The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 13

Resampling-based Methods in Single and Multiple Testing for Equality of Covariance/ Correlation Matrices

Yang Yang, *University of Florida*
Victor DeGruttola, *Harvard University*

Recommended Citation:
Yang, Yang and DeGruttola, Victor (2012) "Resampling-based Methods in Single and Multiple Testing for Equality of Covariance/Correlation Matrices," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 13.
DOI: 10.1515/1557-4679.1388

Resampling-based Methods in Single and Multiple Testing for Equality of Covariance/ Correlation Matrices

Yang Yang and Victor DeGruttola

Abstract

Traditional resampling-based tests for homogeneity in covariance matrices across multiple groups resample residuals, that is, data centered by group means. These residuals do not share the same second moments when the null hypothesis is false, which makes them difficult to use in the setting of multiple testing. An alternative approach is to resample standardized residuals, data centered by group sample means and standardized by group sample covariance matrices. This approach, however, has been observed to inflate type I error when sample size is small or data are generated from heavy-tailed distributions. We propose to improve this approach by using robust estimation for the first and second moments. We discuss two statistics: the Bartlett statistic and a statistic based on eigen-decomposition of sample covariance matrices. Both statistics can be expressed in terms of standardized errors under the null hypothesis. These methods are extended to test homogeneity in correlation matrices. Using simulation studies, we demonstrate that the robust resampling approach provides comparable or superior performance, relative to traditional approaches, for single testing and reasonable performance for multiple testing. The proposed methods are applied to data collected in an HIV vaccine trial to investigate possible determinants, including vaccine status, vaccine-induced immune response level and viral genotype, of unusual correlation pattern between HIV viral load and CD4 count in newly infected patients.

KEYWORDS: resampling, covariance, correlation, multiple testing, robust test

Author Notes: This work was supported by the National Institute of Health grant R01-AI51164. We thank Dr. Peter Gilbert at Fred Hutchinson Cancer Research Center for sharing the VAX004 data with us.

1 Introduction

For HIV-infected patients, there is interest in understanding how clinical responses covary over time. Although increasing viral load tends to be associated with decreasing CD4 T lymphocyte count, the reasons for variability in this relationship are not understood. The inverse relationship arises because viral replication is associated with death of CD4 T lymphocytes, but Yu *et al.*(1994) showed that the replication rate does not entirely determine the death rate of CD4+ T lymphocytes. Below we investigate genetic factors that may account for some of this variability, by developing statistical tests to detect determinants associated with unusual covariation of viral RNA level and CD4 cell count.

The problem of testing for homogeneity in covariance/correlation matrices across pre-defined groups arises in multivariate analysis of variance and in discriminant analysis; asymptotic tests based on likelihood ratio for Gaussian data have been extensively studied (Box, 1949; Bartlett, 1951; Manly and Rayner, 1987). However, asymptotic tests are known to be incapable of controlling type I error for heavy-tailed error distributions with worsening performance as sample size increases (Brown, 1974). More recent development within this framework involves robust estimation for dispersion, such as winsorization and deviation from the median (Tiku and Balakrishnan, 1985; O'brien, 1992) designed to reduce the sensitivity of the tests to nonnormality in data.

Resampling-based approaches for testing covariance and correlation have received much less attention than their counterparts for testing mean structures. Zhang and Boos (1992) proposed the use of the Bartlett statistic, a function of determinants of sample covariance matrices and closely related to the log likelihood ratio for Gaussian data, and a bootstrap procedure that resamples centered responses. The resampling approach does not require an assumption of normality. Zhu *et al.*(2002) suggested an alternative test statistic, which essentially compares the eigen-values of sample covariance matrices across groups. They showed the superiority of the eigenvalue-based statistics over the determinant-based statistic for responses of relatively high dimension. In both methods, responses are centered by group-specific sample means, but are not standardized by group-specific sample covariance matrices.

In their discussion of testing equal univariate variances, Westfall and Young (1993) pointed out the importance of standardizing centered responses with group-specific sample standard deviations, so that the resampling-based null distributions remain correct regardless of whether the hypothesis under consideration is true or false. This suggestion can be viewed as an extension

to hypothesis testing of the general resampling method for interval estimation proposed by Wu (1986), in which standardized residuals are resampled to induce the bias-robustness of an estimator against variance heteroscedasticity of the errors. Generalizing their previous approach to a broader class of null hypotheses, Zhang and Boos (1993) developed a procedure called "separate bootstrap plan", which indeed resamples residuals standardized by group-specific sample covariance matrices. However, this approach appears to inflate type I errors at least for small-size samples or heavy-tailed error distributions.

Resampling standardized residuals is particularly appealing in the context of multiple-testing of homogeneity in covariance or correlation matrices. These residuals are free of the data-generating second-order moments, and are therefore exchangeable across groups under the assumption that the residuals share higher moments after standardization. As a result, all residuals can be used throughout a stepwise procedure for multiple comparisons, regardless of which hypotheses are rejected. In contrast, centered responses from groups for which the null hypotheses have been rejected during previous steps can not be used in subsequent steps, leading to a shrinking pool of useable residuals for resampling along a stepwise procedure.

We propose to improve the "separate bootstrap plan" of Zhang and Boos (1993) with robust estimation for group-specific means and covariance matrices, to attain better control of type I error in the test of equal covariance across groups. In this approach, both the test statistic and residuals are calculated from robust moment estimates. We also modify the eigenvalue-based test statistic of Zhu *et al.*(2002) in such a way that, under the null hypothesis, the new statistic can be written as a function of standardized random errors whose distribution is independent of the moments under testing.

After replacing the sample covariance with the sample correlation matrices, both statistics are extended to the problem of testing equal correlation matrices across groups. Simulation studies are conducted to evaluate the performance of the proposed robust bootstrap test For both single- and multiple-testing settings, simulation studies compare the performance of the proposed robust bootstrap test that resamples standardized residuals to that of the traditional bootstrap test that resamples centered responses, as well as the performance of the determinant-based statistic to that of the modified eigenvalue-based statistic. The proposed approaches are applied to data from an HIV vaccine trial to identify possible genetic patterns associated with unusual correlation between viral load and CD4 response in new infections.

2 Methods

Consider a sample stratified into J groups, each of size n_j, $j = 1, \ldots, J$, with a total sample size of $n = \sum_{j=1}^{J} n_j$. There are L clinical responses, denoted by $\boldsymbol{Y}_{jk} = (Y_{jk1}, Y_{jk2}, \ldots, Y_{jkL})^\tau$, observed for the k^{th} individual in the j^{th} group, $k = 1, \ldots, n_j$, $j = 1, \ldots, J$. For the methods we consider, $L \ll \min_j n_j$. The class of models we consider is

$$\boldsymbol{Y}_{jk} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_{jk}, k = 1, \ldots, n_j, j = 1, \ldots, J, \tag{1}$$

where $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jL})^\tau = E(\boldsymbol{Y}_{jk})$ is the mean responses in group j, and $\boldsymbol{\epsilon}_{jk} = (\epsilon_{jk1}, \ldots, \epsilon_{jkL})^\tau$ is the vector of random errors. The errors are assumed independent across individuals. Let $\text{Diag}(\boldsymbol{x})$ denote a diagonal matrix with diagonal elements given by $\boldsymbol{x}$ if $\boldsymbol{x}$ is a vector or by the diagonal elements of $\boldsymbol{x}$ if $\boldsymbol{x}$ is a matrix. We further make the following assumptions about moments of errors:

(a) $E(\boldsymbol{\epsilon}_{jk}) = \boldsymbol{0}$;
(b) $Cov(\boldsymbol{\epsilon}_{jk}, \boldsymbol{\epsilon}_{jk}) = \boldsymbol{V}_j = \boldsymbol{D}_j \boldsymbol{R}_j \boldsymbol{D}_j$, where $\boldsymbol{\sigma}_j = (\sigma_{j1}, \ldots, \sigma_{jL})^\tau$ is the vector of standard deviations, $\boldsymbol{D}_j = \text{Diag}(\boldsymbol{\sigma}_j)$, and $\boldsymbol{R}_j$ is the correlation matrix. $\boldsymbol{V}_j$ and $\boldsymbol{R}_j$ are positive definite for all j;
(c) The distributions of the errors, after appropriate standardization, are identical for all j and k. The standardization can be based on, for example, the Cholesky or spectral decomposition of $\boldsymbol{V}_j$'s.
(d) The fourth moment is finite.

2.1 Testing homogeneity in covariance

The hypothesis to be tested is

$$H_0 : \boldsymbol{V}_1 = \boldsymbol{V}_2 = \ldots = \boldsymbol{V}_J = \boldsymbol{V} \qquad \text{vs.} \qquad H_a : \boldsymbol{V}_j \neq \boldsymbol{V}_k, \exists j, k. \tag{2}$$

Let $\boldsymbol{U}_j$ and $\boldsymbol{U}$ be the lower triangle Cholesky decomposition of $\boldsymbol{V}_j$ and $\boldsymbol{V}$ respectively, that is, $\boldsymbol{V}_j = \boldsymbol{U}_j \boldsymbol{U}_j^\tau$ and $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{U}^\tau$. Let $\boldsymbol{V}_j^{-1/2}$ and $\boldsymbol{V}^{-1/2}$ be the usual square roots of $\boldsymbol{V}_j^{-1}$ and $\boldsymbol{V}^{-1}$ respectively based on spectral decomposition. We define the following quantities derived from observed data:

- $\widehat{\boldsymbol{\epsilon}}_{jk} = \boldsymbol{Y}_{jk} - \overline{\boldsymbol{Y}}_j = \boldsymbol{\epsilon}_{jk} - \frac{1}{n_j} \sum_{l=1}^{n_j} \boldsymbol{\epsilon}_{jl}$, where $\overline{\boldsymbol{Y}}_j = \frac{1}{n_j} \sum_{l=1}^{n_j} \boldsymbol{Y}_{jl}$, the residual.
- $\widehat{\boldsymbol{V}}_j = \frac{1}{n_j - 1} \sum_k \widehat{\boldsymbol{\epsilon}}_{jk} \widehat{\boldsymbol{\epsilon}}_{jk}^\tau$, the sample covariance matrix for group j, $j = 1, \ldots, J$.

- $\widehat{\boldsymbol{V}} = \frac{1}{n-J}\sum_j (n_j - 1)\widehat{\boldsymbol{V}}_j$, the estimate for the common covariance matrix under H_0.
- $\widehat{\boldsymbol{U}}_j$ and $\widehat{\boldsymbol{U}}$, the lower triangle Cholesky decomposition of $\widehat{\boldsymbol{V}}_j$ and $\widehat{\boldsymbol{V}}$.
- $\widehat{\boldsymbol{V}}_j^{-1/2}$ and $\widehat{\boldsymbol{V}}^{-1/2}$, the square roots of $\widehat{\boldsymbol{V}}_j^{-1}$ and $\widehat{\boldsymbol{V}}^{-1}$.

The determinant-based test statistic One plausible statistic for testing H_0 in (2) is the modified Bartlett's determinant-based statistic:

$$T_{Vd} = \sum_{j=1}^{J} (n_j - 1) \log \frac{|\widehat{\boldsymbol{V}}|}{|\widehat{\boldsymbol{V}}_j|}, \tag{3}$$

This statistic is the log likelihood ratio when the error distribution is Gaussian except for the use of unbiased estimates for sample covariances. Large values of the statistic imply departure from the null hypothesis. We only consider data with invertable $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{V}}_j$ for all j. Under H_0 in (2), we have $\boldsymbol{U}_j = \boldsymbol{U}$ for all j and

$$\begin{aligned} T_{Vd} &= \sum_{j=1}^{J} (n_j - 1) \log \frac{|\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}\boldsymbol{U}^{-1\tau}|}{|\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}_j}\boldsymbol{U}^{-1\tau}|} \\ &= \sum_{j=1}^{J} (n_j - 1) \log \frac{|\frac{1}{n-J}\sum_{j,k}(\boldsymbol{U}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk})(\boldsymbol{U}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk})^\tau|}{|\frac{1}{n_j-1}\sum_{k}(\boldsymbol{U}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk})(\boldsymbol{U}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk})^\tau|} \\ &= \sum_{j=1}^{J} (n_j - 1) \log \frac{|\frac{1}{n-J}\sum_{j,k}(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)^\tau|}{|\frac{1}{n_j-1}\sum_{k}(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)^\tau|}, \end{aligned} \tag{4}$$

where $\boldsymbol{\eta}_{jk} = \boldsymbol{U}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk}$ and $\bar{\boldsymbol{\eta}}_j = \frac{1}{n_j}\sum_l \boldsymbol{\eta}_{jl}$. Clearly, $\boldsymbol{\eta}_{jk}$'s are standardized errors and are thus i.i.d. across individuals. Other approaches to standardization could also be used, for example, replacing $\boldsymbol{U}_j^{-1}$ with $\boldsymbol{V}_j^{-1/2}$.

The eigenvalue-based test statistic Let $\Lambda(\boldsymbol{X})$ be the vector of eigenvalues of matrix $\boldsymbol{X}$, and $|\boldsymbol{x}|$ be the average of absolute values of the elements of vector $\boldsymbol{x}$. An alternative test statistic is

$$T_{Ve} = \frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{l=j+1}^{J} \left| \Lambda\big(\widehat{\boldsymbol{U}}^{-1}(\widehat{\boldsymbol{V}}_j - \widehat{\boldsymbol{V}}_l)\widehat{\boldsymbol{U}}^{-1\ \tau}\big) \right|, \tag{5}$$

This statistic is a modified version of the one proposed in Zhu *et al.*(2002) which uses $\Lambda\big(\widehat{\boldsymbol{V}}^{-1/2}(\widehat{\boldsymbol{V}}_j - \widehat{\boldsymbol{V}}_l)\widehat{\boldsymbol{V}}^{-1/2}\big)$. Let Chol$(X)$ be Cholesky decomposition

of matrix X. Under the null,

$$\widehat{\boldsymbol{U}}^{-1}(\widehat{\boldsymbol{V}}_i - \widehat{\boldsymbol{V}}_j)\widehat{\boldsymbol{U}}^{-1\,\tau} = \mathrm{Chol}\big(\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}\boldsymbol{U}^{-1\tau}\big)^{-1} \times \big(\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}_i\boldsymbol{U}^{-1\tau} - \boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}_j\boldsymbol{U}^{-1\tau}\big) \times \mathrm{Chol}\big(\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}\boldsymbol{U}^{-1\tau}\big)^{-1\,\tau}, \tag{6}$$

and $\boldsymbol{U}^{-1}\widehat{\boldsymbol{V}}_j\boldsymbol{U}^{-1\tau} = \frac{1}{n_j-1}\sum_k(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)^\tau$. Hence, this eigenvalue-based statistic under the null is also a function of standardized errors. This formulation is not available for the original statistic in Zhu *et al.*(2002), because

$$\begin{aligned}&\widehat{\boldsymbol{V}}^{-\frac{1}{2}}(\widehat{\boldsymbol{V}}_i - \widehat{\boldsymbol{V}}_j)\widehat{\boldsymbol{V}}^{-\frac{1}{2}}\\ \neq&\big\{\boldsymbol{V}^{-\frac{1}{2}}\widehat{\boldsymbol{V}}\boldsymbol{V}^{-\frac{1}{2}}\big\}^{-\frac{1}{2}} \times \big(\boldsymbol{V}^{-\frac{1}{2}}\widehat{\boldsymbol{V}}_i\boldsymbol{V}^{-\frac{1}{2}} - \boldsymbol{V}^{-\frac{1}{2}}\widehat{\boldsymbol{V}}_j\boldsymbol{V}^{-\frac{1}{2}}\big) \times \big\{\boldsymbol{V}^{-\frac{1}{2}}\widehat{\boldsymbol{V}}\boldsymbol{V}^{-\frac{1}{2}}\big\}^{-\frac{1}{2}}.\end{aligned}$$

The bootstrap procedure The null distribution of a test statistic is obtained by resampling the standardized errors in its formulation under H_0 from the pool of standardized residuals given by

$$\widehat{\boldsymbol{\eta}}_{jk} = \sqrt{\frac{n_j}{n_j-1}}\widehat{\boldsymbol{U}}_j^{-1}\widehat{\boldsymbol{\epsilon}}_{jk}, \tag{7}$$

where the factor $\sqrt{\frac{n_j}{n_j-1}}$ is to force its variance components to be close to 1. Alternatively, the residuals can also be standardized as $\widehat{\boldsymbol{\eta}}_{jk} = \widehat{\boldsymbol{U}}^{-1}\widehat{\boldsymbol{\epsilon}}_{jk}$, which approximate the true errors only under H_0 but will better preserve type I error. We refer to this method as the hypothesis-dependent (H-D) bootstrap test and the previous one (using (7)) as the hypothesis-independent (H-I) bootstrap test.

For a chosen test statistic T, the bootstrap test proceeds as follows:

1. Calculate $\widehat{\boldsymbol{V}}_j$, $\widehat{\boldsymbol{V}}$ and the observed test statistic T.
2. Standardize $\widehat{\boldsymbol{\epsilon}}_{jk}$ with either $\widehat{\boldsymbol{U}}_j$ (H-I) or $\widehat{\boldsymbol{U}}$ (H-D) to obtain $\widehat{\boldsymbol{\eta}}_{jk}$'s.
3. For $i = 1, \ldots, N$, at the i^{th} iteration, bootstrap from

   $$\{\widehat{\boldsymbol{\eta}}_{jk} : k = 1, \ldots, n_j, j = 1, \ldots, J\},$$

   substitute the samples for $\boldsymbol{\eta}_{jk}$'s in the formulation of T under the null, and denote the resulted statistic by $T_i^{\#}$.
4. The p-value is given by $\frac{1}{N}\sum_i I(T_i^{\#} \geq T)$, where $I(\cdot)$ is the indicator function.

The H-D bootstrap test is equivalent to a test based on bootstrapping centered observations in Zhang and Boos (1992). The H-I bootstrap test is nearly equivalent to the "separate bootstrap plan" in Zhang and Boos (1993) except that they bootstrapped standardized errors within groups, whereas we do so across groups. If validity of assumption (c) is of concern, resampling should be done within groups, i.e., $\boldsymbol{\eta}_{jk}$ is resampled from $\{\widehat{\boldsymbol{\eta}}_{jl} : l = 1, \ldots, n_j\}$.

The bootstrap sampling procedure in Zhu *et al.*(2002) using the original eigenvalue-based statistic is similar to the H-D procedure using the modified statistic except that the standardization factor $\hat{\boldsymbol{V}}^{-1/2}$ was fixed at values based on observed data. we expect the H-D bootstrap procedure to perform better than the one in Zhu *et al.*(2002) because the former accounts for uncertainty in the standardization factor.

2.2 Testing homogeneity in correlation

The hypothesis to be tested is

$$H_0 : \boldsymbol{R}_1 = \boldsymbol{R}_2 = \ldots = \boldsymbol{R}_J = \boldsymbol{R} \qquad \text{vs.} \qquad H_a : \boldsymbol{R}_j \neq \boldsymbol{R}_k, \exists j, k. \tag{8}$$

Define $\boldsymbol{\xi}_{jk} = \boldsymbol{D}_j^{-1}\boldsymbol{\epsilon}_{jk}$ as the error vector scaled by the group-specific standard deviation matrix $\boldsymbol{D}_j = \text{Diag}(\sigma_{j1}, \ldots, \sigma_{jL})$ so that $\boldsymbol{\xi}_{jk}$ has variance 1 but preserves the correlation among the components. Let $\boldsymbol{P}_j$ and $\boldsymbol{P}$ be the lower triangle Cholesky decomposition of $\boldsymbol{R}_j$ and $\boldsymbol{R}$ respectively. Define the following sample estimates for the correlation-related quantities:

- $\widehat{\boldsymbol{D}}_j = \text{Diag}(\widehat{\sigma}_{j1}, \ldots, \widehat{\sigma}_{jL})$, where $\widehat{\sigma}^2_{jl} = \frac{1}{n_j - 1}\sum_k \widehat{\epsilon}^2_{jkl}$, $l = 1, \ldots, L$;
- $\widehat{\boldsymbol{\xi}}_{jk} = \widehat{\boldsymbol{D}}_j^{-1}\widehat{\boldsymbol{\epsilon}}_{jk}$, the scaled residual vector.
- $\widehat{\boldsymbol{R}}_j = \frac{1}{n_j - 1}\sum_k \widehat{\boldsymbol{\xi}}_{jk}\widehat{\boldsymbol{\xi}}^{\tau}_{jk}$, the sample correlation matrix for group j, $j = 1, \ldots, J$.
- $\widehat{\boldsymbol{R}} = \frac{1}{n - J}\sum_j (n_j - 1)\widehat{\boldsymbol{R}}_j$, the estimate for the common correlation matrix under H_0.
- $\widehat{\boldsymbol{P}}_j$ and $\widehat{\boldsymbol{P}}$, the lower triangle Cholesky decomposition of $\widehat{\boldsymbol{R}}_j$ and $\widehat{\boldsymbol{R}}$.

Analogous to (3) and (5), two possible test statistic for $H_0 : \boldsymbol{R}_1 = \ldots = \boldsymbol{R}_J = \boldsymbol{R}_0$ are

$$T_{Rd} = \sum_{j=1}^{J}(n_j - 1)\log\frac{|\widehat{\boldsymbol{R}}|}{|\widehat{\boldsymbol{R}_j}|} \tag{9}$$

and

$$T_{Re} = \frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{l=j+1}^{J} \Big| \Lambda\big(\widehat{\boldsymbol{R}}^{-1/2} (\widehat{\boldsymbol{R}}_j - \widehat{\boldsymbol{R}}_l) \widehat{\boldsymbol{R}}^{-1/2}\big) \Big|. \tag{10}$$

Under H_0 in (8), expand $\widehat{\boldsymbol{\xi}}_{jk}$ as the following:

$$\begin{aligned} \widehat{\boldsymbol{\xi}}_{jk} &= \widehat{\boldsymbol{D}}_j^{-1} \widehat{\boldsymbol{\epsilon}}_{jk} = \Big\{ \widehat{\boldsymbol{D}}_j^{-1} \boldsymbol{D}_j \Big\} \boldsymbol{P} \Big\{ \boldsymbol{P}^{-1} \boldsymbol{D}_j^{-1} \widehat{\boldsymbol{\epsilon}}_{jk} \Big\} \\ &= \Big\{ \frac{1}{n_j - 1} \sum_{k=1}^{n_j} \mathrm{Diag}^2 (\boldsymbol{P} \boldsymbol{P}^{-1} \boldsymbol{D}_j^{-1} \widehat{\boldsymbol{\epsilon}}_{jk}) \Big\}^{-1/2} \boldsymbol{P} \Big\{ \boldsymbol{P}^{-1} \boldsymbol{D}_j^{-1} \widehat{\boldsymbol{\epsilon}}_{jk} \Big\} \\ &= \Big\{ \frac{1}{n_j - 1} \sum_{k=1}^{n_j} \mathrm{Diag}^2 \big(\boldsymbol{P} (\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j) \big) \Big\}^{-1/2} \boldsymbol{P} (\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j). \end{aligned} \tag{11}$$

where $\boldsymbol{\eta}_{jk} = \boldsymbol{U}_j^{-1} \boldsymbol{\epsilon}_{jk} = \boldsymbol{P}^{-1} \boldsymbol{D}_j^{-1} \boldsymbol{\epsilon}_{jk}$ is the standardized error. We sample $\boldsymbol{\eta}_{jk}$'s from the residuals

$$\begin{aligned} \widehat{\boldsymbol{\eta}}_{jk} &= \sqrt{n_j/(n_j-1)}\, \widehat{\boldsymbol{P}}_j^{-1} \widehat{\boldsymbol{\xi}}_{jk} \text{ (H-I), or} \\ \widehat{\boldsymbol{\eta}}_{jk} &= \sqrt{n_j/(n_j-1)}\, \widehat{\boldsymbol{P}}^{-1} \widehat{\boldsymbol{\xi}}_{jk} \text{ (H-D)} \end{aligned}$$

over all groups to generate the null distribution of T_{Rd} or T_{Re}. As the statistics depend on the unknown value of $\boldsymbol{P}$ under H_0, we estimate $\boldsymbol{P}$ with $\widehat{\boldsymbol{P}}$ and fix its value throughout the resampling procedure.

2.3 Robust testing

When the data are contaminated with outliers, e.g., the error distribution is heavy-tailed, robust estimation may be necessary for appropriate inference. In particular, outliers are much more influential for the H-I bootstrap test as only the data in each group are available to estimate the group-specific covariance/correlation matrices. Among various robust estimation procedures, we choose the one in Campbell (1980) for its simplicity. We use the test for equal covariance as an example; extension to the test for equal correlation is straightforward. The robust estimator for $\boldsymbol{V}_j$ is given by

$$\ddot{\boldsymbol{V}}_j = \frac{1}{\sum_k \omega_{jk}^2 - 1} \sum_k \omega_{jk}^2 \ddot{\boldsymbol{\epsilon}}_{jk} \ddot{\boldsymbol{\epsilon}}_{jk}^{\tau} \tag{12}$$

where

$$\ddot{\boldsymbol{\epsilon}}_{jk} = \boldsymbol{Y}_{jk} - \widehat{\boldsymbol{f}}(\boldsymbol{X}_{jk}) - \frac{1}{\sum_l \omega_{jl}} \sum_l \omega_{jl}\big(\boldsymbol{Y}_{jl} - \widehat{\boldsymbol{f}}(\boldsymbol{X}_{jl})\big) \approx \boldsymbol{\epsilon}_{jk} - \frac{1}{\sum_l \omega_{jl}} \sum_{l=1}^{n_j} \omega_{jl}\boldsymbol{\epsilon}_{jl} \tag{13}$$

is the robust estimator for $\boldsymbol{\epsilon}_{jk}$. The weights ω_{jk}'s are determined using an iterative procedure described in Campbell (1980) and also in appendix A. It is natural to derive robust estimates for other quantities from $\ddot{\boldsymbol{V}}_j$'s and $\ddot{\boldsymbol{\epsilon}}_{jk}$'s, for example,

- $\ddot{\boldsymbol{V}} = \frac{1}{(\sum_{j,k}\omega_{jk}^2)-J} \sum_j \big((\sum_k \omega_{jk}^2) - 1\big)\ddot{\boldsymbol{V}}_j$;
- $\ddot{\boldsymbol{D}}_j = \mathrm{Diag}(\ddot{\sigma}_{j1}, \ldots, \ddot{\sigma}_{jL})$, where $\ddot{\sigma}_{jl}^2 = \frac{1}{n_j - 1}\sum_k \ddot{\epsilon}_{jkl}^2$, $l = 1, \ldots, L$;
- $\ddot{\boldsymbol{\xi}}_{jk} = \ddot{\boldsymbol{D}}_j^{-1}\ddot{\boldsymbol{\epsilon}}_{jk}$;
- $\ddot{\boldsymbol{R}}_j = \frac{1}{n_j - 1}\sum_k \ddot{\boldsymbol{\xi}}_{jk}\ddot{\boldsymbol{\xi}}_{jk}^{\tau}$;
- $\ddot{\boldsymbol{R}} = \frac{1}{(\sum_{j,k}\omega_{jk}^2)-J} \sum_j \big((\sum_k \omega_{jk}^2) - 1\big)\ddot{\boldsymbol{R}}_j$;

The robust version of test statistics can be defined accordingly. Take for example the determinant-based statistic for testing equal covariance,

$$\ddot{T}_{Vd} = \sum_{j=1}^{J}(n_j - 1)\log\frac{|\ddot{\boldsymbol{V}}|}{|\ddot{\boldsymbol{V}}_j|},$$

Under H_0 in (2), the robust version of (4) is

$$\ddot{T}_{Vd} = \sum_{j=1}^{J}(n_j - 1)\log\frac{|\frac{1}{(\sum_{j,k}\omega_{jk}^2)-J}\sum_{j,k}\omega_{jk}^2(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)^{\tau}|}{|\frac{1}{(\sum_k \omega_{jk}^2)-1}\sum_k \omega_{jk}^2(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)(\boldsymbol{\eta}_{jk} - \bar{\boldsymbol{\eta}}_j)^{\tau}|},$$

where $\bar{\boldsymbol{\eta}}_j = \frac{1}{\sum_k \omega_{jk}}\sum_k \omega_{jk}\boldsymbol{\eta}_{jk}$. The same notation is used for both unweighted and weighted means, but the distinction is clear from the context. Note that the weights ω_{jk}'s depend on the data; therefore, resampling of the $\boldsymbol{\eta}_{jk}$'s and fixing ω_{jk}'s is not appropriate. In Campbell's procedure, the weights depend on data via the Mahalanobis distance $\ddot{\boldsymbol{\epsilon}}_{jk}^{\tau}\ddot{\boldsymbol{V}}_j^{-1}\ddot{\boldsymbol{\epsilon}}_{jk}$; hence, the weights are independent of the underlying true mean and covariance. That implies that the weights remain unaltered for data generated with different first two moments but without changes in other generation mechanisms — a property equivalent

to affine equivariance for the covariance estimator (Wilcox, 2005). As a result, the null distribution of the test statistic can be appropriately obtained by resampling $\boldsymbol{\eta}_{jk}$'s from standardized residuals

$$\ddot{\boldsymbol{\eta}}_{jk} = \left\{ \frac{\left(\sum_l \omega_{jl}\right)^2}{\sum_l \omega_{jl}^2 + \left(\sum_l \omega_{jl}\right)^2 - 2\omega_{jk}\sum_l \omega_{jl}} \right\}^{1/2} \ddot{\boldsymbol{U}}_j^{-1} \ddot{\boldsymbol{\epsilon}}_{jk} \tag{14}$$

and recalculating the weights based on the resampled residuals. We recommend the use of robust techniques for the H-I bootstrap test in all settings, because, as indicated in the simulation study (below), such use helps controlling type I error not only for heavy-tailed distributions, but also for non-heavy-tailed distributions when the sample size is small.

2.4 Multiple comparisons

The bootstrap tests are particularly useful in the setting of multiple comparisons, for which an asymptotic test properly accounting for correlation among test statistics is generally not available. Let us consider simultaneous tests of J hypotheses regarding equal correlation,

$$H_{0j} : \boldsymbol{R}_j = \boldsymbol{R}_0, j = 1, \ldots, J, \tag{15}$$

where R_0 is the correlation matrix for the reference group which is indexed by 0. In genetic analyses, we are often interested in identifying groups, defined by genotype, whose phenotypes differ substantially from those of a referent group. For microbes, the referent group maybe comprised of those with wild-type genotype; for humans, the appropriate referent may be those with the most common genotype in the population(s) of interest. The non-robust test statistics for H_{0j} is

$$\begin{aligned} T_{Rd}^{(j)} &= n_j \log \frac{|\widehat{\boldsymbol{R}}_{j0}|}{|\widehat{\boldsymbol{R}}_j|} + n_0 \log \frac{|\widehat{\boldsymbol{R}}_{j0}|}{|\widehat{\boldsymbol{R}}_0|}, \\ T_{Re}^{(j)} &= \left| \Lambda\big(\widehat{\boldsymbol{P}}_{j0}^{-1}(\widehat{\boldsymbol{R}}_j - \widehat{\boldsymbol{R}}_0)\widehat{\boldsymbol{P}}_{j0}^{-1\ \tau}\big) \right|, \end{aligned}$$

where $\widehat{\boldsymbol{R}}_{j0} = \frac{1}{n_j+n_0}\left\{ n_j \widehat{\boldsymbol{R}}_j + n_0 \widehat{\boldsymbol{R}}_0 \right\}$ and $\widehat{\boldsymbol{P}}_{j0}$ is the lower triangular matrix such that $\widehat{\boldsymbol{R}}_{j0} = \widehat{\boldsymbol{P}}_{j0}\widehat{\boldsymbol{P}}_{j0}^{\tau}$, $j = 1, \ldots, J$. To obtain the joint null distribution, resampling proceeds the same way as in section 2.2 except that $\boldsymbol{P}$ in (11) is fixed by $\widehat{\boldsymbol{P}}_{j0}$.

With the H-I bootstrap, the standardized residuals from different groups share the same distribution asymptotically, under the assumptions (a)–(d). For any given subset of hypotheses, if hypotheses in this subset are true, the joint null distribution of test statistics in the subset obtained by resampling standardized residuals across all groups does not depend on whether hypotheses not in the subset are true – a property called subset pivotality. This property ensures that the free step-down procedure introduced in Westfall and Young (1993) provides asymptotic strong control of type I familywise error rate (FWE) in the resampling setting. With the H-D bootstrap, standardized residuals from different groups share the same asymptotic distribution only when the null hypotheses hold for these groups. Asymptotic strong control is attainable if, at each step of the step-down procedure, resampling is restricted to standardized residuals in the groups for which the null hypotheses have not been rejected yet. This restriction brings two disadvantages: (1) computational burden is increased by performing resampling at each step; and (2) the pool of standardized residuals available for resampling shrinks as more and more hypotheses are rejected, which may affect the performance of the test.

3 Simulation Study

For selected settings of dimension of response, error distribution and sample size, we compare the performances of the four combinations of the two test statistics, determinant-based versus eigenvalue-based, and the two resampling procedures, H-D bootstrap versus robust H-I bootstrap, for testing equal covariance or correlation matrices. The robust estimation procedure in Campbell (1980) involves two control parameters, a and b, for weighting observations in the calculation of moment estimates. Specifically, a sets a threshold for the Mahalanobis distance from an observation to the sample mean beyond which the observed value can be considered as an outlier. The value of b determines how fast the weight decreases as the observed distance increases, with larger values implying less robustness. The parameter setting $a = 2.0$ has an asymptotic interpretation, but the value of b is empirically determined, e.g., $b = 1.25$ was recommended in Campbell (1980). We vary b over 1.25, 2.5 and 5.0 to evaluate the impact of b.

Three error distributions are examined: multivariate normal (MN), multivariate T with 5 degree of freedom (MT) and multivariate Laplace (ML), with the latter two representing heavy-tailed distributions. Normal errors are sampled from $N(\mathbf{0}, \boldsymbol{V}_j)$. Student errors are constructed as $\boldsymbol{Z}_{jk}/W_{jk}$, where $\boldsymbol{Z}_{jk} \sim N(\mathbf{0}, \boldsymbol{V}_j)$ and $W_{jk} \sim \chi^2_5$. Laplace errors are generated by sampling each

component of $\boldsymbol{Z}_{jk} = (Z_{jk1}, \ldots, Z_{jkL})^\tau$ independently from $Laplace(0, 1)$, and setting $\boldsymbol{\epsilon}_{jk} = \boldsymbol{U}_j \boldsymbol{Z}_{jk}/\sqrt{2}$, where $\boldsymbol{U}_j$ is the lower triangle Cholesky decomposition of $\boldsymbol{V}_j$. The MN and MT settings are the same as those used in Zhang and Boos (1992; 1993) and Zhu *et al.*(2002). Let $\boldsymbol{V}_j = \begin{Bmatrix} \sigma_{j1}^2 & \rho_j \sigma_{j1} \sigma_{j2} \\ \rho_j \sigma_{j1} \sigma_{j2} & \sigma_{j2}^2 \end{Bmatrix}$ be the data-generating covariance matrix of group j, $j = 1, \ldots, J$. The exact values of $\boldsymbol{V}_j$'s are given in corresponding figures and tables. Results are based on 5000 simulations, each with $N = 5000$ resamplings. Nominal significance level is set to 0.05.

3.1 Single testing

3.1.1 Testing equal covariance

Figure 1 present the results of testing equal covariance matrices for bivariate responses ($L = 2$). In general, for a given resampling approach, the determinant-based statistic tends to be more conservative (i.e., yielding lower type I errors) than the eigenvalue-based one. The effect is more evident when the error distribution is heavy-tailed. The determinant-based statistic has slightly lower statistical power to detect differences in correlation components but slightly higher power to detect differences in variance components. When the sample size is small and the error distribution is heavy-tailed, the H-I bootstrap needs a higher robustness level, such as b=1.25, to control type I error, whereas the H-D bootstrap always attains reasonable control. This is not surprising, as the H-D bootstrap utilizes twice as much information for the standardization of residuals when the null hypothesis is true. The power of the H-I bootstrap with $b = 1.25$ is comparable to that of the H-D bootstrap. Higher power could be attained for the H-I bootstrap by setting larger values of b, but the price is inflated type I error. With the sample size increased from 40 to 80, a strong robustness level seems unnecessary for the H-I bootstrap, as $b = 5$ gives reasonable type I error and comparable power to the H-D bootstrap.

These observations also hold in general for relatively high-dimensional responses ($L = 5$), as shown in Figure 2. The only difference is that, with a moderately large sample size and heavy-tailed distributions, it is necessary to choose a weak robustness level to avoid too much compromised power. The performance of the determinant-based statistic is sensitive to the data dimension when the sample size is small, with a much lower type I error and somewhat lower statistical power in detecting differences in correlation components than the eigenvalue-based statistics. This phenomenon was noted before in Zhang and Boos (1992; 1993) and Zhu *et al.*(2002). However, even with a

moderately large sample size, the difference in power between the two statistics is small for both the H-D bootstrap and the weakly robust H-I bootstrap. Unlike the low-dimensional case, the increase in sample size did not bring the type I errors much closer to the desired level when the error distribution is t_5. Although the H-D bootstrap controls type I error asymptotically with both statistics (Zhang and Boos, 1992; Zhu *et al.*., 2002), the pattern of type I errors revealed the relative slow convergence of these resampling methods for high-dimensional responses in the presence of a heavy-tailed error distribution. With a sample size of 1000 and a t_5 error distribution, the type I error is 0.024 for the H-D bootstrap with the determinant-based statistic but ranges from 0.045 to 0.051 for other combinations of bootstrap approaches, test statistics and robustness levels.

3.1.2 Testing equal correlation

The results of testing equality of correlation matrices are shown for bivariate responses ($L = 2$) in Figure 3 and for relatively high-dimensional responses ($L = 5$) in Figure 4. Once again, the determinant-based statistic appears to be conservative. With the normal error distribution, the H-D bootstrap is quite conservative. In contrast, the robust H-I bootstrap with $b = 1.25$ controls type I error at the desired level and offers slightly higher power regardless of the type of statistic. Higher values of b (less robust) only mildly inflate the type I error. For the t_5 distribution, the robustness level must be fairly strong, e.g., $b = 1.25$, to control type I error for the H-I bootstrap. The statistical power is comparable between the two resampling approaches. For the Laplace (1) distribution, the H-D bootstrap appears to be too conservative, whereas the H-I bootstrap with a weak robustness level ($b = 5$) controls type I error satisfactorily with substantially improved power. With a larger sample size of 80, the performance is similar between the H-D bootstrap and the H-I bootstrap with $b = 5$. An inflation in type I error is observed for both resampling approaches with the t_5 distribution, and setting $b = 1.25$ lowers the type I error for the H-I bootstrap without much compromising the power. Figure 4 demonstrates that, for relatively high-dimensional responses, nearly all combinations of bootstrap approaches, test statistics and robustness levels are able to control type I error below or near the desired level. A weakly robust H-I bootstrap with the eigenvalue-based statistic attains higher power than other combinations, in particular when the sample size is small.

3.2 Multiple testing

In this simulation study, the four robust resampling-based testing methods to be compared are formed by combining the determinant-based statistic and the eigenvalue-based statistic with $b = 1.25$ and $b = 5.0$. The H-D bootstrap is not considered for multiple testing. Each simulated population has five groups, including a reference group to which other groups are compared; and all five groups share the same mean and variances. The correlation coefficients are set to 0 for three groups including the reference group and 0.5 for the remaining 2, creating 2 true null hypotheses out of 4. Adjusted p-values are evaluated using a step-down procedure described in Westfall and Young (1993). Performance of the methods is assessed by type I and II family-wise errors (FWE) as well as the probability of detecting at least one false null hypothesis to which we refer as power. Type I (II) FWE is defined as the probability of rejecting (accepting) at least one true (false) null hypothesis. One minus type II FWE reflects the statistical power of detecting all false null hypotheses. Tables 1 and 2 summarize the results for testing equal covariance and for testing equal correlation, respectively.

3.2.1 Testing equal covariance

For the normal error distribution, both statistics preserve type I FWE very well, regardless of the robustness level, the sample size or the dimension of response. The eigenvalue-based statistic provides slightly better power than its determinant-based counterpart. In particular, we would not recommend the determinant-based statistic with $b = 1.25$ when the sample size is small ($n_j = 20$, $j = 1, \ldots, 5$), as it has about 10-25% lower power than the other three methods.

For the heavy-tailed distributions, the performance of the methods differs somewhat by sample size. As in the single testing scenario, for a small sample size ($n_j = 20$, $j = 1, \ldots, 5$), a strong robustness level ($b = 1.25$) is needed to control type I FWE for both statistics. The eigenvalue-based statistic gives better power in general, but it still inflates the type I FWE for heavy-tailed error distributions even with $b = 1.25$, especially for high-dimensional responses. When sample size is large ($n_j = 60$, $j = 1, \ldots, 5$), the less robust methods with $b = 5.0$ not only offer better power but also better control of the type I FWE; this pattern is more evident for high-dimensional responses, e.g., for the t_5 distribution at $L = 5$ and $n_j = 60$. This phenomenon suggest that the relationship between the robustness level and the type I FWE is not always monotonic – a fact that is confirmed by further simulations (not given in the

table) showing an increase in the type I FWE with $b = 20$, compared to $b = 5$, for the t_5 distribution at $L = 5$ and $n_j = 60$. The non-robust H-I bootstrap test gives type I FWEs of about 0.1 for the Laplace distribution and 0.2 for the t_5 distribution, confirming the usefulness of robustization. With a large sample size, the eigenvalue-based statistic also attains slightly higher power than the determinant-based statistic, except when the response dimension is high and the robustness level is low.

3.2.2 Testing equal correlation

For all three error distributions, the two statistics have very similar performance for bivariate responses, regardless of sample size, or for 5-dimensional responses when the sample size is large. When the sample size is small and the dimension of response is high, the eigenvalue-based statistic improves the power by 10-20% as compared to the determinant-based statistic, holding the robustness level constant. Different from the testing for covariance, inflation of type I FWE is only seen for the t_5 distribution in the setting of $L = 5$, $n_j = 20$ and $b = 5$, where the eigenvalue-based statistic has much more severe inflation than the determinant-based one.

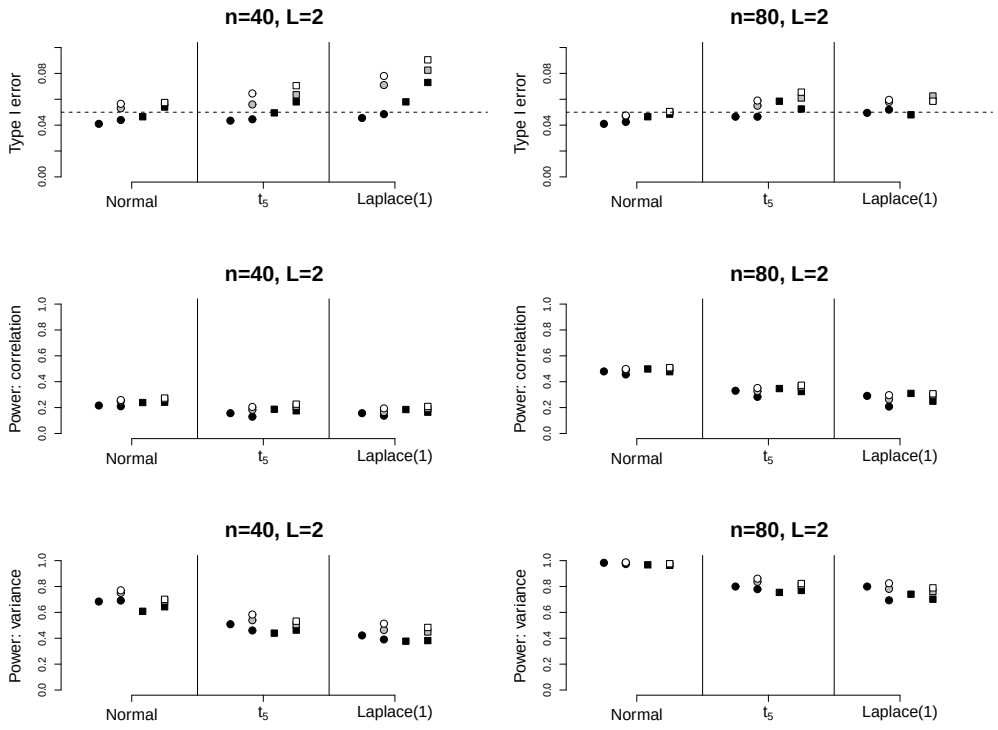

Figure 1: Type I error and power for testing equal covariance matrices of 2-dimensional responses between two groups by error distribution and sample size. For each combination of error distribution and sample size, the four columns from left to right are: (1) H-D bootstrap with the determinant-based statistic (circles), (2) robust H-I bootstrap with the determinant-based statistic (circles) (3) H-D bootstrap with the eigenvalue-based statistic (squares), and (4) robust H-I bootstrap with the eigenvalue-based statistic (squares). For the robust H-I tests, three values are used for the robustness control parameter b, 1.25, 2.5 and 5.0 represented by black, grey and white colors. Type I error is examined in top panels with $\rho_1 = \rho_2 = 0.5$, and $\sigma^2_{jl} = 1$, $j = 1, 2$, $l = 1, 2$. Power in detecting difference in correlation components is examined in middle panels with $\rho_1 = 0.0$, $\rho_2 = 0.5$. and $\sigma^2_{jl} = 1$, $j = 1, 2$, $l = 1, 2$. Power in detecting difference in variance components is examined in bottom panels with $\rho_1 = \rho_2 = 0.0$, $\sigma_{11} = \sigma_{12} = 1$, $\sigma^2_{21} = 2$, and $\sigma^2_{22} = 4$.

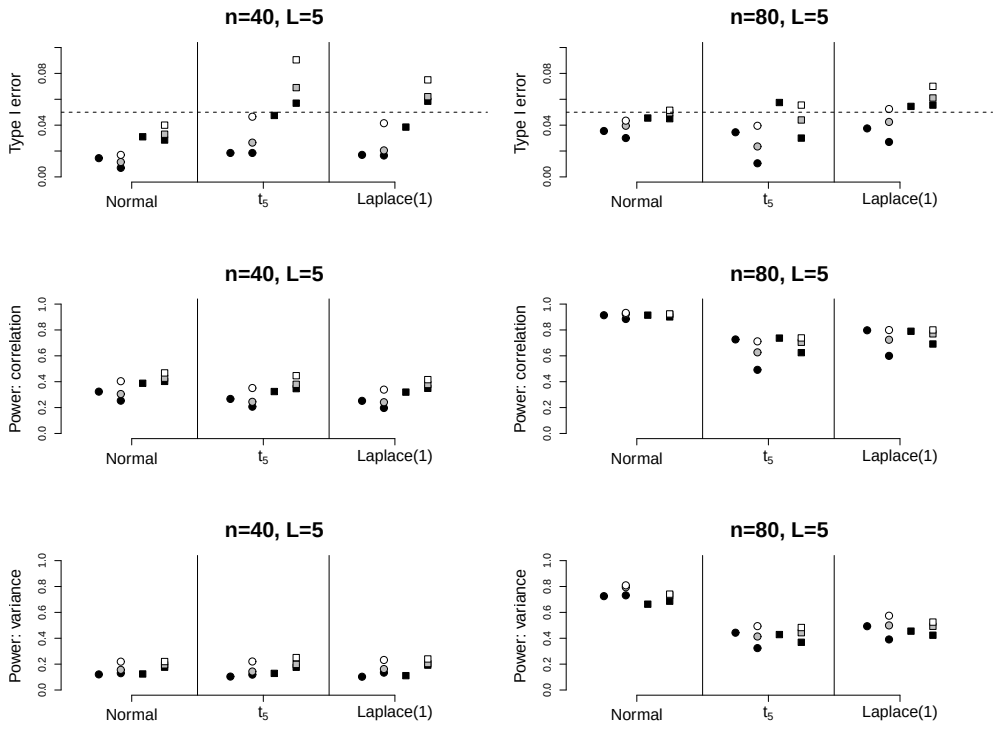

Figure 2: Type I error and power for testing equal covariance matrices of 5-dimensional responses between two groups by error distribution and sample size. Within each combination of error distribution and sample size, the four columns from left to right are: (1) H-D bootstrap with the determinant-based statistic (circles), (2) robust H-I bootstrap with the determinant-based statistic (circles) (3) H-D bootstrap with the eigenvalue-based statistic (squares), and (4) robust H-I bootstrap with the eigenvalue-based statistic (squares). For the robust H-I tests, three values are used for the robustness control parameter b, 1.25, 2.5 and 5.0 represented by black, grey and white colors. Type I error is examined in top panels with $\rho_1 = \rho_2 = 0.5$, and $\sigma_{jl}^2 = 1$, $j = 1, 2$, $l = 1, \ldots, 5$. Power in detecting difference in correlation components is examined in middle panels with $\rho_1 = 0.0$, $\rho_2 = 0.5$. and $\sigma_{jl}^2 = 1$, $j = 1, 2$, $l = 1, \ldots, 5$. Power in detecting difference in variance components is examined in bottom panels with $\rho_1 = \rho_2 = 0.0$, $\sigma_{11} = \ldots = \sigma_5 = 1$, and $\sigma_{21}^2 = \ldots = \sigma_{25}^2 = 2$.

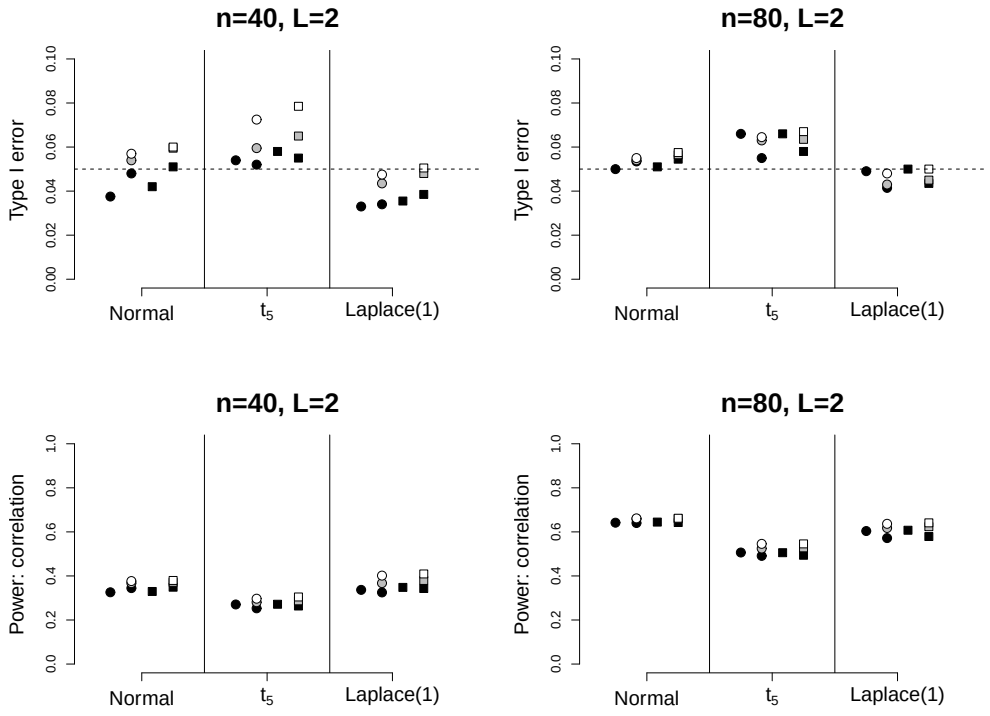

Figure 3: Type I error and power for testing equal correlation matrices of 2-dimensional responses between two groups by error distribution and sample size. For each combination of error distribution and sample size, the four columns from left to right are: (1) H-D bootstrap with the determinant-based statistic (circles), (2) robust H-I bootstrap with the determinant-based statistic (circles) (3) H-D bootstrap with the eigenvalue-based statistic (squares), and (4) robust H-I bootstrap with the eigenvalue-based statistic (squares). For the robust H-I tests, three values are used for the robustness control parameter b, 1.25, 2.5 and 5.0 represented by black, grey and white colors. Type I error is examined in upper panels with $\rho_1 = \rho_2 = 0.5$, and power is examined in lower panels with $\rho_1 = 0.0$ and $\rho_2 = 0.5$. Variances are set to $\sigma^2_{jl} = 1$, $j = 1, 2$, $l = 1, 2$.

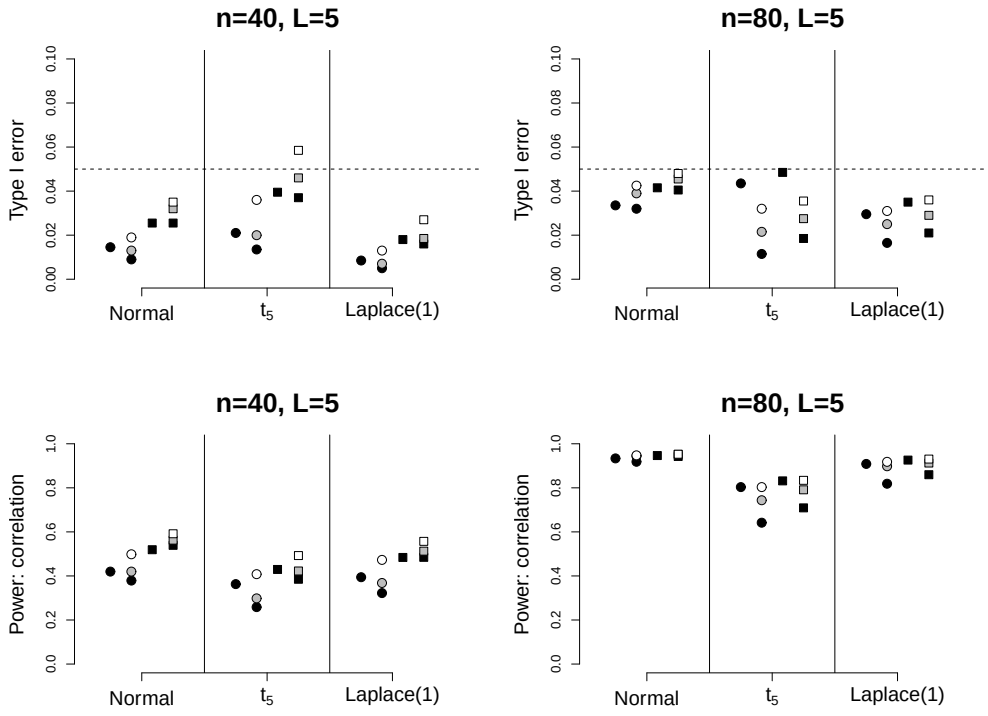

Figure 4: Type I error and power for testing equal correlation matrices of 5-dimensional responses between two groups by error distribution and sample size. Within each combination of error distribution and sample size, the four columns from left to right are: (1) H-D bootstrap with the determinant-based statistic (circles), (2) robust H-I bootstrap with the determinant-based statistic (circles) (3) H-D bootstrap with the eigenvalue-based statistic (squares), and (4) robust H-I bootstrap with the eigenvalue-based statistic (squares). For the robust H-I tests, three values are used for the robustness control parameter b, 1.25, 2.5 and 5.0 represented by black, grey and white colors. Type I error is examined in upper panels with $\rho_1 = \rho_2 = 0.5$, and power is examined in lower panels with $\rho_1 = 0.0$ and $\rho_2 = 0.5$. Variances are set to $\sigma_{jl}^2 = 1$, $j = 1, 2$, $l = 1, \ldots, 5$.

Table 1: *Simulation results for the application of the resampling methods on multiple testing for equal covariance matrices. Comparisons are made between a reference group (indexed by 0) and four other groups (indexed by 1, 2, 3 and 4). The covariance matrices share variances and have an exchangeable correlation structure with pairwise correlation coefficients* $\rho_0 = \rho_1 = \rho_2 = 0.0$ *and* $\rho_3 = \rho_4 = 0.5$. *Results are based on 2000 simulation runs, each with 3000 resamplings.*

				Normal			t_5			Laplace(1)		
			Type of	Type of FWE		Detect	Type of FWE		Detect	Type of FWE		Detect
L	n_j	b	Statistic†	I	II	Prob.	I	II	Prob.	I	II	Prob.
2	20	1.25	D	0.036	0.96	0.17	0.045	0.97	0.13	0.047	0.98	0.099
			E	0.045	0.95	0.21	0.049	0.96	0.17	0.060	0.97	0.13
		5.0	D	0.044	0.95	0.22	0.058	0.95	0.21	0.061	0.96	0.15
			E	0.047	0.94	0.23	0.055	0.95	0.22	0.064	0.97	0.15
	60	1.25	D	0.040	0.63	0.67	0.050	0.81	0.44	0.054	0.90	0.33
			E	0.043	0.63	0.69	0.051	0.79	0.47	0.052	0.89	0.36
		5.0	D	0.041	0.61	0.69	0.048	0.78	0.49	0.048	0.86	0.41
			E	0.041	0.59	0.71	0.051	0.76	0.51	0.043	0.85	0.42
5	20	1.25	D	0.024	0.86	0.41	0.045	0.89	0.32	0.057	0.88	0.37
			E	0.039	0.83	0.47	0.094	0.84	0.45	0.086	0.85	0.45
		5.0	D	0.036	0.80	0.49	0.091	0.80	0.45	0.090	0.80	0.49
			E	0.044	0.82	0.49	0.12	0.81	0.50	0.092	0.83	0.48
	60	1.25	D	0.044	0.022	1.0	0.052	0.25	0.93	0.075	0.19	0.95
			E	0.048	0.024	1.0	0.063	0.25	0.93	0.084	0.20	0.95
		5.0	D	0.041	0.017	1.0	0.045	0.16	0.97	0.071	0.10	0.98
			E	0.050	0.021	1.0	0.055	0.18	0.96	0.072	0.13	0.97

† D:determinant-based, E:eigenvalue-based.

Table 2: *Simulation results for the application of the resampling methods on multiple testing for equal correlation matrices. Comparisons are made between a reference group (indexed by 0) and four other groups (indexed by 1, 2, 3 and 4). The correlation matrices all have an exchangeable strucuture, with the pairwise correlation coefficients* $\rho_0 = \rho_1 = \rho_2 = 0.0$ *and* $\rho_3 = \rho_4 = 0.5$. *Results are based on 2000 simulation runs, each with 3000 resamplings.*

				Normal			t_5			Laplace(1)		
			Type of	Type of FWE		Detect	Type of FWE		Detect	Type of FWE		Detect
L	n_j	b	Statistic†	I	II	Prob.	I	II	Prob.	I	II	Prob.
2	20	1.25	D	0.042	0.89	0.31	0.040	0.92	0.25	0.028	0.91	0.28
			E	0.045	0.88	0.31	0.045	0.91	0.26	0.032	0.90	0.29
		5.0	D	0.043	0.88	0.33	0.050	0.90	0.30	0.032	0.88	0.34
			E	0.045	0.87	0.32	0.054	0.89	0.30	0.034	0.87	0.35
	60	1.25	D	0.046	0.45	0.81	0.048	0.62	0.68	0.040	0.49	0.78
			E	0.047	0.45	0.82	0.048	0.61	0.68	0.041	0.48	0.79
		5.0	D	0.045	0.43	0.83	0.045	0.59	0.71	0.040	0.42	0.82
			E	0.047	0.43	0.83	0.046	0.58	0.71	0.040	0.42	0.83
5	20	1.25	D	0.024	0.79	0.53	0.027	0.87	0.37	0.018	0.84	0.50
			E	0.036	0.73	0.63	0.045	0.84	0.45	0.027	0.78	0.61
		5.0	D	0.032	0.74	0.60	0.061	0.80	0.50	0.030	0.76	0.61
			E	0.041	0.71	0.66	0.076	0.78	0.55	0.034	0.72	0.68
	60	1.25	D	0.048	0.017	1.0	0.047	0.14	0.98	0.040	0.044	1.0
			E	0.045	0.013	1.0	0.049	0.12	0.99	0.044	0.037	1.0
		5.0	D	0.050	0.014	1.0	0.042	0.086	0.99	0.046	0.016	1.0
			E	0.046	0.012	1.0	0.042	0.074	0.99	0.045	0.017	1.0

† D:determinant-based, E:eigenvalue-based.

4 Application

The HIV-1 envelope glycoprotein, gp120, is responsible for viral entry and is a candidate target for vaccine development. A phase III trial (VAX004) of a HIV-1 vaccine (AIDSVAX B/B) was conducted in 5403 subjects at high risk of sexual transmission in North America and the Netherlands from 1998 to 2002 (rgp120 HIV vaccine study group, 2005). From 368 subjects who were infected during the study, post-infection clinical responses, viral RNA load and CD4 cell counts, were measured every four months following the day of diagnosis of infection, with additional measurements at two weeks, one month and two months. The vaccine showed neither protective efficacy against HIV-1 infection (rgp120 HIV vaccine study group, 2005), nor a significant effect on post-infection viral load (Gilbert and Jin, 2010). We investigate whether the correlation among clinical responses is associated with vaccine status and viral genotype among all infected subjects, or with immune responses among infected vaccinees.

Among 368 infected subjects, 239 have the clinical responses measured at both two weeks and four months after diagnosis of infection. We use as bivariate responses the changes of log(viral load) (Y_1) and log(CD4) (Y_2) within this time frame. Departure of the responses from normality is shown in appendix B. Missing clinical responses at four months and later are mainly due to initiation of antiretroviral treatment (ART) – a fact that implies data are not missing at random because ART initiation depends on clinical responses. Gilbert *et al.*(2005) found that pre-ART viral load and CD4 count at one month after diagnosis are independent predictors of ART initiation; and adjusted for the two predictors in the comparison of longitudinal mean clinical responses between vaccine and control groups. Similarly, we adjust for the two predictors in our mean response models to reduce the bias caused by missing data. Scatter plots of the clinical responses over the covariates (not shown) indicate that it is reasonable to assume linear covariate functions. Let x_{jk1} and x_{jk2} be the two predictors for individual k in group j. Let β_{ml} be the coefficient associated with predictor m for response l, $m = 1, 2$ and $l = 1, 2$. Define $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^\tau$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22})^\tau$. The linear coefficients are estimated by fitting the model $\boldsymbol{Y}_{jk} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_{jk} + \boldsymbol{X}_{jk1}\boldsymbol{\beta}_1 + \boldsymbol{X}_{jk2}\boldsymbol{\beta}_{k2}, k = 1, \ldots, n_j, j = 1, \ldots, J$, using iteratively re-weighted least squares, where $\boldsymbol{X}_{jkm} = \mathrm{Diag}\big((x_{jkm}, x_{jkm})\big)$, $m = 1, 2$. In all of the analyses described below, ANOVA was used to test for an effect of the factors under investigation on mean changes in log(CD4) count and in log(RNA), but no effects were detected (not shown).

Overall p-values for single testing and group-specific adjusted p-values

for multiple testing are given in Table **??** for both the determinant-based and eigenvalue-based statistics. We use the robust H-I bootstrap test for this investigation with two levels of robustness, $b = 1.25$ and $b = 5$.

Correlation heterogeneity by vaccine status The sample correlations suggest that vaccine reduced the negative correlation between viral load and CD4 towards 0, but none of the tests show statistically significant effects. The p-values differ in magnitude between the robustness levels.

Correlation heterogeneity by pre-infection immune response Using samples collected prior to infection, two types of vaccine-induced immune responses were measured among infected vaccinees: the binding levels of neutralizing antibodies, and the blocking levels of the CD4's binding with vaccine antigen mixture. While multiple measurements are available on a few subjects, the last measurement before infection is used to characterize the immune response level. The population of infected vaccinees is grouped by quartiles of each immune response. Likely due to the high concordance between the two immune responses (correlation coefficient is 0.81), the results are not very different for these two analyses. Neither the binding levels of neutralizing antibodies nor the CD4 blocking level seem to affect the pattern of correlation among clinical responses.

Correlation heterogeneity by genotype The full genetic sequences of the gp120 surface protein contain nearly 500 codons, necessitating dimension reduction. We first dichotomize each codon into 0 (the most frequent AA) and 1 (otherwise). The dominant AA sequence, the reference in this analysis, and its alignment with the GNE8 strain is shown in Appendix F. To further avoid sparse genotypes, codons whose dominant AA has a frequency less than 30% or greater than 70% are excluded. To attain reasonable statistical power from a sample size of 186 subjects with both gp120 sequence and valid clinical responses, we only consider genotypic groups formed by two codons (site-double) which have a sufficiently large number of observations. To choose a site-double in which there is evidence of heterogeneity, we apply the asymptotic method (Manly and Rayner, 1987) on each site-double to test the single hypothesis in (8). We choose the site-double with the largest p-value from single testing to perform multiple comparisons between genotypes of this site-double. Bootstrap approaches can also be used in this screening procedure; they would, of course, be much more computationally-demanding.

Table 3: Data analysis: overall p-values for single testing and adjusted p-values for multiple testing of equal correlation of changes in log(RNA) and log(CD4) from two weeks to four months after diagnosis across vaccine assignment groups, pre-infection immune response levels among vaccinees, and genotype groups formed by gp120 codons 360 and 426. Reference (0) and mutant (1) amino acids at the two sites are given in the footnote, with the most frequent mutant amino acid marked by $\star$.

				Adjusted P-values			
Grouping		Sample	Sample	$b = 1.25$		$b = 5.0$	
Factor	Group	Size	Correlation	Det.	Eig.	Det.	Eig.
Vaccine Status							
	Placebo	72	-0.45				
	Vaccine	117	-0.34	0.13	0.16	0.27	0.29
Immune Response:							
Antibody Binding							
	0–25%	23	-0.36				
	25–50%	23	-0.25	0.73	0.73	0.72	0.72
	50–75%	37	-0.53	0.71	0.69	0.70	0.69
	75–100%	34	-0.15	0.71	0.69	0.70	0.69
	Overall	117		0.31	0.32	0.29	0.31
Immune Response:							
CD4 Blocking							
	0 –25%	23	-0.39				
	25–50%	29	-0.50	0.83	0.83	0.83	0.82
	50–75%	34	-0.04	0.37	0.38	0.36	0.37
	75–100%	31	-0.50	0.83	0.83	0.83	0.82
	Overall	117		0.14	0.17	0.12	0.15
Genotype:							
Codons 360/426‡							
	0/0	56	-0.50				
	0/1	23	-0.66	0.57	0.58	0.53	0.55
	1/0	73	-0.42	0.66	0.65	0.59	0.58
	1/1	34	0.21	0.003	0.0018	0.0018	0.0006
	Overall	186		0.0012	0.0006	0.001	0.0002

‡ Codon 360: 1={V}, 0={A, D, E, F, H, I*, K, N, Q, R, S};
Codon 426: 1={M}, 0={G, K, L, R*, T, V }

Among all possible site-doubles, the asymptotic test gives the smallest p-value for the pair 360/426, and this site-double is thus chosen to form the viral genotype groups for multiple testing. Codon 360 is near a segment of codons that was found to strongly activate human complement system and thereby enhance disease progression (Susal et al., 1996, codon 360 was numbered 363 in their alignment). Codon 426 is known to affect CD4 binding (Kwong et al., 1998), and it appears quite often in the site-doubles we identified with small p-values. The sample correlation for the group of subjects with the genotype with mutation at the two sites is unusual in that it is positive.

The groupwise difference for this codon pair attains p-values of 0.0012 or less, and the adjusted p-values specific to this mutant group are 0.003 or less, for both test statistics and robustness levels. As we did not adjust for the multiple tests from which we identified this codon pair, the results must be seen as suggestive only; our goal was to illustrate the use of our method.

5 Discussion

We investigate a few resampling-based methods for testing homogeneity in covariance or correlation matrices in both single- and multiple-testing settings. The H-D and nonrobust H-I bootstrap approaches using the determinant-based statistic for testing equal covariance are essentially equivalent to those proposed by Zhang and Boos (1992; 1993), and our simulation results are similar to theirs. Zhu *et al.*used a slightly different H-D bootstrap, which fixes the scaling matrix $\widehat{\boldsymbol{V}}^{1/2}$ at the observed value. We were not able to replicate their simulation results, and therefore no comparison was made between our methods and theirs.

For single testing, the ability of the H-I bootstrap to control type I error is greatly improved by the use of robust sample moments and residuals. The H-I bootstrap with an appropriate robustness level provides in general comparable and sometimes superior performance as compared to the H-D bootstrap in our simulation settings. Robustization of the H-D bootstrap leads to overconservative testing and is therefore not further investigated. More importantly, the robust H-I bootstrap is much easier to implement for multiple testing than the H-D bootstrap, because the residuals are exchangeable across groups under our assumptions, regardless of which hypotheses are true or false. A general guideline for the use of the robust H-I bootstrap is to choose a strong robustness level, or equivalently a small value of b, when the sample size is small and the error distribution is heavy-tailed, and a weak robustness level otherwise. Applying the robust H-I bootstrap test to the data from an HIV vaccine clinical trial, the results suggest that the correlation pattern between short-term changes in HIV viral load and CD4 may vary across viral genotype.

The eigenvalue-based statistic generally provides higher statistical power than the determinant-based statistic in detecting difference in correlation, but such superiority is small, except when the response dimension is high and the sample size is small, and often comes at the price of inflated type I error. We found in additional simulation studies that when the number of groups (J) is 2, the eigenvalue-based statistic, T_{Ve}, is equivalent in performance to a more

intuitive statistic

$$T^{\star}_{Ve} = \frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{l=j+1}^{J} \left| \Lambda\big(\widehat{\boldsymbol{U}}_{jl}^{-1} \widehat{\boldsymbol{V}}_j \widehat{\boldsymbol{U}}_{jl}^{-1\ \tau}\big) - \Lambda\big(\widehat{\boldsymbol{U}}_{jl}^{-1} \widehat{\boldsymbol{V}}_l \widehat{\boldsymbol{U}}_{jl}^{-1\ \tau}\big) \right|,$$

where $\widehat{\boldsymbol{U}}_{jl}^{-1}$ is the inverse of the lower-triangle Cholesky decomposition of the joint sample covariance matrix of groups j and l, that is,

$$\widehat{\boldsymbol{V}}_{jl} = \big((n_j - 1)\widehat{\boldsymbol{V}}_j + (n_l - 1)\widehat{\boldsymbol{V}}_l\big)/(n_j + n_l - 2).$$

This statistic provides a distance measure between eigenvalues of the matrices $\widehat{\boldsymbol{V}}_j$ and $\widehat{\boldsymbol{V}}_l$ after they are standardized by $\widehat{\boldsymbol{U}}_{jl}^{-1}$. More interestingly, empirical evaluation shows that the two standardized matrices share the same eigenvectors and differ only in eigenvalues. In fact, it seems that the inverse of the cholesky decomposition of any $\widehat{\boldsymbol{V}}_{jl} = w\widehat{\boldsymbol{V}}_j + (1-w)\widehat{\boldsymbol{V}}_l$ with $0 < w < 1$ can rotate the samples of the two groups into a space expanded by the same eigenvectors; therefore the new sample covariance matrices only differ in eigenvalues. Such rotation can also be attained using spectral decomposition. When $J > 2$, T_{Ve} and $T^{\star}_{Ve}$ are not equivalent, but their performance is similar as revealed by additional simulations (not shown).

For the H-I bootstrap test, our choice of the robustness level control parameter b is based on studies in Campbell (1980) and our simulation study; this choice may not be appropriate for error distributions not discussed here. In additional simulations with a relatively large number of groups and small or moderate sample size, even a small value of b may not be adequate to control type I error, regardless of test statistic. For example, to test the hypothesis (2) with $(L = 2, J = 8, n_j = 40, b = 1.25)$ using the determinant-based statistic, the type I errors are 0.07 for the normal distribution and 0.11 for the t_5 distribution. The reason may be that the probability of misclassification of regular observations and of outliers among resampled standardized residuals increases with the number of groups, as these residuals are imperfect estimates of the errors, especially with smaller sample sizes. To correct this inflation, one can use a lower robustness level for the calculation of the observed statistics and a higher robustness level for the resampled statistic. In this example, if we use $b = 5$ and $b = 1.25$ for the observed and resampled statistics respectively, the type I errors drop to 0.053 for the normal distribution and 0.061 for the t_5 distribution, and the statistical power is higher than the H-D bootstrap. How to find the optimal differential robustness levels is worth future investigation.

The problem of robust estimation for location and scale is more or less equivalent to that of detecting outliers; therefore, principle component analysis and other search algorithms over appropriately chosen directions in the

sample space may be also used to improve the estimation, especially for high-dimensional responses (Campbell, 1980; Pena and Prieto, 2001; Maronna and Zamar, 2002). These enhanced covariance estimates can be used to standardize residuals. However, the use of robust test statistic in combination with these methods requires scrutiny to assure that affine equivariance holds; otherwise, the bootstrap test may not be valid.

We have assumed that $L \ll \min_j n_j$, or in the asymptotic sense, $L = o(n)$, for the purposes of (1) $\widehat{\boldsymbol{\eta}}_{jk} \xrightarrow{P} \boldsymbol{\eta}_{jk}$ as $n \to \infty$ and $n_j/n \to \rho_j$ for $0 < \rho_j < 1$, $j = 1, \ldots, J$, and (2) the sample covariance or correlation matrices are invertable and thus the test statistics are continuous functions of the $\boldsymbol{\eta}_{jk}$'s under the null. These conditions are needed, together with the moment assumptions (a)–(d), for the test statistics as functions of the $\boldsymbol{\eta}_{jk}$'s and those as functions of bootstrap samples of $\widehat{\boldsymbol{\eta}}_{jk}$'s to converge to the same distribution. When $L \ll \min_j n_j$ does not hold, the sample covariance may be ill-conditioned and not invertable. A possible remedy is to use shrinkage estimators, for example, the optimal linear shrinkage estimator proposed by Ledoit and Wolf (2004). In our case, the optimal linear coefficients for weighting the sample covariance and the identity matrix may be estimated using $\widehat{V}$ for the H-D bootstrap or $\widehat{V}_j$ for the H-I bootstrap, and be treated as fixed during the resampling. This linear shrinkage estimator is a consistent estimator for covariance even if $L = O(n)$. Further study is needed to evaluate the performance of this practice.

How best to handle missing measurements also warrants consideration, because complete-case analysis may be biased unless data are missing completely at random. A possible solution is to weight each completely observed individual by the inverse probability of observation assuming data are missing at random. This assumption implies that, conditioning on observed quantities, the missing mechanism does not depend on the missing values (Tsiatis, 2006). In the calculation of summary statistics, such as $\widehat{\boldsymbol{R}}_j$ and $\widehat{\boldsymbol{V}}_j$, observed individuals would be weighted using inverse probabilities given by an appropriate regression model. Further research is required to establish whether the weights should also be used in resampling standardized residuals. Multiple imputation (Little and Rubin, 2002) may also be used, probably at the price of substantial increase in computational time.

6 Supplementary Materials

6.1 Appendix A: Robust estimation for covariance

We follow the M-estimation procedure in Campbell (1980) for robust estimaiton of the covariance matrix. The procedure is conducted for each group independently.

1. Set $a = 2$, $b = 1.25$ and $\omega_{jk} = 1$, $k = 1, \ldots, n_j$.
2. Let $\ddot{\boldsymbol{\epsilon}}_{jk} = \boldsymbol{Y}_{jk} - \frac{1}{\sum_l \omega_{jl}} \sum_l \omega_{jl} \big(\boldsymbol{Y}_{jl}\big)$.
3. Let $\ddot{\boldsymbol{V}}_j = \frac{1}{\sum_k \omega_{jk}^2 - 1} \sum_k \omega_{jk}^2 \ddot{\boldsymbol{\epsilon}}_{jk} \ddot{\boldsymbol{\epsilon}}_{jk}^{\tau}$.
4. Let $d_0 = \sqrt{L} + a/\sqrt{2}$ and $d_k = \{\ddot{\boldsymbol{\epsilon}}_{jk}^{\tau} \ddot{\boldsymbol{V}}_j^{-1} \ddot{\boldsymbol{\epsilon}}_{jk}\}^{1/2}$. The weights are recalculated with
$$\omega_{jk} = \begin{cases} 1, & \text{if } d_k \le d_0, \\ \frac{d_k}{d_0} \exp\Big\{-\frac{1}{2}\big(\frac{d_k - d_0}{b}\big)^2\Big\}, & \text{otherwise,} \end{cases}$$
5. Repeat steps 2-4 till convergence in $\ddot{\boldsymbol{V}}_j$.

6.2 Appendix B: Departure of clinical responses from normality in the data analysis

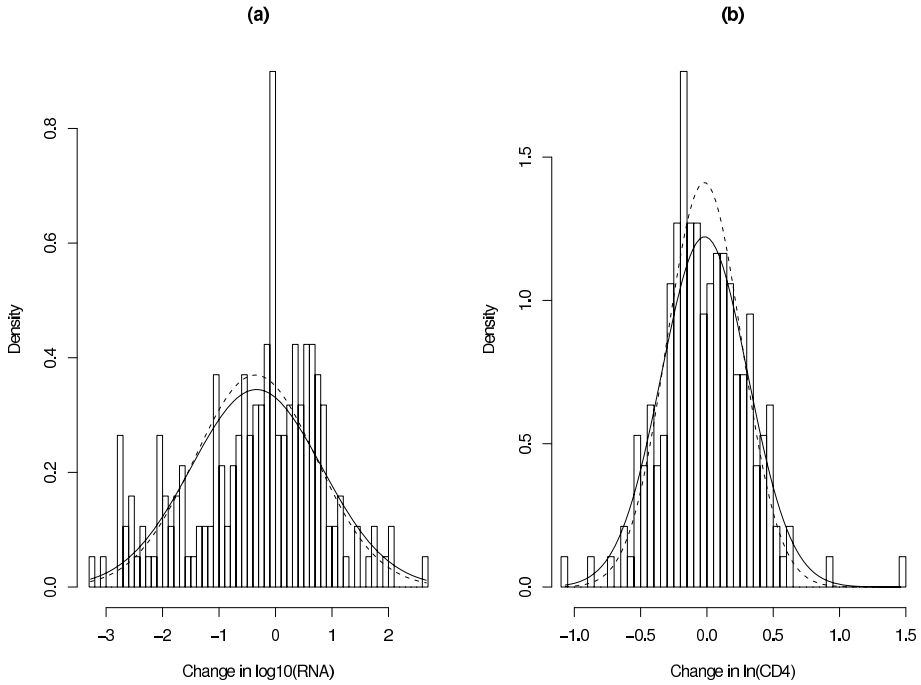

Figure 5: *Histograms of the changes in (a)* $\log_{10}$*(Viral RNA) and (b)* $\ln$*(CD4) from week 2 to month four. The solid and dashed curves show the normal densities with the mean and variance estimated from the data with and without 10% winsorization respectively.*

References

[1] Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, **38**, 337–344.

[2] Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317–346.

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

[4] Campbell, N. A. (1980). Robust procedures in multivariate analysis I: robust covariance estimation. *Applied Statistics*, **29(3)**, 231–237.

[5] Gilbert, P. B. and Jin, Y. (2010). Semiparametric estimation of the average causal effect of treatment on an outcome measured after a

postrandomization event, with missing outcome data. *Biostatistics*, **11**, 34–47.

[6] Gilbert, P. B., Peterson, M.L., Follmann, D., Hudgens, M.G., Francis, D.P., Gurwith, M., Heyward, W.L., Jobes, D.V., Popovic, V., Self, S.G., Sinangil, F., Burke, D. and Berman, P.W. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases*, **191**, 666–677.

[7] Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. and Hendrickson W. A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393**, 648–659.

[8] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

[9] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

[10] Manly, B. F. J. and Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, **74(4)**, 841–847.

[11] Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44(4)**, 307–317.

[12] O'Brien, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics*, **48**, 819–827.

[13] Pena, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* **43(3)**, 286–300.

[14] Susal, C., Kirschfink, M. K., Daniel, V. and Opelz, G. (1996). Identification of complement activation sites in human immunodeficiency virus type-1 glycoprotein gp120. *Blood* **87**, 2329–2336.

[15] Tiku, M. L. and Balakrishnan, N. (1985). Testing the equality of variance-covariance matrices the robust way. *Communications in Statistics - Theory and Methods* **14**, 3033–3051.

[16] Tsiatis, A. A. (2006). *Semiparametric theory And missing data*. New York: Springer-Verlag.

[17] rgp120 HIV Vaccine Study Group Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*, **191**, 654–665.

[18] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.

[19] Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.

[20] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, **14(4)**, 1261–1295.

[21] Yu, X, McLane, M. F., Ratner, L., O'Brien, W., Collman, R., Essex, M. and Lee, T-H. (1994) Killing of primary CD4+ T cells by non-syncytium-inducing macrophage-tropic human immunodeficiency virus type 1. *PNAS*, **91**, 10237–10241.

[22] Zhang, J. and Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices. *Journal of the American Statistical Association*, **87(418)**, 425–429.

[23] Zhu, L. X., Ng, K. W. and Jing, P. (2002). Resampling methods for homogeneity tests of covariance matrices. *Statistica Sinica*, **12(2002)**, 769–783.