Biometrics 73, 960–971
September 2017

DOI: 10.1111/biom.12651

A Generalized Levene's Scale Test for Variance Heterogeneity in the Presence of Sample Correlation and Group Uncertainty

David Soave[1,2,*] and Lei Sun[3,1,]**

[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario M5T 3M7, Canada

[2]Program in Genetics and Genome Biology, Research Institute, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada

[3]Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada

email: david.soave@mail.utoronto.ca

**email:* sun@utstat.toronto.edu

Summary. We generalize Levene's test for variance (scale) heterogeneity between k groups for more complex data, when there are sample correlation and group membership uncertainty. Following a two-stage regression framework, we show that least absolute deviation regression must be used in the stage 1 analysis to ensure a correct asymptotic $\chi^2_{k-1}/(k-1)$ distribution of the generalized scale (gS) test statistic. We then show that the proposed gS test is independent of the generalized location test, under the joint null hypothesis of no mean and no variance heterogeneity. Consequently, we generalize the recently proposed joint location-scale ($gJLS$) test, valuable in settings where there is an interaction effect but one interacting variable is not available. We evaluate the proposed method via an extensive simulation study and two genetic association application studies.

Key words: Association studies; Heteroscedasticity; Joint location-scale test; Scale test.

1. Introduction

Testing for scale (variance) heterogeneity, prior to the main inference of location (mean) parameters, is a common diagnostic method in linear regression to evaluate the assumption of homoscedasticity. In some research areas, such as statistical genetics, testing for heteroscedasticity itself can be of primary interest.

With the goal of detecting a genetic association between a single-nucleotide polymorphism (SNP, G) and a quantitative outcome (phenotype, Y), the traditional approach is to conduct a location test, testing mean differences in Y across the three genotype groups of the SNP ($G = 0$, 1, or 2 copies of the minor allele, the variant with population frequency <0.5). However, it has been noted that a number of biologically meaningful scenarios can lead to variance differences in Y across the genotype groups of a SNP of interest (say G_1). For example, an underlying interaction effect, between G_1 and another SNP G_2 (G_1xG_2) or an environmental factor E (G_1xE), on Y can lead to heteroscedasticity across G_1, if the interacting G_2 or E variable was not collected and the interaction term may not be directly modeled (Pare et al., 2010). Transformations on a phenotype can also result in variance heterogeneity (Sun et al., 2013). This transformation can occur knowingly for statistical purposes, for example, $\log(Y)$, or unknowingly, for example, choosing a phenotype measurement that does not directly represent the true underlying biological outcome of a gene. In each of these scenarios, a scale test can be used either alone to indirectly detect associated SNPs (Pare et al., 2010), or combined with a location test to increase testing power (Cao et al., 2014; Soave et al., 2015). Heteroscedasticity due to interaction effects has also been investigated for variable selection via sliced inverse regression methods (Jiang and Liu, 2014).

Genotype uncertainty is inherent in sequenced and imputed SNP data. For such data, the genotype of a SNP for an individual ($G = 0$, 1, or 2) is represented by three genotype probabilities (p_0, p_1, p_2, and $p_0 + p_1 + p_2 = 1$). For testing methods that require genotype to be known unambiguously, the probabilistic data are typically transformed into the so-called "best-guess" (most likely or hard-call) genotype, selected as the one with the largest probability. In the context of location-testing, several groups have proposed methods that incorporate the probabilistic data and showed that this improves power (Acar and Sun, 2013; Kutalik et al., 2011). The corresponding development for scale-testing, however, is lacking.

Genetic association studies often involve related individuals, where individuals in a sample are correlated or clustered. The correlation structure may be specified based on the known genealogy information or accurately estimated using the genome-wide genetic SNP data collected (Sun and Dimitromanolakis, 2012). A number of generalized location tests allowing for family data have been proposed (Horvath et al., 2001; Jakobsdottir and McPeek, 2013), and their power gain over analyzing only a subset of independent individuals is a direct consequence of the increase in sample size. However,

few scale tests deal with correlated data, with the exception of methods proposed specifically for clustered data present in twin studies (Haseman and Elston, 1970; Iachine et al., 2010). Further, these methods have been reported to have type 1 error issues in the presence of non-normal data or small, unequal group sizes (Iachine et al., 2010), and they have not been extended to incorporate group membership uncertainty.

Both classical statistical tests and graphical procedures have been proposed to investigate heteroscedasticity (Bartlett, 1937; Breusch and Pagan, 1979; Cook and Weisberg, 1983; Levene, 1960; White, 1980). In big data settings, such as genome-wide association studies where possibly millions of SNPs are scanned for association with an outcome, graphical and other computationally burdensome approaches are not ideal. Levene's test (Levene, 1960) is known for its simplicity and robustness to modeling assumptions, and it is perhaps the most popular method for evaluating variance heterogeneity between k groups. Therefore, our development here focuses on Levene's method.

In this article, we extend Levene's test for equality of variances across k groups to allow for both group membership uncertainty and sample correlation. When groups are known, we show that the proposed method outperforms existing methods for clustered twin data. In the presence of group uncertainty, we demonstrate that our test continues to be accurate and has improved power over the "best-guess" approach. This generalized scale test can be used alone for heteroscedasticity diagnostic purposes but with wider applicability. Motivated by the complex genetic association studies described above, we also show that the proposed generalized scale test can be combined with existing generalized location tests using the joint location-scale framework, previously developed for population samples without group uncertainty (Soave et al., 2015), to further improve power. We apply our methods to two genetic association studies, one of HbA1c levels in individuals with type 1 diabetes, and the other of lung disease in individuals with cystic fibrosis (CF).

2. Methodology

We first consider a sample of independent observations with no group uncertainty, and we formulate Levene's test as a regression problem. Using this regression framework, we then extend Levene's test as the generalized scale (gS hereinafter) test to allow for sample dependency and group uncertainty. For clarity in our method comparisons, we also briefly discuss the Iachine et al. (2010) extension of Levene's test, specifically designed for twin pairs without group uncertainty. Finally, we generalize the joint location-scale test of Soave et al. (2015) for the complex data structure considered here ($gJLS$).

2.1. *Notation and Statistical Model*

Let $y_i, i = 1, \ldots, n$, be a sample of independent observations, where each $y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Suppose the y_i's fall into k distinct treatment groups with group-specific variance σ_j^2, $j = 1, \ldots, k$, and let n_j be the sample size for group j, $n = \sum n_j$. Our motivation concerns testing the null hypothesis of equal variance across the k groups:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2. \tag{1}$$

Here, we use σ_j^2 for group-specific variance, $j = 1, \ldots, k$, and σ_i^2 for observation-specific variance, $i = 1, \ldots, n$.

Let x_{ij}, $j = 1, \ldots, k-1$, be the standard dummy variables that define group membership for observation i; group 1 is the reference group. Consider the normal linear model of interest here,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i(k-1)} + \varepsilon_i, \tag{2}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, σ_i^2 corresponds to the variance associated with the group that y_i belongs to, $i = 1, \ldots, n$. In other words, $\sigma_i^2 = \sigma_{j^*}^2$ if $x_{i(j^*-1)} = 1$. In matrix notation,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where X is the design matrix obtained by stacking the $\boldsymbol{x}_i^T$'s $= (1, x_{i1}, x_{i2}, \ldots, x_{i(k-1)})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \Sigma)$, and Σ is the covariance matrix with diagonal elements σ_i^2's.

2.2. *Formulating Levene's Test as a Regression F-Test and Modifications*

The classical formulation of Levene's test first centers the y_i's by their estimated group means and obtains the corresponding absolute deviation d_i's. It then tests for mean differences in the d_i's across the k groups using ANOVA. Let I_{ij}, $j = 1, \ldots, k$ be the group indicator variables, where $I_{ij} = 1$ if individual i belongs to group j. Now, let $\overline{\mu}_{(j)} = \sum_{i=1}^n y_i I_{ij}/n_j$ be the estimated group means of the y_i's, such that an estimate of $E(y_i)$ is $\overline{\mu}_i = \sum_{j=1}^k I_{ij}\overline{\mu}_{(j)}$. The corresponding absolute deviation is

$$d_i = |y_i - \overline{\mu}_i|.$$

Let $\overline{d}_{(j)} = \sum_{i=1}^n d_i I_{ij}/n_j$ be the estimated group means of the d_i's, such that an estimate of $E(d_i)$ is $\overline{d}_i = \sum_{j=1}^k I_{ij}\overline{d}_{(j)}$, and let $\overline{\overline{d}} = \sum_{i=1}^n d_i/n$ be the grand mean. Finally, Levene's test statistic has the following form

$$F(\boldsymbol{d}) = \frac{\sum_{i=1}^n (\overline{d}_i - \overline{\overline{d}})^2/(k-1)}{\sum_{i=1}^n (d_i - \overline{d}_i)^2/(n-k)},$$

where $F(\boldsymbol{d})$ follows approximately an $F(k-1, n-k)$ distribution under the null hypothesis of (1), and a $\chi_{k-1}^2/(k-1)$ distribution asymptotically as $n \to \infty$.

For the purpose of developing a unified approach, we re-formulate Levene's test using the following two-stage regression framework:

- Stage 1.1. Obtain the residuals, $\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$, from the ordinary least squares (OLS) regression of y_i on $\boldsymbol{x}_i$; we refer to this as the *stage 1* regression.
- Stage 1.2. Take the absolute values of these residuals, $d_i = |\widehat{\varepsilon}_i|$.
- Stage 2. Test for an association between the d_i's and $\boldsymbol{x}_i$'s using a regression F-test; we refer to this as the *stage 2* regression *and* test.

The justification for this two-stage regression procedure (Levene's test) being a test of homoscedasticity (1) is as follows. Stage 1 performs OLS regression using a working covariance matrix $\Sigma_{stage\,1} = \sigma_y^2 I$, where I is the identity matrix. Therefore $\widehat{\boldsymbol{y}} = X(X^TX)^{-1}X^Ty = Hy$, $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} \sim \mathcal{N}(\mathbf{0}, \Sigma(I-H))$, and $\widehat{\varepsilon}_i \sim \mathcal{N}(0, \sigma_i^2(1-h_{ii}))$, where h_{ii} is the ith diagonal element of the hat matrix H. Consequently $d_i = |\widehat{\varepsilon}_i|$ follows a folded-normal distribution and its mean is a linear function of σ_i,

$$E(d_i) = \sigma_i\sqrt{2(1-h_{ii})/\pi}.$$

This relationship between d_i and σ_i is approximated by the following stage 2 working model,

$$d_i = \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_{k-1}x_{i(k-1)} + e_i, \tag{4}$$

where $e_i \sim \mathcal{N}(\mathbf{0}, \sigma_d^2)$. In matrix form,

$$\boldsymbol{d} = X\boldsymbol{\theta} + \boldsymbol{e}, \tag{5}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\gamma}^T)^T = (\alpha, \gamma_1, \ldots, \gamma_{k-1})^T$, and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \Sigma_{stage\,2})$, $\Sigma_{stage\,2} = \sigma_d^2 I$. Testing the null hypothesis (1) is now reformulated as testing

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_{k-1} = 0, \tag{6}$$

using the classical OLS regression F-test. Note that although the d_i's are folded normal variables, Levene's variance test takes advantage of the fact that inference from OLS regression is robust to violations of the normality assumption.

This formulation of Levene's test has a similar structure to the score test of Glejser (1969) proposed for testing heterscedasticity associated with continuous covariates. Godfrey (1996) showed that when estimating $\boldsymbol{\beta}$ by OLS in stage 1, the Glejser score statistic derived in stage 2 is not asymptotically distributed as χ_1^2, unless the distribution of $\boldsymbol{\varepsilon}$ is symmetric. To achieve robustness, several modifications have been proposed (Brown and Forsythe, 1974; Im, 2000; Machado and Silva, 2000; Furno, 2005; Gastwirth et al., 2009), among which replacing sample group means with medians in constructing the d_i's is consistently recommended (Conover et al., 1981; Lim and Loh, 1996). It has also been shown analytically that, for non-symmetric $\boldsymbol{\varepsilon}$, centering on the sample group medians will lead to an asymptotically correct Levene's test (Carroll and Schneider, 1985) and correct Glejser's score test (Furno, 2005). In the regression framework, this modification corresponds to estimating $\boldsymbol{\beta}$ by least absolute deviation (LAD) regression instead of OLS regression in stage 1.

2.3. *The Generalized Levene's Scale (gS) Test*

The above regression framework for Levene's test allows us to incorporate group uncertainty by simply replacing the group indicators or dummy variables for each observation, $\boldsymbol{x}_i$, with the corresponding group probabilities. Analogous to dummy variables, the group probabilities for each individual sum to 1, so we omit one of the covariates to ensure model identifiability. Using genetic association as an example again, let $(p_0 = 0.25, p_1 = 0.42, p_2 = 0.33)$ be the genotype probabilities for an individual i at a SNP of interest, then, without loss of generality, we can define $\boldsymbol{x}_i^T = (1, x_{i1}, x_{i2}) = (1, 0.42, 0.33)$. Note that the "best-guess" approach would have the corresponding covariates or dummy variables as $\boldsymbol{x}_i^T = (1, 1, 0)$.

Now, consider correlated data where ε_i and ε_j are no longer independent of each other and the covariance matrix Σ is no longer diagonal. In the stage 1 regression, because we are only interested in obtaining $\widehat{\boldsymbol{\beta}}$ to construct $d_i = |y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}|$, we can continue to use OLS or LAD regression with the misspecified working covariance matrix, $\Sigma_{stage\,1} = \sigma_y^2 I$, to obtain consistent and unbiased $\boldsymbol{\beta}$ estimates.

Stage 2 involves estimating the variance of $\widehat{\boldsymbol{\gamma}}$ to test the null hypothesis of (6), and not accounting for sample dependency can lead to invalid inference. Let $\Sigma_{stage\,2} = \sigma_d^2\Sigma_d$ be the working covariance matrix for $\boldsymbol{d}$, a valid inference can be achieved by using a generalized least squares (GLS) approach when Σ_d is known (Aitken, 1936). When Σ_d is unknown, feasible GLS (FGLS) can be used, with or without iteration, where an estimate of Σ_d is obtained, subject to constraints, and then used in GLS. Alternatively, orthogonal-triangular decomposition methods can be used to obtain a compact representation of the profiled log-likelihood, such that maximum likelihood estimates (MLE's) of all parameters can be obtained jointly through nonlinear optimization (Pinheiro and Bates, 2000).

In many scientific settings, including genetic association studies, the sample correlation structure is often prespecified with constraints on the $n(n-1)/2$ correlations, for example, a single serial correlation ρ for time series or family data with a single relationship type (e.g. sibling data), or cluster-specific correlations ρ's for different clusters. In this case, let $\Sigma_{stage\,2} = \sigma_d^2\Sigma_d(\rho) = \sigma_d^2 C(\rho)C(\rho)^T$ be the Cholesky decomposition, and define

$$\boldsymbol{d}^* = C(\rho)^{-1}\boldsymbol{d},\ \ X^* = C(\rho)^{-1}X,\ \ \boldsymbol{e}^* = C(\rho)^{-1}\boldsymbol{e}.$$

The GLS or FGLS regression, in essence, deals with the transformed model in stage 2

$$\boldsymbol{d^*} = X^*\boldsymbol{\theta} + \boldsymbol{e^*}, \tag{7}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\gamma}^T)^T$. For a fixed ρ, the conditional MLEs for $\boldsymbol{\theta}$ and σ_d^2 are

$$\widehat{\boldsymbol{\theta}} = [X^{*T}X^*]^{-1}X^{*T}\boldsymbol{d}^*,\ \ \widehat{\sigma_d^2} = \frac{1}{n}\left\|\boldsymbol{d}^* - X^*\widehat{\boldsymbol{\theta}}\right\|^2.$$

The MLE of ρ can be obtained by optimizing the profiled log-likelihood,

$$l(\rho) = \text{constant} - n\log\left\|\boldsymbol{d}^*(\rho) - X^*(\rho)\widehat{\boldsymbol{\theta}}(\rho)\right\| - \frac{1}{2}\log|C(\rho)|.$$

Thus, the generalized Levene's scale (gS) test of the null hypothesis of (6), $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, using the regression F-test in stage 2, has the following test statistic:

$$F(\boldsymbol{d}^*) = \frac{\sum_{i=1}^n(\widehat{d_i^*} - \widetilde{d_i^*})^2/(k-1)}{\sum_{i=1}^n(d_i^* - \widehat{d_i^*})^2/(n-k)}, \tag{8}$$

where $\widehat{d}_i^* = (\boldsymbol{x}_i^*)^T\widehat{\boldsymbol{\theta}}$, the predicted values from regression model (7), and $\widetilde{d}_i^* = 1_i^*\widetilde{\alpha}$, the predicted values from the regression of $\boldsymbol{d}^*$ on $\mathbf{1}^*$. Note that $\mathbf{1}^*$ is the first column of the transformed design matrix X^*, and may not be a vector of 1's. When the observations are independent of each other and group membership is known unambiguously, it is easy to verify that $\widehat{d}_i^* = \overline{d}_i$ and $\widetilde{d}_i^* = \overline{\overline{d}}$, and $F(\boldsymbol{d^*})$ reduces to the original form of $F(\boldsymbol{d})$.

Under the normal linear model of (4), the F-statistic (8) is asymptotically $\chi^2_{k-1}/(k-1)$ distributed (Arnold, 1980). However, for non-symmetric $\boldsymbol{\varepsilon}$, we show that this is true only when $\boldsymbol{d}$ is estimated using LAD in the stage 1 regression (Web Appendix A, Theorem 1).

In our simulation and application studies that follow, the examples considered involve known sibling pairs. In that case, we specify a compound symmetric correlation structure with one correlation parameter, ρ, that assumes equal correlation among all within-group errors pertaining to the same pair/cluster. We use GLS regression in R with the *gls*() function in the "nlme" package. The *gls*() function obtains maximum likelihood estimates of the model parameters by using a hybrid expectation maximization and Newton–Raphson algorithm to optimize the profiled likelihood; see Pinheiro and Bates (2000) for details.

2.4. *The Iachine et al. (2010) Scale Test for Twin Pairs and Modifications*

Focusing on paired-observations, Iachine et al. (2010) extended Levene's test to determine if the variance of an outcome differs between monozygotic (MZ) and dizygotic (DZ) twin pairs. The proposed twin (*TW*) test follows Levene's two-stage regression procedure but it makes use of the Huber–White sandwich estimate (White, 1980) of $\text{Var}(\widehat{\gamma}_1)$ in the stage 2 analysis (here $k = 2$ groups, requiring only one dummy variable) to construct an asymptotically χ^2_1 distributed Wald statistic, operationally an F-statistic in finite samples.

Complications with the *TW* test may arise if the number of clusters is small in either group (MZ or DZ) and can be compounded with imbalance between the groups (Iachine et al., 2010). Unfortunately, there is no clear definition of too few clusters (Cameron and Miller, 2015), and empirical type 1 error rates can be inflated for study designs with less than 20 clusters per group, particularly combined with non-symmetric data (see Iachine et al. (2010) and simulation results in Section 3 below).

The original *TW* method assumes that if two observations are from the same pair/cluster they also belong to the same group k. This may not be satisfied in a more general setting such as the genetic association studies discussed above. For example, two individuals from the same DZ pair or familial cluster often have different genotypes at a SNP of interest, so individuals from the same cluster may not share a common σ^2_k. However, the sandwich variance estimator can still be used in this setting. In the presence of group uncertainty, the *TW* method can be modified by replacing the group indicator covariate with corresponding group probabilities.

2.5. *Generalized Joint Location-Scale (gJLS) Testing*

The standard location test of mean differences in an (approximately) normally distributed outcome across covariate values (e.g., the three genotype groups of a SNP in a genetic association study) is testing

$$H_0^{location} : \beta_1 = \cdots = \beta_{k-1} = 0,$$

based on regression model (2). While the location inference tests the β_j's, the scale test discussed here uses only the β estimates from the stage 1 regression of model (2) to obtain $d_i = |y_i - \widehat{y}_i|$ for the stage 2 regression of model (4), and it performs a hypothesis test on the γ_j's, testing

$$H_0^{scale} : \gamma_1 = \cdots = \gamma_{k-1} = 0.$$

A joint location-scale (*JLS*) test is interested in the following global null hypothesis,

$$H_0^{joint} : \beta_j = 0, \text{ and } \gamma_j = 0, \forall \;\; j = 1, \ldots, k-1. \tag{9}$$

One simple yet powerful *JLS* method proposed in Soave et al. (2015) uses Fisher's method to combine p_L and p_S, the p-values of the individual location and scale tests. One can consider other aggregation statistics, for example, the minimum p-value (Derkach et al., 2014); for a review of this topic see Owen (2009). Focusing on Fisher's method, the corresponding test statistic is

$$W_F = -2(\log(p_L) + \log(p_S)).$$

For independent observations with no group uncertainty, Soave et al. (2015) showed that, under H_0^{joint} of (9) and a Gaussian model, p_L and p_S are independent. Thus, W_F is distributed as a χ^2_4 random variable.

In the presence of sample correlation with group uncertainty, we propose to use the same framework but obtain p_L from a generalized location test (e.g., a GLS approach to model (2), where the design matrix X includes the group probabilities, and the covariance matrix, $\Sigma_{stage\,1} = \sigma^2_y\Sigma_y$, incorporates the sample correlation), and p_S from the *gS* test proposed here. We show that the assumption of independence between p_L and p_S continues to hold theoretically under H_0^{joint} of (9) for normally distributed outcomes (Web Appendix B), as well as empirically for approximately normally distributed outcomes in finite samples (Web Figure 1).

3. Simulations

The validity of the generalized joint location-scale (*gJLS*) testing procedure relies on the accuracy of the individual generalized location (*gL*) test and generalized scale (*gS*) test components. The performance of the *gL* test has been established in the literature, therefore, our simulation studies here focused on evaluation of the proposed *gS* test, and when appropriate compared it with Levene's original test (*Lev*) and the *TW* test of Iachine et al. (2010). For completeness, we also numerically demonstrate the power improvement of the proposed *gJLS* test. We use subscripts *OLS* and *LAD* to denote if the stage 1 regression was performed using OLS to obtain group-*mean*-adjusted residuals or LAD for group-*median*-adjusted residuals. Implementation details of each of

the six tests (Lev_{OLS}, Lev_{LAD}, TW_{OLS}, TW_{LAD}, gS_{OLS},gS_{LAD}) is outlined in Web Appendix C.

We considered two main simulation models. Simulation model 1 followed the exact simulation setup of Iachine et al. (2010) to ensure fair comparison. Simulation model 2 extended model 1 by introducing genotype groups for each individual as well as group membership uncertainty. To apply the original *Lev* test for comparison, we ignored the inherent sample correlation in the presence of correlated data. In all simulations, empirical type 1 error and power were evaluated at the 5% significance level using 10,000 replicates, unless otherwise stated.

3.1. *Simulation Model 1*

3.1.1. Model setup. Following the exact simulation study design of Iachine et al. (2010), we simulated correlated outcome values for n_1 MZ and n_2 DZ twin pairs, $n = 2n_1 + 2n_2$, and we tested if the variance of the outcome differed between the two groups of pairs, that is, $\sigma_1^2 = \sigma_2^2$. To study robustness, we simulated outcomes using Gaussian, Students t_4 (heavier tailed), and χ_4^2 (non-symmetric) distributions.

We first generated pairs of observations from independent bivariate normal distributions $BV\mathcal{N}(0, 1, \rho_k), k = 1, 2$, with ρ_1 and ρ_2 corresponding to the correlation within the MZ and DZ twin pairs, respectively. Let w be the variable for an observation, we then applied a transformation $g(\cdot)$ to w to obtain the desired marginal distribution, $y = \sigma_k g(w)$, where the σ_k's induced different variances between the two groups. The choice of $g(\cdot)$ depended on the desired distribution for y:

$$g(w) = \begin{cases} w, & \text{if y} \sim \mathcal{N}(0,1) \\ F_{t_4}^{-1}(\Phi(w)), & \text{if y} \sim \text{t}_4 \\ F_{\chi_4^2}^{-1}(\Phi(w)), & \text{if y} \sim \chi_4^2 \end{cases},$$

where Φ, F_{t_4} and $F_{\chi_4^2}$ are the cumulative distribution functions for the standard normal, Students t_4 and χ_4^2 distributions, respectively.

We varied the sample size ($n_1, n_2 = 5$, 10, or 20 for small samples, and $= 500$, 1000, or 2000 for large samples, and n_1 may or may not equal n_2), and group variances ($\sigma_1^2, \sigma_2^2 = 1$, 2, or 4). The level of correlation within the MZ and DZ twin pairs was $\rho_1 = 0.75$ and $\rho_2 = 0.5$, respectively.

3.1.2. Results. We were able to replicate the simulation results of Iachine et al. (2010) that studied Lev_{OLS}, Lev_{LAD}, TW_{OLS}, and TW_{LAD} (Table 1 and Web Table 1). However, we noticed that results reported in their article for Lev_{LAD} and TW_{LAD} presumably using median-adjusted residuals (labeled as W_{50} and TW_{50}, columns 9 and 12 of Tables 1–4 in Iachine et al. (2010)) were in fact switched with the *Lev* and *TW* results obtained using 10% trimmed mean-adjusted residuals (labeled as W_{10} and TW_{10} in Iachine et al. (2010)). Subsequent conclusions in Iachine et al. (2010) that the *TW* method using the 10% trimmed mean "performed best", therefore, are incorrect and should instead refer to TW_{LAD} using median-adjusted residuals from the stage 1 regression.

Table 1
Type 1 error evaluation under simulation model 1

n_1	n_2	Lev_{OLS}	Lev_{LAD}	TW_{OLS}	TW_{LAD}	gS_{OLS}	gS_{LAD}
				Gaussian			
20	20	0.102	0.087	0.055	0.044	0.058	0.046
5	5	0.115	0.071	0.085	0.041	0.099	0.049
10	20	0.112	0.091	0.085	0.064	0.075	0.054
5	10	0.114	0.079	0.118	0.079	0.092	0.054
				Student's t_4			
20	20	0.102	0.084	0.056	0.043	0.059	0.045
5	5	0.129	0.069	0.086	0.037	0.103	0.046
10	20	0.118	0.093	0.090	0.069	0.078	0.054
5	10	0.123	0.076	0.115	0.071	0.093	0.048
				χ_4^2			
20	20	0.175	0.098	0.112	0.052	0.117	0.054
5	5	0.180	0.083	0.133	0.053	0.153	0.061
10	20	0.181	0.102	0.146	0.079	0.137	0.062
5	10	0.187	0.094	0.178	0.085	0.149	0.064

Six different tests were evaluated, including the original Levene's test, *Lev*, the twin test of Iachine et al. (2010), *TW*, and the proposed generalized scale test, *gS*, with subscripts $_{OLS}$ and $_{LAD}$ denoting whether the stage 1 regression was performed using OLS or LAD. Parameter values included n_1 and n_2 for the number of MZ and DZ twin pairs, respectively, and $\rho_1 = 0.75$ and $\rho_2 = 0.5$ for the corresponding within-pair correlations. Without loss of generality, $\sigma_1^2 = \sigma_2^2 = 1$ for type 1 error rate evaluation. The empirical type 1 error was estimated from 10,000 simulated replicates at the nominal 5% level.

Our results in Table 1 clearly show that

- In the presence of sample correlation, Levene's original method *Lev* that ignores the correlation had severely increased type 1 error rate, even with Gaussian data. That is, *TW* and *gS* performed better than *Lev*.
- When the error structure was non-symmetric (χ_4^2) or the group sizes were small (e.g., n_1 or n_2 less than 20), using OLS in the stage 1 regression for either *TW* or *gS* led to increased type 1 error. That is, TW_{LAD} and gS_{LAD} performed better than TW_{OLS} and gS_{OLS}, respectively.
- When the group sizes were unbalanced and small (e.g., $n_1 = 10, n_2 = 20$), TW_{LAD} had increased type 1 error, even with Gaussian data. That is, gS_{LAD} performed better than TW_{LAD}.

In large samples, the original *Lev* test remained problematic with empirical $\alpha = 0.097$ when $n_1 = n_2 = 2000$ even for Gaussian data (Web Table 1). The accuracy of both TW_{LAD} and gS_{LAD} increased as n increased, with empirical $\alpha = 0.052$ when $n_1 = n_2 = 2000$ even for the non-symmetric χ_4^2 data. The accuracy of both TW_{OLS} and gS_{OLS} also improved as n increased, however, only for symmetric Gaussian or t_4 data. For χ_4^2 data, their empirical α level remained as high as 0.103 even for $n_1 = n_2 = 2000$; this empirical result is consistent with Theorem 1 (Web Appendix A).

Because most of the six tests did not have good type 1 error control in the presence of sample correlation, small samples, unbalanced group sizes, or non-symmetric data, we delay the discussion of power until simulation model 2 below where we focus on methods comparison between TW_{LAD} and gS_{LAD}, and in a more general simulation set-up.

3.2. *Simulation Model 2*

3.2.1. Model setup. The second simulation setup was motivated by genetic association studies as previously discussed. We again considered sibling pairs to introduce sample correlation. However, unlike simulation model 1, here we allowed individuals from the same pair to belong to different groups, where the groups were the three different genotypes of a SNP.

Consider a SNP of interest with minor allele frequency (MAF) of q (= 0.2 or 0.1), we first simulated genotypes for $n/2$ (= 20, 50, 100, 500, or 1000) pairs of siblings. To account for the inherent correlation of genotypes between a pair of siblings, we started with drawing the number of alleles shared identical by decent (IBD), $D = 0$, 1, or 2, from a multinomial distribution with parameters (0.25, 0.5, 0.25), independently for each sib-pair. Given the IBD status D, we then simulated paired genotypes $(G_1, G_2) = (i, j), i, j \in \{0, 1, 2\}$, following the known conditional distribution of $\{(G_1, G_2)|D\}$ (Thompson, 1975; Sun, 2012). The distribution depends on q in a way that smaller q leads to greater imbalance in the genotype group sizes. The probabilities of the numbers of individuals with genotype $G = 0$, 1, and 2 are $(1-q)^2$, $2q(1-q)$, and q^2, respectively.

To introduce group membership uncertainty, we converted the simulated true genotypes G's to probabilistic data X's using a Dirichlet distribution. We used scale parameters a for the correct genotype category and $(1-a)/2$ for the other two; this error model was used previously by Acar and Sun (2013) to study location tests under genotype group uncertainty. We varied a from 1 to 0.5, where $a = 1$ corresponds to no genotype uncertainty and $a = 0.5$ implies that, on average, 50% of the "best-guess" genotypes correspond to the true genotypes. Thus, the genotype group uncertainty level ranged from 0 to 50% in our simulations.

We then simulated outcome data for each sib-pair similarly as for model 1 above. For each of the $n/2$ sib-pairs, we first simulated paired data from $BV\mathcal{N}(0, 1, \rho)$, where $\rho = 0.5$ was the within sib-pair correlation. For each simulated value w, we then applied the $\sigma_k g(w)$ transformation to obtain the desired outcome data y as in simulation model 1 (Gaussian, Student's t_4, and χ^2_4). However, k here refers to the corresponding true underlying genotype group of an individual, and two individuals from the same sib-pair might not have the same genotype. We used $(\sigma_0^2, \sigma_1^2, \sigma_2^2) = (1, 1, 1)$ to study type 1 error control, and $(1, 1.5, 2)$ or $(1, 2, 4)$ to study power; other values such as $(2, 1.5, 1)$ and $(4, 2, 1)$ were also investigated.

It is evident from the earlier simulation results of model 1 that the original Lev test is not valid in the presence of sample correlation, and TW_{OLS} and gS_{OLS} are inferior, respectively, to TW_{LAD} and gS_{LAD}, when the error structure is non-symmetric or the group sizes are small. Therefore, the results presented below focus on comparison between TW_{LAD} and gS_{LAD}. In the presence of genotype group uncertainty, we also considered the "best-guess" approach and used TW^{BG}_{LAD} and gS^{BG}_{LAD} to represent the corresponding results.

Table 2
Type 1 error evaluation under simulation model 2 without group uncertainty

	Gaussian		Student's t_4		χ^2_4	
$n/2$	TW_{LAD}	gS_{LAD}	TW_{LAD}	gS_{LAD}	TW_{LAD}	gS_{LAD}
			MAF = 0.1			
20	0.110	0.040	0.109	0.042	0.113	0.044
50	0.117	0.043	0.140	0.046	0.160	0.044
100	0.092	0.048	0.115	0.049	0.118	0.047
500	0.056	0.048	0.068	0.047	0.070	0.052
1000	0.055	0.050	0.061	0.049	0.058	0.045
			MAF = 0.2			
20	0.068	0.039	0.072	0.040	0.092	0.050
50	0.074	0.042	0.086	0.041	0.095	0.046
100	0.060	0.048	0.072	0.044	0.078	0.051
500	0.055	0.051	0.055	0.047	0.057	0.052
1000	0.051	0.051	0.053	0.051	0.056	0.051

Parameter values included $n/2$ for the number of sib-pairs, $\rho = 0.5$ for the within-pair correlations, and $q = 0.1$ or 0.2 for the minor allele frequency (MAF) of the SNP of interest; on average the expected sizes of the three genotype groups are $n(1-q)^2, n2q(1-q)$ and nq^2. Without loss of generality, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1$ for type 1 error rate evaluation. The empirical type 1 error was estimated from 10,000 simulated replicates at the nominal 5% level.

For completeness, we also compared the power of gS_{LAD} along-side gL and $gJLS$ using sibling data with genotype uncertainty. To do this, after simulating the genotype data for sib-pairs as above ($a = 0.7$ for genotype uncertainty at 70%), we then induced both scale differences $(\sigma_0^2, \sigma_1^2, \sigma_2^2) = (1, 1.5, 2)$ and location differences $(\mu_0, \mu_1, \mu_2) = (0, 0.3, 0.6)$ in the outcome between the three genotype groups; other parameter values were also considered (results not shown). As before, the outcome simulation used the true underlying genotype group membership for each individual, while the test statistics were constructed using either "best-guess" genotypes or incorporating the group membership probabilities as proposed. To numerically demonstrate the loss in power when limiting analysis to independent observations, we also analyzed the "best-guess" genotype dataset after randomly discarding one individual from each sibling pair.

3.2.2. Results. In the presence of sample correlation but with no group uncertainty, the results in Table 2 show that both TW_{LAD} and gS_{LAD} were accurate in large samples, for example, when sample size was 2000 ($n/2 = 1000$ sib-pairs). However, TW_{LAD} had increased type 1 error when group sizes were unbalanced and relatively small, even for Gaussian data. For example, when the MAF is $q = 0.2$ and the number of sib-pairs is $n/2 = 100$, the expected number of observations for the three genotype groups are $n * ((1-q)^2, 2q(1-q), q^2) = (128, 64, 8)$. In that case, the empirical type 1 error of TW_{LAD} was 0.060, 0.072, and 0.078 for Gaussian, t_4 and χ^2_4 data, respectively. The problem was exacerbated by a smaller MAF $q = 0.1$ with empirical type 1 error levels of 0.092, 0.115, and

Table 3
Type 1 error evaluation under simulation model 2 with 30% group uncertainty

	Gaussian				Student's t_4				χ^2_4			
$n/2$	TW^{BG}_{LAD}	TW_{LAD}	gS^{BG}_{LAD}	gS_{LAD}	TW^{BG}_{LAD}	TW_{LAD}	gS^{BG}_{LAD}	gS_{LAD}	TW^{BG}_{LAD}	TW_{LAD}	gS^{BG}_{LAD}	gS_{LAD}
	MAF = 0.1											
20	0.067	0.074	0.036	0.037	0.083	0.079	0.044	0.046	0.088	0.090	0.047	0.050
50	0.066	0.062	0.045	0.045	0.076	0.062	0.046	0.046	0.084	0.076	0.049	0.053
100	0.058	0.057	0.045	0.046	0.064	0.059	0.047	0.046	0.072	0.069	0.051	0.049
500	0.057	0.054	0.054	0.052	0.056	0.053	0.052	0.048	0.055	0.052	0.050	0.048
1000	0.051	0.055	0.052	0.054	0.053	0.050	0.052	0.049	0.054	0.052	0.049	0.047
	MAF = 0.2											
20	0.061	0.062	0.040	0.039	0.065	0.063	0.037	0.044	0.075	0.073	0.047	0.050
50	0.053	0.053	0.046	0.045	0.063	0.059	0.046	0.050	0.069	0.070	0.051	0.053
100	0.049	0.051	0.046	0.045	0.057	0.053	0.047	0.049	0.059	0.058	0.049	0.047
500	0.051	0.049	0.049	0.051	0.051	0.052	0.048	0.050	0.053	0.052	0.048	0.051
1000	0.049	0.046	0.047	0.047	0.052	0.049	0.047	0.049	0.055	0.053	0.050	0.052

Superscript ^{BG} denotes TW_{LAD} and gS_{LAD} applied to the "best-guess" genotype data. The true genotype data were masked using a Dirichlet distribution for the genotype probabilities with scale parameters a for the correct genotype and $(1-a)/2$ for the other two. On average, $a = 0.7$ corresponds to 30% group uncertainty level. See legend of Table 2 for additional simulation details.

0.118, respectively, for the three types of data. In contrast, the proposed gS_{LAD} test remained accurate in most cases and was slightly conservative in small samples when $n/2 < 100$.

Results in Table 3 are characteristically similar to those of Table 2. However, we note that group uncertainty somewhat mitigates the problem of unbalanced group sizes, and consequently the accuracy issue of TW_{LAD}. Nevertheless, it is clear that gS_{LAD} had better type 1 error control than TW_{LAD} across the MAF values and the three outcome distributions. As expected, under the null hypothesis TW^{BG}_{LAD} and gS^{BG}_{LAD} using the "best-guess" genotype group have similar type 1 error control to TW_{LAD} and gS_{LAD} incorporating the group probabilistic data (Table 3 and Web Tables 2 and 3).

Focusing on the accurate gS_{LAD} test, Table 4 and Figure 1 demonstrate the gain in power when incorporating the group probabilistic data into the inference (gS_{LAD}) as compared to the"best-guess" approach (gS^{BG}_{LAD}). For example, at 30% group uncertainty with sample size of 1000 ($n/2 = 500$ sib-pairs), MAF of 0.1, and Gaussian data, the power of gS_{LAD} was 0.613, a 23% increase over the power of 0.495 observed for gS^{BG}_{LAD}; similar gains in efficiency were observed for other sample sizes, MAF, and t_4 and χ^2_2 data (Table 4).

One would expect the relative efficiency gain to increase as uncertainty level increases. However, this is true only if the uncertainty level is not too high. Depending on the model used to induce group uncertainty and the heteroscedasticity alternatives, it is reasonable to assume that the absolute power eventually converges to the type 1 error rate as the uncertainty increases. Consequently, the gain in relative efficiency of gS_{LAD} as compared to gS^{BG}_{LAD} would also diminish and converge to 1. This is consistent with results in Figure 1.

To study the ensuing $gJLS$ joint location-scale test (using the LAD version for the gS component), Web Table 5 illustrates the anticipated power gain when analyzing the full data using all available individuals and incorporating the genotype group probabilities (*Sibs*), as compared to using the "best-guess" genotypes ($Sibs^{BG}$), or using a reduced subset of independent individuals (Ind^{BG}). For completeness, Web Table 5 also includes results for the individual gS and gL testing components for the three Ind^{BG}, $Sibs^{BG}$, and *Sibs* datasets. Within each method, the estimated power increases as the analysis uses more information available from the data; this

Table 4
Power of gS^{BG}_{LAD} and gS_{LAD} under simulation model 2 with 30% group uncertainty

	Gaussian		Student's t_4		χ^2_4	
$n/2$	gS^{BG}_{LAD}	gS_{LAD}	gS^{BG}_{LAD}	gS_{LAD}	gS^{BG}_{LAD}	gS_{LAD}
	MAF = 0.1					
20	0.064	0.066	0.067	0.077	0.050	0.064
50	0.079	0.087	0.077	0.081	0.089	0.089
100	0.124	0.152	0.087	0.112	0.101	0.117
500	0.495	0.613	0.314	0.420	0.376	0.442
1000	0.795	0.885	0.533	0.671	0.634	0.759
	MAF = 0.2					
20	0.050	0.066	0.062	0.058	0.063	0.074
50	0.089	0.120	0.084	0.089	0.091	0.104
100	0.166	0.196	0.114	0.129	0.129	0.160
500	0.668	0.784	0.471	0.582	0.499	0.608
1000	0.939	0.985	0.739	0.846	0.810	0.896

gS^{BG}_{LAD} denotes gS_{LAD} applied to the "best-guess" genotype data. The true genotypes were masked using a Dirichlet distribution for the genotype probabilities with scale parameters a for the correct genotype and $(1-a)/2$ for the other two. On average, $a = 0.7$ corresponds to 30% group uncertainty. Besides the parameters shown in the table, other values include $\rho = 0.5$ for within-pair correlation, and $(\sigma^2_0, \sigma^2_1, \sigma^2_2) = (1, 1.5, 2)$. Power was estimated from 1000 simulated replicates at the 5% level.

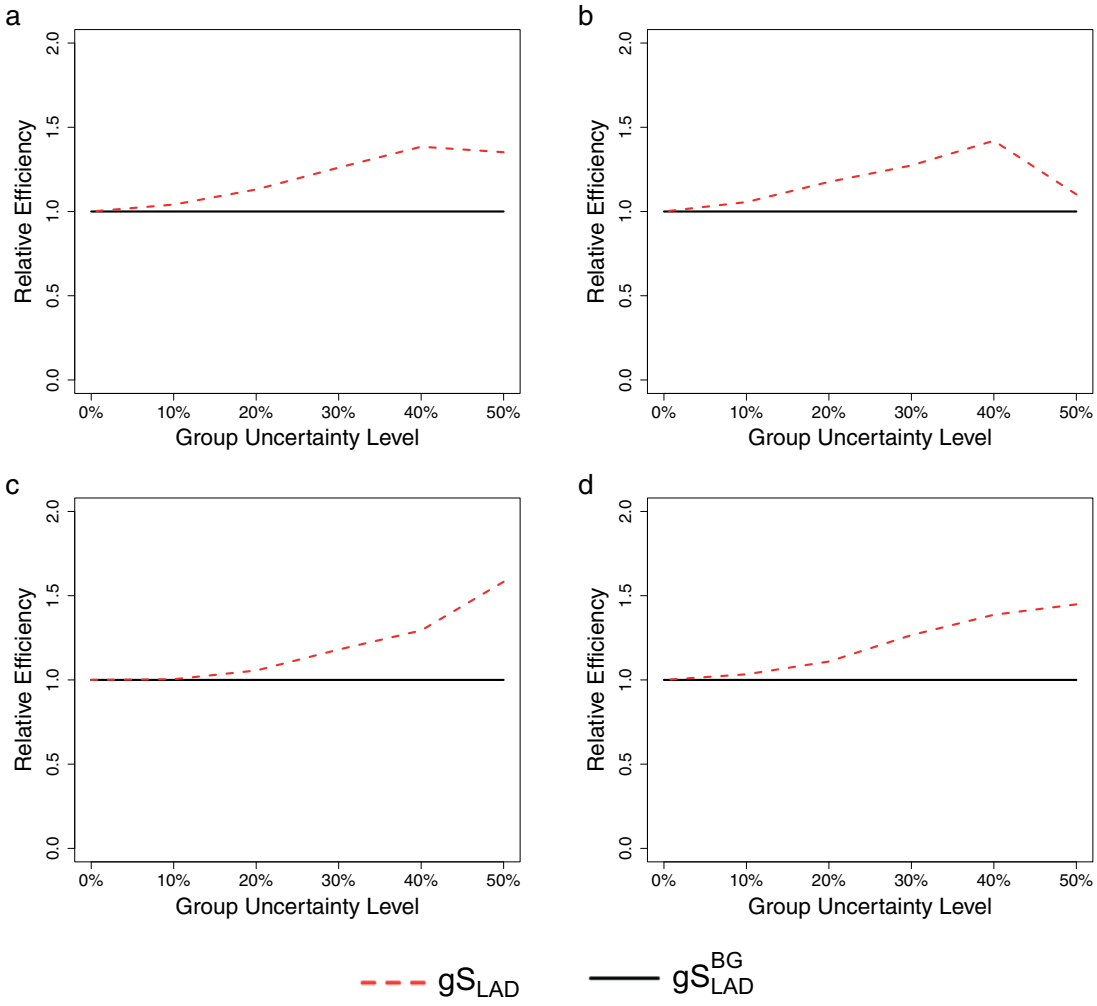

Figure 1. Relative efficiency of gS^{BG}_{LAD} and gS_{LAD} under simulation model 2. gS^{BG}_{LAD} denotes gS_{LAD} applied to the "best-guess" genotype data, and the relative efficiency is the power of gS_{LAD} divided by the power of gS^{BG}_{LAD}. The true genotypes were masked using a Dirichlet distribution with scale parameters a for the correct genotype and $(1-a)/2$ for the other two. On average, $a = 0.5$ corresponds to 50% group uncertainty, and $a = 1$ corresponds to no genotype uncertainty (i.e., 0%). Parameter values included $n/2 = 500$ sib pairs, and $\rho = 0.5$ within-pair correlation. The MAF and $(\sigma_0^2, \sigma_1^2, \sigma_2^2)$ were 0.1 and (1, 1.5, 2) for (a), 0.1 and (2, 1.5, 1) for (b), 0.2 and (1, 1.5, 2) for (c), and 0.2, (2, 1.5, 1) for (d). Power was estimated from 1,000 simulated replicates at the 5% level. The absolute power of gS^{BG}_{LAD} and gS_{LAD} at the 30% group uncertainty level for (a) and (c) are presented in Table 4 under Gaussian data.

was already demonstrated for the proposed gS in the earlier simulations (e.g., Table 3) and for gL in the existing literature. A direct comparison between gL and gS can be difficult to interpret, because the individual power depends on two different sets of parameters, (μ_0, μ_1, μ_2) and $(\sigma_0^2, \sigma_1^2, \sigma_2^2)$. However, it is clear that Fisher's method combines the evidence from gS and gL in a cumulative fashion resulting in $gJLS$ being more powerful than either test alone.

4. Applications

To demonstrate the utility of the proposed generalized scale (gS) test (gS_{LAD} using LAD in the stage 1 regression) and subsequent generalized joint location-scale ($gJLS$) test, we revisited the two genetic association studies considered in Soave et al. (2015), and compared our results with those using only a sample of unrelated individuals with no genotype group uncertainties. We also used application data combined with simulation methods to further empirically validate the performance of the proposed methods.

4.1. *HbA1c Levels in Subjects with Type 1 Diabetes*

We use this application to demonstrate the gain in power by incorporating group uncertainty (probabilistic) data. Details of this dataset were previously reported in Soave et al. (2015). Briefly, the outcome of interest was inverse normal transformed HbA1c levels in $n = 1304$ unrelated subjects with type 1 diabetes, and the SNP of interest was rs1358030 near *SORCS1* on chromosome 10 with MAF of 0.36. With no sample correlation or group uncertainty, the original Lev test

Table 5
Application study of lung function severity in 1,507 patients with cystic fibrosis

					$n_{indep} = 1409$			$n_{all} = 1507$		
Chr	Gene	SNP	bp-Position[a]	MAF	*L*ocation	*S*cale	*JLS*	*gL*	*gS*	*gJLS*
1	*SLC26A9*	rs7512462	204,166,218	0.41	0.30	0.58	0.48	0.30	0.39	0.36
1	*SLC26A9*	rs4077468	204,181,380	0.42	0.53	0.61	0.69	0.45	0.59	0.62
1	*SLC26A9*	rs12047830	204,183,322	0.49	0.55	0.15	0.29	0.52	0.11	0.22
1	*SLC26A9*	rs7419153	204,183,932	0.37	0.50	0.06	0.14	0.73	0.09	0.24
1	*SLC26A9*	rs11240600[b]	204,187,869	0.33	0.14	0.62	0.30	0.11	0.65	0.27
5	*SLC9A3*	rs17563161	550,624	0.26	0.0004	0.02	0.0001	0.0002	0.02	5.6×10^{-5}
5	*SLC9A3*	rs11134081[b]	557,404	0.35	0.0006	0.17	0.001	0.0006	0.05	0.0003
X	*SLC6A14*	rs12839137	115,479,578	0.24	0.02	0.08	0.01	0.01	0.16	0.02
X	*SLC6A14*	rs5905283	115,479,909	0.49	0.009	0.07	0.005	0.005	0.18	0.007
X	*SLC6A14*	rs3788766	115,480,867	0.40	0.001	0.01	0.0002	0.0004	0.02	$9.5 \times 10{-}5$
X	*SLC6A14*	rs62605921[b]	115,475,499	0.24	0.02	0.14	0.02	0.01	0.22	0.01

There were 1,313 singletons, 94 sib-pairs, and two sib-trios in the whole sample, resulting in $n_{indep} = 1,313 + 94 + 2 = 1409$ unrelated individuals, and $n_{all} = 1313 + 94 * 2 + 2 * 3 = 1507$ individuals. In addition to the eight genotyped SNPs originally analyzed in Soave et al. (2015), three imputed SNPs (rs11240600, rs11134081, and rs62605921) were also included in the current analysis. Results using the n_{indep} sample were from Soave et al. (2015) for the eight genotyped SNPs, or obtained here for the three imputed SNPs (using the "best-guess" genotypes), where the standard regression *Location* test, Levene's *Scale* test, and the *JLS* joint location-scale test were used. Results for the n_{all} sample were obtained using the corresponding generalized tests, $gL, gS, gJLS$, incorporating all available information.
[a]hg18 assembly (March 2006; NCBI36).
[b]The three imputed SNPs, imputed using the Beagle software version 4.1 (Browning and Browning, 2016). The imputation quality is measured by allelic R^2 provided by Beagle, and the linkage disequilibrium is measured by r^2 with the adjacent genotype SNP. rs11240600, imputation $R^2 = 0.86$, LD $r^2 = 0.27$ (with rs7512462); rs11134081, imputation $R^2 = 0.72$, LD $r^2 = 0.41$ (with rs17563161); rs62605921, imputation $R^2 = 0.91$, LD $r^2 = 0.42$ (with rs3788766).

was applied and resulted in a significant result with $p = 0.01$ (Soave et al., 2015). Combined with other evidence reported in Paterson et al. (2010), we assume here that the association is real and smaller p-values imply better performance.

To demonstrate the effect of genotype group uncertainty, we masked the true genotypes of rs1358030 using the same Dirichlet distribution as in the simulation studies above, where the value of a ranged from 1 to 0.5, corresponding to no group uncertainty to 50% uncertainty. We then applied gS^{BG} to the "best-guess" data and the proposed gS incorporating the probabilistic data, and obtained the corresponding p-values, $p_{gS^{BG}}$, and p_{gS}. For a given uncertainty level, we repeated the masking process independently 1000 times and obtained averaged p-values on the log10 scale ($10^{\{\text{average of } log10(p)\}}$), $\overline{p}_{gS^{BG}}$, and $\overline{p}_{gS}$. Between the two methods, it was clear that gS was more efficient than gS^{BG}. For example, when $a = 0.75$ for 25% group uncertainty, the gS test remains significant with $\overline{p}_{gS} = 0.048$ as compared to $\overline{p}_{gS^{BG}} = 0.068$. However, regardless of the method used, the power of the scale tests decreased sharply as genotype uncertainty increased, consistent with results for location tests reported in Acar and Sun (2013).

In addition to simulating group uncertainty for genotyped rs1358030, we also analyzed some ungenotyped variants that were imputed in the region surrounding rs1358030. Imputation was done using the IMPUTE2 software (Howie et al., 2009, 2011); see Paterson et al. (2010) for imputation details. Variants were chosen to include different levels of imputation quality (based on "info score" provided by IMPUTE2) and correlation/linkage disequilibrium (LD as measured by r^2) with rs1358030, the original SNP of interest. Web Table 7 presents the results of gL, gS, and $gJLS$ using the "best guess" genotype or incorporating the genotype probabilities. As expected, variants in high LD/correlation with rs1358030 were imputed with high quality (i.e., with little genotype uncertainty), thus they yielded similar results between the two approaches (e.g., rs5787660). In general, we obtained smaller p-values for analysis using the genotype probabilities as compared to the "best-guess" genotypes, with larger differences for SNP of lower imputation quality. However, as discussed earlier, an alternative SNP starts to behave like a null SNP when the genotype uncertainty is too high leading to no power regardless of the methods. These agree with the earlier observations and simulation results in Figure 1.

4.2. *Lung Disease Severity in Individuals with Cystic Fibrosis*

We used this application to demonstrate the gain in power by incorporating all available information including related subjects and genotype probabilities. We also used this dataset combined with permutation methods to further demonstrate the validity of the proposed methods. Details of this dataset were previously reported in Soave et al. (2015). The outcome of interest was a measure of lung function severity based on forced expiratory volume in 1 second, obtained on a total of $n_{all} = 1507$ individuals with CF (1313 singletons, 188 from 94 sib-pairs, and six from two sib-trios).

Focusing on the $n_{indep} = 1313 + 94 + 2 = 1409$ unrelated individuals and eight genotyped SNPs, Soave et al. (2015) performed an association study using the original *Location*, *Scale*, and joint location-scale (*JLS*) tests. These SNPs were from three genes (*SLC26A9*, *SLC9A3*, and *SLC6A14*), chosen

based on association evidence for other CF-related outcomes as reported in Sun et al. (2012) and Li et al. (2014). Soave et al. (2015) concluded that SNPs from *SLC9A3* and *SLC6A14* were associated with CF lung disease, and their results are included in Table 5.

Here, we used all $n_{all} = 1507$ related individuals, and we not only re-analyzed the eight *genotyped* SNPs, but also investigated three *imputed* SNPs previously not studied, using the generalized tests, gL, gS, and $gJLS$ (Table 5). The imputation was previously done using Beagle version 4.1 (Browning and Browning, 2016); see Li et al. (2014) for details. The imputed variants were chosen such that each of the three genes contained a new variant and the imputation quality was reasonably high. Since our analysis involved only (known) siblings and singletons, we used a compound symmetric correlation structure (a single correlation parameter ρ) to model within family dependence for each application of the GLS regression for the gS component.

We first note that the conclusions for the presumed null SNPs from *SLC26A9* did not change, as desired. The conclusions for the presumed associated SNPs from *SLC9A3* and *SLC6A14* did not change either, but using all available data led to overall smaller p-values for the generalized tests. The apparent lack of a large efficiency gain was somewhat disappointing, but it was also expected given the few number of siblings ($n_{omit} = 94 + 2 * 2 = 98$) added to the full sample; see Section 5 for additional comments. Nevertheless, the application clearly demonstrates the advantage of using all the available information (all samples and the genotype probabilistic data). For example, for the imputed *rs*11134081 variant, the p-value of the original Levene's *Scale* test is 0.17 analyzing the smaller sample of independent individuals and their "best guess" genotypes, while the p-value of the gS test is 0.05. Lastly, we note that the JLS framework indeed yields increased power when aggregating evidence from the individual tests; see Soave et al. (2015) for detailed discussions of the motivation and merits of the joint-testing framework.

To further examine the accuracy of the proposed gS and $gJLS$ tests (as well as the gL test for completeness), we generated 10,000 permutation replicates of the outcome to assess the empirical type 1 error control. Permutation was performed separately between singletons and between sib-pairs; see Abney (2015) for permutation techniques for more general family data. Without loss of generality, we focused on SNP rs17563161 from *SLC9A3* (Web Figure 1). Testing the resulting p-values for deviation from the expected Uniform(0,1) distribution using the Kolmogorov–Smirnov test showed that all tests were valid.

5. Discussion

Levene's scale test is widely used as a model diagnostic tool in linear regression, and more recently it has been employed as an indirect test for interaction effects. Increased data complexity due to sample correlation or group uncertainty, however, limits its applicability. Here, we proposed a generalization of Levene's scale test, gS, that has good type 1 error control in the presence of sample correlation, small samples, unbalanced group sizes, and non-symmetric outcome data. We showed that the least absolute deviation (LAD) regression approach to obtain group-*median*-adjusted residuals is needed to ensure robust performance of gS. Based on our results, we recommend the use of gS_{LAD} over gS_{OLS} (and other existing tests) uniformly for all studies analyzed.

In the presence of group membership uncertainty, gS incorporating the probabilistic data increases power compared to using the "best-guess" group data. However, based on the simulations considered here, we note that when the group uncertainty level is moderate (e.g., 30%), the efficiency gain is also moderate (Table 4 and Figure 1). When the group uncertainty is too high, the relative efficiency gain may diminish because the absolute power decreases considerably and eventually converges to the type 1 error rate.

In the presence of sample correlation, the original Lev test is inadequate due to inflated type 1 error. Using a subset of only unrelated individuals would improve the accuracy of Lev but at a cost to the power. The size of the efficiency loss depends on the proportion omitted from the sample as well as the dependency structure. The TW method of Iachine et al. (2010) extends the Lev test for twin data. Their simulation study as well as ours showed that TW has an increased type 1 error rate when group sizes are unbalanced and relatively small, in contrast to the proposed gS. When all group sizes were large, gS and TW were empirically equivalent.

To further study the effect of misspecification of sample correlation structure on type 1 error and power, we conducted an additional simulation study considering three types of misspecifications. Without loss of generality, we revisited simulation model 1 with 20 MZ and 20 DZ twin pairs of Gaussian data, and we focused on the Lev_{LAD}, TW_{LAD}, and gS_{LAD} tests. Web Table 6 clearly shows that in the absence of sample correlation, methods (TW_{LAD} and gS_{LAD}) that model correlation retain correct type 1 error and have negligible power loss. In the presence of correlation, methods (Lev_{LAD}) that ignore correlation are not robust as demonstrated before (cf. Tables 2 and 4 in Iachine et al. (2010)). However, if the true correlation structure is misspecified as considered here, TW_{LAD} and gS_{LAD} can have increased type 1 error rates. Although this type of complete misclassification is unlikely in many practical settings such as the genetic association studies, the potential detrimental effects of correlation misspecification merit further investigation.

In the CF application, although gS yielded comparable or less significant results after incorporating siblings in the analysis, we observed that the corresponding gL test results were more significant. We considered the possibility that even though scale differences existed in the data, the addition of only 98 siblings (7% increase from the independent sample) may not yield a noticeable improvement of gS. Using the setup of simulation model 1, we examined the effect of incorporating only a small proportion of additional related subjects to an otherwise independent sample (Web Table 4). We found that, compared with using a sample of 1000 singletons, using $n = 900$ singletons along with 100 sib-pairs (10% increase) led to a <5% power increase. In contrast, the addition of siblings to all unrelated subjects provided a substantial increase in power (Web Table 4). These results, and the noticeable power gain from the gL test when applied to the same CF data, are consistent with observations in genetic association studies that, larger samples are needed to detect variance compared to mean differences (Visscher and Posthuma, 2010).

The expression of $E(d_i) = \sigma_i \sqrt{\frac{2}{\pi}(1 - h_{ii})}$ in Section 2.2 suggests that the stage 2 regression of (4) could be improved by rescaling the d_i's by $(1 - h_{ii})^{-1/2}$. This adjustment has been shown to improve the type 1 error control of Levene's original test for small samples with group design imbalance (Keyes and Levy, 1997). Examination of this rescaling for gS under simulations involving correlated data, however, led to instances of increased type 1 error (results not shown). Thus, further investigation is required to propose an appropriate adjustment. Another potential improvement to the analysis of regression model (4) is from the recognition that the d_i's are based on residuals, $\widehat{\varepsilon}_i$'s, and thus correlated even when there is no sample correlation among the true disturbances, ε_i's. O'Neill and Mathews (2000) derived expressions for the covariance matrix of $\boldsymbol{d}$ for independent observations with no group uncertainty, showing that the correlation across the d_i's disappears as the group sizes increase. For the complex data scenarios considered here, gS_{LAD} appears robust for even small samples. Nevertheless, the potential for gain in efficiency by accounting for this type of correlation merits additional consideration.

The developments here did not consider additional covariates, $\boldsymbol{z}$, for example, age and sex in genetic association studies. The extension is straightforward if the effects of $\boldsymbol{z}$ on y are strictly on the mean. In that case, including $\boldsymbol{z}$ as part of the design matrix in stage 1 suffices. However, if $\boldsymbol{z}$ also influences the variance of y, not including $\boldsymbol{z}$ as part of the design matrix in stage 2 may lead to increased type 1 error of testing the γ_j's that are associated with the primary covariates of interest. This is the same phenomenon as observed in location-testing where omitting potential confounders can lead to spurious association.

Joint location-scale testing is becoming a popular method for complex outcome-covariate association data, where the conventional location-only analysis may be underpowered. This scenario has received attention in many fields, including our motivating example of genetic epidemiology (Soave et al., 2015). The proposed gS test allows investigators to combine evidence from scale tests with existing generalized location tests via the JLS testing framework of (Soave et al., 2015). The CF application study showed that individual location or scale tests can provide more significant results when utilizing related individuals and incorporating the available group probabilistic data, which in turn may lead to a more powerful $gJLS$ test.

6. Supplementary Materials

Web Appendix Sections A, B, and C, Web Figure 1, and Web Tables S1-S7 referenced in Sections 2, 3, 4, and 5 are available with this article at the Biometrics website on Wiley Online Library. R code is available at https://github.com/dsoave/gJLS.

Acknowledgements

The authors thank Professor Jerry F. Lawless and Dr. Lisa J. Strug for helpful suggestions and critical reading of the original version of the article. The authors thank Dr. Andrew Paterson and Dr. Lisa J. Strug for providing the type 1 diabetes and the cystic fibrosis application data, respectively. This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR) to LS. DS is a trainee of the CIHR STAGE (Strategic Training in Advanced Genetic Epidemiology) training program at the University of Toronto and is a recipient of the SickKids Restracomp Studentship Award and the Ontario Graduate Scholarship (OGS).

References

Abney, M. (2015). Permutation testing in the presence of polygenic variation. *Genet Epidemiol* **39**, 249–258.

Acar, E. F. and Sun, L. (2013). A generalized Kruskal–Wallis test incorporating group uncertainty with application to genetic association studies. *Biometrics* **69**, 427–435.

Aitken, A. C. (1936). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55**, 42–48.

Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *Journal of the American Statistical Association* **75**, 890–894.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **160**, 268–282.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–1294.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association* **69**, 364–367.

Browning, B. L. and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* **98**, 116–126.

Cameron, A. C. and Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources* **50**, 317–372.

Cao, Y., Wei, P., Bailey, M., Kauwe, J. S., Maxwell, T. J., and the Alzheimer's Disease Neuroimaging Initiative (2014). A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology* **38**, 51–59.

Carroll, R. J. and Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statistics & Probability Letters* **3**, 191–194.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* **23**, 351–361.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: A review and some new results. *Statistical Science* **29**, 302–321.

Furno, M. (2005). The Glejser test and the median regression. *Sankhyā: The Indian Journal of Statistics* **67**, 335–358.

Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). *The impact of Levene's test of equality of variances on statistical theory and practice. Statistical Science* **24**, 343–360.

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association* **64**, 316–323.

Godfrey, L. G. (1996). Some results on the Glejser and Koenker tests for heteroskedasticity. *Journal of Econometrics* **72**, 275–299.

Haseman, J. K. and Elston, R. C. (1970). The estimation of genetic variance from twin data. *Behavior Genetics* **1**, 11–19.

Horvath, S., Xu, X., and Laird, N. M. (2001). The family based association test method: Strategies for studying general genotype–phenotype associations. *European Journal of Human Genetics* **9**, 301–306.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–70.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* **5**, e1000529.

Iachine, I., Petersen, H. C., and Kyvik, K. O. (2010). Robust tests for the equality of variances for clustered data. *Journal of Statistical Computation and Simulation* **80**, 365–377.

Im, K. S. (2000). Robustifying Glejser test of heteroskedasticity. *Journal of Econometrics* **97**, 179–188.

Jakobsdottir, J. and McPeek, M. S. (2013). MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics* **92**, 652–666.

Jiang, B. and Liu, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42**, 1751–1786.

Keyes, T. K. and Levy, M. S. (1997). Analysis of Levenes test under design imbalance. *Journal of Educational and Behavioral Statistics* **22**, 227–236.

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., et al. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics* **12**, 1–17.

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics; Essays in Honor of Harold Hotelling*, I. Olkin editor, 278–292. Stanford, California: Stanford University Press.

Li, W., Soave, D., Miller, M. R., Keenan, K., Lin, F., Gong, J., et al. (2014). Unraveling the complex genetic model for cystic fibrosis: Pleiotropic effects of modifier genes on early cystic fibrosis-related morbidities. *Human Genetics* **133**, 151–161.

Lim, T.-S. and Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis* **22**, 287–301.

Machado, J. A. F. and Silva, J. M. C. S. (2000). Glejser's test revisited. *Journal of Econometrics* **97**, 189–202.

O'Neill, M. E. and Mathews, K. (2000). Theory & methods: A weighted least squares approach to Levene's test of homogeneity of variance. *Australian & New Zealand Journal of Statistics* **42**, 81–100.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Annals of Statistics* **37**, 3867–3892.

Pare, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. *PLoS Genetics* **6**, e1000981.

Paterson, A. D., Waggott, D., Boright, A. P., Hosseini, S. M., Shen, E., Sylvestre, M. P., et al. (2010). A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. *Diabetes* **59**, 539–549.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS. Statistics and computing.* New York: Springer.

Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boelle, P. Y., et al. (2015). A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways. *The American Journal of Human Genetics* **97**, 125–138.

Sun, L. (2012). Detecting pedigree relationship errors. In *Statistical Human Genetics: Methods and Protocols*, C. R. Elston, M. J. Satagopan, and S. Sun, (eds), 25–46. Totowa, New Jersey: Humana Press.

Sun, L. and Dimitromanolakis, A. (2012). Identifying cryptic relationships. In *Statistical Human Genetics: Methods and Protocols*, C. R. Elston, M. J. Satagopan, and S. Sun (eds), 47–57. Totowa, New Jersey: Humana Press.

Sun, L., Rommens, J. M., Corvol, H., Li, W., Li, X., Chiang, T. A., et al. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature Genetics* **44**, 562–569.

Sun, X., Elston, R., Morris, N., and Zhu, X. (2013). What is the significance of difference in phenotypic variability across snp genotypes? *The American Journal of Human Genetics* **93**, 390–397.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics* **39**, 173–188.

Visscher, P. M. and Posthuma, D. (2010). Statistical power to detect genetic loci affecting environmental sensitivity. *Behavior Genetics* **40**, 728–733.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Received June 2016. *Revised December* 2016.
Accepted December 2016.