Statistics and Computing (2025) 35:122
https://doi.org/10.1007/s11222-025-10628-4

ORIGINAL PAPER

Fast Bayesian inference in a class of sparse linear mixed effects models

Maria-Zafeiria Spyropoulou[1] · **James G. Hopker**[1] · **Jim E. Griffin**[2]

Received: 14 August 2024 / Accepted: 27 April 2025

Abstract
Linear mixed effects models are widely used in statistical modelling. We consider a mixed effects model with Bayesian variable selection in the random effects using spike-and-slab priors and develop a optimisation-based inference schemes that can be applied to large data sets. An EM algorithm is proposed for the model with normal errors where the posterior distribution of the variable inclusion parameters is approximated using an Occam's window approach. Placing this approach within a variational Bayes scheme allows the algorithm to be extended to the model with skew-t errors. The performance of the algorithm is evaluated in a simulation study and applied to a longitudinal model for elite athlete performance in 100 metres track sprinting and weightlifting.

Keywords Variable selection · Occam's Window · EM · Variational Bayes · Skew-t errors · Longitudinal modelling · Sport performance

1 Introduction

Linear Mixed Effects (LME) models are widely used when we have multiple observations on a sample of individuals, such as repeated measures (e.g. Lindsey 1999,), longitudinal measurements (e.g. Fitzmaurice et al. 2008) or semiparametric regression models (e.g. Ruppert et al. 2003). Suppose that there are M individuals with the i-th individual having observations $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})$, and design matrices $\boldsymbol{X}_i$ $(n_i \times q)$ and $\boldsymbol{S}_i$ $(n_i \times p)$ modelled by

$$\boldsymbol{y}_i = \zeta_0 + \boldsymbol{X}_i\,\boldsymbol{\zeta} + \boldsymbol{S}_i\,\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \tag{1.1}$$

where ζ_0 and $\boldsymbol{\zeta}$ are fixed effects whose values are shared by all individuals, $\boldsymbol{\beta}_i$ are individual-specific zero-mean random effects with covariance matrix $\boldsymbol{\Omega}$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,n_i})$ are i.i.d. zero-mean errors. In a Bayesian analysis of this model, the random effects and the errors are usually assumed to be normally distributed.

In modern applications, either p or q (or both) may be high-dimensional which has led to the use of variable selection in LME models. A prior distribution is given to $\boldsymbol{\zeta}$ and/or $\boldsymbol{\Omega}$ which encourages elements to be shrunk towards zero. Initial methodological and computational developments concentrated on the application of variable selection to both the fixed and random effects by extending methods developed for linear models. Chen and Dunson (2003) suggested using the form $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\lambda})\,\boldsymbol{B}\,\text{diag}(\boldsymbol{\lambda})$ where $\boldsymbol{B}$ is a $(p \times p)$-dimensional matrix and $\boldsymbol{\lambda}$ is p-dimensional vector. Spike-and-slab variable selection priors are placed on $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}$ and inference uses Markov chain Monte Carlo (MCMC) methods. This approach was subsequently developed to use variational Bayes (VB) inference with shrinkage priors by Armagan and Dunson (2011). More recently, work has focused on variable selection in the fixed effects only (and so the random effects are effectively treated as nuisance parameters) using VB methods. Tung et al. (2019) consider using a Bayesian lasso prior (Park and Casella 2008) for the fixed effects. This approach is extended by Degani et al. (2022) to allow for general global-local priors (Bhadra et al. 2019), more sophisticated random effect structures and to take advantages of fast matrix methods.

In contrast to previous work, we focus on Bayesian variable selection in the setting where p is high-dimensional, n_i

✉ Jim E. Griffin
j.griffin@ucl.ac.uk

Maria-Zafeiria Spyropoulou
mzs@kent.ac.uk

James G. Hopker
j.g.hopker@kent.ac.uk

[1] School of Sport and Exercise Sciences, University of Kent, Canterbury, UK

[2] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

Published online: 09 June 2025

can be large and the covariance of the random effects $\boldsymbol{\Omega}$ is assumed to be diagonal. This allows us to consider variable selection at the individual level with potentially different variables used as random effects for each individual. The use of Bayesian variable selection (rather than global-local shrinkage priors) allows the direct calculation of posterior inclusion probabilities for each random effect, which can be summarized using the median model (Barbieri and Berger, 2004). This is difficult with global-local shrinkage priors (see Ray and Bhattacharya xxxx, for a popular approach). We also consider replacing the commonly-used normal distribution for the observational errors $\epsilon_{i,j}$ with the more general skew-t distribution (Azzalini and Capitanio 2003).

Our set-up is motivated by recent work in modelling elite sporting performance over an athlete's career in events such as 100 metres track sprints or weightlifting. Interest focuses on the trajectory of an individual's sporting performance as a function of age (see *e.g.* Berry et al. 1999) and these differ between individuals due to individual physiology, injuries, training, etc. We consider the approach of Griffin et al. (2022) who apply a linear mixed effects model to large databases containing thousands of athletes with potentially hundreds of performances. The fixed effects include polynomial terms for age (providing a population effect of age), as well as environmental conditions (such as wind speed), the month of the event, or the prestige of an event. The difference between an individual's trajectory and the population effect of age is modelled using linear splines as random effects. A spike-and-slab variable selection prior is used to avoid overfitting by the linear splines and suggests that this variable selection should occur at the individual (rather than population) level.

Griffin et al. (2022) used MCMC to fit the model but this can be slow and involve substantial amounts of memory if there are a large number of observations and/or a large number of individuals.

The novelty and contribution of the paper is to develop an EM-based method (Dempster et al. 1977; Meng and van Dyk 1997) for Bayesian inference in the LME model with normal errors and a VB approach (Blei et al. 2017) for this model with skew t errors. We approximate the posterior distribution of the individual variable inclusion parameters using an Occam's window approach (Madigan and Raftery 1994) and show how this can be included in EM-type and VB-type algorithms.

The paper is organized as follows: Section 2 explains how the model is formed. In Section 3, an EM algorithm for inference in the LME model with normal errors is developed including the approximation of the posterior distribution of the variable inclusion indicators using the Occam's window approach. In section 4 an extension to non-normal error is presented which uses the Occam's window approximation in a VB algorithm and is developed for the specific case of skew t errors. Section 5 includes a simulation study and applications to 100 metres track sprinting and weightlifting data with a comparison of the algorithm to the MCMC algorithm in Griffin et al. (2022). Lastly, Section 6 concludes. Appendices gives further details of the algorithms and further results from the simulation study.

2 Sparse Linear Mixed Effects Models

We consider the LME model in (1.1) and initially assume that $\epsilon_{i,j} \overset{ind.}{\sim} \mathcal{N}\left(0, \sigma_i^2\right)$ (we will consider relaxing this assumption to a skew t error distribution in Section 4). The Bayesian variable selection approach introduces individual indicator variables $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \ldots, \gamma_{i,p})$ where $\gamma_{i,j}$ is 1 if the j-th random effect is included in the model and is 0 otherwise for the i-th individual. We write $\boldsymbol{\zeta}^\star = (\zeta_0, \boldsymbol{\zeta})$, $\boldsymbol{S}_i^\gamma$ as the design matrix including only random effects with $\gamma_{i,j} = 1$, $p_i^\gamma = \sum_{j=1}^p \gamma_{i,j}$ for the number of selected random effects and $\boldsymbol{\beta}_i^\gamma$ for the corresponding coefficients. The LME model becomes

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\zeta}^* + \boldsymbol{S}_i^\gamma \boldsymbol{\beta}_i^\gamma + \boldsymbol{\epsilon}_i.$$

The Bayesian model is completed by assigning priors to the parameters. The commonly-used beta-binomial prior is used for the inclusion variables $\boldsymbol{\gamma}_i$ so that $\gamma_{i,k} \overset{i.i.d.}{\sim} Bernoulli(h_i)$ and $h_i \sim \mathcal{B}e\,(a_1, b_1)$. This implies that

$$p(\boldsymbol{\gamma}_i) = \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)} \frac{\Gamma(p_i^\gamma + a_1)\Gamma(p - p_i^\gamma + b_1)}{\Gamma(p + a_1 + b_1)}.$$

The regression coefficients for the included variables are $\beta_{i,1}^\gamma \sim \mathcal{N}(0, \psi\,\sigma_i^2)$ and $\beta_{i,j}^\gamma \mid \gamma_i \sim \mathcal{N}(0, g^2\,\sigma_i^2)$ for $j = 2, \ldots, p_i^\gamma + 1$. We assume a vague prior for the fixed effects, $p(\boldsymbol{\zeta}^*) \propto 1$ but other choices such as a normal prior, g-prior, or global-local shrinkage prior could be used. The error variance is assumed to be $\sigma_i^2 \overset{i.i.d.}{\sim} \mathcal{IG}(a, b)$. The prior is completed by specifying the hyperpriors: $\psi \sim \mathcal{IG}(1, 1)$, where $\mathcal{IG}(a, b)$ represents an inverse-gamma distribution with shape a and mean $b/(a-1)$ if $a > 1$, $\sqrt{g} \sim \mathcal{HC}(1)$, where $\mathcal{HC}(\gamma)$ represents a half-Cauchy distribution with scale γ, $p(a_1, b_1) \propto 1$ and $p(a, b) \propto 1$.

3 EM algorithm with normal errors

We develop an approach to inference in the model in (1.1) when the sample size M is large. The parameters are divided into a set of population-level parameters $\boldsymbol{\chi} = \{\zeta_0, \boldsymbol{\zeta}, \psi, g^2, a, b, a_1, b_1\}$ and a set of individual-level parameters $\boldsymbol{\nu}_i = \{\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2\}$. The maximum a posterior (MAP) estimate of $\boldsymbol{\chi}$ integrating over the individual-level

parameters $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$ is found using the EM algorithm (Dempster et al. 1977; Meng and van Dyk 1997). To simplify the notation, we drop the γ superscript. The EM algorithm iterates the following two steps:

1. **Expectation Step**: Calculate

$$\begin{aligned} Q(\boldsymbol{\chi}) &= \sum_{i=1}^{M} \mathrm{E}_{\boldsymbol{\nu}_i|\boldsymbol{\chi},\boldsymbol{y}_i} \left[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{\nu}_i, \boldsymbol{\chi}\right) + \log P\left(\boldsymbol{\nu}_i|\boldsymbol{\chi}\right)\right] \\ &\quad + \log P\left(\boldsymbol{\chi}\right) \\ &= \sum_{i=1}^{M} \mathrm{E}_{\boldsymbol{\gamma}_i|\boldsymbol{\chi},\boldsymbol{y}_i} \\ &\quad \left[\mathrm{E}_{\boldsymbol{\beta}_i,\sigma_i^2|\boldsymbol{\gamma}_i,\boldsymbol{\chi},\boldsymbol{y}_i} \left[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\chi}\right)\right.\right. \\ &\quad \left.\left. + \log P(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2 \mid \boldsymbol{\chi})\right]\right] + \log P\left(\boldsymbol{\chi}\right) \end{aligned} \tag{3.1}$$

2. **Maximization Step**: Find $\arg\max_{\boldsymbol{\chi}} Q(\boldsymbol{\chi})$

To use the EM algorithm we need to calculate the posterior distributions of the individual-specific parameters $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$. Importantly, $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$ are conditionally independent given $\boldsymbol{\chi}$ under the posterior distribution. Taking the expectation with respect to $\boldsymbol{\nu}_i$ involves a sum over the model space parameter $\boldsymbol{\gamma}_i$ which can be challenging to calculate since it involves 2^p values. The computational time needed to evaluate this sum increases exponentially with p and the sum cannot be fully enumerated for p greater than 30. We use the Occam's window approach (Madigan and Raftery 1994) to approximate this sum, which was introduced for graphical models but can be applied to general Bayesian model averaging problems. Suppose we have L models $m_1^\star, \ldots, m_L^\star$, data $\boldsymbol{D}$, and a quantity of interest Δ with posterior predictive distribution $p(\Delta \mid \boldsymbol{D}, m_l^\star)$ for the l-th model then the Bayesian Model Averaged predictive distribution is

$$\sum_{l=1}^{L} w_l^\star \, p(\Delta \mid \boldsymbol{D}, m_l^\star) \tag{3.2}$$

where $w_l^\star \propto p(m_l^\star)\, p(\boldsymbol{D} \mid m_l^\star)$ is the model weight for the l-th model and the expectation of Δ is

$$\sum_{l=1}^{L} w_l^\star \, \mathrm{E}(\Delta \mid \boldsymbol{D}, m_l^\star) \tag{3.3}$$

where $\mathrm{E}(\Delta \mid \boldsymbol{D}, m_l^\star)$ is the posterior expectation of Δ. If L is large then this sum may contain many models for which w_l is very small and so the sums in (3.2) and (3.3) are approximated by

$$\sum_{l=1}^{K} w_k \, p(\Delta \mid \boldsymbol{D}, m_k)$$

and

$$\sum_{k=1}^{K} w_k \, \mathrm{E}(\Delta \mid \boldsymbol{D}, m_k)$$

respectively, where $K \ll L, m_1, m_2 \ldots, m_K$ are a subset of $m_1^\star, m_2^\star, \ldots, m_L^\star$ and the model weights are

$$w_k = \frac{p(m_k)\, p(\boldsymbol{D} \mid m_k)}{\sum_{t=1}^{K} p(m_t)\, p(\boldsymbol{D} \mid m_t)}.$$

We say that $m_1, m_2, \ldots, m_K$ are the models in Occam's window and by defining $m_k = m_{c_k}^\star$, we can write

$$w_k = \frac{w_{c_k}^\star}{\sum_{k=1}^{K} w_{c_k}^\star}.$$

The most accurate approximation arises when $m_1, m_2, \ldots, m_K$ are chosen to be the K models with the highest values of $w_1^\star, \ldots, w_K^\star$ (the K highest posterior probability models) and the approximation will be accurate if these K models include most of the posterior model probability, *i.e.* $\sum_{k=1}^{K} w_{c_k}^\star$ is close to 1. Although, we cannot guarantee finding these K highest probability models, we will discuss an algorithm in Section 3.2 for finding models with higher posterior probabilities.

We can use Occam's window in the EM algorithm by defining different windows for each individual and applying the approximation to the expectation in $Q(\boldsymbol{\chi})$ in (3.1). The models in Occam's window for the i-th individual are denoted $\boldsymbol{\gamma}_{i,1}, \ldots, \boldsymbol{\gamma}_{i,K}$ with associated parameters $(\boldsymbol{\beta}_{i,1}, \sigma_{i,1}^2), \ldots, (\boldsymbol{\beta}_{i,K}, \sigma_{i,K}^2)$. The expectation of a function $f(\boldsymbol{\gamma}_i)$ is approximated by the sum

$$\sum_{k=1}^{K} w_{i,k} \, f\left(\boldsymbol{\gamma}_{i,k}\right) \tag{3.4}$$

where

$$w_{i,k} = \frac{p\left(\boldsymbol{\gamma}_{i,k} \mid \boldsymbol{\chi}\right) m_i\left(\boldsymbol{\gamma}_{i,k}\right)}{\sum_{t=1}^{K} p\left(\boldsymbol{\gamma}_{i,t} \mid \boldsymbol{\chi}\right) m_i\left(\boldsymbol{\gamma}_{i,t}\right)} \tag{3.5}$$

and

$$\begin{aligned} m_i\left(\boldsymbol{\gamma}_{i,k}\right) &= p\left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\chi}\right) = \int p\left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\beta}_{i,k}, \sigma_{i,k}^2, \boldsymbol{\chi}\right) \\ &\quad p\left(\boldsymbol{\beta}_{i,k}, \sigma_{i,k}^2 \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\chi}\right) d\boldsymbol{\beta}_{i,k}\, d\sigma_{i,k}^2 \end{aligned}$$

is the marginal likelihood of $\boldsymbol{\gamma}_{i,k}$. Substituting the Occam's window approximation in (3.4) into $Q(\boldsymbol{\chi})$ in (3.1) gives

$$\begin{aligned} Q(\boldsymbol{\chi}) = & \sum_{i=1}^{M} \mathrm{E}_{\boldsymbol{\gamma}_i|\boldsymbol{\chi},\boldsymbol{y}_i} \Big[\mathrm{E}_{\boldsymbol{\beta}_i,\sigma_i^2|\boldsymbol{\gamma}_i,\boldsymbol{\chi},\boldsymbol{y}_i} \big[\log P \\ & \left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\chi} \right) + \log P(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2 \mid \boldsymbol{\chi}) \big] \Big] \\ & + \log P(\boldsymbol{\chi}) \\ = & \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma_{i,k}^2|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\ & \Big[\log P \left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\beta}_{i,k}, \sigma_{i,k}^2, \boldsymbol{\chi} \right) \\ & + \log P(\boldsymbol{\gamma}_{i,k}, \boldsymbol{\beta}_{i,k}, \sigma_{i,k}^2 \mid \boldsymbol{\chi}) \Big] + \log P(\boldsymbol{\chi}). \end{aligned} \tag{3.6}$$

To evaluate these expectations, we define residuals $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \, \boldsymbol{\zeta}^*$ and, for the k-th model, define $\boldsymbol{A}_{i,k} = \boldsymbol{B}_{i,k} \, (\boldsymbol{S}_{i,k})^T \, \boldsymbol{r}_i$, $\boldsymbol{B}_{i,k} = \left((\boldsymbol{S}_{i,k})^T \, \boldsymbol{S}_{i,k} + \boldsymbol{\Lambda}_i \right)^{-1}$, $\boldsymbol{C}_{i,k} = \boldsymbol{r}_i^T \left(I_{n_i} - \boldsymbol{S}_{i,k} \, \boldsymbol{B}_{i,k} \, \boldsymbol{S}_{i,k}^T \right) \boldsymbol{r}_i$, $a_{i,k} = a + n_i/2$, $b_{i,k} = b + \boldsymbol{C}_{i,k}/2$, $m_{i,k} = (\boldsymbol{A}_{i,k})_1$, $\boldsymbol{M}_{i,k} = (\boldsymbol{A}_{i,k})_{2:(p_{i,k}+1)}$, $q_{i,k} = (\boldsymbol{B}_{i,k})_{1,1}$ and $\boldsymbol{Q}_{i,k} = (\boldsymbol{B}_{i,k})_{2:(p_{i,k}+1),2:(p_{i,k}+1)}$ where $\boldsymbol{\Lambda}_{i,k} = \mathrm{diag}(\psi^{-1}, \underbrace{g^{-1}, \ldots, g^{-1}}_{p_{i,k}^{\gamma} times})$.

The model weight defined in (3.5) is

$$\begin{aligned} w_{i,k} & \propto p(\boldsymbol{\gamma}_{i,k}) \, p(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\chi}) \\ & = \Gamma \left(p_{i,k} + a_1 \right) \Gamma \left(p - p_{i,k} + b_1 \right) \, (g^2)^{-p_{i,k}^{\gamma}/2} \\ & \quad \mid (\boldsymbol{B}_{i,k})^{-1} \mid^{-1/2} \left(b + \boldsymbol{C}_{i,k}/2 \right)^{-(a+n_i)}. \end{aligned}$$

The posterior distribution $p(\boldsymbol{\beta}_{i,k}, \sigma_{i,k}^2 \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\chi}, \boldsymbol{y}_i)$ can be factorized as $\sigma_{i,k}^2 \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\chi}, \boldsymbol{y}_i \sim \mathcal{IG} \left(a_{i,k}, b_{i,k} \right)$ and $\boldsymbol{\beta}_{i,k} | \boldsymbol{\gamma}_{i,k}, \sigma_{i,k}^2, \boldsymbol{\chi}, \boldsymbol{y}_i \sim \mathcal{N} \left(\boldsymbol{A}_{i,k}, \sigma_{i,k}^2 \boldsymbol{B}_{i,k} \right)$. The posterior expectations that we need to evaluate $Q(\boldsymbol{\chi})$ are

$$\begin{aligned} & \mathrm{E} \left[\frac{\boldsymbol{\beta}_{i,k}}{\sigma_{i,k}^2} \right] = \frac{a_{i,k}}{b_{i,k}} \, \boldsymbol{A}_{i,k}, \quad \mathrm{E} \left[\frac{\boldsymbol{\beta}_{i,k,1}^T \boldsymbol{\beta}_{i,k,1}}{\sigma_{i,k}^2} \right] \\ & \qquad = \left(q_{i,k} + \frac{a_{i,k}}{b_{i,k}} m_{i,k}^2 \right), \\ & \mathrm{E} \left[\frac{\boldsymbol{\beta}_{i,k,2:(p_{i,k}+1)}^T \boldsymbol{\beta}_{i,k,2:(p_{i,k}+1)}}{\sigma_{i,k}^2} \right] \\ & \quad = \left(\mathrm{tr} \left(\boldsymbol{Q}_{i,k} \right) + \frac{a_{i,k}}{b_{i,k}} \left(\boldsymbol{M}_{i,k} \right)^T \left(\boldsymbol{M}_{i,k} \right) \right), \\ & \mathrm{E} \left[\frac{1}{\sigma_{i,k}^2} \right] = \frac{a_{i,k}}{b_{i,k}}, \\ & \mathrm{E} \left[- \log \sigma_{i,k}^2 \right] = \psi(a_{i,k}) - \log b_{i,k}, \end{aligned}$$

$$\mathrm{E} \left[\frac{1}{u} \right] = \frac{g^2}{1+g^2} \tag{3.7}$$

where $\psi(\cdot)$ is the digamma function.

The maximizers of $Q(\boldsymbol{\chi})$ are available in closed-form for some of the parameters. The maximizer of $\boldsymbol{\zeta}^\star$ and ψ are

$$\begin{aligned} \boldsymbol{\zeta}^\star = & \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{X}_i \sum_{k=1}^{K} w_{i,k} \, \mathrm{E} \left[\frac{1}{\sigma_{i,k}^2} \right] \right)^{-1} \\ & \left(\sum_{i=1}^{M} (\boldsymbol{X}_i)^T \sum_{k=1}^{K} w_{i,k} \left(\mathrm{E} \left[\frac{1}{\sigma_{i,k}^2} \right] \boldsymbol{y}_i - \boldsymbol{S}_{i,k} \mathrm{E} \left[\frac{\boldsymbol{\beta}_{i,k}}{\sigma_{i,k}^2} \right] \right) \right) \end{aligned}$$

and

$$\psi = \frac{1}{M+4} \left(\sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \mathrm{E} \left[\frac{\beta_{i,k,1}^2}{\sigma_{i,k}^2} \right] + 2 \right).$$

To find the maximizers of a and b, we solve the following equations:

$$\begin{aligned} & \frac{\Gamma'(a)}{\Gamma(a)} = \log b + \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \mathrm{E} \left[\log \left(\frac{1}{\sigma_{i,k}^2} \right) \right], \\ & b = \frac{a \, M}{\sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \mathrm{E} \left[\frac{1}{\sigma_{i,k}^2} \right]} \end{aligned}$$

In the same way, we update to a_1 to the maximizer of the equation

$$\begin{aligned} & \log \Gamma(a_1 + b_1) - \log \Gamma \left(p + a_1 + b_1 \right) - \log \Gamma(a_1) \\ & + \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \log \Gamma \left(p_{i,k} + a_1 \right), \end{aligned}$$

b_1 to the maximizer of the equation

$$\begin{aligned} & \log \Gamma(a_1 + b_1) - \log \Gamma \left(p + a_1 + b_1 \right) - \log \Gamma(b_1) \\ & + \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \log \Gamma \left(p - p_{i,k} + b_1 \right) \end{aligned}$$

and g to the maximizer of the equation

$$\begin{aligned} & - \log g \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, p_{i,k} - \frac{1}{g} \sum_{i=1}^{M} \sum_{k=1}^{K} w_{i,k} \, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma_{i,k}^2|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\ & \left[\frac{\sum_{j=1}^{p_{i,k}} \boldsymbol{\beta}_{i,k,j}^2}{\sigma_{i,k}^2} \right] - \log g - 2 \log(1+g). \end{aligned}$$

The calculation of these values can be speeded up by only summing over the k such that $w_{i,k} > \epsilon$ where ϵ is chosen

by the user to be small (we used $\epsilon = 0.01$ in the simulation study and real data example). This corresponds to using

$$Q(\boldsymbol{\chi}) = \sum_{i=1}^{M}\sum_{k=1}^{K} \mathrm{I}(w_{i,k} > \epsilon) w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \Big[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{i,k}, \boldsymbol{\beta}_{i,k}, \sigma^2_{i,k}, \boldsymbol{\chi}\right) + \log P(\boldsymbol{\gamma}_{i,k},\ \boldsymbol{\beta}_{i,k}, \sigma^2_{i,k} \mid \boldsymbol{\chi})\Big].$$

3.1 Initialization

The convergence and computational time of the EM algorithm can depend on initialisation of $\boldsymbol{\chi}$. We use the following scheme to initialize the elements of $\boldsymbol{\chi}$:

- $\zeta^\star = \left(\sum_{i=1}^{M} \boldsymbol{X}_i^{\star T} \boldsymbol{X}_i^\star\right)^{-1} \left(\sum_{i=1}^{M} \boldsymbol{X}_i^{\star T} \boldsymbol{y}_i^\star\right)$ where $\boldsymbol{y}_i^\star = \boldsymbol{y}_i - \bar{\boldsymbol{y}}_i$.
- To initialize b, we estimate the LME model

  $$\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\zeta}^\star = \mu_i + \boldsymbol{\epsilon}_i$$

  where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$ and set $a = 10$ and $b = a\hat{\sigma}^2$ where $\hat{\sigma}^2$ is the estimate of σ^2.
- There are $\binom{p}{k}$ possible sub-models of $\boldsymbol{S}_i$ with k variables. We find the smallest $p^\star$ for which $\sum_{k=1}^{p^\star} \binom{p}{k} \geq K$. For $i = 1, \ldots, M$, we calculate the marginal likelihood for model with $1, \ldots, p^\star$ possible variables and initialize Occam's window for the i-th individual with K models with the highest marginal likelihoods.

3.2 Updating Occam's window

The success of the algorithm depends on finding high probability models in Occam's window and a greedy search algorithm is used to achieve this. We define $\tilde{m}_i = \min\{m(\boldsymbol{\gamma}_{i,k})\}$ and update Occam's window for the i-th individual using the following steps:

1. choose a model uniformly at random. Let that model be $\boldsymbol{\gamma}_{i,k}$.
2. choose a variable j, uniformly at random from $1, \ldots, p$.
3. propose a new model $\tilde{\boldsymbol{\gamma}}$ with the values $\tilde{\boldsymbol{\gamma}}_j = 1 - \boldsymbol{\gamma}_{i,k,j}$ and $\tilde{\boldsymbol{\gamma}}_m = \boldsymbol{\gamma}_{i,k,m}$ for $m \neq j$. The new model $\tilde{\boldsymbol{\gamma}}$ will include variable j if it is excluded from $\tilde{\boldsymbol{\gamma}}_{i,k,j}$ or vice versa.
4. check that $\tilde{\boldsymbol{\gamma}}$ is not in Occam's window. If it is not, include $\tilde{\boldsymbol{\gamma}}$ in Occam's window if $m_i(\tilde{\boldsymbol{\gamma}}) > \tilde{m}_i$ by replacing the model corresponding to the value $\tilde{m}_i$ and then re-calculate $\tilde{m}_i$.

We could choose to use same number of updates of Occam's window for each athlete in each iteration of the algorithm. However, we choose to improve efficiency by only updating L models across all athletes in one iteration of the algorithm and biasing the number of updates towards individuals where changes are more likely to be accepted. Define t_i to be the number of updates since the previous successful update and at the start $t_i = 0$ for all athletes. We sample $r_i \sim \mathrm{Ex}(1 + t_i)$ and $p_i = \frac{r_i}{\sum_{j=1}^{M} r_j}$ for $i = 1, \ldots, M$. The probabilities p_i will tend to be larger for individuals who are regularly updated as for those t_i will be set to zero and hence $\mathrm{E}[r_i] = \frac{1}{1+t_i}$. Define l_i to be the number of times that Occam's window for the i-th individuals is updated and generate $l_1, \ldots, l_M$ from a multinomial distribution with total sample size L and probabilities $p_1, \ldots, p_m$.

The algorithm alternates between updating Occam's window and the population-level parameters $\boldsymbol{\chi}$.

4 Extending the algorithm to more complicated models

The algorithm developed in the previous section works for inference in the LME model in (1.1) with normal errors but the need for other error distributions has become clear over time, with skew-normal or skew-t being popular choices, for example, in medical research (Ferede et al. 2024), sports statistics (Griffin et al. 2022) and fire claims (Gong et al. 2023). We will consider skew t distributed errors (Azzalini and Capitanio 2003), which includes the skew-normal distribution as a special case, in a linear mixed model for elite athletic performances due to the occurrence of unusually poor performances. Our Occam's window approach can be extended using a VB algorithm (see Blei et al. 2017, for a review). VB algorithms have been developed for skew-normal and t distributions using latent variable representations. Wand et al. (2011) consider inference about the parameters of the skew-normal and t distributions and Guha et al. (2015) consider inference in inverse problems with skew-t errors.

We denote the skewness and degrees of freedom parameters by c and f. A convenient latent variable representation writes the model in (1.1) with skew t errors in the following form:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\, \boldsymbol{\zeta}^* + \boldsymbol{S}_i^{\gamma}\, \boldsymbol{\beta}_i^{\gamma} + \frac{c}{\sqrt{1+c^2}}\, \boldsymbol{d}_i + \boldsymbol{\epsilon}_i^* \tag{4.1}$$

where

$$\boldsymbol{\epsilon}_i^* \overset{ind.}{\sim} \mathcal{N}\left(0, \frac{\sigma_i^2}{\boldsymbol{\rho}_i} \frac{1}{1+c^2}\right), \quad \boldsymbol{d}_i \overset{ind.}{\sim} \mathcal{TN}_{[0,\infty]}\left(0, \frac{\sigma_i^2}{\boldsymbol{\rho}_i}\right),$$

$$\boldsymbol{\rho}_i \sim \mathcal{G}a\left(\frac{f}{2}, \frac{f}{2}\right)$$

and division by a vector refers to element-wise division. The introduction of latent variables $\boldsymbol{\rho}_i$ and $\boldsymbol{d}_i$ where $\boldsymbol{\rho}_i = (\rho_{i,1}, \ldots, \rho_{i,n_i})$ and $\boldsymbol{d}_i = (d_{i,1}, \ldots, d_{i,n_i})$ respectively leads to a conditionally normal linear model. We will use the prior distributions $f \sim \mathcal{G}a(2, 0.1)$ and $c \sim \mathcal{N}(0, 10^2)$.

Again, we drop the γ superscript notation. The individual-level parameters are now $\boldsymbol{v}_i = \{\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\rho}_i, \boldsymbol{d}_i\}$. The Occam's window approach developed in Section 3 can be used in a VB method for the individual-level parameters. The variational distribution is

$$q(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\rho}_i, \boldsymbol{d}_i \mid \boldsymbol{\chi}, \boldsymbol{y}_i) = q_{\boldsymbol{\eta}_i}(\boldsymbol{\rho}_i, \boldsymbol{d}_i)\, q_{\boldsymbol{\phi}_i}(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2)$$

where $\boldsymbol{\eta}_i$ and $\boldsymbol{\phi}_i$ are variational parameters. We write $\mathrm{E}_{\boldsymbol{\eta}_i}$ and $\mathrm{E}_{\boldsymbol{\phi}_i}$ as expectations with respect to $q_{\boldsymbol{\eta}_i}$ and $q_{\boldsymbol{\phi}_i}$ respectively.

We use the mean-field approximation of the distribution $q_{\psi_i}(\boldsymbol{\rho}_i, \boldsymbol{d}_i)$. Unlike Guha et al. (2015) who factorize this distribution as $q_{\boldsymbol{\rho}_i}(\boldsymbol{\rho}_i)\, q_{\boldsymbol{d}_i}(\boldsymbol{d}_i)$, we derive their joint mean-field distribution which can be factorized as $\prod_{j=1}^{n_i}\left[q_{\eta_{i,j}}(\rho_{i,j} \mid d_{i,j})\, q_{\eta_{i,j}}(d_{i,j})\right]$. The density $q_{\eta_{i,j}}(d_{i,j})$ is proportional to

$$\mathrm{I}(d_{i,j} > 0)\left(\frac{\frac{f+1}{\nu}\left(d_{i,j} - \mu\right)^2}{f+1} + 1\right)^{-\left(\frac{f+1}{2}+\frac{1}{2}\right)}$$

where

$$\mu = \frac{c\sqrt{1+c^2}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_i^2}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_{i,j}^{\gamma}\boldsymbol{\beta}_i^{\gamma}\right)\right]}{(1+c^2)\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_i^2}\right]},$$

$$\frac{1}{\nu} = \frac{1+c^2}{\lambda}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_i^2}\right],$$

which is a truncated t-distribution. The distribution $q_{\eta_{i,j}}(\rho_{i,j} \mid d_{i,j}) = \text{Gamma}\left(a_{i,j}^{\star}, b_{i,j}^{\star}\right)$ where $a = \frac{1+f}{2}$ and

$$b = \frac{1}{2}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_i^2}\left(\sqrt{1+c^2}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_i^{\gamma}\boldsymbol{\beta}_i^{\gamma}\right) - c d_{i,j}\right)^2 + \frac{f}{2}\right].$$

Expectations of functions of $\boldsymbol{\rho}_i$ and $\boldsymbol{d}_i$ needed to update the variational parameter $\boldsymbol{\phi}_i$ are approximated using Monte Carlo averages.

Occam's window can be used to define a variational distribution for $\boldsymbol{\gamma}_i$, $\boldsymbol{\beta}_i$ and σ_i^2 with variational parameters $\boldsymbol{\phi}_i = (\boldsymbol{w}_i, \boldsymbol{a}^{\star}{}_i, \boldsymbol{b}^{\star}{}_i, \boldsymbol{A}_i, \boldsymbol{B}_i)$ which has the form

$$\begin{aligned}
&p(\gamma_i) = w_{i,k}, \qquad k = 1, \ldots, K\\
&\sigma_i^2 \mid \gamma_i = k \sim \mathcal{IG}\left(a_{i,k}^{\star}, b_{i,k}^{\star}\right),\\
&\beta_i \mid \sigma_i^2, \gamma_i = k \sim \mathcal{N}(A_{i,k}, \sigma^2 B_{i,k}).
\end{aligned}$$

The mean-field variational distribution for $q_{\boldsymbol{\phi}_i}(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i, \sigma_i^2)$ is proportional to

$$\begin{aligned}
&-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]\left[\left(\sqrt{1+c^2}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^*\right)\right.\right.\\
&\left.\left.-\sqrt{1+c^2}\boldsymbol{S}_{i,j}^{\gamma}\boldsymbol{\beta}_i^{\gamma} - \frac{1}{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]} c \mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j} d_{i,j}\right]\right]^2\right.\\
&-\frac{1}{2\sigma_i^2}c^2\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j} d_{i,j}^2\right] + \frac{1}{2\sigma_i^2}c^2\sum_{j=1}^{n_i}\frac{\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j} d_{i,j}\right]^2}{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]}\\
&-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j} d_{i,j}^2\right] - n_i \log\sigma_i^2 - \frac{1}{2}\log(\psi\sigma_i^2)\\
&-\frac{1}{2}\frac{(\boldsymbol{\beta}_i^{\gamma})^T\boldsymbol{\Lambda}_i\boldsymbol{\beta}_i^{\gamma}}{\sigma_i^2} - \frac{p_i^{\gamma}}{2}\log(g^2\sigma_i^2).
\end{aligned}$$

This can be approximated using Occam's window by choosing

$$\begin{aligned}
\boldsymbol{B}_{i,k} &= \left(\boldsymbol{S}_{i,j}^{\star}{}^T\boldsymbol{S}_{i,j}^{\star} + \Lambda_i\right)^{-1},\\
\boldsymbol{A}_{i,k} &= \boldsymbol{B}_{i,k}\,\boldsymbol{S}_{i,j}^{\star}{}^T r_{i,j},\\
a_{i,k}^{\star} &= \alpha + n_i,\\
b_{i,k}^{\star} &= \sum_{j=1}^{n_i} r_{i,j}^T\left(I - \boldsymbol{S}_{i,j}^{\star}\,\boldsymbol{B}_{i,k}\,\boldsymbol{S}_{i,j}^{\star}{}^T\right) r_{i,j} + (1+c^2)\\
&\quad\times\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\, d_{i,j}^2\right] - c^2\sum_{j=1}^{n_i}\frac{\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\, d_{i,j}\right]^2}{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]},\\
w_{i,k} &\propto \Gamma(p_{i,k} + a_1)\,\Gamma(p - p_{i,k} + b_1)\, b_{i,k}^{-(\alpha+n_i)}\\
&\quad |\boldsymbol{B}_{i,k}|^{\frac{1}{2}}\,(g^2)^{-p_i^{\gamma}/2}
\end{aligned}$$

where $S_i^{\star}$ is a $(n_i \times p)$-dimensional matrix whose j-th row is $\sqrt{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]\,(1+c^2)}\, S_{i,j}$ and $r_{i,j} = \sqrt{E_{\boldsymbol{\eta}_i}[\rho_{i,j}]}\sqrt{1+c^2}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^*\right) - c\,\frac{E_{\boldsymbol{\eta}_i}[\rho_{i,j}\, d_{i,j}]}{\sqrt{E_{\boldsymbol{\eta}_i}[\rho_{i,j}]}}$ for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$. All expectations needed to update $\boldsymbol{\eta}_i$ can be calculated using the expressions in (3.7).

The global parameters $\boldsymbol{\chi}$ are estimated by maximising the Q function

$$Q(\boldsymbol{\chi}) = \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\mathrm{E}_{\boldsymbol{\phi}_i}\left[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{v}_i, \boldsymbol{\chi}\right) + \log P(\boldsymbol{v}_i \mid \boldsymbol{\chi})\right]\right] + \log P(\boldsymbol{\chi}). \tag{4.2}$$

The algorithm updates parameters in three blocks

1. Update $\boldsymbol{\chi}$ by finding the values that maximize $Q(\boldsymbol{\chi})$ in (4.2).
2. Simulate $\boldsymbol{\rho}_i$ and $\boldsymbol{d}_i$ for $i = 1, \ldots, M$ and calculate Monte Carlo average of functions of $\boldsymbol{\rho}_i$ and $\boldsymbol{d}_i$ needed to evaluate model probabilities.
3. Update Occam's window.

5 Examples and Illustrations

5.1 Simulation Study

To understand the performance of the algorithms. We performed a simulation study using both the model with skew t errors and different values of the model parameters. In each data set, there were $n = 300$ individuals, $q = 6$ fixed effects and $p = 10$ or $p = 20$ random effects. We choose $\zeta_0 = 0$ and $\zeta_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, 5$ as the coefficients of the fixed effects. For the i-th individual we choose between a large number of observations $n_i = 200$ with probability q and a small number of observations $n_i = 50$ otherwise. The regressors are independent with $X_{i,j} \sim \mathcal{N}(0, 1)$ and $S_{i,j} \sim \mathcal{N}(0, 1)$. The error variance is generated $\sigma_i^2 \sim \mathcal{IG}(10, 0.1)$ and the coefficients of the random effects are independent with $\beta_{i,k} \sim s_{i,k}\,\mathcal{N}(0, 1) + (1 - s_{i,k})\,\delta_{\beta_{i,k}=0}$ where $s_{i,k} \sim Bernoulli(h)$.

The data sets are formed by different combinations of the following parameters. We consider settings which have small ($p = 10$) or large ($p = 20$) number of random effects with the sparser ($h = 0.1$) or denser ($h = 0.25$) coefficients. The proportion of larger data sets for individuals is either $q = 0.15$ or $q = 0.3$. We consider a symmetric version ($c = 0$) or fairly heavily skewed ($c = 4$) and heavy-tailed ($f = 5$) and close to normal tails ($f = 20$). This leads to 5 different parameters with 2 possible values leading to 32 combinations. All results are calculated using 30 replicate data sets and with two possible values of Occam's window $K = 30$ and $K = 100$.

We compare the performance of the algorithm for variable selection, prediction and estimation of the skewness c

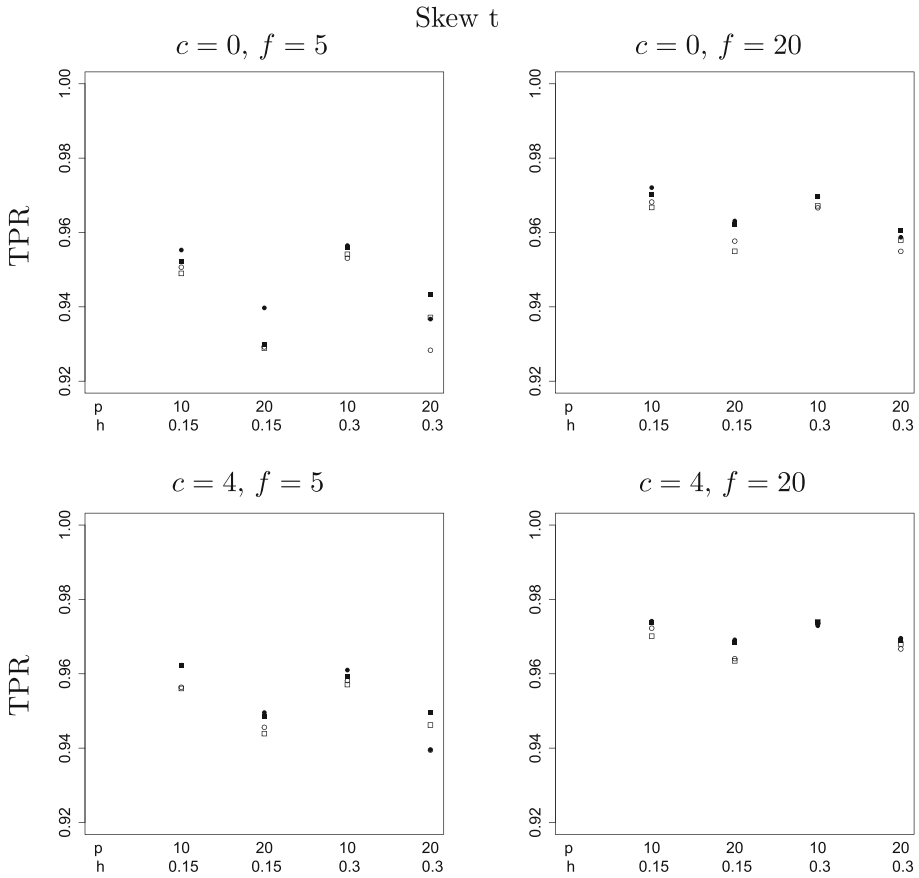

Fig. 1 Simulation study: The average true Positive Rate (TPR) for $\boldsymbol{\gamma}_i$. The symbols represent $q = 0.15$ and window = 30 (□), $q = 0.15$ and window = 100 (○), $q = 0.3$ and window = 30 (■), and $q = 0.3$ and window = 100 (●)

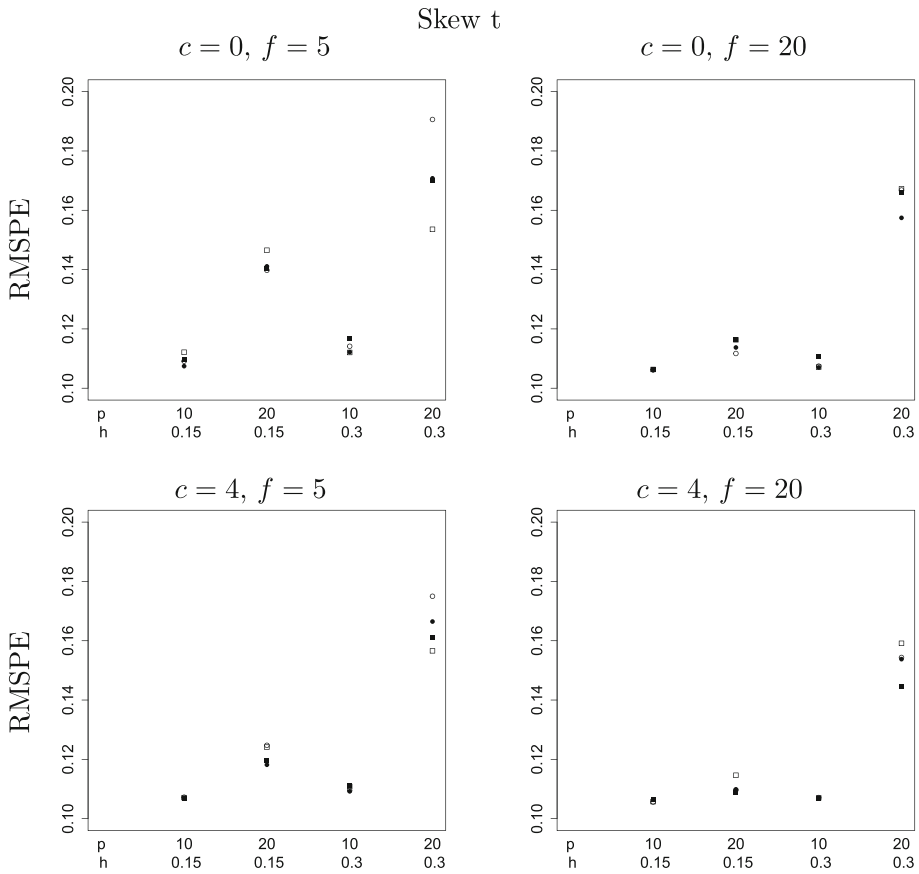

Fig. 2 Simulation study: Average Root Mean Squared Prediction Error. The symbols represent $q = 0.15$ and window = 30 (□), $q = 0.15$ and window = 100 (○), $q = 0.3$ and window = 30 (■), and $q = 0.3$ and window = 100 (●)

and the degrees of freedom f. The variable selection performance is assessed using the median model (Barbieri and Berger 2004) for each individual, that is the model including the k-th variable if $p(\gamma_{i,k} \mid \mathbf{y}^{(j)}) > 0.5$ for the i-th athlete and j-th replicate. To measure the accuracy of the median model, we calculate the true positive rate (TPR), the probability of an important variable being selected in the median model, and the true negative rate (TNR), the probability of a redundant variable not being selected in the median model. We assess predictive performance by generating a validation set of 1000 observations for each athlete in each replicate. We calculate the root mean squared prediction error (RMSPE) averaged over observations, athletes and replicates. The ability to estimate the skewness c and degrees of freedom f are measured using the root mean squared error (RMSE).

The TPR's are shown in Figure 1. It is high (above 0.92) for all settings of simulation parameters. Some of these parameters have stronger effects than others. The TPR is higher with smaller p, larger data sets and lighter tails. These results are not surprising since the variable selection problem becomes easier with smaller p or more information. It is not clear why heavier tails has an effect. The other simulation parameters (skewness, window size and proportion of redundant variables) have very little effect on performance. We also calculated the TNR for all settings of the simulation parameters but these were all very close to one and show the method can effectively identify redundant variables.

The RMSPEs for different simulation parameter settings are shown in Figure 2. The mean error variance is $\mathrm{E}[\sigma_i^2] = 0.011$ and so the RMSPE cannot be lower than approximately $\sqrt{0.011} = 0.105$. The RMSPE are influenced by both variable selection accuracy and estimation of the regression coefficients for important variables. For this reason, the effects of the simulation parameters are different from the effects of the TPR. The method provides RMSPE's close to the approximate minimum for most simulation parameter settings. Unsurprisingly, the number of variables p plays a role here and the RMSPE is larger for larger p. The other simulation parameter play a much smaller role here apart from when $p = 20$ and $h = 0.15$ when more skewness (larger c) and lighter tails (larger f) lead to smaller RMSPE.

The RMSE for the skewness and degrees of freedom are given in Appendix 7. The skewness is well-estimated for both $c = 0$ and $c = 4$. The degrees of freedom is unsurprisingly harder to estimate, but we show good performance when $c = 0$. With $c = 4$, the errors are larger but the estimated error distribution are often similar.

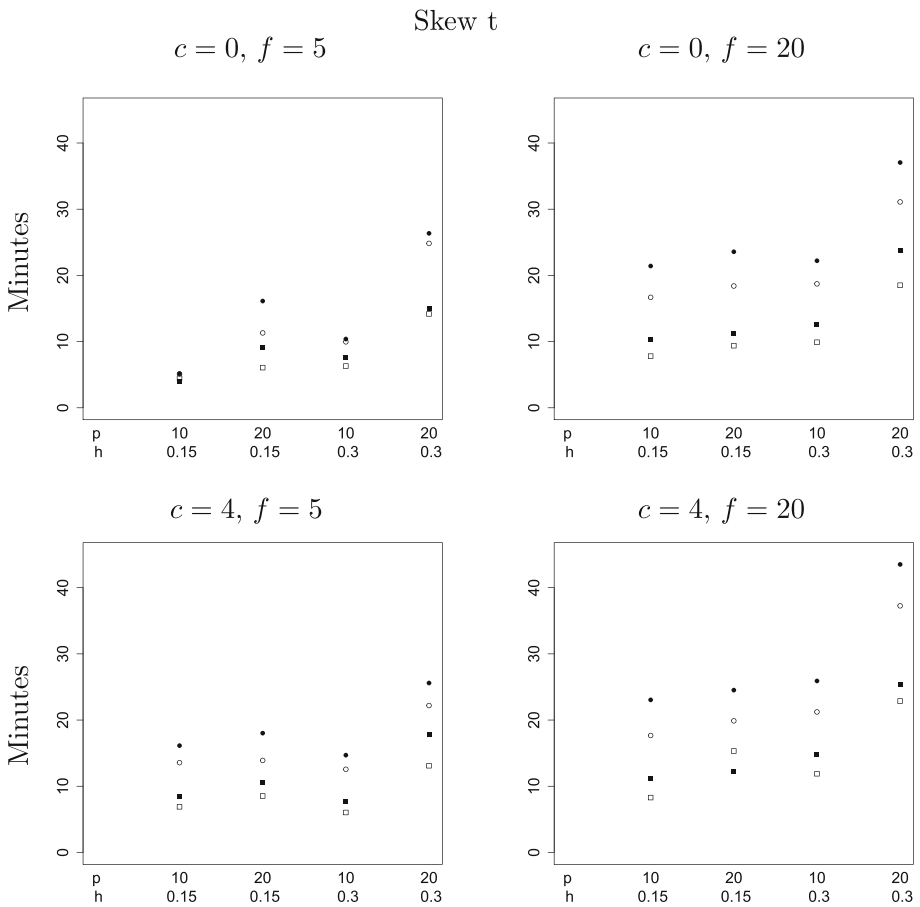

Fig. 3 Simulation study: Average time in minute. The symbols represent $q = 0.15$ and window = 30 (□), $q = 0.15$ and window = 100 (○), $q = 0.3$ and window = 30 (■), and $q = 0.3$ and window = 100 (●)

5.2 Application to modelling elite sporting performance trajectories

Griffin et al. (2022) develop an LME model for elite sporting performance that models the evolution of an athlete's performance as a function of their age. These trajectories have several applications, including identifying athletes with the potential for future success in order to prioritize funding and resource allocation, setting performance goals, or guiding training programs. An LME allows us to account for the irregular intervals between observations, confounders (such as wind speed in track sprinting, seasonality effects, and competition level), and to adjust for selection effects due to ability differences across athletes (for example, exceptional athletes often compete in international competitions at a younger age).

Griffin et al. (2022) use the LME model in (1.1) where the response $y_{i,j}$ is the j-th performance of the i-th athlete. The fixed effects design matrix $\boldsymbol{X}_i$ contains the first four powers of age, which leads to a flexible population-level age effect, and any confounders (for example, wind speed in 100 metres sprints). The random effects design matrix $\boldsymbol{S}_i$ contains linear splines basis function, which allows a flexible form for the individual performance trajectories. Variables selection allows us to use a large number of basis function p (we use $p = 100$) whilst avoiding overfitting. Skew t errors are more appropriate here since athletes usually underperform by a larger amount than they overperform.

We compare the performance of our VB method for the LME model with skew-t errors to the results in Griffin et al. (2022), which used MCMC. The data is from 100 metres track sprinting and weightlifting for both females and males. For 100 metres track sprinting, there are a total of 48,999 observations for 1297 female athletes and a total of 95,376 observations for 2,834 male athletes. For weightlifting there was a total of 11,690 observations for 1,212 female athletes and 14,557 observation for 1609 male athletes. In the 100 metres track data, we also include wind speed and the month of the race as confounders since we can have a large effect on performance. There are no confounders for the weightlifting data. We fit the LME to the data for 100 metres sprinting and weightlifting separately for male and female athletes leading to four regressions.

The computational times of VB and MCMC are provided in Table 1. They show that the VB algorithm is around 5 or 6 times faster than the MCMC method. The table also provides the estimates of the skewness and degrees of freedom from the VB algorithm and the posterior means calculated

Table 1 VB and MCMC estimates of c and f and timings (in minutes) for male and female 100 metres track sprinting and weightlifting

		VB			MCMC		
		Time	c	f	Time	c	f
100 metres	Males	50	0.9	25	> 360	1.21	18.4
	Females	46	0.9	20.4	> 360	1.35	19.6
Weightlifting	Males	59	−1.09	18.4	> 360	−1.91	7.8
	Female	75	−1.1	5.8	> 360	−1.64	6.9

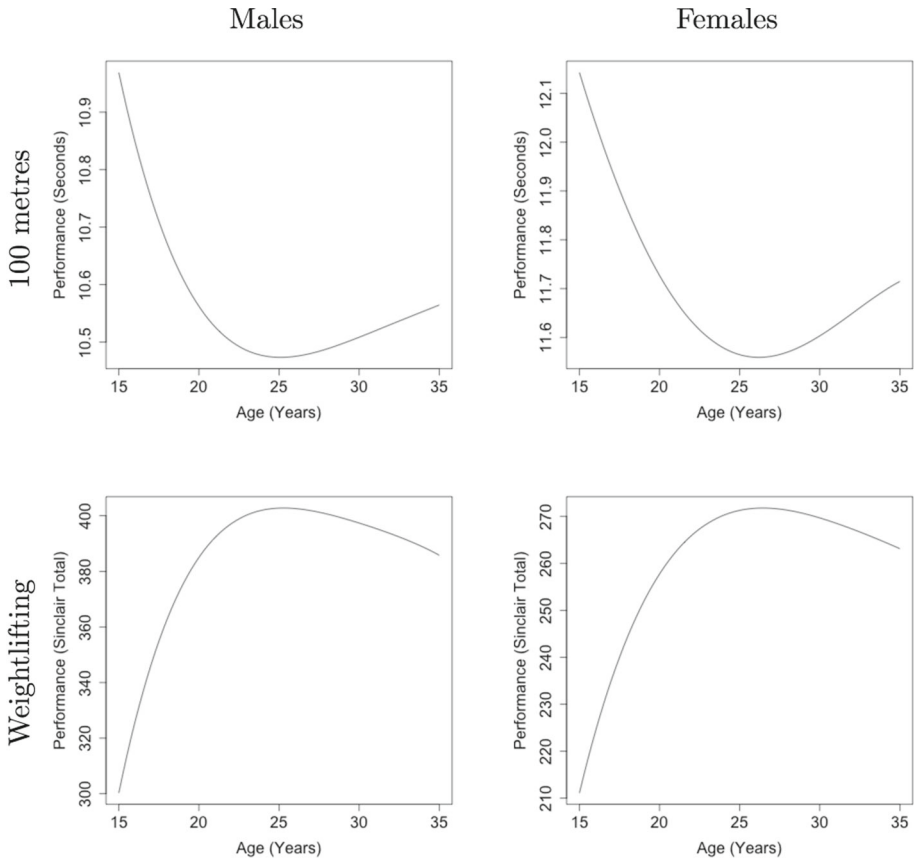

Fig. 4 Fitted population performance trajectories for both males and females in 100 metres (top row) and weightlifting (bottom row) estimated using the VB algorithm with skew t errors

using MCMC. Overall, the estimates are very similar. Specifically, in the 100 metres, the skewness is estimated to be 0.7 for males and 0.9 for females using VB, whereas they are 1.21 for males and 1.35 for females using MCMC. The degrees of freedom is estimated to be 19.05 for males and 20.4 for females using VB which are close to the estimates with MCMC (18.4 for males and 19.6 for females). For the weightlifting dataset, the skewness estimates are -0.6 and -1.1 for males and females respectively with VB whereas, for MCMC, these are -1.91 and -1.64. Although we can observe a difference in estimates between the two methods both give a negative sign. Finally, the degree of freedom estimate for females is close (5.8 with VB and 6.9 with MCMC) but the estimates for men are more different (17 for VB and 7.8 for MCMC). Overall, the VB algorithm tends to give similar results to MCMC for these parameters.

The population performance trajectory, individual level performance trajectories and excess performance play key roles in the modelling approach and are interesting to practitioners. Figure showing inference about these functions using MCMC are given in Griffin et al. (2022).

Plots of the population performance trajectory (modelled by the polynomial terms of age) are shown in Figure 4. Improving performance corresponds to faster times in 100 metres and larger weights lifted in weightlifting. The trajectories show a familiar result, also observed by Griffin et al. (2022), that, on average, performance quickly improves until the mid-twenties when performance starts to slowly deteriorate.

Excess performance measures the difference between an athlete's individual performance trajectory and the population (average) performance trajectory. This can be important to understand how an athlete is performing relative to their

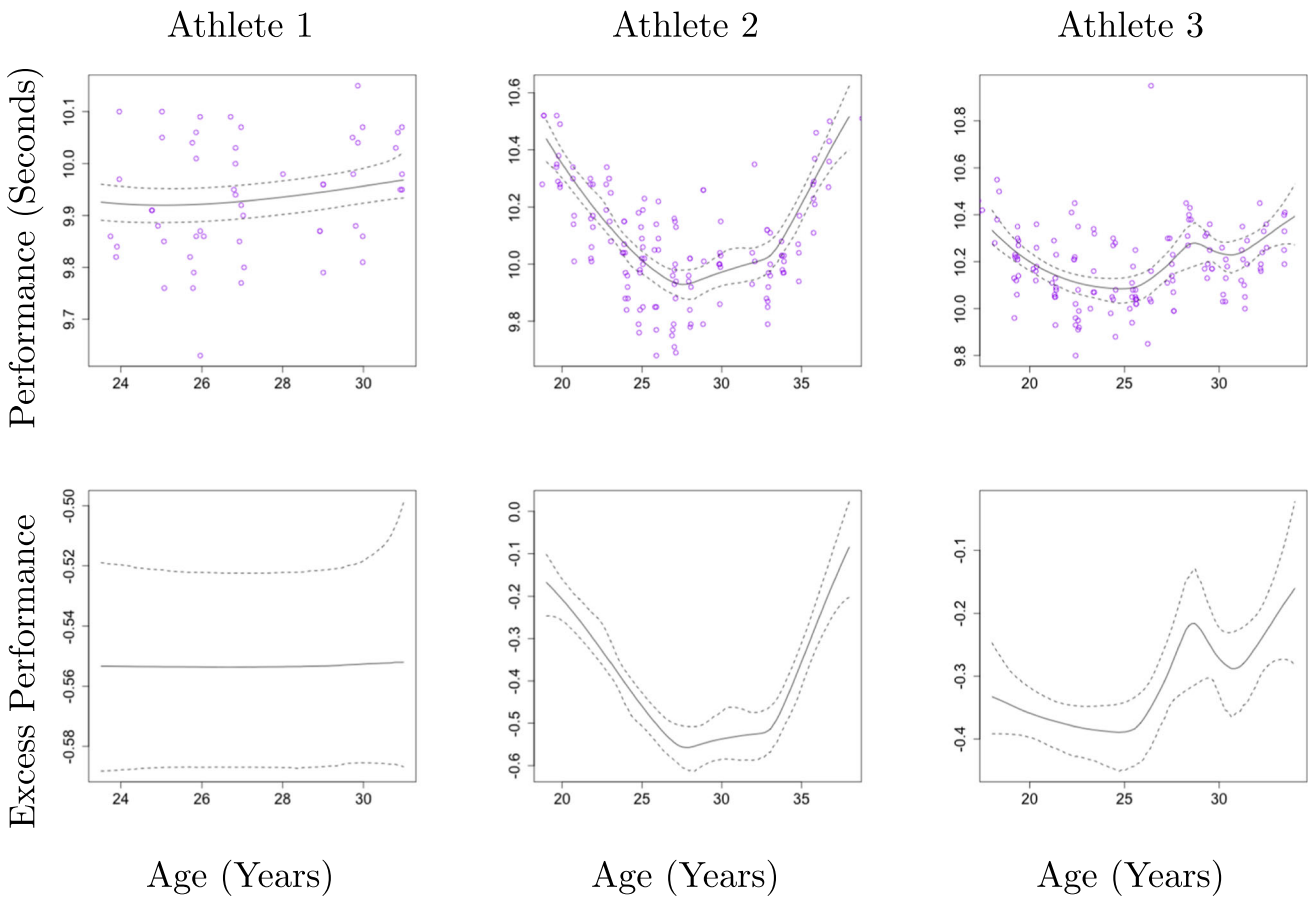

Fig. 5 Individual performance (with observed performances) (top row) and excess performance trajectories (bottom row) for 100 metres for males, shown as posterior median (solid line) and 95% central credible interval (dashed lines)

age-matched peers and is important for evaluating the success of training or for understanding the variability of individual athlete performance around the population trend. In the linear spline model, this is the overall effect of the random effects evaluated at different ages (see Griffin et al. 2022 for more details).

We present fitted individual performance trajectories and excess performance for some athletes in 100 metre sprinting, shown as posterior median with 95% credible interval, in Figures 5 (for males) and 6 (for females). We can see that the VB method is able to account for the differences in the individual performance trajectories, provide appropriate functional fits of the excess performance and sensible posterior credible intervals. Athlete 1's posterior median individual performance trajectory is fairly constant (consistently 0.5 sec faster than the rest of population) but has substantial variability within a year. The observations for Athlete 2 and Athlete 3 cover a longer time span and show more variation in their performance levels. Athlete 2's individual performance trajectory shows faster improvement around the age of 20 and later peak than Athlete 3. The 95% credible intervals give a reasonable measure of uncertainty. Athlete 1 has only 54 observations leading to a much larger credible interval than Athlete 2 (who has 141 observations) and Athlete 3 (who has 139 observations). Figure 6 shows the same results for three female 100m sprinters. These show similar results to the men. The fitted individual trajectories are able to track the observations. There is substantial differences between the excess performance trajectories which indicate that the variable selection at an individual level is needed to model these data using linear splines. The credible intervals also give a sensible measure of uncertainty.

Appendix 8 includes the individual performance and excess performance trajectories for weightlifting in both male and female populations.

6 Conclusions

We develop an EM algorithm for LME with normal errors which is extended to an LME with skew t errors using a VB approach. The latter approach could be extended to other non-normal error distributions using a latent variable representation which leads to a conditional normal linear model. A

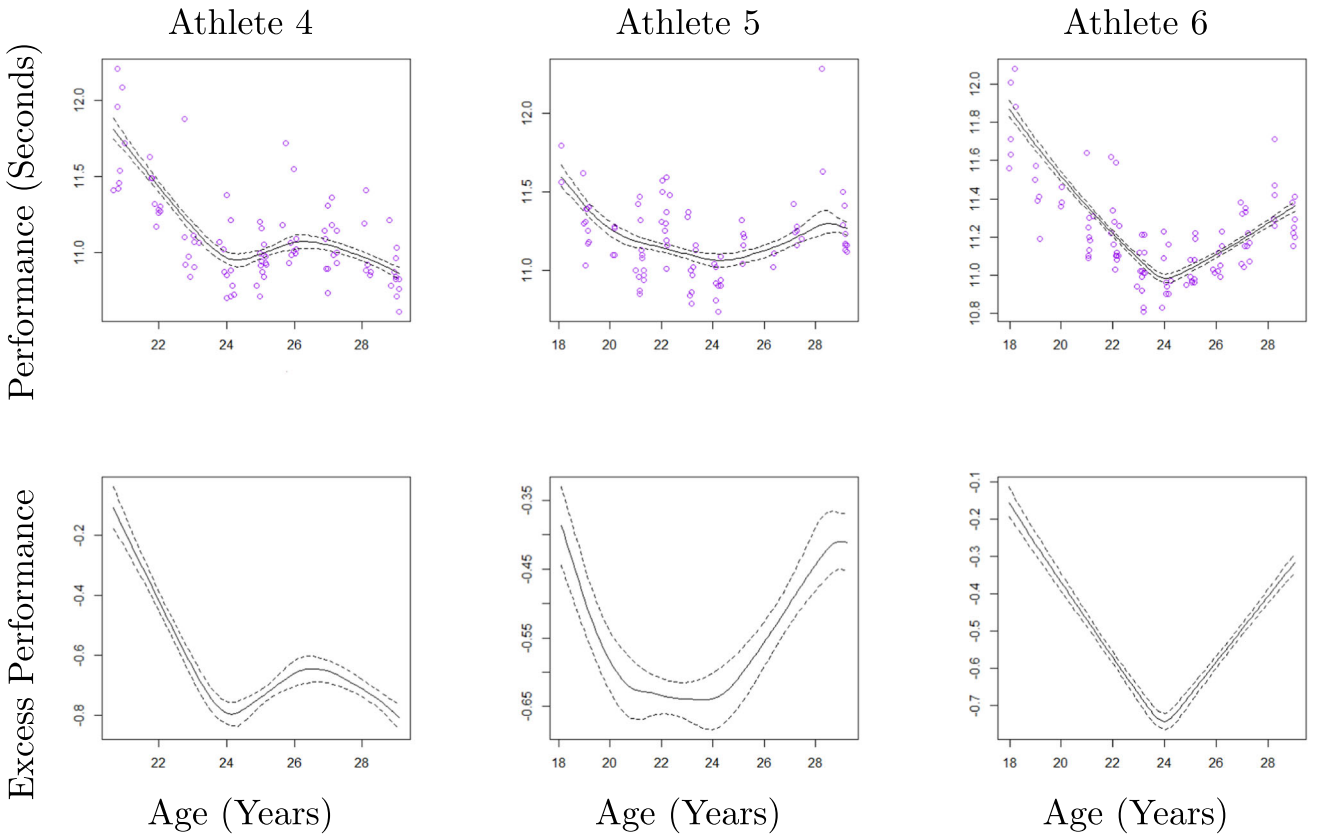

Fig. 6 Individual (with observed performances) (top row) and excess performance trajectories (bottom row) for female 100 metre sprinters, shown as posterior median (solid line) and 95% central credible interval (dashed lines)

simulation study shows that the VB algorithm has good performance for a range of values of the degrees of freedom and skewness. An application of the algorithm to a longitudinal model used in the modelling of elite sporting performance show that the method has similar inference to MCMC on a real-world data set.

Appendix A EM calculations with normal errors

The full expression of (3.6) is

$$
\begin{aligned}
Q(\boldsymbol{\chi}) &= \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\
&\quad \left[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{\beta}_{i,k}, \sigma^2_{i,k}, \boldsymbol{\chi}\right) + \log P\left(\boldsymbol{\beta}_{i,k}, \sigma^2_{i,k} \mid \boldsymbol{\chi}\right)\right] \\
&\quad + \log P(\boldsymbol{\chi}) \\
&= -\frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\
&\quad \left[\frac{1}{\sigma^2_{i,k}}\sum_{j=1}^{n_i}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_{i,j}\boldsymbol{\beta}_{i,k}\right)^2 + \log\sigma^2_{i,k}\right] \\
&\quad - \frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\
&\quad \left[\log(\psi\,\sigma^2_{i,k}) + p_{i,k}\,\log(g\,\sigma^2_{i,k}) + \frac{1}{\sigma^2_{i,k}}(\boldsymbol{\beta}_{i,k})^T\,\boldsymbol{\Lambda}_{i,k}\,\boldsymbol{\beta}_{i,k}\right] \\
&\quad + M\,a\,\log b - M\,\log\Gamma(a) - (a+1) \\
&\quad \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i}\left[\log\sigma^2_{i,k}\right] \\
&\quad - b\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i}\left[\frac{1}{\sigma^2_{i,k}}\right] \\
&\quad + M\left(\log\Gamma(a_1+b_1) - \log\Gamma\left(p + a_1 + b_1\right)\right) \\
&\quad - M\left(\log\Gamma(a_1) + \log\Gamma(b_1)\right) + \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k} \\
&\quad \left(\log\Gamma\left(p^{\gamma}_{i,k} + a_1\right) + \log\Gamma\left(p - p^{\gamma}_{i,k} + b_1\right)\right) \\
&\quad - 2\log\psi - \frac{1}{\psi} - \frac{1}{2}\log g - \log(1+g)
\end{aligned}
$$

where $\boldsymbol{\Lambda}_{i,k} = \text{diag}\left(\psi^{-1}, \underbrace{g^{-1}, \ldots, g^{-1}}_{p_{i,k}^{\gamma} times}\right)$.

To work out the expectations, it's useful to note that if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}, \qquad \mathrm{E}[\boldsymbol{X}^T \boldsymbol{B} \boldsymbol{X}] = \mathrm{tr}(\boldsymbol{B}\boldsymbol{\Sigma}^T) + \boldsymbol{\mu}^T \boldsymbol{B} \boldsymbol{\mu}$$

Appendix B EM calculations with skew-*t* errors

The full expression of (4.2) is

$$\begin{aligned}
Q(\boldsymbol{\chi}) &= \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, \mathrm{E}_{\boldsymbol{\eta}_i}\left[\mathrm{E}_{\boldsymbol{\phi}_i}\left[\log P\left(\boldsymbol{y}_i \mid \boldsymbol{v}_i, \boldsymbol{\chi}\right)\right.\right. \\
&\quad \left.\left. + \log P(\boldsymbol{v}_i)\right]\right] + \log P(\boldsymbol{\chi}) \\
&= -\sum_{i=1}^{M}\sum_{k=1}^{K} \frac{w_{i,k}}{2}\, \mathrm{E}_{\boldsymbol{\eta}_i}\mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_{i,k}^2}\sum_{j=1}^{n_i}\rho_{i,j}\right. \\
&\quad \left.\times \left(\sqrt{1+c^2}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_{i,j}\boldsymbol{\beta}_i\right) - c\, d_{i,j}\right)^2\right] \\
&\quad - \sum_{i=1}^{M}\sum_{k=1}^{K} \frac{w_{i,k}}{2}\, \mathrm{E}_{\boldsymbol{\eta}_i}\mathrm{E}_{\boldsymbol{\phi}_i} \\
&\left[\sum_{j=1}^{n_i}\left(\log\left(\frac{\sigma_{i,k}^2}{(1+c^2)\rho_{i,j}}\right) + \log\left(\frac{\sigma_{i,k}^2}{\rho_{i,j}}\right)\right) + \frac{\boldsymbol{\rho}_i^T \boldsymbol{d}_i^2}{\sigma_{i,k}^2}\right] \\
&- \sum_{i=1}^{M}\sum_{k=1}^{K} \frac{w_{i,k}}{2}\, \mathrm{E}_{\boldsymbol{\phi}_i} \\
&\left[\log(\psi\,\sigma_{i,k}^2) + \frac{\boldsymbol{\beta}_{i,k}^T \boldsymbol{\Lambda}_i \boldsymbol{\beta}_{i,k}}{\sigma_{i,k}^2} + \frac{p_{i,k}}{2}\log(g^2\,\sigma_{i,k}^2)\right] \\
&+ M\,(a\log b - \log\Gamma(a)) - (a+1)\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k} \\
&\mathrm{E}_{\boldsymbol{\phi}_i}\left[\log\sigma_{i,k}^2 - b\,\frac{1}{\sigma_{i,k}^2}\right] + M\,(\log\Gamma(a_1+b_1) \\
&- \log\Gamma\,(p+a_1+b_1) - \log\Gamma(a_1) - \log\Gamma(b_1)) \\
&+ \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\left(\log\Gamma\left(p_{i,k}+a_1\right) + \log\Gamma\left(p-p_{i,k}+b_1\right)\right) \\
&+ \left(\frac{f}{2}\log\left(\frac{f}{2}\right) - \log\Gamma\left(\frac{f}{2}\right)\right)\sum_{i=1}^{M} n_i \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\frac{f-2}{2}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\log\rho_{i,j}\right] - \frac{f}{2}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\right]\right) \\
&- 2\log\psi - \frac{1}{\psi} - \frac{1}{2}\log g - \log(1+g) \\
&- \frac{1}{2}\frac{c^2}{100^2} + \log f - 0.1 f
\end{aligned}$$

where $\boldsymbol{\Lambda}_i = \text{diag}\left(\psi^{-1}, \underbrace{(g^2)^{-1}, \ldots, (g^2)^{-1}}_{p_i^{\gamma} times}\right)$.

It is useful to define $X_{i,j,m}^{\star} = \sqrt{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]\,(1+c^2)}\, X_{i,j,m}$ and $r_{i,j} = \sqrt{E_{\boldsymbol{\eta}_i}[\rho_{i,j}]}\sqrt{1+c^2}\, y_{i,j} - c\, E_{\boldsymbol{\eta}_i}[\rho_{i,j}\, d_{i,j}]$ for $i = 1, \ldots, M$, $j = 1, \ldots, n_i$ and $m = 1, \ldots, q$ and $S_{i,j,m}^{\star} = \sqrt{\mathrm{E}_{\boldsymbol{\eta}_i}[\rho_{i,j}]\,(1+c^2)}\, S_{i,j,m}$ for $i = 1, \ldots, M$, $j = 1, \ldots, n_i$ and $m = 1, \ldots, p_{i,k}$, and $\boldsymbol{r}_i = (r_{i,1}, \ldots, r_{i,n_i})$

The algorithm uses the following updates

$$\begin{aligned}
\boldsymbol{\zeta}^{\star} &= \left(\sum_{i=1}^{M} \boldsymbol{X}_i^{\star T}\boldsymbol{X}_i^{\star}\sum_{k=1}^{K} w_{i,k}\,\mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_{i,k}^2}\right]\right)^{-1} \\
&\left(\sum_{i=1}^{M} \boldsymbol{X}_i^{\star T}\sum_{k=1}^{K} w_{i,k}\left(\mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{1}{\sigma_{i,k}^2}\right]\boldsymbol{r}_i - \mathrm{E}_{\boldsymbol{\phi}_i}\left[\frac{\boldsymbol{S}_{i,k}^{\star}\boldsymbol{\beta}_{i,k}}{\sigma_{i,k}^2}\right]\right)\right)
\end{aligned}$$

and

$$\psi = \frac{1}{M+4}\left(\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\mathrm{E}\left[\frac{\beta_{i,k,1}^2}{\sigma_{i,k}^2}\right] + 2\right),$$

To find the maximizers of a and b, we solve the following equations:

$$\begin{aligned}
\frac{\Gamma'(a)}{\Gamma(a)} &= \log b + \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\mathrm{E}\left[\log\left(\frac{1}{\sigma_{i,k}^2}\right)\right], \\
b &= \frac{a\,M}{\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\mathrm{E}\left[\frac{1}{\sigma_{i,k}^2}\right]}
\end{aligned}$$

In the same way, we update to a_1 to the maximizer of the equation

$$\begin{aligned}
&\log\Gamma(a_1+b_1) - \log\Gamma\,(p+a_1+b_1) - \log\Gamma(a_1) \\
&\quad + \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\log\Gamma\left(p_{i,k}+a_1\right),
\end{aligned}$$

b_1 to the maximizer of the equation

$$\begin{aligned}
&\log\Gamma(a_1+b_1) - \log\Gamma\,(p+a_1+b_1) \\
&\quad - \log\Gamma(b_1) + \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\,\log\Gamma\left(p-p_{i,k}+b_1\right),
\end{aligned}$$

g to the maximizer of the equation

$$
\begin{aligned}
&-\log g \sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k}\, p_{i,k} - \frac{1}{g}\sum_{i=1}^{M}\sum_{k=1}^{K} w_{i,k} \\
&\mathrm{E}_{\boldsymbol{\beta}_{i,k},\sigma^2_{i,k}|\boldsymbol{\gamma}_{i,k},\boldsymbol{\chi},\boldsymbol{y}_i} \\
&\left[\frac{\sum_{j=1}^{p_{i,k}} \boldsymbol{\beta}^2_{i,k,j}}{\sigma^2_{i,k}}\right] - \log g - 2\log(1+g),
\end{aligned}
$$

c to the maximizer of the equation

$$
\begin{aligned}
&-(1+c^2)\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\right]\sum_{k=1}^{K}\frac{w_{i,k}}{2}\mathrm{E}_{\boldsymbol{\phi}_i} \\
&\left[\frac{1}{\sigma^2_{i,k}}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_{i,j}\boldsymbol{\beta}_i\right)^2\right] \\
&+2c\sqrt{1+c^2}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\, d_{i,j}\right]\sum_{k=1}^{K}\frac{w_{i,k}}{2}\,\mathrm{E}_{\boldsymbol{\phi}_i} \\
&\left[\frac{1}{\sigma^2_{i,k}}\left(y_{i,j} - \boldsymbol{X}_{i,j}\boldsymbol{\zeta}^* - \boldsymbol{S}_{i,j}\boldsymbol{\beta}_i\right)\right] \\
&-c^2\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\, d^2_{i,j}\right]\sum_{k=1}^{K}\frac{w_{i,k}}{2}\mathrm{E}_{\boldsymbol{\phi}_i} \\
&\left[\frac{1}{\sigma^2_{i,k}}\right] \\
&+\log\left(1+c^2\right)\sum_{i=1}^{M} n_i - \frac{1}{2}\frac{c^2}{100^2}
\end{aligned}
$$

and f to the maximizer of the equation

$$
\begin{aligned}
&+\left(\frac{f}{2}\log\left(\frac{f}{2}\right) - \log\Gamma\left(\frac{f}{2}\right)\right)\sum_{i=1}^{M} n_i \\
&+\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\frac{f-2}{2}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\log\rho_{i,j}\right] - \frac{f}{2}\mathrm{E}_{\boldsymbol{\eta}_i}\left[\rho_{i,j}\right]\right) \\
&+\log f - 0.1 f.
\end{aligned}
$$

7 Further simulation results

Figures 7 and 8 show the RMSEs for the parameters c and f in the simulation study.

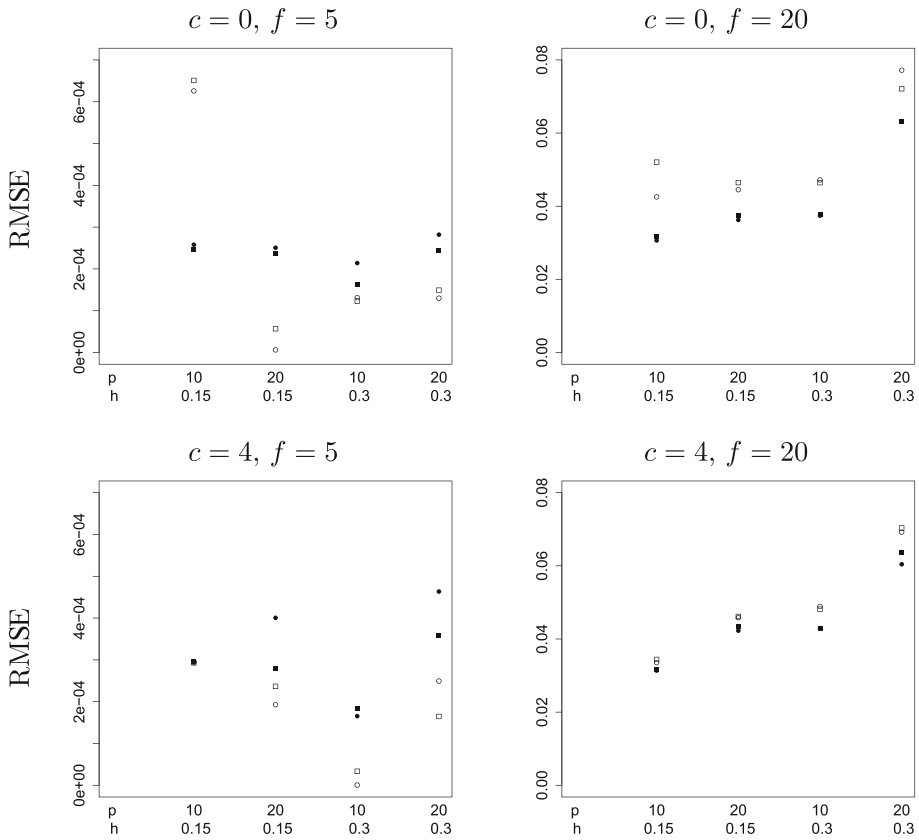

Fig. 7 Simulation study: Average Root Mean Squared Error for c. The symbols represent $q = 0.15$ and window = 30 (□), $q = 0.15$ and window = 100 (○), $q = 0.3$ and window = 30 (■), and $q = 0.3$ and window = 100 (●)

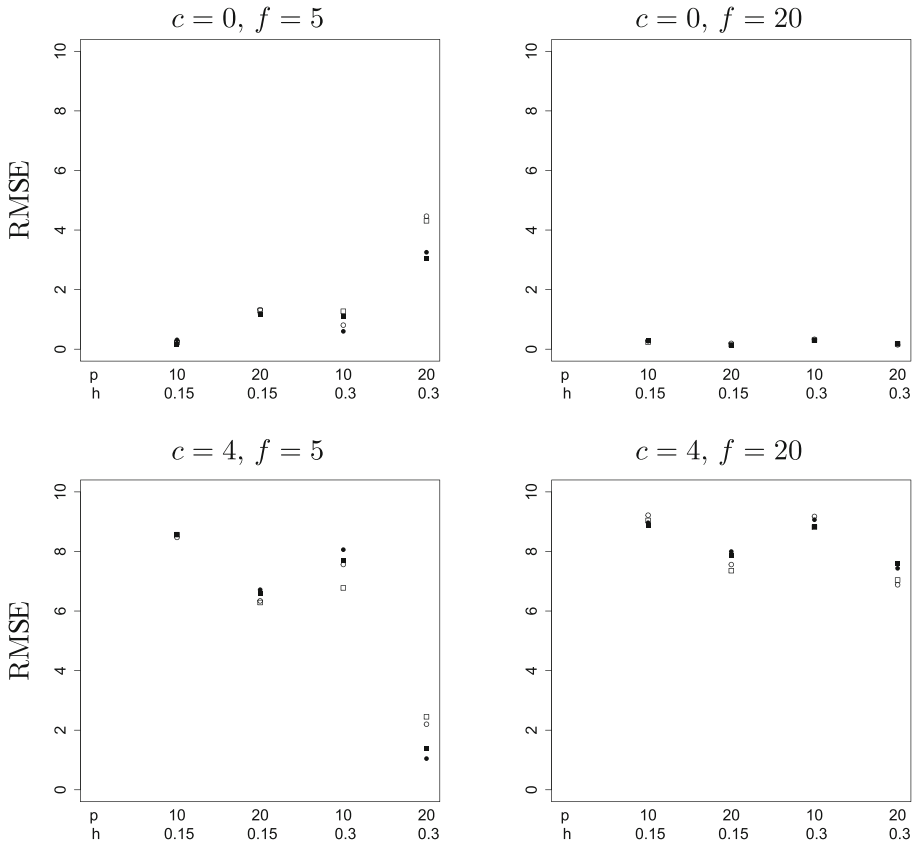

Fig. 8 Simulation study: Average Root Mean Squared Error for f. The symbols represent $q = 0.15$ and window = 30 (□), $q = 0.15$ and window = 100 (○), $q = 0.3$ and window = 30 (■), and $q = 0.3$ and window = 100 (●)

8 Further results of the application to modelling elite sporting performance trajectories

Figures 9 and 10 show individual performance and excess performance trajectories for 12 weightlifters (6 male and 6 female). All three athletes show similar patterns with their performances improving to a peak in their mid-twenties. There are less performance than the 100 metres sprinting example (for example, 20, 16 and 16 observations respectively for the athletes in Figure 10) but the model still able to capture the tends in the data and to provide sensible measures of the uncertainty in the estimation.

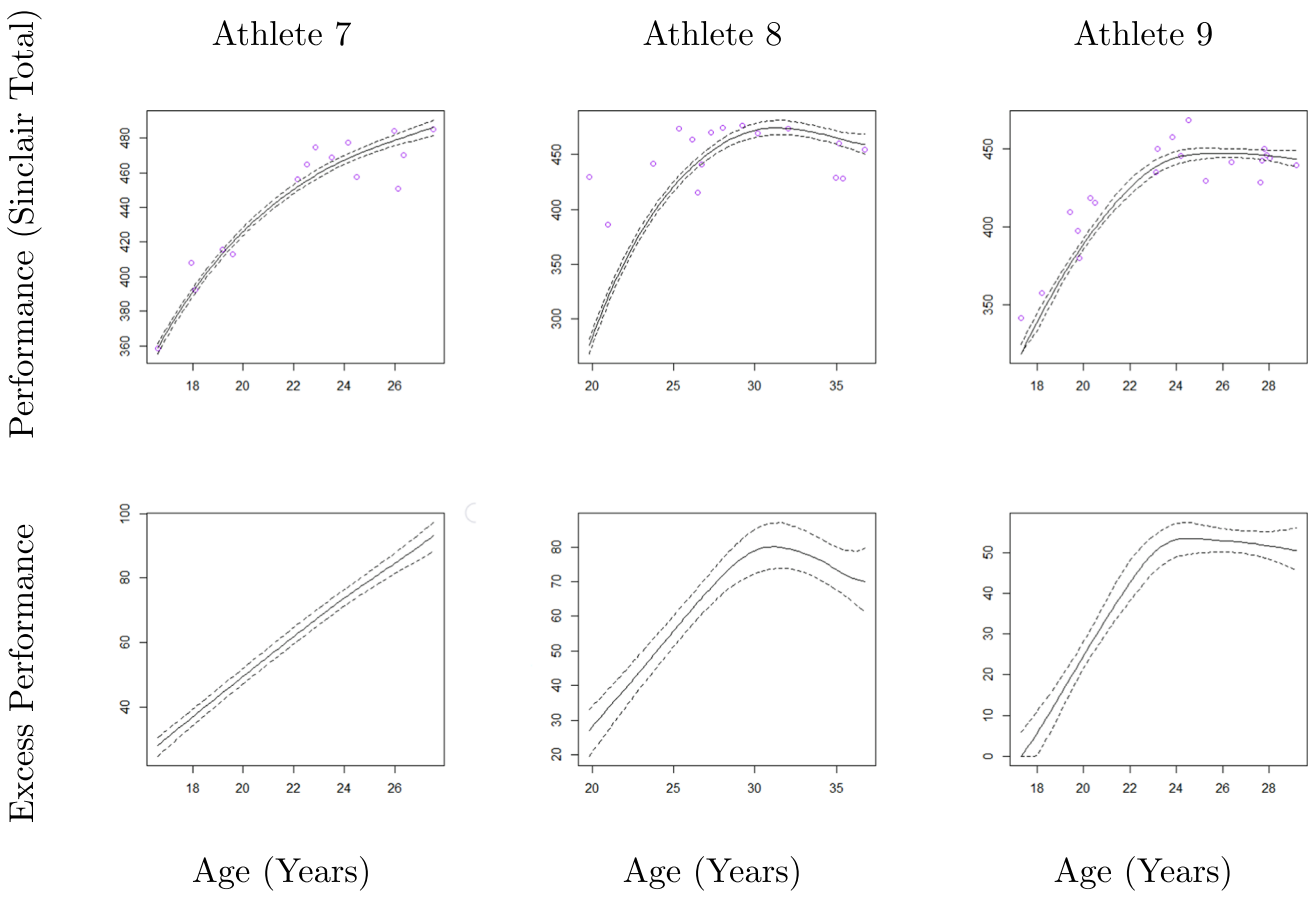

Fig. 9 Individual performance (with observed performances) and excess performance trajectories for male weightlifters, shown as posterior median (solid line) and 95% central credible interval (dashed lines)

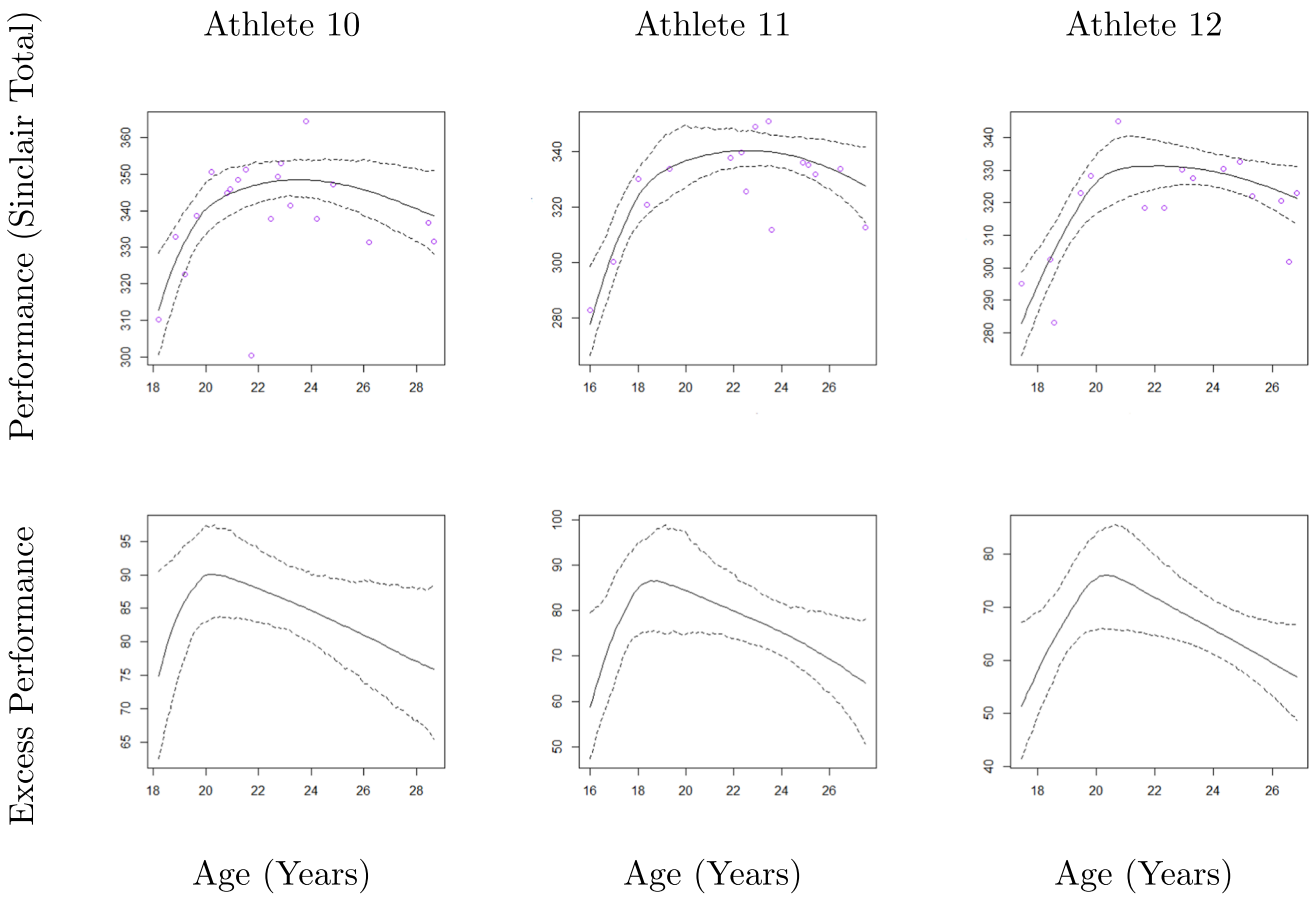

Fig. 10 Individual performance (with observed performances) and excess performance trajectories for female weightlifters, shown as posterior median (solid line) and 95% central credible interval (dashed lines)

Acknowledgements This research was supported by a Partnership for Clean Competition research grant awarded to JH (Grant: 514). Specialist and High Performance Computing systems were provided by Information Services at the University of Kent.

References

Lindsey, J.K.: Models for Repeated Measurements. Oxford University Press, Oxford (1999)

Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G.: Longitudinal Data Analysis. CRC Press, Boca Raton (2008)

Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)

Chen, Z., Dunson, D.B.: Random effects selection in linear mixed models. Biometrics **59**, 762–769 (2003)

Armagan, A., Dunson, D.B.: Sparse variational analysis of linear mixed models for large data sets. Statistics and Probability Letters **81**, 1056–1062 (2011)

Tung, D.T., Tran, M.-N., Cuong, T.M.: Bayesian adaptive lasso with variational bayes for variable selection in high-dimensional generalized linear mixed models. Communications in Statistics - Simulation and Computation **48**, 530–543 (2019)

Park, T., Casella, G.: The Bayesian Lasso. Journal of the American Statistical Association **103**, 681–686 (2008)

Degani, E., Maestrini, L., Toczydlowska, D., Wand, M.P.: Sparse linear mixed model selection via streamlined variational Bayes. Electronic Journal of Statistics **16**, 5182–5225 (2022)

Bhadra, A., Datta, J., Polson, N.G., Willard, B.T.: Lasso meets horseshoe: a survey. Statistical Science **34**, 405–427 (2019)

Ray, P., Bhattacharya, A.: Signal adaptive variable selector for the horseshoe prior arXiv:1810.09004

Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society: Series B **65**, 367–389 (2003)

Berry, S.M., Reese, C.S., Larkey, P.D.: Bridging different eras in sports. Journal of the American Statistical Association **94**, 661–676 (1999)

Griffin, J.E., Hinoveanu, L.C., Hopker, J.G.: Bayesian modelling of elite sporting performance with large databases. Journal of Quantitative Analysis in Sports **18**, 253–268 (2022)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B **39**, 1–38 (1977)

Meng, X.-L., Dyk, D.: The EM algorithm - an old folk-song sung to a fast new tune. Journal of the Royal Statistical Society, Series B **59**, 511–567 (1997)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational Inference: A Review for Statisticians. Journal of the American Statistical Association **112**, 859–877 (2017)

Madigan, D., Raftery, A.E.: Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association **89**, 1535–1546 (1994)

Ferede, M.M., Dagne, G.A., Mwalili, S.M., Bilchut, W.H., Engida, H.A., Karanja, S.M.: Flexible Bayesian semiparametric mixed-effects model for skewed longitudinal data. BMC Medical Research Methodology **24**, 56 (2024)

Gong, M., Mao, Z., Zhang, D., Ren, J., Zuo, S.: Study on Bayesian Skew-Normal Linear Mixed Model and Its Application in Fire Insurance. Fire Technology **59**, 2455–2480 (2023)

Wand, M.P., Ormerod, J.T., Padoan, S.A., Frühwirth, R.: Mean field Variational Bayes for elaborate distributions. Bayesian Analysis **6**, 847–900 (2011)

Guha, N., Wu, X., Efendiev, Y., Jin, B., Mallick, B.K.: A variational Bayesian approach for inverse problems with skew-t error distributions. Journal of Computational Physics **301**, 377–393 (2015)

Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. Annals of Statistics **32**, 870–897 (2004)