

Family Name
First Name
Student Number

UNIVERSITY OF WOLLONGONG
SCHOOL OF MATHEMATICS AND APPLIED STATISTICS

STAT902 : Advanced Data Analysis

Autumn Session Examination 2007

Time Allowed: 3 hours 15 minutes
Number of Questions: 6

DIRECTIONS TO CANDIDATES

1. Each question is to be attempted.
2. The examination paper is printed on both sides.
3. A list of possibly useful facts is provided at the end of the examination paper.

EXAMINATION MATERIALS/AIDS ALLOWED

Pens
Ruler

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM
--

1. (a) Random variables X and Y have joint density function

$$[x, y] = \frac{6(x^2 + 3y)}{11}, \quad 0 < x < 1, 0 < y < 1.$$

Determine the function $g(X)$ that best predicts Y in terms of minimum mean square error: $\text{MSE}\{g(X)\} = E[\{Y - g(X)\}^2]$.

- (b) Let X and Y be two random variables. The best linear predictor (BLP) for Y , given X , is defined to be linear function of X

$$\text{BLP}(Y) = a + bX$$

where a and b are chosen to minimise the mean squared error

$$\text{MSE}(a, b) = E[\{Y - (a + bX)\}^2].$$

Show that the solutions for a and b are:

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad a = E(Y) - bE(X).$$

2. (a) Cyrus Dutta is a statistician in the Loans Department of Fuddrucker's Bank. We would like to make inference about

λ = mean number of errors on home loan applications

at the bank. With the help of an assistant, Kevin Kupper, he collects a simple random sample of 500 applications and counts the number of errors on each one. This leads to the following data set:

```
5 5 7 6 3 9 6 7 6 4 6 4 5 7 4 6 7 4 6 8 6 8 5 3 5 10 5 5 5 5 4 7 8 7 5 7 5 3 7 10 4 6 3 5 7 3 6 9 7 2 7 9 6 4 1 0 7 8 6 9 7 3 6 3 3
12 7 2 9 4 7 3 5 1 2 3 6 2 4 4 5 5 5 1 3 5 5 7 3 7 2 4 3 4 4 2 3 4 2 2 6 7 4 1 4 7 5 6 2 6 3 3 7 3 4 9 7 3 4 4 3 8 1 5 4 3 5 4 5 2 3
5 3 5 6 5 6 5 5 2 3 2 6 9 6 9 4 5 3 2 4 3 8 3 0 11 6 4 4 1 8 6 2 10 2 6 1 5 5 5 7 3 7 2 5 4 7 5 3 3 4 1 5 2 8 3 8 4 6 5 5 5 3 3 4 3
3 2 5 0 4 5 6 3 4 2 4 5 5 2 4 5 5 6 6 4 5 5 5 6 3 5 3 6 9 4 3 1 5 1 1 1 9 2 4 5 3 5 5 2 4 2 9 4 4 4 6 1 6 4 9 7 9 5 3 3 7 8 7 3 6 7
8 7 4 4 2 3 3 7 3 5 5 7 4 3 3 7 3 6 6 4 5 3 6 5 4 6 5 4 2 5 1 6 9 5 3 3 1 6 4 3 4 6 6 1 3 4 1 6 4 8 5 5 3 6 9 8 8 6 6 4 8 7 2 3 6 7 4
9 4 2 6 5 7 7 7 1 9 4 4 4 3 6 3 4 7 5 10 6 4 6 5 5 4 4 6 5 3 2 3 1 10 6 6 3 6 3 3 5 8 6 5 3 4 8 7 3 5 5 4 9 5 7 5 7 3 3 1 5 5 8 4 4
1 4 4 1 7 3 6 0 6 5 4 5 5 2 4 1 6 7 5 10 3 11 9 4 12 6 7 6 7 4 6 6 6 7 3 3 8 1 3 5 2 9 5 4 10 7 3 5 3 3 5 4 4 6 2 7 5 6 6 7 2 3 9 6
6 3 3 7 6 7 5 3 8 5 7 5 1 6 6 6 6 8 7 6 4 6 6 5 9 6 4 5 3 9 1 3 7 3 7 5 3 6 5 4 4 4 8
```

Cyrus decides the model the counts as coming from a Poisson distribution with mean λ :

$$[x; \lambda] = e^{-\lambda} \lambda^x / x!, \quad x = 0, 1, \dots$$

He estimates λ to be $\bar{x} = 4.945$ and a standard error of $\sqrt{\bar{x}/1000} = 0.0703$.

Cyrus then decides to check the Poisson model assumption and starts by forming the following table:

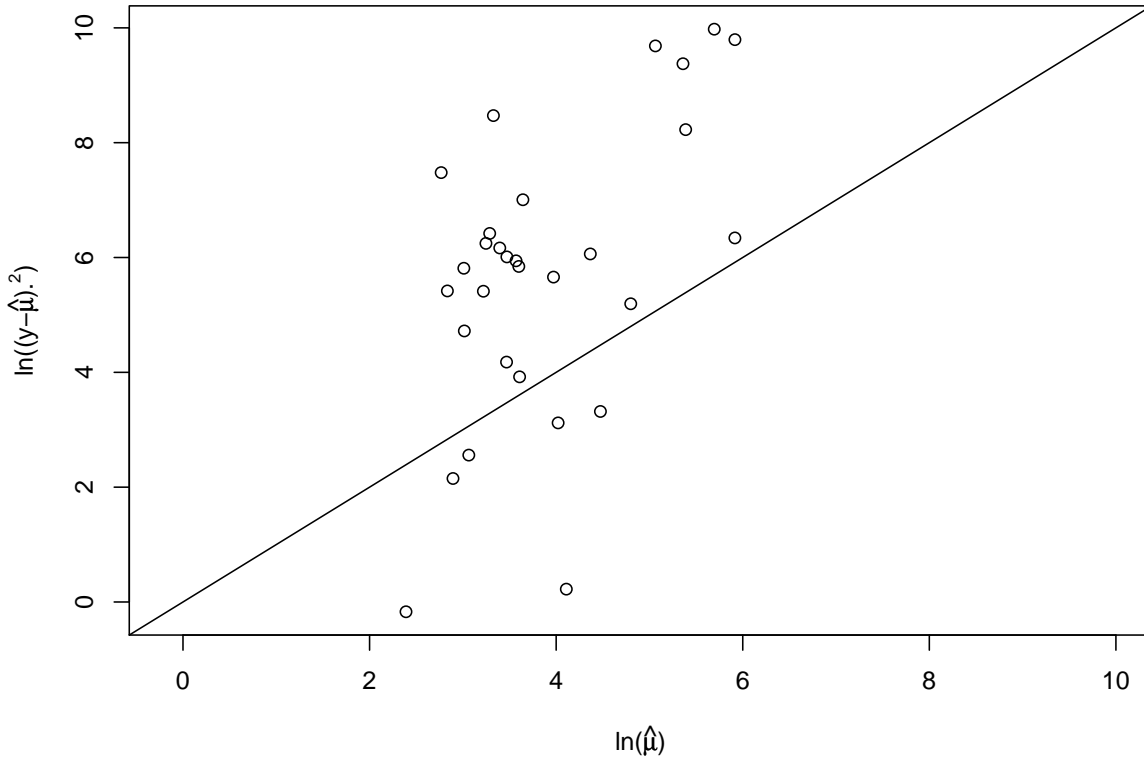
x	proportion	$e^{-4.945} 4.945^x / x!$	$\ln(\text{proportion})$	$\ln(e^{-4.945} 4.945^x / x!)$
0	0.008	0.0073	-4.8	-4.9
1	0.050	0.0361	-3.0	-3.3
2	0.064	0.0887	-2.7	-2.4
3	0.162	0.1452	-1.8	-1.9
4	0.158	0.1784	-1.8	-1.7
5	0.180	0.1753	-1.7	-1.7
6	0.154	0.1436	-1.9	-1.9
7	0.110	0.1008	-2.2	-2.3
8	0.044	0.0619	-3.1	-2.8
9	0.046	0.0338	-3.1	-3.4
10	0.014	0.0166	-4.3	-4.1
11	0.004	0.0074	-5.5	-4.9
12	0.004	0.0030	-5.5	-5.8
13	0.002	0.0011	-6.2	-6.8

The proportion column is the proportion of each x value in the sample (e.g. the proportion of $x = 7$ values is $55/500=0.110$).

The final step of Cyrus' diagnostic check involves making a plot of the last two columns of the above table against one another, with the 1:1 line (i.e. $y = x$) added.

- i. Explain the underlying rationale behind this particular plot, given Cyrus' objective.
- ii. Make the plot using a pen and ruler and comment on the reasonableness of the Poisson assumption.

Figure 1: Plot of $\ln\{(\mathbf{y} - \hat{\boldsymbol{\mu}})^2\}$ versus $\hat{\boldsymbol{\mu}}$ values.



- (b) Four months later Cyrus conducts another statistical analysis for count data on loan defaults, stored in the vector \mathbf{y} ; this time with several predictor variables corresponding to matrix \mathbf{X} . He fits the Poisson regression model:

$$[\mathbf{y}; \boldsymbol{\beta}] = \exp\{\mathbf{y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{1}^T e^{\mathbf{X} \boldsymbol{\beta}} - \mathbf{1}^T \ln(\mathbf{y}!)\}$$

In order to check the suitability of this model Cyrus obtains a plot of $\ln\{(\mathbf{y} - \hat{\boldsymbol{\mu}})^2\}$ versus $\hat{\boldsymbol{\mu}}$ values; where $\hat{\boldsymbol{\mu}} = \exp(\mathbf{X} \hat{\boldsymbol{\beta}})$ is the vector of fitted values on the mean scale; and adds the 1:1 line. The plot is shown in Figure 1.

- i. Explain the underlying rationale behind this particular plot, given Cyrus' objective.
- ii. Explain, in 3-4 sentences (which could include some mathematics), how Cyrus should adjust his model.

3. (a) Consider general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \quad (1)$$

where \mathbf{y} is an $n \times 1$ vector of responses, \mathbf{X} is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients and $\boldsymbol{\varepsilon}$ is a $\sigma_\varepsilon^2 \mathbf{I}$.

- i. The profile log-likelihood for estimation of σ_ε^2 in general linear models (1) is

$$\ell_P(\sigma_\varepsilon^2) = -\text{RSS}/(2\sigma_\varepsilon^2) - \frac{n}{2} \ln(2\pi\sigma_\varepsilon^2)$$

where $\text{RSS} = \mathbf{y}^T \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{y}$ is the residual sum of squares. Show that maximisation of the profile log-likelihood leads to

$$\hat{\sigma}_{\varepsilon, \text{ML}}^2 = \frac{1}{n} \text{RSS}.$$

- ii. The *restricted* profile log-likelihood for estimation of σ_ε^2 in general linear models (1) is

$$\ell_R(\sigma_\varepsilon^2) = \ell_P(\sigma_\varepsilon^2) - \frac{1}{2} \ln |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|$$

where \mathbf{V} is the covariance matrix of \mathbf{y} . Obtain an explicit expression for $\hat{\sigma}_{\varepsilon, \text{REML}}^2$, the maximiser of the residual profile log-likelihood.

- (b) Consider the Gamma generalised linear model (GLM) with canonical link and dispersion parameter ϕ . The linear predictor is

$$\boldsymbol{\eta} = -1/(\mathbf{X}\boldsymbol{\beta})$$

where \mathbf{X} is a design matrix and $\boldsymbol{\beta}$ is a vector of regression coefficients. The log-likelihood for $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}) = \{\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \mathbf{1}^T \ln(-\mathbf{X}\boldsymbol{\beta})\}/\phi + \mathbf{1}^T c(\mathbf{y}, \phi)$$

where $c(\mathbf{y}, \phi)$ depends only on \mathbf{y} and ϕ , but not on $\boldsymbol{\beta}$.

- i. Let $\hat{\boldsymbol{\beta}}_0$ be an initial guess at the maximiser of $\ell(\boldsymbol{\beta})$. Obtain an expression for $\hat{\boldsymbol{\beta}}_1$, the update of $\hat{\boldsymbol{\beta}}_0$ after applying one iteration of the Newton-Raphson procedure.
- ii. Let $\hat{\boldsymbol{\beta}}$ be the maximiser of $\ell(\boldsymbol{\beta})$ (obtained, say, after iteration of the Newton-Raphson procedure until convergence). Describe how to obtain vector of corresponding standard error estimates.

4. Barbara Oakes is an otology clinician interested in determining risk factors for the ear disorder *pyscalbriofrogars*. Left and right ear measurements are taken on 923 patients. You are enlisted to help Barbara with the analysis of her data, and start by postulating the Bayesian logistic mixed model

$$\text{logit}\{P(y_{ij}|\boldsymbol{\beta}, U_i) = \boldsymbol{\beta}^T \mathbf{x}_i + U_i, \quad 1 \leq i \leq 923, j = 1, 2$$

where y_{i1} is an indicator of presence of the disorder in the left ear of the i th patient; y_{i2} is defined similarly for right ears; \mathbf{x}_i is a vector of predictor observations for patient i and $\mathbf{u} = [U_1, \dots, U_{923}]^T$ is a vector of random intercepts. Using the usual matrix notation, the following Bayesian Poisson mixed model is postulated:

$$[\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_U^2] = \exp\{\mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T \ln(1 + e^{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}})\}.$$

$$[\mathbf{u}|\sigma_U^2] \sim N(\mathbf{0}, \sigma_U^2 \mathbf{I})$$

$$[\boldsymbol{\beta}] \sim N(\mathbf{0}, 10^{10}), \quad [\sigma_U^2] \sim \text{IG}(0.001, 0.001).$$

- (a) Determine the distribution of $[\sigma_U^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}]$
 (b) Suppose that

$$[\boldsymbol{\beta}|\sigma_U^2, \mathbf{u}, \mathbf{y}] \sim H(\sigma_U^2, \mathbf{u}, \mathbf{y}) \quad \text{and} \quad [\mathbf{u}|\sigma_U^2, \boldsymbol{\beta}, \mathbf{y}] \sim K(\sigma_U^2, \boldsymbol{\beta}, \mathbf{y})$$

for some distributions H and K . Write down the procedure for generation of samples of size 200, after a burn-in of size 20, from the posterior distributions $[\sigma_U^2|\mathbf{y}]$, $[\boldsymbol{\beta}|\mathbf{y}]$ and $[\mathbf{u}|\mathbf{y}]$ via Gibbs sampling.

- (c) Suppose that the sorted $[\sigma_U^2|\mathbf{y}]$ sample (of size 200) is:

2.140	2.276	2.302	2.509	3.140	3.312	3.324	3.368	3.571	3.573
3.652	3.654	3.720	3.753	3.916	3.927	4.023	4.222	4.284	4.333
4.476	4.579	4.586	4.601	4.619	4.671	4.677	4.727	4.769	4.857
4.965	4.981	5.027	5.032	5.042	5.132	5.136	5.174	5.194	5.199
5.202	5.269	5.347	5.360	5.384	5.429	5.454	5.492	5.529	5.560
5.574	5.587	5.589	5.624	5.625	5.647	5.732	5.776	5.831	5.849
5.899	5.904	5.906	5.928	5.934	5.979	6.020	6.039	6.064	6.084
6.111	6.120	6.194	6.217	6.227	6.267	6.283	6.294	6.352	6.397
6.417	6.474	6.529	6.555	6.556	6.558	6.564	6.582	6.592	6.619
6.734	6.742	6.811	7.020	7.028	7.032	7.033	7.034	7.166	7.168
7.183	7.218	7.282	7.299	7.314	7.320	7.329	7.330	7.336	7.341
7.366	7.373	7.382	7.420	7.429	7.446	7.486	7.520	7.520	7.531
7.535	7.569	7.608	7.641	7.648	7.706	7.722	7.725	7.738	7.754
7.754	7.782	7.819	7.850	7.949	7.950	7.954	8.050	8.053	8.056
8.155	8.157	8.192	8.223	8.234	8.306	8.311	8.311	8.340	8.394
8.405	8.454	8.456	8.461	8.499	8.520	8.561	8.564	8.564	8.579
8.587	8.617	8.676	8.742	8.832	8.964	9.018	9.054	9.062	9.094
9.136	9.287	9.314	9.349	9.365	9.384	9.451	9.460	9.460	9.480
9.752	9.838	9.841	9.864	9.871	9.988	9.997	10.212	10.374	10.490
10.563	10.611	10.708	10.720	10.743	10.764	10.782	11.257	12.489	12.509

Obtain a 95% credible interval for σ_U^2 .

5. The truncated polynomial spline basis functions are of the form

$$1, x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_K)_+^p$$

While most practical spline-based regression analysis involve $p = 1$ or $p = 2$ or $p = 3$, one could use $p = 0$. Consider a regression problem on the interval $[0,1]$. Figure 2 (a) displays the truncated polynomial basis functions corresponding to the knot locations

$$\kappa_1 = \frac{1}{4}, \kappa_2 = \frac{1}{2}, \kappa_3 = \frac{3}{4} \quad \text{for } p = 0.$$

These functions are labelled

$$T_1, T_2, T_3 \quad \text{and} \quad T_4.$$

Note that all functions in Figure 2 only take on the values $-1, 0$ and 1 , but have been subjected to a slight amount of vertical jittering for display purposes. Additionally, all basis functions are defined over $[0, 1]$, but are plotted only over the regions where they are *non-zero*.

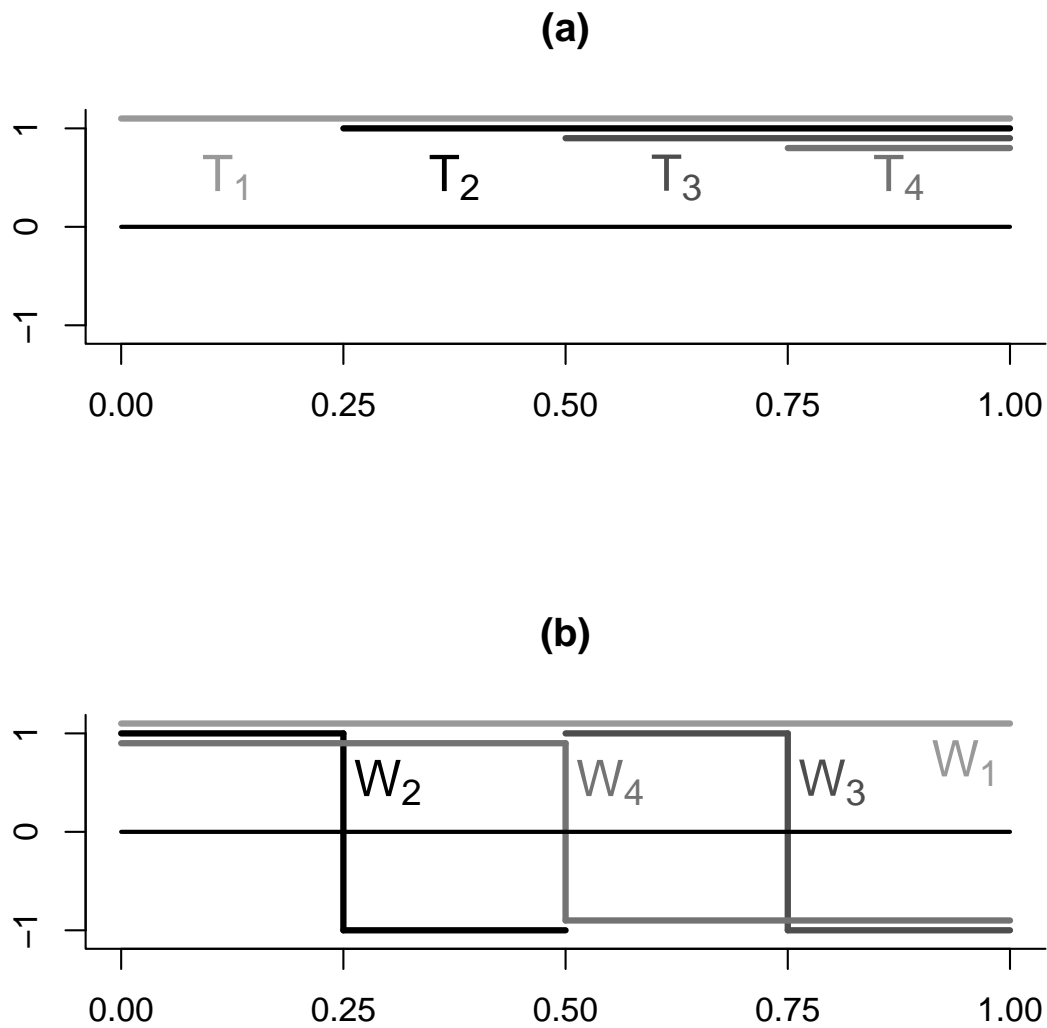
In the usual least regression context this basis can be used to recover any member of the “space” of piecewise step functions on $[0, 1]$, with jumps at $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$.

In Figure 2 (b) an alternative basis for the same space is displayed. These are labelled

$$W_1, W_2, W_3 \quad \text{and} \quad W_4.$$

Here the “W” stands for *wavelet* since the W_j ’s are examples of so-called wavelet functions. An advantage of this basis is that the W_j ’s are mutually orthogonal, which aids computation.

Figure 2: Two sets of basis functions for regression on $[0,1]$.



(a) Clearly

$$W_1 = T_1.$$

One can also show that

$$W_2 = T_1 - 2T_2 + T_3.$$

Write down expressions for W_3 and W_4 in terms of the T_j .

(b) Let (x_i, y_i) , $1 \leq i \leq n$, be the coordinates of a scatterplot, with the x_i 's confined to $[0,1]$. The X-matrix for the regression model

$$y_i = T_1(x_i) + T_2(x_i) + T_3(x_i) + T_4(x_i) + \varepsilon_i \quad (2)$$

is

$$\mathbf{X}_T = [T_1(x_i) \quad T_2(x_i) \quad T_3(x_i) \quad T_4(x_i)]_{1 \leq i \leq n}.$$

Show that the X-matrix, \mathbf{X}_W , for the model

$$y_i = W_1(x_i) + W_2(x_i) + W_3(x_i) + W_4(x_i) + \varepsilon_i \quad (3)$$

is such that

$$\mathbf{X}_W = \mathbf{X}_T \mathbf{L}$$

for some 4×4 matrix \mathbf{L} . Your answer should include the value of \mathbf{L} .

(c) The vector of fitted values for the least squares regression fit to (2) is

$$\hat{\mathbf{y}} = \mathbf{X}_T (\mathbf{X}_T^T \mathbf{X}_T)^{-1} \mathbf{X}_T^T \mathbf{y}$$

Show that $\hat{\mathbf{y}}$ is the same for (3), thereby verifying that least squares regression is invariant to the choice of basis functions (provided that they span the same space).

6. Consider the (non-Bayesian) Poisson random intercept model

$$[y_{ij}|U_i] \stackrel{\text{ind.}}{\sim} \text{Poisson}(e^{\beta_0+U_i}), \quad U_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_U^2)$$

- (a) Obtain an explicit expression for $\text{Var}(y_{ij})$ and simplify as much as possible.
- (b) Obtain an explicit expression for $\text{Cov}(y_{ij}, y_{ij'})$, $j \neq j'$, and simplify as much as possible.
- (c) In the case $\beta_0 = \ln(3)$ and $\sigma_U^2 = \ln(4)$ find the correlation between y_{ij} and $y_{ij'}$ for $j \neq j'$ and simplify as much as possible.

POSSIBLY USEFUL FACTS

If $X \sim \text{Poisson}(\lambda)$ then $E(X) = \lambda$, $\text{Var}(X) = \lambda$ and the moment generating function of X is $m_X(t) = E(e^{tX}) = e^{\lambda(e^t-1)}$.

If $X \sim N(\mu, \sigma^2)$ then $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ and the moment generating function of X is $m_X(t) = E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}$.

The $X \sim \text{IG}(A, B)$ then X has density function

$$[x; A, B] = \frac{B^A}{\Gamma(A)} x^{-A-1} e^{-B/x}, \quad x > 0.$$

If the $d \times 1$ random vector \mathbf{X} has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then the density function of \mathbf{x} is

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

Let X, Y and Z be three random variables. Then

$$\text{Cov}(X, Y) = E\{\text{Cov}(X, Y|Z)\} + \text{Cov}\{E(X|Z), E(Y|Z)\}.$$

Let \mathbf{x} be an $n \times 1$ random vector. The *covariance matrix* of \mathbf{x} is an $n \times n$ matrix, denoted $\text{Cov}(\mathbf{x})$ is

$$\text{Cov}(\mathbf{x}) = E\left[\{\mathbf{x} - E(\mathbf{x})\}\{\mathbf{x} - E(\mathbf{x})\}^T\right].$$

If \mathbf{x} is a random vector, \mathbf{A} is a constant matrix and \mathbf{c} is a constant vector whose dimensions are such that $\mathbf{Ax} + \mathbf{c}$ is defined then

$$E(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}E(\mathbf{x}) + \mathbf{c}$$

and

$$\text{Cov}(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}\text{Cov}(\mathbf{x})\mathbf{A}^T.$$

The updating step for the Newton-Raphson Method for maximising the $S(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^d$ is:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \text{HS}(\mathbf{x}_i)^{-1} \{\text{DS}(\mathbf{x}_i)\}^T.$$

For square matrices \mathbf{A} and \mathbf{B} , $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.

If \mathbf{A} is a $d \times d$ matrix and $a \in \mathbb{R}$ then $|a\mathbf{A}| = a^d|\mathbf{A}|$.