Cluster recovery under smooth field modulation with guarantees

Anonymous Author
Anonymous Institution

Abstract

We consider a structured signal recovery problem where one observes a signal with a finite range which has been modulated by a smooth function. Equivalently, the problem is that of clustering with observations that are modified by a smooth field. The problem is motivated by the bias field removal in MRI but is more broadly applicable. We provide theoretical guarantees for accurately recovering the clusters, highlighting key global topological properties that control the recovery process. We also explore deviation bounds on the estimated signal levels that lead to an identifiability result for the case of zero-mean fields, in the limit of large samples. Moreover, we propose a practical estimation algorithm, assuming the field belongs to a reproducing kernel Hilbert space (RKHS), which is a generalization of the kernel ridge regression and k-means algorithms. We demonstrate the effectiveness of the algorithm by a simulation study.

1 INTRODUCTION

The disentanglement of composite mathematical objects is a problem faced across multiple branches of interdisciplinary data analysis. Commonly faced in applied settings is the problem where an observed quantity of interest has been obfuscated or modified by another extraneous, background process. We refer to this action of disentangling signal from background as recovery.

Other well-known problems can be reframed through the lens of recovery. In the case of regression where observations are modeled as $y = f(x_i) + \varepsilon_i$, the goal is to recover some function f belonging to class $\mathcal{F}$ from the random additive noise ε. In the literature of nonparametric regression, structures such as smoothness (Tsybakov, 2009), sparsity (Wainwright, 2009; Bickel et al., 2009), homogeneity (Ke et al., 2015), and piecewise structures (Kim et al., 2009; Tibshirani, 2014) have been suggested for the function class $\mathcal{F}$. For these noisy regression problems, where structure is contrasted from random noise, the focus is less on whether an object can be recovered and more on the fastest rate at which recovery is possible.

Recovery problems where both the signal and background exhibit different structures is somewhat less studied. Some notable example include deconvolution problems (Meister, 2009) where background ε is given a certain density structure and sparse plus low-rank matrix decomposition (Chandrasekaran et al., 2009) where ε is either sparse or low-rank depending on the context.

Of particular interest to us, will be the be the recovery problem of continuous and rank-limited signals. Mathematically this can be expressed as:

$$y_i = f^*(x_i) + \mu^*_{z^*_i}, \quad \text{for } i = 1, \dots, n \tag{1}$$

where $\mu^* \in \mathbb{R}^M$ is a vector of M values, which we refer to as levels, $z^*_i \in [M]$ are labels to corresponding levels of μ^*, and $f^* \in \mathcal{F}_\omega(\mathcal{X})$ is a real-valued, uniformly continuous function with *modulus of continuity* $\omega : [0, \infty) \to [0, \infty)$ over the metric space $(\mathcal{X}, d)$. Note that the original data process need not follow an additive formulation, only that, after an appropriate transformation, Φ, the resulting y_i can be expressed in terms of (1).

The focus of this paper will be to uncover relevant conditions which guarantee identifiable cluster and level recovery, that is, recovery of (μ^*, z^*). Identifiability of $f^* \in \mathcal{F}_\omega(\mathcal{X})$ can be seen as a separate problem, where the goal is to recover the equivalence class

$$[f^*]_n = \big\{ g \in \mathcal{F}_\omega(\mathcal{X}) : \; g(x_i) = f^*(x_i), \; \forall i \in [n] \big\}.$$

Depending on the additional structure of $\mathcal{F}_\omega(\mathcal{X})$, the f^*-identifiability problem may be refined to recovering some appropriate target representative $\bar{f} \in [f^*]_n$.

1.1 Motivating example

We are motivated by the bias field removal problem of magnetic resonance imaging (MRI). In medical imaging, many

Preliminary work. Under review by AISTATS 2023. Do not distribute.

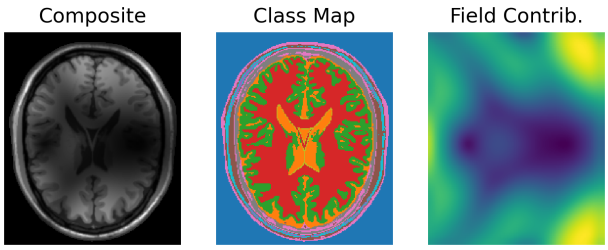


Figure 1: Example of the MRI bias field problem on the BrainWeb phantom.

downstream tasks depend on the correct classification and recovery of tissue quantities. Depending on the patient geometry and radiofrequency (RF) coil setup, modifying fields like the one shown in Figure 1 may cause anomalous tissue identification, ultimately affecting patient outcomes.

The MRI bias field problem can be formulated as such:

$$y(x) = f^*(x) \cdot \mu(x), \quad \text{for } x \in \mathcal{X} \tag{2}$$

where f^* is a positive and smooth multiplicative field on $\mathcal{X}$ and $\mu(x)$ are tissue values at location $x \in \mathcal{X}$. Note for a fixed number of tissues classes M, process (2) can be reformulated in terms of (1) under the transformation $\Phi(x) = \log x$.

The MRI bias field problem has a rich literature of estimation techniques. Earlier iterations focused solely on the presence of the bias field, using nonparametric approaches to filter out low frequency contributions in the Fourier domain of the image (Sled et al., 1998; Tustison et al., 2010). More recent iterations have focused on simultaneous modeling of both tissue and bias contributions through the use of different Bayesian mixture formulations (Wells et al., 1996; Liu and Zhang, 2013; Vinas et al., 2022).

Our contributions In this paper, we make the following contributions to cluster recovery under a smooth field modulation:

1. Identifying conditions under which perfect classification can be achieved in the finite sample setting.

2. Providing a deviation bound on the recovered levels in terms of a novel distance defined on the sample data.

3. Proposing a simple alternating minimization procedure for cluster recovery, that can implemented easily in practice for any function class that admits a representer theorem.

As noted in Corollary 1, the second contribution can be used to produce a limiting identifiability result for the recovered levels in the case of a zero-mean modulating field.

2 METHODS

Given observations $\{y_i\}$ generated by (1), to recover the true triplet (f^*, μ^*, z^*), we consider solving the following optimization problem

$$(\widehat{f}, \widehat{\mu}, \widehat{z}) = \operatorname*{argmin}_{\substack{f \in \mathcal{F}_\omega(\mathcal{X}), \\ \mu \in \mathbb{R}^M,\ z \in [M]^n}} \sum_{i=1}^n (y_i - \mu_{z_i} - f(x_i))^2. \tag{3}$$

In addition, we also consider the *zero-mean* version of (3) where we add a constraint ensuring that f is empirically zero-mean, that is, $\sum_{i=1}^n f(x_i) = 0$.

2.1 Practical versions

Specific function classes of interests will be normable, topological vectors spaces $(\mathcal{F}(\mathcal{X}), \|\cdot\|_{\mathcal{F}})$ that admit a representer theorem for regularized learning problems of the form

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}(\mathcal{X})} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \cdot \psi\big(\|f\|_{\mathcal{F}}\big), \tag{4}$$

for an increasing function $\psi : [0, \infty) \to [0, \infty)$ and regularization parameter $\lambda > 0$. This means that there is a collection of functions $\{\varphi_x, x \in \mathcal{X}\} \subset \mathcal{F}$, only dependent on the structure of $(\mathcal{F}(\mathcal{X}), \|\cdot\|_{\mathcal{F}})$, such that any solution $\widehat{f}$ of (4) can be written as a linear combination of $\varphi_{x_i}, i \in [n]$. Examples of such function spaces include reproducing kernel Hilbert spaces (Schölkopf et al., 2001), Radon spaces (Unser et al., 2017), and locally convex Hausdorff vector spaces (Boyer et al., 2019). The existence of a representer theorem reduces the search over an infinite-dimensional function space to a finite-dimensional one, allowing practical implementations.

Note that for every $\lambda > 0$, there is a corresponding $R > 0$ which allows for an unregularized reformulation of (4) akin to (3) where one replaces $f \in \mathcal{F}_\omega(\mathcal{X})$ with the norm-ball constraint $f \in \mathbb{B}_{\|\cdot\|_{\mathcal{F}}}(R)$.

AltMin Algorithm For practical estimation, we propose a blockwise coordinate descent, with alternating updates on (μ, z) and f. More specfically, in each iteration, the current estimates $(\widehat{f}, \widehat{\mu}, \widehat{z})$, are updated to the new ones $(\widehat{f}^+, \widehat{\mu}^+, \widehat{z}^+)$ by

$$\widehat{f}^+ = \operatorname*{argmin}_{f \in \mathcal{F}(\mathcal{X})} \sum_{i=1}^n (y_i - \widehat{\mu}_{\widehat{z}_i} - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2, \tag{5}$$

$$(\widehat{\mu}^+, \widehat{z}^+) = \operatorname*{argmin}_{\mu \in \mathbb{R}^M,\, z \in [M]^n} \sum_{i=1}^n (y_i - \mu_{z_i} - \widehat{f}^+(x_i))^2. \tag{6}$$

For fixed $\widehat{f}$, optimization (6) can be solved through a k-means procedure. For $\mathcal{F}(\mathcal{X})$ which is an RKHS with kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, optimization (5) has the following

representer solution

$$\widehat{f}^{+} = \sum_{i=1}^{n} \widehat{\alpha}_i^{+} \mathcal{K}(x_i, \cdot), \quad \widehat{\alpha}^{+} := (K + \lambda I_n)^{-1}(y - \widehat{Z}\widehat{\mu})$$

where K is the $n \times n$ kernel matrix with entries $K_{ij} = \mathcal{K}(x_i, x_j)$ and $\widehat{Z} \in \{0,1\}^{n \times L}$ is the label matrix for previous label estimate $\widehat{z}$, constructed by one-hot encoding.

It is worth noting that iterations (5)–(6) reduce to the k-means problem if we fix $\widehat{f} = 0$. Similarly, they reduce to the kernel ridge regression (KRR) when we fix $\widehat{\mu} = 0$, assuming that $\mathcal{F}(\mathcal{X})$ is an RKHS. Thus, the proposed procedure generalizes both the k-means and KRR problems.

3 IDENTIFIABILITY THEORY

We address the identifibaility question for model (1), by showing that the solution of (3) can recover the true clusters and more. We start by recalling the ρ-neighbor graph associated with a point cloud $X = \{x_i\}_{i=1}^n$ in a metric space $(\mathcal{X}, d)$:

Definition 1 (Neighbor Graph)**.** The ρ-neighbor graph $G_\rho(X)$ of X is the graph with vertex set $V = [n]$ and edge set $E = \big\{(i,j) \in [n]^2 : i \neq j \text{ and } d(x_i, x_j) \leq \rho\big\}$.

The ρ-neighbor graph captures some aspect of the topology of the point cloud. In particular, we will see that the connectivity of $G_\rho(X)$, for a sufficiently small ρ, is key in allowing accurate recovery:

Definition 2 (Connectivity)**.** For a point could X, the connectivity is defined as

$$\rho_{\min} := \inf\big\{\rho > 0 : \ G_\rho(X) \text{ is connected}\big\}.$$

Our main result is the following guarantee for lossless cluster recovery:

Theorem 1. *Let $X = \{x_i\}_{i=1}^n$ be points in a metric space $(\mathcal{X}, d)$ and let $\{y_i\}$ follow model (1) with $f^* \in \mathcal{F}_\omega$. Assume that the connectivity $\rho_{\min}$ of X satisfies*

$$\omega(\rho_{\min}) < \frac{1}{2M} \min_{k \neq \ell} |\mu_k^* - \mu_\ell^*|. \tag{7}$$

Then, the labels $\widehat{z}$ produced by (3) have zero misclassification error relative to z^*.

Next, we consider the recovery of the class levels μ_k^*. Let us define the following distance:

Definition 3 (Label Distance)**.** For paired data (X, z^*), let $\mathcal{C}_k = \{i \in [n] : z_i = k\}$ and

$$\delta_{k\ell} := \min_{i \in \mathcal{C}_k, \, j \in \mathcal{C}_\ell} d(x_i, x_j).$$

The label distance for data (X, z^*) is defined by

$$\delta_{\text{lbl}} := \min_{G \in \mathcal{G}_M} \max_{(k,\ell) \in G} \delta_{k\ell}, \tag{8}$$

where $\mathcal{G}_M$ is the set of all connected graphs on $[M]$.

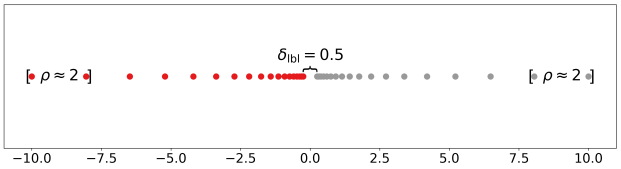


Figure 2: Contrasting example of the ρ-connectivity and label distance δ_{lbl}. The color is used to distinguish the two clusters.

Note that $\delta_{\text{lbl}} \leq \rho_{\min}$. However, as Figure 2 shows, δ_{lbl} can be much smaller than $\rho_{\min}$. Our next result provides guarantees for recovering the levels μ^*, assuming a nearly zero-mean field f^*:

Proposition 1. *Under the assumptions of Theorem 1, let $(\widehat{f}, \widehat{\mu}, \widehat{z})$ be the solution of the zero-mean version of problem (3). Then, we have*

$$\|\mu^* - \widehat{\mu}\|_\infty \leq 2(M-1)\omega(\delta_{\text{lbl}}) + \Big|\frac{1}{n}\sum_{i=1}^{n} f^*(x_i)\Big|. \tag{9}$$

In both Theorem 1 and Proposition 1, we see a recurring theme of deviation bounds given under connectivity constraints. Both $\rho_{\min}$ and δ_{lbl} measure the largest jump required for the vertices of interest, $[n]$ and $[M]$ respectively, to be connected on the sampled data (X, z^*). The presence of any large jump in metric d would allow for multiple $\widehat{f}$ interpolating solutions and thus invite non-identifiability. This is true for both the label and level recovery.

In a related vein, as mentioned earlier, the ρ-connectivity needed in Theorem 1 is in a sense more stringent than the label distance δ_{lbl} the appears in Proposition 1. This could be due to the fact that, while both label and cluster recovery require a notion of connectivity over the data points, candidate labels $\widehat{z}$ are given n degrees of freedom while candidate levels $\widehat{\mu}$ are only given M degrees of freedom for the same sampled data (X, z^*).

As an immediate corollary, we have the following asymptotic consistency result for $\widehat{\mu}$:

Corollary 1. *Consider a data sequence $\{X^{(n)}\}$, with corresponding true labels $\{z^{*(n)}\}$ and class levels $\mu^* \in \mathbb{R}^M$. Assume that the connectivity condition of Theorem 1 is satisfied for all $X^{(n)}, n \in \mathbb{N}$. Let $\delta_{\text{lbl}}^{(n)}$ be the label distance for $(X^{(n)}, z^{*(n)})$ and assume that, as $n \to \infty$,*

$$\delta_{\text{lbl}}^{(n)} = o(1), \quad \frac{1}{n}\sum_{x \in X^{(n)}} f^*(x) = o(1). \tag{10}$$

Then for $(\widehat{f}, \widehat{\mu}, \widehat{z})$ which optimizes the zero-mean version of the recovery problem (3), we have

$$\lim_{n \to \infty} \|\mu^* - \widehat{\mu}\|_\infty = 0.$$

Example 1. Consider the case where $\mathcal{F}$ is an RKHS. Then the natural metric to consider on $\mathcal{X}$ is the so-called *kernel metric*

$$d_{\mathcal{K}}(x,x') = \|\mathcal{K}(x,\cdot) - \mathcal{K}(x',\cdot)\|_{\mathcal{F}} = \sqrt{\mathcal{K}(x,x) - 2\mathcal{K}(x,x') + \mathcal{K}(x',x')}.$$

Using the Cauchy–Schwarz inequality, it is straightforward to show the following Lipschitz property: For any $f \in \mathcal{F}$, $|f(x)-f(x')| \leq \|f\|_{\mathcal{F}}\, d_{\mathcal{K}}(x,x')$ for all $x, x' \in \mathcal{X}$. Letting ω_f denote a modulus of continuity of function f, the above shows that one can take $\omega_f(\rho) = \|f\|_{\mathcal{F}} \cdot \rho$ for all $f \in \mathcal{F}$ in case of an RKHS. This bound might be conservative for specific functions as discussed in Section 4.

4 NUMERICAL EXPERIMENTS

We now present the results of applying variants of ALTMIN algorithm on simulated and real data, and corroborate some the predictions of the theoretical results.

4.1 Simulation Setup

Consider the following M-class data generating process on $\mathcal{X} = [0,1]$ where,

$$z_i^* \sim \text{Unif}([M]), \quad \mu_k^* = k \quad \text{for } k \in [M] \qquad f_\beta^*(x) = 0.75\,\sin(2\pi\beta x). \tag{11}$$

Sampled data will be equally-spaced on $[0,1]$ with $X^{(n)} = \{i/n\}_{i=1}^n$. Shown in Figure 3 are two example processes generated at different (M,β) settings with $n = 75$.

For estimation, we have chosen a Sobolev-1 kernel, namely, $K(x,x') = 1 + \min(x,x')$, due to its sinusoidal eigenfunctions. For the equally-spaced data $X^{(n)}$, the kernel matrix has entries $K_{ij}^{(n)} = \frac{1}{n}(1 + \min(i,j))$.

For the equally-spaced data $X^{(n)}$, the smallest radius ρ that guarantees the connectivity condition of Theorem 2 can be computed as

$$\rho_{\min} = \min_{i \neq j} d_{\mathcal{K}}(x_i, x_j) = \sqrt{(i+1)/n - 2i/n + i/n} = n^{-1/2}.$$

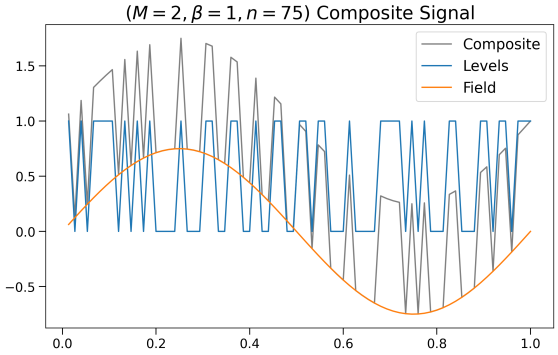
The squared Sobolev-1 norm $||f_\beta^*||_{\mathcal{F}}^2$ can be computed using the inner product $\langle f, g\rangle = \int_0^1 f'(x)\, g'(x)\, dx$. Evaluating this norm gives the following worst-case bound on the modulus of continuity of f_β^*,

$$\omega(\rho_{\min}) \leq \|f_\beta^*\|_{\mathcal{F}} \cdot \rho_{\min} \leq \frac{3\sqrt{2}}{4}\pi\beta n^{-1/2}. \tag{12}$$

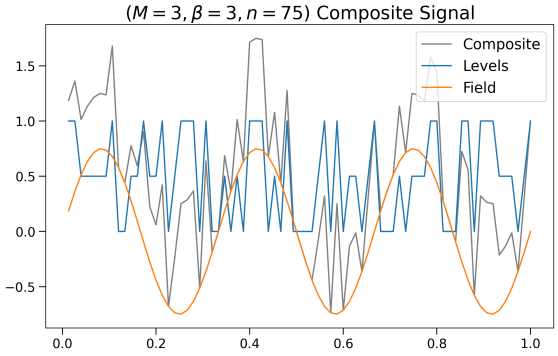
We also consider a noisy recovery setting where observations will be modeled as

$$y_i = \mu_{z_i^*}^* + f_\beta^*(x_i) + \varepsilon_i, \quad \text{for } i \in [n], \tag{13}$$

with noise $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.



(a) Smoother field



(b) Rougher field

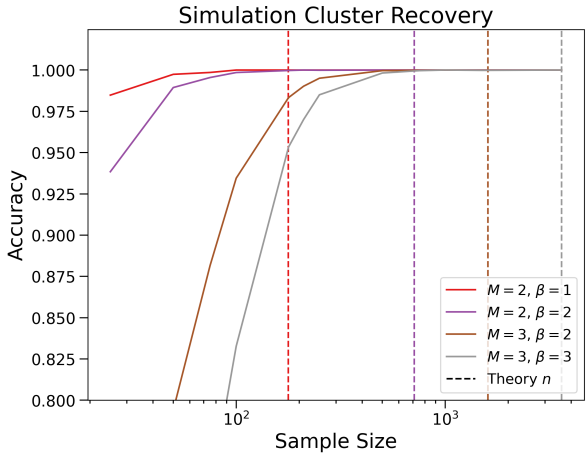
Figure 3: Signal, field and composite observation simulated from (11) for the two and three-class case.

4.2 Simulation Results

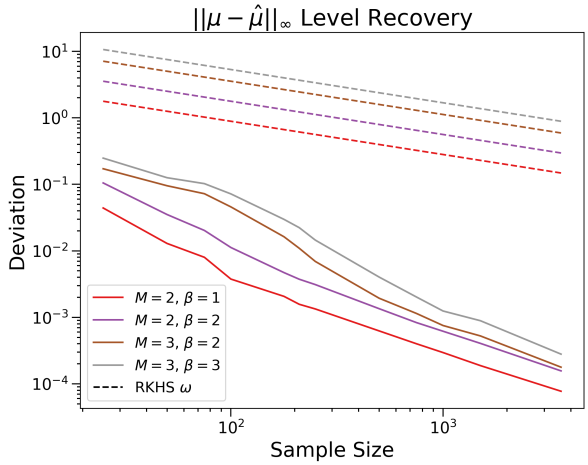
We considered four simulation settings for noiseless recovery: $(M,\beta) = (2,1), (2,2), (3,2)$ and $(3,3)$. The number of data points n was grown in a roughly exponential manner starting at $n = 25$ and ending with $n = 3600$. A total of 100 datasets $X^{(n)}$ were simulated for each n-value. Accuracy and deviation results were calculated using the average of these 100 results.

Noiseless cluster recovery and level deviation results can be found in Figure 4. These figures shows how the simple ALTMIN algorithm can achieve classification and deviation results which match those of Theorem 1 and Proposition 1. In particular, we note that the deviation results of Figure 4b seem to follow an $\mathcal{O}(n^{-1})$ decay rate, which is better than the worst-case $\mathcal{O}(n^{-1/2})$ rate predicted by (12). Interestingly, the modulus of continuity of the particular f_β^* we considered is $\mathcal{O}(n^{-1})$ for equally-spaced data. If we overlay this improved bound in Figure 4b, it is still respected by deviation curves shown, further corroborating Proposition 1. Similar effects in the performance of KRR have been observed for the spectrally truncated case (Amini et al., 2022).

For the noisy recovery results, data setting $(M = 3, \beta = 2)$ is considered at noise levels $\sigma \in \{0, 0.1, 0.2, 0.3\}$. Cluster

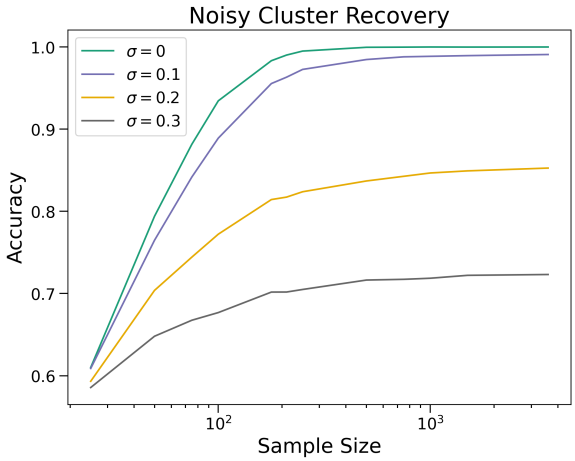


(a) Classification curves

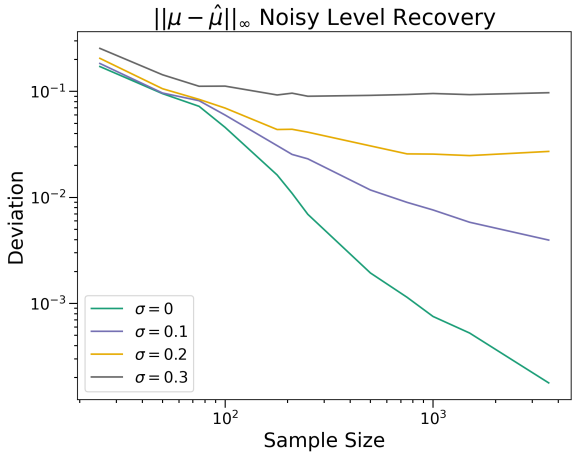


(b) Deviation curves

Figure 4: Results from the ALTMIN Algorithm for noiseless model (11). Theory deviation bounds are shown as dashed lines for each of the different settings.



(a) Classification curves



(b) Deviation curves

Figure 5: Results from the ALTMIN Algorithm for noisy model (13). Setting $(M = 3, \beta = 2)$ is considered at noise levels $\sigma \in \{0, 0.1, 0.2, 0.3\}$.

and level recovery results can be seen in Figure 5. These figures highlight the sensitivity of recovery results for the practical ALTMIN algorithm. Dramatic decreases in accuracy can be seen as noise level is increased linearly from $\sigma = 0$ to $\sigma = 0.2$. These results demonstrate the difficulty of cluster recovery for smooth field modulation under the noisy setting. The small decrease in performance from $\sigma = 0$ to $\sigma = 0.1$ seems to suggest that ALTMIN may still be well-suited for smooth field modulation problems which experience low amounts of noise.

As a passing remark, we bring attention to the subtle change in decay rate past the 100 sample mark for both Figures 4b and 5b. This decay rate change is more prominent in Figure 5b, where noise levels $\sigma \in \{0, 0.1, 0.2\}$ clump around until the 100 sample mark before decaying at different rates. For Figure 4b, there seems to be an interpolation of $\mathcal{O}(n^{-1/2})$ and $\mathcal{O}(n^{-1})$ rates at this point for the $M = 3$ simulation settings. Whether this behavior generalizes or is specific to (11) is left for future work.

4.3 MRI Tissue Recovery

For application, we return to our motivating example of the MRI bias field problem. We consider a simplified 4-class variant modeling the vital areas of the BrainWeb phantom. For the field estimation step, (5), Python smoothing spline routine `csaps` was used, where multidimensional data was fit using an RKHS tensor product of univariate smoothing splines. The relevant smoothing parameter was selected in a post-fitting process. In practice, this parameter would be selected using a validation set for a specific set of coils and MRI scanner. Lastly, initial clusters were chosen through a k-means estimate on the full biased data.

Recovery results for the alternating minimizing procedure on the T1 MRI imaging sequence can be seen in Figure 6 with corresponding error plots shown in Figure 7. Note the

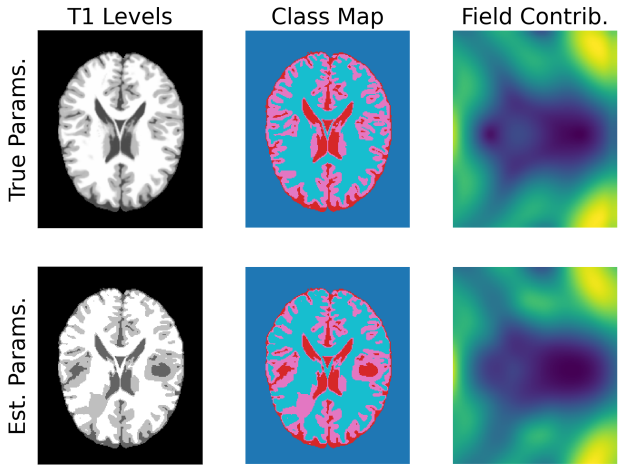


Figure 6: T1 MRI recovery results for the reduced 4-class MRI variant.

tissue discrepancies—that appear in patches—in the estimated parameters in Figure 6 (bottom row). Misclassifications which appear in patches or spatial groups on the patient anatomy have been noted in earlier works of simultaneous tissue and bias modeling (Liu and Zhang, 2013). In practice, these sort of issues may be minimized by baking in additional locality conditions in either the tissue or the bias model, or using multichannel images as we now show.

Modern day MRI machines are able to take multiple imaging sequences at once. It is understood that during scan time, the bias field contribution should not vary much between sequences. If given a total of p sequences, we may consider the following augmented model:

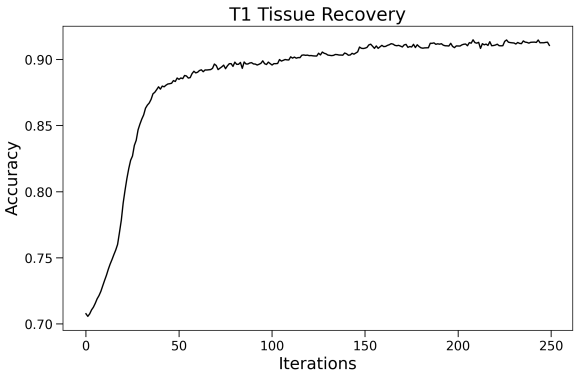
$$y(x) = f^*(x) \cdot \mu(x), \quad \text{for } x \in \mathcal{X}$$

where y is now a vector-valued function in $\mathbb{R}^p$ and levels μ_i take value in $\mathbb{R}^p$. The assumption here is that $f^*(x)$ is still a scalar-valued function.

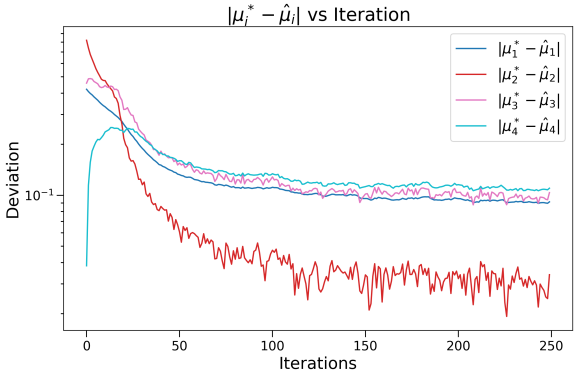
Recovery results for the augmented multi-sequence model are shown in Figures 8 and 9. Important to note is that while the same alternating minimization process is used in the augmented model, the additional sequence data seem to have a significant effect on the performance and stability of the estimation. Note in particular, the absence of anomalous patches in the estimated parameters. These results may give insight as to when the alternating minimizing procedure performs well, ultimately leading to develop error rates for the practical AltMin algorithm.

5 CONCLUSION

In this paper, we formalized the smooth field modulation problem seen in applied fields such as medical imaging. In this formulation, we assumed the modulating field to be



(a) Accuracy vs. iteration.



(b) Level deviations vs. iteration.

Figure 7: T1 MRI cluster and level recovery accuracy. Final cluster accuracy: 91.07%. The initial k-means estimate of the level for cluster 4 corresponded closely with the true T1 level.

uniformly continuous and the signal of interest to be range-limited. We derived recovery conditions, under this model, that make use of the global topology of the data, including connectivity, minimum true level deviation, and the degree of smoothness of the background field.

In addition to theory, we developed a practical algorithm for background fields which admit representer solutions to their regularized learning problems. We have implemented the AltMin algorithm for the case when field f belongs to an RKHS with kernel $\mathcal{K}$. This implementation was then applied to both simulated and real-world data. In the simulation study, AltMin was shown to achieve rates predicted by the identifiability theory in Section 3. We also analyzed the effect of noise on cluster recovery and showed its detrimental effect after a certain noise threshold. For the real-world study, an MRI tissue recovery experiment was conducted showing how the tensor products of smoothing splines can be used to estimate background bias fields. A follow-up study on the MRI recovery experiment was conducted where multiple MRI sequences were considered during the same optimization pass. When given more data for the same bias field, AltMin showed significantly improved clustering performance and overall opti-

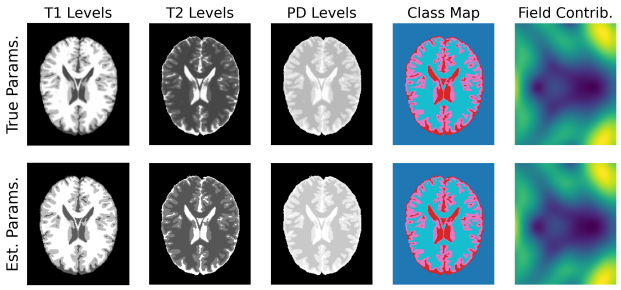


Figure 8: Multi-sequence MRI recovery results for the reduced 4-class MRI variant.

mization stability.

Further work is still required for the noisy version of the recovery problem considered in this paper. In the context of the ALTMIN algorithm, interesting behavior was pointed out in the multi-class and noisy setting. The subtle rate changes in the deviation curves for the low and high sample size regimes were identified. These results when considered in combination with the identifiability theory of Section 3, suggest that ALTMIN is well-suited for data-dense tasks where data is spatially uniform and low in noise. In this light, tasks which are similar in structure to the MRI multi-sequence recovery are good candidates for the application of the ALTMIN algorithm.

6 PROOFS

Any optimal candidate solution $(\widehat{f}, \widehat{\mu}, \widehat{z})$ to (3) must satisfy

$$f^*(x_i) + \mu^*_{z^*_i} = \widehat{f}(x_i) + \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i \in [n]. \tag{14}$$

Since $\widehat{f} - f^* \in \mathcal{F}_{2\omega}(\mathcal{X})$, we may instead consider to analyze the condition

$$g(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i = 1, \ldots, n,$$

for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$. The following result is the main ingredient in the proof of Theorem 1:

Theorem 2. *Suppose for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$ we have $g(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{\widehat{z}_i}$ for all $i \in [n]$ where $z^* = (z^*_i)$ and $\widehat{z} = (\widehat{z}_i)$ both belong to $[M]^n$. Assume the following holds:*

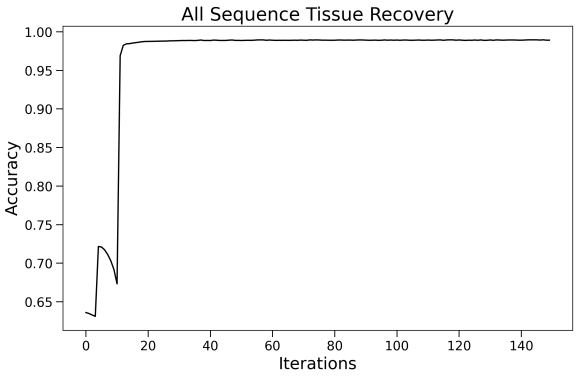
(a) $|\mu^*_k - \mu^*_\ell| \geq \gamma$ for all $k \neq \ell$.

(b) $G_\rho(X)$ is connected for some ρ with $2\,\omega(\rho) < \gamma/M$.

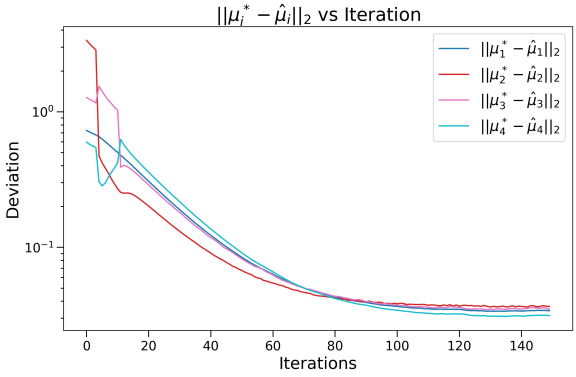
Then for all $i, j \in [n]$ we have

$$\widehat{z}_i = \widehat{z}_j \implies z^*_i = z^*_j. \tag{15}$$

Proof. Start by considering the induction hypothesis that, for any path $\mathcal{P} \subseteq G_\rho(X)$ of length T, all element pairs



(a) Accuracy vs. iteration.



(b) Level deviations vs. iteration.

Figure 9: Multi-sequence cluster and level recovery accuracy. Final cluster accuracy: 98.91%. Level deviations are now measured with respect to the 2-norm distance

$i, j \in \mathcal{P}$ satisfy (15). The base case of $T = 0$ holds trivially with $i = j$.

Throughout the proof, by the label of a node i, we mean its estimated label $\widehat{z}_i$. Consider a general path $\mathcal{P} = \{i_t\}_{t=1}^{T+1}$ of length $T+1$ inside $G_\rho(X)$. As both $\{i_t\}_{t=1}^{T}$ and $\{i_t\}_{t=2}^{T+1}$ are paths of length T, we only need to verify (15) for i_1 and i_{T+1}. Therefore, for our induction step it is sufficient to show that $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$ cannot simultaneously hold for the given assumptions (a) and (b).

For the sake of contradiction, assume $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$. Under this assumption the induction hypothesis guarantees

$$\widehat{z}_{i_t} \neq \widehat{z}_{i_1} \quad \text{for } 1 < t < T+1. \tag{16}$$

Note that if this was not the case with

$$\widehat{z}_{i_1} = \widehat{z}_{i_t} = \widehat{z}_{i_{T+1}} \quad \text{for some } 1 < t < T+1,$$

then the condition $z^*_{i_1} \neq z^*_{i_{T+1}}$ would have caused a contradiction at the earlier induction step $\max\{(T+1)-t, t-1\}$.

Next let $\mathcal{R}$ be the set of labels $\widehat{z}_{i_t}$ on path $\mathcal{P}$. Function $\phi(r)$ will be the index of the last node we see on the path from

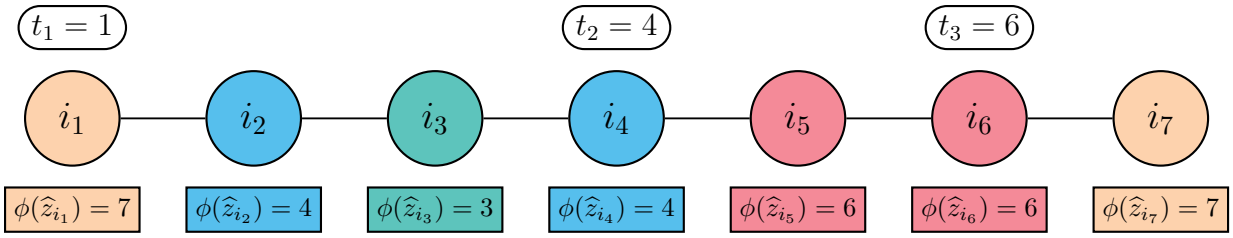


Figure 10: An example demonstrating the $\phi(r)$ function and the construction of the sequence $\{(u_q, v_q)\}_{q=1}^{Q}$ on a 4-class path of length 7. The colors denote the estimated cluster labels. This example has a $\{(u_q, v_q)\}_{q=1}^{Q} = \{(i_1, i_2), (i_4, i_5), (i_6, i_7)\}$ with $Q = 3$.

i_1 to i_{T+1} that has label r, that is,

$$\phi(r) = \max_{t \in [T+1]} \{t : \widehat{z}_{i_t} = r\}.$$

We construct an edge sequence $\{(u_q, v_q)\}_{q=1}^{Q}$—where Q is determined by the construction—recursively as follows: Let $(u_1, v_1) = (i_1, i_2)$ and for $q = 2, \ldots, Q$,

$$(u_q, v_q) = (i_{t_q}, i_{t_q+1}) \quad \text{where} \quad t_q = \phi(\widehat{z}_{v_{q-1}}).$$

The construction continues until $t_Q = T$, so that $(u_Q, v_Q) = (i_T, i_{T+1})$. See Figure 10 for a concrete example. By construction, the labels of v_{q-1} and u_q are the same, while the labels of v_{q-1} and v_q are necessarily different. By this latter property, the labels of $v_1, \ldots, v_{Q-1}$ are distinct elements of $\mathcal{R}$. The added uniqueness condition of (16) gives that the label of v_Q is also distinct from $v_1, \ldots, v_{Q-1}$, hence $Q \le |\mathcal{R}|$.

Using $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$, we obtain the decomposition

$$\widehat{\mu}_{\widehat{z}_{u_1}} - \widehat{\mu}_{\widehat{z}_{v_Q}} = \sum_{q=1}^{Q} (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}}). \tag{17}$$

From the induction hypothesis, $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$ implies $z^*_{v_{q-1}} = z^*_{u_q}$ for $2 \le q \le Q$. This gives the decomposition

$$\mu_{z^*_{u_1}} - \mu_{z^*_{v_Q}} = \sum_{q=1}^{Q} (\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}). \tag{18}$$

Moreover, since u_q and v_q are adjacent on the path, they satisfy $d(x_{u_q}, x_{v_q}) \le \rho$, which by assumption (b) implies

$$|(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})| = |g(x_{u_q}) - g(x_{v_q})| < \gamma/M. \tag{19}$$

By assumption, $\widehat{z}_{u_1} = \widehat{z}_{v_Q}$, hence the LHS of (17) is zero. Then, subtracting decomposition (17) from (18) and using the triangle inequality, we get

$$|\mu^*_{u_1} - \mu^*_{v_Q}| \le \sum_{q=1}^{Q} |(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})| < Q\,\gamma/M,$$

where the second inequality is by (19). If at the same time $z_{i_1} \ne z_{i_{T+1}}$ then $\mu^*_{u_1} \ne \mu^*_{v_Q}$, and by assumption (a), $\gamma \le |\mu^*_{u_1} - \mu^*_{v_Q}|$. Hence,

$$\gamma < Q\,\gamma/M \le |\mathcal{R}|\,\gamma/M.$$

Since $|\mathcal{R}| \le M$, we arrive at a contradiction. This completes the induction step. Applying our induction claim to the connected $G_\rho(X)$ completes the proof. □

Theorem 2 shows that $\widehat{z}$ is a *refinement* of z^*. But since both $\widehat{z}$ and z^* have the same number of classes (M), the classes of $\widehat{z}$ should, in fact, coincide with those of z^*. This is formalized in the following corollary:

Corollary 2. *Suppose every label in $[M]$ is attained by z^* on $[n]$. Then under the conditions of Theorem 2, the misclassification rate between z^* and $\widehat{z}$ is zero.*

Let us now prove Proposition 1. Under the assumptions of Theorem 1, we can relabel the classes of $(\widehat{z}, \widehat{\mu})$ so that $\widehat{z} = z^*$. Then, it follows from (14) that

$$\widehat{f}(x_i) - f^*(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{z^*_i} \quad \text{for all } i \in [n]. \tag{20}$$

Proof of Proposition 1. Let G be minimizing graph in the definition of δ_{lbl}. For every $(k, \ell) \in G$ there exists $i, j \in [n]$ such that $z^*_i = k$, $z^*_j = \ell$ with $d(x_i, x_j) \le \delta_{\mathrm{lbl}}$. As G is connected, every $k, \ell \in [M]$ has a path of nodes $\{(x_{i_t}, x_{j_t})\}_{t=1}^{T}$ such that $z^*_{i_1} = k$ and $z^*_{j_T} = \ell$ and

$$z^*_{j_{t-1}} = z^*_{i_t} \ne z^*_{j_t}$$

with $d(x_{i_t}, x_{j_t}) \le \delta_{\mathrm{lbl}}$. In particular, the condition $z^*_{i_t} \ne z^*_{j_t}$ ensures $T \le M - 1$. With the shorthand $g = \widehat{f} - f^*$ and $\Delta_k = \mu^*_k - \widehat{\mu}_k$, we have $g(x_i) = \Delta_{z^*_i}$ for all $i \in [n]$. Then, the following inequality holds for all $k, \ell \in [M]$,

$$\begin{aligned} |\Delta_k - \Delta_\ell| &\le \sum_{t=1}^{T} |g(x_{i_t}) - g(x_{j_t})| \\ &\le T \cdot 2\omega(\delta_{\mathrm{lbl}}) \le 2(M-1) \cdot \omega(\delta_{\mathrm{lbl}}). \end{aligned}$$

Letting $\pi_k = \frac{1}{n}|\{i : z^*_i = k\}|$ be the proportion of class k,

$$\Big|\Delta_k - \frac{1}{n}\sum_{i=1}^{n} g(x_i)\Big| = \Big|\Delta_k - \sum_{\ell=1}^{M} \pi_\ell \Delta_\ell\Big| \le \sum_{\ell=1}^{M} \pi_\ell |\Delta_k - \Delta_\ell|.$$

Since $\widehat{f}$ is assumed zero-mean, $\frac{1}{n}|\sum_{i=1}^{n} g(x_i)| = \frac{1}{n}|\sum_{i=1}^{n} f^*(x_i)|$. Putting the pieces together, using the the triangle inequality and noting that $\sum_\ell \pi_\ell = 1$ finishes the proof. □

References

Amini, Arash A., Richard Baumgartner, and Dai Feng (2022). *Target alignment in truncated kernel ridge regression*. eprint: https://arxiv.org/abs/2206.14255.

Bickel, Peter J, Ya'acov Ritov, and Alexandre B Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of statistics* 37.4, pp. 1705–1732.

Boyer, Claire et al. (2019). "On Representer Theorems and Convex Regularization". In: *SIAM Journal on Optimization* 29.2, pp. 1260–1281.

Chandrasekaran, Venkat et al. (2009). "Sparse and low-rank matrix decompositions". In: *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 962–967.

Ke, Zheng Tracy, Jianqing Fan, and Yichao Wu (2015). "Homogeneity Pursuit". In: *Journal of the American Statistical Association* 110.509. PMID: 26085701, pp. 175–194.

Kim, Seung-Jean et al. (2009). "ℓ_1 Trend Filtering". In: *SIAM Review* 51.2, pp. 339–360.

Liu, Jun and Haili Zhang (2013). "Image Segmentation Using a Local GMM in a Variational Framework". In: *Journal of Mathematical Imaging and Vision* 46.2, pp. 161–176.

Meister, Alexander (Mar. 2009). *Deconvolution Problems in Nonparametric Statistics*. en. 2009th ed. Lecture notes in statistics. Berlin, Germany: Springer.

Schölkopf, Bernhard, Ralf Herbrich, and Alex J. Smola (2001). "A Generalized Representer Theorem". In: *Computational Learning Theory*. Ed. by David Helmbold and Bob Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 416–426.

Sled, J.G., A.P. Zijdenbos, and A.C. Evans (1998). "A nonparametric method for automatic correction of intensity nonuniformity in MRI data". In: *IEEE Transactions on Medical Imaging* 17.1, pp. 87–97.

Tibshirani, Ryan J. (2014). "Adaptive piecewise polynomial estimation via trend filtering". In: *The Annals of Statistics* 42.1, pp. 285–323.

Tsybakov, Alexandre B (2009). *Introduction to Nonparametric Estimation*. Springer.

Tustison, Nicholas J. et al. (2010). "N4ITK: Improved N3 Bias Correction". In: *IEEE Transactions on Medical Imaging* 29.6, pp. 1310–1320.

Unser, Michael, Julien Fageot, and John Paul Ward (2017). "Splines Are Universal Solutions of Linear Inverse Problems with Generalized TV Regularization". In: *SIAM Review* 59.4, pp. 769–793.

Vinas, Luciano et al. (2022). *LapGM: A Multisequence MR Bias Correction and Normalization Model*. eprint: https://arxiv.org/abs/2209.13619.

Wainwright, Martin J. (2009). "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE Transactions on Information Theory* 55.5, pp. 2183–2202.

Wells, W.M. et al. (1996). "Adaptive segmentation of MRI data". In: *IEEE Transactions on Medical Imaging* 15.4, pp. 429–442.