This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier’s archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights
Model determination and estimation for the growth curve model via group SCAD penalty

Jianhua Hu\textsuperscript{a,}\textsuperscript{*}, Xin Xin\textsuperscript{a,b}, Jinhong You\textsuperscript{a}

\textsuperscript{a}School of Statistics and Management, and Key Laboratory of Mathematical Economics, Shanghai University of Finance and Economics, Shanghai 200433, PR China
\textsuperscript{b}School of Mathematics and Information Sciences, Henan University, Kaifeng, Henan 475000, PR China

\begin{abstract}
The growth curve model is a useful tool for studying the growth problems, repeated measurements and longitudinal data. A key point using the growth curve model to fit data is determining the degree of polynomial profile form, choosing suitable explanatory variables, shrinking some regression coefficients to zero and estimating nonzero regression coefficients. In this paper, we propose a three-level variable selection approach based on weighted least squares with group SCAD penalty to handle the aforementioned problems. Considering the rows and columns of regression coefficient matrix as groups with overlap to control the polynomial order and variables, respectively, our proposed procedure enables us to simultaneously determine the degree of polynomial profile, identify the significant explanatory variables and estimate the nonzero regression coefficients. With appropriate selection of the tuning parameters, we establish the oracle property of the procedure and the consistency of the proposed estimation. We investigate the finite sample performances of our procedure in simulation studies whose results are very supportive, and also analyze a real data set to illustrate the usefulness of our procedure.
\end{abstract}

\section{Introduction}

A growth curve model was first summarized by Potthoff and Roy [21], and studied subsequently by many authors including [9,12,14,16,17,19,22,25]. It is already shown to be very useful, particularly for studying growth problems on short time series, repeated observations often measured over multiple time points on a particular characteristic to investigate the temporal pattern of change on the characteristic [4] and longitudinal data especially with serial correlation [15] in a variety of scientific disciplines, such as medical research, biology, economics, education, forestry and so on. The interested reader can refer to [18,20] for a more detailed discussion and illustration of the usefulness of the growth curve model.

The basic idea of the growth curve model is to introduce some known functions, usually polynomial functions, so as to capture patterns of change for time-dependent measurements. In data analysis, the degree of polynomial profile is often unknown. This allows the possibility of selecting an underfitted (or overfitted) model, leading to biased (or inefficient) estimators and predictions. Meanwhile, if there are too many explanatory variables which are not important, we need to choose suitable explanatory variables for efficiency and accuracy. Therefore, a key point using the growth curve model to fit the data is simultaneously determining the degree of polynomial profile form, selecting suitable explanatory variables, choosing zero regression coefficients and estimating the nonzero regression coefficients.

Usually, a growth curve model can be written as

\[ Y = X\Theta Z^T + \varepsilon \]  \hspace{1cm} (1.1)

\textsuperscript{*}Corresponding author.
E-mail address: frank.jianhuahu@gmail.com (J. Hu).
where \( Y \) is the observation matrix of the response consisting of \( p \) repeated measurements taken on \( n \) individuals, \( X \) is the treatment design matrix with order \( n \times m \), \( Z \) is the profile matrix with order \( p \times q \), and \( \theta \) is the unknown regression coefficient matrix with order \( m \times q \). Assume that observations on individuals are independent, so that the rows of the random error matrix \( \varepsilon \) are independent and identically distributed by a distribution with mean zero and covariance matrix \( \Sigma \).

The treatment design matrix \( X \) discussed in this paper is assumed to comprise of intercept items and explanatory variables. The explanatory variables themselves may be continuous or categorical. Data sets with continuous explanatory variables often appear in real statistical problems. For example, in the problem considering children’s reading recognition, cognitive stimulation and emotional support for children at home will be possible explanatory variables which may be continuous.

Several authors have investigated the problem of determining the degree of polynomial profile form in the model (1.1). Fujikoshi and Rao [8] considered the one of selecting the covariables in the growth curve model, but did not consider the problem of selecting the degree of freedom. Satoh et al. [23] treated determination of the degree of a polynomial growth curve as a problem of variable selection. They proposed the corresponding \( C_p \) and AIC for the situation where the matrix \( X \) is a design matrix across individuals on \( m \) groups. Here it is not necessary to select the columns of \( X \) though the criteria in [23] is also used to determine the degree of polynomial profile and select the explanatory variables simultaneously.

It is well known that the traditional variable selections such as \( C_p \), AIC and Bayesian information criterion (BIC) do not shrink the regression coefficients to zero. This results in the accuracy of the regression coefficient estimators in the final model being somewhat difficult to understand.

To overcome the limitation, we consider an alternative method, to the traditional best subset variable selections, that can simultaneously estimate regression coefficients and shrink some regression coefficients to zero, thereby, removing them from the final model. For linear regression models, Fan and Li [6] novelty proposed a family of variable selection procedures by the smoothly clipped absolute deviation penalty (SCAD). They showed that the proposed method outperforms the best subset variable selection in terms of computational cost and stability. An attractive feature of it is that with a proper choice of regularization parameters, the resulting estimator possesses an oracle property, namely, the true zero regression coefficients are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel were known in advance.

To simultaneously choose the degree of polynomial profile and explanatory variables, we shall propose a three-level variable selection procedure based on group SCAD penalty with a special overlap so that we can simultaneously determine the order of polynomial profile form, identify the significant explanatory variables, shrink some regression coefficients to zero and estimate nonzero regression coefficients.

Our contributions are to divide the rows and columns of regression parametric matrix in the model (1.1) into column and row groups with a special overlap, discover the equivalent relationships between row group and variable selection and between column group and determine of the order of polynomial profile form, then use a smoothly clipped absolute deviation penalty function to penalize row groups, column groups and regression coefficients, and finally minimize the penalized (weighted) least squares with three-level SCAD penalty to lead to a penalized (weighted) least squares estimator of the regression parametric matrix. It will be shown that the oracle property of the procedures and the consistency of the estimation can be achieved under an appropriate selection of the tuning parameters. Our results indicate that our proposed procedure outperforms the best subset selection and the method that SCAD is directly applied to the vector version of model (1.1).

The rest of this paper is as follows. A least squares estimation via a group SCAD penalty with a special overlap is proposed in Section 2. The estimation of the error covariance matrix is provided in Section 3. Based on the estimated covariance, we proposed a three-level variable selection procedure via weighted least squares estimation with a group SCAD penalty in Section 4. The selection of the tuning parameters is discussed in Section 5. Some simulation studies are conducted in Section 6. One practical problem, which can be characterized by the model (1.1), is illustrated in Section 7. Finally, the brief concluding remarks are stated in Section 8. The proof of the main results are collected in the Appendix.

2. Least squares estimation with group SCAD penalty

For convenience of notation, we write \( \Theta \) into \( \Theta = (\Theta_y)_{2 \times 2} \) with \( \Theta_{11} \) a \( m_0 \times q_0 \) matrix. Here \( q_0 (\leq q) \) is the true degree of polynomial profile form and \( m_0 (\leq m) \) is the true number of explanatory variables. Let \( \theta_j \) denote the \( j \)th \((1 \leq j \leq q)\) column of matrix \( \Theta \) and \( \theta_i \) denote the \( i \)th \((1 \leq i \leq m)\) row of matrix \( \Theta \). There is an \( m_0 q_0 \times m_0 q_0 \) elementary transformation matrix \( L_0 = (L_{10}, L_{20}) \) such that \( \text{vec}(\Theta_{11}) = L_{10} \beta_{10} + L_{20} \beta_{20} \), where \( \beta_{10} \) is a \( d_0 \)-dimensional vector with each element nonzero and \( \beta_{20} = 0 \). The following is an example with

\[
\Theta_{11} = \begin{pmatrix}
\theta_{11} & 0 \\
0 & \theta_{32}
\end{pmatrix}, \quad L_{10} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}, \quad L_{20} = \begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}, \quad \beta_{10} = \begin{pmatrix}
\theta_{11} \\
\theta_{12} \\
\theta_{13} \\
\theta_{32}
\end{pmatrix}, \quad \beta_{20} = \begin{pmatrix}
0 \\
0
\end{pmatrix}.
\]

Here the vec operator transforms a matrix into a vector by stacking the rows of the matrix one under another.
The \( j \)th column \( \theta_j \) of matrix \( \Theta \) corresponds to the \( (j - 1) \)th degree item of the polynomial or the \( j \)th column of \( Z \). It is easily seen that \( \theta_j = 0 \) or \( \| \theta_j \| = 0 \) if and only if the \( j \)th item (namely the \( (j - 1) \)th order) of polynomial profile form does not exist. The \( i \)th row \( \theta_i \) of matrix \( \Theta \) corresponds to the \( i \)th column of \( X \). Similarly, it is not difficult to see that \( \| \theta_i \| = 0 \) implies that the \( i \)th explanatory variable is a nonsignificant variable, thereby, removed from the model (1.1).

Including too extra terms in the polynomial profile form causes extra variability and inflated confidence intervals for regression parameters, and a suitable explanatory variable can make the model efficient and accurate. Therefore, based on the structure of the model (1.1) and the above division, grouped variable selection is a promising (maybe indispensable) tool for statistical data analysis by using the model (1.1). Regarding this topic, Yuan and Lin [29] successfully extended the LASSO [24] to select a grouped variable. Huang et al. [11] proposed a group bridge approach for variable selection. Breheiny and Huang [2] addressed some penalized methods for bi-level selection. Wei and Huang [28] studied consistent group selection in high-dimensional linear regression. Huang et al. [10] provide a nice selective review of group selection concerning methodological developments, theoretical properties, and computational algorithms.

At the beginning of this paper, columns and rows of a regression parametric matrix are regarded as groups. They are overlapping groups with prior known group structures. We shall consider extending the nonconcave penalized method to group selection of the variables, determination of the degree of polynomial profile form and individual selection of regression coefficients in the model (1.1). Specially, the proposed three-level variable selection is a combination of the nonconcave penalization for each column group, each row group of \( \Theta \) and individual parameters. A least squares with three-level group nonconcave penalty takes the form

\[
\mathcal{L}(\Theta) = \frac{1}{2} \text{tr}((Y - X\Theta Z')' (Y - X\Theta Z')) + n \sum_{i=1}^{m} p_{e_{k_i}}(\|\theta_i\|) + n \sum_{j=1}^{q} p_{e_{j}}(\|\theta_j\|) + n \sum_{i=1}^{m} \sum_{j=1}^{p} p_{e_{q_{i}}} (\theta_{ij}),
\]

where \( \| \cdot \| \) denotes the \( L_2 \) norm, \( \theta_j \) denotes the \( j \)th column vector and \( \theta_i \) denotes the \( i \)th row vector of \( \Theta \), \( p_{e_{k_i}} (\cdot) \), \( p_{e_{j}} (\cdot) \) and \( p_{e_{q_{i}}} (\cdot) \) are penalty functions, and \((k_i, j, \lambda_{ij})\) are a series of triple tuning parameters. Then

\[
\hat{\Theta}_{gsp} = \arg\min_{\Theta} \mathcal{L}(\Theta),
\]

where the notation \( gsp \) stands for the group SCAD penalty. The set of significant variables and the set of the nonzero items of polynomial profile form can be uniquely determined by solving the optimization problem (2.2).

The tuning parameters control the model complexity and can be selected by some data-driven methods. This will be investigated in Section 5. In addition, many penalty functions, such as the family of \( L_p \)-penalty \((r \geq 0)\), have been used for penalized least squares and penalized likelihood in various parametric models. Antoniadis and Fan [1] and Fan and Li [6] provided various insights into how a penalty function should be chosen. They advocated that a good penalty function should yield an estimator with three properties mentioned in the Introduction. According to [6] a simple penalty function, which results in an estimator with the three desired properties, is the SCAD penalty. Given \( \zeta > 0 \), the SCAD penalty at \( \theta \) is

\[
pe_{\zeta}(\theta) = \begin{cases} 
\zeta |\theta|, & |\theta| \leq \zeta, \\
-\zeta (\theta^2 - 2a\zeta |\theta| + \zeta^2) / [2(a - 1)], & |\theta| > a\zeta.
\end{cases}
\]

One can gain more insight through its first derivative:

\[
pe'_{\zeta}(\theta) = \begin{cases} 
\zeta \text{ sign}(\theta), & |\theta| \leq \zeta, \\
(\zeta a - |\theta|)/(a - 1) \text{ sign}(\theta), & |\theta| > a\zeta.
\end{cases}
\]

Based on \( pe'_{\zeta}(\theta) \) it is easy to see that the SCAD penalty is continuously differentiable on \((-\infty, 0) \cup (0, \infty)\), but singular at 0. Its derivative vanishes outside \([-a\zeta, a\zeta]\). As a consequence, the SCAD penalized method can produce sparse solutions and unbiased estimates for large regression coefficients. The readers can refer to [6] for more details about the SCAD penalty. In addition, \( a = 3.7 \) suggested in [6] will be used in the whole paper.

To study the asymptotic properties of the resulting estimator \( \hat{\Theta}_{gsp} \), we first establish it convergence rate. Denote

\[
\alpha_n = \max_{1 \leq i \leq m, 1 \leq j \leq q} \{ |pe'_{\zeta}(\|\theta_i\|) : \theta_i \neq 0 \}, \quad b_n = \max_{1 \leq i \leq m, 1 \leq j \leq q} \{ |pe'_{\zeta}(\|\theta_j\|) : \theta_j \neq 0 \},
\]

\[
\alpha_n^d = \max_{1 \leq i \leq m, 1 \leq j \leq p} \{ |pe'_{\zeta}(\|\theta_{ij}\|) : \theta_{ij} \neq 0 \}, \quad b_n^d = \max_{1 \leq i \leq m, 1 \leq j \leq p} \{ |pe'_{\zeta}(\|\theta_{ij}\|) : \theta_{ij} \neq 0 \},
\]

where \( \theta \) is a \( d \)-dimensional zero vector.

Corresponding to the partition for \( \Theta \), we write \( \hat{\Theta}_{gsp} = (\hat{\Theta}_{gsp})_{2 \times 2} \), where \( \hat{\Theta}_{11} \) is an \( m_0 \times q_0 \) matrix such that vec(\( \hat{\Theta}_{11} \)) = \( L_{10} \hat{\Theta}_{10} + L_{20} \hat{\Theta}_{20} \). According to the partition of \( \Theta \), we write \( Z' = (Z_1', Z_2')' \) with \( Z_1' \) being a \( p \times q_0 \) matrix; \( X = (X_1, X_2) \) with \( X_1 \) being an \( n \times m_0 \) matrix.
Assumption A. Assume that \( \lim_{n \to \infty} n^{-1}X_i^T X_1 = R_{11} > 0 \) and \( Z_i^T Z_1 > 0 \) hold.

If the row vectors of the design submatrix \( X_1 \) are iid samples from a distribution with covariance \( R_{11} \), then \( X_i^T X_1/n \) converges almost surely to \( R_{11} \) by Kolmogorov’s strong law of large numbers.

**Theorem 2.1.** Under Assumption A, if \( \alpha_i \), \( \beta_i \), \( \alpha^d_i \), \( \beta^d_i \) and \( \beta^d_i \) tend to zero as \( n \to \infty \), then with probability tending to one, there exists a local minimizer \( \hat{\Theta}_{gsp} \) of \( \mathcal{L}(\Theta) \) such that \( \| \hat{\Theta}_{gsp} - \Theta \| = O_p(n^{-1/2} + \alpha_i^2 + \beta_i^2) \), where \( \| \cdot \| \) denotes the trace norm.

Theorem 2.1 demonstrates how the rate of convergence of \( \Theta \) depends on \( \kappa_i \), \( \varsigma_j \) and \( \lambda_{ij} \). To achieve the root-\( n \) convergence rate, we have to take \( \kappa_i \), \( \varsigma_j \) and \( \lambda_{ij} \) small enough so that \( \alpha_i^2 + \beta_i^2 + \beta^d_i \) is of order \( o(n^{-1/2}) \).

Next we establish the oracle property for \( \hat{\Theta}_{gsp} \).

**Theorem 2.2** (Oracle Property). Under Assumption A, if \( \kappa_i \to 0 \), \( \varsigma_j \to 0 \), \( \lambda_{ij} \to 0 \) and \( \sqrt{n}\kappa_i \to \infty \), \( \sqrt{n}\varsigma_j \to \infty \), \( \sqrt{n}\lambda_{ij} \to \infty \) as \( n \to \infty \), then, with probability tending to 1, the root-\( n \) consistent local minimizer \( \hat{\Theta}_{gsp} \) must satisfy

(i) (Sparsity) \( \hat{\beta}_{10} = 0_{m \times (q - m)} \); \( \hat{\Theta}_{21} = 0_{(m - m_2) \times q_0} \); \( \hat{\Theta}_{22} = 0_{(m - m_2) \times (q - q_0)} \).

(ii) (Asymptotic normality)

\[
  \sqrt{n} \Omega_n^{-1/2} (\hat{\beta}_{10} - \beta_{10}) \quad \text{converges in distribution to } N_{m \times q_0}(0, 1),
\]

where

\[
  \Omega_n = \mathbb{E} \left( X_i^T X_1 \otimes Z_i^T Z_1 \right) \left( L_{10} \right)^{-1} \left( X_i^T X_1 \otimes Z_i^T Z_1 \right)^T \left( L_{10} \right)^{-1}
\]

and \( \hat{\Sigma} \) defined in the following section.

\( \hat{\Theta}_{gsp} \) does not take the within cluster correlation into account, hence may not be asymptotically efficient. However, from Theorems 2.1 to 2.2, we can see that they are consistent. So they may serve as initial estimators and more efficient estimators may be constructed based on them.

### 3. Consistent estimation of covariance matrix

Usually, to construct an efficient estimator we need to have a consistent estimator of the error covariance matrix \( \Sigma \). To estimate \( \Sigma \), let

\[
  \hat{\Sigma} = \frac{1}{n} (Y - X \hat{\Theta}_{gsp} Z^\top)^\top (Y - X \hat{\Theta}_{gsp} Z^\top).
\]

**Assumption B.** Assume that the fourth moments are bounded, namely, \( E \| e_i \|^4 < \infty \).

For \( \hat{\Sigma} \) we have the following asymptotic result.

**Theorem 3.1.** Under conditions in Theorem 2.2 and Assumption B, it holds that

\[
  \sqrt{n} \text{Vech}(\hat{\Sigma}) - \text{Vech}(\Sigma) \to_d N \left( 0_{(p(p + 1))/2}, L_p \text{Cov}(e_i \otimes e_i) L^\top_p \right) \quad \text{as } n \to \infty,
\]

where \( e = (e_1, \ldots, e_n)^\top \), Vech is a column stacking operator that stacks only the elements on and below the main diagonal of a matrix, and \( L_p = (L^\top_{p - k + 1}, \ldots, L^\top_p) \) is the \( 1/2 (p + 1) \times p^2 \) elimination matrix with \( L_{pk} = (0_{(p - k + 1) \times (k - 1) (p + 1)}, 1_{p - k + 1}, 0_{(p - k + 1) \times (p - k)}) \) for \( k = 1, \ldots, p \).

Based on \( \hat{\Sigma} \), the efficient variable selection and estimator of the parametric matrix can be constructed. This will be shown in the following section.

### 4. Weighted least squares estimation with group SCAD penalty

The variable selection and estimation procedure in Section 2 may be less efficient since it does not take the within cluster correlation into account. In this section, we will take the within cluster correlation into account and construct more efficient variable selector and estimator. To be more precise, we propose the following weighted least squares with group SCAD penalty

\[
  \mathcal{L}^w(\Theta) = \frac{1}{2} \text{tr}((Y - X \Theta Z^\top) \hat{\Sigma}^{-1} (Y - X \Theta Z^\top)^\top) + n \sum_{i=1}^m p e_i^\top (\| \theta_i \|) + n \sum_{j=1}^q p e_j^\top (\| \theta_j \|) + n \sum_{i=1}^m \sum_{j=1}^q p e_{ij}^\top (\theta_{ij}) \quad (4.1)
\]
where \((\kappa^w, \varsigma_j^w, \lambda^w)\) are another series of triple tuning parameters, which may be different with the series of triple tuning parameters \((\kappa_i, \varsigma_j, \lambda_j)\). Then

\[
\hat{\Theta}_\text{gsp}^w = \arg\min_{\Theta} \mathcal{L}^w(\Theta).
\] (4.2)

Similarly, the set of significant variables and the set of the nonzero items of polynomial profile form can be uniquely determined by solving the optimization problem (4.2).

Denote

\[
a_n^w = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\|\theta_j\|) : \theta_j \neq 0^m_i \}, \quad b_n^w = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\|\theta_j\|) : \theta_j = 0^m_i \},
\]

\[
a_n^w = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\|\theta_j\|) : \theta_j \neq 0^m_j \}, \quad b_n^w = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\|\theta_j\|) : \theta_j = 0^m_j \}
\]

and

\[
a_n^{idw} = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\theta_j) : \theta_j \neq 0 \}, \quad b_n^{idw} = \max_{1 \leq j \leq m} \{ |p_{ij}^w(\theta_j) : \theta_j \neq 0 \}.
\]

For \(\hat{\Theta}_\text{gsp}^w\) we have the following asymptotic result.

**Theorem 4.1.** Under Assumptions A and B, if \(a_n^w, a_n^w, b_n^w, b_n^w, a_n^{idw}\) and \(b_n^{idw}\) tend to zero as \(n \rightarrow \infty\), then with probability tending to one, there exists a local minimizer \(\hat{\Theta}_\text{gsp}^w\) of \(\mathcal{L}^w(\Theta)\) such that \(\|\hat{\Theta}_\text{gsp}^w - \Theta\| = O_p(n^{-1/2} + a_n^w + a_n^w + a_n^{idw})\).

Next we establish the oracle property for \(\hat{\Theta}_\text{gsp}^w\). Corresponding to the partition for \(\Theta\), we write \(\hat{\Theta}_\text{gsp}^w = (\hat{\Theta}_1^w)_2 \times 2\), where \(\hat{\Theta}_1^w\) is an \(m_0 \times q\) matrix. Similarly, there is an \(m_0 q_0 \times m_0 q_0\) elementary transformation matrix \(L_0^w = (L_0^{w_1} L_0^{w_2})\) such that \(\text{vec}(\hat{\Theta}_1^w) = L_0^{w_1} \hat{\beta}_0^w + L_0^{w_2} \hat{\beta}_0^w\).

**Theorem 4.2** (Oracle Property). Under Assumptions A and B, if \(\kappa_i \rightarrow 0, \varsigma_j \rightarrow 0, \varsigma_j \rightarrow 0, \varsigma_j \rightarrow 0\) and \(\sqrt{n} \varsigma_j \rightarrow \infty, \sqrt{n} \varsigma_j \rightarrow \infty, \sqrt{n} \varsigma_j \rightarrow \infty\) as \(n \rightarrow \infty\), then, with probability tending to 1, the root-\(n\) consistent local minimizer \(\hat{\Theta}_\text{gsp}^w\) must satisfy

(i) (Sparsity) \(\hat{\beta}_0^w = 0_{m_0 q_0 - q_0}, \hat{\Theta}_1^w = 0_{m_0 \times (q - q_0)}, \hat{\Theta}_1^w = 0_{(m - m_0) \times q}, \hat{\Theta}_2^w = 0_{(m - m_0) \times (q - q_0)}\).

(ii) (Asymptotic normality)

\[
\sqrt{n} \Sigma_{w_1}^{-1/2} (\hat{\beta}_0^w - \beta_0) \text{ converges in distribution to } N_{m_0 q_0}(0, I),
\]

where \(\Sigma_{w_1} = n(L_1^{w_1} X_1^* X_1 \otimes Z_1^* \Sigma^{-1} Z_1 L_1^{w_1})^{-1}\).

5. Calculation and selection of tuning parameters

We here just consider the calculation of (2.2) and selecting \(\varsigma_j\) and \(\kappa_i\). The calculation of (4.2) and selecting \(\varsigma_j^w\) and \(\tau_j^w\) follow the same line.

Because the SCAD penalty function in (2.2) is irregular at the origin and may not have a second derivative at some points, it is difficult to find \(\hat{\Theta}_\text{gsp}\). Following [6,26], we use an iterative algorithm based on local quadratic approximation to the non-convex penalty \(p_{ij}(\theta_j)\), \(p_{ij}(\|\theta_j\|)\) and \(p_{ij}(\|\theta_j\|)\) to find \(\hat{\Theta}_\text{gsp}\). Given an initial value \(\Theta^0\) with \(\theta_j^0 \neq 0, p_{ij}(\theta_j)\) can be approximated by a quadratic form

\[
p_{ij}(\theta_j) \approx p_{ij}(\theta_j^0) + \frac{p_{ij}'(\theta_j^0)}{2\theta_j^0}(\theta_j - (\theta_j^0)^2),
\]

\(p_{ij}(\|\theta_j\|)\) can be approximated by a quadratic form

\[
p_{ij}(\|\theta_j\|) \approx p_{ij}(\|\theta_j\|) + \sum_{s=1}^{q} \frac{p_{ij}'(\|\theta_j\|)}{2\|\theta_j\|} (\theta_j - (\theta_j^0)^2),
\]

and \(p_{ij}(\|\theta_j\|)\) can be approximated by a quadratic form

\[
p_{ij}(\|\theta_j\|) \approx p_{ij}(\|\theta_j\|) + \sum_{s=1}^{m} \frac{p_{ij}'(\|\theta_j\|)}{2\|\theta_j\|} (\theta_j - (\theta_j^0)^2),
\]
where \( pe'(\cdot) \) is the first derivative of \( pe(\cdot) \). Set

\[
\hat{\Sigma} = \text{diag}(a_{11}, \ldots, a_{1q}, \ldots, a_{m1}, \ldots, a_{mq}) \quad \text{for} \quad a_{ij} = pe'_{\hat{\eta}}(\theta^0_{ij})/(2\theta^0_{ij}).
\]

\[
\Sigma^0 = \text{diag}\left\{ pe'_{\hat{\eta}}(\theta^0_1)/2\theta^0_1, \ldots, pe'_{\hat{\eta}}(\theta^0_m)/2\theta^0_m \right\} \otimes I_q,
\]

\[
\Sigma^0 = I_m \otimes \text{diag}\left\{ pe'_{\hat{\eta}}(\theta^0_1)/2\theta^0_1, \ldots, pe'_{\hat{\eta}}(\theta^0_q)/2\theta^0_q \right\}.
\]

Then, the objective function on the right side of (2.1) can, equivalently, be approximated by

\[
\frac{1}{np} \text{tr}(Y - X\hat{\Theta}Z^\top)^\top(Y - X\hat{\Theta}Z^\top) + \text{vec}(\theta)^\top(\Sigma^0_0 + \Sigma^0_0 + \Sigma^0_0)\text{vec}(\theta).
\]

Given \( \theta^0 \), this approximate objective function is minimized by

\[
\text{vec}(\theta) = \frac{1}{np}\left\{ \frac{1}{np}X^\top X \otimes Z^\top Z + \Sigma^0_0 + \Sigma^0_0 + \Sigma^0_0 \right\}^{-1} (X^\top Z^\top)\text{vec}(Y),
\]

and

\[
(\theta_1, \ldots, \theta_m)^\top = \text{vec}(\hat{\theta}), \quad (\theta_1, \ldots, \theta_q)^\top = \text{vec}(\hat{\Theta}).
\]

The iterative algorithm is outlined as follows:

Step 1. Initialize \( \theta^{(1)} \). For example, set \( \theta^{(1)} = (X^\top X)^{-1}X^\top YZ(Z^\top Z)^{-1} \).

Step 2. For \( k = 1, 2, \ldots \), set \( \theta^{(k)} = \text{vec}(\theta^{(k-1)}) \) by (5.2), and compute \( \text{vec}(\theta^{(k+1)}) \) by the formula (5.1).

Step 3. Iterate Step 2 until convergence and use \( \hat{\Theta} \) to denote the final result.

In Step 2 at any iteration, if some \( \|\theta^{(k)}_j\| \) is smaller than the cutoff value \( 10^{-2} \), we set \( \hat{\theta}_j = 0_m \), if some \( \|\theta^{(k)}_j\| \) is smaller than the cutoff value \( 10^{-2} \), we set \( \hat{\theta}_j = 0 \).

It is worthwhile to note that the cutoff value \( 10^{-2} \) is available. When the minimum of the absolute value of parameters is less than \( 10^{-2} \), we can change the scale via multiplying two sides of the model (2.1) by a suitable number \( 10^k \) in order to avoid the error of computation.

Theorem 2.2 indicates that the proposed variable selection procedure possesses the oracle property. However, this attractive feature relies on the tuning parameters. To this end, we adopt BIC selector proposed in [27] to choose the regularization parameters \( k_1, k_2, \ldots, k_q, \) and \( \lambda_1, \lambda_2, \ldots, \lambda_m \). Since it is computationally expensive to minimize BIC, defined below, with respect to the \( q \)-dimensional regularization parameters, we follow the approach of [7] to set \( \xi_j = \xi \sum_{k=1}^m \text{SE}(\hat{\theta}_j^0) \), where \( \xi \) is the tuning parameter, \( \hat{\Theta} = \{\hat{\theta}_j^0\} \) is the unpenalized least squares estimator of \( \Theta = \{\theta_j\} \) and \( \text{SE}(\hat{\theta}_j^0) \) are the standard errors of the unpenalized least squares estimator \( \theta_j^0 \) of \( \theta_j \). Similarly, we set \( k_1 = k \sum_{k=1}^m \text{SE}(\hat{\theta}_j^0) \) and \( \lambda_j \approx \lambda \cdot \text{SE}(\hat{\theta}_j^0) \), and get a proper \( \lambda \) and a proper \( k \) via following the above mentioned procedure. In the sequel discussion, we let \( k = (k_1, \xi, \lambda) \) and use \( \hat{\Theta}_\xi \) to denote the resulting least squares SCAD estimator \( \hat{\Theta} \) in order to emphasize the relationship between \( \hat{\Theta} \) and \( \xi \). Then, we select \( \xi \) by minimizing the following objective function

\[
\text{BIC}(\xi) = \log(\text{SSE}(\xi)) + \frac{\log(np)}{np} \times \text{DF}(\hat{\Theta}_\xi),
\]

where \( \text{SSE}(\xi) = \text{tr}\left[(Y - X\hat{\Theta}_\xi Z^\top)^\top(Y - X\hat{\Theta}_\xi Z^\top)\right] \), also see [3], denotes the sum of squared error and \( \text{DF}(\hat{\Theta}_\xi) \) is the generalized degrees of freedom given by

\[
\text{DF}(\hat{\Theta}_\xi) = \text{tr}(X^\top Z)(X^\top X \otimes Z^\top Z + \Sigma^0_0 + \Sigma^0_0 + \Sigma^0_0)^{-1}(X^\top Z^\top).
\]

More specifically, we select \( k_1, \xi, \lambda_1 \) to the minimizer separately among a set of grid points over bounded interval \( [0, \xi_{\text{max}}] \) where \( \sqrt{n}\xi_{\text{max}} \rightarrow 0 \) as \( n \rightarrow \infty \), where \( \sqrt{n}\lambda_{\text{max}} \rightarrow 0 \) as \( n \rightarrow \infty \), \( [0, \xi_{\text{max}}] \) and \( [0, \lambda_{\text{max}}] \) where \( \sqrt{n}\lambda_{\text{max}} \rightarrow 0 \) as \( n \rightarrow \infty \). Denote \( \hat{\xi} \) the resulting optimal tuning parameter. Practically, a plot of the BIC(\( \xi \)) against \( \xi \) can be used to determine an appropriate \( \xi_{\text{max}} = (\xi_{\text{max}}, \xi_{\text{max}}, \lambda_{\text{max}}) \) to ensure that BIC(\( \xi \)) reaches its minimum around the middle of the range of \( \xi \). Then, the grid points for \( \xi \) are taken to be evenly distributed over \( ([k_1, \lambda] : k \in [0, \xi_{\text{max}}], \lambda \in [0, \lambda_{\text{max}}], \xi \in [0, \xi_{\text{max}}], \lambda \in [0, \lambda_{\text{max}}]) \) so that they are chosen to be fine enough to avoid multiple minimizers of BIC(\( \xi \)).

To investigate the theoretical properties of the BIC selector, we let \( \hat{\theta}_j \) be an estimator of the true degree of the polynomial profile form, \( m_0 \) be estimator of the true number of variables and \( d_0 \) be estimator of the true number of nonzero elements in \( \hat{\eta} \) by the SCAD procedure with a triple of tuning parameters \( \xi, k_1 \) and \( \lambda_1 \), and set \( \hat{\Sigma}_{ij} = \hat{\Sigma}(Y - X\hat{\Theta}_\xi Z^\top)^\top(Y - X\hat{\Theta}_\xi Z^\top) \), where \( \hat{\Theta}_\xi \) is the estimate matrix of row 1 to i and column 1 to j, \( X^i \) and \( Z^j \) are the corresponding matrices. In addition, we make the following assumption

**Assumption C.** For any \( 1 \leq i \leq m, 1 \leq j \leq q \), \( \hat{\Sigma}_{ij} \rightarrow \Sigma_{ij} > 0 \) as \( n \rightarrow \infty \). In addition, if \( i < m_0, j < q_0 \), we have \( \hat{\Sigma}_{ij} > \Sigma_{m_0q_0} \).
It should be noted that Assumption C is the standard conditions for investigating parameter estimation under model misspecification. See [27] for the details.

**Theorem 5.1.** Under Assumptions A and C, we have

\[
\Pr(\hat{q}_0 = q_0, \hat{m}_0 = m_0, \hat{d}_0 = d_0) \to 1, \quad \text{as} \ n \to \infty.
\]

Theorem 5.1 demonstrates that the BIC tuning parameter selectors enable us to select the true model consistently.

6. A Monte Carlo investigation

In this section, we conduct some simulation studies to show the finite sample performance of the proposed procedures in previous sections.

The data are generated from the following growth curve model

\[
Y = X\Theta Z^\tau + \varepsilon, \quad \varepsilon \sim (0, I_1 \otimes \Sigma)
\]

where \(X = (X_1, \ldots, X_9), (X_1, X_2) = I_2 \otimes 1_{a/2}, X_3 \) comes from the \(\chi^2_1\) distribution, \(X_4, \ldots, X_9\) are all generated from the normal distribution with the standard deviation are 2, \ldots, 6, respectively. \(\Theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})\) where \(\Theta_{11} = \begin{pmatrix} 5 & -7.5 & 0 \\ 5 & 1 & 0.1 \\ 0 & 4 & 0 \end{pmatrix}\). \(\Theta_{12}\) is a 3 \(\times\) 4 zero matrix, \(\Theta_{21}\) is a 5 \(\times\) 3 zero matrix, and \(\Theta_{22}\) is a 5 \(\times\) 4 zero matrix.

\[
Z^\tau = \begin{pmatrix}
-3.5 & -2.5 & -1 & 0.5 & 1 & 2 & 3.5 & 4 & 5.5 & 6 \\
(-3.5)^2 & (-2.5)^2 & (-1)^2 & 0.5^2 & 1 & 2^2 & 3.5^2 & 4^2 & 5.5^2 & 6^2 \\
(-3.5)^3 & (-2.5)^3 & (-1)^3 & 0.5^3 & 1 & 2^3 & 3.5^3 & 4^3 & 5.5^3 & 6^3 \\
(-3.5)^4 & (-2.5)^4 & (-1)^4 & 0.5^4 & 1 & 2^4 & 3.5^4 & 4^4 & 5.5^4 & 6^4 \\
(-3.5)^5 & (-2.5)^5 & (-1)^5 & 0.5^5 & 1 & 2^5 & 3.5^5 & 4^5 & 5.5^5 & 6^5 \\
(-3.5)^6 & (-2.5)^6 & (-1)^6 & 0.5^6 & 1 & 2^6 & 3.5^6 & 4^6 & 5.5^6 & 6^6 \\
(-3.5)^7 & (-2.5)^7 & (-1)^7 & 0.5^7 & 1 & 2^7 & 3.5^7 & 4^7 & 5.5^7 & 6^7 
\end{pmatrix}
\]

and \(\Sigma = (\rho^{-1/2})_{1 \leq i, j \leq 10} + 1/2 \mathbf{1}_{10}\mathbf{1}^\top_{10} + \text{Diag}(1, 0.8, 0.6, 0.4, 0.6, 0.8, 1, 1.2, 1, 0.5)\). In addition, \(\varepsilon_i \sim i.i.d. N(0, \Sigma), \rho = 0.1 \) and 0.5, and we take \(n = 50, 100\) and 200.

In each case the number of simulated samples is 1000. The tuning parameters \(\lambda_{ij}\) and \(\lambda_{ij}\) are selected by the Bayesian information criterion developed in Section 5. Vec SCAD penalty denotes the results that SCAD penalty is directly used to the vector version of the growth curve model (1.1). We first compare the variable selection results in Tables 1 and 2 for different \(n\) and \(\rho\) and over 1000 samples. Since Vec SCAD performs better than Vec LASSO, we do not list the results of Vec LASSO in Tables.

We have the following conclusions from Tables 1 and 2:

- For each fixed \(n\) and \(\rho\), the percentage of the correctly selected models based on weighted least squares with group SCAD penalty is highest, implying that our proposed method is more efficient than the best subset BIC and Vec-SCAD.
- For each method or each fixed \(n\) or \(\rho\), the percentages of the correctly selected degree of polynomial profile form are much higher than ones of the correctly selected explanatory variables.
- All their percentages approach 100% as \(n\) increases. The fastest is the speed of the weighted least squares with group SCAD penalty while the slowest is the speed of the best subset BIC method.

In addition, the mean true positive (TP) and mean false positive (FP) numbers will be reported. Here \(TP = \sum_{i=1}^{q} \sum_{j=1}^{p} I(\hat{\theta}_{ij} \neq 0, \theta_{ij} \neq 0)\) and \(FP = \sum_{i=1}^{q} \sum_{j=1}^{p} I(\hat{\theta}_{ij} = 0, \theta_{ij} \neq 0)\). As suggested by Table 3, the weighted group SCAD gives the smallest mean FP number. While weighted group SCAD and group SCAD both maintain a comparable mean number of TPs at 7.

Together with Tables 1 and 2, our simulation results confirm that the group SCAD variable selection procedure can identify the true model consistently.

For each sample and \((\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})\), we also compute the weighted least squares estimate without the group SCAD penalty (WLSE), the least squares estimate with the SCAD penalty (LSE-sp) for the vector version of model (1.1), the least squares estimate with the group SCAD penalty (LSE-gsp) and the weighted least squares estimate with the group SCAD penalty (WLS-gsp), the estimates of the corresponding standard errors of under the related asymptotic theories and the corresponding 95% confidence intervals also under the related asymptotic theories. Over the 1000 samples in each combination, the sample means (sm) and standard deviations (std) for the parametric estimates, the means for the estimates of the standard errors (se), and the coverage probabilities (cp) for the 95% confidence intervals are considered, see Fig. 1.

From these results we can make the following observations. The WLSE-gsp and LSE-gsp (its results do not appear in Fig. 1) have much smaller standard deviations and standard errors, much closer to the nominal 95% level than the WLSE.
Table 1
The group selection results where $C$, $U$ and $O$ denote, respectively, the numbers of correct-fitting, under-fitting and over-fitting, over 1000 replications.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\rho$</th>
<th>$C_{\text{column}}$</th>
<th>$U_{\text{column}}$</th>
<th>$O_{\text{column}}$</th>
<th>$C_{\text{row}}$</th>
<th>$U_{\text{row}}$</th>
<th>$O_{\text{row}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.1</td>
<td>132</td>
<td>0</td>
<td>868</td>
<td>679</td>
<td>0</td>
<td>321</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>264</td>
<td>0</td>
<td>736</td>
<td>745</td>
<td>0</td>
<td>255</td>
</tr>
<tr>
<td>200</td>
<td>0.1</td>
<td>434</td>
<td>0</td>
<td>566</td>
<td>793</td>
<td>0</td>
<td>207</td>
</tr>
<tr>
<td>50</td>
<td>0.1</td>
<td>992</td>
<td>8</td>
<td>0</td>
<td>873</td>
<td>0</td>
<td>127</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>953</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>200</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>987</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>50</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>972</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>986</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>200</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2
The group element selection results over 1000 replications.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\rho$</th>
<th>$C_{e}$</th>
<th>$U_{e}$</th>
<th>$O_{e}$</th>
<th>$C_{e}$</th>
<th>$U_{e}$</th>
<th>$O_{e}$</th>
<th>$C_{e}$</th>
<th>$U_{e}$</th>
<th>$O_{e}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.1</td>
<td>955</td>
<td>36</td>
<td>9</td>
<td>986</td>
<td>0</td>
<td>14</td>
<td>995</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>996</td>
<td>4</td>
<td>0</td>
<td>999</td>
<td>0</td>
<td>1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>0.1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>974</td>
<td>18</td>
<td>8</td>
<td>982</td>
<td>0</td>
<td>18</td>
<td>995</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>998</td>
<td>1</td>
<td>1</td>
<td>999</td>
<td>0</td>
<td>1</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3
Mean true positive and mean false positive numbers.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\rho$</th>
<th>Vec SCAD</th>
<th>Group SCAD</th>
<th>Weighted group SCAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$TP$</td>
<td>$FP$</td>
<td>$TP$</td>
</tr>
<tr>
<td>50</td>
<td>0.1</td>
<td>6.96(0.19)</td>
<td>0.18(0.45)</td>
<td>7(0)</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>6.99(0.06)</td>
<td>0.05(0.23)</td>
<td>7(0)</td>
</tr>
<tr>
<td>200</td>
<td>0.1</td>
<td>7(0)</td>
<td>0.01(0.11)</td>
<td>7(0)</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>6.98(0.13)</td>
<td>0.27(0.53)</td>
<td>7(0)</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>6.99(0.03)</td>
<td>0.09(0.30)</td>
<td>7(0)</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>7(0)</td>
<td>0.03(0.18)</td>
<td>7(0)</td>
</tr>
</tbody>
</table>

as well as LSE-sp. The improvement is noticeable. This shows the importance of the group SCAD selection. In addition, the WLSE-gsp outperforms the LSE-gsp, especially when the sample is moderate or large (for example $n \geq 100$), or the within cluster correlation is strong (for example $\rho = 1$). The standard deviations and standard errors of the LSE-gsp and WLSE-gsp decrease as $n$ increases.

Our proposed procedure is defined for $p < n$. However, the procedure may be not effective as $p$ approaches to $n$ since some matrices in the procedure may be close to singular ones.

7. An application example

In this section, we will apply the proposed methods in previous sections to a practical data set. The data used here was provided in [5]. The data was drawn from the National Longitudinal Survey of Youth (NLSY), conducted by the U.S. Department of Labor. Assessments of child and mother were administered every other year from 1986 to 1992. There are 221 pairs of children and mothers, who completed interviews at four time points.
In this example, reading recognition was repeatedly measured over the four time points. The child’s reading recognition skill was computed by summing the total number of correct items by children out of 84 items of the Peabody Individual Achievement Reading Recognition subtest.
Besides, cognitive stimulation for children at home, emotional support for children at home, and gender were three other variables. They were measured only once at the initial time point. Cognitive stimulation was obtained as a sum of the mother's responses to 14 items in the cognitive stimulation subscale of the Home Observation for Measurement of the Environment-Short Form (HOME-SF). Emotional support was obtained by summing the mother’s responses to 13 dichotomous items from the HOME-SF. Gender refers to children’s gender. Female child was coded as −1 and male child as 1.

In the analysis, $Y_{ij}$ is the observation value of the reading recognition response variable at the $j$th time point for the $i$th observation. Here $p = 4$. There are $n=221$ observations. What is more, cognitive stimulation, emotional support and gender were considered as explanatory variables $x_1, x_2, x_3$. So the design matrix $X$ is designed as $(1 \ x_1 \ x_2 \ x_3)$ where $1$ represents the intercept item, $x_1$, cognitive stimulation vector, $x_2$, emotional support vector and $x_3$, gender vector.

At the beginning of analyzing the data set, we have no information about the order of growth curves. Since the response variable was merely measured over four time points, we assume quadratic form in time $t$ for both growth curves as the saturated model. Here, the profile matrix and the regression coefficients matrix are, respectively,

$$Z^t = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 9 & 1 & 1 & 9 \end{pmatrix} \quad \text{and} \quad \Theta^{(1)} = \begin{pmatrix} \theta_{11}^{(1)} & \theta_{12}^{(1)} & \theta_{13}^{(1)} \\ \theta_{21}^{(1)} & \theta_{22}^{(1)} & \theta_{23}^{(1)} \\ \theta_{31}^{(1)} & \theta_{32}^{(1)} & \theta_{33}^{(1)} \end{pmatrix}. $$

Applying the method in previous section, the suitable model determined by our procedure is as follows $Y = X_0 \Theta Z^t_0 + \varepsilon$, where $X_0 = (1 \ x_1 \ x_2)$, $\Theta_0 = (\theta_0)_{ij=1}^3$, $Z_0 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \end{pmatrix}$ and $\varepsilon$ is the error matrix. The estimated parameter matrix and covariance matrix are, respectively,

$$\hat{\Theta} = \begin{pmatrix} 3.5599 & 0.3618 \\ 0.0474 & 0 \\ 0.0377 & 0.0193 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{pmatrix} 0.8078 & 0.4850 & 0.4798 & 0.5123 \\ 1.0208 & 0.8755 & 0.8091 & 1.1751 \\ 1.8091 & 0.9867 & 1.3793 \end{pmatrix}. $$

In the equation form, the best fitted function can be expressed as

$$\text{Mean of Reading Recognition} = 3.5599 + 0.3618t + 0.0474 \times \text{Cognitive Stimulation},$$

$$+ (0.0377 + 0.0193t) \times \text{Emotional Support},.$$

From the above results, we can make the following conclusions about the real data set.

Reading recognition skill is a linear function of time and becomes greater and greater as children grow up. Greater cognitive stimulation for children at home can lead to greater reading recognition skill at a positive constant rate. Greater emotional support for children at home can lead to greater reading recognition skill at an increasing rate as children grow up. Gender has no significant effect on reading recognition skill. There also are positive correlation coefficients among the four responses of reading recognition skills at the time points.

8. Concluding remarks

A key point in using the growth curve model for data analysis is determining the order of polynomial profile form and the proper explanatory variables. In this paper, we have proposed three-level variable selection methods based on weighted least squares with group SCAD penalty to adaptively determine the set of significant explanatory variables, the degree of polynomial profile form, shrink and estimate the regression coefficient. With appropriate selection of the tuning parameters, we have established the oracle property of the procedures and the consistency of the estimation. The finite sample performances show that our proposed procedure is very supportive and promising. Our results indicate that the proposed procedure outperforms the best subset selection and the method that SCAD is directly applied to the vector version of model (1.1) in the following four points: (i) higher correct rate of true explanatory variables, (ii) higher correct rate of the degree of polynomial profile form, (iii) higher correct rate of true zero regression coefficients and (iv) smaller bias, much smaller standard deviations and standard error, and more accurate coverage probabilities for nonzero coefficients.

When there are too many explanatory variables, the smoothly clipped absolute deviation method is a suitable method to choose proper variables. In the traditional growth curve models, it is assumed that the number of time points where the measurements are collected is fixed and small compared with the number of individuals. However, repeated measurement data with moderately or extremely long time lengths are the obvious characteristics of many contemporary scientific and economic problems. It is essential to develop novel mathematical tools to handle asymptotic sequences under which two indexes tend to infinity. Especially, when the number of observed time points is large, it is more realistic to regard the degree of polynomial profile form growing with the sample size. Therefore, how to apply the SCAD penalty method to determine a dynamic degree polynomial profile form which is dependent on the time length is an interesting problem. Further work in this direction is ongoing.
Acknowledgments

The authors are deeply grateful to two anonymous referees and the Associate Editor for their valuable comments which led to the improved version of the article.

Hu’s research was supported by National Natural Science Foundation of China (NSFC) Grants 10971126 and by SUFE through Project 211 Phase III and Shanghai Leading Academic Discipline Project B803. Xin’s research was supported by SUFE’s Postgraduate Innovation Fund with Project Number: CXJJ-2011-440. You’s research was supported by National Natural Science Foundation of China (NSFC) Grants 11071154.

Appendix

We shall use the following notation in consequent proofs. Let
\[ \beta = (\beta_{10}^*, \beta_{20}^*, \beta_2^*) \text{ with } \beta_{10}^* = (\beta_{10}^*, \text{vec}(\Theta_{12})^*, \text{vec}(\Theta_{21})^*, \text{vec}(\Theta_{22})^*)^*, \]
where \( \Theta = (\Theta_{ij})_{i,j=1}^2 \), \( \Theta_{11} \) is an \( m_0 \times q_0 \) matrix such that \( \text{vec}(\Theta_{11}) = L(\beta_{10}^*, \beta_{20}^*)^* \) with \( \beta_{10}^* \) a \( d_0 \)-dimensional vector.

In order to prove the main results, we first introduce a lemma.

Lemma A.1. Suppose that, for any pair \( i, j, k_1 \to 0, \kappa_j \to 0, \lambda_{ij} \to 0 \) and \( \sqrt{n}\kappa_1 \to \infty, \sqrt{n}\kappa_j \to \infty, \sqrt{n}\lambda_{ij} \to \infty \) as \( n \to \infty \), then with probability tending to 1, for any given \( \beta_{10}^* \) satisfying \( \| \beta_{10}^* - \beta_{10} \| = O_p(n^{-1/2}) \) and any constant \( c_1, c_2, c_3, c_4 \).

\[ L(\beta_{10}^*, 0) = \min_{\beta_{10} \in \mathcal{H}_0} L(\beta_{10}^*, \beta_{10}^*) \]
where \( \mathcal{H}_0 = \{ \beta_{10}^* : \| \beta_{10}^* \| \leq c_1 n^{-1/2}, \| \Theta_{12}^* \| \leq c_2 n^{-1/2}, \| \Theta_{21}^* \| \leq c_3 n^{-1/2}, \| \Theta_{22}^* \| \leq c_4 n^{-1/2} \} \).

Proof. To prove that the minimizer achieves at \( \beta_{10}^* = 0 \), it suffices to show that as \( n \to \infty \), with probability tending to 1, for the constant \( c = \max\{c_1, c_2, c_3, c_4\} \), if \( \beta \) or \( \Theta^* \) satisfies \( \| \beta_{10}^* - \beta_{10} \| = O_p(n^{-1/2}) \) and \( \beta_{10}^* = \Theta_{ij}^* \in (-cn^{-1/2}, cn^{-1/2}) \) for any element \( \beta_{ij}^* \) of \( \beta_{10}^* \), \( \partial L(\Theta^*)/\partial \beta_{ij}^* \) and \( \beta_{ij}^* \) have the same signs, namely, \( \partial L(\Theta^*)/\partial \beta_{ij}^* \) and \( \beta_{ij}^* \) have the same signs.

Let
\[ \delta(\theta) = \frac{1}{2} \text{tr}((Y - X\Theta Z^*)^T(Y - X\Theta Z^*)). \]

For any nonzero element \( \Theta_{ij}^* \), i.e. \( \beta_{ij}^* \), from \( \beta_{10}^* \),
\[
\frac{\partial L(\Theta^*)}{\partial \theta_{ij}^*} = \frac{\partial \delta(\Theta^*)}{\partial \theta_{ij}^*} + npe'_{\Theta_{ij}^*}(\| \theta_{ij}^* \|) \frac{\theta_{ij}^*}{\| \theta_{ij}^* \|} + npe'_{\Theta_{ij}^*}(\| \theta_{ij}^* \|) \frac{\theta_{ij}^*}{\| \theta_{ij}^* \|} + npe'_{\Theta_{ij}^*}(\| \theta_{ij}^* \|) \frac{\theta_{ij}^*}{\| \theta_{ij}^* \|}.
\]

It is easy to see
\[
\frac{\partial \delta(\Theta^*)}{\partial \theta_{ij}^*} = \frac{-n}{n} \sum_{s=1}^{n} x_{si} (y_{si} - x_{si}^T \Theta_0 z_i) + \frac{n}{n} \sum_{s=1}^{n} x_{si} (\Theta_0 - \Theta) z_i = J_1 + J_2, \quad \text{say},
\]
where \( x_i = (x_{i1}, \ldots, x_{in})^T \) and \( z_i = (z_{i1}, \ldots, z_{ip})^T \). It is easy to verify that \( J_1/n = -\sum_{s=1}^{n} x_{si} \varepsilon s/n \text{vec}(\Theta_{0}) \text{vec}(\Theta_{0}) / \text{vec}(\Theta_{0}) / \text{vec}(\Theta_{0}) = O_p(n^{-1/2}) \) where \( \varepsilon s \) is the st element of the error matrix \( \varepsilon \). In addition, \( \| \Theta^* - \Theta \| = O_p(n^{-1/2}) \) and \( X'X/n = O(1) \) by assumption, we can show that \( J_2 = O_p(n^{-1/2}) \). As a result, \( n^{-1} \partial \delta(\Theta^*)/\partial \theta_{ij}^* \) is of the order \( O_p(n^{-1/2}) \). Thus,
\[
\frac{\partial L(\Theta^*)}{\partial \theta_{ij}^*} = n \left\{ \left| \varepsilon_{ij} \right| / \left| \theta_{ij}^* \right| + \left| \varepsilon_{ij} \right| / \left| \theta_{ij}^* \right| + \left| \varepsilon_{ij} \right| / \left| \theta_{ij}^* \right| \right\} \frac{\theta_{ij}^*}{\| \theta_{ij}^* \|}.
\]

Since \( \kappa_1 \to 0, \kappa_j \to 0, \lambda_{ij} \to 0 \) and \( \sqrt{n}\kappa_1 \to \infty, \sqrt{n}\kappa_j \to \infty, \sqrt{n}\lambda_{ij} \to \infty \) as \( n \to \infty \) there exists a large number \( N \) such that, as \( n > N \),
\[
\| \theta_{ij}^* \| \leq n^{-1/2} n c < \kappa_1 , \quad \frac{\| \theta_{ij}^* \|}{\| \theta_{ij}^* \|} \leq n^{-1/2} n c < \kappa_1 , \quad \frac{\| \theta_{ij}^* \|}{\| \theta_{ij}^* \|} \leq n^{-1/2} n c < \kappa_1 ,
\]

Hence, \( \partial L(\Theta^*)/\partial \theta_{ij}^* \) can be further expressed as
\[
\frac{\partial L(\Theta^*)}{\partial \theta_{ij}^*} = n \left\{ \left( \kappa_1 / \left| \theta_{ij}^* \right| + \kappa_1 / \left| \theta_{ij}^* \right| + \lambda_{ij} \right) \text{sgn}(\theta_{ij}^*) + O_p(n^{-1/2}) \right\},
\]

implying that the sign of the derivative is completely determined by that of \( \theta_{ij}^* \). This completes the proof. \( \square \)
Proof of Theorem 2.1. Denote $\alpha_n = (n^{-1/2} + a_n' + a_n^2 + a_n^{id})$. It is sufficient to show that for any given $\tau > 0$, there exists a large constant $h$ such that

$$P \left\{ \inf_{U \in \mathcal{H}} \mathcal{L}(\Theta + \alpha_n U) \geq \mathcal{L}(\Theta) \right\} \geq 1 - \tau$$

where $U$ is a $m \times q$ constant matrix. This implies, with probability at least $1 - \tau$, that there exists a local minimizer in the ball $\{ \Theta + \alpha_n U : \|U\| \leq h \}$. Define

$$D_n(U) = \mathcal{L}(\Theta + \alpha_n U) - \mathcal{L}(\Theta).$$

Note that $pe_{\alpha}(0) = 0, pe_{\alpha}(0) = 0$ and $pe_{\alpha}(\|\theta\|), pe_{\alpha}(\|\theta\|)$ and $pe_{\alpha}(\|\theta\|)$ is nonnegative. Therefore, it holds that

$$\frac{1}{n}D_n(U) = \frac{1}{2n} \text{tr} \{ Y - X(\Theta + \alpha_n U)Z^T \} + \sum_{i=1}^{m} pe_{\alpha}(\|\theta_i + \alpha_n u_i\|)$$

$$+ \sum_{j=1}^{q} pe_{\alpha}(\|\theta_j + \alpha_n u_j\|) - \frac{1}{n} \mathcal{L}(\Theta).$$

It is easy to see that

$$\text{tr} \{ Y - X(\Theta + \alpha_n U)Z^T \} = \alpha_n^2 \text{tr}(Z U^T X^T U^T Z) - 2 \alpha_n \text{tr}(Z U^T X^T Z) = 2n \alpha_n h, \text{ say.}$$

Due to the fact that $X^T X/n = O(1)$, we know that $f_1$ is of the order $h^2 \alpha_n^2$. Note that $n^{-1/2} \alpha_n = O_p(\alpha_n^2)$. Following the central limit theorem $X^T (Y - X \Theta Z^T)/n = X \varepsilon/n = O_p(n^{-1/2})$, it holds that $f_2 = O_p(n^{-1} \alpha_n \|U\|)$. By choosing a sufficiently large $h$, $f_1$ will dominate the second term uniformly in $\|U\| = h$. Furthermore, it follows from the Taylor expansion and Cauchy–Schwarz inequality that

$$\sum_{j=1}^{q} \{ pe_{\alpha}(\|\theta_j + \alpha_n u_j\|) - pe_{\alpha}(\|\theta_j\|) \}$$

is bounded by

$$\alpha_n a_n^2 \|u\| + \alpha_n^2 b_n^2 \|u\| \leq h \alpha_n^2 (1 + b_n^2).$$

Similarly,

$$\sum_{i=1}^{m} \{ pe_{\alpha}(\|\theta_i + \alpha_n u_i\|) - pe_{\alpha}(\|\theta_i\|) \}$$

is bounded by

$$\alpha_n a_n^2 u_j + \alpha_n^2 b_n^2 u_j \leq h \alpha_n^2 (1 + b_n^2),$$

and

$$\sum_{i=1}^{m} \{ pe_{\alpha}(\|\theta_i + \alpha_n u_i\|) - pe_{\alpha}(\|\theta_i\|) \}$$

is bounded by

$$\alpha_n a_n^2 |u| + \alpha_n^2 b_n^2 |u| \leq h \alpha_n^2 (1 + b_n^2).$$

As $b_n^2 \rightarrow 0$, $b_n^2 \rightarrow 0$ and $b_n^{id} \rightarrow 0$, by taking $h$ sufficiently large, $\alpha_n^2 (1 + b_n^2), \alpha_n^2 (1 + b_n^2)$ and $h \alpha_n^2 (1 + b_n^2)$ are dominated by $f_1$, which completes the proof of the desired result. \( \Box \)

Proof of Theorem 2.2. Part (i) directly follows by Lemma A.1. Now we prove Part (ii). By Lemma A.1 and Theorem 2.1, there exists a root-$n$ consistent local minimizer $\hat{\beta} = (\hat{\beta}_{10}, 0^T)^T$ of $\mathcal{L}(\beta)$ or $\mathcal{L}(\Theta)$ satisfying

$$\frac{\partial \mathcal{L}(\beta)}{\partial \hat{\beta}_l} = 0 \quad \text{for } l = 1, \ldots, d_0.$$

Fix $\Theta_{12} = 0, \Theta_{21} = 0$ and $\Theta_{22} = 0$. Let $T = (X_1 \otimes Z_1)\mathcal{I}_G$. Then the partial derivative of $\mathcal{L}(\beta)$ with respect to $\beta_l$ is calculated as follows,

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_l} = -t_l^T (\text{vec}(Y) - T \beta) + npe_{\alpha}(\|\theta_i\|) \frac{\partial \hat{\beta}_l}{\|\theta_i\|} + npe_{\alpha}(\|\theta_j\|) \frac{\partial \hat{\beta}_j}{\|\theta_j\|} + npe_{\alpha}(\|\theta_l\|).$$
where $\beta_l, l = 1, \ldots, d_0$, are nonzero elements $\theta_{ij}$ in $\Theta_{11}$ and $T = (t_1, \ldots, t_{d_0})^T$. Inserting $\hat{\theta} = (\hat{\beta}_{10}, \mathbf{0}^T)^T$ into above equation and letting both sides of the equation be divided by $n$, we obtain the following equation

$$-\frac{1}{n} T_i^T (\text{vec}(\mathbf{Y}) - T_i \hat{\beta}_{10}) + p e'_s(\|\hat{\theta}_i\|) \frac{\hat{\theta}_i}{\|\hat{\theta}_i\|} + p e'_j(\|\hat{\theta}_j\|) \frac{\hat{\theta}_j}{\|\hat{\theta}_j\|} + p e'_{\kappa_i}(\|\hat{\theta}_i\|) = 0.$$  

Applying the Taylor expansion, we get that

$$p e'_s(\|\hat{\theta}_i\|) = p e'_s(\|\theta_i\|) + \left\{ p e''_s(\|\theta_i\|) + o_p(1) \right\} \left( \|\hat{\theta}_i\| - \|\theta_i\| \right),$$

$$p e'_j(\|\hat{\theta}_j\|) = p e'_j(\|\theta_j\|) + \left\{ p e''_j(\|\theta_j\|) + o_p(1) \right\} \left( \|\hat{\theta}_j\| - \|\theta_j\| \right),$$

$$pe'_{\kappa_i}(\|\hat{\theta}_i\|) = pe'_{\kappa_i}(\|\hat{\theta}_i\|) + \left\{ pe''_{\kappa_i}(\|\hat{\theta}_i\|) + o_p(1) \right\} \left( \|\hat{\theta}_i\| - \theta_i \right).$$

The conditions $b_n^s = o_p(1), b_n^j = o_p(1)$ and $b_n^{\kappa} = o_p(1)$ imply that $p e'_s(\|\theta_i\|) = o_p(1), p e'_j(\|\theta_j\|) = o_p(1)$ and $pe'_{\kappa_i}(\|\theta_i\|) = o_p(1)$. Since $\kappa_i \to 0$, $\xi_i \to 0$ and $\lambda_i \to 0$ as $n \to \infty$, $p e'_s(\|\hat{\theta}_i\|) = 0$, $p e'_j(\|\hat{\theta}_j\|) = 0$ and $pe'_{\kappa_i}(\|\hat{\theta}_i\|) = 0$. These imply that $\|\theta_i\| \to 0$, $\|\theta_j\| \to 0$ and $\|\hat{\theta}_i\| \to 0$. Combining all equations $l = 1, \ldots, d_0$, we have

$$-\frac{1}{n} T_i^T (\text{vec}(\mathbf{Y}) - T_i \hat{\beta}_{10}) + o_p(\|\hat{\beta}_{10} - \beta_{10}\|) = 0,$$

where $T_{ig} = (X_i \otimes Z_i)z_{10}^T$ with $z_{10} = (z_{10}, z_{20})$. Namely,

$$-\frac{1}{n} T_i^T (\text{vec}(\mathbf{Y}) - T_i \hat{\beta}_{10}) + o_p(\|\hat{\beta}_{10} - \beta_{10}\|) = 0.$$

Therefore,

$$\frac{\sqrt{n}(\hat{\beta}_{10} - \beta_{10})}{\hat{\Sigma}} = \left( \frac{T_1^T T_1}{n} \right)^{-1} L_{10}^T (X_i \otimes Z_i^T) \text{vec}(\sqrt{n}(X_i \otimes Z_i^T) - o_p(1)).$$

According to [13], it holds that $\sqrt{n}(X_i \otimes Z_i^T) \to N(\mathbf{0}, R_{11}^{-1} \otimes \Sigma)$. Therefore, $\sqrt{n}(\hat{\beta}_{10} - \beta_{10})$ converges in distribution to $N(\mathbf{0}, \Omega)$, where

$$\Omega = (L_{10}^T (R_{11} \otimes Z_i^T) L_{10}^{-1})^{-1} L_{10}^T (R_{11} \otimes Z_i^T \Sigma Z_i) L_{10}^{-1} (L_{10}^T (R_{11} \otimes Z_i^T) L_{10}^{-1})^{-1}.$$

Let

$$\Omega_n = n (L_{10}^T (X_i \otimes Z_i^T) L_{10}^{-1})^{-1} L_{10}^T (X_i \otimes Z_i^T \Sigma Z_i) L_{10}^{-1} (L_{10}^T (X_i \otimes Z_i^T) L_{10}^{-1})^{-1}.$$

Then, by Slutsky's theorem, $\sqrt{n} \Omega_n^{-\frac{1}{2}} (\hat{\beta}_{10} - \beta_{10})$ converges in distribution to $N(\mathbf{0}, I_{m_0p})$. This completes the proof of Theorem 2.2. □

**Proof of Theorem 3.1.** By the definition of $\hat{\Sigma}$, it holds that

$$\hat{\Sigma} = \frac{1}{n} (\mathbf{Y} - X \hat{\theta}_{gps}Z^T)(\mathbf{Y} - X \hat{\theta}_{gps}Z^T)^T$$

$$= \frac{1}{n} (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T + \varepsilon)^T (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T + \varepsilon)$$

$$= \frac{1}{n} (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T)^T (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T) + \frac{1}{n} \varepsilon^T \varepsilon + \frac{1}{n} (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T)^T \varepsilon + \frac{1}{n} \varepsilon^T (\mathbf{Y} \theta Z^T - X \hat{\theta}_{gps}Z^T)$$

$$= J_1 + J_2 + J_3 + J_4, \text{ say.}$$

By Theorem 2.2, it holds that $J_1 = O_p(n^{-1})$ with probability 1 tends to $\mathbf{0}$ and

$$J_3 = (\hat{\theta} Z^T - \hat{\theta}_{gps}Z^T)^T \frac{1}{n} X^T \varepsilon = O_p(n^{-\frac{1}{2}}) \cdot O_p(n^{-\frac{1}{2}}) = O_p(n^{-1})$$

with probability 1 tends to $\mathbf{0}$. Similarly, $J_3 = O_p(n^{-1})$ with probability 1 tends to $\mathbf{0}$. In addition,

$$J_2 = \frac{1}{n} \varepsilon^T \varepsilon - \Sigma = \frac{1}{n} \sum_{j=1}^n (\varepsilon_i^T \varepsilon_i - \Sigma_i)$$

By the central limit theorem, the proof is complete. □
Proof of Theorems 4.1 and 4.2. The proofs of Theorems 4.1 and 4.2 are same as those of Theorems 2.1 and 2.2. We here omit the details. □

Proof of Theorem 5.1. We first prove that \( Pr(\hat{q}_0 = q_0) \rightarrow 1 \), as \( n \rightarrow \infty \). The proofs of others are similar to the first. Let \( \zeta_n = \log(np) \), \( \xi_j = \zeta_n \sum_{i=1}^{m} \text{SE}(\theta_i) \), \( \kappa_r = 0, \lambda_j = 0 \) and

\[
\text{BIC}_j = \log \left\{ \text{tr}(\text{MSE}_j) \right\} + \frac{\log(np)}{(np)} \times \text{DF}(\hat{\Theta}_j)
\]

with \( \text{MSE}_j = (Y - X\hat{\Theta}_j'Z^*)' (Y - X\hat{\Theta}_j'Z^*)/n \) and \( \hat{\Theta}_j = (\hat{\theta}_j, \ldots, \hat{\theta}_j, 0_{m_{x_k-q_j}}) \). It is easy to see that \( \text{SE}(\hat{\theta}_j) = O(n^{-1/2}) \). Thus, \( \xi_j = O_p(n^{-1/2} \log n) \). Then, by the same techniques as those in [27], we have

\[
Pr\{\text{BIC}(\zeta_n) = \text{BIC}_q_0\} = 1.
\]

(A.1)

Therefore, in order to prove Theorem 5.1, it suffices to show that

\[
Pr \left\{ \inf_{\zeta \in \Omega_- \cup \Omega_+} \text{BIC}(\zeta) > \text{BIC}(\zeta_n) \right\} \rightarrow 1 \text{ as } n \rightarrow \infty
\]

(A.2)

where \( j_0 = \max\{j : \|\hat{\theta}_{q_j}\| \neq 0\} \), \( \Omega_- = \{ \zeta : j_0 < q_0 \} \) and \( \Omega_+ = \{ \zeta : j_0 \neq 0 > q_0 \} \) represent the underfitted and overfitted models, respectively.

To demonstrate (A.2) we consider two separate cases given below.

Case 1: Underfitted model (i.e. the degree of polynomial profile form is less than \( q_0 \)). For any \( \zeta \in \Omega_- \), (A.1) together with Assumption C, implies that, with probability tending to one,

\[
\text{BIC}(\zeta) - \text{BIC}(\zeta_n) = \log(\text{tr}(\text{MSE}(\zeta))) + \frac{\log(np)}{np} \times \text{DF}(\hat{\Theta}_\zeta) - \text{BIC}_q_0 \\
\geq \log(\text{tr}(\text{MSE}(\zeta))) - \text{BIC}_q_0 \geq \log(\text{tr}(\hat{\Sigma}_j)) - \text{BIC}_q_0 \\
\geq \min_{j < q_0} \log(\text{tr}(\hat{\Sigma}_j)) - \log(\text{tr}(\hat{\Sigma}_q_0)) - q_0 \frac{\log(np)}{np} \\
\rightarrow \min_{j < q_0} \log(\text{tr}(\Sigma_j)/\text{tr}(\Sigma_{q_0})) \geq 0.
\]

Case 2: Overfitted model (i.e. the degree of polynomial profile form is greater than \( q_0 \)). For any \( \zeta \in \Omega_+ \), it follows by (A.1) that, with probability tending to one,

\[
np\{\text{BIC}(\zeta) - \text{BIC}(\zeta_n)\} = np \left\{ \text{BIC}(\zeta) - \text{BIC}_q_0 \right\} \\
= np \log \left\{ \text{tr}(\text{MSE}(\zeta))/\text{tr}(\hat{\Sigma}_{q_0}) \right\} + \log(np) \left( \text{DF}(\hat{\Theta}_j) - q_0 \right) \\
\geq np \log \left\{ \text{tr}(\hat{\Sigma}_j)/\text{tr}(\hat{\Sigma}_{q_0}) \right\} + \log(np) \left( \text{DF}(\hat{\Theta}_j) - q_0 \right) \\
= np \left\{ \text{tr}(\hat{\Sigma}_j) - \text{tr}(\hat{\Sigma}_{q_0}) \right\} \left( 1 + o_p(1) \right) + \log(np) \left( \text{DF}(\hat{\Theta}_j) - q_0 \right).
\]

For simplicity, we denote \( \hat{\Sigma}_j = \hat{\Sigma}_{q_0} \). We will show that \( n(\hat{\Sigma}_j - \hat{\Sigma}_{q_0}) = o_p(1) \) and \( \hat{\Sigma}_{q_0} = O_p(1) \).

Let \( \hat{\Theta}_{j} = (\hat{\theta}_{q_j}, \hat{\theta}_{j-1} - \theta_{q_0}) \) and \( \hat{\Sigma} = (Z_{q_j}, Z_{j-1} - \Theta_{q_0}) \). By the definition of \( \hat{\Sigma}_j \) and \( \hat{\Sigma}_{q_0} \) it holds that

\[
n(\hat{\Sigma}_j - \hat{\Sigma}_{q_0}) = (Y - X\hat{\Theta}_j'Z^*)' (Y - X\hat{\Theta}_j'Z^*) - (Y - X\hat{\Theta}_{q_0}Z^{q_0}') (Y - X\hat{\Theta}_{q_0}Z^{q_0}') = J_1 + J_2 + J_3 + J_4 + J_4,
\]

where

\[
J_1 = A' A, \quad A = X\Theta_{q_0}Z^{q_0} - X\hat{\Theta}_{q_0}Z^{q_0} - X\hat{\Theta}_{j-1}Z^{j-1} - X\hat{\Theta}_jZ^{j}, \quad J_2 = A' \varepsilon,
\]
\[
J_3 = (X\Theta_{q_0}Z^{q_0} - X\hat{\Theta}_{q_0}Z^{q_0}) (Y - X\hat{\Theta}_{q_0}Z^{q_0}'), \quad J_4 = (X\Theta_{q_0}Z^{q_0} - X\hat{\Theta}_{q_0}Z^{q_0} )' \varepsilon.
\]

By Theorem 2.2 and the root-\( n \) consistency of the least squares estimator \( \hat{\Theta}_{q_0} \), we have \( J_1 = O_p(1) \) and \( J_4 = O_p(1) \). On the other hand, since \( X' \varepsilon / \sqrt{n} = O_p(1) \), it holds that \( J_2 = O_p(1) \) for \( d = 2, 3, 5 \) and 6. According to the proof of Theorem 3.1, it
is easy to see that $\hat{\Sigma}_q = O_p(1)$. Therefore, $np(\text{tr}(\hat{\Sigma}_q^2) - \text{tr}(\hat{\Sigma}_q^2)) / \text{tr}(\hat{\Sigma}_q^2) = O_p(1)$. On the other hand, for any $\xi \in \Omega_+$, it holds that $DF(\hat{\epsilon}^5) - q_0 \geq 1$. Therefore, $\log(np) \times (DF(\hat{\epsilon}^5) - q_0)$ diverges to $+\infty$. As a result,

$$
\Pr\left\{ \inf_{\xi \in \Omega_+} n \{ BIC(\xi) - BIC(\xi_0) \} > 0 \right\} = \Pr\left\{ \inf_{\xi \in \Omega_+} \{ BIC(\xi) - BIC(\xi_0) \} > 0 \right\} \rightarrow 1
$$

as $n \rightarrow \infty$. By the results for Case 1 and 2, we complete the proof of $Pr(\hat{q}_0 = q_0) \rightarrow 1$, as $n \rightarrow \infty$.

Similarly, we can complete the proof of $Pr(\hat{m}_0 = m_0) \rightarrow 1$, as $n \rightarrow \infty$ and $Pr(\hat{d}_0 = d_0) \rightarrow 1$, as $n \rightarrow \infty$.

Next, let $k_i = \xi_n \sum_{p=1}^{m} \text{SE}(\hat{\theta}_i), \lambda_{ij} = \xi_n \text{SE}(\hat{\theta}_j)$, denote this BIC($\xi_n, \xi_n, \xi_n$) as BIC($\xi_n$) and

$$
\mathrm{BIC}_{ijk} = \log \{ \text{tr}(\text{MSE}_{ijk}) \} + \log(np) \times (DF(\hat{\epsilon}^5_{ik}))
$$

with MSE$_{ijk} = (Y - X\hat{\epsilon}^5_{ik} Z)^\top (Y - X\hat{\epsilon}^5_{ik} Z) / n$ and $\hat{\epsilon}^5_{ik} = \left( \begin{array}{c} \hat{\theta}_{i1} \\ \theta_{i(m-1)} \\ \theta_{i(m-1) \times (q-j)} \end{array} \right)$ where vec($\hat{\theta}_{i1}$) has $k$ nonzero elements.

Thus, for any $\xi = (\kappa, \xi, \lambda)$, with probability tending to 1.

$$
\mathrm{BIC}(\xi) - \mathrm{BIC}(\xi_0) = \mathrm{BIC}(\xi) - \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i}
$$

From the above proof, we know that with probability tending to 1,

$$
\mathrm{BIC}(\xi) - \mathrm{BIC}_{m_{ij}^c,k_i} > 0, \quad \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i} > 0, \quad \mathrm{BIC}_{m_{ij}^c,k_i} - \mathrm{BIC}_{m_{ij}^c,k_i} > 0.
$$

By the results, we complete the proof of the desired result. \qed

References


