JOANNEUM RESEARCH
HEALTH

Automated XCMS parameter optimization

A new approach

Libiseller Gunnar[1], Dvorzak Michaela[3], Gander Edgar[1], Kleb Ulrike[3], Narath Sophie H.[1], Pieber Thomas[1,2], Sinner Frank M.[1,2], Sourij Harald[2], Magnes Christoph[1]

JOANNEUM RESEARCH
HEALTH

1
JOANNEUM RESEARCH
Forschungsgesellschaft mbH

HEALTH
Institute for Biomedicine
and Health Sciences

Christoph Magnes

Elisabethstrasse 5
8010 Graz, Austria

Phone +43 316 876-40 00
Fax +43 316 8769-40 00

health@joanneum.at
www.joanneum.at/health

Medical University of Graz

2
Medical University of Graz

Clinic of Internal Medicine
Division of Endocrinology and
Metabolism

Graz, Austria

JOANNEUM RESEARCH
POLICIES

3
JOANNEUM RESEARCH
Forschungsgesellschaft mbH

POLICIES
Centre for Economic
and Innovation Research

Graz, Austria

References

[1]
Eliasson, M., Rännar, S., Madsen, R., Donten, M.A., Marsden-edwards, E., Moritz, T., Shockcor, J.P., et al. Strategy for Optimizing LC-MS Data Processing in Metabolomics: A Design of Experiments Approach. Anal. Chem, 84(15), 6 869 – 6 876 (2012)

[2]
Smith, C.A. and Want, E.J. and O'Maille, G. and Abagyan, R. and Siuzdak, G.: XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, Analytical Chemistry, 78 : 779 – 787 (2006)

[3]
Ralf Tautenhahn, Christoph Boettcher, Steffen Neumann: Highly sensitive feature detection for high resolution LC / MS BMC Bioinformatics, 9 : 504 (2008)

[4]
H. Paul Benton, Elizabeth J. Want and Timothy M.D. Ebbels Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data Bioinformatics, 26 : 2488 (2010)

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project Met2Net

Introduction

Untargeted metabolomics based on LC / MS relies on automated data processing, such as peak detection, peak picking, retention-time correction or annotation. The software tools available for peak detection and peak picking have many parameters for adjusting algorithms to data set characteristics like MS resolution, chromatographic peak geometry or background noise. The best possible results can rarely be achieved by using the default parameters. Hence, a structured automated workflow to optimize these parameters would be advantageous.

A workflow has been developed that consecutively optimizes all parameters of the subprocedures[1], and tested in XCMS by using a diluted series of pooled samples[2,3,4].

Aim

Since it is now common to periodically analyze QC samples in LC / MS metabolomic studies, we tried to build a workflow for parameter optimization based on these samples.

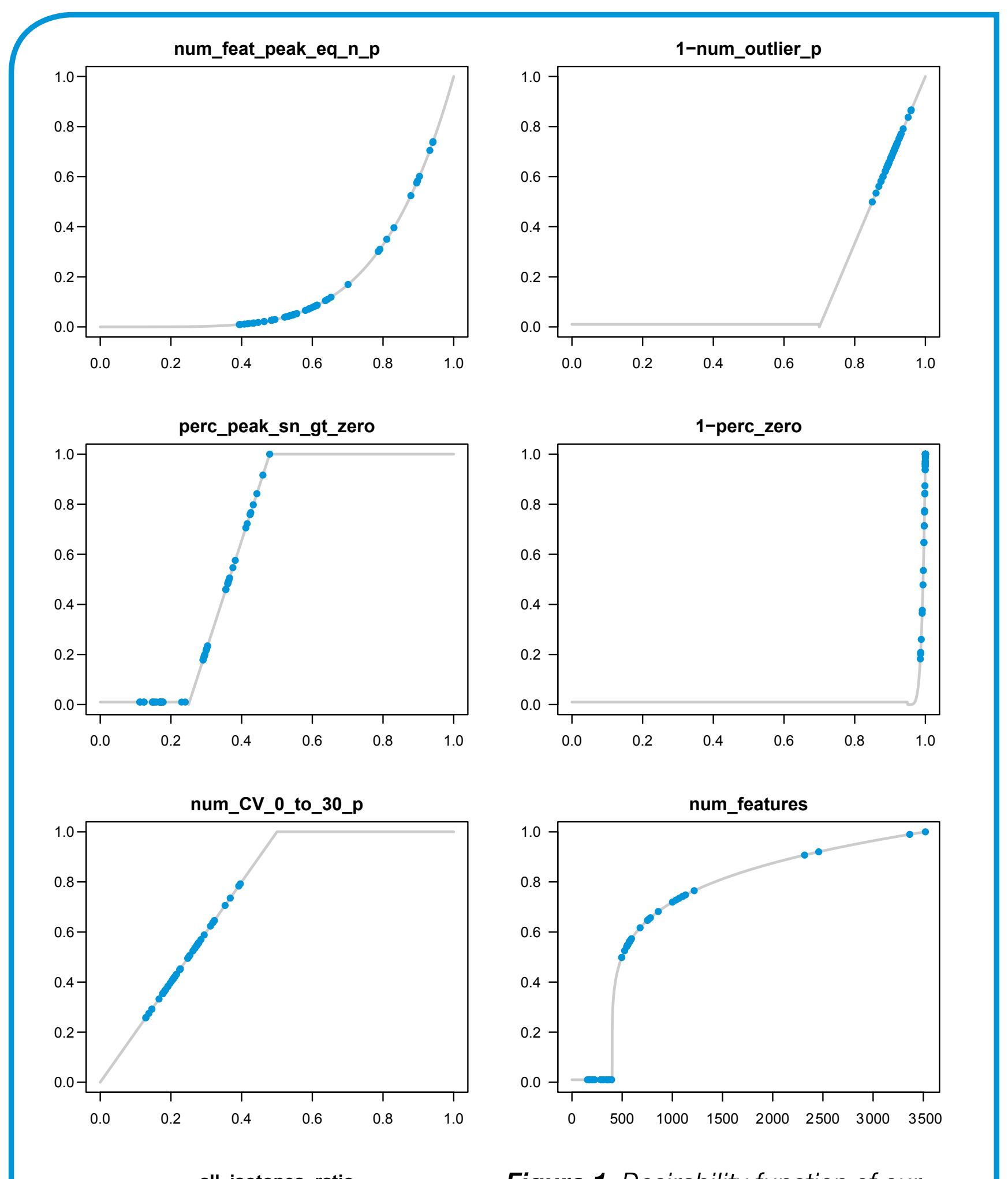

Figure 1: *Desirability function of our target values. These were combined to find the optimal parameters. We transformed the target values to be able to maximize all of them. The functions reflect how we wanted our target values to behave. Some are punished when producing low values, some have a linear or exponential increase from some point and so on.*

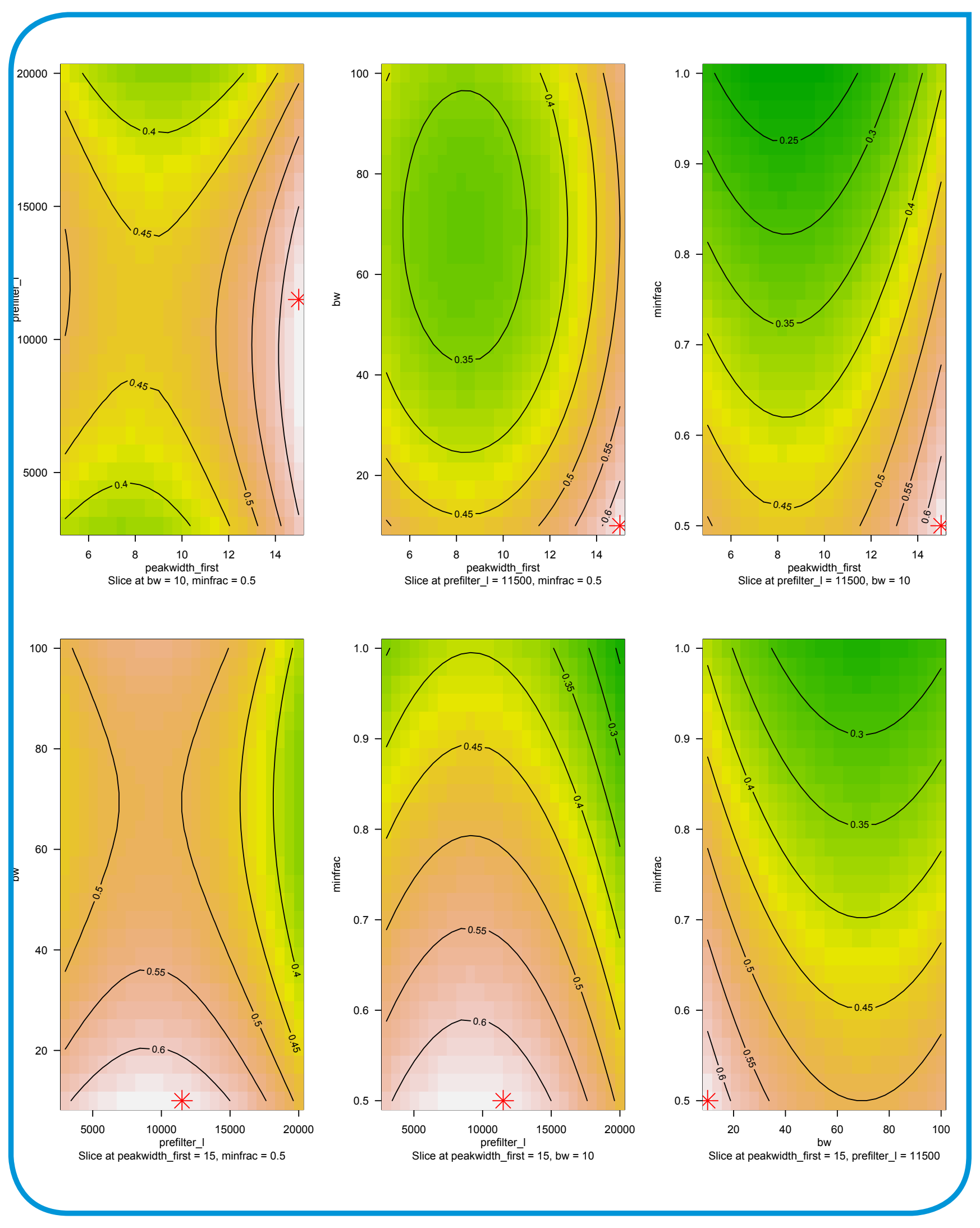

Figure 2: *The response surface model is the result of the optimization step. The remaining parameters peakwidth_first, prefilter_l, bw and minfrac where plotted against each other. The stars indicate the optimal values. For this LC/MS experiment, and the desirability function we choose, the best values would be: peakwidth_first = 15, prefilter_l = 12,000, bw = 10 and minfrac = 0.5. Since some of them are at the outer limits of the parameter settings, these may not reflect the best values, but a good approximation.*

Methods

- Measurements of 6 QC samples analyzed by LC / MS
- Definition of target values which may describe good results (Table 1)
- Screening
  - Used to identify which parameters could be important to explaining target values
  - Design of experiment (DoE) done by a Plackett-Burman-Plan (58 experiments)
  - Analysis using a linear model
- Modeling
  - Aims to further reduce the number of input parameters
  - DoE by Fractional Factorial design (256 experiments)
  - Analysis using a linear model
- Optimization
  - Creation of a desirability function with the remaining target values
  - DoE by Central-Composite-Design (50 experiments)
  - Estimation of response surface model to find optimal settings

Result

The input parameters were reduced from 21 at the beginning of the screening step to 4 at the optimization step (Table 2). We were also able to better evaluate which of our target values may describe reliable peaks.

- Screening
  - Reduction of 21 input parameters to 13
  - Only seven target values were influenced by the input parameters and not correlated with other target values
- Modeling
  - Reduction of 13 input parameters to 4
  - These 4 input parameters influenced our target values significantly (Table 3)
- Optimization
  - Resulting xcms-sets were used to estimate a response-surface-model (Fig. 2)
  - For this LC / MS-experiment, with this desirability function the optimal parameters would be: peakwidth_first = 15, prefilter_l = 12,000, bw = 10 and minfrac = 0.5

Discussion

After the screening step we had to settle for 'centWave' as peak picking method, 'obiwarp' for retention time correction and 'density' for grouping method. This may not have led to optimal results but otherwise the number of necessary experiments would have been much higher. Nevertheless, these experiments helped us to understand the influence of some parameters on our target values.

Outlook

In the future we will conduct experiments to determine whether the number of C13-isotopes relative to the number of peaks would be a good target value for optimization. Fig. 3 shows some trends of this target value for different xcmsSet-parameters-settings. Furthermore, the different steps for xcmsSet generation will be optimized sequentially. This will allow us to test all available methods for peak picking, retention time correction and grouping and make the workflow even faster.

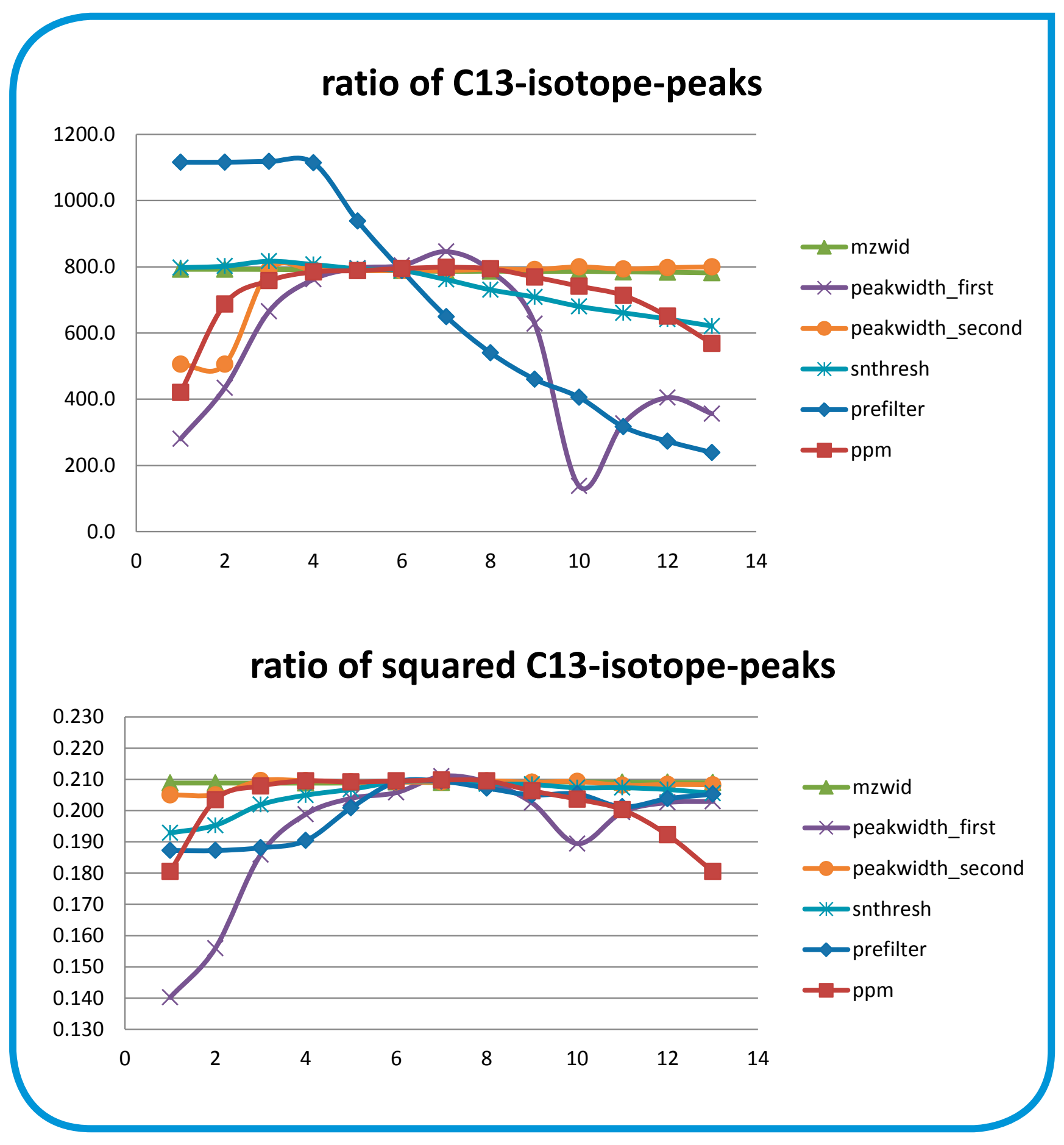

Figure 3: *Trends for different parameter settings for the xcmsSet-method. The upper diagram shows the number of peaks, identified as C13-isotopes, in relation to the number of all peaks; the lower diagram uses the squared number of identified C13-isotopes. We choose these target values, since we think that they might be able to distinguish well between reliable peaks and noise. Some of them also seem to have maxima where an optimization could be beneficial.*

Table 1: *This table shows the target values we had a look at. After the screening step we settled for the 7 most promising.*

Name	Description
num_features	Number of features
num_zeros	Number of peaks with zero intensity
perc_zero	Ratio of peaks with zero intensity relative to num_peaks
num_peaks	Number of peaks
num_peaks_sn_eq_zero	Number of peaks with S/N equal zero, hence from fillPeaks()
num_peaks_sn_gt_zero	Number of peaks with S/N greater zero
perc_peak_sn_gt_zero	Ratio of peaks with S/N greater zero relative to num_peaks
mean_num_peaks_per_feat	Average number of peaks per feature
mean_num_peaks_gt_n	Average number of peaks when less peaks than samples
mean_num_peaks_lt_n	Average number of peaks when more peaks than samples
num_feat_peak_gt_n	Number of features with more peaks than samples
num_feat_peak_lt_n	Number of features with less peaks than samples
num_feat_peak_eq_n	Number of features with as many peaks as samples
std_CV	Standard deviation of coefficients of variation (CV)
mean_CV	Average of CVs
num_CV_0_to_10	Number of CVs ranging from 0 to 10
num_CV_10_to_20	Number of CVs ranging from 10 to 20
num_CV_20_to_30	Number of CVs ranging from 20 to 30
num_CV_30_to_50	Number of CVs ranging from 30 to 50
num_CV_from_50	Number of CVs greater 50
num_outlier	Number of outliers
ratio_C13_isotopes	Ratio of identified C13_isotope_peaks relative to num_peaks
elapsed_runtime	Elapsed runtime of xcmsSet-Generation (just for info)

Table 2: *Tested input parameters. After the screening we decided to use Retcor. obiwarp with no parameters. This reduced the input parameters to 13 and resulted in 256 experiments. For the optimization we only used the four parameters, which influenced our desirability function the most.*

Method	Parameter	Screening	Modeling	Optimization
xcmsSet	ppm	X	X	
	peakwidth_first	X	X	X
	peakwidth_second	X	X	
	snthresh	X	X	
	prefilter_k	X	X	
	prefilter_l	X	X	X
	integrate	X	X	
	mzdiff	X	X	
	profparam_step	X	X	
group	bw	X	X	X
	minfrac	X	X	X
	mzwid	X	X	
	max	X	X	
fillPeaks	fillpeaks	X		
Retcor.obiwarp	profStep	X		
	response	X		
Rector.loess	missing	X		
	extra	X		
	smooth	X		
	span	X		
	family	X		

Table 3: *Influence of the input parameters on our target values. Red background indicates target values to be maximized, blue background indicates minimization. P-values are represented by stars. Blue colored stars, positive algebraic sign: if the input parameter is increased, target value will rise. Red colored stars, negative algebraic sign: inverse behavior of input and output. Last column represents the mean of the other values.*

		num_feat_peak_eq_n_p	num_outlier_p	perc_peak_sn_gt_zero	perc_zero	num_CV_0_to_30_p	num_features	all_isotopes_ratio	D
1	ppm	0	0	***	0	0	***	***	0
2	peakwidth_first	****	****	****	*	****	***	****	****
3	peakwidth_difference	0	0	0	0	0	0	0	0
4	snthresh	**	0	***	0	***	***	0	0
5	prefilter_k	0	0	0	0	0	**	***	*
6	prefilter_l	0	***	0	***	0	****	*	**
7	integrate	0	***	*	0	****	0	***	***
8	mzdiff	0	0	0	0	0	0	0	0
9	profparam_step	0	0	0	**	0	0	*	0
10	bw	****	****	***	***	***	**	0	****
11	minfrac	****	***	****	****	****	****	***	****
12	mzwid	***	0	0	0	0	0	0	0
13	max	0	*	****	***	0	***	0	0
	R²	0.93	0.74	0.93	0.73	0.82	0.78	0.93	0.77

*, **, ***, **** (< 0.05, < 0.01, < 0.001, < 2e-16)