The real-time pandemic influenza model and its implementation(s)

Paul Birrell

September 4, 2017

Introduction

MRC Biostatistics Unit

- Two software packages, obtaining epidemic inference via:
  - MCMC using basic random-walk Metropolis updates.
  - SMC using particle learning *(In development)*.
- In each package, model is implemented with bespoke, standalone C++ code for Bayesian inference from the real-time model.

Additionally...

- R version of the SMC code *(Alice to discuss)*
  - Parallelised using Rmpi.
  - Runs on BSU cluster.
  - Code *slow*, many speed-ups possible.
  - Reduced functionality.

Compilation and Running

MRC Biostatistics Unit

- Within each package, code can be compiled using the command

```
make -f GMakefile
```

- Creates two executable files, one 'fast', one debuggable.

```
rtm_gnu, rtm_gnu_debug          (MCMC)
smc_rtm, smc_rtm_debug          (SMC)
```

- Command line execution (personal preference), e.g.

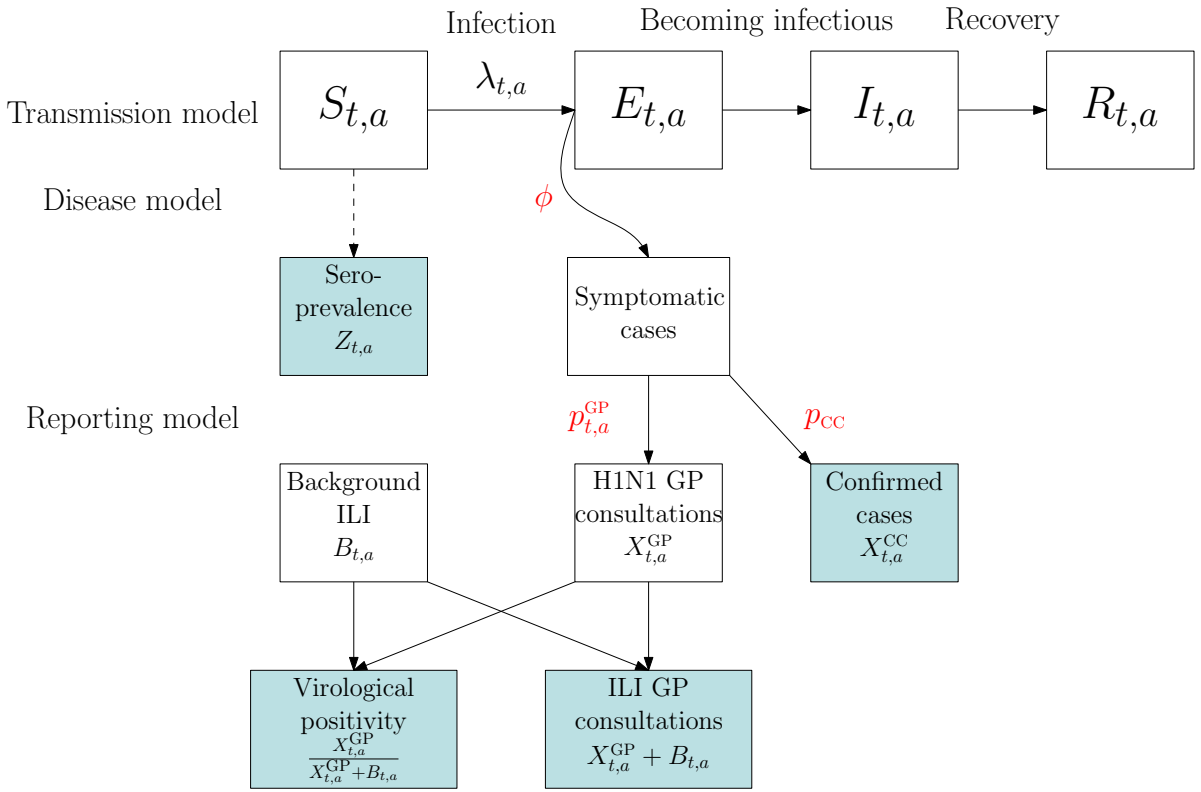
```
nice -19 ./rtm_gnu > outputfile.txt &
```

What does the code do?

MRC Biostatistics Unit

- SEIR model governing transmission.

- Disease model describing disease severity and healthcare seeking behaviours.
- Reporting model. Multiple sources of epidemic data, generated over
  - R parallel regions
  - A strata (usually age-groups)
  - T time points (days/weeks).



Input files

MRC Biostatistics Unit

Unless specified otherwise, the executable will look for two input files `mod_inputs.txt` and `mod_pars.txt`.

`mod_inputs.txt`

Specification of:

- data structure
- likelihood
- regions (population, data files)
- contact model
- MCMC/SMC control

Input files

MRC Biostatistics Unit

Unless specified otherwise, the executable will look for two input files `mod_inputs.txt` and `mod_pars.txt`.

`mod_pars.txt`

Specification of, for each parameter

- initial values (redundant for SMC)
- prior distributions
- dimension
- parameterisation
- proposal parameters

Input files

Unless specified otherwise, the executable will look for two input files `mod_inputs.txt` and `mod_pars.txt`.

Changing the source of inputs

To change the names of either of these input files:

In a file names `smc_rtm_inputs.txt`, set the values of the following variables.

```
model_parameters:  mod_pars_alt.txt
model_inputs:  mod_inputs_alt.txt
```

Would it be preferred if there was the capability to specify these as command line inputs, rather than specified in a .txt file?

Data Structure & Likelihood

MRC Biostatistics Unit

```
GP_consultation_flag 1
Hospitalisation_flag 0
Deaths_flag 0
Sero_data_flag 1
Viro_data_flag 1
gp_count_likelihood 1 ## 0 POISSON, 1 FOR NEGATIVE BINOMIAL DATA.
transmission_kernel 0 ## 0 FOR THE REED-FROST 1 FOR MASS ACTION.
transmission_time_steps_per_day 2
reporting_time_steps_per_day 2
duration_of_runs_in_days 245
duration_of_previous_runs_in_days 164
num_regions 1
num_age_groups 7
Sero_likelihood_start_day 75
Sero_likelihood_end_day 245 ## SPECIFY A SUBSET OF THE DATA TO BE
INCLUDED IN LIKELIHOOD..  USEFUL TO AVOID ZEROS.
```

Green text indicates new options in SMC code

Regions

MRC Biostatistics Unit

```
study_region = { regions_used = London, WestMidlands, North,
South;
regions_population = 119967, 410926, 840309, 989473, 2747554,
1566445, 882256,
68285, 259069, 651598, 730409, 1443924, 1343723, 884837,
227677, 864277, 2217602, 2676783, 5101913, 4823469, 3093895,
224805, 863397, 2251281, 2415331, 5166801, 4927669, 3298357;
```

An $R \times A$ array.

Regions (cont'd)

MRC Biostatistics Unit

```
regions_gp_aggregation = 2, 3, 4, 5, 6;  Ages 1 & 2 should be combined
regions_gp_count_data = Reported_GP_ConsObs_London_28_03_10.txt,
                                          Should contain a T × (A − 1) array.
            Reported_GP_ConsObs_WestMidlands_28_03_10.txt,
            Reported_GP_ConsObs_North_28_03_10.txt,
            Reported_GP_ConsObs_South_28_03_10.txt;
regions_gp_coverage_data = Reported_GP_CovObs_London_28_03_10.txt,
            Reported_GP_CovObs_WestMidlands_28_03_10.txt,
            Reported_GP_CovObs_North_28_03_10.txt,
            Reported_GP_CovObs_South_28_03_10.txt;
}
```

MCMC options - notation

MRC Biostatistics Unit

- The MCMC algorithm aims to sample from a posterior distribution $\pi_t(\theta|\mathbf{x}_t)$.
- Suppose the parameter vector, θ can be partitioned into $(\theta_1, \ldots, \theta_k)$.
- Each parameter component θ_j has a prior distribution $\pi_0(\theta_j)$.
- Start with an initial vector $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$
- At iteration, i, we have $\theta^{(i)}$.
- For $j = 1, \ldots, k$, make a proposal

$$\theta_j^{(*)} \sim N\left(\theta_j^{(i)}, \sigma_i^2 \mathbf{I}\right). \tag{1}$$

- Then accept this proposal as $\theta_j^{(i+1)}$ with the appropriate acceptance probability. Otherwise set $\theta_j^{(i+1)} = \theta_j^{(i)}$.
- Run until $i = N$.

MCMC options (cont'd)

MRC Biostatistics Unit

```
num_iterations 700000
```

N number of iterations.

```
output_type 0
```

0 or 1, for output as MCMC chains, or output as particles

```
thin_output_every 100
```

Stored chains are thinned to every 100^{th} iteration (post burn-in only).

```
thin_stats_every 5000
```

Summary statistics are output every 5000^{th} iteration (post burn-in only).

```
adaptive_phase 100000
```

No. of iterations within the burn-in over which the proposal variances are adaptated.

```
adapt_every 100
```

Proposal variance adaptation occurs every 100^{th} iteration (within the adaptive phase only).

```
burn_in 200000
```

Length of burn-in. All of the first 200000 iterations are discarded.

MCMC options (cont'd)

MRC Biostatistics Unit

```
num_progress_reports 10

mixing_threshold_ub 0.3

mixing_threshold_lb 0.2

prop_var_inflation 1.1

prop_var_shrinkage 0.9

maximum_block_size 9

random_seed 80
max_threads 8
```

Number of updates produced during both the adaptive and posterior phases.

During the adaptive phase, acceptance rates above 0.3 lead to a scale-up of proposal variances

During the adaptive phase, acceptance rates below 0.2 lead to a reduction in proposal variances

Proposal variance scale-up factor

Proposal variance scale-down factor

Maximum number of components of the $\theta^{(j)}$ that are proposed in any one go.

The number of parallel threads used in evaluating the model likelihood.

SMC options - Reweighting

MRC Biostatistics Unit

- We wish to evaluate a posterior density $\pi_{t_1}(\theta|\mathbf{x}_{t_0})$, starting from an initial posterior $\pi_{t_0}(\theta|\mathbf{x}_{t_0})$.
- Code will load in (or generate if $t_0 = 0$) a particle set, $\{\theta^{t_0,1}, \ldots, \theta^{t_0,n}\}$ and weights $\omega_{t_0}^{(1)}, \ldots, \omega_{t_0}^{(n)}$, representing a weighted sample from π_{t_0}.
- At general time $t+1$, assimilate new data in continuous time. This means we consider densities of the form $\pi_{t,\delta}(\theta) = \pi_t(\theta)\{p(\mathbf{x}_{t+1}|\theta)\}^{\delta}$, for $\delta \in [0,1]$. At 'time' $t+\delta$, particle j carries weight

$$\tilde{\omega}_{t+\delta}^{(j)} = \omega_t^{(j)} \left\{ p\left(\mathbf{x}_{t+1}|\theta^{t,j}\right) \right\}^{\delta}$$

.

SMC options - Detecting Impoverishment

MRC Biostatistics Unit

- Define the effective sample size

$$ESS\left(\left\{\omega^{(\cdot)}\right\}\right) = \frac{\left(\sum_{l=1}^{n} \omega^{(l)}\right)^2}{\sum_{j=1}^{n} \omega^{(j)^2}}.$$

- If, for some threshold $\epsilon \in [0, 1]$:
  - $ESS(\{\tilde{\omega}_{t+1}^{(j)}\}) > \epsilon n$. No particle set rejuvenation required. Set $\theta^{t+1,j} = \theta^{t,j}$ and $\omega_{t+1}^{(j)} = \omega_t^{(j)}, \forall j = 1, \ldots, n$.
  - $ESS(\{\tilde{\omega}_{t+1}^{(j)}\}) \leq \epsilon n$. The particle set is impoverished and requires rejuvenation. Pick δ_0, such that $ESS(\{\tilde{\omega}_{t+\delta_0}^{(j)}\}) = \epsilon n$. The particle set is then rejuvenated (see next slide) and weights are reset $\omega_{t+\delta_0}^{(j)} = 1$.
    - Define weights $\tilde{\omega}_{t,\delta_0,\delta}^{(j)} = p(\mathbf{x}_{t+1}|\theta^{t+\delta_0,j})^{\delta-\delta_0}$. Further rejuvenations will be required if this leads to an ESS calculation that once again drops below ϵn while $\delta \leq 1$.

SMC options - Rejuvenation

MRC Biostatistics Unit

Resampling, followed by population MCMC, using proposal distributions weakly dependent upon the starting states (i.e. the particle set requiring rejuvenation).

- Using the weighted sample, make estimates of $\hat{\mu}$ and $\hat{\Sigma}$, the sample mean and covariance.
- Resample the particles according to the particle weights.
- Move the n particles. Effectively running n parallel MCMC chains.. except
  - No requirement to attain convergence - no burn-in phase.
  - No requirement to run long chains to harvest a sample. The parallel chains provide a sample.

SMC options - Rejuvenation (2)

MRC Biostatistics Unit

- **Move***(continued)* Two types of move:
  - Correlated random-walk: Propose

$$\theta^* | \theta_k^{(j)} \sim \mathrm{N}\left(\theta_k^{(j)}, \gamma \bar{\mathbf{\Sigma}}_k\right).$$

  - Approximate Gibbs': Propose

$$\theta^* | \theta_k^{(j)} \sim \mathrm{N}\left(\bar{\theta}_k, \bar{\mathbf{\Sigma}}_k\right)$$

- Moves can be based on the above distributions or their conditionals if not using a full block update.
- Any number of the above types of proposal can be made per iteration, and they can be of a subset of parameter components and they can be full-block or componentwise.
- Parameter γ can be adapted between successive batches.
- *gamma* can also be vector-valued.

SMC options - Rejuvenation (3)

MRC Biostatistics Unit

- **Move***(continued)* Repeat the process until the sample inter-class correlation coefficient, r_A^* reaches some lower threshold.
  - After resampling, $r_A^* = 1$.
  - Provides a measure of how well the chains have collectively forgotten their starting points.
  - Stop making proposals when $r_A^* \leq r_{A,0}$.

SMC options (cont'd)

MRC Biostatistics Unit

`location_particles /saved/`

Where the 'current' particles are stored. Ignored if `duration_of_previous_runs_in_days` *= 0. Defaults to the current directory.*

`num_particles 10000`

Number of particles to use. If saved number is different, this will be the number resampled.

`updaters 110`

3 update steps: 0 for a correlated r-w proposal, 1 for an approx. Gibbs proposal.

`block_updates 110`

One for each of the above. 0 for cpt-wise updates, 1 for a block update

SMC options (cont'd)

MRC Biostatistics Unit

```
ess_threshold 0.5

icc_threshold 0.1

mixing_threshold_ub 0.3

mixing_threshold_lb 0.2

prop_var_inflation 1.1

prop_var_shrinkage 0.9
```

ϵ, the ESS threshold that triggers sample rejuvenation.

$r_{A,0}$, the threshold for stopping the move-step of the rejuvenation.

For correlated r-w updates, acceptance rates above 0.3 lead to a scale-up of γ.

As above, acceptance rates below 0.2 lead to a reduction in γ.

γ scale-up factor

γ scale-down factor

Specifying model parameters - Initial Values

```
prop_symptomatic = { param_value = 0.1, 0.15, 0.2, 0.25;
                   }
```

- Sets initial values for MCMC chain.
- In SMC, redundant, other than to determine the length of the parameter vector - *i.e.* how many parameters govern the proportion symptomatic in the population (over time and ages).

Specifying model parameters - Prior Choice(1)

MRC Biostatistics Unit

```
prop_symptomatic = { param_value = 0.1, 0.15, 0.2, 0.25;
                     prior_distribution = 3, 3, 3, 3;
                     }
```

- Value required for each parameter component.
- If first value indicates a multivariate proposal, it over-rides the rest.
- Values and distributions..
  1. Fixed value.
  2. Gamma distribution.
  3. Beta distribution.
  4. Normal distribution.
  5. Multivariate normal distribution.
  6. Half-normal distribution.
  7. Uniform distribution.

Specifying model parameters - Prior Choice(1)

MRC Biostatistics Unit

```
prop_symptomatic = { param_value = 0.1, 0.15, 0.2, 0.25;
                     prior_distribution = 3, 3, 3, 3;
                     prior_parameters = 0.5, 2, 1, 1, 20, 10, 1,
1;
                     }
```

- Code knows how many parameters each distribution requires.
- Parameters are assigned as a queue - the four parameters components above have prior distributions:
  - $\beta(0.5, 2)$
  - $\beta(1, 1)$
  - $\beta(20, 10)$
  - $\beta(1, 1)$
- A constant parameter component takes no prior parameter values.
- Gamma distribution is in shape/rate parameterisation.
- Multivariate normal $(\mu, \mathbf{\Sigma})$ parameters are specified as a vector $(\mu_1, \Sigma_{11}, \Sigma_{12}, \ldots, \sigma_{1p}, \mu_2, \sigma_{21}, \ldots, \sigma_{pp})$, a $p(p+1)$ length vector.

Specifying model parameters - Parameterisation

MRC Biostatistics Unit

Model parameters are parameters that specify the, for example, proportion symptomatic. We need not be specifying the actual parameter value. Instead the relation is via a generalised linear model set-up.

$$g\left(\underbrace{\theta}_{(r_0 t_0 a_0)\times 1}\right) = \underbrace{\mathbf{X}}_{(r_0 t_0 a_0)\times p}\ \underbrace{\beta}_{p\times 1}$$

- β The parameter vector of length p that we specify in the `mod_pars.txt` file.
- $\mathbf{X}$ A $(r_0 t_0 a_0 \times p)$ design matrix.
- $g(\cdot)$ A link function
- θ A vector of the calculated parameter values, $(\theta_{111}, \theta_{112}, ..., \theta_{r_0 t_0 a_0})$, with, for example, θ_{rta} giving the proportion symptomatic in region r, time t and age a.

Specifying model parameters - Parameterisation(2)

MRC Biostatistics Unit

- Breakpoints: Specify breakpoints to avoid having to specify an $RTA \times p$ design matrix where not necessary (also avoids large matrix multiplication).
- Example: As in regional papers, $R = 4, T = 245, A = 7$. Don't need to always have a matrix with 6860 rows.
- `regional_breakpoints`: true/false. So $r_0 \in \{1, R\}$.
- `age_breakpoints`: true/false or a integer-value vector, taking ascending values in $1, \ldots, A$. If `age_breakpoints = 3`, say, then the first three age-classes can have a different value of the parameter at all times to the older age classes, and $a_0 = 2$. Again, true/false set $a_0 = A$ or 1 respectively.
- `time_breakpoints`: same as `age_breakpoints`, although has no impact for parameters that specify an initial condition, e.g. exponential growth rate.

Specifying model parameters - Parameterisation(3)

```
prop_symptomatic = { param_value = 0.1, 0.15, 0.2, 0.25;
      prior_distribution = 3, 3, 3, 3;
      prior_parameters = 0.5, 2, 1, 1, 20, 10, 1, 1;
      time_breakpoints = 83, 245; Upper bounds for the breakpoints
      age_breakpoints = 3, 7;      are optional and are ignored
      regression_link = 2;
      regression_design = file.design.txt;
      }
```

- `regression_link`: the choice of the function $g(\cdot)$.
  0. $g(x) = x$
  1. $g(x) = \log(x)$
  2. $g(x) = \log(x/(1-x))$
- `regression_design`: the location of a file specifying the design matrix. Elements are read in separated by spaces, in the order X_{11}, $X_{12}, \ldots$. Alternatively, can be specified as a numerical vector here.

Specifying model parameters - SMC/MCMC

MRC Biostatistics Unit

```
prop_symptomatic = { param_value = 0.1, 0.15, 0.2, 0.25;
                     prior_distribution = 3, 3, 3, 3;
                     prior_parameters = 0.5, 2, 1, 1, 20, 10, 1,
1;
                     :
                     proposal_variance = 0.1, 0.1, 0.1, 0.1;
                     update_indicators = 1, 0, 1;
                     }
```

- `proposal_variance` In MCMC, the random-walk proposal variances for each parameter component. In SMC, only the first value is used, it is the value of γ for any component-wise updates.
- `update_indicators` In MCMC, a dummy variable. In SMC, this indicates whether this parameter component should be updated as part of each proposal mechanism, 1 for yes, 0 for no.

Model Parameters - What are they?

MRC Biostatistics Unit

The β's, that specify the θ's

$$g\left(\underbrace{\theta}_{(r_0 t_0 a_0)\times 1}\right) = \underbrace{\mathbf{X}}_{(r_0 t_0 a_0)\times p}\ \underbrace{\beta}_{p\times 1}$$

where θ could represent:

Parameter name	Short Description	Variation
`negbin_overdispersion`	Over-dispersion parameter for Negative Binomially distributed primary care count data.	$r \times t \times a$
`latent_period`	The mean duration of the latent period	$r \times t \times a$
`infectious_period`	The mean duration of the infectious period	$r \times t \times a$
`relative_infectiousness`	The relative infectiousness of state I_2 to state I_1	$r \times t \times a$

Model Parameters - What are they? (2)

MRC Biostatistics Unit

Parameter name	Short Description	Variation
`prop_symptomatic`	Proportion of infections that lead to symptoms (typically ILI).	$r \times t \times a$
`contact_parameters`	Parameters used to scale specified entries of the contact matrices.	$r \times t \times a^*$
`R0_amplitude_kA`	Amplitude of seasonal variation in the value of R_0 (as a fraction of the peak value).	r only.
`R0_seasonal_peakday`	Timing in the year of the peak value of R_0	r only.
`exponential_growth_rate`	Initial rate of growth of the epidemic, proxy of R_0 at epidemic time-0	r only.
`log_p_lambda_0`	Log-initial rate of arrivals in consultation, a proxy for I^* the initial number of infectives.	r only.

*Time and age variation in the contact structure is specified in `mod_inputs.txt`. Any time and age variation specified in `mod_pars.txt` will be quietly ignored.

Model Parameters - What are they? (3)

MRC Biostatistics Unit

Parameter name	Short Description	Variation
`prop_susceptible`	Proportion of individuals susceptible to infection.	$r \times a$.
`prop_HI_32_to_HI_8`	Proportion of immunity that would test positive with a HI of 32 (where immunity is defined by HI 8.	$r \times a$.
`prop_case_to_GP_consultation`	Proportion of symptomatic cases that consult with a GP (or other primary care provider).	$r \times t \times a$
`prop_case_to_hosp`	Proportion of symptomatic cases that will require hospitalisation.	$r \times t \times a$
`prop_case_to_death`	Proportion of symptomatic cases that lead to death.*	$r \times t \times a$

*This data stream currently treated independently of hospitalisations. This may not be reasonable, but it may be possible to use for other count data, such as early case confirmation.

Model Parameters - What are they? (4)

MRC Biostatistics Unit

Parameter name	Short Description	Variation
`importation_rates`	Rate of inflow of new infections.	$r \times t \times a$.
`background_GP`	Rates of background GP consultation (measured in consultations per 100,000 people, per year).	$r \times t \times a$.
`test_sensitivity`	The test sensitivity of the virological swabbing process, taking values in $[0, 1]$.	$t \times a$
`test_specificity`	The test specificity of the virological swabbing process, taking values in $[0, 1]$.	$t \times a$
`day_of_week_effects`	Scaling factors for the expected number of primary care consultations per day.	$r \times t \times a$