2015-04-09

Epidemic Forecasting

EPIDEMIC FORECASTING

Review of the state of the art

Dexter Barrows
April 9 2015
Department of Mathematics and Statistics
McMaster University

Take a deep breath, think about kitties.

Done? Good. Now welcome people.

Introduction
Techniques
Phenomenological
Mechanistic / semimechanistic
Data assimilation
Measuring prediction accuracy

Basics

- Prediction of future values in a time series
- Based on mechanistic understanding, data, mix

Outbreak type

- New disease
  - Scarcity of information is key concern
  - Forecasting extremely difficult
- Established disease
  - Long time series, likely better biological understanding
  - Short-term forecasting is easiest (information plentiful)
  - Long-term forecasting possible, integration of weather/socio-economic factors important

- Prediction of future values in a time series
- Based on mechanistic understanding, data, mix
- Short term:
  - Key concern: scarcity of information (biological, observational)
- Long term
  - Plentiful observational data
  - Integration of weather/socio-economic factors important

2015-04-09

TECHNIQUES

3 main families

- Phenomenological - pure inference from data
- Mechanistic - capture “drivers” of disease spread
- Semi-mechanistic - integration of data into model

ARIMA

- AutoRegressive Integrated Moving Average
- Purely phenomenological
- Assumes linear process, Gaussian distributions
- 3-parameter process $ARIMA(p, d, q)$, indicating order of
  - p - Autoregressive
    $\rightarrow$ Linear combination of past terms
  - d - Integrated
    $\rightarrow$ Used to remove the trend - makes the series stationary
  - q - Moving average
    $\rightarrow$ Dependence on past error
- Orders are determined using
  - *ACF* - works on consecutive elements in series (correlation)
  - *PACF* - works on additional predictor variables (conditional correlation)

[9]

- **A**uto**R**egressive **I**ntegrated **M**oving **A**verage
- Purely phenomenological (**only** data)
- Assumes linear process, Gaussian distributions
- 3-parameter process

  $$ARIMA(p, d, q)$$

  , indicating order of
  - p - Autoregressive
    $\rightarrow$ Linear combination of past terms
  - d - Integrated
    $\rightarrow$ Used to remove the trend - makes the series stationary
  - q - Moving average
    $\rightarrow$ Dependence on past error
- Orders are determined using
  - *ACF* - works on consecutive elements in series (correlation)
  - *PACF* - works on additional predictor variables (conditional correlation)
- General form

  $$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

  where X_t is the time series being considered

SARIMA

- Adaptation of ARIMA used to capture seasonal effects
- Usually expressed as $\mathit{SARIMA}(p, d, q) \times (P, D, Q)_s$
  - P, D, Q are *seasonal* orders
- Orders are determined using
  - *ACF* and *PACF* as before
  - Also Periodic *ACF* (every k elements)

- Adaptation of *ARIMA* used to capture seasonal effects
- Usually expressed as $\mathit{SARIMA}(p, d, q) \times (P, D, Q)_s$
  - P, D, Q are *seasonal* orders
- Orders are determined using
  - *ACF* and *PACF* as before
  - Also Periodic *ACF* (every k elements)

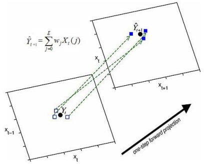
- Construct a “library” of consecutive time lag vectors $\{x_i\}$ of some length E and corresponding forward trajectories $\{y_i\}$
- Use similar past system states with **known** outcomes to project to **unknown** future state
  $\rightarrow$ A weighted linear combination of closest vectors
- Weightings are exponential, function of distance

[2]

- Construct a “library” of consecutive time lag vectors $\{x_i\}$ of some length E and corresponding forward trajectories $\{y_i\}$
- Use similar past system states with **known** outcomes to project to **unknown** future state
  $\rightarrow$ A weighted linear combination of closest vectors
- Weightings are exponential, function of distance

Math

- X_i - neighbour library vectors
- X_t - predictee vector
- $\hat{Y}$ - prediction
- Distances: $d = ||X_i - X_t||$
- Weights: $w(d) = e^{-d}\bar{d}$
- Projection: $\hat{Y} = \sum_{i=1}^{E} w_i Y_i / \sum_{j=1}^{E} w_j$



http://simplex.ucsd.edu/

- Sequentially locally weighted global linear maps (S-map)
- Designed to handle linear, locally nonlinear time series
- Similar to Simplex projection
  → But **all** vectors are used for projection
- Weightings are again exponential

[2][11]

- Sequentially locally weighted global linear maps (S-map)
- Similar to Simplex projection → But **all** vectors are used for projection
- Weightings are again exponential

Procedure

1. Construct a “library” of time lag vectors $\{x_i\}$ of length E and corresponding forward trajectories $\{y_i\}$
2. Choose a state x_t from which you wish to forecast the next system state y_t
3. The estimation $\hat{y}_t$ is evaluated using

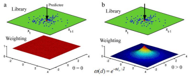
$$\hat{y}_t = \sum_{j=0}^{E} c_t(j) x_t(j).$$

Here c is obtained by solving the system $b = Ac$ where

$$\begin{aligned} b(i) &= w(||x_i - x_t||) y_i \\ A(i,j) &= w(||x_i - x_t||) x_i(j) \end{aligned}$$

where the weights are a function of Euclidean distance

$$w(d) = e^{\frac{-\theta d}{\bar{d}}}$$



http://simplex.ucsd.edu/

- Extensively used model in epidemiology
- Division into classes: **S**usceptible-**I**nfected-**R**emoved
- Transition between states

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N}\\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma\\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

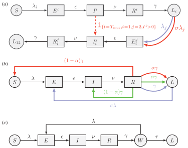
- Many extensions exist
  - Additional classes
  - Additional mechanistic terms

- Extensively used model in epidemiology
- Division into classes: **S**usceptible-**I**nfected-**R**emoved
- Transition between states

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N}\\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma\\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

- Many extensions exist
  - Additional classes
  - Additional mechanistic terms

Can be used to test hypotheses



Camacho et al, 2011

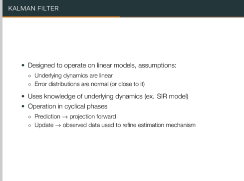
[3]

Example modification

- Part (c) hypothesis: window of reinfection
  - long-term immunity takes time to develop
- Three new classes added
  - *E* - exposed to virus, not yet infective
  - *W* - susceptible to reinfection, long-term immunity not developed
  - *L* - long term immunity developed

- SIR-based models *may* require many parameters to be estimated
  - Not a problem if statistical caution is exercised
- Over-fitting a particular problem - can reduce forecasting ability
- More model complexity = longer time series required
- Iterated filtering methods can estimate parameters in addition to producing forecasts

- SIR-based models may require many parameters to be estimated
- Over-fitting a particular problem - can reduce forecasting ability
- More model complexity = longer time series required
- Iterated filtering methods can estimate parameters in addition to producing forecasts



- Predictive method that operates on noisy data to produce optimal estimate of next system state
- Designed to operate on linear models
- Uses a model of expected system behaviour (for example a system of ODEs) to help with prediction

Prediction / update cycle

- Prediction

$$
\begin{aligned}
\hat{x}_{k|k-1} &= F_k \hat{x}_{k-1|k-1} + B_k u_k \\
P_{k|k-1} &= F_k P_{k-1|k-1} F_k^T + Q_k
\end{aligned}
$$

- Update

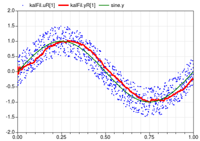
$$
\begin{aligned}
\tilde{y}_k &= z_k - H_k \hat{x}_{k|k-1} \\
S_k &= H_k P_{k|k-1} H_k^T + R_k \\
K_k &= P_{k|k-1} H_k^T S_k^{-1} \\
\hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \tilde{y}_k \\
P_{k|k} &= (I - K_k H_k) P_{k|k-1}
\end{aligned}
$$

2015-04-09

Epidemic Forecasting

└─Techniques

└─Mechanistic / semimechanistic



http://simulationresearch.lbl.gov/modelica/releases/latest/help/Buildings_Utilities_IO_Python27_Examples.html

KALMAN FILTER EXTENSIONS

- Extended Kalman filter (*EKF*)
  - Linearises about the estimate of current mean and covariance
- Ensemble Kalman filter (*EnKF*)
  - Uses a cohort of ensemble members, their sample mean and covariance
  - Still assumes linear process / Gaussian distributions
  - Useful for large number of parameters
- Ensemble Adjustment Kalman filter (*EAKF*)
  - Combination of *EKF* and *EnKF*
    - Linearises as in *EKF*
    - Ensemble members as in *EnKF*

[10]

- Extended Kalman filter (*EKF*)
  - Linearises about the estimate of current mean and covariance
- Ensemble Kalman filter (*EnKF*)
  - Uses a cohort of ensemble members, their sample mean and covariance
  - Still assumes linear process / Gaussian distributions
  - Useful for large number of parameters
- Ensemble Adjustment Kalman filter (*EAKF*)
  - Combination of *EKF* and *EnKF*
    - Linearises as in *EKF*
    - Ensemble members as in *EnKF*

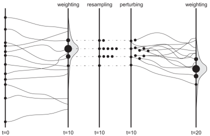
PARTICLE FILTER

- Uses a set of particles, similar to *EnKF* cohort
- Makes no assumption about the distributions involved in the system
- Particle importance using weights
- Problem: Particle degeneracy
  - When one particle accumulates most of the weight
  - Avoided via resampling at each iteration

- Uses a set of particles, similar to *EnKF* cohort
- Makes no assumption about the distributions involved in the system
- Particle importance using weights
- Problem: Particle degeneracy
  - When one particle accumulates most of the weight
  - Avoided via resampling at each iteration

SIS Sequential Importance sampling

- Each of the P particles at time t consists of a weight-state pair $(w_t^{(i)}, x_t^{(i)})$, such that $\sum_{i=1}^{P} w_t^{(i)} = 1$
- Next system state (forecast) is given by the weighted average $\hat{x}_t = \sum_{i=1}^{P} w_{t-1}^{(i)} f(x_{t-1}^{(i)})$
- After forecast, a new observation x_t is assimilated and weights are recalculated based on how accurate the individual projections were



http://www.mdpi.com/sensors/sensors-12-16291/article_deploy/html/images/sensors-12-16291f2-1024.png

Epidemic Forecasting
└─Techniques
└─Mechanistic / semimechanistic
└─Particle filter extensions

2015-04-09

- Maximum likelihood via iterated filtering (MIF or IF1)
  - Uses multiple rounds of particle filtering
  - Stochastic perturbation of parameters
  - Each round pushes the parameter estimates toward ML
- Particle Markov chain Monte Carlo (pMCMC)
  - Uses an MCMC method constrain model parameters
  - Particle filter between each MCMC iteration
- IF2 (MIF2)
  - Evolution of MIF (IF1)
  - Uses stochastic perturbation as before, also data cloning
  - Looks to consistently outperform IF1 and pMCMC

[5][13]

- Maximum likelihood via iterated filtering (MIF)
  - Uses multiple rounds of particle filtering
  - Each round pushes the parameter estimates toward ML
- Particle Markov chain Monte Carlo (pMCMC)
  - Uses an MCMC method constrain model parameters (typically the Metropolis-Hastings algorithm)
  - Particle filter between each MCMC iteration

INCIDENCE DATA
Primary Sources
- Google Flu Trends (GFT)
  - Uses search trend data to infer incidence rates
  - Almost instantaneous, but less accurate
  - Currently up to 29 countries
- Governments ex. Centres for Disease Control (CDC)
  - Regional data (10 regions across the US)
  - Broken down further by age
  - More accurate than GFT, but lag of 1-2 weeks
- WHO
Social Media
- Twitter: Influenza, Korea, 2012
- Social media and informal news: Haiti, Cholera, 2010
[1][6]

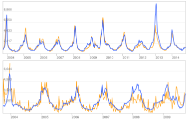
Primary Sources

- Google Flu Trends (GFT)
  - Uses search trend data to infer incidence rates
  - Almost instantaneous, but less accurate
  - Currently up to 29 countries
- Governments ex. Centres for Disease Control (CDC)
  - Regional data (10 regions across the US)
  - Broken down further by age
  - More accurate than GFT, but lag of 1-2 weeks
- WHO

Social Media

- Twitter: Influenza, Korea, 2012
  Very good estimates with 40 keyword markers
- Social media and informal news: Haiti, Cholera, 2010
  Not as good, “Estimates of the reproductive number ranged from 1.54 to 6.89 (informal sources) and 1.27 to 3.72 (official sources) during the initial outbreak growth period, and 1.04 to 1.51 (informal) and 1.06 to 1.73”

GFT VS CDC FLUNET



https://www.google.org/flutrends/about/how.html

- TOP
  - US ILI
  - Blue: GFT
  - Green: Centres for Disease Control (CDC)
- BOTTOM
  - Canada ILI
  - Blue: GFT
  - Green: Public Health Agency of Canada (PHAC)

SEASONALITY AND WEATHER

- Nearly all infectious disease affected by seasonality
  - Contact
  - Susceptibility
  - Influx of susceptibles
  - Resevior dynamics / vector dynamics
- Weather data sources
  - National Oceanic and Atmospheric Administration (NOAA)
  - NASA Jet Propulsion Laboratory (JPL)

[12][13]

- Nearly all infectious disease affected by seasonality
  - Contact (people stay inside during cold weather)
  - Susceptibility (immunity already low in winter from colds, etc.)
  - Influx of susceptibles (schoolchildren)
  - Resevior dynamics / vector dynamics (mosquitoes more active in hotter weather)
- Weather data sources
  - National Oceanic and Atmospheric Administration (NOAA)
  - NASA Jet Propulsion Laboratory (JPL)

ENSO

- **E**l **N**inõ **S**outhern **O**cillation
- Sustained anomalous ocean surface temperature in the Pacific
- Unpredictable
- Many effects on local populations
- Relevant to epidemic outbreaks in Southeast Asian locales
  - Cholera in Bangladesh
  - Dengue fever in Singapore

[4][8]

- **E**l **N**inõ **S**outhern **O**cillation
- Sustained anomalous ocean surface temperature in the Pacific
- Unpredictable
- Many effects on local populations
- Relevant to epidemic outbreaks in southern locales
  - Cholera in Bangladesh (outbreaks in Dhaka highly correlated with ENSO)
  - Malaria epidemics in South America (outbreaks in Colombia, Guyana, Peru, and Venezuela correlation with ENSO)

- What to measure
  - Peak timing / intensity
  - Magnitude
  - Duration
- How to measure
  - Correlation coefficients
  - RMSE
  - Confidence intervals
  - Receiver operating characteristic (ROC) curves

- What to measure
  - Peak timing / intensity
  - Magnitude
  - Duration
- How to measure
  - Correlation coefficients (Pearson, etc.)
  - Root-Mean-Square Error (RMSE)
  - Confidence intervals
  - Receiver operating characteristic (ROC) curves
    - Used to illustrate the performance of binary classifier system
    - Obtained by plotting false positives against false negatives

2015-04-09

- AIC - Akaike Information Criterion
  - Measures relative model quality
  - Rewards goodness-of-fit, penalizes for number of parameters
- BIC - Bayesian Information Criterion
  - Similar to AIC
  - Tends to penalize many parameters more than AIC
- DIC - Deviance Information Criterion
  - Particularly useful when comparing MCMC-based models
- WAIC - Watanabe-Akaike (widely applicable) Information Criterion
  - More "tuned" to prediction

AIC and BIC require calculating the likelihood at its maximum over θ

WAIC uses the whole posterior density more effectively than DIC

[1] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi.
Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.
PLoS ONE, 6(8):1–8, 2011.

[2] S. M. Glaser, H. Ye, and G. Sugihara.
A nonlinear, low data requirement model for producing spatially explicit fishery forecasts.
Fisheries Oceanography, 23:45–53, 2014.

[3] A. L. Graham, A. Camacho, F. Carrat, O. Ratmann, B. Cazelles, and L. E.-e. Mathe.
Explaining rapid reinfections in multiple- wave influenza outbreaks : Tristan da Cunha 1971 epidemic as a case study.
Proc. R. Soc. B, 2016.

[4] Y. Hii, J. Rocklöv, S. Wall, and L. Ng.
Optimal lead time for dengue forecast.
PLoS neglected tropical …, 6(10), 2012.

[5] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. a. King.
Inference for dynamic and latent variable models via iterated, perturbed Bayes maps.
Proceedings of the National Academy of Sciences, 112(3):719–724, 2015.

[6] E.-K. Kim, J. H. Seok, J. S. Oh, H. W. Lee, and K. H. Kim.
Use of hangeul twitter to track and predict human influenza infection.
PloS one, 8(7):e69305, Jan. 2013.

[7] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe.
A systematic review of studies on forecasting the dynamics of influenza outbreaks.
Influenza and other respiratory viruses, 8(3):309–16, May 2014.

[8] R. C. Reiner, A. a. King, M. Emch, M. Yunus, a. S. G. Faruque, and M. Pascual.
Highly localized sensitivity to climate forcing drives endemic cholera in a megacity.
Proceedings of the National Academy of Sciences of the United States of America, 109(6):2033–6, Feb. 2012.

[9] R. Reyburn, D. R. Kim, M. Emch, A. Khatib, L. von Seidlein, and M. Ali.
Climate variability and the outbreaks of cholera in Zanzibar, East Africa: a time series analysis.
The American journal of tropical medicine and hygiene, 84(6):862–9, June 2011.

[10] J. Shaman, W. Yang, S. Kandula, K. S. Inference, and S. Leone.
Inference and Forecast of the Current West African Ebola Outbreak in Guinea , Sierra Leone and Liberia.
pages 1–17, 2014.

[11] G. Sugihara.
Nonlinear Forecasting for the Classification of Natural Time Series.
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 348:477–495, 1994.

[12] J. D. Tamerius, J. Shaman, W. J. Alonso, K. Bloom-Feshbach, C. K. Uejio, A. Comrie, and C. Viboud.
Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates.
PLoS Pathogens, 9(3), 2013.

[13] W. Yang, A. Karspeck, and J. Shaman.
Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics.
PLoS computational biology, 10(4):e1003583, Apr. 2014.