Fitting age-specific fertility rates by a skew-symmetric probability density function

Stefano Mazzuco
Bruno Scarpa
Department of Statistical Sciences, University of Padova
Via Cesare Battisti, 241
35121, Padova, Italy
email: mazzuco@stat.unipd.it, scarpa@stat.unipd.it

September 28, 2011

Abstract

Mixture probability density functions had recently been proposed to describe some fertility patterns characterized by a bi-modal shape. These mixture probability density functions appear to be adequate when the fertility pattern is actually bi-modal but less useful when the shape of age-specific fertility rates is unimodal. A further model is proposed based on skew-symmetric probability density functions. This model is both more parsimonious than mixture distributions and more flexible, showing a good fit with several shapes (bi-modal or unimodal) of fertility patterns.

keywords: fertility rates, skew-symmetric, mixture distributions

1 Introduction

It has been recently observed that patterns of fertility of some developed countries show a deviation from the classical bell shaped curve. Some countries, such as Ireland and UK, exhibit an almost bi-modal shape of age-specific fertility rates that classical fertility models (see Hoem et al., 1981) cannot adequately fit. This kind of pattern can be easily captured by a mixture model, assuming that two populations with different fertility patterns are mixed in one. Chandola et al. (1999) have proposed a mixture Hadwiger model with seven parameters[1]. Another proposal has been recently made by Peristera and Kostaki (2007) who first define a simple model based on normal probability density function but having a different variance parameter for ages before and after the mean age, and then a normal mixture model with 6 parameters. Another proposal have been made by Schmertmann (2003), who proposes a piecewise quadratic spline function. The latter shows a very good fit with wide variety of fertility schedules, but 13 parameters are needed to be estimated for this.

In this article a different solution is proposed, basing on the results on skew-normal distribution and its generalization (Azzalini, 1985, 2005): a 4 parameter model can be defined taking the skew-normal probability density function, where the skewness parameter makes - if needed - the function asymettric (as many fertility patterns). The Skew-normal distribution can be generalized by adding a further parameter and thus allowing a bimodal shape of the distribution. For instance, Ma and Genton (2004) call "Flexible Generalized Skew-Normal" (FGSN) a random variable which pdf is defined by adding a new parameter to Skew-Normal pdf, and can possibly have two modes. Therefore, such a model - which is more parsimonious than those proposed by Chandola et al. (1999), Peristera and Kostaki (2007) and Schmertmann (2003) - is potentially flexible enough to exhibit a good fit both when fertility schedule is unimodal and when it is bi-modal. This can be an advantage with respect to Chandola et al. (1999) and Peristera and Kostaki (2007) mixture models, which work reasonably well when the fertility schedule can be actually seen as a mixture between two patterns, but look greatly over-parameterised when it is unimodal and regular.

In the next section, we briefly review the existing models of age-specific fertility rates, whereas in section 3 the skew-normal and skew-syimmetric

[1] Their initial proposal is a 6 parameter model, but Ortega Osona and Kohler (2000) pointed out that an additional parameter is needed

distributions are introduced and the model we propose is defined. In section 4 we show how all these fertility models fit with some real data and in section 5 we discuss upon the results.

2 Modelling fertility schedules

Following Hoem et al. (1981), a fertility curve can be written in the form

$$g(x; R, \theta_2, \ldots, \theta_r) = R \cdot h(x; \theta_2, \ldots, \theta_r) \tag{1}$$

where $h(\cdot; \theta_2, \ldots, \theta_r)$ is a probability density function on the real line with $r-1$ parameters and R is the r-th parameter representing the total fertility rate (TFR). Several specifications of $h(\cdot; \theta_2, \ldots, \theta_r)$ are exposed by Hoem et al. (1981) using the Hadwiger (inverse Gaussian), the Gamma, the Beta, the Coale-Trussel the Brass and the Gompertz pdfs. Moreover Hoem et al. (1981) define two further models based on regression spline and polynomial functions. The model based on spline functions is that giving the best fit to the data but this comes as no surprise, since ten parameters are used. Among the models with fewer parameters the best fit is given by the Gamma, the Coale-Trussel and in fewer cases by the Hadwiger function.

Recently, it has been observed that these models are not adequate to describe fertility pattern which are arising in some developed countries such as UK, Ireland and US. In particular, a marked hump at early ages has been observed in recent fertility patterns like that of Ireland and showed in figure 1. This hump is even more marked when first order birth are considered, (see Peristera and Kostaki, 2007). The above described fertility models cannot describe such bimodal shape properly - they all are unimodal functions - so if we want an accurate representation of these new schedules, new models should be employed.

A first proposal in this sense, has been made by Chandola et al. (1999) who define a "Hadwiger mixture model". The original Hadwiger function is

$$g(x; a, b, c) = \frac{ab}{c}\left(\frac{c}{x}\right)^{\frac{3}{2}} \exp\left\{-b^2\left(\frac{c}{x} + \frac{x}{c} - 2\right)\right\}. \tag{2}$$

Hoem et al. (1981) define the (2) in a slightly different way starting from the Hadwiger pdf and adding the R parameter. However, this definition is equivalent to that presented by Hoem et al. and if parameter a is multiplied by $\sqrt{\pi}$ we get an estimate of TFR (the R parameter according Hoem et al. specification). The mixture model defined by Chandola et al.

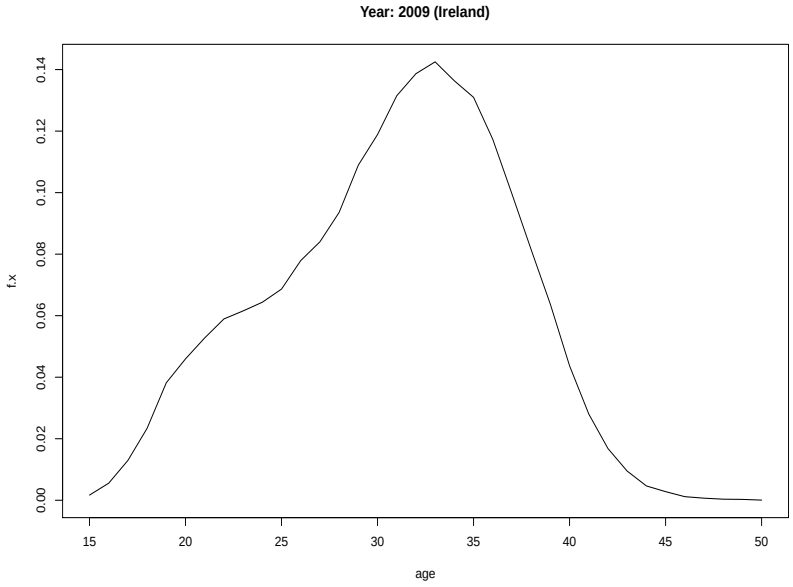

Figure 1: An example of fertility pattern with a marked hump at younger ages. Source: Eurostat.

(1999) is therefore the following

$$g(x;a,m,b_1,b_2,c_1,c_2) = m\frac{ab_1}{c_1}\left(\frac{c_1}{x}\right)^{\frac{3}{2}}\exp\left\{-b_1^2\left(\frac{c_1}{x}+\frac{x}{c_1}-2\right)\right\}+ \\ +(1-m)\frac{ab_2}{c_2}\left(\frac{c_2}{x}\right)^{\frac{3}{2}}\exp\left\{-b_2^2\left(\frac{c_2}{x}+\frac{x}{c_2}-2\right)\right\} \quad (3)$$

where $0 \leq m \leq 1$ is the mixture parameter, determining the sizes of the two underlying populations. The number of parameters is necessarily more than doubled with respect to the (2). The authors show that these parameters can have a demographic interpretation.
Another proposal has been made by Schmertmann (2003) based on quadratic splines

$$g(x;R,\alpha,\beta,\theta_0,\theta_4,t_0,t_4) = R\cdot I(\alpha\leq x\leq\beta)\cdot\sum_{k=0}^{4}\theta_k(x-t_k)^2 \quad (4)$$

where $I(\cdot)$ is the indicator function, α and β the age limits, t_k the spline knots and θ_k the parameters. Thirteen parameters need to be estimated in (4). Schmertmann (2003) also constructed a spline model with only three index ages, thus reducing the number of parameters, but also this

model is thought for fertility schedules with only one mode.
A further model has been proposed by Peristera and Kostaki (2007), basing on normal probability density function with a different variance parameter before and after the mean

$$g(x; c_1, \mu, \sigma_{11}, \sigma_{12}) = c_1 \exp\left\{-\left(\frac{x-\mu}{\sigma(x)}\right)^2\right\} \tag{5}$$

where

$$\sigma(x) = \begin{cases} \sigma_{11} & \text{if} \quad x \leq \mu \\ \sigma_{12} & \text{if} \quad x > \mu \end{cases}$$

Parameter c_1 is related to TFR whereas μ is the location parameter, σ_{11} and σ_{12} are the variances of the distribution before and after μ. Since the (5) cannot capture a two-mode schedule, an extension of it is proposed by Peristera and Kostaki:

$$g(x; c_1, c_2, \mu_1, \mu_2, \sigma_1, \sigma_2) = c_1 \exp\left\{-\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} + c_2 \exp\left\{-\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right\} \tag{6}$$

which is basically a normal mixture model. A third models is suggested, when fertility is stepeer in its left part of the first hump:

$$g(x; c_1, c_2, \mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_2) = c_1 \exp\left\{-\left(\frac{x-\mu_1}{\sigma_1(x)}\right)\right\} + c_2 \exp\left\{-\left(\frac{x-\mu_2}{\sigma_2}\right)\right\} \tag{7}$$

where

$$\sigma_1(x) = \begin{cases} \sigma_{11} & \text{if} \quad x \leq \mu_1 \\ \sigma_{12} & \text{if} \quad x > \mu_1 \end{cases}$$

All the models outlined above are, with a varying degree, appropriate with specific fertility patterns, but it seems that most of them are not adequate for all (or at least the most common) fertility schedules. If Schmertmann (2003) model does not have a good fit with bimodal fertility patterns, Peristera and Kostaki (2007) and Chandola et al. (1999) mixture models may not be adequate (too many and difficult to interpret parameters) when the fertility pattern is unimodal.

3 Skew-normal and skew-symmetric distributions

In this paper, we propose to use a skew-symmetric distribution to fit fertility schedules, and show that this solution is flexible enough for most

fertility patterns, both unimodal and bimodal.
We start defining the skew-normal distribution whose pdf is as follows

$$f(x;\xi,\omega^2,\alpha) = 2\omega^{-1}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left\{\alpha\left(\frac{x-\xi}{\omega}\right)\right\} \tag{8}$$

Properties of (8) have been studied by Azzalini (1985) and by other authors. One interesting feature that has been demonstrated is that (8) is unimodal (Ma and Genton, 2004). Moreover, it clearly appears that, when $\alpha = 0$, (8) reduces to a normal probability density function, which is therefore included as a special case of (8). Figure 2 shows data from US

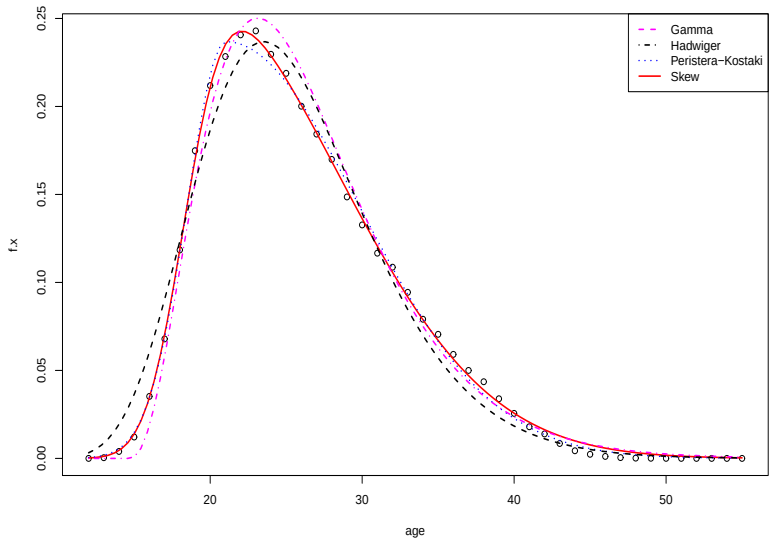

Figure 2: USA Age-specific fertility rates (1963) fitted by several models. Source: Human Fertility Database.

fertility in 1963 (data taken from Human Fertility Database (Human Fertility Database)). The fit provided by the Skew-Normal density is rather good (better than Gamma, Hadwiger, and Peristera-Kostaki ones) suggesting this can be a good model for unimodal fertility schedules. Model (8) can be generalized using the results exposed by Azzalini and Capitanio (2003) and Azzalini (2005). In essence, for any symmetric pdf f_0 and distribution function G with a symmetric density, the function

$$f(x) = 2f_0(x)G\{w(x)\} \tag{9}$$

is a density function for any odd function $w(\cdot)$. if $f_0 = \phi$, $G = \Phi$ and $w(x) = \alpha x$ we get the (8). This result can be used to define what Ma and Genton (2004) call "Flexible Generalized Skew-Normal" (FGSN) distribution:

$$f(x;\xi,\omega^2,\alpha,\beta) = 2\omega^{-1}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left\{\alpha\left(\frac{x-\xi}{\omega}\right)+\beta\left(\frac{x-\xi}{\omega}\right)^3\right\}. \tag{10}$$

Ma and Genton (2004) prove that the pdf (10) can have - at most - two modes and note that in general as the degree of the odd polynomial $w(x)$ increases the number of modes allowed in the pdf increases.

For our purposes, two possible modes are enough, so the (10) is a good candidate to fit bimodal fertility schedules and this generalization is obtained adding only one parameter to the (8). It can be easily shown that if $\beta = 0$ we get again the (8), which is a special case of (10).

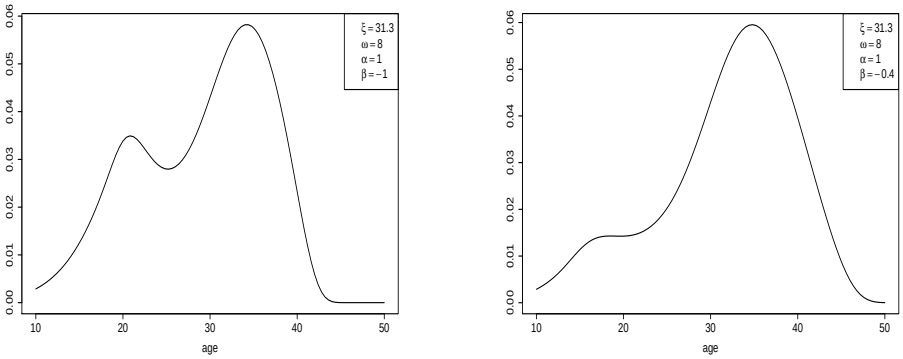

Figure 3: Two examples of FGSN curves.

3.1 Interpretation of parameters

Before fitting the Skew-Symmetric distribution to real data and compare it with other fertility models, we need to explore more in detail the meaning of its parameters and their possible demographic interpretations. The first two parameters (i. e. ξ and ω) are easier to interpret, as they are a location (ξ) and scale (ω) parameters. It should be noted that ξ is not the mean of the distribution (so it cannot be interpreted, as one might be tempted to do, as the average age at childbearing) but it is a function of it, as shown by Arellano-Valle and Azzalini (2008) for the skew-normal.

Similarly, ω is not the variance of the distribution but it is proportional to it.

Interpretation of α and β is more difficult: α is the skewness parameter in the skew-normal distribution, if $\alpha = 0$, but in the skew-symmetric distribution β also contributes to skew the density function. These two parameters are certainly related with the location of the mode (of the two mode), but unfortunately we cannot derive this relation explicitly. We therefore simulated many FGSN distributions keeping ξ and ω fixed and making α and β varying between -5 and +5. For each combination of α and β we look at the locations of the mode(s) and in this way we derive an interpretation of parameters. From our simulations we basically can say that:

- if α and β have the same sign, the resulting pdf has only one mode. In some cases, the pdf shows a small "bulge" but this is never an additional mode
- if α and β have opposite signs, the resulting pdf has two modes
- if the absolute value of β increases, the height of the second modes increases
- the higher the absolute value of α the more distant the two modes between them.

Figure 4 shows 9 the resulting pdfs of nine combinations of values of α and β. Actually, α is kept fixed ($\alpha = 1$) and β varies from 0 to -0.8. In this way we can see how, as β decreases, the second mode increases. The sequence might represent a situation in which the teen-age fertility increases until it creates a second mode in the fertility curve. Basing on these observations, we can draw some conclusions on the possible values that the parameters of the skew-symmetric distribution may assume when fitted with real fertility data.

First, we can expect that if a fertility schedule has an additional hump, like that shown in figure 1, this will likely be not too pronounced. This leads us to expect an absolute value of β not larger than one. Moreover, the additional hump, if exists, will be located in the right side of the distribution, as this is generated by early-age fertility. As far as we know, it is very unlikely to find an additional hump in the left side of the distribution. Therefore we should expect a negative value of β and a positive value of α. Further evaluations can be made by fitting this model to real data.

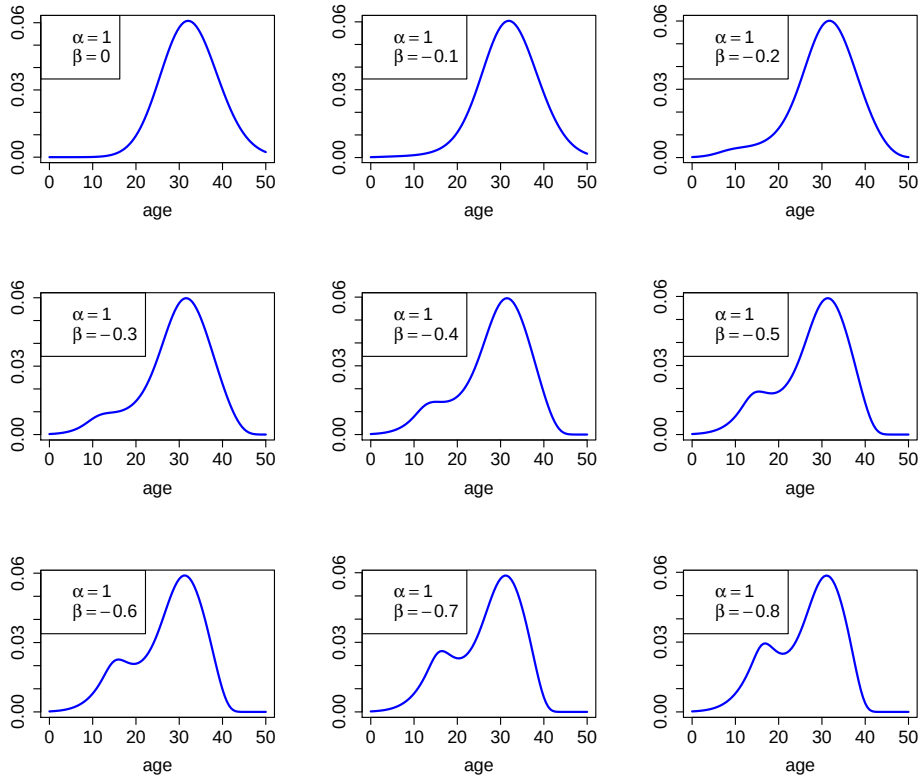

Figure 4: Examples of Skew-symmetric distribution with $\xi = 28$, $\omega = 8$, $\alpha = 1$ and varying values of β

4 Fitting fertility model to real data

Fertility models above described, together with other well-known fertility models, will be fitted to real data form several countries and years, in order to make an evaluation of their quality. We do not consider spline or polynomial models such as those described in Schmertmann (2003) and in Hoem et al. (1981). They undoubtedly provide the best goodness of fit in most of the cases, but, as also Hoem et al. (1981) and Peristera and Kostaki (2007) highlight, the number of parameter they use is too high and too difficult to interpret. We use data from both countries that recently experienced a "bimodal" fertility schedule

(e. g. USA, UK, Ireland) and countries that keep a classic fertility pattern (e. g. Italy, Czech Republic). Data are taken from Human Fertility Database Human Fertility Database (Human Fertility Database), Eurostat (http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/) and Istat (Italian Statistical Institute, see http://demo.istat.it/).
Parameters of fertility models are estimated through non-linear least squares, by minimizing

$$S(R, \theta_2, \ldots, \theta_r) = \sum_{x=b}^{e} \{g(x; R, \theta_2, \ldots, \theta_r) - f_x\}^2 \qquad (11)$$

where f_x is the real age-specific fertility rate, $g(x; R, \theta_2, \ldots, \theta_r)$ is the fertility rate at age x given by the fertility model used, and s and e are the ages at the beginning and at the end of the fertile period, respectively.
We first evaluate the performance of fertility models in a country showing a classical (i. e. with no additional hump) fertility pattern. Italy is among the countries with the lowest rate of teenage fertility and is therefore suited for this first test.

Figure 5 shows the sum of squared residuals of models that have been

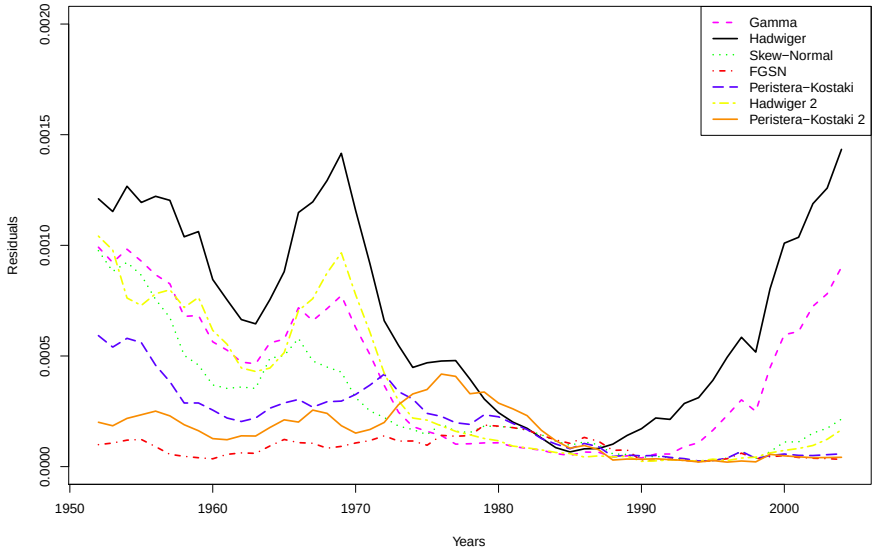

Figure 5: Sum of squared residuals of Fertility models fitted to Italy data (1952-2003)

considered, with respect to Italian fertility data since 1952 up to 2003. In these years, Italian fertility has become of lower intensity (TFR was 2.337 in 1952 and 1.328 in 2003) and mean age at childbearing decreased until the end of 1960s' and increased. Looking at figure 5, it appears that Gamma, Hadwiger and Skew-normal models have a similar pattern but Skew-normal model has almost always a better fit (i. e. a lower sum of square of residuals) especially after 1990, when the shape of Italian fertility pattern starts to become more and more symmetric and therefore more and more similar to a normal distribution. As we said, this is a special case of the skew-normal model ($\alpha = 0$) whereas it is not possible for Gamma and Hadwiger models to become symmetric. Peristera-Kostaki models show a good fit, but the mixture model does not improve a lot the fit of the simplest model. Actually, in some years the mixture model has a slightly worse fit than the simple one. The fit provided by the Hadwiger mixture model is better than that of simple Hadwiger model, although in the first years is worse than Peristera-Kostaki, Skew-Normal and FGSN models.

We fit the same models to fertility data of USA, where, in the last years, a bimodal shape of age pattern has been detected. From the Human Fertility Database we can get fertility data of USA since 1933 up to 2006. In this time span, the age pattern has become less skewed, even though a perfect simmetry has not been reached as occured in Italy. In addition, in the last twenty years, a bimodal shape has appeared, as noticed also by Peristera and Kostaki (2007). Figure 6 shows the sum of squared residuals of models that have been considered, with respect to USA fertility data since 1933 up to 2003. The figures is similar to figure 5 but with significant differences. First, as for figure 5, at a certain point (around 1980) the squared residuals of Gamma and Hadwiger models increase, but this time this is not due to the fact that fertility pattern is becoming symmetric. The reason is that a second hump appears and the Gamma and Hadwiger models are not able to catch it. Also the skew normal and Peristera-Kostaki models are no appropriate fot this shape of fertility rates, and indeed their residuals follow those of Gamma and Hadwiger models. Conversely, the second hump is catched by the mixture models (those proposed by Peristera and Kostaki and that proposed by Chandola et al.) and by the FGSN model, and their residuals does not diverge from the values before 1980.

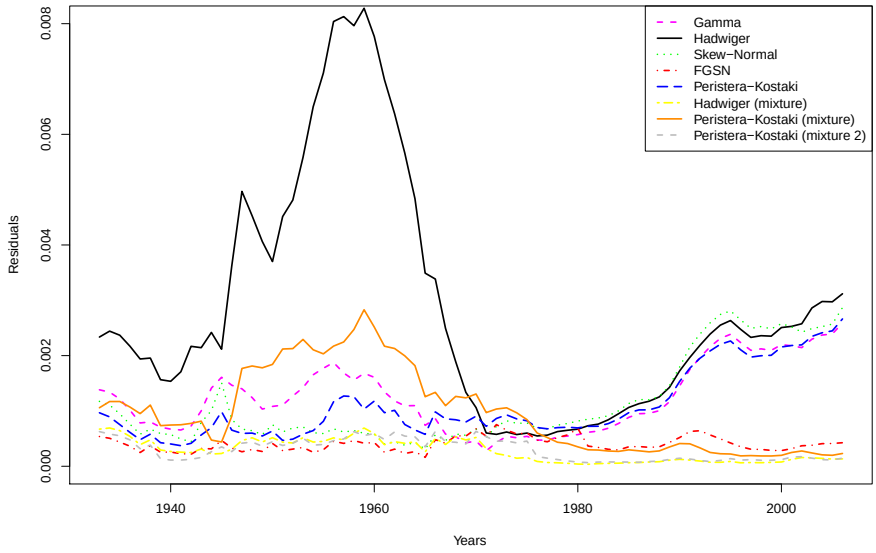

Figure 6: Sum of squared residuals of Fertility models fitted to USA data (1933-2003)

4.1 Intepreting parameters

The quality of a fertility model does not depend only on the goodness of fit with real data. Goodness of fit is certainly important, but a good fertility models also need to provide a useful demographic intepretation of its parameters and their values. This is a particularly necessary feature when the fertility model is used for forecasting. An useful demographic interpretation of parameters can be drawn if their value follows a sensible trend over time, if this does not happen the model is not well specified, at least not for every year the model has been fitted. This is something we have to bear in mind when examining the behaviour of the models outlined before. In appendix we reported the values of parameters of Hadwiger mixture, Peristera-Kostaki mixture and FGSN models fitted with Italy and USA fertility data. By examining these figures, we see that for Italy fertility the FGSN and Peristera-Kostaki mixture models are well specified. All the parameters follow a clear trend without discontinuities, with the exception of β parameter which trend has a discontinuity around 1990.The Hadwiger mixture model, instead, seems to be not suited for Italy data:

in particular, two parameters (a and c_1) remain constant with a jump around 1970. The trend of FGSN parameters tell us that between 1950 and 2003 Italian period fertility has experienced a decreasing intensity (see R parameter) an increasing mean age at childbearing (see ξ parameter) a decreasing heterogeneity (ω) and a "symmetrization" (α falls from 4 to 0). β has increasing trend that needs a more detailed explanation. The fertility pattern of Italy in the fifites (not shown here) shows a significant fertility levels of women aged 30-40. This fertility is not strong enough to create an addiotnal "hump" to the fertility curve but it makes the slope of the second part of the curve smaller. This "late" fertility rapidly declines and this is what generates the trend of β.

For USA data we notice that the behaviour of FGSN and Peristera-Kostaki mixture models parameters is similar to that of Italy. The Hadwiger mixture model parameters do not follow a smooth trend in the first part, but since the 1980s we get a smooth (and interpretable) trend. This behaviour suggests us that the Hadwiger mixture model is a good model when a mixture of fertility behaviours emerges from real data, otherwise it becomes an overparameterised (and possibly not identified) model. The Peristera-Kostaki mixture model shows a smoother trend of its parameters over time. Actually, there is a discontinuity around the eighties but this is an actual discontinuity of fertility data. Indeed, we find a similar discontinuity in the trend of FGSN parameters. From figure 10 we see that in USA the mean age at chidlbearing has constantly increased since 1932 while the variance has decreased. The distribution of births has become more and more symmetric over the years and, more importantly, the trend of β parameter shows that in the last years an additional mode has emerged, a result that was highly expected.

5 Conclusions

In this paper a new fertility model has been proposed, basing on a generalization of the skew-normal distribution (the FGSN model). This generalization allows to detect additional humps that may arise in a fertility distribution as recently has happened in some English-speaking countries. The advantage of the FGSN model is that it is very flexible so that it has a good fit when the fertility pattern is complex (e. g. in several English-speaking countries there is an additional hump at younger ages, due to a high level of teen-age pregnancy) but it is not overparameterised when the

fertility pattern is relatively simple (e.g. in Italy where there is no additional hump). The Hadwiger mixture model is not that flexible, showing a particularly good fit when the fertility pattern is complex, but being overparameterised when the pattern is relatively simple. The main problem is that the Hadwiger mixture model does not seem adequate when in the true data there is no mixture at all. In this sense, the Peristera-Kostaki mixture model works better. It should be noted that the latter is not actually a mixture model – and this is probably the reason for which the model works even when there is no mixture in the true data – but in these cases it is difficult to give a sensible intepretations to its parameters.
It should be also noted, however, that the parameters of FGSN model are not of immediate interpretation, and a reparametrization should be considered. The location parameter is not the mean (but of course it is strictly related to it) and, at the same way, the scale parameter is not the variance so it is difficult for a demographer to say what is a reasonable value of such parameters for a given country. There exists a reparametrization of the skew-normal distribution in which the new parameters are the mean, the variance and the skewness of the distribution (Arellano-Valle and Azzalini, 2008). A possible extension of this work is finding a similar reparametrization for the FGSN distribution in order to have a fertility model with the same flexibility but with parameters easier to intepret.

References

Arellano-Valle, R. B. and A. Azzalini (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis 99*, 1362–1382.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families (with discussion). *Scandinavian Journal of Statistics 32*, 159–188 (C/R 189–200).

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society - Series B 65*(2), 367–389.

Chandola, T., D. Coleman, and R. W. Hiorns (1999). Recent european

fertility patterns: Fitting curves to 'distorted' distributions. *Population Studies 53*, 317–329.

Hoem, J. M., D. Madsen, J. Lovgreen Nielsen, E.-M. Ohlsen, H. O. Hansen, and B. Rennerlman (1981). Experiments in modellin recent danish fertility curves. *Demography 18*(2), 231–244.

Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org, (data downloaded on February 7th, 2011).

Ma, Y. and M. G. Genton (2004). Flexible Class of Skew-Symmetric Distributions. *Scandinavian Journal of Statistics 31*, 459–468.

Ortega Osona, J. A. and H.-P. Kohler (2000). A comment on "recent european fertility patterns: Fitting curves to 'distorted' distributions" by t. chandola, d. a. coleman and r. w. hiorns. *Population Studies 54*, 347–349.

Peristera, P. and A. Kostaki (2007). Modelling fertility in modern populations. *Demographic Research 16*, 141–194.

Schmertmann, C. P. (2003). A system of model fertilty schedules with graphically intuitive parameters. *Demographic Research 9*, 82–110.

Appendix

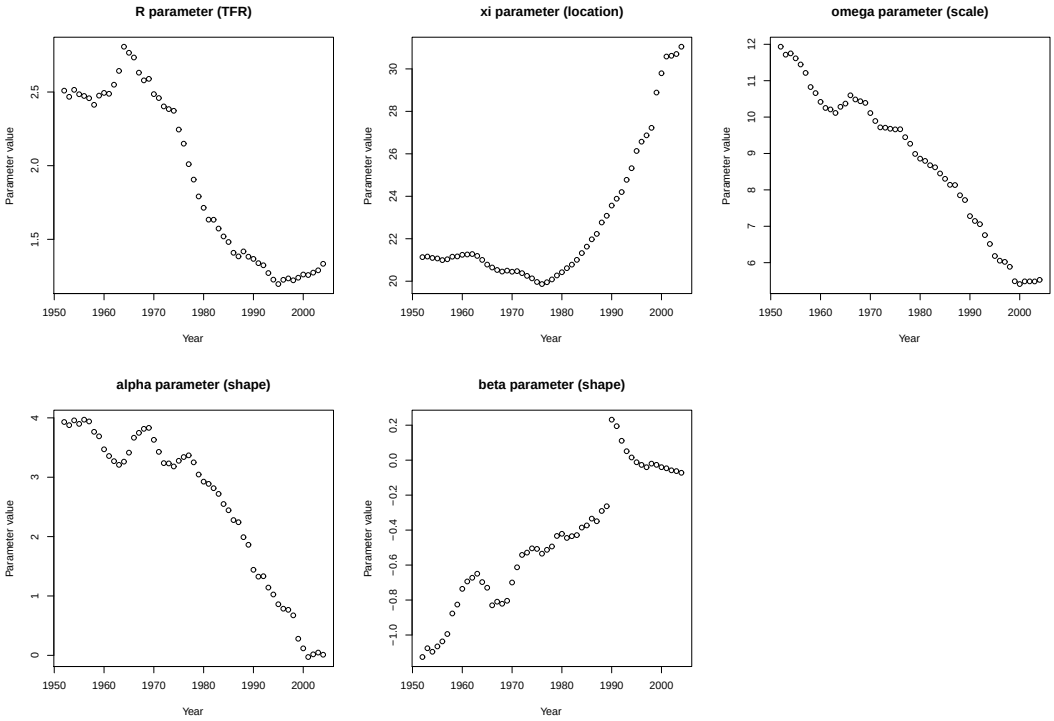

Figure 7: Trends of parameters estimated from the FGSN model. Italy (1952-2003)

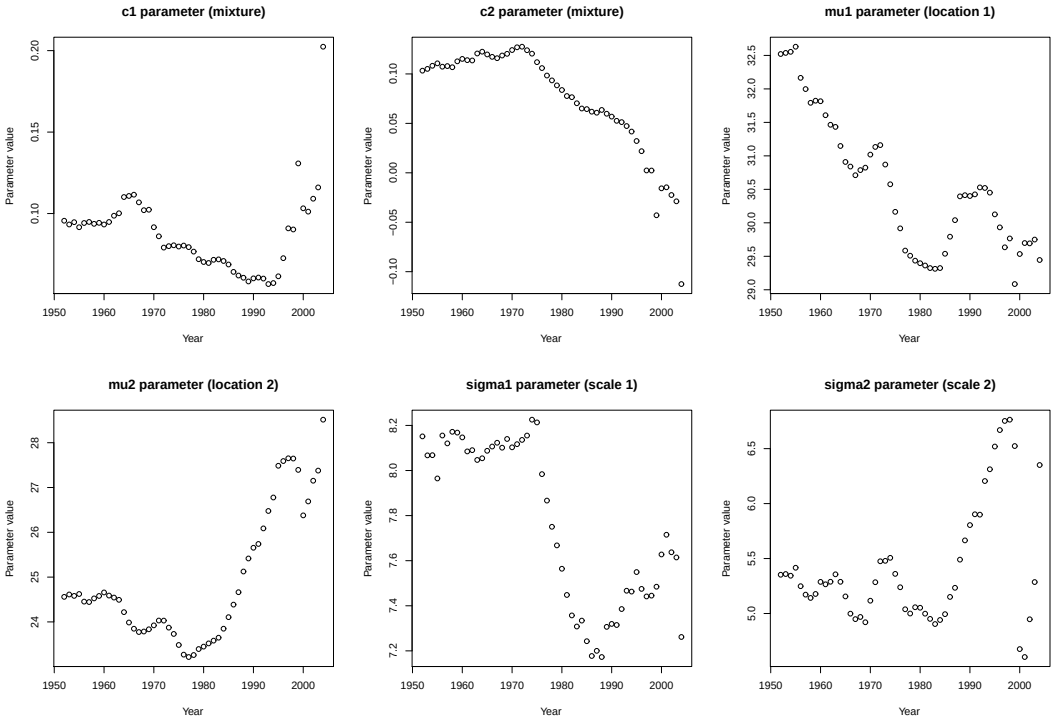

Figure 8: Trends of parameters estimated from the Peristera-Kostaki mixture model. Italy (1952-2003)

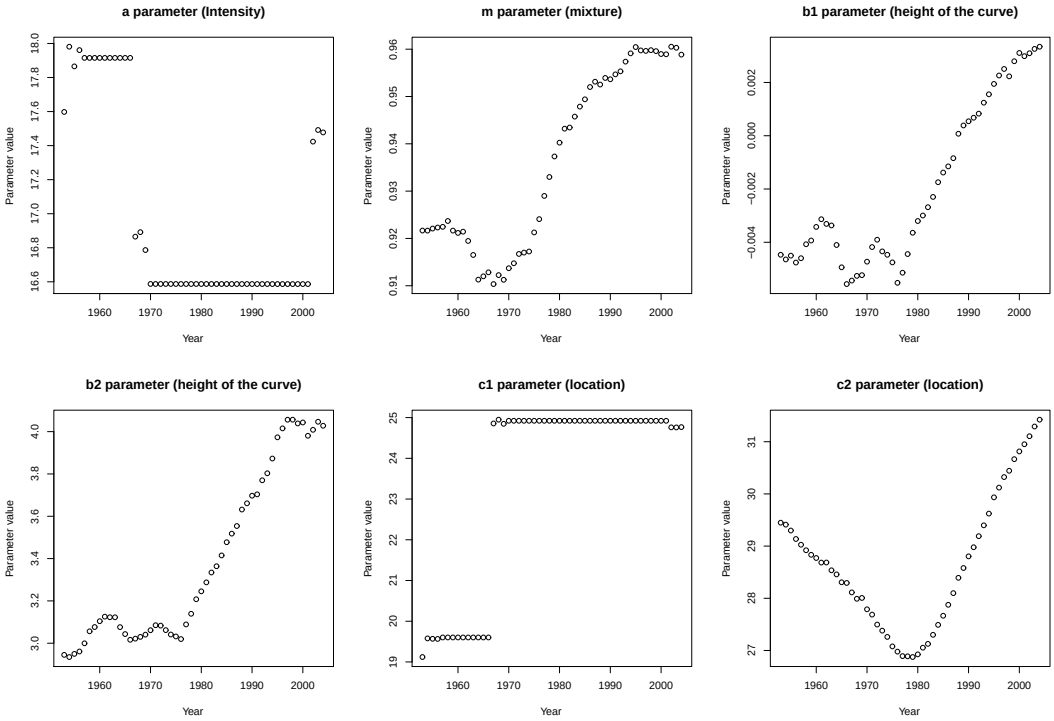

Figure 9: Trends of parameters estimated from the Hadwiger mixture model. Italy (1952-2003)

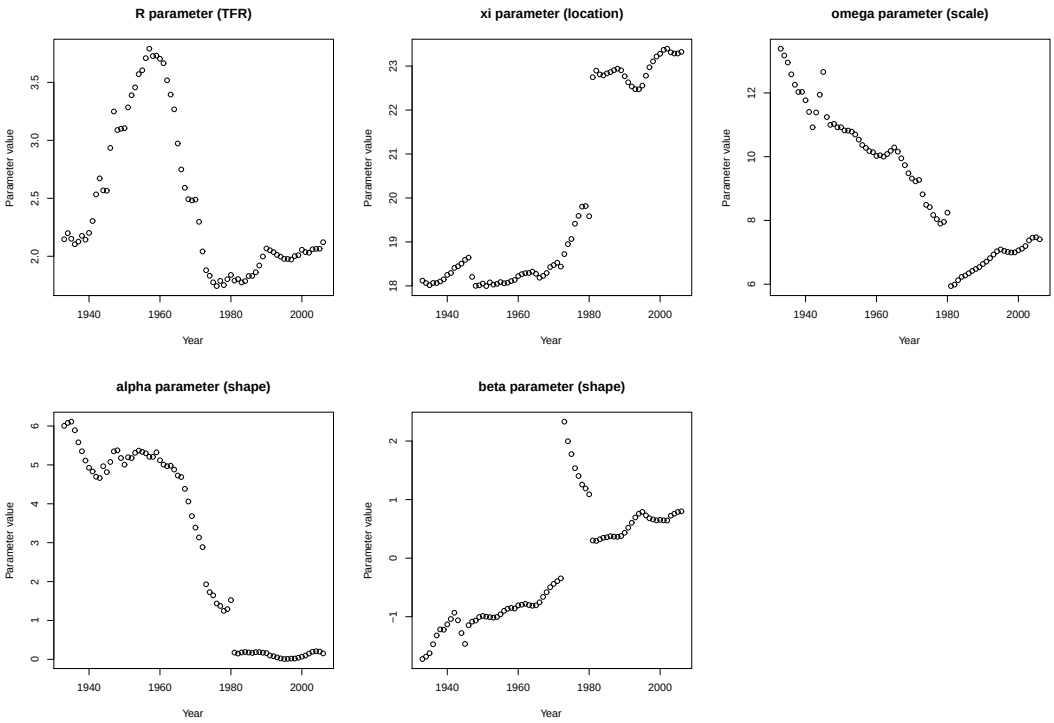

Figure 10: Trends of parameters estimated from the FGSN model. USA (1933-2006)

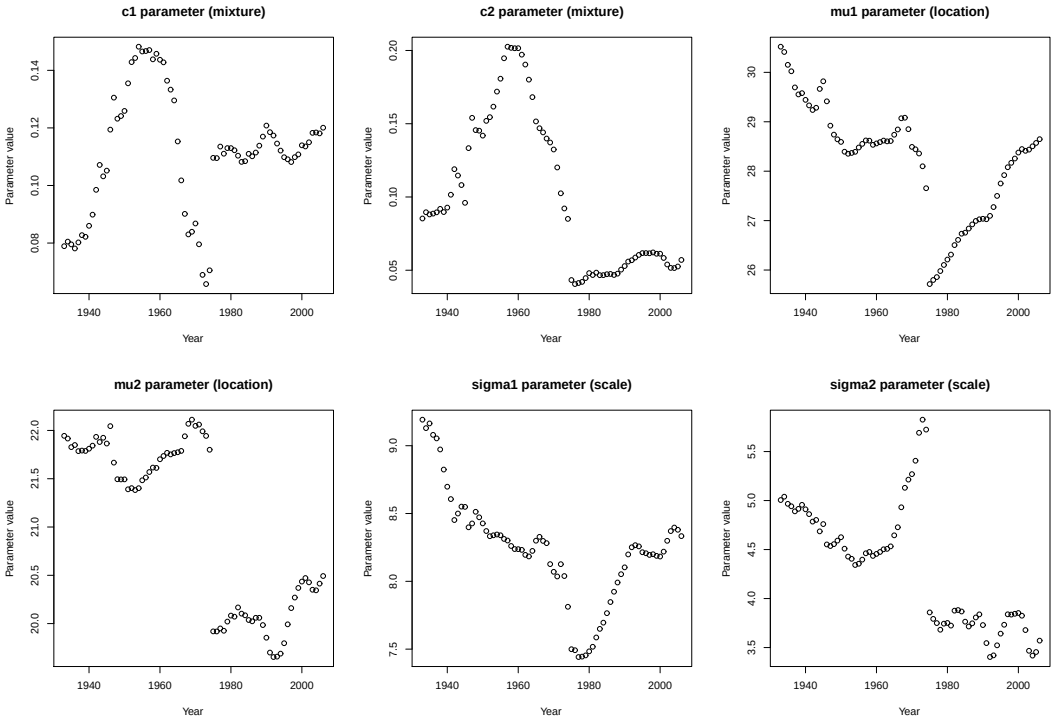

Figure 11: Trends of parameters estimated from the Peristera-Kostaki mixture model. USA (1933-2006)

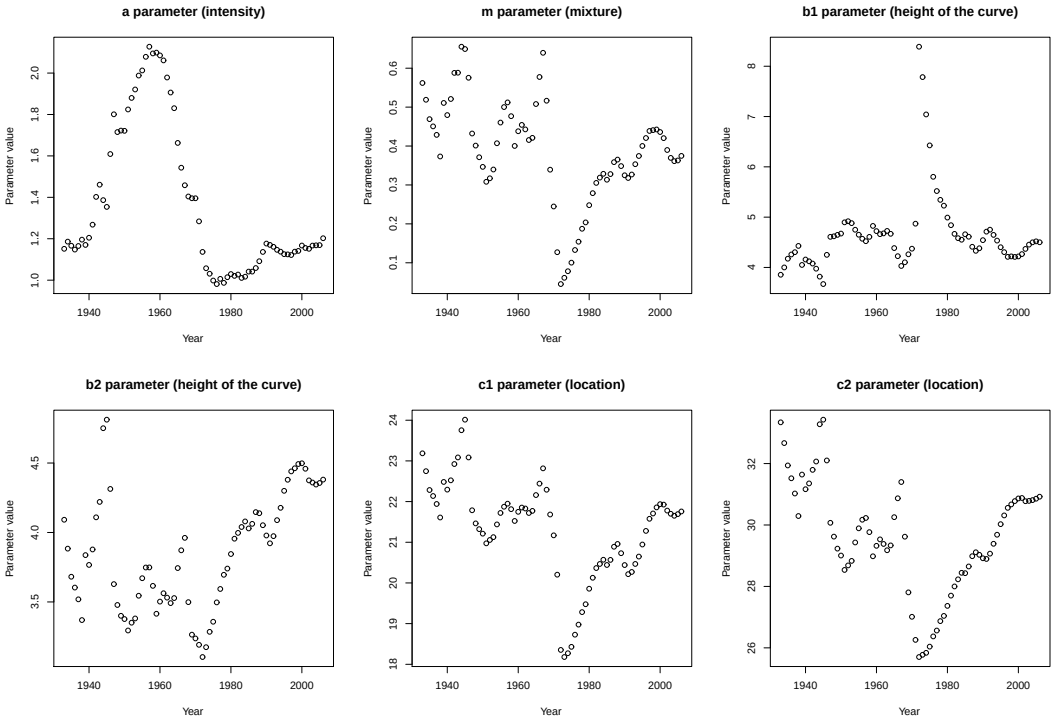

Figure 12: Trends of parameters estimated from the Hadwiger mixture model. USA (1933-2006)