

Nested Functional Clustering Using Basis Representations

Abel Rodriguez¹

Department of Applied Mathematics and Statistics, University of California, Santa Cruz.
e-mail: abel@soe.ucsc.edu

Abstract: This paper discusses clustering procedures for nested samples of curves, where multiple profiles are collected for each subject in the study. We present a model based on the nested Dirichlet processes that uses the information on the distribution of curves to generate the clusters. The method is illustrated using data from the Early Pregnancy Study on hormone profiles along multiple menstrual periods for a cohort of women. The resulting model simultaneously clusters both curves and subjects, allowing us to identify outlier curves for each group of women, as well as outlier women whose distribution of profiles differs from the rest.

Keywords: Nonparametric Bayes; Nested Dirichlet process; Basis Representations; Splines; Functional clustering; Early Pregnancy Study; Hormone profile

1. Introduction

Functional clustering arises in a variety of applications, ranging from genetics and epidemiology to econometrics and oceanography. Most functional clustering methods have been designed for situations where a single curve is observed for each subject; often, they assume that the functions can be represented using a basis expansion, such as splines and wavelets, and then apply a standard clustering procedure to the estimates of the coefficients. Extensions to nested samples typically make the additional assumption that coefficients describing subject-specific curves arise from a parametric distribution, and clustering is applied to the parameters of this subject-specific distribution. The result is typically a procedure that generates clusters of subjects based on their average response curve, which is clearly inappropriate if outliers are expected.

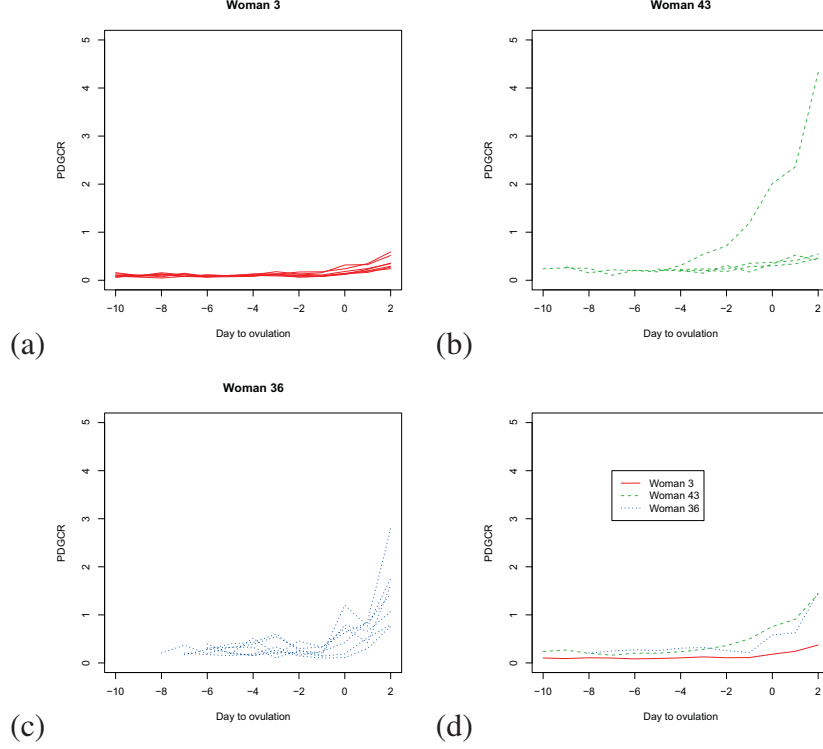
In this paper we discuss hierarchical Bayes models for nested functional data. Our motivation comes from the Early Pregnancy Study (EPS) (Wilcox *et al.*, 1998) conducted by the U.S. National Institutes of Health, where progesterone levels were collected for both conceptive and nonconceptive women along multiple menstrual cycles. Our models use basis expansions along with nonparametric mixture priors to create sparse but flexible representations of the hormone profiles. The resulting model clusters women according to the distribution of their hormonal profiles, rather than based on the average profiles.

In order to motivate the need for distribution based clustering, consider the hormonal profiles from the EPS depicted in Figure 1. Frames (a) to (c) show the hormone profiles for three women in our data set, while frame (d) shows the mean profile corresponding to each one of them, obtained by simply averaging all available observations at a given day within the cycle. When looking at the mean profiles in (d), women 43 and 36 seem to have very similar hormonal responses, which are different from those of woman 3. However, when the individual profiles are considered, it is clear that most of the cycles of woman

¹Address of correspondence: 1156 High Street, Mailstop SOE2, Santa Cruz, CA, 95062.

43 look like those of woman 3 and that the big difference in the means is driven by the single abnormal cycle.

Figure 1: Comparison of hormone profiles for 3 women in the EPS. Frames (a) to (c) show curves for each woman, and frame the (d) shows the average curves.



2. A model for nested functional clustering

Let $\mathbf{y} = (y_{ij1}, \dots, y_{ijT_{ij}})$ be the sequence of progesterone levels for the j -th menstrual cycle of i -th subject collected at times $\mathbf{x} = (x_{ij1}, \dots, x_{ijT_{ij}})$, with $i = 1, \dots, I$ and $j = 1, \dots, n_i$. We model the expected evolution of progesterone in time as a linear combination of basis functions and use a hierarchical specification on the corresponding coefficients to achieve the desired clustering structure. Specifically, we let

$$\mathbf{y}_{ij} = \mathbf{B}(\mathbf{x}_{ij})\mathbf{\Lambda}_{ij}\boldsymbol{\theta}_{ij} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_{ij}^2 \mathbf{I})$$

where $\boldsymbol{\theta}_{ij}$ is a p dimensional parameter of basis coefficients and $\mathbf{\Lambda}_{ij}$ is a 0-1 diagonal matrix that serves to select the subset of basis functions to be used in representing the curve. Next, we assume that the vector of cycle-specific parameters $(\boldsymbol{\theta}_{ij}, \mathbf{\Lambda}_{ij}, \sigma_{ij}^2)$ follows some unknown woman-specific distribution, G_i . It is standard practice in the literature to assume that G_i corresponds to a parametric distribution with subject-specific parameters. Instead, a nonparametric alternative is to assume that the curves are generated by a nested generalized Dirichlet process, (Rodriguez *et al.*, 2008; Hjort, 2000)

$$G_i \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*} \quad G_k^* = \sum_{l=1}^{\infty} w_{lk} \delta_{(\boldsymbol{\theta}_{lk}^*, \mathbf{\Lambda}_{lk}^*, \sigma_{lk}^{2*})} \quad (\boldsymbol{\theta}_{lk}^*, \mathbf{\Lambda}_{lk}^*, \sigma_{lk}^{2*}) \sim G_0$$

where the stick-breaking weights are constructed as,

$$\pi_k = u_k \prod_{s < k} (1 - u_s) \quad w_{lk} = v_{lk} \prod_{s < l} (1 - v_{sk})$$

with $u_k \sim \text{Beta}(a_1, b_1)$ and $v_{lk} \sim \text{Beta}(a_2, b_2)$. That is, we generate groups of women, assigning to each of those groups a common random effect distribution G_k^* . Each one of these distributions is a mixture of different curve shapes given by the atoms $\{\theta_{lk}^*\}_{l=1}^\infty$, $\{\Lambda_{lk}^*\}_{l=1}^\infty$ and $\{\sigma_{lk}^{2*}\}_{l=1}^\infty$. Finally, all curve shapes are drawn from a common prior distribution G_0 , possibly with unknown hyperparameters that allow us to further borrow information across cycles and subjects. Indeed, we set the baseline measure G_0 to

$$G_0 = G_0^{\theta|\Lambda, \sigma} \times G_0^\Lambda \times G_0^\sigma = \mathcal{N}(\mathbf{0}, \sigma^2 \Lambda \Sigma \Lambda') \times \prod_{s=1}^p \text{Ber}(\gamma) \times \text{IG}(\nu_\sigma, \tau_\sigma)$$

with $\Sigma \sim \text{IW}(\nu_\Sigma, \Sigma_0)$, $\tau_\sigma \sim \mathcal{G}(\rho, \psi)$ and $\gamma \sim \text{Beta}(\eta_1, \eta_2)$, providing a conditionally conjugate specification amenable for simple computational implementation. For the precision parameters in the generalized Dirichlet process we set $a_1 \sim \mathcal{G}(\kappa_{a_1}, \tau_{a_1})$, $b_1 \sim \mathcal{G}(\kappa_{b_1}, \tau_{b_1})$, $a_2 \sim \mathcal{G}(\kappa_{a_2}, \tau_{a_2})$ and $b_2 \sim \mathcal{G}(\kappa_{b_2}, \tau_{b_2})$. This specification extends the model of Ray and Mallick (2006) to general basis systems and settings where multiple curves are collected for each subject and clustering among subjects is to be carried out according to the distribution of its curves. Inferences can be performed through an extension of the collapsed Gibbs sampler developed in Rodriguez *et al.* (2008).

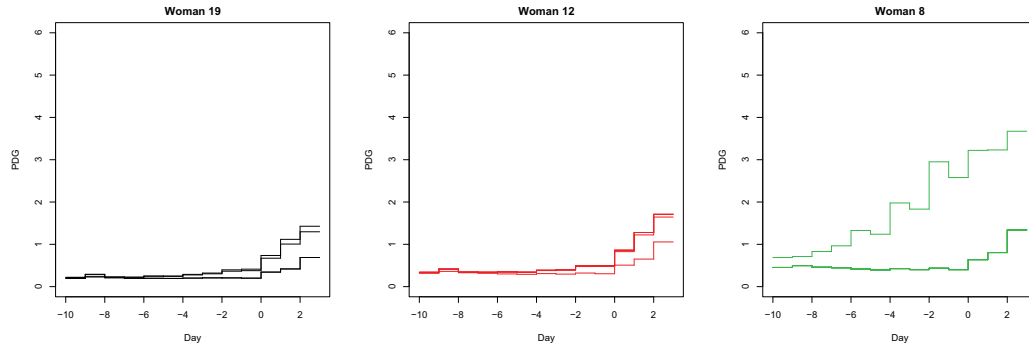
3. An application to the Early Pregnancy Study

Our data, extracted from the EPS, consists of daily creatinine-corrected concentrations of pregnanediol-3-glucuronide (PdG) for 60 women along multiple menstrual cycles, measured in micrograms per milligram of creatinine. We focus on a thirteen-day period extending from 10 days before ovulation to 2 days after ovulation. We considered only non-conceptive cycles and women with at least four cycles in record. Curves are modeled using piecewise constant splines with $q = 13$ nodes, corresponding to each of the days considered in the study, and set $\Lambda_{lk}^* = \mathbf{I}$ for all l and k (i.e. all 13 basis are used).

The posterior distribution on the partitions of subjects was summarized using the average incidence matrix (in which entry (i, j) corresponds to the marginal posterior distribution that women i and j being assigned to the same cluster). A point estimate of the partition can be obtained then using the algorithm described in Lau and Green (2007), while uncertainty in the clustering can be assessed through the visual inspection of the incidence matrix. There is evidence of nine groups of varying size. Clusters 1, 3 and 4 are the largest and best separated clusters, comprising 45 out of 60 women. Clusters 0 and 2 are not clearly separated from the other seven groups. Clusters 5 to 8 are small clusters of one or two women, and correspond to patients with consistently anomalous hormonal profiles. As was suggested in the introduction, women 3 and 43 are clustered together.

Figure 2 shows reconstructed profiles for some representative women on each of the groups. Most profiles are relatively flat before ovulation, when hormone levels start to increase. Also, profiles tend to be relatively consistent for any single woman. However, we can see some outliers, typically corresponding to elevated post-ovulation levels and/or early increases in the hormone levels.

Figure 2: Reconstructed profiles for some representative women. Patient 19 was chosen for cluster 1, patient 12 for cluster 2, patient 8 for cluster 3.



The reconstructed profiles provide insight into the characteristics of the groups. For example, cluster 1 corresponds to women with very low hormonal levels, even after ovulation. These group presents few outliers, and those present are characterized by a slightly larger increase after ovulation. On the other hand Group 3 has a slightly higher baseline level and also present a larger increase in PdG in the luteal phase, reaching a concentration $1.35 \mu\text{g/ml Cr}$. Outlier cycles in this groups are characterized by earlier increases in PdG during the follicular phase, sometimes as early as 5 days before ovulation. Group 2 is a transition cluster, with curves that resemble those in both cluster 1 and 3.

4. Discussion

This paper briefly describes a flexible model for functional clustering in nested datasets. Our current work focuses on extending the analysis of the EPS dataset to more general basis systems and implementing the basis selection algorithm. We also plan to explore the theoretical properties of the prior clustering structure induced by the generalized NDP. Finally, we are interested in exploring strategies for experiment design for epidemiological studies whose outcome can be modeled through our functional clustering approach.

References

- Hjort N. (2000) Bayesian analysis for a generalized Dirichlet process prior, Technical report, University of Oslo.
- Lau J.W. and Green P. (2007) Bayesian model based clustering procedures, *Journal of Computational and Graphical Statistics*, **16**, 526–558.
- Ray S. and Mallick B.K. (2006) Functional clustering by Bayesian wavelet methods, *Journal of the Royal Statistical Society, Series B.*, **68**, 305–332.
- Rodriguez A., Dunson D.B. and Gelfand A.E. (2008) The nested Dirichlet process, with discussion, *Journal of American Statistical Association*, **103**, 1131–1144.
- Wilcox A.J., Weinberg C., O'Connor J.F., Baid D.D., Schlatterer J., E. C.R., Armstrong E.G. and Nisula B.C. (1998) Incidence of early loss of pregnancy, *New England Journal of Medicine*, **319**, 189–194.