

RECOVERING SPARSE LOW-RANK BLOCKS IN MASS SPECTROMETRY*

GRAEME POPE[†], CHRISTOPH STUDER[‡], PEDRO NAVARRO[§], AND RICHARD G. BARANIUK[¶]

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 via tandem mass spectrometry (TMS). To efficiently perform SLoB-based peptide identification, we propose two novel recovery algorithms, an exact iterative method and an approximate greedy algorithm, and provide analytical recovery guarantees. Using experiments with synthetic and real-world TMS data, we demonstrate that the proposed framework and algorithms are capable of substantially outperforming existing sparse signal recovery techniques.

Key words. Sparse signal recovery; nuclear norm; convex optimization; recovery guarantees.

1. Introduction. We consider the identification of peptides (fragments of proteins) in biological samples from data collected with a new tandem mass spectrometry (TMS) technique developed at the IMSB, ETH Zurich, Switzerland [6, 7]. The identification of peptides is key for understanding which proteins are present in biological samples. Moreover, since proteins control the processes of the body, their understanding and identification is a fundamental area of research, including (but not limited to) the fight against cancer [10] and Alzheimer’s disease [8]. The measurement process of [6, 7] first fragments the peptides and then, passes the resulting mixture into a mass spectrometer (MS). The mass spectrometer counts the number of particles that have a particular mass-per-charge, or short m/z, over multiple measurement instants. Since the same fragment types can have different charges, a fragment will have a mass spectrum that is non-zero at a number of m/z values and measurement instants. Put simply, given a list of peptides and possible fragments, our challenge is to detect which of these peptides occur in the sample, at what time instants, and with which intensity.

2. System model and recovery problem. To model the MS measurement process of [6, 7], we assume that the m/z 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 m/z interval. Assume that we also have a list of n peptides (and their fragments) that we are interested in detecting. The spectra of many peptides are known and can, e.g., be found in a database such as the PeptideAtlas [5]. Let the i^{th} peptide have 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 $\mathbf{d}_i^{(j)}$, $j = 1, \dots, f_i - 1$, where each vector \mathbf{d}^j is normalized to have unit ℓ_2 norm. Then, form the $m \times f_i$ dictionary $\mathbf{D}_i = [\mathbf{d}_i^{(0)} \dots \mathbf{d}_i^{(f_i-1)}]$, which characterizes the spectrum of the i^{th} peptide and all its fragments. Now, let \mathcal{S}_j be the set of all precursors that are present at measurement instant t_j , so that the observation \mathbf{z}_j at MS measurement instant j is given by

$$\mathbf{z}_j = \sum_{\ell \in \mathcal{S}_j} \mathbf{D}_\ell \mathbf{x}_j[\ell] + \mathbf{n}_j = \sum_{i=1}^n \mathbf{D}_i \mathbf{x}_j[i] + \mathbf{n}_j, \quad (2.1)$$

*The authors would like to thank R. Aebersold, L. Gillet, and G. Rosenberger for making their data available and H. Bölskei for his support. An extended version of this paper will appear in Chapter 4 of the Ph.D. Thesis of G. Pope, “Structured Sparse Signal Recovery in General Hilbert Spaces”, ETH Zurich, Switzerland, Feb. 2013 [12].

[†]Dept. IT & EE, ETH Zurich, Switzerland; e-mail: gpope@nari.ee.ethz.ch

[‡]Dept. ECE, Rice University, Houston, TX; e-mail: studer@rice.edu

[§]Dept. of Biology, ETH Zurich, Switzerland; e-mail: navarro@imsb.biol.ethz.ch

[¶]Dept. ECE, Rice University, Houston, TX; e-mail: richb@rice.edu

where $\mathbf{x}_i[\ell] \in \mathbb{R}^{f_\ell}$ denotes how much of each fragment ion of the i th peptide is present at measurement j . The vector \mathbf{n}_j models additive measurement noise. Since we observe $j = 1, \dots, T$ spectra over multiple measurement instants, we can rewrite (2.1) as

$$\mathbf{Z} = \sum_{i=1}^n \mathbf{D}_i \mathbf{X}_i + \mathbf{N}, \quad (2.2)$$

with $\mathbf{Z} \in \mathbb{R}^{m \times T}$, $\mathbf{X}_i \in \mathbb{R}^{f_i \times T}$ 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 the following question: Given a collection of observations \mathbf{Z} and the dictionary blocks \mathbf{D}_i , how can we accurately and efficiently recover the \mathbf{X}_i , for $i = 1, \dots, n$, which correspond to the peptides that are present?

A straightforward way is to formulate the recovery problem as 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 . That is, we solve

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

where the parameter $\varepsilon > 0$ needs to be chosen larger than the Frobenius norm of the noise.

We emphasize that the (B-MMV) problem makes no assumption about the data in any of the blocks \mathbf{X}_i . However, for real-world measurements, each of these blocks will—at least ideally—be rank one, and so that we can write $\mathbf{X}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ 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 is non-convex, we relax it to the nuclear norm [1, 3], to obtain the following convex sparse low-rank block (SLoB) recovery problem:

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

This SLoB recovery problem will be our key focus in the remainder of the paper.

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

$$\mu_{\mathcal{D}} = \max_{k, \ell: k \neq \ell} \sup_{\mathbf{X} \neq 0} \frac{\|\mathbf{D}_k^* \mathbf{D}_\ell \mathbf{X}\|_*}{\|\mathbf{X}\|_*} = \max_{k, \ell: k \neq \ell} \sigma_{\max}(\mathbf{D}_k^* \mathbf{D}_\ell). \quad (3.1)$$

With this notion of coherence, we can deploy [12, Thm. 2.6] to prove the following theorem stating when (B-MMV) and (N-MMV) perfectly recover the blocks \mathbf{D}_i from the noiseless observations \mathbf{Z} .¹

THEOREM 1 (B-MMV and N-MMV uniqueness). *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 , $i = 1, \dots, n$. If*

$$s < \frac{1}{2} \left(1 + \frac{1}{\mu_{\mathcal{D}}} \right), \quad (3.2)$$

¹Note that by following the approach of [2], the Theorem 1 can be extended to the case of stable recovery of the blocks \mathbf{D}_i with arbitrary (but bounded) noise, given that (3.2) is satisfied.

then the solutions of (B-MMV) and (N-MMV) using $\varepsilon = 0$ are both unique.

We emphasize that the results for synthetic and real-world data shown in Section 5 demonstrate that (N-MMV) significantly outperforms (B-MMV) in most situations. However, there is no dependence on the rank of the individual blocks in (3.2) of Theorem 1. The reason for this is the fact that both optimization problems suffer from the same worst-case signals for which (3.2) is “just violated.” Specifically, one can create particular instances of blocks \mathbf{X}_i that are either full-rank or rank one, which both (B-MMV) and (N-MMV) cannot distinguish (see [12, Sec. 4.3] for the details). In order to obtain rank-dependent recovery conditions, one needs further assumptions on the signals; the corresponding analysis of such conditions is ongoing work. Nevertheless, Theorem 1 provides insight into the mass/charge 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 must be sufficiently incoherent to enable perfect recovery from the MS measurements contained in \mathbf{Z} .

4. Recovery algorithms. In this section, we briefly outline two different 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 an efficient manner.

4.1. Iterative algorithm. The iterative algorithm outlined next relies on the alternating direction method of multipliers (ADMM) [4, 11]; the algorithm in full detail can be found in [12, Sec. 4.4.1]. We start by reformulating (N-MMV) and introduce the auxiliary matrices: $\hat{\mathbf{Y}}_i = \hat{\mathbf{X}}_i$, $i = 1, \dots, n$, and $\hat{\mathbf{W}} = \mathbf{Z} - \sum_{i=1}^n \mathbf{D}_i \hat{\mathbf{X}}_i$. To arrive at an efficient way of solving N-MMV, we form the augmented Lagrangian as follows:

$$\begin{aligned} & \underset{\hat{\mathbf{X}}_i, \hat{\mathbf{Y}}_i, \forall i, \hat{\mathbf{W}}}{\text{minimize}} && \sum_{i=1}^n \|\hat{\mathbf{X}}_i\|_* + \frac{\beta_1}{2} \sum_{i=1}^n \|\hat{\mathbf{X}}_i - \hat{\mathbf{Y}}_i - \boldsymbol{\Lambda}_i\|_F^2 + \frac{\beta_2}{2} \|\hat{\mathbf{W}} - \mathbf{Z} + \sum_{i=1}^n \mathbf{D}_i \hat{\mathbf{Y}}_i - \boldsymbol{\Omega}\|_F^2 \\ & \text{subject to} && \|\hat{\mathbf{W}}\|_F \leq \varepsilon, \end{aligned}$$

which is an instance of the Douglas-Rachford variable splitting method [11]. The terms β_1 and β_2 as well as the matrices $\boldsymbol{\Lambda}_i \in \mathbb{R}^{f_i \times T}$, $i = 1, \dots, n$, and $\boldsymbol{\Omega} \in \mathbb{R}^{m \times T}$ correspond to Lagrange multipliers.

The resulting recovery algorithm consists of two nested loops. In the inner loop, we iteratively minimize the objective function for the matrices $\hat{\mathbf{X}}_i$, $\hat{\mathbf{Y}}_i$, for $i = 1, \dots, n$, and $\hat{\mathbf{W}}$ individually; this requires singular-value shrinkage operations for $\hat{\mathbf{X}}_i$, least-squares procedures for $\hat{\mathbf{Y}}_i$, and a projection onto the Frobenius-norm ball for $\hat{\mathbf{W}}$. After convergence of the inner loop, the Lagrange multipliers $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Omega}$ are updated in the outer loop of the algorithm. The algorithm continues with the inner loop and then again the outer loop, repeating until it converges.

4.2. Greedy algorithm. We now outline a greedy alternative to solving the convex (N-MMV) problem, which relies on the principles of orthogonal matching pursuit (OMP) [13]; the algorithm in full detail can be found in [12, Sec. 4.4.2]. Concretely, the (N-OMP) algorithm operates as follows. Assume that at the beginning of the i^{th} iteration we have a solution $\{\hat{\mathbf{X}}_1^{(i)}, \dots, \hat{\mathbf{X}}_n^{(i)}\}$ with $r_j^{(i)} = \text{rank}(\mathbf{X}_j^{(i)})$ and $i = \sum_{j=1}^n r_j^{(i)}$. Then, we seek to

1. Identify the block \mathbf{X}_ℓ 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 $\{\hat{\mathbf{X}}_1^{(i+1)}, \dots, \hat{\mathbf{X}}_n^{(i+1)}\}$ that solve the following optimization problem:

$$\underset{\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n}{\text{minimize}} \left\| \mathbf{Z} - \sum_{j=1}^n \mathbf{D}_j \tilde{\mathbf{X}}_j \right\|_F \quad \text{subject to} \quad \text{rank}(\tilde{\mathbf{X}}_j) \leq r_j^{(i+1)} \quad \forall j. \quad (4.1)$$

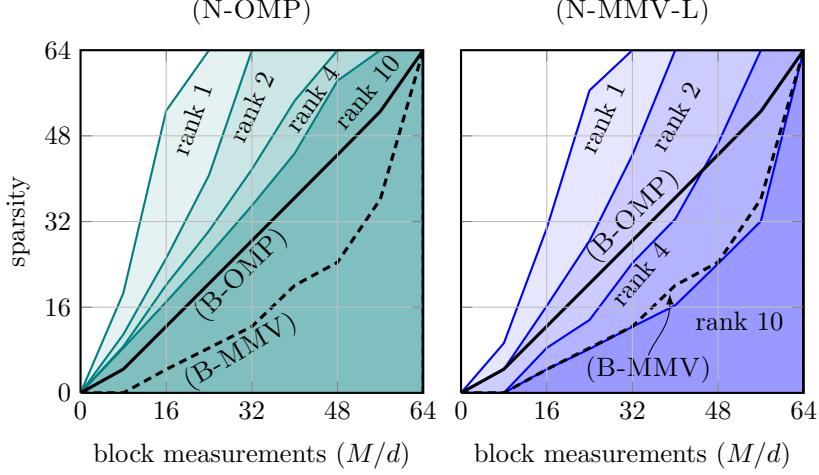


FIGURE 5.1. Empirical phase transition plots showing the region we can recover at least 99% of the signals with Gaussian i.i.d. blocks. Each block \mathbf{X}_i has dimension 10×10 . Exploiting the low-rank structure via (N-OMP) and (N-MMV-L) significantly outperforms the (B-MMV) based algorithms.

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

The first two steps are non-trivial and contrary to first appearances, the solution to (4.1) is not just given by a singular value decomposition. Our own experiments have shown that selecting the dictionary block \mathbf{D}_i most correlated to the residual is an effective strategy.

Our approach to solving (4.1) is based on the Wiberg algorithm [14]. To apply this method to our problem, let $\mathbf{U}_j \Sigma_j \mathbf{V}_j^*$ be the singular value decomposition of $\hat{\mathbf{X}}_j$ (which has rank r_j) and set $\tilde{\mathbf{U}}_j = \mathbf{U}_j \Sigma_j$ and $\tilde{\mathbf{V}}_j = \mathbf{V}_j \Sigma_j$. Then $\sum_{i=1}^n \mathbf{D}_j \mathbf{X}_j = \sum_{i=1}^n \mathbf{D}_j \mathbf{U}_j \Sigma_j \mathbf{V}_j^*$, and furthermore, for each j we have [9, Sec. 4.3]

$$\text{vec}(\mathbf{D}_j \mathbf{X}_j) = \text{vec}(\mathbf{D}_j \mathbf{U}_j \Sigma_j \mathbf{V}_j^*) = (\mathbf{I}_T \otimes \mathbf{D}_j \tilde{\mathbf{U}}_j) \text{vec}(\mathbf{V}_j^*) = (\tilde{\mathbf{V}}_j \otimes \mathbf{D}_j) \text{vec}(\mathbf{U}_j).$$

Therefore, by fixing each of the \mathbf{U}_j and Σ_j , for $j = 1, \dots, n$, we solve the linear least-squares minimization problem

$$\underset{\mathbf{V}_1, \dots, \mathbf{V}_n}{\text{minimize}} \left\| \text{vec}(\mathbf{Z}) - [\mathbf{I}_T \otimes (\mathbf{D}_1 \tilde{\mathbf{U}}_1) \quad \dots \quad \mathbf{I}_T \otimes (\mathbf{D}_n \tilde{\mathbf{U}}_n)] [\text{vec}(\mathbf{V}_1)^T \dots \text{vec}(\mathbf{V}_n)^T]^T \right\|_2, \quad (4.2)$$

to obtain an updated set of right-singular vectors $\{\mathbf{V}_j\}_{j=1}^n$. Note that these problems can be solved efficiently using conjugate gradient methods. Similarly, we can isolate \mathbf{U}_j , $j = 1, \dots, n$, to get a linear least-squares problem in the \mathbf{U}_j . We then alternate between minimizing the set of matrices $\{\mathbf{V}_j\}_{j=1}^n$ and the set of matrices $\{\mathbf{U}_j\}_{j=1}^n$ to arrive at the minimizer of (4.1).

5. Results. We now apply the two SLoB recovery algorithms to synthetic and real data and examine their respective performance for a number of scenarios.

5.1. Synthetic results. We begin by presenting empirical phase transition plots in Figure 5.1. Here we show the regions in which the algorithms are able to recover at least 99% of the tested signals. We generate the i^{th} rank- r (for $r = 1, \dots, 10$) block of size $d_i \times T$ by 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. Gaussian entries. We set $n = 64$, $f_i = 10$

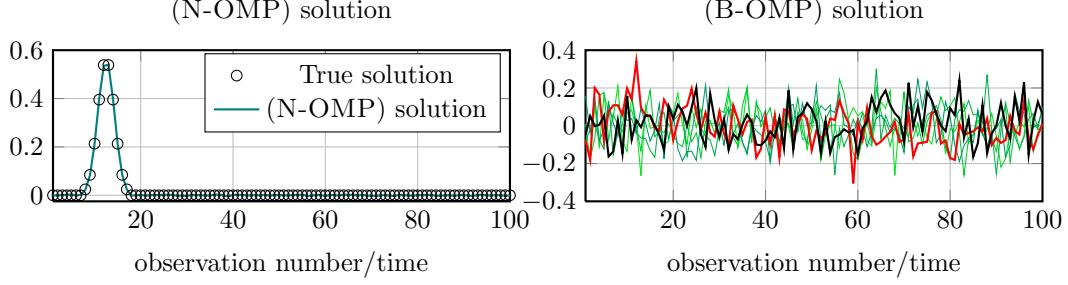


FIGURE 5.2. Right singular vectors for one block. For the (B-OMP) solution, the dominant singular vector is shown in black and the red line is the singular vector (corresponding to the 6th largest singular value) that most resembles the (N-OMP) solution. We see that the rank-aware (N-OMP) method accurately recovers the time/intensity behavior of the true solution, which is in stark contrast to (B-OMP) that ignores the low-rank structure.

(for all i) so that $N = 640$ and we take $T = 10$. We then sweep M from 80 to 640, and vary s , which is the number of non-zero matrices \mathbf{X}_i , from 4 to 64. We refer to M/d as the number of block-measurements. We also compare these results to solving the same problem via (B-OMP) and (B-MMV), which exhibited virtually no rank dependence. More specifically, both of these methods achieved the same recovery performance regardless of the rank of each block.

Figure 5.1 clearly demonstrates that when dealing with low-rank blocks, our nuclear-norm minimization approach is able to recover many more blocks, especially when the rank of the blocks is close to one, than by solving (B-MMV) or using (B-OMP), a greedy variant to find an approximate solution to (B-MMV). Intuitively this makes sense, since fewer parameters are required to specify a rank-one matrix. Hence, we require fewer equations to uniquely specify the solution. Consequently, by exploiting the low-rank structure, we can accurately recover many more non-zero blocks.

5.2. Hybrid real/synthetic experiments. To test our approach, we first create “synthetic experiments” by using a real dictionary and creating an artificial set of observations \mathbf{Z} , by randomly generating the \mathbf{X}_i (see [12, Sec. 4.5.2] for the details). We generated a dictionary from the molecular description of the peptides, with an average of 51 fragments per peptide (the block-length). We then quantized the dictionary and observations uniformly from 200 Th to 1000 Th in steps of 0.025 Th, where Th refers to Thomsons (an m/z measure unit).

Let us take a closer look at a matrix \mathbf{X}_i returned by (N-OMP) and (B-OMP), shown in Figure 5.2. The solution returned by (N-OMP) is rank 1 and we see that the right singular vector, corresponding to the flow rate, matches the true solution. However, the right singular vectors of the (B-OMP) solution bear no resemblance to the true solution. 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 more structure into its solution, which enables us to cope with ill-conditioned blocks and thus, returns a solution that is much closer to the original.

5.3. Experiments with real-world proteomics data. In this section, we test the SLoB framework and recovery methods on actual proteomics data and analyzing three different samples of peptides consisting of 342 known peptides, acquired from the Institute for Molecular Systems Biology at ETH Zurich [6] and measured using the process described in [7] over a period of about 2 hours. Our calculations were performed across $T = 1500$ consecutive time-steps and the dictionary

TABLE 5.1
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%)

from Section 5.2. The results can be seen in Table 5.1 where we give: the number of peptides in the sample, and the number of peptides identified using (B-OMP) and (N-OMP). We clearly see that by exploiting the low-rank structure of the acquired MS measurements we are capable of successfully recovering a significantly higher percentage of the peptides present, i.e., (N-OMP) substantially outperforms (B-OMP) for real-world proteomics data.

6. Conclusion. We 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 significantly fewer time measurements used. This suggests that we can analyze more complicated samples and reduce the physical measurement time, which are key advantages in the field of proteomics. We finally emphasize that the proposed SLoB framework is also applicable hyper-spectral imaging. In particular, one can decompose a 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 SLoBs for hyper-spectral imaging is an interesting research direction.

REFERENCES

- [1] J.-F. CAI, E. J CANDÈS, AND Z. SHEN, *A singular value thresholding algorithm for matrix completion*, J. on Optimization, 20 (2008), pp. 1956–1982.
- [2] T. CAI, L. WANY, AND G. XU, *Stable recovery of sparse signals and an oracle inequality*, IEEE Trans. Inf. Theory, 56 (2010), pp. 3516–3522.
- [3] E. J CANDÈS AND T. TAO, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Trans. Inf. Theory, 56 (2010), pp. 2053–2080.
- [4] J. DOUGLAS AND H. RACHFORD, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., (1956), pp. 421–439.
- [5] F. DESIERE ET AL., *The PeptideAtlas project*, Nucleic Acids Res., 34 (2006), pp. D655–D658.
- [6] H. RÖST ET AL., *OpenSWATH: Automated, targeted analysis of mass spectrometric data generated by data-independent acquisition*, Submitted to Nat. Biotech., (2012).
- [7] L. C. GILLET 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, (2012).
- [8] S. LOVESTONE ET AL., *Proteomics of alzheimer’s disease: understanding mechanisms and seeking biomarkers*, Expert Rev. Proteomics, 4 (2007), pp. 227–38.
- [9] R. A. HORN AND C. JOHNSON, *Matrix analysis*, Cambridge Univ. Press, 1990.
- [10] D. C. LIEBLER, *Proteomics in Cancer Research*, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2005.
- [11] J. NOCEDAL AND S. WRIGHT, *Numerical Optimization*, Springer, 2nd ed., 2000.
- [12] G. POPE, *Structured Sparse Signal Recovery in General Hilbert Spaces*, PhD thesis, D-ITET, ETH Zurich, Zurich, Switzerland, Feb. 2013.
- [13] J. A. TROPP, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inf. Theory, 50 (2004), pp. 2231–2242.
- [14] T. WIBERG, *Computation of principal components when data are missing*, Proc. Sec. Symp. Comp. Stat., (1976), pp. 229–326.