BMC
Bioinformatics

METHODOLOGY ARTICLE   **Open Access**

Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching

Martin Slawski[1*], Rene Hussong[2,3], Andreas Tholey[4], Thomas Jakoby[4], Barbara Gregorius[2,4], Andreas Hildebrandt[2,5] and Matthias Hein[1]

Abstract

Background: The robust identification of isotope patterns originating from peptides being analyzed through mass spectrometry (MS) is often significantly hampered by noise artifacts and the interference of overlapping patterns arising e.g. from post-translational modifications. As the classification of the recorded data points into either 'noise' or 'signal' lies at the very root of essentially every proteomic application, the quality of the automated processing of mass spectra can significantly influence the way the data might be interpreted within a given biological context.

Results: We propose non-negative least squares/non-negative least absolute deviation regression to fit a raw spectrum by templates imitating isotope patterns. In a carefully designed validation scheme, we show that the method exhibits excellent performance in pattern picking. It is demonstrated that the method is able to disentangle complicated overlaps of patterns.

Conclusions: We find that regularization is not necessary to prevent overfitting and that thresholding is an effective and user-friendly way to perform feature selection. The proposed method avoids problems inherent in regularization-based approaches, comes with a set of well-interpretable parameters whose default configuration is shown to generalize well without the need for fine-tuning, and is applicable to spectra of different platforms. The R package `IPPD` implements the method and is available from the Bioconductor platform (http://bioconductor.fhcrc. org/help/bioc-views/devel/bioc/html/IPPD.html).

Background

Mass spectrometry (MS), often in conjunction with high performance liquid chromatography (HPLC), is the de-facto standard analytical tool to derive important biological knowledge about the protein content of whole cells, organelles, or biomedical samples like tumour or blood plasma. Within a typical experimental setup, purified proteins of the sample under study are digested by an enzyme. Before entering the mass spectrometer, peptides are separated chromatographically according to their physico-chemical properties in order to avoid a massive

overlapping of peptide signals within a single scan. Nevertheless, due to the sheer number of peptides present in a sample, interfering patterns still occur frequently, not least because of post-translational modifications such as the deamidation of asparagines or glutamine residues. In order to obtain an unambiguous assignment of the signals, and in particular their isotope patterns, which is a prerequisite for a proper identification and quantification, every data point in m/z dimension is classified either as 'signal' or as 'noise' during the so-called feature detection phase. As this processing lies at the very root of every proteomic application, the quality of feature detection can have dramatic impact on the finally derived results and conclusions. In view of the large amount of data even a single MS experiment can produce, automated analysis is indispensable. However, due to various artifacts arising

*Correspondence: ms@cs.uni-saarland.de
[1] Department of Computer Science, Saarland University, Saarbrücken, Germany
Full list of author information is available at the end of the article

from electric and chemical noise and baseline trends, the identification of isotope patterns is error-prone and time consuming. In addition, severe overlaps of peptide signals within the same mass spectrometric scan can hamper a straightforward analysis furthermore. In recent years, numerous procedures have been developed to process this data (cf., e.g., [1-8]). Within this paper, we propose a novel method that is demonstrated to perform especially well in challenging situations, characterized e.g. by strong local variations in noise and intensity levels or the presence of isotope patterns of different charges exhibiting overlap, which in many cases may be difficult to resolve even for a human expert by visual inspection. Existing software typically depends on a large set of parameters requiring careful fine-tuning, often being rather sensitive to changes in the measurement process like the change of the platform, which makes a proper parameter choice a laboursome task. In contrast, the proposed method has been designed to depend on a comparatively small set of well-interpretable parameters whose default configuration is shown to be robust, yielding mostly excellent, but at least competitive performance on spectra of different platforms. In a nutshell, our method uses non-negative least squares or non-negative least absolute deviation regression to fit a spectrum *s* by a large dictionary of templates mimicking isotope patterns; since true positions and charges of isotope patterns in the spectrum are unknown in advance, regions where the signal exceeds a local measure of noise are identified and then a vast set of templates is placed in those regions. In the spirit of sparse recovery, a small subset of the templates, which reasonably explains the observed signal, is selected by applying hard thresholding with a locally adaptive choice of the threshold to the regression coefficients obtained previously. Our method is related to a formerly proposed template-based approach (`NITPICK`, [3]). As opposed to the present work, `NITPICK` uses ℓ_1-regularized non-negative least squares. Without non-negativity constraints, this procedure is known as the lasso [9]. Reference [10] contains the first application of the lasso to the problem studied in the present paper. Given a dramatic increase in occurrence of high-dimensional datasets in recent years and the resulting need for feature selection, the lasso, due to computationally and theoretically appealing properties, has meanwhile become so popular that it can be regarded as a standard tool of modern data analysis [11]. In this respect, `NITPICK` follows the usual paradigm suggesting that ℓ_1-regularization is the method of choice. In the present paper, we argue for a deviation from that paradigm mainly in view of the following two aspects. First, a major benefit of our fitting+thresholding approach is that parameter choice is more user-friendly, since the threshold can be interpreted in terms of a signal-to-noise ratio. This is unlike the regularization parameter of the lasso, which can

in general not be related directly to the signal. In the presence of heterogeneous noise and model misspecifications, the 'right amount' of regularization is notoriously difficult to choose. Second, there is a substantial body of work showing that non-negativity constraints alone may suffice to recover a sparse target. Non-negative least squares + thresholding is analyzed in [12], where it is shown that it can significantly outperform the usual ℓ_1-approach with respect to sparse recovery. See Section "Sparse recovery with non-negativity constraints: non-negative least squares + thresholding vs. the non-negative lasso" for a detailed discussion.

Methods

A spectrum is understood as a sequence of pairs $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i = m_i/z_i$ is a mass (m_i, measured in Dalton *Da*) to charge (z_i), and y_i is the intensity, i.e. the abundance of a particular mass (modulo charge state), observed at x_i, $i = 1, \ldots, n$, which are assumed to be ordered increasingly.

Template model

The $(y_i)_{i=1}^{n} = \mathbf{y}$ are modeled as a positive combination of templates designed on the basis of prior knowledge about peak shape and composition of isotope patterns. If our model were perfectly correct, we could write

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{\beta}^* = \sum_{c=1}^{C} \mathbf{\Phi}_c \boldsymbol{\beta}_c^*, \ \ \mathbf{\Phi}_c = [\boldsymbol{\varphi}_{c,1} \ldots \boldsymbol{\varphi}_{c,p_c}], \ \ c = 1, \ldots, C,$$

(1)

where $\mathbf{\Phi}$ is a non-negative matrix of templates and $\boldsymbol{\beta}^*$ is a non-negative coefficient vector. Both $\mathbf{\Phi}$ and $\boldsymbol{\beta}^*$ can be arranged according to charge states $c = 1, \ldots, C$. Each sub-matrix $\mathbf{\Phi}_c$ can in turn be divided into columns $\boldsymbol{\varphi}_{c,1}, \ldots, \boldsymbol{\varphi}_{c,p_c}$, where the entries of each column vector store the evaluations of a template $\boldsymbol{\varphi}_{c,j}$, $j = 1, \ldots, p_c$, at the x_i, $i = 1, \ldots, n$. It is assumed that only a small fraction of the templates in $\mathbf{\Phi}$ are needed to represent the signal, i.e. $\boldsymbol{\beta}^*$ is highly sparse. The templates are of the form

$$\varphi_{c,j} = \sum_{k \in \mathbb{Z}} a_{c,j,k} \ \psi_{c,j,k,\boldsymbol{\theta}_{c,j}},$$

(2)

where the $\psi_{c,j,k}$ are functions representing a single peak within an isotope pattern, depending on a location $m_{c,j}$ and a parameter vector $\boldsymbol{\theta}_{c,j}$. In general, peaks can be modeled by Gaussian, Lorentzian, and sech^2 shapes, cf. [13]. Due to their similarity, we restrict ourselves to the Gaussian, but provide in addition the exponentially modified Gaussian (EMG, cf., e.g., [14]), a model for a possibly skewed peak as occuring frequently in MALDI-TOF recordings, where late ion formation in the gas phase

leads to tailed peaks [15]. The EMG is parameterized by $\boldsymbol{\theta}_{c,j} = (\alpha_{c,j}, \sigma_{c,j}, \mu_{c,j})^{\top} \in \mathbb{R}^{+} \times \mathbb{R}^{+} \times \mathbb{R}$ (for $\alpha_{c,j} \downarrow 0$, one obtains a Gaussian)

$$\psi_{c,j,k}(x) = \frac{1}{\alpha_{c,j}} \exp\left(\frac{\sigma_{c,j}^2}{2\alpha_{c,j}^2} + \frac{\mu_{c,j} - (x - m_{c,j,k})}{\alpha_{c,j}}\right) \quad (3)$$
$$\times \left(1 - F\left(\frac{\sigma_{c,j}}{\alpha_{c,j}} + \frac{\mu_{c,j} - (x - m_{c,j,k})}{\sigma_{c,j}}\right)\right),$$
$$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \, du.$$

In (2), the nonnegative weights $a_{c,j,k}$ equal the height of the isotopic peak k within the pattern j of charge state c. These heights are computed according to the averagine model [16]. The $m_{c,j,k}$ are calculated from $m_{c,j}$ as $m_{c,j,k} = m_{c,j} + \kappa \frac{k}{c}$, where κ usually ranges between 1.002 and 1.008 Dalton, see e.g. [17]. Note that in Eq. (2) the location of the most intense peak ($a_{c,j,0} = \max_k a_{c,j,k}$) is taken as characteristic location of the template instead of using the finally reported monoisotopic position: we set $m_{c,j,0} = m_{c,j}$ so that the remaining $m_{c,j,k}$, $k \neq 0$, are computed by shifting $m_{c,j}$ in both directions along the m/z axis. With the normalization $\max_x \varphi_{c,j}(x) = 1$ for all c,j, the entries of $\boldsymbol{\beta}^*$ can be interpreted as intensities of the most intense peaks of the templates. The construction scheme is illustrated in Figure 1.
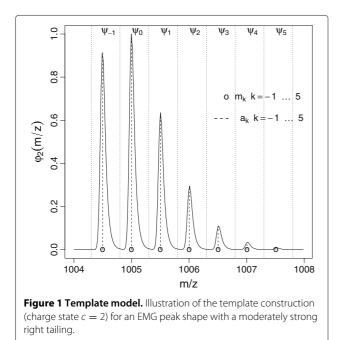
Parameter estimation
The parameters $\theta_{c,j} = (\alpha_{c,j}, \sigma_{c,j}, \mu_{c,j})^{\top}$ of the peaks (3) are unknown in practice. Following a central paradigm of our framework, whic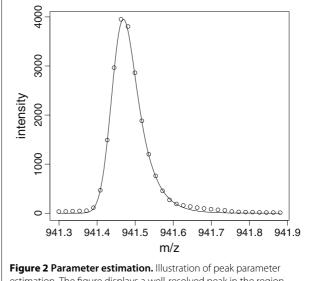h is to relieve the user of performing laboursome fine-tuning of parameters, we have developed a systematic procedure automatically providing estimates of these parameters, which is considerably more efficient and flexible than a grid search. For instance, the parameters may additionally depend on the m/z-position. Our framework for parameter estimation extends a conceptually similar approach in [18] designed for a Gaussian peak shape.

In a first step, we apply a simple peak detection algorithm to the spectrum to identify disjoint regions $\mathcal{R}_r \subset \{1, \ldots, n\}$, $r = 1, \ldots, R$, of well-resolved peaks. For each region, we fit the chosen peak shape to the data $\{(x_i, y_i)\}_{i \in \mathcal{R}_r}$ using nonlinear least squares:

$$\min_{\boldsymbol{\theta}} \sum_{i \in \mathcal{R}_r} (y_i - \psi_{\boldsymbol{\theta}}(x_i))^2, \quad (4)$$

yielding an estimate $\widehat{\boldsymbol{\theta}}_r(\widehat{x}_r)$, where $\widehat{x}_r$ denotes an estimation for the mode of the peak in region $\mathcal{R}_r$. This concept is sketched in Figure 2. The nonlinear least squares problem (4) is solved by using a general purpose nonlinear least squares routine available in most scientific computing environments, e.g. nls in R. Once the sequence of estimators $\{\widehat{\boldsymbol{\theta}}_r(\widehat{x}_r)\}$ has been obtained, they are subject to a suitable aggregation procedure. In the simplest case, one could simply take averages. For spectra where peak shape characteristics, in particular peak width, are known to vary systematically with m/z position, we use the pairs $\{(\widehat{x}_r, \widehat{\boldsymbol{\theta}}_r(\widehat{x}_r))\}$ as input into a linear regression procedure to infer the parameters of pre-specified trend functions. Formally, we model each component θ_l of $\boldsymbol{\theta}$ as a linear

Figure 1 Template model. Illustration of the template construction (charge state $c = 2$) for an EMG peak shape with a moderately strong right tailing.

Figure 2 Parameter estimation. Illustration of peak parameter estimation. The figure displays a well-resolved peak in the region $\mathcal{R} = \{i : 941.3 \, Th \leq x_i \leq 941.9 \, Th\}$. In this example, the size of $\mathcal{R}$ equals 33, i.e. there are 33 pairs $\{(x_i, y_i)\}$ that enter a nonlinear least squares problem of the form (4). Under the assumption of an EMG model, the resulting fit is indicated by a solid line.

combination of known functions $g_{l,m}$ of $x = m/z$ and an error component ε_l, i.e.

$$\theta_l(x) = \sum_{m=1}^{M_l} v_{l,m} g_{l,m}(x) + \varepsilon_l(x), \tag{5}$$

for which a linear trend i.e. $\theta_l(x) = v_{l,1} + v_{l,2}x$, is one of the most common special cases. In [19], a set of instrument-specific models for the peak width is provided, all of which can be fitted by our approach.

We refrain from using least squares regression to determine the parameters in (5) due to its sensitivity to possible outliers, which arise from poorly resolved, wiggly or overlapping isotope patterns, which may affect the quality of the estimates $\widehat{\theta}_r$. Therefore, the linear model is fitted in a robust way by using least absolute deviation regression. Given the resulting estimates of the parameters $\{v_{l,m}\}$, m/z-specific estimates for the parameters in (3) are obtained by evaluating (5).

Template fitting
The computation of the design matrix Φ requires a set of m/z positions at which templates are placed. In general, one has to choose positions from the interval $[x_1, x_n]$. We instead restrict ourselves to a suitable subset of the set $\{x_i\}_{i=1}^n$. The deviations from the positions of the true underlying isotope patterns is then at least in the order of the sampling rate, but this can be improved by means of a postprocessing step described in Section "Postprocessing and thresholding". Using the whole set $\{x_i\}_{i=1}^n$ may be computationally infeasible if n is large and is in fact not necessary since isotope patterns occur very sparsely in the spectrum. Therefore, we apply a preselection step on the basis of what we term 'local noise level' (LNL). The LNL is defined as the median of the intensities y_i falling into a sliding window of fixed width around a specific position. For $x \in [x_1, x_n]$, we define the local noise level based on sliding window width h as

$$\text{LNL}(x) = \text{median}(\{y_i : i \in \mathcal{I}_x\}), \tag{6}$$
$$\mathcal{I}_x = \{i : x_i \in [x - h, x + h]\}.$$

Given the LNL, we place templates at position x_i (one for each charge state) if the corresponding y_i exceeds $\text{LNL}(x_i)$ by a factor `factor.place`. Section "Finding a set of default parameters" describes how we determined defaults for the two parameters h and `factor.place`. In fact, the LNL is a central quantity in our framework, because it does not only influence the placement, but also the selection of templates (see Section "Postprocessing and thresholding" below). Choosing h too small typically has the effect that the LNL is overestimated such that true peaks might be incorrectly classified as noise. Conversely, choosing h too large leads to an underestimation, thereby

increasing the computational burden as well as the number of spurious patterns included in the final list. The advantages of working with the median are obvious: easy computation, robustness and equivariance with respect to monotone transformations. Similar notions of local noise can be found in the literature, see e.g. [8] where a truncated mean is used. Given the positions of the templates, we generate the matrix Φ according to Eqs. (1) and (2). In the fitting step, we compute a non-negative least squares ($q = 2$) or alternatively non-negative least absolute deviation ($q = 1$) fit by determining a minimizer $\widehat{\boldsymbol{\beta}}$ of the criterion

$$\min_{\boldsymbol{\beta} \geq \mathbf{0}} \left\| \mathbf{y} - \Phi \boldsymbol{\beta} \right\|_q^q, \quad q = 1 \text{ or } q = 2, \tag{7}$$

The optimization problem (7) is a quadratic ($q = 2$) or linear ($q = 1$) program and is solved using interior point methods (e.g. [20]). The details are relegated to Appendix "Fitting with non-negativity constraints" section. As far as the choice of q is concerned, we point out that $q = 1$ yields a robust fit that can deal better with deviations from model assumptions, i.e. deviations from the averagine model or from the peak model. However, in general, we are unable to provide any recommendation about how to choose q. Therefore, in our validation, both are evaluated.

Comparison with pepex
In prior work [21], subsequently referred to as 'pepex', non-negative least squares fitting is used as well. An important difference to our approach is that the matrix Φ is not constructed from the convolution of isotope distributions and peak shapes as described in Section "Template model". Instead, peak detection is applied first to reduce the raw intensity data to peak clusters, a step that is usually referred to as centroiding. At the second stage, called de-isotoping, peak clusters are fitted by a design matrix containing isotope distributions themselves, not convolved versions. While the approach is computationally more attractive and avoids estimation of peak shape parameters (cf. Section "Parameter estimation"), the division into centroiding and de-isotoping may lead to poor performance for low resolution and noisy data, or in the presence of overlapping patterns. In these cases, peak detection is little reliable. In our template-based approach, there is no separation of centroiding and de-isotoping. It performs much better in the aforementioned cases, since it operates directly on the data and is hence less affected if single peaks of a pattern are difficult to detect. This reasoning is supported by our evaluation in Section "Results and discussion" as well as that in [3]. At the same time, our approach can in principle be applied to centroided spectra as well. In this case, the columns of the matrix Φ directly represent isotope distributions instead of isotopic patterns.

Postprocessing and thresholding

While indeed a considerable fraction of the entries of $\widehat{\boldsymbol{\beta}}$ are precisely equal to zero, treating all positions for which the corresponding entry differs from zero as locations of isotope patterns would yield a huge number of false positives, at least because of regions, in which noise fitting reduces the objective in (7). Therefore, the fitting step of the previous section is accompanied by a thresholding step, with the aim to separate signal from noise. However, fitting followed by thresholding alone does not lead to a proper output. The strategy could be successful if our template model were free of any kind of misspecification. Even when neglecting possible misfits of the averagine model, we still have to cope with two sources of systematic errors − a limited sampling rate and mismatches in the peak model. These are the main reasons for what we term 'peak splitting', referring to the phenomenon that several templates are used to fit precisely one pattern. Figure 3 illustrates the effect of sampling in a noiseless setting. In the top panel, the signal is sampled in such a way that the top of the peak is lost. When placing two templates at the two sampling points x_l, x_u of maximum signal, non-negative least squares fitting attributes weights $\widehat{\beta}_l, \widehat{\beta}_u$ of roughly equal size to the templates. The postprocessing procedure outlined below yields a suitable correction. One might object that 'peak splitting' is a problem inherent in our entirely fitting-oriented approach (7) not incorporating any form of regularization. The bottom panel of Figure 3 shows the solution path of the non-negative lasso [22] given by $\{\widehat{\boldsymbol{\beta}}(\lambda), \ \lambda \geq 0\}, \quad \widehat{\boldsymbol{\beta}}(\lambda) =$ $\mathrm{argmin}_{\boldsymbol{\beta} \geq \mathbf{0}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \mathbf{1}^\top \boldsymbol{\beta}$. One obtains two nearly parallel trajectories, demonstrating that only a heavily biased fit, which would undesirably lead to the exclusion of additional smaller signals, could accomplish the selection of only one template.

To a large extent, 'peak splitting' can be corrected by means of the following merging procedure, which we regard as postprocessing of the fitting step (7) and which we apply prior to thresholding. Given an estimate $\widehat{\boldsymbol{\beta}}$, we define $\widehat{\mathcal{M}}_c = \{m_{c,j} : \ \widehat{\beta}_{c,j} > 0\} \subset \{x_i\}_{i=1}^n, \ c = 1, \dots, C$, as the set of all template locations where the corresponding coefficient exceeds 0.

1. Separately for each c, divide the sets $\widehat{\mathcal{M}}_c$ into groups $\mathcal{G}_{c,1}, \dots, \mathcal{G}_{c,G_c}$ of 'adjacent' positions. Positions are said to be adjacent if their distance on the m/z scale is below a certain tolerance ppm specified in parts-per-million, cf. Section "Finding a set of default parameters". In the context of 'peak splitting', the templates at locations sharing the same group are assumed to fit precisely one true underlying peak.
2. With the notation of Eq. (2), for each $c = 1, \dots, C$, and for $g = 1, \dots, G_c$, we solve the following optimization problem.
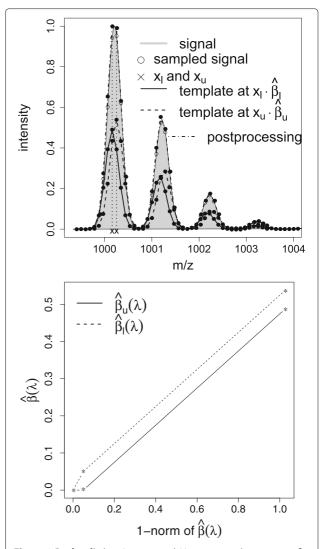

Figure 3 Peak splitting. Lower panel: Non-negative least squares fit of the sampled signal with and without postprocessing. Upper panel: Solution path of the non-negative lasso for the same data.

$$(\widetilde{m}_{c,g}, \widetilde{\beta}_{c,g}) = \underset{\substack{m_{c,g} \\ \beta_{c,g}}}{\mathrm{argmin}} \int_{-\infty}^{\infty} \left(\sum_{m_{c,j} \in \mathcal{G}_{c,g}} \widehat{\beta}_{c,j} \psi_{m_{c,j}}(x) - \beta_{c,g} \psi_{m_{c,g}}(x) \right)^2 dx,$$

(8)

with the aim to find a location $\widetilde{m}_{c,g}$ and a weight $\widetilde{\beta}_{c,g}$ of the most intense peak $\psi_{\widetilde{m}_{c,g}}$ within an isotope pattern $\varphi_{c,g}$ approximating the fit of the most intense peaks $\{\psi_{m_{c,j}} : \ m_{c,j} \in \mathcal{G}_{c,g}\}$ within the isotope patterns $\{\varphi_{c,j} : \ m_{c,j} \in \mathcal{G}_{c,g}\}$ best in a least squares sense.

3. One ends up with sets $\widetilde{\mathcal{M}}_c = \{\widetilde{m}_{c,g}\}_{g=1}^{G_c}$ and coefficients $\{\widetilde{\beta}_{c,g}\}_{g=1}^{G_c}, \ c = 1, \dots, C$.

The additional benefit of solving (8) in step two as compared to the selection of the template with the largest

coefficient within each group as proposed in [3] is that we are able to determine the location of the pattern even more accurately as predetermined by a limited sampling rate, since in (8) we optimize the location over a continuum. The optimization problem (8) can be solved fast and accurately by sampling the integrand on a fine grid of points and then solving a nonlinear least squares problem with optimization variables $m_{c,g}$ and $\beta_{c,g}$.

All candidate positions $(\widetilde{m}_{c,g}, \widetilde{\beta}_{c,g})$ are assigned a signal-to-noise ratio

$$\text{ratio}_{c,g} = \text{GOF}_+(\widetilde{m}_{c,g}) \cdot \frac{\widetilde{\beta}_{c,g}}{\text{LNL}_+(\widetilde{m}_{c,g})}, \qquad (9)$$

where $\text{LNL}_+(\widetilde{m}_{c,g}) = \max\left\{\text{LNL}(\widetilde{m}_{c,g}), \frac{1}{4}\text{median}(\{\text{LNL}(x_i)\}_{i=1}^n)\right\}$ is a truncated version of the local noise level, with a lower bound included to avoid that the denominator in (9) takes on tiny values in low-intensity regions. The factor $\text{GOF}_+(\widetilde{m}_{c,g})$ represents a goodness-of-fit adjustment, a correction which aims at downweighting spurious peaks in low-intensity noise regions. These are not hard to distinguish from signal regions, which, in view of the presence of peak patterns, tend to be considerably regular. In order to spot noise regions, we fit the spectrum by *single* peaks (3) placed at each datum x_i, $i = 1, \ldots, n$, where the peak shape model, the associated peak shape parameters and the parameter q are chosen according to the choice made for template generation (Sections "Template model") and template fitting (Section "Template fitting"), respectively. Denote the residuals of the resulting fit by $\{r_i\}_{i=1}^n$. A local measure of goodness-of-fit is defined by

$$\text{GOF}_+(x) = \min\left\{1 - \frac{\sum_{i \in \mathcal{I}_x} |r_i|^q}{\sum_{i \in \mathcal{I}_x} |y_i|^q}, 0.5\right\},$$

The idea underlying this procedure is that in noise regions, the fit to the data will be poor, and consequently, the size of the residuals is expected to be large relative to the signal, hence leading to a low goodness-of-fit statistic. The truncation at 0.5 limits the influence of this correction. A final list is generated by checking whether the signal-to-noise ratios (9) exceed a 'significance threshold' t specified by the user. We do not give a general guideline for choosing t, because a reasonable choice is very specific to experimental conditions, e.g. the platform used and the composition of the spectrum. It is important to note that while t itself is constant, we take into account that the noise level is heterogeneous, since thresholding is based on the ratios (9), where the local noise level enters.

Finding a set of default parameters

Apart from the signal-to-noise threshold t, we have introduced the parameters `window`, i.e. the width h of the sliding window required for the computation of the

local noise level (6), the template placement parameter `factor.place` and the parts-per-million tolerance `ppm` within which peaks are considered to be merged by the postprocessing procedure. With the exception of the threshold t, we have fixed all parameters to a default setting which we expect to give reasonable (albeit potentially suboptimal) results on spectra different from the ones analyzed here, without the need of manual tuning. In order to find such a default setting, we performed a grid search using only one selected spectrum of those described in Section "Datasets" below. While our default setting, which can be found in the HTML manual of the R package `IPPD`, already performs well, we recommend to do such a calibration to optimize the performance of our method.

Sparse recovery with non-negativity constraints: non-negative least squares + thresholding vs. the non-negative lasso

We believe that our preference for the first alternative is a major methodological contribution that has potential to impact related problems where non-negativity problems come into play. In the present section, we provide, at a high level, a series of arguments rooting in the statistics and signal processing literature that clarify our contribution and support our preference.

Linear models and usual paradigms in statistics

The fact that we favour non-negative least squares + thresholding may seem implausible since it questions or partially even contradicts paradigms about high-dimensional statistical inference. Consider the linear model

$$\boldsymbol{y} \approx \boldsymbol{\Phi}\boldsymbol{\beta}^*, \qquad \boldsymbol{y} \in \mathbb{R}^n, \quad \boldsymbol{\Phi} \in \mathbb{R}^{n \times p}, \qquad (10)$$

which corresponds to model (1), where '$\approx$' is used instead of '$=$' to account for stochastic noise or model misspecifications. Linear models of the form (10) have been and continue to be objects of central interest in statistical modelling.

- Classical work in statistics shows that under mild conditions if the number of sample n grows at a faster rate than the number of features p, the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{\text{ols}} \to \boldsymbol{\beta}^*$ (in probability) as $n \to \infty$.
- Since many contemporary datasets, like the MS datasets of the present paper, are characterized by a large p, which is of the same order as n or even larger, the first bullet has considerably lost relevance. Translated to MS datasets, it provides a statement about the case where the resolution tends to infinity. Therefore, modern statistical theory studies regimes in which p is allowed to grow at a faster rate than n,

with a focus on results that hold for finite sample sizes. These results hinge on some sort of sparsity assumption on $\boldsymbol{\beta}^*$, the simplest being that $\boldsymbol{\beta}^*$ is zero except for some index set (support) of small cardinality. In this context, a multitude of results has been proved (see e.g. [23] for an overview) indicating that the lasso estimate $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ is a statistically optimal procedure in the sense that *if the regularization parameter is chosen in the right way*, the squared Euclidean distance $\|\widehat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_2^2$ is nearly of the same order as that of an estimator one could construct if the non-zeroes of $\boldsymbol{\beta}^*$ were known.

The second bullet provides quite some justification for NITPICK, which is based on the lasso. However, as detailed below, the italicized part can be critical. On the other hand, there are several results that support our approach.

The power of non-negativity constraints

- It turns out that the non-negativity constraint $\boldsymbol{\beta} \geq \mathbf{0}$ imposed in non-negative least squares (NNLS) may lead to a drastically better performance than that of the *ordinary* least squares estimator in 'large p' situations provided $\boldsymbol{\Phi}$ satisfies additional conditions. Roughly speaking, it is shown in [12] that if $\boldsymbol{\Phi}$ has non-negative entries, which is fulfilled for the template matching problem of Section "Template Model", the NNLS estimator $\widehat{\boldsymbol{\beta}}$ does not overfit and is unique even in the singular case ($p > n$). These results indicate that NNLS may behave surprisingly well in a high-dimensional setup, without using ℓ_1-regularization, which is often propagated in the literature as basically the only option ([24], Section 16.2.2).
- There are several recent papers [25-27] in the sparse recovery literature in which it is shown that a sparse, *non-negative* vector can be recovered from few linear measurements $n \ll p$. In [12], these results are extended to a noisy setup. More specifically, it is shown that NNLS + thresholding can consistently recover $\boldsymbol{\beta}^*$ and its support. Very recently, using similar conditions as in [12], Meinshausen [28] has established several guarantees of NNLS in a high-dimensional setup.

One should bear in mind that the non-negativity constraints are essential for our approach. Thresholding the unconstrained ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ in general leads to poor results in the 'large p' situation.

Shortcomings of ℓ_1-regularization in theory

In [12], it is not only shown that NNLS + thresholding is a sound strategy to perform sparse recovery

of a non-negative target, but also examples are given where the non-negative lasso is outperformed even if its regularization parameter is set to match theoretical results and regardless of whether subsequent thresholding as advocated in [29,30] is used or not. In particular, inferiority of the lasso arises in the presence of small, yet significantly non-zero entries in $\boldsymbol{\beta}^*$. These are specifically affected by the non-negligible bias of ℓ_1-regularization [31]. It is important to note that the comparison in [12] does not contradict prior comparisons of the lasso (aka soft thresholding) and (hard) thresholding for *orthonormal designs* ($\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{I}$) in [32,33], where both approaches perform similarly well and non-negativity constraints are not particularly important. Orthonormal designs, which lead to greatly simplified estimation problem are not of interest in the context of the paper, since the template matrix $\boldsymbol{\Phi}$ is far from being orthonormal.

Shortcomings of ℓ_1-regularization in practice

The study in [12] is of more theoretical nature, since all constants of the problem, in particular the noise level, are known, so that the regularization parameter can be set in an optimal fashion. This can realistically not be accomplished in practice. Likewise, the information-theoretic criterion employed in [3] as well as the data-splitting approach of [34] rely on knowledge of the noise level, or a consistent estimate thereof, which is hard to obtain in the 'large p' situation [35]. In any case, the regularization parameter remains a quantity that is hard to grasp and hence hard to set for a practitioner, since it cannot be related directly to the signal. In contrast, the threshold t admits a straightforward interpretation.

Moreover, when using ℓ_1-regularization, data fitting and model selection are coupled. While this is often regarded as advantage, since model selection is performed automatically, we think that it is preferable to have a clear separation between data fitting and model selection, which is a feature of our approach. Prior to thresholding, the output of our fitting approach gives rise to a ranking which we obtain without the necessity to specify any parameter. Selection is completely based on a single fit simply by letting the the threshold vary. On the contrary, if one wants to reduce the number of features selected by the lasso, one resets the regularization parameter and solves a new optimization problem. Note that it is in general not possible to compute the entire solution path of the lasso [22] for the MS datasets used for the present paper, where the dimension of $\boldsymbol{\Phi}$ is in the ten thousands so that the active set algorithm of [22] is prohibitively slow. In this regard, model selection by thresholding is computationally more attractive.

Results and discussion

For the assessment of the pattern picking performance, in total eight spectra generated by two different ionization methods, matrix assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI), respectively, form the basis of the evaluation. While MALDI has been coupled to a time-of-flight (TOF) mass analyzer, ESI MS spectra have been recorded on both a linear ion trap (LTQ) and an Orbitrap mass analyzer. In addition, a series of spectra were prepared with the aim of investigating in detail the method's performance in the presence of overlapping peptides.

Datasets

For MALDI mass spectra (Additional file 1), time of flight mass analysis was performed; spectra were recorded on an ABI MALDI-TOF/TOF 4800 instrument in positive ion mode using α-cyano-4-hydroxy-cinnamic acid (CHCA) as matrix. Nanospray ESI spectra (Additional file 2) were measured in positive ion mode on a Thermo LTQ Orbitrap Velos MS; both high resolution measurements using the Orbitrap mass analyzer (referred to as 'Orbitrap') and, alternatively, low resolution linear ion trap (IT) measurements were performed with this setup. This experiment has been chosen in order to demonstrate the utility of our method at different concentration levels, that it is robust with respect to changes in the data-generating process and that the method is capable of handling singly charged ions, the main form generated by MALDI MS, as well as higher charged ions formed in ESI MS. Tryptic digests (performed in 40 mM ammonium bicarbonate) of model proteins were used as analytes: bovine myoglobin and chicken egg lysozyme (10 and 500 fmol each) for MALDI-TOF experiments, and lysozyme (250 and 1000 fmol) for ESI experiments. Disulfide bonds were reduced with dithiothreitol (DTT) prior to alkylation, free cysteine residues were alkylated by iodacetamide. No further sample pretreatment was performed prior to MS analysis. When referring to these spectra, we omit that tryptic digests are given: e.g., the term 'MALDI-TOF myoglobin spectrum (500 fmol)' means the respective tryptic digest.

To demonstrate explicitly the method's ability to separate strongly overlapping patterns even in the case of badly resolved signals, 22 additional spectra have been generated in positive ion mode on a Bruker Daltonics HCT Ultra Ion Trap MS with an electrospray ion source. Three synthetic peptides (cf. Section "Unmixing of overlaps" for details) with sequences corresponding to tryptic peptides from bovine serum albumin (BSA) were used as analytes. In each measurement two out of three peptides were mixed in different ratios to get overlapping peptide signals, also with different charge states. Two different concentrations (500 fmol/μl and 1000 fmol/μl) were injected into the mass spectrometer via a Cole-Parmer syringe pump.

Validation strategy

Validation of pattern picking is notoriously difficult, because a gold standard which is satisfactory from both statistical and biological points of view is missing. In this context, a major problem one has to account for is that spectra frequently contain patterns whose shape is not distinguishable from those of peptides, but which are in fact various artifacts resulting e.g. from impurities during sample preparation and measurement. These artifacts do not constitute biologically relevant information and are, in this sense, 'false positives'. An important instance are signals derived from the matrix (or from matrix-clusters) frequently observed in MALDI MS. The pattern of these signals is similar to that of peptides; nevertheless, due to their molecular composition, which differs significantly from that of an average peptide, the exact masses can be used to exclude these signals from the data analysis. On the other hand, from a statistical perspective which judges a method according to how well it is able to detect specific patterns in a given dataset, a qualification as 'true positive' is justified. With the aim to unify these aspects, we have worked out a dual validation scheme. In order to reduce the number of artifacts, all automatically generated lists of candidates for peptide masses as well as the lists of a human expert (see below) are postprocessed by a peptide mass filter [36]: only peptides whose monoisotopic mass deviated less than 200 ppm from the closest peptide mass center[a] are used for subsequent evaluation.

Comparison with manual annotation

The first part investigates how well a method is able to support a human expert who annotates the spectra manually. More specifically, the automatically generated lists are matched to the manual annotation such that an entry of the list (potential peptide mass) is declared 'true positive' whenever there is a corresponding mass in the manual annotation deviating by no more than Δ ppm. Otherwise, it is declared 'false positive'. In order to adapt Δ ppm to the resolution of the different mass lines, we used the following strategy: assuming that most of the peptides will have a mass larger than 700 Da, we determined the spacing $\Delta_{m/z}$ between neighboring data points in m/z direction for each mass spectrum in the lower mass range. If we further assume that a simple manual annotation by visual inspection can result in a mass deviation from the 'correct' mass position of at most $\Delta_{m/z}$, we can derive the following tolerance values: $\Delta = 100$ ppm for ion trap recordings, $\Delta = 50$ ppm in the case of MALDI-TOF recordings[b] and $\Delta = 20$ ppm for Orbitrap data.

As the performance of our as well as those of all competing methods depends on a threshold-like parameter

governing, crudely speaking, the trade-off between precision and recall, we explore the performance for a range of reasonable parameter values, instead of fixing an (arbitrary) value, which we believe to be little meaningful. The results are then visualized as ROC curve, in which each point in the (Recall, Precision)-plane corresponds to a specific choice of the parameter. Formally, we introduce binary variables $\{B_i(t)\}$ for each mass i contained in the list of cardinality $\widehat{L}(t)$ when setting the threshold equal to t, where $B_i(t)$ equals 1 if the mass is matched and 0 otherwise, and denote by L the number of masses of the manual annotation. The true positive rate (recall, R), and the positive predictive value (precision, P) associated with threshold t are then defined by $R(t) = \frac{\sum_i B_i(t)}{L}$, $P(t) = \frac{\sum_i B_i(t)}{\widehat{L}(t)}$. An ROC curve results from a sequence of pairs $\{R(t), P(t)\}$ for varying t.

Database query

The second part evaluates the lists in terms of a query to the Mascot search engine [37], version 2.2.04. In particular, we account for a major problem of a manual annotation, namely that peptides yielding weak MS signals might easily be overlooked, but might be detected by methods designed to extract those weak signals. Since we are especially interested in demonstrating the method's ability to separate overlapping patterns, we adapted the standard search parameters of Mascot's peptide mass fingerprint routine to allow two missed cleavage sites and to incorporate the following (variable) post-translational modifications: 'Oxidation (M)', 'Carbamidomethyl (C)', 'Amidated (Protein C-term)', 'Deamidated (NQ)'. In particular, the latter two modifications will frequently trigger MS signals interleaving with the pattern of their unmodified counterpart: in the case of a deamidation the modified ion shows a mass of approx. 0.98 Da more compared to the amidated peptide. The same mass tolerances as for the manual annotation are used. As for the comparison with the manual annotation, we evaluate several lists corresponding to different choices of the threshold. Instead of an ROC curve, which turned out to be visually unpleasant, we display the statistics (score, coverage and fraction of hits) of two lists per method, namely of those achieving the best score and the best coverage, respectively. The complete set of results as well as further details of our evaluation like the manual annotation are contained in Additional file 3.

Competing methods

We compare our method in its two variants depending on the choice of the fitting criterion (cf. Eq. (7)), labelled l_1 ($q = 1$) and l_2 ($q = 2$), respectively, with the following competing methods.

Lasso

The 'lasso' method in this paper serves as surrogate for NITPICK. Since the 'lasso' is embedded into our framework while implementing a methodology that closely resembles NITPICK, we use the 'lasso' for the sake of convenience, to avoid an involved parameter optimization for NITPICK. Our lasso implementation benefits from the improved merging procedure of Section "Postprocessing and thresholding". To accomodate a heterogeneous noise level, NITPICK divides spectra into bins. This can be avoided by determining a minimizer $\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{W})$ of the weighted non-negative lasso problem

$$\min_{\boldsymbol{\beta} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda \boldsymbol{1}^\top \boldsymbol{W}\boldsymbol{\beta}, \quad \lambda > 0, \tag{11}$$

where $\boldsymbol{W}$ is a diagonal matrix with entries $w_{c,j} = \text{LNL}_+(m_{c,j})$, $j = 1, \ldots, p_c$, $c = 1, \ldots, C$, whose purpose is to re-scale the amount of ℓ_1-regularization according to the local noise level. The columns of the template matrix $\boldsymbol{\Phi}$ in (11) are normalized to unit Euclidean norm as it is standard in the literature on the lasso. A grid search over 50 values for λ is performed, where the construction of the grid follows [38]. Different lasso lists are obtained for each active set $A(\lambda_k) = \{c, j : \widehat{\beta}_{c,j}(\lambda_k; \boldsymbol{W}) > 0\}$, $k = 1, \ldots, 50$, which are subsequently merged (see Eq. (8)). The parameter λ here plays the role of the threshold t, cf. Section "Validation strategy".

Pepex

As discussed in Section "Template fitting", pepex performs centroiding and de-isotoping separately. De-isotoping is based on non-negative least squares. Since pepex is limited to detect patterns of charge state one, its performance is only assessed for MALDI-TOF spectra. Accordingly, when comparing the ouptput of pepex with the manual annotation, the few patterns of charge state two are excluded. The parameters nm, pft, mincd, maxcd and nsam were set to optimize performance with respect to manual annotation. The ROC curves are based on peak-lists resulting from ten different choices of the signal-to-noise parameter snr.

Isotope wavelet

As opposed to our method, this approach is not able to handle overlaps. On the other hand, it typically shows strong performance in noisy and low intensity regions or on datasets with extremely low concentrations [39,40]. While the isotope wavelet is not limited to charge one, it is run in charge one only mode for the MALDI-TOF spectra, to achieve more competitive performance. For the sake of fair of comparison, the result of the isotope wavelet on the MALDI-TOF spectra are evaluated in the same way as those of pepex.

Vendor

The parameter setting for the ABI MALDI-TOF/TOF MS software was as follows: Local Noise Width (m/z) 250, Min Peak Width at FWHM 2.9. The Cluster Area Optimization S/N threshold has been dynamically adapted to about three times the S/N threshold as suggested by the ABI documentation. Since the vendor software is limited to charge one, its outputs are evaluated in the same way as those of pepex. Given the disproportionally high effort needed to find an optimal parameter setting of the vendor software for ESI spectra, its performance is not assessed.
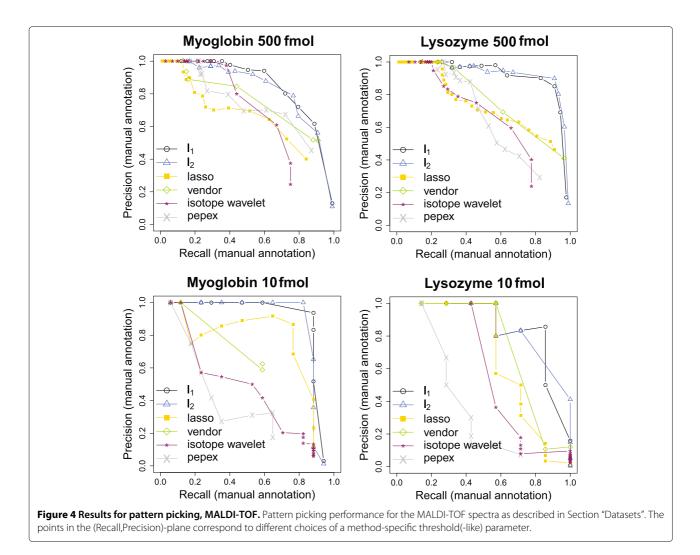
Results

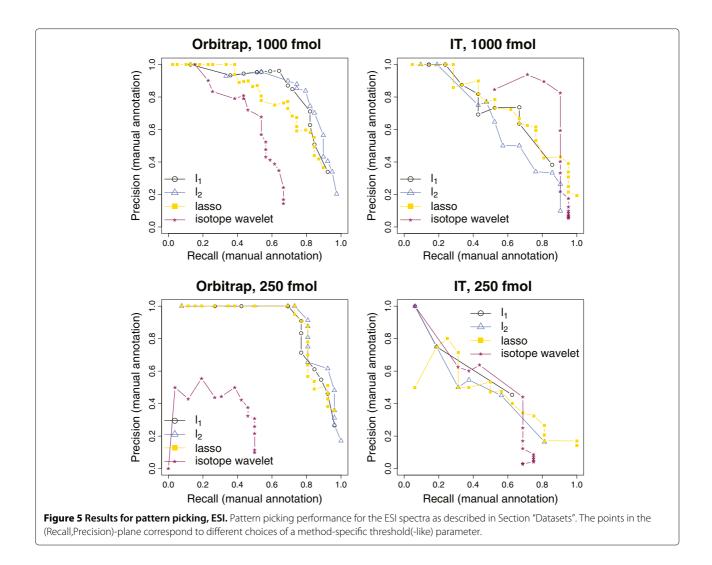#### Manual annotation vs. database query

When inspecting Figures 4 and 5 on the one hand and Table 1 on the other hand, one notices that results of the evaluation based on the manual annotation are not in full accordance with the results of the database query. The difference is most striking for the MALDI-TOF spectra at 500 fmol, where our methods (l_1 and l_2) yield a significant

improvement, which does not become apparent from the database query. This is because only a fraction of the manual annotation is actually confirmed by the database query. The part which is not matched likely consists of artifacts due to contamination or chemical noise as well as of specific modifications not captured by the database query. In light of this, our dual validation scheme indeed makes sense.

Comparison

Figure 4 and Table 1 reveal an excellent performance of our methods (l_1 and l_2) throughout all MALDI-TOF spectra under consideration. For the myoglobin spectra high sequence coverages are attained that clearly stand above those of competing methods. For the spectra at 10 fmol, only the performance of lasso is competetive with that of our methods in terms of the Mascot score; all other competitors, including the vendor software which has been tailored to process these spectra, are significantly weaker. In particular, the strikingly high proportion of

Figure 4 Results for pattern picking, MALDI-TOF. Pattern picking performance for the MALDI-TOF spectra as described in Section "Datasets". The points in the (Recall,Precision)-plane correspond to different choices of a method-specific threshold(-like) parameter.

Figure 5 Results for pattern picking, ESI. Pattern picking performance for the ESI spectra as described in Section "Datasets". The points in the (Recall,Precision)-plane correspond to different choices of a method-specific threshold(-like) parameter.

'hits' ($\geq$ 94%) indicates that even at moderate concentration levels, our methods still distinguish well between signal and noise. This observation is strongly supported by the ROC curves in Figure 4, where the precision drops comparatively slowly with increasing recall. In this regard, our methodology clearly contrasts with approaches like the isotope wavelet that aim at achieving high protein sequence coverage. The latter often requires the selection of extremely lowly abundant peptide signals hidden in noise at the expense of reduced specificity.

For MALDI-TOF spectra at high concentration levels, pepex achieves the best scores and is competitive with respect to sequence coverage. However, the performance of pepex degrades dramatically at lower concentration levels, as it is unambiguously shown by both parts of the evaluation. In particular, the database scores are the worst among all methods compared. This provides some support for our reasoning at the end of Section "Template fitting".

For the ESI spectra, our methods in total fall a bit short of the lasso (particularly for the ion trap spectra), but perform convincingly as well, thereby demonstrating that they can deal well with multiple charge states. This is an important finding, since the presence of multiple charges makes the sparse recovery problem as formulated in model (1) much more challenging, because the number of parameters to be estimated as well as the correlations across templates are increased. In spite of these difficulties, Figure 5 and Table 1 suggest that the performance of our pure fitting approach (7) does not appear to be affected. Using a more difficult set of spectra, the capability to process ESI data with impressive success is additionally shown in the next section.

Additional remarks

- In Figure 4, the area under the curve (AUC) of our methods attained for myoglobin is higher for lower

Table 1 Mascot results

MALDI Myo 500 fmol	Score	cvrg	Hits	Score	cvrg	Hits
l_1	211.0	**0.85**	0.94	96.8	**0.96**	0.04
l_2	211.0	**0.85**	0.94	49.6	**0.96**	0.04
lasso	207.0	**0.85**	**1.00**	142.0	0.91	**0.37**
pepex	**223.0**	**0.85**	**1.00**	142.0	0.90	0.17
vendor	**223.0**	**0.85**	0.94	**174.0**	0.90	0.29
wavelet	207.0	**0.85**	**1.00**	156.0	0.90	0.14
MALDI Lys 500 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	167.0	**0.81**	0.57	133.0	**0.83**	**0.37**
l_2	168.0	0.80	0.64	**144.0**	**0.83**	0.34
lasso	151.0	0.64	0.77	112.0	**0.83**	**0.37**
pepex	**172.0**	0.80	0.63	135.0	**0.83**	0.25
vendor	146.0	0.64	**0.75**	91.4	**0.83**	0.20
wavelet	127.0	0.58	**0.75**	113.0	0.81	0.20
MALDI Myo 10 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	**211.0**	**0.85**	0.94	82.2	**0.95**	0.04
l_2	207.0	0.74	**1.00**	109.0	0.90	0.14
lasso	195.0	0.77	0.87	**146.0**	0.85	0.46
pepex	97.8	0.80	0.22	97.8	0.80	0.22
vendor	123.0	0.62	0.62	123.0	0.62	**0.62**
wavelet	131.0	**0.85**	0.13	131.0	0.85	0.13
MALDI Lys 10 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	**89.0**	0.35	**1.00**	**73.7**	0.54	**0.23**
l_2	**89.0**	0.35	**1.00**	35.4	0.72	0.09
lasso	81.9	**0.46**	0.70	46.0	0.74	0.10
pepex	47.1	0.17	**1.00**	31.2	0.53	0.12
vendor	62.7	0.23	**1.00**	43.2	0.34	0.16
wavelet	55.4	0.23	0.45	43.8	**0.82**	0.10
Orbi Lys 1000 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	149.0	0.70	0.78	138.0	0.80	**0.53**
l_2	139.0	**0.80**	0.50	**139.0**	0.80	0.50
lasso	**159.0**	0.63	**0.87**	120.0	**0.81**	0.29
wavelet	105.0	0.69	0.44	95.1	0.80	0.23
IT Lys 1000 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	78.7	0.63	0.28	70.9	0.74	0.17
l_2	82.1	0.72	0.36	35.4	0.85	0.13
lasso	103.0	**0.84**	0.33	**76.8**	**0.99**	**0.21**
wavelet	**107.0**	0.79	**0.63**	69.8	**0.99**	0.11
Orbi Lys 250 fmol	**Score**	**cvrg**	**Hits**	**Score**	**cvrg**	**Hits**
l_1	107.0	0.63	0.50	100.0	0.80	0.31
l_2	103.0	0.63	0.52	66.9	**0.81**	0.14
lasso	**108.0**	0.63	**0.77**	107.0	0.80	**0.27**
wavelet	80.6	**0.70**	0.22	80.6	0.70	0.22

Table 1 Mascot results (Continued)

IT Lys 250 fmol	Score	cvrg	Hits	Score	cvrg	Hits
l_1	59.4	0.46	0.16	59.4	0.46	0.16
l_2	37.0	0.59	0.14	37.0	0.59	0.14
lasso	**66.3**	**0.84**	0.20	**66.3**	**0.84**	**0.20**
wavelet	56.3	0.59	**0.36**	21.3	0.75	0.12

Corresponding Mascot results for the data shown in Figures 4 and 5. The left halves of the tables report the statistics when choosing the threshold(-like) parameter to optimize the score, the right halves when optimizing the coverage (cvrg, given as fraction). The column 'hits' contains the fraction of masses that could be matched to peptide masses in the database.

concentration. At first glance, this may seem contradictory since an increase in concentration should lead to a simplified problem. However, a direct comparison of the AUCs is problematic, since the number of true positives (17 at 10 fmol, 106 at 500 fmol) is rather different. For instance, there are choices of the threshold that yield 18 true positives and not a single false positive for both of our methods at 500 fmol, yet the AUC is lower.

- The fact that some of the ROCs start in the lower left corner results from outputs containing only false positives.

Unmixing of overlaps
Motivation
One of the main advantages of our method over more simplistic pattern picking methods is the ability to disentangle isotope patterns of overlapping peptide signals, whose presence may lead to a significantly more challenging pattern picking problem as e.g. discussed in [41] in the slightly different context of intact protein mass spectra. Therefore, a potential application for our approach will be the analysis of a certain class of posttranslational modifications, the deamidation of amino acid residues containing a carboxamide side chain functionality. The deamidation of asparagine (Asn) or glutamine (Gln) residues, yielding aspartic acid (Asp) or glutamic acid (Glu) residues, respectively, is an important posttranslational modification, which can have immense effects on the structure of peptides [42] and is of great relevance in a number of pathophysiological events [43]. During the deamidation, the side chain carboxamide is hydrolysed, which is accompanied by a mass increase of 0.98 Da. Thus, in a spectrum of a mixture of the amidated and deamidated form, a direct overlap of both signals can be observed. It has to be noted that additionally to the amidated/deamidated forms, in case of Asn deamidation, a second product containing an iso-peptide bond is formed, too, which has the same molecular behaviour; these two forms can be identified solely by their differential MS/MS behavior.

Results
The peptides analyzed here in order to assess the performance of our approach were synthesized by means of Fmoc-solid phase peptide synthesis; sequences corresponding to tryptic peptides from bovine serum albumin (BSA) with the sequences listed in Table 2 were selected.

In each measurement two out of the three listed peptides were mixed together in different ratios (Additional file 4). Given such a spectrum, we study the question whether our method returns the true underlying composition. We classify the output of our method as correct interpretation of the spectrum if the templates corresponding to the true underlying peptides achieve signal-to-noise-ratios of at least one and these ratios are the two largest among all templates used for fitting. This procedure corresponds to a selection-optimal choice of the threshold based on the knowledge of the true composition of the spectrum. This simplification may be justified in view of the extreme difficulty of the problem as illustrated in Figure 6, in particular in view of lowly resolved spectra with an average m/z-spacing of 0.06 Da. For the remaining parameters, we compare a grid search (performed separately for each spectrum) and the default parameter set (Section "Finding a set of default parameters"). Table 3 and Figure 6 indicate that already the default parameter setting is able to solve successfully a wide range of problem instances. As one would expect, Table 3 and Figure 6 suggest that the higher the concentration and the more balanced the amplitudes of the overlapping peptides, the more likely it is that the overlap can be resolved. On the other hand, the higher the degree of overlap of the peptides, which depends on both their charges and the distance of their positions, the more difficult the problem is.

Table 2 Peptides mixed together

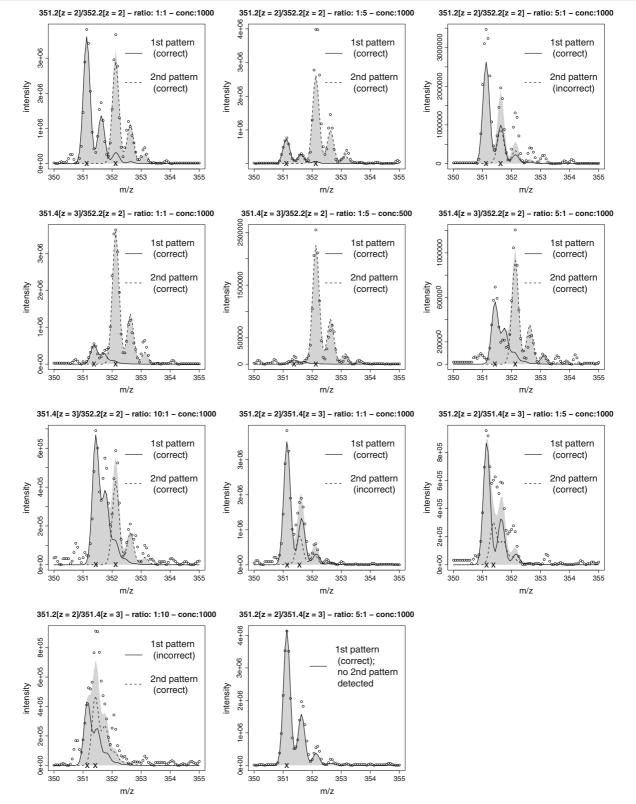| Sequence	Sequence residue no. in BSA	Monoisotopic mass (protonated) / charge
GACLLPK	198-204	351.20437 / +2
CCTKPESER	460-468	351.48816 / +3
VLASSAR	212-218	352.20850 / +2

Figure 6 Unmixing of overlap. Graphical representation of selected overlap problems as tabulated in Table 3. The experimental setups are given in the title of the plots. The dots represent the signal, while solid and dashed lines represent the templates used by our method to match the signal, using the default parameter setting and the choice for the threshold as explained in the text. The grey area represents the overall fit when using the selected templates.

Table 3 Unmixing of overlaps

Peptides		351.2(2)/352.2(2)			351.4(3)/352.2(2)				351.2(2)/351.4(3)			
	Proportion	1:1	1:5	5:1	1:1	1:5	5:1	10:1	1:1	1:5	5:1	1:10
	fmol											
all	500	x	x	—	—	x	—	x	—	—	—	x
	1000	x	x	x	x	x	x	x	x	x	—	—
default	500	—	x	—	—	x	—	—	—	—	—	—
	1000	x	x	—	x	—	x	x	—	x	—	—

Results of the analyses of the series of spectra containing two overlapping target peptides. The first column contains the m/z positions and the charges (in brackets) of the two peptides. The two middle columns indicate whether the corresponding problem is successfully solved ('x') or not ('—') when optimizing over a grid of parameter combinations (column 'all') and when using the default parameter set (column 'default').

This becomes obvious when considering the overlap of the two peptides located at 351.2 and 351.4 Da, respectively.

Conclusion

We have proposed a template matching approach for feature extraction in proteomic mass spectra. The main methodological innovation is a framework for sparse recovery in which sparsity is not promoted explicitly by a regularization term, as it is usually done and was done in previous work. We fully exploit the strength of non-negativity constraints, which permits us to circumvent the delicate choice of a 'proper' amount of regularization, an ever-lasting problem in statistics, and to work with thresholding instead. The latter is not only computationally attractive, because one does not have to repeatedly solve the same optimization problem for different choices of the regularization parameter, but also increases user-friendliness, since the threshold is directly related to the signal-to-noise ratio, the quantity domain experts are interested in. The replacement of a regularization parameter by a threshold is a cornerstone in our conceptual design guided by the principle to relieve the user from laboursome fine tuning of parameters. We believe that a small set of well-interpretable parameters with suitable defaults additionally improves robustness and reproducibility of results. In this context, we would like to emphasize again that apart from the threshold, the user does not have to specify any parameters before running our software.

In a comprehensive experimental study involving instruments of varying resolution and spectra of varying concentration levels, where we comparatively assess the performance of our approach on the basis of an elaborate dual validation scheme, it is demonstrated that the performance for pattern picking is excellent for MALDI-TOF spectra and outstands due to its specificity in selecting signal and only little noise. A major strength of the method is its ability to unmix overlapping peptide signals as shown for a series of ESI spectra. In total, we demonstrate that our approach is broadly applicable to a variety of spectra. While our approach is guided by a concrete application in proteomics, the framework is general enough to be of much of use for related deconvolution problems emerging in other fields — only the templates have to be adjusted according to the specific application.

While in this paper, we have focused on single spectra, the approach can be extended to process whole LC-MS runs, as it has already been implemented in our R package IPPD. More precisely, the sweep line scheme of [44] is used to agglomerate the results from single spectra. To apply our methods on a routine basis, an improved implementation, notably parallelization, is required, since e.g. processing a single spectrum of the Maxquant datasets [2] takes 10s on average on a Unix system equipped with an Intel(R) Core(TM)2 Duo CPU T9400 (2.53GHz) and 4 GB main memory. There is much room for an improvement, since our implementation is based on interpreted R code.

Concerning future directions of research, a question we have not yet answered in a satisfactory way is the choice of the fitting criterion. While both criteria (least squares and least absolute deviation) employed in this paper perform well, their implicit assumption of additive noise might be questionable [45]. It is worth investigating whether a multiplicative noise model could even yield better results. Second, one might ask whether the performance could be further improved when it is used jointly with the isotope wavelet, which is affected by overlaps, but has the potential to achieve higher protein sequence coverage.

Endnotes

[a]Monoisotopic peptide mass centers are modelled by: $1.000485 \cdot m_n + 0.029$, where m_n denotes the nominal mass.
[b]For the MALDI-TOF lysozyme datasets an extended search tolerance of 100ppm was applied due to experimental miscalibration of the MS.

Appendix
Fitting with non-negativity constraints

In the following, we provide the details concerning optimization problem (7). In view of the special structure of $\mathbf{\Phi}$, (7) is computationally tractable even if n and the number

of templates are in the ten thousands. We exploit the sparsity of the problem arising from templates which are highly localized, i.e. the domain on which they are numerically different from zero covers only a small part of the whole m/z range of the spectrum. As a consequence both $\mathbf{\Phi}$ and the Gram matrix $\mathbf{\Phi}^\top\mathbf{\Phi}$, which is crucial in the computation, can conveniently be handled by using software for sparse matrices. For R, such software is available in the `Matrix` package [46].

Non-negative least squares

Consider the quadratic program

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{\beta}\|_2^2 \tag{12}$$

$$\text{subject to } \boldsymbol{\beta} \geq 0. \tag{13}$$

In order to solve (12), we use the so-called log-barrier method which amounts to solving a sequence of an unconstrained nonlinear convex problems in which the constraints $I(\beta_j \geq 0)$, $j = 1,\dots,p$, are taken into account by incorporating log-barrier terms $-\log(\beta_j)/\gamma$ in the objective. As $\gamma \to \infty$, the log-barrier acts like a function which equals $+\infty$ if $\beta_j < 0$ and zero otherwise. Beginning with a moderately sized starting value for γ, we solve the convex problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{\beta}\|_2^2 - \frac{1}{\gamma}\sum_{j=1}^{p}\log(\beta_j) \tag{14}$$

using Newton's method. The gradient and Hessian with respect to $\boldsymbol{\beta}$, respectively, are given by

$$\nabla_{\boldsymbol{\beta}} = -\mathbf{\Phi}^\top(\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{\beta}) - \frac{1}{\gamma}[\,1/\beta_1 \ \dots \ 1/\beta_p\,]^\top.$$

$$\nabla_{\boldsymbol{\beta}}^2 = \mathbf{\Phi}^\top\mathbf{\Phi} + \frac{1}{\gamma}\text{diag}(1/\beta_1^2,\dots,1/\beta_p^2).$$

The Newton descent direction $\boldsymbol{d}_{\boldsymbol{\beta}}$ is obtained from the linear system

$$\nabla_{\boldsymbol{\beta}}^2 \boldsymbol{d}_{\boldsymbol{\beta}} = -\nabla_{\boldsymbol{\beta}}.$$

Solution of linear systems of this structure constitutes the main computational effort to be made. Fast solutions are obtained by using `CHOLMOD` [47], which offers an efficient implementation for computing the Cholesky factorization of sparse symmetric, positive definite matrices. Since the diagonal of $\nabla_{\boldsymbol{\beta}}^2$ changes from one Newton iteration to the next, one Cholesky factorization has to be performed per Newton step. Once we have solved (14) for a specific γ, we solve a new problem of the type (14) for $\gamma \cdot M$, $M > 1$. This is repeated until γ exceeds a predefined maximum value. For a thorough account on the log-barrier method, we refer to [20].

Complexity analysis of non-negative least squares

We here provide the order of magnitude of floating points operations (flops) required per update (i.e. per Newton step) for the specific non-negative least squares problems considered for this paper. In our implementation, we exploit that the templates contained in the matrix $\mathbf{\Phi}$ are highly localized. As a result, after a suitable column permutation, the matrix $\mathbf{\Phi}^\top\mathbf{\Phi}$ is roughly a band matrix with bandwidth k no larger than only few hundreds. The dominant operation is solving the linear system $\nabla_{\boldsymbol{\beta}}^2 \boldsymbol{d}_{\boldsymbol{\beta}} = -\nabla_{\boldsymbol{\beta}}$ with the help of the Cholesky factorization, which can be done in $O(pk^2)$ flops (e.g. [20], p.670). Our algorithm terminates after usually no more than one hundred Newton steps.

Non-negative least absolute deviation

Consider the optimization problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{\beta}\|_1 \tag{15}$$

$$\text{subject to } \boldsymbol{\beta} \geq 0. \tag{16}$$

Problem (15) can be recast as the following linear program.

$$\min_{\boldsymbol{r}} \boldsymbol{r}^\top \mathbf{1} \tag{17}$$

subject to

$$\mathbf{\Phi}\boldsymbol{\beta} - \boldsymbol{y} + \boldsymbol{r} \geq 0,$$

$$\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{\beta} + \boldsymbol{r} \geq 0,$$

$$\boldsymbol{r} \geq 0,$$

$$\boldsymbol{\beta} \geq 0.$$

For its solution, we use the log-barrier method sketched in the previous paragraph. After incorporating log-barrier terms for all constraints, the objectives of the unconstrained convex problems are of the form

$$\boldsymbol{r}^\top \mathbf{1} - \frac{1}{\gamma}\left(\sum_{i=1}^{n}(\log(\xi_i^+) + \log(\xi_i^-) + \log(r_i)) + \sum_{j=1}^{p}\log(\beta_j)\right),$$

where we have used the notational shortcuts

$$\xi_i^+ = (\mathbf{\Phi}\boldsymbol{\beta})_i - y_i + r_i,$$

$$\xi_i^- = y_i - (\mathbf{\Phi}\boldsymbol{\beta})_i + r_i, \ \ i = 1,\dots,n.$$

The gradients w.r.t. $\boldsymbol{r}$ and $\boldsymbol{\beta}$, respectively, are given by

$$\nabla_{\boldsymbol{r}} = \mathbf{1} - \frac{1}{\gamma}\left[\frac{1}{(\xi_1^+ + \xi_1^- + r_1)} \ \cdots \ \frac{1}{(\xi_n^+ + \xi_n^- + r_n)}\right]^\top,$$

$$\nabla_{\boldsymbol{\beta}} = -\frac{1}{\gamma}(\mathbf{\Phi}^\top([\,\mathbf{\Xi}^+]^{-1} - [\,\mathbf{\Xi}^-]^{-1})\mathbf{1} + [\,1/\beta_1 \dots 1/\beta_p]^\top),$$

$$\mathbf{\Xi}^\pm = \text{diag}(\xi_1^\pm,\dots,\xi_n^\pm).$$

Introducing $\boldsymbol{R} = \text{diag}(r_1,\dots,r_n)$ and $\boldsymbol{B} = \text{diag}(\beta_1,\dots,\beta_p)$, the Hessian is given by the block matrix

$$
\begin{bmatrix} \nabla_r^2 & \nabla_{r\,\beta} \\ \nabla_{r\,\beta}^\top & \nabla_\beta^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\gamma}([\,\Xi^+]^{-2}+[\,\Xi^-]^{-2}+R^{-2}) & \frac{1}{\gamma}([\,\Xi^+]^{-2}\,\Phi-[\,\Xi^-]^{-2}\,\Phi) \\ \frac{1}{\gamma}(\Phi^\top[\,\Xi^+]^{-2}-\Phi^\top[\,\Xi^-]^{-2}) & \frac{1}{\gamma}(\Phi^\top([\,\Xi^+]^{-2}+[\,\Xi^-]^{-2})\Phi)+B^{-2}) \end{bmatrix}.
$$

The linear system for the Newton descent directions reads

$$
\begin{bmatrix} \nabla_r^2 & \nabla_{r\,\beta} \\ \nabla_{r\,\beta}^\top & \nabla_\beta^2 \end{bmatrix}\begin{bmatrix} d_r \\ d_\beta \end{bmatrix} = -\begin{bmatrix} \nabla_r \\ \nabla_\beta \end{bmatrix}.
$$

Note that ∇_r^2 is diagonal, so it is a cheap operation to resolve for d_r once d_β is known:

$$
d_r = -(\nabla_r^2)^{-1}(\nabla_{r\,\beta}\,d_\beta + \nabla_r).
$$

Plugging this into the second block of the linear system, one obtains

$$
-\nabla_{r\,\beta}^\top(\nabla_r^2)^{-1}(\nabla_{r\,\beta}\,d_\beta + \nabla_r) + \nabla_\beta^2 d_\beta = -\nabla_\beta
$$

which is equivalent to

$$
(\nabla_\beta^2 - \nabla_{r\,\beta}^\top(\nabla_r^2)^{-1}\nabla_{r\,\beta})d_\beta = -\nabla_\beta + \nabla_{r\,\beta}^\top(\nabla_r^2)^{-1}\nabla_r.
$$

In order to solve the linear system, we proceed as for non-negative least squares. The computational cost of this operation is roughly the same, since the sparse structure of $\Phi^\top\Phi$ can still be exploited. For non-negative least squares, re-computation of the Hessian ∇_β^2 only involves a diagonal update, an operation of negligible computational cost. However, for non-negative least absolute deviation, computation ∇_β^2 involves the matrix multiplication $(\Phi^\top([\,\Xi^+]^{-2}+[\,\Xi^-]^{-2})\Phi)$, i.e. essentially a repeated computation of a scaled Gram matrix. In spite of the special structure of $\Phi^\top\Phi$, the computational cost is of the same order as the solution of the linear system even when using a self-written routine for matrix multiplication tailored to the specific structure.

Additional files

Additional file 1: **MALDI-TOF spectra.**

Additional file 2: **ESI spectra.**

Additional file 3: **Evaluation and results.**

Additional file 4: **Overlapping peptide signals.**

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MS and MH devised the methodology as presented in Section "Methods". MS implemented the Bioconductor package, with contributions by RH and MH. The comparative data analysis was performed by RH, MS, MH and AH; RH and AH performed the MASCOT queries. AT developed the experimental design and provided an interpretation of the MS data. TJ and BG conducted the MS experiments and produced the results of the vendor software. All authors read and approved the final manuscript.

Author details
[1]Department of Computer Science, Saarland University, Saarbrücken, Germany. [2]Center for Bioinformatics, Saarland University, Saarbrücken, Germany. [3]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg. [4]Division for Systematic Proteome Research, Institute for Experimental Medicine, Christian-Albrechts-Universität, Kiel, Germany. [5]Institut für Informatik, Johannes-Gutenberg -Universität, Mainz, Germany.

References
1. Mo F, Mo Q, Chen Y, Goodlett DR, Hood L, Omenn GS, Li S, Lin B: **WaveletQuant, an improved quantification software based on wavelet signal threshold de-noising for labeled quantitative proteomic analysis.** *BMC Bioinformatics* 2010, **11:**219.
2. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.** *Nat Biotechnol* 2008, **26:**1367–1372.
3. Renard B, Kirchner M, Steen H, Steen J, Hamprecht F: **NITPICK: peak identification for mass spectrometry data.** *BMC Bioinformatics* 2008, **9:**355.
4. Hoopmann MR, Finney GL, MacCoss MJ: **High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry.** *Anal Chem* 2007, **79:**5620–5632.
5. Gambin A, Dutkowski J, Karczmarski J, Kluge B, Kowalczyk K, Ostrowski J, Poznanski J, Tiuryn J, Bakun M, Dadlez M: **Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures.** *Int J Mass Spectrom* 2007, **260:**20–30.
6. Mantini D, Petrucci F, Pieragostino D, Del Boccio, P, Di Nicola, M, Di Ilio, C, Federici G, Sacchetta P, Comani S, Urbani A: **LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise.** *BMC Bioinformatics* 2007, **8:**101.
7. Noy K, Fasulo D: **Improved model-based, platform-independent feature extraction for mass spectrometry.** *Bioinformatics* 2007, **23:**2528–2535.
8. Kaur P, O'Connor PB: **Algorithms for automatic interpretation of high resolution mass spectra.** *J Am Soc Mass Spectrom* 2006, **17:**459–468.
9. Tibshirani R: **Regression shrinkage and variable selection via the lasso.** *J R Stat Soc Ser B* 1996, **58:**671–686.
10. Du P, Angeletti R: **Automatic deconvolution of isotope resolved mass spectra using variable Selection and quantized peptide mass distribution.** *Anal Chem* 2006, **78:**3385–3392.
11. Tibshirani R: **Regression shrinkage and selection via the lasso: a retrospective (with discussion).** *J R Stat Soc Ser B* 2011, **73:**273–282.

12. Slawski M, Hein M: **Sparse recovery by thresholded non-negative least squares.** In *Advances in Neural Information Processing Systems 24.* Cambridge, Massachusetts: MIT press; 2011:1926–1934.

13. Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt A: **High-accuracy peak picking of proteomics data using wavelet techniques.** *Pac Symp Biocomput* 2006, **11**:243–254.

14. Schulz-Trieglaff O, Hussong R, Gröpl C, Hildebrandt A, Reinert K: **A fast and accurate algorithm for the quantification of peptides from mass spectrometry data.** In *Proceedings of the Eleventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 2007), Volume 11,* Berlin: Springer; 2007:437–487.

15. Zubarev R: **Accurate monoisotopic mass measurements of peptides: possibilities and limitations of high resolution time-of-flight particle desorption mass spectrometry.** *Rapid Commun Mass Spectrom* 1996, **10**(11):1386–1392.

16. Senko M, Beu S, McLafferty F: **Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions.** *J Am Soc Mass Spectrom* 1995, **6**:229–233.

17. Horn DM, Zubarev RA, McLafferty FW: **Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules.** *J Am Soc Mass Spectrom* 2000, **11**:320–332.

18. Lou X, Renard B, Kirchner M, Koethe U, Graf C, Lee C, Steen J, Steen H, Mayer M, Hamprecht F: **Deuteration distribution estimation with improved sequence coverage for HDX/MS experiments.** *Bioinformatics* 2010, **26**:1535–1541.

19. Suits F, Hoekman B, Rosenling T, Bischoff R, Horvatovich P: **Threshold-avoiding proteomics pipeline.** *Anal Chem* 2011, **83**:7786–7794.

20. Boyd S, Vandenberghe L: *Convex Optimization.* New York: Cambridge University Press; 2004.

21. Samuelsson J, Dalevi D, Levander F, Rognvaldsson T: **Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting.** *Bioinformatics* 2004, **20**:3628–3635.

22. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least Angle Regression (with discussion).** *Ann Stat* 2004, **32**:407–499.

23. van de Geer S, Bühlmann P: **On the conditions used to prove oracle results for the Lasso.** *Electron J Stat* 2009, **3**:1360–1392.

24. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning, 2nd Edition.* New York: Springer; 2008.

25. Bruckstein A, Elad M, Zibulevsky M: **On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations.** *IEEE Trans Inf Theory* 2008, **54**:4813–4820.

26. Wang M, Tang A: **Conditions for a unique non-negative solution to an underdetermined system.** In *Proceedings of Allerton Conference on Communication, Control, and Computing, Volume 49,* Piscataway, New Jersey: IEEE Press; 2009:301–307.

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

28. Meinshausen N: *Sign-constrained least squares estimation for high-dimensional regression. Tech. rep.* Oxford University: Department of Statistics; 2012.

29. Meinshausen N, Yu B: **Lasso-type recovery of sparse representations for high-dimensional data.** *Ann Stat* 2009, **37**:246–270.

30. Zhou S: **Thresholding Procedures for high dimensional variable selection and statistical estimation.** In *Advances in Neural Information Processing Systems 22.* Cambridge, Massachusetts: MIT press; 2009:2304–2312.

31. Zhang T: **Some sharp performance bounds for least squares regression with L_1 regularization.** *Ann Stat* 2009, **37**:2109–2144.

32. Donoho D, Johnstone I: **Ideal spatial adaption by Wavelet shrinkage.** *Biometrika* 1994, **81**:425–455.

33. Fan J, Li R: **Variable selection via nonconcave penalized likelihood and its oracle properties.** *J Am Stat Association* 2001, **97**:210–221.

34. Wasserman L, Roeder K: **High-dimensional variable selection.** *Ann Stat* 2009, **37**:2178–2201.

35. Fan J, Guo S, Hao N: **Variance estimation using refitted cross-validation in ultrahigh dimensional regression.** *J R Stat Soc Ser B* 2012, **74**:37–65.

36. Wolski WE, Farrow M, Emde AK, Lehrach H, Lalowski M, Reinert K: **Analytical model of peptide mass cluster centres with applications.** *Proteome Sci* 2006, **4**:18.

37. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**:3551–3567.

38. Friedman J, Hastie T, Tibshirani R: **Regularized paths for generalized linear models via coordinate descent.** *J Stat Software* 2010, **33**:1–22.

39. Hussong R, Tholey A, Hildebrandt A: **Efficient Analysis of Mass Spectrometry Data Using the Isotope Wavelet.** In *COMPLIFE 2007: The Third International Symposium on Computational Life Science, Volume 940(1),* Edited by Siebes APJM, Berthold MR, Glen RC, Feelders AJ. Melville: AIP; 2007:139–149.

40. Hussong R, Gregorius B, Tholey A, Hildebrandt A: **Highly accelerated feature detection in proteomics data sets using modern graphics processing units.** *Bioinformatics* 2009, **25**:1937–1943.

41. Liu X, Inbar Y, Dorrestein P, Wyne C, Edwards N, Souda P, Whitelegge J, Bafna V, Pevzner P: **Decovolution and database search of complex tandem mass spectra of intact proteins.** *Mol Cell Proteomics* 2010, **9**:2772–2782.

42. Tholey A, Pipkorn R, Bossemeyer D, Kinzel V, Reed J: **Influence of myristoylation, phosphorylation, and deamidation on the structural behavior of the N-Terminus of the Catalytic subunit of CAMP-Dependent protein kinase.** *Biochemistry* 2001, **40**:225–231.

43. Reissner K, Aswad D: **Deamidation and isoaspartate formation in proteins: unwanted alterations or surreptitious signals?** *Cell Mol Life Sci* 2003, **60**:1281–1295.

44. Schulz-Trieglaff O, Hussong R, Gröpl C, Leinenbach A, Hildebrandt A, Huber C, Reinert K: **Computational quantification of peptides from LC-MS data.** *J Comput Biol* 2008, **15**:685–704.

45. Du P, Stolovitzky G, Horvatovich P, Bischoff R, Lim J, Suits F: **A noise model for mass spectrometry based proteomics.** *Bioinformatics* 2008, **24**:1070–1077.

46. Bates D, Maechler M: *Matrix: Sparse and Dense Matrix Classes and Methods*; 2009. [R package version 0.999375-21].

47. Davis T: *CHOLMOD: sparse supernodal Cholesky factorization and update/downdate*; 2005. [http://www.cise.ufl.edu/research/sparse/cholmod/].