Formula

Finn Årup Nielsen

Informatics and Mathematical Modelling
Technical University of Denmark
and
Neurobiology Research Unit
Copenhagen University Hospital Rigshospitalet

April 23, 2010

$Revision: 1.93 $
$Date: 2010/04/23 21:06:44 $

Abstract

Mathematical formulas with special emphasis on statistical distributions. The formulas are sketchy and not necessarily complete or correct. Corrections can be sent to fn@imm.dtu.dk.

Thanks to Berkan Dulek

Contents

1 Distributions **4**
1.1 Normal distributions . . . 4
1.2 Wishart and χ^2 distributions . . . 5
1.3 Generalized matrix T, multivariate Student's t and Cauchy distributions . . . 5
1.4 Normal-inverse Wishart and normal-inverse gamma distribution . . . 6
1.5 Inverse cosh distribution or hyperbolic secant distribution . . . 6
1.6 Multinomial distribution including binomial and Bernoulli . . . 7
1.7 Dirichlet distribution including beta and uniform distributions . . . 7
1.8 Multivariate hypergeometric distribution . . . 7
1.9 Normal linear model distribution . . . 8
1.10 Gaussian mixture distributions . . . 8
1.10.1 Ad hoc variances . . . 9
1.10.2 Univariate Gaussian mixture models . . . 9
1.11 Hyperbolic secant mixture distribution . . . 9
1.12 Dirichlet-multinomial distribution . . . 9
1.12.1 Beta-binomial mixture distribution . . . 10
1.13 Gaussian ratio distribution . . . 10
1.13.1 Gaussian ratio distribution with no correlation . . . 11
1.13.2 Gaussian ratio distribution with no mean . . . 11
1.14 Finite binomial mixture distribution . . . 12
1.15 Multinomial mixture distribution . . . 12
1.16 Mixture of matrix hypergeometric distributions . . . 13

2 Statistical tests and other **14**
2.1 Correlation coefficient . . . 14
2.2 Partial correlation coefficient . . . 14
2.3 Effect sizes . . . 15
2.4 Combining mean and standard deviation . . . 15

3 Matrices **16**
3.1 Definition . . . 16
3.2 Types of matrices . . . 16
3.2.1 Monomial . . . 16
3.2.2 Semiconvergent and convergent matrices . . . 17
3.2.3 **Z**-matrices . . . 17
3.2.4 Nonnegative matrices . . . 17
3.2.5 Normal matrices . . . 18
3.2.6 Hermitian . . . 18
3.2.7 Unitary matrices . . . 18
3.3 Arithmetic operations . . . 19
3.4 Vectorization . . . 19
3.5 Function with scalar output . . . 19
3.5.1 Norms . . . 19
3.5.2 Spectral radius . . . 20

3.6 Inversion . . . 20
3.6.1 Properties . . . 20
3.6.2 Matrix inversion lemma . . . 20
3.6.3 2×2 inversion . . . 21
3.7 Determinant . . . 21
3.8 Trace . . . 21
3.9 Derivatives . . . 21
3.9.1 Determinant . . . 21
3.10 Integrals . . . 22
3.10.1 Transformation with Jacobian . . . 22
3.10.2 Dirac function . . . 22

4 Functions **23**
4.1 Gamma function . . . 23
4.2 Basis functions . . . 23
4.3 Link functions . . . 24

5 Inequalities **25**
5.1 Jensen's inequality . . . 25

6 Optimization **26**
6.1 Gradient descent . . . 26
6.2 Newton method . . . 26

7 Information theory **27**
7.1 Entropies . . . 27
7.2 Mutual information . . . 27

8 Networks **28**
8.1 Types of networks . . . 28
8.2 Graph matrices . . . 28
8.2.1 Eigenvalues and vectors . . . 29
8.3 Eccentricity, radius, diameter . . . 29
8.4 Centralization . . . 29
8.5 Spectral gap and expansion . . . 29
8.6 Bridge, hinge and articulation . . . 30

9 Receptor kinetics **31**
9.1 Distribution volume . . . 32
9.2 Binding potential . . . 32
9.3 Specific binding ratio . . . 33
9.4 Other values . . . 33

Chapter 1

Distributions

1.1 Normal distributions

The "stationary colored noise" matrix normal $\mathbf{X}(N \times P) \sim N_{N\times P}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega})$ (Triantafyllopoulos, 2002, p. 234)

$$p(\mathbf{X}|\mathbf{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = (2\pi)^{-NP/2}|\boldsymbol{\Omega}|^{-P/2}|\boldsymbol{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})^{\mathsf{T}}\right]\right\} \tag{1.1}$$

The probability density function for $\mathbf{X}(N \times P)$ for "white noise" matrix normal, i.e., N independent P-dimensional $\mathbf{x}_n$ with $\boldsymbol{\mu}^{\mathsf{T}}\mathbf{1} = \mathbf{M}$ and $\boldsymbol{\Omega} = \mathbf{I}_N$ (Mardia et al., 1979, section 2.5.5)

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-N/2} \exp\left[-\frac{1}{2}\sum_n^N(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right] \tag{1.2}$$

$$= |2\pi\boldsymbol{\Sigma}|^{-N/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathsf{T}})^{\mathsf{T}}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathsf{T}})\right\}\right] \tag{1.3}$$

with sufficient statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$ (Mardia et al., 1979, equation 4.1.20)

$$= |2\pi\boldsymbol{\Sigma}|^{-N/2} \exp\left[-\frac{N}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) - \frac{N}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right] \tag{1.4}$$

The multinormal distribution (the Gaussian distribution) of a P-dimensional $\mathbf{x} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a single case $N = 1$ (Mardia et al., 1979, equation 2.5.1)

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] \tag{1.5}$$

With an isotropic (co-)variance

$$p(\mathbf{x}|\boldsymbol{\mu}, \sigma^2) = (2\pi\sigma^2)^{-P/2} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu})}{2\sigma^2}\right] \tag{1.6}$$

Univariate $P = 1$

$$p(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] \tag{1.7}$$

Standardized central univariate, $\mu = 0$ and $\sigma = 1$

$$p(x) = (2\pi)^{-1/2} \exp\left(-\frac{x^2}{2}\right). \tag{1.8}$$

1.2 Wishart and χ^2 distributions

The central Wishart distribution for $\mathbf{M}(P \times P)$ where M can be regarded as a degree of freedom parameter (Mardia et al., 1979, equation 3.8.1) (Conradsen, 1984, section 2.5)

$$p(\mathbf{M}|\mathbf{\Sigma}, M) = \frac{1}{2^{MP/2}\pi^{P(P-1)/4}|\mathbf{\Sigma}|^{M/2}\prod_p^P \Gamma[\frac{1}{2}(M+1-p)]}|\mathbf{M}|^{(M-P-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})\right] \tag{1.9}$$

The central Wishart distribution on standard form with $\mathbf{\Sigma} = \mathbf{I}$

$$p(\mathbf{M}|M) = \frac{1}{2^{MP/2}\pi^{P(P-1)/4}\prod_p^P \Gamma[\frac{1}{2}(M+1-p)]}|\mathbf{M}|^{(M-P-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}(\mathbf{M})\right] \tag{1.10}$$

With a one dimensional $P = 1$ central Wishart distribution on standard form the $\mathbf{M} = y \sim \chi^2$ ("chi-square") distribution is attained. The sum of the square of M independently standardized normal distribution random variables is also χ^2 distributed. The χ^2 probability density function (Mardia et al., 1979, section B.3)

$$p_{\chi^2}(y|M) = \left\{2^{M/2}\Gamma(M/2)\right\}^{-1} y^{M/2-1}\exp(-y/2) \tag{1.11}$$

The cumulative distribution function is

$$P(y) = \frac{1}{\Gamma(P/2)}\gamma(y/2, P/2) \tag{1.12}$$

where γ is the lower incomplete gamma function, see section 4.1.

1.3 Generalized matrix T, multivariate Student's t and Cauchy distributions

The type I generalized matrix T distribution (Triantafyllopoulos, 2002, p. 72+), (Triantafyllopoulos, 2003)

$$p(\mathbf{T}|\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{D}) \propto \left|\mathbf{D}^{1/2}\mathbf{\Omega}\mathbf{D}^{1/2} + (\mathbf{T}-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{T}-\mathbf{M})^\mathsf{T}\right|^{-[\mathrm{tr}(\mathbf{D})/P+N+P-1]/2}, \tag{1.13}$$

where $\mathbf{D}$ is a matrix of degrees of freedom. The matrix T distribution is obtained with $\nu = \mathrm{tr}(\mathbf{D})/P +$ constant and rescaling $\mathbf{\Omega}$. With $\mathbf{X}(N \times P) \sim N_P(\mathbf{M}, \mathbf{\Sigma})$ and $\mathbf{B} \sim W_N(\mathbf{\Omega}, \nu)$ then $\mathbf{T}(N \times P) = \mathbf{B}^{-1/2}\mathbf{X}$ is a matrix distributed with the matrix T distribution (Hedibert Freitas Lopes) (Dickey, 1967) ??? (Triantafyllopoulos, 2002, p. 235)

$$p(\mathbf{T}|\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}, \nu) = \pi^{-NP/2}\prod_{p=1}^{P}\frac{\Gamma[(\nu+P-p+1)/2]}{\Gamma[(\nu-p+1)/2]}|\mathbf{\Omega}|^{-\nu/2}|\mathbf{\Sigma}|^{-N/2}\left|\mathbf{\Omega}^{-1} + (\mathbf{T}-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{T}-\mathbf{M})^\mathsf{T}\right|^{-(\nu+P)/2} \tag{1.14}$$

The multivariate t distribution is attained with $N = 1$ and $\mathbf{\Omega} = \mathbf{I} = 1$ (Mardia et al., 1979, section 2.6.5), (O'Hagan, 1994, section 9.5)

$$p_t(\mathbf{t}|\boldsymbol{\mu}, \mathbf{\Sigma}, \nu) = (\pi\nu)^{-P/2}\frac{\Gamma(\frac{\nu+P}{2})}{\Gamma(\nu/2)}\frac{|\mathbf{\Sigma}|^{-1/2}}{\left[1+\nu^{-1}(\mathbf{t}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right]^{(\nu+P)/2}} \tag{1.15}$$

where P is the dimension of $\mathbf{t}$. The multivariate Cauchy distribution is obtained with the degrees of freedom parameter $\nu = 1$

$$p_{\mathrm{Ca}}(\mathbf{t}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \pi^{-P/2}\frac{\Gamma(\frac{1+P}{2})}{\Gamma(1/2)}\frac{|\mathbf{\Sigma}|^{-1/2}}{\left[1+(\mathbf{t}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right]^{(1+P)/2}} \tag{1.16}$$

For the univariate $P = 1$ Cauchy distribution the probability density is

$$p_{\mathrm{Ca}}(t|\mu, \sigma^2) = \frac{1}{\pi\sigma} \frac{1}{1 + (\frac{t-\mu}{\sigma})^2} \tag{1.17}$$

The central standardized univariate Cauchy distribution $\mu = 0$ and $\sigma^2 = 1$

$$p_{\mathrm{Ca}}(t|0, 1) = \frac{1}{\pi} \frac{1}{1 + t^2} \tag{1.18}$$

1.4 Normal-inverse Wishart and normal-inverse gamma distribution

Posterior distribution for joint posterior with an information (natural) conjugate normal-inverse Wishart prior (Mardia et al., 1979, exercise 4.3.2)

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \boldsymbol{\phi}, \mathbf{G}, M) = p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\bar{\mathbf{x}}, \mathbf{S}, \boldsymbol{\phi}, \mathbf{G}, N, M) \tag{1.19}$$

$$\propto |\boldsymbol{\Sigma}|^{-(N+M+1)/2} \mathrm{etr}\left(-\frac{1}{2} \boldsymbol{\Sigma}^{-1} \left[N\mathbf{S} + \mathbf{G} + (\boldsymbol{\mu} - \boldsymbol{\phi})(\boldsymbol{\mu} - \boldsymbol{\phi})^{\mathsf{T}} + N(\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^{\mathsf{T}} \right] \right), \tag{1.20}$$

where $\mathrm{etr}(\cdot)$ is the exponential of the trace: $\exp(\mathrm{tr}(\cdot))$.

The normal-inverse gamma distribution is used as the informative (natural) conjugate prior distribution for the parameters in a normal linear model ($\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}, u \sim N_1(0, \sigma^2)$) in Bayesian inference (O'Hagan, 1994, equation 9.5 and 9.8), (Hansen et al., 2002), (Hansen et al., 2003, eq. 1.6)

$$p_{\mathrm{NIG}}(\mathbf{b}, \sigma^2|a, D, P, \mathbf{m}, \mathbf{V}) = \frac{(a/2)^{D/2}}{(2\pi)^{P/2} |\mathbf{V}|^{1/2} \Gamma(D/2)} (\sigma^2)^{-(D+P+2)/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[(\mathbf{b} - \mathbf{m})^{\mathsf{T}} \mathbf{V}^{-1} (\mathbf{b} - \mathbf{m}) + a \right] \right\} \tag{1.21}$$

$$\propto (\sigma^2)^{-(D+P+2)/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[(\mathbf{b} - \mathbf{m})^{\mathsf{T}} \mathbf{V}^{-1} (\mathbf{b} - \mathbf{m}) + a \right] \right\} \tag{1.22}$$

where P is the dimension of $\mathbf{V}$. The distribution is usually denoted $\mathrm{NIG}(a, D, \mathbf{m}, \mathbf{V})$. The marginal distribution of $\mathbf{b}$ is the multivariate t distribution $t(\mathbf{m}, a\mathbf{V}, D)$. The marginal distribution of σ^2 is the inverse gamma distribution.

Somewhat similar is a distribution for $\theta \equiv \mathbf{b}$ and σ (not σ *squared*) (Box and Tiao, 1992, eq. 2.4.6)

$$p(\theta, \sigma|\mathbf{x}) = \sqrt{\frac{N}{2\pi}} \left[\frac{1}{2} \Gamma\left(\frac{\nu}{2}\right) \right]^{-1} \left(\frac{\nu s^2}{2} \right)^{\nu/2} \sigma^{-(N+1)} \exp\left\{ -\frac{1}{2\sigma^2} \left[\nu s^2 + N(\theta - \bar{x})^2 \right] \right\} \tag{1.23}$$

1.5 Inverse cosh distribution or hyperbolic secant distribution

The inverse hyperbolic cosine, hyperbolic secant distribution or Perks' distribution that is used in independent component analysis (ICA) is heavy-tailed (Smyth, 1994)

$$p(x|\beta) = \frac{1}{2^{\beta-1} B(\beta/2, \beta/2)} \mathrm{sech}^{\beta}(x), \tag{1.24}$$

where B is the beta function. Standardized $\beta = 1$

$$p(x) = \frac{1}{\pi} \mathrm{sech}(x) \tag{1.25}$$

For $\beta = 2$ there is also a nice normalization factor

$$p(x|\beta = 2) = \frac{1}{2} \mathrm{sech}^2(x) \tag{1.26}$$

1.6 Multinomial distribution including binomial and Bernoulli

The discrete multinomial distribution for a discrete variate $\mathbf{n} = [n_1, \ldots, n_P]^\mathsf{T}$ containing counts (Mardia et al., 1979, page 44)

$$P(\mathbf{n}|\mathbf{a}, n) = \frac{n!}{n_1! \ldots n_P!} \prod_p^P a_p^{n_p}, \qquad \sum_p^P n_p = n \tag{1.27}$$

The parameters $\mathbf{a}$ are restricted to be probabilities: $a_p \geq 0$ and $\sum_p^P a_p = 1$. The binomial distribution is obtained with $P = 2$

$$P_{\mathrm{Bin}}(\mathbf{n}|\mathbf{a}, n) = \frac{n!}{n_1! n_2!} \prod_{p=1}^2 a_p^{n_p} = \frac{n!}{n_1! n_2!} a_1^{n_1} (1 - a_1)^{n_2} = \frac{n!}{n_1!(n - n_1)!} a_1^{n_1} (1 - a_1)^{n - n_1} \tag{1.28}$$

$$= \binom{n}{n_1} a_1^{n_1} (1 - a_1)^{n - n_1} \tag{1.29}$$

The Bernoulli distribution appears with $n = 1$, and then $n_1 \in \{0, 1\}$

$$P_{\mathrm{Ber}}(\mathbf{n}|\mathbf{a}) = P(n_1|a_1) = a_1^{n_1} (1 - a_1)^{1 - n_1}. \tag{1.30}$$

...And the important uninomial distribution with $P = 1$

$$P(n|a) = 1, \qquad a = 1 \tag{1.31}$$

1.7 Dirichlet distribution including beta and uniform distributions

The Dirichlet distribution is a kind of "inverse" distribution compared to the multinomial distribution on the bounded continuous variate $\mathbf{x} = [x_1, \ldots, x_P]$ (Mardia et al., 1979, p. 44)

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_p^P \alpha_p\right)}{\prod_p^P \Gamma(\alpha_p)} \prod_p^P x_p^{\alpha_p - 1} \tag{1.32}$$

The beta type I distribution is obtained with a bivariate Dirichlet distribution $P = 2$, $x_1 = x$ and $x_2 = 1 - x$ (Mardia et al., 1979, p. 488)

$$p(x|\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1 - 1} (1 - x)^{\alpha_2 - 1} \tag{1.33}$$

A uniform distribution with limits between 0 and 1 is obtained with $\alpha = \alpha_1 = \alpha_2 = 1$

$$p(x|\alpha) = 1, \qquad \alpha = 1 \tag{1.34}$$

1.8 Multivariate hypergeometric distribution

$$P(\mathbf{n}|\mathbf{m}) = \frac{\prod_{p=1}^P \binom{m_p}{n_p}}{\binom{m}{n}} \tag{1.35}$$

where $\mathbf{m} = [m_1, m_2 \ldots m_P]$, $\mathbf{n} = [n_1, n_2 \ldots n_P]$, $m = m_1 + m_2 + \ldots + m_P$ and $n = n_1 + n_2 + \ldots + n_P$. With $P = 2$ this is the ordinary hypergeometric distribution.

1.9 Normal linear model distribution

The multivariate regression model is (Box and Tiao, 1992, page 439)

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U} \tag{1.36}$$

With the noise as central matrix normal $\mathbf{U}^{\vee} \sim N_{N\times P}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{\Omega})$?

$$\mathbf{B}^{\vee}|\mathbf{X}, \mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Omega} \sim ?N_{Q\times P}\left([\mathbf{X}^{-}\mathbf{Y}]^{\vee}, \mathbf{\Sigma} \otimes \left[\mathbf{X}^{-}\mathbf{\Omega}(\mathbf{X}^{-})^{\mathsf{T}}\right]\right) \tag{1.37}$$

With $\mathbf{U}$ distributed as the white noise matrix normal the maximum likelihood estimates are

$$\widehat{\mathbf{B}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y} = \mathbf{X}^{-}\mathbf{Y} \tag{1.38}$$

$$\widehat{\mathbf{\Sigma}} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^{\mathsf{T}}(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})/N \tag{1.39}$$

With **non-informative** Jeffrey's priors a posterior for $\mathbf{B}$ is the matrix T distribution (Box and Tiao, 1992, page 440), (Mardia et al., 1979, page 180)

$$p(\mathbf{B}|\mathbf{X}, \mathbf{Y}) \propto |N\widehat{\mathbf{\Sigma}} + (\mathbf{B} - \widehat{\mathbf{B}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})|^{-N/2} \tag{1.40}$$

Conjugate priors

$$p(\mathbf{b}|\mathbf{\Sigma}) = N_P(\mathbf{m}, \mathbf{W}) \tag{1.41}$$

$$p(\mathbf{\Sigma}|\mathbf{b}) = W_P \tag{1.42}$$

Posterior of parameters $\mathbf{b}$ for known variance of the residual is a multinormal $\mathbf{\Sigma}_u = \sigma^2$ (O'Hagan, 1994, page 264) (Goutte et al., 2000, equation 6)

$$p(\mathbf{b}|\mathbf{X}, \mathbf{y}, \mathbf{m}, \mathbf{W}, \sigma^2) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{b} - \widehat{\mathbf{b}})^{\mathsf{T}}\widehat{\mathbf{W}}^{-1}(\mathbf{b} - \widehat{\mathbf{b}})\right] \tag{1.43}$$

where the maximum a posteriori (MAP) estimates are

$$\widehat{\mathbf{b}} = (\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}(\mathbf{W}^{-1}\mathbf{m} + \sigma^{-2}\mathbf{X}^{\mathsf{T}}\mathbf{y}), \tag{1.44}$$

$$\widehat{\mathbf{W}} = (\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \tag{1.45}$$

1.10 Gaussian mixture distributions

Gaussian mixture distributions or Mixture of Gaussian (MoG) with a fixed number K of mixtures where $\mathbf{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]^{\mathsf{T}}$ and a data matrix $\mathbf{X}(N \times P)$

$$p(\mathbf{X}|\mathbf{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) = \prod_n^N \sum_k^K p(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\, P(k) \tag{1.46}$$

Heterogeneous isotropic variance (Bishop, 1995, equation 2.77)

$$p(\mathbf{X}|\mathbf{M}, \sigma_1^2, \ldots, \sigma_K^2) = \prod_n^N \sum_k^K (2\pi\sigma_k^2)^{-P/2} \exp\left[-\frac{(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}(\mathbf{x}_n - \boldsymbol{\mu}_k)}{2\sigma_k^2}\right] P(k) \tag{1.47}$$

Equal isotropic variances $\sigma^2 = \sigma_k^2$ with equal mixture weights $P(k) = 1/K$. Number of parameters is $P \times K + 1$.

$$p(\mathbf{X}|\mathbf{M}, \sigma^2) = \frac{1}{K}\prod_n^N \sum_k^K (2\pi\sigma^2)^{-P/2} \exp\left[-\frac{(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}(\mathbf{x}_n - \boldsymbol{\mu}_k)}{2\sigma^2}\right] \tag{1.48}$$

with hard assignment where $\boldsymbol{\mu}_n$ is the mixture component $\boldsymbol{\mu}_k$ assigned exclusively for the $\mathbf{x}_n$ data point, and where $\mathbf{k}(N \times 1)$ contains the assignment indicators for all N objects

$$p(\mathbf{X}|\mathbf{M}, \sigma^2, \mathbf{k}) = \prod_n^N \frac{1}{K} (2\pi\sigma^2)^{-P/2} \exp\left[-\frac{(\mathbf{x}_n - \boldsymbol{\mu}_n)^\mathsf{T} (\mathbf{x}_n - \boldsymbol{\mu}_n)}{2\sigma^2} \right] \tag{1.49}$$

1.10.1 Ad hoc variances

The variance can be computed with a maximum likelihood estimate

$$\sigma^2_{\mathrm{ML}} = \frac{1}{NP} \sum_k^K \sum_n^{N_k} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2. \tag{1.50}$$

Another "unbiased" form (that is not unbiased!)

$$\sigma^2_{\mathrm{UB}} = \frac{1}{P} \sum_k^K \frac{1}{N_k - 1} \sum_n^{N_k} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2. \tag{1.51}$$

This estimate does, however, not work in practice since there might be a small number of objects in a specific cluster such that $N_k = 1$. An *ad hoc* fix with "modified unbiased estimator"

$$\sigma^2_{\mathrm{MUB}} = \frac{1}{P} W \sum_k^K \begin{cases} \frac{1}{N_k - 1} \sum_n^{N_k} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 & \text{if } N_k > 1 \\ 0 & \text{if } N_k \leq 1 \end{cases} \tag{1.52}$$

Where W is "some weight" most simply taken as $W = 1$

$$\tag{1.53}$$

1.10.2 Univariate Gaussian mixture models

The symmetric subgaussian Pearson mixture model as seen in (Lee et al., 1999): $p(s|\mu, \sigma^2) = 1/2 \left(N(\mu, \sigma^2) + N(-\mu, \sigma^2) \right)$

$$p(s|\mu, \sigma^2) = \frac{1}{4\pi\sigma^2} \left[\exp\left(-\frac{(s - \mu)^2}{2\sigma^2} \right) + \exp\left(-\frac{(s + \mu)^2}{2\sigma^2} \right) \right] \tag{1.54}$$

1.11 Hyperbolic secant mixture distribution

The univariate hyperbolic secant mixture distribution is used in independent component analysis to model both heavy-tailed distributions and bimodal distributions (Lee et al., 1999, equation 2.29)

$$p(s|\mu) \propto \operatorname{sech}^2(s - \mu) + \operatorname{sech}^2(s + \mu) \tag{1.55}$$

1.12 Dirichlet-multinomial distribution

With a data matrix $\mathbf{N}(N \times P)$ with N independent distributed samples from a Dirichlet-multinomial distribution where an infinite number of multinomial is mixed and weighted by a Dirichlet distribution

$$P(\mathbf{N}|\boldsymbol{\alpha}, n) = \prod_n^N \int p_{\mathrm{mul}}(\mathbf{N}|\mathbf{a})\, p_{\mathrm{Dir}}(\mathbf{a}|\boldsymbol{\alpha})\, d\mathbf{a} \tag{1.56}$$

Here $\mathbf{a} \geq 0$ contains the parameters for the multinomial(s) and $\boldsymbol{\alpha}$ contains the Dirichlet parameters, i.e., a hyperparameter for the multinomial distribution. Minka calls the Dirichlet-multinomial a "Polya distribution" (Minka, 2003b). The model has been applied in text analysis (Madsen et al., 2005).

1.12.1 Beta-binomial mixture distribution

The beta-binomial mixture distribution (BBM) is continuously mixing via a beta distribution (Griffith, 1973; Erdfelder, 1993), Anders Hald?

$$P(\mathbf{n}|n, \boldsymbol{\alpha}) = P(n_1|n, \boldsymbol{\alpha}) = \int_0^1 P_{\mathrm{Bin}}(\mathbf{n}|a_1, n)\, p_{\mathrm{Dir}}(a_1|\boldsymbol{\alpha})\, da_1 \tag{1.57}$$

$$= \binom{n}{n_1} \frac{B(n_1 + \alpha_1, n - n_1 + \alpha_2)}{B(\alpha_1, \alpha_2)}, \tag{1.58}$$

where $\mathbf{n} = [n_1, n_2]^\mathsf{T}$ and $n = n_1 + n_2$. It might be said not to be a mixture distribution since the beta distribution could be regarded as a Bayesian prior distribution (that in this case is the natural conjugate prior) where $\boldsymbol{\alpha}$ is the hyperparameters. Applications appear in (Lowe, 1999a; Lowe, 1999b).

1.13 Gaussian ratio distribution

The univariate Gaussian ratio distribution constructed from two correlated Gaussian distributions: $p(x|\mu_x, \sigma_x)$ and $p(y|\mu_y, \sigma_y)$ with correlation coefficient ρ. The distribution of the ratio $z = x/y$ is given as (Hinkley, 1969, equations 1 and 2)

$$p(z|\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \frac{b(z)d(z)}{\sqrt{2\pi}\sigma_x\sigma_y a^3(z)} \left[\Phi\left\{ \frac{b(z)}{\sqrt{1-\rho^2}a(z)} \right\} - \Phi\left\{ -\frac{b(z)}{\sqrt{1-\rho^2}a(z)} \right\} \right] \tag{1.59}$$

$$+ \frac{\sqrt{1-\rho^2}}{\pi\sigma_x\sigma_y a^2(z)} \exp\left\{ -\frac{c}{2(1-\rho^2)} \right\}, \tag{1.60}$$

$$a(z) = \left(\frac{z^2}{\sigma_x^2} - \frac{2\rho z}{\sigma_x\sigma_y} + \frac{1}{\sigma_y^2} \right)^{\frac{1}{2}}, \tag{1.61}$$

$$b(z) = \frac{\mu_x z}{\sigma_x^2} - \frac{\rho(\mu_x + \mu_y z)}{\sigma_x\sigma_y} + \frac{\mu_y}{\sigma_y^2}, \tag{1.62}$$

$$c = \frac{\mu_x^2}{\sigma_x^2} - \frac{2\rho\mu_x\mu_y}{\sigma_1\sigma_y} + \frac{\mu_y^2}{\sigma_y^2}, \tag{1.63}$$

$$d(z) = \exp\left\{ \frac{b^2(z) - ca^2(z)}{2(1-\rho^2)a^2(z)} \right\}, \tag{1.64}$$

$$\Phi(u) = \int_{-\infty}^{u} \phi(v)\, dv, \qquad \text{where} \qquad \phi(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}. \tag{1.65}$$

With the form of (Pham-Gia et al., 2006) with a confluent hypergeometric function

$$p(z|\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = K_2 \frac{2(1-\rho^2)\sigma_x^2\sigma_y^2}{\sigma_y^2 z^2 - 2\rho\sigma_x\sigma_y z + \sigma_x^2} \left[{}_1F_1\left(1, 1/2, \theta_2(z)\right) \right] \tag{1.66}$$

$$\theta_2(z) = \frac{\left[\sigma_y^2\mu_x z - \rho\sigma_x\sigma_y(\mu_y z - \mu_x) + \mu_y\sigma_x^2 \right]^2}{\sigma_x\sigma_y 2(1-\rho^2)(\sigma_y^2 z^2 - 2\rho\sigma_x\sigma_y z + \sigma_x^2)} \tag{1.67}$$

$$K_2 = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{\sigma_y^2\mu_x^2 - 2\rho\sigma_x\sigma_y\mu_x\mu_y + \mu_y^2\sigma_x^2}{2(1-\rho^2)\sigma_x^2\sigma_y^2} \right) \tag{1.68}$$

1.13.1 Gaussian ratio distribution with no correlation

With correlation coefficient set to zero $\rho = 0$ in (Hinkley, 1969, equations 1 and 2)

$$p(z|\mu_x, \mu_y, \sigma_x, \sigma_y, \rho = 0) = \frac{b(z)d(z)}{\sqrt{2\pi}\sigma_x\sigma_y a^3(z)} \left[\Phi\left\{\frac{b(z)}{a(z)}\right\} - \Phi\left\{-\frac{b(z)}{a(z)}\right\}\right] + \frac{1}{\pi\sigma_x\sigma_y a^2(z)} \exp\left\{-\frac{c}{2}\right\} \quad (1.69)$$

$$a(z) = \left(\frac{z^2}{\sigma_x^2} + \frac{1}{\sigma_y^2}\right)^{\frac{1}{2}} \quad (1.70)$$

$$b(z) = \frac{\mu_x z}{\sigma_x^2} + \frac{\mu_y}{\sigma_y^2} \quad (1.71)$$

$$c = \frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2} \quad (1.72)$$

$$d(z) = \exp\left\{\frac{b^2(z) - ca^2(z)}{2a^2(z)}\right\} \quad (1.73)$$

With the form of (Pham-Gia et al., 2006) with confluent hypergeometric function

$$p(z|\mu_x, \mu_y, \sigma_x, \sigma_y, \rho = 0) = \frac{K_1}{\sigma_y^2 z^2 + \sigma_x^2} \left[{}_1F_1\left(1, 1/2, \theta_1(z)\right)\right] \quad (1.74)$$

$$\theta_1(z) = \frac{1}{2\sigma_x^2\sigma_y^2} \frac{\left(\sigma_y^2\mu_x z + \mu_y\sigma_x^2\right)^2}{\sigma_y^2 z^2 + \sigma_x^2} \quad (1.75)$$

$$K_1 = \frac{\sigma_x\sigma_y}{\pi} \exp\left(-\frac{1}{2}\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right) \quad (1.76)$$

1.13.2 Gaussian ratio distribution with no mean

With correlation coefficient set to zero $\rho = 0$ and no means $\mu_x = 0$, $\mu_y = 0$ in (Hinkley, 1969, equations 1 and 2)

$$p(z|\mu_x = 0, \mu_y = 0, \sigma_x, \sigma_y, \rho = 0) = \frac{1}{\pi\sigma_x\sigma_y a^2(z)} \quad (1.77)$$

$$a(z) = \left(\frac{z^2}{\sigma_x^2} + \frac{1}{\sigma_y^2}\right)^{\frac{1}{2}} \quad (1.78)$$

$$b(z) = 0 \quad (1.79)$$

$$c = 0 \quad (1.80)$$

$$d(z) = 1 \quad (1.81)$$

$$p(z|\mu_x = 0, \mu_y = 0, \sigma_x, \sigma_y, \rho = 0) = \frac{1}{\pi} \frac{\sigma_x/\sigma_y}{z^2 + (\sigma_x/\sigma_y)^2} \quad (1.82)$$

. . . and this is the Cauchy distribution with zero mean. If $\rho \neq 0$

$$p(z|\mu_x = 0, \mu_y = 0, \sigma_x, \sigma_y, \rho) = \frac{\sqrt{1-\rho^2}}{\pi \sigma_x \sigma_y a^2(z)} \tag{1.83}$$

$$a(z) = \left(\frac{z^2}{\sigma_x^2} - \frac{2\rho z}{\sigma_x \sigma_y} + \frac{1}{\sigma_y^2} \right)^{\frac{1}{2}} \tag{1.84}$$

$$b(z) = 0 \tag{1.85}$$

$$c = 0 \tag{1.86}$$

$$d(z) = 1 \tag{1.87}$$

$$p(z|\mu_x = 0, \mu_y = 0, \sigma_x, \sigma_y, \rho) = \frac{1}{\pi} \frac{\beta}{(z-\alpha)^2 + \beta^2}, \tag{1.88}$$

$$\text{where,} \qquad \alpha = \rho \frac{\sigma_x}{\sigma_y} \tag{1.89}$$

$$\beta = \frac{\sigma_x}{\sigma_y} \sqrt{1-\rho^2}, \tag{1.90}$$

i.e., the Cauchy distribution with non-zero mean (Curtiss, 1941, page 411).

1.14 Finite binomial mixture distribution

The finite binomial mixture distribution is a finite discrete mixture distribution, sometimes refered to as finite binomial mixture (FBM) (Gelfand and Solomon, 1974), (Emrick, 1971)?, Poisson? Pearson (1915)? also discussed in (Everitt and Hand, 1981, pp. 89–97), (Titterington et al., 1985) with K components

$$p(x|N, \mathbf{p}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k P_{\text{bin}}(x|N, p_k); \tag{1.91}$$

where $x \in \{0, \dots, N\}$ and $0 \leq \lambda_k \leq 1$, $\sum_k^K \lambda_k = 1$ are the mixing proportions. The parameters are identifiable with $N \geq 2K - 1$. The likelihood can be written with another sufficient statistics $\mathbf{c}$

$$p(\mathbf{c}|N, \mathbf{p}, \boldsymbol{\lambda}) = \prod_{m=0}^{M} \left[\sum_{k=1}^{K} \lambda_k P_{\text{bin}}(m|N, p_k) \right]^{c_m} \tag{1.92}$$

where $c_m = \sum_{n=1, x_n = m}^{N} 1$ and $m = 0 \dots M$. This might be more efficient to work with if M is small.

$$\tag{1.93}$$

1.15 Multinomial mixture distribution

A data matrix $\mathbf{X}(N \times P)$ with N independent samples from a multinomial mixture distribution with $\mathbf{A}$ containing parameters for the multinomial distributions and $\mathbf{q}$ containing the weighting parameters between the multinomials

$$P(\mathbf{X}|\mathbf{A}, \mathbf{q}) = \prod_n^N \sum_k^K q_k \frac{(\sum_p^P x_{np})!}{\prod_p^P x_{np}!} \prod_p^P a_{kp}^{x_{np}} \tag{1.94}$$

There are several applications of this model in text analysis, e.g., (Rigouste et al., 2005; Nigam et al., 2000).

1.16 Mixture of matrix hypergeometric distributions

With K mixtures and N objects

$$P(\mathbf{X}|\mathbf{Z},\mathbf{q}) = \prod_n^N \sum_k^K \frac{\prod_p^P \binom{z_{kp}}{x_{np}}}{\binom{\sum_p^P z_{kp}}{\sum_p^P x_{np}}} q_k. \tag{1.95}$$

Chapter 2

Statistical tests and other

2.1 Correlation coefficient

With the correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2)}} \tag{2.1}$$

And the hypotheses

$$\begin{array}{rcrcl} H_0 & : & r_{xy} & = & 0 \\ H_1 & : & r_{xy} & \neq & 0 \end{array} \tag{2.2}$$

The test statistic is

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{2.3}$$

and distributed with a t-distribution with $N-2$ degrees of freedom if x and y are identical and independent Gaussian distributed (Gaussian iids).

2.2 Partial correlation coefficient

The partial correlation coefficient between x and y with z as the nuisance

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \tag{2.4}$$

The test statistic is formed as

$$t = r\sqrt{\frac{N-2-K}{1-r^2}} \tag{2.5}$$

Where K is the size of the subspace factored out and N is the number of samples. For Gaussian distributed signals (Gaussian iids) this can be compared to a t-distribution with $N-2-K$ degrees of freedom.

2.3 Effect sizes

Hedges g for the effect size between two groups (Hedges and Olkin, 1985)

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s} \qquad \text{where} \qquad s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}, \tag{2.6}$$

With

$$n^* = n_1 + n_2 - 2 \qquad \text{and} \qquad \tilde{n} = \frac{n_1 n_2}{n_1+n_2}, \tag{2.7}$$

the population value

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{2.8}$$

and the bias correction factor

$$J(a) = \frac{\Gamma(a/2)}{\sqrt{a/2}\Gamma((a-1)/2)} \approx 1 - \frac{3}{4a-1} \tag{2.9}$$

then under Gaussianity of the data $\sqrt{\tilde{n}}g$ follows a noncentral t-distribution with n^* degress of freedom and the noncentrality parameter $\delta\sqrt{\tilde{n}}$, so the expection and variance of g are (Hedges and Olkin, 1985, p. 104)

$$E(g) = \delta / J(n^*) \tag{2.10}$$

$$\approx \delta + \frac{3\delta}{4(n_1+n_2)-9} \tag{2.11}$$

$$\mathrm{Var}(g) = \frac{n^*}{(n^*-2)\tilde{n}} + \delta^2 \left(\frac{n^*}{n^*-2} - \frac{1}{[J(n^*)]^2} \right) \tag{2.12}$$

$$\approx \frac{1}{n^*} + \delta^2 \frac{1}{2(n_1+n_2-3.94)} \tag{2.13}$$

2.4 Combining mean and standard deviation

A data set consisting of two parts where the mean and standard deviation is known for the two part: What are the mean and standard deviation for the full data set?

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}. \tag{2.14}$$

The two means combined in a weighted fashion to the full mean

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_i^{n_1} \mathbf{x}_{1,i} \tag{2.15}$$

$$\bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_j^{n_2} \mathbf{x}_{2,j} \tag{2.16}$$

$$\mathbf{x} = \frac{1}{n} \sum_k^{n} \mathbf{x}_k = \frac{\bar{\mathbf{x}}_1 n_1 + \bar{\mathbf{x}}_2 n_2}{n_1+n_2} \tag{2.17}$$

Unbiased standard deviations

$$\mathbf{S}_1 = \frac{1}{n_1-1} \sum_i^{n_1} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_1)' \tag{2.18}$$

$$\mathbf{S}_2 = \frac{1}{n_2-1} \sum_j^{n_2} (\mathbf{x}_{2,j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2,j} - \bar{\mathbf{x}}_2)' \tag{2.19}$$

$$s^2 = \frac{1}{n-1} \left[s_1(n_1-1) + s_2(n_2-1) + n_1\bar{x}_1^2 + n_2\bar{x}_2^2 - n\bar{x}^2 \right] \tag{2.20}$$

$$\mathbf{S} = \frac{1}{n-1} \left[\mathbf{S}_1(n_1-1) + \mathbf{S}_2(n_2-1) + n_1\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1' + n_2\bar{\mathbf{x}}_2\bar{\mathbf{x}}_2' - n\bar{\mathbf{x}}\bar{\mathbf{x}}' \right] \tag{2.21}$$

Chapter 3

Matrices

3.1 Definition

A matrix is a 2-dimensional table/array of numbers (elements). The numbers can be complex or real. Usually all the elements should be defined, though in some statistical applications one allows for "missing values", — in R denoted by NA ("not available").

3.2 Types of matrices

The matrix can be ordered according to several criteria (from the Matrix Market):

- Non-zero structure: Full, diagonal, tridiagonal, block diagonal, sparse, (lower/upper) triangular.
- Shape: Square, non-square, (row/column) vector, scalar
- Element type: Complex, real, integer, non-negative, positive, binary.
- Numerical symmetric: Hermitian, symmetric, positive definite, ...

3.2.1 Monomial

A *monomial* matrix (also called a *generalized permutation* matrix) contains exactly one non-zero entry in each row and column, i.e. is of full-ranked (Berman and Plemmons, 1994, p. 67). A non-singular monomial matrix has exactly one non-zero entry in each column Thus permutation and identity matrices are specializations.

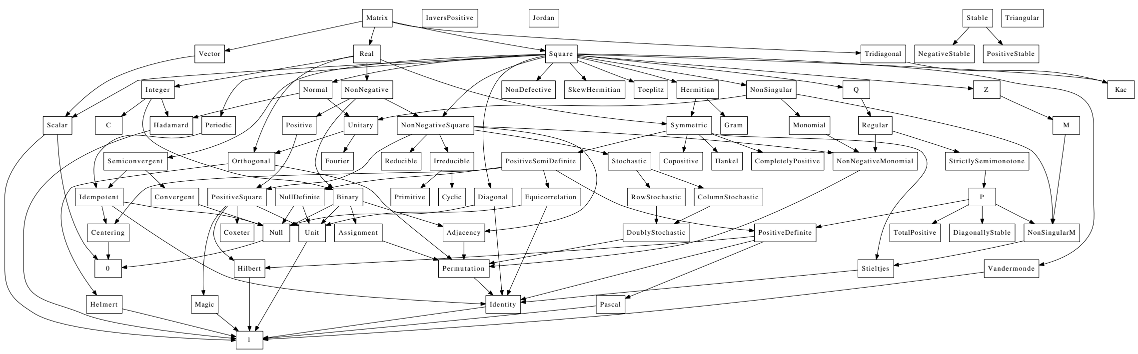


Figure 3.1: Matrix taxonomy

3.2.2 Semiconvergent and convergent matrices

For a *semiconvergent* matrix $\mathbf{C}$ the following applies (Berman and Plemmons, 1994, p. 152)

$$\lim_{k\to\infty} \mathbf{C}^k \text{ exists.} \tag{3.1}$$

A *convergent* matrix $\mathbf{B}$ satisfies (Berman and Plemmons, 1979, p. 9)

$$\lim_{k\to\infty} \mathbf{B}^k = \mathbf{0} \tag{3.2}$$

In both case the convergence is related to the spectral radius: For a convergent $\rho(\mathbf{B}) < 1$.

3.2.3 Z-matrices

A real square matrix is a Z-matrix if all its off-diagonal elements are non-positive (Guo, 2003), and it can be written as

$$\mathbf{A} = s\mathbf{I} - \mathbf{B}, \qquad \mathbf{B} \geq \mathbf{0}. \tag{3.3}$$

M-matrices

An M-matrix is square and has nonnegative diagonal elements and nonpositive off-diagonal elements (Graham, 1987, chapter 5). An M-matrix $\mathbf{A}$ can be decomposed into an positive scaled identity matrix ($s\mathbf{I}$, $s > 0$) and a nonnegative matrix $\mathbf{B}$

$$\mathbf{A} = s\mathbf{I} - \mathbf{B}. \tag{3.4}$$

(Guo, 2003) relates the spectral radius $\rho(\cdot)$ to s: A Z-matrix is an M-matrix if $s \geq \rho(\mathbf{B})$, a singular M-matrix if $s = \rho(\mathbf{B})$ and a non-singular M-matrix if $s > \rho(\mathbf{B})$. The inverse of such a non-singular M-matrix is non-negative (Guo, 2003).[1]

3.2.4 Nonnegative matrices

A *nonnegative matrix* $\mathbf{A}(N \times M)$ is a matrix where all the element are greater than or equal to zero

$$\mathbf{A} \geq \mathbf{0} \qquad \equiv \qquad \forall_{ij} a_{ij} \geq 0 \tag{3.5}$$

And similarly a *positive* matrix $\mathbf{B}(N \times M)$ is a matrix where all the elements are greater than zero

$$\mathbf{B} > 0. \tag{3.6}$$

Doubly nonnegative

A doubly nonnegative square matrix is both nonnegative and positive semi-definite. A doubly nonnegative matrix $\mathbf{A}$ that is *completely positive* can be decomposed with a (not necessarily square) nonnegative matrix $\mathbf{B}$ (Berman and Plemmons, 1994, p. 304)

$$\mathbf{A} = \mathbf{B}\mathbf{B}^{\mathsf{T}} \qquad \mathbf{A} \geq \mathbf{0},\ \mathbf{B} \geq \mathbf{0}. \tag{3.7}$$

Inverse nonnegative matrices

Some non-negative matrices have an inverse that is an M-matrix. Other classes of inverse nonnegative matrices are *strictly ultrametric matrices* (SUM) and *generalized ultrametric matrices* (GUM) (Berman and Plemmons, 1994, p. 306–307).

For a non-singular non-negative matrix $\mathbf{A}$ its inverse $\mathbf{A}^{-1}$ is only non-negative if (and only if) $\mathbf{A}$ is (non-negative) monomial (Berman and Plemmons, 1994, chapter 3, exercise 6.7).

An example of a simple inverse non-negative matrix is

$$\mathbf{A} = \mathbf{A}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{3.8}$$

and this is monomial.

[1] Is Graham's and Guo's M-matrix the same?

3.2.5 Normal matrices

(Graham, 1987, p. 86+).

$$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* \tag{3.9}$$

3.2.6 Hermitian

A complex matrix is *Hermitian* if the following applies

$$\mathbf{A}^* = \mathbf{A}, \tag{3.10}$$

where * denotes the conjugate transpose (*tranjugate*) (Graham, 1987, p. 67).

Symmetric

A real hermitian matrix is symmetric

$$\mathbf{A}^\mathsf{T} = \mathbf{A}, \tag{3.11}$$

i.e., a symmetric matrix is *real*.

3.2.7 Unitary matrices

A complex square matrix $\mathbf{U}(N \times N)$ is unitary if it satisfy (Graham, 1987, p. 67)

$$\mathbf{U}^*\mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^*, \tag{3.12}$$

"Special unitary matrices" are unitary matrices with $\det(\mathbf{X}) = 1$.

Orthogonal matrices

A real matrix $\mathbf{P}$ that is unitary satisfy (Graham, 1987, p. 68)

$$\mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{P}\mathbf{P}^\mathsf{T} = \mathbf{I} \tag{3.13}$$

$$\det(\mathbf{P}) = \pm 1 \tag{3.14}$$

$$\mathbf{P}^{-1} = \mathbf{P}^\mathsf{T} \tag{3.15}$$

Givens matrices

Givens matrices are rotation matrices.

Helmert matrix

A Helmert matrix $\mathbf{A}(N \times N)$ can be used to decorrelate random variables.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{(1)} \\ \vdots \\ \mathbf{a}_{(N)} \end{bmatrix}, \qquad \mathbf{a}_{(n)} = \begin{cases} \left[1/\sqrt{N}, \ldots, 1/\sqrt{N} \right] & \text{if } n = 1 \\ \left[\underbrace{\sqrt{(n-1)n}, \ldots, \sqrt{(n-1)n}}_{(n-1) \text{ elements}}, \quad \underbrace{0, \ldots, 0}_{(N-n) \text{ elements}} \right] & \text{otherwise} \end{cases} \tag{3.16}$$

Permutation matrices

A *permutation matrix* $\mathbf{P}$ is a matrix that can be used to interchange the rows (with $\mathbf{PA}$) or columns (with $\mathbf{AP}$) of an other matrix (here $\mathbf{A}$). An example is (Graham, 1987, p. 43)

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.17}$$

Permutation matrices are orthogonal and binary matrices.

3.3 Arithmetic operations

The row sum of $\mathbf{A}(N \times M)$ gives a column vector $\mathbf{r}(N \times 1)$

$$\mathbf{r} = \sum_m^M a_{nm} = \mathbf{A}\mathbf{1}_M. \tag{3.18}$$

Strangely, in matlab the multiplication '`X*ones(size(X,2),1)`' is faster than a "simple" sum '`sum(X,2)`', — perhaps because of poor cache management (Minka, 2003a).

3.4 Vectorization

A matrix $\mathbf{X}(N \times P)$

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P] \tag{3.19}$$

Vectorization of $\mathbf{X}$ into a column vector $\mathbf{X}^{\mathsf{V}}(NP \times 1)$

$$\mathbf{X}^{\mathsf{V}} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_P \end{bmatrix}. \tag{3.20}$$

Arithmetic operations with $\mathbf{A}(I \times J)$ and $\mathbf{B}(J \times K)$

$$(\mathbf{ABC})^{\mathsf{V}} = (\mathbf{C}^{\mathsf{T}} \otimes \mathbf{A})\mathbf{B}^{\mathsf{V}}, \tag{3.21}$$

$$(\mathbf{AB})^{\mathsf{V}} = (\mathbf{I}_K \otimes \mathbf{A})\mathbf{B}^{\mathsf{V}} \tag{3.22}$$

$$= (\mathbf{B}^{\mathsf{T}} \otimes \mathbf{I}_I)\mathbf{A}^{\mathsf{V}}. \tag{3.23}$$

3.5 Function with scalar output

3.5.1 Norms

The rank of a matrix is the number of non-zero singular values

$$\mathrm{rank}(\mathbf{A}) = \sum_i \delta(\sigma_i(\mathbf{A}) \neq 0). \tag{3.24}$$

The ordinary norm (spectral norm or 2-norm) is the largest singular value.

$$\|\mathbf{A}\|_2 = \sqrt{\max_i(\lambda_i(\mathbf{A}^{\mathsf{H}}\mathbf{A}))}. \tag{3.25}$$

If the matrix is symmetric it is the largest eigenvalue

$$\|\mathbf{A}\|_2 = \max_i(\lambda_i(\mathbf{A})), \tag{3.26}$$

Frobenius norm (Golub and Van Loan, 1996, section 2.3)

$$\|\mathbf{A}\|_F = \sqrt{\sum\sum |a_{ij}|^2} \tag{3.27}$$

$$= \sqrt{\sum_k \sigma_k^2(\mathbf{A})} \tag{3.28}$$

$$= \sqrt{\sum_k \lambda_k(\mathbf{A}^{\mathsf{T}}\mathbf{A})} = \sqrt{\sum_k \lambda_k(\mathbf{A}\mathbf{A}^{\mathsf{T}})} \tag{3.29}$$

$$= \sqrt{\mathrm{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{A})} = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathsf{T}})} \tag{3.30}$$

where the σ_k's are the singular values, and the λ_k's are the eigenvalues.

In some tasks of least squares the cost function is the Frobenius norm and in certain cases it may be faster to compute the cost function in a different way. In the case where the number of columns in $\mathbf{A}$ is large and the number of columns in $\mathbf{B}$ the following may result in fewer computation during an iterative algorithm where $\mathbf{B}$ and $\mathbf{C}$ are optimized:

$$\begin{aligned} \|\mathbf{A}-\mathbf{BC}\|_F^2 &= \mathrm{tr}\left[(\mathbf{A}-\mathbf{BC})(\mathbf{A}-\mathbf{BC})^\mathsf{T}\right] && (3.31)\\ &= \mathrm{tr}\left[\mathbf{A}\mathbf{A}^\mathsf{T}\right] - 2\mathrm{tr}\left[\mathbf{B}\left(\mathbf{C}\mathbf{A}^\mathsf{T}\right)\right] + \mathrm{tr}\left[\mathbf{B}\left(\mathbf{C}\mathbf{C}^\mathsf{T}\right)\mathbf{B}^\mathsf{T}\right] && (3.32)\\ &= \|\mathbf{A}\|_F^2 - 2\mathrm{tr}\left[\mathbf{B}\left(\mathbf{C}\mathbf{A}^\mathsf{T}\right)\right] + \mathrm{tr}\left[\mathbf{B}\left(\mathbf{C}\mathbf{C}^\mathsf{T}\right)\mathbf{B}^\mathsf{T}\right] && (3.33) \end{aligned}$$

The scalar term $\|\mathbf{A}\|_F^2$ only needs to be computed once during the iterations.

3.5.2 Spectral radius

The spectral radius $\rho(\mathbf{A})$ is the largest of the absolute values of the eigenvalues

$$\rho(\mathbf{A}) = \max_i(|\lambda_i(\mathbf{A})|) \tag{3.34}$$

3.6 Inversion

The inverse $\mathbf{A}^{-1}$ of a *square and non-singular* matrix $\mathbf{A}$ satisfies

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{3.35}$$

$$a\mathbf{C} + a^2\mathbf{C}^2 + a^3\mathbf{C}^3 + \ldots + a^k\mathbf{C}^k = (\mathbf{I} - a\mathbf{C})^{-1} - \mathbf{I}, \tag{3.36}$$

which is used in a recursive centrality index computation (Katz, 1953).

3.6.1 Properties

Some properties, see, e.g., (Mardia et al., 1979, section A.2.4) and the matrix cookbook

$$\begin{aligned} \mathbf{A}^{-\mathsf{T}} &= (\mathbf{A}^\mathsf{T})^{-1} = (\mathbf{A}^{-1})^\mathsf{T} && (3.37)\\ (c\mathbf{A})^{-1} &= c^{-1}\mathbf{A}^{-1} && (3.38)\\ (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} && (3.39)\\ \mathbf{A}\mathbf{B}^\mathsf{T} &= \left(\mathbf{B}^{-\mathsf{T}}\mathbf{A}^{-1}\right)^{-1} && (3.40) \end{aligned}$$

Kronecker product

$$(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1} \tag{3.41}$$

3.6.2 Matrix inversion lemma

'Matrix inversion lemma' or ((Bartlett-)Sherman-)Morrison-Woodbury formula (Mardia et al., 1979, equation A.2.4f), (Golub and Van Loan, 1996, section 2.1.3)

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1} \tag{3.42}$$

In simplified versions

$$\begin{aligned} (\mathbf{A}+\mathbf{BD})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1} && (3.43)\\ (\mathbf{A}+\mathbf{a}\mathbf{b}^\mathsf{T})^{-1} &= \mathbf{A}^{-1} - (\mathbf{A}^{-1}\mathbf{a}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1})(1+\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{a})^{-1} && (3.44) \end{aligned}$$

3.6.3 2×2 inversion

With a (2×2)-matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{3.45}$$

the inverse is

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \tag{3.46}$$

i.e., the inversion of a positive (and invertible) (2×2)-matrix will always have an inverse in which two of the elements are negative, — either the off-diagonal or (with negative determinant) the on-diagonal elements.

3.7 Determinant

A determinant only applies for square matrices. With a square $\mathbf{A}(P \times P)$

$$|c\mathbf{A}| = c^P |\mathbf{A}| \tag{3.47}$$
$$|\mathbf{A}^n| = |\mathbf{A}|^n \tag{3.48}$$
$$|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}| \tag{3.49}$$
$$|\mathbf{A} + \mathbf{BC}| = |\mathbf{A}|\,|\mathbf{I}_P + \mathbf{A}^{-1}\mathbf{BC}| \tag{3.50}$$
$$= |\mathbf{A}|\,|\mathbf{I}_N + \mathbf{CA}^{-1}\mathbf{B}| \tag{3.51}$$

3.8 Trace

The trace of a *square* matrix is the sum of the diagonal elements

$$\mathrm{tr}(\mathbf{A}) = \sum_i a_{ii}. \tag{3.52}$$
$$\mathrm{tr}(\mathbf{C}^\mathsf{T}\mathbf{D}) = \sum_i \sum_j c_{ij} d_{ij} \tag{3.53}$$
$$= (\mathbf{C}^\vee)^\mathsf{T} \mathbf{D}^\vee \tag{3.54}$$
$$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \tag{3.55}$$

For a 'squared' matrix the trace can be computed with the Frobenius norm

$$\mathrm{tr}(\mathbf{C}^\mathsf{T}\mathbf{C}) = \|\mathbf{C}\|_F^2 \tag{3.56}$$

3.9 Derivatives

Matrix derivatives are defined as, see, e.g., (Mardia et al., 1979, section A.9)

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left[\frac{\partial f(\mathbf{X})}{\partial x_{ij}} \right]. \tag{3.57}$$

3.9.1 Determinant

$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = \left[\mathbf{X}^\mathsf{T}\right]^{-1}. \tag{3.58}$$

3.10 Integrals

3.10.1 Transformation with Jacobian

Change from one coordinate system $\mathbf{x}$ to another $\mathbf{u}$ by the transformation $\mathbf{x} = \boldsymbol{\phi}(\mathbf{u})$ (Jacobsen, 2004)

$$\int_D f(\mathbf{x})\, d\mathbf{x} = \int_{D'} f(\boldsymbol{\phi}(\mathbf{u}))\, |\det(\mathbf{J})|\, d\mathbf{u}, \tag{3.59}$$

with the Jacobian $\mathbf{J} = \partial\boldsymbol{\phi}(\mathbf{u})/\partial\mathbf{u}$ and the integration region $\boldsymbol{\phi}(D') = D$.

3.10.2 Dirac function

Scalar integral involving Dirac delta function (MacKay, 1999, page 3) relevant for 'no-noise' distributions

$$\int \delta(x - vs)\, f(s)\, ds = \frac{1}{|v|} f(x/v), \tag{3.60}$$

and the matrix version

$$\int \prod_j^J \delta(x_j - V_{ji} s_i) \prod_i^I f_i(s_i)\, d^I\mathbf{s} = \frac{1}{|\mathbf{V}|} \prod_i^I f_i(V_{ij}^{-1} x_j)? \tag{3.61}$$

Chapter 4

Functions

4.1 Gamma function

The gamma function (found in Matlab as `gamma`) defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \, du \tag{4.1}$$

The faculty function, when x is a positive integer

$$\Gamma(x) = (x-1)! \tag{4.2}$$

The lower incomplete gamma

$$\gamma(x, a) = \int_0^x u^{a-1} e^{-u} \, du \tag{4.3}$$

The lower "standardized" incomplete gamma is computed by Matlab's `gammainc`

$$\gamma(x, a) = 1/\Gamma(a) \int_0^x u^{a-1} e^{-u} \, du \tag{4.4}$$

The generalized gamma function (Box and Tiao, 1992, equation 8.2.22)

$$\Gamma(M, N) = \left[\Gamma\left(\frac{1}{2}\right)\right]^{M(M-1)/2} \prod_m^M \Gamma\left(N + \frac{m+M}{2}\right), \tag{4.5}$$

and since $\Gamma(1/2) = \sqrt{\pi}$

$$\Gamma(M, N) = \pi^{M(M-1)/4} \prod_m^M \Gamma\left(N + \frac{m+M}{2}\right). \tag{4.6}$$

The digamma function

$$\Psi(x) = \psi_0(x) = \frac{\Gamma'(x)}{\Gamma(x)} = \frac{d}{dx} \ln \Gamma(x). \tag{4.7}$$

4.2 Basis functions

Chebyshev polynomials of the first kind $T_n(x)$ are orthogonal

$$T_n(x) = \cos\left(n \operatorname{acos}(x)\right) \qquad x \in [-1; 1],\, n = 0 \ldots \tag{4.8}$$

Other orthogonal polynomials are Hermite and Legendre polynomials.

4.3 Link functions

Bradley-Terry-Luce $\mathbf{z}(P \times 1)$

$$\mathbf{y} = \frac{\mathbf{z}^{\alpha}}{\sum_p^P z_p^{\alpha}}, \tag{4.9}$$

where $\mathbf{z}^{\alpha}$ is elementwise.

Softmax transformation

$$\mathbf{y} = \frac{\exp(\alpha \mathbf{z})}{\sum_p^P \exp(\alpha z_p)} \tag{4.10}$$

The logit transform

$$y_1 = \frac{\exp(z_1)}{1 + \exp(z_1)} \tag{4.11}$$

Chapter 5

Inequalities

5.1 Jensen's inequality

$$\log\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i \log(x_i) \qquad \text{If } \lambda_i \geq 0 \text{ and } \sum_i \lambda_i = 1 \tag{5.1}$$

Chapter 6

Optimization

6.1 Gradient descent

Gradient descent (also steepest descent)

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma \boldsymbol{\nabla} f(\mathbf{x}_n) \tag{6.1}$$

Gradient ascent

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \gamma \boldsymbol{\nabla} f(\mathbf{x}_n) \tag{6.2}$$

6.2 Newton method

Minimization:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{H}_{\mathbf{xx}}^{-1} \boldsymbol{\nabla}_{\mathbf{x}} f(\mathbf{x}_n) \tag{6.3}$$

Chapter 7

Information theory

7.1 Entropies

Shannon entropy for a discrete random variable

$$H(X) = -\sum_{x \in X} P_X(x) \log_2 [P_X(x)]. \tag{7.1}$$

Rényi entropy of order α (Cachin, 1997, page 15)

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x \in X} P_X(x)^\alpha, \tag{7.2}$$

where $\alpha \geq 0$ and $\alpha \neq 1$. Rényi entropy of order $\alpha = 2$ aka. quadratic entropy (Torkkola, 2002)

$$H_2(X) = -\log_2 \sum_x [P(x)]^2. \tag{7.3}$$

"Guessing" entropy for ordered probabilities

$$H_G(X) = \log_2 \sum_i i \, p_i. \tag{7.4}$$

Rényi's quadratic entropy for a continuous random variable X

$$H_2(X) = -\log_2 \int_{\mathbf{x}} p(\mathbf{x})^2 \, d\mathbf{x}. \tag{7.5}$$

7.2 Mutual information

The mutual information is a measure of the dependence between two random variables $\mathbf{x}$ and $\mathbf{y}$, i.e., a generalized form of correlation coefficient.

$$I(\mathbf{x}, \mathbf{y}) = \int \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) \, p(\mathbf{y})} \tag{7.6}$$

For independent variables, $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}) \, p(\mathbf{y})$, and the mutual information is $I = 0$. The measure is symmetric in $\mathbf{x}$ and $\mathbf{y}$

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}) \tag{7.7}$$

Expectation value of the mutual information for multinomial variables (Hutter, 2002, eq. 15), see also (Wolf and Wolpert, 1996; Wolf and Wolpert, 1993)

$$\mathsf{E}[I] = \frac{1}{n} \sum_{i,j} n_{ij} \left[\psi(n_{ij} + 1) - \psi(n_i + 1) - \psi(n_j + 1) + \psi(n + 1) \right] \tag{7.8}$$

Chapter 8

Networks

8.1 Types of networks

Table 8.1: Types of graphs

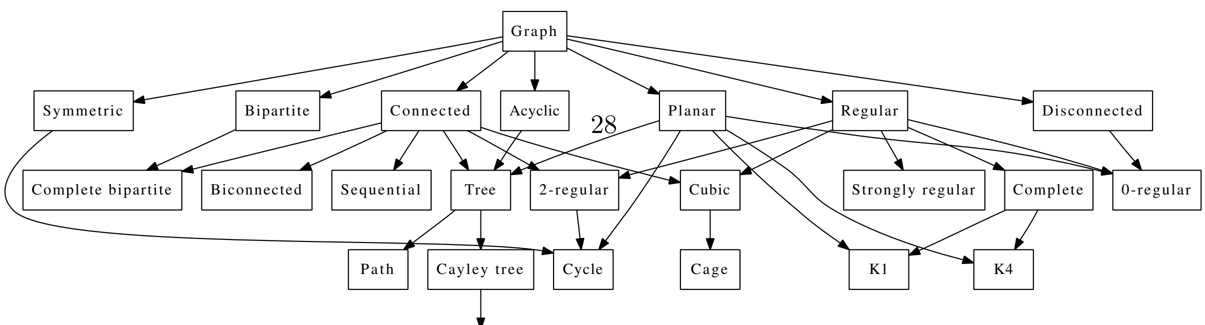
Name	Description	Reference
Causal network	DAG	(Ripley, 1988, p. 262)
Arborescence	Same as "directed tree"	MathWorld
Complete digraph	Digraph with all possible edge	MathWorld
Connected digraph	Can either be weakly or strongly connected	Mathworld
Forest (undirected)	Acyclic graph	
Forest	Polyforest where all nodes have only one parent	David Pennock
Functional graph	Digraph in which each vertex has outdegree one	MathWorld
Moral graph	Undirected graph constructed from a DAG by turning the directed edges to undirected edges and adding edges between parents	(Ripley, 1988, p. 270)
Polyforest	DAG with no undirected cycles	David Pennock
Polytree	Single-connected DAG, i.e., a DAG in which only a single path exists between any vertices. A connected polyforest	(Ripley, 1988, p. 247), David Pennock
Simple directed graph	Digraph with no loops and no multiple edges	MathWorld

8.2 Graph matrices

A graph can be represented as a matrix in a variaty of ways: Adjacency matrix $\mathbf{A}$, incidens matrix $\mathbf{B}$, cycle matrix $\mathbf{C}$ and cut-set matrix (Fould, 1992, chapter 6).

An incidence matrix $\mathbf{B}(N \times P)$ for a graph G with N nodes and P links can be defined in two ways. In the first case the element are set to one if a link p connects to a node n

$$\mathbf{B}_{np} = \begin{cases} 1 & \text{if node } n \text{ and link } p \text{ are incident} \\ 0 & \text{otherwise} \end{cases} \tag{8.1}$$



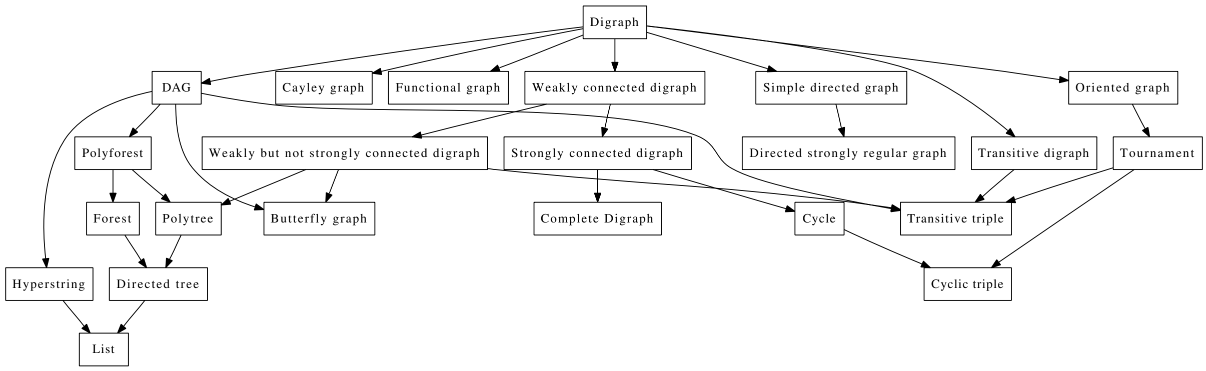


Figure 8.2: Digraph taxonomy

8.2.1 Eigenvalues and vectors

The eigenvalues of a graph is the eigenvalues of its adjacency matrix, and similar for the eigenvectors. The graph sepctrum is the set of eigenvalues, e.g., the vector of eigenvalues λ.

8.3 Eccentricity, radius, diameter

The eccentricity for a node u is the maximum of the shortest distance d to all other nodes

$$e(u) = \max_v d(u,v) \tag{8.2}$$

The diameter of a graph is the maximum eccentricity

$$d = \max_u \left[\max_v d(u,v)\right] \tag{8.3}$$

The radius of a graph is the minimum eccentricity

$$r = \min_u \left[\max_v d(u,v)\right] \tag{8.4}$$

The set of nodes with eccentricity equal to the radius is the graph center.

8.4 Centralization

While centrality is a local measure for an individual node *centralization* is a global measure considering the distribution of links to nodes, e.g., if there is a single node which receives all links the network has a high degree of centralization. On the other hand regular grids and (Poisson-distributed) random network are examples of networks with no centralization.

In economics and ecology the Lorenz curve has been used to describe centrality. The Gini coefficient (or ratio) is one way of summarizing the Lorenz statistics

$$G = \frac{1}{N^2 \bar{x}} \sum_i^N \sum_j^N |x_i - x_j|. \tag{8.5}$$

Here x_j might be the indegree og the j node.

8.5 Spectral gap and expansion

(The **spectral gap** is the smallest non-zero eigenvalue of the Laplacian matrix of a graph.)

8.6 Bridge, hinge and articulation

An **articulation node** (also called articulation vertex, articulation point, cut vertex, cutpoint, hinge, point of articulation) is a node which removal will cause the (remaining) graph to become disconnected or increase the number of disconnected subgraphs. The articulation property is also defined for links with names such as articulation edge, bridge, cut edge, isthmus. Others use the word **bridge** to denote a node that connects to peaks. More generally we might speak of a bridge that is connecting two nodes.

Chapter 9

Receptor kinetics

An introduction to PET and SPECT neuroreceptor kinetics is (Ichise et al., 2001). A general description of compartmental models is (Gunn et al., 2001).

Table 9.1: Receptor kinetics. Partially from (Maguire and Leenders, 2003). A dagger (†) indicates that the symbol is not widely used.

Symbol	Unit	Name	Description, Reference
$B_{\max}$	[mol/l]	Total concentration of receptors	
$B'_{\max}$	[mol/l]	Concentration of available receptors	$B_{\max} - B$ (Innis, 2003, p. 49)
$B'_{\max}$	[fmol/mg]	†Receptor density (after cortical thickness weighting)	The unit is per milligram protein (Forutan et al., 2002)
BP	[ ]	Binding potential	
$\mathrm{BP_G}$	[ ]	†Binding potential (graphical)	BP when determined by a "graphical" method, that is the Logan plot (Biver et al., 1994). To distinguish it from, e.g., $\mathrm{BP_{NLR}}$
C_p	[mol/l]	Concentration in plasma	
f_1	[ ]	Fraction of free ligand in plasma	$f_1 = c_{\text{free}}/c_{\text{total}} = c_{\text{free}}/(c_{\text{free}} + c_{\text{bound}})$ Some part of total ligand might be bound to proteins in the blood
DVR	[ ]	Distribution volume ratio	The ratio between distribution volume in a receptor region and a non-containing receptor (reference) region (Logan et al., 1996)
f_2	[ ]	Fraction of free ligand in tissue	
K_1	[min^{-1}]	Rate (constant) from blood to tissue	
k_2	[min^{-1}]	Rate (constant) from tissue to blood	
k_3	[min^{-1}]	Rate (constant) from free ligand in tissue to bound to specific receptor	
k_4	[min^{-1}]	Rate (constant) from receptor bound to free ligand	
k_5	[min^{-1}]	Rate (constant) from free ligand in tissue to nonreceptor sites (non-specific bound)	
k_6	[min^{-1}]	Rate (constant) from nonreceptor sites to free ligand in tissue	

Symbol	Unit	Name	Description, Reference
K_d	[nM]	Equilibrium dissociation constant. "Affinity" is $1/K_d$.	$K_d = k_{\mathrm{off}}/k_{\mathrm{on}}$ (Innis, 2003, p. 48) (Lammertsma, 2003, p. 63)
k_{off}	[min^{-1}]	Rate of dissociation	(Innis, 2003, p. 48)
k_{on}	[nM^{-1} min^{-1}]	Rate of association	(Innis, 2003, p. 48)
K_i	mol		
R_I	[ ]	"Relative delivery of radiotracer normalized to the cerebellum"	$R_I = K_{1,\mathrm{ROI}}/K_{1,\mathrm{Ref}}$, (Rabiner et al., 2002)
SA	[TBq/mmol]	Specific activity	
V_d	[ ]	Volume of distribution	Also called "distribution volume" (DV) and the same as the partition coefficient C_t/C_p

9.1 Distribution volume

The "distribution volume" or "volume of distribution" is a ratio between tissue concentration and blood concentration of a ligand and can be found abbreviated as V_d or DV

$$V_d = c_t/c_p \tag{9.1}$$

The blood concentration value can be corrected with the fraction of the ligand that is free (Knudsen, 2003, p. 77)

$$V_d = \frac{c_t}{f_1 c_p} \tag{9.2}$$

For a single-tissue compartment model

$$V_d = K_1/k_2 \tag{9.3}$$

For a two-tissue compartment model

$$V_d = \frac{K_1}{k_2}\left(1 + \frac{k_3}{k_4}\right) \tag{9.4}$$

For a three-tissue compartment model (Lammertsma, 2003, p. 66)

$$V_d = \frac{K_1}{k_2}\left(1 + \frac{k_3}{k_4} + \frac{k_5}{k_6}\right) \tag{9.5}$$

This value was refered to as K_{NLR} (for estimation with Non-Linear Regression) in (Biver et al., 1994).

The "distribution volume ratio" is actually an estimate of the binding potential.

9.2 Binding potential

The binding potential (Lammertsma, 2003, p. 64), Mintun 1984

$$\mathrm{BP} = k_3/k_4 \tag{9.6}$$

$$\mathrm{BP} = f_2 B_{\mathrm{max}}/K_d \tag{9.7}$$

The "SRTM" model (Rabiner et al., 2002, p. 622)

$$\mathrm{BP} = \frac{B_{\mathrm{max}} f_2}{K_D\left(1 + \sum f_i/K_i\right)} \tag{9.8}$$

The **distribution volume ratio** is a estimate of the binding potential using a two-tissue model with a reference region with no specific receptor binding

$$\mathrm{DV}_{\mathrm{ratio}} = \frac{\mathrm{DV}}{\mathrm{DV}_{\mathrm{reference}}}, \tag{9.9}$$

where DV is taken from equation 9.4 (as V_d) and the distribution volume of the reference is

$$\mathrm{DV}_{\mathrm{reference}} = \frac{K_1}{k_2}. \tag{9.10}$$

One is substracted to give an estimate of the binding potential

$$\mathrm{BP} = \frac{k_3}{k_4} = \mathrm{DV}_{\mathrm{ratio}} - 1 \tag{9.11}$$

9.3 Specific binding ratio

$$\mathrm{SBR} = \frac{\int_{t_1}^{t_2} C(t)\, dt}{\int_{t_1}^{t_2} C_R(t)\, dt} - 1, \tag{9.12}$$

where $C(t)$ is the radioactive concentration at time t for the region of interest and $C_R(t)$ is for the reference region. This appears in, e.g., (Farde et al., 1997; Okubo et al., 1999).

9.4 Other values

Incorporation coefficient (Sadzot et al., 1995, p. 789) at time T with plasma activity C_p

$$\mathrm{IQ} = \frac{\mathrm{ROI}(T)}{\int_0^T C_p(t)\, dt} \tag{9.13}$$

Bibliography

Berman, A. and Plemmons, R. J. (1979). *Nonnegative matrices in the mathematical sciences*. Computer Science and Applied Mathematics. Academic Press, New York, New York. ISBN 0120922509.

Berman, A. and Plemmons, R. J. (1994). *Nonnegative Matrices in the Mathematical Sciences*. Classics in Applied Mathematics. SIAM, Philadelphia. ISBN 0898713218.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK. ISBN 0198538642.

Biver, F., Goldman, S., Luxen, A., Monclus, M., Forestini, M., Mendlewicz, J., and Fran c. L. (1994). Multicompartmental study of fluorine-18 altanserin binding to brain $5HT_2$ receptors in humans using positron emission tomograpy. *European Journal of Nuclear Medicine*, 21(9):937–946. PMID: 7995287.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, Inc., New York, Wiley classics library edition. ISBN 0471574287.

Cachin, C. (1997). *Entropy Measures and Unconditional Security in Cryptography*. PhD thesis, Swiss Federal Institute of Technology, Zürich, Switzerland. ftp://ftp.inf.ethz.ch/pub/publications/dissertations/th12187.ps.gz.

Conradsen, K. (1984). *En Introduktion til Statistik*, volume 2A. IMSOR, DTH, Lyngby, Denmark, fourth edition. In Danish.

Curtiss, J. H. (1941). On the distribution of the quotient of two chance variables. *The Annals of Mathematical Statistics*, 12(4):409–421.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement*, 8:321–326.

Erdfelder, E. (1993). BINOMIX: A BASIC program for maximum likelihood analyses of finite and beta-binomial mixture distributions. *Behavior Research Methods, Instruments, & Computers*, 25(3):416–418. http://psychonomic.org/search/view.cgi?id=4597.

Everitt, B. S. and Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall, London.

Farde, L., Suhara, T., Nyberg, S., Karlsson, P., Nakashima, Y., Hietala, J., and Halldin, C. (1997). A PET-study of $[^{11}C]$FLB 457 binding to extrastriatal D_2-dopamine receptors in healthy subjects and antiphychotic drug-treated patients. *Psychopharmacology*, 133(4):396–404. PMID: 9372541. DOI: 10.1007/s002130050420. ISSN 0033-3158.

Forutan, F., Estalji, S., Beu, M., Nikolaus, S., Hamacher, K., Coenen, H. H., Vosberg, H., Müller-Gärtner, H.-W., and Larisch, R. (2002). Distribution of $5HT_{2A}$ receptors in the human brain: comparison of data in vivo and post mortem. *Nuklearmedizin*, 4:197–201. http://www.schattauer.de/zs/nukl/-2002/4/pdf/02040197.pdf.

Fould, L. R. (1992). *Graph Theory Applications*. Universitext. Springer-Verlag, New York. ISBN 0387975993.

Gelfand, A. E. and Solomon, H. (1974). Modeling jury verdict in the american legal system. *Journal of the American Statistical Association*, 69(345):32–37.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computation*. John Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, Maryland, third edition. ISBN 0801854148.

Goutte, C., Nielsen, F. Å., and Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Transactions on Medical Imaging*, 19(12):1188–1201.

Graham, A. (1987). *Nonnegative matrices and Applicable Topics in Linear Algebra*. Ellis Horwood, Chichester, England. ISBN 074580182X.

Griffith, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, 29:637–648.

Gunn, R. N., Gunn, S. R., and Cunningham, V. J. (2001). Positron emission tomography. *Journal of Cerebral Blood Flow and Metabolism*, 21:635–652.

Guo, C.-H. (2003). On a quadratic matrix equation associated with an m-matrix. *IMA Journal of Numerical Analysis*, 23(1):11–27. DOI: 10.1093/imanum/23.1.11. http://www.math.uregina.ca/˜chguo/ima1.pdf.

Hansen, L. K., Nielsen, F. Å., and Larsen, J. (2002). Exploring fMRI data for periodic signal components. *Artificial Intelligence in Medicine*, 25:35–44.

Hansen, L. K., Purushotham, A., and Kim, S.-G. (2003). Testing competing hypotheses about single trial fMRI. In Sommer, F., editor, *Explorative methods in Neuroimaging*, chapter 3, pages 49–62. MIT Publishing Co.. http://www.imm.dtu.dk/pubdb/views/edoc_download.php/2883/ps/imm2883.ps.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, Florida. ISBN 0123363802.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56(3):635–639.

Hutter, M. (2002). Distribution of mutual information. In Dietterich, T. G., Becker, S., and Gharamani, Z., editors, *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference*, volume 1, pages 399–406, Cambridge Massachusetts. MIT Press. http://arxiv.org/abs/cs.AI/0112019. CiteSeer: http://citeseer.ist.psu.edu/hutter02distribution.html. ISSN 1049-5258. ISBN 0262042088.

Ichise, M., Meyer, J. H., and Yonekura, Y. (2001). An introduction to PET and SPECT neuroreceptor quantification models. *Journal of Nuclear Medicine*, 42:755–763.

Innis, R. B. (2003). *In vivo* PET imaging understood from the perspective of *in vitro* receptor binding. In Maguire, R. P. and Leenders, K. L., editors, *PET PHarmacokinetic course manual*, chapter 7, pages 47–58. University of Groningen and McGill University, Groningen, The Netherlands and Montreal, Canada.

Jacobsen, D. J. (2004). Integral transformations, Jacobians and featuring the delta function.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43. Suggests a recursive procedure for getting a centrality index in a network.

Knudsen, G. M. (2003). Steady-state measurements of neuroreceptor binding. In Maguire, R. P. and Leenders, K. L., editors, *PET Pharmacokinetic Course Manual*, chapter 8, pages 76–82. University of Groningen and McGill University, Groningen, The Netherlands and Montreal, Canada.

Lammertsma, A. A. (2003). Receptor kinetics — modelling and practical approach. In Maguire, R. P. and Leenders, K. L., editors, *PET Pharmacokinetic Course Manual*, chapter 8, pages 59–76. University of Groningen and McGill University, Groningen, The Netherlands and Montreal, Canada.

Lee, T.-W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441. CiteSeer: http://citeseer.ist.psu.edu/lee97independent.html.

Logan, J., Fowler, J. S., Volkow, N. D., jack Wang, G., Ding, Y.-S., and Alexoff, D. L. (1996). Distribution volume ratios without blood sampling from graphical analysis of PET data. *Journal of Cerebral Blood Flow and Metabolism*, 16:834–840.

Lowe, S. A. (1999a). The beta-binomial mixture model and its application to TDT tracking and detection. In *Proceedings of the DARPA Broadcast News Workshop*. http://www.nist.gov/speech/publications/darpa99/pdf/tdt240.pdf. CiteSeer: http://citeseer.ist.psu.edu/434751.html.

Lowe, S. A. (1999b). The beta-binomial mixture model for word frequencies in documents with applications to information retrieval. In *6TH European Conference On Speech Communication and Technology*, volume 6, pages 2443–2446. http://www.telecom.tuc.gr/paperdb/eurospeech99/PAPERS/S12O2/L043.PDF.

MacKay, D. J. C. (1999). Maximum likelihood and covariant algorithm for independent component analysis. Version 3.8. ftp://www.inference.phy.cam.ac.uk/pub/mackay/ica.ps.gz.

Madsen, R. E., Kauchak, D., and Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, volume 119 of *ACM International Conference Proceeding Series*, pages 545–552, New York, NY, USA. ACM. http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3996/pdf/imm3996.pdf. ISBN 1595931805.

Maguire, R. P. and Leenders, K. L., editors (2003). *PET Pharmacokinetic course manual*. University of Groningen and McGill University, Groningen, The Netherlands and Montreal, Canada.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London. ISBN 0124712525.

Minka, T. (2003a). The lightspeed matlab toolbox. http://www.stat.cmu.edu/˜minka/software/lightspeed/.

Minka, T. P. (2003b). Estimating a Dirichlet distribution.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):1–34. DOI: 10.1023/A:1007692713085. ISSN 0885-6125.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics: Bayesian inference*, volume 2B. Arnold, London. ISBN 0340529229.

Okubo, Y., Olsson, H., Ito, H., Lofti, M., Suhara, T., Halldin, C., and Farde, L. (1999). PET mapping of extrastriatal D2-like dopamine receptors in the human brain using an anatomic standization technique and $[^{11}C]$FLB 457. *NeuroImage*, 10(6):666–674. DOI: 10.1006/nimg.1999.0502. ISSN 1053-8119.

Pham-Gia, T., Turkkan, N., and Marchand, E. (2006). Density of the ratio of two normal random variables and applications. *Communications in Statistics: Theory and Methods*, 35(9):1569–1591. DOI: 10.1080/03610920600683689.

Rabiner, E. A., Messa, C., Sargent, P. A., Husted-Kjaer, K., Montgomery, A., Lawrence, A. D., Bench, C. J., Gunn, R. N., Cowen, P., and Grasby, P. M. (2002). A database of $[^{11}C]$WAY-100635 binding to 5-HT_{1A} receptors in normal male volunteers: normative data and relationship to methodological, demographic, physiological, and behavioral variables. *NeuroImage*, 15(3):620–632. PMID: 11848705.

Rigouste, L., Cappé, O., and Yvon, F. (2005). Evaluation of a probabilistic method for unsupervised text clustering. In Janssen, J. and Lenca, P., editors, *Applied Stochastic Models and Data Analysis*, pages 114–126, Brest Cedex, France. ENST Bretagne. http://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/114.pdf. ISBN 2908849151.

Ripley, B. D. (1988). *Statistical inference for spatial process*. Cambridge University Press, Cambridge, UK. ISBN 0521352347.

Sadzot, B., Lemaire, C., Maquet, P., Salmon, E., Plenevaux, A., Degueldre, C., Hermanne, J. P., Guillaume, M., Cantineau, R., and D. Comar D, e. a. (1995). Serotonin $5HT_2$ receptor imaging in the human brain using positron emission tomography and a new radioligand, [^{18}F]altanserin: results in young normal controls. *Journal of Cerebral Blood Flow and Metabolism*, 15(5):787–797. PMID: 7673371. ISSN 0271-678X.

Smyth, G. K. (1994). A note on modelling cross correlations: hyperbolic secant regression. *Biometrika*, 81:396–402. http://www.statsci.org/smyth/pubs/sech.pdf.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley, Chichester, UK.

Torkkola, K. (2002). Learning dicriminative feature transforms to low dimensions in low dimensions. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference*, volume 2, pages 969–976, Cambridge Massachusetts. MIT Press.

Triantafyllopoulos, K. (2002). *On Observational Variance Learning for Multivariate Bayesian Time Series and Related Models*. PhD thesis, Department of Statistics, University of Warwick, Conventry, UK. http://www.mas.ncl.ac.uk/˜nkt11/research/thesis/phdthesis.pdf.

Triantafyllopoulos, K. (2003). A generalisation of the inverted Wishart distribution. Research Report STA03-08, School of Mathematics and Statistics, University of Newcastle. http://www.ncl.ac.uk/math/research/publications/statistics/STA03-8.pdf.

Wolf, D. R. and Wolpert, D. H. (1993). Estimating functions of probability distributions from a finite set of samples. part II: Bayes estimators for mutual information, chi-squared, covariance, and other statistics. Technical Report LANL LA-UR-93-833, Los Alamos National Laboratory. http://dino.ph.utexas.edu/˜wolf/papers/mutual/wolf_estmut.ps.gz.

Wolf, D. R. and Wolpert, D. H. (1996). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854.

Index

χ^2 distribution, 5

affinity, 32

Bernoulli distribution, 7
binomial distribution, 7
bridge, 30

Cauchy distribution, 6, 12
centralization, 29
convergent matrix, 17
correlation, 14
    partial correlation coefficient, 14
cut edge, 30

derivative, 21
determinant, 21
    derivative, 21
Dirichlet distribution, 7
distribution
    binomial, 7
    Cauchy, 6, 12
    Dirichlet, 7
    Gaussian ratio, 10, 11
    matrix T, 5
    matrix normal, 4
    multinomial, 7
    multivariate Cauchy, 5
    normal-inverse Wishart, 6
    ratio, 10, 11
    type I generalized matrix T, 5

entropy, 27
    Rényi, 27

Gaussian ratio distribution, 10, 11
Gini coefficient, 29

Helmert matrix, 18

incidence matrix, 28
information, 27
    mutual, 27
integration, 22
inversion
    matrix, 20

Jacobian, 22

Kronecker product, 20

M-matrix, 17
matrix
    convergent, 17
    inverse, 20
    M-matrix, 17
    monomial, 16, 17
    nonnegative, 17
    norm, 19
    normal, 18
    permutation, 18
    special unitary, 18
    spectral radius, 20
    symmetric, 18
    unitary, 18
    Z-matrix, 17
matrix T distribution, 5
matrix normal distribution, 4
monomial matrix, 16, 17
multinomial distribution, 7
multivariate Cauchy distribution, 5
mutual information, 27

nonnegative matrix, 17
norm, 19
normal matrix, 18
Normal-inverse Wishart distribution, 6

partial correlation coefficient, 14
permutation matrix, 18

rank, 19
ratio distribution, 10, 11

specific binding ratio, 33
spectral radius, 20
symmetric matrix, 18

test
    correlation coefficient, 14
trace, 21

unitary matrix, 18

vectorization, 19

Wishart distribution, 5

Z-matrix, 17