Statistica Sinica Preprint No: SS-2019-0421

Title	Consistent Fixed-Effects Selection in Ultra-high dimensional Linear Mixed Models with Error-Covariate Endogeneity
Manuscript ID	SS-2019-0421
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0421
Complete List of Authors	Abhik Ghosh and Magne Thoresen
Corresponding Author	Abhik Ghosh
E-mail	abhianik@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

Consistent Fixed-Effects Selection in Ultrahigh-Dimensional Linear Mixed Models with Error-Covariate Endogeneity

Abhik Ghosh[1] and Magne Thoresen[2]

[1] *Indian Statistical Institute, India*

[2] *Department of Biostatsitics, University of Oslo, Norway*

Abstract: Applied sciences, including longitudinal and clustered studies in biomedicine, require analyses of ultrahigh-dimensional linear mixed-effects models, where we need to select important fixed-effect variables from a large pool of available candidates. However, prior studies assume that all available covariates and random-effect components are independent of the model error, which is often violated (endogeneity) in practice. In this study, we first investigate this important issue in ultrahigh-dimensional linear mixed-effects models, focusing particularly on selecting the fixed effects. We study the effects of different types of endogeneity on existing regularization methods, and prove their inconsistencies. Then, we propose a new profiled focused generalized method-of-moments (PFGMM) approach to consistently select fixed effects under "error-covariate" endogeneity, that is, in the presence of a correlation between the model error and the covariates. The proposed method is proved to be oracle consistent with probability tending to one, and works well under most other types of endogeneity too. Additionally, we propose and illustrate several consistent parameter estimators, including those of the variance components, along with variable selection using the PFGMM approach. Empirical simulations and an interesting real-data example further support the claimed utility of the proposed method.

Key words and phrases: Ultrahigh-dimensional Mixed Effects Models; Profiled Focused Generalized Method of Moments; Oracle variable selection; Endogeneity.

1. Introduction

Linear mixed-effects models (LMMs) are widely used to analyze clustered data in econometrics, biomedicine, and other applied sciences. These models include additional random-effect components, along with the usual fixed-effects regression modeling, to account for variability among clusters. In biomedical applications, typical examples are longitudinal studies with repeated measurements within individuals, and multi-center studies with patients clustered within centers. Owing to recent technological advances, such studies often have access to sets of extremely high-dimensional explanatory variables, typically the so-called omics data. Hence, the potential fixed-effects variables are often in the order of millions, even in studies with relatively few patients. Thus, we have to select the important fixed-effects variables from a large pool of available variables under an ultrahigh-dimensional setup. Note that, in most such studies, there are typically few relevant random-effect variables, and their selection is not necessary.

Mathematically, given I groups (e.g., centers) indexed by $i = 1, 2, \ldots, I$, we observe n_i responses in the ith group, denoted by the n_i-dimensional vector $\boldsymbol{y}_i$. The associated fixed- and random-effect covariate values are denoted by the $n_i \times p$ matrix $\boldsymbol{X}_i$ and the $n_i \times q$ matrix $\boldsymbol{Z}_i$, respectively; often, $\boldsymbol{Z}_i$ is a subset of $\boldsymbol{X}_i$. Let $n = \sum_{i=1}^{I} n_i$ denote the total number of observations. Then, the LMM is defined as (Pinheiro and Bates, 2000)

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, ..., I, \tag{1.1}$$

where $\boldsymbol{\beta}$ is the fixed-effects (regression) coefficient vector, and $\boldsymbol{b}_i$ denotes the random effects

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

and $\boldsymbol{\epsilon}_i$ denotes the random error components in the model. We assume that $\boldsymbol{b}_i \sim N_q(0, \boldsymbol{\Psi}_{\boldsymbol{\theta}})$ and $\boldsymbol{\epsilon}_i \sim N_{n_i}(0, \sigma^2 \boldsymbol{I}_{n_i})$, for each $i = 1, \ldots, I$. Furthermore, they are independent of each other and of $\boldsymbol{X}_i$. Here, $\boldsymbol{I}_d$ denotes the identity matrix of order d, and $\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is a model variance matrix defined in terms of a q^*-dimensional (unknown) parameter vector $\boldsymbol{\theta}$; for example, $\boldsymbol{\Psi}_{\boldsymbol{\theta}} = Diag\{\theta_1, \ldots, \theta_q\}$ with $q^* = q$, or $\boldsymbol{\Psi}_{\boldsymbol{\theta}} = \theta_1 \boldsymbol{I}_q$ with $q^* = 1$, and so on. Then, given $\boldsymbol{X}_i$ (and $\boldsymbol{Z}_i$), $\boldsymbol{y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2 \boldsymbol{V}_i(\boldsymbol{\theta}, \sigma^2))$, independently for each i, where $\boldsymbol{V}_i(\boldsymbol{\theta}, \sigma^2) = \sigma^{-2}\boldsymbol{Z}_i \boldsymbol{\Psi}_{\boldsymbol{\theta}} \boldsymbol{Z}_i^T + \boldsymbol{I}_{n_i}$. Stacking the variables in larger matrices, we can rewrite the LMM (1.1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_I^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_I^T)^T$, $\boldsymbol{Z} = \text{Diag}\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I\}$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_I^T)^T \sim N_{qI}(\boldsymbol{0}, \boldsymbol{I}_q \otimes \boldsymbol{\Psi}_{\boldsymbol{\theta}})$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_I^T)^T \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Now, given $\boldsymbol{X}$, $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2))$, with $\boldsymbol{V} = \text{Diag}\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_I\}$. We wish to perform inference about the unknown regression parameter $\boldsymbol{\beta}$ and the variance parameter vector $\boldsymbol{\eta} = (\boldsymbol{\theta}, \sigma^2)$. As described earlier, we assume that $p \gg n$, but that $q \ll p, n$, and prefixed. In particular, we assume an ultrahigh-dimensional setup with $\log p = O(n^\alpha)$, for some $\alpha \in (0, 1)$, which is often the case in analyses of omics data. Then, the total number of parameters is $d := p + q^* + 1 \gg n$, and we need to impose a sparsity condition to estimate $\boldsymbol{\beta}$. This enables us to select the important fixed-effect variables; we assume that the number of such variables is $s \ll n$. Then, the selection of these s important fixed-effect variables and the parameter estimation are done by maximizing a suitably penalized log-likelihood function (Schelldorfer, Buhlmann, and Van de Geer, 2011; Ghosh and Thoresen, 2018; Fan and Li, 2012). The resulting estimator of $(\boldsymbol{\beta}, \boldsymbol{\eta})$ is known as the maximum penalized likelihood estimator (MPLE); see Section 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

Under our ultrahigh-dimensional regime, an important and desired property of the MPLE is the oracle variable selection consistency, which ensures that all (and only) the true important variables are selected, with probability tending to one asymptotically. All prior studies have examined this property of the MPLE under the crucial assumption that the covariates $\boldsymbol{X}$ are independent of the model error $\boldsymbol{\epsilon}$ and the random effects $\boldsymbol{b}$; these independence assumptions are referred to as the "*exogeneity*" of the model. However, they may not always hold in practice. The corresponding situation is referred to as "*endogeneity*" and is defined formally below.

Definition 1. Consider the LMM (1.1) and let the jth covariate be denoted as X_j.

- We have "unit-level endogeneity" or "*level-1 endogeneity*" when X_j is correlated with the model error term ϵ; that is, $Corr(X_j, \epsilon_i) \neq 0$. We also refer to this as "*error-covariate endogeniety*".

- We have "cluster-level endogeneity" or "*level-2 endogeneity*" when X_j is correlated with some random effect b; that is, $Corr(X_j, b) \neq 0$.

- A variable X_i is said to be "*endogenous*" when it is correlated with the model error term ϵ or with some random effect b. □

The problem of endogeneity has been studied extensively for classical low-dimensional settings, and appropriate remedies have been developed using suitable instrumental variables (IVs); see, among many others, Ebbes et al. (2004, 2015), Kim and Frees (2007), Wooldridge (2010,2012), and Bates et al. (2014). In our ultrahigh-dimensional setup, it is practically too demanding to always expect all exogeneity assumptions to hold. In particular, the assumption of independence between the error and all covariates is quite vulnerable and not verifiable for extremely large p. In Section 2, we show that the usual MPLE of the LMM parameters is affected significantly by endogeneity, which also significantly increases the number of false positives in a fixed-effects selection. To the best of our knowledge, no existing studies have examined the effects of such endogeneity and developed appropriate remedies under high-dimensional mixed models. This study aims to tackle this important problem, focusing particularly on level-1 endogeneity. In addition, we propose a new consistent selection procedure for fixed-effect variables, along with the estimation of all parameters, under such endogeneity.

The endogeneity issue in high or ultrahigh-dimensional models was first considered in Fan and Liao (2014) under the usual regression setup. The authors proposed a focused generalized method-of-moments (FGMM) estimator to consistently select and estimate the nonzero regression coefficients. In this study, we extend their FGMM approach to consistently select the important fixed-effect variables under our ultrahigh-dimensional LMM setup and then to estimate the variance parameters in a second stage. The proposed method is shown to satisfy the oracle variable selection consistency property for the fixed effects, even under error-covariate endogeneity. The overall procedure is implemented using an efficient algorithm, and is verified using suitable numerical illustrations.

The main contributions of this study are summarized as follows:

- We investigate the effects of endogeneity on the selection of fixed effects and parameter estimation in ultrahigh-dimensional LMMs. This is the first such attempt for mixed models with an exponentially increasing number of fixed effects, and we prove the inconsistency of the corresponding penalized likelihood procedures under endogeneity.

- We propose a new procedure for selecting important fixed-effects variables in the presence of level-1 endogeneity for the ultrahigh-dimensional LMM. Our method is based on the profiled focused generalized method-of-moments (PFGMM) approach. It handles the endogeneity issue by using appropriate IVs, and uses general nonconcave penalties, such as the smoothly clipped absolute deviation (SCAD) penalty, to carry out sparse variable selection. The problem of unknown variance components is solved by using an appropriate proxy matrix. Our proposed method produces significantly fewer false positives, both in simulations and in a real-data application, compared to the usual penalized likelihood method of Fan and Li (2012) in the presence of endogeneity in the data.

- We rigorously prove the consistency of the estimates of the fixed-effects coefficients $\boldsymbol{\beta}$ and their oracle variable selection property under appropriate verifiable conditions. Our assumptions on the penalty are completely general and cover most common nonconcave penalties, such as the SCAD penalty and the minimax concave penalty (MCP). The proof also allows the important selected variables to be endogenous by allowing the IVs to be completely external to the regression model.

- We also prove, under appropriate conditions, an asymptotic normality result for the estimates of the fixed-effects coefficients obtained by our PFGMM. This will further help us to develop testing procedures in endogenous high-dimensional LMMs in the future.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

- An efficient computational algorithm is also discussed, along with the practical issue of selecting the proxy matrix and the regularization parameter. Together with extensive numerical illustrations, we suggest their choices that are expected to work for most practical data with a strong signal-to-noise ratio. The (unoptimized) MATLAB code is available from the authors upon request.

- Once the important fixed-effects variables are selected consistently, we discuss and illustrate a few second-stage estimation procedures to estimate the variance parameters $\boldsymbol{\eta}$, along with refinements of the fixed-effects coefficients $\boldsymbol{\beta}$.

- Although our primary focus is on level-1 endogeneity, we briefly illustrate the effects of level-2 endogeneity on our proposed PFGMM approach for variable selection. Interestingly, our proposed method works consistently in most such scenarios, a finding we would like to investigate theoretically in subsequent works.

The rest of the paper is organized as follows. We start by describing the usual maximum penalized likelihood approach and its inconsistency in the presence of endogeneity in Section 2. In Section 3, we discuss the proposed PFGMM approach, including its motivation, oracle consistency of variable selection property, asymptotic normality result, and computational aspects, with numerical illustrations. In Section 4, we estimate the variance parameters in a second-stage refinement. The effect of level-2 endogeneity on the proposed PFGMM is examined numerically in Section 5, and a real-data application is presented in Section 6. Finally, Section 7 concludes the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

2. The MPLE under Endogeneity

We begin with a brief description of the MPLE under an ultrahigh-dimensional LMM, considering the notation of Section 1. Using the normality of the stacked response $\boldsymbol{y}$ in our LMM (1.2), the corresponding log-likelihood function of the parameters $(\boldsymbol{\beta}, \boldsymbol{\eta})$ turns out to be

$$l_n(\boldsymbol{\beta}, \boldsymbol{\eta}) \;=\; -\frac{1}{2}\left[n\log(2\pi) + \log|\sigma^2\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)| + \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]. \quad (2.1)$$

Adding an appropriate penalty to each component of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ using a penalty function $P_{n,\lambda}(\cdot)$, the MPLE is defined as the minimizer of the penalized objective function given by

$$Q_{n,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta}) = -l_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} P_{n,\lambda}(|\beta_j|). \quad (2.2)$$

With a suitable regularization parameter λ, the MPLE obtained by minimizing $Q_{n,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\eta})$ simultaneously selects the important (nonzero) components of $\boldsymbol{\beta}$ and estimates $\boldsymbol{\eta}$ consistently. However, the computation is a little tricky for different penalty functions. As a result, several extensions have been proposed. In particular, Schelldorfer, Buhlmann, and Van de Geer (2011) have considered the L_1 penalty in (2.2) under high-dimensionality, whereas Ghosh and Thoresen (2018) have extended the theory for general nonconcave penalties under both low and high-dimensional setups. An alternative two-stage approach has been proposed in Fan and Li (2012), which uses a proxy matrix in place of the unknown $\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)$, and then maximizes the resulting profile likelihood of $\boldsymbol{\beta}$ only, with suitable penalizations, to select the important fixed-effect variables; the estimation and selection of the random-effect variables are considered in a second step. Under certain assumptions, including

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

exogeneity (no endogeneity), these existing approaches to obtaining the MPLE all satisfy the oracle variable selection consistency property; that is, they estimate exactly the true active set (set of nonzero regression coefficients) with probability tending to one.

We now study the effects of different types of endogeneity on the MPLE using a numerical illustration. Here, we use the algorithm proposed by Ghosh and Thoresen (2018) with the well-known SCAD penalty (Antoniadis and Fan, 2001; Fan and Li, 2001); other algorithms indicate the same behavior of the MPLE under endogeneity, and are skipped for brevity. More illustrations are provided in later sections.

Example 2.1. We simulate random samples from the LMM (1.1) with $I = 25$, $n_i = 6$ for each i (such that $n = 150$), $p = 300$, $s = 5$, and $q = 2$, and the random-effects coefficients follow the distribution $N(\mathbf{0}, \boldsymbol{\Psi}_\theta)$, where $\boldsymbol{\Psi}_\theta = Diag\{\theta_1^2, \theta_2^2\}$. The design matrix $\boldsymbol{X}$ has the first column as $\mathbf{1}$, yielding the intercept. The next $(p-1)$ columns are chosen from a multivariate normal distribution with mean $\mathbf{0}_{p-1}$ and a covariance matrix with the (i,j)th element as $\rho^{|i-j|}$, for all $i, j = 1, \ldots, p-1$; the first two columns of $\boldsymbol{X}$ correspond to the two random-effect covariates and are kept nonpenalized. The true values of the parameters $\boldsymbol{\beta}$, σ^2, and θ_i^2 are $\boldsymbol{\beta} = (1, 2, 4, 3, 3, 0, \ldots, 0)^T$, $\sigma^2 = 0.25$, and $\theta_i^2 = 0.56$, for $i = 1, 2$, whereas $\rho = 0.5$ (correlated covariates) is considered. The SCAD penalty is used with tuning parameter $a = 3.7$, and the regularization parameter λ is chosen by minimizing the Bayesian information criterion (BIC) in each replication. We replicate this process 100 times, without endogeneity, to compute the summary measures about the performance of the MPLE, as reported in Table 1.

Next, to study the effects of endogeneity, some covariates X_{ij} are made endogenous with

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

either the model error (ϵ_i) or the kth random effect $\boldsymbol{b}_i$, using the transformations

$$X_{ij} \leftarrow (X_{ij}+1)(\rho_e \epsilon_i + 1), \text{ or } X_{ij} \leftarrow (X_{ij}+1)(\rho_b b_{ik} + 1), \text{ for all } i.$$

These produce correlations of $\frac{\rho_e \sigma}{\sqrt{2\rho_e^2\sigma^2+1}}$ and $\frac{\rho_b \theta_k}{\sqrt{2\rho_b^2\theta_k^2+1}}$, respectively, for the model error and kth random-effect coefficients with the endogenous covariates. The summary performance measures of the resulting MPLE under such endogeneity are reported in Table 1 for $\rho_e = \rho_b = 6$ (strong correlations of 0.688 and 0.698, respectively) and four particular sets of endogenous covariates: (i) Set 1: $X_6, \ldots, X_{15}$, that is, 10 unimportant covariates are endogenous; (ii) Set 2: $X_5, \ldots, X_{15}$, that is, 10 unimportant covariates and one important fixed-effect covariate are endogenous; (iii) Set 3: $X_2, X_6, \ldots, X_{15}$, that is, one important covariate with both a fixed-effect component and a random-effect slope is endogenous, along with 10 unimportant covariates; and (iv) Set 4: $X_6, \ldots, X_p$, that is, all unimportant covariates are endogenous.

The major observations from Table 1 and other similar simulations, not reported here for brevity, are summarized as follows:

- Under endogeneity, we have a significant increase in the number of false positives compared to the ideal exogenous case. The number of such wrongly selected fixed-effect variables further increases with the strength of the endogeneity and/or the number of endogenous variables. Such an effect is more serious for level-1 endogeneity than it is for level-2 endogeneity.

- Under level-1 endogeneity, we are not expected to lose any truly significant fixed-effect

Table 1: Empirical mean, SD, and MSE of the parameter estimates based on penalized MLE with SCAD penalty under different types of endogeneity, along with estimated active set size ($|S(\hat{\boldsymbol{\beta}})|$), number of true positives (TP), and the model prediction error (PE), adjusted for random effects (the column $\boldsymbol{\beta}_N$ denotes the average estimated β_j for $j = 6, \ldots, p$)

Endogenous	covariates	$\lvert S(\hat{\beta})\rvert$	TP	PE	β_1	β_2	β_3	β_4	β_5	β_N	σ^2	θ_1^2	θ_2^2
							No Endogeneity						
None	Mean	6.96	5.00	0.17	1.01	2.05	4.00	3.00	2.99	0.00	0.23	0.36	0.43
	SD	2.91	0.00	0.03	0.37	0.40	0.06	0.07	0.06	0.00	0.04	0.31	0.35
	MSE	–	–	–	0.1335	0.1604	0.0032	0.0047	0.0035	0.0000	0.0018	0.1375	0.1406
							Correlated with error (Level-1 endogeneity)						
Set 1	Mean	10.41	5.00	0.03	0.87	1.99	4.00	3.00	2.98	0.00	0.05	0.44	0.35
	SD	1.64	0.00	0.01	0.40	0.37	0.03	0.04	0.03	0.00	0.01	0.31	0.27
	MSE	–	–	–	0.1718	0.1329	0.0011	0.0013	0.0012	0.0000	0.0406	0.1123	0.1144
Set 2	Mean	10.13	5.00	0.03	0.90	2.02	4.00	2.98	3.03	0.00	0.05	0.35	0.39
	SD	1.95	0.00	0.01	0.34	0.38	0.03	0.03	0.01	0.00	0.01	0.28	0.30
	MSE	–	–	–	0.1222	0.1409	0.0010	0.0013	0.0008	0.0000	0.0420	0.1183	0.1187
Set 3	Mean	8.45	5.00	0.03	0.86	2.06	3.98	3.00	2.99	0.00	0.05	0.46	0.40
	SD	1.50	0.00	0.01	0.32	0.34	0.03	0.03	0.03	0.00	0.01	0.32	0.28
	MSE	–	–	–	0.1250	0.1201	0.0010	0.0009	0.0009	0.0000	0.0414	0.1138	0.1063
Set 4	Mean	23.45	5.00	0.01	0.86	2.00	3.99	3.00	2.99	0.00	0.01	0.49	0.39
	SD	4.64	0.00	0.00	0.38	0.34	0.01	0.02	0.02	0.00	0.00	0.35	0.28
	MSE	–	–	–	0.1607	0.1123	0.0003	0.0003	0.0003	0.0000	0.0586	0.1281	0.1100
							Correlated with random intercept (Level-2 endogeneity)						
Set 1	Mean	7.28	4.96	0.55	1.00	2.00	3.96	2.96	2.98	0.00	0.67	0.45	0.67
	SD	2.93	0.40	3.84	0.39	0.42	0.40	0.30	0.31	0.00	4.46	0.40	2.07
	MSE	–	–	–	0.1487	0.1781	0.1636	0.0937	0.0929	0.0000	19.8562	0.1677	4.2397
Set 2	Mean	7.21	5.00	0.17	0.99	2.01	4.00	3.00	3.00	0.00	0.23	0.43	0.44
	SD	2.54	0.00	0.03	0.33	0.32	0.06	0.06	0.01	0.00	0.04	0.37	0.39
	MSE	–	–	–	0.1075	0.1011	0.0039	0.0033	0.0001	0.0000	0.0018	0.1491	0.1672
Set 3	Mean	5.24	4.64	5.97	0.98	2.08	3.52	2.64	2.64	0.00	6.64	0.51	0.42
	SD	1.63	0.98	15.98	0.63	0.36	1.31	0.98	0.98	0.00	17.61	0.84	0.34
	MSE	–	–	–	0.3983	0.1311	1.9233	1.0834	1.0829	0.0000	347.9638	0.6949	0.1325
Set 4	Mean	12.47	4.28	7.66	0.98	1.62	3.29	2.46	2.43	0.00	8.80	0.43	4.69
	SD	6.78	1.54	16.27	0.36	0.82	1.55	1.16	1.15	0.00	18.63	0.71	9.74
	MSE	–	–	–	1.2489	3.2745	13.1795	6.9913	7.2183	0.1296	416.7764	0.5136	110.9365
							Correlated with random slope (Level-2 endogeneity)						
Set 1	Mean	7.70	5.00	0.17	1.02	1.95	4.00	3.01	2.98	0.00	0.24	0.39	0.42
	SD	3.58	0.00	0.03	0.32	0.38	0.08	0.07	0.06	0.00	0.04	0.31	0.34
	MSE	–	–	–	1.1992	3.9375	15.9995	8.6771	8.9100	0.1296	0.0019	0.1245	0.1353
Set 2	Mean	7.51	5.00	0.16	0.98	1.96	4.00	2.99	3.00	0.00	0.22	0.43	0.34
	SD	2.63	0.00	0.03	0.40	0.35	0.07	0.06	0.01	0.00	0.04	0.32	0.24
	MSE	–	–	–	1.2623	3.9604	16.0081	8.5702	8.9975	0.1296	0.0020	0.1208	0.1060
Set 3	Mean	5.28	4.72	4.75	0.85	1.94	3.64	2.73	2.74	0.00	5.31	0.51	0.41
	SD	1.54	0.90	14.70	0.63	0.42	1.15	0.86	0.87	0.00	16.29	0.70	0.34
	MSE	–	–	–	0.4140	0.1764	1.4431	0.8130	0.8138	0.0000	288.3149	0.4839	0.1350
Set 4	Mean	12.64	4.20	8.69	1.02	1.59	3.20	2.39	2.37	0.00	10.01	0.55	5.62
	SD	7.24	1.61	17.43	0.58	0.85	1.61	1.20	1.19	0.00	19.98	0.73	11.63
	MSE	–	–	–	1.4746	3.2622	12.7758	6.9904	7.0482	0.1296	490.5430	0.5336	159.4848

variables. However, in some cases of level-2 endogeneity, we may lose true positives as well.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

- The model prediction error is reduced in the presence of level-1 endogeneity, because more variables are selected in the final model. However, for level-2 endogeneity, the model prediction error can increase significantly when we lose the few true positives.

- The intercept is estimated with increased bias and MSE for level-1 endogeneity, whereas the estimates of the other fixed effects are affected more by level-2 endogeneity.

- The error variance also becomes severely underestimated in the presence of level-1 endogeneity. Level-2 endogeneity has a mixed effect in this case, producing significantly overestimated values of σ^2 for cases with higher degrees of endogeneity.

- As is well known, the random effect variances are, in general, underestimated, even under ideal exogenous conditions. The effect of endogeneity on these variances is not very clear, but is always moderate, except for level-2 endogeneity with the full set of unimportant variables $X_6, \ldots, X_p$ (Set 4).

Because our motivation is to select the important fixed-effect variables from a large pool of available candidates, in summary, the effect of level-1 endogeneity is more serious and needs proper treatment to decrease the number of false positives; on the other hand, level-2 endogeneity needs to be controlled to ensure no loss in true positives.

Note that we have an idea of the effects of different types of endogeneity on the MPLE; we can now investigate this from a theoretical point of view. In Theorem 1, we first present a set of necessary conditions for the MPLE to be consistent, both for the estimation and the fixed-effects selection, in the LMM (1.1). We then show that at least one of these conditions does not hold under endogeneity; hence, the MPLE is inconsistent under endogeneity.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

Theorem 1 (Necessary conditions for consistency of any sparse estimator in the LMM)**.** *Consider the LMM (1.1), where the estimation is performed by minimizing a general loss function $L_n(\boldsymbol{\beta}, \boldsymbol{\eta})$, which need not be the likelihood loss, along with a general penalty $P_{n,\lambda}(\cdot)$. Assuming sparsity of the true fixed-effect coefficient $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$, let $S = \{j : \beta_{0j} \neq 0\}$ be the (true) active set, $N = \{1, 2, \ldots, p\} \setminus S$, and $s = |S|$, which may or may not depend on the sample size n. Furthermore, assume the following results hold.*

(C1) $L_n(\boldsymbol{\beta}, \boldsymbol{\eta})$ is twice differentiable with respect to its arguments, and the maximum of its second derivatives at the true parameter value $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ is $O_p(1)$.

(C2) There is a local minimizer $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$ of the penalized objective function $L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} P_{n,\lambda}(|\beta_j|)$, which satisfies $P\left(\widehat{\boldsymbol{\beta}}_N = \mathbf{0}_{p-s}\right) \to 1$, $\sqrt{s}||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}|| = o_p(1)$, and $||\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|| = o_p(1)$, as $n \to \infty$, where $\widehat{\boldsymbol{\beta}}_S$ and $\widehat{\boldsymbol{\beta}}_N$ are the elements of $\widehat{\boldsymbol{\beta}}$ corresponding to the indices in S and N, respectively, and $\boldsymbol{\beta}_{0S}$ denotes the nonzero elements of $\boldsymbol{\beta}_0$ with indices in S.

(C3) The penalty function $P_{n,\lambda}$ is nonnegative with $P_{n,\lambda}(0) = 0$, $P'_{n,\lambda}(t)$ is nonincreasing on $t \in (0, u)$ for some $u > 0$, and $\lim_{n\to\infty} \lim_{t\to 0+} P'_{n,\lambda}(t) = 0$.

Then, for any $l \leq p$, we have

$$\left| \frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\beta}_l} \right| \xrightarrow{\mathcal{P}} 0. \tag{2.3}$$

The proof of this theorem is given in the online Supplementary Material. Here, note that Theorem 1 is established without any reference to the exogeneity or endogeneity conditions. It presents a necessary condition (2.3) for the parameter estimates to be consistent, as in (C2), for a general class of loss functions satisfying (C1) and the penalties satisfying (C3). Because it

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

is a necessity result, the rate of consistency in (C2) is not important. Furthermore, Condition 3 about the penalty function is the same as that used in Theorem 2.1 of Fan and Liao (2014), and is quite general. It is satisfied by most common penalties, including the L_1, SCAD, and MCP, by appropriately choosing the sequence of the regularization parameter $\lambda = \lambda_n$. Thus, as in Fan and Liao (2014), our result in Theorem 1 rather provides a necessary condition (2.3) on the loss function for a large class of useful penalty functions. Because (C1) always holds for the likelihood loss, if (2.3) is not satisfied, then the consistency results in (C2) cannot all hold for the resulting MPLE. It is known that (C2), and hence (2.3), must hold for the MPLE under exogeneity. In the following theorem, we show that (2.3) fails to hold under any sort of endogeneity, indicating the inconsistency of the MPLE, in at least one aspect; the proof is given in the online Supplementary Material.

Theorem 2 (Inconsistency of the MPLE in Endogenous LMM)**.** *Consider the LMM (1.1) with the likelihood loss given by (2.1), in negative, and with $P_{n,\lambda}(t)$ satisfying Condition (C3) of Theorem 1. Suppose that at least one X in at least one group i is endogenous (level-1 or level-2), and that the X and the model error ϵ both have finite fourth-order moments. If $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$ denotes a (local) MPLE such that $||\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|| = o_p(1)$, then either*

$$\limsup_{n\to\infty} P(\widehat{\boldsymbol{\beta}}_N = 0) < 1 \quad \textit{or} \quad ||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}|| \neq o_p(1),$$

where $\widehat{\boldsymbol{\beta}}_S$, $\widehat{\boldsymbol{\beta}}_N$, $\boldsymbol{\beta}_{0S}$, and $\boldsymbol{\eta}_0$ are defined as in Theorem 1 for the MPLE.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

3. Focused Selection of Fixed-Effect Variables under Level-1 Endogeneity

Consider the LMM setup described in Section 1. We now propose a new extension of the MPLE of Fan and Li (2012) that leads to consistent oracle selection of important fixed-effect variables, using the FGMM approach with nonconcave penalization. The FGMM loss function, as initially proposed by Fan and Liao (2014) in the context of a high-dimensional linear regression, simultaneously performs sparse selection and applies the IV method against endogeneity. The IV method basically assumes the availability of a vector of observable *instrumental variables* $\boldsymbol{W}$ that is correlated with the covariates $\boldsymbol{X}$, but uncorrelated with the model error; that is, $E[\epsilon|\boldsymbol{W}] = 0$. The choice of a proper IV (s) helps tackle different statistical problems. IVs are often chosen as a function of the covariates, or even as a subset of $\boldsymbol{X}$ and, hence, the above condition can be easily verified using some simple moment conditions. As noted earlier, the IV technique is seen to be extremely useful in addressing the endogeneity issues in classical low-dimensional LMMs; see Hall and Horowitz (2005), Wooldridge (2010), Lin et al. (2015), and Chesher and Rosen (2017) for some recent IV methods.

3.1 The PFGMM with Nonconcave Penalization

Under the LMM setup considered in this study, consistent with many real-life applications, we have assumed that the number of random effects is small enough that their individual analysis is possible in the classical sense. Hence, we assume that the matrix $\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is positive definite (pd). Let us first assume, for the time being, that the variance parameter $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \sigma^2)^T$ is known. Then, based on (2.1), the likelihood of the only parameter $\boldsymbol{\beta}$ becomes

$$L_{\text{profile}}(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}. \tag{3.1}$$

Note that this is also the profiled likelihood of $\boldsymbol{\beta}$ obtained by substituting the MLE of the random-effect vector $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_I^T)^T$, given $\boldsymbol{\beta}$, into the joint likelihood of $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_I^T)^T$ and $\boldsymbol{b}$ given covariates; see Fan and Li (2012) for details. A penalized version of this profile likelihood (in logarithm form) can be maximized for the sparse selection of the fixed effects and the estimation of the corresponding coefficients; Fan and Li (2012) have suggested using a suitable proxy matrix for the unknown $\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)$.

Now, in the presence of endogeneity, we need to additionally apply the IV method to achieve consistency. Let us again assume, for the time being, that $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \sigma^2)^T$ and, hence, that $\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)$ is known. Define the transformed variables

$$\boldsymbol{y}^* = \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1/2}\boldsymbol{y}, \qquad \boldsymbol{X}^* = \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1/2}\boldsymbol{X}, \qquad \boldsymbol{\epsilon}^* = \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1/2}\boldsymbol{\epsilon}.$$

Then, we have $\boldsymbol{\epsilon}^* \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_p)$, and hence $\boldsymbol{y}^* \sim N_n(\boldsymbol{X}^*\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_p)$. Therefore, the profile likelihood of $\boldsymbol{\beta}$, given in (3.1), under the LMM (1.1) is also the ordinary likelihood of $\boldsymbol{\beta}$ under the following linear regression model in the transformed space:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_p). \tag{3.2}$$

Under level-1 endogeneity in the LMM (1.1), we also have endogeneity in the transformed regression (3.2) with $E\left[\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}|\boldsymbol{X}^*\right] \neq \boldsymbol{0}$. Noting that (3.2) is the same model considered by Fan and Liao (2014), our idea is to apply their FGMM approach to this transformed model in the transformed space. Then, we return to the original space of the data to achieve our goal of fixed-effects selection in the LMM (1.1) with endogeneity. This would have been straightforward if the original data were independent and identically distributed (i.i.d.) and $\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)$ were known, but none of these conditions hold in practice. Hence, we need appropriate *nontrivial extensions* to handle the implementation and theoretical derivations. We begin by defining our proposed loss function.

Note that the components of $\boldsymbol{\epsilon}^*$ are i.i.d., and those of $\boldsymbol{y}^*$ are independent with the same variance, but different means. Let us denote the corresponding random variables in the transformed space by ϵ^* and Y^*, respectively. Then, obtaining a consistent solution under endogeneity is based on the availability of an appropriate set of observable IVs $\boldsymbol{W}^*$ in the transformed space, such that

$$E\left[\epsilon^* | \boldsymbol{W}^*\right] = 0.$$

Fan and Liao (2014) achieved variable selection consistency under endogeneity through over-identification by using two sets of sieve functions (Chen, 2007), say, $\boldsymbol{F}^* = (f_1(\boldsymbol{W}^*), \ldots, f_p(\boldsymbol{W}^*))^T$ and $\boldsymbol{H}^* = (h_1(\boldsymbol{W}^*), \ldots, h_p(\boldsymbol{W}^*))^T$, where f_j and h_j are scalar functions. Letting S denote the index set of true nonzero coefficients, the above IV condition implies that, for $\boldsymbol{\beta}_S = \boldsymbol{\beta}_{0S}$, we have the following set of over-identified equations:

$$E\left[(Y^* - \boldsymbol{X}_S^* \boldsymbol{\beta}_S)\boldsymbol{F}_S^*\right] = \mathbf{0}, \quad E\left[(Y^* - \boldsymbol{X}_S^* \boldsymbol{\beta}_S)\boldsymbol{H}_S^*\right] = \mathbf{0}. \tag{3.3}$$

Under these conditions, Fan and Liao (2014) proposed considering the FGMM loss function

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

$$\begin{aligned} L_n(\boldsymbol{\beta}) &= \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \boldsymbol{X}_i^*\boldsymbol{\beta})\boldsymbol{\Pi}_i^*(\boldsymbol{\beta})\right]^T \boldsymbol{J}(\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \boldsymbol{X}_i^*\boldsymbol{\beta})\boldsymbol{\Pi}_i^*(\boldsymbol{\beta})\right] \\ &= \left[\frac{1}{n}\boldsymbol{\Pi}^*(\boldsymbol{\beta})(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})\right]^T \boldsymbol{J}(\boldsymbol{\beta}) \left[\frac{1}{n}\boldsymbol{\Pi}^*(\boldsymbol{\beta})(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})\right], \qquad (3.4) \end{aligned}$$

where $\boldsymbol{\Pi}_i^*(\boldsymbol{\beta}) = (\boldsymbol{F}_i^*(\boldsymbol{\beta})^T, \boldsymbol{H}_i^*(\boldsymbol{\beta})^T)^T$, for all i, and $\boldsymbol{J}(\boldsymbol{\beta})$ is a diagonal weight matrix with nonzero weights corresponding only to the nonzero components of $\boldsymbol{\beta}$. In particular, the nonzero weights of the j components can be chosen as the inverse of the estimated variances of $f_j(\boldsymbol{W}^*)$ and $h_j(\boldsymbol{W}^*)$, respectively. Then, a consistent solution of the transformed problem can be obtained by minimizing the penalized FGMM loss function; see Fan and Liao (2014) for details.

Now, we reconsider our original problem under the LMM (1.1) and map the FGMM loss function (3.4) back into our data space. Note that, through an inverse transformation, we can assume $\boldsymbol{\Pi}^*(\boldsymbol{\beta}) = \boldsymbol{\Pi}(\boldsymbol{\beta})\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1/2}$, for some IV $\boldsymbol{\Pi}$ in the data space. Hence, the FGMM loss function for our mixed model setup has the form

$$L_n(\boldsymbol{\beta}) = \left[\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\beta})\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]^T \boldsymbol{J}(\boldsymbol{\beta}) \left[\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\beta})\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]. \qquad (3.5)$$

Note that, in practice with mixed models, we cannot directly minimize this FGMM loss function or its penalized version, because it depends on the unknown variance parameters through $\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)$. To avoid this problem, we follow the approach of Fan and Li (2012), and propose using $\widetilde{\boldsymbol{V}}_z = \left[\boldsymbol{I}_n + \boldsymbol{Z}^T\mathcal{M}\boldsymbol{Z}\right]$ in place of $\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)$, where $\mathcal{M}$ is some suitable proxy matrix for the unknown variance component matrix $\sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}}$. Therefore, we finally minimize, with respect to $\boldsymbol{\beta}$, the penalized objective function

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= L_n^P(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_{n,\lambda}(|\beta_j|), & (3.6) \\ \text{where} \quad L_n^P(\boldsymbol{\beta}) &= \left[\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\beta})\widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]^T \boldsymbol{J}(\boldsymbol{\beta}) \left[\frac{1}{n}\boldsymbol{\Pi}(\boldsymbol{\beta})\widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]. & (3.7) \end{aligned}$$

We refer to L_n^P as the PFGMM loss function based on its link to the profile likelihood. If we had $\boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)$ for the known variance parameters, the asymptotic consistency results for the resulting estimator would have followed directly from the results of Fan and Liao (2014). However, we here prove that, even using the proxy matrix $\widetilde{\boldsymbol{V}}_z$, we can still achieve variable selection consistency under the LMM, provided that the proxy matrix is not very far away from the truth. We present a rigorous proof, along with the necessary assumptions, in the next subsection.

3.2 Oracle Variable Selection Consistency

Consider the setup of the previous subsection, and assume that the true parameter value $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S}^T, \mathbf{0})^T$ is the unique solution of the set of over-identified IV equations in (3.3), where the nonzero component vector $\boldsymbol{\beta}_{0S} \in \mathbb{R}^s$. Furthermore, we need the following sets of assumptions.

Assumptions on the penalty (P):

The general penalty function $P_{n,\lambda}(t) : [0, \infty) \to \mathbb{R}$ satisfies

(P1) $P_{n,\lambda}(t)$ is concave and nondecreasing on $[0, \infty)$, with $P_{n,\lambda}(0) = 0$,

(P2) $P_{n,\lambda}(t)$ has a continuous derivative $P'_{n,\lambda}(t)$ on $(0, \infty)$, with $\sqrt{s}P'_{n,\lambda}(d_n) = o(d_n)$, where $d_n = \frac{1}{2}\min\{|\beta_{0j}| : \beta_{0j} \neq 0,\ j = 1, \ldots, p\}$ denotes the strength of the signal, (P3) there exists a constant $c > 0$, such that $\sup_{\boldsymbol{\beta} \in B(\boldsymbol{\beta}_{S_0}, cd_n)} \zeta(\boldsymbol{\beta}) = o(1)$, where

$$\zeta(\boldsymbol{\beta}) = \limsup_{\epsilon \to 0+} \max_{j \leq s} \sup_{t_1 < t_2 : (t_1, t_2) \in (|\beta_j| - \epsilon, |\beta_j| + \epsilon)} - \left[\frac{P_{n,\lambda}(t_2) - P_{n,\lambda}(t_1)}{t_2 - t_1} \right]. \tag{3.8}$$

Note that Conditions (P1)–(P3) are quite standard in high-dimensional analysis and used by several authors, including Fan and Liao (2014). These are satisfied by a large class of folded-concave penalties, including L_q with $q \leq 1$, hard-thresholding, SCAD, and MCP, for appropriately chosen tuning parameters. In addition, $\zeta(\boldsymbol{\beta}) \geq 0$, for any $\boldsymbol{\beta} \in \mathbb{R}^s$, by the concavity of the penalty functions. Condition (P2) is related to the signal strength, on which we need the following additional assumptions, depending on the dimension of the problem. These are needed to ensure variable selection consistency, and are satisfied by properly chosen SCAD and MCP penalties for strong signal d_n and small $s \ll n$.

Assumptions on the dimension and signal strength (A):

(A1) $P'_{n,\lambda}(d_n) = o(1/\sqrt{ns})$, $sP'_{n,\lambda}(d_n) + s\sqrt{\log p/n} + s^3 \log s/n = o(P'_{n,\lambda}(0^+))$,
$P'_{n,\lambda}(d_n)s^2 = O(1)$.

(A2) $s\sqrt{\log p/n} = o(d_n)$ and $\sup_{||\boldsymbol{\beta} - \boldsymbol{\beta}_{S0}|| \leq d_n/4} \zeta(\boldsymbol{\beta}) = o(1/\sqrt{s \log p})$.

Next, we assume the following conditions on the IVs $\boldsymbol{F}^*$ and $\boldsymbol{H}^*$, with the notation $F^*_j = f_j(\boldsymbol{W}^*)$ and $H^*_j = h_j(\boldsymbol{W}^*)$, for $j = 1, 2, \ldots, p$. These are motivated from Fan and Liao (2014), and similar justifications hold for their selection; see their Remark 4.1. We use the notation $\lambda_{\min}$ and $\lambda_{\max}$ to denote the smallest and largest eigenvalues, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

Assumptions on the Instruments (I):

(I1) There exists $b_1, b_2, r_1, r_2 > 0$, such that

$$\max_{l \le p} P(|F_l^*| > t) \le e^{-(\frac{t}{b_1})^{r_1}}, \qquad \max_{l \le p} P(|H_l^*| > t) \le e^{-(\frac{t}{b_2})^{r_2}}, \quad \text{for any } t > 0.$$

(I2) $\mathrm{Var}(F_j^*)$ and $\mathrm{Var}(H_j^*)$ are bounded away from both zero and infinity uniformly in $j = 1, \ldots, p$ and $p \ge 1$.

(I3) $\min_{j \in S} \mathrm{Var}\left((Y^* - \boldsymbol{X}^* \boldsymbol{\beta}_0) F_j^*\right)$ and $\min_{j \in S} \mathrm{Var}\left((Y^* - \boldsymbol{X}^* \boldsymbol{\beta}_0) H_j^*\right)$ are bounded away from zero.

(I4) There exist constants $C_1, C_2 > 0$, such that $\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T) < C_1$ and $\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^T) > C_2$, where

$$\boldsymbol{A} = \lim_{n \to \infty} \frac{1}{n} \boldsymbol{\Pi}(\boldsymbol{\beta}_{0S}) \widetilde{\boldsymbol{V}}_z^{-1} \boldsymbol{X}_S.$$

(I5) There exists a constant $C > 0$, such that $\lambda_{\min}(\boldsymbol{\Upsilon}) > C$, where

$$\boldsymbol{\Upsilon} = \lim_{n \to \infty} \frac{\sigma^2}{n} \boldsymbol{\Pi}(\boldsymbol{\beta}_0) \widetilde{\boldsymbol{V}}_z^{-1} \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2) \widetilde{\boldsymbol{V}}_z^{-1} \boldsymbol{\Pi}(\boldsymbol{\beta}_0)^T.$$

Note that Assumptions (I4) and (I5) additionally depend on the choice of proxy matrix $\mathcal{M}$, which appears in $\widetilde{\boldsymbol{V}}_z$. This $\widetilde{\boldsymbol{V}}_z$ is the key to handling the random effects in the loss function by substituting the unknown $V(\boldsymbol{\theta}, \sigma^2)$. However, we do not need $\widetilde{\boldsymbol{V}}_z$ to be consistent for $V(\boldsymbol{\theta}, \sigma^2)$ in our derivations; it is sufficient that the proxy matrix $\mathcal{M}$ is close to $\sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ in the sense of the following assumption.

Assumptions on the Proxy Matrix (M):

(M1) $\lambda_{\min}\left[C_1 \mathcal{M} - \sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}}\right] \ge 0$ and $\lambda_{\min}\left[C_1 \log n (\sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}}) - \mathcal{M}\right] \ge 0$, for some $C_1 > 1$.

(M2) $\max_{j \notin S} ||\boldsymbol{A}_j|| \sqrt{\log s / n} = o(P_{n,\lambda}(0^+))$, where $\boldsymbol{A}_j$ denotes the jth column of the matrix $\boldsymbol{A}$ defined in Assumption (I4).

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

Then, we have the following main theorem. For simplicity in presentation, we defer the proof to the online Supplementary Material.

Theorem 3. *Consider the setup of LMM (1.1), with the true parameter value being* $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$*. Assuming sparsity of* $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$*, let* $S = \{j : \beta_{0j} \neq 0\}$ *be the (true) active set, with size* $s = |S|$ *and* $N = \{1, 2, \ldots, p\} \setminus S$*. Suppose* $s^3 \log p = o(n)$ *and Assumptions (P), (A), (I), and (M) hold. Then, there exists a local minimizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T$ *of the PFGMM objective function* $Q(\boldsymbol{\beta})$ *in (3.6) that satisfies the following properties.*

a) $\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}}_N = \mathbf{0}) = 1$*, where* $\widehat{\boldsymbol{\beta}}_N$ *corresponds to the elements of* $\widehat{\boldsymbol{\beta}}$ *with indices in* N*.*

b) If $\widehat{S} = \{j \leq p : \widehat{\beta}_j \neq 0\}$ *denotes the estimated active set, then* $\lim_{n\to\infty} P(\widehat{S} = S) = 1$*.*

c) For any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$*,* $\sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) \xrightarrow{\mathcal{D}} N(0,1)$*, where* $\widehat{\boldsymbol{\beta}}_S$ *corresponds to the elements of* $\widehat{\boldsymbol{\beta}}$ *with indices in* S*,* $\boldsymbol{\beta}_{0S}$ *denotes the nonzero elements of* $\boldsymbol{\beta}_0$ *with indices in* S*,* $\boldsymbol{\Gamma} = 4\boldsymbol{A}\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{\Upsilon}\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{A}^T$*, and* $\boldsymbol{\Sigma} = 2\boldsymbol{A}\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{A}^T$*.*

d) In addition, the local minimizer $\widehat{\boldsymbol{\beta}}$ *is strict with probability arbitrarily close to one, for all sufficiently large* n*.*

Remark 1. Although we have three types of unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ in the LMM (1.1), the PFGMM loss function depends only on the fixed-effects coefficient parameter $\boldsymbol{\beta}$. Thus, the minimization of the PFGMM objective function $Q(\boldsymbol{\beta})$ in (3.6) only produces an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. This, in turn, selects the important fixed-effects associated with the nonzero coefficients in $\widehat{\boldsymbol{\beta}}$, owing to the use of a sparse nonconcave penalty function. Once $\widehat{\boldsymbol{\beta}}$ is obtained, the other

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

variance parameters $\boldsymbol{\eta} = (\boldsymbol{\theta}, \sigma^2)$ need to be estimated in a second stage, which is described later in Section 4.

Remark 2. Note that although we have proposed the PFGMM approach considering level-1 endogeneity in the LMM (1.1), its oracle consistency results in Theorem 3 are not hampered by the presence of level-2 endogeneity. This is because the transformed model (3.2) and, hence, the PFGMM loss function in (3.7) do not involve the random effects $\boldsymbol{b}$ if the proxy matrix is chosen appropriately. However, the required assumptions might become stricter if endogeneity is also present in the associated random-effect covariates $\boldsymbol{Z}$. Therefore, we expect the proposed PFGMM approach to work well in selecting important fixed-effect variables even, under level-2 endogeneity, in a well-specified LMM; we further illustrate this aspect empirically using simulations in Section 5.

3.3 Computational Aspects

To implement the proposed PFGMM algorithm, we follow the algorithm of Fan and Liao (2014) on the transformed variables $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$. However, these transformed variables leading to the loss function in (3.5) are not known, so we need to use the proxy matrix and the approximated loss given in (3.7). Therefore, given a proxy matrix $\mathcal{M}$, we first compute the matrix $\widetilde{\boldsymbol{V}}_z^{-1}$ and its square root $\widetilde{\boldsymbol{V}}_z^{-1/2}$. To compute the matrix square root, we use the blocked Schur algorithm, as developed by Deadman et al. (2013); its implementation can be found in standard statistical packages, such as MATLAB and R (function named "*sqrtm*" in both). Then, the approximations of the transformed variables are computed as $\widetilde{\boldsymbol{y}}^* = \widetilde{\boldsymbol{V}}_z^{-1/2}\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}}^* = \widetilde{\boldsymbol{V}}_z^{-1/2}\boldsymbol{X}$. Following this, the FGMM loss function based on $\widetilde{\boldsymbol{y}}^*$ and $\widetilde{\boldsymbol{X}}^*$ is nothing but the proposed loss in (3.7). Hence, we can devise an algorithm following the Fan and Liao (2014) approach. The resulting penalized PFGMM objective function is minimized by applying the iterative coordinate algorithm to a smoothed version of the nonsmooth PFGMM minimization problem. More details and justifications can be found in Fan and Liao (2014). For brevity, we present only the crucial considerations related to the choice of the proxy matrix and the choice of the regularization parameter λ in our context.

On the Choice of the Proxy Matrix:

The choice of the proxy matrix $\mathcal{M}$ is not straightforward from the assumed conditions; however, some guidance is provided by Fan and Li (2012, Section 2.3). In particular, assuming the standard nonsingularity conditions involving the random-effect covariates $\boldsymbol{Z}$, one possible choice of $\mathcal{M}$ that can be obtained for large n is $\log(n)$ times the identity matrix. We use this proxy matrix in all of the empirical illustrations presented here.

On the Choice of λ:

For the regression modeling considered in Fan and Liao (2014), the regularization parameter λ can be chosen using cross-validation and, hence, can also be used with any loss function other than the likelihood-loss (e.g., the FGMM loss). However, it is not ideal to apply the cross-validation technique to mixed models. In likelihood-based estimation and variable selection in high or ultrahigh-dimensional LMMs, the usual proposal is to choose λ corresponding to the minimum value of the BIC, given by (Schelldorfer et al., 2011; Delattre et al., 2014; Ghosh and Thoresen, 2018)

$$\mathrm{BIC}(\lambda) = -2l_n\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}\right) + [|\widehat{S}| + dim(\lambda)]\log n.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

When using the proposed PFGMM loss function to estimate $\boldsymbol{\beta}$ in the ultrahigh-dimensional mixed model, we can define a natural extension of the BIC as

$$\mathrm{ExBIC}(\lambda) = -2L_n^P\left(\widehat{\boldsymbol{\beta}}_\lambda^P\right) + |\widehat{S}| \log n,$$

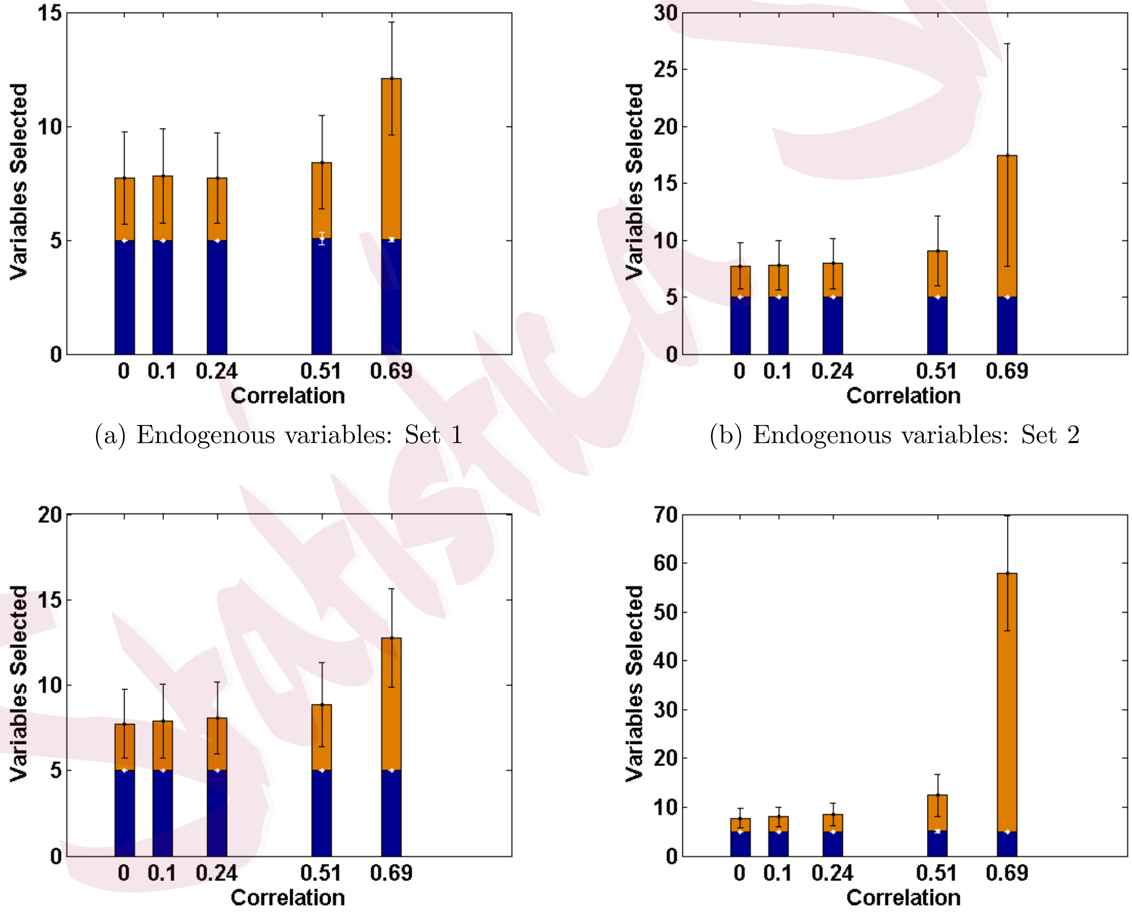
where $\widehat{\boldsymbol{\beta}}_\lambda^P$ is the estimate of $\boldsymbol{\beta}$ obtained using the proposed PFGMM approach with regularization parameter λ. However, in this context, it may be questionable whether the above formulations provide the correct penalty, which clearly needs further investigation. However, it has been observed that the simple choice of $\lambda = 0.1$, as suggested in Fan and Liao (2014), works sufficiently well for all our numerical studies.

3.4 Empirical Illustrations

We consider the same simulation setup as in Example 2.1 with level-1 endogeneity and different values of the underlying parameters, and apply the proposed PFGMM algorithm to select the relevant fixed-effects variables. In particular, we consider the true values of $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (1, 2, 4, 3, 3, 0, \ldots, 0)^T$, representing a strong signal, and the values of other parameters as $\sigma^2 = 0.25$ and $\theta_1^2 = \theta_2^2 = 0.56$, as in Example 2.1. Furthermore, we consider two values of ρ, 0 and 0.5, indicating uncorrelated and correlated covariates, respectively. We also use different values of $\rho_e \in \{0, 0.2, 0.5, 1.5, 6\}$ to represent varying strengths of endogeneity (with correlations of 0, 0.1, 0.24, 0.51, and 0.69, respectively). Note that $\rho_e = 0$ gives the ideal case with no endogeneity. We also studied negative values of ρ_e with the same magnitudes, leading to negative correlations. However, their effects are the same as those of the positive cases (depending only on the magnitudes) and hence, the, results are not reported for brevity. The

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

regularization parameter is set as $\lambda = 0.1$, following the suggestion of Fan and Liao (2014). For comparison, we also apply the profile likelihood proposal of Fan and Li (2012) (referred to here as the PLS method). The average sizes of the estimated active sets obtained by both methods are presented in Figures 1 and 2, respectively, for the correlated and the independent covariate cases.



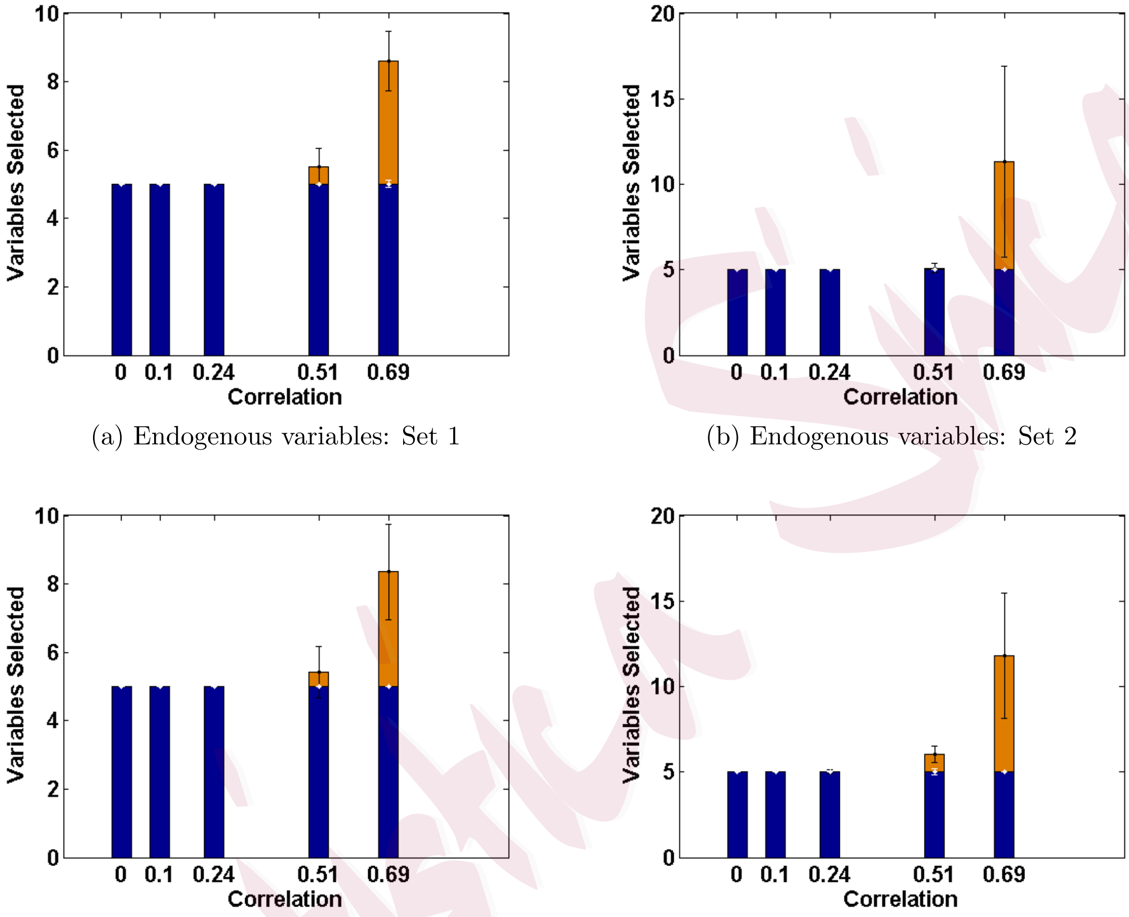
(a) Endogenous variables: Set 1

(b) Endogenous variables: Set 2

(c) Endogenous variables: Set 3

(d) Endogenous variables: Set 4

Figure 1: Sizes of the active sets estimated by the PFGMM (blue) method and the penalized least squares (PLS) method (blue+orange) for different extents of endogeneity and correlated covariates (the standard errors are shown by the respective error bars)



(a) Endogenous variables: Set 1

(b) Endogenous variables: Set 2

(c) Endogenous variables: Set 3

(d) Endogenous variables: Set 4

Figure 2: Sizes of the active set estimated by the PFGMM (blue) method and the PLS method (blue+orange) for different extents of endogeneity and independent covariates (the standard errors are shown by the respective error bars)

These empirical illustrations clearly show the significantly improved performance of the proposed PFGMM method under level-1 endogeneity. In particular, for correlated covariates, even a small amount of endogeneity (small ρ_e) increases the sizes of the active sets estimated by the PLS method. This becomes further damaging as the endogeneity increases, either because there are more endogenous variables or there are higher values of ρ_e. On the other hand, the

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

proposed PFGMM method produces an active set of size almost equal to the original active set size (5) under any extent of endogeneity. In addition, the variation over different replications is negligible compared with that of the PLS method. The results for the independent variables are also similar, although the harmful effect of endogeneity on the PLS method is not significant for smaller values of ρ_e. The proposed PFGMM method still outperforms the PLS method overall, producing the same sets of active variables when the PLS also performs well.

The next section describes the performance of the estimated regression coefficients obtained using the PFGMM, PLS, and further refinements, along with the estimation of the variance components.

4. Estimation of the Variance Parameters

Once we have selected the important fixed-effect variables consistently using the proposed PFGMM algorithm, the problem becomes low dimensional. Let $\widehat{S}$ denote the set of indices of the estimated nonzero coefficients, which is asymptotically the same as the true active set S, with probability tending to one from Theorem 3. Thus, we now have the reduced model

$$\boldsymbol{y}_i = \boldsymbol{X}_{i\widehat{S}}\boldsymbol{\beta}_{\widehat{S}} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, ..., I. \tag{4.1}$$

In addition, we have an estimate $\widehat{\boldsymbol{\beta}}_S$ of $\boldsymbol{\beta}_S$, which is consistent and asymptotically normal, from Theorem 3. Then, the most straightforward and intuitive estimates of $\boldsymbol{\eta}$ can be obtained by applying the maximum likelihood method to the resulting residual (random-effect) model,

$$\widehat{\boldsymbol{r}}_i := \boldsymbol{y}_i - \boldsymbol{X}_{i\widehat{S}}\widehat{\beta}_S = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, ..., I. \tag{4.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

We refer to the resulting estimator, say $\widehat{\boldsymbol{\eta}}^*$, as the PFGMM estimator of $\boldsymbol{\eta}$, in line with the associated PFGMM estimator $\widehat{\boldsymbol{\beta}}_S$ of $\boldsymbol{\beta}_S$. Note that the PFGMME of $\boldsymbol{\eta}$ is also consistent and asymptotically normal by standard results on likelihood-based inferences for the low-dimensional residual model (4.2). Once $\widehat{\boldsymbol{\beta}}_S$ has been computed, as described in Section 3, the PFGMME of $\boldsymbol{\eta}$ can be computed routinely using available software packages for a low-dimensional LMM (e.g., package "*lme4*" in R, function "*fitlme*" in MATLAB).

Alternatively, if we just want to use the proposed PFGMM to select the important fixed effects, in the second stage, we can fine-tune the estimates of $\boldsymbol{\beta}_{\widehat{S}}$ and the estimation of $\boldsymbol{\eta}$ to achieve better finite-sample efficiency. For this purpose, we consider the reduced low-dimensional linear mixed-effect model given in (4.1), containing only the $|\widehat{S}|$ fixed-effect variables from $\widehat{S}$ selected by the PFGMM algorithm. Then, we apply the standard maximum likelihood (ML) or the restricted maximum likelihood (REML) approach to get the new estimates $\widehat{\boldsymbol{\beta}}_{\widehat{S}}$ of $\boldsymbol{\beta}_{\widehat{S}}$ and $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$. Refer to the resulting estimators $((\widehat{\boldsymbol{\beta}}_{\widehat{S}}, \mathbf{0}), \widehat{\boldsymbol{\eta}})$ of $(\boldsymbol{\beta}, \boldsymbol{\eta})$ obtained by the second-stage ML and REML as the 2MLE and 2REMLE, respectively. Their performance in comparison to that of the PFGMM estimator of $(\boldsymbol{\beta}, \boldsymbol{\eta})$ is illustrated below using a simulation.

Example 4.1. We repeat the simulation exercise from Section 3.4, but now we estimate the parameters $(\boldsymbol{\beta}^T, \sigma^2, \theta_1^2, \theta_2^2)$ using the proposed PFGMM, 2MLE, and 2REMLE. The resulting mean values of the estimators, along with their standard deviations (SD) and mean squared errors (MSE), for the cases of exogeneity ($\rho_e = 0$) and extreme endogeneity with $\rho_e = 6$ are reported in Tables 2 and 3, respectively. For comparison, we have also reported the estimates obtained by the PLS method, where the variance parameters are estimated by maximizing the likelihood of the corresponding residual model.

Table 2: Empirical mean, SD, and MSE of different parameter estimates under no endogeneity

Method		β_1	β_2	β_3	β_4	β_5	β_N	θ_1^2	θ_2^2	σ^2
PLS	Mean	0.988	1.984	4.030	3.002	2.985	0.000	0.544	0.550	0.230
	SD	0.162	0.155	0.076	0.062	0.052	0.003	0.163	0.170	0.034
	MSE	0.026	0.024	0.007	0.004	0.003	0.000	0.026	0.029	0.002
PFGMM	Mean	0.987	1.997	4.003	2.995	2.996	0.010	0.544	0.550	0.230
	SD	0.168	0.153	0.071	0.060	0.048	0.000	0.163	0.170	0.034
	MSE	0.028	0.023	0.005	0.004	0.002	0.000	0.026	0.029	0.002
2MLE	Mean	0.990	1.994	4.003	2.995	2.996	0.000	0.540	0.549	0.241
	SD	0.161	0.153	0.070	0.059	0.049	0.000	0.163	0.173	0.035
	MSE	0.026	0.023	0.005	0.004	0.002	0.000	0.027	0.030	0.001
2REML	Mean	0.990	1.994	4.003	2.995	2.996	0.000	0.565	0.575	0.248
	SD	0.161	0.153	0.070	0.059	0.049	0.000	0.170	0.180	0.036
	MSE	0.026	0.023	0.005	0.004	0.002	0.000	0.028	0.032	0.001

One can clearly observe from Table 2 that, under exogeneity, the parameter estimates obtained from the methods are quite similar, although the estimates of the error variance are slightly better using the 2MLE or 2REML approaches. On the other hand, under endogeneity (Table 3), the PLS approach produces biased estimates of the fixed-effect intercepts, with a larger variance, and it significantly underestimates the error variance σ^2. The estimates obtained by the PFGMM method correct the bias of the intercept significantly, but still have a somewhat larger variance for this estimate and also an underestimated value of σ^2. However, the second-stage 2MLE and 2REMLE methods produce highly efficient estimators of both the fixed-effect coefficients and the variance parameters, which are similar to those obtained in the case of an exogenous model, even in the presence of the extreme endogeneity of correlation 0.68 for Sets 1, 2, and 4. Only for Set 3, where a random slope is correlated with the error vector, do our proposed methods still have some significant (negative) bias in estimating the

Table 3: Empirical mean, SD, and MSE of different parameter estimates under high level-1 endogeneity with $\rho_e = 6$ and correlated covariates

Method		β_1	β_2	β_3	β_4	β_5	β_N	θ_1^2	θ_2^2	σ^2
Endogeneity Variable: Set 1										
PLS	Mean	0.877	1.982	4.020	3.015	2.971	0.000	0.545	0.539	0.052
	SD	0.162	0.152	0.034	0.028	0.033	0.002	0.163	0.151	0.011
	MSE	0.041	0.023	0.002	0.001	0.002	0.000	0.027	0.023	0.039
PFGMM	Mean	0.965	1.974	4.004	2.996	2.994	0.010	0.545	0.539	0.052
	SD	0.212	0.254	0.069	0.060	0.050	0.000	0.163	0.151	0.011
	MSE	0.046	0.065	0.005	0.004	0.003	0.000	0.027	0.023	0.039
2MLE	Mean	0.992	1.992	4.003	2.996	2.994	0.000	0.535	0.548	0.240
	SD	0.164	0.155	0.070	0.059	0.050	0.000	0.163	0.173	0.038
	MSE	0.027	0.024	0.005	0.003	0.003	0.000	0.027	0.030	0.001
2REML	Mean	0.992	1.992	4.003	2.996	2.994	0.000	0.560	0.573	0.247
	SD	0.164	0.155	0.070	0.059	0.050	0.000	0.170	0.180	0.039
	MSE	0.027	0.024	0.005	0.003	0.003	0.000	0.028	0.032	0.001
Endogeneity Variable: Set 2										
PLS	Mean	0.873	1.979	4.034	2.990	3.032	0.000	0.546	0.540	0.047
	SD	0.154	0.150	0.048	0.033	0.014	0.004	0.159	0.152	0.011
	MSE	0.040	0.023	0.003	0.001	0.001	0.000	0.025	0.023	0.041
PFGMM	Mean	0.858	1.983	3.996	2.959	3.083	0.010	0.546	0.540	0.047
	SD	0.235	0.246	0.048	0.044	0.010	0.000	0.159	0.152	0.011
	MSE	0.075	0.060	0.002	0.004	0.007	0.000	0.025	0.023	0.041
2MLE	Mean	0.915	1.998	3.996	2.959	3.083	0.000	0.544	0.549	0.125
	SD	0.156	0.151	0.047	0.043	0.010	0.000	0.159	0.159	0.033
	MSE	0.031	0.022	0.002	0.004	0.007	0.000	0.025	0.025	0.017
2REML	Mean	0.915	1.998	3.996	2.958	3.083	0.000	0.568	0.573	0.129
	SD	0.156	0.151	0.047	0.043	0.010	0.000	0.166	0.166	0.033
	MSE	0.031	0.022	0.002	0.004	0.007	0.000	0.027	0.027	0.016
Endogeneity Variable: Set 3										
PLS	Mean	0.868	2.046	4.014	3.014	2.979	0.000	0.549	0.570	0.041
	SD	0.159	0.014	0.029	0.028	0.029	0.002	0.164	0.150	0.010
	MSE	0.042	0.002	0.001	0.001	0.001	0.000	0.027	0.022	0.044
PFGMM	Mean	0.894	2.094	3.998	3.007	2.997	0.010	0.549	0.570	0.041
	SD	0.184	0.009	0.045	0.043	0.036	0.000	0.164	0.150	0.010
	MSE	0.045	0.009	0.002	0.002	0.001	0.000	0.027	0.022	0.044
2MLE	Mean	0.905	2.094	3.998	3.007	2.996	0.000	0.548	0.594	0.105
	SD	0.158	0.009	0.045	0.042	0.036	0.000	0.164	0.165	0.026
	MSE	0.034	0.009	0.002	0.002	0.001	0.000	0.027	0.028	0.022
2REML	Mean	0.905	2.094	3.998	3.007	2.996	0.000	0.572	0.595	0.109
	SD	0.158	0.009	0.045	0.042	0.036	0.000	0.170	0.165	0.027
	MSE	0.034	0.009	0.002	0.002	0.001	0.000	0.029	0.028	0.021
Endogeneity Variable: Set 4										
PLS	Mean	0.842	1.979	4.034	3.018	2.983	0.001	0.538	0.518	0.005
	SD	0.156	0.148	0.019	0.019	0.017	0.002	0.158	0.138	0.002
	MSE	0.049	0.022	0.002	0.001	0.001	0.000	0.025	0.021	0.060
PFGMM	Mean	0.942	1.995	4.004	2.996	2.995	0.010	0.538	0.518	0.005
	SD	0.265	0.155	0.070	0.060	0.050	0.000	0.158	0.138	0.002
	MSE	0.073	0.024	0.005	0.004	0.002	0.000	0.025	0.021	0.060
2MLE	Mean	0.997	1.993	4.004	2.996	2.994	0.000	0.537	0.546	0.241
	SD	0.160	0.155	0.070	0.060	0.050	0.000	0.163	0.172	0.035
	MSE	0.025	0.024	0.005	0.004	0.003	0.000	0.027	0.030	0.001
2REML	Mean	0.997	1.993	4.004	2.996	2.994	0.000	0.562	0.572	0.248
	SD	0.160	0.155	0.070	0.060	0.050	0.000	0.170	0.180	0.036
	MSE	0.025	0.024	0.005	0.004	0.003	0.000	0.028	0.032	0.001

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

fixed intercept and error variance σ^2, although other parameters are estimated with excellent accuracy using 2MLE or 2REMLE. Note that, for the two-stage proposals, we are now again in a situation with endogeneity. See further comments related to this under Remark 3.

The other values of ρ_e give similar results, except for Set 3, and hence are omitted for brevity. In the case of endogenous random slope variables (Set 3) with moderate values of ρ_e (and hence correlations), our method surprisingly underestimates the fixed-effect intercept term to a larger magnitude. This needs further investigation; see Remark 3 below.

In summary, the proposed PFGMM method selects the true positive fixed-effect variables with an extremely small number of false positives under any extent of level-1 endogeneity. However, the resulting estimates of the fixed-effect coefficients are somewhat biased, and the resulting residual model underestimates the variance parameters, especially σ^2. Nevertheless, the second-stage estimators 2MLE or 2REML correct these to yield accurate estimators of all parameters under most level-1 endogeneity, except when the random slope is endogenous.

Remark 3 (When the endogeneous covariate also has a random effect)**.**

As already noted, although providing extremely good results in terms of our main target of fixed-effect selection, the proposed PFGMM and its second-stage refinement cannot fully address the parameter estimation problem (just like the PLS) when the covariates with random effects are endogenous with the error terms. However, because the proposed PFGMM selects the true active sets quite accurately, we can concentrate on the reduced low-dimensional model (using only the selected fixed-effect covariates) to obtain a corrected parameter estimate in the second stage, using a suitably modified approach instead of the 2MLE or 2REML. As

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

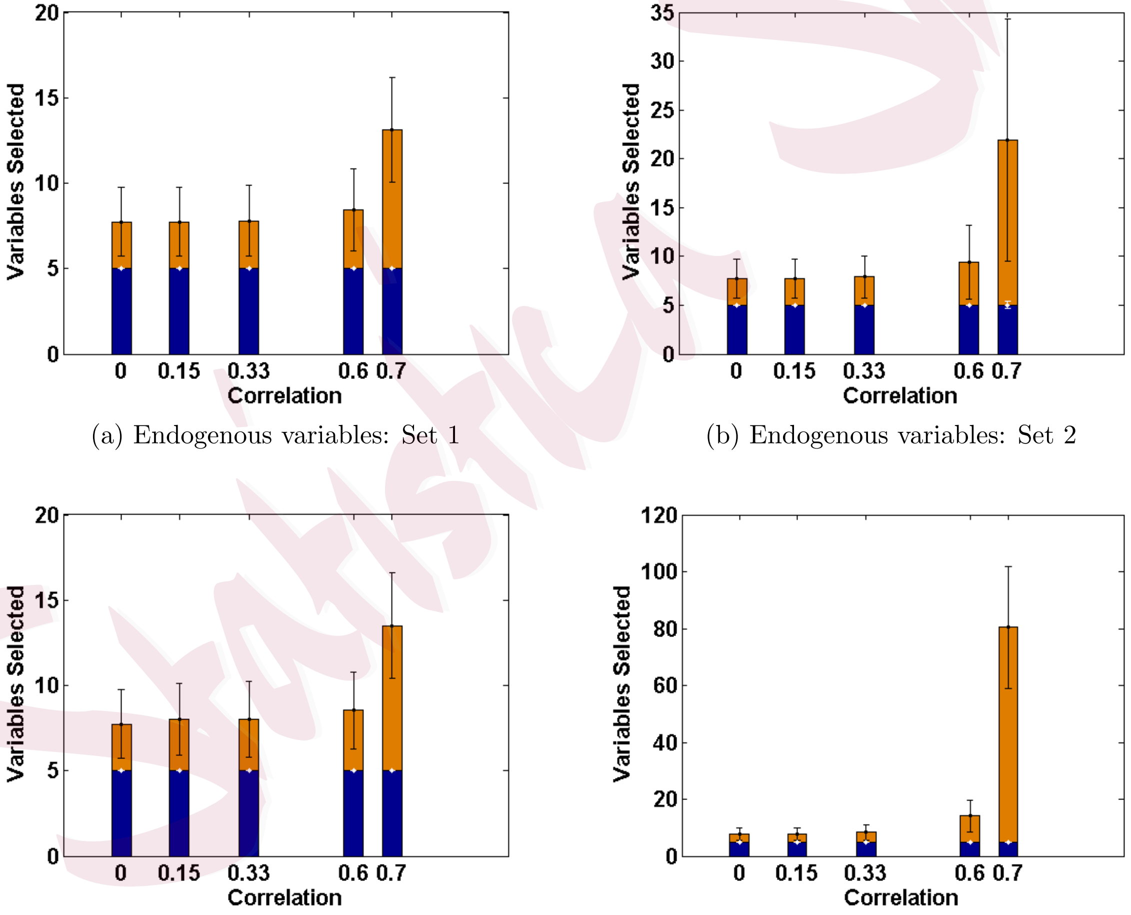
noted in the extensive literature on the endogeneity issue of mixed-effects models, a proper (low-dimensional) IV method (e.g., the two- or three-stage least squares) can be chosen for this purpose for a second-stage refinement to the PFGMM. In order to remain focused on fixed-effects selection in a high-dimensional context, we do not discuss these low-dimensional modifications for parameter estimation in the reduced model here, because they are well covered by the existing literature.

Another important phenomenon is observed in our simulations with Set 3 endogenous covariates and for different values of ρ_e. Surprisingly, the biases of the fixed-effect intercept and random-effect variances decrease with increasing extent of endogeneity, contrary to all other cases and our standard intuition. This contradictory behavior in all of the methods, PLS, PFGMM, 2MLE, and 2REML, indicates a need for further investigation, which we hope to study in future work.

5. What Happens in the Presence of Additional Level-2 Endogeneity?

Although we have developed our proposed method for the consistent selection of fixed-effect variables in an LMM with level-1 endogeneity, it is also of interest to examine how our proposed PFGMM and its second-stage refinements perform in the presence of level-2 endogeneity.

For comparative consistency, we reconsider the simulation setup of Example 2.1, but now with different extents of level-2 endogeneity in both the random intercept and the slope components separately, for $\rho_b \in \{0, 0.2, 0.5, 1.5, 6\}$; this leads to correlations of 0, 0.15, 0.33, 0.6, and 0.7, respectively, in both cases. Because the effect of level-2 endogeneity has already been observed to be significant in the case of correlated covariates, we present only the corresponding results for the selection of the fixed-effect variables (active set sizes) using the usual PLS method (Fan and Li, 2012) and the PFGMM method, for four sets of endogenous covariates, as in Example 2.1. These are shown in Figures 3 and 4, respectively, for the cases of endogenous random intercepts and slopes. From these figures, as well as additional simulations not reported here, we find that our proposed PFGMM method performs extremely well in selecting



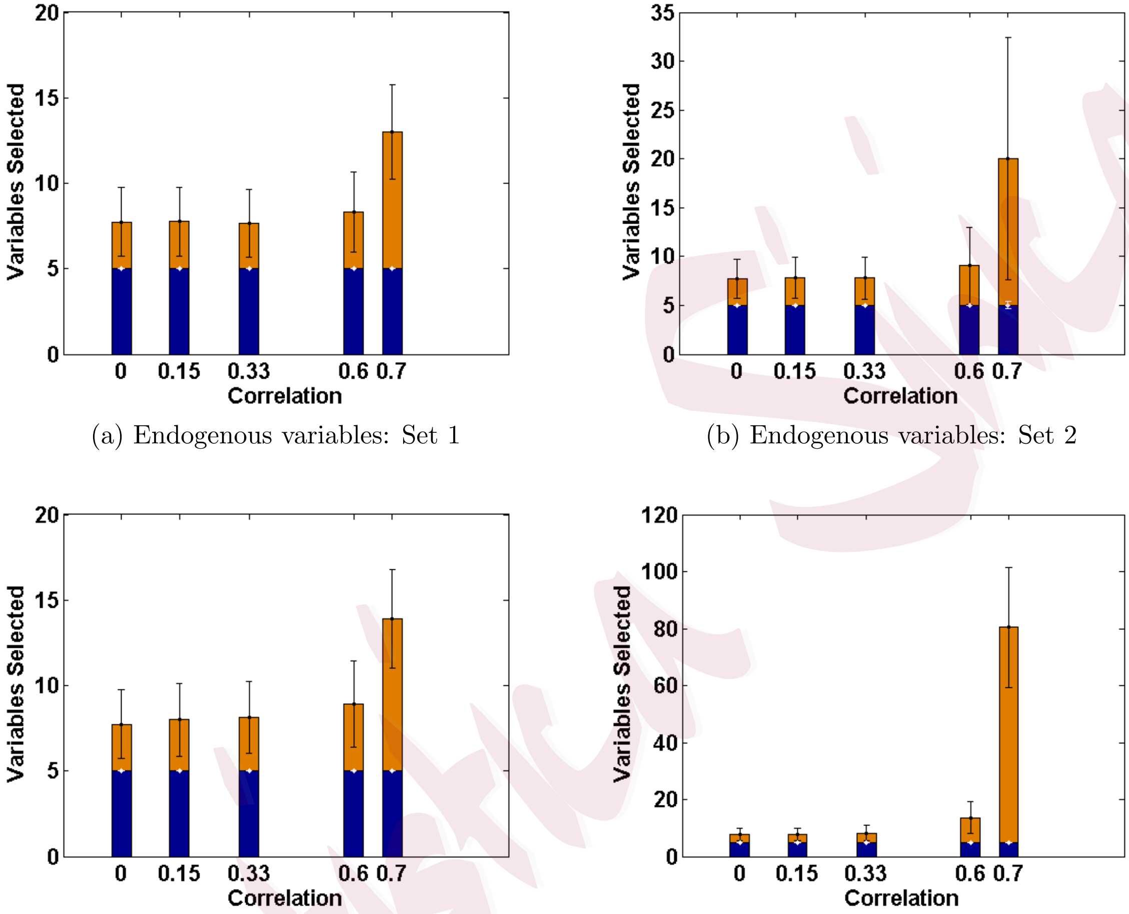
(a) Endogenous variables: Set 1

(b) Endogenous variables: Set 2

(c) Endogenous variables: Set 3

(d) Endogenous variables: Set 4

Figure 3: Sizes of the active sets estimated by the PFGMM (blue) method and the PLS method (blue+orange) for correlated covariates and different extents of level-2 endogeneity in the random intercept (the standard errors are shown by the respective error bars)



(a) Endogenous variables: Set 1

(b) Endogenous variables: Set 2

(c) Endogenous variables: Set 3

(d) Endogenous variables: Set 4

Figure 4: Sizes of the active sets estimated by the PFGMM (blue) method and the PLS method (blue+orange) for correlated covariates and different extents of level-2 endogeneity in the random slope (the standard errors are shown by the respective error bars)

exactly the truly significant variables, as compared with the PLS method, even in these cases of level-2 endogeneity. Except for very high levels of endogeneity, the PFGMM method selects exactly the true active set in most cases, as under level-1 endogeneity (or exogeneity).

Thus, if the main objective is to select important fixed-effect variables, the proposed PFGMM serves the purpose in the presence of any sort of endogeneity in the data, provided

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

the signal is reasonably strong. The requirement of a strong signal is related to the choice of regularization parameter λ, and is also expected from our Assumption (A), required to prove the oracle variable selection consistency of the PFGMM. As noted in Remark 2, the theoretical derivations are justified and linked through numerical illustrations.

We have also studied the effect of level-2 endogeneity on parameter estimation. The results are quite promising, except for Set 3, as in the case of level-1 endogeneity. However, a detailed discussion of these estimation results is beyond the scope of this paper.

6. A Real-Data Application

We analyze data from a randomized controlled cross-over trial with 47 subjects (Hansson et al., 2019). The subjects were exposed to four different meals with similar fat contents. The response was the serum concentration of triglycerids (TG) measured before the meal, and then two, four, and six hours after. It is well known that an elevated level of TG is associated with an increased risk of cardiovascular disease. Thus, it is of interest to understand the individual variation in TG responses and to characterize individuals with an unfavorable response. In this study, we focus on the lipid subclasses. In addition to the primary exposure (meal), we have measurements of lipid subclasses in blood, taken before each meal. Our primary interest is whether the triglyceride response to the meal (say y), as measured over six hours, depends on the level of some of the lipid subclasses (covariates x_js). We analyze this using the mixed

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

model

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i D_i + \sum_{j=1}^{K} \gamma_j x_j + \sum_{i=1}^{3} \sum_{j=1}^{K} \delta_{ij} D_i x_j + b_0 + \sum_{i=1}^{3} b_i D_i + \epsilon, \tag{6.1}$$

where D_i, for $i = 1, 2, 3$, denote dummy variables representing two, four, and six hours, respectively, and $K = 162$ is the number of available lipid subclasses. Here, we have four random-effect coefficients b_i, $i = 0, 1, 2, 3$, corresponding to a random intercept and three time dummy variables. Additionally, we have $4(1 + K) = 652$ fixed-effect coefficients β_i, γ_j and $\delta_{i,j}$, which need to be estimated from repeated (incomplete) observations from only 47 patients. However, we assume that only a few of the available lipid subclasses will influence the triglyceride response significantly, and our goal is to identify these subclasses.

Therefore, we are in a sparse high-dimensional regime, and we can apply the proposed PFGMM method as an alternative to the PLS method to select the important lipid subclasses, assuming $b_i \sim N(0, \sigma_i^2)$, for $i = 0, 1, 2, 3$, and $\epsilon \sim N(0, \sigma^2)$. It is difficult to test for endogeneity in high-dimensional models in practice. However, owing to the high dimensions, one will almost expect endogeneity to arise incidentally. In the current example, important potential confounders are omitted from the model, which would be expected to lead to endogeneity problems. From our simulation studies in Section 3.4, we observed that the PLS method has serious problems with over-selection in situations with endogenous covariates. Hence, we consider a large reduction in the number of selected variables by our method as a sign of endogeneity.

We applied both methods with different values of the regularization parameter λ for the purpose above. In each case, the PFGMM method selects noticeably fewer significant variables compared to the existing PLS approach, giving us fewer false positives; see Table 4 for a few illustrative cases. This clearly indicates the presence of significant endogeneity in the data, and the advantages of our proposed PFGMM approach becomes clear.

Table 4: Estimated active set size $|S|$ and variance components σ_i^2 $(i = 0, 1, 2, 3)$ and σ^2, for the real-data example, obtained by the PLS and the proposed PFGMM combined with the 2REML

$\lambda(\times 10^{-3})$	Method	S	σ_0^2	σ_1^2	σ_2^2	σ_3^2	σ^2
5	PLS	39	0.050	0.057	0.028	0.045	4.42E-07
	PFGMM+2REML	5	0.146	0.126	0.053	0.073	9.66E-08
2	PLS	113	0.030	0.048	0.022	0.039	7.33E-09
	PFGMM+2REML	11	0.146	0.098	0.031	0.063	2.13E-07
1.5	PLS	136	0.026	0.046	0.020	0.033	2.77E-09
	PFGMM+2REML	30	0.046	0.059	0.053	0.047	1.50E-07

In Table 4, the variance estimates obtained using the second-stage refinement 2REML are also reported. Clearly, the error variance reduces as we select additional fixed-effect variables by using lower λ-values. The appropriate model can be chosen via proper justification, along with the biological significance of the resulting model estimates. For example, the model with $\lambda = 2 \times 10^{-3}$ that selects 11 fixed effects looks the best candidate for the present example, because it provides a very low model error and still a rather sparse model. Of particular interest is the selection of seven interaction parameters, pointing to subclasses of interest when it comes to triglyceride responses. Without going into detail about the lipid subclasses, two of the discoveries seem obvious, because they are related to subclasses rich in triglycerides. Furthermore, four parameters point to subclasses related to Hdl cholesterol, a parameter known to be connected to triglycerides. The significance of the last subclass is unclear.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

As an alternative to the model selection above, one can apply a proper extension of the BIC to choose a data-driven value of the regularization parameter λ. Some indications are provided in Section 3.3, because the usual BIC is often affected by the presence of endogeneity in the data. However, further investigation is needed on appropriate BIC extensions under endogeneity.

Finally, note that we have used the unimportant covariates as a general vector of IVs, which performs well in all our simulations; such instruments were also suggested by Fan and Liao (2014) for dealing with high-dimensional regression models with endogeneity. A detailed study on finding an optimal IV is left to future research.

7. Conclusion

We have examined the problem of endogeneity in high-dimensional LMMs, focusing on the selection of important fixed-effect variables under error-covariate endogeneity. We have proved the inconsistency of the usual penalized likelihood approach for such cases, and proposed a new PFGMM approach for the consistent selection of the fixed-effects, combining the generalized method-of-moments, IVs, and proxy matrix for the unknown variance component matrix. The oracle variable selection property and the consistency and asymptotic normality of the estimated fixed-effects coefficients are derived under appropriate assumptions.

This work opens up many new research questions for future research. The immediate follow-up would be a detailed analysis of level-2 endogeneity and its effects on the usual likelihood method and our proposed PFGMM method. We should develop appropriate modifications in such cases, if needed, to establish the variable selection consistency. The second-stage

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

estimators may be further investigated for theoretical optimality. Furthermore, a suitable extension of the BIC should be studied, both theoretically and empirically, to select the regularization parameter from endogenous data. Although we have indicated a possible solution, a detailed analysis required in future work.

Supplementary Material:
The Online Supplementary Material contains proofs for the three theorems.

Acknowledgments

The majority of this work was completed when the first author was visiting the University of Oslo, Norway, through the support of a grant from the Norwegian Cancer Society. The research of the first author was also partially supported by the INSPIRE Faculty Research Grant from Department of Science and Technology, Govt. of India.

References

[1] Antoniadis A. and Fan, J. (2001) Regularization of Wavelets Approximations. *Journal of the American Statistical Association*, 96, 939–967.

[2] Bates, M. D., Castellano, K. E., Rabe-Hesketh, S., and Skrondal, A. (2014). Handling correlations between covariates and random slopes in multilevel models. Journal of Educational and Behavioral Statistics, 39(6), 524-549.

[3] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics VI*. J. J. Heckman and E. E. Leamer, Eds. North-Holland, Amsterdam.

[4] Chesher, A., and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3), 959-989.

[5] Deadman, E., Higham, N. J. and R. Ralha. (2013). *Blocked Schur algorithms for computing the matrix square root*. Lecture Notes in Comput. Sci., 7782, Springer-Verlag, pp. 171–182.

[6] Delattre, M., Lavielle, M., and Poursat, M. A. (2014). A note on BIC in mixed-effects models. *Electronic journal of statistics*, 8(1), 456-475.

[7] Ebbes, P., Böckenholt, U., and Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58(2), 161-178.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0421

[8] Ebbes, P., Wedel, M., Böckenholt, U., and Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3(4), 365-392.

[9] Fan J. and Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

[10] Fan J. and Li R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics*, 40(4), 2043–2068.

[11] Fan J. and Liao Y. (2014). Endogeneity in high dimensions. *Ann. Stat.*, 42(3), 872–917.

[12] Ghosh, A., and Thoresen, M. (2018). Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. *AStA Advances in Statistical Analysis*, 102(2), 179-210.

[13] Hall, P., and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6), 2904-2929.

[14] Hansson, P., Holven, K. B., Oyri, L. K. L., Brekke, H. K., Biong, A. S., Gjevestad, G. O., Raza, G. S., Herzig, K., Thoresen, M., and Ulven, S. M. (2019). Meals with Similar Fat Content from Different Dairy Products Induce Different Postprandial Triglyceride Responses in Healthy Adults: A Randomized Controlled Cross-Over Trial. *Journal of Nutrition*, 149(3), 422-431.

[15] Kim, J. S., and Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika*, 72(4), 505-533.

[16] Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509), 270-288.

[17] Pinheiro J.C. and Bates D.M. (2000). *Mixed-effects models in S and S-plus.* Springer-Verlag, New York.

[18] Schelldorfer J., Buhlmann P. and Van de Geer S. (2011). Estimation for high-dimensional linear mixed-effects models using l_1-penalisation. *Scandinavian Journal of Statistics*, 38, 197–214.

[19] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT press.

[20] Wooldridge J. (2012). Panel Data Models with Heterogeneity and Endogeneity. Online ppt at `https://www.ifs.org.uk/docs/wooldridge%20session%204.pdf`.