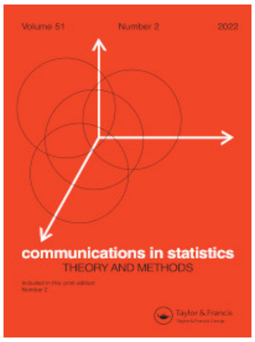

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A robust Hotelling test statistic for one sample case in high dimensional data

Hasan Bulut

To cite this article: Hasan Bulut (2021): A robust Hotelling test statistic for one sample case in high dimensional data, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2021.1996606

To link to this article: https://doi.org/10.1080/03610926.2021.1996606

Published online: 28 Oct 2021.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2021.1996606

Taylor & Francis
Taylor & Francis Group

Check for updates

A robust Hotelling test statistic for one sample case in high dimensional data

Hasan Bulut

Department of Statistics, Ondokuz Mayıs University, Samsun, Turkey

ABSTRACT

The Hotelling T^2 statistic is used to test the hypothesis about the location parameter of multivariate Gaussian distribution, and it is significantly sensitive to outliers. Also, we cannot calculate it when the sample size is less than the number of variables because this statistic needs the inverse of the covariance matrix, and the sample covariance matrix is singular in high dimensional data. Although a new approach, based on shrinkage estimation, was proposed to solve this singularity problem, this estimator is still sensitive to outliers. On the other hand, a robust one sample Hotelling T^2 statistic was proposed by using the minimum covariance determinant (MCD) estimates instead of classical ones. Since the MCD estimates cannot be calculated when $n < p$, this statistic cannot be used in high-dimensional data. This study proposes to use the minimum regularized covariance determinant (MRCD) estimator instead of classical or MCD. The MRCD estimator is a robust location and scatter estimator, which can be calculated in high-dimensional data. We obtain the asymptotic distribution of the proposed test statistic using Monte Carlo simulations and examine the power and robustness properties of the test statistic with simulated datasets. As a result, we show that the approximate distribution of the test statistic is proper, and the proposed robust test statistic can be used to test the hypothesis about the location parameter of contaminated high dimensional data. Finally, we construct an R function in the MVTests package to perform our proposed test statistic.

ARTICLE HISTORY
Received 7 October 2020
Accepted 17 October 2021

KEYWORDS
Robust Hotelling test statistic; one-sample Hotelling test; high-dimensional data; minimum regularized covariance estimators; shrinkage-based diagonal Hotelling test; MVTests

1. Introduction

When we know the scatter parameter $\mathbf{\Sigma}$ of multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, we can use the one-sample Hotelling T^2 statistic given in Equation (1) to test $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

$$T^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_0) \tag{1}$$

where $(.)^T$ is the transpose of a matrix, $\overline{\mathbf{X}}$ and $\mathbf{S}$ are the mean vector and covariance matrix of a sample from the distribution $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, respectively (Hotelling 1992). $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected at the level α (Bulut 2018; Rencher 2003) when

CONTACT Hasan Bulut hasan.bulut@omu.edu.tr Department of Statistics, Ondokuz Mayıs University, Samsun, Turkey.

$$T^2 > \frac{(n-1)p}{n-p} F_{p;n-p;1-\alpha}. \tag{2}$$

If the sample size (n) is less than the number of variables (p), the sample covariance matrix will be singular, and its inverse cannot be obtained. For this reason, classical T^2 statistic cannot be calculated in a high dimensional data in which $n < p$. Generally, there are three approaches in the literature to overcome this problem. The first approach is to remove the sample covariance matrix from the test statistic. This approach generally uses Hotelling's statistic for two samples (Bai and Saranadasa 1996; S. X. Chen and Qin 2010; Zhang and Xu 2009). The second approach uses a regularized covariance matrix instead of the classical one (L. S. Chen et al. 2011). The third approach is to use a diagonal covariance matrix. Dong et al. (2016) proposed the shrinkage-based diagonal Hotelling test statistic for high dimensional data based on the third approach. This test statistic is introduced in Section 5 of this study. Although the shrinkage-based diagonal Hotelling test statistic is helpful in high-dimensional data, it is still sensitive to outliers.

On the other hand, it is well known that the classical location and scatter estimators used in Equation (1) are sensitive to outliers in data (Huber and Ronchetti 2009). Therefore, when the data contains outliers, the Hotelling T^2 test statistic in Equation (1) produces results that are heavily affected by the outliers. To eliminate the effects of outliers on the test statistic, Willems et al. (2002) proposed to use Minimum Covariance Determinant (MCD) estimators instead of classical ones. In their study, they showed that their Hotelling T^2 statistic is not sensitive to outliers.

However, this statistic cannot be computed in high-dimensional data sets because MCD estimators are only used when $n > p$.

To solve this lack of MCD estimators, Boudt et al. (2020) proposed Minimum Regularized Covariance Determinant (MRCD) estimators, which have the properties of MCD estimators in high dimensional datasets. In literature mostly, the MRCD estimators are used to detect outliers in high-dimensional data sets. In one of these studies, Bulut (2020) uses the MRCD estimators to obtain robust Mahalanobis distances in high dimensions.

This study proposes a robust Hotelling T^2 statistic, which can be used in high-dimensional datasets. So, we solve both outlier and high dimensional data problems for Hotelling T^2 statistic. Our statistic uses MRCD estimations in Equation (1) instead of classical or MCD ones. Since the distribution of test statistics based on MRCD estimations differs from the classical one, we obtain approximate distribution using Monte Carlo simulations.

In addition, we construct two R functions for our proposed test statistic in the MVTests package (Bulut 2019). The first function is simRHT2. This function performs a Monte Carlo simulation for different sample sizes and variable numbers given by users to obtain d and q values of the approximate F distribution. Note that there may be differences here as the outputs of the function are based on simulation. The second function is RHT2. This function performs the robust Hotelling T^2 test in high dimensional data, and it needs the d and q values obtained by the simRHT2 function.

The rest of the paper is organized as follows. In Section 2, the MRCD estimators are introduced. In Section 3, we define the proposed robust T^2 statistic and construct its approximate distribution. We investigate whether the approximate distribution is correct for the suggested test statistic using Monte Carlo simulations in Section 4. We introduce shrinkage-based diagonal Hotelling T^2 statistic for one sample in Section 5. Section 6 compares our proposed statistic's power and robustness performance with the shrinkage-based diagonal Hotelling test statistic suggested in (Dong et al. 2016). We give an example of robust inference in Section 7. Finally, we provide conclusions in Section 8.

2. Minimum regularized covariance determinant (MRCD) estimators

The most popular robust estimators of multivariate location and scatters parameters are based on the Minimum Covariance Determinant (MCD) method (Rousseeuw 1985). This method aims to determine the subset with the lowest sample covariance determinant. Let this subset be H_{MCD}. The MCD estimations of location and scatter parameters are the mean vector and the covariance matrix of H_{MCD}, respectively. The robustness properties of MCD estimators were mentioned before (Croux and Haesbroeck 1999; Lopuhaa and Rousseeuw 1991). However, it is well known that the MCD estimators can be estimated when $h \geq p$, otherwise the MCD covariance matrix will be singular. Boudt et al. (2020) proposed Minimum Regularized Covariance Determinant (MRCD) estimators to estimate the robust location and scatter parameters in high dimensional data to overcome this problem. These estimators estimate the location and scatter parameters in high-dimensional data without being affected by the outliers. The MRCD estimator has the good breakdown point properties of the MCD estimator (Bulut 2020).

The MRCD estimations are calculated as follows: First, we standardize data by using the univariate location and scatter estimators. These univariate estimations are the median and Q_n (Rousseeuw and Croux 1993), respectively. Then a target matrix $\boldsymbol{T}$, which is symmetric and positive definite, is chosen. This matrix can be selected in two different structures. The first option is to use the identity matrix as the target matrix, while the second option is an equicorrelation matrix. In this study, the target matrix $\boldsymbol{T}$ is selected as the identity matrix.

The regularized covariance matrix of any subset H, which is obtained from standardized $\boldsymbol{Z}$ data, is calculated as:

$$\boldsymbol{K}(H) = \rho \boldsymbol{T} + (1-\rho) c_\alpha \boldsymbol{S}_Z(H) \tag{3}$$

where ρ is the regularization parameter and c_α is the consistency factor defined by (Croux and Haesbroeck 1999), and,

$$\boldsymbol{S}_Z(H) = \frac{1}{h-1}\left(\boldsymbol{Z}_H - \boldsymbol{\mu}_Z(H)\right)^T\left(\boldsymbol{Z}_H - \boldsymbol{\mu}_Z(H)\right), \;\; \boldsymbol{\mu}_Z(H) = \frac{1}{h}\boldsymbol{Z}_H^T 1_h \tag{4}$$

MRCD estimations are obtained from the subset H_{MRCD}. This subset is obtained by solving the minimization problem given below.

$$H_{MRCD} = \underset{H \in \mathcal{H}}{\operatorname{argmin}} \left[det(\boldsymbol{K}(H))^{1/p} \right] \tag{5}$$

where $\mathcal{H}$ is the set, which consists of all subsets with size h in data. Finally, the MRCD location and scatter estimators are defined as follows:

$$\hat{\boldsymbol{\mu}}_{MRCD} = \boldsymbol{V}_X + \boldsymbol{D}_X \ \boldsymbol{\mu}_Z(H_{MRCD}) \tag{6}$$

$$\hat{\boldsymbol{\Sigma}}_{MRCD} = \boldsymbol{D}_X \ \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\left[\rho\boldsymbol{I} + (1-\rho)c_\alpha \boldsymbol{S}_w(H_{MRCD})\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'\boldsymbol{D}_X \tag{7}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{Q}$ are eigenvalues and eigenvectors matrices of $\boldsymbol{T}$, respectively. On the other hand, $\boldsymbol{S}_w(H_{MRCD})$ is calculated as below:

$$\boldsymbol{S}_w(H_{MRCD}) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}'\boldsymbol{S}_Z(H_{MRCD})\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}. \tag{8}$$

Comprehensive information about MRCD estimators is available (Boudt et al. 2020). In this study, we have used the "rrcov" package in the R software for calculations regarding the MRCD estimators (Todorov and Filzmoser 2009).

3. Robust T^2 Statistic for High Dimensional Data

The classical Hotelling T^2 statistic which is used to test the hypothesis about the location parameter of the multivariate Gaussian distribution can be expressed as below:

$$T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0) = \frac{(n-1)p}{n-p} F_{p;n-p;1-\alpha} \tag{9}$$

We may use this statistic thanks to the following properties,

- $\overline{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$
- $(n-1)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$, where W_p is the Wishart distribution.
- $\overline{\boldsymbol{X}}$ and $\mathbf{S}$ are independent (Bulut 2018; Rencher 2003; Willems et al. 2002).

When $n < p$, the T^2 statistic cannot be calculated since the inverse of matrix $\mathbf{S}$ cannot be obtained. Moreover, the F-distribution in Equation (9) cannot be used because $(n-p)$ is less than zero when $n < p$.

To solve these problems in high dimensional data sets, we construct a robust test statistic using the MRCD location and scatter estimations instead of the classical ones in Equation (9). For n observations from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we define a robust test statistic in high dimensional data as given in Equation (10):

$$T^2_{MRCD} = n\left(\overline{\boldsymbol{X}}_{MRCD} - \boldsymbol{\mu}_0\right)^T \boldsymbol{S}^{-1}_{MRCD}\left(\overline{\boldsymbol{X}}_{MRCD} - \boldsymbol{\mu}_0\right) \tag{10}$$

in this equation $\overline{\boldsymbol{X}}_{MRCD}$ and $\boldsymbol{S}_{MRCD}$ are MRCD location and scatter estimations, respectively. Since the finite-sample distribution of the MRCD estimates is unknown, we should use an approximate distribution as mentioned by (Todorov and Filzmoser 2010; Willems et al. 2002). In this manner, analogously to the F approximation given in Equation (9), we assume the following approximation for T^2_{MRCD} :

$$T^2_{MRCD} \approx dF_{p,q} \tag{11}$$

where d and q are constants. By using the properties of F distribution, the expected value and the variance of T^2_{MRCD} statistic are defined as below:

$$E\left[T^2_{MRCD}\right] = d\frac{q}{q-2} \tag{12}$$

Table 1. Percent above the 5% cutoff.

Distribution	n/p	40	60	80	100	200
$N_p(0, \boldsymbol{I})$	**20**	4.933	4.833	4.667	5.000	5.633
	30	4.433	4.633	4.833	4.967	4.733
	50	5.033	5.300	4.600	4.867	4.833
	100	4.900	5.067	4.800	5.000	4.867
$N_p(0, \boldsymbol{\Sigma})$	**20**	5.500	5.133	5.000	5.000	4.800
	30	4.200	5.233	4.733	4.600	4.867
	50	4.867	4.867	5.100	5.467	5.600
	100	4.467	4.600	4.800	5.000	5.400
$Cauchy_p(0, \boldsymbol{\Sigma})$	**20**	4.567	4.867	5.300	5.000	4.733
	30	5.300	5.067	5.000	4.967	4.633
	50	4.967	4.667	4.400	4.800	5.033
	100	5.133	5.100	5.100	5.600	4.267

$$Var\left[T^2_{MRCD}\right] = d^2 \frac{2q^2(p+q-2)}{p(q-4)(q-2)^2} \tag{13}$$

From Equations (12) and (13), we can obtain the below equations for d and q constants:

$$d = E\left[T^2_{MRCD}\right] \frac{q-2}{q} \tag{14}$$

$$q = \left(\frac{Var\left[T^2_{MRCD}\right]}{E\left[T^2_{MRCD}\right]^2} \frac{p}{2} - 1 \right)^{-1} (p+2) + 4 \tag{15}$$

As mentioned by (Todorov and Filzmoser 2010), the mean and the variance of the T^2_{MRCD} statistic cannot be calculated analytically. For this reason, we select d and q constants by Monte Carlo simulation. We select only one-time the d and q values to obtain the approximate F distribution for any n and p, and we give the d and q values for several n and p values in Table A.1. In addition, we can use the simRHT2 function in the MVTests package to obtain the d and q values for different n and p values from Table A.1. On the other hand, we can use the RHT2 function in the MVTests package to calculate the T^2_{MRCD} value and perform the robust test in high dimensional data using d and q values (Bulut 2019).

4. Approximate Distribution of T^2_{MRCD} Statistic

We have mentioned that we should calculate d and q constants by using Monte Carlo simulation to obtain the approximate F distribution given in Equation (11). For this purpose, we construct three simulation designs because the MRCD estimators are not affine equivariant. So, we want to see whether our proposed test statistic is impressed from distributions of data or not. These designs are defined below:

- Simulation Design-1: Similar to the design used by (Willems et al. 2002), we generate randomly $m = 3000$ samples from standard multivariate Gaussian distribution $N_p(0, \mathbf{I_p})$ for different n and p.
- Simulation Design-2: We generate randomly $m = 3000$ samples from multivariate Gaussian distribution $N_p(0, \ \boldsymbol{\Sigma})$ for different n and p, where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix simulated randomly. Moreover, we assume without loss of generality, the variances in diagonal of $\boldsymbol{\Sigma}$ are all equal to 1. We generate

Table 2. Percent above the 1% cutoff.

Distribution	n/p	40	60	80	100	200
$N_p(0, \boldsymbol{I})$	**20**	0.933	1.067	0.933	0.967	1.000
	30	0.667	0.700	0.767	0.833	0.967
	50	0.667	0.767	0.767	1.033	0.800
	100	0.800	0.700	0.643	0.933	1.067
$N_p(0, \boldsymbol{\Sigma})$	**20**	0.800	0.800	1.167	0.900	0.933
	30	0.800	0.767	0.800	0.933	0.800
	50	0.733	0.733	0.800	0.800	1.100
	100	0.733	0.800	0.667	0.800	0.800
$Cauchy_p(0, \boldsymbol{\Sigma})$	**20**	1.067	0.833	0.833	0.833	0.967
	30	0.700	0.800	0.767	0.700	0.900
	50	0.667	0.867	0.767	0.900	0.933
	100	0.833	0.733	1.033	0.800	0.767

a random covariance matrix $\boldsymbol{\Sigma}$ by first randomly generating eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ for the covariance matrix and then using the columns of a randomly generated orthogonal matrix $\boldsymbol{P}$ as eigenvectors. The covariance matrix $\boldsymbol{\Sigma}$ is then constructed as $\boldsymbol{\Sigma} = \mathbf{P\Lambda}\boldsymbol{P}^{\boldsymbol{T}}$. For this aim, we use the function genPositiveDefMat in the R package "clusterGeneration" (Weiliang and Harry 2015). See Weiliang and Harry (2015) for details.

- Simulation Design-3: We generate randomly $m = 3000$ samples from multivariate Cauchy distribution $Cauchy_p(0, \ \boldsymbol{\Sigma})$ for different n and p, where $\boldsymbol{\Sigma}$ is generated as in Simulation Design-2.

In each design, we obtain MRCD location $\left(\overline{\boldsymbol{X}}_{(i)MRCD}\right)$ and scatter $\left(\boldsymbol{S}_{(i)MRCD}\right)$ estimates for each data matrix $\boldsymbol{X}_i$ $(i = 1, 2, \ldots, m)$ and we calculate $T^2_{(i)MRCD}$ values from Equation (10) by using these estimations. After that, we calculate the mean and variance of $T^2_{(i)MRCD}$ values, respectively.

$$ave\left(T^2_{MRCD}\right) = \frac{1}{m}\sum_{i=1}^{m} T^2_{(i)MRCD} \tag{16}$$

$$var\left(T^2_{MRCD}\right) = \frac{1}{m-1}\sum_{i=1}^{m}\left(T^2_{(i)MRCD} - ave\left(T^2_{MRCD}\right)\right)^2 \tag{17}$$

We use these values as estimations of $E\left[T^2_{MRCD}\right]$ and $Var\left[T^2_{MRCD}\right]$, respectively. Therefore, we obtain d and q values by replacing these estimations instead of $E\left[T^2_{MRCD}\right]$ and $Var\left[T^2_{MRCD}\right]$ in Equations (14) and (15). Finally, we construct the approximate F distribution of T^2_{MRCD} for known n and p by replacing the obtained d and q values in (11).

To investigate the accuracy of approximate F distribution, we compare the actual percentage of T^2_{MRCD} values above the cutoff with significance levels. In each design, we generate $m = 3000$ samples, and we calculate $T^2_{(i)MRCD}$ values for each sample $(i = 1, 2, \ldots, m)$. For each design, we give the percentage of $T^2_{(i)MRCD}$ values above the cutoff $(i = 1, 2, \ldots, m)$ in Table 1 for $\alpha = 0.05$ and Table 2 for $\alpha = 0.01$. According to Tables 1 and 2, the difference between the actual percentage and the determined significance level (5% and 1%, respectively) are unimportant and small. These percents are in two

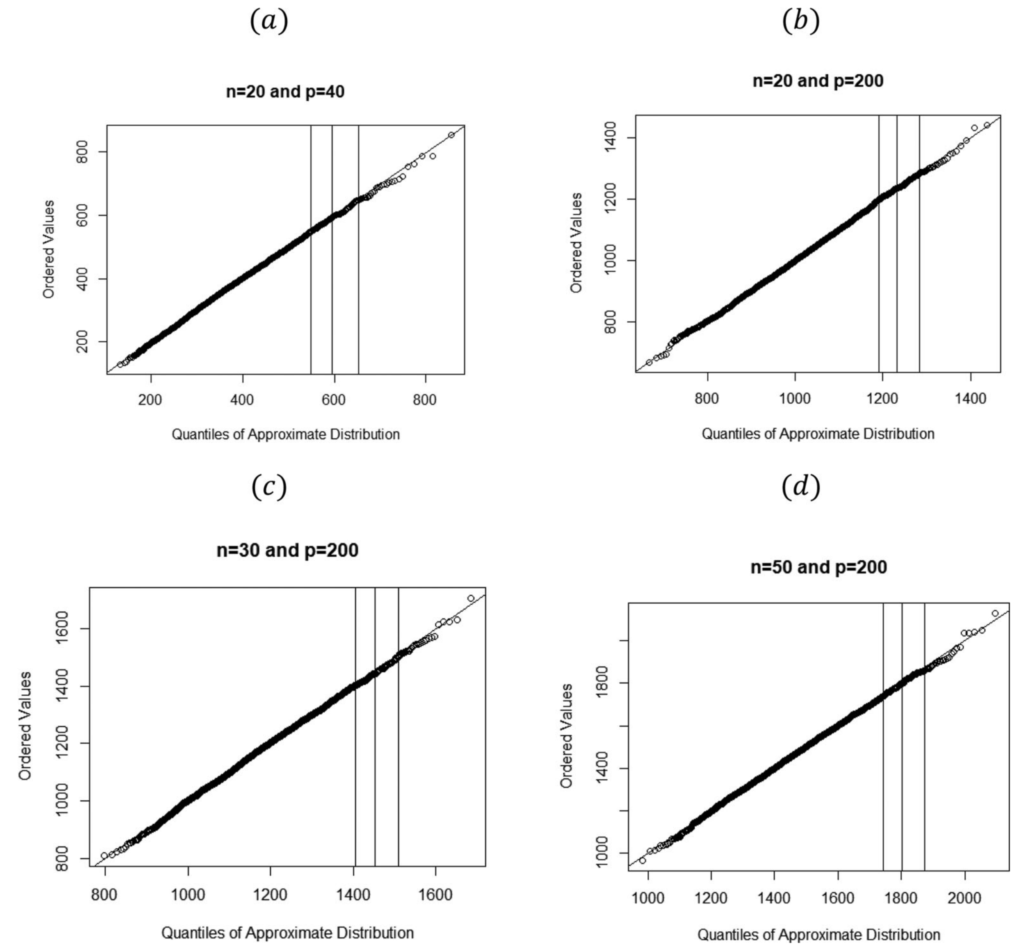

Figure 1. QQ plots for T^2_{MRCD} (The vertical lines indicate 95%, 97.5%, and 99% quantiles of the approximate distribution) (a) $n = 20$ and $p = 40$ (b) $n = 20$ and $p = 200$ (c) $n = 30$ and $p = 200$ (d) $n = 50$ and $p = 200$.

standard deviation intervals around the nominal levels: (4.2%, 5.8%), and (0.64%, 1.36%). Here, the standard error is $\sqrt{\alpha(1-\alpha)/3000}$.

Moreover, we plot the square root of the ordered $T^2_{(i)MRCD}$ values versus the square root of the quantiles of this approximate distribution to compare the empirical distribution of T^2_{MRCD} with the approximate distribution of T^2_{MRCD} given in (11). We provide some of the plots obtained from Simulation Design-1 in Figure 1. Because the plots of other designs are similar, we do not give them in Figure 1. The vertical lines indicate 95%, 97.5%, and 99% quantiles of the approximate distribution in these plots.

When Figure 1 is examined, approximate F distribution based on d and q obtained by Monte Carlo simulation yields a good approximation to the empirical distribution of T^2_{MRCD}.

According to these results, we show that the approximate F distribution is proper for T^2_{MRCD} test statistic and this approximate F distribution is robust in terms of the distribution of data.

5. Shrinkage-based diagonal Hotelling test

As discussed in previous chapters of this study, the classical Hotelling T^2 statistic cannot be used in high-dimensional data. Therefore, the use of a diagonal covariance matrix has been proposed in the literature. (Park and Ayyala 2013; Srivastava 2009; Srivastava and Du 2008; Srivastava, Katayama, and Kano 2013; Wu, Genton, and Stefanski 2006). However, Srivastava, Katayama, and Kano (2013) showed that this approach is not reliable for small n. Dong et al. (2016) proposed an approach based on shrinkage estimators called shrinkage-based diagonal Hotelling's test. This test statistic is defined as follows:

$$T^2_{Shrinkage} = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)^T \tilde{\mathbf{S}}^* (\overline{\mathbf{X}} - \boldsymbol{\mu}_0) \tag{18}$$

where $\tilde{\mathbf{S}}^* = diag\left[\tilde{\sigma}_1^{-2}, \ldots, \tilde{\sigma}_p^{-2}\right]$, $\tilde{\sigma}_j^{2t} = \left(h_{v,p}(t)\ \hat{\sigma}^{2t}_{pool}\right)^{\alpha} \left(h_{v,1}(t)\ \hat{\sigma}_j^{2t}\right)^{1-\alpha}$, $h_{v,p}(t) = \left(\frac{v}{2}\right)^t \left(\frac{\Gamma(v/2)}{\Gamma\left(\frac{v}{2}+\frac{t}{2}\right)}\right)^p$, $\hat{\sigma}^{2t}_{pool} = \prod_{j=1}^p \left(\hat{\sigma}_j^2\right)^{t/p}$, $\Gamma(.)$ is the gamma function, $t = -1$, $v = n - 1$, and $\alpha \in [0, 1]$ is a shrinkage parameter. They obtained an approximate distribution for the $T^2_{Shrinkage}$ as below:

$$T^2_{Shrinkage} \approx c\chi^2_d \tag{19}$$

where

$$c = \frac{(3C_2 - C_3)\hat{\sigma}^{-4\alpha}_{pool} \sum_{j=1}^p \hat{\sigma}_j^{4\alpha} + \left(C_3 - C_1^2\right)\hat{\sigma}^{-4\alpha}_{pool} \left(\sum_{j=1}^p \hat{\sigma}_j^{2\alpha}\right)^2}{2C_1 \hat{\sigma}^{-2\alpha}_{pool} \sum_{j=1}^p \hat{\sigma}_j^{2\alpha}},$$

$$d = \frac{2C_1^2 \hat{\sigma}^{-4\alpha}_{pool} \left(\sum_{j=1}^p \hat{\sigma}_j^{2\alpha}\right)^2}{(3C_2 - C_3)\hat{\sigma}^{-4\alpha}_{pool} \sum_{j=1}^p \hat{\sigma}_j^{4\alpha} + \left(C_3 - C_1^2\right)\hat{\sigma}^{-4\alpha}_{pool} \left(\sum_{j=1}^p \hat{\sigma}_j^{2\alpha}\right)^2},$$

$$C_1 = \frac{h^{\alpha}_{v,p}(-1)\ h^{1-\alpha}_{v,1}(-1)}{h^{p-1}_{v,1}(-\alpha/p) h_{v,1}(-\alpha/p - (1-\alpha))},$$

$$C_2 = \frac{h^{2\alpha}_{v,p}(-1)\ h^{2(1-\alpha)}_{v,1}(-1)}{h^{p-1}_{v,1}(-2\alpha/p) h_{v,1}(-2\alpha/p - 2(1-\alpha))},$$

$$C_3 = \frac{h^{2\alpha}_{v,p}(-1)\ h^{2(1-\alpha)}_{v,1}(-1)}{h^{p-2}_{v,1}(-2\alpha/p) h^2_{v,1}(-2\alpha/p - (1-\alpha))}.$$

When $\frac{T^2_{Shrinkage}}{c} > \chi^2_{d;\alpha}$, the null hypothesis is rejected. For further reading, detailed information about the shrinkage-based diagonal Hotelling Test is available in (Dong et al. 2016).

6. Simulation study

In this part, we compare the power and robustness performance of our proposed statistic T^2_{MRCD} with $T^2_{Shrinkage}$ suggested by Dong et al. (2016) by using a simulation study.

6.1. The power of T^2_{MRCD} statistic

The power of any test statistic is the probability of rejecting H_0 while it is false. We use two simulation designs to compare the power of test statistics.

- Simulation Design-4: Similar to the design used by (Willems et al. 2002), we generate randomly $m = 3000$ samples from multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \mathbf{I_p})$ for different n and p.
- Simulation Design-5: For different n and p, we generate randomly $m = 3000$ samples from multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\Sigma}$ is generated as in Simulation Design-2.

In all designs, $\boldsymbol{\mu}$ consists of $\mu_j \sim Unif(0.1,\ 0.4)$ for $j = 1, 2, \ldots p$. In all cases, we test $H_0 : \boldsymbol{\mu} = 0$ versus $H_1 : \boldsymbol{\mu} \neq 0$ at the 5% significance level. The power is the ratio of the rejection of the null hypothesis within $m = 3000$ samples. We provide the percentages of the rejection for each case in Table 3. According to Table 3, we see that the loss in power is acceptable. Note that the datasets generated are not contaminated.

Moreover, we compare the power of test statistics visually by constructing size-power plots. Davidson and MacKinnon (Davidson and MacKinnon 1998) proposed that the size-power plot should be used to compare test statistics' power when we do not possess their distribution knowledge. As used by (Todorov and Filzmoser 2010), we obtain size-power plots in the following way: (i) Firstly, we generate $m = 3000$ random samples from $N_p(0, \mathbf{I_p})$ distribution under H_0, which is defined as above. After that, we calculate the test statistics for each of them. Then, the values of the test statistics are sorted in ascending order. Let be γ_j is the j_{th} element of these sorting values. When the critical

Table 3. The power of test statistics 5% cutoff.

		40		60		80		100	
Simulation design	n/p	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}
Simulation Design-4	20	58.2	100	92	100	99.5	100	99.8	100
	30	63.7	100	99.4	100	100	100	100	100
	50	53.5	100	99.1	100	100	100	100	100
	100	63.1	100	99.9	100	99.1	100	100	100
Simulation Design-5	20	54.1	100	59	99.7	83	100	96.2	100
	30	63.7	100	76.2	100	96	100	99.4	100
	50	52.2	100	74.4	100	97.7	100	100	100
	100	60.4	100	97.4	100	99.4	100	99.9	100

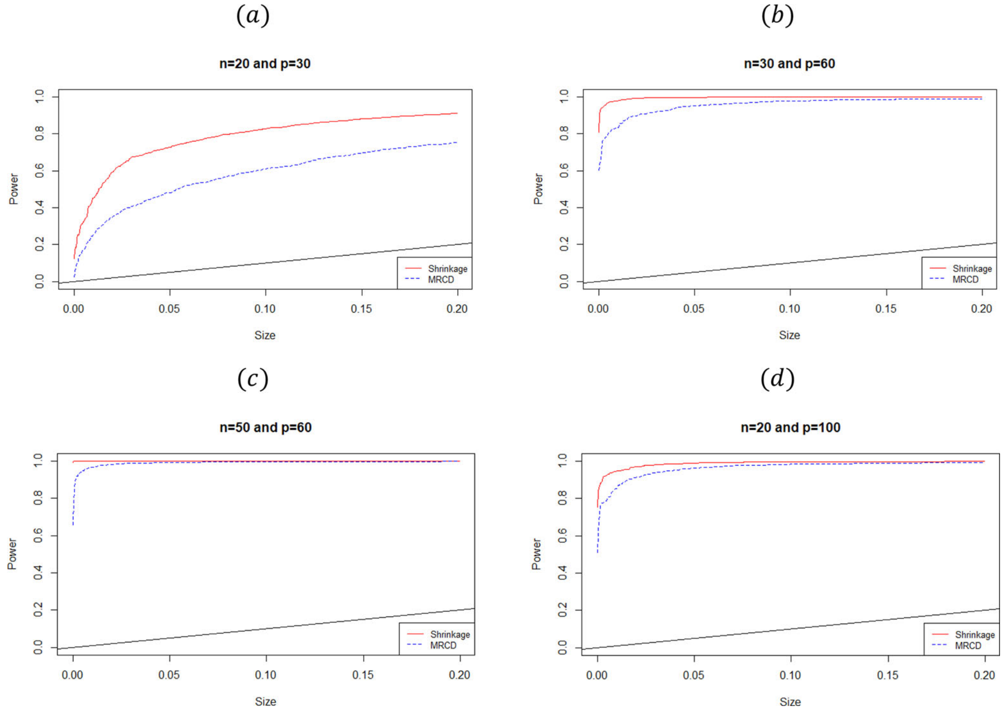

Figure 2. Size-power plots for several cases. The red dashed line means the size-power curve of $T^2_{Shrinkage}$ and the blue line means the size-power curve of T^2_{MRCD} (a) $n = 20$ and $p = 30$ (b) $n = 30$ and $p = 60$ (c) $n = 50$ and $p = 60$ (d) $n = 20$ and $p = 100$.

value is determined as γ_j, the size values s_j is calculated as $s_j = (m - j)/(m + 1)$. (ii) Secondly, we generate $m = 3000$ random samples from $N_p(\boldsymbol{\mu}, \mathbf{I_p})$ distribution under H_1, where $\boldsymbol{\mu}$ consists of $\mu_j \sim Unif(0.1,\ 0.4)$ for $j = 1, 2, ...p$. Then we calculate the test statistics for each of them. The ratio of these values that exceed the critical values γ_j is taken as the power p_j. (iii) Finally, we plot size-power graphs for all the pairs (s_j, p_j). For several cases, the size-power plots are given in Figure 2.

Todorov and Filzmoser (2010) mentioned that the size-power curves should lie above the 45° line, and the increasing area between the size-power curve and the 45° line means better. In Figure 2, the colors of T^2_{MRCD} and $T^2_{Shrinkage}$ curves are blue and red, respectively. According to Figure 2, we see that the size-power curves of both test statistics lie above the 45° line. Moreover, the 45° line does not seem in Figure 2b–d because the powers are considerably high. Although the curve of T^2_{MRCD} statistic is slightly below the shrinkage-based diagonal test statistic one in all cases, the loss of T^2_{MRCD} statistic's power is acceptable because of its robustness.

6.2. The robustness of T^2_{MRCD} statistic

In this subsection, we generate contaminated data to assess the robustness of T^2_{MRCD} statistic. For this purpose, we use two simulation designs:

Table 4. The rates of wrong rejections of the null hypothesis in data with 10% outliers.

		40		60		80		100	
Simulation design	n/p	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}	T^2_{MRCD}	T^2_{Shrin}
Simulation Design-6	20	0.2	98.5	0.1	99.4	0	99.6	0	99.8
	30	0	99.7	0	99.8	0	99.9	0	99.9
	50	0	99.9	0	100	0	100	0	100
	100	0	99.8	0	99.7	0	100	0	100
Simulation Design-7	20	0	98.9	0	99.1	0	99.6	0	99.5
	30	0	99.7	0	99.7	0	99.9	0	99.7
	50	0	99.8	0	99.8	0	100	0	99.8
	100	0	99.9	0	100	0	100	0	100

- Simulation Design-6: We generate 90% of data from the distribution $N_p(0, \mathbf{I_p})$ and remainder 10% of data from the distribution $Cauchy_p(\boldsymbol{\mu}_{out}, \boldsymbol{\Sigma}_{out})$ with $\boldsymbol{\mu}_{out} = (p+5, \ldots, p+5)^T$ and $\boldsymbol{\Sigma}_{out} = 0.1 \times \mathbf{I_p}$.
- Simulation Design-7: We generate 90% of data from the distribution $N_p(0, \boldsymbol{\Sigma})$ and remainder 10% of data from the distribution $Cauchy_p(\boldsymbol{\mu}_{out}, \boldsymbol{\Sigma}_{out})$ with $\boldsymbol{\mu}_{out} = (p+5, \ldots, p+5)^T$ and $\boldsymbol{\Sigma}_{out} = 0.1 \times \boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma}$ is generated as in Simulation Design-2.

In all cases, we test the null hypothesis $H_0 : \boldsymbol{\mu} = 0$ versus $\boldsymbol{H}_1 : \boldsymbol{\mu} \neq 0$ and we give the percentage of the wrong rejection of the null hypothesis for each case in Table 4. The low rate means that the test statistic is robust to outliers. According to Table 4, it is abundantly clear that the T^2_{MRCD} statistic is robust to outliers while $T^2_{Shrinkage}$ is heavily sensitive to outliers.

7. Example

In this section, in order to investigate the performance of T^2_{MRCD} in real data, we use the octane data set since the outliers are well known. The octane data set was described by (Esbensen, SchoníKopf, and Midtgaard 1995) and used by (Boudt et al. 2020; Hubert, Rousseeuw, and Vanden Branden 2005). Since the octane data consists of 39 observations and 226 variables, it is a high dimensional data. Moreover, it is well known that observations 25, 26, 36, 37, 38, and 39 in the data are outliers. So, the outliers rate is 15.38%. For this reason, we test the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is the mean vector of the clean data set obtained by excluding observations 25, 26, 36, 37, 38, and 39. As a result, we know the null hypothesis is true, and we want to see that T^2_{MRCD} fails to reject the null hypothesis without affecting outliers. Moreover, we give the d and q values for several n and p values in Table A.1. However, we may have data sets with different sizes from the ones of the table in practice. In this case, we should use the d and q values based on the closest n and p values. For the octane data, we select the d and q values for $n = 40$ and $p = 200$. We provide these approximate results in the last column of Table 5.

According to Table 5, $T^2_{Shrinkage} = 790.4292$ and the corresponding $p\ value = 1.718E - 53$. As a result, $T^2_{Shrinkage}$ statistic rejects the null hypothesis, although the center of the majority of the data equals to $\boldsymbol{\mu}_0$. Contrary, the robust Hotelling test statistic based on MRCD estimations gives $T^2_{MRCD} = 1.54\text{E} - 27$ and corresponding $p\ value \cong 1$. Hence, we fail to reject the null hypothesis as would be expected.

Table 5. The test results for octane data.

	$T^2_{Shrinkage}$	T^2_{MRCD} $n = 39,\ p = 226$	T^2_{MRCD} based on Table A.1 $n = 40,\ p = 200$
Statistic Value	790.4292	1.54E-27	1.54E-27
p-value	1.718E-53	1	1
Degree of freedom	$d =$ 199.97	$q =$ 1132.99	$q =$ 1069.06
Constant	$c =$ 1.164	$d =$ 1396.59	$d =$ 1454.80

We can obtain these results by using the RHT2 function in the "MVtests" package as below. The d and q values has been calculated by function simRHT2 in "MVTests" package.

```
> library(rrcov)

> library(MVTests)

> data(octane)

> mu.clean<-colMeans(octane[-c(25,26,36,37,38,39),])

> RHT2(data=octane,mu0=mu.clean,alpha=0.84,d=1396.59, q=1132.99)

$T2

                    [,1]

[1,] 1.539677e-27

$Fval

                    [,1]

[1,] 1.102454e-30

$pval

     [,1]

[1,]   1
```

Moreover, T^2_{MRCD}, based on the d and q values in Table A.1, fails to reject the null hypothesis. According to this, it is apparent the T^2_{MRCD} statistic provides a correct decision without needing Monte Carlo simulation even if the data has outliers. We can obtain this result by using the RHT2 function in the "MVtests" package as follows:

```
> RHT2(data=octane,mu0=mu.clean,alpha = 0.84,d=1454.80,q=1069.06)

$T2

                   [,1]

[1,] 1.539677e-27

$Fval

                   [,1]

[1,] 1.058343e-30

$pval

     [,1]

[1,]   1
```

8. Conclusions

One sample Hotelling T^2 statistic is widely used for multivariate inference. However, this statistic is not helpful in the presence of the outliers or case $n < p$. To obtain robust Hotelling T^2 statistic, Willems et al. (2002) proposed a test statistic based on MCD estimations. However, their test statistic cannot be used in high-dimensional data. On the other hand, Dong et al. (2016) proposed a test statistic based on diagonal shrinkage estimations, but their statistic is not robust. In this study, we suggest a robust one sample Hotelling T^2 statistic called T^2_{MRCD} for high dimensional data. This statistic bases on MRCD estimations. We obtain the approximate distribution for our test statistic using Monte Carlo simulation and show that this approximation is proper. Moreover, we investigate our test statistic's power and robustness and show our test statistic possesses the acceptable loss of power compared to the statistic proposed by (Dong et al. 2016). In contaminated data, T^2_{MRCD} is not sensitive to outliers unlike $T^2_{Shrinkage}$, and T^2_{MRCD} possess better performance than $T^2_{Shrinkage}$.

Finally, we construct two R functions to obtain d and q values for approximate F distribution and to perform our robust Hotelling T^2 test in high dimensional data. In this way, readers can perform our proposed test process. We believe T^2_{MRCD} statistic is a valuable contribution to multivariate inference for high dimensional.

Acknowledgments

We are grateful to the referees for their careful review and constructive comments to improve the article.

ORCID

Hasan Bulut http://orcid.org/0000-0002-6924-9651

References

Bai, Z., and H. Saranadasa. 1996. Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* 6:311–29.

Boudt, K., P. J. Rousseeuw, S. Vanduffel, and T. Verdonck. 2020. The minimum regularized covariance determinant estimator. *Statistics and Computing* 30 (1):113–28.

Bulut, H. 2018. *Multivariate statistical methods with r applications*. Ankara: Nobel Academic Publishing.

Bulut, H. 2019. An R package for multivariate hypothesis tests: MVTests. *E-Journal of New World Sciences Academy* 14 (4):132–8. doi:10.12739/NWSA.2019.14.4.2A0175.

Bulut, H. 2020. Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data. *Communications in Statistics - Theory and Methods* 49 (24):5897–11. doi:10.1080/03610926.2020.1719420.

Chen, L. S., D. Paul, R. L. Prentice, and P. Wang. 2011. A regularized Hotelling's T 2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* 106 (496):1345–60. doi:10.1198/jasa.2011.ap10599.

Chen, S. X., and Y.-L. Qin. 2010. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* 38 (2):808–35. doi:10.1214/09-AOS716.

Croux, C., and G. Haesbroeck. 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 71 (2):161–90. doi:10.1006/jmva.1999.1839.

Davidson, R., and J. G. MacKinnon. 1998. Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School* 66 (1):1–26. doi:10.1111/1467-9957.00086.

Dong, K., H. Pang, T. Tong, and M. G. Genton. 2016. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *Journal of Multivariate Analysis* 143:127–42. doi:10.1016/j.jmva.2015.08.022.

Esbensen, K., S. SchoníKopf, &, and T. Midtgaard. 1995. Multivariate analysis in practice: Training package. *Journal of Chemometrics* 9:521–25.

Hotelling, H. 1992. The generalization of Student's ratio. In *Breakthroughs in statistics*, edited by S. Kotz and N. L. Johnson, 54–65. Berlin; Heidelberg; New York: Springer-Verlag.

Huber, P. J., and E. M. Ronchetti. 2009. *Robust statistics*. Hoboken, NJ: John Wiley & Sons, 2.

Hubert, M., P. J. Rousseeuw, and K. Vanden Branden. 2005. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 47 (1):64–79. doi:10.1198/004017004000000563.

Lopuhaa, H. P., and P. J. Rousseeuw. 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* 19 (1):229–48. doi:10.1214/aos/1176347978.

Park, J., and D. N. Ayyala. 2013. A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference* 143 (5):929–43. doi:10.1016/j.jspi.2012.11.001.

Rencher, A. C. 2003. *Methods of multivariate analysis*. Vol. 492. New York: John Wiley & Sons.

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* 8 (283–297):37.

Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88 (424):1273–83. doi:10.1080/01621459.1993.10476408.

Srivastava, M. S. 2009. A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* 100 (3):518–32. doi:10.1016/j.jmva.2008.06.006.

Srivastava, M. S., and M. Du. 2008. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* 99 (3):386–402. doi:10.1016/j.jmva.2006.11.002.

Srivastava, M. S., S. Katayama, and Y. Kano. 2013. A two sample test in high dimensional data. *Journal of Multivariate Analysis* 114:349–58. doi:10.1016/j.jmva.2012.08.014.

Todorov, V., and P. Filzmoser. 2009. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software* 32 (1):1–47.

Todorov, V., and P. Filzmoser. 2010. Robust statistic for the one-way MANOVA. *Computational Statistics & Data Analysis* 54 (1):37–48. doi:10.1016/j.csda.2009.08.015.

Weiliang, Q., and J. Harry. 2015. clusterGeneration: Random cluster generation (with specified degree of separation (version R package version 1.3.4). https://CRAN.R-project.org/package=clusterGeneration

Willems, G., G. Pison, P. Rousseeuw, and S. Van Aelst. 2002. A robust Hotelling test. *Metrika* 55 (1-2):125–38. doi:10.1007/s001840200192.

Wu, Y., M. G. Genton, and L. A. Stefanski. 2006. A multivariate two-sample mean test for small sample size and missing data . *Biometrics* 62 (3):877–85. doi:10.1111/j.1541-0420.2006.00533.x.

Zhang, J., and J. Xu. 2009. On the k-sample Behrens-Fisher problem for high-dimensional data. *Science in China Series A: Mathematics* 52 (6):1285–304. doi:10.1007/s11425-009-0091-x.

Appendix

A.1: The d and q values for several n and p values.

n	p:	10	15	20	25	30	40	60	80	100	200
10	d:	158.947	226.745	271.064	303.630	330.874	386.112	459.963	533.499	593.550	868.433
	q:	12.571	17.055	23.847	24.799	27.850	48.800	59.488	102.346	115.627	232.665
15	d:	102.591	189.750	268.870	316.095	353.932	422.988	520.872	602.593	667.650	946.987
	q:	10.762	18.753	34.346	38.689	51.595	68.629	116.975	185.802	255.516	606.312
20	d:	80.380	166.818	260.197	332.065	384.544	466.369	590.376	682.490	767.408	1065.641
	q:	11.499	20.692	31.217	42.866	63.380	101.619	183.945	272.924	338.045	1221.523
25	d:	67.119	129.950	211.367	298.734	366.039	475.158	623.154	729.913	821.342	1163.838
	q:	11.267	17.696	26.338	39.309	51.093	91.442	183.052	255.956	357.596	959.868
30	d:	61.056	114.178	189.682	275.037	364.127	486.924	663.663	791.846	888.830	1281.136
	q:	9.668	16.281	22.762	34.390	49.709	77.268	159.458	216.786	352.932	1141.608
40	d:	49.900	90.658	144.170	216.645	295.726	466.312	697.192	855.674	989.819	1454.797
	q:	9.027	13.988	21.028	26.142	39.058	65.753	148.623	223.867	317.832	1069.055
50	d:	41.180	81.710	120.791	175.018	248.760	403.305	693.459	896.750	1058.873	471.954
	q:	9.902	13.647	18.490	24.078	30.954	53.609	105.496	200.513	399.196	97.899
75	d:	34.015	56.525	92.684	131.864	172.359	282.003	581.426	880.058	1105.434	1861.921
	q:	16.481	17.289	16.948	21.083	25.176	38.171	79.753	117.806	195.412	832.946
100	d:	32.717	50.312	73.459	110.120	149.991	226.634	470.834	773.047	1091.167	2028.512
	q:	21.640	22.690	23.879	23.627	24.483	35.412	55.967	102.839	154.716	611.301