# 1 Sparse recovery for protein mass spectrometry data

**Martin Slawski** ms@cs.uni-saarland.de
*Saarland University*
*Saarbrücken, Germany*

**Matthias Hein** hein@cs.uni-saarland.de
*Saarland University*
*Saarbrücken, Germany*

*Extraction of peptide masses from a raw protein mass spectrum (MS) is a challenging problem in computational biology, which can be recast as sparse recovery problem. We discuss modifications of standard sparse recovery methods that accomodate non-negativity and heteroscedastic noise, which are characteristic of MS data.*
*The non-negativity constraints are found to be extremely powerful, since an approach combining non-negative least squares fitting and thresholding is shown to outperform competing methods that explicitly promote sparsity via some form of regularization.*
*By means of examples taken from the given data, we discuss that assumptions such as absence of model mis-specifications and an upper bound on the coherence of the dictionary typically made within the usual sparse recovery framework are not met. We show that the resulting gap between theory and practice can be bridged by a suitable post-processing procedure.*

## 1.1 Introduction

In the next two paragraphs, the reader is introduced to the underlying practical problem and its formulation within a sparse recovery framework.

**Background.** 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 pre-processing of MS data all subsequent analyses, like e.g. sample classification, 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 (approximately) 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 1.1). The data are composed of intensities observed for a large number 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.** On a high level, the problem amounts to deconvolution, where, using a representation on a continuous domain, the underlying signal composed of $s$ isotopic patterns is given by

$$y^*(x) = \sum_{k=1}^{s} b_k(\psi \star \iota)(x-\mu_k^*), \quad \iota(x-\mu_k^*) = \sum_{l\in\mathbb{Z}} \alpha_l(\mu_k^*; z_k)\delta\left(x - \mu_k^* - \frac{l}{z_k}\right), \tag{1.1}$$

where $x$ takes values within some specific interval of $m/z$-values, the $\{b_k\}_{k=1}^s$ are positive weights (amplitudes) and $\psi$ is a fixed localized function modeling a smeared peak (the default being a Gaussian), which is convolved with the function $\iota$. The latter represents an isotopic pattern which is modeled as a positive combination of Diracs centered at $m/z$-positions $\{\mu_k^* + \frac{l}{z_k}\}$, where the weights $\{\alpha_l(\mu_k^*; z_k)\}_{l\in\mathbb{Z}}$ are computed according to a well-established model for isotopic abundances (Senko et al., 1995) given the position $\mu_k^*$ of the leading peak (i.e. $\alpha_0(\mu_k^*; z_k) \geq \alpha_l(\mu_k^*; z_k), l \neq 0$) and the charge $z_k$. In terms of model (1.1), the task to be performed is to find the positions $\{\mu_k^*\}_{k=1}^s$ and the corresponding charges $\{z_k\}_{k=1}^s$ as well as the amplitudes $\{b_k\}_{k=1}^s$. For 'benign' spectra, the problem can be solved easily in two steps. First, one detects all peaks $\{\delta(x-\mu_k^*-\frac{l}{z_k})\}$ of a significantly high amplitude ($\alpha_l(\mu_k^*; z_k)$ decays rapidly with $|l|$). Second, nearby peaks are merged to form groups, each group representing an isotopic pattern. The charges $\{z_k\}$ can be inferred from the spacings of the peaks within the same group. For more complicated spectra, this approach is little suitable. When the supports

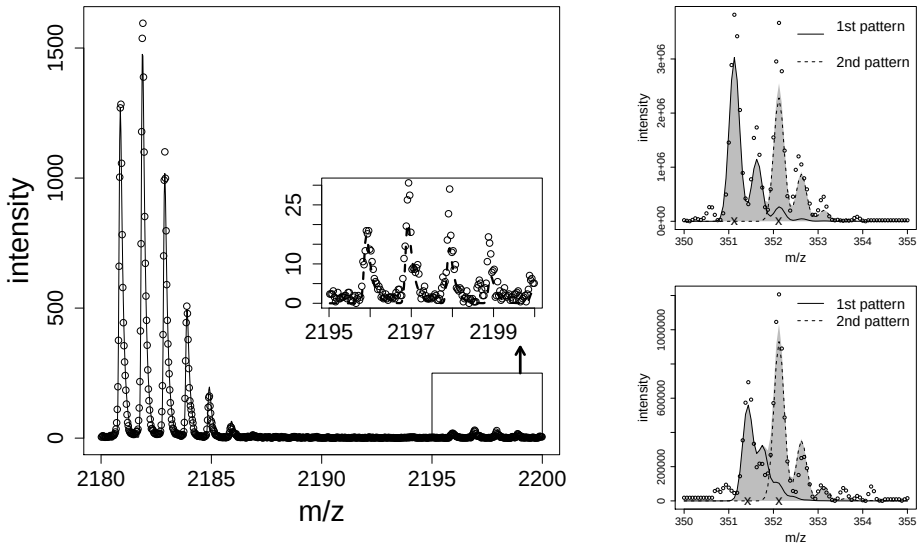


**Figure 1.1**: Left panel: Two isotopic patterns whose intensities differ drastically. Right panel: Two instances of overlapping isotopic patterns.

of multiple patterns corresponding to different peptides overlap (see the right panel of Figure 1.1), peaks are likely to be overlooked in the first step because of the function $\psi$ smearing the peaks out. But even if that does not happen, one cannot hope to correctly assemble detected peaks according to the pattern they belong to in the second step, since nearby peaks may belong to different patterns. Approaches based on *template matching* (see Figure 1.2 for an illustration) circumvent these evident shortcomings by directly tackling the problem at the level of isotope patterns. In essence, template matching involves a sparse regression scheme in which the dictionary consists of templates matching the shape of isotope patterns, exploiting that, as mentioned above, the amplitudes $\{\alpha_l\}$ are known given location and charge. Since the composition of the spectrum is unknown in advance, templates are placed at positions $\{\mu_j\}_{j=1}^{p}$ covering the whole $m/z$-range. This yields 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 $z \in \{1, 2, 3, 4\}$. It then remains to select a small subset of the templates yielding a good fit to the given data. More specifically, after sampling (1.1) at $m/z$-positions $\{x_i\}_{i=1}^{n}$, obtaining intensities $y_i^* = y^*(x_i),\ i = 1, \ldots, n$, the

following model is assumed.

$$y_i^* = \sum_{z=1}^{Z} \sum_{j=1}^{p} \beta_{z,j}^* \phi_{z,j}(x_i), \quad i = 1, \ldots, n, \quad \Longleftrightarrow \; \boldsymbol{y}^* = \boldsymbol{\Phi}\boldsymbol{\beta}^*, \tag{1.2}$$

where

$$\phi_{z,j}(x) = \sum_{l \in \mathbb{Z}} \alpha(z; \mu_j)(\psi \star \delta)\left(x - \mu_j + \frac{l}{z}\right)$$

are the templates. The coefficient vector $\boldsymbol{\beta}^*$ is related to the $\{b_k\}_{k=1}^s$ in (1.1) in the sense that $\beta_{z,j}^* = b_k$ if $\phi_{z,j}(\cdot) = \iota(\cdot - \mu_k^*)$ and $\beta_{z,j}^* = 0$ otherwise. Since one uses much more templates in (1.2) than there are corresponding isotopic patterns in the spectrum, $\boldsymbol{\beta}^*$ is sparse.

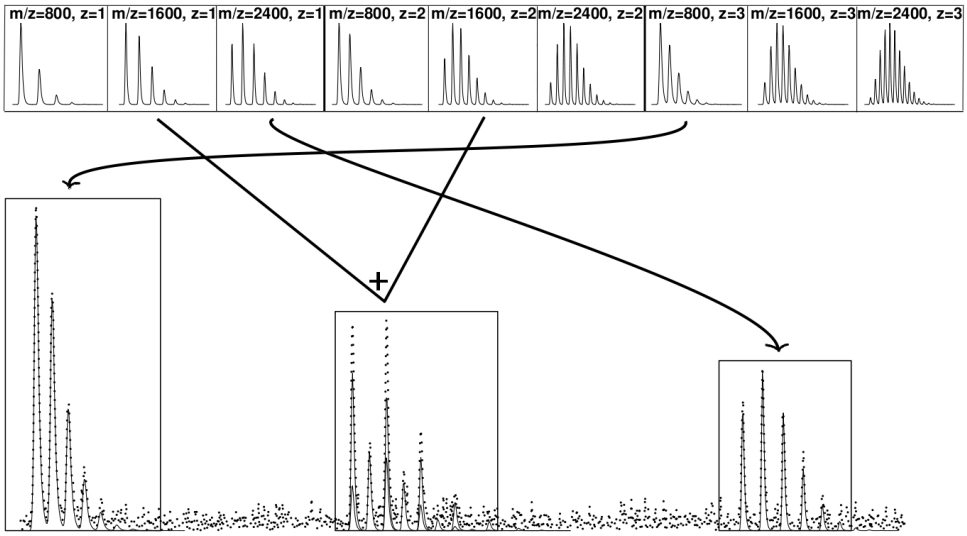


**Figure 1.2**: Illustration of template matching. The boxes in the top part of the figure contain nine templates $\{\phi_{z,j}\}$ whose shape varies in dependency of mass-over-charge ($m/z$) and charge ($z$). The bottom part of the Figure depicts a toy spectrum generated by combining four different templates and adding a small amount of random noise. The arrows indicate how the templates are matched to their counterparts in the spectrum. The signal in the middle is an overlap of two patterns which are accordingly fitted by a combination of templates, which is indicated by a '+'.

In practice, one does not observe $\{y_i^*\}_{i=1}^n$, but instead noisy versions $\{y_i\}_{i=1}^n$. This makes template matching, i.e. finding the support of $\boldsymbol{\beta}^*$, a highly non-trivial task even in the case where $n > p \cdot Z$, because noise can be fitted

by templates whose coefficient in (1.2) is in fact zero (in the sequel, these templates are referred to as 'off-support templates'). Consequently, one has to find a suitable compromise between data fidelity and model complexity as quantified by the number of templates one assigns a coefficient different from zero. According to the paradigm established in recent years, solving the problem by regularized regression with a sparsity-promoting term appears to be a natural approach. One might also think of greedy approximation schemes, where templates are successively added until the fit cannot be significantly improved. For the latter, regularization is performed implicitly.

**Outline.** As discussed in the following chapters, heteroscedastic noise (cf. left panel of Figure 1.1) has to be accommodated. Consequently, modifications of standard algorithms are indispensable. The non-negativity constraint on $\boldsymbol{\beta}^*$ turns out to be extremely powerful. In Section 1.3, we describe an approach combining non-negative least squares and thresholding, which yields excellent results in practice, outperforming competing methods employing regularization. Various modelling issues are discussed in Section 1.4. In particular, the problem of model mis-specifications casts serious doubts on the usefulness of the sparse recovery framework used in theory for the given practical application.

**Notation.** For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, $\boldsymbol{A}_J$ denotes the matrix one obtains by extracting the columns corresponding to an index set $J$. For $j = 1, \ldots, m$, $\boldsymbol{A}_j$ denotes the $j$-th column of $\boldsymbol{A}$. Likewise, for $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{v}_J$ is the sub-vector corresponding to $J$. Its complement is denoted by $J^c$. The notation $\boldsymbol{v} \geq \boldsymbol{0}$ means that all components of $\boldsymbol{v}$ are non-negative.

## 1.2 Adapting sparse recovery methods to non-negativity and heteroscedasticity

Dealing with strong heteroscedasticity is fundamental to a successful analysis of MS data. What happens if heteroscedasticity is ignored can be well understood from Figure 1.3 below. Both signals emerge in different $m/z$-regions of the same spectrum, and both are equally well distinguishable from noise around them. Inspecting the horizontal axes in the two plots, the signal achieves an intensity of around 150, whereas the noise achieves intensities as large as 40 in the left plot. In the right panel, we have intensities of roughly 30 for the signal and less than 10 for the noise. When applying a template matching scheme, this has the consequence that templates just fitting noise in the left panel are assigned larger coefficients than the template matching the signal on the right panel. If, as usually,

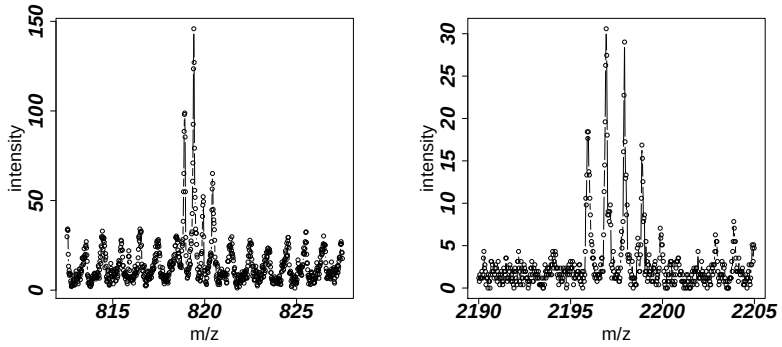


**Figure 1.3**: Heteroscedasticity in MS data. The two panels display two patterns occuring in different $m/z$-regions of the same spectrum. Note the different scalings of the vertical axis.

the selection of templates is based on the size of their coefficients, this has the effect that over-selection is necessary to include the signal of lower intensity. We conclude that absolute signal strength is not meaningful for the data under consideration. Instead, a quantification relative to the local noise level is more appropriate. In the remainder of this section, we have a closer look at two popular sparse recovery methods for which we suggest modifications that take heteroscedasticity into account. The positive effect of these modifications is demonstrated experimentally.

**Adapting the lasso.** In conjunction with a template matching approach similar to our description in Section 1.1, Renard et al. (2008) propose to use the lasso (Tibshirani, 1996) with non-negativity constraints to recover $\boldsymbol{\beta}^*$ in model (1.2). The non-negative lasso is defined as a minimizer of the problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \mathbf{1}^\top \boldsymbol{\beta} \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}, \tag{1.3}$$

with regularization parameter $\lambda \geq 0$. In view of strong local differences in noise and intensity levels, 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 $\widehat{\sigma}_j = \widehat{\sigma}(\mu_j),\ j = 1, \ldots, p$, of the noise level for the $m/z$-positions $\{\mu_j\}_{j=1}^p$ at which a template is placed, we minimize the

weighted non-negative lasso criterion

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{z=1}^{Z} \sum_{j=1}^{p} \widehat{\sigma}_j \beta_{z,j} \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}. \tag{1.4}$$

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. Needless to say, one might employ more sophisticated techniques to obtain these estimates. By making the amount of regularization proportional to the noise level in a component-specific way, we aim at preventing over-selection in high-noise regions and ensuring detection of small signals in low-noise regions. The modification can be employed in connection with any sparsity-promoting regularizer in a generic way.

**Adapting orthogonal matching pursuit.** Orthogonal matching pursuit (OMP, Algorithm 1.1) generates a sparse approximation in a greedy way. Its properties are analyzed, among others, in Tropp (2004) and Zhang (2009). The rather close connection between $\ell_1$-regularization (1.3) and OMP is unveiled in Efron et al. (2004).

**Algorithm 1.1** Orthogonal matching pursuit (OMP)

**Input:** $\boldsymbol{\Phi}$, $\boldsymbol{y}$, tolerance $\varepsilon \geq 0$, positive integer $s \leq \min\{n, p \cdot Z\}$.
  $\mathcal{A} \leftarrow \emptyset$, $\boldsymbol{r} \leftarrow \boldsymbol{y}$, $\widehat{\boldsymbol{\beta}} \leftarrow 0$.
  **while** $\|\boldsymbol{\Phi}^\top \boldsymbol{r}\|_\infty > \varepsilon$ **and** $|\mathcal{A}| < s$ **do**
    $\widehat{j} \leftarrow \operatorname{argmax}_{j \in \mathcal{A}^c} |\boldsymbol{\Phi}_j^\top \boldsymbol{r}|$, $\mathcal{A} \leftarrow \mathcal{A} \cup \{\widehat{j}\}$.
    $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} \leftarrow (\boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{\Phi}_{\mathcal{A}})^{-1} \boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{y}$.
    $\boldsymbol{r} \leftarrow \boldsymbol{y} - \boldsymbol{\Phi}_{\mathcal{A}} \widehat{\boldsymbol{\beta}}_{\mathcal{A}}$.
  **end while**
  **return** $\widehat{\boldsymbol{\beta}}$

Algorithm 1.1 is not a suitable answer to the template matching problem for the aforementioned reasons. We here present a modification of OMP that integrates both heteroscedasticity and non-negativity of $\boldsymbol{\beta}^*$. As for the lasso in the preceding paragraph, we assume that we are given estimates of the local noise levels $\{\widehat{\sigma}_j\}_{j=1}^p$. Comparing Algorithms 1.1 and 1.2, there are two major differences. First, the active set $\mathcal{A}$ is augmented by the index which maximizes $\boldsymbol{\Phi}_j^\top \boldsymbol{r}/\widehat{\sigma}_j$ instead of $|\boldsymbol{\Phi}_j^\top \boldsymbol{r}|$. The division by $\widehat{\sigma}_j$ integrates heteroscedasticity by preventing off-support templates in high noise regions from being included into $\mathcal{A}$. The absolute value is omitted because of the non-negativity constraint imposed on $\widehat{\boldsymbol{\beta}}$: it is not hard to verify that after $\widehat{j}$ has been included into $\mathcal{A}$, the corresponding sign of the corresponding least

**Algorithm 1.2** Weighted non-negative orthogonal matching pursuit

**Input:** $\boldsymbol{\Phi}$, $\boldsymbol{y}$, tolerance $\varepsilon \geq 0$, positive integer $s \leq \min\{n, p \cdot Z\}$.
  $\mathcal{A} \leftarrow \emptyset$, $\boldsymbol{r} \leftarrow \boldsymbol{y}$, $\widehat{\boldsymbol{\beta}} \leftarrow 0$.
  **while** $\max_j \boldsymbol{\Phi}_j^\top \boldsymbol{r} / \widehat{\sigma}_j > \varepsilon$ **and** $|\mathcal{A}| < s$ **do**
    $\widehat{j} \leftarrow \operatorname{argmax}_{j \in \mathcal{A}^c} \boldsymbol{\Phi}_j^\top \boldsymbol{r}$, $\mathcal{A} \leftarrow \mathcal{A} \cup \{\widehat{j}\}$.
    $\widetilde{\boldsymbol{\beta}} \leftarrow \widehat{\boldsymbol{\beta}}$
    $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} \leftarrow (\boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{\Phi}_{\mathcal{A}})^{-1} \boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{y}$.
    *% Backward loop*
    **while** $\exists j : \widehat{\beta}_j < 0$ **do**
      Set $\alpha_j \leftarrow \widetilde{\beta}_j / (\widetilde{\beta}_j - \widehat{\beta}_j)$ if $\widetilde{\beta}_j > 0$ and $\alpha_j \leftarrow 0$ otherwise, $j = 1, \ldots, p \cdot Z$.
      $j^* \leftarrow \operatorname{argmin}\{j : \alpha_j > 0\}$, $\widehat{\alpha} \leftarrow \alpha_{j^*}$.
      $\widetilde{\boldsymbol{\beta}} \leftarrow \widetilde{\boldsymbol{\beta}} + \widehat{\alpha}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})$
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j^*\}$
      $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} \leftarrow (\boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{\Phi}_{\mathcal{A}})^{-1} \boldsymbol{\Phi}_{\mathcal{A}}^\top \boldsymbol{y}$
    **end while**
    *% End of backward loop*
    $\boldsymbol{r} \leftarrow \boldsymbol{y} - \boldsymbol{\Phi}_{\mathcal{A}} \widehat{\boldsymbol{\beta}}_{\mathcal{A}}$.
  **end while**
  **return** $\widehat{\boldsymbol{\beta}}$

squares coefficient $\widehat{\beta}_{\widehat{j}}$ equals the sign of $\boldsymbol{\Phi}_{\widehat{j}}^\top \boldsymbol{r}$. While $\widehat{\beta}_{\widehat{j}}$ is guaranteed to be feasible, this is not necessarily the case for the whole sub-vector $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$. If $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ fails to be feasible, a backward loop is entered whose construction is adopted from the Lawson-Hanson active set algorithm (Lawson and Hanson, 1987) for solving the non-negative least squares problem (cf. (1.6) below). In fact, the proposed Algorithm 1.2 coincides with the Lawson-Hanson algorithm if the $\{\widehat{\sigma}_j\}_{j=1}^p$ are constant, $\varepsilon = 0$ and $s = \min\{n, p \cdot Z\}$. The backward loop can be understood as follows. Given a current iterate $\widetilde{\boldsymbol{\beta}}$, one performs an update of the form $\widetilde{\boldsymbol{\beta}} + \alpha(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})$, where $\alpha \in (0, 1]$ is a step size. A step size of $\alpha = 1$ corresponds to the least squares solution restricted to the active set. Since the latter may not be feasible, one proceeds into the direction $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$ until one of the coefficients of the active set drops to zero. The procedure is repeated with a reduced active set. The possibility of backward steps allows the algorithm to correct itself by dropping elements that have been included into the active set at previous iterations. This is unlike the standard OMP, which is a pure forward selection scheme.

**Illustration.** To demonstrate that the set of modifications can yield a drastic improvement, we present the result of an experiment, where we generate random artificial spectra of the form

$$y_i = 2\phi_1(x_i) + \phi_2(x_i) + 0.5\phi_3(x_i) + \sigma(x_i)\epsilon_i, \tag{1.5}$$

where the sampling points $\{x_i\}_{i=1}^n$, $n = 5000$, are placed evenly along the $m/z$-range $[1000, 1150]$. The functions $\{\phi_j\}_{j=1}^3$ represent isotopic patterns of charge $z = 1$ placed at the $m/z$-positions $\{1025, 1075, 1125\}$. The random variables $\{\epsilon_i\}_{i=1}^n$ constitute an additive error component. They are drawn i.i.d. from a truncated Gaussian distribution supported on $[0, \infty)$ with standard deviation 0.2. Heteroscedasticity is induced by the positive function $\sigma(x)$ which is constant on the sub-intervals $[1000, 1050)$, $[1050, 1100)$, $[1100, 1150]$. Figure 1.4 displays one instance of such a spectrum. The aim is to recover $\{\phi_j\}_{j=1}^3$ from a dictionary of 600 templates placed evenly in the range $[1000, 1150]$, that is to find the support of $\boldsymbol{\beta}^*$ after re-writing (1.5) as

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\beta}^* + \boldsymbol{\xi} = [\boldsymbol{\Phi}_1\boldsymbol{\Phi}_2\boldsymbol{\Phi}_3 \;\; \boldsymbol{\Phi}_4 \ldots \boldsymbol{\Phi}_{600}] \, [\; 2 \quad 1 \quad 0.5 \quad 0 \quad \ldots \quad 0 \;]^\top + \boldsymbol{\xi},$$
$$\boldsymbol{y} = (y_i), \;\; \boldsymbol{\Phi}_j = (\phi_j(x_i)), \; j = 1, \ldots, p, \;\; \boldsymbol{\xi} = (\sigma(x_i)\epsilon_i).$$

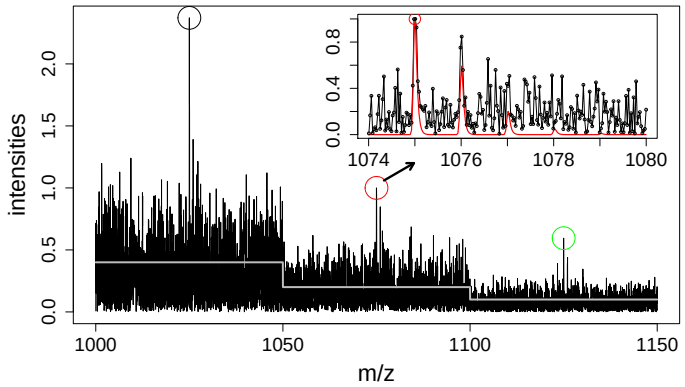


**Figure 1.4**: An artificial mass spectrum generated randomly according to (1.5). The coloured circles indicate the positions of the initial peak of the patterns ($\phi_1$ = black, $\phi_2$ = red, $\phi_3$ = green). The function $\sigma$ is drawn in grey.

By construction, $\phi_j$ is centered at the $j$-th sub-interval on which $\sigma$ is constant, $j = 1, \ldots, 3$, while the amplitudes $\{2, 1, 0.5\}$ have been chosen such that the corresponding signal-to-noise ratios are equal. We generate 100 random spectra from (1.5). For each instance, we compute the solution paths (Efron et al. (2004)) of both the non-negative lasso (1.3) and its weighted counterpart (1.4) as well as all intermediate solutions of OMP and its modification given by Algorithm 1.2. For simplicity the $\{\widehat{\sigma}_j\}$ are obtained by evaluating the function $\sigma$. The results of the experiments displayed in Figure 1.5 show unambiguously that $\phi_3$ cannot be distinguished from the off-support templates $\phi_4, \ldots, \phi_{600}$ if heteroscedastic noise is ignored. The proposed modifications turn out to be an effective means to counteract that

problem, since on the right halves of the plots, $\phi_3$ clearly stands out from the noise.

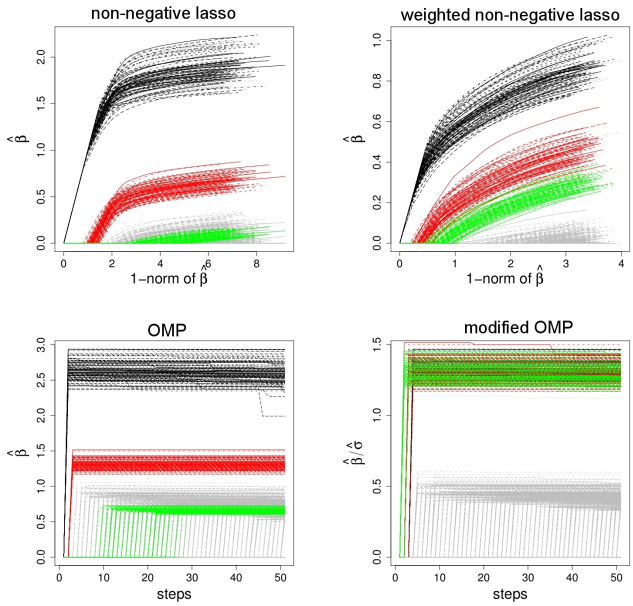


**Figure 1.5**: Upper panel: Solution paths of the non-negative lasso (1.3) (left), solution paths of the weighted non-negative lasso (1.4) (right). Lower panel: Output of OMP (Algorithm 1.1, left) and output of its modification (Algorithm 1.2, right) after running the (outer) while loop *steps* times, where *steps* ranges from 1 to 50. Note that for Algorithm 1.2 (right), $\{\widehat{\beta}_j/\widehat{\sigma}_j\}$ is on the vertical axis. Colours: $\phi_1$ = black, $\phi_2$ = red, $\phi_3$ = green, off-support templates $\phi_4, \ldots, \phi_{600}$= grey.

## 1.3 A pure fitting approach and its advantages

An alternative to conventional sparse approximation schemes as discussed in the preceding section is a pure fitting approach applied with great success in Slawski et al. (2012), in which the $\ell_1$-regularizer is discarded from (1.4), and a sparse model is enforced by subsequently applying hard thresholding with a threshold depending on an estimate of the local noise level, i.e. given a minimizer $\widehat{\boldsymbol{\beta}}$ of the non-negative least squares criterion

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 \quad \text{subject to } \boldsymbol{\beta} \geq \mathbf{0}, \tag{1.6}$$

and a threshold $t \geq 0$, we obtain $\widehat{\boldsymbol{\beta}}(t)$ defined component-wise by

$$\widehat{\beta}_{z,j}(t) = \begin{cases} \widehat{\beta}_{z,j} & \text{if } \widehat{\beta}_{z,j} \geq t\widehat{\sigma}_j \\ 0 & \text{otherwise} \end{cases}, \quad z = 1, \ldots, Z,\ j = 1, \ldots, p,$$

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 the usual effect of canncellation of large positive and negative terms.

**Sparse recovery by non-negativity constraints.** The empirical success (see Figure 1.6 below) of the fitting-plus-thresholding approach to perform sparse recovery of non-negative signals is not a coincidence. To make our exposition self-contained, the goal of this paragraph is to provide the reader the main concepts an analysis of that approach is based upon. To this end, we follow the lines of Slawski and Hein (2011), where we provide a solid theoretical basis for the idea of of recovering a sparse, non-negative signal *without* regularization even in the presence of noise, thereby extending prior work addressing the noiseless case (Bruckstein et al., 2008; Wang and Tang, 2009; Donoho and Tanner, 2010; Wang et al., 2011). In these papers, the authors study uniqueness of non-negative solutions of underdetermined linear systems of equations

$$\boldsymbol{\Phi}\boldsymbol{\beta} = \boldsymbol{y} \text{ subject to } \boldsymbol{\beta} \geq \boldsymbol{0} \tag{1.7}$$

given the existence of a sparse solution $\boldsymbol{\beta}^*$ with support set $S = \{j :\ \beta_j^* > 0\}$ of cardinality $s$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, $\boldsymbol{A}\mathbb{R}_+^m = \{\boldsymbol{A}\boldsymbol{x} :\ \boldsymbol{x} \in \mathbb{R}_+^m\}$ denotes the polyhedral cone generated by the columns of $\boldsymbol{A}$. In geometrical terms, the condition for uniqueness is then given by the following statement.

**Proposition 1.1.** *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* (1.7) *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 $\boldsymbol{\beta}^* + \boldsymbol{\delta},\ \boldsymbol{\delta} \neq \boldsymbol{0}$. 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 $\boldsymbol{\beta}_{S^c}^* = \boldsymbol{0}$, 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 $p$. In order to extend Proposition 1.1 to a noisy setup with i.i.d. zero-mean sub-Gaussian error terms $\{\epsilon_i\}_{i=1}^n$, Slawski and Hein (2011) introduce an incoherence constant that naturally builds upon the notion of a face. Recall that the cone $\boldsymbol{\Phi}_S\mathbb{R}_+^s$ generated by the columns of the support is a face if there is a hyperplane separating it from the rest of the cone. The idea of the separating hyperplane constant $\widehat{\tau}(S)$ is to quantify separation. It is defined as the optimum value of the following quadratic program (it is assumed that $\|\boldsymbol{\Phi}_j\|_2 = O(\sqrt{n})$ for all $j$).

$$\widehat{\tau}(S) = \max_{\tau,\boldsymbol{w}} \tau$$
$$\text{subject to } \frac{1}{\sqrt{n}}\boldsymbol{\Phi}_S^\top \boldsymbol{w} = 0, \;\; \frac{1}{\sqrt{n}}\boldsymbol{\Phi}_{S^c}^\top \boldsymbol{w} \geq \tau\mathbf{1}, \;\; \|\boldsymbol{w}\|_2 \leq 1.$$

In geometric terms, $\widehat{\tau}(S)$ equals the distance of the subspace spanned by the columns of $\boldsymbol{\Phi}_S$ and the convex hull of the columns of $\boldsymbol{\Phi}_{S^c}$. Intuitively, the stronger the separation as indicated by the size of $\widehat{\tau}(S)$, the less sparse recovery will be affected by noise. Accordingly, a rough version of the main result in Slawski and Hein (2011) is as follows.

**Theorem 1.2.** *(Slawski and Hein, 2011). Consider the linear model* $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$*, where the entries* $\{\epsilon_i\}_{i=1}^n$ *of* $\boldsymbol{\epsilon}$ *are i.i.d. zero-mean sub-Gaussian with parameter* $\sigma > 0$*, and* $\boldsymbol{\beta}^*$ *is as in Proposition 1.1. Consider* $\widehat{\boldsymbol{\beta}}(t)$ *obtained by thresholding a non-negative least squares estimator as defined in* (1.6)*. If* $t > \frac{2\sigma}{\widehat{\tau}^2(S)}\sqrt{\frac{2\log p}{n}}$ *and* $\min_{j\in S}\beta_j^* > \widetilde{t}$*,* $\widetilde{t} = tC(S)$*, for a constant* $C(S)$*,* $\widehat{\boldsymbol{\beta}}(t)$ *satisfies* $\|\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*\|_\infty \leq \widetilde{t}$*, and* $\{j : \widehat{\beta}_j(t) > 0\} = S$*, with high probability.*

To the best of our knowledge, this is the first result about sparse recovery by non-negative least squares in a high-dimensional statistical inference framework. Yet, the result bears some resemblance with a similar result of Wainwright (2009) (Theorem 1) about support recovery of the lasso.

**Performance in practice.** With regard to the template matching problem one encounters for MS data, the fitting-plus-thresholding approach offers several advantages over $\ell_1$-regularized fitting.

- With the normalization $\sup_x \phi_{z,j}(x) = 1$ for all $j, z$, the coefficient $\widehat{\beta}_{z,j}$ equals the estimated amplitude of the highest peak of the template, such that $\widehat{\beta}_{z,j}/\widehat{\sigma}_j$ may be interpreted as signal-to-noise ratio and thresholding

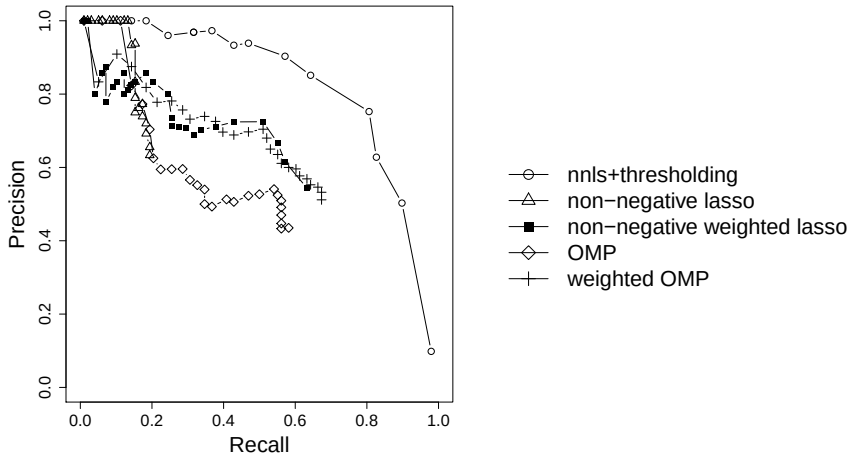


**Figure 1.6**: Precision-recall plot for the Myoglobin spectrum as described in the text.

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 $\ell_\infty$-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 $\ell_1$-regularized criteria (1.3) and (1.4), 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 $\lambda$ are required.

These aspects lead to an excellent performance in practice. We here present the results obtained on a MALDI-TOF spectrum of Myoglobin and compare them to those of $\ell_1$-regularization without (1.3) and with weights (1.4) as wells as to OMP and its weighted counterpart. 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 in Figure 1.6. 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)).

## 1.4 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 mis-specifications of the linear model (1.2) with regard to sparse recovery. In a second paragraph, we discuss alternatives to squared loss (1.6) one could argue for in view of specific properties of MS data. This issue deserves some attention, since for our problem 'denoising' and 'sparse recovery' are tightly connected, such that the choice of the loss function has considerable influence on the performance.

**Effects of sampling and misspecified templates.** Let us look more closely at the transition from the continuous model formulation (1.1) to the discrete one. Sampling yields pairs $\{(x_i, y_i^*)\}_{i=1}^n$, which are related by the linear model (1.2). However, (1.2) can hold only if the positions $\{\mu_j\}_{j=1}^p$ comprise the positions $\{\mu_k^*\}_{k=1}^s$ of the isotopic patterns. In practice, the $\{\mu_j\}_{j=1}^p$ are chosen as a subset of the sampling points, so that sampling at the unknown $m/z$ positions at which there is actually a peptide in the spectrum would be required, i.e. $\{x_i\}_{i=1}^n \supset \{\mu_k^*\}_{k=1}^s$ would have to hold true. We conclude that the matrix $\mathbf{\Phi}$ is not correctly specified in practice due to imprecision induced by sampling. Placing densely templates at subset of all positions $\{x_i\}_{i=1}^n$ which have been sampled leads to a phenomenon we refer to as 'peak splitting'. Consider an isotopic pattern of amplitude $\beta^*$ located at $\mu^*$ and let $\mu_l, \mu_r$, $\mu_l < \mu^* < \mu_r$, be the $m/z$-positions of templates in the dictionary closest to $\mu^*$ from the left and right, respectively. One observes that the corresponding non-negative least squares coefficients $\widehat{\beta}_l, \widehat{\beta}_r$ are both assigned positive values which are roughly proportional to the distances $|\mu_l - \mu^*|, |\mu_r - \mu^*|$ and $\beta^*$. In particular, if $|\mu_l - \mu^*| \approx |\mu_r - \mu^*|$ is small, the weight $\beta^*$ is divided into two weights $\widehat{\beta}_l, \widehat{\beta}_r$ of about the same size. Consequently, *any* sparse recovery method is very likely to select *both* templates located at $\mu_l$ and $\mu_r$. The situation is mimicked in Figure 1.7. The plot suggests that the lasso (1.3) 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 reason for 'peak splitting' is mis-specification of the function $\psi$ (cf. model (1.1)) that defines the shape of the smeared out peaks emerging in the spectrum. The function $\psi$ implicitly depends on a parameter controlling its spread, which may additionally be position-dependent. While Slawski et al. (2012) have developed a reliable procedure for estimating the spreads in a data-driven way, the estimates may yield a poor fit at some places of the

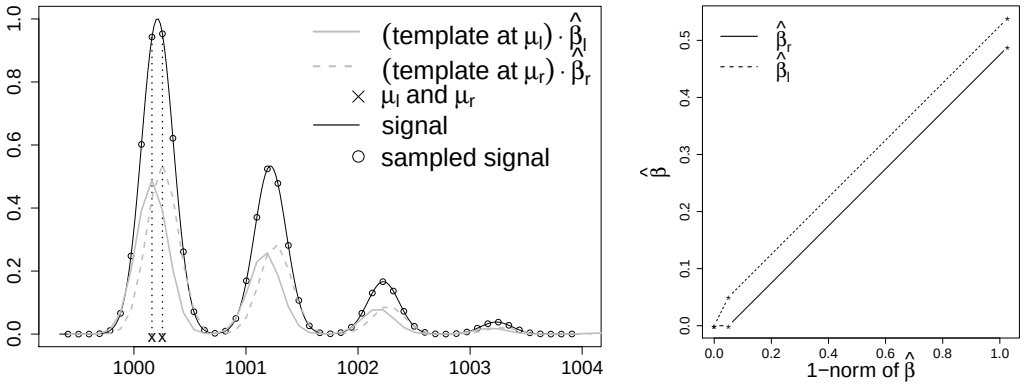


**Figure 1.7**: Systematic errors in the template model: consequences of a limited sampling rate. The right half of the plot displays the solution path of the non-negative lasso.

spectrum. Figure 1.8 shows the consequences of an underestimation of the spread. In order to avoid the effect arising from sampling, we work within an idealized setting where the true $m/z$-position of the pattern (denoted by 'correct' template in Figure 1.8) is included in the set of positions $\{\mu_j\}_{j=1}^p$ the templates of the dictionary are placed at. Again, $\ell_1$-regularization (1.3) would hardly save the day, because the selection of only one template would underestimate the true amplitude at least by a factor of two, as can be seen from the lower right panel. For the situations depicted in Figures 1.7 and 1.8, noise is not present. The issues raised here are caused by a wrong specification of $\mathbf{\Phi}$. The presence of noise may lead to an amplification of the effects one observes here.

Due to its frequent occurrence, 'peak splitting' requires a correction − otherwise, the output of *any* sparse recovery scheme would be only of limited practical use. The only possible way to address this issue within the sparse recovery framework would be to place templates less densely. However, this would come at the expense of reduced accuracy in estimating the positions $\{\mu_k^*\}_{k=1}^s$, which is not an option since it could hamper the biological validation of the output. In Slawski et al. (2012), a post-processing procedure is proposed that not only corrects peak-splitting, but that also tries to obtain even more accurate estimations for the positions. As detailed in Algorithm 1.3, all selected templates of the same charge that are within a neighbourhood whose size is proportional to the average spacing of two sampling points are merged to form a group. For each group of templates, precisely one new template is returned that comes closest to the fit when combining all templates of the group, thereby reducing the number of templates returned to

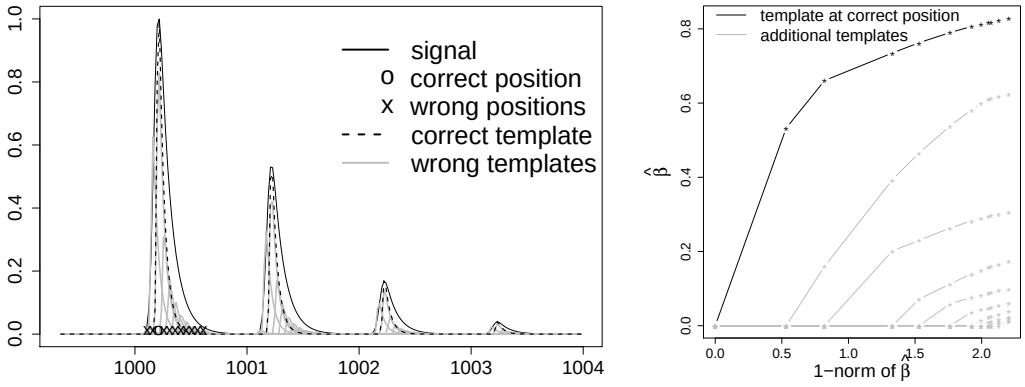


**Figure 1.8**: Systematic errors in the template model: consequences of an incorrectly specified spread. The right half of the plot displays the solution path of the non-negative lasso.

**Algorithm 1.3** Post-processing

**Input:** Output $\widehat{\boldsymbol{\beta}}$ of a sparse recovery algorithm obtained from a template matrix $\boldsymbol{\Phi}$ and intensities $\boldsymbol{y}$.

$\widehat{S}_z \leftarrow \{j : \widehat{\beta}_{z,j} > 0\}$, $z = 1, \ldots, Z$.
**for** $z = 1, \ldots, Z$ **do**
  $\overline{\boldsymbol{\mu}}_z \leftarrow \mathbf{0}, \quad \overline{\boldsymbol{\beta}}_z \leftarrow \mathbf{0}$.
  Partition $\widehat{S}_z$ into $G_z$ groups $\mathcal{G}_{z,1}, \ldots, \mathcal{G}_{z,G_z}$ by merging adjacent positions $\{\mu_j : \in S_z\}$.
  **for** $m = 1, \ldots, G_z$ **do**
    Using numerical integration, solve the nonlinear least squares problem

$$(\overline{\mu}_{z,m}, \overline{\beta}_{z,m}) = \operatorname*{argmin}_{\mu,\beta} \left\| \beta \cdot \phi_{z,\mu} - \sum_{l \in \mathcal{G}_m} \widehat{\beta}_{z,l} \phi_{z,l} \right\|_{L_2}^2 ,$$

    where $\phi_{z,\mu}(x) = (\psi \star \iota)(x - \mu)$ is a template at position $\mu$.
  **end for**
**end for**
**return** $\{\overline{\boldsymbol{\mu}}_z\}_{z=1}^Z$ and $\{\overline{\boldsymbol{\beta}}_z\}_{z=1}^Z$.

only one per detected pattern. By taking into account the coefficients of the templates assigned before post-processing, the accuracy of the position estimates can be considerably improved. For the situation depicted in Figure 1.7, the post-processing procedure returns a position roughly in the middle of the interval defined by the two sampling points. By choosing the size of the neighbourhood of a magnitude that is of the same order as the spacing between two sampling points, we ensure that the procedure does not erroneously merge templates that actually belong to different patterns, i.e. no

false negatives are introduced at that stage.

**Scope of the standard sparse recovery framework.** As argued in the introduction, applying sparse modeling techniques is a reasonable way to perform feature extraction for protein mass spectra. On the other hand, in view of the preceding discussion, the notion of support recovery commonly considered in theory (cf. Theorem 1.2) is not meaningful. But even if the linear model were free of any kind of mis-specification, the incoherence conditions employed for the analysis of non-negative least squares (cf. Section 1.3) and the lasso (Wainwright, 2009) would require a constant distance of the positions of the support templates from those of the off-support templates. As the sampling rate increases, however, one can hope to locate the positions of the patterns more accurately by accordingly placing templates more densely, so that incoherence approaches zero as $n$ tends to infinity − an obvious contradiction. On the other hand, suitable post-processing in form of Algorithm 1.3 permits us to overcome this limitation of the standard sparse recovery framework.

**Choice of the loss function.** MS data are contaminated by various kinds of noise arising from sample preparation and the measurement process. Apparently, the assumption of additive random noise with zero mean is not realistic, since the intensities $\{y_i\}_{i=1}^n$ are non-negative. Second, chemical noise generates a baseline which is much more regular than random noise. For this reason, we have not made explicit the relation between the intensities $\{y_i^*\}_{i=1}^n$ in (1.2) and their noisy counterparts $\{y_i\}_{i=1}^n$. Finding a realistic noise model is out of the scope of the paper, yet we would like to discuss alternatives to squared loss.

**Robust loss.** Eventually, model mis-specifications as described above can be absorbed by a general additive error term. The fact that drastic mis-specifications are not rare may make absolute loss a more suitable choice than squared loss which is known not to be robust to gross errors.

**Additive vs. multiplicative noise.** Both squared loss and absolute loss rely 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

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{(\boldsymbol{\Phi}\boldsymbol{\beta})_i - y_i \log((\boldsymbol{\Phi}\boldsymbol{\beta})_i)\}\,. \tag{1.8}$$

with the convention $0 \cdot \log(0) = 0$. Noting that $y_i \geq 0,\ i = 1, \ldots, n$, $L$ is seen to be convex with domain $\{\boldsymbol{\beta} : (\boldsymbol{\Phi}\boldsymbol{\beta})_i > 0\ \forall i \text{ with } y_i > 0\}$, which fits well into a non-negativity framework. Assuming that the $\{y_i\}_{i=1}^n$ are integers following a Poisson distributions with means $\{(\boldsymbol{\Phi}\boldsymbol{\beta})_i\}_{i=1}^n$, (1.8) equals the resulting negative log-likelihood. While the intensities $\{y_i\}_{i=1}^n$ are actually real-valued, they are obtained from a detector which basically counts the number of arriving molecules within a certain period. 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 postulates that $\mathbf{E}[y_i|\boldsymbol{\Phi}, \boldsymbol{\beta}] = \mathrm{var}[y_i|\boldsymbol{\Phi}, \boldsymbol{\beta}] = (\boldsymbol{\Phi}\boldsymbol{\beta})_i,\ i = 1, \ldots, n$, 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. (2013). 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 $\{y_i\}_{i=1}^n$ result from a combination of two templates and a perturbation by multiplicative noise, that is for $i = 1, \ldots, n = 600$,

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

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 $\phi_1$ and $\phi_2$ 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 1.3, once using non-negative least squares, once the Poisson-like loss (pll) given in (1.8), where $\widehat{\boldsymbol{\beta}}^{\mathrm{pll}}$ 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 $\phi_2$. This may not be accomplished in cases where the inclusion of off-support templates serves to compensate for misfit in $\phi_1$ arising from noise as shown in Figure 1.9. Table 1.1 suggests that the Poisson-like loss is preferable in this regard. For the real world MALDI-TOF Myoglobin spectrum (cf. Figure 1.6), we do not observe any improvement, as shown in Figure 1.10. This conforms to the hypothesis that the structure of the noise is too complex to be modelled welll by a simple multiplicative error term.

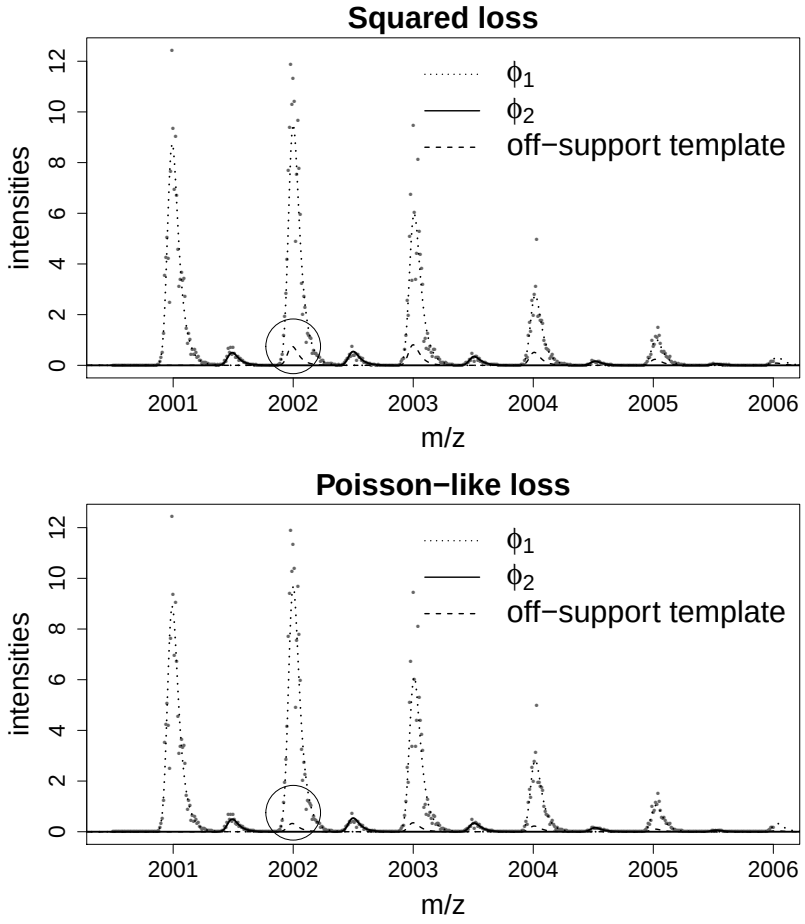


**Figure 1.9**: An instance of the experiment comparing squared loss and Poisson-like loss in the presence of low multiplicative noise. Top: Fit of non-negative least squares. Bottom: Fit of the Poisson-like loss. In the upper panel, the coefficient of the off-support template exceeds that of $\phi_2$ such that sparse recovery via thresholding is not possible.

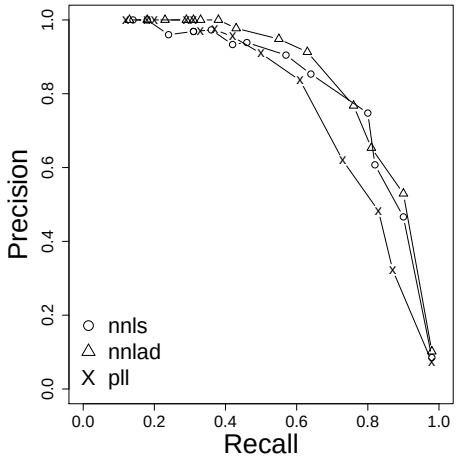


**Figure 1.10**: Performance of the three loss functions for the MALDI-TOF Myoglobin spectrum: squared loss (nnls), absolute loss (nnlad) and poisson-like loss (pll) in conjunction with the fitting-plus-thresholding approach. The precision-recall curve for nnls identical to that in Figure 1.6.

| | $\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$ | $\|\widehat{\boldsymbol{\beta}}_{S^c}\|_\infty$ | $I(\|\widehat{\boldsymbol{\beta}}_{S^c}\|_\infty > \widehat{\beta}_2)$ |
|---|---|---|---|
| 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 1.1**: Results of the experiment comparing squared loss and Poisson-like loss in the presence of low multiplicative noise. 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.

**Summary.** In this chapter, we have discussed how to apply sparse recovery methods to feature extraction for protein mass spectra. While the chapter addresses a specific kind of data, we believe that various issues raised in our exposition have implications for other fields of applications where deconvolution, sparsity in connection with non-negativity and heteroscedasticity, which are the dominant themes of the chapter, play an important role.

**Acknowledgments.** We thank our collaborators Rene Hussong and Andreas Hildebrandt (Junior research group for Computational Proteomics & Protein-Protein-Interactions at the Center for Bioinformatics, Saarland University) and Andreas Tholey, Thomas Jakoby and Barbara Gregorius (Divi-

sion for Systematic Proteome Research, Institute for Experimental Medicine, Universität Kiel). The project was funded by the cluster of excellence 'Multimodal Computing and Interaction' (MMCI) of Deutsche Forschungsgemeinschaft.

## 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. *The Annals of Statistics*, 32:407–499, 2004.

J. Jia, K. Rohe, and B. Yu. The Lasso under Poisson-like heteroscedasticity. *Statistica Sinica*, 23:99–118, 2013.

R. Lawson and C. Hanson. *Solving least squares problems*. SIAM Classics in Applied Mathematics, 1987.

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. Sparse recovery by thresholded non-negative least squares. In *Advances in Neural Information Processing Systems 24*, pages 1926–1934. 2011.

M. Slawski, R. Hussong, A. Tholey, T. Jakoby, B. Gregorius, A. Hildebrandt, and M. Hein. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics*, 13:291, 2012.

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. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

M. Wang and A. Tang. Conditions for a Unique Non-negative Solution to an Underdetermined System. In *47th Annual Allerton Conference on Communication, Control, and Computing*, pages 301–307, 2009.

M. Wang, W. Xu, and A. Tang. A unique nonnegative solution to an undetermined system: from vectors to matrices. *IEEE Transactions on Signal Processing*, 59:1007–1016, 2011.

T. Zhang. On the Consistency of Feature Selection using Greedy Least Squares Regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

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