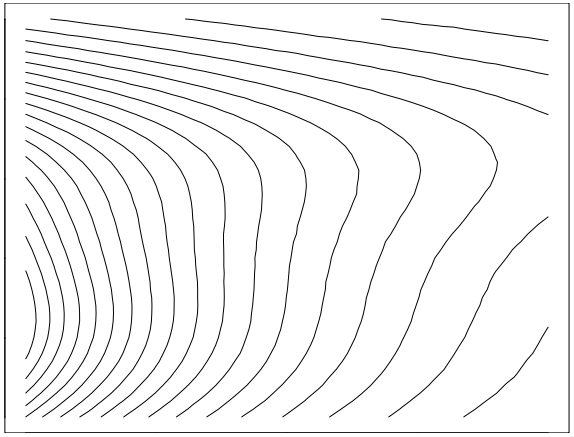
Standard error estimation and multi-dataset modeling with TIMP

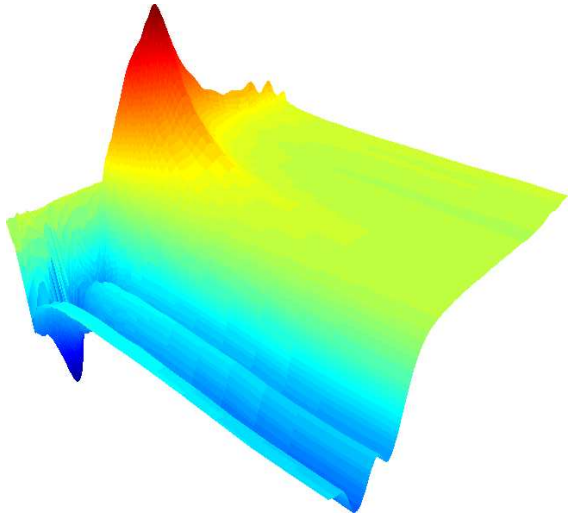
K. M. Mullen, L. J. G. W. van Wilderen, M. L. Groot and I. H. M. van Stokkum
Department of Physics and Astronomy, Vrije Universiteit Amsterdam, The Netherlands
{kate,luuk,marloes,ivo}@few.vu.nl

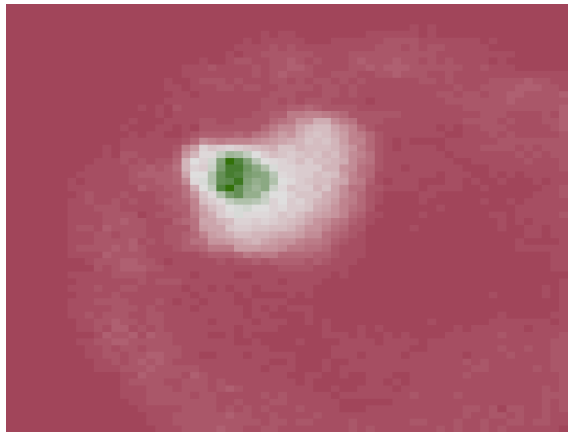
Supported by The Netherlands Organisation for Scientific Research (NWO) grant 635.000.014

TIMP is a package for fitting superposition models that has been applied to measurements arising in

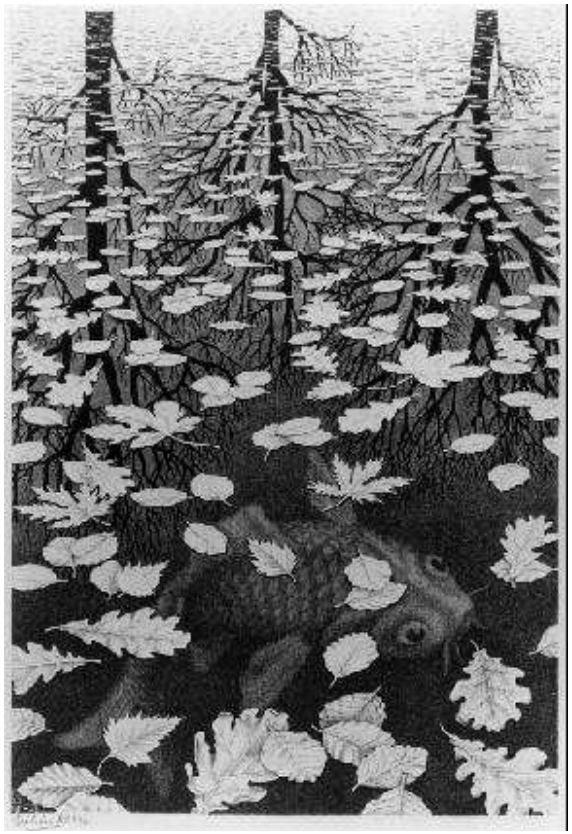
- time (and/or temperature, polarization, pH)-resolved spectroscopy
- fluorescence lifetime imaging microscopy (FLIM)







superposition of below-surface, on-surface, and above-surface states resolved with respect to location (say, pixel number):



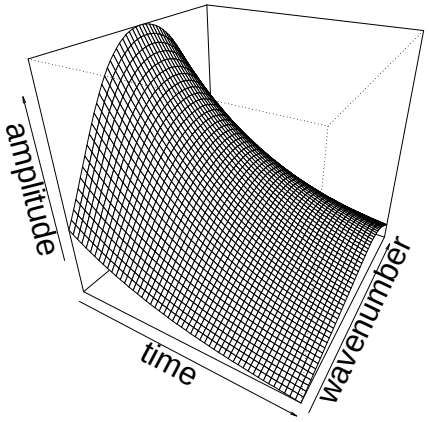
in general, measurement may be with respect to many independent variables

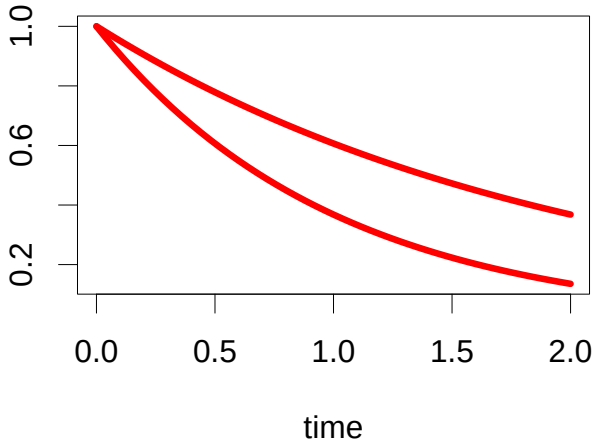
In many experiments (e.g., those in spectroscopy, fluorescent lifetime image measurement, mass spectrometry), the measurement also can be described by a superposition of the contribution of states in 2 or more independent variables

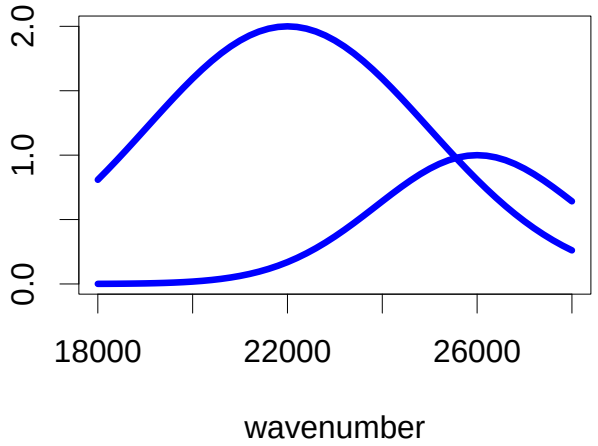
basic equation for 2-way data representing n_{comp} components:

$$\begin{aligned} \text{Measurement} &= \sum_{l=1}^{n_{\text{comp}}} c_l \epsilon_l \\ \Psi &= \left[\begin{array}{c|c|c|c} 1 & 2 & \cdots & n_{\text{comp}} \\ \vdots & \vdots & \ddots & \vdots \end{array}\right] \left[\begin{array}{c|c|c|c} 1 & 2 & \cdots & n_{\text{comp}} \\ \vdots & \vdots & \ddots & \vdots \end{array}\right]^T \\ \Psi &= CE^T \end{aligned}$$

often can postulate a parametric model for the measurement with respect to a subset of the independent variables, e.g., for 2-way data







$$\begin{aligned} \Psi &= CE^T \\ &= C(\theta)E^T \end{aligned}$$

for 2-component C, where Ψ represents timepoints $t_1, t_2, \ldots, t_x$, an example model with $\theta = \{\theta_1, \theta_2\}$:

$$C(\theta) = \left[\begin{array}{c|c} \exp(-\theta_1 t_1) & \exp(-\theta_2 t_1) \\ \exp(-\theta_1 t_2) & \exp(-\theta_2 t_2) \\ \vdots & \vdots \\ \exp(-\theta_1 t_x) & \exp(-\theta_2 t_x) \end{array}\right]$$

bilinear form of model $\Psi = C(\theta)E^T$ allows solving for least-squares estimates of E given estimates for θ as

$$E = (C(\theta)^T C(\theta))^{-1} C(\theta)^T \Psi = C(\theta)^+ \Psi$$

the estimation problem is then

$$\text{Minimize} \parallel vec(I - C(\theta)C(\theta)^+)\Psi) \parallel_2$$

reducing dimension of nonlinear parameter space to just length(θ)
(as opposed to length(θ) + (`dim(E)[1] * dim(E)[2]`) if solving for the entries of E as nonlinear parameters)

[Parameter estimation problem]

TIMP fits separable nonlinear models

given

- the number of contributing components
- a parametric model for each component with respect to subset of independent variables

obtain

- estimates for nonlinear parameters
- the evolution of components with respect to the independent variables lacking a parametric model solved for as conditionally linear parameters

[Variable projection]

the core of variable projection:
iteratively move $\hat{\theta}$ in a direction determined by approximating $\frac{d(I-C(\theta)C^{+}(\theta))}{d\theta}\Psi$

Available approximations:

- Golub-Pereyra exact analytical solution based on $\frac{dC^{+}}{d\theta}$
- Kaufman approximation of analytical solution
- finite-difference based approximation

The `nls` function contains a variable projection algorithm

- uses Golub-Pereyra solution
- accessible via the option `algorithm="plinear"`

Golub, G.H., Pereyra, V. (1973), The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate, SIAM J. Num. Anal., 10, 413-432.

Kaufman, L. (1975), A variable projection method for solving separable nonlinear least squares problems, BIT, 15, 49-57.

many models for C vary in the other independent variables with which the data are resolved ...

in the 2-way case where E is $n \times n_{\text{ncomp}}$, there are often n different models for C

recalling $vec(XYZ) = (Z^T \otimes X)vec(Y)$, where $\otimes$ is the Kronecker product,
model for Ψ is:

$$vec(\Psi) = vec(C_{\text{super}} E_{\text{super}}^T I_n) = (I_n \otimes C_{\text{super}})vec(E_{\text{super}}^T)$$

where

$$I_n \otimes C_{\text{super}} = \begin{bmatrix} C_1 & & & & \\ & \ddots & & & \\ & & C_p & & \\ & & & \ddots & \\ & & & & C_n \end{bmatrix}$$

Forming $I_n \otimes C_{\text{super}}$ requires large memory resources

[Partitioned variable projection]

Solution via partitioning:

1. get the residual $vec(I - C(\theta)C(\theta)^{+})\psi_p$ for each of the n models for C
2. concatenate these residual pieces and minimize the result with respect to θ ...

Resulting residual is the same as via standard variable projection implementation, but without the need to store and manipulate large matrices

partitioned variable projection allows application of variable projection method to modeling

- datasets resolved with respect to many independent variables
- many datasets simultaneously

without large memory resources

TIMP implements partitioned variable projection

Mullen KM, van Stokkum IHM (2007), **TIMP**: an R package for modeling multi-way spectroscopic measurements, Journal of Statistical Software, 18(3).

have model

$$vec(\Psi) = vec(CE^T I_n) = (I_n \otimes C)vec(E^T)$$

where n is the number of conditionally linear parameters.

Jacobian of the model function is

$$J = \begin{bmatrix} \frac{\partial}{\partial \theta} \\ \frac{\partial}{\partial vec(E^T)} \end{bmatrix} (I_n \otimes C)vec(E^T) = \begin{bmatrix} \frac{\partial}{\partial \theta}(I_n \otimes C)vec(E^T) \\ I_n \otimes C \end{bmatrix}$$

the linear approximation covariance matrix of both intrinsically nonlinear and conditionally linear parameters is then

$$cov\left(\begin{bmatrix} \theta \\ vec(E^T) \end{bmatrix}\right) = \hat{\sigma}^2 (J^T J)^{-1}$$

with $\hat{\sigma}^2 = SSE(\hat{\theta})/df$.

but need a lot of memory to get standard error estimates this way

can also get standard error estimates in a partitioned manner

get standard error estimates for the nonlinear parameters θ, where J is Jacobian of partitioned model function, as

$$cov\left(\left[\begin{array}{c}\theta\end{array}\right]\right) = \hat{\sigma}^2(J^T J)^{-1}$$

and standard error estimates for the conditionally linear parameters as

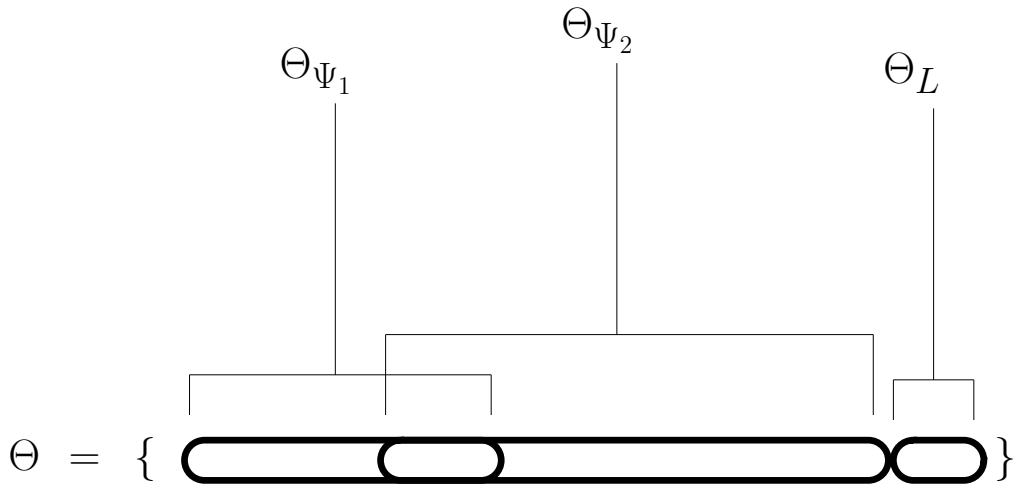
$$\begin{aligned} cov(\epsilon_p) &= \sigma^2(C_p^+ C_p^{+^T}) + G cov(\theta) G^T \\ &= \sigma^2(R_p^{-1} R_p^{-T}) + G R_J^{-1} R_J^{-T} G^T \end{aligned}$$

where G consists of columns $C^+ \frac{dC_p}{d\theta_i}\epsilon_p$, R_p results from the QR decomposition of C, and R_J results from the QR decomposition of J.

Ivo H. M. van Stokkum (2005), Global and target analysis of time-resolved spectra, lecture notes for the troisième cycle de la physique en suisse romande. Technical report.

[New TIMP options: facilitating multidataset model specification]

- the model for each dataset results in a residual vector
- concatenating these residual vectors results in a residual vector for a multidataset model
- multidataset models fit by minimizing total residual vector with respect to all parameters
- parameters may be used in the model for multiple datasets or to scale the residuals of a single dataset



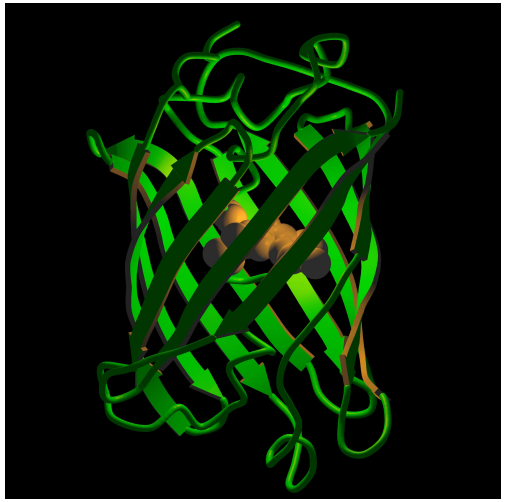
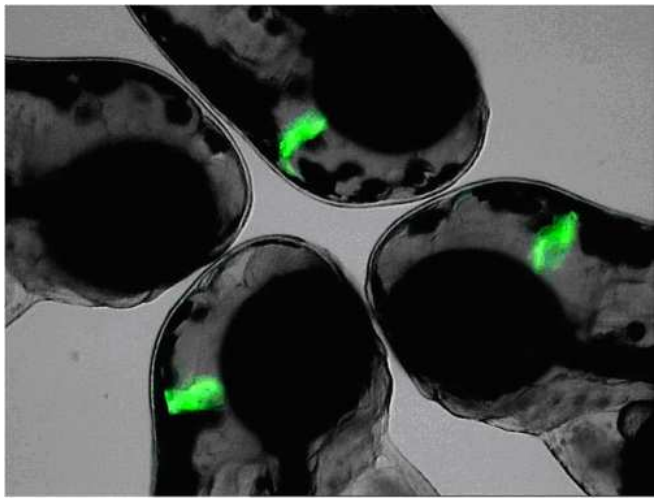
multidataset model specification in **TIMP**:

- previously based on a single model and specification of per-dataset differences
- now also can also map each dataset to a (possibly) separate model

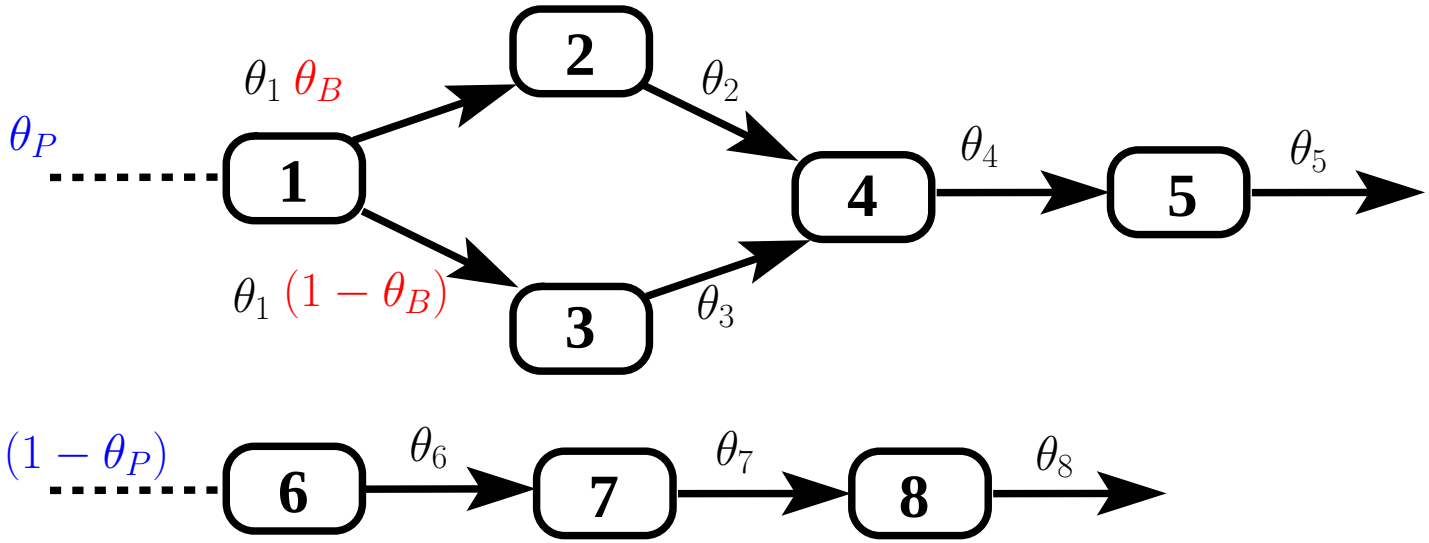
[Case study on GFP]

green fluorescent protein (GFP):

- widely used biomarker
- studied in our group via ultrafast visible/mid-infrared pump-probe spectroscopy
- detailed kinetic model for fluorescence decay sheds light on proton transfer pathway

van Wilderen LJGW, van Stokkum IHM, Mullen KM, Arents JC, Kennis JTM, Hellingwerf KJ, van Grondelle R, Groot ML (2007). "The pathway for proton transfer in Green Fluorescent Protein". Submitted.



A compartmental model for the kinetics of GFP

- structure describes kinetics in both H_2O and D_2O buffers
- θ_P and the kinetic rates $\theta_1 - \theta_8$ estimated separately for the group of datasets representing each buffer
- θ_B estimated using all datasets

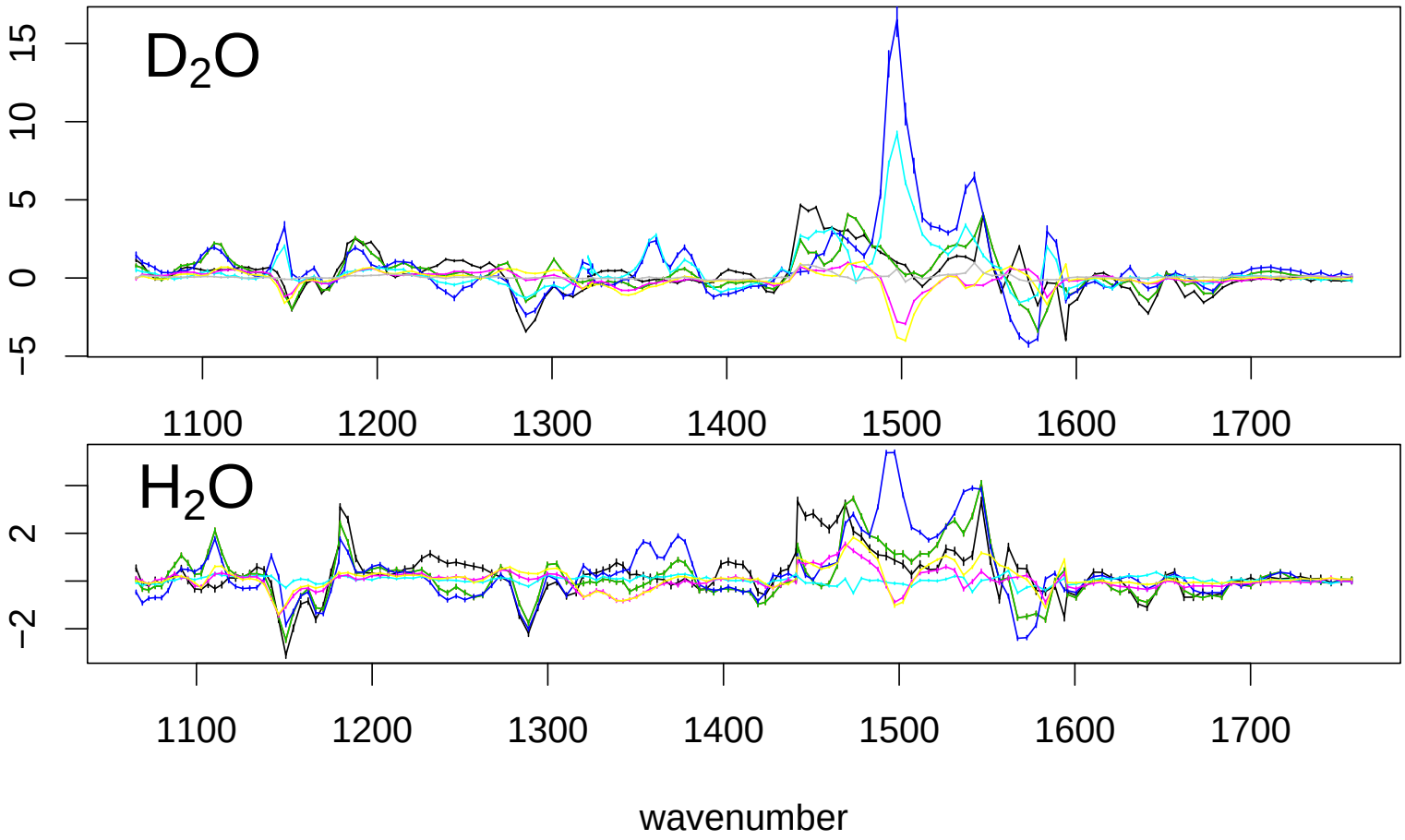
	kinetic rates	branching θ_B	θ_P	IRF
39 H_2O datasets	8	1	7	20
40 D_2O datasets	8		7	20

θ_P and IRF parameters are estimated separately for datasets in different wavenumber ranges

total number of

- nonlinear parameters: 71
- datapoints: 282,000

spectra for the H_2O and D_2O datasets estimated as conditionally linear parameters, with standard error bars:



The two sets of spectra represent a total of 2091 conditionally linear parameters.

new options for multidataset model specification in **TIMP** make it easy to:

- specify a model
- assign H_2O datasets to copy 1 of the model
- assign D_2O datasets to copy 2 of the model
- link the parameter θ_B between all datasets
- fit model parameters to all 79 datasets simultaneously

validate results using

- knowledge of physically plausible parameter values
- standard error estimates for nonlinear and conditionally linear parameters
- SVD of residuals

[Conclusions and outlook]

- package **TIMP** fits superposition models
- includes new options to
  - estimate standard errors of conditionally linear parameters
  - facilitate specification of models for multiple datasets
- package used to perform elaborate case studies, e.g., on GFP measurements

outlook:

- develop options for (largely) automated model-based analysis of mass spectrometry data
- develop java-based GUI

obtain **TIMP** from:

- R-Forge: `https://r-forge.r-project.org/projects/timp/`
- CRAN: `http://cran.r-project.org/src/contrib/Descriptions/TIMP.html`

[Acknowledgments]

Ivo H. M. van Stokkum, Vrije Universiteit Amsterdam, project leader

Sergey Laptenok, Belarusian State University and Wageningen University, FLIM modeling

Joris Snellenburg, Vrije Universiteit Amsterdam, java-based GUI

Biophysics group, Vrije Universiteit Amsterdam, data, testing, model ideas