Stony Brook University

Adaptive Fitting of Mixed-Effects Models with Correlated Random-effects

A Dissertation Presented

by

Guangxiang Zhang

to

The Graduate School

in Partial fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2011

Stony Brook University

The Graduate School

Guangxiang Zhang

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree,
hereby recommend acceptance of this dissertation.

John J. Chen (Dissertation Advisor), Associate Professor,
Departments of Preventive Medicine & Applied Mathematics and Statistics

Nancy R. Mendell (Chairperson of Defense), Professor,
Department of Applied Mathematics and Statistics

Stephen J. Finch, Professor,
Department of Applied Mathematics and Statistics

Barbara Nemesure, Associate Professor,
Department of Preventive Medicine

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Adaptive Fitting of Mixed-Effects Models with Correlated Random-effects

by

Guangxiang Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2011

Linear mixed-effects model (LMM) has been widely used in hierarchical and longitudinal data analyses. In practice, the fitting algorithm can fail to converge because of boundary issues of the estimated random-effects covariance matrix, i.e., being near-singular, non-positive definite, or both. The traditional grand mean centering technique cannot generally improve the numerical stability and may even increase the correlation between random-effects. Also, current available algorithms are not computationally optimal because the condition number of random-effects covariance matrix is unnecessarily increased when the random-effects correlation estimate is not zero.

To improve the convergence of data with such boundary issue, we propose an adaptive fitting (AF) algorithm using an optimal linear transformation of the random-effects design matrix. It is a data-driven adaptive procedure, aiming at reducing subsequent random-effects correlation estimates down to zero in the optimal transformed estimation space. Extension of the AF algorithm to multiple random-effects models is also discussed. The

AF algorithm can be easily implemented with standard software and be applied to other mixed-effects models.

Simulations show that the AF algorithm significantly improves the convergence rate, and reduces the condition number and non-positive definite rate of the estimated random-effects covariance matrix, especially under small sample size, relative large noise, and high correlation settings. We also propose a new two-step modeling strategy for LMM fitting and random-effects selection. This parsimonious LMM with uncorrelated random-effects in the optimal transformed space is favored by the likelihood ratio test and Akaike Information Criterion. Two real life longitudinal data sets are used to illustrate the application of this AF algorithm implemented with software package R (nlme).

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply grateful to my advisor, Professor John J. Chen, for his great encouragement and guidance in the past five years. The fruitful and inspiring research work I have done with Professor Chen truly helped open a new frontier for my future career development. I also thank my committee members, Professor Nancy R. Mendell, Professor Stephen J. Finch and Professor Barbara Nemesure. Their advice and help were indispensable for my dissertation and other research projects.

I thank the Department of Applied Mathematics and Statistics for providing me with an excellent academic environment. I also appreciate the help and friendship of all the graduate students in our department, especially my officemate Hyeong Jun Ahn.

I thank all of my colleagues in the Department of Preventive Medicine, especially Dr. Liming Dong for proofing the writing for my prelim and Dr. Zhongming Yang for his latex template.

I thank the research supports and opportunities from the Biostatistical Consulting Core and my collaborators, including Drs. Daniel Baram, Frank A. Cervo, Sherry Courtney, Paul Impellizzeri, David Keller, Zhengrong Liang, Susmita Pati, Shinya Shibutani, Mark E. Wagshul, Thomas A. Wilson, and many others.

I thank my parents and wife for their sacrifice and support during my academic pursuit.

Chapter 1

Introduction

1.1 Background

Linear mixed-effects model (LMM) has been widely used for the analysis of hierarchical and longitudinal data (Laird and Ware, 1982). By incorporating random-effects into an ordinary regression model, LMM accommodates correlations among multiple observations made on the same unit (e.g., subject, group, cluster, classroom, center) and allows for unbalanced designs where all units do not require an equal number of observations and/or the same data collection occasions. LMM is also known as multilevel model (Goldstein, 1986), random coefficient model (Longford, 1993) or hierarchical linear model (Raudenbush and Bryk, 2002) in different substantive fields.

The maximum likelihood estimation of LMM parameters can be implemented by various numerical optimization algorithms, such as expectation-maximization (EM) (e.g., Laird and Ware, 1982; Dempster et al., 1984; Laird et al., 1987; Liu and Rubin, 1994; Meng and van Dyk, 1998), Newton-Raphson (e.g., Jennrich and Schluchter, 1986; Thompson and Meyer, 1986; Lindstrom and Bates, 1988; Callanan and Harville, 1991), iterative generalized least squares (IGLS) (Goldstein, 1986), or Fisher scoring (Longford, 1987), and is implemented in many standard or specialized softwares (see West et al., 2006 for details). However, an important practical problem of applying mixed-effects models is the slow or non-convergence issue during the nonlinear iterative maximizing likelihood process. When random-effects are highly correlated, it can sometimes be difficult to achieve convergence using these available

packages, or alternatively, an algorithm can converge to a random-effects covariance matrix that is near-singular or non-positive definite (PD). Several papers reported that the non-convergence rate can vary from a few percents to more than 50% in simulation studies (Meng and van Dyk, 1998; van Dyk, 2000; Browne and Draper, 2000; Berkhof and Snijders, 2001; Shieh and Fouladi, 2003), depending on algorithms used, simulation scale and settings, e.g., sample size, relative size of random-effects variance to the residual variance, and correlation level. It has also been shown that the estimated covariance matrix for random-effects could be non-PD (Meng and van Dyk, 1998; Mikulich et al., 1999; Browne and Draper, 2000; West et al., 2006; Pryseley et al., 2011) and the estimated correlation between random-effects could be close to unity (Meng and van Dyk, 1998; Pinheiro and Bates, 2000; Gurrin et al., 2001; Solaro and Ferrari, 2007). The non-convergence problem has been shown to get worse as the correlation level increased, either between random-effects (Browne and Draper, 2000) or between fixed-effects (Shieh and Fouladi, 2003).

During LMM fitting, the traditional grand mean centering technique, which transforms a predictor covariate around its grand mean, has been extensively used in practice (e.g., Kreft et al., 1995; van der Leeden et al., 1996; Morrell et al., 1997; Browne and Draper, 2000; Gurrin et al., 2001; Zhang and Davidian, 2001). It is even built in as a default or an optional setting for some softwares (Kreft et al., 1995) because of its ability to facilitate parameter interpretation and the possibility of improving numerical stability. For independent data, centering can remove the non-essential ill-conditioning in ordinary regressions, e.g., the sample covariance between intercept and slope being zero after centering (e.g., Marquardt and Snee, 1975; Bradley and Srivastava, 1979; Draper and Smith, 1998, ch. 16). However, for hierarchical or correlated data modeled by LMM, the computational consequences of centering on random-effects are complicated and not well understood. The reported correlation estimate between random intercept and random slope might be partially reduced after centering, but centering did not necessarily eliminate the collinearity problems between random-effects (e.g., Kreft et al., 1995; Pinheiro and Bates, 2000). After centering, both non-convergence and non-PD could still occur (Pinheiro and Bates, 2000; Browne and Draper, 2000). Therefore, grand mean centering is not a computationally optimal linear transformation for mixed-effects models in general.

Illustrative Example The following concrete numerical example illustrates the problem of the centering transformation. For simulated longitudinal continuous data fitted by a

LMM with two random-effects (b_{0i}, b_{1i}), where b_{0i} and b_{1i} are the random intercept and slope for the i-th subject, respectively, the possible impacts of different recoding methods for the observed raw time variable can be examined. Let

$$y_{ij} = \beta_0 + t_{ij}\beta_1 + b_{0i} + t_{ij}b_{1i} + \varepsilon_{ij}, \qquad i = 1, \cdots, m; \qquad j = 1, \cdots, n, \tag{1.1.1}$$

$$(b_{0i}, b_{1i})^T \sim N(0, G_{2\times 2}), \qquad \varepsilon_{ij} \sim N(0, \sigma_e^2), \qquad b_{0i} \perp \varepsilon_{ij}, \qquad b_{1i} \perp \varepsilon_{ij}.$$

The balanced longitudinal data set had continuous measurements Y_{ij} at the same five occasions $t_{ij} \in (1, 2, 3, 4, 5)$ (i.e., from 1 to 5 years) for the i-th subject in a sample of 50. To take into account the correlations among repeated measurements made on the same subject, a LMM with random intercept and random slope (RIS) model was applied at different time scales. Table 1.1 summaries the results of the RIS model examined at three usual time scales, namely, $t_{ij}^* = t_{ij} + \delta$, where δ was a location shift taking on the values of $\delta = 0, -3, -1$, which corresponded to raw, centering and follow-up scales, respectively. With a total of 250 data points, the RIS model failed to converge for the three time scales due to the estimated correlation between random-effects being on the boundary of the parameter space, i.e., correlation estimates closed to ± 1. However, after the introduction of a new location shift ($\delta = -1.0416$), RIS model could achieve convergence, and the new estimated correlation is reduced to 0.167.

Table 1.1: RIS model fitting results based on different recoding methods

Method	Raw	Centering	Follow-up	Optimal
Location shift (δ)	0	-3	-1	-1.0416
Predictor ($t_{ij} + \delta$)	(1,2,3,4,5)	(-2,-1,0,1,2)	(0,1,2,3,4)	t_{ij}^*
Converged?	No	No	No	Yes
$\widehat{Cor(b_{0i}, b_{1i})}$	-1.000	1.000	-0.990	0.167

$$t_{ij}^* = (1, 2, 3, 4, 5) + (-1.0416) = (-0.0416, 1.9584, 2.9584, 3.9584)$$

How to construct the location shift which can reduce the correlation in two and multiple random-effects cases will be described in Section 3.2 and 5.3.

1.2 Research goals

The numerical example above shows that there is possible link between convergence status and the correlation being estimated which is in turn dependent on three specific location shifts of the data. For a general linear transformation, the invariant nature of LMM inferences before and after transformation has been discussed (Longford, 1993; Morrell et al., 1997). Longford (1993) also identified an optimal linear transformation with zero correlation between random-effects in the search for minimum variance for response variable. Little work has been done on this optimal transformation with respect to its improving fitting of mixed-effects models. To the best of our knowledge, this optimal transformation has not been linked to the non-convergence issue during numerical estimation process nor been used as a replacement for the traditional centering technique.

The primary goal of my dissertation is to utilize the optimal linear transformation of the predictor variable to improve the numerical stability of LMM fitting. Focusing on the potential computational benefits from using this optimal transformation, we formulate an adaptive fitting (AF) algorithm, aiming to improve the non-convergence problems during the fitting of mixed-effects models with highly correlated random-effects. The second goal is to propose an AF-enhanced uncorrelated RIS model and compare it with other competing models based on several model selection criteria.

The dissertation is organized as follows. We first review the LMM and describe its linear transformation in Chapter 2. In Chapter 3, the AF algorithm is proposed, and its correlation and condition number reduction properties are provided. The performance of AF is studied through simulations (Chapter 4). The extension of AF to multiple random-effects cases is described in Chapter 5. After incorporating AF, a new two-step modeling strategy for LMM fitting and random-effects selection is proposed and investigated in Chapter 6. The application of the AF algorithm on LMM fitting and modeling are shown by two real life examples (Chapter 7). Discussions about this AF procedure are provided in Chapter 8.

Chapter 2

LMM and Linear Transformation

This chapter will review LMM, its general linear transformation and the associated model equivalency.

2.1 LMM and its estimation

LMM (Laird and Ware, 1982; Pinheiro and Bates, 2000) can be specified as:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \qquad i = 1, \cdots, m \tag{2.1.1}$$

$$b_i \sim N_q(0, G), \qquad \varepsilon_i \sim N(0, \sigma_e^2 R_i), \qquad b_i \perp \varepsilon_i,$$

where y_i is the observed response vector for ith subject, with scalar components y_{ij}, $j = 1, \ldots, n_i$; $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are design matrices of known covariates; β are the $p \times 1$ fixed-effects coefficients modeling the population-average effects; b_i are the $q \times 1$ random-effects modeling subject-specific effects and are assumed to be normally distributed with mean zero and a general covariance matrix G; ε_i are the $n_i \times 1$ unexplained errors and $R_i = I_{n_i}$ is usually assumed; b_i and ε_i are assumed to be independent.

If a LMM has two random-effects $b_i^T = (b_{0i}, b_{1i})$ with the corresponding $Z_i = (\mathbf{1}, x_i)$, where $\mathbf{1}$ is a constant vector of one and x_i is the observed slope predictor vector with scalar components x_{ij}, then it is a random intercept and slope (RIS) model. If the random slope b_{1i} term is dropped from a RIS model, then it becomes a random intercept only (RI) model.

Following the notations of Jacqmin-Gadda et al. (2007), let $\theta \triangleq (\beta, \phi)$, where $\phi \triangleq (G, \sigma_e^2)$ is the vector of covariance parameters. The vector θ can be estimated by maximizing the log-likelihood function of (2.1.1),

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{m}\{log(|V_i|) + (y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta) + n_i log(2\pi)\}, \tag{2.1.2}$$

where V_i is the marginal covariance matrix of response variable,

$$V_i = Cov(y_i) = Z_i G Z_i^T + \sigma_e^2 I. \tag{2.1.3}$$

The MLE of fixed-effects parameters β can be obtained after solving the score function $\partial l(\theta)/\partial\theta = 0$,

$$\widehat{\beta}(\widehat{\phi}) = (\sum_i X_i^T V_i(\widehat{\phi})^{-1} X_i)^{-1} \sum_i X_i^T V_i(\widehat{\phi})^{-1} y_i, \tag{2.1.4}$$

which is a generalized least squares estimator. However, the MLEs of covariance parameters ϕ do not have a general closed form solution and must be estimated iteratively by maximizing the log-likelihood function after plugging $\widehat{\beta}$ in (2.1.4) into β in (2.1.2). The asymptotic covariance matrix of the MLEs is estimated by the inverse of the Hessian matrix at the optimum $-\partial l^2(\theta)/\partial\theta\partial\theta^t$. The MLEs of β and ϕ are asymptotically independent given that $E(\partial l^2(\theta)/\partial\beta\partial\theta) = 0$. The MLEs of covariance parameters ϕ are biased and can be corrected by the restricted likelihood (REML) method, adjusting for the loss of degree of freedom due to estimating the fixed-effects in β. The corresponding REML log-likelihood differs from the full likelihood (ML) function in (2.1.2) only by a constant term plus $-0.5\sum_i log(|X_i^T V_i^{-1} X_i|)$.

There are several iterative optimization algorithms for fitting LMM. The EM algorithm will always converge to a local maximum of the likelihood surface but may need a very large number of iterations. The EM algorithm can be used to provide starting values for other algorithms (e.g., in R, Stata and HLM). Unlike the EM algorithm, the convergence of the Newton-Raphson algorithm is not guaranteed. However, the Newton-Raphson algorithm and its variations are the most commonly used algorithms to fit LMM, where an observed Hessian matrix is needed. The Iterative generalized least squares (IGLS) algorithm and the Fisher scoring algorithm are mathematically equivalent under normality assumption (Goldstein, 2002). The Fisher scoring algorithm uses the expected Hessian matrix and can be considered as a modification of the Newton-Raphson algorithm.

The main computational difficulty in applying LMM is the estimation of the covariance matrix (West et al., 2006, pp.30-31). It is well know that the Newton-Raphson algorithm

does not guarantee convergence or can converge to a non-PD covariance matrix (e.g., Lindstrom and Bates, 1988; Mikulich et al., 1999; van Dyk, 2000; West et al., 2006, ch. 6).

2.2 Linear transformation

To improve the numerical stability of LMM fitting, the traditional grand mean centering technique has been extensively used in practice, with the expectation of a computational benefit similar to that in ordinary regressions for independent data (e.g., van der Leeden et al. 1996, pp. 600; Morrell et al. 1997, pp. 339). However, for clustered or hierarchical data, the impacts of centering become more complicated, and the computational benefits in ordinary regression may no longer exist. In the context of RI model, there exists an optimal linear transformation of the response variable which can convert the correlated random vector (intercept and error) into an uncorrelated one to facilitate testing for independent normality (Hwang and Wei, 2006). For RIS model, an orthogonal linear transformation for random-effects has been identified but not proposed for numerical optimization purpose (Longford 1993). Reparameterization of the design matrix $[X_i, Z_i]$ by an orthogonal-triangular form has also been shown to reduce the complexity of computation and improve the LMM fitting (Lindstrom and Bates, 1988), although it is not a linear transformation method for the design matrix.

The general linear transformation for LMM has been discussed by Longford (1993) and Morrell et al. (1997). Let the original design matrices X_i and Z_i be transformed to $X_i^* = X_i A_1$ and $Z_i^* = Z_i A_2$, respectively. Namely,

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \varepsilon_i \\ &= X_i A_1 A_1^{-1}\beta + Z_i A_2 A_2^{-1} b_i + \varepsilon_i \\ &= X_i^*\beta^* + Z_i^* b_i^* + \varepsilon_i, \end{aligned} \tag{2.2.1}$$

where $\beta^* = A_1^{-1}\beta$, $b_i^* = A_2^{-1} b_i$, with both $A_1(p \times p)$ and $A_2(q \times q)$ assumed invertible. Morrell et al. (1997) showed that the likelihood function in (2.1.2) is invariant under linear transformations and that the REML function is only affected by a constant related to the determinant of A_1 if the scaling of fixed-effects exists (i.e., $|A_1|^2 \neq 1$). Thus REML is also invariant for a location shift linear transformation, i.e., matrix A_1 being a unit upper triangular matrix with all diagonal elements of one. Random-effects design matrix Z_i is

usually a subset of its fixed-effects counterpart X_i since the mean of random-effects $E(b_i)$ is assumed zero (Pinheiro and Bates, 2000). For convenience and simplicity, we assume here the non-singular linear transformation matrices for X_i and Z_i are the same in (2.2.1), i.e., $A_1 = A_2 = A$. Thus,

$$G^* = Cov(b_i^*) = Cov(A^{-1}b_i) = A^{-1}Cov(b_i)(A^{-1})^T = A^{-1}G(A^{-1})^T. \tag{2.2.2}$$

How to construct an optimal transformation matrix A in two and multiple random-effects cases will be described in Section 3.2 and 5.3.

Morrell et al. (1997) also illustrated that a linear transformation should be applied on a LMM which follows the hierarchical principle. The hierarchical principle requires that a lower order term should be kept in the model no matter whether it is statistically significant, as long as a higher order term which it is involved with appears in the model. Using a LMM with random slope only (denoted as "RS" model) as an example, Morrell et al. (1997) showed that the RS model could not be one-to-one linear transformation back into the original space.

2.3 Model equivalency

We follows the definition of Kreft et al. (1995) in the multilevel modeling setting. If two different models produce the same set of mean and variability profiles for the outcome variable Y_{ij}, they are *equivalent*. For LMM, we only need to check the response mean $E(Y_{ij})$ and response variance $Var(Y_{ij})$, because two normal distributions are identical if and only if they share the same mean vector and covariance matrix. Two equivalent models may have a different set of parameters describing them. For equivalent models, some parameterizations can be more parsimonious than others . If the transformation between two different parameterizations is one-to-one, then the models formulated by the two parameterizations are equivalent. The numerical results of estimates can be used to verify the equivalency of models. In general, fitting two equivalent models may not produce the same parameter estimates. But if the same maximum likelihood estimation procedure is used, one can expect that the estimates should be the same for equivalent models due to the invariance property of maximum likelihood estimation.

Chapter 3

The Proposed Adaptive Fitting (AF) Algorithm

3.1 Introduction

When random-effects are highly correlated, the iterative fitting process can be slow or non-convergent. To improve the convergence condition due to estimated correlation between random-effects on the boundary, we propose an adaptive fitting (AF) algorithm through transforming a random-effect covariate using the optimal location shift.

In this chapter, we discuss the proposed AF algorithm for a LMM with two random-effects (i.e., RIS model). The extension to a LMM with multiple random-effects scenarios will be described in Chapter 5. Section 3.4.2 provides the rationale on why the traditional grand mean centering technique cannot generally improve the numerical stability and may even increase the correlation between random-effects. Section 3.4.3 addresses why current available LMM fitting algorithms are not computationally optimal because the condition number of random-effects covariance matrix is unnecessarily increased when the random-effects correlation estimate is not zero.

For a RIS model with random intercept b_0 and random slope b_1, let $Z_i = (\mathbf{1}, x_i)$. Given a non-singular linear transformation (A) of observed data Z_i, we have $Z_i^* = Z_i A$. The corresponding implicit change in random-effects is $b_i \Rightarrow b_i^* = A^{-1} b_i$.

3.1.1 Notations for two random-effects cases

δ : a general location shift which may reduce or increase the original correlation.

d : the optimal location shift which drives the original correlation to zero.

A_δ : $A = A_\delta \triangleq \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix}$, where A is a general location shift transformation matrix specified by a scalar location shift index of δ, corresponding to the transformation, $x_{ij} \Rightarrow x^*_{ij} = x_{ij} + \delta$.

$G(\rho)$: $G = G(\rho) = Cov(b_i) \triangleq \begin{pmatrix} \sigma^2_{b_0} & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma^2_{b_1} \end{pmatrix}$, $|\rho| < 1$, where G is the random-effects covariance matrix in the original space, with the corresponding correlation ρ between random intercept and random slope.

$G^*_\delta(\rho^*)$: $G^* = G^*_\delta(\rho^*) = Cov(b^*_i) = Cov(A^{-1}_\delta b_i)$, where G^* is the random-effects covariance matrix in a transformed space ($Z_i \Rightarrow Z^*_i = Z_i A_\delta$), with the corresponding new correlation ρ^*. If $\delta = 0$, then $G^* = G$ and $\rho^* = \rho$.

3.2 Optimal linear transformation

Applying a general location shift transformation (equation 2.2.2) to a RIS model, we have,

Lemma 3.2.1. *For a general initial covariance matrix $G = Cov(b_i)$ and a general transformation matrix $A = A_\delta$, the new covariance matrix in the transformed space is,*

$$G^* = Cov(b^*_i) = A^{-1}G(A^{-1})^T = \begin{pmatrix} \sigma^2_{b_0} - 2\delta\rho\sigma_{b_0}\sigma_{b_1} + \delta^2\sigma^2_{b_1} & \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma^2_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma^2_{b_1} & \sigma^2_{b_1} \end{pmatrix}.$$

We define the optimal shift d to be a location shift such that the resulting matrix G^* above becomes diagonal in the transformed space. To diagonalize G^*, set $\rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma^2_{b_1} = 0$, and we get $\delta = \rho\sigma_{b_0}/\sigma_{b_1}$. Thus, the optimal shift d is the ratio of the covariance between random intercept and random slope divided by the variance of random slope, i.e.,

$$d \triangleq \rho\frac{\sigma_{b_0}}{\sigma_{b_1}} = \frac{Cov(b_0, b_1)}{Var(b_1)}. \tag{3.2.1}$$

The optimal covariance matrix G^* can be obtained after replacing the general δ in Lemma 3.2.1 with the optimal shift d. Thus,

Theorem 3.2.2. *After a location shift by A_δ, the new covariance matrix in the transformed space G^* will become diagonal (denoted as G_d^* or Ω), i.e.,*

$$G^* = Cov(b_i^*) = G_\delta^*(\rho^* = 0) = \begin{pmatrix} \sigma_{b_0}^2(1-\rho^2) & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix} \triangleq \Omega,$$

if and only if $\delta = d$.

In other words, correlated random-effects will become uncorrelated after the optimal linear transformation $Z_i = (\mathbf{1}, x_i) \Rightarrow Z_i^* = Z_i A = (\mathbf{1}, x_i + d \cdot \mathbf{1})$, where matrix $A = A_d$ is specified by the d in equation 3.2.1. For the longitudinal data settings, this is a location shift transformation changing the origin of time variable.For example, a longitudinal study with predictor time variable x_{ij} for i-th subject at j-th occasion, transforming of the slope covariate $x_{ij} \Rightarrow x_{ij}^* = x_{ij} + d$ will lead to a zero correlation among random intercept and slope.

3.3 AF algorithm

Inspired by the equation 3.2.1, we propose an adaptive fitting (AF) algorithm to reduce the subsequent estimated correlation numerically closer to zero in the transformed estimation space, by transforming the slope covariate using the optimal location shift d adaptively. This process can be repeated and the convergence can be improved iteratively until the current location shift d estimate is down to zero.

For a RIS model, the steps of an AF algorithm can be described as:

1. Fit the model using the untransformed initial covariate x_{ij}. Let iteration index $k = 0$, and denote the initial $x_{ij} = x_{ij}^0$ and the initial location shift $d_0 = 0$.

2. Obtain random-effects covariance matrix estimate $\widehat{G}$ from current fitting outputs. For non-convergent cases, results from the last iteration can be used. Denote the initial $\widehat{G} = \widehat{G}_0$.

3. Calculate current observed optimal location shift d_{k+1} from current covariance matrix estimate $\widehat{G}_k$,

$$d_{k+1} = \frac{\widehat{Cov(b_0, b_1)}}{\widehat{Var(b_1)}} = \widehat{\rho}_k \frac{\widehat{\sigma_{b_0,k}}}{\widehat{\sigma_{b_1,k}}} \tag{3.3.1}$$

4. If d_{k+1} is not zero, go to the next step; otherwise stop and go to Step 7.

5. Adaptively fit the model using updated transformed covariate $x^*_{ij} = x^{k+1}_{ij}$,

$$x^{k+1}_{ij} = x^k_{ij} + d_{k+1},$$

and obtain updated parameter estimates in the transformed space, such as $\widehat{G^*}$ and denote as $\widehat{G}_{k+1}$.

6. Update $k = k + 1$ and go to Step 3.

7. Obtain the final estimates for (β, ϕ) in the original space by taking the one-to-one inverse transformation. Based on the fitted $\widehat{G^*}$ at the last iteration, the corresponding random-effects covariance matrix estimate $\widetilde{G}$ in the original space can be obtained by

$$\widetilde{G} = A\widehat{G^*}A^T, \tag{3.3.2}$$

where A is specified by the sum of cumulative shifts $(\sum_{i=0}^{k+1} d_i)$ relative to the original space.

In Step 4 above, whether another AF step is needed depends on whether the current observed location shift d_{k+1} is close enough to zero. Empirically, the stopping rule with a magnitude at 0.001 provides reasonable good results. Step 7 may not be needed for those parameters or measures which are invariant with a location shift, e.g., the determinant of a estimated random-effects covariance matrix.

3.4 AF algorithm properties

We first illustrate the general impact of a location shift δ on the random-effects covariance matrix in the transformed space in Section 3.4.1, and then establish several Theorems for the optimal shift d in Section 3.4.2 and 3.4.3. Theorem 3.4.2 addresses the question on what kind of location shift transformation will be able to reduce the correlation level and why

traditional centering approach can actually increase the correlation level. Theorems 3.4.3, 3.4.7 and 3.4.9 show whether and how the condition number of random-effects covariance matrix is reduced in the optimal transformed space.

Recall that in the original estimation space with observed data (Y, X), we have covariance matrix G and random-effects $b_i = (b_0, b_1)^T$. After a non-singular transformation of the original data X by A, the model is refitted in the transformed space using pseudo-new observed data $(Y, X^* = XA)$, with new matrix G^* and transformed random-effects $b_i^* = A^{-1}b_i$.

3.4.1 Impact of a general location shift

Lemma 3.4.1. *Assume a diagonal initial covariance matrix,* $G = Cov(b_i) = G(\rho_0 = 0) = \begin{pmatrix} \sigma_{b_0}^2 & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix}$ *and a general* $A = A_\delta$. *Then* $G^* = Cov(b_i^*) = A^{-1}G(A^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 + \delta^2\sigma_{b_1}^2 & -\delta\sigma_{b_1}^2 \\ -\delta\sigma_{b_1}^2 & \sigma_{b_1}^2 \end{pmatrix}$, *and*

1. $\forall\ \delta \neq 0 \ \Rightarrow\ \rho^* = Cor(b_0^*, b_1^*) \neq 0$;
2. $|\rho^*| \to 1$, *as* $|\delta| \to \infty$; $|\rho^*|$ *is a monotonically increasing function of* $|\delta|$.

Proof. The first result is obvious. We just need to prove the second result in the Lemma above. By definition, $\rho^* = -\frac{\delta\sigma_{b_1}}{\sqrt{\sigma_{b_0}^2+\delta^2\sigma_{b_1}^2}} = -\frac{\delta}{\sqrt{\sigma^2+\delta^2}} = -\frac{\delta}{h} = f(\delta, h)$, where $\sigma = \frac{\sigma_{b_0}}{\sigma_{b_1}}$ and $h = \sqrt{\sigma^2 + \delta^2}$.

Taking first order partial derivative, we have
$\frac{\partial f}{\partial \delta} = -\frac{h-\delta h'}{h^2} = \frac{\delta^2/h - h}{h^2} = \frac{\delta^2 - h^2}{h^3} = -\frac{\sigma^2}{h^3} < 0.$

Thus, $|\rho^*| = \frac{|\delta|}{\sqrt{\sigma^2+\delta^2}}$ monotonically goes to unity, as $|\delta| \to \infty$ for a fixed ratio $\sigma_{b_0}/\sigma_{b_1}$.

□

Lemma 3.4.1 can be illustrated by plotting the new correlation between random-effects in a transformed space as a function of location shift across various variance component ratios of the initial diagonal matrix G (Fig. 3.1). After a location shift perturbation, the new

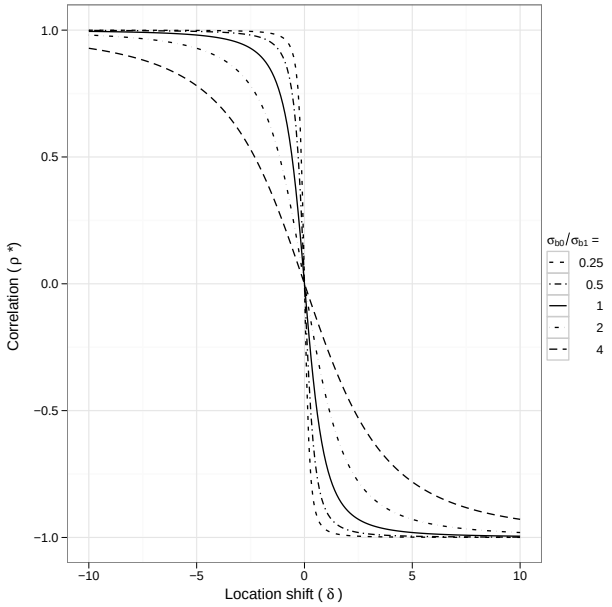

Figure 3.1: Random-effects correlation in the transformed space as a function of location shift (δ) from a setting with an initial correlation (ρ_0) of zero, for different variance component ratios

correlation will move away from the initial zero correlation. The correlation can approach the boundaries (± 1.0) after a large enough location shift for different variance component combinations. This trend can also be illustrated by taking absolute values of correlations (Fig. 3.2).

Lemma 3.4.1 also shows that the new random intercept variance in the transformed space also strictly increases with the increase in the location shift level, if the initial $G = G(\rho_0 = 0)$. It is possible that for a large amount of δ, not only the new correlation is going to unity, but also the new intercept variance $\sigma^2_{b_0^*} = \sigma^2_{b_0} + \delta^2\sigma^2_{b_1}$ is larger than the range of y_{ij}. This can cause numerical instability. If the origin of x_{ij} is too far away from the zero-correlation position(i.e., $|\delta| \gg 0$), then the numerical stability to estimate such a random-effects covariance matrix may be poor, as indicated by both the random-effects correlation and the magnitude of variance of random intercept. A good location shift should change the origin of x_{ij} such that it reduces the correlation between random-effects.

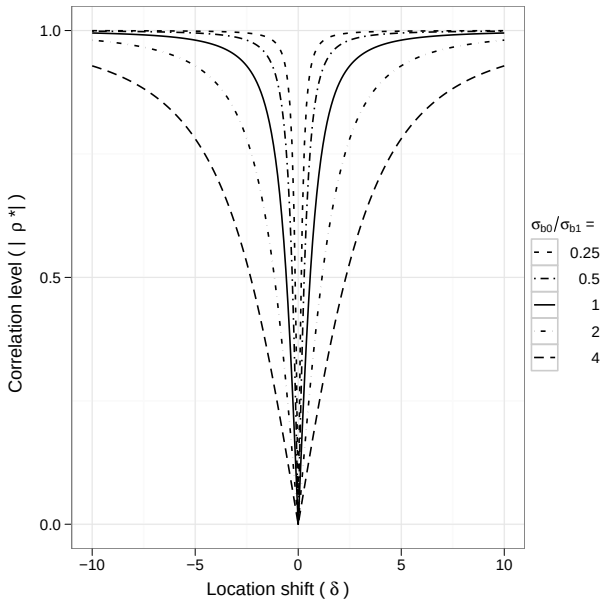

Figure 3.2: Absolute correlation level approaches limits for relatively large shifts

3.4.2 Correlation reduction

We define a neighborhood of an optimal shift d to be δ_τ, a set of location shifts around the optimal shift, i.e.,

$$\delta_\tau \in \{0 \leq |\delta - d| < |d|; \forall\ d \neq 0\}. \tag{3.4.1}$$

Theorem 3.4.2. *There exists a neighborhood of the optimal shift such that a location shift within this neighborhood will reduce the initial correlation level in the transformed space, i.e., given the initial* $\rho \neq 0$,

$$\delta \in \delta_\tau \quad \Longleftrightarrow \quad |\rho^*| < |\rho|.$$

Specifically,

1. *if the initial* $\rho = \rho_0 > 0$ *(thus* $d > 0$*), then*

$$\delta_\tau \in \{0 < \delta < 2d\} \quad \Longleftrightarrow \quad |\rho^*| < |\rho_0|;$$

2. *if the initial* $\rho = \rho_0 < 0$ *(thus* $d < 0$*), then*

$$\delta_\tau \in \{2d < \delta < 0\} \quad \Longleftrightarrow \quad |\rho^*| < |\rho_0|;$$

3. *beyond the optimal shift neighborhood, new correlation becomes larger in absolute value than the initial value,*

$$|\delta - d| > |d| \quad \Longleftrightarrow \quad |\rho^*| > |\rho_0|.$$

Proof. By Lemma 3.2.1, in general, $G^* = \begin{pmatrix} \sigma_{b_0}^2 - 2\delta\rho\sigma_{b_0}\sigma_{b_1} + \delta^2\sigma_{b_1}^2 & \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 \\ \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 & \sigma_{b_1}^2 \end{pmatrix}$.

First, it is straightforward to show that doubling the optimal shift will change the sign of correlation, by replacing general δ in above G^* with the value of $d = \rho\sigma_{b_0}/\sigma_{b_1}$. That is,

$$\delta = 2d \iff G^* = G_{2d}^* = A_{2d}^{-1} G (A_{2d}^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 & -\rho\sigma_{b_0}\sigma_{b_1} \\ -\rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix} \tag{3.4.2}$$

By reparameterizing, $A = A_\delta = \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \delta - d \\ 0 & 1 \end{pmatrix} = A_d A_{\delta - d}$. Obviously both A_d and $A_{\delta-d}$ are invertible location shift transformation matrices.

$$\begin{aligned} G^* &= Cov(b_i^*) \\ &= A^{-1} G (A^{-1})^T \\ &= (A_d A_{\delta-d})^{-1} G (A_{\delta-d}^{-1} A_d^{-1})^T \\ &= A_{\delta-d}^{-1} A_d^{-1} G (A_d^{-1})^T (A_{\delta-d}^{-1})^T \\ &= A_{\delta-d}^{-1} [A_d^{-1} G (A_d^{-1})^T] (A_{\delta-d}^{-1})^T \\ &= A_{\delta-d}^{-1} [\Omega] (A_{\delta-d}^{-1})^T \qquad \text{(by Theorem 3.2.2)} \end{aligned}$$

$\Longrightarrow$ Applying Lemma 3.4.1 to G^* relative to the diagonal matrix Ω, we have,

(i) $|\rho^*|$ monotonically increases with $|\delta - d|$;

(ii) G^* with $\rho^* = 0$, as long as no shift from Ω, e.g., $|\delta - d| = 0$ and $\delta = d$;

(iii) G^* with $|\rho^*| = |\rho_0| \neq 0$, as long as current shift, $|\delta - d| = |d| = |\rho_0\sigma_{b_0}/\sigma_{b_1}|$, thus, $\delta = 0$ or $2d$.

$\Longrightarrow$ Theorem 3.4.2 is proved after combining (i), (ii) and (iii).

□

Fig. 3.3 illustrates the existence of a neighborhood of the optimal shift in the case of the initial correlation $\rho_0 > 0$. Shifting to the right of ρ_0 but still within the neighborhood of the optimal shift will ensure the reduction of correlation level in the transformed space. For the the case with a negative initial correlation, by Lemma 3.4.1, the shifting direction will be to the left of ρ_0 (Fig. 3.4).

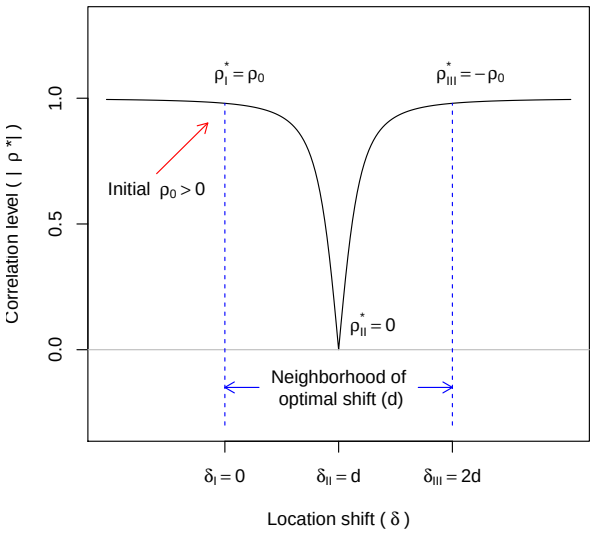

Figure 3.3: Random-effects correlation level in the transformed space, given initial positive correlation, showing reduction within the neighborhood of the optimal shift (d) (move to the right)

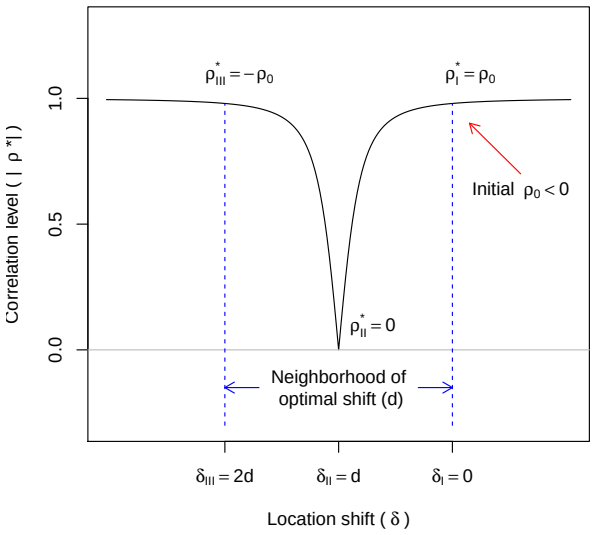

Figure 3.4: Random-effects correlation level in the transformed space, given initial negative correlation, showing reduction within the neighborhood of the optimal shift (d) (move to the left)

Note that the traditional centering will not work if the origin after centering transformation does not fall in the neighborhood of the optimal shift. For example, if the initial $\rho_0 > 0$, then $d > 0$. Centering which shifts the origin in the opposite direction will always increase ρ^* in the transformed space.

3.4.3 Condition number improvement

A common measure of numerical stability and singularity of a matrix is the condition number (CN) (Belsley and Oldford, 1986; Sengupta and Bhimasankaram, 1997; Trefethen and Bau, 1997). The CN of a matrix M is defined as the square root of the ratio of the maximal eigenvalue to the minimal eigenvalue of MM^T,

$$CN(M) = \sqrt{\frac{\lambda_{max}(MM^T)}{\lambda_{min}(MM^T)}}. \tag{3.4.3}$$

Larger condition number corresponds to less numerical stability. If the condition number is one $CN(M) = 1$, M is said to be perfectly conditioned. If a matrix M is near-singular, $CN(M)$ can be very large and M is ill-conditioned. When the condition number is huge and the inverse of the condition number of a matrix becomes comparable to computer round-off error, one can expect the computing quality will be poor for a nontrivial matrix operation on this matrix, such as inverse operation. The rates of convergence of many iterative algorithms are strongly influenced by the size of condition number (Yuan and Chan, 2008). For example, the condition number of the correlation matrix of the discrete Fourier transform vector has been directly linked to the rate of converge of algorithm (Chen et al., 2006).

3.4.3.1 Smaller condition number and larger minimal eigenvalue after AF

For a general G matrix, we can show that the condition number will become smaller after AF as long as the original population correlation is not zero. But first, this trend of CN reduction can be illustrated in Fig. 3.5 for various G matrices covering 5 levels of correlations and 5 levels of variance component ratios. Fig. 3.5 shows the scatter plots of the theoretical condition numbers in the optimal transformed space against those in the original space, stratified by various variance component ratio settings. A dashed concordance line with an intercept of zero and a slope of one is also provided to help visual comparison for each

scatter plot. More deviation below the dashed line indicates a larger reduction of condition number after AF.

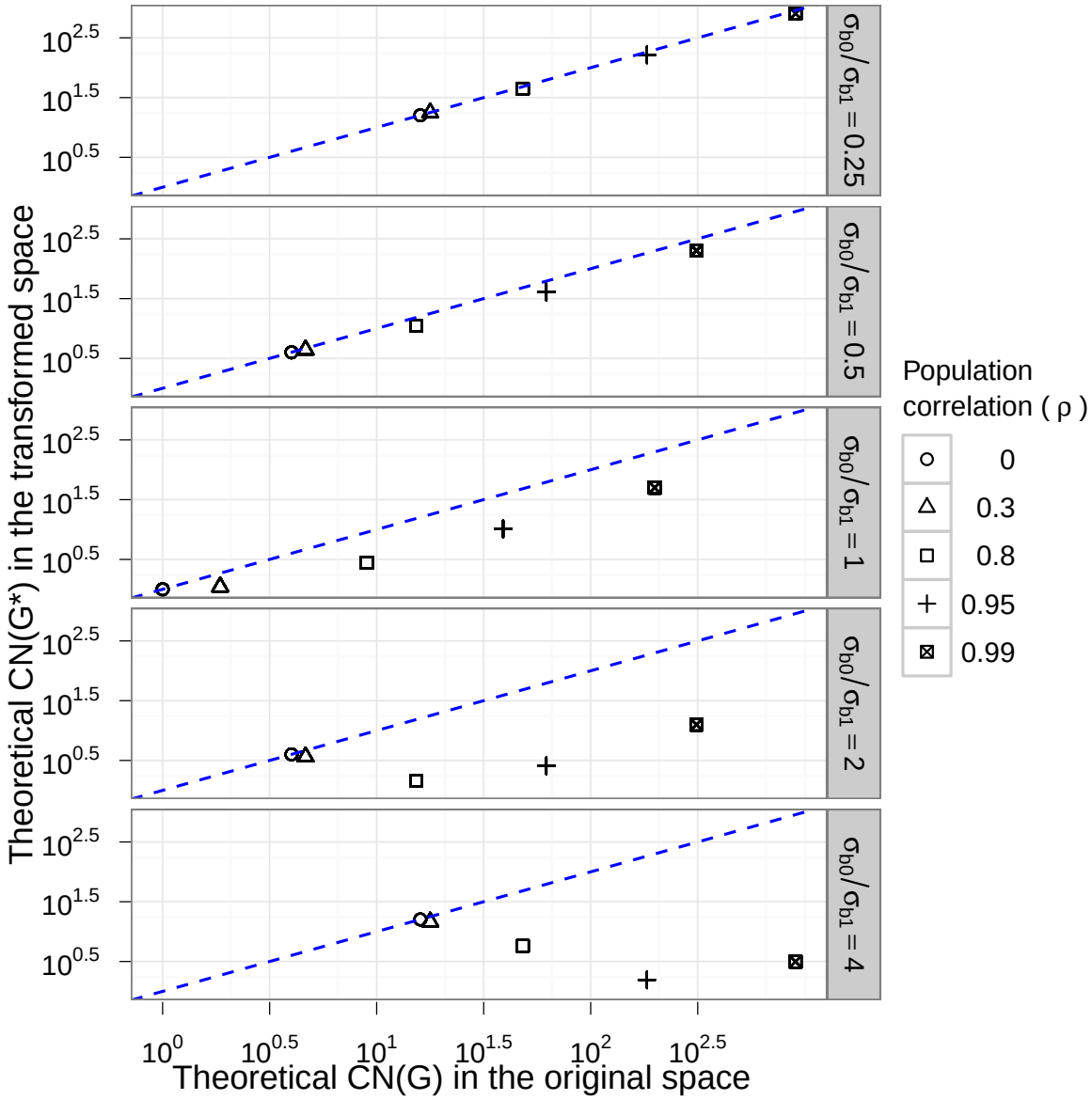

Figure 3.5: Scatter plots of theoretical condition number of random-effects covariance matrix (G) in the transformed space against that in the original space, stratified by variance component ratio, for different correlation levels

Theorem 3.4.3. *Let* $G = \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$, *where* $0 < |\rho| < 1$, *with eigenvalues* $\lambda(G)$*:* $\lambda_1 > \lambda_2 > 0$. *Let* $G^* = G_d^* = \Omega = \begin{pmatrix} \sigma_{b_0}^2(1-\rho^2) & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix}$, *with eigenvalues* $\lambda(G^*)$*:* $\lambda_1^* \geq \lambda_2^* > 0$. *Compared to the initial* G, G^* *in the transformed space has*

(i) *more clustered eigenvalues,* $\lambda_2 < \lambda_2^* \leq \lambda_1^* < \lambda_1$*;*

(ii) *smaller condition number,* $CN(G^*) < CN(G)$*.*

Obviously, by Theorem 3.2.2, the matrix $G^* = \Omega$ is obtained in the optimal transformed space, after applying the optimal location shift $A = A_d$ on the original matrix G. The proof for the first result (i) in Theorem 3.4.3 will be provided after the following Lemma 3.4.4 - 3.4.6. The result (ii) can also be derived from (i), since $CN(G^*) = \lambda_1^*/\lambda_2^*$ and $CN(G) = \lambda_1/\lambda_2$.

Lemma 3.4.4. *For a square symmetric matrix* $M_{n\times n}$*, relative to a scalar* $\alpha \neq 0$*, the condition number of* M *is invariant, i.e.,* $CN(\alpha M) = CN(M)$*; eigenvalues of* M *are not invariant and* $\lambda(\alpha M) = \alpha\lambda(M)$*.*

Proof. $M^T = M \Rightarrow$ eigenvalue $\lambda(M^T) = \lambda(M) \Rightarrow \lambda(MM^T) = [\lambda(M)]^2$. By the definition of CN in (3.4.3), for a square symmetric matrix M,

$$CN(M) = \sqrt{\frac{\lambda_{max}(MM^T)}{\lambda_{min}(MM^T)}} = \frac{|\lambda_{max}(M)|}{|\lambda_{min}(M)|}.$$

Then,

$$\begin{aligned}
\lambda = eigen(M) &\Longleftrightarrow det(M - \lambda I_n) = 0 \\
&\Longleftrightarrow \alpha^n det(M - \lambda I_n) = 0, \forall\ \alpha \neq 0 \\
&\Longleftrightarrow det(\alpha M - \alpha\lambda I_n) = 0 \\
&\Longleftrightarrow eigen(\alpha M) = \alpha\lambda
\end{aligned}$$

Also,

$$CN(\alpha M) = \frac{|\lambda_{max}(\alpha M)|}{|\lambda_{min}(\alpha M)|} = \frac{|\alpha\lambda_{max}(M)|}{|\alpha\lambda_{min}(M)|} = CN(M).$$

□

Lemma 3.4.5. *For a positive definite matrix* $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$*, where* $\sigma > 0$ *and* $0 < |\rho| < 1$*, denote diagonal elements* $G_{11} = \sigma^2$*,* $G_{22} = 1$*, and eigenvalues* $\lambda(G)$*,* $\lambda_1 > \lambda_2 > 0$*. Then the two eigenvalues are less clustered than the two diagonal elements, i.e.,*

$$\lambda_1 > G_{22} > \lambda_2 > 0,\ and\ \lambda_1 > G_{11} > \lambda_2 > 0, \quad for\ \forall\ \sigma > 0,\ \forall\ 0 < |\rho| < 1.$$

Proof. The characteristic equation is $det(G - \lambda I_2) = 0$, then

$$\lambda^2 - (\sigma^2 + 1)\lambda + \sigma^2(1 - \rho^2) = 0,$$

then,

$$\triangle = (\sigma^2 + 1)^2 - 4\sigma^2(1 - \rho^2) = (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 \geq 0. \tag{3.4.4}$$

Thus,

$$\lambda_1 = \frac{\sigma^2 + 1 + \sqrt{\triangle}}{2}, \lambda_2 = \frac{\sigma^2 + 1 - \sqrt{\triangle}}{2}. \tag{3.4.5}$$

Note that

$$\lambda_1 = \lambda_2 \Leftrightarrow \triangle = 0 \Leftrightarrow \sigma = 1 \; and \; \rho = 0 \Leftrightarrow G = I_2 \Leftrightarrow \lambda_1 = \lambda_2 = 1 = G_{22}.$$

It means that non-zero correlation will result in unequal eigenvalues and that neither of the two eigenvalues will be equal to a diagonal element G_{22},

$$\rho \neq 0 \Leftrightarrow G_{22} \neq \lambda_1 \neq \lambda_2 \neq G_{22}.$$

Thus, Lemma 3.4.5 is obviously true if $\sigma = 1$.

If $\sigma > 1$, then $\lambda_1 > 1 = G_{22}$, by (3.4.5). If $\sigma < 1$,

$$\begin{aligned}
&\lambda_1 = \frac{\sigma^2 + 1 + \sqrt{\triangle}}{2} > G_{22} = 1; \; \forall \; 0 < \sigma < 1 \\
&\Longleftrightarrow \sqrt{\triangle} > 1 - \sigma^2 \\
&\Longleftrightarrow \sqrt{\triangle} > 1 - \sigma^2 \geq 0 \\
&\Longleftrightarrow \triangle > (1 - \sigma^2)^2 \\
&\Longleftrightarrow (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 > (1 - \sigma^2)^2 \\
&\Longleftrightarrow 4\rho^2\sigma^2 > 0; \; \forall \; \rho \neq 0, \; \forall \; \sigma \neq 0.
\end{aligned}$$

If $\sigma < 1$, then $\lambda_2 < 1 = G_{22}$, by (3.4.5). If $\sigma > 1$,

$$\begin{aligned}
&\lambda_2 = \frac{\sigma^2 + 1 - \sqrt{\triangle}}{2} < G_{22} = 1;\ \forall\ \sigma > 1 \\
&\iff \sqrt{\triangle} > \sigma^2 - 1 \\
&\iff \sqrt{\triangle} > \sigma^2 - 1 \geq 0 \\
&\iff \triangle > (\sigma^2 - 1)^2 \\
&\iff (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 > (\sigma^2 - 1)^2 \\
&\iff 4\rho^2\sigma^2 > 0;\ \forall\ \rho \neq 0,\ \forall\ \sigma \neq 0.
\end{aligned}$$

Similarly, we can prove $\lambda_1 > G_{11} > \lambda_2$.

□

Based on above Lemma 3.4.4 and Lemma 3.4.5, we have

Lemma 3.4.6. *For a positive definite matrix* $G = \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$, *where* $\sigma_{b_0} > 0$, $\sigma_{b_1} > 0$ *and* $0 < |\rho| < 1$, *denote two eigenvalues* $\lambda(G)$, $\lambda_1 \geq \lambda_2 > 0$. *Then the two eigenvalues are more separated than the two diagonal elements, i.e., for* $\forall\ 0 < |\rho| < 1$,

$$\lambda_1 > \sigma_{b_1}^2 > \lambda_2 > 0,\ \textit{and}\ \lambda_1 > \sigma_{b_0}^2 > \lambda_2 > 0$$

$$\iff \lambda_1 > max(\sigma_{b_0}^2, \sigma_{b_1}^2)\ \&\ \lambda_2 < min(\sigma_{b_0}^2, \sigma_{b_1}^2).$$

Proof. (of Theorem 3.4.3 (i))

Recall $\lambda_1^* = \max(\sigma_{b_0}^2(1-\rho^2), \sigma_{b_1}^2)$ and $\lambda_2^* = \min(\sigma_{b_0}^2(1-\rho^2), \sigma_{b_1}^2)$ for diagonal G^*.

If $\sigma_{b_1}^2 \leq \sigma_{b_0}^2(1-\rho^2)$, then

$$\begin{aligned}
&\sigma_{b_1}^2 \leq \sigma_{b_0}^2(1-\rho^2);\ 0 < |\rho| < 1 \\
&\iff \lambda_2^* = \sigma_{b_1}^2 \leq \sigma_{b_0}^2(1-\rho^2) = \lambda_1^* \\
&\iff \sigma_{b_1}^2 = \lambda_2^* \leq \lambda_1^* = \sigma_{b_0}^2(1-\rho^2) < \sigma_{b_0}^2 \\
&\implies \lambda_2 < \sigma_{b_1}^2 = \lambda_2^* \leq \lambda_1^* < \sigma_{b_0}^2 < \lambda_1 \quad \text{(by Lemma 3.4.6)} \\
&\iff \lambda_2 < \lambda_2^* \leq \lambda_1^* < \lambda_1.
\end{aligned}$$

If $\sigma_{b_1}^2 > \sigma_{b_0}^2(1-\rho^2)$, then

$$
\begin{aligned}
&\sigma_{b_0}^2(1-\rho^2) < \sigma_{b_1}^2;\ 0 < \lambda_2^* < \lambda_1^* \\
&\iff \lambda_2^* = \sigma_{b_0}^2(1-\rho^2) < \sigma_{b_1}^2 = \lambda_1^* \\
&\iff \lambda_2^* < \lambda_1^* = \sigma_{b_1}^2 \\
&\implies \lambda_2^* < \lambda_1^* = \sigma_{b_1}^2 < \lambda_1 && \text{(by Lemma 3.4.6)} \\
&\iff \lambda_2^* < \lambda_1^* < \lambda_1 && \text{(i)} \\
&\implies \lambda_1^* < \lambda_1 \\
&\iff \lambda_1^*\lambda_2 < \lambda_1\lambda_2 \\
&\iff \lambda_1^*\lambda_2 < \lambda_1\lambda_2 = \lambda_1^*\lambda_2^*, && (\det(G) = \det(G^*)) \\
&\iff \lambda_2 < \lambda_2^* && \text{(ii)} \\
&\implies \lambda_2 < \lambda_2^* < \lambda_1^* < \lambda_1 && \text{(i \& ii).}
\end{aligned}
$$

□

3.4.3.2 Two properties of CN reduction after AF

To measure the extent of CN reduction, we define the condition number ratio as

$$CNR = CNR[G] \triangleq \frac{CN(G)}{CN(G^*)}. \tag{3.4.6}$$

If the CN is reduced in the transformed space, the CNR will be greater than one.

The CN reduction after AF are further illustrated by Fig. 3.6 and Fig. 3.7. Note that the scales of the vertical axes are different across panels for Fig. 3.6 and Fig. 3.7. Both figures shows that all CNRs are greater than one (above the dashed line) if the random-effects correlation in the original space is not zero. Given the relative size of two random-effect variances, Fig. 3.6 demonstrates that a higher CNR can be achieved for a higher level of correlation. The magnitude of CNR tends to increase as the variance component ratio increases, especially when the variance component ratio is larger than one, namely, the random slope has smaller variance relative to random intercept. This observation will be formally supported by the following Theorem 3.4.7. On the other hand, given the level of correlation, Fig. 3.7 shows that the magnitude of CNR still generally increases across panels as the level of correlation increases, while there may not be a monotonic relationship between

the CNR and the variance component ratio when the ratio is larger than one. However, the CNR at a variance component ratio larger than one seems generally larger than that obtained at the inverse counterpart of the ratio (e.g., $\sigma_{b_0}/\sigma_{b_1} = 2$ vs. $\sigma_{b_0}/\sigma_{b_1} = 0.5$). This conjecture will be confirmed by Theorem 3.4.9.

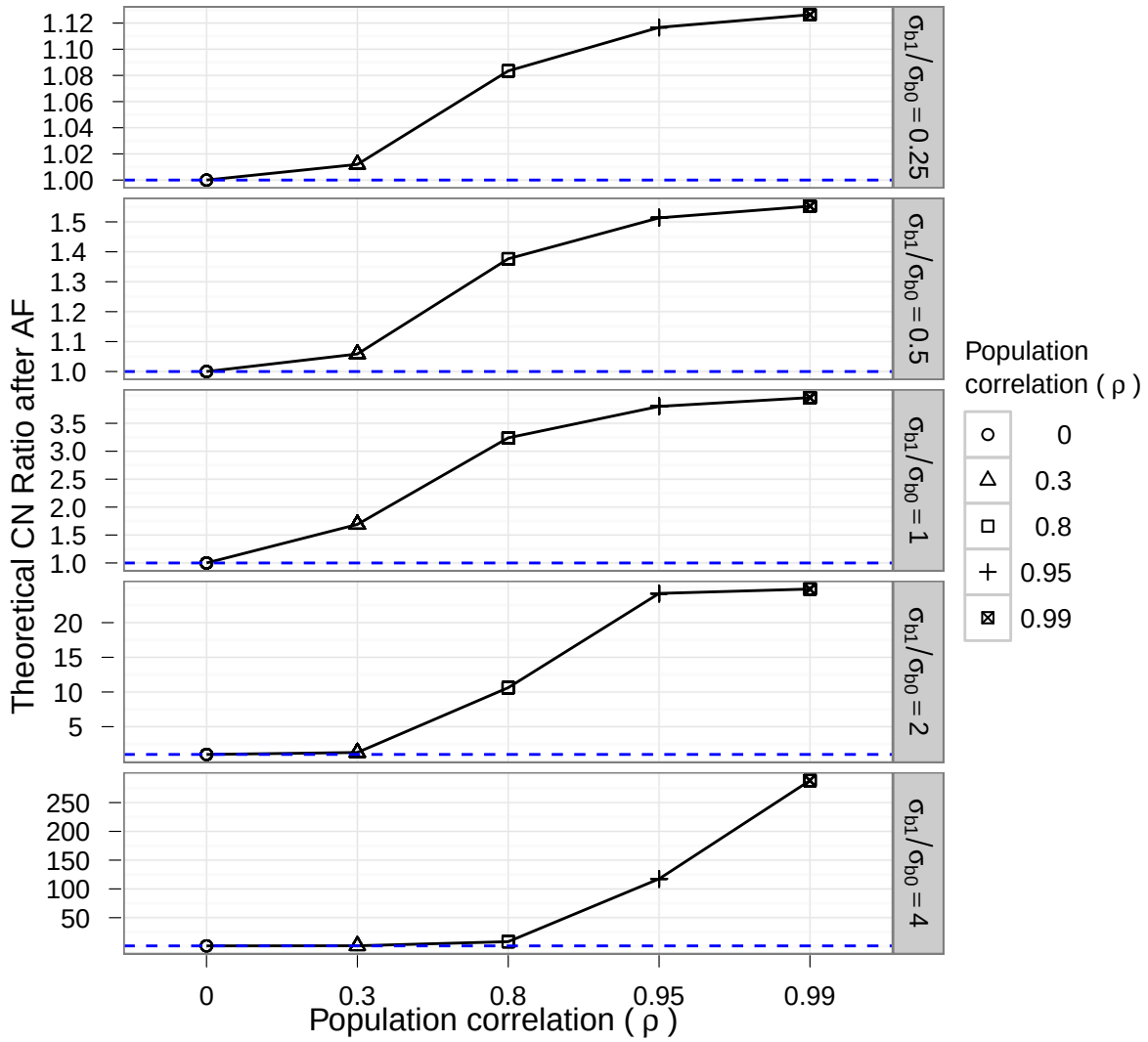

Figure 3.6: Scatter plots of theoretical condition number ratio after AF against different correlation levels, stratified by variance component ratio

- **Property one**

Theorem 3.4.7. *Larger CN reduction will be obtained for a setting with higher initial*

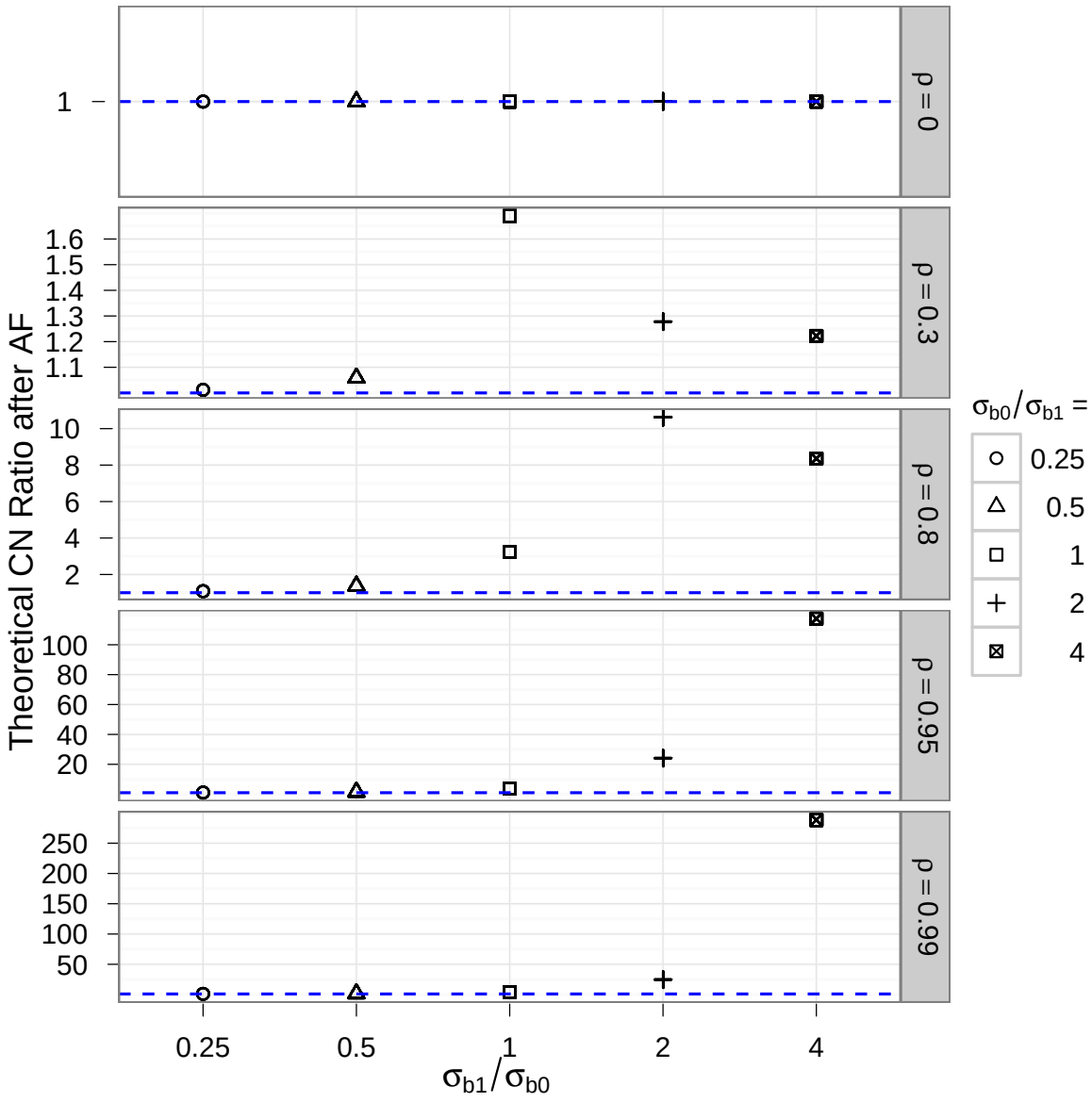

Figure 3.7: Scatter plots of theoretical condition number ratio after AF against variance component ratios, stratified by correlation level

correlation, given the same initial variance component ratio, i.e., if $|\rho_1| > |\rho_2|$*, then*

$$CNR\left[\begin{pmatrix} \sigma_{b_0}^2 & \rho_1\sigma_{b_0}\sigma_{b_1} \\ \rho_1\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}\right] > CNR\left[\begin{pmatrix} \sigma_{b_0}^2 & \rho_2\sigma_{b_0}\sigma_{b_1} \\ \rho_2\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}\right].$$

Theorem 3.4.7 will be proved after the following Lemma 3.4.8.

Lemma 3.4.8. *For a positive definite matrix* $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$*, the condition number ratio* $CNR[G]$ *is either the square of the largest eigenvalue of G, or the inverse of the square of the smallest eigenvalue of G, i.e.,*

$$CNR[G] \triangleq \frac{CN(G)}{CN(G^*)} = \begin{cases} [\lambda_{max}(G)]^2, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \frac{1}{[\lambda_{min}(G)]^2}, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases}$$

Proof. Denote the eigenvalues of the matrix G to be $\lambda(G)$: $\lambda_1 > \lambda_2 > 0$, and the eigenvalues of the corresponding matrix $G^* = \begin{pmatrix} \sigma^2(1-\rho^2) & 0 \\ 0 & 1 \end{pmatrix}$ in the optimal transformed space to be $\lambda(G^*)$: $\lambda_1^* \geq \lambda_2^* > 0$. By definition,

$$\begin{cases} \lambda_1^* = 1, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \lambda_2^* = 1, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases} \tag{3.4.7}$$

By definition, $det(G^*) = det(G)$, i.e., $\lambda_1\lambda_2 = \lambda_1^*\lambda_2^*$.

Therefore,

$$CNR[G] \triangleq \frac{CN(G)}{CN(G^*)} = \frac{\lambda_1/\lambda_2}{\lambda_1^*/\lambda_2^*} = (\frac{\lambda_1}{\lambda_1^*})^2 = (\frac{\lambda_2^*}{\lambda_2})^2 \tag{3.4.8}$$

Thus, Lemma 3.4.8 is proved after combining equation 3.4.8 with 3.4.7. Combining Lemma 3.4.8 with Lemma 3.4.4 and Lemma 3.4.5, we also have $CNR[G] > 1$ for a general covariance matrix G.

□

Proof. (of Theorem 3.4.7)

Given matrix $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ with eigenvalues: $\lambda_1 > \lambda_2 > 0$, by Lemma 3.4.5, we have

$$\lambda_1 = \frac{\sigma^2 + 1 + \sqrt{\triangle}}{2}, \lambda_2 = \frac{\sigma^2 + 1 - \sqrt{\triangle}}{2}, \text{ where } \triangle = (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 \geq 0.$$

For a fixed σ, if $|\rho|$ increases, then $\triangle$ strictly increases, and therefore λ_1 strictly increases while λ_2 strictly decreases. By Lemma 3.4.8, $CNR[G]$ will increase with the change in either λ_1 or λ_2. Thus, given $|\rho_1| > |\rho_2|$, we have

$$CNR\left[\begin{pmatrix} \sigma^2 & \rho_1\sigma \\ \rho_1\sigma & 1 \end{pmatrix}\right] > CNR\left[\begin{pmatrix} \sigma^2 & \rho_2\sigma \\ \rho_2\sigma & 1 \end{pmatrix}\right].$$

Together with Lemma 3.4.4, Theorem 3.4.7 is thus proved.

□

• **Property two**

Theorem 3.4.9. *Let two settings have the same initial correlation but reversed variance component ratios. The setting with larger variance component ratio will have larger CN reduction after an optimal transformation, i.e., given $\rho \neq 0$ and $\sigma = \sigma_{b_0}/\sigma_{b_1} > 1$, we have*

$$CNR\left[\begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right] > CNR\left[\begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right].$$

Proof. Denote covariance matrix $G_\sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ with eigenvalues: $\lambda_1 > \lambda_2 > 0$. Denote covariance matrix $G_{1/\sigma} = \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}$. Then $CN(G_{1/\sigma}) = CN(G_\sigma) = \lambda_1/\lambda_2$ by Lemma 3.4.5. We also have the corresponding new matrix $G^*_{1/\sigma} = \begin{pmatrix} 1-\rho^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$ in the optimal transformed space. Obviously, $CN(G^*_{1/\sigma}) = \sigma^2/(1-\rho^2)$, given $\sigma > 1$. Thus

$$CNR[G_{1/\sigma}] \triangleq \frac{CN(G_{1/\sigma})}{CN(G^*_{1/\sigma})} = \frac{\lambda_1/\lambda_2}{\lambda^*_1/\lambda^*_2} = \frac{\lambda_1/\lambda_2}{\sigma^2/(1-\rho^2)} \tag{3.4.9}$$

By Lemma 3.4.8,

$$CNR[G_\sigma] = \begin{cases} [\lambda_1]^2, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \frac{1}{[\lambda_2]^2}, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases} \tag{3.4.10}$$

Taking the ratio and noting that $\lambda_1\lambda_2 = \sigma^2(1-\rho^2)$, we have

$$\frac{CNR[G_\sigma]}{CNR[G_{1/\sigma}]} = \begin{cases} \lambda_1\lambda_2 \times \frac{\sigma^2}{1-\rho^2} = \sigma^4 > 1, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \frac{1}{\lambda_1\lambda_2} \times \frac{\sigma^2}{1-\rho^2} = \frac{1}{(1-\rho^2)^2} > 1, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases} \tag{3.4.11}$$

Thus $CNR[G_\sigma] > CNR[G_{1/\sigma}]$ is always true, regardless of the relative size of σ^2 and $\frac{1}{1-\rho^2}$. By (3.4.11), we have

$$\frac{CNR[G_\sigma]}{CNR[G_{1/\sigma}]} = min(\sigma^4, \frac{1}{(1-\rho^2)^2}), \text{ given } \sigma > 1.$$

□

Thus, a large reduction of CN can be expected under challenging data fitting situations, e.g., with extreme correlation and/or large variance component ratio. Such challenging scenarios also produce a wide neighborhood of the optimal shift, where the proposed AF algorithm is expected to perform well.

Chapter 4

Simulation Study

This chapter examines the performance of the proposed AF algorithm for RIS models in 405 simulation settings.

4.1 Convergence performance measures

We used non-convergence rate to measure the numerical convergence performance of the AF algorithm. The non-convergence rate was defined as the observed proportion of non-convergent runs, i.e., the number of failed runs divided by the total number runs for a simulation setting. Although a statistical software package may not signal any error messages related to the convergence status, a nominally convergent run may produce a non-positive definite covariance matrix estimate. The non-PD here was defined for random-effects covariance matrix rather than for the covariance matrix of outcome variable (Browne and Draper, 2000; Pryseley et al., 2011). As long as the estimated covariance matrix G had a negative eigenvalue, it was a non-PD run. Due to the skewed distribution of observed CNs, we used the geometric mean instead of arithmetic mean to measure the condition number estimate for a simulation setting.

4.2 Softwares

The simulated data sets were fitted by routine *lme*(*nlme*) (Pinheiro et al., 2009) in the *R* environment using REML estimation method. By package default, the fitting function first ran 25 steps of EM iterations to provide initial values before entering Newton-Raphson iterations. The intermediate fitting results were always available at any iteration step by setting option $returnObject = TRUE$ in the *lmeControl* list of the *lme* routine, no matter whether the convergence was obtained or not. To simplify the convergence diagnosis, two default options in *lmeControl* list, the maximum number of Newton-Raphson iterations ($msMaxIter$, default 50) and the maximum number of evaluations of the objective function permitted for nlminb ($msMaxEval$, default 200), were also both increased to 500. As a result, the relevant warning messages for these two options did not occur in our simulations. The same *lme* routine was used both before and after AF.

All simulations and analyses were conducted on a 2.13 GHz Intel Core (TM) 2 CPU 6400 processor on the Windows XP Professional (version 2002) platform.

4.3 Simulation settings

We fitted a RIS model to series of balanced datasets generated from the following LMM.

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + b_{0i} + x_{ij}b_{1i} + \varepsilon_{ij}, \qquad i = 1, \cdots, m; \qquad j = 1, \cdots, n, \tag{4.3.1}$$

where y_{ij} being scalar, the jth observations within ith group; fixed-effects $\beta_0 = -1$ and $\beta_1 = 1$; both predictor variable x_{ij} and residual ε_{ij} were sampled from standard normal distribution, $N(0, 1)$.

Sample sizes ($N = m \times n$) varied from 125 to 2000, with number of groups $m \in$ (25, 50, 100) and group size $n \in$ (5, 10, 20), a balanced design. Each random effect had three levels, σ_{b0} (intercept) or σ_{b1} (slope) $\in$ (1, 1/2, 1/4). The correlation between random-effects covered five levels, $\rho \in$ (0.99, 0.95, 0.80, 0.30, 0.00), indicating extremely high, very high, high, moderate and zero correlation, respectively. Thus, there were a total of 405 scenarios (9 sample size settings $\times$ 9 variance combinations $\times$ 5 correlation settings).

The number of replication runs per simulation design scenario was set at 1000. For

Table 4.1: Summary of convergence improvements after adaptive fitting (AF)

Estimation Space	# of AF	# (%) of scenarios** with failed runs	Total # of failed runs***	non-convergence rate per scenario mean (range)
Untransformed	0	324 (80.0%)	69,787	17.23% (0.0%, 60.3%)
Transformed	1	52 (12.8%)	165	0.041% (0.0%, 1.3%)
Transformed	2	2 (0.5%)	2	0.00049% (0.0%, 0.1%)

* Total 405 scenarios were simulated and each scenario with 1000 replication runs
** Scenarios with at least one non-converged run
*** Out of all 405 X 1000 runs

each simulated dataset, the RIS model was fitted both in the original space without AF and in the transformed space with AF.

4.4 Simulation results

4.4.1 Non-convergence rate

The AF procedure was found to improve the non-convergence rate significantly (Table 4.1). Across 405 scenarios, the average non-convergence rate was as high as 17.23% before AF, but was reduced to close to zero level (0.00049%) in the transformed space after AF. Correspondingly, the percentage of scenarios with non-convergence issues was 80.0% and 0.5%, before and after AF, respectively. The number of non-convergent runs out of 405,000 simulated runs dropped from 69,787 to 165 and 2, before AF, after single and two AF steps, respectively. The following discussions about the performance of AF algorithm are all based on the fitting results after two AF steps.

Before AF, the average non-convergence rate increased from 3.88%, 4.62%, 13.47%, 27.14% to 37.05% as the population correlation level increased from $\rho = 0.00$, 0.30, 0.80, 0.95 to 0.99, respectively. However, the population correlation level had little impact on the efficiency of AF or the number of AF steps needed. After the first AF step, there were 165 non-convergent runs, evenly distributed in 52 scenarios, with 12, 12, 11, 10, and 7 scenarios where population $\rho = 0.00$, 0.30, 0.80, 0.95 and 0.99, respectively. After the second AF step, there were two failed runs, one from zero and the other from moderate correlation (0.30) scenarios. Therefore, the strong impact of high correlation on non-convergence in the original space before AF disappeared in the optimal transformed space. The AF algorithm

was generally applicable and especially effective when the correlation was extremely high.

Small random slope variance ($\sigma^2_{b_1}$) also had a significant impact on the non-convergence rate before AF (Fig. 4.1 and 4.2). When the random slope variance was reduced (Fig. 4.2, from left panels to right panels), the non-convergence rate increased across all three levels of random intercept variance ($\sigma^2_{b_0}$). Controlling random slope at the same level (e.g., $\sigma_{b_1} = 0.25$), however, when the random intercept variance became larger (Fig. 4.2, from top to bottom), the non-convergence rate was reduced slightly. Fig. 4.2 also confirms that small random slope variance had a larger impact on the non-convergence rate than small random intercept variance across various sample size settings. On the other hand, the two factors of sample size, the number of groups (m) and group size (n), seemed to have comparable impact on the non-convergence rate. Fig. 4.3 shows the non-convergence rates for all simulated scenarios.

The non-convergence issue could become more severe as the variances combination decreased, namely, relative noise increased. For the 45 scenarios with the highest level of variance component considered ($\sigma_{b_0} = \sigma_{b_1} = 1$), only 25 scenarios had all 1000 converged runs, with a mean (range) non-convergence rate of 5.06% (0.0%, 35.9%) across the 45 scenarios. On the other hand, for the 45 scenarios with the smallest level of variance component examined ($\sigma_{b_0} = \sigma_{b_1} = 0.25$), only 2 scenarios had a zero non-convergence rate, with a mean (range) non-convergence rate of 32.56% (0.0%, 60.3%) across the 45 scenarios.

It is not surprising that the non-convergence rate could be improved by increasing sample size. For the 45 scenarios with the smallest number of observations ($m \times n = 25 \times 5 = 125$), none of the scenarios had a zero non-convergence rate before AF, with a mean (range) non-convergence rate of 30.06% (1.0%, 57.7%) across the 45 scenarios. On the other hand, for the 45 scenarios with the maximum number of observations ($100 \times 20 = 2000$), only 22 scenarios had a zero non-convergence rate before AF, with a mean (range) non-convergence rate of 0.81% (0.0%, 44.9%) across the 45 scenarios.

4.4.2 Non-positive definite

Besides the non-convergence rate, the utility of AF could also be illustrated by the improvement in the fraction of positive definite estimates of random-effects covariance matrix (Table 4.2). Across all 405 scenarios, there was only one run with the non-PD issue after

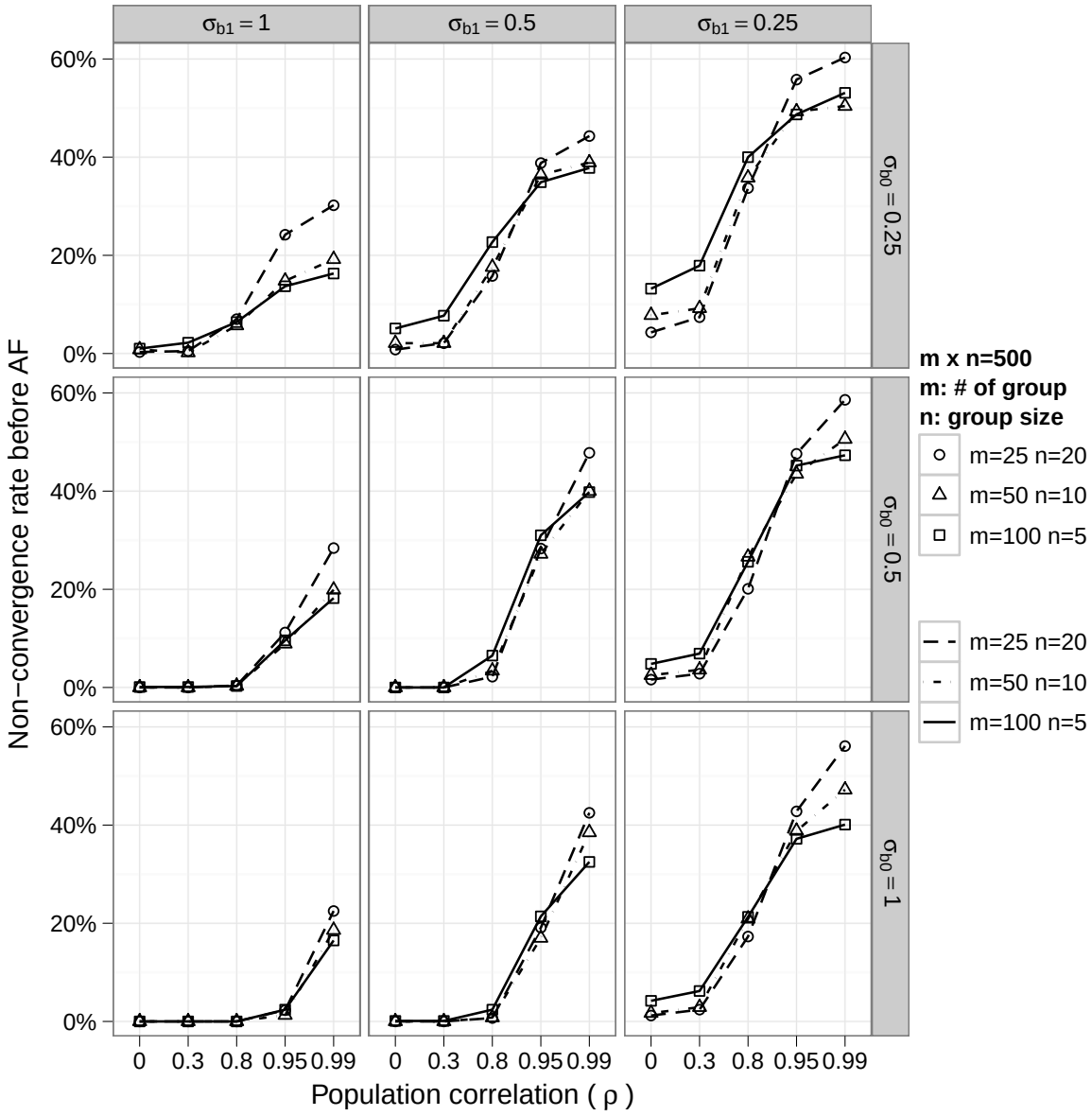

Figure 4.1: The non-convergence rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine variance component combinations, controlling the total number of observations for three sample size combinations to 500

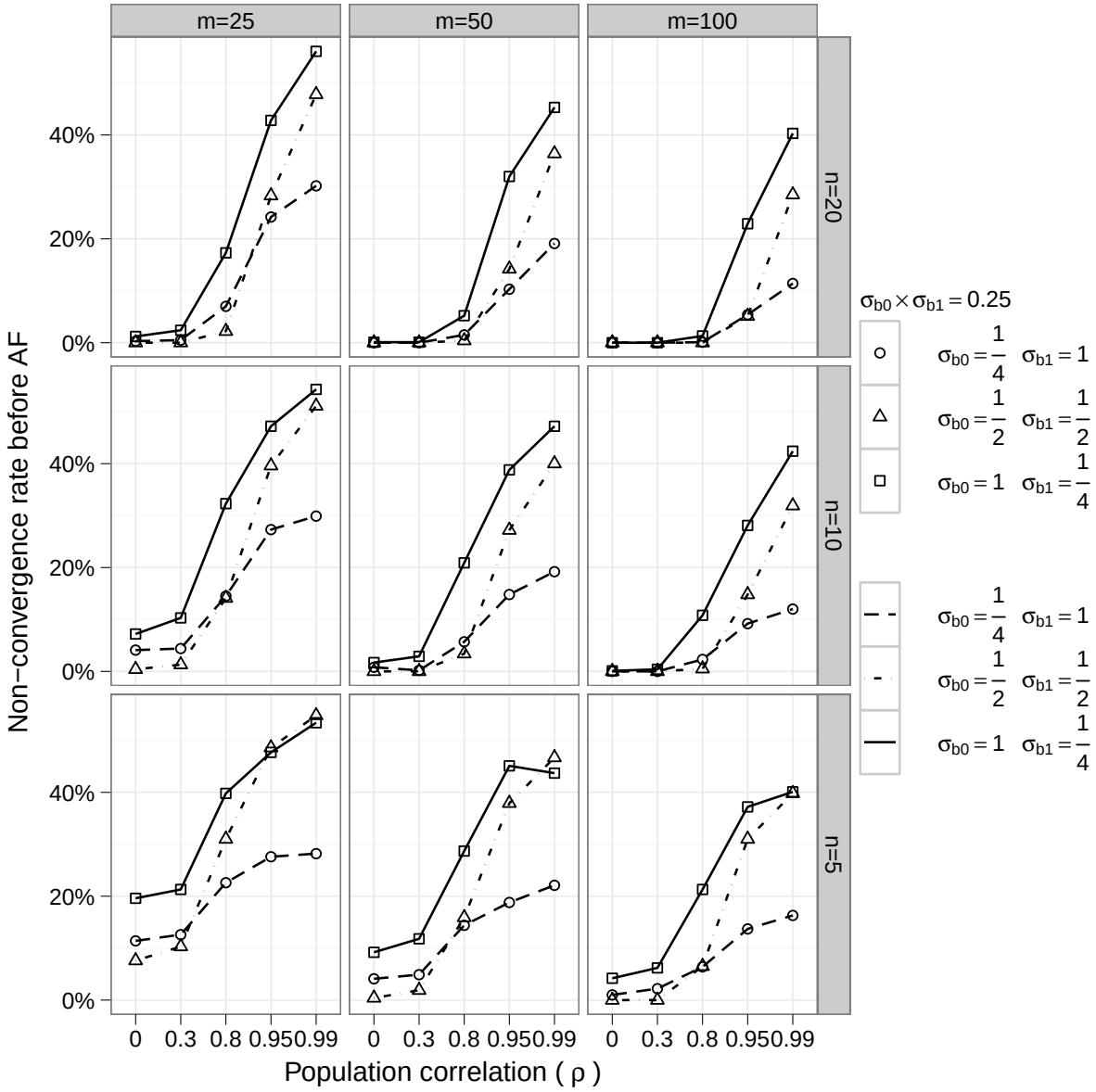

Figure 4.2: The non-convergence rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine sample size combinations, controlling the product of two variance components for three variance component combinations to 0.25

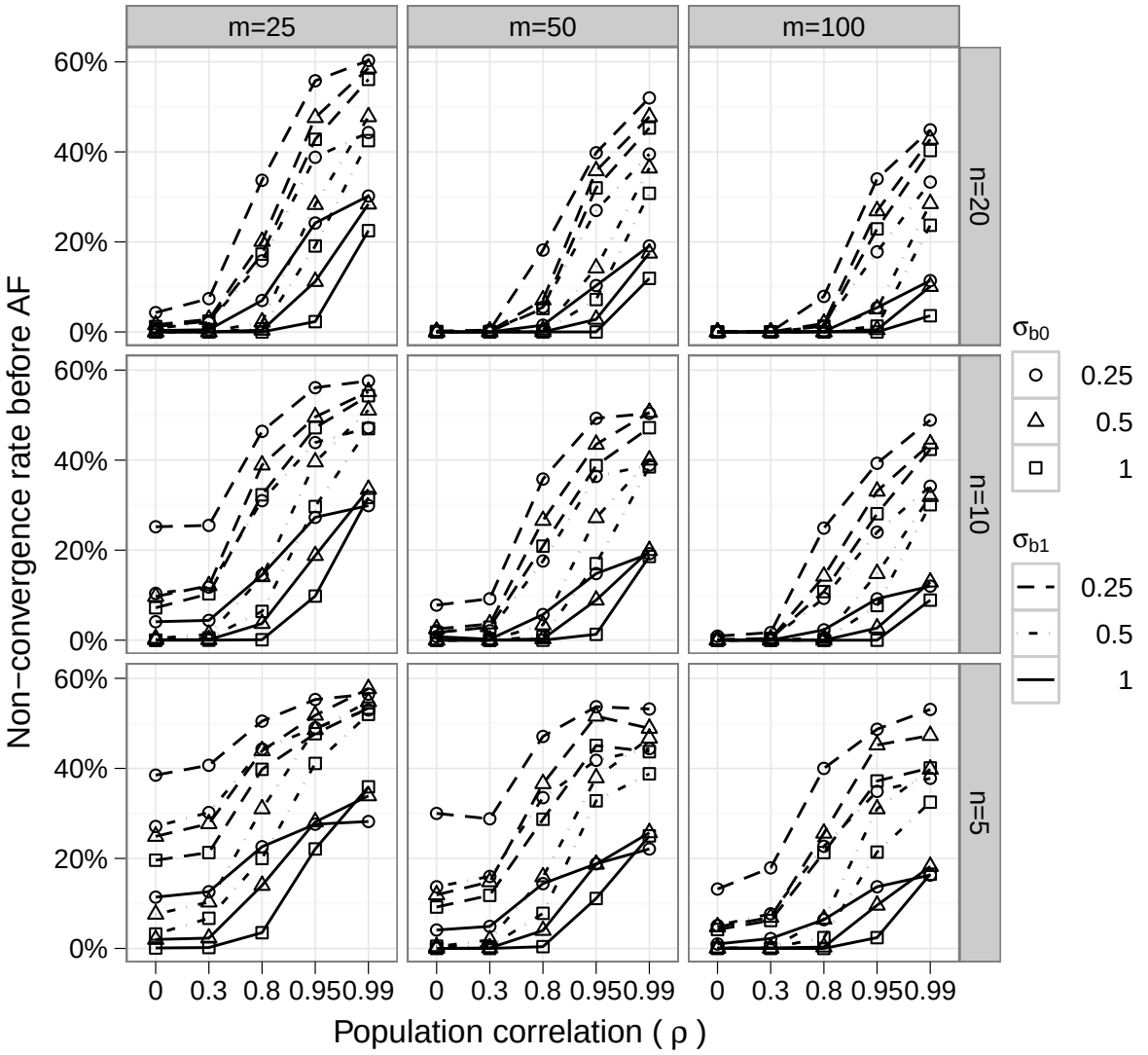

Figure 4.3: The non-convergence rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine sample size combinations, and with each of the nine curves of a panel corresponding to one of the nine variance component combinations

AF. However, before AF, there were 48,278 (11.92%) runs with non-PD issues, which were distributed in 321 of the 405 scenarios. Importantly, the 11.92% non-PD runs was composed of 8.56% non-convergent and 3.36% "nominally convergent" runs. In other words, there were on average more than 33 "implicit" non-PD runs per scenario, which were considered as convergent runs without signaling any error message by the software package. Thus, although the average nominally convergent rate was 82.77% before AF (see Section 4.4.1), there was only 79.41% runs which were not only nominally convergent, but also had PD covariance matrix estimates across all 405 scenarios. If stratified by the convergence status, the conditional non-PD rate per run was much higher within failed runs than that within converged runs (49.67% vs. 4.06%). Fig. 4.4 plots the conditional non-PD rate given a nominally convergent run as a function of population correlation level. It seemed that a large sample size (e.g., $2000 = 100 \times 20$) might not alleviate the non-PD issue, especially for the settings with small variance component for random intercept relative to random slope (e.g., $\sigma_{b_0} = 0.25, \sigma_{b_1} = 1$). Therefore, non-PD issue might still be a numerical issue even for a nominally convergent run before AF but not in the optimal transformed space.

4.4.3 Change in correlation

The above overall improvement in the non-convergence rate and positive definite property could be further understood by observing the change in estimated random-effects correlation before and after AF (Table 4.2). On average, the near-zero correlation ($|\widehat{\rho}| < 0.10$) rate improved from 9.59% before AF to 85.56% after AF. More importantly, extreme correlation ($|\widehat{\rho}| \geq 0.99$) almost disappeared in the optimal transformed space. The average extreme correlation rate dropped from 26.4% before AF down to 0.0042% after AF. Our simulations show that the correlation estimate had a 99.3% empirical rate to become smaller after AF.

The near-zero correlation after AF could also be illustrated by plotting the observed correlation estimate $\widehat{\rho^*}$ in the transformed space against the nominal non-convergence rate before AF, and against the original population correlation setting (Fig. 4.5 and 4.6, respectively). The $\widehat{\rho^*}$ scattered around zero and had a tight range of (-0.104, 0.166) across all simulated 405 scenarios. The magnitude of $\widehat{\rho^*}$ might slightly increase with the increase in the non-convergence rate or the increase in the original population correlation level. The value of $\widehat{\rho^*}$ might have larger variability for higher correlation settings, but overall the AF procedure performed very well even for extremely high correlation. For those 81 scenarios

Table 4.2: Properties of estimated random-effects covariance matrix in the original and optimal transformed spaces

Properties	Estimation space	
	Original	Optimal transformed
Non-positive definite (PD)		
I. Overall non-PD**		
# of scenarios involved***	321	1
Total # runs involved****	48,278	1
Rate per scenario		
range	(0.0%, 34.8%)	(0.0%, 0.1%)
mean	11.92%	0.00025%
II. "Nominally converged" but non-PD		
# of scenarios involved***	303	0
Total # runs involved****	13,614	0
Rate per scenario		
range	(0.0%, 19.8%)	0.0%
mean	3.36%	0.00%
Random-effects correlation		
III. Observed |rho|>=0.99		
# of scenarios involved***	334	13
Total # runs involved****	106,520	17
Rate per scenario		
range	(0.0%, 66.4%)	(0.0%, 0.2%)
mean	26.36%	0.0042%
IV. Observed |rho|<0.10		
# of scenarios involved***	240	405
Total # runs involved****	38,840	346,522
Rate per scenario		
range	(0.0%, 64.2%)	(45.0%, 100.0%)
mean	9.59%	85.56%

* Observed values at last iteration

** A non-PD run can be either a non-converged run or a "nominally converged" run

*** Scenarios with at least one run assoicated with that specified event

**** Out of all 405 X 1000 runs

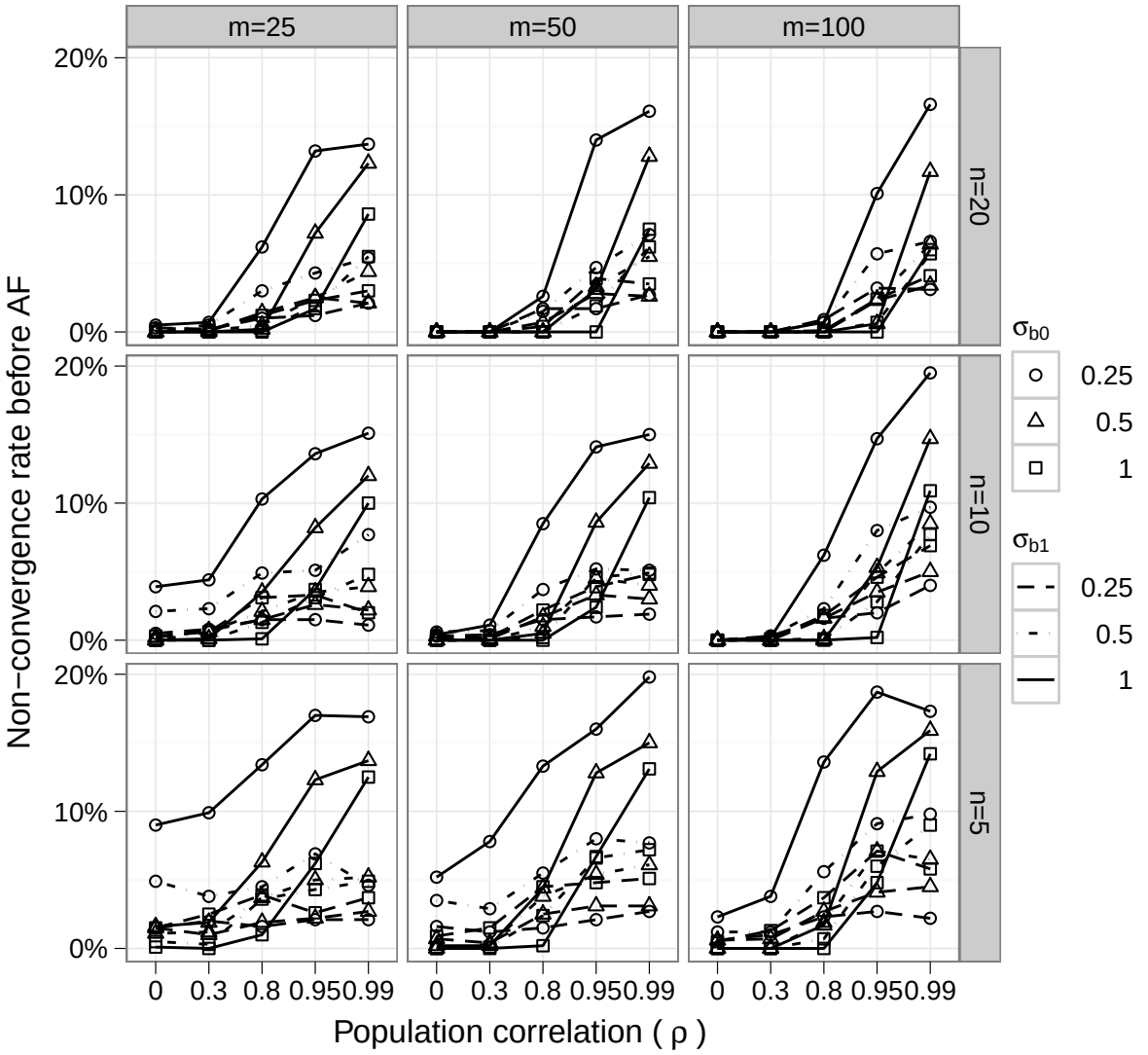

Figure 4.4: The conditional non-PD rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine sample size combinations, and with each of the nine curves of a panel corresponding to one of the nine variance component combinations

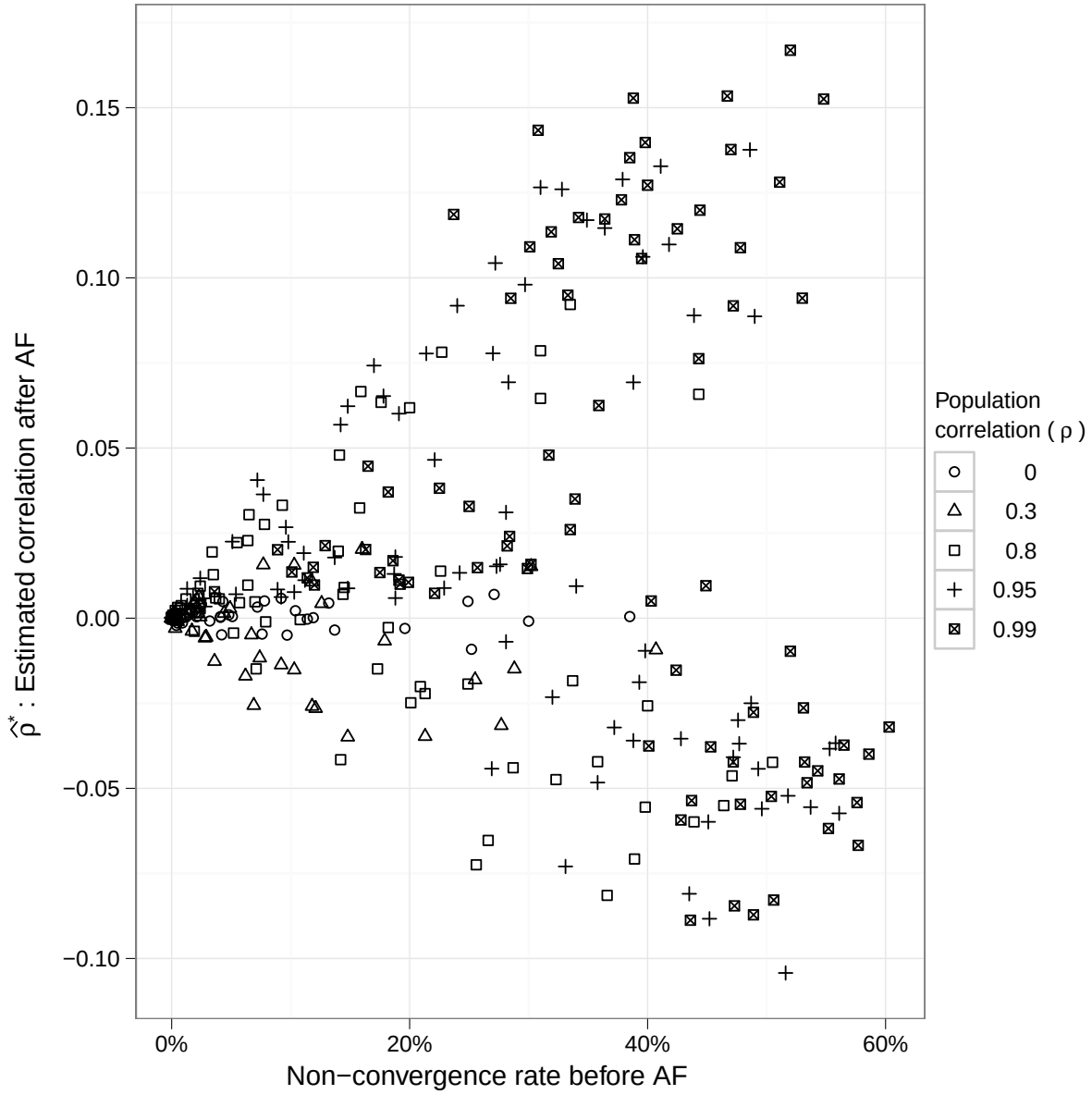

Figure 4.5: The scatter plot of correlation estimate in the transformed space after adaptive fitting (AF) against the non-convergence rate before AF across all simulation scenarios for various population correlations

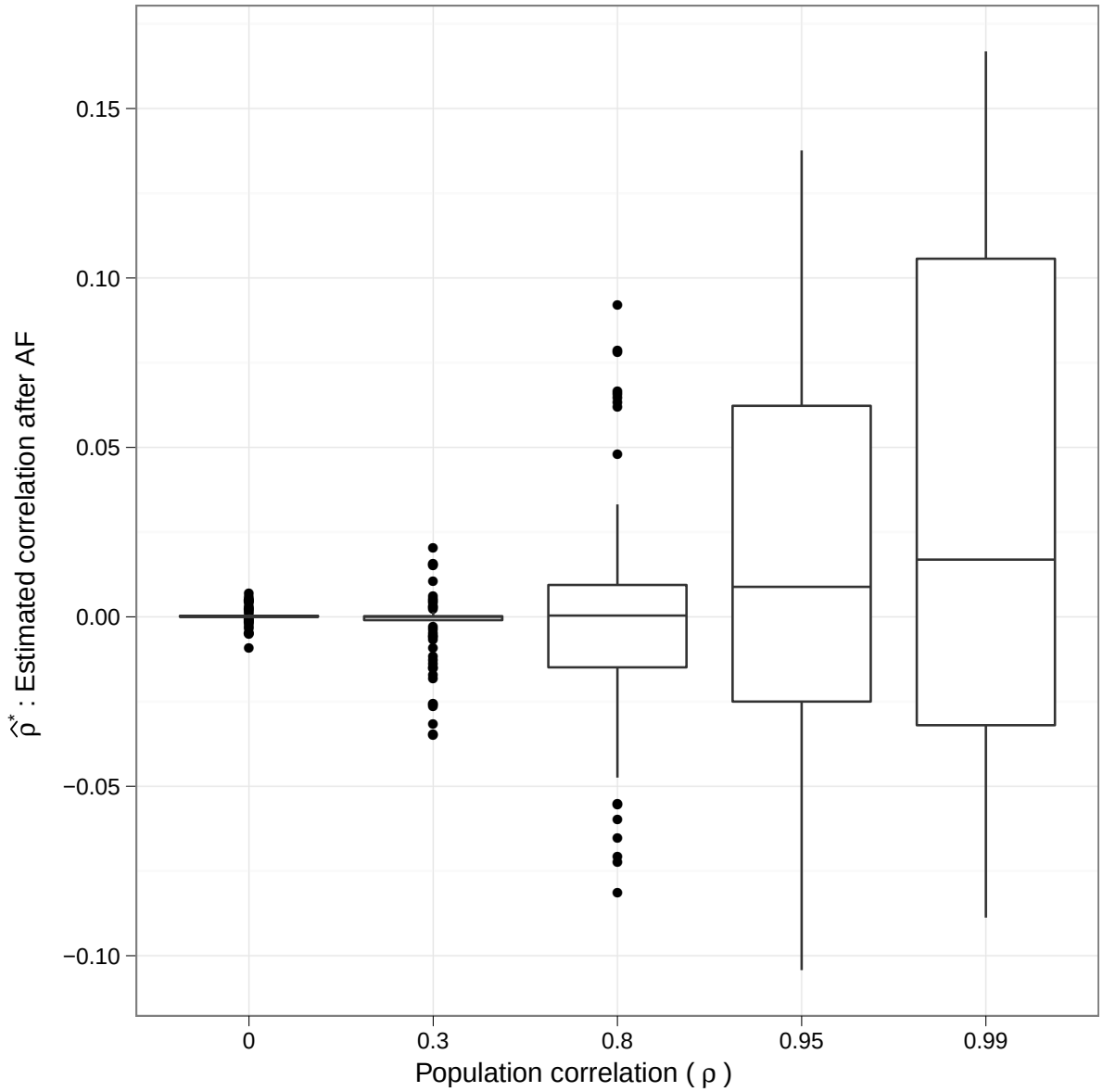

Figure 4.6: The box-plot of correlation estimate in the transformed space after adaptive fitting (AF) against population correlation level across all simulation scenarios

with zero non-convergence rate (Fig. 4.5), the observed random-effects correlation in the optimal transformed space were very close to zero, with a range of (-0.00021, 0.00041).

Therefore, to improve the non-convergence rate, the estimated correlation after AF is not required to be exactly zero, which is the ideal theoretical value after the proposed optimal transformation. The crucial factor is to move the estimated correlation away from the extreme boundary region.

4.4.4 Change in condition number

The theoretical and observed condition numbers in the transformed space against those in the original space were shown by scatter plots (Fig. 4.7 and 4.8, respectively), both stratified by various population correlation settings. To help visual comparison, we provide a dashed concordance line with an intercept of zero and a slope of one to each scatter plot. More deviation below the dashed line indicates a larger reduction of condition number after AF.

Fig. 4.7 analytically predicts that the condition number would become smaller after AF as long as the original population correlation was not zero, namely, 80.0% of 405 scenarios were expected to have smaller condition numbers after AF. The theoretical condition number ratio ($CNR = \frac{CN(G)}{CN(G^*)}$) has a geometric mean (range) of 2.829 (1.000, 288.400) for all 405 scenarios. Larger reduction of condition number should occur for settings with higher correlation and smaller random slope variance relative to random intercept. For example, at $\rho = 0.95$, or 0.99, and $\sigma = \frac{\sigma_{b0}}{\sigma_{b1}} = 2$, or 4, the corresponding theoretical CNR values are all greater than 20 and are well bellow the concordance line on Fig. 4.7. This advantage of large condition number reduction also helped explain why the AF algorithm could perform very well under challenging settings with high correlation and small random slope variance, where both factors tended to significantly increase the non-convergence rate before AF.

Fig. 4.8 demonstrates the empirical condition number reduction pattern for observed results at last iteration, namely, $CN(\widehat{G}^*)$ against $CN(\widehat{G})$. For the settings with very high correlation and relatively small random slope variance, all the corresponding points were well below the dashed lines. Across all 405 scenarios, 389 points (96.0%) were below the dashed lines and only 16 points (4.0%) were slightly above the dashed lines. We note that these 16 scenarios tended to have not only theoretical CNR values very close to one (smaller than 1.6 in 15 out of 16 cases), but also non-convergence rate higher than 20%

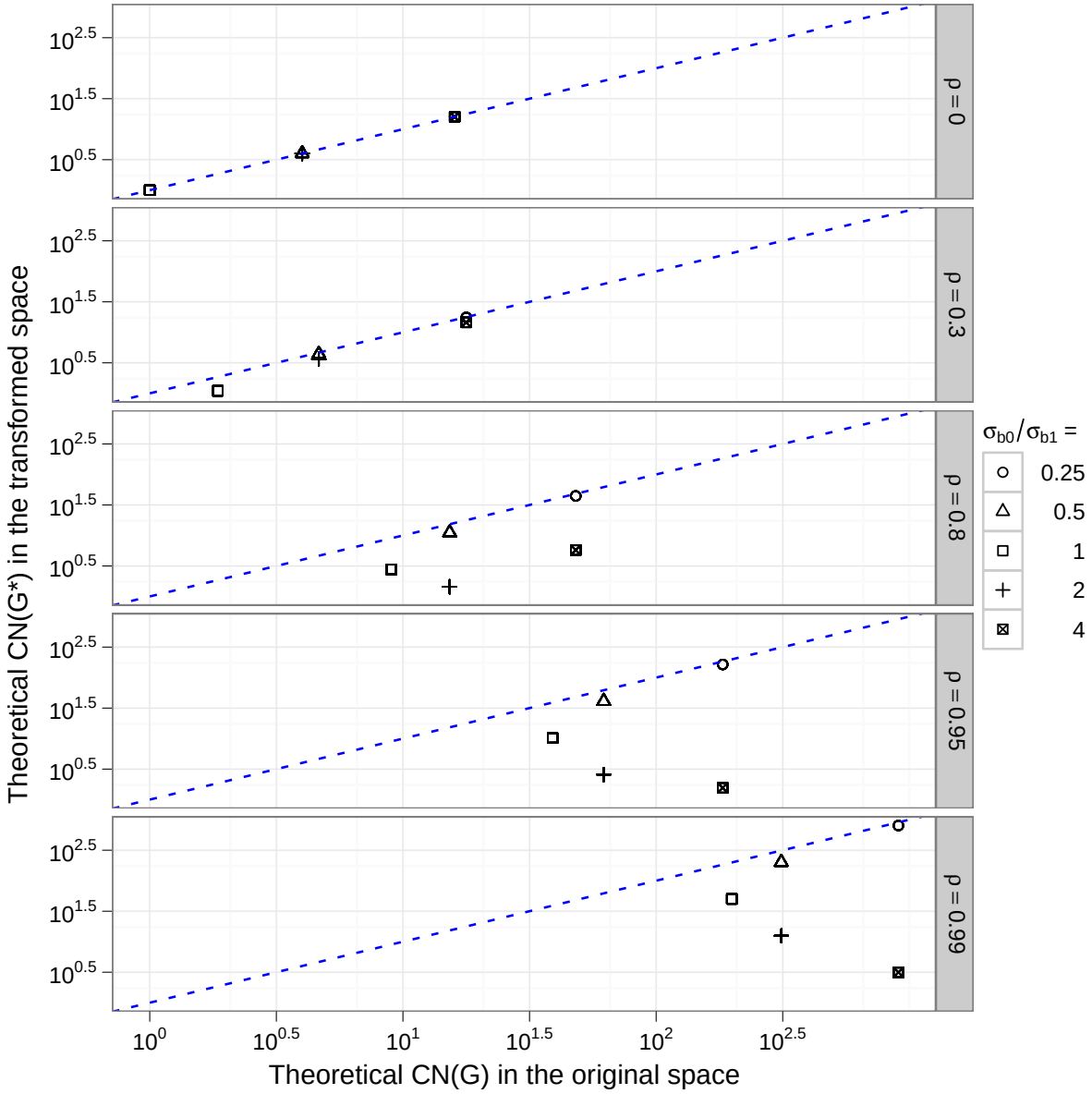

Figure 4.7: Scatter plots of theoretical condition number of random-effects covariance matrix (G) in the transformed space against that in the original space, stratified by correlation level, for different variance component ratios

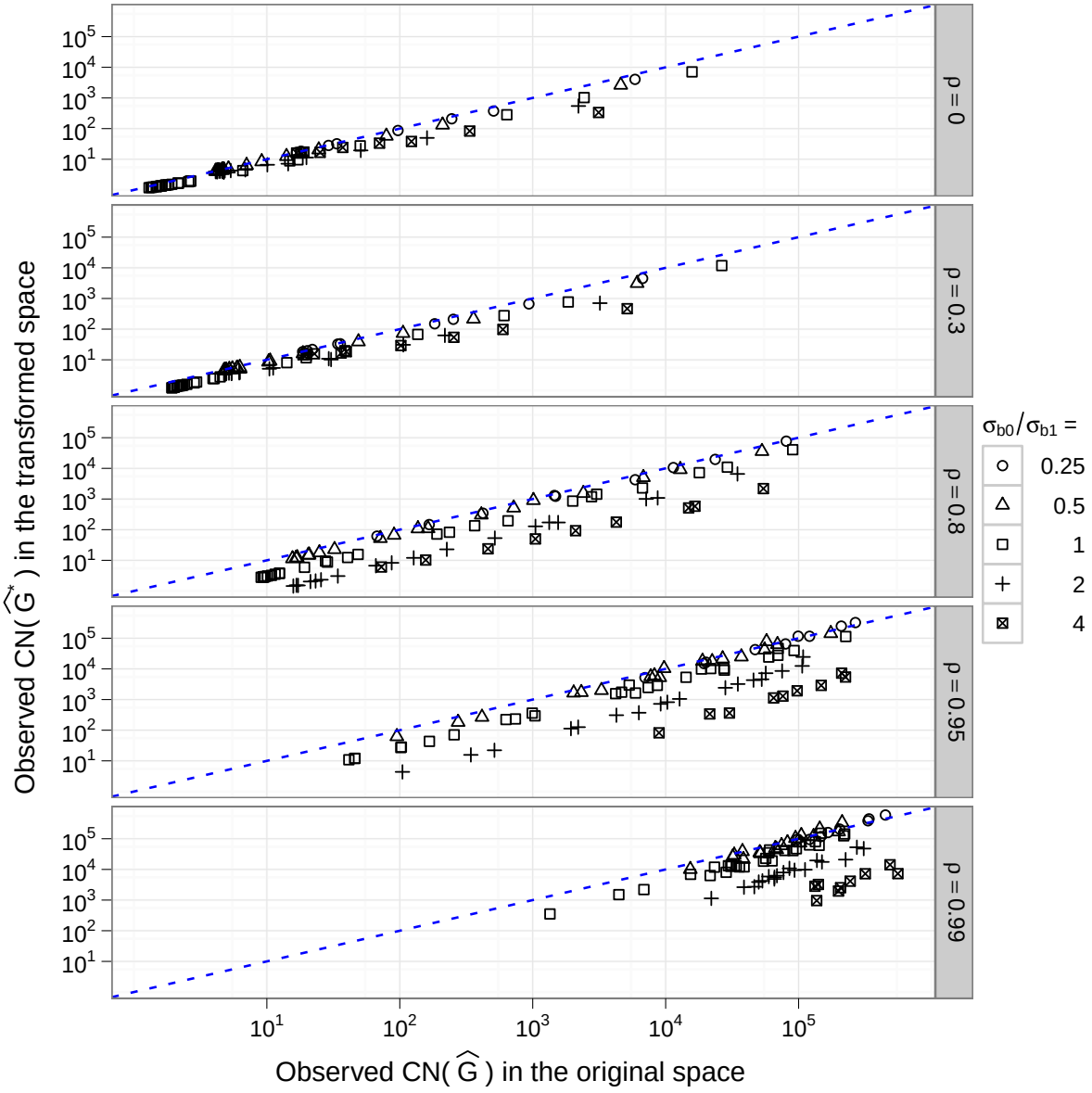

Figure 4.8: Scatter plots of observed average condition number of random-effects covariance matrix (G) in the transformed space against that in the original space across all simulation scenarios, stratified by correlation level, for different variance component ratios

(13 out of 16 cases). In addition, $CN(\widehat{G})$ on the x-axis of Fig. 4.8 was directly observed at last iteration in the original space without AF and thus contained some intermediate fittings from non-convergent runs. It is reasonable to argue that a non-convergent run could give an inaccurate estimate for a covariance matrix and potentially under-estimate the value of $CN(G)$ since the corresponding iterative fitting might have been stopped earlier before reaching more singular level due to non-convergence. Our simulations show that the condition number estimate had a 91.3% empirical rate to become smaller after AF across 405,000 simulated runs. Therefore, the numerical optimization in the optimal transformed space was significantly improved in both the intuitive correlation measure and the more essential condition number measure.

The magnitude of average observed CN of a scenario might be much larger than the theoretical value, indicated by the CN scale differences between Fig. 4.8 and Fig. 4.7. For example, in the original space, the theoretical values of CN had a geometric-mean (range) of 3.43 (1, 16) and 340.4 (199.0, 905.7) across 81 scenarios for $\rho = 0$ and $\rho = 0.99$, respectively, while the corresponding observed values were 13.97 (1.30, 2.016 $\times 10^4$) and 1.277 $\times 10^5$ (1.398 $\times 10^3$, 1.150 $\times 10^6$), respectively. In the optimal transformed space, it is also not surprising that the magnitude of observed CN might increase more than one order of magnitude compared to the theoretical value. However, the observed CN ratio (CNR) due to AF was still generally greater than one, which had been implied by Fig. 4.8. As the number of observations increased, the limiting behavior of condition number ratio was illustrated by Fig. 4.9. The observed CNR tended to approach the theoretical value asymptotically. It is interesting to see that the observed CNR might be larger than the theoretical value, especially for the settings with small sample sizes at $\rho = 0$, where the theoretical value of CNR is one (left and top panel). This means that AF could be better and effective even it was not theoretically expected to have improvement. For the settings with extreme high correlation ($\rho = 0.99$) and small sample sizes, the observed CNR was still generally much larger than one, although it could be relatively smaller than the theoretical value (left and bottom panel).

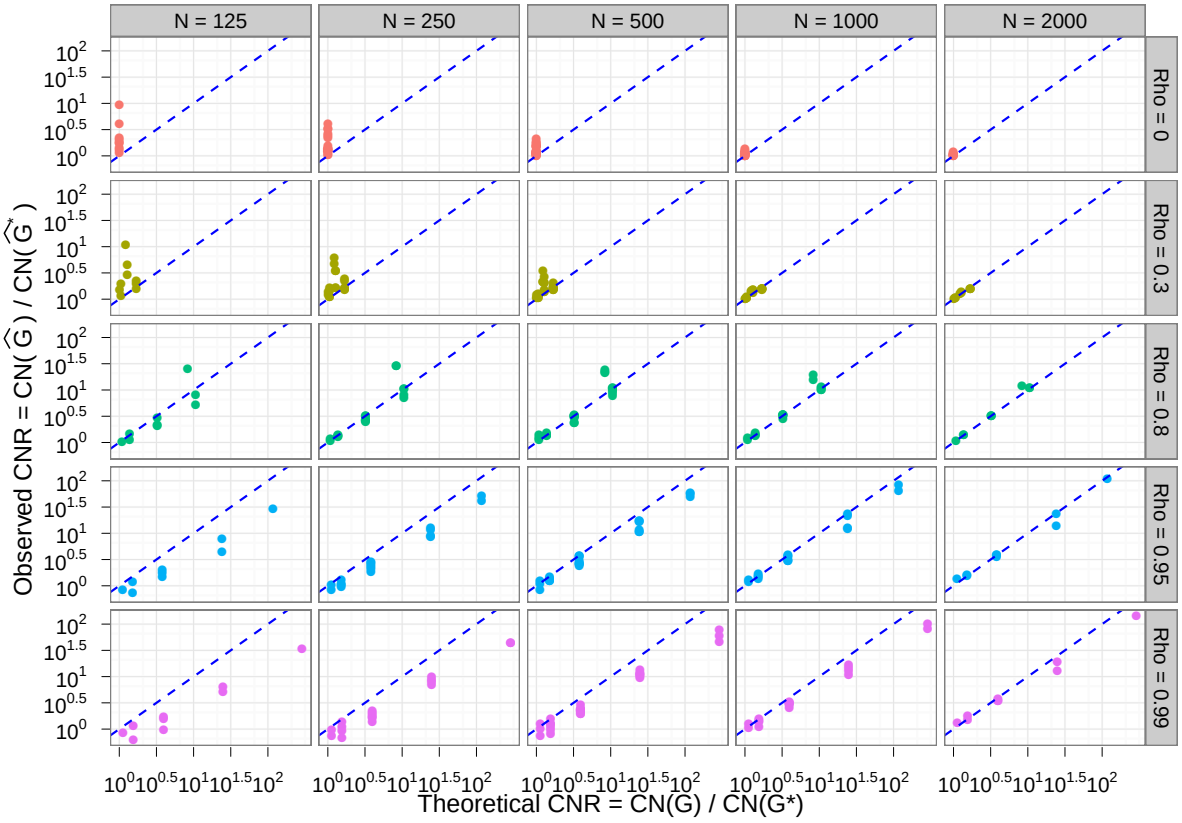

Figure 4.9: Scatter plots of observed average condition number ratio (CNR) of random-effects covariance matrix (G) in the original space over that in the transformed space, against theoretical CNR value across all simulation scenarios, stratified by correlation level (Rho), showing the limiting behavior of CNR as the number of observations (N) increases from left to right

4.4.5 Other simulation results

4.4.5.1 Number of iterations

We illustrate how the improved condition number after AF translated into an accelerated convergence in terms of the number of iteration steps reached at last iterations. Fig. 4.10 plots the geometric mean of the number of iteration steps over 1,000 runs of a scenario before and after AF. The geometric mean (range) of the numbers of iteration steps was 14.06 (1, 85.56) before AF and 7.42 (1, 20.25) after AF across all simulation scenarios. Similar reduction of iterations was also observed for those less-challenging runs which could converge before AF (Fig. 4.11).

4.4.5.2 Parameter estimates

The RIS fittings in the original and a non-singular linear transformed space are theoretically identical but not necessarily numerically identical. To describe the extent and direction of bias of the model parameter estimates in the two estimation spaces, we use the measure of relative bias (RB) to conduct a brief descriptive analysis across various scenarios. The relative bias was defined as the percentage deviation of the estimate away from the theoretical value, $RB = \frac{\widehat{\theta}-\theta}{\theta} \times 100\%$. The analysis was limited to those runs that converged in the original space (before AF) and in the optimal transformed space (after AF).

The relative biases of β_0, β_1, and σ_e estimates were small across all simulated scenarios, ranging from -3.07% to 1.60% before AF, and from -2.07% to 1.26% after AF. Before AF, the average relative biases of the β_0, β_1, and σ_e estimates were negligible, 0.042%, 0.024% and -0.55%, respectively. After AF, the corresponding averages were 0.041%, 0.011% and -0.35%, respectively.

The relative biases of σ_{b0} and σ_{b1} estimates were positive and might be relatively large, with means of 6.25% and 7.10% before AF, 3.83% and 3.61% after AF, respectively. Across various scenarios, except at $\rho = 0$, the average relative biases of $Cov(b_0, b_1)$ estimates were negative (-3.16% before AF and -1.65% after AF), ranging from -40.59% to 4.31% before AF and from -22.33% to 7.07% after AF.

Overall, our simulations show that the estimates of fixed-effects were unbiased while the

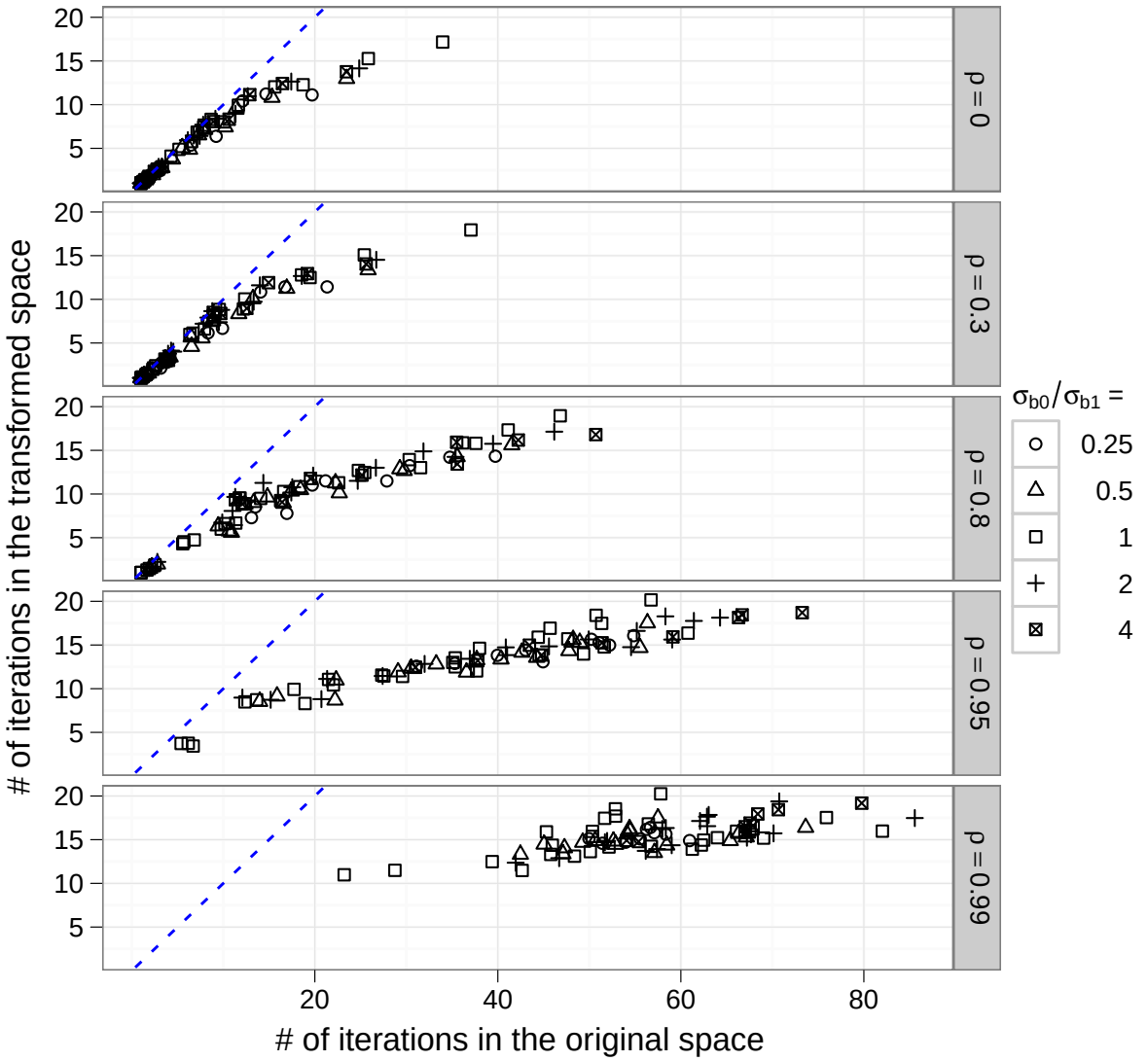

Figure 4.10: Scatter plots of observed average number of iterations in the transformed space against that in the original space across all simulation scenarios, stratified by correlation level, for different variance component ratios

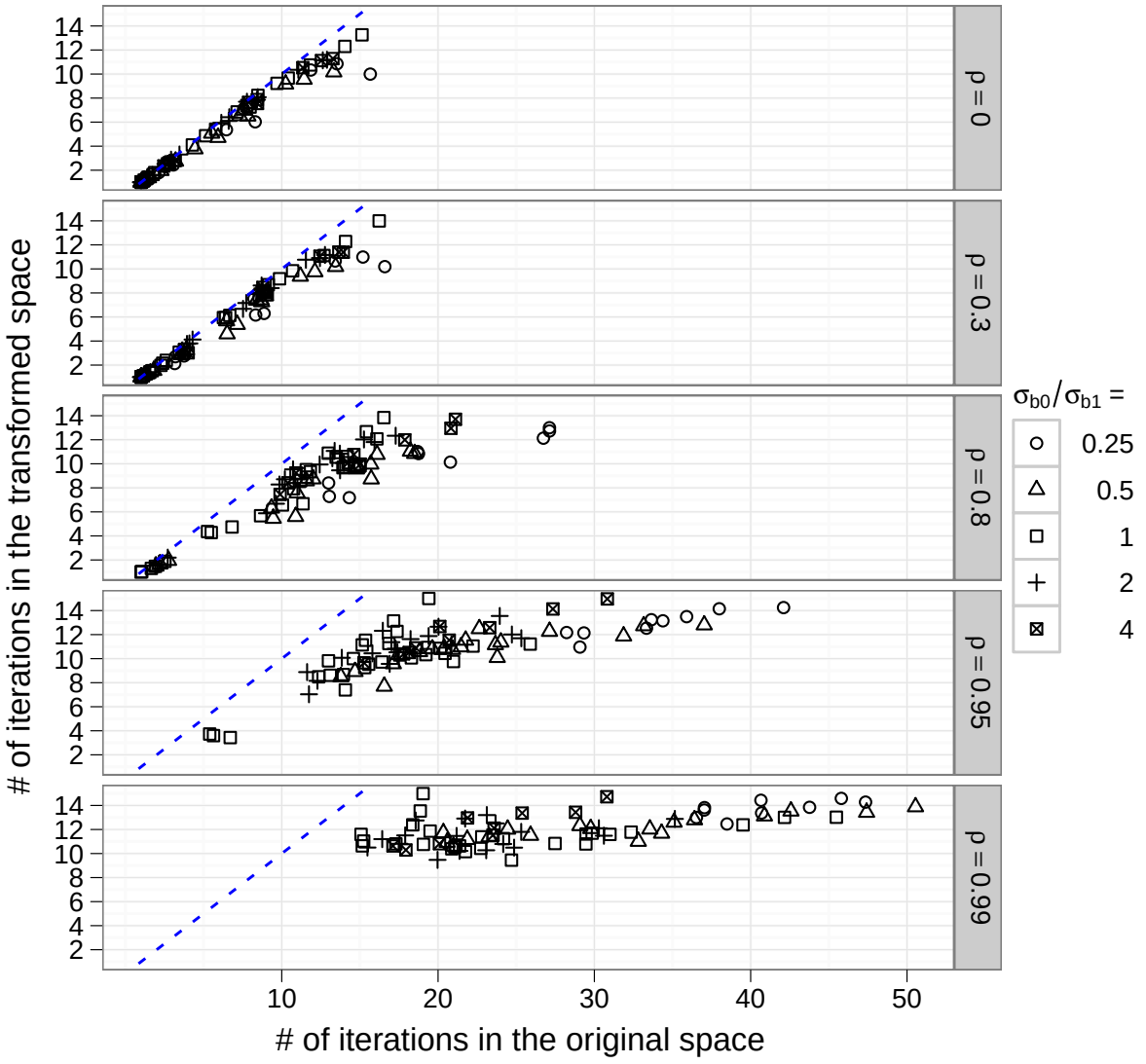

Figure 4.11: Scatter plots of conditional number of iterations (averaging over those subset runs of a scenario which could converge before AF) in the transformed space against that in the original space across all simulation scenarios, stratified by correlation level, for different variance component ratios

covariance matrix of random-effects were relatively poorly estimated under many scenarios, which are similar to those results reported for LMMs with different variance component specifications (Ferron et al., 2002; Murphy and Pituch, 2009).

There were not obvious patterns and differences in the parameter estimates between those non-PD and PD runs. This might be attributed to the high numerical accuracy of the current R(nlme) package, which was considered as a better procedure than SAS Proc Mixed by some researchers (Lange and Laird, 1989). This seemed to be supported by our additional exploration of some simulated data between these two packages. The determinant of the non-PD covariance matrix estimate was not very negative (>-0.0001) when the R(nlme) package was used, but it could be relatively more negative (< -1.0) when the SAS Proc mixed was used to fit the same data set using the same RIS model in the original space.

There were two interesting findings for the covariance parameter estimates of our simulations. First, the directions of biases for σ_ϵ and $Cov(b_0, b_1)$ were negative for many scenarios, while the corresponding biases for σ_{b0} and σ_{b1} tended to be positive. Secondly, for a scenario that had non-convergent runs before AF, those runs which failed to converge before AF tended to have bias in the opposite direction for the four variance components compared to those runs which converged before AF. Thus if those runs which failed to converge before AF were not included in the comparisons, the estimated bias of a scenario is expected to be larger.

4.4.5.3 Starting values based on AF

We further explore the impact of a non-default starting value by refitting the RIS model in the original space for the simulated data in all 405 scenarios. Any iterative optimization algorithm needs initial values to start the iteration process. The initial variance-covariance matrix of random effects is usually assumed to be a diagonal matrix, such as in package R lme (nlme)(Pinheiro and Bates (2000)). Thus, it is possible that the non-convergence issue in the original space before AF is associated with the package default diagonal structure of starting matrix G, which is too simplified and can prevent the algorithm from achieving convergence. As we have known from the simulations, a non-convergence run was generally a run with high correlation estimate between random-effects, where the resulting matrix G at last iteration was not diagonal at all. It is reasonable to argue that it will be easier

for an iterative algorithm to converge if the starting value is similar to the final estimate. Such a starting value of matrix G in the original space could be the estimated $\widetilde{G} = A\widehat{G^*}A^T$ (equation 3.3.2 on page 12), where $\widehat{G^*}$ was available after AF. This was almost always from a *convergent* run (hence $\widetilde{G}$ potentially provided better information than the package default starting value in the original space). Note that both matrices $\widetilde{G}$ and $\widehat{G^*}$ were only available after AF, but they corresponded to different estimation spaces, where $\widehat{G^*}$ was directly obtained at the last iteration after AF and $\widetilde{G}$ was indirectly available from a run fitted in the transformed space.

We applied such a fitting strategy, which was pre-specified by $\widetilde{G}$ in the original space, for all 405 simulation scenarios in this chapter and the obtained results were denoted as "RIS (specified $\widetilde{G}$)". We use "RIS (default G_0)" to denote those results available from the RIS fittings in the original space using the package default starting values. Recall that RIS (default G_0) results had an average empirical rate of 20.59 % to result in problematic runs, which were composed of a 17.23% rate of non-convergent runs and a 3.36% rate of nominally convergent but non-PD runs. The corresponding numbers for RIS (specified $\widetilde{G}$) results were 13.04% = 1.05% + 11.99%, respectively. Namely, RIS (specified $\widetilde{G}$) showed a largely decreased nominal non-convergence rates, with a mean (range) of 1.05% (0, 8.90%) (totally 4,247 failed runs distributed in 268 scenarios). However, the non-PD issue for a nominally convergent run in the original space was still there (more than half scenarios (204/405 = 50.4%) with more than 10% non-PD rate). Overall, using the AF-introduced starting value $\widetilde{G}$ reduced problematic runs from 20.59 % to 13.04%, but did not alleviate the non-PD issue in the original space. This fact strengthens the importance of conducting LMM fitting in the optimal transformed space.

Numerically, specifying the starting random-effects covariance matrix for the LMM fitting in the original space using an estimated matrix $\widetilde{G}$ did not mean that the algorithm did not need any further iteration and could immediately converge just because $\widetilde{G}$ had been calculated from a converged run in the transformed space. Therefore, after pre-specifying with $\widetilde{G}$, the algorithm would proceed as if it started with the default starting value and could fail to converge or even nominally converge to a non-PD estimate of matrix G, even when the specified staring $\widetilde{G}$ was PD and "converged" in some sense.

Chapter 5

Extension to Multiple Random-effects Cases

5.1 Introduction

It is expected to be more complicated to fit a LMM with more than two random-effects since the estimation involves more complex PD requirements. Estimates of each component of a covariance matrix will place limits on the other components. If we think of a covariance matrix as a correlation matrix, it is a necessary but not a sufficient condition for correlations within their domain $(-1, 1)$, to meet the PD constraints of a high dimensional covariance matrix. We illustrate this complexity using a 4-dimensional correlation matrix $M(\rho) = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & 0 & 0 \\ \rho & 0 & 1 & 0 \\ \rho & 0 & 0 & 1 \end{pmatrix}$. The matrix $M(\rho)$ is PD at $\rho = 0.5$, but becomes non-PD at $\rho = 0.6$, where the corresponding minimal eigenvalues are 0.13 and -0.039, and $CN(M)$ are 14 and 52, respectively.

For a mixed-effects model with random intercept b_0 and other random-effects $(b_k; k = 1, \ldots, q-1$ and $q \geq 2)$, let $Z_i = (\mathbf{1}, x_{i1}, \ldots, x_{ik})$ and given a non-singular linear transformation (A) of observed data Z_i, we have $Z_i^* = Z_i A$. The corresponding change in random-effects is $b_i \Rightarrow b_i^* = A^{-1} b_i$.

5.2 Notations incorporating multiple random-effects cases

δ : a general location shift vector, $\delta = (\delta_1, \ldots, \delta_k)^T$, $k \geq 1$, which may reduce or increase the original k of random-intercept related correlations.

d : the optimal location shift vector, $d = (d_1, \ldots, d_k)^T$, which drives the original k of random-intercept related correlations to zero.

A_δ : $A = A_\delta \triangleq \begin{pmatrix} 1 & \delta^T \\ 0 & I_k \end{pmatrix}$, where A is a general location shift transformation matrix specified by a scalar location shift column vector index of $\delta = (\delta_1, \ldots, \delta_k)^T$. Thus $A^{-1} = \begin{pmatrix} 1 & -\delta^T \\ 0 & I_k \end{pmatrix}$.

$G(k)$: $G = Cov(b_i) \triangleq \begin{pmatrix} \sigma_{b_0}^2 & v^T \\ v & \Phi_{k\times k} \end{pmatrix}$, where G is the random-effects covariance matrix in the original space, with the corresponding covariance matrix Φ among random-effects except for random intercept and the covariance v between random intercept and other random-effects.

$G_\delta^*(k)$: $G^* = Cov(b_i^*) = Cov(A_\delta^{-1} b_i)$, where G^* is the random-effects covariance matrix in a transformed space ($Z_i \Rightarrow Z_i^* = Z_i A_\delta$). If $\delta = \mathbf{0}$, then $G^* = G$.

5.3 Optimal linear transformations incorporating multiple random-effects cases

Applying a general location shift transformation to a mixed-effects model with multiple random-effects, by equation (2.2.2) on page 8, we have,

Lemma 5.3.1. *For a general initial covariance matrix $G = Cov(b_i)$ and a general transformation matrix $A = A_\delta$, the new covariance matrix in the transformed space is,*

$$G^* = Cov(b_i^*) = A^{-1} G (A^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 - \delta^T v - v^T \delta + \delta^T \Phi \delta & v^T - \delta^T \Phi \\ v - \Phi\delta & \Phi \end{pmatrix}.$$

We define the optimal shift $d = (d_1, \ldots, d_k)$ to be a location shift vector such that the resulting matrix G^* above has zero covariance between random intercept and other random-effects in the transformed space. Set $v - \Phi\delta = \mathbf{0}$ and we get

$$\delta = \Phi^{-1} v. \tag{5.3.1}$$

If $k = 1$, this reduces to the case of RIS model (Section 3.2), where the optimal shift $d = (d_1)$ is the ratio of the covariance between random intercept and random slope divided by the variance of random slope, i.e.,

$$d \triangleq \rho \frac{\sigma_{b_0}}{\sigma_{b_1}} = \frac{Cov(b_0, b_1)}{Var(b_1)}.$$

However, if $k > 1$, such simple relationship is not necessarily true. For example, $k = 2$, $G = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$, then

$$d_1 = \frac{\sigma_3^2\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \tag{5.3.2}$$

$$d_2 = \frac{\sigma_2^2\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2}. \tag{5.3.3}$$

Thus, to obtain $d_1 = \frac{\sigma_{12}}{\sigma_2^2}$, we need the covariance between two random slopes either to be zero $\sigma_{23} = 0$, or $\sigma_{23} = \frac{\sigma_2^2\sigma_{13}}{\sigma_{12}}$. Similarly, to obtain $d_2 = \frac{\sigma_{13}}{\sigma_3^2}$, we need the covariance between two random slopes either to be zero $\sigma_{23} = 0$, or $\sigma_{23} = \frac{\sigma_3^2\sigma_{12}}{\sigma_{13}}$.

If $k = 3$,$G = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}$, then

$$d_1 = \left\{ -\frac{-\sigma_4^2\sigma_{12}\sigma_3^2 + \sigma_{14}\sigma_{24}\sigma_3^2 + \sigma_{12}\sigma_{34}^2 + \sigma_4^2\sigma_{13}\sigma_{23} - \sigma_{14}\sigma_{23}\sigma_{34} - \sigma_{13}\sigma_{24}\sigma_{34}}{\sigma_2^2\sigma_3^2\sigma_4^2 - \sigma_{23}^2\sigma_4^2 - \sigma_3^2\sigma_{24}^2 - \sigma_2^2\sigma_{34}^2 + 2\sigma_{23}\sigma_{24}\sigma_{34}} \right\}$$

$$d_2 = \left\{ -\frac{\sigma_4^2\sigma_{13}\sigma_2^2 - \sigma_{14}\sigma_{34}\sigma_2^2 - \sigma_{13}\sigma_{24}^2 - \sigma_4^2\sigma_{12}\sigma_{23} + \sigma_{14}\sigma_{23}\sigma_{24} + \sigma_{12}\sigma_{24}\sigma_{34}}{-\sigma_2^2\sigma_3^2\sigma_4^2 + \sigma_{23}^2\sigma_4^2 + \sigma_3^2\sigma_{24}^2 + \sigma_2^2\sigma_{34}^2 - 2\sigma_{23}\sigma_{24}\sigma_{34}} \right\}$$

$$d_3 = \left\{ -\frac{\sigma_3^2\sigma_{14}\sigma_2^2 - \sigma_{13}\sigma_{34}\sigma_2^2 - \sigma_{14}\sigma_{23}^2 - \sigma_3^2\sigma_{12}\sigma_{24} + \sigma_{13}\sigma_{23}\sigma_{24} + \sigma_{12}\sigma_{23}\sigma_{34}}{-\sigma_2^2\sigma_3^2\sigma_4^2 + \sigma_{23}^2\sigma_4^2 + \sigma_3^2\sigma_{24}^2 + \sigma_2^2\sigma_{34}^2 - 2\sigma_{23}\sigma_{24}\sigma_{34}} \right\}$$

The optimal covariance matrix G^* can be obtained after replacing the general δ in Lemma 5.3.1 with the optimal shift vector d. Thus,

Theorem 5.3.2. *After a location shift by A_δ, the new covariance matrix in the transformed space G^* will become diagonal (denoted as G_d^* or Ω), i.e.,*

$$G^* = Cov(b_i^*) = G_\delta^*(v^* = \mathbf{0}) = \begin{pmatrix} \sigma_{b_0}^2 - v^T\Phi^{-1}v & \mathbf{0} \\ \mathbf{0} & \Phi \end{pmatrix} \triangleq \Omega,$$

if and only if $\delta = d$.

Compared to the original covariance matrix G, the new matrix Ω is more sparse and tends to have smaller condition number. For a general covariance matrix, we assume that the lower and upper bounds exist for both its condition number and its minimal and maximal eigenvalues.

Theorem 5.3.3. *Let G be the random-effects covariance matrix in the original space with covariances between random intercept and other random-effects $v = \{Cov(b_0, b_k)\}$, and Ω be the corresponding covariance matrix in the optimal transformed space. Denote two upper bounds satisfying $CN(G) \leq U_G$ and $CN(\Omega) \leq U_\Omega$, then we have*

$$\forall\ Cov(b_0, b_k) \neq 0 \quad \Longrightarrow \quad U_G < U_\Omega.$$

Proof. First, the matrix Ω has a smaller upper bound for its maximal eigenvalue than its counterpart matrix G, since the matrix Ω has smaller maximum absolute row sum. For a symmetric covariance matrix M, $\lambda_{max}(M) \leq$ maximum absolute row sum (Grenander and Szego, 2001). It can also be proved by the fact that the matrix Ω has smaller trace for its squared matrix. $\lambda_{max}(M) = \lambda_{max}^{1/2}(MM^T) \leq Trace^{1/2}(MM^T)$; $\sigma_{b_0^*}^2 = \sigma_{b_0}^2 - v^T\Phi^{-1}v < \sigma_{b_0}^2$, since quadratic form $v^T\Phi^{-1}v > 0, \forall\ v \neq \mathbf{0}$.

Second, the matrix Ω tends to have a larger lower bound for its minimum eigenvalue than its counterpart matrix G. Note that $det(\Omega) = det(G)$ and $\lambda_{min}^{-1}(M) = \lambda_{max}(M^{-1})$. □

The extension of AF to multiple random-effects cases will be illustrated by an actual data set in Chapter 7.

Chapter 6

Uncorrelated RIS (URIS) Model

6.1 Introduction

In practice, when a LMM encounters a non-convergence problem, one may attempt using another LMM with fewer covariance parameters or a LMM with the same number of covariances but a different parameterization. In this chapter we propose an AF-enhanced uncorrelated RIS model and compare it with other competing models based on several model selection criteria using simulation studies. It will also be illustrated by a real life data set in Chapter 7.

6.1.1 RIS, URIS, and RI models

Let $G = Var(b_i) = \begin{pmatrix} \sigma_{b0}^2 & \rho\sigma_{b0}\sigma_{b1} \\ \rho\sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 \end{pmatrix}$ be the random-effects covariance matrix for a general RIS model. If both $\sigma_{b1} = 0$ and $\rho = 0$ are assumed, namely, random slope to be excluded as a random effect, then a RIS model will be reduced to a RI model. If only $\rho = 0$ is assumed, then a RIS model becomes a Uncorrelated Random Intercept and Slope (URIS) model, which keeps the random slope term but imposes zero correlation assumption between random intercept and random slope. Obviously, in the order of RI, URIS and RIAS models, the relative model complexity increases. The relative complexity of the iterative fitting process for a data set is also expected to increase in the same order.

We don't cover the LMM with random slope only (RS) due to two reasons. First, a RS model assumes that every subject is from a homogeneous initial point but with varying growth velocities, where the assumption has to be very carefully justified (West et al., 2006). Second, a RI model follows the hierarchical principle and is a nested model of a RIS model, while a RS model does not. There is a danger when applying a linear transformation on a model which does not follow the hierarchical principle (Morrell et al., 1997).

In practice, when a RIS model encounters non-convergence problems, a simpler model may be attempted and it is usually a RI model. Recently, a URIS model has also been briefly described as a candidate model to handle the non-convergence issue, due to the estimated correlation falling on its boundary and with liberal 95% CI, i.e., (-1.000, 1.000), for a specific data set (Pinheiro and Bates, 2000). The general feasibility of URIS model has not been documented as the uncorrelated assumption is forced upon the data which may not be realistic.

6.1.2 AF-enhanced URIS model in the transformed space (URIS.t)

In this chapter, we propose a new two-step modeling strategy to address the non-convergence problems associated with RIS modeling. The first step is the usual AF process. A RIS model is first fitted in the optimal transformed space (denoted as "RIS.t" model), where the estimated correlation between random-effects is expected to be closer to zero than its counterpart "RIS" model in the original space. The second step is to run an uncorrelated RIS model in the optimal transformed space which has been built by the previous RIS.t modeling process. The obtained URIS model is denoted as "URIS.t" model. Similarly, "RI.t" model will be used to denote a RI model fitted in the same optimal transformed space built by the RIS.t model.

6.1.3 LRT, AIC and BIC

Selecting a covariance matrix for random-effects can sometimes be tricky because a variance component might be tested on the boundary of the parameter space. For two nested LMMs, let the "reduced" model with q correlated random-effects specified by $q(q+1)/2$ covariance parameters, and the "full" model with $q+1$ correlated random-effects thus $(q+1)(q+2)/2$

covariance parameters. Then the difference in the number of covariance parameters between these two models will be $q+1$, because the full model has one additional variance parameter and q new covariance parameters compared to its reduced model. When comparing these two nested LMMs using Likelihood ratio test (LRT), the usual chi-squared distribution with $q+1$ degrees of freedom is no longer valid for such a comparison. Instead, the null distribution of LRT has been proposed as a 50:50 mixture of chi-squared distributions with q and $q+1$ degrees of freedom (Lange and Laird, 1989; Verbeke and Molenberghs, 2000; Pinheiro and Bates, 2000; Fitzmaurice et al., 2004; West et al., 2006).

For example, a RI model has $q = 1$ random-effect, and a RIS model has $q + 1 = 2$ random-effects corresponding to 3 covariance parameters. Then the difference in the number of covariance parameters between the "full" RIS model and the "reduced" RI model will be $3 - 1 = 2$. Specifically, to compare these two models using LRT, the mixture chi-squared distributions $0.5\chi^2_1 + 0.5\chi^2_2$ has a cut point of 5.14 at $\alpha = 0.05$, while the corresponding cut point is 5.99 for χ^2_2, the usual chi-squared distribution with 2 degrees of freedom. For another example, when testing a single variance parameter, such as comparing a URIS model with a RI model, the mixture distributions will be $0.5\chi^2_0 + 0.5\chi^2_1$ and with a cut point of 2.71 at $\alpha = 0.05$, while the usual $\chi^2_{1(\alpha=0.05)} = 3.84$. These two examples illustrate that the usual chi-squared distribution χ^2_{q+1} is less powerful than its mixture counterpart $0.5\chi^2_q + 0.5\chi^2_{q+1}$. Failure to take this issue into account can result in selecting a model for covariance matrix that is too parsimonious. Consequently, the selected model might be too simplistic, and the important structure in the covariance matrix might be ignored. When the usual likelihood ratio test (LRT) is used for such non-standard testings, an ad hoc solution is also recommended by some researchers (Fitzmaurice et al., 2004), i.e., using $\alpha = 0.1$ instead of $\alpha = 0.05$ to determine the statistical significance of the LRT.

Alternatively, the model selection can also be based on some"information criteria", such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Given a pool of candidate models for the covariance matrix, if AIC criterion is used, one should pick the model that minimizes

$$\begin{aligned} AIC &= -2(\textit{maximized loglikelihood}) + 2(\textit{number of parameters}) \\ &= -2(\widehat{l} - c), \end{aligned} \tag{6.1.1}$$

where $\widehat{l}$ is the maximized restriced log-likelihood and c is the number of model parameters ($c = p+q^*$, where p and q^* are the dimension of fixed-effects β and the number of covariance

parameters, respectively).

Similarly, if the BIC criterion is applied, one should select the model that minimizes

$$\begin{aligned} BIC &= -2(maximized\ loglikelihood) + logN^*(number\ of\ parameters) \\ &= -2(\widehat{l} - log\sqrt{N^*}\ c), \end{aligned} \tag{6.1.2}$$

where N^* is the "effective sample size" and is $N-p$ for REML estimation. There are various versions of AIC and BIC where the definition of likelihood, N^* and also c may be different (Pinheiro and Bates, 2000; Fitzmaurice et al., 2004; Littell et al., 2006).

These two information criteria can be used to compare non-nested models for the covariance matrix selection. In general, BIC imposes a very large penalty for the estimation of each additional covariance parameter. BIC is usually not recommended to be used for covariance model selection (Fitzmaurice et al., 2004, pp. 177) since it entails a high risk of favoring a model that is too simple for the data at hand. Guerin and Stroup (2000) reported that BIC tends to favor a more parsimonious model but with worse Type I error rate control than AIC, using SAS 8.0 Proc Mixed procedure. It can also be of interest to compare models with the same dimension of covariance parameters, although it is actually equivalent to compare the fitted likelihood values which are "maximized" under different conditions. Note that REML log-likelihood is not valid to compare random-effects if two LMMs have different number of fixed-effects.

6.2 Simulation settings

A series of balanced datasets generated from the following RIS model,

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + b_{0i} + x_{ij}b_{1i} + \varepsilon_{ij}, \qquad i = 1, \cdots, m; \qquad j = 1, \cdots, n, \tag{6.2.1}$$

where y_{ij} was the jth observation within ith group; fixed-effects $\beta_0 = -1$ and $\beta_1 = 0$; predictor variable x_{ij} was sampled from the standard normal distribution, $N(0, 1)$; residual ε_{ij} was sampled from the standard normal distribution, $N(0, 1)$.

Sample sizes ($N = m \times n$) varied from 100 to 2000, with number of groups $m \in$ (20, 50, 100) and group size $n \in$ (5, 20), a balanced design. Each random effect had three levels, σ_{b0} (intercept) or σ_{b1} (slope) $\in$ (1, 1/2, 1/4). The correlation between random-effects had two

levels, $\rho \in (0.00, 0.99)$. Thus, there were totally 54 scenarios (6 sample sizes $\times$ 9 variance combinations) at each correlation level.

The number of replication runs per simulation design scenario was set at 1000. Each of RIS-simulated data would be fitted by 3 models (RIS, URIS and RI) in the original space and another 3 models (RIS.t, URIS.t and RI.t) in the optimal transformed space.

6.3 Simulation results

6.3.1 Convergence

For the general RIS model in the original space at $\rho = 0$, the non-convergence rate had a mean (range) of 4.36% (0.00%, 35.20%) across all 54 scenarios (Fig. 6.1). Half of scenarios had non-zero non-convergence rates, with a mean (range) of 8.72% (0.1%, 35.20%) across these 27 scenarios. For the general RIS model in the original space at $\rho = 0.99$, the non-convergence rate had a mean (range) of 42.61% (0.00%, 78.00%) across all 54 scenarios (Fig. 6.2), where a zero non-convergence rate only occurred for one setting ($m = 100$, $n = 20$, $\sigma_{b0} = 0.25$, $\sigma_{b1} = 1$ and $\rho = 0.99$).

Unlike RIS model, such non-convergence issues were not observed for other five models (URIS and RI models in the original space; RIS.t, URIS.t and RI.t models in the optimal transformed space) for all 108 simulated scenarios.

Although the non-convergence problem was actually only associated with the RIS model, it is still meaningful to examine the performances of other five LMM fittings against the non-convergence rate of the RIS model since they were all fitted on the same series of RIS-simulated data.

6.3.2 Log-likelihood

Fig. 6.3 illustrates all the pairwise comparisons among the 6 models in terms of the fitted log-likelihood values in a setting ($m = 100$, $n = 5$, $\sigma_{b0} = \sigma_{b1} = 0.5$ and $\rho = 0$), where all 6 models had all 1000 converged runs. The RIS model seemed to have the same fitted

log-likelihood values as its counterpart (RIS.t) in the optimal transformed space. This seemed also the case between the RI and RI.t models, but not between the URIS and URIS.t models. The URIS.t model was likely to have better fitted log-likelihood values than the URIS model (Fig. 6.4). All 1000 points on the Fig. 6.4 were above the 45-degree concordance line, with 1000, 661, 177 and 68 out of 1000 points having a larger than 0.001, 0.1, 1.0 and 2.0 increase in the fitted log-likelihood, respectively. Actually, across all 108 simulation scenarios, a better fitted log-likelihood was generally observed for the URIS.t model compared to its counterpart in the original space (Fig. 6.5). Fig. 6.5 shows that the median of the log-likelihood gap between the URIS.t and URIS models in a scenario had a mean (range) of 0.23 (0.19, 0.27) at $\rho = 0$, and 18.41 (0.39, 103.80) at $\rho = 0.99$, where the maximal median occurred for the setting ($m = 100$, $n = 20$, $\sigma_{b0} = \sigma_{b1} = 1$ and $\rho = 0.99$). The improvement of the fitted log-likelihood values for the URIS.t model when it was compared to the URIS models will be expected to affect the model selection results determined by both AIC and BIC criteria.

With the same model specification, the log-likelihood is expected to be the same between RI.t and RI, or between RIS.t and RIS. This was confirmed by our simulations. The RI and RI.t models had a more than 90% empirical rate to be "identical" and otherwise not distinguishable numerically across all simulated scenarios, where the maximal difference in the fitted log-likelihood out of all 108,000 runs was 1.179×10^{-7}, which was two order less than the error tolerance, the convergence criteria ($1e^{-5}$) of the Newton-Raphson algorithm. To avoid redundancy, we can use the RI model to fully represent the RI.t model results.

Similarly, the RIS and RIS.t models had almost identical fitted log-likelihood, implied by those 27 scenarios without the non-convergence issue for the RIS model at $\rho = 0$, where the average difference in the fitted log-likelihood in a scenario was at most at the order of 10^{-6}, and the maximal difference in the fitted log-likelihood out of 27,000 runs was 7.319×10^{-3}. To alleviate the complexity due to non-convergence runs associated with the RIS model, only the RIS.t model results will be directly used for the comparisons with other model candidates in the following analysis. Recall that the RIS.t model was fitted in the optimal transformed space with zero non-convergence rate, and can be one-to-one transformed back into the original space. Thus, we can focus our analyses on only 4 models (URIS, RI, RIS.t and URIS.t).

The URIS.t and RIS.t models also had very similar fitted log-likelihood across all 108

scenarios (Fig. 6.6). Fig. 6.6 shows that the maximum of the log-likelihood gap between the URIS.t and RIS.t models in a scenario was all less than 0.01 but one exceptional run for a scenario at $\rho = 0.99$.

On the other hand, the URIS and RIS models in the original space was not expected to have similar fitted log-likelihood. This was the natural conclusion by considering the difference between the URIS and URIS.t models shown on Fig. 6.5, the similarity between the URIS.t and RIS.t models shown on Fig. 6.6, and the equivalency between the RIS.t and RIS models, in terms of the fitted log-likelihood. Thus the uncorrelated assumption assumed by both URIS and URIS.t models might significantly change the fitted log-likelihood values in the original space, but had little impact in the optimal transformed space.

6.3.3 AIC, BIC and LRT

6.3.3.1 Three LMMs with random slopes against RI model

Across three fit criteria, the proposed URIS.t model after AF was obviously the most favored model among three LMMs with random slopes when compared to the RI model (Fig. 6.7). For example, when the fit criteria was AIC and correlation was zero (left and top panel on Fig. 6.7), the empirical rate against the RI model were on average 85.35%, 87.54% and 90.94%, with the lowest rate of 16.90%, 21.50% and 39.10%, for RIS.t, URIS and URIS.t models, respectively. Since the data were simulated from RIS models with random slopes, one may expect that the RI model fitting should not be more favorable compared to those LMMs with random slopes. Our simulations shows that the RI model could be more favored by any of three fit criteria, where a $< 50\%$ empirical rate on Fig. 6.7 indicated that a RI model was the favored model. This would occur for those scenarios with the non-convergence issue for a RIS model in the original space (Fig. 6.8 and Fig. 6.9). For example, when the fit criteria was AIC and correlation was zero (top three panels on Fig. 6.8), there were 3 out of 54 scenarios (5.6%) where the RI model was dominant over the URIS.t model, while the corresponding numbers were doubled when compared with RIS.t or URIS models, both at a rate of 6/54 (11.1%). This trend was held at $\rho = 0.99$ too (Fig. 6.9). For example, either for AIC or LRT criteria (top or bottom three panels on Fig. 6.9, respectively), the number of scenarios where the RI model was the favored model was only one for the URIS.t-RI comparison, but it was larger than four for both the URIS-RI and RIS.t-RI comparisons.

Among three fit criteria, the RI model was generally most favored by BIC, and least favored by AIC. When LRT was used to pairwise compare the RIS.t, URIS and URIS.t model against the RI model, the empirical better rates were on average 73.51%, 79.64% and 81.97% at $\rho = 0$, and 90.51%, 83.77% and 94.39% at $\rho = 0.99$.

Besides the impacts of the non-convergence rate, the comparisons with the RI models might also be affected by how small the random slope variance, the number of group (m) and the group size (n) were (e.g., Fig. 6.8, Fig. 6.10 and Fig. 6.11).

6.3.3.2 Among three LMMs with random slopes

The pairwise comparisons between three LMMs with random slopes were determined by AIC and BIC (Fig. 6.12). When compared to the RIS.t model, the URIS model was the favored model at $\rho = 0$, with an average empirical rate of 83.77% by AIC and of 98.35% by BIC, but it was generally not at $\rho = 0.99$, with an average empirical rate of 8.67% by AIC and of 18.22% by BIC. This suggests that the URIS model may be a poor model selection if the random-effects correlation is not zero, although the URIS model has showed better convergence rate than the RIS model in the original space.

The URIS.t model was always better than a general RIS.t model even though data were simulated with random-effects correlation at $\rho = 0.99$. This is consistent with the very similar fitted likelihood between these two models (shown on Fig. 6.6).

The URIS.t model was generally better than its counterpart model in the original space (URIS). The URIS.t model was favored by an average empirical rate of 98.13% at $\rho = 0$ and of 99.15% at $\rho = 0.99$, where both AIC and BIC criteria produced the same rates due to the same number of covariance parameters between the two models. Recall that the URIS.t model also had comparable empirical rates against the RI model at two correlation levels, 90.94% at $\rho = 0$ and 89.30% at $\rho = 0.99$. Thus, the performance of the URIS.t model seemed not to be strongly affected by the random-effects correlation level, but it was not the case for the URIS model. On the other hand, none of runs showed that the URIS model was better than the URIS.t model beyond the error tolerance, no matter whether the non-convergence rate was zero or not.

The AIC comparison between the URIS model with RIS.t model or with URIS models

might be slightly affected by simulation settings (e.g., Fig. 6.13 and Fig. 6.14), especially for small sample size and small random slope variance. The pattern was similar for BIC (data not shown).

When LRT was used for the URIS-RIS.t comparison, the usual chi-squared distribution χ^2_1 was applied since the correlation parameter was not tested on its boundary (Fig. 6.15). By LRT at $\rho = 0$ settings, it is interesting to see that the RIS.t model had a mean (range) empirical rate of 5.3% (3.0%, 7.5%) to be better than the URIS model across all 54 scenarios, and corresponding numbers were 5.2% (4.0%, 7.5%) for those 27 scenarios without the non-convergence issue, which were all close to the nominal 5% level. The LRT results might be poor at $\rho = 0.99$ settings, with a mean of 85.91% and five-number summary of (8.9%, 82.7%, 99.9%, 100.0%, 100.0%). When LRT was used for the URIS.t-RIS.t comparison, the URIS.t model was favorable due to similar fitted log-likelihood and more parsimonious parameterization. LRT was not applicable to compare the URIS.t and URIS models.

6.3.4 Non-coverage of β_1

The non-coverage rates of fixed-effect slope β_1 were summarized in Fig. 6.16. The RI model might have a more than 25% empirical non-coverage rate of β_1 for either correlation level. No obvious differences were observed for the three LMMs with random slopes.

6.4 Summary

The proposed URIS.t model is generally better than the URIS model regardless of the correlation level between random-effects. This is true based on either Log-likelihood, AIC or BIC criterion, at both $\rho = 0$ and $\rho = 0.99$ settings.

If non-convergence is an issue, then the RI model could be better than the three LMMs with random slopes in some scenarios, especially when BIC is used. At $\rho = 0$, the URIS model is generally favored by AIC compared to RIS.t, while RIS.t may be preferred at $\rho = 0.99$.

Theoretically, the URIS.t and RIS.t models are equivalent models regardless of the

original correlation level between random-effects, but the URIS and RIS models are only equivalent when the RIS model also assumes no correlation between random-effects in the original space.

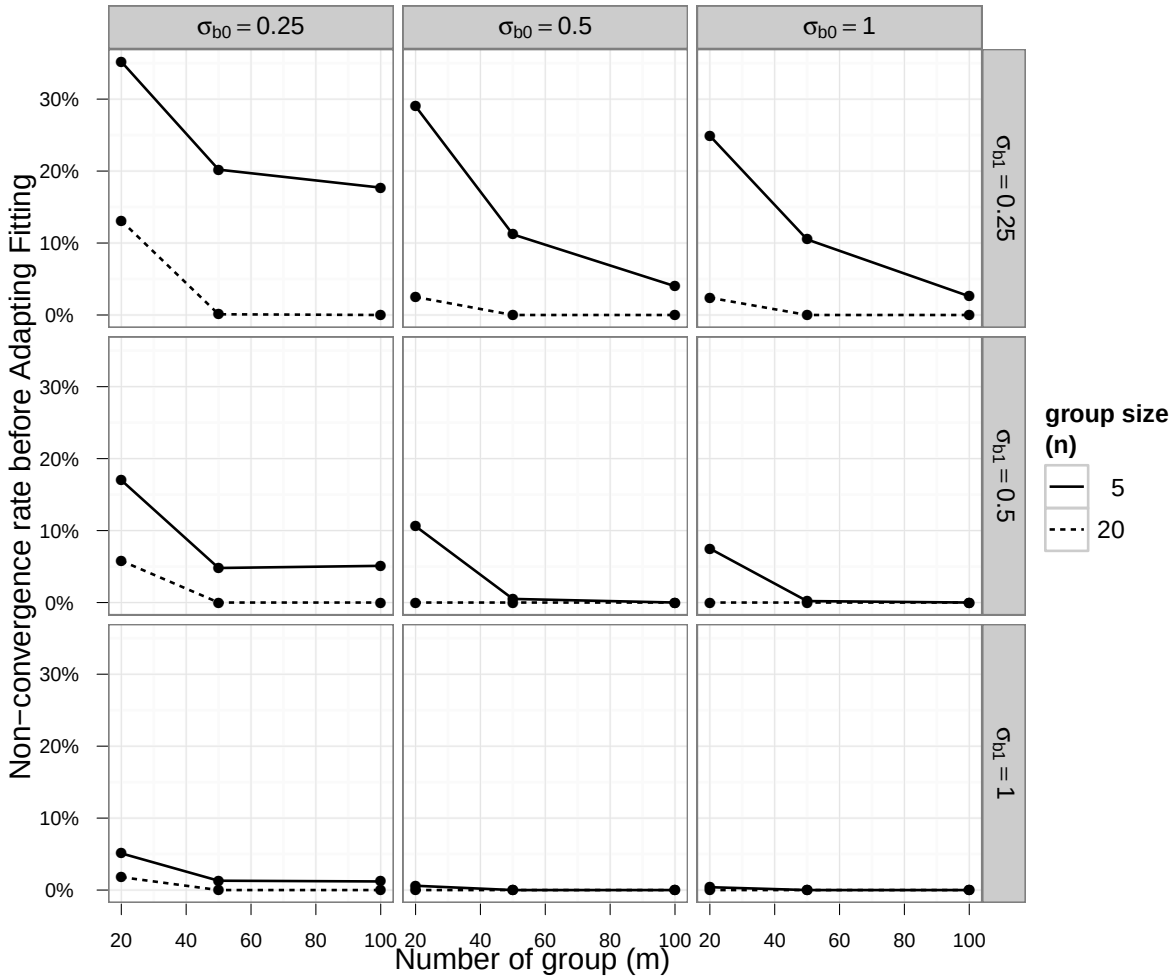

Figure 6.1: The non-convergence rate before adaptive fitting (AF) as a function of the number of group (m), with each of the nine panels corresponding to one of the nine variance component combinations, and with each of the two curves of a panel corresponding to each of two group sizes (n), given random-effects correlation at $\rho = 0$

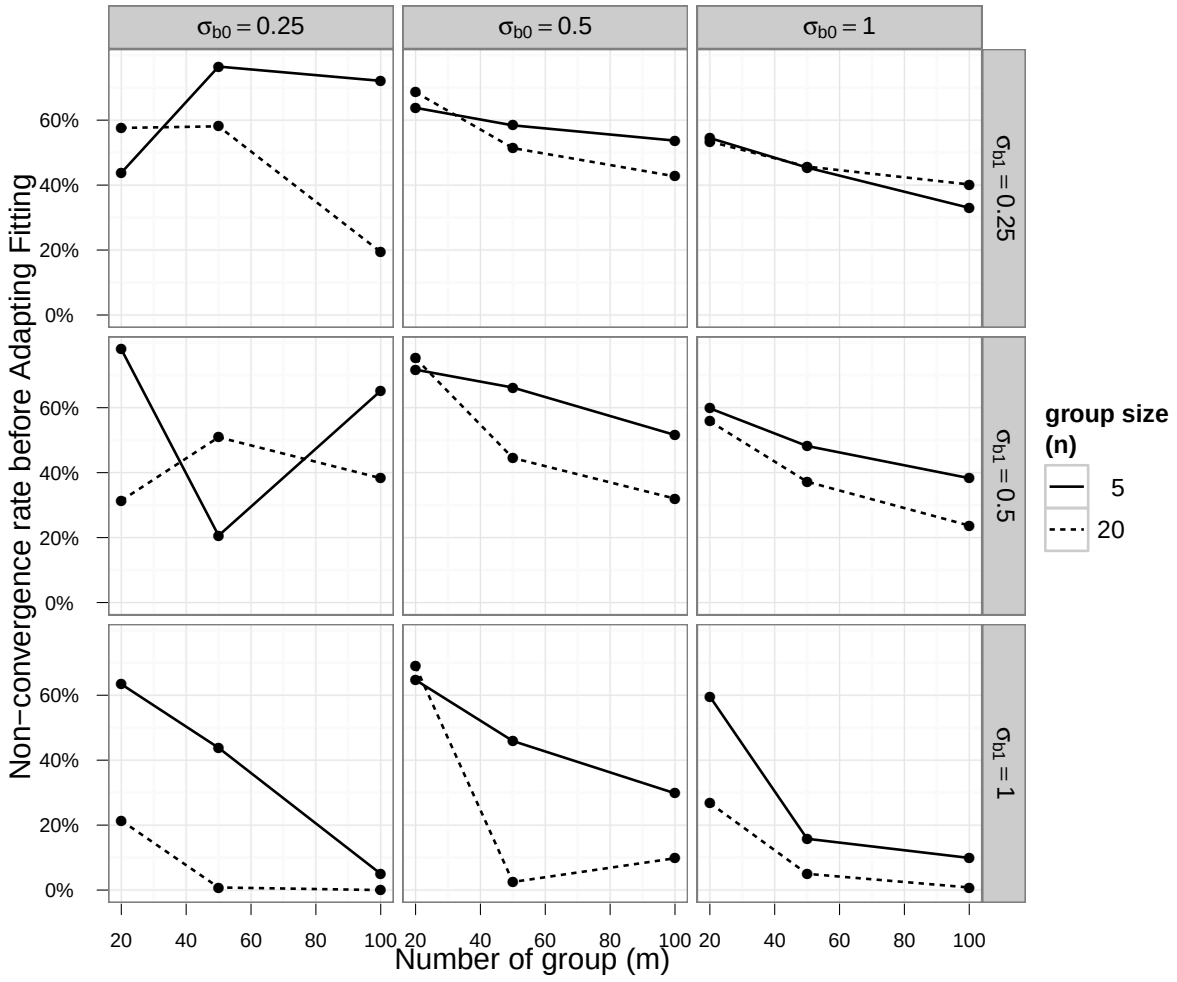

Figure 6.2: The non-convergence rate before adaptive fitting (AF) as a function of the number of group (m), with each of the nine panels corresponding to one of the nine variance component combinations, and with each of the two curves of a panel corresponding to each of two group sizes (n), given random-effects correlation at $\rho = 0.99$

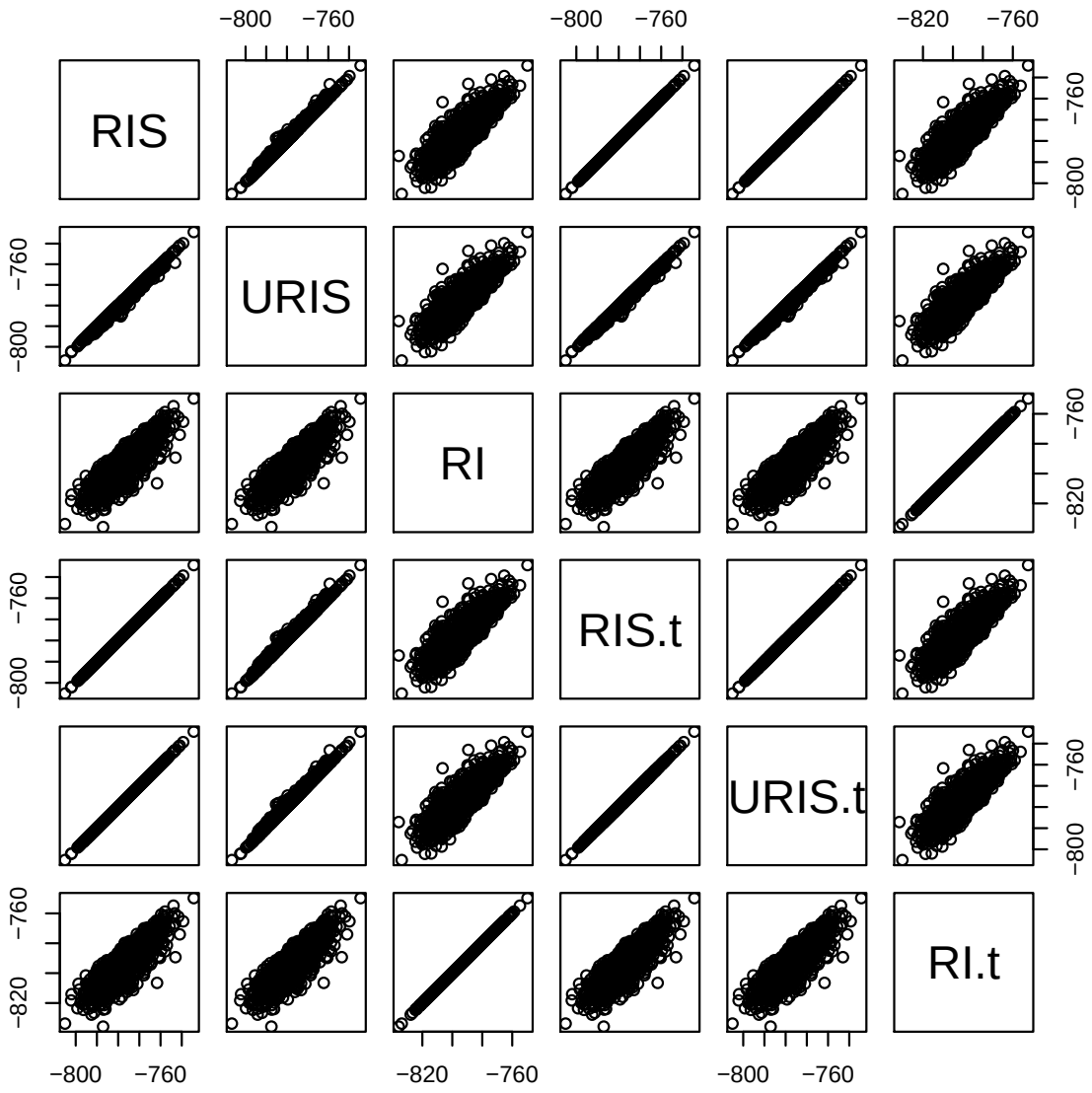

Figure 6.3: Scatter plots of the fitted log-likelihood for 3 models (RIS, URIS and RI) fitted in the original space and 3 models (RIS.t, URIS.t and RI.t) in the optimal transformed space, with the setting ($m = 100$, $n = 5$, $\sigma_{b0} = \sigma_{b1} = 0.5$ and $\rho = 0$)

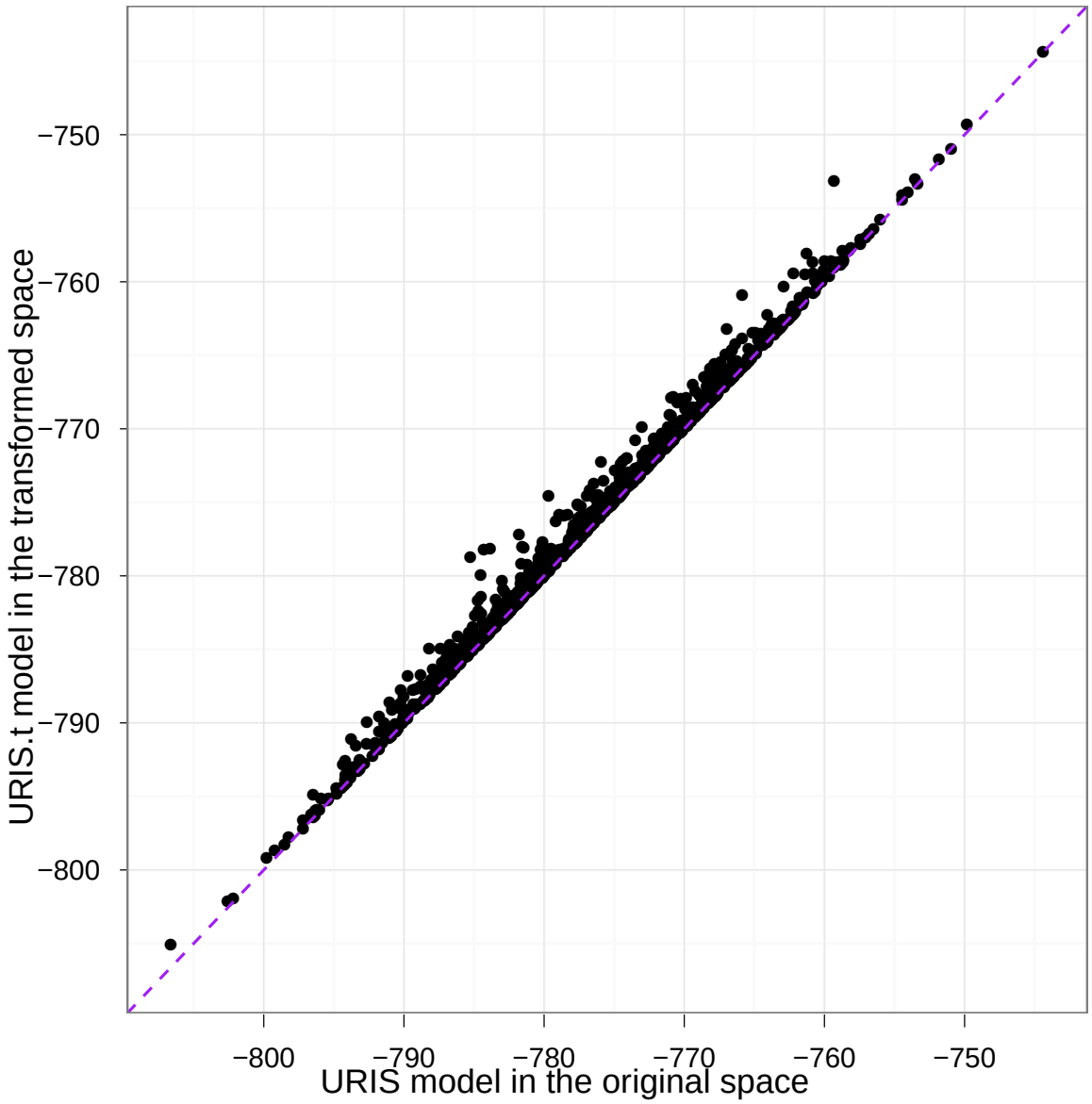

Figure 6.4: Scatter plots of the fitted log-likelihood for the URIS.t model which was fitted in the optimal transformed space against the URIS model in the original space, with the setting ($m = 100$, $n = 5$, $\sigma_{b0} = \sigma_{b1} = 0.5$ and $\rho = 0$)

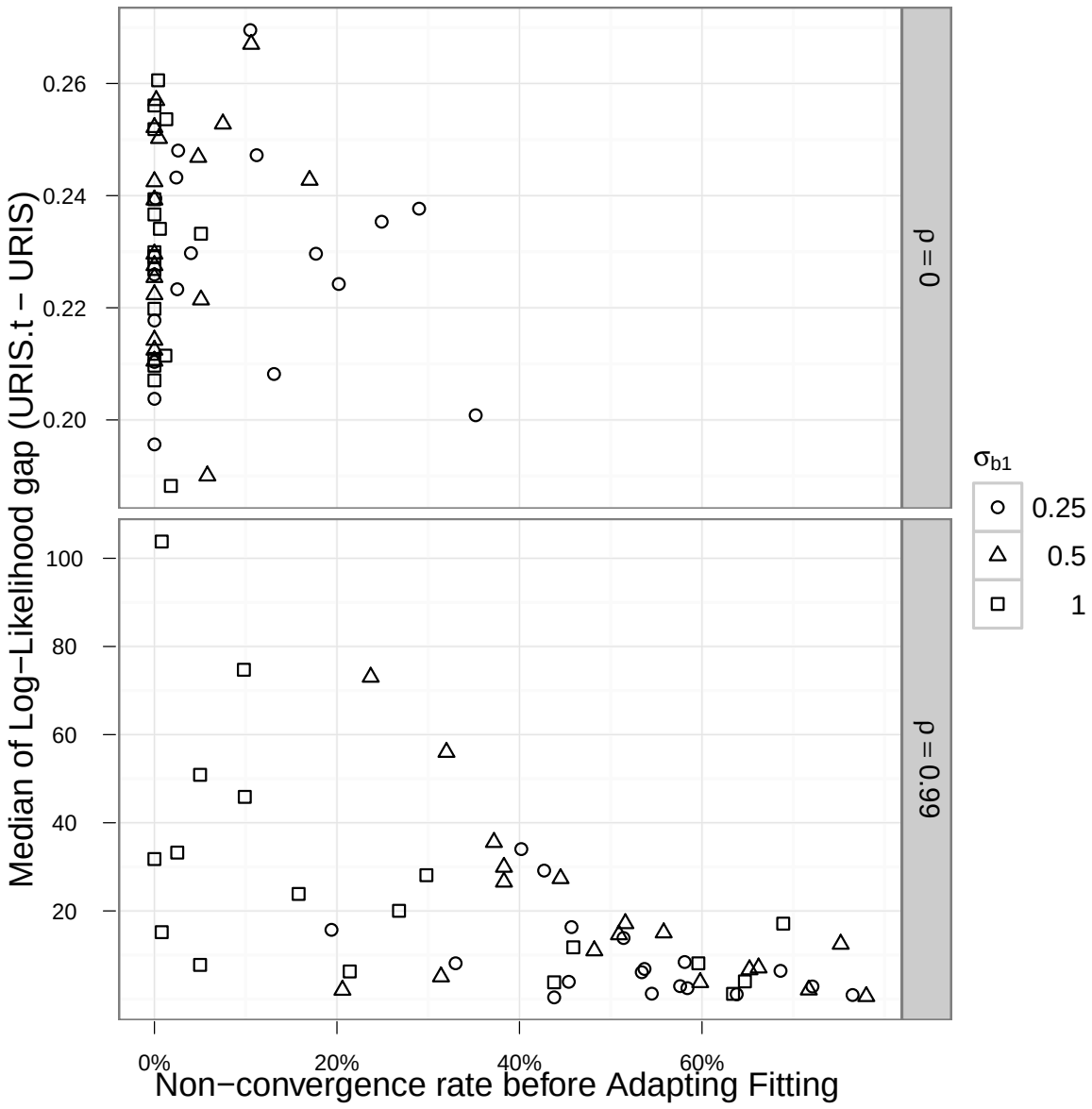

Figure 6.5: Scatter plots of the median of differences in the fitted log-likelihood values between the URIS.t model and the URIS model, as a function of non-convergence rate before Adapting Fitting for the RIS model in the original space across all 108 scenarios, stratified by random-effects correlation levels

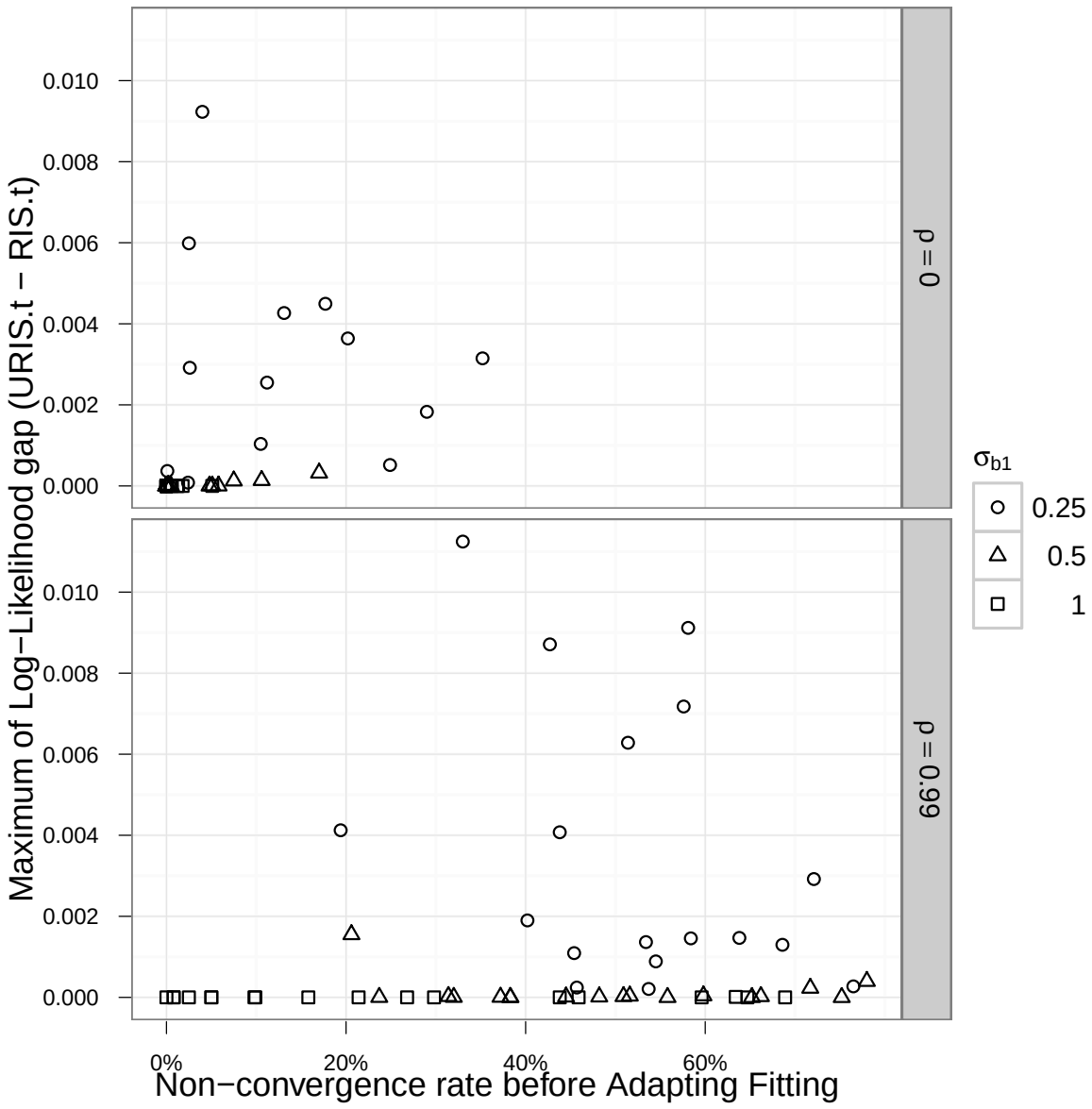

Figure 6.6: Scatter plots of the maximum of differences in the fitted log-likelihood values between the URIS.t model and the RIS.t model, as a function of non-convergence rate before Adapting Fitting for the RIS model in the original space across all 108 scenarios, stratified by random-effects correlation levels

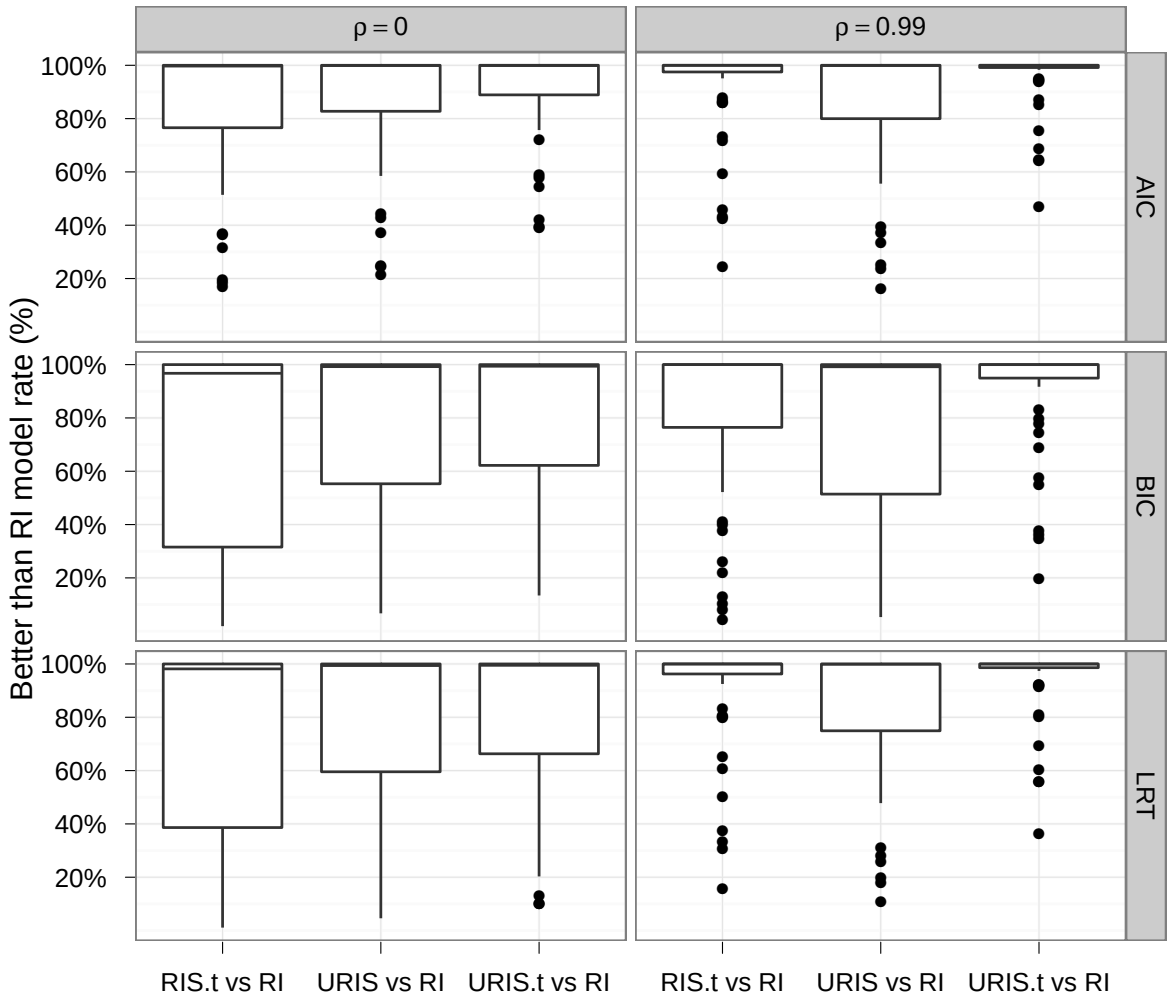

Figure 6.7: Boxplots of comparing the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t) against the LMM with random intercept only (RI), in term of AIC, BIC and LRT, where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0$ or $\rho = 0.99$

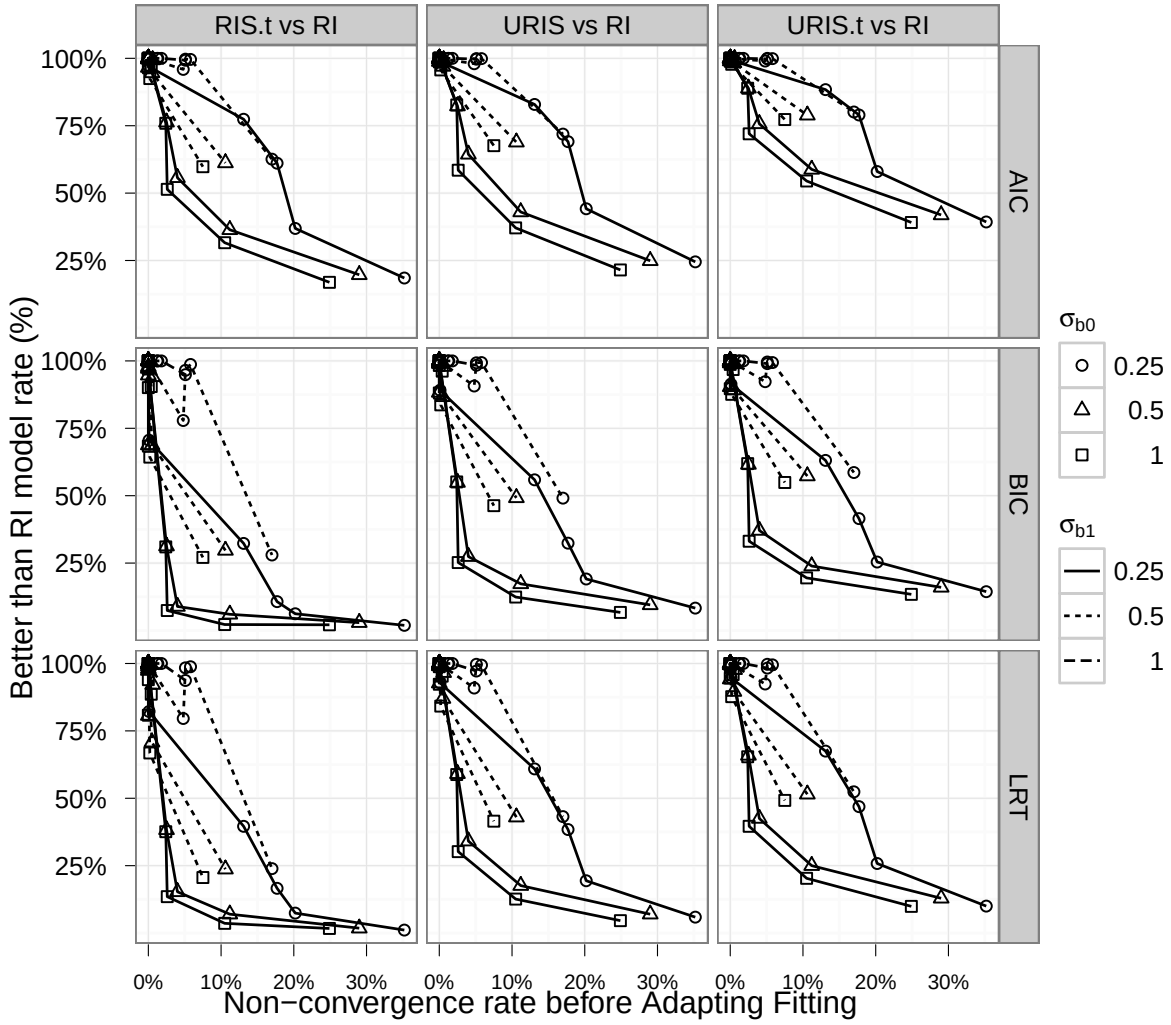

Figure 6.8: Comparing each of the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t) against the LMM with random intercept only (RI), in term of AIC, BIC and LRT criteria as a function of non-convergence rate before Adapting Fitting for the RIS model in the original space, where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0$

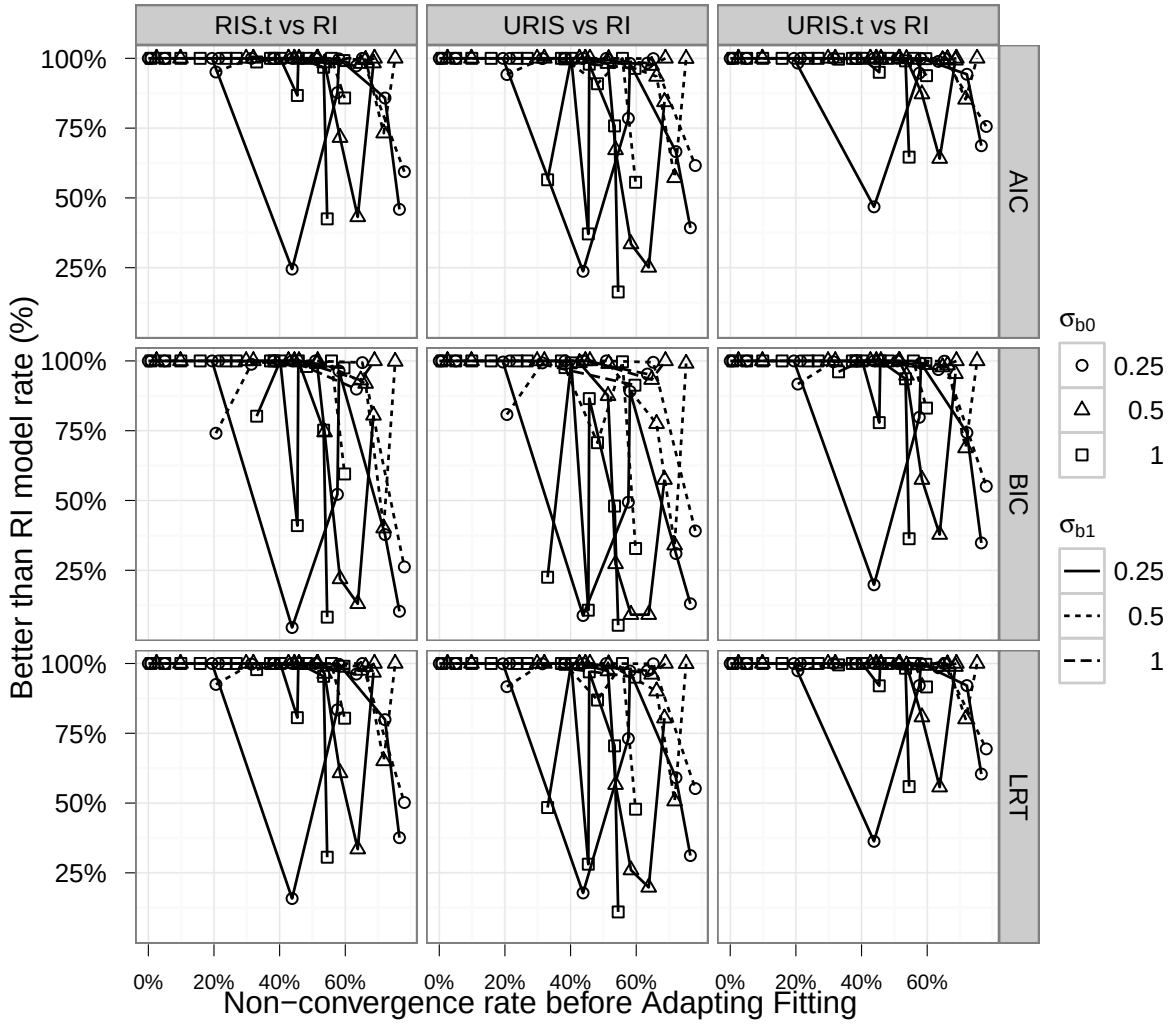

Figure 6.9: Comparing each of the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t) against the LMM with random intercept only (RI), in term of AIC, BIC and LRT criteria as a function of non-convergence rate before Adapting Fitting for the RIS model in the original space, where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0.99$

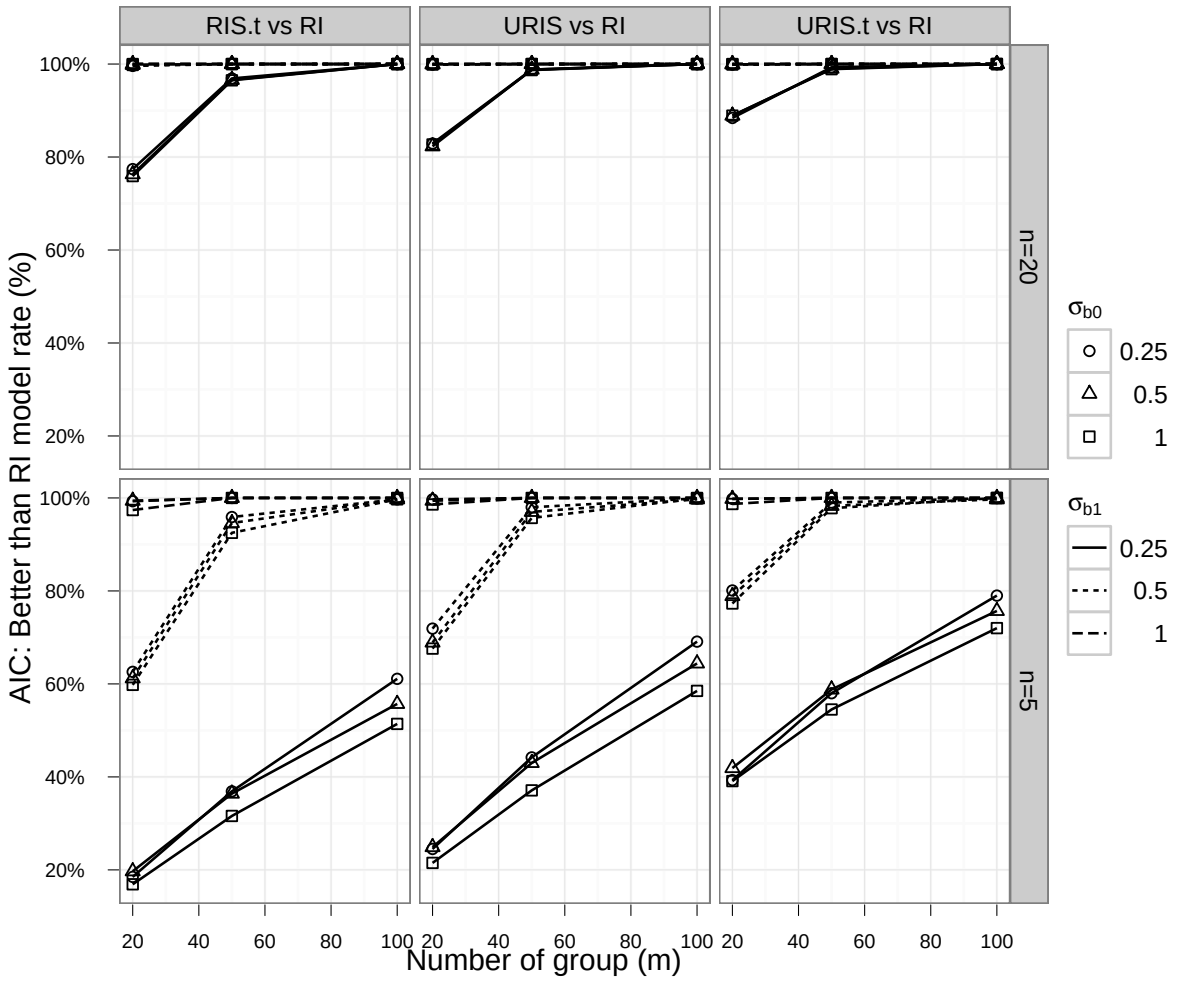

Figure 6.10: Comparing each of the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t) against the LMM with random intercept only (RI), in term of AIC criteria as a function of number of group (m), where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0$

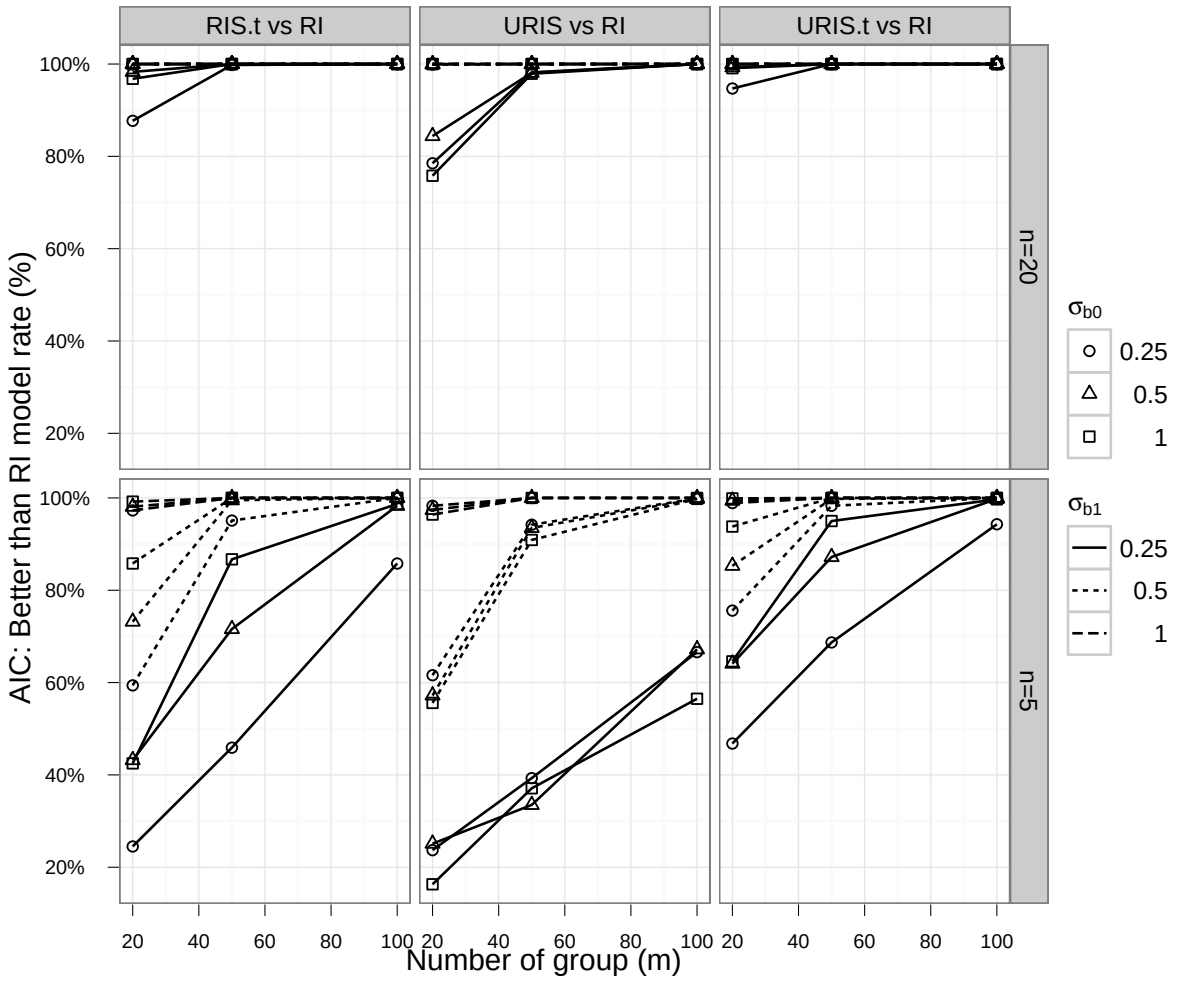

Figure 6.11: Comparing each of the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t) against the LMM with random intercept only (RI), in term of AIC criteria as a function of number of group (m), where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0.99$

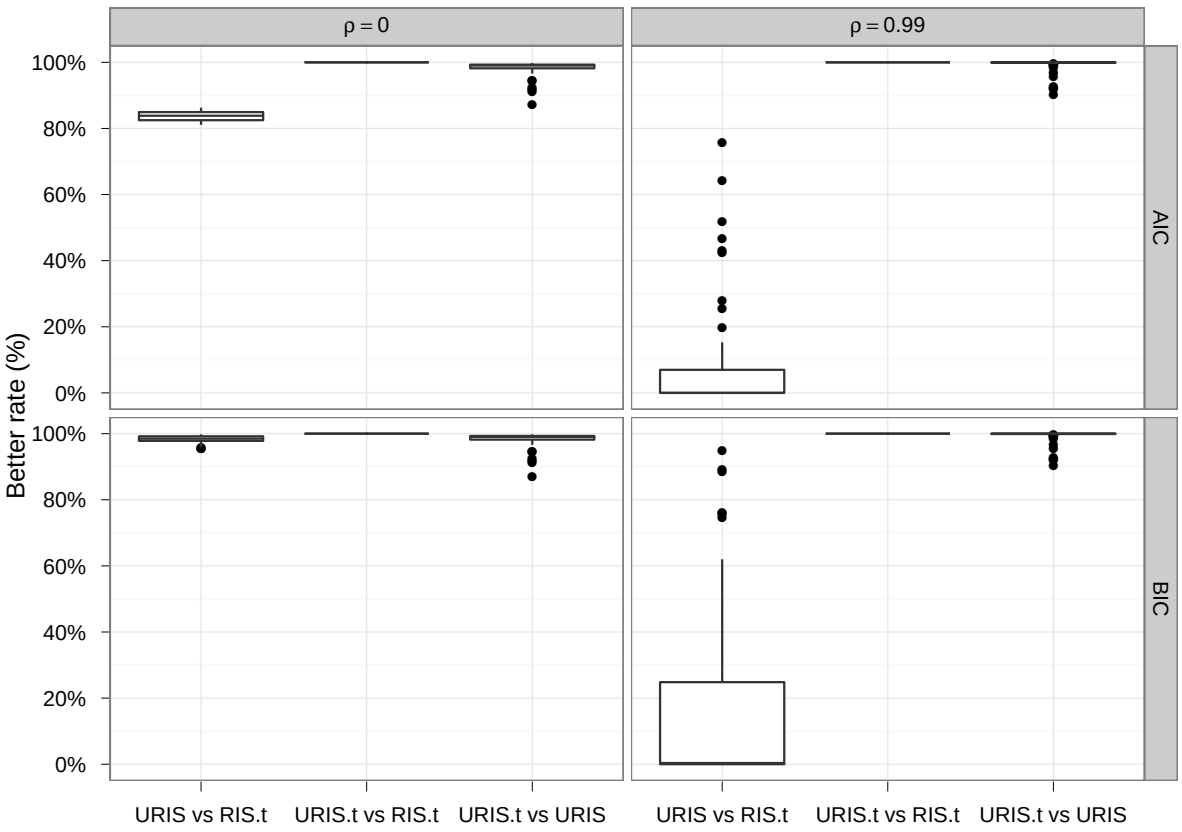

Figure 6.12: Boxplots of comparisons between the three LMMs with random intercept and slope (RIS.t, URIS, URIS.t), in term of AIC and BIC, where all scenarios were simulated from RIS models with random-effects correlation at $\rho = 0$ or $\rho = 0.99$, respectively

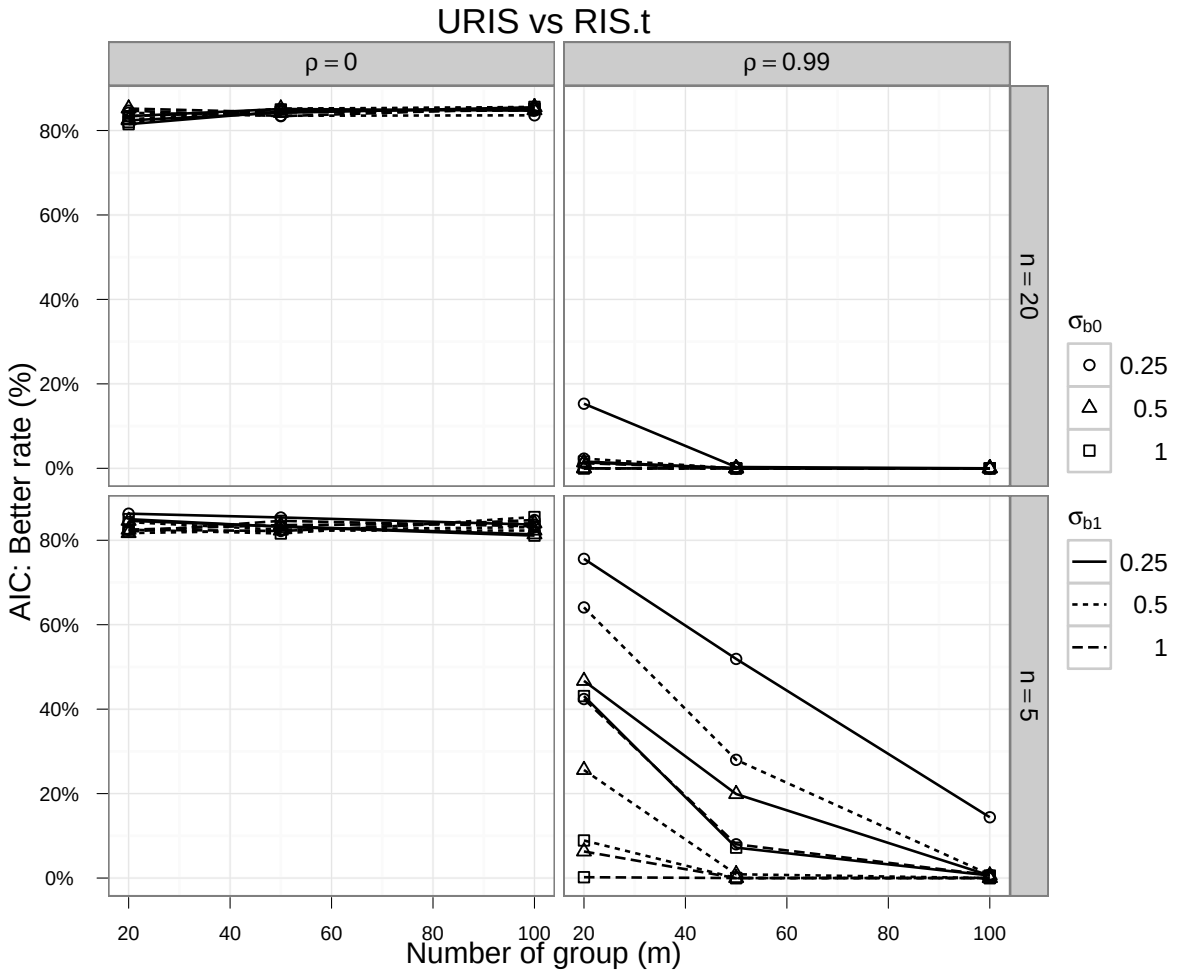

Figure 6.13: Comparisons of URIS vs. RIS.t by AIC criteria as a function of number of group (m), where all scenarios were simulated from RIS models

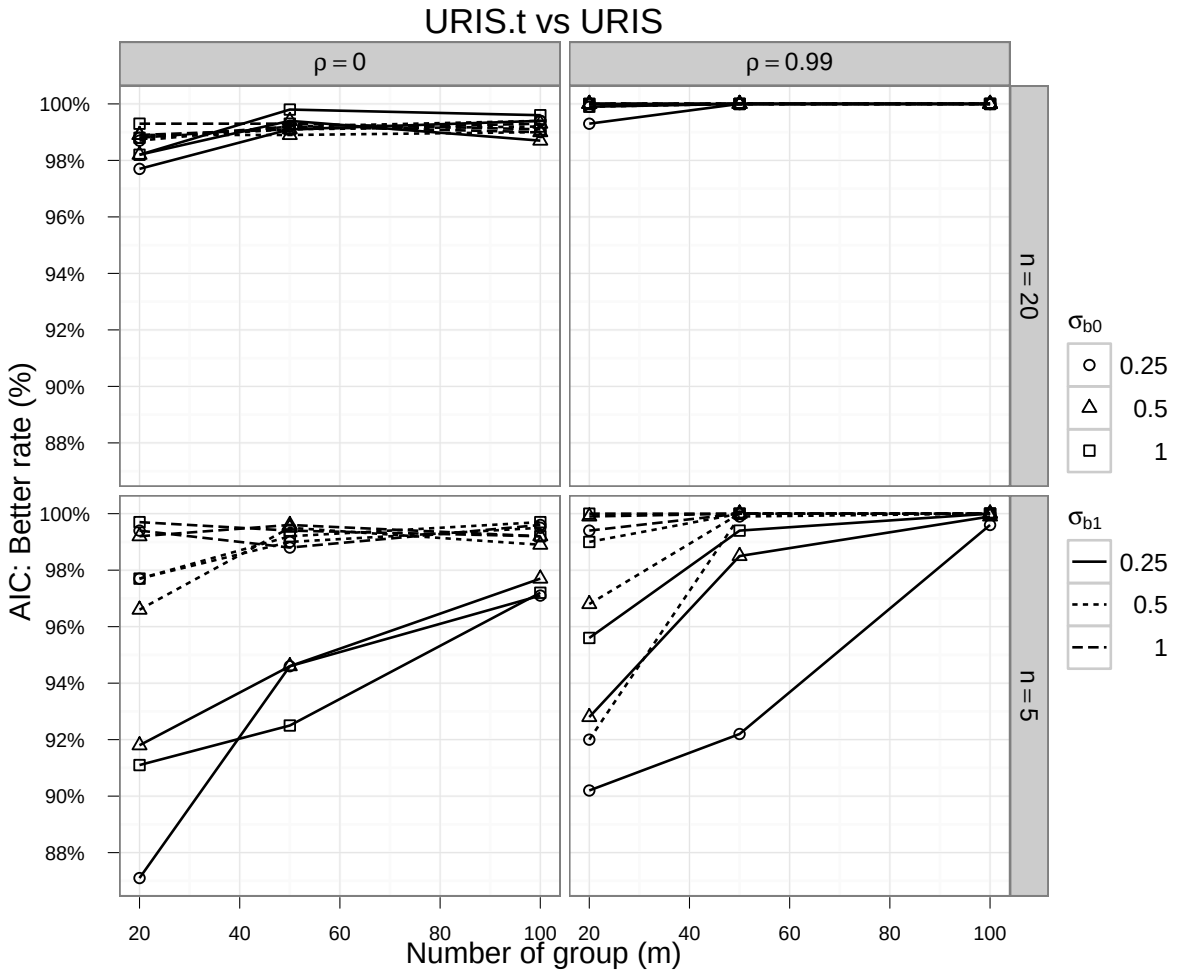

Figure 6.14: Comparisons of URIS.t vs. URIS by AIC criteria as a function of number of group (m), where all scenarios were simulated from RIS models

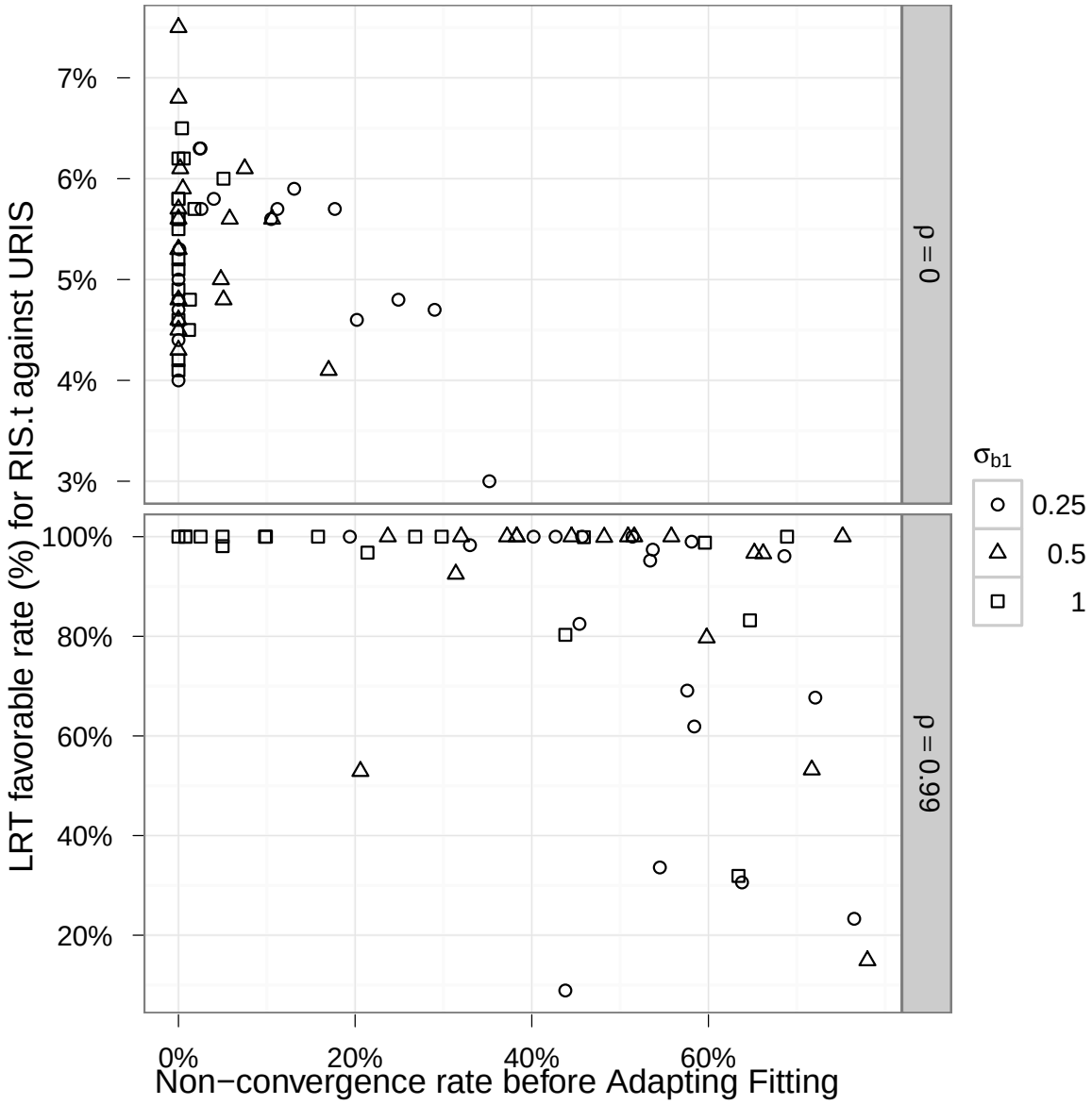

Figure 6.15: Scatter plots of the LRT favorable rate for the RIS.t model against the URIS.t model, as a function of non-convergence rate before Adapting Fitting for the RIS model in the original space across all 108 scenarios, stratified by random-effects correlation levels

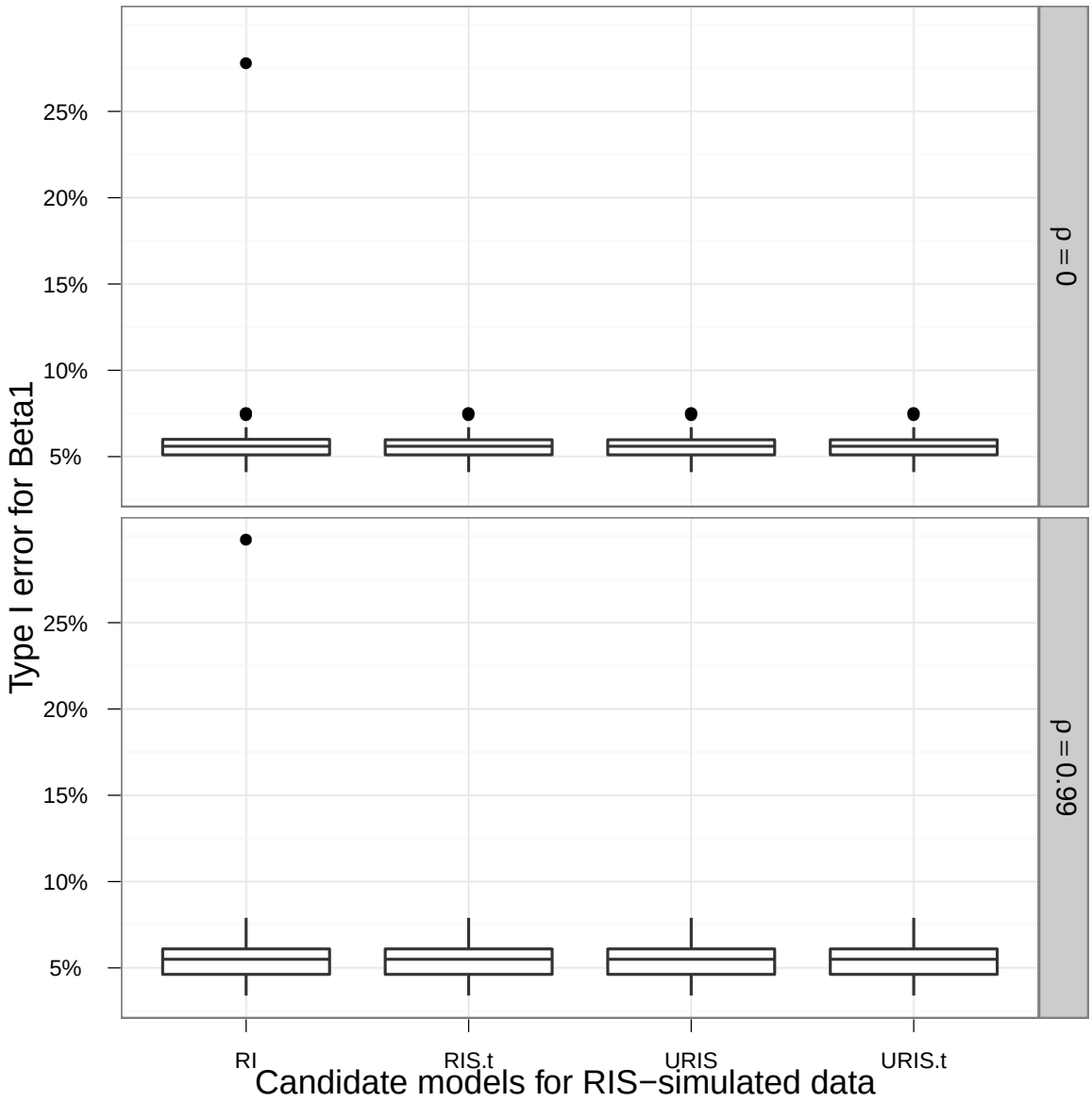

Figure 6.16: Boxplots of the Type I error for the nominal coverage of fixed-effect slope $beta_1$, for 3 models with both random intercept and slope (RIS.t, URIS, URIS.t) and 1 model without random slope (RI), stratified by random-effects correlation levels

Chapter 7

Application Examples

Two real data sets will be discussed in this chapter. In Section 7.1, we use IGF data to further demonstrate the performance of AF algorithm for a RIS model in the optimal transformed space, and to illustrate the model selection results involving the URIS.t model proposed in Chapter 6. The FEV_1 data in Section 7.2 is used to illustrate the feasibility of AF algorithm extended to multiple random-effects case described in Chapter 5.

7.1 A longitudinal data with non-convergence issue for RIS model: IGF data

Davidian and Giltinan (1995) described a data set obtained during quality control radioimmunoassays for radioactive tracer used to calibrate the Insulin-like Growth Factor (IGF-I) protein concentration measurements. The data contained 237 measurements from 10 lots during 1 to 50 days, each lot with 4 to 39 measurements (Fig. 7.1). The median and mean of collection time points were 22 and 22.45 days, respectively.

This publicly available IGF data is also included in the R package *nlme* and has been fitted by RIS, URIS and RI models (Pinheiro and Bates, 2000). It failed to converge using a RIS model either in the original or grand mean centered space, where the correlation estimates at last iteration were -1.000 and 1.000, respectively. Because the approximate 95% confidence intervals for the estimated correlations above by the RIS model are very

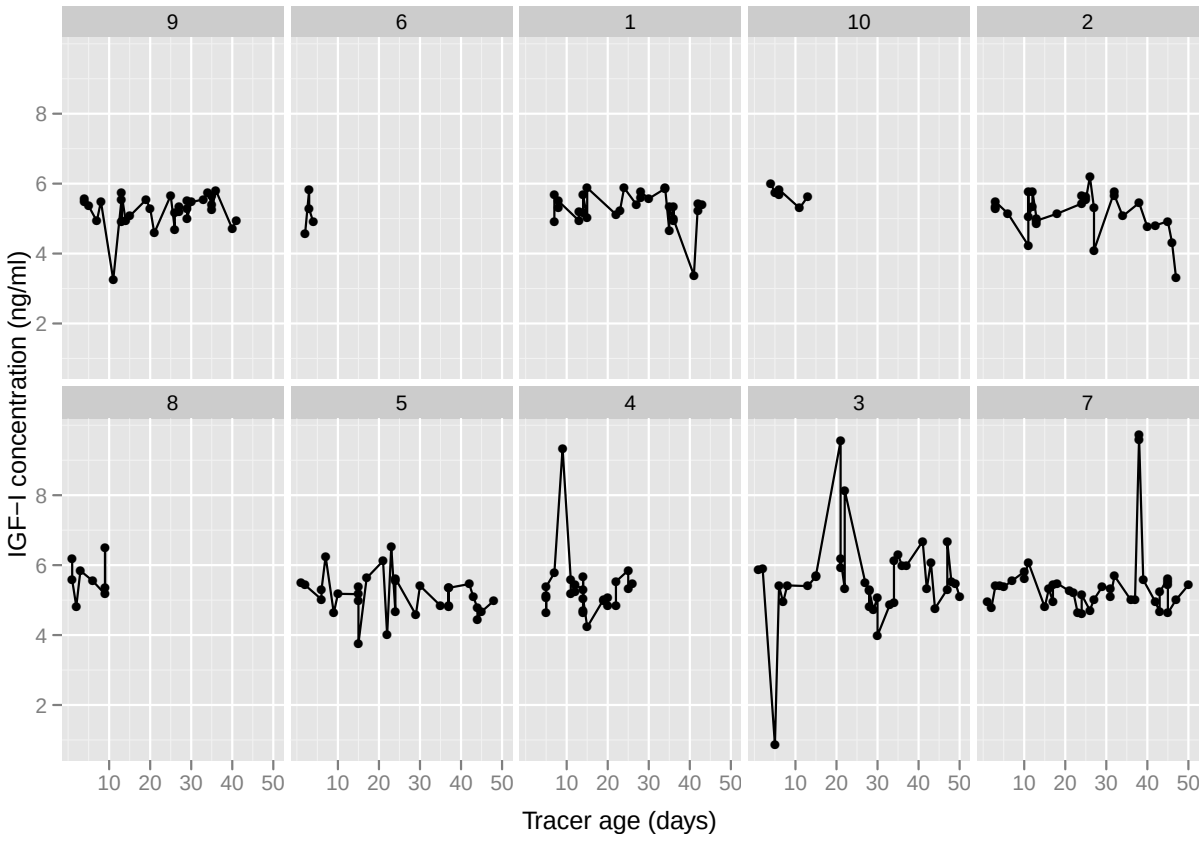

Figure 7.1: Time plot, with joined line segments, of IGF-1 concentration versus tracer age in days for all 10 lots from the IGF data set

liberal (e.g., $(-1.000, 1.000)$), the RI and URIS models have also been attempted.

In this section, we show the results of two proposed models which were both fitted in the optimal transformed space, one with unstructured random-effects (RIS.t model: last column on Table 7.1), the other with uncorrelated assumption between random-effects (URIS.t model: last column on Table 7.2).

7.1.1 RIS model fitting in the original, centering, and optimal transformed spaces

Table 7.1 summarizes the fitting results of IGF data for three models (RIS, RIS.c and RIS.t), which were obtained in the original, grand mean centering and optimal transformed spaces. For this specific data, centering could not alleviate the random-effects collinearity problem. The reversing sign of correlation is an indication of an over-shifting after centering. Compared to the original space, the centering results were slightly worse in terms of fitted log-likelihood, minimal eigenvalue and condition number.

Table 7.1: LMM fitting results of IGF data in the original, centering and optimal transformed spaces

Fitting results	Estimation space		
	Original	Centering	Optimal transformed
I. Diagnosis information			
Convergence	No	No	Yes
Positive definite of G*	No	No	Yes
Random-effects correlation	-1.000	1.000	-0.187
Minimal eigenvalue of G	-1.590E-11	-2.838E-12	1.961E-09
Condition number of G	4.309E+08	3.487E+09	3.335E+04
Condition number ratio relative to the transformed space	1.290E+04	1.046E+05	1
Log-likelihood	-297.1831	-297.1832	-297.1831
AIC	606.3663	606.3664	606.3662
BIC	627.1238	627.1239	627.1237
# of iterations	75	64	22
II. Parameter estimates at last iteration			
Beta1(standard error)	-0.002535 (0.005045)	-0.002535 (0.005044)	-0.002534 (0.005043)
Var(b1)	6.5495E-05	6.5493E-05	6.5401E-05
Var(b0)	6.7855E-03	9.8314E-03	2.0322E-09
Cov(b0,b1)	-6.6575E-04	8.0243E-04	-6.8174E-08
Var(noise)	0.6734	0.6734	0.6734

* Estimated random-effects covariance matrix

Among the three estimation spaces, the proposed transformed space produced the best fitting diagnosis results. For instance, the convergence was achieved with positive definite covariance matrix estimate, and the condition number was reduced more than 10,000 times under the optimal transformed space. After AF, the estimated correlation was relatively small (-0.187), and the AIC, BIC and number of Newton-Raphson iteration steps needed were also improved slightly.

Theoretically, it is invariant for the parameters of fixed effect slope β_1, random noise and random slope variance after a non-singular location shift linear transformation of RIS model, but the intercept-related parameters (e.g., random intercept variance and covariance) could change under different spaces. This theoretical expectation could be observed by comparing these parameter estimates at last iteration across three estimation spaces, although some of them were from non-converged fittings.

We note that the random-effects variance estimate was relatively small compared to the random noise, and the minimal eigenvalue of random-effects covariance matrix was negative but very close to zero in both the original and centering spaces. Thus, the non-PD issue in a non-optimal space could be due to larger round-off error during the nonlinear iteration process.

7.1.2 Proposed URIS.t model in the optimal transformed space

Table 7.2 summarizes the fitting results of three models (RI, URIS, URIS.t) for IGF data. All three models converged without non-PD issue. Compared to the results of the RIS.t model in Table 7.1, the uncorrelated assumption between random-effects had little impact on the URIS.t model fitting, but affected the URIS model fitting, in terms of log-likelihood, fixed effect slope β_1, random noise, and random slope variance. Among the three models in Table 7.2, together with the best model (RIS.t) in Table 7.1, the URIS.t model was most favored by AIC criterion, while the RI model was favored by BIC. In addition, the criterion of AIC also slightly favors the URIS model against the RIS.t model. The difference of fitted -2Log-likelihood values between the URIS.t and RI model is approaching to be significant based on LRT ($596.8771 - 594.3662 = 2.5049 < 2.71 = (0.5\chi^2_0 + 0.5\chi^2_1)_{\alpha=0.05}$).

Table 7.2: LMM fitting results of IGF data in the original space, without (RI) and with (URIS) random slope, and in the optimal transformed space with random slope (URIS.t)

Fitting results	Model (Estimation space)		
	RI (Original)	URIS (Original)	URIS.t (Transformed)
I. Diagnosis information			
Convergence	Yes	Yes	Yes
Positive definite of G*	Yes	Yes	Yes
-2Log-likelihood	596.8711	594.8006	594.3662
AIC	604.8711	604.8006	604.3662
BIC	618.7094	622.0985	621.6641
II. Parameter estimates			
Beta1(standard error)	-0.00082 (0.00397)	-0.00193 (0.00457)	-0.00253 (0.00504)
t-value of Beta1	-0.206	-0.422	-0.502
Var(b1)		2.89E-05	6.54E-05
Var(b0)	0.00512	1.31E-09	7.72E-10
Var(noise)	0.6889	0.6754	0.6735

* Estimated random-effects covariance matrix

7.2 Application to LMM model with three random-effects: FEV_1 data

The data for pulmonary function FEV_1 is publicly available (Fitzmaurice et al., 2004, pp. 210-216). The data contains a cohort of 299 girls who were born in or after 1967 and lived in Topeka, Kansas. Most girls were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were available annually until graduation from high school or loss to follow-up. The FEV_1 data consists of annual measurements of FEV_1, height and age, with a minimum of one and a maximum of twelve observations per girl over time. The means (medians) of height and age of total 1993 observations are 1.498 (1.540) and 12.568 (12.597) respectively.

Several LMMs have been fitted for this data using SAS Proc Mixed. The data set, related codes and outputs are available online from the website `http://biosun1.harvard.edu/~fitzmaur/ala/`. The candidate LMMs include a model with three random-effects, where two random slopes are age and log(height), denoted as RI2S model here. We use RI2S model for the FEV_1 data to illustrate the feasibility of AF algorithm in multiple random-effects cases. We are not claiming that RI2S model is the best LMM fitting for

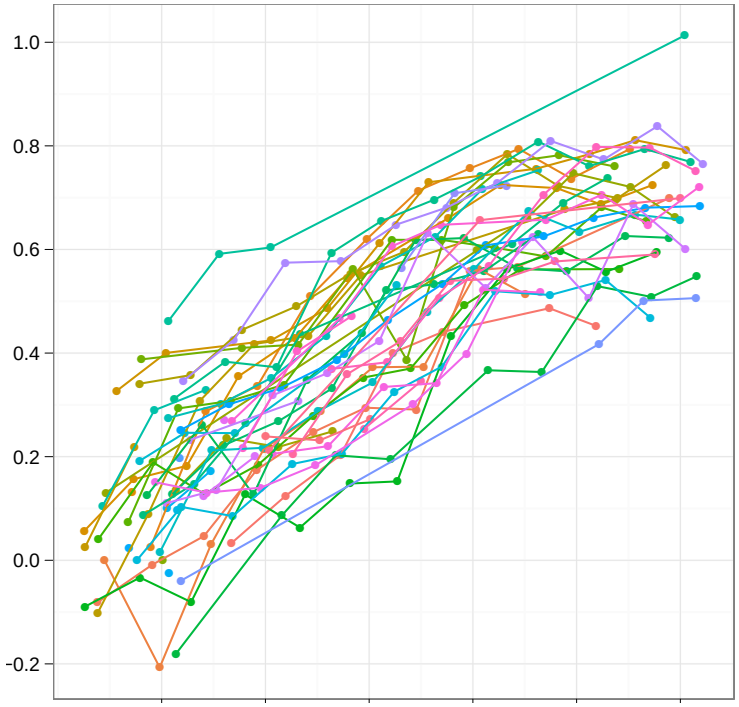

Figure 7.2: Time plot, with joined line segments, of log(FEV_1/height) versus age in years for 50 randomly selected girls from the FEV_1 data set

this specific data set. Actually, the RIS model with log(height) as random slope was the selected model (Fitzmaurice et al., 2004, pp. 216).

Based on software package R lme(nlme), we conducted similar analyses (including the data profile in Fig. 7.2) and obtained similar fitting results as shown in the book. For the RI2S model fittings for FEV_1 data (Table 7.3), the convergence was achieved for all three estimation spaces: original, centering and AF optimal transformed, with AF being the fastest among three estimation spaces. As expected, AF reduced the correlations between random intercept and each of two random slopes to zero level ($<$0.001). Under the optimal transformed space, the condition number was slightly reduced and the minimal eigenvalue was slightly larger, away from zero. This is not surprising since the correlation levels before AF were not high.

The estimates of log-likelihood, AIC, BIC, unexplained noise, and fixed-effects slopes were almost identical among the three estimation spaces. There were small variations (in third or more decimal point) for the estimates for the random-effects variances and the covariance of two random slopes.

Table 7.3: LMM fitting results of FEV_1 data in the original, centering and optimal transformed spaces

Fitting results	Estimation space		
	Original	Centering	Optimal transformed
I. Diagnosis information			
Convergence	Yes	Yes	Yes
Positive definite of G*	Yes	Yes	Yes
Random-effects correlation (Intercept, Age)	-0.165	-0.185	**0.001**
Random-effects correlation (Intecept, Log(Height))	-0.509	0.388	**-0.001**
Minimal eigenvalue of G	8.149E-06	1.013E-05	1.017E-05
Condition number of G	1.028E+04	8.034E+03	7.856E+03
Condition number ratio relative to the transformed space	1.31	1.02	1
Log-likelihood	2294.9500	2294.9500	2294.9500
AIC	-4565.8990	-4565.8990	-4565.8990
BIC	-4498.7610	-4498.7610	-4498.7610
# of iterations	64	36	27
time (sec)	29.45	21.32	12.16
II. Parameter estimates - Fixed-effects			
Slope - Age (standard error)	0.02344 (0.001278)	0.02344 (0.001278)	0.02344 (0.001278)
Slope - Log(Height) (standard error)	2.2476 (0.04692)	2.2476 (0.04692)	2.2476 (0.04692)
III. Parameter estimates - Random-effects			
Intercept	0.01337	0.00938	0.00795
Slope - Age	1.179700E-05	1.179600E-05	1.183000E-05
Slope - Log(Height)	0.07984	0.07984	0.07992
Cor(Age, Log(Height))	-0.373	-0.373	-0.374
Var(noise)	0.003516	0.003516	0.003516

Chapter 8

Discussion

Mixed-effects models have been widely used to model correlated data from many research fields. In practice it is important to alleviate non-convergence issues during the iterative optimization process. Non-PD may also be an issue and can be implicitly masked as a "nominally convergent" run by the default criteria of a software package. The proposed AF algorithm provides a straightforward technique to achieve these goals by reducing the collinearity between random-effects of LMM.

Simulations show that both the non-convergence rate and the non-PD rate are significantly reduced to zero level after AF, using the optimal linear transformation of the random slope variable. The AF procedure is generally effective across various settings, including those challenging scenarios with very high correlation, relatively large noise or small random-effects variance, and small sample size.

The computational advantages of AF algorithm are demonstrated by several measures. The reduction of random-effects correlation down to zero provides an intuitive measure of the benefits of AF. The smallest eigenvalue of random-effects covariance matrix in the optimal transformed space is increased and less susceptible to boundary issue. The condition number of random-effects covariance matrix is also generally reduced in the transformed space. Larger reduction of condition number coincides with a setting with high non-convergence rate, i.e., extremely high correlation and near-zero random slope variance.

The core idea of AF algorithm The core idea of our proposed AF algorithm is to utilize the existence of an optimal linear transformation where the resulting zero correlation is the farthest away from the parameter boundary, i.e., -1 or $+1$. Such a strategy is expected to further reduce the possibility for correlation estimate to converge near or out of the parameter boundary than that of explicitly direct reparameterization of model parameters, e.g., using some nonlinear transformations, such as log or logit, to force variance and correlation estimates "within" their domains (e.g., Pinheiro and Bates, 1996 for linear and non-linear mixed-effects models which had been implemented in the $lme(nlme)$ routine; Abellana et al., 2006 for generalized linear mixed-effects model in disease mapping setting). Such direct reparameterization methods can still encounter the convergence on the boundary issue if estimated random-effects are highly correlated. For example, our simulations show that there is only a 79.41% empirical rate to converge with PD covariance matrix across all simulation scenarios.

The feasibility of AF algorithm The AF approach is feasible for several reasons. First, conceptually, the existence of a neighborhood of the optimal shift d provides a working window for AF algorithm to reduce the random-effects correlation. We can expect that the AF algorithm has more tolerant to an non-accurate estimate of d if the underlying neighborhood is wider. In our simulations, non-convergent runs all had very high observed correlations between random-effects before AF even when the initial population correlation was zero. Thus both d and its neighborhood were determined mainly by the variance ratio. Second, by Lemma 3.4.1, when random-effects are extremely correlated, the random intercept variance is also expected to be large, thus very likely to be larger than the random slope variance. In real data analysis using a RIS model, smaller random slope variance relative to random intercept one was usually observed (e.g., Kreft et al., 1995; Verbeke and Lesaffre, 1997; Browne and Draper, 2000; Gurrin et al., 2001; Zhang and Davidian, 2001; Jacqmin-Gadda et al., 2007). Overall, a non-convergent run before AF will be strongly associated with a non-zero optimal shift, a wide neighborhood of the shift, and a large CN reduction potential after AF. Third, note that a location shift matrix A used in AF, even not optimal or estimated with some error, will not introduce extra computational error during the transformation of observed data, since $det(A) = det(A^{-1}) = 1$ and $CN(A) = CN(A^{-1}) = 1$ for any δ. Lastly, the AF algorithm can be straightforwardly implemented in existing LMM routines without modifying the internal optimization algorithm. The only needed coding effort in using AF is to extract the covariance matrix estimate from the

algorithm outputs and calculate the optimal shift d. For a non-convergent run, the estimate of optimal shift d can be available from the last iteration output. Even if a highly accurate estimate of d from previous challenging fitting may be not realistic, an approximate estimate can be adequate as long as the subsequent correlation estimate becomes smaller and is not near the parameter boundary.

AF algorithm compared to centering The proposed AF procedure differs from traditional grand mean centering technique in several important aspects. First, the random-effects correlation and the relative size of random-effects variances are taken into account by AF while centering only uses first-moment information. Centering is not a numerically optimal linear transformation for LMM. If the new origin of a slope variable after centering is not within the proposed optimal shift neighborhood, centering may even increase the correlation level between random-effects. Second, the optimal location shift $d = \rho\sigma_{b0}/\sigma_{b1}$ is a unit-free scalar. It cannot be directly calculated from the observed data itself before estimating a random-effects covariance matrix. Third, AF can be iterative while centering is a one step approach. Lastly, centering is a one-direction location shift method and the new origin is always within the observed range of slope covariate. However, as pointed out by Longford (1993), the direction of an location shift can be either positive or negative, and even beyond the range of the observed raw data. In real longitudinal data analysis, the calculated value of d may be close to the observed mean of time variable, e.g., for a fetal growth data (Gurrin et al., 2001), or the two may not be close to each other as in the IGF data in section 7.1. In our simulations, it is possible for the new origin of slope covariate to go beyond the range of the data during numerical optimization process, especially when the random intercept variance is larger than the random slope variance in high correlation settings. Since the AF algorithm focuses on the numerical optimization process rather than the inference of model parameters, this is acceptable. If a run can converge both before and after AF, the parameter estimates in the original space will be almost identical.

The implications of AF algorithm for RIS models The numerical optimization of a RIS model can be much more challenging than that of a RI model. When a RIS model fails to converge due to highly correlated random-effects, it is a common practice to remove random slope from the RIS model and fit a RI model. However, the simple compound symmetric covariance structure and the constant correlation assumption of RI model may be unrealistic

or not adequate for actual data modelings (Verbeke et al., 1998). RIS model is probably the most standard tool to study changes over time in longitudinal growth curve modeling, as it allows to model non-stationary covariance structure and to investigate the growth velocity variations across different subjects using subject-specific random slope estimates. Several studies have showed that the maximum likelihood inference on fixed-effects is more robust to mis-specification of the covariance structure under a RIS model, compared to a RI model (Lange and Laird, 1989; Jacqmin-Gadda et al., 2007; Schielzeth and Forstmeier, 2009). Thus, the excellent convergence property of AF algorithm is useful not only for simulation studies but also for real data analyses.

The proposed URIS.t model has been shown to have several advantages over other competing models, in terms of log-likelihood and AIC criteria. Selecting a model that is too parsimonious has a more severe impact on Type I error rate than selecting a model that is too complex. Our model comparison results based on the fitting from R(nmle) are consistent with those obtained from SAS Proc Mixed (Guerin and Stroup, 2000).

The general application of AF algorithm The AF procedure can be considered as a general purpose and efficient algorithm for mixed-effects models with correlated random-effects, including the multiple random-effects case. Although Newton-Raphson algorithm is the main iterative optimization tool used in our simulations, EM algorithm is also found to be faster after AF based on our limited simulations (results not shown). Furthermore, the linear transformation used by the AF procedure does not impose any distribution assumption and thus can be generalized to other mixed-effects models, e.g., with non-normality error or binary outcome, where the convergence can be even more difficult. We recommend that the AF procedure should be considered as a routine collinearity diagnostic and sensitivity analysis tool during the fitting of mixed-effects models, especially when there are reasons to suspect that convergence on the boundary issue is present.

Limitation and future work There are several potential limitations for the current study. First, whether a non-convergent run achieves convergence after AF can be influenced by the magnitude of the calculated optimal shift d. If the automatically estimated shift is at zero level, the AF procedure cannot further improve the optimization and thus stop with non-convergence. This situation occurred in only two runs in our simulations. If manually introducing a small (e.g., less than one) location shift once or twice, these two failed runs

could also converge.

Second, the performance of AF algorithm may not be the same across different software packages when the covariance matrix estimate is near-singular or non-PD. Some packages may force the variance or covariance estimate to be zero, or do not provide the random-effects covariance matrix estimate, thus the optimal shift cannot be calculated from the output.

Third, it is worthwhile to conduct similar and more extensive simulation studies comparing the 6 candidate models discussed in Chapter 6 using data simulated from various models. Using a RIS model to fit a RI-simulated data is expected to be much more difficult than using a RI model to a RIS-simulated data.

Lastly, the number of iteration required and the performance of AF procedure need to be further evaluated for various challenging settings for LMM, such as unbalanced data structures and very high dimension of random-effects, and for generalized mixed-effects models. The sensitivity study of AF algorithm relative to large change of origin of random covariate can also be examined by simulations.

Bibliography

Abellana, R., Carrasco, J., Jover, L., Ascaso, C., 2006. Improving the convergence rate in conditional autoregressive models. Computational Statistics & Data Analysis 50, 1153–1163.

Belsley, D.A., Oldford, R.W., 1986. The general problem of ill conditioning and its role in statistical analysis. Computational Statistics & Data Analysis 4, 103–120.

Berkhof, J., Snijders, T., 2001. Variance component testing in multilevel models. Journal of Educational and Behavioral Statistics 26, 133–152.

Bradley, R.A., Srivastava, S.S., 1979. Correlation in polynomial regression. American Statistician 33, 11–14.

Browne, W.J., Draper, D., 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Statistics 15, 391–420.

Callanan, T.P., Harville, D.A., 1991. Some new algorithms for computing restricted maximum likelihood estimates of variance components. Journal of Statistical Computation and Simulation 38, 239–259.

Chen, W.W., Hurvich, C.M., Lu, Y., 2006. On the correlation matrix of the discrete fourier transform and the fast solution of large toeplitz systems for long-memory time series. Journal of the American Statistical Association 101, 812–822.

Davidian, M., Giltinan, D.M., 1995. Nonlinear Models for Repeated Measurement Data. Chapman and Hall, London.

Dempster, A.P., Selwyn, M.R., Patel, C.M., Roth, A.J., 1984. Statistical and computational aspects of mixed model analysis. Applied Statistics 33, 203–214.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis. Wiley-Interscience. 3rd edition.

Ferron, J., Dailey, R., Yi, Q., 2002. Effects of misspecifying the first-level error structure in two-level models of change. Multivariate Behavioral Research 37, 379–403.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2004. Applied Longitudinal Analysis. Wiley-Interscience, USA. Http://biosun1.harvard.edu/ fitzmaur/ala/.

Goldstein, H., 1986. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73, 43–56.

Goldstein, H., 2002. Multilevel Statistical Models. Wiley. 3rd edition.

Grenander, U., Szego, G., 2001. Toeplitz Forms and Their Applications. AMS Chelsea, New York. 2nd edition.

Guerin, L., Stroup, W., 2000. A simulation study to evaluate proc mixed analysis of repeated measures data, in: Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture, Kansas State University, Manhattan, KS.

Gurrin, L., Blake, K., Evans, S., Newnham, J., 2001. Statistical measures of foetal growth using linear mixed models applied to the foetal origins hypothesis. Statistics in Medicine 20, 3391–3409.

Hwang, Y.T., Wei, P.F., 2006. A novel method for testing normality in a mixed model of a nested classification. Computational Statistics & Data Analysis 51, 1163–1183.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.M., Thiebaut, R., 2007. Robustness of the linear mixed model to misspecified error distribution. Computational Statistics & Data Analysis 51, 5142–5154.

Jennrich, R., Schluchter, M., 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42, 805–820.

Kreft, I.G.G., de Leeuw, J., Aiken, L., 1995. The effect of different forms of centering in hierarchical linear modeling. Multivariate Behavioral Research 30, 1–20.

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

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Lange, N., Laird, N.M., 1989. The effect of covariance structure on variance-estimation in balanced growth-curve models with random parameters. Journal of the American Statistical Association 84, 241–247.

Lindstrom, M.J., Bates, D.M., 1988. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 83, 1014–1022.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., Schabenberber, O., 2006. SAS for Mixed Models. SAS Publishing. 2nd edition.

Liu, C., Rubin, D.B., 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika 81, 633–648.

Longford, N.T., 1987. A fast scoring algorithm for maximum-likelihood-estimation in unbalanced mixed models with nested random effects. Biometrika 74, 817–827.

Longford, N.T., 1993. Random Coefficient Models. Oxford: Clarendon Press.

Marquardt, D.W., Snee, R.D., 1975. Ridge regression in practice. American Statistician 29, 3–20.

Meng, X.L., van Dyk, D., 1998. Fast EM-type implementations for mixed effects models. Journal of the Royal Statistical Society Series B-Statistical Methodology 60, 559–578.

Mikulich, S.K., Zerbe, G.O., Jones, R.H., Crowley, T.J., 1999. Relating the classical covariance adjustment techniques of multivariate growth curve models to modern univariate mixed effects models. Biometrics 55, 957–964.

Morrell, C.H., Pearson, J.D., Brant, L.J., 1997. Linear transformations of linear mixed-effects models. American Statistician 51, 338–343.

Murphy, D.L., Pituch, K.A., 2009. The performance of multilevel growth curve models under an autoregressive-moving average process. Journal of Experimental Education 77, 255–282.

Pinheiro, J.C., Bates, D.M., 1996. Unconstrained parametrizations for variance-covariance matrices. Statistics and Computing 6, 289–296.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

Pinheiro, J.C., Bates, D.M., DebRoy, S., Sarkar, D., R Development Core Team, 2009. nlme: Linear and Nonlinear Mixed Effects Models. R 2.9.2.

Pryseley, A., Tchonlafi, C., Verbeke, G., Molenberghs, G., 2011. Estimating negative variance components from gaussian and non-gaussian data: A mixed models approach. Computational Statistics & Data Analysis 55, 1071–1085.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications. 2nd edition.

Schielzeth, H., Forstmeier, W., 2009. Conclusions beyond support: overconfident estimates in mixed models. Behavioral Ecology 20, 416–420.

Sengupta, D., Bhimasankaram, P., 1997. On the roles of observations in collinearity in the linear model. Journal of the American Statistical Association 92, 1024–1032.

Shieh, Y., Fouladi, R., 2003. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. Educational and Psychological Measurement 63, 951–985.

Solaro, N., Ferrari, P.A., 2007. Robustness of parameter estimation procedures in multilevel models when random effects are MEP distributed. Statistical Methods & Applications 16, 51–67.

Thompson, R., Meyer, K., 1986. Estimation of variance components: What is missing in the EM algorithm. Journal of Statistical Computation and Simulation 24, 215–230.

Trefethen, L.N., Bau, D., 1997. Numerical Linear Algebra. SIAM: Society for Industrial and Applied Mathematics.

van der Leeden, R., Vrijburg, K., de Leeuw, J., 1996. A review of two different approaches for the analysis of growth data using longitudinal mixed linear models. Computational Statistics & Data Analysis 21, 583–605.

van Dyk, D.A., 2000. Fitting mixed-effects models using efficient EM-type algorithms. Journal of Computational and Graphical Statistics 9, 78–98.

Verbeke, G., Lesaffre, E., 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Computational Statistics & Data Analysis 23, 541–556.

Verbeke, G., Lesaffre, E., Brant, L.J., 1998. The detection of residual serial correlation in linear mixed models. Statistics in Medicine 17, 1391–1402.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

West, B.T., Welch, K.B., Galecki, A.T., 2006. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman and Hall/CRC, Boca Raton, FL.

Yuan, K.H., Chan, W., 2008. Structural equation modeling with near singular covariance matrices. Computational Statistics & Data Analysis 52, 4842–4858.

Zhang, D., Davidian, M., 2001. Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics 57, 795–802.