Noname manuscript No.
(will be inserted by the editor)

Fitting monotone polynomials in mixed effects models

Joshua J. Bon · Kevin Murray · Berwin A. Turlach

This article has been accepted for publication in Statistics and Computing. Please cite with:

Bon, J.J., Murray, K. & Turlach, B.A. Stat Comput (2017).
https://doi.org/10.1007/s11222-017-9797-8

Received: date / Accepted: date

Abstract We provide a method for fitting monotone polynomials to data with both fixed and random effects. In pursuit of such a method, a novel approach to least squares regression is proposed for models with functional constraints. The new method is able to fit models with constrained parameter spaces that are closed and convex, and is used in conjunction with an expectation-maximisation algorithm to fit monotone polynomials with mixed effects. The resulting mixed effects models have constrained mean curves, and have the flexibility to include either unconstrained or constrained subject-specific curves. This new methodology is demonstrated on real world repeated measures data with an application from sleep science. Code to fit the methods described in this paper are available online.

Keywords monotone polynomials · monotone regression · mixed effects · random effects · shape constraints

1 Introduction

Statistical practitioners often require fitting procedures that adhere to the context of their application, namely some constraint on the regression curve to be fitted. This context represents *a priori* information and is often driven by physical necessity. For example, when estimating human growth curves the height of a person should increase over time, or in other words the fitted curve should be monotonically increasing. Ensuring the fitted regression curve adheres to a monotonicity constraint can be challenging, as standard procedures will often determine the 'best' fit in violation of such a constraint. This has prompted much research into developing methodology and algorithms for modelling monotone relationships.

Monotonic relationships occur in many real-world examples. As such there has been application driven research into monotonic fitting across several diverse areas. Examples include growth curve modelling (Zimmerman and Núñez-Antón, 2001), dose-response curves (Kelly and Rice, 1990), calibration in probabilistic classification (Zadrozny and Elkan, 2002), dependent variable transformations (Ramsay, 1998), and probability density estimation (Ramsay, 1998). Furthermore, imposing *a priori* restrictions into a regression can be very important for reasonable and practical predictions. Hence, adding a constraint can produce a more realistic result for the application at hand.

Fitting monotonic relations has been addressed by nonparametric methods including isotonic regression, monotonic spline regression, and kernel smoothing. Isotonic regression (Barlow and Brunk, 1972; Barlow et al, 1972) was devised specifically for fitting monotonic relations to ordered data and smoothing was added by Friedman and Tibshirani (1984). Monotonicity in spline estimation was introduced and refined by a number of authors (Hornung, 1978; Dierckx, 1980; Utreras, 1982, 1985), and shape constrained spline procedures have been implemented more recently (see reviews in Turlach (2005) and Hazelton and Turlach (2011) for examples). However, nonparametric smoothing can result in unrealistic flat stretches (Dette et al, 2006), and often have functional forms that are difficult for subsequent post-processing of the fitted functions, including derivative calculations (Murray et al, 2013). In particular, whilst n-degree B-splines (De Boor, 1978) are $n-1$ times differen-

J. J. Bon
Centre for Applied Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009.
E-mail: joshuajbon@gmail.com

K. Murray
School of Population and Global Health (M431), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009.

B. A. Turlach
Centre for Applied Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009.

tiable, they are still defined piecewise, complicating post-processing. Moreover, subject-specific curves from mixed effect B-splines models are only $r-2$ differentiable where r is the number of random effects per spline segment. As such modelling using monotone polynomials can be motivated by the need to detect derivative based quantities, such as inflection points.

Parametrising polynomial coefficients to ensure monotonicity with few constraints (or none at all) has been the key concept underpinning monotonic polynomial estimation. The idea was spawned by Elphinstone (1983) whilst working on non-parametric density estimation. As noted by Hawkins (1994), the Elphinstone parametrisation of a monotone polynomial is highly non-linear in its parameters. As such Hawkins developed a semidefinite quadratic programming approach to estimate monotone polynomials, which used the typical polynomial parametrisation and reduced the number of constraints by constraining only the horizontal inflection points. Murray et al (2013, 2016) explored the Elphinstone parametrisations, Hawkins' method, and developed additional parameterisations which can impose monotonicity over a (semi-)compact subset of the real line. Murray et al's work employed modern computational routines for old and new parameterisations, reported on the substantial computational limitations of previous methods and provided instructions for best-practice monotone regression in polynomial models. The package `MonoPoly` (Turlach and Murray, 2016) provides an implementation of their fitting procedures in the statistical programming language `R` (R Core Team, 2016).

The methodology presented in this paper for fitting monotone polynomials to data, deviates greatly from the established procedures of this kind (Hawkins, 1994; Murray et al, 2013, 2016), in that it neither uses a semidefinite quadratic programming approach, nor non-linear parameterisations of the polynomial, to achieve monotonicity. The main drawbacks of these approaches are; (i) the latter necessitates non-linear optimisation for fitting procedures, (ii) they both are applicable only to monotonic constraints, and (iii) neither can be easily extended to mixed effects models. By way of contrast, we adopt an orthonormal representation of the polynomial to induce a better conditioned optimisation problem for which a simple, but effective, coordinate descent algorithm can be applied. These methods address all of the aforementioned concerns.

The remainder of this paper is structured as follows: in Section 2 we introduce a novel method for constrained least squares regression when the feasible set is closed and convex, and only an oracle is given for set membership. We extend this methodology to a mixed model framework in Section 3 and describe estimation procedures through the use of an expectation–maximisation (EM) algorithm in Section 4. We provide a demonstration of these techniques on real data in Section 5, followed by some concluding remarks in Section 6. Generic code for fitting the proposed models are available as an `R` package on Github (github.com/bonStats/gcreg). An additional example on human growth data is given in Appendix F.

2 Fixed effects methodology

In order to fit monotone polynomials with mixed effects we first propose a novel method for fitting monotone polynomials in fixed effects models. This method no longer relies on non-linear reparametrisation and will enable the addition of random effects.

Since the monomial parameterisation of a polynomial in a regression model tends to create highly correlated explanatory variables its use leads to computational and inferential difficulties (Bradley and Srivastava, 1979; Narula, 1979). The latter advocated the use of discrete orthogonal polynomials generated by the data to address ill-conditioning, correlation in tests of significance for parameters, and inaccuracy, whilst providing faster computing time than that of procedures with a monomial basis. Earlier examples of the generation and use of orthogonal polynomial bases for regression appeared in Wong (1935) and Forsythe (1957), whilst Emerson (1968) proposed a general orthonormal polynomial generating procedure which relaxed assumptions of frequency and spacing of the data. We make use of Emerson's procedure for generating a discrete orthonormal polynomial basis from a dataset that results in an orthonormal design matrix, i.e. $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$.

Using the orthonormalisation procedure in Emerson (1968), we wish to estimate the coefficients, $\boldsymbol{\beta} = \left[\beta_0\, \beta_1\, \cdots\, \beta_q\right]^\top$, of a polynomial, $p(x;\boldsymbol{\beta})$, given by

$$p(x;\boldsymbol{\beta}) = \sum_{j=0}^{q} \beta_j p_j(x), \tag{1}$$

that is constructed from a set of discrete orthonormal polynomials, having the form

$$p_j(x) = \sum_{i=0}^{j} \psi_{j,i} x^i \tag{2}$$

where each $\psi_{j,i}$ is determined by the design of the explanatory variable. Restricting the polynomial to be monotonically increasing or decreasing results in a closed, convex parameter space for $\boldsymbol{\beta}$, which we denote as $\Omega_{\boldsymbol{\beta}}$. Hence the least squares problem may be described as

$$\min_{\boldsymbol{\beta}} \{\mathrm{RSS}(\boldsymbol{\beta})\} \text{ s.t. } \boldsymbol{\beta} \in \Omega_{\boldsymbol{\beta}} \tag{3}$$

where $\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ and the response variable $\boldsymbol{Y}$ is a column vector of length n. The i^{th} row of the design matrix $\boldsymbol{X}$ corresponds to the observation x_i evaluated at

each of the orthonormal polynomials, specifically the i^{th} row is $\left[p_0(x_i)\ p_1(x_i)\ \cdots\ p_q(x_i)\right]$.

2.1 Constrained orthogonal least squares regression

To describe the optimisation of (3) under monotonicity constraints, we proceed, without loss of generality, with the case of monotone increasing polynomials over the real line or over a (semi-)compact set, which we denote by $S \subseteq \mathbb{R}$. The constrained parameter space is

$$\Omega_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} \text{ s.t. } p'(x;\boldsymbol{\beta}) \geq 0, \forall x \in S\}. \tag{4}$$

We propose a coordinate descent approach to (3) under the parameter space in (4) since this optimisation procedure adapts well to feasible regions that are closed and convex (Cassioli et al, 2013). This optimisation approach solves a multivariate optimisation problem iteratively by reducing the problem to a succession of univariate optimisation problems, optimising in each of these sub-problems over one parameter, whilst holding all other parameters fixed. This enables the boundary of the feasible region to easily be found in each one dimensional sub-problem. Furthermore, using an orthonormal design matrix speeds up convergence since the derivatives of the RSS with respect to the β_is are functionally independent of each other, as the derivative of the RSS with respect to a single coefficient can be written as

$$\frac{\partial \text{RSS}}{\partial \beta_i} = 2\left(\beta_i - \boldsymbol{X}_i^\top \boldsymbol{Y}\right) = 2\left(\beta_i - \hat{\beta}_i^U\right) \tag{5}$$

where $\hat{\beta}_i^U = \boldsymbol{X}_i^\top \boldsymbol{Y}$ is the unconstrained optimum. This motivates the use of a line search algorithm, and hence coordinate descent, for optimising RSS with respect to a β_i. The algorithm searches between the current β_i and $\hat{\beta}_i^U$ for the minimal RSS subject to the monotonicity constraint. Values closer to $\hat{\beta}_i^U$ will have derivative closer to zero and therefore a smaller RSS value, so the RSS does not have to be evaluated. To describe the constrained optimisation we first define the line search in Algorithm 1. This algorithm is used to find a coordinate's optimal value within the permissible region. Algorithm 1 is described recursively for the sake of brevity, but the implementation need not be. Note that $\boldsymbol{\beta}_{(i)}$ denotes all elements of $\boldsymbol{\beta}$ except the i^{th} element.

For monotone fitting, the oracle function in Algorithm 1 is

$$I(\boldsymbol{\beta}, \boldsymbol{\Psi}^\top) = \begin{cases} 1 & \text{if } \boldsymbol{\beta} \in \Omega_{\boldsymbol{\beta}} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

indicating if a particular $\boldsymbol{\beta}$ is monotone increasing over S. The auxiliary variable, in this case $\boldsymbol{\Psi}^\top$, converts the orthonormal basis to the monomial basis to calculate the derivative of $p(x;\boldsymbol{\beta})$ and test for monotonicity.

Algorithm 1 Constrained line search.

Require: $I(\boldsymbol{\beta}, A)$ is an indicator function (passed as an argument) for $p(\boldsymbol{\beta})$ belonging to some closed convex region. It may require an auxiliary variable, A. The vector $\boldsymbol{\beta}^{\text{curr}}$ contains the starting values of $\boldsymbol{\beta}$, and is in the permissible region. The end point of the line search is β_i^{aim}.

procedure LINESEARCH($\boldsymbol{\beta}^{\text{curr}}, \beta_i^{\text{aim}}, i, I(\boldsymbol{\beta}, A)$)
  $\boldsymbol{\beta}'_{(i)} \leftarrow \boldsymbol{\beta}''_{(i)} \leftarrow \boldsymbol{\beta}^{\text{curr}}_{(i)}$
  $\beta_i' \leftarrow (\beta_i^{\text{curr}} + \beta_i^{\text{aim}})/2$
  $\beta_i'' \leftarrow \beta_i^{\text{aim}}$
  if $I(\boldsymbol{\beta}'', A) = 1$ **then**
    return $\boldsymbol{\beta}''$
  else if $I(\boldsymbol{\beta}', A) = 1$ **then**
    return LINESEARCH($\boldsymbol{\beta}', \beta_i'', i, I(\boldsymbol{\beta}, A)$)
  else
    return LINESEARCH($\boldsymbol{\beta}^{\text{curr}}, \beta_i', i, I(\boldsymbol{\beta}, A)$)

Algorithm 2 describes the operation of the coordinate descent routine. It takes the data $\boldsymbol{Y}$ and $\boldsymbol{X}$ as inputs, the matrix $\boldsymbol{\Psi}^\top$, and $d = q + 1$ where q is the maximal degree of the polynomial. Note that if monotonicity is required over the entire real line, q must be an odd number. Initialising the coordinate descent algorithm with $\boldsymbol{\beta}^{[0]}$ such that $p(x;\boldsymbol{\beta}^{[0]})$ is monotonic allows the procedure to maintain this monotonicity. A simple initial point for the polynomial is a straight line with slope equal to one. Hence the vector $\boldsymbol{\beta}$ can be given an initial value of $\boldsymbol{\beta}^{\text{init}} = (\boldsymbol{\Psi}^\top)^{-1}\left[0\ 1\ 0\ \cdots\ 0\right]^\top$ where this vector will have length d.

Algorithm 2 Monotone OLS fitting via coordinate descent and line search.

Require: $\boldsymbol{Y}$ is the vector of observations, $\boldsymbol{X}$ is orthonormal polynomial design matrix, with d columns. $\boldsymbol{\Psi}^\top$ is the auxiliary variable for conversion between orthonormal and monomial bases. $\boldsymbol{\beta}^{\text{init}}$ is initial value (in orthonormal basis). T is the maximum iterations and $\epsilon > 0$ is the convergence criteria.

procedure FIT.COLS($\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}^\top, \boldsymbol{\beta}^{\text{init}}, d, T, \epsilon$)
  $\hat{\boldsymbol{\beta}}^U \leftarrow \boldsymbol{X}^\top \boldsymbol{Y}$
  $\boldsymbol{\beta}^{[0]} \leftarrow \boldsymbol{\beta}^{\text{init}}$
  $\beta_0^{[0]} \leftarrow \hat{\beta}_0^U$ ▷ update intercept.
  for $t = 1$ to T **do**
    $i \leftarrow t \mod d$ ▷ i, index of current element.
    $\boldsymbol{\beta}_{(i)}^{[t]} \leftarrow \boldsymbol{\beta}_{(i)}^{[t-1]}$
    $\beta_i^{[t]} \leftarrow$ LINESEARCH($\boldsymbol{\beta}^{[t-1]}, \hat{\beta}_i^U, i, I(\boldsymbol{\beta}, \boldsymbol{\Psi}^\top)$)
    if $t > d$ and $\|\boldsymbol{\beta}^{[t]} - \boldsymbol{\beta}^{[t-d]}\| < \epsilon$ **then return** $\boldsymbol{\beta}^{[t]}$
return "did not converge"

Finally, we note that although we use an orthonormal design matrix, the dependence between the parameters is not completely removed since the constraints on each coordinate of $\boldsymbol{\beta}$ vary based on the values of the other coordinates. This necessitates that the coordinate descent algorithm optimises over each coordinate at least once, but more typically several times.

Henceforth we refer to Algorithm 2 as *constrained orthogonal least squares* (COLS). It is a flexible procedure

that can easily be used with other kinds of constraints so long as the feasible space is closed and convex. For other constraints, the main change to the COLS procedure would be the removal of line 4, and the provision of an appropriate initial vector and oracle function.

In terms of computational cost, the COLS algorithm has some benefits and drawbacks compared to other iterative algorithms. Using an orthonormal design matrix removes one of the two dimensions of dependence between coordinates – dependence is only induced by the constraint. However, it does operate coordinate-wise which is slow compared to methods that act on multiple coordinates simultaneously. As seen in Algorithm 2, once $\hat{\boldsymbol{\beta}}^U = \boldsymbol{X}^\top \boldsymbol{Y}$ is calculated the algorithm run-time does not depend on data size, but does depend on the number of parameters.

The COLS algorithm can be used directly in mixed effects model estimation by replacing line 2 of Algorithm 2 with the unconstrained minimum of the objective function (which can be found using a Newton-Raphson algorithm or otherwise). As an alternative, a penalised version of the COLS algorithm is given in Appendix A. This method, *penalised constrained orthogonal least squares* (pCOLS), avoids directly finding the unconstrained minimum by taking steps based on the derivative of the objective function. The pCOLS algorithm operates in a very similar manner to COLS, hence a close understanding is not a prerequisite for the rest of the paper.

3 Mixed effects methodology

In this section we consider mixed effects models to facilitate the modelling of non-independence of observations in the data. In particular, we seek a monotone estimation procedure which extends to random effects on repeated measures and longitudinal data. For these types of data and models, the random effects can be interpreted as modelling deviations of subject-specific curves from a population curve. In this way the random effects also incorporate covariance of observations belonging to the same individual. We introduce random effects to define a mixed effects model using the conditional distribution of the observations as

$$\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}} = \boldsymbol{U} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{ZU}, \boldsymbol{R}) \qquad \boldsymbol{\mathcal{U}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}) \tag{7}$$

with a normal distribution for the random effects. In this specification there are a total of n observations from g groups (individuals) that are modelled by d covariates and r random effects. The column vector $\boldsymbol{\mathcal{Y}}$ is of length n, $\boldsymbol{X}$ is an $n \times d$ fixed design matrix, $\boldsymbol{\beta}$ is a column vector of fixed effects parameters with length d, $\boldsymbol{Z}$ is the random effects design matrix with size $n \times rg$, and $\boldsymbol{\mathcal{U}}$ is the random effects column vector with length rg. The variances of the random effects and measurement error, $\boldsymbol{G}$ and $\boldsymbol{R}$, are positive definite matrices with size $rg \times rg$ and $n \times n$ respectively. We can also write a joint normal distribution for $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ by

$$\begin{bmatrix} \boldsymbol{\mathcal{Y}} \\ \boldsymbol{\mathcal{U}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X\beta} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{R} + \boldsymbol{ZGZ}^\top & \boldsymbol{ZG} \\ \boldsymbol{GZ}^\top & \boldsymbol{G} \end{bmatrix} \right). \tag{8}$$

From this joint distribution the derivation of the marginal distribution of $\boldsymbol{\mathcal{Y}}$ is trivial and the conditional distribution $\boldsymbol{\mathcal{U}} \mid \boldsymbol{\mathcal{Y}}$ can be derived as

$$\begin{aligned} &\boldsymbol{\mathcal{U}} \mid \boldsymbol{\mathcal{Y}} = \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}, \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}) \text{ where} \\ &\quad \boldsymbol{M}_{\boldsymbol{\mathcal{U}}} = \boldsymbol{GZ}^\top (\boldsymbol{ZGZ}^\top + \boldsymbol{R})^{-1} (\boldsymbol{Y} - \boldsymbol{X\beta}) \text{ and} \\ &\quad \boldsymbol{V}_{\boldsymbol{\mathcal{U}}} = \boldsymbol{G} - \boldsymbol{GZ}^\top (\boldsymbol{ZGZ}^\top + \boldsymbol{R})^{-1} \boldsymbol{ZG}. \end{aligned} \tag{9}$$

As we will see below, the conditional distribution $\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}$ in (7) is useful for partitioning the likelihood, whilst $\boldsymbol{\mathcal{U}} \mid \boldsymbol{\mathcal{Y}}$ in (9) is required for the EM algorithm used in Section 4.

As we model subject-specific curves our model does not use crossed random effects between subjects and within each subject no further nesting of random effects is considered. This imposes further structure on $\boldsymbol{G}$ and $\boldsymbol{Z}$ as both matrices can be decomposed into block-diagonal matrices. The matrix $\boldsymbol{Z}$ can be written as $\boldsymbol{Z} = \operatorname{diag}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_g\right)$ where $\boldsymbol{Z}_i$ is an $n_i \times r$ matrix, a subset of the covariates in $\boldsymbol{X}$ for the individual, and n_i is the number of observations for individual i such that $\sum_{j=1}^{g} n_j = n$. The variance-covariance matrix $\boldsymbol{G}$ consists of a subject's covariance matrix $\boldsymbol{H}$ which is equal across subjects. Therefore $\boldsymbol{G}$ can be written as

$$\boldsymbol{G} = \boldsymbol{I}_g \otimes \boldsymbol{H}$$

where $\boldsymbol{H}$ has size $r \times r$. The covariance matrices $\boldsymbol{R}$ and $\boldsymbol{H}$ have parameter sets $\boldsymbol{\phi}_{\boldsymbol{R}}$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$ respectively, hence $\boldsymbol{\phi}_{\boldsymbol{H}}$ also parameterises $\boldsymbol{G}$. Note that the size of vectors $\boldsymbol{\phi}_{\boldsymbol{R}}$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$ will depend on the structure of $\boldsymbol{R}$ and $\boldsymbol{H}$, respectively.

The model in (7) has a pseudo-likelihood with latent variable $\boldsymbol{\mathcal{U}}$ that can be partitioned as

$$\begin{aligned} &L_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}} \mid \boldsymbol{\mathcal{Y}} = \boldsymbol{Y}, \boldsymbol{\mathcal{U}}) = \\ &\qquad L_{\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}}(\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}} \mid \boldsymbol{Y}, \boldsymbol{\mathcal{U}}) \times L_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}} \mid \boldsymbol{\mathcal{U}}). \end{aligned} \tag{10}$$

The first component of the pseudo-likelihood in (10) under (7) is

$$L_{\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}}(\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}} \mid \boldsymbol{Y}, \boldsymbol{\mathcal{U}}) = \left((2\pi)^n |\boldsymbol{R}|\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1} \boldsymbol{\mathcal{E}} \right\} \tag{11}$$

where the stochasticity of $\boldsymbol{\mathcal{E}} = \boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Z\mathcal{U}}$ depends only on $\boldsymbol{\mathcal{U}}$ now that we condition on observing $\boldsymbol{\mathcal{Y}} = \boldsymbol{Y}$. The second component of the pseudo-likelihood is

$$L_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}} | \boldsymbol{\mathcal{U}}) = \left((2\pi)^{rg} |\boldsymbol{G}|\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1} \boldsymbol{\mathcal{U}} \right\}. \tag{12}$$

Taking the logarithm of Equation (10) with components (11) and (12) we show that the pseudo-log-likelihood is

$$\begin{aligned} & l_{\mathcal{Y},\mathcal{U}}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}) \\ & \quad = l_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}) + l_{\mathcal{U}}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{\mathcal{U}}) \\ & \quad = -\frac{1}{2}\left[c + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1}\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1}\boldsymbol{\mathcal{U}}\right] \end{aligned} \tag{13}$$

where c is the logarithm of the normalising constant in the pseudo-likelihood. Model (7) can be estimated using an EM Algorithm to maximise the pseudo-likelihood in (10). In the maximisation step it is possible to use constrained optimisation to ensure the mean curve of the fitted model is monotone, which we discuss later. However, model (7) is misspecified if each individual's curve is to be monotone. Direct optimisation of the likelihood in (10) with constraints on the random effects is possible, but would not be equivalent to maximising the likelihood resulting from postulating a suitable constrained distribution for the random effects.

Specifically, to constrain each individual's curve to be monotone, the underlying random effects distribution should be changed. In keeping with standard mixed effects models a truncated multivariate normal distribution is used, and the truncation is chosen so that the random effects are restricted to a space that retains the monotonicity of all individuals' curves. We rewrite model (7) with the random effects truncated as

$$\begin{aligned} \boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}_T = \boldsymbol{U} &\sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{ZU}, \boldsymbol{R}) \\ \boldsymbol{\mathcal{U}}_T &\sim \mathcal{N}_{T(\boldsymbol{\beta})}(\boldsymbol{0}, \boldsymbol{G}) \end{aligned} \tag{14}$$

where $\boldsymbol{\mathcal{U}}_T$ is a truncated multivariate normal random variable, as denoted by $\mathcal{N}_{T(\boldsymbol{\beta})}$, with support constricted to the region $T(\boldsymbol{\beta}) \subset \mathbb{R}^{rg}$ written to emphasise the dependence of the region T on $\boldsymbol{\beta}$. Note that $T(\boldsymbol{\beta})$ is also dependent on the variance parameters of the random effects in $\boldsymbol{\phi_H}$. The mean of $\boldsymbol{0}$ and variance matrix $\boldsymbol{G}$ are the moments of the underlying normal distribution, before truncation.

Using the definition in (14), we write the joint likelihood of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}_T$ as

$$\begin{aligned} & L_{\mathcal{Y},\mathcal{U}_T}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}_T) = \\ & \qquad L_{\mathcal{Y}|\mathcal{U}_T}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}_T) \times L_{\mathcal{U}_T}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{\mathcal{U}}_T) \end{aligned} \tag{15}$$

without having to derive the functional form of the unpartitioned joint distribution. In Appendix B we show that from (15) the joint density of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}_T$ is a truncated normal distribution. Taking the logarithm of Equation (15) gives

$$\begin{aligned} & l_{\mathcal{Y},\mathcal{U}_T}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}) = \\ & \quad -\frac{1}{2}\left[c + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1}\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1}\boldsymbol{\mathcal{U}}\right] \\ & \quad -\log\left(\int_{T(\boldsymbol{\beta})} ((2\pi)^{rg}|\boldsymbol{G}|)^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{U}^\top \boldsymbol{G}^{-1}\boldsymbol{U}\right\} \mathrm{d}\boldsymbol{U}\right). \end{aligned} \tag{16}$$

Equation (16) can be rewritten as a function of the unconstrained likelihood (13) and the normalising term $\eta(\boldsymbol{\beta}) = \log(P(\boldsymbol{\beta}))$, giving

$$l_{\mathcal{Y},\mathcal{U}_T}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}) = l_{\mathcal{Y},\mathcal{U}}(\boldsymbol{\beta},\boldsymbol{\phi_R},\boldsymbol{\phi_H} \mid \boldsymbol{Y},\boldsymbol{\mathcal{U}}) - \eta(\boldsymbol{\beta}) \tag{17}$$

where $P(\boldsymbol{\beta}) = \int_{T(\boldsymbol{\beta})} ((2\pi)^{rg}|\boldsymbol{G}|)^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{G}^{-1}\boldsymbol{W}\right\} \mathrm{d}\boldsymbol{W}$. Note that if the individuals' curves are not constrained then $T(\boldsymbol{\beta}) = \mathbb{R}^{rg}$, $P(\boldsymbol{\beta}) = 1$ and hence $\eta(\boldsymbol{\beta}) = 0$. Therefore the likelihood in (16) is correct when the individuals' curves are constrained, but also when they are not. This will simplify the discussion in Section 4.

An analogous result to (9) is needed for the conditional distribution $\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}$. This will allow the calculation of the mean and variance from this distribution in the expectation step of the EM algorithm. Appendix B shows the conditional random variable $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}$ truncated to the region $T(\boldsymbol{\beta})$ is the desired distribution for $\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}$. Once the mean and variance of the equivalent nontruncated distribution is found, the density (and hence moments) for the constrained distribution are computable. The difficulty in calculating these truncated moments then becomes evaluating high-dimensional integrals over $T(\boldsymbol{\beta})$. Appendix B contains more details regarding the conditional truncated distribution of the random effects.

For simple cases of $T(\boldsymbol{\beta})$ evaluating these integrals and hence deriving the moments can be done analytically. Tallis (1961) derives the moment generating function of point-truncated multivariate normal distributions for example. For $r = 1$ the truncation is not needed since a random intercept will not affect monotonicity. For $r = 2$, $T(\boldsymbol{\beta})$ is defined by one-sided point truncation, and the R package `tmvtnorm` (Wilhelm and Manjunath, 2015) can be used to find the mean and variance of $\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}$ based on the truncation of $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}$. When $r > 2$, Monte Carlo simulation is needed to evaluate both the expectation of the conditional $\boldsymbol{\mathcal{U}}$ and the normalising term, $\eta(\boldsymbol{\beta})$. The general methodology is presented next in Section 4 whilst the various scenarios, outlined above, are discussed in Appendix C and D.

4 Mixed effects estimation

An EM algorithm is implemented to handle the additional complexity of random effects, that may or may not be truncated, in the monotonic polynomial models. The usefulness of the EM algorithm for mixed models with repeated-measures is demonstrated in Laird et al (1987) and Lindstrom and Bates (1988), while the benefits of the algorithm for flexible random effects are shown in Booth and Hobert (1999) and Chen et al (2002). Additionally, even with the difficult region of monotonicity for individual curves when $r > 2$, it is feasible to employ the EM algorithm with Monte Carlo expectations (an MCEM algorithm, see for example

Levine and Casella (2001)). Other optimisers, such as the Newton-Raphson method, do not afford the same level of flexibility in random effect specification. The downside of EM algorithms for mixed effects models is their relative speed compared to Newton-Raphson methods (Lindstrom and Bates, 1988). In general, we have found that warm starting the EM algorithm using constrained fixed effects estimates from the COLS algorithm results in very reasonable run times. In following subsections we outline the general expectation and maximisations steps in the algorithm for our constrained mixed effects model. Technical details for specific instances of these models are detailed in Appendix C and D.

4.1 Expectation step

Denote the set of parameters governing the likelihood as $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}, \boldsymbol{\phi}_{\boldsymbol{H}}\}$. The vector of parameters, $\boldsymbol{\Theta}$, at the t^{th} iterate is $\boldsymbol{\Theta}^{[t]}$. For $t = 0$, we specify some initial parameters $\boldsymbol{\Theta}^{[0]}$, and note there are two random quadratic forms in the log-likelihood, $\boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1} \boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1} \boldsymbol{\mathcal{U}}$, whose conditional expectation is needed for all EM algorithms in the remainder of this section.

The conditional expectation of $\boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1} \boldsymbol{\mathcal{U}}$ is

$$\mathbb{E}_{\boldsymbol{\mathcal{U}}}\left(\boldsymbol{\mathcal{U}}^\top \boldsymbol{G}^{-1} \boldsymbol{\mathcal{U}} \mid \boldsymbol{Y}, \boldsymbol{\Theta}^{[t]}\right) = \operatorname{tr}\left[\boldsymbol{G}^{-1} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] + \left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{G}^{-1} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$$

where $\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} = \mathbb{E}_{\boldsymbol{\mathcal{U}}}\left(\boldsymbol{\mathcal{U}} \mid \boldsymbol{Y}, \boldsymbol{\Theta}^{[t]}\right)$, $\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} = \mathbb{V}_{\boldsymbol{\mathcal{U}}}\left(\boldsymbol{\mathcal{U}} \mid \boldsymbol{Y}, \boldsymbol{\Theta}^{[t]}\right)$, and $\operatorname{tr}(\cdot)$ is the trace operator. The conditional expectation of the quadratic form $\boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1} \boldsymbol{\mathcal{E}}$ is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mathcal{U}}}\left(\boldsymbol{\mathcal{E}}^\top \boldsymbol{R}^{-1} \boldsymbol{\mathcal{E}} \mid \boldsymbol{Y}, \boldsymbol{\Theta}^{[t]}\right) &= \operatorname{tr}\left[\boldsymbol{R}^{-1} \boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top\right] \\ &\quad + \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{R}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right).\end{aligned}$$

The conditional expectation of the general pseudo-log-likelihood in (16), denoted by $q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{[t]}\right)$, can then be calculated using the expectations of the quadratic forms to give

$$\begin{aligned}q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{[t]}\right) &= \mathbb{E}_{\boldsymbol{\mathcal{U}}}\left(l_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}\left(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{\mathcal{U}}\right) \mid \boldsymbol{Y}, \boldsymbol{\Theta}^{[t]}\right) \\ &= -\frac{1}{2}\Big[c + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \operatorname{tr}\left[\boldsymbol{R}^{-1} \boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top\right] \\ &\quad + \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{R}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right) \\ &\quad + \operatorname{tr}\left[\boldsymbol{G}^{-1} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] + \left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{G}^{-1} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\Big] - \eta(\boldsymbol{\beta}).\end{aligned} \tag{18}$$

If we can calculate $\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$ and $\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}$ then we are just left with the maximisation step. When we have the unconstrained random effects the above moments are easily computed from Equation (9) for any reasonable r. However, constraining the random effects (for monotonic individuals' curves) requires a specific truncated version of this distribution which is difficult to calculate analytically over the support $T(\boldsymbol{\beta})$. A random intercept ($r = 1$) is an exception, it will have no bearing on monotonicity so the moments of a normal distribution may be used. With a random intercept and slope ($r = 2$) the calculation of $\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$ and $\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}$ is possible with current methods for calculating box-truncated multivariate normal moments. For $r > 2$, the truncation is no longer linear and simulation is needed. We use rejection and Markov chain Monte Carlo sampling for this purpose. The technical details for incorporating a random intercept, random slope, and higher degree random effects are discussed in Appendix C.

4.2 Maximisation step

Next, we describe the general maximisation step on the EM algorithm. For simplicity we drop the functional notation of the expectation and express q, as in (18), on the deviance scale so that $q_{\text{dev}}^{[t]} = -2 \times q\left(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{[t]}\right)$. Since we have negated q we aim to minimise $q_{\text{dev}}^{[t]}$ in the M-step. To minimise $q_{\text{dev}}^{[t]}$ in a given step we must optimise over each parameter in $\boldsymbol{\Theta}$. The general optimisation strategy is developed in this section, whilst the details for particular classes of models are given in Appendices D.1 to D.3.

The deviance-scale expected pseudo-log-likelihood, $q_{\text{dev}}^{[t]}$, can be written as

$$\begin{aligned}q_{\text{dev}}^{[t]} &= \log|\boldsymbol{R}| + \operatorname{tr}\left[\boldsymbol{R}^{-1} \boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top\right] \\ &\quad + \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{R}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right) \\ &\quad + \log|\boldsymbol{G}| + \operatorname{tr}\left[\boldsymbol{G}^{-1} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] + \left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{G}^{-1} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} \\ &\quad + 2\eta(\boldsymbol{\beta})\end{aligned} \tag{19}$$

after dropping all constant terms in (18). The partial derivatives of $q_{\text{dev}}^{[t]}$ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\phi}_{\boldsymbol{R},i}$, and $\boldsymbol{\phi}_{\boldsymbol{H},i}$ are

$$\frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{W}^{[t]} + 2\frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{20}$$

$$\begin{aligned}\frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\phi}_{\boldsymbol{R},i}} &= \operatorname{tr}\left[\boldsymbol{R}^{-1} \tfrac{\partial \boldsymbol{R}}{\partial \boldsymbol{\phi}_{\boldsymbol{R},i}}\right] - \operatorname{tr}\left[\boldsymbol{R}^{-1} \tfrac{\partial \boldsymbol{R}}{\partial \boldsymbol{\phi}_{\boldsymbol{R},i}} \boldsymbol{R}^{-1} \boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top\right] \\ &\quad - \left(\boldsymbol{W}^{[t]}\right)^\top \boldsymbol{R}^{-1} \tfrac{\partial \boldsymbol{R}}{\partial \boldsymbol{\phi}_{\boldsymbol{R},i}} \boldsymbol{R}^{-1} \boldsymbol{W}^{[t]}\end{aligned} \tag{21}$$

$$\begin{aligned}\frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}} &= \operatorname{tr}\left[\boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}\right] - \operatorname{tr}\left[\boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}} \boldsymbol{G}^{-1} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] \\ &\quad - \left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^\top \boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}} \boldsymbol{G}^{-1} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} + 2\frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}\end{aligned} \tag{22}$$

where $\boldsymbol{W}^{[t]} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$ are the "working" residuals and the subscript i denotes the i^{th} component of the respective parameter vectors. In this general form, the optimisation of $q_{\text{dev}}^{[t]}$ over $\boldsymbol{\beta}, \boldsymbol{\phi}_{\boldsymbol{R}}$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$ are interrelated. More precisely when $\eta(\boldsymbol{\beta}) \neq 0$ we can see from the derivatives (20) to (22) that the optimisation of

- $\boldsymbol{\beta}$ depends on $\boldsymbol{\phi}_{\boldsymbol{R}}$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$,

– $\boldsymbol{\phi_R}$ depends on $\boldsymbol{\beta}$, and
– $\boldsymbol{\phi_H}$ depends on $\boldsymbol{\beta}$

in a single M-step[1]. Additionally, closed form solutions may not exist for the equations obtained by setting (20) to (22) to zero. A natural approach to this optimisation step uses an iteratively reweighted least squares algorithms (see, among others Holland and Welsch, 1977) that switches between coefficient and variance estimates until convergence within a single M-step.

We apply the pCOLS method to the mixed effects model estimation by noting that $\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$ can be written as $\boldsymbol{Y}_*^{[t]} - \boldsymbol{X\beta}$ where $\boldsymbol{Y}_*^{[t]} = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$. Hence in the pCOLS estimation procedure we replace $\boldsymbol{Y}$ with $\boldsymbol{Y}_*^{[t]}$. The derivative to be used is

$$\frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{R}^{-1}\left(\boldsymbol{Y}_*^{[t]} - \boldsymbol{X\beta}\right) + 2\,\frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\,. \tag{23}$$

When the random effects are not constrained, $\eta(\boldsymbol{\beta}) = 0$, the optimisation of $\boldsymbol{\beta}$ can employ COLS instead of pCOLS.

In a single M-step, when $\eta(\boldsymbol{\beta}) = 0$, we see that the optimisation of

– $\boldsymbol{\beta}$ depends on $\boldsymbol{\phi_R}$,
– $\boldsymbol{\phi_R}$ depends on $\boldsymbol{\beta}$, and
– $\boldsymbol{\phi_H}$ has no dependence on the other parameter sets.

In this case an iteratively reweighed scheme can be used for $\boldsymbol{\beta}$ and $\boldsymbol{\phi_R}$ and when convergence is reached the parameters of $\boldsymbol{\phi_H}$ can be optimised for a single M-step. Appendix D explores further simplifications to the maximisation step so that useful (and analytically tractable) results for the M-step can be derived. Monte Carlo approximations to $\eta(\boldsymbol{\beta})$ and its partial derivatives are also proposed for use when $r > 2$.

5 Example on sleep study data

We use a subset of data from a sleep deprivation study (Belenky et al, 2003) to demonstrate our methodology for several types of constrained polynomial models. The technical details of the fitting implementation are described in Appendices C and D. The data has observations from 18 individuals who, over a 10 day period, are subjected to limited sleep (three hours per night). Individual reaction times are recorded for a series of tests each day and the average daily reaction time is recorded and will be used as the response variable. Participants are said to be in recovery for the final three observations and consequently sleep eight hours. These data are available in the `lme4` package (Bates et al, 2015) in R.

[1] After multiple M-steps there is dependence through the use of updated parameters in the E-step of the EM algorithm.

In such studies it seems reasonable to assume that consecutive evenings of insufficient sleep will lead to increasingly poor reaction times. Hence, a monotonically increasing mean curve, and possibly monotonically increasing subject-specific curves may be beneficial for modelling purposes. We demonstrate our methodology by initially fitting degree eight polynomials to the data with various random effects scenarios. Figure 1 shows the estimated mean curves for degree eight polynomials with either two or three random effects and different constraint structures fitted to the data. In the key, q indicates the degree, r the number of random effects, and an asterisk (*) indicates which of the components have been constrained to ensure monotonicity over the whole 10 days of the study. Note that the unconstrained mean fitted curve with two and three random effects coincide.

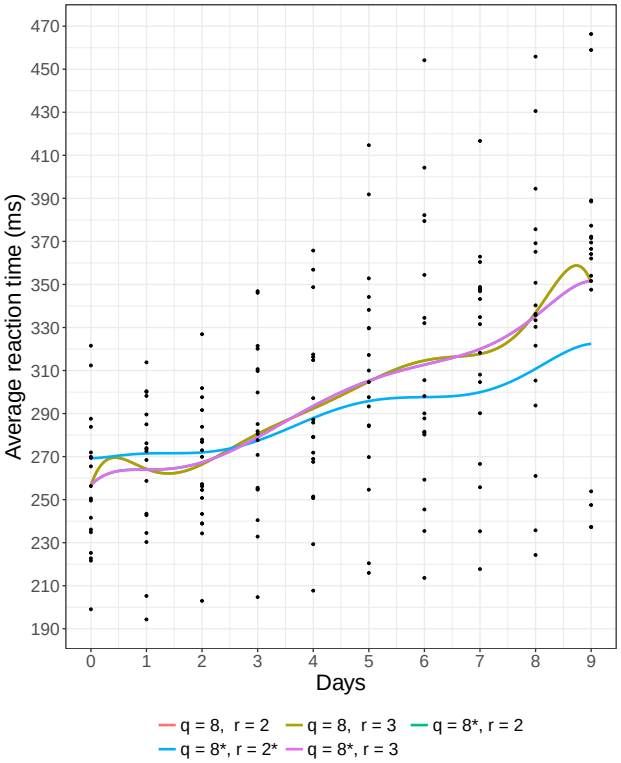

Fig. 1 Mean fitted curves for degree $q = 8$ polynomials with varying numbers of random effects r under different constraint structures (* indicating which components, fixed or random, are constrained to ensure monotonicity). The unconstrained mean curves coincide, as do the constrained mean curves without random effect constraints.

The constrained mean curve with unconstrained subject-specific curves ($q = 8^*, r = 2, 3$) lie in the a similar region to the completely unconstrained curves ($q = 8, r = 2, 3$), more so in days 0 to 6 than 7 to 9. The completely unconstrained curves seem to be overfitting between day 8 and 9, which can be seen in Figure 2 since the data does not support

a turning point between these days. Constraining the mean curves to be monotonic does have the effect of decreasing over-fitting in this instance, without having to use a lower degree polynomial.

The constrained subject-specific mean curve ($q = 8^*, r = 2^*$) lies well below the other curves. The best fit under this dually constrained model lies on the boundary of the parameter space – observe the horizontal inflection point (within numerical precision) between day 1 and 2. Since each subject-specific curve is the result of a linear movement in the intercept and the gradient, each individuals' curve must have a gradient greater than that of the mean curve, over the constrained region. Hence, the y-intercept of the mean curve compensates by reducing in magnitude to respect the constraint on the random slope and provide the best fit under these conditions.

Figure 2 shows that for the most part, the constrained subject-specific model fits the data poorly when there is strong evidence of non-monotonicity. Besides individual 332 with a random quadratic term, the unconstrained individual curve fits are very similar, demonstrating the flexibility of the random effects to fit data when constraints are imposed on the mean curve.

In Figures 3 and 4 we show a lower order polynomial (degree 4) to illustrate the situation where we define the monotonicity constraint over a closed interval that does not extend over the entire range of the data and show an example where the estimated constrained subject-specific mean does not contain a stationary point. The region of the monotonicity restriction to be over days 2 to 6 is chosen to reflect the nature of the data, given that reaction times could recover after participants were allowed an eight hour sleeping pattern in days 7 through 9. Subject 335 is removed since they have a negative trend which clearly violates the monotonically increasing assumption that we are trying to demonstrate.

The mean fitted curves in Figure 3 are very similar, with the constrained individuals' curve ($q = 4^*, r = 2^*$) having a very close fit to that of the unconstrained curves ($q = 4, r = 2, 3$) which are indistinguishable. The (scaled) estimated coefficients for the degree 4 fits are given in Table 1, along with applicable standard errors. The standard errors for the constrained fits were calculated using case bootstrapping (by subject), whilst the estimated coefficients and standard errors of the unconstrained fits were calculated with `lme4`.

A subset of individuals that exhibit varying levels of different subject-specific curves are given in Figure 4, and again the random effects are shown to be very flexible at accounting for the differences in the constrained and unconstrained mean curves.

We note that a limitation of the constrained random effects models may be that for estimated mean curves that are very close to the boundary, these curves may no longer be interpretable as estimates of the population mean curves since the random effect distribution is no longer symmetric. Appendix F contains an example on human growth data and constrained subject-specific curves when $r > 2$.

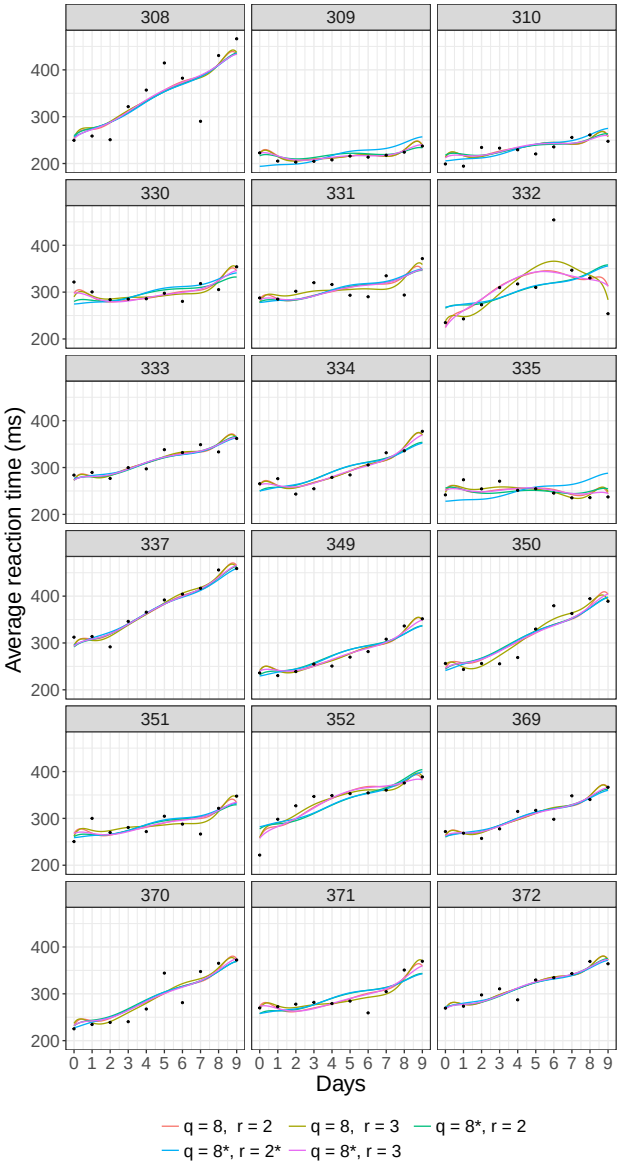

Fig. 2 Subject-specific fitted curves for degree 8 polynomials with varying random effects and constraint structures.

6 Discussion

In order to achieve a monotone constrained mixed effects model estimation procedure, a new fixed effects method, no longer based on a non-linear reparameterisation, was developed. *Constrained orthogonal least squares* speeds up the estimation of monotonic polynomial fixed effects models, and is applicable to any polynomial least squares estimation

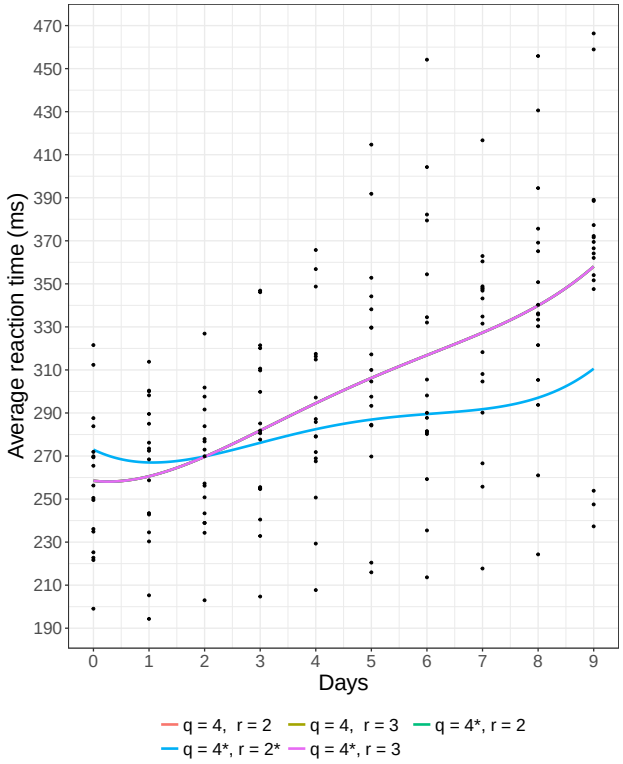

Fig. 3 Mean fitted curves for degree 4 polynomials with varying random effects and constraint structures. All curves, but $q = 4^*, r = 2^*$, coincide.

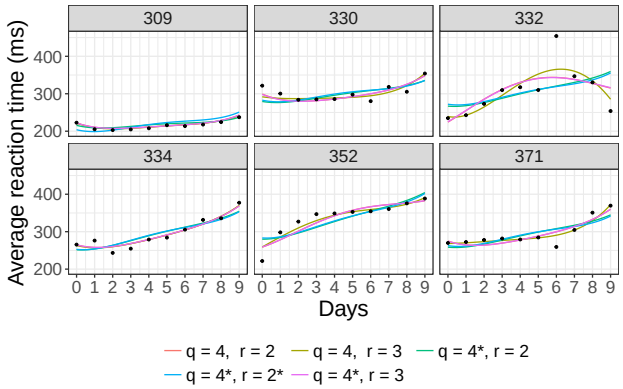

Fig. 4 Subject-specific fitted curves for degree 4 polynomials with varying random effects and constraint structures.

with a closed convex constraint. Moreover, employing a QR decomposition can potentially extend the use of COLS to a variety of transformations of the regressor variables, not just polynomials. This is possible due to the minimal assumptions behind COLS estimation. A closed convex constraint set, oracle function, orthonormal design matrix, and a linear transformation matrix to the standard coefficients is all that is needed for least squares and penalised least squares constrained regressions.

	β_0	β_1	β_2	β_3	β_4	σ
$q = 4$,	-0.22	0.39	-0.10	-0.02	0.15	0.17
$r = 2$	(0.08)	(0.07)	(0.15)	(0.07)	(0.13)	(0.01)
$q = 4$,	-0.22	0.39	-0.10	-0.02	0.15	0.16
$r = 3$	(0.08)	(0.09)	(0.14)	(0.10)	(0.12)	(0.01)
$q = 4^*$,	-0.22	0.39	-0.10	-0.02	0.15	0.19
$r = 2$	(0.08)	(0.09)	(0.13)	(0.09)	(0.13)	(0.03)
$q = 4^*$,	-0.33	0.15	-0.17	-0.01	0.22	0.19
$r = 2^*$	(0.07)	(0.09)	(0.17)	(0.09)	(0.15)	(0.03)
$q = 4^*$,	-0.22	0.39	-0.10	-0.02	0.15	0.17
$r = 3$	(0.08)	(0.09)	(0.13)	(0.09)	(0.13)	(0.03)

Table 1 Estimated mean parameters and residual (standard deviation with standard errors in parenthesis) for degree 4 polynomials with varying random effects and constraint structures. Standard errors for constrained fits were calculated with case bootstrapping ($N = 1000$), whereas unconstrained fits are standard output from `lme4`. The data has been scaled to $[-1, 1]$.

Using COLS for fixed effects models has several advantages over non-parametric alternatives. For example, smooth fits and predicted values, simple detection of other derivative based quantities such as inflection points, and the ability to interpret results in the well-known parametric framework. The COLS procedure will also handle joint constraints, since the intersection of a set of closed convex sets will also be closed and convex. This enhances the ability of statistical practitioners to incorporate *a priori* information into their statistical models.

There is still work to do on monotonic and general constrained fits determined by COLS. For example, a theory of standard errors for the coefficients of constrained models needs to be developed. At present, bootstrapping seems like the best available option for reasonable standard errors, and Murray et al (2016) discuss this in detail. We implemented case bootstrapping for the sleep study mixed effects models in Section 5. Encouragingly, the results were in line with those reported from `lme4`. A Bayesian approach could be another option for assessing the uncertainty in the parameter estimates. Determining the degrees of freedom when some or all of the parameters are constrained would also be useful to reduce approximation in the hypothesis testing and variance estimation.

Uncertainty in the degrees of freedom also complicates standard model selection techniques, such as choosing the degree of the polynomial. Information criteria and hypothesis testing are not possible without an appropriate degree of freedom. Stratified m out of n bootstrapping to select polynomial degree has proven successful in fixed effects models Murray et al (2016), and can be extended to the mixed model case, but is computationally demanding.

Two cases for constrained mixed effects models were considered, (i) where the mean curve is constrained but not each subject-specific curve, and (ii) where both the mean curve and subjects' curves were constrained. For case (i), a general fitting procedure was given for an arbitrary num-

ber of random effects, so long as the number of random effects were less than the number of coefficients. For case (ii), the fitting procedure for a random intercept and slope was derived, and a Monte Carlo EM algorithm was proposed and tested for higher order random effects. This is the first time, to the authors knowledge, that a complex constraint has been incorporated into mixed effects models. Furthermore, the methodology is readily extendible to other closed convex parameter sets.

A numerically stable and versatile method for fitting constrained models, with fixed or mixed effects, of this kind has not yet been proposed in the statistical literature. The applicability and usefulness to scientists and statistical practitioners is compounded by the method's basis in likelihood theory. The developed methodology is complemented by the publicly available R code which will help statisticians incorporate *a priori* information into models and predictions.

A Penalised constrained orthogonal least squares regression

A generalisation of Algorithm 2 can be useful for constrained penalised regression models and within a mixed effects model framework. Specifically, in mixed effects models the $\hat{\beta}_i^U$ are not available in closed form. For notational simplicity, we discuss the necessary changes to Algorithm 2 within a penalised regression framework.

For our purposes the penalised regression takes the form

$$\min_{\boldsymbol{\beta}} \{\text{RSS}(\boldsymbol{\beta}) + \eta(\boldsymbol{\beta})\} \text{ s.t. } \boldsymbol{\beta} \in \Omega_{\boldsymbol{\beta}} \tag{24}$$

where $\eta(\boldsymbol{\beta})$ is a continuously differentiable penalty function. Let $\text{RSS}^*(\boldsymbol{\beta}) = \text{RSS}(\boldsymbol{\beta}) + \eta(\boldsymbol{\beta})$ be the function to be minimised. The partial derivative of RSS^* with respect to β_i is

$$\frac{\partial \text{RSS}^*}{\partial \beta_i} = 2\left(\beta_i - \boldsymbol{X}_i^\top \boldsymbol{Y}\right) + \frac{\partial \eta}{\partial \beta_i} \tag{25}$$

for which we will seek the roots, i.e. the value for which

$$2\left(\beta_i - \boldsymbol{X}_i^\top \boldsymbol{Y}\right) + \frac{\partial \eta}{\partial \beta_i} = 0 \tag{26}$$

for each of the β_i. If the penalty term, $\eta(\boldsymbol{\beta})$, inhibits a closed form solution to (26) we must adjust Algorithm 2 to handle this additional complexity. In essence this adjustment can be implemented by using a Newton-Raphson (NR) type step before conducting the line search. Due to the inexact nature of this step, more iterations will be needed for convergence. This algorithm is a simple but effective extension to COLS and is detailed in Algorithm 3. As an alternative, the NR algorithm could be run before conducting any line search. However, Algorithm 3 does not require the unconstrained solution and so this may be unnecessary.

Algorithm 3 Monotone penalised regression fitting via coordinate descent, a Newton-Raphson step, and a line search.

Require: As in Algorithm 2.

procedure FIT.PCOLS($\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi}^\top, \boldsymbol{\beta}^{\text{init}}, d, T, \epsilon$)
  $\boldsymbol{\beta}^{[0]} \leftarrow \boldsymbol{\beta}^{\text{init}}$
  for $t = 0$ to T **do**
    $i \leftarrow t \mod d$
    $\boldsymbol{\beta}_{(i)}^{[t]} \leftarrow \boldsymbol{\beta}_{(i)}^{[t-1]}$
    $\hat{\beta}_i^{NR} \leftarrow \beta_i^{[t-1]} + h(\beta_i^{[t-1]})$ ▷ NR step or approx-NR step.
    $\beta_i^{[t]} \leftarrow$ LINESEARCH($\boldsymbol{\beta}^{[t-1]}, \hat{\beta}_i^{NR}, i, I(\boldsymbol{\beta}, \boldsymbol{\Psi}^\top)$)
    if $t > d$ and $\|\boldsymbol{\beta}^{[t]} - \boldsymbol{\beta}^{[t-d]}\| < \epsilon$ **then return** $\boldsymbol{\beta}^{[t]}$
return "did not converge"

The Newton-Raphson step uses the function $h(b)$, see (27) for example, which may be approximated based on the difficulty of finding the second derivative, $\frac{\partial^2 \eta}{\partial \beta_i^2}$. If the second derivative is available then $h(b)$ can take the form

$$h(b) = -\left.\frac{\partial \text{RSS}^*}{\partial \beta_i}\left(\frac{\partial^2 \text{RSS}^*}{\partial \beta_i^2}\right)^{-1}\right|_{\beta_i = b} \tag{27}$$

where $\frac{\partial^2 \text{RSS}^*}{\partial \beta_i^2} = 2 + \frac{\partial^2 \eta}{\partial \beta_i^2}$ when $\boldsymbol{X}$ is an orthonormal design matrix. If $\frac{\partial^2 \eta}{\partial \beta_i^2}$ is difficult to evaluate, a suitable approximation can be made. For example, it may be known that $\frac{\partial^2 \eta}{\partial \beta_i^2}$ is small compared to 2 so we may express $\frac{\partial^2 \text{RSS}^*}{\partial \beta_i^2} \approx 2 + e$ where e is a positive number used to shrink the step size so that we do not overshoot the optimal value or get stuck on a boundary. A quasi-NR step is also possible. In testing, values of e between 1 and 2 demonstrated the most potential. Reducing the step size was also effective for Algorithm 2. Henceforth we refer to the optimisation procedure in Algorithm 3 as *penalised constrained orthogonal least squares* (pCOLS).

B Conditional truncated normal distribution

The first and second order moments of the conditional distributions $\boldsymbol{\mathcal{U}} \mid \boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}_T \mid \boldsymbol{\mathcal{Y}}$ are needed for use in the EM algorithm, for the unconstrained random effects and constrained random effects respectively. Equation (9) gives the mean and variance in the unconstrained scenario. Finding the moments when the random effects are truncated is a non-trivial problem. We consider the general case of an arbitrary number of random effects, although generally $rg < d \ll n$. We have to start by finding the joint distribution of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ denoted here by $\boldsymbol{\mathcal{K}} = \begin{bmatrix} \boldsymbol{\mathcal{Y}} \\ \boldsymbol{\mathcal{U}}_T \end{bmatrix}$ where the model for $\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}_T$ and $\boldsymbol{\mathcal{U}}_T$ is

$$\begin{aligned} \boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}_T = \boldsymbol{U} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}, \boldsymbol{R}) \\ \boldsymbol{\mathcal{U}}_T &\sim \mathcal{N}_{T(\boldsymbol{\beta})}(\boldsymbol{0}, \boldsymbol{G}) \end{aligned}$$

as in equation (14) and $T(\boldsymbol{\beta}) \subset \mathbb{R}^{rg}$. We note that $\boldsymbol{\mathcal{U}}_T$ may be defined using the underlying normal distribution of $\boldsymbol{\mathcal{U}}$ by

$$\begin{aligned} &\boldsymbol{\mathcal{U}}_T = \boldsymbol{\mathcal{U}} \mid \boldsymbol{\mathcal{U}} \in T(\boldsymbol{\beta}), \text{ where } \boldsymbol{\mathcal{U}} \sim \mathcal{N}(0, \boldsymbol{G}) \\ &\text{and } f_{\boldsymbol{\mathcal{U}}_T}(\boldsymbol{U}) = 0 \text{ if } \boldsymbol{U} \notin T(\boldsymbol{\beta}) \end{aligned} \tag{28}$$

where $f_{\boldsymbol{\mathcal{U}}_T}(\boldsymbol{U})$ is the density of $\boldsymbol{U}_T$. Let $\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{U}}_T$ and $\boldsymbol{\mathcal{U}}$ have the densities $f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}_T}(\boldsymbol{Y}|\boldsymbol{U})$, and $f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U})$ respectively. The density $f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U})$ of $\boldsymbol{\mathcal{K}}$ can be written as the product of the density of $\boldsymbol{\mathcal{Y}}$ given $\boldsymbol{\mathcal{U}}_T$ with the density of $\boldsymbol{\mathcal{U}}_T$ as

$$f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U}) = f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}_T}(\boldsymbol{Y}|\boldsymbol{U}) f_{\boldsymbol{\mathcal{U}}_T}(\boldsymbol{U}) \tag{29}$$

with $\boldsymbol{U} \in T(\boldsymbol{\beta})$. We may rewrite the density $f_{\boldsymbol{\mathcal{U}}_T}(\boldsymbol{U})$, the truncated normal distribution, in terms of the non-truncated density of $\boldsymbol{\mathcal{U}}$ by

$$f_{\boldsymbol{\mathcal{U}}_T}(\boldsymbol{U}) = \frac{f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U})}{\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W}} \tag{30}$$

with $\boldsymbol{U} \in T(\boldsymbol{\beta})$ and where $\boldsymbol{W}$ is a dummy variable for the rg-dimensional integration. Substituting equation (30) into (29) allows the joint density to be written as

$$f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U}) = \frac{f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}_T}(\boldsymbol{Y}|\boldsymbol{U}) f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U})}{\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W}} \tag{31}$$

with $\boldsymbol{U} \in T(\boldsymbol{\beta})$. Consider the functional form of $f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}_T}(\boldsymbol{Y}|\boldsymbol{U})$. It is normally distributed without truncation since the random effects term is given. Therefore it has the same distribution (and density) as $f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{Y}|\boldsymbol{U})$ where the random effects are non-truncated. We may rewrite equation (31) using this identity as

$$f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U}) = \frac{f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{Y}|\boldsymbol{U}) f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U})}{\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W}} \tag{32}$$

$$= \frac{f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{Y}, \boldsymbol{U})}{\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W}} \tag{33}$$

hence the joint distribution of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ is a truncated normal distribution, with $\boldsymbol{U} \in T(\boldsymbol{\beta})$. The second step in equation (33) is valid from recognising the product of $f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{Y}|\boldsymbol{U}) f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U})$ as the unconstrained joint density. Note that in $f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U})$ no truncation is imposed on $\boldsymbol{\mathcal{Y}}$ and potentially some elements of $\boldsymbol{\mathcal{U}}$. At a minimum the random intercept has

no effect the monotonicity and so can be integrated out of the denominator of equation (33).

Conditioning $f_{\boldsymbol{\mathcal{K}}}(\boldsymbol{Y}, \boldsymbol{U})$ in equation (33) on $\boldsymbol{\mathcal{Y}} = \boldsymbol{Y}$ the density $f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U})$ (written to emphasise conditioning on $\boldsymbol{Y}$) can be described up to proportionality as

$$\begin{aligned} f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U}) &\propto f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{Y}, \boldsymbol{U}) \\ &\propto f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U}) \end{aligned} \tag{34}$$

since both $f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}$ and $f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}$ are proportional to $f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}$ for fixed $\boldsymbol{Y}$. Hence $f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U})$ can be written as

$$f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U}) = \frac{f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U})}{\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{W})\,\mathrm{d}\boldsymbol{W}} \tag{35}$$

by re-normalising the density in equation (34), for $\boldsymbol{U} \in T(\boldsymbol{\beta})$. This indicates that the conditional distribution of $\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}$ is also a truncated multivariate normal where the distribution before truncation has the same properties as the unconstrained (non-truncated) case, $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}$. A similar result for a conditional point-truncated multivariate normal distribution appears in Horrace (2005).

The well know result for the conditional distribution, $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}$, is restated here from equation (9), as $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{Y}) \sim \mathcal{N}(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}, \boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}})$ where

$$\begin{aligned} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}} &= \boldsymbol{G}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \text{ and} \\ \boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}} &= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R})^{-1}\boldsymbol{Z}\boldsymbol{G} \end{aligned}$$

Using the corresponding density we replace $f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U})$ in equation (35) to obtain the general constrained conditional density of the random effects as

$$f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{U}) = \frac{((2\pi)^{rg}|\boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}|)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{U} - \boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}})^\top \boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}^{-1}(\boldsymbol{U} - \boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}})\right\}}{\int_{T(\boldsymbol{\beta})} ((2\pi)^{rg}|\boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}|)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{W} - \boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}})^\top \boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}^{-1}(\boldsymbol{W} - \boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}})\right\} \mathrm{d}\boldsymbol{W}}. \tag{36}$$

The evaluation of the both the above density and moments of $\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}$ is hindered by the form of the truncation $T(\boldsymbol{\beta})$. In general the integrals $\int_{T(\boldsymbol{\beta})} f_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{W})\,\mathrm{d}\boldsymbol{W}$ and $\int_{T(\boldsymbol{\beta})} \boldsymbol{W}^m f_{\boldsymbol{\mathcal{U}}_T|\boldsymbol{\mathcal{Y}}=\boldsymbol{Y}}(\boldsymbol{W})\,\mathrm{d}\boldsymbol{W}$ are rg dimensional integrals with dependent components. We are helped by independence of individuals but this problem is difficult because of the complex nature of $T(\boldsymbol{\beta})$.

If $T(\boldsymbol{\beta})$ could be defined by a set of box constraints[2] the work of Tallis (1961) and Lee (1979) to establish the analytical results for the moments of point-truncated multivariate normal densities and their numerical calculation (Leppard and Tallis, 1989) can be used. As discussed in Section 4 this only occurs for $r = 2$. In this case the R package, `tmvtnorm` (Wilhelm and Manjunath, 2015), is able to provide the truncated distribution's mean and variance based of the unconstrained distribution of $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}$ with mean $\boldsymbol{M}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}$ and variance $\boldsymbol{V}_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}}$.

For $r > 2$ Monte Carlo simulation is necessary. One could implement a routine similar to that of Damien and Walker (2001) but replace the indicator function for standard truncation with one that adheres to the constraints of our application. This would simulate a $T(\boldsymbol{\beta})$-truncated multivariate normal distribution which could then be used to evaluate the mean and variance. We outline our proposed methods in Appendix C.3 and D.4.

[2] Box constraints occur when $T(\boldsymbol{\beta}) = \left\{(u_1, u_2, \ldots, u_{rg}) \in \mathbb{R}^{rg} : a_{i,1} \leq u_i \leq a_{i,2}, i = 1, 2, \ldots, rg\right\}$ where the $a_{i,j}$'s are constants.

C Technical details of expectation steps

C.1 Random intercepts

To complete the derivation of the expectation step for a random intercept model, we need to find the conditional expectation and variance of the random effects terms $\boldsymbol{\mathcal{U}}$. These expectations and variances are conditional on observing $\boldsymbol{\mathcal{Y}} = \boldsymbol{Y}$ and their calculations are simplified as the random intercepts do not impact the monotonicity of individuals' polynomials. Furthermore, the random effects' variance structure is now defined by $\boldsymbol{\phi}_{\boldsymbol{H}} = \{\sigma^2_{\boldsymbol{H}}\}$ so that $\boldsymbol{H} = [\sigma^2_{\boldsymbol{H}}]$ and $\boldsymbol{G} = \sigma^2_{\boldsymbol{H}}\boldsymbol{I}_g$. Additionally, we assume the error terms have homogenous variance so that $\boldsymbol{R} = \sigma^2_{\boldsymbol{R}}\boldsymbol{I}_n$. Therefore from (9) we see that the conditional mean is

$$\begin{aligned} \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}} &= \boldsymbol{G}^{[t]}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{G}^{[t]}\boldsymbol{Z}^\top + \boldsymbol{R}^{[t]})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}) \\ &= \boldsymbol{Z}^\top\left(\boldsymbol{Z}\boldsymbol{Z}^\top + \frac{(\sigma^2_{\boldsymbol{R}})^{[t]}}{(\sigma^2_{\boldsymbol{H}})^{[t]}}\boldsymbol{I}_n\right)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}) \end{aligned}$$

while the conditional variance is given by

$$\begin{aligned} \boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}} &= \boldsymbol{G}^{[t]} - \boldsymbol{G}^{[t]}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{G}^{[t]}\boldsymbol{Z}^\top + \boldsymbol{R}^{[t]})^{-1}\boldsymbol{Z}\boldsymbol{G}^{[t]} \\ &= (\sigma^2_{\boldsymbol{H}})^{[t]}\left(\boldsymbol{I}_g - \boldsymbol{Z}^\top\left(\boldsymbol{Z}\boldsymbol{Z}^\top + \frac{(\sigma^2_{\boldsymbol{R}})^{[t]}}{(\sigma^2_{\boldsymbol{H}})^{[t]}}\boldsymbol{I}_n\right)^{-1}\boldsymbol{Z}\right). \end{aligned}$$

C.2 Random slope

When random slopes are included in the model we need to impose a suitable truncation on the distribution of the random effects, when we also wish to constrain the individuals' curves. The first step is to derive the form of truncation needed when we have two random effects present. Define $p(x; \boldsymbol{\beta}, u_{0,i}, u_{1,i})$ as the i^{th} individuals' orthonormal polynomial curve to be estimated in the random intercepts and slopes model. We may write this equation by extending (1) so that

$$\begin{aligned} &p(x; \boldsymbol{\beta}, u_{0,i}, u_{1,i}) = \\ &(\beta_0 + u_{0,i})p_0(x) + (\beta_1 + u_{1,i})p_1(x) + \sum_{j=2}^{q}\beta_j p_j(x) \end{aligned} \tag{37}$$

where $u_{0,i}$ and $u_{1,i}$ are the random intercepts and slopes and the p_j are defined in (2). Note also that $p_0(x) = \psi_{0,0}$ and $p_1(x) = \psi_{1,0} + \psi_{1,1}x$, where $\psi_{i,j}$ are elements of $\boldsymbol{\Psi}$. The derivative of (37) with respect to x is

$$p'(x; \boldsymbol{\beta}, u_{1,i}) = \psi_{1,1}(\beta_1 + u_{1,i}) + \sum_{j=2}^{q}\beta_j p'_j(x) \tag{38}$$

whereby monotonicity for each individual i is maintained whilst $p'(x; \boldsymbol{\beta}, u_{1,i}) \geq 0$ for all x in the set S. When estimating the constrained random effects we take $\boldsymbol{\beta}$ as fixed in our optimisation process for a given iteration, in accordance with the E-step of the EM algorithm. After fixing $\boldsymbol{\beta}$, the monotonicity for each individuals' curve is determined additively by $u_{1,i}$ and the shape of the curve. To determine the necessary truncation for $u_{1,i}$ the minimum of the derivative of the individual's curve as a function of $u_{1,i}$ is sought. Maintaining this minimum above zero will provide the correct truncation. Given (38) we see that

$$\min_{x \in S} p'(x; \boldsymbol{\beta}, u_{1,i}) = \min_{x \in S} p'(x; \boldsymbol{\beta}) + \psi_{1,1}u_{1,i} \tag{39}$$

where the derivative of the mean curve is $p'(x;\boldsymbol{\beta}) = \sum_{j=1}^{q} \beta_j p'_j(x)$. We denote the scaled minimum of this polynomial as

$$c(\boldsymbol{\beta}) = \psi_{1,1}^{-1} \min_{x \in S} p'(x;\boldsymbol{\beta}) \tag{40}$$

which can be calculated for a given $\boldsymbol{\beta}$ by applying standard root-solving routines to the derivative of $p'(x;\boldsymbol{\beta})$. The fixed coefficient $\psi_{1,1}$ that appears in (40) can be expressed as

$$\psi_{1,1} = \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-1/2} \tag{41}$$

as a result of the orthonormality of $p_0(x)$ and $p_1(x)$ over the set of observations $\{x_1, x_2, \ldots, x_n\}$. Once $c(\boldsymbol{\beta})$ is determined, the monotonicity of the individuals' curves in the random intercept and slope model is guaranteed by one of the following equivalent statements

$$\begin{aligned} \min_{x \in S} p'(x;\boldsymbol{\beta}, u_{1,i}) &\geq 0 \\ \psi_{1,1} c(\boldsymbol{\beta}) + \psi_{1,1} u_{1,i} &\geq 0 \\ u_{1,i} &\geq -c(\boldsymbol{\beta}). \end{aligned}$$

The last statement uses $\psi_{1,1} > 0$ by its definition in (41). Hence, we have a one sided point truncation of $u_{1,i}$ which can be used to define the truncation region $T(\boldsymbol{\beta})$ as

$$\begin{aligned} T(\boldsymbol{\beta}) = \Big\{ \boldsymbol{\mathcal{U}}_T = \left[u_{0,1}\, u_{1,1}\, u_{0,2}\, u_{1,2} \cdots u_{0,g}\, u_{1,g} \right]^\top \in \mathbb{R}^{2g} \\ \text{s.t. } u_{i,1} > -c(\boldsymbol{\beta}), i = 1, 2, \ldots, g \Big\} \end{aligned} \tag{42}$$

when we have a random intercept and slope.

Having the form of the truncation in (42), we use the result of Appendix B for $\boldsymbol{\mathcal{U}}_T \,|\, \boldsymbol{\mathcal{Y}}$ which tells us the conditional distribution is a truncated multivariate normal random variable. The R package `tmvtnorm` (Wilhelm and Manjunath, 2015) can be used to find the expectation and variance of this distribution for use in the E-step because $T(\boldsymbol{\beta})$ defines a point truncation.

C.3 Higher order random effects

To evaluate the expectation and variance of the conditional random effects when $r > 2$ we use Monte Carlo integration. There are two simple Monte Carlo methods we consider for this task, rejection sampling and the Metropolis-Hastings algorithm (Hastings, 1970).

The rejection sampler proceeds by drawing samples from the multivariate normal distribution governing the unconstrained conditional random effects. As in the general case, the relevant unconstrained mean and variance are given by

$$\begin{aligned} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} &= \boldsymbol{G}^{[t]} \boldsymbol{Z}^\top (\boldsymbol{Z} \boldsymbol{G}^{[t]} \boldsymbol{Z}^\top + \boldsymbol{R}^{[t]})^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}) \quad \text{and} \\ \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} &= \boldsymbol{G}^{[t]} - \boldsymbol{G}^{[t]} \boldsymbol{Z}^\top (\boldsymbol{Z} \boldsymbol{G}^{[t]} \boldsymbol{Z}^\top + \boldsymbol{R}^{[t]})^{-1} \boldsymbol{Z} \boldsymbol{G}^{[t]}. \end{aligned}$$

However, we decompose these matrices into subject-specific means and variances assuming no crossed random effects, i.e. $\boldsymbol{G}$ is a block-diagonal matrix. This way we can sample from g r-dimensional distributions rather than one rg-dimensional distribution. It is also easily parallelised.

A random sample is drawn from the underlying unconstrained multivariate normal distribution for each subject. The realisations are rejected when the resulting subject-specific curve is not monotone. The remaining set are effectively samples drawn from the truncated multivariate normal distribution and the approximate expectation and variance is calculated from this set.

Rejection sampling is adequate if a sufficient number of realisations of the unconstrained sample are monotone. If not, it may be very slow to draw realisations from the constrained distribution. In these cases, a random walk Metropolis algorithm (Metropolis et al, 1953) is a good alternative. In order to adhere to the constrained space an additional step is added to the algorithm that rejects non-monotonic proposals. This algorithm is in the class of random walk Metropolis-Hastings (RWMH) samplers.

It is sensible to define the random walk in the RWMH to have variance equal to $\tau \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}$ where $\tau \in (0, 1)$ is a constant used to scale down the step size. Using this variance incorporates the inherent correlation of the random effects which is only altered by the truncation from the extra rejection step. The only tuning parameter is then τ, which can be chosen in the warm-up phase or be pre-determined. Implementations should consider thinning the samples to reduce autocorrelation in the Markov chain.

To reduce computation time we use a combination of rejection sampling and RWMH. Initially, we draw a preliminary sample of each subjects' constrained random effects using rejection sampling and calculate individual acceptance ratios. The subjects with low acceptance ratios are transferred to the RWHM algorithm to draw samples – the more efficient option. As with the rejection sampler, the approximate expectation and variance is then calculated from the constrained random samples of each individual.

D Technical details of maximisation steps

D.1 Homogenous observational variance

In the special case where $\boldsymbol{R} = \sigma_{\boldsymbol{R}}^2 \boldsymbol{I}_n$ so that $\boldsymbol{\phi}_{\boldsymbol{R}} = \left\{ \sigma_{\boldsymbol{R}}^2 \right\}$, the equation $\frac{\partial q_{\text{dev}}^{[t]}}{\partial \sigma_{\boldsymbol{R}}^2} = 0$ can be solved analytically. The derivative of $q_{\text{dev}}^{[t]}$ with respect to $\sigma_{\boldsymbol{R}}^2$ can be simplified to

$$\begin{aligned} \frac{\partial q_{\text{dev}}^{[t]}}{\partial \sigma_{\boldsymbol{R}}^2} &= \text{tr}\left[\frac{1}{\sigma_{\boldsymbol{R}}^2} \boldsymbol{I}_n \right] - \text{tr}\left[\frac{1}{\sigma_{\boldsymbol{R}}^4} \boldsymbol{I}_n \boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top \right] - \left(\boldsymbol{W}^{[t]} \right)^\top \frac{1}{\sigma_{\boldsymbol{R}}^4} \boldsymbol{I}_n \boldsymbol{W}^{[t]} \\ &= \frac{n}{\sigma_{\boldsymbol{R}}^2} - \frac{1}{\sigma_{\boldsymbol{R}}^4} \text{tr}\left[\boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top \right] - \frac{1}{\sigma_{\boldsymbol{R}}^4} \left(\boldsymbol{W}^{[t]} \right)^\top \boldsymbol{W}^{[t]}. \end{aligned} \tag{43}$$

Equating (43) to zero we find that, the optimal $\sigma_{\boldsymbol{R}}^2$ for a given $\boldsymbol{\beta}$ is

$$\sigma_{\boldsymbol{R}}^2 = \frac{1}{n} \left[\text{tr}\left[\boldsymbol{Z} \boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]} \boldsymbol{Z}^\top \right] + \left(\boldsymbol{W}^{[t]} \right)^\top \boldsymbol{W}^{[t]} \right]. \tag{44}$$

The denominator in (44) could be replaced by the degrees of freedom in the model. If this were a fixed effects model we would replace n by $n - d$ where d is the number of coefficients. However, calculating the degrees of freedom in a constrained mixed effects model is complicated in two ways. Firstly, d describes the upper limit for the number of fixed effects parameters in the context of degrees of freedom. This is because the number d does not account for the constraint over the parameter space. Secondly, we need to account for the random effects which do not constitute parameters in the general sense but are estimated and reduce the degrees of freedom in the model[3]. We defer this choice for future consideration, and for the time being conservatively use n.

In the optimisation of $\boldsymbol{\beta}$, with $\boldsymbol{R} = \sigma_{\boldsymbol{R}}^2 \boldsymbol{I}_n$, we can simplify (20) to

$$\begin{aligned} \frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\beta}} &= -\frac{2}{\sigma_{\boldsymbol{R}}^2} \boldsymbol{X}^\top \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} \right) + 2 \frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= -\frac{2}{\sigma_{\boldsymbol{R}}^2} \left(\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{\beta} - \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} \right) + 2 \frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \end{aligned} \tag{45}$$

[3] For a detailed discussion on the degrees of freedom in a mixed effects model see Chapter 2.2 of Hodges (2013).

and optimisation still needs to be iterative in the M-step since the dependence on $\sigma^2_{\boldsymbol{R}}$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$ is retained (the parameters in $\boldsymbol{\phi}_{\boldsymbol{H}}$ enter (45) through $\eta(\boldsymbol{\beta})$). Therefore the parameter dependence structure has not changed the scenario described in Section 4.2 when $\eta(\boldsymbol{\beta}) \neq 0$. To find $\boldsymbol{\beta}$ in each step, pCOLS can still be used iteratively in combination with optimisation of the variance parameters.

When the individual curves are not constrained ($\eta(\boldsymbol{\beta}) = 0$), equating (45) to a vector of zeros gives $\hat{\boldsymbol{\beta}}^U = \boldsymbol{X}^\top \boldsymbol{Y}^{[t]}_*$, which replaces the standard unconstrained solution in Algorithm 2 for the COLS optimisation. The dependence on $\sigma^2_{\boldsymbol{R}}$ can be ignored in the coordinate descent routine, since it is just a multiplicative constant to the derivative. Hence the M-step dependence structure with homogenous observational variance and no constraint on the random effects ($\eta(\boldsymbol{\beta}) = 0$) is simplified. The optimisation of

- $\boldsymbol{\beta}$ has no dependence on the other parameter sets,
- $\boldsymbol{\phi}_{\boldsymbol{R}}$ depends on $\boldsymbol{\beta}$, and
- $\boldsymbol{\phi}_{\boldsymbol{H}}$ has no dependence on the other parameter sets

in any one M-step. So to estimate this model we carry out the expectation step, maximise over $\boldsymbol{\beta}$ followed by $\boldsymbol{\phi}_{\boldsymbol{R}}$. The variance parameters $\boldsymbol{\phi}_{\boldsymbol{H}}$ can be optimised at any stage. Then we repeat the expectation and maximisation steps until convergence is reached.

D.2 A random intercept

Having only a random intercept model simplifies the optimisation step considerably. It reduces the variance vector $\boldsymbol{\phi}_{\boldsymbol{H}}$ to $\boldsymbol{\phi}_{\boldsymbol{H}} = \left\{\sigma^2_{\boldsymbol{H}}\right\}$ so that $\boldsymbol{H} = \left[\sigma^2_{\boldsymbol{H}}\right]$ and $\boldsymbol{G} = \sigma^2_{\boldsymbol{H}} \boldsymbol{I}_g$. We continue to assume $\boldsymbol{R}$ has homogenous variance so $\boldsymbol{R} = \sigma^2_{\boldsymbol{R}} \boldsymbol{I}_n$, which still provides simplification if this is not the case. Including only a random intercept dictates that a subject's curve is monotone if the mean curve is and so there is no need for truncation of the random effects' normal distribution. The random intercept is normally distributed and the normalising term, $\eta(\boldsymbol{\beta})$, is null.

Updating the derivative of $q^{[t]}_{\text{dev}}$ with respect to $\sigma^2_{\boldsymbol{H}}$ from (22) with $\eta(\boldsymbol{\beta}) = 0$, we have

$$\begin{aligned}
\frac{\partial q^{[t]}_{\text{dev}}}{\partial \sigma^2_{\boldsymbol{H}}} &= \operatorname{tr}\left[\boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{H}}}\right] - \operatorname{tr}\left[\boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{H}}} \boldsymbol{G}^{-1} \boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}}\right] \\
&\quad - \left(\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right)^\top \boldsymbol{G}^{-1} \tfrac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{H}}} \boldsymbol{G}^{-1} \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}} \\
&= \operatorname{tr}\left[\sigma^{-2}_{\boldsymbol{H}} \boldsymbol{I}_g \boldsymbol{I}_g\right] - \operatorname{tr}\left[\sigma^{-2}_{\boldsymbol{H}} \boldsymbol{I}_g \boldsymbol{I}_g \sigma^{-2}_{\boldsymbol{H}} \boldsymbol{I}_g \boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}}\right] \\
&\quad - \left(\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right)^\top \sigma^{-2}_{\boldsymbol{H}} \boldsymbol{I}_g \boldsymbol{I}_g \sigma^{-2}_{\boldsymbol{H}} \boldsymbol{I}_g \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}} \\
&= g\sigma^{-2}_{\boldsymbol{H}} - \sigma^{-4}_{\boldsymbol{H}} \operatorname{tr}\left[\boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}}\right] - \sigma^{-4}_{\boldsymbol{H}} \left(\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right)^\top \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}.
\end{aligned}$$

Setting the derivative to zero we find that

$$\sigma^2_{\boldsymbol{H}} = \frac{1}{g}\left[\operatorname{tr}\left[\boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}}\right] + \left(\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right)^\top \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right]. \tag{46}$$

Since $\eta(\boldsymbol{\beta}) = 0$, the unconstrained solution for $\boldsymbol{\beta}$ can be found analytically from (45), and takes the form

$$\boldsymbol{\beta}^U = \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}} \tag{47}$$

in a given iterate. Consequently, the $\sigma^2_{\boldsymbol{R}}$ term can be ignored in the coordinate descent optimisation when only a random intercept is present.

This gives us the final components of the updating equations for our variance components. Using the COLS methods for $\boldsymbol{\beta}$, (44) for $\sigma^2_{\boldsymbol{R}}$, and (46) for $\sigma^2_{\boldsymbol{H}}$, the EM updating equations are

$$\begin{aligned}
\boldsymbol{\beta}^{[t+1]} &= \text{ COLS (Algorithm 2) with (47)} \\
\left(\sigma^2_{\boldsymbol{R}}\right)^{[t+1]} &= \frac{1}{n}\left[\operatorname{tr}\left[\boldsymbol{Z} \boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}} \boldsymbol{Z}^\top\right] + \left(\boldsymbol{Y}^{[t]}_* - \boldsymbol{X}\boldsymbol{\beta}^{[t+1]}\right)^\top \left(\boldsymbol{Y}^{[t]}_* - \boldsymbol{X}\boldsymbol{\beta}^{[t+1]}\right)\right] \\
\left(\sigma^2_{\boldsymbol{H}}\right)^{[t+1]} &= \frac{1}{g}\left[\operatorname{tr}\left[\boldsymbol{V}^{[t]}_{\boldsymbol{\mathcal{U}}}\right] + \left(\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right)^\top \boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}\right]
\end{aligned} \tag{48}$$

in this order, where $\boldsymbol{Y}^{[t]}_* = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{M}^{[t]}_{\boldsymbol{\mathcal{U}}}$. This completes the EM algorithm for models with just a random intercept.

D.3 A random slope

Upon introducing a random slope into the model we must consider the normalising term and its derivatives, as well as the variance structure of the random effects. These will affect the way we undertake the maximisation step in each iterate of the EM algorithm. The first consideration is the definition and derivative of the normalising term $\eta(\boldsymbol{\beta})$. Alternatively to (42), define the support by grouping the random effects by individual as

$$T_i(\boldsymbol{\beta}) = \left\{\boldsymbol{\mathcal{U}}_{T,i} = [u_{0,i}\, u_{1,i}]^\top \in \mathbb{R}^2 : u_{i,1} > -c(\boldsymbol{\beta})\right\}$$

for $i = 1, 2, \ldots, g$, where $c(\boldsymbol{\beta})$ is defined in (40). In the case of random slopes, $\eta(\boldsymbol{\beta})$ can be simplified and evaluated analytically. Additionally, independence of the groups of random effects, and the marginalising over the random intercept, w_1, allows us to write $\eta(\boldsymbol{\beta})$ as

$$\begin{aligned}
\eta(\boldsymbol{\beta}) &= \log\left(\int_{T(\boldsymbol{\beta})} \left((2\pi)^{2g}|\boldsymbol{G}|\right)^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{U}^\top \boldsymbol{G}^{-1} \boldsymbol{U}\right\} \mathrm{d}\boldsymbol{U}\right) \\
&= \log\left(\left[\int_{T_i(\boldsymbol{\beta})} \left((2\pi)^2|\boldsymbol{H}|\right)^{-1/2} \exp\left\{-\frac{1}{2}\begin{bmatrix} u_1 & u_2 \end{bmatrix} \boldsymbol{H}^{-1} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\right\} \mathrm{d}u_1\, \mathrm{d}u_2\right]^g\right) \\
&= g \log\left(\int_{-c(\boldsymbol{\beta})}^{\infty} \int_{-\infty}^{\infty} \left((2\pi)^2|\boldsymbol{H}|\right)^{-1/2} \exp\left\{-\frac{1}{2}\begin{bmatrix} u_1 & u_2 \end{bmatrix} \boldsymbol{H}^{-1} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\right\} \mathrm{d}u_1\, \mathrm{d}u_2\right) \\
&= g \log\left(\int_{-c(\boldsymbol{\beta})}^{\infty} \left((2\pi)\sigma^2_{\boldsymbol{H},1}\right)^{-1/2} \exp\left\{-\frac{u_2^2}{2\sigma^2_{\boldsymbol{H},1}}\right\} \mathrm{d}u_2\right).
\end{aligned}$$

Finally, the normalising term can be written as

$$\eta(\boldsymbol{\beta}) = g \log\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{\boldsymbol{H},1}}\right)\right] \tag{49}$$

by noting its relation to the standard normal cumulative distribution function, denoted by $\Phi(z)$.

The expression of the normalising term in (49) allows us to evaluate the derivatives of $\eta(\boldsymbol{\beta})$ analytically. The derivative with respect to $\boldsymbol{\beta}$ is somewhat involved, requiring envelope theorem[4] (Milgrom and Segal, 2002) to evaluate the derivative[5] $\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{\beta}}$, in addition to successive chain rules. We find that

$$\frac{\partial \eta}{\partial \boldsymbol{\beta}} = \left(\frac{\Phi'\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{\boldsymbol{H},1}}\right)}{\sigma_{\boldsymbol{H},1}\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{\boldsymbol{H},1}}\right)\right]}\right) \frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{\beta}} \tag{50}$$

where $\Phi'(z)$ is just the density of the standard normal distribution. Using the results of Appendix E, the component form of (50) can be written as

$$\frac{\partial \eta}{\partial \beta_i} = \left(\frac{\Phi'\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{\boldsymbol{H},1}}\right)}{\sigma_{\boldsymbol{H},1}\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{\boldsymbol{H},1}}\right)\right]}\right) \frac{p_i'(x^*)}{\psi_{1,1}} \tag{51}$$

where $p_i'(x)$ is the derivative of the i^{th} polynomial in the orthonormal basis for $0 \leq i \leq q$, $x^* = \arg\min_{x \in S} p'(x; \boldsymbol{\beta})$, and $\psi_{1,1}$ is defined in (41).

The second consideration in the M-step when $r = 2$ is finding the derivative of $\eta(\boldsymbol{\beta})$ with respect to the random effects variance components. The variance-covariance matrix, $\boldsymbol{G}$, becomes a block-diagonal

[4] Appendix E contains details on the envelope theorem.

[5] The full derivative is used here because $c(\boldsymbol{\beta})$ is a multivariate composite function where each component is dependent on $\boldsymbol{\beta}$. See Appendix E for more details.

matrix made up of g positive definite matrices, $\boldsymbol{H}$, with dimension 2×2. The variance parameters of $\boldsymbol{H}$ have general derivative given by (22) restated here as

$$\frac{\partial q_{\text{dev}}^{[t]}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}} = \text{tr}\left[\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}\right] - \text{tr}\left[\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}\boldsymbol{G}^{-1}\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] - \left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^{\top}\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}\boldsymbol{G}^{-1}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} + 2\frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\phi}_{\boldsymbol{H},i}}.$$

Generally $\boldsymbol{H}$ is considered to have the form

$$\boldsymbol{H} = \begin{bmatrix}\sigma^2_{\boldsymbol{H},0} & \rho_{\boldsymbol{H}} \\ \rho_{\boldsymbol{H}} & \sigma^2_{\boldsymbol{H},1}\end{bmatrix}$$

and $\boldsymbol{G} = \boldsymbol{I}_g \otimes \boldsymbol{H}$. However, to estimate these parameters we need to transform $\boldsymbol{H}$, and hence $\boldsymbol{G}$, to induce an unconstrained optimisation problem. Although several matrix reparameterisations that preserve positive definiteness exist (Pinheiro and Bates, 1996), we use the log-Cholesky decomposition. The Cholesky parameterisation decomposes a positive definite matrix into the product of two triangular matrices, thereby ensuring positive definiteness when working with this form, but to ensure uniqueness we enact a log-transform on the diagonals. The log-Cholesky decomposition can be described as $\boldsymbol{H} = \boldsymbol{L}_{\boldsymbol{H}}\boldsymbol{L}_{\boldsymbol{H}}^{\top}$, where $\boldsymbol{L}_{\boldsymbol{H}}$ is a lower triangular matrix with diagonal raised to the exponential number.

Due to the dependence of normalising term, $\eta(\boldsymbol{\beta})$, on $\sigma^2_{\boldsymbol{H},1}$ we change this parameterisation to use an upper triangular matrix rather than the standard lower triangular matrix. The upper triangular matrix is therefore defined as

$$\boldsymbol{J}_{\boldsymbol{H}} = \begin{bmatrix}\exp\{\omega_1\} & \omega_3 \\ 0 & \exp\{\omega_2\}\end{bmatrix} \tag{52}$$

where the redefined $\boldsymbol{H}$ becomes

$$\boldsymbol{H} = \boldsymbol{J}_{\boldsymbol{H}}\boldsymbol{J}_{\boldsymbol{H}}^{\top} = \begin{bmatrix}\exp\{\omega_1\} & \omega_3 \\ 0 & \exp\{\omega_2\}\end{bmatrix}\begin{bmatrix}\exp\{\omega_1\} & 0 \\ \omega_3 & \exp\{\omega_2\}\end{bmatrix} \tag{53}$$

$$\begin{bmatrix}\sigma^2_{\boldsymbol{H},0} & \rho_{\boldsymbol{H}} \\ \rho_{\boldsymbol{H}} & \sigma^2_{\boldsymbol{H},1}\end{bmatrix} = \begin{bmatrix}\exp\{2\omega_1\} + \omega_3^2 & \omega_3\exp\{\omega_2\} \\ \omega_3\exp\{\omega_2\} & \exp\{2\omega_2\}\end{bmatrix} \tag{54}$$

which ensures that $\sigma^2_{\boldsymbol{H},1} = \exp\{2\omega_2\}$ is a function of just one of the new parameters rather than two.

Let $\boldsymbol{J}_{\boldsymbol{G}} = \boldsymbol{I}_g \otimes \boldsymbol{J}_{\boldsymbol{H}}$, replacing $\boldsymbol{G}$ with $\boldsymbol{J}_{\boldsymbol{G}}$ the derivative of $q_{\text{dev}}^{[t]}$ with respect to the parameters of $\boldsymbol{J}_{\boldsymbol{H}}$ (and hence $\boldsymbol{J}_{\boldsymbol{G}}$) becomes

$$\frac{\partial q_{\text{dev}}^{[t]}}{\partial \omega_i} = 2\text{tr}\left[\boldsymbol{J}_{\boldsymbol{G}}^{-1}\frac{\partial \boldsymbol{J}_{\boldsymbol{G}}}{\partial \omega_i}\right] - 2\text{tr}\left[\left(\boldsymbol{J}_{\boldsymbol{G}}\boldsymbol{J}_{\boldsymbol{G}}^{\top}\right)^{-1}\frac{\partial \boldsymbol{J}_{\boldsymbol{G}}}{\partial \omega_i}\boldsymbol{J}_{\boldsymbol{G}}^{-1}\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\right] - 2\left(\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}\right)^{\top}\left(\boldsymbol{J}_{\boldsymbol{G}}\boldsymbol{J}_{\boldsymbol{G}}^{\top}\right)^{-1}\frac{\partial \boldsymbol{J}_{\boldsymbol{G}}}{\partial \omega_i}\boldsymbol{J}_{\boldsymbol{G}}^{-1}\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]} + 2\frac{\partial \eta(\boldsymbol{\beta})}{\partial \omega_i} \tag{55}$$

where a standard derivative based routine can solve for (55) set to zero, and $\boldsymbol{M}_{\boldsymbol{\mathcal{U}}}^{[t]}$ and $\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}$ can be found using `tmvtnorm` (Wilhelm and Manjunath, 2015) as discussed in Section C.2. The normalising term from (49) can now be represented using ω_2 instead of $\sigma_{\boldsymbol{H},1}$ as

$$\eta(\boldsymbol{\beta}) = g\log\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\exp\{\omega_2\}}\right)\right] \tag{56}$$

for which the derivative of $\eta(\boldsymbol{\beta})$ with respect to ω_2 is

$$\frac{\partial \eta}{\partial \omega_2} = -g\left(\frac{\Phi'\left(\frac{-c(\boldsymbol{\beta})}{\exp\{\omega_2\}}\right)}{1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\exp\{\omega_2\}}\right)}\right)\frac{c(\boldsymbol{\beta})}{\exp\{\omega_2\}} \tag{57}$$

and zero for ω_1 and ω_3, thus fully specifying Equation (55).

In summary, with a random intercept and slope, the M-step becomes

$\boldsymbol{\beta}^{[t+1]}$ = pCOLS (Algorithm 3) using (45) and (51)

$$\left(\sigma^2_{\boldsymbol{R}}\right)^{[t+1]} = \frac{1}{n}\left[\text{tr}\left[\boldsymbol{Z}\boldsymbol{V}_{\boldsymbol{\mathcal{U}}}^{[t]}\boldsymbol{Z}^{\top}\right] + \left(\boldsymbol{Y}_*^{[t]} - \boldsymbol{X}\boldsymbol{\beta}^{[t+1]}\right)^{\top}\left(\boldsymbol{Y}_*^{[t]} - \boldsymbol{X}\boldsymbol{\beta}^{[t+1]}\right)\right]$$

$(\omega_i)^{[t+1]}$ = numerical optimisation using (19), (55), and (57) when appropriate.

In a single M-step these individual optimisations should be carried out iteratively to account for the dependence of the parameters detailed in Section 4.2. It is possible to undertake each optimisation once if a conditional EM algorithm is used (Meng and Rubin, 1993), however further investigation of the merits in this method are left for future x.

D.4 Higher order random effects

When $r > 2$ it is not clear how to analytically derive the partial derivatives of $\eta(\boldsymbol{\beta})$ given the space $T(\boldsymbol{\beta})$ is no longer box-constrained. Numerical integration is possible, but computationally intensive compared to the methods for $r \leq 2$ previously described.

Rather than working with an rg-dimensional integral, when the groups of random effects are independent (no crossed random effects) we can decompose the integral into g identical parts. Let $\boldsymbol{u} = \begin{bmatrix}u_1 & u_2 & \cdots & u_r\end{bmatrix}^T$ be a dummy variable for integration representing any of the subjects' random effects. The constrained region for each (subject-specific) set of the random effects can be written as

$$T_I(\boldsymbol{\beta}) = \left\{\boldsymbol{u} \in \mathbb{R}^r \text{ s.t. } \boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{u}^T & 0 & \cdots & 0\end{bmatrix}^T \in \Omega_{\boldsymbol{\beta}}\right\} \tag{58}$$

or in other words, the subject-specific polynomial created from the mean polynomial and the subjects' random effects must be monotone. With these we can write $\eta(\boldsymbol{\beta})$ with only an r-dimensional integral as

$$\begin{aligned}\eta(\boldsymbol{\beta}) &= \log\left(\int_{T(\boldsymbol{\beta})}((2\pi)^{rg}|\boldsymbol{G}|)^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{U}^{\top}\boldsymbol{G}^{-1}\boldsymbol{U}\right\}\,\mathrm{d}\boldsymbol{U}\right) \\ &= g\log\left(\int_{T_I(\boldsymbol{\beta})}((2\pi)^{r}|\boldsymbol{H}|)^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{u}^{\top}\boldsymbol{H}^{-1}\boldsymbol{u}\right\}\,\mathrm{d}\boldsymbol{u}\right).\end{aligned} \tag{59}$$

We use Monte Carlo integration to evaluate $\eta(\boldsymbol{\beta})$ and its' partial derivatives. The integrand of (59) is a probability density function, hence to approximate the integral we can generate samples from the distribution (multivariate normal) and count how many are in the integrable region. The approximate value of $\eta(\boldsymbol{\beta})$ is the proportion of these iterates belonging to $T_I(\boldsymbol{\beta})$.

Once the value of $\eta(\boldsymbol{\beta})$ can be approximated, numerical differentiation techniques can be used to find $\frac{\partial \eta}{\partial \boldsymbol{\beta}}$ and $\frac{\partial \eta(\boldsymbol{\beta})}{\partial \omega_i}$. However, some optimisations should be made first. The derivative of $\eta(\boldsymbol{\beta})$ with respect to the relevant parameter α can be written as

$$\frac{\partial \eta}{\partial \alpha} = g\frac{\frac{\partial}{\partial \alpha}\int_{T_I(\boldsymbol{\beta})}((2\pi)^r|\boldsymbol{H}|)^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{u}^{\top}\boldsymbol{H}^{-1}\boldsymbol{u}\right\}\,\mathrm{d}\boldsymbol{u}}{\int_{T_I(\boldsymbol{\beta})}((2\pi)^r|\boldsymbol{H}|)^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{u}^{\top}\boldsymbol{H}^{-1}\boldsymbol{u}\right\}\,\mathrm{d}\boldsymbol{u}} \tag{60}$$

and the numerator in (60) can then be numerically differentiated. Without any further adjustments, all realisations will need to be tested for membership in $T_I(\boldsymbol{\beta})$ every time the integral is calculated during numerical integration. To reduce this computational burden, we make two further approximations

1. Fix the realisations of the multivariate normal distribution for the entire EM algorithm (but use a large number of samples);
2. Only use a sub-sample of the realisations when approximating the derivative.

More specifically, the sub-sample should be the realisations from the underlying distribution that are sufficiently close to the monotone boundary, for a given $\boldsymbol{\beta}$. These are the only samples which will impact calculation of the derivative, as they lose and gain membership in $T_l(\boldsymbol{\beta})$ based on very small changes to $\boldsymbol{\beta}$ and $\omega_1, \omega_2, ...$ respectively. Realisations that are sufficiently far away from the boundary will not change membership during the derivative calculation. Conveniently, we can calculate how close the samples are to the monotone boundary using $c(\boldsymbol{\beta})$ in (40) and take a subset by smallest absolute value. The size of the sub-sample will affect accuracy of the derivative and computational effort. But in our testing we have found it is a worthwhile trade-off.

Once approximated, the derivatives and function can be used in the M-step as described in Appendix D.3.

E Derivative of the normalising term

In Section 4.2 we are faced with finding the derivative of the normalising term of the form

$$\eta(\boldsymbol{\beta}) = \log\left(\int_{T(\boldsymbol{\beta})} ((2\pi)^{rg}|\boldsymbol{G}|)^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{G}^{-1}\boldsymbol{W}\right\} \mathrm{d}\boldsymbol{W}\right). \tag{61}$$

In general, finding the derivative of (61) with respect to $\boldsymbol{\beta}$ requires use of the chain rule to handle the logarithm, and the Fundamental Theorem of Calculus (FTC) for the integral. However as it stands the integral is too general to make use of the FTC. The difficulty being the integrable region, $T(\boldsymbol{\beta})$. In the general case, we suggest the use of numerical integration techniques to find this derivative. Since it is the integrable region that is complicated, it may be that Monte Carlo methods are most appropriate. For now we use the normalising term in the case of Section D.3 where we have a random intercept and slope. In this case the normalising term, from (49) is

$$\eta(\boldsymbol{\beta}) = g \log\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{H,1}}\right)\right]$$

and we derive the derivative, in (50), as

$$\frac{\partial \eta}{\partial \boldsymbol{\beta}} = \left(\frac{\Phi'\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{H,1}}\right)}{\sigma_{H,1}\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{H,1}}\right)\right]}\right)\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{\beta}}$$

where $\sigma_{H,1} = \exp\{\omega_2\}$ when using the log-Cholesky decomposition. The difficulty in (50) is finding $\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{\beta}}$, for which we turn to envelope theorem[6]. The function $c(\boldsymbol{\beta})$ is the linear distance from the minimum value of $p'(\boldsymbol{\beta})$ to the x-axis. It was defined in (40) as

$$c(\boldsymbol{\beta}) = \psi_{1,1}^{-1} \min_{x \in S} p'(x; \boldsymbol{\beta}).$$

Standard envelope theorems generally apply to $\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{\beta}}$ when $S = \mathbb{R}$, however the case where S is an arbitary set is covered by Theorem 1 of Milgrom and Segal (2002). As a result the derivative of $c(\boldsymbol{\beta})$ can be written as

$$\frac{\mathrm{d}c(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = \frac{1}{\psi_{1,1}} \frac{\partial p'(x; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{x = x^*(\boldsymbol{\beta})} \tag{62}$$

[6] Schmidt (2004) attributes the origins of the envelope theorem to Auspitz and Lieben (1889) (in German) in their review, and acknowledges that these theorems are not well know outside of economics and sensitivity analysis. An early English publication of the envelope theorem is contained within Samuelson (1947), which was further extended by Afriat (1971) and others.

where $x^*(\boldsymbol{\beta}) = \arg\min_{x \in S} p'(x; \boldsymbol{\beta})$ which can be calculated for a given $\boldsymbol{\beta}$. Recall that $p'(x; \boldsymbol{\beta}) = \sum_{i=1}^{q} \beta_i p_i'(x)$ where the p_i are the orthonormal polynomials. Therefore the component form of (62) is

$$\begin{aligned}\frac{\mathrm{d}c(\boldsymbol{\beta})}{\mathrm{d}\beta_i} &= \frac{1}{\psi_{1,1}} \frac{\partial p'(x; \boldsymbol{\beta})}{\partial \beta_i}\bigg|_{x = x^*(\boldsymbol{\beta})} \\ &= \frac{p_i'(x)}{\psi_{1,1}}\bigg|_{x = x^*(\boldsymbol{\beta})} \\ &= \frac{p_i'(x^*)}{\psi_{1,1}}.\end{aligned} \tag{63}$$

Having found the derivative of $c(\boldsymbol{\beta})$ in (63) we can state (50), in component form, as

$$\frac{\partial \eta}{\partial \beta_i} = \left(\frac{\Phi'\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{H,1}}\right)}{\sigma_{H,1}\left[1 - \Phi\left(\frac{-c(\boldsymbol{\beta})}{\sigma_{H,1}}\right)\right]}\right)\frac{p_i'(x^*)}{\psi_{1,1}} \tag{64}$$

where $\psi_{1,1} = \left(\sum_{x \in D_w} (x - \bar{x})^2\right)^{-1/2}$ and $\sigma_{H,1}$ can be replaced by $\exp\{\omega_2\}$ when using a log-Cholesky decomposition.

F Berkeley growth data

The Berkeley growth dataset (BGD) (Tuddenham and Snyder, 1954) is well known in the area of growth curve analysis. It contains a set of repeated height measurements for 39 male and 54 female children over the ages 1 to 18. In all there are 2,883 observations. The measurements were taken at unequal intervals, a total of 31 times for each participant. The BGD is used to illustrate the monotonic mixed effect models with higher degree of mean and random effects (with constrained subject-specific curves). The fitted curves should be monotonic as heights of children should not decrease over time. We demonstrate the constrained fitting methods on the males in the dataset.

Figure 5 shows the degree 12 fits with subject-specific curves defined by 6 random effects. The mean curve is constrained to be monotonic over the ages 0 to 18, whilst the subject-specific curves are unconstrained. The subject-specific curves are plotted along with the mean curve and raw data for 9 male subjects. The differences between the subject-specific curves and mean curve demonstrates the flexibility that linear and parametric random effects can add.

The particular individuals in Figure 5 were chosen to demonstrate that some of the 38 subject-specific curves are non-monotonic between the ages of 16 and 18 (boy 35 is given as an example of a monotonic subject-specific fit). The data for each individual appear clearly monotonically increasing, but in the subject-specific terms we observe a decrease. This may be due to a marginally lower height recorded as one of the last observations, such as in boy 15's height data. It could also be under-fitting in that area of the data. In any case, we are able to correct for the non-monotonic fits by introducing constrained random effects adhering to monotonicity.

Figure 6 contains the previous model structure ($q = 12$, $r = 6$) but now the subject-specific curves are constrained to be monotonically increasing over the ages 0 to 18. This model is more computationally intensive than its' lower degree counterparts in Section 5, specifically because the evaluation of the random effects' mean and variance, as well as the penalty term, $\eta(\boldsymbol{\beta})$, requires Monte Carlo integration.

Incorporating this subject constraint has induced subject monotonicity but also flatter behaviour in the curves between the ages 16 to 18. This is a beneficial result as we would expect growth curves to flatten out in later years as children stop growing. It is worthwhile noting that boy 35's subject-specific fit is almost unchanged, as we would expect.

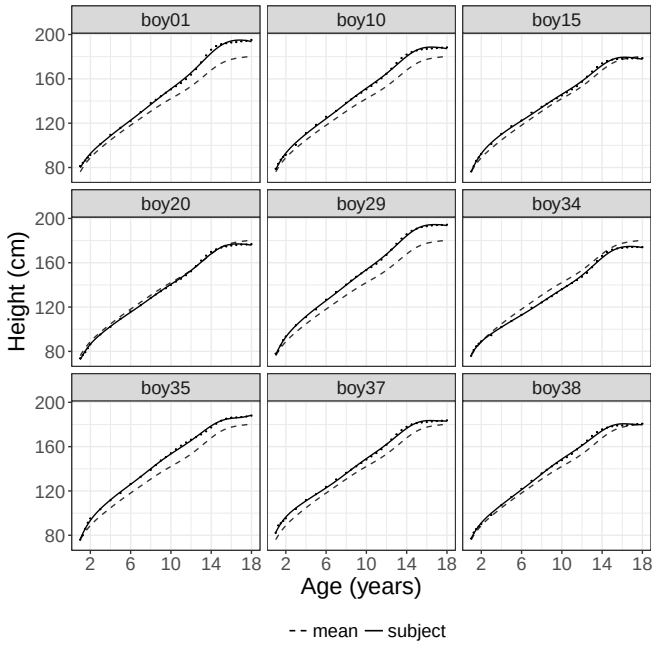

Fig. 5 Mean versus subject-specific fitted curves for degree 12 polynomials with 6 random effects terms overlaid on data points. The mean is monotone-constrained for ages [0, 18].

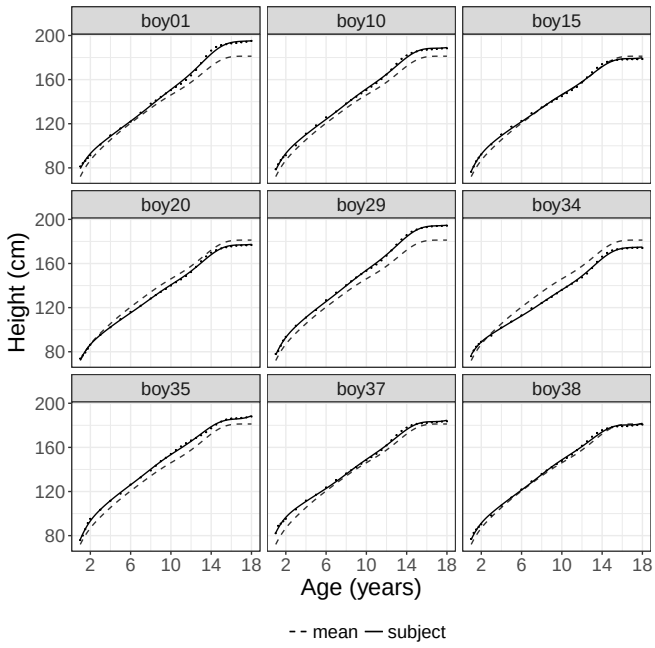

Fig. 6 Mean versus subject-specific fitted curves for degree 12 polynomials with 6 random effects terms overlaid on data points. The mean and subject-specific curves are monotone-constrained for ages [0, 18].

References

Afriat S (1971) Theory of maxima and the method of Lagrange. SIAM Journal on Applied Mathematics 20(3):343–357

Auspitz R, Lieben R (1889) Untersuchungen über die Theorie des Preises. Duncker & Humblot

Barlow RE, Brunk HD (1972) The isotonic regression problem and its dual. Journal of the American Statistical Association 67(337):140–147

Barlow RE, Bartholomew DJ, Bremner J, Brunk HD (1972) Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression. Wiley, New York

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48

Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, Russo MB, Balkin TJ (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. Journal of Sleep Research 12(1):1–12

Booth JG, Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society Series B (Statistical Methodology) 61(1):265–285

Bradley RA, Srivastava SS (1979) Correlation in polynomial regression. The American Statistician 33(1):11–14

Cassioli A, Lorenzo DD, Sciandrone M (2013) On the convergence of inexact block coordinate descent methods for constrained optimization. European Journal of Operational Research 231(2):274–281

Chen J, Zhang D, Davidian M (2002) A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics 3(3):347–360

Damien P, Walker SG (2001) Sampling truncated normal, beta, and gamma densities. Journal of Computational and Graphical Statistics 10(2):206–215

De Boor C (1978) A practical guide to splines. Springer-Verlag, New York

Dette H, Neumeyer N, Pilz KF, et al (2006) A simple nonparametric estimator of a strictly monotone regression function. Bernoulli 12(3):469–490

Dierckx IP (1980) An algorithm for cubic spline fitting with convexity constraints. Computing 24(4):349–371

Elphinstone CD (1983) A target distribution model for nonparametric density estimation. Communications in Statistics - Theory and Methods 12(2):161–198

Emerson PL (1968) Numerical construction of orthogonal polynomials from a general recurrence formula. Biometrics 24(3):695–701

Forsythe GE (1957) Generation and use of orthogonal polynomials for data-fitting with a digital computer. Journal of the Society for Industrial and Applied Mathematics 5(2):74–88

Friedman J, Tibshirani R (1984) The monotone smoothing of scatterplots. Technometrics 26(3):243–250

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109

Hawkins DM (1994) Fitting monotonic polynomials to data. Computational Statistics 9(3):233–247

Hazelton ML, Turlach BA (2011) Semiparametric regression with shape-constrained penalized splines. Computational Statistics & Data Analysis 55(10):2871–2879

Hodges JS (2013) Richly parameterized linear models: Additive, time series, and spatial models using random effects. CRC Press, Boca Raton

Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. Communications in Statistics - Theory and Methods 6(9):813–827

Hornung U (1978) Monotone spline interpolation. In: Collatz L, Meinardus G, Werner H (eds) Numerische Methoden der Approximationstheorie, Birkhäuser, Basel, pp 172–191

Horrace WC (2005) Some results on the multivariate truncated normal distribution. Journal of Multivariate Analysis 94(1):209–221

Kelly C, Rice J (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics 46(4):1071–1085

Laird N, Lange N, Stram D (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. Journal of the American Statistical Association 82(397):97–105

Lee LF (1979) On the first and second moments of the truncated multi-normal distribution and a simple estimator. Economics Letters 3(2):165–169

Leppard P, Tallis GM (1989) Algorithm AS 249: Evaluation of the mean and covariance of the truncated multinormal distribution. Journal of the Royal Statistical Society Series C (Applied Statistics) 38(3):543–553

Levine RA, Casella G (2001) Implementations of the Monte Carlo EM algorithm. Journal of Computational and Graphical Statistics 10(3):422–439

Lindstrom MJ, Bates DM (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 83(404):1014–1022

Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80(2):267–278

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21(6):1087–1092

Milgrom P, Segal I (2002) Envelope theorems for arbitrary choice sets. Econometrica 70(2):583–601

Murray K, Müller S, Turlach BA (2013) Revisiting fitting monotone polynomials to data. Computational Statistics 28(5):1989–2005

Murray K, Müller S, Turlach BA (2016) Fast and flexible methods for monotone polynomial fitting. Statistical Computation and Simulation 86(15):2946–2966

Narula SC (1979) Orthogonal polynomial regression. International Statistical Review 47(1):31–36

Pinheiro JC, Bates DM (1996) Unconstrained parametrizations for variance-covariance matrices. Statistics and Computing 6(3):289–296

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.r-project.org`

Ramsay JO (1998) Estimating smooth monotone functions. Journal of the Royal Statistical Society Series B (Statistical Methodology) 60(2):365–375

Samuelson PA (1947) Foundations of Economic Analysis. Harvard University Press, Cambridge, Massachusetts

Schmidt T (2004) Really pushing the envelope: Early use of the envelope theorem by Auspitz and Lieben. History of Political Economy 36(1):103–129

Tallis GM (1961) The moment generating function of the truncated multi-normal distribution. Journal of the Royal Statistical Society Series B (Methodological) 23(1):223–229

Tuddenham RD, Snyder MM (1954) Physical growth of California boys and girls from birth to eighteen years. Publications in child development University of California, Berkeley 1(2):183–364

Turlach BA (2005) Shape constrained smoothing using smoothing splines. Computational Statistics 20(1):81–104

Turlach BA, Murray K (2016) MonoPoly: Functions to Fit Monotone Polynomials. URL `http://cran.r-project.org/package=MonoPoly`, R package version 0.3-8

Utreras FI (1982) Convergence rates for monotone cubic spline interpolation. Journal of Approximation Theory 36(1):86–90

Utreras FI (1985) Smoothing noisy data under monotonicity constraints existence, characterization and convergence rates. Numerische Mathematik 47(4):611–625

Wilhelm S, Manjunath BG (2015) tmvtnorm: Truncated Multivariate Normal and Student t Distribution. URL `http://cran.r-project.org/package=tmvtnorm`, R package version 1.4-10

Wong Y (1935) An application of orthogonalization process to the theory of least squares. The Annals of Mathematical Statistics 6(2):53–75

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 694–699

Zimmerman DL, Núñez-Antón (2001) Parametric modelling of growth curve data: An overview. Test 10(1):1–73