr}(\mathbf{X}^2) &= 2\mathrm{tr}((\mathrm{d}\mathbf{X})^2) = 2\mathrm{vec}(\mathrm{d}\mathbf{X}')'\mathrm{vec}(\mathrm{d}\mathbf{X})\\
&= 2\mathrm{vec}(\mathrm{d}\mathbf{X})'\mathbf{K}_{nn}\mathrm{vec}(\mathrm{d}\mathbf{X})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'}\log|\mathbf{X}| &= -\mathbf{K}_{nn}(\mathbf{X}^{-\mathrm{T}}\otimes\mathbf{X}^{-1}) && (107)\\
\text{because } \mathrm{d}\log|\mathbf{X}| &= \mathrm{tr}(\mathbf{X}^{-1}\mathrm{d}\mathbf{X})\\
\mathrm{d}^2\log|\mathbf{X}| &= -\mathrm{tr}(\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1}\mathrm{d}\mathbf{X})\\
&= -\mathrm{vec}(\mathrm{d}\mathbf{X}')'\mathrm{vec}(\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1})\\
&= -\mathrm{vec}(\mathrm{d}\mathbf{X}')'(\mathbf{X}^{-\mathrm{T}}\otimes\mathbf{X}^{-1})\mathrm{vec}(\mathrm{d}\mathbf{X})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'}\mathrm{tr}(\mathbf{X}^{-1}) &= \mathbf{K}_{nn}(\mathbf{X}^{-2\mathrm{T}}\otimes\mathbf{X}^{-1} + \mathbf{X}^{-2}\otimes\mathbf{X}^{-\mathrm{T}}) && (108)\\
\text{because } \mathrm{dtr}(\mathbf{X}^{-1}) &= -\mathrm{tr}(\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-1})\\
\mathrm{d}^2\mathrm{tr}(\mathbf{X}^{-1}) &= 2\mathrm{tr}((\mathrm{d}\mathbf{X})\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-2})
\end{aligned}$$

$$
\begin{aligned}
&= 2\mathrm{vec}(\mathrm{d}\mathbf{X}')'\mathrm{vec}(\mathbf{X}^{-1}(\mathrm{d}\mathbf{X})\mathbf{X}^{-2}) \\
&= 2\mathrm{vec}(\mathrm{d}\mathbf{X}')'(\mathbf{X}^{-2\mathrm{T}} \otimes \mathbf{X}^{-1})\mathrm{vec}(\mathrm{d}\mathbf{X}) \\
&= \mathrm{vec}(\mathrm{d}\mathbf{X}')'(\mathbf{X}^{-2\mathrm{T}} \otimes \mathbf{X}^{-1} + \mathbf{X}^{-2} \otimes \mathbf{X}^{-\mathrm{T}})\mathrm{vec}(\mathrm{d}\mathbf{X}) \text{ (for symmetry)}
\end{aligned}
$$

Constraints Sometimes we want to compute the Hessian of a function whose argument must be symmetric. In this case, the conversion from canonical form is

$$
\begin{aligned}
\mathrm{d}^2 y &= \mathrm{dvec}(\mathbf{X})'\mathbf{H}\mathrm{dvec}(\mathbf{X}) \Rightarrow & (109) \\
\frac{\mathrm{d}^2 y}{\mathrm{dvec}(\mathbf{X})\mathrm{dvec}(\mathbf{X})'} &= (\mathbf{I}_{n^2} + \mathbf{K}_{nn} - \mathrm{diag}(\mathrm{vec}(\mathbf{I}_n)))\mathbf{H}(\mathbf{I}_{n^2} + \mathbf{K}_{nn} - \mathrm{diag}(\mathrm{vec}(\mathbf{I}_n))) & (110) \\
&= \mathbf{D}_n\mathbf{D}_n'\mathbf{H}\mathbf{D}_n\mathbf{D}_n' & (111)
\end{aligned}
$$

where $\mathbf{K}_{nn}$ is the commutation matrix discussed earlier and $\mathbf{D}_n$ is defined below. Since ∂x_{ij} and ∂x_{ji} are identical, this formula adds together $\frac{\partial y}{\partial x_{ij}^2}$, $\frac{\partial y}{\partial x_{ij}\partial x_{ji}}$, $\frac{\partial y}{\partial x_{ji}\partial x_{ij}}$, and $\frac{\partial y}{\partial x_{ji}^2}$, when $i \neq j$. To derive it, we make $\mathrm{d}\mathbf{X}$ symmetric by substituting $\mathrm{d}\mathbf{X} + \mathrm{d}\mathbf{X}' - (\mathrm{d}\mathbf{X} \circ \mathbf{I})$ (cf (21)) and get

$$
\begin{aligned}
\mathrm{vec}(\mathrm{d}\mathbf{X}) &\Rightarrow \mathrm{vec}(\mathrm{d}\mathbf{X} + \mathrm{d}\mathbf{X}' - (\mathbf{I} \circ \mathrm{d}\mathbf{X})) & (112) \\
&= (\mathbf{I}_{n^2} + \mathbf{K}_{nn} - \mathrm{diag}(\mathrm{vec}(\mathbf{I}_n)))\mathrm{vec}(\mathrm{d}\mathbf{X}) & (113)
\end{aligned}
$$

by (99) and (66).

However, we may not want the full Hessian, but only the submatrix corresponding to unique elements of $\mathbf{X}$. That is, we want $\frac{d^2 y}{\mathrm{dvech}(\mathbf{X})\mathrm{dvech}(\mathbf{X})'}$, where $\mathrm{vech}(\mathbf{X})$ (Searle, 1982) is $\mathrm{vec}(\mathbf{X})$ with elements above the diagonal deleted. For example,

$$
\mathrm{vec}\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix} \qquad (114)
$$

$$
\mathrm{vech}\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix} \qquad (115)
$$

To convert between $\mathrm{vec}(\mathbf{X})$ and $\mathrm{vech}(\mathbf{X})$, Magnus and Neudecker (1988) define the *duplication matrix* $\mathbf{D}_n$, which is the permutation matrix satisfying

$$
\mathbf{D}_n\mathrm{vech}(\mathbf{X}) = \mathrm{vec}(\mathbf{X}) \qquad (116)
$$

where $\mathbf{X}$ is n by n. This leads to the rule

$$
\begin{aligned}
\mathrm{d}^2 y &= \mathrm{dvec}(\mathbf{X})'\mathbf{H}\mathrm{dvec}(\mathbf{X}) & (117) \\
&= \mathrm{dvech}(\mathbf{X})'\mathbf{D}_n'\mathbf{H}\mathbf{D}_n\mathrm{dvech}(\mathbf{X}) & (118) \\
&\Rightarrow \frac{\mathrm{d}^2 y}{\mathrm{dvech}(\mathbf{X})\mathrm{dvech}(\mathbf{X})'} = \mathbf{D}_n'\mathbf{H}\mathbf{D}_n & (119)
\end{aligned}
$$

Furthermore, it can be shown that the matrix $\mathbf{I}_{n^2} + \mathbf{K}_{nn} - \mathrm{diag}(\mathrm{vec}(\mathbf{I}_n))$ above is equal to $\mathbf{D}_n\mathbf{D}_n'$.

Example: Hessian of a Gaussian likelihood Let $l(\mathbf{m}, \mathbf{V})$ be the logarithm of a Gaussian likelihood at $\mathbf{x}$:

$$l(\mathbf{m}, \mathbf{V}) = -\frac{1}{2}\log|2\pi\mathbf{V}| - \frac{1}{2}(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}) \tag{120}$$

By (107), the first term has Hessian

$$\frac{1}{2}\mathbf{D}_n\mathbf{D}_n'(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1})\mathbf{D}_n\mathbf{D}_n' \tag{121}$$

where the symmetry of $\mathbf{V}$ has been invoked.

The differential of the second term is

$$(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}\mathrm{d}\mathbf{m} + \frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}\mathrm{d}\mathbf{V}) \tag{122}$$

and the second differential is

$$\begin{aligned}
\mathrm{d}^2 &= -\mathrm{d}\mathbf{m}'\mathbf{V}^{-1}\mathrm{d}\mathbf{m} - \mathrm{tr}((\mathrm{d}\mathbf{V})\mathbf{V}^{-1}(\mathrm{d}\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}) \\
&\quad -\mathrm{tr}((\mathrm{d}\mathbf{V})\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}(\mathrm{d}\mathbf{V})\mathbf{V}^{-1}) && (123) \\
&= -\mathrm{d}\mathbf{m}'\mathbf{V}^{-1}\mathrm{d}\mathbf{m} - \mathrm{vec}(\mathrm{d}\mathbf{V})'(\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}) \otimes \mathbf{V}^{-1})\mathrm{vec}(\mathrm{d}\mathbf{m}) \\
&\quad -\mathrm{vec}(\mathrm{d}\mathbf{V})'(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1})\mathrm{vec}(\mathrm{d}\mathbf{V}) && (124)
\end{aligned}$$

So the full Hessian is

$$\begin{aligned}
\frac{\mathrm{d}^2 l(\mathbf{m}, \mathbf{V})}{\mathrm{d}\mathbf{m}\mathrm{d}\mathbf{m}'} &= -\mathbf{V}^{-1} && (125) \\
\frac{\mathrm{d}^2 l(\mathbf{m}, \mathbf{V})}{\mathrm{d}\mathbf{m}\mathrm{dvec}(\mathbf{V})'} &= -\mathbf{D}_n\mathbf{D}_n'(\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}) \otimes \mathbf{V}^{-1}) && (126) \\
\frac{\mathrm{d}^2 l(\mathbf{m}, \mathbf{V})}{\mathrm{dvec}(\mathbf{V})\mathrm{dvec}(\mathbf{V})'} &= \mathbf{D}_n\mathbf{D}_n'(\mathbf{V}^{-1} \otimes (\frac{1}{2}\mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}))\mathbf{D}_n\mathbf{D}_n' && (127)
\end{aligned}$$

And the reduced Hessian involving unique elements of $\mathbf{V}$ is

$$\begin{aligned}
\frac{\mathrm{d}^2 l(\mathbf{m}, \mathbf{V})}{\mathrm{d}\mathbf{m}\mathrm{dvech}(\mathbf{V})'} &= -\mathbf{D}_n'(\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}) \otimes \mathbf{V}^{-1}) && (128) \\
\frac{\mathrm{d}^2 l(\mathbf{m}, \mathbf{V})}{\mathrm{dvech}(\mathbf{V})\mathrm{dvech}(\mathbf{V})'} &= \mathbf{D}_n'(\mathbf{V}^{-1} \otimes (\frac{1}{2}\mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}))\mathbf{D}_n && (129)
\end{aligned}$$

Acknowledgements

Tony Jebara and Martin Szummer helped clarify the presentation.

References

[1] C. T. J. Dodson and T. Poston. *Tensor geometry: the geometric viewpoint and its uses.* Springer-Verlag, 1991.

[2] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[3] Harold Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, third edition, 1961.

[4] Peter Lancaster and Miron Tismenetsky. *The theory of matrices.* Academic Press, 1985.

[5] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics.* John Wiley & Sons, 1988.

[6] David H. Marimont and Brian A. Wandell. Linear models of surface and illuminant spectra. *Journal of the Optical Society of America*, 9(11):1905–1913, November 1992.

[7] Barak A. Pearlmutter and Lucas C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In Michael C. Mozer, Michael Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems 9*, Cambridge, MA, 1997. MIT Press. `http://www.cs.unm.edu/~bap/publications.html`.

[8] Viktor V. Prasolov. *Problems and Theorems in Linear Algebra.* American Mathematical Society, 1991.

[9] G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two images. *Proc. Royal Society of London*, B(244):21–26, 1991.

[10] Shayle R. Searle. *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York, NY, 1982.

[11] Joshua B. Tenenbaum and William T. Freeman. Separating style and content. In Michael C. Mozer, Michael Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems 9*, Cambridge, MA, 1997. MIT Press.