Supplement for accelerated fitting of joint models with cumulative variations

Yan Gao[1*], Rodney A. Sparapani[1] and Sergey Tarima[1]

[1]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, 53226, Wisconsin, U.S.A.

*Corresponding author(s). E-mail(s): yagao@mcw.edu;

1 Data notations and visualization

The mathematical symbols associated with data and statistical modeling are presented in Table 1. The illustration of the survival and longitudinal data are displayed in Tables 2 and 3, respectively.

2 Matrix algebra for joint models

Appendix 2 elaborates matrix algebra for joint models. The detailed operations in the longitudinal submodel will be introduced in the subsections 2.1 and 2.2, followed by the details of the survival submodel. Note that i' and j' in the matrix notation refer to a matrix entry's row and column indices, respectively. And k' can index a vector or a matrix entry based on the context.

Notation	**Description**
Data-associated symbols	
i	subject index
j	time point index
t_i	observed survival time
z_{ij}	observed longitudinal outcome for the ith subject at jth time point
n_i	number of time points for the ith subject
n	number of subjects
$N = \sum_{i=1}^{n} n_i$	number of observations
Model-associated symbols	
k	spline-function index
K	number of spline functions
d	GK node index
D	number of GK nodes
r	grid index for a partition
R	number of grids for a partition per subject
$\dot{R} = n \times R$	number of grids for all subjects
$\dot{K} = n \times R \times K$	number of all grids and spline functions for all subjects
$\ddot{R} = R - 1$	reduced number of grids
$\widetilde{R} = \dot{R} \times D$	number of grids and GK nodes per subject

Table 1 List of notations.

i	t_i	δ_i	$\boldsymbol{x}_{i1}$	$\boldsymbol{x}_{i2}$	...	$\boldsymbol{x}_{iP}$
1	t_1	δ_1	$\boldsymbol{x}_{11}$	$\boldsymbol{x}_{12}$	...	$\boldsymbol{x}_{1P}$
2	t_2	δ_2	$\boldsymbol{x}_{21}$	$\boldsymbol{x}_{22}$	...	$\boldsymbol{x}_{2P}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	t_n	δ_n	$\boldsymbol{x}_{n1}$	$\boldsymbol{x}_{n2}$	...	$\boldsymbol{x}_{nP}$
n	t_n	δ_n	$\boldsymbol{x}_{n1}$	$\boldsymbol{x}_{n2}$	...	$\boldsymbol{x}_{nP}$

Table 2 Illustration of survival data.

i	z_{ij}	s_{ij}
1	z_{11}	s_{11}
1	z_{12}	s_{12}
1	z_{1n_1}	s_{1n_1}
2	z_{21}	s_{21}
2	z_{22}	s_{22}
2	z_{32}	s_{32}
2	z_{2n_2}	s_{2n_2}
⋮	⋮	⋮
n	z_{n1}	s_{n1}
n	z_{n2}	s_{n2}
n	z_{n3}	s_{n3}
n	z_{n4}	s_{n4}
n	z_{nn_n}	s_{nn_n}

Table 3 Illustration of longitudinal data.

2.1 CSR structure in longitudinal submodel

To store the design matrix $\boldsymbol{B} = (B_{i'j'})$ and take advantage of the sparse matrices, $\mathbf{B}^{\boldsymbol{s}}$, we adopt the popular CSR (Compressed Sparse Row) data structure. Instead of the common term "array" under many contexts (Saad, 2003), we use the matrix term "vector" to align with our representations because they are equivalent in computation. Three vectors are constructed in the CSR structure $\mathbb{S}^B(\boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{\omega})$ to support efficient matrix-vector multiplication : 1) $\boldsymbol{v} = \{v \in \mathbb{R} : v \text{ is an nonzero entry of } \boldsymbol{B}\}$ with $dim(\boldsymbol{v}) = a$ (the number of nonzeros, NNZ) ; 2) $\boldsymbol{\tau} = \{\tau \in \mathbb{N} : \tau \text{ is the column index of the nonzero entry of } \boldsymbol{B}\}$ with $dim(\boldsymbol{\tau}) = a$; 3) $\boldsymbol{\omega} = \{\omega \in \mathbb{N} : \omega \text{ is the row extent of nonzero entry of } \boldsymbol{B}\}$ with $dim(\boldsymbol{\omega}) = N + 1$ and the last element, a. That is, if $v_{k'}$ stores $B_{i'j'}$, then $\tau_{k'} = j'$ and $\omega_{i'} \leq k' < \omega_{i'+1}$. The memory footprint and the complexity for the CSR format are of the order $\mathcal{O}(n)$ instead of $\mathcal{O}(n^2)$ as in a dense matrix (Hoffman, 2021).

2.2 Cholesky factorization in longitudinal submodel

To further accelerate the computation, the covariance matrix is decomposed into the product of the diagonal matrix $\text{diag}(\boldsymbol{c})$ and the correlation matrix $\Lambda = (\Lambda_{i'j'})$: $\boldsymbol{\Sigma} = (\Sigma_{i'j'}) = \text{diag}(\boldsymbol{c}) \times \Lambda \times \text{diag}(\boldsymbol{c})$, where $\text{diag}(\boldsymbol{c})$ consists of the vector of the standard deviations $\boldsymbol{c} = \{c_1, \ldots, c_K\}$. This mapping can be reversed by the following formula: $c_{k'} = \sqrt{\Sigma_{k'k'}}$; and $\Lambda_{i'j'} = \frac{\Sigma_{i'j'}}{c_{i'}c_{j'}}$. Given that the correlation matrix is a symmetric positive-definite matrix, it has a Cholesky factorization: $\Lambda = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L} = (l_{i'j'})$ denotes a real lower triangular matrix whose diagonal elements are positive. Each component can be derived

based on the following formulas (Hoffman, 2021)

$$l_{i'i'} = \pm\sqrt{\Lambda_{i'j'} - \sum_{k'=1}^{i'-1} l_{i'k'}^2}$$
$$l_{i'j'} = \frac{1}{l_{i'j'}}\left(\Lambda_{i'j'} - \sum_{k'=1}^{j'-1} l_{i'k'}l_{j'k'}\right), \qquad i' > j'$$

which follows from

$$\begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \dots & \Lambda_{1K} \\ \Lambda_{21} & \Lambda_{22} & \dots & \Lambda_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{K1} & \Lambda_{K2} & \dots & \Lambda_{KK} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{K1} & l_{K2} & \dots & l_{KK} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \dots & l_{K1} \\ 0 & l_{22} & \dots & l_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & l_{KK} \end{bmatrix}$$
$$= \begin{bmatrix} l_{11}^2 & l_{21}l_{11} & \dots & l_{K1}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & \dots & l_{K1}l_{21} + l_{K2}l_{22} \\ \vdots & \vdots & \ddots & \vdots \\ l_{K1}l_{11} & l_{K1}l_{21} + l_{K2}l_{22} & \dots & l_{K1}^2 + \dots + l_{KK}^2 \end{bmatrix}$$

3 Convergence plots in numerical examples

The chordal approximation method converge to the true value immediately for S1 because it is actually a straight line and no estimation errors occur. For the other three functions S2-S4, the chordal approximation method also converge to the true or similar values when the number of grids are less than 200.

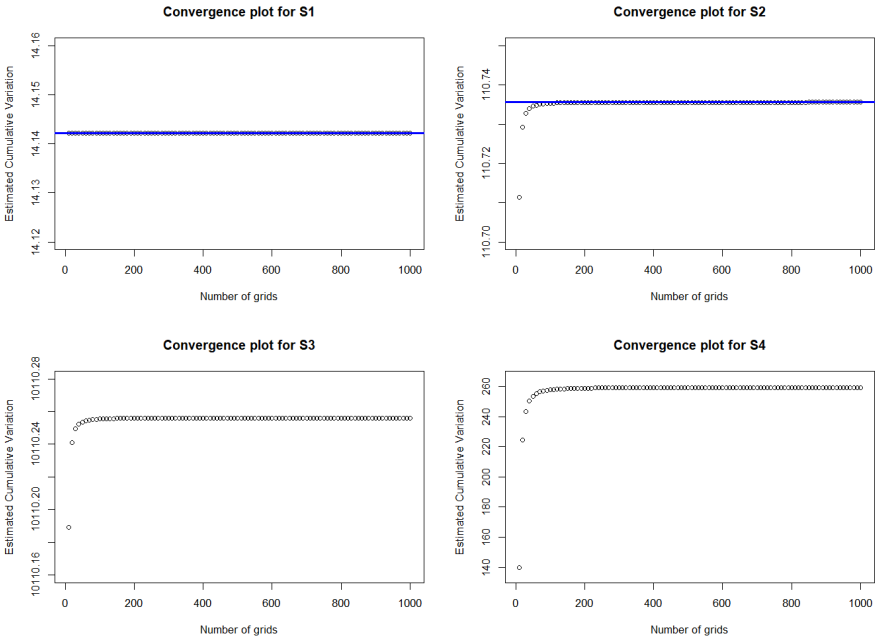

Fig. 1 The convergence plots for the estimated cumulative variation. For S1 and S2, the blue line is the closed-form exact values for the cumulative variation per Table **??**. For S3 and S4, no closed-form solutions exist for the cumulative variation.

4 Application in primary biliary cirrhosis clinical study

Figure 2 provides the four B-spline basis functions. Figure 3 demonstrate that the parameters are converged well. The pair plots in Figure 4 indicate that the Markov chain explores the parameter space well.

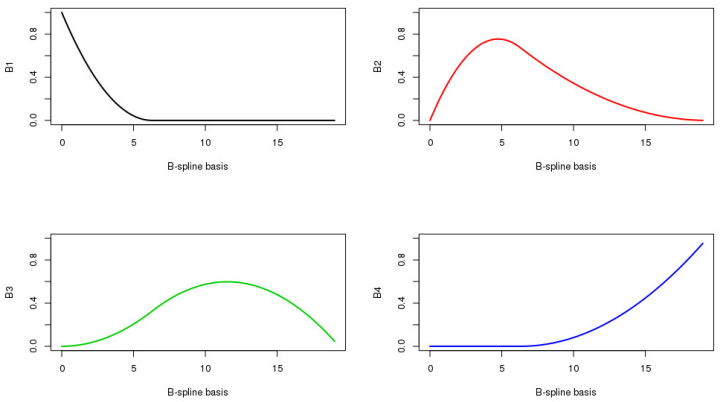

Fig. 2 Illustration of four B-spline basis.

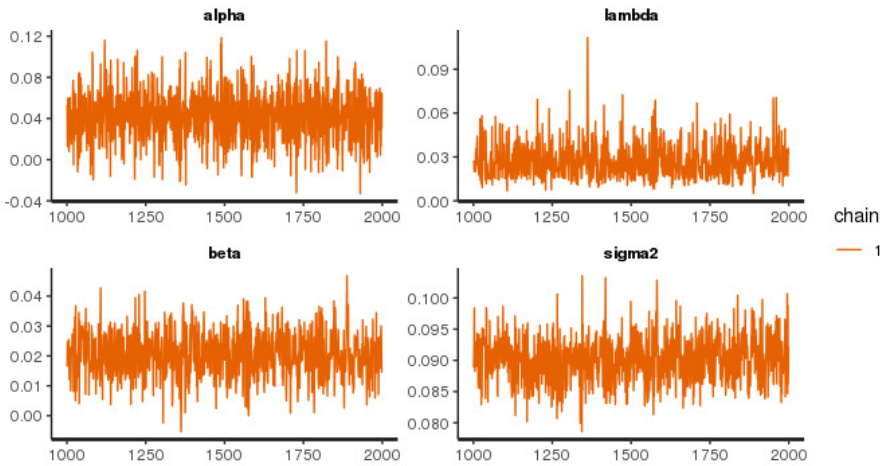

Fig. 3 Trace plots for the posterior estimation in the PBC study.

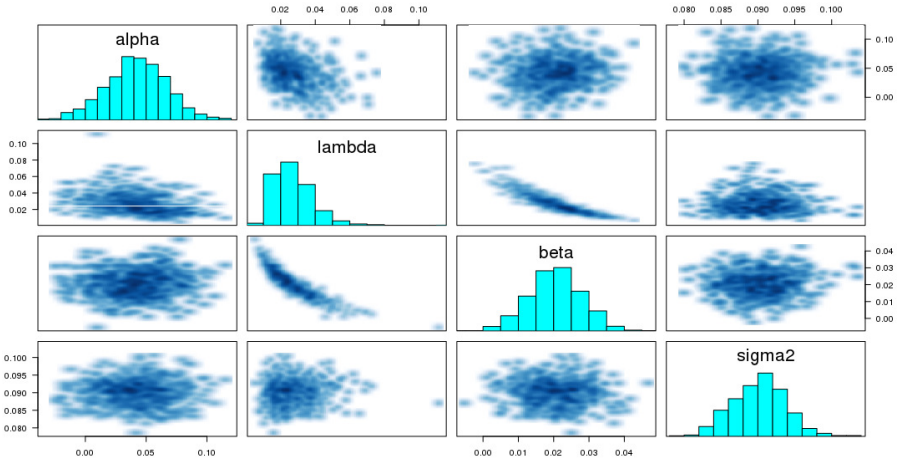

Fig. 4 Pair plots for the posterior estimation in the PBC study.

5 Application in the HIV/AIDS clinical trial

This open-label randomized clinical trial aimed to compare the efficacy and safety of two antiretroviral drugs, zalcitabine (ddC) and didanosine (ddI), in treating HIV patients for whom the standard zidovudine (AZT) therapy did not succeed (Abrams et al, 1994). A total of 467 HIV-infected patients were enrolled from Dec 1990 through Sep 1991, and 230 and 237 patients were randomized to ddI and ddC, respectively. This trial measured CD4 counts (the number of functioning CD4 cells measuring the status of a person's immune system) per cubic millimeter at baseline, 2, 6, 12, and 18 months. The analysis data set consisted of 1405 observations, 9 variables, and 188 deaths. Figure 5 suggests CD4 displays similar trends between the two treatment groups. CD4 demonstrates certain variations within and between patients. Figure 6 shows that the median survival time for ddC is about 19.1 months. For ddI, the median survival time is not attained given the time restriction of the study. Our analysis here aims to assess the association of the cumulative variation

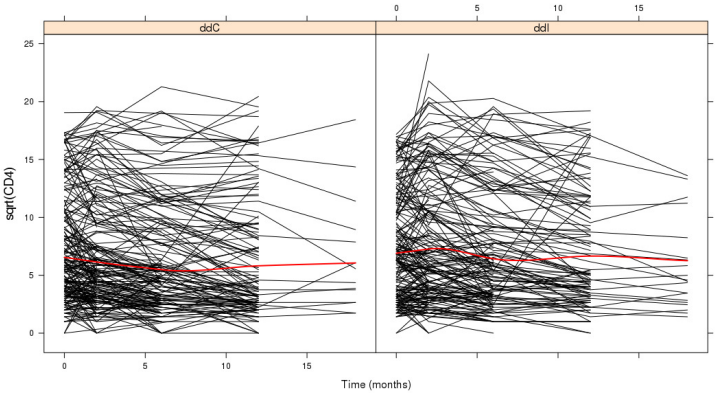

Fig. 5 Spaghetti plot for CD4 counts over time.

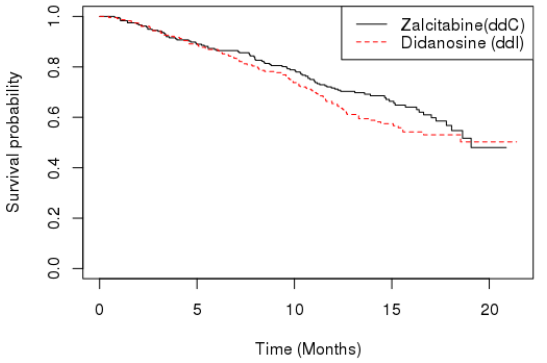

Fig. 6 Kaplan-Meier curve for survival data.

of the CD4 count on the hazard rate while adjusting for the treatment effect and measurement errors in HIV-infected patients. We formulate the following joint model to address this research question.

$$h_i(t) = \lambda \exp\left\{ x_i\beta + \alpha \int_0^t \sqrt{1 + \left\{ \sum b_{ik} B_k'(s) \right\}^2} \, \mathrm{d}s \right\} \tag{1a}$$

$$z_i(t) = Q_i(t) + \varepsilon_i(t) = \sum_{k=1}^{4} b_{ik} B_k(t) + \varepsilon_i(t) \tag{1b}$$

For the survival submodel (1a), we let (i) $\boldsymbol{x} = x$ (treatment); and (ii) $h_0(t) = \lambda$ (a constant baseline hazard). For the longitudinal submodel (1b), four B-spline basis functions $\{B_k : k = 1, \ldots, 4\}$ are considered. There are 3 knots involved, including 1 inner knot (the observed median survival time: 13.2), and 2 boundary knots [0, 26.4], where the right endpoint is the maximum survival time. The order of the B-splines is fixed at 3, leading to the degree of 2 given the formula of Degree = Order - 1. The total number of B-spline basis functions = number of knots (3) + degree (2) - 1 = 4, and the multivariate normal distribution is assumed:

$$\begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{pmatrix} \sim \text{MVN} \left[\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} \right]$$

Table 4 presents the posterior analysis results by the No-U-Turn sampler method. Successful MCMC convergence is guaranteed by the result that all $\hat{R}$ values < 1.01. The 95% credible interval of β, [-0.091, 0.478] suggests that the two treatments did not demonstrate a significant difference on mortality, and the posterior mean of the association parameter α, 0.049, along with its 95% credible interval [0.021, 0.077], indicate strong statistical evidence of an association between the $\sqrt{CD4}$ cumulative variation and the hazard. The positive posterior values suggest that a greater $\sqrt{CD4}$ cumulative variation could increase the hazard when adjusting for treatment. Figure 7 displays the fitting curves of the longitudinal data for four subjects. The estimated cumulative variation and observed survival time are denoted by $\hat{V}(t)$ and t, respectively. The vertical bar represents the time equal to t (death: red solid line; censor: black dashed line). The blue dashed line is the fitting curve for the individual

Table 4 Posterior analysis results from the No-U-Turn sampler for the HIV/AIDS clinical trial. There are 5,000 sampling iterations per chain drawn by the NUTS method, 1,000 burn-in, and 4 chains. The 95% credible interval covers [2.5%, 97.5%]. ESS represents Effective Sample Size. Rhat is the Gelman-Rubin convergence diagnostic factor. Given its symmetric property of the variance matrix Σ, only the lower triangular components are presented.

Parameters	Mean (SD)	2.5%	97.5%	ESS	Rhat
Survival Submodel					
Constant baseline hazard (λ)	0.019 (0.003)	0.014	0.026	12230	1.000
Coefficient of treatment (β)	0.196 (0.146)	-0.091	0.478	16555	1.000
Cumulative Variation					
Association coefficient (α)	0.049 (0.014)	0.021	0.077	14530	1.000
Longitudinal Submodel					
Variance of random error (σ^2)	2.279 (0.196)	2.42	3.188	1901	1.002
Mean of $\boldsymbol{b}_1$ (μ_1)	7.214 (0.221)	6.785	7.653	1458	1.005
Mean of $\boldsymbol{b}_2$ (μ_2)	6.041 (0.341)	5.366	6.702	2700	1.002
Mean of $\boldsymbol{b}_3$ (μ_3)	4.536 (0.602)	3.360	5.726	8143	1.001
Mean of $\boldsymbol{b}_4$ (μ_4)	6.559 (3.952)	-1.185	14.341	11024	1.000
Variance of $\boldsymbol{b}_1$ (σ_1^2)	20.332 (1.482)	17.585	23.376	2673	1.006
Variance of $\boldsymbol{b}_2$ (σ_2^2)	28.689 (3.291)	22.651	35.582	2985	1.002
Variance of $\boldsymbol{b}_3$ (σ_3^2)	30.955 (8.026)	17.568	48.755	1255	1.003
Variance of $\boldsymbol{b}_4$ (σ_4^2)	148.461 (153.749)	0.419	563.156	2460	1.002
Covariance (σ_{21})	21.662 (1.781)	18.379	25.303	3000	1.004
Covariance (σ_{31})	14.114 (2.628)	9.134	19.443	7848	1.001
Covariance (σ_{32})	14.985 (3.554)	7.252	21.349	1777	1.002
Covariance (σ_{41})	10.25 (13.119)	-12.318	39.003	9702	1.000
Covariance (σ_{42})	18.610 (19.644)	-10.155	63.911	3275	1.001
Covariance (σ_{43})	-26.834 (31.082)	-106.756	10.778	1932	1.003

patient, where different points represent the observed CD4 values over time. The length of the blue line up to the vertical bar is approximated cumulative variation. Although the observed survival time for Subject 121 is longer than that for Subject 447 ($t_{121} = 12.50$ vs. $t_{447} = 12.47$), the approximated cumulative variation of transformed CD4 is greater in Subject 447 ($\hat{V}_{121} = 12.68$ vs. $\hat{V}_{447} = 12.74$). This is consistent with the observed longitudinal measurements for these two patients as shown in Figure 7. This particular case especially highlights the importance of the cumulative variation as a good measurement of the variation and risks regardless of the duration of the observed time.

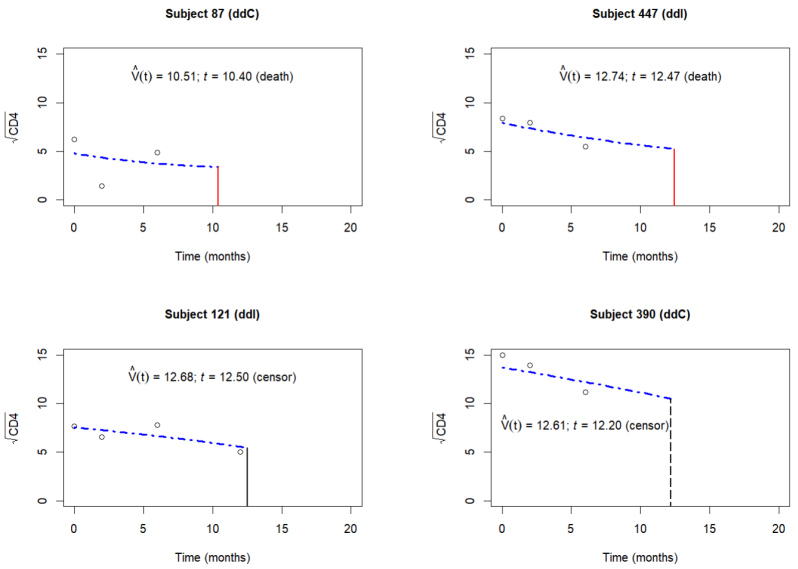

Fig. 7 CD4 fitting curves for four patients in the HIV/AIDS clinical trial.

References

Abrams DI, Goldman AI, Launer C, et al (1994) A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. New England Journal of Medicine 330(10):657–662

Hoffman J (2021) Methods in computational science. Society for Industrial and Applied Mathematics, Philadelphia, PA

Saad Y (2003) Iterative methods for sparse linear systems, 2nd edn. Society for Industrial and Applied Mathematics