Recovering Sparse Low-rank Blocks in Tandem Mass Spectrometry

Christoph Studer[1], Graeme Pope[2], Pedro Navarro[2], and Richard G. Baraniuk[1]

[1]Rice University, Houston, TX, USA; e-mail: {studer, richb}@rice.edu
[2]ETH Zürich, Zürich, Switzerland; e-mail: graemepope@gmail.com, navarro@imsb.biol.ethz.ch

***Abstract*—We develop a novel sparse low-rank block (SLoB) signal recovery framework that simultaneously exploits sparsity and low-rankness to accurately identify peptides (fragments of proteins) from biological samples observed using tandem mass spectrometery (TMS). To efficiently perform SLoB-based peptide identification, we propose two novel recovery algorithms: An exact iterative method based on the alternating method of multipliers (ADMM) and an approximate greedy algorithm that extends orthogonal matching pursuit (OMP) to the SLoB framework. For the exact ADMM algorithm, we provide analytical conditions as to when the underlying convex optimization is capable of detecting the peptides in a given sample. We finally demonstrate that the SLoB framework and the proposed algorithms substantially outperform existing sparse signal recovery techniques for peptide detection with synthetic and real-world TMS data.**

I. Introduction

A. Peptide identification and proteomics

The identification of peptides (fragments of proteins) is key for understanding which proteins are present in biological samples. Since proteins control the processes of the human body, their identification and understanding is a fundamental area of research, including the fight against cancer [2], [3] and Alzheimer's disease [4]. The study of peptides is known as *proteomics* and one standard approach for protein identification is to split the proteins into its peptide fragments [5]–[8] and then, to identify these peptides using tandem mass spectrometery (TMS).

In this paper, we focus on a specific measurement process developed in [9], [10], and we show how one can accurately and efficiently identify the peptides in a biological sample from a series of TMS measurements taken over time. The proposed approach relies on finding sparse low-rank blocks (SLoBs) in measured data that—together with a known dictionary of the peptides—accurately model the TMS measurement process. The generality of the proposed framework and the associated recovery methods find potential use in related areas of compressive sensing [11], [12], sparse signal recovery [13], and matrix completion [14]–[16].

B. Contributions

We present a novel peptide detection framework that relies on sparse signal recovery and low-rank methods, as an alternative to the block multiple-measurement vector (B-MMV) problem [17]. To arrive at low complexity methods for peptide identification, we propose two recovery algorithms: (i) A convex optimization procedure, that relies on the alternating direction method of multipliers (ADMM) [18], [19], and (ii) a greedy algorithm, which is well-suited for peptide identification in large datasets. For the convex optimization approach, we develop a coherence-based recovery guarantee, which provides analytical conditions as to when the underlying convex optimization problem is capable of detecting the peptides in a given sample. Finally, we present numerical experiments with both synthetic data and real world TMS measurements, to demonstrate the efficacy of the proposed SLoB framework.

C. Notation

Lowercase and uppercase boldface letters stand for vectors and matrices, respectively. For the matrix $\mathbf{A}$, we denote its transpose by $\mathbf{A}^T$. The j^{th} column and the entry in the i^{th} row and j^{th} column of a matrix $\mathbf{A}$ is designated by $\mathbf{a}_j$ and $[\mathbf{A}]_{i,j}$, respectively. We write $\mathbf{A}\otimes\mathbf{B}$ and $\mathbf{A}\odot\mathbf{B}$ for the Kronecker and the Hadamard product (i.e., the entry-wise product) of $\mathbf{A}$ and $\mathbf{B}$, respectively. The Frobenius and nuclear norm of a matrix are defined as $\|\mathbf{A}\|_F=\sqrt{\sum_{i,j}[\mathbf{A}]_{i,j}^2}$ and as $\|\mathbf{A}\|_*=\sum_{i=1}^r\sigma_i(\mathbf{A})$ with $r=\text{rank}(\mathbf{A})$ and the singular values $\sigma_i(\mathbf{A})$ for $i=1,\ldots,r$ of of $\mathbf{M}$, respectively. The maximum and minimum singular values of $\mathbf{A}$ are designated by $\sigma_{\text{max}}(\mathbf{A})$ and $\sigma_{\text{min}}(\mathbf{A})$, respectively.

D. Outline of the paper

The remainder of the paper is organized as follows. Section II introduces the signal and system model. Section III summarizes the analytical recovery guarantee. Section IV details two peptide identification algorithms. Section V discusses experimental results. We conclude in Section VI.

Parts of this paper have appeared previously in G. Pope's Ph.D. Thesis [1].

The authors would like to thank S. Christen for inspiring discussions, R. Aebersold, L. Gillet, and G. Rosenberger for making their data available, and H. Bölcskei for his support.

This work was supported in part by the Grants NSF CCF-0431150, CCF-0728867, CCF-0926127, DARPA/ONR N66001-08-1-2065, N66001-11-1-4090, N66001-11-C-4092, ONR N00014-08-1-1112, N00014-10-1-0989, AFOSR FA9550-09-1-0432, ARO MURIs W911NF-07-1-0185 and W911NF-09-1-0383, and by the Texas Instruments Leadership University Program.

II. System Model and Recovery Problem

To model the TMS measurement process of [9], [10], we assume that the $\mathrm{m/z}$ (mass over charge) spectrum of a precursor (or peptide fragment) can be represented as a vector in $\mathbb{R}^m$, where each entry of the vector corresponds to the number of particles measured in a particular $\mathrm{m/z}$ interval. Assume that we also have a list of n peptides (and their fragments) that we are interested in detecting. Note that the spectra of many peptides are known and can, for example, be found in a database such as PeptideAtlas [20]. Let the i^{th} peptide have $\mathrm{m/z}$ spectrum $\mathbf{d}_i^{(0)}$ (where we use 0 to denote that this is the parent peptide) and its $f_i - 1$ fragments will have spectra $\mathbf{d}_i^{(j)}$, $j = 1, \ldots, f_i - 1$, where each vector $\mathbf{d}_i^{(j)}$ is normalized to have unit ℓ_2 norm. Then, form the $m \times f_i$ dictionary $\mathbf{D}_i = \left[\mathbf{d}_i^{(0)} \cdots \mathbf{d}_i^{(f_i-1)}\right]$, which characterizes the spectrum of the i^{th} peptide and all its fragments. Now, let $\mathcal{S}_j$ be the set (with cardinality n) of all precursors that are present at measurement instant j, so that the observation $\mathbf{z}_j$ is given by the following input-output relation:

$$\mathbf{z}_j = \sum_{\ell \in \mathcal{S}_j} \mathbf{D}_\ell \mathbf{x}_j[\ell] + \mathbf{n}_j = \sum_{i=1}^{n} \mathbf{D}_i \mathbf{x}_j[i] + \mathbf{n}_j, \qquad (1)$$

where $\mathbf{x}_j[\ell] \in \mathbb{R}^{f_\ell}$ denotes how much of each fragment of the i^{th} peptide is present at measurement instant j, also referred to as the intensity. The vector $\mathbf{n}_j$ models additive measurement noise occurring in the TMS measurement process. Since we observe $j = 1, \ldots, T$ TMS spectra over multiple measurement instants, we can rewrite (1) as

$$\mathbf{Z} = \sum_{i=1}^{n} \mathbf{D}_i \mathbf{X}_i + \mathbf{N}. \qquad (2)$$

Here, $\mathbf{Z} \in \mathbb{R}^{m \times T}$, $\mathbf{X}_i \in \mathbb{R}^{f_i \times T}$ and $\mathbf{N} \in \mathbb{R}^{m \times T}$ are matrices containing as columns the vectors $\mathbf{z}_j$, $\mathbf{x}_j[i]$, and $\mathbf{n}_j$ as appropriate. We are now interested in finding the solution to the following problem: Given a collection of measurements $\mathbf{Z}$ and the peptide dictionary blocks $\mathbf{D}_i$, recover the intensities $\mathbf{X}_i$, for $i = 1, \ldots, n$, which correspond to the peptides that are present in the TMS measurements.

A straightforward way to solve this problem is to deploy a combination of a multiple-measurement vector (MMV) problem with block sparsity. Specifically, instead of taking the $\ell_{2,1}$-norm of the (vector) blocks (occurring in both the (MMV) and block-sparse recovery problem), we take the Frobenius norm of the matrix blocks $\mathbf{X}_i$. Concretely, one may solve the following block MMV problem:

$$\text{(B-MMV)} \quad \begin{cases} \underset{\hat{\mathbf{X}}_1, \ldots, \hat{\mathbf{X}}_n}{\text{minimize}} & \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i\|_F \\ \text{subject to} & \|\mathbf{Z} - \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{X}}_i\|_F \leqslant \varepsilon. \end{cases}$$

Here, the parameter $\varepsilon \geq 0$ needs to be chosen in accordance to the Frobenius norm of the measurement noise.

We emphasize that the (B-MMV) problem makes no assumption about the intensity values in any of the blocks $\mathbf{X}_i$. However, for real-world measurements, each of these blocks will—at least ideally—be rank one, so that we can write $\mathbf{X}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ where $\mathbf{u}_i$ contains the ratio of the fragmented ions and $\mathbf{v}_i$ can be regarded as a vector describing the flow rate of a precursor over time and all of its fragment ions. The scalar σ_i then gives the intensity after $\mathbf{u}_i$ and $\mathbf{v}_i$ are normalized to unit ℓ_2-norm. However, since a rank constraint would result in a non-convex optimization problem, we relax this constraint to the nuclear norm [14]–[16], to obtain the following convex sparse low-rank block (SLoB) recovery problem:

$$\text{(N-MMV)} \quad \begin{cases} \underset{\hat{\mathbf{X}}_1, \ldots, \hat{\mathbf{X}}_n}{\text{minimize}} & \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i\|_* \\ \text{subject to} & \|\mathbf{Z} - \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{X}}_i\|_F \leqslant \varepsilon. \end{cases}$$

This nuclear-norm MMV (N-MMV) recovery problem will be our main focus in the remainder of the paper.

III. Recovery Guarantee

In order to gain insight into the recovery performance of (B-MMV) and (N-MMV), we start by defining an appropriate notion of coherence. By assuming that each block $\mathbf{D}_i$ is normalized to have $\sigma_{\min}(\mathbf{D}_i) = 1$, we obtain a SLoB coherence parameter $\mu_{\mathcal{D}}$ defined as

$$\mu_{\mathcal{D}} = \max_{k,\ell \colon k \neq \ell} \sup_{\mathbf{X} \neq 0} \frac{\|\mathbf{D}_k^* \mathbf{D}_\ell \mathbf{X}\|_*}{\|\mathbf{X}\|_*} = \max_{k,\ell \colon k \neq \ell} \sigma_{\max}(\mathbf{D}_k^* \mathbf{D}_\ell). \qquad (3)$$

With this notion of coherence, we can deploy [1, Thm. 2.6] to prove the following theorem stating when (B-MMV) and (N-MMV) perfectly recover the blocks $\mathbf{D}_i$ and the associated intensities $\mathbf{X}_i$ from the noiseless observations $\mathbf{Z}$.[1]

Theorem 1: Let $\mathbf{Z} = \sum_{i=1}^{n} \mathbf{D}_i \mathbf{X}_i$ and $s \leq n$ be the number of non-zero blocks $\mathbf{X}_i$, $i = 1, \ldots, n$. If

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu_{\mathcal{D}}}\right), \qquad (4)$$

then the solutions of (B-MMV) and (N-MMV) using $\varepsilon = 0$ are unique and equal to the ground truth intensities $\mathbf{X}_i$.

We emphasize that the results for synthetic and real-world data shown in Section V demonstrate that (N-MMV) significantly outperforms (B-MMV) in most situations. However, there is no dependence on the rank of the individual blocks in the condition (4) of Theorem 1. The reason for this behavior is the fact that both optimization problems suffer from the same worst-case signals for which (4) is "just violated." Specifically, one can design particular instances of blocks $\mathbf{X}_i$ that are either full-rank or rank one, which both (B-MMV) and (N-MMV) cannot distinguish between (see [1, Sec. 4.3] for the details). In order to obtain rank-dependent recovery guarantees, one needs further assumptions on the signals; the corresponding analysis of such assumptions and recovery guarantees is part of ongoing work. Nevertheless, Theorem 1 provides insight into the $\mathrm{m/z}$ spectra of peptides (and their fragments) that can be recovered via the SLoBs framework. In particular, Theorem 1 states that the dictionary blocks $\mathbf{D}_i$

[1] Note that by following the approach of [21], [22], Theorem 1 can be extended to the case of stable recovery of the blocks $\mathbf{D}_i$ and the intensities $\mathbf{X}_i$ with arbitrary (but bounded) noise, if condition (4) is satisfied.

must be sufficiently incoherent to enable perfect recovery from the TMS measurements contained in $\mathbf{Z}$. In other words, the spectral signatures of the peptides to be detected must be sufficiently distinct.

IV. Recovery Algorithms

In this section, we detail two distinct methods for solving (N-MMV). The first method is an iterative algorithm that exactly solves the convex optimization problem (N-MMV); the second method is a greedy algorithm that finds an approximate solution in a computationally efficient manner.

*A. Iterative algorithm (*N-MMV-L*)*

The iterative algorithm detailed next relies on the alternating direction method of multipliers (ADMM) [18], [19].

1) Reformulating the optimization problem: In order to solve the (N-MMV) problem directly, we next describe an efficient iterative method based on the alternating direction method of multipliers (ADMM) [18], [19]. To do so, we reformulate (N-MMV) by introducing the following auxiliary matrices: $\hat{\mathbf{W}} = \mathbf{Z} - \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{X}}_i$ and $\hat{\mathbf{Y}}_i = \hat{\mathbf{X}}_i$, for $i = 1, \ldots, n$. The reformulated (and equivalent) optimization problem is given by

$$
(\text{N-MMV}^\star) \quad \begin{cases} \underset{\hat{\mathbf{X}}_i, \hat{\mathbf{Y}}_i, \forall i, \hat{\mathbf{W}}}{\text{minimize}} & \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i\|_* \\ \text{subject to} & \|\hat{\mathbf{W}}\|_F \leqslant \varepsilon, \\ & \hat{\mathbf{W}} = \mathbf{Z} - \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{Y}}_i, \\ & \hat{\mathbf{X}}_i = \hat{\mathbf{Y}}_i,\ i = 1, \ldots, n. \end{cases}
$$

To arrive at an efficient way of solving (N-MMV*), we relax the linear constraints to form its augmented Lagrangian

$$
(\text{N-MMV-L}) \quad \begin{cases} \underset{\hat{\mathbf{X}}_i, \hat{\mathbf{Y}}_i, \forall i, \hat{\mathbf{W}}}{\text{minimize}} & \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i\|_* + \\ & \frac{\beta_1}{2} \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i - \hat{\mathbf{Y}}_i - \boldsymbol{\Lambda}_i\|_F^2 + \\ & \frac{\beta_2}{2} \|\hat{\mathbf{W}} - \mathbf{Z} + \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{Y}}_i - \boldsymbol{\Omega}\|_F^2 \\ \text{subject to} & \|\hat{\mathbf{W}}\|_F \leqslant \varepsilon. \end{cases}
$$

This augmented Lagrangian problem can be viewed as an instance of the Douglas-Rachford variable splitting method commonly used in convex optimization [18], [19]. In (N-MMV-L), the matrices $\boldsymbol{\Lambda}_i \in \mathbb{R}^{f_i \times T}$, $i = 1, \ldots, n$, and $\boldsymbol{\Omega} \in \mathbb{R}^{m \times T}$ correspond to (scaled) Lagrange multipliers.

*2) Solving (*N-MMV-L*) via ADMM:* We now describe an iterative method that finds the solution to (N-MMV-L) using two nested loops. Rather than jointly optimizing for the $\hat{\mathbf{X}}_i$, $\hat{\mathbf{Y}}_i$ for $i = 1, \ldots, n$ and for $\hat{\mathbf{W}}$, we optimize each individually in the inner iterations, while keeping all others fixed. This procedure has the advantage of separating (N-MMV-L) into three main (inner) sub-problems each of which can be solved efficiently. In the outer loop, we update the Lagrange multipliers $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Omega}$ in standard ADMM fashion.

The proposed ADMM procedure is given as follows. Let the variables $k = 1, 2, \ldots$ and $\ell = 1, 2, \ldots$ denote the inner and outer iteration counters, respectively. For $k = \ell = 1$ the algorithm is initialized with $\hat{\mathbf{Y}}_i^{(1)} = \mathbf{0}$, $\forall i$, $\hat{\mathbf{W}}^{(1)} = \mathbf{0}$, ${\boldsymbol{\Lambda}_i}^{(1)} = \mathbf{0}$, $\forall i$, and $\boldsymbol{\Omega}^{(1)} = \mathbf{0}$. We then perform the following three steps until one of the following stopping criteria is reached: (i) the maximum number of K_{in} inner iterations is achieved, or (ii) if the objective function of (N-MMV-L) decreases by less than τ_{in} from one inner iteration to the next.

Step 1: Update the matrices $\hat{\mathbf{Y}}_i$: For each matrix $\hat{\mathbf{Y}}_i^{(k)}$, we fix the matrices $\hat{\mathbf{X}}_i^{(k)}$, for $i = 1, \ldots, n$, $\hat{\mathbf{Y}}_j^{(k)}$, for $j \neq i$, and $\mathbf{W}$. Then, by only considering the terms in the summand of the objective function that depend on $\hat{\mathbf{Y}}_i^{(k)}$, we obtain a new estimate $\hat{\mathbf{Y}}_i^{(k+1)}$ by solving the following (unconstrained) optimization problem:

$$
\hat{\mathbf{Y}}_i^{(k+1)} \leftarrow \underset{\tilde{\mathbf{Y}}_i}{\arg\min} \bigg\{ \frac{\beta_1}{2} \left\| \hat{\mathbf{X}}_i^{(k)} - \tilde{\mathbf{Y}}_i - {\boldsymbol{\Lambda}_i}^{(\ell)} \right\|_F^2 + \frac{\beta_2}{2} \left\| \hat{\mathbf{W}}^{(k)} - \mathbf{Z} + \mathbf{P}_i + \mathbf{D}_i \tilde{\mathbf{Y}}_i - \boldsymbol{\Omega}^{(\ell)} \right\|_F^2 \bigg\}.
$$

Here, $\mathbf{P}_i = \sum_{\ell \neq i} \mathbf{D}_\ell \hat{\mathbf{Y}}_\ell^{(k)}$ and $\tilde{\mathbf{Y}}_i$ is the minimization variable. We emphasize that the squared Frobenius norm of a matrix $\mathbf{A}$ can be rewritten as $\|\mathbf{A}\|_F^2 = \sum_\ell \|\mathbf{a}_\ell\|_2^2$ and hence, by considering each column of $\tilde{\mathbf{Y}}_i$ separately, one can solve the above problem using T *independent* least-squares (LS) problems. To achieve low computational complexity, we solve these LS problems using off-the-shelf conjugate gradient methods (CGLS) [23].

Step 2: Update the matrices $\hat{\mathbf{X}}_i$: For each matrix $\hat{\mathbf{X}}_i^{(k)}$, $i = 1, \ldots, n$, we only consider the terms in the summand that involve $\hat{\mathbf{X}}_i^{(k)}$. Thus we obtain a new estimate $\hat{\mathbf{X}}_i^{(k+1)}$ by solving the following (unconstrained) optimization problem:

$$
\hat{\mathbf{X}}_i^{(k+1)} \leftarrow \underset{\tilde{\mathbf{X}}}{\arg\min} \left\{ \left\| \tilde{\mathbf{X}} \right\|_* + \frac{\beta_1}{2} \left\| \tilde{\mathbf{X}} - \hat{\mathbf{Y}}_i^{(k+1)} - {\boldsymbol{\Lambda}_i}^{(\ell)} \right\|_F^2 \right\}.
$$

This optimization problem admits an efficient closed-form solution via the singular value shrinkage operator [14], [24]. Specifically, set $\mathbf{Q}_i = \hat{\mathbf{Y}}_i^{(k+1)} + {\boldsymbol{\Lambda}_i}^{(\ell)}$ and perform the singular value decomposition $\mathbf{Q}_i = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $[\mathbf{S}]_{\ell,\ell} = \sigma_\ell$ are the singular values of $\mathbf{Q}_i$. With this decomposition, the updated matrix $\hat{\mathbf{X}}_i^{(k+1)}$ is given by [14], [24]

$$
\hat{\mathbf{X}}_i^{(k+1)} \leftarrow \mathbf{U}\eta(\mathbf{S})\mathbf{V}^T
$$

with the shrinkage operator $[\eta(\mathbf{S})]_{k,\ell} = 0$, for $k \neq \ell$ and

$$
[\eta(\mathbf{S})]_{\ell,\ell} = \max\left\{ \sigma_\ell - \frac{1}{\beta_1}, 0 \right\}, \forall \ell.
$$

Step 3: Update the matrix $\hat{\mathbf{W}}$: In this step, we only consider the terms in the objective function that involve $\hat{\mathbf{W}}$ to obtain a new estimate $\hat{\mathbf{W}}^{(k+1)}$ given by the solution of the following (constrained) optimization problem:

$$
\hat{\mathbf{W}}^{(k+1)} \leftarrow \underset{\tilde{\mathbf{W}}, \|\tilde{\mathbf{W}}\|_F \leqslant \varepsilon}{\arg\min} \left\| \tilde{\mathbf{W}} - \mathbf{Z} + \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{Y}}_i^{(k+1)} - \boldsymbol{\Omega}^{(\ell)} \right\|_F^2.
$$

This problem has a closed-form solution, which is obtained by first calculating

$$
\mathbf{E} = \mathbf{Z} - \sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{Y}}_i^{(k+1)} + \boldsymbol{\Omega}^{(\ell)},
$$

followed by projecting the resulting matrix $\mathbf{E}$ onto the Frobenius ball of radius ε as follows

$$\hat{\mathbf{W}}^{(k+1)} \leftarrow \begin{cases} \mathbf{E} & \text{if } \|\mathbf{E}\|_F \leqslant \varepsilon \\ \mathbf{E}/\|\mathbf{E}\|_F & \text{otherwise.} \end{cases}$$

After the stopping criteria of the inner loop is met, we proceed with the outer iteration where we update the Lagrange multipliers $\boldsymbol{\Lambda}_\ell$, $\ell = 1, \ldots, n$, and $\boldsymbol{\Omega}$ in the standard manner [18]:

$$\boldsymbol{\Lambda}_i{}^{(\ell+1)} \leftarrow \boldsymbol{\Lambda}_i{}^{(\ell)} - \mu\Big(\hat{\mathbf{X}}_i^{(k+1)} - \hat{\mathbf{Y}}_i^{(k+1)}\Big), \forall i$$

$$\boldsymbol{\Omega}^{(\ell+1)} \leftarrow \boldsymbol{\Omega}^{(\ell)} - \mu\Big(\hat{\mathbf{W}}^{(k+1)} - \mathbf{Z} + \textstyle\sum_{i=1}^{n} \mathbf{D}_i \hat{\mathbf{Y}}_i^{(k+1)}\Big).$$

Here, the parameter $\mu \geq 0$ is an appropriate step-size. The algorithm detailed above continues to perform the inner iterations, followed by updating the Lagrange multipliers until either a maximum number of outer iterations K_{out} is reached or if the objective function of (N-MMV-L) converges. The convergence behavior of this ADMM method, as well as its computational complexity, are affected by the parameters K_{out}, K_{in}, β_1, β_2 and μ (see, e.g., [18] for additional details). In all experiments shown in Section V, we set $K_{\text{out}} = 100$, $K_{\text{in}} = 5$, $\beta_1 = 1$, $\beta_2 = 1$ and $\mu = (1+\sqrt{5})/2$, which delivers excellent recovery performance at low computational complexity.

*B. Greedy algorithm (*N-OMP*)*

Orthogonal matching pursuit (OMP) [25], [26] was developed as an iterative greedy alternative to the classical basis pursuit algorithm [27]. Analogously, we next propose a greedy alternative that attempts to find an approximate solution to the (N-MMV-L) problem.

1) Algorithm outline: The general form of such a greedy algorithm performs the following steps until either a predetermined number of iterations is reached, or the residual (given by $\mathbf{R}^{(i)} = \mathbf{Z} - \sum_j \mathbf{D}_j \hat{\mathbf{X}}_j^{(i)}$, where $\hat{\mathbf{X}}_j^{(i)}$ is the current approximation to $\mathbf{X}_j$) is sufficiently small. To this end, assume that prior to the i^{th} iteration we have a solution $\{\hat{\mathbf{X}}_1^{(i)}, \ldots, \hat{\mathbf{X}}_n^{(i)}\}$ with $r_j^{(i)} = \text{rank}(\mathbf{X}_j^{(i)})$ and $i = \sum_{j=1}^{n} r_j^{(i)}$. Then, we perform the following steps:

1) Identify the matrix $\mathbf{X}_\ell$ in which we want to increment the rank. Then, set $r_\ell^{(i+1)} = r_\ell^{(i)} + 1$ and $r_j^{(i+1)} = r_j^{(i)}$ for all $j \neq \ell$.
2) Find the matrices $\{\hat{\mathbf{X}}_1^{(i+1)}, \ldots, \hat{\mathbf{X}}_n^{(i+1)}\}$ that solve the following optimization problem:

$$\begin{cases} \underset{\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_n}{\text{minimize}} & \Big\|\mathbf{Z} - \sum_{j=1}^{n} \mathbf{D}_j \tilde{\mathbf{X}}_j\Big\|_F \\ \text{subject to} & \text{rank}(\tilde{\mathbf{X}}_j) \leqslant r_j^{(i+1)} \; \forall j. \end{cases} \tag{5}$$

3) Update the residual as $\mathbf{R}^{(i+1)} = \mathbf{Z} - \sum_{k=1}^{n} \mathbf{D}_k \hat{\mathbf{X}}_k^{(i+1)}$.

The first two steps are non-trivial and contrary to first appearances, the solution to (5) is not just given by a singular value decomposition. Hence, we propose the following heuristic: Find the block $\mathbf{D}_\ell$ of the dictionary that is most correlated with the residual, i.e., find the block that maximizes $\big\|\mathbf{D}_\ell^* \mathbf{R}^{(i)}\big\|_F$. Note that this would be the block that would most reduce the objective function of (5) without considering the rank constraints. This approach resembles the Wiberg algorithm [28] which is used to solve the problem

$$\underset{\hat{\mathbf{U}},\hat{\mathbf{V}}}{\text{minimize}} \; \|\mathbf{W} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V})\|_F \,, \tag{6}$$

given $\mathbf{W}$ and $\mathbf{Y}$, and where $\mathbf{U}$ and $\mathbf{V}$ are rank-r matrices. The Wiberg algorithm is an alternating least squares (ALS) approach relying on the observation that by alternately fixing the matrices $\mathbf{U}$ and $\mathbf{V}$, one can transform the problem into separate LS minimization problems.

2) The Wiberg algorithm: To apply the Wiberg algorithm to our problem, let $\mathbf{U}_i \boldsymbol{\Sigma}_i \mathbf{V}_i^T$ be the singular value decomposition of $\hat{\mathbf{X}}_i$ (which has rank r_i) and set $\tilde{\mathbf{U}}_i = \mathbf{U}_i \boldsymbol{\Sigma}_i$ and $\tilde{\mathbf{V}}_i = \mathbf{V}_i \boldsymbol{\Sigma}_i$ to be the scaled version of the left and right singular vectors of $\hat{\mathbf{X}}_i$. Then, $\sum_{i=1}^{n} \mathbf{D}_i \mathbf{X}_i = \sum_{i=1}^{n} \mathbf{D}_i \mathbf{U}_i \boldsymbol{\Sigma}_i \mathbf{V}_i^T$, and furthermore, for each i we have

$$\text{vec}(\mathbf{D}_i \mathbf{X}_i) = \text{vec}(\mathbf{D}_i \mathbf{U}_i \boldsymbol{\Sigma}_i \mathbf{V}_i^T) \overset{(a)}{=} \Big(\mathbf{I}_T \otimes \mathbf{D}_i \tilde{\mathbf{U}}_i\Big) \text{vec}(\mathbf{V}_i^T),$$

where (a) follows since $\text{vec}(\mathbf{A}\mathbf{B}) = (\mathbf{I} \otimes \mathbf{A}) \, \text{vec}(\mathbf{B})$ (see [29, Sec. 4.3] for the details). Consequently, by fixing each of the matrices $\mathbf{U}_i$ and $\boldsymbol{\Sigma}_i$, we can rewrite the problem as

$$\underset{\mathbf{V}_1,\ldots,\mathbf{V}_n}{\text{minimize}} \; \Big\| \text{vec}(\mathbf{Z}) - \big[\mathbf{I}_T \otimes (\mathbf{D}_1 \tilde{\mathbf{U}}_1) \cdots \mathbf{I}_T \otimes (\mathbf{D}_n \tilde{\mathbf{U}}_n)\big] \times \big[\text{vec}(\mathbf{V}_1)^T \cdots \text{vec}(\mathbf{V}_n)^T\big]^T \Big\|_2^2, \tag{7}$$

which corresponds to a LS problem in the matrices $\mathbf{V}_i$ that can be solved efficiently. Analogously, one can isolate the terms $\mathbf{U}_i$ and solve the corresponding LS problem to update the matrices $\mathbf{U}_i$.

For large problem sizes, explicit calculation of the Kronecker product in (7) may require significant amounts of memory. However, the use of CGLS [23] enables us to solve the problem at low complexity. In particular, we directly compute the matrix vector products $(\mathbf{I}_T \otimes \mathbf{D}_i \tilde{\mathbf{U}}_i) \, \text{vec}(\mathbf{V}_i^T)$ as $\text{vec}(\mathbf{D}_i \mathbf{X}_i)$ by exploiting the vectorization properties of the Kronecker product [29, Sec. 4.3].

We note that for the case of a single pair of factors $\mathbf{U}$ and $\mathbf{V}$, the Wiberg algorithm has been shown to converge to the desired solution, albeit slowly [30]. However, there are currently no theoretical guarantees for convergence when dealing with multiple $\mathbf{U}_i$ and $\mathbf{V}_j$ terms, as in the problem at hand. Nevertheless, our own simulations have shown that this Wiberg-based approach delivers excellent performance for various datasets analyzed using our method.

V. Results

We now apply the two proposed recovery algorithms to synthetic and real measured TMS data, and we examine their respective performance for a number of scenarios.

A. Synthetic results

We begin by presenting empirical phase transition plots in Figure 1. Here we show the regions in which the algorithms are able to recover at least 99% of the test signals. We generate the i^{th} rank-r (for $r = 1, \ldots, 10$) block of size $d_i \times T$ by

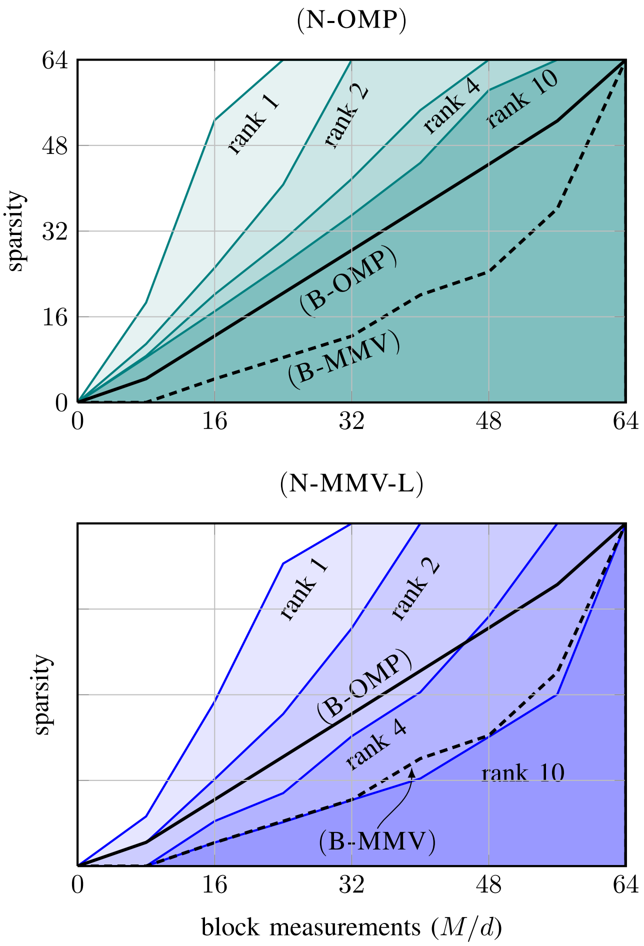

Figure 1. Empirical phase transition plots showing the region we can recover at least 99% of the signals with Gaussian i.i.d. blocks. Each block $\mathbf{X}_i$ has dimension 10×10. In the graph, "rank" refers to the rank of the non-zero matrices $\mathbf{X}_i$. Exploiting the low-rank structure via (N-OMP) and (N-MMV-L) significantly outperforms the (B-MMV) based algorithms.

multiplying together two matrices $\mathbf{A} \in \mathbb{R}^{d_i \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times T}$ with i.i.d. zero-mean Gaussian entries with variance $1/r$. We set $n = 64$, $f_i = 10$ (for all i) so that $N = 640$ and we take $T = 10$. We then sweep M from 80 to 640, and vary s from 4 to 64, which corresponds to the number of non-zero matrices $\mathbf{X}_i$. We refer to M/d as the number of block-measurements. We also compare these results to solving (B-MMV) or using (B-OMP), a greedy variant that finds an approximate solution to (B-MMV). Both of these methods ignore the low-rank structure in the matrices $\mathbf{X}_i$ and, hence. they exhibit virtually no rank dependence.

Figure 1 demonstrates that when dealing with low-rank blocks, the proposed nuclear-norm minimization approach is able to recover a substantially larger number of blocks, especially when the rank of the blocks is close to one, compared to solving (B-MMV) or using (B-OMP). Intuitively, this behavior makes sense, since fewer parameters are required to specify a rank-one matrix. In other words, fewer equations are required to uniquely specify the solution. Consequently, by exploiting the low-rank structure in the matrices $\mathbf{X}_i$, we can consistently recover a larger number of non-zero blocks.

Table I
Number of peptides identified via (B-OMP) and (N-OMP)

Sample	# peptides	(B-OMP)	(N-OMP)
L120224	342	140 (41%)	**317 (93%)**
L120225	342	156 (46%)	**321 (94%)**
L120227	342	140 (41%)	**319 (93%)**

B. *Hybrid real/synthetic experiments*

To further evaluate the performance of the proposed methods, we perform a mixture of real and synthetic experiments that use a real peptide dictionary and an artificial set of observations $\mathbf{Z}$, by randomly generating the matrices $\mathbf{X}_i$ (see [1, Sec. 4.5.2] for additional details). We generated a dictionary from the molecular description of the peptides, with an average of 51 fragments per peptide (corresponding to the block-size). We then quantized the dictionary and observations uniformly from $200\,\mathrm{Th}$ to $1000\,\mathrm{Th}$ in steps of $0.025\,\mathrm{Th}$, where Th refers to Thomsons, which is a common $\mathrm{m/z}$ measure unit. In total, there are $32,000$ quantization bins.

Let us take a closer look at a matrix $\mathbf{X}_i$ returned by (N-OMP) and (B-OMP) in Figure 2. The solution returned by (N-OMP) is rank 1 and we see that the right singular vector, which models to the flow rate, matches the ground truth. However, the right singular vectors of the (B-OMP) solution, which ignores the low-rank structure in the intensity matrices $\mathbf{X}_i$, bear no resemblance to the ground truth. The reason for this discrepancy between the dominant singular vector of the (B-OMP) solution and the original solution stems from the fact that each $\mathbf{D}_i$, despite being a tall matrix, is ill-conditioned. So although (B-OMP) is able to identify the present peptides in the mixture, it cannot accurately decompose the sample into its constituents. However, (N-OMP) imposes a low-rank structure into its solution, which enables us to cope with ill-conditioned dictionaries and thus, returns a solution that is much closer to the original signal.

C. *Experiments with real-world proteomics data*

We now evaluate the proposed SLoB framework and the associated recovery methods on actual proteomics data and analyze three different samples of peptides consisting of 342 known peptides. These samples have been acquired at the Institute for Molecular Systems Biology at ETH Zürich [10] and were measured using the TMS process described in [9] over a period of approximately 2 hours. Our calculations were performed across $T = 1500$ consecutive time-steps and use the same peptide dictionary as in Section V-B.

The results are shown in Table I, where we give: (i) the number of peptides in the sample, and (ii) the number of peptides identified using (B-OMP) and (N-OMP). We clearly see that by exploiting the low-rank structure of the acquired TMS measurements via (N-OMP), one is capable of successfully recovering a significantly higher percentage of the peptides present in real-world measurements, i.e., (N-OMP) substantially outperforms (B-OMP) for real-world proteomics

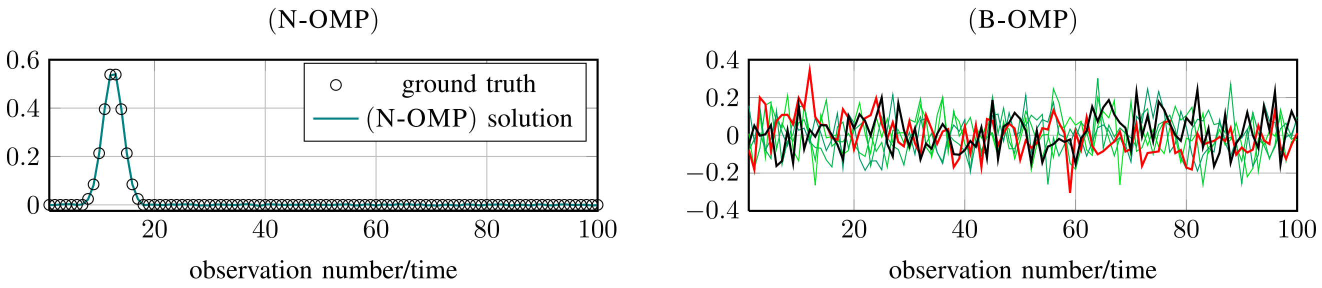

Figure 2. Right singular vectors for one block. For the (B-OMP) solution, the dominant singular vector is shown in black and the red line is the singular vector (corresponding to the 6th largest singular value) that most resembles the (N-OMP) solution. We see that the rank-aware (N-OMP) method accurately recovers the time/intensity behavior of the true solution, which is in stark contrast to (B-OMP) that ignores the low-rank structure.

data. Note that the performance of (N-MMV-L) is not shown here, as it delivers similar performance to (N-OMP).

VI. Conclusion

We have have developed a novel sparse low-rank block (SLoB) framework and corresponding recovery algorithms that are able to identify a large number of peptides in real-world biological samples—significantly more than by using a naïve sparsity-based approach. Our experimental results show that we can successfully distinguish overlapping peptides, even with a small number of measurements. These preliminary results suggest that we can analyze more complicated samples and simultaneously reduce the physical measurement time, which are two key advantages in the field of proteomics.

We emphasize that the proposed SLoB framework is also applicable to hyper-spectral imaging. In particular, one can decompose a particular material into its constituent parts, i.e., the left singular vectors would describe the mixture of materials and the right, the spatial locality. Investigating the capabilities of the proposed SLoB framework for hyper-spectral imaging is an interesting open research direction.

References

[1] G. Pope, "Structured sparse signal recovery in general Hilbert spaces," Ph.D. dissertation, D-ITET, ETH Zurich, Zurich, Switzerland, Feb. 2013.

[2] D. C. Liebler, *Proteomics in Cancer Research*. Hoboken, NJ, USA: John Wiley & Sons, Inc, 2005.

[3] S. Hanash and A. Taguchi, "The grand challenge to decipher the cancer proteome," *Nat. Rev. Cancer*, vol. 10, no. 9, pp. 652–60, Sep. 2010.

[4] S. L. et al., "Proteomics of alzheimer's disease: understanding mechanisms and seeking biomarkers," *Expert Rev. Proteomics*, vol. 4, no. 2, pp. 227–38, Apr. 2007.

[5] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198—207, Mar 2003.

[6] M. J. MacCoss and D. E. Matthews, "Quantitative ms for proteomics: teaching a new dog old tricks," *Anal. Chem.*, vol. 77, no. 15, pp. 294A–302A, Aug. 2005.

[7] X. Han, A. Aslanian, and J. R. Yates, "Mass spectrometry for proteomics," *Curr. Opin. Chem. Biol.*, vol. 12, no. 5, pp. 483—90, Oct. 2008.

[8] T. C. Walther and M. Mann, "Mass spectrometry-based proteomics in cell biology," *J. Cell. Biol.*, vol. 190, no. 4, pp. 491–500, Aug. 2010.

[9] L. C. G. et al., "Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis," *Mol. Cell. Proteomics*, Jan. 2012.

[10] H. R. et al., "OpenSWATH: Automated, targeted analysis of mass spectrometric data generated by data-independent acquisition," *Submitted to Nat. Biotech.*, 2012.

[11] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure and Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[13] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[14] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, Jan. 2008.

[15] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Apr. 2009.

[16] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[17] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Sig. Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[18] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. Springer, 2000.

[19] J. Douglas and H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Trans. Amer. Math. Soc.*, no. 82, pp. 421–439, 1956.

[20] F. D. et al., "The PeptideAtlas project," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D655–D658, Jan. 2006.

[21] T. Cai, L. Wany, and G. Xu, "Stable recovery of sparse signals and an oracle inequality," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3516–3522, Jul. 2010.

[22] C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Submitted to App. Comp. Harm. Anal.*, 2011.

[23] J. Meijerink and H. van der Vorst, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric m-matrix," *Math. Comp.*, vol. 31, no. 137, pp. 148–162, 1977.

[24] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Math. Prog.*, vol. 128, no. 1, pp. 321–353, 2011.

[25] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. of 27th Asilomar Conf. on Signals, Systems, and Comput.*, Pacific Grove, CA, USA, Nov 1993, pp. 40–44.

[26] S. G. Mallat, G. Davis, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE J. of Optical Engineering*, vol. 33, pp. 2183–2191, July 1994.

[27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[28] T. Wiberg, "Computation of principal components when data are missing," *Proc. Sec. Symp. Comp. Stat.*, pp. 229–326, 1976.

[29] R. A. Horn and C. Johnson, *Topics in matrix analysis*. New York, NY, USA: Cambridge Univ. Press, 1994.

[30] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix factorization with missing data," in *Proc. of IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, San Diego, CA, USA, June 2005, pp. 316–322.