Chapter 8. Case study: Panel data on dynamic variation in sexual contact rates

Objectives

1. Discuss the use of partially observed Markov process (POMP) methods for panel data, also known as longitudinal data.
2. See how POMP methods can be used to understand the outcomes of a longitudinal behavioral survey on sexual contact rates.
3. Introduce the R package **panelPomp** that extends **pomp** to panel data.

Introduction to panel data

- Panel data consist of a collection of time series having no dynamic coupling.
- Each time series is called a **unit**
- If each unit contain insufficient information to estimate model parameters, we infer **shared parameters** by pooling across the whole panel.
- We may have **Unit-specific parameters**, taking distinct values for each unit.
- The goals of developing, fitting and criticizing mechanistic models for panel data are similar to analysis of a single time series.

Heterogeneities in sexual contacts

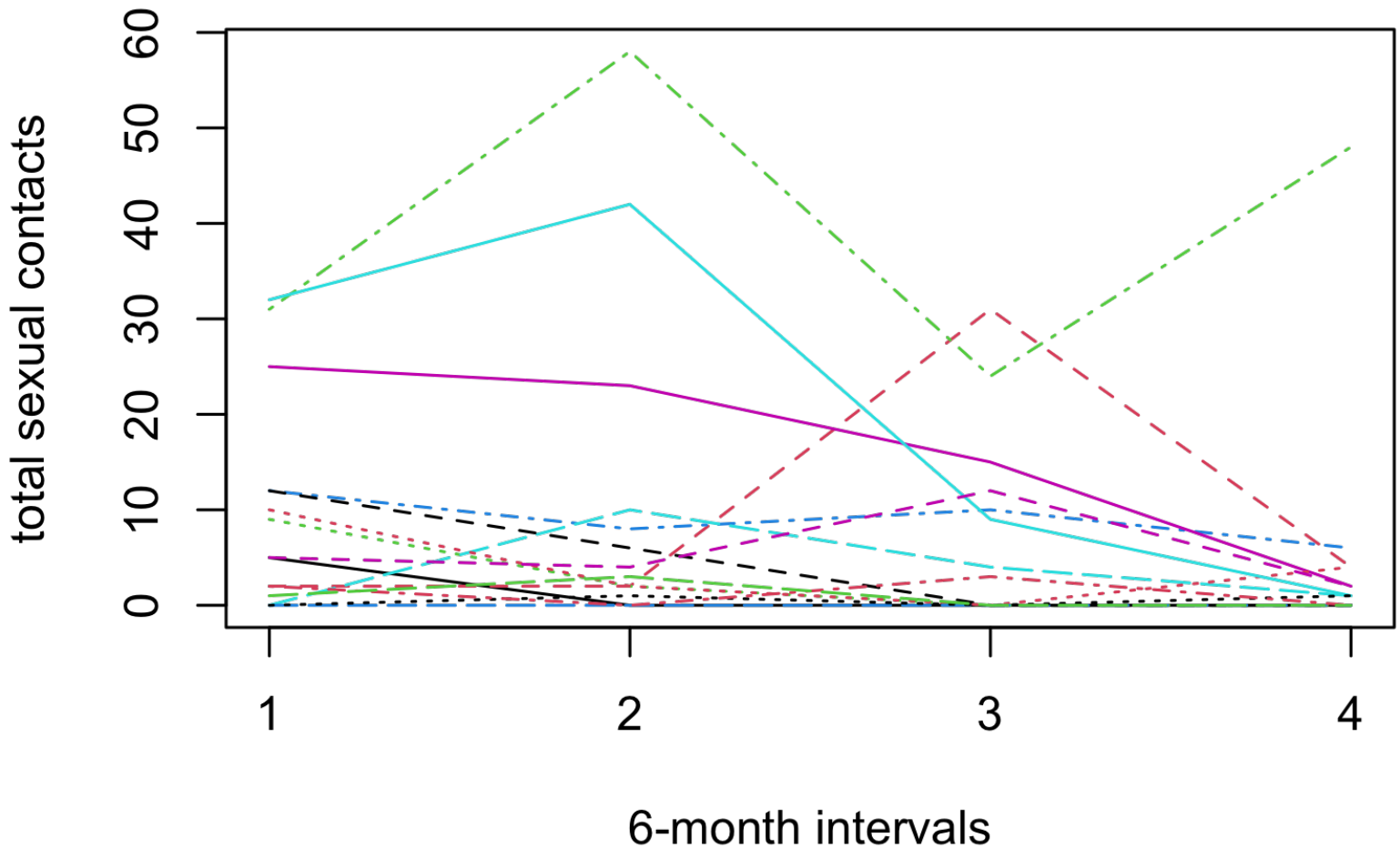
- Basic epidemiological models suppose equal contact rates for all individuals in a population.
- Sometimes these models are extended to permit rate heterogeneity between individuals.
- Rate heterogeneity within individuals, i.e., dynamic behavioral change, has rarely been considered.
- There have been some indications that rate heterogeneity plays a substantial role in the HIV epidemic.

Data from a prospective study

- Romero-Severson et al. (2015) investigated whether dynamic variation in sexual contact rates are a real and measurable phenomenon.
- They analyzed a large cohort study of HIV-negative gay men in 3 cities (Vittinghoff et al.; 1999).
- In a simple model for HIV, with a fully mixing population of susceptible and infected individuals, the fitted variation found by Romero-Severson et al. (2015) can explain the observed prevalence history in the US despite the low per-contact infectivity of HIV.
- Here, we consider the longitudinal data from Vittinghoff et al. (1999) on total sexual contacts over four consecutive 6-month periods, for the 882 men having no missing observations.

- Plotted is a sample of 15 time series from contacts.csv.

```
contact_data <- read.table(file="contacts.csv",header=TRUE)
matplot(t(contact_data[1:15,1:4]),
        ylab="total sexual contacts",xlab="6-month intervals",
        type="l",xaxp=c(1,4,3))
```



Types of contact rate heterogeneity

We want a model that can describe all sources of variability in the data:

1. Differences between individuals
2. Differences within individuals over time
3. Over-dispersion: variances exceeding that of a Poisson model

A model for dynamic variation in sexual contact rates

- We use the model of Romero-Severson et al. (2015), with each individual making contacts at a latent rate $X_i(t)$.
- Each data point, y_{ij}, is the number of reported contacts for individual i between time t_{j-1} and t_j, where $i = 1, \ldots, 882$ and $j = 1, \ldots, 4$.
- The unobserved process $\{X_i(t)\}$ is connected to the data through the expected number of contacts for individual i in reporting interval j, which we write as

$$C_{ij} = \alpha^{j-1} \int_{t_{j-1}}^{t_j} X_i(t)\, dt,$$

where α is an additional secular trend that accounts for the observed decline in reported contacts.

Overdispersion relative to Poisson variation

- A basic stochastic model for homogeneous count data models y_{ij} as a Poisson random variable with mean and variance equal to C_{ij} (Keeling and Rohani; 2009).
- However, the variance in the data are much higher than the mean of the data (Romero-Severson et al.; 2012).
- Therefore, we model the data as negative binomial, a generalization of a Poisson distribution that permits variance larger than the mean:

$$y_{ij} \sim \text{NegBin}\left(C_{ij}, D_i\right),$$

  with mean C_{ij} and variance $C_{ij} + C_{ij}^2/D_i$.
- Here, D_i is called the dispersion parameter, with the Poisson model being recovered in the limit as D_i becomes large.
- The dispersion, D_i, can model increased variance (compared to Poisson variation) for individual contacts, but cannot explain observed autocorrelation between measurements on an individual over time.

Autocorrelation and individual-level effects

- To model autocorrelation, we suppose that individual i has behavioral episodes within which $X_i(t)$ is constant, but the individual enters new behavioral episodes at rate R_i. At the start of each episode, $X_i(t)$ takes a new value drawn from a Gamma distribution with mean μ_X and variance σ_X,

$$X_i(t) \sim \text{Gamma}(\mu_X, \sigma_X).$$

- To complete the model, we also assume Gamma distributions for D_i and R_i,

$$D_i \sim \text{Gamma}(\mu_D, \sigma_D),$$

$$R_i \sim \text{Gamma}(\mu_R, \sigma_R).$$

The parameters, σ_X, σ_D and σ_R control individual-level differences in behavioral parameters allowing the model to encompass a wide range of sexual contact patterns.

Parameter interpretation and identifiability

- The distinction between the effects of the rate at which new behavioral episodes begin, R_i, and the dispersion parameter, D_i, is subtle since both model within-individual variability.
- The signal in the data about distinct behavioral episodes could be overwhelmed by a high variance in number of reported contacts resulting from a low value of D_i.
- Whether the data are sufficient to identify both R_i and D_i is an empirical question.

Consequences of dynamic behavior in an SI model for HIV

- 3 cases where contact rates are either (a) constant; (b) vary only between individuals; (c) vary both between and within individuals.
- In each case, parameterize the model by fitting the behavioral model above, and supplying per-contact infection rates from the literature.
- This simple model shows a potential role for dynamic variation.

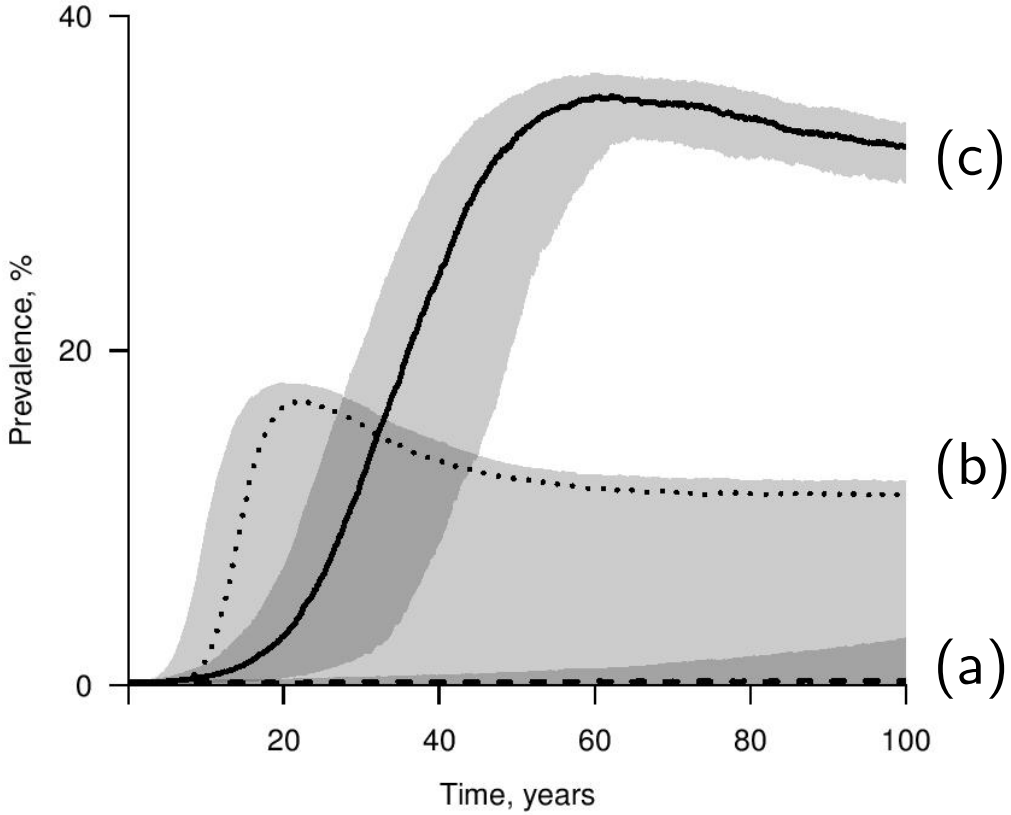


Fig 4 of Romero-Severson et al. (2015). The median of 500 simulations are shown as lines and the 75^{th} and 25^{th} quantiles are shown as gray envelopes.

- 'Homogeneous' (dashed line): the epidemic was simulated where μ_X is estimated by the sample mean (1.53 month^{-1}) without any sources of between-individual or within-individual heterogeneity.
- 'Between Heterogeneity' (dotted line): the epidemic was simulated where μ_X is estimated by the sample mean (1.53 month^{-1}) and σ_X is estimated by the sample standard deviation (3.28 month^{-1})
- 'Within+Between Heterogeneity' (solid line): the epidemic was simulated where each parameter is set to the estimated maximum likelihood estimate for total contacts.
- For all situations, the per contact probability of transmission was set to 1/120, the average length of infection was set to 10 years, and the infection-free equilibrium population size was set to 3000. The per contact probability was selected such that the basic reproduction number in the the 'Homogeneous' case was 1.53. In the 'Homogeneous', 'Between Heterogeneity', 'Within+Between Heterogeneity' cases respectively 239/500 and 172/500, 95/500 simulations died out before the 100 year mark.

PanelPOMP models as an extension of POMP models

- A PanelPOMP model consists of independent POMP models for a collection of **units**.
- The POMP models are tied together by shared parameters.
- Here, the units are individuals in the longitudinal survey.
- In general, some parameters may be **unit-specific** (different for each individual) whereas others are **shared** (common to all individuals).
- Here, we only have shared parameters. The heterogeneities between individuals are modeled as **random effects** with distribution determined by these shared parameters.
- **pomp** methods were extended to PanelPOMP models by Bretó et al. (2019).

Using the **panelPomp** R package

- The main task of **panelPomp** beyond **pomp** is to handle the additional book-keeping necessitated by the unit structure.
- PanelPOMP models also motivate methodological developments to deal with large datasets and the high dimensional parameter vectors that can result from unit-specific parameters.
- A `panelPomp` object for the above contact data and model is provided by `pancon` in **panelPomp**.

```
library(panelPomp)
contacts <- panelPompExample(pancon)
```

- The implementation of the above model equations in `contacts` can be found in the `panelPomp` source code on github.

- Let's start by exploring the `contacts` object

```
class(contacts)

## [1] "panelPomp"
## attr(,"package")
## [1] "panelPomp"

slotNames(contacts)

## [1] "unit.objects" "shared"       "specific"

class(unitobjects(contacts)[[1]])

## [1] "pomp"
## attr(,"package")
## [1] "pomp"
```

- We see that an object of class `panelPomp` is a list of `pomp` objects together with a parameter specification permitting shared and/or unit-specific parameters.
- The POMP models comprising the PanelPOMP model do not need to have the same observation times for each unit.

Exercise 8.1. Suppose a PanelPOMP model has all its parameters unit-specific. Is there anything useful to be gained from the PanelPOMP structure, or is it preferable to analyze the data as a collection of POMP models?

Exercise 8.2. Methods for panelPomps.

- How would you find the **panelPomp** package methods available for working with a `panelPomp` object?

Worked solution

Likelihood evaluation for panelPomps

- PanelPOMP models are closely related to POMPs, and particle filter methods remain applicable.
- `contacts` contains a parameter vector corresponding to the MLE for total contacts reported by Romero-Severson et al. (2015):

```
coef(contacts)
```

```
##    mu_X sigma_X    mu_D sigma_D    mu_R sigma_R   alpha
##    1.75    2.67    3.81    4.42    0.04    0.00    0.90
```

- `pfilter(contacts,Np=1000)` carries out a particle filter computation at this parameter vector.

Exercise 8.3. Consider what happens when we `pfilter` a `panelPomp`.

- Describe what you think `pfilter(contacts,Np=1000)` should do.
- Hypothesize what might be the class of the resulting object? What slots might this object possess?
- Check your hypothesis.

Worked solution

Replicated likelihood evaluations

- As usuual for Monte Carlo calculations, it is useful to replicate the likelihood evaluations, both to reduce Monte Carlo uncertainty and (perhaps more importantly) to quantify it.

```
stew("pfilter1.rda",{
  pf1_results <- foreach(i=1:20) %dopar% {
    library(pomp)
    library(panelPomp)
    pf <- pfilter(contacts,Np= if(DEBUG) 10 else 2000)
    list(logLik=logLik(pf),
      unitLogLik=sapply(unitobjects(pf),logLik))

  }
})
```

- This took 0.6 minutes using 2 cores.

Combining Monte Carlo likelihood evaluations for PanelPOMPs

- We have a new consideration not found with pomp models. Each unit has its own log likelihood arising from an independent Monte Carlo computation.
- The basic pomp approach remains valid:

```
loglik1 <- sapply(pf1_results,function(x) x$logLik)
logmeanexp(loglik1,se=T)

##                           se
## -9555.9363699     0.7327691
```

- Can we do better, using the independence of units? It turns out we can (Bretó et al.; 2019).

logmeanexp versus panel_logmeanexp

```
pf1_loglik_matrix <- sapply(pf1_results,function(x) x$unitLogLik)
panel_logmeanexp(pf1_loglik_matrix,MARGIN=1,se=T)
```

```
##                          se
## -9555.1035400     0.6107453
```

- The improvement via `panel_logmeanexp` is small in this case, since the number of observation times is small.
- For longer panels, the difference becomes more important.

Exercise 8.4. The difference between `panel_logmeanexp` and `logmeanexp`

- The basic `pomp` approach averages the Monte Carlo likelihood estimates after aggregating the likelihood over units.
- The `panel_logmeanexp` averages separately for each unit before combining.
- Why does the latter typically give a higher log likelihood estimate with lower Monte Carlo uncertainty?
- Either reason at a heuristic level or (optionally) develop a mathematical argument.

Worked solution

Writing a PanelPOMP as a POMP

- If we can formally write a PanelPOMP as a POMP, we can use methods such as `mif2` for inference.
- We could stack the panel models in different ways to make a large POMP model.
- A naive way to do inference for a PanelPOMP model as a POMP is to let an observation for the POMP be a vector of observations for all units in the PanelPOMP at that time. This gives a high-dimensional observation vector which is numerically intractable via particle filters.
- Instead, we concatenate the panels into one long time series, with dynamic breaks where the panels are glued together.

Likelihood maximization using the PIF algorithm

- The panel iterated filtering (PIF) algorithm of Bretó et al. (2019) applies the IF2 algorithm to a POMP model constructed by concanenating the collection of panels.
- PIF is implemented in **panelPomp** as the `mif2` method for class `panelPomp`.
- Comparing `?panelPomp::mif2` with `?pomp::mif2` reveals that the only difference in the arguments is that the `params` argument for `pomp::mif2` becomes `shared.start` and `specific.start` for `panelPomp::mif2`.
- As an example of an iterated filtering investigation, let's carry out a local search, starting at the current estimate of the MLE.

```
stew("mif1.rda",{
  mif_results <- foreach(i=1:10) %dopar% {
    library(pomp); library(panelPomp)
    mf <- mif2(contacts,
      Nmif = if(DEBUG) 2 else 50,
      Np = if(DEBUG) 5 else 1000,
      cooling.fraction.50=0.1,
      cooling.type="geometric",
      transform=TRUE,
      rw.sd=rw.sd(mu_X=0.02, sigma_X=0.02, mu_D = 0.02, sigma_D=0.0
        mu_R=0.02, sigma_R =0.02, alpha=0.02
      )
    )
    list(logLik=logLik(mf),params=coef(mf))
  }
})
```

- This is a relatively quick search, taking 20.4 minutes.

Some considerations for likelihood evaluations

Similar likelihood evaluation issues arise for **panelPomp** as for pomp.

- The preliminary likelihood estimated as a consequence of running `mif2` and extracted here by `sapply(m2,logLik)` does not correspond to the actual, fixed parameter, model. It is the sequential Monte Carlo estimate of the likelihood from the last filtering iteration, and therefore will have some perturbation of the parameters.
- One typically requires fewer particles for each filtering iteration than necessary to obtain a good likelihood estimate—stochastic errors can cancel out through the filtering iterations, rather than within any one iteration.
- For promising new parameter values, it is desirable to put computational effort into evaluating the likelihood sufficient to make the Monte Carlo error small compared to one log unit.

```
stew("mif1-lik-eval.rda",{
  mif_logLik <-  sapply(mif_results,function(x)x$logLik)
  mif_mle <- mif_results[[which.max(mif_logLik)]]$params
  pf3_loglik_matrix <- foreach(i=1:10,.combine=rbind) %dopar% {
    library(pomp)
    library(panelPomp)
    unitlogLik(pfilter(contacts,shared=mif_mle,Np=if(DEBUG) 50 else
  }
})
panel_logmeanexp(pf3_loglik_matrix,MARGIN=2,se=T)

##                          se
## -9565.7187699     0.8275629
```

- This took 0.3 minutes.

Acknowledgments and License

- This tutorial is prepared for the Simulation-based Inference for Epidemiological Dynamics module at the 12th Annual Summer Institute in Statistics and Modeling in Infectious Diseases, SISMID 2020.
- Previous versions were presented at SISMID by ELI and AAK in 2015–2019. Carles Breto assisted in 2018.
- Produced with R version 4.0.1 and **pomp** version 3.0.1.1.

References I

Bretó, C., Ionides, E. L. and King, A. A. (2019). Panel data analysis via mechanistic models, *Journal of the American Statistical Association* pp. pre–published online.

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

Romero-Severson, E. O., Alam, S. J., Volz, E. M. and Koopman, J. S. (2012). Heterogeneity in number and type of sexual contacts in a gay urban cohort, *Statistical Communications in Infectious Diseases* **4**(1).

Romero-Severson, E. O., Volz, E., Koopman, J. S., Leitner, T. and Ionides, E. L. (2015). Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men, *American Journal of Epidemiology* p. kwv044.

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.