An active set approach to the elastic-net and its applications in mass spectrometry

Theodore Alexandrov, Oliver Keszöcze, Dirk A. Lorenz, Stefan Schiffler, Klaus Steinhorst

An active set approach to the elastic-net and its applications in mass spectrometry

Theodore Alexandrov, Oliver Keszöcze, Dirk A. Lorenz, Stefan Schiffler and Klaus Steinhorst
University of Bremen
Bibliotheksstraße 2, D-28359 Bremen, Germany
Telephone: (0049) 421–218 63822
E-Mail: schiffi@math.uni-bremen.de

***Abstract*—This paper uses the framework of a Mass Spectrometry application to introduce a new method of peak picking as well as two active set methods for the minimization of the elastic-net-functional.**

The application of peak picking is essential in mass spectrometry and is often based on mean spectra. In contrast our procedure uses a set of spectra obtained from a basis learning method. Our procedure utilizes the well known ℓ^1-minimization and corresponding active set algorithms but comprises ill conditioned operators such that regularization is required. We show, that the elastic-net gives a natural justification for Tikhonov-Philips-regularization in the used algorithms. Therefore we introduce adaptions of known active set algorithms for ℓ^1-minimization to the elastic net. Furthermore, we emphasize the differences of the algorithms for ℓ^1 and the elastic-net in numerical examples.

I. Introduction

Mass spectrometry (MS) is an important technique for profiling of chemical substances. In medicine, MS contributes to clinical research and the diagnosis by identification of biomarkers (relevant peptides) of a disease. This problem is usually solved by classifying the spectra of ill and healthy individuals and identifying the peptides corresponding to the features (peaks) used in the classification. Thus, the problem of peak picking is essential in MS analysis and still attracts considerable attention (see e.g. [10]) since the conventional methods are still far from ideal.

We developed a new procedure of peak picking based on a sparse coding algorithm. The idea is that an MS-spectrum y can be modeled as a linear combination of a set of basis spectra. To find suitable basis spectra we use the efficient sparse coding algorithm from [9] which captures the succinct features common for a set of spectra. The features found are represented in a set of few basis vectors (or basis spectra) $\{\varphi_i \in L^2 : i \in \mathbb{N}\}$. The basis spectra contain the peaks which are in evidence of the majority of the spectra processed.

Second, considering the basis spectra found, we determine the positions of their peaks with a simple method. The procedure is similar to a peak picking using mean spectra but instead of taking mean spectrum, we consider a sequence of the basis spectra. The new procedure was motivated by recent successes of sparsity-oriented algorithms in many fields, in particular in MS [10] [2].

Our procedure uses the linear syntheses operator $K : \ell^2 \to L^2$

$$Kx = \sum_{i \in \mathbb{N}} x_i \varphi_i$$

and minimizes the ℓ^1-functional

$$\Psi_\alpha(x) = \frac{1}{2}\|Kx - y\|^2 + \alpha\|x\|_1$$

for the given MS-data y. In contrast to conventional methods this approach significantly increases the number of correctly detected peaks. In practice an ill conditioned operator K may be problematic for various ℓ^1-minimization algorithms. Thus, in our application, we often need to use further regularization.

There is a wide range of well known algorithms to minimize Ψ_α, e.g. iterated thresholding [3], gradient projection [6], fixpoint continuation [8] and many more. All of them perform differently well on various applications. We found that especially the active set methods – semismooth Newton method (SSN) [7] and feature sign search algorithm (FSS) [9] – perform favorably for our application. Both algorithms exploit the optimality condition

$$K^*Kx \in K^*y - \alpha\,\mathrm{Sign}(x)$$

for the minimization of Ψ_α on a so called active set. On this active set the operator K^*K has to be inverted. In the case of the peak picking problem the operator K^*K is often close to singular on the active set and hence, the inversion is unstable. Thus, we use Tikhonov-Phillips regularization in this step, that is replacing the inversion of K^*K by the inversion of $K^*K + \beta\,\mathrm{id}$ with a small $\beta > 0$.

In this paper we will give a natural justification of Tikhonov-Phillips regularization for active set algorithms minimizing Ψ_α. We will show that regularization can be interpreted as adding a penalty term to Ψ_α, which is known as the elastic-net-functional (cf. [11], [4]):

$$\Phi_{\alpha,\beta}(x) = \frac{1}{2}\|Kx - y\|^2 + \alpha\|x\|_1 + \frac{\beta}{2}\|x\|^2$$

Using sub-differential calculus the optimality condition for this functional turns out to be

$$0 \in \partial\Phi_{\alpha,\beta}(x) = \partial\Psi_\alpha(x) + \beta x. \tag{1}$$

The similarity of the optimality conditions for classical ℓ^1- and elastic-net-minimization yields a straight forward adaption of

existing ℓ^1-algorithms. We will derive adaptions of the SSN [7] and FSS [9] and show that these algorithms are regularized versions of the ℓ^1-algorithms. In the end we give numerical results for the MS-application.

II. Regularized SSN (RSSN)

We will now derive a semismooth Newton based method for the elastic-net-functional $\Phi_{\alpha,\beta}$. Afterwards we see that the method coincides with the described regularization of the SSN [7] and hence will be called RSSN.

We start with a more precise formulation of (1) which is

$$-K^*(Kx-y)-\beta x \in \alpha \operatorname{Sign}(x) \tag{2}$$

with the set valued Sign function. This condition can be formulated equivalently with the help of the soft-shrinkage function

$$(S_\alpha(x))_i = \begin{cases} 0, & |x_i| \le \alpha \\ x_i-\alpha, & x_i > \alpha \\ x_i+\alpha, & x_i < -\alpha \end{cases}, \quad i \in \mathbb{N}.$$

Lemma. *A vector x solves (2) if and only if*

$$F(x) := \beta x - S_\alpha(-K^*(Kx-y)) = 0. \tag{3}$$

Proof: We reformulate (2) to

$$-\frac{K^*(Kx-y)}{\beta} \in x + \frac{\alpha}{\beta}\operatorname{Sign}(x).$$

Since the soft-shrinkage function is given by $S_\alpha = (\mathrm{id} + \alpha \operatorname{Sign})^{-1}$ (see, e.g. [7]) this leads to

$$x = S_{\alpha/\beta}(-K^*(Kx-y)/\beta).$$

Finally, using $S_{c\alpha}(cx) = cS_\alpha(x)$ for $c>0$ proves the claim. ∎

From [7] we know the following:

Lemma. *A Newton derivative of S_α is given by*

$$G(x) = \begin{pmatrix} \mathrm{id}_{\{k\in\mathbb{N}:|x_k|>\alpha\}} & 0 \\ 0 & 0 \end{pmatrix}$$

and for any linear operator $T : \ell^2 \to \ell^2$ and any $b \in \ell^2$ a Newton derivative of $S_\alpha(Tx+b)$ is given by $G(Tx+b)T$.

Hence, we obtain a Newton derivative of F:

$$D(x) = \beta\,\mathrm{id} - G(-K^*(Kx-y))K^*K$$

According to G, we can split K^*K for a given set $A \subset \mathbb{N}$ into

$$K^*K = \begin{pmatrix} M_A & M_{AA^c} \\ M_{A^cA} & M_{A^c} \end{pmatrix}$$

and hence for $A_x := \{k \in \mathbb{N} : (|K^*(Kx-y)|)_k > \alpha\}$ we find that

$$D(x) = \begin{pmatrix} \beta id_{A_x} + M_{A_x} & M_{A_x A_x^c} \\ 0 & \beta id_{A_x^c} \end{pmatrix}.$$

Applying the semi-smooth Newton iteration

$$x_{k+1} = x_k - D(x_k)^{-1}F(x_k)$$

to find solutions of (3) a straight forward calculation leads to the RSSN:

1) Initialize: $k=0$, $x_0 = 0$
2) Choose active set:
   $A_{x_k} = \{n \in \mathbb{N} : |K^*(Kx_k - y)|_n > \alpha\}$
   and calculate
   $$s_n^k = \begin{cases} 1, & [-K^*(Kx_k-y)]_n > \alpha \\ -1, & [-K^*(Kx_k-y)]_n < -\alpha \\ 0, & \text{else} \end{cases}.$$
3) Update:
   $$x_{k+1}|_{A_{x_k}} = (\beta\,\mathrm{id} + M_{A_{x_k}})^{-1}(K^*y - s^k\alpha)|_{A_{x_k}}$$
   $$x_{k+1}|_{A_{x_k}^c} = 0$$
4) Check stop criteria: Return x_{k+1} or set $k \leftarrow k+1$ and continue at step 2.

This is exactly the SSN [7] with regularization. Note that the next iterate x_{k+1} depends on x_k only through the active set.

Remark. Multiplying (2) by $\gamma > 0$ and adding x we obtain

$$x - \gamma K^*(Kx-y) - \gamma\beta x \in x + \gamma\alpha \operatorname{Sign}(x) \tag{4}$$

and hence we have another characterization of a minimizer of $\Psi_{\alpha,\beta}$:

$$x - S_{\gamma\alpha}(x - \gamma K^*(Kx-y) - \gamma\beta x) = 0.$$

This equation leads to a similar algorithm but with a different active set, namely:

$$A_x^1 = \{n \in \mathbb{N} : |x - \gamma K^*(Kx-y) - \gamma\beta x|_n > \gamma\alpha\}$$

A third choice of the active set can be derived from (4) by rewriting it as

$$x - \gamma K^*(Kx-y) \in (1+\gamma\beta)x + \gamma\alpha\operatorname{Sign}(x)$$

leading to

$$(1+\gamma\beta)x - S_{\gamma\alpha}(x - \gamma K^*(Kx-y)) = 0.$$

Hence, the third choice of the active set is

$$A_x^2 = \{n \in \mathbb{N} : |x - \gamma K^*(Kx-y)|_n > \gamma\alpha\}$$

These choices may affect the convergence behavior.

III. Regularized FSS (RFSS)

In the same fashion we can derive a modification of the FSS [9] for elastic net-minimization. The method is similar to the SSN [7] but uses a more modest active set strategy. The basic framework of the RFSS is the following:

1) Initialize: $k=0$, $x_0 = 0$, $A = \emptyset$
2) Increase the active set by one element i_0 $A \leftarrow A \cup \{i_0\}$
3) Update:

$$x_{k+1}|_A = (\beta\,\mathrm{id} + M_A)^{-1}(K^* y \pm \alpha)|_A$$
$$x_{k+1}|_{A^c} = 0$$

4) If necessary: Remove indices from the active set, set $k \leftarrow k+1$ and continue at step 3.
5) Check stop criteria: Return x_{k+1} or set $k \leftarrow k+1$ and continue at step 2.

The full details of the algorithm would be beyond the scope of this paper and can be found in [1]. Now we will shortly discuss the differences of FSS [9] compared to RFSS.

Step 2: This is a greedy step based on the optimality condition of Ψ_α for FSS [9] or $\Phi_{\alpha,\beta}$ for RFSS resp. These conditions are exploited on A^c – which is the set of all indices i where $x_{k,i} = 0$ – and hence they coincide.

Step 3: In contrast to FSS [9] the RFSS regularizes the equation. This is the only essential difference between both algorithms.

Step 5: might differ slightly for FSS [9] and RFSS if one uses the optimality condition of Ψ_α or $\Phi_{\alpha,\beta}$ resp. as stopping criteria.

IV. Analytical properties

First of all the results proved for SSN in [7] and FSS in [9] also hold for the regularized versions, see [1] for more details. Hence we have

Lemma. *The RSSN converges locally super linearly.*

Lemma. *In the finite dimensional case i.e. $K : \mathbb{R}^n \to \mathbb{R}^n$ the following holds: RFSS converges globally in finitely many steps. Moreover every iteration strictly reduces the value of $\Phi_{\alpha,\beta}$.*

Another interesting property is the following

Lemma. *The minimizers $x_{\alpha,\beta}$ of $\Phi_{\alpha,\beta}$ depend continuously on α and β, that is for $\alpha, \beta > 0$:*

$$\lim_{(\alpha_n,\beta_n)\to(\alpha,\beta)} x_{\alpha_n,\beta_n} = x_{\alpha,\beta}$$

V. Numerical experiments

We compared the procedures using RFSS and FSS applying them to the problem of peak picking for the following simulated spectra. The data set consists of two classes each of 25 spectra of length 110. Indeed, the data set simulates not the whole spectra but only parts thereof. The spectra of each class have three peaks at prespecified positions with slightly variable height. The white gaussian noise is added together with small randomly distributed spurious peaks. An example of simulated data is shown in Figure 1. The peak picking is done for all 50 spectra taken altogether without using information about the spectra class.

Given the data set as a matrix with spectra in columns, we use FSS for calculating the matrix of coefficients x. Then, given the coefficients, we learn the basis vectors φ_i using the Lagrange dual, see [9] for more details.

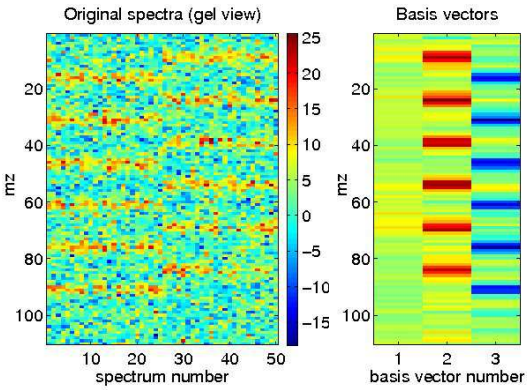

Fig. 1. The simulated data showed in gel view **Was ist "gel view"?** (left) and the basis vectors φ_i learned with $\beta = 10^{-10}$ (right).

β	Acc.	#FP	#BV	#It.	skips
–	95.8%	4.25	5.8	4	5 of 9
1	90.7%	5.6	57.2	14.1	–
10^{-5}	98.1%	4.6	5.9	3	–
10^{-10}	100%	4.2	5.3	2.3	–

TABLE I
The summary characteristics of processing of the simulated data with procedures using FSS (without β) and RFSS using different values of β.

We measured the following characteristics: the accuracy of the procedure, i.e. the ratio of numbers of correctly found peaks to all peaks, denoted in Table I as "Acc."; the number of false positive peaks or not peaks detected as peaks, "#FP"; the number of basis vectors φ_i found, "#BV"; the number of iterations of the FSS or RFSS resp., "#It.". The characteristics have been calculated for different β values and fixed α values nine times, Table I presents the averaged results. For the procedure using FSS we also show the number of runs when the FSS cannot be applied because of ill-conditioning of the operator matrix, "skips". One can see on Figure 1 how typical basis vectors learned from the data look like and that they seem to represent the peaks of the spectra well.

As can be seen in Table I FSS fails often and already a small regularization, i.e. a small β, suffices to rectify this problem. It is remarkable that even if FSS does not fail the RFSS gives better algorithmic results, that is higher sparsity while taking fewer iterations. It also illustrates our heuristic experience that it is desirable to choose beta as small as possible such that the RFSS still converges. One can see that RFSS provides not only better algorithmic results but also the procedure based on RFSS works better since the accuracy is higher and the number of false positives is lower.

VI. Efficient implementation

Since in RFSS the active set changes in a modest way we can use this to efficiently implement step 3. Our observations from numerical experiments are that in many applications the active set is growing by one element in almost every iteration (Step 2). In contrast the shrinking step (Step 4) is performed

rarely. Hence, the symmetric matrix $(\beta\,\mathrm{id} + M_{A_x})$ in the equation in step 3 changes only by one row and one column in almost every iteration. This fact can be exploited using the Cholesky factorization (CF) which has some benefits:

1) The CF factorizes a symmetric positive definite Matrix M as
$$M = L \cdot L^*$$
with a lower triangular matrix L. If L is known the triangular form makes it computationally easy to solve equations of the type $Mx = b$.
2) It can be easily updated if the matrix grows by one row and one column only. In combination with the first item and the active set strategy this gives us a fast method for the calculations in step 3.
3) One only has to keep a triangular matrix in memory (L instead of M) which reduces the memory needs to almost one half. Anyway the computational effort for calculating Mx or $L \cdot L^* x$ is roughly the same if one uses the triangular structure.
4) It is known to be robust for solving equations $Mx = b$ with almost singular matrices M.

VII. Conclusion

This paper discussed regularized versions of the semismooth Newton method and the feature-sign-search method. It has been shown that these regularization indeed minimize the elastic net functional. This functional now involves two regularization parameter: α for the sparsity constraint and β for the Tikhonov-Phillips regularization of the update step. Hence, both parameters has to be chosen in some way. In this task it is important to notice that the corresponding minimizer depends continuously on both parameters. Moreover we suggest to choose β as small as possible while the choice of α is still a delicate issue.

References

[1] Theodore Alexandrov, Bangti Jin, Dirk Lorenz, and Stefan Schiffler. Active set methods for the elastic-net – a regularization for l1-minimization. preprint, 2009.

[2] Kristian Bredies, Theodore Alexandrov, Jens Decker, Dirk Lorenz, and Herbert Thiele. Sparse deconvolution for peak picking and ion charge estimation in mass spectrometry. *Submitted to Proc. ECMI'08*, 2008.

[3] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications in Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[4] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *ArXiv e-prints*, 807, July 2008.

[5] David L. Donoho and Yaakov Tsaig. Fast solution of ℓ^1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

[6] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Applications to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 4:586–597, 2007.

[7] Roland Griesse and Dirk A. Lorenz. A semismooth Newton method for Tikhonov functionals with sparsity constraints. *Inverse Problems*, 24(3):035007 (19pp), 2008.

[8] Elaine T. Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for l1-minimization: Methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008.

[9] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *Proceedings of the Neural Information Processing Systems (NIPS)*, volume 19, 2006.

[10] B.Y. Renard, M. Kirchner, H. Steen, J.A. Steen, and F.A. Hamprecht. Nitpick: peak identification for mass spectrometry data. *BMC Bioonf*, 9(1):355, 2008.

[11] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B*, 67:301–320, 2005.