Computational Statistics and Data Analysis 115 (2017) 250–266

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

On hyperbolic transformations to normality

Arthur C. Tsai *, Michelle Liou *, Maria Simak, Philip E. Cheng

Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan

HIGHLIGHTS

- A family of hyperbolic transformations towards normality is proposed/constructed.
- The family is effective in transforming skewed/platykurtic distributions to normal.
- The matching quantile approach is used for initial parameter estimates.
- The new family outperforms other well-known transformations in a simulation.
- Data examples on mathematics test scores and DNA microarrays are illustrated.

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 30 March 2017
Accepted 5 June 2017
Available online 27 June 2017

Keywords:
Behrens–Fisher problem
Bimodal distribution
Box–Cox transformation
Levene test
Welch test

ABSTRACT

In biological and social sciences, it is essential to consider data transformations to normality for detecting structural effects and for better data representation and interpretation. An array of transformations to normality has been derived for data exhibiting skewed, leptokurtic and unimodal shapes, but is less amenable to data exhibiting platykurtic shapes, such as a nearly bimodal distribution. This study proposes and constructs a new family of hyperbolic power transformations for improving normality of raw data with varying degrees of skewness and kurtosis. An advantage this new family has is its effectiveness in transforming platykurtic or bimodal data distributions to normal. A simulation study and a real data example on mathematics achievement test scores are used to illustrate the wide-ranging applications of the proposed family of transformations. As a cautionary note, usefulness and limitations of the proposed method will be discussed for stabilizing the variance of DNA microarray data and for symmetrizing the data distribution towards normality. The empirical applications also illustrate an example of conservative t- and ANOVA F-tests when the assumption of normality is violated.

1. Introduction

In biological and social sciences, researchers can be concerned by the presence of nonnormally distributed variables since commonly employed parametric modeling and analysis methods are derived under the assumption of population normality. Empirical and Monte Carlo studies have provided evidence in support of the robustness of parametric inference in small to large samples under the violation of the normality assumption (Sawilowsky and Blair, 1992; Schmider et al., 2010; Rasch et al., 2011). Nevertheless, this does not preclude a usable alternative, namely, data transformation to normality. There are many valid reasons for utilizing data transformations, including improvement of normality, variance stabilization, and conversion of scales to interval measurement (Osborne, 2002; Liermann et al., 2004; Greenacre, 2009; D'Haese et al., 2011;

* Corresponding authors.
E-mail addresses: arthur@stat.sinica.edu.tw (A.C. Tsai), mliou@stat.sinica.edu.tw (M. Liou).

http://dx.doi.org/10.1016/j.csda.2017.06.001

Hou et al., 2011; Pattyn et al., 2011). An array of transformations to normality has been derived from mixing concave and convex functions in order to adjust for both kurtosis and skewness in the data (see Sakia, 1992 for a review of the early literature). In these transformations, a location parameter is defined to adjust for varied skewness in the two tails of a data distribution. The transformations are suitable for data exhibiting skewed, leptokurtic and unimodal shapes as they are similar mixtures of concave and convex functions. It is, however, unclear whether they are practical in application to data exhibiting platykurtic shapes, including the commonly encountered bimodal distribution, which is itself often a mixture of two normal distributions.

This study contributes to the important literature on transformations to normality by introducing a new family of hyperbolic power transformations, hereafter referred to as the HP family. The HP transformation is constructed by implementing a pair of power and scale parameters in a product of hyperbolic functions. Similar to a few existing transformations, two pairs of these parameters are used to adjust for varied shapes of skewness and kurtosis appropriate for general data distributions, but the usual location parameter is not needed. Thus, the HP family incorporates four essential types of transformations in a single formula, which applies concave and convex functions simultaneously to both sides of the sample median.

The paper proceeds as follows. Section 2 introduces the proposed HP family in detail along with a technical review on the Box–Cox transformation (hereafter, the BC family; Box and Cox, 1964) and its extended methods. To find maximum likelihood (ML) parameter estimates of the HP transformation, a method of initial parameter estimation is introduced by matching pairs of selected quantiles in the normalized raw data distribution to the corresponding pairs in the standard normal distribution. This elementary matching quantile approach may facilitate the search for the ML estimates, and often secure compatible ML estimation of the HP parameters. Section 3 provides a simulation to compare the performance of the HP transformation with the BC, gpower, modulus and *sinh–arcsinh* transformations for a range of nonnormal distributions, including the beta, Cauchy, gamma, Laplace, lognormal, Weibull, uniform, and bimodal distributions. The evaluation criteria are the skewness and kurtosis of the transformed distributions along with the results on testing the null hypothesis of normality. Section 4 contains an empirical example on mathematics achievement test scores to demonstrate that a nearly bimodal distribution can be transformed into a normal distribution with the HP transformation in the context of a conservative two-sample t-test and nonrobust ANOVA F-test under bimodality. Section 5 presents the usefulness and limitations of the HP family for stabilizing the variance of DNA microarray data as well as for symmetrizing the data distribution towards normality. Finally, a discussion is offered on further research and applications pertinent to the HP family.

2. The hyperbolic power transformation

The BC family of power transformations is defined on the positive real line ($x > 0$) as

$$\psi^{BC}(x, \lambda) = \begin{cases} (x^{\lambda} - 1)/\lambda & \lambda \neq 0, \\ \log(x) & \lambda = 0, \end{cases}$$

where λ denotes the transformation power parameter. This family transforms skewed data distributions towards normality, and is defined on the positive side of the real line, as depicted in Fig. 1(a). Flexible transformations without such a domain restriction include the exponential transformations by Manly (1976; see Fig. 1(b)), and the extended power transformations by Yeo and Johnson (2000; see Fig. 1(c)). These revised transformations are monotonically concave or convex, making them particularly suitable for data exhibiting a skewed or unimodal shape, but inapplicable to data with platykurtic, leptokurtic, or bimodal shapes.

Useful transformations have also been derived through mixing concave and convex functions in order to incorporate adjustments of kurtosis in the data distribution. Some examples of these families include the signed power transformation (Fig. 1(d); Bickel and Doksum, 1981), inverse hyperbolic sine transformation (Fig. 1(e); Burbidge et al., 1988), and modulus transformation (Fig. 1(f); John and Draper, 1980). Recently, the *sinh–arcsinh* transformation (Fig. 1(g); Jones and Pewsey, 2009) and the gpower transformation (Fig. 1(h); Kelmansky et al., 2013) were proposed to treat data with a peaked sample mode and with heavier or lighter tails than the normal distribution.

As noted before, the BC power transformation has been generalized in the literature to include a variety of concave and convex functions in order to adjust for both kurtosis and skewness in data. An analogous treatment of the kurtosis of raw data distribution is the hyperbolic tangent function $\psi(x) = \tanh(x)$, which has been used as a transfer function in 'infomax' algorithms for characterizing a source density with specified kurtosis (Bell and Sejnowski, 1995; Lee et al., 1999; Hyvärinen et al., 2001). The hyperbolic tangent function is an essential mathematical tool for describing the rate of action potential firing in a neural cell, which is often applied to simulate the dynamic process of input current intensity. However, the hyperbolic tangent function ignores the treatment of data skewness. This important pitfall is the motivation behind this study, which explores the potential utility of treating both kurtosis and skewness based on the hyperbolic tangent function appropriate for general applications.

Without loss of generality, assume that the median of the raw data is located at $x = 0$. The HP transformation is defined as

$$\psi(x, \theta) = \alpha \sinh(\beta x)\mathrm{sech}^{\lambda}(\beta x)/\beta, \tag{1}$$

where $\theta = \{\alpha, \beta, \lambda\}$, $\alpha, \beta > 0$, and $\lambda \leq 1.0$. Thus, a power parameter λ, a scale parameter β, and a slope parameter α (dependent on the other parameters) are implemented in a product of two hyperbolic functions to yield the HP family in

Eq. (1). Note that the scale parameter β also plays a role as the shape parameter as it features in the exponent parameters of both factors in (1), and it is referred to as the scale parameter for simplicity. Moreover, the parameter α, the slope of the transformation function ψ at the median, is a normalizing constant of the underlying probability density as well as a function of the parameters (β, λ).

To accommodate the various shapes of both skewness and kurtosis, two pairs of parameters are required to construct an effective family of transformations: (β_-, λ_-) and (β_+, λ_+), defined on the negative and positive sides of the median, respectively. Both power parameters λ_- and λ_+ cannot be larger than 1.0, otherwise the transformation is no longer monotonic. The smoothness of the transformation is ensured by requiring a continuous second derivative of Eq. (1) with respect to x at the median of the raw data (Yeo and Johnson, 2000). In fact, the first two derivatives at the median are equal to the constants α and zero, respectively, and independent of the parameters $\{\beta_-, \beta_+, \lambda_-, \lambda_+\}$ used with $\psi(x, \theta)$ on both sides of the median. Except for the Laplace distribution that is nondifferentiable at the median, combinations of the parameter values in the proposed HP transformation are capable of modifying both skewness and kurtosis for a wide class of data distributions, and thus generally yield satisfactory performances.

Of particular note is the HP family's ability to remove the bimodality of a distribution on the two sides of the median. Fig. 2 depicts plots of these transformations with specified power and scale parameters. Fig. 2(a), (d), (e), and (g) demonstrate the HP family as inclusive of concave, convex, and interchanges between concave and convex functions as x changes signs with various combinations of λ_- and λ_+. In Fig. 2(c), the parameter λ controls the distribution shape such that positive values of λ between 0.4 and 1.0 are useful for transforming leptokurtic distributions to normality, and values less than 0.4 are useful for transforming platykurtic distributions. The latter includes both asymmetric and symmetric bimodal distributions when λ is small or negative, respectively, as shown in Fig. 2(g)–(i). In addition to scaling by β, combinations of λ and β are shown to adjust for different degrees of kurtosis and skewness in the data. In general, Fig. 2(a)–(c) illustrate the transformation effect on the kurtosis of a data distribution; Fig. 2(d)–(f) illustrate the transformation effect on both kurtosis and skewness. Finally, Fig. 2(g)–(i) indicate the specific effect on bimodal distributions.

2.1. Initial parameter estimation

To estimate the parameters of different families of transformations to normality, standard methods include the ML estimation (Box and Cox, 1964), method of moments (Baker, 1934) and method of quantile points (Johnson, 1949; Forbes et al., 2011). The Newton–Raphson iterative estimation procedure of the ML estimation is infeasible given the difficulty of solving a set of intricate nonlinear equations. This naturally leads to considering the standard simplex method (Nelder and Mead, 1965; Lagarias et al., 1998) or the line search method for finding the ML estimates when proper initial parameter estimates are acquired. For the proposed HP transformation, the ML estimates can be formulated and uniquely solved for the basic functional relations between the quantiles of observed data and corresponding quantiles of transformed data to yield appropriate sets of tentative parameter estimates. While this approach is not defined using functions of sufficient statistics, it can incorporate perturbations comparable to the method of simulated annealing so as to expedite convergence towards the ML estimates of the parameters.

The proposed matching quantile approach is designed to equate a few quantiles of the observed data to the corresponding quantiles in the standard normal distribution. Let x_q, q in $(0, 1)$, denote the qth quantile or $100q$th percentile of the data distribution. Since Eq. (1) preserves the ordering of the observations, the qth quantile of the data distribution is mapped onto the qth quantile of the target distribution, denoted by $\psi(x_q, \theta)$. Assume that the transformed data $\psi(x_q, \theta)$ approximates a normal distribution with mean μ and standard deviation σ. Then

$$\psi(x_q, \theta) = \sigma z_q + \mu \tag{2}$$

where z_q denotes the qth quantile of the standard normal distribution. Suppose that the observed data is centered at the sample median and standardized by an estimated normalizing constant α, corresponding to $\mu = 0$ and $\sigma = 1$. The simplex procedure for estimating θ is initialized with the following three steps:

Step 1. Find initial percentile estimates for β_+ and β_-, respectively. To estimate β_+, for instance, select from the centered data two sample quantiles x_p and $x_q = 2x_p$ on the positive side of the data, where the values of the pair (p, q) can be arbitrary. Suppose p is not greater than 0.75 and q is close to 0.95. It is then easy to verify whether z_q is less than or greater than $2z_p$ on the standard normal scale. In general, the sample kurtosis and sample quantiles of the raw data can be measured such that z_q is either less than or greater than $2z_p$. Then a proper initial estimate of the power parameter is $\lambda = 1$ or $\lambda = 0$, respectively. It follows from Eq. (1) and the quantiles used in Eq. (2) that either

$$\tanh(x_p\beta)z_q = \tanh(2x_p\beta)z_p$$

for the case of higher-valued kurtosis, or

$$\sinh(x_p\beta)z_q = \sinh(2x_p\beta)z_p$$

for the case of lower-valued kurtosis. The two separate equations yield either

$$\beta|x_p| = \operatorname{arccosh}\left(\frac{z_p}{z_q - z_p}\right) \tag{3}$$

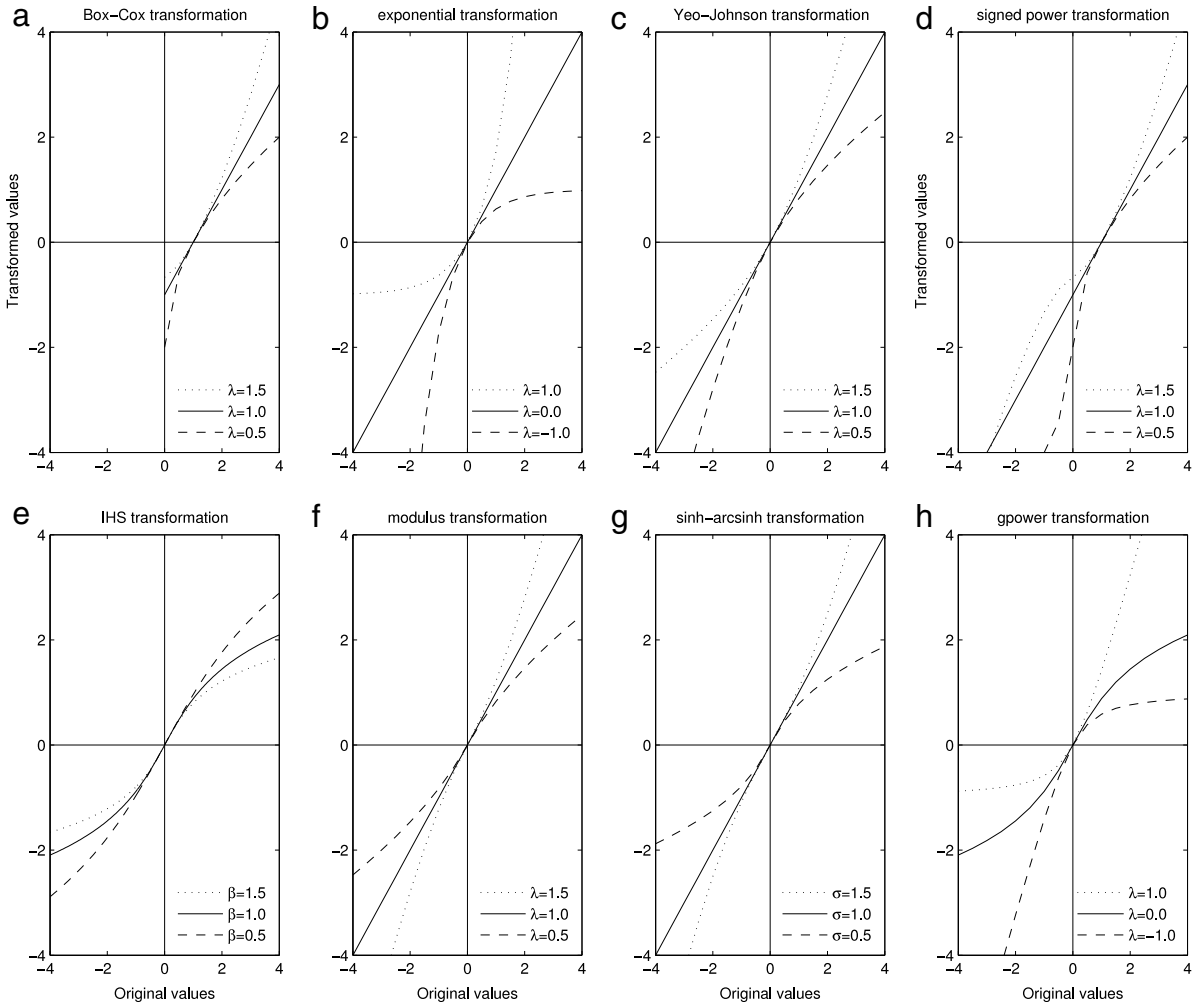


Fig. 1. Transformation graphs (a)–(c) are concave or convex functions, (d)–(g) are mixed concave and convex functions, and (h) includes both functional forms. The graphs depict: (a) BC transformation, (b) exponential transformation: $\psi(x,\lambda) = (\exp(\lambda x)-1)/\lambda$, $\lambda \neq 0$; $= x$, $\lambda = 0$, (c) Yeo–Johnson power transformation: $\psi(x,\lambda) = \left((x+1)^{\lambda}-1\right)/\lambda$, $\lambda \neq 0$, or $\log(x+1)$, $\lambda = 0$, where $x \geq 0$; $\psi(x,\lambda) = -\left((-x+1)^{2-\lambda}-1\right)/(2-\lambda)$, $\lambda \neq 2$, or $-\log(-x+1)$, $\lambda = 2$, where $x<0$, (d) signed power transformation: $\psi(x,\lambda) = \left(\text{sign}(x)|x|^{\lambda}-1\right)/\lambda, \lambda > 0$, (e) inverse hyperbolic sine (IHS) transformation: $\psi(x,\beta) = \sinh^{-1}(\beta x)/\beta$, (f) modulus transformation: $\psi(x,\lambda) = \text{sign}(x)\left((|x|+1)^{\lambda}-1\right)/\lambda$, $\lambda \neq 0$; $= \text{sign}(x)(\log(|x|+1))$, $\lambda = 0$, (g) *sinh–arcsinh* transformation: $\psi(x,\sigma,\gamma) = \sinh\{\sigma \sinh^{-1}(\frac{x-\gamma}{\sigma}) - \gamma\}$, and (h) gpower transformation: $\psi(x,\lambda) = \left((x+\sqrt{x^2+1})^{\lambda}-1\right)/\lambda$, $\lambda \neq 0$; $= \log(x+\sqrt{x^2+1})$, $\lambda = 0$. Note: The horizontal and vertical axes are the standardized scales of the original and transformed data, respectively.

or

$$\beta|x_p| = \text{arccosh}\left(\frac{z_q}{2z_p}\right), \tag{4}$$

respectively. Appropriate usage of either Eq. (3) or (4) according to high or low kurtosis of the raw data, respectively, can be empirically assessed without additional computational effort. A similar analysis can be used to provide the initial estimate of the parameter β_- defined on the negative side of the sample median.

Step 2. Find an initial estimate for the parameter λ to fine-tune the degree of concavity or convexity by the transformation. It follows from the initial estimate of β in Step 1, and a pair of sample quantiles x_s and x_t based on (2) that

$$\sinh(\beta x_s)\text{sech}^{\lambda}(\beta x_s)z_t = \sinh(\beta x_t)\text{sech}^{\lambda}(\beta x_t)z_s,$$

and an initial estimate of λ can be derived as

$$\lambda = \frac{\log(z_s/z_t) + \log\sinh(\beta x_t) - \log\sinh(\beta x_s)}{\log\text{sech}(\beta x_s) - \log\text{sech}(\beta x_t)}. \tag{5}$$

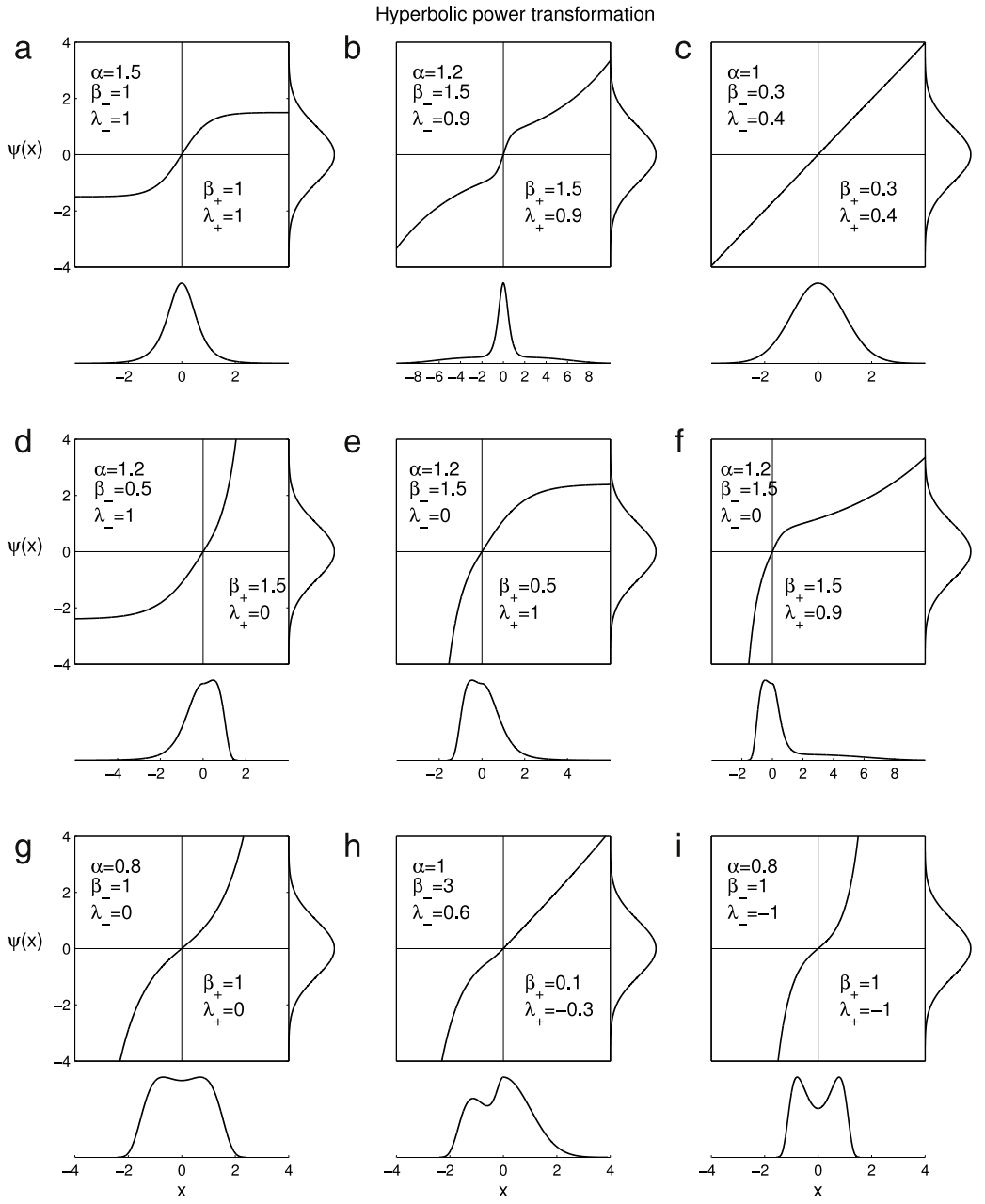


Fig. 2. Graphs of the hyperbolic power functions defined on the line using the slope parameter α and pairs of scale and power parameters (β_-, λ_-) and (β_+, λ_+) on the left-hand side and right-hand side of zero, respectively. The probability density function (pdf) plotted below each graph is transformed to the standard normal pdf along the right horizontal axis in each graph.

Step 3. The slope parameter α of the transformation at $x = 0$ can be estimated for the transformed data with zero mean and unit variance as follows. Apply the above estimates of β and λ to (2), and a pair of quantile points, say x_{r-} and x_{r+}, which are close to one another on either side of zero. It follows by unit variance that an appropriate estimate of α is

$$\alpha = \frac{\beta_+ z_{r+} - \beta_- z_{r-}}{\sinh(\beta_+ x_{r+})\mathrm{sech}^{\lambda_+}(\beta_+ x_{r+}) - \sinh(\beta_- x_{r-})\mathrm{sech}^{\lambda_-}(\beta_- x_{r-})}. \tag{6}$$

The above three steps together form a fast and simple procedure for computing the initial solutions. The sample quantiles of steps 1 and 2 can be conveniently selected quantiles, for example, 7%, 25%, 75% and 93%, which were used in the simulation and the empirical studies in Sections 3 and 4. Other choices of quantiles are also valid.

2.2. Maximum likelihood estimation

The probability density function defined on the HP transformation in (1) can be written as

$$p(x|\theta) = \phi(y(x))|J(x)|. \tag{7}$$

Here, $y(x) = \psi(x, \theta)$, $\phi(y(x)) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\psi^2(x, \theta)/2\right\}$ is the standard normal probability density, and $|J(x)| = |\partial y(x)/\partial x| = \alpha\left(1 - \lambda\tanh^2(\beta x)\right)\text{sech}^{\lambda-1}(\beta x)$ is the Jacobian of the transformation. In estimating the parameters, raw data on the negative and positive sides of the sample median are divided and separately analyzed to find the ML estimates of the parameter pairs, (β_-, λ_-) and (β_+, λ_+), respectively. The log-likelihood function $L(\theta|x) = \log p(\theta|x)$, omitting the arguments for notational simplicity, can be expressed as

$$L(\theta|x) = const - \frac{1}{2}\sum_i \psi^2 + n\log\alpha + \sum_i \log(1 - \lambda\tanh^2) + (\lambda - 1)\sum_i \log\text{sech} \tag{8}$$

for the observed sample $x = \{x_i, i = 1, \ldots, n\}$, where $\psi(x_i, \theta)$ is abbreviated as ψ, and likewise, $\tanh(\beta x_i)$, $\text{sech}(\beta x_i)$, $\cdots$, and $\sinh(\beta x_i)$, are abbreviated as tanh, sech, $\cdots$, and sinh, respectively.

The ML estimates of $\theta = \{\alpha, \beta_-, \beta_+, \lambda_-, \lambda_+\}$ are the solutions to the score equation of the first-order partial derivatives $\partial L(\theta|x)/\partial\theta = 0$ from (8), where the solution to α, given the estimates of β and λ, is

$$\alpha = \left(\frac{1}{n}\sum_i (\sinh\ \text{sech}^{\lambda}/\beta)^2\right)^{-1/2}. \tag{9}$$

Ideally, the Newton–Raphson algorithm iteratively updates the parameter estimates using the first-order derivatives $\partial L(\theta|x)/\partial\theta = 0$ and the Hessian matrix of second-order derivatives (cf. Theorem 6.3.10 in Lehmann and Casella, 1998 and proof of Lemma 1 in Appendix). Numerical approximation to the solutions of $\partial L(\theta|x)/\partial\theta = 0$ derived from formula (8) often leads to the locations of maximum likelihood and saddle point(s) where the sample Hessian matrices are checked to be negative or positive. This is known as the weaker version of the Cramér consistency of ML parameter estimation (Lehmann and Casella, 1998) which is stated as Lemma 1 in Appendix along with its proof. In Lemma 1, the open set $\Theta = \{(\beta, \lambda) : \beta > 0, \lambda < 1\}$ is the parameter space defined for continuous partial differentiability with respect to the parameters (β, λ). The boundary set defined with "$\lambda = 1.0$" (and its accompany β) is usually expected to be the solution of the maximum likelihood formula (8) when the raw data appears to show a leptokurtic distribution. In practice, approximation to the desired location of maximum likelihood can be achieved using the simplex method (Nelder and Mead, 1965; Lagarias et al., 1998) instead of Newton–Raphson algorithm. In this approach, initial parameter estimates obtained as solutions to the three-step percentile equations of Section 2.1 can be effectively implemented. An iterative scheme is employed to warrant approximation to the maximum of (8) through updating the parameter estimates by perturbing the slope parameter estimate α of (9).

3. Simulation study

We compare the performance of the proposed HP family of transformations against four other well-known and popular families of transformations, namely, BC, modulus, *sinh–arcsinh* and gpower transformations. While the literature has examined the theoretical properties of each of these transformations, there exists no comparison study on the effects of these transformations. Thus, the simulation work in this section additionally contributes to the literature by offering the first comprehensive comparison study among these four transformations, as well as the proposed HP transformation. Using MATLAB (The MathWorks, Inc.) functions, data was simulated from the following nine families of distributions: lognormal, Weibull, gamma, exponential, Laplace, beta, uniform, bimodal and Cauchy distributions (Rakhshan and Pishro-Nik, 2014). These families represent typical nonnormal distributions that exhibit skewed, leptokurtic and platykurtic shapes; five distributions bear standard defining parameter values, and two beta distributions, with especially small and large kurtosis values. The bimodal distribution is defined as a mixture of two normal distributions using the mean values 0 and 8, variances 2 and 3, and the mixing proportions 0.2 and 0.8, respectively. The five transformations were applied to each of the simulated random samples of the ten nonnormal distributions. The raw and the transformed data were tested against the null hypothesis of normality using three modern tests: the robustified Jarque–Bera (RJB) test (Gel and Gastwirth, 2008), a skewness-based Z_2^*-test and its counterpart kurtosis-based Z_3^*-test (Mudholkar et al., 2002; Stehlík et al., 2014). These three tests were constructed on the basis of testing sample skewness and kurtosis of the normalized data against those of standard normal distribution, 0 and 3, respectively.

The present simulation study comprised of 5000 random samples of size $n = 100$ for each member of the selected ten distributions. To carry out the comparisons using the five families of transformations, the following layouts are given. First, values of skewness and kurtosis of the simulated raw data and the five transformed data are presented by using standard box plots in Fig. 3. Next, results from testing against the null hypothesis of standard normality (using the RJB test, the Z_2^*-test and the Z_3^*-test) are reported in Tables 1–3, respectively. In calculating the test effects of these three tests, the test value is defined as zero if the null hypothesis of normality is accepted; otherwise, it is defined as one. Technical details of the transformation families in Fig. 1 of Section 2 are stated, and results of the simulation study are presented in Fig. 3 and Tables 1–3.

Table 1
Average rejection rates on the transformed data by the robust Jarque–Bera (RJB) test for normality. Bold entries show the rejection rates lower than 5% (<0.05) against the null hypothesis that the transformed data are normally distributed.

Distribution	Raw data	Box–Cox	Modulus	sinh–arcsinh	gpower	HP
log-normal (0, 1)	1.000	**0.010**	0.106	**0.015**	**0.006**	**0.000**
Weibull (2, 1)	1.000	**0.001**	0.125	**0.019**	**0.031**	**0.000**
Gamma (2, 2)	0.984	**0.005**	0.125	0.139	0.050	**0.001**
Exp (1)	1.000	**0.001**	0.212	**0.001**	**0.037**	**0.000**
Laplace (0, 1)	0.893	0.772	**0.003**	**0.002**	**0.000**	**0.016**
Beta (5, 1)	0.992	**0.013**	**0.009**	0.890	0.413	**0.000**
Beta (2, 2)	**0.000**	**0.000**	**0.034**	**0.000**	**0.000**	**0.000**
Uniform (0, 1)	**0.043**	**0.015**	**0.024**	**0.022**	**0.040**	**0.000**
bimodal (0, 2; 8, 3)	0.160	**0.001**	**0.003**	0.315	0.132	**0.001**
Cauchy (0, 1)	1.000	1.000	0.164	0.244	**0.009**	**0.004**

Note: Simulation results based on 5000 random samples of size 100.

Table 2
Average rejection rate on the transformed data by Z_2^* test for normality. Bold entries show the rejection rates lower than 5% (<0.05) against the null hypothesis that the transformed data are normally distributed.

Distribution	Raw data	Box–Cox	Modulus	sinh–arcsinh	gpower	HP
log-normal (0, 1)	1.000	**0.000**	0.094	**0.010**	**0.004**	**0.000**
Weibull (2, 1)	1.000	**0.000**	0.186	**0.044**	**0.036**	**0.000**
Gamma (2, 2)	1.000	**0.000**	0.133	0.061	0.053	**0.000**
Exp (1)	1.000	**0.000**	0.268	**0.018**	**0.038**	**0.000**
Laplace (0, 1)	0.294	**0.000**	**0.017**	**0.012**	**0.000**	**0.002**
Beta (5, 1)	1.000	0.419	**0.045**	0.998	0.413	**0.000**
Beta (2, 2)	**0.021**	**0.001**	0.064	**0.005**	**0.007**	**0.000**
Uniform (0, 1)	**0.044**	0.409	**0.009**	**0.002**	**0.015**	**0.000**
bimodal (0, 2; 8, 3)	0.537	**0.022**	**0.002**	0.141	0.426	**0.000**
Cauchy (0, 1)	0.872	0.250	0.106	0.209	**0.000**	**0.000**

Table 3
Average rejection rate on the transformed data by Z_3^* test for normality. Bold entries show the rejection rates lower than 5% (<0.05) against the null hypothesis that the transformed data are normally distributed.

Distribution	Raw data	Box–Cox	Modulus	sinh–arcsinh	gpower	HP
log-normal (0, 1)	0.998	0.211	0.170	0.190	**0.018**	**0.003**
Weibull (2, 1)	0.896	0.374	0.374	0.650	0.143	**0.042**
Gamma (2, 2)	0.635	0.269	0.205	0.145	0.138	**0.002**
Exp (1)	0.893	0.353	0.292	0.848	0.140	**0.036**
Laplace (0, 1)	0.815	0.721	**0.005**	**0.021**	**0.027**	**0.012**
Beta (5, 1)	0.419	1.000	0.851	0.195	0.264	**0.045**
Beta (2, 2)	0.979	0.963	0.499	0.974	0.976	**0.005**
Uniform (0, 1)	1.000	1.000	0.867	1.000	0.999	0.364
bimodal (0, 2; 8, 3)	0.513	0.734	0.120	0.523	0.482	**0.013**
Cauchy (0, 1)	1.000	1.000	0.123	0.178	**0.028**	**0.004**

In Fig. 3, the first column presents the box plots of the skewness (3a) and kurtosis (3b) of the simulated raw data. The second column lists the plots of skewness and kurtosis for the BC transformed data, where the BC power and location parameters are estimated using the MATLAB functions from Strauss (2012). The third column exhibits these values derived from the modulus transformation

$$\psi^{MT}(x, \nu, \delta, \sigma, \gamma, \lambda) = \nu + \delta \text{sign}\left(\frac{x-\gamma}{\sigma}\right) \frac{(|\frac{x-\gamma}{\sigma}| + 1)^{\lambda} - 1}{\lambda},$$

where λ is the shape parameter, and four additional parameters ν, δ, γ and σ are used as standard auxiliary location and scale parameters. The simplex algorithm can be used for a direct search of the parameters through maximizing the log-likelihood, which is expressed as

$$L = const - \frac{1}{2}\sum_i (\psi^{MT})^2 + n\log\delta - n\log\sigma + (\lambda - 1)\sum_i \log\left(\left|\frac{x-\gamma}{\sigma}\right| + 1\right).$$

The arguments of the transformation are omitted for brevity. The fourth column displays plots for data under the four-parameter *sinh–arcsinh* transformation

$$\psi^{SINH-ARCSINH}(x, \nu, \delta, \sigma, \gamma) = \nu + \delta \sinh\left\{\sigma \sinh^{-1}\left(\frac{x-\gamma}{\sigma}\right) - \gamma\right\},$$

whose log-likelihood function is

$$L = const - \frac{1}{2}\sum_i (\psi^{SINH-ARCSINH})^2 + n\log\delta$$
$$-\frac{1}{2}\sum_i \log\left(1+\left(\frac{x-\gamma}{\sigma}\right)^2\right) + \sum_i \log\cosh\left(\sigma\sinh^{-1}\left(\frac{x-\gamma}{\sigma}\right)-\gamma\right).$$

The fifth column yields plots for data under the gpower transformation

$$\psi^{GPOWER}(x,\nu,\delta,\sigma,\gamma,\lambda) = \begin{cases} \nu + \delta\dfrac{\left(\frac{x-\gamma}{\sigma}+\sqrt{1+\left(\frac{x-\gamma}{\sigma}\right)^2}\right)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \nu + \delta\log\left(\dfrac{x-\gamma}{\sigma}+\sqrt{1+\left(\dfrac{x-\gamma}{\sigma}\right)^2}\right), & \text{if } \lambda = 0. \end{cases}$$

The log-likelihood function can be expressed as

$$L = const - \frac{1}{2}\sum_i (\psi^{GPOWER})^2 + n\log\delta - n\log\sigma - \frac{1}{2}\sum_i \log\left(1+\left(\frac{x-\gamma}{\sigma}\right)^2\right)$$
$$+\lambda\sum_i \log\left(\left(\frac{x-\gamma}{\sigma}\right)+\sqrt{1+\left(\frac{x-\gamma}{\sigma}\right)^2}\right).$$

The last (right-hand side) column refers to the transformed data using the proposed HP transformation. The simplex method was employed to compute the ML estimates in (8), where Eqs. (3)–(6) were used to find initial parameter estimates in accordance with the sample skewness and kurtosis of the data distribution. The six columns in Tables 1–3 follow the same order of presentation as Fig. 3.

The plots in Fig. 3 indicate that the HP transformation outperforms its competitors in most cases, except for two distributions, namely, Laplace (0, 1) and Cauchy (0, 1). This is because the Laplace density function is not smooth at the median, and the Cauchy distribution has undefined moments. In general, these two distributions are utilized for special purpose without the need of transformations, and were included in the simulation to give more complete illustrations. The results in Tables 1–3 indicate that the proposed HP transformation yields significantly better performance than the other transformations for most of the nonnormal distributions.

The matching quantile approach proceeds with centering the raw data at the sample median so that the transformation to a normal distribution does not require the use of a location parameter. This is in direct contrast to the BC and many other transformations (cf. simulation study in Section 3). The initial parameter estimation by the percentile approach is effective in expediting the approximation to the ML estimates for both parameters β and λ, because the method can be iterated with increased likelihood by perturbing the slope parameter estimate α at the sample median. This technical match in using the sample median without using a location parameter is a practical advantage of the proposed HP transformation. In applications, $\lambda > 0.4$ is recommended for data with heavy tails, and $\lambda \leq 0.4$ is more suitable for lighter tails. Unequally valued pairs, for the shape parameter λ and the scale parameter β, can be separately selected on the two sides of the sample median, respectively, for skewed data (cf. Fig. 2(h)), while equally valued pairs are suitable for symmetric data distribution with a leptokurtic or platykurtic shape.

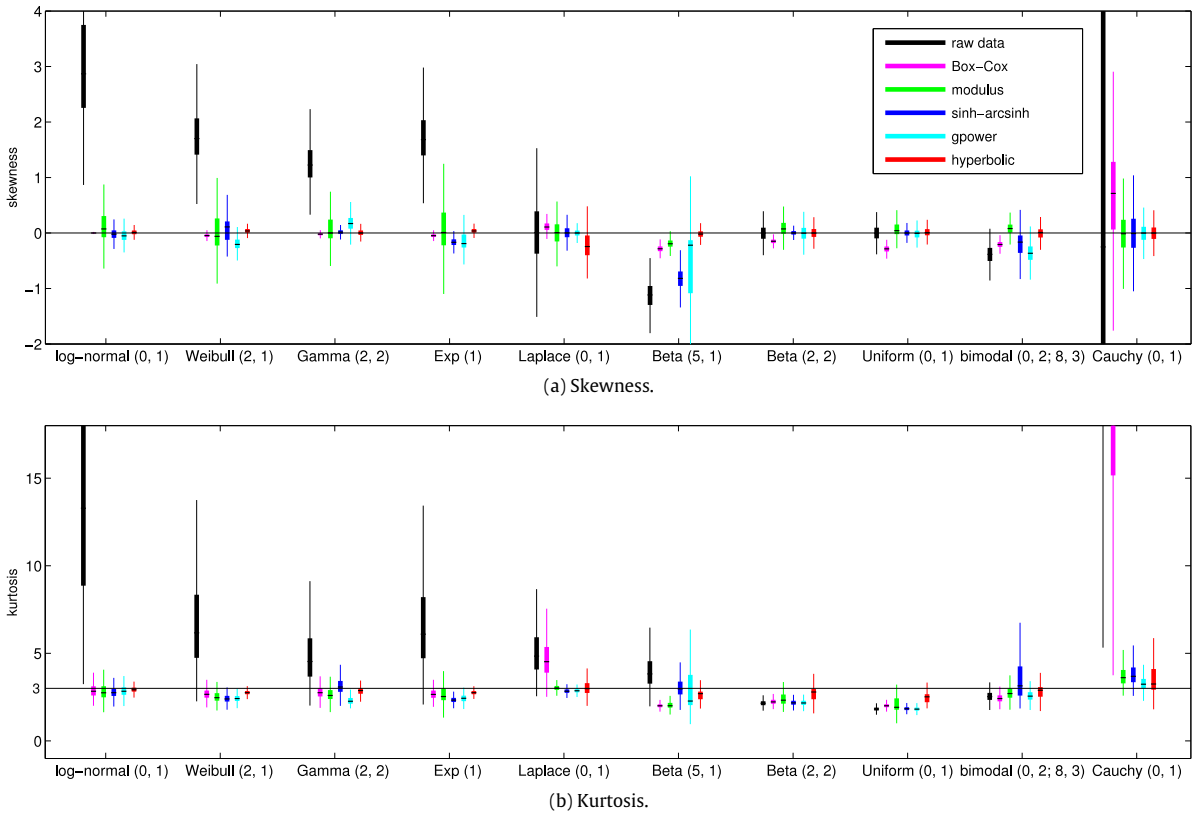
4. Mathematics achievement test scores

In this section, a data set on scores of a mathematics achievement test from Taiwan is analyzed using the parametric t- and F-tests. We investigate the effect of the basic normality assumption in the application of parametric statistics to real nonnormal data. The acquired test effects are compared between the raw and transformed data obtained from the proposed HP transformation and the well-known families: BC, modulus, *sinh–arcsinh*, and gpower transformations. The achievement test was a part of the mathematical examination administered by the Taiwan Basic Achievement Test Center in 2009. The data set contains scores on a 34-item test for measuring ability to solve practical problems using 9th grade algebra and geometry knowledge and are listed according to the following demographic factors: gender (male vs. female) and residency (urban vs. rural). Questions were graded as 0 points for an incorrect answer, and 1 point for a correct answer. A sample of 7928 individuals and their scores from the data set were randomly selected. Male students performed significantly better, as demonstrated through a likelihood ratio test ($p < 0.01$; Cheng et al., 2008), in 14 of 34 questions (mainly in geometry). Female students performed significantly better than their male counterparts ($p < 0.01$) on 8 questions out of the 34 (mainly in algebra).

Table 4 shows the two-way ANOVA tests on either the raw or transformed scores classified according to gender and residency. Histograms of the residuals are given in Fig. 4 for the raw and transformed scores after fitting the ANOVA model.

Table 4
Statistics of the mathematics scores from the Taiwan Basic Achievement Test Center before and after transformation by BC, modulus, *sinh–arcsinh*, gpower, and hyperbolic power transformation. Bold entries indicate the results from transformed scores that male students score significantly ($p<0.01$) higher than female students based on two-sample t-test in conjunction with the significant F-test.

	Raw data				BC				Modulus			
Parameters					$\lambda = 1.52$, $c = 3.58$ (location parameter)				$\nu = -0.42, \delta = 26.32$, $\sigma = 60.25, \gamma = -0.65$, $p = 63.03$			
Skewness	−0.40				−0.21				−0.03			
Kurtosis	1.98				1.82				2.63			
RJB test	<0.001				<0.001				<0.001			
Male ($n = 4098$)	mean = 22.17 SD = 8.47				mean = 22.22 SD = 8.48				mean = 22.32 SD =8.60			
Female ($n = 3830$)	mean = 21.88 SD = 7.81				mean = 21.82 SD = 7.79				mean = 21.72 SD =7.64			
Welch t-test	$t(7924.52) = \frac{0.292}{0.183} = 1.597$ $p = 0.110$				$t(7921.52) = \frac{0.402}{0.183} = 2.202$ $p = 0.028$				$t(7900.02) = \frac{0.598}{0.182} = 3.277$ $p = 0.001$			
ANOVA test: source	d.f	Mean Sq.	F	$p > F$	d.f	Mean Sq.	F	$p > F$	d.f	Mean Sq.	F	$p > F$
Gender	1	98.26	1.53	0.22	1	218.16	3.40	0.07	1	588.11	9.19	**0.002**
Residency	1	18398.20	286.62	< 0.001	1	18911.10	294.93	< 0.001	1	19664.96	307.29	< 0.001
Gender× Residency	1	13.11	0.20	0.65	1	10.46	0.16	0.69	1	2.86	0.04	0.83
Error	7924	64.19			7922	64.12			7919	63.99		
	sinh–arcsinh				gpower				HP			
Parameters	$\nu = -0.48, \delta = 0.71$, $\sigma = 1.461e+5$, $\gamma = -0.23$				$\nu = -0.94, \delta = 358.32$, $\sigma = 435.59, \gamma = -0.88$, $p = 108.39$				$\alpha = 0.23, \beta_- = 0.17$, $\lambda_- = -1.89, \beta_+ = 1.03$, $\lambda_+ = -0.11$			
Skewness	−0.096				−0.1				−0.001			
Kurtosis	2.31				1.81				2.9			
RJB test	<0.001				<0.001				0.27			
Male ($n = 4098$)	mean = 22.28 SD = 8.57				mean = 22.26 SD = 8.50				mean = 22.30 SD = 8.63			
Female ($n = 3830$)	mean = 21.75 SD = 7.68				mean = 21.78 SD = 7.77				mean = 21.73 SD = 7.60			
Welch t-test	$t(7908.67) = \frac{0.532}{0.183} = 2.914$ $p = 0.004$				$t(7917.11) = \frac{0.473}{0.183} = 2.584$ $p = 0.010$				$t(7893.39) = \frac{0.570}{0.182} = 3.122$ $p = 0.002$			
ANOVA test: source	d.f	Mean Sq.	F	$p > F$	d.f	Mean Sq.	F	$p > F$	d.f	Mean Sq.	F	$p > F$
Gender	1	442.24	6.91	**0.009**	1	320.49	5.00	0.025	1	533.77	8.33	**0.004**
Residency	1	19573.30	305.73	< 0.001	1	19202.40	299.60	< 0.001	1	19233.10	300.24	< 0.001
Gender× Residency	1	4.00e-4	6.25e-6	0.998	1	7.23	0.11	0.74	1	4.01	0.06	0.80
Error	7920	64.02			7919	64.09			7919	64.06		



(a) Skewness.

(b) Kurtosis.

Fig. 3. Simulation results based on 5000 random samples of size 100 from lognormal, Weibull, gamma, exponential, Laplace, beta, uniform, bimodal, and Cauchy distributions: (a) skewness, (b) kurtosis. Stem plots were made using the MATLAB *boxplot* function with default parameters. The median is denoted by the black horizontal line segment marked on each solid box, which denotes the interquartile range of the raw data and transformed data distributions. The range of each vertical line includes the percentile values $q_3 + 1.5(q_3 - q_1)$ and $q_1 + 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles. Note: The red plots are results from the HP transformation.

The raw scores clearly exhibit a bimodal shape, which is unsurprising as a few geometry questions were designed to be very difficult and intended for high achievement students. The ANOVA test on raw scores also suggests that students in urban schools score significantly higher than those in rural schools. In this example, the classical ANOVA F-test of the gender effect yields $F(1, 7924) = 1.53$ ($p = 0.22$) on the raw scores, and $F(1, 7919) = 8.33$ ($p = 0.004$) on the HP transformed scores. In general, population normality is not considered as essential as equality of variances among the groups in an unbalanced ANOVA F-test (Glass et al., 1972; Tomarken and Serlin, 1986; Leys and Schumann, 2010). Both Levene and Brown–Forsythe tests for the equal-variance assumption (Levene, 1960; Brown and Forsythe, 1974; Gastwirth et al., 2009) are rejected in both raw and transformed score distributions ($p < 0.01$). However, the HP transformed scores give smaller F-values on both equal-variance tests than the raw scores (i.e., variances of transformed scores are less heterogeneous; for example, the Brown–Forsythe test gives $F(3, 7924) = 30.03$ on the raw scores and $F(3, 7919) = 27.26$ on the HP transformed scores). Note that the variances of the BC and gpower transformed scores, with kurtoses 1.82 and 1.81 respectively, are also less heterogeneous than those of the raw scores; the two methods return less significant p-values for testing the gender effect compared with other transformation methods. By comparing the skewness and kurtoses of raw scores with those of the transformed scores in Table 4 and by referring to the Levene and Brown–Forsythe test results, it is reasonable to suggest that the conservativeness of the ANOVA F-test on raw scores is mainly attributable to violation of normality.

As a comparison, Table 4 also presents the results of a Welch two-sample t-test which is known to be robust to normality and equal-variance assumptions (Welch, 1947; Sawilowsky and Blair, 1992; Rasch et al., 2011). The independent two-sample t-test yields $t = 1.60$ ($p = 0.11$) on the raw scores, indicating no evidence that male students ($n = 4098$; mean $= 22.17$; SD $= 8.47$) scored higher than female students ($n = 3830$; mean $= 21.88$; SD $= 7.81$). Nevertheless, the same t-test on the HP transformed scores suggests that male students do indeed score higher than female students ($t = 3.12$, $p = 0.002$). Table 4 shows that a significant gender effect is also found with other transformed scores, especially with those from the modulus and *sinh–arcsinh* methods ($p < 0.01$). We make two important observations. First, the HP transformation is a unique and capable method for transforming a bimodal distribution to normal according to the RJB test ($p = 0.27$; cf.

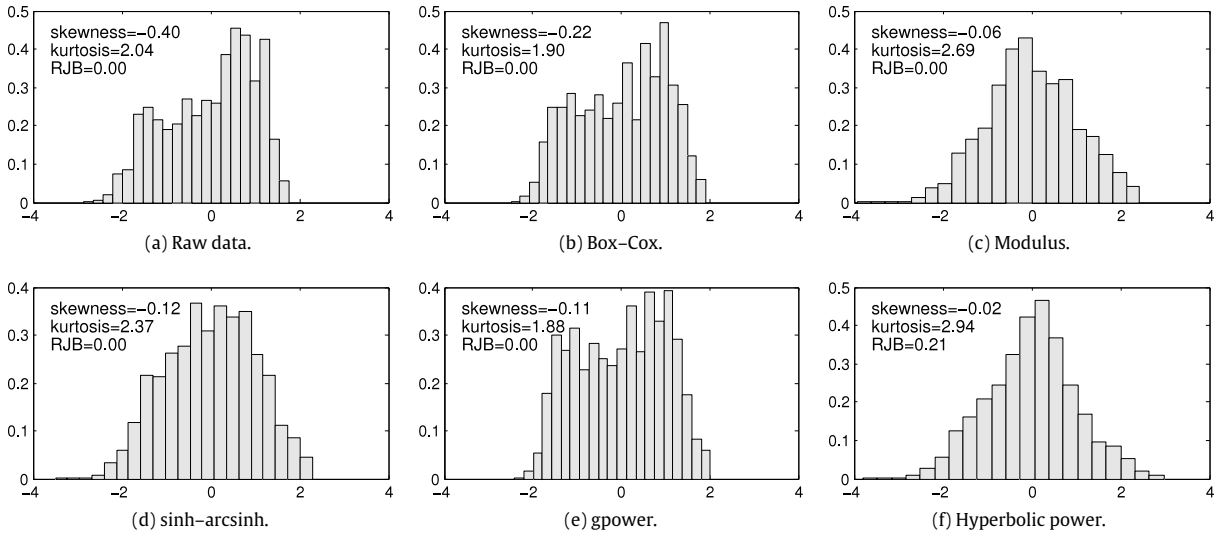


Fig. 4. Histograms of the ANOVA residuals of the (a) raw data, and the transformed data including the (b) BC, (c) modulus, (d) *sinh–arcsinh*, (e) gpower and (f) HP transformations.

Fig. 4 and Table 4). The example suggests a case where the Welch t-test can be conservative under nonnormal distributions in large samples. Second, the significant gender effect found with the transformed scores is essentially attributable to the enlarged difference between group means, because the Levene and Brown–Forsythe tests for inhomogeneous variances of both raw and transformed data are equally significant. In this example, it is also of interest to compare the t-test with its nonparametric counterpart: the Wilcoxon two-sample rank test (cf. Lehmann and D'Abrera, 2006) of the gender effect gives $p = 0.004$ on the raw scores, and $p = 0.005$ on the HP transformed scores.

In the basic ANOVA framework, the regressors in the linear model are categorical variables (e.g., gender and residency) for testing the mean differences. In this elementary case, bimodality in the raw scores also exists in the residuals of fitting linear models to data. As mentioned earlier, the gender difference cannot be detected by the two-sample t-test or ANOVA F-test on the raw scores, but can be detected by the t-test, rank test, or the ANOVA F-test on the HP transformed scores. This example suggests that the so-called robust t- and F-tests to mild nonnormality (such as mild skewness) in ANOVA may not apply to the case of ill-conditioned kurtosis, even in sufficiently large samples.

5. Microarray data

DNA microarrays contribute a high-throughput screening tool for obtaining expression profiles via a single assay, and are known to produce large amounts of data that are surprisingly resistant to analysis by standard statistical techniques (Durbin et al., 2002; Parrish et al., 2009). This is because these techniques require the assumption that data come from a normal or at least symmetric distribution. A transformation-based approach, for example log_2 or the generalized logarithmic (*glog*) transformation, is commonly used for stabilizing the variance of microarray data expressed at high levels, and is also useful for making the data more symmetric (Munson, 2001; Durbin et al., 2002; Huber et al., 2002; Leiva et al., 2009; Vilca et al., 2013). In this example, we show the usefulness and limitations of the HP transformation prior to statistical analysis.

The MicroArray Quality Control (MAQC) project (Shi et al., 2006; Wen et al., 2010) provides data resources that help build consensus on the use of microarrays (Gentleman et al., 2004; Ambroise et al., 2011). Data acquired with the Agilent one-color (AG1) platform from Sites 1 and 3 were downloaded from the Gene Expression Omnibus (GEO) repository (GEO accession: GSE5350, Barrett and Edgar, 2006) with five replicate assays in a given sample. Data points labeled as 'Absent' or 'Marginal' were dropped from analysis. Among the samples measured in the MAQC study, samples A and B were used for illustration. Sample A corresponds to the Universal Human Reference RNA from Stratagene, and sample B corresponds to the Human Brain Reference RNA from Ambion.

The Welch two-sample t-test for differential expressions between samples A and B returns a p-value for each probe. A probe designed on an array is used to measure a transcript expression level for a specific gene. Normalized data from the Taqman quantitative PCR were also downloaded from the GEO repository with four replicate assays to compute fold-changes and p-values with 1044 probes between samples A and B. These values are regarded as gold-standard fold-changes and p-values (Ambroise et al., 2011).

Figs. 5(a) and 6(a) present the histogram of raw intensities and scatter plot of the mean intensity of the five replicates versus their standard deviation of the 1439 probes (corresponding to the 1044 distinct Tagman RNAs) acquired from the

Table 5
Comparison between the transformation methods by sensitivity, specificity, classification accuracy and Kappa.

	log_2	*glog*	HP
Sensitivity	0.774	0.762	0.789
Specificity	0.811	0.826	0.825
Classification accuracy	0.796	0.796	0.809
Kappa	0.589	0.590	0.612

Note: The Kappa index is a chance corrected proportion of classification accuracy. The volcano plots (Allison et al., 2006) of fold-changes and corresponding *p*-values for the Tagman quantitative PCR and AG1 platforms are available from the first author.

AG1 platform. The number of included probes on the AG1 array differs from that in the Tagman assay because, based on the available annotation files, a single transcript may be mapped by multiple probes on the AG1 array, whereas transcripts and probes are in one-to-one correspondence in Tagman assays. Figs. 5(b)–5(d) and 6(b)–6(d) present histograms and scatter plots of means versus standard deviations of transformed AG1 microarray intensities using the log_2, *glog*, and HP transformation methods respectively.

Differentially expressed RNAs on the Tagman and AG1 platforms were decided on by the criteria: (1) an absolute log_2 fold-change higher than 1 and (2) a *p*-value in the *t*-test lower than 10^{-3} (Allison et al., 2006). Table 5 summarizes the specificity, sensitivity, and classification accuracy computed by comparing between transformed AG1 and Tagman microarray data using the log_2, *glog* and HP transformation methods. The histograms in Fig. 5 depict all the probe values of the entire transformed AG1 and Tagman data, and the plots indicate that the HP transformed distribution is closer to normal than the other transformed distributions. According to Parsons et al. (2007), stabilizing the variance can improve the classification accuracy. The plots in Fig. 6 suggest that both log_2 and *glog* transformations are better variance stabilizers than the HP transformation especially for data expressed at low levels in this example.

It is known that microarray data have a complicated error structure (Durbin et al., 2002). If it can be assumed that all the probes on an array are separately independent observations (Purdom and Holmes, 2005), the HP transformation may bring the distribution of their expression values to normality such that the assumption underlying the two-sample *t*-test can be justified. Results in Table 5 suggest that the HP transformation yields the highest sensitivity (0.789) (i.e., differentially expressed RNAs are closer in the transformed Tagman and AG1 data sets), and gives the maximal classification accuracy (0.809) and Kappa index (0.612) with a slightly lower specificity (0.825) than that of the *glog* transformation (0.826). Perhaps the gain in sensitivity and classification accuracy is more likely attributable to using the standard linear model with (nearly) normal residuals, and less likely attributable to variance stabilization. To validate the findings in Table 5, however, a more thorough comparison between variance stabilizers and data transformation methods must be required with more microarray data sets.

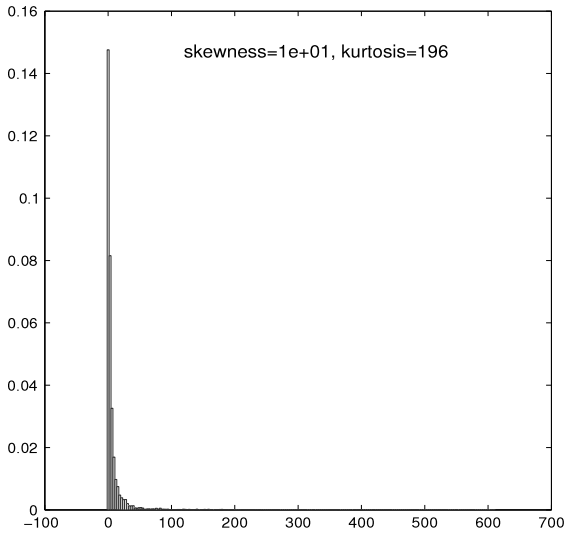
6. Conclusion

In linear regression analysis and ANOVA, it is accepted that the assumption of normally distributed residuals must be examined prior to testing hypotheses and constructing confidence intervals for parameter estimates in small samples. A common cure for failing this assumption is to use a modified method or acquire sufficient amount of data (for using a nonparametric method) such that inference with approximate normality can be legitimate. As an exceptional contrast, the example in Section 4 presents a practical case of an ANOVA application where a univariate distribution exhibits a bimodal shape. Even with a fairly large sample size, neither the standard *F*-test nor the two-sample *t*-test detects the gender difference in the raw score distribution. The proposed HP transformation provides a remedy to yield the correct inference; that is, the *F*-test based on the transformed scores is rectified to yield a significant gender effect, which is consistent with the result of the nonparametric two-sample rank test.

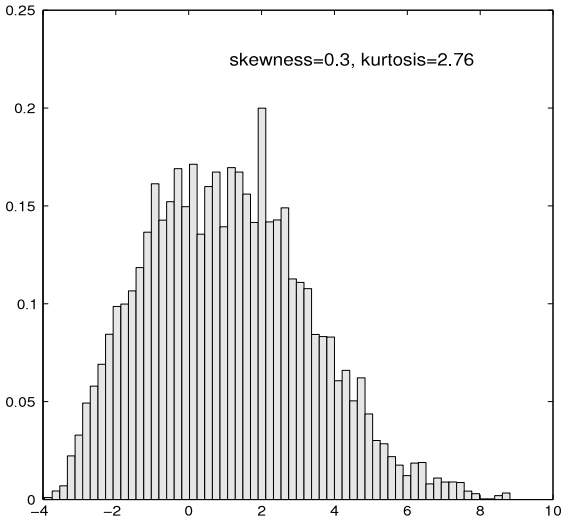
In the literature, the proposed HP transformation appears to be the first family that engages all possible cross combinations of concave and convex functions in the separate transformations on both sides of the median. Without using an extra location parameter, it is able to adjust for both skewness and kurtosis in the data, and also to transform bimodal (or mixture) distributions to normal as shown in the simulation study in Section 3, and as illustrated with the empirical examples in Sections 4 and 5. These key features of the HP transformation offer significant advantages over the existing families of transformations and suggest wider-ranging applications. For general multivariate data, it appears that transformation to multivariate normality via the HP transformation of marginal variables will be a challenging topic for future research.

Acknowledgments

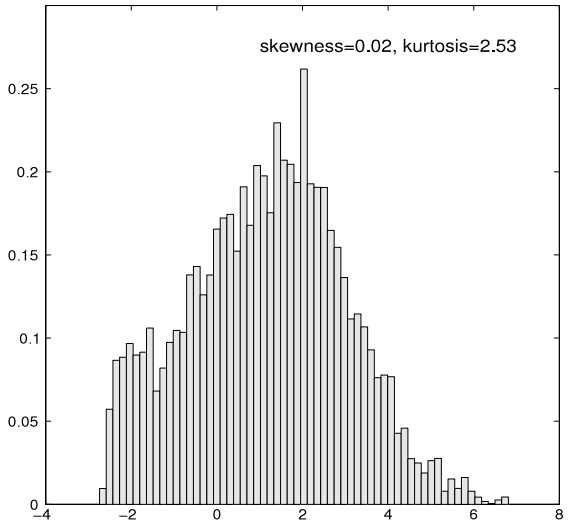
We would like to thank the Co-Editor, Associate Editor and Reviewers for helpful and constructive comments, which have helped with improving this study considerably. We also thank Mr. Shih-Kai Chu for helpful discussions and Mr. Chii-Shyang



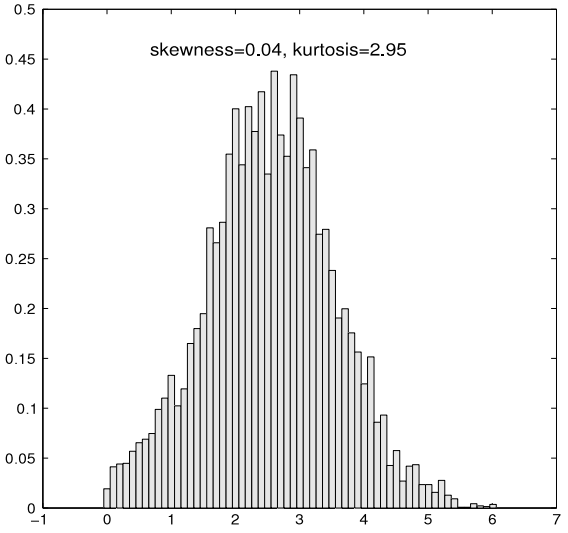
(a) Intensity of spotted microarray data.



(b) Intensity after $\log_2$ transformation.



(c) Intensity after glog transformation.



(d) Intensity after hyperbolic transformation.

Fig. 5. Histogram of (a) intensity of raw AG1 microarray data, intensities after (b) log_2, (c) *glog*, and (d) HP transformation, respectively. The Jarque–Bera test suggests that the null hypothesis of normality of (a)–(c) is rejected, each with a p-value $<.001$. The *glog* transformation is determined by the equation: $glog(x, \alpha, \lambda) = log[x - \alpha + \sqrt{(x-\alpha)^2 + \lambda}]$ with parameters $\alpha = 0.242$ and $\lambda = 0.029$ estimated by using the method and software developed by Geller et al. (2003). The HP transformation is performed on the log_2 transformed data with parameter estimates $\alpha = 0.64$, $\beta_- = 1.5$, $\lambda_- = 0.41$, $\beta_+ = 0.00063$, $\lambda_+ = 0.22$ and the Jarque–Bera test gives the p-value .101.

Kuo for help with analyzing the microarray data. This work was supported by the Taiwan Ministry of Science and Technology under research grants 103-2410-H-001-058-MY2, 104-2118-M-001-012, and 105-2118-M-001-011.

Appendix

The proof and theory of Lemma 1 below are confined to the open set $\Theta = \{(\beta, \lambda) : \beta > 0, \lambda < 1\}$ of the entire parameter space. The boundary subset $\lambda = 1.0$ (with its accompanying β values) only arises as desired solutions to approximating the maximum likelihood of formula (8) when using the simplex method. The parameter space Θ actually represents the original parameters $\theta = \{\alpha, \beta_-, \beta_+, \lambda_-, \lambda_+\}$ on both sides of the median, and α is a function of (β, λ) by formula (9).

Lemma 1. *Let $X_1, X_2, \ldots X_n$ be an i.i.d. random sample from $p(x, \theta)$ defined for all x on the real line. It can be shown that* (i) *the distributions $p(x, \theta)$ (including the additional boundary set $\lambda = 1.0$) are distinct, and* (ii) *for almost all x, $p(x, \theta)$ is differentiable*

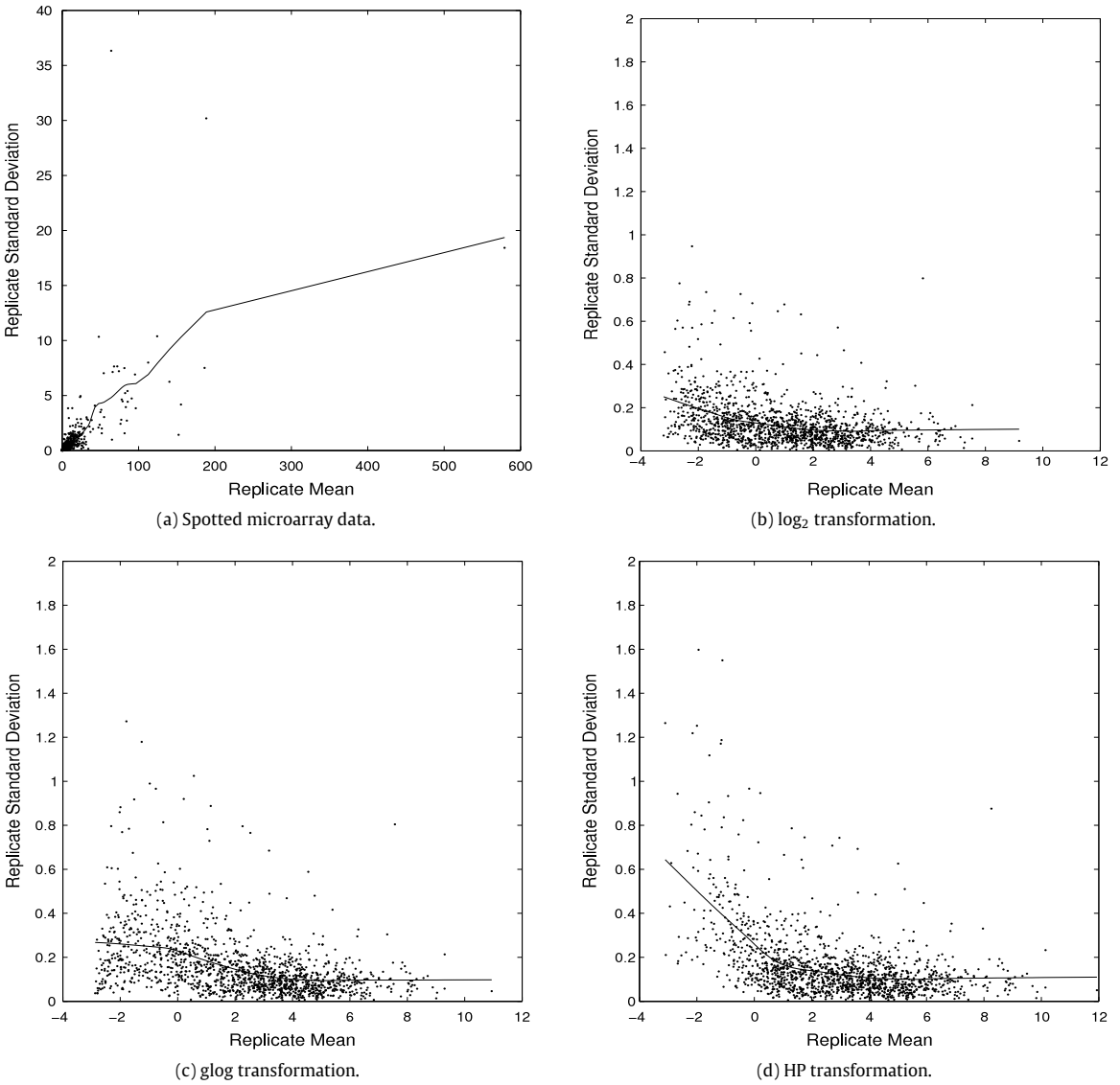


Fig. 6. Mean and standard deviation of replicates for the (a) raw AG1 microarray data, (b) log_2, (c) *glog*, and (d) HP transformed observations, respectively. Each point shows the mean and standard deviation of 5 replicates of sample A. The continuous black curve shows the overall trend line as estimated by LOESS regression.

with respect to θ *in the open set* $\Theta = \{(\beta, \lambda) : \beta > 0, \lambda < 1\}$. *Then, with probability tending to* 1 *as* $n \to \infty$, *the log-likelihood equation*

$$\frac{\partial}{\partial\theta} L(\theta|x) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} p(x_i|\theta)}{p(x_i|\theta)} = 0$$

(of partial derivatives with respect to (β, α)*) has a root* $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$ *such that* $\hat{\theta}_n(x_1, x_2, \ldots, x_n)$ *tends to the true value* $\theta_0 \in \Theta$ *(assumed to lie in* Θ*) in probability.*

Proof. It suffices to show the following two effects for the family of distributions $p(x, \theta)$, that is,
(i) $p(x, \theta_1) = p(x, \theta_2)$ for all x if, and only if, $\theta_1 = \theta_2$ in the entire parameter space $\{(\beta, \lambda) : \beta > 0, \lambda \leq 1\}$; and
(ii) continuous partial derivatives of $p(x, \theta)$ with respect to each x and θ exist in the open set $\Theta = \{(\beta, \lambda) : \beta > 0, \lambda < 1\}$.

To prove (i) with formula (7), the identifiability of $p(x, \theta)$ is valid if the identifiability of the absolute-valued Jacobian functions is verified. Thus, $p(x, \theta_1) = p(x, \theta_2)$ iff

$$\alpha_1 \left(1 - \lambda_1 \tanh^2(\beta_1 x)\right) \operatorname{sech}^{\lambda_1 - 1}(\beta_1 x) = \alpha_2 \left(1 - \lambda_2 \tanh^2(\beta_2 x)\right) \operatorname{sech}^{\lambda_2 - 1}(\beta_2 x), \tag{A.1}$$

for all x where $\theta_i = \{\alpha_i, \beta_i, \lambda_i\}$, $\alpha_i > 0$, $\beta_i > 0$, and $\lambda_i \leq 1$, for $i = 1, 2$. First, let $x = 0$ and Eq. (A.1) is reduced to the equation that $\alpha_1 = \alpha_2$. Thus, the constant factors α_i can be omitted from (A.1). We will show that, for any two pairs of parameters (β_1, λ_1) and (β_2, λ_2) satisfying (A.1) $\beta_1 = \beta_2$ and $\lambda_1 = \lambda_2 (\leq 1)$, respectively. Thus, it follows from (A.1) that

$$\frac{1 - \lambda_1 \tanh^2(x)}{1 - \lambda_2 \tanh^2(\beta x)} = \frac{\operatorname{sech}^{\lambda_2 - 1}(\beta x)}{\operatorname{sech}^{\lambda_1 - 1}(x)}, \tag{A.2}$$

where, without loss of generality, we may assume that $0 < \beta_1 < \beta_2$ and $\beta = \beta_2 / \beta_1$ by rescaling the arguments $\beta_1 x$ and $\beta_2 x$. That is, in (A.2), it is assumed that $\beta > 1$, and the next two distinct undesirable cases will be shown to be invalid.

(a) Assume $\lambda_1 < \lambda_2 \leq 1$. We will show that the validity of (A.2) for all x leads to a contradiction.

(b) Assume $\lambda_2 < \lambda_1 \leq 1$. We will also show that (A.2) fails to be valid.

Case (a). If $\lambda_1 < \lambda_2 \leq 1$, and $\beta > 1$, then $\tanh^2(\beta x) > \tanh^2(x)$, and the left hand side of (A.2) is greater than 1 for all $|x|$. By convexity of $\operatorname{sech}^{\lambda - 1}(x)$, $\lambda \leq 1$, the ratio on the right hand side of (A.2) is less than 1 for $|x|$ sufficiently large, and the right-hand side of (A.2) is less than 1. This contradicts (A.2).

Case (b). If $\lambda_2 < \lambda_1 \leq 1$, and $\beta > 1$, then, by similar argument, the left hand side of (A.2) is less than 1, for $|x|$ sufficiently small, but the right hand side of (A.2) is greater than 1, at least for a range of small $|x|$. This also contradicts (A.2).

From cases (a) and (b), we conclude that validity of (A2) implies that $\lambda_1 = \lambda_2$ and $\beta_1 = \beta_2$. Therefore, the HP family of transformations $\psi(x, \theta)$ in (1) are well-defined and distinct, that is, the members are uniquely defined by the pairs (β, λ) plus α.

Next, in view of Theorem 6.3.7 (Lehmann and Casella, 1998) it remains to prove (ii). Thus, it suffices to check that the individual summands of the partial derivatives given below are continuous functions on the real line:

$$\frac{\partial L}{\partial \lambda} = (1 - \psi^2) \log \operatorname{sech} + \frac{1}{\lambda - \coth^2}, \tag{A.3}$$

$$\frac{\partial L}{\partial \beta} = \psi^2 (x\lambda \tanh - x \coth + \beta^{-1}) - \frac{2x\lambda}{1 - \lambda \tanh^2} \tanh \operatorname{sech}^2 + x(1 - \lambda) \tanh, \tag{A.4}$$

and the elements in the Hessian matrix

$$\frac{\partial^2 L}{\partial \lambda^2} = -2(\psi \ln sech)^2 - \frac{1}{(\lambda - coth^2)^2} \tag{A.5}$$

$$\frac{\partial^2 L}{\partial \lambda \partial \beta} = x \tanh \left(\psi^2 - 1 - \frac{2\operatorname{sech}^2}{(1 - \lambda \tanh^2)^2} \right) + 2\psi^2 \log \operatorname{sech}(x\lambda \tanh - x \coth + \beta^{-1}). \tag{A.6}$$

$$\begin{aligned} \frac{\partial^2 L}{\partial \beta^2} &= -2\psi^2 (xcoth - x\lambda tanh - \beta^{-1})^2 + \psi^2 (x^2 csch^2 + x^2 \lambda sech^2 - \beta^{-2}) \\ &\quad + \frac{2x^2 \lambda sech^4}{1 - \lambda tanh^2} \left(\frac{2\lambda}{\lambda - coth^2} - 1 + 2sinh^2 \right) + x^2 (1 - \lambda) sech^2. \end{aligned} \tag{A.7}$$

The Newton–Raphson algorithm using the partial derivatives (A.3)–(A.7) constructs the essential computation of the MLEs in Section 2.2. The algorithm is quite effective but known to be sensitive to local extrema. The complicated forms of the second partial derivatives of the log-likelihood indicate that the MLE via Newton–Raphson iteration can be intricate. We have empirically conducted numerical solutions of the stationary points in the log-likelihood equation at $\partial L(\theta|x)/\partial\theta = 0$ for random samples of size 200 from lognormal, Weibull, gamma, exponential, Laplace, beta, uniform, bimodal, and Cauchy distributions. The negative definiteness of the Hessian matrix at stationary points of the log-likelihood equations has been checked. With the Weibull, gamma, beta, uniform, bimodal, or Cauchy distribution, the solution set exhibits two or more stationary points, among them only one is the global maximum. With the lognormal, exponential, or Laplace distributions, the global maximum may be located at the boundary of the parameter space ($\lambda = 1$) where $\partial L(\theta|x)/\partial\theta \neq 0$ and the conventional Newton–Raphson algorithm fails to find the maximum. Thus, the standard simplex method can be employed to facilitate the numerical approximation using a range of starting parameter estimates by perturbing estimates of the slope parameter α to ensure effective convergence to the global maximum. In the empirical study of these simulated and real data examples, the case of multiple local maxima on the log-likelihood surfaces was not encountered.

References

Allison, D., Cui, X., Page, G., Sabripour, M., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet 7 (1), 55–65.

Ambroise, J., Bearzatto, B., Robert, A., Govaerts, B., Macq, B.M., Gala, J.-L., 2011. Impact of the spotted microarray preprocessing method on fold-change compression and variance stability. BMC Bioinform. 12, 413.

Baker, G.A., 1934. Transformation of non-normal frequency distributions into normal distributions. Ann. Math. Statist. 5 (2), 113–123.
Barrett, T., Edgar, R., 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 411.
Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7 (6), 1129–1159.
Bickel, P., Doksum, K., 1981. An analysis of transformations revisited. J. Amer. Statist. Assoc. 76 (374), 296–311.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with discussion). J. R. Stat. Soc. Ser. A 26, 211–252.
Brown, M.B., Forsythe, A.B., 1974. Robust tests for the equality of variances. J. Amer. Statist. Assoc. 69 (346), 364–367.
Burbidge, J.B., Magee, L., Robb, A.L., 1988. Alternative transformations to handle extreme values of the dependent variable. J. Amer. Statist. Assoc. 83 (401), 123–127.
Cheng, P.E., Liou, M., Aston, J.A., Tsai, A.C., 2008. Information identities and testing hypotheses: Power analysis for contingency tables. Stat. Sinica 18 (2), 535.
D'Haese, S., De Meester, F., De Bourdeaudhuij, I., Deforche, B., Cardon, G., 2011. Criterion distances and environmental correlates of active commuting to school in children. Int. J. Behav. Nutr. Phys. Act 8 (1), 88.
Durbin, B., Hardin, J., Hawkins, D., Rocke, D., 2002. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18, S105–110.
Forbes, C., Evans, M., Hastings, N., Peacock, B., 2011. Statistical Distributions. Wiley, New York.
Gastwirth, J.L., Gel, Y.R., Miao, W., 2009. The impact of Levene's test of equality of variances on statistical theory and practice. Stat. Sci. 343–360.
Gel, Y.R., Gastwirth, J.L., 2008. A robust modification of the Jarque–Bera test of normality. Econ. Lett. 99 (1), 30–32.
Geller, S.C., Gregg, J.P., Hagerman, P., Rocke, D.M., 2003. Transformation and normalization of oligonucleotide microarray data. Bioinformatics 19 (14), 1817–1823.
Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Li, F.L.C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 5, R80.
Glass, G.V., Peckham, P.D., Sanders, J.R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev. Educ. Res. 237–288.
Greenacre, M., 2009. Power transformations in correspondence analysis. Comput. Statist. Data Anal. 53 (8), 3107–3116.
Hou, Q., Mahnken, J.D., Gajewski, B.J., Dunton, N., 2011. The Box-Cox power transformation on nursing sensitive indicators: Does it matter if structural effects are omitted during the estimation of the transformation parameter? BMC Med. Res. Methodol. 11 (1), 118.
Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., Vingron, M., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18, S96–104.
Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons.
John, J.A., Draper, N.R., 1980. An alternative family of transformations. Appl. Statist. 29, 190–197.
Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36 (1/2), 149–176.
Jones, M.C., Pewsey, A., 2009. Sinh–arcsinh distributions. Biometrika 96 (4), 761–780.
Kelmansky, D.M., Martínez, E.J., Leiva, V., 2013. A new variance stabilizing transformation for gene expression data analysis. Stat. Appl. Genet. Mol. Biol. 12 (6), 653–666.
Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. SIAM J. Optim. 9 (1), 112–147.
Lee, T., Girolami, M., Sejnowski, T., 1999. Independent component analysis using an extended infomax algorithm for mixed subGaussian and superGaussian sources. Neural Comput. 11 (2), 417–441.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer, Science & Business Media.
Lehmann, E.L., D'Abrera, H.J., 2006. Nonparametrics: Statistical methods based on ranks. Springer, New York.
Leiva, V., Sanhueza, A., Kelmansky, D.M., Martínez, E.J., 2009. On the glog-normal distribution and its association with the gene expression problem. Comput. Statist. Data Anal. 53 (5), 1613–1621.
Levene, H., 1960. Robust tests for equality of variances. In: Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 2, pp. 278–292.
Leys, C., Schumann, S., 2010. A nonparametric method to analyze interactions: The adjusted rank transform test. J. Exp. Soc. Psychol. 46 (4), 684–688.
Liermann, M., Steel, A., Rosing, M., Guttorp, P., 2004. Random denominators and the analysis of ratio data. Environ. Ecol. Stat. 11 (1), 55–71.
Manly, B.F, 1976. Exponential data transformation. Statistician, 25. pp. 37–42.
Mudholkar, G.S., Marchetti, C.E., Lin, C.T., 2002. Independence characterizations and testing normality against restricted skewness–kurtosis alternatives. J. Statist. Plann. Inference 104 (2), 485–501.
Munson, P., 2001. A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In: GeneLogic Workshop of Low Level Analysis of Affymetrix GeneChip Data, Bethesda, MD.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. (ISSN: 1460-2067) 7 (4), 308–313.
Osborne, J., 2002. Notes on the use of data transformations. Pract. Assess. Res. Eval. 8 (6), 1–8.
Parrish, R.S., Spencer, H.J., Xu, P., 2009. Distribution modeling and simulation of gene expression data. Comput. Statist. Data Anal. 53 (5), 1650–1660.
Parsons, H.M., Ludwig, C., Günther, U.L., Viant, M.R., 2007. Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinform. 8 (1), 234.
Pattyn, E., Van Praag, L., Verhaeghe, M., Levecque, K., Bracke, P., 2011. The association between residential area characteristics and mental health outcomes among men and women in Belgium. Arch. Public Health 69 (1), 1–11.
Purdom, E., Holmes, S.P., 2005. Error distribution for gene expression data. Stat. Appl. Genet. Mol. Biol. 4 (1).
Rakhshan, A., Pishro-Nik, H., 2014. Introduction to simulation using MATLAB. In: Pishro-Nik, H. (Ed.), Introduction to Probability, Statistics, and Random Processes: Statistics and Random Processes. Kappa Research, LLC, ISBN: 9780990637202, pp. 703–723 (Chapter 12).
Rasch, D., Kubinger, K.D., Moder, K., 2011. The two-sample t test: pre-testing its assumptions does not pay off. Stat. Pap. 52 (1), 219–231.
Sakia, R., 1992. The Box-Cox transformation technique: a review. Statistician 41 (2), 169–178.
Sawilowsky, S.S., Blair, R.C., 1992. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychol. Bull. 111 (2), 352–360.
Schmider, E., Ziegler, M., Danay, E., Beyer, L., Bühner, M., 2010. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology 6, 147–151.
Shi, L., Reid, L., Jones, W., Shippy, R., Warrington, J., Baker, S., Collins, P., de Longueville, F., et al., 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnol. 24 (9), 1151–1161.
Stehlík, M., Střelec, L., Thulin, M., 2014. On robust testing for normality in chemometrics. Chemom. Intell. Lab. Syst. 130, 98–108.
Strauss, R.E., 2012. MATLAB statistical functions [computer software]. http://www.faculty.biol.ttu.edu/Strauss/Matlab/matlab.htm.
Tomarken, A.J., Serlin, R.C., 1986. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol. Bull. 99 (1), 90–99.
Vilca, F., Rodrigues-Motta, M., Leiva, V., 2013. On a variance stabilizing model and its application to genomic data. J. Appl. Statist. 40 (11), 2354–2371.
Welch, B.L., 1947. The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34, 28–35.

Wen, Z., Wang, C., Shi, Q., Huang, Y., Su, Z., Hong, H., Tong, W., Shi, L., 2010. Evaluation of gene expression data generated from expired Affymetrix GeneChip microarrays using MAQC reference RNA samples. BMC Bioinform. 11 (Suppl. 6).
Yeo, I.-K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87 (4), 954–959.