AGE-SPECIFIC PROBABILITY OF CHILDBIRTH

SMOOTHING VIA BAYESIAN NONPARAMETRIC MIXTURE OF ROUNDED SKEW-NORMAL

Antonio Canale and Bruno Scarpa
University of Turin and Collegio Carlo Alberto, University of Padua
antonio.canale@unito.it, scarpa@stat.unipd.it

The municipality of Milan is one of the most important areas in Italy being the center of many economic activities and the destination of strong national and international immigration. In this context, policy makers are interested in understanding socio-demographical and economical differences among the different urban areas. In this paper we concentrate in estimating differences in fertility among the nine areas of Milan. The knowledge of age-specific fertility indicators, indeed, is extremely useful in order to decide where to build a new nursery-school, where to increase obstetrics departments in hospitals, or which kind of services can be offered to families. To estimate the age-specific probabilities of childbirths in the municipality of Milan, we use opendata on the births residents in Milan in 2011. It has recently been observed that the patterns of fertility of developed countries show a deviation from the classic right-skewed shape due to the fact that women tend to have children later. Also, when a large component of immigrants is present, the age-specific fertility rate exhibits an almost bimodal shape, the curve shows a little hump between 20 and 25 years of the woman, presumably due to the presence of subpopulations. To deal with this phenomena and to compare fertility between the nine urban areas of the municipality of Milan, we introduce a novel Bayesian nonparametric location-scale-shape mixture model which can account for skewness and multimodality and we estimate the age-specific probability of childbirth.

INTRODUCTION

Milan is the main industrial, commercial and financial center in Italy by hosting the headquarters of the largest national companies and banks. Its municipality is the second largest municipality in Italy with almost 1.3 millions of residents (ISTAT, 2013) while its urban area is the largest in Italy. It is also a multiethnic city, being the destination of, national and international, immigration with almost the 20% of the total resident population made of foreign-born residents.

In such a large and multicentric city, the different areas may be characterized by different sub-population with different economical status or social behavior. Clearly, in such a context, policy makers are interested in understanding socio-demographical and economical differences among areas, in order to choose correct decisions. For example, Milan is divided in nine areas (*zone di decentramento*) having partial political autonomy, which may require an accurate knowledge of specific population social needs.

In this paper we concentrate on estimating differences in fertility among the nine areas of Milan. The knowledge of age-specific fertility indicators in different areas of large urban centers is extremely useful in order to make informed political decision. For example, it may be useful to decide where to build a new nursery-school, where to increase obstetrics departments in hospitals, or which kind of services can be offered to mothers and families.

We use the opendata made available by the Municipality of Milan, counting all the childbirths in Milan in 2011 divided by areas and mother's age and we estimate the different age-specific probabilities of childbirths.

Even in a big cities, the simple, and quite used for large populations, empirical estimator of fertility curves, based on the counting of childbirths in a given year by area and mother's age is affected by the typical variability related to random noise. Therefore a statistical model is needed to describe fertility in detail and discuss the differences.

A large number of models have been proposed in demographical literature to describe fertility curves of large populations (see for example Mazzuco and Scarpa, 2013, for a recent review); however, much less attention has been given to models for local fertility curves, where we expect a wider variety of patterns than for the country level. In large populations, fertility curves for developed countries are moving from the classical right-skewed shape to a symmetric one due to the fact that women with higher educational level tend to delay childbirth. Also, is some developed countries, fertility curves exhibit an almost bimodal shape, due to a hump appearing around 20 years; this could be related to the presence of different subpopulations. For example US, UK and Ireland fertility curves show this pattern due to higher levels of young women pregnancy in lower classes.

Although for city level we do not have specific studies to describe fertility, in some areas of Milan we may expect to observe the symmetric pattern, particularly, in those areas where most of the resident women have middle-high educational level. On the other side in some other peripheral areas we may observe the bimodal behavior, given the presence of a subpopulations of foreigners, with different average ages at childbirth. Moreover, in other areas of a city as Milan we may also expect different patterns, such as, for example, a left-skewed curve, related to a generalized very long delay in childbirth.

Given this variety of possible patterns, a nonparametric approach seems appropriate to both smooth the random noise affecting the curves, and to account for different patterns. Skewness and multimodality can be modeled via mixture models. It is known, for example, that a mixture of Gaussians kernels can consistently estimate the shape of almost any continuous distribution. As discussed by many authors (Chandola et al., 1999; Ortega Osona and Kohler, 2000; Peristera and Kostaki, 2007; Schmertmann, 2003), mixture models are clearly appropriate when two populations with different age-specific fertility rates are present.

In the following we use a Bayesian nonparametric mixture model to fit age-specific probabilities of childbirths. However, since the open data on childbirth are rounded, in the sense that we only have the mother's age in years, we propose a model which account for this discrete scale in which data are available. In the next section we present the model and, in the following one, we show and discuss some results for the Milan data.

1 The model

Let y be the age of the mother at childbirth and assume that we want to model the probability distribution $p(y)$. In fact, even if age is ideally continuous, it is typically rounded to the lower integer when recorded, and so are the opendata available for Milan. Hence $p(y)$ is a probability mass function defined on the positive integers.

We propose to estimate $p(y)$ with a Bayesian nonparametric approach. Bayesian nonparametrics is a relatively young area of research which has recently received abundant attention in the statistical literature. The considerable degree of flexibility it ensures, if compared to standard parametric alternatives, and the recent development of

new and efficient computational tools, have pushed its concrete use in a number of complex real world problems. Some of the most successful models in Bayesian nonparametrics are Dirichlet process (DP) mixture models. A DP mixture model for probability mass function estimation assumes

$$p(y) = \int K(y;\theta)dP(\theta), \;\; P \sim DP(\alpha, P_0),$$

where $K(\cdot;\theta)$ is a discrete kernel parametrized by a parameter vector θ and P is a random probability mixing measure which has a DP prior (Ferguson, 1973, 1974). The DP is parametrized by α, a scalar precision parameter, and a base measure P_0. The DP can be seen as a probability measure over the space of probability measures. DP mixtures are widely used in continuous density estimation and in particular using Gaussian kernel in place of $K(\cdot;\theta)$ (Lo, 1984; Escobar and West, 1995). This DP mixture of Gaussians is computationally convenient and has nice theoretical properties. Marginalizing out P, from equation above, one can obtain

$$p(y) = \sum_{h=1}^{\infty} \pi_h K(y;\theta_h), \qquad \theta_h \overset{iid}{\sim} P_0, \qquad \pi = \{\pi_h\} \sim \text{Stick}(\alpha)$$

where Stick(α) denotes the stick-breaking process of Sethuraman (1994). This stick-breaking representation shows that mixture models can be useful for estimate data made of different sub-populations.

Although Bayesian nonparametric mixture models for continuous data are well developed, there is a limited literature on related approaches for discrete data. Following Canale and Dunson (2011), we assume that $y = h(y^*)$, where $h(\cdot)$ is a rounding function defined so that $h(y^*) = j$ if $y^* \in (j-1, j]$, for $j = 0, 1, \ldots$. This assumption, introduced in a more general way, for computational and theoretical convenience by Canale and Dunson (2011), matches the data generating process that we are considering. Under this setting the probability mass function p of y is $p = g(f)$, where $g(\cdot)$ is the rounding function having the simple form

$$p(j) = g(f)[j] = \int_{j-1}^{j} f(y^*)dy^* \quad j \in N.$$

A prior over the space of probability mass functions is obtained specifying a prior for the distribution of the latent y^*. As proposed by Canale and Scarpa (2013) we assume

$$\begin{aligned} y = h(y^*), \quad y^* \sim f^*, \\ f^*(y) = \textstyle\sum_{h=1}^{\infty} \pi_h f_{SN}(y^*; \xi_h, \omega_h, \lambda_h) \end{aligned} \tag{1}$$

with $\pi \sim$ Stick(α), $(\xi_h, \omega_h, \lambda_h) \sim P_0$ and f_{SN} being the Azzalini (1985) skew-normal distribution defined as

$$f_{SN}(X;\xi,\omega,\lambda) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\frac{x-\xi}{\omega}\right), \tag{2}$$

where $\phi(x)$ is the density function of a standard normal and $\Phi(\cdot)$ is the distribution function of a standard normal, $\xi \in \Re$, $\omega \in \Re^+$ and $\lambda \in \Re$. The skew normal distribution accounts for different negative and positive asymmetric shapes and includes the Gaussian as a special case (namely when $\lambda = 0$).

The use of the skew-normal kernel has a particular advantage in this context. If data present different sub-populations which are not symmetric, a single skew-normal can be able to fit each of them. Otherwise if data exhibit the presence of a single asymmetric population, a single skew-normal in the mixture model (1) may be sufficient to obtain a satisfactory fit.

To complete the prior specification we assume $\alpha \sim \text{Ga}(1/2, 1/2)$ as Escobar and West (1995) and P_0 to be a diffuse normal-inverse-gamma for the location and scale parameter and an independent diffuse normal centered in zero for the shape parameter. This choice reflects low prior information on the values of the parameters of each mixture component, which is a common practical choice in Bayesian nonparametrics.

This class of location-scale-shape mixture models has been extensively discussed in Canale and Scarpa (2013). To fit the model to real data the authors propose a Gibbs sampling algorithm for efficient simulation from the posterior which we use in the next section. Also, among the theoretical properties of the models, the authors show large support of the prior and strong posterior consistency which basically ensures that the posterior concentrates in a small neighborhood of the true data generating process as the sample size increases.

2 Fertility in Milan municipality

Data on births on the municipality of Milan divided in areas are available on the website `dati.comune.milano.it` for the years 2003-2011 divided by neighborhoods and areas. We consider here the more recent data and the subdivision by areas. The nine areas of Milan include the following neighborhoods: area 1 - historical center; area 2 - central station, Gorla, Turro, Greco, Crescenzago; area 3 - Città Studi, Lambrate, Venezia; area 4 - Vittoria, Forlanini; area 5 - Vigentino, Chiaravalle, Gratosoglio; area 6 - Barona, Lorenteggio; area 7 - Baggio, De Angeli, San Siro; area 8 - Fiera, Gallaratese, San Leonardo, Quarto Oggiaro; area 9 - Garibaldi station, Niguarda. A map with the 9 areas is reported in Figure 1.

To implement the Gibbs sampler of Canale and Scarpa (2013), the first 1,000 iterations were discarded as a burn-in and the next 5,000 samples were used to calculate the posterior mean of the probability mass function for $j = 15, \ldots, 50$. As posterior estimate, we consider the mean probability mass functions in the nine areas, reported in Figure 1 over the maps of Milan. From this figure it is clear that the distributions of the different areas have different shapes. For example, area 1, 3, and 5 are almost symmetric with, in area 3, only mild left skewness. These probability mass functions clearly show a delay in childbirth, with respect to classical curves, but suggest also the presence of a common fertility behaviour inside these areas. Other areas, instead, present a small hump around 20-25 years. In area 4 and 6 this is clearly evident, while in area 8 and 9 this is only partially noticeable. The former areas are likely to have at least two subpopulations, with the smaller consisting in women anticipating the childbirth. Most of the estimated probability mass functions exhibit moderate skewness to the left, sign of a general trend of the majority of women in the area to postpone the age at childbirth, but also indicator of the presence of subgroups that anticipate it.

Our procedure allows for borrowing of information across the age of childbirth. Indeed, from Figure 2, which compares for each area our estimates along with the empirical estimates, it is clear that the mean of the posterior probability mass function is smoother than the empirical estimate, which has an erratic behaviour by chance. However, our procedure is also able to catch the shape of each probability mass function, which, as we already discuss, is quite different one from another.

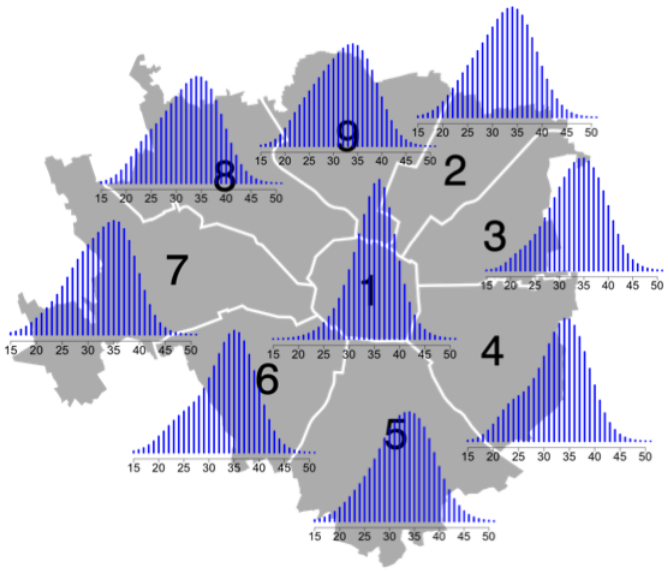

Figure 1: Posterior mean probability mass function for the age of the mother at childbirth in the nine Milan areas.

In Table 1, we report some interesting posterior point estimates along with 95% credible intervals. The posterior predictive mean represent the estimated average in each area. This is generally high, over 31 years, but there are considerable differences between different areas. The higher mean is recorded in the central area 1 and is 34.86 with a posterior credible interval containing 35, while the lower is recorded in area 9, being almost 32 years. The widths of the credible intervals suggest that the differences between areas are often statistically significative. The probability of deliver in young age is very low for every area, and again the "oldest" mothers are in area 1 (pr($y \leq 25$) $= 0.03$ and pr($y \geq 40$) $= 0.21$) and the "youngest" in area 9 (pr($y \leq 25$) $= 0.15$ and pr($y \geq 40$) $= 0.14$).

Because of the likely presence of sub-populations in some areas, an interesting posterior quantity is given by the average number of occupied clusters in the mixtures. This is reported in the last column of Table 1. At a first glance it may surprise to see that the unimodal and symmetric probability of area 1, has more than 3 occupied clusters on average; however, the posterior variability associated to this, is very high, with a posterior credible interval ranging from 1 to 6: with such a large variability, it seems that mixture is used simply to better fit the data without any claim of interpretation. On the other side, for example, area 6 has an average number of occupied clusters equal to 2.36, with a quite narrow posterior credible interval (between 2 and 4); as interpretation this suggests the cohabitation in the area of a small number (two or three) of groups of women with different behaviours in terms of fertility. This is also indicated by the hump on the left (see Figure 1 and 2). To better perceive the presence of the two possible subpopulations, Figure 3 shows the posterior density along with the two most populated clusters.

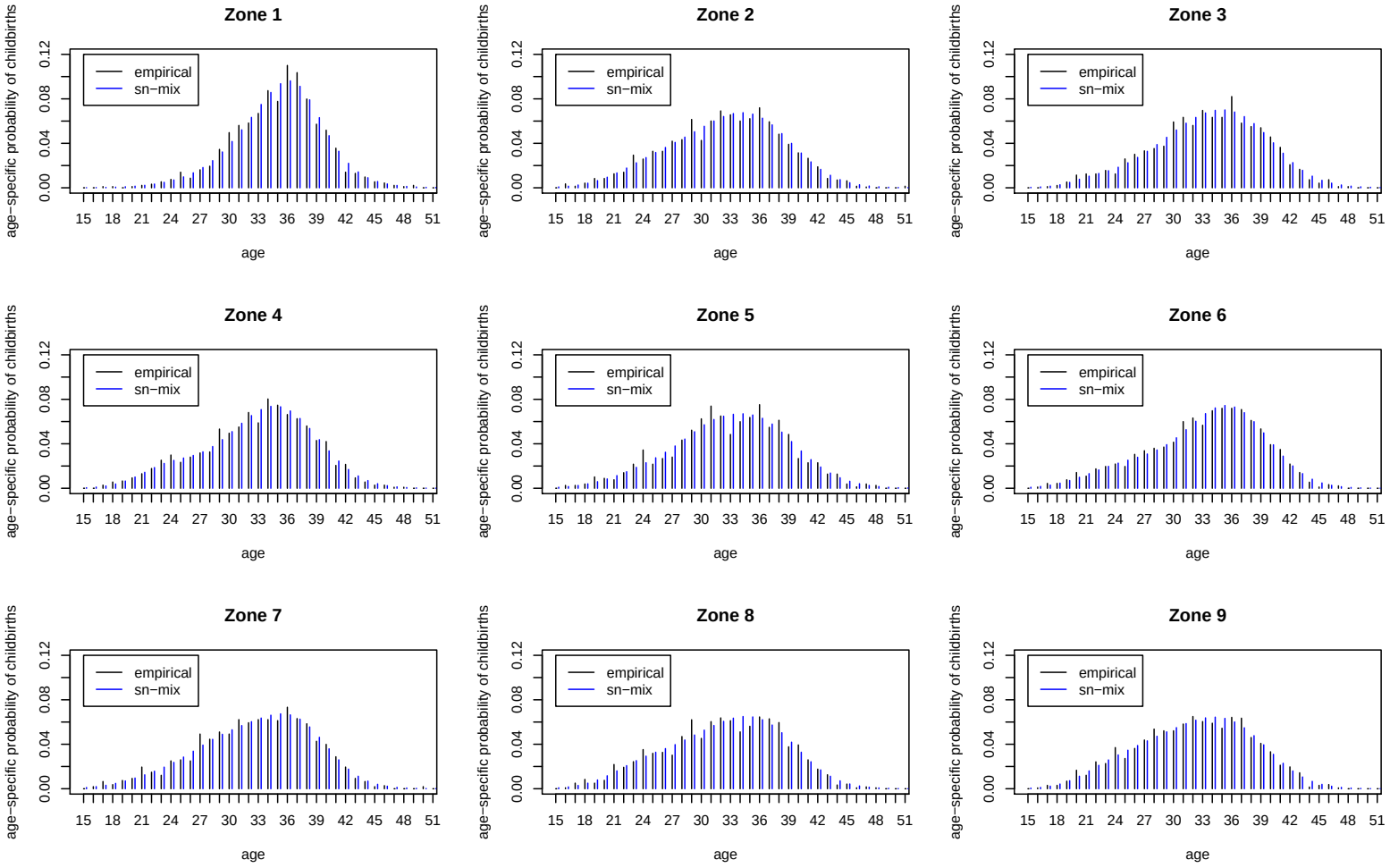

Figure 2: Posterior mean probability mass function and empirical probability mass function for the age of the mother at childbirth in the nine Milan areas.

Table 1: Posterior summaries in the nine different administratives areas of the municipality of Milan. Posterior predictive mean and variance, posterior predictive probability of young and old age at childbirth, averege number of occupied mixture components. Posterior means and, in parenthesis, posterior credible intervals.

	$E(y_{n+1}\|\hat{p})$	$\text{Var}(y_{n+1}\|\hat{p})$	$\text{pr}(y_{n+1} \leq 19\|\hat{p})$	$\text{pr}(y_{n+1} \leq 25\|\hat{p})$	$\text{pr}(y_{n+1} \geq 40\|\hat{p})$	# of clusters
Area 1	34.86 (34.27, 35.43)	4.69 (4.32, 5.15)	0.0025 (0.0005, 0.0064)	0.03 (0.02, 0.05)	0.21 (0.16, 0.25)	3.35 (1, 6)
Area 2	32.27 (31.72, 32.8)	5.87 (5.55, 6.20)	0.0170 (0.0097, 0.0263)	0.13 (0.11, 0.16)	0.14 (0.12, 0.17)	2.46 (2, 5)
Area 3	33.36 (32.71, 34.01)	5.80 (5.40, 6.23)	0.0114 (0.0051, 0.0200)	0.10 (0.07, 0.12)	0.19 (0.15, 0.23)	2.64 (2, 5)
Area 4	32.64 (32.12, 33.13)	5.77 (5.36, 6.21)	0.0178 (0.0088, 0.0301)	0.12 (0.09, 0.15)	0.15 (0.12, 0.17)	1.77 (1, 4)
Area 5	32.64 (31.69, 33.64)	6.00 (5.53, 6.47)	0.0176 (0.0078, 0.0303)	0.12 (0.08, 0.16)	0.16 (0.12, 0.21)	2.51 (2, 5)
Area 6	33.00 (32.44, 33.56)	5.92 (5.47, 6.39)	0.0185 (0.0091, 0.0308)	0.12 (0.09, 0.15)	0.17 (0.14, 0.20)	2.36 (2, 4)
Area 7	32.54 (31.92, 33.16)	5.91 (5.54, 6.30)	0.0188 (0.0105, 0.0296)	0.13 (0.10, 0.16)	0.15 (0.12, 0.19)	2.41 (2, 4)
Area 8	32.14 (31.62, 32.66)	6.03 (5.63, 6.43)	0.0202 (0.0112, 0.0313)	0.15 (0.12, 0.18)	0.15 (0.12, 0.17)	2.19 (2, 3)
Area 9	31.98 (31.50, 32.48)	5.92 (5.59, 6.26)	0.0173 (0.0098, 0.0267)	0.15 (0.12, 0.18)	0.14 (0.11, 0.16)	2.21 (2, 3)

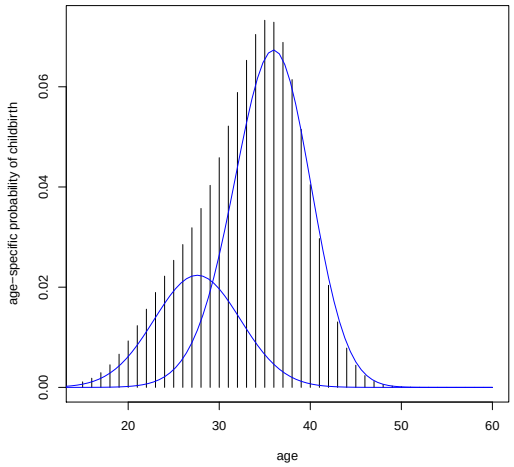

Figure 3: Posterior mean probability mass function of area 6 with the two principal mixture components.

To conclude, we want to notice that a simpler parametric model could be used to estimate the probability of age at childbirth in each single area, for the symmetric shapes the choice of the parametric model being quite easy, and for the skew and almost bimodal one being more tricky. However, the choice of a nonparametric flexible approach is preferable in absence of prior information on the shape of the distributions and is very flexible to catch all the shapes of the different areas.

3 Discussion

Open data are a formidable way to empower citizens, help small businesses, create value in positive and constructive ways. The Municipality of Milan initiative to diffuse administrative data and start collaborative projects is certainly the beginning of a fascinating and challenging path to improve education and to help government and policy makers to better exploit the available information.

We presented a statistical model to describe fertility curves by areas. The proposed model describes the probability of childbirth and deals with the different shapes observed in the nine areas of the city of Milan. The nonparametric characteristic of our model allows for smoothing the random variability due to the small size of the data for some age and area, but its specific formulation, based on almost certainly finite mixture of skew shaped kernels, enables for a clear interpretation of the results.

Interesting results describing the variability of fertility between the nine Milan areas are sketched and discussed, leading to hints for further investigation of demographical and socio-economical interpretations.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12:171–178.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539.

Canale, A. and Scarpa, B. (2013). Bayesian nonparametric location-scale-shape mixtures. Technical report, under preparation.

Chandola, T., Coleman, D., and Hiorns, R. W. (1999). Recent European fertility patterns: Fitting curves to 'distorted' distributions. *Population Studies*, 53:317–329.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–629.

ISTAT (2013). Popolazione residente al maggio 2013. `www.demo.istat.it`.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12:351–357.

Mazzuco, S. and Scarpa, B. (2013). Fitting age-specific fertility rates by a flexible generalized skew-normal probability density function. *Journal of the Royal Statistical Society: Series A*.

Ortega Osona, J. A. and Kohler, H.-P. (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 Kostaki, A. (2007). Modelling fertility in modern populations. *Demographic Research*, 16:141–194.

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.