View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by VU Research Portal

VU UNIVERSITEIT AMSTERDAM

VU Research Portal

Algorithms for separable nonlinear least squares with application to modelling time-resolved spectra

Mullen, Katharine; van Stokkum, Ivo; Vengris, Mikas

published in
Journal of Global Optimization
2007

DOI (link to publisher)
10.1016/0895-4356(94)00152-G

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Mullen, K., van Stokkum, I., & Vengris, M. (2007). Algorithms for separable nonlinear least squares with application to modelling time-resolved spectra. *Journal of Global Optimization*, *38*(2), 201-213. https://doi.org/10.1016/0895-4356(94)00152-G

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 27. May. 2021

J Glob Optim (2007) 38:201–213
DOI 10.1007/s10898-006-9071-7

ORIGINAL PAPER

Algorithms for separable nonlinear least squares with application to modelling time-resolved spectra

Katharine M. Mullen · Mikas Vengris · Ivo H. M. van Stokkum

Received: 15 December 2005 / Accepted: 27 July 2006 / Published online: 29 March 2007
© Springer Science+Business Media B.V. 2007

Abstract The multiexponential analysis problem of fitting kinetic models to time-resolved spectra is often solved using gradient-based algorithms that treat the spectral parameters as conditionally linear. We make a comparison of the two most-applied such algorithms, alternating least squares and variable projection. A numerical study examines computational efficiency and linear approximation standard error estimates. A new derivation of the Fisher information matrix under the full Golub-Pereyra gradient allows a numerical comparison of parameter precision under variable projection variants. Under the criteria of efficiency, quality of standard error estimates and parameter precision, we conclude that the Kaufman variable projection technique performs well, while techniques based on alternating least squares have significant disadvantages for application in the problem domain.

Keywords Separable nonlinear models · Time-resolved spectra · Variable projection · Alternating least squares · Fisher information

1 Introduction

State-of-the-art dynamical experiments in photophysics result in huge datasets of time-resolved spectra. Such data represent a spectral property associated with a photophysical system at m times and n wavelengths by an $m \times n$ matrix Ψ. For typical

K. M. Mullen (✉) · I. H. M. van Stokkum
Department of Physics and Astronomy, Vrije Universiteit Amsterdam,
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
e-mail: kate@few.vu.nl

I. H. M. van Stokkum
e-mail: ivo@nat.vu.nl

M. Vengris
Department of Quantum Electronics, Vilnius University,
Sauletekio 10, LT10223 Vilnius, Lithuania
e-mail: mikas.vengris@ff.vu.lt

experiments, m and n are of order 10^3. With such an overwhelming amount of data a model-based analysis is mandatory for interactive validation of hypotheses regarding physicochemical mechanisms of the underlying system. The basic kinetic model applied to Ψ is

$$\Psi = CE^T + \Xi = \sum_{l=1}^{n_{\text{comp}}} c_l e_l^T + \Xi = \sum_{l=1}^{n_{\text{comp}}} exp(-\phi_l t) e_l^T + \Xi \tag{1}$$

where column l of C represents the concentration in time of a spectrally distinct subsystem contributing a component to Ψ, column l of E describes the spectrum of that subsystem, n_{comp} is the number of contributing components, and Ξ is a residual matrix with spherical Gaussian distribution. Elements of Ψ, C, and E are in $\mathbb{R}$, but no other constraints are enforced in the general case. Estimation of parameters ϕ under least-squares criteria is thus a multiexponential analysis problem, the difficulty of which is well-known [3,28]. Problems in multiexponential analysis are ubiquitous in physics applications in which data is modelled by the solution of first-order differential equations, as Istratov and Vyvenko review [18].

The estimation problem associated with estimating ϕ in Model (1) under least-squares criteria is

$$\text{Minimize } \| \text{ vec}(C(\phi)E^T - \Psi) \|_2, \tag{2}$$

which is an instance of the unconstrained optimization problem Minimize $\gamma(x), x \in \mathbb{R}^n$ in which the variables separate into $x = (y, z)$ with $y \in \mathbb{R}^p$, $z \in \mathbb{R}^q$, $p + q = n$, and the subproblem

$$\text{Minimize } \gamma(y, z), \tag{3}$$

is easy to solve for fixed z, and, more generally, of the bilinear programming problem [1,9,17]. Separating the parameters reduces the n-dimensional unconstrained optimization problem to the q-dimensional unconstrained problem

$$\text{Minimize } \gamma(y(z), z), \tag{4}$$

where $y(z)$ denotes a solution of (3). In the considered application $y(z)$ is solved as the solution of a linear-least squares problem for fixed z, there are hundreds more conditionally linear parameters y than intrinsically nonlinear parameters z, and linear approximation standard error estimates about estimates for z are desired for model validation. These structural features of the problem and the requirement for standard error estimates make gradient-based algorithms that exploit the conditional linearity of Problem (2) attractive, though a variety of other algorithms, e.g., Branch and Cut methods [2], evolutionary search [35], or Prony-based methods [22] are also applicable. The development of gradient-based methods for the separable Problem (4) is chronicled in, e.g., [13,23,29]. The gradient-based algorithms most commonly applied to Problem (2) are based on alternating least squares [6,7,10,19] or variable projection [14,21,33]. These techniques have been numerically compared in [4] for a single nonlinear parameter, and in [11] for small datasets (<70 datapoints). Theoretical comparisons of gradient-based methods for separable problems have been made in [4,8,20,23,27]. In this paper we extend the literature comparing gradient-based methods for separable nonlinear optimization problems to Problem (2), the central estimation problem in fitting parametric kinetic models to time-resolved spectra.

A comparison of techniques in the photophysical modelling application domain is desirable due to the difficulty of Problem (2), which is not identifiable [34] and

sensitive to starting values [24,30]. Convergence issues due to ill-conditioning when two or more decay rate parameters ϕ_l are close are well-known [22,25], and are partially dependent on the choice of gradient. The stochastic noise term contained in measured Ψ introduces a further source of difficulty by complicating the sum-of-square error parameter surface of ϕ with local minima. The performance of alternating least squares and variable projection variants is studied here in such a way as to expose the vulnerabilities and strengths of the algorithms in the face of these difficulties as they occur in typical photophysical model fitting problems. To the best of our knowledge this is the first such comparison in the literature.

Alternating least squares and variable projection variants are presented in terms of their gradients in Sect. 2. The ability of the algorithms to deal with degeneracy in the case of similar decay rate parameters ϕ_l, ϕ_j is studied theoretically in Sect. 3 by comparison of Fisher information matrices (FIM) associated with parameter estimates under variable projection variants. This Sect. contains a new derivation of the FIM under the full Golub-Pereyra variable projection functional. Sect. 4 discusses the simulation of realistic datasets of time-resolved spectra to be used in numerical comparison. A numerical study is made in Sect. 5 to highlight convergence issues and sensitivity to starting values. Section 5.2 contains a numerical comparison of variable projection techniques using FIMs as rate constants vary in such a way to make Problem (2) more nearly-degenerate.

2 Gradient-based algorithms for separable nonlinear least squares

Gradient-based algorithms for solution of Problem (2) estimate E as $\hat{E}^T(\phi) = C^+\Psi$ where $+$ is the Moore-Penrose pseudoinverse, so that Problem (2) may be written as

$$\text{Minimize} \parallel (I - C(\phi)C^+(\phi))\Psi \parallel_2 . \tag{5}$$

The gradient-based techniques most often applied to Problem (5) are based on either the alternating least squares (ALS) or variable projection functionals. ALS was introduced by Wold [36] as NIPALS and has a simple functional form which neglects the derivative of the pseudoinverse C^+. The variable projection gradient (GP) makes use of the derivative of C^+ due to Golub and Pereyra [15,16]. The approximation to the full GP functional (KAUF) introduced by Kaufman [20] is more efficient to compute and for simple models has been shown to return nearly as precise parameter estimates as the full functional [4,11].

In order to make clear the core differences between algorithms, we present ALS, KAUF and GP and a finite difference approximation of $(I - C(\phi)C^+(\phi))\Psi$ in terms of the gradient in ϕ-space, using the notation of [4]. The derivative of C with respect to the nonlinear parameters is denoted $C_\phi = \frac{dC}{d\phi^T}$. Applying the QR decomposition, $C = QR = [Q_1\ Q_2][R_{11}\ 0]^T$, where Q is $m \times m$ and orthogonal and R is $m \times n_{\text{comp}}$. Assuming C is of full column rank, $C^+ = R_{11}^{-1}Q_1^T$. Then, where "convergence" is some appropriate stopping criterion and the iteration subscript s is suppressed, we have

Algorithms ALS, KAUF, GP, NUM:
1. Choose starting ϕ approximately
2. *For* $s := 1, 2 \ldots$ *until* convergence *do*
Determine the gradient in ϕ-space according to:

$$\begin{aligned}
&\text{NUM} := \text{finite difference approximation of } \tfrac{d(I-CC^{+})}{d\phi}\Psi \\
&\text{GP} := -Q_2Q_2^TC_\phi C^+\Psi - Q_1R_{11}^{-T}C_\phi^TQ_2Q_2^T\Psi \\
&\text{KAUF} := -Q_2Q_2^TC_\phi C^+\Psi \\
&\text{ALS} := -C_\phi C^+\Psi \\
&\phi_{s+1} := \text{STEP}(\phi_s, \text{gradient}, \ldots)
\end{aligned}$$

In the presentation of the algorithms above, STEP refers to the method of determining the step-size, which is not further discussed. This allows for a clear description and separates the question of which STEP method is optimal from the differences between the gradients.

For a numerical comparison, we consider two varieties of ALS differing in the STEP method. The first (ALS-GN) makes a Gauss-Newton step given the ALS gradient. The second (ALS-LS) makes a Gauss-Newton step augmented by a line search until the sum-square error (SSE) is seen to increase. KAUF, GP, and NUM are considered under a Gauss-Newton step. Simulation studies indicate that for the numerical problems considered in Sect. 5, replacement of the Gauss-Newton step with a Levenberg-Marquardt step does not appreciably alter the performance for any of the algorithms considered.

Implementation is straightforward using library subroutines for QR decomposition, finite difference derivatives, and nonlinear least squares. Such subroutines are found, for instance, in the *base* and *stats* packages of the R language and environment for statistical computing [26], where we base the implementation for numerical comparison. An analytical expression for C_ϕ is used for models based on a sum-of-exponentials. Under more complicated models for C a finite difference approximation of C_ϕ is often desirable.

We now summarize some prior results comparing subsets of the algorithms under consideration. Ruhe and Wedin [27] have shown that for starting ϕ close to the solution, the asymptotic convergence rates of KAUF and GP are superlinear whenever application of Gauss-Newton to the unseparated parameter set $(\phi + E)$ has a superlinear rate of convergence, and that ALS always has only a linear rate of convergence. Bates and Lindstrom [4] demonstrated that for a simple model having a single nonlinear parameter the performance of KAUF and GP was similar. Gay and Kaufman [11] also performed a comparison of KAUF and GP on several small datasets, ($<$70 data points), demonstrating that the time to compute KAUF was about 25% less than the time to compute GP for the range of problems considered.

3 Parameter precision under variable projection variants

The precision of nonlinear parameter estimates ϕ is a means of evaluating the performance of algorithms on Problem (2) of special interest on nearly-degenerate problems, i.e., when optimal estimates for two or more nonlinear parameters are close, so that the data are well-approximated by a lower-order sum-of-exponentials. Sect. 4.1 further elaborates the importance of parameter precision in solving nearly-degenerate problems.

A means of quantifying the precision of a vector of parameter estimates is found in the FIM. The structure of the FIM provides insight into contributions to parameter precision, and FIMs may be numerically compared under different gradients, as in

Sect. 5.2. The resolution limit of exponential analysis has been oft-studied in terms of FIMs and other information-theoretic metrics, as discussed in [18]. Badu and Bresler [3] have studied the connection between the stochastic stability of nonlinear least squares problems and the FIM with attention to separable problems such as Problem (2).

Definition 3.1 Where J is the gradient of the residual function with respect to the nonlinear parameters ϕ and the model error σ^2 is determined as $\sigma^2 = SSE(\phi)/df$, with df the degrees of freedom of the model, and where, as throughout, the noise Ξ is assumed to have spherical Gaussian distribution, the FIM M may be defined as

$$M = \sigma^{-2}\text{vec}(J)^T\text{vec}(J) = \sigma^{-2}\widetilde{M}. \tag{6}$$

When M is positive definite the covariance estimate of any unbiased estimator of parameter vector ϕ is bounded below by the inverse of M (the Cramér-Rao Bound), so that

$$Cov[\hat{\phi}] \geq M^{-1}. \tag{7}$$

We will now give functions for $\widetilde{M}$ under the variable projection algorithms *KAUF* and *GP*.

Proposition 3.1

$$\widetilde{M}_{\text{KAUF}} = \text{vec}(C_\phi)^T(E^TE \otimes P)\text{vec}(C_\phi). \tag{8}$$

Proof J_{KAUF} is given as

$$J_{\text{KAUF}} = Q_2Q_2^TC_\phi C^+\Psi = PC_\phi E^T, \tag{9}$$

where $P = Q_2Q_2^T$.

Writing J_{KAUF} in vectorized form,

$$\begin{aligned} \text{vec}(J_{\text{KAUF}}) &= \text{vec}(PC_\phi E^T) &(10)\\ &= (E \otimes P)\text{vec}(C_\phi). &(11)\end{aligned}$$

Then from [32],

$$\begin{aligned} \widetilde{M}_{\text{KAUF}} &= \text{vec}(J_{\text{KAUF}})^T\text{vec}(J_{\text{KAUF}}) &(12)\\ &= ((E \otimes P)\text{vec}(C_\phi))^T((E \otimes P)\text{vec}(C_\phi)) &(13)\\ &= \text{vec}(C_\phi)^T(E^TE \otimes P)\text{vec}(C_\phi). &(14)\end{aligned}$$

It is often convenient to consider $\widetilde{M}$ by entry $\widetilde{M}_{ij}$. This is

$$(\widetilde{M}_{\text{KAUF}})_{ij} = \text{vec}(C_{\phi_i})^TE^TE \otimes P\text{vec}(C_{\phi_j}), \tag{15}$$

where $\text{vec}(C_{\phi_i})$ is the vector representation of $\frac{dC}{d\phi_i}$.

For a two column matrix C in which $c_l = \exp(\phi_l t)$, $\text{vec}(C_{\phi_1}) = \begin{pmatrix} g_1 \\ 0 \end{pmatrix}$ and $\text{vec}(C_{\phi_2}) = \begin{pmatrix} 0 \\ g_2 \end{pmatrix}$, where $g_i = -t\exp(-\phi_i t)$. For this case the expression for $\widetilde{M}_{\text{KAUF}}$ simplifies to

$$(\widetilde{M}_{\text{KAUF}})_{ij} = \epsilon_i^T\epsilon_j g_i^TPg_j. \tag{16}$$

Proposition 3.2 *Writing $\widetilde{M}_{\mathrm{GP}}$ per entry,*

$$(\widetilde{M}_{\mathrm{GP}})_{ij} = (\widetilde{M}_{\mathrm{KAUF}})_{ij} + \mathrm{vec}(C_{\phi_i}^T)^T (P\Psi)(P\Psi)^T \otimes C^+ (C^+)^T \mathrm{vec}(C_{\phi_j}^T). \tag{17}$$

Proof The gradient J_{GP} of the residuals with respect to the nonlinear parameters contains the extra term $Q_1 R_{11}^{-T} C_\phi^T Q_2 Q_2^T \Psi$ as compared to J_{KAUF}, so that

$$J_{\mathrm{GP}} = Q_2 Q_2^T C_\phi C^+ \Psi + Q_1 R_{11}^{-T} C_\phi^T Q_2 Q_2^T \Psi \tag{18}$$

$$= J_{\mathrm{KAUF}} + (C^+)^T C_\phi^T P\Psi. \tag{19}$$

Vectorizing J_{GP},

$$\mathrm{vec}(J_{\mathrm{GP}}) = (E \otimes P)\mathrm{vec}(C_\phi) + (P\Psi)^T \otimes (C^+)^T \mathrm{vec}(C_\phi^T), \tag{20}$$

and vectorizing J_{GP}^T,

$$\mathrm{vec}(J_{\mathrm{GP}})^T = \mathrm{vec}(C_\phi)^T (E^T \otimes P) + \mathrm{vec}(C_\phi^T)^T (P\Psi) \otimes C^+. \tag{21}$$

Then, writing $\widetilde{M}_{\mathrm{GP}}$ per entry,

$$(\widetilde{M}_{\mathrm{GP}})_{ij} = (\widetilde{M}_{\mathrm{KAUF}})_{ij} + \mathrm{vec}(C_{\phi_i}^T)^T (P\Psi)(P\Psi)^T \otimes C^+ (C^+)^T \mathrm{vec}(C_{\phi_j}^T). \tag{22}$$

where we have used the orthogonality of J_{KAUF} and $(C^+)^T C_\phi P\Psi$.

For a two column matrix C in which $c_l = \exp(\phi_l)$, the expression for $\widetilde{M}_{\mathrm{GP}}$ simplifies to

$$(\widetilde{M}_{\mathrm{GP}})_{ij} = (\widetilde{M}_{\mathrm{KAUF}})_{ij} + g_i^T P\Psi (P\Psi)^T g_j (R_{11}^T R_{11})_{ij}^{-1}. \tag{23}$$

The extra term in $\widetilde{M}_{\mathrm{GP}}$ as compared to $\widetilde{M}_{\mathrm{KAUF}}$ is associated with the more accurate representation of the Hessian of Problem (2) under J_{GP} as compared to under J_{KAUF}. The extent to which this extra term is of benefit in solving Problem (2) in practice is evaluated numerically in Sect. 5.2.

4 Data for a simulation study

For a simulation study we used a model giving rise to a multiexponential analysis problem involving two exponentials with rate constant parameters $\phi = \{k_1, k_2\}$. The generative model for the C matrix of concentrations is then $c_l = \exp(-k_l t)$, where t is a vector of times and $n_{\mathrm{comp}} = 2$ (Fig. 1).

The spectra E associated with the exponential decays are modelled as a mixture of Gaussians in the wavenumber $\bar{\nu}$ (reciprocal of wavelength) domain, so that

$$e_l(\mu_{\bar{\nu}}, \Delta_{\bar{\nu}}) = a_l \bar{\nu}^5 \exp(-ln(2)(2(\bar{\nu} - \mu_{\bar{\nu}})/\Delta_{\bar{\nu}})^2), \tag{24}$$

where e_l is column l of E describing the lth spectrum, with parameters $\mu_{\bar{\nu}}$, $\Delta_{\bar{\nu}}$, and a_l, for the location, full width at half maximum (FWHM), and amplitude, respectively. This underlying model for E is chosen because it is a simple model capable of representing real spectra in practice [32], and because the use of Gaussians to represent spectral shapes is wide-spread, (see, e.g., [33] and references therein). The algorithms presented in Sect. 2 to solve Problem (2) treat the entries of E as conditionally linear parameters so that the spectral shapes are recoverable without specification of an underlying parametric model. This is often desirable because the set of parameters

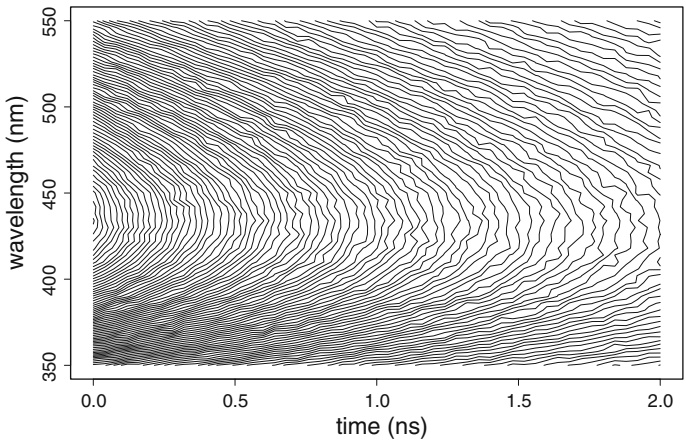

Fig. 1 Contour map of typical simulated data Ψ used in computational study. Model fitting will resolve the two contributing components

Table 1 Rate constants, spectral parameters, and amplitudes for simulated Ψ

component	k	$\mu_{\bar{\nu}}$	$\sigma_{\bar{\nu}}$	a
1	0.5	22	9	1
2	0.6	18	8	2

necessary to adequately describe the spectra of photophysical systems of interest is often large and more difficult to determine in comparison to the small and relatively simple parameterization ϕ of the concentrations C.

Given these models for C and E, data was generated with the parameter values in Table 1. Values for kinetic parameters k_1 and k_2 are similar and the spectral parameters represent overlapping spectral shapes. $n = 51$ time points equidistant in the interval 0–2 ns and $m = 51$ wavelengths equidistant in the interval 350–550 nm. These parameter values are inspired by real data ([32] and references cited therein).

4.1 Degeneracy and multimodality due to noise

Measured time-resolved spectra Ψ always contain stochastic noise. The presence of noise may introduce stationary points where $\frac{dJ(\phi)}{d\phi} = 0$ at ϕ distinct from those values underlying the deterministic model, so that the algorithms presented in Sect. 2 are sensitive to starting values. This numerical identifiability problem is well-known in kinetic modeling [12], (as is the problem of structural, i.e., deterministic model-based, lack of identifiability). In the case of convergence to a local minimum introduced by noise, estimates for kinetic parameters and spectra are often implausible from physicochemical first principles. Uninterpretable parameter estimates typically allow spurious solutions to be recognized and discarded.

In fitting Model (1) to measured time-resolved spectra the signal-to-noise ratio may be such that degeneracy is a significant issue. That is, optimal estimates for two or more rate constant parameters in the vector of nonlinear parameters ϕ may be close enough that noise disrupts the SSE surface in such a way that the globally optimal solution

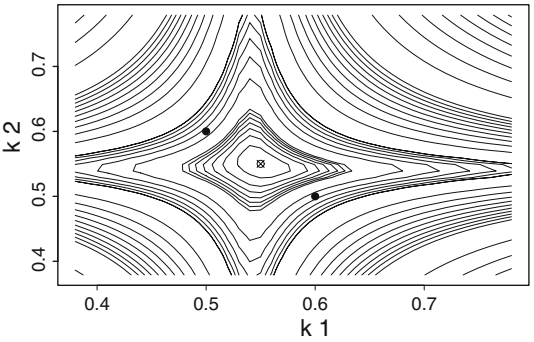

Fig. 2 The dataset described in Sect. 4 with a stochastic noise term with Gaussian distribution and zero mean having width Δ equal to 7×10^{-3} the maximum of the data. The parameter values $\phi = \{0.5, 0.6\}$ or symmetrically $\phi = \{0.6, 0.5\}$ (closed circles) underly the deterministic part of the data, and would be the globally optimal parameter estimates save for the effect of noise, which makes the lower order solution $\phi = \{0.55, 0.55\}$ (crossed circle) globally optimal

is a sum of less than n_{comp} exponentials, as reviewed in [31]. Then the least-squares solution yields estimates with $k_1 = k_2$ for $\{k_1, k_2\} \in \phi$. In nearly-degenerate cases the least squares solution is with $k_1 \approx k_2$, and the parameters may be resolved if the precision with which they are estimated is sufficiently high, as is studied numerically under the KAUF and GP algorithms in Sect. 5.2. For the simulated dataset described in Sect. 4 degeneracy is probable for noise with width of about 7×10^{-3} the maximum of the data. The SSE surface of parameters ϕ for a noise realization that results in near-degeneracy is shown in Fig. 2.

5 Computational results

Model (1) was fit to the data described in Sect. 4 with a stochastic noise term with Gaussian distribution and zero mean having width Δ equal to 3×10^{-3} the maximum of the data using each of the algorithms described in Sect. 2. The convergence criterion was reduction of sum square error (SSE) $||\text{vec}(\Psi - CE^T)||^2$ by a factor of less than $1/2^{10}$ between iterations. Estimated spectra found as conditionally linear parameters under KAUF, GP, ALS-LS or NUM well-represent the spectra used in generating the simulated data, as shown in Fig. 3.

To visualize the progress of the algorithms per iteration, the SSE as rate constants k_1, k_2 vary is evaluated, with the result being the surface shown in Fig. 4. Figure 4 also

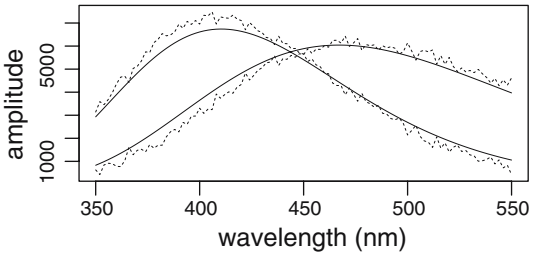

Fig. 3 Estimated spectra (dashed lines) as found with KAUF, GP or NUM by fitting the simulated dataset depicted in Fig. 1 with the two-component kinetic model described in Sect. 5. Spectra used to generate the deterministic part of the dataset (solid lines) are shown for comparison

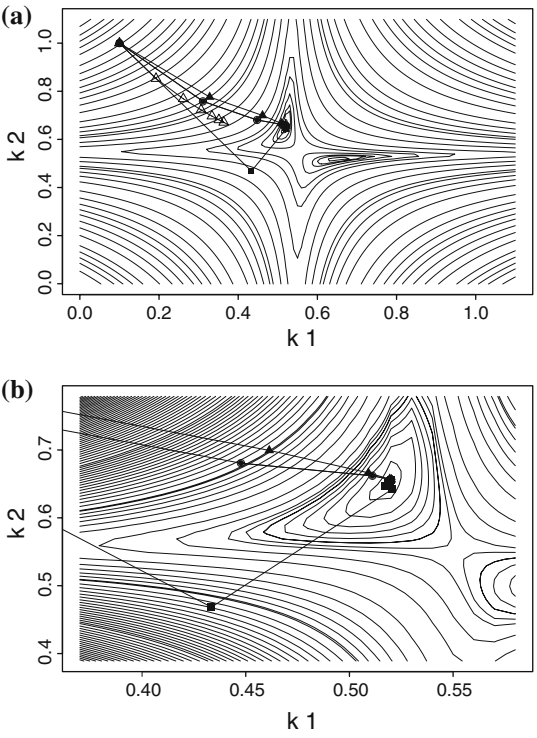

Fig. 4 Contour map of the sum square of residuals $||vec(\Psi - CE^T)||^2$ as rate constants k_1, k_2 vary, at a relatively large (**a**) and relatively small (**b**) scale. The progress of ALS-GN (unfilled triangle), ALS-LS (square), KAUF (filled triangle), and GP/NUM (filled and unfilled circles) is depicted from starting values $k_1 = 0.1, k_2 = 1$; rate constant estimates are marked with the symbol associated with each algorithm after each iteration. Spacing between contour lines is not uniform

shows the values found by each algorithm under consideration for each of 50 iterations from the starting values $k_1 = 0.1, k_2 = 1$. KAUF, GP, ALS-LS and NUM converge on the same (globally optimal) solution in 4 iterations. ALS-GN is not generally convergent in many hundreds of iterations, and from this case study and others we conclude that the Gauss-Newton step coupled with the ALS-gradient is not sufficient for the solution of typical estimation problems in this domain.

Performance from a range of starting values and on variants of the dataset under different noise realizations was examined. For cases in which globally optimal parameter values are located at the end of a valley with respect to the starting values, the performance of ALS-LS is very much hampered in terms of iterations required to convergence in comparison to KAUF, GP, and NUM. A plot of the SSE surface (as in Fig. 4) in this case shows that ALS-LS follows a zig-zagging path between the walls of the valley toward a globally optimal solution.

We conclude that the ALS gradient coupled with a line search and both variable projection methods KAUF and GP solve this problem for the considered data realizations. The KAUF algorithm typically requires the same number of iterations as the GP algorithm. ALS with line search converges in a greater or equal number of iterations as compared to KAUF and GP. The iterations required for ALS-LS are greater than for KAUF and GP when the globally optimal parameter values are at the end of a valley in SSE with respect to starting parameter estimates. Therefore in terms of iterations to convergence and sensitivity of computational efficiency to starting values, the variable projection-based algorithms demonstrate the best performance.

5.1 Standard error estimates

In order to examine the properties of linear approximation standard error estimates as returned by the algorithms under consideration, 1000 realizations of the dataset described in Sect. 4 were simulated. For each realization, the deviation$(k) = \hat{k} - k$, where $\hat{k}$ is the estimated rate constant value, and k is the value used in simulation, the linear approximation standard error ($\hat{\sigma}_{\hat{k}}$), derived from $cov(\hat{\phi}) = \hat{\varsigma}^2(J^TJ)^{-1}$, where $\hat{\varsigma}^2$ denotes the estimated variance and J is the gradient evaluated at $\hat{\phi}$, and the ratio of these two calculations, the studentized parameter deviation [5,28,32], was calculated. Table 2 reports root mean square (RMS) results.

At the level of precision collated in Table 2, results for NUM, KAUF and GP are identical. NUM and GP only differ from KAUF in the 3rd decimal place of RMS (deviation/$\hat{\sigma}_{\hat{k}}$), and from each other in the 6th.

RMS (deviation/$\hat{\sigma}_{\hat{k}}$) is expected to be 1 in linear models, and hence the degree to which this ratio approximates 1 can be used as a measure of the applicability of the linear approximation standard error returned by the respective algorithms. Under the ALS gradient, $\hat{\sigma}_{\hat{k}}$ is much too small, and not useful as a measure of confidence in parameter value estimates.

Likelihood-based confidence regions may be constructed around parameter estimates based on the likelihood ratio between the sum square of residuals $S(\hat{\phi}) = ||\text{vec}(\Psi - CE^T)||^2$ at the solution and at values $S(\phi)$ around the solution as $\phi = \{k_1, k_2\}$ is varied. The confidence level $1 - \alpha$ is determined as

$$1 - \alpha = F\left(P, N - P, (N - P)/P\frac{S(\phi) - S(\hat{\phi})}{S(\hat{\phi})}\right) \tag{25}$$

where F is the cumulative F-distribution, $P = n_{\text{comp}} = 2$, and $N = (\text{times} - n_{\text{comp}})(\text{wavelengths}) = (51-2)(51)$ [5], [28]. The resulting contour plot of confidence regions about the parameter estimates is shown in Fig. 5(a). For comparison, the linear approximation confidence regions calculated from $cov(\phi)$ for KAUF, GP, or NUM are shown in Fig. 5(b). Note that the linear approximation confidence regions are slightly too small as compared to the likelihood-based confidence regions, which is consistent with the slight underestimation of $\hat{\sigma}_{\hat{k}}$ in Table 2, as measured by the overshoot of deviation/$\hat{\sigma}_{\hat{k}}$ to 1.

In conclusion, the standard error estimates returned by both variable projection variants are usable as a measure of confidence in the associated parameter estimates, and allow, e.g., the construction of confidence regions about parameter estimates. The standard error estimates returned by ALS with line search are so poor as to prohibit

Table 2 Root mean square deviation and standard error of nonlinear parameters

		ALS-LS	KAUF/GP/NUM
RMS deviation ($\hat{k} - k$)	k_1	0.022	0.022
	k_2	0.025	0.025
RMS $\hat{\sigma}_{\hat{k}}$	k_1	0.00033	0.021
	k_2	0.00048	0.027
RMS (deviation/$\hat{\sigma}_{\hat{k}}$)	k_1	55	1.3
	k_2	37	1.2

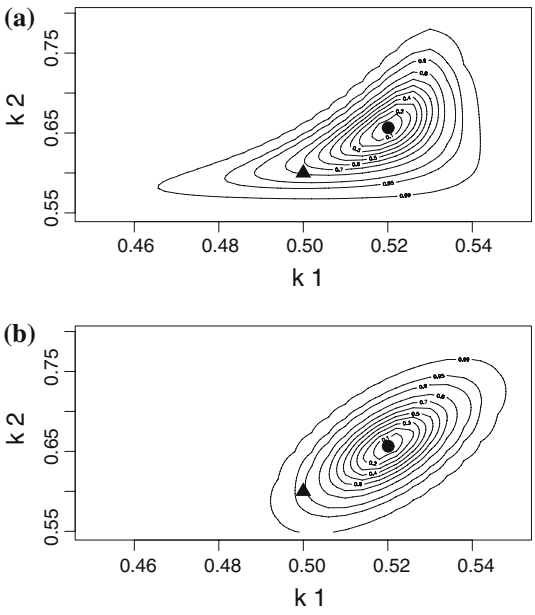

Fig. 5 For the dataset depicted in Fig. 4, (**a**) contour map of confidence levels $1-\alpha$ as determined by Eq. 25 as rate constants k_1, k_2 vary, (**b**) linear approximation confidence regions as found using KAUF, GP, or NUM for the same levels as at left. In both (**a**) and (**b**) a triangle marks the rate constant values used in simulation, and a circle marks the globally optimal values found by KAUF, GP, NUM, and ALS-LS

inference regarding the associated parameter estimates. Hence the variable projection-based algorithms also demonstrate better performance relative to ALS-based algorithms under the criteria of goodness of standard error estimates.

5.2 Numerical comparison of Fisher information matrices

The functional forms for the FIM are useful in accessing the loss of parameter precision under KAUF as compared to GP for typical problems. Relation (7) allows standard error bounds under both algorithms to be numerically compared. This comparison is of particular interest for estimation problems associated with a SSE surface of the nonlinear parameters ϕ with multiple closely spaced global minima.

For fitting Model (1) to the dataset described in Sect. 4 realized with a noise distribution having width 1×10^{-4} the maximum of the data, we studied the standard error bounds returned by KAUF and GP using Relation (7). We varied the separation between rate constants $k_2 - k_1$ by letting $k_1 = 0.5$ and varying k_2 between 1 and 0.5075. The standard error bounds under KAUF never increased by more than 5×10^{-4} percent in comparison to the bounds under GP, even when the separation $k_2 - k_1$ became very small. Hence the decrease in parameter precision under KAUF as compared to under GP is negligible even for nearly-degenerate instances of Problem (2). Since KAUF is faster to compute it may therefore be preferred for application.

6 Conclusions

Gradient-based algorithms for separable nonlinear least squares based on alternating least squares and variable projection were compared for an application in multi-exponential analysis that is common and important in fitting photophysical kinetic

models to time-resolved spectra. The efficiency of the variable projection algorithms was found to be less sensitive to starting values as compared to the efficiency of algorithms based on alternating least squares. The linear approximation confidence regions about parameter estimates using the variable projection gradients were furthermore found to well-approximate likelihood-based confidence regions, while those based on an alternating least squares gradient did not. Using a new derivation of the Fisher information matrix under the Golub-Pereyra variable projection gradient, parameter precision under variable projection techniques was compared numerically. The loss of precision under the Kaufman approximation as compared to the Golub-Pereya variable projection functional was found to be acceptable even on nearly-degenerate problems, so that the faster Kaufman approximation algorithm can be recommended for application to the problem in photophysical modelling considered here.

Acknowledgements This research was funded by Computational Science grant #635.000.014 from the Netherlands Organization for Scientific Research (NWO).

References

1. Al-Khayyal, F.: Jointly constrained bilinear programs and related problems: An overview. Comput. Math. Appl. **19**, 53–62 (1990)
2. Audet, C., Hansen, P., Jaumard, B., Savard, G.: A branch and cut algorithm for non-convex quadratically constrained quadratic programming. Math. Program. **87**, 131–152 (2000)
3. Basu, S., Bresler, Y.: Stability of Nonlinear Least Squares Problems and the Cramer-Rao Bound. IEEE T. Signal Proces. **48**, 3426–3436 (2000)
4. Bates, D.M., Lindstrom, M.J.: Nonlinear Least Squares with Conditionally Linear Parameters. In Proceedings of the Statistical Computing Section. 152–157. American Statistical Association, New York (1986)
5. Bates, D.M., Watts, D.G.: Nonlinear regression analysis and its applications. John Wiley & Sons, New York (1988)
6. Bijlsma, S., Boelens, H.F.M., Hoefsloot, H.C.J., Smilde, A.K.: Estimating reaction rate constants: comparison between traditional curve fitting and curve resolution. Anal. Chimica Acta, **419**, 197–2007 (2000)
7. Bijlsma, S., Boelens, H.F.M., Hoefsloot, H.C.J. Smilde A.K.: Constrained least squares methods for estimating reaction rate constants from spectroscopic data. J. Chemometr. **16**, 28–40 (2002)
8. Böckmann, C.: A modification of the trust-region Gauss-Newton method to solve separable nonlinear least squares problems. J. Math. Syst. Estimat. Control. **5**, 111–115 (1995)
9. Brimberg, J., Hansen, P., Mladenović, N.: A note on reduction of quadratic and bilinear programs with equality constraints. J. Global Optim. **22**, 39–47 (2002)
10. Dioumaev, A.K.: Evaluation of intrinsic chemical kinetics and transient product spectra from time-resolved spectroscopic data. Biophys. Chem. **67**, 1–25 (1997)
11. Gay, D., Kaufman, L.: Tradeoffs in Algorithms for Separable and Block Separable Nonlinear Least Squares. In Vichnevetsky, R., Miller, J.J.H. (eds.) IMACS '91, Proceedings of the 13th World Congress on Computational and Applied Mathematics, pp.157–158. Criterion Press, Dublin (1991)
12. Godfrey, K.: Compartmental Models and Their Application. Academic Press, London (1983)
13. Golub, G., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. Inverse Problems **19**, R1–R26 (2003)
14. Golub, G.H., LeVeque, R.J.: Extensions and uses of the variable projection algorithm for solving nonlinear least squares problems. Proc. Army Num. Anal. Comp. Conf. ARO Report 79–3, 1–12 (1979)
15. Golub, G.H., Pereyra, V.: The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, Tech. rep., Stanford University, Department of Computer Science (1972)
16. Golub, G.H., Pereyra, V.: The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. SIAM J. Num. Anal. **10**, 413–432 (1973)

17. Horst, R., Tuy, H.: Global Optimization: Deterministic Approaches. Springer-Verlag, Berlin, 3rd edn (1996)
18. Istratov, A.A., Vyvenko, O.F.: Exponential analysis in physical phenomena. Rev. Sci. Instrum. **70**, 1233–1257 (1999)
19. Jandanklang, P., Maeder, M., Whitson, A.C.: Target transform fitting: a new method for the nonlinear fitting of multivariate data with separable parameters. J. Chemometr. **15**, 886–9383 (2001)
20. Kaufman, L.: A variable projection method for solving separable nonlinear least squares problems. BIT. **15**, 49–57 (1975)
21. Nagle J.F.: Solving Complex Photocycle Kinetics – Theory and Direct Method. Biophys. J. **59**, 476–487 (1991)
22. Osborne, M.R., Smyth, G.K.: A modified Prony algorithm for exponential function fitting. SIAM J. Sci. Comput. **16**, 119–138 (1995)
23. Parks, T.A.: Reducible Nonlinear Programming Problems, Tech. Rep. Technical Report TR85-08, Department of Computational and Applied Mathematics, Rice University, USA (1985)
24. Petersson, J., Holmström, K.: A Review of the Parameter Estimation Problem of Fitting Positive Exponential Sums to Empirical Data, Tech. Rep. Technical Report IMa-TOM-1997-08, Department of Mathematics and Physics, Märlardalen University, Sweden (1997)
25. Petersson, J., Holmström K.: Initial Values for the Exponential Sum Least Squares Fitting Problem, Tech. Rep. Technical Report IMa-TOM-1998-01, Department of Mathematics and Physics, Märlardalen University, Sweden (1998)
26. R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0 (2004)
27. Ruhe, A., Wedin, P.A.: Algorithms for Separable Nonlinear Least Squares Problems. SIAM Rev. **22**, 318–337 (1980)
28. Seber, G.A.F., Wild, C.J.: Nonlinear regression. John Wiley & Sons, New Jersey (2003)
29. Smyth, G.K.: Partitioned algorithms for maximum likelihood and other nonlinear estimation. Stat. Computing. **6**, 201–216 (1996)
30. Bos, A. van den : A class of small sample nonlinear least squares problems. Automatica. **16**, 487–490 (1980)
31. van den Bos, A., Swarte J.H.: Resolvability of the parameters of multiexponentials and other sum models. IEEE T. Signal Process. **41**, 313–322 (1993)
32. van Stokkum, I.H.M.: Parameter Precision in Global Analysis of Time-Resolved Spectra. IEEE T. Instrum. Measure. **46**, 764–768 (1997)
33. van Stokkum, I.H.M., Larsen, D.S., van Grondelle, R.: Global and target analysis of time-resolved spectra. Biochim. Biophys. Acta. **1657**, 82–104, and erratum, **1658**, 262 (2004)
34. Varah, J.: On Fitting Exponentials by Nonlinear Least Squares. SIAM J. Sci. Stat. Comput. **1**, 30–44 (1985)
35. Wohlleben, W., Buckup, T., Herek, J.L., Cogdell, R.J., Motzkus, M.: Multichannel Carotenoid Deactivation in Photosynthetic Light Harvesting as Identified by an Evolutionary Target Analysis. Biophys. J. **85**, 442–450 (2003)
36. Wold, H., Lyttkens, E.: Nonlinear iterative partial least squares (NIPALS) estimation procedures. Bulletin. Int. Stat. Institut. **43**, 29–51 (1969)