Chapter 2

smoothEM: a new approach for the simultaneous assessment of smooth patterns and spikes

2.1 Introduction and Motivation

The past decade has witnessed an increasing interest in functional data, where one or more variables arise from underlying functions and often possess additional structures of interest. The vast majority of past and current literature focuses on functional data obeying certain smoothness conditions [2, 3]. This kind of data can be effectively represented through basis functions (e.g., Fourier basis, spline basis or polynomial basis) and in the majority of applications a penalty is employed to ensure such representation – an approach commonly referred to as regularized smoothing [4–8]. When functions have discontinuities and/or sporadic behaviors, wavelet bases are often preferred due to their ability to yield information about the function at multiple resolution [9]. Somewhere in between, [10] consider functions with globally smooth components and rough components that are smooth at local scales.

In this work, we are interested in a composite structure where data is generated from a smooth curve with additive errors – except at certain locations, where irregular behaviors occur in the form of additive spikes further superimposed to the "noisy curve". In symbols, we are interested in data $(x_i, y_i)_{i=1}^n$ of the form

$$y_i = f(x_i) + \mu^* \cdot \mathbb{1}(x_i \in \mathbb{S}) + \epsilon_i \tag{2.1}$$

where x_i, $i = 1 \ldots, n$ are locations in a domain (which we map in $[0, 1]$ without loss of

generality), $f \in C^m[0,1]$ is a smooth function, $\mathbb{S}$ is a set of locations affected by spikes of size μ^* such that $\mathbb{P}(X \in \mathbb{S}) = \alpha^*$, and ϵ_i, $i = 1 \ldots, n$ are independent random errors. Assuming a fixed design, our main goal is to capture the underlying curve and to identify the spikes, i.e. to estimate f and the classification vector $\mathbb{1}_{\mathbb{S}} = (\mathbb{1}(x_i \in \mathbb{S}))_{i=1}^n$.

This structure is rather common in applications. To provide some meaningful examples, we consider data on extreme temperature indexes (an annual time series from the United States) and electricity consumption (a weekly time series from Ireland). In both these examples occasional spikes occur on top of a smooth underlying trend. We are interested in producing a estimate of such trend and in identifying the spikes, which can be analyzed (possibly using additional assumptions on their distribution) to gain insight into their frequency, location, magnitude and spread.

For this type of data a simple application of regularized smoothing, e.g., with a spline basis, can be misleading; without taking into account the "spikes" traditional smoothing can generate an inaccurate approximation of the underlying curve f in (2.1). [10] considered functions that possess rough variations, but assumed smoothness at a local scale, which does not allow for discontinuity in the regression function $\mathbb{E}(Y_i|X_i = x_i)$. In contrast, we propose a procedure that utilizes regularized smoothing splines, thus preserving assumptions about the smoothness of f, and combines them with an *Expectation-Maximization* (EM) algorithm [11–15] which, under some conditions, allows us to identify the spikes. Because we preserve the smoothness of f, we are able to leverage existing results on penalized smoothing spline estimators [16–22]. The result is a good approximation of f, on par with that one would obtain in the absence of spikes.

With the additional assumption that errors are Gaussian, we rewrite (2.1) as $y_i = f(x_i) + \xi_i$ and model the departure from f as a mixture of Gaussians; namely

$$\xi_i \sim \alpha^* N(0, \sigma^{*2}) + (1 - \alpha^*) N(\mu^*, \sigma^{*2}). \tag{2.2}$$

Thus, with probability $\alpha^* \in [0,1]$, the departure is Gaussian with mean 0 and variance σ^{*2}, but with probability $(1 - \alpha^*)$ it is "spiked" by a scalar amount μ^*. In this setting, MLE estimates of α^*, μ^* and σ^* can be obtained through the EM algorithm. Previous work on EM convergence often assume that the only parameter to be estimated is μ^*, with σ^{*2} and α^* taken as known; see [23] and [24]. Drawing inspiration from the latter, we study convergence when all parameters μ^*, σ^{*2} and α^* are unknown – proving that $\nu-$strong concavity, Lipschitz smoothness and Gradient smoothness conditions hold for our Gaussian mixture model. Guarantees on the convergence rate are harder to establish

when all three parameters are treated as unknown, but can in fact be provided if the "contamination" level α^* is taken as known. We use simulations to demonstrate the practical effectiveness of our approach notwithstanding this shortcoming in theoretical guarantees.

The remainder of this chapter is organized as follows. Section 2 provides some technical background on smoothing splines and EM algorithm. Section 3 details our approach and the conditions under which it performs well. Section 4 provides convergence guarantees for the EM algorithm. Sections 5 and 6 demonstrate the performance of our proposal through simulations and real data analyses. Section 7 contains final remarks.

2.2 Technical background

2.2.1 Smoothing splines

Suppose that the data $(x_i, y_i)_{i=1}^n$ are generated according to

$$y_i = f(x_i) + \epsilon_i \tag{2.3}$$

where $x_i \in [0,1], i = 1, \cdots n$ are either fixed or random, and the ϵ_i's represent white noises (independent and Gaussian random errors). Assuming that f has p continuous derivatives on $[0,1]$, i.e. $f \in C^p[0,1]$, we are interested in approximating f from $(x_i, y_i)_{i=1}^n$. In a spline approximation [22, 25] the estimator $\hat{f}$ is restricted to lie in the space $\$_{m,\boldsymbol{t}}$ of spline functions of order m with knots sequence $\boldsymbol{t} = \{0 = t_0 < t_1 < \cdots < t_{K_0+1}\}$. Functions in this space have the representation $\sum_j a_j N_{k,\boldsymbol{t}}^{[m]}$, where

$$N_{k,\boldsymbol{t}}^{[m]}(x) = (t_k - t_{k-m})[t_{k-m}, \cdots, t_k](\cdot - x)_+^{m-1}, \quad 1 \le k \le K = K_0 + m$$

is the k^{th} B-spline function. The notation $[t_{k-m}, \cdots, t_k](\cdot - x)_+^{m-1}$ represents the m^{th} divided difference of $(\cdot - x)_+^{m-1}$ at sites $t_{k-m}, \cdots, t_k$. This divided difference is the leading coefficient of the polynomial g of order $m+1$ that agrees with $(\cdot - x)_+^{m-1}$ at the sequence $t_{k-m}, \cdots, t_k$. Thus if the sequence $t_{k-m}, \cdots, t_k$ is non-distinct, say with n_{t^*} values coinciding at t^*, then for $g(t)$ such that

$$(t^* - x)_+^{m-1,(i-1)} = g^{(i-1)}(t^*) \quad \text{for } i = 1, \cdots, n_{t^*}$$

$[t_{k-m}, \cdots, t_k](\cdot - x)_+^{m-1}$ is the coefficient of the term t^m in g. From here on, where they are obvious, we will suppress $\boldsymbol{t}, m$ from the notation $N_{k,\boldsymbol{t}}^{[m]}$.

Depending on the number of bases or, equivalently, the number of knots, $\hat{f}$ can either underfit or overfit the data. Solving for the optimal number (and locations) of knots can prove very challenging [17]. An alternative approach is to use a large number of knots K and regularize $\hat{f}$ by placing a penalty on its higher-order derivatives [2, 22, 26]; that is, to minimize

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{K} a_k N_k(x_i) \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_{k=1}^{K} a_k N_k^{(q)}(x) \right\}^2 dx \tag{2.4}$$

where λ is a tuning parameter whose optimal value can be found using cross validation.

We now briefly describe a matrix representation of (2.4) introduced by [22]. Let $N(x) = [N_1(x), \cdots, N_K(x)]^T \in \mathbb{R}^K$ and $N = [N(x_1), \cdots, N(x_n)]^T \in \mathbb{R}^{n \times K}$. Define $\Delta_{K,1} \in \mathbb{R}^{(K-1)\times K}$ so that, for any $\theta \in \mathbb{R}^K$, $\Delta_{K,1}\theta = (\theta_2 - \theta_1, \cdots, \theta_K - \theta_{K-1})^T$, and define recursively the q^{th} order difference operator $\Delta_{K,q} = \Delta_{K-1,q-1}\Delta_{K,1}$, for $1 < q < K$. In addition, let $W_K^{[m]} \in \mathbb{R}^{(K-1)\times(K-1)}$ be a diagonal matrix whose k^{th} diagonal entry is equal to $(m-1)(t_k - t_{k-m+1})^{-1}$. Using this notation, we have that

$$\frac{dN^{[m]}(x)}{dx} = \Delta_{K,1}^T W_K^{[m]} N^{[m-1]}(x) \ .$$

For higher order derivatives, let $\tilde{\Delta}_{K,1,m} = W_K^{[m]}\Delta_{K,1} \in \mathbb{R}^{(K-1)\times K}$ and define recursively $\tilde{\Delta}_{K,q,m} = \tilde{\Delta}_{K-1,q-1,m-1}\tilde{\Delta}_{K,1,m}$. Then

$$\frac{d^q N^{[m]}(x)}{dx^q} = \tilde{\Delta}_{K,q,m}^T N^{[m-q]}(x) \ .$$

We are now in the position to rewrite (2.4) as

$$\min_{\boldsymbol{a}} \left(\frac{1}{n} \left\| Y - N\boldsymbol{a} \right\|_2^2 + \lambda \boldsymbol{a}^T P_q \boldsymbol{a} \right) \tag{2.5}$$

where $P_q = \tilde{\Delta}_{K,q,m}^T G^{[m-q]} \tilde{\Delta}_{K,q,m}$ with $G^{[m-q]} = \int N^{[m-q]}(x) N^{[m-q],T}(x) dx \in \mathbb{R}^{n\times K}$. Equation (2.5) can be solved explicitly; if we set $H_n := N^T N/n + \lambda P_q$, then the solutions are

$$\begin{aligned} \hat{\boldsymbol{a}} &= H_n^{-1}(N^T Y/n) \\ \hat{f}(x) &= N^T(x) H_n^{-1}(N^T Y/n) \end{aligned}$$

2.2.2 EM algorithm for Gaussian mixtures

Let $\xi \in \Xi$ and $z \in Z$ be random variables whose joint density function is ϕ_{θ^*}, where θ^* belongs to a (non-empty) convex parameter space Ω. Suppose we can observe data $(\xi_i)_{i=1}^n$, while the $(z_i)_{i=1}^n$ are unobservable, and that, for example, $(\xi_i|z_i = j) \overset{iid}{\sim} G_j$ where the G_j's are Gaussian distributions. Our goal is to estimate the unknown θ^* using Maximum Likelihood; that is, to find $\hat{\theta}$ that maximizes

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} log\left(\int_Z \phi_\theta(\xi_i, z_i)dz_i\right) .$$

In practice, the function ℓ_n is usually hard to optimize. The EM algorithm provides a way of searching for such maximum indirectly through the maximization of another function $Q_n : \Omega \times \Omega \to \mathbb{R}$ defined as

$$Q_n(\theta|\theta') = \frac{1}{n}\sum_{i=1}^{n}\left(\int_Z k_{\theta'}(z|\xi_i)\log\phi_\theta(\xi_i, z)dz\right)$$

where $k_\theta(z|\xi)$ is the conditional density of z given ξ. Given this function and a current estimate $\theta_{n,t}$, the sample EM update is defined as

$$\theta_{n,t+1} = \theta_{n,t} + \alpha\nabla Q_n(\theta|\theta_{n,t})\Big|_{\theta=\theta_{n,t}}, \qquad t = 0, 1, \cdots .$$

To study convergence of the EM to a (neighborhood of) the global optimum, [24] define the population level versions ℓ of ℓ_n and Q of Q_n as

$$\begin{aligned}\ell(\theta) &= \int_\Xi log\left(\int_Z \phi_\theta(\xi_i, z_i)dz_i\right)g_{\theta^*}(\xi)d\xi \\ Q(\theta|\theta') &= \int_\xi\left(\int_Z k_{\theta'}(z|\xi_i)\log\phi_\theta(\xi_i, z)dz\right)g_{\theta^*}(\xi)d\xi .\end{aligned}$$

Correspondingly, one has a population version of the EM update

$$\theta_{t+1} = \theta_t + \alpha\nabla Q(\theta|\theta_t)\Big|_{\theta=\theta_t}, \qquad t = 0, 1, \cdots .$$

Based on this, since θ^* maximizes $\ell(\theta)$, to prove that the sample EM update converges to (a neighborhood of) θ^*, one needs to prove that (i) the population EM update converges to (a neighborhood of) θ^*; and (ii) the sample EM update tracks closely the population update (this is precisely what we do in Proposition 2.4.1 and Proposition 2.4.2 below).

2.3 The *smoothEM* approach

Let us consider again data as in Equation (2.2),

$$y_i = f(x_i) + \xi_i$$
$$\xi_i \sim \alpha^* N(0, \sigma^{*2}) + (1 - \alpha^*) N(\mu^*, \sigma^{*2})$$

where the design (the x's) is taken as fixed. If we knew f, the EM algorithm could be used to search for the MLE of the mixture parameters $\theta^* = (\alpha^*, \mu^*, \sigma^{*2})$ and to estimate membership (i.e. posterior) probabilities for each point, and thus the classification vector $1_{\mathbb{S}}$. In reality, we do not know f, so we use the EM with an $\hat{f}$ that is as close as possible to f. Such an $\hat{f}$ is the result of the following procedure.

2.3.1 Iterated smoothing

The regularized smoothing technique in (2.4) declares as optimal an $\hat{f}$ that minimizes both sum of squared errors and degree of roughness. The tuning parameter λ, chosen by cross validation, determines the balance between these two competing criteria. In the case of (2.1), a sufficiently large μ^* causes the sum of squared errors term to dominate the roughness criterion. This in turns causes the cross validation procedure to be biased towards small values of λ. This is the key observation that gives rise to our iterated regularized smoothing procedure, which we demonstrate here through a simple illustrating example. Figure 2.1 plots $n = 500$ data points observed across the [0,1] domain. The majority of such points are scattered about a fourth degree polynomial, but some form three *"hovering clouds"* of different denseness at three different locations along the domain. Here, we use $\epsilon_i \overset{iid}{\sim} N(0, 1)$ and $\mu^* = 12$. Purple dots are observations from the noisy smooth component $f(x_i) + \epsilon_i$, and pink dots are observations where spikes have been added on top of it, i.e. from $f(x_j) + \mu^* + \epsilon_j$. The true underlying smooth curve f is not plotted, and the estimates $\hat{f}$ are plotted in different shades of green and pink, depending on the values of λ, here ranging from 10^{-5} to 1. To calculate $\hat{f}$, our example uses 300 cubic spline bases with penalty of order 1. At first glance, none of the $\hat{f}$'s approximates f well, though some give better fits than the others. In particular, note that the presence of spikes, especially when coupled with larger λ, affects estimation at both spike and non-spike locations. Also, a "denser" spikes site (leftmost cloud) distorts $\hat{f}$ to a higher degree – especially for small λ values. Generalized cross-validation here selects $\lambda = 10^{-4}$, which still results in a highly distorted $\hat{f}$. Suppose now we were to

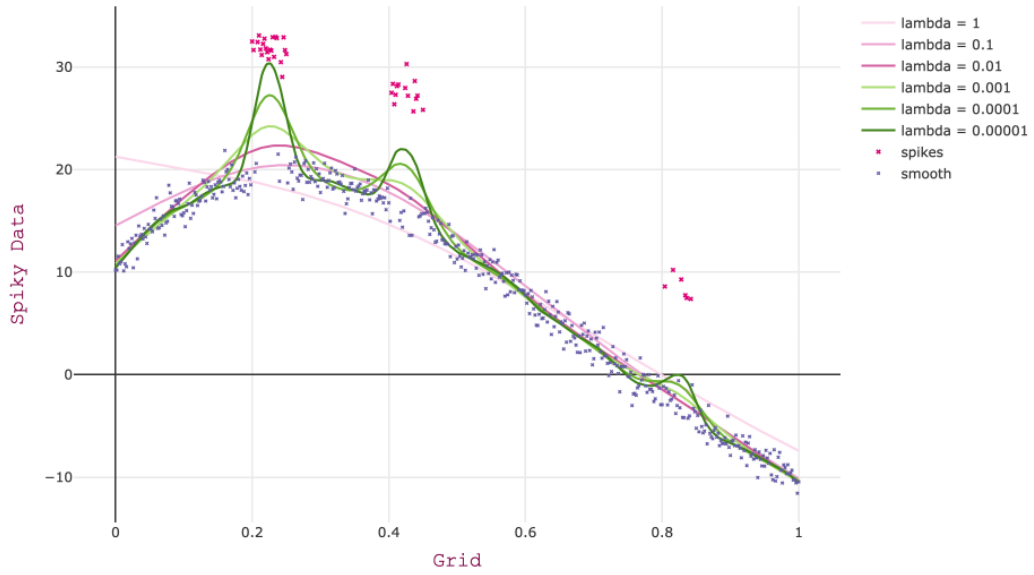

Figure 2.1: Spline fit for observed data with $\lambda = 1, 0.1, 0.01, 0.001, 0.0001, 0.00001$

identify spikes and do inference on the parameters $\alpha^*, \mu^*, \sigma^{*2}$ through the residuals from $\hat{f}$. Such a distorted estimate would generate misleading residuals. On the other hand, just from visual inspection, $\lambda = 10^{-1}$ and $\lambda = 10^{-2}$ produce much more reasonable $\hat{f}$'s and, correspondingly, residuals which are a much better approximation of the underlying ϵ_i's, giving some hope that one may be able to identify spikes based on their magnitudes. If we identify and filter out the spikes, and then repeat regularized smoothing, we can obtain a much improved fit to f. Based on this reasoning, we propose an *iterated smoothing procedure* comprising three steps:

S1 Fit a regularized smoothing spline over a grid of λ values, and obtain the residuals from each fit;

S2 Choose a value of λ based on how well the corresponding residuals' magnitudes separate the spikes, and filter them out (see Section 2.3.2 below);

S3 Fit again a regularized smoothing spline over a grid of λ values, but excluding the spikes identified in Step 2 and select λ by generalized cross-validation to obtain the final $\hat{f}$.

The example above represents a rather "ideal" situation, in which f is sufficiently smooth, the ratio $\frac{\mu^*}{\sigma^*}$ is large, the number of observations n is large, and the denseness at spike locations is not too detrimental. The smoothness of f facilitates interpolation at

locations where spikes are removed; a large $\frac{\mu^*}{\sigma^*}$ facilitates the identification of the spikes; and a limited denseness at spike locations means that, even in the first round of smoothing, for some λ value $\hat{f}$ will not be too distorted and will produce residuals that allow one to pinpoint spikes. In less ideal scenarios, one might have to iterate smoothing and spike removal several times – and indeed if f is highly irregular, $\frac{\mu^*}{\sigma^{*2}}$ is too small, or spike locations are too dense, the procedure might not work at all. Before further articulating our proposal in Section 2.3.2, we discuss here at some more length the interplay between smoothness of f, λ, $\frac{\mu^*}{\sigma^*}$, n, and the denseness of spikes, in determining the effectiveness of iterated smoothing.

Suppose $(x_i, y_i)_{i=1}^n$ are generated as at the beginning of Section 3.3. For a given realization of $\mathbb{1}(x_i \in \mathbb{S})$ (i.e. with the spikes location fixed), the only randomness is from the noise $(\epsilon_i)_{i=1}^n$. Assuming without loss of generality that $\mu^* > 0$, let $r(x_p)$ and $r(x_s)$ be residuals from Step 1 at spike location x_p and smooth location x_s. The difference

$$\begin{aligned} r(x_p) - r(x_s) &= Y_p - \hat{f}(x_p) - (Y_s - \hat{f}(x_s)) \\ &= f(x_p) + \mu^* + \epsilon_p - N(x_p) H_n^{-1} N^T (f(\boldsymbol{x}) + \mu^* \cdot \mathbb{1}_{\mathbb{S}} + \boldsymbol{\epsilon})/n \\ &\quad - [f(x_s) + \epsilon_s - N(x_s) H_n^{-1} N^T (f(\boldsymbol{x}) + \mu^* \cdot \mathbb{1}_{\mathbb{S}} + \boldsymbol{\epsilon})/n] \end{aligned}$$

can be decomposed as $r(x_p) - r(x_s) = C_1 + C_2 + C_3$, where

$$\begin{aligned} C_1 &= f(x_p) - N(x_p) H_n^{-1} N^T (f(\boldsymbol{x}) + \boldsymbol{\epsilon})/n \\ &\quad - [f(x_s) - N(x_s) H_n^{-1} N^T (f(\boldsymbol{x}) + \boldsymbol{\epsilon})/n] \\ C_2 &= \mu^* - N(x_p) H_n^{-1} N^T \mu^* \cdot \mathbb{1}_{\mathbb{S}}/n \\ &\quad - [0 - N(x_s) H_n^{-1} N^T \mu^* \cdot \mathbb{1}_{\mathbb{S}}/n] \\ C_3 &= \epsilon_p - \epsilon_s \end{aligned}$$

The first component is $C_1 = r'(x_p) - r'(x_s)$, the difference of the residuals from fitting a spline to the noisy smooth component only. As long as the underlying true $f(x)$ is $\in C^p[0,1]$ for $p \leq m$, where m is the order of the smoothing spline, existing literature [22] guarantees that under appropriate conditions $\max_x r'(x)$ is of order $o\Big\{\Big(\frac{\log n}{n}\Big)^{\frac{-m}{2m+1}}\Big\}$. The second component is $C_2 = r''(x_p) - r''(x_s)$, the difference of the residuals from fitting a spline to the n-discretized piecewise constant $\mu^* \cdot \mathbb{1}_{\mathbb{S}}$. As spline smoothing is highly localized by nature, intuitively, if the number of spikes in a neighborhood of x_p is sufficiently small and λ is sufficiently large, the smoothing spline will prioritize approximation of the constant line $g(\cdot) = 0$, causing $r''(x_p) - r''(x_s)$ to be near μ^*. As more spikes gather

around x_p, this magnitude decreases, making spikes less distinguishable. It is important to note that the same λ is used to fit the noisy $f(x)$ in C_1 and $\mu^* \cdot \mathbb{1}_{\mathbb{S}}$ in C_2. As a consequence, if $f(x)$ is rather rugged and thus the optimal λ to fit it small, one may easily overfit $\mu^* \cdot \mathbb{1}_{\mathbb{S}}$ – especially when spikes are dense in a neighborhood of x_p (see Figure 2.2). The third component is simply $C_3 = \epsilon_p - \epsilon_s$. Under ideal conditions, C_1

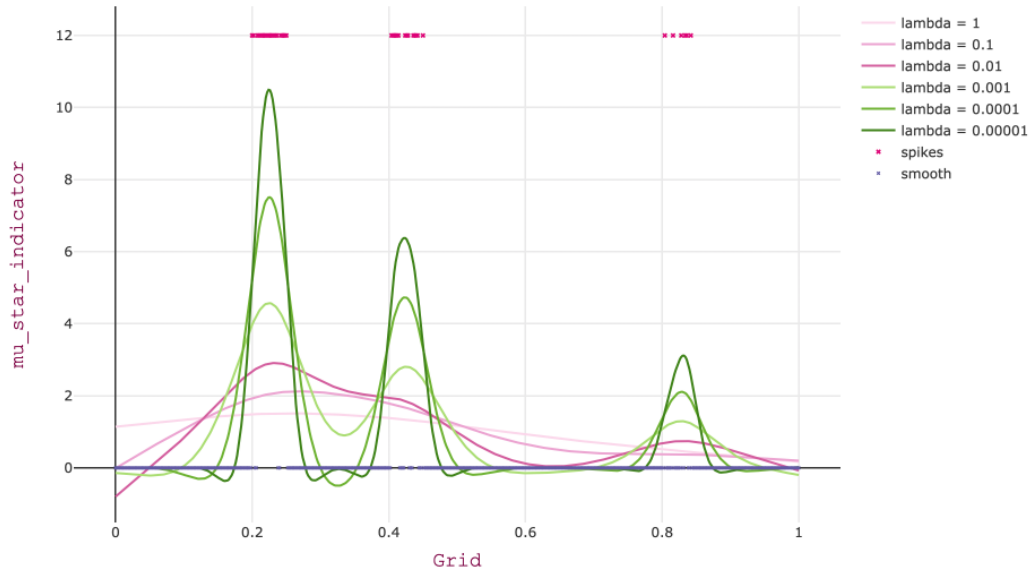

Figure 2.2: Spline fit for $\mu^* \cdot \mathbb{1}_{\mathbb{S}}$ with $\lambda = 1, 0.1, 0.01, 0.001, 0.0001, 0.00001$

will be close to 0, C_2 will be close to μ^*, and as long as $\frac{\mu^*}{\sigma^{*2}}$ is large, the effect of C_3 will be negligible.

2.3.2 EM

Here we go back to the iterated smoothing procedure introduced in Section 3.3.1, better articulating how residuals from the first smoothing round can be used both to identify spikes and to produce inferences about the parameters $(\alpha^*, \mu^*, \sigma^{*2})$. These two goals can be pursued simultaneously with the use of the EM algorithm. Specifically, the steps in our procedure will be modified as follow:

S'1 Fit a regularized smoothing spline over a grid of λ values, and obtain the residuals $\boldsymbol{\xi}(\lambda) = \{\xi(\lambda)_i\}_{i=1}^n$;

S'2 For each pair $(\lambda, \boldsymbol{\xi}(\lambda))$, classify $\boldsymbol{\xi}(\lambda)$ into two groups based on their magnitudes. This can be done in various ways, e.g., running a 1-dimensional K-means algorithm,

or based on the largest sequential difference of ordered residuals. Label as spikes the group with higher magnitudes, provided its cardinality does not exceed 30% of the total number n of observations. If no group of large residuals can be identified, or if one can be identified but it contains more than 30% of the observations, simply label all observations as "smooth" (none as spikes). The output of this step are binary group memberships $\{M_i(\lambda)\}_{i=1}^n$ for each value of λ ($M_i(\lambda) = 0, 1$ means the observation (x_i, y_i) is classified as smooth or spike, respectively, when using λ in Step 1).

S'3 For each pair $(\lambda, \boldsymbol{M}(\lambda))$, fit a second regularized smoothing spline using only observations with $M_i(\lambda) = 0$, and compute updated residuals $\boldsymbol{\xi'}(\lambda)$ for all observations

S'4 For each λ, run an EM algorithm on $\boldsymbol{\xi'}(\lambda)$, using $\boldsymbol{M}(\lambda)$ as initialization. Choose as best the λ^* that maximizes the log likelihood $\lambda^* = \arg\max_\lambda \ell(\lambda; \hat{\alpha}, \hat{\mu}, \hat{\sigma}^2)$. Obtain estimates $(\hat{\alpha}, \hat{\mu}, \hat{\sigma}^2)_{\lambda^*}$ and update the binary memberships to $\boldsymbol{M'}(\lambda^*)$, obtained by thresholding the EM posterior probabilities.

S'5 Repeat (S1'), fitting again a regularized smoothing spline over a grid of λ values, but using only the points with $M_i'(\lambda^*) = 0$.

The procedure is iterated until $\boldsymbol{M'}(\lambda^*)$ no longer changes, or when $\boldsymbol{M'}(\lambda^*)$ is identical to one from the previous iterations. Obviously the first rule is preferable, but there are times when $\boldsymbol{M'}(\lambda^*)$ loops around a number of possible membership sets, and the second rule can be used to prevent looping (to chose among looping membership sets, we can again use the log likelihood).

An important observation here is that the performance of the procedure depends critically on the grid of λ values chosen for its implementation. We can choose a relatively small grid based on prior knowledge about the smoothness of f and/or visual inspection. If there is no prior knowledge and/or visual inspection is not particularly informative, we must choose a sufficiently large grid that will likely include a good λ. However, a larger grid means we are fitting more smoothing splines, which comes at greater computational cost. Another avenue that needs prior input from the user is the threshold value in step S'4. A straightforward approach would be to classify a data point as spike if its probability of belonging to the spike group exceed 0.5. Our algorithm automatically chooses from threshold values ranging between 0.5 and 1 such that the resulting classification maximizes the log-likelihood.

2.4 Theoretical results for EM

We start by stating three conditions that are needed to guarantees good properties for the EM algorithm. In the following, let $q(\theta) = Q(\theta|\theta^*)$ (see Section 2.2.2). The notation $\mathbb{B}_2(r;\theta^*)$ below is used to denote an L_2 ball centered at θ^* with radius r.

C1 (ν-strong concavity) There is some $\nu > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \leq -\frac{\nu}{2} \|\theta_1 - \theta_2\|_2^2 \quad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{B}_2(r;\theta^*)$$

C2 (Lipschitz-smoothness) There is some $L > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \geq -\frac{L}{2} \|\theta_1 - \theta_2\|_2^2 \quad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{B}_2(r;\theta^*)$$

C3 (Gradient Smoothness). For an appropriately small parameter $\gamma > 0$,

$$\|\nabla q(\theta) - \nabla Q(\theta|\theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \quad \text{for all } \theta \in \mathbb{B}_2(r;\theta^*)$$

The Theorem below, proved in [24], utilizes these conditions to formulate guarantees for the population level EM.

Theorem 2.4.1 ***(Balakrishnan, Wainwright & Yu, 2017)***
For some radius $r > 0$, and a triplet (γ, ν, L) such that $0 \leq \gamma < \nu \leq L$, suppose that Conditions 1, 2, and 3 hold, and suppose that the stepsize is chosen as $s = \frac{2}{\mu+\nu}$. Then given any initialization $\theta_0 \in \mathbb{B}_2(r;\theta^*)$, with probability $1 - \delta$, the population first order EM iterates satisfy the bound

$$\|\theta_k - \theta^*\|_2 \leq \left(1 - \frac{2\nu - \gamma}{L + \nu}\right)^k \|\theta_0 - \theta^*\|_2 \quad \textit{for all } k = 1, 2, \cdots$$

Based on this Theorem, we have the following Proposition concerning our problem (a proof is provided in the Appendix).

Proposition 2.4.1 ***(Population level guarantees for the Gaussian Mixture)***
The ϵ below denotes an arbitrarily small, positive number. It's not to be confused with noise in the model. Provided $\alpha^* \in [0.7, 1]$, our Gaussian Mixture model in Eq. (2.2) satisfies Conditions 1, 2 and 3 with

- $\nu = \min\{(\frac{1}{(\alpha^*+r)^2} \vee 1), \frac{\sigma^{*2}-r}{2(\sigma^{*2}+r)^3} - \frac{(1-\alpha^*)r}{(\sigma^{*2}-r)^2}, \frac{1-\alpha^*}{\sigma^{*2}+r} - \frac{(1-\alpha^*)r}{(\sigma^{*2}-r)^2}\}$.

- $L = \max\{\frac{\alpha^*}{(\alpha^* - r \vee 0.7)^2} + \frac{1-\alpha^*}{(1-\alpha^* - r \vee \epsilon)^2}, \frac{(1-\alpha^*)\sigma^{*2}}{(\sigma^{*2}-r)^2}, \frac{\sigma^{*2}+r}{2(\sigma^{*2}-r)^3} + \frac{1-\alpha^*}{\sigma^{*2}-r}\}$.
- $\gamma(\alpha^*, \mu^*, \sigma^{*2}) \sim O\left(\frac{\mu^{*5}}{\sigma^{*8}} \exp(-\frac{\mu^* - r}{\sigma^{*2}+r}\epsilon_0)\right)$, *which goes to* 0 *exponentially fast with large* $\frac{\mu^*}{\sigma^{*2}}$.

Remark 1. ν decreases as r increases; this creates a trade off, as ideally we would want both to be large. A larger r allows for a larger basin of attraction for convergence, whereas a larger ν hastens the convergence rate. One would wish to find the largest values of r at which $\nu = 0$, but this requires solving a 4th degree polynomial. As a bypass, table 2.1 demonstrates some admissible $(r; \nu)$ pairs for various parameters settings. We propose to allow r to grow with σ^{*2}; taking $r = \frac{\sigma^*}{3}$ leads to a positive value of ν in most cases.

Remark 2. Analyzing Condition 2, we see that even if such a positive ν exists, the convergence rate can be prohibitively slow due to a large Lipschitz smoothness constant L. We note that a large value of σ^{*2} also leads to a smaller ν, which negatively affects the convergence rate unless the radius r is chosen to be small – in a lot of cases, too small to be meaningful.

Remark 3. Regarding the constant L in Condition 2, $\frac{1-\alpha^*}{(1-\alpha^* - r \vee \epsilon)^2}$ can be arbitrarily large, thus making the rate $\left(1 - \frac{2\nu - \gamma}{L+\nu}\right)^k$ unacceptably slow. As a way to enforce an upper bound for L, we can constrain α^* to be upper-bounded by a constant, say 0.95. However, even if we do constrain α^*, the benefit is meager; even under ideal circumstances, we can only manage to get the convergence rate down to 0.999^k, which is still very slow. There might be other approaches to guarantee fast convergence with unknown α^* and σ^{*2}. While we are unaware of such approaches, we take comfort in the fact that our simulations seem to provide evidence of reasonable convergence rates with our procedure (see Section 2.5).

Remark 4. The majority of current literature assumes that both σ^{*2} and α^* are known. If this is the case, then $\nu = L$ and we can prove that population updates converge geometrically to the true population parameters, provided that the ratio $\frac{\mu^*}{\sigma^{*2}}$ is large enough (Condition 3). If we assume only that α^* is known (so both σ^{*2} and μ^* are unknown) we can obtain a reasonable convergence rate using

- $r = \frac{\sigma^*}{3}$
- $\nu = \min\{\frac{\sigma^{*2}-r}{2(\sigma^{*2}+r)^3} - \frac{(1-\alpha^*)r}{(\sigma^{*2}-r)^2}, \frac{1-\alpha^*}{\sigma^{*2}+r} - \frac{(1-\alpha^*)r}{(\sigma^{*2}-r)^2}\}$.
- $L = \max\{\frac{(1-\alpha^*)\sigma^{*2}}{(\sigma^{*2}-r)^2}, \frac{\sigma^{*2}+r}{2(\sigma^{*2}-r)^3} + \frac{1-\alpha^*}{\sigma^{*2}-r}\}$.

(conv. rate; k)		(σ^*, r)				
		(1.1,0.37)	(2.1,0.7)	(3.1,1.03)	(4.1,1.37)	(5.1,1.7)
$1-\alpha^*$	.1	(0.984,40)	(0.795,17)	(0.807,17)	(0.852,23)	(0.894,32)
	.05	(0.991,70)	(0.795,17)	(0.667,10)	(0.702,15)	(0.752,18)

Table 2.1: Population updates convergence rates for different values of σ^*, α^*, r

- $\gamma(\alpha^*, \mu^*, \sigma^{*2}) \sim O\left(\frac{\mu^{*5}}{\sigma^{*8}} \exp(-\frac{\mu^*-r}{\sigma^{*2}+r}\epsilon_0)\right)$ which is negligible if $\frac{\mu^*}{\sigma^*}$ is assumed to be large enough.

Table 2.1 provides values of r, $\left(1-\frac{2\nu-\gamma}{L+\nu}\right)$, and k (number of iterations) that guarantee a rate $\left(1-\frac{2\nu-\gamma}{L+\nu}\right)^k < 0.0001$.

Next, we consider a Theorem for the sample EM, also proved in [24].

Theorem 2.4.2 ***(Balakrishnan, Wainwright & Yu, 2017)*** *For a given size n and tolerance parameter $\delta \in (0,1)$, let $\epsilon_Q^{unif}(n,\delta)$ be the smallest scalar such that with probability at least $1-\delta$*

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \|\nabla Q_n(\theta|\theta) - \nabla Q(\theta|\theta)\|_2 \leq \epsilon_Q^{unif}(n,\delta)$$

Suppose that, in addition to the conditions of Theorem (2.4.1), the sample size n is large enough to ensure that

$$\epsilon_Q^{unif}(n,\delta) \leq (\nu-\gamma)r$$

Then with probability at least $1-\delta$, given any initial vector $\theta_0 \in \mathbb{B}_2(r;\theta^*)$, the finite sample EM iterates $\{\theta_k\}_{k=0}^{\infty}$ satisfy the bound

$$\|\theta_t - \theta^*\|_2 \leq \left(1-\frac{2\nu-2\gamma}{L+\nu}\right)^t \|\theta_0-\theta^*\|_2 + \frac{\epsilon_Q^{unif}(n,\delta)}{\nu-\gamma} \ .$$

Based on this Theorem, we have the following Proposition concerning our problem (a proof is provided in the Appendix).

Proposition 2.4.2 ***(Sample level guarantees for the Gaussian Mixture)*** *For our Gaussian Mixture model in Eq (2.2), $\epsilon_Q^{unif}(n,\delta) \to 0$ almost surely.*

2.5 Simulation results

In the following simulations, $(x_i)_{i=1}^n$ are equispaced along the interval $[0, 1]$. We look at how the procedure performs under $n = 2000, 1000, 500, 200$, α^* ranging from 0.8 to 0.98 (i.e. for spike contamination levels between 0.2 and 0.02), and the STN $\frac{\mu^*}{6\sigma^*}$ ranging from 0.2 to 2 (σ^* is fixed at 1). Note that the factor $6\sigma^*$ used to define the spike size parameter represents the width of the 95% Gaussian noise interval at any given location. Thus, for instance, if (x_i, y_i) lies $3\sigma^*$ *below* $f(x_i)$, and if $x_i \in \mathbb{S}$, then $\mu^* = 2 \cdot 6\sigma^*$ practically brings (x_i, y_i) to a new height well separated from points that are $3\sigma^*$ distance *above* the graph of f. Due to the local nature of spline basis, different distributions of spike locations skew traditional spline smoothing results differently; we therefore consider both an "even" and a "clumped" simulation scenario.

2.5.1 Uniformly distributed spikes

Figure 2.3 depicts a scenario where the spike locations are uniformly distributed across the domain. The blue curve represents the true underlying smooth component. Figure

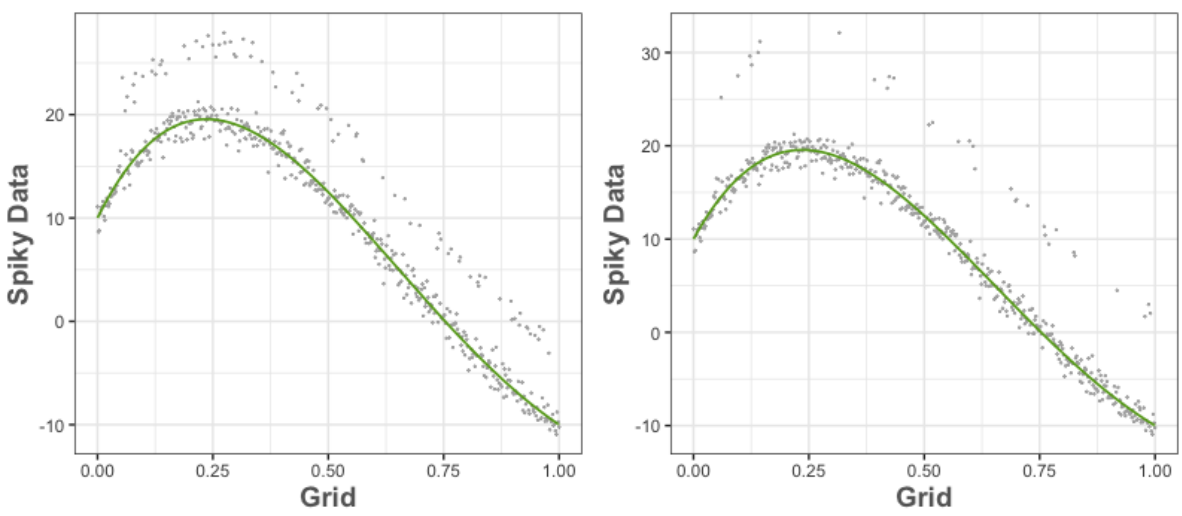

Figure 2.3: *Simulated data with uniformly distributed spikes. The green curve represents the true smooth component.* $n = 500, \sigma^* = 1$. *left:* $\alpha^* = 0.85, \mu^* = 7.2$*; right:* $\alpha^* = 0.94, \mu^* = 12$.

2.4 contains contour plots of $\left\|\hat{f} - f\right\|_2$ when $n = 200, 500, 1000, 2000$. Each plot shows how $\left\|\hat{f} - f\right\|_2$ varies with different spike percentages (α^*) and relative magnitudes ($\frac{\mu^*}{6\sigma^*}$). For each parameter setting, the results shown are averages over 20 replicate of the

simulation. Figure 2.4 is more informative when analyzed together with Figure 2.5, which contains similar plots for the False Negative Rate (FNR) in spike identification. We are not reporting results for the False Positive Rate (FPR), whose maximum over all parameter settings is 0.02, which implies high power of the procedure. We observe that,

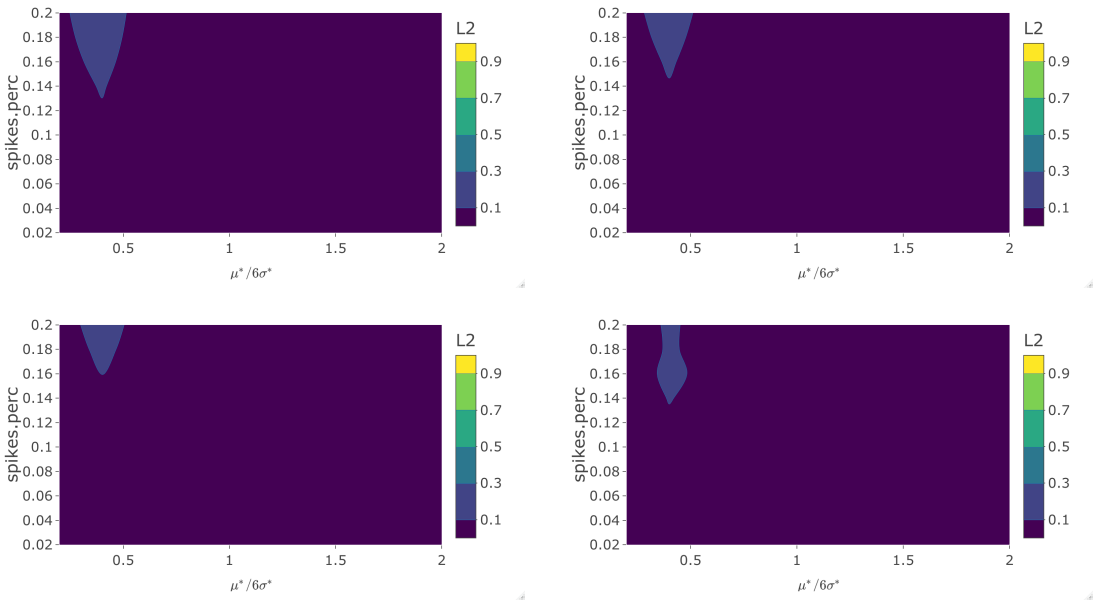

Figure 2.4: *(from left to right, top to bottom) L2-error of smooth component estimate in the case of uniformly distributed spikes when $n = 200$, $n = 500$, $n = 1000$, $n = 2000$.*

when n is large, a sufficiently large signal ($\frac{\mu^*}{6\sigma^*} \geq 1$) corresponds to low FNR (i.e. good classification), and thus low error in estimating the smooth component. When signal is small, so that spikes are not well separated, our procedure naturally has a harder time recognizing them. However, estimation of the smooth component doesn't suffer much in these settings, as smaller spikes don't distort the fit substantially. Notably though, both spike identification and estimation of the smooth component do improve with larger spike size and lower α^*, which is in line with our previous discussion in section 3.3.1. Figure 2.6 plots the sum of squared error of parameter estimates $\left\|\hat{\theta} - \theta^*\right\|_2$.

2.5.2 Nonhomogeneous Poisson spikes

Figure 2.7 depicts a scenario where spikes are "clumped" instead of being distributed in a uniform manner along the domain. Figures A.1, A.2, A.3 in the Appendix are the analogs of Figures 2.4, 2.5, 2.6 – and they look very similar to the latter. This suggest that our algorithm performs similarly well, and thus that it possesses a degree

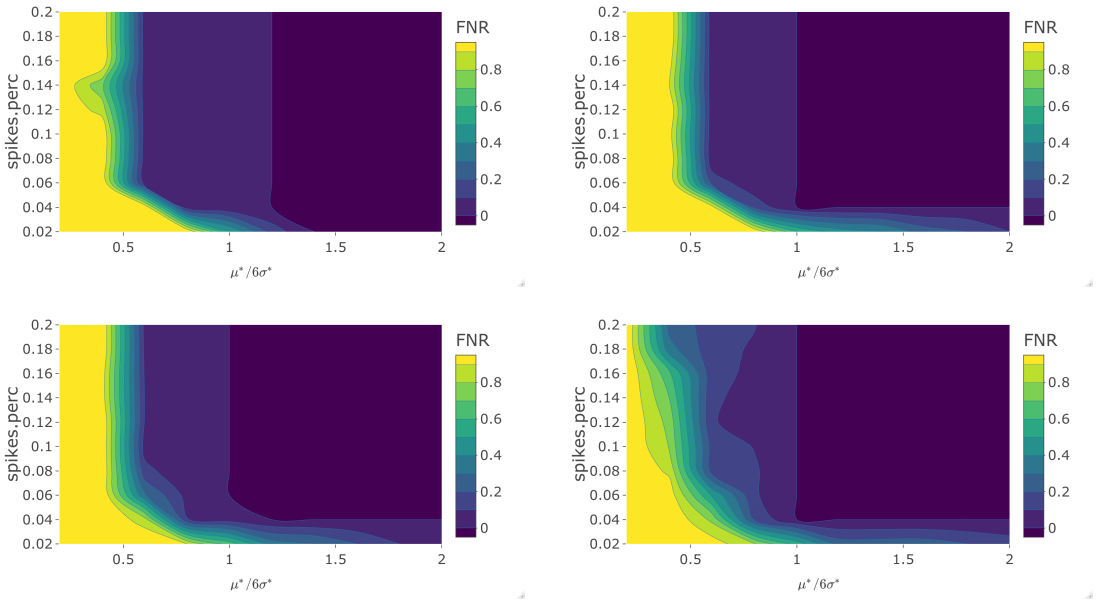

Figure 2.5: *(from left to right, top to bottom) FNR of spikes identification in the case of uniformly distributed spikes when $n = 200$, $n = 500$, $n = 1000$, $n = 2000$.*

of robustness to the different ways spikes may be distributed across the domain.

2.6 Data applications

2.6.1 Smart meter electricity data

For our first application we considered data from the Smart Meter Electricity project, which was conducted during 2009 and 2010 by the Commission for Energy Regulation (CER) in Ireland. Over 5000 Irish households and businesses participated. The main goal of the project was to study the costs and benefits of a smart metering system as opposed to the existing electro-mechanical and diaphragm metering system. As smart meters are capable of recording and communicating detailed energy statements and electricity bills to end consumers, policy makers were interested in understanding whether the availability of this richer information would lead to a reduction in overall measured electricity consumption. We did not address the goal of the policy makers. Rather, sought a meaningful statistical representation of the electricity consumption behavior, which may of course also be useful to policy makers.

Assuming that electricity consumption predominantly follows a smooth pattern with occasional spiked activity – for example during the times when multiple electrical devices

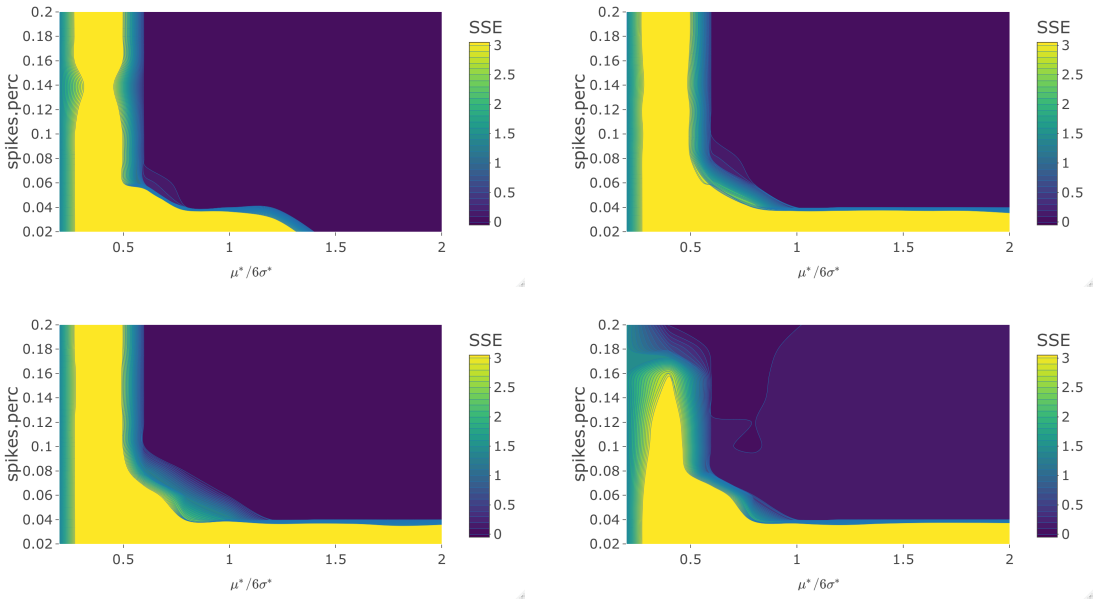

Figure 2.6: *(from left to right, top to bottom) SSE of parameter estimates in the case of uniformly distributed spikes when $n = 200$, $n = 500$, $n = 1000$, $n = 2000$.*

are turned on simultaneously – the Smart Meter Electricity data can be effectively analyzed with our framework. The data contains daily measurements of electricity consumed, collected at 30 minute intervals in kWh, for each household in the study. As an illustration, we ran our procedure on data from one household (meter ID 1392) and two consecutive days in the Summer of 2009. A visual inspection of the data in Figure 2.8, suggests that imposing the variances around the smooth component and the spikes to be the same may be too restrictive here. If we do, the algorithm returns a single point as outlier. Because of this, we allowed errors to have not only a mean shift but also inflated variance at spike locations; that is, we used the model

$$g_{\theta^*}(\xi) = \alpha^* \phi(\xi; 0, \sigma_\epsilon^{*2}) + (1 - \alpha^*)\phi(\xi; \mu^*, \sigma_\epsilon^{*2} + \sigma_h^{*2}) \ .$$

The inflated variance adds a new variance parameter to be estimated in EM algorithm, which is not covered in our theoretical treatment in Section 2.4. However, the algorithm converged very fast. The (somewhat arbitrary) choice of grid for the tuning parameter λ was (50,40,30,20,10). Given this grid, the algorithm yielded the spike identification and estimated smooth curve shown in red and green, respectively, in Figure 2.8. We clearly see that power consumption increases around night time and decreases during the day – when people are more likely not to be at home. We also clearly see spikes on top of this

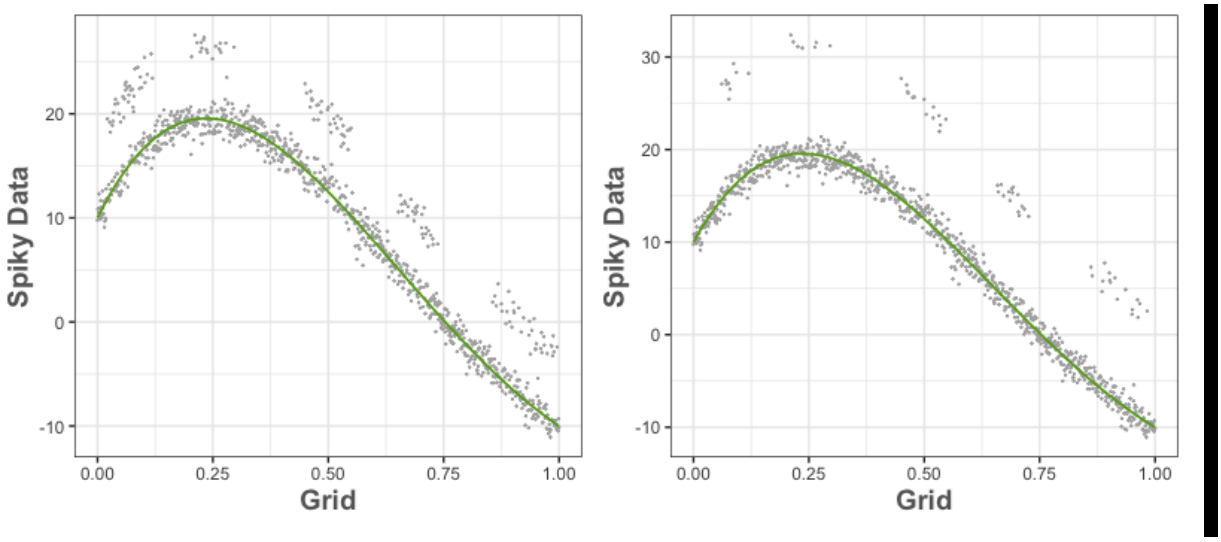

Figure 2.7: *Simulated data with spikes that occur as a nonhomogeneous Poisson process. The green curve represents the true smooth component.* $n = 500, \sigma^* = 1$. *left:* $\alpha^* = 0.85, \mu^* = 7.2$*; right:* $\alpha^* = 0.94, \mu^* = 12$.

trend at night. For this data, the estimated values for $\alpha^*, \mu^*, \sigma_\epsilon^{*2}, \sigma_h^{*2}$ were, respectively, $0.07, 0.78, 0.03, 0.71$ for the first day and $0.05, 0.21, 0.02, 0.71$ for the second. Notably, the shift parameter, expressing the magnitude of the spikes in the two days is substantially different. An interesting way to extend the analysis presented here could be to consider a collection of different days (for one household) and/or different households (for the same day), to understand what range and patterns may exist in their spike behaviors – i.e. in the estimates of the parameters expressing their prevalence, magnitude and variance.

2.6.2 Extreme temperatures in US

For our second application, we considered the time series of the annual heatwave index in the US from 1910 to 2015. This index classifies as a heatwave any period of four or more days with an unusually high average temperature (the threshold for qualification is an average temperature that is expected to occur once every 10 years). The index value is a function of geographical spread and frequency of heatwaves. We also considered the time series of the annual percentage of US land area with unusually high summer temperatures from 1910 to 2015. Probably because geographical spread is part of the definition of the heatwave index, the two time series show rather similar underlying trends. Our procedure was able to detect the 1936 North American heat wave, one of the most intense in modern history, both in terms of heat index and of US area affected

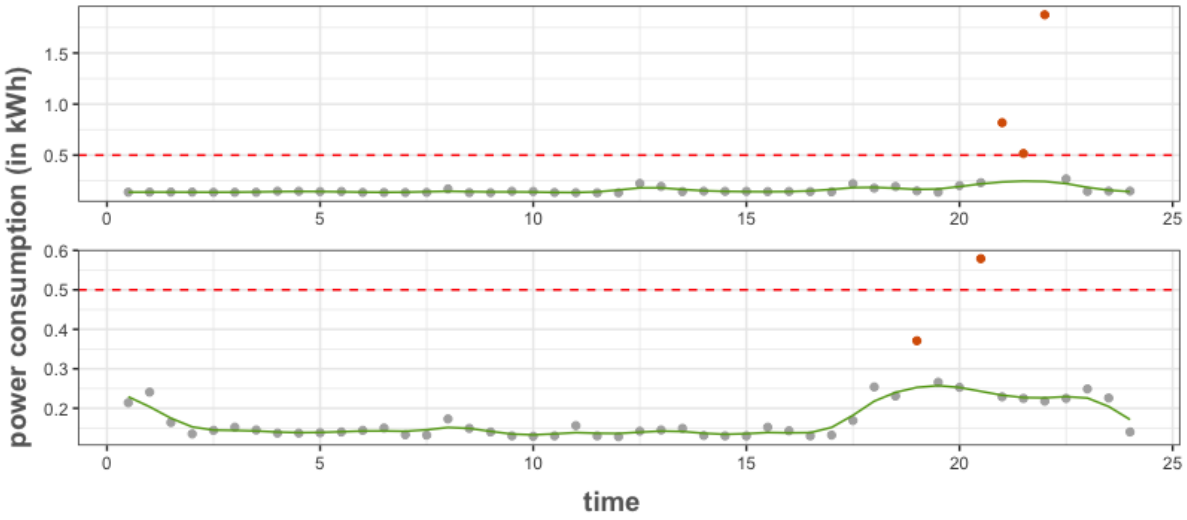

Figure 2.8: *Electricity consumption by a household on Wed, Jul 15, 2009 (top) and Thu, Jul 16, 2009 (bottom). The estimated smooth component is plotted in green, whereas spikes are identified by the red points. The horizontal dashed red lines help visualize the different magnitudes of the flagged spikes on the two days.*

by high temperatures – along with a number of other spikes. Interestingly, none but one such spikes (in terms of US area) concerns the most recent decades. But this is due to the fact that in recent decades the smooth component estimate exhibits an upward trend in both time series.

Using the same settings as in Section 2.6.1, the estimated values for $\alpha^*, \mu^*, \sigma_\epsilon^{*2}, \sigma_h^{*2}$ are respectively $0.08, 30.83, 5.77, 30.79$ for the heatwave index, and $0.07, 23.36, 7.07, 1.41$ for the US area.

2.7 Discussion

We propose a procedure that simultaneously performs estimation of a smooth component and identification of spikes interspersed with such signal. Our procedure uses regularized spline smoothing techniques and the EM algorithm. It is suited to analyze data with discontinuous irregularities superimposed to a smooth curve. This type of data occurs in many applications. We lay out conditions for the procedure to work, and prove asymptotic convergence properties of the EM to a neighborhood of the global optimum under certain restricted conditions. We also demonstrate the effectiveness of the procedure under departures from such restricted conditions through simulations and two real data applications.

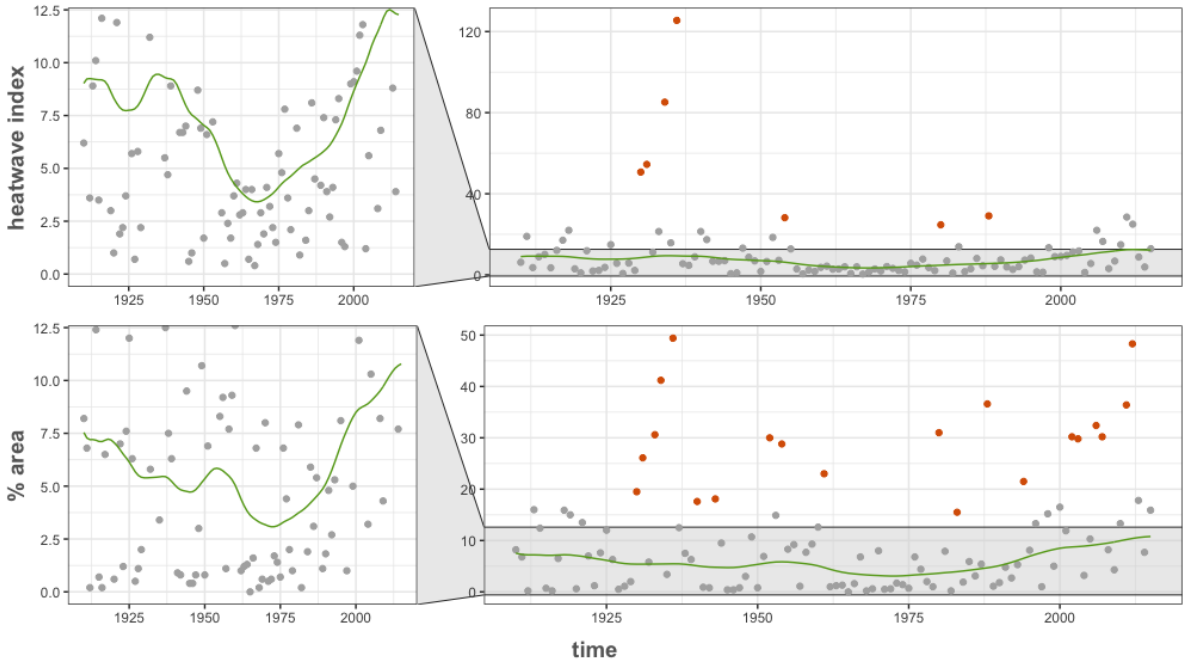

Figure 2.9: *Annual heatwave index in the United States (top) and share of US land with unusually high simmer temperature (bottom). The estimated smooth component is plotted in green, whereas spikes are identified by the red points. For each plot, a vertical zoom (on the left) allows us to better appreciate the upward trend in the smooth components of both signals in recent decades.*

We do tune the smoothing parameter λ, but the grid of explored λ values is usually fixed before hand. This grid choice is critical, and it differed in our simulations and real data applications. For "peculiar" data, even exploring a large grid of λ values might not guarantee good results – leaving a role for prior knowledge and/or preliminary visual inspection of the data to judge how smooth the underlying curve may reasonably be.

Even though our proposal falls within the Functional Data Analysis framework, so far we have only dealt with individual "spiky" curves. In the future, we hope to extend the procedure to allow the analysis of multiple curves that come from the same distribution, calculating the EM log likelihood based on all of them. By borrowing information from a number of curves, approximation of the smooth component will likely improve, and parameters inference will also be more accurate. Moreover, utilizing multiple instances of the same data generating mechanism (if said instances are available) may also be a way to reduce the dependence on the grid choice for the tuning of λ.

Since it separates spikes and smooth component, our procedure could also be used as a way to pre-process functional predictors in a regression context. Instead of introducing in a regression a functional predictor obtained though traditional spline smoothing of the

row data, one could apply our procedure and introduce two distinct predictors; namely, the estimated $\hat{f}$ and, separately, the flagged spike locations. We shall leave these and other possibilities for future work.

Finally, we note that neither in our simulation study nor in the data applications we compared our approach to other existing methods. This is because, to our knowledge, none exists that targets the same type of data structure and problem. Perhaps the closest method known to us is [10], in which the authors also assume an additive structure comprising smooth and "rough" components. However, their"rough" component is assumed to be smooth at finer scales. Thus does the approach in [10] not allow for discontinuous spikes like ours does and, as a consequence, cannot tackle spike identification.

2.8 Acknowledgements

The extreme temperature data was obtained from the National Oceanic and Atmospheric Administration (NOAA) via the Environmental Protection Agency (APE) website. We thank the Irish Social Science Data Archive for having generously provided access to the Smart Meter Electricity data.

Bibliography

[1] Bullmore, E., J. Fadili, M. Breakspear, R. Salvador, J. Suckling, and M. Brammer (2003) "Wavelets and statistical analysis of functional magnetic resonance images of the human brain," *StatisticalMethods in Medical Research*, **12**, pp. 375–399.

[2] Ramsay, J. O. and B. Silverman (2007) *Applied functional data analysis: methods and case studies*, Springer.

[3] Kokoszka, P. and M. Reimherr (2017) *Introduction to Functional Data Analysis*, CRC Press.

[4] Yao, F. and T. C. M. Lee (2006) "Penalized spline models for functional principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), pp. 3–25.

[5] Xiao, L., V. Zipunnikov, D. Ruppert, and C. Crainiceanu (2016) "Fast covariance estimation for high-dimensional functional data," *Statistics and Computing*, **26**(1), pp. 409–421.

[6] Xiao, L., C. Li, W. Checkley, and C. Crainiceanu (2018) "Fast covariance estimation for sparse functional data," *Statistics and Computing*, **28**(3), pp. 511–522.

[7] Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011) "Penalized functional regression," *Journal of Computational and Graphical Statistics*, **20**(4), pp. 830–851.

[8] Wood, S. (2006) "Low-rank scale-invariant tensor product smooths for generalized additive mixed models," *Biometrics*, **62**, pp. 1025–1036.

[9] Nason, G. P. (2008) *Wavelet Methods in Statistics with R*, Springer.

[10] Descary, M. H. and V. M. Panaretos (2019) "Functional data analysis by matrix completion," *The Annals of Statistics*, **47**(1), pp. 1–38.

[11] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977) "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**, pp. 1–38.

[12] Liu, C. and D. B. Rubin (1994) "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence," *Biometrika*, **81**, pp. 633–648.

[13] Louis, T. A. (1982) "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **44**, pp. 226–233.

[14] Meilijson, I. (1989) "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **51**, pp. 127–138.

[15] Meng, X. and D. B. Rubin (1993) "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, **80**, pp. 267–278.

[16] Claeskens, G., T. Krivobokova, and J. D. Opsomer (2009) "Asymptotic properties of penalized spline estimators," *Biometrika*, **96**(3), pp. 529–544.

[17] Eilers, P. and B. Marx (1996) "Flexible smoothing with B-splines and penalties (with Discussion)," *Statistical Science*, **11**, pp. 89–121.

[18] Eilers, P., B. Marx, and M. Durban (2015) "Twenty years of p-splines," *SORT*, **39**(2), pp. 149–186.

[19] Hall, P. and J. D. Opsomer (2015) "Theory for penalised spline regression," *Biometrika*, **92**(1), pp. 105–118.

[20] Kauermann, G., T. Krivobokova, and L. Fahrmeir (2009) "Some asymptotic results on generalized penalized spline smoothing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), pp. 487–503.

[21] Wang, X., J. Shen, and D. Ruppert (2011) "On the asymptotics of penalized spline smoothing," *Electronic Journal of Statistics*, **5**, pp. 1–17.

[22] Xiao, L. (2019) "Asymptotic theory of penalized splines," *Electronic Journal of Statistics*, **13**, pp. 747–794.

[23] Wu, C., C. Yang, H. Zhao, and J. Zhu (2017) "On the Convergence of the EM Algorithm: A Data-Adaptive Analysis," ArXiv:1611.00519v2.

[24] Balakrishnan, S., M. J. Wainwright, and B. Yu (2017) "Statistical guarantees for the EM algorithm: from population to sample-based analysis," *The Annals of Statistics*, **45**(1), pp. 77–120.

[25] de Boor, C. (1978) *A practical guide to splines*, Springer.

[26] O'Sullivan, F. (1986) "A statistical perspective on ill-posed inverse problems," *Statistical Science*, **1**(4), pp. 502–518.

[27] Lazar, N. (2008) *The statistical analysis of functional MRI data*, Springer.

[28] Power, J., K. Barnes, A. Snyder, B. Schlaggar, and S. Petersen (2012) "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *NeuroImage*, **59**(3), pp. 2142–2154.

[29] Satterthwaite, T., D. Wolf, J. Loughead, K. Ruparel, M. Elliott, H. Hakonarson, R. Gur, and R. Gur (2012) "Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth," *NeuroImage*, **60**(1), pp. 623–632.

[30] Van Dijk, K., M. Sabuncu, and R. Buckner (2012) "The influence of head motion on intrinsic functional connectivity MRI," *NeuroImage*, **59**(1), pp. 431–438.

[31] Mowinckel, A., T. Espeseth, and L. Westlye (2012) "Network-specific effects of age and inscanner subjectmotion: a resting-state fMRI study of 238 healthy adults," *NeuroImage*, **63**(3), pp. 1364–1373.

[32] Bright, M. and K. Murphy (2013) "Removing motion and physiological artifacts fromintrinsic BOLD fluctuations using short echo data," *NeuroImage*, **64**, pp. 526–537.

[33] Satterthwaite, T., M. Elliott, R. Gerraty, K. Ruparel, J. Loughead, M. Calkins, S. Eickhoff, H. Hakonarson, R. Gur, R. Gur, and D. Wolf (2013) "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data," *NeuroImage*, **64**, pp. 240–256.

[34] Yan, C., B. Cheung, C. Kelly, S. Colcombe, R. Craddock, A. Di, Q. Li, X. Zuo, F. Castellanos, and M. Milham (2013) "A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics," *NeuroImage*, **76**, pp. 183–201.

[35] Tyszka, J., D. Kennedy, L. Paul, and R. Adolphs (2014) "Largely typical patterns of restingstate functional connectivity in high-functioning adults with autism," *Cerebral Cortex*, **24**(7), pp. 1894–1905.

[36] Fox, M., A. Snyder, J. Vincent, M. Corbetta, D. Van Essen, and M. Raichle (2005) "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," *Proceedings of the National Academy of Sciences - PNAS*, **102**(27), pp. 9673–9678.

[37] Fox, M., M. Corbetta, A. Snyder, J. Vincent, and M. Raichle (2006) "Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems," *Proceedings of the National Academy of Sciences - PNAS*, **103**(26), pp. 10046–10051.

[38] Fox, M., D. Zhang, A. Snyder, and M. Raichle (2009) "The Global Signal and Observed Anticorrelated Resting State Brain Networks," *Journal of Neurophysiology*, **101**(6), pp. 3270–3283.

[39] Weissenbacher, A., C. Kasess, F. Gerstl, R. Lanzenberger, E. Moser, and C. Windischberger (2009) "Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies," *NeuroImage*, **47**(4), pp. 1408–1416.

[40] Patel, A., P. Kundu, M. Rubinov, P. Jones, P. Vértes, K. Ersche, J. Suckling, and E. Bullmore (2014) "A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series," *NeuroImage*, **95**, pp. 287–304.

[41] Mallat, S. and W. Hwang (1992) "Singularity detection and processing with wavelets," *IEEE transactions on information theory*, **38**(2), pp. 617–643.

[42] Mallat, S. (1989) "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE transactions on pattern analysis and machine intelligence*, **11**(7), pp. 674–693.

[43] ——— (2009) *A wavelet tour of signal processing: the sparse way*, Elsevier /Academic Press.

[44] Percival, D. (2000) *Wavelet methods for time series analysis*, Cambridge University Press.

[45] Snoek, L., M. Van der Mieson, T. Beemsterboer, A. Van der Leij, A. Eigenhuis, and H. Scholte (2021) "The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses," *Scientific Data*, **8**(1), p. 85.

[46] Jahfari, S., L. Waldorp, K. R. Ridderinkhof, and H. S. Scholte (2015) "Visual information shapes the dynamics of corticobasal ganglia pathways during response selection and inhibition," *Journal of cognitive neutoscience*, **27**(7), pp. 1344–1359.

[47] Fadili, M. and E. Bullmore (2002) "Wavelet-Generalized Least Squares: A New BLU Estimator of Linear Regression Models with 1/f Errors," *NeuroImage*, **15**, pp. 217–232.

[48] Donoho, D. (1995) "De-noising by soft-thresholding," *IEEE transactions on information theory*, **41**(3), pp. 613–627.

[49] Donoho, D. and I. Johnstone (1995) "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, **90**(432), pp. 1200–1224.

[50] Pizurica, A. and W. Philips (2003) "A versatile wavelet domain noise filtration technique for medical imaging," *IEEE transactions on medical imaging*, **22**(3), pp. 323–331.

[51] Johnstone, I. and S. B.W. (1997) "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **59**(2), pp. 319–351.

[52] Anderson, C., S. Lowen, and P. Renshaw (2006) "Emotional task-dependent low-frequency fluctuations and methylphenidate: Wavelet scaling analysis of 1/f-type fluctuations in fMRI of the cerebellar vermis," *Journal of neuroscience methods*, **151**(1), pp. 52–61.

[53] Ciuciu, P., P. Abry, and B. He (2014) "Interplay between functional connectivity and scale-free dynamics in intrinsic fMRI networks," *NeuroImage*, **95**, pp. 248–263.

[54] Risk, B., D. Matteson, R. Spreng, and D. Ruppert (2016) "Spatiotemporal mixed modeling of multi-subject task fMRI via method of moments," *NeuroImage*, **142**, pp. 280–292.

[55] Costafreda, S., G. Barker, and M. Brammer (2009) "Bayesian wavelet-based analysis of functional magnetic resonance time series," *Magnetic Resonance Imaging*, **27**, pp. 460–469.

[56] Varin, C., N. Reid, and D. Firth (2011) "An overview of composite likelihood methods," *Statistical Sinica*, **21**, pp. 5–42.

[57] Tak, H., K. You, S. Ghosh, B. Su, and J. Kelly (2020) "Data transforming augmentation for heteroscedastic models," *Journal of Computational and Graphical Statistics*, **29**(3), pp. 659–667.

[58] Bertsekas, D. P. (1995) *Nonlinear Programming*, Athena Scientific.

[59] Pollard, D. (1984) *Convergence of stochastic processes*, Springer-Verlag.

[60] Vershynin, R. (2018) *High-dimensional probability*, Cambridge University Press.