Statistica Sinica 1(1991), 389-400

AN ADAPTIVE NONLINEAR STATE SPACE MODEL APPLIED TO MODELLING EPIDEMICS

Richard H. Jones and Des F. Nicholls

University of Colorado and The Australian National University

Abstract: A state space model is developed for a system of nonlinear differential equations with observations consisting of nonlinear functions of the state variables. This is applied to modelling gonorrhea transmission in a heterosexual population. Variable transformations are used to keep the incidence rates in the interval zero to one and the unknown parameters in the proper ranges. A refinement of the model allows adaptively varying contact rates. The Kalman filter is used to calculate an approximate likelihood, and nonlinear optimization is used to obtain approximate maximum likelihood estimates.

Key words and phrases: State space models, nonlinear differential equations, Kalman filter, epidemics, adaptive estimation.

1. Introduction

Hethcote and Yorke (1984) discuss deterministic models for the spread of gonorrhea. Heathcote and Nicholls (1990) use least-squares to estimate the contact rates in the case of discrete time nondeterministic generalizations of these models. This paper uses a state space approach and generalizes these models by including random inputs. This generates a system of stochastic nonlinear differential equations. By integrating the nonlinear differential equations numerically, using small time intervals, linearization methods can be used to obtain approximate propagation of the state covariance matrix. The Kalman (1960) filter is used to calculate the approximate $-2\ln$ likelihood (Schweppe (1965)), and numerical optimization is used to obtain estimates of the unknown parameters. The method will be developed for a particular bivariate model.

2. The Basic Model

A bivariate model considered by Hethcote and Yorke (1984, p.26) for the heterosexual transmission of gonorrhea is

$$\begin{aligned}\frac{d}{dt}\{N_1 I_1(t)\} &= \lambda_{12}\{1 - I_1(t)\}N_2 I_2(t) - N_1 I_1(t)/d_1 \\ \frac{d}{dt}\{N_2 I_2(t)\} &= \lambda_{21}\{1 - I_2(t)\}N_1 I_1(t) - N_2 I_2(t)/d_2,\end{aligned} \tag{1}$$

where N_1 is the number of females in the population, and $I_1(t)$ is the proportion of females infected at time t. Similarly, N_2 is the number of males in the population, and $I_2(t)$ is the proportion of males infected at time t. Assuming that there is only heterosexual transmission of the infection, λ_{12} is the contact rate between a susceptible female and an infected male. This is interpreted as the average number of adequate contacts per unit time between an infected male and a susceptible female, where an adequate contact is one in which the disease is transmitted from the infected to the susceptible. The unit of time in the example presented here is one year. λ_{21} is the contact rate between a susceptible male and an infected female. d_1 is the mean length of time that a female remains infected before returning to the population as a susceptible, and d_2 is the mean length of time that a male remains infected. The interpretation of these equations is that the rate of change of the number of infected females is proportional to the contact rate for females with infected males multiplied by the proportion of females who are susceptible times the number of infected males. The last term with the negative sign reflects the rate at which infected females re-enter the population as susceptibles. Assuming that the population is constant over time, Equation (1) can be written

$$\begin{aligned}\frac{d}{dt}I_1(t) &= (\lambda_{12}/r)\{1 - I_1(t)\}I_2(t) - I_1(t)/d_1 \\ \frac{d}{dt}I_2(t) &= r\lambda_{21}\{1 - I_2(t)\}I_1(t) - I_2(t)/d_2,\end{aligned}$$

where $r = N_1/N_2$ is the ratio of females to males in the population.

If random inputs are added to these equations, the rates of infection may not remain in the interval zero to one. One possibility is to use a logistic transformation on the $I_i(t)$, $i = 1, 2$, which maps them to the real line. This transformation is

$$u_i(t) = \ln \frac{I_i(t)}{1 - I_i(t)}$$

with the inverse transformation

$$I_i(t) = \frac{\exp\{u_i(t)\}}{1 + \exp\{u_i(t)\}}.$$

Since $\frac{d}{dt}I_i(t) = I_i(t)\{1 - I_i(t)\}\frac{d}{dt}u_i(t)$, the equations can be written

$$\begin{aligned}\frac{d}{dt}u_1(t) &= (\lambda_{12}/r)I_2(t)/I_1(t) - [d_1\{1 - I_1(t)\}]^{-1}\\ \frac{d}{dt}u_2(t) &= r\lambda_{21}I_1(t)/I_2(t) - [d_2\{1 - I_2(t)\}]^{-1}.\end{aligned}$$

These equations can be integrated forward in time using a small time step δt,

$$\begin{aligned}u_1(t+\delta t) &= u_1(t) + \big((\lambda_{12}/r)I_2(t)/I_1(t) - [d_1\{1 - I_1(t)\}]^{-1}\big)\delta t\\ u_2(t+\delta t) &= u_2(t) + \big(r\lambda_{21}I_1(t)/I_2(t) - [d_2\{1 - I_2(t)\}]^{-1}\big)\delta t.\end{aligned} \tag{2}$$

Equation (2) can be used to predict the elements of the state vector, $u_1(t)$ and $u_2(t)$ forward in time. It is important that δt be chosen small enough so that the results approximate the solution of the nonlinear differential equations. A reasonable value of δt can be obtained by experimentation. If a value of δt is chosen and the numerical integration carried out, the results should not differ much if the value of δt is cut in half. In the example presented in this paper with yearly data, a value of δt equal to one year was much too large. After some experimentation, a value of $\delta t = .05$ was chosen.

While predictions of the state vector can be carried out using numerical integration, the updating of the state covariance matrix after a prediction over a time interval of δt can only be approximated because the equation is nonlinear. The approximate propagation of the state covariance matrix can be carried out by expanding these equations in a Taylor series about the values at time t. If $f_i(t)$ represents the right hand side of equation i, element ij of the linearized state transition matrix is

$$\Phi_{ij}(t) = \frac{\partial}{\partial u_j}f_i(t).$$

Dropping the argument t for convenience, the linearized state transition matrix is

$$\begin{bmatrix} 1 - \left(\frac{\lambda_{12}I_2(1-I_1)}{rI_1} + \frac{I_1}{d_1(1-I_1)}\right)\delta t & \frac{\lambda_{12}I_2(1-I_2)}{rI_1}\delta t \\ \frac{r\lambda_{21}I_1(1-I_1)}{I_2}\delta t & 1 - \left(\frac{r\lambda_{21}I_1(1-I_2)}{I_2} + \frac{I_2}{d_2(1-I_2)}\right)\delta t \end{bmatrix}. \tag{3}$$

The data to be used in the example were obtained from the Center for Disease Control[1] and are shown in Table 1. These rates will be used as the observations and denoted by $y_1(t)$ for females and $y_2(t)$ for males.

[1] STD Statistics No. 135. Source: Form CDC 73.2638:HHS:PHS:CDC:DSTD:ESSB:SSS: Atlanta, Georgia 30333, U.S.A.

3. The Kalman Filter

The use of Kalman filter for calculating the exact likelihood for linear state space models with Gaussian errors will be reviewed here and related to our nonlinear model. The state equation contains the dynamics of the system being studied. For a linear system the state equation is

$$s(t+\delta t) = \Phi(t+\delta t; t)s(t) + G(t)w(t),$$

where $s(t)$ is the state vector, $\Phi(t+\delta t; t)$ is the state transition matrix from time t to time $t+\delta t$, and $G(t)$ is a matrix multiplying the vector of random inputs that are assumed to have a normal distribution with zero mean and covariance matrix equal to the identity matrix. Correlation is introduced through the matrix $G(t)$ which is assumed, in the example presented here, to be lower triangular so the GG' is the Cholesky factorization of the random input covariance matrix. Parameterizing the input covariance matrix by using the factor G ensures that the input covariance matrix will remain non-negative definite during the nonlinear optimization search. The random inputs $w(t)$ are assumed to be uncorrelated at different times. The observations are specified by the observation equation

$$y(t_j) = H(t_j)s(t_j) + v(t_j),$$

where $y(t_j)$ is a vector of observations at time t_j which are linear combinations of the state vector, specified by the matrix $H(t_j)$, plus a random observational error vector $v(t_j)$. These errors are assumed to be Gaussian with zero mean and covariance matrix $R(t_j)$, uncorrelated at different times and uncorrelated with the state noise $w(t)$. The matrices $\Phi(t+\delta t; t)$, $G(t)$, $H(t_j)$ and R may contain unknown parameters to be estimated by maximum likelihood. The initial state vector is specified as $s(0|0)$ with initial state covariance matrix $P(0|0)$. The notation $s(t_j|t_i)$ indicates the optimal estimate of the state at time t_j given observations up to time t_i. These initial conditions may either be specified as known properties of the system or estimated by maximum likelihood. When $P(0|0)$ is estimated by maximum likelihood, this is referred to as an empirical Bayes procedure.

The Kalman recursion as presented here is similar to the procedure used by Jones (1980) for missing observations. There are small steps between observations used for the purpose of numerical integration, as follows:

1. Calculate a one step prediction $s(t+\delta t) = \Phi(t+\delta t; t)s(t|t)$.
2. Calculate the covariance matrix of this prediction

$$P(t+\delta t|t) = \Phi(t+\delta t; t)P(t|t)\Phi'(t+\delta t; t) + G(t)G'(t),$$

where the $'$ denotes the transposed matrix. When we are integrating forward using a small step size between observations, these first two steps are repeated until the next observation point is reached.

3. Predict the next observation vector, $y(t_j|t_j - \delta t) = H(t_j)s(t_j|t_j - \delta t)$.
4. The innovation vector is the difference between the observation vector and the predicted observation vector, $\epsilon(t_j) = y(t_j) - y(t_j|t_j - \delta t)$.
5. The innovation covariance matrix is $V(t_j) = H(t_j)P(t_j|t_j-\delta t)H'(t_j)+R(t_j)$.
6. For the purpose of calculating $-2\ln$ likelihood at the end of the recursion, $\epsilon'(t_j)V^{-1}(t_j)\epsilon(t_j)$ and $\ln|V(t_j)|$ are accumulated over all the observations. Here $|V(t_j)|$ denotes the determinant of the innovation covariance matrix.
7. The state vector is now updated to reflect the new information obtained from the new observation vector,
$$s(t_j|t_j) = s(t_j|t_j - \delta t) + P(t_j|t_j - \delta t)H'(t_j)V^{-1}(t_j)\epsilon(t_j).$$
8. The updated state covariance matrix is
$$P(t_j|t_j) = P(t_j|t_j - \delta t) - P(t_j|t_j - \delta t)H'(t_j)V^{-1}(t_j)H(t_j)P(t_j|t_j - \delta t).$$

Now return to step one until the end of the data is reached.

The value of $-2\ln$ likelihood is calculated at the end of the recursion as

$$\ell = \sum_j \{\ln(2\pi|V(t_j)|) + \epsilon'(t_j)V^{-1}(t_j)\epsilon(t_j)\}.$$

It is possible to concentrate one of the unknown variances out of the likelihood by setting it equal to a constant (usually 1) in the recursion. In this case the variance set equal to 1 is estimated as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_j \{\epsilon'(t_j)V^{-1}(t_j)\epsilon(t_j)\},$$

where n is the number of observation times, and $-2\ln$ likelihood becomes

$$\ell = n\{1 + \ln(2\pi\hat{\sigma}^2)\} + \sum_j \ln|V(t_j)|.$$

In the nonlinear problem considered in this paper, the state vector is

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix},$$

and $G(t)$ is a constant lower triangular 2×2 matrix with arbitrary elements to be estimated by nonlinear optimization. The initial values, $u(0|0)$, are assumed to be unknown and will also be estimated by nonlinear optimization, and $P(0|0)$ will be set equal to zero. This states that the initial conditions are fixed but unknown. Step 1 in the Kalman recursion is calculated from Equation (2) after linearization using the transition matrix (3). Step 2 is calculated using the transition matrix (3). Steps 1 and 2 are repeated until the next observation is reached.

Since the observations are not linear functions of the state variables, Step 3 needs some modification. The actual observation equation is

$$\begin{bmatrix} y_1(t_j) \\ y_2(t_j) \end{bmatrix} = \begin{bmatrix} I_1(t_j)N_1 \\ I_2(t_j)N_2 \end{bmatrix} + \begin{bmatrix} v_1(t_j) \\ v_2(t_j) \end{bmatrix}.$$

It will be assumed that the observational error covariance matrix $R = \sigma^2 I$. The female and male population sizes are not known even if the total population for the age group is known since not everyone in the population is at risk, and the proportion at risk may vary with time. N_1 and N_2 can be thought of as the effective population at risk and known to be between the number of cases and the total population. Since the observations are nonlinear functions of the state variables, an approximation to the innovation covariance matrix can be obtained by linearization. Again omitting time for convenience, if the observations are $y_i = g_i(s)$ where s denotes the state vector, element ij of the linearized H matrix is

$$H_{ij} = \frac{\partial}{\partial s_j} g_i(s).$$

In our example $\frac{\partial}{\partial u_i} I_i = I_i(1 - I_i)$; so the linearized H matrix is

$$H = \begin{bmatrix} I_1(1-I_1)N_1 & 0 \\ 0 & I_2(1-I_2)N_2 \end{bmatrix}.$$

4. Results for Non-Adaptive Model

The model as specified has potentially 11 nonlinear parameters, the two initial incidences, $I_1(0)$ and $I_2(0)$, the two contact rates, λ_{12} and λ_{21}, the three elements of the matrix G, G_{11}, G_{21}, G_{22}, the female population size, N_1, the ratio of females to males, r, and the two mean length of infections, d_1 and d_2. After some preliminary calculations and using information in Hethcote and Yorke (1984), the last four parameters were held fixed. The data do contain information about all eleven parameters, but these can not be estimated with any precision from the 64 data values. If the entire population were sexually active and randomly mixing, N_1 would be 100. Here we set $N_1 = 20$, and the ratio of active females to males at $r = 0.5$. From Hethcote and Yorke, d_1 was taken to be 80 days and d_2, 20 days. Since the unit of time in the analysis is the year, $d_1 = 80/365$ and $d_2 = 20/365$. The final estimates of the other parameters are

$$I_1(0) = .315, \quad I_2(0) = 1.05, \quad \lambda_{12} = 2.79, \quad \lambda_{21} = 33.3,$$
$$G_{11} = .291, \quad G_{21} = .236, \quad G_{22} = .574.$$

The mean square error which estimates the observational error variance is 0.171, and $-2 \ln$ likelihood $= 45.53$. The results of the fit are shown in Figure 1. The top graph shows the actual data with circles for females and triangles for males. The dashed lines are the one step predictions from the previous time point, so the difference between each data curve and the corresponding dashed curve gives the innovations. The bottom curve shows the estimated λ_{12} and λ_{21} which are assumed to be constant over time.

The top curves in Figure 1 show that the model is not a good fit to the data. The innovations should be random over time with zero mean and the curves should not show strong systematic differences. For the females, the infected proportion is being over predicted for the first half of the curve and under predicted for the second half of the curve. This leaves the possible conclusion that the λ's may not be constant over time.

5. An Adaptive Model

Wecker and Ansley (1983) developed a method of nonparametric modelling using a state space model which assumed that the function was an integrated random walk observed with error. Assuming a noninformative prior for the initial conditions, this approach generates a smoothing spline fit to the data. The order of the spline is determined by the number of times the random walk is integrated. A single integration produces a cubic spline. The smoothing parameter is estimated by maximum likelihood. Anderson et al. (1990) generalized this to vector processes. The basic model in Section 2 can be generalized to allow the λ's to vary with time. If the $\ln \lambda$'s are assumed to be integrated random walks so that the λ's remain positive, the state vector contains four additional elements,

$$u_3(t) = \ln \lambda_{12}(t), \quad u_4(t) = \ln \lambda_{21}(t), \quad u_5(t) = \frac{d}{dt} \ln \lambda_{12}(t), \quad u_6(t) = \frac{d}{dt} \ln \lambda_{21}(t).$$

The continuous time state equation for these four elements is

$$d \begin{bmatrix} u_3(t) \\ u_4(t) \\ u_5(t) \\ u_6(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_3(t) \\ u_4(t) \\ u_5(t) \\ u_6(t) \end{bmatrix} dt + d \begin{bmatrix} 0 \\ 0 \\ W_1(t) \\ W_2(t) \end{bmatrix}.$$

Here $W_1(t)$ and $W_2(t)$ are correlated Wiener processes and the W's are assumed to have a general 2 x 2 covariance matrix. Any random errors input to the first two elements of the state are uncorrelated with these random inputs. Variations of this model allow the random walks to be integrated any number of times.

The discrete form of the second part of the state equation when integrated over a small time interval δt is (Jones and Tryon (1987)),

$$\begin{bmatrix} u_3(t+\delta t) \\ u_4(t+\delta t) \\ u_5(t+\delta t) \\ u_6(t+\delta t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \delta t & 0 \\ 0 & 1 & 0 & \delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_3(t) \\ u_4(t) \\ u_5(t) \\ u_6(t) \end{bmatrix} dt + \begin{bmatrix} 0 \\ 0 \\ \eta_1(t) \\ \eta_2(t) \end{bmatrix} .$$

The linearized state transition matrix is now

$$\begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} & 0 & 0 & 0 \\ \Phi_{21} & \Phi_{22} & 0 & \Phi_{24} & 0 & 0 \\ 0 & 0 & 1 & 0 & \delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

where Φ_{11}, Φ_{12}, Φ_{21} and Φ_{22} are from Equation (3). The elements Φ_{13} and Φ_{24} are obtained by differentiating the first equation of (2) with respect to $u_3 = \ln \lambda_{12}$, and the second equation with respect to $u_4 = \ln \lambda_{21}$, which gives

$$\Phi_{13} = r^{-1}\lambda_{12}(I_2/I_1)\delta t, \quad \Phi_{24} = r\lambda_{21}(I_1/I_2)\delta t.$$

The calculations for the adaptive method are the same as the non-adaptive method except that the state vector is larger. The nonlinear parameters λ_{12} and λ_{21} in the non-adaptive method are replaced by the initial values of the last four elements of the state vector. These four parameters are the initial values of the logs of the contact rates and the initial values of their derivatives. These are set equal to unknown constants to be estimated by nonlinear optimization. A question remains as to whether the random input to the state should be on the first two elements of the state vector, the last two elements of the state vector, or both. If the random input is on both and it is assumed that the first two are uncorrelated with the last two, three more nonlinear parameters are introduced. If no randomness were introduced into the contact rates, they would be straight lines determined by the initial values and slopes.

6. Results for the Adaptive Model

The parameters N_1, r, d_1 and d_2 were fixed at the same values as in the non-adaptive model. The random input was tried on the first two and the last two elements of the state vector. A better fit was obtained with the random input on the last two elements of the state vector with no significant improvement when random inputs were applied to both the first two and the last two elements of the state vector. The final estimates of the parameters are

$$I_1(0) = .332, \quad I_2(0) = 1.25, \quad \lambda_{12}(0) = 1.25, \quad \lambda_{21}(0) = 70.5,$$
$$\frac{d}{dt}\ln\lambda_{12}(0) = .355, \quad \frac{d}{dt}\ln\lambda_{21}(0) = -.351,$$
$$G_{11} = .00774, \quad G_{21} = .0380, \quad G_{22} = .0107.$$

The mean square error is 0.0324, and $-2\ln$ likelihood = 0.46. The results of the fit are shown in Figure 2.

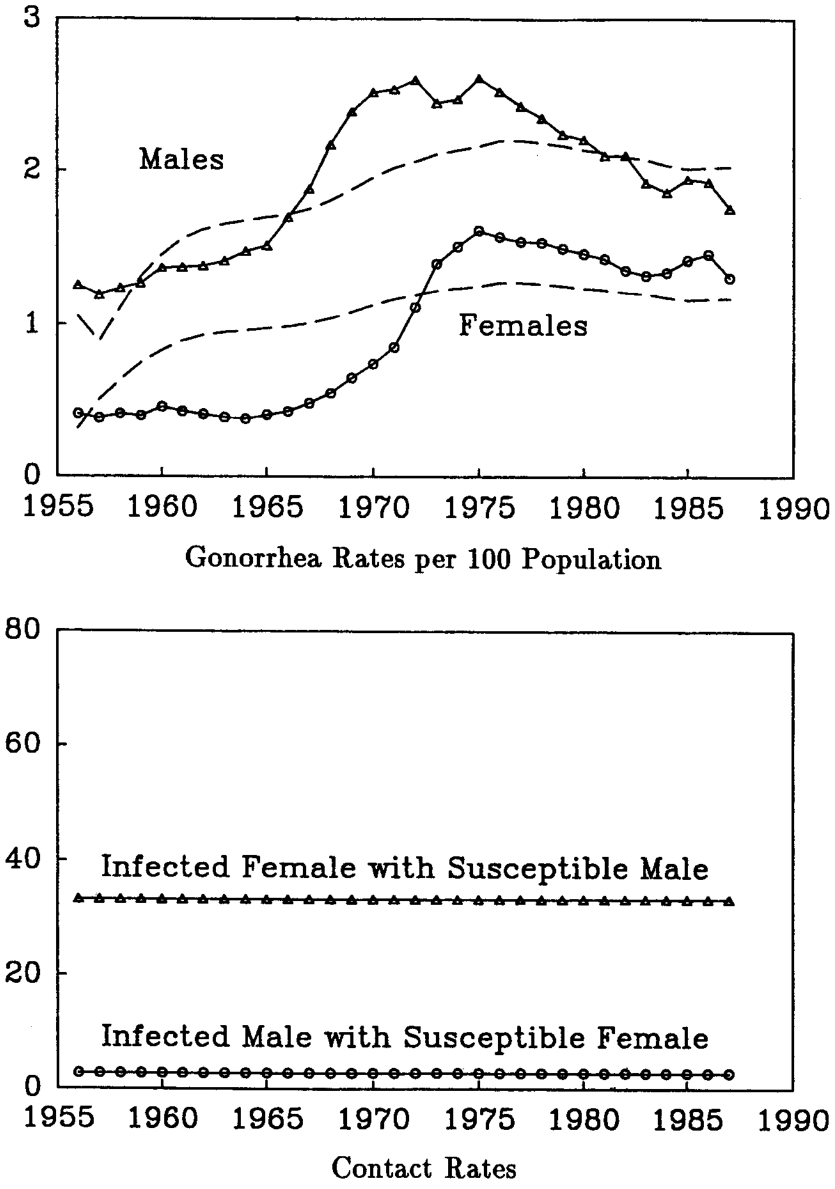
While the fit is significantly better than the non-adaptive fit, there are still systematic errors in the innovations. The improvement in the fit is caused mainly by a continual decrease of the contact rate between infected females with susceptible males.

Acknowledgment

This paper was written while the first author was on sabbatical leave at the Australian National University and partially supported by the U.S. Public Health Service research grant GM38519 from the National Institute of General Medical Sciences.

Table 1. United States gonorrhea rates per 100 population for age group 20–24

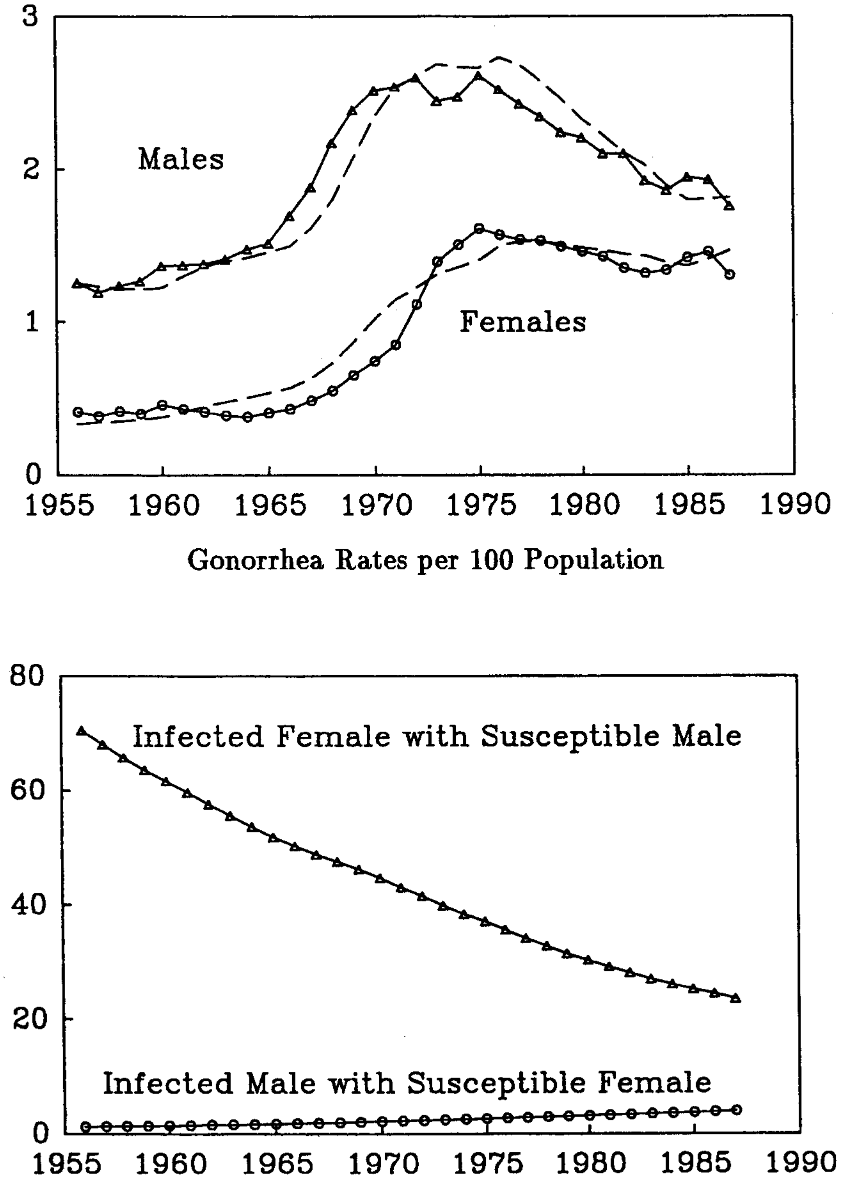
Year	Females	Males	Year	Females	Males
1956	.4095	1.2536	1972	1.1130	2.5969
1957	.3824	1.1912	1973	1.3948	2.4452
1958	.4114	1.2347	1974	1.5063	2.4707
1959	.3972	1.2657	1975	1.6108	2.6122
1960	.4560	1.3669	1976	1.5696	2.5213
1961	.4273	1.3706	1977	1.5384	2.4247
1962	.4081	1.3796	1978	1.5325	2.3447
1963	.3861	1.4102	1979	1.4939	2.2389
1964	.3778	1.4743	1980	1.4607	2.2041
1965	.4044	1.5125	1981	1.4281	2.1012
1966	.4270	1.6945	1982	1.3536	2.1026
1967	.4819	1.8806	1983	1.3200	1.9237
1968	.5473	2.1692	1984	1.3404	1.8607
1969	.6486	2.3856	1985	1.4206	1.9477
1970	.7405	2.5124	1986	1.4610	1.9311
1971	.8495	2.5365	1987	1.3063	1.7585



Gonorrhea Rates per 100 Population

Contact Rates

Figure 1. Non-adaptive model with fixed contact rates. Upper graph shows the observed gonorrhea rates with circles for females and triangles for males. One step predictions from the state space model are shown by the dashed lines. The lower graph shows the estimated contact rates.



Gonorrhea Rates per 100 Population

Contact Rates

Figure 2. Adaptive model with time varying contact rates.

References

Anderson, S. J., Jones, R. H. and Swanson, G. D. (1990). Smoothing polynomial splines for bivariate data. *SIAM J. Sci. Statist. Comput.* **11**, 749–766.

Heathcote, C. R. and Nicholls, D. F. (1990). Least-squares estimation of the contact rate in models for the spread of infectious diseases. *Biometrika* **77**, 161–168.

Hethcote, H. W. and Yorke, J. A. (1984). *Gonorrhea Transmission Dynamics and Control.* Lecture Notes in biomath. **56**, Springer-Verlag, Berlin.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389–395.

Jones, R. H. and Tryon, P. V. (1987). Continuous time series models for unequally spaced data applied to modeling atomic clocks. *SIAM J. Sci. Statist. Comput.* **8**, 71–81.

Kalman R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME* (*J. Basic Engineering*) **82D**, 35–45.

Schweppe, F. C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory* **11**, 61–70.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78**, 81–89.

Department of Preventive Medicine and Biometrics, School of Medicine, Box B-119, University of Colorado, Health Sciences Center, Denver, Colorado 80262, U.S.A.

Department of Statistics, The Faculty of Economics and Commerce, The Australian National University, GPO Box 4, Canberra ACT 2601, Australia.

(Received January 1990; accepted January 1991)