An off-the-grid approach to multi-compartment magnetic resonance fingerprinting

Mohammad Golbabaee¹,* and Clarice Poon²

¹ Computer Science Department, University of Bath, United Kingdom
² Mathematics Department, University of Bath, United Kingdom

E-mail: m.golbabaee@bath.ac.uk

Received 27 September 2021, revised 28 February 2022
Accepted for publication 18 May 2022
Published 15 June 2022

Abstract
We propose a novel numerical approach to separate multiple tissue compartments in image voxels and to estimate quantitatively their nuclear magnetic resonance (NMR) properties and mixture fractions, given magnetic resonance fingerprinting (MRF) measurements. The number of tissues, their types or quantitative properties are not a-priori known, but the image is assumed to be composed of sparse compartments with linearly mixed Bloch magnetisation responses within voxels. Fine-grid discretisation of the multi-dimensional NMR properties creates large and highly coherent MRF dictionaries that can challenge scalability and precision of the numerical methods for (discrete) sparse approximation. To overcome these issues, we propose an off-the-grid approach equipped with an extended notion of the sparse group Lasso regularisation for sparse approximation using continuous (non-discretised) Bloch response models. Furthermore, the nonlinear and non-analytical Bloch responses are approximated by a neural network, enabling efficient back-propagation of the gradients through the proposed algorithm. Through numerical experiments on simulated and in vivo healthy brain MRF data, we demonstrate the effectiveness of the proposed scheme compared to baseline multi-compartment MRF methods.

*Author to whom any correspondence should be addressed.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Keywords: inverse problems, compressed sensing, quantitative MRI imaging, magnetic resonance fingerprinting, off-the-grid compressed sensing, sparse approximation

Supplementary material for this article is available online (Some figures may appear in colour only in the online journal)

1. Introduction

Multi-compartment (MC) effects, also known as partial volume effects, occur when more than one tissue type occupies a single image voxel. This effect is common in medical images e.g. in MRI [1–5] due to the images’ finite spatial resolution. Estimating MC effects is crucial for obtaining accurate segmentation and estimation of the tissue volumes and their contents e.g. for studies related to several brain disorders such as Alzheimer’s disease, multiple sclerosis, or Schizophrenia [6–8].

Magnetic resonance fingerprinting (MRF) [9, 10] is an emerging technology that enables quantitative mapping of several tissues’ physical properties in short and clinically feasible scan times. The MC effects also occur in MRF, and if left unmodelled, they can produce false and blurry mappings e.g. at the tissue boundaries [11, 12]. On the other hand, a multi-component MRF (MC-MRF) analysis can potentially help to distinguish more tissues more reliably than e.g. contrast-weighted MRI, because multiple quantitative tissue parameters are measured in (naturally) co-registered mapped images.

Current numerical methods for solving MC-MRF rely on sparse approximations of linear mixtures in a large discretised dictionary of simulated Bloch responses (fingerprints). This can lead to several numerical issues: the accuracy of the estimated compartments depends on a fine-grid discretisation of the tissue properties that amounts to exponentially-large dictionaries in multi-parametric MRF applications and creating storage bottlenecks. Fine-grid discretisation also increases the coherence of dictionary atoms (fingerprints) which fundamentally limits the precision of sparse approximation. Furthermore, the precision of fast (first-order) shrinkage solvers such as FISTA [13] was observed to be inadequate (despite long iterations) to tackle the over-redundancy of the MRF dictionary [11, 12], demanding instead higher precision and more computationally-involved sparse solvers.

1.1. Contributions

In this study, we propose the first numerical approach for off-the-grid MC-MRF estimation. Motivated by the promising use of continuous off-the-grid methods in inverse problems [14–16], our approach adopts continuous (non-discretised) models of Bloch responses for sparse approximation, in order to address the (storage) non-scalability issue of the state-of-the-art methods as well as offering more precise MC estimations with continuous mapping of tissues’ multi-dimensional quantitative properties. We do so by modelling the sought-after MC as a sparse measure made up of a sum of Dirac masses, and unlike previous approaches where the sparse measure has scalar amplitudes, our model is a vector-valued measure (since each component is an image, represented as a vector). We present a regularisation for promoting a hybrid notion of group-and-pixel sparsity in the MC-MRF solutions. This regularisation extends the sparse group Lasso model of [17] to non-discretised dictionaries. We also extend previous theoretical analysis on identifiability and stability to this sparse-group setting.
a new shrinkage algorithm based on Frank–Wolfe iterations is proposed to solve this optimisation problem, and the nonlinear and non-analytical Bloch magnetic responses are approximated by a neural network to enable efficient back-propagation of the gradients through this algorithm.

2. Related works on MC analysis

MC analysis based on regularised statistical models have been applied in various settings, including: conventional contrast-weighted MRI [2–5]; quantitative MRI relaxometry for splitting multi-exponential Bloch responses e.g. for an individual [18, 19] or joint mapping [20–22] of the T1 and T2 relaxations times; and for measuring the white/grey matters volumes, the myelin content in the brain or the cerebral blood flow [23–26]. MC-MRF analysis has also been performed in the context of multi-parametric tissue mappings e.g. T1/T2 relaxometry [11, 12, 27, 28]. MC-MRF baselines typically use large dictionaries of simulated Bloch responses over fine-grid discretisation of the multi-dimensional parameter space. For example, methods [11, 12, 27] iteratively solve (nonconvex) reweighted $\ell_2$ and reweighted nonnegative (squared) $\ell_1$ regularisations, respectively, for promoting sparsity in the mixture weights. High-precision (e.g. second-order [12]) optimisation methods were employed to solve each reweighted iteration, this was done in conjunction with empirical dictionary pruning heuristics [11, 27] to reduce long runtimes. Method [27] promotes group-sparsity to cluster the whole image into sparse compartments which is shown (in vitro and in vivo) to be more accurate and easy-to-visualise than the pixel-wise sparsity [11, 12]. Other group-sparse models [28, 29] based on $k$-means clustering assume an a priori known number of compartments and that most image pixels are 100% pure (single-compartment). For instance [28], applies $k$-means on single-compartment mappings obtained via MRF dictionary matching [30] to cluster tissues’ T1/T2s, followed by an additional (combinatorially large) dictionary matching step for estimating mixture fractions. Our approach differs from these MC-MRF baselines: for improved numerical precision and scalability, we use continuous Bloch response models for sparse approximation rather than discretised (gridded) dictionaries. Although single-compartment continuous MRF mapping were proposed in [31, 32] without sparsity penalties, this work proposes a different numerical approach rooted in the off-the-grid sparse approximation literature [14, 15] for promoting sparsity in the MC-MRF solutions.

3. Multi-compartment quantitative MRI model

We are interested in quantifying tissues nuclear magnetic resonance (NMR) properties given a time-series of magnetisation images (TSMI) $X = [x_1, x_2, \ldots, x_v] \in \mathbb{R}^{T \times v}$ with $v$ voxels and $T$ timeframes\(^1\). We assume that the per-voxel magnetisation signal $x \in \mathbb{R}^T$ resulted by MCs follows a linear mixture model, with $m \in \mathbb{N}$ components [11]:

$$x = \sum_{s=1}^{m} c_s \phi(\theta_s),$$

where $c_s \geq 0$ are the mixture weights and $\phi : \mathcal{T} \to \mathbb{R}^T$ is the Bloch magnetisation response model that maps $d$-dimensional NMR properties $\theta_s \in \mathcal{T} \subseteq \mathbb{R}^d$ to time signals. We focus on

\(^1\)The TSMI is obtained from possibly undersampled multi-coil $k$-space data using a reconstruction algorithm and a phase-correction step for mapping into real-valued images (see section 6.1) before multi-compartment analysis.
the case of $d = 2$, where $\theta = (T_1, T_2)$ represents the T1 and T2 relaxometry with the constraint $\mathcal{T} = \{(T_1, T_2) | T_1 > T_2\}$. We can extend this formulation to all TSMI voxels by writing:

$$X = \sum_{s=1}^{m} \phi(\theta_s) c_s^\top$$

where the vector $c_s \in \mathbb{R}_+^{v}$ correspond to the $\theta_s$-dependent mixture weights for all voxels i.e. the mixture map images.

For factorising the TSMI and estimating tissue properties and mixture maps $\{\theta_s, c_s\}$, the MC-MRF baselines proposed to quantise the space of NMR properties by a dense grid $\Theta \subset \mathcal{T}$, and form an exponentially-large MRF dictionary

$$D_\Theta \triangleq (\phi(\theta))_{\theta \in \Theta}$$

of $n = \text{Card}(\Theta) = \mathcal{O}(e^d)$ fingerprints ($e > 1$). This has lead to the following discretised or gridded formulation:

$$X = D_\Theta C^\top,$$

where one aims to recover the very large-sized matrix $C \in \mathbb{R}_+^{v \times n}$ containing the mixture maps for every possible fingerprint. To handle this ill-posed inverse problem, one typically imposes sparsity (or a column-sparse) constraints on the mixture map $C$, as well as non-negativity constraints $[27, 28]$.

4. Sparse group Beurling Lasso (SGB-Lasso)

We introduce our off-the-grid approach by rewriting (2) as a continuous linear model over the space of vector-valued measures on $\mathcal{T}$, which we denote by $\mathcal{M}(\mathcal{T}; \mathbb{R}^v)$. Elements of $\mathcal{M}(\mathcal{T}; \mathbb{R}^v)$ are measures $m$ that map all measurable subsets of $\mathcal{T}$ to $\mathbb{R}^v$. The variation of a measure $m$ is defined relative to a norm $\| \cdot \|$ on $\mathbb{R}^v$ as follows:

$$|m|(\mathcal{T}) \triangleq \sup_{A_i \in \mathcal{P}} \sum_{i=1}^{n} \|m(A_i)\|$$

where the supremum is over all partitions $\{A_i\}_{i=1}^n$ of $\mathcal{T}$. Our measure of interest $m \in \mathcal{M}(\mathcal{T}; \mathbb{R}^v)$

$$m = \sum_{s=1}^{m} c_s^\top \delta(\theta - \theta_s)$$

is characterised jointly by the mixture maps $c_s \in \mathbb{R}_+^{v}$ and the tissue properties through the weighted sum of Dirac’s mass function $\delta(\theta - \theta_s)$ at positions $\theta_s$. We now define a linear operator $\Phi : \mathcal{M}(\mathcal{T}; \mathbb{R}^v) \rightarrow \mathbb{R}^{T \times v}$ by

$$\Phi m \triangleq \int \phi(\theta) d m(\theta),$$

so that equation (2) becomes $X = \Phi m$ for the $m$ given in (6). We therefore consider the infinite dimensional linear inverse problem of estimating the measure $m$ from the TSMI $X$. 

We regularise the underlying measure \((6)\) to be sparse i.e. composed of few compartments in the sum and additionally, each mixture map \(c_i\) itself to be a sparse image. For this we propose to solve \((7)\) by the following variational formulation, coined as the SGB-Lasso:

\[
\arg\min_{m \in \mathbb{M}(\mathbb{R}^{T \times v})} \frac{1}{2} \|X - \Phi m\|_F^2 + \alpha \|m\|_\beta \quad (\text{SGB - Lasso})
\]

Where \(\|\cdot\|_F\) denotes the matrix Frobenius norm. The regularisation parameter \(\alpha > 0\) balances between a TSM fidelity term and a sparsity-promoting norm. We define the sparse group total variation (SGTV) norm \(\|m\|_\beta\) as follows: let the variation norm in \((5)\) relative to the vector one-norm \(\|y\|_1 = \sum_{i=1}^v |y_i|\) be denoted by \(|m|_1(T)\), and let the variation norm in \((5)\) defined via the vector two-norm \(\|y\|_2 = \sqrt{\sum_{i=1}^v |y_i|^2}\) be denoted by \(|m|_2(T)\). We can then define the SGTV norm as:

\[
\|m\|_\beta \overset{\text{def}}{=} (1 - \beta)|m|_1(T) + \beta \sqrt{T}|m|_2(T).
\]

With a slight abuse of notation, we also define a matrix norm:

\[
\|C\|_\beta \overset{\text{def}}{=} \sum_s (1 - \beta)\|c_s\|_1 + \beta \sqrt{T}\|c_s\|_2,
\]

where \(c_s\) denotes the \(s\)th column of a matrix \(C\) containing \(MC\) mixture maps. With respect to \((6)\), we have the identities \(|m|_1 = \sum_{s}\|c_s\|_1\), \(|m|_2 = \sum_{s}\|c_s\|_2\), and therefore \(\|m\|_\beta = \|C\|_\beta\).

Note that \(\|m\|_\beta\) is a continuous extension of the sparse group regularisation introduced in [17] for discretised dictionaries. The term \(|m|_1(T)\) promotes group sparsity in a sense that few compartments \((\theta_s)\) should contribute to approximate the entire TSM across all voxels \((2)\). The term \(|m|_2(T)\) in addition promotes spatial sparsity within the mixture map/image \((C_s)\) of each contributing compartment i.e., a voxel should usually receive contributions from fewer compartments than those composing the entire TSM, which is related to having some level (not necessarily 100\%) of voxel purity that helps cross-compartment spatial separability of the mixture maps. The parameter \(0 \leq \beta \leq 1\) provides a degree of freedom to balance between these two terms and to promote the corresponding sparsity forms jointly and at appropriate levels in the solutions.

5. Identifiability of the mixtures via SGB-Lasso

The convexity of the SGB-Lasso regularisation problem allows for obtaining guarantees on the identifiability of mixtures based on convex analysis tools. While identifiability of sparse models have been theoretically analysed in previous works (e.g. [14, 15, 33]), these results are for the case of scalar-valued measures. In this section, we present the analogous theoretical results in the vector-valued setting, where our vector norm allows for sparsity within each component.

Suppose that \(X = \Phi m + W\) where \(m\) is a \(m\)-sparse \((m\)-compartment\) measure from \((6)\) and \(W \in \mathbb{R}^{T \times v}\) is some bounded additive noise \(\|W\|_F \leq \epsilon\). Below we describe a condition (certificate) that guarantees stable estimation of \(m\) i.e., stable demixing of \(\{\theta_s, c_s\}_{s=1}^v\) given a
noisy TSMI $X$. In the following, we denote the adjoint of $\Phi$ by $\Phi^*$. Note that $\Phi^*$ maps matrices in $\mathbb{R}^{2n \times d}$ to the space of vector-valued continuous functions, denoted by $\mathcal{C}(T, \mathbb{R}^n)$. It has the explicit form $(\Phi^* Q)(\theta) = Q^\top \phi(\theta) = (q_i(\theta))^\top_{i=1}$ where $q_i$ is the $i$th column of a matrix $Q \in \mathbb{R}^{2n \times d}$.

**Definition 1.** Let $f[Q_*] \equiv \frac{(\Phi^* Q_* + \beta - 1)}{\sqrt{\gamma^2 + \epsilon}} \in \mathcal{C}(T; \mathbb{R})$ where

$$Q_* \equiv \arg \min_{Q \in \mathbb{R}^{2n \times d}} \|Q\|_2 \quad \text{s.t.} \quad f = \frac{(\Phi^* Q + \beta - 1)}{\sqrt{\gamma^2 + \epsilon}} \in \mathcal{K}. \quad (10)$$

The constraints of $\mathcal{K}$ are $\forall s \in \{1, \ldots, k\}$: $[f(\theta)]_{I_s} = [c_s]_{I_s}/\|c_s\|_2$ and $\nabla \|f(\theta)\|_2^2 = 0$. Here, $I_s$ is the support (indices) of the non-zero elements of $c_s$, and the subscript of a vector with $I_s$ denote the restriction to $I_s$, and subscript with $+$ denotes restriction to positive part. We then define the stability certificate:

$$g_s(\theta) \equiv \|f[Q_s](\theta)\|_2^2. \quad (11)$$

By definition $g_s(\theta) = 1$, $\forall s = 1, \ldots, m$. We call $g_s$ a non-degenerate certificate if it satisfies:

(a) (non-saturation) $g_s(\theta) < 1$, $\forall \theta \notin \{\theta_s\}_{s=1}^m$.

(b) (Curvature) the Hessian $\nabla^2 g_s(\theta)$ is negative definite, $\forall s = 1, \ldots, k$.

**Remark 1.** The non-degeneracy condition is a typical condition used to analyse stability and identifiability properties of the Lasso. In particular, the vector $Q_s$ in (10) is the analogous definition of the minimal norm pre-certificate presented in [15], extended to the sparse group setting. Under the non-degeneracy condition, this object is in fact the minimal norm solution to the dual problem of (SGB-Lasso) as $\alpha \to 0$ and from which, one can deduce properties on solution to (SGB-Lasso) when $\alpha$ is sufficiently small. We finally remark that the constraints of (10) can be written as a linear system and the condition can therefore be checked numerically. We numerically illustrate this condition in section 5.1 for our neural network embedded Bloch response model $\phi(\cdot)$.

The result below shows that these conditions are sufficient for stable demixing. Following similar arguments in [15, proposition 8] one could also expect these conditions are tight, so that if $g_s(\theta) > 1$, then demixing is necessarily unstable.

**Theorem 1.** Suppose that the matrix with columns made up of $\phi(\theta_s)$ and $\mathbb{J}_0(\theta_s)$, the Jacobian of $\phi$ at $\theta_s$, is full rank. If the certificate $g_s$, associated to $m$, is nondegenerate with respect to $\{c_i\}$ and $\theta_s$, then there exists a constant $\gamma > 0$ such that by setting $\alpha \leq \gamma$ and $\epsilon/\alpha \leq \gamma$, the (SGB-Lasso) has a unique solution of the form $\sum_{i=1}^m c_i \delta(\theta - \theta_s)$ with bounded errors $\sum_{i=1}^m \|\hat{e}_i - c_i\|_2 = O(\epsilon)$ and $\sum_{i=1}^m \|\hat{\theta}_i - \theta_s\|_2 = O(\epsilon^2)$.

**Remark 2. (Idea of proof).** The proof is in appendix B and it follows the proof techniques introduced in [15] for the case of $\beta = 0$. The main difference to previous works is the $f[Q]$ which comes from characterisations of $\|\cdot\|_2$ for $\beta > 0$. By convex duality of the problem (SGB-Lasso), one can show that the solution to (SGB-Lasso) is related to the dual solution $Q_\alpha \in \mathbb{R}^{2n \times d}$ by $Q_\alpha = (X - \Phi m_\alpha)/\alpha$ and $\text{Supp}(m_\alpha) \subset \{\theta \in T \|f[Q_\alpha](\theta)\|_2 = 1\}$. In particular, $Q_\alpha$ describes the support of $m_\alpha$: provided that there are only $m$ points $\theta \in T$ for $\|f[Q_\alpha](\theta)\|_2 = 1$, the measure $m_\alpha$ will be supported on at most $m$ components. The full rank
condition ensures uniqueness of solutions. For establishing support stability for small $\alpha$, the typical approach is to study a surrogate to $Q$, which can be easily computed. This surrogate is precisely the object in the nondegeneracy condition.

**Remark 3. (Insights from the certificate).** The certificate can be computed by solving a linear system. Although this analysis requires knowing the ground-truth demixed targets which are usually not available for *in vivo* data, one can use it for understanding the identifiability conditions of the operator $\Phi$ and by using (simulated) models for the tissues e.g. the brain phantom. This is further illustrated in the next section.

5.1. **Numerical illustration of the certificate**

Illustrations below adopt a neural network embedded Bloch response model $\phi(\cdot)$ used in our experiments for encoding $\theta = (T_1, T_2)$ NMR relaxation properties (see sections 6.2.3 and 7.2 for model details). We numerically examine the certificate nondegeneracy in theorem 1, particularly the non-saturation condition, on several two-compartment examples $m = c_1^\top \delta(\theta - \theta_1) + c_2^\top \delta(\theta - \theta_2)$ in order to highlight the following points:

1. The certificate is defined with linear constraints and can be numerically computed by solving a linear system (see section appendix B.4.1 for details), moreover, in the examples below, the corresponding linear system is indeed full rank.

2. Stable demixing requires a minimum separation between the compartments’ $T_1/T_2$ values. Consider simulating mixtures with $\theta_1 = (784, 77)$ ms and $\theta_2 = \theta_1 + \Delta \theta$ where $\Delta \theta = (1216, 96) - \theta_1$ and the separation is denoted by $\Delta \in \mathbb{R}$. The $T_1/T_2$ values considered here are associated to the two left-most mixture maps of the brain phantom, bottom row figure 6. For $\beta = 10^{-3}$, we plot the certificate $g_*$ for different values of $\Delta$ in figure 1. We can observe that $g_*$ becomes degenerate when $\Delta$ is too small (for the plots shown, we have nondegeneracy when $\Delta \geq 0.4$). This means, there is a minimum separation distance below which demixing becomes unstable.

3. **Choice of $\beta$ depends on the sparsity of the mixture maps.** We simulate several mixtures where $\theta_1 = (719, 80)$, $\theta_2 = (1190, 98)$ ms, and the mixture weights $c_1, c_2 \in \mathbb{R}^v = 1024$ are randomly generated from i.i.d. normal distribution (followed by taking the absolute value to ensure positivity) with at most $\lceil \rho v \rceil$ nonzero entries. In figure 2, we display the values of parameter $\beta$ and sparsity level ratios $\rho \in [0,1]$ for which the nondegeneracy condition is satisfied. As expected, $\beta$ should be taken smaller for sparser mixture maps/weights.

4. **Group sparsity is essential.** We finally highlight the issue that, the case of $\beta = 0$ (which relates to pure pixel sparsity framework e.g. [12]) is numerically unstable, and taking $\beta > 0$ is necessary to ensure stable mixture separation. Consider a mixture example with $\theta_1 = (784, 77)$, $\theta_2 = (1216, 96)$ ms, and almost pure (except one pixel) mixtures maps $c_1, c_2$ in figure 3(a). We compute and plot in figure 3(b) two certificates $g_*$ for when SGB-Lasso would use $\beta = 0$ and $\beta = 10^{-3}$. We observe a discontinuity in the SGB-Lasso’s behaviour as per removing or adding the non-smooth group-sparsity penalty through changing $\beta = 0$ to $\beta > 0$ (see further details in supplementary section SIV (https://stacks.iop.org/IP/38/085002/mmedia)). As expected $g_*(\theta_1) = g_*(\theta_2) = 1$ in both cases (red points), but for certificate $g_*(\theta)$ to be nondegenerate its value should not exceed 1 at $\theta \notin \{\theta_1, \theta_2\}$. In this case, switching from $\beta = 0$ to $\beta > 0$ results in a drastic difference: the certificate is nondegenerate only when $\beta > 0$.

---

2 Note that this is not an exhaustive list and other factors including the Bloch response model, the number of compartments, and the mixture weights values also play roles in determining the certificate behaviour in definition 1.
Figure 1. The certificate $g_\star(\theta)$ values across the $\theta = (T_1, T_2)$ plane, shown for six two-compartment mixture examples where the compartments’ $T_1/T_2$ values (red points) are separated by distance $\Delta$. A certificate is nondegenerate if $g_\star < 1$ elsewhere than red points, which holds for cases $\Delta \geq 0.4$.

Figure 2. The white region satisfies nondegeneracy $\max_{\theta} g_\star(\theta) \leq 1$, and guarantees stable demixing by SGB-Lasso. We see that for small levels of sparsity, $\beta$ should be chosen small.
6. Algorithms

6.1. TSMI reconstruction

Prior to mixture separation, the (coil-combined) TSMI is computed from MRF’s undersampled multi-coil k-space measurements using the LRTV algorithm [34, equation (10)]. This method is dictionary-matching-free and does not limit the reconstruction accuracy to the finite resolution of a discretised MRF dictionary. By exploiting the TSMI’s spatiotemporal structures, LRTV can efficiently remove aliasing artefacts and is shown more accurate than the Fourier backprojection scheme SVD-MRF [30] used by most MC-MRF baselines. LRTV exploits the low-rank structure of TSMI (see also [30, 35–38]) for subspace dimensionality reduction and accelerated reconstructions. For many MRF sequences, including fast imaging steady state precession (FISP) [10] in our experiments, the Bloch responses and TSMIs lie approximately on a low-dimensional subspace:

\[ \phi(.) \approx \mathbf{V} \mathbf{V}^T \phi(.) \quad \text{and} \quad \mathbf{X} \approx \mathbf{V} \mathbf{V}^T \mathbf{X}, \]  

(12)

where \( \mathbf{V} \) is a \( T \times \tau \) tall matrix representing the subspace of dimension \( \tau \ll T \). The raw TSMIs (complex-valued) are then phase-corrected and mapped to real-valued images before being fed to the mixture separation step. This is particularly important for imposing the non-negativity constraint in (2) and SGB-Lasso. For FISP sequence with constant TE per-voxel signal evolution has a constant complex-valued phase [10, 27]. This phase can be estimated from the first (principal) image component of the dimension-reduced TSMI and removed from the image [34, 39, 40].

6.2. Solving SGB-Lasso by Frank–Wolfe iterations

One of the typical algorithms used for solving the Lasso is the Frank–Wolfe algorithms: as explained in [16, 33, 41], (SGB-Lasso) can be viewed as the minimisation of a differentiable
Algorithm 1. Frank–Wolfe iterations to solve (SGB-Lasso).

1: Inputs: TSMI X, Bloch model ϕ(), params α, β
2: Outputs: NMR parameters θ, mixture weights C
3: Initialise: i = 0, (θ₀, C₀) = {} \{η(θ) \equiv \frac{1}{α} Φ^∗ X
4: Repeat
5: \(\theta^t = \arg\max_{\theta \in \mathcal{T}} \|η(\theta) + β - 1\|_2^2\)
6: \(\theta^{t+1} = \theta^t \cup \{\theta^t\}\)
7: \(C^{t+1}_i = \arg\min_{C \geq 0} \frac{1}{2} \|X - D_{\theta^t} \theta^t C\|_F^2 + \alpha \|C\|_β\)
8: Initialised by \(C^{t+1/2}\) and \(\theta^{t+1/2}\), solve \((C^{t+1}, \theta^{t+1}) = \arg\min_{θ \in \mathcal{T}}, C \geq 0} \frac{1}{2} \|X - \sum_{i=1}^{t+1} \phi(\theta_i) c_i\|_F^2 + \alpha \|C\|_β\)
9: \(\eta(\theta) \equiv \frac{1}{α} Φ^∗ (X - D_{\theta^t+1}(C^{t+1})^\top)\)
10: \(t = t + 1\)
11: Until \(\max_{\theta \in \mathcal{T}} \|η(\theta) + β - 1\|_2^2 < αβ^2\)

function on a bounded convex set in the space of measures, for which the Frank–Wolfe algorithms can be directly applied. Following this line of work, we propose to solve (SGB-Lasso) using algorithm 1 which generalises the sliding Frank–Wolfe algorithm of [41] beyond \(β = 0\). Algorithm 1 recovers one compartment per iteration (steps 5 and 6) and undergoes refinement steps 7 and 8 to update the current solution \(\{\theta^t, c^t\}\). Below we highlight the key algorithm steps, more details about the algorithm derivation from the standard Frank–Wolfe algorithm is in appendix A.

6.2.1. Nonconvex steps. Algorithm 1 (lines 5 and 8) optimises nonconvex objectives (due to the nonlinearity of the Bloch responses) with respect to the NMR parameters \(θ\) or a set of them \(θ = (θ_i)\). We use the L-BFGS algorithm\(^3\) to locally solve these steps, constrained that the mixture maps \(C\) are non-negative and the parameters e.g. \(θ = (T_1, T_2)\) satisfy \(T = \{T_1 > T_2\}\). Line 5 is initialised by a coarse grid-search over a small subset of discretised NMR parameters in \(T\) i.e. a coarsely gridded dictionary used for enumerating \(η(θ)\) on this subset. Step 8 is initialised by the outcomes of lines 6 and 7 for the NMR parameters and mixture weights, correspondingly.

6.2.2. Convex step. Algorithm 1 recovers one compartment \(θ^t\) per iteration \(t\), simulates its Bloch response and adds it to a dictionary \(D_{θ^t+1/2} = D_{θ^t+1/2} \cup \phi(θ^t)\). In line 7, this \(D\) is used for solving discrete sparse group Lasso [17] by the (restarted) fast shrinkage algorithm FISTA [42] and the shrinkage operator:

\[
\text{Prox}_∥∥_β(C) = \arg\min_{W \geq 0} \frac{1}{2} \|W - C\|_F^2 + \|W\|_β = G_{\lambda, 2} \circ S_{1 - \beta}(C).
\]

where for a \(λ > 0\) and matrix \(C\), operator \(S_{1} (C) = (C - λ)\) is the element-wise positive soft thresholding, and \(G_{\lambda, 2} (c_i) = \frac{c_i}{∥c_i∥_2} \left(∥c_i∥_2 - λ\right)\) is the group soft thresholding for every column \(s\) of \(C\). Note that \(D\) has \(t+1\) columns at each Frank–Wolfe iteration. In practice/our numerical results, algorithm 1 is convergent before a maximum 30 iterations. Hence \(D\) is

\(^3\) https://github.com/stephenbeckr/L-BFGS-B-C.
small, leading to fast and accurate FISTA updates without facing numerical issues of dictionary over-redundancy.

6.2.3. Neural network approximation. Solving lines 5 and 8 requires computing derivatives of the Bloch responses with respect to the NMR parameters i.e. the Jacobian matrix $\partial \phi(\theta)/\partial \theta$. While this can be analytically computed for simple forms of Bloch responses (e.g. MRF sequences in [31, 32], or the exponential models in classical quantitative MRI), such approach would not extent to more general non-analytical response models e.g. those like FISP simulated by the extended phase graph (EPG) formalism [43]. To circumvent this issue, we leverage on universal approximation property of neural networks [44, 45] that enables embedding complicated functions in conveniently differentiable surrogates via back-propagation mechanism. We train a neural network $\tilde{\phi} : T \rightarrow \mathbb{R}^\tau$ to approximate the dimension-reduced Bloch responses (see section 7.2 for details, also [46] which applied this idea for single-compartment MRF reconstruction). This idea greatly accelerates algorithm 1 in steps requiring function evaluation and differentiating Bloch responses: for solving (SGB-Lasso), $X$ and $\phi(.)$ are replaced by

$$\tilde{X} \approx V^T X \quad \text{and} \quad \tilde{\phi}(\theta) \approx V^T \phi(\theta)$$

i.e. the dimension-reduced TSMI and the neural approximation of the (compressed) Bloch responses, respectively.

7. Experiments

Computations were conducted using MATLAB on an Intel Xeon gold CPU core and 32 GB RAM\(^5\). All experiments (simulated and in vivo) adopted a joint T1/T2-encoding MRF excitation sequence similar to the FISP protocol [10] with the same flip angle schedule, fixed repetition/echo times $TR/TE = 10/1.9$ ms, and the inversion time $18$ ms. This sequence had the length of $T = 1000$ timepoints (repetitions).

7.1. Tested algorithms

We compared SGB-Lasso (algorithm 1) to the MC-MRF baselines SPIJN [27], Bayesian-MRF [11] and PVMRF [28]. All algorithms used subspace dimensionality reduction $\tau = 10$ [30, 34]. Reconstructed TSMIs were phase-corrected and mapped to real-valued images before applying mixture separation. Baselines work with fine-gridded MRF dictionary. The regularisation parameters of BayesianMRF and SPIJN were $\mu_B = \{0.001, 1\}$ and $\lambda_{SPIJN} = \{0.03, 0.5\}$ for Dirichlet phantoms and the simulated/in vivo brain experiments, respectively. BayesianMRF used shape parameters $\alpha_B = 1.75, \beta_B = 0.1$. PVMRF used parameter $k_{PVMRF} = 11$ for the $k$-means. To stabilise PVMRF, $k$-means steps were repeated 10 times (randomly initialised) and result scoring lowest model error (4) was selected. SGB-Lasso used

\(^4\)To be specific, after replacements we solve $\arg\min_{m \in \mathcal{M}} \|X - \tilde{\Phi} m\|_2^2 + \alpha \|m\|_1$ by algorithm 1.

\(^5\)Source codes are available at https://github.com/mgolbabaee/SGB-Lasso-for-partial-volume-quantitative-MRI.
parameters \( \alpha = 10^{-6}, \beta = 0.9 \) for the Dirichlet phantoms experiment and \( \alpha = 0.8, \beta = 10^{-3} \) for the simulated and in vivo brain experiments.

7.2. Embedding Bloch responses by a neural network

A neural network was trained to approximate Bloch responses. For training and evaluation 95,143 Bloch responses were simulated using the EPG formalism [43] over a \((T1, T2) \in [10,6000] \times [4,4000] \text{ ms} \) grid discretised by logarithmically spaced values of \( T1 \) and \( T2 \) (400 points each) with \( T1 > T2 \) constraint. Following [30], we adopted PCA to compress Bloch responses’ (also TMSI’s) temporal dimensions to a \( \tau = 10 \) dimensional subspace i.e. the \( V \) matrix in (14). Data was randomly split in 80–10–10 percent ratios for training, validation and testing sets, correspondingly. For our application we followed [34] and used a (convolutional) network with \( 1 \times 1 \) filters for pixel-wise processing: two-channel inputs for \( T1 \) and \( T2 \) values, ten-channel (linear) outputs for the compressed Bloch responses, and one hidden layer of 500 channels with nonlinear ReLU activations. For training we minimised the MSE loss between the EPG-generated and network-predicted (dimension-reduced) Bloch responses using ADAM optimiser ran for 100 epochs with the initial learning rate 0.005, learning rate/gradient decay factors 0.95/0.95, and minibatch size 100. Training, validation and testing normalised RMSEs for approximating Bloch responses were \( \{8.9397, 8.9742, 8.9561\} \times 10^{-3} \), correspondingly. We also examined deeper networks with more hidden layers (and various numbers of filter channels) but did not observe a notable improvement in the overall MC-MRF performance (i.e. accuracy or runtime) compared to the chosen architecture. Also note that our approach is distinguished from neural models proposed for the MRF (e.g. [47–50]) based on mapping time-responses to the (single compartment) NMR properties. Our model instead generates the Bloch time-responses given the NMR properties as inputs.

SGB-Lasso used only neural network approximations. For the initialising grid-search in step 5 of algorithm 1, a fixed 64-atom dictionary was simulated over 10 logarithmically-spaced points per \( T1 \) and \( T2 \), and selecting \( T1 > T2 \), whereas the tested baselines used the actual EPG dictionary comprising 8540 fingerprints: 120 values per \( T1 \) and \( T2 \), respecting \( T1 > T2 \).

7.3. Numerical Dirichlet phantoms experiment

Several phantoms were simulated according to the mixture model (2) for creating TSMIs of \( 20 \times 20 \) pixels spatial resolution. In these phantoms three compartments were used with relaxation properties \( T1/T2 = \{784/77, 1216/96, 4083/1394\} \) ms. Per-pixel mixture weights were drawn randomly from i.i.d. Dirichlet distributions parametrised by various values of \( a > 0 \). This parameter governs the mixture levels such that large \( a \) values result in highly mixed pixels that receive similar contributions from all compartments and hence are more difficult for demixing [51]. Instead, small \( a \) creates sparse (pixel-pure) mixture maps i.e. pixels receive contributions from fewer compartments, making demixing task easier.

7.3.1. Different mixture levels.

We created a dataset of such phantoms from different mixture distributions with values \( a \in \{0.1, 0.5, 1, 2, 4\} \) and for each distribution we simulated 10 phantoms independently at random (see exemplar mixtures in supplementary figure C1). We used this dataset to compare the MC-MRF baselines to the SGB-Lasso algorithm. The mean absolute percentage errors (MAPE) of the estimated compartments’ \( T1 \) and \( T2 \) values were measured...
and reported in figure 4 using MATLAB’s boxchart tool (the inside box line, box edges, whiskers and circles represent the mean, quartiles, extreme values and outliers, correspondingly). The T1 MAPE (similarly for T2) is defined as $\sum_s |\hat{T}_1^s - T_1^s| / T_1^s$ where $\hat{T}_1^s$ and $T_1^s$ denote the estimated and true T1 values for the compartment $s \in \{1, 2, 3\}$. The SGB-Lasso outperforms tested baselines with accurate T1/T2 predictions i.e. less than 5% average errors for all tested mixture distributions. For $a = 0.1$ where the mixtures have the highest pixel-purity the PVMRF and SPIJN methods perform comparably well. However SGB-Lasso shows robustness for separating less pixel-pure mixtures i.e. larger $a$ values, where the gap between baselines and SGB-Lasso in terms of both the mean and variation of the MAPE errors increases. The supplementary figure C2 similarly reports the PSNRs of the estimated mixture maps. The SGB-Lasso’s mixture map PSNRs also on average outperform the baselines. It can be observed that more accurate T1/T2 predictions generally help the estimated mixture maps to also become more precise.

### 7.3.2. Noise stability and grid-search size.

In another experiment we created a dataset of phantoms by setting the Dirichlet parameter $a = 0.5$, and corrupt the TSMLs by additive white Gaussian noises of various SNRs $\in \{10, 20, 30, 40, 60\}$ dB. We simulated 10 noisy phantoms per SNR value. We used this dataset to measure the sensitivity of SGB-Lasso to noise and also the size of the grid-search used in algorithm 1 (step 5). For this we increased the storage requirement of SGB-Lasso using grid-sizes $\{64, 991, 2237\}$ corresponding to sampling $\{10, 40, 60\}$ logarithmically-spaced points per T1 and T2, $T_1 > T_2$. The estimated T1 and T2 MAPEs are reported in figure 5. For SNRs $\geq 20$ dB the average errors are less than 5%. These then increase to 10%–15% in the lowest tested SNR regime 10 dB. Further, it can be observed that SGB-Lasso’s overall performance per SNR regime has small variations with respect to the tested grid-sizes (less than 3% difference on average MAPEs) which could motivate adoption of small grid-sizes to overcome the memory inefficiency (challenge of dimensionality) of the dictionary-based MC-MRF approaches. Throughout other experiments, SGB-Lasso used the smallest grid-size 64 that is 2 orders of magnitude smaller and more memory-efficient than the size (8540 fingerprints) of the MRF dictionary used by the baselines.
7.4. Numerical brain phantom experiment

This experiment used the Brainweb’s anatomical model of healthy brain [1, 52]. This data includes fuzzy segmentations of the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) and provides precise control on the ground truth mixture maps (figure 6). The assigned relaxations for these tissues were $T_1/T_2 = \{784/77, 1216/96, 4083/1394\}$ ms, and TSMI was constructed according to the model (2).
Table 1. The SGB-Lasso’s (joint) T1/T2 MAPE errors and mixture maps PSNRs for estimating the simulated brain’s compartments in figure 6.

|                | WM | GM | CSF | WM | GM | CSF |
|----------------|----|----|-----|----|----|-----|
| Full           | 3.09 | 1.72 | 1.43 | 35.05 | 35.03 | 53.70 |
| LRTV           | 3.61 | 1.64 | 1.71 | 29.92 | 29.44 | 43.34 |
| SVD-MRF        | 3.39 | 1.62 | 2.33 | 24.69 | 23.69 | 34.08 |

We used this data to study SGB-Lasso for the case where TSMIs were un-compressed (fully sampled) and compare it to the case where acquisitions were accelerated using (compressed sensing) the same subsampled k-space readouts as for our in vivo data. Gaussian noise (50 dB SNR) were added to the measurements. Where compressed sensing applied, we adopted the reconstruction schemes LRTV and SVD-MRF to estimate TSMIs before the demixing step for comparison. Figure 6 illustrate the reconstructed mixture maps (weights) and their differences (errors) to the ground truth. The joint T1/T2 MAPEs and the mixture maps’ reconstruction PSNRs are reported in table 1. We observe accurate T1/T2 estimations with less than 4% MAPE for all compartments using the full and compressed-sampleddata. Estimated mixture maps have larger errors using subsampled data, but this error is smaller using LRTV than SVD-MRF that produces subsampling (aliasing) artefacts (figure 6).

7.5. In vivo brain experiment

To demonstrate feasibility of the proposed approach in vivo, compressed-sampled MRF data was acquired from a healthy volunteer’s brain with an informed consent obtained. Acquisition used a 1.5T GE HDxT scanner with eight-channel receive-only head RF coil, 230 × 230 mm² field-of-view, 230 × 230 image pixels, 5 mm slice thickness, with the above-mentioned FISP MRF protocol. A variable density spiral trajectory was used for k-space sampling. The total number of spiral interleaves were 377. One spiral arm sampled 920 k-space locations per TR/timeframe and this pattern was rotated for the next TR. The overall acquisition for \( T = 1000 \) timeframes was about 10 seconds.

7.5.1. The effects of parameter \( \beta \). We first used this data to examine the effects of parameter \( \beta \) in SGB-Lasso on mixture separation. LRTV algorithm (\( \lambda_{\text{LRTV}} = 4 \times 10^{-5} \)) was used to reconstruct coil-combined TSMI from undersampled k-space data (coil sensitivity maps were estimated using [53]). We then ran SGB-Lasso for different values of \( \beta = \{1, 10, 100, 500\} \times 10^{-4} \) (the algorithm has increased sensitivity towards small \( \beta \) values) and fixed regularisation weight \( \alpha = 0.8 \). The estimated compartments (T1/T2 values) are scatter plotted in figures 7(a)–(d). Mixture maps of these compartments are also shown in supplementary figure B1. As can be observed, interpolating between the two extreme cases of pure pixel vs. group sparsity (i.e. \( \beta \rightarrow 0 \) or 1) creates different demixing solutions across the T1/T2 values and the spatial mixture maps. For small \( \beta \) values tissues are decomposed into spatially finer/detailed compartments (figure B1), while larger \( \beta \) values hierarchically cluster together tissue compartments around fewer T1/T2 values. Decompositions at a moderate value \( \beta = 10^{-3} \) indicates a single component for WM, whereas the boundaries of GM and CSF can be decomposed into few additional compartments with long relaxation properties.

---

6 We thank GE Healthcare (Munich, Germany) for providing this data.
Inverse Problems 38 (2022) 085002
M Golibabaee and C Poon

75.2. Comparison to the baselines. We compare the performances of the SGB-Lasso and MC-MRF baselines for WM and GM estimation. We used $\beta = 10^{-3}$, $\alpha = 0.8$ parameters for SGB-Lasso. For visualisation outcomes of all tested methods we hard thresholded to exactly three groups: the T1/T2 values of top two compartments with highest energy ($\ell^2$ norm) mixture maps were reported as WM and GM. Remaining compartments with longer T1/T2 times were averaged (weighted by their mixture maps energies) and mapped to a third group. Given these T1/T2 estimates, the (thresholded) maps were computed by simulating a three-atom discrete dictionary and solving a small-size nonnegative least squares for (4). Figure 8 compares the thresholded mixture maps of the SGB-Lasso and baselines. All methods used the LRTV reconstruction before demixing. In separate figures 7(b) and (e)–(h) the T1/T2 values of all estimated compartments before thresholding are scatter plotted. Table 2 compares the estimated T1/T2 values of (thresholded) WM and GM to their literature values at 1.5 T, and table 3 summarises the algorithms runtimes.

In figure 8 SGB-Lasso outperforms the baselines in terms of the visual appearance of the mixture maps, their spatial separability and correspondence to the anatomical structures of WM, GM and CSF. Further, the estimated T1/T2 values for WM and GM are within the range of literature values (table 2). The CSF relaxations underestimated the literature values and were excluded from comparisons. This issue was previously reported in all MC-MRF baselines (see e.g. [28]) due to the in vivo pulsations that are not captured by the MRF’s signal model [10]. BayesianMRF does not exploit group sparsity and results in many unclustered compartments (figure 7(g)) and poor mixture maps visualisation. Baselines SPIJN and PVMRF improve this thanks to the compartment grouping, however the WM map using PVMRF is not well separated from the CSF region, and additionally in SPIJN the WM map is not well separated from the GM region (figure 8). SPIJN overestimated GM’s T1/T2 and underestimated WM’s T2 relaxations. BayesianMRF underestimated GM’s relaxations, and PVMRF slightly overestimated GM’s T2 value. The non-iterative PVMRF was the fastest method, followed by the SGB-Lasso which is the fastest amongst tested model-based iterative algorithms (table 3). Notably, SGB-Lasso implemented at a high level of accuracy (objective tolerance $= 10^{-8}$ and the longest runtime, outputs unclustered compartments (figure 7(h)) and poor mixture maps (figure 8), highlighting
the significance of an off-the-grid alternative to overcome the fundamental limitation of sparse approximation in a highly coherent discretised dictionary.

Similar comparisons are illustrated in supplementary figures A1 where LRTV is replaced by SVD-MRF [30] reconstruction i.e. the method previously adopted by all baselines. SGB-Lasso also outperforms baselines, but as can be observed, SVD-MRF produces undersampling artefacts that propagate to the demixing step and perturb the mixture maps. SPIJN demixing favours SVD-MRF albeit outputting noisy (aliased) mixture maps indicating its high sensitivity to TSMI variations, and that the LRTV’s spatial smoothing favours enforcing pixel-sparsity besides group-sparsity (as in SGB-Lasso) at the demixing step.

Table 2. Estimated T1/T2 values for in vivo WM and GM compartments using MC-MRF algorithms compared to the 1.5 T literature values [54] \([55]^{+}\).

| Tissue | Literature | SGB-Lasso | PVMRF | SPIJN | BayesianMRF |
|--------|------------|-----------|-------|-------|-------------|
|        |            | T1 (ms)   |       |       |             |
| WM     | 694–862\(^+\) | 829       | 806   | 699   | 821         |
| GM     | 1074–1174\(^+\) | 1143      | 1165  | 1483  | 874         |
|        |            | T2 (ms)   |       |       |             |
| WM     | 68–87\(^+\) | 81        | 80    | 51    | 77          |
| GM     | 87–103\(^+\) | 102       | 105   | 164   | 82          |
8. Discussions

We introduced a novel off-the-grid approach to address the non-scalability of the dictionary-based MC-MRF baselines. We observed that a voxel-sparsity alone (BayesianMRF) results in inferior demixing performance compared to the group-sparse models (PVMRF, SPIJN) that cluster the entire image into few compartments. The proposed SGB-Lasso improves upon both approaches by simultaneously promoting both sparsity types via SGTV regularisation. This regularisation provides the flexibility of promoting a desired level of spatial sparsity (i.e. certain level of pixel purity) within the mixture maps of the estimated sparse compartments. In our simulations (section 7.3) we also observed that SGB-Lasso was more robust than other baselines for separating less pixel-pure (more challenging) mixture distributions. Further, the SGB-Lasso was able to separate the WM, GM and CSF regions of healthy brain in vivo more precise than the baselines. Estimated T1/T2s for the WM and GM were consistent with the literature. The WM region was separated in a single compartment whereas the boundaries of GM and CSF (pre-thresholded) were decomposed into few additional compartments with long relaxation properties (also reported in [27]). Further in vivo validations are required to confirm repeatably of the results and their usage for clinical applications.

A T1/T2-encoding MRF sequence was used in our experiments. Encoding more parameters could potentially separate more complex e.g. pathology-related compartments. It will be interesting to examine potentials of this work in applications encoding larger number of parameters e.g. [56, 57], where dictionary-based gridding could create a major precision vs storage bottleneck to the MC-MRF framework. The potential applications of SGB-Lasso to a wider range of quantitative MRI sequences with other forms of Bloch responses than MRF remains an interesting direction for future studies. Some imaging protocols e.g. the exponential models [18], may have analytically differentiable Bloch response models that could be directly accommodated by algorithm 1 without requiring neural network approximations. Also, current implementation of SGB-Lasso uses the L-BFGS quasi Newton method which is an accurate but slow nonlinear fitting solver. Stochastic gradient methods [58–60] could be an interesting way forward to accelerate the SGB-Lasso’s computations.

Previous works [11, 27, 28] reported great sensitivity (e.g. in terms of noise amplification in mixture maps) to TSMIs’ noise and under-sampling artefacts. For this we replaced SVD-MRF by a spatiotemporally regularised model-based reconstruction LRTV for enhancing demixing results. While further numerical advances for MRF reconstruction will benefit the current work, we believe that future works combining tasks of reconstruction and mixture separation could be more efficient way forward (e.g. see multi-task compressed sensing examples [61–64]) in order to numerically tackle shorter and aggressively under-sampled acquisition protocols. Finally, in this work, there were two parameters to tune, the regularisation parameter $\lambda > 0$ and the parameter $\beta$ which balances spatial and component sparsity. Automated strategies for turning these parameters is a venue for future work, possible approaches to automatically adapt to the noise level include the extension of square-root Lasso [65] or the use of SURE [66, 67].
9. Conclusion

We introduced a novel off-the-grid approach for the MC-MRF problem. The proposed SGB-Lasso algorithm is an accurate and importantly a scalable alternative to the MC-MRF baselines because it does not rely on fine-gridded multi-parametric MRF dictionaries. The method was theoretically described and its basic feasibility was demonstrated and compared to other baselines in simulations and \textit{in vivo} healthy brain measurements.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Derivation of algorithm 1

We follow the presentation of [41] where the sliding Frank–Wolfe was presented for scalar-valued measures. The Frank–Wolfe algorithm seeks to solve minimisation problems of the form

$$\min_{x \in K} F(x)$$

where $F : \mathcal{X} \rightarrow \mathbb{R}$ is a continuously differentiable function defined on some Banach space $\mathcal{X}$, and $K \subseteq \mathcal{X}$ is a bounded convex set. To derive our algorithm, we follow [41] by rewriting (SGB-Lasso) as the minimisation of a differentiable function over a bounded convex set. Note that the optimal measure satisfies

$$\frac{1}{2}\|\Phi m - X\|_F^2 + \alpha\|m\|_{\beta} \leq \|X\|_F^2 / 2$$

Recalling the definition of $\|\cdot\|_{\beta}$, we can write $\alpha\|m\|_{\beta} = \lambda_1|m|_1 + \lambda_2|m|_2$ where $\lambda_1 = \alpha(1 - \beta)$ and $\lambda_2 = \alpha\beta\sqrt{v}$. So in particular, $|m|_1 \leq \|X\|_F^2 / (2\lambda_1)$. We can therefore equivalently write this as

$$\min_{m \in \mathcal{M}_+} F(t_1, t_2, m) \triangleq \lambda_1 t_1 + \lambda_2 t_2 + \frac{1}{2}\|\Phi m - y\|^2$$

subject to $(t_1, t_2, m) \in K$

where $K \triangleq \{(t_1, t_2, m) | |m|_1 \leq t_1, |m|_2 \leq t_2, t_2 \leq t_1 \leq \|X\|_F^2 / (2\lambda_1)\}$. The differential of $F$ is a bounded linear operator on $\mathbb{R} \times \mathbb{R} \times \mathcal{M}_+(T; \mathbb{R}^p)$. Writing $s \triangleq (t_1, t_2, m)$ and $s' \triangleq (t_1', t_2', m')$, we have

$$[dF(s)]s' = \lambda_1 t_1' + \lambda_2 t_2' + \int [\Phi'(\Phi m - y)](\theta)dm'(\theta)$$

The Frank–Wolfe algorithm is an iterative algorithm consisting of the following two steps: denoting the $k$th iterate by $x^k$, do

(a) $x^k \in \arg\min_{s \in K} dF(x^k)(s)$
Appendix B. Proof of theorem 1

B.1. Notations

For \( k \in \mathbb{N} \), let \([ k ] \defeq \{1, \ldots, k \} \). For \( k, d \in \mathbb{N} \), let \( \mathbf{0}_k \) denote the vector of \( k \) zeros, \( \mathbf{0}_{k \times d} \) as the matrix of size \( k \times d \) with all zero entries and \( \text{Id}_d \) denotes the identity matrix on \( \mathbb{R}^d \). Given a matrix \( \mathbf{Q} \in \mathbb{R}^{n \times m} \), let \( \text{Vec}_{n,m}(\mathbf{Q}) \in \mathbb{R}^{nm} \) be its vectorized version with columns stacked vertically, let \( \text{Vec}_{n,m}^{-1} \) be the inverse operation, so that \( \text{Vec}_{n,m}^{-1}(\text{Vec}_{n,m}(\mathbf{Q})) = \mathbf{Q} \). Given \( \beta > 0 \), we
define the soft-thresholding operator by $S_{\beta} : \mathbb{R}^v \rightarrow \mathbb{R}^v$ is defined by

$$S_{\beta}(\xi)_i = \begin{cases} 
\xi_i - \beta & \text{if } \xi_i > \beta, \\
\xi_i + \beta & \text{if } \xi_i < -\beta, \\
0 & \text{if } |\xi_i| \leq \beta.
\end{cases} \quad (B.1)$$

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 entries indexed $I$. Given a point $x \in \mathbb{R}^n$ and $r > 0$, we denote by $B(x, r) \overset{\text{def}}{=} \{ z \mid \|x - z\| < r \}$ the open ball of radius $r$ around $x$. Given $x \in \mathbb{R}^n$, $x_+$ is the positive part of $x$. The adjoint of $\Phi$ is denoted by $\Phi^* : \mathbb{R}^{T \times v} \rightarrow \mathcal{C}(T; \mathbb{R}^v)$ which maps matrices into the space of vector-valued continuous functions. Given a (e.g. TSMI) matrix $X \in \mathbb{R}^{T \times v}$ with columns $x_i$:

$$\Phi^* X \overset{\text{def}}{=} (\eta_i)_{i=1}^v \quad \text{where} \quad \eta_i(\theta) = \langle \phi(\theta), x_i \rangle. \quad (B.2)$$

Given a proper, convex, lower semicontinuous function $f$ on $T$, we denote by $\partial f(\theta)$ its subdifferential at $\theta \in T$. Given a smooth function $f : T \rightarrow \mathbb{R}$, we denote its Hessian at $\theta \in T$ by $\nabla^2 f(\theta)$. Given a smooth function $f : T \rightarrow \mathbb{R}^T$, denote by $J_f(\theta) \in \mathbb{R}^{T \times d}$ the Jacobian of $f$ at $\theta \in T$.

To simplify notation, we denote the Dirac at position $\theta$ by $\delta_{\theta} = \delta(\cdot - \theta)$. Moreover, let

$$\lambda_1 \overset{\text{def}}{=} (1 - \beta) \quad \text{and} \quad \lambda_2 \overset{\text{def}}{=} \sqrt{v\beta}, \quad (B.3)$$

so

$$\|m\|_{\beta} = \lambda_1|m|_1 + \lambda_2|m|_2. \quad (B.4)$$

**Outline of proof.** Appendix B.2 describe the dual problem of (SGB-Lasso) and appendix B.3 describes how dual solutions can be used to study support stability. These are the analogous results to [15] in the case of vector-valued measures. In particular, following [15], nondegeneracy of the limit dual solution as $\alpha \rightarrow 0$, denoted by $Q_0$, guarantees support stability (see proposition 2) and $Q_0$ precisely coincides with $Q_*$ from definition 1 under the nondegeneracy conditions.

**B.2. Duality and variational formulation of the sparse-group norm**

We first review some duality results for vector norms described in [68]: let

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

defined for $x \in \mathbb{R}^n$. For $\epsilon \in (0, 1)$, the $\epsilon$-norm denoted by $\|x\|_{\epsilon}$ for $x \in \mathbb{R}^n$, is defined as the unique $\nu > 0$ such that

$$\sum_i (|x_i| - (1 - \epsilon)\nu)^2 + (\epsilon\nu)^2 = 0. \quad (B.5)$$
We can also write: for all \( \lambda_1, \lambda_2 > 0 \),

\[
\{ x + y \mid \| x \|_2 \leq \lambda_1 \nu, \| y \|_\infty \leq \lambda_2 \nu \} = \left\{ z \mid \| S_{\lambda_2}(z) \|_2^2 \leq (\lambda_1 \nu)^2 \right\}
\]

These two norms are dual to each other, as it is shown in [69, appendix E, lemmas 1 and 2] (see also [68]) that

\[
\{ x + y \mid x, y \in \mathbb{R}^d, \| x \|_2 \leq \epsilon \nu, \| y \|_\infty \leq (1 - \epsilon) \nu \} = \left\{ z \in \mathbb{R}^d \mid \| z \|_\epsilon \leq \nu \right\}
\]

and hence, since \( J \) is the support function of the set in (B.6), the dual norm of \( J \) is the \( \epsilon \)-norm. Moreover, we have the unique \( \epsilon \)-decomposition

\[
\xi = S_\epsilon(\xi) + (\xi - S_\epsilon(\xi)) \quad (B.7)
\]

with \( \| S_\epsilon(\xi) \|_2 = (1 - \epsilon) \| \xi \|_2 \) and \( \| \xi - S_\epsilon(\xi) \|_\infty = \epsilon \| \xi \|_\infty \), where we recall that \( S_\epsilon \) is the soft-thresholding operator. Therefore, by considering the dual norms of \( |m|_1 \) and \( |m|_2 \), the following holds

\[
\| m \|_\beta = \sup_{f \in K_0} \langle f, m \rangle \quad (B.8)
\]

where

\[
K_0 = \left\{ f \in C(T; \mathbb{R}^v) \mid \sup_{\theta \in T} \| S_{\lambda_1}(f(\theta)) \|_2^2 \leq \lambda_2^2 \right\}
\]

**Proposition 1. (Dual problem).** For \( \alpha > 0 \), the dual problem to (SGB-Lasso) is

\[
\underset{Q \in \mathcal{K}}{\sup} \langle X, Q \rangle_F - \alpha \| Q \|_F^2 \quad (D_\alpha(X))
\]

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

\[
\mathcal{K} \overset{\text{def}}{=} \left\{ Q \mid \sum_{i=1}^{v} \| \Phi^\ast Q(\theta)_i - \lambda_1 \|^2_2 \leq \lambda_2^2 \right\} \quad (B.10)
\]

The primal and dual problems are related by \( m \) solves (SGB-Lasso) if and only if \( Q = \frac{X - \Phi m}{\alpha} \) solves (D_\alpha(X)). Moreover, \( \Phi^\ast Q \in \partial \| m \|_\beta \).
In the case of $\alpha = 0$, the dual of the limit problem

$$
\min_m \|m\|_\beta \ s.t. \ \Phi m = X
$$

is $(D_\alpha(X))$ with $\alpha = 0$. Moreover, if $Q$ solves $(D_\alpha(X))$ and $m$ solves $(P_0(X))$, then $\Phi^*Q \in \partial\|m\|_\beta$

**Proof.** In (B.9), we can restrict the set $K_0$ to positive functions $K_+ = K_0 \cap C(\mathcal{T}; \mathbb{R}_+^n)$ since $m$ is a positive measure. Therefore, the convex conjugate of $\|m\|_\beta$ is $\iota_{K_+}$, the indicator function on the set $K_+$. The result now follows by applying the Fenchel–Rockafellar duality theorem [70, theorem 4.2].

**B.3. Support stability**

Given a dual solution $Q_\alpha$ to $(D_\alpha(X))$, the function

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

(B.11)

characterizes the support of any primal solution $m_\alpha$ of (SGB-Lasso) in the following sense:

**Lemma 1.** Any solution $m_\alpha$ to (SGB-Lasso) satisfies

$$
\text{Supp}(m_\alpha) \subseteq \{\theta \in \mathcal{T} : \|f_\alpha(\theta)\| = 1\}.
$$

(B.12)

If $m_\alpha = \sum_i c_i \delta(\theta - \theta_i)$ is a discrete measure, then for each $s$, $\text{Supp}(c_s) \subseteq \{j \in [n] : |(\Phi^*Q_\alpha)(\theta_i)|_j > \lambda_1\}$ and $f_\alpha(\theta_i) = c_i/\|c_i\|_2$.

**Proof.** From proposition 1, $\Phi^*Q_\alpha \in \partial\|m\|_\beta = \lambda_1 \partial|m|_1 + \lambda_2 \partial|m|_2$. From (B.9), if $\xi$ satisfies $\sum_i (\xi - \lambda_1)^2 \leq \lambda_2^2$ then $\|S_{\lambda_1}(\xi)\|_2 \leq \lambda_2$ and $\|\xi - S_{\lambda_1}(\xi)\|_\infty \leq \lambda_1$. So

$$
S_{\lambda_1}(\Phi^*Q_\alpha) \in \lambda_2 \partial|m|_2 \quad \text{and} \quad \Phi^*Q_\alpha - S_{\lambda_1}(\Phi^*Q_\alpha) \in \lambda_1 \partial|m|_1
$$

(B.13)

The first inclusion follows from the first equality. For the second inclusion, note that $\Phi^*Q_\alpha - S_{\lambda_1}(\Phi^*Q_\alpha) \in \lambda_1 \partial|m|_1$ means that given $s \in [m]$ and $I_s = \text{Supp}(c_s)$,

$$
(\Phi^*Q_\alpha(\theta_i) - \max \{(\Phi^*Q_\alpha)(\theta_i) - \lambda_1, 0\})_{I_s} = \lambda_1 \text{ sign}(c_s)_{I_s}
$$

(B.14)

where given a vector $z \in \mathbb{R}^n$, $\text{sign}(z_i) = z_i/|z_i|$ for $i \in [n]$. If $\Phi^*Q_\alpha(\theta_i)_{I_s} < \lambda_1$ for $j \in I_s$, then this equation implies that $\Phi^*Q_\alpha(\theta_i)_{I_s} = \lambda_1 \text{ sign}(c_s)_{I_s}$ which is a contradiction. Therefore, $I_s \subset \{j \in I_s : \Phi^*Q_\alpha(\theta_i)_{I_s} > \lambda_1\}$. \qed

Note that $(D_\alpha(X))$ has a unique solution, since it can be seen as the projection of $X/\alpha$ onto the closed convex set $K$. Moreover, the previous lemma shows that its solution characterises the support of any primal solution $m_\alpha$ of (SGB-Lasso). Therefore, to understand the structure of solutions to (SGB-Lasso) with $X = \Phi m + W$ with $\|W\|_F \leq \epsilon$, it suffices to study the solution of the dual problem $(D_\alpha(X))$, which we denote by $Q_{\alpha,\epsilon}$. Following [15], we can show that $Q_{\alpha,\epsilon}$ has a limit as $\epsilon/\alpha$ and $\alpha$ converge to 0: define

$$
Q_0 \in \arg \min \left\{\|Q\|_F \ | \eta \overset{\text{def}}{=} \Phi^*Q \in K, \langle \eta, m \rangle = \|m\|_\beta \right\}.
$$

(B.15)
**Lemma 2.** Consider \( \mathbf{Q}_{\alpha, \epsilon} \) and \( \mathbf{Q}_0 \) defined above. We have \( \| \mathbf{Q}_{\alpha, \epsilon} - \mathbf{Q}_0 \|_F \to 0 \) as \( \alpha \to 0 \), and

\[
\| \mathbf{Q}_{\alpha, \epsilon} - \mathbf{Q}_{\alpha, 0} \|_F \leq \epsilon / \alpha. \tag{B.16}
\]

**Proof.** The proof is omitted as it is verbatim the proof of proposition 1 in [15]. \( \square \)

The minimal norm element \( \mathbf{Q}_0 \) satisfies \( \mathbf{Q}_0 \) is a solution to the dual problem \( (D_{\alpha}(\mathbf{X})) \) with \( \alpha = 0 \) and \( \mathbf{X} = \Phi \mathbf{m} \). Moreover, from lemma 1, for a discrete measure \( \mathbf{m} = \sum_i c_i^* \delta_{\theta_i} \), we in fact have

\[
\mathbf{Q}_0 \in \arg \min \left\{ \| \mathbf{Q} \|_F \left| \sup_{\theta \in \mathcal{T}} \| f_Q(\theta) \|_2 \leq 1, \quad \forall s, f_Q(\theta_s) = \frac{c_i^*}{\| c_i \|_2} \right. \right\}, \tag{B.17}
\]

where we denote \( f_Q \equiv \frac{1}{\lambda_2} (\Phi^* \mathbf{Q} - \lambda_1)_+ \) inside the constraint.

**Definition 2.** Define \( f_{Q_0} = \frac{1}{\lambda_2} (\Phi^* \mathbf{Q}_0 - \lambda_1)_+ \) and \( g_0(\theta) \equiv \| f_{Q}(\theta) \|_2^2 \). We call \( \mathbf{g}_0 \) is nondegenerate minimal norm certificate with respect to the sparse measure \( \mathbf{m} = \sum_{s} c_i^* \delta_{\theta_s} \) if \( \mathbf{g}_0 \)

(a) (non-saturation) \( g_0(\theta) < 1 \) for all \( \theta \notin \{ \theta_i \} \)

(b) (curvature) \( \nabla^2 g_0(\theta_s) \prec 0 \) for all \( s \in [m] \).

Note that by definition, \( g_0(\theta_s) = 1 \) for all \( s \in [m] \), so this condition says that \( \mathbf{g}_0 \) saturates at its maximum value 1 only on \( \{ \theta_s \}_{s \in [m]} \), and (ii) is a curvature condition on \( \mathbf{g}_0 \) at these saturation points.

**B.4. Precertificates**

We show below that nondegeneracy of \( \mathbf{g}_0 \) implies support stability. However, in general, \( \mathbf{g}_0 \) 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 \( \mathbf{g}_* \) [15], a candidate certificate which could be computed by solving a linear system. This precertificate precisely leads to the certificate conditions described in definition 1.

Notice that since \( g_0(\theta) \leq 1 \) for all \( \theta \) and \( g_0(\theta_s) = 1 \) for all \( s \in [m] \), it is necessary that \( \nabla g_0(\theta_s) = 0 \). Replacing the constraint of \( \| f(\theta) \|_2 \leq 1 \) for all \( \theta \in \mathcal{T} \) with \( \nabla \| f(\theta_s) \|_2^2 = 0 \) for all \( s \in [m] \) leads to the definition of \( \mathbf{g}_* \) in definition 1 (recall that \( \mathcal{I}_c \) denotes the support of \( \mathbf{c}_* \)). Notice that if \( g_*(\theta) \leq 1 \) for all \( \theta \in \mathcal{T} \), then it follows that \( \mathbf{Q}_* = \mathbf{Q}_0 \) is the minimal norm solution from (B.17). Clearly, if \( \mathbf{g}_* \) is nondegenerate, then \( \mathbf{g}_0 = \mathbf{g}_* \) is also nondegenerate.

**B.4.1. The precertificate as a least squares solution.** The attractiveness of \( \mathbf{Q}_* \) stems from the fact that it is defined via \( \sum_i |\mathcal{I}_c| + md \) linear equations, and hence, \( \mathbf{Q}_* \) can be computed by solving a linear system.

Observe that the constraints \( [f(\theta_s)]_{s \in [m]} \equiv \frac{c_i^*}{\| c_i \|_2} \) for all \( s \in [m] \) in definition 1 can be written as

\[
\mathcal{P}_\mathcal{T} \mathbf{Vec}(D_{\theta}^* \mathbf{Q}) = \mathcal{P}_\mathcal{T}[\mathbf{Id}_c \otimes D_{\theta}^*] \mathbf{Vec}(\mathbf{Q}) = \mathbf{u}_0 \tag{B.18}
\]
where $\otimes$ is the Kronecker product and

$$
u_0 = (\lambda_1 + \lambda_2 [c_i, ||c_i||])^{m_i} \in \mathbb{R}^{\sum_{s=1}^{|\theta|}},$$  

(B.19)

$D_{\theta}$ is the matrix with columns $\phi(\theta_j)$, and $P_T: \mathbb{R}^n \rightarrow \mathbb{R}^{\sum_{s=1}^{|\theta|}}$ is the subsampling operator given by which selects the nonzero entries of $\{I_j\}_{j \in [m]}$, so that given a matrix $Z \in \mathbb{R}^{m \times n}$ with ith row $Z_s \in \mathbb{R}^n$ for $s \in [m]$, $P_T \text{Vec}(Z) = ([Z_s]_i)_{i \in [m]}$.

The constraints $\nabla||f(\theta_j)||^2_2 = 0$ for all $s \in [m]$ can be written as

$$
\begin{align*}
\theta_d &= \lambda_2 \sum_{i \in I_s} f_i(\theta) \nabla f_i(\theta) = \frac{1}{||c_i||^2} \sum_{i=1}^v (c_i, \nabla \Phi^T \Phi)(\theta) \\
&= \Phi(\theta) \Phi^T = \frac{1}{||c_i||^2} [c_i, \Omega_{\Phi}(\theta)]_{\Phi^T}(\Phi(\theta))
\end{align*}

(B.20)

We can therefore define the $T \times (\sum_{s=1}^m |I_s| + md)$ matrix

$$
\Gamma = \left[(\text{Id} \otimes D_{\theta})P_T^T, \frac{c_1}{||c_1||^2} \otimes \Omega_{\Phi}(\theta_1), \ldots, \frac{c_k}{||c_k||^2} \otimes \Omega_{\Phi}(\theta_k)\right]

(B.21)

and write

$$Q = \text{Vec}_{\Gamma}^{-1} \left(\left[I_0, 0_{md}\right]\right).$$

(B.22)

Note that $\Gamma$ depends on $\theta$ and $\{c_i/||c_i||\}$. To make this dependence clear, we will sometimes write $\Gamma_{\theta}$ in place of $\Gamma$.

**B.5. Stability**

**Proposition 2. (Result on stability).** If $g_0$ is nondegenerate and $D_{\theta}$ is full rank, then $(P_{\theta}(X))$ with $X = \Phi \left(\sum_{j=1}^m c_j \delta_{\theta_j}\right)$ has a unique solution and provided that $\epsilon/\alpha$ is sufficiently small, the solution to $(SGB-Lasso)$ with $X = \Phi \left(\sum_{j=1}^m c_j \delta_{\theta_j}\right) + W$ and $||W||_{\epsilon} \leq \epsilon$ is unique and of the form $m_{\alpha, \epsilon} = \sum_{j=1}^m c_j \delta_{\theta_j}$ where $\text{Supp}(c_j) = \text{Supp}(c_i)$. Moreover, $c_j \rightarrow c_j$ and $\hat{\theta}_j \rightarrow \theta_j$ as $\alpha$ and $\epsilon/\alpha \rightarrow 0$.

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

$$||Q_{\alpha, \epsilon} - Q_{\alpha, \epsilon}|| \leq ||P_{\mathcal{K}}(X/\alpha) - P_{\mathcal{K}}(X + W)/\alpha|| \leq ||W||_{\epsilon}/\alpha,$$

(B.23)

we have that $v_{\alpha, \epsilon} \rightarrow v_0$ in the uniform norm as $\alpha$ and $\epsilon/\alpha$ converge to 0. So, if $g_0$ is non-degenerate, then given any $r > 0$, provided that $||W||_{\epsilon}/\alpha$ and $\alpha$ are sufficiently small, letting $\eta(\theta) \equiv \frac{1}{s} \left(||v_{\alpha, \epsilon}(\theta) - \lambda_v||\right)^2$, we have $\eta(\theta) < 1$ for all $\theta < \cup B(\theta, r)$, and for all $\theta \in B(\theta, r)$, $\nabla^2 \eta(\theta) < 0$. So, there are at most $k$ points for which $\eta(\theta) = 1$. So, by lemma 1, given data $X = \Phi m + W$, we recover at most $k$ components with $m_{\alpha, \epsilon} = \sum_{j=1}^m c_j \delta_{\theta_j}$. Moreover, uniform convergence of $v_{\alpha, \epsilon}$ to $v_0$ also ensures that $\text{Supp}(c_j) \subseteq \text{Supp}(c_i)$ for $\alpha$ and $\epsilon/\alpha$
sufficiently small. Finally, since $D_\theta$ is full rank, the solution $\mathcal{m}_0 = \sum_{j=1}^{m} c_j^T \delta_j$ to $(P_\theta(X))$ is unique and the solutions to (SGB-Lasso) converge weak-$*$ to $\mathcal{m}_0$ [15]. So, $\mathcal{m}_{\alpha,\epsilon}$ consists precisely of $k$ spikes and $\text{Supp}(\hat{c}_j) = \text{Supp}(c_j)$ provided that $\alpha$ and $\epsilon/\alpha$ sufficiently small. Note that convergence of the support $\hat{\theta}_j \rightarrow \theta_j$ follows due to convergence of the dual certificates, while convergence of $\hat{c}_j \rightarrow c_j$ follow due to weak-$*$ convergence of $\mathcal{m}_{\alpha,\epsilon}$ towards $\mathcal{m}_0$. Uniqueness of solutions to (SGB-Lasso) is because $D_\theta$ is also full rank whenever $\epsilon$ is sufficiently small.

The proposition above establishes that for $\alpha$ and $\epsilon/\alpha$ sufficiently small, $\mathcal{m}_{\alpha,\epsilon}$ consists of precisely $k$ components. Theorem 1 is now a consequence of the implicit function theorem. By applying the implicit function theorem, we can show that there is a $C^1$ function $G$ which maps the parameter $\alpha$ and noise $W$ to solutions of (SGB-Lasso). That is, $G : (W, \alpha) \in \mathbb{R}^{T \times v} \times \mathbb{R}_+ \mapsto (\tilde{C}, \tilde{\theta}) \in \mathbb{R}_+^N \times T^m$, so that $(\tilde{C}, \tilde{\theta}) = G(W, \alpha)$ correspond to the solution $\mathcal{m}_0$ of (SGB-Lasso) with regularisation parameter $\alpha$ and data $\Phi \mathcal{m}_0 + W$ for $(W, \alpha)$ on a neighbourhood around $0$. The stated error bounds then follow since the recovered parameters $\tilde{C}$ and $\tilde{\theta}$ are continuously differentiable with respect to $\alpha$ and $W$, since

$$
\left\| (\tilde{C}, \tilde{\theta}) - (C, \theta) \right\| = \| G(W, \alpha) - G(0, 0) \| \leq A \| (W, \alpha) - (0, 0) \|
$$

where $A > 0$ bounds the Jacobian of $G$, and we write $\| (C, \theta) \| = \| C \|_p + \| \theta \|_p$.

To this end, let $N = \sum_{i} |I_i|$ and define a function

$$
F : \mathbb{R}^N_+ \times T^m \times \mathbb{R}^{T \times v} \times \mathbb{R}_+ \rightarrow \mathbb{R}^{N + kd}
$$

so that given $\hat{C} = \{ \hat{c}_i \}_{i \in [m]}$ with $\hat{c}_i \in \mathbb{R}^{\left| I_i \right|}$, $\hat{\theta} \in T^m$, $W \in \mathbb{R}^{T \times v}$ and $\alpha \in \mathbb{R}_+$,

$$
F(\hat{C}, \hat{\theta}, W, \alpha) = \left[ \left( g_s(\hat{C}, \hat{\theta}, W, \alpha) \right)_{s=1}^m \right] \left[ \left( h_s(\hat{C}, \hat{\theta}, W, \alpha) \right)_{s=1}^m \right]
$$

(B.25)

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

$$
g_s(\hat{C}, \hat{\theta}, W, \alpha)^T = \left( \phi(\hat{\theta}_s)^T [D_\theta \hat{C}^T - D_\theta C^T - W] \right)_{I_s} + \alpha \left( \lambda_1 + \lambda_2 \frac{\hat{c}_s^T}{\| \hat{c}_s \|} \right)
$$

(B.26)

and

$$
h_s(\hat{C}, \hat{\theta}, W, \alpha) = J_{\phi}(\hat{\theta}_s)^T \left( D_\theta \hat{C}^T - D_\theta C^T - W \right) \frac{\hat{c}_s}{\| \hat{c}_s \|_2}
$$

(B.27)

Here, $\hat{C} \in \mathbb{R}^{T \times m}$ is the matrix with $x$th column satisfying $(\hat{c}_i)_x = \hat{c}_i$ and $(\hat{c}_i)_x = 0$. Observe that if $(\hat{C}, \hat{\theta})$ correspond to a solution of (SGB-Lasso) with data $X = D_\theta C^T + W$, then $F(\hat{C}, \hat{\theta}, W, \alpha) = 0$, since $(g_s)_s = 0$ correspond to the condition that the dual certificate should take values $\hat{c}_s/\| \hat{c}_s \|$ on the support $\theta_s$, and $(h_s)_s = 0$ correspond to the condition that the gradient of the dual certificate is 0.
The partial derivatives of $g \equiv (g_{\ell})$ and $h \equiv (h_{\ell})$ are as follows: define

$$Z \equiv D_{g} C^{\top} - D_{g'}(C')^{\top} - W,$$

then

$$\partial C g = P_{I} (\text{Id}_{n} \otimes D_{g}^{\top} D_{g}) P_{I}^{\top} + \alpha \lambda_{2} \text{diag} \left( \frac{1}{\|e_{i}\|_{2}} \text{Id}_{|d|} - \frac{\mathbf{c}_{i}}{\|e_{i}\|_{2}} \right)_{x \in [m]}$$

$$\partial g = \text{diag} \left( [Z_{\ell}]^{\top} J_{\phi}(\hat{\theta}_{\ell}))_{x \in [m]} + \left( \frac{\mathbf{c}}{\|e_{i}\|_{2}} \right)^{\top} J_{\phi}(\hat{\theta}_{\ell}))_{x \in [m]} \right)_{x \in [m]}$$

$$\partial \phi = \left( \lambda_{1} + \lambda_{2} \frac{\mathbf{c}}{\|e_{i}\|_{2}} \right)$$

$$\partial w = P_{I} (\text{Id}_{n} \otimes D_{g}^{\top})$$

where $Z_{\ell}$ denotes the matrix $Z$ with columns restricted to those indexed by $I_{\ell}$.

Let $\mathbb{H}_{\phi}(\hat{\theta})^{\top} \in \mathbb{R}^{d \times d \times T}$ so that is $(i, j, n)$ entries with $i, j \in [d]$ for the Hessian of $\phi_{n}(\hat{\theta}_{s})$. So, given a vector $z = (z_{i})_{n=1}^{T}, \mathbb{H}_{\phi}(\hat{\theta})^{\top} z = \sum_{m=1}^{T} z_{m} \nabla^{2} \phi(\hat{\theta}) \in \mathbb{R}^{d \times d}$. Then,

$$\partial C h = \text{diag} \left( J_{\phi}(\hat{\theta})^{\top} Z_{\ell} (\mathbf{c}_{i}^{\top} \text{Id}_{|d|} - \frac{\mathbf{c}_{i}^{\top}}{\|e_{i}\|_{2}})_{x \in [m]} \right)$$

$$+ \left( \frac{1}{\|e_{i}\|_{2}} \mathbf{c}_{i}^{\top} \otimes J_{\phi}(\hat{\theta})^{\top} \right) \left( \text{Id}_{n} \otimes D_{g} \right)_{x \in [m]}$$

$$\partial h = \text{diag} \left( \mathbb{H}_{\phi}(\hat{\theta})^{\top} Z \frac{\mathbf{c}_{i}}{\|e_{i}\|_{2}} + \left( \frac{1}{\|e_{i}\|_{2}} J_{\phi}(\hat{\theta})^{\top} J_{\phi}(\hat{\theta}) \mathbf{c}_{i}^{\top} \right)_{x \in [m]} \right)_{x \in [m]}$$

$$\partial \phi h = 0$$

$$\partial w = -\left( \frac{1}{\|e_{i}\|_{2}} \mathbf{c}_{i}^{\top} \otimes J_{\phi}(\hat{\theta})^{\top} \right)_{x \in [m]}$$

We therefore have

$$\partial \ell_{\phi, \theta} F = (\Gamma_{\phi}^{\top} \Gamma_{\theta} + Y) \left( \text{Id}_{N} \otimes \text{diag} \left( \|e_{i}\|_{2} \right)_{i=1}^{m} \otimes \text{Id}_{d} \right)$$

where

$$Y \equiv \begin{pmatrix}
\alpha \lambda_{2} \text{diag} \left( \frac{1}{\|e_{i}\|_{2}} \text{Id}_{|d|} - \frac{\mathbf{c}_{i}}{\|e_{i}\|_{2}} \right)_{x \in [m]} \\
\text{diag} \left( \mathbb{H}_{\phi}(\hat{\theta})^{\top} Z_{\ell} (\text{Id}_{|d|} \frac{\mathbf{c}_{i}}{\|e_{i}\|_{2}}) \right)_{x \in [m]}
\end{pmatrix}$$

(B.32)
and

\[ \partial_{(\alpha, W)} F = \begin{bmatrix} \left( \frac{\lambda_1 + \lambda_2}{\|\hat{c}\|_2} \right) \cdot \hat{c}, \Gamma_\theta^\top \end{bmatrix} \]  

(B.33)

Note that \( \Gamma_\theta \) is full rank provided that

\[ [D_\theta \ J_\phi(\theta_1) \ldots J_\phi(\theta_k)] \]  

(B.34)

is full rank. In particular, \( \partial_{(\hat{C}, \hat{\theta})} F(C, \theta, 0, 0) \) is invertible. Hence, the implicit function theorem applies, and there is a \( C^1 \) function \( G : (W, \alpha) \in \mathbb{R}^{T \times v} \times \mathbb{R}_+ \mapsto (\hat{C}, \hat{\theta}) \in \mathbb{R}_+^{N} \times T^m \), such that

\[ F(\hat{C}, \hat{\theta}, W, \alpha) = 0. \]

For \( \alpha \) and \( \|W\|/\alpha \) sufficiently small, this coincides with solutions to (SGB-Lasso) since \( F(\hat{C}, \hat{\theta}, W, \alpha) = 0 \) along with the nondegeneracy of \( Q_0 \) imply that the certificate corresponding the first order optimality condition holds for \( \hat{C}, \hat{\theta} \).

**ORCID iDs**

Mohammad Golbabaee \( \text{https://orcid.org/0000-0001-5822-2990} \)

**References**

[1] Collins D L, Zijdenbos A P, Kollokian V, Sled J G, Kabani N J, Holmes C J and Evans A C 1998 Design and construction of a realistic digital brain phantom IEEE Trans. Med. Imaging 17 463–8

[2] Tohka J, Zijdenbos A and Evans A 2004 Fast and robust parameter estimation for statistical partial volume models in brain MRI Neuroimage 23 84–97

[3] Van Leemput K, Maes F, Vandermeulen D and Suetens P 2003 A unifying framework for partial volume segmentation of brain MR images IEEE Trans. Med. Imaging 22 105–19

[4] Cuadra M B, Cammoun L, Butz T, Cuisenaire O and Thiran J-P 2005 Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images IEEE Trans. Med. Imaging 24 1548–65

[5] Manjon J V, Tohka J and Robles M 2010 Improved estimates of partial volume coefficients from noisy brain MRI using spatial context Neuroimage 53 480–90

[6] Jack C R et al 2004 Comparison of different MRI brain atrophy rate measures with clinical disease progression in ad Neurology 62 591–600

[7] Lladó X, Oliver A, Cabezás M, Freixenet J, Vilanova J C, Quiles A, Valls L, Ramió-Torrentà L and Rovira A 2012 Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches Inf. Sci. 186 164–85

[8] Shenton M E, Dickey C C, Frumin M and McCarley R W 2001 A review of MRI findings in schizophrenia Schizophr. Res. 49 1–52

[9] Ma D, Gulani V, Seiberlich N, Liu K, Sunshine J L, Duerr J L and Griswold M A 2013 Magnetic resonance fingerprinting Nature 495 187–92

[10] Jiang Y, Ma D, Seiberlich N, Gulani V and Griswold M A 2015 MR fingerprinting using fast imaging with steady state precession (FISP) with spiral readout Magn. Reson. Med. 74 1621–31

[11] McKeown D, Deshmukh A, Jiang Y, Ma D, Badve C, Sloan A, Gulani V and Griswold M 2018 Bayesian estimation of multicomponent relaxation parameters in magnetic resonance fingerprinting Magn. Reson. Med. 80 159–70

[12] Tang S, Fernandez-Granda C, Lannuzel S, Bernstein B, Lattanzi R, Cloos M, Knoll F and Assländer J 2018 Multicompartment magnetic resonance fingerprinting Inverse Problems 34 094005

[13] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems SIAM J. Imaging Sci. 2 183–202

[14] Candes E J and Fernandez-Granda C 2014 Towards a mathematical theory of super-resolution Commun. Pure Appl. Math. 67 906–56
[15] Duval V and Peyré G 2015 Exact support recovery for sparse spikes deconvolution Found. Comput. Math. 15 1315–55
[16] Boyd N, Schiebinger G and Recht B 2017 The alternating descent conditional gradient method for sparse inverse problems SIAM J. Optim. 27 616–39
[17] Simon N, Friedman J, Hastie T and Tibshirani R 2013 A sparse-group Lasso J. Comput. Graph. Stat. 22 231–45
[18] Whittall K P and MacKay A L 1989 Quantitative interpretation of NMR relaxation data J. Magn. Reson. 84 134–52
[19] Ahlgren A, Wirestam R, Ståhlberg F and Knutsson L 2014 Automatic brain segmentation using fractional signal modeling of a multiple flip angle, spoiled gradient-recalled echo acquisition Magn. Reson. Mater. Phys. Biol. Med. 27 551–65
[20] West J, Warntjes J B M and Lundberg P 2012 Novel whole brain segmentation and volume estimation using quantitative MRI Eur. Radiol. 22 998–1007
[21] Andica C, Hagiwara A, Horii M, Nakazawa M, Goto M, Koshino S, Kamagata K, Kumamaru K K and Aoki S 2018 Automated brain tissue and myelin volumetry based on quantitative imaging with various in-plane resolutions J. Neuroradiol. 45 164–8
[22] Deoni S C L, Matthews L and Kolind S H 2013 One component? Two components? Three? The effect of including a nonexchanging ‘free’ water component in multicomponent driven equilibrium single pulse observation of $T_1$ and $T_2$ Magn. Reson. Med. 70 147–54
[23] Bouhrara M and Spencer R G 2017 Rapid simultaneous high-resolution mapping of myelin water fraction and relaxation times in human brain using BMC-mcDESPOT Neuroimage 147 800–11
[24] Andica C et al 2019 Gray matter alterations in early and late relapsing-remitting multiple sclerosis evaluated with synthetic quantitative magnetic resonance imaging Sci. Rep. 9 8747
[25] Kim H G, Moon W-J, Han J and Choi J W 2017 Quantification of myelin in children using multiparametric quantitative MRI Magn. Reson. Med. 69 101945
[26] McGivney D F, Pierre E, Ma D, Jiang Y, Saybasili H, Gulani V and Griswold M A 2014 SVD compression for magnetic resonance fingerprinting in the time domain IEEE Trans. Med. Imaging 33 2311–22
[27] Nagtegaal M, Konen P, Amthor T and Doneva M 2020 Fast multi-component analysis using a joint sparsity constraint for MR fingerprinting Magn. Reson. Med. 83 521–34
[28] Deshmale A, McGivney D F, Ma D, Jiang Y, Badve C, Gulani V, Seiberlich N and Griswold M A 2019 Partial volume mapping using magnetic resonance fingerprinting NMR Biomed. 32 e4082
[29] Duarte R, Repetti A, Gómez P A, Davies M and Wiaux Y 2020 Greedy approximate projection for magnetic resonance fingerprinting with partial volumes Inverse Problems 36 035015
[30] McGivney D F, Pierre E, Ma D, Jiang Y, Saybasili H, Gulani V and Griswold M A 2014 SVD compression for magnetic resonance fingerprinting in the time domain IEEE Trans. Med. Imaging 33 2311–22
[31] Sbrizzi A, Bruijnjen T, van der Heide O, Luijten P and van den Berg C A 2017 Dictionary-free MR fingerprinting reconstruction of balanced-GRE sequences (arXiv:1711.08905)
[32] Dong G, Hintermüller M and Papafitsoros K 2019 Quantitative magnetic resonance imaging: from fingerprinting to integrated physics-based models SIAM J. Imaging Sci. 12 927–71
[33] Bredies K and Pikkarainen H K 2013 Inverse problems in spaces of measures ESAIM: Control, Optim. Calc. Var. 19 190–218
[34] Golbabaee M, Buonincontri G, Pirki C M, Menzel M I, Menze H B, Davies M and Gómez P A 2020 Compressive MRI quantification using convex spatiotemporal priors and deep encoder-decoder networks Med. Image Anal. 69 101945
[35] Zhao B, Lu W, Hitchens T K, Lam F, Ho C and Liang Z-P 2015 Accelerated MR parameter mapping with low-rank and sparsity constraints Magn. Reson. Med. 74 489–98
[36] Zhao B et al 2018 Improved magnetic resonance fingerprinting reconstruction with low-rank and subspace modeling Magn. Reson. Med. 79 933–42
[37] Mazor G, Weizman L, Tal A and Eldar Y C 2018 Low-rank magnetic resonance fingerprinting Med. Phys. 45 4066–84
[38] Lima da Cruz G, Bustin A, Jaubert O, Schneider T, Botnar R M and Prieto C 2019 Sparsity and locally low rank regularization for mr fingerprinting Magn. Reson. Med. 81 3530–43
[39] Cline C C, Chen X, Mailhe B, Wang Q, Pfeuffer J, Nittka M, Griswold M A, Speier P and Nadar M S 2017 AIR-MRF: accelerated iterative reconstruction for magnetic resonance fingerprinting Magn. Reson. Imaging 41 29–40
[40] Golbabaee M, Chen Z, Wuia Y and Davies M 2019 CoverBLIP: accelerated and scalable iterative matched-filtering for magnetic resonance fingerprint reconstruction Inverse Problems 36 015003
[41] Denoyelle Q, Duval V, Peyré G and Soubies E 2019 The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy Inverse Problems 36 014001
[42] O’donoghue B and Candès E 2015 Adaptive restart for accelerated gradient schemes Found. Comput. Math. 15 715–32
[43] Weigel M 2015 Extended phase graphs: dephasing, RF pulses, and echoes—pure and simple J. Magn. Reson. Imaging 41 266–95
[44] Cybenko G 1989 Approximation by superpositions of a sigmoidal function Math. Control Signals Syst. 2 303–14
[45] LeCun Y, Bengio Y and Hinton G 2015 Deep learning Nature 352 436–44
[46] Chen D, Davies M E and Golbabaee M 2020 Compressive MR fingerprinting reconstruction with neural proximal gradient iterations Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI) pp 13–22
[47] Cohen O, Zhu B and Rosen M S 2018 MR fingerprinting deep reconstruction network (DRONE) Magn. Reson. Med. 80 885–94
[48] Golbabaee M, Chen D, Gómez P A, Menzel M I and Davies M E 2019 Geometry of deep learning for magnetic resonance fingerprinting Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing pp 7825–9
[49] Balsiger F, Shridhar Konar A, Chikop S, Chandran V, Scheidegger G, Geethanath S and Reyes M 2018 Magnetic resonance fingerprinting reconstruction via spatiotemporal convolutional neural networks Int. Workshop Machine Learning for Medical Image Reconstruction pp 39–46
[50] Gómez P A et al 2020 Rapid three-dimensional multiparametric MRI with quantitative transient-state imaging Sci. Rep. 10 13769
[51] Bioucas-Dias J M 2009 A variable splitting augmented Lagrangian approach to linear spectral unmixing IEEE Workshop Hyperspectral Image and Signal Processing: Evolution in Remote Sensing pp 1–4
[52] Collins D L, Zijdenbos A P, Kollokian V, Sled J G, Kabani N J, Holmes C J and Evans A C 1998 BrainWeb: Simulated Brain Database http://brainweb.bic.mni.mcgill.ca/brainweb/
[53] Walsh D O, Gmitro A F and Marcellin M W 2000 Adaptive reconstruction of phased array MR imagery Magn. Reson. Med. 43 682–90
[54] Sled J G and Pike G B 2001 Quantitative imaging of magnetization transfer exchange and relaxation properties in vivo using MRI Magn. Reson. Med. 46 923–31
[55] Stanisz G J, Odobina E E, Pun J, Escaravage M, Graham S J, Bronskill M J and Henkelman R M 2005 T1, T2 relaxation and magnetization transfer in tissue at 3T Magn. Reson. Med. 54 507–12
[56] Wright K L, Jiang Y, Ma D, Noll D C, Griswold M A, Gulani V and Hernandez-Garcia L 2018 Estimation of perfusion properties with MR fingerprinting arterial spin labeling Magn. Reson. Imaging 50 68–77
[57] Jiang Y et al 2017 Simultaneous T1, T2 and diffusion quantification using multiple contrast prepared magnetic resonance fingerprinting Proc. Int. Society for Magnetic Resonance in Medicine
[58] Bottou L 2010 Large-scale machine learning with stochastic gradient descent Proc. COMPSTAT pp 177–86
[59] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[60] Tang J, Golbabaee M and Davies M 2017 Gradient projection iterative sketch for large scale constrained least-squares Proc. Int. Conf. Machine Learning vol 70 pp 3377–86
[61] Ji S, Dunson D and Carin L 2008 Multitask compressive sensing IEEE Trans. Signal Process. 57 92–106
[62] Golbabaee M, Arberet S and Vandergeynst P 2013 Compressive source separation: theory and methods for hyperspectral imaging IEEE Trans. Image Process. 22 5096–110
[63] Golbabaee M, Arberet S and Vandergeynst P 2010 Distributed compressed sensing of hyperspectral images via blind source separation Proc. 44th Asilomar Conf. Signals, Systems and Computers (IEEE) pp 196–8
[64] Duarte-Carvajalino J, Lenglet C, Ugurbil K, Moeller S, Carin L and Sapiro G 2012 A framework for multi-task Bayesian compressive sensing of DW-MRI Proc. MICCAI Computational Diffusion MRI pp 1–13
[65] Belloni A, Chernozhukov V and Wang L 2011 Square-root Lasso: pivotal recovery of sparse signals via conic programming Biometrika 98 791–806
[66] Deledalle C-A, Vaiter S, Fadili J and Peyré G 2014 Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection SIAM J. Imaging Sci. 7 2448–87
[67] Poon C and Peyré G 2019 Degrees of freedom for off-the-grid sparse estimation (arXiv:1911.03577)
[68] Burdakov O and Merkulov B 2001 On a new norm for data fitting and optimization problems Technical Report LiTH-MAT (Linköping: Linköping University)
[69] Ndiaye E, Fercoq O, Gramfort A and Salmon J 2016 Gap safe screening rules for sparse-group Lasso Advances in Neural Information Processing Systems pp 388–96
[70] Ekeland I and Temam R 1999 Convex Analysis and Variational Problems (Philadelphia, PA: SIAM)