Supplementary Material: "Modeling subject-specific nonautonomous dynamics"

Siyuan Zhou, Debashis Paul and Jie Peng

1 Shields Avenue, Department of Statistics, University of California, Davis, CA 95616

S.1 Bias and variance terms in Theorem 2

Expression for bias terms

In (25), the functions f_ℓ's are defined as follows:

$$f_1(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0) = \Xi^{12}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1}\Psi^{-1}\boldsymbol{\theta}_i^* \tag{S.1}$$

$$\begin{aligned} f_2(a_i, \theta_i^*, \boldsymbol{\gamma}_0) &= -\sigma_\varepsilon^2\Xi^{12}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1} \\ &\quad \cdot \mathbb{E}(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1}\nabla_\theta X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)|a_i, \theta_i^*) \end{aligned} \tag{S.2}$$

$$f_3(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0) = \sigma_\varepsilon^2\mathbb{E}(\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1}\nabla_{\boldsymbol{\theta}} X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)|a_i, \boldsymbol{\theta}_i^*) \tag{S.3}$$

$$\begin{aligned} f_4(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0) &= -\sigma_\varepsilon^2 \left(\mathbb{E}\left(\nabla_{\boldsymbol{\theta}^T} X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1}\right.\right. \\ &\quad \cdot \left.\left. R_k(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)\Xi^{22}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)^{-1}\nabla_\theta X(T; a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0)|a_i, \boldsymbol{\theta}_i^*\right)\right)_{k=1}^M, \end{aligned} \tag{S.4}$$

where

$$R_k(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0) = m_i^{-1}\mathbb{E}(R_{i,k}^*(\boldsymbol{\beta}_0)|a_i, \boldsymbol{\theta}_i^*).$$

and $R_{i,k}^*(\boldsymbol{\beta}_0)$ is as in (S.42).

Estimate of $\Gamma_n(\boldsymbol{\gamma}_0)$

We can estimate $\Gamma_n(\boldsymbol{\gamma}_0)$ by

$$\widehat{\Gamma}_n = \frac{1}{N_n}\sum_{i=1}^{n} \mathbf{C}^T[\widehat{B}_i - \widehat{\xi}_i\widehat{W}_i^{-1}\widehat{\xi}_i^T]\mathbf{C},$$

where

$$\begin{aligned}
\widehat{B}_i &:= \sum_{j=1}^{m_i} \nabla_{\boldsymbol{\beta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\beta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}), \qquad \widehat{\xi}_i = \sum_{j=1}^{m_i} \nabla_{\boldsymbol{\beta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}) \\
\widehat{W}_i &= \sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}) + \widehat{\Psi}.
\end{aligned}$$

Here, we use $\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})$ as a shorthand for $X_i(T_{ij}; a_i, \widehat{\boldsymbol{\theta}}_i(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\beta}})$.

Estimate of $\mathfrak{b}_n(\boldsymbol{\gamma}_0)$

We use estimate $\mathfrak{b}_n(\boldsymbol{\gamma}_0)$ by

$$\widehat{\mathfrak{b}}_n = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{\ell=1}^{4} \widehat{f}_{\ell,i}$$

where

$$\begin{aligned}
\widehat{f}_{1,i} &= \widehat{\xi}_i\widehat{W}_i^{-1}\widehat{\Psi}^{-1} \\
\widehat{f}_{2,i} &= -\widehat{\sigma}_\varepsilon^2\widehat{\xi}_i\widehat{W}_i^{-1}\left[\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\widehat{W}_i^{-1}\nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\right] \\
\widehat{f}_{3,i} &= \widehat{\sigma}_\varepsilon^2\left[\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\widehat{W}_i^{-1}\nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\right] \\
\widehat{f}_{4,i} &= -\widehat{\sigma}_\varepsilon^2\left(\sum_{j=1}^{m_i}(\nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}))^T\widehat{W}_i^{-1}\widehat{R}_{k,i}\widehat{W}_i^{-1}\nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\right)_{k=1}^{M}
\end{aligned}$$

where, for $k = 1, \ldots, M$, and $i = 1, \ldots, n$,

$$\widehat{R}_{k,i} = \sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\beta_k\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}) + \sum_{j=1}^{m_i} \nabla_{\beta_k\boldsymbol{\theta}}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}) + \sum_{j=1}^{m_i} \nabla_{\beta_k}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}\widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}).$$

S.2 Details on the proof of identifiability

In the proof of identifiability in Section A.1, the assertion that "(A.5) holds only if $\mathbf{d}_{i_k}^T q(t) = c_k$ for all k" is trivially true when $p = 1$. Therefore, let $p > 1$.

Let $f_{\mathbf{h}}(x) := \mathbf{h}^T \mathbf{C}^T \Phi(x)$. Then we can rewrite (A.5) as

$$f_{\mathbf{h}}(X_{i_k}(t)) = g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) \mathbf{d}_{i_k}^T q(t), \qquad \text{for all } t \in [0,1]. \tag{S.5}$$

Since both $f_{\mathbf{h}}$ and $g_{\boldsymbol{\beta}(\boldsymbol{\gamma})h}$ are represented in the same spline basis, there exists an interval $(\underline{t}_k, \bar{t}_k) \subset [0,1]$ such that both $f_{\mathbf{h}}$ and $g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}$ are polynomials on $(\underline{x}_k, \overline{x}_k)$ where $\underline{x}_k = X_{i_k}(\underline{t}_k)$ and $\overline{x}_k = X_{i_k}(\bar{t}_k)$. Differentiating both sides of (S.5) with respect to t, and invoking (1), we have

$$\begin{aligned} e^{\boldsymbol{\theta}_{i_k}^T q(t)} g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) f'_{\mathbf{h}}(X_{i_k}(t)) &= e^{\boldsymbol{\theta}_{i_k}^T q(t)} g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) g'_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) \mathbf{d}_{i_k}^T q(t) \\ &\quad + g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) \mathbf{d}_{i_k}^T q'(t), \qquad \text{for } t \in (\underline{t}_k, \bar{t}_k), \end{aligned}$$

or,

$$e^{\boldsymbol{\theta}_{i_k}^T q(t)} \left(f'_{\mathbf{h}}(X_{i_k}(t)) - g'_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) \mathbf{d}_{i_k}^T q(t) \right) = \mathbf{d}_{i_k}^T q'(t), \qquad \text{for } t \in (\underline{t}_k, \bar{t}_k). \tag{S.6}$$

Now, notice that since $p > 1$, by **F1**.2, $e^{\boldsymbol{\theta}_{i_k}^T q(t)}$ is not a constant. Therefore, (S.5) and (S.6) are two polynomial equations for $X_{i_k}(t)$, the former with coefficients that are polynomials in t, and the latter with coefficients that are polynomials in t and $e^{\boldsymbol{\theta}_{i_k}^T q(t)}$. The dependence on $e^{\boldsymbol{\theta}_{i_k}^T q(t)}$ is nontrivial unless the right hand side of (S.6) is zero, which can only happen if $\mathbf{d}_{i_k}^T q(t)$ is a constant. Otherwise, $X_{i_k}(t)$ cannot simultaneously satisfy both (S.5) and (S.6). This establishes the fact that (S.5) (i.e., (A.5)) can hold only if $\mathbf{d}_{i_k}^T q(t) = c_k$ for all k.

S.3 Likelihood and profiling

Our working assumption is that $\boldsymbol{\theta}_i$'s are i.i.d. $N(\boldsymbol{\mu}, \Sigma)$ for a $p \times 1$ unknown vector $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_\theta$ and a $p \times p$ positive definite matrix $\Sigma \equiv \Sigma_\theta$, which is assumed known. We also assume that ε_{ij}'s are i.i.d. $N(0, \sigma_\varepsilon^2)$ where σ_ε^2 is also considered known, though we can also estimate both Σ and σ_ε^2 from the data. Define $\Psi = (1/\sigma_\varepsilon^2)\Sigma$. Let $\mathbf{H}$ be an $M \times M$ positive semi-definite matrix. Then for any $\lambda \geq 0$, we define the penalized *negative log generalized hierarchical likelihood* for $(\boldsymbol{\Theta}, \boldsymbol{\beta})$ where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ as

$$L^H(\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\mu}) \equiv \ell^H_{n,\lambda}(\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \ell^H_{i,\lambda}(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\mu}), \tag{S.7}$$

where

$$\begin{aligned} \ell^H_{i,\lambda}(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\mu}) &= \frac{1}{2\sigma_\varepsilon^2}\left(\sum_{j=1}^{m_i}(Y_{ij} - X(T_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\beta}))^2 + (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T\Psi^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}) + \frac{\lambda}{n}\boldsymbol{\beta}^T\mathbf{H}\boldsymbol{\beta}\right) \\ &\quad + \frac{1}{2}\log|\Sigma| + \frac{m_i}{2}\log\sigma_\varepsilon^2 + \frac{m_i}{2}\log(2\pi) \\ &= \frac{1}{\sigma_\varepsilon^2}L^H_i(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\mu}) + \frac{1}{2}\log|\Sigma| + \frac{m_i}{2}\log\sigma_\varepsilon^2 + \frac{m_i}{2}\log(2\pi), \end{aligned} \tag{S.8}$$

In the second line of the above equation, we dropped the suffix λ for notational convenience. The phrase *generalized* refers to the fact that, though Σ and σ_ε are considered known, they can be any positive definite matrix and positive scalar, respectively, even if the latter are not the true variances of $\boldsymbol{\theta}_i$'s and ε_{ij}'s, respectively.

Then, obtaining the maximum generalized H-likelihoood estimate $(\widehat{\boldsymbol{\theta}}_H, \widehat{\boldsymbol{\beta}}_H, \widehat{\boldsymbol{\mu}}_H)$ is equivalent to

$$\min_{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\mu}} \sum_{i=1}^{n} L^H_i(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\mu}). \tag{S.9}$$

The minimization in (S.9) can be broken into two steps:

$$\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}) := \arg\min_{\boldsymbol{\theta}_i} L_i^H(\boldsymbol{\theta}_i,\boldsymbol{\beta},\boldsymbol{\mu}), \quad i=1,\ldots,n, \tag{S.10}$$

and

$$(\widehat{\boldsymbol{\beta}}_H,\widehat{\boldsymbol{\mu}}_H) := \arg\min_{\boldsymbol{\beta},\boldsymbol{\mu}} \sum_{i=1}^{n} L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) \tag{S.11}$$

where

$$L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu}). \tag{S.12}$$

Finally, $\widehat{\boldsymbol{\Theta}}_H = (\widehat{\boldsymbol{\theta}}_1(\widehat{\boldsymbol{\beta}}_H,\widehat{\boldsymbol{\mu}}_H),\ldots,\widehat{\boldsymbol{\theta}}_n(\widehat{\boldsymbol{\beta}}_H,\widehat{\boldsymbol{\mu}}_H))$. The expression $L_i^P(\boldsymbol{\beta})$ in (S.11) (or more appropriately, $(1/\sigma_\varepsilon^2)L_i^P(\boldsymbol{\beta},\boldsymbol{\mu})$) can be termed the penalized negative "profile log-likelihood" with respect to $(\boldsymbol{\beta},\boldsymbol{\mu})$.

S.3.1 Likelihood equations and identifiability

The following sets of equations characterize the maximum H-likelihoood estimates under the given set up. First, since $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})$ minimizes $L_i^H(\boldsymbol{\theta}_i,\boldsymbol{\beta},\boldsymbol{\mu})$ for any given $(\boldsymbol{\beta},\boldsymbol{\mu})$, we have

$$\nabla_\theta L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu}) := \frac{\partial}{\partial\theta} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu}) = 0. \tag{S.13}$$

Here and afterwards, we follow the convention

$$\nabla_y f(y_0,z_0) = \frac{\partial}{\partial y} f(y_0,z_0) := \frac{\partial}{\partial y} f(y,z) \mid_{y=y_0,z=z_0},$$

and, assuming $z(y)$ to be a differentiable function of y, make use of the chain rule,

$$\begin{aligned} \frac{d}{dy} f(y,z(y)) &= \nabla_y f(y,z(y)) + \frac{dz(y)}{dy} \nabla_z f(y,z(y)) \\ &:= \frac{\partial}{\partial y} f(y,z) \mid_{z=z(y)} + \frac{dz(y)}{dy} \frac{\partial}{\partial z} f(y,z) \mid_{z=z(y)}. \end{aligned}$$

Also, we define $\nabla_{xy^T} := \frac{\partial^2}{\partial x \partial y^T}$.

For the rest of this subsection, for notational convenience, we use $\nabla_{\boldsymbol{\theta}} L_i^{H,r}$, $\nabla_{\boldsymbol{\beta}} L_i^{H,r}$ and $\nabla_{\mu} L_i^{H,r}$ to mean $\nabla_{\boldsymbol{\theta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu})$, $\nabla_{\boldsymbol{\beta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu})$ and $\nabla_{\boldsymbol{\mu}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu})$, respectively, by suppressing the dependence on $(\boldsymbol{\beta},\boldsymbol{\mu})$, with analogous notations for second order mixed partial derivatives.

By the Implicit Function Theorem, $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})$ is differentiable with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, and thus,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) &= \frac{d}{d\boldsymbol{\beta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu}) \\ &= \nabla_{\boldsymbol{\beta}} L_i^{H,r} + \frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \nabla_{\theta} L_i^{H,r} \;=\; \nabla_{\boldsymbol{\beta}} L_i^{H,r}, \end{aligned} \tag{S.14}$$

where the last equality is due to (S.13). Similarly,

$$\frac{\partial}{\partial \boldsymbol{\mu}} L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}} L_i^{H,r}. \tag{S.15}$$

Differentiating (S.13) with respect to $\boldsymbol{\beta}$, we have

$$0 = \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^{H,r} + \frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^{H,r},$$

so that

$$\frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}} = -\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu}) \left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu}),\boldsymbol{\beta},\boldsymbol{\mu})\right]^{-1}. \tag{S.16}$$

Differentiating (S.14) one more time with respect to $\boldsymbol{\beta}$, we have

$$\begin{aligned}\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) &= \frac{\partial}{\partial \boldsymbol{\beta}} (\nabla_{\boldsymbol{\beta}} L_i^{H,r})^T \\ &= \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^{H,r} + \frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \nabla_{\boldsymbol{\theta}\boldsymbol{\beta}^T} L_i^{H,r} \\ &= \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^{H,r} - \frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^{H,r}\right] \left(\frac{\partial \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}}\right)^T, \\ &= \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^{H,r} - \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^{H,r} \left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^{H,r}\right]^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\beta}^T} L_i^{H,r}, \end{aligned} \tag{S.17}$$

where the last two equalities follow by making use of (S.16). Similar derivations yield

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = \nabla_{\boldsymbol{\beta}\boldsymbol{\mu}^T}L_i^{H,r} - \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T}L_i^{H,r}\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^{H,r}\right]^{-1}\nabla_{\boldsymbol{\theta}\boldsymbol{\mu}^T}L_i^{H,r}, \tag{S.18}$$

$$\frac{\partial^2}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}\boldsymbol{\mu}^T}L_i^{H,r} - \nabla_{\boldsymbol{\mu}\boldsymbol{\theta}^T}L_i^{H,r}\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^{H,r}\right]^{-1}\nabla_{\boldsymbol{\theta}\boldsymbol{\mu}^T}L_i^{H,r}. \tag{S.19}$$

Notice that, from (S.8), we also get

$$\nabla_{\boldsymbol{\beta}\boldsymbol{\mu}^T}L_i^{H,r} = 0, \qquad \nabla_{\boldsymbol{\mu}\boldsymbol{\mu}^T}L_i^{H,r} = \frac{1}{\sigma_\varepsilon^2}\Psi^{-1}, \qquad \nabla_{\theta\boldsymbol{\mu}^T}L_i^{H,r} = -\frac{1}{\sigma_\varepsilon^2}\Psi^{-1}. \tag{S.20}$$

From (S.17) – (S.19) and (S.20), it is clear that

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = O_P(\overline{m}), \quad \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = O_P(1), \quad \frac{\partial^2}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) = O_P(1),$$

where the O_P terms can be made uniform in i with an additional factor of $\log n$.

An important consequence of the above is that the main contribution of subject i to the information matrix for $\boldsymbol{\beta}$, namely,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu}) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu})\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu})\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\mu}^T}L_i^P(\boldsymbol{\beta},\boldsymbol{\mu})\right)^T,$$

comes from the first term (involving Hessian with respect to $\boldsymbol{\beta}$). Along with the identifiability condition (6), this establishes the asymptotic nonsingularity of the observed information matrix with respect to $\boldsymbol{\beta}$, and hence the asymptotic identifiability of $\boldsymbol{\beta}$, even when $\boldsymbol{\mu}$ is treated as an unknown parameter. As a further consequence, the discussion here also indicates that we can prove consistency of the estimator of $\boldsymbol{\beta}$ under the condition $\sum_{j=1}^{M}\beta_j = 1$ even when $\boldsymbol{\mu}$ is estimated from the data.

S.4 Details on the proof of Theorem 1

S.4.1 Gradients and Hessians of L_i^H

Define, for $j = 1, \ldots, m_i$; $i = 1, \ldots, n$,

$$
\begin{aligned}
X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= X_i(T_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\beta}), \qquad \nabla_{\boldsymbol{\theta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) = \nabla_{\boldsymbol{\theta}} X_i(T_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\beta}) \\
\nabla_{\boldsymbol{\beta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} X_i(T_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\beta}), \qquad \widehat{X}_{ij}(\boldsymbol{\beta}) = X_i(T_{ij}; \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) \\
\nabla_{\boldsymbol{\theta}} \widehat{X}_{ij}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\theta}} X_i(T_{ij}; \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}), \qquad \nabla_{\boldsymbol{\beta}} \widehat{X}_{ij}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} X_i(T_{ij}; \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) \\
X^*_{ij}(\boldsymbol{\beta}) &= X_i(T_{ij}; \boldsymbol{\theta}^*_i, \boldsymbol{\beta}), \qquad \nabla_{\boldsymbol{\theta}} X^*_{ij}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\theta}} X_i(T_{ij}; \boldsymbol{\theta}^*_i, \boldsymbol{\beta}) \\
\nabla_{\boldsymbol{\beta}} X^*_{ij}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} X_i(T_{ij}; \boldsymbol{\theta}^*_i, \boldsymbol{\beta})
\end{aligned}
$$

etc.

Direct calculations yield,

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= -\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}))\nabla_{\boldsymbol{\theta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) + \Psi^{-1}\boldsymbol{\theta}_i && \text{(S.21)}\\
\nabla_{\boldsymbol{\beta}} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= -\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}))\nabla_{\boldsymbol{\beta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) + n^{-1}\lambda\mathbf{H}\boldsymbol{\beta} && \text{(S.22)}\\
\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= -\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}))\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\\
&\quad + \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\theta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\nabla_{\boldsymbol{\theta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) + \Psi^{-1} && \text{(S.23)}\\
\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= -\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}))\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\\
&\quad + \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\beta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\nabla_{\boldsymbol{\theta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) && \text{(S.24)}\\
\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\beta}) &= -\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}))\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\\
&\quad + \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\beta}} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta})\nabla_{\boldsymbol{\beta}^T} X_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}) + n^{-1}\lambda\mathbf{H}. && \text{(S.25)}
\end{aligned}$$

S.4.2 Expansion of $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)$

Since $\varepsilon_{ij} = Y_{ij} - X_{ij}^*(\boldsymbol{\beta}_0)$, from (S.23), we get

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0) = -\sum_{j=1}^{m_i}\varepsilon_{ij}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + W_i^*(\boldsymbol{\beta}_0) = -P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} + W_i^*(\boldsymbol{\beta}_0), \qquad \text{(S.26)}$$

and so from (S.35) we get,

$$
\begin{aligned}
& \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^* \\
= \; & -W_i^*(\boldsymbol{\beta}_0)^{-1} \nabla_{\boldsymbol{\theta}} L_i^H(\boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0) + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) - W_i^*(\boldsymbol{\beta}_0)^{-1} r_{1,i} \\
= \; & W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* \\
& + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) - W_i^*(\boldsymbol{\beta}_0)^{-1} r_{1,i} \\
= \; & W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} \\
& - W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* \\
& + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) - W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} r_{1,i}, \\
= \; & W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} + r_{2,i},
\end{aligned}
$$

where we have used (S.21) in the second step. From the expression for $r_{2,i}$ and the bound for $r_{1,i}$ in (S.36), we obtain (S.41).

S.4.3 Behavior of the gradient of L_i^P

By (S.14), (A.12), (S.22) and (S.24), we have

$$\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}}L_i^P(\boldsymbol{\beta}_0) \;=\; \nabla_{\boldsymbol{\beta}}L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0),\boldsymbol{\beta}_0)\\
=\;& \nabla_{\boldsymbol{\beta}}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)+\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*)\\
&+\;\Big((\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*)^T\nabla_{\beta_k\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*)\Big)_{k=1}^M+\tilde{r}_{3,i}\\
=\;& -p_{i,\boldsymbol{\beta}}+n^{-1}\lambda\mathbf{H}\\
&+\;(-P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T}+\xi_i^*(\boldsymbol{\beta}_0))\times(W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}-W_i^*(\boldsymbol{\beta}_0)^{-1}\Psi^{-1}\boldsymbol{\theta}_i^*+W_i^*(\boldsymbol{\beta}_0)^{-1}P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}+r_{2,i})\\
&+(p_{i,\boldsymbol{\theta}}^TW_i^*(\boldsymbol{\beta}_0)^{-1}R_{i,k}^*(\boldsymbol{\beta}_0)W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}})_{k=1}^M+r_{3,i}\\
=\;& -p_{i,\boldsymbol{\beta}}+\xi_i^*(\boldsymbol{\beta}_0)W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}\\
&-\;\xi_i^*(\boldsymbol{\beta}_0)W_i^*(\boldsymbol{\beta}_0)^{-1}\Psi^{-1}\boldsymbol{\theta}_i^*+\xi_i^*(\boldsymbol{\beta}_0)W_i^*(\boldsymbol{\beta}_0)^{-1}P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}-P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}\\
&+\;(p_{i,\boldsymbol{\theta}}^TW_i^*(\boldsymbol{\beta}_0)^{-1}R_{i,k}^*(\boldsymbol{\beta}_0)W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}})_{k=1}^M+r_{4,i},
\end{aligned}$$

where

$$\begin{aligned}
r_{4,i} \;=\;& P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}\Psi^{-1}\boldsymbol{\theta}_i^*-P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}W_i^*(\boldsymbol{\beta}_0)^{-1}p_{i,\boldsymbol{\theta}}+n^{-1}\lambda\mathbf{H}\\
&-P_{i,\beta\boldsymbol{\theta}^T}r_{2,i}+\xi_i^*(\boldsymbol{\beta}_0)r_{2,i}+r_{3,i}.
\end{aligned}$$

It can be shown that $\max_{1\le i\le n}\parallel r_{3,i}\parallel=\widetilde{O}((\log n)^2\underline{m}^{-1/2})$ and from this, **A5** and (S.41) we can deduce (S.46).

S.4.4 Behavior of the Hessian of L_i^P

Using calculations similar to those in Sections S.4.2 and S.4.3, and using the expansions

$$\widehat{X}_{ij}(\boldsymbol{\beta}_0) - X^*_{ij}(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i) + O(\| \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i \|^2) \quad \text{(S.27)}$$

$$\nabla_{\boldsymbol{\theta}}\widehat{X}_{ij}(\boldsymbol{\beta}_0) - \nabla_{\boldsymbol{\theta}} X^*_{ij}(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i) + O(\| \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i \|^2) \quad \text{(S.28)}$$

$$\nabla_{\boldsymbol{\beta}}\widehat{X}_{ij}(\boldsymbol{\beta}_0) - \nabla_{\boldsymbol{\beta}} X^*_{ij}(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i) + O(\| \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i \|^2), \quad \text{(S.29)}$$

we can isolate the leading order terms in the following quantities

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0), \qquad \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) \qquad \text{and} \qquad \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0).$$

From (S.27), (S.28), (S.23) and (A.12), we have

$$\begin{aligned}
&\frac{1}{m_i}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) \\
= \; & -\frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0) + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}} X^*_{ij}(\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0) + \frac{1}{m_i}\Psi^{-1} \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i)\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0) - \frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij}\langle \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T\boldsymbol{\theta}} X^*_{ij}(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i\rangle \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}} X^*_{ij}(\boldsymbol{\beta}_0)(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i))^T \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}^*_i)\nabla_{\boldsymbol{\theta}^T} X^*_{ij}(\boldsymbol{\beta}_0) + \widetilde{O}\left(\frac{\log n}{\underline{m}}\right) \\
= \; & \frac{1}{m_i} W^*_i(\boldsymbol{\beta}_0) + \widetilde{O}\left(\sqrt{\frac{\log n}{\underline{m}}}\right). \qquad \text{(S.30)}
\end{aligned}$$

Similarly, by (S.27), (S.29), (S.24) and (A.12), we have

$$\begin{aligned}
&\frac{1}{m_i}\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) \\
= \; & -\frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_\beta X_{ij}^*(\boldsymbol{\beta}_0) \nabla_{\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) - \frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij} \langle \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T\boldsymbol{\theta}} X_{ij}^*(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^* \rangle \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_\beta X_{ij}^*(\boldsymbol{\beta}_0)(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*))^T \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) \nabla_{\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \widetilde{O}\left(\frac{\log n}{\underline{m}}\right) \\
= \; & \frac{1}{m_i}\xi_i^*(\boldsymbol{\beta}_0) + \widetilde{O}\left(\sqrt{\frac{\log n}{\underline{m}}}\right). \qquad \text{(S.31)}
\end{aligned}$$

Finally, by (S.27), (S.29), (S.25) and (A.12), we have

$$\begin{aligned}
&\frac{1}{m_i}\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) \\
= \; & -\frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_\beta X_{ij}^*(\boldsymbol{\beta}_0) \nabla_{\boldsymbol{\beta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \frac{\lambda}{n}\mathbf{H} \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} X_{ij}^*(\boldsymbol{\beta}_0) - \frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij} \langle \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{\theta}} X_{ij}^*(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^* \rangle \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_\beta X_{ij}^*(\boldsymbol{\beta}_0)(\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*))^T \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^*) \nabla_{\boldsymbol{\beta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \widetilde{O}\left(\frac{\log n}{\underline{m}}\right) \\
= \; & \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla_\beta X_{ij}^*(\boldsymbol{\beta}_0)(\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0) + \widetilde{O}\left(\max\{\frac{1}{n}, \sqrt{\frac{\log n}{\underline{m}}}\}\right). \qquad \text{(S.32)}
\end{aligned}$$

S.4.5 Details of the derivation of (23)

Based on the derivations in Section S.3.1, the following expressions are valid for $i = 1, \ldots, n$:

$$\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}), \qquad \text{(S.33)}$$

and using (S.17),

$$\begin{aligned}\frac{d^2}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T}L_i^P(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T}L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}),\boldsymbol{\beta}) \\ &\quad -\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T}L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}),\boldsymbol{\beta})\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}),\boldsymbol{\beta})\right]^{-1}\nabla_{\boldsymbol{\theta}\boldsymbol{\beta}^T}L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}),\boldsymbol{\beta}) \quad \text{(S.34)}\end{aligned}$$

In the following, we use the notations and expressions of gradients and Hessians given in the Supplementary Material.

Our next step in proving (23) is to obtain a first order expansion for $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)$ by making use of (A.7) and (A.8). Expanding the right hand side of (A.7) in Taylor series around $\boldsymbol{\theta}_i^*$, we have

$$0 = \nabla_{\boldsymbol{\theta}}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)+\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*)+r_{1,i}, \quad \text{(S.35)}$$

where

$$r_{1,i}=\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^H(\widetilde{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0),\boldsymbol{\beta}_0)-\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T}L_i^H(\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)\right](\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*)$$

and $\parallel \widetilde{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^* \parallel \leq \parallel \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^* \parallel$, which implies that

$$\max_{1\leq i\leq n}\parallel r_{1,i}\parallel = \widetilde{O}((\log n)^{1/2}\overline{m}\parallel \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)-\boldsymbol{\theta}_i^*\parallel^2)=\widetilde{O}((\log n)^{3/2}). \quad \text{(S.36)}$$

Next, define

$$\begin{aligned}B_i^*(\boldsymbol{\beta}) &= \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\beta}}X_{ij}^*(\boldsymbol{\beta})\nabla_{\boldsymbol{\beta}^T}X_{ij}^*(\boldsymbol{\beta}), \qquad \xi_i^*(\boldsymbol{\beta}) = \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\beta}}X_{ij}^*(\boldsymbol{\beta})\nabla_{\boldsymbol{\theta}^T}X_{ij}^*(\boldsymbol{\beta}) \\ W_i^*(\boldsymbol{\beta}) &= \sum_{j=1}^{m_i}\nabla_{\boldsymbol{\theta}}X_{ij}^*(\boldsymbol{\beta})\nabla_{\boldsymbol{\theta}^T}X_{ij}^*(\boldsymbol{\beta})+\Psi^{-1}.\end{aligned}$$

Also, by **A0**-**A5** and (13)–(15), we have

$$\frac{1}{m_i}B_i^*(\boldsymbol{\beta}) = \Xi^{11}(a_i,\boldsymbol{\theta}_i^*,\boldsymbol{\beta})+\widetilde{O}\left(\sqrt{\log n}\underline{m}^{-1/2}\right) \quad \text{(S.37)}$$

$$\frac{1}{m_i}\xi_i^*(\boldsymbol{\beta}) = \Xi^{12}(a_i,\boldsymbol{\theta}_i^*,\boldsymbol{\beta})+\widetilde{O}\left(\sqrt{\log n}\underline{m}^{-1/2}\right) \quad \text{(S.38)}$$

$$\frac{1}{m_i}W_i^*(\boldsymbol{\beta}) = \Xi^{22}(a_i,\boldsymbol{\theta}_i^*,\boldsymbol{\beta})+\widetilde{O}\left(\sqrt{\log n}\underline{m}^{-1/2}\right). \quad \text{(S.39)}$$

where the $\widetilde{O}$ terms are uniform in i.

Further, define

$$p_{i,\boldsymbol{\theta}} = \sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\theta}} X_{ij}^*(\boldsymbol{\beta}_0), \quad p_{i,\boldsymbol{\beta}} = \sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\beta}} X_{ij}^*(\boldsymbol{\beta}_0), \quad P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} = \sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0),$$
$$P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T} = \sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}_0), \quad P_{i,\boldsymbol{\beta}\boldsymbol{\beta}^T} = \sum_{j=1}^{m_i} \varepsilon_{ij} \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} X_{ij}^*(\boldsymbol{\beta}_0).$$

Then, as is shown in Section S.4.2, we have the following expansion of $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0)$:

$$\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^* = W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} + r_{2,i}, \quad \text{(S.40)}$$

where

$$\max_{1 \le i \le n} \| r_{2,i} \| = \widetilde{O}\left(\frac{(\log n)^2}{\underline{m}^{3/2}} \right). \quad \text{(S.41)}$$

The next step is to obtain an expansion for $\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}_0)$. Let

$$R_{i,k}^*(\boldsymbol{\beta}) = \sum_{j=1}^{m_i} \nabla_{\boldsymbol{\theta}} X_{ij}^*(\boldsymbol{\beta}) \nabla_{\beta_k \boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}) + \sum_{j=1}^{m_i} \nabla_{\beta_k \boldsymbol{\theta}} X_{ij}^*(\boldsymbol{\beta}) \nabla_{\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}) + \sum_{j=1}^{m_i} \nabla_{\beta_k} X_{ij}^*(\boldsymbol{\beta}) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} X_{ij}^*(\boldsymbol{\beta}). \quad \text{(S.42)}$$

Then, from the derivation in Section S.4.3, we have

$$\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}_0) = V_i^{(1)}(\boldsymbol{\beta}_0) + V_i^{(2)}(\boldsymbol{\beta}_0) + r_{4,i} \quad \text{(S.43)}$$

where

$$V_i^{(1)}(\boldsymbol{\beta}_0) = -p_{i,\boldsymbol{\beta}} + \xi_i^*(\boldsymbol{\beta}_0) W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}}, \quad \text{(S.44)}$$

which contributes primarily to the asymptotic variance of $\widehat{\boldsymbol{\gamma}}$, and

$$\begin{aligned} V_i^{(2)}(\boldsymbol{\beta}_0) &= \xi_i^*(\boldsymbol{\beta}_0) W_i^*(\boldsymbol{\beta}_0)^{-1} \Psi^{-1} \boldsymbol{\theta}_i^* \\ &\quad + \xi_i^*(\boldsymbol{\beta}_0) W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - P_{i,\boldsymbol{\beta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} \\ &\quad + (p_{i,\boldsymbol{\theta}}^T W_i^*(\boldsymbol{\beta}_0)^{-1} R_{i,k}^*(\boldsymbol{\beta}_0) W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}})_{k=1}^M, \end{aligned} \quad \text{(S.45)}$$

which contributes primarily to the asymptotic bias of $\widehat{\boldsymbol{\gamma}}$, and the remainder terms $r_{i,4}$ satisfy

$$\max_{1\le i\le n} \parallel r_{4,i} \parallel = \widetilde{O}\left(\max\{\lambda/n, (\log n)^2 \underline{m}^{-1/2}\}\right). \tag{S.46}$$

From (S.38) and (S.39) and using (S.21)–(S.23), we can easily derive that $\parallel N_n^{-1} V_i^{(1)}(\boldsymbol{\beta}_0) \parallel = \widetilde{O}(\sqrt{\log n/(n\underline{m})})$ and $\parallel N_n^{-1} V_i^{(2)}(\boldsymbol{\beta}_0) \parallel = \widetilde{O}(\underline{m}^{-1})$, uniformly in i, while the contribution of $r_{4,i}$ can be neglected asymptotically.

Next, combining (S.34) with (S.30), (S.31) and (S.32), derived in Section S.4.4, we have

$$\begin{aligned}
\frac{1}{m_i}\frac{d^2}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} L_i^P(\boldsymbol{\beta}_0) &= \frac{1}{m_i} B_i^*(\boldsymbol{\beta}_0) - \left(\frac{1}{m_i}\xi_i^*(\boldsymbol{\beta}_0)\right)\left(\frac{1}{m_i} W_i^*(\boldsymbol{\beta}_0)\right)^{-1}\left(\frac{1}{m_i}\xi_i^*(\boldsymbol{\beta}_0)\right)^T \\
&\quad + \widetilde{O}\left(\max\left\{\frac{1}{n}, \sqrt{\frac{\log n}{\underline{m}}}\right\}\right) \\
&= \Xi^{1|2}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0) + \widetilde{O}\left(\max\left\{\frac{\lambda}{n}, \sqrt{\frac{\log n}{\underline{m}}}\right\}\right),
\end{aligned} \tag{S.47}$$

where the second equality is obtained by using (S.37)–(S.39) and (17).

S.5 Levenberg-Marquardt procedure for model fitting

In this subsection, we provide some detail on the estimation of $(\boldsymbol{\Theta}, \boldsymbol{\beta})$ using the Levenberg-Mardquardt procedure. We use the notations introduced in Section 3 to denote the current estimates of $(\boldsymbol{\Theta}, \boldsymbol{\beta})$ at any specific iteration of the Levenberg-Marquardt algorithm. Accordingly, the values of $(\boldsymbol{\Theta}, \boldsymbol{\beta})$ are updated by solving the following equations,

$$\begin{aligned}
\left(\mathbf{J}_{i,\boldsymbol{\theta}_i^c}^T \mathbf{J}_{i,\boldsymbol{\theta}_i^c} + \lambda_\theta \Sigma_\theta^{-1}\right)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^c) &= \mathbf{J}_{i,\boldsymbol{\theta}_i^c}^T \tilde{\boldsymbol{\varepsilon}}_i - \lambda_\theta \Sigma_\theta^{-1}(\boldsymbol{\theta}_i^c - \boldsymbol{\mu}_\theta) \\
\left[\mathbf{J}_{\boldsymbol{\beta}^c}^T \mathbf{J}_{\boldsymbol{\beta}^c} + \lambda_\beta \mathbf{H} + \eta_\beta \mathrm{diag}(\mathbf{J}_{\boldsymbol{\beta}^c}^T \mathbf{J}_{\boldsymbol{\beta}^c})\right](\boldsymbol{\beta} - \boldsymbol{\beta}^c) &= \mathbf{J}_{\boldsymbol{\beta}^c}^T \tilde{\boldsymbol{\varepsilon}} - \lambda_\beta \mathbf{H}\boldsymbol{\beta}^c, \quad i = 1, \ldots, n,
\end{aligned}$$

where $\tilde{\boldsymbol{\varepsilon}}_i = (\tilde{\varepsilon}_{ij})_{j=1}^{m_i}$, $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\boldsymbol{\varepsilon}}_i)_{i=1}^n$ and $\tilde{\varepsilon}_{ij} = Y_{ij} - X_i(t_{ij}; \boldsymbol{\theta}_i^c, \boldsymbol{\beta}^c)$ (current residuals) with $\boldsymbol{\theta}_i^c$ and $\boldsymbol{\beta}^c$ denoting the values of the estimates of $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}$, respectively, at the current iteration.

Here, η_β is a sequence of real numbers decreasing to 0 with iterations, used to increase the stability of the algorithm. Also, $\mathbf{J}_{i,\boldsymbol{\theta}_i^c}$ and $\mathbf{J}_{\boldsymbol{\beta}^c}$ denote the Jacobians of the negative log H-likelihoood with respect to the parameters. The latter are defined as

$$\mathbf{J}_{i,\theta_i^c} := \left[\mathbf{J}_{i,\theta_{i1}^c} : \cdots : \mathbf{J}_{i,\theta_{ip}^c}\right], \quad \text{where} \quad \mathbf{J}_{i,\theta_{ik}^c} = \left(\frac{\partial}{\partial \theta_{ik}^c}\tilde{X}_i(t_{ij};\boldsymbol{\theta}_i^c,\boldsymbol{\beta}^c)\right)_{j=1}^{m_i}, \quad k=1,\ldots,p;$$

and

$$\mathbf{J}_{\boldsymbol{\beta}^c} := \left[\mathbf{J}_{\beta_1^c} : \cdots : \mathbf{J}_{\beta_M^c}\right], \quad \text{where} \quad \mathbf{J}_{\beta_k^c} = \left(\frac{\partial}{\partial \beta_k}X_i(t_{ij};\boldsymbol{\theta}_i^c,\boldsymbol{\beta}^c)\right)_{j=1,i=1}^{m_i,n}, \quad k=1,...,M.$$

Computation of $\mathbf{J}_{i,\boldsymbol{\theta}^c}$'s and $\mathbf{J}_{\boldsymbol{\beta}^c}$, requires evaluating $X_i(\cdot;\boldsymbol{\theta}_i,\boldsymbol{\beta})$ and its partial derivatives with respect to $\boldsymbol{\theta}_i$'s and $\boldsymbol{\beta}$. Since these are not available in close forms, a 4^{th} order Runge-Kutta method is used to evaluate these functions on a fine grid. The details of the implementation of Runge-Kutta method are similar to those in the Appendix of Paul et al. (2011).

S.6 Approximate CV score $\widetilde{CV}$

Here we describe a computationally efficient approximate leave-one-subject-out cross-validation score that is similar to that proposed in Paul et al. (2011).

The leave-one-subject-out estimates $\tilde{\boldsymbol{\beta}}^{(-i)}$ are computed approximately by a first order Taylor expansion around the estimate $\hat{\boldsymbol{\beta}}$ (using complete data). Then, we get the leave-one-subject-out prediction of $\boldsymbol{\theta}_i$ by:

$$\tilde{\boldsymbol{\theta}}_i^{(-i)} = \arg\min_{\boldsymbol{\theta}_i} \sum_{j=1}^{m_i}(Y_{ij} - X_i(t_{ij},\boldsymbol{\theta}_i,\tilde{\boldsymbol{\beta}}^{(-i)}))^2 + \sigma_\varepsilon^2(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)^T\Sigma_\theta^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta).$$

The approximate leave-one-subject-out cross-validation score is then defined as :

$$\widetilde{CV} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\ell_{ij}(\tilde{\boldsymbol{\theta}}_i^{(-i)},\tilde{\boldsymbol{\beta}}^{(-i)}). \tag{S.48}$$

Here, $\tilde{\boldsymbol{\theta}}_i^{(-i)}$'s and $\tilde{\boldsymbol{\beta}}^{(-i)}$'s, are obtained by neglecting the higher order terms in their expansions around $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{\beta}}$, respectively. Below we give a detailed derivation of the approximate CV score.

First we obtain an approximation for $\tilde{\boldsymbol{\beta}}^{(-i)}$ by using a Taylor expansion around $\hat{\boldsymbol{\beta}}$ in the equation

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}=0.$$

Our approximation is of the form

$$\tilde{\boldsymbol{\beta}}^{(-i)}:=\widehat{\boldsymbol{\beta}}+\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\partial^2\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}+\lambda_\beta\mathbf{H}\right]^{-1}\left(\sum_{j=1}^{m_i}\frac{\partial\ell_{ij}(\hat{\boldsymbol{\theta}}_i,\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right). \tag{S.49}$$

This approximation ignores the identifiability constraint on $\boldsymbol{\beta}$. In order to obtain the correct form under the identifiability constraint, we use the reparametrization of $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$, that is, $\boldsymbol{\beta}=\mathbf{C}\boldsymbol{\gamma}+\frac{1}{M}\mathbf{1}_M$, where $\mathbf{C}$ is an $M\times(M-1)$ or rank $M-1$ satisfying $\mathbf{1}_M^T\mathbf{C}=0_{M-1}$, and $\boldsymbol{\gamma}\in\mathbb{R}^{M-1}$. Then we have,

$$\begin{aligned}\frac{\partial\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}} &= \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\gamma}}\frac{\partial\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\beta}}\\ \frac{\partial^2\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} &= \mathbf{C}^T\frac{\partial^2\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\mathbf{C}\end{aligned}$$

so that

$$\tilde{\boldsymbol{\gamma}}^{(-i)}=\widehat{\boldsymbol{\gamma}}\quad+\quad\left[\mathbf{C}^T\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\partial^2\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}+\lambda_\beta\mathbf{H}\right)\mathbf{C}\right]^{-1}\mathbf{C}^T\left(\sum_{j=1}^{m_i}\frac{\partial\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right).$$

Then the approximate leave-one-subject-out estimate of $\tilde{\boldsymbol{\beta}}^{(-i)}$ becomes

$$\tilde{\boldsymbol{\beta}}^{(-i)}:=\widehat{\boldsymbol{\beta}}+\mathbf{C}\left[\mathbf{C}^T\left(\sum_{i=1}^{n}\sum_{j=1}^{m_{il}}\frac{\partial^2\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}+\lambda_\beta\mathbf{H}\right)\mathbf{C}\right]^{-1}\mathbf{C}^T\left(\sum_{j=1}^{m_i}\frac{\partial\ell_{ij}(\widehat{\boldsymbol{\theta}}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right) \tag{S.50}$$

The leave-one-subject-out estimate $\tilde{\boldsymbol{\theta}}_i^{(-i)}$ of $\boldsymbol{\theta}_i$ is derived as follows.

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_i^{(-i)} &:= \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{m_i} (Y_{ij} - X_i(T_{ij}, \boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}^{(-i)}))^2 + \lambda_\theta (\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^T \Sigma_\theta^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) \\
&\Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{j=1}^{m_i} \ell_{ij}(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}^{(-i)}) \right) + 2\lambda_\theta \Sigma_\theta^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)|_{\tilde{\boldsymbol{\theta}}_i^{(-i)}} = 0 \\
&\Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{j=1}^{m_i} \ell_{ij}(\tilde{\boldsymbol{\theta}}_i^{(-i)}, \tilde{\boldsymbol{\beta}}^{(-i)}) \right) + 2\lambda_\theta \Sigma_\theta^{-1} (\tilde{\boldsymbol{\theta}}_i^{(-i)} - \boldsymbol{\mu}_\theta) = 0
\end{aligned}$$

We obtain an approximation to the above equation by expanding around $\widehat{\boldsymbol{\theta}}_i$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \tilde{\boldsymbol{\beta}}^{(-i)}) \right) &+ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \tilde{\boldsymbol{\beta}}^{(-i)}) \right)^T (\tilde{\boldsymbol{\theta}}_i^{(-i)} - \widehat{\boldsymbol{\theta}}_i) \\
&+ 2\lambda_\theta \Sigma_\theta^{-1} (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_\theta) + 2\lambda_\theta \Sigma_\theta^{-1} (\tilde{\boldsymbol{\theta}}_i^{(-i)} - \widehat{\boldsymbol{\theta}}_i) = 0
\end{aligned}$$

Let $\delta_i = \tilde{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}}$, so that $\tilde{\boldsymbol{\beta}}^{(-i)} = \widehat{\boldsymbol{\beta}} + \delta_i$. We approximate the above equation again by expanding around $\widehat{\boldsymbol{\beta}}$ and ignoring the term with $\delta_i(\tilde{\boldsymbol{\theta}}_i^{(-i)} - \widehat{\boldsymbol{\theta}}_i)$

$$\begin{aligned}
&\frac{\partial \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right)}{\partial \boldsymbol{\theta}} + \left(\frac{\partial^2 \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} \right) \delta_i + 2\lambda_\theta \Sigma_\theta^{-1} (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_\theta) \\
&\qquad + \left(\frac{\partial^2 \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\beta}) \right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + 2\lambda_\theta \Sigma_\theta^{-1} \right) (\tilde{\boldsymbol{\theta}}_i^{(-i)} - \widehat{\boldsymbol{\theta}}_i) = 0.
\end{aligned} \tag{S.51}$$

Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right) + 2\lambda_\theta \Sigma_\theta^{-1} (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_\theta) = 0,$$

equation (S.51) reduces to

$$\begin{aligned}
0 &= \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} \sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right) \delta_i \\
&\quad + \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right) + 2\lambda_\theta \Sigma_\theta^{-1} \right)^T (\tilde{\boldsymbol{\theta}}_i^{(-i)} - \widehat{\boldsymbol{\theta}}_i),
\end{aligned}$$

so that

$$\tilde{\boldsymbol{\theta}}_i^{(-i)} = \widehat{\boldsymbol{\theta}}_i - \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left(\sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right) + 2\lambda_\theta \Sigma_\theta^{-1} \right)^{-1} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} \sum_{j=1}^{m_i} \ell_{ij}(\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}}) \right) \delta_i,$$

or,

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_i^{(-i)} = \widehat{\boldsymbol{\theta}}_i - & \left[\sum_{j=1}^{m_i} \left(\varepsilon_{ij} \frac{\partial^2 X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - \frac{\partial X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta}} \frac{\partial X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta}^T} \right) + \lambda_\theta \Sigma_\theta^{-1} \right]^{-1} \\ & \left[\sum_{j=1}^{m_i} \left(\varepsilon_{ij} \frac{\partial^2 X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} - \frac{\partial X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta}} \frac{\partial X_i(T_{ij}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right) \right] \cdot \delta_i. \end{aligned}$$

S.7 Additional Figures and Tables

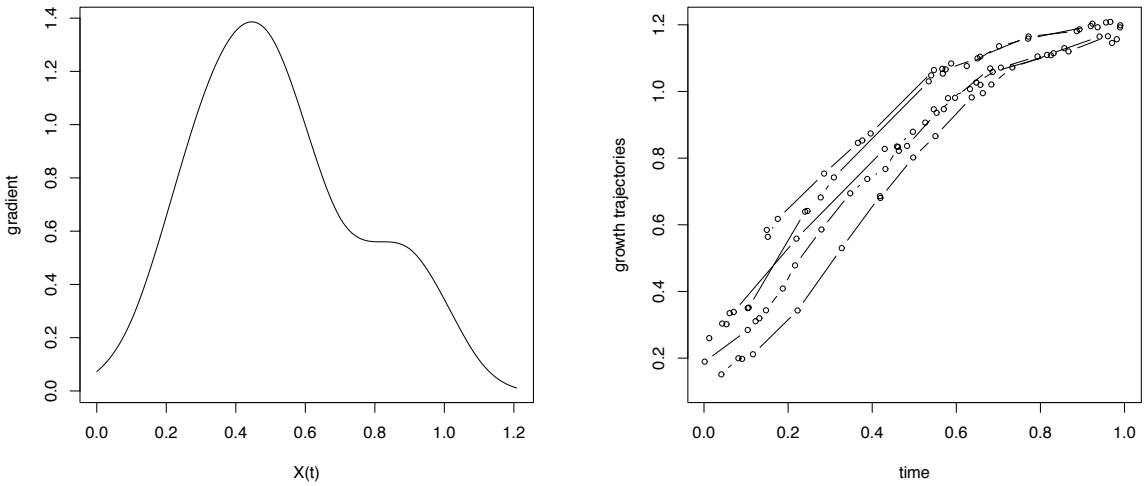


Figure S.1: The "Two-peak" gradient function (left panel) and the corresponding growth trajectories and observations (dots) under `dense` case with $\sigma_\varepsilon = 0.01$ (right panel).

Table S.1: Simulation with "Two-peak" gradient function g and linear ($p = 2$) $Z(\cdot, \boldsymbol{\theta_i})$s. $M = 30$ cubic B-spline basis functions are used for the fitting. "Ideal" model selection is used for selection of λ_β.

Sampling rate	μ_2	σ_ε	Mean(ISE)	SD(ISE)	Median(ISE)	MAD(ISE)
$ISE(\widehat{g})$						
sparse	0	0.01	0.000995	0.000801	0.000780	0.000594
	2	0.01	0.002606	0.001912	0.002202	0.001338
dense	2	0.01	0.001114	0.001414	0.000745	0.000556
	2	0.02	0.001584	0.001165	0.001269	0.000813
$1000 \times ISE(\widehat{X})$						
sparse	0	0.01	0.050654	0.033898	0.047118	0.013144
	2	0.01	0.081423	0.701797	0.046949	0.012836
dense	2	0.01	0.038872	0.701015	0.006609	0.001236
	2	0.02	0.021748	0.004430	0.021222	0.004431

Table S.2: Simulation with "Two-peak" baseline gradient function g and linear ($p = 2$) $Z(\cdot, \theta_i)$s. $M = 30$ cubic B-spline basis functions are used for the fitting. Summary statistics for $ISE(\widetilde{g}) := \int (s_\beta \widehat{g}(x) - g(x))^2 dx$. $\widetilde{CV}$ is used for selection of $\lambda_{\boldsymbol{\beta}}$.

Sampling rate	μ_2	σ_ε	Mean(ISE)	SD(ISE)	Median(ISE)	MAD(ISE)
sparse	0	0.01	0.007634	0.006347	0.006167	0.004536
	2	0.01	0.005099	0.004786	0.003858	0.002970
dense	2	0.01	0.001853	0.001836	0.001339	0.001217
	2	0.02	0.003642	0.004631	0.002546	0.002077

Table S.3: Simulation with "Two-peak" baseline gradient function g and constant ($p_{tr} = 1$) or linear ($p_{tr} = 2$) $Z(\cdot, \boldsymbol{\theta}_i)$'s. Results are under the fitting procedure (with $M = 30$ cubic B-spline basis functions) using $p = p_{tr}$ and the "ideal" model selection criterion for λ_β.

True model	Sampling rate	Mean(ISE($\widehat{g}$))	SD(ISE($\widehat{g}$))	Median(ISE($\widehat{g}$))	MAD(ISE($\widehat{g}$))
$p_{tr} = 1$	sparse	0.000493	0.000197	0.000540	0.000204
	dense	0.000209	0.000165	0.000148	0.000137
$p_{tr} = 2$	sparse	0.000701	0.000416	0.000570	0.000310
	dense	0.000415	0.000354	0.000306	0.000270

Table S.4: Simulation with "Two-peak" baseline gradient function g and constant ($p_{tr} = 1$) or linear ($p_{tr} = 2$) $Z(\cdot, \boldsymbol{\theta}_i)$'s. Results are under the fitting procedure (with $M = 30$ cubic B-spline basis functions) using $\widetilde{CV}$ for selection of both λ_β and p. In almost all cases, $p = 2$ is selected.

True model	Sampling rate	Mean(ISE($\widehat{g}$))	SD(ISE($\widehat{g}$))	Median(ISE($\widehat{g}$))	MAD(ISE($\widehat{g}$))
$p_{tr} = 1$	sparse	0.001902	0.001450	0.001571	0.000614
	dense	0.000375	0.000231	0.000398	0.000248
$p_{tr} = 2$	sparse	0.001400	0.000998	0.001016	0.000539
	dense	0.000556	0.000436	0.000417	0.000279

Figure S.2: Simulation with "Two-peak" gradient function g and linear $Z(\cdot, \boldsymbol{\theta}_i)$ under the `sparse` case with $\boldsymbol{\mu}_\theta = (0, 0)^T$ and $\sigma_\varepsilon = 0.01$. **Left panel:** X-axis: x; Solid line: true $g(x)$; Dotted line: point-wise mean of $\widetilde{g}(x) = s_\beta \widehat{g}(x)$; Dash-dotted line : point-wise 5% and 95% percentiles of $\widetilde{g}(x)$. **Right panel:** X-axis: mean of $\overline{X}(t)$; Solid line: point-wise mean of $e^{\mu(t)} g(\overline{X}(t))$; Dotted line: point-wise mean of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$; Dash-dotted line : point-wise 5% and 95% percentiles of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$.

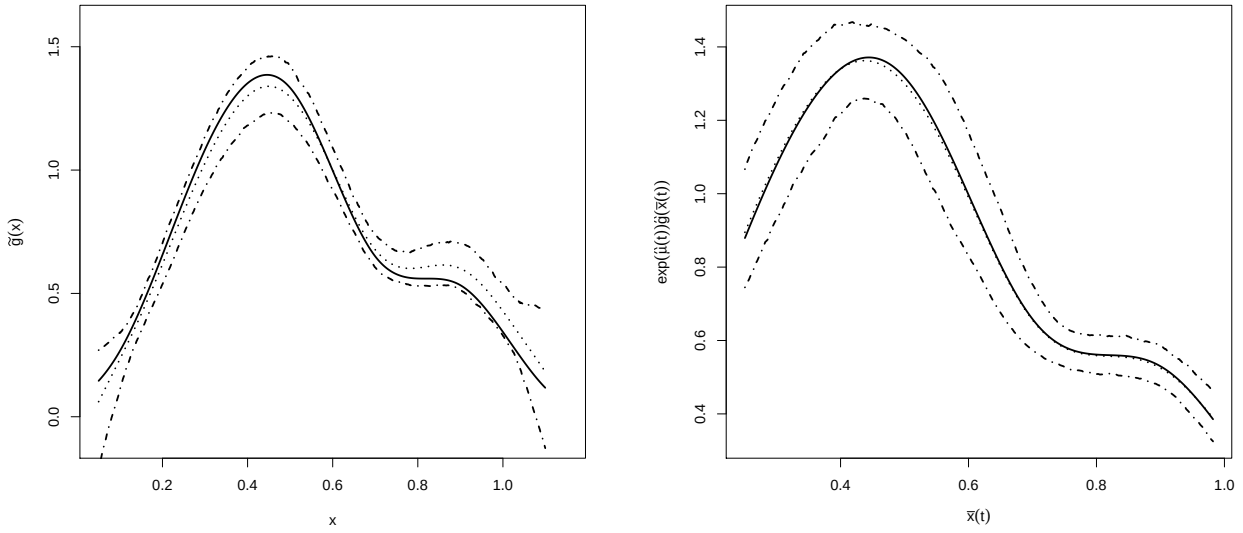


Figure S.3: Simulation with "Two-peak" gradient function g and linear $Z(\cdot, \boldsymbol{\theta}_i)$ under the `sparse` case with $\boldsymbol{\mu}_\theta = (0, 2)^T$ and $\sigma_\varepsilon = 0.01$. **Left panel:** X-axis: x; Solid line: true $g(x)$; Dotted line: point-wise mean of $\widetilde{g}(x) = s_\beta \widehat{g}(x)$; Dash-dotted line : point-wise 5% and 95% percentiles of $\widetilde{g}(x)$. **Right panel:** X-axis: mean of $\overline{X}(t)$; Solid line: point-wise mean of $e^{\mu(t)} g(\overline{X}(t))$; Dotted line: point-wise mean of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$; Dash-dotted line : point-wise 5% and 95% percentiles of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$.

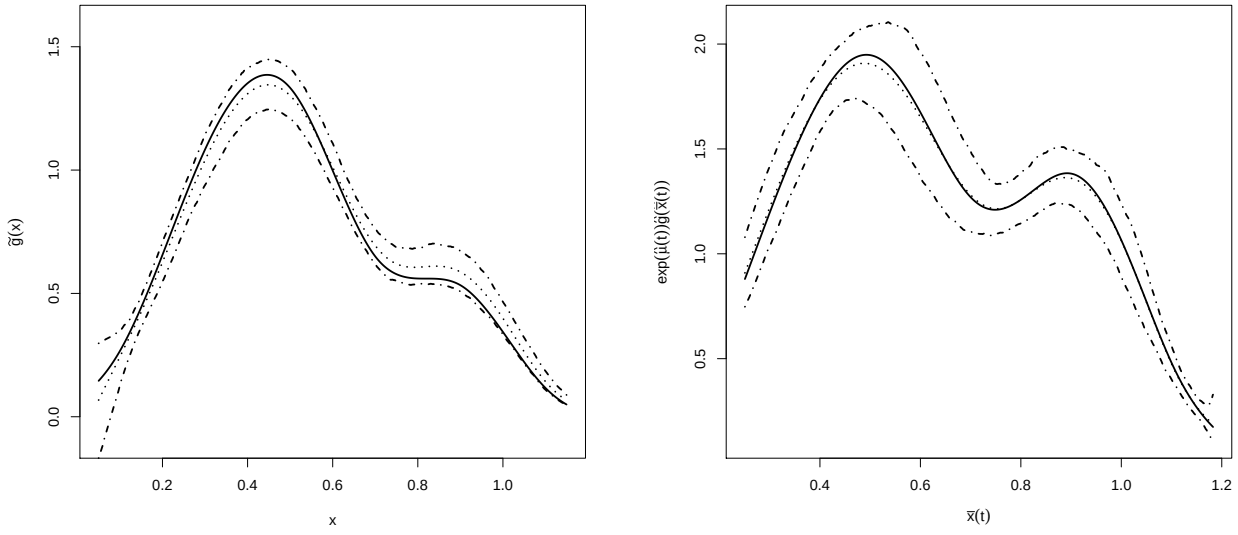
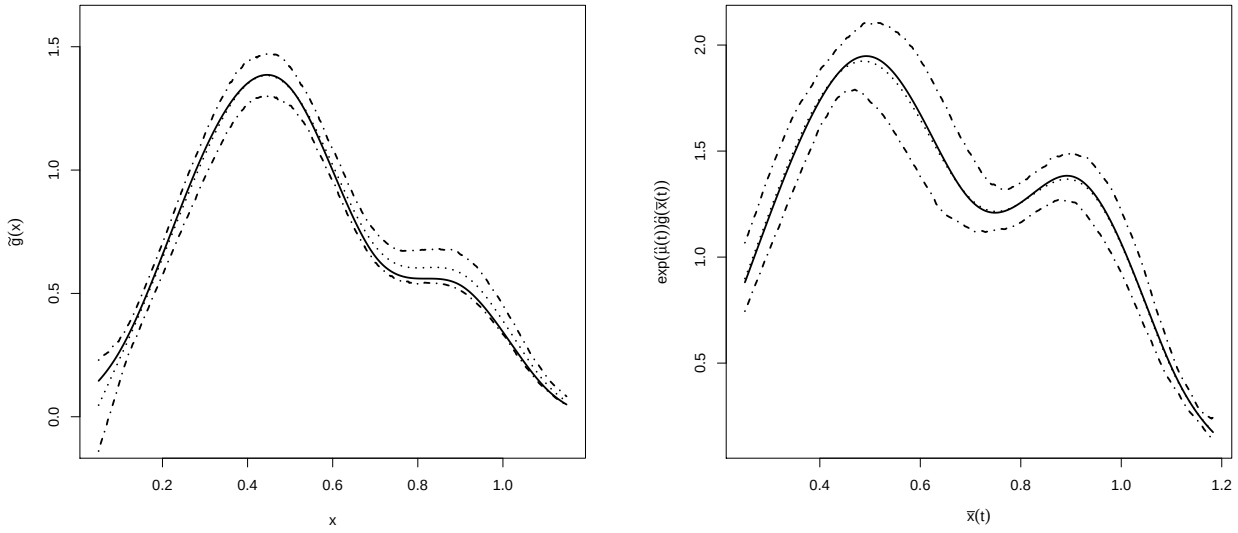


Figure S.4: Simulation with "Two-peak" gradient function g and linear $Z(\cdot, \boldsymbol{\theta}_i)$ under the `dense` case with $\boldsymbol{\mu}_\theta = (0, 2)^T$ and $\sigma_\varepsilon = 0.02$. **Left panel:** X-axis: x; Solid line: true $g(x)$; Dotted line: point-wise mean of $\widetilde{g}(x) = s_\beta \widehat{g}(x)$; Dash-dotted line : point-wise 5% and 95% percentiles of $\widetilde{g}(x)$. **Right panel:** X-axis: mean of $\overline{X}(t)$; Solid line: point-wise mean of $e^{\mu(t)} g(\overline{X}(t))$; Dotted line: point-wise mean of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$; Dash-dotted line : point-wise 5% and 95% percentiles of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$.



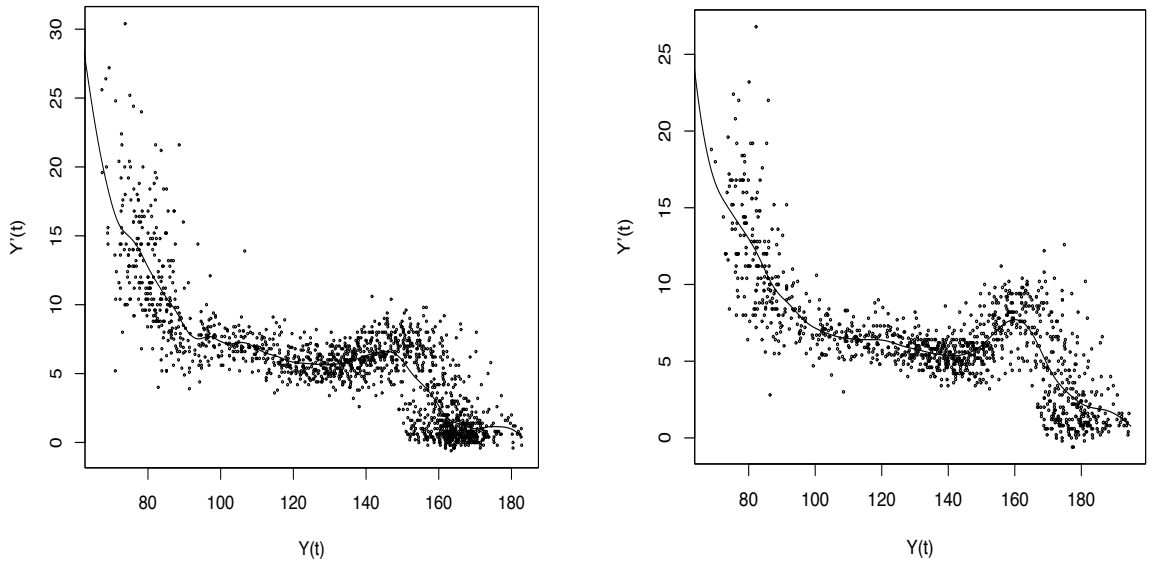


Figure S.5: Berkeley growth data. Empirical derivatives (divided differences) $Y'(t)$ against height measurements $Y(t)$ for female group (left panel) and male group (right panel).

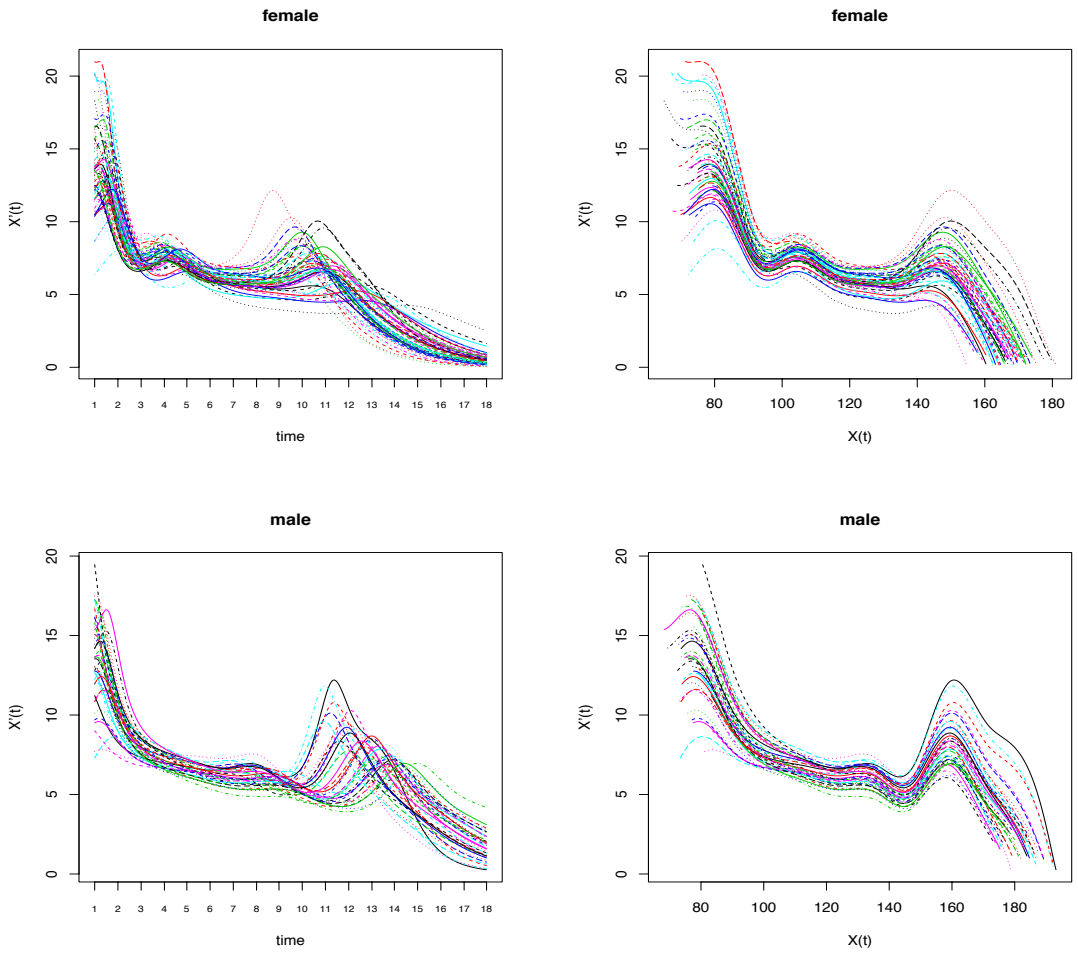


Figure S.6: Berkeley growth data. Fitted individual growth rate $\widehat{X}_i'(t)$ under the quadratic subject-specific effects ($p = 3$) against time t (left panel) and against heights $\widehat{X}_i(t)$ (right panels) for the female group (upper panel) and for the male group (lower panel).

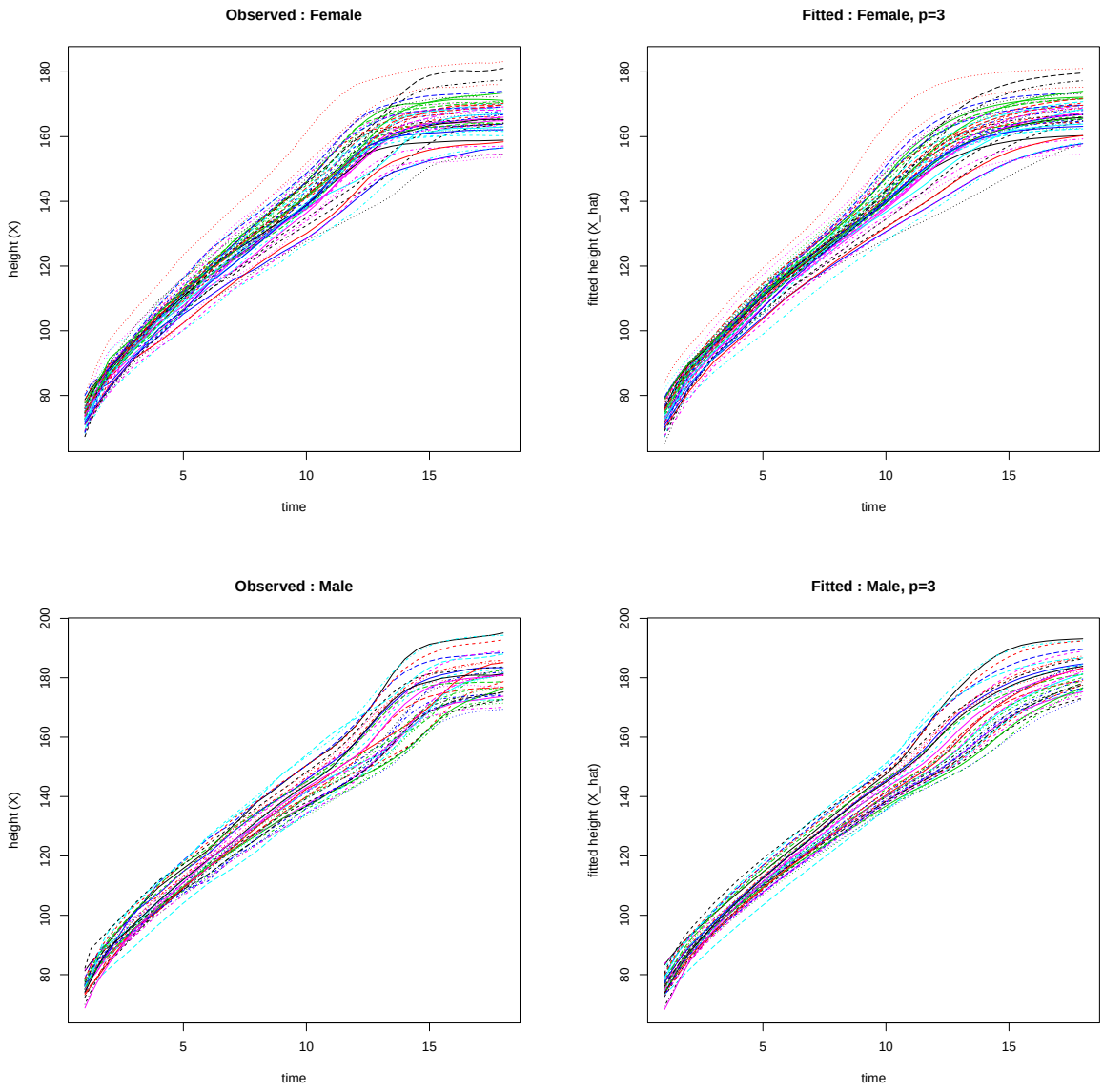


Figure S.7: Berkeley growth data. Observed (left panel) and fitted trajectories (right panel) under the quadratic subject-specific effects $p = 3$ for the female group (upper panel) and for the male group (bottom panel).

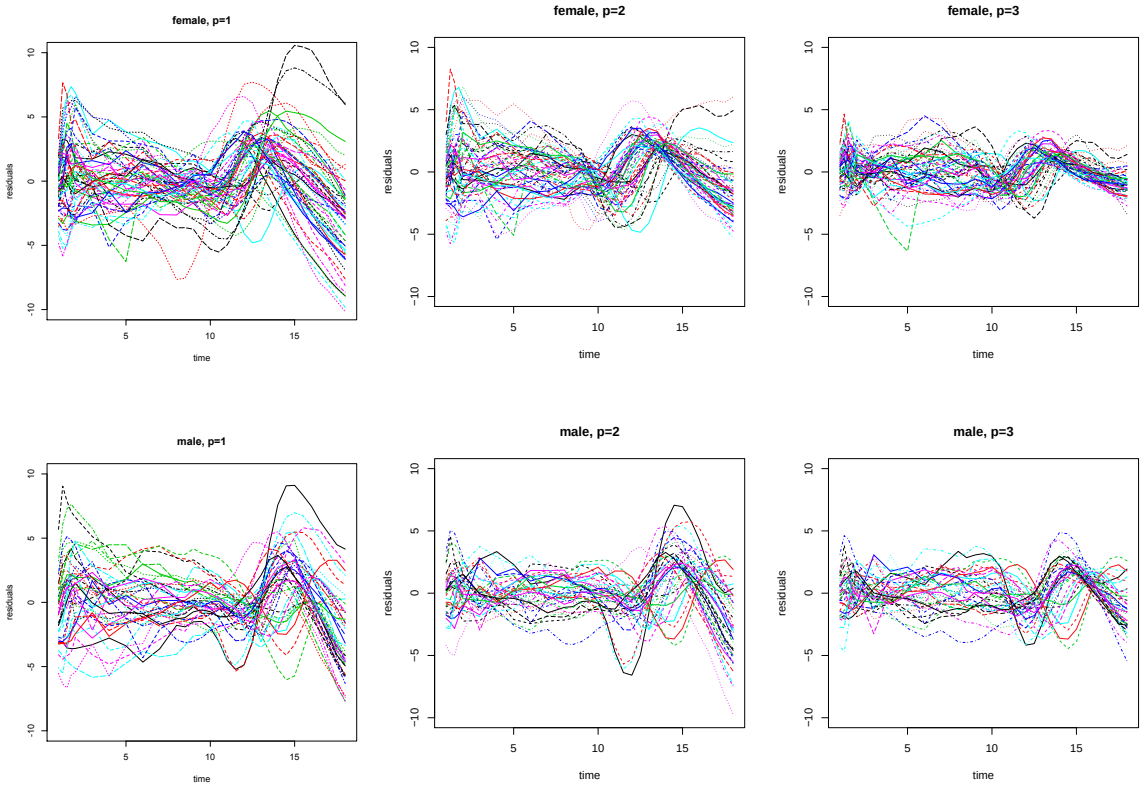


Figure S.8: Residuals of fitted trajectories based on full data for the female group with $p = 1, 2, 3$ (upper panel) and for the male group with $p = 1, 2, 3$ (lower panel).

S.8 Additional model selection results

Following the suggestions from one of the reviewers, we carried out additional numerical studies to further explore the model selection issue.

Simulated study. We expanded the simulation study under the same settings ($p_{tr} = 1, 2$) as reported in the paper and fit the models for $p = 1, 2, 3, 4$, fixing the parameter σ_{θ_k} to be 0.2 for $k \geq 2$, and $\sigma_{\theta_1} = 0.1$. CV scores corresponding to a set of representative simulation runs are reported in Figures S.9 (for $p_{tr} = 1$) and S.10 (for $p_{tr} = 2$). Notably, the CV curves flatten out for $p > p_{tr}$ in all the cases. For the `dense` setting (right panels), the CV scores are very close to each other for all values of $p \geq p_{tr}$, while for the `sparse` setting (left panels), there is a slight downward trend in most curves as p increases. The model selection performance based on the CV scores is summarized in Tables S.5 (for $p_{tr} = 1$) and S.6 (for $p_{tr} = 2$). These results show that there is a tendency for the CV criterion to select larger models, and larger models are more likely to be selected for sparse case. This observation is related to the fact that, for sparse case it is harder to distinguish the features of the baseline dynamics g from the subject-specific time-dependent effects $Z(t, \boldsymbol{\theta}_i)$, resulting in a certain degree of "practical lack of model identifiability" and over-fitting.

Table S.5: $p_{tr} = 1$: Percentages of selected p based on minimizing CV scores.

Sampling rate	$p = 1$	$p = 2$	$p = 3$	$p = 4$
sparse	0	16	42	42
dense	0	30	10	60

Another noticeable effect is the presence of a pronounced "elbow" in the CV score curves

Table S.6: $p_{tr} = 2$: Percentages of selected p based on minimizing CV scores.

Sampling rate	$p = 1$	$p = 2$	$p = 3$	$p = 4$
sparse	0	0	24	76
dense	0	35	40	25

at the true p when $p_{tr} = 2$ (Figure S.10). This clearly indicates that the CV criterion is rather sensitive to under-specification of p. It also suggests that the location of such an "elbow" could provide a more accurate estimate of the dimension of the random effects than that deduced from CV scores alone. Motivated by this, we quantify the relative changes in the CV scores and use a threshold $\tau > 0$ to detect significant changes. Accordingly, using $CV_{(k)}$ to denote the CV score for the model with $p = k$, if the fraction $|CV_{(k-1)} - CV_{(k)}|/CV_{(k-1)}$ is less than τ, then we treat the change in the CV scores between $p = k - 1$ and $p = k$ as insignificant. The largest k for which a significant change occurs is chosen as the optimal p. We tried small values of τ and, as a rule of thumb and for illustrations, use $\tau = 0.05$. The corresponding model selection results are reported in Tables S.7 (for $p_{tr} = 1$) and S.8 (for $p_{tr} = 2$). For both values of p_{tr}, model selection through this approach is very effective in the `dense` setting, and quite reasonable in the `sparse` setting. Comparing with results in Tables S.5 and S.6, we conclude that this approach is more reliable than the one based purely on the CV scores. Specifically, both the approaches guard very well against under-specification of p. However, the approach based on relative changes in the CV scores also tends to disregard models with p larger than p_{tr}, especially so for relatively dense samples.

Table S.7: $p_{tr} = 1$: Percentages of selected p based on relative change in CV scores (threshold $\tau = 0.05$).

Sampling rate	$p = 1$	$p = 2$	$p = 3$	$p = 4$
sparse	0	79	5	16
dense	100	0	0	0

Table S.8: $p_{tr} = 2$: Percentages of selected p based on relative change in CV scores (threshold $\tau = 0.05$).

Sampling rate	$p = 1$	$p = 2$	$p = 3$	$p = 4$
sparse	0	70	18	12
dense	0	100	0	0

Figure S.9: $p_{tr} = 1$: CV scores corresponding to $p = 1, 2, 3, 4$ for a set of simulation runs. Left panel: `sparse` setting, Right panel: `dense` setting.

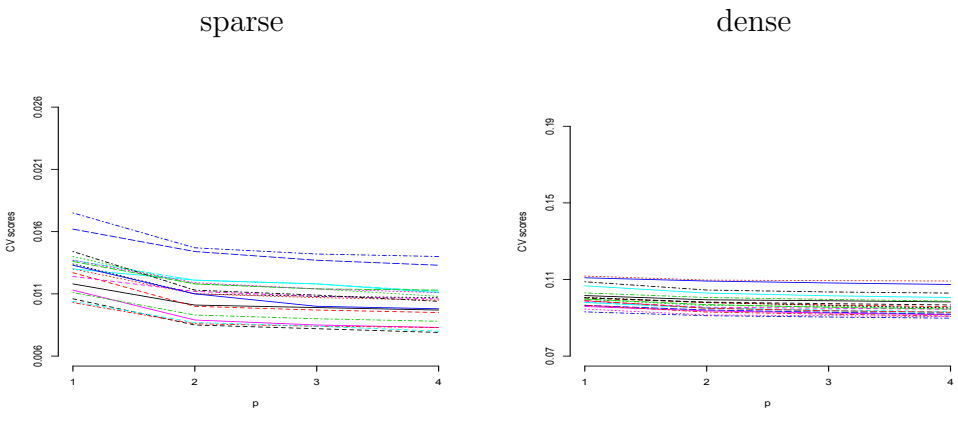
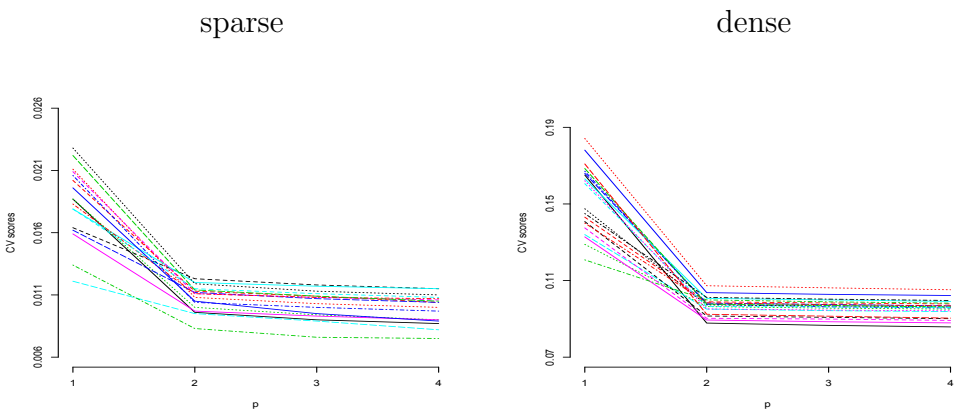


Figure S.10: $p_{tr} = 2$: CV scores corresponding to $p = 1, 2, 3, 4$ for a set of simulation runs. Left panel: `sparse` setting, Right panel: `dense` setting.



For the **Berkeley Growth Data**, we fitted the heights of both female and male subjects for $p = 4$. As is to be expected in such nonlinear models, higher value of p leads to some degree of instability. To address this, we set $\sigma_{\theta_4} = 2$, and $\sigma_{\theta_k} = 5$ for $k = 1, 2, 3$. As before, Σ_θ is chosen to be diagonal with diagonal elements $(\sigma^2_{\theta_k})_{k=1}^4$. This specification enables a straightforward comparison with the $p = 3$ case. We also keep all the other parameters the same as in the $p = 3$ case. Based on the CV score, the model with $p = 4$ is preferred. For a closer inspection, we compare the MISEs for each group. For the female group, the drop in MISE from $p = 3$ to $p = 4$ is modest (about 25%), while for the male group, the drop in MISE is a bit more significant (about 41%). We also compare the residuals for the two models for each group, as illustrated in Figures S.11 (female group) and S.12 (male group). These plots show that even though the overall spread of the residuals is not much different between $p = 3$ and $p = 4$, there is a reduction in the spread towards the right end point (beyond the age of 15 years) for both groups under $p = 4$ model, with more pronounced reduction for the male group. This reduction in errors, though moderate, is nevertheless reflected in the selection of the model with $p = 4$ over that with $p = 3$ by the CV criterion. However, as a further comparison, we also consider the estimation of the baseline gradient function g, which is of primary interest to us. In Figure S.13, we plot the adjusted baseline gradient function $e^{\widehat{\mu}(t)}\widehat{g}(\overline{X}(t))$ (where $\hat{\mu}(t) = Z(t, \widehat{\mu}_\theta)$) against $\overline{X}(t)$ (left panel: female group, right panel: male group). In each plot, the black curve corresponds to fit for $p = 3$ and the red curve corresponds to fit for $p = 4$. It can be seen that the two curves are nearly overlapping for both groups, indicating that there is very little difference between the fits of the gradient function corresponding to $p = 3$ and $p = 4$. Thus we decided

to report the result corresponding to $p = 3$ in the main manuscript for simplicity.

Figure S.11: Female group : comparison of residuals. Left panel: $p = 3$, Right panel: $p = 4$.

$p = 3$ $p = 4$

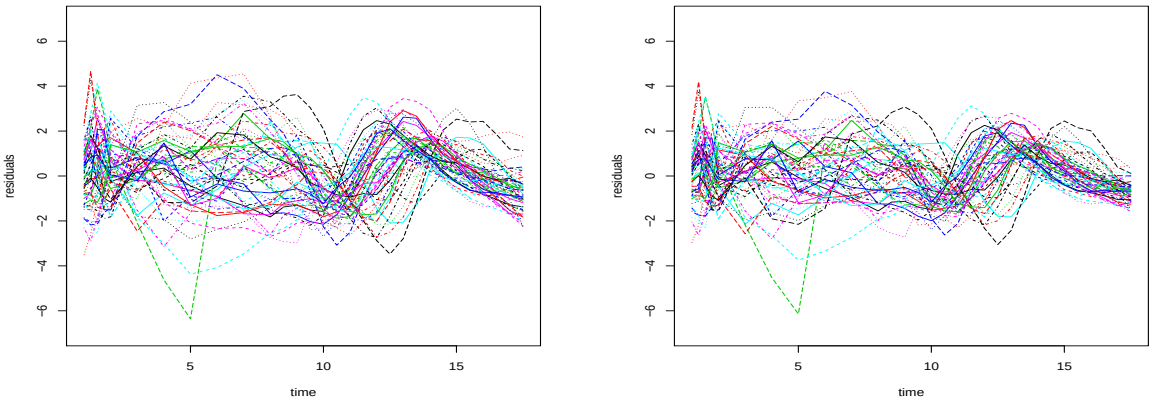


Figure S.12: Male group : comparison of residuals. Left panel: $p = 3$, Right panel: $p = 4$.

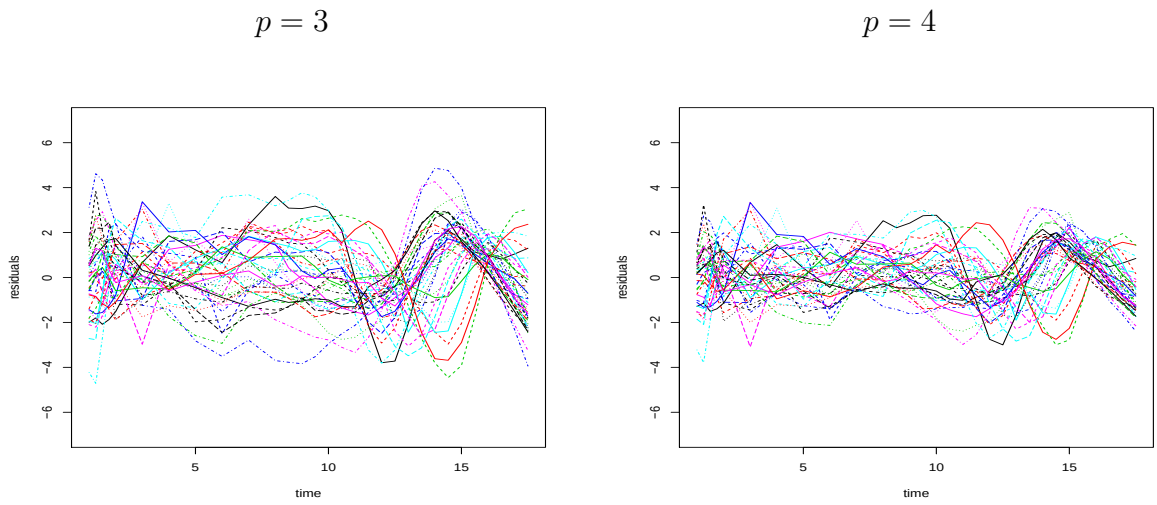


Figure S.13: Plot of adjusted baseline gradient function against average growth trajectory. Black curve: $p = 3$, Red curve: $p = 4$. Left panel: female group, Right panel: male group.