Panel data analysis via mechanistic models

Edward Ionides
University of Michigan, Department of Statistics

Based on joint work with Carles Bretó and Aaron King.

Lecture 6 at Wharton Statistics Department
Thursday 11th May, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

Panel data

- Panel data, also known as longitudinal data, consists of a collection of time series.
- Each time series, which could itself be multivariate, consists of a sequence of observations measured on a distinct unit.
- The investigator may be interested in properties shared by units and/or the differences between units.
- Inferences that are unclear on any one unit may become unambiguous when analyzing the entire panel.

Mechanistic panel models

- Mechanistic modeling involves writing down scientifically motivated equations describing the collection of dynamic systems giving rise to the observations on each unit.
- A defining characteristic of panel systems is that the dynamic interaction between units should be negligible.
- Panel models therefore consist of a collection of independent stochastic processes, generally linked through shared parameters while also having unit-specific parameters.
- We develop a framework for inference on panel data using nonlinear, partially observed panel processes.
- Our methodology extends iterated filtering techniques for likelihood-based inference from nonlinear time series via partially observed Markov process models.
- We address inferential and computational issues arising due to the high dimensionality of panel data (relative to time series data).

Previous methodology

- Practical Markov chain Monte Carlo (MCMC) methodology for panel data is a research topic (Chen et al., 2016)
- Scaling MCMC and Monte Carlo Expectation-Maximization (MCEM) is a challenge.
- Numerical methods such as expectation propagation (EP) (Gelman et al., 2014) and variational Bayes (Hoffman et al., 2013) are effective for some model classes, but are not readily applicable to general nonlinear mechanistic models.

PanelPOMP models

- Each unit, $u \in 1:U$, in the panel has a corresponding POMP model with latent process $\{X_u(t), t_{u,0} \le t \le T_u\}$ and observable process $Y_{u,1:N_u} = \{Y_{u,1}, \ldots, Y_{u,N_u}\}$.
- $Y_{u,n}$ models data $y^*_{u,n}$ collected on unit u at time $t_{u,n}$, with $t_{u,0} \le t_{u,1} < \cdots < t_{u,N_u} \le T_u$.
- The POMP specification requires that $\{X_u(t)\}$ is a Markov process and $Y_{u,n}$ is conditionally independent of $\{X_u(t)\}$ and $\{Y_{u,k}, k \ne n\}$, given $X_u(t_n)$.
- We write $X_{u,n} = X_u(t_{u,n})$.
- A PanelPOMP is required to have $\big(X_{u,0:N_u}, Y_{u,1:N_u}\big)$ independent of $\big\{\big(X_{v,0:N_v}, Y_{v,1:N_v}\big) : v \ne u\big\}$
- The likelihood for unit u is $\ell_u(\theta) = f_{Y_{u,1:N_u}}(y^*_{u,1:N_u}\,;\theta)$.
- The likelihood for the entire panel is $\ell(\theta) = \prod_{u=1}^{U} \ell_u(\theta)$.

Toward iterated filtering for PanelPOMPs

- A PanelPOMP model can be represented as a POMP model.
- Three different POMP representations of a PanelPOMP were noted by Romero-Severson et al. (2015).
- Deciding which representation to use involves understanding how SMC and IF2 scale with the dimension of the latent variable and the length of the data.
- We must address a **curse of dimensionality**: SMC generally requires exponentially many particles as the latent variable dimension grows (Bengtsson et al., 2008).

Writing a PanelPOMP as a POMP: Representation R1

- For a panel in which each unit is observed over the same time interval, we can write $X^{[R1]}(t) = \big(X_1(t), X_2(t), \ldots, X_U(t)\big)$.
- This constructs a POMP by concatenating the latent state vectors for each separate unit of the PanelPOMP.
- The dimension of the resulting latent process increases with the number of panel units, U.
- This representation is therefore useful for SMC based methods only when U is small.

Writing a PanelPOMP as a POMP: Representation R2

- We can define an equivalent integer-time POMP model,
  $X^{[R2]}(u) = \big(X_{u,0}, X_{u,1}, \ldots, X_{u,N_u}\big), \quad u \in 1:U,$
  concatenating all observations on unit i into a single vector-valued measurement at "time" i.
- For reasons of dimensionality, this representation is useful for SMC based methodology only when $N_1, \ldots, N_U$ are small.
- The dynamics in this POMP model are trivial: $X^{[R2]}(i)$ is independent of $X^{[R2]}(j)$ for $i \neq j$.
- General POMP methodology must be able to work with general latent variable models, so may be useful even with trivial dynamics.
- Representation R2 can provide a simple way to apply existing POMP methodology to panel data: it was adopted by Romero-Severson et al. (2015) for that reason.

Writing a PanelPOMP as a POMP: Representation R3

- We can concatenate the time series for each unit,
  $X^{[R3]}(t) = X_u\left(t - t_{u,0} - T_{u-1}^{\mathrm{cum}}\right)$ for $T_{u-1}^{\mathrm{cum}} \le t < T_u^{\mathrm{cum}}$,
  where T_u^{cum} is the cumulative latent process time for all panels up to unit u, given by
  $$T_u^{\mathrm{cum}} = \sum_{k=1}^{u} \left(T_k - t_{k,0}\right).$$
- Representation R3 is appropriate for SMC based methodology, since under general **mixing** conditions SMC behaves favorably for long time series.

Panel Iterated Filtering (PIF)

- Denote by PIF an application of IF2 to Representation R3 of a PanelPOMP.
- PIF is implemented in the R package `panelPomp` (Bretó et al., 2017) which extends the package `pomp` (King et al., 2016).
- For larger datasets, it is helpful to use Monte Carlo adjusted profiles (see Lecture 6).

Unit-specific fixed effects and random effects

- The parameter vector θ may consist of a component shared between units, θ^{shared}, and a collection of components specific to each unit, $\{\theta_u^{\text{unit}}\}$.
- For the sexual contacts model below, there are no unit-specific parameters. Instead, each unit (individual) has some randomly drawn time-constant characteristics which we call **unit-specific random effects**.
- Unit-specific parameters are called **unit-specific fixed effects**.
- Fixed and random effects each have distinct numerical challenges.
  - Random effects are non-mixing components of the dynamic model. Therefore, they can lead to filtering difficulties if each panel member is not short. For long time series, fixed effects may be preferable.
  - Fixed effects result in the dimension of the parameter space growing with the number of units. This adds numerical difficulties to likelihood maximization, solved by coordinate descent combined with Monte Carlo profile methods that are robust to imperfect maximization.

Scaling panel iterated filtering (PIF)

- For a large panel dataset, Monte Carlo error in the likelihood will necessarily be large.
  - Monte Carlo adjusted profile (MCAP) methodology is useful—see lecture 5.
- When there are many unit-specific parameters, the dimension of the parameter space can become large.
  - This high dimensionality increases Monte Carlo error in maximization. Maximization error is distinct from likelihood evaluation error, but both are addressed by MCAP methodology.
  - Iterations of PIF can be applied marginally to each unit, with the shared parameters fixed. This coordinate ascent adds numerical stability to the estimation of unit-specific parameters.
- There are not yet theorems explaining the practical successes of these scaling approaches.

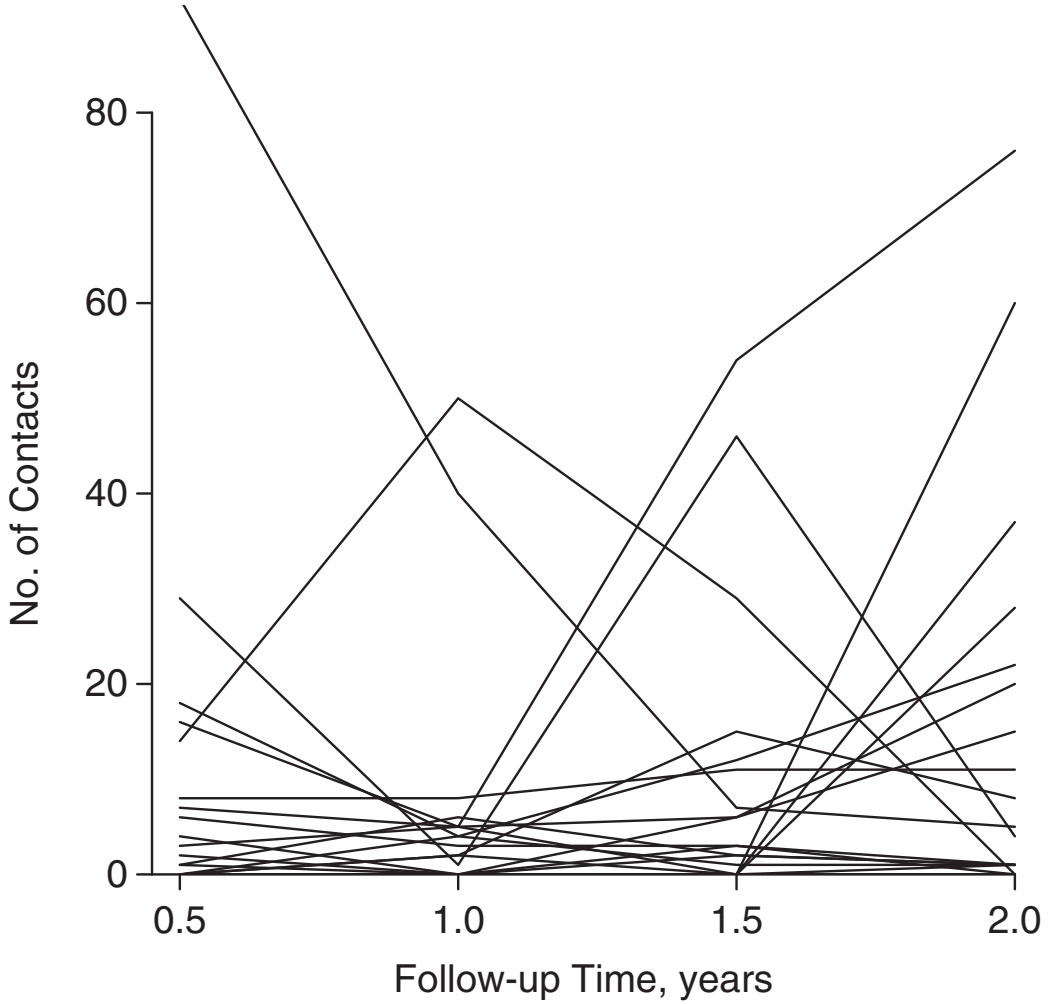
HIV: dynamic variation in sexual individual contact rates

- Basic epidemiological models suppose equal contact rates for all individuals in a population (Keeling and Rohani, 2009).
- Sometimes these models are extended to permit heterogeneity between individuals. Heterogeneity within individuals over time has rarely been considered
- Models have indicated that dynamic changes in sexual contact behavior, known as **episodic risk**, may play a substantial role in the HIV epidemic (Zhang et al., 2012).
- That raises the question: Is episodic risk a real, measurable phenomenon?
- To address this, Romero-Severson et al. (2015) developed an approach to quantify dynamic changes in sexual behavior, by fitting a model for dynamic variation in sexual contact rates to panel data from a large cohort study of HIV-negative gay men (Vittinghoff et al., 1999).

Mechanistic models vs phenomenological models

- For behavioral studies, we interpret "mechanistic model" broadly to mean a mathematical model describing phenomena of interest via interpretable parameters.
- In this context, we want a model that can describe:
  1. Differences between individuals.
  2. Differences within individuals over time;
  3. Flexible relationships between mean and variance of contact rates.
- Romero-Severson et al. (2015) developed a PanelPOMP model capturing these phenomena.

Total sexual contacts in 6 month intervals



- Time series for 15 units from a panel of 882 gay men who completed a 2 year longitudinal study (Romero-Severson et al., 2015).
- Sexual contacts were reported in various categories: oral, anal, protected, unprotected, etc. Here, we show total reported contacts.

A model with dynamic variation in sexual behavior

- Suppose that each individual $u \in 1:U$ has a latent rate $X_u(t)$ of making a sexual contact.
- Each data point, $y^*_{u,n}$, is the number of reported contacts for individual u between time t_{n-1} and t_n, for $u \in 1:U$ and $n \in 1:N$.
- The unobserved process $\{X_u(t)\}$ is connected to the data through the expected number of contacts for individual u in reporting interval n,

$$C_{u,n} = \alpha^{n-1} \int_{t_{n-1}}^{t_n} X_u(t)\, dt,$$

- α is a secular trend that accounts for the observed decline in contacts.
- A basic Poisson model for homogeneous count data has conditional mean and variance of $Y_{u,n}$ equal to $C_{u,n}$ (Keeling and Rohani, 2009).
- Here, the variance in the data are much higher than the mean.
- Negative binomial processes provide a route to modeling dynamic over-dispersion (Bretó and Ionides, 2011).

Modeling dynamic behavior, continued

- We suppose that $Y_{u,n}|C_{u,n}, D_u \sim \text{NegBin}\ (C_{u,n}, D_u)$, a conditional negative binomial distribution with mean $C_{u,n}$ and variance $C_{u,n} + C_{u,n}^2/D_u$.
- D_u is the dispersion random effect, with the Poisson model being recovered in the limit as D_u becomes large.
- D_u can model increased variance compared to the Poisson distribution for individual contacts, but does not result in autocorrelation between measurements on an individual over time.
- To model autocorrelation, we let individual u have behavioral episodes within which $X_u(t)$ is constant. Individual u enters a new behavioral episode at rate R_u.
- For each episode, $X_u(t)$ takes a new value drawn from a Gamma distribution with mean μ_X and standard deviation σ_X, $X_u(t) \sim \text{Gamma}(\mu_X, \sigma_X)$.
- To complete the model, we also assume Gamma distributions for D_u and R_u, $D_u \sim \text{Gamma}(\mu_D, \sigma_D)$, $R_u \sim \text{Gamma}(\mu_R, \sigma_R)$.

Modeling dynamic behavior, summary

- The model has a parameter vector $\theta = (\mu_X, \sigma_X, \mu_D, \sigma_D, \mu_R, \sigma_R, \alpha)$

μ_X	mean contact rate for an episode
σ_X	SD of contact rate for an episode
μ_D	mean individual overdispersion parameter
σ_D	SD of individual overdispersion
μ_R	mean individual rate of new episodes
σ_R	SD of individual rate of new episodes
α	secular trend shared by all individuals

- One might suspect that some of these parameters may be only weakly informed by the data. That is an empirical question, resolved while investigating the likelihood surface and constructing profile likelihood functions.

Results

- We can estimate parameters by maximum likelihood, implemented as the maximum of a smoothed Monte Carlo profile.
- Confidence intervals are constructed via a cutoff on this Monte Carlo profile.
- Models can be compared by likelihood ratio tests or Akaike's information criterion (AIC).
- We find strong statistical evidence for $\sigma_D \neq 0$, $\sigma_X \neq 0$ and $\mu_R \neq 0$. The data were consistent with $\sigma_R = 0$.
- To assess **practical significance** rather than statistical significance, we see if the estimated behavioral heterogeneity parameters can contribute to resolve the paradox that known per-contact transmission struggles to explain the observed HIV epidemic.
- We include these heterogeneities in a simple SI population model: individuals enter a population, pass from susceptible (S) to infected (I) and subsequently leave.

Consequences in a simple epidemic model

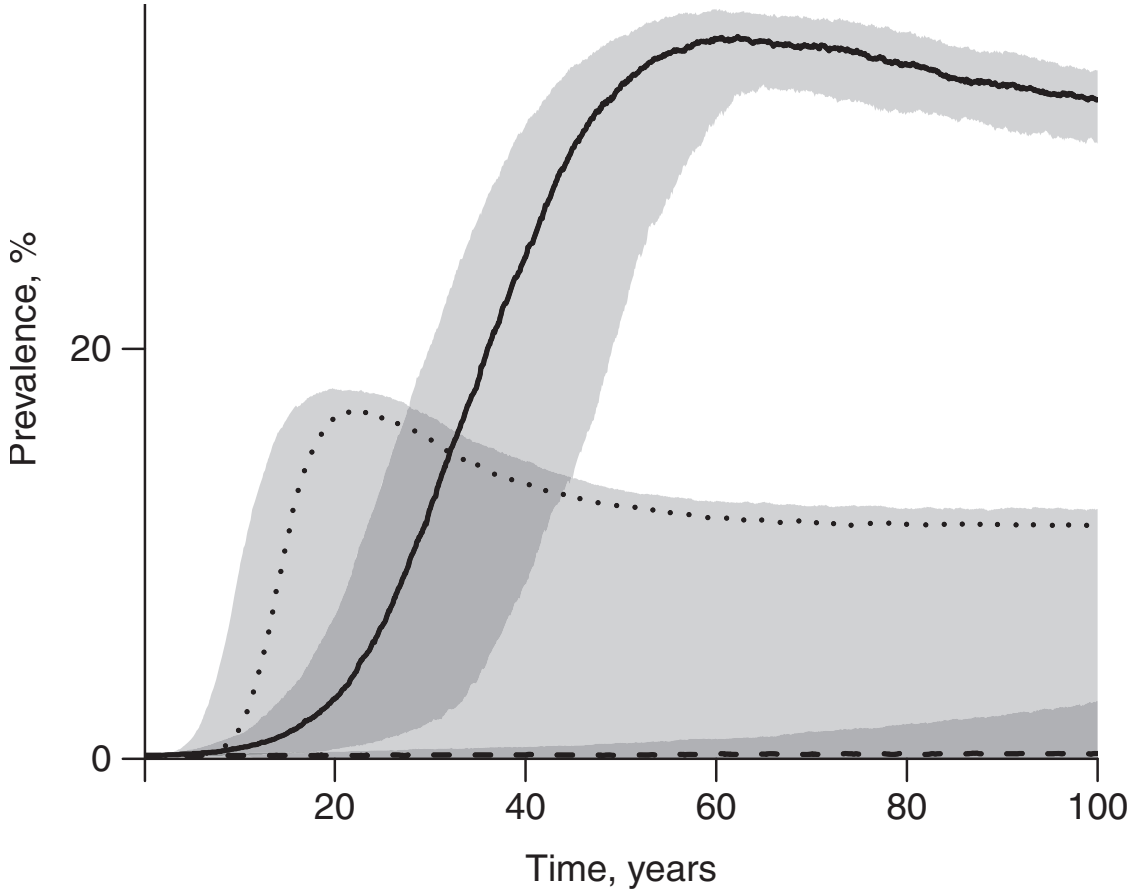


Fig 4 of Romero-Severson et al. (2015). Lines show the median of 500 simulations; gray envelopes are 75^{th} and 25^{th} quantiles. Parameters other than behavioral heterogeneity were fixed.

Dashed line: Homogeneous behavior, $\mu_X = 1.53\,\mathrm{mon}^{-1}$, $\sigma_X = \mu_R = 0$.

Dotted line: Heterogeneity between individuals, $\mu_X = 1.53\,\mathrm{mon}^{-1}$, $\sigma_X = 3.28\,\mathrm{mon}^{-1}$, $\mu_R = 0$.

Solid line: Heterogeneity within and between individuals, behavioral parameters set to the maximum likelihood estimate for total contacts.

References I

Bengtsson, T., Bickel, P., and Li, B. (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Speed, T. and Nolan, D., editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, Beachwood, OH.

Bretó, C. and Ionides, E. L. (2011). Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, 121:2571–2591.

Bretó, C., Ionides, E. L., and King, A. A. (2017). Panel data analysis via mechanistic models. *In preparation*.

Chen, Y., Shen, K., Shan, S.-O., and Kou, S. (2016). Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *Journal of the American Statistical Association*, 111(515):951–966.

References II

Gelman, A., Vehtari, A., Jylänki, P., Robert, C., Chopin, N., and Cunningham, J. P. (2014). Expectation propagation as a way of life. *ArXiv:1412.4869*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Keeling, M. and Rohani, P. (2009). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, NJ.

King, A. A., Nguyen, D., and Ionides, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, 69:1–43.

Romero-Severson, E., Volz, E., Koopman, J., Leitner, T., and Ionides, E. (2015). Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *American Journal of Epidemiology*, 182:255–262.

References III

Vittinghoff, E., Douglas, J., Judon, F., McKiman, D., MacQueen, K., and Buchinder, S. P. (1999). Per-contact risk of human immunodificiency virus tramnsmision between male sexual partners. *American Journal of Epidemiology*, 150(3):306–311.

Zhang, X., Zhong, L., Romero-Severson, E., Alam, S. J., Henry, C. J., Volz, E. M., and Koopman, J. S. (2012). Episodic HIV risk behavior can greatly amplify HIV prevalence and the fraction of transmissions from acute HIV infection. *Statistical Communications in Infectious Diseases*, 4:1–25.