SMF SOLUTIONS TO EXAMINATION. 2012

Q1. (i) **Theorem (Cramér-Rao Inequality)**. Let $Y = u(\mathbf{X})$ be any unbiased estimator of θ. Then the minimum variance bound for $var\ Y$ is

$$var\ Y \geq 1/I(\theta, \mathbf{X}) = 1/(nI(\theta)),$$

where $I(\theta)$ is the information per reading. **[2]**

An unbiased estimator is *efficient* if it attains the Cramér-Rao lower bound. **[2]**

(ii) *Iterative solution of the Likelihood Equation*

It may not be possible to solve the Likelihood Equation (LE) $\ell' = 0$ for the desired MLE $\hat{\theta}$ in closed form. In such cases, we have to proceed iteratively. Begin by drawing a rough graph of ℓ. By inspection, find a rough approximation to the desired root. Call this trial value t. Then (with $s = \ell'$)

$$0 = s(\hat{\theta}) = s(t) + (\hat{\theta} - t)s'(\theta^*),$$

with θ^* between t and $\hat{\theta}$. Solving,

$$\hat{\theta} = t - s(t)/s'(\theta^*). \qquad (*)$$

Fisher's method of scoring. Here we replace $s'(\theta^*)$ by $E[s'(t)] = 1/I(t)$. We know that the MLE $\hat{\theta}$ is strongly consistent: $\hat{\theta} \to \theta_0$ as $n \to \infty$, so for large n $\hat{\theta} \sim \theta_0$; so if t is close enough to θ_0 (all iterations need a close enough starting value), $t \sim \theta_0$, so also $\theta^* \sim \theta_0$. So $I(t) \sim I(\theta_0)$, so by $(*)$,

$$\hat{\theta} \sim t - s(t)/E[s'(t)] = t + s(t)I(t). \qquad (**)$$

This is our next (better) approximation. **[8]**

(iii) *Cauchy location family.*

$$f(x;\mu) = \frac{1}{\pi(1 + (x-\mu)^2)}, \qquad \ell = \log f = c - \log[1 + (x-\mu)^2],$$

$$\ell' = \frac{2(x-\mu)}{1 + (x-\mu)^2}, \qquad s(\mu) := \ell'(\mathbf{x};\mu) = 2\sum_1^n \frac{(x_i - \mu)}{1 + (x_i - \mu)^2}.$$

The information per reading here is constant:

$$I(\mu) = E[(\ell')^2] = \int (\partial f/\partial\mu)^2 f = \frac{4}{\pi}\int \frac{(x-\mu)^2}{[1 + (x-\mu)^2]^3}dx = \frac{4}{\pi}\int \frac{x^2}{[1+x^2]^3}dx = \frac{4}{\pi}I,$$

say, which we can evaluate by Complex Analysis as $\frac{1}{2}$. We can then use these values of $s(\mu)$, $I(\mu) = \frac{1}{2}$ in $(**)$. **[8]**

(Seen – lectures)

Q2. (i) *The delta method.* Given

$$\sqrt{n}(T_n - \theta) \to N(0, \sigma(\theta)^2),$$

with T_n the MLE $\hat{\theta}$ based on a sample of size n and $\sigma^2(\theta) = 1/I(\theta)$. Then by the Mean Value Theorem

$$g(T_n) - g(\theta) = (T_n - \theta)(g'(\theta) + \epsilon_n) = (T_n - \theta)g'(\theta^*),$$

with ϵ_n a (random) error term and θ^* between T_n and θ. From the given result, $T_n \to \theta$, so $\theta^* \to \theta$ also. So by continuity of g', $g'(\theta^*) \to g'(\theta)$ (and $\epsilon_n \to 0$). So

$$g(T_n) - g(\theta) \sim (T_n - \theta)g'(\theta).$$

Since $var(cX) = c^2\ var\ X$,

$$\sqrt{n}(g(T_n) - g(\theta)) \to N(0, [g'(\theta)\sigma(\theta)]^2) \qquad [\mathbf{8}]$$

(ii) *Invariance and Jeffreys priors.* Suppose we work with a parameter θ, with information per reading $I(\theta) = E[(\ell'(\theta)^2] = \int((\log f)_\theta)^2 f(\theta)$. If we reparametrise to $\phi := g(\theta)$, then as $\partial/\partial\phi = (d\theta/d\phi)(\partial/\partial\theta)$,

$$I(\phi) = (d\theta/d\phi)^2 I(\theta).$$

As in maximum-likelihood estimation, we choose a prior which is large where the information is large; the Jeffreys prior is

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

Then

$$\pi(\phi)d\phi \propto \sqrt{I(\phi)}d\phi = \sqrt{I(\theta)}d\theta \propto \pi(\theta)d\theta : \qquad \pi(\phi)d\phi = \pi(\theta)d\theta$$

(both sides integrate to 1, so we can take equality here). So the Jeffreys prior is invariant under reparametrisation. **[8]**

(iii) The variance adds over independent (or even uncorrelated) summands, so has much better mathematical properties than its square root, the standard deviation (SD). But, the SD has the same dimensions (units) as the data, and this is much better for data involving dimensions (anything except scalars). So it is better to use both, and convenient to be able to pass between them as above. **[4]**

((i) and (ii): Seen, lectures)

Q3 (*Sufficiency for the multivariate normal*). Given a sample $x_1, \ldots, x_n$ from a multivariate distribution, form the *sample mean* (vector) and the *sample covariance matrix* as in the one-dimensional case:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad [\mathbf{2}]$$

$$S := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T. \qquad [\mathbf{2}]$$

(i) The *multivariate normal distribution* (in d dimensions) $N(\mu, \Sigma)$ (μ a d-vector, Σ an $d \times d$ symmetric positive definite matrix) has density (Edgeworth's Theorem)

$$f(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{1}{2}d}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}.$$

The likelihood for a sample of size 1 is

$$L(x|\mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\},$$

so the likelihood for a sample of size n is

$$L = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\frac{1}{2}\sum_{1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}.$$

Writing $x_i - \mu = (x_i - \bar{x}) - (\mu - \bar{x})$,

$$\sum_{1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = \sum_{1}^{n}(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x}) + n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)$$

(the cross-terms cancel as $\sum(x_i - \bar{x}) = 0$). The summand in the first term on the right is a scalar, so is its own trace. Since $trace(AB) = trace(BA)$ and $trace(A + B) = trace(B + A)$,

$$trace(\sum_{1}^{n}(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x})) = trace(\Sigma^{-1}\sum_{1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T)$$
$$= trace(\Sigma^{-1}.nS) = n\ trace(\Sigma^{-1}S).$$

Combining,

$$L = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\frac{1}{2}n[trace(\Sigma^{-1}S) + (\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)]\}. \quad [\mathbf{12}]$$

So by the Fisher-Neyman Theorem, $(\bar{X}, S)$ is sufficient for (μ, Σ). **[4]**
(Seen – lectures)

Q4. $ARMA(1,1)$.

$$X_t = \phi X_{t-1} + \epsilon_t + \theta\epsilon_{t-1}: \qquad (1-\phi B)X_t = (1+\theta B)\epsilon_t.$$

Condition for stationarity and invertibility: $|\phi| < 1$; $|\theta| < 1$. **[2, 2]**

Assuming these:

$$X_t = (1-\phi B)^{-1}(1+\theta B)\epsilon_t = (1+\theta B)(\sum_0^\infty \phi^i B^i)\epsilon_t$$

$$= \epsilon_t + \sum_1^\infty \phi^i B^i \epsilon_t + \theta\sum_0^\infty \phi^i B^{i+1}\epsilon_t = \epsilon_t + (\theta+\phi)\sum_1^\infty \phi^{i-1}B^i\epsilon_t:$$

$$X_t = \epsilon_t + (\phi+\theta)\sum_{i=1}^\infty \phi^{i-1}\epsilon_{t-i}.$$

Variance: lag $\tau = 0$. Square and take expectations. The ϵs are uncorrelated with variance σ^2, so

$$\gamma_0 = varX_t = E[X_t^2] = \sigma^2 + (\phi+\theta)^2\sum_1^\infty \phi^{2(i-1)}\sigma^2$$

$$= \sigma^2 + \frac{(\phi+\theta)^2\sigma^2}{(1-\phi^2)} = \sigma^2(1-\phi^2+\phi^2+2\phi\theta+\theta^2)/(1-\phi^2):$$

$$\gamma_0 = \sigma^2(1+2\phi\theta+\theta^2)/(1-\phi^2) \qquad [8]$$

Covariance: lag $\tau \geq 1$.

$$X_{t-\tau} = \epsilon_{t-\tau} + (\phi+\theta)\sum_{j=1}^\infty \phi^{j-1}\epsilon_{t-\tau-j}.$$

Multiply the series for X_t and $X_{t-\tau}$ and take expectations:

$$\gamma_\tau = cov(X_t, X_{t-\tau}) = E[X_tX_{t-\tau}],$$

$$= \{[\epsilon_t + (\phi+\theta)\sum_{i=1}^\infty \phi^{i-1}\epsilon_{t-i}].[\epsilon_{t-\tau} + (\phi+\theta)\sum_{j=1}^\infty \phi^{j-1}\epsilon_{t-\tau-j}]\}.$$

The ϵ_t-term in the first [.] gives no contribution. The i-term in the first [.] for $i = \tau$ and the $\epsilon_{t-\tau}$ in the second [.] give $(\phi+\theta)\phi^{\tau-1}\sigma^2$. The product of the i term in the first sum and the j term in the second contributes for $i = \tau + j$; for $j \geq 1$ it gives $(\phi+\theta)^2\phi^{\tau+j-1}.\phi^{j-1}.\sigma^2$. So

$$\gamma_\tau = (\phi+\theta)\phi^{\tau-1}\sigma^2 + (\phi+\theta)^2\phi^\tau\sigma^2\sum_{j=1}^\infty \phi^{2(j-1)}.$$

The geometric series is $1/(1-\phi^2)$ as before, so for $\tau \geq 1$

$$\gamma_\tau = \frac{(\phi+\theta)\phi^{\tau-1}\sigma^2}{(1-\phi^2)}.[1-\phi^2+\phi(\phi+\theta)]: \qquad \gamma_\tau = \sigma^2(\phi+\theta)(1+\phi\theta)\phi^{\tau-1}/(1-\phi^2).$$

This decreases geometrically beyond the first term, and this behaviour is indicative of $ARMA(1,1)$. **[8]**

(Seen – lectures and problems)

Q5. (i) The joint MGF is

$$M(u,v) := E\exp\{u^TAx + iv^TBx\} = E\exp\{(A^Tu + B^Tv)^Tx\}.$$

This is the MGF of x at argument $t = A^Tu + B^Tv$, so

$$M(u,v) = \exp\{(u^TA+v^TB)\mu+\frac{1}{2}[u^TA\Sigma A^Tu+u^TA\Sigma B^Tv+v^TB\Sigma A^Tu+v^TB\Sigma B^Tv]\}.$$

This factorises into a product of a function of u and a function of v iff the two cross-terms in u and v vanish, that is, iff $A\Sigma B^T = 0$ and $B\Sigma A^T = 0$; by symmetry of Σ, the two are equivalent. **[4]**
(ii) $P^2 = A(A^TA)^{-1}A^T.A(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T = P$;
$(I-P)^2 = I - 2P + P^2 = I - 2P + P = I - P$.
So P, $I - P$ are both (symmetric) projections. **[4]**
(iii) Recall that $tr(A + B) = tr(A) + tr(B)$, and that $tr(AB) = tr(BA)$. So

$$trace(I - AC^{-1}A^T) = trace(I) - trace(AC^{-1}A^T).$$

But $trace(I) = n$ (as here I is the $n\times n$ identity matrix), and as $trace(AB) = trace(BA)$, $trace(AC^{-1}A^T) = trace(C^{-1}A^TA) = trace(I) = p$, as here I is the $p \times p$ identity matrix. So $trace(I - AC^{-1}A^T) = n - p$. **[4]**
(iv) If λ is an eigenvalue of B, with eigenvector x, $Bx = \lambda x$ with $x \neq 0$. Then

$$B^2x = B(Bx) = B(\lambda x) = \lambda(Bx) = \lambda(\lambda x) = \lambda^2 x,$$

so λ^2 is an eigenvalue of B^2 (always true – i.e., does not need idempotence). So

$$\lambda x = Bx = B^2x = \ldots = \lambda^2 x,$$

and as $x \neq 0$, $\lambda = \lambda^2$, $\lambda(\lambda - 1) = 0$: $\lambda = 0$ or 1. The trace is the sum of the eigenvalues, which is r if there are r eigenvalues 1, i.e. when the rank is r. So trace = rank. **[4]**
(v) Because P is a projection of rank r, it has r eigenvalues 1 and the rest 0. We can diagonalise it by an orthogonal transformation to a diagonal matrix with r 1s on the diagonal, followed by 0s. So the quadratic form x^TPx can be reduced to a sum of r squares of standard normal variates, $y_1, \ldots, y_r$. These are independent $N(0,\sigma^2)$ (if $y = Ox$ with O orthogonal and the x_i iid $N(0,1)$, then the y_i are also iid $N(0,1)$: for, the joint density of the x_i involves only $\|x\|$, which is preserved under an orthogonal transformation). So $x^TPx = y_1^2 + \ldots y_r^2$ is σ^2 times a $\chi^2(r)$-distributed random variable. **[4]**
(Seen – lectures)

Q6. In principal components analysis (PCA), we seek a dimension reduction, say from p to k. The covariance (or correlation) matrix Σ can be written by Spectral Decomposition as

$$\Sigma = \Gamma\Lambda\Gamma^T,$$

where $\Lambda = diag(\lambda_i)$ with $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ are the eigenvalues of Σ and Γ is an orthogonal matrix of corresponding normalised eigenvectors. Then $y_1 := \gamma_1^T(x-\mu)$ is the standardised linear combination (SLC – sums of squares of coefficients = 1) of x with largest variance (λ_1), ...,

$$y_k := \gamma_k^T(x-\mu)$$

the SLC of largest variance (λ_k) uncorrelated with $y_1, \ldots, k_{k-1}$. Then the proportion of the total variability explained by the first k PCs is

$$(\lambda_1 + \ldots + \lambda_k)/(\lambda_1 + \ldots + \lambda_p).$$

We continue to retain PCs until we are satisfied that this fraction is acceptably high. We then use these k PCs as a parsimonious summarisation in k dimensions of the data in p dimensions. **[10]**

We need to choose, *before* doing PCA, whether to work with covariances or with correlations. One prefers covariances when the units in which the data are measured are meaningful, correlations otherwise. **[2]**

Examples with correlations. Typically, data are given in terms of prices, and these are meaningful – they are expressed directly in terms of money. But what matters to an investor now is whether the stock will appreciate or depreciate. The actual amounts he cares about are the amounts he will *invest* in the various candidate stocks, and the number of stocks he holds in the company is simply the ratio of his stake to the stock price. Similarly, with foreign exchange, the units of currency in different countries may be of different orders of magnitude. Similarly for an investor dividing his holdings between different economic sectors: what counts here is proportions. **[4]**

Examples with covariances. Examples where the units are meaningful include the internal accounts of a company, where different departments, or activities, contribute to the overall company accounts and balance sheet: all entries are in terms of money, and relate directly to profit and loss.

Empirical evidence suggests that in managing a portfolio of a range of stocks, covariances are better than correlations. **[4]**

(Seen – lectures) NHB