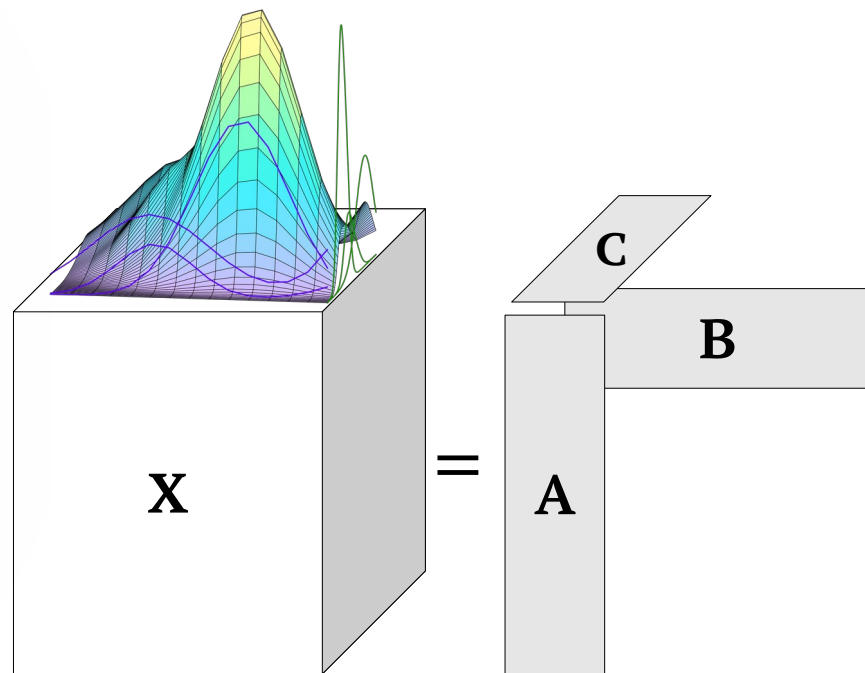
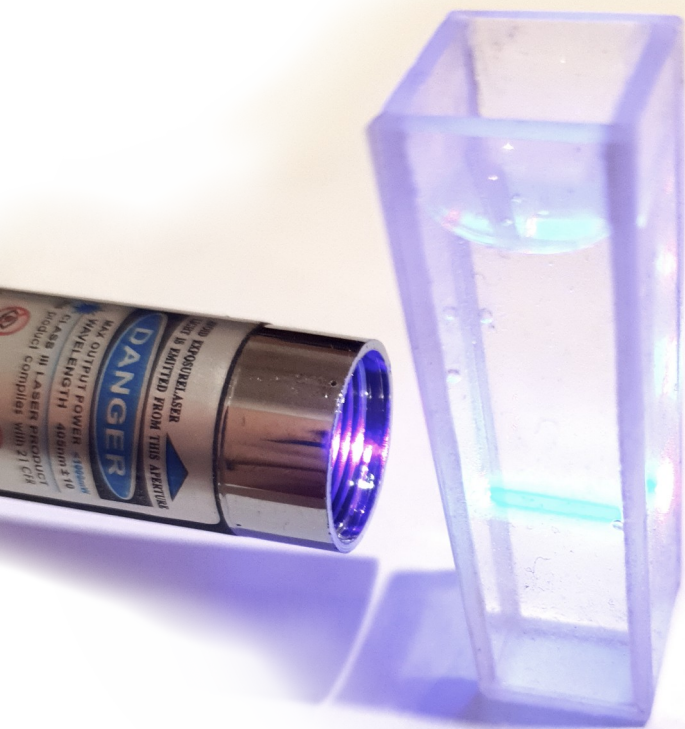
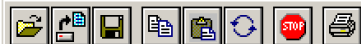


Tensor decompositions of fluorescence spectra:

A case study in R





R Console

```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> help()
starting httpd help server ... done
> q()
```

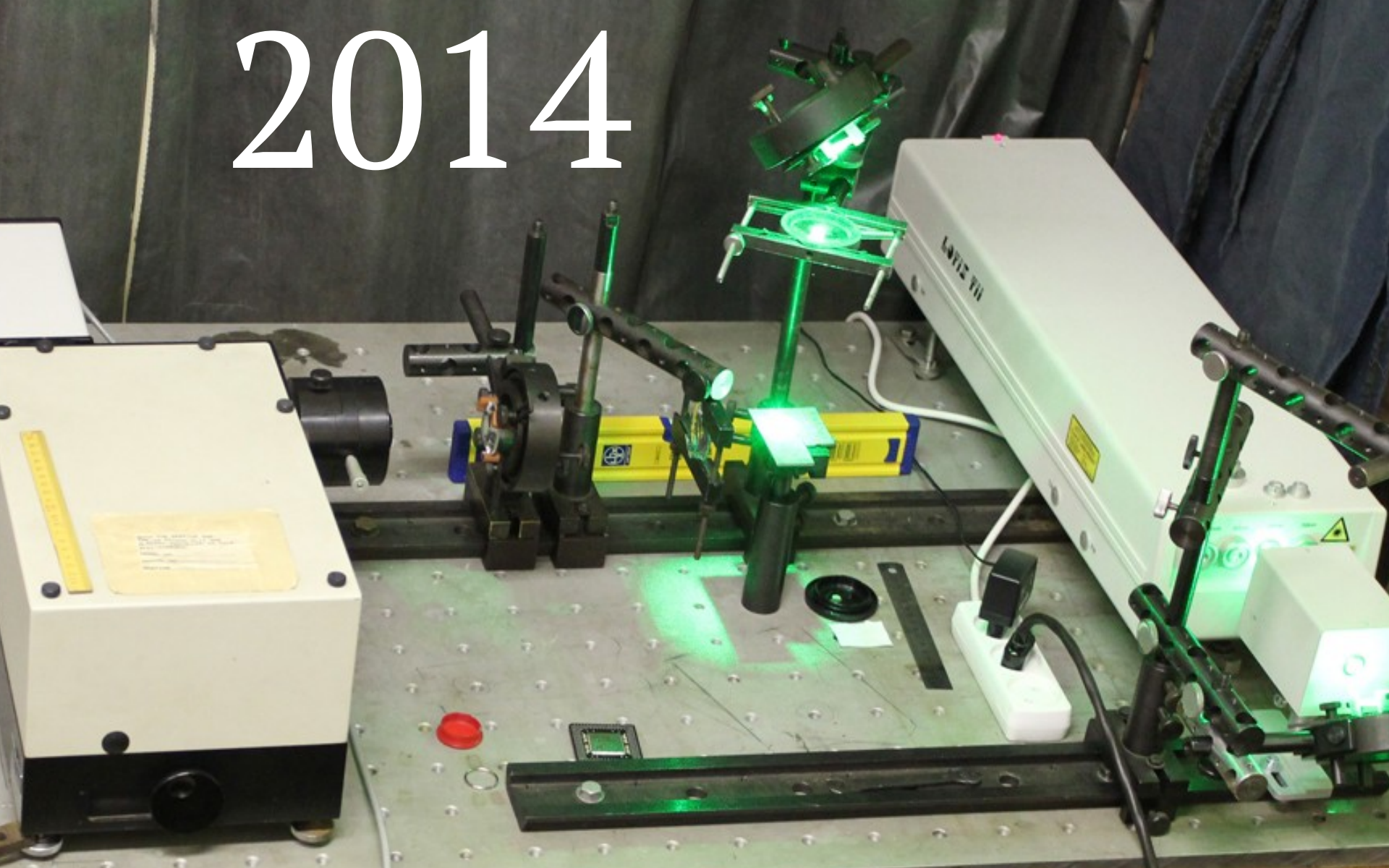
Question

Save workspace image?

Да Нет Отмена

2010

2014





2016

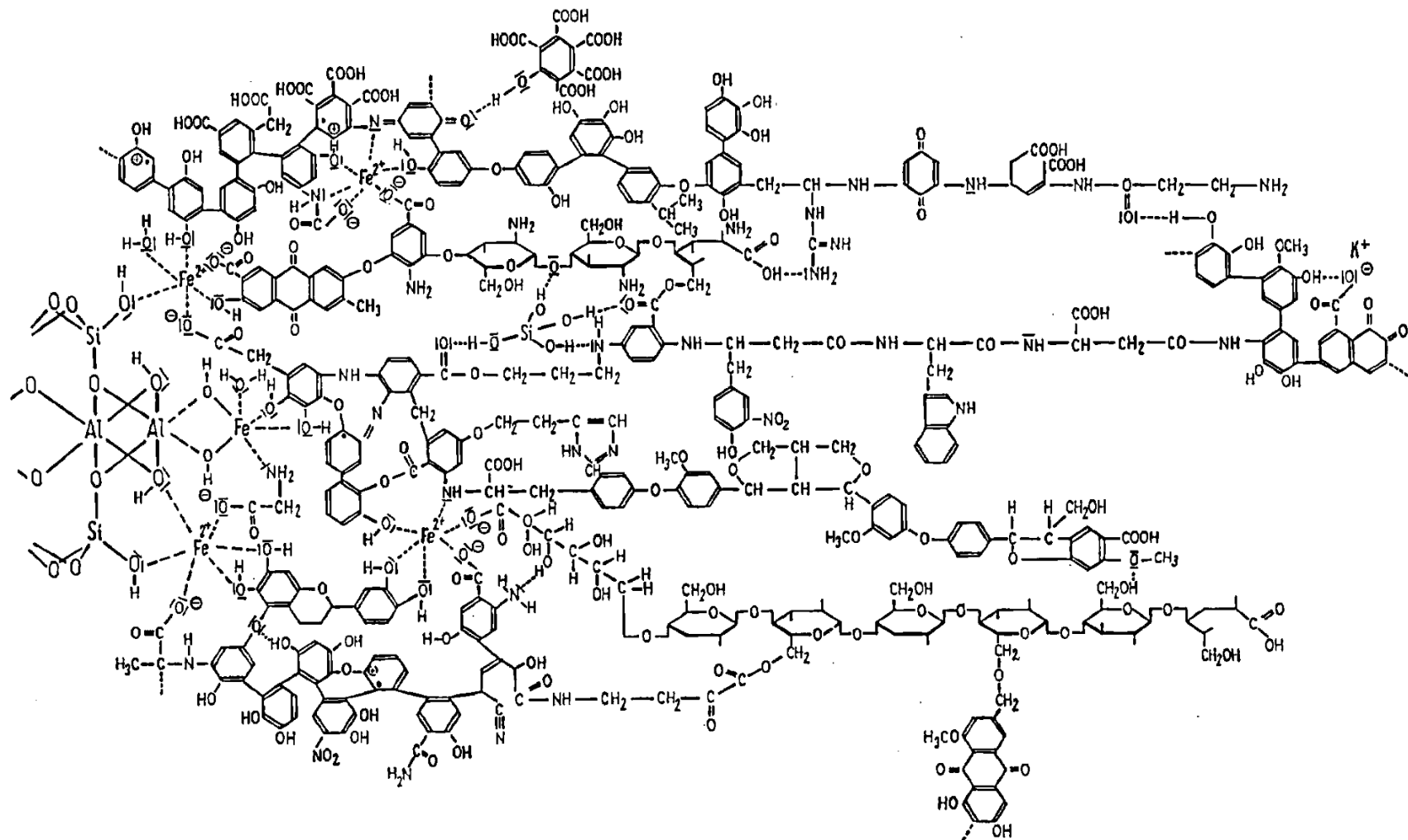
```
==1730== Invalid read of size 1
==1730==    at 0x10915D: main (in /home/ivan/foo)
==1730==    Address 0x4a6206a is 0 bytes after a block of size 42 alloc'd
==1730==    at 0x48417B4: malloc (vg_replace_malloc.c:381)
==1730==    by 0x10914A: main (in /home/ivan/foo)
==1730==
==1730==
==1730== More than 10000000 total errors detected. I'm not reporting any more.
==1730== Final error counts will be inaccurate. Go fix your program!
==1730== Rerun with --error-limit=no to disable this cutoff. Note
==1730== that errors may occur in your program without prior warning from
==1730== Valgrind, because errors are no longer being displayed.
==1730==
==1730==
==1730== HEAP SUMMARY:
==1730==    in use at exit: 42 bytes in 1 blocks
==1730==    total heap usage: 1 allocs, 0 frees, 42 bytes allocated
==1730==
==1730== LEAK SUMMARY:
==1730==    definitely lost: 42 bytes in 1 blocks
==1730==    indirectly lost: 0 bytes in 0 blocks
==1730==    possibly lost: 0 bytes in 0 blocks
==1730==    still reachable: 0 bytes in 0 blocks
==1730==         suppressed: 0 bytes in 0 blocks
==1730== Rerun with --leak-check=full to see details of leaked memory
==1730==
==1730== For lists of detected and suppressed errors, rerun with: -s
```

2017



Dissolved organic matter

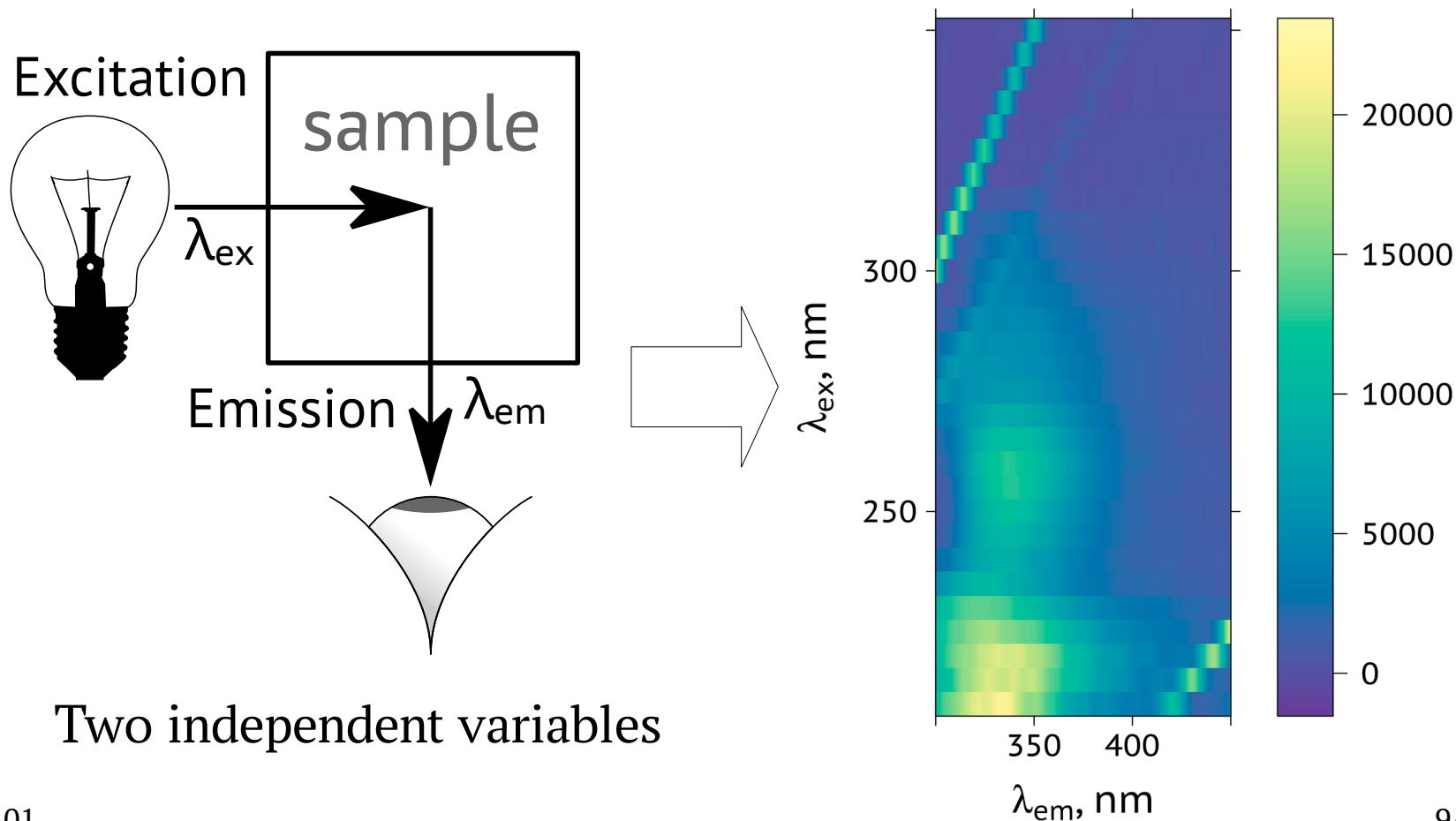
- Sources
 - Terrestrial runoff
 - Local production in rivers and oceans
 - Melting permafrost
- Sinks
 - Photochemical bleaching
 - Local biodegradation
- Part of the global carbon cycle
 - flux ≈ 50 Gt C/year
- Goal: monitoring



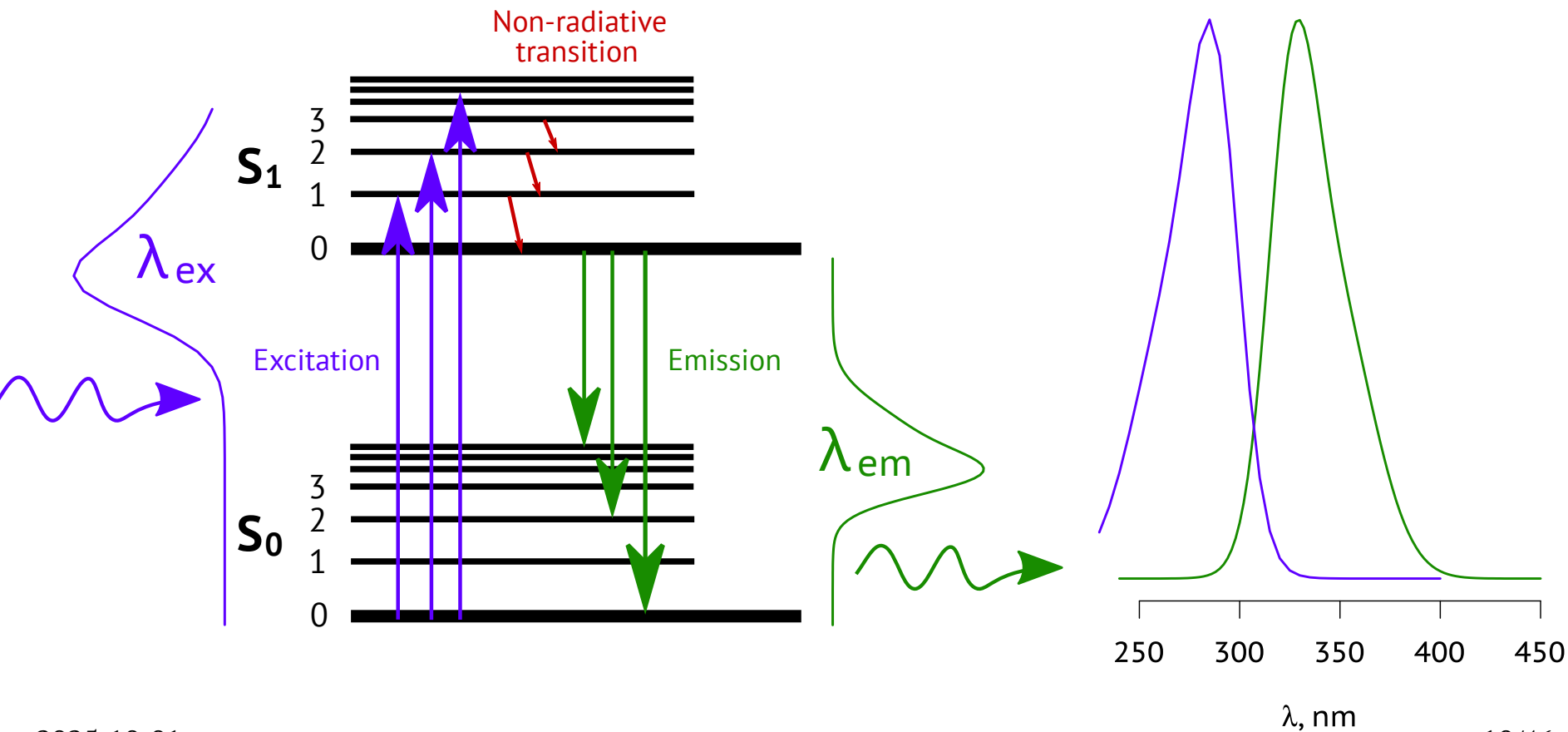
(Kleinhempel, 1970)

Abb. 1

Fluorescence spectroscopy

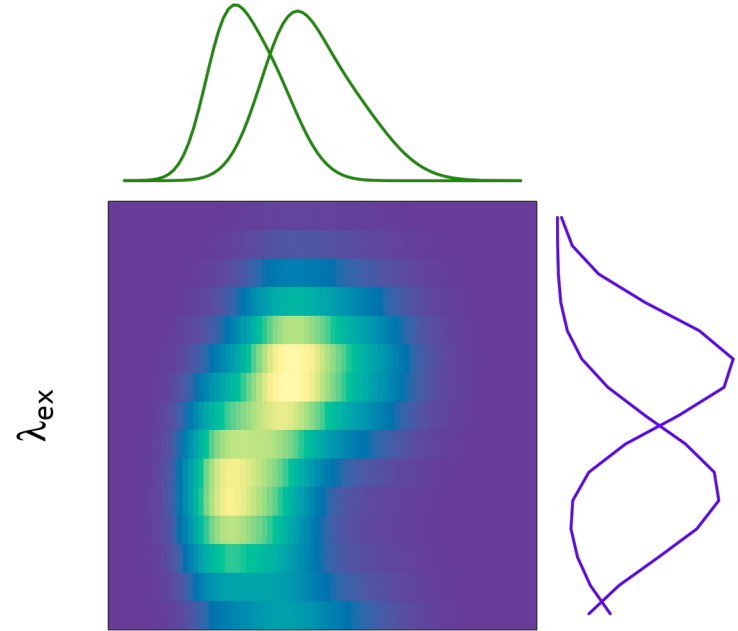


Excitation and emission spectra



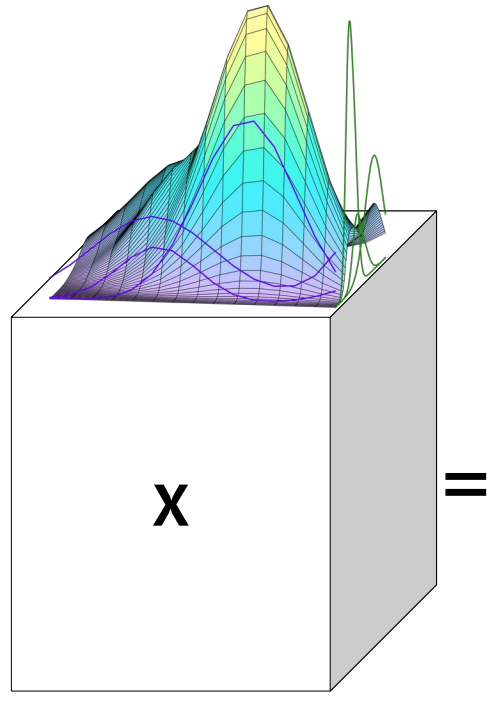
Fluorescence as an outer product

- Emission spectrum independent from excitation spectrum

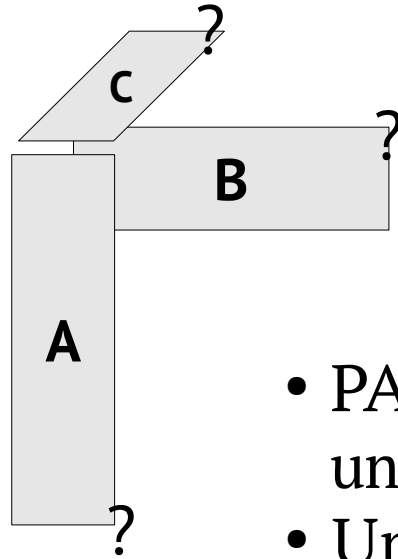


$$F_k(\lambda_{\text{em}}, \lambda_{\text{ex}}) = \sum_r C_{k,r} f_r^{\text{em}}(\lambda_{\text{em}}) f_r^{\text{ex}}(\lambda_{\text{ex}})$$

Second order advantage



$$X_{i,j,k} = \sum_r A_{i,r} B_{j,r} C_{k,r}$$



- PARAFAC decomposition is unique*
- Unknown interferences handled

Fitting the model

- Alternating least squares

$$\min_{\mathbf{A}} \sum_{i,j,k} W_{i,j,k} \left(X_{i,j,k} - \sum_r A_{i,r} B_{j,r} C_{k,r} \right)^2 \quad \mathbf{A} \leftarrow \mathbf{X}_A (\mathbf{C} * \mathbf{B})^+$$

$$\min_{\mathbf{B}} \sum_{i,j,k} W_{i,j,k} \left(X_{i,j,k} - \sum_r A_{i,r} B_{j,r} C_{k,r} \right)^2 \quad \mathbf{B} \leftarrow \mathbf{X}_B (\mathbf{A} * \mathbf{C})^+$$

$$\min_{\mathbf{C}} \sum_{i,j,k} W_{i,j,k} \left(X_{i,j,k} - \sum_r A_{i,r} B_{j,r} C_{k,r} \right)^2 \quad \mathbf{C} \leftarrow \mathbf{X}_C (\mathbf{B} * \mathbf{A})^+$$

- Nonlinear optimisation

Deviations from the PARAFAC model



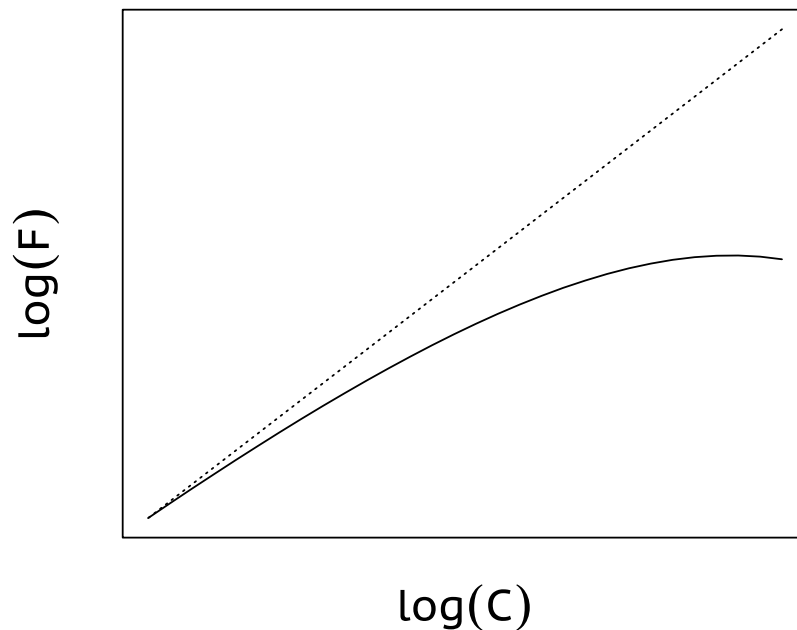
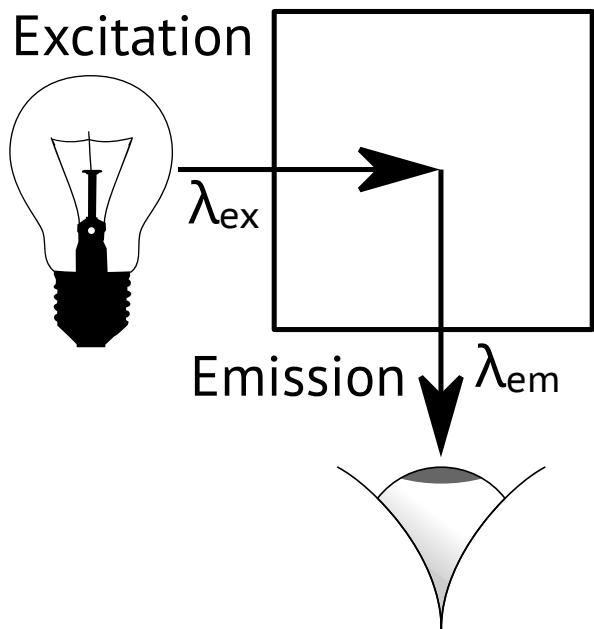
Inner filter effect
Multiplying the whole EEM due to absorbance



Scattering signal
Adding non-trilinear components

https://en.wikipedia.org/wiki/File:Why_is_the_sky_blue.jpg

Inner filter effect



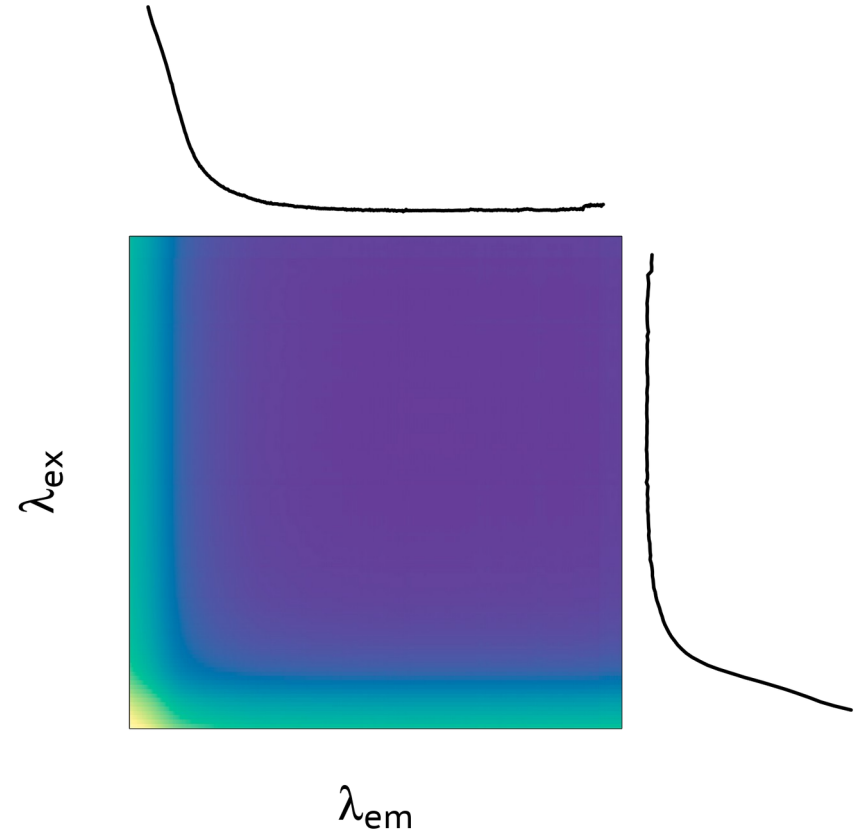
Both excitation and emission beams get partially reabsorbed by the solution

Absorbance-based correction

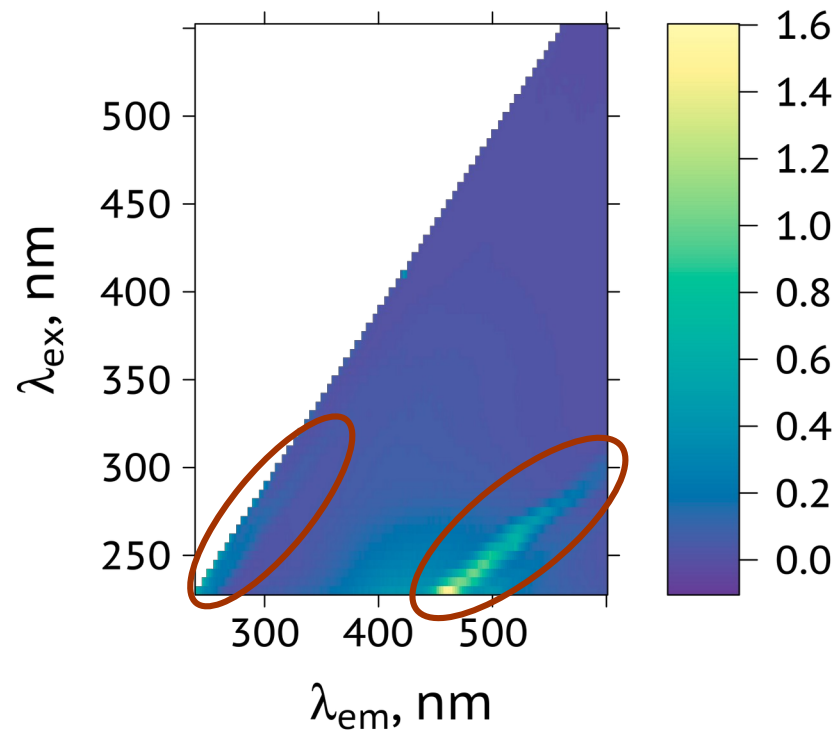
- We observe partially absorbed signal

$$F_0(\lambda_{\text{em}}, \lambda_{\text{ex}}) \cdot 10^{-\frac{a(\lambda_{\text{em}}) + a(\lambda_{\text{ex}})}{2}}$$

- Multiply it back by corresponding absorbance factor



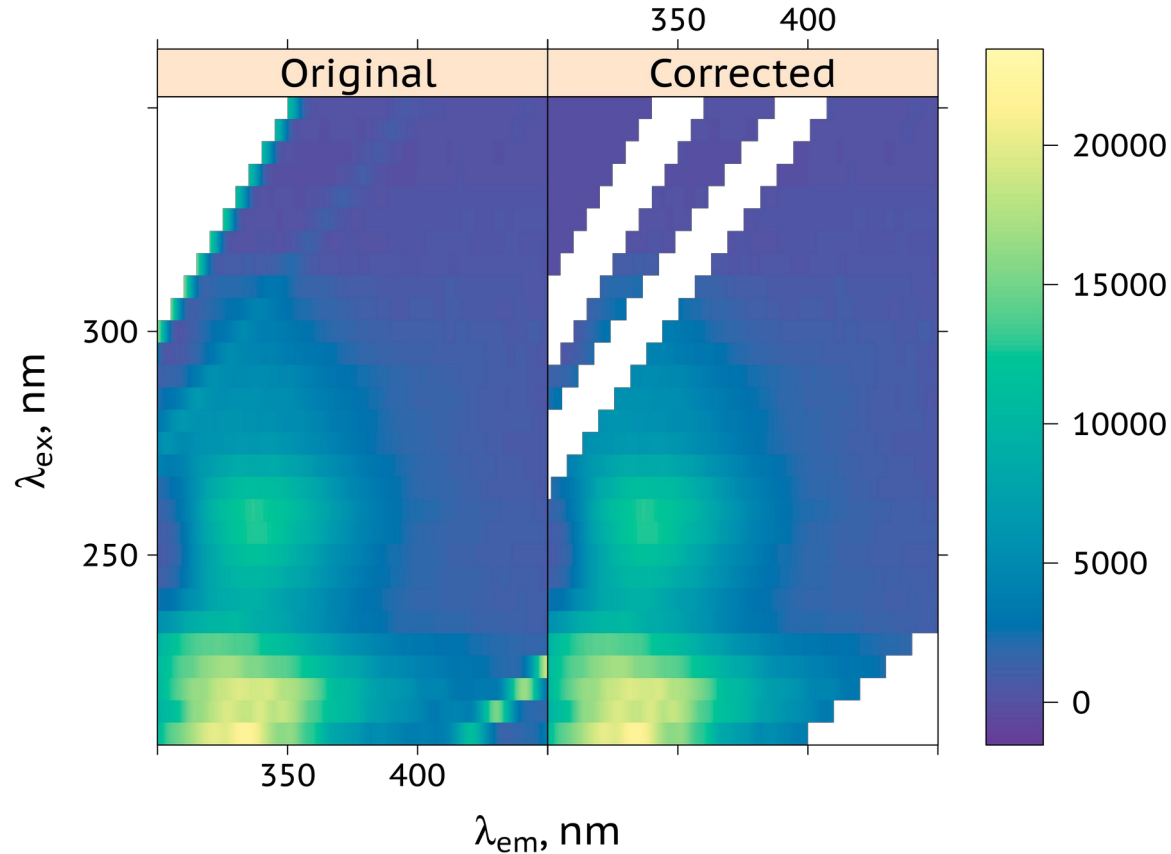
Scattering signal



Scattering signal is not trilinear and must be handled separately

Removal of scattering signal

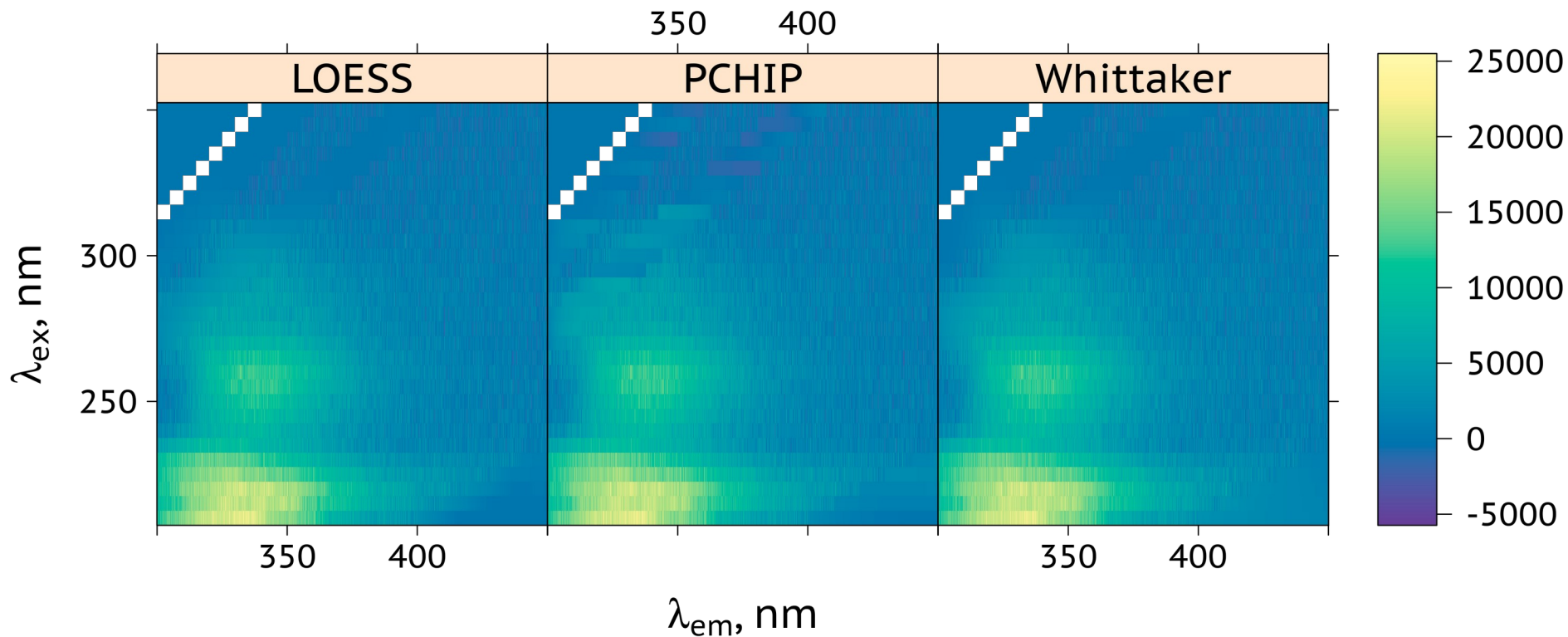
- Zero residual weights for missing data
- May cause convergence problems, nonsense solutions



Interpolation of scattering signal

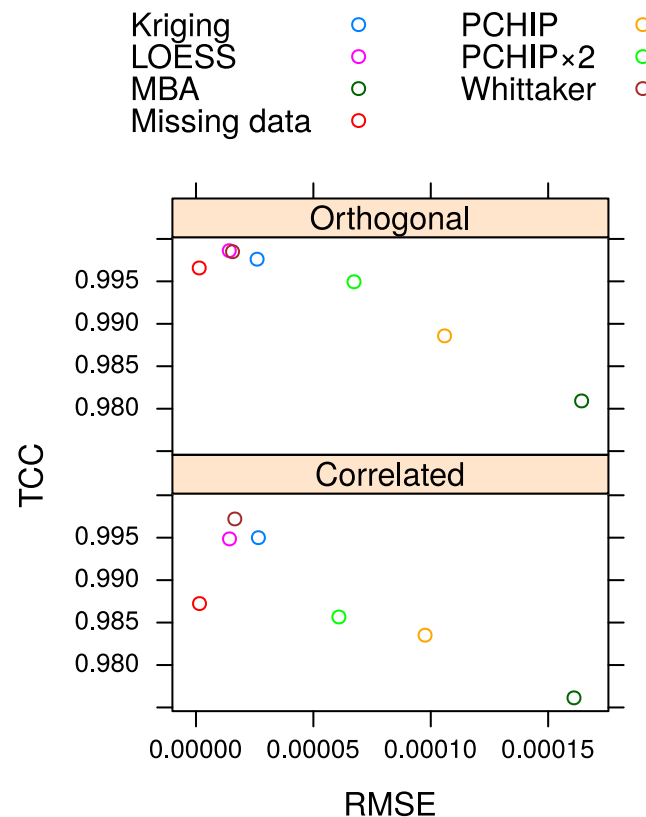
- Wide choice of methods
 - linear, PCHIP, LOESS, Kriging, B/P-splines, Whittaker...
- Prevents local minima, swamps
 - Fewer degrees of freedom in the model
- Choice of parameters not obvious
 - Typically empirical

Interpolation of scattering signal

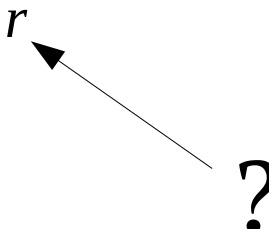


Whittaker interpolation

- Empirically, need 1:1000 first order:second order penalty for ideal interpolated surface



How many components?

$$X_{i,j,k} = \sum_r A_{i,r} B_{j,r} C_{k,r}$$


“Most tensor problems are NP-hard” (Hillar et al., 2013)

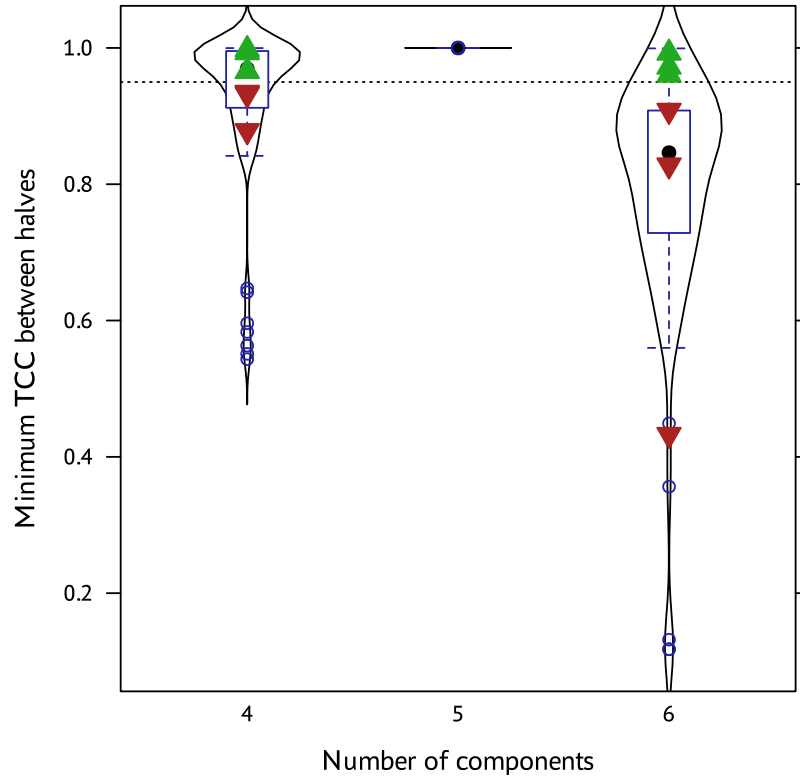
How to cross-validate an unsupervised learning model?

Split-half validation

- Divide dataset into non-intersecting halves
 - Group repeats together (Harshman & De Sarbo, 1984)
- Fit both models
- Reorder components to match
- Measure similarity
 - Tucker's congruence coefficient

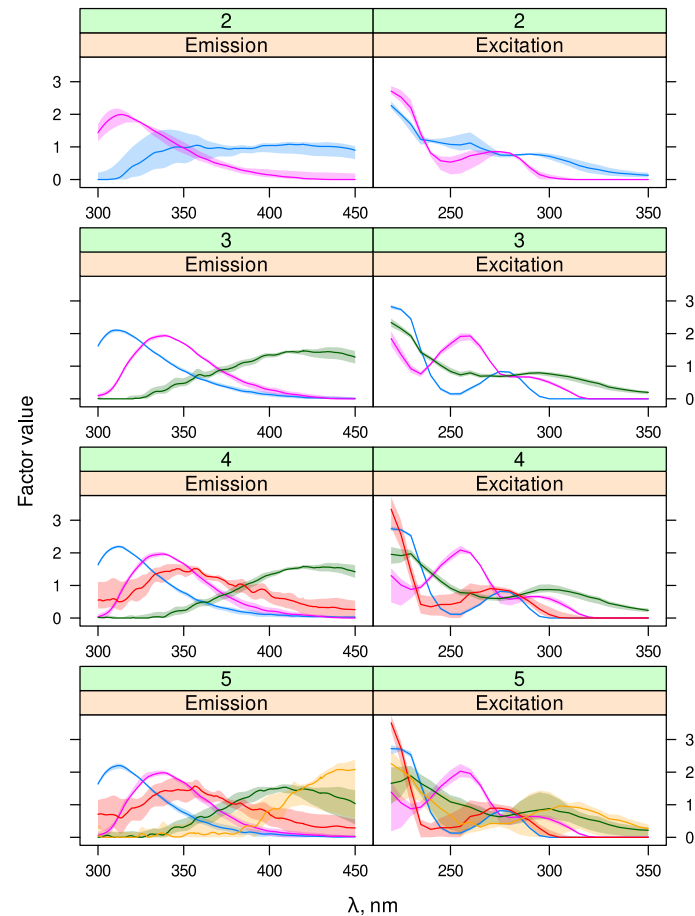
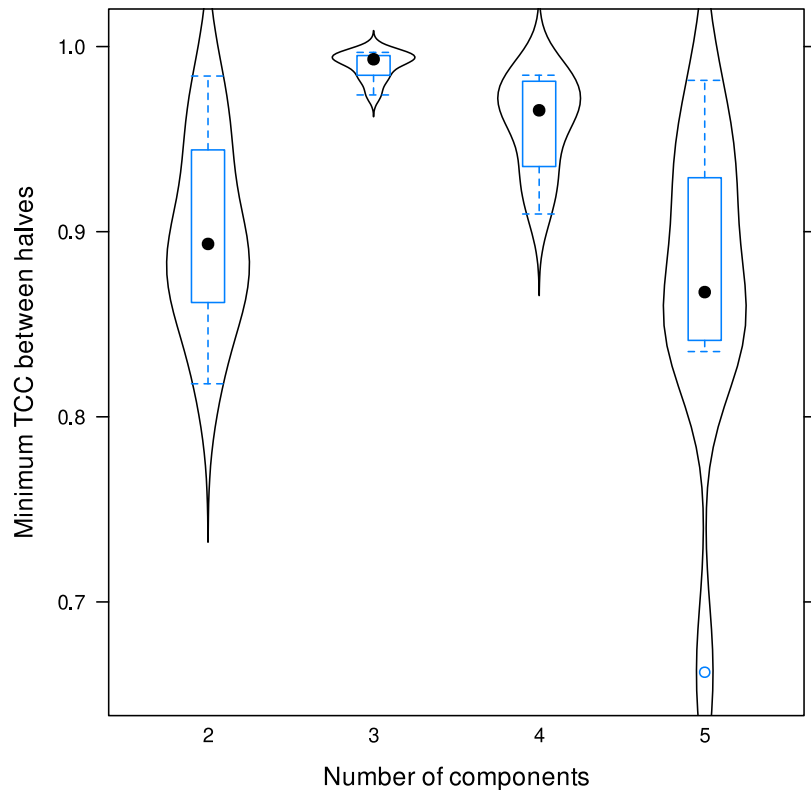
$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

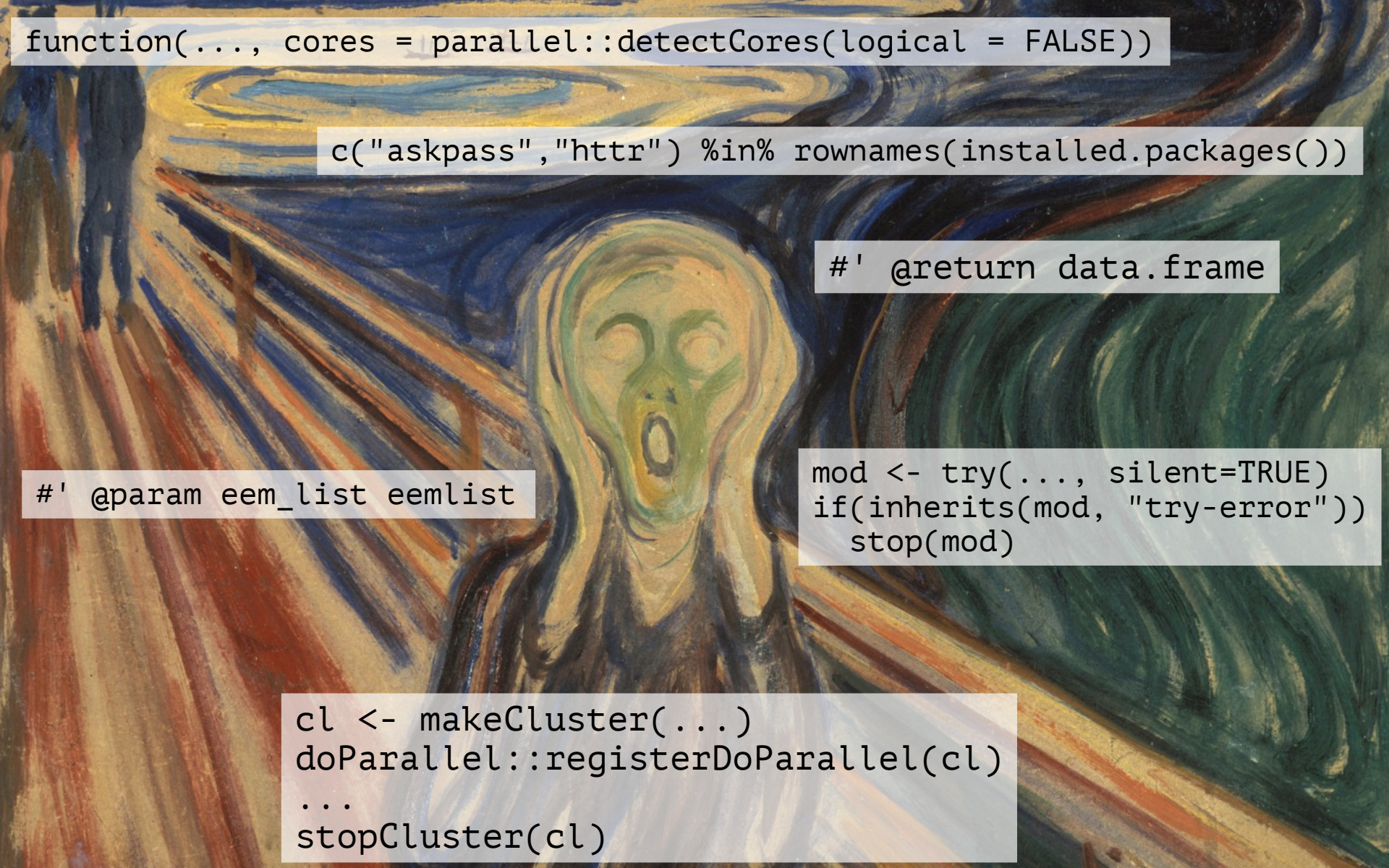
How much to split?



- “S₄C₆T₃” methodology
 - (Murphy, 2014)
- Shuffle a ground truth 5-component dataset
- By chance some validation results are wrong
- More resampling \Rightarrow better?

Validated model



The background of the image is a reproduction of the painting 'The Scream' by Edvard Munch. It depicts a figure in the center with a pale, greenish-yellow face and an open mouth in a scream, set against a turbulent, swirling background of blue, yellow, and red. The figure is framed by dark, swirling lines that suggest a storm or intense emotional distress.

```
function(..., cores = parallel::detectCores(logical = FALSE))
```

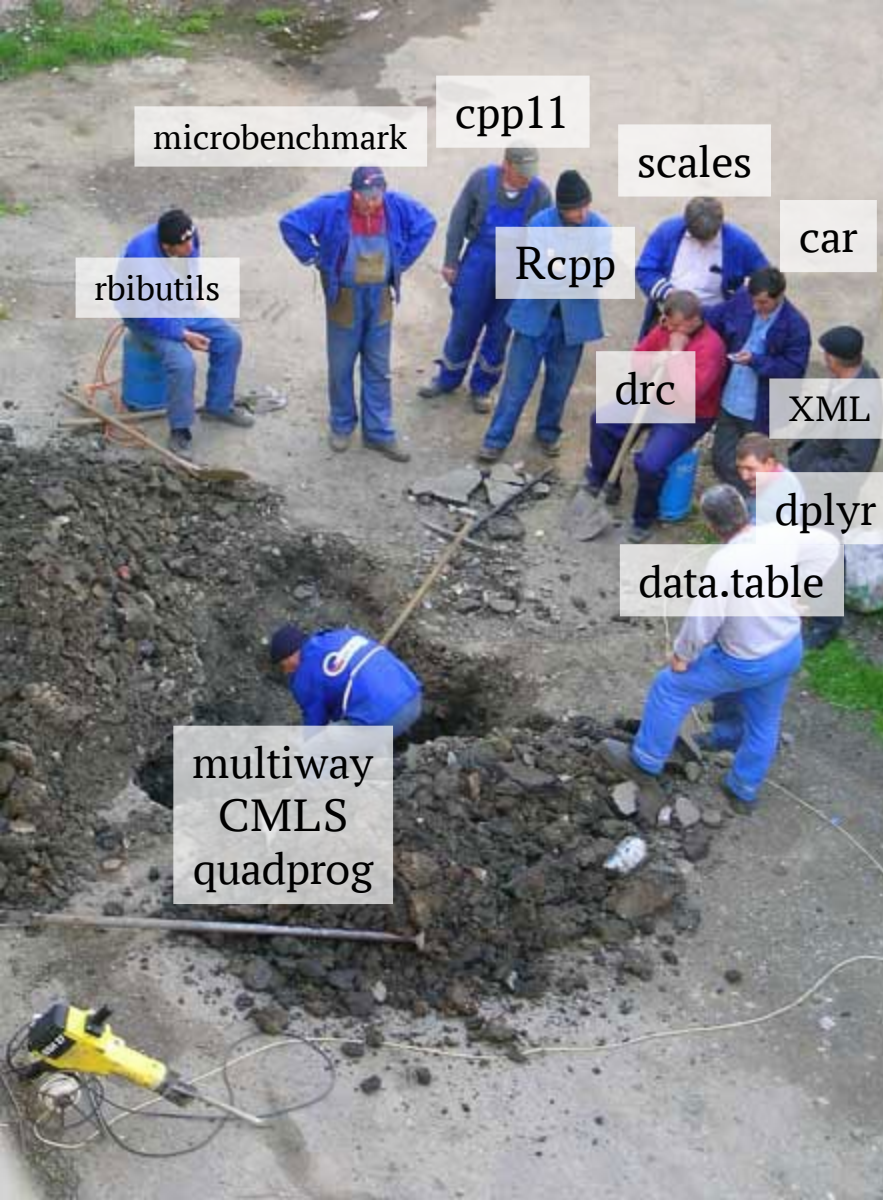
```
c("askpass","httr") %in% rownames(installed.packages())
```

```
#' @return data.frame
```

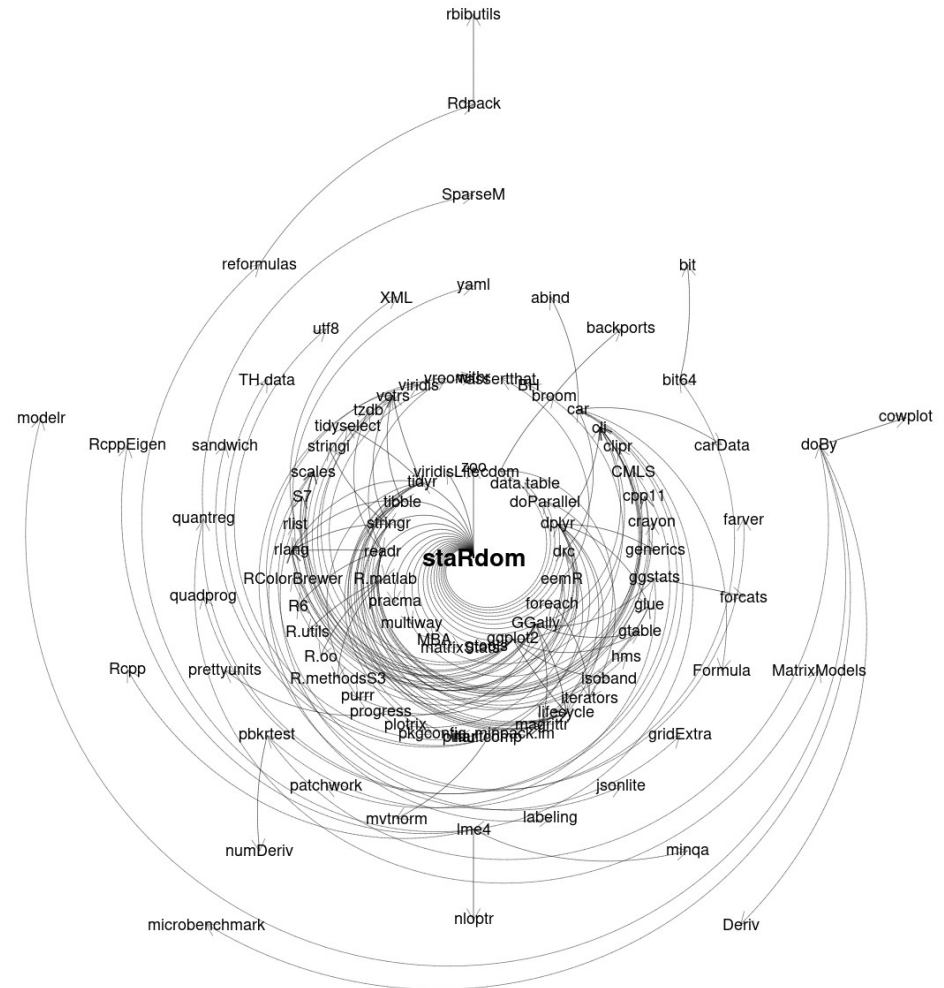
```
#' @param eem_list eemlist
```

```
mod <- try(..., silent=TRUE)  
if(inherits(mod, "try-error"))  
  stop(mod)
```

```
cl <- makeCluster(...)  
doParallel::registerDoParallel(cl)  
...  
stopCluster(cl)
```

```
lengths(package_dependencies("staRdom", ...)) == 100
```





Generating figures in Rd

```
\Sexpr[results=rd,stage=build]{  
  if (!dir.exists('man/figures'))  
    dir.create('man/figures')  
  {  
    # draw the figure  
  }  
  ,  
}}}
```

Current directory actually not guaranteed,
will break in
R CMD Rd2pdf <package directory>

Workaround for long-fixed bugs in all variants of results=
(e.g. r80718)

Hand-written HTML equations

```
\newcommand{\eqn3}{  
  \ifelse{html}{\out{<i>#3</i>}}  
    {\eqn{#1}{#2}}  
}
```

R<3.6.0: sometimes show up empty

```
\eqn3{  
  A = \log_{10}\frac{I_0}{I}  
}{A = log10(I0 / I)}{  
  A = log<sub>10</sub>(  
    <sup>I<sub>0</sub></sup>/<sub>I</sub>  
  )}
```

Fixing hand-written HTML equations

- `\newcommand{\forcebuild}{
 \Sexpr[results=hide,stage=build]{}
}`

Use pre-cached Rd database, don't parse anew

- `\newcommand{\eqn3}{
 \forcebuild
 \ifelse{
 \Sexpr[results=rd,stage=render]{...}
 }{...}{\eqn{#1}{#2}}
}`

Don't force ugly HTML on R \geq 4.2

Progress bar for parallel operations?

- `doSNOW`: make your own `parLapply()` using `sendCall()` & friends
- `pbapply`: split workload into chunks, wait for each chunk, update progress bar
- `futureverse`: sneak custom messages into communication channel using `dynGet()`

Progress bar for parallel operations

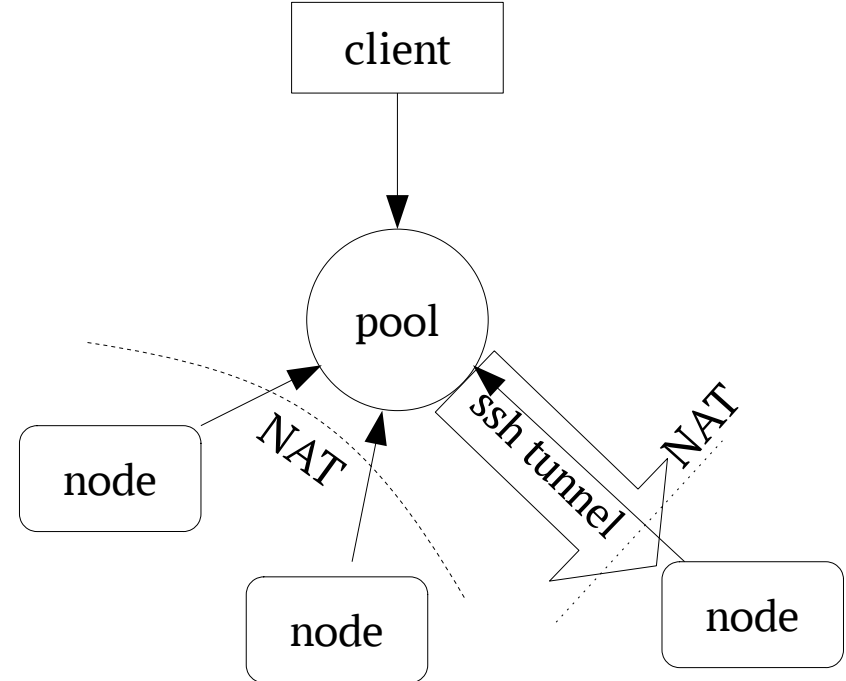
- Wrap a `parallel` cluster object
- Register methods for `recvData`, `recvOneData`
- Estimate the number of calls to happen
- Increment progress bar when called

“Cooperative computing”

- Challenging network conditions
 - NAT between rooms
 - TCP connections disappear without keep-alive
- Log in: nested remote desktop
- Availability: sometimes

Poor man's cluster...

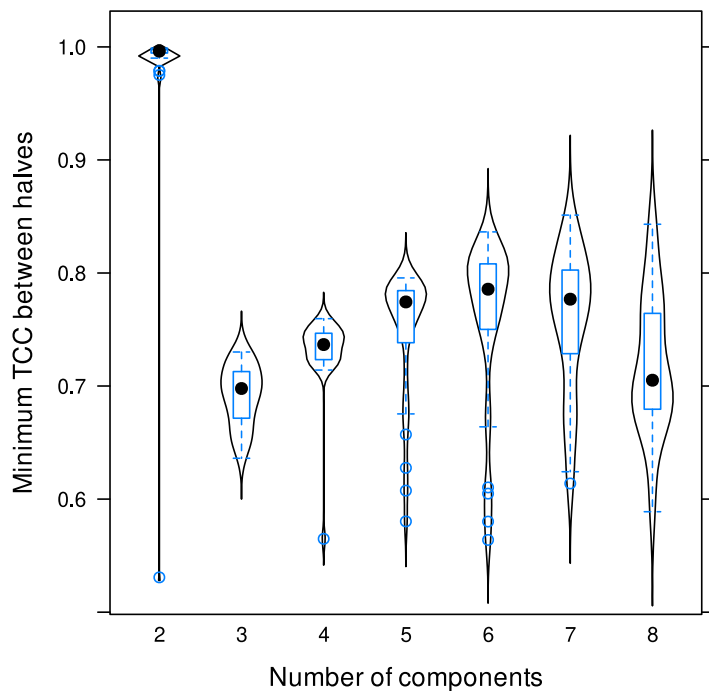
- Only the pool server needs to be reachable
- Nodes can join and leave
- Pool can restart
- Implements the `sendData()` / `recvData()` / `recvOneData()` interface
 - `clusterEvalQ()` problematic



...in base R

- The protocol is `serialize()`
 - Clients & nodes are trusted
 - Doesn't work with non-blocking socket connections
 - Must be half-duplex
- Interrupting the *client* still loses a day's work
 - Caching? cf. `depcache`
 - How to compute the key?
 - Needs a reasonable eviction policy

Rare samples with unique components



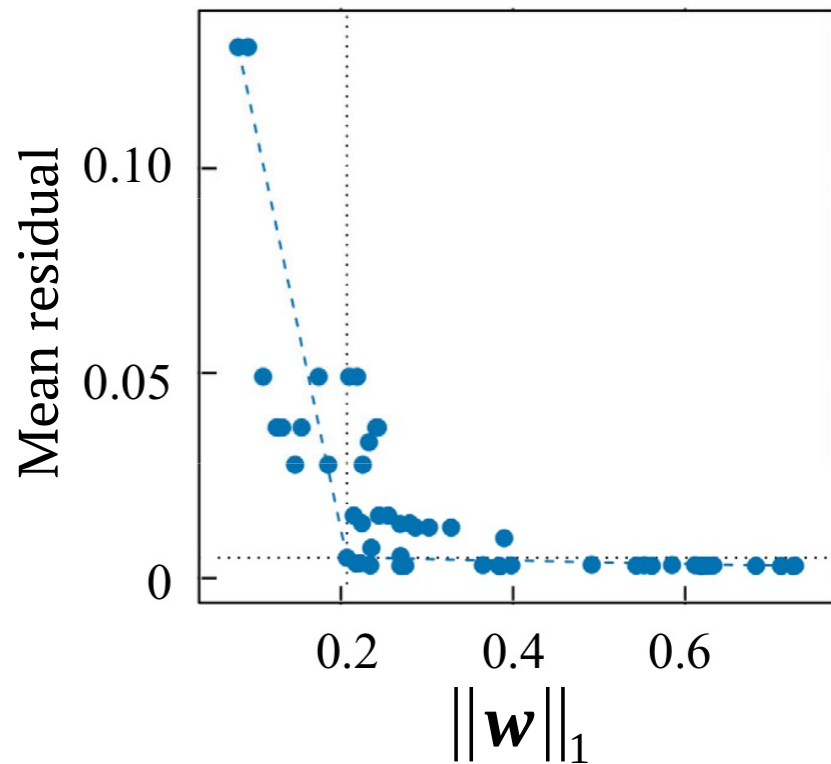
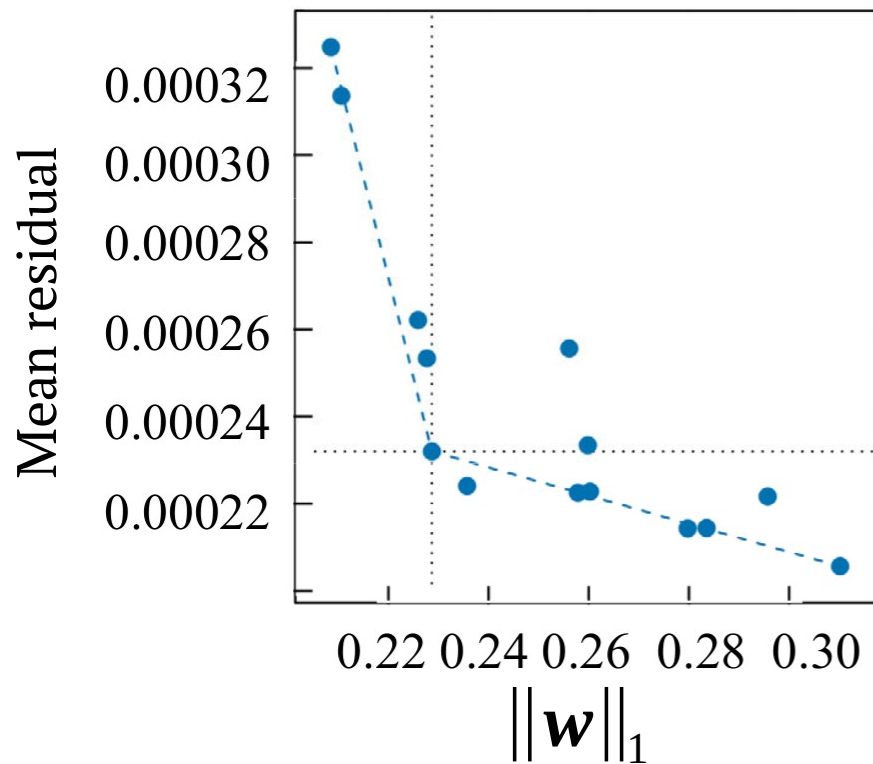
- Split-half only validates a two-component model

Sparse PARAFAC

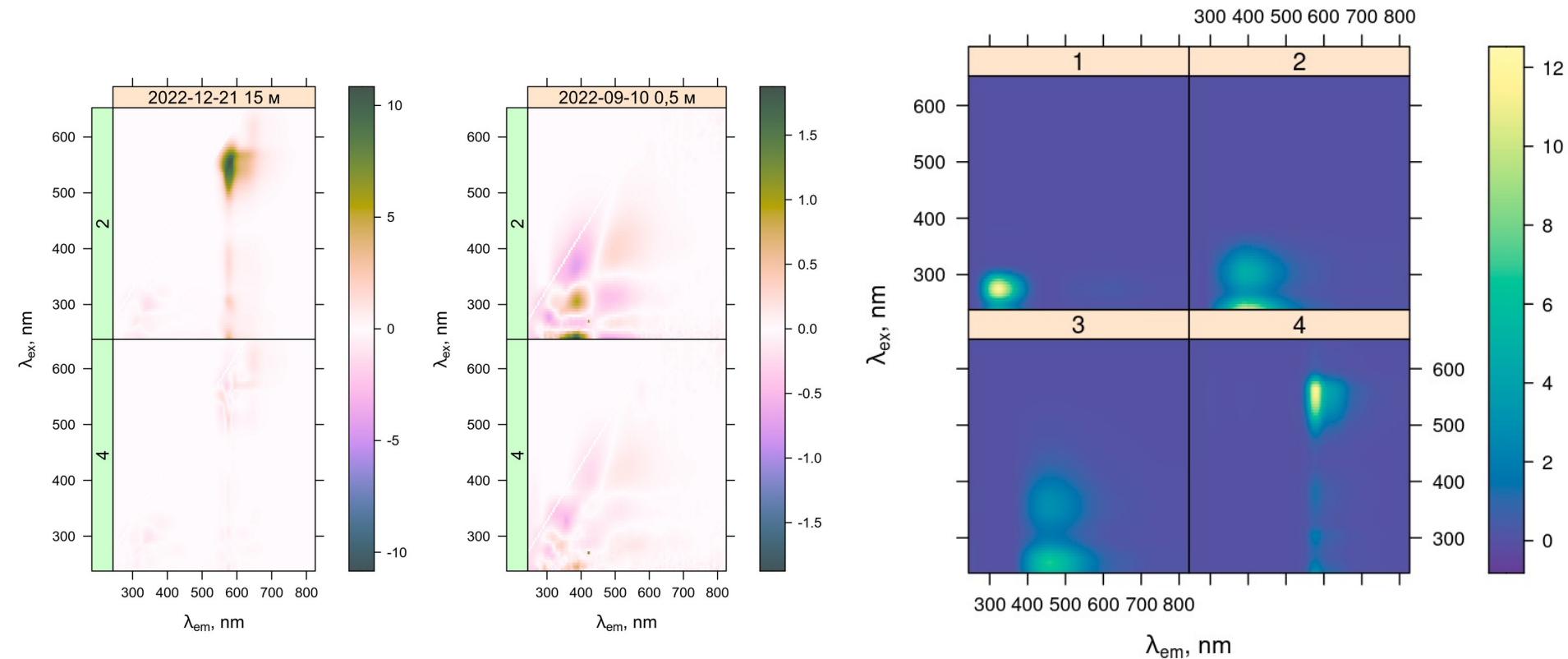
$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{w}} & \left(\frac{1}{2 N_F} \sum_{i,j,k} \left(F_{i,j,k} - \sum_r w_r A_{i,r} B_{j,r} C_{k,r} \right)^2 + \lambda_1 \|\mathbf{w}\|_1 \right) \\ \text{s.t.} \quad & \|\mathbf{A}\|_2^2 = N_A, \|\mathbf{B}\|_2^2 = N_B, \|\mathbf{C}\|_2^2 = N_C \end{aligned}$$

- Implementation: `alabama::auglag`
- Tune the regularisation coefficient instead of the number of non-zero components

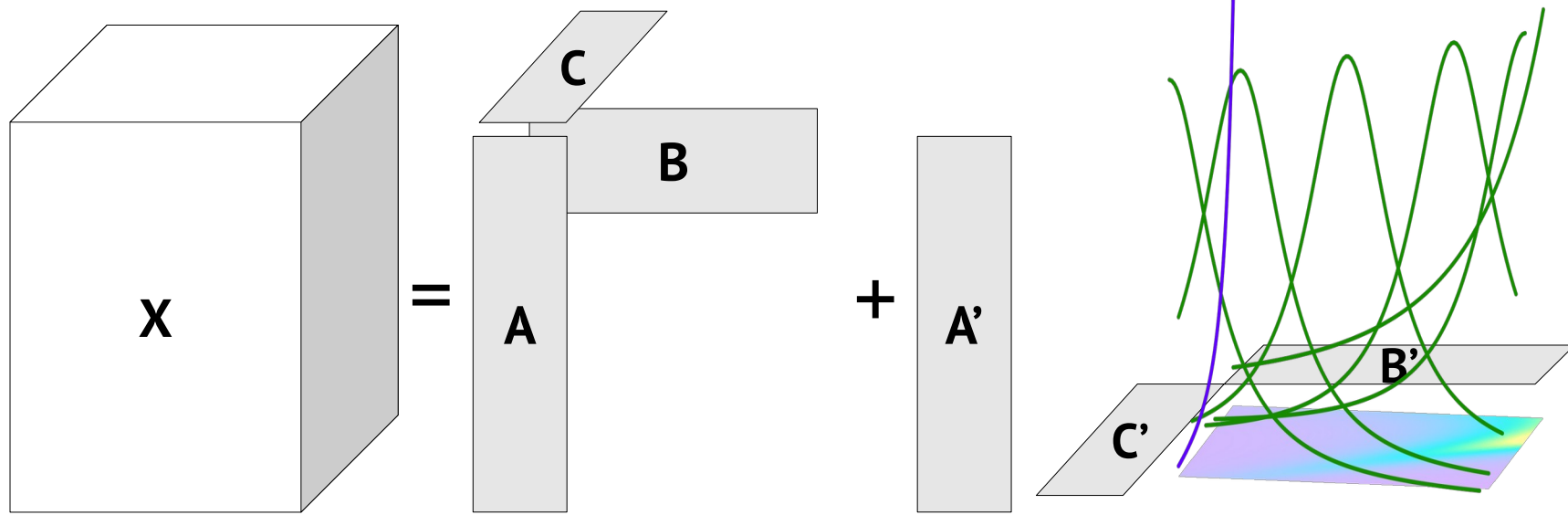
Choosing the penalty



Sparse 4-component solution

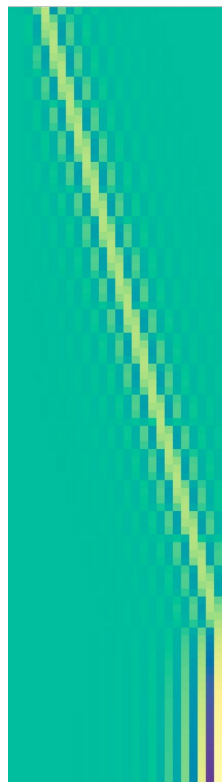


Can scattering signal have tensor structure?



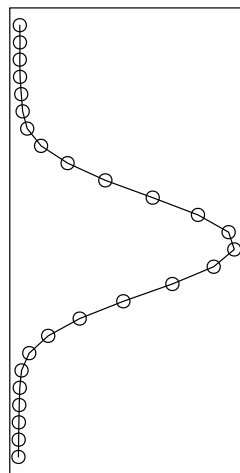
Shifting the scattering spectra

interpolation
matrix



H

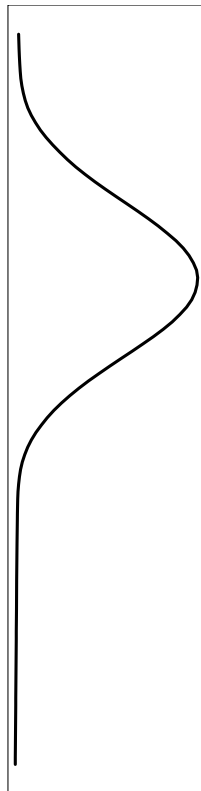
×



emission
loadings

=

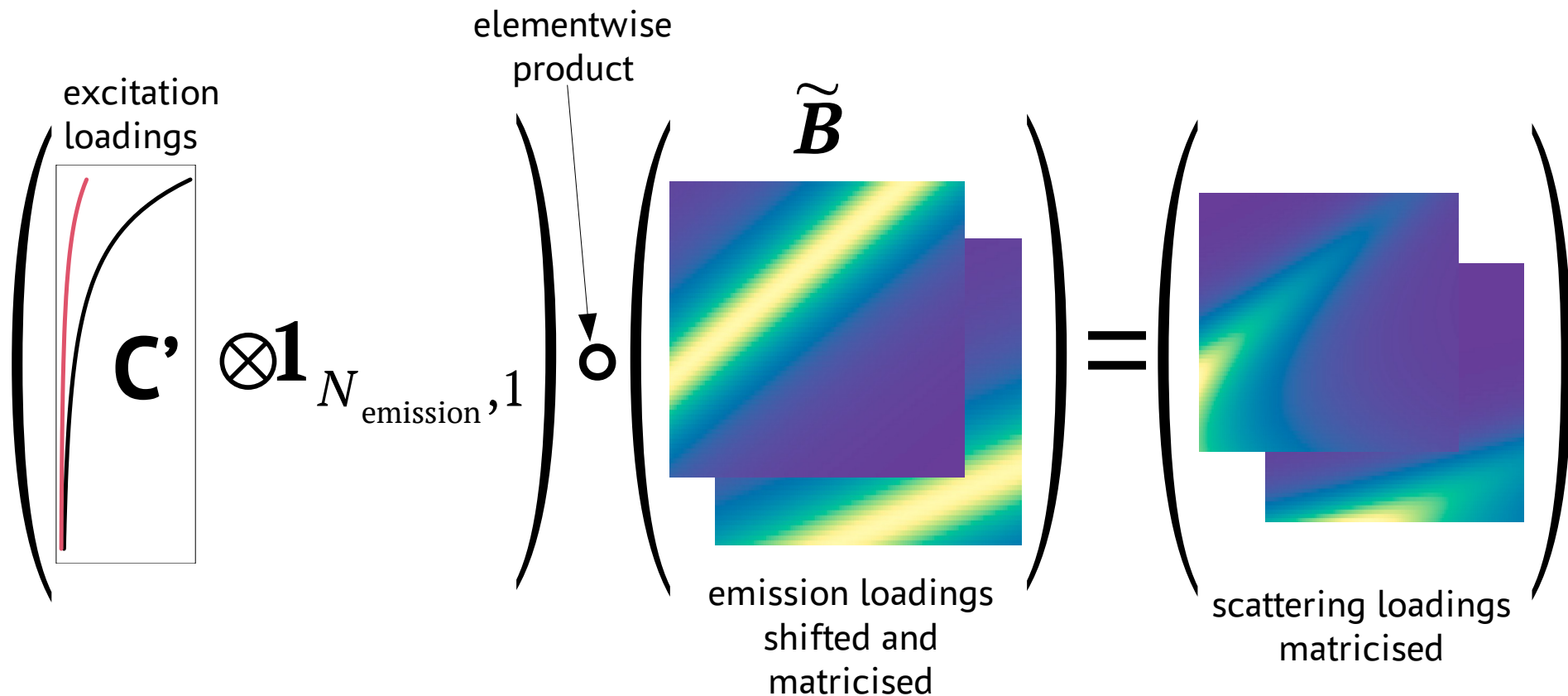
scattering at
given λ_{ex}



- Stack multiple hat matrices on top of each other to get the whole scattering EEM as a vector

$$\begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_{N_{\text{ex}}} \end{pmatrix} \mathbf{b}_{\cdot, j} \rightarrow \text{vec}(\tilde{\mathbf{B}})$$

Constructing the scattering loadings



Putting it all together

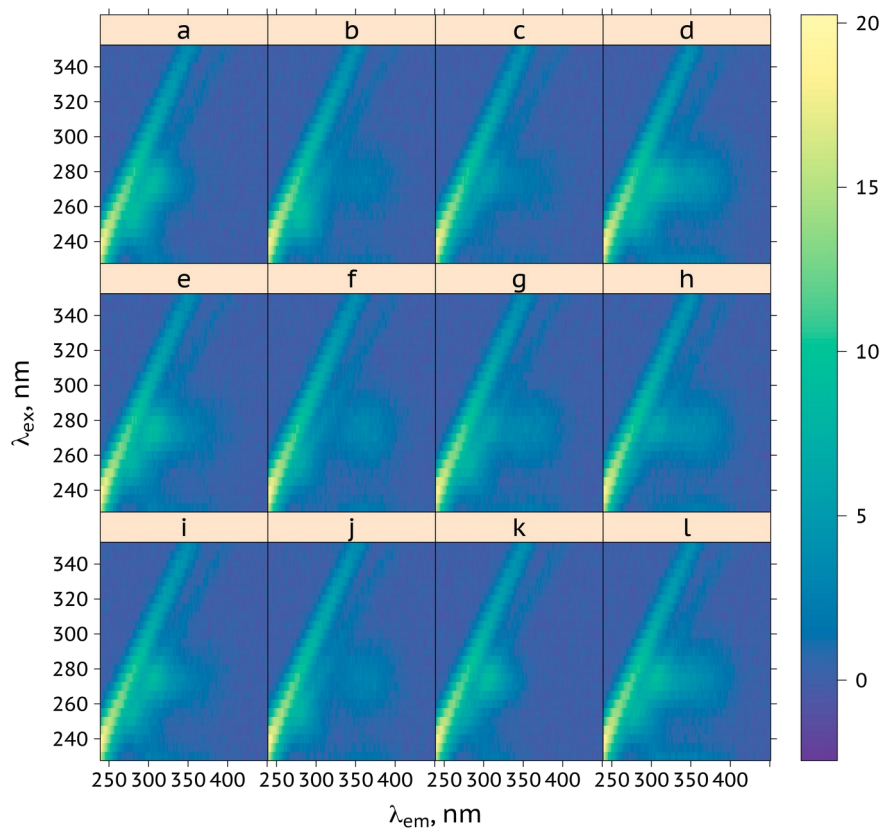
$$\text{matricise}(\tilde{\mathbf{X}}) \approx \underbrace{\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T}_{\text{PARAFAC}} + \mathbf{A}' \left((\mathbf{C}' \otimes \mathbf{1}_{N_{\text{emission}}, 1}) \circ (\mathbf{H}_{\text{stacked}} \mathbf{B}') \right)^T$$

Scattering scores

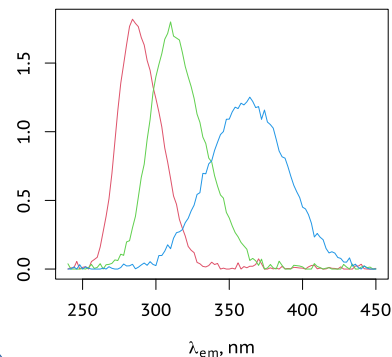
Excitation loadings replicated for every
emission wavelength

Emission scores replicated, shifted,
and interpolated at every excitation
wavelength

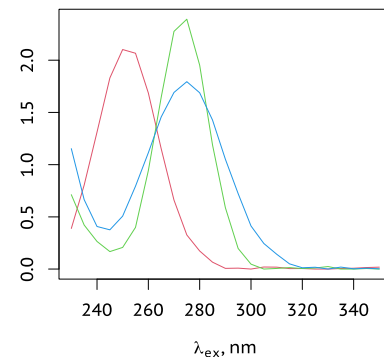
Preliminary results



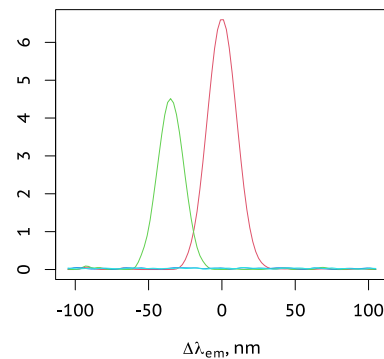
Fluorescence emission



Fluorescence excitation



Scattering emission



Scattering excitation

