The sparse-group Beurling-Lasso

Clarice Poon [*] Mohammad Golbabaee [†]

November 28, 2020

Abstract

The Beurling-Lasso is an off-the-grid optimization problem for dealing with non-linear least squares problem, where one aims to recover both mixture weights and the parameters of a nonlinear function. Existing works have been limited to cases where the mixture weights are scalars. In this work, we consider the case of vector-valued weights and extend the Beurling-Lasso to incorporate a sparse-group variation norm. This promotes both sparsity in the number of mixture weights, and also sparsity within each mixture weights. Our main result establishes a numerically verifiable 'certificate' condition which guarantees support stability.

1 Introduction

Many problems in science and engineering require fitting observations to non-linear models. This involves solving the following non-linear inverse problem:

$$X = \sum_{j=1}^{k} \varphi(\theta_j) C_j^\top \in \mathbb{R}^{L\times v}, \quad \text{where} \quad C \geqslant 0$$

where the observations matrix X has v columns representing data population or the number of samples, and L rows representing the dimension of each data sample. The observations are formed by linearly

[*]Dept. of Mathematical Sciences, University of Bath. Email: cmhsp20@bath.ac.uk
[†]Dept. of Computer Science, University of Bath. Email: mg2105@bath.ac.uk

combining the responses of a non-linear function $\varphi(\theta_j)$ for k parameters θ_j in the parameter space $\mathcal{T} \subset \mathbb{R}^d$. The problem is to estimate the underlying parameters θ_j, and the non-negative mixture weights $C_j \in \mathbb{R}^p$.

This problem has numerous applications in biomedical imaging, such as magnetic resonance (MR) spectroscopic imaging [1], quantifying multi-compartment tissue relaxation/decay times in MRI relaxometry [2], the modelling of magneto-encephalogram (MEG) [3], and also in parameter identification in engineering applications [4–6].

In general, the number of components k in the summation is a-priori unknown. Typical approaches would be to guess the number of parameters k and to solve the nonlinear least squares problem [7]

$$\min_{\theta, C} \|X - \sum_{j=1}^{k} \varphi(\theta_j) C_j^\top \|_F^2, \quad \text{where} \quad C \geqslant 0 \tag{1}$$

or to form a discrete dictionary $D_\Theta = (\varphi(\theta))_{\theta \in \Theta}$ by finely discretizing the space $\mathcal{T}$ as Θ, solving

$$\min_{C \geqslant 0} \frac{1}{2} \|X - D_\Theta C^\top\|_F^2 + \alpha \mathcal{R}(C) \tag{2}$$

where $\mathcal{R}$ is a sparsity enforcing regularizer and $\alpha > 0$ is a regularization parameter [8–10]. On one hand, (1) is a nonconvex problem, and requires a-priori assumptions on the number of components, and on the other hand, while (2) is a convex optimization problem, D_Θ is potentially a very large matrix (many columns) and hence, this is computational expensive. Moreover, fine discretizatons typically lead to high coherence in D_Θ (the columns are almost identical), which, even in the presence of the regularizer R, will lead to problems with identifying sparse supports. We refer to [11] for an example where fine discretizations with ℓ_1 regularization will always lead to the recovery of a larger support and fails to identify the sparse support.

In this work, we consider an off-the-grid formulation where we seek to recover the sparse vector-valued measure $\mathbf{m}^* \stackrel{\text{def.}}{=} \sum_{j=1}^k C_j^\top \delta_{\theta_j}$ from observations $X = \Phi \mathbf{m}^* = \sum_{j=1}^k \varphi(\theta_j) C_j^\top \in \mathbb{R}^{L \times v}$, where Φ is a linear operator defined by

$$\Phi : \mathcal{M}_+(\mathcal{T}; \mathbb{R}^{1 \times v}) \to \mathbb{R}^{L \times v}, \mathbf{m} \mapsto \int \varphi(\theta) \mathrm{d}\mathbf{m}(\theta), \quad \text{where} \quad \varphi \in \mathcal{C}(\mathcal{T}; \mathbb{R}^L).$$

Throughout, δ_θ denotes the Dirac mass centred at θ. This approach is introduced in [12, 13], and has been studied in a series of articles.

The idea is that by lifting the problem to the space of measures, one can instead consider the following convex but infinite-dimensional optimization problem (in the scalar-valued setting):

$$\min_{\mathbf{m} \in \mathcal{M}_+(\mathcal{T};\mathbb{R})} \alpha \|\mathbf{m}\|_V + \frac{1}{2}\|x - \Phi \mathbf{m}\|_2^2 \tag{3}$$

given data $x = \Phi \mathbf{m}^* + w = \int \varphi(\theta) \mathrm{d}\mathbf{m}^*(\theta) + w \in \mathbb{R}^L$, and $\alpha > 0$ is a regularisation parameter which balances the data fidelity ℓ_2 term and the variation norm regularisation

$$\|\mathbf{m}\|_V = \sup_{\{\mathcal{A}_i\}_i \in \Pi(\mathcal{T})} \sum_i |\mathbf{m}(\mathcal{A}_i)|.$$

where $\Pi(\mathcal{T})$ is the set of all measurable partitions of $\mathcal{T}$. The authors of [12] named problem (3) the Beurling-Lasso, in acknowledgement to the work [14] of mathematician, Andre Beurling, where the method was first proposed in the context of Fourier measurements. Properties of solutions to (3) have been extensively studied in the literature for *scalar-valued* measures, see [12, 13, 15–18].

We stress that while (3) is related to the Lasso [8] since $\| \sum_j c_j \delta_{\theta_j} \|_V = \sum_j |c_j|$, it is a fundamentally different approach to (2) or (1). First, there is no need to a-priori specify the number of components k. Second, the problem is now convex and allows for deriving strong theoretical results on its recovery properties. In particular, if $\mathbf{m}^*$ is composed of k Diracs, then for sufficiently small noise levels $\|w\|$ and regularization parameter λ, one can prove *sparsistency*, that is, (3) recovers precisely k components [16]. Finally, the formulation (3) lends itself to the development of new algorithms which respect the infinite-dimensional nature of this problem, see for instance: [15,19] for semi-definite programming approaches; [20,21] for conditional gradient descent approaches; and [22, 23] for analysis of particle optimization methods (where one simultaneously optimizes over a fixed number of Dirac positions and mixture weights) as a means of solving (3).

In this work, we study a formulation of (3) in the case of *vector-valued measures*. In the vector-valued setting, there is choice in the underlying norm when defining the variation norm. For this, we introduce the so-called sparse-group variation norm in Section 2. We establish analogous support stability results under a nondegeneracy condition: under sufficiently small noise level and regularisation parameters, one recovers exactly k spikes if the underlying measure defining observations X consist of k spikes. In the following subsection, we describe some practical examples in which the case of vector-valued measures is of interest.

1.1 Examples

The case of complex-valued measures This setting encompasses the case of complex-valued measures, studied in [24], since we have the equivalence between $\mathcal{M}(\mathcal{T};\mathbb{C})$ and $\mathcal{M}(\mathcal{T};\mathbb{R}^2)$. If $\mathbf{m}_\mathbb{C} = \sum_{j=1}^k z_j \delta_{\theta_j}$ where $z_j \in \mathbb{C}$ and we are given measurements $\Phi\mathbf{m}_\mathbb{C}$, then writing $C_j = (\mathrm{Re}(z_j), \mathrm{Im}(z_j))$, we can equivalently consider the recovery of $\mathbf{m}_\mathbb{R} \stackrel{\text{def.}}{=} \sum_j C_j^\top \delta_{\theta_j}$ from $\Phi\mathbf{m}_\mathbb{R} = \begin{bmatrix}\mathrm{Re}(\Phi\mathbf{m}_\mathbb{C}) & \mathrm{Im}(\Phi\mathbf{m}_\mathbb{C})\end{bmatrix}$.

Multicompartment analysis in imaging One of the interests in (2) arises in multicompartment analysis for imaging problems, such as quantitative magnetic resonance imaging [25]: At each image voxel $i = 1, \ldots, v$, one has some time series data of L time points, $x_i \in \mathbb{R}^L$, with

$$x_i = \sum_{s=1}^k c_{i,s}\varphi(\theta_s).$$

Nuclear magnetic resonance (NMR) properties of the tissues at voxel i are driven by mixtures of $\varphi(\theta_s)$, which represents the time dynamics parameterized by these NMR properties. For example, φ could be an exponentially decaying time signal (function) in MRI relaxometry [26] or rather a complicated time response in MR Fingerprinting applications [27]. By aggregating this information, we are led to consider precisely (2) with x_i being the columns of X and $C_s = (c_{i,s})_{i=1}^v$.

Non-stationary modulation processes In [28,29], the authors present a model for non-stationary modulation processes, which is relevant to blind deconvolution or self-calibration problems. The observation model is of the form

$$y = \sum_{j=1}^k c_j H_j \varphi(\theta_j)$$

where $\varphi(\theta_j) \in \mathbb{R}^L$ is an atom from a dictionary (parameterized by θ), and we seek to recover the parameters θ_j, the unknown modulation matrices $H_j \in \mathbb{R}^{L\times L}$, and the unknown coefficients $c_j \in \mathbb{R}$. One might assume that the modulation matrices are represented from a low dimensional subspace so that $H_j = \operatorname{diag}(Bh_j)$ where $B = \begin{bmatrix} b_1 & \dots & b_v \end{bmatrix} \in \mathbb{R}^{L\times v}$ is a known basis and the unknown modulations are simply the vectors $h_j \in \mathbb{R}^v$. The goal is therefore to recover the parameters $\{\theta_j\}_{j=1}^k$ and the modulations

$$Z = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_k \end{bmatrix} \in \mathbb{R}^{v\times k}, \quad \text{where} \quad Z_j \overset{\text{def.}}{=} c_j h_j.$$

By writing $\mathbf{m} = \sum_{j=1}^k c_j h_j^\top \delta_{\theta_j}$, the forward problem becomes

$$y = \mathcal{L}(\Phi \mathbf{m}),$$

where $\mathcal{L} : \mathbb{R}^{L\times v} \to \mathbb{R}^L$ is the linear operator $\mathcal{L}(X) = \sum_{\ell=1}^v \operatorname{diag}(b_\ell) X_\ell$ [1].

2 The sparse-group Beurling-Lasso

Given a measure $\mathbf{m}$ taking values in a normed vector space $\mathcal{V}$ with norm $\|\cdot\|_\mathcal{V}$, its variation is defined as

$$|\mathbf{m}|_\mathcal{V}(\mathcal{V}) \overset{\text{def.}}{=} \sup\left\{ \sum_i \|\mathbf{m}(\mathcal{A}_i)\|_\mathcal{V} \setminus \{\mathcal{A}_i\}_i \text{ partitions } \mathcal{V} \right\}.$$

Let $\beta \in (0,1]$. By considering the variations by vector space $\mathbb{R}^v$ with $\|\cdot\|_1$ and $\|\cdot\|_2$, we define the β-sparse-group norm of $\mathbf{m} \in \mathcal{M}(\mathcal{T}, \mathbb{R}^v)$ as follows

$$\|\mathbf{m}\|_\beta \overset{\text{def.}}{=} (1-\beta)|\mathbf{m}|_1(\mathcal{T}) + \beta\sqrt{v}|\mathbf{m}|_2(\mathcal{T}) \tag{4}$$

Given data $X = \mathbb{R}^{L\times v}$, we consider solutions of the following minimisation problem:

$$\min_{\mathbf{m}\in\mathcal{M}_+(\mathcal{T};\mathbb{R}^p)} \frac{1}{2}\|\Phi\mathbf{m} - X\|_F^2 + \alpha\|\mathbf{m}\|_\beta \tag{$\mathcal{P}_\alpha(X)$}$$

Remark 1. We do not consider $\beta = 0$, since in this case, the problem becomes separable, and one can simply consider v optimisation problems: for $\ell \in [v]$,

$$\mathbf{m}_\ell \in \operatorname*{argmin}_{\mathbf{m}\in\mathcal{M}_+(\mathcal{T};\mathbb{R})} \frac{1}{2}\|\Phi\mathbf{m} - X_\ell\|_2^2 + \alpha|\mathbf{m}|_1 \tag{5}$$

and solutions to $(\mathcal{P}_\alpha(X))$ is simply the measure defined $\mathbf{m}(\mathcal{A}) = (\mathbf{m}_\ell(\mathcal{A}))_{\ell\in[v]}$. Therefore, this is covered by previous studies on Beurling-Lasso.

[1] For simplicity, we do not consider the composition with a linear operator $\mathcal{L}$ here, although our results can be extended to this setting.

Relationship to the sparse-group lasso If $\mathbf{m} = \sum_j C_j^\top \delta_{\theta_j}$ is a sparse measure, then

$$\|\mathbf{m}\|_\beta = (1-\beta)\sum_j \|C_j\|_1 + \beta\sqrt{v}\sum_j \|C_j\|_2.$$

This is precisely the sparse-group regularisation term introduced in [10], with the ℓ_1 term promoting sparsity within each vector C_j, and the ℓ_2 term promoting group sparsity. In this sense, $(\mathcal{P}_\alpha(X))$ can be seen as a continuous extension of the sparse-group lasso.

2.1 Main result

Our main result is a support stability result on the solution of $(\mathcal{P}_\alpha(X))$ under a nondegenerate precertificate assumption which is numerically verifiable. We first describe this precertificate.

Given a sparse measure $\mathbf{m} = \sum_s \delta_{\theta_s} C_s^\top$, we define the vanishing derivatives pre-certificate as follows: Define

$$Q_V \overset{\text{def.}}{=} \operatorname*{argmin}_{Q\in\mathbb{R}^{T\times v}} \left\{ \|Q\|_F \setminus f \overset{\text{def.}}{=} \frac{(\Phi^* Q - (1-\beta))}{\sqrt{v}\beta} \in \mathcal{K} \right\}$$

where $\mathcal{K} \subset \mathcal{C}(\mathcal{T};\mathbb{R}^v)$ is

$$\mathcal{K} \overset{\text{def.}}{=} \left\{ f \setminus \forall s \in [k],\ [f(\theta_s)]_{I_s} = \frac{[C_s]_{I_s}}{\|C_s\|_2}, \nabla \|f_{I_s}\|_2^2(\theta_s) = \mathbf{0}_d \right\}. \tag{6}$$

and $I_s \overset{\text{def.}}{=} \mathrm{Supp}(C_s)$ is the position of the non-zero elements of C_s.

Definition 1. *Define* $\eta_V(\theta) \overset{\text{def.}}{=} \|f_V(\theta)\|^2$*, where* $f_V(\theta) \overset{\text{def.}}{=} \frac{1}{\sqrt{v}\beta}(\Phi^* Q_V(\theta) - (1-\beta))_+$*. We call* η_V *a vanishing derivatives precertificate (with respect to the sparse measure* $\mathbf{m} = \sum_j C_j \delta_{\theta_j}$*) and say it is nondegenerate if*

(i) (non-saturating) $\eta_V(\theta) < 1$ *for all* $\theta \notin \{\theta_s\}_{s=1}^k$*.*

(ii) (curvature) $\nabla^2 \eta_V(\theta_s) \prec 0$ *for all* $s \in [k]$*.*

Note that the vanishing derivatives precertificate depends only on $\{\varphi(\theta_s), \mathbb{J}_\varphi(\theta_s)\}_{s\in[k]}$ and the sign pattern $\{\frac{C_s}{\|C_s\|_2}\}_{s\in[k]}$. As discussed in Section 3.5.1, this precertificate can be computed by solving a linear system and hence, the nondegeneracy condition is numerically verifiable. We refer to [25] for numerical validations of this certificate for the problem of multicomponent analysis in quantitative MRI. When the precertificate is nondegenerate, it corresponds directly to a dual solution of $(\mathcal{D}_\alpha(X))$ when $\alpha = 0$ (See Proposition 1).

Our main result shows that under the assumption that η_V is nondegenerate, we have support stability whenever the noise level is $\varepsilon = \mathcal{O}(\alpha)$ and the regularisation parameter is $\alpha = \mathcal{O}(c_{\min}^2/c_{\max})$. In the following, denote the Jacobian of φ at θ by $\mathbb{J}_\varphi(\theta) \in \mathbb{R}^{T\times d}$ and let $c_{\min} = \min_s \|C_s^*\|$ and $c_{\max} \overset{\text{def.}}{=} \max_s \|C_s^*\|$. We also write $\Theta = \{\theta_s\}_s$.

Theorem 1. *Let $\varepsilon > 0$ and $X = \Phi\mathbf{m}^* + W$ where $W \in \mathbb{R}^{T\times v}$ satisfies $\|W\|_F \leqslant \varepsilon$ and $\mathbf{m}^* = \sum_{s=1}^k C_s^*\delta_{\theta_s^*}$. Assume that*

$$\begin{bmatrix}\varphi(\theta_1^*) & \cdots & \varphi(\theta_k^*) & \mathbb{J}_\varphi(\theta_1^*) & \cdots & \mathbb{J}_\varphi(\theta_k^*)\end{bmatrix}$$

is full rank, and the vanishing derivatives precertificate η_V is nondegenerate with respect to $\mathbf{m}^*$. Then, there exists constants $\rho_1, \rho_2, \rho_3 > 0$ such that for all $\varepsilon/\alpha \leqslant \rho_1$ and $\alpha \leqslant \rho_2 c_{\min}^2/c_{\max}$, $(\mathcal{P}_\alpha(X))$ recovers a unique solution of the form $\sum_{s=1}^k C_s\delta_{\theta_s}$ with

$$\|C^* - C\|_F + c_{\min}\|\Theta^* - \Theta\|_F \leqslant \rho_3\alpha \tag{7}$$

The constant ρ_i for $i = 1, 2, 3$ depend only on $\{\varphi(\theta_s^*), J_\varphi(\theta_s^*)\}_{s\in[k]}$ and the sign pattern $\{C_s^*/\|C_s^*\|_2\}_{s\in[k]}$

2.2 Links to previous works

The work which is closest in nature to this work is [16], where the notion of support stability and sparsistency was studied for deconvolution problems in the case of scalar-valued measures under a nondegeneracy condition (see also [24] for the case of complex-valued measures and the general operator setting). This work can be seen as an extension of their results to the vector-valued setting, where we provide sparsistency results under the corresponding nondegeneracy condition assumption. We also highlight that in [16], the support stability result is non-quantitative (for *sufficiently* small noise, one can guarantee support stability), while in this work, we describe how α should scale with respect to the underlying 'amplitudes' $c_{\min}$ and $c_{\max}$. Our proof is largely inspired by a proof technique introduced in [30].

In the case of vector-valued measures, there is a choice to be made in the definition of the variation norm (i.e. the norm of the underlying vector space). In this work, we investigate the sparse-group norm. In the discrete setting, the sparse-group norm was introduced by [10] for enforcing sparsity within groups, and properties of this norm was studied in [31], where connections to the so-called epsilon-norm [32] were made.

One could of course analyse precise conditions under which the nondegeneracy condition holds, this has been done in the scalar valued setting in [15], [16], and a general result in the multivariate setting was investigated in [18, 24]. Compressed sensing results were also derived in [17] in the univariate random Fourier setting and in [18], for a wide class of operators which encompasses non-translational invariant operators such as the Laplace transform. In general, one requires sufficient separation of the underlying spikes, we expect that similar results can also be attained on our vector-valued measures setting, however, precise analysis of this is beyond the scope of this work.

3 Proof of Theorem 1

3.1 Notations

Given a matrix $Q \in \mathbb{R}^{n\times m}$, let $\mathrm{Vec}_{n,m}(p) \in \mathbb{R}^{nm}$ be its vectorized version with columns stacked vertically, let $\mathrm{Vec}_{n,m}^{-1}$ be the inverse operation, so that $\mathrm{Vec}_{n,m}^{-1}(\mathrm{Vec}_{n,m}(p)) = p$. Given $\beta > 0$, we define the soft-thresholding operator by $\mathcal{S}_\beta : \mathbb{R}^v \to \mathbb{R}^v$ is defined by

$$\mathcal{S}_\beta(\xi)_i = \begin{cases} \xi_i - \beta & \xi_i > \beta, \\ \xi_i + \beta & \xi_i < -\beta, . \\ 0 & |\xi_i| \leqslant \beta. \end{cases}$$

Given a matrix or a tensor, we write $\|\cdot\|$ without subscript to denote the operator norm with respect to the vector norm $\|\cdot\|_2$. Given an index set I and a vector V, we denote by V_I the restriction of V to the index set I. Given a point $x \in \mathbb{R}^n$ and $r > 0$, we denote by $\mathcal{B}(x,r) \overset{\text{def.}}{=} \{z \setminus \|x - z\| < r\}$ the open ball of radius r around x. Given $x \in \mathbb{R}^n$, x_+ is the positive part of x.

Outline of this section Section 3.2 describe the dual problem of $(\mathcal{P}_\alpha(X))$ and Section 3.4 describes how dual solutions can be used to study support stability. These are the analogous results to [16] in the case of vector-valued measures. The main novelty is in Section 3.6 where we prove Theorem 1.

3.2 Duality

To simplify notation, throughout this section and the next, we let $\lambda_1 \overset{\text{def.}}{=} (1-\beta)$ and $\lambda_2 \overset{\text{def.}}{=} \sqrt{v}\beta$, so $\|\mathbf{m}\|_\beta = \lambda_1|\mathbf{m}|_1 + \lambda_2|\mathbf{m}|_2$.

3.3 Variational formulation of the sparse-group norm

We first mention a duality result, described in [32], between the vector norm

$$J(x) \overset{\text{def.}}{=} (1-\varepsilon)\|x\|_1 + \varepsilon\|x\|_2$$

defined for $x \in \mathbb{R}^n$ and $\varepsilon \in (0,1)$, and the so-called ε-norm, which is defined for $\xi \in \mathbb{R}^n$ as $\nu = \|\xi\|_\varepsilon$ is the unique $\nu > 0$ such that

$$\sum_i (|\xi_i| - (1-\varepsilon)\nu)_+^2 - (\varepsilon\nu)^2 = 0.$$

It is shown in [31, Appendix E, Lemmas 1 and 2] (see also [32]) that [2]

$$\begin{aligned}&\left\{x + y \setminus x, y \in \mathbb{R}^d, \|x\|_2 \leqslant \varepsilon\nu, \|y\|_\infty \leqslant (1-\varepsilon)\nu\right\}\\ &= \left\{\xi \in \mathbb{R}^d \setminus \|\xi\|_\varepsilon \leqslant \nu\right\}\end{aligned} \tag{8}$$

and hence, since J is the support function of the set in (8), the dual norm of J is the ε-norm. Moreover, we have the *unique* ε-decomposition

$$\xi = \mathcal{S}_\varepsilon(\xi) + (\xi - \mathcal{S}_\varepsilon(\xi))$$

with $\|\mathcal{S}_\varepsilon(\xi)\|_2 = (1-\varepsilon)\|\xi\|_\varepsilon$ and $\|\xi - \mathcal{S}_\varepsilon(\xi)\|_\infty = \varepsilon\|\xi\|_\varepsilon$, where we recall that $\mathcal{S}_\varepsilon$ is the soft-thresholding operator. Therefore, by considering the dual norms of $|\mathbf{m}|_1$ and $|\mathbf{m}|_2$, the following holds

$$\begin{aligned}\|\mathbf{m}\|_\beta &= \sup_{\sup_\theta \|f(\theta)\|_\infty \leqslant \beta} \langle f, \mathbf{m}\rangle + \sup_{\sup_\theta \|g(\theta)\|_2 \leqslant \lambda_2} \langle g, \mathbf{m}\rangle\\ &= \sup\left\{\langle f+g, \mathbf{m}\rangle \setminus \forall\theta,\ \|f(\theta)\|_\infty \leqslant \lambda_1, \|g(\theta)\|_2 \leqslant \lambda_2\right\}.\end{aligned}$$

[2] which of course can be written as: for all $\lambda_1, \lambda_2 > 0$,

$$\{x + y \setminus \|x\|_2 \leqslant \lambda_1\nu, \|y\|_\infty \leqslant \lambda_2\nu\} = \left\{\xi \setminus \|\mathcal{S}_{\lambda_2}(\xi)\|_2^2 \leqslant (\lambda_1\nu)^2\right\}$$

From (8),

$$\left\{x+y \in \mathbb{R}^d \setminus \|x\|_2 \leqslant \lambda_2, \|y\|_\infty \leqslant \lambda_1\right\} = \left\{\xi \in \mathbb{R}^d \setminus \|\mathcal{S}_{\lambda_1}(\xi)\|_2^2 \leqslant \lambda_2\right\}$$

and hence, we have the following variational formulation of the norm $\|\cdot\|_\beta$:

$$\|\mathbf{m}\|_\beta = \sup_{f \in \mathcal{K}_0} \langle f, \mathbf{m}\rangle \tag{9}$$

where

$$\mathcal{K}_0 = \left\{f \in \mathcal{C}(\mathcal{T}; \mathbb{R}^v) \setminus \sup_{\theta \in \mathcal{T}} \|\mathcal{S}_{\lambda_1}(f(\theta))\|_2^2 \leqslant \lambda_2^2\right\}.$$

Proposition 1 (Dual problem)**.** *For $\alpha > 0$, the dual problem to $(\mathcal{P}_\alpha(X))$ is*

$$\sup_{Q \in \mathcal{K}} \langle X, Q\rangle_F - \alpha \|Q\|_F^2 \tag{$\mathcal{D}_\alpha(X)$}$$

where $\mathcal{K} \subseteq \mathbb{R}^{T \times v}$ is defined as

$$\mathcal{K} \stackrel{\text{def.}}{=} \left\{Q \setminus \sum_{i=1}^{v} ([\Phi^* Q(\theta)]_i - \lambda_1)_+^2 \leqslant \lambda_2^2\right\}$$

The primal and dual problems are related by $\mathbf{m}$ solves $(\mathcal{P}_\alpha(X))$ if and only if $Q = \frac{X - \Phi\mathbf{m}}{\alpha}$ solves $(\mathcal{D}_\alpha(X))$. Moreover, $\Phi^* Q \in \partial\|\mathbf{m}\|_\beta$.

In the case of $\alpha = 0$, the dual of the limit problem

$$\min_{\mathbf{m}} \|\mathbf{m}\|_\beta \ \textit{s.t.}\ \Phi\mathbf{m} = X \tag{$\mathcal{P}_0(X)$}$$

is $(\mathcal{D}_\alpha(X))$ with $\alpha = 0$. Moreover, if Q solves $(\mathcal{D}_\alpha(X))$ and $\mathbf{m}$ solves $(\mathcal{P}_0(X))$, then $\Phi^* Q \in \partial\|\mathbf{m}\|_\beta$

Proof. In (9), we can restrict the set $\mathcal{K}_0$ to positive functions $\mathcal{K}_+ \stackrel{\text{def.}}{=} \mathcal{K}_0 \cap \mathcal{C}(\mathcal{T}; \mathbb{R}_+^v)$ since $\mathbf{m}$ is a positive measure. Therefore, the convex conjugate of $\|\mathbf{m}\|_\beta$ is $\iota_{\mathcal{K}_+}$, the indicator function on the set $\mathcal{K}_+$.

The result now follows by applying the Fenchel-Rockafellar duality theorem [33, Thm 4.2].

□

3.4 Support stability

Given a dual solution Q_α to $(\mathcal{D}_\alpha(X))$, the function

$$f_\alpha(\theta) \stackrel{\text{def.}}{=} \frac{1}{\lambda_2}[(\Phi^* Q_\alpha)(\theta) - \lambda_1]_+$$

characterizes the support of any primal solution $\mathbf{m}_\alpha$ of $(\mathcal{P}_\alpha(X))$ in the following sense:

Lemma 1. *Any solution* $\mathbf{m}_\alpha$ *to* $(\mathcal{P}_\alpha(X))$ *satisfies*

$$\mathrm{Supp}(\mathbf{m}_\alpha) \subseteq \{\theta \in \mathcal{T} \setminus \|f_\alpha(\theta)\| = 1\}.$$

If $\mathbf{m}_\alpha = \sum_s C_s^\top \delta(\theta - \theta_s)$ *is a discrete measure, then for each* s, $\mathrm{Supp}(C_s) \subseteq \{j \in [v] \setminus [\Phi^* Q_\alpha(\theta_s)]_j > \lambda_1\}$ *and* $f_\alpha(\theta_s) = C_s/\|C_s\|_2$.

Proof. We know $\Phi^* Q_\alpha \in \partial|\mathbf{m}|_\beta = \lambda_1 \partial|\mathbf{m}|_1 + \lambda_2 |\mathbf{m}|_2$. From (9), if ξ satisfies $\sum_i (\xi - \lambda_1)_+^2 \leqslant \lambda_2^2$ then $\|\mathcal{S}_{\lambda_1}(\xi)\|_2 \leqslant \lambda_2$ and $\|\xi - \mathcal{S}_{\lambda_1}(\xi)\|_\infty \leqslant \lambda_1$. So $\mathcal{S}_{\lambda_1}(\Phi^* Q_\alpha) \in \lambda_2 \partial|\mathbf{m}|_2$ which gives the first inclusion. For the second,

$$\Phi^* Q_\alpha - \mathcal{S}_{\lambda_1}(\Phi^* Q_\alpha) \in \lambda_1 \partial|\mathbf{m}|_1$$

which means that given $s \in [k]$ and $I_s = \mathrm{Supp}(C_s)$,

$$(\Phi^* Q_\alpha(\theta_s) - \max\{(\Phi^* Q_\alpha)(\theta_s) - \lambda_1, 0\})_{I_s} = \lambda_1 \,\mathrm{sign}(C_s)_{I_s}$$

where given a vector $V \in \mathbb{R}^n$, $\mathrm{sign}(V)_i = V_i/|V_i|$ for $i \in [n]$, where division is in a pointwise sense. If $\Phi^* Q_\alpha(\theta_s)_j < \lambda_1$ for $j \in I_s$, then this equation implies that $\Phi^* Q_\alpha(\theta_s)_j = \lambda_1 \,\mathrm{sign}(C_s)_j$ which is a contradiction. Therefore, $I_s \subset \{j \setminus \Phi^* Q_\alpha(\theta_s)_j > \lambda_1\}$. □

Note that $(\mathcal{D}_\alpha(X))$ has a unique solution, since it can be seen as the projection of X/α onto the closed convex set $\mathcal{K}$. Moreover, the previous lemma shows that its solution characterises the support of any primal solution $\mathbf{m}_\alpha$ of $(\mathcal{P}_\alpha(X))$. Therefore, to understand the structure of solutions to $(\mathcal{P}_\alpha(X))$ with $X = \Phi\mathbf{m} + W$ with $\|W\|_F \leqslant \varepsilon$, it suffices to study the solution of the dual problem $(\mathcal{D}_\alpha(X))$, which we denote by $Q_{\alpha,\varepsilon}$. Following [16], we can show that $Q_{\alpha,\varepsilon}$ has a limit as ε/α and α converge to 0: Define

$$Q_0 \in \operatorname{argmin}\left\{\|Q\|_F \setminus \eta \stackrel{\text{def.}}{=} \Phi^* Q \in \mathcal{K},\ \langle \eta, \mathbf{m}\rangle = \|\mathbf{m}\|_\beta\right\}. \tag{10}$$

Lemma 2. *If* $(\mathcal{D}_\alpha(X))$ *has a solution with* $\alpha = 0$*, then we have* $\|Q_{\alpha,0} - Q_0\|_F \to 0$ *as* $\alpha \to 0$*, and*

$$\|Q_{\alpha,\varepsilon} - Q_{\alpha,0}\|_F \leqslant \varepsilon/\alpha.$$

Proof. The proof is omitted as it is verbatim the proof of Proposition 1 in [16] □

The minimal norm element Q_0 is a solution to the dual problem $(\mathcal{D}_\alpha(X))$ with $\alpha = 0$ and $X = \Phi\mathbf{m}$. Moreover, from Lemma 1, for a discrete measure $\mathbf{m} = \sum_s C_s^\top \delta_{\theta_s}$, we in fact have

$$Q_0 \in \operatorname{argmin} \left\{ \|Q\|_F \setminus \sup_{\theta \in \mathcal{T}} \|f_Q(\theta)\|_2 \leqslant 1 \ f_Q(\theta_s) = \frac{C_s}{\|C_s\|_2} \right\}. \tag{11}$$

where we denote $f_Q \stackrel{\text{def.}}{=} \frac{1}{\lambda_2}(\Phi^* Q - \lambda_1)_+$ inside the constraint.

Definition 2. *Define $f_{Q_0} = \frac{1}{\lambda_2}(\Phi^* Q_0 - \lambda_1)_+$ and $\eta_0(\theta) \stackrel{\text{def.}}{=} \|f_0(\theta)\|_2^2$. We call η_0 is nondegenerate minimal norm certificate with respect to the sparse measure $\mathbf{m} = \sum_{s=1}^k C_s \delta_{\theta_s}$ if η_0 satisfies satisfies*

(i) (non-saturation) $\eta_0(\theta) < 1$ for all $\theta \notin \{\theta_i\}$

(ii) (curvature) $\nabla^2 \eta_0(\theta_s) \prec 0$ for all $s \in [k]$.

Note that by definition, $\eta_0(\theta_s) = 1$ for all $s \in [k]$, so this condition says that η_0 saturates at its maximum value 1 only on $\{\theta_s\}_{s\in[k]}$, and (ii) is a curvature condition on η_0 at these saturation points.

Proposition 2 (Non-quantitiative result on stability)**.** *If η_0 is nondegenerate, then provided that ε/α and α are sufficiently small, the solution to $(\mathcal{P}_\alpha(X))$ is of the form $\mathbf{m}_{\alpha,\varepsilon} = \sum_{j=1}^k \hat{C}_j \delta_{\hat{\theta}_j}$ where $\operatorname{Supp}(\hat{C}_j) \subseteq \operatorname{Supp}(C_j)$.*

Proof. From $(\mathcal{D}_\alpha(X))$, we see that the dual solution to $(\mathcal{D}_\alpha(X))$ can be written as the projection of X/α onto the set $\mathcal{K}$, denote this by $\mathcal{P}_{\mathcal{L}}$. So, from

$$\|Q_{\alpha,\varepsilon} - Q_{\alpha,0}\| \leqslant \|\mathcal{P}_{\mathcal{K}}(X/\alpha) - \mathcal{P}_{\mathcal{K}}((X+W)/\alpha)\| \leqslant \|W\|_F/\alpha,$$

we have that $v_{\alpha,\varepsilon} \stackrel{\text{def.}}{=} \Phi^* Q_{\alpha,\varepsilon} \to v_0 \stackrel{\text{def.}}{=} \Phi^* Q_0$ in the uniform norm as α and ε/α converge to 0. So, if η_0 is non-degenerate, then given any $r > 0$, provided that $\|W\|_F/\alpha$ and α are sufficiently small, letting $\eta(\theta) \stackrel{\text{def.}}{=} \frac{1}{\lambda_2^2}\|(v_{\alpha,\varepsilon}(\theta) - \lambda_1)_+\|^2$, we have $\eta(\theta) < 1$ for all $\theta \notin \cup_j \mathcal{B}(\theta_j, r)$, and for all $\theta \in \mathcal{B}(\theta_j, r)$, $\nabla^2 \eta(\theta) \prec 0$. So, there are at most k points for which $\eta(\theta) = 1$. So, by Lemma 1, given data $X = \Phi\mathbf{m} + W$, we recover at most k components with $\mathbf{m}_{\alpha,\varepsilon} = \sum_{j=1}^k \hat{C}_j \delta_{\hat{\theta}_j}$. Finally, uniform convergence of $v_{\alpha,\varepsilon}$ to v_0 also ensures that $\operatorname{Supp}(\hat{C}_j) \subseteq \operatorname{Supp}(C_j)$ for α and ε/α sufficiently small. $\square$

3.5 Precertificates

To establish support stability, it suffices to show that η_0 is nondegenerate. However, in general, η_0 does not have a closed form expression and can be hard to compute and analyse. It is now standard practice in these situations to consider a *precertificate* η_V [16], a candidate certificate which could be computed by solving a linear system.

Notice that since $\eta_0(\theta) \leqslant 1$ for all θ and $\eta_0(\theta_s) = 1$ for all $s \in [k]$, it is necessary that $\nabla \eta_0(\theta_s) = 0$. Replacing the constraint of $\|f(\theta)\|_2 \leqslant 1$ for all $\theta \in \mathcal{T}$ with $\nabla \|f(\theta_s)_{I_s}\|_2^2 = 0$ for all $s \in [k]$ leads to the definition of η_V in Definition 1 (I_s denotes the support of C_s). Notice that if $\eta_V(\theta) \leqslant 1$ for all $\theta \in \mathcal{T}$, then it follows that $Q_V = Q_0$ is the minimal norm solution from (11). Clearly, if η_V is nondegenerate, then $\eta_V = \eta_0$ is also nondegenerate.

3.5.1 The precertificate as a least squares solution

The attractiveness of Q_V stems from the fact that it is defined via $\sum_s |I_s| + kd$ linear equations, and hence, Q_V can be computed by solving a linear system.

Observe that the constraints $[f(\theta_s)]_{I_s} = \frac{[C_s]_{I_s}}{\|C_s\|_2}$ for all $s \in [k]$ in (6) can be written as

$$\mathcal{P}_{\mathbf{I}} \mathrm{Vec}\left(D_\Theta^\top Q\right) = \mathcal{P}_{\mathbf{I}}[\mathrm{Id}_v \otimes D_\Theta^\top]\mathrm{Vec}(Q) = u_0$$

where

$$u_0 = (\lambda_1 + \lambda_2 [C_s]_{I_s} / \|C_s\|_2)_{s=1}^k \in \mathbb{R}^{\sum_s |I_s|},$$

D_Θ is the matrix with columns $\varphi(\theta_s)$, and $\mathcal{P}_{\mathbf{I}} : \mathbb{R}^{kv} \to \mathbb{R}^{\sum_s |I_s|}$ is the subsampling operator given by which selects the nonzero entries of $\{I_s\}_{s\in[k]}$, so that given a matrix $Z \in \mathbb{R}^{k\times v}$ with s^{th} row $Z_s \in \mathbb{R}^v$ for $s \in [k]$, $\mathcal{P}_{\mathbf{I}}\mathrm{Vec}(Z) = ([Z_s]_{I_s})_{s\in[k]}$.

The constraints $\nabla \|f(\theta_s)_{I_s}\|_2^2 = 0$ for all $s \in [k]$ can be written as

$$\begin{aligned}
\mathbf{0}_d = \lambda_2 \sum_{i\in I_s} f_i(\theta_s) \nabla f_i(\theta_s) &= \frac{1}{\|C_s\|_2} \sum_{i=1}^v (C_s)_i \nabla [\Phi^* Q](\theta_s) \\
&= \mathbb{J}_\varphi(\theta_s)^\top Q \frac{C_s}{\|C_s\|_2} = \frac{1}{\|C_s\|_2} [C_s^\top \otimes \mathbb{J}_\varphi(\theta_s)^\top] \mathrm{Vec}_{T,v}(P)
\end{aligned}$$

We can therefore define the $Tv \times (\sum_{s=1}^k |I_s| + kd)$ matrix

$$\Gamma = \left[(\mathrm{Id}_v \otimes D_\Theta)\mathcal{P}_{\mathbf{I}}^*, \frac{C_1}{\|C_1\|_2} \otimes \mathbb{J}_\varphi(\theta_1), \cdots, , \frac{C_k}{\|C_k\|_2} \otimes \mathbb{J}_\varphi(\theta_k) \right] \tag{12}$$

and write

$$Q_\Theta = \mathrm{Vec}_{T,v}^{-1}\left((\Gamma^*)^\dagger \begin{pmatrix} u_0 \\ \mathbf{0}_{kd} \end{pmatrix}\right).$$

Note that Γ depends on Θ and $\{C_s/\|C_s\|_2\}_s$. To make this dependence clear, we will sometimes write Γ_Θ in place of Γ.

3.6 A quantitative result on support stability

To prove Theorem 1, we rely on the implicit function theorem. The classical implicit function theorem is as follows:

Proposition 3 (Implicit function theorem)**.** *Let $u_0 \in \mathbb{R}^m$, $v_0 \in \mathbb{R}^n$. Let $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be such that $F(u_0, v_0) = 0$ and $\partial_u F(u_0, v_0)$ is invertible. Then, there exists a neighbourhood V of v_0 and a neighbourhood U of u_0, and a continuously differentiable function $G : V \to U$ such that*

$$F(u, v) = 0 \iff u = G(v).$$

Moreover, for all $v \in V$, the Jacobian of G is

$$\mathbb{J}_G(v) = (\partial_u F(G(v), v))^{-1}\,\partial_v F(G(v), v).$$

Typical quantitative versions of the implicit function theorem require showing invertibility of $\partial_u F(u, v)$ and obtaining norm bounds on the partial derivatives of F in some neighbourhood of U of u_0 and V of v_0. A quantitative version is proved in [30, Section 4.3], which requires to look at $\partial_u F(u, v)$ only when $F(u, v) = 0$. We present their arguments below and restate their result in greater generality.

Proposition 4. *Let $n, m, k \in \mathbb{N}$ with $k < m$, and $r_a, r_\theta, R > 0$. Let $v_0 \in \mathbb{R}^n$, $u_0 = (a_0, \theta_0) \in \mathbb{R}^k \times \mathbb{R}^{m-k}$ and let $U_0 = \mathcal{B}_{r_a}(a_0) \times \mathcal{B}_{r_\theta}(\theta_0) \subset \mathbb{R}^k \times \mathbb{R}^{m-k}$ be an open neighbourhood of u_0. Let $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be such that $F(u_0, v_0) = 0$ and for all $u \in U_0$ and $v \in \mathcal{B}(v_0, R)$, $F(u, v) = 0$ implies that the following two conditions hold:*

(i) $\partial_u F(u, v)$ is invertible

(ii) $J \overset{\text{def.}}{=} \partial_u F(u, v)^{-1}\partial_v F(u, v)$ satisfies $\|P_a J\| \leqslant M_a$ and $\|P_\theta J\| \leqslant M_\theta$, for some $M_a, M_\theta > 0$.

Then, the conclusions of Proposition 3 hold with

$$V \supset \mathcal{B}(v_0, \min(r_a/M_a, r_\theta/M_\theta, R))$$

and $\|P_a \mathbb{J}_G(v)\| \leqslant M_a$, $\|P_\theta \mathbb{J}_G(v)\| \leqslant M_\theta$. Here, we denote $P_a J = (J_i)_{i=1}^k$ and $P_\theta J = (J_i)_{i=k+1}^m$.

Proof. Define $V^* \overset{\text{def.}}{=} \bigcup_{V\in\mathcal{V}} V$ where $\mathcal{V}$ is the collection of all open sets such that

(I) $v_0 \in V$

(II) V is star-shaped with respect to v_0,

(III) $V \subset \mathcal{B}(v_0, R)$

(IV) there exists a $\mathcal{C}^1$ function $G : V \to \mathbb{R}^m$ such that $G(v_0) = u_0$ and for all $v \in V$, $F(G(v), v) = 0$.

(V) $G(v) \subset U_0$.

Note that $\mathcal{V}$ is non-empty since we can apply the implicit function theorem to F at u_0, v_0 to obtain a set V and function G which satisfies (I) to (V). The collection $\mathcal{V}$ is stable by union and we can define G^* on V^* by

$$G^*(v) = G(v), \qquad \text{if} \quad v \in V,\ V \in \mathcal{V},\ G \text{ is the corresponding function.}$$

We simply need to show that $V^* \supset \mathcal{B}(v_0, \min\{r_a/M_a, r_\theta/M_\theta, R\})$.

Let $v \in V$ be of norm 1, and define

$$r \overset{\text{def.}}{=} \sup\left\{r > 0 \setminus v_0 + rv \in V^*\right\}.$$

Then, $r \in (0, R]$ by (III). Assume that $r < R$. Let $v_r \overset{\text{def.}}{=} v_0 + rv \in V^*$ and we can define $G^*(v_r) = \lim_{r'\to r} G^*(v_0 + r'v)$. Since $G^*(v_0 + r'v) \in U_0$ for all $r' < r$, we have $u_r \overset{\text{def.}}{=} G^*(v_r) \in \overline{U_0}$. We claim that $u_r \in \partial U_0$ is on the boundary. Suppose $u_r \in U_0$. Then, by assumption, $F(G^*(v_r), v_r) = 0$ and $\partial_u F(G^*(v_r), v_r)$ is invertible, we can therefore apply the IFT to construct neighbourhoods U' around $G^*(v_r)$, V' around v_r to define a $\mathcal{C}^1$ function $G : V' \to \mathbb{R}^m$ such that $G(v_r) = G^*(v_r)$ and for all $v \in V'$, $F(G(v), v) = 0$. We can therefore extend the set V^* to V_r so that V_r contains V', but this would mean that $V^* \subsetneq V_r$. This is a contradiction to the maximality of V^*. So, $u_r \in \partial U_0$. In particular, either $P_a(u_r) \in \partial B(a_0, r_a)$ or $P_\theta(u_r) \in \partial B(\theta_0, r_\theta)$.

Note that for all $t \in [0, 1)$, by (II), $v_0 + trv \in V^* \subset \mathcal{B}(v_0, R)$, so $G^*(v_0 + trv) \in U_0$. Moreover, the Jacobian of G^* at $v_0 + trv$ is $\partial_u F(u, v)^{-1}\partial_v F(u, v)$. So, by assumption (ii), $\|P_a \mathbb{J}_{G^*}(v_0 + trv)\| \leqslant M_a$ and $\|P_\theta \mathbb{J}_{G^*}(v_0 + trv)\| \leqslant M_\theta$. Therefore, either $P_a(u_r) \in \partial B(a_0, r_a)$ and

$$r_a \leqslant \|P_a(G^*(v_r) - u_0)\| = \left\|\int_0^1 P_a \mathbb{J}_{G^*}(v_0 + trv)(rv)\mathrm{d}t\right\| \leqslant M_a r$$

or $P_\theta(u_r) \in \partial B(\theta_0, r_\theta)$ and

$$r_\theta \leqslant \|P_\theta(G^*(v_r) - u_0)\| = \left\| \int_0^1 P_\theta \mathbb{J}_{G^*}(v_0 + trv)(rv)\mathrm{d}t \right\| \leqslant M_\theta r$$

Therefore, $r \geqslant \min\left(\frac{r_a}{M_a}, \frac{r_\theta}{M_\theta}, R\right)$.

$\square$

Proof of Theorem 1. The goal is to define, given noise W and regularisation parameter α, a $\mathcal{C}^1$ function $G : (W, \alpha) \mapsto (C, \Theta)$ such that (C, Θ) corresponds to a solution of $(\mathcal{P}_\alpha(X))$ with data

$$X = \Phi \mathbf{m}^* + W = D_{\Theta^*}(C^*)^\top + W.$$

We first use the implicit function theorem to define such a G, then show that it does indeed define a solution to $(\mathcal{P}_\alpha(X))$.

To this end, let $N = \sum_s |I_s|$ and define a function

$$F : \mathbb{R}_+^N \times \mathcal{T}^k \times \mathbb{R}^{T \times v} \times \mathbb{R}_+ \to \mathbb{R}^{N+kd}$$

so that given $C = \{C_s\}_{s \in [k]}$ with $C_s \in \mathbb{R}^{|I_s|}$, $\Theta \in \mathcal{T}^k$, $W \in \mathbb{R}^{T \times v}$ and $\alpha \in \mathbb{R}_+$,

$$F(C, \Theta, W, \alpha) = \begin{bmatrix} (g_s(C, \Theta, W, \alpha))_{s=1}^k \\ (h_s(C, \Theta, W, \alpha))_{s=1}^k \end{bmatrix}$$

where $g_s(C, \Theta, W, \alpha) \in \mathbb{R}^{|I_s|}$ and $h_s(C, \Theta, W, \alpha) \in \mathbb{R}^d$ are given by

$$g_s(C, \Theta, W, \alpha)^\top = \left(\varphi(\theta_s)^\top [D_\Theta \bar{C}^\top - D_{\Theta^*}(C^*)^\top - W]\right)_{I_s} + \alpha \left(\lambda_1 + \lambda_2 \frac{C_s^\top}{\|C_s\|}\right)$$

and

$$h_s(C, \Theta, W, \alpha) = J_\varphi(\theta_s)^\top \left(D_\Theta \bar{C}^\top - D_{\Theta^*}(C^*)^\top - W\right) \frac{\bar{C}_s}{\|C_s\|_2}.$$

Here, $\bar{C} \in \mathbb{R}^{v \times k}$ is the matrix with s^{th} column satisfying $(\bar{C}_s)_{I_s} = C_s$ and $(\bar{C}_s)_{I_s^c} = 0$. Observe that if (C, Θ) correspond to a solution of $(\mathcal{P}_\alpha(X))$ with data $X = D_{\Theta^*}(C^*)^\top + W$, then $F(C, \Theta, W, \alpha) = 0$, since $(g_s)_s = 0$ correspond to the condition that the dual certificate should take values $C_s/\|C_s\|$ on the support θ_s, and $(h_s)_s = 0$ correspond to the condition that the gradient of the dual certificate is 0. Note in particular that $F(C^*, \Theta^*, 0, 0) = 0$.

The partial derivatives of $g \stackrel{\text{def.}}{=} (g_s)$ and $h \stackrel{\text{def.}}{=} (h_s)$ are as follows: Define

$$Z \stackrel{\text{def.}}{=} D_\Theta \bar{C}^\top - D_{\Theta^*}(C^*)^\top - W, \tag{13}$$

then

$$\partial_c g = \mathcal{P}_{\mathbf{I}} \left(\mathrm{Id}_v \otimes \Phi_\Theta^\top D_\Theta \right) \mathcal{P}_{\mathbf{I}}^* + \alpha\lambda_2 \operatorname{diag} \left(\frac{1}{\|C_s\|} \mathrm{Id}_{|I_s|} - \frac{C_s C_s^\top}{\|C_s\|^3} \right)_{s\in[k]}$$

$$\partial_\theta g = \operatorname{diag}([Z_{(:,I_s)}]^\top \mathbb{J}_\varphi(\theta_s))_{s\in[k]} + \left(C_j \varphi(\theta_s)^\top \mathbb{J}_\varphi(\theta_j) \right)_{s,j\in[k]}$$

$$\partial_\alpha g = \left(\lambda_1 + \lambda_2 \frac{C_s}{\|C_s\|} \right)_{s\in[k]}$$

$$\partial_w g = \mathcal{P}_{\mathbf{I}}(\mathrm{Id}_v \otimes \Phi_\Theta^\top)$$

Let $\mathbb{H}_\varphi(\theta)^\top \in \mathbb{R}^{d\times d\times T}$ so that is (i,j,n) entries with $i,j \in [d]$ for the Hessian of $\varphi_n(\theta_s)$. So, given a vector $z \stackrel{\text{def.}}{=} (z_n)_{n=1}^T$, $\mathbb{H}_\varphi(\theta)^\top z = \sum_{n=1}^T z_n \nabla^2 \varphi_j(\theta) \in \mathbb{R}^{d\times d}$. Then,

$$\begin{aligned}\partial_c h = &\operatorname{diag} \left(\mathbb{J}_\varphi(\theta_s)^\top Z_{(:,I_s)} \left(\frac{1}{\|C_s\|_2} \mathrm{Id}_{|I_s|} - \frac{C_s C_s^\top}{\|C_s\|^3} \right) \right)_{s\in[k]} \\ &+ \left([\frac{1}{\|C_s\|_2} C_s^\top \otimes \mathbb{J}_\varphi(\theta)^\top][\mathrm{Id}_v \otimes D_\Theta] \right)_{s\in[k]}\end{aligned}$$

$$\partial_\Theta h = \operatorname{diag} \left(\mathbb{H}_\varphi(\theta_s)^\top Z \frac{\bar{C}_s}{\|C_s\|_2} \right)_s + \left(\frac{1}{\|C_j\|_2} \mathbb{J}_\varphi(\theta_j)^\top \mathbb{J}_\varphi(\theta_s) \bar{C}_j^\top \bar{C}_s \right)_{j,s\in[k]}$$

$$\partial_\alpha h = \mathbf{0}_{kd}$$

$$\partial_w h = -\left(\frac{1}{\|C_s\|_2} C_s^\top \otimes \mathbb{J}_\varphi(\theta_s)^\top \right)_{s\in[k]}$$

We therefore have

$$\partial_{(C,\Theta)} F = \left(\Gamma_\Theta^\top \Gamma_\Theta + Y \right) \begin{pmatrix} \mathrm{Id}_N & \mathbf{0}_{N\times kd} \\ \mathbf{0}_{kd\times N} & \operatorname{diag}\left(\|C_s\|_2\right)_{s=1}^k \otimes \mathrm{Id}_d \end{pmatrix}$$

where

$$Y \stackrel{\text{def.}}{=} \begin{pmatrix} \alpha\lambda_2 \operatorname{diag}\left(\frac{1}{\|C_s\|} \mathrm{Id}_{|I_s|} - \frac{C_s C_s^\top}{\|C_s\|^3} \right) & \operatorname{diag}\left(\frac{1}{\|C_s\|_2} \mathbb{J}_\varphi(\theta_s)^\top Z_{(:,I_s)} \right)_{s\in[k]}^\top \\ \operatorname{diag}\left(\mathbb{J}_\varphi(\theta_s)^\top Z_{(:,I_s)} \frac{1}{\|C_s\|_2} \left(\mathrm{Id}_{|I_s|} - \frac{C_s C_s^\top}{\|C_s\|_2^2} \right) \right)_{s\in[k]} & \operatorname{diag}\left(\mathbb{H}_\varphi(\theta_s)^\top Z \frac{C_s}{\|C_s\|_2^2} \right)_s \end{pmatrix},$$

and

$$\partial_{(\alpha,W)} F = \left[\begin{pmatrix} \left(\lambda_1 + \lambda_2 \frac{C_s}{\|C_s\|^2} \right)_s \\ \mathbf{0}_{kd} \end{pmatrix}, \quad \Gamma_\Theta^\top \right]$$

Application of implicit function theorem to obtain a candidate solution.

To apply the quantitative implicit function theorem, we first bound $\|\Gamma_\Theta\|$ and $\|Y\|$: Define

$$S_s \stackrel{\text{def.}}{=} C_s/\|C_s\|_2 \quad \text{and} \quad S_s^* \stackrel{\text{def.}}{=} C_s^*/\|C_s^*\|_2$$

and write $S = \{S_s\}_{s=1}^k$ and $S^* = \{S_s^*\}_{s=1}^k$.

i) **Bound on $\|\partial_{\alpha,W} F\|$:**

Note that by Taylor's theorem, $\|D_\Theta - D_{\Theta^*}\| \leqslant \|\Theta - \Theta^*\|_F \max_\theta \|\mathbb{J}_\varphi(\Theta)\|$ and

$$\begin{aligned}&\|S_s \otimes \mathbb{J}_\varphi(\theta_s) - S_s^* \otimes \mathbb{J}_\varphi(\theta_s^*)\| \leqslant \|\left(S_s - S_s^*\right) \otimes \mathbb{J}_\varphi(\theta_s)\| + \|S_s^* \otimes \left(\mathbb{J}_\varphi(\theta_s^*) - \mathbb{J}_\varphi(\theta_s)\right)\| \\ &\leqslant \|S_s - S_s^*\|_2 \max_\theta \|\mathbb{J}_\varphi(\theta)\| + \|\theta_s - \theta_s^*\| \|S_s^*\|_2 \max_\theta \|\mathbb{H}_\varphi(\theta)\|.\end{aligned}$$

Therefore

$$\begin{aligned}\|\Gamma_\Theta - \Gamma_{\Theta^*}\|^2 &\leqslant \|D_\Theta - D_{\Theta^*}\|^2 + \sum_s \|S_s \otimes \mathbb{J}_\varphi(\theta_s) - S_s^* \otimes \mathbb{J}_\varphi(\theta_s^*)\|^2 \\ &\leqslant \|\Theta - \Theta^*\|_F^2 \max_\theta \|\mathbb{J}_\varphi(\Theta)\|^2 + \|S - S^*\|_F^2 \max_\theta \|\mathbb{J}_\varphi(\theta)\|^2 \\ &\quad + \|\Theta - \Theta^*\|^2 \max_\theta \|\mathbb{H}_\varphi(\theta)\|^2 \\ &\leqslant A_1^2 \left(\|S - S^*\|_F^2 + \|\Theta - \Theta^*\|_F^2\right)\end{aligned}$$

where

$$A_1^2 \stackrel{\text{def.}}{=} \max_\theta \|\mathbb{H}_\varphi(\theta)\|^2 + \max_\theta \|\mathbb{J}_\varphi(\Theta)\|^2.$$

We can apply the bounds in i) to deduce that

$$\|\partial_{\alpha,w} F\| \lesssim \|\Gamma_{\Theta^*}\| + A_1 \left(\|S - S^*\|_F + \|\Theta - \Theta^*\|_F\right) = \mathcal{O}(1) \tag{14}$$

ii) Bounds for $\partial_{(C,\Theta)} F$ when $F(C, \Theta, W, \alpha) = 0$. We first bound $\|Y\|$:

$$\|Y\| \lesssim \max_s \{\frac{1}{\|C_s\|_2}\} \cdot \max_s \{\alpha\lambda_2, \|\mathbb{J}_\varphi(\theta_s)^\top Z\|, \|\mathbb{H}_\varphi(\theta_s)^\top Z S_s\|\}$$

Let $U \stackrel{\text{def.}}{=} \begin{pmatrix}(\lambda_1 + \lambda_2 C_s/\|C_s\|_2) \\ \mathbf{0}_{kd}\end{pmatrix}$. Then, since $F(C, \Theta, W, \alpha) = 0$,

$$\Gamma_\Theta^\top Z + \alpha U = 0.$$

By applying $\Gamma_\Theta(\Gamma_\Theta^\top \Gamma_\Theta)^{-1}$ to both sides, we obtain

$$\begin{aligned}0 &= Z - \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} Z + \alpha(\Gamma_\Theta^\top)^\dagger U \\ &= Z + \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \Gamma_{\Theta^*} \begin{pmatrix} C^* \\ \mathbf{0}_{kd} \end{pmatrix} + \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} W + \alpha(\Gamma_\Theta^\top)^\dagger U\end{aligned}$$

Therefore,

$$\|Z\| \leqslant \|\mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} D_{\Theta^*} \begin{pmatrix} C^* \\ \mathbf{0}_{kd} \end{pmatrix}\| + \|W\| + \alpha\|(\Gamma_\Theta^\top)^\dagger U\|$$

Note that

$$\begin{aligned}&\mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} D_{\Theta^*} C^* = \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \varphi\left(\theta_s^*\right) \left(C_s^*\right)^\top \\ &= \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \left(\varphi\left(\theta_s\right) \left(C_s^*\right)^\top + (\theta_s - \theta_s^*) \mathbb{J}_\varphi\left(\theta_s\right) \left(C_s^*\right)^\top\right) + \mathcal{O}(c_{\max}\|\Theta - \Theta^*\|_F^2) \\ &= \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \left(\varphi\left(\theta_s\right) \left(C_s^*\right)^\top + (\theta_s - \theta_s^*) \mathbb{J}_\varphi\left(\theta_s\right) C_s^\top\right) \\ &\qquad + \mathcal{O}(\|C^* - C\|_F \|\Theta - \Theta^*\|_F) + \mathcal{O}(\|\Theta - \Theta^*\|_F^2 c_{\max}) \\ &= \mathcal{O}(\|\Theta - \Theta^*\|_F^2 c_{\max} + \|C^* - C\|_F \|\Theta - \Theta^*\|_F)\end{aligned}$$

Moreover, $(\Gamma_\Theta^\top)^\dagger U = Q^* + \mathcal{O}(\|\Theta - \Theta^*\|_F) + \mathcal{O}(\|S - S^*\|_F)$ where

$$Q^* \overset{\text{def.}}{=} (\Gamma_{\Theta^*}^\top)^\dagger \begin{pmatrix} \lambda_1 + \lambda_2 C_s^* / \|C_s^*\|_2 \\ \mathbf{0}_{kd} \end{pmatrix}.$$

Therefore,

$$\|Z\| = \mathcal{O}\left(\|\Theta - \Theta^*\|_F^2 c_{\max} + \|C^* - C\|_F \|\Theta - \Theta^*\|_F + \alpha + \|W\|\right).$$

So,

$$\|Y\| = \mathcal{O}\left(c_{\min}^{-1} \cdot \left(c_{\max}\|\Theta - \Theta^*\|_F^2 + \|C^* - C\|_F \|\Theta - \Theta^*\|_F + \alpha + \lambda_2\alpha + \|W\|\right)\right)$$

To show that $\partial_{(C,\Theta)} F$ is invertible, note that given square matrices A, E where A is invertible, $(A+E)$ is also invertible with $\|(A+E)^{-1}\| \leqslant 2\|A^{-1}\|$ provided that $\|E\| \leqslant \frac{1}{2\|A^{-1}\|}$. We therefore require that

$$\alpha + \|W\| + \|C - C^*\|_F = \mathcal{O}(c_{\min}) \quad \text{and} \quad \|\Theta - \Theta^*\|_F = \mathcal{O}(c_{\min}/c_{\max})$$

We can therefore apply Proposition 4 with $u_0 = (C^*, \Theta^*)$, $v_0 = (0, \mathbf{0}_{k\times d})$, $r_a = c_{\min}$, $r_\theta = c_{\min}/c_{\max}$, $R = \mathcal{O}({c_{\min}}^2/c_{\max})$.

$$U_0 = \mathcal{B}(C^*, r_a) \times \mathcal{B}(\Theta^*, r_\theta)$$

We can therefore define

$$G : \mathcal{B}(v_0, R_0) \to \mathbb{R}^{N+kd}, \quad \text{where} \quad R_0 = \mathcal{O}(c_{\min}^2/c_{\max} \cdot \|(\Gamma_{\Theta^*}^\top \Gamma_{\Theta^*})^{-1}\|^{-1})$$

so that $G(\alpha, W) = (C, \Theta)$ if and only if $F(C, \Theta, \alpha, W) = 0$, and

$$\|C - C^*\|_F = \mathcal{O}(\alpha) \quad \text{and} \quad \|\Theta - \Theta^*\|_F = \mathcal{O}(\alpha/c_{\min})$$

Verifying the candidate solution. Finally, it remains to check that $G(\alpha, W) = (C, \Theta)$ does indeed correspond to a solution: it suffices to check that

$$Q \stackrel{\text{def.}}{=} \frac{-1}{\alpha} Z = \frac{-1}{\alpha}(D_\Theta \bar{C} - D_{\Theta^*} C^* - W)$$

satisfies the primal dual relationships (see Proposition 1). In particular, we need to check that $\Phi^* Q$ satisfies

$$\sup_{\theta \in \mathcal{T}} \|\frac{1}{\lambda_2}(\Phi^* Q(\theta) - \lambda_1)_+\|_2 \leqslant 1.$$

Note that $F(C, \Theta, \alpha, W) = 0$ can be rewritten as

$$\Gamma_\Theta^\top Z = -\begin{pmatrix} \alpha(\lambda_1 + \lambda_2 \frac{C_s}{\|C_s\|})_{s\in[k]} \\ \mathbf{0}_{kd} \end{pmatrix}.$$

By applying $\Gamma_\Theta (\Gamma_\Theta^\top \Gamma_\Theta)^{-1}$ to this equation and recalling that

$$\mathcal{P}_{\mathcal{R}(\Gamma_\Theta)} \stackrel{\text{def.}}{=} \Gamma_\Theta (\Gamma_\Theta^\top \Gamma_\Theta)^{-1} \Gamma_\Theta^\top$$

is the orthogonal projection onto the range of Γ_Θ, we obtain

$$-\frac{1}{\alpha} Z = (\Gamma_\Theta^\top)^\dagger u_0 - \frac{1}{\alpha} \mathcal{P}_{\mathcal{R}(\Gamma_\Theta)}^\perp (D_{\Theta^*} C^* + W)$$

It therefore follows that $Q = Q_V - \frac{1}{\alpha}\mathcal{P}_{\mathcal{R}(\Gamma_\Theta)}^\perp (D_{\Theta^*} C^* + W)$. We need to show that $g(\theta) \stackrel{\text{def.}}{=} \|\frac{1}{\lambda_2}[(\Phi^* Q)(\theta) - \lambda_1]_+\|_2^2$ satisfies

i) $g(\theta) < 1$ for all $\theta \notin \Theta$

ii) $\nabla^2 g(\theta_s) \prec 0$ for all $s \in [k]$.

iii) $g(\theta_s) = 1$ for all $s \in [k]$.

Note that since $|(\Phi^*Q)(\theta) - (\Phi^*Q_V)(\theta)| \leqslant \max_\theta \|\varphi(\theta)\|_2 \|Q - Q_V\|_F = \|Q - Q_V\|_F$ and

$$|\nabla^2(\Phi^*Q)(\theta) - \nabla^2(\Phi^*Q_V)(\theta)| \leqslant \max_\theta \|\mathbb{H}_\varphi(\theta)\| \|Q - Q_V\|_F,$$

provided that $\|Q - Q_V\|_F$ is sufficiently small, we have $\eta_i(\theta) > \lambda_1$ whenever $(\Phi^*Q_V)_i(\theta) > \lambda_1$ and g satisfies i) , ii), iii) since η_V is nondegenerate.

It is enough to show that $\|Q_V - Q\|_F \leqslant \rho$ for a sufficiently small constant ρ (which depends only on η_V, and in particular, $\min_{i \in I_s} [\Phi^* Q_V(\theta_s^*)]_i - \lambda_1$, $\|\nabla^2 \eta_V(\theta_s^*)\|$ and $1 - \eta_V(\theta)$ for $\theta \notin \cup_s \mathcal{B}(\theta_s^*, r)$ where r is such that $\min_{\theta \in \mathcal{B}(\theta_s, r)} \|\nabla^2 \eta_V(\theta)\| \geqslant \frac{1}{2} \|\nabla \eta_V(\theta_s)\|$). Note that

$$\begin{aligned}
\mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} D_{\Theta^*} C^* &= \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \varphi\left(\theta_s^*\right) \left(C_s^*\right)^\top \\
&= \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \left(\varphi\left(\theta_s\right) \left(C_s^*\right)^\top + (\theta_s - (\theta_0)_s) \mathbb{J}_\varphi\left(\theta_s\right) \left(C_s^*\right)^\top \right) \\
&\qquad + \mathcal{O}(c_{\max} \|\Theta - \Theta^*\|_F^2) \\
&= \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} \sum_s \left(\varphi\left(\theta_s\right) \left(C_s^*\right)^\top + (\theta_s - (\theta_0)_s) \mathbb{J}_\varphi\left(\theta_s\right) C_s^\top \right) \\
&\qquad + \mathcal{O}(\|C^* - C\|_F \|\Theta - \Theta^*\|_F) + \mathcal{O}(c_{\max} \|\Theta - \Theta^*\|_F^2) \\
&= \mathcal{O}(c_{\max} \|\Theta - \Theta^*\|_F^2 + \|C^* - C\|_F \|\Theta - \Theta^*\|_F).
\end{aligned}$$

So,

$$\begin{aligned}
\alpha^{-1} \mathcal{P}^\perp_{\mathcal{R}(\Gamma_\Theta)} D_{\Theta^*} C^* &= \mathcal{O}(\alpha^{-1} \|\Theta - \Theta^*\|_F^2 c_{\max} + \alpha^{-1} \|C^* - C\|_F \|\Theta - \Theta^*\|_F) \\
&= \mathcal{O}(\|\Theta - \Theta^*\|_F c_{\max}/c_{\min} + \|\Theta - \Theta^*\|_F) = \mathcal{O}(1).
\end{aligned}$$

So, $Q = Q_V + \mathcal{O}(\|W\|_F/\alpha) + \mathcal{O}(1)$, and so, C, Θ define a solution provided that $\|W\|_F/\alpha = \mathcal{O}(1)$.

□

4 Conclusion

In this work, we considered the nonlinear least squares problem where the mixture weights are vectors. We introduce the sparse-group Beurling-Lasso, which is an off-the-grid convex optimization problem, and our regularisation promotes the recovery of both sparse measures and sparsity within each mixture weight. Our theoretical analysis establish support stability under the existence of a nondegenerate precertificate.

References

[1] Leentje Vanhamme, Aad van den Boogaart, and Sabine Van Huffel. Improved method for accurate and efficient quantification of mrs data with use of prior knowledge. *Journal of magnetic resonance*, 129(1):35–43, 1997.

[2] Sean CL Deoni, Lucy Matthews, and Shannon H Kolind. One component? two components? three? the effect of including a nonexchanging "free" water component in multicomponent driven equilibrium single pulse observation of t1 and t2. *Magnetic resonance in medicine*, 70(1):147–154, 2013.

[3] Alexandre Gramfort, Daniel Strohmeier, Jens Haueisen, Matti S Hämäläinen, and Matthieu Kowalski. Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013.

[4] Travis Askham and J Nathan Kutz. Variable projection methods for an optimized dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 17(1):380–416, 2018.

[5] Linda Kaufman, Garrett S Sylvester, and Margaret H Wright. Structured linear least-squares problems in system identification and separable nonlinear data fitting. *SIAM Journal on Optimization*, 4(4):847–871, 1994.

[6] Alexander R Borden and Bernard C Lesieutre. Variable projection method for power system modal identification. *IEEE Transactions on Power Systems*, 29(6):2613–2620, 2014.

[7] Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1, 2003.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[9] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[10] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[11] Vincent Duval and Gabriel Peyré. Sparse regularization on thin grids i: the lasso. *Inverse Problems*, 33(5):055008, 2017.

[12] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[13] Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

[14] Arne Beurling. Sur les intégrales de fourier absolument convergentes et leur applicationa une transformation fonctionnelle. In *Ninth Scandinavian Mathematical Congress*, pages 345–366, 1938.

[15] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[16] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[17] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[18] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*, 2018.

[19] Paul Catala, Vincent Duval, and Gabriel Peyré. A low-rank approach to off-the-grid sparse deconvolution. *IEEE Trans. Information Theory*, 63(1):621–630, 2017.

[20] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

[21] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank–wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.

[22] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

[23] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

[24] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. Support localization and the fisher metric for off-the-grid sparse regularization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1341–1350, 2019.

[25] Mohammad Golbabaee and Clarice Poon. An off-the-grid approach to multi-compartment magnetic resonance fingerprinting. *arXiv preprint arXiv:2011.11193*, 2020.

[26] Kenneth P Whittall and Alexander L MacKay. Quantitative interpretation of nmr relaxation data. *Journal of Magnetic Resonance (1969)*, 84(1):134–152, 1989.

[27] D Ma, V Gulani, N Seiberlich, K Liu, J Sunshine, J Durek, and M Griswold. Magnetic resonance fingerprinting. *Nature*, 495(7440):187–192, 2013.

[28] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[29] Youye Xie, Michael B Wakin, and Gongguo Tang. Support recovery for sparse signals with non-stationary modulation. *arXiv preprint arXiv:1910.13104*, 2019.

[30] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. Support recovery for sparse super-resolution of positive measures. *Journal of Fourier Analysis and Applications*, 23(5):1153–1194, 2017.

[31] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. In *Advances in neural information processing systems*, pages 388–396, 2016.

[32] O Burdakov and Boris Merkulov. On a new norm for data fitting and optimization problems. *Linköping University, Linköping, Sweden, Tech. Rep. LiTH-MAT*, 2001.

[33] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.