Is Less More? Data-Driven Dimensionality Reduction in Parametric ASFR Models

paper submitted to the
Population Association of America 2011 Annual Meeting
to be held in Washington, DC, March 31 – April 2, 2011

Bilal F. Barakat
Vienna Institute of Demography
bilal.barakat@oeaw.ac.at

Extended Abstract

Introduction

The timing of fertility behaviour can be summarized by the schedule of age-specific fertility rates (ASFR). Numerous parametric models for indexing empirical ASFR schedules by a moderate number of parameters have been proposed, allowing for more convenient interpolation, inference, and specification of projection assumptions. Reducing the number of parameters involves a number of trade-offs, not least between expressiveness and parsimony.

Recent proposals have essentially converged conceptually to using 4 parameters to specify the scale, location of peak, and the shape of the ascending and descending slopes. The difference rests largely in whether the slopes are specified as polynomial splines [7], exponential [4], or logistic [6]. A competitive 3-parameter ASFR model, however, has remained elusive, because the most obvious path to eliminate one of the generic parameters above, namely assuming a symmetric or at least deterministic relationship between the two slopes, clearly fails to capture the diversity in empirical schedules.

Motivated by the insight that a more concise model would need to exploit complex dependencies between the above parameters, and that unaided human cognition is notoriously poor at high-dimensional reasoning, I propose to push for more parsimonious ASFR models through algorithmic methods. Concretely, I explore a novel application of a data-driven dimensionality reduction technique, namely Independent Component Analysis (ICA), to *embed* a 'reduced' ASFR model of low parametric dimension within an arbitrary parametric ASFR 'full' model of higher dimension. This yields two functions: one to project the parameters of the full model onto a subspace, and another to conversely express a model described by the ICs in terms of the original, higher-dimensional parametrization.

Doing so combines the statistical benefits of parsimony of the reduced model with the expressiveness and transparent interpretation of the full one. This approach frees human experts from the intrinsically algorithmic task of identifying and avoiding statistical redundancy. Instead, they may focus on devising ASFR models that are meaningful with respect to the substantive demographic concepts and able to reproduce a wide range of human fertility timing behavior, without excessive concern for concision. The subsequent data-driven techniques can exploit statistical dependencies between parameters in the full model much more efficiently than manual 'fine tuning'.

In contrast to other efforts to identify the 'optimal' number of parameters for ASFR modeling according to general model selection criteria such as the AIC [1], the aim is not to balance fit (to empirical ASFR schedules) and model dimension, but to maximize fit given the constraint that the number of parameters should be 2 or 3 at most. This reflects a particular kind of application, where a robustly estimable model with reasonable fit is preferable to a better fitting model that cannot be reliably estimated in practice. In particular, this includes situations where fertility rates are provided only for five-year age groups, implying the need to estimate the model given only 7 data points, or possibly even fewer, and if the data are messy and incomplete, say historical data from developing countries.

Methods

A range of models of ASFR schedules are fitted to a set of empirical schedules. For each model, an ICA is performed on the resulting set of parameters, using a Maximum Likelihood algorithm in the statistical computing language 'R' [8, 5]. Popular in signal processing, ICA is related to the more widely known technique of Principal Component Analysis (PCA) [3]. Simply put, the difference is that PCA

seeks uncorrelated factors, whereas ICA seeks statistically independent factors, a stronger condition. Statistical independence is highly desirable in parsimonious model parameters, because it implies zero mutual information between them, eliminating redundancy. The result is a smaller number of ICs that 'span' the subspace of the full QS parameter space that contains (or is close to) the majority of empirical schedules, leaving out unrealistic combinations of QS parameter values.
The empirical (period) ASFR schedules on which the fits of different models are assessed are drawn from the US Census Bureau's International Data Base (IDB). The cover 226 populations and for each country/territory, 7 data-points are provided, for each five-year age group between 15 and 49. These are the same as those used by Schmertmann in the paper introducing his QS model [7].
The focus rests on reducing *fully parametric* models. This excludes semi-parametric models, such as the Brass relational Gompertz model and the Coale-Trussell model, that parametrize deviations from a 'standard' schedule. The reason is that in comparisons of parsimony, semi-parametric models 'muddy the waters': it depends on the application whether the rates defining the 'standard' should be counted as parameters. To some extent, it is merely a matter of convention that in application, users feel free to create customized Brass-type models by supplying their own standard schedule, while the Coale-Trussell standard is normally taken as given.
Not all fully parametric models that have been proposed are included; the aim here is not a competitive comparison of the greatest possible number of models. The purpose is to study dimensionality reduction, so a range of established parametric models with different numbers of parameters were considered. These are Hoem's Beta model (five parameters) [2], Hoem's Gamma, Schmertmann's Quadratic Splines (QS) and the Peristera and Kostaki [4] exponential model (four parameters), as well as the Hadwiger model (three parameters) [9].
For reasons of space, only the QS model is discussed here, both because of its widespread application and because it is designed to be parsimonious but to have maximally interpretable parameters (youngest age of non-zero fertility, age of peak fertility, age of half-way decline from peak, level of peak). This creates the greatest conceptual distance to the purely synthetic parameters of the IC reduced model, thus clarifying the discussion.
The reduced models for QS are QS.IC3 and QS.IC2, where the four parameters of QS are reduced to 3 and 2 ICs respectively, as well as QS.IC2+1, where the three shape parameters of QS are reduced to 2 ICs and the scale parameter remains as is.
In fitting continuous ASFR models to the five-year empirical schedules, the function value for each age group is taken to be the average across mid-ages in each group. The optimization criterion is the minimal (unweighted) sum of squared errors across age groups. However, the qualitative assessment and graphs are based on relative error (RE); the scale of the latter makes the differences between models and the practical significance of these differences more easily interpretable.
Noisy and incomplete schedules are also fitted. For the former, the empirical schedules are perturbed by multiplicative Gaussian noise. For the curves fit to the perturbed data, the error is calculated relative to the noiseless data. The average error over 100 random runs is calculated.
The sensitivity of different models to missing data is assessed by fitting to the empirical schedules with one data point deleted. This results in seven estimated schedules per empirical schedule — one per omitted age group. The fit to the empirical schedule is then evaluated by calculating the error at each age using the predicted value from the schedule that was estimated with that age omitted.
Technically, the data are being used twice in the IC models, because these are a function of the set of QS parameters it was estimated on, which in turn depend on the data. A 10-fold cross-validation (in other words, performing the ICA on the ten 90% (random) subsets of countries such that each country is included nine times) shows only minimal variation in the QS.IC models. This implies that the models' performance on a given empirical schedule is not overestimated significantly if the initial ICA is performed on the complete set of empirical schedules.

Results

The results confirm the usefulness of the novel, data-driven approach.
To begin with, despite being the output of a 'black box' algorithm, the independent components of the reduced QS model, for example, lend themselves readily to meaningful interpretation. The axes of the components in the QS parameter space are illustrated by curves well along either direction of the central schedule (the 'origin' of the IC coordinate system). The three main axes in this case correspond to: 'early start' vs 'late start', 'early peak' vs 'late peak' timing, and 'high rapid rise, slow decline' vs 'low slow rise, rapid decline' shape dimensions respectively. These dimensions are not pure, but that is of course

precisely the point: they overlap in an optimally efficient way.
In terms of fitting performance, on the clean set of schedules, the reduced models such as QS.IC3 by construction cannot match the fit of the corresponding full models (QS in this case), because they only index a proper subset of the latter's parameter space. Nevertheless, QS.IC3 is competitive with *other* parametric models with a larger number of parameters. In particular, at around 4% RE, 3-parameter QS.IC3 matches the fit of the 4-parameter Hoem Gamma model (as well as the fit reported by Schmertmann [7] for the 4-parameter Coale-Trussell model on the same data, not reimplemented here). It also outperforms the other 3-parameter model in the comparison, namely Hadwiger, by a wide margin. On the noisy data, 3-parameter QS.IC3 is competitive with full QS, although it still exhibits more cases of outlying very poor fits (but see the discussion of outliers below). In fact, the reduced model slightly outperforms the full model on the vast majority of schedules. Figure 1 illustrates these results by way of example.
The gap diminishes further in the comparison with missing data. Here, even the 2-parameter IC model performs only somewhat worse than QS and exceeds its fit in a large number of cases.
The issue of outliers, as in complete failures to achieve a reasonable fit, are discussed. While QS.IC3 performs reasonably well overall, there is seemingly very poor performance in individual cases, such as North Korea, where the relative error on clean data of QS is 6.2% while that of QS.IC3 is 37.4%. However, the difference in RE is misleading: in fact both models' fits are qualitatively unacceptable. We see that the empirical schedule shows a very steep increase in the first half. The standard QS model achieves a very close fit, but in doing so overshoots considerably. The QS.IC model on the other hand undershoots, because it cannot produce such behavior within its parameter space. This results in a large RE, which is almost entirely due to the discrepancy at the mode.
Despite the large difference in RE, it is not clear that for all purposes the estimates produced by the QS model are actually preferable in the Korean case. The implied schedule with its sudden extreme drop at a ≪threshold age≫ is implausible. Indeed, this case illustrates one of the benefits of the reduced IC model: by construction, it will fail to achieve a good fit on schedules that are far from the 'centre' of the set of schedules it was originally estimated on. This means that large errors are guaranteed to be associated with atypical schedules.
The QS.IC model is not per se unable to produce narrowly peaked schedules such as the Korean ones. However, by construction, reducing the number of parameters and using the remaining degrees of freedom efficiently requires a focus on typical data, and the most 'outlying' schedules will be matched least well. The Korean schedules are outliers in terms of their value of $p - a$, but schedules outlying in terms of any other combination of parameters too would, in principle, be poorly matched. If narrow peaks were more typical of ASFR schedules, they would be included in the schedule space spanned by the three QS.IC parameters.

Discussion

In terms of application, the results suggest that the choice between the full and reduced model depend both on the quality of the available data and on the intended application.
The standard QS model, with its closer fit under ideal circumstances and transparent parameters, is more appropriate when the aim is to interpolate high quality data or as a specification of assumptions for projections, and similar applications.
The QS.IC3 models presented here — with less tight, but still reasonable fit — which is more robust and whose errors correlate with the representativeness of the data to be fitted, and its more parsimonious but in-transparent parameters, is, in principle, more suited to problems of inference from imperfect data.
With imperfect data, there appears to be little or no benefit to using more than 3 degrees of freedom to estimate ASFR schedules, at least when faced with data for 5-year age groups and with respect to the relative error criterion.
Despite the in-transparent parameters of the reduced models, a comparative analysis of their performance does allow substantive conclusions. For example, the relative performance of QS.IC2+1 where the scale parameter r is treated independently compared to QS.IC3, where the shared information between r and the other parameters is taken into account, confirms that scale is indeed not independent from shape in ASFR schedules. Note that this is not a tautology. The 'level' here is not a measure such as TFR that can only attain its maximum if childbearing starts early. The parameter R is merely the maximum of the ASFR schedule; clearly a schedule of any shape can in principle be scaled to any level.
In the full paper I discuss in detail the difference between projecting onto a subset of the ICs on the one hand, and estimating a smaller number of ICs to begin with on the other, as well as why the former

approach is preferable for present purposes.
The present analysis can be seen as a first step in an Empirical Bayes approach to ASFR modeling. Not necessarily in the sense of combining different ASFR models (although this is certainly possible), but in the estimation of a given model such as QS. Because the ICs are by construction maximally independent (linear combinations of the original parameters), it is much more straightforward to specify a Bayesian QS model in terms of independent priors on the ICs, rather than as a complicated joint prior on the original parameters. Such IC priors could be based on their empirical distribution in the set of reference ASFR schedules. In this context, the present QS.IC3 and QS.IC2 models can be interpreted as corresponding to crude priors that are non-informative on a 3-dimensional and 2-dimensional (hyper)plane respectively, and zero outside of it. While this perspective awaits formalization in a follow-up paper, it is discussed here because it already sheds light on issues such as the outliers discussed above as representing schedules with low prior probability.

References

[1] Gayawan, E. et al.: "Modeling fertility curves in Africa". In: *Demographic Research* 22.10 (2010), pp. 211–236.

[2] Hoem, Jan M. et al.: "Experiments in Modelling Recent Danish Fertility Curves". In: *Demography* 18.2 (May 1981), pp. 231–244. URL: `http://www.jstor.org/stable/2061095`.

[3] Hyvaerinen, A.; Karhunen, J and Oja, E: *Independent Component Analysis*. New York (Wiley) 2001.

[4] Peristera, P. and Kostaki, A.: "Modeling fertility in modern populations". In: *Demographic Research, Volume 16* 16 (2008), p. 141.

[5] R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2010. URL: `http://www.R-project.org`.

[6] Rueda-Sabater, Cristina and Alvarez-Esteban, Pedro C.: "The analysis of age-specific fertility patterns via logistic models". In: *Journal of Applied Statistics* 35.9 (2008), pp. 1053–1070.

[7] Schmertmann, C.P.: "A system of model fertility schedules with graphically intuitive parameters". In: *Demographic Research* 9.5 (2003), pp. 82–110.

[8] Teschendorff, Andrew: *mlica: Independent Component Analysis using Maximum Likelihood*. R package version 0.6.1.

[9] Yntema, L.: "On Hadwiger's fertility function". In: *Statistical Review of the Swedish National Central Bureau of Statistics, III* 7 (1969), pp. 113–117.

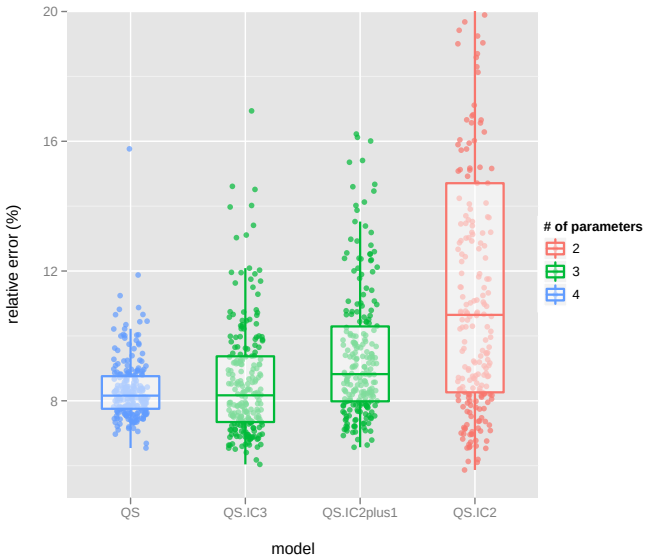

Figure 1: performance of Quadratic Spline full and reduced models on noisy data