Robust Tests for the Equality of Two Normal Means based on the Density Power Divergence

A. Basu · A. Mandal · N. Martin · L. Pardo

September 28, 2014

Abstract Statistical techniques are used in all branches of science to determine the feasibility of quantitative hypotheses. One of the most basic applications of statistical techniques in comparative analysis is the test of equality of two population means, generally performed under the assumption of normality. In medical studies, for example, we often need to compare the effects of two different drugs, treatments or preconditions on the resulting outcome. The most commonly used test in this connection is the two sample $t$-test for the equality of means, performed under the assumption of equality of variances. It is a very useful tool, which is widely used by practitioners of all disciplines and has many optimality properties under the model. However, the test has one major drawback; it is highly sensitive to deviations from the ideal conditions, and may perform miserably under model misspecification and the presence of outliers. In this paper we present a robust test for the two sample hypothesis based on the density power divergence measure (\(\Phi\)), and show that it can be a great alternative to the ordinary two sample $t$-test. The asymptotic properties of the proposed tests are rigorously established in the paper, and their performances are explored through simulations and real data analysis.

AMS 2001 Subject Classification: 62F35, 62F03.

keywords and phrases: Robustness, Density Power Divergence, Hypothesis Testing.

1 Introduction: Motivation and Background

In many scientific studies, often the main problem of interest is to compare different population groups. In medical studies, for example, the primary research problem could be to test for the difference between the location parameters of two different populations receiving two different drugs, treatments or therapy, or having two different preconditions. The normal distribution often provides the basic setup for statistical analyses in medical studies (as well as in other disciplines). Inference procedures based on the sample mean, the standard deviation and the one and two-sample $t$-tests are often the default techniques for the scenarios where they are applicable. In particular, the two
sample \(t\)-test is the most popular technique in testing for the equality of two means, performed under the assumption of equality of variances. Its applicability in real life situations is, however, tempered by the known lack of robustness of this test against model perturbations. Even a small deviation from the ideal conditions can make the test completely meaningless and lead to nonsensical results. This problem is caused by the fact that the \(t\)-test is based on the classical estimates of the location and scale parameters (the sample mean and the sample standard deviation). Large outliers tend to distort the mean and inflate the standard deviation. This may lead to false results of both types, i.e. detecting a difference when there isn’t one, and failing to detect a true significance.

In this paper we are going to develop a class of robust tests for the two sample problem which evolves from an appropriate minimum distance technique in a natural way. This class of tests is indexed by two real parameters \(\beta\) and \(\gamma\), and we will constrain each of these parameters to lie within the \([0, 1]\) interval. Our general minimum distance approach will allow us to study the likelihood ratio test in an asymptotic sense, as the likelihood ratio test is asymptotically equivalent to the test generated by the parameters \(\beta = \gamma = 0\). Normally we will work with the one parameter family of test statistics corresponding to \(\beta = \gamma\); the outlier stability of the proposed tests increase with the tuning parameter \(\gamma\).

Let \(X\) and \(Y\) be independent random variables whose distributions are modeled as normals having unknown means \(\mu_1\) and \(\mu_2\), respectively, with an unknown but common variance \(\sigma^2\). We are interested in testing the null hypothesis \(H_0 : \mu_1 = \mu_2\) against \(H_1 : \mu_1 \neq \mu_2\), (1) under the above set up. It is well known that the exact two sample \(t\)-test (which is equivalent to the likelihood ratio test) rejects the null hypothesis in (1) if and only if

\[
t = \frac{|\bar{X} - \bar{Y}|}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2),
\]

where \(\bar{X}\) and \(\bar{Y}\) are the sample means corresponding to the random samples \(X_1, X_2, \ldots, X_{n_1}\) and \(Y_1, Y_2, \ldots, Y_{n_2}\) obtained from the two distributions,

\[
S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},
\]

\[
S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2,
\]

and \(t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)\) is the 100\((1 - \frac{\alpha}{2})\)-th quantile of the \(t\)-distribution with \(n_1 + n_2 - 2\) degrees of freedom. The \(t\)-test is the uniformly most powerful unbiased and invariant test for this hypothesis. Testing the equality of means of independent normal populations with unknown variances which are not necessarily equal, is referred to as the Behrens-Fisher problem.

In this paper we will use the density power divergence (DPD) measure (\(?\)), which provides a natural robustness option for many standard inference problems. The density power divergence and its variants have been successfully used by many authors in a variety of inference problems; see, eg. \(?\), \(?\), \(?\), \(?\). However, the two sample problem requires a non-trivial extension of the currently existing techniques. Our purpose in this paper is to derive the asymptotic properties of the class of two sample tests based on the density power divergence and demonstrate their robust behavior in practical situations.

**Example 1 (Cloth Manufacturing data):** In order to emphasize the need for applications early, we now present a motivational example. This example illustrates the use of quality control methods practiced in a clothing manufacturing plant. Levi-Strauss manufactures clothing from cloth supplied
Table 1 Cloth Manufacturing data.

| Mill   | 0.12 | 1.01 | –0.20 | 0.15 | –0.30 | –0.07 | 0.32 | 0.27 | –0.32 | –0.17 | 0.24 |
|--------|------|------|--------|------|--------|--------|------|------|--------|--------|------|
|        | 0.03 | 0.35 | –0.08  | 2.94 | 0.28   | 1.30   | 4.27 | 0.14 | 0.30   | 0.24   | 0.13 |

| Mill   | 1.64 | –0.60 | –1.16 | –0.13 | 0.40   | 1.70   | 0.38 | 0.43 | 1.04   | 0.42   | 0.85 |
|--------|------|-------|--------|--------|--------|--------|------|------|--------|--------|------|
|        | 0.63 | 0.90  | 0.71   | 0.43   | 1.97   | 0.30   | 0.76 | 7.02 | 0.85   | 0.60   | 0.29 |

Fig. 1 The $p$-values of the DPD tests for the Cloth Manufacturing data for different values of $\gamma$. The solid line represents the full data analysis, while the dashed line represents the outlier deleted case.

by several mills. The data used in this example (see Table 1) are for two of these mills and were obtained from the quality control department of the Levi plant in Albuquerque, New Mexico (?, p. 86). In order to maintain the anonymity of these two mills we have coded them $A$ and $B$. A measure of wastage due to defects in cloth and so on is called run-up. It is quoted as percentage of wastage per week and is measured relative to computerized layouts of patterns on the cloth. Since the people working in the plant can often beat the computer in reducing wastage by laying out the patterns by hand, it is possible for run-up to be negative. From the viewpoint of quality control, it is desirable not only that the run-up be small but that the quality from week to week be fairly consistent. There are 22 measurements on run-up for each of the two mills and they are presented in Table 1. The $t$-test for the equality of the two means against the two-sided alternative has a $p$-value of 0.3428 and fails to reject the null hypothesis; however, when the presumed outliers (presented in bold fonts in Table 1) are removed from the dataset, the same two-sample $t$-test produces a $p$-value of 0.0308, leading to clear rejection. Choosing $\beta = \gamma$ to be the only parameter, the $p$-values of the DPD tests (to be developed in the next section) for testing the same hypotheses are presented in Figure 1 as a function of $\gamma$. It is observed that the $p$-values of the tests with the full data and those with the outlier deleted data are practically identical for $\gamma = 0.2$ or larger, and lead to solid rejection. Thus, while the outliers mask the significance in case of the two sample $t$-test, the more robust DPD tests are able to capture the same.

Our primary motivation for studying the alternatives of the two sample $t$-test has been the need for developing such a test in the context of examples relating to medical data. However, examples abound in practically all scientific disciplines showing that this is a real necessity which is certainly not restricted to the medical field. The example considered above is one such, where the context does not have anything directly to do with a medical problem, but the importance of the problem and the need for a robust solution can immediately be appreciated.
The rest of the paper is organized as follows: In Section 2 the asymptotic distribution of the minimum DPD estimators in the two sample situation is described. In Section 3 we introduce our robust two sample test statistic and develop the necessary theory. A large number of real data examples and extensive simulation results are presented in Section 4. Finally Section 5 has some concluding remarks.

2 The Minimum DPD Estimator: Asymptotic Distribution

For any two probability density functions \( f \) and \( g \), the density power divergence measure is defined, as the function of a single tuning parameter \( \beta \geq 0 \), as

\[
d_{\beta}(g, f) = \begin{cases} 
\int f^{1+\beta}(x) - \left(1 + \frac{1}{\beta}\right) f^{\beta}(x)g(x) + \frac{1}{\beta} g^{1+\beta}(x) \, dx, & \text{for } \beta > 0, \\
\int g(x) \log \left( \frac{g(x)}{f(x)} \right) \, dx, & \text{for } \beta = 0.
\end{cases}
\]

Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from a \( N(\mu, \sigma^2) \) distribution, where both parameters are unknown. Let \( f_{\mu, \sigma}(x) \) represent the density function of a \( N(\mu, \sigma^2) \) variable. For a given \( \beta \), we get the minimum density power divergence estimators (MDPDEs) \( \tilde{\mu}_\beta \) and \( \tilde{\sigma}_\beta \) of \( \mu \) and \( \sigma \) by minimizing the following function over \( \mu \) and \( \sigma \)

\[
\int f_{\mu, \sigma}^{1+\beta}(x) \, dx - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_{\mu, \sigma}^\beta(X_i), \quad \text{for } \beta > 0,
\]

and

\[-\frac{1}{n} \sum_{i=1}^{n} \log f_{\mu, \sigma}(X_i), \quad \text{for } \beta = 0.
\]

For \( \beta = 0 \), the objective function in (4) is the negative of the usual log likelihood and has the classical maximum likelihood estimator as the minimizer. For a normal density the function in (3) simplifies to

\[
h_{n,\beta}(\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \left( \frac{1}{(1 + \beta)^{3/2}} - \frac{1}{n\beta} \sum_{i=1}^{n} \exp \left( -\frac{1}{2} \frac{(X_i - \mu)^2}{\sigma} \right) \beta \right).
\]

In order to get \( \tilde{\mu}_\beta \) and \( \tilde{\sigma}_\beta \), we have to solve the estimating equation

\[
h'_{n,\beta}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta) = \begin{pmatrix} 1h'_{n,\beta}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta) \\ 2h'_{n,\beta}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta) \end{pmatrix} = 0_2,
\]

where

\[
h'_{n,\beta}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta) = \frac{\partial h_{n,\beta}(\mu, \sigma)}{\partial \mu} \bigg|_{\mu=\tilde{\mu}_\beta}, \quad 2h'_{n,\beta}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta) = \frac{\partial h_{n,\beta}(\mu, \sigma)}{\partial \sigma} \bigg|_{\sigma=\tilde{\sigma}_\beta},
\]

and \( 0_2 \) represents a zero vector of length 2. We denote

\[
H_{n,\beta}(\mu_0, \sigma_0) = \begin{pmatrix} 11h''_{n,\beta}(\mu_0, \sigma_0) & 12h''_{n,\beta}(\mu_0, \sigma_0) \\ 21h''_{n,\beta}(\mu_0, \sigma_0) & 22h''_{n,\beta}(\mu_0, \sigma_0) \end{pmatrix},
\]

where

\[
11h''_{n,\beta}(\mu_0, \sigma_0) = \frac{\partial^2 h_{n,\beta}(\mu, \sigma)}{\partial \mu^2} \bigg|_{\mu=\mu_0}, \quad 12h''_{n,\beta}(\mu_0, \sigma_0) = \frac{\partial^2 h_{n,\beta}(\mu, \sigma)}{\partial \mu \partial \sigma} \bigg|_{\mu=\mu_0, \sigma=\sigma_0},
\]

\[
21h''_{n,\beta}(\mu_0, \sigma_0) = \frac{\partial^2 h_{n,\beta}(\mu, \sigma)}{\partial \sigma^2} \bigg|_{\sigma=\sigma_0}.
\]
Using a Taylor series expansion of the function in equation (5), it is easy to show that

$$\sqrt{n} \left( \frac{\hat{\beta}}{\hat{\sigma}^2} - \frac{\mu_0}{\sigma_0} \right) = \sqrt{n} H_{n,\beta}(\mu_0, \sigma_0) h_{n,\beta}(\mu_0, \sigma_0) + o_p(1)$$

$$= \sqrt{n} J_{\beta}(\sigma_0) h_{n,\beta}^T(\mu_0, \sigma_0) + o_p(1),$$

where

$$J_{\beta}(\sigma_0) = \lim_{n \to \infty} H_{n,\beta}(\mu_0, \sigma_0) = \frac{1}{\sqrt{1 + \beta^2 (2\pi)^{3/2} \sigma_0^{2+\beta}}} \begin{pmatrix} 1 & 0 \\ \frac{\beta^2 + 1}{(1+\beta)^2} & 0 \end{pmatrix}.$$  

The joint distribution of $\hat{\beta}$ and $\hat{\sigma}_0$ then follows (see ?) from the result that

$$\sqrt{n} h_{n,\beta}^T(\mu_0, \sigma_0) \xrightarrow{c} N \left( 0_2, K_{\beta}(\sigma_0) \right),$$

where

$$K_{\beta}(\sigma_0) = (K_{ij,\beta}(\sigma_0))_{i,j=1,2} = \frac{1}{\sigma_0^{2+2\beta}} \left( \frac{1}{(1 + \beta^2)^{3/2}} \right) \begin{pmatrix} 1 & 0 \\ 0 & \frac{\beta^2 + 1}{(1+\beta)^2} \end{pmatrix}.$$  

We will use the above results to obtain the MDPDEs of the parameters in the two sample setup mentioned below.

Suppose $X_1, X_2, \ldots, X_{n_1}$ is a random sample of size $n_1$ from $X$ which has a $N(\mu_1, \sigma^2)$ distribution, and $Y_1, Y_2, \ldots, Y_{n_2}$ is a random sample of size $n_2$ from $Y$ which has a $N(\mu_2, \sigma^2)$ distribution; all three parameters are unknown. Let $f_{\mu_1, \sigma}(x)$ and $f_{\mu_2, \sigma}(y)$ be the density functions of $X$ and $Y$ respectively. Let us denote the set of unknown parameters by $\eta = (\mu_1, \mu_2, \sigma)^T$. The MDPDE of $\eta$, denoted by $\hat{\eta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})^T$, is obtained by minimizing the following function

$$h_{n_1,n_2,\beta}(\eta) = \frac{1}{n_1 + n_2} \left( n_1 h_{n_1,\beta}(\mu_1, \sigma) + n_2 h_{n_2,\beta}(\mu_2, \sigma) \right).$$

It may be noticed that $\hat{\mu}_1$ is based only on the first term of the above function, and similarly $\hat{\mu}_2$ depends only on the second term. Therefore, the estimating equations are given by $h_{n_1,\beta}(\mu_1, \sigma) = 0$, $i = 1, 2$, and $2h_{n_2,\beta}(\eta) = 0$, where

$$2h_{n_1,n_2,\beta}(\eta) = \frac{\partial h_{n_1,n_2,\beta}(\eta)}{\partial \sigma} = \frac{1}{n_1 + n_2} \left( n_1 2h_{n_1,\beta}(\mu_1, \sigma) + n_2 2h_{n_2,\beta}(\mu_2, \sigma) \right).$$

For $\beta = 0$, the above equations can be explicitly solved to get the MDPDEs for this case. It is easily seen that $\hat{\mu} = X$ and $\hat{\mu} = Y$. Moreover, using equation (4) we get from (11)

$$h_{n_1,n_2,\beta=0}(\hat{\eta}_0) = -\frac{1}{n_1 + n_2} \left( n_1 \log \prod_{i=1}^{n_1} f_{\hat{\mu}_{10}, \sigma_0}(X_i) + n_2 \log \prod_{i=1}^{n_2} f_{\hat{\mu}_{20}, \sigma_0}(Y_i) \right)$$

$$= \frac{1}{n_1 + n_2} \left( (n_1 + n_2) \log \hat{\sigma}_0 + \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{2\hat{\sigma}_0^2} + \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{2\hat{\sigma}_0^2} + (n_1 + n_2) \log \sqrt{2\pi} \right).$$

So,

$$2h_{n_1,n_2,\beta}(\hat{\eta}_0) = \frac{1}{\hat{\sigma}_0} - \frac{1}{\hat{\sigma}_0^2(n_1 + n_2)} \left( (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \right),$$

which leads to the solution

$$\hat{\sigma}_0 = \left( \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2} \right)^{\frac{1}{2}}.$$

Therefore, for $\beta = 0$ the MDPDEs turn out to be the MLEs of the corresponding parameters. The following theorem gives the asymptotic distribution of the MDPDE of $\eta$ for a given $\beta$. 
Theorem 1 We consider two normal populations with unknown means $\mu_1$ and $\mu_2$ and unknown but common variance $\sigma^2$. Let

$$w = \lim_{n_1, n_2 \to \infty} \frac{n_1}{n_1 + n_2}$$

be the limiting proportion of observations from the first population in the whole sample. We assume that $w \in (0, 1)$. Then, the minimum density power divergence estimator $\hat{\eta}_β$ of $\eta$ has the asymptotic distribution given by

$$\sqrt{n_1 n_2} (\hat{\eta}_β - \eta_0) \overset{D}{\to} \mathcal{N} (0, \Sigma_{w, \beta}(\sigma_0)),$$

where $\eta_0 = (\mu_{10}, \mu_{20}, \sigma_1)^T$ is the true value of $\eta$, and

$$\Sigma_{w, \beta}(\sigma_0) = \sigma_0^2 \begin{pmatrix}
(1 - w) \frac{(\beta + 1)^3}{(2\beta + 1)^2} & 0 & 0 \\
0 & w \frac{(\beta + 1)^3}{(2\beta + 1)^2} & 0 \\
0 & 0 & w (1 - w) \frac{(\beta + 1)^2}{(2\beta + 1)^2} \end{pmatrix}.$$ \hspace{1cm} (15)

Proof See Appendix.

3 The Asymptotic Distribution of the DPD Test Statistic

Let $f_{\mu_1, \sigma_1}(x)$ and $f_{\mu_2, \sigma_2}(y)$ be the density functions of $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ respectively. The density power divergence measure between the densities of $X$ and $Y$, for $\gamma > 0$, is given by

$$d_\gamma(f_{\mu_1, \sigma_1}, f_{\mu_2, \sigma_2}) = \frac{1}{\sigma_2^2 \sqrt{1 + \gamma(2\pi)^{\gamma/2}}} + \frac{1}{\gamma \sigma_1^2 \sqrt{1 + \gamma(2\pi)^{\gamma/2}}}$$

$$- \frac{\gamma \sigma_2^{-1}(\gamma \sigma_1^2 + \sigma_2^2)^{1/2}}{(2\pi)^{\gamma/2}}$$

$$\times \exp \left\{ \frac{1}{2} \left[ - \left( \frac{\mu_2}{\sigma_2^2} \right)^2 + \frac{\mu_1^2}{\sigma_1^2} \right] + \left( \frac{\sigma_2^2 + \mu_1 \left( \frac{\mu_2}{\sigma_2^2} \right)^2}{\sigma_1^2 + \left( \frac{\mu_2}{\sigma_2^2} \right)^2} \right)^2 \right\},$$

and for $\gamma = 0$

$$d_\gamma(f_{\mu_1, \sigma_1}, f_{\mu_2, \sigma_2}) = \log \frac{\sigma_2}{\sigma_1} = \frac{1}{2} + \frac{\sigma_2^2}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}(\mu_1 - \mu_2)^2. $$

To test the null hypothesis given in (1), under the assumption that $\sigma_1 = \sigma_2 = \sigma$, we will consider the divergence between the two normal populations with the estimated parameters; this yields

$$d_\gamma(f_{\hat{\mu}_1, \hat{\sigma}}, f_{\hat{\mu}_2, \hat{\sigma}}) = \begin{cases}
\frac{\sqrt{1 + \gamma}}{\gamma \sqrt{2\pi \hat{\sigma}^2}} \left[ 1 - \exp \left\{ - \frac{\gamma}{2(\gamma + 1)} \left( \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}} \right)^2 \right\} \right], & \text{for } \gamma > 0, \\
\frac{1}{2} \left( \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}} \right)^2, & \text{for } \gamma = 0.
\end{cases} \hspace{1cm} (17)$$

Naturally, we will reject the null hypothesis for large values of $d_\gamma(f_{\hat{\mu}_1, \hat{\sigma}}, f_{\hat{\mu}_2, \hat{\sigma}})$. To propose the test in a very general setup we have considered two possibly distinct tuning parameters $\gamma$ and $\beta$ in the above expression; the parameter $\gamma$ represents the tuning parameters of the divergence, and the parameter $\beta$ represents the tuning parameter of the MDPDEs. In order to determine the critical region of this test we will find (later in Theorem 3) the asymptotic null distribution of the test statistic based on (17), standardized with a suitable scaling constant involving $n_1$ and $n_2$. 
Theorem 2 For \( \gamma > 0 \), let us define \( t_\gamma (\eta) = (t_{\gamma,1}(\eta), t_{\gamma,2}(\eta), t_{\gamma,3}(\eta))^T \), with
\[
t_{\gamma,1}(\eta) = \frac{\mu_1 - \mu_2}{\sqrt{1 + \gamma} (\sqrt{2\pi})^{\gamma/2} \sigma^{\gamma+1}} \exp \left\{ -\frac{1}{2 (\gamma+1)} \left( \frac{\mu_1 - \mu_2}{\sigma} \right)^2 \right\},
\]
\[
t_{\gamma,2}(\eta) = -t_1(\eta),
\]
\[
t_{\gamma,3}(\eta) = -\frac{(\gamma+1)}{(\sqrt{2\pi})^{\gamma+1}} \left[ 1 - \left( \frac{1}{\gamma+1} \left( \frac{\mu_1 - \mu_2}{\sigma} \right)^2 \right) \exp \left\{ -\frac{1}{2 (\gamma+1)} \left( \frac{\mu_1 - \mu_2}{\sigma} \right)^2 \right\} \right].
\]
Then, for \( w \in (0,1) \) as defined in (14) we have
\[
\sqrt{\frac{n_1 n_2}{n_1 + n_2}} (d_\gamma(f_{\tilde{\mu}_{12},\tilde{\sigma}_\beta}, f_{\tilde{\mu}_{22},\tilde{\sigma}_\beta}) - d_\gamma(f_{\mu_{10},\sigma_0}, f_{\mu_{20},\sigma_0})) \xrightarrow{n_1,n_2 \to \infty} N(0, \sigma_\gamma^2),
\]
where
\[
\sigma_\gamma^2 = t_\gamma^T(\eta_0) \Sigma_{w,\beta}(\sigma_0) t_\gamma(\eta_0),
\]
and \( \Sigma_{w,\beta}(\sigma_0) \) is given in (16).

Proof See Appendix.

Notice that \( t_\gamma^T(\eta_0) \Sigma_{w,\beta}(\sigma_0) t_\gamma(\eta_0) \geq 0 \). If \( \mu_{10} \neq \mu_{20} \), we observe that \( t_\gamma(\eta_0) \neq 0 \), and since \( \Sigma_{w,\beta}(\sigma_0) \) is positive definite matrix, we have \( t_\gamma^T(\eta_0) \Sigma_{w,\beta}(\sigma_0) t_\gamma(\eta_0) > 0 \). But for \( \mu_{10} = \mu_{20} \), \( t_\gamma(\eta_0) = 0 \), and hence \( t_\gamma^T(\eta_0) \Sigma_{w,\beta}(\sigma_0) t_\gamma(\eta_0) = 0 \). Therefore, to get the asymptotic distribution of the test statistic under the null hypothesis we need a higher order scaling involving \( n_1 \) and \( n_2 \) to the quantity given in (17).

Theorem 3 Let \( w \in (0,1) \) as defined in (14) and \( \gamma > 0 \). Then, under the null hypothesis, we have
\[
S_\gamma(\tilde{\mu}_{12}, \tilde{\mu}_{22}, \tilde{\sigma}_\beta) = \frac{2 n_1 n_2}{n_1 + n_2} \frac{d_\gamma(f_{\tilde{\mu}_{12},\tilde{\sigma}_\beta}, f_{\tilde{\mu}_{22},\tilde{\sigma}_\beta})}{\lambda_{\beta,\gamma}(\tilde{\sigma}_\beta)} \xrightarrow{n_1,n_2 \to \infty} \chi^2(1),
\]
where
\[
\lambda_{\beta,\gamma}(\tilde{\sigma}_\beta) = \frac{(\beta + 1)^3 (2\beta + 1)^{-\frac{3}{2}}}{\tilde{\sigma}_\beta^2 (2\pi)^{\frac{3}{2}} (\gamma + 1)^{\frac{3}{2}}},
\]

Proof See Appendix.

The above result indicates that the density power divergence test for the hypothesis in (1) can be based on the statistic \( S_\gamma(\tilde{\mu}_{12}, \tilde{\mu}_{22}, \tilde{\sigma}_\beta) \), where the critical region corresponding to significance level \( \alpha \) is given by the set of points satisfying
\[
S_\gamma(\tilde{\mu}_{12}, \tilde{\mu}_{22}, \tilde{\sigma}_\beta) > \chi^2_{\alpha}(1).
\]

Using the result of Theorem 2 we can get an approximation of the power function of the test statistic. We consider \( \mu_{10} \neq \mu_{20} \). In the following we will let \( \lambda \) denote the quantity defined in equation (24) to keep the notation simple. The power function is then given by
\[
\eta_{\gamma,\beta}(\mu_{10}, \mu_{20}, \sigma_0) = P(S_\gamma(\tilde{\mu}_{12}, \tilde{\mu}_{22}, \tilde{\sigma}_\beta) > \chi^2_{\alpha}(1))
\]
\[
= P \left( \frac{2}{\lambda n_1 + n_2} d_\gamma(f_{\tilde{\mu}_{12},\tilde{\sigma}_\beta}, f_{\tilde{\mu}_{22},\tilde{\sigma}_\beta}) > \chi^2_{\alpha}(1) \right)
\]
\[
= P \left( \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (d_\gamma(f_{\tilde{\mu}_{12},\tilde{\sigma}_\beta}, f_{\tilde{\mu}_{22},\tilde{\sigma}_\beta}) - d_\gamma(f_{\mu_{10},\sigma_0}, f_{\mu_{20},\sigma_0})) > \frac{\lambda}{2} \sqrt{\frac{n_1 + n_2}{n_2}} \left( \chi^2_{\alpha}(1) - \frac{2 n_1 n_2}{n_1 + n_2} d_\gamma(f_{\mu_{10},\sigma_0}, f_{\mu_{20},\sigma_0}) \right) \right)
\]
\[
= 1 - \Phi_n \left( \frac{\lambda}{2 \sigma_\gamma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left( \chi^2_{\alpha}(1) - \frac{2 n_1 n_2}{n_1 + n_2} d_\gamma(f_{\mu_{10},\sigma_0}, f_{\mu_{20},\sigma_0}) \right) \right),
\]
where $\Phi_n$ is a sequence of distributions functions tending uniformly to the standard normal distribution function $\Phi$, and $\sigma_\gamma$ is defined in (22). We observe that if $\mu_{10} \neq \mu_{20}$
\[
\lim_{n_1,n_2 \to \infty} n_1 \beta n_2 (\mu_{10}, \mu_{20}, \sigma_0) = 1. 
\] (25)
Therefore, the test is consistent in the Fraser’s sense (?).

**Corollary 1** Let $w \in (0, 1)$ as defined in (14) and $\gamma = \beta = 0$. Then, under the null hypothesis defined in (1), we have
\[
S_0(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0) = \frac{n_1 n_2 (\bar{X} - \bar{Y})^2}{\sigma_0^2} \xrightarrow{n_1,n_2 \to \infty} \chi^2(1). 
\] (26)

The proof of the corollary is straightforward. The test statistic given in the above corollary is closely related to the likelihood ratio test. This correspondence is described in the next corollary.

**Corollary 2** For a given sample the value of the test statistic $S_0(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0)$, defined in (26), does not exactly match the value of the likelihood ratio test statistic
\[
-2 \log A(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0) = (n_1 + n_2) \log \left(1 + \frac{n_1 n_2 (\bar{X} - \bar{Y})^2}{(n_1 + n_2)^2} \right),
\]
where $\hat{\sigma}_0^2$ is defined in (13). However, as $n_1,n_2 \to \infty$, and $w \in (0, 1)$ as defined in (14), both test statistics are asymptotically equivalent.

**Proof** Let us denote $\Theta_0 = \{(\mu, \mu, \sigma)^T : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$, $\Theta = \{(\mu_1, \mu_2, \sigma)^T : \mu_1, \mu_2 \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$. The likelihood function is given by
\[
\mathcal{L}(\mu_1, \mu_2, \sigma) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} f_{\mu_1, \sigma}(X_i) f_{\mu_2, \sigma}(Y_j).
\]

It can be shown that
\[
A(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0) = \sup_{(\mu_1, \mu_2, \sigma) \in \Theta_0} \mathcal{L}(\mu_1, \mu_2, \sigma) = \left( \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2} \right)^{-1}.
\]
where $\hat{\mu} = \frac{n_1}{n_1 + n_2} \bar{X} + \frac{n_2}{n_1 + n_2} \bar{Y}$. Therefore, asymptotically, the likelihood ratio test rejects the null hypothesis $H_0$ if
\[
-2 \log A(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0) = (n_1 + n_2) \log \left( \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2} \right) > \chi^2(1).
\]

Now
\[
\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2 + n_1 (\bar{X} - \hat{\mu})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 + n_2 (\bar{Y} - \hat{\mu})^2 \\
= \sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{X} - \bar{Y})^2.
\]
So

\[-2 \log A(\mu_{10}, \mu_{20}, \sigma_0) = (n_1 + n_2) \log \left( \frac{\sum_{i=1}^{n_1} (X_i - \bar{\mu})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{\mu})^2}{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2} \right) \]

\[= (n_1 + n_2) \log \left( 1 + \frac{n_1 n_2}{(n_1 + n_2)^2} \frac{(\bar{X} - \bar{Y})^2}{\sigma_0^2} \right) \]

\[= \frac{n_1 n_2}{(n_1 + n_2)} (\bar{X} - \bar{Y})^2 / \sigma_0^2 + R_{n_1, n_2}, \]

where \(R_{n_1, n_2} \to 0\) in probability as \(n_1, n_2 \to \infty\) and \(w \in (0, 1)\). Thus, the test statistics \(-2 \log A(\mu_{10}, \mu_{20}, \sigma_0)\) and \(S_0(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0)\) are asymptotically equivalent.

4 Numerical Studies

4.1 Simulation Study

In this section we study the performance of our proposed test statistics through simulated data. We have generated two random samples \(X_1, X_2, \ldots, X_{n_1}\) and \(Y_1, Y_2, \ldots, Y_{n_2}\) from \(N(\mu_1, \sigma_2)\) and \(N(\mu_2, \sigma_2)\) respectively; thus the total sample size is \(n = n_1 + n_2\). The value of \(w\) in (14) is taken to be 0.6, and the sample size from the first population is \(n_1 = \lfloor wn \rfloor + 1\), where \(\lfloor x \rfloor\) denotes the integer part of \(x\). Our aim is to test the null hypothesis given in (1). We have taken \(\sigma^2 = 1\) in this study.

We have compared the results of the ordinary two sample \(t\)-test and the density power divergence tests with four different values of the tuning parameter \(\gamma = \beta = 0, 0.05, 0.1\) and 0.15; let \(DPD(\gamma)\) represent the DPD test with tuning parameter \(\gamma\). The nominal level of the tests are 0.05, and all tests are replicated 1,000 times.

In the first case we have taken \(\mu_1 = \mu_2 = 0\). Plot (a) in Figure 2 shows the observed levels of the five test statistics for different values of the sample size (obtained as the proportion of test statistics, in the 1000 replications, that exceed the nominal \(\chi^2\) critical value at 5% level of significance). It is seen that the observed levels of the \(t\)-test are very close to the nominal level. On the other hand, the DPD tests are slightly liberal for very small sample sizes and lead to somewhat inflated observed levels. However, as the sample size increases the levels settle down rapidly around the nominal level.

Next, we have generated data with \(\mu_1 = 0\) but \(\mu_2 = 1\). The observed power of the tests are presented in plot (b) of Figure 2. There is not much difference among the observed powers in this plot. The DPD tests have slightly higher power than the \(t\)-test in very small sample sizes. This, however, must be a consequence of the fact that the observed levels of these tests are higher than the nominal level (and higher than the observed level of the \(t\)-statistic) in small samples.

Now we check the performance of the tests under contaminated data. So, we have generated \(n_2\) observations \(Y_1, Y_2, \ldots, Y_{n_2}\) from \(0.95N(\mu_2, 1) + 0.05N(-10, 1)\), whereas the \(n_1\) observations representing the first population come from the pure \(N(\mu_1, 1)\) distribution. To evaluate the stability of the level of the tests for testing the hypothesis in (1), we have taken \(\mu_1 = \mu_2 = 0\). Figure 2 (c) presents the levels for different values of the sample sizes. It may be observed that there is a drastic inflation in the levels for the \(t\)-test and DPD(0) test statistic, but the levels of the other DPD test statistics remain stable.

Figure 2 (d) shows the power of the tests under the contaminated setup considered in the previous paragraph, when \(\mu_1 = 0\) and \(\mu_2 = 1\). Here, the presence of the outliers lead to a sharp drop in power for the \(t\)-test and the DPD(0) test. On the other hand, the other tests are clearly more resistant, and hold their power much better as \(\gamma\) increases.

On the whole, therefore, it appears that in comparison to the \(t\)-test, many of our DPD tests are quite competitive in performance when the data come from the pure model. Under contaminated
data, however, the robustness properties of the DPD tests appear to be far superior, and they do much better at maintaining the stability of the level and the power in such cases.

4.2 Comparison with Other Robust Tests

In this section we compare the DPD test with some other popular robust tests. For comparison we have used a parametric test – the two sample trimmed \( t \)-test proposed by ?, as well as two non-parametric tests – the Kolmogorov-Smirnov test (KS-test) and the Wilcoxon two-sample test (which
Table 2 Lead Measurement data.

|     | First Lake |   | Second Lake |   |
|-----|------------|---|-------------|---|
|     | −1.48      | 0.93| 1.32        | 2.86|
|     | 1.25       | 0.51| 1.81        | 0.47|
|     | −0.51      | 1.11| −0.54       | −0.42|
|     | 0.46       | −0.17| 2.68       | 0.16|
|     | 0.60       | −0.79| 2.27       | 0.69|
|     | −4.27      | −1.02| 2.70       | 0.78|
|     | 0.63       | −0.91| 0.78       | 1.72|
|     | −0.14      | 0.10| 1.88       | 1.57|
|     | −0.38      | 0.41| 0.86       | 2.14|
|     | 1.28       | 1.11| 1.62       |     |

is also known as the Mann-Whitney $U$-test). For the two sample trimmed $t$-test we have trimmed 20% extreme observations from each of the data sets of $X$ and $Y$. The set up, the parameters taken for the simulation and the level of contamination are exactly the same as in the previous section. For comparison we have used only one DPD test in this case, that corresponding to tuning parameter 0.1. To emphasize the robustness properties of these tests we have also included the two sample $t$-test in this investigation. The results are presented in Figure 3.

Figure 3 (a) shows that the observed levels of all the robust tests are very close to the nominal level of 0.05 for the pure normal data. The same result is observed in Figure 3 (c) for the contaminated data. On the other hand, if we consider the observed power of the tests the DPD test is much more powerful than the other tests. Specifically, for the contaminated data, the DPD test does significantly better than the others in holding on to its power. Therefore, on the whole, the DPD tests are not only superior to the two sample $t$-test under contamination, but they also appear to be competitive or better than the other popular robust tests as far as this simulation study is concerned.

4.3 Real Data Examples

**Example 2 (Lead Measurement data):** In Table 2 the lead measurement data (7, p. 280) are presented. The numbers represent the values of $10(x − 2)$, $x$ being the level of lead in the water samples from two lakes at randomly chosen locations. To test whether the average pollution levels of the two lakes are equal, we perform tests for equality of the means of the populations represented by the two different samples. The $p$-values of the DPD tests are plotted in Figure 4; the solid line represents the $p$-values for the full data, while the dashed line represents for the $p$-values for the outlier deleted data. The less robust tests (corresponding to very small values of $\gamma$) register only borderline significance under full data, and for very small values of $\gamma$ the tests would fail to reject the equality hypothesis at the 1% significance level. However, for all value of $\gamma$, the tests would soundly reject the null hypothesis when the obvious outliers (displayed with bold fonts in Table 2) are removed from the dataset. For higher values of $\gamma$ (0.2 or larger), the $p$-values with or without the outliers are practically identical, demonstrating that the outliers have little effect in such cases. The $p$-values for the two-sample $t$-test with and without the outliers are 0.02397 and 0.0004 respectively. As in Example 1, the presence of the outliers masks the significance of the two-sample $t$-test and the small $\gamma$ DPD tests, but the large $\gamma$ DPD tests successfully discount the effect of the outliers.

**Example 3 (Ozone Control data):** 7 report data from a study design to assess the effects of ozone on weight gain in rats. The experimental group consisted of 22 rats, each 70-day old kept in an ozone environment for 7 days. A control group of 23 rats, of the same age, were kept in an ozone-free environment. The weight gains, in grams, are listed in Table 3. We want to test for the equality of the means of the two groups. The $p$-values of the DPD tests are plotted in Figure 5. The $p$-values of the two-sample $t$-test for the full data and the outlier deleted data are 0.0168 and $3.4721 \times 10^{-6}$ respectively. The conclusions of this example are similar to those of Examples 1 and 2.

**Example 4 (Newcomb’s Light Speed data):** In 1882 Simon Newcomb, an astronomer and mathematician, measured the time required for a light signal to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back. The total distance
Fig. 3 (a) Simulated levels of different tests for pure data; (b) simulated power of different tests for pure data; (c) simulated levels of different tests for contaminated data; (d) simulated power of different tests for contaminated data. 

Table 3 Ozone Control data

| X  | 41.0 | 38.4 | 24.4 | 25.9 | 21.9 | 18.3 | 13.1 | 27.3 | 28.5 | -16.9 | 26.0 | 17.4 |
|----|------|------|------|------|------|------|------|------|------|-------|------|------|
|    | 21.8 | 15.4 | 27.4 | 19.2 | 22.4 | 17.7 | 26   | 29.4 | 21.4 | 26.6  | 22.7 |
| T  | 10.1 | 6.1  | 20.4 | 7.3  | 14.3 | 15.5 | -9.9 | 6.8  | 28.2 | 17.9  | -9.0 | -12.9 |
|    | 14.0 | 6.6  | 12.1 | 15.7 | 39.9 | -15.9| 54.6 | -14.7| 44.1 | -9.0  |      |      |
was 7443.73 meters. Table 4 contains these measurements from three samples, as deviations from 24,800 nanoseconds. For example, for the first observation, 28, means that the time taken for the light to travel the required 7443.73 meters is 24,828 nanoseconds. The data comprises three samples, of sizes 20, 20 and 26, respectively, corresponding to three different days. These data have been analyzed previously by a number of authors including ? and ?. The p-values of the DPD statistics for the test of the equality of means between Day 1 and Day 2, and Day 1 and Day 3 are plotted in Figure 6, and 7 respectively.

The p-values for the two-sample t-tests for the (Day 1, Day 2) comparison are 0.1058 for the full data case, and 0.3091 for the outlier deleted case. The same for the (Day 1, Day 3) comparison are 0.0970 and 0.2895 respectively. However, for the large γ, the results from the DPD tests are clearly insignificant with or without the outliers. In this example, therefore, the outliers are forcing the outcome of the two-sample t-test (and the DPD tests for small γ) to the borderline of significance, but the robust tests give insignificant results with or without the outliers, preventing the false
Table 4 Newcomb’s Light Speed data.

| day 1 | 28 | 26 | 33 | 24 | 34 | -44 | 27 | 16 | 40 | -2 | 29 | 22 | 24 | 21 | 25 |
|-------|----|----|----|----|----|-----|----|----|----|----|----|----|----|----|----|
| day 2 | 24 | 20 | 36 | 32 | 36 | 28 | 25 | 21 | 28 | 29 | 37 | 25 | 28 | 26 | 30 |
| day 3 | 36 | 23 | 27 | 27 | 28 | 27 | 31 | 27 | 26 | 26 | 26 | 32 | 32 | 32 | 24 | 39 |

Fig. 6 The p-values of the DPD tests for Newcomb’s Light Speed data (Day 1 versus Day 2) for different values of γ. The solid line represents the full data analysis, while the dashed line represents the outlier deleted case.

Fig. 7 The p-values of the DPD tests for Newcomb’s Light Speed data (Day 1 versus Day 3) for different values of γ. The solid line represents the full data analysis, while the dashed line represents the outlier deleted case.

significance that is produced by the outliers in the t-test; this is unlike the previous three examples where the robust tests overcame a masking effect. These examples demonstrate that the robust DPD tests can give protection against spurious conclusions in both directions.
Robust Tests for the Equality of Two Normal Means based on DPD

Table 5 Na Intake data.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
|   | 114.6 | 64.6 | 70.4 | 61.2 | **297** | 69.9 | 73.7 | 15.7 | 53.3 |
| Y | 14.2 | 3.2 | 3.7 | 0.0 | 73.6 | 56.6 | 97.2 | 2.4 | 0.0 | 4.8 | 0.0 |

Fig. 8 The p-values of the DPD tests for Na Intake for different values of γ. The solid line represents the full data analysis, while the dashed line represents the outlier deleted case.

Table 6 Sri Lanka Zinc Content data.

|   | Urban (X) |   |   |   |   |   |   |   |   | Rural (Y) |
|---|---|---|---|---|---|---|---|---|---|---|---|
|   | 1120 | 230 | **4200** | 1200 | 1400 | 750 | 2101 | 430 | 690 | 600 | 834 |
| X | 3619 | 1104 | 243 | 658 | 673 | 598 | 648 | 918 | 133 | 289 | 250 |
|   | 304 | 555 | 640 | 933 | | | | | | |

Example 5 (Na Intake data): Sodium chloride preference was determined in ten patients with essential hypertension and in 12 normal volunteers. All exhibited normal detection and recognition thresholds for the taste of sodium chloride. All were placed on a constant dry diet containing 9 mEq of Na+ and given, as their only source of fluids, a choice of drinking either distilled water or 0.15 M sodium chloride. Patients with essential hypertension consumed a markedly greater proportion of their total fluid intake as saline (38.2% vs 10.6%, average daily preference over one week) and also showed a greater total fluid intake (1,269 ml vs 668 ml, average daily intake over one week). The hypertensive patients consumed more than four times as much salt as did the normal volunteers. The data are given in Table 5. The p-values of the tests for the equality of means are plotted in Figure 8. The findings are similar to examples 1, 2 and 3.

Example 6 (Sri Lanka Zinc Content data): The impact of a polluted environment on the health of the residents of an area is a common environmental concern. Large amounts of heavy metals in the body may signal a serious health threat to a community. One study, performed in Sri Lanka, sought to compare rural Sri Lankans with their urban counterparts in terms of the zinc content of their hair. A collection of individuals from rural Sri Lanka was recruited, samples of their hair were taken, and the zinc content in the hair was measured. An independent collection of students from an urban environment was studied, with the zinc content in samples of their hair being measured as well. The data are given in Table 6. The p-values of the tests for the equality of the means are plotted in Figure 9. The results again indicate that the presence of outliers can mask the true significance in case of the two sample t-test and DPD tests for small values of γ, but for the large γ DPD tests are much more stable in such situations.
5 Concluding Remarks

Without any doubt, the two sample $t$-test is one of the most frequently used tools in the statistics literature. It allows the experimenter to perform tests of the comparative hypotheses, which are the default requirements to be passed before one may declare that a new drug or treatment is an improvement over an existing one. The two sample $t$-test is simple to implement and has several optimality properties. In spite of such desirable attributes, this test is deficient on one count, which is that it does not retain its desired properties under contamination and model misspecification. As few as one, single, large outlier can turn around the decision of the test, and can make the resulting inference meaningless. In this paper we have introduced a test based on the density power divergence; the theoretical properties of the test have been rigorously determined. More importantly, we have demonstrated, through several real data examples, that the DPD test is capable of uncovering both kinds of masking effects caused by outliers – blurring the true difference when one exists, and detecting a difference when there is actually none. The test is simple to use and easy to understand, and we trust that it has the potential to become a powerful tool for the applied statistician.

Acknowledgments This work was partially supported by Grants MTM-2012-33740 and ECO-2011-25706. The authors gratefully acknowledge the suggestions of two anonymous referees which led to an improved version of the paper.

Appendix

Proof of Theorem 1: As $\hat{\mu}_{i,\beta}$ is the solution of the estimating equation $1h'_{n_{i,\beta}}(\mu_{i}, \sigma) = 0$, we get from equation (7)

$$\sqrt{n_i}(\hat{\mu}_{i,\beta} - \mu_{i0}) = \sqrt{n_i} J^{-1}_{11,\beta}(\sigma_0) 1h'_{n_{i,\beta}}(\mu_{i0}, \sigma_0) + o_p(1), \quad i = 1, 2.$$ 

Hence, using (9) we get

$$\sqrt{n_i}(\hat{\mu}_{i,\beta} - \mu_{i0}) \xrightarrow{\text{n}_i \to \infty} N \left( 0, K_{11,\beta}(\sigma_0)J^{-2}_{11,\beta}(\sigma_0) \right), \quad i = 1, 2,$$  

(27)
Similarly we get
\[ K_{11, \beta}(\sigma_0)J_{11, \beta}^2(\sigma_0) = \sigma_0^2 (\beta + 1)^3 (2\beta + 1)^{-\frac{3}{2}}. \]  
(28)

It is clear that \( \hat{\mu}_{1, \beta} \) and \( \hat{\mu}_{2, \beta} \) are based on two independent set of observations, hence, \( \text{Cov}(\hat{\mu}_{1, \beta}, \hat{\mu}_{2, \beta}) = 0 \). As \( 2h'_{n_1, n_2, \beta}(\eta_0) = 0 \), taking a Taylor series expansion around \( \eta_0 \) we get
\[
2h'_{n_1, n_2, \beta}(\eta_0) + \frac{\partial}{\partial \mu_1} 2h'_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} (\hat{\mu}_{1, \beta} - \mu_0) \\
+ \frac{\partial}{\partial \mu_2} 2h''_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} (\hat{\mu}_{2, \beta} - \mu_2) \\
+ \frac{\partial}{\partial \sigma} 2h''_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} (\hat{\sigma}_{\beta} - \sigma_0) + o_p \left( (n_1 + n_2)^{-1/2} \right) \\
= 0. 
\]  
(29)

Notice that
\[
\lim_{n_1, n_2 \to \infty} \frac{\partial}{\partial \mu_1} 2h'_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} = \lim_{n_1, n_2 \to \infty} \frac{\partial}{\partial \mu_1} \left( \frac{n_1}{n_1 + n_2} 2h'_{n_1, \beta}(\mu_{10}, \sigma_0) + \frac{n_2}{n_1 + n_2} 2h'_{n_2, \beta}(\mu_{10}, \sigma_0) \right) \\
= \lim_{n_1, n_2 \to \infty} \frac{n_1}{n_1 + n_2} \lim_{n_1, n_2 \to \infty} \frac{\partial}{\partial \mu_1} h'_{n_1, \beta}(\mu_1, \sigma_0) \bigg|_{\mu_1 = \mu_{10}} \\
= wJ_{12, \beta}(\sigma_0) = 0. 
\]  
(30)

Similarly we get
\[
\lim_{n_1, n_2 \to \infty} \frac{\partial}{\partial \mu_2} 2h'_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} = 0. 
\]  
(31)

Moreover,
\[
\lim_{n_1, n_2 \to \infty} \frac{\partial}{\partial \sigma} 2h''_{n_1, n_2, \beta}(\eta) \bigg|_{\eta=\eta_0} = \lim_{n_1, n_2 \to \infty} \frac{n_1}{n_1 + n_2} 2h''_{n_1, \beta}(\mu_{10}, \sigma_0) + \lim_{n_1, n_2 \to \infty} \frac{n_2}{n_1 + n_2} 2h''_{n_2, \beta}(\mu_{20}, \sigma_0) \\
= wJ_{22, \beta}(\sigma_0) + (1 - w)J_{22, \beta}(\sigma_0) = J_{22, \beta}(\sigma_0). 
\]  
(32)

Therefore, using equations (30), (31) and (32) we get from equation (29)
\[
\sqrt{n_1 + n_2} (\hat{\sigma}_{\beta} - \sigma_0) = -J_{22, \beta}^{-1}(\sigma_0) \sqrt{n_1 + n_2} h'_{n_1, n_2, \beta}(\eta_0) + o_p(1). 
\]  
(33)

Applying (9) and (14) we get
\[
\lim_{n_1, n_2 \to \infty} E \left[ \sqrt{n_1 + n_2} h'_{n_1, n_2, \beta}(\eta_0) \right] \\
= \lim_{n_1, n_2 \to \infty} \frac{\sqrt{n_1 + n_2}}{n_1 + n_2} E \left[ n_1 h'_{n_1, \beta}(\mu_{10}, \sigma_0) + n_2 h'_{n_2, \beta}(\mu_{20}, \sigma_0) \right] \\
= \lim_{n_1, n_2 \to \infty} \sqrt{n_1} \lim_{n_1, n_2 \to \infty} E \left[ \sqrt{n_1} h'_{n_1, \beta}(\mu_{10}, \sigma_0) \right] \\
+ \lim_{n_1, n_2 \to \infty} \sqrt{n_2} \lim_{n_1, n_2 \to \infty} E \left[ \sqrt{n_2} h'_{n_2, \beta}(\mu_{20}, \sigma_0) \right] \\
= 0. 
\]
Similarly we also have
\[
\lim_{n_1, n_2 \to \infty} \frac{n_1 + n_2}{n_1 + n_2} Var \left[ \sqrt{n_1 + n_2} h_{n_1, n_2, \beta}(\eta) \right]
\]
\[
= \lim_{n_1, n_2 \to \infty} \frac{n_1}{n_1 + n_2} V_{\beta} \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \left( n_1 h_{n_1, \beta}(\mu_{10}, \sigma_0) + n_2 h_{n_2, \beta}(\mu_{20}, \sigma_0) \right) \right]
\]
\[
+ \lim_{n_1, n_2 \to \infty} \frac{n_2}{n_1 + n_2} V_{\beta} \left[ \sqrt{n_1} h_{n_1, \beta}(\mu_{10}, \sigma_0) \right]
\]
\[
\frac{n_2}{n_1 + n_2} V_{\beta} \left[ \sqrt{n_2} h_{n_2, \beta}(\mu_{20}, \sigma_0) \right]
\]
\[
= wK_{22, \beta}(\sigma_0) + (1 - w)K_{22, \beta}(\sigma_0)
\]
\[
= K_{22, \beta}(\sigma_0).
\]

Hence,
\[
\sqrt{n_1 + n_2} h_{n_1, n_2, \beta}(\eta) \xrightarrow{n_1, n_2 \to \infty} \mathcal{N}(0, K_{22, \beta}(\sigma_0)).
\]

Now, from equation (33) we get
\[
\sqrt{n_1 + n_2} (\hat{\sigma}_\beta - \sigma_0) \xrightarrow{n_1, n_2 \to \infty} \mathcal{N}(0, K_{22, \beta}(\sigma_0)J^{-2}_{22, \beta}(\sigma_0)).
\]

where
\[
K_{22, \beta}(\sigma_0)J^{-2}_{22, \beta}(\sigma_0) = \sigma_0^2(\beta + 1)^2 \left( \frac{4\beta^2 + 2}{(1 + 2\beta)^{3/2}} - \frac{\beta^2}{(1 + \beta)^3} \right).
\]

As \(J_{12, \beta}(\sigma_0) = J_{21, \beta}(\sigma_0) = 0\), it is clear that
\[
\lim_{n_1, n_2 \to \infty} \frac{h_{n_1, n_2, \beta}(\eta)}{\mu_{10}, \sigma_0} \bigg|_{\eta = \eta_0} = 0,
\]

Therefore, \(cov(\hat{\mu}_{1\beta}, \hat{\sigma}_\beta) = cov(\hat{\mu}_{2\beta}, \hat{\sigma}_\beta) = 0\). Moreover, \(cov(\hat{\mu}_{1\beta}, \hat{\mu}_{2\beta}) = 0\). Combining the results in (27) and (34) we get the covariance matrix of \(\sqrt{n_1 + n_2} \hat{\eta}_\beta\) as follows
\[
\Sigma_{\omega, \beta}(\sigma_0) = \begin{pmatrix}
(1 - w)K_{11, \beta}(\sigma_0)J^{-2}_{11, \beta}(\sigma_0) & 0 & 0 \\
0 & wK_{11, \beta}(\sigma_0)J^{-2}_{22, \beta}(\sigma_0) & 0 \\
0 & 0 & w(1 - w)K_{22, \beta}(\sigma_0)J^{-2}_{22, \beta}(\sigma_0)
\end{pmatrix},
\]

where the values of the diagonal elements are given in (28) and (35). Hence, the theorem is proved.  

Proof of Theorem 2: A Taylor expansion of \(d_\gamma(f_{\hat{\mu}_{1\beta}, \hat{\sigma}_\beta}, f_{\hat{\mu}_{2\beta}, \hat{\sigma}_\beta})\) around \(\eta_0\) gives
\[
d_\gamma(f_{\hat{\mu}_{1\beta}, \hat{\sigma}_\beta}, f_{\hat{\mu}_{2\beta}, \hat{\sigma}_\beta}) = d_\gamma(f_{\mu_{10}, \sigma_0}, f_{\mu_{20}, \sigma_0}) + t_{\gamma}^\top(\eta_0)(\hat{\eta}_\beta - \eta_0) + o_p(\|\hat{\eta}_\beta - \eta_0\|),
\]

where \(t_{\gamma}(\eta_0) = \frac{w}{\eta_0} d_\gamma(f_{\mu_{1\beta}, \sigma_0}, f_{\mu_{2\beta}, \sigma_0})\); the expressions of the components \(t_{\gamma,i}(\eta_0), i = 1, 2, 3\), are given in (18)-(20). Hence, the result directly follows from Theorem 1.

Proof of Theorem 3: If \(\mu_{10} = \mu_{20}\), it is obvious that \(d_\gamma(f_{\mu_{1\beta}, \sigma_0}, f_{\mu_{2\beta}, \sigma_0}) = 0\), and \(t_{\gamma}(\eta_0) = 0\). Hence, a second order Taylor expansion of \(d_\gamma(f_{\hat{\mu}_{1\beta}, \hat{\sigma}_\beta}, f_{\hat{\mu}_{2\beta}, \hat{\sigma}_\beta})\) around \(\eta_0\) gives
\[
2d_\gamma(f_{\hat{\mu}_{1\beta}, \hat{\sigma}_\beta}, f_{\hat{\mu}_{2\beta}, \hat{\sigma}_\beta}) = (\hat{\eta}_\beta - \eta_0)^TA_{\gamma}(\sigma_0)(\hat{\eta}_\beta - \eta_0) + o_p(\|\hat{\eta}_\beta - \eta_0\|^2),
\]

(36)
where $A_\gamma(\sigma_0)$ is the matrix containing the second derivatives of $d_\gamma(f_{\mu_1,\sigma}, f_{\mu_2,\sigma})$ evaluated at $\mu_{10} = \mu_{20}$. It can be shown that

$$A_\gamma(\sigma_0) = \ell_\gamma(\sigma_0) \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where

$$\ell_\gamma(\sigma_0) = \sigma^{-\gamma} \left( \frac{1}{2} \right)^{\frac{\gamma}{2}} (\gamma + 1)^{-\frac{\gamma}{2}}.$$

Therefore, equation (36) simplifies to

$$2d_\gamma(f_{\hat{\mu}_1, \hat{\sigma}}, f_{\hat{\mu}_2, \hat{\sigma}}) = \left( \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} - \begin{pmatrix} \mu_{10} \\ \mu_{20} \end{pmatrix} \right)^T A^*_{\gamma}(\sigma_0) \left( \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} - \begin{pmatrix} \mu_{10} \\ \mu_{20} \end{pmatrix} \right) + o_p \left( \|\hat{\eta}_\beta - \eta_0\|^2 \right),$$

where

$$A^*_{\gamma}(\sigma_0) = \ell_{\gamma}(\sigma_0) \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}.$$

From Theorem 1 we know that

$$\sqrt{n_1 n_2} \left( \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} - \begin{pmatrix} \mu_{10} \\ \mu_{20} \end{pmatrix} \right)^T L \xrightarrow{\mathcal{L}} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^*_{w,\beta}(\sigma_0) \right),$$

where

$$\Sigma^*_{w,\beta}(\sigma_0) = K_{11,\beta}(\sigma_0) J_{11,\beta}^{-2}(\sigma_0) \begin{pmatrix} 1 & -w & 0 \\ -w & 0 & w \end{pmatrix}.$$