Modelling fertility rates by age, time, and birth order from coarsely grouped data: a penalized composite link model approach

Diego Ayma*[1] and Carlo G. Camarda†[2]

[1]Department of Statistics, Universidad Carlos III de Madrid, Leganés, Spain
[2]INED - Institut National d'Études Démographiques, Paris, France

14th June 2016

Abstract

In practice, fertility and population data can be only available in age classes and calendar year classes, which may have different levels of aggregation. Moreover, these coarse data can be recorded along a third dimension, making their analysis more challenging. Data aggregation process may, however, hinder the visualization of the underlying distribution that follows the data. In this paper, we propose the use of the penalized composite link model to estimate the latent trend behind these coarsely grouped data. This model only assumes that the underlying distribution is smooth, making it suitable for other applications in which disaggregation is required. We illustrate our proposal using a Canadian fertility dataset, which is recorded by age of the mother, calendar year, and birth order.

Keywords. Penalized composite link models; Grouped data; Vital rates.

1 Introduction

Coarsely grouped data often appear in areas such as demography, epidemiology, and public health. Some examples of grouped data are age group-specific disease incidence rates, abridged life tables, and bivariate histograms with wide bins. A more complex situation occurs when data are registered by age groups, calendar year groups, and another covariate that can be aggregated or not. In this case, data structure can be viewed as a three-dimensional (coarse) array.

In general, data aggregation is done to protect the privacy of patients, to facilitate compact presentation, or to make it comparable with other datasets. Another reasons can be derived from the lack of good vital registration systems or regular population censuses in some countries. This data aggregation process may, however, hinder the visualization of the underlying distribution behind these data. In this paper, we propose the use of the penalized composite link model (PCLM, Eilers, 2007) to estimate such latent distribution, from coarsely grouped data, at a finer resolution. We assume here the underlying distribution at the fine resolution is smooth. The flexibility of the model is given by the use of B-spline bases, together with a penalty on the regression coefficients, following the P-spline methodology (Eilers and Marx, 1996). We address here the case when grouped data can be arranged as a three-dimensional array (for applications in the one- and two-dimensional cases, see Lambert and Eilers, 2009; Rizzi et al., 2015; Lambert, 2011; Ayma et al., 2015). We illustrate our approach using a Canadian fertility dataset, registered by ages, years, and birth order.

2 The 3-d penalized composite link model

2.1 The model

Let $\boldsymbol{y} = \text{vec}(\mathbf{Y})$ be the vector of observed aggregated counts, where $\mathbf{Y}$ denotes a three-dimensional array of dimension $n_1 \times n_2 \times n_3$. Assume that $\boldsymbol{y}$ follows a Poisson distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^{n_1 n_2 n_3}$. These counts can be seen as indirect observations of a latent process that we want to estimate. The PCLM approach

*dayma@est-econ.uc3m.es
†carlo-giovanni.camarda@ined.fr

of Eilers (2007) offers an elegant way to do this, by considering $\boldsymbol{\mu}$ as composed of latent expectations. Denoting $\boldsymbol{x}_d \in \mathbb{R}^{m_d}$, for $d = 1, 2, 3$, as the covariates that determine the dimension of the latent process, the Poisson PCLM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}(\boldsymbol{e}_{\mathrm{f}} * \exp(\mathbf{B}\boldsymbol{\theta})), \tag{1}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{m_1 m_2 m_3}$ represents the mean vector of the latent process at fine resolution, $\mathbf{C}$ is the composition matrix that describes how these latent expectations are combined to yield $\boldsymbol{\mu}$, $\mathbf{B}$ is the full regression matrix constructed from the covariates $\boldsymbol{x}_d$, $\boldsymbol{\theta} \in \mathbb{R}^{c_1 c_2 c_3}$ is a vector of regression coefficients, and $\boldsymbol{e}_{\mathrm{f}}$ is the vector of exposures at the fine resolution (which allows to analyse rates instead of counts). Under this framework, the matrices $\mathbf{C}$ and $\mathbf{B}$ in Eq. (1) can be expressed as $\mathbf{C} = \mathbf{C}_1 \otimes \mathbf{C}_2 \otimes \mathbf{C}_3$ and $\mathbf{B} = \mathbf{B}_1 \otimes \mathbf{B}_2 \otimes \mathbf{B}_3$, respectively, where $\mathbf{C}_d$ is the marginal composition matrix (of dimension $n_d \times m_d$) and $\mathbf{B}_d = \mathbf{B}(\boldsymbol{x}_d)$ is the marginal B-spline basis (of dimension $m_d \times c_d$), for $d = 1, 2, 3$. Smoothness is achieved by imposing a penalty on the regression coefficients in the form $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$, where $\mathbf{P}$ is the penalty matrix of dimension $c_1 c_2 c_3 \times c_1 c_2 c_3$. Here we choose an anisotropic penalization, i.e., a different amount of smoothing for each $\boldsymbol{x}_d$ ($d = 1, 2, 3$). Thus the penalty matrix is given by:

$$\mathbf{P} = \lambda_1 \mathbf{P}_1 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_3} + \lambda_2 \mathbf{I}_{c_1} \otimes \mathbf{P}_2 \otimes \mathbf{I}_{c_3} + \lambda_3 \mathbf{I}_{c_1} \otimes \mathbf{I}_{c_2} \otimes \mathbf{P}_3,$$

where $\mathbf{I}_k$ denotes an identity matrix of dimension $k \times k$, λ_d is a smoothing parameter that controls the amount of smoothing along the covariate $\boldsymbol{x}_d$, and $\mathbf{P}_d = \mathbf{D}_d'\mathbf{D}_d$ is the marginal penalty matrix based on the matrix $\mathbf{D}_d$ that computes q_d-th differences, i.e, $\Delta^{q_d}\boldsymbol{\theta} = \mathbf{D}_d\boldsymbol{\theta}$ ($d = 1, 2, 3$).

2.2 Parameter estimation

For fixed values of λ_d, $d = 1, 2, 3$, the estimation the regression coefficients $\boldsymbol{\theta}$ in Eq. (1) are obtained by iteratively solving:

$$(\breve{\mathbf{B}}'\tilde{\mathbf{W}}\breve{\mathbf{B}} + \mathbf{P})\hat{\boldsymbol{\theta}} = \breve{\mathbf{B}}'(\boldsymbol{y} - \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{W}}\breve{\mathbf{B}}\tilde{\boldsymbol{\theta}}), \tag{2}$$

where $\tilde{\mathbf{W}} = \mathrm{diag}(\tilde{\boldsymbol{\mu}})$ and $\breve{\mathbf{B}} = \tilde{\mathbf{W}}^{-1}\mathbf{C}\tilde{\boldsymbol{\Gamma}}\mathbf{B}$, with $\tilde{\boldsymbol{\Gamma}} = \mathrm{diag}(\tilde{\boldsymbol{\gamma}})$. Here a tilde, as in $\tilde{\boldsymbol{\theta}}$, indicates the current approximation to the solution, and $\hat{\boldsymbol{\theta}}$ denotes the updated estimate of $\boldsymbol{\theta}$. The expression in Eq. (2) corresponds to a modified version of the iteratively reweighted least squares algorithm, as it was shown previously by Eilers (2007).

The PCLM fit requires an optimal selection of the smoothing parameters λ_d (for $d = 1, 2, 3$). We choose to minimize the Akaike Information Criterion (AIC), which combines the deviance and effective dimension of the model as follows:

$$\mathrm{AIC} = \mathrm{Dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + 2\mathrm{ED} = 2\left(\boldsymbol{y}'\left(\log(\boldsymbol{y}) - \log(\hat{\boldsymbol{\mu}})\right) - \mathbf{1}'_{n_1 n_2 n_3}\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right)\right) + 2\mathrm{ED},$$

where $\mathrm{ED} = \mathrm{trace}((\breve{\mathbf{B}}'\hat{\mathbf{W}}\breve{\mathbf{B}} + \mathbf{P})^{-1}\breve{\mathbf{B}}\hat{\mathbf{W}}\breve{\mathbf{B}})$.

A simple algorithm for the search of optimal smoothing parameters is to consider a three-dimensional array of $\log(\lambda_d)$ values ($d = 1, 2, 3$) and choose the one that minimizes AIC. Since grouped data have an array structure, the parameter estimation procedure can be speed up by considering the so-called GLAM methods (Currie et al., 2006) in the computations.

3 An illustration

To illustrate our proposal, we use a Canadian fertility dataset (downloaded from http://www.humanfertility.org/) that consists of birth counts and age-specific fertility rates recorded by age (which varies from 15 to 54 years), calendar year (from 1944 to 2009), and birth order (from 1st to 4th). The female population exposure can be then derived using these data. To show how our methodology works, we have grouped the birth counts and the exposures into 5-year age classes, and calendar year classes of different lengths (starting with four classes of length 10, followed by five classes of length 5). Figures 1(a), 1(c), 1(e), and 1(g) show the fertility rates resulting from this aggregation for the 1st, 2nd, 3rd, and 4th birth order, respectively.

Now we apply the 3-d PCLM approach to the previous grouped counts, in order to obtain detailed trends at a fine resolution. To do that, let us consider the female population exposure at the fine resolution as $\boldsymbol{e}_{\mathrm{f}}$ in Eq. (1) (if the exposure vector is only available at an aggregated level, we can apply the 3-d PCLM approach to disaggregate them, and then use the resulting estimates as $\boldsymbol{e}_{\mathrm{f}}$). To set up the 3-d PCLM formulation, we use $\boldsymbol{x}_1 = (15, ..., 54)'$ (for ages), $\boldsymbol{x}_2 = (1945, ..., 2009)'$ (for years), and $\boldsymbol{x}_3 = (1, ..., 4)'$ (for birth orders) as covariates at the fine resolution. We choose 10, 16, and 3 internal equally-spaced knots for the marginal cubic B-spline bases $\mathbf{B}_d$, respectively, and second order penalties, i.e., $q_d = 2$, for $d = 1, 2, 3$. About the composition matrix $\mathbf{C}$ in Eq. (1), since we have not aggregated the data by birth order, the marginal composition matrix $\mathbf{C}_3$ is equal to $\mathbf{I}_4$. The dimensions of the other marginal composition matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ are 8×40 and 9×65, respectively.

Figures 1(b), 1(d), 1(f), and 1(h) show the resulting PCLM estimates for the fertility rates along ages and years, for the 1st, 2nd, 3rd, and 4th birth order, respectively. We observe more detailed impressions of the Canadian fertility, delineating clearly lower and higher fertility rates. Figure 2(a) shows the estimated fertility rates using the 3-d PCLM approach at age 36, for 1st, 2nd, 3rd, and 4th birth orders. We observe the patterns that the solid curves exhibit follow closely the true distribution of the fertility rates (dot points), except perhaps in the left extremes of some curves where we have less information (recall the aggregation procedure previously done). Figure 2(b) shows the estimated fertility rates using the 3-d PCLM approach for the year 1990. In this case, we have calculated the true and the estimated total fertility rates, obtaining similar results. Note here that, in the case of the 1st birth order, the red curve departs slightly from the true distribution from ages 15 to 27. This can be improved if we incorporate more information in the left part of the curve, i.e., fertility information below 15 years old.

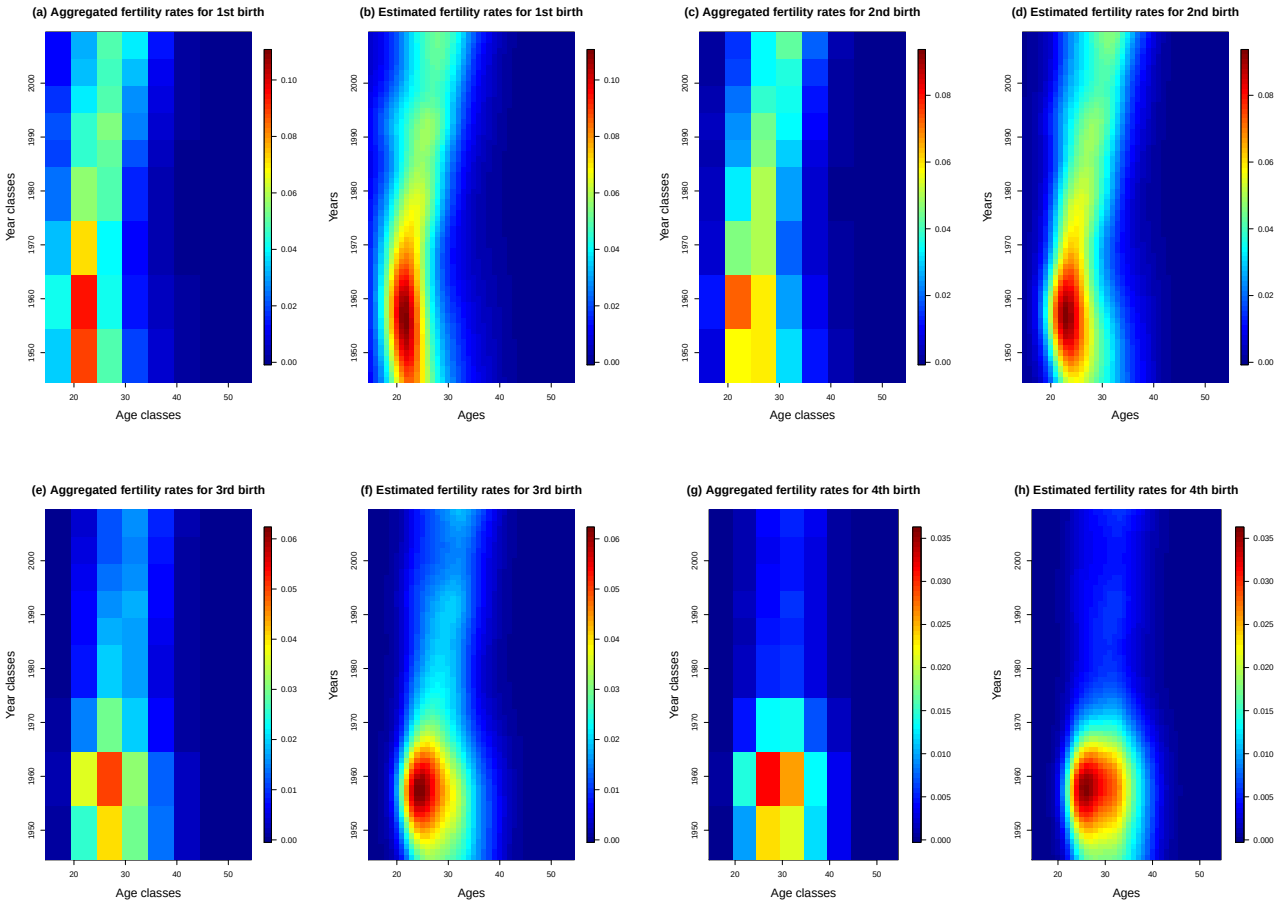

Figure 1: Grouped Canadian fertility rates for 1st, 2nd, 3rd, and 4th birth orders (left side) and estimated fertility rates using the 3-d PCLM approach (right side).

4 Concluding remarks

In this paper, the penalized composite link model for 3-d array (coarse) data was developed and applied to the disaggregation of fertility rates. The PCLM provides a flexible tool for epidemiological studies, when the aim is to obtain estimations of vital rates at a fine resolution. We used the statistical software R (R Core Team, 2015) for data analysis with the 3-d PCLM approach. Our plan is to make accessible efficient R codes for the use of our proposal.

The estimation procedure presented in Section 2.2 is not so computational efficient, in the sense that we have to search estimates for the smoothing parameters on a fine array of values. A way to improve the computational

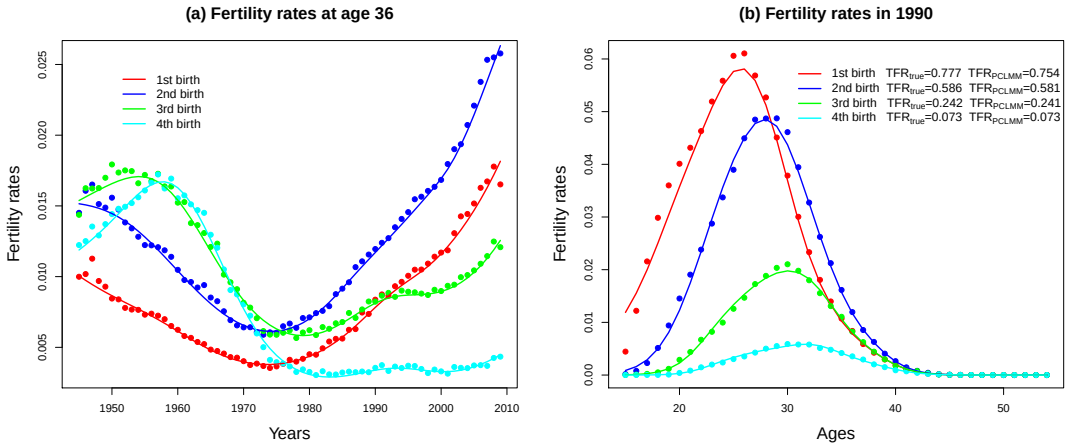

Figure 2: True (dot points) and estimated (solid lines) fertility rates (a) at age 36; and (b) for year 1990.

speed is to reformulate the model in Eq. (1) as a mixed model (in fact, as a generalized linear mixed model). In this new context, the smoothing parameters are seen as ratios of variance components, i.e., $\lambda_d = \phi/\tau_d^2$ $(d = 1, 2, 3)$ with $\phi = 1$ (in the Poisson case). Thus, we can use an adaptation of the SAP (separation of anisotropic penalties, Rodríguez-Álvarez et al., 2015) algorithm in the PCLM framework to estimate the parameters of the model.

References

Ayma, D., Lee, D.-J., Durbán, M., and Eilers, P. H. C. (2015). Penalized composite link mixed models for two-dimensional count data. Technical report, Universidad Carlos III de Madrid.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):259–280.

Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3):239–254.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, 11(2):89–121.

Lambert, P. (2011). Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics & Data Analysis*, 55:429–445.

Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53:1388–1399.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizzi, S., Gampe, J., and Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, 182(2):138–147.

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. H. C. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, 25(5):941–957.