Modelling Fertility Rates via Scoring Rule Minimisation

David Huk

1 Introduction

Human fertility is a complex process influenced by biological, social and economic factors, making it difficult to forecast. However, being able to accurately forecast fertility levels is crucial in assessing demand to predicting population size and age structure. As such, fertility forecasts have severe implications on the private and public sector. Therefore, to help manage the risks associated with forecast deviations, it is considered advantageous to consider methodologies able to capture the uncertainty surrounding fertility levels.

In Hilton et al. [2019], the authors employ a Bayesian approach to the problem by decomposing the forecast in two parts. In one part, they apply a mixture model of smooth parametric curves to capture the pattern of fertility with respect to age. In a second part, the evolution of parameters of the aforementioned mixtures as well as the level of fertility specific to a cohort is modelled with a time-series propagating them through time. Their parametric approach preserves parsimony in the model while also accounting for uncertainty. Additionally, it improves on existing work as the mixtures and time series models are fitted simultaneously.

In this short essay we follow in the steps of the aforementioned paper. Using data from the Human Fertility Database, we will model fertility levels in the USA by a mixture model with time-varying parameters. We will follow the same approach of fitting jointly the two parts under a Bayesian paradigm. However, we use a different estimation procedure, relying on Approximate Bayesian Computation (ABC) in order to infer parameters. Furthermore, we use a Scoring Rule-based criterion to assess data discrepancy inside the ABC algorithm, motivated by recent works on the use of divergences coupled with ABC. In the following section, we introduce ABC as a simulation-based approach to inference. We discuss common issues with ABC methods and motivate the use of Scoring Rules as statistical divergences within that context, resulting in our proposed method. In section 3, we apply this estimation procedure to model fertility data. Finally, in section 4, we discuss our results.

2 ABC: Inference by Simulations

The general idea is as follows. For a given parametric model, we wish to infer some parameter of interest θ. We have observed data $\boldsymbol{y}$ with an assumed distribution $p(y|\theta)$ and prior $p(\theta)$, both of which we can sample from. We begin by drawing a sample parameter θ_1 from the prior followed by a sample realisation y' conditional on that parameter. Next, we compare the simulated y' against observed data $\boldsymbol{y}$ ultimately deciding whether or not to retain the parameter θ_1 based on the similarity between simulated and observed data. When repeating this process many times, we end up with a set of parameter values $\{\theta_1, \ldots, \theta_m\}$ which are representative of our posterior density of θ as we restrict the threshold of acceptance for proposed θ_is. (See eg. Lintusaari et al. [2017] for further details.)

This method gives us a way of performing inference while avoiding any explicit likelihood computations; all we are doing is comparing samples of data. As it may not be obvious from the start how many acceptances one might expect, a good strategy is to run the algorithm for a fixed number of iterations and accept the best 5 or 10% of parameters based on how computationally expensive it is. Note that "similar enough" can mean many different ways of comparing data samples.

A common choice for comparing the simulations to observed data are summary statistics, which can lead to information loss. On the other hand, using more data can introduce unwanted variance into the inference. A trade-off ensues between reducing information loss and reducing the variation between simulation-observation pairs Fearnhead and Prangle [2012]. In search for a compromise, the authors in Bernton et al. [2019] introduce a method based on the Wasserstein distance in order to compare simulations to observations. In Jiang [2018] and Wang et al. [2022], the Kullback-Leibler divergence is used as a discrepancy measure for comparison of simulated and observed data. Another such attempt is done by Park et al. [2016], where the authors use the maximum mean discrepancy as a dissimilarity measure. All of these are examples of divergences. But first, let us discuss what a divergence is and how to use it on data.

Statistical Divergences for Sample Comparison Assume we have observed data from distribution $\mathcal{P}^*$ and want to choose a parameter $\boldsymbol{\theta}$ which parameterises a second distribution $\mathcal{P}^{\boldsymbol{\theta}}$ such that the two distributions are as close to each other as possible. A divergence D is then defined as a function of two distributions such that (i) $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) \geq 0$ and (ii) $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) = 0 \iff \mathcal{P}^* = \mathcal{P}^{\boldsymbol{\theta}}$. Therefore, a divergence can be used to optimise a parameter to recover the best possible model $\mathcal{P}^{\boldsymbol{\theta}}$ for the data generating distribution $\mathcal{P}^*$.
An example divergence is the maximum mean discrepancy (MMD), which for a given choice of kernel k is defined as:

$$\mathrm{MMD}^2\left(\mathcal{P}_x, \mathcal{P}_y\right) = \mathbb{E}_{X_1}\mathbb{E}_{X_2}k\left(X_1, X_2\right) + \mathbb{E}_{Y_1}\mathbb{E}_{Y_2}k\left(Y_1, Y_2\right) - 2\mathbb{E}_{X_1}\mathbb{E}_{Y}k(X_1, Y_1), \tag{1}$$

where X_1, X_2 and Y_1, Y_2 are distributed according to two distributions $\mathcal{P}_x$ and $\mathcal{P}_y$ respectively. Due to the expectation with respect to the distributions, it is straightforward to extend the MMD to an empirical approximation in order to compare samples. As motivated in Bernton et al. [2019] and Park et al. [2016], using these divergences in an ABC algorithm avoids the issues of information loss due to the use of summaries while also being generalisable to high dimensions.

Another possible choice of divergences is scoring rules (SRs). As defined in Gneiting and Raftery [2007], a scoring rule $S(\mathcal{P}, \mathbf{x})$ is a function between a distribution $\mathcal{P}^{\boldsymbol{\theta}}$ and observed data $\mathbf{x}$ as a realisation of a random variable $X \sim \mathcal{P}^*$. Then the *expected scoring rule* is defined as $S(\mathcal{P}, \mathbf{x}) := \mathbb{E}_{Y\sim\mathcal{P}^*}S(\mathcal{P}^{\boldsymbol{\theta}}, Y)$. The SR is termed *proper* if relative to a set of distributions $\mathbf{P}$, if the expected SR is minimised when $\mathcal{P}^* = \mathcal{P}^{\boldsymbol{\theta}}$:

$$S\left(\mathcal{P}^*, \mathcal{P}^*\right) \leq S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall\, \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P}.$$

Furthermore, a SR is termed *strictly proper*, if the minimisation above is unique:

$$S\left(\mathcal{P}^*, \mathcal{P}^*\right) < S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall\, \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P} \text{ s.t. } \mathcal{P}^* \neq \mathcal{P}^{\boldsymbol{\theta}}.$$

By considering the quantity $D_{SR}(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) := S\left(\mathcal{P}^*, \mathcal{P}^{\boldsymbol{\theta}}\right) - S\left(\mathcal{P}^*, \mathcal{P}^*\right)$ for a strictly proper SR, one can see that it defines a divergence. Indeed, (i) is verified by the SR being proper and (ii) is verified by the additional requirement of being strictly proper. This permits the use of strictly proper SRs as divergences to do inference on parameters of a distribution. For instance, one such strictly proper scoring rule is the kernel score, which for an appropriate choice of kernel is identical to the

MMD. As for the choice of SR we will use, we introduce the Energy Score as:

$$S_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right) = 2 \cdot \mathbb{E}_{\mathbf{X}' \sim \mathcal{P}^{\boldsymbol{\theta}}} \|\mathbf{X}' - \mathbf{x}\|_2^{\beta} - \mathbb{E}_{\mathbf{X}'_1, \mathbf{X}'_2 \sim \mathcal{P}^{\boldsymbol{\theta}}} \|\mathbf{X}'_1 - \mathbf{X}'_2\|_2^{\beta}$$

where $\beta \in (0, 2)$ is a hyperparameter regulating the severity of the divergence for incorrect distributions $\mathcal{P}^{\boldsymbol{\theta}}$. The energy SR is a strictly proper SR for the class of $\mathbf{P}$ such that $\mathbb{E}_{\mathbf{X}' \sim P} \|\mathbf{X}'\|^{\beta} < \infty$, see Gneiting and Raftery [2007]. It is possible to obtain unbiased estimates of $S_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right)$ by repeated sampling from $\mathcal{P}^{\boldsymbol{\theta}}$, as:

$$\hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right) = \frac{2}{m} \sum_{j=1}^{m} \|\mathbf{x}_j - \mathbf{y}\|_2^{\beta} - \frac{1}{m(m-1)} \sum_{\substack{k=1 \\ k \neq j}}^{m} \|\mathbf{x}_j - \mathbf{x}_k\|_2^{\beta}$$

where $\mathbf{y}$ are observations and $\mathbf{x}$ are samples. As such, we have a method for inferring parameters of a distribution by comparing the observations to simulated samples. This is equivalent to the following optimisation problem for choosing $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right)$ since the first part of D_{SR} is constant in $\boldsymbol{\theta}$.

Proposed Approach Our idea is rather simple and follows from the previous works using divergences in ABC. What we want to do is perform ABC-based inference by using divergences as a discrepancy measure between simulated and observed data. By comparing samples to observed data, we can get a score estimate corresponding to set of parameters. As the solution to the optimisation problem is the θ which minimises the SR, we then pick the sets of parameters associated with the lowest SR estimates. Effectively, we replace the similarity assessment by instead evaluating the SR and using that to rank the samples.

As we are interested in a probabilistic model, the use of divergences is alluring as it operates on the probability distributions directly. Furthermore, Scoring Rules are a known tool in the forecasting literature to assess precision and uncertainty, be it in precipitation (eg. Harris et al. [2022]) or fertility (eg .Ellison et al. [2020]). A common diagnostic is the Continuous Ranked Probability Score (CRPS), which can be understood as a generalisation of the absolute error metric, to which it reduces when the forecast is deterministic (Gneiting and Raftery [2007]). The energy score is a generalisation of the CRPS to which it reduces when $\beta = 1$ with a univariate distribution. In this work, we follow the convention set in previous applications of the energy score as in Pacchiardi and Dutta [2021] and Pacchiardi et al. [2021] by fixing $\beta = 1$.

Intuitively, we take advantage of recent advances in ABC methodology to use well-known probabilistic forecasting tools as optimisation objectives to perform inference.

3 Application to Fertility Data

The data we use is taken from the Human Fertility Database. It contains age-specific fertility rates (ASFR) in the USA by year. The age brackets start at 12 years and below and go up to 55 years and above. We will restrict this study to the last 10 years period (2012-2021) of the data to facilitate computations, the goal being to assess the effectiveness of the modelling approach rather than provide a complete study of fertility rates. We follow the findings of Hilton et al. [2019], assuming a Gamma-Gamma mixture model, formulating our model as follows.

For given year t and age a, we model fertility $y_{t,a}$ using a mixture of Gamma-Gamma densities as

$$f(y_{t,a}) = \frac{\gamma_t \cdot \mathcal{G}_1(y_{t,a} - 12; \alpha_t^{(1)}, \beta_t^{(1)}) + (1 - \gamma t) \cdot \mathcal{G}_2(y_{t,a} - 12; \alpha_t^{(2)}, \beta_t^{(2)})}{\eta_t}$$

where γ_t is the mixture for the Gamma densities $\mathcal{G}_1, \mathcal{G}_2$ and η_t is the sum of ASFRs for that year to normalise to 1 (we do not model η_t). We deviate from the aforementioned study by simplifying the assumptions of the model and omitting explicit cohort modelling of the parameters in favor of time-dependence only. In order to ensure that our parameters are on the correct scale, we use the following transformations to obtain the vector $\boldsymbol{\theta}_t = (ln(\frac{\gamma_t}{1-\gamma_t}), ln(\alpha_t^1), ln(\beta_t^1), ln(\alpha_t^2), ln(\beta_t^2)) = (\theta_t^{(1)}, \theta_t^{(2)}, \theta_t^{(3)}, \theta_t^{(4)}, \theta_t^{(5)}) \in \mathbb{R}^5$, which are the quantities we actually model. We do not model the evolution of parameters through time (An attempt was made, but the results in figure 4 indicate that either more ABC iterations are needed to converge to the posteriors, or more care needs to be taken when specifying the priors). Finally, we assume $\mathcal{N}(2.9, 1)$ and $\mathcal{N}(3.1, 1)$ priors for θ_2 and θ_4 respectively to center the locations at appropriate values, along with $\mathcal{N}(0, 1)$ priors for the remaining parameters. In order to avoid identifiability issues with the mixtures, we enforce $\theta_2 > \theta_4$, $\forall t$.

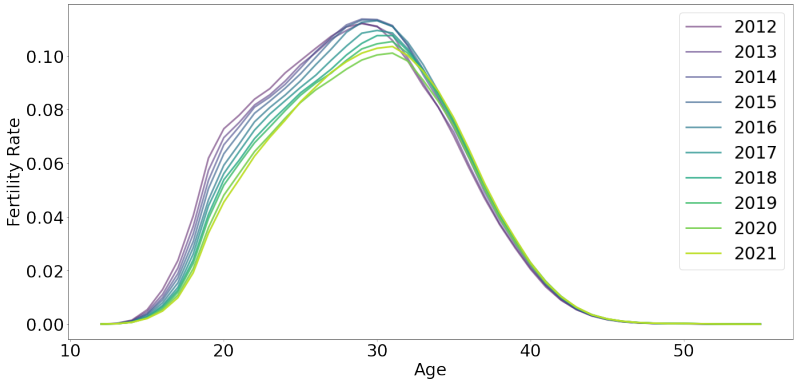

Figure 1: Observed ASFR for years 2012 to 2021.

In figure 1, we show the observed ASFR for the period considered. There is a hint of multi-modality, but it is not very pronounced.

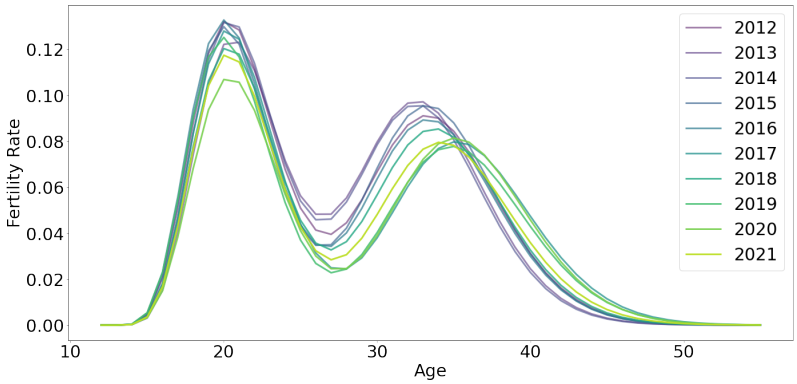

Figure 2: Modelled ASFR for years 2012 to 2021.

In figure 2, we show our models ASFR where the parameters used for the mixtures are medians of posterior approximations, that is medians of retained values for each parameter and year combination.

Lastly, in figure 3, we show all posteriors for the 5 parameters across time, with medians shown in orange.

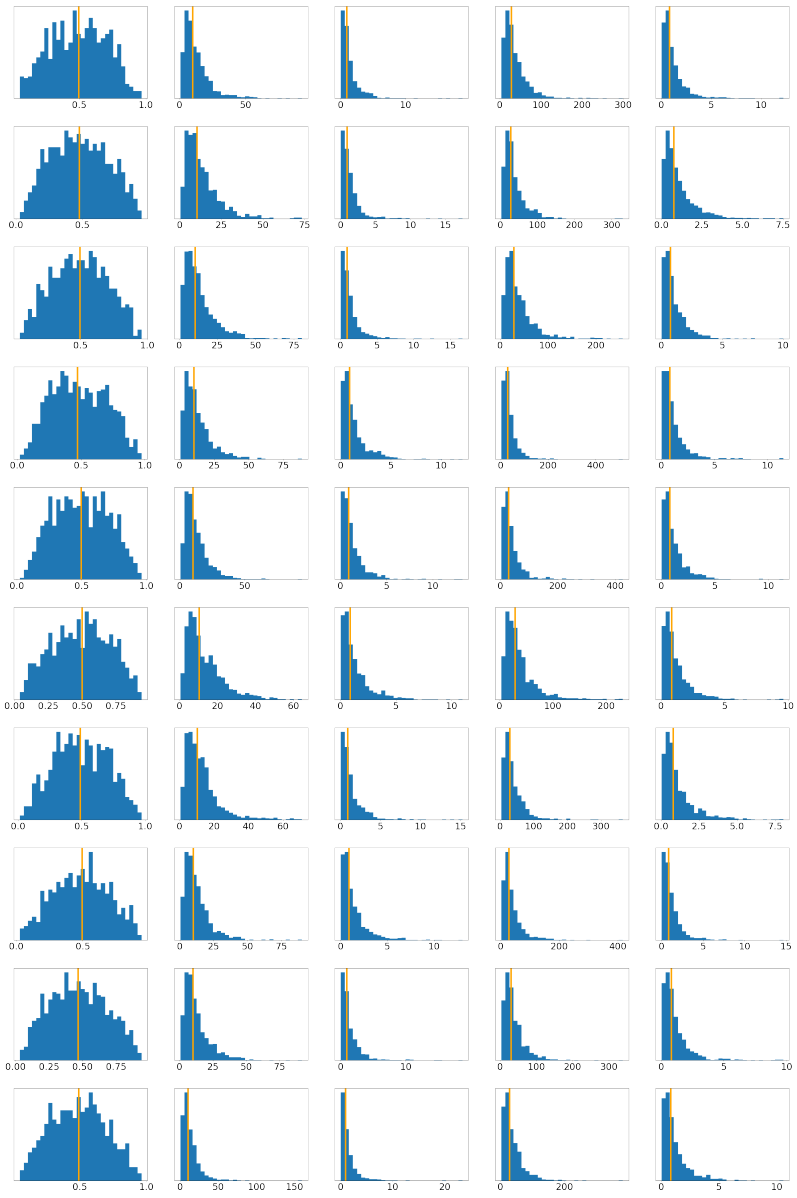

Figure 3: Posterior approximations for each parameter obtained with the ABC with SR approach. Medians are shown in orange.

4 Conclusion

In this work, we explored the use of SRs as a measure of data discrepancy within an ABC algorithm. We applied our approach to fertility data in the USA over a 10-year period. We made multiple simplifying assumptions for the model compared to what was done in Hilton et al. [2019], however the modelling task remains non-trivial. The choice of priors seems to have a great impact on the posterior values, as from our results, it seems the data was overwhelmed by the prior and more iterations are needed to correctly approximate the posteriors.

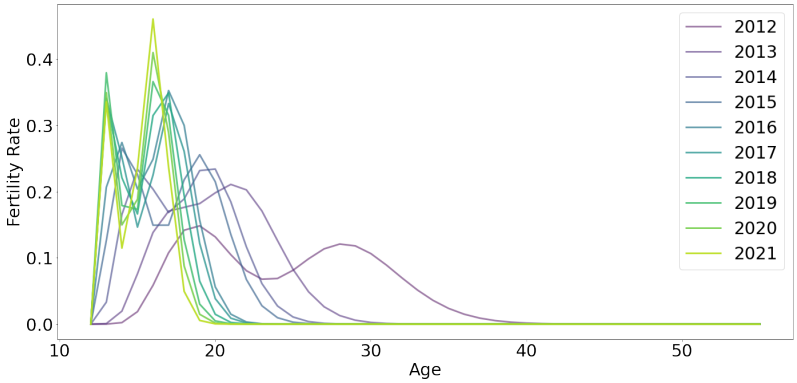

Figure 4: Modelled ASFR for years 2012 to 2021 when modelling evolution through time.

References

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate bayesian computation with the wasserstein distance. *arXiv preprint arXiv:1905.03747*, 2019.

J. Ellison, E. Dodd, and J. J. Forster. Forecasting of cohort fertility under a hierarchical bayesian approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3): 829–856, 2020.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL `https://doi.org/10.1198/016214506000001437`.

L. Harris, A. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.

J. Hilton, E. Dodd, J. J. Forster, P. W. Smith, and J. Bijak. Forecasting fertility with parametric mixture models. *arXiv preprint arXiv:1909.09545*, 2019.

B. Jiang. Approximate bayesian computation with kullback-leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics*, pages 1711–1721. PMLR, 2018.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82, 2017.

L. Pacchiardi and R. Dutta. Generalized bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*, 2021.

L. Pacchiardi, R. Adewoyin, P. Dueben, and R. Dutta. Probabilistic forecasting with conditional generative networks via scoring rule minimization. *arXiv preprint arXiv:2112.08217*, 2021.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-abc: Approximate bayesian computation with kernel embeddings. In *Artificial intelligence and statistics*, pages 398–407. PMLR, 2016.

Y. Wang, T. Kaji, and V. Rockova. Approximate bayesian computation via classification. *Journal of Machine Learning Research*, 23(350):1–49, 2022.