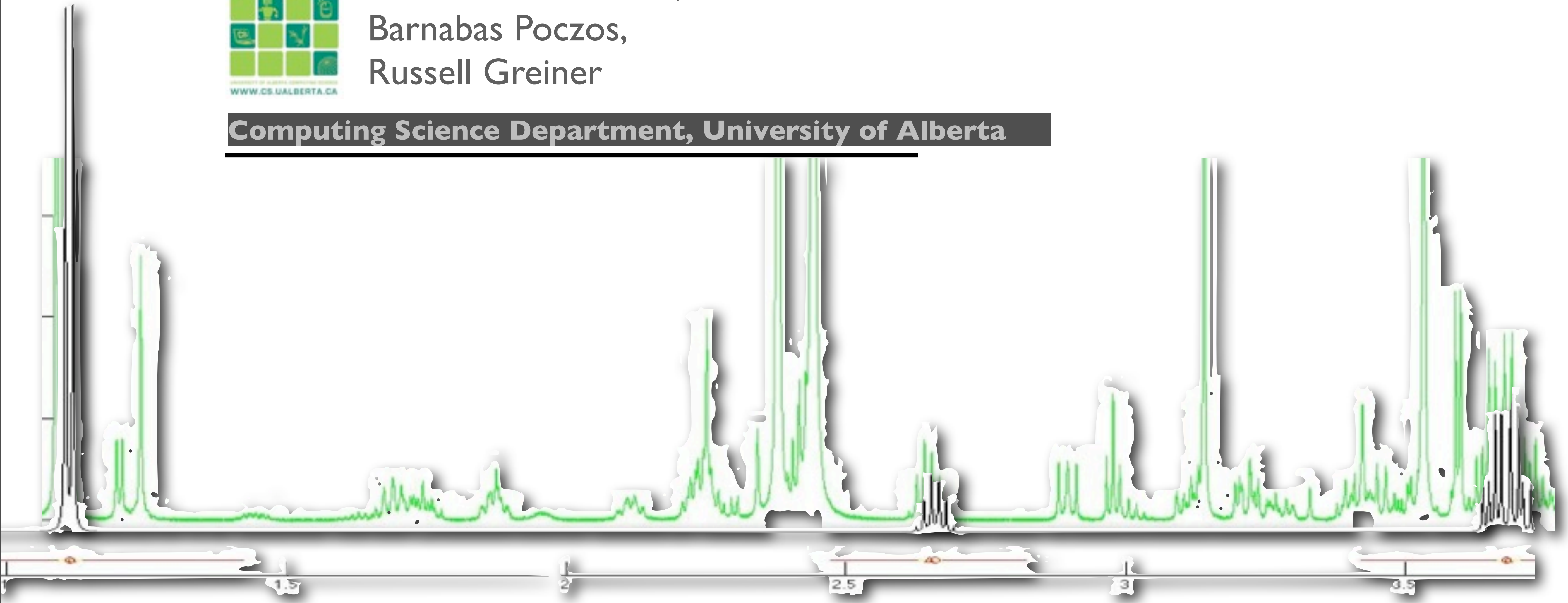
A Cross-Entropy Method that Optimizes Partially Decomposable Problems:

A New Way to Interpret NMR Spectra

Siamak Ravanbakhsh,
Barnabas Poczos,
Russell Greiner

Computing Science Department, University of Alberta



Metabolomics & NMR Spectroscopy

Metabolomics study of chemical fingerprints that cellular processes leave behind

Metabolomics & NMR Spectroscopy

Metabolomics study of chemical fingerprints that cellular processes leave behind

Metabolites: *end products of gene expression*

Metabolomics & NMR Spectroscopy

Metabolomics study of chemical fingerprints that cellular processes leave behind

Metabolites: *end products of gene expression*

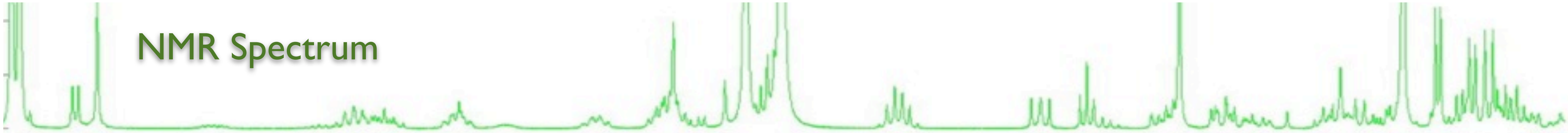
Toolbox:

- Gas chromatography
- High performance liquid chromatography
- Capillary electrophoresis
- Mass Spectrometry
- **Nuclear Magnetic Resonance (NMR) Spectroscopy**

source: wikipedia

Nuclear Magnetic Resonance (NMR) Spectroscopy

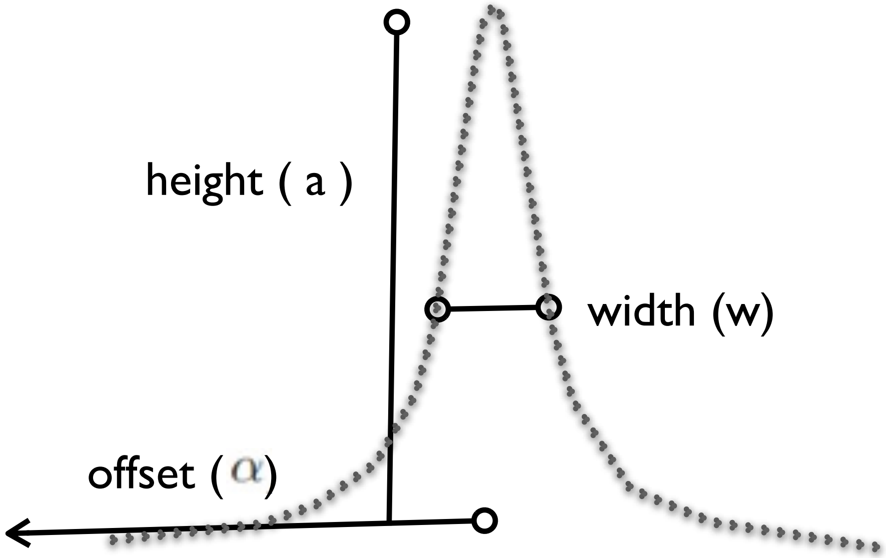
Nuclear Magnetic Resonance (NMR) Spectroscopy



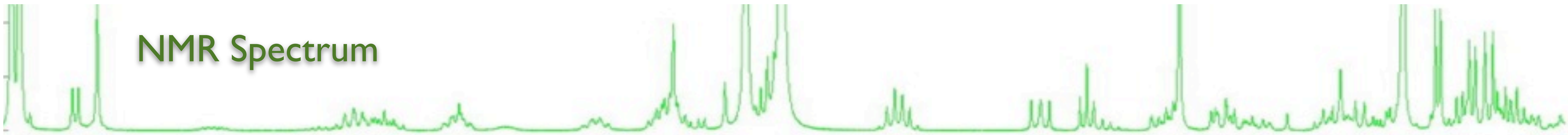
Nuclear Magnetic Resonance (NMR) Spectroscopy

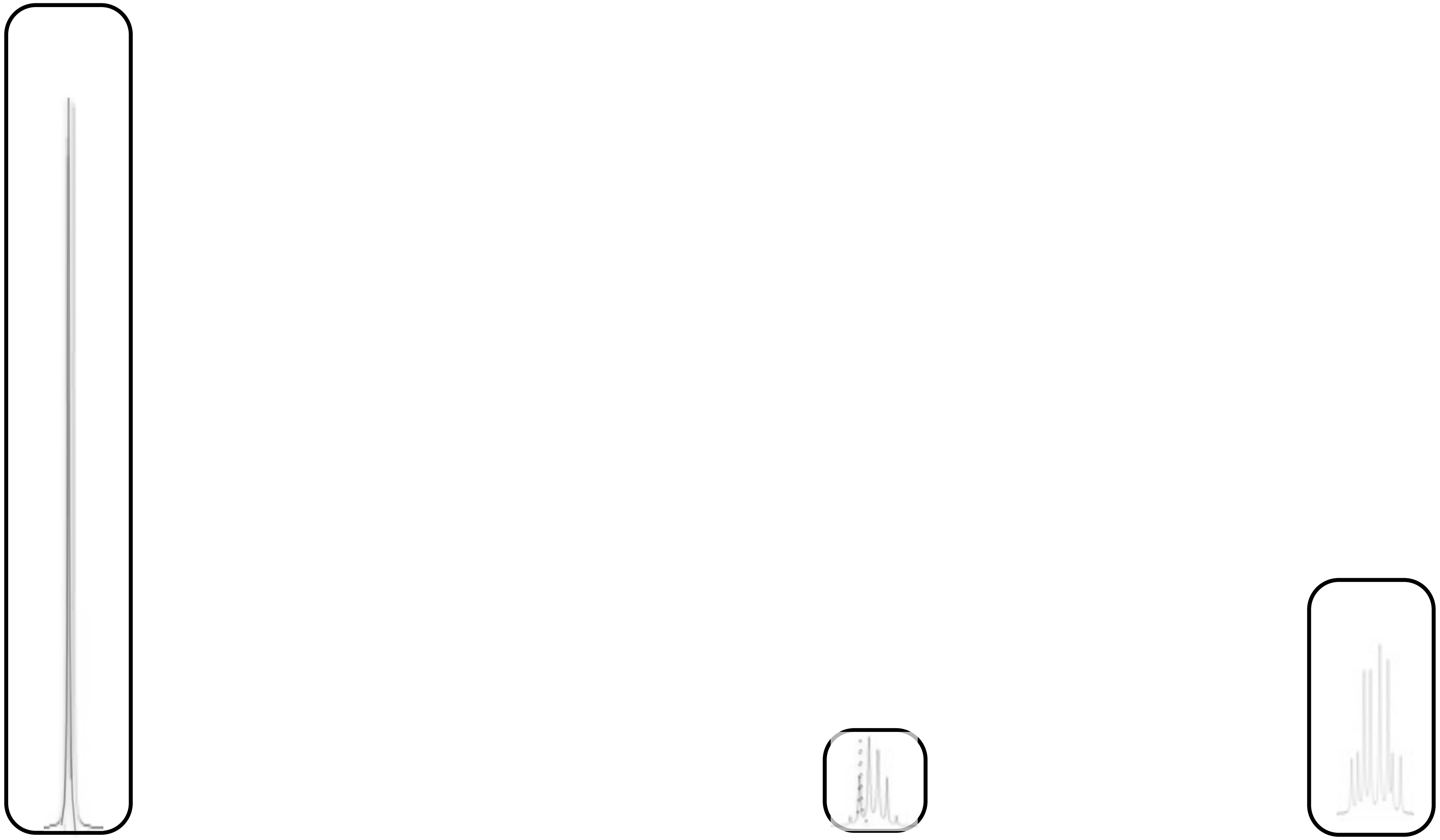
Spectrum is made of many

Lorentzian peaks

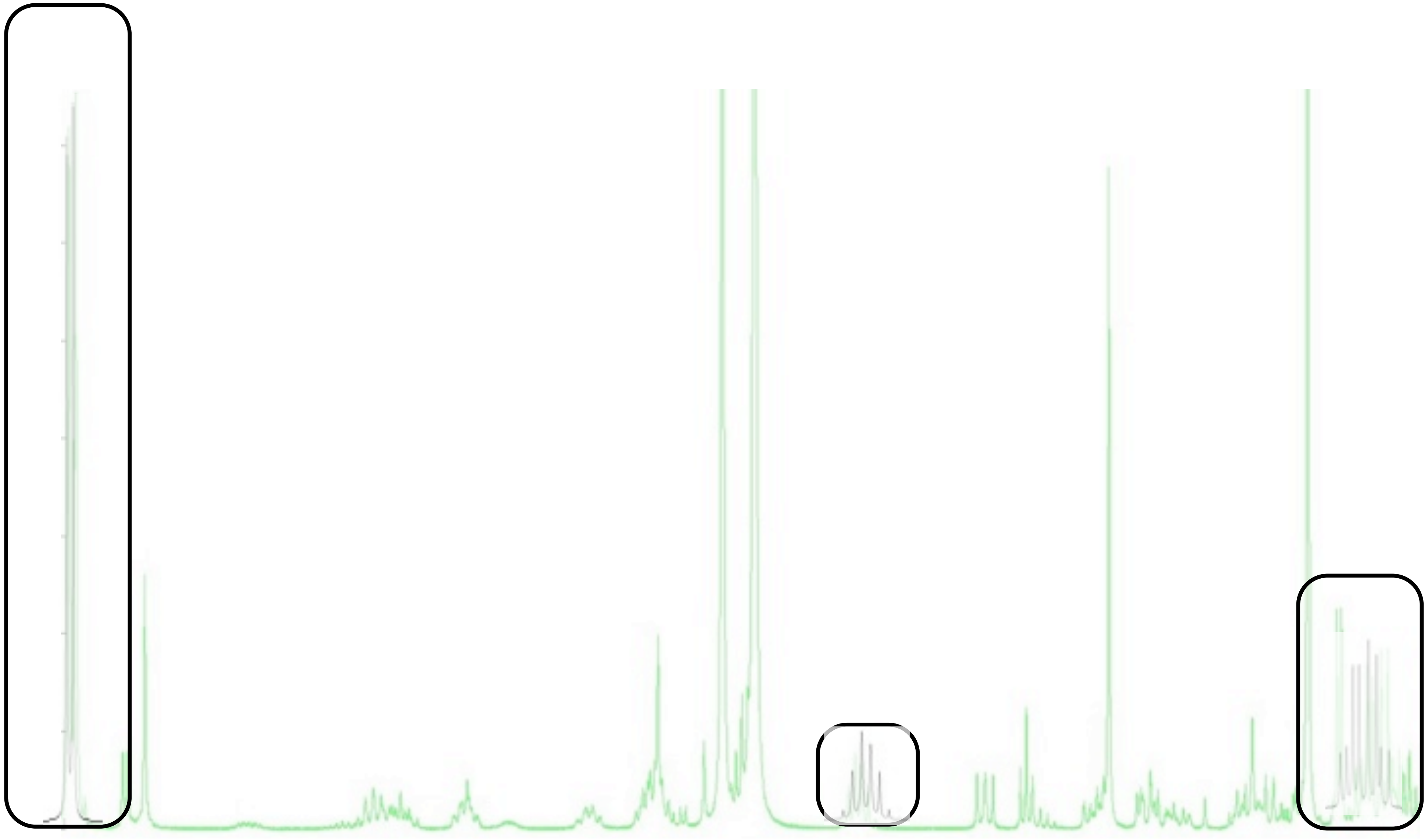


$$S(y) = \frac{aw^2}{w^2 + 4(y-\alpha)^2}$$

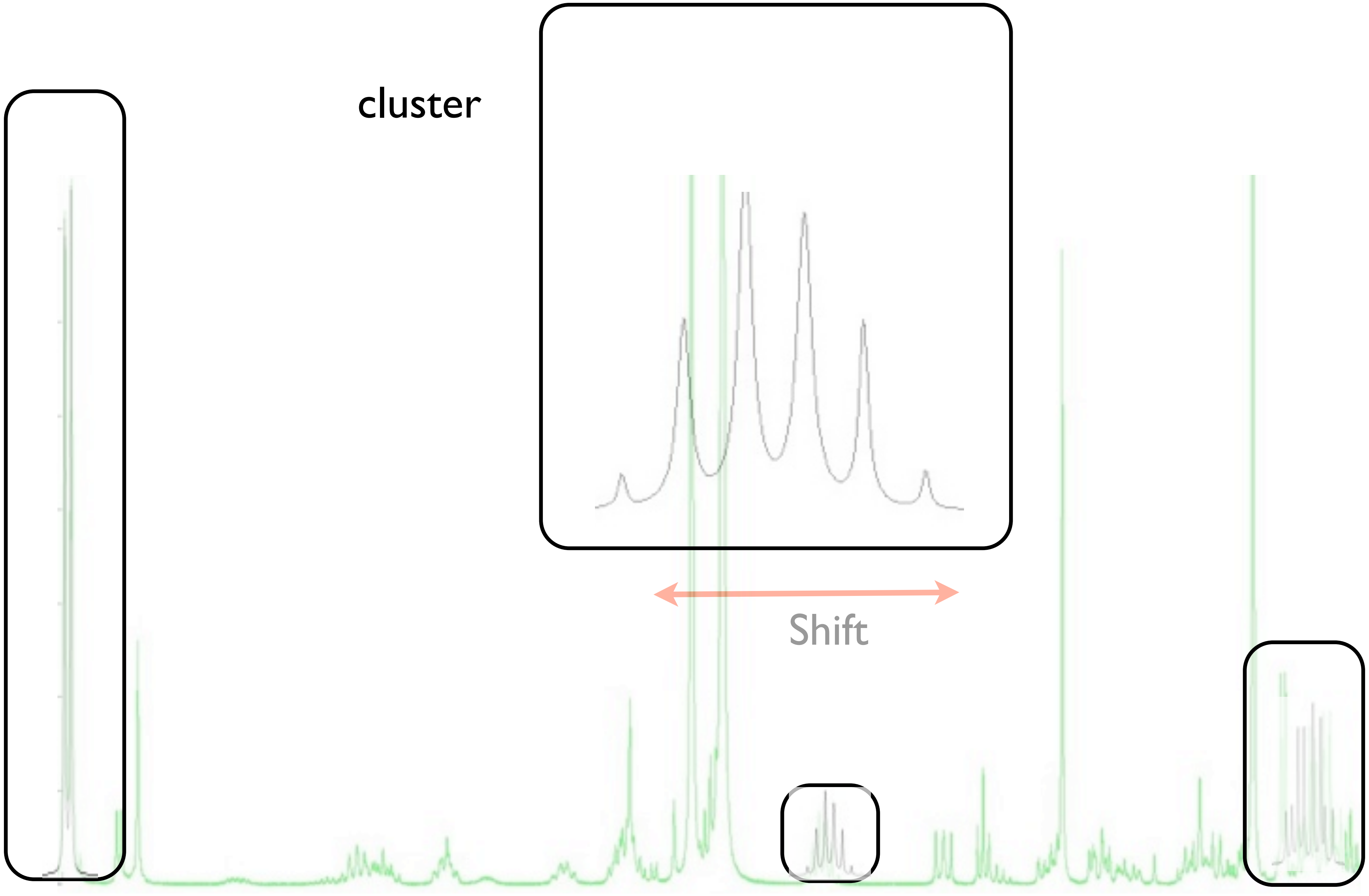

NMR Spectrum



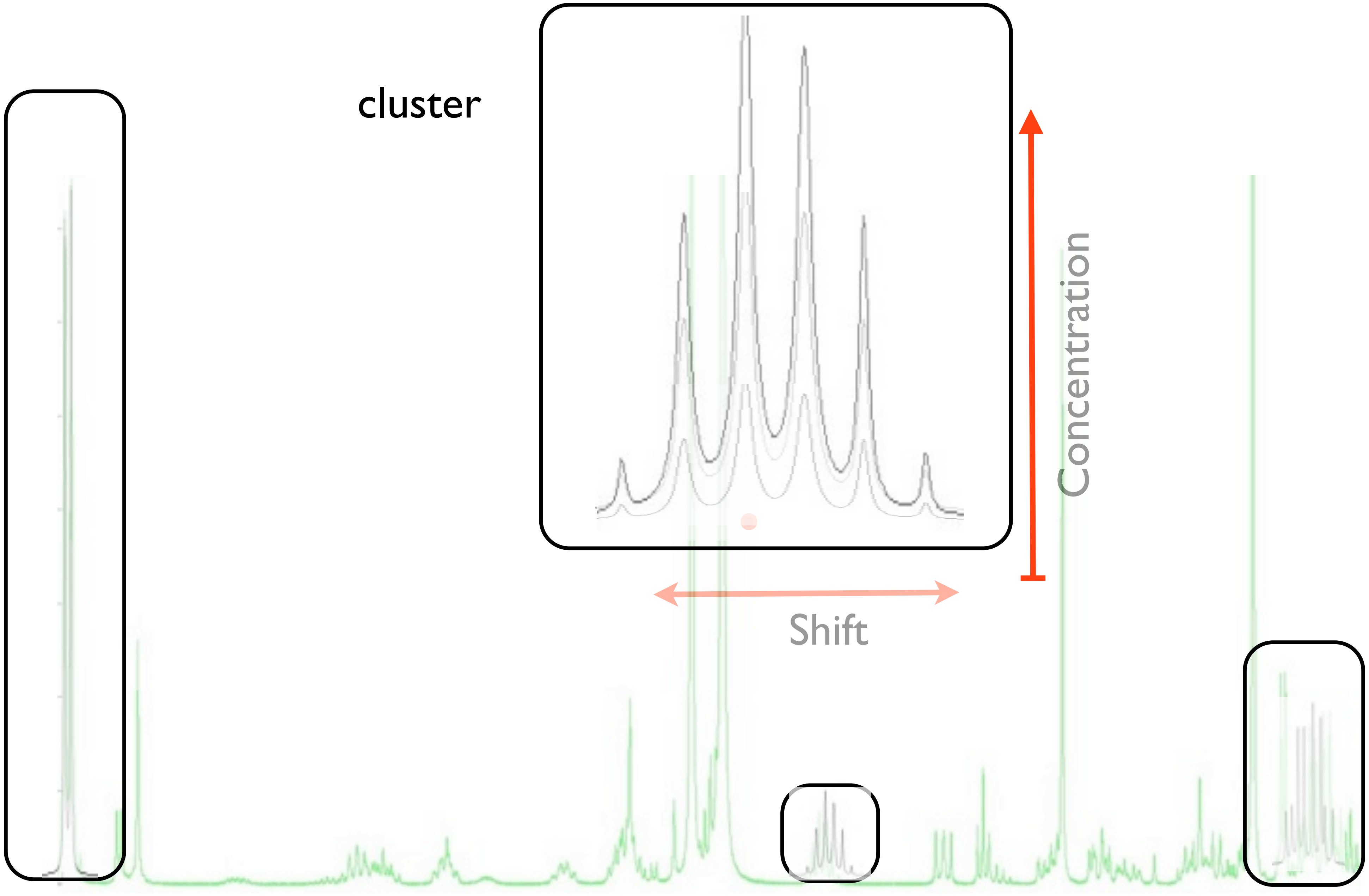
3-Hydroxyisobutyric acid



3-Hydroxyisobutyric acid



3-Hydroxyisobutyric acid



3-Hydroxyisobutyric acid

The Goal is:

Given The library of metabolite Signatures
Find Corresponding Concentration

The Goal is:

Given The library of metabolite Signatures
Find Corresponding Concentration

Optimization variables:

- Metabolite Concentrations
- Chemical Shifts

The Goal is:

Given The library of metabolite Signatures
Find Corresponding Concentration

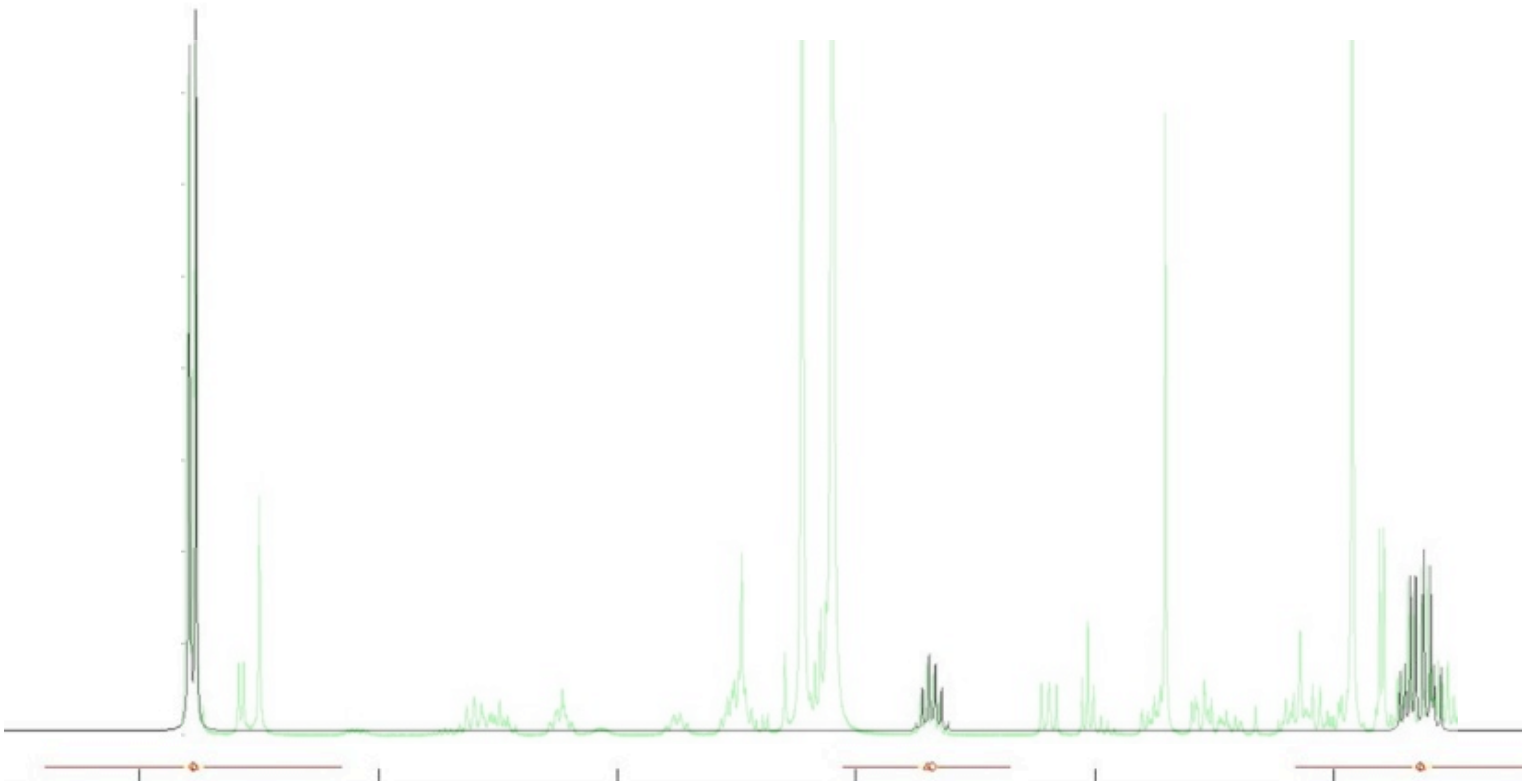
Optimization variables:

- Metabolite Concentrations
- Chemical Shifts

It's a difficult Problem

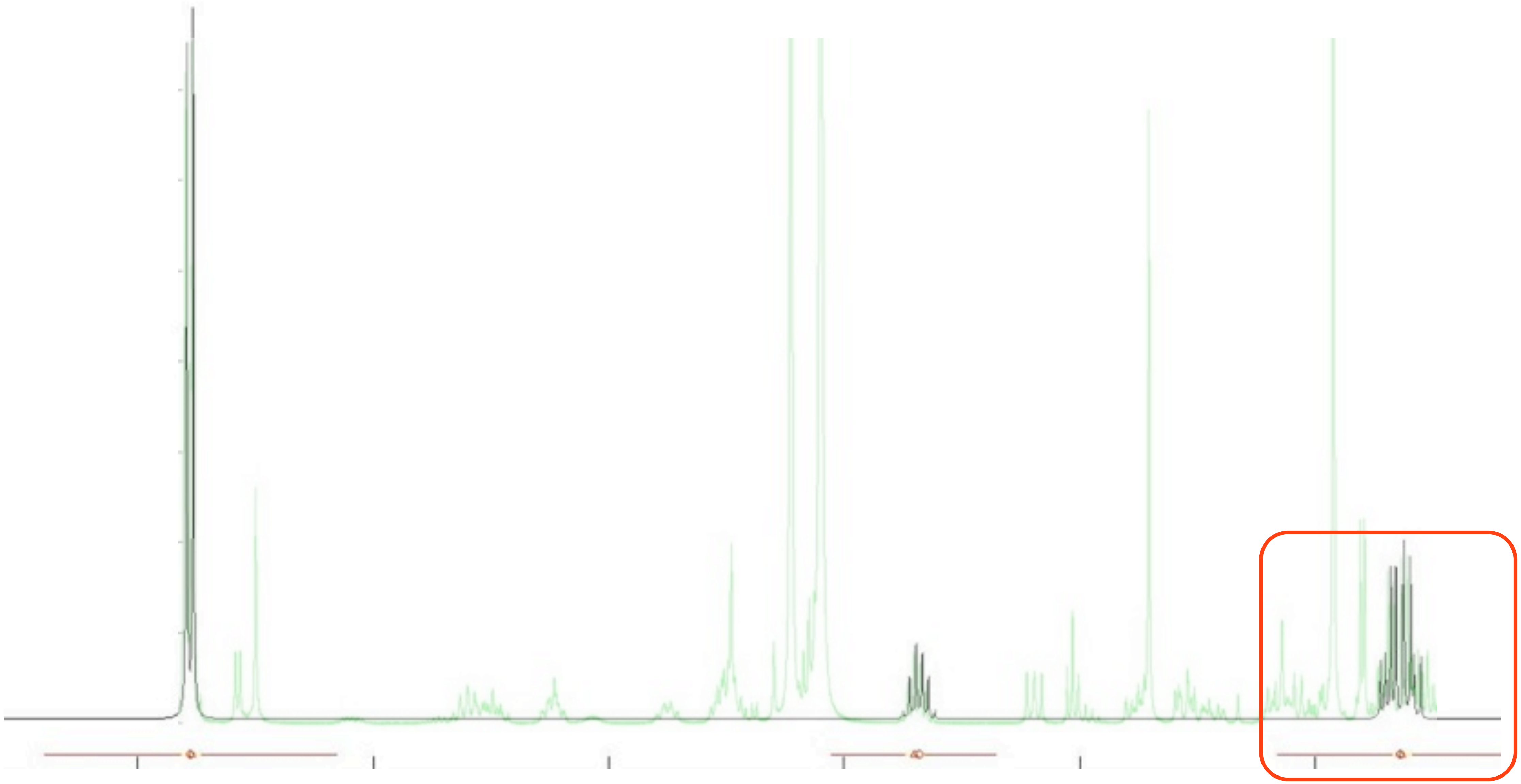
- Nonlinear in shift variables
- Involves hundreds of variables
- Loss is very expensive to evaluate (100K points)
- Non-Convex even around local optima
- Incomplete library results in over-fitting

Exploiting the structure



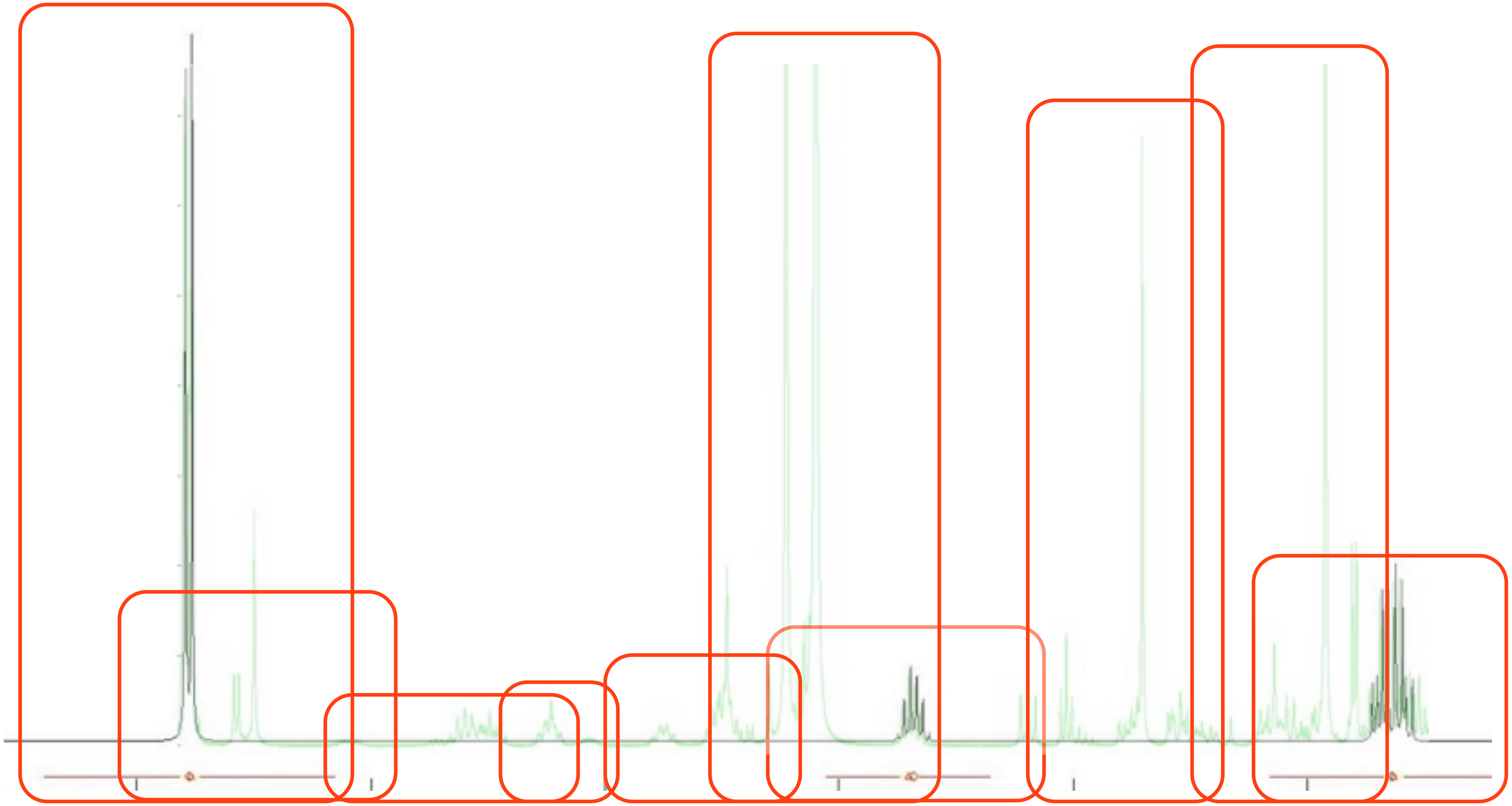
Exploiting the structure

Sub-problems are easier to solve but they share variables

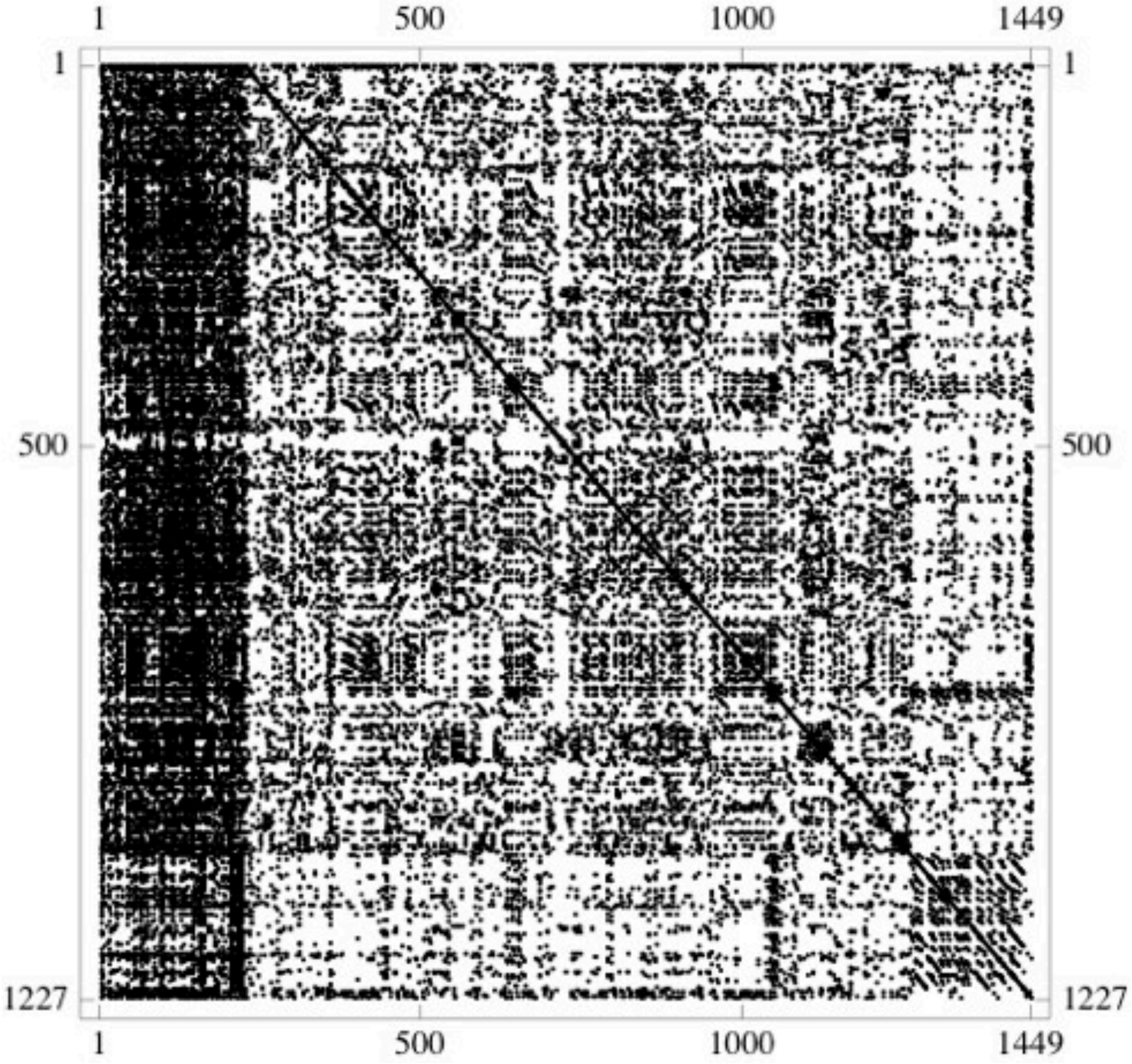


Exploiting the structure

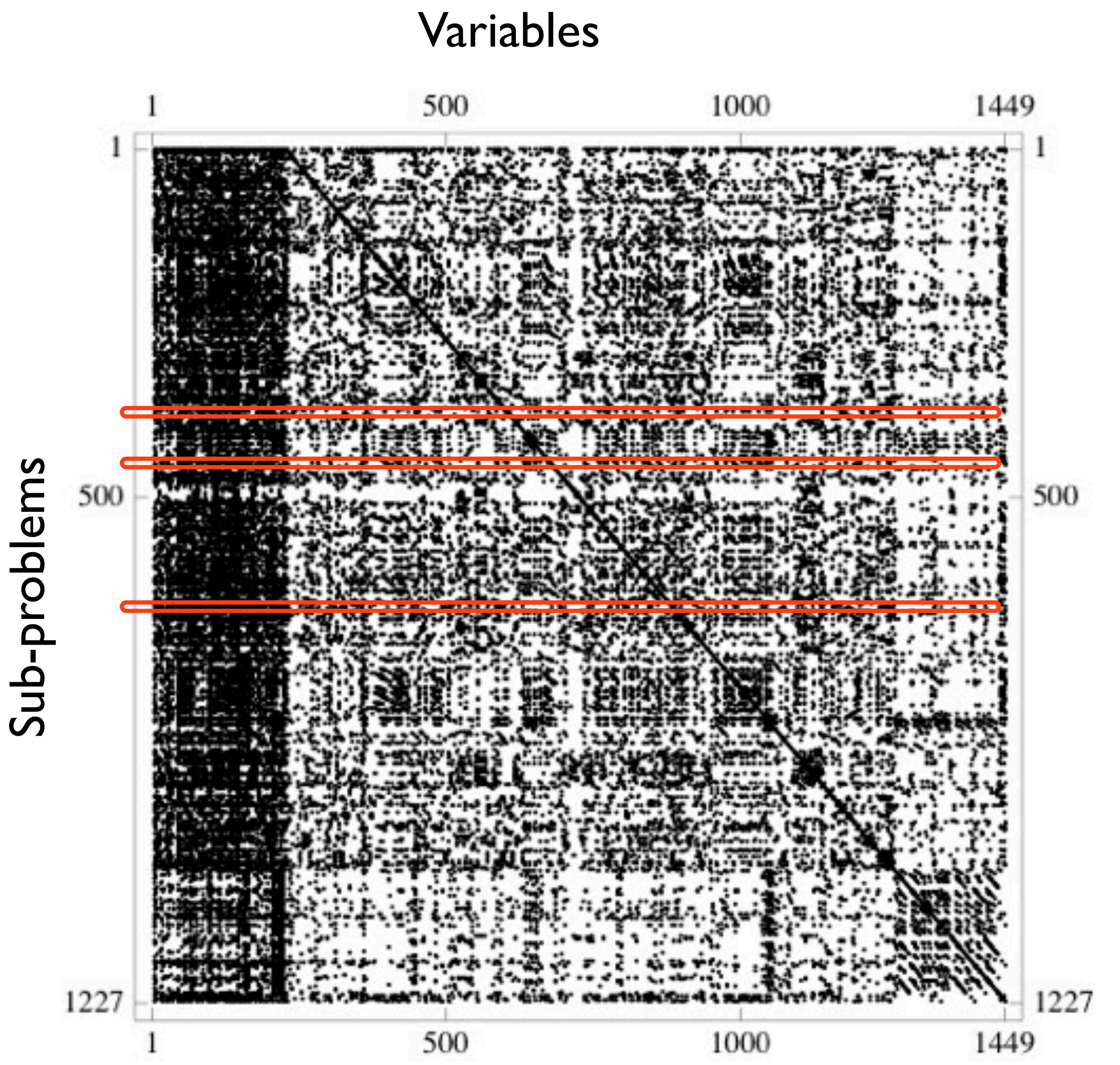
Sub-problems are easier to solve but they share variables



Coupling Matrix represents the structure



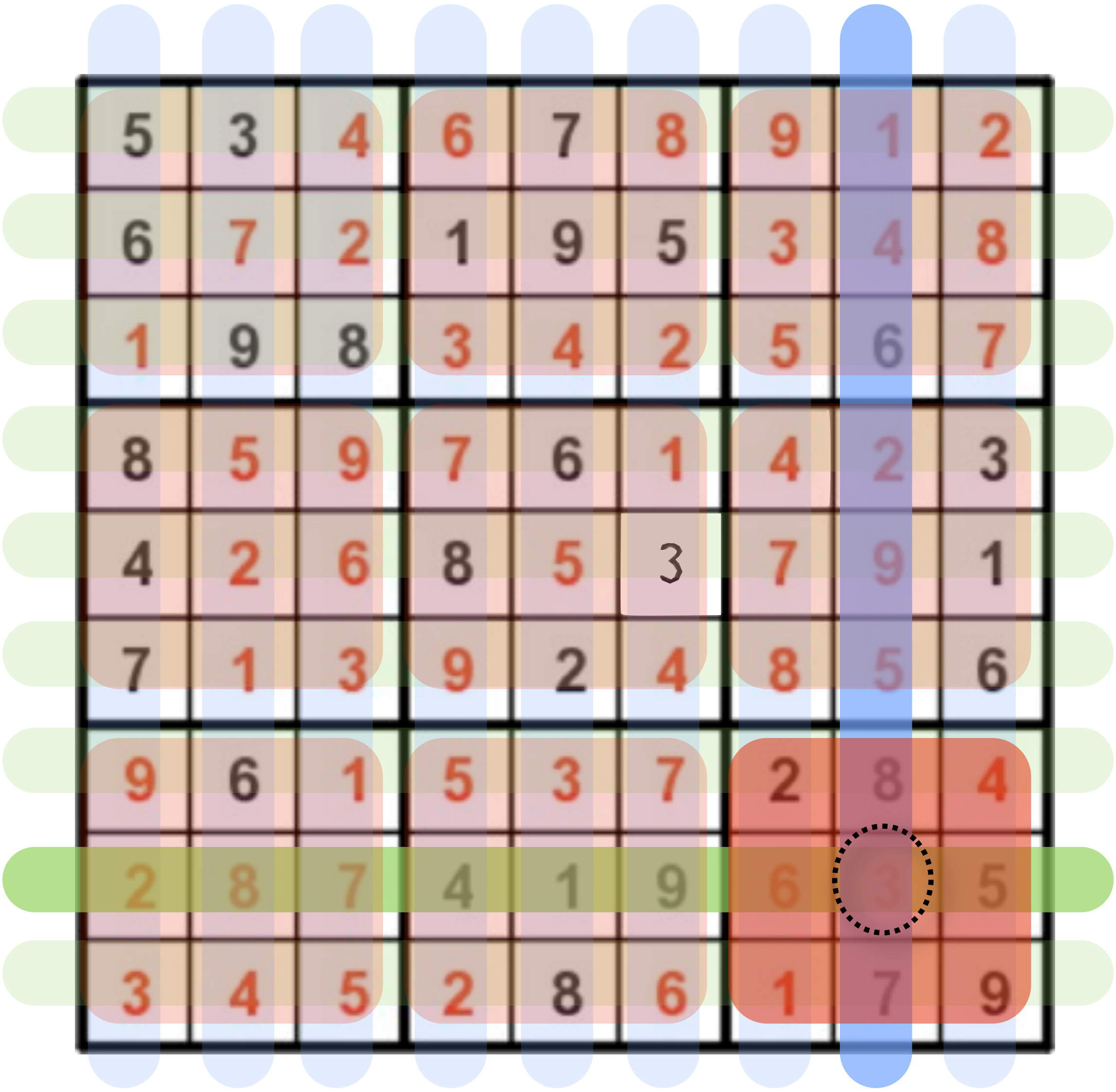
Coupling Matrix represents the structure



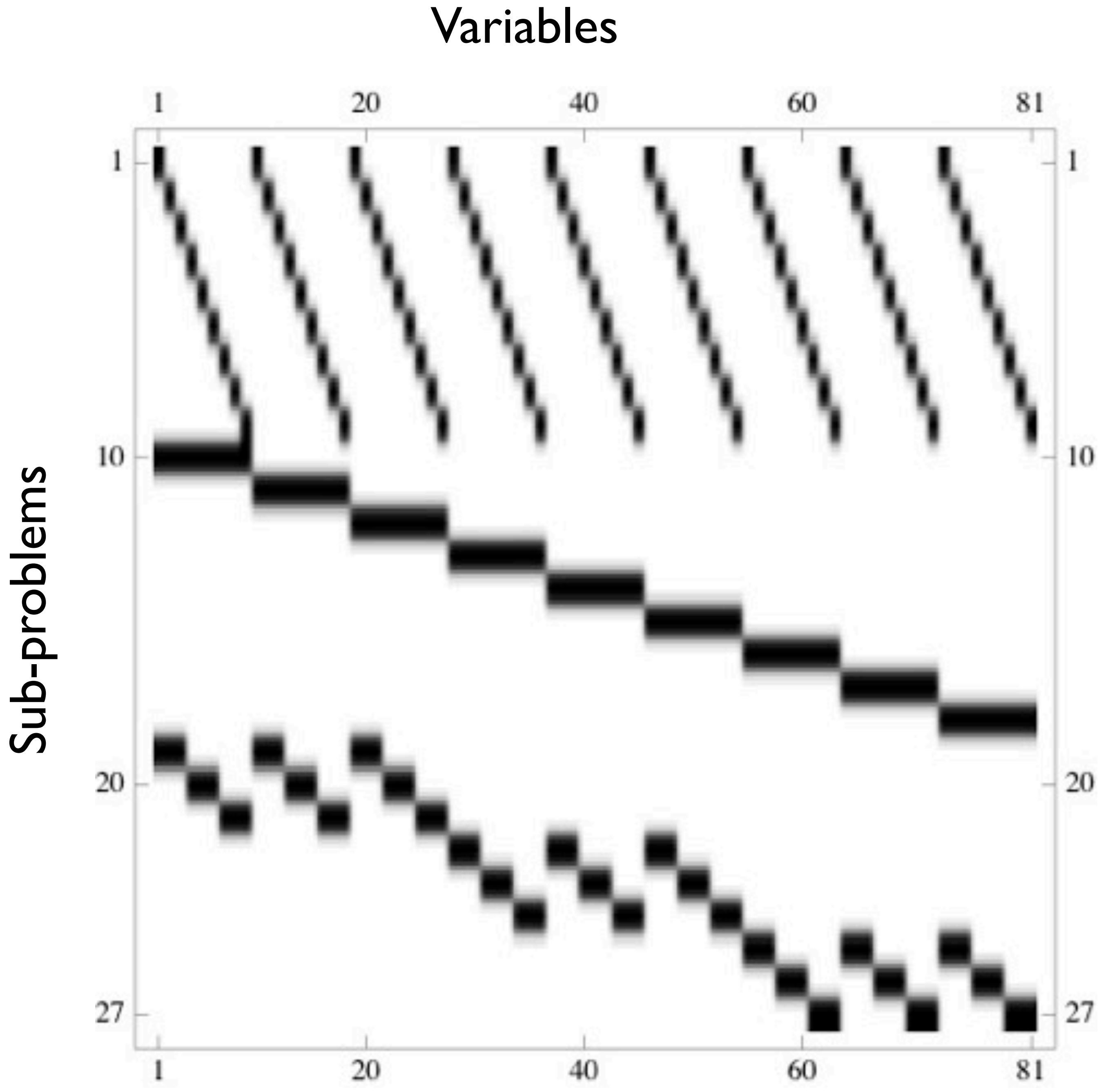
Coupling Matrix

represents the structure

Sudoku Puzzle



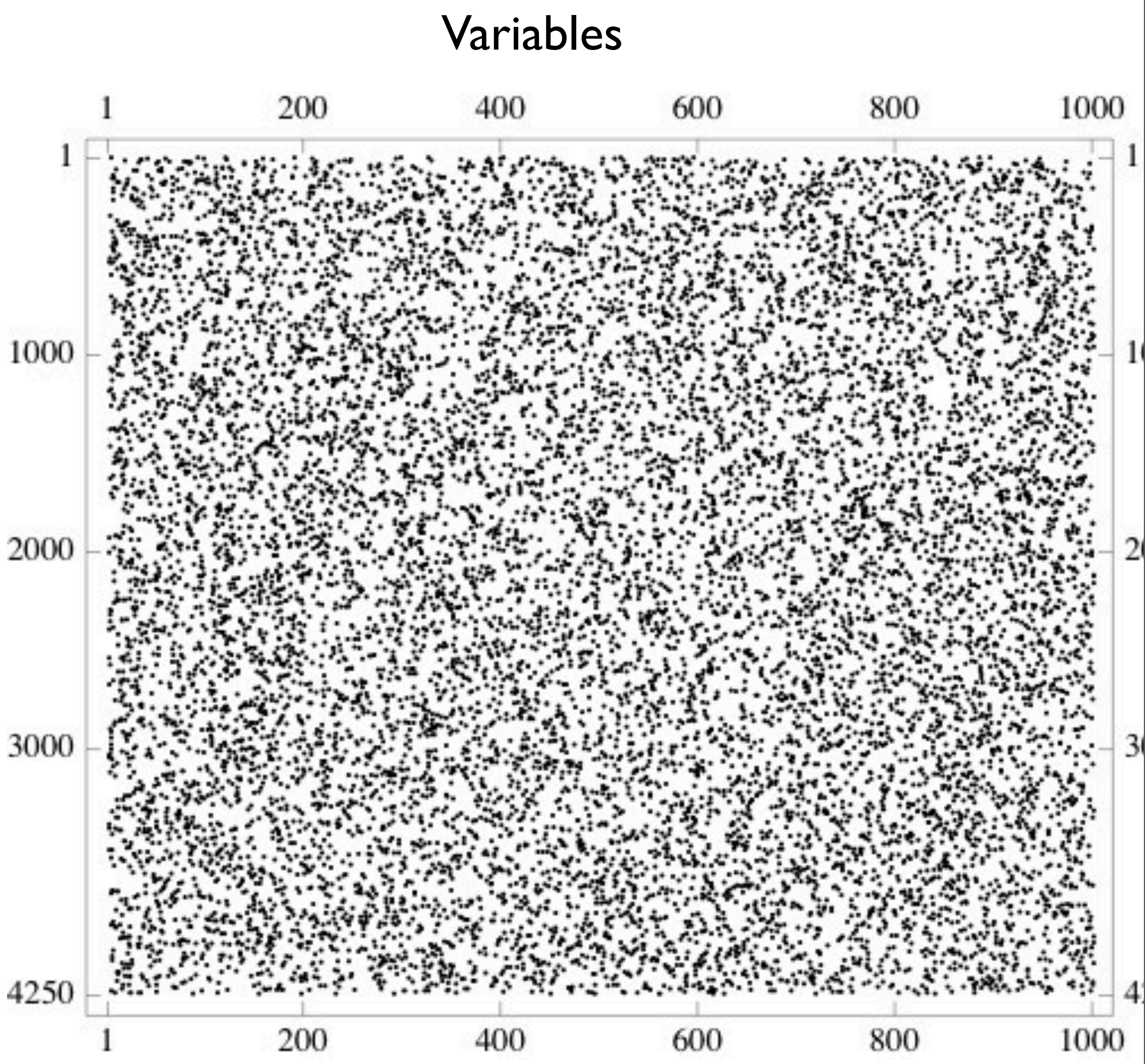
5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9



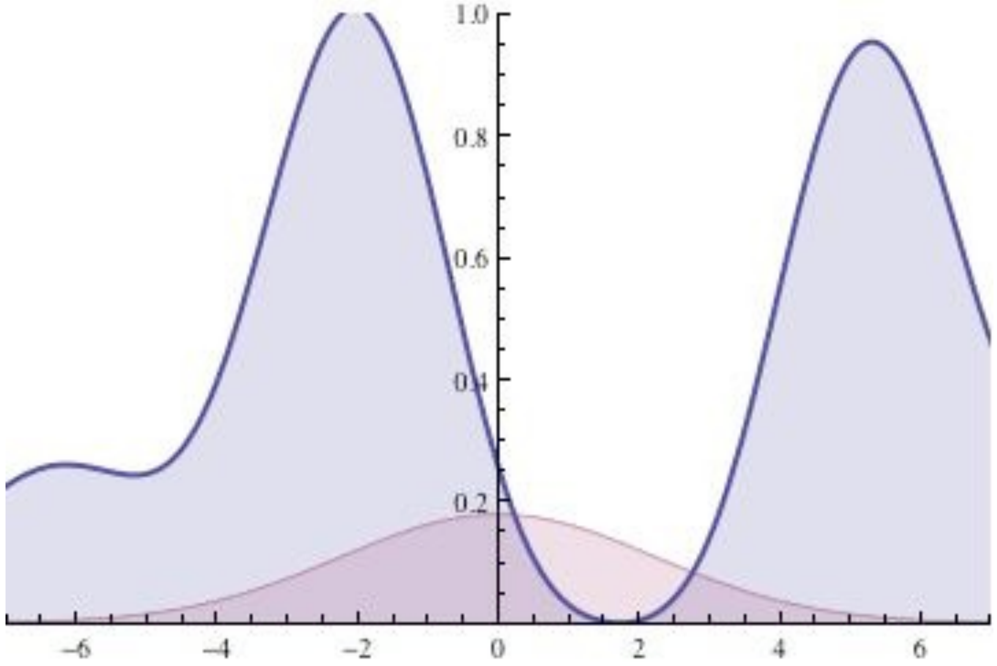
Coupling Matrix represents the structure

SATisfiability Problem

$$(A \vee B) \wedge (\neg B \vee C \vee \neg D) \wedge (D \vee \neg E)$$

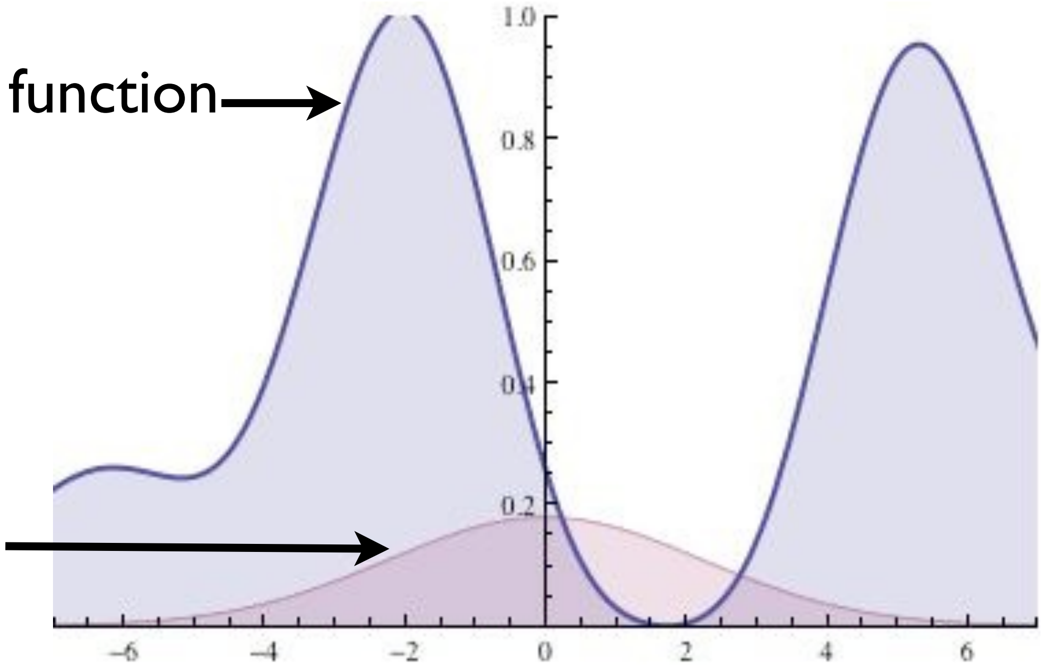


Cross Entropy (CE) Method for Optimization



Cross Entropy (CE) Method for Optimization

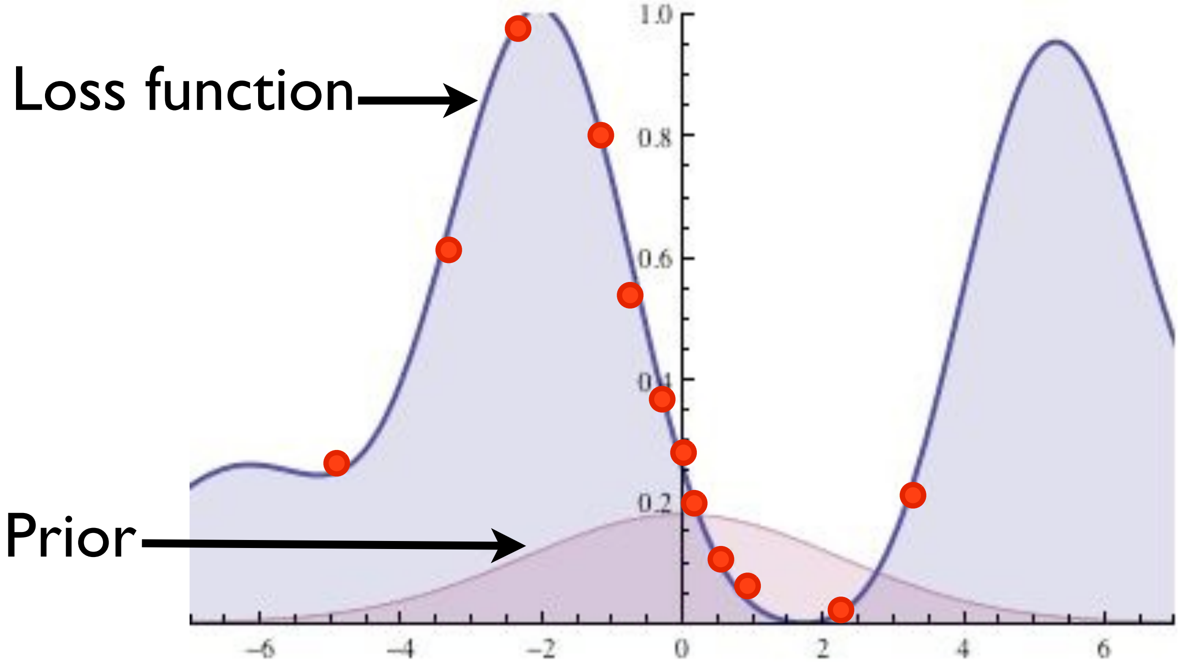
Start from a prior



Loss function

Prior

Cross Entropy (CE) Method for Optimization


Loss function

Prior

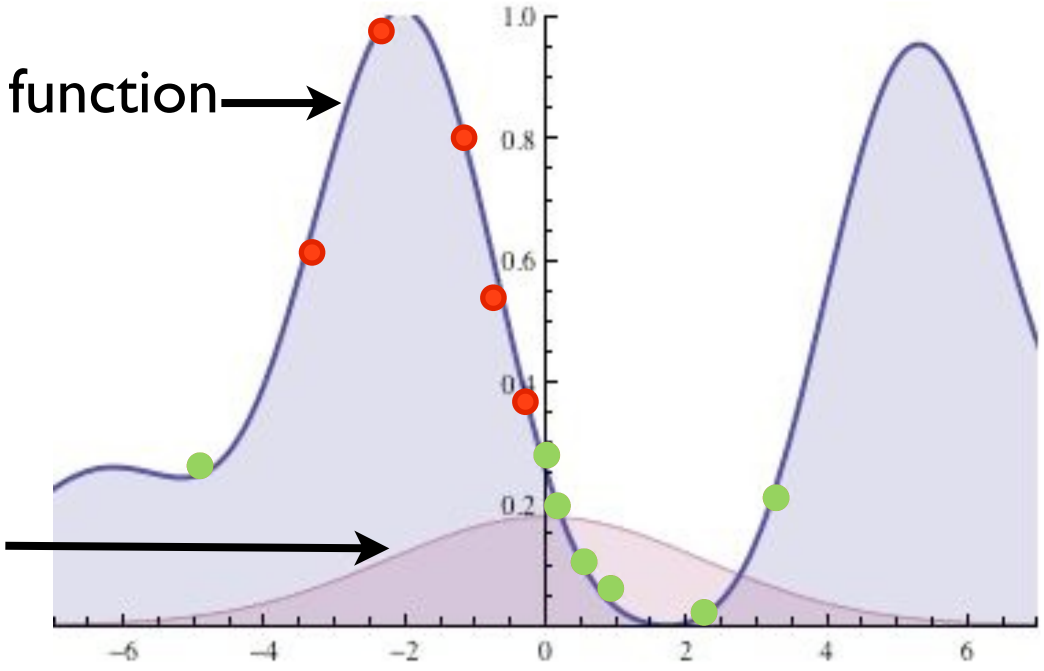
Start from a prior

Repeat until convergence

Take samples from current dist.

Calculate the loss for samples

Cross Entropy (CE) Method for Optimization



Loss function

Prior

Start from a prior

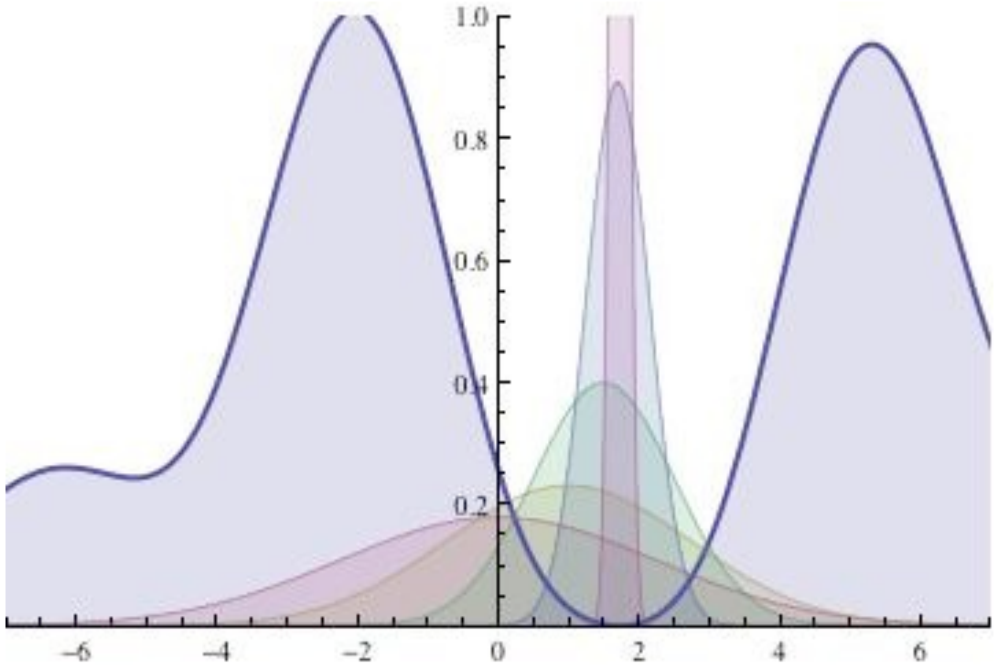
Repeat until convergence

Take samples from current dist.

Calculate the loss for samples

Select Elite samples

Cross Entropy (CE) Method for Optimization

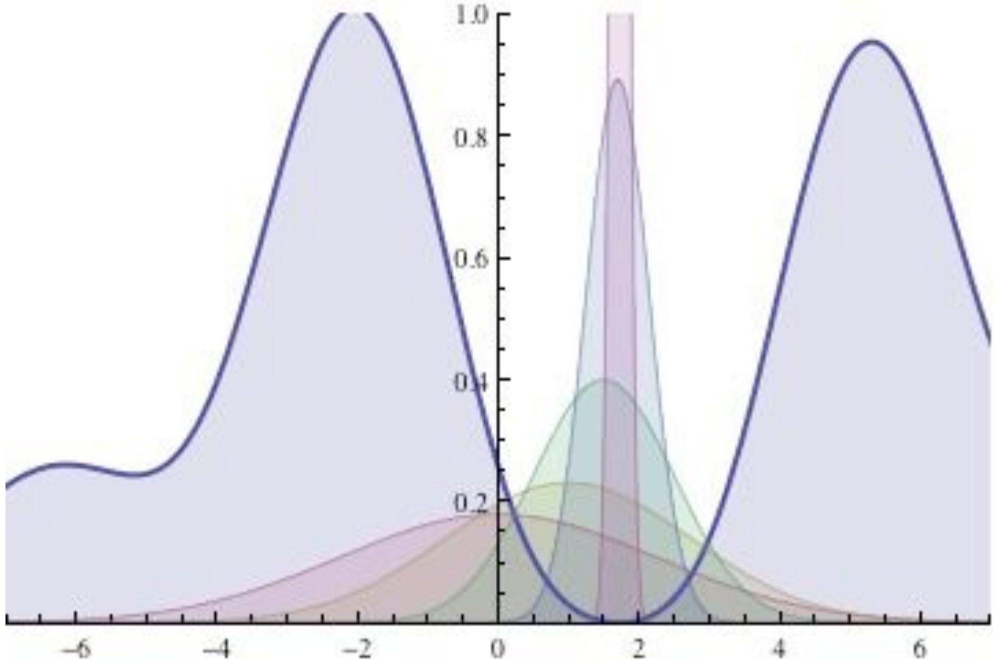


Start from a prior

Repeat until convergence

- Take samples from current dist.
- Calculate the loss for samples
- Select Elite samples
- Find maximum likelihood dist. for Elites

Cross Entropy (CE) Method for Optimization



Start from a prior

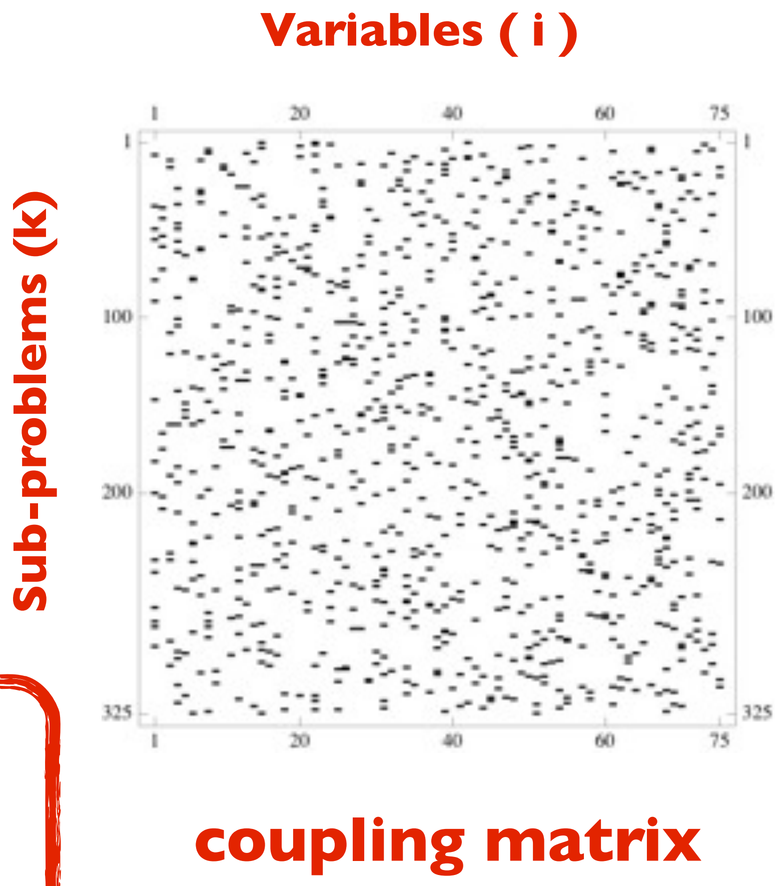
Repeat until convergence

- Take samples from current dist.
- Calculate the loss for samples
- Select Elite samples
- Find maximum likelihood dist. for Elites

A subroutine to be used again

Variation of CE

that Exploits Decomposability (**CEED**)



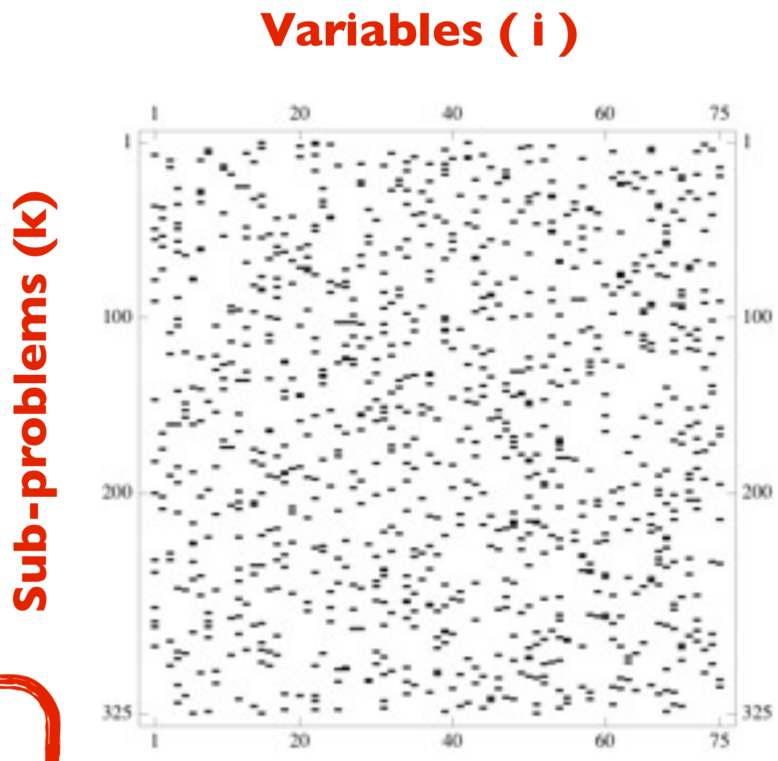
Variation of CE

that Exploits Decomposability (CEED)

Start from a prior

Repeat,

Until convergence

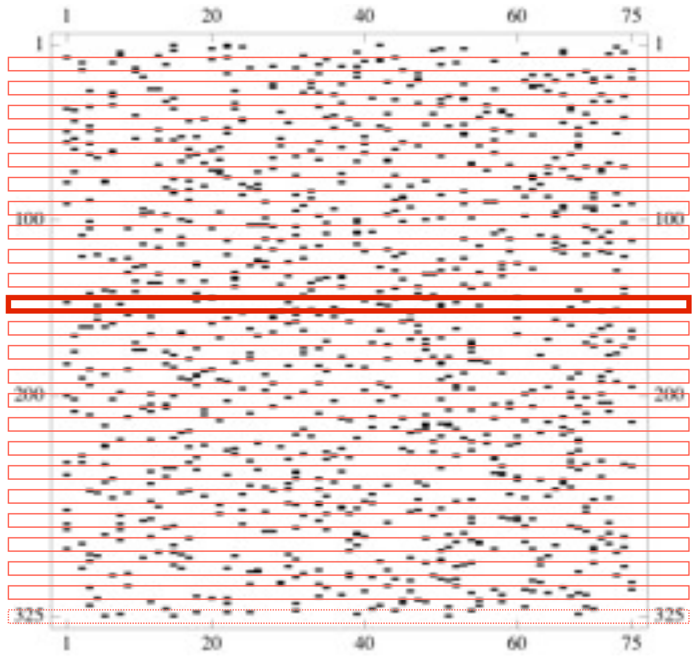


coupling matrix

Variation of CE

that Exploits Decomposability (CEED)

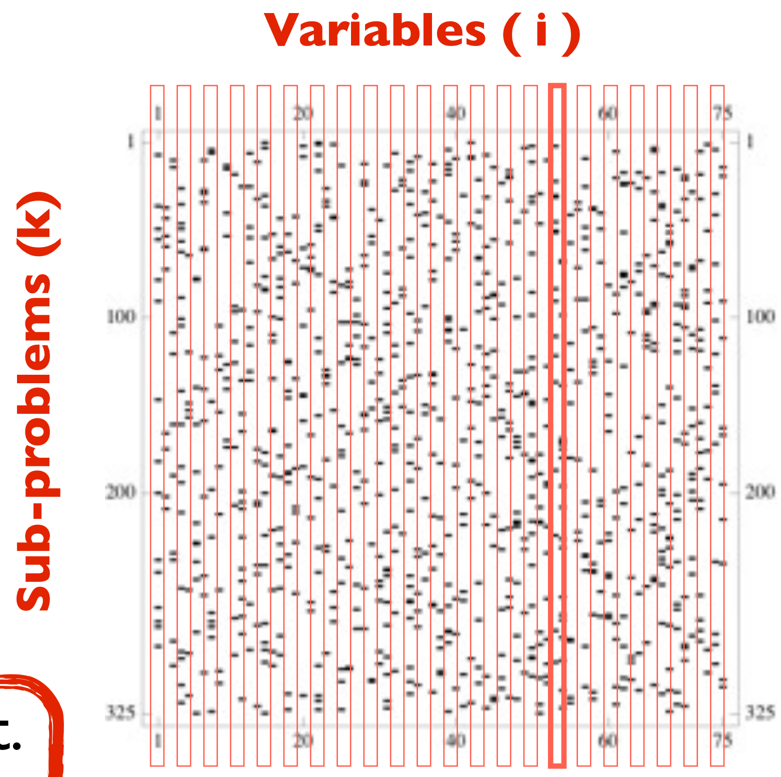
```
Start from a prior
Repeat,
    For each sub-problem k
        Draw samples from marginal of current dist.
        Calculate the loss for samples
        Select Elite samples
        Find maximum likelihood dist. for Elites
Until convergence
```



coupling matrix

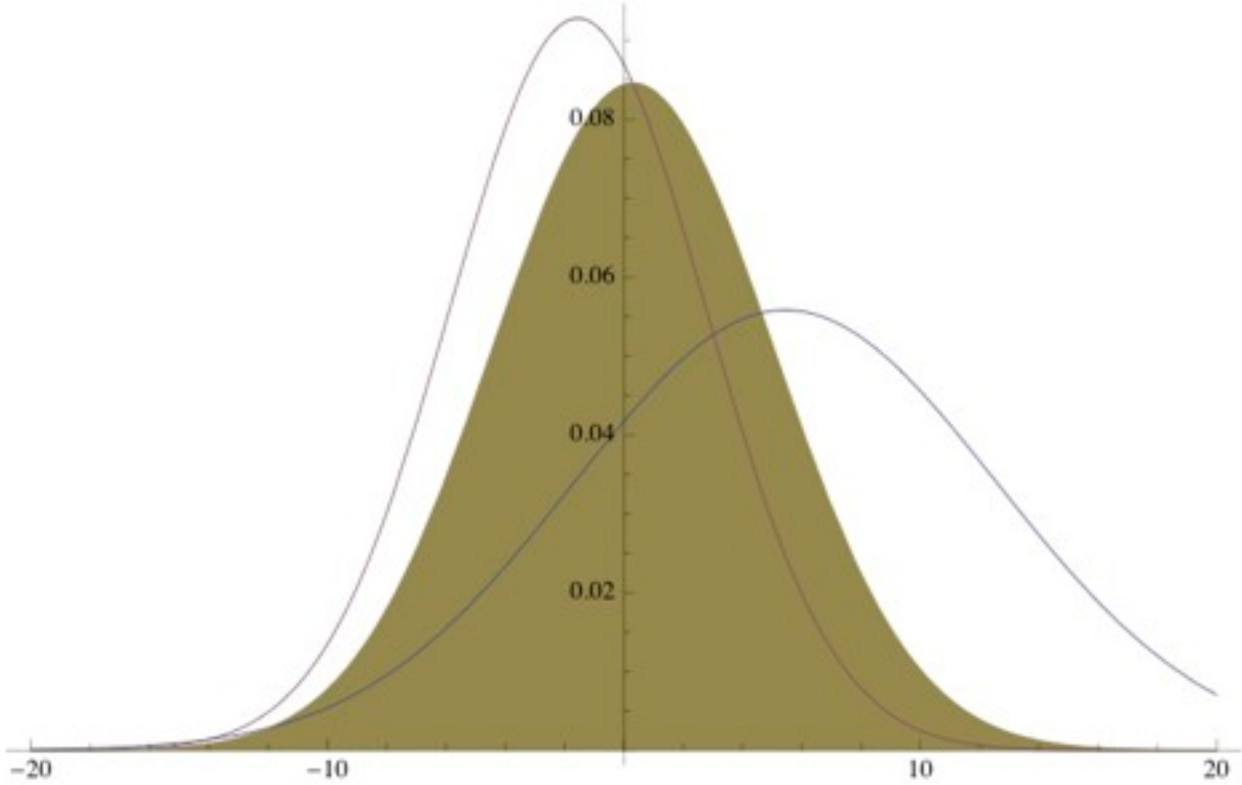
Variation of CE
that Exploits Decomposability (CEED)

```
Start from a prior
Repeat,
    For each sub-problem k
        Draw samples from marginal of current dist.
        Calculate the loss for samples
        Select Elite samples
        Find maximum likelihood dist. for Elites
    For each variable i
        Combine   dist's from related sub-problems
Until convergence
```



coupling matrix

How to Combine ML Distributions



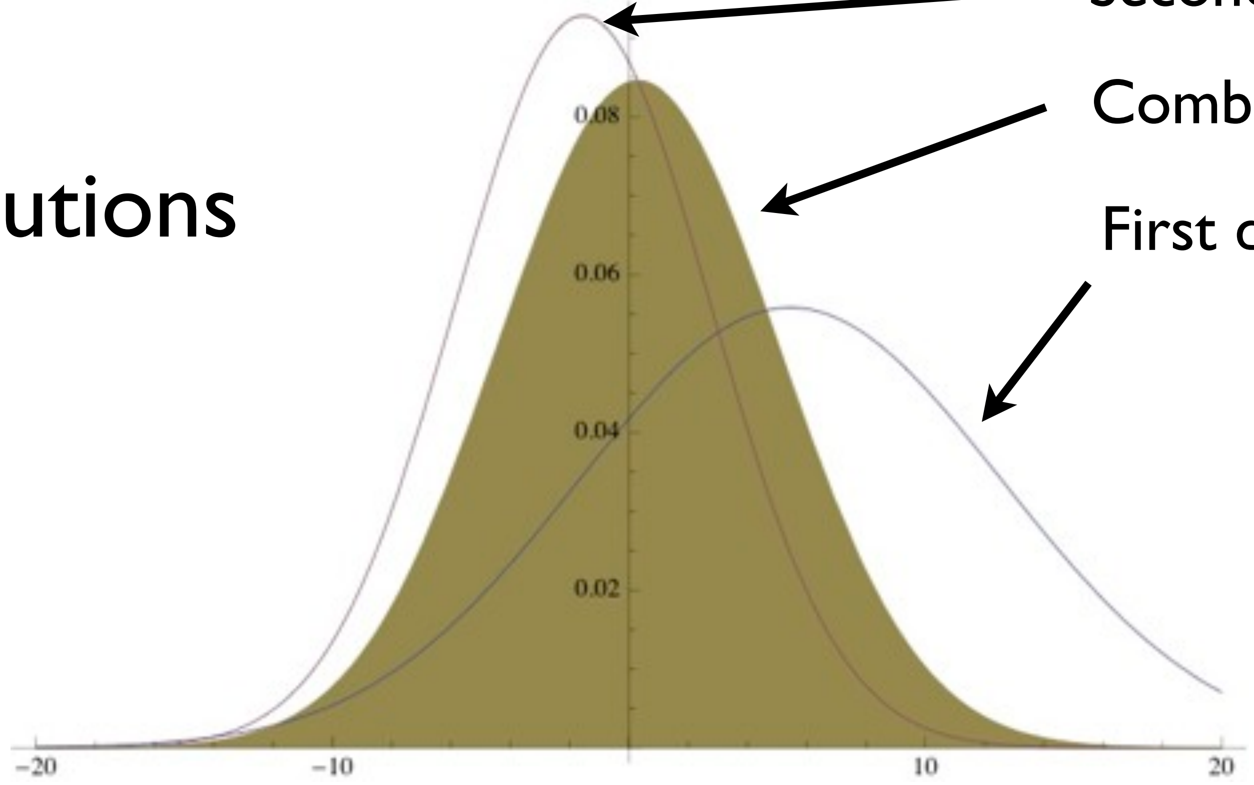
How to Combine ML Distributions

We use linear combination by Fisher Information

How to Combine ML Distributions

We use linear combination by Fisher Information

Gaussian Distributions



Second distribution

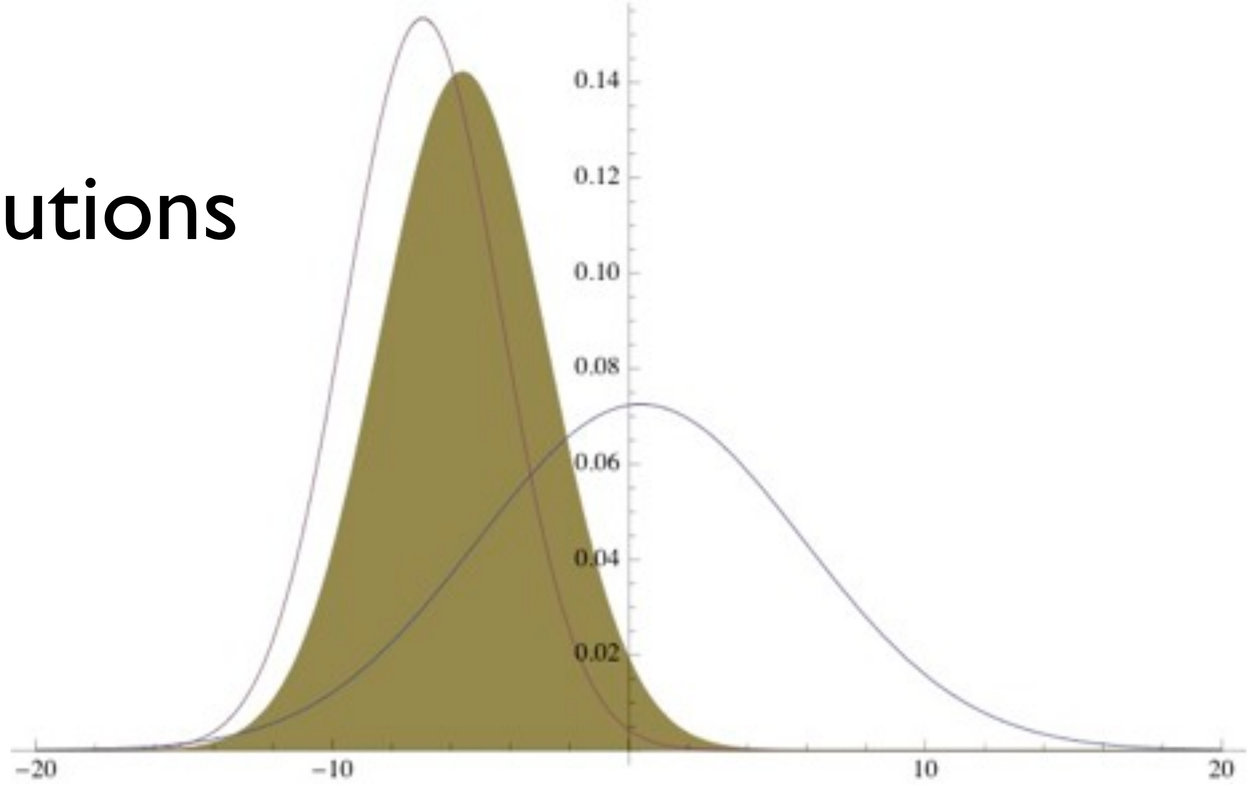
Combined distribution

First distribution

How to Combine ML Distributions

We use linear combination by Fisher Information

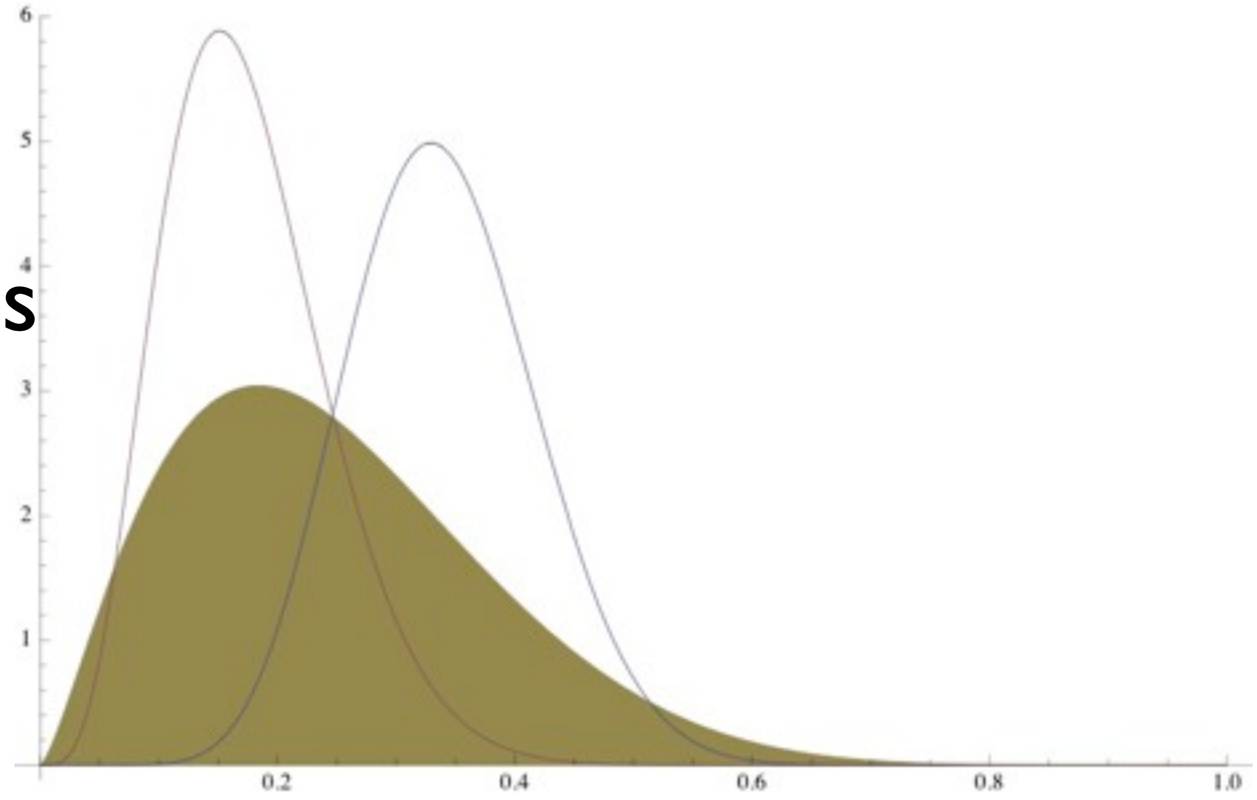
Gaussian Distributions



How to Combine ML Distributions

We use linear combination by Fisher Information

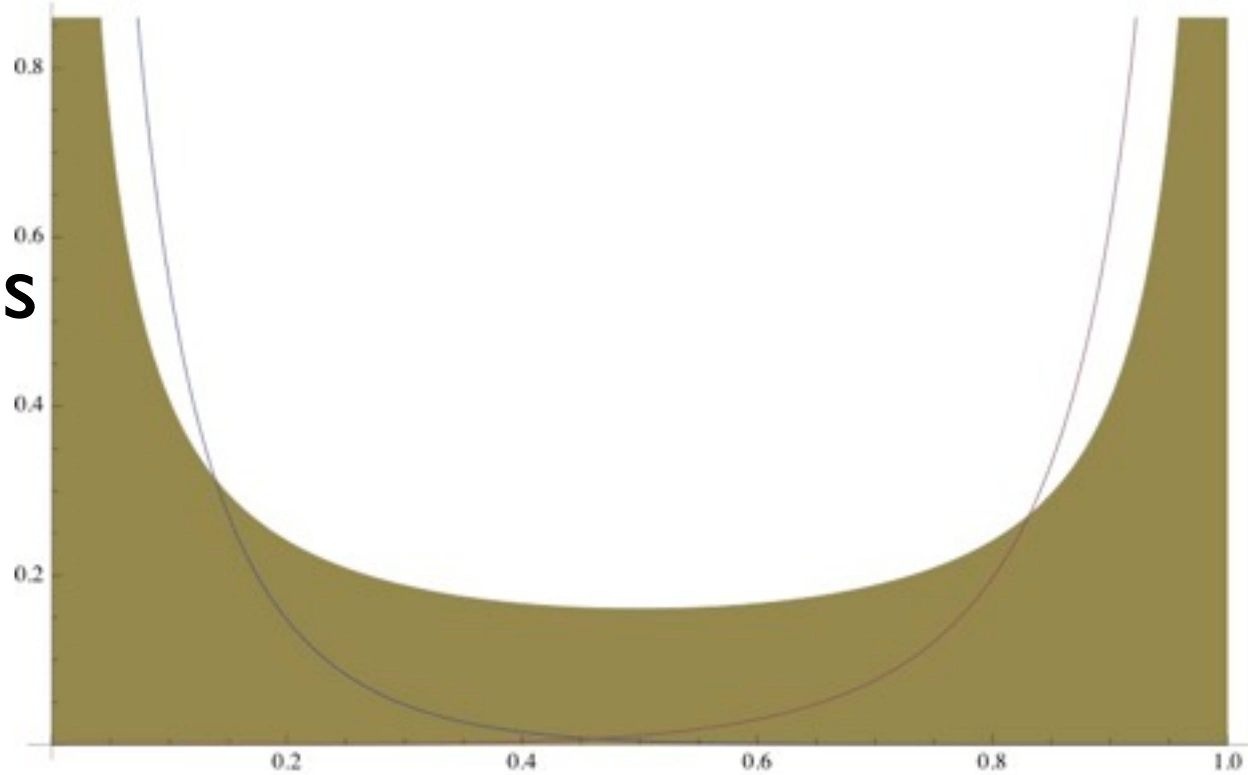
Beta Distributions



How to Combine ML Distributions

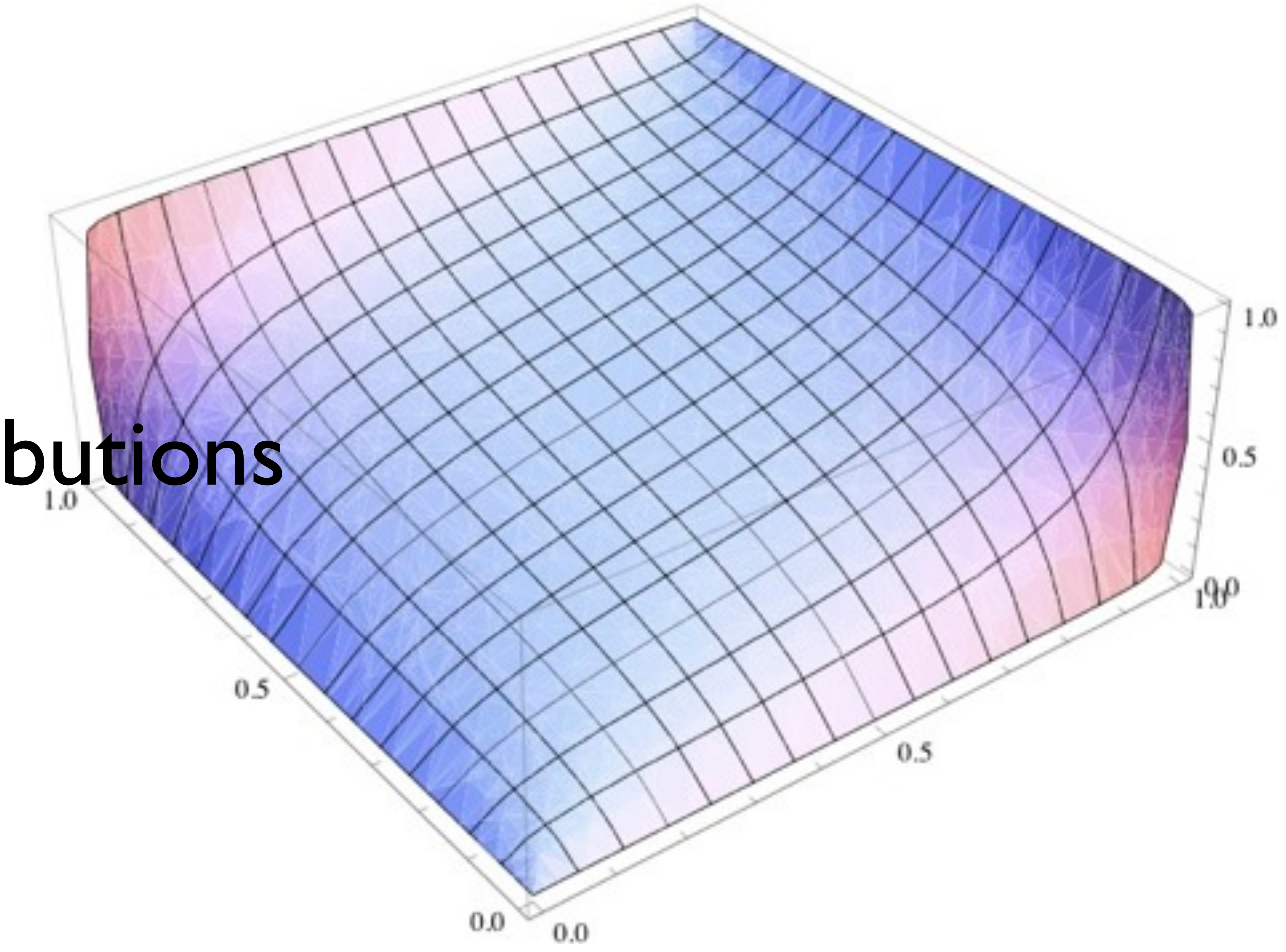
We use linear combination by Fisher Information

Beta Distributions



How to Combine ML Distributions

We use linear combination by Fisher Information

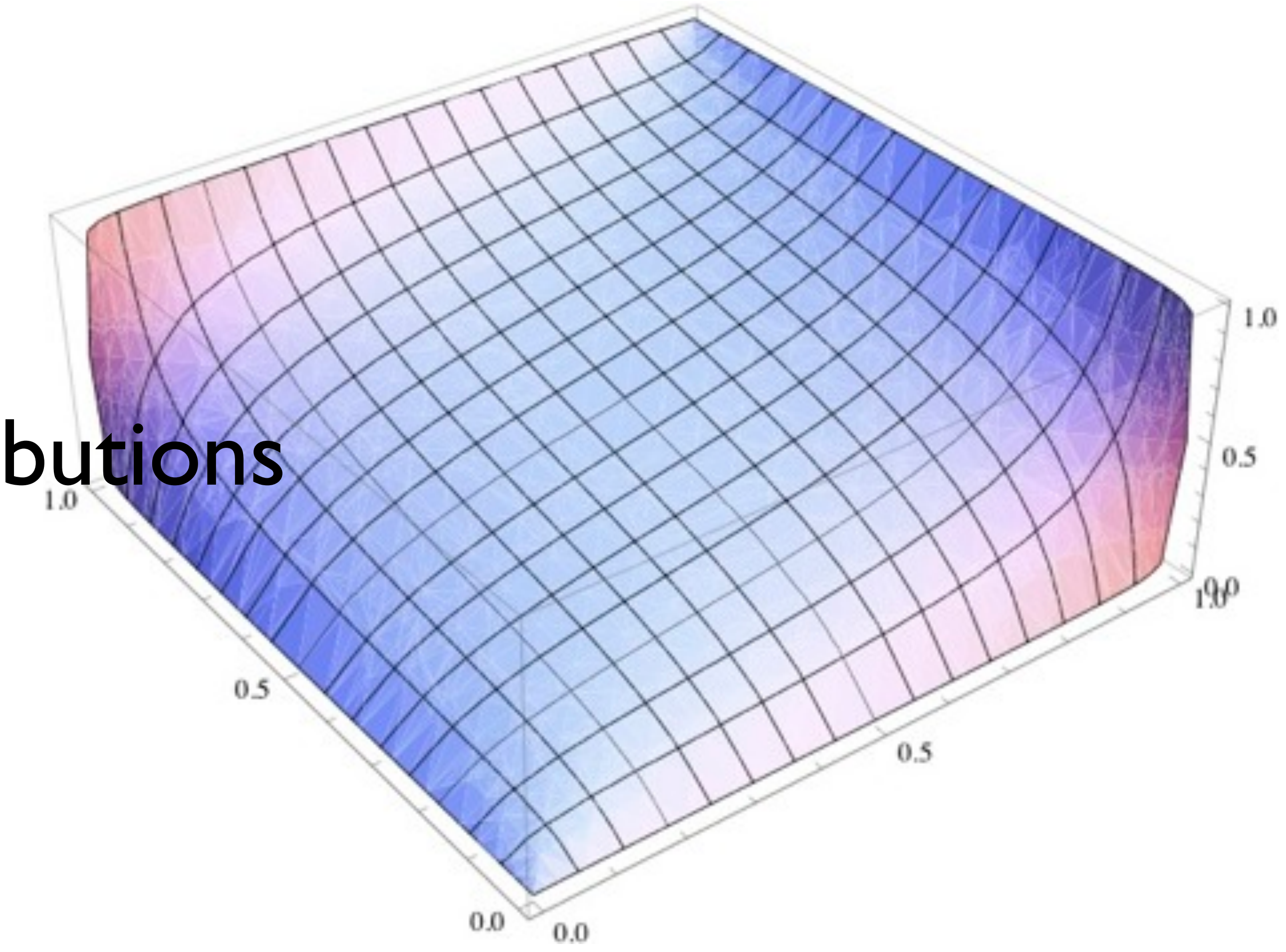


Bernoulli Distributions

How to Combine ML Distributions

We use linear combination by Fisher Information

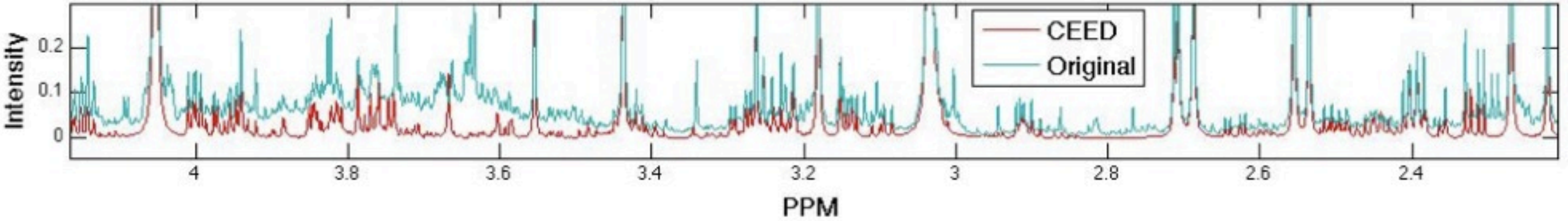
Bernoulli Distributions

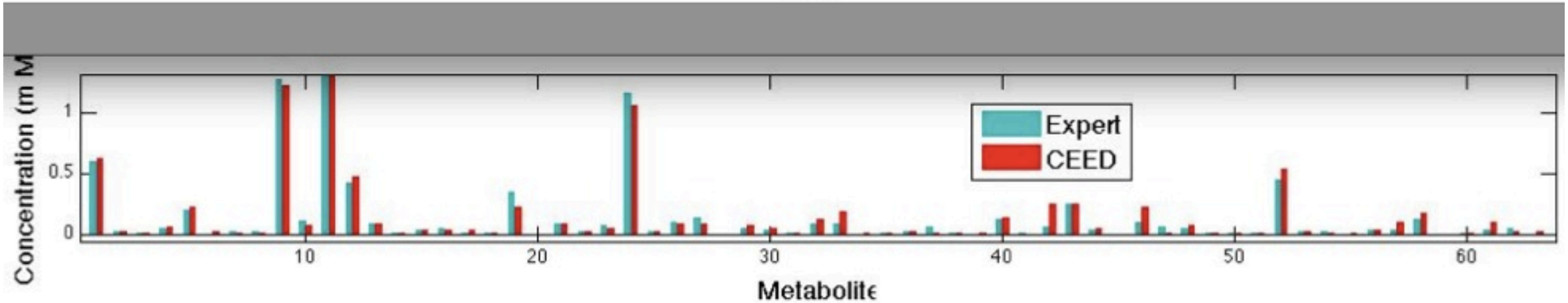


A Tweak for NMR problem

Experimental Results

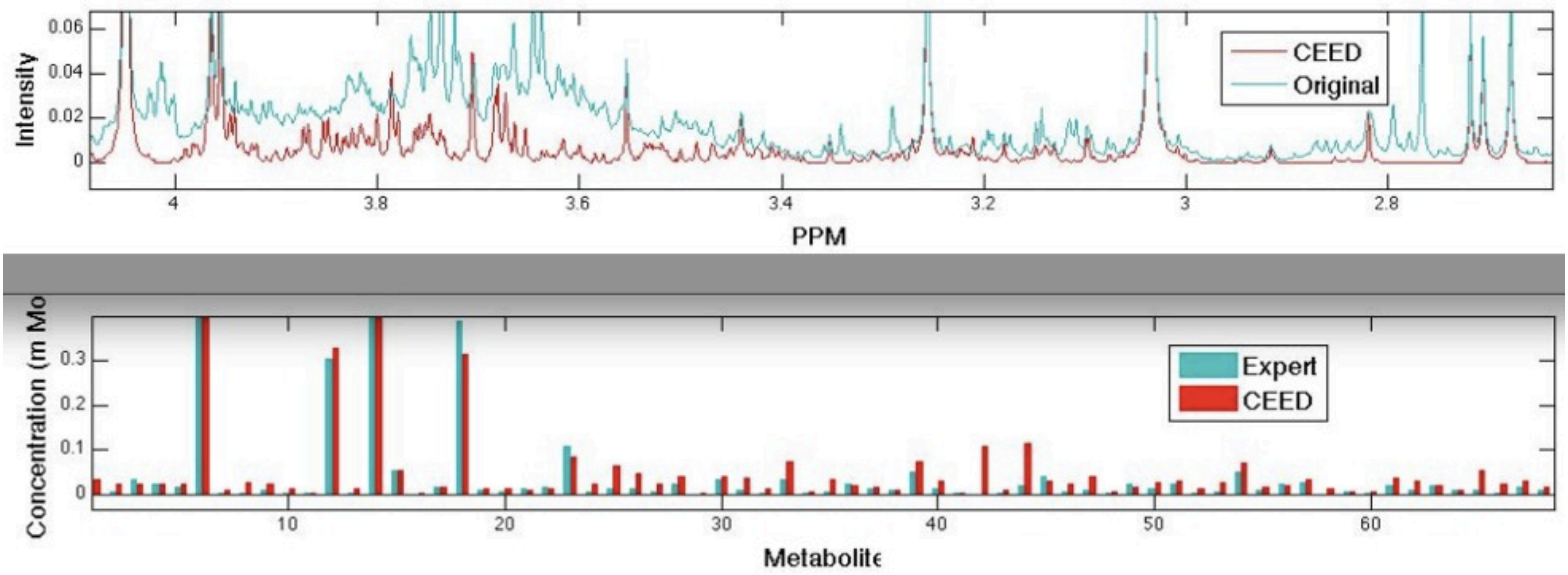
800 MHz Spectra





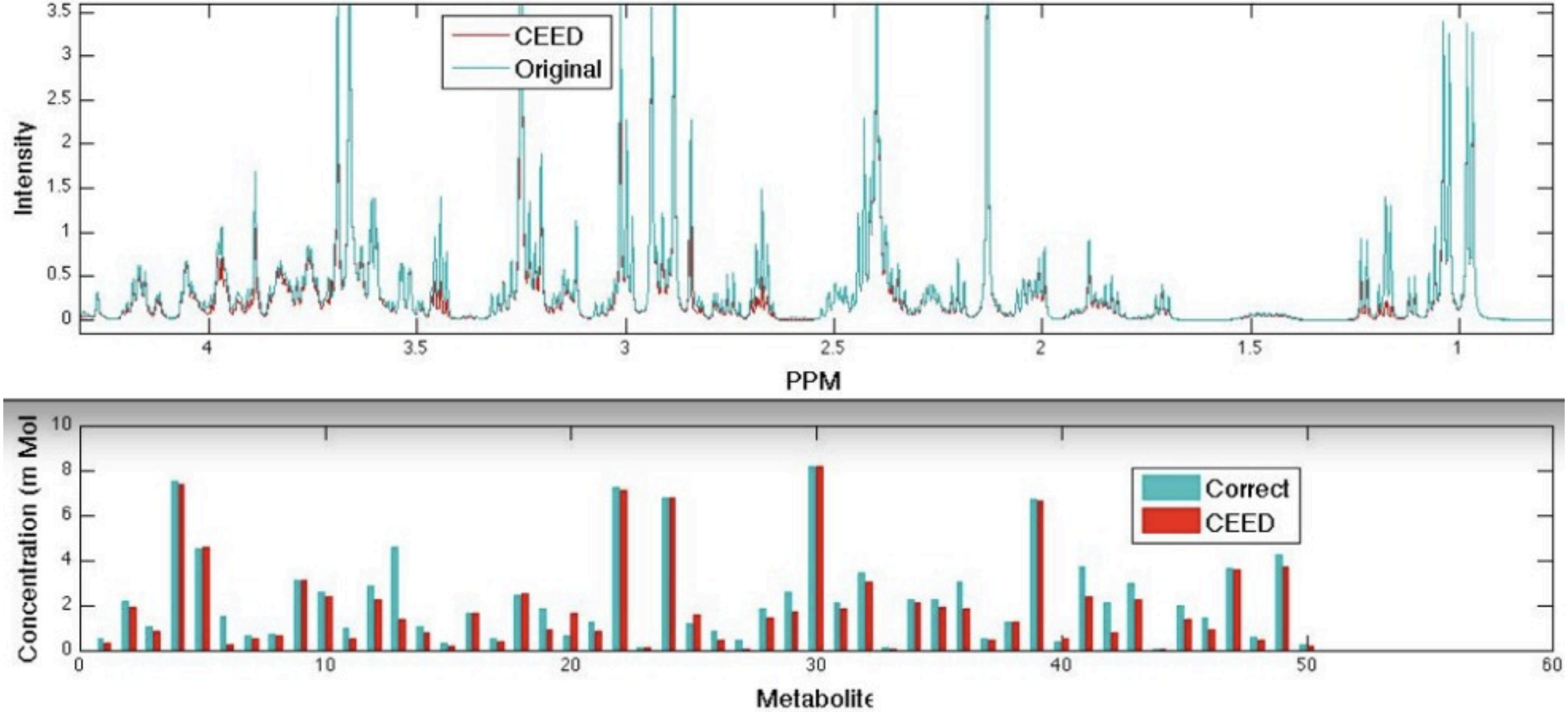
Experimental Results

500 MHz Spectra



Experimental Results

Simulated Spectra

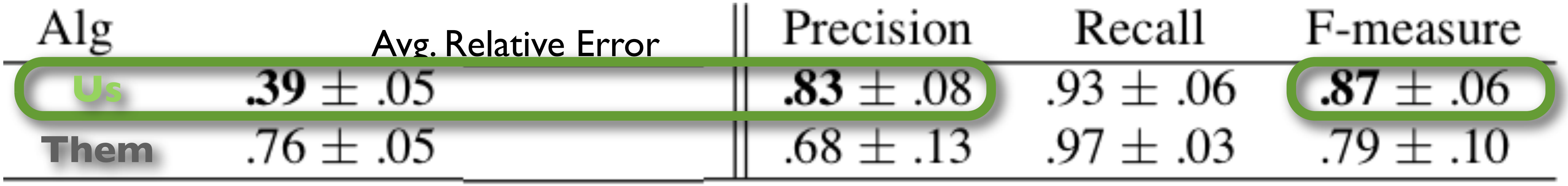


Experimental Results

Comparison with ChenomX Inc. automated fitting software

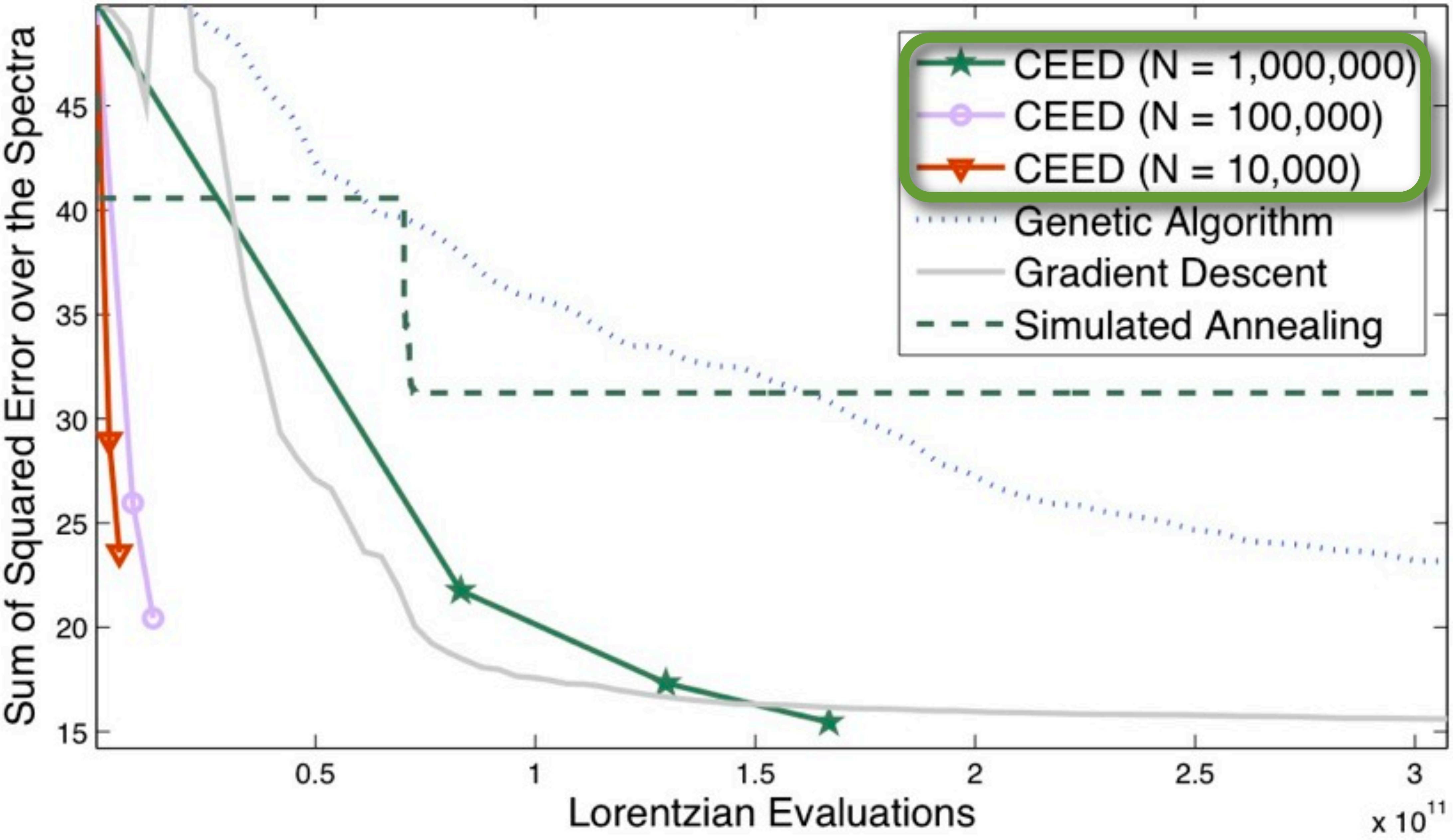
Quantification Task

Metabolite Detection Task
Threshold: .02 mMol



Alg	Avg. Relative Error	Precision	Recall	F-measure
Us	**.39** ± .05	**.83** ± .08	.93 ± .06	**.87** ± .06
Them	.76 ± .05	.68 ± .13	.97 ± .03	.79 ± .10

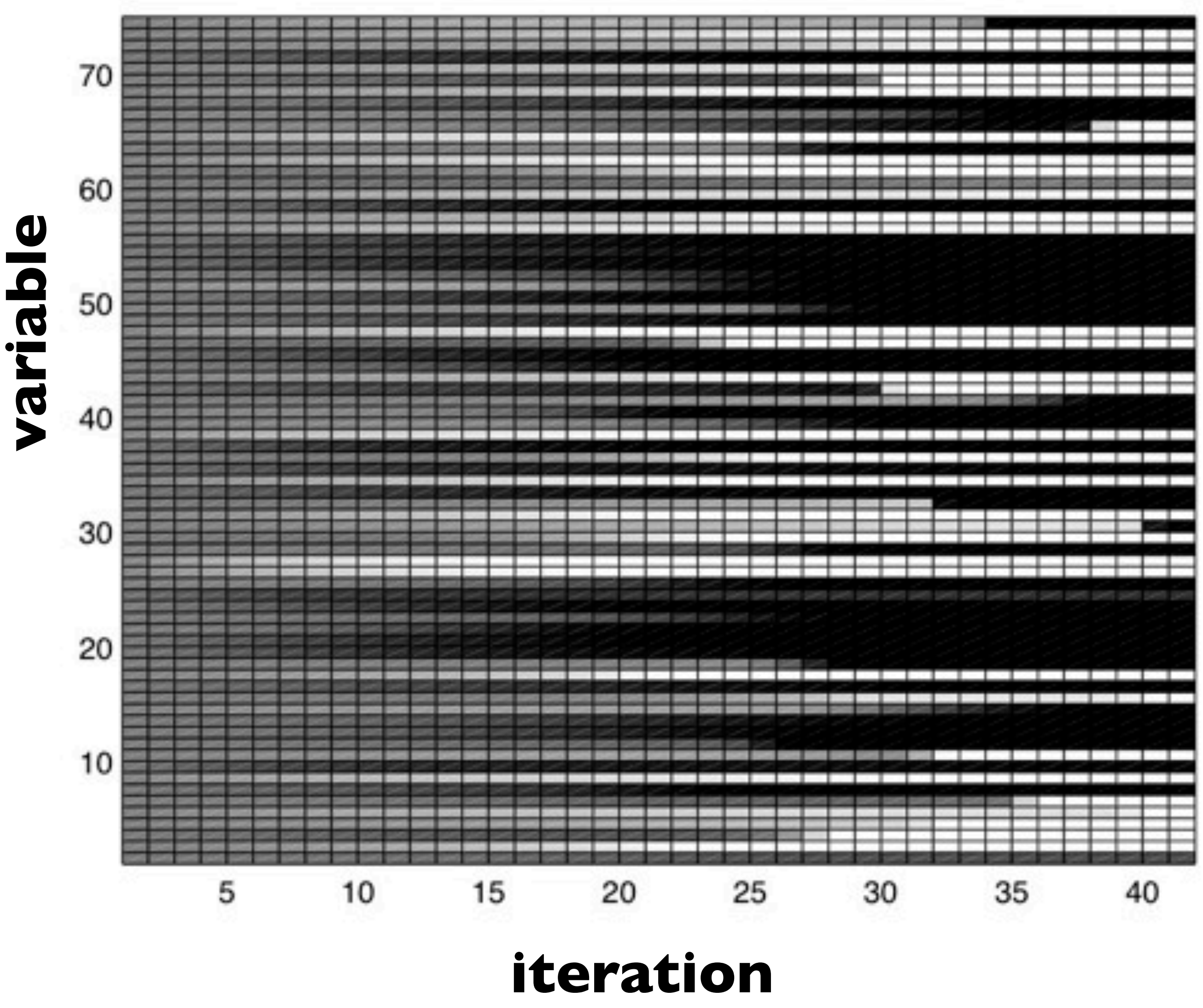
Experimental Results



MaxSAT

Analytically update of dist's is possible.

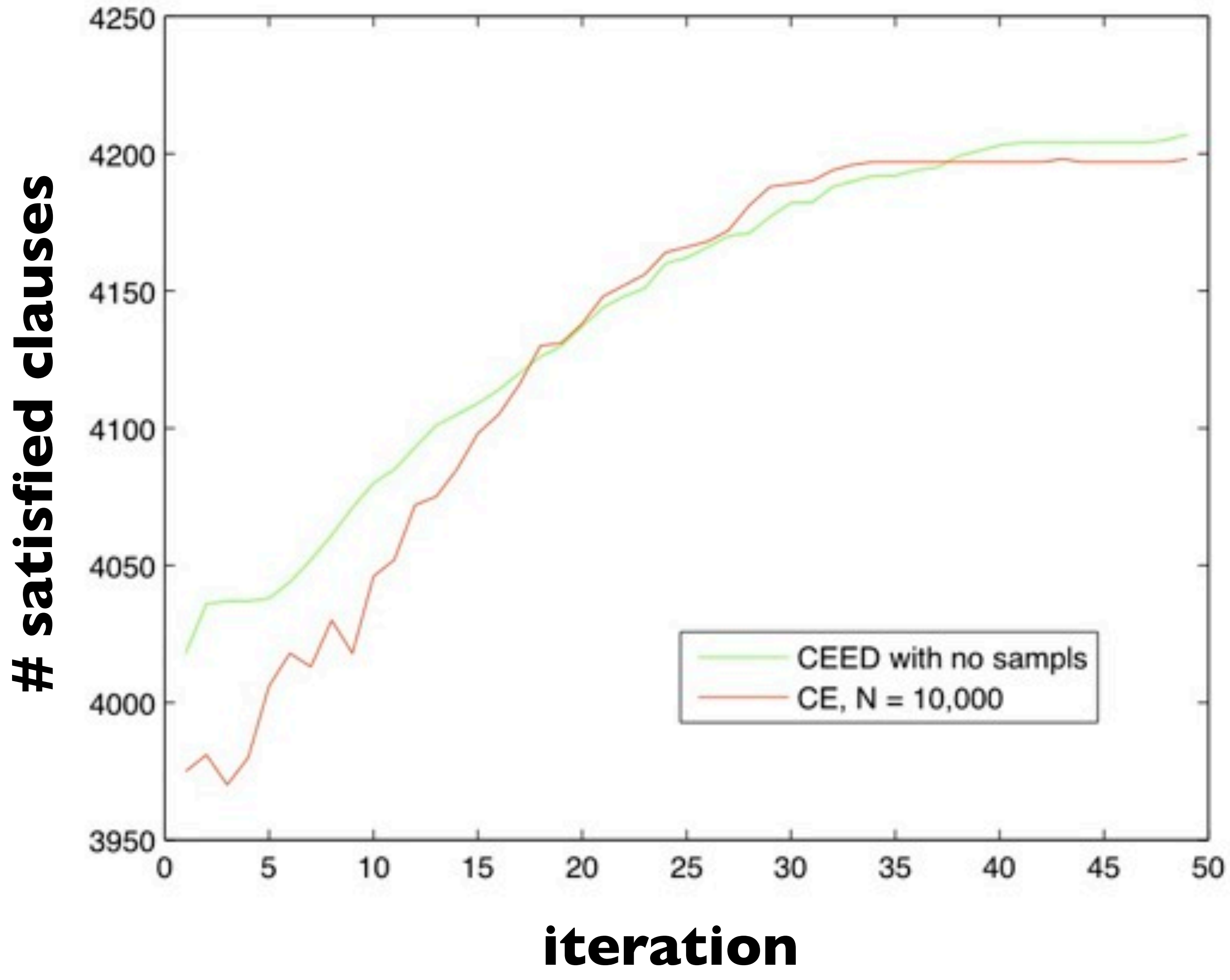
Convergence of distributions to the correct assignment



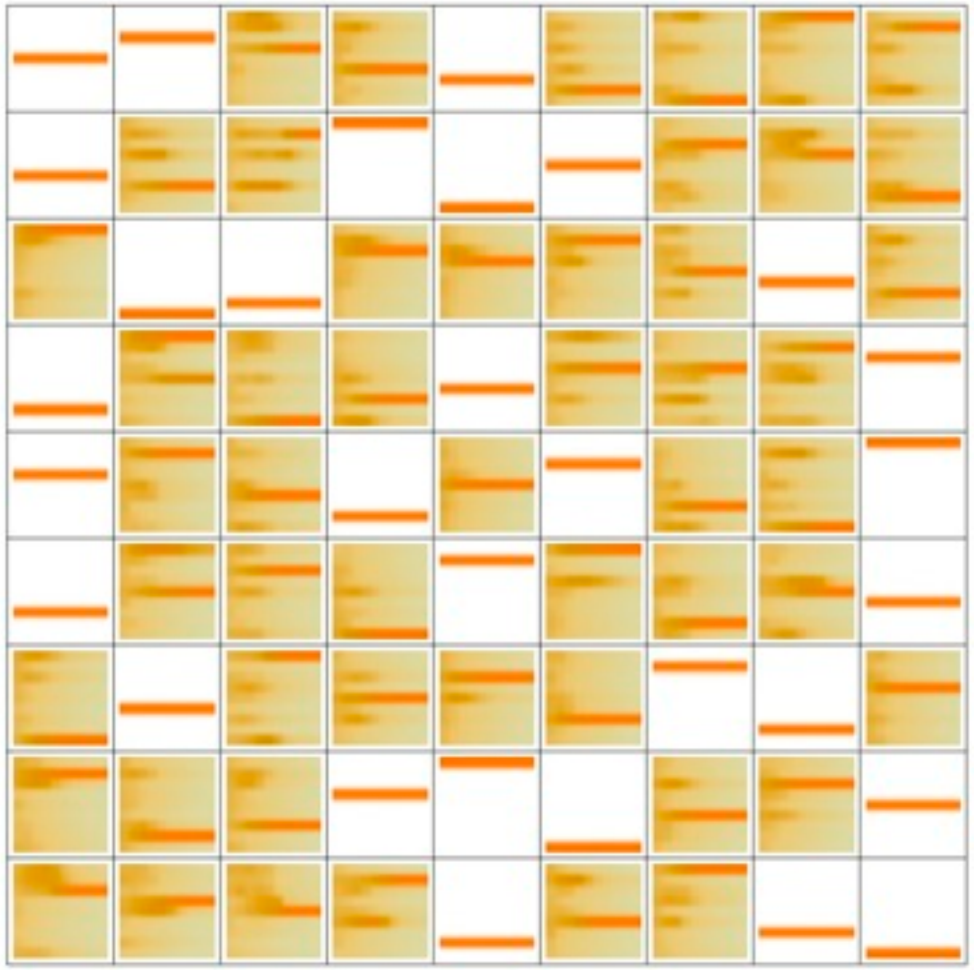
MaxSAT

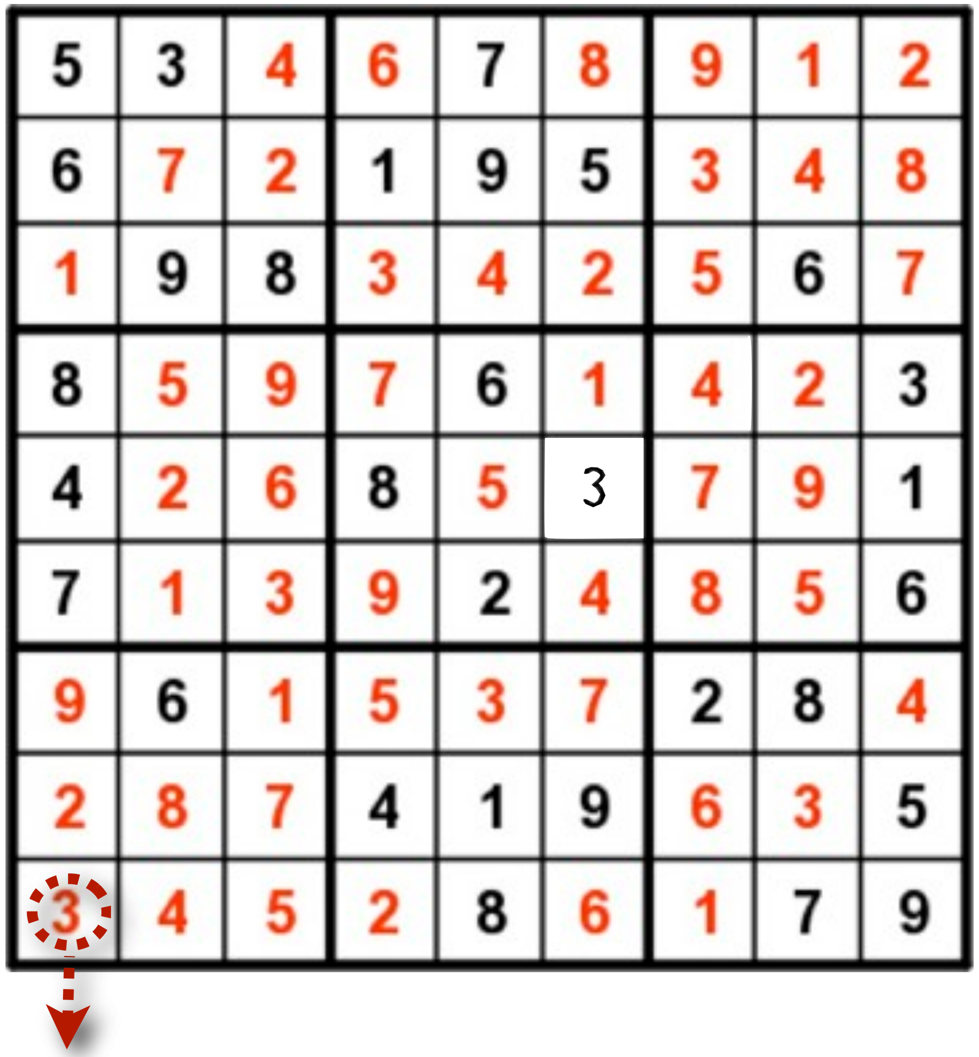
Convergence of CE and CEED

#clauses	4250
#vars	1000



Sudoku





5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

solution in red

Using categorical distribution
Our Alg. is **5** times faster than **CE**

Conclusion

Our method successfully exploits partial decomposability in targeted profiling of NMR spectra
as well as some combinatorial problems (i.e. SAT & Sudoku)

Maximum likelihood estimates could be linearly combined based on their certainty using their Fisher Information.

Thank you!

Computing Science Dept. University of Alberta

Alberta Ingenuity Centre for Machine Learning