Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 00.0000/00000000

Analysis of AneuRisk65 data: warped logistic discrimination*

Daniel Gervini

Department of Mathematical Sciences
University of Wisconsin–Milwaukee
e-mail: gervini@uwm.edu

Abstract: We analyze the AneuRisk65 curvature functions using a likelihood-based warping method for sparsely sampled curves, and combine it with logistic regression in order to discriminate subjects with aneurysms at or after the terminal bifurcation of the internal carotid artery (the most life-threatening) from subjects with no aneurysms or aneurysms along the carotid artery (the less serious). Significantly lower misclassification rates are obtained when the warping functions are included in the logistic discrimination model, rather than being treated as mere nuisance parameters.

Keywords and phrases: Karhunen–Loève decomposition, Missing data, Monotone Hermite splines, Random-effect models.

1. Introduction

The organizers of this section of the workshop are to be congratulated for their choice of data. Without being overly complicated, the AneuRisk65 data (Sangalli *et al.*, 2013) presents many non-trivial challenges for analysis. For example: the 65 angiographic images are misaligned due to the different placement of the patients with respect to the image-capturing device; the images have different lengths, with the origin corresponding to a physiologically recognizable landmark but the endpoints being arbitrary; and the main feature of interest, the syphon (Piccinelli *et al.*, 2011), varies in shape and location from person to person.

My analysis of the data was done on the curvature functions, not on the 3D images themselves; this avoids the problem of rotating and translating the 3D curves to remove subject-placement artifacts, but does not remove the inherent variability in shape and location of the artery syphon, corresponding to the peaks around $t = -40$ and $t = -20$ in Figure 1 (the variable t is negative arc length in this parametrization, so the curves run "backwards"). The problem of unequal endpoints is also present whether we analyze the original 3D images or the one-dimensional curvature functions. My approach here is to treat the shorter curves as incomplete curves (which they are). Since the curves for patients with aneurysms at or after the terminal bifurcation of the internal carotid artery (the "upper" group) rarely extend beyond $t = -80$ (i.e. the data

*This research was partially supported by NSF grant DMS 10-06281. The author also thanks the Mathematical Biosciences Institute (MBI) for funding his participation in the workshop.

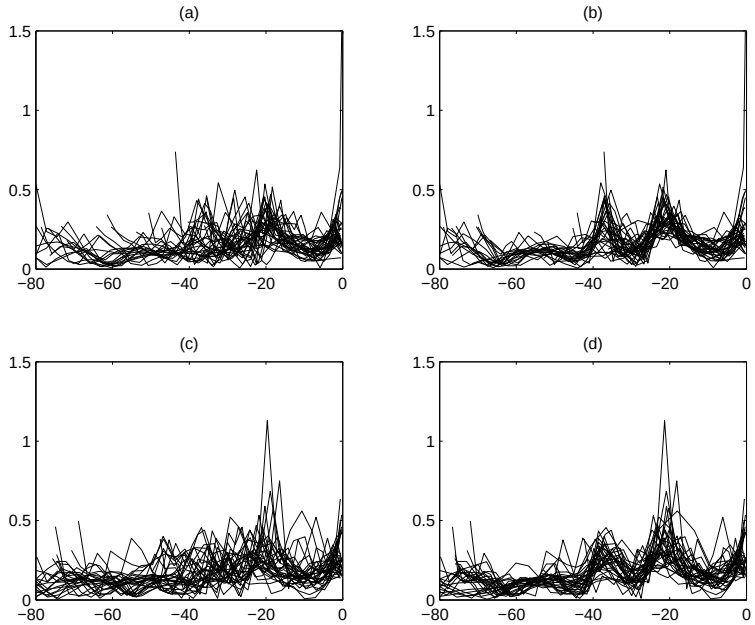

FIG 1. *Curvature functions, down-sampled to 30 measurements per curve, for (a) the "upper" group of patients and (c) the "lower" and no-aneurysm groups of patients. The corresponding warped curves are shown in (b) for the "upper" group and in (d) for the "lower" and no-aneurysm groups.*

is *not* missing at random), we truncated the curves at $t = -80$ in order to avoid artifacts. But many curves were shorter than this, so the problem of unequal endpoints persists; we deal with this by introducing a model that can handle missing data, as explained below.

2. The model

Let $f_1, \ldots, f_n$ be the (complete, unobserved) curvature functions, $f_i : I \to \mathbb{R}$ with $I = [-80, 0]$. The data actually observed is of the form

$$y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n, \tag{1}$$

for different grids $\{t_{i1}, \ldots, t_{im_i}\}$ and random errors $\{\varepsilon_{ij}\}$ (the errors could be assumed to be zero because the curves were pre-smoothed, but the εs are still a useful slack variable to capture the random variation not explained by model (3) below). The variability in location of the syphon will be accounted for by

the warping functions $h_i : I \to I$. We assume, then, that

$$f_i(t) = \tilde{f}_i\{h_i^{-1}(t)\}, \tag{2}$$

where $\tilde{f}_1, \ldots, \tilde{f}_n$ are functions that, loosely speaking, possess only amplitude variability and can therefore be modeled with a parsimonious principal-component decomposition,

$$\tilde{f}_i(t) = \mu(t) + \sum_{k=1}^{p} z_{ik}\xi_k(t), \tag{3}$$

where the ξ_ks are orthonormal functions in $\mathbb{L}^2(I)$ and the z_{ik}s are uncorrelated with decreasing variances. In fact, we will assume $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip}) \sim N_p(\mathbf{0}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p > 0$. We will denote by $\mathcal{F}$ the family of functions spanned by (3), generally referred to as "the template" in the warping literature. The ξ_ks, λ_ks and μ will be estimated from the data; we will assume μ and the ξ_ks are spline functions, thus reducing the estimation problem to a common multivariate problem: given e.g. a B-spline basis $\{\phi_1, \ldots, \phi_q\}$, we assume $\mu(t) = \sum_{k=1}^{q} a_l\phi_l(t)$ and $\xi_k(t) = \sum_{l=1}^{q} c_{kl}\phi_l(t)$ for parameters $\mathbf{a} = (a_1, \ldots, a_q)$ and $\mathbf{c}_k = (c_{k1}, \ldots, c_{kq})$ to be estimated from the data.

For the warping functions h_i we also specify a family of functions $\mathcal{H}$ that is parsimonious but flexible enough to accommodate phase variability at the salient features of the curves. The family of monotone interpolating Hermite splines (Fritsch and Carlson, 1980) is very convenient to work with. Given a knot vector $\boldsymbol{\tau}_0$ of "locations on interest" (for example, $\boldsymbol{\tau}_0 = (-60, -40, -20)$ in our case) and any $\boldsymbol{\tau}_i$ with monotone increasing coordinates, there exists an $h_i \in \mathcal{H}$ such that $h_i(\boldsymbol{\tau}_0) = \boldsymbol{\tau}_i$; this interpolating property provides all the warping flexibility we want at the features of interest, without increasing the dimension of $\mathcal{H}$ unnecessarily. The monotonicity of Hermite splines is very easy to enforce for *any* $\boldsymbol{\tau}_i$s; see Fritsch and Carlson (1980). The individual $\boldsymbol{\tau}_i$s could be either specified by the researcher (as in landmark registration) or treated as unobservable random effects, as we will do here. Since the coordinates of the $\boldsymbol{\tau}_i$s must be strictly increasing and fall within the range I, it is more convenient to transform them into unconstrained vectors $\boldsymbol{\theta}_i$ using e.g. the Jupp transform, and assume $\boldsymbol{\theta}_i \sim N_r(\boldsymbol{\theta}_0, \boldsymbol{\Sigma})$ with $\boldsymbol{\theta}_0$ the Jupp transform of $\boldsymbol{\tau}_0$ and $\boldsymbol{\Sigma}$ a covariance matrix to be estimated from the data. Therefore, our warping functions will be parameterized as $h_i(t) = g(t, \boldsymbol{\theta}_i)$ for a fixed function g that depends only on $\boldsymbol{\tau}_0$ (its exact form does not matter here).

A brief digression: the decomposition (2) is clearly not unique; given any f_i and any arbitrary monotone function h_i, one can always define $\tilde{f}_i = f_i \circ h_i$ and then the decomposition $f_i(t) = \tilde{f}_i\{h_i^{-1}(t)\}$ trivially follows. So it does not make sense to talk about "the" warping component h_i and "the" amplitude component $\tilde{f}_i$ for a given f_i. Nevertheless, for a *given* template $\mathcal{F}$ and a *given* warping family $\mathcal{H}$, the decomposition (2) *is* identifiable (except for the usual indeterminacy on the sign of the ξ_ks). But different combinations of templates and warping models can give rise to essentially equivalent fits. The usual example is the random shift: if $f_i(t) = \mu(t - \tau_i)$, a simple Taylor approximation

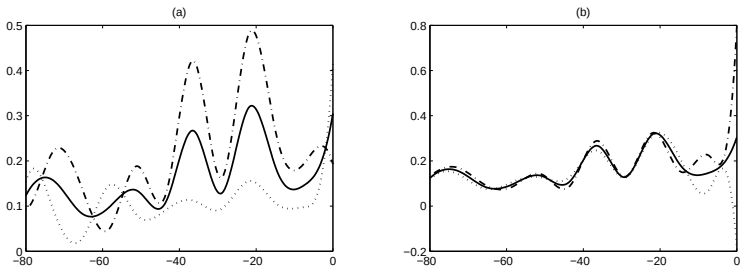

FIG 2. *Amplitude principal components. Mean function (solid line), mean plus principal component (dash-dot line), and mean minus principal component (dotted line), for first [(a)] and second [(b)] principal component.*

yields $f_i(t) \approx \mu(t) - \tau_i \mu'(t)$, so the f_is could be modeled by a one-amplitude-component model without warping just as well. Therefore, when we talk about "the" amplitude component and "the" warping component in this paper, it is always in the context of a specific pair $(\mathcal{F}, \mathcal{H})$.

Going back to the original problem: putting together (1), (2), $\mathcal{F}$ and $\mathcal{H}$, and assuming the ε_{ij}s are i.i.d. $N(0, \sigma^2)$, we obtain the following random-effects model for the raw data $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})$:

$$\begin{aligned} \mathbf{y}_i|(\boldsymbol{\theta}_i, \mathbf{z}_i) &\sim N_{m_i}\{\boldsymbol{\Phi}_i(\boldsymbol{\theta}_i)(\mathbf{a} + \mathbf{C}\mathbf{z}_i), \sigma^2 \mathbf{I}_{m_i}\}, \\ \boldsymbol{\theta}_i &\sim N_r(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}), \\ \mathbf{z}_i &\sim N_p(\mathbf{0}, \boldsymbol{\Lambda}), \end{aligned} \tag{4}$$

with $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_p]$ and $\boldsymbol{\Phi}_i(\boldsymbol{\theta}_i)$ the $m_i \times q$ matrix with elements $[\boldsymbol{\Phi}_i(\boldsymbol{\theta}_i)]_{jl} = \phi_l\{g^{-1}(t_{ij}, \boldsymbol{\theta}_i)\}$ (the inverse of g is taken with respect to the variable t for each $\boldsymbol{\theta}_i$.) The model parameters $\mathbf{a}$, $\mathbf{C}$, σ^2, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ are estimated by maximum likelihood using the EM algorithm. A drawback of this approach is that it was developed for sparse and irregular time grids, and it becomes infeasible for large m_is; therefore we down-sampled the curves so that $m_i = 30$ for all i. Some high-definition features were lost, but the main peaks are still clearly visible in Figure 1.

The random-effect approach to warping described in this section is still unpublished for univariate samples, but a similar approach in the context of functional regression is described in Gervini (2012), where the interested reader may find more technical details.

3. Results

We fitted several models with warping knots $\boldsymbol{\tau}_0 = (-60, -40, -20)$ and different numbers of amplitude components p ranging from 0 (mean-only model) to 5. We used cubic B-splines with 10 equispaced knots for μ and the ξ_ks. The warped

functions for $p = 2$ are shown in Figures 1(b) and 1(d). Plots of $\hat{\mu}$ plus/minus $\hat{\xi}_1$ and $\hat{\xi}_2$ are shown in Figure 3. The first principal component is mostly associated with amplitude variation at the syphon peaks, while the second component is mostly associated with amplitude variation at the origin. Can they be used to discriminate patients with aneurysms at or after the terminal bifurcation of the internal carotid artery (the "upper" group) from patients with no-aneurysms or with aneurysms along the carotid artery (the "lower" group)?

To answer this question we first tried logistic discrimination based on the registered curves $\tilde{f}_1, \ldots, \tilde{f}_n$. Introducing a binary variable y, with $y_i = 1$ indicating the "upper" group and $y_i = 0$ the rest of the patients, the logistic model assumes that

$$p(y_i = 1|\tilde{f}_i) = \text{logist}\left[\alpha + \int_I \beta(t)\{\tilde{f}_i(t) - \mu(t)\}dt\right] \tag{5}$$

for parameters $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{L}^2(I)$. Without loss of generality we can assume $\beta \in \text{span}\{\xi_1, \ldots, \xi_p\}$, since in view of (3) the part of β orthogonal to $\text{span}\{\xi_1, \ldots, \xi_p\}$ will also be orthogonal to $\tilde{f}_i - \mu$. Then we have $\beta(t) = \sum_{k=1}^p b_k \xi_k(t)$ and we can re-write (5) as

$$p(y_i = 1|\tilde{f}_i) = \text{logist}\left(\alpha + \mathbf{b}^T \mathbf{z}_i\right), \tag{6}$$

which is just a common multivariate logistic model. The parameters α and $\mathbf{b}$ were estimated by conditional maximum likelihood, as usual. The crossvalidated misclassification rates for each p are given in Table 1 (first column). The lowest one is attained at $p = 4$, but in the interest of parsimony we choose the second-best, the two-component model, for which the misclassification rate is only slightly larger at 38.5%.

This high misclassification rate is disappointing, and we wonder if the warping process may not contain additional information that could be useful for discrimination. An easy way to answer this question is to augment model (6) with the $\boldsymbol{\tau}_i$s and assume that

$$p(y_i = 1|f_i) = \text{logist}\left(\alpha + \mathbf{b}^T \mathbf{z}_i + \mathbf{d}^T \boldsymbol{\tau}_i\right). \tag{7}$$

Estimating the parameters by conditional maximum likelihood as before, the crossvalidated misclassification rates we now obtain (Table 1, second column) are considerably lower, in particular for the optimal two-component model, which is 24.6%. The parameter estimators are $\hat{\mathbf{b}} = (-8.12, -6.43)$ and $\hat{\mathbf{d}} = (-.15, .22, .27)$. The sign of $\hat{\mathbf{b}}$ indicates that the probability of being in the "upper" group decreases as the height of the peaks at $t = -40$, $t = -20$ and $t = 0$ increases (this is somewhat visible to the naked eye in Figure 1(b) and 1(d).) The signs of the last two coefficients of $\hat{\mathbf{d}}$ also indicate that for patients in the "upper" group the peaks at $t = -40$ and $t = -20$ tend to occur closer to the origin; a caveat is that this could be an artifact of the image-capturing process and not a feature of artery shape, although the negative sign of $\hat{d}_1$ seems to rule this out (because, if the whole curve had been shifted, $\hat{d}_1$ would also be

	CMRs (%)	
p	without $\boldsymbol{\tau}$s	with $\boldsymbol{\tau}$s
0	—	41.5
1	49.2	35.4
2	38.5	24.6
3	47.7	35.4
4	36.9	35.4
5	58.5	46.1

TABLE 1

Crossvalidated misclassification rates for models with p amplitude components, with and without warping parameters included in the model.

positive). Either way, this example shows that the warping process sometimes does contain useful information for classification and discrimination that should not be neglected.

There are a number of ways in which this analysis could be refined. For example, instead of the two-step process followed above, where estimation of amplitude principal components and warping functions is done separately from discrimination, both steps could be brought together by maximizing the likelihood of model (7) instead of (4). The principal components and warping functions thus obtained would have been optimized for discrimination and may yield lower misclassification rates than the two-step process; the author is currently investigating this approach. The other important issue is the handling of incomplete curves. The approach in this analysis was to down-sample the curves and apply a likelihood-based method originally developed for sparsely sampled curves, but in doing so, the sharpest peaks of the curves are dulled or lost entirely; that did not matter much for these data, but in other situations the impact may be significant. The existing registration methods that handle densely sampled curves usually involve functional inner products and norms that require computation of integrals over the whole range I, which cannot be done with incomplete curves (not in an elegant way at least, i.e. avoiding artificial truncations or extrapolations). Finding a way around this problem would be an interesting contribution to the registration literature.

References

[1] FRITSCH, F.N. and CARLSON, R.E. (1980). Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **17** 238–246.

[2] GERVINI, D. (2012). Warped functional regression. *ArXiv 1203.1975.*

[3] PICCINELLI, M., BACIGALUPPI, S., BOCCARDI, E., ENE-IORDACHE, B., REMUZZI, A., VENEZIANI, A., and ANTIGA, L. (2011). Influence of internal carotid artery geometry on aneurysm location and orientation: a computational geometry study. *Neurosurgery* **68** 1270–1285.

[4] SANGALLI, L.M., SECCHI, P., and VANTINI, S. (2013). AneuRisk65: a dataset of three-dimensional cerebral vascular geometries. *Special Section, Electronic Journal of Statistics.*