A Review of Functional Data Analysis Through Density Regression

Zach Horton - Stat 222 Project

Abstract

In this paper we review topics surrounding functional data analysis (FDA) using nonparametric Dirichlet process mixtures, inspired by the work of Rodriguez et. al. (2009). We briefly cover functional data and associated ideas. We discuss foundational material such as the finite Dirichlet process mixture model and its role density regression. The study culminates in an exploration how various FDA inferential goals can be acheived through the use of dependent Dirichlet process priors and a demonstration of curve clustering.

1. Introduction

Functional data analysis (FDA) is a tool used to model data that come in the form of curves. Growth curves, financial trends, and continuous light measurements are examples of commonly used functional data. Many disciplines rely on FDA to make statistically founded conclusions. For example, kinesiology researchers are often concerned with how pain affects movement. FDA could reveal to them how and where movement curves differ between the pain and no pain groups. Such a dataset is displayed in Figure 1. These data come from a study by Seeley et. al. (2020) in which subjects were asked to perform a standing jump. Knee force measurements were recorded over time during the landing phase, resulting in an observed force curve for each subject.

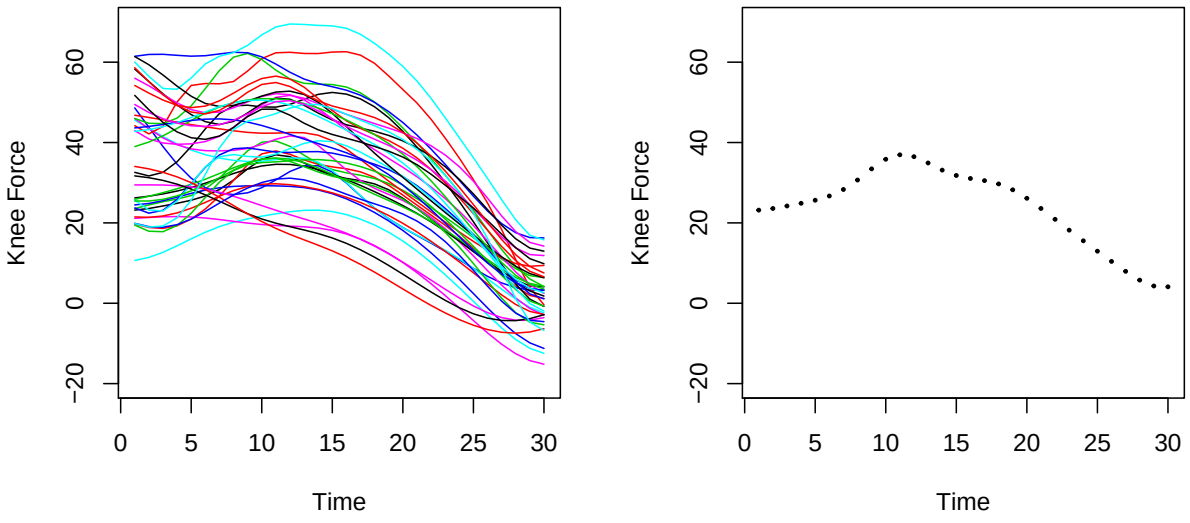

Figure 1: Functional dataset of knee forces measured during a jump landing (left) with a selected curve's points (right)

Ramsay and Silverman (2005) provide a very thorough collection of analysis techniques for functional data, including very popular basis representations such as splines and wavelets. The majority of these methods are parametric, reducing FDA down to multivariate modeling or employing thoughtful penalizations to infinite dimensional models. The multivariate linear model is appealing due to the rich body of methods available for estimation and inference. However, Ramsay and Silverman (2005) argue that there is valuable information contained in a functional object that is lost when projected into a finite dimensional vector space. Consequently, much effort has been devoted to deriving and estimating purely functional models which do not rely on arbitrary (and usually parametric) approximations. Purely functional methods are difficult to use for a number of reasons. Curves are never truly measured beyond a finite number of points, and modern computation is inherently discrete. As well as practical concerns, inferential tasks are difficult to accomplish as functional analogs of variability and asymptotics are complicated mathematical notions.

Bayesian nonparametric (BNP) methods offer a promising solution to this struggle; they avoid the trouble of selecting arbitrary basis approximations while also being more ammenable to performing inference and uncertainty quantification by way of MCMC sampling. Rodriguez et. al. (2009) develop this idea into a powerful method which we will explore. In the remainder this paper, we review primary elements of their work which includes a recap of density regression, the foundational method of BNP curve estimation, a brief discussion of extensions using dependent Dirichlet process priors, and finally a demonstration of curve clustering using the nested Dirichlet process.

2. Density Regression

A landmark paper on BNP curve estimation comes from Müller et. al. (1996) where the fundamentals of density regression and its application to curve fitting are outlined. To develop and demonstrate this concept, we consider fitting a single curve density regression to an observed function, shown in the right panel of Figure 1.

Suppose $\boldsymbol{y} = (y_1, \ldots, y_m)$ denotes the vector of response (knee force) values and let $\boldsymbol{x} = (x_1, \ldots, x_m)$ denote the vector of covariates (time points) corresponding to $\boldsymbol{y}$. Let $\boldsymbol{z}$ be the paired collection of $\boldsymbol{y}$ and $\boldsymbol{x}$ such that $z_j = (y_j, x_j)$. Then consider the following truncated Dirichlet process mixture model representation:

$$
\begin{aligned}
z_j &\sim N_q(\boldsymbol{\theta}_{L_j}, \Sigma_{L_j}) \\
(\boldsymbol{\theta}_{L_j}, \Sigma_{L_j}) &\sim NIW(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \Sigma_0) \\
\boldsymbol{\theta}_0 &\sim N_q(\boldsymbol{\theta}_{00}, D_{00}) \\
\kappa_0 &\sim Ga(a_k, b_k) \\
\Sigma_0 &\sim W(\gamma, \Sigma_{00}) \\
L_j &\sim \sum_{\ell=1}^{N} p_\ell \delta_\ell(L_j) \\
\boldsymbol{p} &\sim Stick(1, \alpha) \\
\alpha &\sim Ga(a_a, b_a)
\end{aligned}
$$

where N_p denotes a p-dimensional normal distribution, $Ga(a, b)$ denotes a gamma distribution with mean $\frac{a}{b}$, W denotes a Wishart distribution, NIW denotes a normal inverse-Wishart distribution, and $Stick(a_\ell, b_\ell)$ denotes a truncated stick-breaking process sequantially derived from $Beta(a_\ell, b_\ell)$ distributions for $\ell = 1, \ldots, L$. In this example we have $q = 2$, but we note that regression with several covariates ($q > 2$) can be done using the same framework. Rodriguez et. al. (2009) suggest specific values for the hyperparameters $\nu_0, \boldsymbol{\theta}_{00}, D_{00}, \gamma, \Sigma_{00}, a_k, b_k, a_a, b_a$, but they also note that results seem robust so long as the scale of the data is taken into consideration.

To fit this model, Gibbs sampling can be used. Specifically, the blocked Gibbs sampler is both useful and common. The complete conditionals from which iterative samples can be drawn are given below:

$$
\begin{aligned}
[\alpha|-] &\sim Ga(N + a_a - 1, b_a - \log p_N) \\
[\boldsymbol{p}|-] &\sim Stick(1 + n_\ell, \alpha + \sum_{r=\ell+1}^{N} n_r); \quad \ell = 1, \ldots, N-1 \\
p(L_j = \ell|-) &\propto p_\ell N_q(z_j|\boldsymbol{\theta}_{L_j}, \Sigma_{L_j}) \\
[\kappa_0|-] &\sim Ga(N^* q/2 + a_k, b_k + \frac{1}{2}\sum_{\ell \in \boldsymbol{L}^*} (\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_0)^T \Sigma_\ell^{-1} (\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_0) \\
[\Sigma_0|-] &\sim W(N^*\nu_0 + \gamma, (\Sigma_{00}^{-1} + \sum_{\ell \in \boldsymbol{L}^*} \Sigma_\ell^{-1})^{-1}) \\
[\boldsymbol{\theta}_0|-] &\sim N(V_0^{-1}\boldsymbol{m}_0, V_0^{-1}) \\
V_0 &= D_{00}^{-1} + \kappa_0 \sum_{\ell \in \boldsymbol{L}^*} \Sigma_\ell^{-1} \\
\boldsymbol{m}_0 &= D_{00}^{-1}\boldsymbol{\theta}_{00} + \kappa_0 \sum_{\ell \in \boldsymbol{L}^*} \Sigma_\ell^{-1}\boldsymbol{\theta}_\ell \\
[(\boldsymbol{\theta}_{L_j}, \Sigma_{L_j})|-] &\sim NIW(\boldsymbol{m}, \kappa_0 + n_\ell, \nu_0 + n_\ell, V) \\
\boldsymbol{m} &= \frac{1}{n_\ell + \kappa_0}(\bar{z}_\ell n_\ell + \kappa_0 \boldsymbol{\theta}_0) \\
V &= \Sigma_0 + \sum_{j:L_j=\ell} (z_j - \bar{z}_\ell)(z_j - \bar{z}_\ell)^T + \frac{\kappa_0 n_\ell}{\kappa_0 + n_\ell}(\bar{z}_\ell - \boldsymbol{\theta}_0)(\bar{z}_\ell - \boldsymbol{\theta}_0)^T
\end{aligned}
$$

where n_ℓ denotes the number of elements in component ℓ, N^* denotes the number of components which contain data, $\boldsymbol{L}^*$ is the vector of components numbers which contain data, and $\bar{z}_\ell$ denotes the mean of the elements in component ℓ. We note that similar expressions are given in Rodriguez et. al. (2009), with differences arising from different Wishart parametrizations.

This model provides a flexible density estimate for the joint distribution of $g(y_j, x_j)$. In attempting to estimate the function $f(x_j)$ which describes the curve of observed y_j values, consider a general linear regression mode of the form $y_j = f(x_j) + \epsilon_j$ where the ϵ_j are distributed with zero mean. Note that in this setting we have $f(x_0) = E(y|x_0)$. A convenient result from the model above which relates a bivariate density estimation to a conditional expectation is that

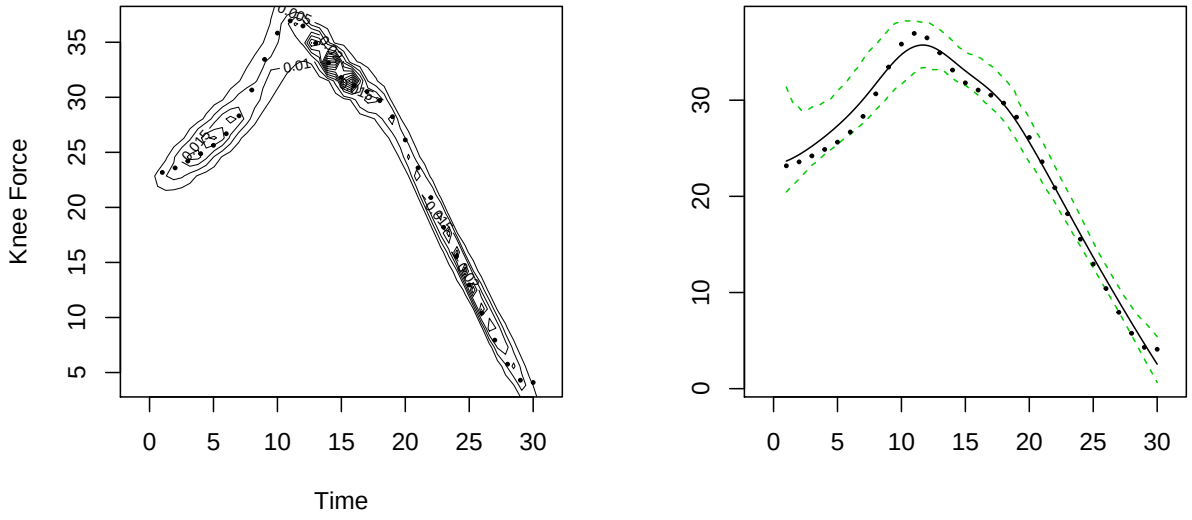

Figure 2: Estimated density plot (left) and fitted function with uncertainty bounds (right).

$$E(y|x_0) = \sum_{\ell=1}^{N} p_\ell N(x_0|\boldsymbol{\theta}_\ell^{(x)}, \Sigma_\ell^{(xx)})(\boldsymbol{\theta}_\ell^{(y)} + \Sigma_\ell^{(yx)}\Sigma_\ell^{(xx)^{-1}}(x_0 - \boldsymbol{\theta}_\ell^{(x)}))$$

where $\boldsymbol{\theta}_\ell^{(x)}$ denotes the subset of $\boldsymbol{\theta}$ corresponding to x, and similar subsetting conventions apply to $\Sigma_\ell^{(xx)}$. By evalutating this expression at every MCMC iterate across a grid of x_0, a posterior distribution for the entire function can be constructed. Although not explored here, observe that this model not only provides means to flexibly model a regression function, but also leads to flexible error distributions as well. Figure 2 shows the fitted density and regression function for our example observed function.

We note that, in the abscence of noise in the data, fitting highly stretched normal mixture components will be somewhat sensitive to prior values surrounding α and Σ_0. This explains why the stimated regression function is not a perfect fit to the data.

3. Hierarchical Models for Multiple Curves

When a single regression function is needed, the method given previously is sufficient. However, functional data analysis often involves analyzing several curves, making independent curve estimation somewhat orthogonal to any functional inference. One naive approach is to consider employing the previous model where $z_j = (y_{1j}, \ldots y_{nj}, x_j)$ and $q = (n+1)$. However, this model will quickly run into posterior consistency problems as the dimensionality of the mixture components gets larger. Additionally, this model could be used to estimate functional relationships between observed curves, which will likely be complex and wasteful computationally if not needed.

Rodriguez et. al. (2009) proposes that dependency be induced by imposing a dependent Dirichlet process (DDP) prior on the collection of G_i mixing distributions. In mathematical terms this roughly be expressed by:

$$z_{ij} \sim \int N(z_{ij}|\boldsymbol{\theta}_{L_j}, \Sigma_{L_j}) dG_i; \quad G_i \sim DDP(\cdot)$$

The authors point out that different inferential goals can be accomplished using different DDP priors. A common-weights DDP prior may have the effect of inducing similar curve complexity or potentially inducing common smoothing. A common-atoms DDP prior may have the effect of curves sharing similar shapes. Through thoughtful consideration of the role the mixture components play in estimating a curve and how a specific DDP prior will affect those, practically any functional inference is possible.

The power in this method cannot be understated. Here we have a fully nonparametric, fully Bayesian method of estimating multiple interdependent regression functions with uncertainty bounds as well as flexibly modeling the error distributions which may vary dynamically over the covariates.

4. Curve Clustering through the Nested Dirichlet Process

We now consider fitting a model to the full dataset shown in Figure 1 in order to cluster curves together. Clustering in a kinesiology setting may be valuable in creating treatment plans. Often physical therapy treatments are either constructed for an individual, which is expensive and time-consuming, or are designed for universal use, which may be insufficient at effectively treating all cases. Clustering may reveal a (hopefully small) number of subpopulations within the overall population. If these subgroups correspond to meaningful physical differences (such as those produced by weight, bone structure, or movement pattern), then a somewhat effective treatment middle-ground can be acheived by designing treatment plans which are universal within a subpopulation. Thus the goal of this functional clustering analysis is to both produce clusters and determine if they correspond to meaningful physical differences. Rodriguez et. al. (2009) suggests that curve clustering can be acheived as a byproduct of density regression if a nested Dirichlet process (NDP) prior is used.

The nested DP was originally proposed by Rodriguez et. al. (2008) and can be expressed as a prior for an unknown distribution G_i by $G_i \sim DP(\alpha, DP(\beta, H))$. This expression makes clear that the basemeasure of an NDP is itself a DP (not a realization from a DP as in the hierarchical DP). The stick-breaking representation makes the clustering effect more obvious:

$$G_i \sim DP(\alpha, DP(\beta, H)) \quad \equiv \quad G_i = \sum_{k=1}^{\infty} \omega_k \delta_{G_k^*}, G_k^* = \sum_{\ell=1}^{\infty} \pi_{\ell k} \delta_{\theta_{\ell k}}$$

Notice that the G_k^* are discrete distributions realized from a DP with base measure H and concentration β and that these are the atoms from which each G_i is drawn. Thus, there is nonzero probability that G_i and $G_{i'}$ are identical distributions. Therefore, mixture components and weights for two seperate curves $f_i(\cdot)$ and $f_{i'}(\cdot)$ can be equivalent, which suggests the curves belong in the same cluster.

In the case of density regression, we consider using the following model formulation which incorporates a finite truncation of the NDP:

$$
\begin{aligned}
z_i|\zeta_i = k, \xi_{ij} = l &\sim N_q(\boldsymbol{\theta}_{lk}, \Sigma_{lk}) \\
(\boldsymbol{\theta}_{lk}, \Sigma_{lk}) &\sim NIW(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \Sigma_0) \\
\boldsymbol{\theta}_0 &\sim N(\boldsymbol{\theta}_{00}, D_{00}) \\
\kappa_0 &\sim Ga(a_k, b_k) \\
\Sigma_0 &\sim W(\gamma, \Sigma_{00}) \\
\zeta_i &\sim \sum_{k=1}^{K} w_k \delta_k(\zeta_i) \\
\xi_{ij} &\sim \sum_{l=1}^{L} \pi_{lk} \delta_{lk}(\xi_{ij}) \\
\boldsymbol{w} &\sim Stick(1, \alpha) \\
\boldsymbol{\pi}_k &\sim Stick(1, \beta) \\
\alpha &\sim Ga(a_a, b_a) \\
\beta &\sim Ga(b_a, b_b)
\end{aligned}
$$

where $\zeta_i = k$ denotes curve i belongs to curve cluster k, $\xi_{ij} = l$ denotes that point j within curve i belongs to component l within cluster k, and $\boldsymbol{\theta}_{lk}$ denotes the $\boldsymbol{\theta}$ vector for the lth component within the kth cluster. Much like the single curve model, the posterior distribution can be sampled from using a blocked Gibbs sampling algorithm, and indeed it is the only algorithm published for the NDP. The full conditionals for this sampling scheme are given by:

$$
\begin{aligned}
[\alpha|-] &\sim Ga(a_a + K - 1, b_a - \log w_K) \\
[\beta|-] &\sim Ga(a_b + K(L-1), b_b - \sum_{k=1}^{K} \log \pi_{Lk}) \\
[\boldsymbol{w}|-] &\sim Stick(1 + m_k, \alpha + \sum_{r=k+1}^{K} m_r)) \\
[\boldsymbol{\pi}_k|-] &\sim Stick(1 + n_{lk}, \beta + \sum_{r=l+1}^{L} n_{rk})) \\
p(\zeta_j = k|-) &\propto w_k \prod_{i=1}^{n_j} \sum_{l=1}^{L} \pi_{lk} N_q(z_i|\boldsymbol{\theta}_{lk}, \Sigma_{lk}) \\
p(\xi_{ij} = l|-) &\propto \pi_{l\zeta_j} N_q(z_i|\boldsymbol{\theta}_{l\zeta_j}, \Sigma_{l\zeta_j}) \\
[\kappa_0|-] &\sim Ga(|n^*_{lk}|q/2 + a_k, b_k + \frac{1}{2} \sum_{l,k:n_{lk}\neq 0} (\boldsymbol{\theta}_{lk} - \boldsymbol{\theta}_0)^T \Sigma_{lk}^{-1} (\boldsymbol{\theta}_{lk} - \boldsymbol{\theta}_0) \\
[\Sigma_0|-] &\sim W(|n^*_{lk}|\nu_0 + \gamma, (\Sigma_{00}^{-1} + \sum_{l,k:n_{lk}\neq 0} \Sigma_{lk}^{-1})^{-1}) \\
[\boldsymbol{\theta}_0|-] &\sim N(V_0^{-1}\boldsymbol{m}_0, V_0^{-1}) \\
V_0 &= D_{00}^{-1} + \kappa_0 \sum_{l,k:n_{lk}\neq 0} \Sigma_{lk}^{-1} \\
\boldsymbol{m}_0 &= D_{00}^{-1}\boldsymbol{\theta}_{00} + \kappa_0 \sum_{l,k:n_{lk}\neq 0} \Sigma_{lk}^{-1}\boldsymbol{\theta}_{lk} \\
[(\boldsymbol{\theta}_{lk}, \Sigma_{lk})|-] &\sim NIW(\boldsymbol{m}_{lk}, \kappa_0 + n_{lk}, \nu_0 + n_{lk}, V_{lk}) \\
\boldsymbol{m}_{lk} &= \frac{1}{n_{lk} + \kappa_0}(\bar{z}_{lk} n_{lk} + \kappa_0 \boldsymbol{\theta}_0) \\
V_{lk} &= \Sigma_0 + \sum_{i,j:\zeta_j = k, \xi_{ij} = l} (z_i - \bar{z}_{lk})(z_i - \bar{z}_{lk})^T + \frac{\kappa_0 n_{lk}}{\kappa_0 + n_{lk}} (\bar{z}_{lk} - \boldsymbol{\theta}_0)(\bar{z}_{lk} - \boldsymbol{\theta}_0)^T
\end{aligned}
$$

where m_k denotes the number of curves assigned to cluster k, n_{lk} denotes the number of points within cluster k assigned to component l, and $|n^*_{lk}|$ denotes total the number of components which are assigned data. As before, we note these differe comewhat from Rodriguez et. al. (2009) in that different Wishart parametrizations are used. To perform clustering with this model, on might look at the posterior distribution of cluster membership. Figure 3 shows results of fitting this model to the knee force dataset where curves have been grouped using the posterior cluster membership median.

It is interesting to note the cluster structure here. Curves with overall larger force likely correspond to patients with higher weights. Thus it would seem reasonable that, should an injury occur with the knee, different treatment protocols ought to consider weight as a factor. Another feature of interest is the presence (or lack) of an initial dip followed by the global max. These may have a more nuanced interpretation, such as an indicator of prior injury or a sign of pronation in the foot.

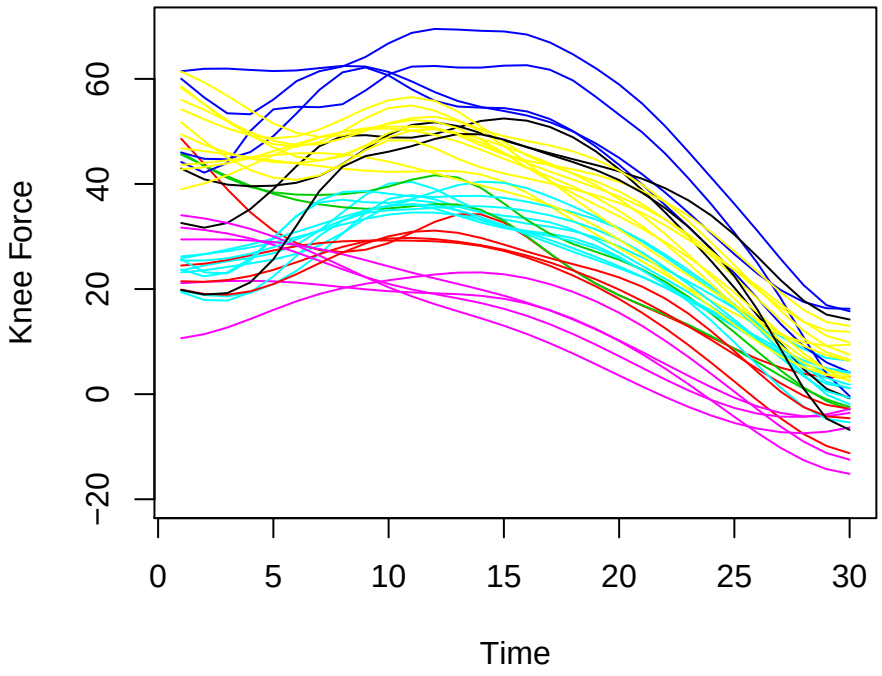

Figure 3: Knee force data with colors representing different clusters

Whatever scientific interpretation may exist, it is gratifying to see that density regression using a nested DP prior is able to infer these groups.

5. Conclusions

Functional data analysis may hold valuable information for many disciplines such as those in the exercise sciences. Analyzing those data may require certain arbitrary decisions which may leave a practitioner unsatisfied. Bayesian nonparametric density regression offers a nice solution in a framework that is quite flexible. When combined with dependent Dirichlet process priors, density regression becomes a powerful model for performing a wide variety of functional inference tasks. The knee force data provide a compelling demonstration of curve clustering and the valuable conclusions that can be drawn from such a versatile model.

REFERENCES

Gelman, A., Carlin, J. P., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, Florida: Chapman & Hall/CRC Texts in Statistical Science.

Müller, P., Erkanli, A., and West, M. (1996), *Bayesian curve fitting using multivariate normal mixtures*. Biometrika, 83(1), 67-79,

R Core Team (2019), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J., and Silverman, B.W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer Series in Statistics.

Rodriguez A., Dunson D.B., and Gelfand, A.E. (2008), *The Nested Dirichlet Process*. Journal of the American Statistical Association, 103:483, 1131-1154.

Rodriguez A., Dunson D.B., and Gelfand, A.E. (2009), *Bayesian nonparametric functional data analysis through density estimation*. Biometrika, 96(1), 149-162.

Seeley, M.K., Denning, W.M., Garner, K., Park, J., Horton, Z., and Hopkins, J.T. (2020), *Anterior knee pain independently alters landing and jumping biomechanics*. Clinical Biomechanics. In Review.