Analysis of modelling techniques used in the HIV epidemic

Mres in Biomedical research

Department of Surgery and Cancer

Imperial College London

Joshua D'Aeth

Supervisors: Dr. Jeff Eaton, Dr. Tara Mangal & Prof. Tim Hallet

10th May 2018

Contents

1 Introduction

UNAIDS currently uses the Estimation and projection package (EPP) to evaluate and predict trends in HIV incidence and prevalence within countries. This model has evolved markedly over the years, incorporating Bayesian melding for parameter estimation and using various techniques to estimate the transmission parameter.

In this report we aim to compare two of the most commonly used methods for modelling the transmission parameter and incidence: penalized B splines and the gaussian random walk. We systematically evaluate how each technique performs under different data configurations, to better inform future modelling directions for the EPP package.

2 Methods

We will simulate data for a HIV epidemic from our deterministic simple EPP model. This models the transmission parameter as a logistic curve through time and we can incorporate ART treatment into this framework.

We will initially test the goodness of fit to simulated data, from our deterministic model, of both first order and second penalized splines, and first and second order penalized gaussian random walks, with complete data for prevalence from the beginning of the epidemic. This will be performed over a set of different sample sizes from the population: 100, 500, 1,000 and 5,000 people. These random samples from the population will be repeated 100 times in each case and each of the four technqiues wil fit to the same set of four samples.

We will test how well our modelled fitted values match the true epidemic curve via a comparison of the root meas sqaured error (RMSE) of the true epidemic to the fitted values. This will be compared for the three output curves from our model representing: prevalence, incidence and the transmission parameter kappa. We will compare this over the whole time period of the epidemic and specifically in the last 5 years of the epidemic when we have no sample data, to assess which methods if likely best for predicting future trends.

Furthermore we will evaluate how often the true values for the epidemic parameters fall within our 95% confidence interval produced from the model fitting, again this will be performed for each of the three parameters: Prevalence, incidence and the transmission rate.

Finally we will also attempt to understand how the four different models maybe overfitting to the data. We attempt to understand this via a comparison of the RMSE between the model fitted values and the true epidemic values, and the RMSE between the model fitted values and the values of the prevalence used in the sample to fit the data with.

3 Results

3.1 Plots of mean results

First we will plot the mean output over the 100 runs of the sampled data for each of the four techniques for the model fitting, below is the mean fit with respect to prevalence, the red line represents the true epidemic and the blue line represents the fitted values with their 95% confidence intervals in shaded blue.

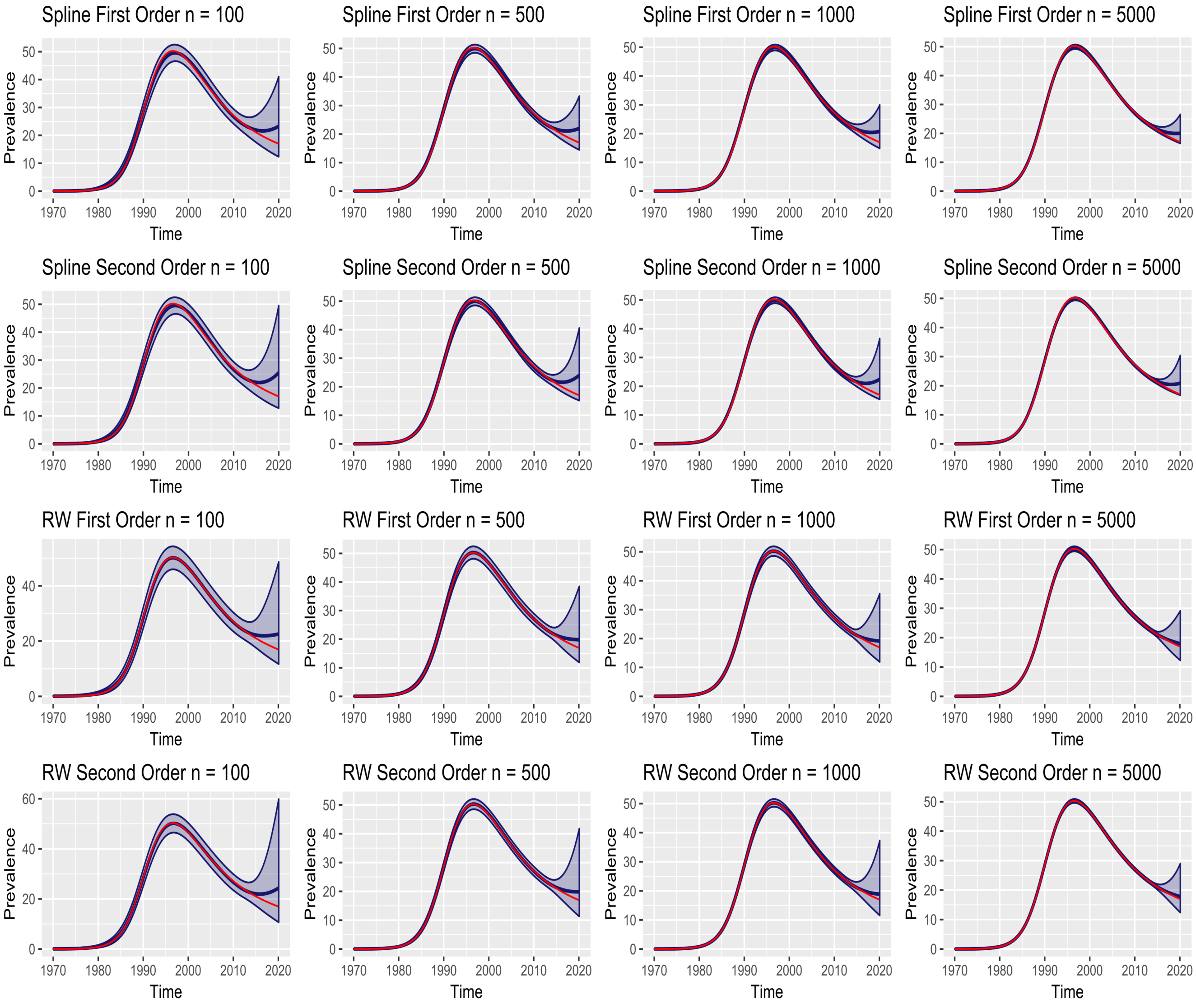

Figure 1: Prevalence mean for the 4 modelling techniques for 4 different sample sizes

From this plot we can initially see that all the different fitting methods capture the prevalence of the true epidemic very well during the phase from 1970 to 2015 when we have data to fit to. All models also have a narrowing of their credibility bounds as the sample size goes up from left to right. During the prediction period however the models differ in how well they seem to match the true epidemic. Both spline models seem to predict a levelling off of the prevalence or a slight uptick depending on their sample sizes, whereas in reality the prevalence is seen to decrease during this period. The RW models begin to more closely match the epidemic qualitatively with larger sample sizes, with the RW second order model at sample size 5000 very similar indeed.

The mean values over the 100 runs on the different data sets are depicted below, again the blue line is the fitted data while the red line indicates the true value from the simulated epidemic.

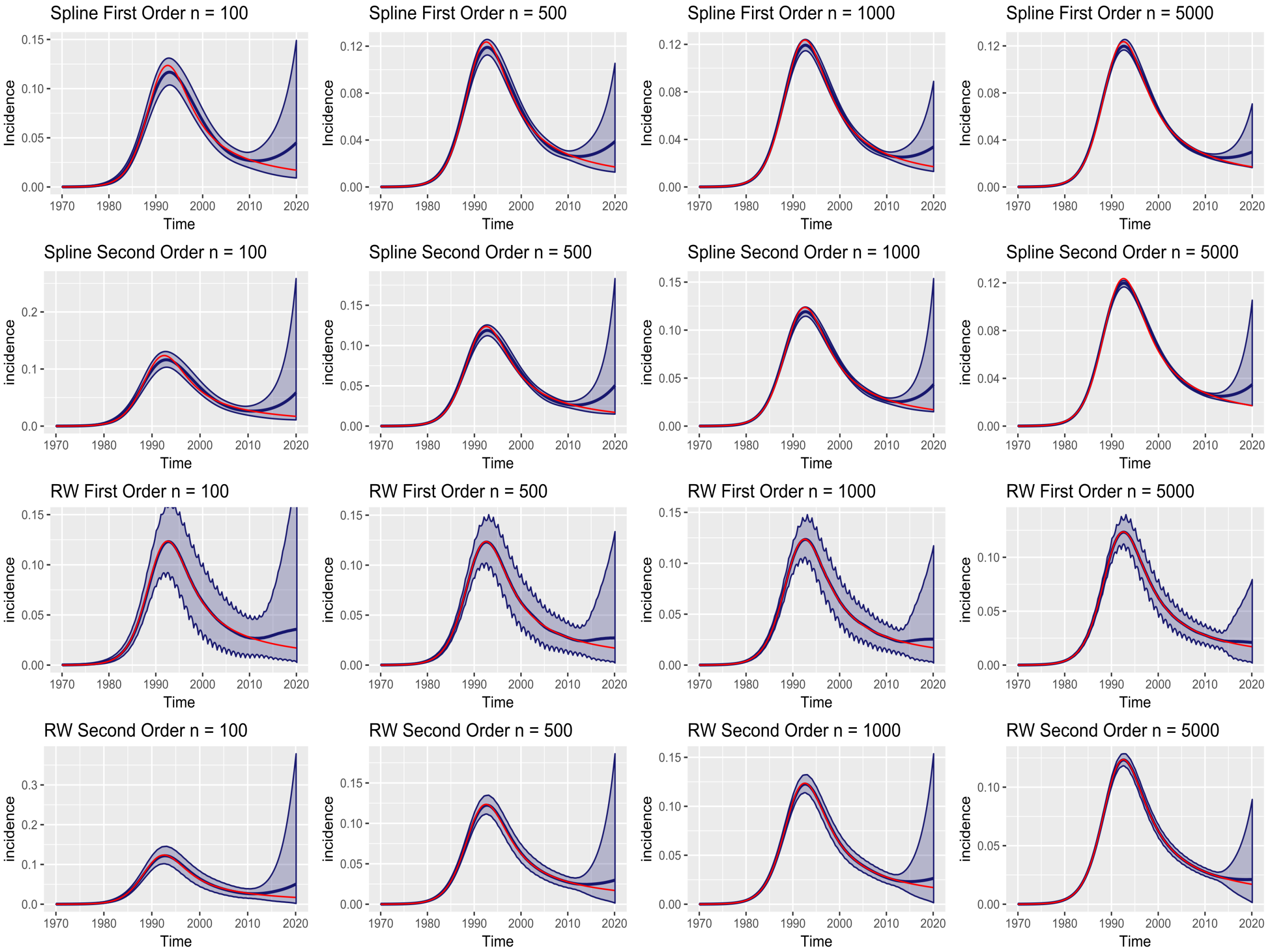

Figure 2: Incidence plots for the four different fitting methods over four different sample sizes from the population

From this we see that all four methods generally fit well to the true epidemic. Both spline methods however seem to underestimate the peak incidence reached during the early to mid 1990s. The credibility intervals for the RW methods also remain fairly constant, especially for the first order RW, despite an increae in sample size from the population. Once again the RW methods also seem to more closely match the true epidemic incidence during our prediction period from 2015 to 2020. The Second order RW again matches most closely the slight decline seen in the true simulated epidemic.

Finally the transmission parameter, here descirbed as kappa, plots are seen below. Again the red line indicates the true value and the blue line the fitted value with assoicated shaded region the 95% credible interval.

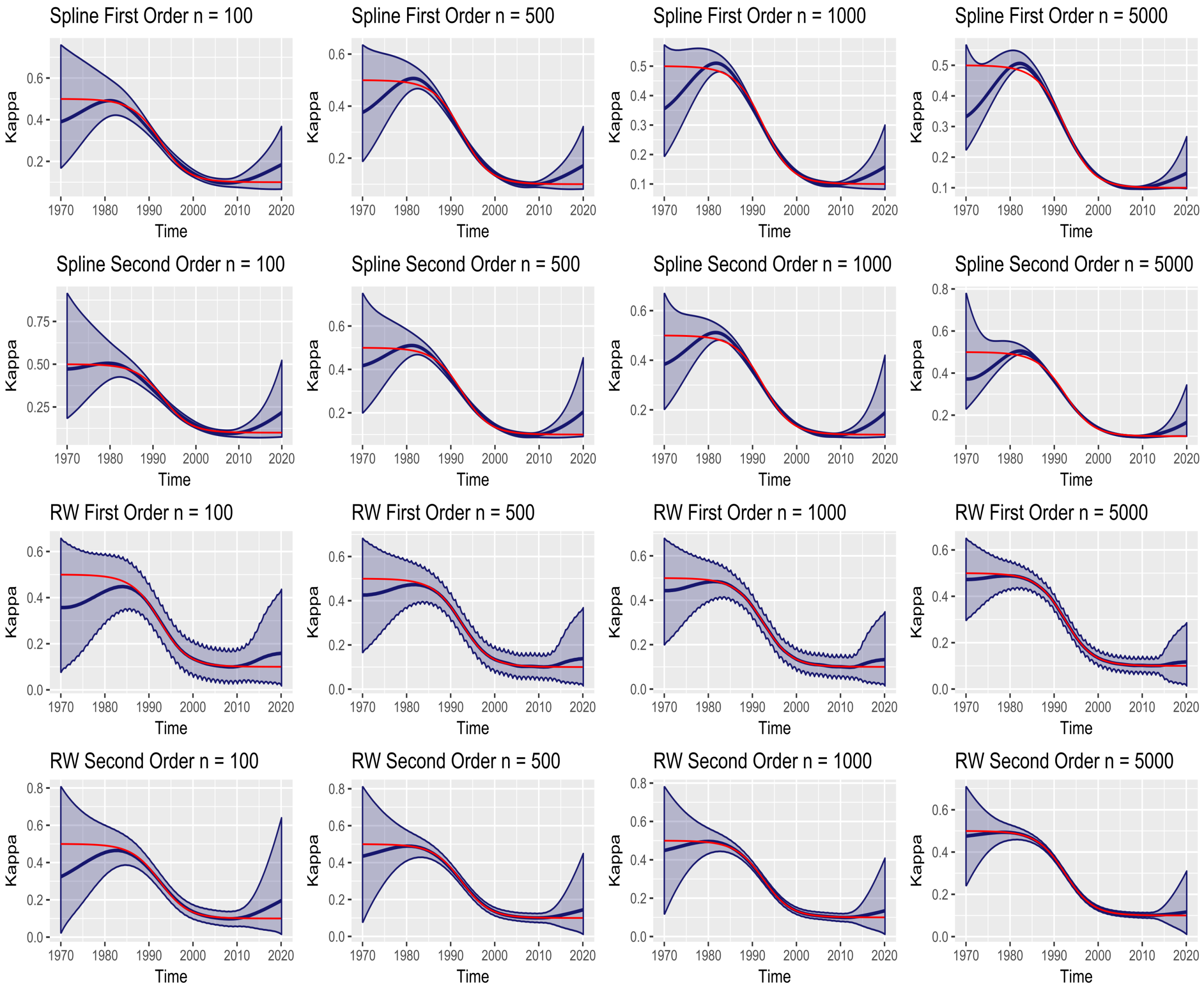

Figure 3: Kappa plots produced for the four different fitting methods, over four different sample sizes from the population

The fitted values from the model produced for kappa are much less accurate than our previous two parameters. All methods are seen to underestimate the initial kappa value, with first order splines and in particular RW second order fitting to a sample size of 100 very low estimates. All techniques though fit well during the decline phase of kappa from the late 1980s to the early 2000s. During the prediction period, both spline methods predict an uptick in kappa, with this being most pronounced in smaller sample sizes. The RW methods are seen to predict a levelling off of the kappa during this period, with the RW second order fitting particularly well at sample size 5000.

3.2 Quantitative measures of fit

3.2.1 RMSE to true values

Below in table (1) is the RMSE values for prevalence of the true epidemic against the fitted median values for the whole data series, averaged over 100 iterations. One broad trend is that as the sample size increases, we see a better fit to the true epidemic as witnessed by a decrease in RMSE. In general at lower sample sizes we see the first order penalized splines and RW outperform their second order versions, while at higher sample sizes both RW techniques fitting better to the true epidemic than the spline methods, with second order RW fitted with a sample size of 5,000 having the lowest RMSE value.

Table 1: The complete RMSE for prevalence

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0185343	0.0214740	0.0182192	0.0210485
500	0.0120273	0.0146318	0.0097615	0.0100131
1000	0.0093654	0.0113620	0.0078701	0.0077898
5000	0.0062764	0.0073870	0.0040916	0.0038942

In table (2) we see the RMSE values over the whole course of the epidemic for the true incidence value against the fitted incidence values averaged over 100 iterations. Similar to for prevalence we see that the first order methods are slightly better at fitting with lower sample sizes, while at the highest sample sizes, RW methods fit to true incidence better, while the best fit to incidence is the RW second order at a sample size of 5,000.

Table 2: The complete RMSE for incidence

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0081221	0.0102019	0.0078942	0.0093847
500	0.0056484	0.0074060	0.0049852	0.0044005
1000	0.0044308	0.0058095	0.0044844	0.0035941
5000	0.0031688	0.0038922	0.0029155	0.0019608

For the transmission parameter, kappa, the RMSE values for the whole time series of the epidemic averaged over 100 iterations are seen in table (3). This is an interesting set of results, for both splines we see no real decrease in RMSE with an increae in sample size, as we had seen for both incidence and prevalence. Indeed if anything first order splines decrease in predictive power with increasing sample size. With RW there is a clear improvement with increasing sample size, with the best fit to the true value being the First order RW with a sample size of 5000. These values though are still the same order of magnitude as the values Spline results for RMSE.

Table 3: The complete RMSE for Kappa

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0432773	0.0450542	0.0606701	0.0633587
500	0.0427531	0.0451453	0.0332154	0.0362085
1000	0.0457103	0.0457409	0.0284724	0.0337876
5000	0.0481981	0.0434053	0.0173404	0.0185758

3.2.2 RMSE to true for prediction period

In table (4) we have the RMSE values of the true prevalence against the fitted values for prevalence during the last five year period in which no data were sampled, these are averaged values across 100 iterations of sampled data. In general from this we can seen that both RW methods produce closer fits to the true data, with RW first order at a sample size of 5,000 having the best average fit. At lower sample sizes the different methods are quite similar, with RW first order having a slightly better fit each sample size.

Table 4: RMSE of Prevalence for fitted against true values during 2015:2020 prediction period

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0413941	0.0512600	0.0386404	0.0497123
500	0.0292720	0.0383811	0.0199103	0.0220041
1000	0.0233138	0.0301396	0.0166608	0.0180795
5000	0.0167389	0.0205310	0.0086108	0.0091037

Incidence predictions during 2015:2020 are seen in table (5). Here once again we see that the RW fitting technique produces a closer fit across the different sample sizes than the equivalent penalized spline. Additionally the first order penalized RW produces the best fit to the true data among the different fitting techniques across the range of sample sizes during this five year prediction period. It is interesting to note that the first order penalized methods in this scenario always outperform their second order counterparts. This could be due to the fact that in the absence of data a first order penalty will mean the value plateaus, something which we see the incidence parameter of the true epidemic begins to do during this prediction period. Further testing with different epidemic trajectories will allow us to test whether this better fitting of first order methods is merely an artefact from this true epidemic curve.

For the kappa paramter, the average RMSE values during this prediction period are displayed in table (6). These results are similar to the results for prevalence and incidence during this prediction period, in that at each sample size the best fitting technique is the first order penalized random walk, while the first order techniques also outperform their respective second order penalized versions. The improvement with increasing sample size

Table 5: RMSE of Incidence for fitted against true values during 2015:2020 prediction period

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0184568	0.0253233	0.0144984	0.0218784
500	0.0139638	0.0199397	0.0079612	0.0093225
1000	0.0108674	0.0156428	0.0067635	0.0076223
5000	0.0081842	0.0106920	0.0036823	0.0040593

isn’t as clear cut as the previous two parameters, with all RMSE values the same order of magnitude for each of the sample sizes in this case.

Table 6: RMSE of Kappa for fitted against true values during the 2015:2020 prediction period

Sample size	Spline First Order	Spline Second Order	RW first order	RW second order
100	0.0608326	0.0807562	0.0495436	0.0692751
500	0.0496149	0.0689270	0.0307835	0.0343041
1000	0.0400896	0.0569184	0.0269582	0.0287582
5000	0.0322347	0.0420635	0.0153424	0.0166802