Modelling Multiple Time Series with Missing Observations

By

Cheung King Chau

Student# 9151426

Supervisors

Dr. Don Poskitt and Dr. Ray Chambers

SUBMITTED AS A PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF STATISTICS
at the
AUSTRALIAN NATIONAL UNIVERSITY
1st June 1993

This dissertation is the original work of the author. All sources, authors or other material to which reference was made have been duly acknowledged.

Simon Cheung (Mr.)

Acknowledgement

Having spent over one year studying Statistics in the Australian National University, I have had big improvement in theoretical thinking, mathematical skills and computing skills. Being extremely interested in time series analysis, the working of this thesis allows me to go one step further to this interesting and important topic.

I must attribute the success of this thesis to my supervisors, Dr. Don Poskitt and Dr. Ray Chambers. They have been a consistent source of encouragement, criticism and guidance, without whose patience and support this thesis would not have been possible.

I wish to express my gratitude to Dr. W. K. Li, my first supervisor of my undergraduate studies at the University of Hong Kong. His advice has cultivated my interest in the field of time series analysis.

This thesis is dedicated to my parents to whom I owe my lifelong indebtedness.

Abstract

This thesis introduces an approach to the state space modelling of time series that may possess missing observations. The procedure starts by estimating the autocovariance sequence using an idea proposed by Parzen(1963) and Stoffer(1986). Successive Hankel matrices are obtained via Autoregressive approximations. The rank of the Hankel matrix is determined by a singular value decomposition in conjunction with an appropriate model selection criterion. An internally balanced state space realisation of the selected Hankel matrix provides initial estimate for maximum likelihood estimation. Finally, theoretical evaluation of the Fisher information matrix with missing observations is considered.

The methodology is illustrated by applying the implied algorithm to real data. We consider modelling the white blood cell counts of a patient who has Leukaemia. Our modelling objective is to be able to describe the dynamic behaviour of the white blood cell counts.

Contents

1. An Overview and a Summary of the thesis

1.1 Summary of the Thesis

The aim of this thesis is to develop a method of fitting state space models to multivariate time series data containing missing observations. The model is illustrated by using it to model the dynamic behaviour of the white blood cell counts of a Leukaemia patient. These data consist of measurements of white blood cell counts as well as other auxiliary information, all of which are categorical in nature. These data are defined in Table S.1 below and figure 1.1 presents the graphs for the blood cell count data y_t. (see also Chap.4 on P.67) The proportion of missing observations of all three main variables are about 33.5%. However, they are not necessarily missing at the same time, i.e. partial missing is possible.

Since the three white blood cell counts are correlated, a multivariate model for y_t=(NC,LC,OWBC)' where OWBC=TWBC-NC-LC is developed in what follows. A major problem with the modelling process is the fact that there are missing data in these cell counts.

1.2 Treatment of Missing Data

Throughout it is assumed that the missing data generating process is ignorable. By ignorability, we mean that the conditional expectation of the missing data flag defined as

$$R_t = \begin{cases} 1 \text{ if } y_t \text{ is missing} \\ 0 \text{ if } y_t \text{ is observed} \end{cases} \qquad \text{(F.1)}$$

given the auxiliary information Z_t is independent of $y_t|Z_t$ (see Rubin and Little(1987)). The assumption of ignorability enables us to partition the joint log-likelihood of $(y_t,R_t)|Z_t$ into the log-likelihood of $y_t|Z_t$ plus that of $R_t|Z_t$ provided that they have

Table S.1

Main Variables	Total White Blood Cell Count (TWBC) Neutrophyll Count (NC) Lymphocyte Count (LC)
Auxiliary Information	Infection Observed (0: no infection, 1: infected) Stage of Treatment (Stage from 1 to 4) Cycle within Stage (at most 8 cycles per stage) Week within Cycle Treatment Break (0: no break, 1: break) Time elasped since beginning of treatment

distinct parameters. We, thus, are able to 'ignore' the information contained in the log-likelihood due to $R_t|Z_t$. However, a drawback of this assumption is that all future analysis must be conditional on the auxiliary information Z_t. (see section 3.1 on P.41)

One simple way to justify the assumption of ignorability is to check whether the distribution of the conditional variable $R_t|Z_t$ is constant or nearly constant. This can be based on a preliminary logistic analysis of the missing data flag in terms of the auxiliary variables. A logistic analysis of the missing data flag for the white blood cell counts data in terms of the auxiliary variables indicates a reduction of 20%-30% in residual deviance. Although this reduction is not outstanding, it is sufficient to indicate that the assumption that the missing data generating process is ignorable is a reasonable one.

(see P.45-46; section 4.1 on P.68)

Definition of the Model

In the following development, the data will be assumed to follow the state space model

$$\begin{aligned} y_t &= \beta Z_t + Cx_t + \varepsilon_t \\ x_{t+1} &= Ax_t + B\varepsilon_t \end{aligned} \qquad \text{(F.2)}$$

where

1. Z_t is the vector of auxiliary variables at time t,
2. x_t is the state vector of unknown dimension,
3. $\varepsilon_t \sim N(0,\Omega)$ is i.i.d., and
4. $x_0 \sim N(\mu,\Sigma)$ independent of ε_t.

The development begins by describing a procedure for determining the dimension of the state vector x_t and for computing an initial estimate of the parameter (β,A,B,C,Ω) of the model (F.2). Maximum likelihood estimation based on a modified version of Newton's algorithm is then used to obtain the final model. Finally, confidence intervals for the model parameters are computed based on large sample maximum likelihood theory.

1.3 Removal of the Deterministic Component

Before the state space model (F.2) can be fitted, the data must be 'detrended'. In particular, a robust regression procedure is used to sweep out the effect of the auxiliary variables Z_t from the square-root transformed time series

$$y_t = (\sqrt{NC},\ \sqrt{LC},\ \sqrt{OWBC})'.$$

The use of the square root transformation is to stablise the residual variance. Selection of the auxiliary variables to include as regressors is based on the following statistic which is analogous to the F-statistic in least squares analysis (see Huber(1981), page 196-197):

$$\text{(F.3)} \qquad \frac{\frac{1}{p-q}\,\|\hat{y}_p - \hat{y}_q\|^2}{\frac{1}{n-p}\cdot\frac{K^2\sum\psi(r_i/\hat{\sigma})^2\hat{\sigma}}{\left(\frac{1}{n}\sum\psi'(r_i/\hat{\sigma})\right)^2}}.$$

Here

1. $\hat{y}_p$ and $\hat{y}_q$ are vectors of fitted values using the models which include p auxiliary variables (p-model) and q auxiliary variables (q-model) respectively.
2. r_i is the residual of the p-model.
3. n is the number of observations.
4. $\psi(x) = \min\{c,\max(-c,x)\}$ for some constant $c > 0$.
5. $K = 1 + \frac{p}{n}\cdot\frac{1-m}{m}$, m = the relative frequency of r_i satisfying $|r_i| < c$.
6. $\hat{\sigma}$ is the median absolute deviation of $|r_i|$.

According to Huber(1981), the distribution of the statistic (F.3) is well approximated by that of a χ^2_{p-q} variable.

Application of the above procedure to the blood cell count data was carried out using the robust regression function 'rreg' in Splus with option 'method=Huber'. In doing so, models with different auxiliary variables were compared using the statistic (F.3) and assuming a chi-square distribution as described above. The final set of auxiliary variables selected for Z_t can be seen in page 62-63 of the thesis and the associated

parameter estimates were then used to provide an initial estimate of β. (see section 4.2 on P.70-74)

1.4 Successive AR approximations

After determining an initial estimate of β along the lines described above, the method proceeds by determining successive AR approximations to the detrended time series $\tilde{y}_t = y_t - \beta Z_t$. The pth order AR approximation is defined by fitting the following model to the data:

(F.4) $$\tilde{y}_t = A_1\tilde{y}_{t-1} + A_2\tilde{y}_{t-2} + \cdots + A_p\tilde{y}_{t-p} + \varepsilon_t^p$$

where $\varepsilon_t^p \sim N(0,S_p)$ is i.i.d.. It can be shown that the ε_t's appearing in (F.4) are the same as that appearing in the (F.2)(see section 2.6 on P.25) (see Hannan and Deistler(1988)). There are two standard ways of estimating the coefficients $A_1,\ldots,A_p$ making up this model. These are the Levinson-Whittle (LW) algorithm and Burg's algorithm. Although Burg's algorithm is typically preferred, the presence of missing data implies that only viable option for modelling these data is the LW algorithm, which only requires the autocovariance sequence as input. The LW algorithm leads to estimates of $A_1,\ldots,A_p$ and S_p which we write as $A_1^{(p)},\ldots,A_p^{(p)}$ and S_p^f and which satisfy the Yule Walker equations

$$\sum_{k=0}^{p} A_k^{(p)}\hat{R}_{j-k} = \begin{cases} 0 & \text{if } j > 0 \\ S_p^f & \text{if } j = 0 \end{cases}$$

where $\hat{R}_k$ is the estimated lag-k autocovariance of $\tilde{y}_t$. This algorithm can be found, for example, in Jones(1978) and Hannan(1986). (see section 5.1 on P.76)

The estimated lag-k autocovariance $\hat{R}_k$ is given by for $i,j = 1,2,\ldots,h$,

(F.5) $$\hat{R}_k(i,j) = \frac{\sum_{t=1}^{T-k} a_{t+k}(i)a_t(j)\{\tilde{y}_{t+k}(i)-\underline{\tilde{y}}(i)\}\{\tilde{y}_t(j)-\underline{\tilde{y}}(j)\}}{\sum_{t=1}^{T-k} a_{t+k}(i)a_t(j)}.$$

Here,

$h=\dim(y_t)$, $a_k(i)=\begin{cases} 1 & \text{if } y_k(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$ and $\tilde{\underline{y}}(k)=\dfrac{\sum_{t=1}^{T} a_t(k)\tilde{y}_t(k)}{\sum_{t=1}^{T} a_t(k)}$.

This estimate is known to be consistent when some of the data are missing and the time series $\{a_k\}$ is asymptotically stationary, see Parzen(1963) and Stoffer(1986). (see section 3.3 on P.52-54)

For the blood cell count data, the above procedure was used to fit an AR(p) approximation to the data for p=1,2,...,max where max ≤ loglogT was determined as the estimated dimension of the state vector x_t stopped changing. For each value of p, hp balanced state space representations of AR(p) were then constructed along the lines described below and, finally, all hp×max models thus fitted were compared using the Hannan-Quinn(1980) criterion to determine the final dimension of the state vector x_t.

1.5 Balanced State Space Representation of the AR Models

For each values of p=1,2,..., the AR(p) approximation can be written as $A^{(p)}(z^{-1})\tilde{y}_t = \varepsilon_t^p$ where

$$A^{(p)}(z^{-1}) = I - A_1^{(p)}z^{-1} - \cdots - A_p^{(p)}z^{-p}.$$

By employing an algorithm described in Robinson(1967), p.160-162, $\mathrm{adj}A^{(p)}(z^{-1})$ and $\det A^{(p)}(z^{-1})$ can be determined where

(F.6) $$\left[A^{(p)}(z^{-1}) \right]^{-1} = \frac{\mathrm{adj}\ A^{(p)}(z^{-1})}{\det A^{(p)}(z^{-1})}$$

A method based on partial fraction, see Tranter(1960), can then be used to invert the polynomial $\det A^{(p)}(z^{-1})$ by assuming that it does not possess any repeated roots. Polynomial multiplication can then be used to calculate the coefficients $\pi_r^{(p)}$, $r \geq 0$ of the power series expansion

(F.7) $$\left[A^{(p)}(z^{-1}) \right]^{-1} = \sum_{r\geq 0} \pi_r^{(p)} z^{-r}.$$ (see section 5.2 on P.77-79)

The pth order Hankel matrix $\mathcal{H}_p^{(p)}$ corresponding to the AR(p) approximation is then given by

(F.8)
$$\mathcal{H}_p^{(p)} = \begin{bmatrix} \pi_1^{(p)} & \pi_2^{(p)} & \cdots & \pi_p^{(p)} \\ \pi_2^{(p)} & \pi_3^{(p)} & \cdots & \pi_{p+1}^{(p)} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_p^{(p)} & \pi_{p+1}^{(p)} & \cdots & \pi_{2p-1}^{(p)} \end{bmatrix}$$

In this setting, Hannan and Deistler(1988) proved that the rank of $\mathcal{H}_p^{(p)}$ is $\leq$ hp. Furthermore, Otter(1985) showed that a state space realisation with state vector dimension equal to the rank of $\mathcal{H}_p^{(p)}$ exists provided that this rank is finite. Thus, we require an estimate of the rank v_p, $1 \leq v_p \leq hp$, of the Hankel matrix $\mathcal{H}_p^{(p)}$ which, in turn, will allow us to estimate the dimension of the state vector x_t.

A convenient way of estimating the rank of a matrix is to carry out a singular value decomposition of the matrix. Let $\mathcal{H}_p^{(p)} = U_p S_p V_p'$ denote the singular value decomposition on $\mathcal{H}_p^{(p)}$. Here U_p and V_p are orthonormal matrices (i.e. $U_p'U_p = V_p'V_p = I$) and $S_p = \text{diag}(s_1,\ldots,s_{hp})$ is the diagonal matrix of singular values of $\mathcal{H}_p^{(p)}$. Due to the noisy nature of the data, the magnitude of these singular values typically decreases to zero very smoothly. We therefore require a criterion for deciding how many of these singular values can be considered effectively zero. Since the ultimate objective is to provide a best fit to the data, the criterion used for this purpose is the well-known consistent criterion suggested by Hannan and Quinn(1980), see Hannan(1986). Effectively, each one of the possible hp balanced state space models corresponding to the AR(p) approximation were computed and that model (and hence value v_p) chosen which minimised this criterion. (see P.81-85)

For v = 1,2,...,hp, each balanced state space model was defined via the steps:

1. Put $S_{1v} = \text{diag}(s_1,\ldots,s_v)$.
2. define the corresponding partitions $U_p = (U_{1p}\ U_{2p})$ and $V_p = (V_{1p}\ V_{2p})$.

3. Put

$$\mathcal{H}_B^{(p)}=\begin{bmatrix}\pi_1^{(p)}\\ \pi_2^{(p)}\\ \vdots\\ \pi_p^{(p)}\end{bmatrix},\quad \mathcal{H}_A^{(p)}=(\pi_1^{(p)}\ \cdots\ \pi_p^{(p)}) \text{ and } \mathcal{H}_p^{(p)\uparrow}=\begin{bmatrix}\pi_2^{(p)} & \cdots & \pi_{p+1}^{(p)}\\ \pi_3^{(p)} & \cdots & \pi_{p+2}^{(p)}\\ \vdots & & \vdots\\ \pi_{p+1}^{(p)} & \cdots & \pi_{2p}^{(p)}\end{bmatrix}.$$

A balanced state space representation of the AR(p) approximation to $\tilde{y}_t$ which assumes $\text{rank}(\mathcal{H}_p^{(p)}) = v$ can then be constructed by setting

$$\text{(F.9)}\qquad \begin{aligned} x_{t+1} &= A_{1v}x_t + B_{1v}\varepsilon_t\\ \tilde{y}_t &= C_{1v}x_t + \varepsilon_t \end{aligned}$$

where

1. $v = \dim(x_t)$
2. $A_{1v} = S_{1v}^{-1/2}U'_{1p}\mathcal{H}_p^{(p)\uparrow}V_{1p}S_{1v}^{-1/2}$
3. $B_{1v} = S_{1v}^{-1/2}U'_{1p}\mathcal{H}_B^{(p)}$
4. $C_{1v} = \mathcal{H}_C^{(p)}V_{1p}S_{1v}^{-1/2}$.

It can be shown that all models constructed in this way are numerically stable, controllable and observable so that the effect due to uncertainty of the initial state x_0 is minimised, see Crabb and Young(1989). For each value of p, therefore, there are hp balanced state space models representing the AR(p) process with the dimension of the state vector x_t ranging from 1 to hp.

The appropriate value of v_p is then chosen so that the Hannan and Quinn(1979) (HQ) criterion

$$\text{(F.10)}\qquad HQ(v) = -2\log\text{-likelihood} + 2v\lceil\log\log T\rceil$$

is minimised among these hp models, where $\lceil\cdot\rceil$ is the ceiling function and -2log-likelihood is evaluated using the Kalman Filter, see Balakrishnan(1984), appropriately modified to allow for missing data. Details of the algorithm are set out below. A result from Hannan(1986) indicates that the HQ criterion is consistent in picking the correct order when applied to state space models.

Applying this procedure for each value of p, values v_p, p=1,2,...,max, are obtained. The order of the state space model (F.2) finally chosen to represent the data is then such that $HQ(v_p)$ is minimum. This balanced state space model is also used to provide initial estimates of the structural parameters A, B, and C. The initial estimate of Ω is provided by the value of S_p^f for the chosen value of v_p.

1.6 Maximum Likelihood Estimation

Having determined the dimension of the state vector of the model (F.2), maximum likelihood estimation of the parameters

$$\theta = (\text{vec}(A)', \text{vec}(B)', \text{vec}(C)', \text{vec}(\Omega^{-1})', \text{vec}(\beta)')'$$

of this model is carried out using numerical optimisation. In turn, this requires an initial estimate of θ as well as algorithms for generating the likelihood surface and its gradient.

The initial estimate of θ is computed along the lines described in the previous sections. Thus, the initial estimate of β is obtained from a robust regression of y_t in terms of the auxiliary variables Z_t, while the initial estimates of the other components of θ are obtained as a by-product of the procedure used to determine the dimension of the state vector x_t.

The computation of the likelihood surface for the model (F.2) is based on the following modified version of the Kalman Filter which allows for missing data:

Modified Kalman Filter

Given an initial estimate of θ, assume that $\mu = 0$ and $\Sigma = \{I-A\otimes A\}^{-1}\text{vec}(B\Omega B')$. The stationarity and stability conditions imply that Σ satisfies $\Sigma = A\Sigma A'+B\Omega B'$ and the effect of setting $\mu=0$ is minimised by assuming that the model (F.2) is observable and controllable.

In order to minimise notational complexity in specifying the Kalman Filter, the following notation has been used

1. $\tilde{y}_1(t) = \begin{cases} \text{Observed part of } \tilde{y}_t \text{ if } \tilde{y}_t \text{ is partially or fully observed} \\ \text{null if } \tilde{y}_t \text{ is completely missing} \end{cases}$
2. $\varepsilon_1(t)$ is the subvector of ε_t corresponding to the observed part of $\tilde{y}_t$.
3. $\hat{x}_t = E\{x_t | \tilde{y}_1(n),\ 1 \le n \le t\}$, $\bar{x}_t = E\{x_t | \tilde{y}_1(n),\ 1 \le n < t\}$.
4. $H_{t-1} = E\{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'\}$
5. $\Omega_{1t} = E\{\varepsilon_1(t)\varepsilon_t'\}$ and $\Omega_{11t} = E\{\varepsilon_1(t)\varepsilon_1(t)'\}$.
6. C_{1t} is the rows of C corresponding to the observed part of $\tilde{y}_t$.
7. λ_n is the -2log-likelihood of the model (F.2) at the given parameter values for the first n observations.

Starting Conditions: $\bar{x}_1 = A\mu$, $H_0 = A\Sigma A' + B\Omega B'$ and $\lambda_0 = 0$.

For n = 1,2,...,T, do the following recursion:

Case 1: When some, not necessarily all, of the components of $\tilde{y}_n$ are observed,

$$Q_n = \Omega_{11n} + C_{1n} H_{n-1} C_{1n}'$$

$$K_n = H_{n-1} C_{1n}' Q_n^{-1}$$

$$P_n = (I - K_n C_{1n}) H_{n-1}$$

$$\hat{x}_n = \bar{x}_n + K_n(\tilde{y}_1(n) - C_{1n}\bar{x}_n)$$

$$\bar{x}_{n+1} = A\hat{x}_n + B\Omega_{1n}' Q_n^{-1}(\tilde{y}_1(n) - C_{1n}\bar{x}_n)$$

$$H_n = AP_nA' + B\Omega B' - AK_n\Omega_{1n}B' - B\Omega_{1n}'K_n'A' - B\Omega_{1n}'Q_n^{-1}\Omega_{1n}B'$$

$$\lambda_n = \lambda_{n-1} + \log|Q_n| + (\tilde{y}_1(n) - C_{1n}\bar{x}_n)'Q_n^{-1}(\tilde{y}_1(n) - C_{1n}\bar{x}_n)$$

Case 2: When $\tilde{y}_n$ is completely missing,

$$\hat{x}_n = \bar{x}_n$$

$$P_n = H_{n-1}$$

$$\bar{x}_{n+1} = A\hat{x}_n$$

$$H_n = AP_nA' + B\Omega B'$$

end the for loop.

The final value of λ, λ_T, is then the -2log-likelihood of the model at the given parameter values. Without missing data, the matrices Ω_{11n}, Ω_{1n}, $\tilde{y}_1(n)$ and C_{1n} appearing in the above recursion are replaced by Ω, $\tilde{y}_t$ and C respectively. In this case, the above recursion is the same as the Kalman Filter described in

Balakrishnan(1984). That is, this modified Kalman Filter is consistent with the Kalman Filter given in Balakrishnan(1984) in the sense that the dimensions of all quantities have been modified to be consistent with the observed data. (see P.30 and P.59 for comparison)

Observe that if we regard the states x_t, t=0,1,...,T, as potentially 'observable', then the 'complete-data' likelihood may be expressed as

$$(F.11) \qquad -2\log L = T\log|\Omega| + \sum_{t=1}^{T} (y_t-\beta Z_t-Cx_t)'\Omega^{-1}(y_t-\beta Z_t-Cx_t).$$

A similar idea was used by Shumway and Stoffer(1982). The score function (or the gradient of the likelihood) is then computed via matrix differentiation of (F.11) and the formula

$$(F.12) \qquad Sc(\theta|\mathcal{O}_S) = E\{Sc(\theta|\mathcal{O}_C)|\mathcal{O}_S\}$$

where $\mathcal{O}_S$ and $\mathcal{O}_C$ are the observed data and complete data sets respectively and $Sc(\cdot|\mathcal{O}_S)$ and $Sc(\cdot|\mathcal{O}_C)$ are the score functions defined by these 'observed' data and 'complete' data sets. (see Louis(1982) and Tanner(1990)) (see section 6.1 on P.88-91)

In order to use (F.12), we need to evaluate the expectation of the 'complete' data score function given the observed data. In turn, this requires that we evaluate quantities such as $\hat{x}_n(T) = E\{x_n|\mathcal{O}_S\}$ and $P_n(T) = E\{(x_n-\hat{x}_n(T))(x_n-\hat{x}_n(T))'\}$. This is carried out using the Kalman Smoother: (see P.30-33)

Kalman Smoother

Initial Conditions: $\hat{x}_T(T) = \hat{x}_T$ and $P_T(T) = P_T$.

For n = T-1,...,0, do the following recursion:

$$S_n = P_n A' H_n^{-1}$$

$$\hat{x}_n(T) = \hat{x}_n + S_n(\hat{x}_{n+1}(T)-\hat{x}_{n+1})$$

$$P_n(T) = P_n + S_n(P_{n+1}(T)-H_n)S_n'$$

end the for loop.

Further details on computation of the score function (F.12) are given in Chapter 6 of the thesis.

Finally, a quasi-Newton optimisation algorithm, see Davidon, Fletcher and Powell(1963), is used to locate the MLE. The advantage of the algorithm over the more conventional Newton-Raphson algorithm is that it does not require inversion of the Hessian matrix at each step, see Rao(1977). (see P.87-88)

An immediate by-product of the Kalman Filter and Kalman Smoother is that one step ahead predictions and interpolated values for missing data are easily computed via the formulae: $\hat{y}_t = \hat{\beta}Z_t + C\bar{x}_t$ and $\hat{y}_t(T) = \hat{\beta}Z_t + C\hat{x}_t(T) + \varepsilon_t(T)$ respectively. Here, $\hat{y}_t = E\{y_t \mid y_1(n), 1 \le n < t\}$, $\hat{y}_t(T) = E\{y_t \mid \mathcal{O}_S\}$ and $\varepsilon_t(T) = E\{\varepsilon_t \mid \mathcal{O}_S\}$. Graphs of these predicted series for the blood cell count data are shown in figure 1.2.

1.7 Estimating the accuracy of the MLE

It is well known that the asymptotic variance of the MLE is proportional to the inverse of the Fisher Information Matrix

(F.13) $$\mathcal{I} = E\{Sc(\theta \mid \mathcal{O}_S)Sc(\theta \mid \mathcal{O}_S)'\}.$$

One disadvantage of this approach is that this expectation must be computed element-wise, which is numerically inefficient. There is scope, therefore, for further research on developing a more efficient algorithm for this problem. For more detail on computation of $\mathcal{I}$, see Chapter 6 and Appendix B of the thesis.

1.8 Discussion

The essential differences between 'complete data' case and 'incomplete data' case are that

1. Large scale matrix computation is a fact of life when missing data are present while this can be largely avoided with complete data. As a consequence, numerical errors can accumulate in the incomplete data case.
2. The presence of missing data introduces time-varying elements into the analysis, increasing the dimensionality of the problem and further increasing problems associated with numerical stability when implementing Kalman Filter.
3. The presence of missing data reduces the amount of information available for fitting the state space model (F.2) and therefore

increases the prediction error of the fitted model.

1.9 Diffuse Kalman Filtering

The regression parameter β in the state space model (F.2) is fixed. In some applications, it may be more reasonable to consider this parameter to be a realisation of a random vector. This leads to the modified model

$$\begin{aligned} y_t &= \beta Z_t + G x_t + \varepsilon_t \\ x_{t+1} &= F x_t + H \varepsilon_t \end{aligned} \tag{F.14}$$

where

1. $\varepsilon_t \sim N(0,\Omega)$ is i.i.d.
2. $x_0 = 0$ and $\beta = b + B\gamma$ with $\gamma \sim N(0,C)$.
3. γ and ε's are uncorrelated.
4. C is nonsingular unless C = 0.

A special case of model (F.14) has been called 'diffuse' by Jong(1991). This is when $C^{-1} = 0$ which, in a Bayesian sense, is similar to the case of a noninformative prior and reflects our ignorance about the parameter β. This section is concerned with maximum likelihood estimation of model (F.14) in such a 'diffuse' situation. In order to carry out the maximum likelihood estimation in such a situation, we require an algorithm for generating the likelihood and an initial estimate of the model parameter

$$\theta = (\text{vec}(F)', \text{vec}(G)', \text{vec}(H)', \text{vec}(b)', \text{vec}(B)', \text{vec}(\Omega)')'.$$

Initial estimates for F, G, H, Ω and b are available from maximum likelihood estimation of the model (F.2). That is, the initial estimates of F, G, H and Ω are provided by the fitted balanced state space model (F.2) while that of b is provided by the robust regression of y_t on the auxiliary variables Z_t. In order to allow the effect of the random variable γ, an initial estimate of B equal to a matrix of 1's is arbitrarily chosen.

Having determined an initial estimate of θ, the likelihood of model (F.14) at a given parameter is evaluated using the diffuse Kalman Filter developed by Jong(1991), again, with appropriate modifications to allow for missing data. As with the ordinary Kalman Filter, these modifications mainly consist of

appropriately reducing the dimensions of the various matrices involved to take account of the missing data, in a way that is consistent with the proof of the diffuse Kalman Filter. (see P.109-110)

Again, for ease of explanation, we adopt the following notation:

1. $y_1(t)$ is the observed part of y_t.
2. B_t and b_t are the rows of B and b respectively which correspond to the observed part of y_t.
3. G_{1t} is the rows of G corresponding to the observed part of y_t.
4. $\varepsilon_1(t)$ is a subvector of ε_t corresponding to the observed part of y_t.
5. $\Omega_{1t} = E\{\varepsilon_1(t)\varepsilon_t'\}$ and $\Omega_{11t} = E\{\varepsilon_1(t)\varepsilon_1(t)'\}$.

The modified diffuse Kalman Filter which allows for missing data is then defined by the following recursion:

Modified Diffuse Kalman Filter

Initial Conditions: $A_1=0$, $Q_1=0$, $\omega=0$ and $\text{vec}P_1=\{I-F\otimes F\}^{-1}\text{vec}(H\Omega H')$.

For t = 1,2,...,T, do the following recursion:

Case 1: When some of the components of y_t are observed,

1. $e_t = (B_tZ_t, y_1(t)-b_tZ_t) - G_{1t}A_t$
2. $D_t = G_{1t}P_tG_{1t}' + \Omega_{11t}$
3. $K_t = (FP_tG_{1t}' + H\Omega_{1t}')D_t^{-1}$
4. $A_{t+1} = FA_t + K_te_t$
5. $P_{t+1} = (F-K_tG_{1t})P_tF'+H\Omega H'-K_t\Omega_{1t}H'$
6. $\omega = \omega + \log|D_t|$
7. $Q_{t+1} = Q_t + e_t'D_t^{-1}e_t$

Case 2: When y_t is completely missing,

1. $A_{t+1} = FA_t$
2. $P_{t+1} = FP_tF' + H\Omega H'$

end the for loop.

After the recursion, -2log-likelihood of the model (F.14) is equal to $\log|S| + \omega + (q-s'S^{-1}s)$ where q, s and S are defined by the partition

(F.15) $$Q_{T+1} = \begin{bmatrix} S & s' \\ s & q \end{bmatrix}$$

with q a scalar.

Given the above algorithm for generating the likelihood and an initial estimate of θ, the quasi-Newton algorithm is invoked again to locate the MLE of the model (F.14). However, in this case, the gradient of the log-likelihood is estimated by the finite difference method, i.e. by numerical approximation.

By defining

(F.16) $$Q_t = \begin{bmatrix} S_t & s_t' \\ s_t & q_t \end{bmatrix},$$

the predicted series based on the model (F.14) fitted by ML can be evaluated by means of the formula

(F.17) $$\hat{y}_t = (b_t + B_t\hat{\gamma}_t)Z_t + C\bar{x}_t$$

where $\hat{\gamma}_t = S_t^{-1}s_t$ and $\bar{x}_t = A_t(-\hat{\gamma}_t\ 1)'$. Graphs of these predicted series for the blood cell count data are shown in figure 1.3. (see P.113)

Discussion

A feature of this 'diffuse' analysis is that the fitted series is extremely similar to that obtained when the ordinary Kalman Filter is used. This implies that, practically speaking, the choice between the ordinary Kalman Filter and the diffuse Kalman Filter is really a matter of preference. However, for the specific data set considered in this thesis, the ordinary Kalman Filter is clearly preferrable since

1. These data correspond to the white blood cell counts of just one patient. It is difficult to justify a diffuse distribution for β in this case.
2. The decomposition of β into the parameters b and B implied by (F.14) reduces the degrees of freedom for fitting the model so much that it tends to result in numerical instability.

On the other hand, if our data correspond to the white blood cell counts of a group of patients, then the diffuse Kalman Filter can be justified because

1. Although different patients may have different values of β, it may be reasonable to assume that there is a 'universal' β for all patients which we may regard as 'diffuse' in order to reflect our ignorance about its value.
2. More patients lead to more data. In this case the loss of degrees of freedom through the inclusion of additional parameters represents a reasonable price to pay for the additional explanatory power provided by these parameters.

1.10 Conclusion

This thesis demonstrates that state space modelling is a convenient way of modelling time series with ignorable missing data. However, the presence of missing observations further increases the non-linearity of the likelihood function. Therefore, as Jones(1986) indicated, it can take a lot of effort to compute the MLE of the model (F.2) or model (F.14). By appropriately reducing the dimensions of the matrices involved, the modified Kalman Filter and the modified Kalman Smoother proved to be a very efficient way to generate the likelihood surface of model (F.2) and model (F.14) respectively. Finally, consideration has been given to estimating the accuracy of the resulting MLE. Although the calculation is not efficient, it represents the first attempt to compute the accuracy of the MLE when missing data are present.

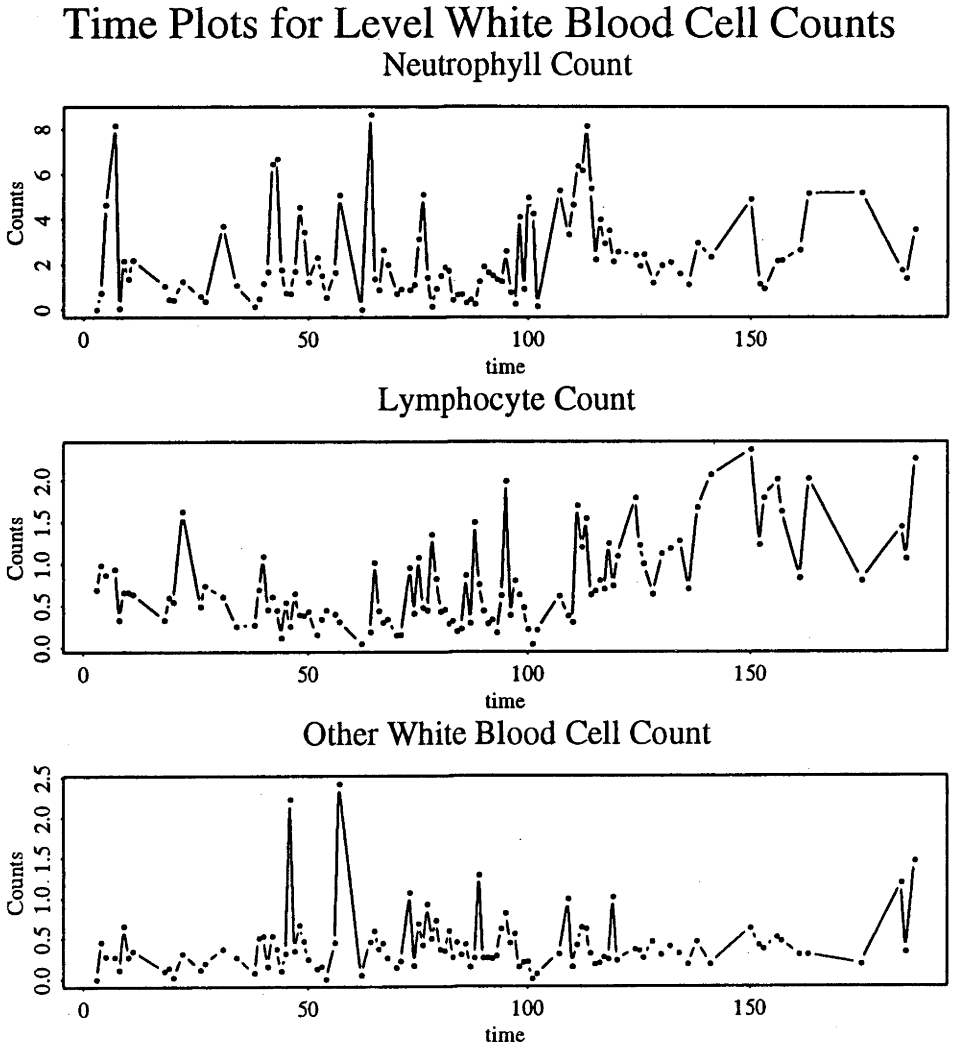
Time Plots for Level White Blood Cell Counts
Neutrophyll Count
Counts
0
2
4
6
8
0
50
100
150
time
Lymphocyte Count
Counts
0.0
0.5
1.0
1.5
2.0
0
50
100
150
time
Other White Blood Cell Count
Counts
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
150
time

Figure 1.1

Predicted Series for Square Root White Blood Cell Counts
generated by Kalman Filtering

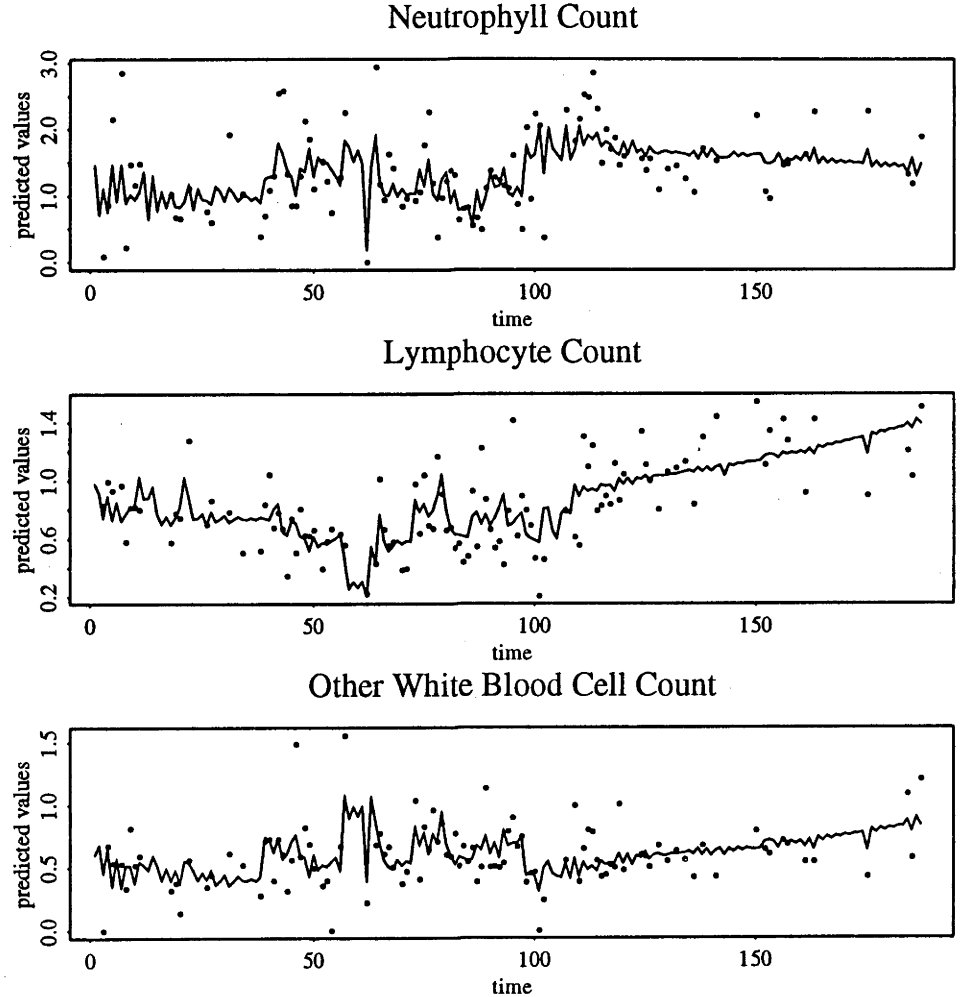

Figure 1.2

Predicted Series for Square Root White Blood Cell Counts
generated by Diffuse Kalman Filtering

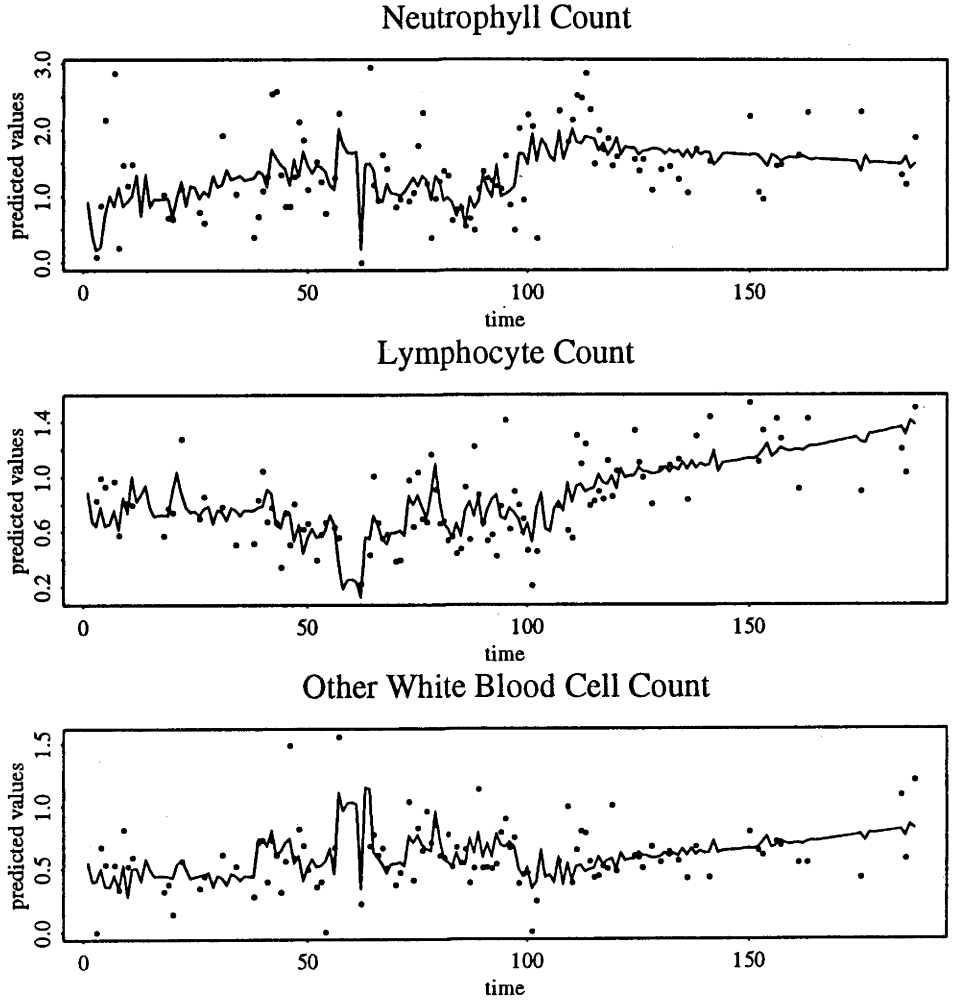

Figure 1.3

2. State Space Modelling

Since Box and Jenkins(1969) advocated the use of ARIMA modelling, we have witnessed an increasing acceptance of another family of models derived from engineering literature called 'State Space Modelling'. State Space Modelling is essentially a Markovian representation of output series given perhaps an input/output map (Kalman(1969)). Kalman(1968) proved the existence of state space realisation for given transfer function or impluse response mapping. This, in effect, justifies the use of state space modelling in modelling most time series data. Moreover, State Space Modelling (SSM) also offers several flexibilities which are not shared by ARIMA modelling. They can be described as follows:

1. SSM is vector in nature as compared with univariate extension of Vector ARMA modelling, see Aoki(1991).
2. SSM allow time varying parameters.
3. SSM allows irregularly observed data (Bucy and Joseph(1968)).

To summarising this interesting and important subject, we begin by describing transfer function model and state space model. The equivalence between the two families of models are shown by a minimal realisation theorem given in Kalman(1968). Then, different forms of state space models are introduced but the most important of all is the prediction error form which will be concentrated in future analysis. Conversion between ARMA models and State space models are discussed. We, then, come to the Kalman Filter and the Kalman Smoother. Finally, maximum likelihood estimation as well as the calculation of the Fisher information matrix are discussed.

2.1 Definitions of Transfer Function Model

Definition 2.1.1

A dynamical input-output system or Transfer function model ($\Sigma I/O$) is defined by the existence of an input-output map F which maps the input into the output.

Remarks:

1. A $\Sigma I/O$ is called linear if F is linear.
2. If the underlying time index is subset of $\mathbb{Z}$ it is called discrete.
3. A $\Sigma I/O$ is called time-invariant if $FS_t=S_tF$ where $S_t u(\tau)=u(t+\tau)$ and u represents input.

It is clear from definition (2.1.1) that the system function F provides the key between inputs and outputs. In general, for a linear, discrete, time-invariant ΣI/O, F can be described in the following as

$$y(t) = \sum_{j=-\infty}^{\infty} L(j)u(t-j) \tag{2.1.2}$$

where $y(t)$ are outputs, $u(t)$ are inputs and $L(j)$ are coefficients of the transfer function $\ell(z)=\sum L(j)z^{-j}$, $z\in\mathbb{C}$. For simplicity, we can write $F=\{L(j)\,|\,j\in\mathbb{Z}\}$ which is also called impulse response function. In addition, a system is called causal if $L(j)=0$ for $j<0$. Those readers who are familiar with time series analysis would understand that all stationary discrete systems admit causal transfer function (see Hannan and Deistler(1988)). We shall be mainly concerned with causal transfer function only.

Definition 2.1.3

A dynamical system with state space is defined by the existence of the state evolution function $\varphi:T_+^2\times\chi\times U\to\chi$, where χ is the set of all possible states, $T_+^2 = \{(t_1,t_0)\in T^2\,|\,t_1>t_0\}$, U is the set of all inputs and T is the time index, which satisfies the following properties:

(i) $\varphi(t_0,t_0,x_0,u)=x_0$, $x_0\in\chi$

(ii) $\varphi(t_2,t_1,\varphi(t_1,t_0,x_0,u),u)=\varphi(t_2,t_0,x_0,u)$

(iii) if for $u_1,u_2\in U$, $u_1(t)=u_2(t)$ for $t_0 <= t < t_1$ then
$\varphi(t_1,t_0,x_0,u_1)=\varphi(t_1,t_0,x_0,u_2)$

(iv) r is the reading function with $r:\chi\times U\times T\to Y$.

In words, $\varphi(t_1,t_0,x_0,u)$ is the state obtained at time t_1 by starting from x_0 and applying input u.

Remark:

1. Σm is called linear if $\varphi(t_1,t_0,\ldots):\chi\times U\to\chi$ is linear and $r(.,.,t):\chi\times U\to Y$ is linear for all t.
2. Σm is called stationary if $\varphi(t_1+t,t_0+t,x_0,u)=\varphi(t_1,t_0,x_0,S_t u)$ and $r(x,u,t)$ is independent of t.

If $T=\{t_k\in\mathbb{Z}\,|\,k=0,1,2,\ldots\}$ we can define the one-step evolution function as $\varphi(t_{k+1},t_k,x(t_k),u(t_k))=f_k(x(t_k),u(t_k))$. Then the following recursions would determine the model completely:

$$x(t_{k+1})=f_k(x(t_k),u(t_k))$$
$$x(t_0)=x_0 \qquad (2.1.4)$$
$$y(t_k)=r(x(t_k),u(t_k),t_k)$$

Moreover, if Σm is linear then f_k and $r(.,.,t)$ can be represented by matrices and we have the following model description:

$$x(t_{k+1})=A(k)x(t_k)+B(k)u(t_k)$$
$$x(t_0)=x_0 \qquad (2.1.5)$$
$$y(t_k)=C(k)x(t_k)+D(k)u(t_k)$$

We call (2.1.5) state space model with time varying parameters. However, as far as our purpose is concerned, we are only interested in constant parameters. i.e. A(k)=A, B(k)=B, C(k)=C and D(k)=D.

2.2 Concepts of Controllability and Observability

Definition 2.2.1

An event (τ,x) is controllable iff there exists a $t\in T$ and an $u\in U$ such that $\varphi(t,\tau,x,u)=0$. In words, an event is controllable iff it can be transferred to 0 in finite time by an appropriate choice of input function u.

Definition 2.2.2

An event (τ,x) is reachable iff there is a $s\in T$ and an $u\in U$ such that $x=\varphi(\tau,s,0,u)$.

Results

The following results assume a real, discrete-time, n-dimensional, linear, constant dynamical system $\Sigma=(A,B,-)$:

(2.2.3) A reachable state is always controllable and the converse is always true whenever $\det(A)\neq 0$.

(2.2.4) A state x of Σ is reachable iff $x \in \text{span}(B,AB,..,A^{n-1}B)$ and hence Σ is completely reachable iff $\text{rank}(B,AB,..,A^{n-1}B)$ is n.

(2.2.5) The set of all reachable states form a vector subspace.

As easily inferred from the above results, reachability is equivalent to controllability in our context although they represent different conceptual meanings. In fact controllability requires that the map from input functions to states be surjective and we can write

$$x(t) = \sum_{j=1}^{t} A^{j-1}Bu(t-j)+A^t x(0). \qquad (2.2.6)$$

Thus, if the inputs u is under "control" then we can move x(t) into any state given any initial state. (see Bucy and Joseph(1968) and Hannan and Deistler(1988)) This also implies that the effect due to uncertainty in the initial state x(0) would be minimal if our system is controllable.

Definition 2.2.7

An event (τ,x) in a real, finite dimensional, linear dynamical system $\Sigma=(A,-,C)$ is unobservable iff $CA^kx=0$ for all $k >= \tau$.

Definition 2.2.8

With respect to the same system, an event (τ,x) is unconstructible iff $CA^kx=0$ for all $k <= \tau$.

Results

The following results are referred to a real, discrete-time, n-dimensional, linear, constant dynamical system $\Sigma=(A,-,C)$:

(2.2.9) The set of all unobservable states forms a vector subspace. Hence, we say that Σ is completely observable iff the subspace of unobservable states contains only the zero element.

(2.2.10) An unobservable state is always unconstructible and vice versa whenever $\det(A)\neq 0$. Thus, Complete observability is always equivalent to complete constructability when $\det(A)\neq 0$.

(2.2.11) The system Σ is completely observable iff $\text{rank}(C',A'C',..,A'^{n-1}C')$ is n.

Observability ensures that the map from present states to present and past output functions be injective. Thus, observability also means that any given pair of input-output sequences (u_k,y_k), k=0,1,... uniquely determine an initial state x(0) by solving the system of equations $CA^tx(0)=y(t)$, t=0,1,2,..,n-1. Together observability and controllability makes sure that our modelling/filtering problem is well posed.

Theorem (2.2.12) (Canonical Decomposition)(Kalman(1968))

Every real (continuous or discrete time), finite dimensional, constant, linear dynamical system may be canonically decomposed into four parts, of which only one part, that which is completely controllable and completely observable, is involved in the input/output behaviour of the system.

Hence, as far as what is observed is concerned, it is no

restriction to consider only the part which corresponds to completely observable and completely controllable. Thus, we can assume complete observability and complete controllability when performing our data analysis. This is important as they jointly guarantee not only minimum dimensionality but also the existence of state space realisation.

2.3 Realisation Theory

Statement of the problem:

Given the impluse response matrix or transfer function of ΣI/O, can we find a state space model with the same behaviour as the given system?

This problem was solved by Kalman(1968) via the following minimum realisation theorem for real, continuous-time, finite dimensional, linear dynamical system:

Theorem 2.3.1(Kalman(1968))

Given the impulse response matrix W of a real, continuous-time, finite dimensional, linear dynamical system, there exists a real, continuous time, finite dimensional, linear dynamical system Σ_W which

(a) realises W: i.e. the impluse response matrix of Σ_W is equal to W;

(b) has minimal dimension in the class of linear systems satisfying (a);

(c) is completely controllable and completely observable;

Corollary 2.3.2

If W comes from a constant system, there is a constant Σ_W which satisfies (a), (b) and (c).

Corollary 2.3.3

All claims of Corollary 2.3.2 continues to hold if W is replaced by transfer function matrix of a constant, finite dimensional system.

Although there exists state space realisation of given transfer function, it is by no means unique. Consider the following state space system (2.3.4):

$$x(t+1)=Ax(t)+Bu(t)$$
$$y(t)=Cx(t)+u(t)$$

Its transfer function can be shown to be $z^{-1}C(I-Az^{-1})^{-1}B+I$ where z^{-1} is the lag operator. For any non-singular matrix T with compatible dimension, the system (TAT^{-1},TB,CT^{-1}) can be seen to have the same transfer function. Therefore, the existence of state space realisation is not unique but it is unique if we restrict to equivalence classes defined by the following equivalence relation: $(A_1,B_1,C_1) \sim (A_2,B_2,C_2)$ iff they have the same transfer function.

We consider a causal transfer function representation of input-output model:

$$y(t) = \sum_{i=0}^{\infty} L(i)u(t-i) \tag{2.3.5}$$

where $y(t)\in\mathbb{R}^p$ is the output vector and $u(t)\in\mathbb{R}^q$ is input vector. Then we immediately have the following relation:

$$\begin{pmatrix} y(k+1) \\ y(k+2) \\ \vdots \end{pmatrix} = \begin{pmatrix} G_1 & G_2 & G_3 & \cdots \\ G_2 & G_3 & \cdots & \\ G_3 & \cdots & & \\ \vdots & & & \end{pmatrix} \begin{pmatrix} u(k) \\ u(k-1) \\ \vdots \end{pmatrix} + \begin{pmatrix} G_0 & 0 & 0 & \cdots \\ G_1 & G_0 & 0 & \cdots \\ G_2 & G_1 & G_0 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} u(k+1) \\ u(k+2) \\ \vdots \end{pmatrix}$$

or $Y^+ = \mathcal{H}u^- + \mathcal{T}u^+$ where $\mathcal{H}$ and $\mathcal{T}$ are called Hankel matrix and Toeplitz matrix respectively. It is clear from the above relation that $\mathcal{H}$ represents the relationship between future outputs and past inputs. If $\text{rank}(\mathcal{H})=n < \infty$ this relationship can be described by an n-dimensional memory vector called state vector.

Result (2.3.6) (Hannan and Deistler(1988))

1. The rank of $\mathcal{H}$ is finite iff the transfer function is rational.
2. If the transfer function is rational and can be written as $\tilde{a}^{-1}\tilde{b}$ where $\tilde{a}$ and $\tilde{b}$ are polynomial matrices then the degree n of the polynomial $\det(\tilde{a})$ is equal to the number of linearly independent rows of $\mathcal{H}$. If $\mathcal{H}$ has rank n, then $\mathcal{H}_n^n$ also has rank n where

$$\mathcal{H}_n^n = \begin{pmatrix} G_1 & G_2 & G_3 & \cdots & G_n \\ G_2 & G_3 & \cdots & & G_{n+1} \\ \cdot & \cdot & \cdot & & \cdot \\ G_n & G_{n+1} & \cdots & & G_{2n-1} \end{pmatrix}$$

Moreover, we have the following theorem given in Otter(1985):

Theorem (2.3.7)

Given the impulse response matrix or transfer function $\{G_k\}$ of a real, discrete-time, finite dimensional, constant, linear dynamical system, we have

(1) there is (A,B,C) of system (2.3.4) such that $\dim(x)<\infty$ and the transfer function of (2.3.4) is equal to $\{G_k\}$ iff $\text{rank}(\mathcal{H})$ is finite.

(2) there is a realisation with $\dim(x)=\text{rank}(\mathcal{H})$.

(3) Suppose $\text{rank}(\mathcal{H})$ is finite, $\dim(x)=\text{rank}(\mathcal{H})$ iff the realisation is observable and controllable iff the realisation is minimal.

An immediate consequence of theorem (2.3.7) and result (2.3.6) is that any causal rational transfer function can be expressed as $z^{-1}C(I-Az^{-1})B+I$ where (A,B,C) is observable and controllable. This result is important because there is always a minimal state space realisation for given causal rational function which approximates the transfer function of our data series. It also justifies the use of state space modelling in approximating the dynamics of our data series. Moreover, the Hankel matrix $\mathcal{H}$ is extremely important in not only formalising the relationship between the future outputs and past inputs but also determining the dimension of state space realisation.

2.4 Stability

We shall not discuss stability in great depth such as different concepts of stability except to recognise that stability is a pre-requisite condition to have a convergent causal rational transfer function. This can easily be seen from expression of the transfer function of system (2.3.4) defined as $z^{-1}C(I-Az^{-1})^{-1}B+I$ and $(I-Az^{-1})^{-1}=\sum_{j=0}^{\infty}A^jz^{-j}$. The inverse exists iff $\det(I-Az^{-1})\neq 0$ for $|z|\geq 1$ which is clearly equivalent to $\det(A-\lambda I)\neq 0$ for $|\lambda|\geq 1$ and this is equivalent to $\max_i(|\lambda_i|)\leq 1$ where λ_i is the ith eigenvalue of A. Therefore we have the following result:

Result 2.4.1

The time-invariant linear discrete system (2.3.4) is stable in the sense that it has a convergent transfer function outside the unit disk iff $|\lambda_i| \leq 1$ for all i.

Remark: for continuous system the condition would be $\text{Re}(\lambda_i) \leq 0$ for all i instead of $|\lambda_i| \leq 1$ for all i.

By expanding the transfer function defined above, we have the steady state solution of system (2.3.4) written as follows:

$$x(t) = \sum_{j=1}^{\infty} A^{j-1}B\varepsilon(t-j) \tag{2.4.2}$$

$$y(t) = \sum_{j=1}^{\infty} CA^{j-1}B\varepsilon(t-j) + \varepsilon(t) \tag{2.4.3}$$

Moreover, stability is crucial in proving the existence of solution of the Liapunov equation $\mathbf{P-FPF'=Q}$ which provides the starting condition for our filtering algorithm developed by Kalman and Bucy(1961). By assuming $\mathbf{P}$ is stable, the solution of Liapunov equation can be written as follows (2.4.4):

$$\text{vec}(\mathbf{P}) = \{\mathbf{I} - (\mathbf{F}\otimes\mathbf{F})\}^{-1}\text{vec}(\mathbf{Q}), \text{ or}$$

$$\mathbf{P} = \sum_{j=0}^{\infty} \mathbf{F}^j\mathbf{Q}\mathbf{F}'^j.$$

2.5 Prediction Error Form and State Space Models

In the following we shall consider input series as unobserved error and regard them as white noise. We refer that $\varepsilon(t)$ is white noise iff $E(\varepsilon(t))=0$ and $E(\varepsilon(t)\varepsilon(s)')=\delta_{ts}\Omega$. Given the binary nature of our observed input series, we will consider including them into our state space model in later section and would like to ignore them at the moment. Throughout the literature, state space representation can come in different forms. For example Shumway and Stoffer(1982) considered application of EM algorithm to identify the following state space model (2.5.1):

$$x(k+1) = Ax(k) + \upsilon(k)$$
$$y(k) = Cx(k) + \omega(k)$$

where
1. $y(k)\in\mathbb{R}^p$ is the observed output series
2. $x(k)\in\mathbb{R}^q$ is the unobserved state vector with $x(0)$ is a $N(\mu,\Sigma)$ r.v.
3. $\upsilon(k),\omega(s)$ are independent white noise
4. $x(0)$ and $\upsilon(k)$ or $\omega(s)$ are independent.

However, Aoki(1991) often referred to the following state space model (2.5.2):

$$x(k+1) = Ax(k) + B\varepsilon(k)$$
$$y(k) = Cx(k) + \varepsilon(k)$$

where 1. $y(k)$ and $x(k)$ are similarly defined

2. $\varepsilon(k)$ is independent white noise with $E(\varepsilon(t)\varepsilon(s))=\delta_{ts}\Omega$

3. $x(0)$ and $\varepsilon(k)$ are uncorrelated.

For more variety of state space representation, we refer to Otter(1985). In fact, state space model can, in general, be written in the following form (2.5.3):

$$x(t+1) = Ax(t) + \xi(t)$$
$$y(t) = Cx(t) + \eta(t)$$

where 1. $y(t)$ and $x(t)$ are similarly defined

2. $\xi(t)$ and $\eta(t)$ are errors satisfying

$$E(\xi(t)',\eta(t)')=0$$
$$E\begin{pmatrix}\xi(s)\\ \eta(s)\end{pmatrix}(\xi(t)',\eta(t)')=\delta_{st}\begin{pmatrix}Q & S\\ S' & R\end{pmatrix}.$$

We immediately see that system (2.5.3) is more general than system (2.5.1) considered in Shumway and Stoffer(1982). On the other hand, system (2.5.2) is called Prediction Error Form and can be shown that every state space system (2.5.3) can be transformed into this form. The following details are extraction from Hannan and Deistler(1988):

Suppose the stability condition of (2.4.1) is satisfied and we have the following steady state solution analogous to (2.4.2):

$$x(t) = \sum_{j=1}^{\infty} A^{j-1}\xi(t-j) \tag{2.5.4}$$

and hence $y(t) = \sum_{j=1}^{\infty} CA^{j-1}\xi(t-j) + \eta(t)$

Define $H_y(t) = \text{span}\{y(j)\,|\,j\le t\}$ and let $x(t|s)$ be the orthogonal projection of $x(t)$ on $H_y(s)$. Then from the specification of system (2.5.3), we have $x(t+1|t)=Ax(t|t)=Ax(t|t-1)+A\{x(t|t)-x(t|t-1)\}$. Moreover, we define $\varepsilon(t)$ by $y(t)=Cx(t|t-1)+\varepsilon(t)$. Since $\varepsilon(t)$ is orthogonal to $H_y(t-1)$ and $H_y(t)$ is also spanned by $\varepsilon(t)$ and $y(j)$, $j<t$, we see that $x(t|t)-x(t|t-1) \in \text{span}(\varepsilon(t))$. Thus there exists B such that system (2.5.3) is transformed into

$$x(t+1|t) = Ax(t|t-1) + B\varepsilon(t)$$
$$y(t) = Cx(t|t-1) + \varepsilon(t)$$

a form of system (2.5.2).

Furthermore, from the setting of the prediction error form, we understand that $\varepsilon(t)$ is the innovation of the vector $y(t)$ and also

appears in the transfer function representation (2.3.5) of $y(t)$. Because of all of these advantages we are going to concentrate mainly on Prediction Error Form.

2.6 State Space Model and Causal ARMA model

In this section, we shall discuss the relationship or connection between State Space Models and more traditional (to Statisticians) causal ARMA models. Consider the following Causal ARMA(p.q) model ($p \geq q$) (2.6.1):

$$y(t) = \sum_{j=1}^{p} A_j y(t-j) + \varepsilon(t) + \sum_{j=1}^{q} B_j \varepsilon(t-j)$$

where $y(t)$ is the output process and $\varepsilon(t)$ is standard white noise. We define

$$\begin{aligned}
x_p(t) &= A_p x_1(t-1) + K_p \varepsilon(t-1) \\
x_{p-1}(t) &= A_{p-1} x_1(t-1) + x_p(t-1) + K_{p-1} \varepsilon(t-1) \\
&\cdots\cdots\cdots \\
x_2(t) &= A_2 x_1(t-1) + x_3(t-1) + K_2 \varepsilon(t-1) \\
x_1(t) &= A_1 x_1(t-1) + x_2(t-1) + K_1 \varepsilon(t-1) \\
x_1(t) &= y(t) - \varepsilon(t)
\end{aligned}$$

where

$$K_i = \begin{cases} B_i + A_i & i=1,2,\ldots,q \\ A_i & i=q+1,\ldots,p \end{cases}$$

Then the ARMA model (2.6.1) can be represented by the following state space model (2.6.2) in prediction error form:

$$\begin{aligned}
x(t+1) &= Ax(t) + K\varepsilon(t) \\
x(0) &= 0 \\
y(t) &= Cx(t) + \varepsilon(t)
\end{aligned}$$

where $x(t)=(x_1(t)',\ldots,x_p(t)')'$, $C=(I,0,\ldots,0)$, $K=(K_1',\ldots,K_p')'$,

$$A = \begin{pmatrix} A_1 & I & \cdots & 0 \\ \vdots & & & \\ & & & I \\ A_p & 0 & \cdots & 0 \end{pmatrix}$$

Conversely, if given a controllable and observable, i.e. minimal, state space model with $\dim(x)=n$:

$$\begin{aligned}
x(t+1) &= Ax(t) + B\varepsilon(t) \\
y(t) &= Cx(t) + \varepsilon(t)
\end{aligned}$$

we have the following relationship:

$$\begin{pmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+n} \end{pmatrix} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^n \end{pmatrix} x(t) + \begin{pmatrix} I & & & \\ CB & I & & \\ \vdots & \vdots & \ddots & \\ CA^{n-1}B & \cdots & CB & I \end{pmatrix} \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t+1} \\ \vdots \\ \varepsilon_{t+n} \end{pmatrix}$$

Since, by observability, $(C',(CA)',\ldots,(CA^{n-1})')'$ is already of full rank n, there exists $a_1,\ldots,a_n$ such that

$$CA^n = a_n C + a_{n-1} CA + \cdots + a_1 CA^{n-1}$$

Then

$$y(t+n)-a_1 y(t+n-1)- \cdots -a_n y_t = \varepsilon_{t+n}+Q_1\varepsilon_{t+n-1}+ \cdots +Q_n\varepsilon_t$$

for some $Q_1,Q_2,\ldots,Q_n$.

Thus, we see that discrete-time state space models is in fact equivalent to Causal ARMA models. However, with regard to allowing irregularly observed data, state space models offer a very efficient representation in terms of both evaluating the likelihood function and of acting as on-line procedure as data comes in.

2.7 Kalman Filter and Kalman Smoother

This section is concerned with understanding and derivation of the celebrated Kalman Filter and its Smoother developed by Kalman and Bucy(1961). As Hannan and Deistler(1988) pointed out, Kalman Filtering introduced a new era and found a lot of real time applications especially when the filter is not restricted to stationary series. Moreover, Kalman(1968) described that Wiener-Kolmogorov Filter + theory of finite-dimensional linear dynamical systems = Kalman Filter. Thus, we see that Kalman Filter is a very interesting and powerful tools in resolving problems of statistical prediction and filtering.

Kalman Filtering, in one view, is an algorithm to produce (asymptotically) optimal state estimates and, in another view, can be regarded as a procedure to effect a Gram-Schmidt orthogonalisation of our data series. Based on the latter view, Kalman Filter can also be used to generate the likelihood surface given the knowledge of parameters of our model. This is important because a descent optimisation algorithm can be invoked to find the (locally) maximum likelihood estimates of our parameters.

However, there are few authors who actually addressed the problem of evaluating the asymptotic variance of the parameters estimate. One of them is Mehra(1976) who considered expanding the dimension of state vector to evaluate the Fisher information matrix the inverse of which determines the asymptotic variance of MLE. With regard to variation for missing observations or evaluation of Fisher information matrix with missing observations, we shall discuss them in later chapter.

As sufficed to our purpose, we shall concentrate on studying a stable and stationary state space system (2.7.1):

$$x(k+1) = Ax(k) + B\varepsilon(k)$$
$$y(k) = Cx(k) + \varepsilon(k)$$

where

1. $y(k)\in\mathbb{R}^n$ is the output vector
2. $x(k)\in\mathbb{R}^m$ is the unobserved state vector
3. $x(0)$ is distributed as $N(\mu,\Sigma)$
4. $\varepsilon(k)$ is independent white noise normally distributed with mean 0 and variance Ω.

Statement of Problem:

Given state space system (2.7.1) and for $t \le t_1$ find an optimal estimator $\hat{x}(t_1)$ of the unobserved state vector $x(t_1)$ in the sense that its mean square error $E\{(x(t_1)-\hat{x}(t_1))(x(t_1)-\hat{x}(t_1))'\}$ is minimum.

The above statement is also called a filtering problem. The answer is very simple: the optimum estimator is simply given by $E\{x(t_1)|\mathcal{F}(t_1)\}$ where $\mathcal{F}(t_1)$ is the σ-algebra generated by $y(t)$, $t \le t_1$ (see Bucy and Joseph(1968)). If we are only interested in theoretically solving the filtering problem, we are finished. But we also would like to have a computational possible algorithm to evaluate the expectations. This alogrithm is provided by Kalman Filter. Although the derivation of Kalman Filter can be found in lot of text, we would like to use (1) the same logic of deriving the Kalman Filter to derive the Kalman Smoother and (2) the derivation to explain the adjustment idea to allow missing observations in next chapter. Therefore, we shall discuss how Kalman Filter was derived and how it can effect a Gram-Schmidt orthogonalisation of our data series .

Define

$$\hat{x}(n) = E(x(n)|y(1),\ldots,y(n))$$

$$\hat{x}(0) = \mu$$

1.The Gram-Schmidt orthogonalisation of $\{y(t)\}$ is given as

$$\omega(n) = y(n) - E(y(n)|y(1),\ldots,y(n-1))$$

$$\omega(1) = y(1) - E(y(1))$$

If we define $\bar{x}(n) = E(x(n)|y(1),\ldots,y(n-1))$, then

$$\omega(n) = y(n) - E(Cx(n)+\varepsilon(n)|y(1),\ldots,y(n-1))$$

$$= y(n) - C\bar{x}(n)$$

2. The Gram-Schmidt orthogonalisation of $\{\hat{x}(n)\}$ is given as

$$\omega^s(n) = \hat{x}(n) - E(\hat{x}(n)|\hat{x}(1),\ldots,\hat{x}(n-1))$$

$$\omega^s(1) = \hat{x}(1) - A\mu$$

Step 1

Show that $E(\hat{x}(n)|\hat{x}(1),\ldots,\hat{x}(n-1)) = E(\hat{x}(n)|y(1),\ldots,y(n-1))$

We observe that $\hat{x}(n)-E(\hat{x}(n)|y(1),\ldots,y(n-1))$ is uncorrelated with $y(1),\ldots,y(n-1)$. Since $\hat{x}(1),\ldots,\hat{x}(n-1)$ are functions of $y(1),y(2),\ldots,y(n-1)$, $\hat{x}(n)-E(\hat{x}(n)|y(1),\ldots,y(n-1))$ is uncorrelated with $\hat{x}(1),\ldots,\hat{x}(n-1)$. Hence, we have

$$E\left\{ \hat{x}(n)-E(\hat{x}(n)|y(1),\ldots,y(n-1))|\hat{x}(1),\ldots,\hat{x}(n-1)\right\} = 0$$

and the result follows.

Step 2

Show that $E(\varepsilon(t)|y(1),\ldots,y(t)) = \Omega Q_t^{-1}\omega(t)$ where $Q_t = E(\omega(t)\omega(t)')$.

Since $\omega(1),\ldots,\omega(t)$ are Gram-Schmidt orthogonalisation of $y(1),\ldots,y(t)$, $E(\varepsilon(t)|y(1),\ldots,y(t))=E(\varepsilon(t)|\omega(1),\ldots,\omega(t))$. Since $\varepsilon(t)$ is independent of $y(1),\ldots,y(t-1)$ and $\omega(t)$ represents the innovation on $y(t)$, $E(\varepsilon(t)|\omega(1),\ldots,\omega(t))=E(\varepsilon(t)|\omega(t))$. From standard conditional expectation result,

$$E(\varepsilon(t)|\omega(t)) = \text{cov}(\varepsilon(t),\omega(t))E(\omega(t)\omega(t)')^{-1}\omega(t)$$

$$= \text{cov}(\varepsilon(t),\varepsilon(t))Q_t^{-1}\omega(t)$$

$$= \Omega Q_t^{-1}\omega(t)$$

and the result follows.

Step 3

Show that $E(\hat{x}(n)|y(1),\ldots,y(n-1)) = \bar{x}(n) = A\hat{x}(n-1)+B\Omega Q_{n-1}^{-1}\omega(n-1)$.

$$E(\hat{x}(n)|y(1),\ldots,y(n-1)) = E(E(x(n)|y(1),..,y(n))|y(1),..,y(n-1))$$

$$= E(x(n)|y(1),\ldots,y(n-1))$$

$$= \bar{x}(n)$$

$\bar{x}(n) = E(x(n)\,|\,y(1),\ldots,y(n-1))$

$= E(Ax(n-1)+B\varepsilon(n-1)\,|\,y(1),\ldots,y(n-1))$

$= A\hat{x}(n-1) + BE(\varepsilon(n-1)\,|\,y(1),\ldots,y(n-1))$

$= A\hat{x}(n-1) + B\Omega Q_{n-1}^{-1}\nu(n-1)$ from Step 2.

From results in step 1 and step 3, $\nu^s(n) = \hat{x}(n) - \bar{x}(n)$.

Step 4

Show that there is K_n such that $\nu^s(n)=K_n\nu(n)$.

Note that

$E(\nu^s(n)\nu(n-k)') = E((\hat{x}(n)-x(n)+x(n)-\bar{x}(n))\nu(n-k)')$

$= 0$ for $k \geq 1$ by optimality of $\hat{x}(n)$ and $\bar{x}(n)$.

i.e. $\nu^s(n)$ is correlated only with $\nu(n)$. However, since $\nu^s(n)$ can be expressed linearly in terms of $\nu(1),\ldots,\nu(n)$, there exists K_n such that $\nu^s(n) = K_n\nu(n)$. Moreover, K_n can be defined as solution of the Wiener-Hopf equation: $E(\nu^s(n)\nu(n)')=K_nE(\nu(n)\nu(n)')$ as $E(\nu^s(n)\,|\,\nu(n))=K_n\nu(n)$.

It follows that $\hat{x}(n) = \bar{x}(n) + K_n\nu(n)$.

Define $H_{n-1} = E\{(x(n)-\bar{x}(n))(x(n)-\bar{x}(n))'\}$

Step 5

Show that $Q_n = \Omega + CH_{n-1}C'$ and $K_n = H_{n-1}C'Q_n^{-1}$.

$Q_n = E(\nu(n)\nu(n)')$

$= E\{(C(x(n)-\bar{x}(n))+\varepsilon(n))(C(x(n)-\bar{x}(n))+\varepsilon(n))'\}$

$= \Omega + CH_{n-1}C'$

It suffices to show that $E(\nu^s(n)\nu(n)')=H_{n-1}C'$. This is because

$E(\nu^s(n)\nu(n)') = E\{(\hat{x}(n)-x(n)+x(n)-\bar{x}(n))\nu(n)'\}$

$= E\{(x(n)-\bar{x}(n))\nu(n)'\}$

$= E\{(x(n)-\bar{x}(n))(C(x(n)-\bar{x}(n))+\varepsilon(n))'\}$

$= H_{n-1}C'$

From $K_n = E(\nu^s(n)\nu(n)')Q_n^{-1}$ the result follows.

Define $e(n)=x(n)-\hat{x}(n)$ and $P_n=E(e(n)e(n)')$

Step 6

Show that $P_n=(I-K_nC)H_{n-1}$.

Since $e(n)=x(n)-\hat{x}(n)=(I-K_nC)(x(n)-\bar{x}(n))-K_n\varepsilon(n)$,

$P_n = (I-K_nC)H_{n-1}(I-K_nC)'+K_n\Omega K_n'$

$= (I-K_nC)H_{n-1}-\{H_{n-1}C'-K_n(CH_{n-1}C'+\Omega)\}K_n'$

$= (I-K_nC)H_{n-1}$ by results in step 5.

Step 7

Show that $H_{n-1} = AP_{n-1}A'+B\Omega B'-AK_{n-1}\Omega B'-B\Omega K'_{n-1}A'-B\Omega Q_{n-1}^{-1}\Omega B'$

$$
\begin{aligned}
H_{n-1} &= E\{(x(n)-\bar{x}(n))(x(n)-\bar{x}(n))'\} \\
&= E\{(Ae(n-1)+B\varepsilon(n-1)-B\Omega Q_{n-1}^{-1}\omega(n-1))(\cdot)'\} \\
&= AP_{n-1}A'+B\Omega B'-AK_{n-1}\Omega B'-B\Omega K'_{n-1}A'-B\Omega Q_{n-1}^{-1}\Omega B'.
\end{aligned}
$$

Assuming stationary, Σ satisfies $\Sigma = A\Sigma A'+B\Omega B'$. From (2.4.4), we see that $vec(\Sigma) = \{I-(A\otimes A)\}^{-1}vec(B\Omega B')$ where $vec(\cdot)$ is defined as the stack of column vectors of $\cdot$. Hence we have the following (forward) Kalman Filter:

Initial conditions:

$\hat{x}(0) = \mu$, $H_0 = B\Omega B'+A\Sigma A'$, Σ is given by (2.4.4)

for given data series $\{y(1),\ldots,y(T)\}$ and for n from 1 to T, the following equations are calculated:

$$Q_n = \Omega + CH_{n-1}C'$$

$$K_n = H_{n-1}C'Q_n^{-1}$$

$$P_n = (I-K_nC)H_{n-1}$$

$$\bar{x}(n) = A\hat{x}(n-1)+B\Omega Q_{n-1}^{-1}(y(n-1)-C\bar{x}(n-1)),\ \bar{x}(1)=A\hat{x}(0)$$

$$\hat{x}(n) = \bar{x}(n) + K_n(y(n)-C\bar{x}(n))$$

$$H_n = AP_nA'+B\Omega B'-AK_n\Omega B'-B\Omega K'_nA'-B\Omega Q_n^{-1}\Omega B'$$

Remark:

Since $\omega(n)=y(n)-C\bar{x}(n)$'s are orthogonalisation of our data series and Q_n is defined as $E(\omega(n)\omega(n)')$, we see that:

1. As a by-product, Kalman Filter produces the orthogonalised sequence $\omega(n)$,n=1,..,T as well.
2. Define $v = (\omega(1)',\ldots,\omega(T)')'$ a column vector,

   $$E(v)=0 \text{ and } Cov(v)=diag(Q_1,\ldots,Q_T)$$

   since $E(\omega(t)\omega(s)')=0$ if $t\neq s$.
3. By assuming Guassian noises, the likelihood function apart from a constant is defined as:

   $$-2\log L = \sum_{t=1}^{T} \log\det(Q_t) + \sum_{t=1}^{T} \omega(t)'Q_t^{-1}\omega(t).$$

4. It is importance to understand that we are never able to construct optimal Kalman Filter because of the lack of knowledge of μ and P_n. However, due to the assumption of

controllability, the initial data effect is forgetable.

Therefore, the Kalman Filter stated above would be asymptotically optimal.

In the following we are going to discuss the derivation of Kalman Smoother of system (2.7.1):

Define $\hat{x}^T(n)=E\{x(n)\,|\,y(1),\ldots,y(T)\}$. From the orthogonal nature of $\{\nu(1),\ldots,\nu(T)\}$, we can write

$$\hat{x}^T(n) = \sum_{m=1}^{T} A_{n,m}\nu(m) + E(x(n)) \quad \text{and}$$

$$\hat{x}(n) = \sum_{m=1}^{n} A_{n,m}\nu(m) + E(x(n))$$

By comparing coefficients on the equation:

$$\hat{x}(n) = A\hat{x}(n-1)+B\Omega Q_{n-1}^{-1}\nu(n-1)+K_n\nu(n),$$

we obtain the following relations:

$$A_{n,n} = K_n$$
$$A_{n,n-1} = AA_{n-1,n-1}+B\Omega Q_{n-1}^{-1}$$
$$A_{n,m} = AA_{n-1,m},\ m=1,\ldots,n-2$$

Step a

Show that $A_{n,n+1} = P_nA'C'Q_{n+1}^{-1}$

Since $A_{n,n+1}E(\nu(n+1)\nu(n+1)')=E((x(n)-\hat{x}(n))\nu(n+1)')$ (by optimality of $\hat{x}(n)$ and orthogonal structure of $\nu(s),s=1,\ldots,T$) and $\nu(n+1)=\varepsilon(n+1)+C(Ae(n)+B\varepsilon(n)-B\Omega Q_n^{-1}\nu(n))$, $A_{n,n+1}Q_{n+1}=P_nA'C'$. Hence the result follows.

Since H_n is defined as covariance matrix of $x(n-1)-\bar{x}(n-1)$, we can assume H_n is non-singular. From the result $H_nC'=K_{n+1}Q_{n+1}$ of step 5, we can write (2.7.2) $C'=H_n^{-1}K_{n+1}Q_{n+1}$.

Step b

Show that $\hat{x}^T(n)-\hat{x}(n)=S_n(\hat{x}^T(n+1)-A\hat{x}(n)-B\Omega Q_n^{-1}\nu(n))$ where $S_n=P_nA'H_n^{-1}$.

The result will be shown by proving that:

(2.7.3) $E\{(\hat{x}^T(n)-\hat{x}(n)-S_n(\hat{x}^T(n+1)-A\hat{x}(n)-B\Omega Q_n^{-1}\nu(n)))\nu(m)'\}=0\ \forall m\geq 1$

Case 1: $m \leq n$

We can show that

$$\hat{x}^T(n)-\hat{x}(n) = \sum_{j=n+1}^{T} A_{n,j}\nu(j)$$

$$\hat{x}^T(n+1)-A\hat{x}(n) = \sum_{j=n+1}^{T} A_{n+1,j}\nu(j) + B\Omega Q_n^{-1}\nu(n)$$

Therefore (2.7.3) is OK.

Case 2: m > n

$$E\{(\hat{x}^T(n)-\hat{x}(n)-S_n(\hat{x}^T(n+1)-A\hat{x}(n)-B\Omega Q_n^{-1}\nu(n)))\nu(m)'\}$$

$$= E\{(\hat{x}^T(n)-S_n\hat{x}^T(n+1))\nu(m)'\}$$

$$= E\{(x(n)-S_n x(n+1))\nu(m)'\} \text{ (by optimality of } \hat{x}^T(n))$$

Therefore it suffices to show that $E\{(x(n)-S_n x(n+1))\nu(n+p)'\}=0$ for all $p \geq 1$. We shall show this by induction.

For <u>p=1</u>

Using (2.7.2) and Step a, $A_{n,n+1}=P_n A'H_n^{-1}A_{n+1,n+1}$ since $A_{n,n}=K_n$.

Thus, $A_{n,n+1} = S_n A_{n+1,n+1}$ and hence p=1 is true.

Suppose $E\{(x(n)-S_n x(n+1))\nu(n+p)'\} = 0$ for $p \geq 1$.

Since $\nu(n)=C[Ae(n-1)+B\varepsilon(n-1)-B\Omega Q_{n-1}^{-1}\nu(n-1)]+\varepsilon(n)$, we see that

$$E\{(x(n)-S_n x(n+1))\nu(m)'\}$$

$$=E\{(x(n)-S_n x(n+1))(e(m-1)'A'C'-\nu(m-1)'Q_{m-1}^{-1}\Omega B'C')\} \text{ for } m\geq n+2.$$

Therefore, induction hypothesis also implies that

for $p \geq 2$,

(2.7.4) $E\{(x(n)-S_n x(n+1))(e(n+p-1)'A'-\nu(n+p-1)'Q_{n+p-1}^{-1}\Omega B')C'\}=0$

for p = 1

(2.7.5) $E\{(x(n)-S_n x(n+1))(e(n)'A'+\varepsilon(n)'B'-\nu(n)'Q_n^{-1}\Omega B')C'\}=0.$

Consider

1) for $p \geq 2$,

$E\{(x(n)-S_n x(n+1))\nu(n+p+1)'\}$

$= E\{(x(n)-S_n x(n+1))((e(n+p-1)'A'+\varepsilon(n+p-1)'B'-\nu(n+p-1)'Q_{n+p-1}^{-1}\Omega B'-$
$\nu(n+p)'K_n')A'-(\varepsilon(n+p)'+(e(n+p-1)'A'-\nu(n+p-1)'Q_{n+p-1}^{-1}\Omega B'+$
$\varepsilon(n+p-1)'B')C')Q_{n+p}^{-1}\Omega B')C'\}$

= 0 by (2.7.4)

2) for p = 1,

$E\{(x(n)-S_n x(n+1))\nu(n+2)'\}$

$= E\{(x(n)-S_n x(n+1))((e(n)'A'+\varepsilon(n)'B'-\nu(n)'Q_n^{-1}\Omega B'-\nu(n+1)'K_n')A'-$
$(\varepsilon(n+1)'+(e(n)'A'+\varepsilon(n)'B'-\nu(n)'Q_n^{-1}\Omega B')C')Q_{n+1}^{-1}\Omega B')C'\}$

= 0 by (2.7.5).

Hence, by induction, (2.7.3) is true for case 2.

This completes the proof of Step b. In particular, $A_{n,m}=S_n A_{n+1,m}$ for $m \geq n+1$.

Define $e^T(n)=x(n)-\hat{x}^T(n)$ and $P_n^T=E\{e^T(n)e^T(n)'\}$. Then from result in step b, we have the following relation:

$$S_n\hat{x}^T(n+1)+e^T(n) = e(n)+S_nA\hat{x}(n)+S_nB\Omega Q_n^{-1}\psi(n)$$

Hence,

(2.7.6)

$$S_nE\{\hat{x}^T(n+1)\hat{x}^T(n+1)'\}S_n+P_n^T$$

$$= P_n+S_nAE\{\hat{x}(n)\hat{x}(n)'\}A'S_n'+S_nB\Omega Q_n^{-1}\Omega B'S_n'+S_nB\Omega K_n'A'S_n'+S_nAK_n\Omega B'S_n'.$$

Step c

Show that $P_n^T=P_n+S_n(P_{n+1}^T-H_n)S_n'$

Define $R_n=E\{x(n)x(n)'\}=AR_{n-1}A'+B\Omega B'$ and $R_0=\Sigma+\mu\mu'$. We can show that $E(\hat{x}^T(n)\hat{x}^T(n))=R_n-P_n^T$ and $E(\hat{x}(n)\hat{x}(n)')=R_n-P_n$. Therefore, from (2.7.6) and from result in step 7, we have our result.

Given the results in step a, b and c, we can define the following recursions to find $\hat{x}^T(n)$ and P_n^T and they are called Kalman Smoother.

Initial conditions: $\hat{x}^T(T)=\hat{x}(T)$ and $P_T^T=P_T$

for n from T-1 to 0, we do the following recursions:

$$S_n = P_nA'H_n^{-1}$$

$$\hat{x}^T(n) = \hat{x}(n) + S_n(\hat{x}^T(n+1)-\bar{x}(n+1))$$

$$P_n^T = P_n + S_n(P_{n+1}^T-H_n)S_n'$$

The importance of Kalman Smoother is its application to missing data. The missing data score function or information matrix requires evaluation of conditional expectation of x(n) given the entire data vector. We shall discuss missing data score function and information matrix in later chapter.

2.8 Maximum Likelihood Estimation

As we have suggested in last section, Kalman Filtering is an effective means to evaluate the likelihood function given the model parameters. In this section, we shall consolidate this idea on the determination of maximum likelihood estimator and, for completeness, introduce the estimation of information matrix (by Mehra(1976)). Finally, some well-known asymptotics of MLE will be reviewed.

We shall consider the following system (2.8.1):

$$x(t+1) = Ax(t) + B\varepsilon(t)$$
$$y(t) = Cx(t) + \varepsilon(t)$$

where 1. $x(0) \sim N(\mu,\Sigma)$.

2. $\varepsilon(t)$'s are independent white noise normally distributed with mean 0 and variance Ω.

3. $\varepsilon(t)$'s and $x(0)$ are independent.

The parameters of interest of system (2.8.1) is $\theta = (A,B,C,\Omega)$. Note that μ and Σ are not included into our parameters list because (1) under stationary condition, Σ can be determined by (2.4.4) and (2) μ can be estimated by $\hat{x}^T(0)$ using the Kalman Smoother. However, we do assume $\hat{x}(0)=0$ to start off the Kalman Filter. Given system (2.8.1), the likelihood function is given as follow:

$$(2.8.2) \qquad -2\log(L) = \sum_{j=1}^{T} \{ \log(\det Q_t) + \mathrm{tr}(Q_t^{-1} \omega(t)\omega(t)') \}$$

where 1. $\bar{x}(t) = E(x(t) \mid y(1),\ldots,y(t-1))$

2. $\omega(t) = y(t) - C\bar{x}(t)$

3. $Q_t = E(\omega(t)\omega(t)')$

All quantities appearing above can be calculated using Kalman Filter provided the parameters are known a-priori. Therefore, before we are able to determine the minimum point of (2.8.2) defined as maximum likelihood estimate, we must have available an algorithm to obtain a descent initial parameters estimate. One of the contribution of this thesis is to introduce a method to obtain such initial estiamte. Therefore details of deriving the initial estimate will be discussed later but the procedures can be outlined as AR approximation, Hankel matrix approximation and model reduction via balanced realisation. The final parameters estimate provided by balanced realisation is used to act as initial estimate to find the maximum likelihood estimate (MLE).

Suppose an initial estimate θ_0 is obtained, we can select an optimisation algorithm to search for $\hat{\theta}$ which corresponds to at least local minimum of (2.8.2). One of the popular alogrithm is Quasi-Newton Method. The quasi-newton method is, in a sense, an approximation to the traditional Newton method. However, it has the advantage of avoiding the computation of Hessian matrix as

well as its inverse. For detail we refer to Rao(1985).

It might be interesting if we can determine the gradient vector directly. One of the methods was provided by Mehra(1976) assuming that parameter C is constant. We define the parameter vector by $\theta=(vec(A)',vec(B)',vec(\Omega)')'$ and denote the ith component of θ by θ_i. Since $e(t)=Q_t^{-1/2}v(t)$ is white noise with constant variance, i.e. $E\{e(t)\}=0$ and $E\{e(t)e(s)\}=I\delta(t,s)$, we can write the log-likelihood as follows (2.8.3):

$$\log L = -\frac{1}{2}\sum_{t=1}^{T}\{\, 2\log|\Gamma_t| + e(t)'e(t)\,\}$$

where $\Gamma_t = Q_t^{1/2}$. By differentiating (2.8.3) with respect to θ_i, we have (2.8.4)

$$\frac{\partial \log L}{\partial \theta_i} = -\sum_{t=1}^{T}\{\, tr(\Gamma_t^{-1}\frac{\partial \Gamma_t}{\partial \theta_i}) + e(t)'\frac{\partial e(t)}{\partial \theta_i}\,\}$$

Thus, $\frac{\partial e(t)}{\partial \theta_i}$ is crucial in determining the gradient vector. From $e(t)=\Gamma_t^{-1}v(t)$,

$$\frac{\partial e(t)}{\partial \theta_i} = \frac{\partial \Gamma_t^{-1}}{\partial \theta_i}v(t) - \Gamma_t^{-1}C\frac{\partial \bar{x}(t)}{\partial \theta_i}. \tag{2.8.5}$$

However, from expressions in Kalman Filter, $\bar{x}(t+1)=A\bar{x}(t)+K(t)v(t)$ where $K(t)=(AH_tC'+B\Omega)Q_t^{-1}$. Therefore it follows that

$$\frac{\partial \bar{x}(t+1)}{\partial \theta_i} = (A-K(t)C)\frac{\partial \bar{x}(t)}{\partial \theta_i} + \frac{\partial K(t)}{\partial \theta_i}v(t) \tag{2.8.6}$$

with $\frac{\partial \bar{x}(0)}{\partial \theta_i} = 0$. Hence we can determine (2.8.4) for every i except the derivatives of Γ_t and Γ_t^{-1} with respect to θ_i which might be very difficult to evaluate. Besides the above method introduced by Mehra(1976), Shumway and Stoffer(1982) considered an alternate form of log-likelihood function of system (2.5.1) to develop a more explicit expression for determining the maximum likelihood estimators via application of EM algorithm. We shall talk about this when discussing ways to evaluate missing data score function in later chapters.

After developing algorithm for the gradient vector, Mehra(1976) went on to the calculation of Fisher information matrix defined as $E\{(\frac{\partial \log L}{\partial \theta})(\frac{\partial \log L}{\partial \theta})'\} = (M_{ij})$. We can show that

$$M_{ij} = \sum_{t=1}^{T}\sum_{s=1}^{T} tr(\Gamma_t^{-1}\frac{\partial \Gamma_t}{\partial \theta_i})tr(\Gamma_s^{-1}\frac{\partial \Gamma_s}{\partial \theta_j}) + \sum_{t=1}^{T} tr\{E(\frac{\partial e(t)}{\partial \theta_i}\cdot\frac{\partial e(t)'}{\partial \theta_j})\}.$$

From (2.8.5),

$$E(\frac{\partial e(t)}{\partial \theta_i}\cdot\frac{\partial e(t)'}{\partial \theta_j}) = \Gamma_t^{-1}\frac{\partial \Gamma_t}{\partial \theta_i}\cdot\frac{\partial \Gamma_t}{\partial \theta_j}\Gamma_t^{-1} + \Gamma_t^{-1}CE(\frac{\partial \bar{x}(t)}{\partial \theta_i}\cdot\frac{\partial \bar{x}(t)'}{\partial \theta_j})C'\Gamma_t^{-1}$$

The expressions for $E(\frac{\partial \bar{x}(t)}{\partial \theta_i}\cdot\frac{\partial \bar{x}(t)'}{\partial \theta_j})$ may be obtained by the augmented state vector

$$\bar{x}_A(t) = \begin{pmatrix} \bar{x}(t) \\ \frac{\partial \bar{x}(t)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \bar{x}(t)}{\partial \theta_m} \end{pmatrix}$$

From previous relations, we understand that

$$\bar{x}_A(t+1) = \Phi_A(t)\bar{x}_A(t) + K_A(t)\omega(t)$$

where

$$\Phi_A(t) = \begin{bmatrix} A & 0 & \cdots & 0 \\ \frac{\partial A}{\partial \theta_1} & A-K(t)C & \cdots & 0 \\ & \cdots\cdots\cdots & & \\ \frac{\partial A}{\partial \theta_m} & 0 & \cdots & 0 \end{bmatrix} \text{ and } K_A(t) = \begin{bmatrix} K(t) \\ \frac{\partial K(t)}{\partial \theta_1} \\ \vdots \\ \frac{\partial K(t)}{\partial \theta_m} \end{bmatrix}$$

Define $H_A(t) = E\{\bar{x}_A(t)\bar{x}_A(t)'\}$, we have

$$E\{\bar{x}_A(t+1)\} = \Phi_A(t)\Phi_A(t-1)\cdots\Phi_A(0)E\{\bar{x}_A(0)\}$$

and

$$H_A(t+1) = \Phi_A(t)H_A(t)\Phi_A(t)' + K_A(t)Q_tK_A(t)'.$$

If we can assume $E\{\bar{x}_A(0)\}=0$ and $H_A(0)=diag\{\Sigma,0,\ldots,0\}$ then the (i,j) block matrix will give us the required expressions. Hence this completes the derivation of Fisher information matrix introduced by Mehra(1976). We shall talk more on derivation of Fisher information matrix with missing data in later chapters.

Instead of considering the asymptotic properties of maximum likelihood estimator, Otter(1985) considered general prediction

error estimation of the form $y(t)=f(y^{(t-1)},t,\theta)+\varepsilon(t,\theta)$, $t\geq 1$ where $y(t)$ is the output vector; $y^{(t-1)}$ is the vector consisting of previous values of output vector and previous values of input values; θ is the parameter vector. Let the prediction model be given by $\hat{y}(t|\theta)=f(y^{(t-1)},t,\theta)$ then the prediction error is $\varepsilon(t,\theta)=y(t)-y(t|\theta)$. Let $\ell(\varepsilon(t,\theta),\theta,t)$ be some scalar measure then a criterion for the validity of the prediction model after T observations of the output vector is given by

$$V_T(\theta,y^{(T)}) = \frac{1}{T}\sum_{t=1}^{T} \ell(\varepsilon(t,\theta),\theta,t).$$

For example:

1. Least square: $\ell(\varepsilon(t,\theta),\theta,t) = \varepsilon(t,\theta)^2$
2. log-likelihood: $\ell(\varepsilon(t,\theta),\theta,t) = -\log p_{\varepsilon}(\varepsilon(t,\theta))$ where p_{ε} is the pdf of $\varepsilon(t,\theta)$.

Let $\hat{\theta}(T)$ be such that $V_T(\hat{\theta}(T),y^{(T)})$ is minimum then assuming the following conditions:

1. the input sequence $\{u(t)\}$ is "persistently exciting", i.e.

   $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u(t)u(t-j)' = r(j)$ exists for all j. Under such conditions, we can show that the impulse response matrix can be determined uniquely from the input-output statistics.
2. The limit

   $$\bar{V}(\theta) = \lim_{T\to\infty} E_{y^{(T)}|\theta}\{V_T(\theta,y^{(T)})\}$$

   exists.

Then we have the following result due to Ljung and Soderstrom(1983):

Theorem 2.8.7

Under weak regularity conditions $\hat{\theta}(T)$ converges with probability one to θ^* such that $\bar{V}(\theta^*)$ is a minimum of $\bar{V}(\theta)$ as T tends to infinity. It means that a model with as estimate $\hat{\theta}(T)$ gives the "best" description of the data, "best" measured in terms of the expected value of the criterion. Moreover, if $\hat{\theta}(T)$ converges to θ^* such that $\bar{V}''(\theta^*) = \frac{\partial^2\bar{V}(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\theta^*}$ is invertible then

$$\sqrt{T}\,(\hat{\theta}(T) - \theta^*) \xrightarrow{L} N(0,P)$$

where $P = \bar{V}''(\theta^*)^{-1} \lim_{T\to\infty} TE\{V_T'(\theta^*)V_T'(\theta^*)^T\}\bar{V}''(\theta^*)^{-1}$, $V_T'(\theta)=\frac{\partial V_T(\theta)}{\partial\theta}$ and T denote the matrix transposition.

With regard to maximum likelihood estimation, P can be shown to be equal to Fisher Information Matrix per sample. Therefore the maximum likelihood estimator has a limiting variance which is equal to the Fisher information matrix per sample.(see Otter(1985)), i.e.

$$\sqrt{T}\ (\hat{\theta}(T) - \theta^*) \xrightarrow{L} N(0,(\lim_{N\to\infty} \frac{1}{N} I_N(\theta^*))^{-1})$$

where $I_N(\theta^*)$ is the Fisher information matrix for N observations.

3. Missing Observations in time series analysis

Missing observations are an inevitable part of our data set. It has been troubling statisticians whenever there are data. One can easily imagine the ease of getting "holes" in a large scale survey, the backbone of most data sources. With respect to time series data, missing observations can also occur. For example, a manufacturer who would like to collect information on daily production of one of its plants is highly probable to find that a fraction of the daily production data are missing. So, how and what can we do with missing observations?

For independent data, early treatment of missing observations could be one of the followings:

1. ignore the missing observations and proceed as if there were no missing observations.
2. impute the missing observations by the mean of observed data and proceed if there were no missing data. An alternative to mean imputation is to impute the observed data mean plus a random noise drawn from a suitably chosen distributions.
3. impute the missing observations by some predicted value using auxiliary information and proceed as if there were no missing data. Similarly, besides imputing the predicted values, we can add noise as well.

As you might have noticed, methods (1) and (2) regards the observed data as random sub-sample of the original complete data sample while that of method (3) is true only when conditional on auxiliary information. However, in terms of not disrupting the original distribution of data, method (3) is the best among the above three methods. In fact, all methods share the following points:

1. the intention to retain the maximum use of complete data algorithms so that standard statistical package can be used without much modifications.
2. the representation of idea to get around the problem without formal justification of correctness.

Despite the points, people had used the described or similar methods for many years until attention was brought by Rubin(1976) and the coming of more theoretical results. A rigorous treatment

of missing observations will be discussed later.

On the other hand, as far as time series data are concerned, it is impossible to simply ignore the missing data points and proceed if there were no missing observations because of an additional time dimension. The difficulty lies on the fact that we only have one data for each time point (time series is interpreted as a series of random variables of time). Early treatment of missing observations are normally one of the following methods:

1. By ignoring the missing observations and part of observed data, least square criteria is used to obtain some estimate of parameters. For example, consider modelling data with AR(1) model: $y(t)=\rho y(t-1)+\varepsilon(t)$, we find estimate of parameter ρ by minimising the least square criteria: $\sum \{y(t)-\rho y(t-1)\}^2$ where the summation is performed whenever $y(t)$ and $y(t-1)$ are both observed. Then by employing some "optimal" interpolation procedure we can, with the estimate of ρ, interpolate the missing observations and re-estimate the parameter again.
2. Thanks to Parzen(1963) who developed consistent estimate of autocovariances sequence when the missing data indicators interpreted as time series are stationary, we are able to use Durbin-Levinson algorithm to produce initial AR parameters estimate, model comparsion, etc.. However, this method is limited to univariate time series analysis until Stoffer(1986) who generalised the idea to multivariate case and it may have difficulty in estimating the MA part of ARMA models since most existing algorithms, like Hannan-Rissanen algorithm, often need complete data to construct estimates.
3. Instead of considering the original time series with missing observations, we can consider parameters estimation of a new series which has zeros at missing data points. (see Brockwell and Davies(1991) and Jong(1992)) Although this method represents a very interesting way of getting around the problem, it is by no means a reasonable method. Afterall imputing zeros to missing data points could create unnecessary fluctuation of original time series and thereby promote instability of our parameter estimates. Nevertheless, we shall

talk about this method later in this section.

However, as state space modelling becomes more popular, there are increasing tendency to use state space modelling to solve missing data problems in time series. The major reason is because state space model, in fact, represents a decomposition of joint probability distribution of our data series into product of conditional distributions, i.e.

$$f(y(t_1),\ldots,y(t_T)) = f(y(t_1))f(y(t_2)|y(t_1))\cdots f(y(t_T)|y(t_1),..,y(t_{T-1}))$$

Moreover, practically speaking, it provides a very simple but still theoretically justifiable way of solving this missing data problems. Using the fact that stationary stable state space models are in fact equivalence to causal ARMA models (i.e. there exists one-one correspondence between the two types of models), we can see that state space modelling can be an efficient method of estimating MA part of ARMA models even in the presence of missing observations. Adjustment procedure to missing observations will be discussed shortly.

No matter what types of data we are talking about, statisticains are always looking for adjustment methods to missing observations without much modifications of existing complete data algorithms and major theoretical results. It is also this reason that people often have the intention of simply ignoring the missing observations or, more specifically, ignoring the underlying missing data generating mechanism. That why Rubin(1976) asked when this is the proper way to go.

3.1 When Missing data is ignorable

To express the question more plainly, we use the following notations:

1. Let $Y = (Y_1,\ldots,Y_T)$ be vector random variable with probability density $f(\cdot|\theta)$, where θ is the vector parameter of this density.
2. Let $M = (M_1,\ldots,M_T)$ be the associated "missing data indicator" vector random variable, where each M_1 takes the value 0 or 1 corresponding to Y_1 is missing or observed respectively. The probability that M takes the value $m = (m_1,\ldots,m_T)$ given that Y takes the value $y = (y_1,\ldots,y_T)$ is $g(m|y,\phi)$ where ϕ is the

nuisance vector parameter of the distribution.

Our objective is to make inferences about θ but not ϕ. The conditional distribution $g(m|y,\phi)$ corresponds to "the process that causes missing data". In addition, what we have observed is not y but the pair (y_0,m) where $y = (y_0,y_M)$ with y_0 the observed part of y.

Suppose that $\tilde{y}$ and $\tilde{m}$ are particular sample realisation of Y and M respectively. The practice of ignoring the process that causes missing data means proceeding by (a) fixing the random variable M at the observed pattern of missing data, $\tilde{m}$, and (b) assuming that the values of the observed data $\tilde{y}_0$ arose from the marginal density of the random variable (3.1.1) Y_0: $\int f(y|\theta)dy_M$. Then central question, in Rubin(1976) term, is concerned with the weakest simple conditions on g such that ignoring the process that causes missing data will always yield proper inferences about θ. Although the answer still depends on the types of inferences we are pursuing, it is rather simple if we have the following classification of missing observations.

Definition 3.1.2

The missing data are missing at random (MAR) if for each value of ϕ, $g(\tilde{m}|\tilde{y},\phi)$ takes the same value for all y_M, i.e.

$$g(\tilde{m}|\tilde{y},\phi) = g(\tilde{m}|\tilde{y}_0,\phi).$$

Definition 3.1.3

The observed data are observed at random (OAR) if for each value of ϕ and y_M, $g(\tilde{m}|y,\phi)$ takes the same value for all y_0, i.e.

$$g(\tilde{m}|y,\phi) = g(\tilde{m}|y_M,\phi).$$

Definition 3.1.4

The data are missing completely at random (MCAR) if the missing data are MAR and the observed data are OAR. (see Rubin and Little(1987))

Definition 3.1.5

The parameter ϕ is distinct from θ if their joint parameter space factorises into a ϕ-space and a θ-space, i.e. $\Omega_{\phi,\theta} = \Omega_\phi \times \Omega_\theta$.

3.1(a) Sampling Distribution Inference

A sampling distribution inference is an inference that results solely from comparing the observed value of a statistic with the sampling distribution of that statistic under various hypothesized underlying distributions. Let the statistic be $S(y)$ and suppose our objective is to make inference on θ, then ignoring the process that causes missing data means deriving the sampling distribution of $S(y_0)$ from the marginal density (3.1.1) of Y_0. However, the correct sampling distribution of $S(y_0)$ should be the conditional density of Y_0 given that $M = \tilde{m}$, i.e. (3.1.6)

$$\int \{f(y|\theta)g(\tilde{m}|y,\phi)/k(\tilde{m}|\theta,\phi)\}dy_M$$

where $k(\tilde{m}|\theta,\phi) = \int f(y|\theta)g(\tilde{m}|y,\phi)dy$ is the marginal probability that M takes the value $\tilde{m}$. We see that (3.1.1) is equal to (3.1.6) when the value of g in numerator can be cancelled from that in denominator. Therefore we have the following result:

Theorem (3.1.7) (Rubin(1976))

Suppose that (a) the missing data are missing at random and (b) the observed data are observed at random. Then the sampling distribution of $S(y_0)$ under $f(\cdot|\theta)$ ignoring the process that causes missing data, i.e. calculated from (3.1.1), equals the correct conditional sampling distribution of $S(y_0)$ given $\tilde{m}$ under $f(\cdot|\theta)g(\cdot|y,\phi)$, i.e. calculated from (3.1.6) assuming $k(\tilde{m}) > 0$.

Note that, with assumptions (a) and (b), $g(\tilde{m}|y,\phi) = g(\tilde{m}|\phi)$ and therefore the result follows.

Thus, only when the data are MCAR ignoring the process that causes missing data is valid in sampling distribution inference.

3.1(b) Likelihood Inference

A likelihood inference is an inference that results solely from ratios of the likelihood function for various values of parameter. In this case θ and ϕ take values in a joint parameter space $\Omega_{\theta,\phi}$. Ignoring the process that causes missing data when making likelihood inference on θ means defining a parameter space Ω_θ for θ and taking ratios, for various $\theta \in \Omega_\theta$, of the likelihood function based on density (3.1.1):

$$\mathcal{L}(\theta|\tilde{y}_0) \propto \int f(\tilde{y}|\theta)dy_M. \qquad (3.1.8)$$

However, the correct likelihood should be the joint likelihood of the observed data $\tilde{y}_0$ and $\tilde{m}$:

(3.1.9) $$\mathcal{L}(\theta,\phi|\tilde{y}_0,\tilde{m}) \propto \int f(\tilde{y}|\theta)g(\tilde{m}|\tilde{y},\phi)dy_M.$$

By comparing expressions in (3.1.8) and (3.1.9), we may see that the likelihood ratio derived from (3.1.8) equals that derived from (3.1.9) when $g(\tilde{m}|\tilde{y},\phi)$ takes the same value for all y_M, i.e. the missing data are missing at random. Thus, we have the following result:

Theorem 3.1.10 (Rubin(1976))

Suppose (a) that the missing data are missing at random, and (b) that ϕ is distinct from θ. Then the likelihood ratio ignoring the process that causes missing data, i.e. $\mathcal{L}(\theta_1|\tilde{y}_0)/\mathcal{L}(\theta_2|\tilde{y}_0)$, equals the correct likelihood ratio, i.e. $\mathcal{L}(\theta_1,\phi|\tilde{y}_0,\tilde{m})/\mathcal{L}(\theta_2,\phi|\tilde{y}_0,\tilde{m})$, for all $\phi \in \Omega_\phi$ such that $g(\tilde{m}|\tilde{y},\phi) > 0$.

3.1(c) Bayesian Inference

A Bayesian inference is an inference that results solely from posterior distributions corresponding to a specified prior distribution. We shall in general use $p(\cdot)$ to signify prior density. Ignoring the process that causes missing data means choosing $p(\theta)$ and assuming that the observed data $\tilde{y}_0$ arose from density (3.1.1). Thus this implies the posterior distribution of θ ignoring the process that causes missing data is proportional to

(3.1.11) $$p(\theta)\int f(\tilde{y}|\theta)dy_M.$$

However, similar to other cases, the correct marginal posterior distribution of θ should be derived from the joint posterior distribution of (θ,ϕ) and then integrating out ϕ. It is easy to see that the joint posterior distribution of (θ,ϕ) is proportional to

(3.1.12) $$p(\theta,\phi) = p(\theta)p(\phi|\theta)\int f(\tilde{y}|\theta)g(\tilde{m}|\tilde{y},\phi)dy_M.$$

Hence the correct posterior distribution of θ is proportional to

(3.1.13) $\int p(\theta,\phi)d\phi$. We may therefore see that (3.1.11) equals to (3.1.13) when g takes the same value for all y_M, i.e. when the missing data is MAR and $p(\phi|\theta) = p(\phi)$. In essence we have the following theorem given in Rubin(1976):

Theorem (3.1.14)

Suppose (a) that the missing data are missing at random, and (b) that ϕ is distinct from θ. Then the posterior distribution of θ ignoring the process that causes missing data, i.e. calculated from (3.1.11), equals the correct posterior distribution of θ, that is calculated from (3.1.12), and the posterior distributions for θ and ϕ are independent.

With assumptions (a) and (b) (3.1.12) becomes

$$\{p(\theta)\int f(\tilde{y}|\theta)dy_M\}\{p(\phi)g(\tilde{m}|\tilde{y}_o,\phi)\}$$

and therefore the result follows.

From the above results, we see that concepts of MAR and MCAR are critical for ignoring the process that causes missing data to be valid. Under MCAR, the observed data can be easily seen as a random sub-sample of complete data and therefore ignoring the process that causes missing data does not affect the theoretical validity of ultimate result. Because of the lack of knowledge of missing observations, practical justification of MCAR is very difficult if not impossible unless we know explicitly what g is. On the other hand, as opposite to MCAR, the weaker condition MAR is somewhat difficult to be provided a proper physical interpretation except the understanding that the probability of missingness does not depend on the missing data. However, if we have an auxiliary information X with sample realisation $\tilde{x}$ such that $g(\tilde{m}|\tilde{y},\tilde{x})=g(\tilde{m}|\tilde{x})$ then the data is missing at random and the $\tilde{y}$ values defined by sub-classes of $\tilde{x}$ form a random sub-sample within sub-classes. Hence ignoring the process that causes missing data would be appropriate if our inference is conditional on $\tilde{x}$. Again, practical justification of the relation $g(\tilde{m}|\tilde{y},\tilde{x})=g(\tilde{m}|\tilde{x})$ may be difficult unless M can be explained perfectly by X. Although the justification of MCAR or MAR may be hard, it could be "partly" done with a depth understanding of actual data collection process.
Since the missing data indicator is binary, one statistical way to justify the relationship between the indicator and some auxiliary variable X (i.e. to justify MAR) is via linear logit analysis. The linear logistic model can be written as

$$\Pr(R=1|\tilde{x}) = \exp(\alpha+\beta'\tilde{x})/\{1+\exp(\alpha+\beta'\tilde{x})\} \qquad (3.1.15)$$

where R is the missing data indicator and $\tilde{x}$ is the observed value of X. Assume that R is a bernoulli random variable with parameter π. Hence for T observations, the log-likelihood function of R is given as

$$(3.1.16) \qquad l(\pi;R(i),i=1,..,T) = \Big(\sum_{i=1}^{T} R(i)\Big)\log\Big\{\frac{\pi}{1-\pi}\Big\} + T\log(1-\pi).$$

Using (3.1.15), (3.1.16) becomes (3.1.17)

$$l(\beta;R(i),i=1,..,T) = \sum_{i}\sum_{j} R(i)\tilde{x}_{ij}\beta_j - \sum_{i} \log\{1+\exp\sum_{j}\tilde{x}_{ij}\beta_j\}.$$

where $\tilde{x}_{ij}$ is the jth observation of X_i and β_j the jth parameter.

Therefore, given (3.1.17), maximum likelihood estimation of β can be performed and by looking at the decrease in residual deviance, we can observe how much the missing data indicator can be explained by X. This will be elaborated in subsequent chapters when we apply logit analysis to real data. For more detail please see McCullagh and Nelder(1989).

As a summary, if we are given the data Y, with missing observations, and auxiliary information X then the data is MCAR, MAR or Missing not at random are according to whether the process of response is independent of both X and Y, depends on X but not on Y or depends on Y and possibly on X respectively. For more details, we refer to Rubin and Little(1987) and Rubin(1976).

3.2 Maximum Likelihood Estimation and EM algorithm

In this section, a general presentation of maximum likelihood estimation with missing observations are considered without refering to particular application and, for completeness, literature of EM algorithm are also presented although we did not apply it to real data (Note: we shall discuss the reasons in appropriate section.).

Consider a random vector $Y \in \mathbb{R}^T$ with particular realisation $\tilde{y}$. Suppose that $\tilde{y}$ is not fully observed and can be written as $(\tilde{y}_0', \tilde{y}_M')'$ where $\tilde{y}_0$ denotes the observed part and $\tilde{y}_M$ denotes the missing part. Let M be vector of missing data indicator with ith component defined as

$$M_i = \begin{cases} 1 \text{ if } Y_i \text{ is missing} \\ 0 \text{ otherwise} \end{cases}.$$

and let $\tilde{m}$ be particular realisation of M. Furthermore, we assume that Y follows a distribution with density $f(\cdot|\theta)$ and that the conditional density of M given y is written as $g(\cdot|y,\phi)$.

Given the above setting, we have parameters (θ,ϕ) and our objective is to find $(\hat{\theta},\hat{\phi})$ such that the joint likelihood defined in (3.1.9) is maximised. We write $l(\theta,\phi|\tilde{y}_0,\tilde{m})$ to be $\log\mathcal{L}(\theta,\phi|\tilde{y}_0,\tilde{m})$. The problem may be equivalently posed as maximising the log-likelihood function. Since we have the identity $\mathcal{L}(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})=\mathcal{L}(\theta,\phi|\tilde{y}_0,\tilde{m})p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)$, the following relation are established:

$$\text{(3.2.1)}\qquad \frac{\partial l(\theta,\phi|\tilde{y}_0,\tilde{m})}{\partial\vartheta} = \frac{\partial l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta} - \frac{\partial \log p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)}{\partial\vartheta}$$

where ϑ can be either θ or ϕ.

This also implies that

$$\frac{\partial l(\theta,\phi|\tilde{y}_0,\tilde{m})}{\partial\vartheta} = \int \frac{\partial l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta} p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)dy_M$$
$$- \int \frac{\partial \log p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)}{\partial\vartheta} p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)dy_M.$$

Since the last term is easily seen to be zero, we have the expression for missing data score function:

$$\text{(3.2.2)}\qquad \frac{\partial l(\theta,\phi|\tilde{y}_0,\tilde{m})}{\partial\vartheta} = E\{ \frac{\partial l(\theta,\phi|y,\tilde{m})}{\partial\vartheta} | \tilde{y}_0,\tilde{m} \},$$

where the expectation is taken over the missing data y_M conditional on $\tilde{y}_0$ and $\tilde{m}$. Moreover, given sufficient regularity conditions,

$$\frac{\partial^2 l(\theta,\phi|\tilde{y}_0,\tilde{m})}{\partial\vartheta^2} = \int \frac{\partial^2 l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta^2} p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)dy_M$$
$$+ \int \frac{\partial l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta} \cdot \frac{\partial \log p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)}{\partial\vartheta} p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)dy_M.$$

From (3.2.1), we have

$$\text{(3.2.3)}\qquad \frac{\partial^2 l(\theta,\phi|\tilde{y}_0,\tilde{m})}{\partial\vartheta^2} = \int \frac{\partial^2 l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta^2} p(y_M|\tilde{y}_0,\tilde{m},\theta,\phi)dy_M$$
$$+ \text{var}\left\{ \frac{\partial l(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})}{\partial\vartheta} \middle| \tilde{y}_0,\tilde{m} \right\}$$

Therefore, we have the following important result for inference with missing observations:

Theorem (3.2.4)

(a) The missing data score function is given by

$$Sc(\theta,\phi|\tilde{y}_0,\tilde{m}) = E\{\ Sc(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})|\tilde{y}_0,\tilde{m}\ \}$$

where $Sc(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})$ is the complete data score function.

(b) The observed information matrix with missing observations is given by

$$Info(\theta,\phi|\tilde{y}_0,\tilde{m}) = E\{\ Info(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})|\tilde{y}_0,\tilde{m}\ \} - var\{\ Sc(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})|\tilde{y}_0,\tilde{m}\ \}$$

where $Info(\theta,\phi|\tilde{y}_0,y_M,\tilde{m})$ is the complete data information matrix.

Similar theorem was already given by Louis(1982), Tanner(1990), Breckling, Chambers, Dorfman, Tam and Welsh(1990) and was also already hinted by Fisher(1925) who showed that $\mathbf{DL}^y(\theta)=E_\theta(\mathbf{DL}^x(\theta)|y)$ where $y(x)$ is some statistic of x and $\mathbf{D}$ is the differential operator.

Moreover, if (a) the missing data is missing at random (MAR) and (b) θ is distinct from ϕ then we see the following relation:

$$l(\theta,\phi|\tilde{y}_0,\tilde{m}) = \log\{f(\tilde{y}_0|\theta)\} + \log\{g(\tilde{m}|\tilde{y}_0,\phi)\}$$

Since our main objective is to find maximum likelihood estimate of θ, it therefore suffices to concentrate on $l(\theta|\tilde{y}_0)$. However, we would like to warn reader the fact that adopting this likelihood function doesn't mean we are ignoring the missing data and proceed as if there were no missing observations. In fact, since the joint log-likelihood is decomposed into two parts, one for θ and the other for ϕ, we thus see that the score function and the observed information matrix with missing observations (with respect to θ only) are determined by the following equations:

$$Sc(\theta|\tilde{y}_0) = E\{\ Sc(\theta|\tilde{y}_0,y_M)|\tilde{y}_0\ \}$$

$$Info(\theta|\tilde{y}_0) = E\{Info(\theta|\tilde{y}_0,y_M)|\tilde{y}_0\} - var\{Sc(\theta|\tilde{y}_0,y_M)|\tilde{y}_0\}$$

under assumptions (a) and (b).

We also remark that the score function and information matrix jointly define the Newton-Raphson sequence $\theta^{(1)}$ which represents successive approximation to the maximum likelihood estimator.

Given $\theta^{(0)}$, the (i+1)th iteration is defined as:

$$\theta^{(i+1)} = \theta^{(i)} + \lambda^{(i)} \mathrm{Info}(\theta^{(i)} | \tilde{y}_0)^{-1} \mathrm{Sc}(\theta^{(i)} | \tilde{y}_0)$$

where $0 < \lambda^{(i)} \leq 1$ is a sequence of step-lengths included to ensure convergence of successive iterations. For methods to determine suitable step-lengths, we refer to Ortega and Rheinboldt(1970). In coming section, we shall try to justify the asymptotic distribution of MLE with missing observations in time series case.

Interestingly enough, there is method to determine MLE without directly deriving the missing data likelihood function. This method is collectively called EM algorithm by Dempster, Laird and Rubin(1977). Although the major application of EM algorithm is confined in solving missing data problems, it also has application to complete data problems. For example, Shumway and Stoffer(1982) used EM algorithm to find MLE of parameters of special state space model. As given in Dempster, Laird and Rubin(1977) we have the following definitions:

Definition (3.2.5)

The EM iteration $\theta^{(p)} \mapsto \theta^{(p+1)}$ is defined as follows:

E-step: Compute $Q(\theta | \theta^{(p)})$;

M-step: Find $\theta^{(p+1)}$ to be a value $\theta \in \Omega_\theta$ which maximises $Q(\theta | \theta^{(p)})$ where

$$Q(\theta' | \theta) = E\{ \log f(\tilde{y}_0, y_M | \theta') | \tilde{y}_0, \theta \}$$

Moreover, we write $\theta^{(p+1)} = EM(\theta^{(p)})$.

Definition (3.2.6)

The Generalised EM iteration $\theta^{(p)} \mapsto \theta^{(p+1)}$ is defined as:

E-step: Compute $Q(\theta | \theta^{(p)})$;

M-step: Find $\theta^{(p+1)}$ to be a value $\theta \in \Omega_\theta$ such that

$$Q(\theta^{(p+1)} | \theta^{(p)}) \geq Q(\theta^{(p)} | \theta^{(p)}).$$

Moreover we write $\theta^{(p+1)} = GEM(\theta^{(p)})$.

Remark:

1. EM algorithm is a special case of GEM algorithm.
2. Define the conditional density of complete data $(\tilde{y}_0', y_M')'$ given the data $\tilde{y}_0$ by $k(\tilde{y}_0, y_M | \tilde{y}_0, \theta)$ and, for convenience, we write

   $$H(\theta' | \theta) = E\{ \log k(\tilde{y}_0, y_M | \tilde{y}_0, \theta') | \tilde{y}_0, \theta \}$$

   Then for any pair $(\theta', \theta) \in \Omega_\theta, \times \Omega_\theta$, $l(\theta' | \tilde{y}_0) = Q(\theta' | \theta) - H(\theta' | \theta)$

and $H(\theta'|\theta) \le H(\theta|\theta)$.

3. Any instance of GEM algorithm $\{\theta^{(p)}\}$ has the property: $l(\theta^{(p+1)}|\tilde{y}_0) \ge l(\theta^{(p)}|\tilde{y}_0)$. This property is followed from remark (2) and was given as Theorem 1 in Dempster, Laird and Rubin(1977).

Problems:

Since any EM sequence $\{\theta^{(p)}\}$ defines the sequence $\{l(\theta^{(p)}|\tilde{y}_0)\}$, what is the convergent conditions for the sequence? Moreover, if there exists l^* such that $l(\theta^{(p)}|\tilde{y}_0) \to l^*$, does it imply that l^* is some maximum point of likelihood function and does it imply that $\theta^{(p)} \to \theta^*$ such that $l(\theta^*|\tilde{y}_0) = l^*$? If not, under which conditions they will?

All of the above questions are interesting and was provided in detail by Wu(1982). The followings are summary of his results:

Results:

(3.2.7) If $\{l(\theta^{(p)}|\tilde{y}_0)\}$ is bounded above where $\{\theta^{(p)}\}$ is any instance of EM algorithm then the sequence converges to some l^*.

(3.2.8) Suppose Q satisfies the continuity condition: $Q(\psi|\phi)$ is continuous in both ψ and ϕ. Then all the limit points of any instance $\{\theta^{(p)}\}$ of an EM algorithm are stationary points of l and $l(\theta^{(p)}|\tilde{y}_0)$ converges monotonically to $l^* = l(\theta^*|\tilde{y}_0)$ for some stationary point θ^*.

(3.2.9) Suppose Q satisfies the continuity condition stated above and $\sup\{ Q(\theta'|\theta) | \theta' \in \Omega_\theta \} > Q(\theta|\theta)$ for any θ which is a stationary point but not a (local) maxima. Then all the limit points of any instance $\{\theta^{(p)}\}$ of an EM algorithm are local maxima of l and $l(\theta^{(p)}|\tilde{y}_0)$ converges monotonically to $l^* = l(\theta^*|\tilde{y}_0)$ for some local maximum θ^*.

(3.2.10) Suppose $\mathbf{D}^{10}Q(\theta'|\theta)$ is continuous in θ' and θ. Then $\theta^{(p)}$ converges to a stationary point θ^* with $l(\theta^*|\tilde{y}_0) = l^*$, the limit of $l(\theta^{(p)}|\tilde{y}_0)$, if either (a) $\mathcal{L}(l^*)=\{\theta \in \Omega_\theta: l(\theta|\tilde{y}_0)=l^*\} = \{\theta^*\}$, or (b) $\{\theta^{(p)}\}$ is a cauchy sequence and $\mathcal{L}(l^*)$ is discrete. Consequently, if $l(\theta|\tilde{y}_0)$ is unimodal with θ^* being the only stationary point then any EM sequence $\{\theta^{(p)}\}$ converges to θ^*.

(3.2.11) If the set of stationary points (local maxima) with a given l value, denoted $\mathcal{S}(l)$, is not discrete and $\{\theta^{(p)}\}$ is cauchy then $\theta^{(p)}$ converges to a compact, connected component of $\mathcal{S}(l)$ but

not necessarily to a point.

Unfortunately, only convergent to stationary points are guaranteed in most cases. In fact, even if we know that the EM sequence converges to a maximum, local maxima is most likely to be the case. (this also depends on the complexity of the likelihood.) However, this phenomenon is also common in other optimisation algorithms. Finally convergent of $\{l(\theta^{(p)}|\tilde{y}_0)\}$ does not imply convergence of the corresponding $\theta^{(p)}$ unless more stringent conditions are given. But, nonetheless, this is not as important as the convergence of the likelihood sequence.

Dempster, Laird and Rubin(1977) demonstrated that the convergent rate of EM algorithm depends directly on the relative sizes of $\mathbf{D}^{20}l(\theta^*|\tilde{y}_0)$ and $\mathbf{D}^{20}H(\theta^*|\theta^*)$ which is linear and represents the proportion of information about θ that is observed. Therefore, the convergent rate of EM algorithm is not as rapid as Newton-Raphson algorithm but this is compensated by not requiring inversion at each steps. In addition, Orchard and Woodbury(1972) developed the so called missing information principle which are stated as:

Observed Information = Complete Information - Missing Information.

where the missing information is equal to $\dfrac{-\partial^2 H(\theta|\phi)}{\partial\theta^2}$.

Tanner(1990) presented methods to numerically approximate this missing information and hence the observed information can be numerically evaluated.

We finish discussion of EM algorithm by considering an example of modelling ARMA models to time series with missing observations.

Given a time series $\{y_1,y_2,\ldots,y_T\}$ and assume that only $\{y(t_1),\ldots,y(t_n)\}$ are observed. Suppose we would like to consider ARMA(p,q): $\alpha(z^{-1})y(t) = \beta(z^{-1})\varepsilon(t)$ where, for simplicity, $\varepsilon(t) \sim N(0,1)$. Then our parameter would be (α,β) where α and β represent the coefficients of $\alpha(z^{-1})$ and $\beta(z^{-1})$ respectively.

At (p+1)th iteration given $\alpha^{(p)}$ and $\beta^{(p)}$,

E-Step:

Form covariance matrix Σ of $\{y(t_1),..,y(t_n),y(i_1),..,y(i_m)\}$ where $\{i_1,\ldots,i_m\}=\{1,2,\ldots,T\}\backslash\{t_1,\ldots,t_n\}$. This is theoretically possible given the current parameters estimate. Then it follows

that

$$\begin{pmatrix} y(i_1) \\ \vdots \\ y(i_m) \end{pmatrix} \Big| \begin{pmatrix} y(t_1) \\ \vdots \\ y(t_n) \end{pmatrix} \sim N(\Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix} y(t_1) \\ \vdots \\ y(t_n) \end{pmatrix}, \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$ are the corresponding partition.

Therefore we can determine $E(y(i_k)|y(t_1),..,y(t_n))$ and $E(y(i_h)y(i_k)|y(t_1),..,y(t_n))$.

Since the complete data log-likelihood is given as follow:

(3.2.12) $$-2\log f(y_1,..,y_T|\alpha,\beta) \propto \log|\Sigma| + tr\{\Sigma^{-1}yy'\}$$

where $y = (y_1,\ldots,y_T)'$.

Hence expected log-likelihood given the observed data in fact depends on the following substitutions to $yy' = (y(h)y(k))_{h,k}$:

1. When both y(h) and y(k) are missing, the product y(h)y(k) are replaced by its conditional expectation.
2. When either y(h) or y(k) is missing, only the missing value is replaced by its conditional expectation.

This completes the E-Step.

M-Step:

We find $\alpha^{(p+1)}$ and $\beta^{(p+1)}$ such that they jointly maximised the likelihood function (3.2.12) after the substitutions described in the E-Step was made. This completes the M-Step.

After presenting general statistical theory concerning missing observations, we are going to discuss literature of tackling missing observations in time series.

3.3 Parzen's idea to estimate covariance sequence

Parzen(1963) studied the relationship between autocovariance sequence of a time series $\{y(t),\ t \in \mathbb{N}\}$ and that of its "amplitude modulating" series $\{x(t)=g(t)y(t),\ t \in \mathbb{N}\}$ where $g(\cdot)$ is some bounded function such that the limit, for $v = 0,1,2,\ldots,$

$$R_g(v) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T-v} g(t)g(t+v)$$

exists. By defining $R_y(v)$ to be $E\{y(t)y(t+v)\}$, the autocovariance function of y(t) of lag v. Parzen(1963) showed that $x(\cdot)$ is not covariance stationary but asymptotically stationary with

autocovariance function $R_x(v)$ given by

$$R_x(v) = R_g(v)R_y(v)$$

when $y(\cdot)$ is stationary. (Note: that $y(\cdot)$ is mean zero has been assumed.) Moreover, if $y(\cdot)$ is an ergodic normal process then $x(\cdot)$ is also ergodic. (Ergodic means the sample autocovariance function is a consistent in quadratic mean estimate of R(v).) Therefore, we have the following interesting result from Parzen(1963):

Theorem 3.3.1 (Parzen(1963))

Let $\{y(t),\ t \in \mathbb{N}\}$ be stationary and normal with zero means and autocovariance function $R_y(v)$ satisfying

$$\lim_{T\to\infty} \sum_{t=1}^{T} R_y^2(v) = 0$$

so that $y(\cdot)$ is ergodic.

Suppose that the time series $y(\cdot)$ is not directly observed. Rather one observes a time series $x(\cdot)$ which is amplitude modulated version of $y(\cdot)$: $x(t) = g(t)y(t)$, $t=1,2,\ldots$, where $g(\cdot)$ is a non-random function possessing an autocovariance function $R_g(v)$ defined above. If $R_g(v) \neq 0$, $v = 0,1,2,\ldots$, a consistent in quadratic mean estimate of the covariance function $R_y(v)$ of the unobserved time series $y(\cdot)$ is given as

$$\hat{R}_y(v) = \hat{R}_x(v)/\hat{R}_g(v)$$

where

$$\hat{R}_x(v) = \frac{1}{T}\sum_{t=1}^{T-v} x(t)x(t+v) \text{ and } \hat{R}_g(v) = \frac{1}{T}\sum_{t=1}^{T-v} g(t)g(t+v).$$

The application of this theorem to time series with missing observations is immediate. Associated with each stationary and normal time series $y(\cdot)$ with missing observations, we can define a "amplitude modulating" series $x(\cdot)$ using the missing data indicator $M(\cdot)$, i.e. if $M(i) = \begin{cases} 1 & \text{if } y(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$, then $x(i) = M(i)y(i)$ is called the "amplitude modulating" series defined by $M(\cdot)$. Moreover, if $M(i)$ possesses finite autocovariances sequence then a consistent estimate of $R_y(v)$ is given as $\hat{R}_x(v)/\hat{R}_M(v)$. However, the result was proved in the univariate case. The multivariate case was proved by Stoffer(1985) and are described as

follows:

Assume that $y(\cdot) \in \mathbb{R}^p$ is a stationary zero-mean time series with autocovariances $R_y(v) = E\{y(t)y(t+v)'\}$. Let $M_t = \text{diag}(M_{1t},..,M_{pt})$ be such that M_{it} is the missing data indicator of $y_i(t)$, i=1,..,p. Assume that

$$\lim_{T\to\infty} \hat{R}_M(v) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T-v} M_t 1_p 1_p' M_{t+v} = R_M(v)$$

exists in mean square sense for $v \geq 0$ where $1_p=(1,1,\ldots,1)'$. Consider the time series $x(t) = M_t y(t)$. Then the autocovariance between the ith and jth observation at lag $v \geq 0$ is

$$R_{xij}(v) = E\{x_i(t)x_j(t+v)\} = M_{it}M_{jt+v}R_{yij}(v)$$

where $R_{yij}(v)$ is the (i,j)th element of $R_y(v)$. It can be shown that $x(\cdot)$ is asymptotic stationary in the sense that

$$\lim_{T\to\infty} \hat{R}_x(v) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T-v} x(t)x(t+v)' = R_x(v)$$

exists in mean square for $v \geq 0$. Then the (i,j)th element of $R_x(v)$ can be written as $R_{xij}(v) = R_{Mij}(v)R_{yij}(v)$. Hence if $R_{Mij}(v)\neq 0$ then $\hat{R}_{yij}(v) = \hat{R}_{xij}(v)/\hat{R}_{Mij}(v)$ would be mean square consistent for $R_{yij}(v)$. Moreover, Stoffer(1985) also made the remark that replacing $\hat{R}_{Mij}(v)$ by T or, possibly, some constant proportion of T would not make any difference asymptotically (and therefore still be consistent estimate). One further comment we would like to make is the potential deficient of the estimation strategy. Because of the possibility of unbalanced situations, the resulting autocovariances sequence may not be positive definite. One way to get around the problem may be to perform an eigenvalue analysis of the resulting autocovariances sequence and replace the zero eigenvalues by a small positive quantities. (see Dunsmuir and Robinson(1981))

3.4 Brockwell and Davies Estimation Strategy

Brockwell and Davies(1991) considered application of state space modelling to and determination of likelihood function of irregularly observed time series. This, at first, seems to be similar to what other people was proposed. In fact the underlying idea is significantly different from them.

Consider the following state space model:

$$Y(t) = C(t)X(t) + W(t)$$
$$X(t+1) = A(t)X(t) + V(t) \qquad (3.4.1)$$

where

$$EU_t = E\begin{bmatrix} V_t \\ W_t \end{bmatrix} = 0,\ E(U_t U_t') = \begin{bmatrix} Q_t & S_t \\ S_t' & R_t \end{bmatrix} \text{ and } E\{X(0)U_t'\} = 0.$$

Now, given an irregularly observed realisation of (3.4.1): $\{y(i_1),\ldots,y(i_r)\}$, our aim is to formulate the likelihood function of this observed series. The idea from Brockwell and Davies(1990) is to consider the following alternative state space model:

$$Y^*(t) = C^*(t)X(t) + W^*(t)$$
$$X(t+1) = A(t)X(t) + V(t) \qquad (3.4.2)$$

where

$$C^*(t) = \begin{cases} C(t) & \text{if } t\in\{i_1,..,i_r\} \\ 0 & \text{otherwise} \end{cases},\quad W^*(t) = \begin{cases} W(t) & \text{if } t\in\{i_1,..,i_r\} \\ N(t) & \text{otherwise} \end{cases}$$

and N_t is i.i.d. with

$$N_t \sim N(0,I),\ N_s \perp\!\!\!\perp X(0),\ N_s \perp\!\!\!\perp U_t \text{ for all } s,t.$$

It is clear from model (3.4.2) that $Y^*(t)$ equals to $Y(t)$ whenever $t \in \{i_1,\ldots,i_r\}$ and takes random values independent of $\{Y(t)\}$ at other times. Therefore we can consider a particular realisation of $Y^*(t)$ by defining $y^*(t) = \begin{cases} y(t) & \text{if } t \in \{i_1,\ldots,i_r\} \\ 0 & \text{otherwise} \end{cases}$. Brockwell and Davies(1990) then argue that the likelihood of $\{y(i_1),\ldots,y(i_r)\}$ is proportional to that of $\{y^*(1),\ldots,y^*(T)\}$. Since the likelihood surface for $y^*(t)$, t=1,..,T can be easily generated by Kalman Filter, the likelihood for $\{y(i_1),\ldots,y(i_r)\}$ can also be evaluated as well.

3.5 Shumway and Stoffer's Estimation Strategy

As similar to Brockwell and Davies(1990), Shumway and Stoffer(1982) also considered application of state space modelling to time series. They, first of all, applied the techniques to complete data and then considered extension to time series of missing observations. Unfortunately, the extension is only restricted to special cases.

Consider the following state space model:

$$y(t) = C(t)x(t) + v(t)$$
$$x(t+1) = Ax(t) + w(t) \qquad (3.5.1)$$

where $x(0)\sim N(\mu,\Sigma)$, $w(t)\sim N(0,Q)$, $v(t)\sim N(0,R)$ and $E\{v(t)w(s)'\}=0$. The parameters which are concerned are μ, Σ, A, R and Q. C(t), t=1,..,T are assumed to be known in advance. The justification could be that the transformation from ARMA model to state space model often leads to known C(t), t=1,...,T. (see chapter 2) By considering x(t), t=0,1,...,T are missing, the complete data likelihood can be written as follows:

$$-2\log L = \log|\Sigma| + tr\{\Sigma^{-1}(x(0)-\mu)(x(0)-\mu)'\}$$
$$+ T\log|Q| + \sum_{t=1}^{T} tr\{Q^{-1}(x(t)-Ax(t-1))(x(t)-Ax(t-1))'\}$$
$$+ T\log|R| + \sum_{t=1}^{T} tr\{R^{-1}(y(t)-C(t)x(t))(y(t)-C(t)x(t))'\}.$$

Application of EM algorithm is then considered. Therefore, we look at expectation of -2logL given the data y(1),...,y(T):

$$E\{-2\log L|y(1),..,y(T)\}$$
$$=\log|\Sigma|+tr\{\Sigma^{-1}(P_0^T+(x_0^T-\mu)(x_0^T-\mu)')\}$$
$$+T\log|Q|+tr\{Q^{-1}(H-KA'-AK'+AMA')\}$$
$$+T\log|R|+tr\{R^{-1}\sum_{t=1}^{T}((y(t)-C(t)x_t^T)(y(t)-C(t)x_t^T)'+C(t)P_t^TC(t)')\}$$

where

$$M = \sum_{t=1}^{T}(P_{t-1}^T + x_{t-1}^T x_{t-1}^{T'}),\quad K = \sum_{t=1}^{T}(P_{t,t-1}^T + x_t^T x_{t-1}^{T'}),\quad H = \sum_{t=1}^{T}(P_t^T + x_t^T x_t^{T'}),$$

$x_n^T = E\{x(n)|y(1),..,y(T)\}$, $P_n^T = E\{(x(n)-x_n^T)(x(n)-x_n^T)'\}$ and $P_{n,n-1}^T = E\{(x(n)-x_n^T)(x(n-1)-x_{n-1}^T)'\}$.

From the above expression we can deduce the EM algorithm and it is described as follows:

Given the current estimate of parameters: μ^k, Σ^k, A^k, R^k and Q^k,

E-Step:

Apply Kalman Filter described in last chapter to find x_n^T and P_n^T for n=0,1,2,...,T. The algorithm for determining $P_{n,n-1}^T$, n=T,...,1

can be developed by following similar arguments when we derived the Kalman Smoother in last chapter and are presented as follows:

Define

$$e^T(t) = x(t) - x_t^T,\ e(t) = x(t) - \hat{x}(t) \text{ and } S_t = P_t A' H_t^{-1}$$

for model (3.5.1) where

$\hat{x}(t) = E\{x(t)\,|\,y(1),..,y(t)\}$, $P_t = E\{(x(t)-\hat{x}(t))(x(t)-\hat{x}(t))'\}$ and $H_{t-1} = E\{(x(t)-A\hat{x}(t-1))(x(t)-A\hat{x}(t-1))'\}$.

With respect to model (3.5.1), we note that $H_{t-1}=AP_{t-1}A'+Q$. Since

$$e^T(t) = S_t e^T(t+1)+(I-S_t A)e(t)-S_t w(t) \text{ and}$$

$$e(t) = (I-K_t C(t))(Ae(t-1)+w(t-1))-K_t v(t)$$

where $K_t = H_{t-1}C(t)'(R+C(t)H_{t-1}C(t)')^{-1}$, we can consider

$$\begin{aligned} P^T_{t-1,t-2} &= E\{e^T(t-1)e^T(t-2)'\} \\ &= E\{S_{t-1}e^T(t)e^T(t-1)'S'_{t-2}+S_{t-1}e^T(t)e(t-2)'(I-S_{t-2}A)'- \\ &\quad S_{t-1}e^T(t)w(t-2)'S'_{t-2}+(I-S_{t-1}A)e(t-1)e^T(t-1)'S'_{t-2}+ \\ &\quad (I-S_{t-1}A)e(t-1)e(t-2)'(I-S_{t-2}A)'- \\ &\quad (I-S_{t-1}A)e(t-1)w(t-2)'S'_{t-2}-S_{t-1}w(t-1)e^T_{t-1}S'_{t-2}- \\ &\quad S_{t-1}w(t-1)e(t-2)'(I-S_{t-2}A)'+S_{t-1}w(t-1)w(t-2)'S'_{t-2}\} \\ &= S_{t-1}P^T_{t,t-1}S'_{t-2} + (I-S_{t-1}A)P_{t-1}S'_{t-2}. \end{aligned}$$

Hence for t=T,T-1,...,1, the above equation can be used to determined $P^T_{t,t-1}$. Consequently, M, K, H as well as the expected log-likelihood can also be determined.

M-Step:

This step is to find μ, Σ, A, R and Q such that the expected log-likelihood evaluated in the E-step is maximised. For detail, we refer to Shumway and Stoffer(1982).

When discussing the case when missing observations are present, Shumway and Stoffer(1982) said the Kalman Filtering algorithm is still valid after imputing zeros to the missing observations and putting zeros to the corresponding component of C(t) and the maximisation step is still valid when there are no correlations among missing and observed components of y(t). However, they did not discuss in any further the appropriate steps if this is not

the case.

3.6 Jones Estimation Strategy

Again, State space modelling is employed as a bridge to solve missing data problems in time series. The basic state space model we would like to investigate is the prediction error form which is described as follows:

$$y(t) = Cx(t) + \varepsilon(t)$$
$$x(t+1) = Ax(t) + B\varepsilon(t) \qquad (3.6.1)$$

where

$x(0) \sim N(\mu,\Sigma)$, $\varepsilon(t) \sim N(0,\Omega)$ is i.i.d. and $x(0) \perp \varepsilon(t)\ \forall\ t$.

In essence, the idea is to reduce the dimension of the matrices appearing in Kalman Filter appropriately and with such adjustment, likelihood function is evaluated at the given parameters value. In model (3.6.1), the parameters may include (A,B,C,Ω) with μ and Σ estimated by x_0^T and P_0^T using the Kalman Smoother.

For each $y(t)$, we define $y(t) = (y_1(t)',y_2(t)')'$ where $y_1(t)$ and $y_2(t)$ are the observed part and the missing part of $y(t)$ respectively. Hence if $y(t)$ is completely observed, $y_1(t)=y(t)$ and $y_2(t)$ is null. On the other hand, if $y(t)$ is completely missing, $y_1(t)$ is null and $y_2(t)=y(t)$. Moreover, when $\dim(y(t))=n$ and $\dim(y_1(t))=n_1(t) \le n$, we define the following notations:

1. $\Omega_{1t} = E\{\varepsilon_1(t)\varepsilon(t)'\}$ is $n_1(t)\times n$ component of Ω where $\varepsilon_1(t)$ corresponds to the observed component of $y(t)$, i.e. $y_1(t)$.
2. $\Omega_{11t} = E\{\varepsilon_1(t)\varepsilon_1(t)'\}$ is the $n_1(t)\times n_1(t)$ component of Ω.
3. C_{1t} is the $n_1(t)\times m$ component of C, i.e. $C = \begin{bmatrix} C_{1t} \\ C_{2t} \end{bmatrix}$, where the partition corresponds to the observed part of $y(t)$ and $\dim(x(t)) = m$.

The Kalman Filter to evaluate -2log-likelihood:

Let $\mathcal{O}$ be the index set of partially observed or completely observed $y(t)$, $t \in \{1,\ldots,T\}$. Define

$$\hat{x}(t) = E\{x(t) \mid y_1(k),\ k \in \mathcal{O} \cap \{1,..,t\}\}$$
$$\bar{x}(t) = E\{x(t) \mid y_1(k),\ k \in \mathcal{O} \cap \{1,..,t-1\}\}$$
$$v_1(t) = y_1(t) - C_{1t}\bar{x}(t) \text{ (when } y_1(t) \text{ is not null).}$$

Then the steps for Kalman Filter can be described as follows:

Starting conditions: $\bar{x}(1) = A\mu$, $H_0 = A\Sigma A'+B\Omega B'$ and $\lambda=0$

For n = 1,2,...,T do the following:

Case 1: When $y_1(n)$ is not null,

$$Q_n = E\{v_1(t)v_1(t)'\} = \Omega_{11n} + C_{1n}H_{n-1}C'_{1n}$$

$$K_n = H_{n-1}C'_{1n}Q_n^{-1} \text{ since } E\{(\hat{x}(n)-\bar{x}(n))v_1(n)'\}=K_nQ_n$$

$$P_n = (I - K_nC_{1n})H_{n-1}$$

$$\hat{x}(n) = \bar{x}(n) + K_n(y_1(n)-C_{1n}\bar{x}(n))$$

$$\bar{x}(n+1) = A\hat{x}(n) + BE\{\varepsilon(n)\,|\,y_1(k),\ k \in \mathcal{O} \cap \{1,..,n\}\}$$

$$= A\hat{x}(n) + BE\{\varepsilon(n)\,|\,v_1(n)\}$$

$$= A\hat{x}(n) + B\Omega'_{1n}Q_n^{-1}(y_1(t)-C_{1n}\bar{x}(n))$$

$$H_n = E\{(x(n+1)-\bar{x}(n+1))(x(n+1)-\bar{x}(n+1))'\}$$

$$= E\{(Ae(n)+B\varepsilon(n)-B\Omega'_{1n}Q_n^{-1}v_1(n))(\cdot)'\}$$

$$= AP_nA'+B\Omega B'-AK_n\Omega_{1n}B'-B\Omega'_{1n}K'_nA'-B\Omega'_1Q_n^{-1}\Omega_1B'$$

$$\lambda = \lambda + \log|Q_n| + v_1(n)'Q_n^{-1}v_1(n)$$

where $e(n) = x(n)-\hat{x}(n) = (I-K_nC_{1n})(x(n)-\bar{x}(n))-K_n\varepsilon_1(n)$.

Case 2: When $y_1(n)$ is null,

$$\hat{x}(n) = \bar{x}(n)$$

$$P_n = H_{n-1}$$

$$\bar{x}(n+1) = A\hat{x}(n)$$

$$H_n = AP_nA'+B\Omega B'$$

end the do loop.

With such adjustment, the resulting λ would give us the value of -2log-likelihood at the given parameters. Therefore maximum likelihood estimation are possible to perform via a descent algorithm. However, Jones(1983) commented that it could take a long time to reach the maximum likelihood estimator and the choice of initial estimate is essential. In fact, the likelihood function is not well-behaved so that local maxima can occur because of ripples caused by rounding off error. Although Jones' approach changes the dimension of matrices as time changes, we can see that the sequence $\{v_1(t),\ t \in \mathcal{O} \cap \{1,...,T\}\}$ is indeed an orthogonal

sequence of $\{y_1(t),\ t \in \mathcal{O} \cap \{1,\ldots,T\}\}$. In addition, the exposition of logic is easier to understand than the strategy given by Shumway and Stoffer(1982) or Brockwell and Davies(1991) who simply put zeros to where missing observations occur. Based on this view, we have chosen Jones approach as our maximum likelihood estimation strategy.

3.7 Kohn and Ansley's Strategy for differencing

With complete data, Box and Jenkins approach would use differencing to remove non-stationarity and then apply ARMA modelling to the residual series. But this does not work when missing observations are present. One suggestion to get around differencing was presented in Kohn and Ansley(1985). We believe that this problem is very interesting as well as important to practical time series modelling. Therefore, it could be worthwhile to spend a section to discuss their idea. However, although Kohn and Ansley(1985) also presented general strategy to identify and estimate parameters of state space model for nonstationary time series, we are not going to discuss them. For detail, please see Kohn and Ansley(1985).

We consider the ARIMA(p,d,q) model, in which

(3.7.1) $$\phi(z^{-1})(1-z^{-1})^d y(t) = \theta(z^{-1})\varepsilon(t)$$

where

$$\phi(z^{-1}) = 1 - \sum_{j=1}^{p} \phi_j z^{-j} \text{ and } \theta(z^{-1}) = 1 - \sum_{j=1}^{q} \theta_j z^{-j}.$$

The following assumptions were made:

1. $\phi(\cdot)$ has all its roots outside the unit circle.
2. $\varepsilon(t)$ is a sequence of independent $N(0,\sigma^2)$ random variables.
3. We observed $y(t_1),\ldots,y(t_N)$ with $1=t_1 < t_2 <\cdots< t_N=T$.

Define $u(t) = (1-z^{-1})^d y(t)$. If we can write $(1-z^{-1})^d = \sum_{j=1}^{d} \delta_j z^{-j}$, then $y(t) = \sum_{j=1}^{d} \delta_j y(t-j) + u(t)$. Let $\eta = \{y(1-d),\ldots,y(0)\}'$, we understand that $y(t) = a(t)'\eta+\omega(t)$, $t \geq 1$, where the $D\times 1$ vector $a(t)$ depends only on $\delta_1,\ldots,\delta_d$ and not on ϕ, θ or σ^2 and $\omega(t)$ is linear combination of $u(1),\ldots,u(t)$ and does not depend on η. By writing our observed data as $y = (y(t_1),\ldots,y(t_N))'$ and ω =

$(\omega(t_1),\ldots,\omega(t_N))'$, $y = A\eta + \omega$ where A is a N×d matrix with ith row defined by $a(t_i)'$. Since $u(\cdot)$ is Gaussian mean zero process defined by (3.7.1), ω is also Gaussian mean zero with covariance matrix depending only on ϕ, θ and σ^2. However, because $y(\cdot)$ is non-stationary, the distribution of η is not defined. The idea fron Kohn and Ansley(1985) is clear in that a suitable transformation of y(t) would eliminate dependence on η. Let the rank of A be d' ≤ d and let J be N×N matrix with (1) det(J)=1 and (2) JA consists exactly of d' non-zero rows. Then J can be decomposed into J_1 and J_2 where J_1A is the d' non-zero rows of JA and $J_2A = 0$. Hence we have the following:

$$\omega_1 = J_1A\eta + J_1\omega \text{ and } \omega_2 = J_2\omega.$$

Now, ω_2 depends only on $u(\cdot)$ and its density is well-defined. Kohn and Ansley(1985) therefore define the likelihood as the density of ω_2. By identifying such a transformation, we clearly have reduced the series $y(\cdot)$ into a stationary one.

When there are no missing observations, we can define J_1 and J_2 as the d×T and (T-d)×T matrices respectively so that $J_1=(I_d,0)$ and

$$J_2 = \begin{bmatrix} -\delta_d & \cdots & -\delta_1 & 1 & & \\ 0 & -\delta_d & \cdots & -\delta_1 & 1 & \\ & & \cdots & & & \\ & & -\delta_d & \cdots & -\delta_1 & 1 \end{bmatrix}$$

Then $J_1y = (y(1),\ldots,y(d))'$ and $J_2y = (u(d+1),\ldots,u(T))'$. This immediate shows that the idea is in fact equivalent to differencing when there are no missing observations. However, we also note that the transformation depends very much on the assumed ARIMA model and may be difficult to construct in practice.

3.8 Some Comments on Asymptotic distribution of MLE when missing observations are present

In this section, we consider the asymptotic distribution of the maximum likelihood estimator of a given Markov model when only irregularly observed series $\{x(t_i),\ i \in \{1,\ldots,T\}\}$ are available. The reason for studying Markov process is that the estimated state vector given current and past information is in fact a Markov process:

$$\hat{x}(t) = A\hat{x}(t-1) + B\Omega Q_t^{-1}v(t)$$

where v(t) may be viewed as innovation due to current time with the property $E\{v(t)|y(1),..,y(t-1)\}=0$. This fact may be important and could be extended in the future.

Statement of problem: Given an observed Markov process $\{x(t_i)|i\in\{1,...,N\}\}$ with transition density $f(\cdot,\cdot;\theta)$ where θ is some unknown parameter. Our aim is to show the existence of maximum likelihood estimator and its asymptotic distribution.

Define the transition density between $x(t_k)$ and $x(t_{k+1})$ by $f_k(x(t_k),x(t_{k+1});\theta)$. Then we have the following relation:

$$f_k(x(t_k),x(t_{k+1});\theta)$$
$$=\int_{\chi_k} f(x(t_k),x(t_k+1);\theta)\cdots f(x(t_{k+1}-1),x(t_{k+1});\theta)dx(t_k+1)\cdots dx(t_{k+1}-1)$$

where χ_k is the product $\chi\times\cdots\times\chi$ ($t_{k+1}-t_k-1$ times) and χ is the state space.

Apart from the initial distribution (which does not affect the asymptotic property), the log-likelihood function would be given as follows:

$$L_N(\theta) = \sum_{k=1}^{N} \log f_k(x(t_k),x(t_{k+1});\theta)$$

Assuming θ is r-dimensional, the MLE $\hat{\theta}$ satisfies $\frac{\partial}{\partial\theta_u}L_N(\theta)\Big|_{\theta=\hat{\theta}}$ for u=1,...,r. For convenience, we write $g_k(x,y;\theta) = \log f_k(x,y;\theta)$. In the course of proving asymptotic properties of Markov process, Billingsley(1961) assumes the following Central Limit Theorem for martingale:

Theorem (3.8.1)

Let u_1, u_2, ... be random variables with moments of order $2+\delta$, $\delta > 0$, and let $\mathcal{F}_0$, $\mathcal{F}_1$, ... be a nondecreasing sequence of Borel fields such that $E\{u_n|\mathcal{F}_{n-1}\} = 0$ with probability one, n=1,2,.... Suppose that

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} E\{u_k^2|\mathcal{F}_{k-1}\} = \beta^2$$

and

$$\lim_{n\to\infty} n^{-1-\delta/2}\sum_{k=1}^{n} E\{|u_k|^{2+\delta}|\mathcal{F}_{k-1}\} = 0$$

with probability one, where β^2 is a nonnegative constant. Then

$$(3.8.2) \qquad n^{-1/2} \sum_{k=1}^{n} u_k \xrightarrow{\mathcal{L}} N(0,\beta^2).$$

where $\mathcal{L}$ denotes convergent in distribution.

In the sequel, we define $\mathcal{F}_{n-1}$ equals to the σ-field generated by $x(t_1),\ldots,x(t_n)$ and regard the following conditions as regularity conditions:

Condition 1.1 For any k and ξ, the set of η for which $f_k(\xi,\eta;\theta)>0$ does not depend on θ. For any k, ξ and η, $f_{uk}(\xi,\eta;\theta)$, $f_{uvk}(\xi,\eta;\theta)$ and $f_{uvwk}(\xi,\eta;\theta)$ exist and are continuous throughout Θ where $\frac{\partial}{\partial\theta_u} f_k(\xi,\eta;\theta) = f_{uk}(\xi,\eta;\theta)$, $\frac{\partial^2}{\partial\theta_u\partial\theta_v} f_k(\xi,\eta;\theta) = f_{uvk}(\xi,\eta;\theta)$ and etc. and Θ is the space for θ. For any $\theta \in \Theta$, there exists a neighbourhood N of θ such that for any k, u, v, w, ξ,

$$\int_{\chi} \sup_{\theta'\in N} |f_{uk}(\xi,\eta;\theta')| \lambda(d\eta) < \infty$$

$$\int_{\chi} \sup_{\theta'\in N} |f_{uvk}(\xi,\eta;\theta')| \lambda(d\eta) < \infty$$

$$E_\theta\{ \sup_{\theta'\in N} |g_{uvwk}(x(t_k),x(t_{k+1});\theta')|\} < \infty.$$

where λ is some suitable measure.

Finally, for u = 1,2,...,r and for all k,

$$E_\theta\{|g_{uk}(x(t_k),x(t_{k+1});\theta)|^2\} < \infty,$$

and if $\sigma_{uv}(\theta)$ is defined by

$$(3.8.3)\ \sigma_{uv}(\theta)$$

$$= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E\{g_{uk}(x(t_k),x(t_{k+1});\theta) g_{vk}(x(t_k),x(t_{k+1});\theta)\}$$

then the r×r matrix $\sigma(\theta) = (\sigma_{uv}(\theta))$ is non-singular.

Condition 1.2

(i) For each $\theta \in \Theta$, the stationary distribution, which by assumption exists and is unique, defined as

$$p_\theta(\xi,A) = \int_A f(\xi,\eta;\theta)\lambda(d\eta) \text{ and } p_\theta(A) = \int_{\chi} p_\theta(\xi,A) p_\theta(d\xi),$$

has the property that for each $\xi \in \chi$, $p_\theta(\xi,\cdot)$ is absolutely continuous with respect to $p_\theta(\cdot)$.

(ii) There is some $\delta > 0$ such that for u = 1,...,r and for all k,

$$E_\theta\{|g_{uk}(x(t_k),x(t_{k+1});\theta)|^{2+\delta}\} < \infty.$$

Theorem 3.8.4

Regularity conditions imply that for any $\theta \in \Theta$ and any initial distribution, the random vector $y(n) = (y_1(n),..,y_r(n))'$ where

$$y_u(n) = n^{-1/2} \sum_{k=1}^{n} g_{uk}(x(t_k),x(t_{k+1});\theta)$$

converges in law to $N(0,\sigma(\theta))$.

Proof: It suffices to show that for any set $z_1,...,z_r$ of real numbers,

$$n^{-1/2} \sum_{k=1}^{n} u_k \xrightarrow{\mathcal{L}} N(0,\beta^2)$$

where

$$u_k = \sum_{u=1}^{r} z_u g_{uk}(x(t_k),x(t_{k+1});\theta) \text{ and } \beta^2 = \sum_{u=1}^{r}\sum_{v=1}^{r} z_u z_v \sigma_{uv}(\theta).$$

Since $\int_\chi f_k(\xi,\eta;\theta)\lambda(d\eta) = 1$, $\int_\chi f_{uk}(\xi,\eta;\theta)\lambda(d\eta) = 0$. This implies that $E\{g_{uk}(x(t_k),x(t_{k+1});\theta)|x(t_k)\} = 0$. Then

$$E\{u_k|\mathcal{F}_{k-1}\} = \sum_{u=1}^{r} z_u E\{g_{uk}(x(t_k),x(t_{k+1});\theta)\} = 0.$$

Theorem (3.8.1) implies that

$$\beta^2 = \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E\{u_k^2|\mathcal{F}_{k-1}\}$$

$$= \sum_{u=1}^{r}\sum_{v=1}^{r} z_u z_v \lim_{vn\to\infty} n^{-1} \sum_{k=1}^{n} E\{g_{uk}(x(t_k),x(t_{k+1});\theta)g_{vk}(x(t_k),x(t_{k+1});\theta)\}$$

$$= \sum_{u=1}^{r}\sum_{v=1}^{r} z_u z_v \sigma_{uv}(\theta).$$

where with probability one,

$$\sigma_{uv}(\theta) \sim \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E\{g_{uk}(x(t_k),x(t_{k+1});\theta)g_{vk}(x(t_k),x(t_{k+1});\theta)|\mathcal{F}_{k-1}\}$$

was from Hall and Hedye(1980).

This completes the proof.

We assume that there is some $\theta^0 \in \Theta$ which represents the true value of θ. The following results can be established:

Result 3.8.5

Regularity conditions and conditions found in Theorem 2.23 of Hall and Hedye(1980) imply that there exists a sequence $\{\hat{\theta}_n\}$ of random vectors in Θ, each one being a function $\hat{\theta}_n = \hat{\theta}(x(t_1),..,x(t_{n+1}))$ of the observation, such that $\hat{\theta}_n$ converges in probability to θ^0, the true value, and such that $\hat{\theta}_n$ is a solution of the system: $\frac{\partial}{\partial\theta_u} L_n(\theta) = 0$, u=1,...,r with probability going to one as $n \to \infty$. Thus there is a consistent maximum likelihood estimator of θ^0. Moreover, $\hat{\theta}_n$ is a local maximum of $L_n(\theta)$ with probability going to one. Finally if $\bar{\theta}_n$ is also a local maximum of $L_n(\theta)$ then the probability that $\hat{\theta}_n = \bar{\theta}_n$ goes to one as $n \to \infty$.

Proof: From previous discussion, we understand that for all k,

$$E_\theta\{g_{uk}(x(t_k),x(t_{k+1});\theta)\} = 0$$

and

$$E_\theta\{g_{uvk}(x(t_k),x(t_{k+1});\theta)\,|\,x(t_k)\}$$
$$= - E_\theta\{g_{uk}(x(t_k),x(t_{k+1});\theta)g_{vk}(x(t_k),x(t_{k+1});\theta)\,|\,x(t_k)\}.$$

This demonstrates that

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E_\theta\{g_{uvk}(x(t_k),x(t_{k+1});\theta)\,|\,x(t_k)\} = -\sigma_{uv}(\theta).$$

By weak law of large number for martingale (see Hall and Hedyes(1980)), we have

$$\operatorname{plim}_{n\to\infty} n^{-1} \sum_{k=1}^{n} g_{uk}(x(t_k),x(t_{k+1});\theta^0) = 0.$$

Hall and Hedyes(1980)'s conditions imply that

$$\text{(3.8.6)} \qquad \operatorname{plim}_{n\to\infty} n^{-1} \sum_{k=1}^{n} g_{uvk}(x(t_k),x(t_{k+1});\theta^0) = -\sigma_{uv}(\theta^0).$$

Then by using mean value theorem and similar arguments as shown in Billingsley(1961), we can establish the existence of consistent maximum likelihood estimator and the negative-definiteness of $n^{-1}\frac{\partial^2}{\partial\theta_u\partial\theta_v} L_n(\theta)$ for every θ in some neighbourhood of θ^0.

Given the existence of MLE, we now can talk about its asymptotic properties. In fact the following well-known result:

Let $l(n) = (l_1(n),\ldots,l_r(n))'$ where $l_u(n) = n^{-1/2}(\hat{\theta}_n(u)-\theta^0_u)$. Then

$$l(n) \xrightarrow{\mathcal{L}} N(0,\sigma^{-1}(\theta^0)).$$

The arguments are also similar to Billingsley(1961). However, we would like to put one remark. (3.8.6), in essence, implies that the Fisher information matrix is equal to $\sigma(\theta^0)$ which is also asymptotically equivalent to $\sigma(\hat{\theta}_n)$ as $n \to \infty$.

From the above discussion on Markov process with irregularly observations, we can expect state space model to behave similarly. However, the derivation for state space model could be more difficult.

4. Data Description, Ignorability and Trend Analysis

Having discussed in details the theoretical background of state space modelling and adjustment idea to allow missing observations, it is time to introduce real data into our journey. The data is a hospital record of a patient who has Leukaemia. They are obtained directly from measurement records of that hospital and are taken by qualified specialists when the patient sees the doctor. The patient or The group of patients with Leukaemia is taken measurements regularly (about once in every 7 days) at the beginning of the treatment until the condition improves. Subsequent measurement taking is still regular but less frequently (about once in every 14 days). However, interruption occurs when the patient did not come in a pre-scheduled week but the week after that. Also, when the doctor thinks that the patient is healthy enough, he/she may allow for even longer period between meetings. That's why we have the missing observations. Partial missing occurs when the measurement people feel that interim data are not required at that moment. The proportion of missing observations from all major variables are about 33.5%. Although there are many patients with Leukaemia, we only use data from a particular patient to illustrate the modeling techniques we have developed in this thesis. Our data set consists of the three main variables and six auxiliary variables which are described in the followings:

Main Variables:

1. Total White Blood Cell Count (TWBC)
2. Neutrophyll Count (NC)
   - a main group of white blood cells constituting of about 30%-70% of TWBC
3. Lymphocyte Count (LC)
   - another main group of white blood cells also constituting of about 30%-70% of TWBC.

Remark:

1. A more accurate observation is that NC + LC $\approx$ 95%$\times$TWBC and an obvious constraint among the three main variables are NC + LC $\leq$ TWBC.
2. Since they are all count variables, all three variables assume values greater than or equal to zero.
3. Because they are all white blood cells, we believe that they should have similar response when they are subjected to similar external influence.

Auxiliary Variable:

1. Infection Observed
   - an indicator showing whether the patient was infected by other disease or not
   - 0 = no infection; 1 = infected.
2. Stage of Treatment
   - a categorical variable showing the stage of treatment.
   - Stages from 1 to 4.
3. Cycle within Stage
   - a categorical variable from 1 to 8 (in this case) showing the current repetition of every stage.
4. Week within Cycle
   - the number of weeks within a given cycle.

5. Treatment Break
   - an indicator showing whether treatment was given according to schedule or not.
   - 0 = given to schedule; 1 = not given (or possibly ceased).
6. Time (in week) elapsed since beginning of treatment.

The objective of this exercise is to modelling the white blood cell counts. However, because of the presence of missing observations, it is necessary to check the ignorability of the process which caused missing observations according to Rubin(1976). If the process is ignorable (possibly MAR) then we don't need to pay attention to modelling the process. (Remark: it does not mean that missing observations can be thrown away.)

4.1 Ignorability Analysis

By "ignorability", we mean that y_t is independent of R_t with or without conditioning on the auxiliary information where y_t is the white blood cell counts at time t and $R_t = \begin{cases} 1 & \text{if } y_t \text{ is missing} \\ 0 & \text{otherwise} \end{cases}$. The way we use to justify ignorability is based on the fact that a random variable is always independent to a constant. If we can explain R_t satisfactorily, at least, conditional on auxiliary information then R_t given the auxiliary information can be regarded effectively as a constant and hence ignorable. However, if ignorability is proved in this way, future analysis will also be conditional on auxiliary information. Furthermore, since we only got one realisation of $\{R_t\}$, we need the assumption that R_t is generated by a stationary process. With this assumption, $E(R_t)$ or the probability of missing at time t is a constant invariant with time. Therefore it enables the whole series to be used.

Given the binary nature of R_t, one of the standard ways is to perform a logit analysis of R_t against the auxiliary information. Our logistic model can be described as follows:

$$(4.1.1) \qquad \pi = \Pr(R_t \mid x) = \exp(\alpha+\beta'x)/\{1+\exp(\alpha+\beta'x)\}$$

where x is the auxiliary information and α and β are parameters of the model. Therefore, the log-likelihood function may be written in the form:

$$(4.1.2) \qquad l(\pi;\ R_t, t=1,..,T) = \sum_{t=1}^{T} \{R_t \log(\frac{\pi}{1-\pi}) + \log(1-\pi)\}.$$

Substituting (4.1.1) into (4.1.2) gives

$$(4.1.3)\quad l(\beta;\ R_t, t=1,..,T) = \sum_{t=1}^{T}\sum_{j=1}^{p} R_t x_{tj}\beta_j - \sum_{t=1}^{T} \log\{1+\exp(\sum_{j=1}^{p} x_{tj}\beta_j)\}$$

Now, maximum likelihood estimation can be done via differentiation (4.1.3) with respect to β_j, j=1,..,p. We can show that

$$(4.1.4)\qquad \frac{\partial l}{\partial \beta_j} = \sum_{t=1}^{T} x_{tj}(R_t - \pi),\ j=1,2,\ldots,p.$$

The objective is then to find the zeros of the system (4.1.4). In fact we can differentiate the log-likelihood further to obtain the Fisher information matrix which is easily to be equal to X'WX where $X = (x_{tj})$ and $W = \mathrm{diag}\{\pi(1-\pi)\}$. Then Newton-Raphson algorithm can be used to obtain the maximum likelihood estimator. Given the simple nature of the likelihood function, convergent would clearly not be a problem. However, we still need to make model comparisons in order to eliminate unnecessary variables. This may be done by comparing the difference in deviances of the two models under consideration. Deviance, in this context, is defined as $-2l(\hat{\pi};\ R_t, t=1,\ldots,T)$ where $\hat{\pi}$ is the given parameter value and reduces with increasing number of auxiliary variables. Moreover, the asymptotic distribution of the difference in deviances (which is chi-square) is very rough and can be relied on only when the p-value is extreme enough.

In table 4.1.5, we present our resullts of performing logit analysis on the missing data indicators on NC, LC and TWBC-NC-LC respectively. Interestingly, if we can regard the deviance of the null model, i.e. no auxiliary variable, as full deviance, approximately 20% of the full deviance can be explained by our auxiliary information. Clearly, the auxiliary information has explanatory power over the process that caused missing observations. The question is "is that mean the process which caused missing data ignorable?". A paper from Cox(1992) said that low explanatory power could be a fact for binary response and apparently low explanatory power doesn't mean no existence of essential relationship. Therefore a generally reduction in 20% of residual deviance over the "full" deviance is in fact a quite good result. We thus believe that the process which caused missing

observations may be regarded as ignorable but only conditional on auxiliary information, i.e. MAR in Rubin(1976) sense.

Table 4.1.5(a)

	Coef	se(Coef)	Deviance	df	change
Intercept	-6.979000	0.5042	137.0	185	
$I(Stage_t=4)$	-3.169000	0.9066			
Time	0.0439200	0.02015			
$Time^2$	-0.00009664	0.00007786	102.2	182	34.8

* logit analysis of missing data indicator of Neutrophyll count.

Table 4.1.5(b)

	Coef	se(Coef)	Deviance	df	change
Intercept	-6.26827	0.268782	135.2	185	
LC_{t-1}	0.270380	0.256610			
$I(Stage_t=4)$	-1.59972	0.501463			
Time	0.012710	0.003664	111.6	182	23.6

* logit analysis of missing data indicator of Lymphocyte count.

Table 4.1.5(c)

	Coef	se(Coef)	Deviance	df	change
Intercept	-7.2304621	0.576700	136.5	185	
$I(Stage_t=4)$	-3.3671416	0.974200			
$I(Treatment_{t-1}=1)$	-0.5654033	0.516600			
Time	0.0507775	0.022910			
$Time^2$	-0.0001055	0.00008509	103.9	181	32.5

* logit analysis of missing data indicator of residual white blood cell counts which is equal to TWBC - NC - LC.

Note that $I(\cdot)$ is the indicator function.

4.2 Trend Analysis

Generally speaking, trend analysis are concerned with measuring the "long term" tendency of white blood cell counts data. With trend analysis, we can produce better interpolation of the endogeneous variable(s) as well as reduce the data to a stationary (approximately) mean zero process for further analysis. In addition, conditional on auxiliary information can guarantee the ignorability of the process that caused missing observations.

Similar to most statistical data analysis, the useful first step is to look at time plots of the data. In fig. A.1, we present four graphs each representing TWBC, NC, LC and the residual count

(OWBC = TWBC-NC-LC) respectively and number of observations have been noted as shown below:

1. All graphs demonstrate change in level after treatment was ceased.
2. LC shows a strong quadratic trend in time.
3. OWBC also shows a significant quadratic trend but not as strong as that of LC.

The overall objective of this trend analysis is simple. Basically, we would like to define a proper exogeneous variable Z_t such that $y_t = \beta Z_t + \tilde{y}_t$ where β is the regression parameter, y_t = (NC, LC, OWBC)', $\tilde{y}_t$ is the (approximately mean zero and stationary) residual series and Z_t is the vector of explanatory variables which contains an entry of 1 for the intercept term. However, the exact logical steps taken needs a little bit more explanation. Therefore before proceeding to the actual steps of analysis, we would like to present the following "algorithmic-like" statements:

```
repeat {
    Select explanatory vector Z_t
    repeat {
        determine β such that
            Σ_{t=1}^{T} ρ(y_kt - Σ_{j=1}^{p} β_kj Z_jt) = min! ∀ k = 1,...,dim(y_t)
        if (all variables are significant) then
            ALLOK = true
        else
            ALLOK = false
            Model Reduction
        end if
    } until (ALLOK is true)
    if (residual plots show nonconstant mean or variance) then
        RPOK = false
        Select appropriate transformation for y_t
    else
        RPOK = true
    end if
} until (RPOK is true)
```

use final β estimate as initial estimate to maximum likelihood estimation.

Stop.

where ρ is a convex, bounded and differentiable of sufficiently high order function and $\rho'(r_i)$ represents the Winsorized version of residual r_i, i.e. bounded residual. (see Huber(1981))

To begin the analysis, we define the following variables for consideration of including into our final explanatory vector Z_t:

1. $\text{Stage(i)}_t = I(\text{Stage}_t = i)$, i=2,3,4
2. $\text{Cycle(i)}_t = I(\text{Cycle}_t = i)$, i=2,3,...,8
3. $I_t = I(\text{Infection}_t = 1)$
4. $T_t = I(\text{Treatment}_t = 1)$
5. $EC_t = I(t=62)$

Remark:

In case 62, the level of WBC of the patient dropped to exceptionally low level and the patient was also infected. Although we don't know whether the low level of WBC casued infection or vice versa, we believe that the patient was infected by a serious virus. In addition, treatment was also forced to terminate as well. Therefore, we treat this case as a special case.

Hence our full vector of explanatory variables Z_t would be

$Z_t = (1,\text{Stage(i)}_t,i=2,3,4,\text{Cycle(i)}_t,i=2,..,8,I_t,T_t,T_{t-1},EC_t,t,t^2)'$

where 1 represents the intercept term.

However, the way we used to estimate β such that $y_t = \beta Z_t + \tilde{y}_t$ where $\tilde{y}_t$ is some zero-mean stationary time series is not ordinary least square because of the presence of missing observations. Instead we estimate β by minimising a robustified function of residuals due to Huber(1981). Such a robustified function is less sensitive to contamination of data and, therefore, may be able to provide better estimate of β given the lack of prior knowledge of missing observations.

The robust regression was performed using "rreg" with method=wt.huber from Splus. Afterwards, the method we used to compare a p-parameters model and a q-parameters model where p > q was based on the method given in Huber(1981) p.197. As a summary,

let $\hat{y}_p$ and $\hat{y}_q$ be vectors of fitted value using the p-model and the q-model respectively. Then model comparison was done via the following statistic:

$$\frac{\frac{1}{p-q}\,\|\hat{y}_p - \hat{y}_q\|^2}{\frac{1}{n-p}\,\frac{K^2\sum \psi(r_i/\hat{\sigma})^2\hat{\sigma}}{\left[\frac{1}{n}\sum \psi'(r_i/\hat{\sigma})\right]^2}} \sim \chi^2_{p-q}$$

where

1. r_i is the residual of the p-model.
2. n = number of observations.
3. $\psi(x) = \min\{c, \max(-c, x)\}$ for some $c > 0$.
4. $K = 1 + \frac{p}{n}(\frac{1-m}{m})$, m=relative frequency of r_i satisfying $|r_i| < c$.
5. $\hat{\sigma}$ = median absolute deviation of $|r_i|$.
6. $\|x\|^2 = x'x$.

Moreover, following models comparison and parameter estimation, the covariance of the estimated parameter β can be determined as

$$\hat{cov}(\hat{\beta}) = K^2 \frac{\frac{1}{n-p}\sum \psi(r_i)^2}{\left[(1/n)\sum \psi'(r_i)\right]^2}(Z'Z)^{-1}$$

where Z is the matrix of explanatory variables.

After computing the β and model selection, we found that the residual plot of Lymphocyte Count exhibits pattern of nonconstant variance (as can be seen in fig. A.2). Therefore, we tried square root transformation on y_t hoping that residual variances can be stabilised, i.e. $y_t = (\sqrt{NC}, \sqrt{LC}, \sqrt{OWBC})'$. Using this new response variables, the resulting model showed a much better residual plots despite the fact that heavy occurrence of missing data might distort or hidden the intrinsic dynamic of the original data. (see fig. A.4)

Thus using the square root transformation, our resulting models given as

	NC	LC	OWBC
	Coef(se)	Coef(se)	Coef(se)
Intercept	1.303051(0.073)	0.9547731(0.021)	0.4847417(0.01)
Time	-0.018690(0.003)	-0.0057328(0.0003)	-0.0027111(0.0005)

	NC	LC	OWBC
	Coef(se)	Coef(se)	Coef(se)
$Time^2$	0.000045(0.00001)	0.0000378(0.000002)	0.000019(0.000002)
$Stage(4)_t$	0.7624096(0.013)	0.000000000	0.2393647(0.018)
$Cycle(2)_t$	0.18110(0.0921)	-0.08327043(0.018)	-0.1115114(0.014)
$Cycle(3)_t$	1.294243(0.164)	-0.3856895(0.0263)	0.000000000
$Cycle(4)_t$	0.000000000	-0.07808983(0.019)	-0.0823534(0.014)
$Cycle(5)_t$	0.2546841(0.089)	0.1639946(0.017)	0.1155861(0.013)
$Cycle(6)_t$	0.000000000	0.0000000000	-0.0503651(0.013)
$Cycle(7)_t$	0.3976013(0.095)	0.0000000000	0.0000000000
$Cycle(8)_t$	1.097918(0.102)	0.0000000000	-0.2435395(0.014)
I_t	0.000000000	-0.07929592(0.01)	0.0000000000
T_t	0.000000000	0.08974408(0.013)	0.0000000000
T_{t-1}	0.200656(0.067)	0.1745385(0.013)	0.1321626(0.009)
EC_t	-2.375385(0.263)	0.0000000000	-0.405563(0.031)

* 0.000000000 means "the corresponding variable is not selected".

The $\hat{\beta}$ presented above was served as initial estimate to optimisation algorithm to compute the maximum likelihood estimator.

5. Balanced Realisation and Model Reduction

After obtaining an initial estimate of β, we are going to consider the following family of models:

$$\tilde{y}_t = y_t - \beta Z_t = Cx(t) + \varepsilon(t)$$
$$x(t+1) = Ax(t) + B\varepsilon(t)$$

where $\varepsilon(t) \sim N(0,\Omega)$ is i.i.d. and $x(0) \sim N(\mu,\Sigma)$.

However, we shall impose the following assumptions (5.a):

1. $\tilde{y}_t$ has a causal rational transfer function so that state space modelling can be applied.
2. the eigenvalues of A has absolute value less than 1, i.e. the state space model is stable.
3. the mapping φ: $(A,B,C) \to z^{-1}C(I-Az^{-1})^{-1}B$ is injective for the purpose of identifiability.

Our problems in this section are:

1. to determine suitable dimension of state vector x(t);
2. to obtain initial parameters estimate of the selected model.

The steps we use to achieve the objective can be simply put into the following statements:

1. apply AR approximations to $\tilde{y}_t$ to obtain successive approximation of Hankel matrix $\mathcal{H}$.
2. apply Singular Value Decomposition (SVD) to the estimated Hankel matrices to obtain an internally balanced realisation of $\tilde{y}_t$ and then to estimate suitable dimension of x(t).

We are going to discuss each steps seperately:

5.1 AR approximations

for each order p = 1,2,...,loglogT where T is the number of observations, we estimate matrices $A_1,\ldots,A_p$ such that

$$\tilde{y}_t = A_1\tilde{y}_{t-1} + A_2\tilde{y}_{t-2} + \cdots + A_p\tilde{y}_{t-p} + \varepsilon(t). \tag{5.1.1}$$

Equation (5.1.1) is called p-th AR approximation to $\tilde{y}_t$. In the literature, the best algorithm which does the job is undoubtly the multivariate generalisation of the Burg's algorithm (see Jones(1978)). This algorithm makes use of all data via consideration of maximum entropy and was proved to guarantee stationary solution as well. Unfortunately, the presence of missing observations makes this algorithm, if not possible, very difficult to employ. We, therefore, work with a more traditional

Levinson Whittle algorithm which requires only the sequence of autocovariances as input. This algorithm produces coefficient matrices estimate which satisfy the Yule-Walker equations. However, one of the disadvantage of Levinson-Whittle algorithm is the presence of small sample bias toward zero when sample size is small. Nevertheless, Levinson-Whittle algorithm still serves as a reasonable method to provide AR coefficients estimate.

Since inputs to Levinson-Whittle algorithm is the sequence of autocovariances of $\tilde{y}_t$, we define

$$a_k(i) = \begin{cases} 1 & \text{if } \tilde{y}_k(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}, \text{ for } i=1,\ldots,h$$

where $h = \dim(\tilde{y}_t)$ and $k = 1,\ldots,T$.

As we have seen in chapter 3, Parzen(1963) and Stoffer(1986) provided a consistent estimators of the sequence of autocovariances and the estimate of autocovariance at lag k is defined by the following statistic:

$$(5.1.2)\quad \hat{R}_k(i,j) = \frac{\sum_{t=1}^{T-k} a_{t+k}(i)a_t(j)(\tilde{y}_{t+k}(i)-\bar{\tilde{y}}(i))(\tilde{y}_t(j)-\bar{\tilde{y}}(j))}{\sum_{t=1}^{T-k} a_{t+k}(i)a_t(j)}$$

for $i,j = 1,\ldots,h$ and $\bar{\tilde{y}}(k) = \dfrac{\sum_{t=1}^{T} a_t(k)\tilde{y}_t(k)}{\sum_{t=1}^{T} a_t(k)}$. (5.1.2) is mean-square consistent provided that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T-k} a_{t+k}(i)a_t(j) = \Theta_k(i,j) \text{ for } i,j = 1,\ldots,h$$

exist.

The Levinson-Whittle algorithm can be described as the following recursions:

Define

$$S_0^f = S_0^b = \hat{R}_0$$
$$A_1^{(1)} = \hat{R}_1(S_0^b)^{-1}$$
$$B_1^{(1)} = \hat{R}_1'(S_0^f)^{-1}$$

for p = 2,3,...,loglogT, we evaluate

$$A_p^{(p)} = \{ \hat{R}_p - \sum_{k=1}^{p-1} A_k^{(p-1)} \hat{R}_{p-k} \} (S_{p-1}^b)^{-1}$$

$$B_p^{(p)} = \{ \hat{R}'_p - \sum_{k=1}^{p-1} B_k^{(p-1)} \hat{R}'_{p-k} \} (S_{p-1}^f)^{-1}$$

$$A_k^{(p)} = A_k^{(p-1)} - A_p^{(p)} B_{p-k}^{(p-1)}, \quad k=1,\ldots,p-1$$

$$B_k^{(p)} = B_k^{(p-1)} - B_p^{(p)} A_{p-k}^{(p-1)}, \quad k=1,\ldots,p-1$$

$$S_p^f = (I - A_p^{(p)} B_p^{(p)}) S_{p-1}^f$$

$$S_p^b = (I - B_p^{(p)} A_p^{(p)}) S_{p-1}^b$$

end for loop.

This algorithm can also be seen in Jones(1978). Whittle(1963) proved that the resulting coefficients estimate $A_k^{(p)}$, k=1,...,p and for p = 1,2,... satisfy the Yule Walker equations which may be written as

$$\sum_{k=0}^{p} A_k^{(p)} \hat{R}_{j-k} = \begin{cases} 0 & \text{if } j > 0 \\ S_p^f & \text{if } j = 0 \end{cases}.$$

For each order p = 1,2,...,loglogT, the resultant pth order AR approximation is given by

$$\tilde{y}_t = A_1^{(p)} \tilde{y}_{t-1} + A_2^{(p)} \tilde{y}_{t-2} + \cdots + A_p^{(p)} \tilde{y}_{t-p} + \varepsilon_t^{(p)}$$

with estimated residual covariance S_p^f.

After obtaining AR approximations of sufficient order, we are ready to go on to next step.

5.2 SVD and Internally Balanced Realisation

General Theory:

Given a pth order AR approximation to $\tilde{y}_t$, we can write $A_p(z^{-1})=\varepsilon_t$ where $A_p(z^{-1}) = I - A_1^{(p)} z^{-1} - \cdots - A_p^{(p)} z^{-p}$. Since we have made assumptions (5.a), $A_p(z^{-1})$ is invertible. Therefore, we may write

$$A_p(z^{-1})^{-1} = \sum_{r \geq 0} \pi_r^{(p)} z^{-r}$$

and observe that $\pi_r^{(p)} \longrightarrow \pi_r$ as $p \to \infty$.

The method to compute $\pi_r^{(p)}$ follows algorithm proposed by Robinson(1967). Basically, $A_p(z^{-1})^{-1}$ is computed by means of the

formula $A_p(z^{-1})^{-1} = \dfrac{\text{adj}A(z^{-1})}{\det A(z^{-1})}$ and the algorithm to evaluate $\text{adj}A(z^{-1})$ and $\det A(z^{-1})$ can be described as follows:

$$p_1 = \text{tr}(A_p(z^{-1})) \qquad B_1 = A_p(z^{-1}) - p_1 I$$

$$A_2 = B_1 A_p(z^{-1}) \qquad p_2 = \text{tr}A_2 \qquad B_2 = A_2 - p_2 I$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$A_h = B_{h-1} A_p(z^{-1}) \qquad p_h = \frac{1}{h} \cdot \text{tr}A_h \qquad B_h = A_h - p_h I$$

Then $\det A_p(z^{-1}) = (-1)^{h-1} p_h$, $\text{adj}A_p(z^{-1}) = (-1)^{h-1} B_{h-1}$ and hence $A_p(z^{-1})^{-1} = \dfrac{B_{h-1}}{p_h}$. Moreover, we note that deg $p_h \leq$ ph and degree of elements of B_{h-1} are $\leq (h-1)p$.

By considering B_{h-1} as matrix of polynomials, the next task is to do the polynomials division which may be described in the following:

We consider a ploynomial $f(x)$ with complex roots $a_k \pm ib_k$, k=1,..,m and real roots c_k, k=1,..,n. For simplicity, we assume no repeated roots on $f(x)$. Therefore,

$$f(x) = \alpha \prod_{k=1}^{m} \{(x - a_k)^2 + b_k^2\} \prod_{k=1}^{n} (x - c_k)$$

After some algebra,

$$\frac{1}{f(x)} = \sum_{k=1}^{m} \frac{(A_k+B_k)(x-a_k)+i(A_k-B_k)b_k}{(x-a_k)^2+b_k^2} + \sum_{k=1}^{n} \frac{C_k}{x-c_k}$$

where

$$A_j = \left\{ \alpha \prod_{\substack{k=1 \\ k \neq j}}^{m} \{(a_j-a_k)+i(b_j-b_k)\} \prod_{k=1}^{m} \{(a_j-a_k)+i(b_j+b_k)\} \prod_{k=1}^{n} \{(a_j-c_k)+ib_j\} \right\}^{-1}$$

$$B_j = \left\{ \alpha \prod_{k=1}^{m} \{(a_j-a_k)-i(b_j+b_k)\} \prod_{\substack{k=1 \\ k \neq j}}^{m} \{(a_j-a_k)-i(b_j-b_k)\} \prod_{k=1}^{n} \{(a_j-c_k)-ib_j\} \right\}^{-1}$$

$$C_j = \left\{ \alpha \prod_{k=1}^{m} \{(c_j-a_k)^2+b_k^2\} \prod_{\substack{k=1 \\ k \neq j}}^{n} (c_j-c_k) \right\}^{-1}$$

Using the identity $\{(x-a_k)^2+b_k^2\}^{-1} = \sum_{s=0} \xi_{k,s} x^s$ where

$$\xi_{k,s} = \sum_{\substack{r+v=s \\ 0\le v\le r\le s}} \frac{1}{(a_k^2+b_k^2)^{r+1}}(-1)^v C_v^r (2a_k)^{r-v},$$

we get the formula for inverse of $f(x)$:

$$(*)\ \frac{1}{f(x)} = -\sum_{k=1}^{m} \xi_{k,0}\{a_k(A_k+B_k)-i(A_k-B_k)b_k\} + \sum_{s\ge 1}\left\{\sum_{k=1}^{m} \xi_{k,s-1}(A_k+B_k) - \sum_{k=1}^{m} \xi_{k,s}\{a_k(A_k+B_k)-i(A_k-B_k)b_k\}\right\}x^s$$

Thus, $(\det A_p(z^{-1}))^{-1}$ can be computed from formula (*).

Finally, suppose

$$\mathrm{adj}A_p(z^{-1}) = \sum_{\mu=0}^{(h-1)p} K_\mu z^{-\mu} \text{ and } (\det A_p(z^{-1}))^{-1} = \sum_{v\ge 0} \omega_v z^{-v}$$

where K_μ's are scalar matrices and ω_v's are scalars. Then

$$A_p(z^{-1})^{-1}(s,t) = \sum_{\delta\ge 0} \sum_{(\mu,v)\in S_\delta} K_\mu(s,t)\omega_v z^{-\delta}$$

where $S_\delta = \{(\mu,v):\ \mu+v = \delta,\ 0 \le \mu \le \min\{(h-1)p,\delta\},\ 0 \le v \le \delta\}$.

This implies that $\pi_r^{(p)}(s,t) = \sum_{(\mu,v)\in S_\delta} K_\mu(s,t)\omega_v$ for $s,t = 1,..,h$.

Given $A_p(z^{-1})^{-1} = \sum_{r\ge 0} \pi_r^{(p)} z^{-r}$, the Hankel matrix approximated by the pth order AR approximation is defined as the following matrix:

$$\mathcal{H}^{(p)} = \begin{bmatrix} \pi_1^{(p)} & \pi_2^{(p)} & \pi_3^{(p)} & \cdots \\ \pi_2^{(p)} & \pi_3^{(p)} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

With this setting, Hannan and Deistler(1988) proved that $\mathrm{rank}(\mathcal{H}^{(p)})$ would be equal to the degree of the polynomial $\det A_p(z^{-1})$ which is $\le hp$. Moreover, if the rank of $\mathcal{H}^{(p)}$ is finite (which is the case) then $\mathrm{rank}(\mathcal{H}^{(p)}) = \mathrm{rank}(\mathcal{H}_p^{(p)})$ where

$$\mathcal{H}_p^{(p)} = \begin{bmatrix} \pi_1^{(p)} & \pi_2^{(p)} & \cdots & \pi_p^{(p)} \\ \pi_2^{(p)} & \pi_3^{(p)} & \cdots & \pi_{p+1}^{(p)} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_p^{(p)} & \pi_{p+1}^{(p)} & \cdots & \pi_{2p-1}^{(p)} \end{bmatrix}.$$

It therefore suggests that we should concentrate on $\mathcal{H}_p^{(p)}$ and our objective becomes to obtain the best approximation of $r_p^{(p)}=\text{rank}(\mathcal{H}_p^{(p)})$ for p = 1,2,...,max. Since our ultimate goal is to identify a suitable model for the data, the rank should be selected in terms of the explanatory power of the family of models with the rank. However, by incorporating a higher rank, one usually can get a better explanatory power. Thus, the rank selection should favour a smaller rank than a large one when their families of models have similar explanatory power. Fortunately, this job can be fulfilled by following arguments from Crabbe and Young(1989) who considered application of Singular Value Decomposition (SVD) and Internally Balanced Realisation to model reduction. Basically, their idea is simple. Representative members from each family of models of different ranks are selected and compared. The rank at which the corresponding representative produces the lowest value of a criterion function is selected. At this point, we would like to put a remark for those readers who thinks that the models identified by maximum likelihood estimators would do the job. This, in fact, may not be the case. For one thing, maximum likelihood estimation requires initial estimates which is what we would like to find. Another problem is that the models identified by maximum likelihood estimators may not be numerically well-behaved. In addition, they may not possess similar properties. Therefore it may be difficult to single out the effect due to increase in dimension.

To formulate the idea from Crabbe and Young(1989), we need the following definitions:

We consider the following state space model:

$$\begin{aligned} \tilde{y}(t) &= Cx(t) + \varepsilon(t) \\ x(t+1) &= Ax(t) + B\varepsilon(t) \end{aligned} \tag{5.2.1}$$

Definition 5.2.2

For state space model (5.2.1), we define the observability and the controllability as $\mathcal{O} = (C'\ (CA)'\ \cdots\ )'$ and $\mathcal{C} = (B\ AB\ A^2B\ \cdots\ )$ respectively. For convenience we write, for $n \in \mathbb{N}$,

$$\mathcal{O}_n = (C'\ (CA)'\ \cdots\ (CA^{n-1})')'$$
$$\mathcal{C}_n = (B\ AB\ \cdots\ A^{n-1}B).$$

Remark:

1. System (5.2.1) is observable and controllable if $\mathcal{O}_n$ and $\mathcal{C}_n$ have full rank n where n = dim(x(t)).
2. $\mathcal{H}_p^{(p)} = \mathcal{O}_p \mathcal{C}_p$

Definition 5.2.3

The observability and controllability grammians of system (5.2.1) are defined as $\mathcal{O}_n'\mathcal{O}_n$ and $\mathcal{C}_n'\mathcal{C}_n$ respectively where n = dim(x(t)).

Definition 5.2.4

The system (5.2.1) is called internally balanced if and only if its observability and controllability grammians are equal.

Given the dimension n of x(t) and the truncated Hankel matrix $\mathcal{H}_n$, we can identify the internally balanced realisation in the following way:

Step 1

Perform a SVD on $\mathcal{H}_n$ to obtain $\mathcal{H}_n = US_nV'$ where U and V are orthonormal matrices, i.e. U'U=V'V=I and $S_n=\text{diag}(s_1,\ldots,s_{nh})$ where h = dim(y(t)).

Define

$$(5.2.5)\quad \mathcal{H}_B=\begin{bmatrix}\pi_1\\ \vdots\\ \pi_n\end{bmatrix} \quad \mathcal{H}_C=(\pi_1 \cdots \pi_n) \quad \text{and} \quad \mathcal{H}_n^{\uparrow}=\begin{bmatrix}\pi_2 & \pi_3 & \cdots & \pi_{p+1}\\ \pi_3 & \pi_4 & \cdots & \pi_{p+2}\\ \cdots & \cdots & \cdots & \cdots\\ \pi_{p+1} & \pi_{p+2} & \cdots & \pi_{2p}\end{bmatrix}$$

Step 2

Write

$$A = S_n^{-1/2}U'\mathcal{H}_n^{\uparrow}VS_n^{-1/2},\ B = S_n^{-1/2}U'\mathcal{H}_B \text{ and } C = \mathcal{H}_CVS_n^{-1/2}.$$

Then we can show that the observability and controllability grammians of the model defined above are both equal to S_n. Moreover, we have the following nice properties with respect to the model defined in the above way:

Results (Crabbe and Young(1989))

1. A model identified by singular values is internally balanced.
2. All internally balanced models are stable if the time series itself is stable.
3. The quantity $\max\{\mu(\mathcal{O}),\mu(\mathcal{C})\}$ achieves its minimum for internally balanced model where $\mu(M) = s_{max}/s_{min}$ and s_{max} and s_{min} are the

largest and smallest singular values of M respectively.

* Note that the higher the condition number $\mu(\cdot)$ the more ill-condition the problem is.

The above results suggest that all internally balanced models are numerically well-behaved and guarantee stability. Moreover, if the internally balanced models were identified by singular values then they are guaranteed to be observable and controllable, i.e. of minimal dimension. Considering models of such qualities can single out the effect due to difference in the dimension of state vector.

For p = 1,2,...,max., pth order autoregression generates a Hankel matrix approximation $\mathcal{H}_p^{(p)}$ with rank $\leq$ hp. Suppose that performing SVD on $\mathcal{H}_p^{(p)}$ identify $s_1,\ldots,s_{hp}$ as singular values. Let $1 \leq n \leq hp$, define $S_1 = \text{diag}(s_1,..,s_n)$ and $S_2 = \text{diag}(s_{n+1},..,s_{hp})$ with $U = (U_1\ U_2)$ and $V = (V_1\ V_2)$ the corresponding partition. Then we can consider the internally balanced model identified by $s_1,\ldots,s_n$ determined in the following way:

$$\tilde{y}(t) = C_1 x(t) + \varepsilon(t)$$

$$x(t+1) = A_1 x(t) + B_1 \varepsilon(t)$$

where

1. $n = \dim(x(t))$
2. $A_1 = S_1^{-1/2} U_1' \mathcal{H}_p^{(p)\uparrow} V_1 S_1^{-1/2}$
3. $B_1 = S_1^{-1/2} U_1' \mathcal{H}_B^{(p)}$
4. $C_1 = \mathcal{H}_C^{(p)} V_1 S_1^{-1/2}$

Remark: $\mathcal{H}_p^{(p)\uparrow}$, $\mathcal{H}_B^{(p)}$, $\mathcal{H}_C^{(p)}$ are defined in (5.2.5).

By comparing values of a criterion functions of each of the internally balanced models for n = 1,2,...,hp and p = 1,2,..,max., we can select the dimension $\hat{n}$ which will give the best fit to our data. In the literature, the most commonly employed criteria for order selection in time series analysis are AIC, BIC and HQ which have the following general form:

(5.2.6) $$CF(n) = -2\text{log-likelihood} + C_T d(n)$$

where n is the dimension under consideration, C_T is a balancing constant (to be discussed) and d(n) is an increasing function of

n. AIC, BIC and HQ correspond to $C_T = 2$, $C_T = \log T$ and $C_T = 2c\log\log T$, where $c > 1$, respectively. The rationale behind (5.2.6) is that although -2log-likelihood can be regarded as a measure of goodness of fit, its value decreases as the dimension increases. Hence, an additional penalty term should be added to prevent the selection of larger than necessary order. One of the prominent criteria for choosing C_T is that CF(n) will eventually pick the correct order when more and more information are coming. If we can use this as our principal criteria for choosing C_T, we have the following result due to Hannan(1984):

Theorem (5.2.7)

Assume that there is a true n_0 and $\hat{n}$ minimises (5.2.6) while, as $T \longrightarrow \infty$, $C_T/T \longrightarrow 0$. Then the following holds, where a.s. stands for "almost surely".

(i) If $\lim_{T\longrightarrow\infty} \inf\ C_T/(2\log\log T) > 1$ then $\hat{n} \longrightarrow n_0$ a.s..

If $\lim_{T\longrightarrow\infty} \sup\ C_T/(2\log\log T) < 1$ then $\hat{n}$ does not converge a.s. to n_0.

(ii) If $\lim_{T\longrightarrow\infty} \inf\ C_T = \infty$ then $\hat{n} \longrightarrow n_0$ in probability.

If $\lim_{T\longrightarrow\infty} \sup\ C_T < \infty$ then $\lim_{\delta\to 0} \lim_{T\to\infty} P(\hat{n} > n_0) = 1$.

We observe the following facts for AIC, BIC and HQ:

1. AIC is not a consistent criteria and would tend to overestimate the true order of the system.
2. BIC and HQ are both strongly consistent. However, using Quinn(1980)'s words, BIC will tend to underestimate the true order in small samples, relative to the HQ criteria, which is also strongly consistent, but minimally so.

If we can judge different criteria from their consistency then the choice is really between BIC and HQ (assuming confined to the three criteria). But practical experience demonstrated that BIC will pick an order which is equal to or less than the order picked by HQ. In this sense, HQ tends to be more conservative. We are, therefore, inclined to HQ criteria. Another important decision we have to make is the choice of d(n). Usually, d(n) will be chosen to be the number of parameters. The problem is that we are talking about multivariate modelling. The number of parameters in a multivariate models are normally proportional to n^2, i.e.

$d(n) \propto n^2$ which may be too heavy a penalty and may lead to the choice of lower than necessary order. To be conservative, we simply use $d(n) = n$. Therefore, the criterion function we would use is HQ(n) = -2log-likelihood + $2n\lceil \text{loglogT} \rceil$ where $\lceil \alpha \rceil$ is the ceiling of α.

When there are no missing observation, Kalman Filtering provides us a convenient way to calculate the -2log-likelihood. Our trouble is the existence of missing observations. Therefore, -2log-likelihood are evaluated by means of Kalman Filtering with adjustment idea to allow missing observations proposed by Jones(1980). For detail, we refer to Jones(1980) or Chapter 3 of this thesis.

As a summary, the following algorithmic-like statements present the sequence of works to achieve order selection of state space modelling:

for (p = 1,2,...,max.) {

using pth order AR approximation to construct $\mathcal{H}_p^{(p)}$.

perform SVD on $\mathcal{H}_p^{(p)}$ with singular values $s_1,\ldots,s_{hp}$.

for (n = 1,2,...,hp) {

the internally balanced model identified by $s_1,..,s_n$ is constructed.

evaluate $HQ_p(n)$ via Kalman Filtering with Jones(1980) adjustment idea to allow missing observations.

}

set n_p the lowest order at which $\min_{1 \le n \le hp} HQ_p(n)$ is attained.

}

We choose $\hat{n}$ such that $\min_{\substack{1 \le n \le hp \\ 1 \le p \le max}} HQ_p(n)$ is attained.

In addition, the way to choose max is to look at the sequence $\{n_p : p = 1,2,\ldots\}$. If the sequence converges as p increases then we stop and choose the largest p value we have checked so far as the max. However, if the sequence doesn't converge then we would choose loglogT as our max.

Finally, before presenting our results, we would like to add some general comments:

1. As indicated by Crabbe and Young(1989), Moore(1978)'s idea of identifying large break in singular values to select state space dimension often fails to work in practice. It is because

singular values are usually declined smoothly.

2. Otter(1985) also pointed out that we may regard singular values as canonical correlation coefficients. Therefore the null hypothesis of $s_{n+1} = s_{n+2} = \cdots = s_{hp} = 0$ can be tested by means of the statistic

$$-\left[T - \frac{1}{2}(2hp+1) \right] \log \prod_{i=n+1}^{hp} (1-s_i^2) \sim \chi^2_{(hp-n)(hp-n)}$$

But this method does not work as well because of the increase in value of the statistic cannot keep up with the increase in degree of freedom of the chi-square distribution under the null hypothesis. In addition, Crabbe and Young(1989) critised that the chosen dimension still doesn't guarantee that important dynamic of full model has been captured even the statistic works.

The following table displays the results of this section and final choice of state vector dimension:

Table 5.2.8: HQ values of Internally Balanced Models

n	p = 1	p = 2	p = 3	p = 4
1	341.8516*	344.90326*	351.59243*	362.2592*
2	347.23556	351.35297	357.52401	367.54403
3	352.22501	356.15408	363.59732	373.13926
4		361.13938	370.58027	379.21088
5		365.12553	376.59218	385.32595
6		369.15325	381.44019	390.13367
7			386.03004	396.92352
8			390.29925	401.80347
9			394.33423	405.87364
10				409.92521
11				413.89731
12				418.08524

* indicates the n value at which HQ is minimised for given p.

Therefore, our result suggests the dimension of state vector should be one. This result also seems to be reasonable as components of $\tilde{y}(t)$ are all detrended white blood cell counts of the same patient. They are suspected to produce similar response when subjected to similar external influence. Hence our estimate of dim(x(t)) is 1. This dimension will also be assumed for

subsequent data analysis. This is the so called model identification stage. Based on this result, we can proceed to model estimation and then model diagnosis. (Also see P.2).

Since maximum likelihood estimation often requires initial estimate as part of the inputs, we would use the internally balanced model or order one as our initial estimate. The model is given as follows:

$$A = -0.35003557$$

$$B = (0.04417205 \quad -0.25872236 \quad 0.83188602)$$

$$C = (0.048742754 \quad 0.070962494 \quad 0.241522314)'$$

$$\Omega = \begin{pmatrix} 0.244927 & 0.007205 & 0.016049 \\ 0.007205 & 0.043026 & 0.011359 \\ 0.016049 & 0.011359 & 0.043429 \end{pmatrix}$$

6. Maximum Likelihood Estimation and Its Asymptotic Distribution

Everything we did to prepare the initial estimate is to perform the maximum likelihood estimation in this section. Since the likelihood function is a highly non-linear function, a reasonable initial estimate is very important. This is because most optimisation algorithm searches the optimum closest to the initial estimate given. In this section, we are going to introduce the optimisation algorithm we employ and method to theoretically derive the missing data score function as well as the Fisher information matrix.

Consider the following state space model with regression parameter:

$$\begin{aligned} y(t) &= \beta Z_t + Cx(t) + \varepsilon(t) \\ x(t+1) &= Ax(t) + B\varepsilon(t) \end{aligned} \tag{6.a}$$

where

1. $y(t) \in \mathbb{R}^n$, $x(t) \in \mathbb{R}^m$ and $\varepsilon(t) \in \mathbb{R}^n \sim N(0,\Omega)$
2. $x(0) \sim N(\mu,\Sigma)$
3. $\theta = (vec(A),vec(B),vec(C),vec(\Omega^{-1}),vec(\beta))$

We remark that although Ω^{-1} is included into our parameter, only the upper diagonal and the diagonal elements are actually required because Ω is symmetric. Since Kalman Filter (with adjustment to allow missing observations) can generate the likelihood surface given the model, it can be used to obtain the maximum likelihood estimator of the parameters via an optimisation algorithm. One of the popular optimisation routine is Newton-Raphson algorithm. However, The requirement of computing the inverse of observed/Fisher information matrix at each steps render this procedure impractical. This is because the procedure we develop to compute the Fisher information matrix is not efficient enough for performing an optimisation. Therefore, we consider the alternative Quasi-Newton algorithm developed by Davidon-Fletcher-Powell(1963) which requires only the first derivative. Loosely speaking, when we would like to find a minimum x of f(x) with derivative g(x), the Quasi-Newton procedure can be stated as:

(i) Start with an initial point x_1 and a n×n positive definite symmetric matrix H_1. Usually H_1 is taken as the identity matrix I. Set iteration number as i=1.

(ii) Compute $S_i = - H_i g(x_i)$

(iii) Find the optimum step length λ_i in the direction S_i and set

$$x_{i+1} = x_i + \lambda_i S_i$$

(iv) Test the new point x_{i+1} for optimality. If x_{i+1} is optimal, terminate the iterative process. Otherwise go to next step.

(v) Update H as $H_{i+1} = H_i + M_i + N_i$

where

$$M_i = \lambda_i \frac{S_i S_i'}{S_i' Q_i}, \quad N_i = - \frac{(H_i Q_i)(H_i Q_i)'}{Q_i' H_i Q_i} \quad \text{and} \quad Q_i = g(x_{i+1}) - g(x_i)$$

(vi) Set the new iteration number as i+1 and go to step (ii).

Thus, the only work is to derive the missing data score function. It then serves as the gradient function in conjunction with the likelihood function to perform maximum likelihood estimation. Of course, the gradient at any given parameters can be estimated using finite difference method. It would be even more appropriate if we can develop computation procedure directly.

6.1 Theoretical Determination of Missing Data Score Function

The basic idea to compute the score function is to regard x(0), x(1),..., x(T) as well as the unobserved data y_M and use the formula $Sc(\theta|\tilde{y}_0) = E\{Sc(\theta|\tilde{y}_0, y_M)|\tilde{y}_0\}$. Note that this is also the idea of Shumway and Stoffer(1982). With this setting, the -2log-likelihood of (6.a), apart from the initial distribution, can be written as

$$(6.1.1) \qquad -2\log L = T\log|\Omega| + \sum_{t=1}^{T} \mathrm{tr}\left(\Omega^{-1}(y_t - \beta Z_t - C x_t)(y_t - \beta Z_t - C x_t)'\right).$$

By differentiating (6.1.1) wrt θ, we have

$$\frac{\partial(-2\log L)}{\partial \beta'} = 2\sum_{t=1}^{T}\left((Z_t Z_t')\beta' - Z_t(y_t - Cx_t)'\right)\Omega^{-1}$$

$$\frac{\partial(-2\log L)}{\partial C'} = 2\sum_{t=1}^{T}\left((x_t x_t')C' - x_t(y_t - \beta Z_t)'\right)\Omega^{-1}$$

$$\frac{\partial(-2\log L)}{\partial \Omega^{-1}} = 2\left(\sum_{t=1}^{T}(y_t - \beta Z_t - Cx_t)(y_t - \beta Z_t - Cx_t)' - T\Omega\right)$$

$$\frac{\partial(-2\log L)}{\partial A} = \sum_{t=1}^{T}\left(\frac{\partial x_t'}{\partial A}\right)\left(I_m \otimes C'\Omega^{-1}(y_t - \beta Z_t - Cx_t)\right)$$

$$\frac{\partial x_{t+1}'}{\partial A} = \frac{\partial x_t'}{\partial A}(I_m \otimes A') + (I_m \otimes x_t')\left(\sum_{r=1}^{m}\sum_{s=1}^{m} E_{rs} \otimes E_{rs}'\right)$$

$$\frac{\partial(-2\log L)}{\partial B} = \sum_{t=1}^{T}\left(\frac{\partial x_t'}{\partial B}\right)\left(I_n \otimes C'\Omega^{-1}(y_t - \beta Z_t - Cx_t)\right)$$

$$\frac{\partial x_{t+1}'}{\partial B} = \frac{\partial x_t'}{\partial B}(I_n \otimes A') + (I_m \otimes (y_t - \beta Z_t - Cx_t)')\left(\sum_{r=1}^{m}\sum_{s=1}^{n} U_{rs} \otimes U_{rs}'\right)$$

where E_{rs} and U_{rs} are m×m and m×n matrices respectively with everywhere zeros except the (r,s)-entry which is unity.

They can be used to construct the complete data score function. However, in order to compute the missing data score function, we need the following lemmas:

Lemma (6.1.2)

Given model (6.a) and various definitions from section (3.6), define $\{v_1(n): n \in \mathcal{O}\}$ to be the orthogonal sequence of the data output from the Kalman Filter where $\mathcal{O}$ is the index set of completely and partially observed data and $y_1(t)$ represents the observed part of y(t). Then

$$\begin{aligned}
&\mathrm{Cov}(v_1(n), \varepsilon_t) \\
&= 0 \qquad\qquad \text{if } t > n \\
&= \Omega_{1n} \qquad\qquad \text{if } t = n \\
&= C_{1n}(B\Omega - V_{n-1}\Omega_{1n-1}) \qquad \text{if } t = n-1 \\
&= C_{1n}\left(A^{n-t-1}B\Omega - \sum_{k\in\mathcal{O}\cap N_{t+1}} W_{n-1}\cdots W_{k+1}V_kC_{1k}A^{k-t-1}B\Omega - W_{n-1}\cdots W_{t+1}V_t\Omega_{1t}\right) \\
&\qquad\qquad \text{if } 1 \le t \le n-2 \\
&= C_{1n}\left(A^{n-1}B\Omega - \sum_{k\in\mathcal{O}\cap N_1} W_{n-1}\cdots W_{k+1}V_kC_{1k}A^{k-1}B\Omega\right) \qquad \text{if } t = 0
\end{aligned}$$

where

$$V_n = \begin{cases} AK_n + B\Omega_{1n}Q_n^{-1} & \text{if } y_1(n) \text{ is observed} \\ 0 & \text{if } y_1(n) \text{ is null} \end{cases}$$

$$W_n = \begin{cases} A(I - K_nC_{1n}) - B\Omega_{1n}Q_n^{-1}C_{1n} & \text{if } y_1(n) \text{ is observed} \\ A & \text{if } y_1(n) \text{ is null} \end{cases}$$

and

$$N_z = \{z, \ldots, n-1\}.$$

Proof: (in Appendix B)

Since the orthogonal sequence $\tilde{v}_0 = \{v_1(n): n \in \mathcal{O}\}$ has covariance $\Gamma = \mathrm{diag}\{Q_n, n \in \mathcal{O}\}$,

(6.1.3) $$E\{\varepsilon_t|\tilde{v}_0\} = Cov(\varepsilon_t,\tilde{v}_0)\Gamma^{-1}\tilde{v}_0$$

where $\alpha_t = Cov(\varepsilon_t,\tilde{v}_0) = (Cov(v_1(n),\varepsilon_t)': n \in \mathcal{O})$.

In the sequel, we define $(\cdot)^T = E(\cdot|\tilde{v}_0)$.

Lemma (6.1.4)

If $E(x_n\varepsilon_v'|\tilde{v}_0) = x_n^T\varepsilon_v^{T,} + \gamma_{n,v}^T$ then

$$\gamma_{n,v}^T = \sum_{k=1}^{n} A^{k-1}B\{I(v=n-k)\Omega - \alpha_{n-k}\Gamma^{-1}\alpha_v'\}$$

Proof: in Appendix B.

With Lemma (6.1.2) and (6.1.4), the missing data score function is constructed by evaluating the following conditional expectations:

$$E\left(-\frac{\partial(-2\log L)}{\partial\beta'}|\tilde{v}_0\right) = 2\left(\sum_{t=1}^{T} Z_t\varepsilon_t^{T,}\right)\Omega^{-1}$$

$$E\left(-\frac{\partial(-2\log L)}{\partial C'}|\tilde{v}_0\right) = 2\left(\sum_{t=1}^{T}\{x_t^T\varepsilon_t^{T,}+\gamma_{t,t}^T\}\right)\Omega^{-1}$$

$$E\left(\frac{\partial(-2\log L)}{\partial\Omega^{-1}}|\tilde{v}_0\right) = 2\left(\sum_{t=1}^{T}\{\varepsilon_t^T\varepsilon_t^{T,} + \Omega-\alpha_t\Gamma^{-1}\alpha_t'\} - T\Omega\right)$$

$$E\left(\frac{\partial(-2\log L)}{\partial A}|\tilde{v}_0\right) = \sum_{t=1}^{T}\left(E\{\frac{\partial x_{t-1}'}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_t^T)\right.$$

$$\left.+C'\Omega^{-1}(\varepsilon_t^Tx_{t-1}^{T,}+\gamma_{t-1,t}^{T,})\right)$$

$$E\left(\frac{\partial x_{t+1}'}{\partial A}|\tilde{v}_0\right) = E\left(\frac{\partial x_t'}{\partial A}|\tilde{v}_0\right)(I_m\otimes A')+(I_m\otimes x_t^{T,})\left(\sum_{r=1}^{m}\sum_{s=1}^{m}E_{rs}\otimes E_{rs}'\right)$$

$$E\left(\frac{\partial(-2\log L)}{\partial B}|\tilde{v}_0\right) = \sum_{t=1}^{T}\left(E\{\frac{\partial x_{t-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_t^T)\right.$$

$$\left.+C'\Omega^{-1}(\varepsilon_t^T\varepsilon_{t-1}^{T,}-\alpha_t\Gamma^{-1}\alpha_{t-1}')\right)$$

$$E\left(\frac{\partial x_{t+1}'}{\partial B}|\tilde{v}_0\right) = E\left(\frac{\partial x_t'}{\partial B}|\tilde{v}_0\right)(I_n\otimes A')+(I_m\otimes\varepsilon_t^{T,})\left(\sum_{r=1}^{m}\sum_{s=1}^{n}E_{rs}\otimes E_{rs}'\right)$$

with

$$E\left(\frac{\partial x_0'}{\partial A}|\tilde{v}_0\right) = E\left(\frac{\partial x_0'}{\partial B}|\tilde{v}_0\right) = 0$$

Hence, the missing data score function can be written as

$$(6.1.5) \qquad Sc(\theta|\tilde{v}_0) = \begin{pmatrix} vec\left(E\{\frac{\partial logL}{\partial \beta'}|\tilde{v}_0\}'\right) \\ vec\left(E\{\frac{\partial logL}{\partial C'}|\tilde{v}_0\}'\right) \\ vec\left(E\{\frac{\partial logL}{\partial \Omega^{-1}}|\tilde{v}_0\}\right) \\ vec\left(E\{\frac{\partial logL}{\partial A}|\tilde{v}_0\}\right) \\ vec\left(E\{\frac{\partial logL}{\partial B}|\tilde{v}_0\}\right) \end{pmatrix}$$

Using (6.1.5) and the likelihood function generated by Kalman Filter, we can apply the Quasi-Newton algorithm to search for the maximum likelihood estimator for θ. However, we would like to emphasis that it is highly unlikely that the maximum likelihood estimator that we got is an absolute maximum given the complexity of our problem. Nevertheless the resultant (local) maximum likelihood estimator still can provide sufficient knowledge of the time series we are studying. The resultant maximum likelihood estimator is presented as follows:

$$A = -0.826435$$

$$B = (1.220327 \quad -1.0355834 \quad -0.1390808)$$

$$C = \begin{pmatrix} 0.2076028 \\ -0.0466558 \\ -0.0858516 \end{pmatrix}$$

$$\Omega = \begin{pmatrix} 1.0641004 & 0.0901966 & 0.1669614 \\ 0.0901966 & 0.3008799 & 0.0677159 \\ 0.1669614 & 0.0677159 & 0.2753071 \end{pmatrix}$$

	NC	LC	OWBC
Intercept	1.12038883	0.83005156	0.52600002
Time	-0.00439112	-0.00371089	-0.00436932
$Time^2$	-0.000000684	0.000032265	0.00002879
$Stage(4)_t$	0.46445116	-0.02913479	0.28264870
$Cycle(2)_t$	0.12054024	-0.12609582	-0.13655230
$Cycle(3)_t$	0.29513262	-0.41505453	0.28903085
$Cycle(4)_t$	-0.23176494	-0.12070265	-0.10199792
$Cycle(5)_t$	-0.01955215	0.11745007	0.09964845
$Cycle(6)_t$	-0.29253325	-0.05196021	-0.06101628
$Cycle(7)_t$	-0.02817001	0.00053197	0.03605690
$Cycle(8)_t$	0.60887323	-0.11708711	-0.22177885

	NC	LC	OWBC
I_t	-0.17509676	0.11313516	-0.03417007
T_t	0.25888047	0.16914444	0.11406385
T_{t-1}	-0.11471028	-0.04873878	-0.00879848
EC_t	-1.16726202	-0.13046037	-0.48700271

In addition, the mean μ of the initial distribution estimated by $x^T(0)$ is equal to -1.5894453.

The above number does not mean anything without presenting how predictable the model is. We recall that the predicted series is given by the formula $\hat{y}(t) = \hat{\beta}Z_t + C\bar{x}_t$. The graph of $y(t)'y(t)$ and $\hat{y}(t)'\hat{y}(t)$ are presented in Fig. A.6. (note that $y(t)'y(t)$ represent the total white blood cell counts) In the graph, the dots represent the true time series while the continuous curve represents the predicted series. We immediately see that the predicted series fluctuates in a much greater degree in the left hand of the graph than that in the right hand side. This is due to the high concentration of missing observations in the right and side of the graph (i.e. after the treatment is effectively ceased). Our final model is effectively predicting the mean level on the right hand region of the graph. The fact was also noted by Jones(1986). When there are long sequence of consecutive missing observations, the Kalman Filter will return to the initial conditions. This is because $A^n \longrightarrow 0$ as $n \to \infty$ if A is stable. When there are long sequence of missing observations, the Kalman Filter will be kept at the initial conditions and thereby always predicts the mean level.

The following table present the interpolaion of missing observations using the model identified by the maximum likelihood estimator:

Table (6.1.7)

Time	Interpolated value	Time	Interpolated value
6	2.2156089	13	1.3082855
15	1.4528858	16	1.8443138
17	1.5133487	23	1.3942716
24	1.9522409	29	1.9239400
30	1.7398371	32	1.7542331
33	1.5995066	35	1.6898852

Time	Interpolated value	Time	Interpolated value
36	1.6002237	37	1.5550082
51	3.2277080	58	3.9213289
60	3.7440563	69	1.6941932
103	5.0234864	104	3.5429072
106	3.3288833	108	3.0029529
121	4.1146683	123	3.9631638
129	4.2084826	131	4.1082674
133	4.0778179	135	4.0466077
137	3.9266340	139	4.1238640
140	4.4613222	142	4.1395639
144	4.1963945	145	4.3351203
146	4.2431988	147	4.3016673
148	4.2857680	149	4.4156185
154	4.1155785	155	4.4181845
158	4.1426616	159	4.6237761
160	4.3091850	162	4.6077391
164	4.2411435	165	4.5764014
166	4.2567938	168	4.2534283
169	4.6919796	170	4.2302241
171	4.8061348	172	4.1861729
173	4.9796078	174	4.9734674
176	4.4372436	177	4.6725351
178	4.4827085	179	4.6978288
180	4.5165550	181	4.7407716
182	4.5395487	183	4.5628733

We recall that the interpolation is done via the formula $y^T(t) = \hat{\beta}Z_t + Cx_t^T + \varepsilon_t^T$. It is clear that the model is predicting the mean level at the end of the series where a lot of data are missing. Therefore, the suggestion of imputing the interpolated values to where the data are missing before modelling is unjustified.

6.2 Theoretical Evaluation of the Fisher Information Matrix

The determination of Fisher information matrix is considered and is approximated by the information matrix calculated at the maximum likelihood estimator. A well-known result showed that the two information matrix are closed to each other and the inverse of

the true Fisher information matrix provides the asymptotic variance of the resultant maximum likelihood estimator. Therefore, although the calculation involved is not efficient enough, it is still worthwhile to have some idea about the accuracy of the estimator we obtained in last section. Before the matrix can be evaluated, we need the following lemmas:

Lemma (6.2.1)

Let $e = (e_1,\ldots,e_K)'$ be a normal random vector with mean 0 and covariance I. Let $A = (A_1,\ldots,A_K)$ and $B = (B_1,\ldots,B_K)$ be $r\times K$ matrices. Then

(i) $$E\{ee'B'Aee'\} = \left(I(i=j)\{3B_i'A_i + \sum_{\substack{g=1\\ g\neq i}}^{K} B_g'A_g\} + I(i\neq j)\{B_i'A_j + B_j'A_i\}\right)_{\substack{i=1,K\\ j=1,K}}$$

(ii) $E\{ee'B'Ae\} = 0$

Proof: in Appendix B.

In the sequel, we write $K_{a,b}(r,s) = E\{\tilde{e}_0\tilde{e}_0'\Gamma^{-1/2}\alpha_r^{a,}\alpha_s^{b}\Gamma^{-1/2}\tilde{e}_0\tilde{e}_0'\}$ where $\tilde{e}_0 = \Gamma^{-1/2}\tilde{v}_0$ and α_r^i denotes the ith row of α_r.

Lemma (6.2.2)

$E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_h^T x_w^{T,}\}$ is equal to

Case 1: h=0, w=0

$$E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_0^T x_0^{T,}\} = (\alpha_r^i\Gamma^{-1}\alpha_s^{j,}\mu^l\mu^k + \varphi_0^l\Gamma^{-1/2}K_{i,j}(r,s)\Gamma^{-1/2}\varphi_0^{k,})_{l,k=1,m}$$

Case 2: h > 0, w=0

$$E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_h^T x_0^{T,}\} = \sum_{v=1}^{h} A^{v-1}B(\alpha_r^l\Gamma^{-1/2}K_{j,l}(s,h-v)\Gamma^{-1/2}\varphi_0^{k,})_{\substack{l=1,n\\ k=1,m}} + A^hE\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_0^T x_0^{T,}\}$$

Case 3: h=0, w > 0

$$E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_0^T x_w^{T,}\} = E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_w^T x_0^{T,}\}'$$

Case 4: h > 0, w > 0

$$E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_h^T x_w^{T,}\}$$

$$= \sum_{v=1}^{h}\sum_{z=1}^{w} A^{v-1}B(\alpha_r^l\Gamma^{-1/2}K_{j,l}(s,h-v)\Gamma^{-1/2}\alpha_{w-z}^{k,})_{l,k=1}^{n}B'A^{z-1,} + \sum_{v=1}^{h} A^{v-1}B(\alpha_r^l\Gamma^{-1/2}K_{j,l}(s,h-v)\Gamma^{-1/2}\varphi_0^{k,})_{\substack{l=1,n\\ k=1,m}}A^{w,}$$

$$+\sum_{z=1}^{w} A^{h}(\alpha_{r}^{l}\Gamma^{-1/2}K_{j,l}(s,w-z)\Gamma^{-1/2}\varphi_{0}^{k,})'_{\substack{l=1,n\\k=1,m}} B'A^{z-1},$$

$$+A^{h}(\alpha_{r}^{l}\Gamma^{-1}\alpha_{s}^{j,}\mu^{l}\mu^{k}+\varphi_{0}^{l}\Gamma^{-1/2}K_{i,j}(r,s)\Gamma^{-1/2}\varphi_{0}^{k,})_{\substack{l=1,m\\k=1,m}} A^{w},$$

where $\varphi_0 = \mathrm{Cov}(x(0),\tilde{v}_0)$ which is evaluated in Appendix B.

Proof: in Appendix B.

Lemma (6.2.3)

1. $E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_s^{T,})\} = (\alpha_r\Gamma^{-1}\alpha_s^{i,})_{i=1}^{n}$

2. $E\{\mathrm{vec}(x_r^T\varepsilon_r^{T,})\mathrm{vec}(x_s^T\varepsilon_s^{T,})'\} = (E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_r^T x_s^{T,}\})_{\substack{i=1,n\\j=1,n}}$

3. $E\{\mathrm{vec}(x_r^T\varepsilon_s^{T,})\}=(I_n\otimes A^r)(\varphi_0\Gamma^{-1}\alpha_s^{i,})_{i=1}^{n}+\sum_{z=1}^{r}(I_n\otimes A^{z-1}B)(\alpha_{r-z}\Gamma^{-1}\alpha_s^{i,})_{i=1}^{n}$

4. $E\{\mathrm{vec}(\varepsilon_r^T x_v^{T,})\}=(A^v\otimes I_n)(\alpha_r\Gamma^{-1}\varphi_0^{i,})_{i=1}^{m}+\sum_{z=1}^{v}(A^{z-1}B\otimes I_n)(\alpha_r\Gamma^{-1}\alpha_{v-z}^{i,})_{i=1}^{n}$

5. $E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_v^{T,})\mathrm{vec}(\varepsilon_s^T\varepsilon_z^{T,})'\}=(E\{\varepsilon_{vi}^T\varepsilon_{zj}^T\varepsilon_r^T\varepsilon_s^{T,}\})_{i,j=1,n}$

   $E\{\varepsilon_{vi}^T\varepsilon_{zj}^T\varepsilon_r^T\varepsilon_s^{T,}\}=(\alpha_v^l\Gamma^{-1/2}K_{j,l}(z,r)\Gamma^{-1/2}\alpha_s^{k,})_{i,k=1,n}$

6. $E\{x_{vi}^T x_{wj}^T\varepsilon_r^T\varepsilon_s^{T,}\}=\left((E\{\varepsilon_{ri}^T\varepsilon_{sk}^T x_v^T x_w^{T,}\})_{ij}\right)_{i,k=1,n}$

7. $E\{\mathrm{vec}(\varepsilon_r^T x_v^{T,})\mathrm{vec}(\varepsilon_s^T x_w^{T,})'\}=(E\{x_{vi}^T x_{wj}^T\varepsilon_r^T\varepsilon_s^{T,}\})_{i,j=1,m}$

8. $E\{\mathrm{vec}(Z_r\varepsilon_r^{T,})\mathrm{vec}(x_h^T\varepsilon_s^{T,})'\}=(Z_r E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_h^{T,}\})_{i,j=1,n}$

   $E\{\varepsilon_{ri}^T\varepsilon_{sj}^T x_h^T\} = A^h(\alpha_r^l\Gamma^{-1}\alpha_s^{j,}\mu^k)_{k=1,m}$

9. $E\{\mathrm{vec}(Z_r\varepsilon_r^{T,})\} = 0$ and $E\{\mathrm{vec}(Z_r\varepsilon_r^{T,})\mathrm{vec}(\varepsilon_h^T\varepsilon_s^{T,})'\} = 0$

10. $E\{\mathrm{vec}(Z_r\varepsilon_r^{T,})\mathrm{vec}(\varepsilon_s^T x_v^{T,})'\}=(Z_r E\{\varepsilon_{ri}^T x_{vj}^T\varepsilon_s^{T,}\})_{\substack{i=1,n\\j=1,m}}$

    $E\{\varepsilon_{ri}^T x_{vj}^T\varepsilon_s^T\}=\left((E\{\varepsilon_r^T x_v^{T,}\varepsilon_{sk}^T\})_{ij}\right)_{k=1,n}$

    $E\{\varepsilon_r^T x_v^{T,}\varepsilon_{sk}^T\} = (\alpha_s^k\Gamma^{-1}\alpha_r^{p,}\mu^q)_{\substack{p=1,n\\q=1,m}} A^v,$

11. $E\{\mathrm{vec}(x_r^T\varepsilon_w^{T,})\mathrm{vec}(\varepsilon_s^T\varepsilon_v^{T,})'\}=(E\{\varepsilon_{wi}^T\varepsilon_{vj}^T x_r^T\varepsilon_s^{T,}\})_{i,j=1,n}$

    $E\{\varepsilon_{wi}^T\varepsilon_{vj}^T x_r^T\varepsilon_s^{T,}\}=A^r(\alpha_w^i\Gamma^{-1/2}K_{j,k}(v,s)\Gamma^{-1/2}\varphi_0^{l,})_{\substack{i=1,m\\k=1,n}}$

$$+\sum_{z=1}^{r} A^{z-1}B(\alpha_w^{l}\Gamma^{-1/2}K_{j,l}(v,r-z)\Gamma^{-1/2}\alpha_s^{k,})_{l,k=1,n}$$

12. $E\{vec(x_h^T\varepsilon_r^{T,})vec(\varepsilon_s^T x_k^{T,})'\}=\Big(A^h E\{\varepsilon_{rl}^T x_{kj}^T x_0^T \varepsilon_s^{T,}\}$

$$+\sum_{z=1}^{h} A^{z-1}BE\{\varepsilon_{rl}^T x_{kj}^T \varepsilon_{h-z}^T \varepsilon^{T,}\}\Big)_{\substack{l=1,n\\ j=1,m}}$$

(i) $E\{\varepsilon_{rl}^T x_{kj}^T x_0^T \varepsilon_s^{T,}\} = \Big((E\{\varepsilon_r^T x_k^{T,} x_{0l}^T \varepsilon_{sd}^T\})_{lj}\Big)_{\substack{l=1,m\\ d=1,n}}$

$$E\{\varepsilon_r^T x_k^{T,} x_{0l}^T \varepsilon_{sd}^T\}=(\alpha_r^p\Gamma^{-1}\alpha_s^{d,}\mu^q\mu^l+\varphi_0^q\Gamma^{-1/2}K_{p,d}(r,s)\Gamma^{-1/2}\varphi_0^{l,})_{\substack{p=1.n\\ q=1,m}}A^k,$$

$$+\sum_{w=1}^{k}(\alpha_r^p\Gamma^{-1/2}K_{q,d}(k-w,s)\Gamma^{-1/2}\varphi_0^{l,})_{\substack{p=1,n\\ q=1,n}}B'A^{w-1},$$

(ii) $E\{\varepsilon_{rl}^T x_{kj}^T \varepsilon_{h-z}^T \varepsilon^{T,}\} = \Big((E\{\varepsilon_r^T x_k^{T,} \varepsilon_{h-z,l}^T \varepsilon_{sd}^T\})_{lj}\Big)_{l,d=1,n}$

$$E\{\varepsilon_r^T x_k^{T,} \varepsilon_{h-z,l}^T \varepsilon_{sd}^T\}=(\varphi_0^q\Gamma^{-1/2}K_{p,l}(r,h-z)\Gamma^{-1/2}\alpha_s^{d,})_{\substack{p=1,n\\ q=1,m}}A^k,$$

$$+\sum_{w=1}^{k}(\alpha_r^p\Gamma^{-1/2}K_{q,l}(k-w,h-z)\Gamma^{-1/2}\alpha_s^{d,})_{\substack{p=1,n\\ q=1,n}}B'A^{w-1},$$

13. $E\{vec(\varepsilon_h^T\varepsilon_r^{T,})vec(\varepsilon_s^T x_k^{T,})'\} = (E\{\varepsilon_{rl}^T x_{kj}^T \varepsilon_h^T \varepsilon_s^{T,}\})_{\substack{l=1,n\\ j=1,m}}$

which can be evaluated by replacing h-z by h in the above equations.

14. $E\{vec(\varepsilon_s^T x_k^{T,})vec(\varepsilon_h^T\varepsilon_r^{T,})'\} = E\{vec(\varepsilon_h^T\varepsilon_r^{T,})vec(\varepsilon_s^T x_k^{T,})'\}'$

Proof: in Appendix B.

Given the above lemma, we are going to refer the expectation appearing in (i) as $P_i(a,b,..)$ where i=1,2,...,14 and a,b,... are indices of ε's or x's according to their order of appearance. For example,

$$P_4(r,v) = E\{vec(\varepsilon_r^T x_v^{T,})\}$$

$$P_6(v,i,w,j,r,s) = E\{x_{vl}^T x_{wj}^T \varepsilon_r^T \varepsilon_s^{T,}\}$$

$$P_{14}(s,k,h,r) = E\{vec(\varepsilon_s^T x_k^{T,})vec(\varepsilon_h^T\varepsilon_r^{T,})'\}$$

etc..

Since the Fisher information matrix is symmetric, we, in fact, need to evaluate 15 expectations. They can be presented in the

followings:

$$E\left\{\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)'\right\}$$

$$= (\Omega^{-1}\otimes I_{nvar})\left(\sum_{r=1}^{T}\sum_{s=1}^{T}(I_n\otimes Z_r)\alpha_r\Gamma^{-1}\alpha_s'(I_n\otimes Z_s')\right)(\Omega^{-1}\otimes I_{nvar})$$

$$E\left\{\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)'\right\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}\Big(P_2(r,r,s,s)+P_3(r,r)\mathrm{vec}(\gamma^T_{s,s})'+\mathrm{vec}(\gamma^T_{r,r})P_3(s,s)'$$

$$+\mathrm{vec}(\gamma^T_{r,r})\mathrm{vec}(\gamma^T_{s,s})'\Big)(\Omega^{-1}\otimes I_m)$$

$$E\left\{\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)'\right\}$$

$$= T^2\mathrm{vec}(\Omega)\mathrm{vec}(\Omega)' - \sum_{s=1}^{T}T\mathrm{vec}(\Omega)\left(P_1(s,s)+\mathrm{vec}(\Omega-\alpha_s\Gamma^{-1}\alpha_s')\right)'$$

$$-T\sum_{r=1}^{T}\left(P_1(r,r)+\mathrm{vec}(\Omega-\alpha_r\Gamma^{-1}\alpha_r')\right)\mathrm{vec}(\Omega)'$$

$$+\sum_{r=1}^{T}\sum_{s=1}^{T}\Big(P_5(r,r,s,s)+P_1(r,r)\mathrm{vec}(\Omega-\alpha_s\Gamma^{-1}\alpha_s')'$$

$$+\mathrm{vec}(\Omega-\alpha_r\Gamma^{-1}\alpha_r')P_1(s,s)'+\mathrm{vec}(\Omega-\alpha_r\Gamma^{-1}\alpha_r')\mathrm{vec}(\Omega-\alpha_s\Gamma^{-1}\alpha_s')'\Big)$$

$$E\left\{4\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)'\right\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T}\left\{\sum_{v=0}^{r-2}\sum_{w=0}^{s-2}(A^{r-1-v},C'\Omega^{-1}P_6(v,i,w,j,r,s)\Omega^{-1}CA^{s-1-w})_{i,j=1.m}\right.$$

$$+\sum_{v=0}^{r-2}(I_m\otimes A^{r-1-v},C'\Omega^{-1})\Big(P_7(r,v,s,s-1)(I_m\otimes\Omega^{-1}C)$$

$$+P_4(r,v)\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{s-1,s})'\Big)$$

$$+\sum_{w=0}^{s-2}\Big(P_7(s,w,r,r-1)(I_m\otimes\Omega^{-1}C)+P_4(s,w)\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})'$$

$$\Big)'(I_m\otimes\Omega^{-1}CA^{s-1-w})$$

$$+(I_m\otimes C'\Omega^{-1})P_7(r,r-1,s,s-1)(I_m\otimes\Omega^{-1}C)$$

$$+(I_m\otimes C'\Omega^{-1})P_4(r,r-1)\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{s-1,s})'$$

$$+ \text{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})P_4(s,s-1)'(I_m\otimes\Omega^{-1}C)$$

$$+ \text{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})\text{vec}(C'\Omega^{-1}\gamma^{T,}_{s-1,s})'\Big\}.$$

$$E\left\{ 4\text{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)\text{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)' \right\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T}\Bigg\{\sum_{v=0}^{r-2}\sum_{z=0}^{s-2}(I_n\otimes A^{r-1-v,}C'\Omega^{-1})P_5(r,v,s,z)(I_n\otimes\Omega^{-1}CA^{s-1-z})$$

$$+ \sum_{v=0}^{r-2}(I_n\otimes A^{r-1-v,}C'\Omega^{-1})\Big(P_5(r,v,s,s-1)(I_n\otimes\Omega^{-1}C)$$

$$-P_1(r,v)\text{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\Big)$$

$$+ \sum_{w=0}^{s-2}\Big(P_5(s,w,r,r-1)(I_n\otimes\Omega^{-1}C)-P_1(s,w)\text{vec}(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha'_{r-1})'$$

$$\Big)'(I_n\otimes\Omega^{-1}CA^{s-1-w})$$

$$+ (I_n\otimes C'\Omega^{-1})P_5(r,r-1,s,s-1)(I_n\otimes\Omega^{-1}C)$$

$$- \text{vec}(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha'_{r-1})P_1(s,s-1)'(I_n\otimes\Omega^{-1}C)$$

$$- (I_n\otimes C'\Omega^{-1})P_1(r,r-1)\text{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'$$

$$+ \text{vec}(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha'_{r-1})\text{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\Bigg\}$$

$$E\left\{ \text{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)\text{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)' \right\}$$

$$= (\Omega^{-1}\otimes I_{nvar})\sum_{r=1}^{T}\sum_{s=1}^{T}P_8(r,r,s,s)\ (\Omega^{-1}\otimes I_m)$$

$$E\left\{ \text{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)\text{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)' \right\} = 0$$

$$E\left\{ -2\text{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)\text{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)' \right\}$$

$$= (\Omega^{-1}\otimes I_{nvar})\sum_{r=1}^{T}\sum_{s=1}^{T}\Bigg(\sum_{v=0}^{s-2}P_{10}(r,r,s,v)(I_m\otimes\Omega^{-1}CA^{s-1-v})$$

$$+ P_{10}(r,r,s,s-1)(I_m\otimes\Omega^{-1}C)\Bigg)$$

$$E\left\{ -2\text{vec}\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)\text{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)' \right\} = 0$$

$$E\left\{ \operatorname{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)' \right\}$$

$$= (\Omega^{-1}\otimes I_m)\left\{T\cdot\sum_{r=1}^{T}\left(P_3(r,r)+\operatorname{vec}(\gamma^T_{r,r})\right)\operatorname{vec}(\Omega)'\right.$$

$$-\sum_{r=1}^{T}\sum_{s=1}^{T}\left(P_{11}(r,r,s,s)+P_3(r,r)\operatorname{vec}(\Omega-\alpha_s\Gamma^{-1}\alpha_s')'\right.$$

$$\left.\left.+\operatorname{vec}(\gamma^T_{r,r})P_1(s,s)'+\operatorname{vec}(\gamma^T_{r,r})\operatorname{vec}(\Omega-\alpha_s\Gamma^{-1}\alpha_s')'\right)\right\}$$

$$E\left\{ -2\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)' \right\}$$

$$=(\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}\left\{\sum_{v=0}^{s-2}\left(P_{12}(r,r,s,v)+\operatorname{vec}(\gamma^T_{r,r})P_4(s,v)'\right)(I_m\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$+P_{12}(r,r,s,s-1)(I_m\otimes\Omega^{-1}C)+P_3(r,r)\operatorname{vec}(C'\Omega^{-1}\gamma^{T,}_{s-1,s})'$$

$$\left.+\operatorname{vec}(\gamma^T_{r,r})P_4(s,s-1)'(I_m\otimes\Omega^{-1}C)+\operatorname{vec}(\gamma^T_{r,r})\operatorname{vec}(C'\Omega^{-1}\gamma^{T,}_{s-1,s})'\right\}$$

$$E\left\{ -2\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)' \right\}$$

$$=(\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}\left\{\sum_{v=0}^{s-2}\left(P_{11}(r,r,s,v)+\operatorname{vec}(\gamma^T_{r,r})P_1(s,v)'\right)(I_n\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$+P_{11}(r,r,s,s-1)(I_n\otimes\Omega^{-1}C)+\operatorname{vec}(\gamma^T_{r,r})P_1(s,s-1)'(I_n\otimes\Omega^{-1}C)$$

$$\left.-P_3(r,r)\operatorname{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'-\operatorname{vec}(\gamma^T_{r,r})\operatorname{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\right\}$$

$$E\left\{ 2\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)' \right\}$$

$$=\sum_{r=1}^{T}\sum_{s=1}^{T}\left\{\sum_{v=0}^{s-2}\left(P_{13}(r,r,s,v)-\operatorname{vec}(\alpha_r\Gamma^{-1}\alpha_r')P_3(s,v)'\right)(I_m\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$\left.+\left(P_{13}(r,r,s,s-1)-\operatorname{vec}(\alpha_r\Gamma^{-1}\alpha_r')P_3(s,s-1)'\right)(I_m\otimes\Omega^{-1}C)\right\}$$

$$E\left\{ 2\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\operatorname{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)' \right\}$$

$$=\sum_{r=1}^{T}\sum_{s=1}^{T}\left\{\sum_{v=0}^{s-2}\left(P_5(r,r,s,v)-\operatorname{vec}(\alpha_r\Gamma^{-1}\alpha_r')P_1(s,v)'\right)(I_n\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$\left.+\left(P_5(r,r,s,s-1)-\operatorname{vec}(\alpha_r\Gamma^{-1}\alpha_r')P_1(s,s-1)'\right)(I_n\otimes\Omega^{-1}C)\right\}$$

$$E\left\{ 4\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)' \right\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T}\left\{ \sum_{v=0}^{r-2}\sum_{w=0}^{s-2}(I_m \otimes A^{r-1-v},C'\Omega^{-1})P_{14}(r,v,s,w)(I_n \otimes \Omega^{-1}CA^{s-1-w}) \right.$$

$$+\sum_{v=0}^{r-2}(I_m \otimes A^{r-1-v},C'\Omega^{-1})\Big(P_{14}(r,v,s,s-1)(I_n \otimes \Omega^{-1}C)$$

$$-P_4(r,v)\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\Big)$$

$$+\sum_{w=0}^{s-2}\Big((I_m \otimes C'\Omega^{-1})P_{14}(r,r-1,s,w)$$

$$+\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})P_1(s,w)'\Big)(I_n \otimes \Omega^{-1}CA^{s-1-w})$$

$$+(I_m \otimes C'\Omega^{-1})P_{14}(r,r-1,s,s-1)(I_n \otimes \Omega^{-1}C)$$

$$-(I_m \otimes C'\Omega^{-1})P_4(r,r-1)\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'$$

$$+\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})P_1(s,s-1)'(I_n \otimes \Omega^{-1}C)$$

$$\left.-\mathrm{vec}(C'\Omega^{-1}\gamma^{T,}_{r-1,r})\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\right\}$$

The proof of the above equations for the information matrix are given in Appendix B.

We remind that the above calculation refers to the information matrix of the parameter

$$\theta = (\mathrm{vec}(A)',\mathrm{vec}(B)',\mathrm{vec}(C)',\mathrm{vec}(\Omega^{-1})',\mathrm{vec}(\beta)')'$$

and the fact that $\hat{\theta}(T)$ is asymptotically distributed as $N(\theta,\frac{1}{T}\ I(\theta)^{-1})$ where $I(\theta)$ is the Fisher information matrix per sample.

Since the matrix is quite large: 58×58 matrix, we present the eigenvalues of the matrix in increasing order and the asymptotic variance of the parameters as follows:

Eigenvalues of the Fisher Information Matrix:

0.160987225	0.243225331	0.324122216	0.657275020
1.190409585	2.111308722	2.796357324	3.296238969
3.912426872	5.026245128	5.552639450	8.136217922
8.298575918	9.994533496	10.193077178	10.980596860
12.031620888	12.374024183	13.067746240	15.487698795
16.867763729	17.841798196	17.937949571	20.631631492
29.100637571	31.286928108	32.669548589	36.850000370

Eigenvalues of the Fisher Information Matrix:

38.683602364	43.231704042	44.263335112	45.785988210
47.594828126	50.511160288	53.016278475	56.988070244
57.904193321	66.128420448	70.158590048	76.183903705
83.649216817	86.975886386	89.369032551	117.099352864
141.042293876	143.582955887	178.229626174	221.6812059482
249.521971004	529.124136420	848.516172821	1530.6178612867
1.348×10^{5}	4.530×10^{5}	7.9326×10^{5}	2.55141×10^{10}
8.50195×10^{11}	1.484764×10^{11}		

Parameter	Asy. Var of θ	Upper 95% CL	Lower 95% CL
A(1,1)	0.002598168	-0.726529543	-0.926340676
B(1,1)	0.016506490	1.472143425	0.968511289
B(1,2)	0.010639381	-0.833414510	-1.237752221
B(1,3)	0.014341684	0.095642467	-0.373804071
C(1,1)	0.009963591	0.403245673	0.011959940
C(2,1)	0.002360885	0.048578565	-0.141890075
C(3,1)	0.003305943	0.026843220	-0.198546314
$\Omega^{-1}(1,1)$	0.000375748	1.085793072	1.009806928
$\Omega^{-1}(1,2)$	0.000242313	-0.150589844	-0.211610156
$\Omega^{-1}(1,3)$	0.000213263	-0.562277087	-0.619522912
$\Omega^{-1}(2,2)$	0.000742563	3.603110018	3.496289982
$\Omega^{-1}(2,3)$	0.000682336	-0.712001738	-0.814398261
$\Omega^{-1}(3,3)$	0.001453939	4.253035882	4.103564118
$\beta(1,1)$	0.001353768	1.192504254	1.048273396
$\beta(1,2)$	6.1691×10^{-6}	0.000477062	-0.009259306
$\beta(1,3)$	2.1756×10^{-10}	2.82261×10^{-5}	-2.95936×10^{-5}
$\beta(1,4)$	0.001975219	0.551560294	0.377342025
$\beta(1,5)$	0.000952114	0.181018676	0.060061805
$\beta(1,6)$	0.001270104	0.364984118	0.225281125
$\beta(1,7)$	0.001313378	-0.160733449	-0.302796434
$\beta(1,8)$	0.001457838	0.055283878	-0.094388168
$\beta(1,9)$	0.001765263	-0.210183785	-0.374882714
$\beta(1,10)$	0.001892309	0.057091322	-0.113431336
$\beta(1,11)$	0.002455968	0.658430949	0.559315502
$\beta(1,12)$	0.000331931	-0.139387576	-0.210805953
$\beta(1,13)$	0.000329698	0.294469343	0.223291598
$\beta(1,14)$	0.000244859	-0.084040257	-0.145380304

Parameter	Asy. Var of θ	Upper 95% CL	Lower 95% CL
$\beta(1,15)$	0.004394438	-1.037332529	-1.297191513
$\beta(2,1)$	0.001947074	0.916537859	0.743565266
$\beta(2,2)$	9.2933×10^{-6}	0.002264157	-0.009685929
$\beta(2,3)$	2.6872×10^{-10}	6.43947×10^{-5}	1.35363×10^{-7}
$\beta(2,4)$	0.004222807	0.098232146	-0.156501718
$\beta(2,5)$	0.000241781	-0.095619177	-0.156572467
$\beta(2,6)$	0.000609793	-0.366654314	-0.463454743
$\beta(2,7)$	0.000431639	-0.079981844	-0.161423465
$\beta(2,8)$	0.000685611	0.168771048	0.066129083
$\beta(2,9)$	0.000893884	0.006639661	-0.110560079
$\beta(2,10)$	0.000997425	0.062432761	-0.061368819
$\beta(2,11)$	0.000830059	-0.060618051	-0.173556170
$\beta(2,12)$	0.000186326	0.139889412	0.086380909
$\beta(2,13)$	0.000101791	0.188919182	0.149369703
$\beta(2,14)$	7.0432×10^{-5}	-0.032289718	-0.065187837
$\beta(2,15)$	0.001412356	-0.056800977	-0.204119772
$\beta(3,1)$	0.000436399	0.566944739	0.485055292
$\beta(3,2)$	1.3560×10^{-6}	-0.002086949	-0.006651687
$\beta(3,3)$	2.8899×10^{-11}	3.93282×10^{-5}	1.8255×10^{-5}
$\beta(3,4)$	0.000725574	0.335444200	0.229853195
$\beta(3,5)$	0.000221578	-0.107376724	-0.165727869
$\beta(3,6)$	0.000636804	0.338491406	0.239570301
$\beta(3,7)$	0.000330671	-0.066356568	-0.137639265
$\beta(3,8)$	0.000372232	0.137463348	0.061833553
$\beta(3,9)$	0.000459704	-0.018992492	-0.103040071
$\beta(3,10)$	0.000534429	0.081367632	-0.009253832
$\beta(3,11)$	0.000657442	-0.171523213	-0.272034490
$\beta(3,12)$	8.3486×10^{-5}	-0.016261417	-0.052078722
$\beta(3,13)$	0.000105383	0.134184465	0.093943226
$\beta(3,14)$	9.323051×10^{-5}	0.010126486	-0.027723444
$\beta(3,15)$	0.001274745	-0.417023705	-0.556981704

Since, so far, we only have the information matrix for Ω^{-1}, it might be important if we can obtain the information matrix of Ω. The derivation can be easily seen as follows:

we note that

(6.2.4) $$\frac{\partial \log L}{\partial \Omega} = \left(\frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)\left(I_n \otimes \frac{\partial \log L}{\partial vec(\Omega^{-1})}\right)$$

by chain rule of matrix differentiation.

Since logL is a scalar,

(6.2.5) $$\frac{\partial \log L}{\partial vec(\Omega^{-1})} = vec\left(\frac{\partial \log L}{\partial \Omega^{-1}}\right).$$

Hence, combining (6.2.4) and (6.2.5) and taking expectation, we get

(6.2.6) $$E\left(\frac{\partial \log L}{\partial \Omega}|\tilde{v}_0\right) = \left(\frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)\left(I_n \otimes vec\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\right)$$

The Fisher information matrix of Ω is defined as

$$E\left\{vec\left(E\{\frac{\partial \log L}{\partial \Omega}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial \log L}{\partial \Omega}|\tilde{v}_0\}\right)'\right\}$$

$$= E\left\{\left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)vec\left(I_n \otimes vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})\right)\right.$$
$$\left.vec\left(I_n \otimes vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})\right)'\left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)'\right\}$$

$$= E\left\{\left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)\left(\bar{U} \otimes vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})'\right)\right.$$
$$\left.\left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)'\right\}$$

$$= \left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)\left(\bar{U} \otimes E\left[vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})'\right]\right)$$
$$\left(I_n \otimes \frac{\partial vec(\Omega^{-1})'}{\partial \Omega}\right)'$$

where $\bar{U} = \sum_{r=1}^{n}\sum_{s=1}^{n} E_{rs} \otimes E_{rs}$ and E_{rs} is n×n matrix with everywhere zeros except (r,s) entry which is unity.

Note that the quantity $E\left[vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})vec(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\})'\right]$ is the information matrix of Ω^{-1}. The remaining quantity to calculate is $\frac{\partial vec(\Omega^{-1})'}{\partial \Omega}$ which can be seen as equal to $\left(vec(\frac{\partial \Omega^{-1}}{\partial \Omega_{rs}})'\right)_{r,s=1,n}$ where Ω_{rs} is the (r,s) element of Ω. Therefore it immediately follows that

(6.2.7) $$\frac{\partial vec(\Omega^{-1})'}{\partial \Omega} = \left(vec\left[\left(\frac{\partial \Omega^{-1}}{\partial \Omega}\right)_{rs} \right]' \right)_{r,s=1,n}$$

where $\left(\frac{\partial \Omega^{-1}}{\partial \Omega}\right)_{rs}$ is the (r,s) block matrix of $\frac{\partial \Omega^{-1}}{\partial \Omega}$ and $\frac{\partial \Omega^{-1}}{\partial \Omega}$ is equal to $-(I_n \otimes \Omega^{-1})\bar{U}(I_n \otimes \Omega^{-1})$.

Based on the above derivation, the information matrix of Ω are computed and the following table presents the asymptotic variance and confidence limits for Ω:

Parameter	Asy. Var of Ω	Upper 95% CL	Lower 95% CL
$\Omega(1,1)$	5.32914×10^{-5}	1.078408619	1.049792243
$\Omega(1,2)$	8.62463×10^{-6}	0.095952657	0.084440514
$\Omega(1,3)$	1.05950×10^{-5}	0.173341216	0.160581609
$\Omega(2,2)$	4.58579×10^{-6}	0.305077105	0.296682634
$\Omega(2,3)$	2.68477×10^{-6}	0.070927400	0.064504371
$\Omega(3,3)$	4.11952×10^{-6}	0.279285235	0.271328972

The above confidence limits demonstrate that most of our parameters are significant which also partly justifies our modelling results.

6.3 General Comments

6.3.1 On the information matrix

When there is no missing observations, it is well-known that the information matrix of the mean μ and covariance Σ for independent data is block diagonal. However, the presence of missing observations destroys this nice simple structure of the matrix. Although the information matrix of MLE of parameters of state space model does not possess block diagonal structure, the situation where there is no missing observations would be simpler than the situation where missing data are present. For example,

(i) $\varepsilon^T(t)$ can now be written as $y(t)-\beta Z_t-Cx^T(t)$. This is impossible when $y(t)$ is missing and we must deal with large matrix computations.

(ii) $E(\varepsilon_t^T \varepsilon_s^{T,}) = -CP_{t,s}^T C'$ which is in a much simpler form than where missing observations are present.
($P_{t,s}^T = E\{(x(t)-x_t^T)(x(s)-x_s^T)'\}$)

(iii) $E(x_t^T \varepsilon_s^{T,}) = P_{t,s}^T C'$

etc..

The essential feature is that large matrix computation can be avoided under complete data case while this is the fact of life when missing observations are present. Another important observation is that missing data introduces time-varying elements into our model thereby eliminating the simple expressions that would normally possess. The fact that the presence of missing observations reduce the amount of information available to us would decrease the predictability of our resultant model and would potentially increase the variability of our parameters estimate. Moreover, That the evaluation of the information matrix with missing data requires more matrix computations and summations can accumulate small numerical errors into a large one. Fortunately, the division of the matrix by T to get the asymptotic variance of the resultant MLE would hopefully average out numerical errors. But, nonetheless, we must bear in mind all of above stated facts when actually computing the informaiton matrix.

6.3.2 On the condition number and its implication

The condition number of the resultant information matrix is defined as

$$\text{Cond}(I(\hat{\theta})) = \frac{\max_{\sigma} \sigma(I(\hat{\theta}))}{\min_{\sigma} \sigma(I(\hat{\theta}))} = 9.223\times10^{11}$$

where $\sigma(A)$ is the set of all eigenvalues of A.

The larger the condition number the more ill-condition is the problem since large condition number would suggest "near" singularity of the information matrix. When we invert a "near" singular matrix, some elements of the inverse of the matrix may become very large while other elements may be very small. Because of the fact that the inverse of the information matrix is the asymptotic variance of the resultant MLE, some parameters might be difficult to be "precisely" estimated in the sense that small pertubations of the data would generate large change in the resultant MLE. In addition, large condition number also suggests the existence of multicollinearity among the parameters. This might then imply some degree of over-parameterization.

However, the question of whether the condition number we got is large or small should be related to the accuracy of the computer routine we are using. Since we are using double precision

and Fortran programming, the inversion of a matrix having condition number of order 10^{12} is still easy to be handled by the computer. In this sense the condition number we got might be regarded as "tolerable". From the experience of performing maximum likelihood estimation, we understood that the regression parameter β is quite imprecise in the sense that different β could return similar likelihood values. Therefore, we suspect that large condition number is basically contributed by this parameter. In order to check this, we extract the submatrices of the information matrix: one corresponds to parameters (A,B,C,Ω^{-1}) and the other corresponds to the parameter β and then we look at their condition numbers: 1860.6 for the former parameters and 1.989×10^{11} for the latter parameter. It is clear that β causes problems. We believe that β, in this sense, is over-parameterised and is sensitive to changes in the data such as missing observations. To ascertain this fact, we conduct a simple experiment by doing a series of robust regressions in the following ways:

1. performing robust regression using the full data set.
2. performing robust regression after deleting the last 10 observations from the full data set.
3. performing robust regression after deleting the first 10 observations from the full data set.

The results are presented as follows:

Robust regression of Neutrophyll Count:

Variable	Method 1	Method 2	Method 3
Intercept	1.517664	1.208377	1.968725
Time	-0.02901509	-0.01150007	-0.0482302
$Time^2$	0.00008175	-0.0000538	0.00014882
$Stage(4)_t$	0.8277621	0.6196104	0.9539674
$Cycle(2)_t$	0.3687869	0.331184	0.5927351
$Cycle(3)_t$	1.570074	1.464443	1.977053
$Cycle(4)_t$	0.3644192	0.3149468	0.7853358*
$Cycle(5)_t$	0.5444561	0.5030243	0.9983883*
$Cycle(6)_t$	0.3562458	0.3826191	0.8608703*
$Cycle(7)_t$	0.7861587	0.8365339	1.407742*
$Cycle(8)_t$	1.625328	1.878715	2.486768
I_t	-0.1176541	0.0117856	-0.127912

Variable	Method 1	Method 2	Method 3
T_t	0.2121458	0.2392529	0.09443031*
T_{t-1}	-0.03467786	-0.09598981	0.07491938
EC_t	-2.278478	-2.288462	-2.428543

*** indicates a large change in parameter value**

Robust Regression of Lymphocyte Count

Variable	Method 1	Method 2	Method 3
Intercept	0.8880246	0.9693024	0.8307554
Time	-0.006514467	-0.01370897	-0.004505752
$Time^2$	0.00004118	0.0001144	0.000033321
$Stage(4)_t$	0.01868138	0.1180292*	-0.01649054
$Cycle(2)_t$	-0.0819256	-0.08131741	-0.08177412
$Cycle(3)_t$	-0.3585921	-0.2772415	-0.4326497
$Cycle(4)_t$	-0.07209091	-0.1075976	-0.09400357
$Cycle(5)_t$	0.1704516	0.1118862	0.1398458
$Cycle(6)_t$	-0.0057674	-0.1347797*	-0.05760165*
$Cycle(7)_t$	0.01198965	-0.1277085	-0.009052041
$Cycle(8)_t$	-0.01366096	-0.2693194*	-0.1019663*
I_t	0.1063091	0.04089525	0.1210783
T_t	0.1797428	0.09714966	0.2232403
T_{t-1}	-0.08639844	-0.04694136	-0.04048935
EC_t	-0.09881993	-0.1701984	-0.0873267

*** indicates large change in parameter value**

Robust Regression of OWBC

Variable	Method 1	Method 2	Method 3
Intercept	0.6121939	0.5981516	0.6586402
Time	-0.008121073	-0.007915518	-0.009917787
$Time^2$	0.000038734	0.000041631	0.0000461507
$Stage(4)_t$	0.2518527	0.2482623	0.2639189
$Cycle(2)_t$	0.005527853	0.002134049	0.01628845
$Cycle(3)_t$	0.5179667	0.5056138	0.5311766
$Cycle(4)_t$	0.06324554	0.05077456	0.08748825
$Cycle(5)_t$	0.2938493	0.2741133	0.3241875
$Cycle(6)_t$	0.1577063	0.1265185	0.1833834
$Cycle(7)_t$	0.2293405	0.1998561	0.272100
$Cycle(8)_t$	0.02022268	-0.02124715	0.07491306
I_t	-0.04180376	-0.04744431	-0.05929698

Variable	Method 1	Method 2	Method 3
T_t	0.1326148	0.1327162	0.137184
T_{t-1}	-0.05240194	-0.03314368	-0.024566
EC_t	-0.7095874	-0.71710	-0.7087664

We observe that the change in β values in OWBC are less drastic than that in other white blood cell counts. Although not all β values show drastic changes, the above experiment demonstrates that β is really quite difficult to track. That's why it causes a lot of problems when we perform maximum likelihood estimation. However, it should be reminded that the above experiment does not represent any conclusive evidence for the claim that β is unstable.

7. Diffuse Kalman Filtering

In this section, we consider the following state space model:

$$\begin{aligned} y_t &= \beta Z_t + G x_t + \varepsilon_t \\ x_{t+1} &= F x_t + H \varepsilon_t \end{aligned} \tag{7.1}$$

where

1. $\varepsilon_t \sim N(0,\Omega)$ is i.i.d..
2. $x(0)=0$ and $\beta=b+B\gamma$ where $\gamma \sim (0,C)$.
3. γ and ε_t's are uncorrelated.
4. C is nonsingular unless C=0.

Note that model (7.1) differs from models we studied previously the inclusion of a random element γ. The diffuse Kalman Filter deals with model (7.1) when $C \longrightarrow \infty$. This can be simply interpreted as representing imprecise knowledge on the regression parameter β. The topic has been studied extensively by Piet de Jong(1990-1992). The followings are summaries of Jong(1991)'s results with adjustment to allow the presence of missing observations:

Let $\tilde{y}_0 = (y_{1t},\ t \in \mathcal{O})$ where y_{1t} represents the observed part of y_t and $\mathcal{O}$ is the index set of the completely or partially observed data. Then

Result (7.2)

$$\mathrm{vec}(\tilde{y}_0) = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_T \end{pmatrix} \mathrm{vec}(Z) + \mathrm{vec}(\upsilon)$$

where T = no. of observations, β_t is the components of β which corresponds to the observed part of y_t, $Z = (Z_t,\ t=1,\ldots,T)$, $\mathrm{vec}(\upsilon) \sim (0,\Sigma)$ with Σ non-singular where υ is a $n \times T$ matrix whose components are linear combinations of ε_t's and $\mathrm{Cov}(\mathrm{diag}(\beta_1,..,\beta_T),\mathrm{vec}(\upsilon)) = 0$.

Therefore the -2log-likelihood of model (7.1) assuming normality should be given as

$$\begin{aligned} -2\log L = {} & \log|C| + \gamma^2/C + \log|\Sigma| \\ & + (\mathrm{vec}(\tilde{y}_0) - \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_T \end{pmatrix} \mathrm{vec}(Z) - \begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_T \end{pmatrix} \mathrm{vec}(Z)\gamma)'\Sigma^{-1}(\cdot) \\ & - \log|(C^{-1}+S)^{-1}| - \left(\gamma - (C^{-1}+S)^{-1}s \right)'(C^{-1}+S)\left(\cdot \right) \end{aligned}$$

where

$$S = \text{vec}(Z)' \begin{pmatrix} B_1' & & \\ & \ddots & \\ & & B_T' \end{pmatrix} \Sigma^{-1} \begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_T \end{pmatrix} \text{vec}(Z)$$

and

$$s = \text{vec}(Z)' \begin{pmatrix} B_1' & & \\ & \ddots & \\ & & B_T' \end{pmatrix} \Sigma^{-1} \left(\text{vec}(\tilde{y}_0) - \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_T \end{pmatrix} \text{vec}(Z) \right).$$

This would simplifies to

$$(7.3) \quad -2\log L = \log|C| + \log|C^{-1}+S| + \log|\Sigma| + \{q - s'(S+C^{-1})^{-1}s\}$$

where

$$q = \left(\text{vec}(\tilde{y}_0) - \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_T \end{pmatrix} \text{vec}(Z) \right)' \Sigma^{-1} \left(\cdot \right).$$

For detail of above derivation, we refer to Jong(1991).

In the above derivation, C is assumed to be finite. We then can let C tend to infinity to see

Result (7.4)

Suppose that $\tilde{y}_0$ follows the state space model (7.1) and γ and $\tilde{y}_0$ are normally distributed. If S is nonsingular, then as $C \longrightarrow \infty$, $-2\log L - \log|C|$ converges to

$$(7.5) \qquad \log|S| + \log|\Sigma| + \{q - s'S^{-1}s\}$$

We define (7.5) as the proper likelihood for state space model (7.1) when the random variable γ is diffuse, i.e. $C \longrightarrow \infty$. Jong(1991) also showed that (7.5) is the -2loglikelihood of a linear transformation of the data. This, in fact, justifies the sense that (7.5) is "proper". In the following we would start discussing the diffuse Kalman Filter which can be described simply as a convenient algorithm to compute the likelihood (7.5) given the state space model (7.1).

The diffuse Kalman Filter for model (7.1) with adjustment to allow missing observations is defined as the following recursions:

Initial conditions: $A_1 = 0$, $Q_1 = 0$, $\lambda = 0$ and

$$\text{vec}P(1) = \{I - F \otimes F\}^{-1} \text{vec}(H\Omega H')$$

For t = 1,2,...,T do {

if y_{1t} is not null then

1. $e(t) = (B_t Z_t,\ y_{1t} - b_t Z_t) - G_{1t} A_t$
2. $D(t) = G_{1t} P(t) G'_{1t} + \Omega_{11t}$
3. $K(t) = (FP(t)G'_{1t} + H\Omega'_{1t}) D_t^{-1}$
4. $A_{t+1} = FA_t + K(t)e(t)$
5. $P(t+1) = (F - K(t)G_{1t})P(t)F' + H\Omega H' - K(t)\Omega_{1t}H'$
6. $\lambda = \lambda + \log|D(t)|$
7. $Q_{t+1} = Q_t + e(t)'D(t)^{-1}e(t)$

else (if y_{1t} is null)

1. $A_{t+1} = FA_t$
2. $P(t+1) = FP(t)F' + H\Omega H'$

end if

}

Then $Q_{T+1} = \begin{pmatrix} S & s' \\ s & q \end{pmatrix}$ with q a scalar and the -2log-likelihood is equal to $\log|S| + \lambda + (q - s'S^{-1}s)$. (proof of this result can be seen in Jong(1991) and the extension to include the case where missing observations are present is trivial.)

Therefore, if the dimension of the state vector which was selected is appropriate for this case, then maximum likelihood estimation can be performed by selecting a suitable optimisation algorithm and a suitable initial estimate. The optimisation algorithm we use is also Quasi-Newton algorithm. The difference this time is that the score function is estimated numerically. Furthermore, the initial estimate are provided via the balanced state space realisation calculated previously with the additional parameter $\hat{B} = (\underset{\sim}{1}\ \underset{\sim}{1}\ \underset{\sim}{1})'$ where $\underset{\sim}{1}$ is a column of ones. The reason for picking $\hat{B}$ as initial estimate of B is to allow the effect of "diffuse" to come into play though it is purely arbitrary. Our result is as follows:

$F = -0.16051024$

$H = (0.227565 \ -0.4318305 \ 0.4455304)$

$$G = \begin{pmatrix} 0.682541 \\ -0.382653 \\ -0.356860 \end{pmatrix}$$

$$\Omega = \begin{pmatrix} 1.1882064 & 0.0585012 & 0.0863229 \\ 0.0585012 & 0.3973338 & 0.0733945 \\ 0.0863229 & 0.0733945 & 0.3769363 \end{pmatrix}$$

MLE of b

Variable	NC	LC	OWBC
Intercept	0.72885357	0.13470258	0.073450711
Time	-0.0467857	-0.00888054	-0.006178684
$Time^2$	0.00014476	0.00002431	0.0000154532
$Stage(4)_t$	0.45251177	0.02845532	-0.013388443
$Cycle(2)_t$	0.95838241	0.24097937	0.228827503
$Cycle(3)_t$	1.03684315	0.23350959	0.156513250
$Cycle(4)_t$	1.37486061	0.37381460	0.318694681
$Cycle(5)_t$	1.56841224	0.37170208	0.307199776
$Cycle(6)_t$	1.76911824	0.42748566	0.368912162
$Cycle(7)_t$	1.97756861	0.45792621	0.385783580
$Cycle(8)_t$	2.39914014	0.53977062	0.453431817
I_t	-0.02289524	-0.0009008	-0.000524081
T_t	0.06412896	0.04586134	0.040127768
T_{t-1}	0.03885622	-0.0004205	0.006944403
EC_t	0.23865714	0.17202869	0.181377411

MLE of B

Variable	NC	LC	OWBC
Intercept	0.67653242	0.73216234	0.430236409
Time	0.02732713	0.00160544	0.001197684
$Time^2$	-0.0000965	0.00001688	0.000011952
$Stage(4)_t$	0.10214520	-0.0097453	0.281959246
$Cycle(2)_t$	-0.55009632	-0.2868103	-0.28617713
$Cycle(3)_t$	-0.35800038	-0.6080889	0.249382000
$Cycle(4)_t$	-1.17196990	-0.3462447	-0.32806975
$Cycle(5)_t$	-1.08082111	-0.1211409	-0.13551286
$Cycle(6)_t$	-1.51334418	-0.3225347	-0.32376610

Variable	NC	LC	OWBC
$Cycle(7)_t$	-1.37675188	-0.2854432	-0.25248056
$Cycle(8)_t$	-1.00320601	-0.4650583	-0.54615673
I_t	-0.19175136	0.12943413	-0.05643648
T_t	0.207706707	0.15103276	0.118865380
T_{t-1}	-0.16449796	-0.0374457	0.000805565
EC_t	-1.18551628	-0.3562519	-0.77406679

Again, only presenting the numbers may not mean anything at all. Therefore, we need to consider the predictability of the model (7.1) identified by the MLE. The quantities

$$\hat{\gamma}_t = E\{\gamma \,|\, y_{1t}, t\in\mathcal{O}\cap(1,..,t-1)\}$$

and

$$\bar{x}_t = E\{x(t) \,|\, y_{1t}, t\in\mathcal{O}\cap(1,2,\ldots,t-1)\}$$

are important since $\hat{y}_t = (b_t+B_t\hat{\gamma}_t)Z_t+C\bar{x}_t$. However, from the condition distribution of γ given y's, we can immediately deduce that

$$\hat{\gamma}_t = S_t^{-1}s_t$$

where the notation

$$Q_t = \begin{pmatrix} S_t & s_t' \\ s_t & q_t \end{pmatrix} \text{ is used.}$$

Moreover,

$$\begin{aligned}\bar{x}_t &= E\left(E\{x(t) \,|\, \gamma, y_{1t}, t\in\mathcal{O}\cap(1,\ldots,t-1)\} \,|\, y_{1t}, t\in\mathcal{O}\cap(1,..,t-1)\right) \\ &= E\left(A_t\begin{pmatrix} -\gamma \\ 1 \end{pmatrix} \,|\, y_{1t}, t\in\mathcal{O}\cap\{1,\ldots,t-1\}\right) \\ &= A_t\begin{pmatrix} -\hat{\gamma}_t \\ 1 \end{pmatrix}\end{aligned}$$

since $A_t(-\hat{\gamma}_t,1)'$ satisfies the ordinary Kalman Filtering equation. Hence we use the predicted series defined as $\hat{y}_t = (b_t+B_t\hat{\gamma}_t)+C\bar{x}_t$ to obtain a plot of $\hat{y}_t'\hat{y}_t$ and $y_t'y_t$ against time where $y_t'y_t$ is the total white blood cell counts. The plot is shown in Figure A.8. One prominent feature is that this graph is extremely similar to that obtained from the predicted series of ordinary Kalman Filter. The choice between diffuse Kalman Filter and ordinary Kalman Filter is really a preference of the user and does not come from

the power of one method over the other if only judged from the two predicted series. However, we do believe that the assumption of a "diffuse" β in our case may not be valid since our data describes the white blood cell counts of only one patient and the inclusion of an extra parameter uses up a lot of degree of freedom which may pose serious numerical problems. Therefore, our preference still favours ordinary Kalman Filtering. On the contrary, if our data describes the white blood cell counts of more than one patient then diffuse Kalman Filter may be better in the following sense:

1. different patients may demonstrate different β empirically and if there is a β for all patients, then we may regard it as "diffuse" to reflect our innocent about this parameter.
2. More patients give us more data. Therefore, inclusion of an additional parameter may actually help explain the data without putting too much burden.

Nevertheless, diffuse Kalman Filtering is a very efficient and powerful tools to modelling wide range of time series.

8. Conclusions

The major contribution of this thesis are the introduction of an approach to modelling multiple time series with missing observations and the theoretical derivation of the Fisher information matrix. Our proposed procedure starts by estimating the autocovariance sequence using an idea proposed by Parzen(1963) and Stoffer(1986). The idea utilises an "amplitude modulating" series of the original time series and the consistency of the procedure relies on the assumption that the missing data generating mechanism is stationary. We then obtain successive AR approximations via the Levinson-Whittle algorithm. The potential deficiencies of the procedure are that (1) the resulting error covariance sequence may not be positive definite and the existence of bias toward zero when sample size is small. This is because the presence of missing observations creates an unbalance situation and distorts the nice structure of the resulting estimates. Without missing observations, the preferred option would be the multivariate generalisation of the Burg's algorithm. The Burg's algorithm guarantees that the resulting error covariance matrices are positive definite but requires the use of the whole data series. Thus, the presence of missing observations prevent the implementation of the algorithm. Dunsmuir and Robinsion(1981) proposed that we may use Burg's algorithm if we have a long consecutive sequence of observed series. This is, in fact, impossible in our situation because (1) there exists a change in mean level in our data series and (2) the missing observations are scattered around the whole time span though a relatively high concentration is in the rear part of the series. Nevertheless, Levinson-Whittle algorithm does provide us a good starting point to prepare for further analysis. Given the AR approximations, Hankel matrices which represents the relationship between future output and past input are constructed by inverting the matrix polynomial of autoregression. Hannan and Deistler(1988) proved that the rank of the Hankel martix constructed by a pth order AR approximation is less than or equal to np where n is the dimension of the time series. With the knowledge, there is no need to consider rank which is bigger than np. Therefore, for

p=1,2,..,max, a singular value decomposition of the pth Hankel matrix (the Hankel matrix approximated by a pth order Autoregression) is performed to identify singular values $s_1,\ldots,s_{np}$. Then for $v = 1,2,\ldots,np$, balanced state space realisation of the pth Hankel matrix assuming $s_{v+1} = \cdots = s_{np} = 0$ is constructed and its likelihood is also evaluated via Kalman Filtering with adjustment to allow missing observations. The pair (p_0,v_0) which gives the minimum of a model selection criterion developed by Hannan and Quinn(1980) is then selected. The balanced realisation of the selected Hankel matrix provides an initial estimate for maximum likelihood estimation. In order to perform maximum likelihood estimation efficiently, we employ a formula developed by Louis(1982) and Breckling, Chambers, etc.(1990) to compute the missing data score function. This enables us to use Quasi-Newton algorithm to search for (local) maximum of the likelihood function. Finally, theoretical evaluation of missing data information matrix is considered thereby allowing the computation of the asymptotic accuracy of the resulting MLE.

We believe that the above proposed procedure represents an interesting approach to time series modelling with missing observations. It improves on Jones(1980) idea to allow partial missing observations and makes possible the identification of prediction error form thus extending the EM approach proposed by Shumway and Stoffer(1982) which identifies a restricted family of state space models where the observation noise and the process noise are independent. One interesting computational aspect in determining the rank of the Hankel matrix is that the statistic proposed by Otter(1985) does not work. This statistic is derived by regarding the singular values as canonical correlations. We then resort to balanced state space realisation proposed by Crabbed and Young(1989) which is theoretically well-behaved to help solve the problem. Although the computation of information matrix seems complex, it invariably represents the first attempt to provide the asymptotic accuracy of the maximum likelihood estimator. We believe the procedure can be simplified to allow a more efficient computation.

With respect to the implementation of numerical algorithm, we find that it took a long time to search for the maximum likelihood estimator of β. This fact is also illuminated from the large condition number of the information matrix of the MLE of β. Large condition number implies that the parameter β is probably over-parameterised and is quite sensitive to change in data. This also supports the role of robust regression to obtain initial estimate of β. On the contrary, other parameters behave much more stable. This, in fact, demonstrates the stability of the structural parameters of state space model. The small extension to allow β to be "diffuse" illustrates how diffuse Kalman Filter (with adjustment to allow missing observations) works. The diffuse Kalman Filter is proposed and developed by Jong(1991) and represents an interesting extension to ordinary Kalman Filter. Our adjustment idea to allow missing observations for diffuse Kalman Filter in a large part follows Jones(1980) and is different from what Jong(1991) was proposed. We believe that this adjustment idea is theoretically sound and easy to understand.

All of the analysis in this thesis assumes that the missing data generating mechanism is ignorable. This assumption is quite difficult to either justify or to disprove. For one thing, the most popular method to justify ignorability is to do a logit analysis of the missing data indicator against given auxiliary information. When the missing data mechanism can be explained perfectly using our auxiliary information, we understand that the missing data mechanism is missing at random and therefore ignorable if the analysis is conditional on the auxiliary information. However, we can never obtain a perfect fit from real data and one of the characteristics of logit analysis is that the residuals are often large even when the auxiliary information is a very good explanator of the missing data mechanism. Thus, it is extremely difficult to judge whether the missing data mechanism is ignorable or not. On the other hand, it is very difficult to justify non-ignorability because the information is not available at where the data are missing. Moreover, if we believe the missing data mechanism is non-ignorable then the missing data mechanism cannot be determined and therefore must be assumed. It infers that

one possible extension of this thesis is to consider non-ignorability of missing data mechanism. However, the assumption of the missing data mechanism should be realistic to the data collection process.

Finally, we would like to comment on the choice of modeling procedure in this thesis. One reason we employ the modeling procedure is because of its smoothness and completeness. The procedure begins by doing ignorability analysis and model identification. Then it goes to model estimation and diagnosis. Moreover, the resulting form of the score function expression allows us to derive the information directly (although it is not an algorithmic approach). This is better than method introduced by Kohn and Ansley(1985) where we need to know the model first before we can "difference" the data. Another reason is that all adjustments are made directly in consistent with the original proof. No need to impute zeros in place where missing data occurs as suggested by Shumway and Stoffer(1982) and Brockwell and Davies(1991) (although, in some cases, the two may be just the same). Lastly, the idea of imputing missing observations using the existing data is completely rejected.

Appendix A

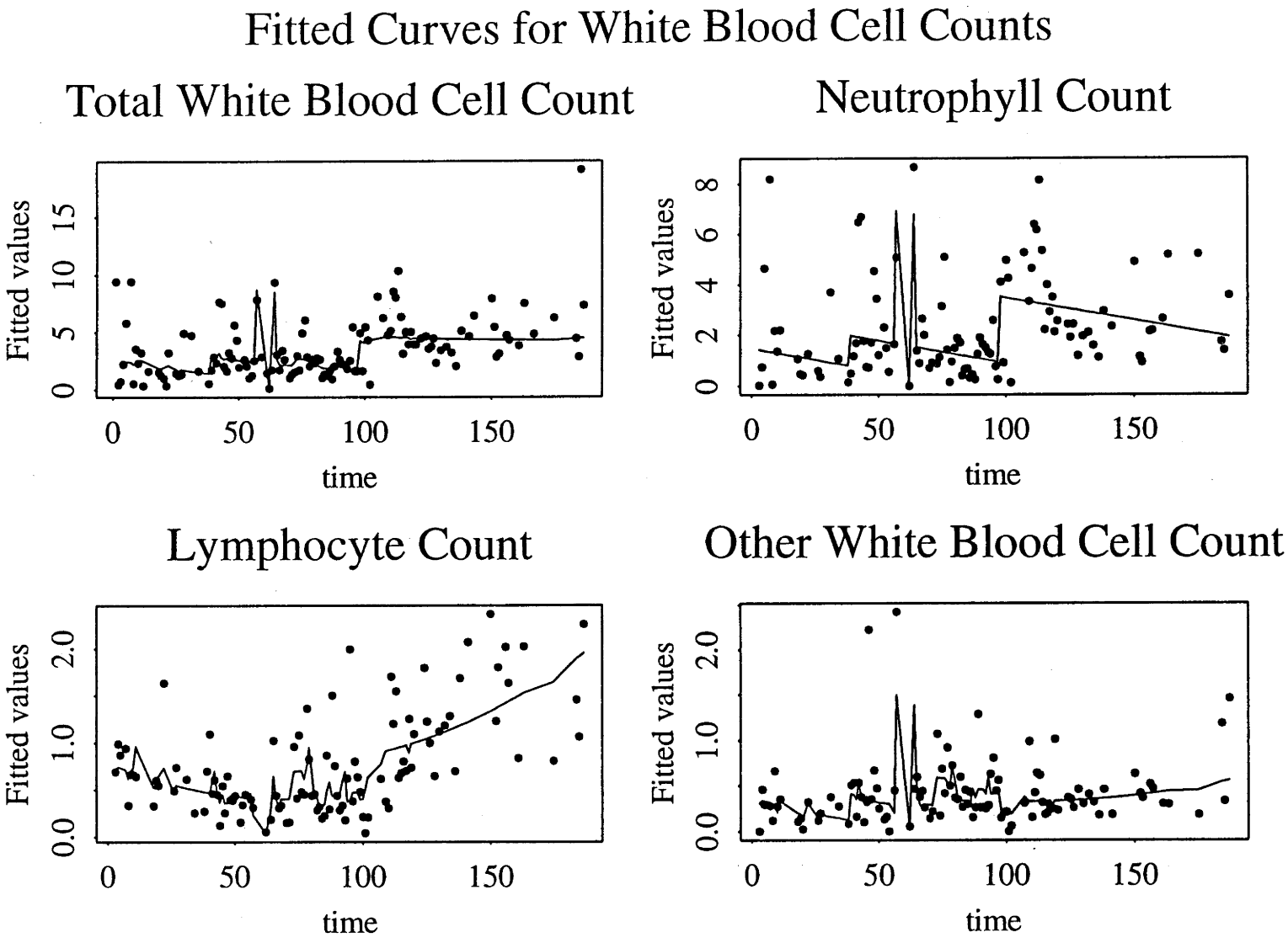

Figure A.1 - Huber Robust Regression used

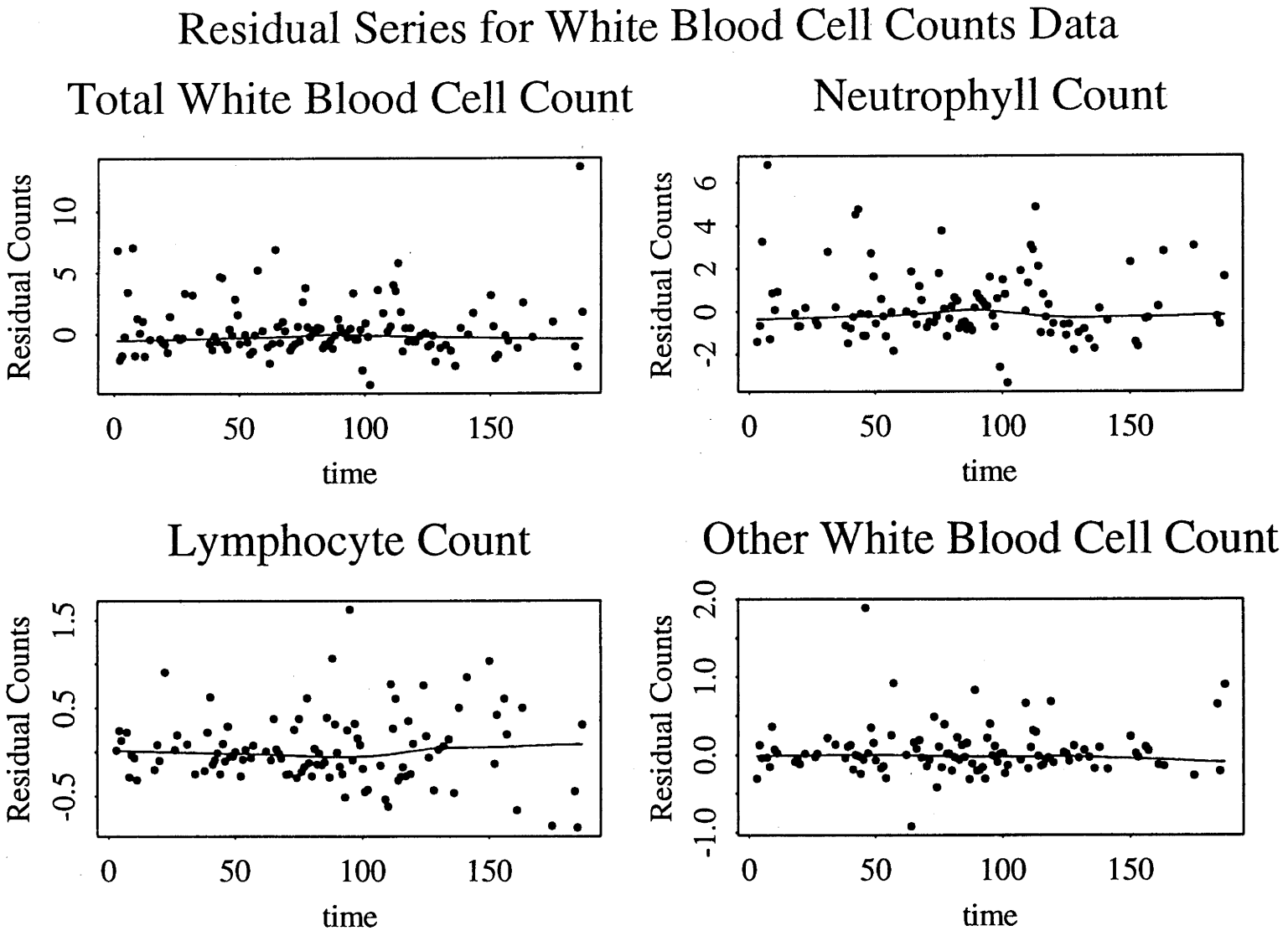

Figure A.2 - Lowess lines are drawn

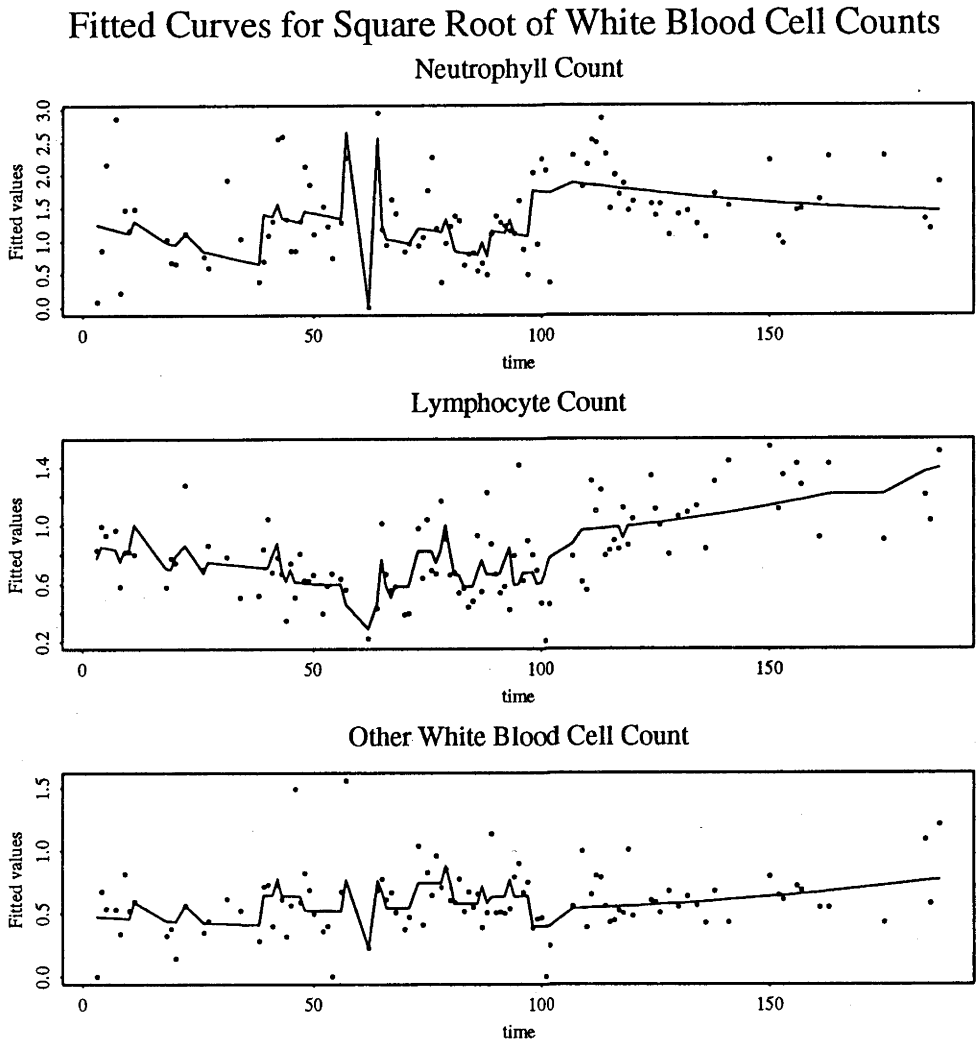

Figure A.3 - Huber Robust Regression used

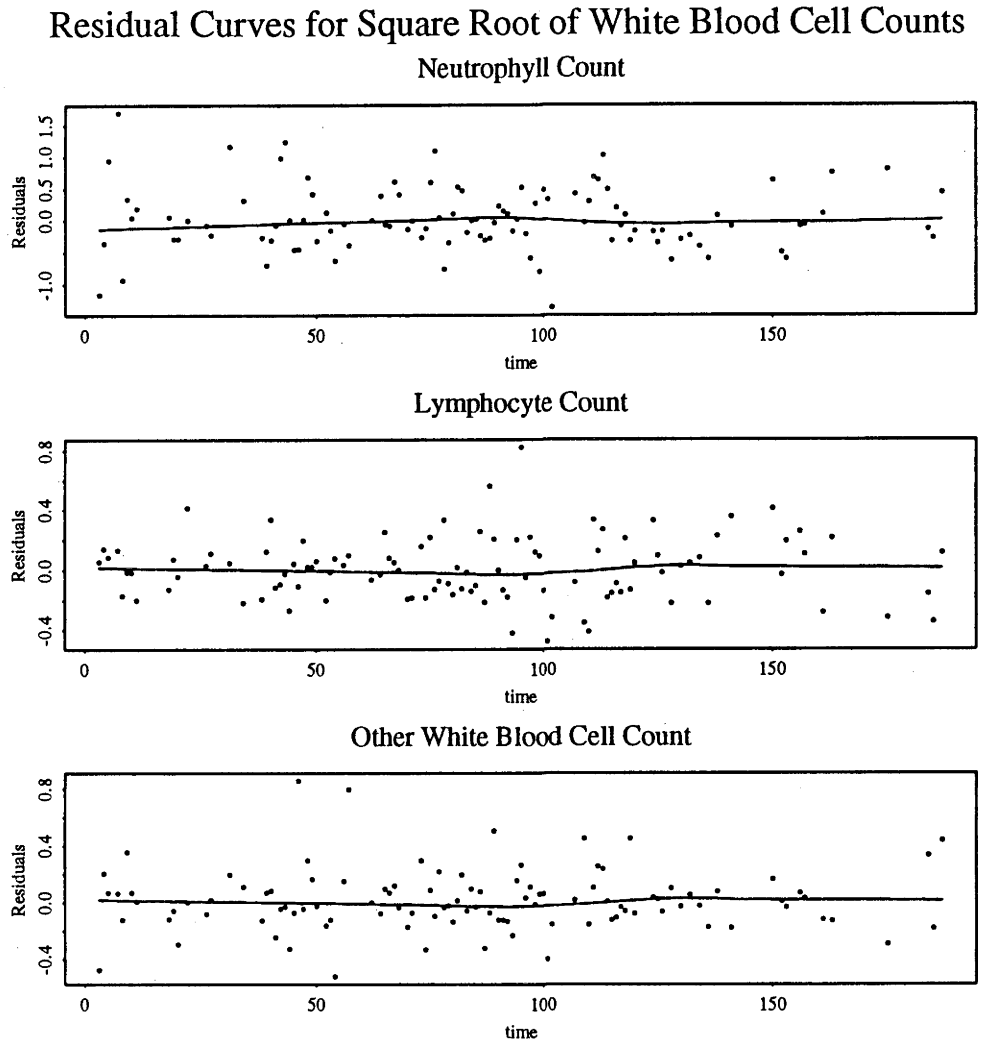

Figure A.4 - Lowess Lines are drawn

Predicted Series for White Blood Cell Counts generated by Kalman Filtering

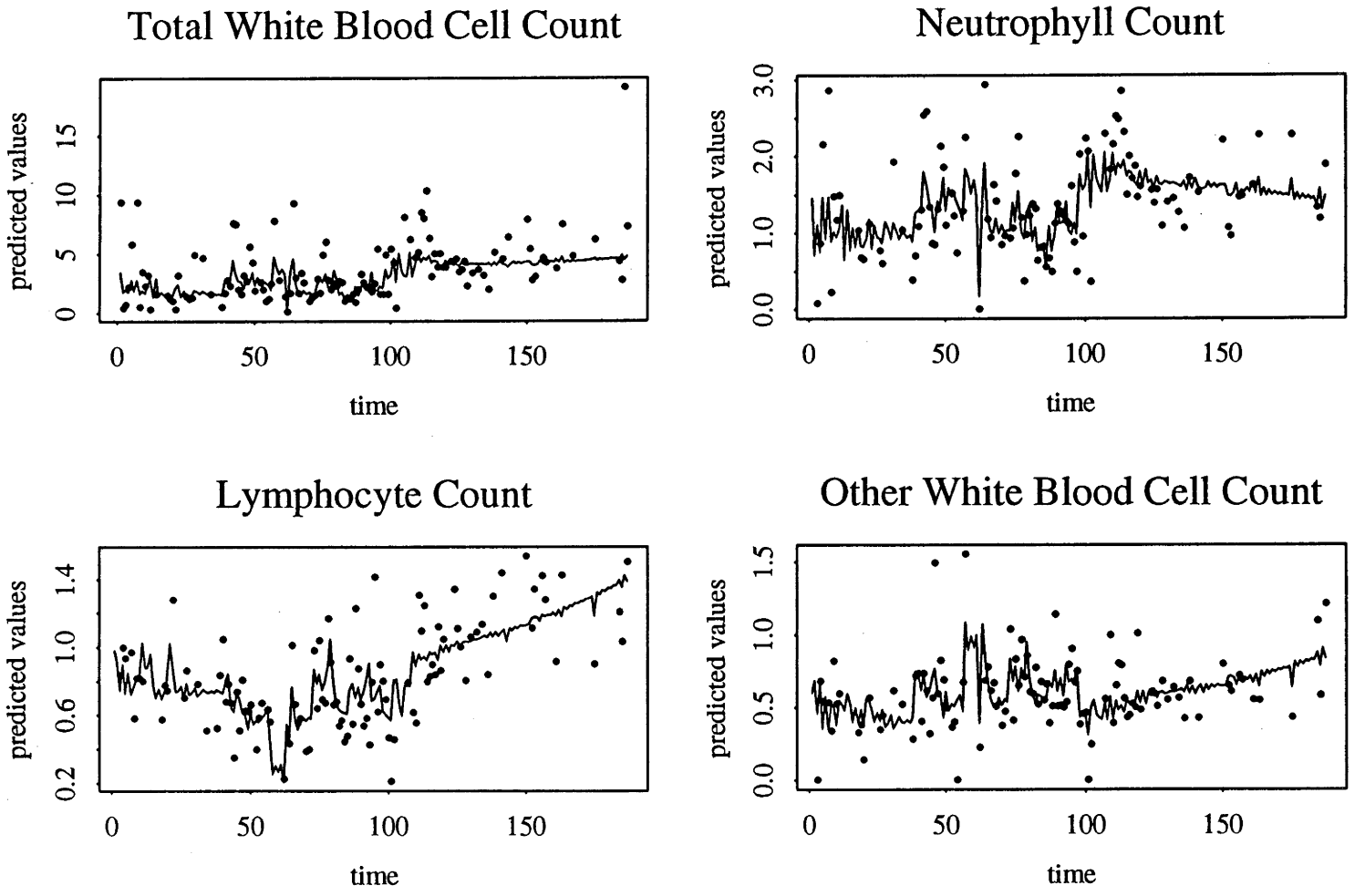

Figure A.5 - Total count is level data and components are square root data

Predicted Series for Total White Blood Cell Counts generated by Kalman Filtering

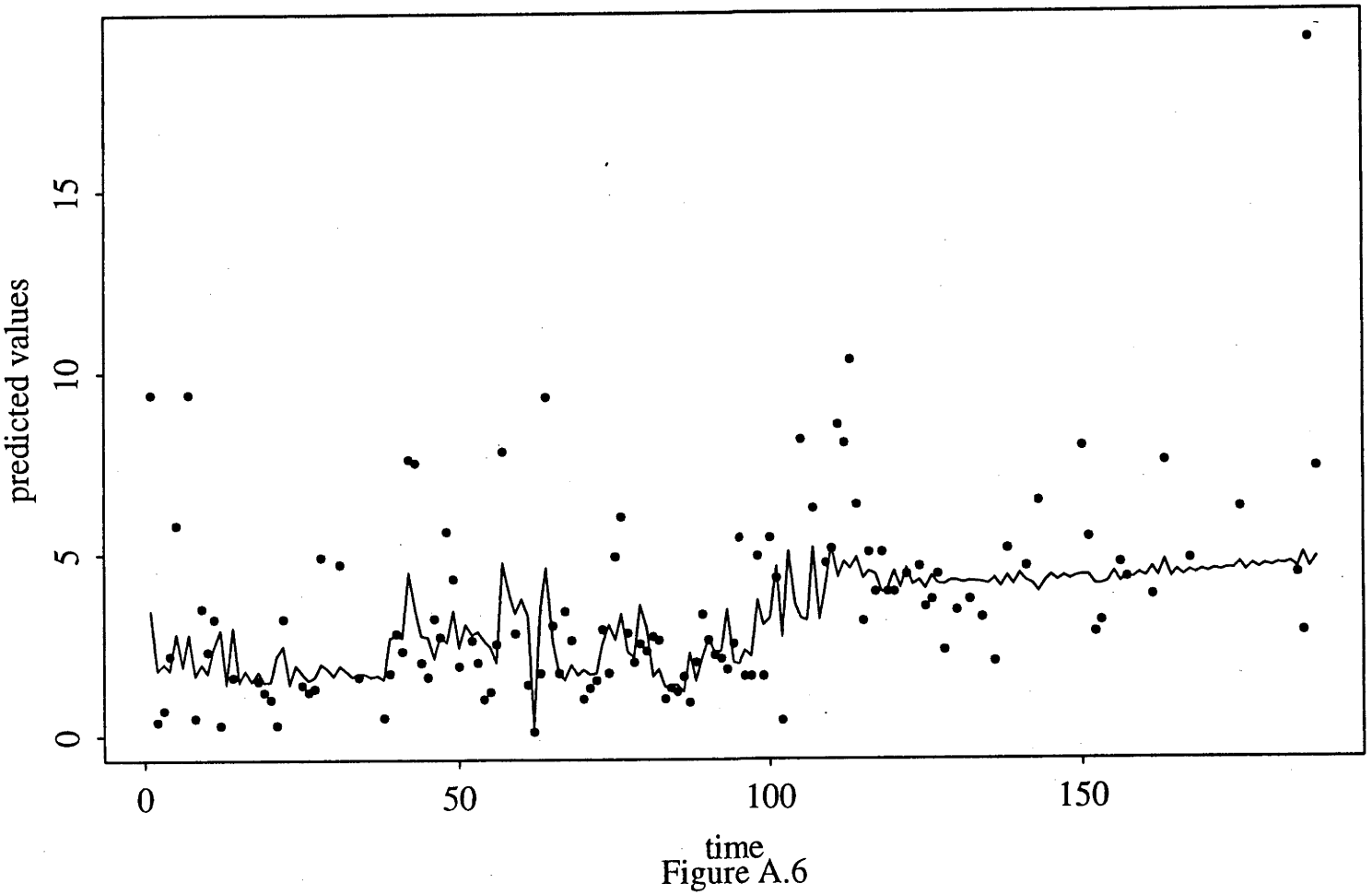

Figure A.6

Predicted Series for White Blood Cell Counts generated by Diffuse Kalman Filtering

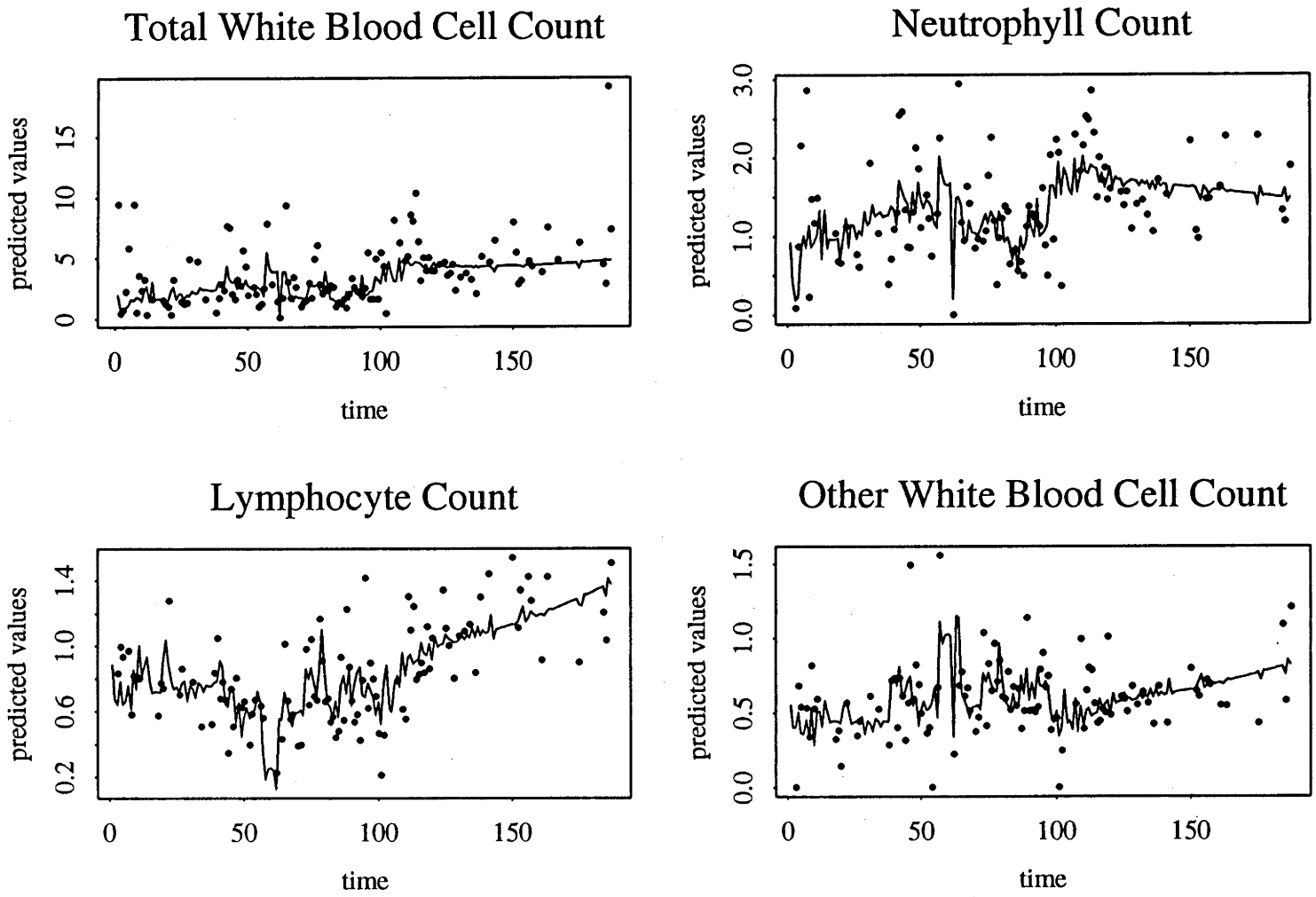

Figure A.7 - Total count is level data and components are square root data

Predicted Series for Total White Blood Cell Counts
generated by Diffuse Kalman Filtering

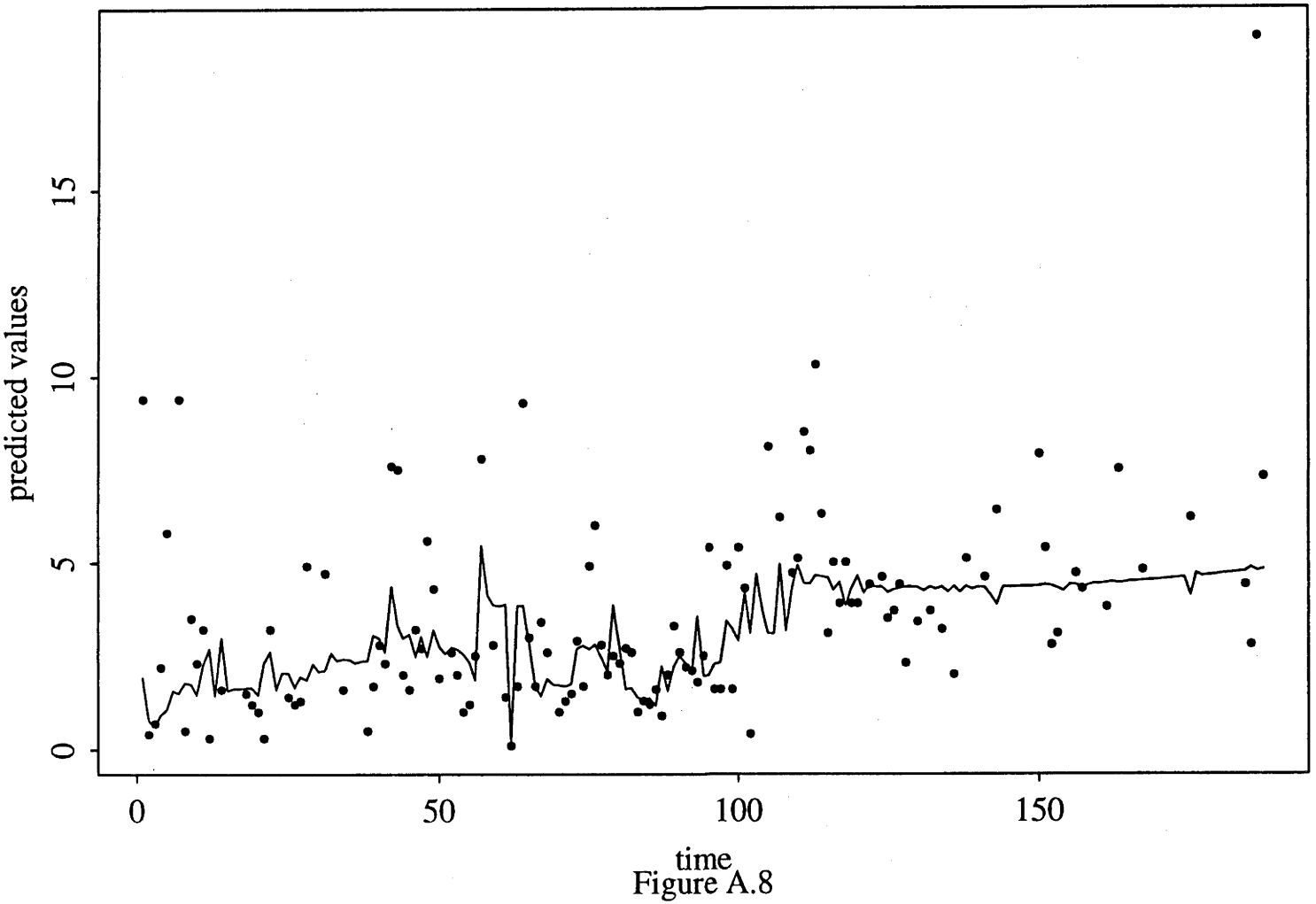

Figure A.8

Appendix B

This appendix presents the proof of the results from chapter 6. We, for convenience, assume the notations from chapter 6.

Proof of Lemma 6.1.2

We note that, given model (6.a),

1. $y_1(n) = C_{1n}A^n x(0) + \sum_{i=1}^{n} C_{1n}A^{n-i}B\varepsilon_{i-1} + \varepsilon_{1n}$ and

2. $\bar{x}(n) = V_{n-1}y_1(n-1) + W_{n-1}\bar{x}(n-1)$

$$= \sum_{i=1}^{n-1} W_{n-1}\cdots W_{n-i+1}V_{n-i}y_1(n-i) + \left(\prod_{i=1}^{n-1} W_i\right)A\mu$$

Then

$$v_1(n) = y_1(n) - C_{1n}\bar{x}(n)$$

$$= C_{1n}\left(A^n - \sum_{k\in\mathcal{O}\cap N_1} W_{n-1}\cdots W_{k+1}V_k C_{1k}A^k\right)x(0) + \varepsilon_{1n}$$

$$+ \sum_{k=1}^{n} C_{1n}A^{n-k}B\varepsilon_{k-1} - C_{1n}\sum_{z=1}^{n-1}\sum_{k\in\mathcal{O}\cap N_z} W_{n-1}\cdots W_{k+1}V_k C_{1k}A^{k-z}B\varepsilon_{z-1}$$

$$- C_{1n}\sum_{k\in\mathcal{O}\cap N_1} W_{n-1}\cdots W_{k+1}V_k\varepsilon_{1k} - C_{1n}\left(\prod_{k=1}^{n-1} W_k\right)A\mu$$

This implies our result.

Proof of Lemma 6.1.4

we note that, given model (6.a),

$$x(n) = \sum_{k=1}^{n} A^{k-1}B\varepsilon_{n-k} + A^n x(0)$$

Therefore,

$$E(x(n)\varepsilon_v' \mid \tilde{v}_0) = \sum_{k=1}^{n} A^{k-1}B\left(I(v=n-k)\Omega - \alpha_{n-k}\Gamma\alpha_v' + \varepsilon_{n-k}^T\varepsilon_v^T\right) + A^n x_0^T\varepsilon_v^T$$

$$= \sum_{k=1}^{n} A^{k-1}B\left(I(v=n-k)\Omega - \alpha_{n-k}\Gamma\alpha_v'\right) + x_n^T\varepsilon_v^T$$

Hence we have the result.

Proof of Lemma 6.2.1

Since $z \sim N(0,1)$ implies that $E(z^3) = 0$ and $E(z^4) = 3$,

$$E\{ee'B'Aee'\} = \left(\sum_{g=1}^{K}\sum_{h=1}^{K} E(e_i e_g B'A_h e_h e_j) \right)_{i,j=1,K}$$

$$= \left(\sum_{g=1}^{K}\sum_{h=1}^{K} B'A_h E(e_i e_g e_h e_j) \right)_{i,j=1,K}$$

$$= \left(I(i=j) \sum_{g=1}^{K} \left\{ I(g=i)3B'A_g + I(g\neq i)B'A_g \right\} \right.$$

$$\left. + I(i\neq j) \sum_{g=1}^{K}\sum_{h=1}^{K} \{B'_i A_j + B'_j A_i\} \right)_{i,j=1,K}$$

and

$$E\{ee'B'Ae\} = \left(\sum_{g=1}^{K}\sum_{h=1}^{K} B'A_h E(e_i e_g e_h) \right)_{i=1,K} = 0$$

This implies the result.

From the proof of lemma 6.1.2, we can see that

$$\varphi_0 = (\mathrm{cov}(x(0),v_1(n)) : n \in \mathcal{O})$$

where

$$\mathrm{cov}(v_1(n),x(0)) = C_{1n}\left\{ A^n - \sum_{k\in\mathcal{O}\cap N_1} W_{n-1}\cdots W_{k+1}V_k C_{1k}A^k \right\}\Sigma$$

Proof of Lemma 6.2.2

we note that

$$x_0^T = \mu + \varphi_0\Gamma^{-1}\tilde{v}_0 \text{ and } x_n^T = \sum_{v=1}^{n} A^{v-1}B\varepsilon_{n-v}^T + A^n x_0^T$$

Writing $K = \tilde{e}_0\tilde{e}_0'$ where $\tilde{e}_0 = \Gamma^{-1/2}\tilde{v}_0$

Case 1: h=0, w=0

$$E(\varepsilon_{r1}^T\varepsilon_{sj}^T x_0^T x_0^{T,}) = (E\{\varepsilon_{r1}^T\varepsilon_{sj}^T x_{01}^T x_{0k}^T\})_{\substack{l=1,m\\k=1,m}}$$

$$= (E\{\varepsilon_{r1}^T\varepsilon_{sj}^T(\mu_l+\varphi_{0l}\Gamma^{-1}\tilde{v}_0)(\mu_k+\varphi_{0k}\Gamma^{-1}\tilde{v}_0)\})_{l,k=1,m}$$

$$= (\alpha_{rl}\Gamma^{-1}\alpha'_{sj}\mu_l\mu_k + \mu_l\alpha_{rl}\Gamma^{-1/2}E(K\Gamma^{-1/2}\alpha'_{sj}\varphi_{0k}\Gamma^{-1/2}\tilde{e}_0)$$

$$+ \mu_k\alpha_{rl}\Gamma^{-1/2}E(K\Gamma^{-1/2}\alpha'_{sj}\varphi_{0l}\Gamma^{-1/2}\tilde{e}_0)$$

$$+ \alpha_{rl}\Gamma^{-1/2}E(K\Gamma^{-1/2}\alpha'_{sj}\varphi_{0l}\Gamma^{-1/2}K)\Gamma^{-1/2}\varphi'_{0k})_{\substack{l=1,m\\k=1,m}}$$

Hence we have the result after applying lemma 6.2.1.

Case 2: h > 0, w=0

$$E\{\varepsilon_{r1}^T\varepsilon_{sj}^T x_h^T x_0^{T,}\} = E\{\varepsilon_{r1}^T\varepsilon_{sj}^T(\sum_{v=1}^{h} A^{v-1}B\varepsilon_{h-v}^T x_0^{T,} + A^h x_0^T x_0^{T,})\}$$

$$= \sum_{v=1}^{h} A^{v-1}B(E(\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v,l}^T x_{0k}^T))_{\substack{l=1,n \\ k=1,m}} + A^h E(\varepsilon_{r1}^T \varepsilon_{sj}^T x_0^T x_0^{T,})$$

where

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v,l}^T x_{0k}^T\} = \alpha_{r1}\Gamma^{-1/2}E(K\Gamma^{-1/2}\alpha'_{sj}\alpha_{h-v,l}\Gamma^{-1/2}K)\Gamma^{-1/2}\varphi'_{0k}$$
$$+ \alpha_{r1}\Gamma^{-1/2}E(K\Gamma^{-1/2}\alpha'_{sj}\alpha_{h-v,l}\Gamma^{-1/2}\tilde{e}_0)\mu_k$$

Hence the result follows after applying lemma 6.2.1.

Case 3: h=0, w > 0

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_0^T x_w^{T,}\} = E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_w^T x_0^{T,}\}'$$

Case 4: h > 0, w > 0

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_h^T x_w^{T,}\} = E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \Big(\sum_{v=1}^{h}\sum_{z=1}^{w} A^{v-1}B\varepsilon_{h-v}^T \varepsilon_{w-s}^{T,} B'A^{z-1,}$$
$$+ \sum_{v=1}^{h} A^{v-1}B\varepsilon_{h-v}^T x_0^{T,}A^{w,} + \sum_{z=1}^{w} A^h x_0^T \varepsilon_{w-z}^{T,} B'A^{z-1,}$$
$$+ A^h x_0^T x_0^{T,}A^{w,}\Big)\}$$

Hence it only requires the derivation of the following expectations:

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v}^T \varepsilon_{w-z}^{T,}\}$$
$$= (E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v,l}^T \varepsilon_{w-z,k}^T\})_{\substack{l=1,n \\ k=1,n}}$$
$$= (\alpha_{r1}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{sj}\alpha_{h-v,l}\Gamma^{-1/2}K\}\Gamma^{-1/2}\alpha'_{w-z,k})_{\substack{l=1,n \\ k=1,n}}$$

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v}^T x_0^{T,}\}$$
$$= (\alpha_{r1}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{sj}\alpha_{h-v,l}\Gamma^{-1/2}K\}\Gamma^{-1/2}\varphi'_{0k})_{\substack{l=1,n \\ k=1,m}}$$

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_0^T \varepsilon_{w-z}^{T,}\} = E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{w-z}^T x_0^{T,}\}'$$

Therefore the result follows.

This completes the proof of lemma 6.2.2.

Proof of Lemma 6.2.3

1. $E\{\mathrm{vec}(\varepsilon_r^T \varepsilon_s^{T,})\} = \mathrm{vec}(E\{\varepsilon_r^T \varepsilon_s^{T,}\}) = \mathrm{vec}(\alpha_r \Gamma^{-1}\alpha'_s) = (\alpha_r \Gamma^{-1}\alpha'_{sl})_{l=1}^n$

2. $E\{\mathrm{vec}(x_r^T \varepsilon_r^{T,})\mathrm{vec}(x_s^T \varepsilon_s^{T,})\}$
$$= E\{(\varepsilon_r^T \varepsilon_s^{T,}) \otimes (x_r^T x_s^{T,})\} = (E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_r^T x_s^{T,}\})_{l,j=1,n}$$

3. $E\{vec(x_r^T \varepsilon_s^{T,})\}$

$$= E\{vec\left((A^r x_0^T + \sum_{z=1}^{r} A^{z-1} B \varepsilon_{r-z}^T)\varepsilon_s^{T,}\right)\}$$

$$= E\{vec(A^r x_0^T \varepsilon_s^{T,}) + \sum_{z=1}^{r} vec(A^{z-1} B \varepsilon_{r-z}^T \varepsilon_s^{T,})\}$$

$$= E\{(I_n \otimes A^r) vec(x_0^T \varepsilon_s^{T,}) + \sum_{z=1}^{r} (I_n \otimes A^{z-1} B) vec(\varepsilon_{r-z}^T \varepsilon_s^{T,})\}$$

$$= (I_n \otimes A^r)(\varphi_0 \Gamma^{-1} \alpha'_{sl})_{l=1}^{n} + \sum_{z=1}^{r} (I_n \otimes A^{z-1} B)(\alpha_{r-z} \Gamma^{-1} \alpha'_{sl})_{l=1}^{n}$$

4. $E\{vec(\varepsilon_s^T x_v^{T,})\}$

$$= E\{vec(\varepsilon_s^T x_0^{T,} A^{v,}) + \sum_{z=1}^{v} (A^{z-1} B \otimes I_n) vec(\varepsilon_s^T \varepsilon_{v-z}^{T,})\}$$

$$= (A^v \otimes I_n)(\alpha_s \Gamma^{-1} \varphi'_{0l})_{l=1}^{n} + \sum_{z=1}^{v} (A^{z-1} B \otimes I_n)(\alpha_s \Gamma^{-1} \alpha'_{v-z,l})_{l=1}^{n}$$

5. $E\{vec(\varepsilon_r^T \varepsilon_v^{T,}) vec(\varepsilon_s^T \varepsilon_z^{T,})'\}$

$$= E\{(\varepsilon_v^T \varepsilon_z^{T,}) \otimes (\varepsilon_r^T \varepsilon_s^{T,})\}$$

$$= (E\{\varepsilon_{vl}^T \varepsilon_{zj}^T \varepsilon_r^T \varepsilon_s^{T,}\})_{\substack{l=1,n \\ j=1,n}}$$

$$E\{\varepsilon_{vl}^T \varepsilon_{zj}^T \varepsilon_r^T \varepsilon_s^{T,}\} = (E\{\varepsilon_{vl}^T \varepsilon_{zj}^T \varepsilon_{rl}^T \varepsilon_{sk}^T\})_{\substack{l=1,n \\ k=1,n}}$$

$$= (\alpha_{vl} \Gamma^{-1/2} E\{K \Gamma^{-1/2} \alpha'_{zj} \alpha_{rl} \Gamma^{-1/2} K\} \Gamma^{-1/2} \alpha'_{sk})_{\substack{l=1,n \\ k=1,n}}$$

Hence we have the result.

6. $E\{x_{vl}^T x_{wj}^T \varepsilon_r^T \varepsilon_s^{T,}\} = E\{(x_{vl}^T x_{wj}^T \varepsilon_{rl}^T \varepsilon_{sk}^T)_{\substack{l=1,n \\ k=1,n}}\}$

$$= \left((E\{\varepsilon_{rl}^T \varepsilon_{sk}^T x_v^T x_w^{T,}\})_{ij}\right)_{\substack{l=1,n \\ k=1,n}}$$

7. $E\{vec(\varepsilon_r^T x_v^{T,}) vec(\varepsilon_s^T x_w^{T,})'\}$

$$= E\{(x_v^T x_w^{T,}) \otimes (\varepsilon_r^T \varepsilon_s^{T,})\}$$

$$= (E\{x_{vl}^T x_{wj}^T \varepsilon_r^T \varepsilon_s^{T,}\})_{\substack{l=1,m \\ j=1,m}}$$

8. $E\{vec(Z_r \varepsilon_r^{T,}) vec(x_h^T \varepsilon_s^{T,})'\}$

$$= E\{(\varepsilon_r^T \varepsilon_s^{T,}) \otimes (Z_r x_h^{T,})\}$$

$$= (Z_r E\{\varepsilon_{rl}^T \varepsilon_{sj}^T x_h^T\}')_{\substack{l=1,n \\ j=1,n}}$$

$$E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_h^T\} = E\{\varepsilon_{r1}^T \varepsilon_{sj}^T (A^h x_0^T + \sum_{v=1}^{h} A^{v-1} B \varepsilon_{h-v}^T)\}$$

$$= A^h E\{\varepsilon_{r1}^T \varepsilon_{sj}^T x_0^T\} \text{ since } E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_{h-v,k}^T\}=0 \text{ for any } k$$

$$= (\alpha_{r1} \Gamma^{-1} \alpha'_{sj}) A^h \mu$$

9. $E\{\text{vec}(Z_r \varepsilon_r^{T,})\} = (I_n \otimes Z_r) E\{\text{vec}(\varepsilon_r^{T,})\} = (I_n \otimes Z_r) E\{\varepsilon_r^T\} = 0$

$$E\{\text{vec}(Z_r \varepsilon_r^{T,}) \text{vec}(\varepsilon_h^T \varepsilon_s^{T,})'\}$$

$$= E\{(\varepsilon_r^T \varepsilon_s^{T,}) \otimes (Z_r \varepsilon_h^{T,})\}$$

$$= (Z_r E\{\varepsilon_{r1}^T \varepsilon_{sj}^T \varepsilon_h^T\}')_{\substack{l=1,n \\ j=1,n}}$$

$$= 0$$

10. $E\{\text{vec}(Z_r \varepsilon_r^{T,}) \text{vec}(\varepsilon_s^T x_v^{T,})'\}$

$$= E\{(\varepsilon_r^T x_v^{T,}) \otimes (Z_r \otimes \varepsilon_s^{T,})\}$$

$$= (Z_r E\{\varepsilon_{r1}^T x_{vj}^T \varepsilon_s^T\}')_{\substack{l=1,n \\ j=1,m}}$$

$$E\{\varepsilon_{r1}^T x_{vj}^T \varepsilon_s^T\} = (E\{\varepsilon_{r1}^T x_{vj}^T \varepsilon_{sk}^T\})_{k=1,n}$$

$$= \left((E\{\varepsilon_r^T x_v^{T,} \varepsilon_{sk}^T\})_{ij}\right)_{k=1,n}$$

$$E\{\varepsilon_r^T x_v^{T,} \varepsilon_{sk}^T\} = E\{\varepsilon_{sk}^T \varepsilon_r^T (A^v x_0^T + \sum_{h=1}^{v} A^{h-1} B \varepsilon_{v-h}^T)'\}$$

$$= E\{\varepsilon_{sk}^T \varepsilon_r^T x_0^{T,}\} A^{v,}$$

$$= (\alpha_{sk} \Gamma^{-1} \alpha'_{rp})_{p=1,n} \mu' A^{v,}$$

11. $E\{\text{vec}(x_r^T \varepsilon_w^{T,}) \text{vec}(\varepsilon_s^T \varepsilon_v^{T,})'\}$

$$= E\{(\varepsilon_w^T \varepsilon_v^{T,}) \otimes (x_r^T \varepsilon_s^{T,})\}$$

$$= (E\{\varepsilon_{wi}^T \varepsilon_{vj}^T x_r^T \varepsilon_s^{T,}\})_{\substack{i=1,n \\ j=1,n}}$$

$$E\{\varepsilon_{wi}^T \varepsilon_{vj}^T x_r^T \varepsilon_s^{T,}\}$$

$$= E\{\varepsilon_{wi}^T \varepsilon_{vj}^T (A^r x_0^T + \sum_{z=1}^{r} A^{z-1} B \varepsilon_{r-z}^T) \varepsilon_s^{T,}\}$$

$$= A^r (E\{\varepsilon_{wi}^T \varepsilon_{vj}^T x_{0l}^T \varepsilon_{sk}^T\})_{\substack{l=1,m \\ k=1,n}}$$

$$+ \sum_{z=1}^{r} A^{z-1} B (E\{\varepsilon_{wi}^T \varepsilon_{vj}^T \varepsilon_{r-z,l}^T \varepsilon_{sk}^T\})_{\substack{l=1,n \\ k=1,n}}$$

$$= A^r (\alpha_{wi} \Gamma^{-1/2} E\{K \Gamma^{-1/2} \alpha'_{vj} \alpha_{sk} \Gamma^{-1/2} K\} \Gamma^{-1/2} \varphi'_{0l})_{\substack{l=1,m \\ k=1,n}}$$

$$+\sum_{z=1}^{r} A^{z-1}B(\alpha_{wl}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{vj}\alpha_{r-z,l}\Gamma^{-1/2}K\}\Gamma^{-1/2}\alpha'_{sk})_{\substack{l=1,n\\k=1,n}}$$

12. $E\{vec(x_h^T\varepsilon_r^{T,})vec(\varepsilon_s^T x_k^{T,})'\}$

$$= E\{(\varepsilon_r^T x_k^{T,})\otimes(x_h^T\varepsilon_s^{T,})\}$$

$$= (E\{\varepsilon_{rl}^T x_{kj}^T(A^h x_0^T + \sum_{z=1}^{h} A^{z-1}B\varepsilon_{h-z}^T)\varepsilon_s^{T,}\})_{\substack{l=1,n\\j=1,m}}$$

$$= (A^h E\{\varepsilon_{rl}^T x_{kj}^T x_0^T\varepsilon_s^{T,}\} + \sum_{z=1}^{h} A^{z-1}BE\{\varepsilon_{rl}^T x_{kj}^T\varepsilon_{h-z}^T\varepsilon_s^{T,}\})_{\substack{l=1,n\\j=1.m}}$$

(i) $E\{\varepsilon_{rl}^T x_{kj}^T x_0^T\varepsilon_s^{T,}\}$

$$= \left((E\{\varepsilon_r^T x_k^{T,}x_{0l}^T\varepsilon_{sd}^T\})_{lj}\right)_{\substack{l=1,m\\d=1,n}}$$

$E\{\varepsilon_r^T x_k^{T,}x_{0l}^T\varepsilon_{sd}^T\}$

$$= E\{\varepsilon_r^T(A^k x_0^T + \sum_{z=1}^{k} A^{z-1}B\varepsilon_{k-z}^T)'x_{0l}^T\varepsilon_{sd}^T\}$$

$$= (E\{\varepsilon_{rp}^T x_{0q}^T x_{0l}^T\varepsilon_{sd}^T\})_{\substack{p=1,n\\q=1,m}}A^{k,}$$
$$+\sum_{z=1}^{k}(E\{\varepsilon_{rp}^T\varepsilon_{k-z,q}^T x_{0l}^T\varepsilon_{sd}^T\})_{\substack{p=1,n\\q=1,n}}B'A^{z-1,}$$

$E\{\varepsilon_{rp}^T x_{0q}^T x_{0l}^T\varepsilon_{sd}^T\}$

$$= \alpha_{rp}\Gamma^{-1}\alpha'_{sd}\mu_q\mu_l + \varphi_{0q}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{rp}\alpha_{sd}\Gamma^{-1/2}K\}\Gamma^{-1/2}\varphi'_{0l}$$

$E\{\varepsilon_{rp}^T\varepsilon_{k-z,q}^T x_{0l}^T\varepsilon_{sd}^T\}$

$$= \alpha_{rp}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{k-z,q}\alpha_{sd}\Gamma^{-1/2}K\}\Gamma^{-1/2}\varphi'_{0l}$$

(ii) $E\{\varepsilon_{rl}^T x_{kj}^T\varepsilon_{h-z}^T\varepsilon_s^{T,}\} = \left((E\{\varepsilon_r^T x_k^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\})_{lj}\right)_{\substack{l=1,n\\d=1,n}}$

$E\{\varepsilon_r^T x_k^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\}$

$$= E\{\varepsilon_r^T x_0^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\}A^{k,} + \sum_{a=1}^{k} E\{\varepsilon_r^T\varepsilon_{k-a}^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\}B'A^{a-1,}$$

$E\{\varepsilon_r^T x_0^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\}$

$$= (\varphi_{0q}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{rp}\alpha_{h-z,l}\Gamma^{-1/2}K\}\Gamma^{-1/2}\alpha'_{sd})_{\substack{p=1,n\\q=1,m}}$$

$E\{\varepsilon_r^T\varepsilon_{k-a}^{T,}\varepsilon_{h-z,l}^T\varepsilon_{sd}^T\}$

$$= (\alpha_{rp}\Gamma^{-1/2}E\{K\Gamma^{-1/2}\alpha'_{k-a,q}\alpha_{h-z,l}\Gamma^{-1/2}K\}\Gamma^{-1/2}\alpha'_{sd})_{\substack{p=1,n\\q=1,n}}$$

13. $E\{vec(\varepsilon_h^T \varepsilon_r^{T,})vec(\varepsilon_s^T x_k^{T,})'\}$

$$= E\{(\varepsilon_r^T x_k^{T,})\otimes(\varepsilon_h^T \varepsilon_s^{T,})\}$$

$$= (E\{\varepsilon_{rl}^T x_{kj}^T \varepsilon_h^T \varepsilon_s^{T,}\})_{\substack{l=1,n \\ j=1,m}}$$

14. trivial.

This completes the proof of lemma 6.2.3.

Derivation of the Fisher Information Matrix

The followings present the details of the derivation of the information matrix:

$$E\{vec\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\left(\sum_{r=1}^{T} Z_r \varepsilon_r^{T,}\right)\Omega^{-1}\right)vec\left(\left(\sum_{s=1}^{T} Z_s \varepsilon_s^{T,}\right)\Omega^{-1}\right)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar})\left(\sum_{r=1}^{T}\sum_{s=1}^{T}(I_n \otimes Z_r)E\{\varepsilon_r^T \varepsilon_s^{T,}\}(I_n \otimes Z'_s)\right)(\Omega^{-1}\otimes I_{nvar})$$

$$= (\Omega^{-1}\otimes I_{nvar})\left(\sum_{r=1}^{T}\sum_{s=1}^{T}(I_n \otimes Z_r)\alpha_r \Gamma^{-1}\alpha'_s(I_n \otimes Z'_s)\right)(\Omega^{-1}\otimes I_{nvar})$$

$$E\{vec\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\left(\sum_{r=1}^{T}(x_r^T \varepsilon_r^{T,}+\gamma_{r,r}^T)\right)\Omega^{-1}\right)vec\left(\left(\sum_{s=1}^{T}(x_s^T \varepsilon_s^{T,}+\gamma_{s,s}^T)\right)\Omega^{-1}\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\left(\sum_{r=1}^{T}\sum_{s=1}^{T}E\left(vec(x_r^T \varepsilon_r^{T,}+\gamma_{r,r}^T)vec(x_s^T \varepsilon_s^{T,}+\gamma_{s,s}^T)'\right)\right)(\Omega^{-1}\otimes I_m)$$

$$E\{vec\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(T\Omega - \sum_{r=1}^{T}\left(\varepsilon_r^T \varepsilon_r^{T,}+ \Omega-\alpha_r \Gamma^{-1}\alpha'_r\right)\right)vec\left(T\Omega - \sum_{s=1}^{T}\left(\varepsilon_s^T \varepsilon_s^{T,}+ \Omega-\alpha_s \Gamma^{-1}\alpha'_s\right)\right)'\}$$

$$= T^2 vec(\Omega)vec(\Omega)' - \sum_{s=1}^{T} Tvec(\Omega)\left(E\{vec(\varepsilon_s^T \varepsilon_s^{T,})\}+vec(\Omega-\alpha_s \Gamma^{-1}\alpha'_s)\right)'$$

$$- \sum_{r=1}^{T} T\left(E\{vec(\varepsilon_r^T \varepsilon_r^{T,})\}+vec(\Omega-\alpha_r \Gamma^{-1}\alpha'_r)\right)vec(\Omega)'$$

$$+ \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\left(vec(\varepsilon_r^T \varepsilon_r^{T,})+vec(\Omega-\alpha_r \Gamma^{-1}\alpha'_r)\right)\left(vec(\varepsilon_s^T \varepsilon_s^{T,})+vec(\Omega-\alpha_s \Gamma^{-1}\alpha'_s)\right)'\}$$

$$E\{4vec\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\sum_{r=1}^{T}\left\{E\{\frac{\partial x'_{r-1}}{\partial A}|\tilde{v}_0\}(I_m \otimes A'C'\Omega^{-1}\varepsilon_r^T)+C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right\}\right)$$

$$\text{vec}\left(\sum_{s=1}^{T}\left\{E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right\}\right)'\}$$

$$= E\{\sum_{r=1}^{T}\sum_{s=1}^{T}\text{vec}\left(E\{\frac{\partial x'_{r-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_r^T)+C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)$$

$$\text{vec}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

From the recursive algorithm of $E\{\frac{\partial x'_t}{\partial A}|\tilde{v}_0\}$, we understand that

$$E\{\frac{\partial x'_t}{\partial A}|\tilde{v}_0\} = \sum_{r=0}^{t-1}(I_m\otimes x_r^{T,})S_{mm}(I_m\otimes A^{t-1-r,})$$

where $S_{mm} = \sum_{r=1}^{T}\sum_{s=1}^{T}(E_{rs}\otimes E'_{rs})$

Moreover, we can easily derive that

$$(I_m\otimes a')S_{mm}(I_m\otimes c) = ca'$$

where $a = (a_1,\ldots,a_m)'$ and $c = (c_1,\ldots,c_m)'$.

Therefore,

$$E\{\text{vec}\left(E\{\frac{\partial x'_{r-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_r^T)\right)\text{vec}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)'\}$$

$$= E\{\sum_{v=0}^{r-2}\sum_{w=0}^{s-2}\left(x_v^T x_w^{T,}\right)\otimes\left(A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T\varepsilon_s^{T,}\Omega^{-1}CA^{s-1-w}\right)\}$$

$$= \sum_{v=0}^{r-2}\sum_{w=0}^{s-2}\left(E\{x_{vl}^T x_{wj}^T A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T\varepsilon_s^{T,}\Omega^{-1}CA^{s-1-w}\}\right)_{l,j=1,m}$$

$$E\{\text{vec}\left(E\{\frac{\partial x'_{r-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_r^T)\right)\text{vec}\left(C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

$$= E\{\text{vec}\left(\sum_{v=0}^{r-2}A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T x_v^{T,}\right)\text{vec}\left(C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

$$= \sum_{v=0}^{r-2}(I_m\otimes A^{r-1-v'}C'\Omega^{-1})\Big\{E\{\text{vec}(\varepsilon_r^T x_v^{T,})\text{vec}(\varepsilon_s^T x_{s-1}^{T,})'\}(I_m\otimes\Omega^{-1}C)$$

$$+ E\{\text{vec}(\varepsilon_r^T x_v^{T,})\}\text{vec}(C'\Omega^{-1}\gamma_{s-1,s}^{T,})'\Big\}$$

$$E\{\text{vec}\left(C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)\text{vec}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)'\}$$

$$= E\{\text{vec}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)\text{vec}\left(C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)'\},$$

$$E\{\text{vec}\left(C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)\text{vec}\left(C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

$$= E\{\mathrm{vec}\left(C'\Omega^{-1}\varepsilon_r^T x_{r-1}^{T,}\right)\mathrm{vec}\left(C'\Omega^{-1}\varepsilon_s^T x_{s-1}^{T,}\right)'$$
$$+ \mathrm{vec}\left(C'\Omega^{-1}\varepsilon_r^T x_{r-1}^{T,}\right)\mathrm{vec}\left(C'\Omega^{-1}\gamma_{s-1,s}^{T,}\right)'$$
$$+ \mathrm{vec}\left(C'\Omega^{-1}\gamma_{r-1,r}^{T,}\right)\mathrm{vec}\left(C'\Omega^{-1}\varepsilon_s^T x_{s-1}^{T,}\right)'$$
$$+ \mathrm{vec}\left(C'\Omega^{-1}\gamma_{r-1,r}^{T,}\right)\mathrm{vec}\left(C'\Omega^{-1}\gamma_{s-1,s}^{T,}\right)'\}$$
$$= (I_m\otimes C'\Omega^{-1})E\{\mathrm{vec}(\varepsilon_r^T x_{r-1}^{T,})\mathrm{vec}(\varepsilon_s^T x_{s-1}^{T,})'\}(I_m\otimes\Omega^{-1}C)$$
$$+ (I_m\otimes C'\Omega^{-1})E\{\mathrm{vec}(\varepsilon_r^T x_{r-1}^{T,})\}\mathrm{vec}(C'\Omega^{-1}\gamma_{s-1,s}^{T,})'$$
$$+ \mathrm{vec}(C'\Omega^{-1}\gamma_{r-1,r}^{T,})E\{\mathrm{vec}(\varepsilon_s^T x_{s-1}^{T,})\}'(I_m\otimes\Omega^{-1}C)$$
$$+ \mathrm{vec}(C'\Omega^{-1}\gamma_{r-1,r}^{T,})\mathrm{vec}(C'\Omega^{-1}\gamma_{s-1,s}^{T,})'$$

$$E\{4\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)'\}$$
$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\Big\{\mathrm{vec}\left(E\{\frac{\partial x_{r-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_r^T)+C'\Omega^{-1}(\varepsilon_r^T\varepsilon_{r-1}^{T,}-\alpha_r\Gamma^{-1}\alpha_{r-1}')\right)$$
$$\mathrm{vec}\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)'\Big\}$$

Similarly, the recursive algorithm of $E\{\frac{\partial x_t'}{\partial B}|\tilde{v}_0\}$ suggests that

$$E\{\frac{\partial x_t'}{\partial B}|\tilde{v}_0\} = \sum_{r=0}^{t-1}(I_m\otimes\varepsilon_r^{T,})U_{mn}(I_n\otimes A^{t-1-r,})$$

where $U_{mn} = \sum_{r=1}^{m}\sum_{s=1}^{n} U_{rs}\otimes U_{rs}'$.

Therefore,

$$E\{\mathrm{vec}\left(E\{\frac{\partial x_{r-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_r^T)\right)\mathrm{vec}\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)'\}$$
$$= \sum_{v=0}^{r-2}\sum_{z=0}^{s-2} E\{\mathrm{vec}\left(A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T\varepsilon_v^{T,}\right)\mathrm{vec}\left(A^{s-1-z'}C'\Omega^{-1}\varepsilon_s^T\varepsilon_z^{T,}\right)'\}$$
$$= \sum_{v=0}^{r-2}\sum_{z=0}^{s-2}(I_n\otimes A^{r-1-v'}C'\Omega^{-1})E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_v^{T,})\mathrm{vec}(\varepsilon_s^T\varepsilon_z^{T,})'\}(I_n\otimes\Omega^{-1}CA^{s-1-z})$$
$$E\{\mathrm{vec}\left(E\{\frac{\partial x_{r-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_r^T)\right)\mathrm{vec}\left(C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)'\}$$
$$= \sum_{v=0}^{r-2}\mathrm{vec}(A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T\varepsilon_v^{T,})\left(\mathrm{vec}(C'\Omega^{-1}\varepsilon_s^T\varepsilon_{s-1}^{T,})'-\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}')'\right)$$
$$= \sum_{v=0}^{r-2}(I_n\otimes A^{r-1-v'}C'\Omega^{-1})\left(E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_v^{T,})\mathrm{vec}(\varepsilon_s^T\varepsilon_{s-1}^{T,})'\}(I_n\otimes\Omega^{-1}C)\right.$$
$$\left. - E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_v^{T,})\}\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}')'\right)$$

$$E\{vec\left(C'\Omega^{-1}(\varepsilon_r^T\varepsilon_{r-1}^{T,}-\alpha_r\Gamma^{-1}\alpha_{r-1}')\right)vec\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)'\}$$

$$= E\{vec\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right)vec\left(C'\Omega^{-1}(\varepsilon_r^T\varepsilon_{r-1}^{T,}-\alpha_r\Gamma^{-1}\alpha_{r-1}')\right)'\}'$$

$$E\{vec\left(C'\Omega^{-1}(\varepsilon_r^T\varepsilon_{r-1}^{T,}-\alpha_r\Gamma^{-1}\alpha_{r-1}')\right)vec\left(C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)'\}$$

$$= E\Big\{vec\left(C'\Omega^{-1}\varepsilon_r^T\varepsilon_{r-1}^{T,}\right)vec\left(C'\Omega^{-1}\varepsilon_s^T\varepsilon_{s-1}^{T,}\right)'$$

$$- vec\left(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha_{r-1}'\right)vec\left(C'\Omega^{-1}\varepsilon_s^T\varepsilon_{s-1}^{T,}\right)'$$

$$- vec\left(C'\Omega^{-1}\varepsilon_r^T\varepsilon_{r-1}^{T,}\right)vec\left(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}'\right)'$$

$$+ vec\left(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha_{r-1}'\right)vec\left(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}'\right)'\Big\}$$

$$= (I_n\otimes C'\Omega^{-1})E\{vec\left(\varepsilon_r^T\varepsilon_{r-1}^{T,}\right)vec\left(\varepsilon_s^T\varepsilon_{s-1}^{T,}\right)'\}(I_n\otimes\Omega^{-1}C)$$

$$- vec\left(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha_{r-1}'\right)E\{vec(\varepsilon_s^T\varepsilon_{s-1}^{T,})\}'(I_n\otimes\Omega^{-1}C)$$

$$- (I_n\otimes C'\Omega^{-1})E\{vec(\varepsilon_r^T\varepsilon_{r-1}^{T,})\}vec\left(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}'\right)'$$

$$+ vec\left(C'\Omega^{-1}\alpha_r\Gamma^{-1}\alpha_{r-1}'\right)vec\left(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}'\right)'$$

$$E\{vec\left(E\{\frac{\partial logL}{\partial\beta'}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial logL}{\partial C'}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\left(\sum_{r=1}^{T}Z_r\varepsilon_r^{T,}\right)\Omega^{-1}\right)vec\left(\sum_{s=1}^{T}\left(x_s^T\varepsilon_s^{T,}+\gamma_{s,s}^T\right)\Omega^{-1}\right)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar})\sum_{r=1}^{T}\sum_{s=1}^{T}\Big\{E\{vec\left(Z_r\varepsilon_r^{T,}\right)vec\left(x_s^T\varepsilon_s^{T,}\right)'\}$$

$$+E\{vec\left(Z_r\varepsilon_r^{T,}\right)\}vec(\gamma_{s,s}^T)'\Big\}(\Omega^{-1}\otimes I_m)$$

$$E\{vec\left(E\{\frac{\partial logL}{\partial\beta'}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial logL}{\partial\Omega^{-1}}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\left(\sum_{r=1}^{T}Z_r\varepsilon_r^{T,}\right)\Omega^{-1}\right)vec\left(T\Omega-\sum_{s=1}^{T}\left(\varepsilon_s^T\varepsilon_s^{T,}+\Omega-\alpha_s\Gamma^{-1}\alpha_s'\right)\right)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar})E\{vec\left(\sum_{r=1}^{T}Z_r\varepsilon_r^{T,}\right)\left(Tvec(\Omega)'-\sum_{s=1}^{T}vec(\varepsilon_s^T\varepsilon_s^{T,}+\Omega-\alpha_s\Gamma^{-1}\alpha_s')'\right)\}$$

$$= 0$$

$$E\{-2vec\left(E\{\frac{\partial logL}{\partial\beta'}|\tilde{v}_0\}\right)vec\left(E\{\frac{\partial logL}{\partial A}|\tilde{v}_0\}\right)'\}$$

$$= E\{vec\left(\left(\sum_{r=1}^{T}Z_r\varepsilon_r^{T,}\right)\Omega^{-1}\right)vec\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x_{s-1}'}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T\right)\right.$$

$$+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,} + \gamma_{s-1,s}^{T,})\Big)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar}) \sum_{r=1}^{T}\sum_{s=1}^{T} E\{vec(Z_r\varepsilon_r^{T,})\Big(\sum_{v=0}^{s-2} vec(A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T x_v^{T,})$$
$$+ vec(C'\Omega^{-1}\varepsilon_s^T x_{s-1}^{T,}) + vec(C'\Omega^{-1}\gamma_{s-1,s}^{T,})\Big)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar}) \sum_{r=1}^{T}\sum_{s=1}^{T} E\{vec(Z_r\varepsilon_r^{T,})\Big(\sum_{v=0}^{s-2} (I_m\otimes A^{s-1-v'}C'\Omega^{-1})vec(\varepsilon_s^T x_v^{T,})$$
$$+ (I_m\otimes C'\Omega^{-1})vec(\varepsilon_s^T x_{s-1}^{T,}) + vec(C'\Omega^{-1}\gamma_{s-1,s}^{T,})\Big)'\}$$

$$E\{-2vec\Big(E\{\frac{\partial \log L}{\partial \beta'}|\tilde{v}_0\}\Big)vec\Big(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\Big)'\}$$

$$= E\{vec\Big(\Big(\sum_{r=1}^{T} Z_r\varepsilon_r^{T,}\Big)\Omega^{-1}\Big)vec\Big(\sum_{s=1}^{T}\Big(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)$$
$$+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,} - \alpha_s\Gamma^{-1}\alpha_{s-1}')\Big)\Big)'\}$$

$$= (\Omega^{-1}\otimes I_{nvar}) \sum_{r=1}^{T}\sum_{s=1}^{T} E\{vec(Z_r\varepsilon_r^{T,})\Big(\sum_{v=0}^{s-2} vec(A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T\varepsilon_v^{T,})'$$
$$+ vec\Big(C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,} - \alpha_s\Gamma^{-1}\alpha_{s-1}')\Big)'\Big)\}$$

$$= (\Omega^{-1}\otimes I_{nvar}) \sum_{r=1}^{T}\sum_{s=1}^{T} E\Big\{vec(Z_r\varepsilon_r^{T,})\Big(\sum_{v=0}^{s-2} vec(\varepsilon_s^T\varepsilon_v^{T,})'(I_n\otimes\Omega^{-1}CA^{s-1-v})$$
$$+ vec(\varepsilon_s^T\varepsilon_{s-1}^{T,})'(I_n\otimes\Omega^{-1}C) - vec(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}')'\Big)\Big\}$$

$$E\{vec\Big(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\Big)vec\Big(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\Big)'\}$$

$$= E\{vec\Big(\sum_{r=1}^{T}(x_r^T\varepsilon_r^{T,} + \gamma_{r,r}^T)\Omega^{-1}\Big)vec\Big(T\Omega - \sum_{s=1}^{T}(\varepsilon_s^T\varepsilon_s^{T,} + \Omega - \alpha_s\Gamma^{-1}\alpha_s')\Big)'\}$$

$$= (\Omega^{-1}\otimes I_m)E\Big\{vec\Big(\sum_{r=1}^{T}(x_r^T\varepsilon_r^{T,} + \gamma_{r,r}^T)\Big)\Big(Tvec(\Omega)'$$
$$-vec\Big(\sum_{s=1}^{T}(\varepsilon_s^T\varepsilon_s^{T,} + \Omega - \alpha_s\Gamma^{-1}\alpha_s')\Big)'\Big)\Big\}$$

$$= (\Omega^{-1}\otimes I_m)\Big\{T\sum_{r=1}^{T}\Big(E\{vec(x_r^T\varepsilon_r^{T,})\} + vec(\gamma_{r,r}^T)\Big)vec(\Omega)'$$
$$- \sum_{r=1}^{T}\sum_{s=1}^{T}\Big(E\{vec(x_r^T\varepsilon_r^{T,})vec(\varepsilon_s^T\varepsilon_s^{T,})'\} + E\{vec(x_r^T\varepsilon_r^{T,})\}vec(\Omega - \alpha_s\Gamma^{-1}\alpha_s')'$$
$$+vec(\gamma_{r,r}^T)E\{vec(\varepsilon_s^T\varepsilon_s^{T,})\}' + vec(\gamma_{r,r}^T)vec(\Omega - \alpha_s\Gamma^{-1}\alpha_s')\Big)\Big\}$$

$$E\{-2\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)'\}$$

$$= E\{\mathrm{vec}\left(\sum_{r=1}^{T}(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T)\Omega^{-1}\right)\mathrm{vec}\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right.\right.$$
$$\left.\left.+\ C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T)\mathrm{vec}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right.$$
$$\left.+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T)\left(\sum_{v=0}^{s-2}\mathrm{vec}(A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T x_v^{T,})\right.$$
$$\left.+\mathrm{vec}\left(C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T)\left(\sum_{v=0}^{s-2}\mathrm{vec}(\varepsilon_s^T x_v^{T,})'(I_m\otimes\Omega^{-1}CA^{s-1-v})\right.$$
$$\left.+\mathrm{vec}(\varepsilon_s^T x_{s-1}^{T,})'(I_m\otimes\Omega^{-1}C)+\mathrm{vec}(C'\Omega^{-1}\gamma_{s-1,s}^{T,})'\right)\}$$

$$E\{-2\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial C'}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)'\}$$

$$= E\{\mathrm{vec}\left(\sum_{r=1}^{T}(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T)\Omega^{-1}\right)\mathrm{vec}\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x'_{s-1}}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right.\right.$$
$$\left.\left.+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha'_{s-1})\right)\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}\left(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T\right)\mathrm{vec}\left(\sum_{v=0}^{s-2}A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T\varepsilon_v^{T,}\right.$$
$$\left.+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha'_{s-1})\right)'\}$$

$$= (\Omega^{-1}\otimes I_m)\sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}\left(x_r^T\varepsilon_r^{T,}+\gamma_{r,r}^T\right)\left(\sum_{v=0}^{s-2}\mathrm{vec}(\varepsilon_s^T\varepsilon_v^{T,})'(I_n\otimes\Omega^{-1}CA^{s-1-v})\right.$$
$$\left.+\mathrm{vec}(\varepsilon_s^T\varepsilon_{s-1}^{T,})'(I_n\otimes\Omega^{-1}C)-\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha'_{s-1})'\right)\}$$

$$E\{2\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)'\}$$

$$= E\{\mathrm{vec}\left(\sum_{r=1}^{T}(\varepsilon_r^T\varepsilon_r^{T,}+\Omega-\alpha_r\Gamma^{-1}\alpha'_r)-T\Omega\right)\mathrm{vec}\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x'_{s-1}}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right.\right.$$
$$\left.\left.+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T}E\{\mathrm{vec}\left(\varepsilon_r^T\varepsilon_r^{T,}-\alpha_r\Gamma^{-1}\alpha'_r\right)\mathrm{vec}\left(\sum_{v=0}^{s-2}A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T x_v^{T,}\right.$$
$$\left.+C'\Omega^{-1}(\varepsilon_s^T x_{s-1}^{T,}+\gamma_{s-1,s}^{T,})\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\mathrm{vec}\left(\varepsilon_r^T\varepsilon_r^{T,}-\alpha_r\Gamma^{-1}\alpha_r'\right)\left(\sum_{v=0}^{s-2}\mathrm{vec}(\varepsilon_s^T x_v^{T,})'(I_m\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$\left.+\mathrm{vec}(\varepsilon_s^T x_{s-1}^{T,})'(I_m\otimes\Omega^{-1}C)+\mathrm{vec}(C'\Omega^{-1}\gamma_{s-1,s}^{T,})'\right)\}$$

* note that $E\{\mathrm{vec}(\varepsilon_r^T\varepsilon_r^{T,})\} = \mathrm{vec}(\alpha_r\Gamma^{-1}\alpha_r')$

$$E\{2\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial \Omega^{-1}}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)'\}$$

$$= E\{\mathrm{vec}\left(\sum_{r=1}^{T}\left(\varepsilon_r^T\varepsilon_r^{T,}+\Omega-\alpha_r\Gamma^{-1}\alpha_r'\right)-T\Omega\right)\mathrm{vec}\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)\right.\right.$$

$$\left.\left.+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\mathrm{vec}\left(\varepsilon_r^T\varepsilon_r^{T,}-\alpha_r\Gamma^{-1}\alpha_r'\right)\mathrm{vec}\left(\sum_{v=0}^{s-2}A^{s-1-v'}C'\Omega^{-1}\varepsilon_s^T\varepsilon_v^{T,}\right.$$

$$\left.+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\mathrm{vec}\left(\varepsilon_r^T\varepsilon_r^{T,}-\alpha_r\Gamma^{-1}\alpha_r'\right)\left(\sum_{v=0}^{s-2}\mathrm{vec}(\varepsilon_s^T\varepsilon_v^{T,})'(I_n\otimes\Omega^{-1}CA^{s-1-v})\right.$$

$$\left.+\mathrm{vec}(\varepsilon_s^T\varepsilon_{s-1}^{T,})'(I_n\otimes\Omega^{-1}C)-\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}')'\right)\}$$

$$E\{4\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial A}|\tilde{v}_0\}\right)\mathrm{vec}\left(E\{\frac{\partial \log L}{\partial B}|\tilde{v}_0\}\right)'\}$$

$$= E\{\mathrm{vec}\left(\sum_{r=1}^{T}\left(E\{\frac{\partial x_{r-1}'}{\partial A}|\tilde{v}_0\}(I_m\otimes A'C'\Omega^{-1}\varepsilon_r^T)+C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)\right)$$

$$\mathrm{vec}\left(\sum_{s=1}^{T}\left(E\{\frac{\partial x_{s-1}'}{\partial B}|\tilde{v}_0\}(I_n\otimes A'C'\Omega^{-1}\varepsilon_s^T)+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\mathrm{vec}\left(\sum_{v=0}^{r-2}A^{r-1-v'}C'\Omega^{-1}\varepsilon_r^T x_v^{T,}+C'\Omega^{-1}(\varepsilon_r^T x_{r-1}^{T,}+\gamma_{r-1,r}^{T,})\right)$$

$$\mathrm{vec}\left(\sum_{w=0}^{s-2}A^{s-1-w'}C'\Omega^{-1}\varepsilon_s^T\varepsilon_w^{T,}+C'\Omega^{-1}(\varepsilon_s^T\varepsilon_{s-1}^{T,}-\alpha_s\Gamma^{-1}\alpha_{s-1}')\right)'\}$$

$$= \sum_{r=1}^{T}\sum_{s=1}^{T} E\{\left(\sum_{v=0}^{r-2}(I_m\otimes A^{r-1-v'}C'\Omega^{-1})\mathrm{vec}(\varepsilon_r^T x_v^{T,})+(I_m\otimes C'\Omega^{-1})\mathrm{vec}(\varepsilon_r^T x_{r-1}^{T,})\right.$$

$$\left.+\mathrm{vec}(C'\Omega^{-1}\gamma_{r-1,r}^{T,})\right)\left(\sum_{w=0}^{s-2}\mathrm{vec}(\varepsilon_s^T\varepsilon_w^{T,})'(I_n\otimes\Omega^{-1}CA^{s-1-w})\right.$$

$$\left.+\mathrm{vec}(\varepsilon_s^T\varepsilon_{s-1}^{T,})'(I_n\otimes\Omega^{-1}C)-\mathrm{vec}(C'\Omega^{-1}\alpha_s\Gamma^{-1}\alpha_{s-1}')'\right)\}$$

Remark: The above derivations only briefly describe the necessary steps to reach the information matrix. However, the rest of the calculations requires trivial expansions of the equations shown above.

This completes our derivation of the Fisher information matrix and appendix B.

References

Aoki(1990) *State Space Modelling of Time Series.* Springer Verlag.

Balakrishnan(1984) *Kalman Filtering Theory.* Optimisation Software, Inc..

Breckling, Chambers, Dorfman, Tam and Welsh(1990) Maximum Likelihood Inference from Sample Survey Data. *Preprint.*

Billingsley(1961) *Statistical Inference for Markov Processes.* University of Chicago Press.

Brockwell and Davies(1991) *Time Series Analysis: Theory and Methods.* Springer Verlag.

Crabbe and Young(1989) Model Reduction via the Internally Balanced State Space Representation. *J. of Stat. Comp. Simul.*, 33, 233-248.

Dempster, Laird and Rubin(1977) Maximum Likelihood from Incomplete Data via EM algorithm. *J. of Royal Stat. Soc. Series B*, 39, 1-38.

Dunsmuir and Robinson(1981) Estimation of Time Series Models in the Presence of Missing Data. *J. of Amer. Stat. Assoc.*, 76, 560-568.

Evangelisti(1968) *Controllability and Observability.* C.I.M.E.

Graham(1981) *Kronecker Products and Matrix Calculus with Applications.* Ellis Horwood Ltd.

Hall and Hedye(1980) *Martingale Limit Theory and its applications.* Academic Press.

Hannan and Deistler(1988) *The Statistical Theory of Linear Systems.* John Wiley and Sons.

Hannan, Krishnaiah and Rao(1985) *Handbook of Statistics Vol. 5: Time Series in the Time Domain.* Elsevier Science Publishers B.V.

Huber(1980) *Robust Statistics.* John Wiley and Sons.

Jones(1979) Maximum Likelihood Fitting of ARIMA Models to Time Series with Missing Observations. *Technometrics*, 22, 389-395.

Jones(1978) Multivariate Autoregression Estimation using Residuals. *Applied Time Series Analysis*, 139-162.

Jong Piet de(1991) The Diffuse Kalman Filter. *The Annals of Statistics*, 19, 1073-1083.

Kohn and Ansley(1986) Estimation, Prediction and Interpolation for ARIMA Models with Missing Data. *J. of Amer. Stat. Assoc.*, 81,

751-761.

Louis(1982) Finding the Observed Information Matrix when using the EM algorithm. *Journal of Royal Stat. Soc.*, Series B(**44**),226-233.

McCullagh and Nelder(1989) *Generalised Linear Models*. Chapman and Hall.

Mehra and Lainiotis(1976) *System Identification Advances and Case Studies*. Academic Press.

Otter(1985) *Dynamic Feature Space Modelling, Filtering and Self-Tuning Control of Stochastic Systems*. Springer Verlag.

Parzen(1983) *Time Series Analysis of Irregularly Observed Data*. Lecture Notes Series, Springer Verlag.

Parzen(1967) *Time Series Analysis Papers*. Holden-Day.

Rao(1984) *Optimisation: Theory and Applications*. Wiley Eastern Ltd.

Rubin(1976) Inference and Missing Data. *Biometrika*, **63**, 581-592.

Sergio Bittanti(1986) *Time Series and Linear Systems*. Springer Verlag.

Shumway and Stoffer(1982) An Approach to Time Series Smoothing and Forecasting using EM algorithm. *Journal of Time Series Analysis*, 3,253-264.

Stoffer(1986) Estimation and Identification of Space-Time ARMAX Models in the Presence of Missing Data. *J. of Amer. Stat. Assoc.*, **81**, 762-772.

Whittle(1963) On the Fitting of Multivariate Autoregressions and the Approximate Canonical Factorization of a Spectral Density Matrix. *Biometrika*, **50**, 129-134.

Wu(1983) On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*. **11**, 95-103.

Moore(1978) in Time Series and Linear Systems, Springer Verlag, Sergio and Bittanti(1986).

751-761.

Louis(1982) Finding the Observed Information Matrix when using the EM algorithm. *Journal of Royal Stat. Soc.*, Series B(**44**),226-233.

McCullagh and Nelder(1989) *Generalised Linear Models*. Chapman and Hall.

Mehra and Lainiotis(1976) *System Identification Advances and Case Studies*. Academic Press.

Otter(1985) *Dynamic Feature Space Modelling, Filtering and Self-Tuning Control of Stochastic Systems*. Springer Verlag.

Parzen(1983) *Time Series Analysis of Irregularly Observed Data*. Lecture Notes Series, Springer Verlag.

Parzen(1967) *Time Series Analysis Papers*. Holden-Day.

Rao(1977) *Optimisation: Theory and Applications*. Wiley Eastern Ltd.

Rubin(1976) Inference and Missing Data. *Biometrika*, **63**, 581-592.

Sergio Bittanti(1986) *Time Series and Linear Systems*. Springer Verlag.

Shumway and Stoffer(1982) An Approach to Time Series Smoothing and Forecasting using EM algorithm. *Journal of Time Series Analysis*, **3**,253-264.

Stoffer(1986) Estimation and Identification of Space-Time ARMAX Models in the Presence of Missing Data. *J. of Amer. Stat. Assoc.*, **81**, 762-772.

Whittle(1963) On the Fitting of Multivariate Autoregressions and the Approximate Canonical Factorization of a Spectral Density Matrix. *Biometrika*, **50**, 129-134.

Wu(1983) On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*. **11**, 95-103.

Parzen(1963) in *Time Series Analysis Paper* Holden Day, Parzen(1967).

Rubin and Little(1987) *Statistical Analysis with Missing Data*, New York; John Wiley.

Hannan(1986) in *Time Series and Linear Systems* Springer Verlag, Sergio Bittanti(1986)

Hannan and Quinn(1979) *The determination of the order of an autoregression*, J. Roy. Stat. Soc. B. 41; P.190-195.

Davidon, Fletcher and Powell(1963) in *Optimisation: Theory and Applications*, Wiley Eastern Ltd., Rao(1984).

Box and Jenkins(1970) *Time Series Analysis: Forecasting and Control*, San Francisco; Holden Day.

Kalman(1968) *On Partial Realisations of a Linear input/output map*, Guillemin Anniversary Volume, Holt, Winston and Rinehart.

Kalman(1969) *Lectures on Algebraic System Theory*, Springer Lecture Notes.

Bucy and Joseph(1968) *Filtering for Stochastic Processes with Applications to Guidance*, New York; Interscience Publishers.

Kalman and Bucy(1969) *New results in linear prediction and filtering theory*, J. Basic Engr. (Trans. ASME, Ser. D), 83D, P.95-100.

Piet de Jong and SingFat Chu-Chun-Lin(1992) *Stationary and Non-Stationary State Space Models*, J. of Time Series Analysis, to appear.

Fisher(1925) *Theory of Statistical Estimation*, Proc. Camb. Phil. Soc., 22, P.700-725.

Jones(1983) in *Handbook of Statistics Vol. 5: Time Series in Time Domain*, Elsevier Science Publishers B.V., Hannan, Krishnaiah and Rao(1985).

Cox and Wermuth(1992) *A Comment on the Coefficient of Determination for Binary Responses*, The American Statistician, Vol. 46, No. 1, P.1-3.

Ljung and Soderstrom(1983) *Theory and Practice of Recursive Identification*, Cambridge, Mass., MIT Press.

Robinson(1967) *Multichannel Time Series Analysis*, San Francisco, Holden-Day.

Ortega and Rheinboldt(1970) in *System Identification: Advances and Case Studies*, Academic Press, Mehra and Lainiotis(1976).