IDENTIFIABILITY CRITERIA IN SPARSE REGRESSION

MICHAEL EICKENBERG, SUPERVISOR: GABRIEL PEYRE

Abstract. The lasso functional is introduced in the context of solving noisy ill-posed linear inverse problems. Some properties of its solutions are exhibited. Signal dependent recovery and identifiability criteria are introduced which guarantee the correct identification of the sign (Fuchs identifiability criterion) and the support (Tropp exact recovery criterion) of the solution vector. A bound on l_2 error on the estimation of the solution and the prediction of measurements due to Grasmair is introduced and placed in context. An application of the Fuchs identifiability criterion to super-resolution in magnetoencephalographic (MEG) measurements is elaborated on an idealized example. This work is a review of these well-established criteria, placed in a uniform notation and setting. The last point is original and a preliminary tentative towards finding a bound on resolution in a realistic MEG setting. Everything presented should be accessible to a first year master student of mathematics.

Contents

1. Introduction

The presence of sparsity in a high-dimensional setting can often dramatically reduce the complexity of a problem and can be treated with dedicated algorithms yielding tractable performance, where a dense setting may be hopeless. Here we study inverse problems of the type

$$\text{Find or approximate } x^0 \text{ given } y = \Phi x^0 + w,$$

where $\Phi \in \mathbb{R}^{Q\times N}$ is a matrix and we know that $x^0 \in \mathbb{R}^N$ is sparse, i.e. that $|\operatorname{supp}(x^0)| << N$, where the support is the set of non-zero coordinates, and $w \in \mathbb{R}^Q$ is a noise vector and N is potentially very large. This inverse problem crops up in many different settings from signal processing to statistics and machine learning, but its employment bears a lot of similarities.

Solving it is not straightforward. Denote by $\|x\|_0^0 := |\operatorname{supp}(x)|$ the so-called l_0-*pseudonorm* of x. Several ways of enforcing sparsity in the solution of the linear inverse problem come to mind: One can move around in low-dimensional subspaces spanned by columns of Φ:

$$\operatorname{argmin}_x \|x\|_0^0 \text{ subject to } \Phi x = y.$$

In order to obtain an approximate sparse solution we can pose a constraint, i.e. solve for $k \in \mathbb{N}$

$$\operatorname{argmin}_x \|y - \Phi x\|_2^2 \text{ subject to } \|x\|_0^0 \leq k.$$

or we can penalize the use of a large support by solving for $\lambda > 0$

$$\operatorname{argmin}_x \|y - \Phi x\|_2^2 + \lambda \|x\|_0^0.$$

As it turns out, these problems are combinatorial in nature and extremely hard (NP) to solve exactly as soon as dimensionality increases [26]. The watershed between tractable and intractable optimization problems is very often the presence of convexity in the problem [2]. In certain cases a problem can be *convexified*, yielding a new problem that is easy to optimize, sometimes at the cost of possibly not finding the optimal solution for the initial problem. In our situation, a way to convexify the l_0 pseudonorm is to replace it by the "closest" convex norm, the l_1-norm (one can imagine this as a transition $\|x\|_p^p, 0 \xrightarrow{p} 1$). The sudden tractability of this problem with the advent of better methods for convex optimization has led to a very wide use of l_1 relaxed methods to obtain sparse solutions.

1.1. **Ubiquity of the linear inverse problem with sparsity assumption.** In signal processing sparsity is cardinal to representing a signal in a concise way as a linear combination of *atoms* where the set of atoms spans the signal space in a potentially highly overcomplete manner. Consider the following general setup: A signal vector $\bar{x} \in \mathbb{R}^P$ is represented as a linear combination of vectors well adapted to describe this type of signal $\bar{x} = \Psi x$, where $\Psi \in \mathbb{R}^{P \times N}$. Now allow some noisy, potentially incomplete measurements of the signal $\bar{x}$, by writing $y = \Xi \bar{x} + w = \Xi \Psi x + w = \Phi x + w$ for $\Xi \in \mathbb{R}^{Q \times P}$ and $\mathbb{R}^{Q \times N} \ni \Phi = \Xi \Psi$.

In a first setting consider a sparse signal $\bar{x} \in \mathbb{R}^N$ (a *spike train*) that is measured by a low-pass convolution Ξ. In this situation $\Psi = \mathrm{Id}$ (the signal is sparse in the canonical basis) and $\Phi = \Xi$, i.e. $\bar{x} = x$ and $\Xi x = h * x$ for some other discrete function h, e.g. a Gaussian bell shape or any other function with high Fourier energy in the low frequencies and potentially zero energy in the high frequencies. In many applications it is of essence to determine as well as possible the locations and amplitudes of the spikes in x given the measurements $y = \Xi x + w = h * x + w$. This problem is called *sparse spike deconvolution.* An l_1-regularization approach is documented in geophysics as early as 1981 [23] and 1986 [31]. Donoho uses it in 1990 [9]. Deconvolution is at the basis of superresolution, a topic which will be briefly introduced later, in order to explain our application.

Next we consider for a moment a complete, but potentially noisy measurement of the signal, i.e. $\Xi = \mathrm{Id}$ and $\Phi = \Psi$. In this case, given $y = \Phi x$ or $y = \Phi x + w$ we would like to find a "good" representation of y as a linear combination of columns of Φ. What is a "good" representation? If Φ is an orthogonal basis, such as for example one of the many wavelet bases in the literature (see the book by Stéphane Mallat [24] for an excellent overview and reference), then there is not much choice for the representation. A thresholding operation can be used to reduce noise [24]. However, some types of signals are not well represented in bases. For a given signal class $C \subset \mathbb{R}^Q$, this is to be understood in the sense that while the representation of typical signals $x \in C$ in a basis may need all the basis vectors for a full description, the same signals may be representable in a possibly only slightly larger linear system with very few non-zero coefficients. In order to elucidate this, consider the following extreme example. Let $N > 0$ and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T, 1 \leq i \leq N$ be the cartesian unit vectors and $f_j = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{\frac{2\pi i j k}{N}}$ the Fourier vectors. A vector e_i has full support in the basis of the f_j and vice versa. Now imagine a signal space in which a typical signal y consists of a superposition of few sinusoids f_j with some spikes e_i, such that $y = \sum_{i=1}^N a_i e_i + \sum_{j=1}^N b_j f_j$ with very few coefficients $a_i, b_j \neq 0$, e.g. $\bar{x} = e_1 + f_1$. This vector has full support in the cartesian basis and in the Fourier basis due to the mixing, but would be very concisely represented in a concatenation of the two bases. Indeed, Donoho and Huo studied the problem of representing signals in pairs of bases and found that under certain *incoherence* conditions on the two bases Φ_1, Φ_2, a signal y that has a sparse representation in the concatenation $\Phi = [\Phi_1, \Phi_2]$ can be recovered by l_1 methods. The result is refined by Elad and Bruckstein [13] and Feuer and Nemirovsky [15]. Later it is generalized to general matrices Φ by Donoho and Elad [8] and Gribonval and Nielsen [21]. They give a guarantee that if the solution is "very sparse" it corresponds to the solution of the corresponding l_0 problem.

In *compressed sensing* the emphasis is placed on the incompleteness of the measurement Ξ. For example, if Ψ is the identity and $x = \bar{x}$ is a sparse signal, then for Ξ drawn from a certain random matrix ensemble, such as the Gaussian ensemble or the ensemble of partial Fourier coefficients, one can reconstruct with very high probability the signal by solving a simple linear program. See the seminal works [5], [3], [35] for details. They make use of a *restricted isometry property* in the measurement matrix, which means that all subsets of a certain number of columns in the matrix must be close to orthogonal. For a given matrix the verification of this property is combinatorial, but certain random matrix ensembles can be shown to verify this property with high probability.

The common denominator of the inverse problems mentioned so far is that they can all be satisfactorily solved by finding

$$\operatorname{argmin}_x \|x\|_1 \text{ subject to } \Phi x = y,$$

in the noiseless case and with

$$\operatorname{argmin}_x \|x\|_1 \text{ subject to } \|\Phi x - y\|_2^2 \leq \varepsilon^2,$$

in the case of a noisy measurement with noise $w \in \mathbb{R}^Q$, $\|w\|_2 \leq \varepsilon$. The first optimization is called *Basis Pursuit* [6] and the second is called *Basis pursuit denoising.*

Typically in statistics the matrix $\Phi \in \mathbb{R}^{Q \times N}$ is a data matrix following an associated statistical data distribution and representing Q observed data vectors of dimension N. x^0 is then a weight vector suited to linearly explain an effect y with the data Φ: $y = \Phi x^0$. Often only few of the regressor columns of Φ explain the observation y - this is especially the case if one is in the process of searching a model for y and the data matrix contains potentially unrelated (uncorrelated or even statistically independent) columns. In 1996, Tibshirani introduced the LASSO (least absolute shrinkage and selection operator) [32] which shrinks many small coefficients in the weight vector to zero, leaving only coefficients large enough in absolute value, thus yielding a sparse vector. It is obtained by solving for a parameter $\mu > 0$ the convex optimization problem

$$\min_x \|y - \Phi x\|_2^2 \text{ subject to } \|x\|_1 \leq \mu.$$

In the context of statistics the size of the data matrix is permitted to vary according to the number of samples which generally come from a distribution. It may also be corrupted with noise. It is thus of interest to study stability results when the amount of data (lines) tends to infinity and potentially also the amount of features (columns). This results in consistency results ([41], [39] and others) and oracle inequalities ([42] and others). These results are outside the scope of this work.

In the context of machine learning, searching for and enforcing sparsity can largely be seen as a means to an end if we permit a machine learning algorithm to be defined as a mechanism that ideally makes a good *prediction* of an effect given new data. For a *training data* matrix Φ and a *training target* y, a linear predictor will attempt to find a weight vector $\hat{x}$ that will be used for predicting an unseen target given new data. For a new data matrix and observation (Φ', y') drawn from the same distribution as (Φ, y) a linear predictor generates a prediction $\hat{y}' = \Phi' \hat{x}$. Its performance is evaluated by comparing $\hat{y}'$ with y' and not by comparing a potential underlying sparse ground truth x^0, which reflects the exact relation between data and observation, with the estimate $\hat{x}$. However, given the sometimes very high dimensional data vectors that occur in a machine learning setting, enforcing sparsity is useful when prior knowledge exists that the weight vector $\hat{x}$ should be sparse. This happens in supervised learning where one tries to predict an effect from a big data matrix with only very few relevant features. Ng shows in [27] that for logistic regression l_1-regularization can asymptotically deal with exponentially many irrelevant features with respect to the number of relevant features and shows that for rotationally symmetric regularization such as l_2 can deal with only a linear relation between irrelevant and relevant features in the worst case.

To sum up this introduction, we note that the unconstrained Lagrangian formulations of the basis pursuit denoising and the LASSO problems are equivalent to solving for a certain $\lambda > 0$ the following problem

$$\operatorname{argmin}_x \|y - \Phi x\|_2^2 + \lambda \|x\|_1.$$

Studying the optimality conditions and recovery conditions of this functional is thus informative to all of the fields mentioned thus far.

For completeness let us mention here that the l_1-operator can be exchanged in different ways to obtain different forms of sparsity adapted to the signal. This work has presented the so-called *synthesis* setting. The corresponding *analysis* setting employs a dictionary $D \in \mathbb{R}^{N \times P}$ and the method penalizes using $\|D^* x\|_1$ instead of $\|x\|_1$. In this way one can for example penalize total variation of the signal [30]. In general the analysis setting is not equivalent to the synthesis setting, but for special matrices D (full rank, invertible) the models are equivalent [14, 25]. Also, variables can be grouped together using a so-called $l_1 - l_2$ norm $\sum_{g \in \mathcal{G}} \|x_g\|_2$, where the elements of $\mathcal{G}$ are sets that disjointly tile $\{1 \dots N\}$. This is useful if there exists prior knowledge that both variables should either be identically zero or active together [40].

1.2. **Recovery criteria.** In the introduction, two criteria guaranteeing recovery properties have already been mentioned. Donoho's *coherence* and the *restricted isometry property.* The latter concerns mostly random matrices and is out of the scope of this work. The former is a measure of coherence of the columns of the matrix. For a matrix Φ with unit norm columns it is defined by $\mu(\Phi) := \max_{i \neq j} |\langle \Phi_i, \Phi_j \rangle|$ where Φ_i is the i-th column of Φ. If a vector $x^0 \in \mathbb{R}^N$ satisfies $\|x^0\|_0^0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right)$ then x^0 is the solution of the basis pursuit [8]

$$\operatorname{argmin}_x \|x\|_1 \text{ subject to } \Phi x = \Phi x^0.$$

If we set $M(x^0) := \frac{\|x^0\|_0^0 \mu(\Phi)}{1 - (\|x^0\|_0^0 - 1)\mu(\Phi)}$ then Donoho's criterion is fulfilled if and only if $M(x^0) < 1$. In practise, as soon as two columns of the matrix Φ are rather correlated, this criterion has very little use. Even if signals of most supports are well recovered, the criterion cannot indicate this. From this we deduce an interest in signal dependent criteria, since the matrix Φ may perform very differently on signals of different support. In this work we present three of these criteria. The criterion of Tropp [34] called the *exact recovery principle* (ERC) depends on the support of the signal vector $I := \operatorname{supp}(x^0)$. If $ERC(I) < 1$ the support of a basis pursuit denoising is included in I for arbitrary noise levels. The identifiability criterion IC according to Fuchs depends on the *sign* of the entries of the signal x^0, making it even more specific and sharper. If $IC(\operatorname{sign}(x^0)) < 1$ then the sign of the solution is recovered by a basis pursuit denoising even if the measurement is perturbed by a small noise. We present a converse that makes sign recovery impossible if $IC(\operatorname{sign}(x^0)) > 1$ provided we have a sufficient bound on the noise. The last criterion in this line that we present has been dubbed $IC_0(\operatorname{sign}(x^0))$. If it is strictly less than 1 it guarantees a bound on the l_2 error $\|x^0 - \hat{x}\|_2$, where $\hat{x}$ is the solution of a basis pursuit denoising. The true support is not necessarily discovered.

In this document we will detail the definitions of these signal dependent criteria and give sketches of the proofs that underlie them. Since they rely heavily on the optimality conditions of the lasso functional, these conditions shall be given in form of an introductory section.

1.3. **Superresolution.** The concept of superresolution is important in a number of signal processing settings. In general it can be described as the attempt to extract signal information at a resolution that is finer than the original acquisition resolution. In imaging this can amount to extracting or changing sub-pixel features, usually given several images of a same scene or a video sequence [29], but it is also possible on a single image, see [17]. The aperture shape of an acquisition device such as a camera poses fundamental limits on precision as well as the pixel size in imaging. In microscopy and astronomy there are fundamental physical limits to acquisition precision such as the diffraction limit, where objects become impossible to resolve because they are to small with respect to the wavelength employed. However, sometimes it is possible to surpass this limit if some signal properties are known. Candès and Fernandez-Granda describe in a recent publication [4] how this may be formalized mathematically. If $x(t_1, t_2)$ is an object to be measured, the measurement amounts to a low pass filtering with a point spread function $h(t_1, t_2)$

$$y(t_1, t_2) = (h * x)(t_1, t_2),$$

where the transfer function $\hat{h}$ is of compact support. This means that the high frequencies of the object x are completely lost. According to Candès it is the job of superresolution to *extrapolate* the missing spectrum. This is in contrast to for example compressed sensing using random Fourier coefficients, where the missing part of the spectrum must be *interpolated.* The case of spike deconvolution mentioned earlier is included in this context. Candès and Granda prove a surprising theorem in a continuous setting which is stated here informally as a reference for motivating the next section.

Let $\mathbb{T}$ be the $[0, 1]$-circle with $0 \sim 1$. For points $t_i \in \mathbb{T}$ and $a_i \in \mathbb{C}$ define a signed atomic measure $x = \sum_i a_i \delta_{t_i}$. Define $\mathcal{F}_n x(k) := \int_0^1 e^{-2\pi i k t} x(dt) = \sum_i a_i e^{-2\pi i k t_i}, k \in \mathbb{Z}, |k| \leq f_c, n = 2f_c + 1$. These are the the n Fourier coefficients under the cutoff frequency f_c. Further define for $T \subset \mathbb{T}$ a collection of points, the *minimum separation* $\Delta(T) := \inf_{t \neq t' \in T} |t - t'|$, where the distance is seen as the shortest on the circle.

Their Theorem goes as follows. Denote by $\| \cdot \|_{TV}$ the total variation of a signed measure. Let $f_c \geq 128$, $n = 2f_c + 1$ and $y = \mathcal{F}_n x$. If $T = \{t_i\}$ is the support of x and $\Delta(T) \geq 2/f_c$, then

$$\operatorname{argmin}_{\hat{x}} \|\hat{x}\|_{TV} \text{ subject to } \mathcal{F}_n \hat{x} = y$$

recovers x *exactly*. In other words, we can infer the exact, continuous positions and amplitudes of any spike train that is minimally separated from a finite number of Fourier coefficients.

The idea of this type of superresolution is enticing for the recovery of MEG currents. It will be discussed as possible future work in the Outlook. Next we explain magnetoencephalography.

1.4. **Magnetoencephalography.** *Magnetoencephalography* or MEG is a non-invasive measurement technique for electrical currents in the brain through the magnetic field they induce in their surrounding. These currents and magnetic fields are very small and are thus captured using superconducting quantum interference devices (SQUIDs) placed around the outside of the head [7], making noise treatment a very important aspect of the acquisition.

In the magnetostatic regime the Maxwell equations dictate that the magnetic field quantities measured in the detectors are linear functions of the current density present in the brain. For discretized source locations and an activation vector $x^0 \in \mathbb{R}^N$ we can thus write a linear forward model $B = \Phi x^0 + w$ where B is the measured magnetic field vector. The design Φ is also called the *gain matrix*. We present an example of how this matrix can be obtained from the simplified case of sources and detectors in vacuum by using the Biot-Savart formula

$$\vec{B}(\vec{R}) := \frac{\mu_0}{4\pi} \int_{\mathbb{R}^3} \vec{\jmath} \times \frac{\vec{R} - \vec{r}}{\|\vec{R} - \vec{r}\|_2^3} d\vec{r}.$$

The general methods used to retrieve source positions from the measurements B are *dipole fitting* and minimum norm estimates of currents. Dipole fitting is a non-linear procedure attempting to fit a forward model by optimizing the location and orientation/intensity of a previously fixed number of current dipoles in the brain. Minimum norm estimate methods use a lasso-type problem to solve a regularized inverse problem under the assumption that the total sum of current flow is minimal [22]. Norms more elaborate than l_1 are used in order to satisfy prior knowledge that a source may stay active over a longer period [18]. It is even possible to create norms that favor transient signals in localized time-frequency bands [19].

In this section we analyze a stylized MEG setting where sources are placed on the z-axis in 3D space, all pointing in one direction along this axis. The MEG magnetic field is measured on a cylinder which is concentric around the sources, or, equivalently, on a line parallel to the z-axis. We analyze the behaviour of IC in a two sources setting and study the regions where it is violated by evaluating $\Phi_J^* \Phi_I^{*,+} \operatorname{sign}(x^0)$ in a continuous setting, the limit of infinite measurements on the cylinder. We suppose that taking this limit can only add information and numerically we find this to be true. Yet even in the limit case we find that IC can *never* be less than 1 if the discretization on the source grid is chosen too finely with respect to the distance $R > 0$ at which we are measuring. In order to study this phenomenon closely, we establish a continuous function $\eta : \mathbb{R} \to \mathbb{R}$ such that for a given source discretization grid $(z_i)_i$, IC evaluates to $\max_i |\eta(z_i)|$. Using this function, we can establish a quantity dependent on the distance of the detectors $R > 0$ (which we shall dub *imaging scale*) and the distance between the two sources $d > 0$ (the *minimum source separation*), which we call the *minimum sampling distance* $\Delta(d, R) \geq 0$. If we discretize at a step greater or equal to $\Delta(d, R)$, IC holds in the two-source setting. We observe numerically that $\Delta(d, R) > 0$ for all $d, R > 0$ and can value $+\infty$ if d is too small with respect to R, leading to instability in support recovery for any discretization grid.

2. Some properties of the lasso functional

2.1. **Notation and preliminaries.** In the following we will note $x \in \mathbb{R}^N$ an N-dimensional real vector, for a real number a its absolute value is denoted by $|a| := \max\{a, -a\}$. The *sign* of a real number $a \in \mathbb{R}$ is defined by

$$\operatorname{sign}(a) = \begin{cases} \frac{a}{|a|} & \text{if } a \neq 0 \\ 0 & \text{if } a = 0 \end{cases} = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \\ 0 & \text{if } a = 0 \end{cases},$$

and the sign of a vector $x \in \mathbb{R}^N$ is defined by $(\operatorname{sign}(x))_i := \operatorname{sign}(x_i)$. The *support* of a vector is defined by $\operatorname{supp}(x) := \{i | x_i \neq 0\} \subset \{1 \dots N\}$. For a set $I \subset \{1 \dots N\}$ we denote I^C its complementary in $\{1 \dots N\}$. The notation Φ_I for a matrix Φ is shorthand for the matrix consisting of the columns of Φ indexed in I (in ascending order). For $i \in \{1 \dots N\}$, Φ_i represents the ith column of the matrix Φ. Let $V_I := \operatorname{span} \Phi_I$. The l_1-norm of a vector x is defined by $\|x\|_1 := \sum_{i=1}^N |x_i|$, the l_2-norm is noted $\|x\|_2 := \sqrt{\sum_{i=1}^N |x_i|^2}$ and the maximum or l_∞-norm is $\|x\|_\infty := \max_i |x_i|$. For a matrix $\Phi \in \mathbb{R}^{Q \times N}$ the adjoint matrix is denoted

$\Phi^* \in \mathbb{R}^{N\times Q}$. The Moore-Penrose pseudoinverse is written Φ^+ and if $\Phi^*\Phi$ is invertible it can be written $\Phi^+ = (\Phi^*\Phi)^{-1}\Phi^*$, or if $\Phi\Phi^*$ is invertible we have $\Phi^+ = (\Phi^*(\Phi\Phi^*)^{-1})^*$. For a convex function $f : \mathbb{R}^N \to \mathbb{R}$ and $x \in \mathrm{dom} f$, the subgradient is defined by $\partial f(x) := \{\alpha | f(y) \geq f(x) + \alpha^*(y - x) \text{ for all } y \in \mathbb{R}^N\}$. Let Π_C represent the projection operator on a convex $C \subset \mathbb{R}^M$.

2.2. Properties of the lasso functional.

Definition 2.1 (The lasso functional). Let $\Phi \in \mathbb{R}^{Q\times N}$, $y \in \mathbb{R}^Q$, $\lambda > 0$ and $x \in \mathbb{R}^N$. Then the lasso functional is defined as

$$L_{\lambda,y}(x) := \frac{1}{2}\|\Phi x - y\|_2^2 + \lambda\|x\|_1.$$

The problem of finding its minimum in x is called the *lasso problem* and shall be denoted $P_\lambda(y)$.

Lemma 2.2 (Optimality condition). *Any minimizer $\hat{x} \in \mathbb{R}^N$ of $L_{\lambda,y}$ must verify $0 \in \partial L_{\lambda,y}(\hat{x})$, where $\partial L_{\lambda,y}(\hat{x})$ is the subderivative of $L_{\lambda,y}$ in $\hat{x}$. This is a necessary and sufficient condition. It amounts to*

$$\Phi^*(y - \Phi\hat{x}) = \lambda\gamma,$$

where $\gamma \in \partial\|\cdot\|_1(\hat{x})$. Denote by I the support of $\hat{x}$. Then, noting that $\Phi\hat{x} = \Phi_I\hat{x}_I$, this condition can be expressed as

$$\begin{array}{rcl} \Phi_I^*(y - \Phi_I\hat{x}_I) & = & \lambda\,\mathrm{sign}(\hat{x}_I), \\ \Phi_{I^C}^*(y - \Phi_I\hat{x}_I) & \in & [-\lambda, \lambda]. \end{array}$$

Proposition 2.3 (Existence and uniqueness of minimizers). *Let $\Phi \in \mathbb{R}^{Q\times N}$, $y \in \mathbb{R}^Q$ and $\lambda > 0$. Then we have the following results:*

a) the set of solutions to $P_\lambda(y)$ is non-empty, convex and compact.
b) Let $\hat{x}$ be a solution to $P_\lambda(y)$ and $K := \{k|\Phi_k^*(y - \Phi\hat{x}) = \pm\lambda\}$. Then, if Φ_K is injective, the solution is unique and can be written as

$$\hat{x} = \Phi_K^+ y - \lambda(\Phi_K^*\Phi_K)^{-1}\,\mathrm{sign}\,(\hat{x}_K).$$

This result can be found in [16, 33, 28] and many others.

Remark 2.4. The solution to the lasso problem is not always unique. Let

$$\Phi := \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \end{pmatrix}, \quad y := \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Then for $0 < \lambda < 1$, $s_0^\lambda = (1 - \lambda)(0, 0, 2)^T$ and $s_1^\lambda = (1 - \lambda)(1, 1, 0)^T$ are minimizers of the lasso functional. This motivates the following proposition:

Proposition 2.5 (Characterization of minimizers). *Let $\Phi \in \mathbb{R}^{Q\times N}$, $y \in \mathbb{R}^Q$ and $\lambda > 0$. Further let $\hat{x} \in \mathrm{argmin}_x L_{\lambda,y}(x)$, $K := K(\hat{x}) := \{k|\Phi_k^*(y - \Phi\hat{x}) = \pm\lambda\}$ and $\gamma := \gamma(\hat{x}) := \frac{1}{\lambda}\Phi^*(y - \Phi\hat{x})$. For any (possibly other) minimizer $\tilde{x} \in \mathrm{argmin}\, L_{\lambda,y}$ we have the following results*

(1) There exists $b \in \ker\Phi$ such that $\tilde{x} = \hat{x} + b$.
(2) We have $\|\tilde{x}\|_1 = \|\hat{x}\|_1$, $\gamma(\tilde{x}) = \gamma(\hat{x})$ and $K(\tilde{x}) = K(\hat{x})$.
(3) For $b \in \ker\Phi$, $\tilde{x} := \hat{x}+b$ is a solution to $P_\lambda(y)$ if and only if $b_{K^C} = 0$ and for all $k \in K$, $\gamma_k(\hat{x}_k+b_k) \geq 0$
(4) There exists a minimizer $\tilde{x}$ of $L_{\lambda,y}$ with support $I := \mathrm{supp}\,\tilde{x}$ such that Φ_I is injective. This immediately entails the existence of a solution such that $|I| \leq \min(Q, N)$.

This is a collection of results mentioned in the following literature [16, 33, 28], among many others.

The solution $\hat{x}$, if unique, can be studied as a function of y, λ. Except for a certain transition space, essentially characterizable by $K(\hat{x}) \neq \mathrm{supp}(\hat{x})$ and an injectivity property, the solution $\hat{x}$ is locally affine in y, λ.

Definition 2.6 (Transition space). Let $I \subset \{1 \dots N\}$ such that Φ_I injective. Define for $j \notin I$

$$\mathcal{H}_{I,j} := \big\{(y, \lambda) \in \mathbb{R}^Q \times \mathbb{R}_+^* \,\big|\exists x_I \in \mathbb{R}^I \text{ such that } \Phi_I^*(y - \Phi_I x_I) = \lambda\,\mathrm{sign}(x_I)\ ,$$
$$\text{and } \Phi_j^*(y - \Phi_I x_I) = \pm\lambda\big\}\,.$$

Define the *transition space* as

$$\mathcal{H} := \bigcup_{\substack{I\in\{1\dots N\}\\ \Phi_I \text{ injective}}} \ \bigcup_{\substack{j\notin I\\ \Phi_j\notin\operatorname{span}(\Phi_I)}} \mathcal{H}_{I,j}.$$

Theorem 2.7 (Variation of the solution with y and λ)**.** *For $(y,\lambda)\notin\mathcal{H}$, let $\hat{x}$ be a minimizer of $L_{\lambda,y}$ and $I := \operatorname{supp}(\hat{x})$ its support. We suppose Φ_I injective. Then there exists a neighbourhood $\mathcal{U}$ of (y,λ) and a function $\varphi : \mathcal{U}\to\mathbb{R}^N$ such that for all $(\bar{y},\bar{\lambda})\in\mathcal{U}$, $(\bar{y},\bar{\lambda})$ is a minimizer of $L_{\lambda,y}$ with support I and we have $\varphi(y,\lambda)=\hat{x}$. We have*

$$\partial_{\bar{y}}\varphi(\bar{y},\bar{\lambda})_I = \Phi_I^+ \qquad \partial_{\bar{\lambda}}\varphi(\bar{y},\bar{\lambda})_I = -(\Phi_I^*\Phi_I)^{-1}\operatorname{sign}(\hat{x}_I),$$

and

$$\partial_{\bar{y}}\varphi(\bar{y},\bar{\lambda})_{I^C} = \partial_{\bar{\lambda}}\varphi(\bar{y},\bar{\lambda})_{I^C} = 0.$$

This theorem can be found in [11]. Another proof can be formulated as the adaptation of the same statement for the group lasso by Vaiter et al. in [37]. The proof uses the implicit function theorem. Outside the transition space the sign function (and the corresponding generalized sign function for groups) are smooth with Jacobi determinant non zero. Other proofs have been put forward in Osborne et al. [28] and [36] and elsewhere.

Remark 2.8 (Homotopy Algorithm). The piecewise linearity of the lasso solution as a function of λ can be exploited to create efficient solver algorithms that have been dubbed *homotopy method* [28, 10] and *LARS lasso* (from *least angle regression*, see [12]). The algorithm keeps track of the signal support and iteratively finds the transition points where a new variable is added to the support or leaves the support. Let $\Phi\in\mathbb{R}^{Q\times N}$ and $y\in\mathbb{R}^Q$. The algorithm starts with the zero solution and its corresponding maximal penalty $\lambda_{\max} = \|\Phi^* y\|_\infty$. The first index $1\leq i\leq N$ entering the support is $i = \operatorname{argmax}_j |\Phi_j^* y|$. Using the given support I one determines the locally constant $\gamma_I(\lambda) = \frac{1}{\lambda}(\Phi_I^*(y-\Phi_I\hat{x}_I)) = \operatorname{sign}(\hat{x}_I)$. The next transition point is linearly extrapolated using the local parametrization of the solution $\hat{x}_I(\lambda) = \Phi_I^+ y - \lambda(\Phi_I^*\Phi_I)^+\operatorname{sign}(\gamma_I)$ by determining for which λ a coefficient i in is equal to zero $\hat{x}_i(\lambda)=0$ or a variable $j\in I^C$ enters the support by verifying $\Phi_j^*(y-\Phi_I\hat{x}_I(\lambda)) = \pm\lambda$.

3. Identifiability criteria for sparse solutions of underdetermined linear systems

In the following we will expose several criteria permitting recovery of sparse vectors transformed by underdetermined linear measurements corrupted with noise.

3.1. The Fuchs identifiability criterion (IC).

The identifiability criterion due to Fuchs [16] is a criterion based on the matrix Φ and sign of the signal to be recovered. It is thus identical for two signals having the same sign. The criterion is derived directly from the optimality conditions (2.2) and the proof uses the criterion in order to verify the optimality conditions. If $IC<1$ and the noise vector $w\in\mathbb{R}^Q$ is sufficiently bounded in the l_2 sense.

Definition 3.1 (The identifiability criterion (IC) due to Fuchs [16])**.** Let $\Phi\in\mathbb{R}^{Q\times N}$, $x^0\in\mathbb{R}^N$ a signal with support $I := \operatorname{supp}(x^0)$ and set $d_0 := \Phi_I^{*,+}\operatorname{sign}(x^0)$. Define

$$\operatorname{IC}(\operatorname{sign}(x^0)) := \max_{j\in I^C}|\langle\Phi_j, d_0\rangle| = \|\Phi_{I^C}^* d_0\|_\infty = \|\Phi_{I^C}^*\Phi_I^{*,+}\operatorname{sign}(x^0)\|_\infty$$

Proposition 3.2 (Sign recovery using IC)**.** *Suppose that* $\operatorname{IC}(\operatorname{sign}(x^0))<1$ *and that Φ_I has full rank. Then there exist two constants $c_I, \tilde{c}_I$ such that if*

$$\frac{\|w\|_2}{\min_{i\in I}|x_i^0|} < \frac{\tilde{c}_I}{c_I},$$

and if

$$c_I\|w\|_2 < \lambda < \tilde{c}_I\min_{i\in I}|x_i|,$$

then the unique solution to the lasso problem verifies

$$\hat{x}_I = x_I^0 + \Phi_I^+ w - \lambda(\Phi_I^*\Phi_I)^{-1}\operatorname{sign}(x_I^0),$$

and we have equality of the signs

$$\operatorname{sign}(\hat{x}) = \operatorname{sign}(x^0).$$

This is a formulation adapted from [38] to the sparse synthesis case. In words: if the smallest non-zero entry of x^0 is large enough compared to the noise, then solving $P_\lambda(y)$ for a well chosen $\lambda > 0$ yields a solution $\hat{x}$ that has the same sign as x^0.

Sketch of proof. We "guess" the solution $\hat{x}_I := x^0 + \Phi_I^+ w - \lambda(\Phi_I^* \Phi_I)^{-1} \text{sign}(x_I^0)$ and $\hat{x}_{I^C} = 0$. This is inspired by the optimality condition. One begins by finding a criterion for exact sign recovery. It is based on the fact that $a, b \in \mathbb{R}$ have the same sign if $|a| > |a - b|$. Proving $\|\hat{x} - x^0\|_\infty < \min_{i \in I} |x_i^0|$ is thus a sufficient condition for sign recovery and is done using operator inequalities. Given the equality of signs we can check the optimality condition on $\hat{x}$. If we manage to strictly bound $\|\Phi_J(y - \Phi\hat{x})\|_\infty$ by λ we have shown that $\hat{x}$ is the unique solution to the lasso problem. This is done using the fact that $IC < 1$. The two preceding conditions constitute an upper and a lower bound for λ respectively. A third step must be undertaken to show that there exists an interval of possible λ that verifies both bounds.

The first condition furnishes an l_2 bound of the order $O(\|w\|_2)$ on the estimation error if λ is chosen proportional to $\|w\|_2$. □

There exists a converse statement to (3.2), making this criterion in some sense tight.

Proposition 3.3 (Impossibility to recover sign if IC > 1)**.** *Let* $\Phi \in \mathbb{R}^{Q \times N}$ *and* $x^0 \in \mathbb{R}^N$ *of support* $I := \text{supp}(x^0)$*, with* $\text{IC}(\text{sign}(x^0)) > 1$*. Let* $w \in \mathbb{R}^Q$ *a noise vector and* $y := \Phi x^0 + w$*. If* $\frac{1}{\lambda}\|\Phi_{I^C}^* \Pi_{V_I^\perp} w\|_\infty < \text{IC}(\text{sign}(x^0)) - 1$ *then for no choice of* $\lambda > 0$ *the solution* $\hat{x}$ *of* $P_\lambda(y)$ *verifies* $\text{sign}(\hat{x}) = \text{sign}(x^0)$*.*

Sketch of proof. The proof is based on the following contradiction. Suppose we have sign equality. This entails support equality. Let $I := \text{supp}(\hat{x}) = \text{supp}(x^0)$. Then by the optimality conditions on $\hat{x}$ we must have $\lambda \geq \|\Phi_{I^C}^*(\Phi_I x_I^0 + w - \Phi_I \hat{x}_I)\|_\infty = \|\Phi_{I^C}^*((\text{Id} - \Phi_I \Phi_I^+)w + \lambda \Phi_I^+ \text{sign}(x^0))\|_\infty \geq \lambda\, \text{IC}(\text{sign}(x^0)) - \|\Phi_{I^C}^* \Pi_{V_I^\perp}(w)\|_\infty$ □

3.2. **The exact recovery principle (ERC) of Tropp.** The next criterion we present is independent of signal sign, but dependent on signal support. It is called *exact recovery criterion* and is due to Tropp [34]. Its main idea is that if the criterion is verified, the support of the solution to $P_\lambda(y)$ will be included in the support of x^0 for a sufficiently large $\lambda > 0$. Note that the recovered support may be strictly included in the original one. This can include the zero solution and can be restricted to only the zero solution if the noise level $\|w\|_2$ is too high.

Definition 3.4 (The exact recovery criterion according to Tropp)**.** Let $\Phi \in \mathbb{R}^{Q \times N}$ and $x^0 \in \mathbb{R}^N$ a sparse signal with $I := \text{supp}(x^0)$. Define

$$\text{ERC}(I) := \|\Phi_{I^C}^* \Phi_I^{*,+}\|_{\infty,\infty}.$$

Proposition 3.5 (Support recovery using ERC)**.** *Let* $\Phi \in \mathbb{R}^{Q \times N}$*,* $x^0 \in \mathbb{R}^N$ *a sparse signal with* $I := \text{supp}(x^0)$ *and* $w \in \mathbb{R}^Q$ *a noise vector. Then, if* Φ_I *has full rank and if* $\lambda > \frac{\|\Phi_{I^C}^* w\|_\infty}{1 - ERC(I)}$*, we have* $\text{supp}(\hat{x}) \subset I$*.*

Sketch of proof. One projects the original lasso problem onto the image of Φ_I by multiplying with $\Phi_I \Phi_I^+$. The solution is extended by zeros to N dimensions. Then the optimality conditions are checked outside I using ERC, inside I we distinguish the support of the projected solution from the cosupport, i.e. as mentioned above, the support of the solution $\hat{x}$ may be strictly included in the support of the original signal x^0. On both we check the optimality conditions. By injectivity of Φ_I the solution is unique. □

Proposition 3.6 (Error bound using ERC)**.** *Suppose that ERC*$(I) < 1$*. Then the following error bound holds:*

$$\|x_I^0 - \bar{x}_I\|_2 \leq \|(\Phi_I^* \Phi_I)^{-1}\|_{2,2} \left(\sqrt{|I|}\lambda + \|\Phi_I^*\|_{2,2}\|w\|_2\right),$$

3.3. **An error bound due to Grasmair.** The following criterion differs a lot in the assertion that it makes and the approach it takes to prove it. It can be seen as a proof that if $IC < 1$ holds, one not only has sign robustness to small noise, but also an l_2 bound on the error of the estimate for arbitrary noise (i.e. $\|x^0 - \hat{x}\|_2$ is bounded for arbitrary noise). However, the result that is stated is actually more general since the criterion does not directly rely on IC. Instead it works with arbitrary *dual certificates* η, provided that they verify the two conditions $\eta \in \text{Im}\, \Phi_{I^C}^*$ and $\eta \in \partial\| \cdot \|_1(x^0)$. The specific certificate for IC, $\eta := \Phi_{I^C} \Phi_I^{*,+} \text{sign}\, x^0$ verifies both these properties, but may not be the only one. There may be others smaller in norm, yielding a weaker criterion. The following proposition is adapted from the work of Grasmair et al. [20]

Proposition 3.7 (l_2 error bound due to Grasmair [20])**.** *Let $x^0 \in \mathbb{R}^N$ with support $I := \operatorname{supp}(x^0)$. Suppose that there exists $\eta \in \operatorname{Im}\Phi_J^* \cap \partial\|\cdot\|_1(x^0)$ with $\|\eta\|_\infty < 1$. Further let $d(x^0) \in {\Phi_{I^C}^*}^{-1}(\{\eta\})$, suppose that Φ_I has full rank and that $\|w\|_2 \leq \varepsilon$. Let $\alpha := \frac{\lambda}{\varepsilon}$. Then for $\hat{x}$ lasso solution we have*

$$\|\Phi\hat{x} - y\|_2 \leq c_1\varepsilon,$$

and

$$\|\hat{x} - x^0\|_2 \leq c_2\varepsilon,$$

with

$$\begin{aligned} c_1 &= 1 + \alpha\|d(x^0)\|_2, \\ c_2 &= \|\Phi_I^+\|_{2,2}(1 + c_1) + \frac{1 + \|\Phi_I^+\|_{2,2}\|\Phi_{I^C}\|_{2,2}}{\alpha(1 - \|\eta\|_\infty)}\left(1 + \alpha\frac{\|d(x^0)\|_2}{2}\right)^2. \end{aligned}$$

Remark 3.8. If $\operatorname{IC}(\operatorname{sign}(x^0)) < 1$ holds, we can apply (3.7) with $\eta := \Phi_J^*\Phi_I^{*,+}\operatorname{sign}(x^0)$.

We give a definition of the smallest possible certificate in maximum norm that can be used for (3.7)

Definition 3.9 (IC_0)**.** Let $\Phi \in \mathbb{R}^{Q\times N}$ and $x^0 \in \mathbb{R}^N$ a signal with support $I := \operatorname{supp} x^0$. Let $d_0 := \Phi_I^{*,+}\operatorname{sign}(x^0)$ We define

$$\mathrm{IC}_0(\operatorname{sign}(x^0)) = \min_{u \in \ker\Phi_I^*}\|d_0 + u\|_\infty$$

Corollary 3.10 (Error bound with IC_0)**.** *Let $\Phi \in \mathbb{R}^{Q\times N}$ and $x^0 \in \mathbb{R}^N$ with support $I := \operatorname{supp}(x^0)$. Then if $\mathrm{IC}_0(\operatorname{sign}(x^0)) < 1$ we have the l_2 error bounds of (3.7).*

4. Brain imaging with magnetoencephalography : A super-resolution problem

4.1. **Obtaining the design matrix.** For this subsection we will briefly resort to typical notation in physics in order to derive the design we will be working on.

Derived directly from the Maxwell equations, the Biot-Savart formula returns the magnetic field at a given location $\vec{R} \in \mathbb{R}^3$ due to a current density $\vec{\jmath}\colon \mathbb{R}^3 \to \mathbb{R}^3$

Assuming a finite number of strongly localized point sources, they may be modeled by a distribution of Diracs. Let $\vec{r}_n \in \mathbb{R}^3, 1 \leq n \leq N$ be the locations of these point sources and $\vec{\jmath}_n \in \mathbb{R}^3$ the finite integral point densities located at $\vec{r}_n$. The total current density can then be written as $\vec{\jmath}(\vec{r}) = \sum_{n=1}^N \vec{\jmath}_n\delta_{\vec{r}_n}(\vec{r})$. In MEG, a number of $Q > 0$ detectors, localized at $\vec{R}_q \in \mathbb{R}^3$ can measure the magnetic field in a fixed direction $\vec{\xi}_q \in \mathbb{S}^2$. The currents that are measured are assumed to be originating from the surface of the brain, the *cerebral cortex*. This surface has a wavy structure with so-called *gyri* (ridges) and *sulci* (troughs) and can be modeled by a $2D$ mesh grid $\vec{r}_n \in \mathbb{R}^3$. For a given mesh grid, point sources $\vec{\jmath}_n \in \mathbb{R}^3$ are assumed to be placed on its vertices $\vec{r}_n \in \mathbb{R}^3$. Since most of the measured current flows normal to this surface, it is possible to impose the current direction on the model, by fixing $\vec{\nu}_n \in \mathbb{S}^2$ and letting $\vec{\jmath}_n := j_n\vec{\nu}_n$.

Definition 4.1 (Biot-Savart with fixed current and detector orientations)**.** Let $N > 0$ and $\vec{r}_n, 1 \leq n \leq N$ be a mesh grid of possible source locations and $\vec{\nu}_n \in \mathbb{R}^3$ their fixed current orientations. Let $Q > 0$ and $\vec{R}_q, 1 \leq q \leq Q$ be the detector positions and $\vec{\xi}_n \in \mathbb{R}^3$ its measuring orientations. Then, for a number of point currents $\vec{\jmath}_n = j_n\vec{\nu}_n$, the magnetic field measured by detector q reads

$$B_q = \sum_{n=1}^N j_n\left\langle\frac{\vec{R}_q - \vec{r}_n}{\|\vec{R}_q - \vec{r}_n\|_2^3}, \vec{\xi}_q \times \vec{\nu}_n\right\rangle.$$

Let $\Phi_{q,n} := \left\langle\frac{\vec{R}_q - \vec{r}_n}{\|\vec{R}_q - \vec{r}_n\|_2^3}, \vec{\xi}_q \times \vec{\nu}_n\right\rangle$. Then we can write

$$B = \Phi j.$$

Remark 4.2. In SI units there is a factor $\frac{\mu_0}{4\pi}$ in front of the integral, which we will ignore for the sake of notational simplicity.

4.2. **Continuous measurements for IC.** We propose a first test for superresolution capacities of the MEG measurements by studying the signal sign recovery properties in the case that there are only two locations with non-zero current by using IC.

An important quantity is the

Definition 4.3 (Covariance of measurements due to the sources)**.** Let $\Phi \in \mathbb{R}^{Q\times N}$ be the design matrix. Then let

$$C := \Phi^*\Phi \in \mathbb{R}^{N\times N}.$$

Further let $K, L \subset \{1 \dots N\}$. Then denote by $C_{K,L}$ the matrix $(C_{kl})_{(k,l)\in K\times L}$ (indices ordered as in $\mathbb{N}$)

Lemma 4.4 (IC from covariance matrix)**.** *Let $\Phi \in \mathbb{R}^{Q\times N}$ a design and $x^0 \in \mathbb{R}^N$ a signal with support $I := \operatorname{supp}(x^0)$ and suppose Φ_I injective. Let $C := \Phi^*\Phi$ the covariance matrix. We can then write*

$$\mathrm{IC}(\operatorname{sign}(x^0)) = \|\eta_0\|_\infty,$$

with

$$\eta_0 = C_{I^C,I}C_{I,I}^{-1}\operatorname{sign}(x^0).$$

For a finite source support $|I| < \infty$, this notation provides the possibility to pass to an infinite number of measurements since $C_{ij} = \langle \Phi_i, \Phi_j\rangle$ is nothing but a scalar product and can be replaced by a continuous counterpart if available. We thus propose to study the situation where the magnetic field is known everywhere on a 2D surface, e.g. an entire sphere engulfing the head, containing all the sources.

Definition 4.5 (Scalar product)**.** Let $S \subset \mathbb{R}^3$ be a smooth 2D surface, parametrized by $\Omega \subset \mathbb{R}^2$ using a smooth mapping $\vec{R} : \Omega \to S$. For two fixed sources, characterized by locations $\vec{r}_i, \vec{r}_j \in \mathbb{R}^3$ and $\vec{j}_i, \vec{j}_j$, and $\vec{R} \in \mathbb{R}^3$ write the magnetic field according to Biot-Savart as $\vec{B}(\vec{R}, \vec{r}, \vec{j}) = \vec{j} \times \frac{\vec{R}-\vec{r}}{\|\vec{R}-\vec{r}\|_2^3}$. Then let

$$\bar{C}_{ij} := \int_\Omega \left|\frac{\partial \vec{R}(x)}{\partial x}\right| dx \langle \vec{B}(\vec{R}(x), \vec{r}_i, \vec{j}_i), \vec{B}(\vec{R}, \vec{r}_j, \vec{j}_j)\rangle$$

where $\left|\frac{\partial \vec{R}(x)}{\partial x}\right| = \|\frac{\partial \vec{R}(x)}{\partial x_1} \times \frac{\partial \vec{R}}{\partial x_2}\|_2$ is the area of the surface element. This is the continuous approximation to the covariance matrix C_{ij} by assuming an infinite number of measurement points on a certain surface around the sources.

4.3. **Sources on 1D line, measurements on concentric cylinder.** We now propose to study the extremely simplified setting where sources are placed on a line and oriented in identical directions, parallel to the line. The measurements of the magnetic field are taken on a concentric cylinder of radius $R > 0$. To fix ideas, imagine the sources placed on the z-axis, at $\vec{r}_i = (0, 0, z_i)^T$. The continuous measurement locations are parametrized by $[0, 2\pi) \times \mathbb{R} \ni (\phi, h) \mapsto (R\cos\phi, R\sin\phi, h)$.

Remark 4.6. We propose to study this setting for its high symmetry. In fact, in this case the magnetic field is always tangential to the cylinder and will be of equal magnitude at a given height at all points on the circumference. Measuring on a cylinder is thus equivalent to measuring on a line that is contained in the cylinder which is parallel to the z-axis.

This leads to $\bar{C}_{ij}$ for the cylinder only depending on the distance $d = z_i - z_j$ of the sources, making the magnetic field measurement is a convolution of the source signal. Since it only depends on the absolute value of the distance between sources, the convolution kernel is symmetric. Our problem can thus be seen in a classic noisy deconvolution setting.

However, note that this function is not 1-homogeneous in R. Although R in some sense describes the *imaging scale*, doubling the distance between Diracs will not necessarily be enough for deconvolution at $2R$.

Proposition 4.7 ($\bar{C}_{ij}$ for line and cylinder)**.** *Let $z_i, z_j \in \mathbb{R}$ be two locations of sources on the z-axis and $R > 0$ the radius of the measurement cylinder. Then we have*

$$C_{i,j}^R = c(z_i - z_j),$$

where

$$c^R(d) = \frac{16}{3\pi}R^5\int_{\mathbb{R}} \frac{1}{\sqrt{\left(w^2 + R^2 - \left(\frac{d}{2}\right)^2\right)^2 + R^2d^2}^{\,3}}dw,$$

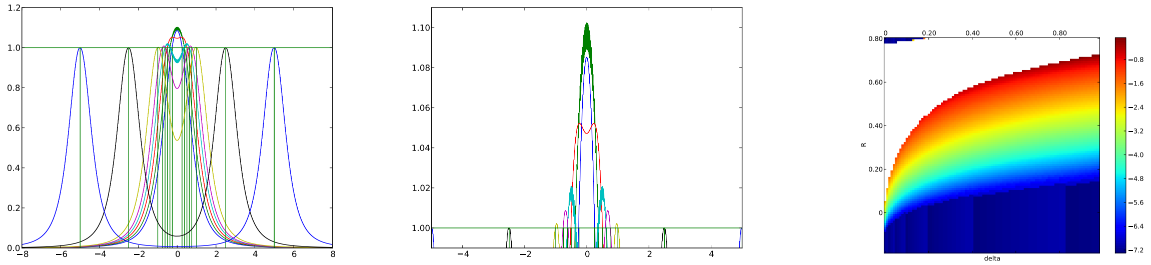

FIGURE 1. $R = 0.5$: distance values of [.5, .75, 1., 1.25, 1.5, 2., 5., 10.]. Left: η as a function of position for different distances between diracs. Middle: zoom in on the peaks at the origin. The smeared parts are due to numerical inexactness due to the sampling method. Right: log plot of discretization step Δ necessary to guarantee $\eta < 1$

which is the scalar product of the normalized magnetic fields measured on S, induced by sources in z_i, z_j.

Lemma 4.8 (IC for two Diracs)**.** *Let x^0 be a signal of support $I := \operatorname{supp}(x^0)$ and support cardinality $|I| = 2$. Let $I = \{i_1, i_2\}$ and $s_I := \operatorname{sign}(x^0_I)$. Then for $s_I = (+1, +1)^T$ we have $\operatorname{IC}(s_I) = \|\eta\|_\infty$ with*

$$\eta_j = \frac{C_{i_1 j} + C_{i_2 j}}{1 + C_{i_1 i_2}}, j \in I^C.$$

For $s_I = (+1, -1)^T$ we have $\operatorname{IC}(s_I) = \|\eta\|_\infty$ with

$$\eta_j = \frac{C_{i_1 j} - C_{i_2 j}}{1 - C_{i_1 i_2}}, j \in I^C.$$

With the observation in lemmas (4.4) and (4.8) that IC only depends on the entries of the covariance matrix, we can define a function $\eta : \mathbb{R} \to \mathbb{R}$ that represents the η-vector from the preceding proposition in every real point.

Definition 4.9 (Continuous η-vector for IC on positive sign)**.** Let $R > 0$ and $z_1, z_2 \in \mathbb{R}$. Then, for $z \in \mathbb{R}$ define, using the function $c : \mathbb{R} \to \mathbb{R}$ from (4.4)

$$\eta^R(z) := \frac{c^R(z - z_1) + c^R(z - z_2)}{1 + c^R(z_1 - z_2)}$$

Remark 4.10. One observes that this function is an interpolator which uses c^R as a kernel. This can be generalized to several Diracs using the definition of IC in the continuous setting and yields a rational function in different evaluations of the function c^R.

Figure 1 shows the values η can take for different distances between sources. The behaviour of this function immediately poses constraints on the allowed discretization grid on which one may place the sources. If one discretizes too finely, then already in the two-delta case with sources for example in z_{i_1}, z_{i_2}, there will exist a grid point z_j such that $\eta(z_j) > 1$. This violates IC and under a certain bound on the noise will lead to false source identification when using the lasso. We formalize this in the

Numerical observation 4.11. *Let $R > 0$ and $z_1, z_2 \in \mathbb{R}, z_1 < z_2$. Then for η^R defined as in (4.9) we have the following observation: There exists $\delta > 0$ such that $\eta(z_1 + \delta) = \eta(z_2 - \delta) > 1$.*

This observation rules out super-resolution with exact signal sign recovery at arbitrary discretizations. A minimum distance between potential sources is required in order to guarantee sign recovery in the 2-delta case. This is not uncommon, since we can imagine many convolution kernels that may produce the same effect. We can define

Definition 4.12 (Minimum discretization step)**.** Let $R > 0$ and $d > 0$. Let $z_1 = 0$ and $z_2 = d$. Then define

$$\Delta(d, R) := \left\{ \begin{array}{cc} \infty & \text{if } \min\{\delta > 0 | \eta^R(\delta) = 1\} \geq d \\ \min\{\delta > 0 | \eta^R(\delta) = 1\} & \text{otherwise} \end{array} \right.$$

Given and imaging resolution $R > 0$ and a minimum source distance $d > 0$, any source space discretization under $\Delta(d, R)$ will lead to violation of IC and thus the impossibility of correct support and sign recovery for two Diracs. The function is depicted in Figure 1 on the right. The white region depicts $\Delta(d, R) = +\infty$ indicating recovery failure due to a too small source separation. The lower border line to recovery failure and the iso-colour-lines vary superlinearly with R, showing a severe degradation in resolution with increasing R. This indicates the necessity of placing the detectors as close as possible to the sources.

5. Conclusions and Outlook

We have presented some classical results on sign and support dependent recovery criteria due to Fuchs, Tropp and Grasmair. Based on the tightness of the IC criterion for sign stability in a low-noise setting, we set out to study an idealized setting of MEG measurement, imposing high symmetry by placing sources on a straight line and measuring on a cylinder of circumference $R > 0$. Taking the limit of infinite measurements on this cylinder we find ourselves in the setting of spike deconvolution. Using Fuchs IC we assess the sign recovery properties as a function of imaging scale $R > 0$ and minimum source separation $d > 0$ and find a minimum discretization step $\Delta(d, R) > 0$, which is always strictly positive, under which sign recovery of two Diracs necessarily fails.

In real life, however, we do not have infinite measurement power on a surface surrounding the head. Indeed, simulations show a degradation in recovery capacity, significantly enlarging the minimum source separation with respect to the continuous setting. With a finite or discrete number of measurements the translation invariance in source position is also disturbed. Even if we did have infinite measurements, given the structure of the brain surface, we would not be working in a convolution setting with a constant convolution kernel. In several aspects, we are dealing with a more difficult problem. Several questions emerge.

1. Can we create a source grid adapted to the brain surface that under realistic situations permits stable recovery of the signal sign? Answering this question takes more than analysing two Diracs on toy data. A realistic forward model is needed based on the given non-vacuum setting along with an estimate of the noise. The possible source locations grid must then be very well chosen, since it should ideally correspond in a best possible manner to actual source locations. The minimum source separation may potentially be high.
2. Can we do a similar theoretical analysis on sources of variable orientation? The setting with free orientation analog to the one described above would entail the use of a sign dependent criterion for stable recovery using the group lasso [40], since coordinate orientations need to be treated together. As briefly mentioned above, the group lasso uses a different penalty, *grouping* several coefficients together: Let $\mathcal{G}$ be a disjoint collection of subsets of $\{1 \dots N\}$ such that their union is $\{1 \dots N\}$. Then the penalty reads $\sum_{g \in \mathcal{G}} \|x_g\|_2$. The variables in a group $g \in G$ are activated or set to 0 together. The notion of a discrete sign must be abandoned and replaced by the *generalized sign* $\frac{x_g}{\|x_g\|_2}$. Although an equivalent to ERC exists in the group setting (see [1], where it is used to prove consistency), the generalized-sign-dependent criterion is a subject of current research.
3. Can we create a setting in which the theory of super-resolution of Candès and Granda may be applied to this problem? If so, this would entail the possibility of localizing sources based only on a minimum distance criterion.
4. In connection with the preceding question: can we place detectors in a more optimal way than is presently done? This question can be addressed using simulations with IC or with a super-resolution theory.
5. Is the prior information that sources are sparse, which we implicitly impose, correct enough? Would other signal priors compatible with the superresolution theory, for example a more general analysis prior as explained in [14, 25] be better adapted?

References

[1] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[3] Emmanuel Candès and Terence Tao. Near optimal signal recovery from random projections: Universal encoding strategies?, 2004.

[4] Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *CoRR*, abs/1203.5871, 2012.

[5] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, 2004.

[6] Shaobing Chen and David Donoho. Basis pursuit. Technical report, Statistics Department, Stanford University, 1994.

[7] D. Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, August 1968.

[8] David L. Donoho and Michael Elad. Elad m 2003 optimally sparse representation in general (non-orthogonal) dictionaries via 1 minimization. In *Proc. Natl Acad. Sci. USA 100 2197202*, 2003.

[9] David L. Donoho and B. F. Logan. Signal recovery and the large sieve. *SIAM Journal of Applied Mathematics*, 1992.

[10] David L. Donoho and Yaakov Tsaig. Fast solution of 1-norm minimization problems when the solution may be sparse. 2006.

[11] Charles Dossal, Maher Kachour, Jalal Fadili, Gabriel Peyré, and Christophe Chesneau. The degrees of freedom of the Lasso for general design matrix. Previously entitled ”The degrees of freedom of penalized l1 minimization”, August 2011.

[12] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[13] M. Elad and A. M. Bruckstein. On sparse representation. In *Proceedings of the International Conference on Image Processing (ICIP) 2001*, 2001.

[14] Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. Technical report, 2005.

[15] Feuer and Nemirovsky. On sparse representations in pairs of bases. *IEEE Transactions on Information Theory*, June 2003.

[16] Jean Jacques Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Th*, page 1344, 2004.

[17] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, 2009.

[18] Alexandre Gramfort, Matthieu Kowalski, and Matti Hämäläinen. Mixed-norm estimates for the m/eeg inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57(7):1937–1961, Mar 2012.

[19] Alexandre Gramfort, Daniel Strohmeier, J. Haueisen, M. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *submitted to IPMI*, 2012.

[20] Markus Grasmair, Otmar Scherzer, and Markus Haltmeier. Necessary and sufficient conditions for linear convergence of 1-regularization. *Communications on Pure and Applied Mathematics*, 64(2):161–182, 2011.

[21] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[22] Olaf Hauk, Daniel G. Wakeman, and Richard Henson. Comparison of noise-normalized minimum norm estimates for meg analysis using multiple resolution metrics. *Neuroimage*, February 2011.

[23] Shlomo Levy and Peter K. Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, September 1981.

[24] Stéphane Mallat. *A Wavelet Tour of Signal Processing.* AP Professional, London, 1997.

[25] Sangnam Nam, Michael E. Davies, Michael Elad, and Rémi Gribonval. Cosparse analysis modeling - uniqueness and algorithms. In *ICASSP*, pages 5804–5807, 2011.

[26] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[27] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *In ICML*, 2004.

[28] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999.

[29] Sung C. Park, Min K. Park, and Moon G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, May 2003.

[30] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, November 1992.

[31] Fadil Santosa and William W. Symes. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

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

[33] Ryan Tibshirani. The lasso problem and uniqueness. *ArXiv*, June 2012.

[34] Joel A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise, 2006.

[35] Yaakov Tsaig and David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[36] S. Vaiter, C. Deledalle, G. Peyré, C. Dossal, and J. Fadili. Local behavior of sparse analysis regularization: Applications to risk estimation. Technical report, CEREMADE, Université Paris 9 Dauphine, 2012.

[37] S. Vaiter, C. Delledalle, G. Peyré, J. Fadili, and C. Dossal. The degrees of freedom of the group lasso. *ICML2012*, 2012.

[38] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili. Robust sparse analysis regularization. Technical report, Preprint Hal-00627452, 2011.

[39] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programmming (lasso), 2006.

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

[41] Peng Zhao and Bin Yu. On model selection consistency of lasso, 2006.

[42] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006.

Inria Parietal, Neurospin, Bât. 145, CEA Saclay, 91191 Gif-sur-Yvette
E-mail address: michael.eickenberg@nsup.org