The Power of a Few Large Blocks: A credible assumption with incredible efficiency

Dongyu Lin and Dean P. Foster*

Abstract

The most powerful assumption in data analysis is that of *independence*. Unfortunately, it is almost impossible to check, so it often must stay as an assumption. One way out of this trap is to assume less independence than usual: Instead of assuming all n observations are *i.i.d.* (independently and identically distributed), it is often more credible to assume that several blocks of the data are independent of each other. This gain in credibility, however, might come with a loss of statistical power compared to an analysis that correctly assumed (if it did) the full *i.i.d.* assumption. The main goal of this paper is to address this trade-off. We show that in the true *i.i.d.* case, little power is lost if the weaker block independence is assumed, as long as there are at least a handful of independent blocks; on the other hand, if block-dependent data is treated independently, the risk of making type I errors can be severe. Hence, we advocate using the weaker block independence assumption rather than the stronger full independence assumption, if the former leads to a gain in credibility.

KEY WORDS: Data dependence; Robustness; t-test; Test power; Type I error.

*Dongyu Lin is Postdoctoral Fellow, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 (E-mail: *dongyu@wharton.upenn.edu*). Dean P. Foster is William H. Lawrence Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *dean@foster.net*).

1 Introduction

The motivation of this paper arose from a lunchtime discussion: Adi described a data set he had seen analyzed in a legal case. A company had claimed their motor oil helped the gas mileage of a car. They showed many measurements of gas mileage with and without the use of this oil, using different cars, different drivers, under different driving conditions. All in all, there were hundreds of observations. Under the classical *i.i.d.* assumption about the error structure, the two-sample t-statistic for oil was significant. It seemed like a clear win for the company justifying its claim.

However, with further discussion, Adi pointed out that there were only four different cars used in the whole study. Furthermore, the car variable had a very significant effect. Thus, an alternative analysis would be to treat each car as an independent observation and then perform a one-sample t-test with only four mileage differences obtained for each car. This, like most discussions here at Penn, lead to a long argument. Were there hundreds of observations, or were there four observations? We will call these two models the full *i.i.d.* model and the blocks/cars model hereafter. Clearly, the company would like to argue for the full *i.i.d.* model for its benefit of higher power, whereas the prosecution wants to use the cars model for its control of type I error.

This paper will show there is a middle ground. Such a middle ground would have to analyze the data so that both sides of the legal battle will believe the result under their own preferred model. For the company, it would have to have about as high power as their full *i.i.d.* analysis. For the prosecution, it would have to control the type I error if the cars model is correct. We will argue that such a middle ground already exists: a blockwise t-test. In particular, we will show that the loss of power of using the blockwise t-test when the full *i.i.d.* model is correct is small. Furthermore, it has the right size under the cars model. Thus, both parties should accept this the answers generated by this test.

This kind of arguments is also common in group randomized trials (Cornfield, 1978) in public health research. In such study designs, interventions are operated at a school, hospital, or community level. The number of groups is typically small in most studies (Murray *et al.*, 2004). Since treatment policies differ at each site, the "errors" for each site can be correlated. Therefore, similar to the cars model, group randomized trials are characterized by two features: small number of groups and positive intra-cluster correlations. Lack of attention to these issues would lead to "underpowered" designs (not enough data collected) and "overstated" power (Feng *et al.*, 2001).

Longitudinal data also typically behaves this way (Diggle *et al.*, 2002). In the COPSAC$_{2000}$ (Copenhagen Studies on Asthma in Childhood) birth cohort study (Bisgaard, 2004), medical researchers recruited 411 eligible children born to asthmatic mothers between 1998 and 2001 in order to investigate complex effects in the origin of atopic diseases. The children were assessed at the research unit regularly and have sequential records of follow-up diagnoses. The data is thus composed by $K = 411$ groups of correlated observations. In reality, due to complicated situations, for example, deficient cases as in studies of rare disease, or data validity issues such as data missing and measurement errors, the effective number of independent groups can be small in the observed longitudinal data.

In biostatistics there are already common techniques for analyzing these cars model like data. Common modeling includes treating the within group correlation using random effects (Laird and Ware, 1982). Alternatively, we can avoid having to model them if we treat the entire group as an observational unit. Since we have independence between groups, this blocking structure can then be used to generate an efficient estimator. Correct analysis of this sort of data started with Liang and Zeger (1986). But typically the sandwich estimator in Liang and Zeger (1986) requires a relatively large number of groups. Small sample corrections are usually not easy to compute (Fay and Graubard, 2001; Mancl and DeRouen, 2001). We will argue that using the simple blockwise t-test we advocate, even if there are only a small number of groups, there is very little power loss compared to making the bold assumption that each patient is independent from all others.

In the rest of the paper, we will show that little power is lost when we use blocks instead of a full *i.i.d.* model (Section 2). Alternatively, if the full *i.i.d.* model is wrong, it might make a large type I error (Section 3). So we argue using a blocks model if it there is any doubt about the full *i.i.d.* model. We then compare our simple approach with linear mixed effect models using standard software in Section 4 and conclude with some discussions in Section 5.

2 Comparisons of Power

In this section, the data will always be generated using the full *i.i.d.* model. We primarily analyze this data using an estimator that is designed for block independence. We then compare the power of this test statistic to that of the classical one which in this case has correctly made the full *i.i.d.* assumption.

Notationally we need blocks, but the true model, that we have is n *i.i.d.* observations from a normal population with mean μ and variance σ^2, ignores the blocks. To make it consistent, instead of the usual notation:

$$Y_1, \ldots, Y_n \sim_{i.i.d.} N(\mu, \sigma^2)$$

which draws attention to the full *i.i.d.* structure, we change the names of the random variables and use the following:

$$Y_{1,1} \cdots Y_{1,n_1}\ ,\ \ldots\ ,\ Y_{i,1} \cdots Y_{i,n_i}\ ,\ \ldots\ ,\ Y_{K,1} \cdots Y_{K,n_K} \quad \sim_{i.i.d.} \quad N(\mu, \sigma^2)$$

where $\sum_{i=1}^{K} n_i = n$. Regardless of which notation is used, the maximum likelihood estimator (MLE) for μ on this data is simply the grand mean $\overline{Y}$, namely the sum of all the $Y_{i,j}$'s divide by total number of observations n.

The blocks model supposes that $Y_{i,j}$'s with difference indices i are independent of each other, but that if the i's are the same there may be a correlation between the $Y_{i,j}$'s; so $Y_{1,1}$ and $Y_{3,1}$ are independent, but $Y_{1,1}$ and $Y_{1,3}$ are correlated. The blocks model makes no claims as to what this correlation is and only assumes:

$$Y_{i,j} \sim N(\mu, \sigma_{ij}^2), \quad j = 1, \ldots, n_i, i = 1, \ldots, K. \tag{1}$$

with $\mathrm{cov}(Y_{i,j}, Y_{i,k})$ being allowed to be non-zero. In total, there are $\sum_{i=1}^{K} \binom{n_i}{2}$ covariances that it would allow to be non-zero. Hence, it is a much larger null model than the *i.i.d.* model which assumes that all of these off diagonal covariances are exactly zero. Even a random effects model will often assume the variance-covariance structure of each block to be the same and model the structure by way of reducing these $\sum_{i=1}^{K} \binom{n_i}{2}$ parameters to only a few ones.

Now we want to know whether or not $H_0 : \mu = \mu_0$ holds. In the full *i.i.d.* model where $Y_{i,j} \sim_{i.i.d.} N(\mu, \sigma^2)$, we may utilizes the classic one-sample t-test with statistic

$$T_{\text{iid}} = \frac{\overline{Y} - \mu_0}{SE_{\text{iid}}}. \tag{2}$$

Here $\overline{Y}$ is the grand mean and SE_{iid} is the classical standard error of $\overline{Y}$, namely $SE_{\text{iid}} = s/\sqrt{n}$, where s is the sample standard deviation of all $\{Y_{i,j}\}$. The test statistic T_{iid} has a student's t-distribution with $n - 1$ degrees of freedom under the normality assumption.

If one admits that there is dependence within a block, one might be tempted to model this depedence. An alternative and simple approach is to consider looking at the $\overline{Y}_{i\cdot}$'s, where $\overline{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{i,j}/n_i$ is the average of the ith block. These averages, by our block independence assumption, are indepedent and thus are more amenable classical analysis. For example, to test whether their mean is μ_0 or not, we may use the statistic

$$T_{\text{block}} = \frac{(\sum_{i=1}^{K} \overline{Y}_{i\cdot})/K \; - \; \mu_0}{SE_{\text{block}}}, \tag{3}$$

where the $SE_{\text{block}} = s_{\overline{Y}_{i\cdot}}/\sqrt{K}$, and $s_{\overline{Y}_{i\cdot}}$ is the standard deviation of $\{\overline{Y}_{i\cdot}\}$. This leads to a t-distributed statistic with degrees of freedom $K - 1$ under the normality assumption.

One advantage of looking at $\{\overline{Y}_{i\cdot}\}$ is that we can analyze the properties using the so-called self-normalized sums which make few distributional assumptions. If the distributions of $Y_{i,j}$'s are unknown, then the block model only informs us that the $\overline{Y}$'s are IID. Since we might have as few as two block, using the central limit theorem here would raise eyebrows. But if we can add just a bit more, powerful theorems from about self normalized sums (for example Bertail *et al.* (2008)) come to the rescue. If the $\overline{Y}$'s can be modelled as being symetric, then we get the following theorem:

Theorem 1. *(Self-normalized sums theorem, Bertail et al., 2008) Under the null hypothesis $H_0 : \mu = 0$, if $\overline{Y}_{i\cdot}$'s are symmetric, then for $K > 1$:*

$$P\left(\frac{\left|\sum_{i=1}^{K} \overline{Y}_{i\cdot}\right|}{\sqrt{\sum_{i=1}^{K} \overline{Y}_{i\cdot}^2}} \geq t\right) \leq 2\exp\left(-\frac{t^2}{2}\right),$$

He also has a statement under a 4th moment assumption instead of under symmetry.[1]

In this paper, we will take the low road and simply assume normality. The statistics T_{block} (3) thus has a t-distribution with degrees of freedom $K - 1$.

We will compare the power of these two test statistics and investigate the power loss when the more credible blocks model is assumed. Define $1 - \beta_{\text{iid}}(\mu)$ to be the power of using $T_{\text{iid}} \sim t_{n-1}$ and $1-\beta_{\text{block}}(\mu)$ to be the power of using $T_{\text{block}} \sim t_{K-1}$. To see how much these two power functions differ from each other, shown below are several plots on the power difference

$$\beta_{\text{block}}(\mu) \; - \; \beta_{\text{iid}}(\mu). \tag{4}$$

We compare this power loss with different test levels α, block sizes n_0 and numbers of blocks K. Without loss of generation, we assume $\sigma = 1$. On this scale, we will show values of the power loss as a function in μ varying from 0 to 2 .

In the first two figures, we compare the power loss at test level $\alpha = 0.05$. In Figure 1, the number of observations per block, n_0, equals 10. Clearly, if there is only one block, the SE_{block} for the blocks model has no ability to estimate the variability. When there are only two block, namely $K = 2$, there can be an power loss up to an 86%. Since the maximal loss is 95% this is not doing very well. However, the power loss decreases fast as K increases. When $K = 8$, $\beta_{\text{block}}(\mu) \; - \; \beta_{\text{iid}}(\mu) < 0.1$ for all μ. In fact, the supremum of power losses over all μ,

$$\sup_{\mu} \; \{ \; \beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu) \; \}$$

[1] He continues (2) In general, if the kurtosis of $\overline{Y}_{i\cdot}$'s, $\gamma_4 < \infty$, for any $a > 1$,

$$P\left(\frac{\left(\sum_{i=1}^{K} \overline{Y}_{i\cdot}\right)^2}{\sum_{i=1}^{K} \overline{Y}_{i\cdot}^2} \geq t\right) \leq 2\exp\left(1 - \frac{t}{2(1+a)}\right) + \exp\left(-\frac{K}{2\gamma_4}\left(1 - \frac{1}{a}\right)^2\right).$$

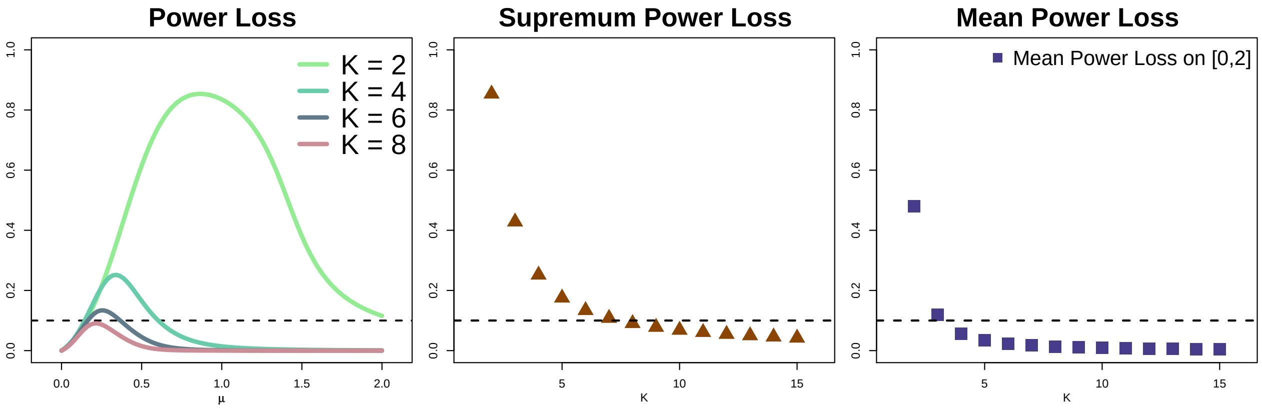

Figure 1: Power loss at test level $\alpha = 0.05$ with $n_0 = 10$. *Left*: Power losses when the number of independent blocks $K = 2, 4, 6$ and 8. When $K \geq 6$, not much power is lost at most μ's. *Middle*: The supremum of power losses for each K, $2 \leq K \leq 15$. This maximum power loss is less than 0.1 when $K \geq 8$. *Right*: The average power loss over $\mu \in [0, 2]$ is generally small except in the extreme case of $K = 2$.

decays exponentially as shown in the middle panel of Figure 1. On average, the power loss is fairly small as long as we have more than three independent blocks.

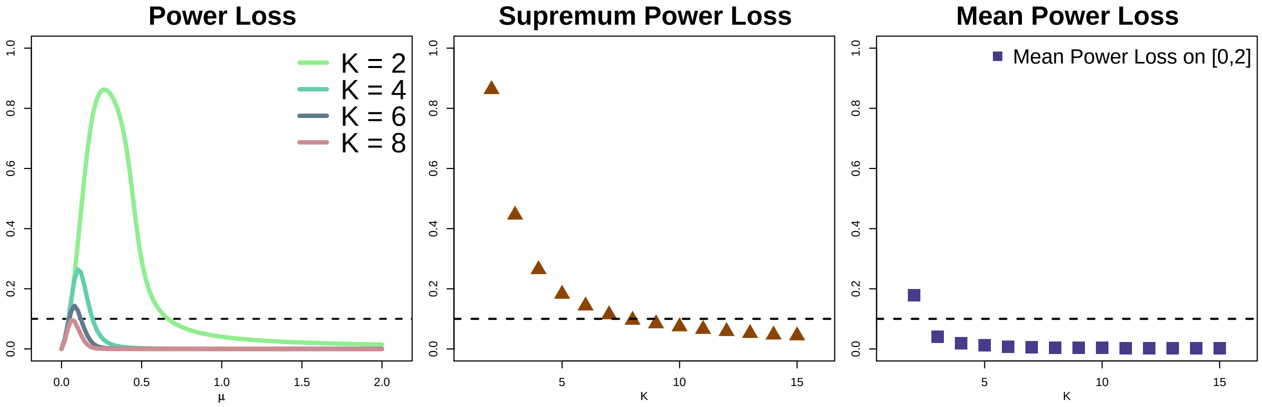

Figure 2: Power loss at test level $\alpha = 0.05$ with $n_0 = 100$. *Left*: For most μ the lost power does not exceed 0.1. *Middle*: The supremum power loss for each K, $2 \leq K \leq 15$. *Right*: The average power loss over $\mu \in [0, 2]$ is small in general.

Figure 2 tells a similar story when the block size $n_0 = 100$. A more pronounced fact revealed by this graph is that although the supremum power loss $\sup_\mu\{\beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu)\}$ will not be less than 0.1 until $K \geq 8$, most μ's have $\beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu) < 0.1$ even when K is small. This can be also seen from the right panel of Figure 2: except the most extreme case when $K = 2$, the average power loss is ignorable.

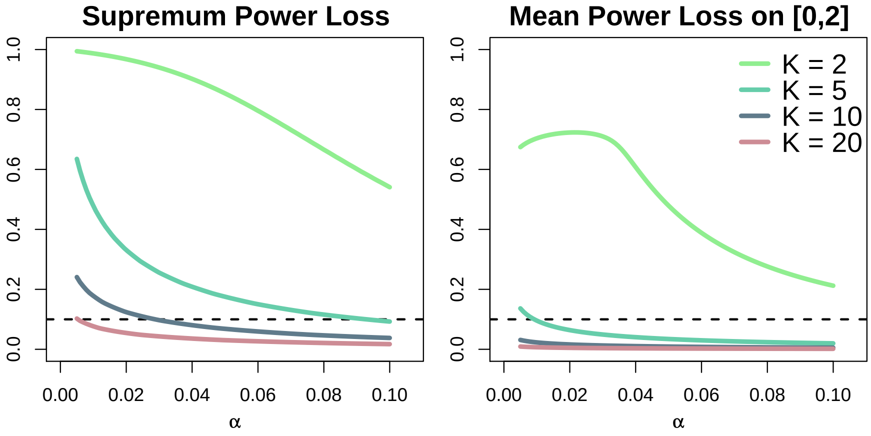

Figure 3: Power loss when α varies from 0.005 to 0.10. The block size $n_0 = 10$. *Left*: Although up to more than 50% power is lost at $K = 2$, the supremum power loss alleviates quickly as K gets larger. When $K = 10$, the maximum power loss goes less than 0.1 at test levels $\alpha \geq 0.03$. *Right*: The average power loss over $\mu \in [0, 2]$ is small and negligible except the extreme case with only two blocks.

To control the power loss, we will need more independent blocks when the test levels are small. In Figure 3 we display the decay speed of the supremum power loss $\sup_\mu\{\beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu)\}$ at the test levels varying from 0.005 to 0.10. Note that these are *the worst cases*, similar patterns hold as in Figure 1 and 2 that the power loss is very small at must μ, which is also implied by the right panel, a plot of the mean power loss $\text{avg}_{\mu\in[0,2]}\{\beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu)\}$ at different test levels.

As expected, the supremum power loss decreases slowly when K is extremely small, say 2. Even at $\alpha = 0.1$, $\sup_\mu\{\beta_{\text{block}}(\mu) - \beta_{\text{iid}}(\mu)\} > 0.5$. But the decay trace shows a fast decay pattern when K is slightly larger. When $K = 10$, at most 21% power is lost at level $\alpha = 0.005$ but no more than 10% power is lost when $\alpha \geq 0.03$.

In summary, not much power is lost in general except in the most extreme case $K = 2$; at the common test level $\alpha = 0.05$, very little power is lost at *every* μ when the number of independent blocks $K \geq 8$; on average, approximately with $K \geq 4$ we will not lose detectable power; for most test levels α, a medium K can guarantee the maximum power loss not exceeding 0.1.

3 Type I Errors

The previous section showed that using the block model based test maintained about as much power as the *i.i.d.* based test. The dual question then is, how well does the *i.i.d.* based test perform in the presence of dependence? This section addresses this question from the perspective of type I error. We will show that for some dependence structures the size of the *i.i.d.* based test is not close to the target of $\alpha = .05$.

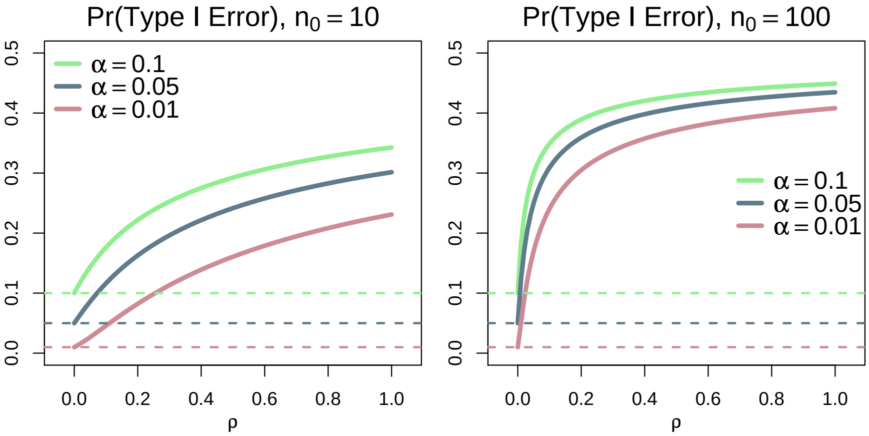

Figure 4: The true probability of type I error. *Left*: Block size $n_0 = 10$. In general, these probabilities are much larger than the designed levels of 0.10, 0.05 and 0.01. *Right*: Block size $n_0 = 100$. More information on dependence being lost, the probabilities of type I error jump up to a high level quickly. There is at least 30% chance of making type I errors even with a small $\rho > 0$, if the wrong test procedure has been taken.

Assume that the blocks model (see equation (1)) holds and with fixed block sizes, so

$n_i \equiv n_0$. Furthermore, assume

$$\text{Corr}(Y_{i,j}, Y_{i,k}) = \rho, \quad 0 \le \rho \le 1, \tag{5}$$

for any $1 \le j \ne k \le n_0$. As shown in Figure 4, the size of the test, namely the probability of making type I error, is often much larger than the designed test levels. For high correlation, the $\alpha(\rho)$ is worse–this makes sense since the higher the correlation the further we are from independence. Likewise as the number of observations in a block increases the α gets worse. This is because the more observations there are in a block the larger the divergence between the *i.i.d.* model and blocks model. This substantially increases the chance of making type I errors, as shown in the right panel of Figure 4, where the block size $n_0 = 100$. Even with a mild correlation ρ, the chance of making type I errors is large.

In a word, the price of ignoring the dependence is very expensive.

4 Comparison with Linear Random Effect Model

The *i.i.d.* model assumes that the $\sum_{i=1}^{K} \binom{n_i}{2}$ correlations are exactly zero whereas the blocks model completely unconstrains them. A compromize between these two extremes is to assume a model which allows for some correlations, but not any possible correlation. For example, the illustrative model in Section 3 we used has the same correlation structure as linear random effect models (Laird and Ware, 1982). Specifically, model (1) can be also modeled via a random effect model:

$$Y_{i,j} = \mu + \gamma_i + \epsilon_{ij} \tag{6}$$

where γ_i and ϵ_{ij} have mean 0 and variance τ^2 and σ^2, respectively. If ϵ_{ij}'s are *i.i.d.* with variance $1 - \rho$, and α_i are *i.i.d.* with $\tau^2 = \rho$, it reduces to the equal intra-cluster correlation case specified in (5). It is thus of interest to compare testing using the simple t-test with testing using standard implementations for random effect models.

The `lme` function in R package `nlme` is one of the popular functions that conveniently implements linear mixed effect models. It has been widely used in medical research (Fitzmaurice *et al.*, 2008) and educational research (Lockwood *et al.*, 2003). For testing fixed effects, it has been noticed in the literature (See e.g., Pinheiro and Bates, 2002; Faraway, 2005) that likelihood ratio tests based on maximum likelihood estimations tend to be anti-conservative and has smaller p-values. Permutation tests are sometimes recommended. Pinheiro and Bates (2002, Section 2.4) has thorough comparisons of the degrees of freedoms of test statistics and suggests that the conditional F or t test based on provide more reliable results. We thus compare the p-values of the conditional t-test for intercepts provided in `lme` with our proposed t-test.

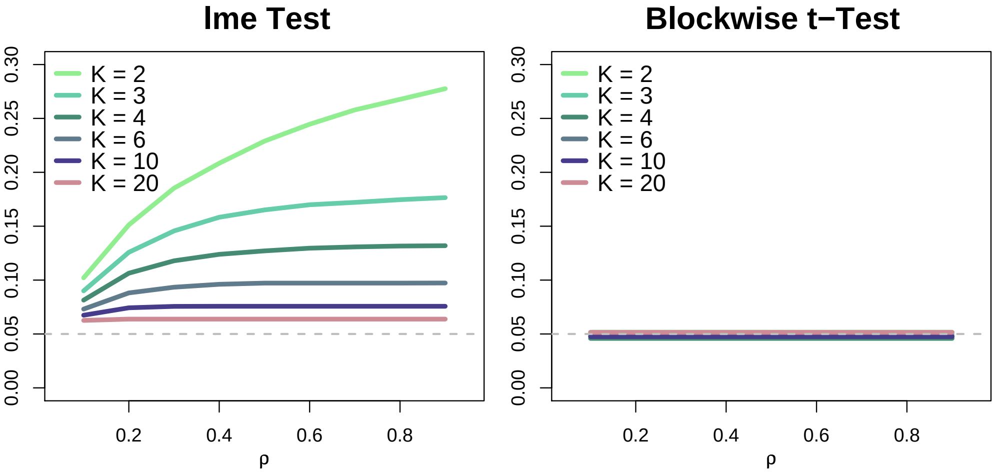

Figure 5: Estimated type I error rates based on 10,000 simulation replications. We simulated data sets from model (1) with $\mu = 0$ and $n_0 = 10$. The intra-cluster correlation ρ varied from 0.1 to 0.9, and the number of blocks K varied from 2 to 20. We then performed a two-sided test for the $H_0 : \mu = 0$ using the R lme function and the block t-test proposed in Section 2. The nominal test size is 0.05. For the lme test, we used the p-value for the fixed effect provided by the function. For both tests, the estimated type I error rates were computed as the proportion of rejections among the 10,000 replications. *Left*: The type I errors are in general inflated and severely inflated when the intra-cluster correlation is large and the number of independent blocks is small. When $K = 20$, the effect from ρ is ignorable and the type I error rates platforms. However, it is still clear that the real test size (0.064) was larger than the nominal 0.05. *Right*: The test size is well maintained by the block t-test.

We notice that the p-value provided by `lme` is biased and tends to be too small under the null hypothesis. The statistical inference for fixed effect relies on the asymptotic normal theory, which is known to be inadequate for small sample data (Kenward and Roger, 1997). More importantly, `lme` uses $n - K$ as the degrees of freedom of μ in its statistical inference, which is correct assymptotically, but ignores the contribution of the random effects themselves. It has been controversial in deciding the degrees of freedom in the statistical inference for fixed effect in linear mixed effect models (See e.g., Pinheiro and Bates, 2002). It is suggested (Feng *et al.*, 2001; Small *et al.*, 2008) that the degree of freedom for the fixed effect should be the number of random groups minor one. We thus corrected the p-values using $K-1$ as the degree of freedoms. Shown in Figure 6 are the Q-Q plot of the uncorrected/corrected `lme` p-values and the proposed simple t-test. Unfortunately, this $K - 1$ rule is now too conservative when K is small.

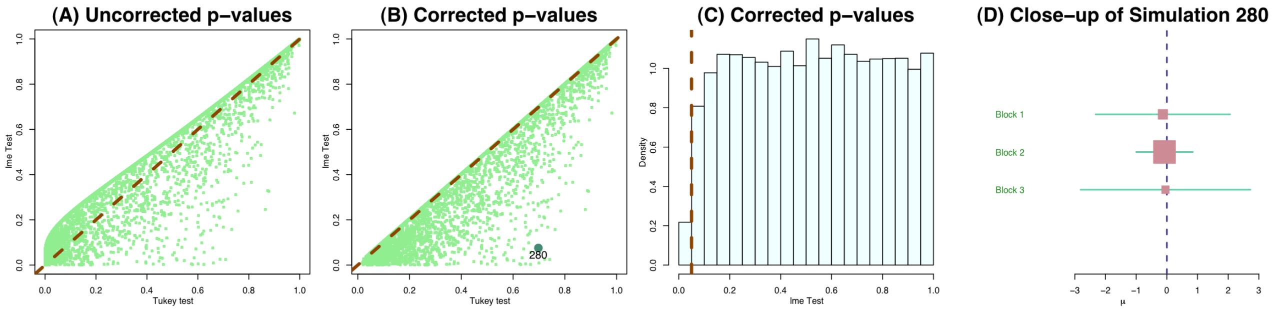

Figure 6: Q-Q plots of the lme p-values and the proposed t-test p-values. Data was simulated in a similar scheme as in Figure 5 with $K = 3$ and $\rho = 0.2$. (A). The original p-value output of lme. (B). p-values corrected by using degree of freedoms $K - 1$. (C). Histogram of the corrected p-values. The distribution does not look normal, and relatively fewer points falls close to the lower end. (D). The three blocks in sample 280. The between group variance is small, and $MSA = SSA/(K - 1) = 0.02$, while the within group variance is relative large with $MSE = SSE/K(n_0 - 1) = 1.16$. In this case, the standard error of $\overline{Y}$ is estimated by $\sqrt{MSE/n} = 0.196$, much smaller than the standard error used in the t-test $\sqrt{MSA/n_0} = 0.04$.

What goes wrong? The standard errors provided in `lme` are estimated using REML. In the balance design case, REML estimates can be illustrated by ANOVA (Faraway, 2005, Section

8.1). Simple algebra reveals that fact that REML tends to over-estimate the standard errors. REML estimates τ^2 and σ^2 separately. In order to estimate τ^2 and σ^2, note that the within group sum of squares $SSE = \sum_i^K \sum_j^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$ has expectation $K(n_0 - 1)\sigma^2$, and the between group sum of squares $SSA = n_0 \sum_i^K (\overline{Y}_{i\cdot} - \overline{Y})^2$ has expectation $(K-1)(\sigma^2 + n_0\tau^2)$. Hence, we may estimate σ^2 as $\hat{\sigma}^2 = SSE/K(n_0 - 1)$ and τ^2 as $\hat{\tau}^2 = \{SSA/(K-1) - \hat{\sigma}^2\}/n_0$. Unfortunately, the resulting $\hat{\tau}^2$ may have negative values. The maximum likelihood procedure indeed has the estimate $\hat{\tau}_+^2 = \hat{\tau}^2 \cdot I_{\{\hat{\tau^2} > 0\}}$. It is straightforward to see that the standard error for $\overline{Y}$ is thereby inflated. By contrast, the simple t-test does not separately estimate σ^2 and τ^2, has an honestly unbiased estimate of the standard errors, and thus retains the test size.

Indeed, since the maximum likelihood estimate, namely, $\hat{\tau}^2$, may be on the boundary of the parameter space, special asymptotic properties should be taken into account (Self and Liang, 1987). In some cases, an approximate scaled chi-square distribution might be derived for the standard error of $\overline{Y}$ (See e.g., Lin, 1997).

Fortunately, when the number of independent blocks K is large, intra-cluster correlation ρ is relatively large, *or* the effect size μ is large, the over-coservativeness of the $K-1$ correction rule does not affect the performance of `lme` test. In Figure 7, we illustrate the power difference between the blockwise t-test and `lme` test, $\beta_{\text{lme}}(\mu) - \beta_{\text{block}}(\mu)$, as functions in ρ and $\mu = 0.3, 0.5, 1,$ and 2. When $K \geq 10$, the two tests essentially have the same power; when $K \leq 4$ and the true μ is large enough, the `lme` test is more powerful than the blockwise t-test; the closer the ρ is to 1, the less conservative the `lme` test is; that the blockwise t-test is slightly less powerful close to the end of $\rho \uparrow 1$ is probably due to the fact that even the $K-1$ correction rule cannot completely control the inflated type I error rates of `lme` test under the null (data not shown).

Another popular R function that implements linear mixed effect model is `lmer` in package `lme4`. The author is aware of the problem of degrees of freedom and does not provide a p-value in his function. We did not find significant difference between the t-statistics for μ provided by these two functions (results not shown), hence we do not discuss `lmer` separately.

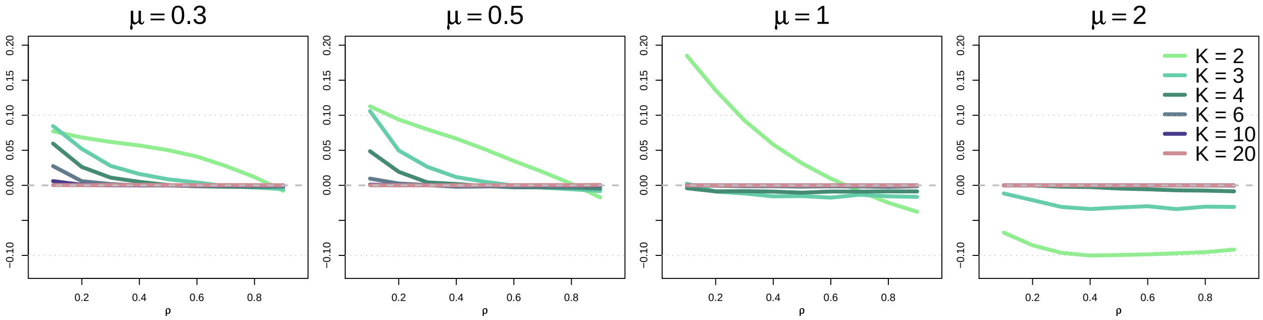

Figure 7: Estimated power difference between the blockwise t-test and lme test based on 10,000 simulation replications. Simulation details are the same as in Figure 5. The nominal test size is 0.05.

,t

5 Discussion

Powerful statistical inference highly relies on the validity of its assumptions. Violation of these assumptions may lead to unreliable or even false analysis. Robustness is thus of crucial importance in statistical analysis. Statisticians have studied the robustness of error distribution (Huber, 1981), heteroskedasticity (MacKinnon and White, 1985), model misspecification (White, 1982), data dependence (Liang and Zeger, 1986), and other aspects.

In this paper, we studied the robustness against the independence assumption, which is often the one that leads to the power of most statistical methodologies. We relaxed the *i.i.d.* assumption and took a more credible block independence assumption. We have showed that only the first few degrees of freedom are necessary–with a handful of independent blocks, the blocks model can pertain most power.

On the other hand, if correlated date are treated independently, the risk of making type I errors could be high even with very mild dependence. Since avoiding type I error is a more important concern in many studies, we suggest that it would not hurt to assume some kind of dependence, without substantial loss of test power.

Group structured data is very common. The structure may come with the data by design as in group randomized trials, or by the sampling scheme as in family-based studies, where relatives in a pedigree will be recruited in the study. Sometimes, group-wise in/dependence might also be implicit, for instance, due to the population or spatial mixture nature of the data. Another kind of implicit group structure may also be caused by the different sources of the data. For example, It is very common in education research and medical research to combine data from difference resources and then perform meta-analysis (Egger and Smith, 1997; Pena, 1997). In genetic and genomic research, pooling sequencing data together can not only help strengthen the signal (for example, van Vliet *et al.*, 2008) and sometimes achieve a genome-wide significance under the Bonferroni correction, but also help impute untyped SNPs that will lead to exciting findings (Willer *et al.*, 2008; Li *et al.*, 2009). However, because of the difference locations of studies or different genotyping platforms used in each study, simply pooling the data together might cause dependence. Unaware of these group structures might be dangerous and be in high risk of overstating significance.

In this paper, we advocated a simple block-wise t-test, which uses the sample standard deviation of the group means to estimate the standard error of the estimate. A more careful choice of the standard error is to use the sandwich estimator (Liang and Zeger, 1986)

$$s = \frac{1}{n}\sqrt{\sum_{i=1}^{K}\sum_{j,k=1}^{n_i}(Y_{i,j} - \overline{Y})(Y_{i,k} - \overline{Y})}$$

as an estimate of the standard deviation of group means. This approach has been widely applied in analyzing correlated data. In some cases knowing the covariance structures can lead to more powerful tests (for example, Thornton and McPeek, 2007). However, with small number of independent groups, the sandwich estimator is biased and underestimates the standard deviation. Hence, the sandwich estimator based statistical inference should be careful with small sample (Fay and Graubard, 2001; Mancl and DeRouen, 2001).

A Appendix

A.1 The power functions in Section 2

Without loss of generation, presume that the hypotheses are $H_0 : \mu_0 = 0$ versus $H_a : \mu_0 > 0$; $n_i \equiv n_0$ and $\sigma_i^2 \equiv \sigma^2$, $i = 1, \ldots, K$.

Notice that the rejection region for a level α test is $\{T \geq t_{n-1,\alpha}\}$. The power of this test with $s = s_1/\sqrt{n}$ at $\mu_0 = \mu$ is

$$1 - \beta_{\text{iid}}(\mu) = 1 - F_{n-1}\left(t_{n-1,\alpha} - \frac{\mu}{s_1/\sqrt{n}}\right), \tag{7}$$

where $F_{n-1}(\cdot)$ and $t_{n-1,\alpha}$ are respectively the cumulative distribution function, and the $1-\alpha$ quantile of a student's t distribution with degrees of freedom $n-1$.

Taking $s = s_2/\sqrt{K}$, the power of a level α test at $\mu_0 = \mu$ is

$$1 - \beta_{\text{block}}(\mu) = 1 - F_{K-1}\left(t_{K-1,\alpha} - \frac{\mu}{s_2/\sqrt{K}}\right), \tag{8}$$

where $F_{K-1}(\cdot)$ and $t_{K-1,\alpha}$ are respectively the cumulative distribution function, and the $1-\alpha$ quantile of a student's t distribution with degrees of freedom $K-1$.

A.2 The type I error probabilities in Section 3

In general the sample distribution of $(n-1)s_1^2 = \sum_{i,j}(Y_{i,j} - \overline{Y})^2$ does not have a χ^2 distribution unless $\rho = 0$. Hence, the distribution of the T_{iid} is not easy to formulate. To simplify this calculation, we suppose σ is known.

Suppose some careless statistician performs an α level test on $H_0 : \mu = 0$ versus $H_a : \mu > 0$, mistakenly using statistic

$$Z_w = \frac{\overline{Y}}{\sigma/\sqrt{n}} \tag{9}$$

with rejection region $\{Z_w \geq z_\alpha\}$.

Since under $H_0 : \mu = 0$,

$$\overline{Y}/\sigma \sim N\left(0, \frac{1}{n} + \frac{(n_0 - 1)\rho}{n}\right),$$

the true type I error probability is

$$\alpha(\rho) = 1 - \Phi\left(\frac{z_\alpha}{\sqrt{1 + (n_0 - 1)\rho}}\right). \tag{10}$$

References

Bertail, P., Gautherat, E., and Harari-Kermadec, H. (2008). Exponential Bounds for Multivariate Self-normalized Sums. *Electronic Communications in Probability*, **13**, 628–640.

Bisgaard, H. (2004). The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): Design, Rationale, and Baseline Data from a Longitudinal Birth Cohort Study. *Annals of Allergy Asthma and Immunology*, **93**, 381–389.

Cornfield, J. (1978). Randomization by Group. *American Journal of Epidemiology*, **108**, 100–102.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford, second edition.

Egger, M. and Smith, G. D. (1997). Meta-analysis: Potentials and Promise. *BMJ*, **315**, 1371–1374.

Faraway, J. J. (2005). *Extending the Linear Model with R (Texts in Statistical Science)*. Chapman & Hall/CRC.

Fay, M. P. and Graubard, B. I. (2001). Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. *Biometrics*, **57**, 1198–1206.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. (2001). Selected Statistical Issues in Group Randomized Trials. *Annu. Rev. Public Health*, **22**, 167–187.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Kenward, M. G. and Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, **53**, 983–997.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963–974.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annu. Rev. Genom. Human Genet.*, **10**, 387–406.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lin, X. (1997). Variance Component Testing in Generalized Linear Models with Random Effects. *Biometrika*, **84**, 309–326.

Lockwood, J. R., Doran, H., and McCaffrey, D. F. (2003). Using R for Estimating Longitudinal Student Achievement Models. *R News*, **3**, 17–23.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**, 305–325.

Mancl, L. A. and DeRouen, T. A. (2001). A Covariance Estimator for GEE with Improved Small-Sample Properties. *Biometrics*, **57**, 126–134.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004). Design and Analysis of Group-Randomized Trials: A Review of Recent Methodologies Developments. *American Journal of Public Health*, **94**(3), 423–432.

Pena, D. (1997). Combining Information in Statistical Modeling. *The American Statistician*, **51**(4), 326–332.

Pinheiro, J. C. and Bates, D. M. (2002). *Mixed Effects Models in S and S-Plus*. Springer.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *J. Amer. Statist. Assoc.*, **82**(398), 605–610.

Small, D., Ten Have, T., and Rosenbaum, P. (2008). Randomization Inference in a Group Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects. *J. Amer. Statist. Assoc.*, **103**(481), 271–279.

Thornton, T. and McPeek, M. S. (2007). Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *Am. J. Hum. Genet.*, **81**, 321–337.

van Vliet, M. H., Reyal, F., Horlings, H. M., van de Vijver, M. J., Reinders, M. J. T., and Wessels, L. F. A. (2008). Pooling Breast Cancer Datasets Has A Synergetic Effect on Classification Performance and Improves Signature Stability. *BMC Genomics*, **9**, 375.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50**(1), 1–25.

Willer, C. J., Shanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., and et al. (2008). Genome-wide Association Scans Identify Novel Loci That Influence Lipid Levels and Risks of Coronary Artery Disease. *Nat. Genet.*, **40**, 161–169.