Estimation of error propagation and prediction intervals in Multivariate Curve Resolution Alternating Least Squares using resampling methods

Joaquim Jaumot, Raimundo Gargallo and Romà Tauler*

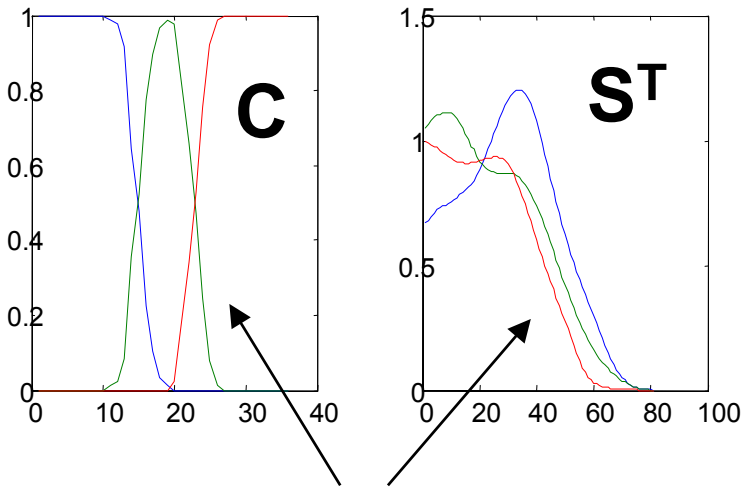
Department of Analytical Chemistry

University of Barcelona

Outline:

- Introduction
- Rotational ambiguities and feasible bands
- Error propagation and resampling methods
- Results
- Conclusions

Multivariate (*Soft*) *Self Modeling* *Curve* Resolution



NC NC

C S^T E

NR + NR

NC

NR D

$$d_{ij} = \sum_{k=1}^{N} c_{ik} s_{kj} + e_{ij}$$

Bilinearity!

Multivariate (*Soft*) *Self Modeling Curve* Resolution

- **Multivariate Curve Resolution (MCR)** methods have been shown to be powerful self-soft-modeling tools able to investigate complex chemical systems with a minimum number of assumptions.

- **Alternating Least Squares (ALS)** has become a popular method for Multivariate Curve Resolution (MCR) due to its flexibility in constraint implementation during the optimization of resolved profiles.

Multivariate (*Soft*) *Self Modeling Curve* Resolution

- **What are the reliability of MCR-ALS estimations?**
  - Do the MCR-ALS solutions have rotational and scale freedom?
  - Are they unique solutions or exist instead a band of feasible solutions?
  - How errors and noise are propagated from experimental data to ALS estimations?

Goals of this study

- Find the reliability of ALS resolved profiles in multivariate curve resolution.

- Estimate prediction error intervals for ALS profiles

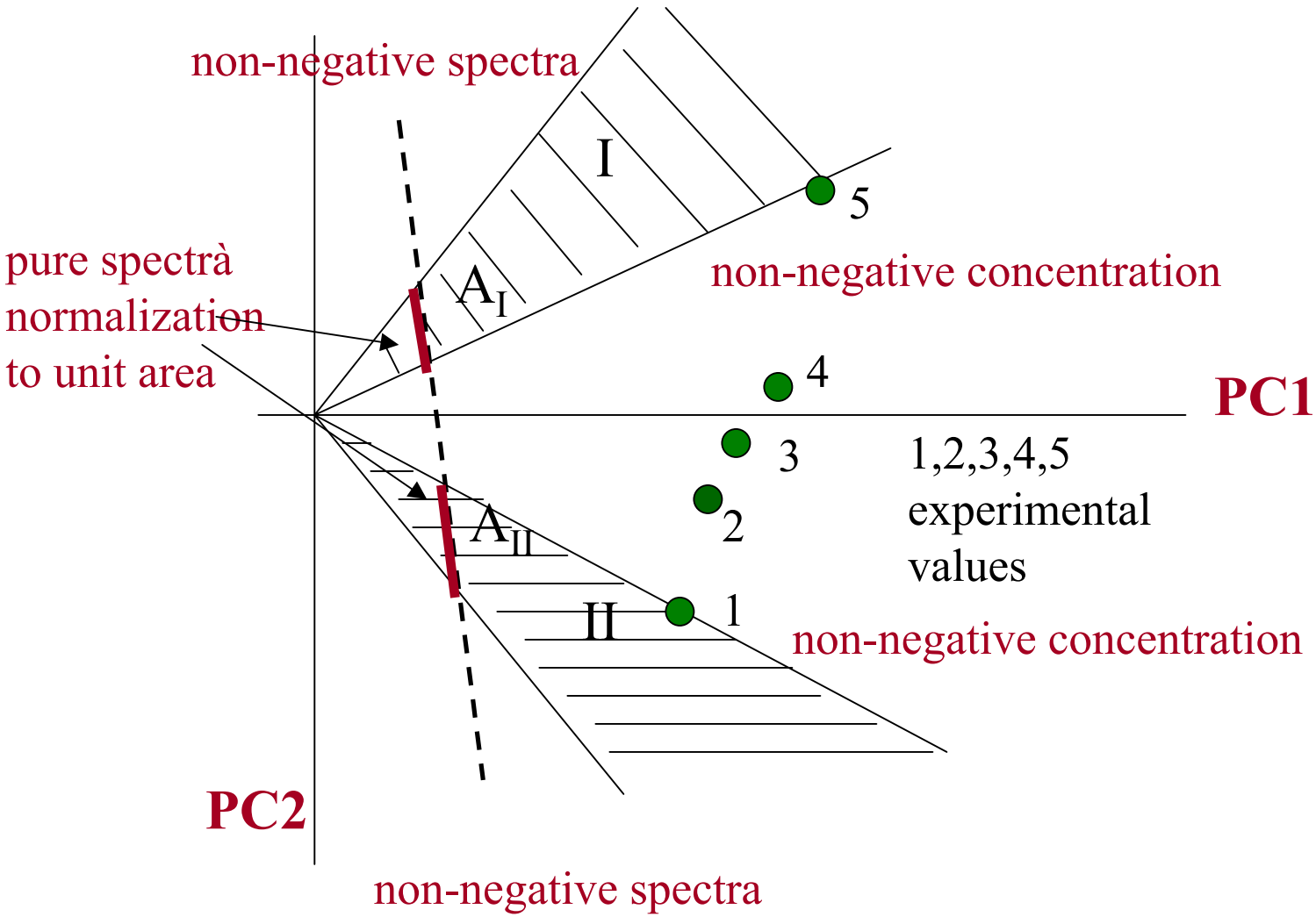
- Estimate prediction error intervals for parameters calculated from MCR-ALS resolved profiles

- Investigate the interaction between propagation of errors and rotational ambiguities (noise effects on rotational ambiguities and constraints).

Outline:

- Introduction
- **Rotational ambiguities and calculation of feasible bands**
- Error propagation and resampling methods
- Results
- Conclusions

Lawton and Sylvestre feasible bands



W.H. Lawton and EA Sylvestre, Technometrics 1971, 13, 617-33

Rotational Ambiguities

Factor Analysis (PCA) Data Matrix Decomposition

$$D = U V^T + E$$

'True' Data Matrix Decomposition

$$D = C S^T + E$$

$$D = U T T^{-1} V^T + E = C S^T + E$$

$$C = U T; \quad S^T = T^{-1} V^T$$

How to find the rotation matrix T?

Matrix decomposition is not unique!

T(N,N) is any non-singular matrix

There is rotational freedom for **T**

Rotational Ambiguities

Because of rotational ambiguities instead of unique solutions, a set of ***feasible solutions*** are obtained

Feasible solutions are different solutions that fit equally well the data under a set of constraints

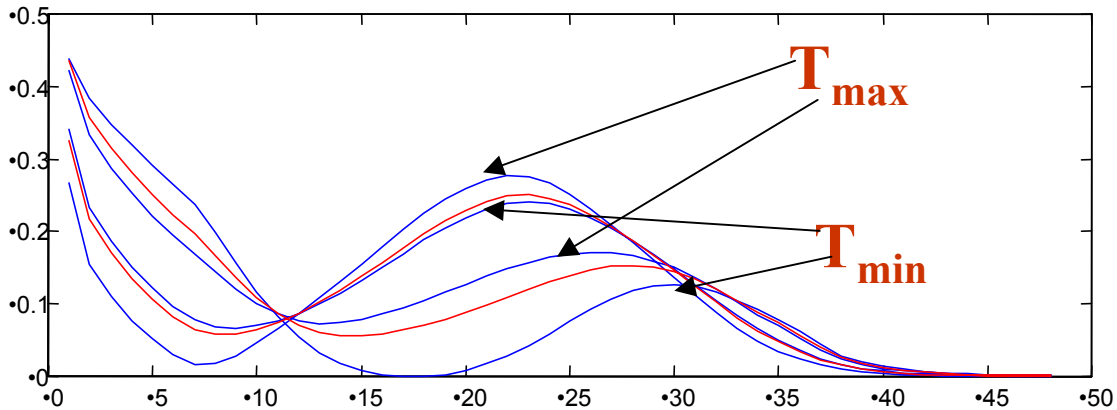
For a particular system under a set of constraints, feasible solutions are defined from a set of possible **T** values.

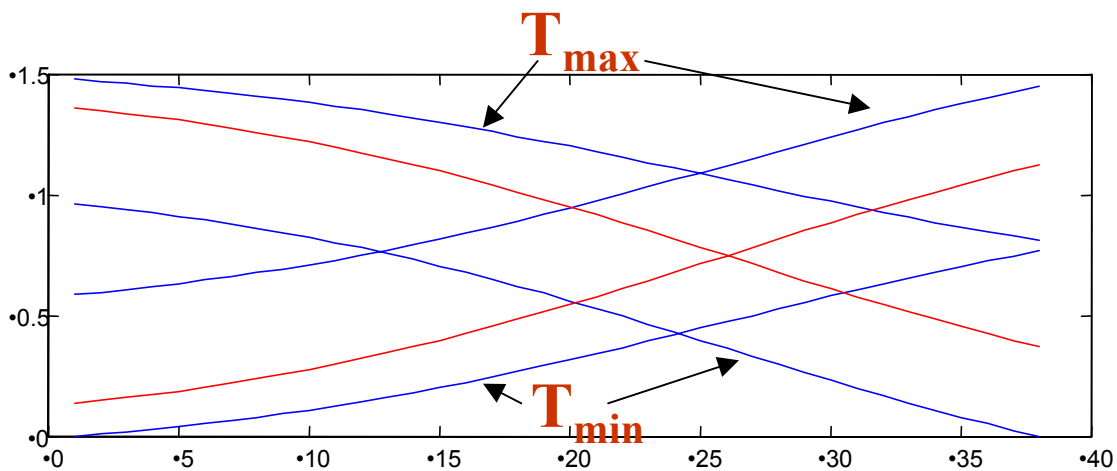
Rotational Ambiguities

- **T** values define the ***band of feasible solutions* or *feasible bands***
  - How to define the ***boundaries*** of these feasible bands?
  - How to ***represent graphically*** these boundaries?

Is it possible to define band boundaries (T_{max} and T_{min})?

How to calculate T_{max} and T_{min}?





How to define and find the band boundaries?

- What are the **T** values giving the ***maximum/outer and minimum/ inner boundaries*** of the feasible bands under a set of constraints?

$$\mathbf{D}^{*} = \mathbf{C}_{inic}\,\mathbf{S}^{T}_{inic} =$$
$$= \mathbf{C}_{inic}\,\mathbf{T}_{min}\,\mathbf{T}^{-1}_{min}\,\mathbf{S}^{T}_{inic} = \mathbf{C}_{min}\mathbf{S}^{T}_{min} =$$
$$= \mathbf{C}_{inic}\,\mathbf{T}_{max}\,\mathbf{T}^{-1}_{max}\,\mathbf{S}^{T}_{inic} = \mathbf{C}_{max}\mathbf{S}^{T}_{max}$$

where: $\mathbf{D}(NR,NC)$, $\mathbf{C}(NR,N)$, $\mathbf{S}^{T}(N,NC)$, $\mathbf{T}(N,N)$

How to define and evaluate $\mathbf{T}_{max}$ and $\mathbf{T}_{min}$?

Evaluation of boundaries of feasible bands: Previous studies

- W.H.Lawton and E.A.Sylvestre, Technometrics, 1971, 13, 617-633
- O.S.Borgen and B.R.Kowalski, Anal. Chim. Acta, 1985, 174, 1-26
- K.Kasaki, S.Kawata, S.Minami, Appl. Opt., 1983 (22), 3599-3603
- R.C.Henry and B.M.Kim (Chemomet. and Intell. Lab. Syst., 1990, 8, 205-216)
- P.D.Wentzell, J-H. Wang, L.F.Loucks and K.M.Miller (Can.J.Chem. 76, 1144-1155 (1998))
- P. Gemperline (Analytical Chemistry, 1999, 71, 5398-5404)
- R.Tauler (J.of Chemometrics 2001, 15, 627-46)
- M.Legger and P.D.Wentzell, Chemomet and Intell. Lab. Syst., 2002, 171-188

Definition of band boundaries

The whole measured signal is:

$$\mathbf{D} = \sum \mathbf{D_i} = \sum \mathbf{c_i}\ \mathbf{s_i}^{\mathbf{T}}$$

The contribution of each species to the whole signal is:

$$\mathbf{D_i} = \mathbf{c_i s_i}^{\mathbf{T}}$$

Solving the Optimization Problem:

max/outer boundary: Find $\mathbf{T_{max}}$ that makes $\mathbf{c_i}\ \mathbf{s_i}^{\mathbf{T}}$ maximum

min/inner boundary: Find $\mathbf{T_{min}}$ that makes $\mathbf{c_i}\ \mathbf{s_i}^{\mathbf{T}}$ minimum

Constrained Non-Linear Optimization Problem (NCP)

Find **T** which makes:

$$\min_{\mathbf{T}} / \max f(\mathbf{T}) \quad \text{subject to} \quad g_e(\mathbf{T}) = 0 \quad \text{and to} \quad g_i(\mathbf{T}) \leq 0$$

where **T** is the matrix of variables, f(**T**) is a non-linear scalar function of **T** and **g(T)** is the vector of constraints (non-linear function of **T**)

Matlab Optimizarion Toolbox ***fmincon*** function

1) What are the variables of the problem?

T (rotation matrix),

$$\mathbf{D = C\ T\ T^{-1}\ S^{T}}$$

2) What is the objective function f(T) to be optimized?

For each species i = 1,..,ns

$$f_i(\mathbf{T}) = \frac{\|c_i s_i\|}{\|C\ S^T\|} \quad \text{or} \quad f_i(\mathbf{T}) = \frac{\sum_j c_{ij} s_{ij}}{\sum_{i,j} c_{ij} s_{ij}}$$

This gives the relative signal contribution of species i respect the global measured signal !

f(T) is scalar value between 0 and 1!

3) What are the constraints g(T)?

The following constraints may be considered:

normalization/closure	g_{norm}/g_{clos}
non-negativity	g_{cneg}/g_{sneg}
known values/selectivity	g_{known}/g_{sel}
unimodality	g_{unim}
trilinearity (three-way data)	g_{tril}

Are they equality or inequality constraints?

4) What are the initial estimates of C, S^T?

- Initial estimates of C and S^T are obtained by MCR-ALS
- Initial estimates are feasible solutions fulfilling the constraints of the system (*non-negativity, unimodality, closure, selectivity, local rank,…)*

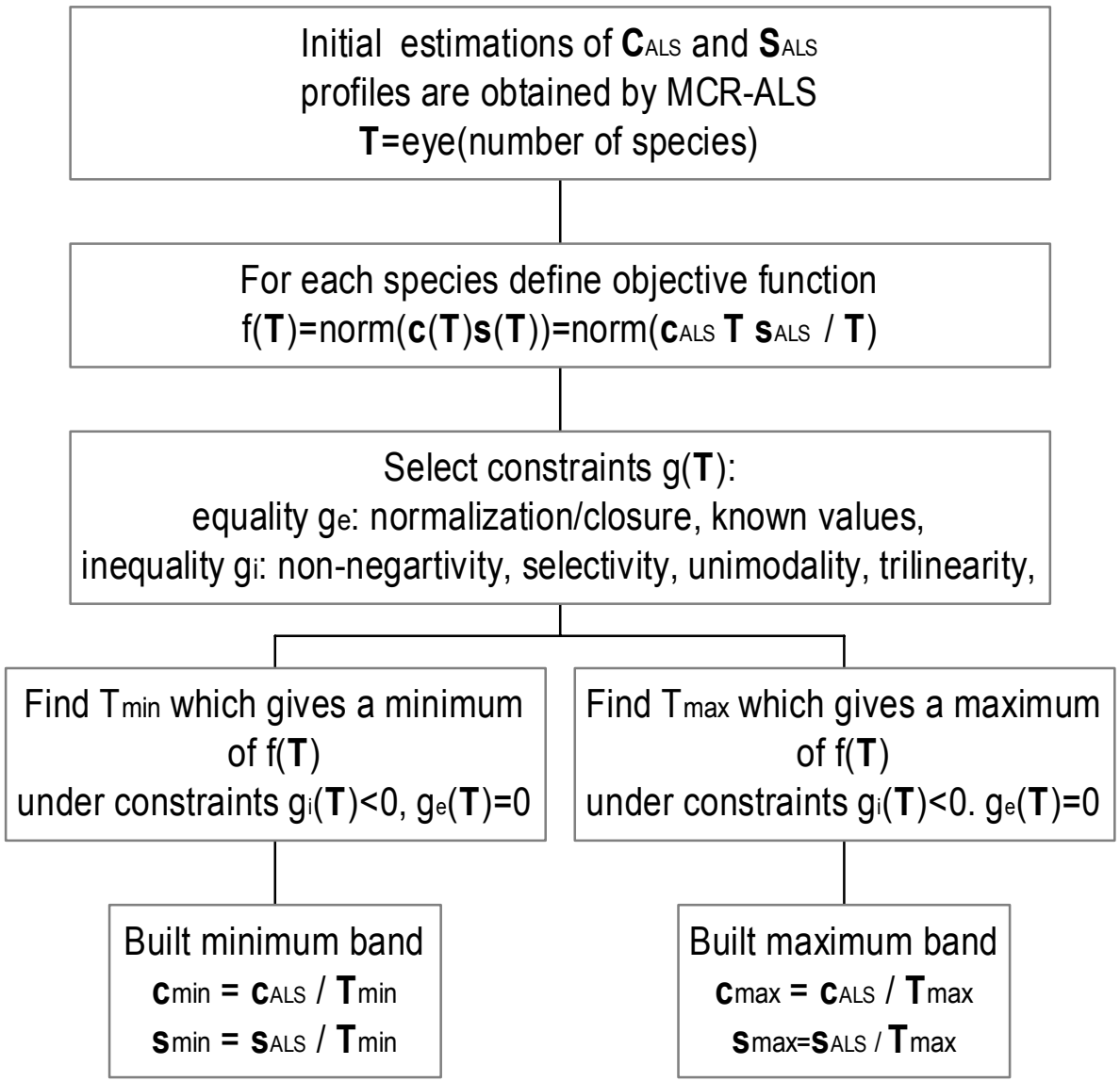
5) What are the initial values of T?

- NCP depends on initial estimates of **T**! (local minima, convergence, speed …)

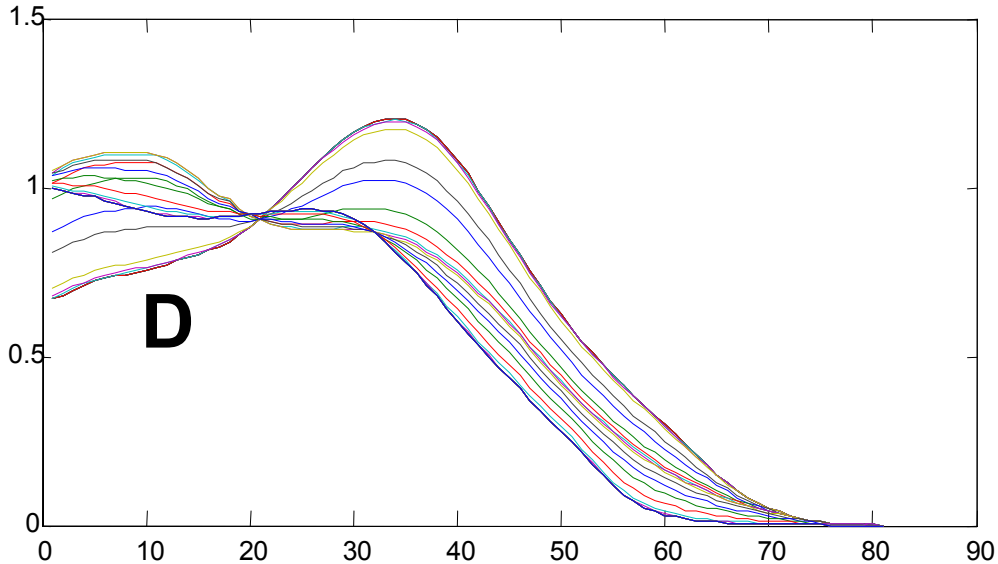
$$\mathbf{T_{ini} = eye(N) =} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Optimization algorithm

•R.Tauler (J.of Chemometrics 2001, 15, 627-46)

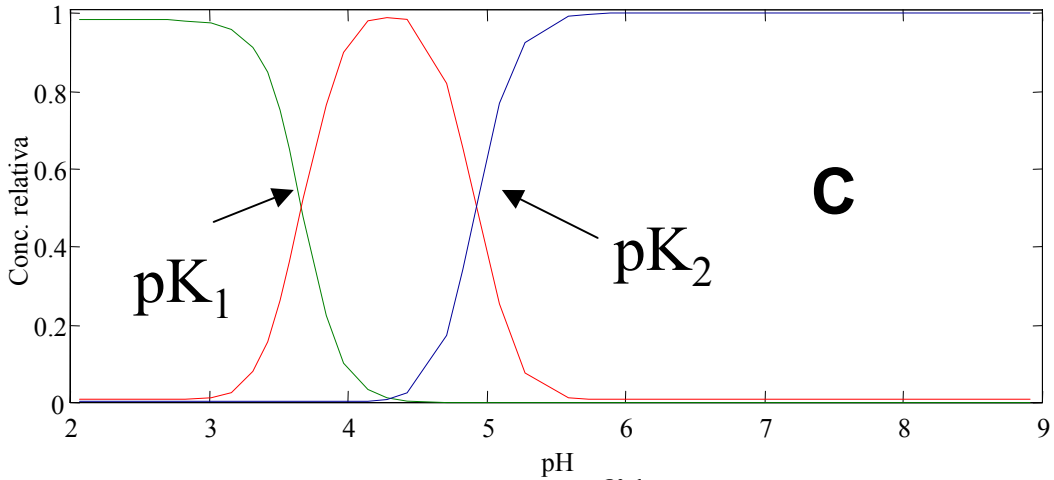


Experimental data system under study

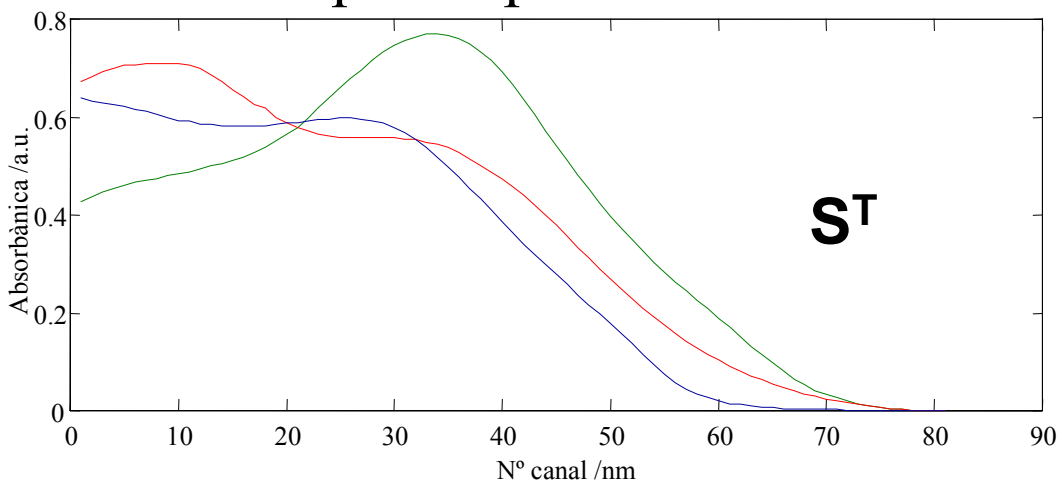


Acid-base spectrophotometric titration of the double stranded heteropolynucleotide polyinosinic-polycytidylic acid. Spectral region between 240-320 nm and pH region between pH 2 and pH 9

concentration profiles



spectra profiles



Application of MCR-ALS to the experimental data matrix D

Applied constraints in ALS were:

a) non-negative spectra
b) non-negative concentrations
c) closure in concentrations

Initial estimates were obtained from purest variables

- **This system has selectivity! local rank resolution conditions!**
- **Initial estimates from pure variable detection methods provide good initial estimates that produce solutions close to the true profiles**

Parameter estimation

Mass-action law is only assumed at the site level and not for the whole polynucleotide molecule

Evaluation of constants
from intersection profiles

pK_1	3.6660
pK_2	4.9244

Proposed species:

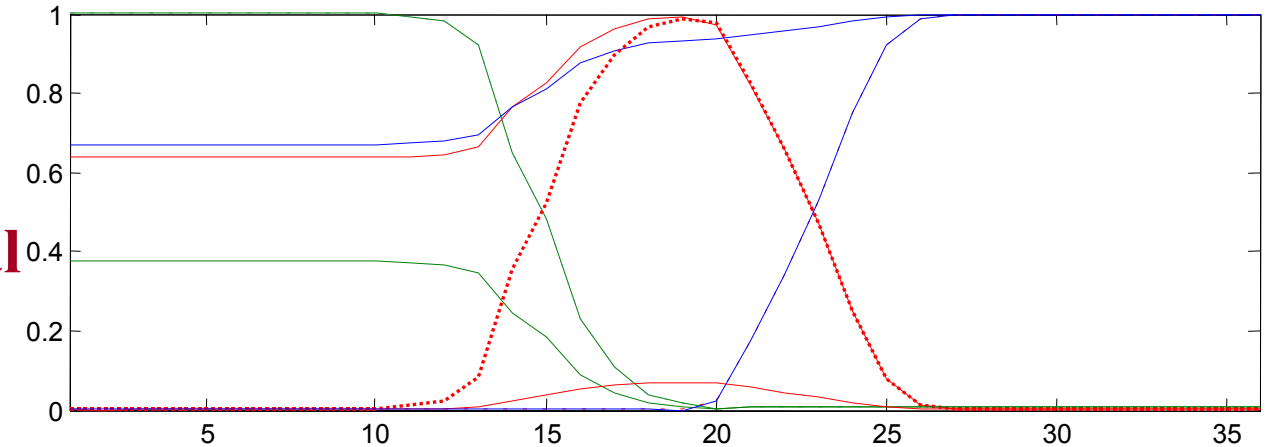
poly(I)-poly(C+) ⇔

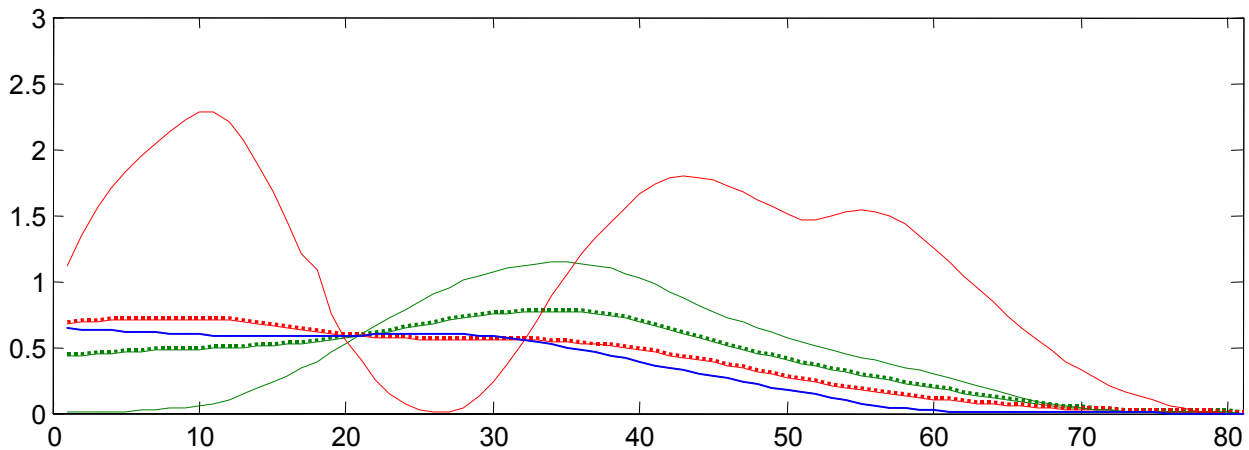
poly(I)-poly(C)-poly(C+) + H ⇔

poly(I)-poly(C) + H

Estimation of band boundaries (max/min contribution of each species) of feasible solutions

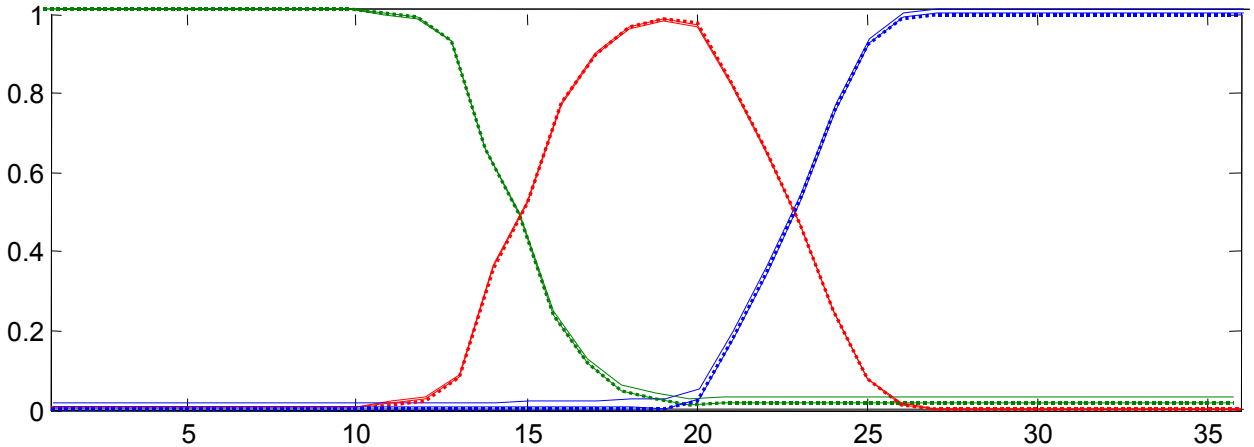
Large Rotational ambiguities were present when constraints applied were only closure non-negativity!!!

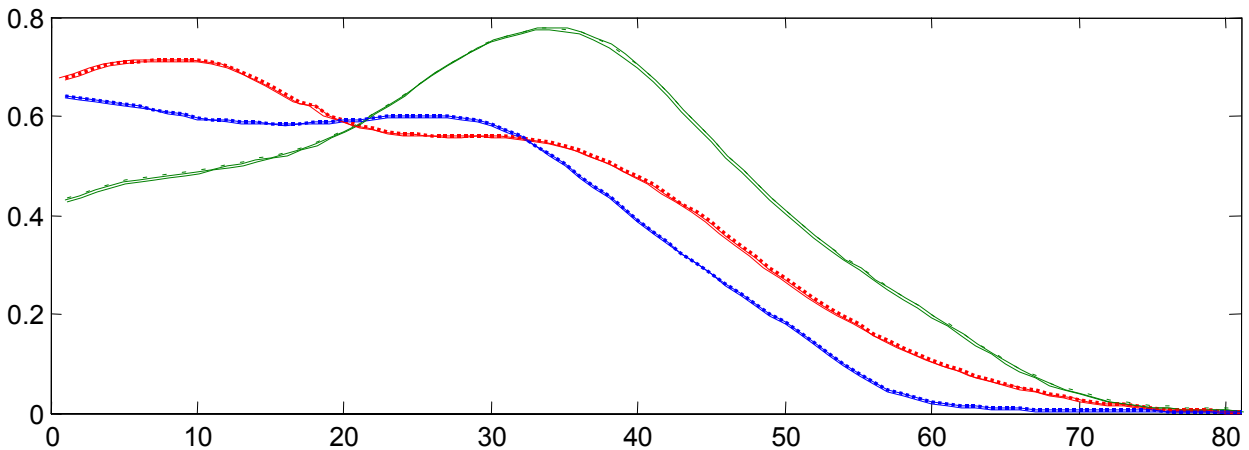




Estimation of band boundaries (max/min contribution of each species) of feasible solutions

Rotational ambiguities nearly dissappear when selectivity constraint was applied!!!





Outline:

- Introduction
- Rotational ambiguities and feasible bands
- Error propagation and resampling methods
- Results
- Conclusions

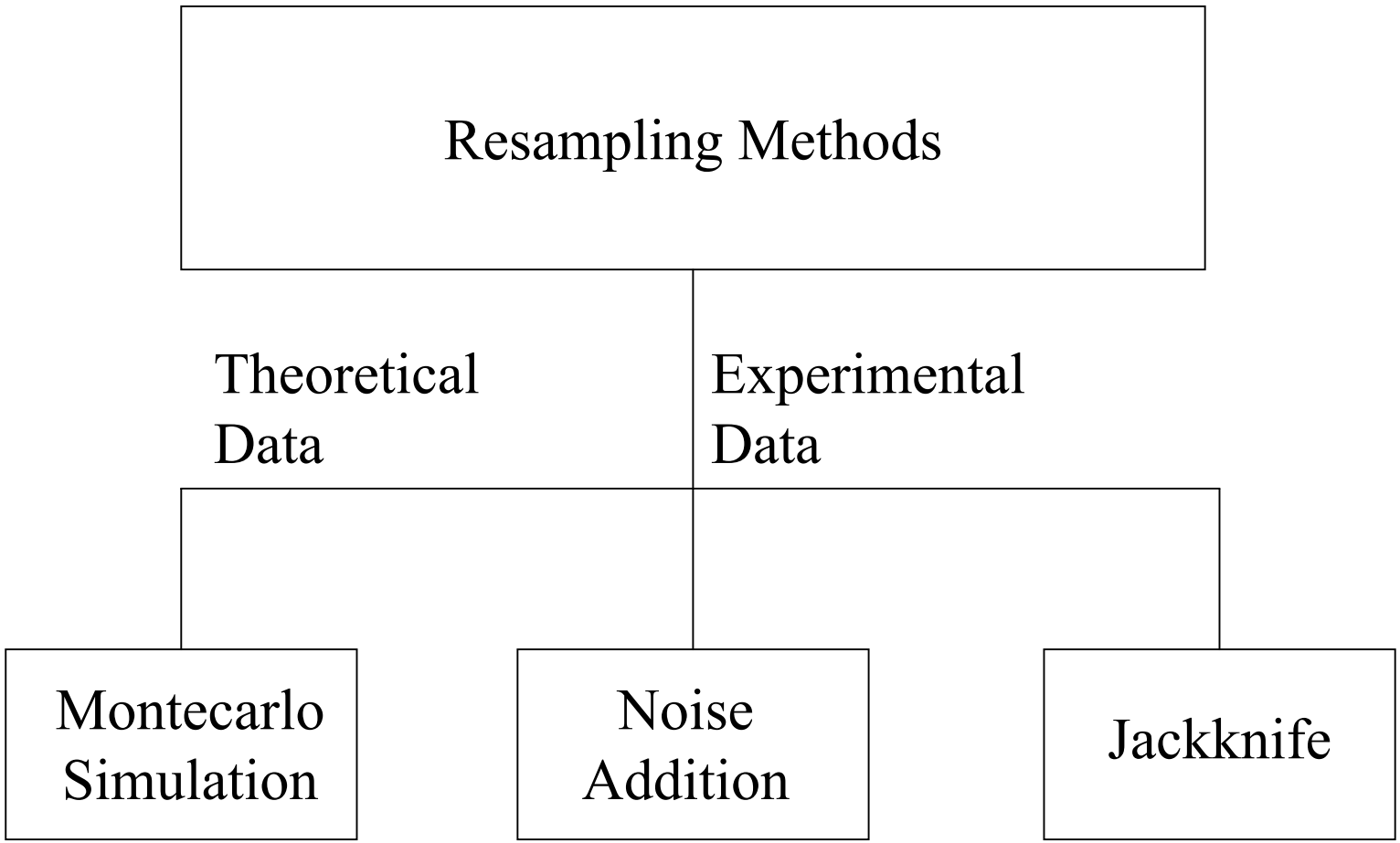
Error propagation and resampling methods

- How experimental error/noise in the input data matrices affects MCR-ALS results?

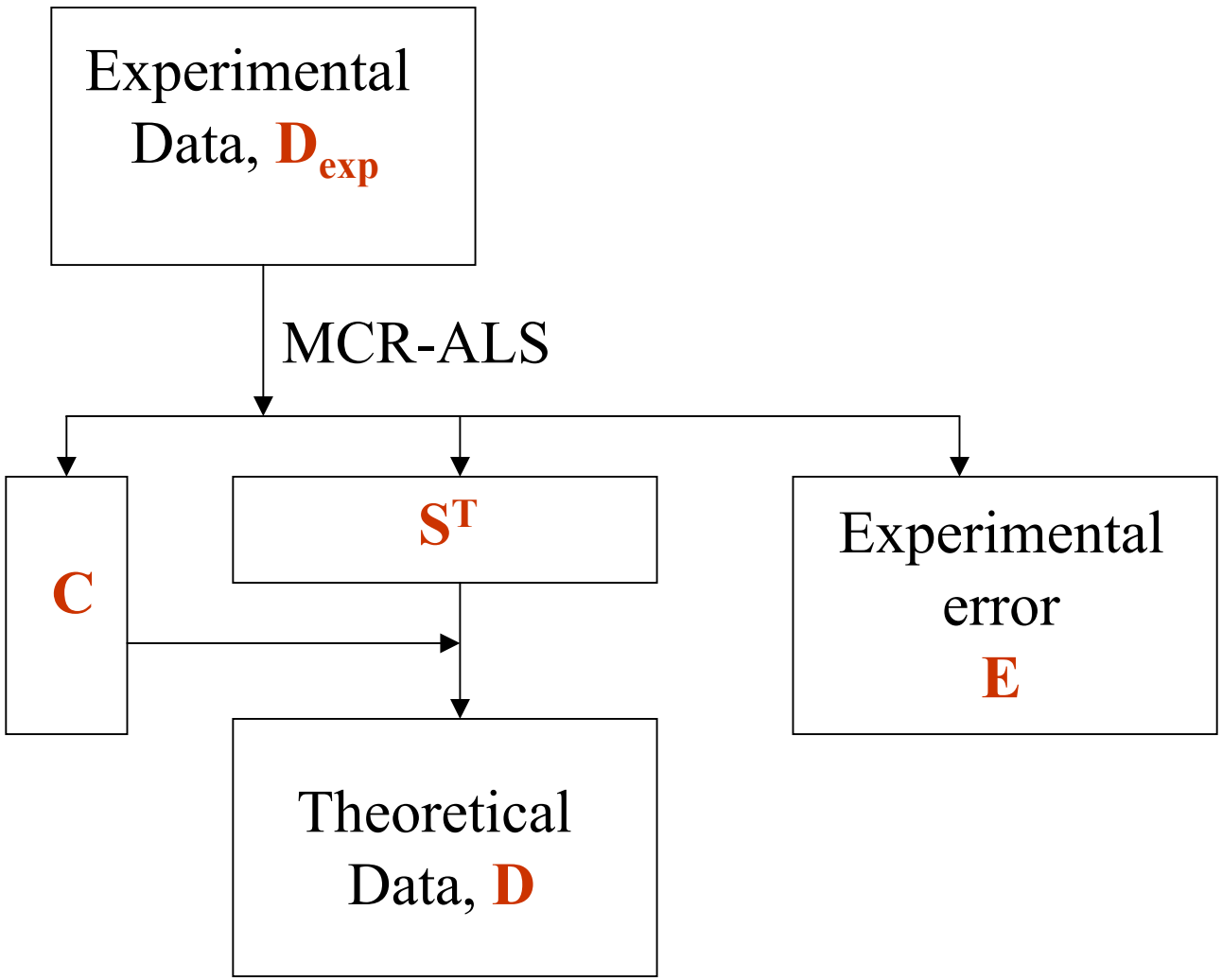
  - For ALS calculations there is no known analytical formula to calculate error estimations. (i.e. like in linear lesast-squares regressions)

  - Bootstrap estimations using resampling methods is attempted

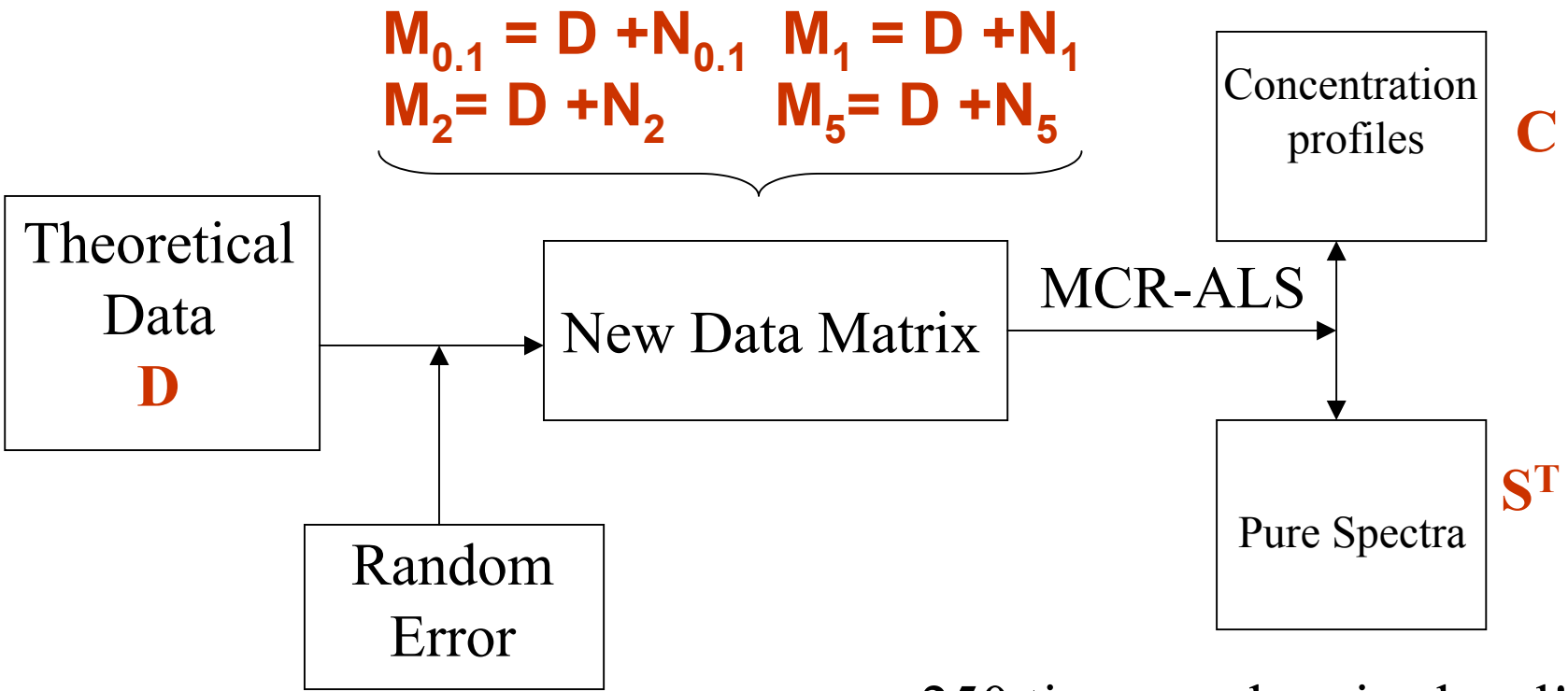
Resampling Methods



Building theoretical data



Montecarlo Simulations

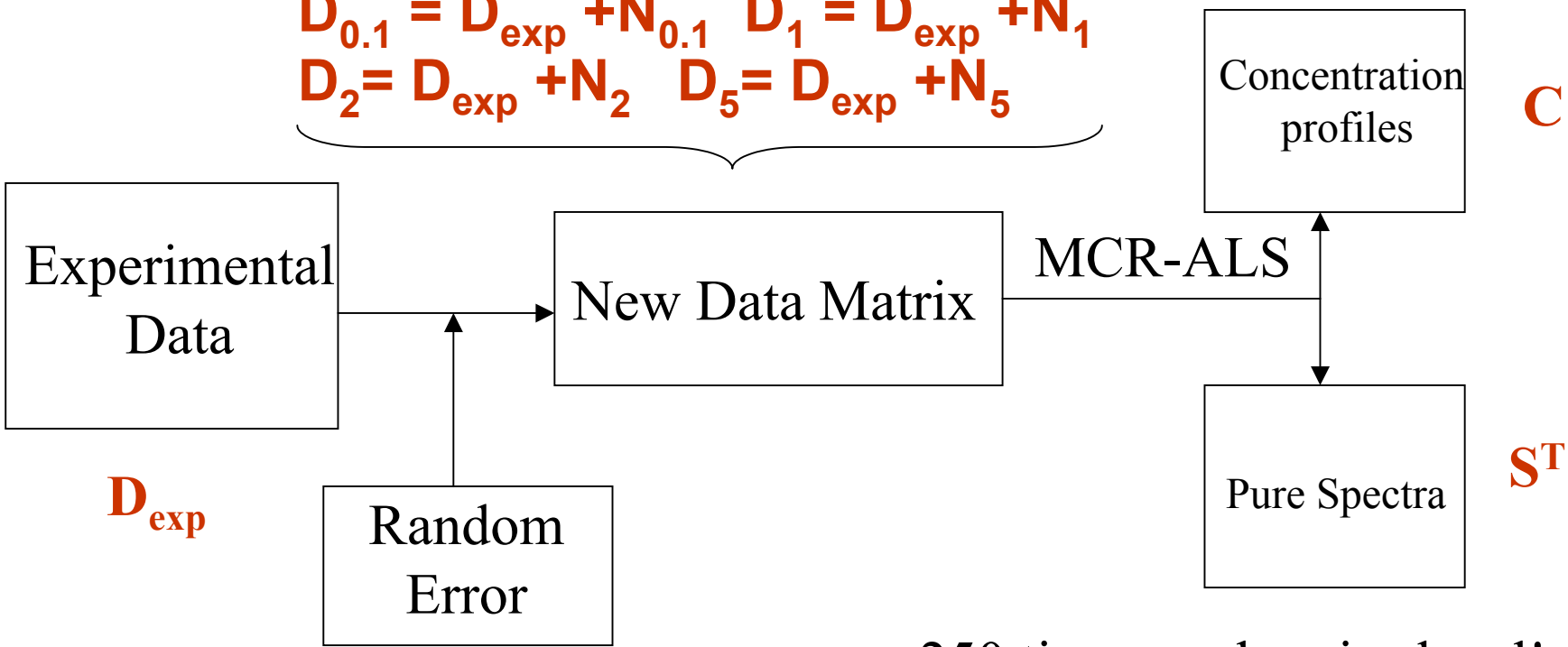


250 times each noise level!
1000 simulations!

MATLAB function ***randomn*** with zero mean and relative sd 0.1%, 1%, 2% and 5% of maximum signal in **D**

Noise Addition Simulations

$D_{0.1} = D_{exp} + N_{0.1}$ $D_1 = D_{exp} + N_1$
$D_2 = D_{exp} + N_2$ $D_5 = D_{exp} + N_5$

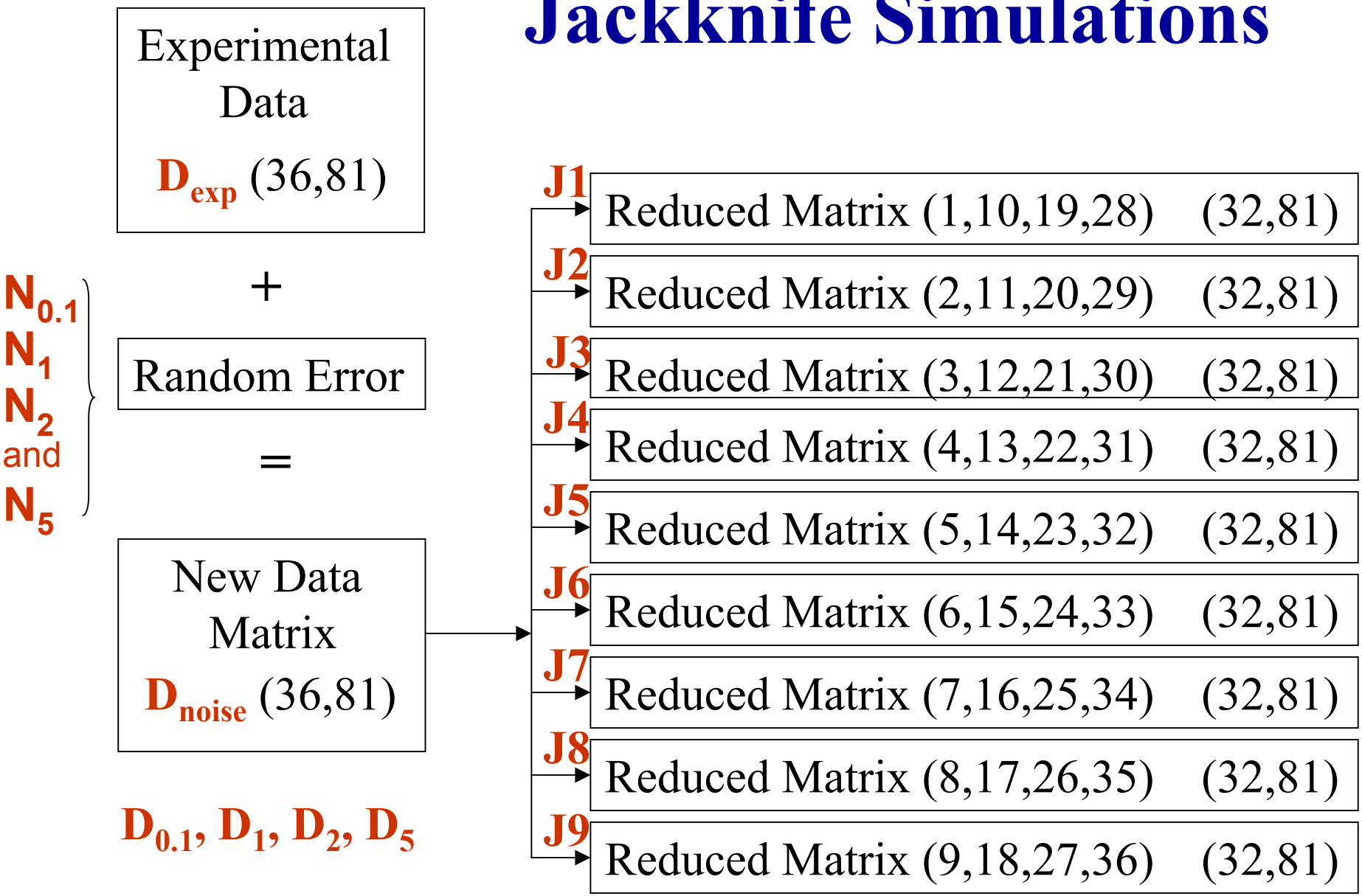


$N_{0.1}$, N_1, N_2 and N_5

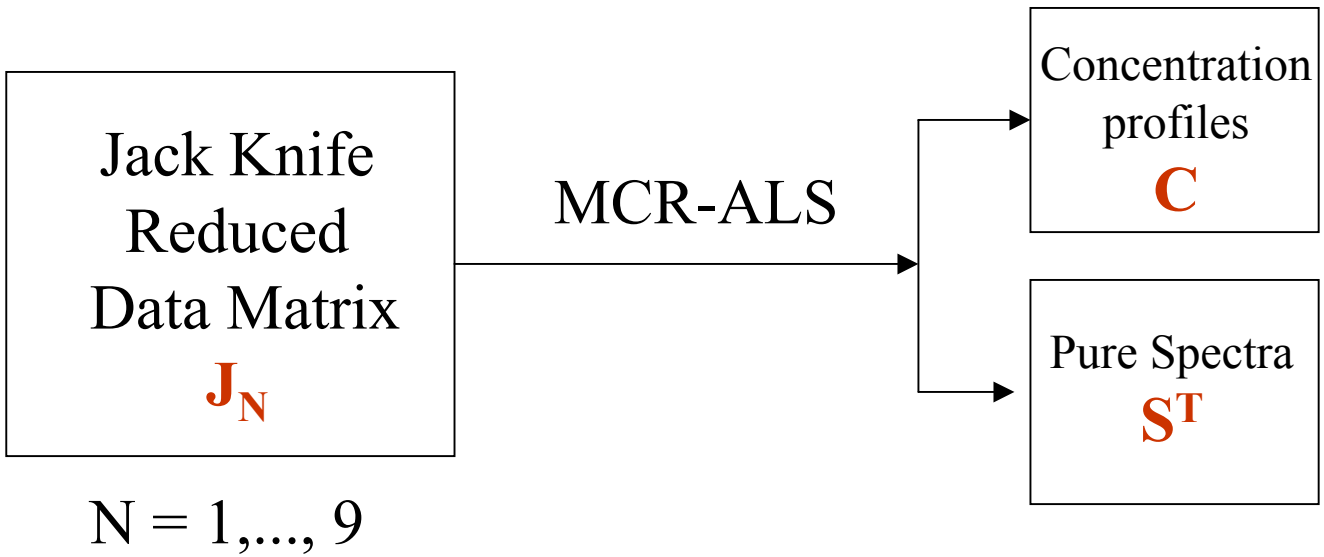
250 times each noise level!
1000 simulations!

MATLAB function ***randomn*** with zero mean and relative sd 0.1%, 1%, 2% and 5% of maximum signal in **D**

Jackknife Simulations



Jackknife Simulations



N = 1,..., 9

Outline:

- Introduction
- Rotational ambiguities and feasible bands
- Error propagation and resampling methods
- **Results**
- Conclusions

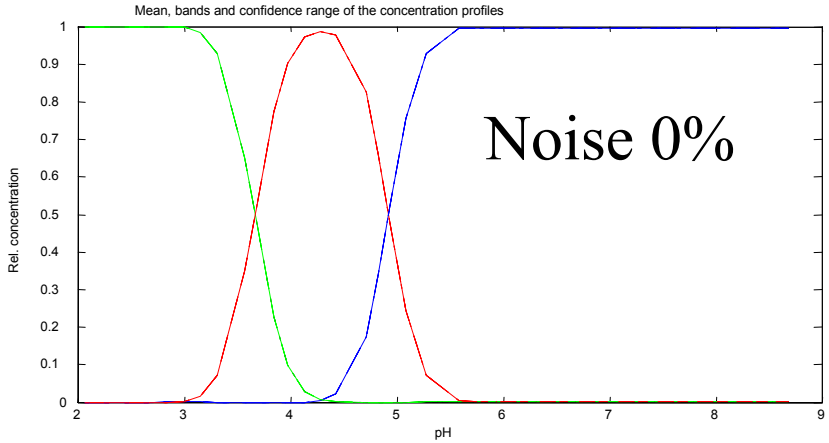
Presentation of Results

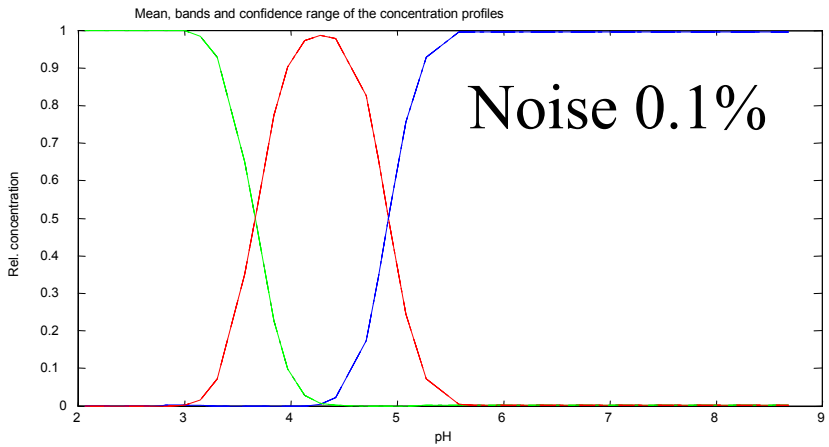
1. Calculation of **species profiles error bands**: Mean profile, maximum and minimum profiles, standard deviation profiles and confidence range profiles

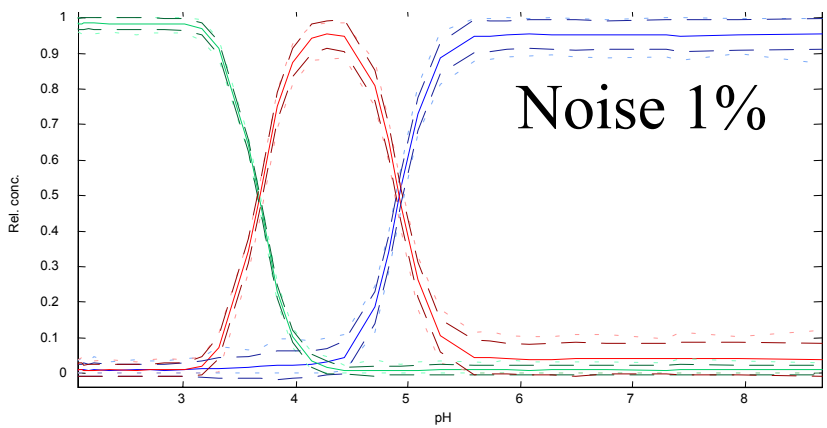
2. **pKa (parameter) error estimations**

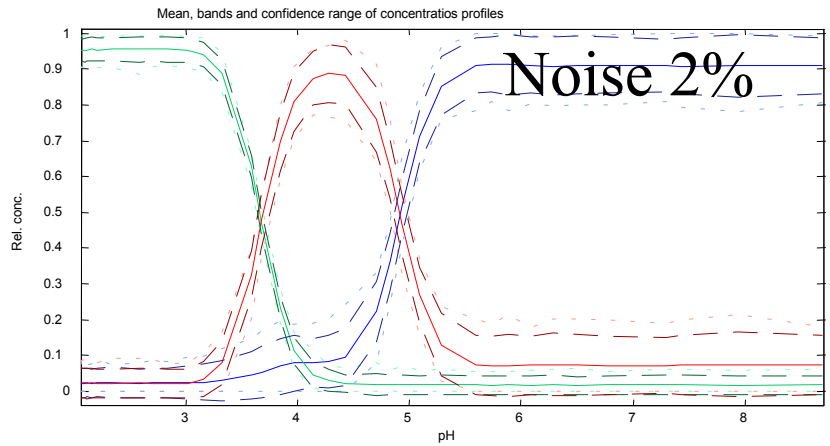
3. Rotational ambiguity effects on error estimates from resampling methods. Calculation of **boundaries of feasible bands from mean species profiles error bands**

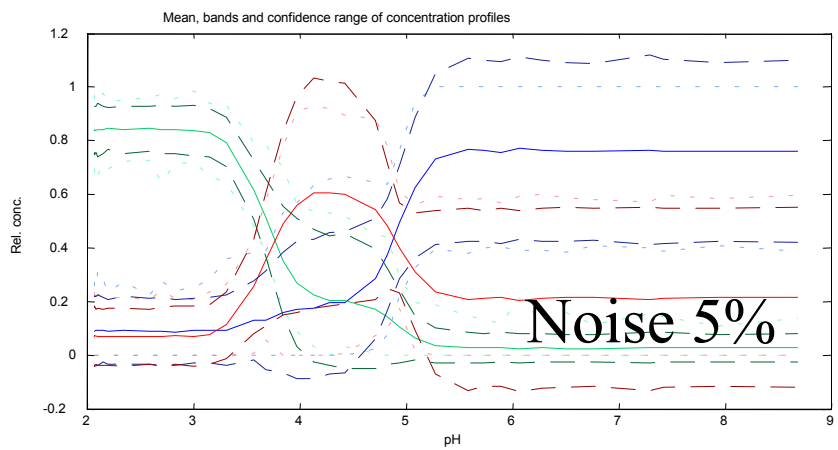
Monte Carlo Simulations
Concentration profiles:
Mean max and min profiles
Confidence range profiles



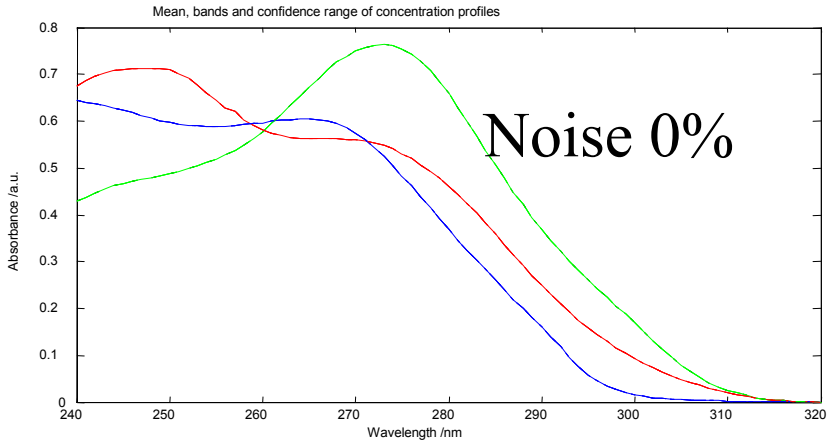


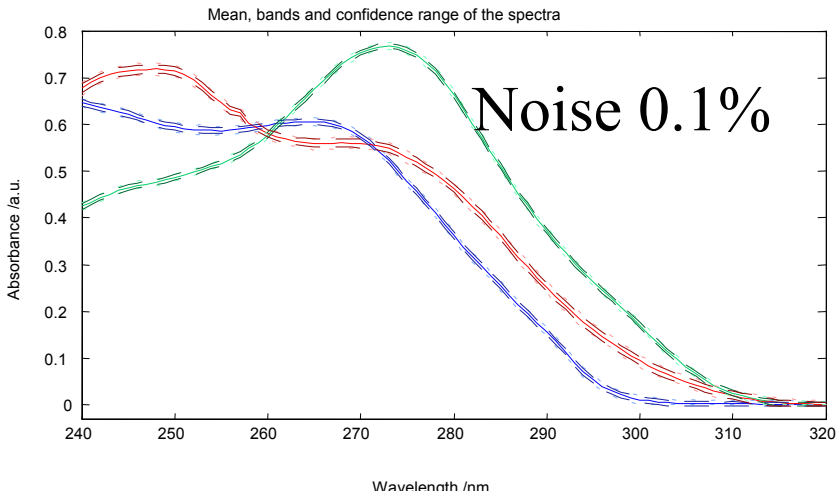


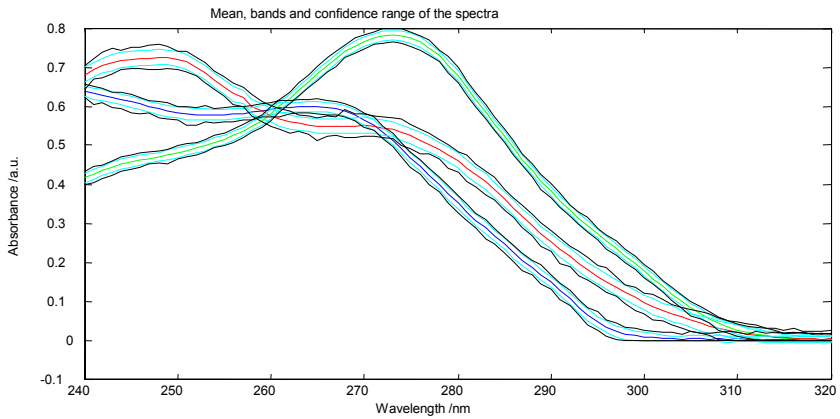


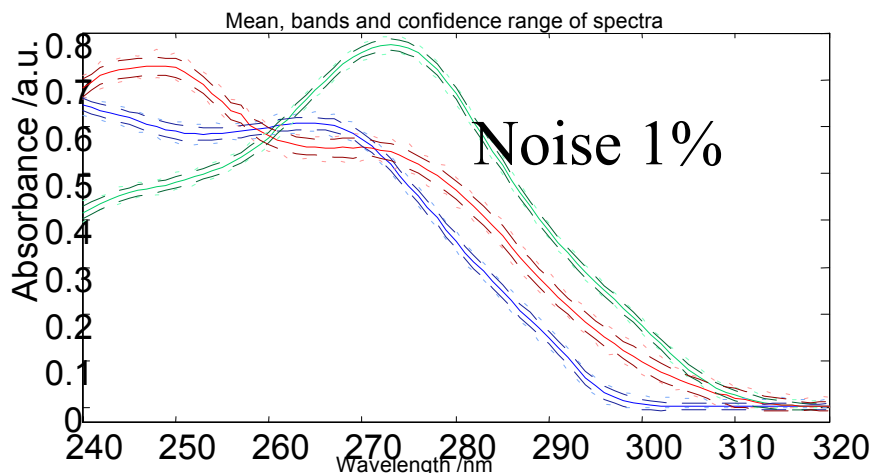


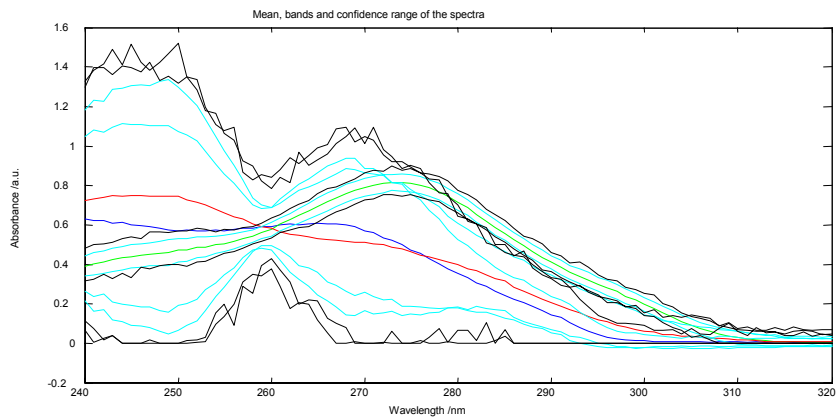
Monte Carlo Simulations
Spectra profiles:
Mean max and min profiles
Confidence range profiles







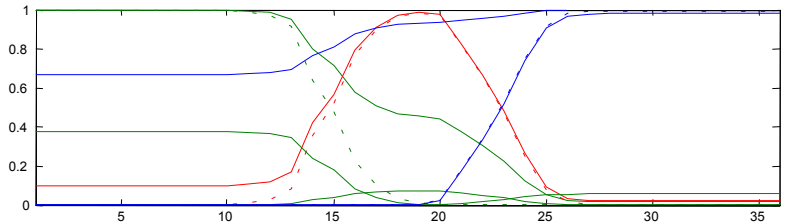


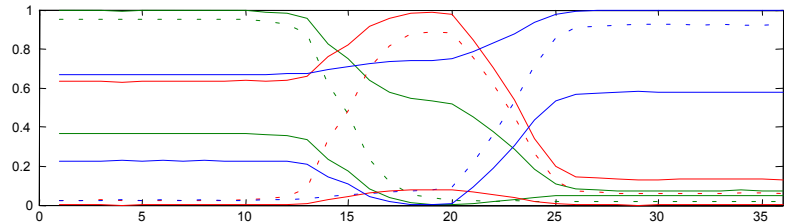


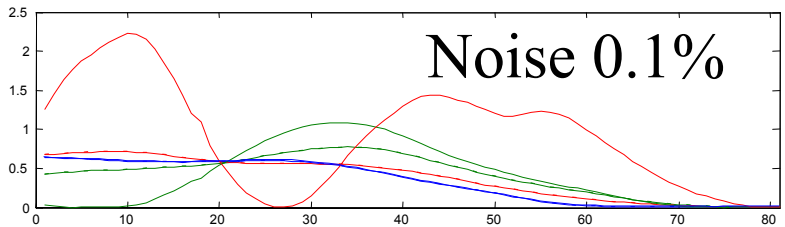
Monte Carlo Simulations
pKa error estimations

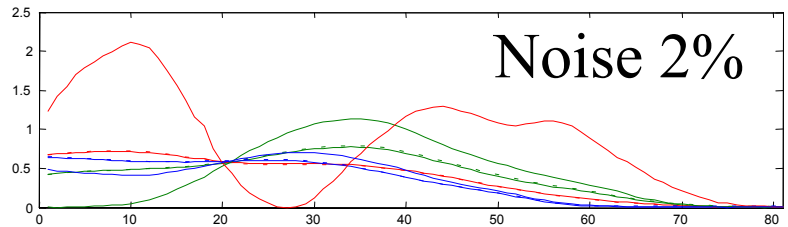
	pK_1		pK_2	
Noise added	Value	Std. dev	Value	Std. dev
0 %	3.6660	4e-15	4.9244	9e-15
0.1 %	3.6662	6e-4	4.9243	0.0012
1 %	3.6696	0.0065	4.9262	0.0128
2 %	3.6761	0.0127	4.9173	0.0245
5 %	3.9762	0.4349	5.0745	0.7595

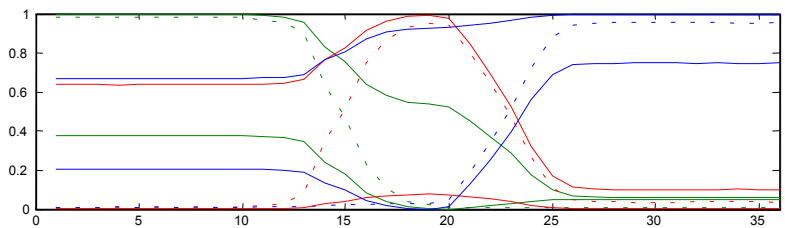
Calculation of band boundaries from mean species profiles error bands (under non-negativity and closure constraints)

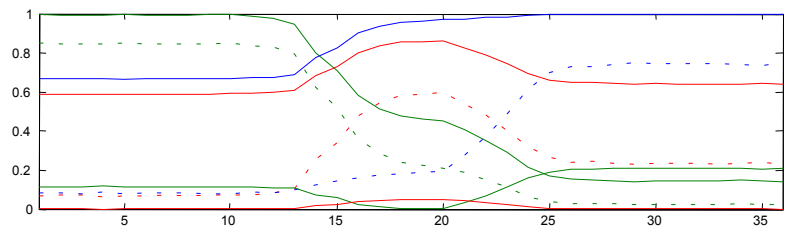


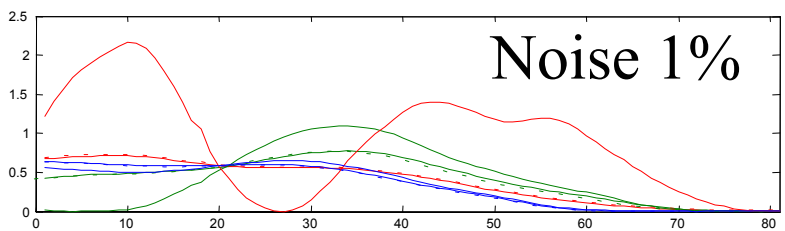


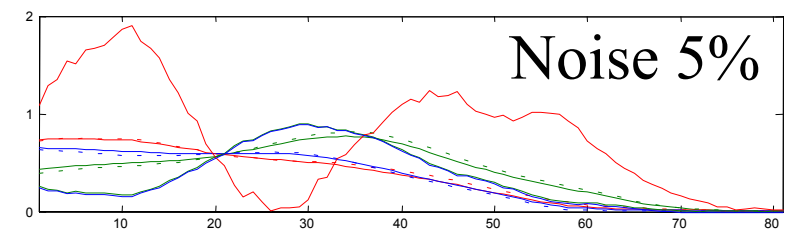




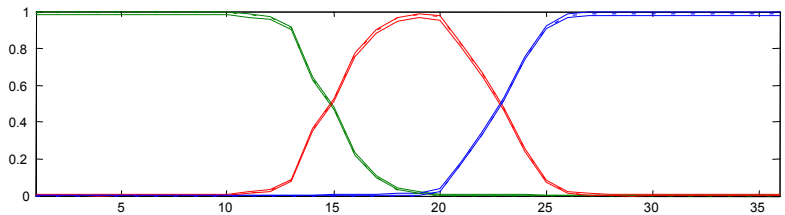


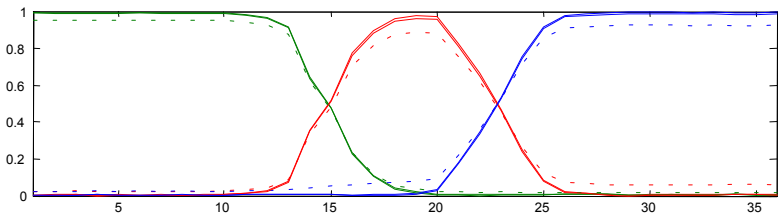


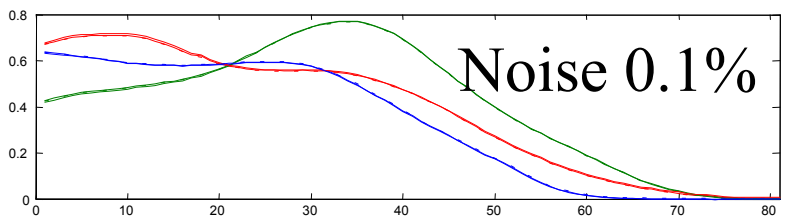


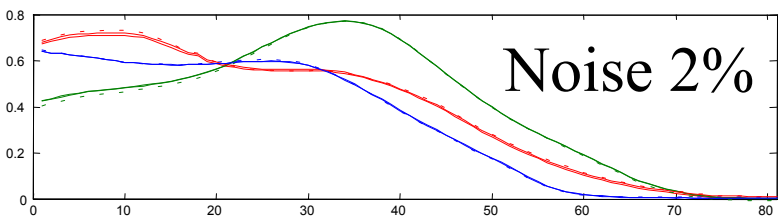


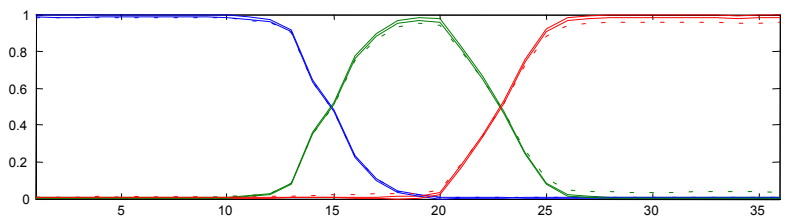
Calculation of band boundaries from mean profile error bands (under non-negativity, closure and selectivity constraints)

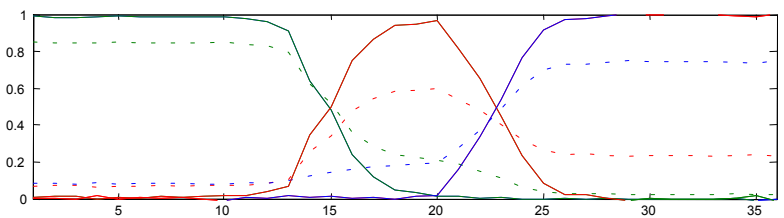


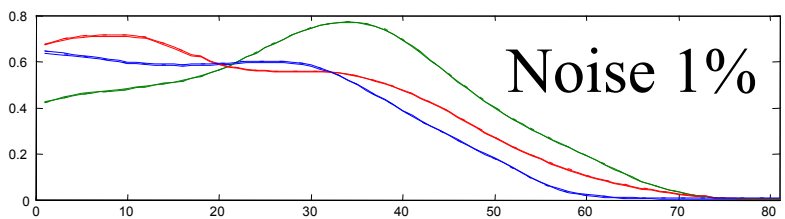


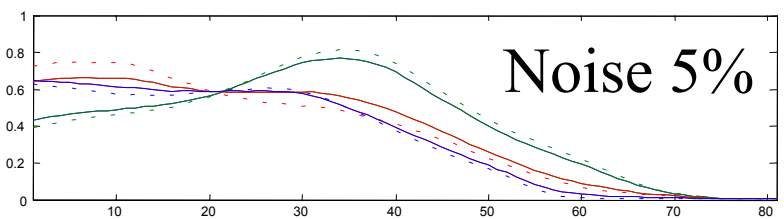


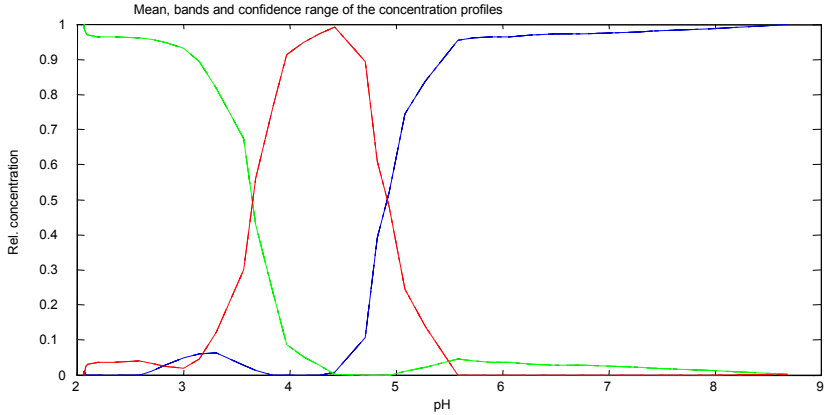




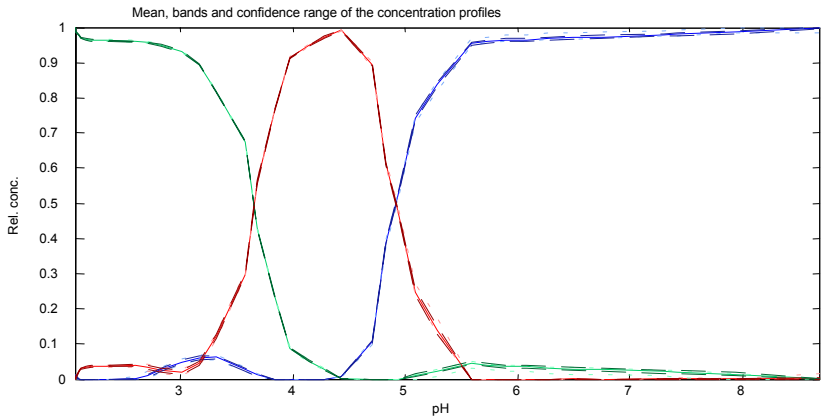


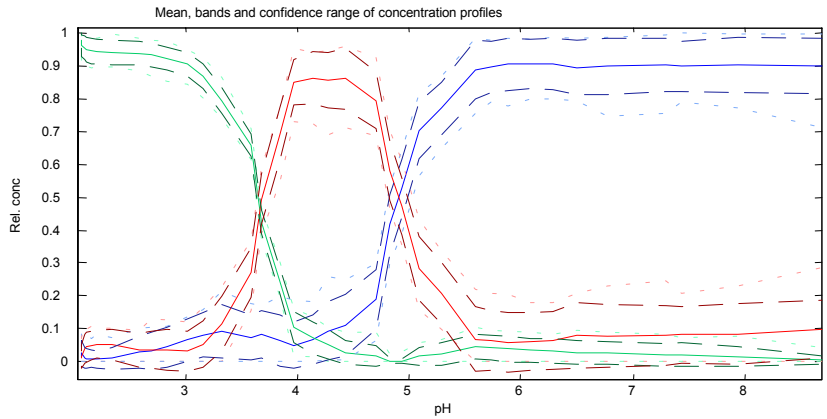


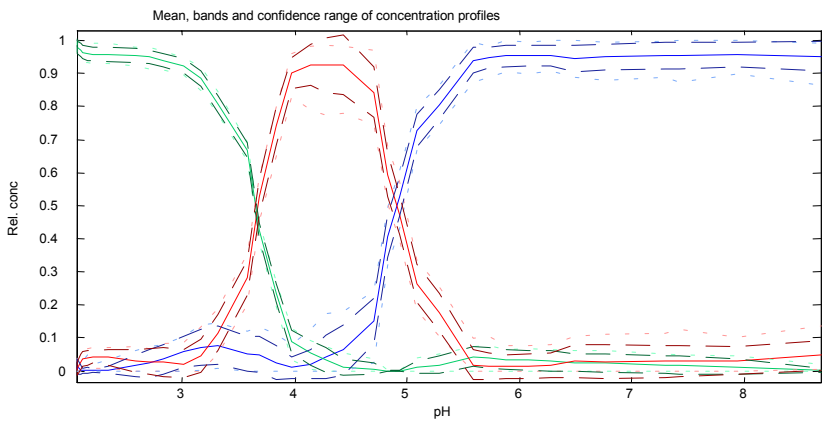


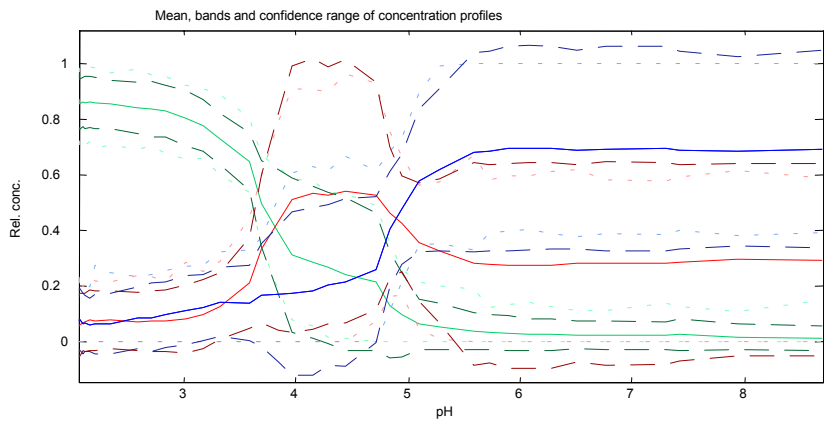


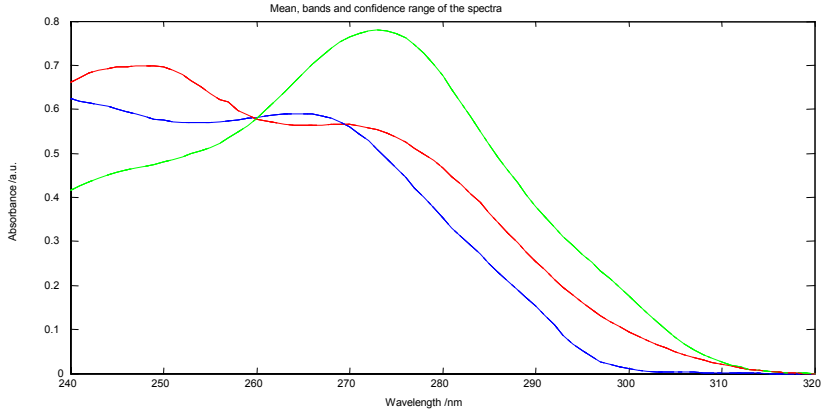
Noise Addition Simulations
Concentration profiles:
Mean max and min profiles
Confidence range profiles



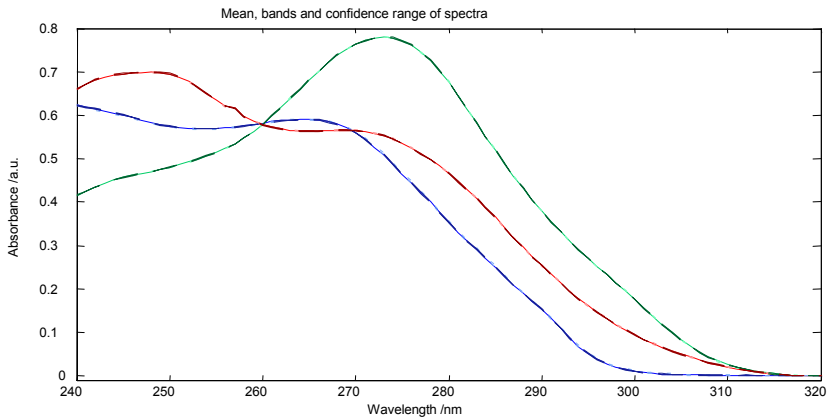


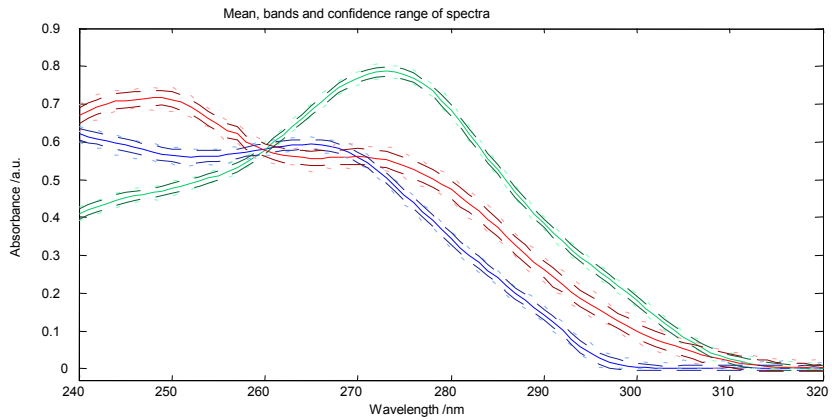


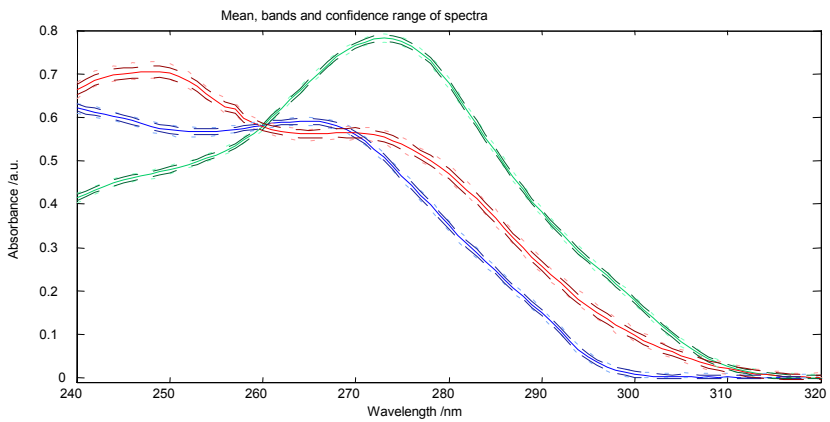


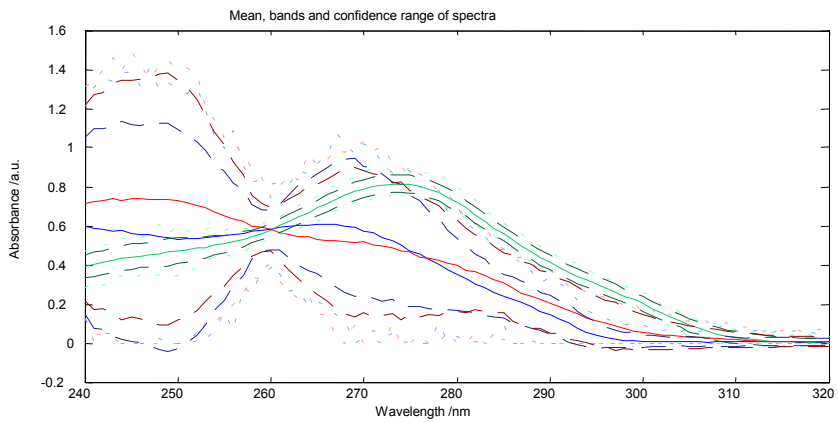


Noise Addition Simulations
Spectra profiles:
Mean, max and min profiles
Confidence range profiles





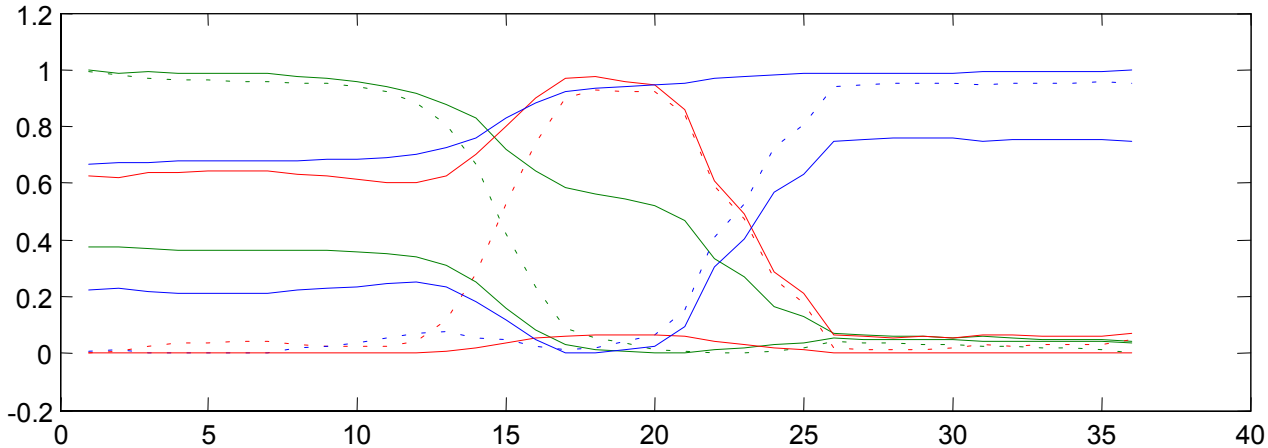


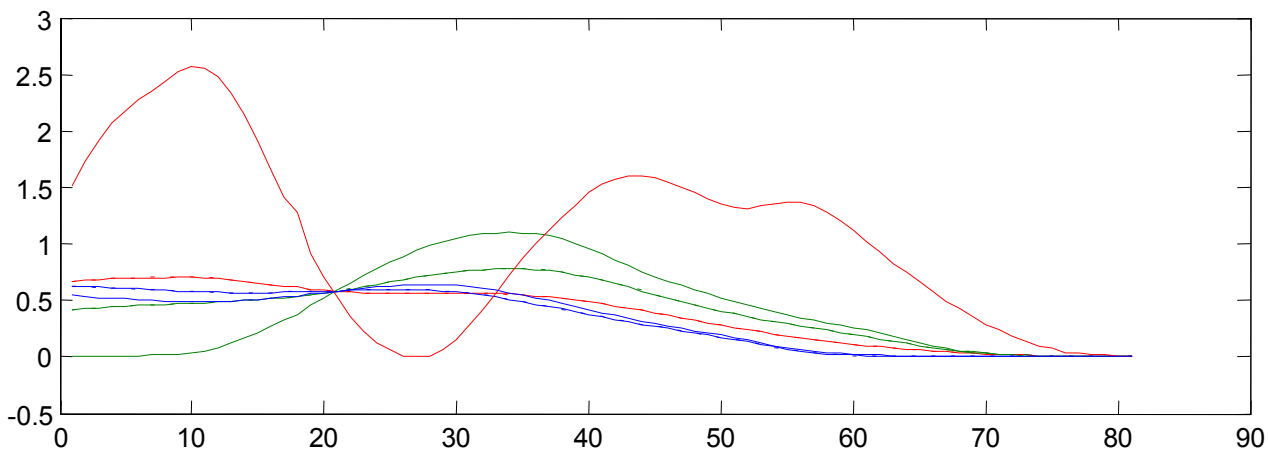


Noise Addition Simulations
pKa error estimations

	pK_1		pK_2	
Noise added	Value	Std. dev	Value	Std. dev
0 %	3.6539	2e-14	4.9238	2e-14
0.1 %	3.6540	6e-4	4.9226	0.0022
1 %	3.6592	0.0061	4.9134	0.0264
2 %	3.6656	0.0101	4.9100	0.0409
5 %	4.0754	0.4873	5.3308	1.1217

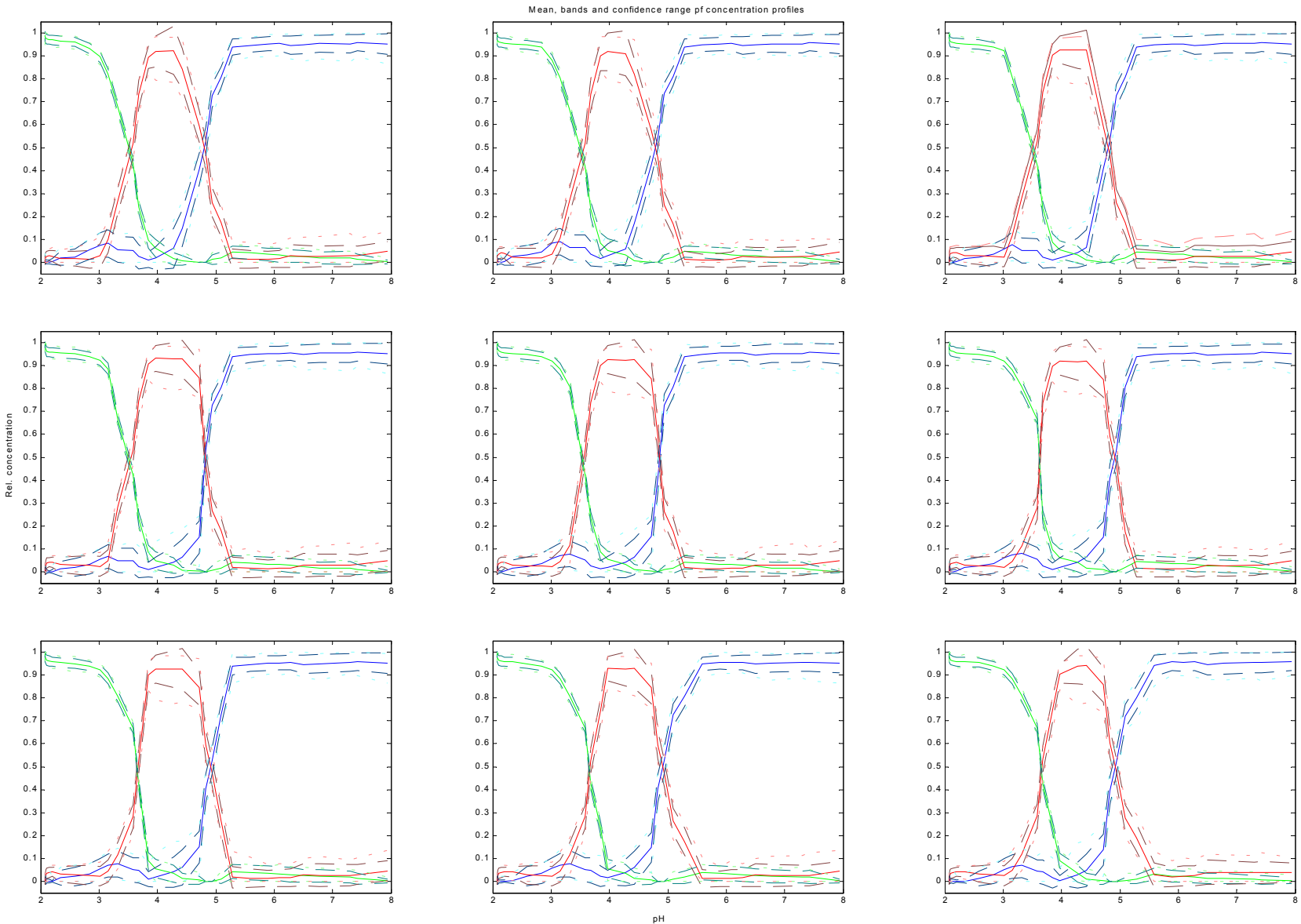
Calculation of band boundaries from mean profile error bands (under non-negativity and closure constraints) at 1% error noise addition



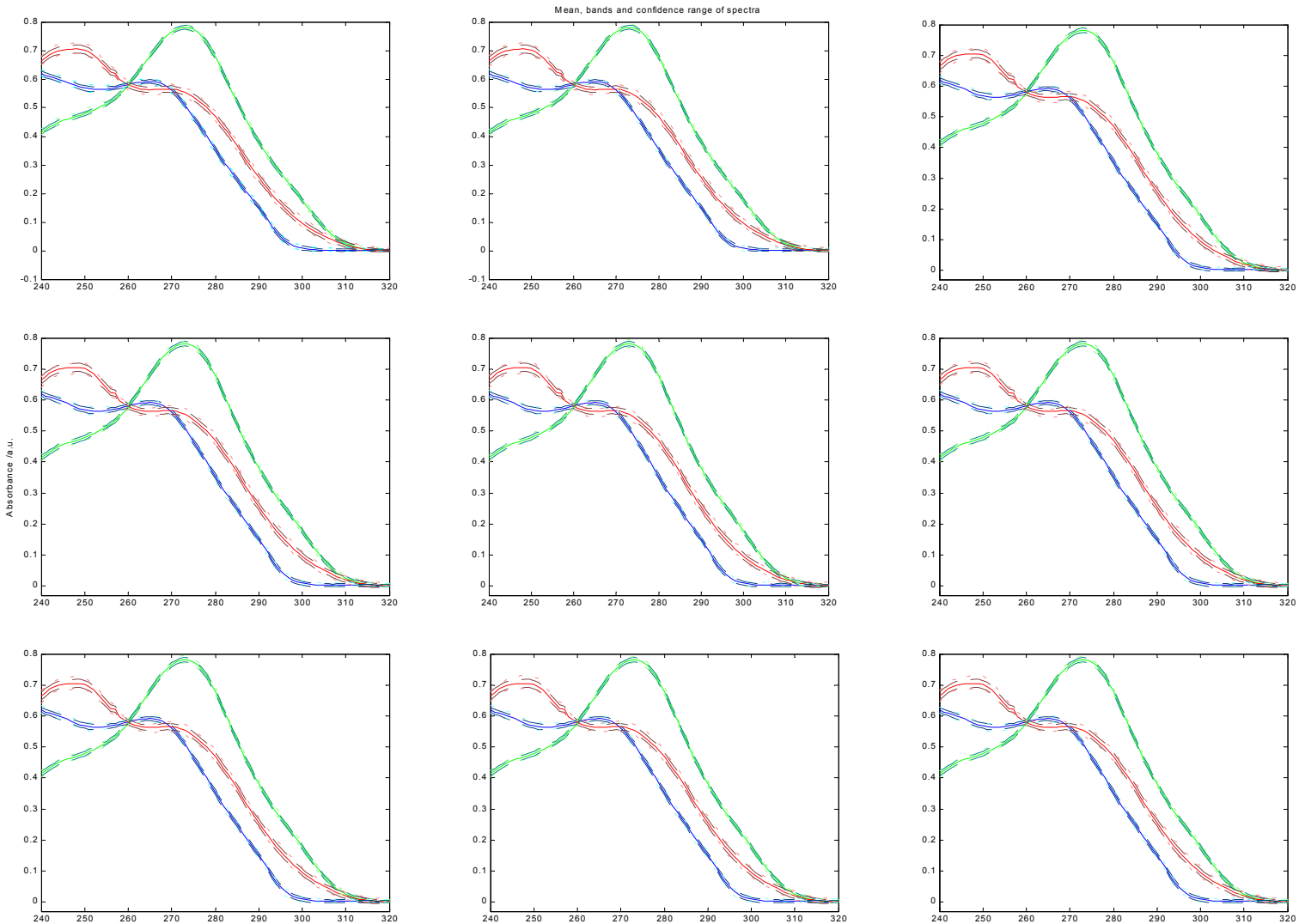


Jackknife Simulations at 1% noise; Concentration profiles:

Mean max and min profiles and confidence range profiles



Jackknife Simulations at 1% noise; spectra profiles:
Mean max and min profiles and confidence range profiles



Jackknife Simulations
pKa error estimations
at 1% noise level

Nº exp	pK_1	pK_2
1	3.6629 ± 0.0066	4.9135 ± 0.0277
2	3.6601 ± 0.0074	4.8989 ± 0.0221
3	3.6590 ± 0.0059	4.9122 ± 0.0261
4	3.6580 ± 0.0056	4.9221 ± 0.0189
5	3.6333 ± 0.0130	4.9018 ± 0.0236
6	3.6882 ± 0.0198	4.9144 ± 0.0267
7	3.6591 ± 0.0064	4.9144 ± 0.0256
8	3.6592 ± 0.0059	4.9144 ± 0.0253
9	3.6582 ± 0.0065	4.9233 ± 0.0239

Parameter Estimation
Summary of results

		Real		0.1 %		1 %		2 %		5 %	
		pk1	pk2	pk1	pk2	pk1	pk2	pk1	pk2	pk1	pk2
Theoretical Value	Value	3.6660	4.9244	-	-	-	-	-	-	-	-
MonteCarlo Simulations	Value	-	-	3.6662	4.9244	3.6696	4.9262	3.6761	4.9173	3.9762	5.0745
	Stand. dev.	-	-	0.0006	0.0012	0.0065	0.0128	0.0127	0.0245	0.4349	0.7595
Noise Addition	Value	-	-	3.6540	4.9226	3.6592	4.9134	3.6656	4.9100	4.0754	5.3308
	Stand. dev.	-	-	0.0006	0.0022	0.0061	0.0264	0.0101	0.0409	0.4873	1.1217
JackKnife	Value	-	-	3.6546	4.9199	3.6598	4.9128	3.6673	4.9131	4.0822	5.3292
	Stand. dev.	-	-	0.0038	0.0032	0.0086	0.0244	0.0124	0.0471	0.5145	1.0906

Outline:

- Introduction
- Rotational ambiguities and feasible bands
- Error propagation and resampling methods
- Experimental system and simulations
- Results
- Conclusions

Summary

•Different approaches for calculation of error propagation and prediction intervals of estimations have been compared including: Monte Carlo simulations, Noise addition resampling approaches and Jackknife based methods.

•The obtained results allowed a preliminary investigation of the noise effects on MCR-ALS resolved profiles and on parameters from them estimated, and allowed also a preliminary investigation of noise effects on rotational ambiguities.

•The study has been shown for the resolution of a three-component equilibrium system with overlapping concentration and spectra profiles

Conclusions

-Rotational ambiguity effects on species profiles depend on the structure and constraints of the data system.

-Rotational ambiguities effects at low noise levels in a system with low selectivity are more important than error propagation effects

-However, at high noise levels (≥ 5%), error propagation effects became larger than rotational ambiguities effects and they are both mixed and undistinguishable

- Obviously the best is to have a system with enough selectivity (low rotational ambiguities) and with low noise levels (low error propagation)

Poster presentations of the Chemometrics group from the University of Barcelona at CAC2002

ELUCIDATION OF THE STRUCTURE OF A **PROTEIN FOLDING INTERMEDIATE** (MOLTEN GLOBULE STATE) USING MULTIVARIATE CURVE RESOLUTION ALTERNATING LEAST SQUARES (MCR-ALS)
Susana Navea, Anna de Juan and Romà Tauler

MULTIVARIATE CURVE RESOLUTION ALTERNATING LEAST SQUARES ANALYSIS OF THE **CONFORMATIONAL EQUILIBRIA OF THE OLIGONUCLEOTIDE d<TGCTCGCT>**
Joaquim Jaumot, Núria Escaja, Raimundo Gargallo, Enrique Pedroso and Romà Tauler

HARD AND SOFT MODELLING OF ACID-BASE CHEMICAL EQUILIBRIA OF **BIOMOLECULES** USING **^{1}H-NMR**

Joaquim Jaumot, Montserrat Vives, Raimundo Gargallo and Romà Tauler

IDENTIFICATION AND DISTRIBUTION OF **MICROCONTA-MINANTS SOURCES OF NONIONIC SURFACTANTS**, THEIR DEGRADATION PRODUCTS AND LINEAR ALKYLBENZENE SULFONATES IN COASTAL WATERS AND SEDIMENTS IN SPAIN BY MEANS OF CHEMOMETRIC METHODS

Emma Peré-Trepat, Mira Petrovic, Damià Barceló and Romà Tauler

MULTIWAY DATA ANALYSIS OF **ENVIRONMENTAL CONTAMINATION SOURCES IN SURFACE NATURAL WATERS** OF CATALONIA (SPAIN)

Emma Peré-Trepat, Mónica Flo, Montserrat Muñoz, Manel Vilanova, Josep Caixach, Antoni Ginebreda, Romà Tauler