HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

J Am Stat Assoc. 2015 June 1; 110(510): 837–849. doi:10.1080/01621459.2014.934826.

A Two-Sample Test for Equality of Means in High Dimension

Karl Bruce Gregory,
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 kbgregory@stat.tamu.edu

Raymond J. Carroll,
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 carroll@stat.tamu.edu

Veerabhadran Baladandayuthapani, and
Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77230-1402, USA veera@mdanderson.org

Soumendra N. Lahiri
Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695-8203 snlahiri@ncsu.edu

Abstract

We develop a test statistic for testing the equality of two population mean vectors in the "large-p-small-n" setting. Such a test must surmount the rank-deficiency of the sample covariance matrix, which breaks down the classic Hotelling T^2 test. The proposed procedure, called the generalized component test, avoids full estimation of the covariance matrix by assuming that the p components admit a logical ordering such that the dependence between components is related to their displacement. The test is shown to be competitive with other recently developed methods under ARMA and long-range dependence structures and to achieve superior power for heavy-tailed data. The test does not assume equality of covariance matrices between the two populations, is robust to heteroscedasticity in the component variances, and requires very little computation time, which allows its use in settings with very large p. An analysis of mitochondrial calcium concentration in mouse cardiac muscles over time and of copy number variations in a glioblastoma multiforme data set from The Cancer Genome Atlas are carried out to illustrate the test.

Keywords

large p; copy number variation; heteroscedasticity; two-sample

1 Introduction

In many applications it is desirable to test whether the means of high-dimensional random vectors are the same in two populations. Often, the number of components in the random vectors exceeds the number of sampled observations, the so-called "large-p-small-n" problem, and conventional test statistics become unviable. Given the steadily growing availability and interest in high-dimensional data, particularly in biological applications, test statistics that are viable for high-dimensional data are in increasing demand. The challenge

when $p > > n$ is to model the structure of dependence among the p components without estimating each of the $p(p + 1)/2$ unique entries in the full covariance matrix. The classical test for equal mean vectors between two populations is Hotelling's T^2 test, but the test statistic is undefined when p is larger than the sum of the sample sizes (minus 2), because it involves inverting the $p \times p$ sample covariance matrix. Several procedures are available which circumvent full covariance matrix estimation. We acheive this in the important case in which the p components admit an ordering in time, space, or in another index, such that the dependence between two components is related to their displacement. When measurements are taken along a chromosome, for example, the location of each measurement is recorded, providing an index over which dependence may be modeled, affording gains in power. For concreteness, it is here assumed that the components admit a unidirectional ordering.

To fix notation, let $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$ and $Y_1, Y_2, \ldots, Y_m \in \mathbb{R}^p$ be independent identically distributed random samples from two populations having $p \times 1$ mean vectors μ_1 and μ_2 and $p \times p$ covariance matrices Σ_1 and Σ_2, respectively. The hypotheses of interest become $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$.

There are some methods available for testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ in the "large-p-small-n" setting. Srivastava (2007) presented a modification of Hotelling's T^2 statistic which handles the singularity of the sample covariance matrix by replacing its inverse with the Moore-Penrose inverse. Wu et al. (2006) proposed the pooled component test, for which the test statistic is the sum of the squared univariate pooled two-sample t-statistics for all p vector components, which they assumed to follow a scaled chi-square distribution. Bai & Saranadasa (1996) presented a test statistic which uses only the trace of the sample covariance matrix and performs well when the random vectors of each population can be expressed as linear transformations of zero-mean i.i.d. random vectors with identity covariance matrices. Each of these methods assumes a common covariance matrix between the two populations, that is that $\Sigma_1 = \Sigma_2$.

More recently, under a setup similar to that of Bai & Saranadasa (1996), but which accommodates unequal covariances, Chen & Qin (2010) introduced a method (hereafter called the Ch-Q test), which allows $\Sigma_1 \neq \Sigma_2$ and sidesteps covariance matrix estimation altogether. Srivastava & Kubokawa (2013) proposed a method (hereafter called the SK test) for multivariate analysis of variance in the large-p-small-n setting, of which the high-dimensional two-sample problem is an instance. Cai et al. (2014) presented a test (hereafter called the CLX test) based upon the supremum of standardized differences between the observed mean vectors, and offer an illuminating discussion about the conditions under which supremum-based tests are likely to outperform sum-of-squares-based tests, which include the Ch-Q and SK tests as well as the test we introduce in this paper. If the differences between μ_1 and μ_2 are rare, but large where they occur, i.e. the signals are sparse but strong, a supremum-based test should have greater power than a sum-of-squares-based test. The reason is that tests which sum the differences across a large number of indices will not be greatly influenced by a very small number of large differences. If, however, there are many differences between μ_1 and μ_2, but these differences are small, i.e. the signals are dense but weak, the supremum of the differences across all the indices will not likely be

extreme enough to arouse suspicion of the null. A sum-of-squares based test statistic, however, will represent an accumulation of the large number of weak signals, and will have more power. Dense-but-weak signal settings do exist, for example in the analysis of copy number variations, where mildly elevated or reduced numbers of DNA segment copies in cancer patients are believed to occur over regions of the chromosome rather than at isolated points (Olshen et al. (2004), Baladandayuthapani et al. (2010)). It is for such cases that our test is designed.

Section 2 describes the GCT test statistic and Section 3 gives its asymptotic distribution. Section 4 presents a simulation study of the GCT, comparing its performance with that of the Ch-Q, SK, and CLX tests in terms of power and maintenance of nominal size. Section 5 implements the GCT as well as the Ch-Q, SK, and CLX tests on a copy number data set and a time series data set. Concluding remarks appear in Section 7 and the Appendix provides proofs of the main results. Full details for the proofs may be found in the **Supplementary Material**.

2 Test Statistic

The GCT statistic is computed as follows. Let $T_n = p^{-1}\left(t_{n1}^2 + t_{n2}^2 + \cdots + t_{np}^2\right)$, where

$$t_{nj}^2 = \left(\overline{X}_{nj} - \overline{Y}_{mj}\right)^2 \left(s_{nj}^2/n + \vartheta_{mj}^2/m\right)^{-1} \quad (1)$$

for $j = 1, \ldots, p$, where $\overline{X}_{nj}$ and $\overline{Y}_{mj}$ are the sample means of the j^{th} vector component and s_{nj}^2 and ϑ_{mj}^2 are the sample variances of the j^{th} vector component for the X and Y samples, respectively. Thus T_n is the mean of the squared unpooled univariate two-sample t-statistics t_{nj}^2 over all components $j = 1, \ldots, p$.

The GCT statistic is a centered and scaled version of T_n defined as $G_n \equiv p^{1/2}\left(T_n - \hat{\xi}_n\right)/\hat{\zeta}_n$, where $\hat{\xi}_n$ and $p^{1/2}/\hat{\zeta}_n$ are described below. The equal means hypothesis is rejected at level α when $|G_n| > \Phi^{-1}(1 - \alpha/2)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

In what shall be called the *moderate-p* version of the test, $\hat{\xi}_n \equiv 1$, so that $G_n^{(M)} \equiv p^{1/2}\left(T_n - 1\right)/\hat{\zeta}_n$. For the *large-p* version, higher-order expansions suggest a centering of the form $\hat{\xi}_n \equiv 1 + n^{-1}\hat{a}_n + n^{-2}\hat{b}_n$, so that

$$G_n^{(L)} \equiv p^{1/2}\left\{T_n - \left(1 + n^{-1}\hat{a}_n + n^{-2}\hat{b}_n\right)\right\}/\hat{\zeta}_n. \quad (2)$$

The quantities $\hat{a}_n$ and $\hat{b}_n$ are defined as $\hat{a}_n \equiv \left(\hat{c}_{n1} + \cdots + \hat{c}_{np}\right)/p$ and $\hat{b}_n \equiv \left(\hat{d}_{n1} + \cdots + \hat{d}_{np}\right)/p$, where $\hat{c}_{nj}$ and $\hat{d}_{nj}$ are obtained by plugging sample moments into the expressions given in Lemma 1 for c_{nj} and d_{nj} for each of the components $j = 1, \ldots, p$.

Though T_n is a mean of squared marginal two-sample t-statistics, the construction of the scaling will account for the dependence among them. In both the moderate- and large-p versions of the test statistic, the scaling $p^{1/2}/\hat{\zeta}_n$ is the same. Let

$$\hat{\gamma}(k) = (p-k)^{-1} \sum_{j=1}^{p-k} \left(t_{nj}^2 - T_n\right)\left(t_{n(j+k)}^2 - T_n\right), \quad (3)$$

which is the sample autocovariance function of the squared t-statistics. Then the scaling $\hat{\zeta}_n$ is defined such that

$$\hat{\zeta}_n^2 \equiv \sum_{|k|<L} w(k/L)\,\hat{\gamma}(k), \quad (4)$$

where $w(x)$ is an even, piecewise function of x such that $w(0) = 1$, $|w(x)| \leq 1$ for all x, and $w(x) = 0$ for $|x| > 1$, and L is a user-selected lag window size.

The choices of the lag window $w(\cdot)$ considered here are the Parzen window

$$w_p(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & |x| < 1/2 \\ 2(1-|x|)^3. & 1/2 \leq x \leq 1 \\ 0, & |x| > 1 \end{cases}$$

found in Brockwell & Davis (2009) and the trapezoid window

$$w_T(k/r) = \begin{cases} 1, & |k| < [L/2] \\ 1 - \left(\frac{k-[L/2]}{r-[L/2]}\right), & [L/2] \leq k \leq L \\ 0, & |k| > L \end{cases}$$

from Politis & Romano (1995), where $[x]$ denotes the largest integer not exceeding x.

3 Main Results

Let $\alpha(r) = sup\left\{\alpha\left(\mathscr{F}_1^k, \mathscr{F}_{k+r}^p\right) : 1 \leq k \leq p-r\right\}$, where $\mathscr{F}_a^b \equiv \mathscr{F}_{a,n}^b = \sigma\langle\{t_{nj} : a \leq j \leq b\}\rangle$ and where for any σ-fields, $\mathscr{F}$ and $\mathscr{G}$,

$$\alpha(\mathscr{F}, \mathscr{G}) = sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathscr{F}, B \in \mathscr{G}\}$$

denotes the strong mixing coefficient between $\mathscr{F}$ and $\mathscr{G}$. Then the following conditions are assumed in deriving the asymptotic distribution of the test statistic T_n.

(C.1) For some $\delta \in (0, \infty)$, (i) $\sum_{r=1}^{\infty} \alpha(r)^{\delta/(2+\delta)} < \infty$, and (ii) $E|t_{nj}^2|^{2r+\delta} < c < \infty$ for all $j = 1, \ldots, p$ for some integer $r \geq 1$.

(C.2) The limit $lim_{n\to\infty} \frac{1}{p-k} \sum_{j=1}^{p-k} Cov\left(t_{nj}^2, t_{n(j+k)}^2\right) = \gamma(k)$ exists for all $k > 0$.

(C.3)

- **i.** $\max\{E|X_{1j}|^{16}, E|Y_{1j}|^{16}, j = 1, \ldots, p\} = O(1)$.
- **ii.** $\min\{\mathrm{Var}(X_{1j}), \mathrm{Var}(Y_{1j})\} > c > 0$.

The following theorem establishes the asymptotic normality of the test statistic under the appropriate centering and scaling.

Theorem 1 *Suppose that* $p \equiv p_n = o(n^6)$ *and (C.1)–(C.3) hold with* $r = 1$ *in (C.1). Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(T_n - 1<x\right) - \Phi\left\{\sqrt{p}\left(x - n^{-1}a_n - n^{-2}b_n\right)/\tau_\infty\right\}\right| = o(1),$$

where $\tau_\infty^2 = \gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k)$ *and* $a_n = (c_{n1} + \cdots + c_{np})/p$ *and* $b_n = (d_{n1} + \cdots + d_{np})/p$, *where* c_{nj} *and* d_{nj} *for* $j = 1, \ldots, p$ *are as in Lemma 1 in the Appendix.*

Remark 1 *Theorem 1 shows that* $G_n \equiv p^{1/2}\left(T_n - \hat{\xi}_n\right)/\hat{\zeta}_n \to^d$ *Normal*(0, 1) *as* $n \to \infty$.

3.1 Technical Details

The choice of the centering quantity $\hat{\xi}_n$ comes from noting that $ET_n = 1 + O(n^{-1})$ as $n \to \infty$. This follows from the fact that t_{nj} converges in distribution to Z, where $Z \sim$ Normal(0, 1), for all $j = 1, \ldots, p$, and $EZ^2 = 1$. Thus $E\{\sqrt{p}(T_n - 1)\} = \sqrt{p}O\left(n^{-1}\right)$, so that when $\hat{\xi}_n \equiv 1$, the expectation of the test statistic differs from zero by $\sqrt{p}O\left(n^{-1}\right)$, restricting p to grow at a rate such that $p = o(n^2)$. When $\hat{\xi}_n \equiv 1 + n^{-1}\hat{a}_n + n^{-2}\hat{b}_n$, the expectation of the test statistic is $\sqrt{p}O\left(n^{-3}\right)$, allowing $p = o(n^6)$. Hence the "moderate-" and "large-p" designations. One may also consider an intermediate-p version of the test which uses only $n^{-1}\hat{a}_n$ in the centering correction, allowing $p = o(n^4)$, but its performance is not investigated here.

While the large-p test allows for $p = o(n^6)$, an advantage of the moderate-p test is its robustness to outliers. The centering correction in the large-p test involves high-order sample moments which are volatile when the data come from a very heavy-tailed distribution, in which case the centering value of 1 is preferable.

The formulation of $\hat{\zeta}_n$ rests on the assumption that the p components admit a logical ordering such that their dependence is autocovarying and diminishing as components are further removed—that is, that the covariance between components may be described with an autocovariance function that decays sufficiently fast. In the proof of Theorem 1, the asymptotic variance of $p^{1/2}T_n$ under some regularity conditions is shown to be $\sum_{h=-\infty}^{\infty}\gamma(h)$, which is equal to 2π times the spectral density $f(\cdot)$ of the sequence $(t_{n1}^2, t_{n2}^2, \ldots)$ evaluated at 0. Thus $\hat{f}(0) = (2\pi)^{-1}\sum_{|k|<L} w(k/L)\hat{\gamma}(k)$ provides the scaling in (4).

3.2 Power of the Generalized Component Test

In order to compute the asymptotic power of the GCT, the expected value of $T_n = p^{-1}\left(t_{n1}^2, \ldots, t_{np}^2\right)$ must be computed under the alternative $H_1 : \mu_{1j} - \mu_{2j} = \delta_j$ for $j = 1, \ldots, p$ where $\delta_j \neq 0$ for at least one j. Let $\xi_n^{(1)}$ denote $E(T_n|H_1 \text{ true})$. Then the power of the GCT, which is $P\left(\left|p^{1/2}\left(T_n - \hat{\xi}_n\right)/\hat{\zeta}_n\right| > z_{\alpha/2} \,|H_1 \quad \text{true}\right)$, is equal to

$$1 - P\left(-z_{\alpha/2} - p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\hat{\zeta}_n < p^{1/2}\left(T_n - \xi_n^{(1)}/\hat{\zeta}_n\right) < z_{\alpha/2} - p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\hat{\zeta}_n | H_1 \quad \text{true}\right).$$

Under conditions (C.1)–(C.3) we can invoke the asymptotic normality of $p^{1/2}\left(T_n - \xi_n^{(1)}\right)/\hat{\zeta}_n$ and the consistency of $\hat{\zeta}_n$ for ζ and approximate the power with

$$1 - \left\{\Phi\left(z_{\alpha/2} - p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\zeta\right) - \Phi\left(-z_{\alpha/2} - p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\zeta\right)\right\}$$

so that it is a function of $p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\zeta$.

Given the tedium of computing $\xi_n^{(1)} = E\left\{p^{-1}\sum_{j=1}^p t_{nj}^2 | H_1 \quad \text{true}\right\} - np^{-1}\sum_{j=1}^p E\left[\left(\overline{X}_{nj} - \overline{Y}_{mj}\right)^2 / \left\{s_{nj}^2 + (n/m)\,\vartheta_{mj}^2\right\} | H_1 \quad \text{true}\right]$ to within $O(n^{-3})$ of its true value as was done for $\hat{\xi}_n$ under the null hypothesis (cf. Lemma 1), we replace s_{nj}^2 and ϑ_{mj}^2 with their population values σ_{1j}^2 and σ_{2j}^2 and $\xi_n^{(1)} \approx 1 + n(\mu_{1j} - \mu_{2j})^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}$.

If we may replace n, $p^{1/2}\left(\xi_n^{(1)} - \hat{\xi}_n\right)/\zeta$ with $np^{-1/2}\sum_{j=1}^p \delta_j^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}/\zeta$, then the power may be expressed

$$1 - \left(\Phi\left[z_{\alpha/2} - np^{-1/2}\sum_{j=1}^p \delta_j^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}/\zeta\right] - \Phi\left[-z_{\alpha/2} - np^{-1/2}\sum_{j=1}^p \delta_j^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}/\zeta\right]\right).$$

From this expression we note that under $p = o(n^2)$

$$\text{Power} \to \begin{cases} 1, & p^{1/2}n^{-1} = o\left(\sum_{j=1}^p \delta_j^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}\right) \\ \alpha, & \sum_{j=1}^p \delta_j^2 / \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}\right) = o\left(p^{1/2}n^{-1}\right) \end{cases}$$

For example, if $\delta_j = \delta p^{-1/2}$ for $j = 1, \ldots, p$ for some $\delta > 0$ then the power will converge to 1, but if $\delta_j = \delta p^{-(1/2+\varepsilon)}$ for $j = 1, \ldots, p$ the test will have "nonpower" above the significance level as $n, p \to \infty$.

4 Simulation Studies

The performances of the GCT, Ch-Q, SK, and CLX tests were compared in terms of size control and power under various settings. For the sample sizes $(n, m) = \{(45, 60), (90, 120)\}$ with $p = 300$, two-sample data were generated such that for each subject the p components were (i) independent (IND), (ii) ARMA dependent, or (iii) long-range (LR) dependent. For each dependence structure, the innovations used to generate each subject series were (a) Normal(0,1), (b) skewed innovations, coming from a gamma(4, 2) distribution centered at zero, thus having mean zero and variance $4(2)^2 = 16$, and (c) heavy-tailed innovations from a Pareto(a, b) distibution with distribution function $F(x) = 1 - (1 + x/b)^{-1/a}$ where the density was shifted to the origin and reflected across the vertical axis to form a "double" Pareto distribution. Under this double Pareto distribution,

$$E|X|^r = \begin{cases} \infty, & r \geq a \\ b^r \Gamma(a-r)\,\Gamma(1+r)\,/\Gamma(a), & r < a. \end{cases}$$

Once a zero-mean series was generated for each subject, it was added to the $p \times 1$ mean vector μ_1 or μ_2, depending on the population to which the subject belonged. Under IND, the zero-mean series consisted of p independent identically distributed innovations from the chosen innovation distribution. For the ARMA dependence structure, p-length series from an ARMA process with AR parameters $\varphi_1 = \{0.4, -0.1\}$ and MA parameters $\theta_1 = \{0.2, 0.3\}$ were used for both populations. Under the LR structure, realizations of zero-mean, long-range-dependent processes with self-similarity parameter $H_1 = (1/2)(2 - 0.75) = 0.625$ were used. The algorithm used for generating vectors of long-range dependent random variables is found in Hall et al. (1998).

At each sample size, dependence structure, and innovation distribution combination, a simulation was run in which $\Sigma_1 = \Sigma_2$ and in which $\Sigma_2 = 2\Sigma_1$, where the unequal covariance setting was imposed by scaling the zero-mean series for the population 2 subjects by $\sqrt{2}$.

For the CLX test, which features an equal-covariances and an unequal-covariances version, Cai et al. (2014) suggest first testing $H_0 : \Sigma_1 = \Sigma_2$ using a test from Cai et al. (2013) and then choosing the version of the CLX test accordingly. Since in practice it is generally not known whether $\Sigma_1 = \Sigma_2$ holds, the test of $H_0 : \Sigma_1 = \Sigma_2$ was performed in each simulation run to determine which version of the CLX test would be used. The CLX test requires an estimate for the precision matrix $\Omega = \Sigma_1^{-1} \left(= \Sigma_2^{-1}\right)$ or $\Omega = \{\Sigma_1 + (n/m)\Sigma_2\}^{-1}$ for the unequal-covariances version. Of the two methods the authors suggest for estimating , that which is presented in Cai et al. (2011) and provided in the R package fastclime (Pang et al. (2013)) was chosen and implemented under default settings.

For power simulations, the alternate hypotheses were that $\mu_1 = 0$ and $\mu_2 = \left[\delta 1'_{\beta p}, (0)\, 1'_{(1-\beta)p}\right]'$, where 1_k was a $k \times 1$ vector of ones, p was the number of components, and $\beta \in [0, 1]$ was the proportion of the p components for which the difference in means was nonzero. The number of components p was fixed at 300 and the power was

simulated for $\beta \in \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 1\}$. The difference or signal δ was chosen such that the signal to noise ratio δ/σ was equal to 1/8, where σ was the standard deviation of the innovations used to construct each series (each p-variate observation); thus $\delta = \sigma/8$ was used.

Full factorial simulation results for $\{(45, 60), (90, 120)\} \times \{\text{IND, ARMA, LR}\} \times \{\text{Normal, Skewed, Heavy-tailed}\} \times \{\Sigma_1 = \Sigma_2, \Sigma_2 = 2\Sigma_1\}$ are given in the **Supplementary Material** and selected results are highlighted here. In addition to the factorial simulation, the tests were evaluated under heteroscedastic component variances and ultra-heavy tailed (infinite-variance) innovations.

4.1 Performance under normality

Table 1 displays the simulated Type I error rates of the four tests under the sample sizes (n, m) = (45, 60), (90, 120) across the three dependence structures under Normal(0, 1) innovations and for $\Sigma_1 = \Sigma_2$. For the GCT, results are given for the Parzen and trapezoid lag windows at lag window sizes L = 10, 15, 20 for the moderate-p (upper panel) and the large-p (lower panel) choice of the centering. The Ch-Q, SK, and CLX Type I error rates are duplicated in the upper and lower panels as the moderate- and large-p versions of the GCT were applied to the same 500 simulated data sets.

The Ch-Q and SK tests maintained very close-to-nominal Type I error rates. The CLX test exhibited slightly inflated Type I error rates under the IND and LR dependence structures for the smaller sample sizes (n, m) = (45, 60), but maintained close-to-nominal rates for (n, m) = (90, 120). For the GCT, the Parzen window appeared to control the Type I error rate slightly better than the trapezoid window, and the Type I error rates were similar for the three choices of the lag window size.

Power simulation results under normal innovations appear in Figure S.6 of the Supplementary Material.

4.2 Effect of skewness

The results of the Type I error simulation with skewed innovations were similar to those in the Normal(0, 1) case and can be seen in Table S.3 of the **Supplementary Material**. For the power simulation, Figure 1 plots the proportion of rejections across 500 simulation runs against the proportion β of the p = 300 components in which μ_1 and μ_2 differed, where $\beta \in \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 1\}$. The three panels show the power curves of the four tests under the IND, ARMA, and LR dependence structures, respectively, when the innovations came from the centered gamma(4, 2) distribution and when the sample sizes were (n, m) = (90, 120). The four tests exhibited similar performance under these settings, though under independence the size of the CLX test was somewhat inflated, yet its power increased more rapidly in β than that of the other tests under ARMA dependence.

4.3 Effect of heavy-tailedness

The results for the heavy-tailed simulation with innovations coming from the double Pareto(16.5, 8) distribution did not differ greatly from those of the normal- and skewed-

innovations simulations. Full results may be found in the **Supplementary Material**. In order to assess the robustness of the GCT to violations of its moment conditions, ultra-heavy tailed data were simulated using innovations from a double Pareto(1.5, 1) distribution, which has infinite variance. Since the centering corrections $\hat{a}_n$ and $\hat{b}_n$ in the large-p GCT are computed using higher order sample moments, only the moderate-p GCT was here considered, as its centering of 1 gives it stability. Under these settings, the signal, which was set to $\delta = .5$, is very weak relative to the noise, such that as the proportion β of non-null mean differences goes to 1, a dense-but-weak signal structure is simulated. The resulting power curves are shown in Figure 2, in which the Ch-Q test is seen to have much less power than the others; the CLX also suffers a reduction in power under ARMA and LR dependence. Under LR dependence, the size of the GCT was somwhat inflated, but it was very close to nominal for the IND and ARMA cases. In the ARMA case, the GCT exhibited greater power than the other tests across the range of alternatives.

4.4 Effect of heteroscedasticity

The effect of heteroscedasticity on the GCT may be anticipated by noting that t_{nj}^2 from (1) can be expressed

$$t_{nj}^2 = \left[\frac{\sqrt{n}\left\{ \left(\overline{X}_{nj} - \mu_{1j} \right) - \left(\overline{Y}_{mj} - \mu_{2j} \right) \right\}}{\sqrt{s_{nj}^2 + (n/m)\,\vartheta_{mj}^2}} + \frac{\sqrt{n}\delta_j}{\sqrt{s_{nj}^2 + (n/m)\vartheta_{mj}^2}} \right]^2 \quad (5)$$

where $\delta_j = \mu_{1j} - \mu_{2j}$, for $j = 1, \ldots, p$. The second term is equal to zero under H_0. Under H_1, for a fixed difference δ_j, the variances σ_{1j}^2 and σ_{2j}^2 affect the magnitude of t_{nj}^2 such that very small values for σ_{1j}^2 and σ_{2j}^2 promote very large values of t_{nj}^2. Since th scaling $\hat{\zeta}_n$ for T_n is a function of $\hat{\gamma}_j(\cdot)$, the estimated autocovariance function of $t_{n1}^2, t_{n2}^2, \ldots, t_{np}^2$, as seen from (3) and (4), extreme values of t_{nj}^2 will pull $\hat{\zeta}_n$ upward, shrinking T_n toward zero. Extreme values of t_{nj}^2 will tend to occur if σ_{1j}^2 and σ_{2j}^2 are very small when $\delta_j \neq 0$. Although smaller variances ought to ensure a greater likelihood of rejecting H_0, if $\hat{\zeta}_n$ is inflated by extreme values of t_{nj}^2, the GCT statistic will be close to zero, and the test will fail to reject, hence condition (C.3) (ii). Large values of σ_{1j}^2 and σ_{2j}^2 when $\delta_j = \neq 0$ will tend to reduce t_{nj}^2, but since it is bounded below by zero, extreme values will not occur. The size of the test should be robust to any scaling of the variances, as the second term in (5) will be zero when H_0 is true.

To investigate the impact of heterscedasticity on the performance of the four tests, the standard deviations of the components were each scaled by a realization from the exponential distribution with mean 1/2 shifted to the right by 1/2 such that the average scaling was 1 and so that the scaled variances were bounded away from 0. The power simulation with centered gamma(4, 2) innovations was repeated under these heteroscecastic conditions with $(n, m) = (45, 60)$. Figure 3 exhibits a dramatic reduction in the power of the Ch-Q test due to heteroscedasticity. The CLX test exhibited somewhat inflated size under

the IND and LR dependence structures, while the SK test and the GCT demonstrated robustness to the heteroscedstic variance scalings.

4.5 Effect of unequal covariance matrices

Of the four tests, the SK test is the only one which assumes a common covariance matrix for the two populations. Cai et al. (2014) suggest first testing $H_0 : \Sigma_1 = \Sigma_2$ with a test from Cai et al. (2013) and implementing the equal or unequal covariances version of the CLX test accordingly. The Ch-Q and the GCT do not require any assumption or testing of equality between the covariance matrices. The SK is thus anticipated to perform more poorly than the others when the covariance matrices are unequal.

To impose inequality between Σ_1 and Σ_2, the zero-mean sequences for each subject from population two were scaled by $\sqrt{2}$ before the signal μ_2 was added. This imposed the condition that $\Sigma_2 = 2\Sigma_1$.

Figure 5 displays results for a simulation in which the variances of the second population were scaled by two and in which the variances in both populations were heteroscedastic. The SK lost much of its power under these settings, which was expected given its assumption of a common covariance matrix in the two populations. The Ch-Q test exhibited low power as before owing to the heteroscedasticity, but performed none the worse for the unequally scaled variances. The GCT achieved the greatest power under the LR dependence structure, having less power than the CLX test in the ARMA case.

Lastly, under the ultra heavy-tailed innovation distribution with unequally scaled covariances between the two populations, the GCT exhibited superior power to the Ch-Q, SK, and CLX tests under all three dependence stuctures at the $(n, m) = (90, 120)$ sample sizes. Although the size of the GCT was somewhat inflated under the LR dependence structure, it maintained the nominal Type I error rate in the ARMA case, under which it achieved roughly 60% power when $\beta = 0.4$ while the CLX test achieved only about 10% power.

5 Copy Number Variation Example

The GCT, Ch-Q, SK, and CLX tests were each applied to a data set from The Cancer Genome Atlas containing copy number measurements at chromosomal copy number locations in 92 long-term-surviving patients, who survived for more than two years after their initial diagnosis and 138 short-term-surviving patients, who survived for fewer than 2 years after their initial diagnosis of a brain cancer called glioblastoma multiforme (GBM). Pinkel & Albertson (2005) suggest that the numbers of copies of certain DNA segments within a cell may be associated with cancer development and spread. It is thus of interest to identify regions along the genome in which high numbers of copies are associated with the incidence or severity of cancer, as such regions may harbor cancer-causing or tumor-suppressor genes. In studies having relatively few patients, several thousand copy number measurements are taken along each arm of each chromosome, which makes identifying regions for which two patient groups have different mean copy number profiles a high-dimensional problem. Additionally, it is believed that copy number variations between

patient groups will occur over stretches of the chromosome (spanning multiple probes) rather than at isolated points (singleton probe locations) (Olshen et al. (2004), Baladandayuthapani et al. (2010)), suggesting a serial dependence over the chromosome as well as the presence of a dense-but-weak rather than a sparse-but-strong signal structure.

We restricted our analysis to the q arm of chromosome 1, the longest chromosome, on which there are 8,895 copy number measurements. Each measurement is a log-ratio of the number of copies at each location over 2, where 2 is the number of copies found in normal DNA. Positive measurements thus indicate duplications and negative measurements indicate deletions. The measurements, in conformity with the assumption of the GCT that the components of interest admit a logical ordering, are recorded along with their locations given in the number of base pairs from the end of the DNA strand. For many of the 8,895 components, there are a few missing values in either or both of the samples such that the average proportion of missing values per component is 0.0276 for the long-term survivors and 0.0273 for the short-term survivors. Prior to analysis, each missing value was replaced with the mean of the non-missing values for the same component in the same sample.

Although a test may reject $H_0 : \mu_1 = \mu_2$ when μ_j is the 8895 × 1 vector of copy number means for $j = 1, 2$, a wholesale conclusion for the entire arm of the chromosome is of little use if it is desired to identify particular regions in which copy number differences lie. In order to break the chromosome arm into meaningful regions in which the equal means hypothesis is of interest, we performed a method of segmentation called circular binary segmentation (CBS) from Olshen et al. (2004). This procedure locates change points in the copy number sequence for a single sequence of copy number values, and is implemented in the R package DNAcopy (Seshan & Olshen (2013)). In order to segment the q arm of chromosome 1 for equal means hypothesis testing when multiple patients are observed, the CBS procedure was applied to the 8895 × 1 vector of differences in means $X - \bar{Y}$ using weights equal to $s_j^2/n + \vartheta_j^2/m$ for $j = 1, \ldots, 8895$. Before computing X, $\bar{Y}$, and s_j^2 and ϑ_j^2 for $j = 1, \ldots, p$, each series was smoothed using the function smooth.CNA() from the DNAcopy package. The CBS procedure provided 26 segments of varying lengths at the edges of which change points were detected in the vector of differences in means. As a set of 7 contiguous segments contained small numbers of markers (44, 14, 26, 39, 26, 21, 27) they were collapsed into a single segment having 197 markers, which left 20 regions within which the number of probes p ranged from 73 to 1811. Such splitting of the chromosome into windows or segments has been widely used in genome-wide association studies in searching for chromosomal regions in which genetic variants are associated with a continuous or dichotomous clinical outcome, as in Wu et al. (2011).

The large- and moderate-p GCT with lag window size $L = (2/3)p^{1/2}$ and the Ch-Q, SK, and CLX tests were applied to each of the twenty segments identified by the CBS procedure to test $H_{0k} : \mu_{1k} = \mu_{2k}$ for $k = 1, \ldots, 20$ (Though smoothing was used in identifying the segments, the analysis was carried out on the raw, unsmoothed data). Since the equal-means hypothesis was tested for twenty different regions simultaneously, the sets of p-values which each of the four tests generated were compared with the Benjamini & Hochberg (1995) discovery rate (FDR) threshold. For m tests of hypotheses, the m p-values are ordered $p_{(1)} \leq$

$p_{(2)} \leq \ldots \leq p_{(m)}$ and then the hypothesis to which $p_{(i)}$ corresponds is rejected if $i \leq k$, where $k = \max\{j : p_{(j)} \leq (j/m)q\}$. This procedure was originally shown to control the FDR at q for m independent hypothesis tests, though Benjamini & Yekutieli (2001) showed that for many common types of positive dependence among the m test statistics, the same procedure still adequately controls the FDR. The procedure was therefore applied to the twenty p-values computed from each test.

Figure 6 summarizes the analysis. The left panel displays the univariate two-sample t-statistics, which are the t_{nj} values for $j = 1, \ldots, 8895$, against their locations in base pairs along the q arm of chromosome 1. The vertical line at zero marks the value around which the t-statistics would be centered under the null hypotheses, and the horizontal dotted lines delineate the CBS-selected segments of the chromosome arm. The numbers of copy number markers p within each segment appear on the right. Rejections acheived by the tests are marked with symbols appearing on the left, where rejections for each test are determined by the Benjamini & Hochberg (1995) FDR procedure.

The upper right panel of Figure 6 displays the estimated autocorrelation function of the squared two-sample univariate t-statistics, the t_{nj}^2 values for $j = 1, \ldots, 8895$, along the q arm of chromosome 1. The 95% confidence bounds using the large-lag standard error described in Anderson (1977) are shown, which suggest that dependence decays in conformity with (C.1) (i).

The lower right panel of Figure 6 shows the results of the FDR procedure. The upward sloping line is given by $y = (x/m)q$, which is the Benjamini & Hochberg (1995) FDR rejection threshold. The p-values for all four tests are shown, but are ordered according to the ranking of the large-p GCT p-values (The rejection decisions were the same for the moderate- and large-p versions of the GCT). The SK and CLX tests did not achieve any rejections; the Ch-Q test achieved one rejection, and the GCT rejected equal means for fifteen of the twenty regions.

Figure 7 offers an explanation of the additional power demonstrated by the GCT. The upper and lower panels show the estimated standard deviation at each of the 8,895 copy number locations across the q arm of chromosome 1 for the 92 long-term and 138 short-term survivors, respectively. Both panels exhibit spikes at shared locations as well as prominent humps around 2.0×10^8 Mbps, suggesting that the variances are not constant across the chromosome; nor are the humps at equal heights for the two groups of patients. The boxplots of the 8,895 standard deviations for each group reveal significant right skewness, suggesting heavy-tailedness of some of the component distributions. The minimum estimated standard deviations for the long- and short-term survivors were 0.1314 and 0.1123, respectively, indicating that the component variances are bounded sufficiently away from zero. The severe heteroscedasticity as well as the inequality of variances between the two samples appear to have attenuated the power of the Ch-Q and SK tests just as in the simulation.

None of the univariate two-sample t-statistics in the lefthand panel of Figure 6 are very extreme, the largest of their magnitudes being 3.607. This suggests that the difference

between the copy number profiles of short- and long-term survivors consists of smaller differences distributed over a larger number of components rather than larger differences over a smaller number of components. That is, the signals appear to be dense but weak rather than sparse but strong. In such a setting the CLX test will likely have low power.

It is worth discussing the computation time of the four tests. For this analysis, in which each test was implemented twenty times at various values of the dimension p, the moderate-p GCT finished in 1.75 seconds and the large-p GCT finished in 6.60 seconds. The Ch-Q and SK tests finished in 2.32 and 2.68 minutes, respectively, and the CLX took 2.79 hours to run on a MacBook Air with a 1.86 GHz Intel Core 2 Duo processor with 4 GB of memory. The SK procedure involves matrix operations which can be quite slow for large p, and the Ch-Q test involves a cross-validation type sum of inner products which becomes slow for large sample sizes. The CLX method must first test whether $\Sigma_1 = \Sigma_2$ and then directly estimate Σ^{-1} or $\{\Sigma_1 + (n/m)\Sigma_2\}^{-1}$ under sparsity assumptions. Estimating these large matrices quickly becomes computationally burdensome. The GCT requires only a summation over p components and computation of the sample autocovariance function of a p-length series, making it very fast to compute.

6 Mitochondrial Calcium Concentration

Ruiz-Meana et al. (2003) subjected cells from cardiac muscles in mice to conditions which simulated reduced blood flow for a period of one hour. To a treatment group, a dose of cariporide was administered, which is believed to inhibit cell death due to oxidative stress. The investigators measured the mitochondrial concentration of Ca^{2+} every ten seconds during the hour. The experiment was run twice, once on intact cells and once on cells with permeabilized membranes. The data have been made available by Febrero-Bande & Oviedo de la Fuente (2012) in the R package fda.usc.

The mean percent increase of the calcium concentration over its initial value for the treatment and control in both the experiments is plotted against time in Figure 8, where the sample sizes for each curve are shown. The first 180 seconds of the data are removed, given the erratic behavior of the curves, leaving $p = 342$ time points. The four tests were applied to both the intact and permeabilized data to test for equality between the true treatment and control mean curves. The p-values for the four tests are given in Table 2.

For the intact cells, the Ch-Q test and the GCT strongly rejected the null, while the CLX test, after failing to reject equality of the covariance matrices, produced a p-value of 0.086 under the equal covariances assumption, and the SK test failed to reject. For the permeabilized experiment the Ch-Q test and the GCT again strongly rejected the null. The CLX test again failed to reject equality of the covariance matrices, which is a dubious assumption for either the intact or permeabilized experiments given the plot in Figure 9 of s_j^2/ϑ_j^2 for $j = 1, \ldots, 342$ for each set of data. In this plot the variance of the treatment group measurements for the intact cells is well over twice as high as in the control group for the first ten minutes (fluctuating wildly), and for the permeabilized cells the variance of the treatment group measurements remains at roughly twice that of the control group

measurements after half an hour has elapsed. The low power of the SK test apparently owes to the variance inequality depicted here.

The inability of the CLX test to reject what appears to be an implausible null hypothesis likely owes to a difference in mean functions which is characterized by gradual separation rather than by spikes in one function or the other. The large number of small differences are unable to produce a maximum which will exceed the CLX rejection threshold. However, the Ch-Q test and the GCT are able to register the large number of small differences cumulatively and reject the equal means hypothesis.

This example illustrates the applicability of our test in functional data contexts, in which each observation consists of a function observed at points over some domain. When it is of interest to compare the mean functions in two populations, the assumptions of the GCT are likely to apply.

7 Conclusions

The test we present for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, called the *generalized component test*, was shown to be competitive in the $p > > n$ setting when the p components admit an ordering allowing the dependence between two components to be modeled according to their displacement. Moderate- and large-p versions of the test were given for $p = o(n^2)$ and $p = o(n^6)$, respectively. The test requires very little computation time and is easily scalable to very-large p settings.

The moderate-p version of our test is robust to ultra heavy-tailedness, and both the moderate- and large-p versions are robust to heteroscedasticity and highly unequal covariance matrices. The Chen and Qin (Ch-Q) test lost most of its power in the presence of heavy-tailedness or heteroscedasticity; the Srivastava and Kubokawa (SK) test lost much of its power when the covariance matrices were unequally scaled. The Cai, Liu, and Xia (CLX) test performed well under a variety of settings, proving to be robust to heteroscedasticity and to unequally scaled covariance matrices; however, when the data were very heavy-tailed, which rendered the signals very weak, the CLX lost considerable power. Also, since the CLX test requires estimating the $p \times p$ precision matrix, it is computationally much slower than the other tests, requiring over 2.5 hours to complete the copy number data analysis which the SK and Ch-Q tests completed in under 3 minutes and the GCT in under 10 seconds.

For the copy number analysis, the GCT exhibited superior power over the other three tests. This was likely due to heteroscedasticity in the component variances, under which the Ch-Q would lose power, unequally scaled variances between the two populations, under which the SK test would lose power, and likely to the presence of a dense-but-weak rather than a sparse-but-strong signal structure, under which the CLX test would have low power.

For the mitochondrial calcium concentration data set, only the Ch-Q test and the GCT were able to reject the equal means hypothesis. The SK test appears to have lost power due to unequal variances and the CLX supremum-based test was unable to detect the smooth

separation of the two mean functions over time, which was characterized by small differences in many components rather than by large differences in a few.

Software

We created the R package highD2pop for implementing the GCT as well as the Ch-Q, SK, and CLX tests. A source version, highD2pop.zip, is available for download. The package includes copy number data for the CBS-selected segment of the q arm of chromosome 1 having $p = 400$ copy number probes. See package documentation in highD2pop-manual.pdf.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Veerabhadran Baladandayuthapani's work is partially supported by NIH grant R01 CA160736 and the Cancer Center Support Grant (CCSG)(P30 CA016672). The content is solely the responsibility of the authors and does not necessarily represent the o cial views of the National Cancer Institute or the National Institutes of Health.

Appendix: Proofs of main results

Proof 1 (Asymptotic Normality of Test Statistic) By an adaptation of the big-block-little-block argument to the triangular array it can be shown that $p^{-1/2}\sum_{j=1}^{p}\left[t_{nj}^2 - Et_{nj}^2\right] \to Normal\left(0, \tau_\infty^2\right)$, *where*

$$\begin{aligned} \tau_2^\infty = & \lim_{n\to\infty} Var\left(p^{-1/2}\sum_{j=1}^{p} t_{nj}^2\right) = \lim_{n\to\infty} p^{-1}\sum_{k=0}^{p-1}\sum_{|j_1-j_2|=k} Cov\left(t_{nj_1}^2, t_{nj_2}^2\right) \\ = & \gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k), \end{aligned} \tag{A.1}$$

where $\gamma(k) = lim_{n\to\infty}(p-k)^{-1}\sum_{j=1}^{p-k} Cov\left(t_{nj}^2, t_{n(j+k)}^2\right)$, $k \geq 0$. To prove (A.1), use the moment and α-mixing conditions to show that for any M ≥1,

$$\begin{aligned} p^{-1}\sum_{k=M+1}^{p-1}\sum_{|j_1-j_2|=k}\left|Cov\left(t_{nj_1}, t_{nj_2}\right)\right| \leq & \; 2\sum_{k>M} p^{-1}(p-k)\left\{\alpha(k)^{\delta/(2+\delta)}\bigvee_{j=1}^{p}\left(E|t_{nj}|^{2+\delta}\right)^{\frac{2}{2+\delta}}\right\} \\ \leq & \; C\sum_{k=M+1}^{\infty}\alpha(k)^{\delta/(2+\delta)} \to 0 \end{aligned}$$

as $M \to \infty$. *Thus,*

$$\begin{aligned} & \sup_{x\in\mathbb{R}}\left|P\left(\sqrt{p}\left[T_n - p^{-1}\sum_{j=1}^{p}E\left(t_{nj}^2\right)\right] \leq x\right) - \Phi\left(x/\tau_\infty\right)\right| = o(1) \\ \Rightarrow \; & \sup_{x\in\mathbb{R}}\left|P\left(T_n - p^{-1}\sum_{j=1}^{p}E\left(t_{nj}^2\right) \geq x\right) - \Phi\left(\sqrt{p}x/\tau_\infty\right)\right| = o(1) \\ \Rightarrow \; & \sup_{x\in\mathbb{R}}\left|P\left(T_n - 1 \leq x\right) - \Phi\left(\sqrt{p}\left[x - n^{-1}a_n - n^{-2}b_n\right]/\tau_\infty\right)\right| = o(1), \end{aligned}$$

where a_n and b_n are bounded sequences such that

$$p^{-1}\sum_{j=1}^{p} E\left(t_{nj}^2\right) = 1 + n^{-1}a_n + n^{-2}b_n + O\left(n^{-3}\right). \quad \text{(A.2)}$$

Lemma 1 provides c_{nj} and d_{nj} for $j = 1, \ldots, p$ such that $a_n = (c_{n1} + \cdots + c_{np})/p$ and $b_n = (d_{n1} + \cdots + d_{np})/p$ satisfy (A.2).

Lemma 1 *Let $X_{1j}, \ldots, X_{nj}$ and $Y_{1j}, \ldots, Y_{mj}$ be independent identically distributed random samples with $Var(X_{1j}) = \sigma_{1j}^2$ and $Var(Y_{1j}) = \sigma_{2j}^2$ and $EX_{1j} = EY_{1j}$ for all $j = 1, \ldots, p$. Assume that $\max\{E|X_{1j}|^{16}, E|Y_{1j}|^{16}, j = 1, \ldots, p\} = O(1)$ and that $min\left\{\sigma_{1j}^2, \sigma_{2j}^2\right\} > c > 0$* (The first moment condition may be reduced further by means of truncation, but this would considerably lengthen the proof. The discussion of heteroscedasitiy in Section 4.4 illustrates the importance of bounding the component variances away from zero). Let $t_{nj}^2 = n\left(\overline{X}_{nj} - \overline{Y}_{mj}\right)^2 \left\{s_{nj}^2 + (n/m)\,\vartheta_{mj}^2\right\}^{-1}$, *where s_n^2 and ϑ_m^2 are the two samples variances and let $m \sim n$ as $n \to \infty$. Then $E\left(t_{nj}^2\right) = 1 + n^{-1}c_{nj} + n^{-2}d_{nj} + O\left(n^{-3}\right)$ for*

$$c_{nj} = \tau_{nj}^{-2}\left\{\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2\right\} + 2\tau_{nj}^{-5}\left\{\mu_{3j}^{'} + (n/m)^2\eta_{3j}^{'}\right\}^2 \quad \text{(A.3)}$$

and

$$\begin{aligned}
d_{nj} = \;& \tau_{nj}^{-4}\left[\left\{\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2\right\} - \left\{\left(\mu_{4j}^{'} - 3\sigma_{1j}^4\right) + (n/m)^4\left(\eta_{4j}^{'} - 3\sigma_{2j}^4\right)\right\}\right] \\
+ \;& \tau_{nj}^{-6}\left\{\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2\right\}\left\{\left(\mu_{4j}^{'} - \sigma_{1j}^4\right) + (n/m)^3\left(\eta_{4j}^{'} - \sigma_{2j}^4\right)\right\} \\
- \;& 4\tau_{nj}^{-6}\left\{\mu_{3j}^{'} + (n/m)^2\eta_{3j}^{'}\right\}\left\{\mu_{3j}^{'} + (n/m)^3\eta_{3j}^{'}\right\} \\
- \;& \tau_{nj}^{-6}\left\{\left(\mu_{3j}^{'}\right)^2 + (n/m)^5\left(\eta_{3j}^{'}\right)^2\right\} \\
- \;& 6\tau_{nj}^{-8}\left\{\mu_{3j}^{'} + (n/m)^2\eta_{3j}^{'}\right\}\left\{\mu_{5j}^{'} - 2\mu_{3j}^{'}\sigma_{1j}^2 + (n/m)^4\left(\eta_{5j}^{'} - 2\eta_{3j}^{'}\sigma_{2j}^2\right)\right\} \quad \text{(A.4)} \\
- \;& 3\tau_{nj}^{-8}\left\{\left(\mu_{4j}^{'} - \sigma_{1j}^4\right) + (n/m)^3\left(\eta_{4j}^{'} - \sigma_{2j}^4\right)\right\}^2 \\
+ \;& 6\tau_{nj}^{-8}\left\{\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2\right\}\left\{\mu_{3j}^{'} + (n/m)^2\eta_{3j}^{'}\right\}^2 \\
+ \;& 3\tau_{nj}^{-10}\left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}\left\{\left(\mu_{4j}^{'} - \sigma_{1j}^4\right) + (n/m)^3\left(\eta_{4j}^{'} - \sigma_{2j}^4\right)\right\}^2 \\
+ \;& 12\tau_{nj}^{-10}\left\{\mu_{3j}^{'} + (n/m)^2\eta_{3j}^{'}\right\}^2\left\{\left(\mu_{4j}^{'} - \sigma_{1j}^4\right) + (n/m)^3\left(\eta_{4j}^{'} - \sigma_{2j}^4\right)\right\},
\end{aligned}$$

where $\tau_{nj}^2 = \left\{\sigma_{1j}^2 + (n/m)\,\sigma_{2j}^2\right\}$ and $\mu_{kj}^{'}$ and $\eta_{kj}^{'}$ are the kth central moments of X_{1j} and Y_{1j}, respectively.

Proof 2 (Proof of Lemma 1) For ease of syntax, ignore the subscript j, and, without loss of generality, assume that $EX_{1j} = EY_{1j} = 0$. *Let $\Delta_n = s_n^2 - \sigma_1^2 + (n/m)\left(\vartheta_m^2 - \sigma_2^2\right)$ and let t_n^2 be approximated by the expansion*

$$\tilde{t}_n^2 = n\left(\overline{X}_n - \overline{Y}_m\right)^2\left(\tau_n^{-2} - \tau_n^{-4}\Delta_n + \tau_n^{-6}\Delta_n^2 - \tau_n^{-8}\Delta_n^3 + \tau_n^{-10}\Delta_n^4\right). \quad \text{(A.5)}$$

so that $t_n^2 = \tilde{t}_n^2 + O_p\left(n^{-3}\right)$. *An expression for the expected value* $E\left(\tilde{t}_{nj}^2\right)$ *would thus involve the quantities* $n\tau_n^{-2k} E\left(\overline{X}_n - \overline{Y}_m\right)^2 \Delta_n^{k-1}$ *for* $k = 1, \ldots, 5$. These expectations must be computed such that they retain terms out to the order of $O(n^{-3})$.

Let $\chi_{|B|}(\{X_j : j \in B\})$ represent the joint cumulant of the random variables in the set $\{X_j : j \in B\}$, *where* $|B|$ *is the cardinality of B. Then the formula*

$$E(X_1 \ldots X_k) = \Sigma_\pi \Pi_{B \in \pi} \chi_{|B|}(\{X_j : j \in B\}) \quad \text{(A.6)}$$

from Leonov & Shiryaev (1959) gives the expected value of a product of random variables in terms of joint cumulants, where Σ_π *denotes summation over all possible partitions of* $\{X_1, \ldots, X_k\}$, *and* $\Pi_{B \in \pi}$ denotes the product over all cells of the partition π. Using (A.6) to compute $E\left(\overline{X}_n - \overline{Y}_m\right)^2 \Delta_n^{k-1}$ *to within* $O(n^{-4})$ *of their true values for* $k = 1, \ldots, 5$ involves the joint cumulants tabulated below, where $\Delta \equiv \Delta_n$, $\bar{X} \equiv \bar{X}_n$, and $\bar{Y} \equiv \bar{Y}_m$.

	0	1	2
0		$\chi_1(\overline{X} - \overline{Y})$	$\chi_1(\overline{X} - \overline{Y}, \overline{X} - \overline{Y})$
1	$\chi_1(\Delta)$	$\chi_2(\Delta, \overline{X} - \overline{Y})$	$\chi_3(\Delta, \overline{X} - \overline{Y}, \overline{X} - \overline{Y})$
2	$\chi_2(\Delta, \Delta)$	$\chi_3(\Delta, \Delta, \overline{X} - \overline{Y})$	$\chi_4(\Delta, \Delta, \overline{X} - \overline{Y}, \overline{X} - \overline{Y})$
3	$\chi_3(\Delta, \Delta, \Delta)$	$\chi_4(\Delta, \Delta, \Delta, \overline{X} - \overline{Y})$	$\chi_5(\Delta, \Delta, \Delta, \overline{X} - \overline{Y}, \overline{X} - \overline{Y})$

If $\kappa(i, j)$ denotes the ijth member of the table of joint cumulants, then (A.6) gives

$$E\left(\overline{X} - \overline{Y}\right)^2 = \kappa(0,2) + O\left(n^{-4}\right) \quad \text{(A.7)}$$

$$E\left(\overline{X} - \overline{Y}\right)^2 \Delta = \kappa(1,2) + \kappa(0,2)\,\kappa(1,0) + O\left(n^{-4}\right) \quad \text{(A.8)}$$

$$E\left(\overline{X} - \overline{Y}\right)^2 \Delta^2 = \kappa(2,2) + 2\kappa(1,0)\,\kappa(1,2) + \kappa(0,2)\,\kappa(2,0) + 2\kappa^2(1,1) + \kappa(0,2)\,\kappa^2(1,0) + O\left(n^{-4}\right) \quad \text{(A.9)}$$

$$\begin{aligned} E\left(\overline{X} - \overline{Y}\right)^2 \Delta^3 &= \kappa(0,2)\,\kappa(3,0) \\ &+ 6\kappa(1,1)\,\kappa(2,1) \\ &+ 3\kappa(2,0)\,\kappa(1,2) \\ &+ 3\kappa(2,0)\,\kappa(1,2) \\ &+ 3\kappa(1,0)\,\kappa(2,0)\,\kappa(0,2) \\ &+ 6\kappa(1,0)\,\kappa^2(1,1) + O\left(n^{-4}\right) \end{aligned} \quad \text{(A.10)}$$

$$E\left(\overline{X}-\overline{Y}\right)^2\Delta^4=3\kappa\left(0,2\right)\kappa^2\left(2,0\right)+12\kappa^2\left(1,1\right)\kappa\left(2,0\right)+O\left(n^{-4}\right) \quad \text{(A.11)}$$

after removing cumulant products of order smaller than O(n^{-4}) *and noting that* $\kappa(0, 1) = 0$. Each cumulant is simplified using rules found in Brillinger (1981), and the formula

$$\chi_k\left(X_1,\ldots,X_k\right)\Sigma_\pi(-1)^{(|\pi|-1)}\left(|\pi|-1\right)!\Pi_{B\in\pi}E\left(\Pi_{i\in B}X_i\right) \quad \text{(A.12)}$$

from Leonov & Shiryaev (1959) provides expressions for the simplified cumulants in terms of moments. Each cumulant is computed exactly or is approximated to within the order necessary for the cumulant products in (A.7)–(A.11) to lie within O(n^{-4}) *of their true values. Two examples are worked out below.*

$$\begin{aligned}
\kappa\left(1,1\right) &= \chi_2\left(\Delta,\overline{X}-\overline{Y}\right)\\
&= \chi_2\left(\overline{X^2},\overline{X}\right)-\chi_2\left(\overline{X^2},\overline{X}\right)+\left(n/m\right)\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y}\right)\\
&= n^{-1}\chi_2\left(X_1^2,X_1\right)-n^{-3}\chi_2\left(\Sigma_iX_i^2+\Sigma_{i\neq j}X_iX_j,\Sigma_jX_i\right)+\left(n/m\right)\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y}\right)\\
&= n^{-1}\mu_3^{'}-n^{-2}\chi_2\left(X_1^2+X_1\Sigma_{j=2}^nX_j,X_1\right)+\left(n/m\right)\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y}\right)\\
&= \left(n^{-1}-n^{-2}\right)\mu_3^{'}+n^{-2}\left(n-1\right)\chi_2\left(X_1X_2,X_1\right)+\left(n/m\right)\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y}\right)\\
&= \left(n^{-1}-n^{-2}\right)\mu_3^{'}+\left(n/m\right)\left(m^{-1}-m^{-2}\right)\eta_3^{'}
\end{aligned}$$

$$\begin{aligned}
\kappa\left(2,0\right) &= \chi_2\left(\Delta,\Delta\right)\\
&= \chi_2\left(\overline{X^2}-\overline{X}^2,\overline{X^2}-\overline{X}^2\right)+\left(n/m\right)^2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y^2}-\overline{Y}^2\right)\\
&= \chi_2\left(\overline{X^2},\overline{X^2}\right)-2\chi_2\left(\overline{X^2},\overline{X^2}\right)+\left(n/m\right)^2\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y^2}-\overline{Y}^2\right)\\
&= n^{-1}\chi_2\left(X_1^2,X_1^2\right)-2n^{-3}\left\{\chi_2\left(\Sigma_iX_i^2,\Sigma_iX_i^2\right)+\chi_2\left(\Sigma_{i\neq j},X_iX_j,\Sigma_iX_i^2\right)\right\}+n^{-1}\left\{\chi_2\left(\Sigma_iX_i^2,\Sigma_iX_i^2\right)-2\chi_2\left(\Sigma_{i\neq j}X_iX_j,\Sigma_j\right.\right.\\
&= \left(n^{-1}-2n^{-2}+n^{-3}\right)\left(\mu_4^{'}-\sigma_1^4\right)+2\left(n-1\right)n^{-3}\sigma_1^4+\left(n/m\right)^2\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y^2}-\overline{Y}^2\right)\\
&= \left(n-1\right)^2n^{-3}\mu_4^{'}-\left(n-1\right)\left(n-3\right)n^{-3}\sigma_1^4+\left(n/m\right)^2\chi_2\left(\overline{Y^2}-\overline{Y}^2,\overline{Y^2}-\overline{Y}^2\right)\\
&= \left(n^{-1}-2n^{-2}\right)\mu_4^{'}-\left(n^{-1}-4n^{-2}\right)\sigma_1^4+\left(n/m\right)^2\left\{\left(m^{-1}-2m^{-2}\right)\eta_4^{'}-\left(m^{-1}-4m^{-2}\right)\sigma_2^4\right\}+\left(n^{-3}\right)
\end{aligned}$$

Once all the cumulant expressions are obtained, they may be plugged into (A.7)–(A.11). Then, adding and subtracting (A.7)–(A.11) according to the expansion in (A.5) and gathering terms out of which n^{-1} *and* n^{-2} can be factored yields c_n from (A.3) and d_n from (A.4), respectively, which completes the proof.

References

Anderson, OD. Time Series Analysis and Forecasting: The Box-Jenkins Approach. Butterworths; 19 Cummings Park, Woburn, MA, 01801: 1977.

Bai Z, Saranadasa H. Effect of high dimension: By an example of a two sample problem. Statistica Sinica. 1996; 6:311–329.

Baladandayuthapani V, Ji Y, Talluri R, Nieto-Barajas LE, Morris JS. Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. Journal of the American Statistical Association. 2010; 105:1358–1375. [PubMed: 21512611]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57:289–300.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 2001; 29:1165–1188.

Brillinger, D. Holden-Day series in time series analysis. Holden-Day; 1981. Time Series: Data Analysis and Theory..

Brockwell, P.; Davis, R. Springer Series in Statistics. Springer; New York: 2009. Time Series: Theory and Methods..

Cai T, Liu W, Luo X. A constrained l-1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011; 106:594–607.

Cai T, Liu W, Xia Y. Two-sample covariance matrix testing and support recovery. Journal of the American Statistical Association. 2013; 108:265–277.

Cai TT, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. J. R. Statist. Soc. B. 2014; 76:349–372.

Chen SX, Qin YL. A two sample test for high dimensional data with applications to gene-set testing. The Annals of Statistics. 2010; 38:808–835.

Febrero-Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software. 2012; 51:1–28. [PubMed: 23504300]

Hall P, Jing B-Y, Lahiri SN. On the sampling window method for long-range dependent data. Statistica Sinica. 1998; 8:1189–1204.

Leonov VP, Shiryaev AN. On a method of calculation of semi-invariants. Theory of Probability and Its Applications. 1959; 4:319–329.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics. 2004; 5:557–572. [PubMed: 15475419]

Pang H, Liu H, Vanderbei R. fastclime: A fast solver for parameterized lp problems and constrained l1 minimization approach to sparse precision matrix estimation. R package version 1.2.3. 2013

Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nature Genetics Supplement. 2005; 37:S11–S17.

Politis DN, Romano JP. Bias-corrected nonparametric spectral estimation. J. Time Ser. Anal. 1995; 16:67–104.

Ruiz-Meana M, Garcia-Dorado D, Pina P, Inserte J, Agulló L, Soler-Soler J. Cariporide preserves mitochondrial proton gradient and delays atp depletion in cardiomyocytes during ischemic conditions. Am J Physiol Heart Circ Physiol. 2003; 285:H999–H1006. [PubMed: 12915386]

Seshan VE, Olshen A. DNAcopy: DNA copy number data analysis. R package version 1.36.0. 2013

Srivastava M. Multivariate theory for analyzing high dimensional data. J. Japan Statist. Soc. 2007; 37:53–86.

Srivastava MS, Kubokawa T. Tests for multivariate analysis of variance in high dimension under non-normality. Journal of Multivariate Analysis. 2013; 115:204–216.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

Wu Y, Genton MC, Stefanski LA. A multivariate two-sample test for small sample size and missing data. Biometrics. 2006; 62:877–885. [PubMed: 16984331]

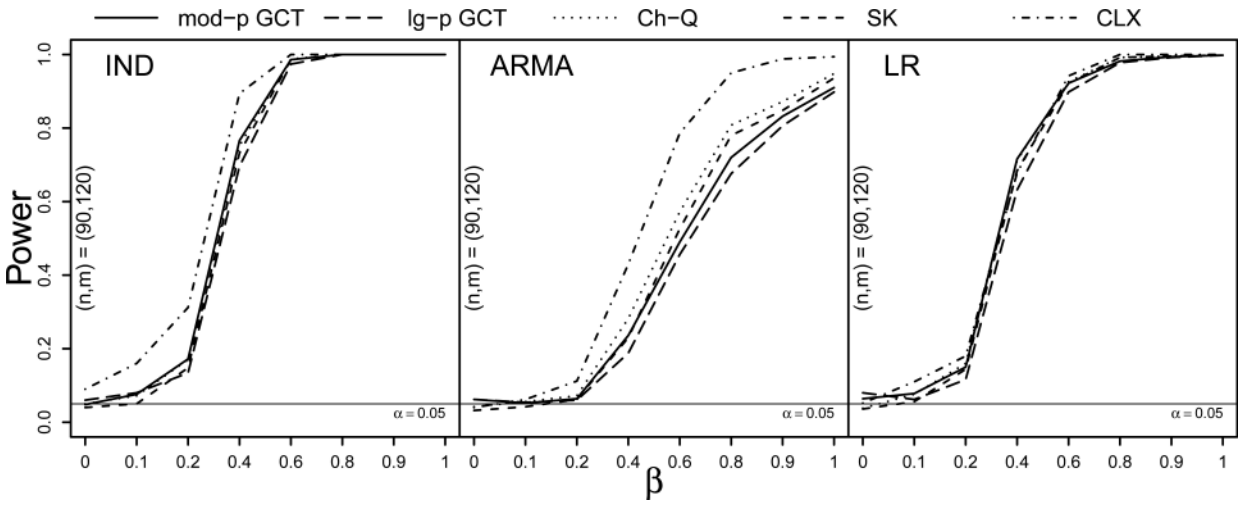

Figure 1.

Power curves at sample sizes (n, m) = (90, 120) for the moderate- and large-p GCT, Ch-Q, SK, and CLX tests against the proportion of nonzero mean differences β under IND, ARMA, and LR dependence (left to right) with centered gamma(4, 2) innovations and $\Sigma_1 = \Sigma_2$. Based on S = 500 simulations.

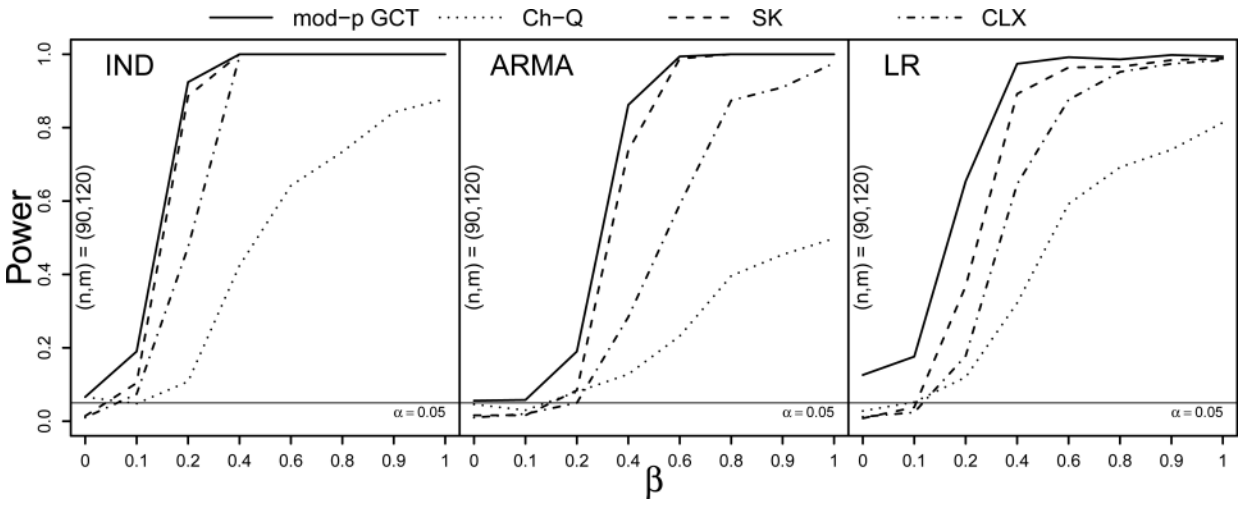

Figure 2.

Power curves at sample sizes (n, m) = (90, 120) for the large-p GCT, Ch-Q, SK, and CLX tests against the proportion of nonzero mean differences β under IND, ARMA, and LR dependence (left to right) with double Pareto(1.5,1) innovations and $\Sigma_1 = \Sigma_2$. Based on S = 500 simulations.

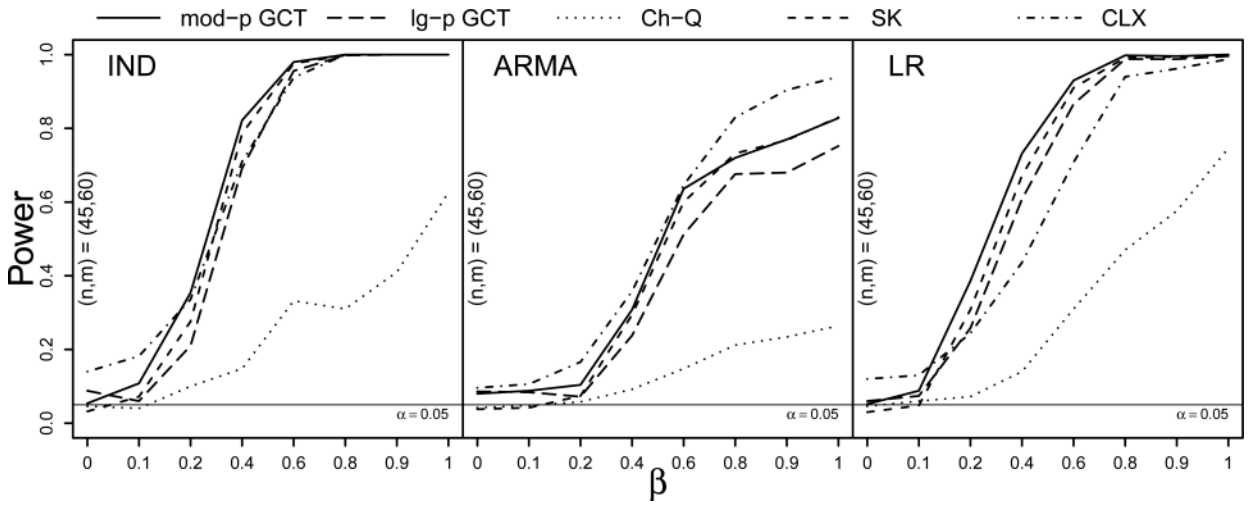

Figure 3.

Power curves at sample sizes (n, m) = (45, 60) for the moderate- and large-p GCT, Ch-Q, SK, and CLX tests against the proportion of nonzero mean differences β under IND, ARMA, and LR dependence (left to right) with heteroscedastic centered gamma(4, 2) innovations and $\Sigma_1 = \Sigma_2$. Based on $S = 500$ simulations.

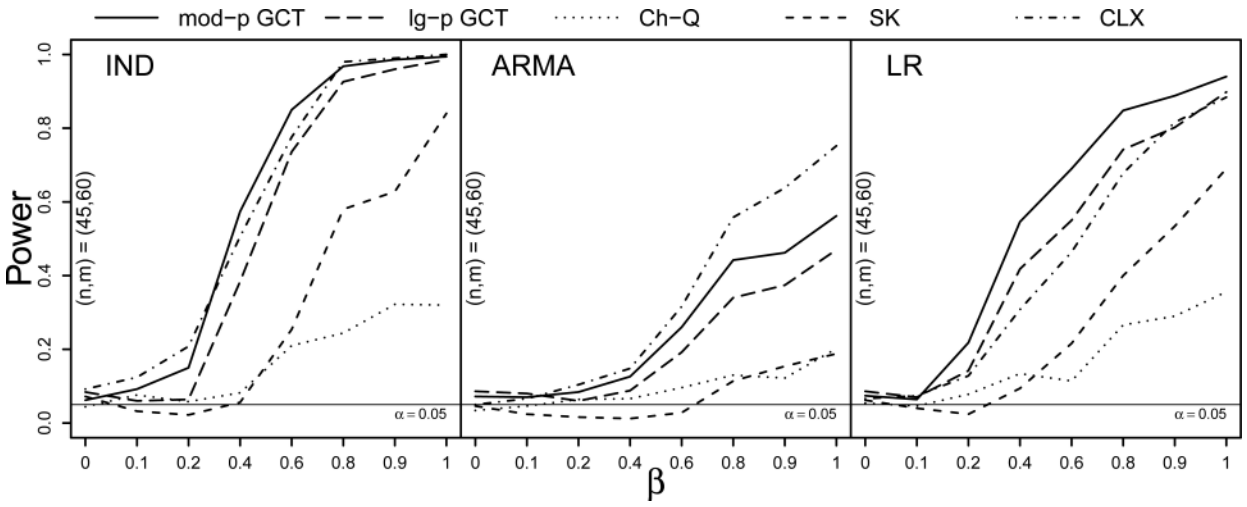

Figure 4.

Power curves at sample sizes (n, m) = (45, 60) for the moderate- and large-p GCT, Ch-Q, SK, and CLX tests against the proportion of nonzero mean differences β under IND, ARMA, and LR dependence (left to right) with heteroscedastic centered gamma(4, 2) innovations and $\Sigma_2 = 2\Sigma_1$. Based on $S = 500$ simulations.

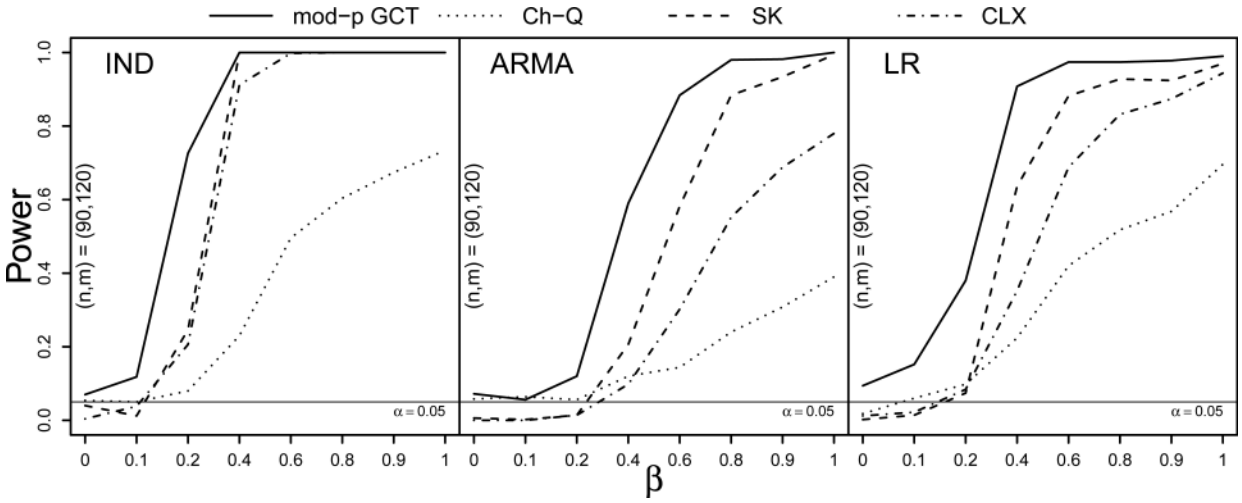

Figure 5.

Power curves at sample sizes (n, m) = (90, 120) for the moderate-p GCT, Ch-Q, SK, and CLX tests against the proportion of nonzero mean differences β under IND, ARMA, and LR dependence (left to right) with double Pareto(1.5,1) innovations and $\Sigma_2 = 2\Sigma_1$. Based on S = 500 simulations.

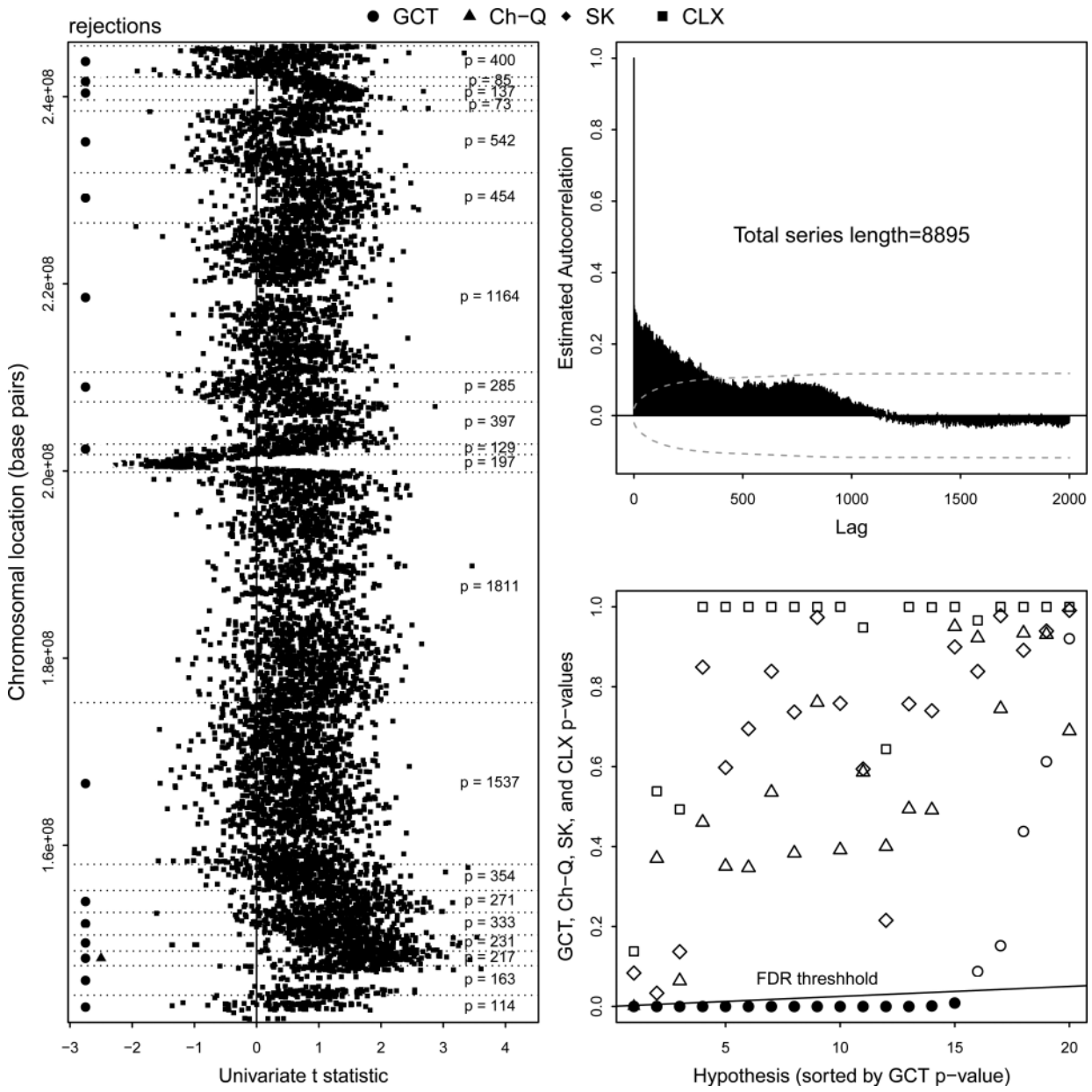

Figure 6.

(Left) Univariate *t*-statistics (t_{nj}) plotted against base-pair location on q arm of chromosome 1. Filled symbols denote rejections from FDR procedure for the GCT, Ch-Q, SK, and CLX tests. The number of components *p* within each CBS-selected chromosomal region is shown. (Upper right) Estimated autocorrelation function for squared univariate *t*-statistics along q arm of chromosome 1 with large-lag confidence bands. (Lower right) FDR results, hypotheses sorted by GCT *p*-values. FDR rejection threshold shown with filled symbols denoting rejections.

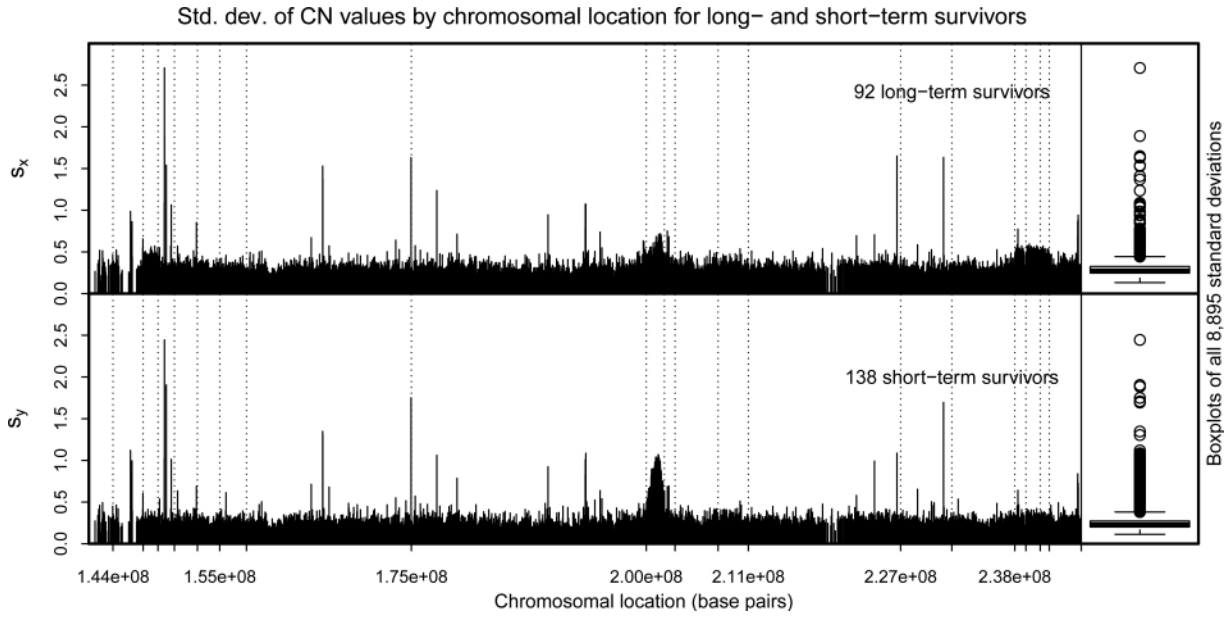

Figure 7.

Sample standard deviations of copy number at all 8,894 copy number locations for long- and short-term survivors with boxplots at right. Gaps occur at chromosomal locations where no copy number measurements were taken. Vertical dashed lines delineate the twenty CBS-selected regions in which the equal means hypothesis was tested.

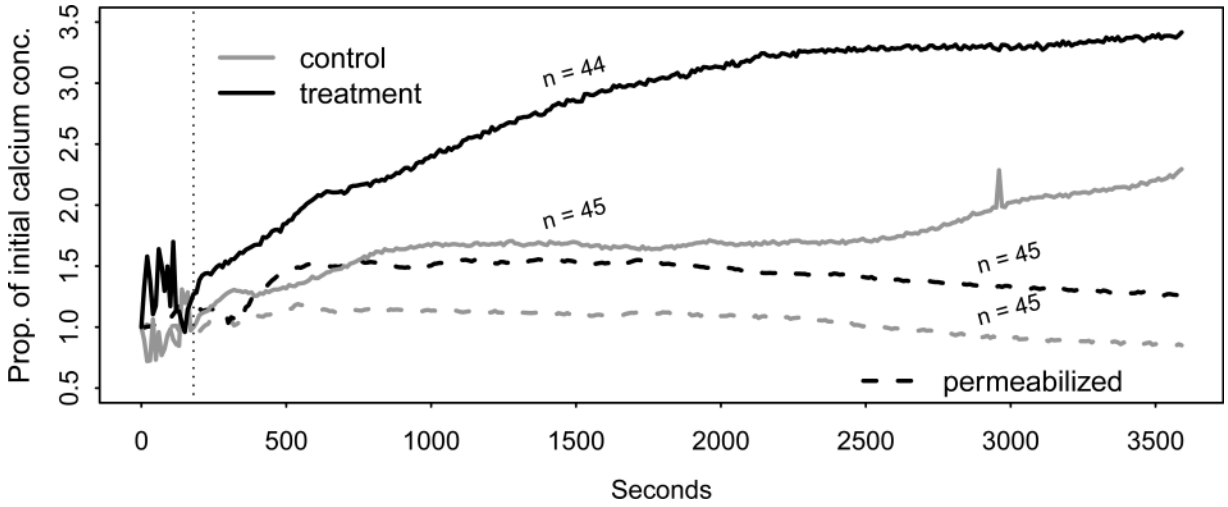

Figure 8.

Mean curves of the proportional increase in calcium concentration over initial value in intact and permeabilized cells from cardiac muscles in mice over one hour with and without cariporide treatment. First 180 seconds removed from analysis.

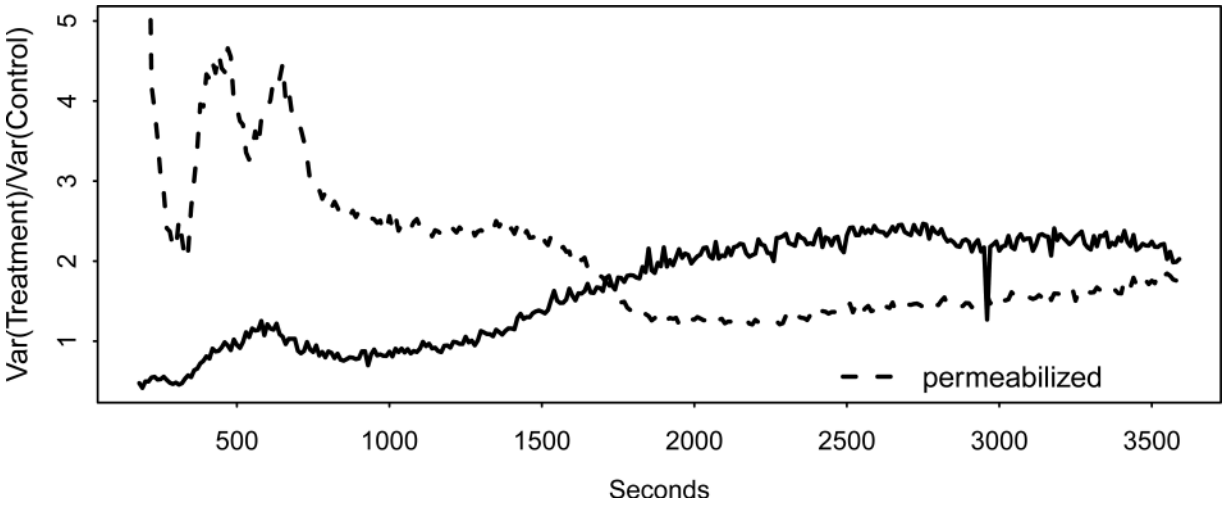

Figure 9.

Ratios of the variances of the proportional increase in calcium concentration for the treatment versus control group plotted against time for the intact and permeabilized data sets.

Table 1

Type I error rates over $S = 500$ simulations with nominal size $\alpha = .05$ for the moderate- and large-p GCT under the Parzen and trapezoid lag windows at lengths $L = 10, 15, 20$ and for the Ch-Q, SK, CLX tests under Normal(0, 1) innovations with $\Sigma_1 = \Sigma_2$.

$p = 300, \Sigma_1 = \Sigma_2$			Normal(0, 1) Innovations								
$\xi_n \equiv 1$						Parzen Window			Trapezoid Window		
n	m	Cov	Ch-Q	SK	CLX	L = 10	L = 15	L = 20	L = 10	L = 15	L = 20
45	60	IND	0.07	0.04	0.09	0.06	0.07	0.07	0.06	0.08	0.07
		ARMA	0.06	0.04	0.08	0.06	0.07	0.07	0.07	0.08	0.07
		LR	0.05	0.04	0.10	0.06	0.06	0.07	0.08	0.09	0.07
90	120	IND	0.05	0.04	0.07	0.06	0.06	0.06	0.07	0.08	0.06
		ARMA	0.05	0.04	0.06	0.07	0.07	0.08	0.08	0.09	0.08
		LR	0.03	0.03	0.07	0.05	0.05	0.07	0.06	0.08	0.07

$\xi_n \equiv 1 + a_n/n + b_n/n^2$						Parzen Window			Trapezoid Window		
n	m	Cov	Ch-Q	SK	CLX	L = 10	L = 15	L = 20	L = 10	L = 15	L = 20
45	60	IND	0.07	0.04	0.09	0.07	0.07	0.07	0.07	0.08	0.07
		ARMA	0.06	0.04	0.08	0.07	0.07	0.07	0.07	0.08	0.07
		LR	0.05	0.04	0.10	0.07	0.07	0.08	0.08	0.09	0.08
90	120	IND	0.05	0.04	0.07	0.06	0.06	0.07	0.07	0.08	0.07
		ARMA	0.05	0.04	0.06	0.08	0.08	0.08	0.08	0.09	0.08
		LR	0.03	0.03	0.07	0.06	0.06	0.06	0.07	0.09	0.06

Table 2

The *p*-values produced by the four tests for equality between the treatment and control calcium concentration curves in the intact and permeabilized experiments.

	Ch-Q	SK	CLX	mod-*p* GCT	lg-*p* GCT
Intact	0.000	0.118	0.086	0.000	0.000
Permeabilized	0.001	0.358	0.817	0.000	0.000