Data and text mining

Regression analysis and modelling of data acquisition for SELDI-TOF mass spectrometry

Martin Sköld1,*, Tobias Rydén1, Viktoria Samuelsson2, Charlotte Bratt3, Lars Ekblad3, Håkan Olsson3 and Bo Baldetorp3

1Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden, 2Oncological Centre, University Hospital, Klinikgatan 22, SE-221 85 Lund, Sweden and 3Department of Oncology, Lund University, Barngatan 2:1, SE-221 85 Lund, Sweden

Received on August 31, 2006; revised on February 22, 2007; accepted on March 10, 2007

Advance Access publication March 25, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Pre-processing of SELDI-TOF mass spectrometry data is currently performed on a largely ad hoc basis. This makes comparison of results from independent analyses troublesome and does not provide a framework for distinguishing different sources of variation in data.

Results: In this article, we consider the task of pooling a large number of single-shot spectra, a task commonly performed automatically by the instrument software. By viewing the underlying statistical problem as one of heteroscedastic linear regression, we provide a framework for introducing robust methods and for dealing with missing data resulting from a limited span of recordable intensity values provided by the instrument. Our framework provides an interpretation of currently used methods as a maximum-likelihood estimator and allows theoretical derivation of its variance. We observe that this variance depends crucially on the total number of ionic species, which can vary considerably between different pooled spectra. This variation in variance can potentially invalidate the results from naive methods of discrimination/classification and we outline appropriate data transformations. Introducing methods from robust statistics did not improve the standard errors of the pooled samples. Imputing missing values however—using the EM algorithm—had a notable effect on the result; for our data, the pooled height of peaks which were frequently truncated increased by up to 30%.

Contact: martins@maths.lth.se

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The use of mass spectrometry for profiling serum proteome has shown great potential, but there are still many unresolved problems related to the processing of data and quantification of results. In particular, results have shown poor reproducibility over time and across laboratories. For a review of the technology and previous work, see the recent commentary in Coombes et al., (2005a) and references therein.

Much research has been devoted to post-processing of data, such as clustering and classification of samples, while issues related to the pre-processing steps of normalization and baseline subtraction have received less attention and are commonly done on an ad hoc basis. It is often argued that, in order to increase reproducibility and decrease the amount of false discoveries, it is necessary to form a complete understanding of the statistical properties of a SELDI-TOF experiment and the pre-processing steps. This article aims to take a first step in this direction. We have analysed the building blocks of each spectra, the single-shot spectra obtained from each individual shot from the laser. In doing so, a number of more or less known anomalies in the spectra, including machine artefacts and those resulting from the irregular spatial structure of coated chips are revealed. Analysing single-shots also allows us to empirically verify statistical properties of their pooled versions, in particular we find a close to linear relation between mean-level and variance of the response.

Based on our empirical findings, we construct a full statistical model within the framework of heteroscedastic linear regression. This framework allows explicit expressions for a maximum-likelihood estimator and its variance. Our results indicate that the necessary assumptions behind naive methods for classification/discrimination may fail and we outline appropriate transformations of data. The framework also allows straightforward extension to handle censored data (due to limited recording span of the instrument) and introduce robust methods, both are discussed in detail in this article.

This study is important in the search of reliable methods for measuring biomarkers for early detection of e.g. cancer disease.

2 RESULTS AND DISCUSSION

A ProteinChip® array consists of 8 or 16 spots, to which sample and matrix are applied. In a typical configuration, each spot is irradiated \(n_{s/p} = 12\) times at each of \(n_{pos} = 16\) positions, resulting in \(n = n_{s/p} \times n_{pos} = 192\) ‘single-shot spectra’ which define our raw data. (Here, ‘s/p’ in \(n_{s/p}\) stands for ‘shots per
position'). Before peak detection, raw data is typically processed in the following steps.

(1) Pooling: The \( n \) single shots are combined, by averaging, into a ‘raw spectra’ (the default Ciphergen setting also performs a local smoothing). This step is performed by the instrument software and the individual single shots are by default not available for inspection.

(2) Base-line subtraction: The ‘base-line’—consisting of chemical noise (e.g. contribution from matrix molecules) and machine artefacts—is removed, usually through a local constrained smoothing algorithm. Sauve and Speed (2004) use a morphological filter, Coombes et al. (2005b) use wavelets and a monotone local minimum, Tan et al. (2006) use a non-Gaussian mixed-model-based robust smoothing, Malyarenko et al. (2005) use time-series methods.

(3) Normalization: The base-line subtracted spectra are normalized, usually through division with total ion current (area under curve), in order to allow pairwise comparisons.

Our interest in this article lies in the first step, where we aim to introduce a statistical framework that allows straightforward extensions in a number of important directions. It should be noted that we are not considering the problem of base-line subtraction, since we do not attempt to distinguish between chemical noise and proteins. Hence, the result of our analysis may still be in need of further pre-processing through steps 2 and 3 above.

In order to provide a complete understanding of the statistical properties of a SELDI-TOF experiment we have—through non-standard machine settings—extracted all 192 single-shot spectra in replicates and under a number of different conditions. We have in particular studied experiments on blank chips without coating, chips coated with matrix only and chips coated with serum proteins and matrix. For a detailed description of our experiment, see the Appendix. Two of these single-shot spectra from the same spot containing serum proteins and matrix are shown in Figure 1.

As in Malyarenko et al. (2005) we prefer to work on the timescale rather than the mass/charge \((m/z)\) scale. In a single-shot spectrum, intensity measurements are recorded at a range of time points (Fig. 1). We refer to the time parameter as ‘time-of-flight’ (TOF), and in this article we disregard the actual timescale and measure TOF as an integer corresponding to the sequential ordering of observations within the spectra. Hence, the unit of measurement of TOF is 4 ns, corresponding to the time between recordings of the particular instrument used.

The instrument provides integer intensity measurement in the range 0–255 (due to 8-bit analogue-to-digital conversion). In our experiments, the values varied between a constant offset around 80 and the maximum recordable value of 255, providing roughly 200 unique recordable intensity values (Fig. 2). Some of these values were ‘less favoured’ by the instrument though (Fig. 2).

A potential problem comes from misalignments of peaks on the timescale between shots. In a SELDI-TOF experiment, the effect of triggering errors is negligible (2 ns according to ProteinChip System Users Guide, 2000, p. D-2). Instead, the major source of misalignment is the varying thickness of material on the spot, see e.g. Malyarenko et al., (2005), Önnerfjord et al. (1999). In our study, peaks were well aligned for single shots within each position (as also observed in Malyarenko et al., 2005), though a few positions were slightly misaligned with the majority. To make a more quantitative statement, we studied a peak around TOF 7655 on a spot coated with matrix and serum. For each position in which this peak was pronounced (10 positions in total; other positions were almost empty in this region) we searched for the TOF in the range 7645–7665 at which the recorded value was maximal.

![Fig. 1. Two single shots from Figure 3.](image)

![Fig. 2. Stemplot of the number of recordings of each intensity value out of the total 192×13470 recorded values from one spot (top frame). The bottom frame shows part of the top frame. The values 87 and 89 are recorded roughly 4×10⁵ and 2×10⁴ times, respectively, while there are only 76 recordings of 88.](image)
We then performed a two-way ANOVA of these 10 × 12 values, with ‘position’ and ‘single-shot index’ as covariates. In this ANOVA, 73% of the total sum of squares in peak location was explained by differences between positions and only 7% by differences within position. Although this result should not be taken too literally (errors are certainly not normal, for instance), it clearly indicates that peak misalignment is primarily caused by differences between positions. When comparing the alignment across 128 positions (aligning positions using maximal cross-correlation), we found a maximum shift of five TOF between any two positions. Since this value is small in comparison to the observed width of peaks, we choose not to take the influence from misalignment into account in the present study. Note also that since variation in alignment mainly occurs between positions, using data at position-level seems crucial in order to be able to increase resolution of spectra by correcting for misalignment.

Blank chips The blank chips should ideally give a constant spectrum, and deviations from a constant can be attributed to electronic noise from the equipment and variations in lab environment. This noise showed to be relatively low, with a standard deviation around 0.60 and increasing to 1 with a period of roughly 800 TOF (3.2 ms; Fig. 1 in Supplementary Material). This could be an artefact related to the clock similar to that observed by Baggerly et al. (2003), however they reported a period of 4096 with a different instrument. A corresponding change in mean level was not observed. The noise was slightly coloured, with a first-lag auto-correlation of around 0.35 quickly vanishing for higher lags. It should be noted that Malyarenko et al. (2005), who also analysed blank single-shots from a PBS II instrument, seem to record a higher standard deviation than we did (cf. Fig. 2A in their article) possibly due to different instrument settings.

Coated chips The top two frames of Figure 3 show heat maps of the log single shots recorded at two spots coated with serum proteins and matrix. It is clear from the banded structure that we are not recording 2 × 192 independent and identically distributed single shots. There is a strong dependence both within each position (group of 12 shots) and within each spot, confirming a large spatial variation in the number of successfully bound/emitted ionic species on the surface of the chip. This variation is a well-known problem with techniques that use the so-called dried-droplet method of preparing chips, see e.g. Önnerfjord et al. (1999). At some positions (e.g. the third position in the left frame, corresponding to shots 25–36), the shots are practically empty and at others (e.g. the fourth position in the left frame, corresponding to shots 37–48) the intensities have been censored/truncated at the maximum recording value of 255 (cf. Fig. 1). The censoring is especially troublesome since it will bias the high peaks downwards and low-abundance peaks upwards in the pooled and normalized spectra. Moreover, based on the default instrument output we cannot tell whether censoring has taken place or not (unless all 192 single shots are censored at some specific TOF). We will later discuss how this could be corrected using the EM algorithm.

We did not observe the shelf-like increase in baseline after a detector overload event reported in Malyarenko et al. (2005).
A natural assumption is that, given a total of \( N_i \) charged particles recorded by the detector, the expected value of \( Y_{it} \) is proportional to \( N_i \), i.e.

\[
E(Y_{it}|N_i) = s_i N_i,
\]

where \( s_i \) is the unnormalized proportion of particles that, on the average, hit the detector at TOF \( i \) (unnormalized since we do not know the unit of measurement of \( y_{it} \)). Further, based on this assumption, the area under the curve (AUC) of the \( i \) th single-shot equals

\[
x_i = \sum_{t=1}^{T} y_{it} = N_i \sum_{t=1}^{T} s_i + \sum_{t=1}^{T} (y_{it} - s_i N_i) \approx N_i \sum_{t=1}^{T} s_i,
\]

where the final step follows by the central limit theorem; the second sum is of order \( \sqrt{T} \) which is small in comparison with the first sum which is of order \( T \). The above two equations now define a linear regression problem with zero intercept through

\[
y_{it} = p_i x_i + \varepsilon_{it},
\]

with \( p_i = s_i / \sum_{t=1}^{T} s_i \) the unknown proportion of interest and \( \varepsilon_{it} \) representing a zero-mean random variation. The relation (3) might seem naive in the sense that it assumes the same shape of baseline for all positions. However, since we are only concerned with the very first step of the pre-processing, we will allow it to form the basis for our further investigations.

There are good reasons to assume heteroscedasticity, i.e. that \( \sigma^2_{i} = \text{Var}(\varepsilon_{it}) \) depends on the mean level \( \mu_{it} = p_i x_i \). A plausible assumption is that \( Y_{it} \) is proportional to a Poisson random variable, since we would expect the number of charged particles recorded by the detector in a small time interval to follow a Poisson distribution (which in turn could be viewed as an approximation of a multinomial likelihood). This implies that variance is proportional to the mean, i.e. \( \sigma^2_{it} = c p_i x_i \), where \( c \) can be interpreted as the absolute contribution of a single charge to the recording (cf. below). Note that this variance model ignores any additive instrument noise unrelated to charges hitting the detector, i.e. it assumes \( \sigma^2_{it} = 0 \) when \( p_i = 0 \).

While this is not entirely correct, our analysis of blank shots suggests that additive noise is small and hence we choose to ignore it. Nevertheless, care should be taken when interpreting results for very small mean levels.

By assuming \( x_i \) to be roughly constant over each position, an assumption supported by Figure 3, we can verify the relation between mean and variance by plotting log-empirical position means and variances. That is, plotting pairs \((\log(m_j), \log(v_j))\) where

\[
m_j = \frac{1}{n_{ij}/p} \sum_{t \in I_j} y_{it} \quad \text{and} \quad v_j = \frac{1}{n_{ij}/p} \sum_{t \in I_j} (y_{it} - m_j)^2,
\]

\( I_j \) is the set of indices that constitute the \( j \) th position \((I_j = (12(j-1) + 1, \ldots, 12j) \) in our experiments\) and \( n_{ij}/p \) the number of single shots at each position \( (12 \) in our experiments). This is done for a few selected TOFs, corresponding to visible ‘peaks’, in Figure 5. Here, the observed linear relation suggests

---

**Fig. 5.** Plots of log-empirical within-position means against log-empirical within-position variances for four different TOF’s (TOF 7335 = \( m/z \) 5903, TOF 7655 = \( m/z \) 6432, TOF 8406 = \( m/z \) 7762, TOF 9256 = \( m/z \) 9418). Note the effects of the discrete nature of measurements in the lower left ends of the plots. Groups that contain censored values are represented by crosses. Straight lines represent the line \( x = y \). The points in the figure correspond to the measurements at each position of eight spots with the same serum.
a model of the form $\sigma^2 = c(p_0x_i)^\alpha$. To verify that $\alpha = 1$, we have performed an individual linear regression analysis of log-empirical variance against log-empirical mean for each TOF. The resulting estimated slopes (i.e. $\alpha$s) are shown in Figure 6. It seems as if $\alpha$ is fairly stable across the time axis, and only slightly larger than 1. However, the intercept log($c$) clearly depends on TOF, suggesting that the contribution to the recording made by a single particle varies with time. This may be due to multiple charged matrix molecules hitting the detector, an explanation supported by the shape of log($c$) as a function of TOF which resembles the spectrum baseline. Another plausible explanation is charge accumulation in the detector as discussed in Malyarenko et al. (2005). In order to check whether the varying intercept is due to the removal of a constant baseline only, we repeated the analysis after first (for each position) removing a time-varying baseline computed using a local median smoother. For this case, neither slopes nor intercepts were found to be constant as functions of TOF.

It is interesting to note that the commonly used approach to pooling of taking the average at each TOF divided by the total mass average is equivalent to maximizing a Poisson likelihood. That is, assuming $Y_i \sim Po(p_0x_i)$ and solving

$$\hat{p}_i = \arg \max_x \sum_{i=1}^n \left\{ y_{it} \log(p_{xi}) - p_{xi} - \log(y_{it}) \right\} = \frac{\sum_{i=1}^n Y_{i t}}{\sum_{i=1}^n x_{it}}$$

(5)

where $n$ denotes the number of single shots to be pooled (192 in our experiment). In the bottom two frames of Figure 3, we have displayed the pooled versions of the heat maps in the same figure. We remark that the use of a Poisson likelihood only makes sense for non-negative integer-valued observations. However, writing $y_{it} = c\tilde{y}_{it}$ and $x_i = cN_i$, where $\tilde{y}_{it}$ and $N_i$ represent the total number of proteins recorded at TOF $t$ and over the entire single-shot, respectively [cf. Equation (1)] and $c$ is a constant given by the instrument, and assuming that $\tilde{y}_{it}$ is an observation from a Poisson distribution with mean $p_{ti}N_i$, leads to a similar likelihood with the same maximizer $\hat{p}_i$ (irrespective of $c$).

In the model formulation $x_i = cN_i$ etc., it holds that $E(Y_i | x_i) = p_{ti}x_i$ and $\text{Var}(Y_i | x_i) = c p_{ti} x_i$. We also note that a normal approximation, $\tilde{y}_{it} \sim N(p_{ti}x_i, \tilde{c}_{xi})$ for some (unknown) constant $\tilde{c}_i$ unrelated to $p_i$ leads to the MLE $\hat{p}_i$ as well. This approximation will later be used to take censoring of values that are truncated at the maximum recordable 255 into account.

2.1.1 Variance of regression estimators Using the variance assumption $\text{Var}(Y_i | x_i) = c p_{ti} x_i$ and independence of the $Y_{it}$ across single-shot indices $i$, we can derive the variances of the estimator $\hat{p}_i$ as

$$\text{Var}(\hat{p}_i) = \text{Var} \left( \frac{\sum_{i=1}^n Y_{it}}{\sum_{i=1}^n x_{it}} \right) = \frac{\sum_{i=1}^n c p_{ti} x_{it}}{\left( \sum_{i=1}^n x_{it} \right)^2} = \frac{c p_0}{\sum_{i=1}^n x_{it}}.$$  

(6)

It is important to note that this variance depends on $\sum_{i=1}^n x_{it}$, the sum of AUCs. Moreover, our initial data analysis suggested that this value can vary greatly from spot to spot. For example, spots A and B displayed in Figure 3 differ by a factor 7 in their sum of AUCs.

In order to empirically verify the relation between pooled spectra variance and $\sum_{i=1}^n x_{it}$, we have also estimated the variance using a non-parametric bootstrap resampling approach. This approach is based on the assumption that single-shot spectra are statistically exchangeable within each position and that positions (groups of single-shots within a position) are exchangeable within a spot (see Appendix for a description of the algorithm). From the resampled spots, we can then empirically estimate the standard deviation of $\hat{p}_i$ and compare with the theoretical value suggested by (6). We do this by plotting pairs $\left( \sqrt{\hat{p}_i^2 / \sum_{i=1}^n x_{it}}, \sigma^* \right)$, where $\sigma^*$ denotes the bootstrap standard deviation for a particular configuration of positions and $\hat{p}_i^2 / \sum_{i=1}^n x_{it}$ is computed from the same configuration. In Figure 7, we show the result for the previously analysed four TOFs. The observed linear relations confirm the relation between sum of AUC and pooled spectrum variance. However, the variation in slope of the lines suggests that $c$ is not constant over the time scale (c.f. the plot of log($c$) in Fig. 6). Note in particular, as predicted by theory, the large difference in variance between spots.

2.1.2 Variance and baseline subtraction The above section provides explicit formulae for the variance of $\hat{p}_i$, the pooled spectra with only a constant baseline subtracted. Since the shape of the baseline tend to vary between spectra, it is usually necessary to remove also a time-variable baseline in order to facilitate pair wise comparison of protein content. Furthermore, baseline-subtracted spectra need to be renormalized. Denoting estimated baseline by $\hat{b}_i$, it is thus of interest to also derive the variance of $\hat{p}_i - \hat{b}_i$ (1 - $\sum_i \hat{b}_i$). Here we argue that since most methods used for estimating the baseline $\hat{b}_i$ are based on averaging over a large range of TOF, the variance of $\hat{b}_i$ will be negligible in comparison to that of $\hat{p}_i$.  

Fig. 6. Estimates of $\alpha$ (top) and log($c$) (bottom) based on ordinary linear regression on log-mean/log-variance plots. Measurements from eight spots with the same serum were used in the construction of the figure. Segments of high variability, in particular below 2000 and above 9000 TOF, correspond to areas where most spectra are practically empty and hence provide little information on $\alpha$ and log($c$).
Returning to the problem of testing for an equal mean, let $\hat{p}_h^k$ and $\hat{p}_d^k$ be the weighted averages of estimates from the healthy and diseased, respectively, $A^k = \sum_{i \in I_h} A_k$ and $A^d = \sum_{i \in I_d} A_i$. According to the null model $\hat{p}_{k;i} \sim N(p_i, c/p_i/\hat{A}_k)$ for all $k$, and hence $\hat{p}_h^k - \hat{p}_d^k \sim N(0, \frac{c}{A^h + A^d}/1 + \frac{1}{A^d})$. Furthermore, $Q_h = \sum_{k \in I_h} A_k (\hat{p}_{k;i} - \hat{p}_{h;i})^2$ has the distribution of $c_i$ times a $\chi^2$-distribution with $N_i - 1$ degrees of freedom (d.f.) and is independent of $\hat{p}_d^k$; corresponding statements hold true for the diseased group. As a result, the pooled variance $\hat{c}_i = (Q_h + Q_d)/f$, where $f = N_h + N_d - 2$, is an unbiased estimator of $c_i$, and under the null hypothesis $T = (\hat{c}_i/A^h + \hat{c}_i/A^d)^{-1/2}(\hat{p}_h^k - \hat{p}_d^k)$ has a $t$-distribution with $f$ d.f. In conclusion, the null hypothesis is rejected if $|T| > t_{1-\alpha/2}(f)$, where $t_{1-\alpha/2}(f)$ is the upper $(1 - \alpha/2)$ quantile of the $t$ distribution with $f$ d.f. and $\alpha$ is the level of the test.

2.3 Taking censoring into account

Censored values are generally observed around the matrix-hump and at peaks of the most abundant proteins. While the problem could be alleviated by reducing the laser intensity, this seems difficult due to the extreme variations in intensity observed at different TOFs of the spectra. In our experiment, the total number of censored values in each spot (a spot consisting of 192 x 13470 values) varied from 229 to 25 776 corresponding to Chip II Spots B and A displayed in Figure 3. The top frame of Figure 9 shows the proportion of censored values at each TOF for the latter spot.

By assuming the normal model $Y_{i;k} \sim N(p_i x_i, \hat{c}_i x_i)$, the missing/censored values can be imputed and the maximum-likelihood estimate computed by the EM algorithm; for a full description, see the Appendix. In Figure 8, we have plotted pooled spectra from eight different spots over a range where, for some spots, censoring was frequent. For the spots with several censored values, the peak appears both higher and ‘sharper’ when censoring is properly taken into account. The bottom frame of Figure 9 examines the observed relative increase in the pooled spectra when the EM algorithm is applied. We see that the pooled spectrum can be as much as 30% larger when using the EM algorithm to take censoring into account.

Note that censoring also affects the sum of AUCs. While ignoring censoring will underestimate the height of censored peaks in the spectra it will overestimate uncensored peaks after normalization. In our experiment, the largest observed increase in sum of AUCs taking censoring into account was 2%. Hence, ignoring censoring would overestimate the height of uncensored peaks by roughly the same amount for this spectra.

2.4 Robust regression

Based on the variance model $x_i^T \vartheta_i \propto x_i$ or $\sigma_i^2 \propto p_i x_i$, one may compute normalized residuals $(y_{i;k} - \hat{p}_i x_i)/x_i^{1/2}$ where $\hat{p}_i$ is an estimate of $p_i$. Given either variance model, these residuals should roughly have the same variance. In a plot of such residuals, it is however typical that a few have magnitudes notably larger than the other ones and may therefore be thought of as outliers. Thus, it is natural to ask if a robust regression procedure can improve the performance of the estimates.
MLE is, as noted above, given by
\[ p \]
Ordinary least squares linear regression, corresponding to
\[ p \]
Robust regression,
to large residuals under the variance model
\[ p \]
Two fixed values for
\[ p \]
Poisson likelihood regression, corresponding to
\[ p \]
model:
Here we compare a
\[ p \]
visual inspection of the top frame in Figure 6.

\[ p \]
we attempted a 'robustification’ of its MLE. For \( \alpha = 1 \) the MLE is, as noted above, given by \( \hat{p}_1 \), and in general it is
\[ p \]
Our robust method, estimators denoted \( \hat{p}^{2,\alpha} \), is based on reweighted least squares (see Appendix) and gives low weights to large residuals under the variance model \( \sigma_x \propto x^\beta \). We tried two fixed values for \( \alpha \); \( \alpha = 1 \) and \( \alpha = 1.2 \), the latter based on visual inspection of the top frame in Figure 6.

### 2.4.1 Comparison of regression methods

Here we compare a number of different regression methods based on the proposed model:

- **M1** Poisson likelihood regression, corresponding to \( \hat{p}_1 \).
- **M2** Robust regression, \( \alpha = 1 \), corresponding to \( \hat{p}^{2,1} \).
- **M3** Robust regression, \( \alpha = 1.2 \), corresponding to \( \hat{p}^{2,1.2} \).
- **M4** Ordinary least squares linear regression, corresponding to \( \hat{p}_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} \).

As a measure of performance, we estimated the variance \( \text{Var}(\hat{p}_i) \) of the estimators using a bootstrap algorithm described in the Appendix. The results showed that all methods perform comparably and that there was no improvement in introducing robust methods. For a graphical display, see Figure 3 in the Supplementary Material.

### 3 CONCLUSIONS

By studying single-shot spectra, we have observed a large variation in the amount of successfully emitted ionic species in a SELDI-TOF experiment. This variation was found both between positions within the same spot and between spots. In order to accurately take this variation into account, we have constructed a statistical framework for pooling single-shot spectra over a spot. This framework involves heteroscedastic linear regression, and gives a simple expression for the variance of a pooled spectra as a function of the sum of AUCs. We also observe that this value can show considerable variation between spots. Hence, the sum of AUCs is an important covariate to take into account when comparing pooled spectra. Apart from providing statistically motivated estimators, using a full statistical model allow straightforward extensions in a number of directions.

For example, it allows a simple modification of the standard two-sample \( t \)-test that would otherwise be incorrect due to the differing variances of pooled spectra. It also allows taking instrument censoring of high peaks into account in a straightforward manner using the EM algorithm. When applying this to peaks corresponding to high-abundance proteins and peaks in the area of high baseline, we obtained estimates of peak heights that were up to 30% larger than what was obtained when censoring was ignored. We attempted at introducing robust methods in order to reduce variance of the estimators. However, in an empirical comparison of the variances of robust and non-robust estimators we found them to be similar. This gives some indication that the observed high variance is not due to a few outlying observations, since such values would be ignored by the robust methods.

Using a linear regression framework also has the advantage of being straightforwardly extended to a multivariate context. Here, extra covariates may involve e.g. instrument settings in order to combine/compare experiments under possibly different settings or done by different labs.

Finally, we would like to remark that this study is only a first step towards a full statistical understanding of a SELDI-TOF experiment. Further work is needed in order to accommodate the fact that relative height of baselines may vary between shots and take into account bias introduced in the baseline-subtraction step. We believe that this analysis need to be done at the resolution of a position on the spot and that much important information is lost in the current default instrument output of spot averages. We also believe that extensions of our framework to other platforms like MALDI-TOF would be straightforward, provided the spot-reading protocol is defined in a similar manner.

### ACKNOWLEDGEMENTS

This work has been supported by grants from the Swedish Research Council, The Royal Physiographic Society in Lund, IngaBritt and Arne Lundberg Foundation, Gunnar Nilsson.
Cancer Foundation, Hospital of Lund Foundations, Mrs. Berta Kamprad Foundation, King Gustav V Jubilee Foundation and Franke and Margaretha Bergqvist Foundation.

Conflict of Interest: none declared.

REFERENCES


Coombes,K.R. et al. (2005b) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics, 5, 4107–4117.


REFERENCE

M.Sköld et al.

APPENDIX

Description of experiment

The spectra were recorded in a PBS-II from Ciphergen (Freemont, CA) with the following settings: high mass 20000, optimized from 2000 to 20000, laser intensity 180, detector sensitivity 7, focus mass 8000 and mass deflector 1500. In all experiments, the protein chip CM10 with a weak cationic coating was used. Every position was warmed by two shots with laser intensity 210 after which twelve single shots were recorded. In this way, 16 positions were analysed, producing 192 spectra from each spot. The blank chips were analysed without pre-treatment. To the matrix chips, 1 μL 100% sinapinic acid in 50% acetonitrile and 0.5% trifluoroacetic acid was added per spot and left to air-dry for 15 min before analysis. The serum proteins from healthy volunteers were solubilized in 9 parts of 9 M urea, 2% CHAPS and 50 mM Tris–HCl, pH 9, and then adsorbed to the protein chip spots in 0.1 M sodium acetate, pH 4.0, (CM low stringency buffer) using a Biomek2000 robot (Beckman Coulter, Inc., Fullerton, CA) equipped with a bioprocessor from Ciphergen according to the instructions from the manufacturer. In this way, proteins from 1 μL serum were applied to each spot. The protein chips were air dried for 20 min and then matrix was added as above.

Removal of constant baseline

Before analysis, the constant baseline was estimated as the most abundant value in the higher region of each spectrum and subtracted from data. This value turned out to be identical for all single shots in our experiment (i.e. 86).

Robust regression algorithm

The robust procedure is based on reweighted least squares, and proceeds as follows for any fixed t:

(i) Choose a starting value \( p_{t}^{(0)} \), e.g. \( p_{t}^{(0)} = \frac{x_{t}^{2}}{\sum_{i=1}^{n} x_{i}^{2}} \).

(ii) Repeat for \( k = 0, 1, 2, \ldots \) until convergence.

\( (k.1) \) Compute residuals \( r_{t}^{(k)} = (y_{t} - p_{t}^{(k)} x_{t})/\sqrt{\sum_{i=1}^{n} x_{i}^{2}} \), compute an estimate \( \hat{c}_{t}^{(k)} \) of their scale \( c_{t} \) as their median absolute deviation (MAD) and compute weights \( w_{t}^{(k)} = W(r_{t}^{(k)}/(c_{t}^{(k)})^{\tau}) \), where \( W \) is a weight function and \( \tau \) is a tuning parameter. We used Tukey’s bisquare weight function \( W(r) = (1 - r^{2})^{2} \) for \( |r| < 1 \) and \( \tau = 4.685 \).

\( (k.2) \) Compute the reweighted estimate

\[
\hat{p}_{t}^{(k+1)} = \frac{\sum_{i=1}^{n} y_{t} x_{i}^{1-t} w_{t}^{(k)}}{\sum_{i=1}^{n} x_{i}^{2-t} w_{t}^{(k)}}.
\]

We notice that the regression step combines the variance model \( \sigma_{y}^{2} \propto x_{t}^{2} \) with the weights. The weight functions assign low weights to observations with large residuals, which hence may suspicious to be outliers.

A hierarchical bootstrap analysis

In order to evaluate the performance of the regression methods, we followed a bootstrap resampling approach. It is important to note that direct sampling with replacement of the single-shots within each spot is not appropriate due to the strong dependence between AUCs within each position (cf. Fig. 3).

Thus, we adopted a hierarchical approach based on the assumption that shots are statistically exchangeable within each position (group of single-shots) and that positions are exchangeable within each spot. Hence, resampling from a spot containing \( n \) single-shots grouped in \( n_{\text{pos}} \) positions of \( n_{\text{pos}} \) shots each proceeds as follows (as before \( I_{k} \) denotes the set of indices of spectra at the \( k \) th position):

(i) Draw bootstrapped position indices \( \tilde{i}_{1}, \ldots, \tilde{i}_{n_{\text{pos}}} \) independently from the uniform distribution on the set \( \{1, \ldots, n_{\text{pos}}\} \).

(ii) For each \( k \in \{1, \ldots, n_{\text{pos}}\} \) and each \( i \in I_{k} \), draw the bootstrapped single-shot \( y_{j}^{*} \) uniformly among the original shots at position \( \tilde{i}_{k} \), that is, uniformly from the set \( \{y_{i,j} \mid j \in I_{k}\} \).

This procedure gives a set \( y^{*} = (y_{1}^{*}, \ldots, y_{n}^{*}) \) of \( n \) resampled single-shots, and a corresponding set \( x^{*} = (x_{1}^{*}, \ldots, x_{n}^{*}) \) of resampled totals.

In Fig. 7, each point (there are 100 points for each TOF and spot) in the figure corresponds to one iteration of step (i). The theoretical variance \( \sigma_{y}^{2}/\sum_{i=1}^{n} x_{i}^{2} \) was computed from the drawn position indices, keeping the original single shots within each position. For each set of position...
indices (points in the figure), we repeated step (ii) 100 times, and these 100 resampled spots were used to empirically estimate \( \sigma^* \).

We repeated the procedure 500 times [i.e. one iteration of each of (i) and (ii)] for each of the two spots analysed and for each bootstrap sample computed the estimates corresponding to M1–M4 over the observed range of TOFs. These values were then used to empirically compute bootstrap variances.

**EM algorithm**

By assuming a normal likelihood, \( Y_t \sim N(p_t x_i, \epsilon_t x_i) \), the model with censored data can be viewed as a missing data problem. As before, we denote by \( y_t \) the observed measurements after constant baseline subtraction. We further let \( m_t \) be the number of censored values and re-order in such a way that the first \( m_t \) values, \( y_{1t}, \ldots, y_{m_t t}, \) are censored. Denoting by \( Z_{1t}, \ldots, Z_{m_t t} \) the corresponding unobserved ‘true intensities’, we have that \( Z_{it} \) is distributed as \( Y_t | Y_t > y_t \), where \( Y_t \) is distributed as \( N(p_t x_i, \epsilon_t x_i) \). The EM algorithm for computing the MLE of \( p_t \) now proceeds as follows:

(i) Choose starting values \( p_t^{(0)} \) and \( \epsilon_t^{(0)} \), e.g.

\[
p_t^{(0)} = \frac{1}{n} \sum_{i=1}^n y_{it} / \sum_{i=1}^n x_i.
\]

(ii) Repeat for \( k = 0, 1, 2, \ldots \) until convergence. Compute

\[
p_t^{(k+1)} = \frac{\sum_{i=1}^{m_t} M_t^{(k)} + \sum_{i=m_t+1}^n y_{it}}{\sum_{i} x_i}
\]

and

\[
\epsilon_t^{(k+1)} = \frac{1}{n} \left( \sum_{i=1}^{m_t} V_t^{(k)} + n \sum_{i=m_t+1}^n y_{it} \right) /
\]

where

\[
m_t^{(k)} = E(Z_{it}|p_t^{(k)} \epsilon_t^{(k)})
= p_t^{(k)} x_i + (\epsilon_t^{(k)} x_i)^{1/2} H(y_t - p_t^{(k)} x_i)(\epsilon_t^{(k)} x_i)^{1/2}
\]

and

\[
V_t^{(k)} = E(Z_{it}^2|p_t^{(k)} \epsilon_t^{(k)})
= (p_t^{(k)} x_i)^2 + \epsilon_t^{(k)} x_i
+ (\epsilon_t^{(k)} x_i)^{1/2}(y_t + p_t^{(k)} x_i)
\times H(y_t - p_t^{(k)} x_i)(\epsilon_t^{(k)} x_i)^{1/2}.
\]

Here, \( H(x) = \phi(x)/(1 - \Phi(x)) \) and \( \phi, \Phi \) are the standard normal density and distribution functions, respectively. Note that we have ignored the effect of censoring on the \( x_i \) since this seems negligible.

**AUTHORS’ CONTRIBUTIONS**

T.R., V.S. and M.S. developed the statistical methodology; T.R. and M.S. wrote the main part of the article; L.E. and C.B. performed the SELDI-TOF experiments and took part in the data collection; L.E., C.B. and B.B. contributed to the experimental design; B.B. and H.O. contributed reagents/materials/analysis tools and critically revised the manuscript. All authors read and approved the final manuscript.