FRODO: A novel approach to micro-macro multilevel regression

Shaun McDonald, Dave Campbell

July 13, 2022

Abstract

Within the field of hierarchical modelling, little attention is paid to micro-macro models: those in which group-level outcomes are dependent on covariates measured at the level of individuals within groups. Although such models are perhaps underrepresented in the literature, they have applications in economics, epidemiology, and the social sciences. Here, we present a new empirical Bayesian technique for fitting such models called FRODO. The method jointly infers group-specific densities for multilevel covariates and uses them as functional predictors in a functional linear model. The power and versatility of FRODO are demonstrated on several simulated datasets, showcasing its ability to accommodate a wide variety of covariate distributions and regression models.

1 Introduction

Hierarchically structured data is quite common in statistics, with a litany of resources and methodology available for almost every imaginable configuration. Books such as [15] provide comprehensive reviews on the subject of multilevel data. For the purposes of this chapter, it will suffice to consider data organized in a two-level hierarchy. Data will be observed from "groups", each of which is comprised of multiple "individuals", with variables measured at either the group level (i.e. one measurement per group) or individual level (i.e. one measurement for each individual within each group).

Multilevel data structures can be broadly categorized into two types: *macro-micro*, in which an individual-level outcome is predicted from group-level covariates; and *micro-macro*, which is the opposite [34]. Although substantial attention has been given to the former structure (random effects models being one example of the macro-micro framework), the micro-macro paradigm is the subject of much less discussion [12], despite the occurrence of such datasets in health sciences [10], sociology [3], and economics [2]. Among the relatively few papers on the subject is the one by Croon and van Veldhoven [9], one of the earliest papers to devise a method specifically for micro-macro regression.

The data structure they considered (hereafter described as "classical") is as follows. Letting subscripts i and ij denote, respectively, the i^{th} group and the j^{th} individual within that group, the basic structure is

$$Y_i = \alpha + \beta\xi_i + \beta_Z Z_i + \epsilon_i, \tag{1}$$

$$\xi_i = X_{ij} + \nu_{ij}. \tag{2}$$

Assuming group i contains n_i individuals, the observed data from that group is $\{Y_i, Z_i, X_{i1}, \ldots, X_{in_i}\}$. In words, Y_i is a group-level response variable (with regression error ϵ_i), Z_i is a group-level scalar covariate, and the X_{ij}'s are individual-level measurements of some "latent" unobserved covariate ξ_i with errors ν_{ij}. One can think of the model as two "parts": a regression part specified by (1), and a covariate observation part specified by (2). The linearity of the regression and additivity of the covariate error justify the "classical" moniker for this structure.

Although micro-macro modelling literature is relatively scarce, the structure implied by (1–2) is essentially equivalent to (a version of) the much better-studied *classical measurement error model* [chapter 1 of 8, and references therein]. The main difference is conceptual: in a micro-macro model, replicate covariate measurements correspond to distinct individuals within a group; while in a measurement error model, they are merely repeated noise-corrupted observations of some true explanatory variable for the i^{th} observational unit. There is another practical difference: most measurement error literature assumes smaller n_i's (the number of covariate measurements per group) than one tends to encounter in a "true" micro-macro setting.

The simplest approach to modelling such data is the "naive" one: simply using the sample means $\bar{X}_i = n_i^{-1} \sum_j X_{ij}$ as proxies for the latent ξ_i's. However, it is well-known [e.g. chapter 3 of 8, and references therein] that such a failure to account for the uncertainty in the X_{ij}'s biases estimates of the regression parameters. Most notably, it creates *attenuation* in the estimate of β: letting $\hat{\beta}$ denote such an estimate, we will have $|\widehat{\beta}| < |\beta|$, even as the number of groups grows asymptotically. In intuitive terms, this "attenuation" happens because the noise in the covariates stretches the regression line on the horizontal axis. Thus, a plethora of both frequentist and Bayesian methods have been proposed to account for covariate uncertainty in a way that produces less biased estimation and inference for the regression part of the model. A comprehensive review of measurement error methodology is beyond the scope of this chapter, but the interested reader may refer to books such as [6, 8] or the review paper of Schennach [31].

Many real-world datasets do not obey the "classical" framework of (1–2) [e.g. Section 6.4 of 6, and references therein], and there are two ways to transcend it: by replacing the linear terms $\beta\xi_i$ and $\beta_Z Z_i$ in (1) with arbitrary regression functions, or by generalizing the additive covariate structure in (2). There are few micro-macro modelling papers with generalizations of either type, aside from the discrete variable methods of Bennink et al. [3, 4]. Thus, we focus our attention here on the measurement error literature instead. Beyond the compre-

hensive review sources mentioned above, the most generalized framework which is relevant to this chapter is that of Hu and Schennach [18]. They assumed each observational unit i only had a single covariate measurement $X_i \sim f_{X|\xi=\xi_i}$, but also had a single replicate measurement or *instrumental variable* W_i, assumed to provide further information about ξ_i. They also allowed a very general form for the regression function in which Y only depended on the unobserved ξ, with only some technical assumptions on the distributions of $Y \mid \xi$, $X \mid \xi$, and $\xi \mid W$. Their assumptions on the covariate structure were very broad, requiring only that there exists a functional M such that $M\left[f_{X|\xi}\left(\cdot \mid \xi\right)\right] \equiv \xi$ for all ξ. Examples of such functionals include the mode, as well as any quantile or moment. With this framework, the authors proposed a sieve likelihood estimator for the regression parameters and the densities of $X \mid \xi$ and $\xi \mid W$. To our knowledge, there are no established Bayesian methods that accommodate this level of generality. Sarkar et al. [30] proposed a Bayesian model which used Dirichlet Process mixtures to achieve a great deal of flexibility in modelling the regression function, latent covariates, and error terms; but it still assumed an additive error structure of the form (2).

Neither of the aforementioned papers (or, indeed, any measurement error literature we have seen) gives much consideration to the "unit-specific" covariate distributions $f_{X|\xi=\xi_i}$ — specifically, to any differences between them across units. This is understandable, as most errors-in-variables problems have no more than a single-digit number of covariate measurements available per unit, making any such differences irrelevant. However, in an explicitly multilevel setting, there are typically many more individuals per group [e.g. 9, 2], and it may be of interest to explicitly consider the group-specific covariate densities in inference. We believe that the Bayesian paradigm (or, at the very least, the empirical Bayesian paradigm) is the most natural setting in which to achieve this.

With all of the above considerations in mind, our goals in this chapter are threefold. First, we seek to develop a(n empirical) Bayesian model with generality comparable to that of Hu and Schennach [18]. Second, we wish to apply this model in the micro-macro multilevel setting, providing an ability to accommodate "non-classical" data structures which we believe is sorely missing in that literature. Our final goal is to leverage the data sizes characteristic of micro-macro situations in order to focus our inference not only on the regression part of the model, but also the distributions of "individual-level" covariates within each group.

To achieve these goals, we propose **FRODO** (Functional Regression On Densities of Observations), a method which unifies density estimation and functional regression in a joint empirical Bayesian model. Although the core idea of FRODO is a fairly straightforward combination of well-established methods in principle, it allows for a remarkable degree of generality in data structures, and its design proves to be far from trivial.

Before describing FRODO, we first give an overview of necessary functional data analysis concepts in Section 2. We then give a general overview of the FRODO model and its assumed data structure in Section 3, followed by a de-

tailed description of its prior and likelihood components, as well as its practical implementation. In Sections 4 and 5, we show several simulation studies which demonstrate the potential generality of FRODO in both the regression and covariate observation parts of a micro-macro model.

2 A brief review of key functional data analysis concepts

Broadly speaking, *functional data analysis* (FDA) is a field of statistics in which the fundamental units of interest. A detailed overview of the field is beyond the scope of this chapter, but the interested reader may find one in the excellent book by Ramsay and Siverman [27]. Here we discuss only the concepts necessary to establish notation and motivation for FRODO.

2.1 Scalar-on-function functional regression

As the name implies, scalar-on-function regression concerns the modelling of a real-valued univariate (or "scalar") response variable with predictors that are functions [27, Section 12.3]. This is achieved by using integrals in place of the sums which define scalar regression models. For example, consider a simple case in which our data are pairs $\{Y_i, f_i^*\}$, $i = 1, \ldots, N$, where Y_i is a continuous-valued scalar responses and f_i^* is an almost everywhere continuous function on $[0, 1]$. For this data, a *functional linear model* would be of the form

$$Y_i = \alpha + \int_0^1 \beta^*(x) f_i^*(x) \mathrm{d}x + \epsilon_i, \tag{3}$$

with i.i.d. errors $\epsilon_i \sim \mathcal{N}(0, \sigma_Y)$. The *coefficient function* β^* weighting the integral is analogous to regression coefficients in a fully scalar regression model.

2.2 Basis function expansions

Because function spaces are infinite-dimensional, a core component of FDA is the representation of functions of interest in finite-dimensional spaces [27]. Typically, this is achieved by modelling functions as linear combinations of finitely many *basis functions* [27, Section 3.3]. Throughout this chapter, we will use f^* to denote a function of interest, and remove the asterisk to denote a relevant basis function approximation f.

Several types of functional bases exist, including those based on functional principal components, Fourier series, and splines [27]. Attention here is restricted to the latter, and in particular the P-splines of Eilers and Marx [11]. For our purposes, it suffices to know that a P-spline representation of a function f^* on a compact interval $[a, b]$ has the form

$$f(x) = \sum_{k=1}^{K} c_k B_k(x), \tag{4}$$

where the spline basis functions B_k are piecewise polynomials with supports defined by a set of equally-spaced "knots" in $[a, b]$. More detailed explanations of splines can be found in both [27] and Eilers and Marx [11]. In the frequentist setting, the coefficients $c = (c_1, \ldots, c_K)$ can be fit with a penalized likelihood method. Common penalties force f (as defined in (4)) to adhere to desirable shapes by penalizing "roughness", as measured with a suitable linear differential operator [see 27, Chapter 5]. Eilers and Marx [11] modified this idea by instead using a penalty based on *finite differences* between coefficients. Their penalty defines the notion of *P-splines* and is of the form

$$\lambda \sum_{k=r+1}^{K} \left[(\Delta^r c)_{k-r} \right]^2, \tag{5}$$

for a positive integer r, where Δ^r denotes the r^{th}-order finite difference penalty and $(\Delta^r c)_{k-r}$ denotes the $(k-r)^{\text{th}}$ element of the $(K-r)$-dimensional vector $(\Delta^r c)$. For instance,

$$\begin{aligned}
\left(\Delta^1 c\right)_1 &= c_2 - c_1, \\
\left(\Delta^2 c\right)_1 &= c_3 - 2c_2 + c_1, \text{ and} \\
\left(\Delta^3 c\right)_1 &= c_4 - 3c_3 + 3c_2 - c_1.
\end{aligned}$$

When the *smoothing parameter* $\lambda > 0$ is large, (5) dominates the penalized likelihood. Eilers and Marx [11] noted that the sum in this penalty is a good approximation to the r^{th} derivative of f when the knots defining the spline basis are equally spaced, especially for large dimensionality K. Thus, for large λ the estimated f is forced to take the approximate shape of a polynomial of degree $r-1$.

Lang and Brezger [20] devised a Bayesian version of P-splines, based on the notion that a penalized likelihood function is analogous to a posterior distribution on the log scale, with the penalty term assuming the role of the prior. The penalty 5 is the log density of an r^{th}-order Gaussian *random walk*:

$$(\Delta^r c)_{k-r} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{2\lambda}}\right) \tag{6}$$

for $k = r, r+1, \ldots, K$. Lang and Brezger [20] gave the first r components of c (which we call "*free parameters*" in contrast with the last $K-r$ components, whose behaviour is restricted by (6)) flat priors. However, we adopt the philosophy that such priors are unreasonable because they give equal weight to all values, no matter how extreme Betancourt [e.g. the case study of 5], and we have also found such priors to result in extremely poor MCMC sampling behaviour in our models. Our priors on the free parameters in the various P-spline components of FRODO are described in Sections 3.2–3.3.

As noted by Eilers and Marx [11], one can use P-splines to model a density g^* by letting $f^* = \log g^*$ and approximating f^* with (4). The imposition of a polynomial shape on f then leads to density estimate which is close to the

exponentiation of the corresponding polynomial. For instance, using a penalty of order $r = 3$ (in either the frequentist or Bayesian setting) forces f towards a quadratic shape, and therefore the resulting density estimate will be similar in shape to a Gaussian.

3 The FRODO model

3.1 General overview

Having reviewed the necessary functional data analysis concepts, we are now ready to describe the FRODO approach to micro-macro modelling. Assume the data is organized into N groups, with the i^{th} group containing n_i individuals. In the simplest case (assumed in the remainder of this section for ease of exposition), data i^{th} group consists of a group-level response variable Y_i, and individual-level observations of a covariate X, $(X_{i1}, \ldots, X_{in_i})$. Although we assume real-valued Gaussian Y_i's throughout this chapter for the sake of simplicity, in principle the following methodology could be extended to any response type for which generalized linear modelling is possible. As in Section 1, the model is comprised of both a regression part and a covariate observation part, but we assume a much greater level of generality than in (1–2). Our only assumption for the covariate density part is that, for the i^{th} group, $X_i := (X_{i1}, \ldots, X_{in_i})$ (where an omitted subscript means the collection of all elements across that subscript) is an i.i.d. sample from an unobserved or "latent" group-specific covariate density f_i^*. The regression part of the model defines the "novel" idea at the core of FRODO: the use of these densities (technically, basis expansion estimators thereof) as predictors in a functional linear regression. In mathematical terms, the regression part of the model is

$$Y_i = \alpha + \int \beta^*(x) f_i^*(x) \mathrm{d}x + \epsilon_i \tag{7}$$

$$= \alpha + \mathbb{E}_i^* \left[\beta^*(X)\right] + \epsilon_i, \tag{8}$$

$$\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_Y),$$

where $\mathbb{E}_i^* \left[\beta^*(X)\right]$ denotes the expectation of $\beta^*(X)$ with respect to the density f_i^*. The equivalence between (7) and (8) is the key to FRODO's utility: by simply using densities as predictors in a functional linear regression, the resulting model is essentially a GAM. Thus, FRODO allows for a fully nonparametric approach to both regression functions and covariate structures.

It must be noted that the regressor in (8), $\mathbb{E}_i^* \left[\beta^*(X)\right]$, is the "expectation of the regression function". In general, this is *not* equal to $\beta^*\left(\mathbb{E}_i^*[X]\right)$ — the "regression on the expectations" — unless $\tilde{\beta}$ is linear. Use of the latter is perhaps more "standard" in the measurement error literature, where it is typically assumed that the X_{ij}'s within each unit i are noise-corrupted versions of some "true" covariate ξ_i [see 8, or any standard reference on measurement error]. Although it is not always assumed that $\mathbb{E}_i^*[X] = \xi_i$ (e.g. the general linear error structures described in Section 6.4 of [6], and references therein), typically

the target is estimation of $\beta^*(\xi_i)$, possibly marginalized over an estimate of the "posterior" $f_{\xi|X_i}$ [e.g. 17, 21]. We are not aware of any literature which explicitly uses "expectations of the regression" in the way that FRODO does.

In the next two subsections, we detail the priors and likelihoods comprising FRODO. Recall that we approximate β^* and the f_i^*'s with basis function expansions, use of which will be denoted without asterisks. In a slight abuse of notation, we consider the model

$$Y_i = \alpha + \int \beta(x) f_i(x) \mathrm{d}x + \epsilon_i \tag{9}$$

$$= \alpha + \mathbb{E}_i\left[\beta\left(X\right)\right] + \epsilon_i, \tag{10}$$

as a proxy to (7–8), where β and f_i are the basis function approximations to their "true" counterparts, and $\mathbb{E}_i$ denotes expectation w.r.t. f_i.

Before exploring the details of FRODO, some final technical and notational points are in order. We recommend standardizing the data so that default prior choices are weakly informative [14, Sections 2.9 and 16.3]. Keeping with our convention of using omitted subscripts to mean the collection of all elements across that subscript, let $Y = (Y_1, \ldots, Y_N)$ and $X = \{X_1, \ldots, X_N\}$, where X_i was defined above. In what follows, we will assume that Y and X have both been standardized to have zero mean and unit variance. Note that for X, this standardization is "marginal", meaning that it is done across groups *and* individuals within groups. We will overload notation and use f_i^* and f_i to refer to, respectively, the true density and its basis function approximation for the standardized version of X_i. For technical reasons, it is necessary to assume that β and the f_i's are all defined over a common compact interval. This will be denoted by $[a, b]$ on the standardized scale, and when it is necessary to speak about the domain of the covariates on the original (unstandardized) scale, it will be denoted by $[a', b']$. Assuming X has been standardized as recommended above, we have $a = \left(a' - \bar{X}\right)/\sigma(X)$ and $b = \left(b' - \bar{X}\right)/\sigma(X)$, and $[a', b']$ can be chosen so that its endpoints are (nearly) equal to the unscaled extrema of the covariates.

3.2 The density model

For computational convenience — and because it suffices for the ordinal covariates which are common in real micro-macro datasets [e.g. 9, 2, 10] — the f_i's are modelled as histograms. In practical terms, this means that they are linear combinations of constant basis functions:

$$f_i(x) = \sum_{k=1}^{K} \phi_{ik} I_k(x), \tag{11}$$

where I_k is the k^{th} subinterval $[a + (k-1)h, a + kh)$, and $h = (b-a)/K$ is the bin width. The density coefficients ϕ_{ik} are scaled "softmax" transformations of

Gaussian random variables θ_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, K$:

$$\phi_{ik} = \frac{e^{\theta_{ik}}}{h \sum_{j=1}^{K} e^{\theta_{ij}}}, \tag{12}$$

where, for all i, $\theta_{i1} \equiv 0$ to ensure identifiability. Equivalently, we may say that the ϕ_i's are (up to the scaling factor h) logistic normal random vectors [1].

The priors for the θ's are chosen in order to impose useful constraints on the behaviour of the densities. In particular, for some positive integer r we will impose an r^{th}-order Gaussian random walk prior on $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$ for all i. Since the logarithms of the f_i's are also piecewise constant, this structure means that $\log f_i$ is a Bayesian P-spline of degree zero, with r^{th}-order penalty, for all i. Recall from Section 2.2 that an r^{th}-order random walk prior on θ_i,

$$\left(\Delta^r \theta_i\right)_{k-r} \sim \mathcal{N}\left(0, \tau_i\right), \; k \geq r+1 \tag{13}$$

forces $\log f_i$ towards the approximate shape of a $(r-1)^{\text{th}}$-degree polynomial when the smoothing parameter τ_i is small[1]. .

Note that (13) completely determines the conditional distributions of θ_{ik} for $k > r$ given θ_{ik} for $k \leq r$. In the case $r > 1$, it remains to set the priors on the "free parameters" θ_{ik} for $2 \leq k \leq r$: the "initial values" of the random walk. A seemingly sensible and simple choice would be diffuse, mean-zero, independent Gaussian priors. Unfortunately, this turns out not to be entirely suitable for FRODO. For $r > 1$, imposing fully independent priors on the densities[2] causes bias in the posterior mean coefficient function, $\widehat{\beta}$. For instance, if the true β is a linear function, the magnitude of the slope of $\widehat{\beta}$ will be biased downward, just as in the "naive" approach to modelling described in Section 1. In the Bayesian hierarchical setting, this "attenuation" problem can be solved by putting priors on the covariates which introduce dependence between them and "pool" each group's measurements towards a latent group-level variable. The solution here is similar.

To expand on this, first note that with $\theta_{i1} \equiv 0$ for all i, we have

$$\theta_{ik} = \log\left(h^{-1} \int_{a+(k-1)h}^{a+kh} f_i(x)\mathrm{d}x\right) - \log\left(h^{-1} \int_{a}^{a+h} f_i(x)\mathrm{d}x\right) \tag{14}$$

$$\approx \log f_i^*\left(a + h\left(k - \frac{1}{2}\right)\right) - \log f_i^*\left(a + \frac{h}{2}\right), \tag{15}$$

recalling that f_i^* is the "true" density for group i.

[1] Henceforth, the phrase "smoothing parameter" will refer to the standard deviation of the random walk prior (τ), instead of its precision as in Section 2.2 (where it was denoted by $\lambda = \tau^2/2$

[2] When discussing the model itself, we will typically write "the densities" to refer to the histograms f_i which are actually part of the model. When it is necessary to invoke the f_i^*'s, we will specify them as the "*true* densities".

Suppose f_i^* is that of a $\mathcal{N}(\xi_i, \sigma_i)$ random variable[3]. This corresponds to the limiting case for $r = 3$ as $\tau_i \to 0$, and it can be shown that (15) in this case reduces to

$$\theta_{ik} \approx \frac{h(k-1)}{\sigma_i^2}\left(\xi_i - \left(a + \frac{kh}{2}\right)\right) \tag{16}$$

For $r = 3$, this approximation motivates our choice of priors for the "free parameters". For each i and $k = 2, 3$, we take them to be Gaussian with mean given by the right side of (16) and standard deviation τ_i. Thus, τ_i controls f_i's adherence to the limiting Gaussian shape in two respects: by controlling the free parameters' deviations from their means, and by scaling the random walk behaviour in (13).

We now set priors on ξ_i and σ_i. When the true covariate densities are Gaussian, the structure of the data is analogous to that of the "classical" micro-macro model, with ξ_i being a "latent group-level covariate" and σ_i controlling the level of Gaussian noise for each group's individual-level covariate measurements. In keeping with natural choices for that setting, we first assign the ξ_i's a $\mathcal{N}(\mu_\xi, \sigma_\xi)$ prior. Recalling that $[a', b']$ denotes the assumed domain of the covariate densities on the original (unstandardized) scale, the mean μ_ξ is given a $\mathcal{N}\left((a'b - b'a)/(b' - a'), 15/K^2\right)$ hyperprior. This corresponds to a mean-zero hyperprior on the original covariate scale, with the empirically-determined standard deviation $15/K^2$ accounting for the discretization error from approximation (15). The scale σ_ξ is given a standard half-normal prior, which will be fairly uninformative if the X_{ij}'s have been scaled to have unit marginal variance. It will often be reasonable that the covariate densities are homoscedastic: $\sigma_i \equiv \sigma_X$ for all i. A standard half-normal prior is a sensible choice in this case. If one wishes to explicitly model heterogeneity, then each σ_i can be given its own half-normal prior, perhaps sharing a common scale parameter with its own hyperprior.

Now, suppose f_i^* is instead a (shifted) Exponential(λ_i) density. This corresponds to the limiting case for the random walk with $r = 2$, and here (15) reduces to

$$\theta_{ik} = -\lambda_i (k-1) h. \tag{17}$$

Note that for an exponential density, there is no discretization error, so (14) and (15) are equal. Thus, analogously to the $r = 3$ case described above, when $r = 2$ we assume the "free parameters" θ_{i2} are Gaussian with mean given by the right side of (17) and standard deviation τ_i. A natural choice of prior for the "latent rates" λ_i is Gamma$(\alpha_\lambda, \alpha_\lambda/\mu_\lambda)$. The mean μ_λ is given a standard half-normal prior (which should be only weakly informative if the covariates have been standardized), while the shape parameter α_λ is given a more diffuse half-normal prior with scale 10. Note that this parameterization of the Gamma

[3] Assuming the covariates have been standardized as recommended in Section 3.1, most of f_i^*'s mass presumably lies in $[a, b]$, and $a < \xi_i < b$.

in terms of shape and mean, rather than the more conventional shape and rate, proved computationally advantageous.

By defining the "free parameters" in terms of latent group-level variables with their own hyperpriors, we introduce the necessary dependence and "pooling" to prevent bias in the regression part of the model, just as one might do in the scalar case. For any order r, the density model is completed with priors on the smoothing parameters τ_i, which we take to be exponentials with rates δ_i^{-1}. The scales are assumed to be fixed data, chosen empirically based on heuristics and the properties of the X_{ij}'s in the absence of more meaningful prior information. Such choices place FRODO in the category of "empirical Bayesian" methods, but we have found that sampling behaviour and posterior results can become poor when the δ_i's are not chosen carefully. If group sizes are moderate (n_i's roughly between 20 and 60) and one doesn't expect any of the covariate densities to deviate too seriously from the shape implied by the r^{th}-order random walk prior, $\delta_i = 0.1$ for all i seems to be a good default choice based on preliminary empirical results. Smaller groups tend to require smaller δ_i's, and it may also be advantageous to shrink them when the basis dimension K is very large, especially relative to the n_i's.

Finally note that, because the densities are piecewise constant, the likelihood $X_i \sim f_i$ is equivalent to $m_i := (m_{i1}, \ldots, m_{iK}) \sim \text{Multinomial}\,(n_i, \phi_i)$, where m_{ik} is the bin count $|\{j : X_{ij} \in I_k\}|$. In summary, the model for the densities,

assuming an r^{th}-order random walk prior structure (for $r \leq 3$), is

$$
\begin{aligned}
m_i &\sim \text{Multinomial}\left(n_i, \phi_i\right) \\
\phi_{ik} &= \frac{e^{\theta_{ik}}}{h \sum_{j=1}^{K} e^{\theta_{ij}}} \\
\theta_{i1} &\equiv 0
\end{aligned}
$$

$$
\left.
\begin{aligned}
\theta_{i2} &\sim \mathcal{N}\left(-\lambda_i h, \tau_i\right) \\
\lambda_i &\sim \text{Gamma}\left(\alpha_\lambda, \frac{\alpha_\lambda}{\mu_\lambda}\right) \\
\alpha_\lambda &\sim \text{Half-Normal}(0, 10) \\
\mu_\lambda &\sim \text{Half-Normal}(0, 1)
\end{aligned}
\right\} \quad r = 2
$$

$$
\left.
\begin{aligned}
\theta_{ik} &\sim \mathcal{N}\left(\frac{h(k-1)}{\sigma_i^2}\left(\xi_i - \left(a + \frac{kh}{2}\right)\right), \tau_i\right) \ (k = 2, 3) \\
\xi_i &\sim \mathcal{N}\left(\mu_\xi, \sigma_\xi\right) \\
\mu_\xi &\sim \mathcal{N}\left(\frac{a'b - b'a}{b' - a'}, \frac{15}{K^2}\right) \\
\sigma_\xi &\sim \text{Half-Normal}(0, 1) \\
\sigma_X &\sim \text{Half-Normal}(0, 1)
\end{aligned}
\right\} \quad r = 3
$$

$$
\begin{aligned}
\left(\Delta^r \theta_i\right)_{k-r} &\sim \mathcal{N}\left(0, \tau_i\right), \ k > r \\
\tau_i &\sim \text{Exp}(\delta_i^{-1})
\end{aligned}
$$

3.3 The regression model

Here we detail the regression part of FRODO. Recall that we have restricted our attention in this chapter to continuous real-valued responses Y_i with i.i.d. errors $\epsilon_i \sim \mathcal{N}\left(0, \sigma_Y\right)$, The following priors on α and β would require only minor changes to accommodate more general response types (e.g. different scaling may be in order to ensure plausible effect sizes in a logistic regression; see Section 16.3 of Gelman et al. [14]), and the prior on the dispersion parameter could easily be changed as necessary.

The error scale σ_Y is given a half-T prior with 4 degrees of freedom and scale $1/\sqrt{2}$, so that σ_Y has a prior mean of $1/\sqrt{2}$. Recalling the assumption from Section 3.1 that Y has been standardized to have unit variance, this scale (in informal terms) loosely corresponds to a prior expectation that roughly half of the variation in the response values is due to regression error (assuming that the errors and regressors are independent, which we do here). This seems to be a sensible approach for a "default" prior, unless one has prior domain knowledge which would allow for context-specific prior beliefs about the regression error.

Both α and β are given hierarchical priors with scales proportional to σ_Y. This can be shown to ensure unimodality in some penalized Bayesian regression models [25], and we also found that it improved sampling behaviour. The intercept α is given a diffuse $\mathcal{N}(0, 20\sigma_Y)$ prior.

We take the coefficient function β to be piecewise constant, with the same dimensionality K as the densities. This is quite computationally convenient, as the integral in (9) then reduces to the inner product between the coefficients of β and f_i, scaled by the bin width h. Because the functional predictors all have unit integral, adding a constant shift to β does not change the model: for any $c \in \mathbb{R}$, the model is identical if β and α are replaced by $\beta + c$ and $\alpha - c$, respectively. Thus, we impose the identifiability constraint $\mathbb{E}[\beta(X)] := \int_a^b \widehat{f}_{\text{Cent}}(x)\beta(x)\mathrm{d}x = 0$, where $\widehat{f}_{\text{Cent}}$ is the *empirical central density*:

$$\widehat{f}_{\text{Cent}}(x) := \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} m_{ik}}{\sum_{l=1}^{K}\sum_{i=1}^{N} m_{il}}. \tag{18}$$

Essentially, $\widehat{f}_{\text{Cent}}$ is the "marginal histogram" of all covariate data across groups. Presumably, the total number of covariate observations $\sum_i n_i$ will be large enough in most data sets to ensure that $\widehat{f}_{\text{Cent}}$ is reasonably "smooth", so that it is a good approximation to the "marginal" covariate density (i.e. marginalized across groups) for large K. Note that we use the *empirical* central density mainly for computational convenience: an "inferred central density" like $N^{-1}\sum_i f_i$ would certainly be "smoother", but this would add needless complexity to the gradients used in NUTS when the empirical version is sufficient to ensure identifiability.

This constraint amounts to centering the inferred regressors $\mathbb{E}_i[\beta(X)]$. In practice, the constraint is achieved by defining a piecewise constant function

$$\beta^0(x) := \sum_{k=1}^{K} \beta_k^0 I_k(x) \tag{19}$$

and taking $\beta = \beta^0 - \int \widehat{f}_{\text{Cent}}\beta^0$. In keeping with `[litany of Bayesian FLR sources]`, we put a second-order random walk prior on the coefficients of β^0, with the first coefficient set to 0 for identifiability:

$$\begin{aligned} \beta_1^0 &\equiv 0, \\ \beta_2^0 &\sim \mathcal{N}(0, 20h\sigma_Y), \\ \left(\Delta^2\beta^0\right)_{k-2} &\sim \mathcal{N}(0, \tau_\beta\sigma_Y). \end{aligned}$$

The smoothing parameter τ_β controls the extent to which β deviates from the random-walk behaviour. As $\tau_\beta \to 0$, β is forced towards a stepwise approximation to a straight line, and the regression model (9) is therefore forced towards a linear regression. In this limiting case, the "slope" of β, $h^{-1}\beta_2^0$, is equivalent to the regression coefficient in a scalar linear model. Thus, using a scale factor of $20\sigma_Y h$ in β_2^0's prior can be considered roughly analogous to placing a

$\mathcal{N}(0, 20\sigma_Y)$ prior on the coefficient in the scalar case, which should be reasonably diffuse if the covariates have been scaled as recommended above [e.g 35, Section 25.12 of User's Guide]. Finally, τ_β is given an exponential prior with rate 2 (equivalently, scale 0.5). In contrast to the smoothing parameters for the densities, we found that τ_β did not require a careful selection of prior scale in order to ensure good model performance.

3.4 Implementation

The FRODO model is implemented in the Stan programming language [7], which provides exceptional power, flexibility, and efficiency through its use of the No-U-Turns Sampling (NUTS) variant of Hamiltonian Monte Carlo [16]. For each of the below simulation studies, four parallel chains were run with fairly diffuse starting values, with sufficiently many sampling iterations to ensure effective sample sizes of at least 500 for all parameters [see 14, Section 11.5]. All model runs were devoid of divergent transitions [35], and the overwhelming majority of parameters in all simulations had $\hat{R}$ values (where $\hat{R}$ is a diagnostic which helps to assess model convergence, see Vehtari et al. [39]) below 1.01, with only a single parameter in the model of Section 4.2 having a value very slightly above this threshold. All of the simulation studies below were conducted using R [26], interfacing with Stan via the RStan package [36].

4 Simulation studies

As discussed in previous sections, FRODO is uniquely powerful in theory because it is "doubly nonparametric": it can capture arbitrary unknown structures in both the covariate densities and the regression model. In the following subsections, we put this to the test with a wide variety of simulated datasets. We will assess FRODO's ability to harness location, scale, and shape information from covariate densities and use it to recover true regression relationships. In each study, FRODO will be compared to two simpler models:

1. a "naive" scalar regression model using only the sample means of the covariate measurements (or of some suitable transformation, where applicable); and

2. a "hierarchical" scalar regression model, where the form of the regression function and covariate distributions are assumed known, with only the actual parameter values unknown.

More detail will be provided in the following subsections.

Because FRODO does not assume any parametric form for either the regression or covariate parts of the model, all that is required are choices of an appropriate random walk order r, dimensionality K, (unstandardized) density domain $[a', b']$, and set of density scaling factors $\delta = (\delta_1, \ldots, \delta_N)$. These choices must be made assuming that the true data-generating mechanisms are not known

a priori. One could use subject-specific domain knowledge if it is available. Otherwise, an "empirical Bayesian" approach based on informal inspections of the data is acceptable, and this is the approach we will use for all simulation studies in this chapter. Visual inspection of default histograms or KDE's suffices to this end. From a strictly Bayesian perspective on inference, one could argue that this data dependence in the prior is not philosophically sound. However, an empirical Bayesian approach to nonparametric modelling is certainly not without precedent [e.g. 29, 37]. Serra and Krivobokova [32] devised an empirical Bayesian method for determining both the smoothing parameter and penalty order in spline fitting; our strategy could be viewed as a crude, heuristic approximation of such a method.

4.1 Gaussian covariate densities, linear regression model

We begin with the "classical" structure from Section 1, where the individual-level measurements within groups are Gaussian deviations from a latent group-level covariate, itself Gaussian:

$$\xi_i \sim \mathcal{N}(0, \sigma_\xi), \tag{20}$$

$$X_{ij} \sim \mathcal{N}(\xi_i, \sigma_X). \tag{21}$$

The regression model is also linear:

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\xi_i + \epsilon_i, \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i, \end{aligned} \tag{22}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_Y). \tag{23}$$

Note that the second line explicitly restates the regression model in the form of (8), with the regression function $\beta^*(x) = \tilde{\beta}x$ ($\tilde{\beta} \in \mathbb{R}$ throughout Section 4). The true parameter values *before* standardizing the data as described in Section 3.2[4] are $\sigma_\xi = 2$, $\sigma_X = 3$, $\alpha = 0.3$, $\tilde{\beta} = 0.4$, and $\sigma_Y = 0.5$. The result is a data set with moderate amounts of noise in both the regression and the covariate measurements. The number of groups is $N = 275$ and each group contains covariate samples for $n = 20$ individuals.

Upon inspecting the data as recommended in the introduction to this section (not shown), we find that an assumption of roughly Gaussian density shape (corresponding to $r = 3$) is reasonable for these data. Because the densities are moderately wide but relatively close together (as the between-density variability σ_ξ, is somewhat smaller than the within-density variability, σ_X), a modest basis of size $K = 10$ should suffice without substantial loss of information. For this simulated data we have $\min_{i,j} X_{ij} = -13.54922$ $\max_{i,j} X_{ij} = 10.87845$, so we extend this range slightly by the same amount in each direction to arrive at

[4]Throughout this section, all parameter values and results will be presented on the original (unstandardized) scale of the given data. The standardization only occurs "internally", during the fitting of the FRODO model.

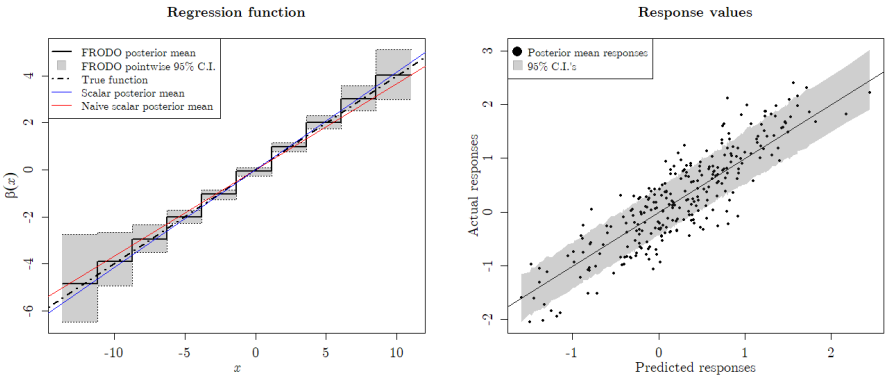

Figure 1: Results of FRODO applied to Gaussian covariate data with a linear regression response. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a "naive" Bayesian scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

an assumed density domain[5] of $[a', b'] = [-13.67077, 11]$. Finally, the default choice of $\delta_i = 0.1$ for all i recommended in Section 3.2 is used here.

As stated at the beginning of this section, we compare FRODO to two simpler models. The first is simply a standard Bayesian linear regression, with (20) omitted and the group-level sample covariate means $\bar{X}_i$ treated as the "true" covariates. The second is a scalar micro-macro Bayesian regression, implemented in the "obvious" way: namely, (20)–(23) are assumed to be the known form of the model, with all parameters (including the latent ξ_i's) unknown and inferred. Recall that the estimate of $\tilde{\beta}$ from the "naive" model will be smaller in magnitude than the "true" value, which the hierarchical scalar model will presumably recover more effectively.

Figure 1 shows results for the regression part of the model. In the left plot, the stepwise estimator of the regression function is shown with its pointwise 95% credible interval. Superimposed on the plot are the true regression function, as well as the posterior means from the hierarchical and naive scalar models (both of which assume a known linear form for the regression, unlike FRODO, which only controls adherence to a linear regression through τ_β). Because the within-group variability is not too much larger than the across-group variability and the sample sizes are reasonable, only a small amount of attenuation is caused by using the naive model, so the estimated regression functions for both scalar models are entirely within the pointwise CI from FRODO. However, the "slope"

[5]Henceforth, the "assumed domain" will be stated on the unstandardized scale of the original data (i.e. $[a', b']$), with the standardization to $[a, b]$ left unstated.

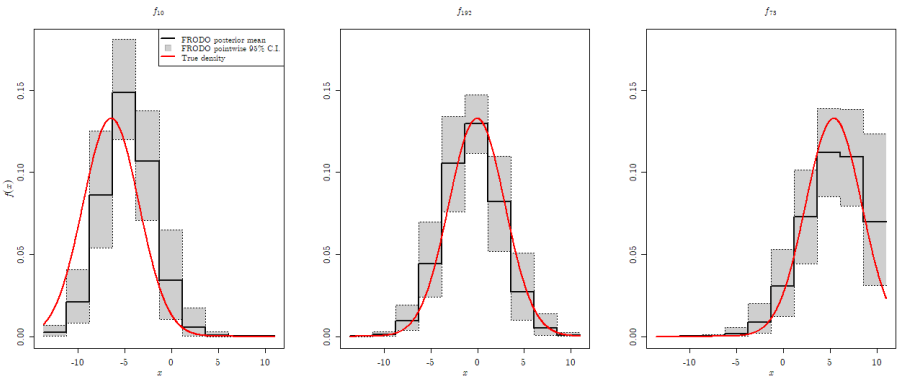

Figure 2: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

of the mean regression function from FRODO seems to be closer to those of the true function and the hierarchical scalar estimate, rather than that of the naive estimate.

Another way to assess FRODO's ability to infer the "true" regression (rather than the incorrect one implied by the naive model) is by checking the posterior for the regression error scale, σ_Y. Because of the additional noise in the individual-level covariate measurements, the naive model's estimate for σ_Y will be biased upward [e.g. 8, Section 3.2.1]. Indeed, the posterior mean for this parameter from the naive scalar model is 0.5559 (95% CI $(0.5104, 0.6012)$), while the posterior means from FRODO and the hierarchical scalar model are 0.4944 (95% CI $(0.4417, 0.5505)$) and 0.4901 (95% CI $(0.4363, 0.5494)$), respectively. Because the FRODO estimate is much closer to the true value of 0.5 than it is to the "naive estimate", we are satisfied that we have avoided the attenuation problem inherent in the naive model. The right plot of Figure 1 shows the posterior means and 95% CI's for the predicted responses $\mathbb{E}[Y_i \mid \alpha, \beta, f_i]$ alongside the observed responses.

Figure 2 shows the estimated f_i's, along with their pointwise 95% CIs, for the group with the smallest (left) and largest (right) ξ_i's, as well as the group whose ξ_i is closest to the sample mean (middle). The middle and right fits are satisfactory, with the inference effectively capturing the true covariate densities (shown in red). The left plot shows that there is something of a mismatch between the inferred and true densities for the group with the lowest ξ_i, with the former shifted slightly too far to the right. Given that the model appears to perform well in all other respects, this is not a significant concern. We did not observe this problem in other datasets generated with the same parameter values (not shown), and therefore assume it is simply an unfortunate quirk of this particular data.

4.2 Gaussian covariate densities, nonlinear regression model

Here, we test FRODO's ability to handle nonlinear regression functions. The covariates adhere to the same Gaussian structure as in Section 4.1, but the regression model is now quadratic:

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\left(\xi_i^2 + \sigma_X^2\right) + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X^2\right] + \epsilon_i. \end{aligned}$$

Because the true covariate densities all have common variance σ_X^2, the difference between $\mathbb{E}_i\left[X^2\right]$ and $\left(\mathbb{E}_i\left[X\right]\right)^2$ is constant and can therefore be absorbed into the intercept. Here, the regression function is $\beta^*(x) = \tilde{\beta}x^2$.

The same parameter values $\left(\sigma_\xi, \sigma_X, \alpha, \tilde{\beta}, \sigma_Y\right) = (2, 3, 0.3, 0.4, 0.5)$ and number of groups $N = 275$ are used as in Section 4.1 although the data is not strictly the same as we used a different seed for pseudorandom number generation in this study. Because the values of ξ^2 span a wider interval than those of ξ, the "relative" level of regression error is lower than in Section 4.1, since the "signal" is larger in scale than the "noise".

As before, we compare FRODO to two scalar models, one hierarchical and one naive. Here, however, it is assumed known in the hierarchical model that the regression is quadratic in the latent covariates ξ, with no linear term. The naive scalar model here is a GAM rather than a linear model, with the covariates taken to be the group-level sample means and the unknown regression function modelled as a cubic P-spline with second-order penalty.

Because the regression function is not one-to-one, an interesting difficulty arises in this framework when the group sizes n_i are too small. On the regression side, the distributions are unchanged if ξ_i is replaced with $-\xi_i$ in a given group. When n_i is small and the true ξ_i is close to zero, the available measurements X_i may not be informative enough to distinguish between these possibilities[6]. This creates multimodality in the posterior (for the hierarchical scalar model, and for FRODO to a somewhat lesser extent) with all of its associated difficulties, including poor HMC sampling behaviour and posterior mean estimates that are not particularly meaningful. Thus, larger group sizes are required if one wants meaningful inference on the covariate parameters as well as the regression parameters. Here, we increase the group size in the simulated data from the $n = 20$ used in Section 4.1 to $n = 50$ for all i. The X_{ij}'s range from -13.76074 to 14.0043, and we expand this range by a small amount in each direction for an assumed density domain of $[-13.80644, 14.05]$. As before, we find $K = 10$ and $\delta_i = 0.1\ \forall i$ to be suitable choices here.

Results for the regression part of the model are shown in Figure 3. At first glance, it may appear as though the FRODO estimate of the regression function is too attenuated, as it is closer to the estimate from the naive scalar model at

[6]This appears to also depend on the amount of covariate variability withing the group relative to the size of its regression error, although it is not currently clear exactly how this dependence works.

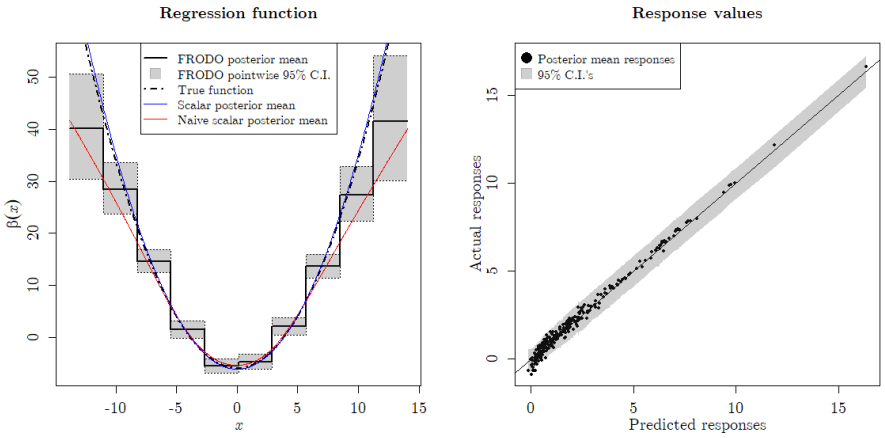

Figure 3: Results of FRODO applied to Gaussian covariate data with a quadratic regression response. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

the endpoints than it is to the true function and the hierarhical scalar estimate. Note, however, that over 95% of the X_{ij}'s lie within the middle six bins, and over 95% of the true latent ξ_i's within the middle four. In those regions, the FRODO estimate is quite close to the true quadratic regression function. Towards the endpoints where the X_{ij}'s are very sparse, there is much less information with which to estimate value of the regression function. This is readily seen in several examples in this manuscript by observing that the pointwise CI's for β are wider in regions with few covariate estimates. In the linear example of Section 4.1, this did not create noticeable bias in the actual posterior mean for β near the endpoints. Presumably this is because — in somewhat informal terms — the covariates in the middle of the domain were sufficiently informative to constrain the posterior for β to a linear shape with high probability, which results in the smoothing parameter τ_β being small with high probability, which, in turn, enforces a linear shape in β with fairly high probability throughout the rest of the domain. In this example, we do not penalize β towards a quadratic shape — only away from a linear shape. As such, it is not surprising that the posterior for β is biased away from the truth near the endpoints, as neither the prior nor the likelihood are very informative there. In principle, one could specify a third-order random walk prior for β in order to ensure a more genuinely quadratic shape, provided one had sufficient reason *a priori* to assume this was an appropriate choice. However, we argue that the second-order random walk prior used here is more intuitive, as it is formulated in terms of deviations from a linear model. At any rate, the heightened bias and uncertainty in the FRODO regression function near the endpoints does not create any seriously adverse consequences for the rest of the inference. In particular, the FRODO posterior mean for σ_Y is 0.4715 (95% CI $(0.3848, 0.6662)$), much closer to the true value of 0.5 than the estimate from the naive model (0.8848, 95% CI $(0.8150, 0.9620)$), suggesting that FRODO is successfully recovering the true regression model and not the biased naive version. The plot of estimate vs. true responses on the right of Figure 3 shows an overall good fit, although there is a small amount of bias in the estimates of the lowest responses.

Figure 4 shows a sample of covariate densities, once again for the group with the smallest and largest ξ_i's, and the ξ_i closest to the sample mean. With larger group sizes, FRODO successfully approximates the true densities for each group shown here.

4.3 Exponential covariate densities, linear regression model

Although it is useful to model arbitrary regression functions, doing so with Gaussian covariate distributions is a capability shared by many methods. In fact, authors such as Sarkar et al. [30] have developed Bayesian methods which allow for even more general structures of the form $X_{ij} = \xi_i + \nu_{ij}$. The true advantage of FRODO lies in its ability to handle covariates that are not based on any kind of *additive* error structure. To demonstrate this, here we use an

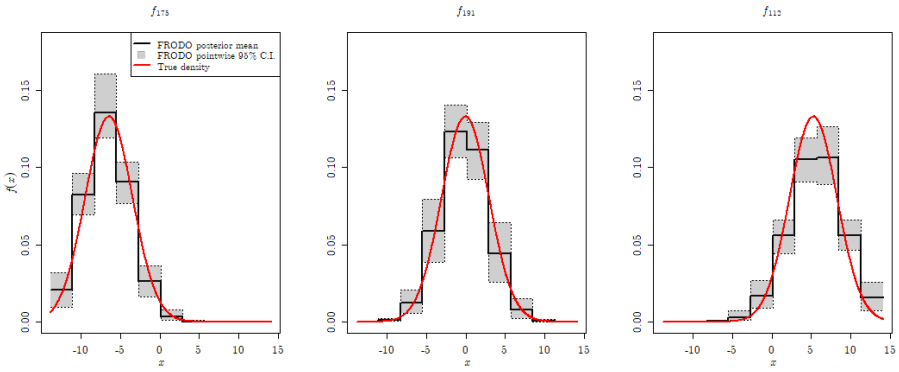

Figure 4: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

exponential covariate structure:

$$\begin{aligned} \lambda_i &\sim \text{Gamma}\,(10, 10) \\ X_{ij} &\sim \text{Exponential}\,(\lambda_i)\,; \end{aligned}$$

and a linear regression model

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\lambda_i^{-1} + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i, \end{aligned}$$

where we have not restated the distribution for the error variance since it is identical to (23) for all subsequent studies.

It is worth contrasting this framework with that of Section 4.1. There, the true covariate densities were Gaussians with equal variances, so the group-level responses depended on their *locations*. With exponential covariate distributions, the linear regression model implies responses that instead vary with the *scales* of the densities. This turns out to be a somewhat challenging type of model for FRODO, due to its treatment of β and the f_i's as piecewise constant functions on bins of equal width. When the true densities are exponential, for any group i it is highly probable that most of the X_{ij}'s will be near 0, with a few very large measurements in the groups with small rates λ_i. If the dimensionality (equivalently, the number of bins) K is taken too small, then the groups with large rates will all have estimated f_i's with probability mass near one in the first bin, and mass near zero in the rest. Thus, it is necessary to use a fairly large K in order to capture the differences between these densities. However, this introduces an opposing challenge due to the sparsity of large X_{ij}'s: near the right end of the domain, many of the bins will not contain any covariate

measurements, so there is little information with which to estimate the densities — and therefore, the regression function — in that region. In summary, when the density scales differ to this extent, the "resolution" of the data varies throughout the domain.

The use of unequal-width bins would perhaps mitigate this problem, but recall from Section 2.2 that the P-spline constructions used here are predicated on an assumption of equally-spaced "knots" (which, with splines of degree zero, are simply the bin endpoints). Without these, the unaltered finite-difference penalties on the coefficients no longer serve as approximations to derivatives of a suitable order. It then becomes nontrivial to penalize the f_i's towards some predetermined "smooth" shape, although Li and Cao [22] proposed a method of modifying the P-spline penalty in the presence of uneven knots. We do not pursue this here, acknowledging that FRODO in its current state has slightly more difficulty using scale information in the covariate densities than it does using location or shape information.

For this dataset ($N = 200$ groups, each of size $n = 50$), we use $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.1, -0.9, 0.1)$. A preliminary visual inspection of KDE's or histograms of the covariate data (and the observation that they are all strictly positive and highly concentrated near zero) justifies a random walk prior of order $r = 2$ on the densities. In order to capture the "high-resolution" differences between covariate measurements near zero as described above, we use a moderately large basis of size $K = 20$. With no reason to suspect severe deviations from this shape we once again set $\delta_i = 0.1$ for all groups. The observed covariates range from 1.3232×10^{-4} to 16.3810. Zero is a natural choice for the left endpoint of the assumed domain, and because there are so few large values, we simply take the right endpoint to be the overall sample maximum 16.3810.

The regression results in the left plot of Figure 5 represent the most significant example of the phenomenon discussed in Section 4.2; namely, the heightened uncertainty in the regression function in regions where covariate measurements are sparse. Here, 99.73% of the observed X_{ij}'s lie in the left half of the domain, while all of the latent λ_i^{-1}'s lie within the first 3 bins. Thus, the pointwise 95% credible interval for β is quite narrow near zero — where most of the covariates are concentrated — and becomes significantly wider moving from left to right. Once again, we compare FRODO to two scalar models: a naive linear regression using the of the $\bar{X}_i$'s as fixed covariates, and a hierarchical linear model in which the latent λ_i's are jointly inferred with the regression parameters. Once again, the estimated regression function from the hierarchical model is very close to the true function, and the FRODO estimate approximates it quite well. Some attenuation bias occurs in the right half of the domain, but because all of the covariate densities have such small mass in this region, this does not seem to adversely affect the regression inference in any other significant way. Indeed, the right plot of Figure 5 shows that the predicted responses closely align with the true Y_i's.

As in previous studies, we compare inferred and true covariate densities for multiple groups in Figure 6. FRODO appears to do a good job of capturing

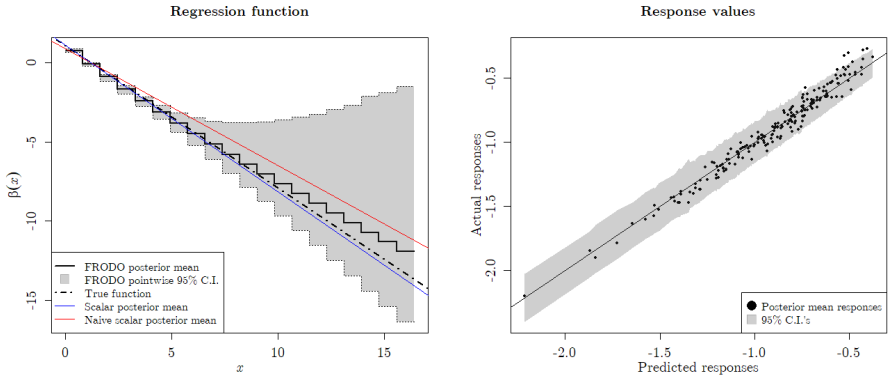

Figure 5: Results of FRODO applied to exponential covariate data with a linear regression response. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a "naive" Bayesian scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

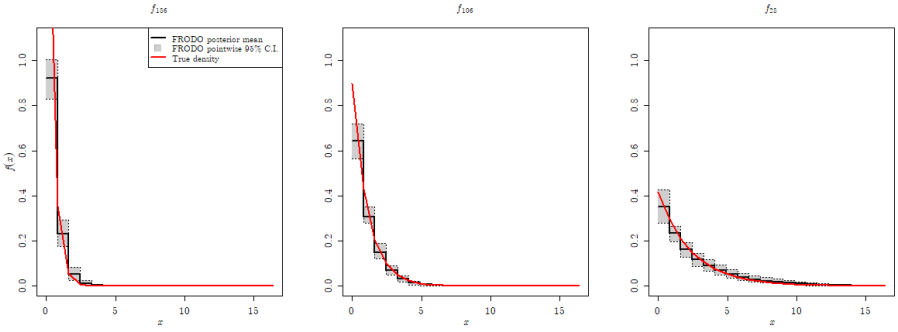

Figure 6: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

the true densities for small, moderate, and large λ_i's, although with no real deviations from the shape imposed by the random walk prior, this is perhaps not surprising.

4.4 Beta covariate densities, linear regression model

In the following two sections, we demonstrate FRODO's ability to capture regression relationships that are encapsulated in the shapes of the covariate densities, rather than their locations or scales. Whereas the covariate densities in preceding examples were governed by group-level latent parameters which were random themselves, here those parameters are deterministic, allowing us to better control the range of shapes we see. In particular, for this section we take $\xi = (\xi_1, \ldots, \xi_N)$ to be a mesh of equally-spaced points from 1/10 to 9/10, and

$$X_{ij} \sim \text{Beta}\left(\xi_i, 1 - \xi_i\right).$$

The regression model is

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\xi_i + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i. \end{aligned}$$

The true densities f_i^* are bimodal for all i, with peaks at 0 and 1 and minima at 1/2. For small i with $\xi_i < 1 - \xi_i$, the peak on the left is wider than the one on the right, so f_i is skewed towards 0 and $\mathbb{E}_i[X] < 1/2$. The opposite is true for large i, and for i near $N/2$ the densities are roughly symmetric.

For this simulation, we use $N = 250$ groups. Because the beta densities have relatively low variance (for the parameter values used here, all of them have variance below 1/8), we use relatively small groups of size $n_i = 15$ for all i, so that the difference between the "true" and "naive" regression functions is more pronounced[7]. The true regression parameters are $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.2, 1, 0.05)$.

Upon inspection of the available covariate data, would see that all covariate measurements are constrained to the unit interval, with the minimum and maximum measurements being extremely close to 0 and 1, respectively. Thus, $[a', b'] = [0, 1]$ is a sensible choice for the assumed domain. Quick visual assessment of KDE or histogram estimates for the group-specific covariate densities reveals that they are neither Gaussian nor exponential. This observation, combined with the strong evidence that the densities are supported only on the unit interval, may lead one to believe that the covariates within each group are, indeed, roughly Beta-distributed. This justifies a random walk prior of order $r = 1$ on the densities, for which the limiting shape is a uniform distribution. Note, however, that unlike the examples above for which we used second- and third-order random walk priors, here the limiting behaviour is *unique*, in the sense that there is only one uniform density on the chosen domain. Thus, if all groups

[7] With large groups, the "naive" regression with group-level covariate sample means would be quite close to the true model, making it difficult to tell which one FRODO was capturing.

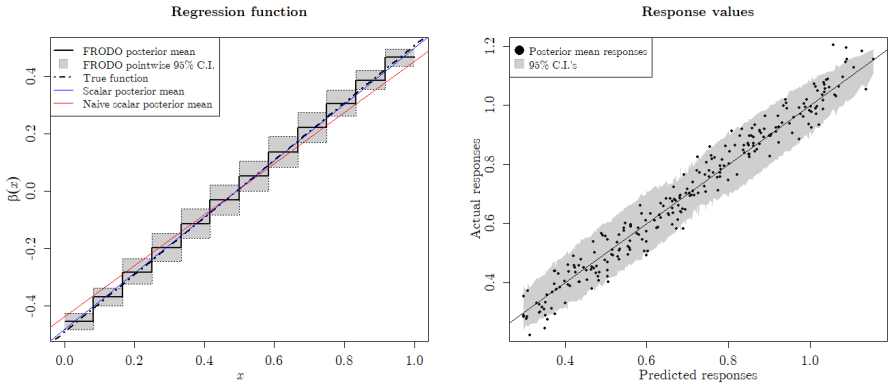

Figure 7: Results of FRODO applied to beta-distributed covariate data with a linear regression response. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a "naive" Bayesian scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

had small smoothing parameter scales δ_i (corresponding to a prior assumption that no severe deviations from the limiting shape occurred), the FRODO estimates of the covariate densities all would be nearly identical, thereby suppressing the differences between groups and compromising the model's ability to extract meaningful regression information. With an assumed first-order random walk prior, one should therefore expect that the covariate densities will exhibit larger deviations from the limiting shape than they would in a situation where $r > 1$ was appropriate (especially since a bimodal shape will be apparent for at least some of the groups upon preliminary visual inspection). Thus, rather than the default $\delta_i = 0.1$ used in previous examples, here we take $\delta_i = 1$ for all groups. Finally, since several groups have most of their covariate measurements near the endpoints (necessitating bins which are narrow enough to capture differences in densities within these regions), we use $K = 12$ bins: more than the 10 used in the Gaussian examples, but less than the 20 used in Section 4.3 since we do not have enough covariate measurements per group to support such a large number of bins (especially since "roughness", or deviation from the random walk shape, is penalized less severely here).

Once again, the regression component of the model is visualized in Figure 7, alongside posterior mean estimates from naive and hierarchical scalar models In contrast to previous datasets, here there are more covariate measurements at each endpoint of the domain than there are in the middle, leading to a slight "bulge" in the pointwise 95% credible interval around 0.5. However, each bin is relatively well-populated with observations, compared to the large

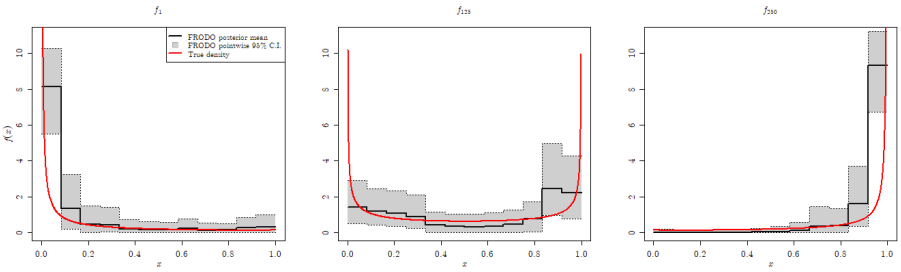

Figure 8: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

differences in concentration seen in previous examples. It is visually obvious that FRODO captures the true regression function and not the naive one. The plot of predicted vs. true responses on the right of Figure 7 provides further confirmation that FRODO's regression inference is satisfactory here.

Figure 8 shows that FRODO has more difficulty inferring the true densities here than for previous examples. Although the asymmetrical shapes for ξ_i's near 0.1 or 0.9 are captured, the steep curvature of the true densities near the endpoints in these cases results in them being near the edges of the model's pointwise 95% credible intervals — if not excluded altogether — in these regions. From the middle plot, we see that the model imposes a somewhat excessive degree of uniformity on the nearly-symmetric densities for which ξ_i is near 0.5. These difficulties are not surprising: given the small group sizes and the fairly large values used for K and the δ_i's, neither the prior nor the likelihood make very strong implications about the density shapes. Aside from collecting more covariate measurements for each group (i.e. strengthening the likelihood), the only other possible mitigation for this would be to strengthen the prior: either by using smaller δ_i's to more strictly enforce the uniform shape, or by using a smaller K to reduce the dimensionality of the problem. However, as discussed above, both of these options would result in an obfuscation of any information that does exist in the available covariate data. Thus, the most prudent choice seems to be accepting that FRODO's density inference in this example is necessarily limited to some degree. Fortunately, this limitation does not adversely affect any of the inference on the regression side of the model. Furthermore, despite the relative "roughness" of the FRODO density estimates[8], they are certainly improvements over, say, "raw" histograms (corresponding to $\delta_i \to \infty$), for which the low amount of covariate data would result in even less interpretable shapes.

[8]Note that this is an inherent difficulty in any dataset for which the first-order random walk prior is justified, because imposing smoothness in this case is inseparable from forcing all of the densities towards being identical.

4.5 Beta covariate densities, nonlinear regression model

Although the previous example shows that FRODO can extract relationships based on the shapes of covariate densities, the regression model itself still ultimately depended only on the means of the covariate measurements. The nonadditive structure of the X_{ij}'s would pose a challenge for many established multilevel methods, but it is conceivable that one could devise a nonparametric, hierarcical Bayesian method which jointly inferred the $\mathbb{E}_i[X]$'s while using them to recover the correct regression parameters, subverting the need for full functional regression on the densities. When the regression is not lienar, this may not be the case. Thus, in this section we combine a nonadditive covariate structure with a nonlinear regression model to demonstrate the full generality of FRODO. Once again ξ is a mesh of equally-spaced points, this time from 1/10 to 2, and

$$\begin{aligned} X_{ij} &\sim \text{Beta}(\xi_i, \xi_i), \\ Y_i &= \alpha + \tilde{\beta}\left(1 + \frac{1}{2\xi_i + 1}\right) + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[4\tilde{\beta}\left(X - \frac{1}{2}\right)^2\right] + \epsilon_i. \end{aligned} \tag{24}$$

Here, the regression function is $\beta^*(x) = 4\tilde{\beta}(x - 1/2)^2$. The f_i^*'s are all symmetric: bimodal and U-shaped for i near 1, roughly uniform for i near $N/2$, and peaked at 1/2 for i near N. For positive $\tilde{\beta}$, the expected response $\mathbb{E}_i[Y]$ is higher for "more bimodal" covariate densities and lower for "more unimodal" ones. The regression is therefore entirely dependent on the shapes of the densities, not their locations or scales. Furthermore, because the densities are all symmetric it holds that $\mathbb{E}_i^*[X] = 1/2$ for all i. Thus, any modelling approach targeting $\beta(\mathbb{E}_i[X])$ ("regression on the expectaton") will be unsuitable here[9], as opposed to FRODO with its use of "the expectation of the regression", $\mathbb{E}_i[\beta(X)]$. In every aspect, this particular data structure is decidedly "non-classical", and FRODO seems uniquely well-suited to handle such a structure.

Because the true covariate densities all have expectation equal to 1/2, the regression function is actually not unique: indeed, when the f_i^*'s are all symmetric Beta densities, (24) is equivalent to $\alpha + \mathbb{E}\left[4\tilde{\beta}X^2\right] + \epsilon_i$, up to a term which is constant with respect to i. This does not seem to be a problem in practice, however: even when HMC chains are explicitly initialized such that β is close to the latter form, they converge to a posterior which is consistent with (24). We conjecture that the FRODO posterior concentrates around the form of the regression function with "lowest error": empirically, we observed that the

[9] In theory, one could invoke a measurement error method with more general assumptions on the covariate structure. Recall from 1 that the frequentist approach of Hu and Schennach [18] described in Section 1 assumed a general functional mapping the f_i^*'s to the ξ_i's. Although higher-order moments should be permissible under their assumptions, the authors required a *known* functional. Thus, even with their level of generality it would still be necessary to assume quadratic regression *a priori*.

within-group sample means of $(X_{ij} - 1/2)^2$ values provide much more accurate estimates of their population analogues than the within-group sample means of the X_{ij}^2's.

For this example, we simulated a dataset with $N = 250$ groups, each containing $n = 60$ covariate measurements. The true regression parameters were $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.7, 1, 0.1)$. As in Section 4.4, the observed range of the covariate measurements provides strong evidence that $[0, 1]$ is a good choice for the assumed density domain. Here, the range of shapes in preliminary histograms or KDE's (from bimodal, to roughly uniform, to unimodal) gives further justification for a random walk prior of order $r = 1$. As in the previous section, we take $\delta_i = 1$ for all i to allow a greater degree of deviation from the limiting (uniform) shape of the prior. Because the data is highly concentrated near the endpoints for the groups whose ξ-values are low (even moreso than in Section 4.4's dataset), we use a basis of size $K = 15$.

Due to the aforementioned uselessness of methods involving "regression on expectations" here, constructing scalar models to compare with FRODO is nontrivial. We cannot use a "naive GAM" as we did for the Gaussian quadratic model in Section 4.2. There, $\mathbb{E}_i\left[X^2\right]$ and $\left(\mathbb{E}_i\left[X\right]\right)^2$ differed by a constant, but this is not the case here. Thus, the naive scalar model we use for comparisons is somewhat contrived: a linear regression model, using the within-group sample means of the $(X_{ij} - 1/2)^2$ values as covariates. As always, the hierarchical scalar model assumes the true forms of the regression function and covariate densities are all known, jointly inferring the ξ_i's and all regression parameters.

Because of the relatively large group sizes, and the fact that the quadratic form of the regression function was assumed known in both scalar models, the naive model does not suffer from any appreciable attenuation bias. As shown on the left of Figure 9, both it and the hierarchical scalar model approximate the true regression function almost perfectly. Some bias is apparent in the FRODO estimate, particularly near the vertex at $1/2$, but its pointwise 95% credible interval almost completely captures the true function. On the right side of Figure 9, we see a moderate "clumping" of predicted responses just over 2.0, where the variability in the actual Y_i's exceeds that of the mean predictions from FRODO. These values correspond to groups with ξ-values near 1 (i.e. those whose true covariate densities f_i^* are close to uniform). For this dataset, it appears that FRODO has a small amount of difficulty capturing small shape differences between nearly-uniform densities. Nevertheless, the overall fit appears largely satisfactory, especially considering that the true forms of the regression function and covariate densities may not be known *a priori*.

Figure 10 shows that FRODO roughly captures all three types of density shapes present in this data, although some excess noise and bias is evident in the posterior estimates. This is particularly evident for the unimodal density in the right plot. Although the true density is fully contained in the pointwise 95% credible interval, the posterior mean is perhaps somewhat too flat. The true unimodal densities in this dataset certainly differ more subtly from the uniform shape than the bimodal ones (contrast the true density in the left plot of Figure

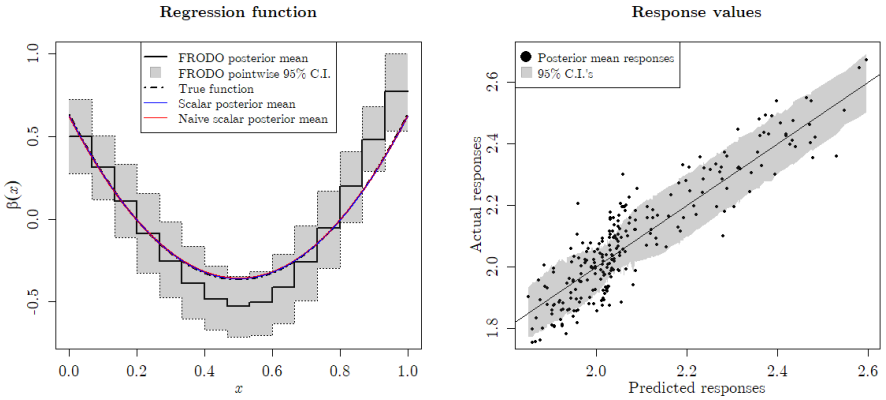

Figure 9: Results of FRODO applied to beta-distributed covariate data with a quadratic regression response. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

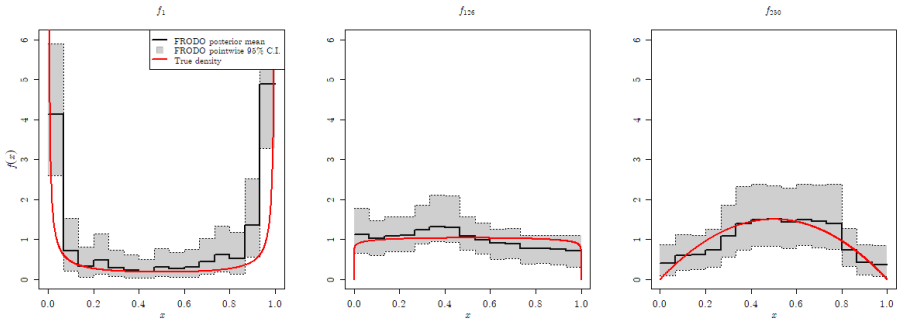

Figure 10: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

10 with that on the right) — since the prior on densities here is structured only in terms of "deviations from uniformity", this slight deficiency is not entirely unexpected. As in Section 4.4, some of the excess noise in the density inference is an unavoidable consequence of the larger values of K and δ necessary to capture the shapes and fine structure of the true densities with the first-order prior.

5 Extended simulation study: FRODO with varying group sizes and a group-level covariate

As a final "application" of FRODO, we recreate the simulated data considered by Croon and van Veldhoven [9]. This is very much a "classical" model, with Gaussian covariate data and a linear regression function much like the one considered in Section 4.1. However, there are three unique features here which were absent from the "toy" examples explored above. First (recalling the notation of (20–23)), the parameter values are $\left(\sigma_\xi, \sigma_X, \alpha, \tilde{\beta}, \sigma_Y\right) = \left(1, 3, 0.3, 0.3, \sqrt{0.35}\right)$: not only is the within-group variability of the X_{ij}'s much greater than the between-group variability of the true ξ_i's, but the regression error is also quite high, accounting for just under 65% of the variability in the Y_i's. Overall, the amount of "signal" in the data — at both the covariate and regression levels — is low relative to the amount of noise. Second, there are varying group sizes, some of which are quite small: out of $N = 100$ groups, roughly 50% (randomly selected with probability 1/2) contain $n_i = 10$ covariate measurements, and the rest contain $n_i = 40$. Finally, the actual regression model is altered from the basic FRODO form considered thus far, with the inclusion of a "scalar" group-level covariate Z:

$$Y_i = \alpha + \tilde{\beta}\xi_i + \beta_Z Z_i + \epsilon_i. \tag{25}$$

The covariate values Z_i are generated from a standard Normal distribution, independently of ξ, and are treated as fixed observations.

It is straightforward to extend FRODO to accommodate for Z by specifying a $\mathcal{N}(0, 20\sigma_Y)$ prior on β_Z, conditionally independent from the prior for β (which still denotes the regression function corresponding to the group-specific densities of the X_{ij}'s). We use a third-order random walk prior on the f_i's with $K = 10$ bins as in Section 4.1, since the available data gives no reason to suspect that finer structures need to be captured. Due to the relatively small amount of covariate measurements, we simply take the assumed domain $[a', b']$ to be the range of observed X_{ij}-values, which in this case is $[-12.0365, 11.2258]$. For the groups of size $n_i = 40$, the default smoothing prior scale choice δ_i is appropriate, but with only $n_i = 10$ observations in the smaller groups, a tighter prior is necessary to ensure posterior density estimates with useful shape information. Thus, we set $\delta_i = 0.05$ for the small groups.

The actual method proposed by Croon and van Veldhoven [9] for micro-macro modelling is frequentist and involves a stepwise estimation procedure.

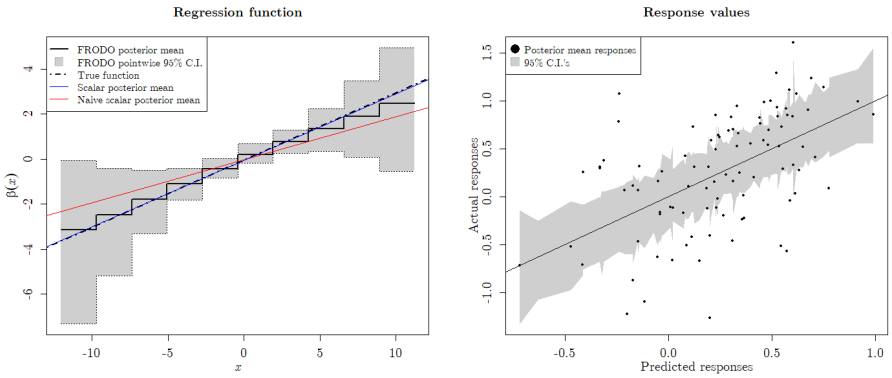

Figure 11: Results of FRODO applied to Gaussian covariate data with a linear regression response and an additional group-level scalar covariate. Left: the regression function for the multilevel covariate estimated by FRODO, alongside its pointwise 95% credible region, the true function, the posterior mean estimate from a "naive" Bayesian scalar linear regression, and the posterior mean estimate from a scalar version of the true hierarchical model. Right: responses $\hat{Y}_i$ predicted by FRODO (alongside their 95% credible intervals) vs. the true response values.

An R implementation exists [23], but here we are only interested in comparing FRODO to analogous scalar Bayesian methods. Thus, as in the studies of Section 4 we compare it to both a naive and hierarchical scalar model, trivially extended to accommodate Z and place a prior on β_Z. These results are shown in the left plot of 11. Note how much wider the pointwise 95% credible interval is — particularly near the endpoints — than the one in the similar model of Figure 1, owing to the higher noise and smaller amount of available covariate data here. It appears that the posterior for FRODO has concentrated somewhere in between the true and naive regressions. Indeed, FRODO's posterior mean for σ_Y is 0.5975 (95% C.I. $(0.5152, 0.6945)$), in contrast with 0.5856 from the hierarchical scalar model (95% C.I. $(0.5014, 0.6835)$) and 0.6128 from the naive scalar model (95% C.I. $(0.5332, 0.7029)$). Given that the dataset is fairly small and high in noise, it is perhaps unsurprising that FRODO struggles more than it did in previous studies. However, this seems to be a problem of variability, not of bias: other simulated datasets with the exact same parameters, group sizes, and number of groups resulted in FRODO estimates with differing amounts of attenuation (not shown). Even the scalar hierarchical model proved quite variable with other datasets, as its estimate of the regression function did not always align as closely with the true function as it does here. Although the high degree of noise in the right plot of Figure 11 appears troubling, this is reflective of the actual amount of noise in the data: a plot of predicted vs. actual responses from a frequentist multiple linear regression using the true ξ_i's appears similar.

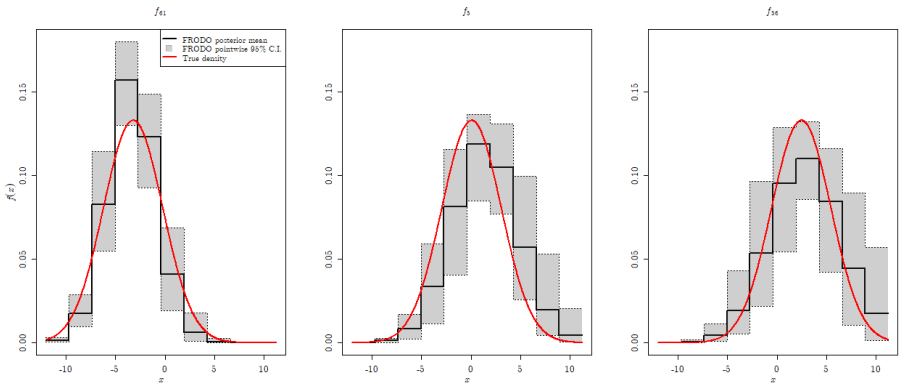

Figure 12: For a selection of groups, the FRODO estimate of the group-specific covariate density, alongside their pointwise 95% credible regions. The true densities are superimposed as red lines.

The usual density plots are shown in Figure 12. Note that the group in the left plot contains 40 individuals, and the other two contain only 10. It is intuitive that the smaller groups would have wider pointwise credible intervals for their densities (on further inspection, this pattern also seemed to hold for other groups not shown here), although it is somewhat noteworthy that the smaller δ_i-values for these groups do not seem to neutralize this effect. Some bias in the model is evident, particularly in the middle plot, but overall the inference provided by FRODO seems reasonable.

6 Discussion and future work

In this chapter, we have presented a new approach for micro-macco modelling which combines density estimation and functional data analysis into a unified hierarchical Bayesian framework. Although FRODO is relatively simple in principle due to its use of step functions and only *linear* functional regression terms, it is deceptively powerful in its ability to use these elements for approximation of generalized additive models. Beyond the generality of the regression component of the model, FRODO is also quite flexible in terms of the individual-level covariate structures it can accommodate. Whereas many Bayesian methods for GAM's with measurement error or micro-macro structure assume a Gaussian — or at the very least, additive — error structure in the X_{ij}'s, FRODO has no such limitation, allowing for covariate densities which influence the group-level regression responses through their locations, scales, or shapes. All that is required is the selection of a suitable prior structure for the densities, based on either prior domain knowledge, or — if this is not possible and an empirical Bayesian approach is required — a preliminary heuristic examination of the data. Although FRODO's inference on the covariate densities is generally

more accurate when the true densities adhere to the specified "smooth shape" encoded in the prior, this is not a strict *requirement* provided hyperparameters are chosen carefully.

The simulation studies conducted above show that the power and generality of FRODO translate from theory to practice, providing reasonable inference for a variety of data structures. However, the potential for improvements and extensions to the model is vast. The most immediate potential for this is in the density part of the model, as described in Section 3.2. Here we have not considered r^{th}-order random walk priors for any integer $r > 3$. These would result in densities being penalized towards exponentiated polynomials of higher degree: with an r^{th}-order random walk prior, $\log f_i(x)$ is close to a polynomial of degree $r-1$ when the smoothing parameter λ_i is small. Such limiting smooth shapes correspond to *generalized error distributions* [38] (or folded versions thereof) with shape parameter $r-1$, of which the Normal, Laplace, and uniform distributions are special cases. For $r > 2$, the generalized error distribution has lighter tails than a Gaussian. It is not certain how useful such higher-order random walk priors would be in practice (i.e. how often one might expect covariate densities to be similar to, say, an exponentiated quartic), but one challenge in implementing these would be determining suitable distributions for the "free parameters" θ_{ik}, $2 \leq k \leq r$. Equivalent derivations of the type carried out for $r = 2$ and 3 in Section 3.2 would be much more complex.

There is even room for generalization within the confines of the third-order (resp. second-order) random walk priors considered here. Although the construction in Section 3.2 was explicitly tailored in terms of Gaussian (resp. exponential) distributions, in principle it could be adapted for *any* densities whose logarithms are roughly quadratic (resp. linear) in shape. Folded or truncated Normal distributions may be a useful shape to accommodate with a third-order random walk prior; one could even modify it to allow for densities f such that $\log f$ is approximately quadratic with *positive* leading coefficient, not negative as for a Gaussian. This may be useful for modelling "U-shaped" densities, such as the Beta distributions considered in Section 4.4. Similarly, the second-order structure could be generalized to allow for positively-sloped densities (i.e. "reversed" exponentials), or Laplace densities whose logarithms are *piecewise* linear. Furthermore, it may be useful to combine differing random walk orders within the same model. For instance, the example in Section 4.5 might have benefited if we used a third-order random walk prior for the unimodal densities (since symmetric Beta densities are close to Gaussians in shape for large parameter values), a first-order R.W. prior for the flatter densities, and perhaps an "inverted" third-order R.W. prior for the U-shaped densities as suggested above.

Further investigation of the relationships between n, r, K, and δ would also be useful, particularly how best to set the latter two in terms of the former two. Although the empirical heuristic methods employed here worked well in practice, a more formal approach might result in better performance and generalization. Appeals to asymptotics could guide derivation of mathematical relationships between the hyperparameters: for instance, an expression for an "optimal" δ_i in

terms of r, K, and n_i, based on the "big-O" relationships shown by Silverman [33] to guarantee convergence of penalized density estimators in the frequentist setting. The choice of the assumed domain for the densities may also have an effect on any such expressions.

There is also significant potential for generalizations on the regression side of the model. The most immediate of these is the realization of our proposed extension to non-Gaussian responses such as count or categorical data. Just as the regression part of FRODO for the Gaussian responses considered here is nothing more than a functional linear model, allowing for other response types is simply a matter of using functional GLM machinery.

Perhaps the most useful immediate extension to FRODO would be the incorporation of multiple multilevel covariates. Indeed, many real-world micro-macro datasets include several covariates measured at the individual level within groups [e.g 9, 2, 10]. Of course, this would increase the computational complexity of FRODO, as the number of parameters to infer grows roughly linearly in the number of multilevel covariates. Note, however, that real-world micro-macro datasets commonly include ordinal covariates with a small number of levels [e.g. 2, 10]. Modelling the distributions for these covariates requires only as many basis functions as there are levels, which would mitigate computational difficulty to some extent in practice.

A powerful yet challenging improvement would be modelling more complex relationships amongst covariates. For instance, Croon and van Veldhoven [9] considered a version of the simulation study replicated in Section 5 where the latent and observed group-level covariates (ξ and Z, respectively) were correlated [see also measurement error literature such as 28]. Accounting for dependence between multilevel and "scalar" covariates in FRODO will be highly nontrivial, especially if one wishes to maintain flexibility in the shapes of the inferred densities. For instance, if the multilevel data is Gaussian as in Section 5, the most obvious way to account for correlation between ξ and Z is to explicitly include it in the prior for the ξ_i's (see Section 3.2). However, we have found in practice that the ξ_i's inferred by FRODO are often poor approximations for the actual latent group means of the X_{ij}'s, unless a Gaussian shape is heavily enforced on the f_i's by deliberately taking very small δ_i's. This was not a problem for the examples in Sections 4.1 and 4.2, as the posterior density estimates ended up being close enough to the true Gaussians that there were no major difficulties in the inference. If such latent density parameters are required more explicitly to model correlations with scalar covariates, this inaccuracy may become problematic. The potential for dependence between distinct *multilevel* covariates is arguably even more interesting. Presumably this would require regression on multiple integrals over their joint densities. However, even with the degree-zero splines considered here, this would result in a substantial increase in computational complexity. Indeed, the number of coefficients required to model the joint density of d multilevel covariates for a single group in this way is exponential in d. Therefore, some type of simplification would likely be required to make interactions between multilevel covariates viable. See Lambert and Eilers [19] for a discussion of multivariate density estimation with splines in the case of a

single density.

We conclude by acknowledging potential shortcomings in FRODO for which there are likely no solutions, either due to the inherent properties of the model or the excessive computational difficulty that would be required to solve them. First, one may question the use of piecewise constant basis functions, since higher-order splines would certainly result in smoother and better-behaved density estimates. However, recall from Section 3.3 that this choice was made partially for computational convenience: it ensures that the integral of $\beta \cdot f_i$ is simply the inner product of the two functions' coefficients. This is no longer the case with higher-order splines, for which the integrals are more complicated expressions involving products between neighbouring coefficients. Beyond the heightened complexity, we also found in preliminary experiments that the resulting posterior geometry was extremely difficult to navigate with NUTS. Note that these experiments modelled the densities themselves with higher-degree splines, requiring (among other things) a potentially costly softmax transformation of each θ_i vector. The other possibility is modelling the *logarithms* of the densities with splines [e.g. 24]. These approaches are equivalent for degree-zero splines, but with higher degrees the logarithmic approach requires approximate numerical integration to normalize the f_i's, which are exponentiated piecewise polynomials. These numerical integrals, in turn, depend on the spline coefficients in complex ways which would likely complicate the posterior geometry even further. Thus, unless a radically different approach is used to fit the model, higher-order splines do not seem to be worth the effort, given the satisfactory results obtained with piecewise constant functions and the prevalence of ordinal covariates in real-world micro-macro data.

In earlier experiments (not shown), we found problems with bias and sampling efficiency when the within-group covariate noise was large relative to either the regression noise or between-group covariate scale. In the notation of the Gaussian model, problems occurred when the n_i's were small and σ_X was large relative to either σ_ξ or σ_Y, especially when the magnitude of the effect size $\tilde{\beta}$ was large. This problem also affected hierarchical scalar models — suggesting that there is innate difficulty in the posteriors induced by such datasets — but FRODO did seem slightly more sensitive to it, in the sense that some parameter combinations were problematic for FRODO but not for a scalar model. These problems could be mitigated with different prior choices such as a zero-avoiding prior for σ_Y, but these can create bias [13]. Fortunately, we suspect that the relative noise levels which tend to create problems are unlikely to occur in practice, as they imply either extremely low-error regression models or high-error covariate groups.

Finally, it bears repeating that FRODO only models responses in terms of *expectations of functions of covariates*: any regression relationship that cannot be expressed in the form (8), or some multivariate extension thereof, is incompatible with this methodology. In particular, responses which depend on the medians or modes of densities cannot be modelled with FRODO, requiring other methods specifically suited for those purposes [e.g. 18]. Its current inability to

model *functions of expectations* may also be a shortcoming. For instance, if the data in Section 4.2 was modified so that the covariate densities had unequal variances and the group-level responses were proportional to these variances, FRODO would not be usable due to the nonconstant $(\mathbb{E}_i[X])^2$ term in the regression. One could potentially augment (8) with an "outer function", using terms of the form $g(\mathbb{E}_i[\beta(X)])$ with some unknown function g to be modelled with a basis function expansion. However, this would likely create a litany of problems with unidentifiability.

Despite these challenges, we believe that FRODO's power and flexibility make it a strong addition to the field of micro-macro regression modelling, especially as improvements and extensions are developed to handle an even broader variety of data structures.

References

[1] J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980. ISSN 00063444. URL `http://www.jstor.org/stable/2335470`.

[2] Adalgiso Amendola, Cristian Barra, and Roberto Zotti. Does graduate human capital production increase local economic development? An instrumental variable approach. *Journal of Regional Science*, 60(959-994), 2020. doi: 10.1111/jors.12490.

[3] Margot Bennink, Marcel A. Croon, and Jeroen K. Vermunt. Micro-Macro Multilevel Analysis for Discrete Data: A Latent Variable Approach and an Application on Personal Network Data. *Sociological Methods & Research*, 42(4):431–457, 2013. doi: 10.1177/0049124113500479.

[4] Margot Bennink, Marcel A. Croon, Brigitte Kroon, and Jeroen K. Vermunt. Micro-macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*, 10:139–154, 2016. doi: 10.1007/s11634-016-0234-1.

[5] Michael Betancourt. How the shape of a weakly informative prior affects inferences, 2017. URL `https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes`.

[6] John P. Buonaccorsi. *Measurement error: Models, methods, and applications*. CRC Press, jan 2010. ISBN 9781420066586. doi: 10.1201/9781420066586. URL `https://www-taylorfrancis-com.proxy.lib.sfu.ca/https://www-taylorfrancis-com.proxy.lib.sfu.ca/books/mono/10.1201/9781420066586/measurement-error-john-buonaccorsi`.

[7] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li,

and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[8] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, jun 2006. doi: 10.1201/9781420010138. URL https://www-taylorfrancis-com.proxy.lib.sfu.ca/https://www-taylorfrancis-com.proxy.lib.sfu.ca/books/mono/10.1201/9781420010138/measurement-error-nonlinear-models-raymond-carroll-david-ruppert-leonard-stefanski-cipr

[9] Marcel A. Croon and Marc J.P.M. van Veldhoven. Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, 12(1):45–57, mar 2007.

[10] Oumou Salama Daouda, Mounia N. Hocine, and Laura Temime. Determinants of healthcare worker turnover in intensive care units: A micro-macro multilevel analysis. *Plos One*, 16(5):e0251779, may 2021. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0251779. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251779.

[11] Paul H C Eilers and Brian D Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121, 1996. URL https://projecteuclid.org/download/pdf_1/euclid.ss/1038425655.

[12] Lynn Foster-Johnson and Jeffrey D. Kromrey. Predicting group-level outcome variables: An empirical comparison of analysis strategies. *Behavior Research Methods*, 50(6):2461–2479, mar 2018. ISSN 1554-3528. doi: 10.3758/S13428-018-1025-8. URL https://link.springer.com/article/10.3758/s13428-018-1025-8.

[13] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006. URL http://www.stat.columbia.edu/$\sim$gelman/research/published/taumain.pdf.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

[15] Harvey Goldstein. *Multilevel statistical models*. John Wiley & Sons, fourth edition, 2010. ISBN 9780470973400.

[16] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[17] Cheng Hsiao. Consistent estimation for some nonlinear errors-in-variables models. *Journal of Econometrics*, 41(1):159–185, may 1989. ISSN 03044076. doi: 10.1016/0304-4076(89)90047-X.

[18] Yingyao Hu and Susanne M. Schennach. Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, 76(1):195–216, 2008. URL http://www.econometricsociety.orgor.

[19] Philippe Lambert and Paul HC Eilers. Bayesian multi-dimensional density estimation with p-splines. In *Proceedings of the 21st International Workshop on Statistical Modelling*, pages 313–320, 2006.

[20] Stefan Lang and Andreas Brezger. Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, mar 2004. ISSN 1061-8600. doi: 10.1198/1061860043010. URL http://www.tandfonline.com/doi/abs/10.1198/1061860043010.

[21] Tong Li. Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, 110:1–26, 2002. URL www.elsevier.com/locate/econbase.

[22] Zheyuan Li and Jiguo Cao. General p-splines for non-uniform b-splines, 2022. URL https://arxiv.org/abs/2201.06808.

[23] Jackson G Lu, Elizabeth Page-Gould, and Nancy R Xu. *MicroMacro-Multilevel: Micro-Macro Multilevel Modeling*, 2017. URL https://CRAN.R-project.org/package=MicroMacroMultilevel. R package version 0.4.0.

[24] Finbarr O'Sullivan. Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2):363–379, 1988. ISSN 0196-5204. doi: 10.1137/0909024. URL http://www.siam.org/journals/ojsa.php.

[25] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337. URL https://people.eecs.berkeley.edu/$\sim$jordan/courses/260-spring09/other-readings/park-casella.pdf.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

[27] James O. Ramsay and Bernard W. Siverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2005. ISBN 0-387-40080-X. doi: 10.1007/b98888. URL http://link.springer.com/10.1007/b98888.

[28] Sylvia Richardson and Walter R. Gilks. Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error. *Statistics in Medicine*, 12:1703–1722, 1993.

[29] Judith Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *The Annual Review of Statistics and Its Applications*, 3:211–231, 2016. doi: 10.1146/annurev-statistics-041715-033523. URL www.annualreviews.org.

[30] Abhra Sarkar, Bani K. Mallick, and Raymond J. Carroll. Bayesian Semiparametric Regression in the Presence of Conditionally Heteroscedastic Measurement and Regression Errors. *Biometrics*, 70:823–834, 2014. doi: 10.1111/biom.12197.

[31] Susanne M. Schennach. Recent Advances in the Measurement Error Literature. *Annual Review of Economics*, 8:341–377, oct 2016. ISSN 19411391. doi: 10.1146/annurev-economics-080315-015058. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-080315-015058.

[32] Paulo Serra and Tatyana Krivobokova. Adaptive Empirical Bayesian Smoothing Splines. *Bayesian Analysis*, 12(1):219 – 238, 2017. doi: 10.1214/16-BA997. URL https://doi.org/10.1214/16-BA997.

[33] B.W. Silverman. On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method. *The Annals of Statistics*, 10(3):795–810, 1982. URL https://www-jstor-org.proxy.lib.sfu.ca/stable/pdf/2240905.pdf?refreqid=excelsior%3A973c855a2c118096e73d9457f4facd2c.

[34] Tom AB Snijders and Roel J Bosker. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage, 2011.

[35] Stan Development Team. Stan modeling language users guide and reference manual, version 2.30.0, 2018. URL http://mc-stan.org/.

[36] Stan Development Team. RStan: the R interface to Stan, 2021. URL https://mc-stan.org/. R package version 2.21.3.

[37] Mark A. van de Wiel, Dennis E. Te Beest, and Magnus M. Münch. Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 46:2–25, 2019. doi: 10.1111/sjos.12335. URL https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12335.

[38] Mahesh K Varanasi and Behnaam Aazhang. Parametric generalized gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415, 1989.

[39] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.