Sparse recovery for Protein Mass Spectrometry data

Martin Slawski and Matthias Hein
Department of Computer Science
Saarland University
Campus E 1.1, Saarbrücken, Germany
{ms,hein}@cs.uni-saarland.de

Abstract

Extraction of peptide masses from a raw protein mass spectrum (MS) is a challenging problem in computational biology. We discuss several structural characteristics of MS data, notably non-negativity and heteroscedastic noise that make standard sparse recovery methods exhibit inferior performance as compared to more targeted approaches. In particular, we suggest non-negative least squares followed by thresholding as a simple, user-friendly alternative, which yields very promising results in practice.

1 Introduction

In recent years, protein mass spectrometry (MS) has become a popular technology in systems biology and clinical research, where it is used, among other things, to discover bio-markers and to enhance the understanding of complex diseases. A central step in the analysis of MS data all subsequent analyses depend on is the extraction of the biologically relevant components (peptides) from the raw spectrum. Peptides emerge as isotopic patterns: the chemical elements serving as building blocks of peptides naturally occur as isotopes differing in the number of neutrons and hence by an integer of atomic mass units, such that a peptide produces a signal at multiple mass positions, which becomes manifest in a series of regularly spaced peaks (see Figure 1). The data are composed of intensity information for a large number n of mass-per-charge (m/z) positions, which is typically in the ten to the hundred thousands. The feature selection problem is to detect those m/z-positions at which a peptide is located and to assign charge states (z) resulting from ionization. In combination, one obtains a list of peptide masses.

Formulation as sparse recovery problem.

The emergence of peptides in the form of isotopic patterns makes mere peak detection a less suitable approach, because it is not clear a priori how to assemble detected peaks according to the pattern they belong to, in particular, when the supports of multiple patterns corresponding to different peptides overlap (see Figure 1, right panel). The fact that the peak heights within an isotopic pattern (for a given mass) and the spacings (for a given charge) can be calculated on the basis of a biochemical model prompts a regression scheme in which the dictionary consists of templates closely matching the shape of isotopic patterns. Each template is a weighted combination of highly localized functions, Gaussians being the default, used as models for a single peak, where the weights (peak heights) are inferred from a well-established model for isotopic abundances (Senko et al. [1995]). Since the composition of the spectrum is unknown in advance, templates are spread over the whole m/z-range, yielding a dictionary of size $p \cdot Z$, where p is in the order of the number n of m/z-positions and Z equals the number of possible charge states, typically $z \in \{1, 2, 3, 4\}$. Representing the raw data by (m/z, intensity)-pairs $\{(x_i, y_i)\}_{i=1}^n$, the underlying model is given by

$$y_i^* = \sum_{z=1}^{Z} \sum_{j=1}^{p} \beta_{z,j}^* \phi_{z,j}(x_i), \quad i = 1, \ldots, n, \quad \Longleftrightarrow \; \boldsymbol{y}^* = \boldsymbol{\Phi}\boldsymbol{\beta}^*. \tag{1}$$

where the $\{y_i^*\}_{i=1}^n$ denote the noise-free counterparts of the $\{y_i\}_{i=1}^n$ and the sum is over templates $\{\phi_{z,j}\}$ and coefficients $\{\beta_{z,j}^*\}$ indexed by charge state (z) and m/z-position (j). The coefficient vector $\boldsymbol{\beta}^*$ is assumed to be sparse, the number of nonzero entries being equal to the number of peptides contained in the spectrum, which is assumed to be considerably smaller than the number of templates in the dictionary. Templates $\phi_{z,j}$ for which $\beta_{z,j}^* = 0$ will be referred to as 'off-support templates'. As the spectrum consists of intensity data, it is non-negative. Therefore, it makes sense to choose the design matrix $\boldsymbol{\Phi}$ such that all of its entries are non-negative and to assume that all entries of $\boldsymbol{\beta}^*$ are non-negative as well.

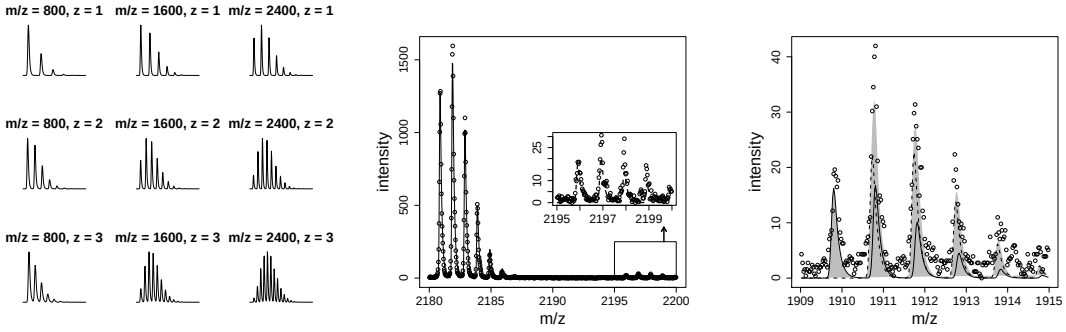

Figure 1: Left panel: Examples of templates $\{\phi_{z,j}\}$ at different m/z positions for different charge states. Middle panel: Example of strongly heterogeneous intensities. Right panel: Example of two overlapping isotopic patterns, indicated by solid and dashed lines; the grey area represents the overall non-negative least squares fit (see Section 3).

2 How to modify the lasso in the presence of heterogeneous noise

Renard et al. [2008] propose to use the non-negative lasso (Tibshirani [1996]) to recover $\boldsymbol{\beta}^*$ in model (1), which is defined as a minimizer of the problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \mathbf{1}^\top \boldsymbol{\beta} \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}, \tag{2}$$

with regularization parameter $\lambda \geq 0$. We will argue that ℓ_1-regularization is subject to severe problems which can be circumvented by a pure fitting approach discussed in Section 3. First of all, the choice of the regularization parameter turns out to be an extremely delicate matter for the problem under consideration because of local differences in noise and intensity levels. An example is given in the middle panel of Figure 1. Consequently, choosing the amount of regularization globally yields poor results. Renard et al. [2008] attack this problem by cutting the spectrum into pieces and fitting each piece separately. While this strategy partially solves the issue, it poses new problems arising from the division of the spectrum. We instead propose to use a more direct adjustment related to the adaptive lasso (Zou [2006]), albeit the motivation is a different one. Given local estimations $\{\widehat{\sigma}_j\}_{j=1}^p$ of the noise level, we minimize the weighted non-negative lasso criterion

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{z=1}^{Z} \sum_{j=1}^{p} \widehat{\sigma}_j \beta_{z,j} \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}. \tag{3}$$

The estimates $\{\widehat{\sigma}_j\}_{j=1}^p$ are obtained as the median of the intensities within a sliding window, whose size constitutes a tuning parameter. The idea is illustrated in Figure 2, which summarizes the result of the following experiment replicated 100 times. For $i = 1, \ldots, n = 5000$, we generate data

$$y_i = 2\phi_1(x_i) + \phi_2(x_i) + 0.5\phi_3(x_i) + \sigma(x_i)\epsilon_i, \quad \{x_i\}_{i=1}^n \text{ equi-spaced in } [1000, 1150],$$

where the isotopic patterns $\{\phi_j\}_{j=1}^3$ are placed at the m/z-positions $\{1025, 1075, 1125\}$ ($z = 1$), $\sigma(x)$ is a piecewise constant function, and the $\{\epsilon_i\}_{i=1}^n$ are drawn i.i.d. from a zero-truncated Gaussian distribution supported on $[0, \infty)$ with standard deviation 0.2. A dictionary is formed by placing 600 templates evenly in the range $[1000, 1150]$. We compute the solution paths (Efron et al. [2004]) of both the non-negative lasso (2) and the weighted non-negative lasso (3), where for simplicity the $\{\widehat{\sigma}_j\}$ are directly obtained from the function σ. Without adjustment, the template ϕ_3 cannot be distinguished from the off-support templates.

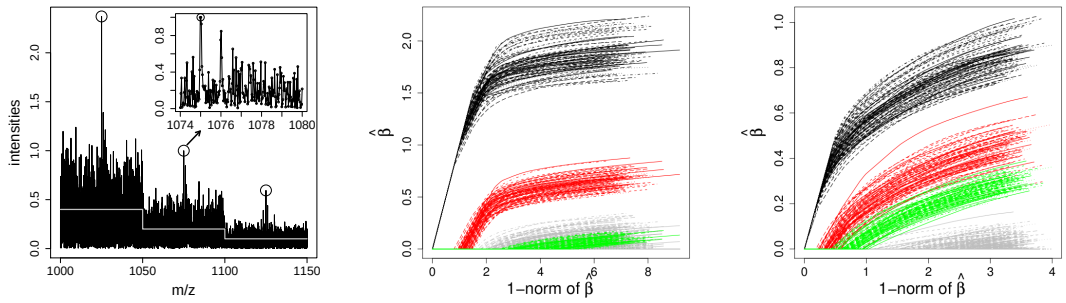

Figure 2: Left panel: Simulated data as described in the text. The circles indicate the positions of the initial peak of the patterns $\{\phi_j\}_{j=1}^3$. The function σ is drawn in grey. Middle panel: Solution paths of the criterion (2). Right panel: Solution paths of the criterion (3). Colours: ϕ_1 = black, ϕ_2 = red, ϕ_3 = green, off-support templates = grey.

3 A pure fitting approach and its advantages

An alternative which has proved to be very successful in practice (Slawski et al. [2010+]) is a pure fitting approach, in which the ℓ_1-regularizer is discarded from (3), and a sparse model is enforced by applying hard thresholding with a threshold depending on an estimate of the local noise level, that is, given a minimizer $\widehat{\boldsymbol{\beta}}^{\text{nnls}}$ of the non-negative least squares criterion

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}, \quad \text{we obtain}$$

$$\widehat{\beta}_{z,j} = I(\widehat{\beta}_{z,j}^{\text{nnls}} > t \cdot \widehat{\sigma}_j)\widehat{\beta}_{z,j}^{\text{nnls}}, \quad z = 1, \ldots, Z, \; j = 1, \ldots, p,$$

where I denotes the indicator function, $t \geq 0$ is the threshold and $\{\widehat{\sigma}_j\}_{j=1}^p$ are, as in the previous section, local estimates of the noise level, computed as medians of the intensities within a sliding window. At first glance, this approach seems to be entirely naive, since in the absence of a regularizer, one would expect over-adaption to the noise, making sparse recovery via subsequent thresholding a hopeless task. This turns out not to be the case, because non-negativity of both $\boldsymbol{\Phi}$ and $\boldsymbol{\beta}$ prevents a common phenomenon one would observe without non-negativity: large positive and negative terms are used to represent a quantity close to zero. Additional background on sparse recovery with non-negativity is given in the Appendix.
The fitting-plus-thresholding approach has several advantages over ℓ_1-regularized fitting.

- With the normalization $\sup_x \phi_{j,z}(x) = 1$ for all j, z, the coefficient $\widehat{\beta}_{z,j}^{\text{nnls}}$ equals the estimated amplitude of the highest peak of the template, such that $\widehat{\beta}_{z,j}^{\text{nnls}}/\widehat{\sigma}_j$ may be interpreted as signal-to-noise ratio and thresholding amounts to discarding all templates whose signal-to-noise ratio falls below a specific value. This makes the parameter choice easier compared to that of a non-intuitive regularization parameter, notably for MS experts.
- The ℓ_∞-normalization of the templates is natural, since it enhances interpretability of the coefficients. The pure fitting approach allows one to choose the most convenient normalization freely, as opposed to regularized fitting where the normalization may cause an implicit preference for specific elements of the dictionary.
- Thresholding is computationally attractive, since it is applied to precisely one non-negative least squares fit. For the ℓ_1-regularized criteria (2) and (3), the entire solution path cannot be computed in a reasonable amount of time: with both n and p in the several ten thousands, an active set algorithm is simply too slow, such that different algorithms in combination with a grid search for λ are required.

In addition, the practical performance is rather encouraging. We present the results obtained on a MALDI spectrum of Myoglobin and compare them to those of ℓ_1-regularization in the variants (2) and (3) as well as to those of orthogonal matching pursuit (OMP, Tropp [2004]) and marginal regression (Genovese et al. [2009]). A manual annotation of the spectrum by an MS expert is used to classify selected templates either as true or false positives, which yields the Precision-Recall curve displayed in Figure 3. Each point in the (Recall, Precision)-plane corresponds to a specific choice of the central tuning parameter, which is specific to the method employed (threshold, regularization parameter, number of iterations (OMP)).

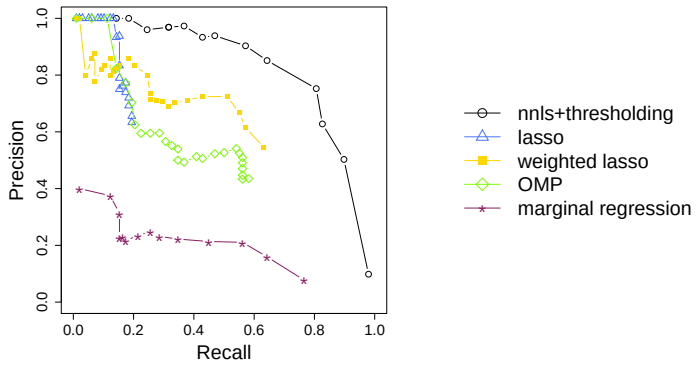

Figure 3: Precision-recall plot for the Myoglobin spectrum.

4 Systematic and random error

In theory, one conventionally assumes that the model is correctly specified – an ideal situation rarely encountered in practice. We discuss the consequences of two common misspecifications of the linear model (1) with regard to sparse recovery. In a second subsection, we consider one specific deviation from the standard additive model for the random error.

Effects of sampling and misspecified templates. Our model formulation (1) assumes that some of the templates of the dictionary are placed at precisely those m/z-positions at which there is actually a peptide in the spectrum. This assumption is not met in practice, since these (unknown) m/z positions live on a continuum, whereas the size of the dictionary is finite. Placing templates at a subset of all positions which have been sampled leads to a phenomenon we refer to as 'peak splitting'. Denoting by x^*, β^* true position and intensity of the peptide, respectively, and by x_l, x_u, $x_l < x^* < x_u$, the two closest m/z-positions of templates in the dictionary, one observes that the corresponding non-negative least squares coefficients $\widehat{\beta}_l, \widehat{\beta}_u$ are both assigned positive values depending on the distances $|x_l - x^*|, |x_u - x^*|$ and β^*. In particular, if $|x_l - x^*| \approx |x_u - x^*|$ is small, the weight β^* is divided into two weights $\widehat{\beta}_l, \widehat{\beta}_u$ of about the same size. As a consequence, *any* sparse recovery method is very likely to select both of x_l, x_u. The situation is mimicked in the top two panels of Figure 4. The top right panels suggests that the lasso (2) is not an answer to the problem, since only a high amount of regularization leading to a poor fit would achieve a selection of only one template.
A second more obvious reason for 'peak splitting' is an incorrect configuration of parameters of the templates controlling width and tailing of the peak patterns. The lower panels of Figure 4 show the consequences of a too low peak width in an idealized setting where the true peptide position is known. Again, ℓ_1-regularization (2) would hardly save the day, because the selection of only one template would underestimate the true intensity at least by a factor of two, as can be seen from the lower right panel. Note that Figure 4 displays noiseless settings.

Additive vs. multiplicative noise. In addition, MS data are contaminated by various kinds of noise arising from sample preparation and the measurement process. Finding a realistic noise model is out of the scope of the paper, yet we would like to discuss an alternative to squared loss. The latter relies on an additive noise model. In view of strong local discrepancies of noise and intensity levels, it might be more adequate to think in terms of relative instead of absolute error. In this direction, we have experimented with a Poisson-like model belonging to the family of generalized linear models (McCullagh and Nelder [1989]). The corresponding loss function reads

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{(\boldsymbol{\Phi\beta})_i - y_i \log((\boldsymbol{\Phi\beta})_i)\}\,, \tag{4}$$

with the convention that $0 \cdot \log(0) = 0$. Recalling that $y_i \geq 0,\ i = 1, \ldots, n$, L is seen to be convex with domain $\{\boldsymbol{\beta} :\ (\boldsymbol{\Phi\beta})_i > 0 \text{ for all } i \text{ with } y_i > 0\}$, which fits well into our non-negativity framework. From the expression one obtains for the gradient of L, one can deduce (McCullagh and Nelder [1989], Chapter 2.2) that the model underlying the loss function

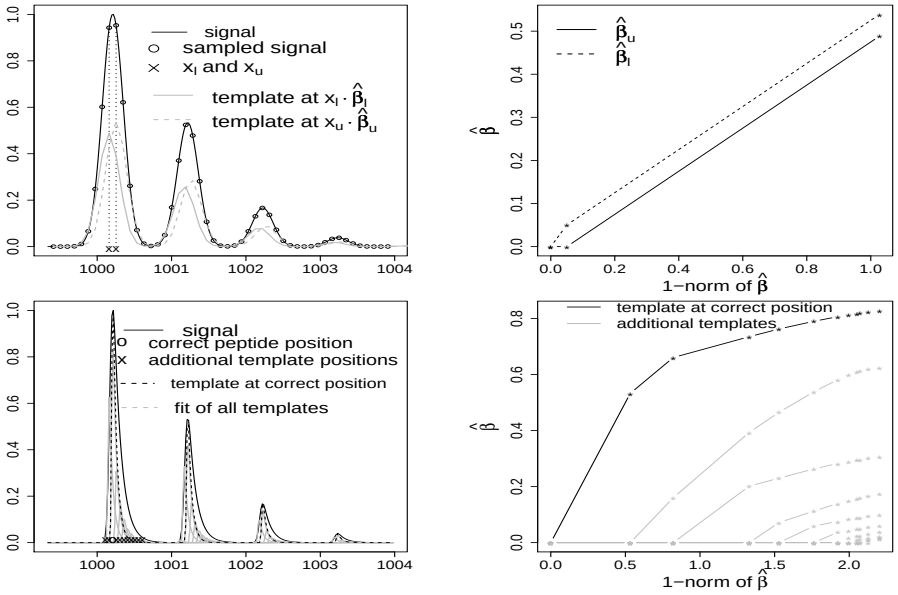

Figure 4: Systematic errors in the template model. Top panels: Consequences of a limited sampling rate. Lower panels: Consequences of an incorrectly specified peak width. The right panels display the corresponding solution paths of the non-negative lasso.

postulates that $\mathbf{E}[y_i|\mathbf{\Phi},\boldsymbol{\beta}] = \mathrm{Var}[y_i|\mathbf{\Phi},\boldsymbol{\beta}] = (\mathbf{\Phi}\boldsymbol{\beta})_i,\ i = 1,\ldots,n$, that is the variance grows linearly with the mean. The influence of a similar error model on the performance of the lasso has recently been studied in Jia et al. [2009]. In that paper, the authors show that sparse recovery fails if the ratio of the maximum to the minimum non-zero entry of the target $\boldsymbol{\beta}^*$ is large in absolute value. In an experiment where this ratio equals 20, we generate an artificial spectrum in which the $\{y_i\}$ result from a combination of two templates and a perturbation by multiplicative noise, that is for $i = 1,\ldots,n = 600$,

$$y_i = (10\phi_1(x_i) + 0.5\phi_2(x_i))(1+\epsilon_i),\quad \{x_i\}_{i=1}^n \text{ equi-spaced in } [2000, 2006],$$

where the $\{\epsilon_i\}_{i=1}^n$ are drawn from a Gaussian distribution with standard deviation 0.3. The data are fitted with a dictionary of templates placed evenly in $[2000, 2006]$ with a spacing of 0.25. The highest peaks of the templates ϕ_1 and ϕ_2 are located at 2002 and 2002.5, respectively. The aim is to find the correct sparse representation by using the fitting-plus-thresholding approach of Section 3, once using non-negative least squares, once the Poisson-like loss (pll) given in (4), where $\widehat{\boldsymbol{\beta}}^{\text{pll}}$ is determined as a minimizer of $L(\boldsymbol{\beta})$ subject to the non-negativity constraint $\boldsymbol{\beta} \geq \mathbf{0}$. A necessary condition for thresholding to succeed is that the coefficients of the noise templates included in the dictionary are smaller than the one of ϕ_2. This may not be accomplished in cases where the inclusion of off-support templates serves to compensate for misfit in ϕ_1 arising from noise as shown in Figure 5. Table 1 suggests that the Poisson-like loss is preferable in this regard. For the real world Myoglobin spectrum, we do not observe any improvement (Figure 6).

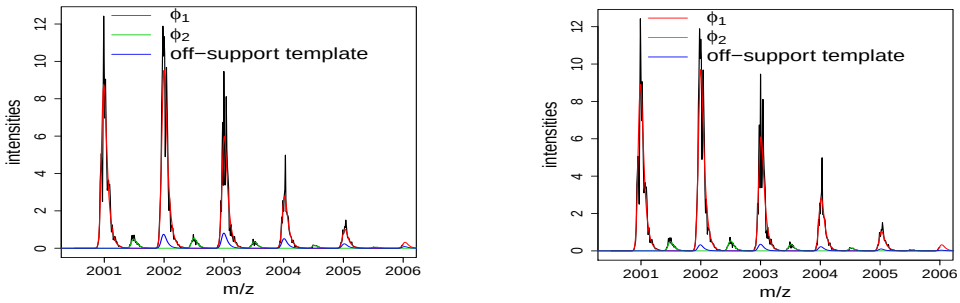

Figure 5: An instance of the experiment described in the text. Left panel: Fit of non-negative least squares. Right panel: Fit of the Poisson-like loss. In the left panel, the coefficient of the noise template exceeds that of ϕ_2 such that sparse recovery via thresholding is not possible.

	$\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$	$\|\widehat{\boldsymbol{\beta}}_{S^c}\|_\infty$	$I(\|\widehat{\boldsymbol{\beta}}_{S^c}\|_\infty > \widehat{\beta}_2)$
nnls	0.36 (0.04)	0.33 (0.04)	0.26 (0.04)
pll	0.17 (0.02)	0.15 (0.02)	0.10 (0.03)

Table 1: Results of the experiment described in the text. We denote by $\widehat{\boldsymbol{\beta}}_{S^c}$ the coefficient vector of the off-support templates. Displayed are averages over 100 iterations, with standard errors in parentheses. The right column indicates that sparse recovery fails in a considerably higher fraction of cases when squared loss is used.

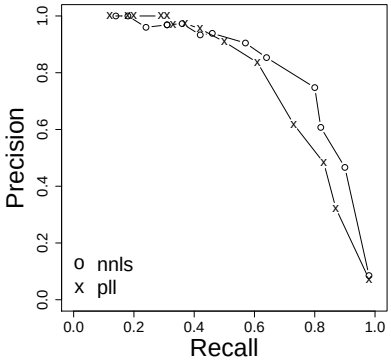

Figure 6: Performance of the two loss functions in conjunction with the fitting-plus-thresholding approach for the Myoglobin spectrum. The precision-recall curve for non-negative least squares is identical to that in Figure 3.

A Sparse recovery by non-negativity constraints

A series of recent papers (Bruckstein et al. [2008], Wang and Tang [2009], Donoho and Tanner [2010], Wang et al. [2010]) have studied uniqueness of non-negative solutions of undetermined linear systems of equations

$$\boldsymbol{\Phi}\boldsymbol{\beta} = \boldsymbol{y} \text{ subject to } \boldsymbol{\beta} \geq \mathbf{0} \tag{5}$$

given the existence of a sparse solution $\boldsymbol{\beta}^*$ with support set $S = \{j : \beta_j^* > 0\}$ of cardinality s. Its complement is denoted by S^c. The sub-matrices of the $n \times p$ matrix $\boldsymbol{\Phi}$ obtained by extracting columns associated with S and S^c are denoted by $\boldsymbol{\Phi}_S$ and $\boldsymbol{\Phi}_{S^c}$, respectively, and $\boldsymbol{\Phi}_j$ denotes the j-th column of $\boldsymbol{\Phi}$. Likewise, the subscripts S or S^c attached to a vector denote the sub-vectors with components in S or S^c, respectively. In geometrical terms, the condition for uniqueness is given in the following statement.

Proposition 1. *If $\boldsymbol{\Phi}_S\mathbb{R}_+^s$ is a face of $\boldsymbol{\Phi}\mathbb{R}_+^p$ and the columns of $\boldsymbol{\Phi}$ are in general position in $\mathbb{R}^n$, then the constrained linear system* (5) *has $\boldsymbol{\beta}^*$ as its unique solution.*

Proof. By definition, since $\boldsymbol{\Phi}_S\mathbb{R}_+^s$ is a face of $\boldsymbol{\Phi}\mathbb{R}_+^p$, there is a hyperplane separating $\boldsymbol{\Phi}_S\mathbb{R}_+^s$ from $\boldsymbol{\Phi}_{S^c}\mathbb{R}_+^{p-s}$, i.e. there exists a $\boldsymbol{w} \in \mathbb{R}^n$ such that $\langle \boldsymbol{\Phi}_j, \boldsymbol{w}\rangle = 0,\ j \in S,\ \ \langle \boldsymbol{\Phi}_j, \boldsymbol{w}\rangle > 0,\ j \in S^c$. Assume that there is a second solution $\boldsymbol{\beta}^* + \boldsymbol{\delta},\ \boldsymbol{\delta} \neq \mathbf{0}$. Expand $\boldsymbol{\Phi}_S(\boldsymbol{\beta}_S^* + \boldsymbol{\delta}_S) + \boldsymbol{\Phi}_{S^c}\boldsymbol{\delta}_{S^c} = \boldsymbol{y}$. Multiplying both sides by $\boldsymbol{w}^\top$ yields $\sum_{j\in S^c} \langle \boldsymbol{\Phi}_j, \boldsymbol{w}\rangle\, \delta_j = 0$. Since $\boldsymbol{\beta}_{S^c}^* = \mathbf{0}$, feasibility requires $\delta_j \geq 0,\ j \in S^c$. All inner products within the sum are positive, concluding that $\boldsymbol{\delta}_{S^c} = \mathbf{0}$. General position implies $\boldsymbol{\delta}_S = \mathbf{0}$. □

This statement suggest that there are situations where sparse recovery is possible by enforcing non-negativity. In fact, Donoho and Tanner [2010] (Corollary 4.1, Theorem 4.1) give explicit examples of $\boldsymbol{\Phi}$ allowing for sparse recovery for a support size s proportional to p. First results for a noisy setup are derived in Slawski and Hein [2010+].

References

A. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54: 4813–4820, 2008.

D. Donoho and J. Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete and Computational Geometry*, 43:522–541, 2010.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression (with discussion). *The Annals of Statistics*, 32:407–499, 2004.

C. Genovese, J. Jin, and L. Wasserman. Revisiting Marginal Regression. Technical report, Carnegie Mellon University, 2009.

J. Jia, K. Rohe, and B. Yu. The Lasso under heteroscedasticity. Technical report, University of California at Berkeley, 2009.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

B. Renard, M. Kirchner, H. Steen, J. Steen, and F. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9:355, 2008.

M. Senko, S. Beu, and F. McLafferty. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *Journal of the American Society for Mass Spectrometry*, 6:229–233, 1995.

M. Slawski and M. Hein. Positivity vs. Sparsity − Feature Selection for non-negative data. In preparation, 2010+.

M. Slawski, R. Hussong, A. Tholey, A. Hildebrandt, and M. Hein. Peak pattern deconvolution for Protein Mass Spectrometry by Non-Negative Least Squares/Least Absolute Deviation template matching. In preparation, 2010+.

R. Tibshirani. Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:671–686, 1996.

J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004.

M. Wang and A. Tang. Conditions for a Unique Non-negative Solution to an Underdetermined System. In *Proceedings of Allerton Conference on Communication, Control, and Computing*, 2009.

M. Wang, W. Xu, and A. Tang. A unique nonnegative solution to an undetermined system: from vectors to matrices. Technical report, Cornell University, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.