Supplementary details on experiments of effective reproduction number estimation with trend filtering

Jiaping Liu, Zhenglun Cai, Paul Gustafson, and Daniel J. McDonald

A.1 Derivation of Kullback Leibler divergence for accuracy comparison

A.2 Supplmentary details on experimental settings

We compare the accuracy of the estimated effective reproduction numbers using Kullback Leibler (KL) divergence (with Poisson distributional assumption on incidence) averaged per coordinate across our `RtEstim` and several alternative methods including `EpiEstim` with weekly and monthly sliding windows, `EpiLPS`, `EpiFilter`, `EpiNow2`, and `RtEstim` with degrees k=0,1,2,3, which yields 9 methods in total. We consider two lengths of epidemics with $n = 50, 300$ timepoints respectively. Since `EpiNow2` runs too long (specifically, for one long epidemic, it takes ???TBA), we only compare it with other methods for short epidemics.

We consider serial interval (SI) distributions of measles and SARS to generate long synthetic epidemics, and flu for short epidemics. Incident cases in synthetic measles epidemics are relatively low (within 1000 at the peak overall), and SARS incident cases are relatively large (between 15000 and 20000 at the peak overall). We specifically consider a reasonably large overdispersion level of negative Binomial incidence is of size 5. Figure A.2.1 displays the ratio of standard deviation over mean (called, sigma to mean ratio) of incidence across different settings using the same set of sample epidemics in the manuscript. Compared to the counterpart of Poisson incidence (which decreases quickly to be under 0.25 and close to 0 at most timepoints) per $\mathcal{R}_t$ scenario for each epidemic, the negative Binomial incidence appears to have an apparently larger sigma to mean ratio (staying at around 0.5 at most timepoints) implying a distinguishable overdispersion level.

In model fitting, we use both true and misspecified serial interval (SI) distributions. The misspecification of serial interval distributions are either mild or major, where, in the major misspecification, we use SI of SARS to solve long measles epidemics and SI of measles to solve short flu epidemics. While, in the mild SI misspecification, we consider a shaped (mean increased by 2) and scaled (standard deviation increased by 10) parameters for both short flu and long measles epidemics, denoted as `flu_ss` and `measles_ss` respectively. These settings result in 7 pairs of SI distributions (for epidemic generating, and for model fitting), i.e., (`measles`, `measles`), (`SARS`, `SARS`), (`measles`, `measles_ss`), (`measles`, `SARS`) for long epidemics and (`flu`, `flu`), (`flu`, `flu_ss`), (`flu`, `measles`) for short epidemics. Figure A.2.2 displays the SI distributions (`measles`, `measles_ss`, `SARS`, `flu`, and `flu_ss`) used in the experiments.

Table 1 summarizes the aforementioned experimental setting for accuracy comparison. Poisson and negative Binomial (NB) distributions for incidence and 4 $\mathcal{R}_t$ scenarios are used for all settings. Each experimental

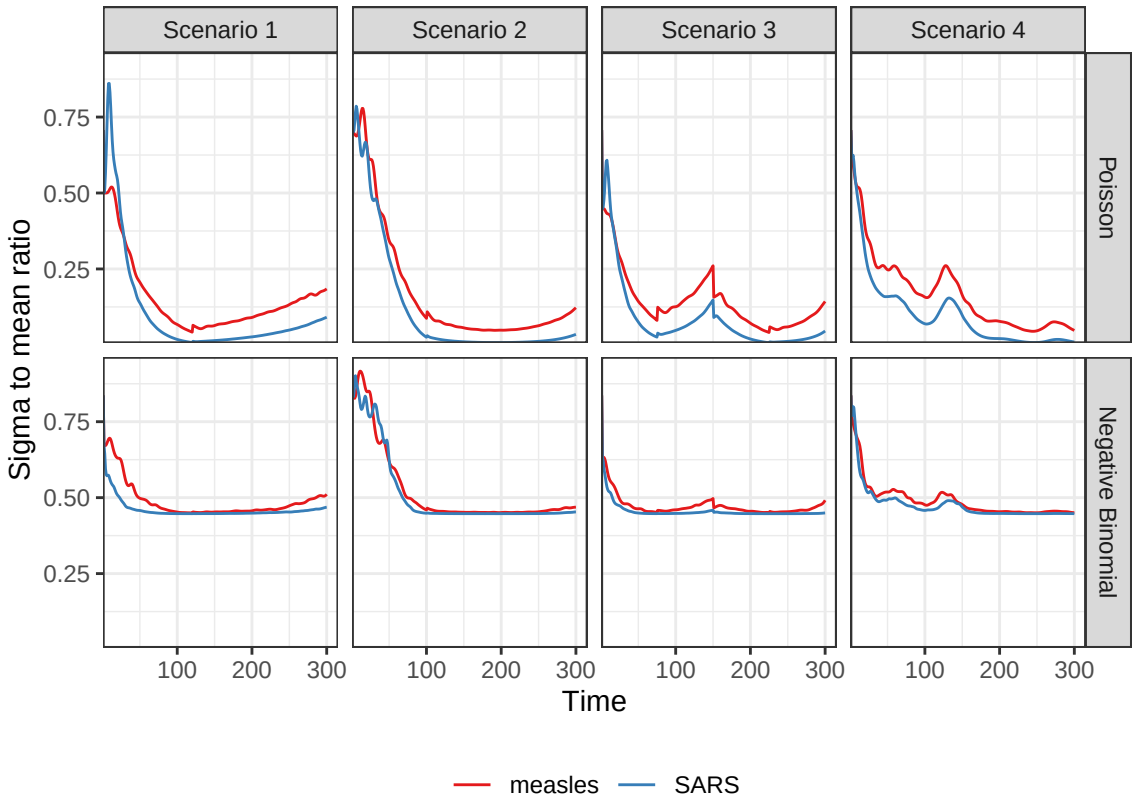


Figure A.2.1: Dispersion level of incidence of sample epidemics

Table 1: Summary of experimental setting on accuracy comparison

Length	SI	Rt scenario	Incidence	SI for modelling	Method
300	measles	1-4	Poisson, NB	measles, measles_ss, SARS	8 methods
300	SARS	1-4	Poisson, NB	SARS	8 methods
50	flu	1-4	Poisson, NB	flu, flu_ss, measles	9 methods

setting is replicated for 50 times, which yields 12800 experiments with long epidemics and 2700 with short epidemics.

Here we list the hyperparameters used in modelling for each method. Most of them are the experimental settings used in the papers where they were proposed and deemed as the "best" tuned ones. We consider both weekly and monthly sliding windows in `EpiEstim`, 40 basis functions in `EpiLPS` with the NelderMead method to maximize the hyperparameter posterior distribution. We input 2000 grid size in `EpiFilter` with 0.1 diffusion noise and uniform prior on $\mathcal{R}_t$ with mean $1/2000$, and use the smoothed $\mathcal{R}_t$ as estimates. We run 10-fold cross validation (CV) to choose the best tuning parameter from the candidate set of size 50, i.e., $\boldsymbol{\lambda} = \{\lambda_1, \cdots, \lambda_{50}\}$. Specifically, we divide all samples (except the first and last entries) into ten folds evenly and randomly, and build models on each sample set by leaving a fold out across all hyperparameters. We select the tuning parameter that gives the lowest averaged ***deviance*** between the estimated incidence and the observed samples averaged over all folds.

We visualize the selected key results of accuracy comparison using long synthetic epidemics in Section 3.2 in the manuscript. Other main experimental results are displayed as follows.

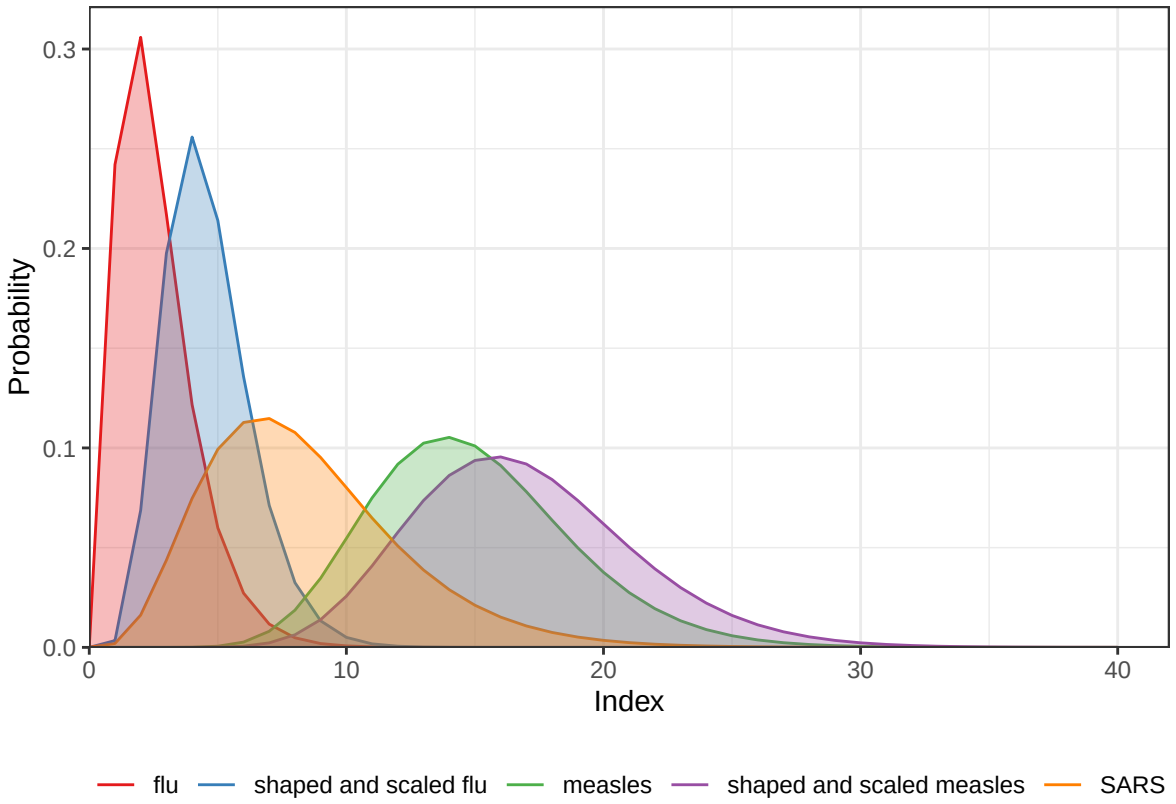


Figure A.2.2: Density curves of serial interval distributions used in the experiments

A.3 Supplementary experimental results on accuracy comparison

Figures A.3.1 and A.3.2 display the KL divergence excluding the first months for measles and SARS epidemics respectively. To compare `EpiEstim` with ***monthly*** sliding windows with other methods, we average the KL divergence per coordinate excluding the timepoints in the first months for all approaches, since `EpiEstim` estimates with the monthly sliding windows are not available until the second months. The y-axis is displayed on a logarithmic scale for a better visualization, since a few values are much larger than others.

The relative performance of `EpiEstim` with monthly sliding windows, in general, is not as good as its weekly sliding window based on the relative positions of its boxes and the counterparts of the other methods, except for the Scenario 2 with negative Binomial incidence. It can be explained by `EpiEstim` with longer sliding windows assume similarity of neighbouring $\mathcal{R}_t$ across longer periods, and thus, is smoother and less accurate compared to the one with shorter sliding windows.

Figures A.3.3 and A.3.4 display the KL divergence for short epidemics averaged per coordinate excluding the first weeks and months resepctively.

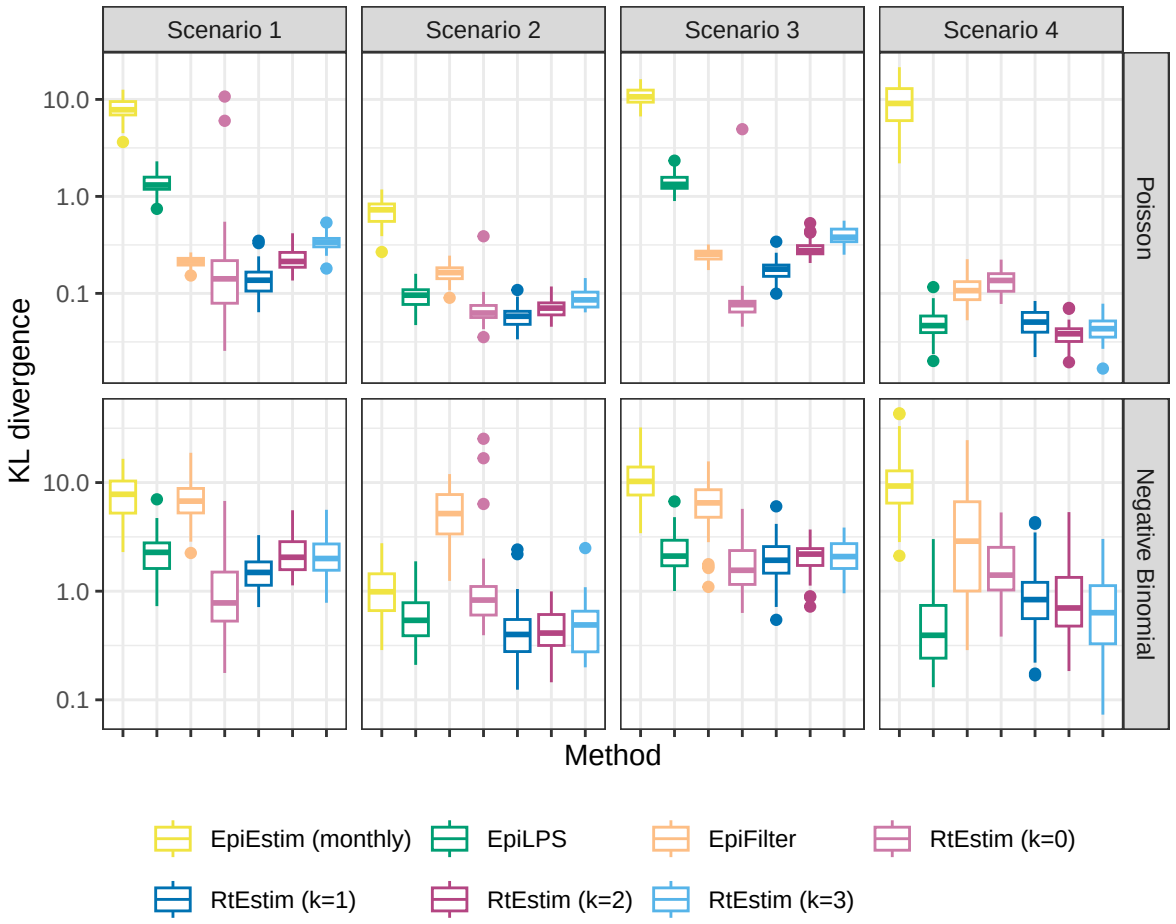


Figure A.3.1: KL divergence excluding the first months for measles epidemics. Y-axis is on a logarithmic scale.

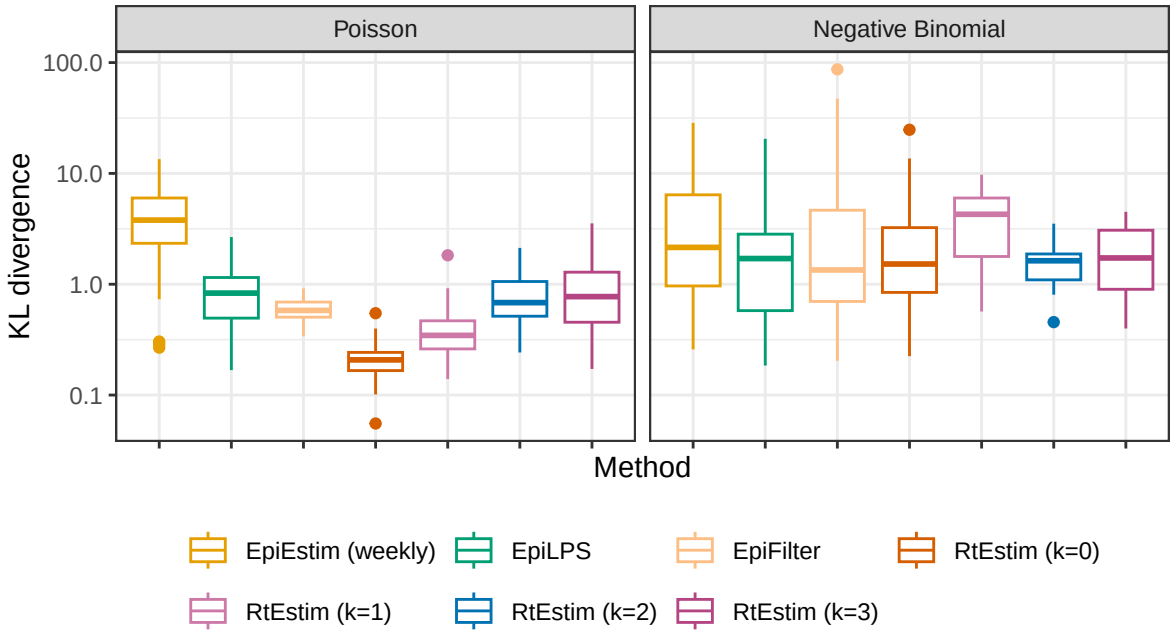


Figure A.3.3: KL divergence excluding the first weeks for flu epidemics. Y-axis is on a logarithmic scale.

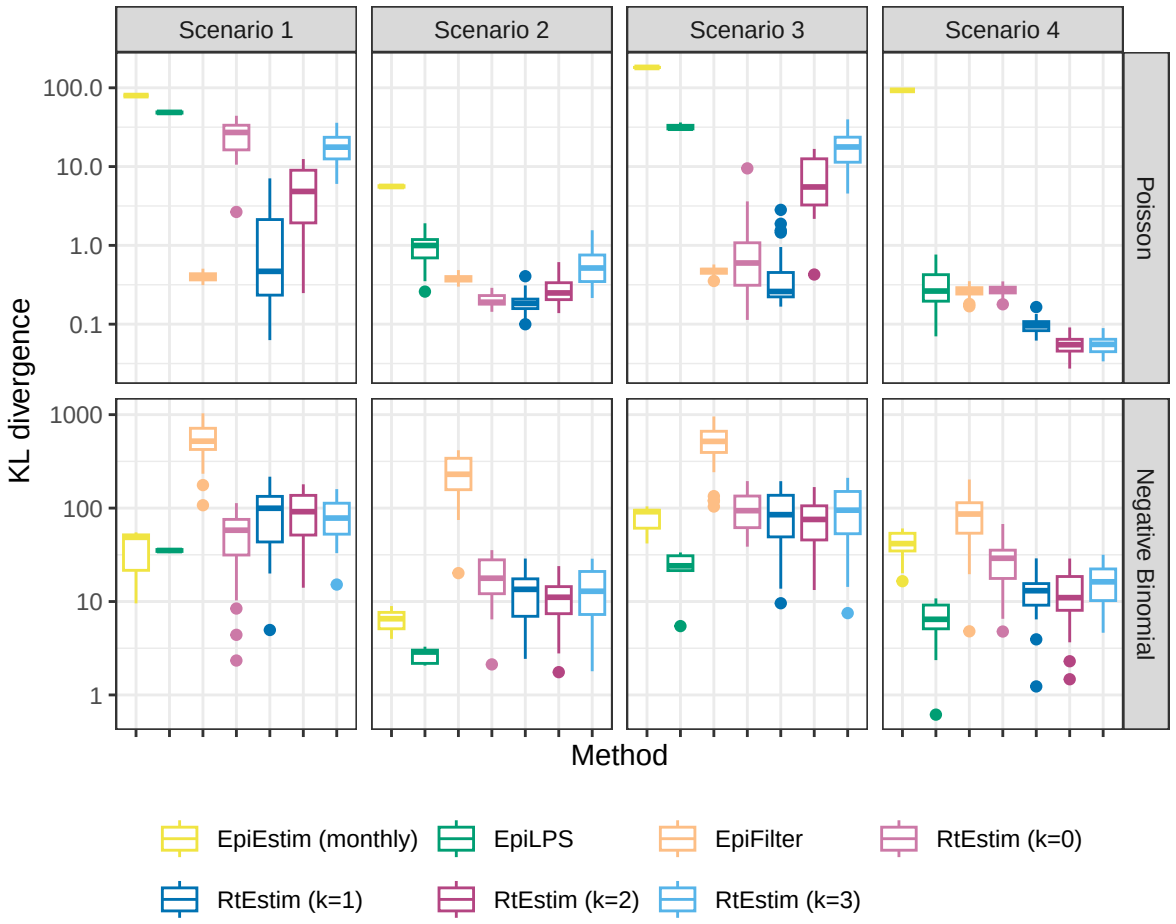


Figure A.3.2: KL divergence excluding the first months for SARS epidemics. Y-axis is on a logarithmic scale.

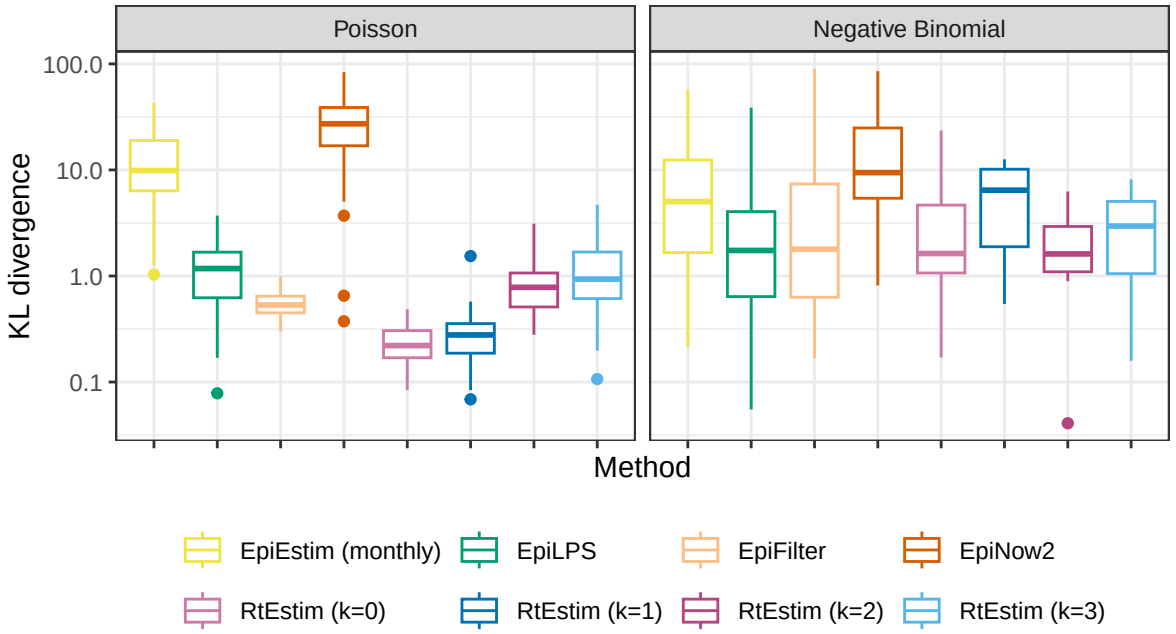


Figure A.3.4: KL divergence excluding the first months for flu epidemics. Y-axis is on a logarithmic scale.

A.4 Experimental results on accuracy under misspecification of serial interval distributions

A.4.1 Mild misspecification

Figures x and x display KL divergence (excluding first weeks and months respectively) for all 8 methods with shaped and scaled `measles` SI parameters with long `measles` epidemics across all settings.

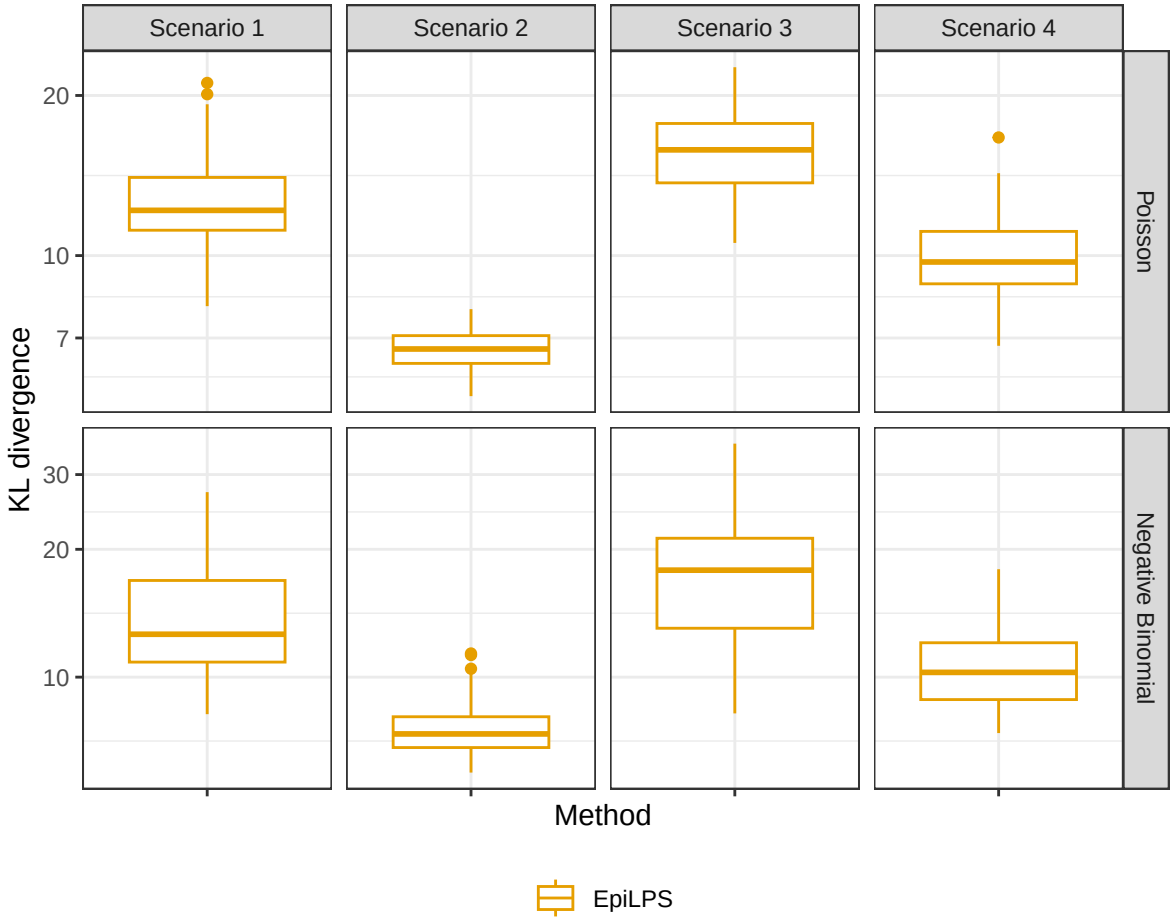


Figure A.4.1: KL divergence excluding the first weeks for measle epidemics with mild SI misspecification. Y-axis is on a logarithmic scale.

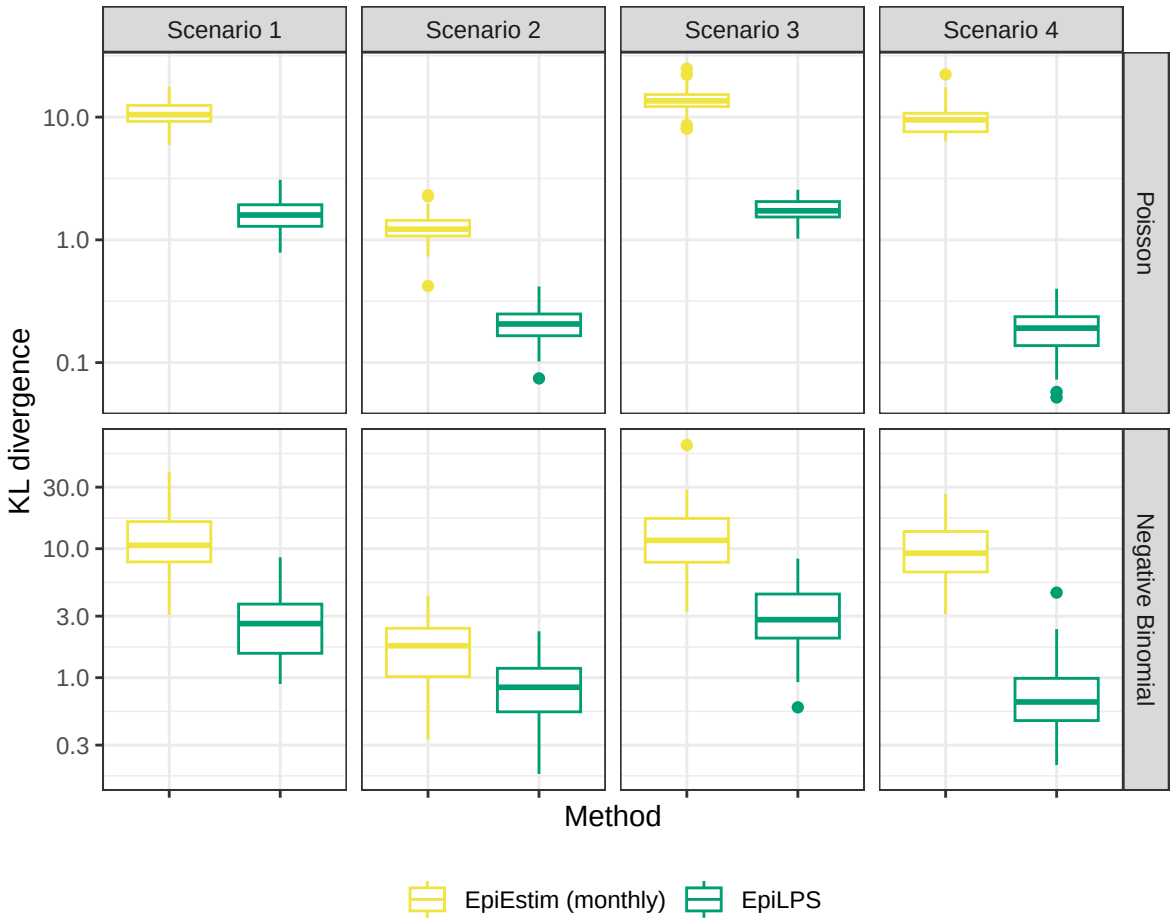


Figure A.4.2: KL divergence excluding the first months for measle epidemics with mild SI misspecification. Y-axis is on a logarithmic scale.

Figures x and x display KL divergence (excluding first weeks and months respectively) for all 9 methods with shaped and scaled `flu` SI parameters with short `flu` epidemics across all settings.

A.4.2 Major misspecification

Figures A.4.3 and A.4.4 display KL divergence (excluding first weeks and months respectively) for all 8 methods with `SARS` SI parameters with long `measles` epidemics across all settings.

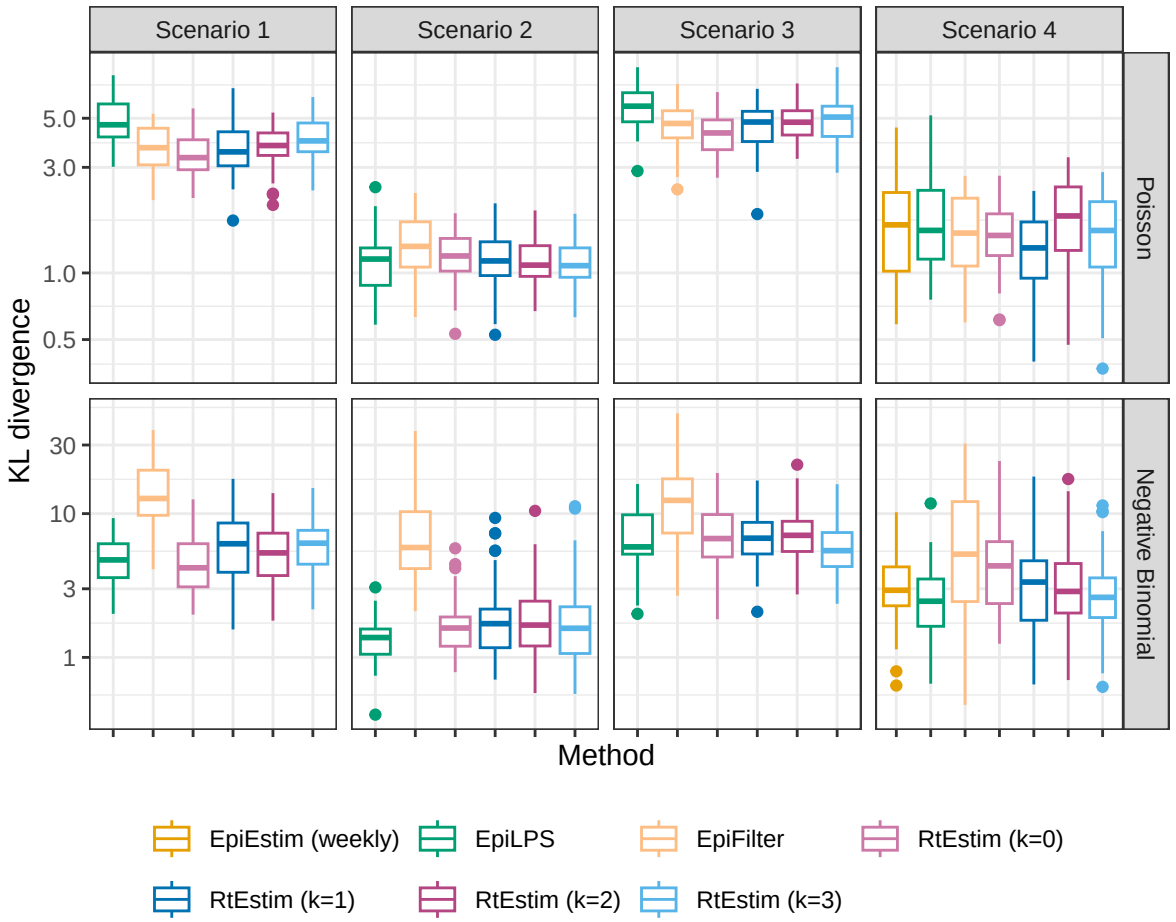


Figure A.4.3: KL divergence excluding the first weeks for measle epidemics with major SI misspecification. Y-axis is on a logarithmic scale.

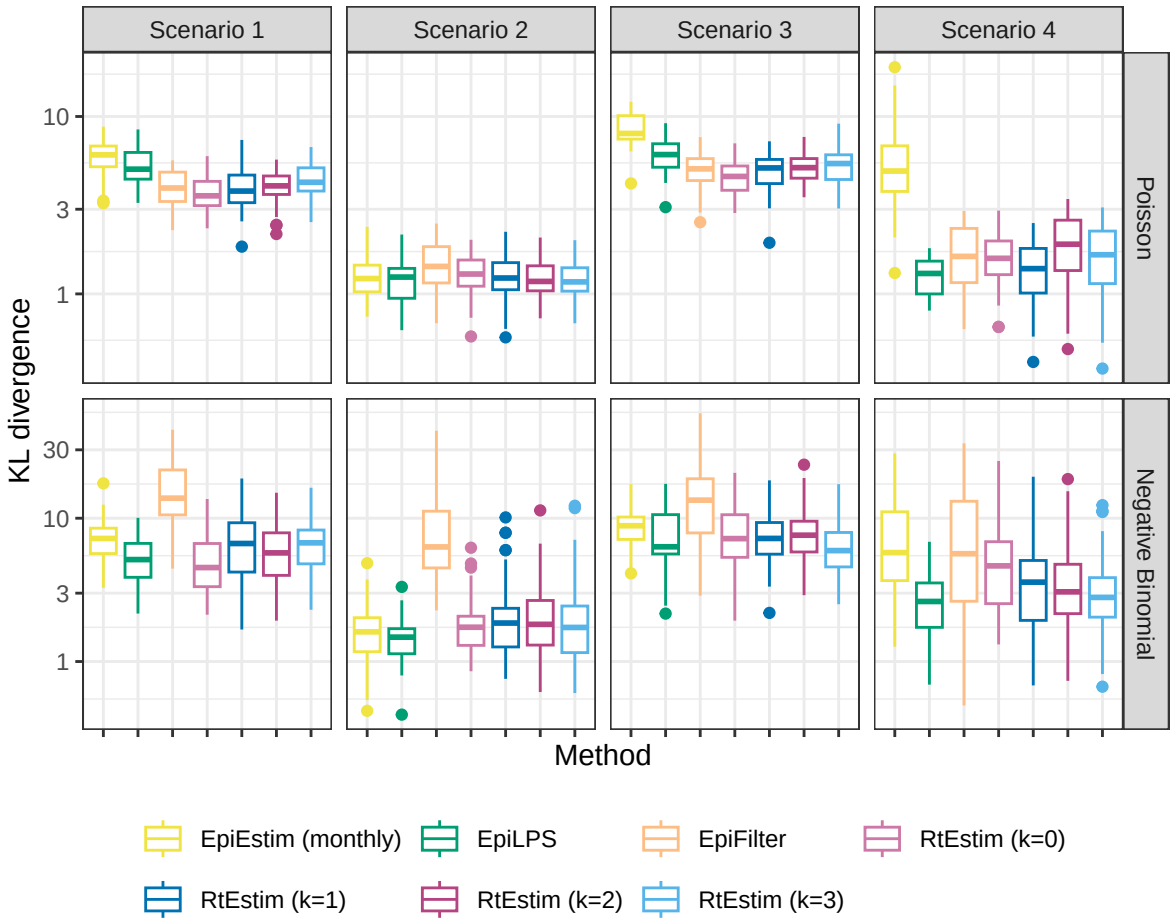


Figure A.4.4: KL divergence excluding the first months for measle epidemics with major SI misspecification. Y-axis is on a logarithmic scale.

Figures A.4.5 and A.4.6 display KL divergence (excluding first weeks and months respectively) for all 9 methods with `measles` SI parameters with short `flu` epidemics across all settings.

A.5 Confidence interval coverage

A.5.1 Display estimates and confidence intervals for sample epidemics

Let's review the estimated $\mathcal{R}_t$ with 95% confidence intervals for the sample long epidemics by all methods in the manuscript in Figure xx.

\begin{figure}[H]

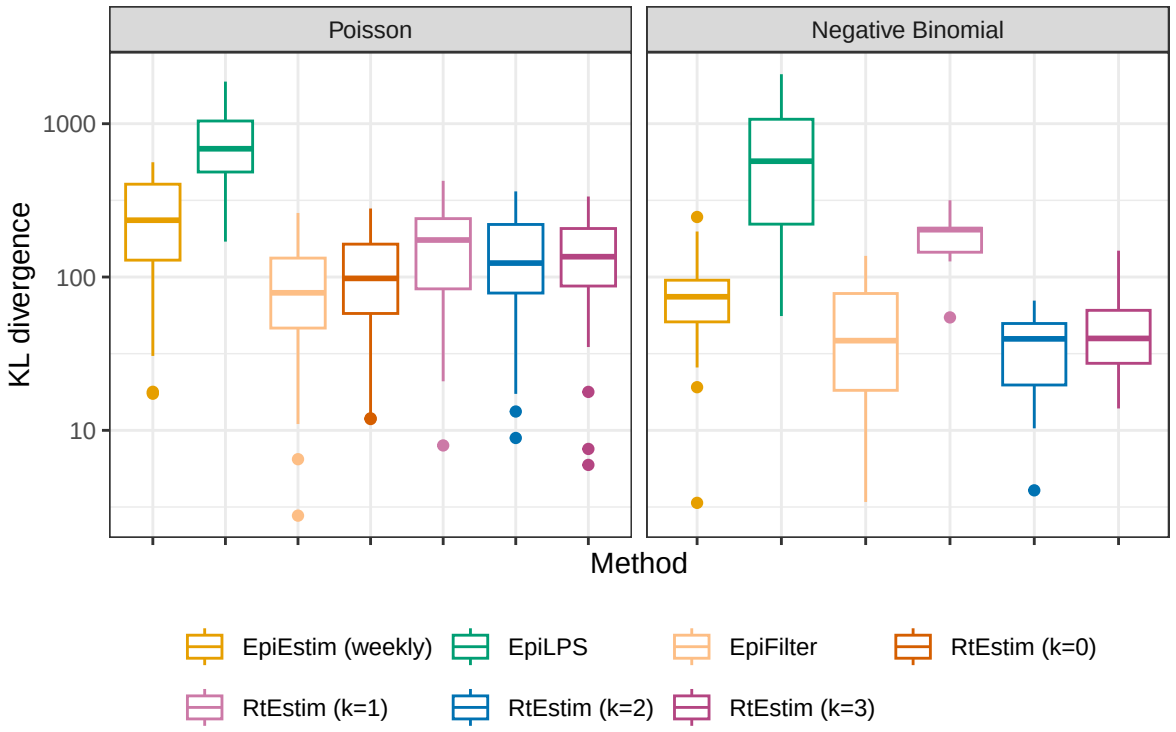


Figure A.4.5: KL divergence excluding the first weeks for flu epidemics with major SI misspecification. Y-axis is on a logarithmic scale.

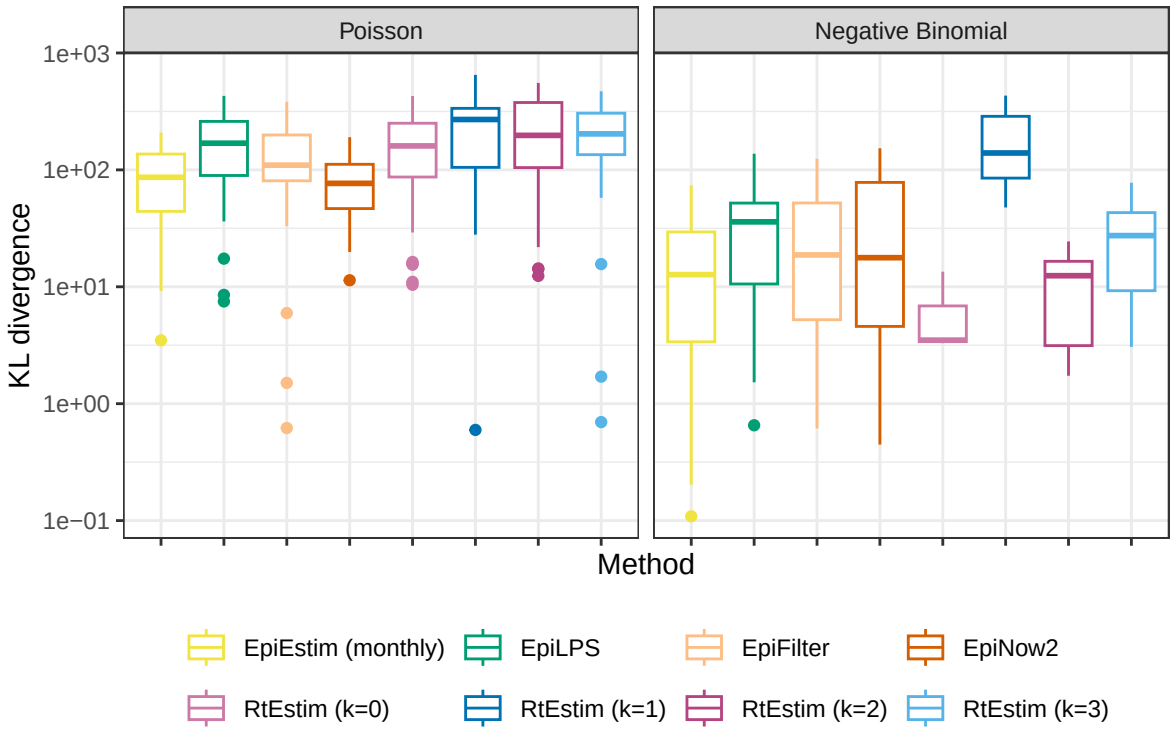
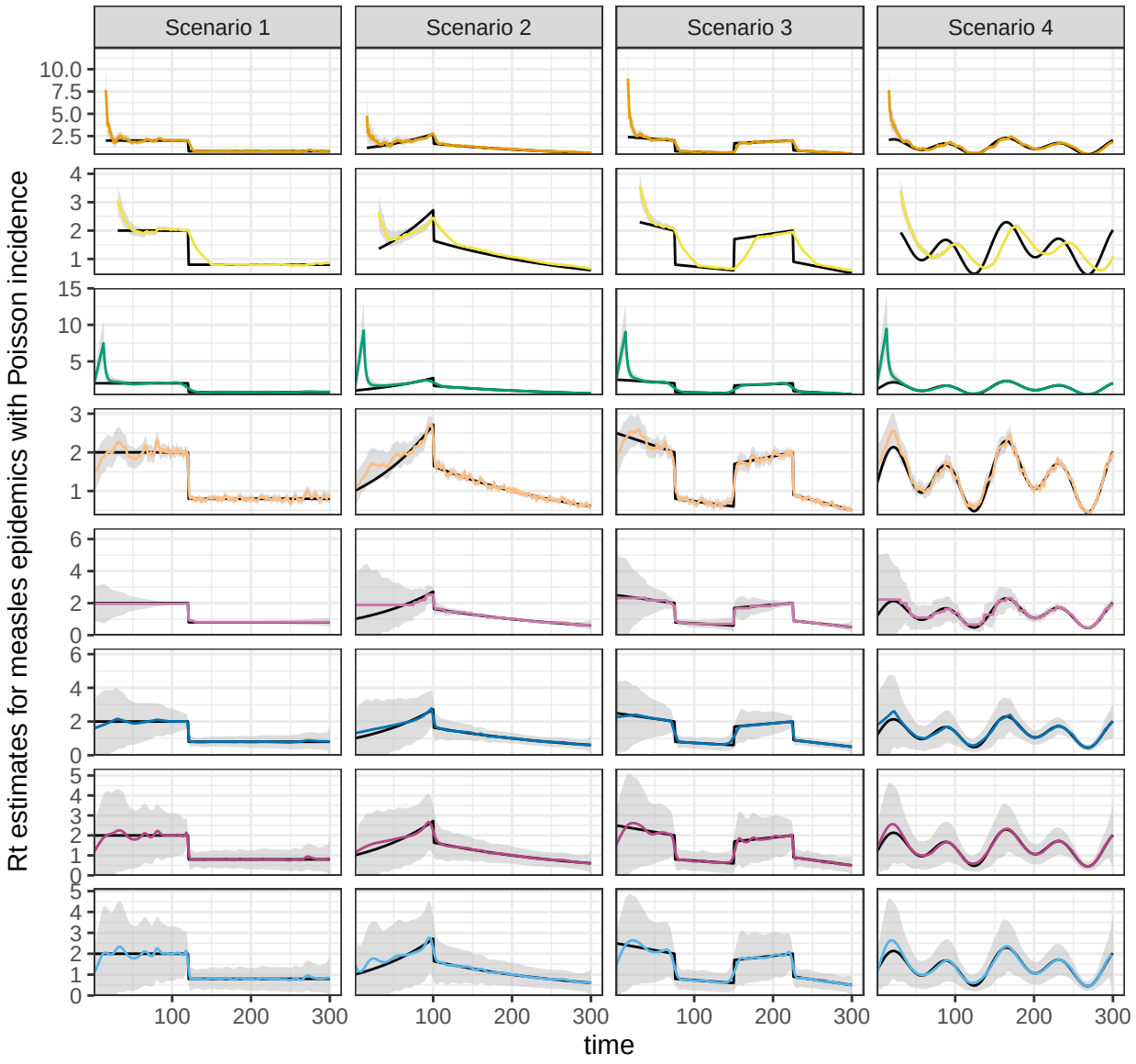


Figure A.4.6: KL divergence excluding the first months for flu epidemics with major SI misspecification. Y-axis is on a logarithmic scale.

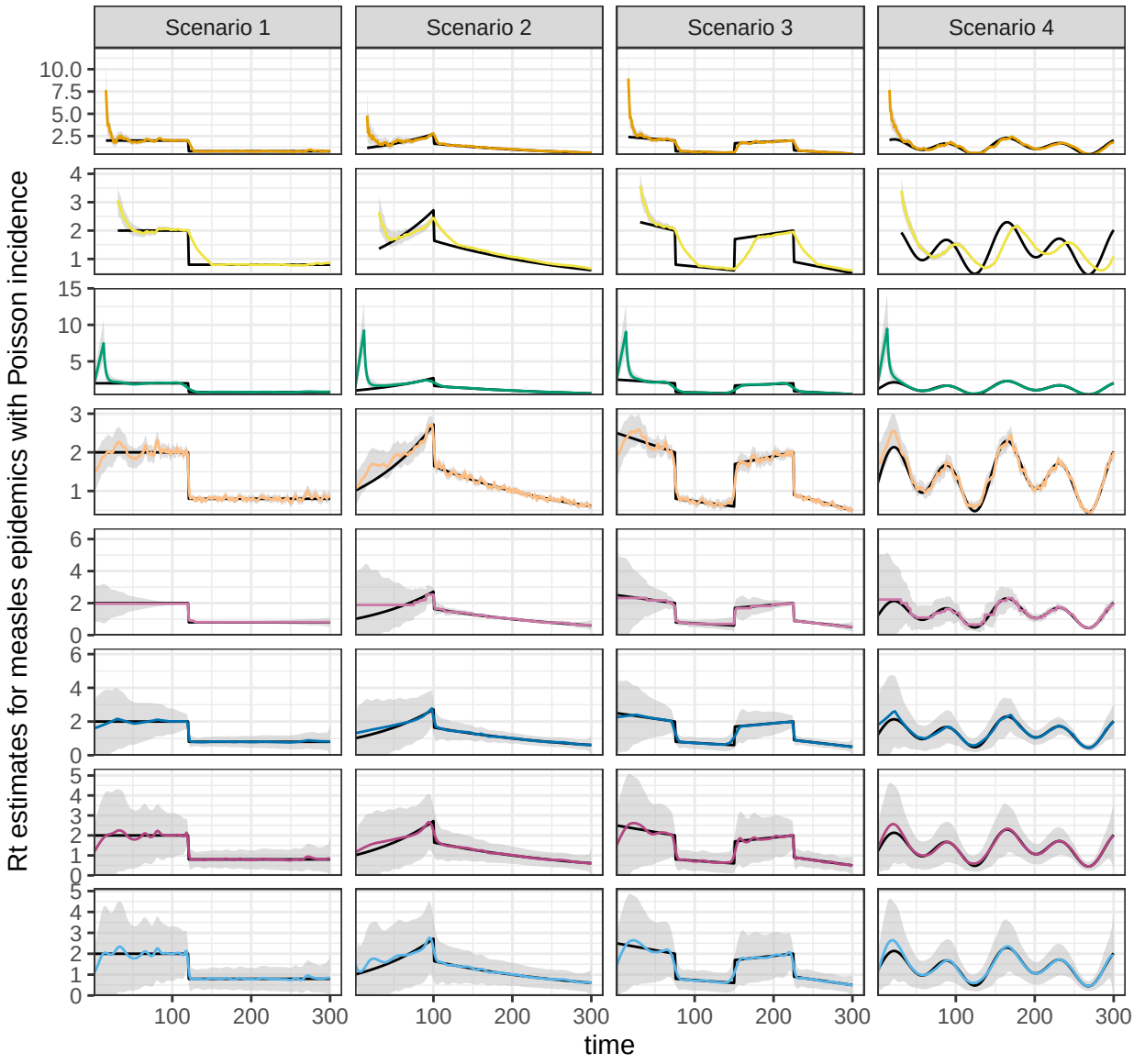


{

}

\caption{Fitted Rt with 95% confidence intervals for measles epidemics with Poisson incidence.} \end{figure}

\begin{figure}[H]

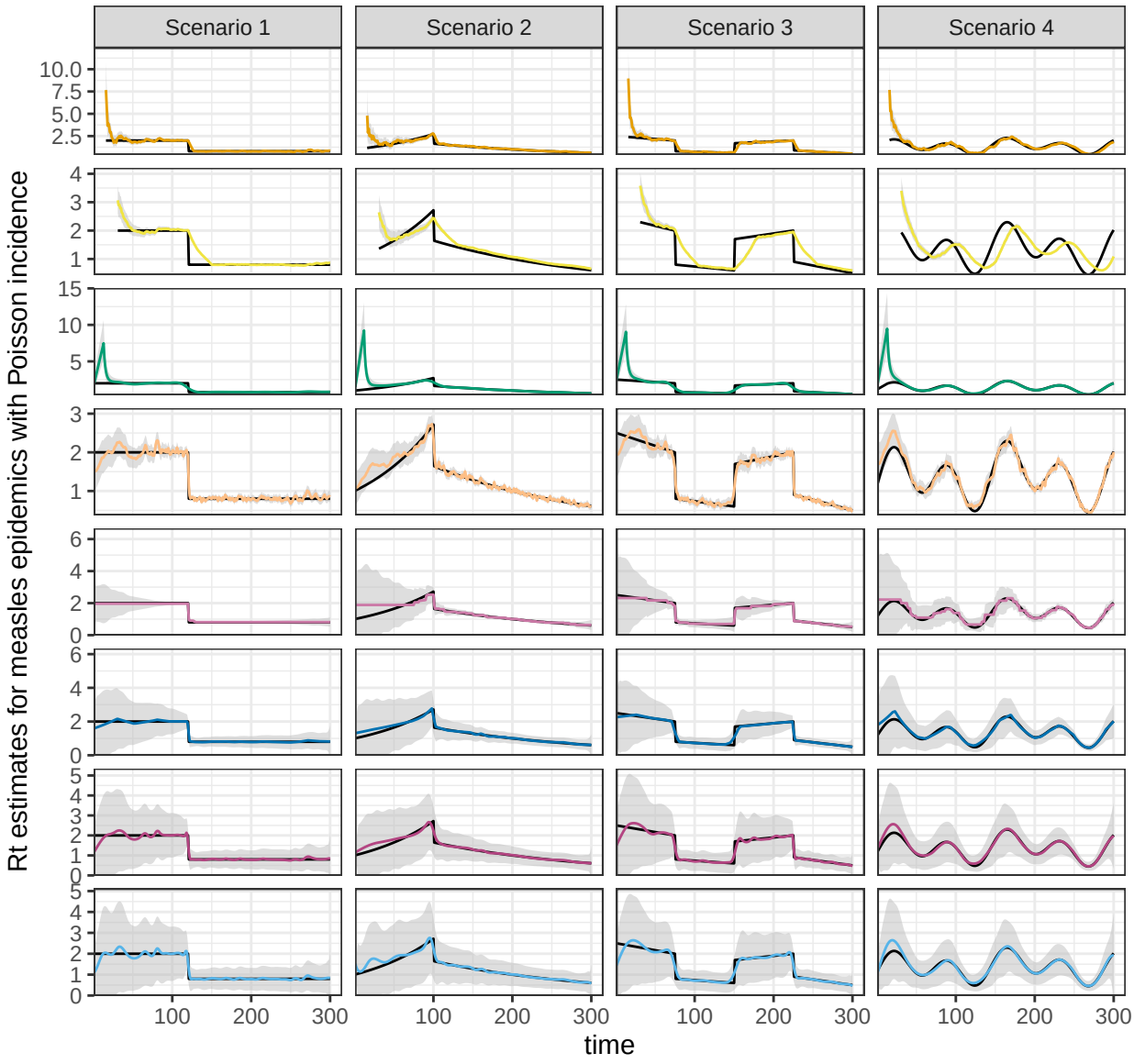


{

}

\caption{Fitted Rt with 95% confidence intervals for measles epidemics with negative Binomial incidence.}
\end{figure}

\begin{figure}[H]

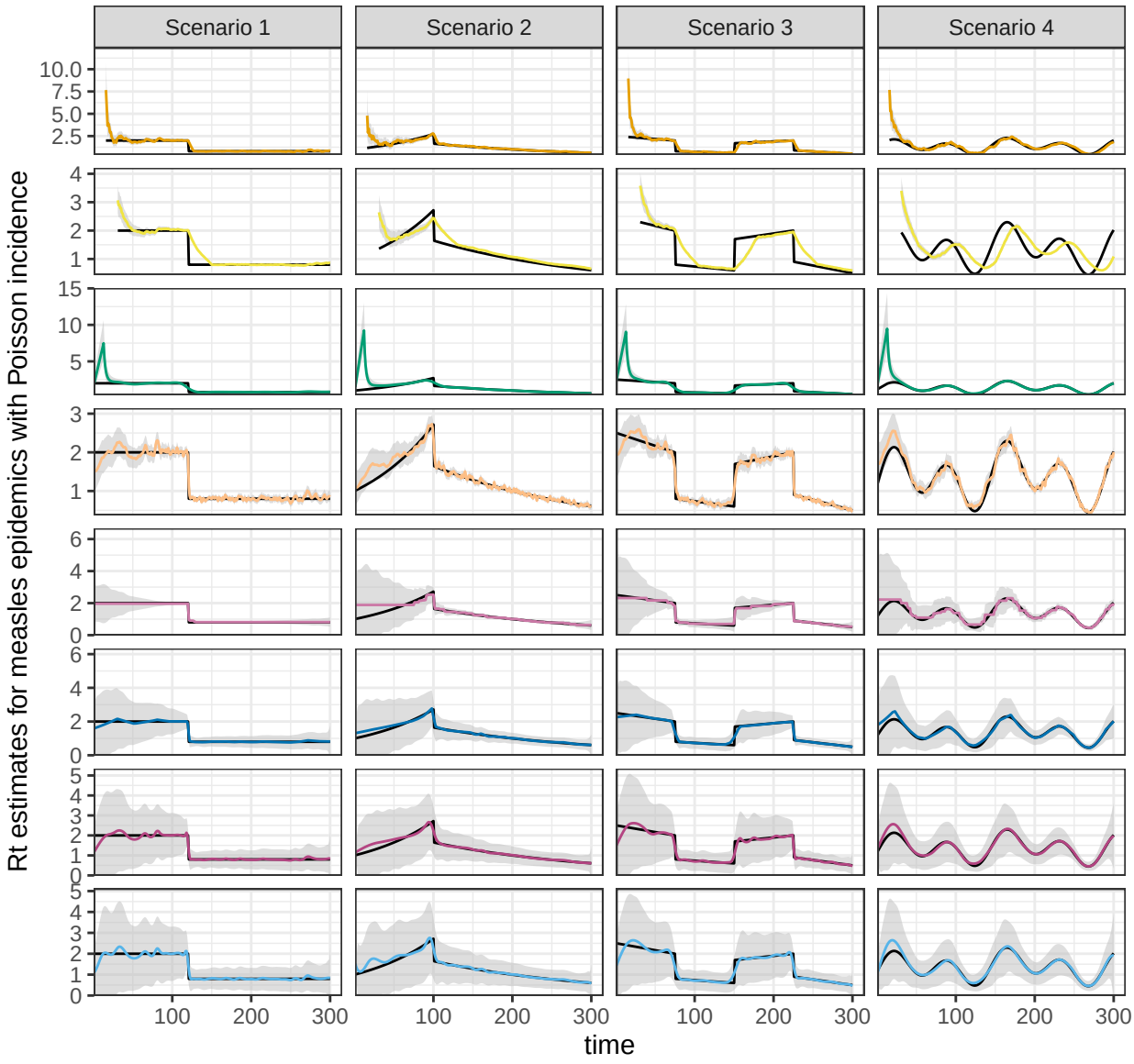


{

}

\caption{Fitted Rt with 95% confidence intervals for SARS epidemics with negative Binomial incidence.}
\end{figure}

\begin{figure}[H]



{

}

\caption{Fitted Rt with 95% confidence intervals for SARS epidemics with nengative Binomial incidence.}
\end{figure}

A.5.2 Experimental settings on coverage level comparison of confidence intervals

We focus on a specific $\mathcal{R}_t$ scenario, the piecewise linear case, and only long epidemics to compare the coverage of 95% confidence intervals across all 8 methods. We use the true serial interval distributions in modelling. Table 2 summarizes the experimental settings.

Table 2: Summary of experimental setting on coverage of confidence intervals

Length	SI	Rt scenario	Incidence	SI for modelling	Method
300	measles	3	Poisson, NB	measles	8 methods
300	SARS	3	Poisson, NB	SARS	8 methods

A.5.3 Experimental results

Figures A.5.1 and A.5.2 displays the percentages of coverage of 95% CI per coordinate over 50 random samples for measles and SARS epidemics respectively.

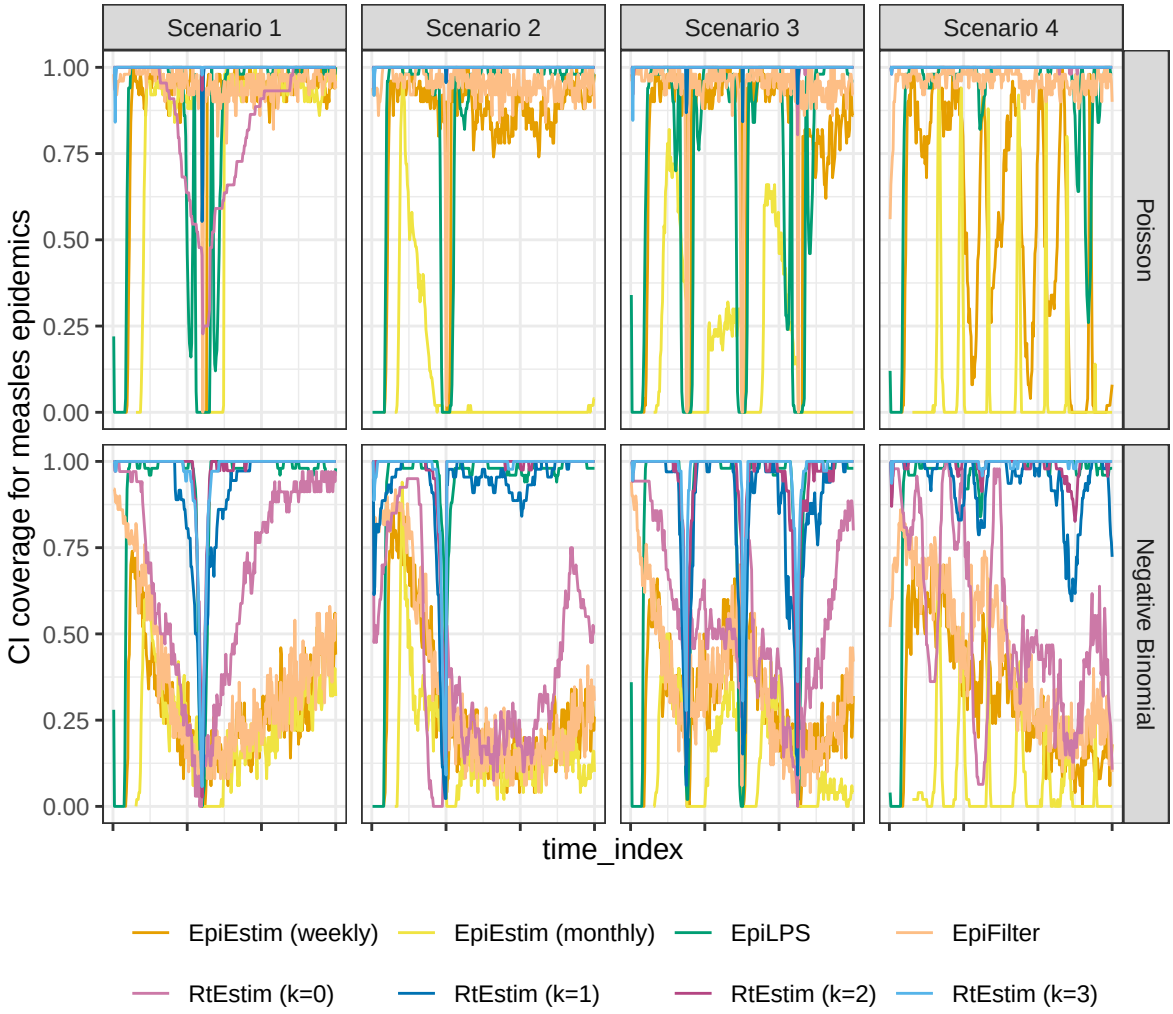


Figure A.5.1: Averaged coverage of CI per coordinate with measles epidemics.

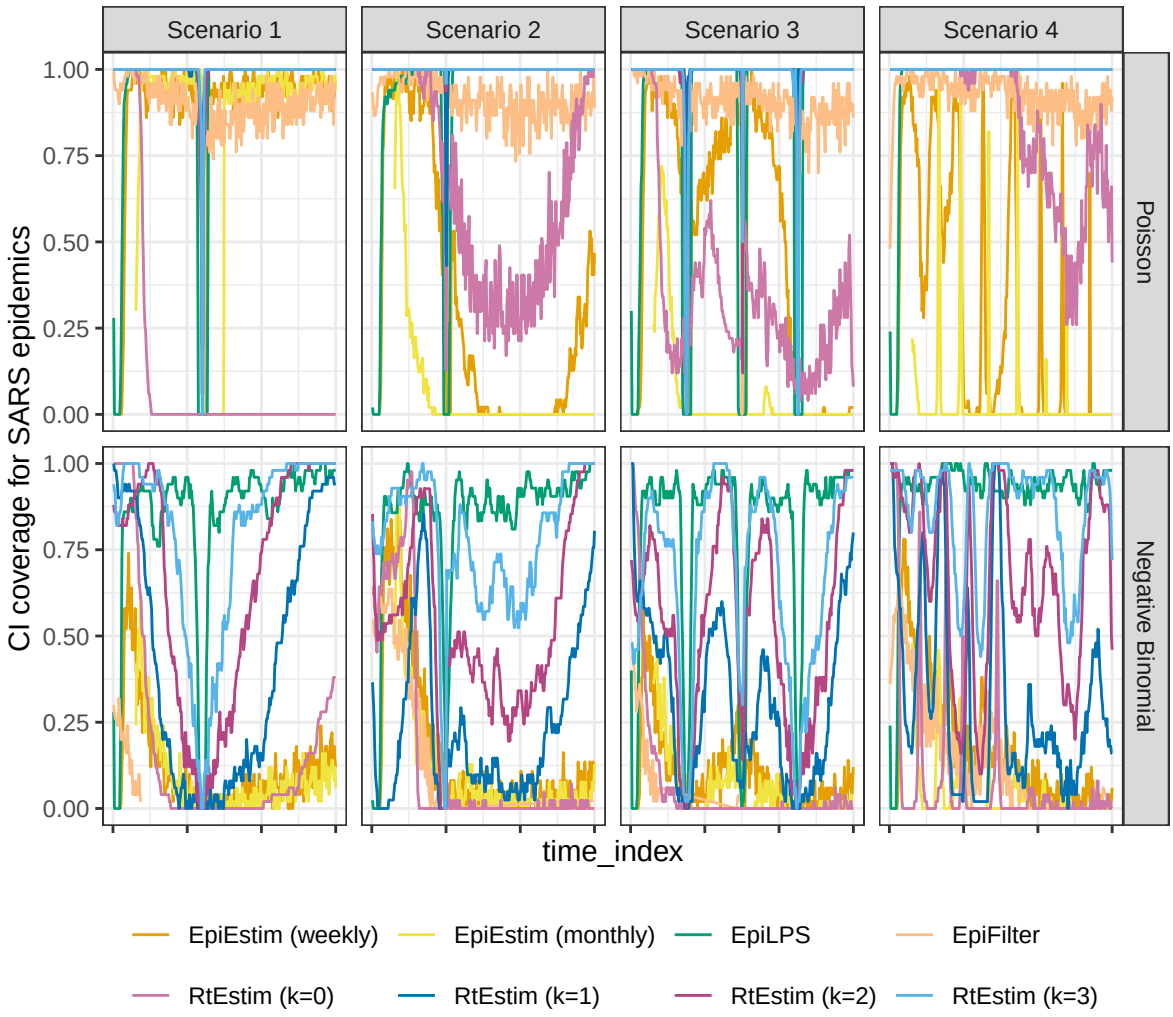
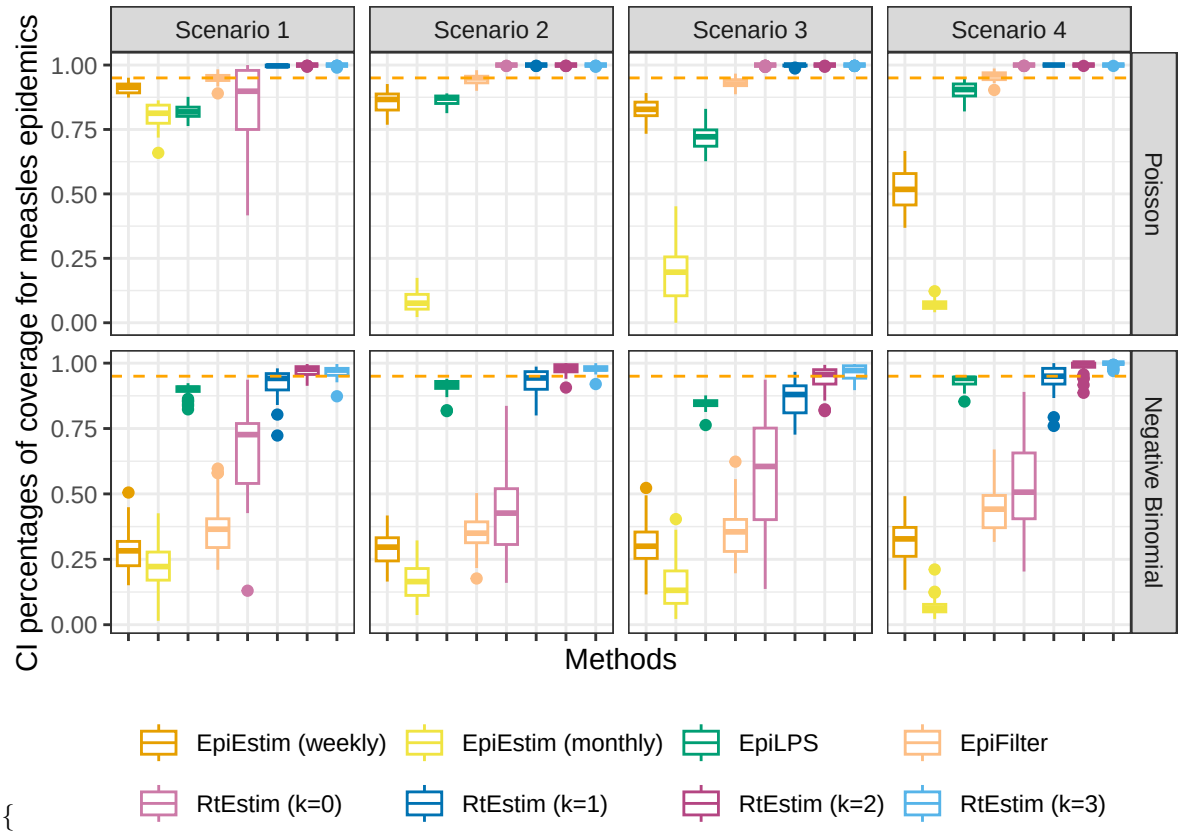


Figure A.5.2: Averaged coverage of CI per coordinate with SARS epidemics.

Figures A.5.3 and A.5.3 displays the percentages of coverage of 95% CI across all timepoints averaged over 50 random measles and SARS epidemics respectively.

\begin{figure}[H]



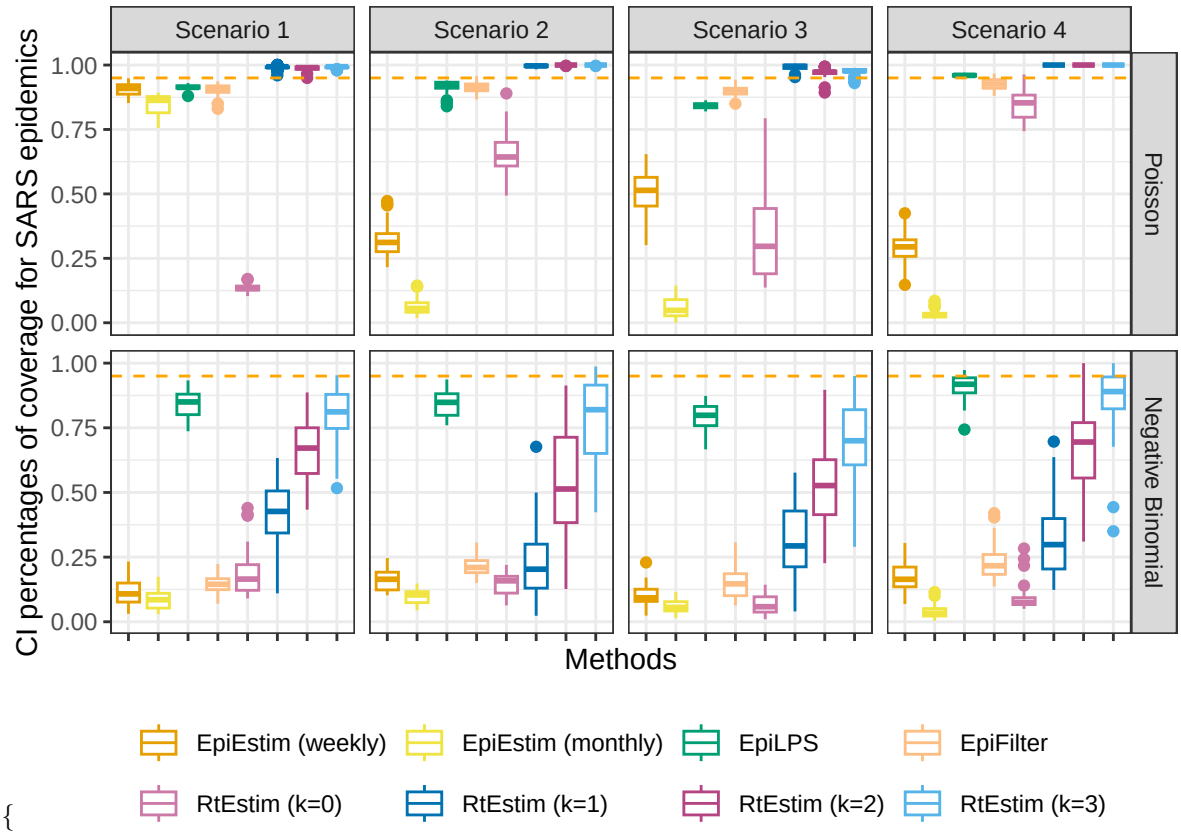
{

}

\caption{Averaged percentages of 95% CI with measles epidemics.} \end{figure}

"

\begin{figure}[H]



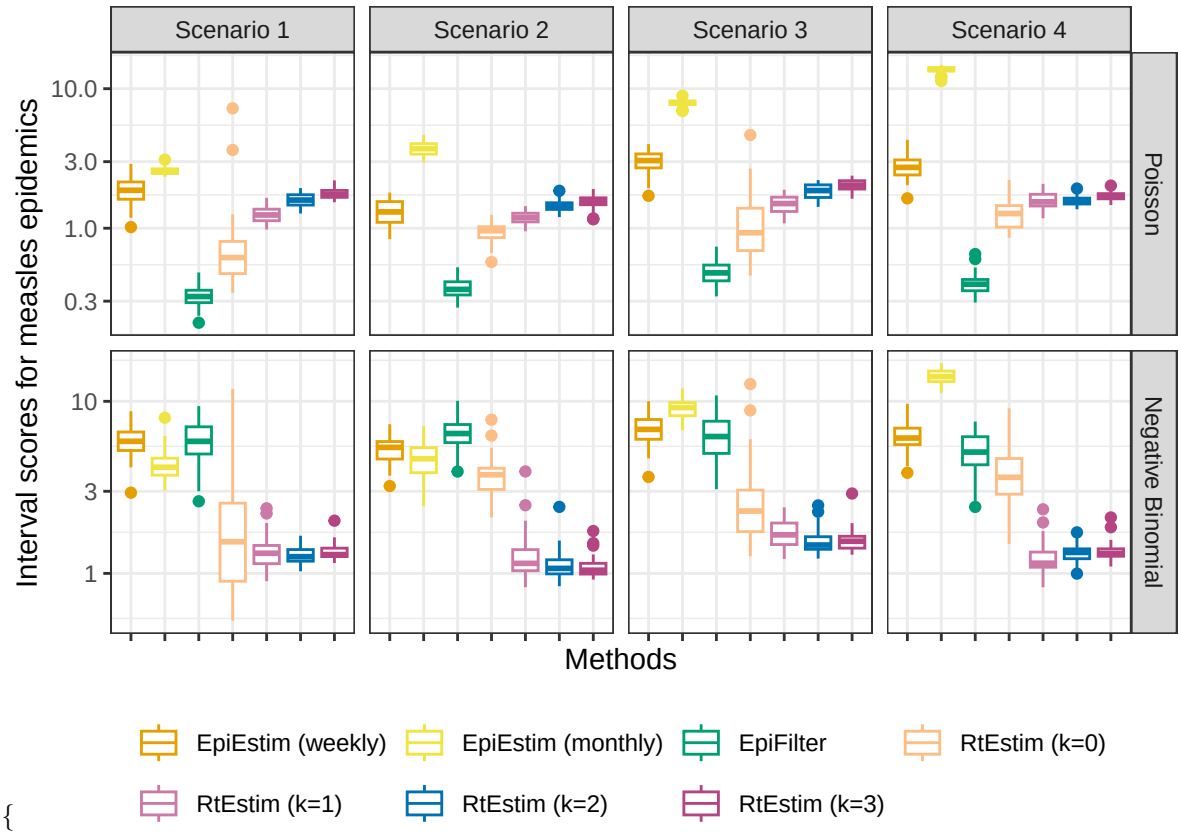
{

}

\caption{Averaged percentages of 95% CI with SARS epidemics.} \end{figure}

Figures A.5.3 and A.5.3 displays the interval scores of 95% CI averaged over 50 random measles and SARS epidemics respectively.

\begin{figure}[H]

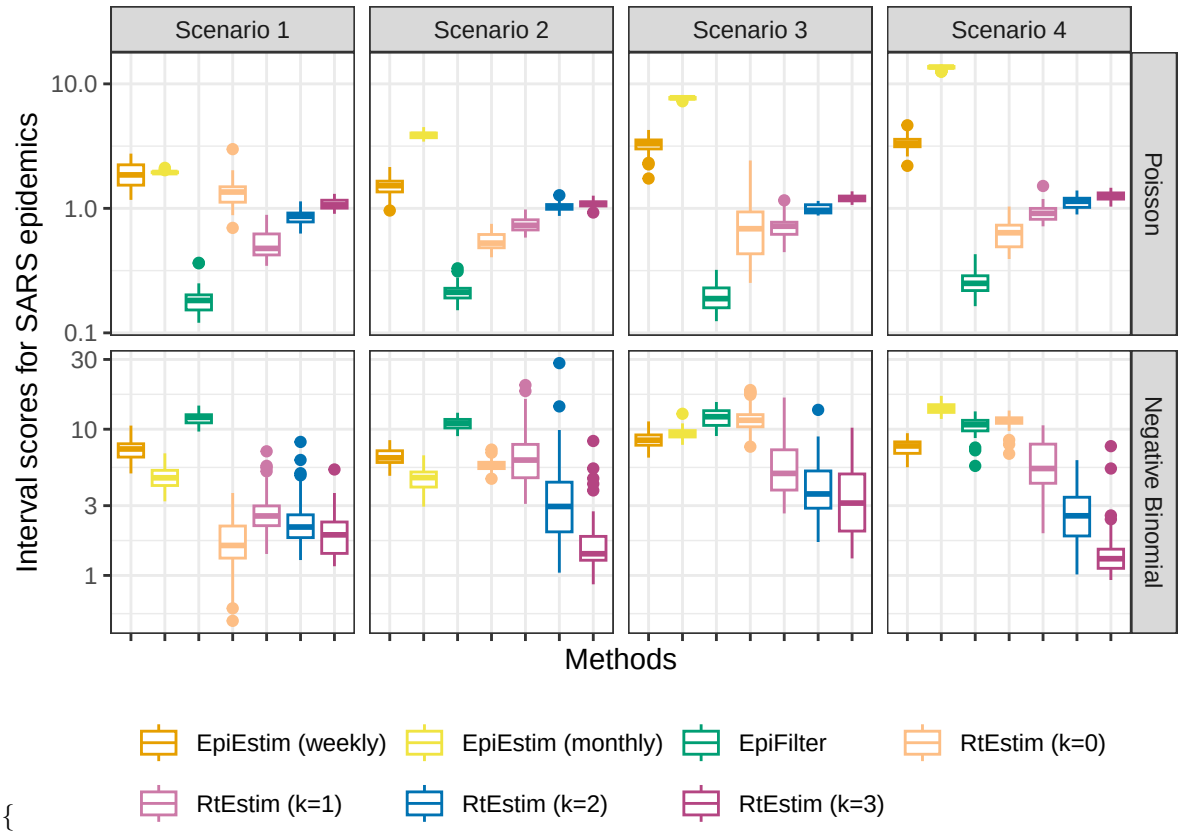


{

}

\caption{Averaged interval scores of 95% CI with measles epidemics.} \end{figure}

\begin{figure}[H]



{

}

\caption{Averaged interval scores of 95% CI with SARS epidemics.} \end{figure}

A.6 Time comparisons of methods for Section 3.2

Figures A.6.1 and A.6.2 show the time comparisons across all methods. `EpiEstim` with both weekly and monthly sliding windows are very fast and converge in less than 0.1 seconds. Piecewise constant `RtEstim` (with k=0) estimates can be generated within 0.1 seconds as well. `EpiLPS` is slightly slower, but still very fast and within 1 second for all experiments. Piecewise linear and cubic `RtEstim` (with $k = 1$ and $k = 3$ respectively) are slower, but mostly within 10 seconds.

It is remarkable that our `RtEstim` computes 50 lambda values with 10-fold CV for each experiment, which results in 550 times of modelling per experiment (including modelling for all folds). The running times are no more than 10 seconds for most of the experiments, which means the running time for each time of modelling is very fast, and on average can be less than 0.02 seconds. The other two methods only run once for a fixed set of hyperparameters for each experiment.

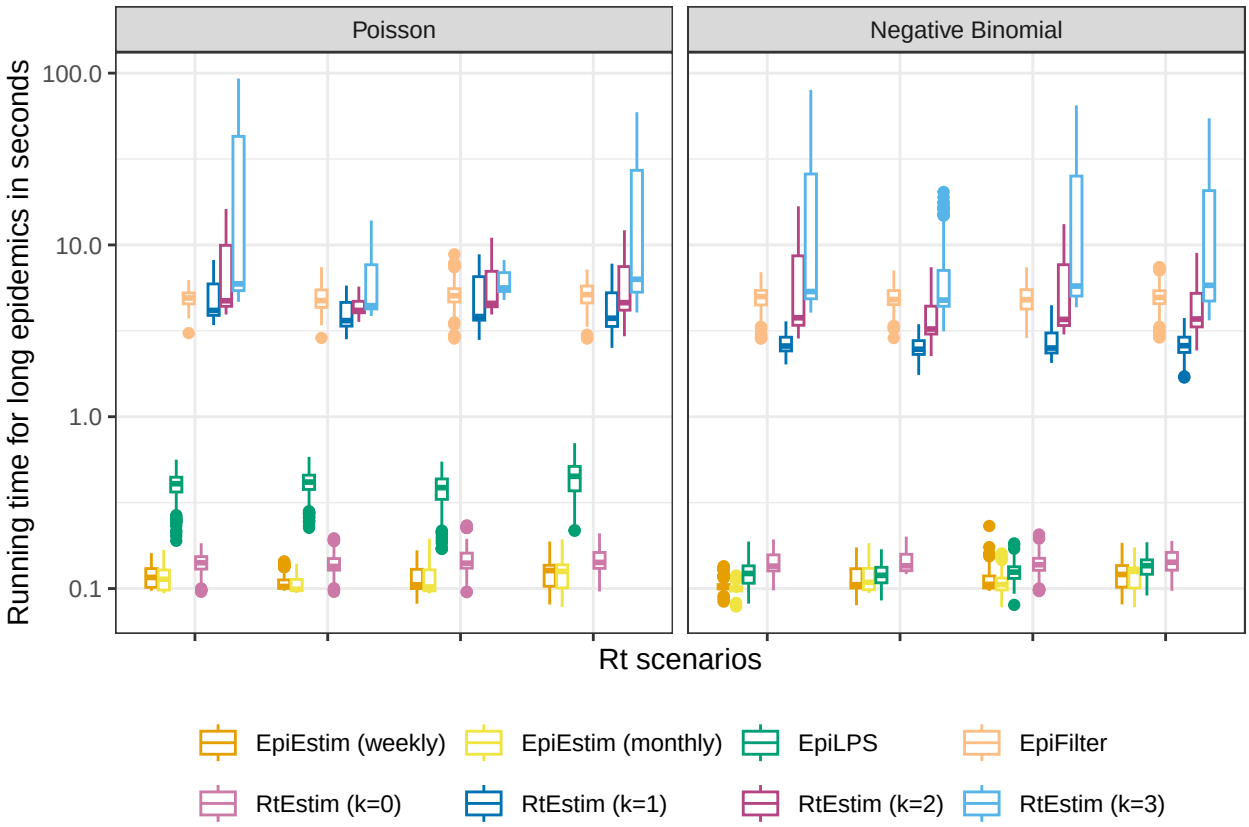


Figure A.6.1: Time comparisons of methods (excluding one outlier of 'RtEstim (k=1)' in Scenario 2 with negative Binomial incidence). Y-axis is on a logarithmic scale.

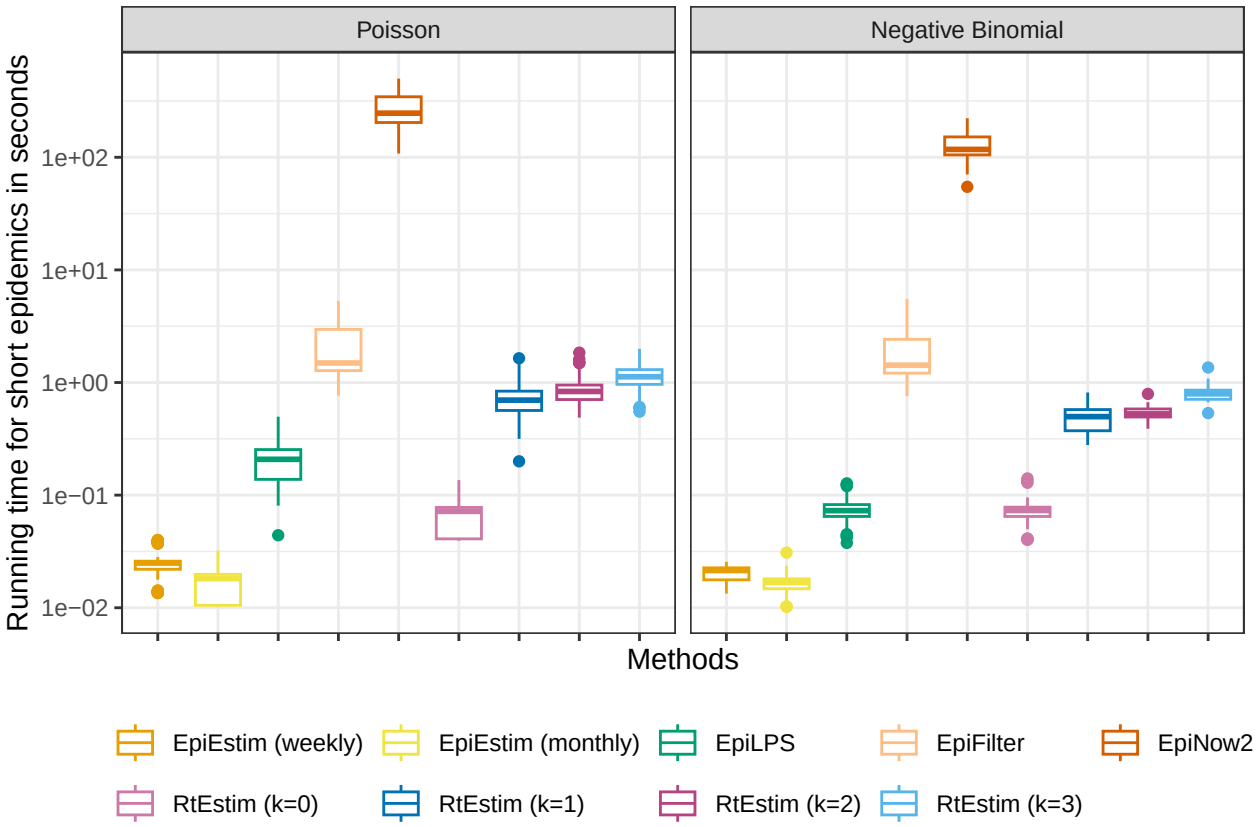


Figure A.6.2: Time comparisons of methods (excluding one outlier of 'RtEstim (k=1)' in Scenario 2 with negative Binomial incidence). Y-axis is on a logarithmic scale.