Supplement to "Panel data analysis via mechanistic models"

Carles Breto[1*], Edward L. Ionides[1], Aaron A. King[2,3,4]

1 Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
2 Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA
3 Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan, USA
4 Department of Mathematics, University of Michigan, Ann Arbor, Michigan, USA
∗ E-mail: cbreto@umich.edu

Supplementary Content

S1 Alternative POMP representations of a PanelPOMP **S-2**

S2 Estimators for the likelihood of a PanelPOMP **S-2**

S3 Considerations for likelihood shortfall **S-4**

S4 Model misspecification and model selection **S-5**

S5 Graphs for subsets of the polio and contacts data **S-6**

S6 Algorithmic parameters **S-7**

S1 Alternative POMP representations of a PanelPOMP

Section 2.1 of the main text developed a partially observed Markov process (POMP) representation of a PanelPOMP, which we call construction R1. The following constructions, R2 and R3, provide two alternative ways to write a PanelPOMP as a POMP.

(R2) For a panel in which each unit is observed over the same time interval, we can write

$$X^{[2]}(t) = \big(X_1(t), X_2(t), \dots, X_U(t)\big).$$

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. Sequential Monte Carlo (SMC) methods struggle with high-dimensional latent processes (Bengtsson et al. 2008). This representation is therefore anticipated to be useful for SMC based methodology only when U is small.

(R3) The latent process for a PanelPOMP model need only be specified at the observation times. Therefore, we can define an equivalent integer-time POMP model,

$$X^{[3]}(u) = \big(X_{u,0}, X_{u,1}, \dots, X_{u,N_u}\big).$$

The dynamics in this POMP model are trivial: $X^{[3]}(i)$ is independent of $X^{[3]}(j)$ for $i \neq j \in 1\!:\!U$. Due to the 'curse of dimensionality' for importance sampling, this representation is useful for SMC based methodology only when all of $N_1, \dots, N_U$ are small. This representation can provide a simple way to apply existing POMP methodology to panel data, and for that reason it was adopted by Romero-Severson et al. (2015).

The only reasons of which we are aware to give preference to R2 or R3 over R1 are small potential gains in conceptual and coding simplicity. However, the scaling difficulties faced by both R2 and R3 make them inappropriate for general-purpose methodology and software based on SMC.

S2 Estimators for the likelihood of a PanelPOMP

Consider $R \geq 2$ independent particle filters, each with $J \geq 1$ particles, which give independent Monte Carlo likelihood estimators $L_u^{(r)}, r \in 1\!:\!R$, for each unit. We work with a constant parameter value θ and write $\ell_u(\theta) = \ell_u$. The Monte Carlo estimator is unbiased and has finite variance, written as

$$\mathbb{E}[L_u^{(r)}] = \ell_u, \qquad \mathrm{Var}[L_u^{(r)}] = \sigma_u^2 < \infty.$$

A corresponding estimator of the full panel likelihood based on replication r is

$$L^{(r)} = \prod_{u=1}^{U} L_u^{(r)},$$

which has mean and variance given by

$$\mathbb{E}[L^{(r)}] = \ell\,, \qquad\qquad \mathrm{Var}[L^{(r)}] = \prod_{u=1}^{U}\Big\{\sigma_u^2 + \ell_u^2\Big\} - \ell^2.$$

A natural approach to combining the R independent likelihood estimators for estimation of the likelihood of a single panel unit u is

$$\bar{L}_u = \frac{1}{R}\sum_{r=1}^{R} L_u^{(r)},$$

which has mean and variance given by

$$\mathbb{E}[\bar{L}_u] = \ell_u\,, \qquad\qquad \mathrm{Var}[\bar{L}_u] = R^{-1}\sigma_u^2.$$

However, it is not immediately clear how to combine the unit-level likelihood estimators to estimate the likelihood of the entire panel. We consider two estimators,

$$\widetilde{L} = \frac{1}{R}\sum_{r=1}^{R} L^{(r)}, \qquad\qquad \widehat{L} = \prod_{u=1}^{U}\bar{L}_u.$$

While both estimators are unbiased, $\widetilde{L}$ is less efficient than $\widehat{L}$. To see this, consider first their variances,

$$\begin{aligned}
\mathrm{Var}[\widetilde{L}] &= \frac{1}{R}\left[\prod_{u=1}^{U}\Big\{\sigma_u^2 + \ell_u^2\Big\} - \ell^2\right] && \text{(S1)}\\
\mathrm{Var}[\widehat{L}] &= \mathbb{E}\left[\prod_{u=1}^{U}\bar{L}_u\right]^2 - \ell^2 \\
&= \prod_{u=1}^{U}\Big\{\frac{\sigma_u^2}{R} + \ell_u^2\Big\} - \ell^2. && \text{(S2)}
\end{aligned}$$

Expanding the product in (S1) yields

$$\mathrm{Var}[\widetilde{L}] = \frac{1}{R}\left[\sum_{k_{1:U}\in\{0,1\}^U}\left\{\prod_{u=1}^{U}\sigma_u^{2k_u}\ell_u^{2(1-k_u)}\right\} - \ell^2\right]. \qquad \text{(S3)}$$

The term ℓ^2 in (S3) cancels with the summand for $k_{1:U} = (0, 0, \ldots, 0)$, giving

$$\mathrm{Var}[\widetilde{L}] = \sum_{k_{1:U}\in\{0,1\}^U\setminus\{0\}^U}\left\{\frac{1}{R}\prod_{u=1}^{U}\sigma_u^{2k_u}\ell_u^{2(1-k_u)}\right\}. \qquad \text{(S4)}$$

An analogous expression for $\mathrm{Var}[\widehat{L}]$, derived from (S2), is

$$\mathrm{Var}[\widehat{L}] = \sum_{k_{1:U}\in\{0,1\}^U\setminus\{0\}^U}\left\{\left[\prod_{u=1}^{U}\frac{1}{R^{k_u}}\right]\prod_{u=1}^{U}\sigma_u^{2k_u}\ell_u^{2(1-k_u)}\right\}. \qquad \text{(S5)}$$

Comparing equivalent terms in the sums for (S4) and (S5) we see that, supposing that either variance is strictly positive (which, from the unbiasedness of the likelihood estimator, implies that σ_u^2 and $\ell_u(\theta)$ are strictly positive) and given that $R > 1$,

$$\mathrm{Var}[\widehat{L}] < \mathrm{Var}[\widetilde{L}].$$

For a quantitative comparison of $\widetilde{L}$ and $\widehat{L}$, consider the situation with constant likelihood

$$\ell_u = \ell_\bullet$$

and constant variance, $\sigma_u^2 = \sigma^2$. Then,

$$\mathrm{Var}[\widetilde{L}] = \frac{1}{R}\Big[\big(\sigma^2 + \ell_\bullet^2\big)^U - \ell_\bullet^{2U}\Big], \qquad \mathrm{Var}[\widehat{L}] = \Big(\frac{\sigma^2}{R} + \ell_\bullet^2\Big)^U - \ell_\bullet^{2U}. \tag{S6}$$

Further, suppose we are interested in the relative likelihood so we can scale to $\ell_\bullet = 1$. Now, if $R = cU$ for some constant c, then we see from (S6) that $\mathrm{Var}[\widehat{L}]$ is stable as $U \to \infty$ whereas $\mathrm{Var}[\widetilde{L}]$ increases exponentially.

S3 Considerations for likelihood shortfall

The shortfall is expected to be hard to calculate outside of toy problems. To quantify this, we would have to know the actual profile likelihood function, but the motivation for using this methodology is that it provides us with the best available approximation to this function. Nevertheless, theoretical and empirical approaches can give us some relevant insights.

As a profile interval becomes narrower, the variation of the bias across the confidence interval converges to zero since the Monte Carlo optimization and likelihood evaluation problems solved for each profile point become increasingly similar. In the limit when the confidence interval collapses toward a single point, the bias becomes trivially constant over the relevant part of the profile. The interval width for a shared parameter in large panels can be expected to be short, since the accumulated information over a large number of panels is anticipated to result in profiles with tall narrow peak. On the other hand, the level of Monte Carlo noise in small panels can be expected to be closer to that in multivariate time series. For time series analysis, a body of existing empirical analyses (reviewed by Bretó 2018) suggests it is often feasible to apply sufficient computational effort to make Monte Carlo error small, avoiding the need for any Monte Carlo adjustment.

From an empirical perspective, simulation studies can address coverage of the constructed confidence intervals. Simulations can provide a combined assessment of all the assumptions and computational approximations involved in a methodology. We have not focused on this holistic assessment in the current article, since we are primarily concerned with how to compute likelihood-based inferences rather than the topic of statistical properties such as coverage, size and power for likelihood-based inferences. This has led us to present results that check the particular computational issues of profile likelihood shortfall, in a specific case where this can readily be done, rather than considering overall statistical performance.

S4 Model misspecification and model selection

The generality of the PIF algorithm gives the scientist many options for diagnosing model misspecification and developing improved models. Since PIF is not constrained to a particular model, the data analyst is encouraged to consider and compare a wide range of models. Likelihood-based techniques such as Akaike's information criterion and likelihood ratio tests are available for model selection. Models can also be compared by assessing whether simulations from the fitted model capture features of interest.

As PIF builds on sequential Monte Carlo (SMC), existing diagnostic methods for this methodology are available. A widely used SMC diagnostic tool is to compute an effective Monte Carlo sample size for each data point (King et al. 2016). Observations with low effective sample size may be outliers, or may be hard to predict for some other reason. Effective sample size also plays a role for diagnosing successful Monte Carlo convergence, since one of the symptoms of an outlier is that the measurement is rare under the postulated model and so a large Monte Carlo sample is needed to accommodate the unexpected observation.

Finding models with appropriate stochasticity to explain the data can be a critical aspect of effective data analysis. Both the latent process and the measurement model are open to critical assessment and improvement within the general PanelPOMP framework permitted by PIF. Some relevant issues on this are discussed by He et al. (2010) and Bretó & Ionides (2011) in the context of time series analysis.

For panel data, the growing size of datasets can add difficulty to viewing diagnostic plots; sometimes one must look for creative summary statistics of the full set of diagnostics for each data point. The `panelPomp` R package provides a software environment to facilitate model development and diagnostics, extending the capabilities of `pomp` (King et al. 2016).

S5 Graphs for subsets of the polio and contacts data

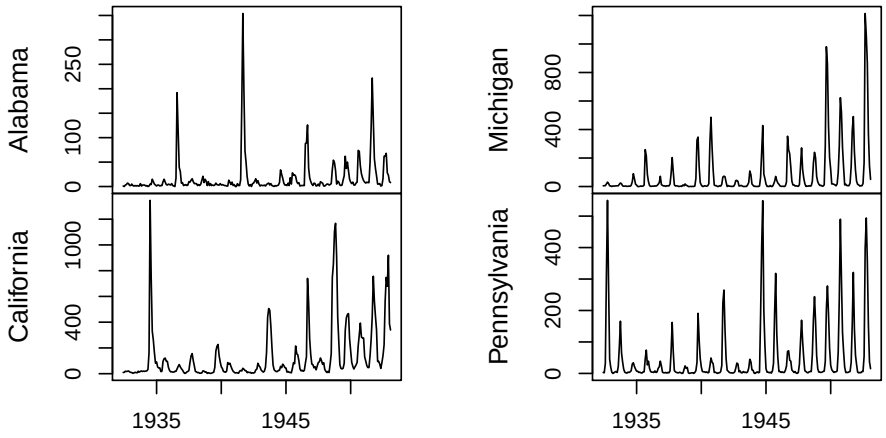

Figure S-1: Selection of 4 time series from the panel dataset from Martinez-Bakker et al. (2015) giving USA monthly of acute paralysis from polio from May 1932 through January 1953 for the 48 continuous US states and Washington D.C. Birth data are missing for South Dakota (before January 1933) and Texas (before January 1934) and so these states were modeled over a reduced time interval. The full data can be accessed from the `panpol` panelPomp object included in the `panelPomp` package.

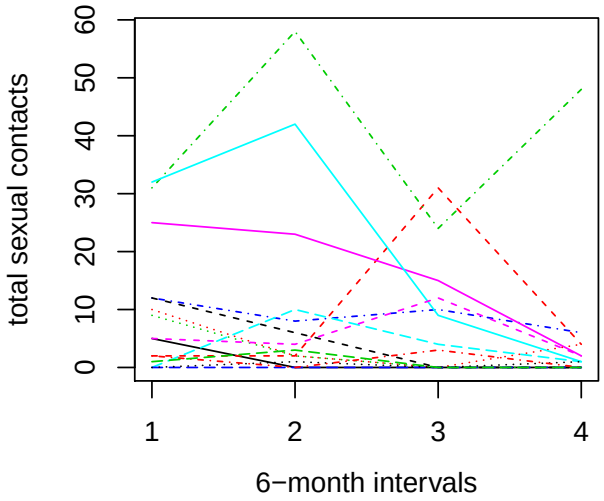

Figure S-2: Sample of 15 time series in the panel dataset from Vittinghoff et al. (1999) on total sexual contacts over four consecutive 6-month periods, for the 882 men having no missing observations. The full data can be accessed from the `pancon` panelPomp object included in the `panelPomp` package.

S6 Algorithmic parameters

The following tables specify all the algorithmic parameters used for our examples. Algorithmic parameters were chosen by assessing diagnostic plots, together with consideration of total run time and quantification of Monte Carlo error in the final results. Diagnosis via convergence plots and effective sample size plots was carried out as described by King et al. (2016).

	Gompertz	polio	polio (MCAP interval)	contacts
J_{pf}	4000	5000	5000	4000
R_{pf}	10	10	10	10
J_{if}	2000	4000	4000	4000
R_{if}	13	19	27	13
M	100	236	236	200
$J_{if,u}$	1000	6000	6000	–
$R_{pf,u}$	4	2	3	–
M_u	50	118	118	–
λ_{loess}	0.9	0.6	0.9	0.9

Table S-1: Algorithmic parameter values used to produce plots in the main text. J_{pf} particles were used for each of R_{pf} replicated particle filter Monte Carlo log likelihood estimates. These replicates were averaged using $\widehat{L}$ (defined in Section S2). The resulting log likelihood estimates correspond to parameter values with maximum likelihood that were reached initializing the joint step of the panel iterated filtering algorithm at R_{if} different parameter starting values ($\Theta^0_{1:J}$ in the description of the PIF algorithm in the main text), running the algorithm for M iterations. These R_{if} convergence points from the joint step were used to initialize $R_{pf,u}$ marginal steps with M_u iterations and $J_{if,u}$ particles. Monte Carlo profiles were obtained by applying loess smoothing to the profile evaluations with smoothing parameter λ_{loess}. A smaller value of λ_{loess} was used for the initial exploratory polio profile than for the MCAP confidence intervals.

	lower bound	upper bound	σ_0	$\sigma_{u,0}$
r	0.05	0.20	0.00125	–
τ_u	0.05	0.20	0.05000	0.05

Table S-2: Starting values, parameter transformations and perturbation specifications for applying PIF to the Gompertz model. The first two columns give the lower and upper bounds of a hyper-rectangle sampled uniformly to generate a value used to initialize all particles $\Theta^0_{1:J}$ for each independent PIF replicate. For the joint maximization, the perturbation sequence used was $\sigma_m = \sigma_0 0.5^{m/50}$. For the marginal maximization, $\sigma_m = \sigma_{u,0} 0.25^{m/50}$ was used instead. These random perturbations were carried out as Gaussian random walks after applying a logarithmic transformation to ensure non-negativity constraints were met.

	lower bound	upper bound	transformation	σ_0	$\sigma_{\mathrm{u},0}$
σ_{dem}	0.0	0.50	log	0.02	–
ψ	0.0	0.10	log	0.02	–
τ	0.0	0.10	log	0.02	–
$b_{u,1}$	-2	8.00	–	0.02	0.02
$b_{u,2}$	-2	8.00	–	0.02	0.02
$b_{u,3}$	-2	8.00	–	0.02	0.02
$b_{u,4}$	1.0	11.0	–	0.02	0.02
$b_{u,5}$	-2	8.00	–	0.02	0.02
$b_{u,6}$	-2	8.00	–	0.02	0.02
$\sigma_{u,\mathrm{env}}$	0.0	1.00	log	0.02	0.02
$\tilde{S}^O_{u,0}$	0.0	1.00	logit	0.10	0.10
$\tilde{I}^O_{u,0} \times 10^4$	0.0	4.00	logit	0.20	0.20

Table S-3: Starting values, parameter transformations and perturbation specifications for applying PIF to the polio model. The first two columns give the lower and upper bounds of a hyper-rectangle sampled uniformly to generate a value used to initialize all particles $\Theta^0_{1:J}$ for each independent PIF replicate. For the joint maximization, the perturbation sequence used was $\sigma_m = \sigma_0 0.5^{m/50}$. For the marginal maximization, $\sigma_m = \sigma_{u,0} 0.25^{m/50}$ was used instead. Some of these random perturbations were carried out as Gaussian random walks after applying a transformation to ensure that non-negativity and unit-interval constraints were met, and these transformations are given in the third column.

	lower bound	upper bound	transformation	σ_0
μ_X	0.80	3.00	log	0.01
σ_X	1.40	5.00	log	0.01
μ_D	1.80	7.00	log	0.01
σ_D	2.00	8.50	log	0.01
α	0.70	0.99	logit	0.01

Table S-4: Starting values, parameter transformations and perturbation specifications for applying PIF to the contacts model. The first two columns give the lower and upper bounds of a hyper-rectangle sampled uniformly to generate a value used to initialize all particles $\Theta^0_{1:J}$ for each independent PIF replicate. For the joint maximization, the perturbation sequence used was $\sigma_m = \sigma_0 0.5^{m/50}$. These random perturbations were carried out as Gaussian random walks after applying the transformations in the third column to ensure that non-negativity and unit-interval constraints were met. Marginal maximization was not applicable in this example since all parameters were shared.

Supplementary References

Bengtsson, T., Bickel, P. & Li, B. (2008), Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, *in* T. Speed & D. Nolan, eds, 'Probability and Statistics: Essays

in Honor of David A. Freedman', Institute of Mathematical Statistics, Beachwood, OH, pp. 316–334.

Bretó, C. (2018), 'Modeling and inference for infectious disease dynamics: a likelihood-based approach', *Statistical Science* **33**(1), 57–69.

Bretó, C. & 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.

He, D., Ionides, E. L. & King, A. A. (2010), 'Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study', *Journal of the Royal Society Interface* **7**, 271–283.

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

Martinez-Bakker, M., King, A. A. & Rohani, P. (2015), 'Unraveling the transmission ecology of polio', *PLoS Biology* **13**(6), e1002172.

Romero-Severson, E., Volz, E., Koopman, J., Leitner, T. & Ionides, E. (2015), 'Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men', *American Journal of Epidemiology* **182**, 255–262.

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