HAL
open science

Statistical properties of the quantile normalization method for density curve alignment

Santiago Gallón, Jean-Michel Loubes, Elie Maza

To cite this version:

Santiago Gallón, Jean-Michel Loubes, Elie Maza. Statistical properties of the quantile normalization method for density curve alignment. Mathematical Biosciences, 2013, vol. 242 (n° 2), pp. 129-142. 10.1016/j.mbs.2012.12.007 . hal-01135106

HAL Id: hal-01135106

https://hal.science/hal-01135106v1

Submitted on 24 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive Toulouse Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: http://oatao.univ-toulouse.fr/
Eprints ID: 13737

Identification number: DOI: 10.1016/j.mbs.2012.12.007
Official URL: http://dx.doi.org/10.1016/j.mbs.2012.12.007

To cite this version:
Gallón, Santiago and Loubes, Jean-Michel and Maza, Elie *Statistical properties of the quantile normalization method for density curve alignment*. (2013) Mathematical Biosciences, vol. 242 (n° 2). pp. 129-142. ISSN 0025-5564

Any correspondence concerning this service should be sent to the repository administrator:
staff-oatao@inp-toulouse.fr

Statistical properties of the quantile normalization method for density curve alignment

Santiago Gallón [a,b], Jean-Michel Loubes [b,*], Elie Maza [c]

[a] Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia, cl. 67 #53-108, blq. 13, 050010 Medellín, Colombia
[b] Institut de Mathámatiques de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
[c] Genomic & Biotechnology of the Fruit Laboratory, UMR 990 INRA/INP-ENSAT, Toulouse, France

A B S T R A C T

The article investigates the large sample properties of the quantile normalization method by Bolstad et al. (2003) [4] which has become one of the most popular methods to align density curves in microarray data analysis. We prove consistency of this method which is viewed as a particular case of the structural expectation procedure for curve alignment, which corresponds to a notion of barycenter of measures in the Wasserstein space. Moreover, we show that, this method fails in some case of mixtures, and we propose a new methodology to cope with this issue.

Keywords:
Curve registration
Manifold registration
Microarray data analysis
Normalization
Order statistics
Structural expectation
Wasserstein distance

1. Introduction

We consider a density estimation problem in the particular situation where the data are samples of density curves, observed with some variations which are not directly correlated to the studied phenomenon. This situation occurs often in biology, for example when considering gene expression data obtained from microarray technologies, which is used to measure genome wide expression levels of genes in a given organism. A microarray may contain thousands of spots, each one containing a few million copies of identical DNA molecules that uniquely correspond to a gene. From each spot, a measure is obtained and then one of the most popular applications is to compare gene expression levels on different conditions, which leads to millions of measures of gene expression levels on technical and biological samples. However, before performing any statistical analysis on such data, it is necessary to process raw data in order to remove any systematic bias inhering to the microarray technology: differential efficiency of the two fluorescent dyes, different amounts of starting mRNA material, background noise, hybridization reactions and conditions. A natural way to handle this phenomena is to try remove these variations in order to align the measured densities, which proves difficult since the densities are unknown. In bioinformatics and computational biology, a method to reduce this kind of variability is known as normalization.

Among the normalization methods, the quantile normalization proposed by Bolstad et al. [4] has received a considerable interest. The procedure consists in assuming that there exists an underlying common distribution followed by the measures. Then, for $i = 1, \ldots, m$ samples of $j = 1, \ldots, n$ i.i.d random variables $X_{i,j}$, the *mean* distribution is achieved by projecting the jth empirical vector of sample quantiles, $\hat{\mathbf{q}}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{m,j})^\top$, onto the vector $\mathbf{d} = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$. This gives $\mathrm{proj}_{\mathbf{d}}\hat{\mathbf{q}}_j = (m^{-1}\sum_{i=1}^m \hat{q}_{i,j}, \ldots, m^{-1}\sum_{i=1}^m \hat{q}_{i,j})^\top$, which is such that if all m data vectors $X_i, i = 1, \ldots, m$, share the same distribution, then the plot of the quantiles gives a straight line along the line $\mathbf{d}$. We refer to [4,12] for some applications of this method. An example of this method is given in Fig. 1, where the densities of a sample of 18 two-color microarrays are plotted after normalization of the expression log-ratios within two-color arrays (see [24]). The dot-dashed and solid lines through densities corresponds to cross-sectional mean and quantile normalization of the log intensities across the arrays, respectively. The quantile normalization method has the advantages to be simple and quick with respect to others normalization procedures and yet providing very good estimation results. However its statistical properties have not been derived yet up to our knowledge.

Actually, normalization of density samples may be seen as the empirical version of a warping problem between distribution functions. This issue has received a growing attention in the last decade

* Corresponding author.
E-mail addresses: santiagog@udea.edu.co (S. Gallón), jean-michel.loubes@math.univ-toulouse.fr (J.-M. Loubes), elie.maza@ensat.fr (E. Maza).

http://dx.doi.org/10.1016/j.mbs.2012.12.007

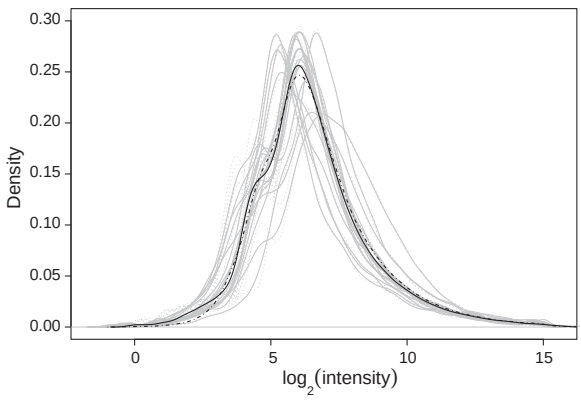

Fig. 1. Densities for individual-channel intensities for two-color microarray data after normalization within arrays. Dotted and solid gray lines correspond to the "green" and "red" color arrays, respectively.

where many authors tackle the problem of recovering an unknown curve observed with both amplitude (variation in the y-axis) or phase (variation in the x-axis) variations, which prevent any direct extraction of classical statistics such as the mean or the median. Indeed the classical cross-sectional mean does not provide a consistent estimate of the function of interest when the phase variations are ignored since it fails to capture the characteristics of the sample of curves as quoted in [18]. Therefore curve registration (also called curve alignment, structural averaging, and time warping) methods have been proposed in the statistical literature, among them we refer, for example, to [14,20,11,28,29,18,17,10,13,15,9] and references therein.

Hence, in this paper we point out that the quantile normalization can be seen as a particular case of the structural mean procedure, described in [9], which corresponds to a notion of barycenter of measures in the Wasserstein space as described in Boissard et al. [3]. We study the large sample properties of the quantile normalization method. In addition, when this procedure fails, using the analogy with warping issues, we propose a variation of this method to still recover a mean density and thus improving one pointed drawback of the quantile normalization method.

The outline of this article is as follows. In Section 2, we describe a nonparametric warping functional model which will be used to relate with the quantile normalization method. In Section 3 we present the quantile estimation and derive the asymptotic properties of the quantile normalization. Section 4 presents the relationship between normalization and distribution function alignment, which enables to improve quantile normalization method. Simulations are shown in Section 5. In Section 6, we apply the methods to normalize two-channel spotted microarray densities and evaluate its utility to identify differentially expressed genes. Finally, some conclusions are given in Section 7. All the proofs are gathered in Appendix A.

2. Statistical model for density warping

Let $X_{i,j},\ i=1,\dots,m,\ j=1,\dots,n_i$ be a sample of m independent real valued random variables of size n_i with density function $f_i:\mathbb{R}\to[0,+\infty)$ and distribution function $F_i:\mathbb{R}\to[0,1]$. We assume without loss of generality that $n_i=n$ for all units $i=1,\dots,m$. The random variables are assumed to model the same phenomena with a variation effect modeled as follows.

Each distribution function F_i is obtained by warping a common distribution function $F:\mathbb{R}\to[0,1]$ by an invertible and differentiable warping function H_i, of the following manner:

$$F_i(t)=\Pr(X_{i,j}\leqslant t)=F\circ H_i^{-1}(t) \tag{1}$$

where H_i is random, in the sense that $(H_1,\dots,H_m)$ is an i.i.d random sample from a (non parametric) warping stochastic process $\mathcal{H}:\Omega\to\mathcal{C}(\mathbb{R})$ defined on an unknown probability space $(\Omega,\mathcal{A},\mathbf{P})$, while $\mathcal{C}(\mathbb{R})$ denotes the space of all continuous functions defined on $\mathbb{R}$. Define ϕ its mean and let ϑ be its variance which is assumed to be finite. This model is also considered in [10,9].

Since the model (1) to estimate the function f is not identifiable (see [9]), we consider the *structural expectation (SE)* of the quantile function to overcome this problem as

$$q_{SE}(\alpha):=F_{SE}^{-1}(\alpha)=\phi\circ F^{-1}(\alpha),\quad 0\leqslant\alpha\leqslant 1. \tag{2}$$

Inverting Eq. (1) leads to

$$q_i(\alpha)=F_i^{-1}(\alpha)=H_i\circ F^{-1}(\alpha),\quad 0\leqslant\alpha\leqslant 1, \tag{3}$$

where $q_i(\alpha)$ is the population quantile function (the left continuous generalized inverse of F_i), $F_i^{-1}:[0,1]\to\mathbb{R}$, given by

$$q_i(\alpha)=F_i^{-1}(\alpha)=\inf\left\{x_{ij}\in\mathbb{R}:F_i(x_{ij})\geqslant\alpha\right\},\quad 0\leqslant\alpha\leqslant 1.$$

Hence the natural estimator of the structural expectation (2) is given by

$$\overline{q_m(\alpha)}=\frac{1}{m}\sum_{i=1}^{m}q_i(\alpha),\quad 0\leqslant\alpha\leqslant 1. \tag{4}$$

In order to get the asymptotic behavior of the estimator, the following assumptions on the warping process $\mathcal{H}$ and on the distribution function F are considered:

A1. There exists a constant $C_1>0$ such that for all $(\alpha,\beta)\in[0,1]^2$, we have

$$\mathbf{E}\left[\left|H(\alpha)-\mathbf{E}H(\alpha)-(H(\beta)-\mathbf{E}H(\beta))\right|^2\right]\leqslant C_1|\alpha-\beta|^2.$$

A2. There exists a constant $C_2>0$ such that, for all $(\alpha,\beta)\in[0,1]^2$, we have

$$\mathbf{E}\left[\left|F^{-1}(\alpha)-F^{-1}(\beta)\right|^2\right]\leqslant C_2|\alpha-\beta|^2.$$

The following theorem deals with the asymptotic behavior of the estimator (3).

Theorem 1. *The estimator $\overline{q_m(\alpha)}$ is consistent is the sense that*

$$\left\|\overline{q_m(\alpha)}-\mathbf{E}\left(\overline{q_m(\alpha)}\right)\right\|_\infty=\left\|\overline{q_m(\alpha)}-q_{SE}(\alpha)\right\|_\infty\underset{m\to\infty}{\overset{a.s.}{\to}}0.$$

Moreover, under Assumptions A1 and A2, the estimator is asymptotically Gaussian, for any $K\in\mathbb{N}$ and fixed $(\alpha_1,\dots,\alpha_K)\in[0,1]^K$,

$$\sqrt{m}\begin{bmatrix}\overline{q_m(\alpha_1)}-q_{SE}(\alpha_1)\\ \vdots\\ \overline{q_m(\alpha_K)}-q_{SE}(\alpha_K)\end{bmatrix}\underset{m\to\infty}{\overset{\mathcal{D}}{\to}}\mathcal{N}_K(\mathbf{0},\mathbf{\Sigma})$$

where $\mathbf{\Sigma}$ is the asymptotic variance–covariance matrix whose (k,k')-element is given by $\Sigma_{k,k'}=\vartheta(q(\alpha_k),q(\alpha_{k'}))$ for all $(\alpha_k,\alpha_{k'})\in[0,1]^2$ with $\alpha_k<\alpha_{k'}$.

3. Quantile estimation and the quantile normalization method

The distribution function is not observed and only random samples $X_{i,1},\dots,X_{i,n}$ from $F_i(x)$ for $i=1,\dots,m$ are observed. The ith empirical quantile function is a natural estimator of F_i^{-1} when there is not any information on the underlying distribution function F_i. Consider the order statistics $X_{i,1:n}\leqslant X_{i,2:n}\leqslant\cdots\leqslant X_{i,n:n}$, hence the estimation of the quantile functions, $q_i(\alpha)$, is obtained by

$$\hat{q}_{i,n}(\alpha) = \mathbb{F}_{i,n}^{-1}(\alpha) = \inf\{x_{ij} \in \mathbb{R} : \mathbb{F}_{i,n}(x_{ij}) \geqslant \alpha\} = X_{i,j:n}$$
$$\text{for} \quad \frac{j-1}{n} < \alpha \leqslant \frac{j}{n}, \quad j = 1, \ldots, n,$$

where $\mathbb{F}_{i,n}^{-1}$ is the ith empirical quantile function.

Finally, the estimator of the structural quantile is given by

$$\bar{\hat{q}}_j = \frac{1}{m}\sum_{i=1}^{m}\hat{q}_{i,j} = \frac{1}{m}\sum_{i=1}^{m}X_{i,j:n}, \quad j = 1, \ldots, n. \tag{5}$$

Note that, this procedure corresponds to the so-called quantile normalization method proposed by Bolstad et al. [4].

Based on sample quantiles we can obtain a "mean" distribution by mean the projection of the empirical quantile vector of the jth sample quantiles, $\hat{\mathbf{q}}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{m,j})^\top$, onto the m-vector $\mathbf{d} = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$, given by $\text{proj}_d\hat{\mathbf{q}}_j = (m^{-1}\sum_{i=1}^{m}\hat{q}_{i,j}, \ldots, m^{-1}\sum_{i=1}^{m}\hat{q}_{i,j})^\top$. The quantile normalization method can be understood as a quantile–quantile plot extended to m dimensions such that if all m data vectors share the same distribution, then the plot of the quantiles gives a straight line along the line $\mathbf{d}$.

The asymptotic behavior of the quantile normalization estimator (5) is established by the next theorem.

Theorem 2. *The quantile normalization estimator $\bar{\hat{q}}_j$ is strongly consistent, $\bar{\hat{q}}_j \underset{m,n\to\infty}{\overset{a.s}{\to}} q_{SE}(\alpha_j)$, $j = 1, \ldots, n$. Also, under the assumptions of compactly central data, $|X_{i,j:n} - \mathbf{E}(X_{i,j:n})| \leqslant L < \infty$ for all i and j, and $\sqrt{m}/n \to 0$, it is asymptotically Gaussian. Actually, for any $K \in \mathbb{N}$ and fixed $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$,*

$$\sqrt{m}\begin{bmatrix} \bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1) \\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K) \end{bmatrix} \underset{m,n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}_K(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the asymptotic variance–covariance matrix whose (k,k')-element is given by $\Sigma_{k,k'} = \vartheta(q(\alpha_k), q(\alpha_{k'}))$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$.

This theorem relies on the asymptotic behavior of the quantile estimator, $\hat{q}_{i,n}(\alpha)$, given by the following proposition.

Proposition 1. *Assume F_i is continuously differentiable at the αth population quantile $q_i(\alpha)$ which is the unique solution of $F_i(q_i(\alpha)-) \leqslant \alpha \leqslant F_i(q_i(\alpha))$, and $f_i(q_i(\alpha)) > 0$ for a fixed $0 < \alpha < 1$. Also assume $n^{-1/2}(j/n - \alpha) = o(1)$. Then, for $i = 1, \ldots, m$, the estimator $\hat{q}_{i,n}(\alpha)$ is strongly consistent, $\hat{q}_{i,n}(\alpha) \underset{n\to\infty}{\overset{a.s.}{\to}} q_i(\alpha)$; and asymptotically Gaussian*

$$\sqrt{n}(X_{i,j:n} - H_i \circ q(\alpha)) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{\left(f \circ H_i^{-1}(H_i \circ q(\alpha)) \cdot \left(H_i^{-1}\right)'(H_i \circ q(\alpha))\right)^2}\right),$$

where $\left(H_i^{-1}\right)'(z) = dH_i^{-1}(z)/dz = \left\{H_i' \circ H_i^{-1}(z)\right\}^{-1}$.

4. Density alignment as a registration problem

As we have seen in the previous sections, quantile normalization amounts to finding a mean distribution that fits the data density. Indeed, if the distribution function were known, hence, given respectively F_i's the distribution functions and μ_i's the distributions of the i.i.d sample $X_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, the problem consists in finding a distribution function F and a probability μ which plays the role of a *mean* function but close enough to the data. This corresponds to the usual registration problem of the F_i's function restricted to the set of distribution functions.

One of the major issue in registration problem is to find the fitting criterion which enables to give a sense to the notion of mean of a sample of points. A natural criterion is in this framework given by the Wasserstein distance and this problem can be rewritten as finding a measure μ which minimizes

$$\mu \mapsto \frac{1}{m}\sum_{i=1}^{m}W_2^2(\mu, \mu_i), \tag{6}$$

where W_2^2 stands for the 2-Wasserstein distance

$$W_2^2(\mu, \mu_i) = \int \left|F_i^{-1}(\alpha) - F^{-1}(\alpha)\right|^2 d\alpha.$$

The existence and the uniqueness of such a minimizer is a difficult task in a general framework, which has been proved very recently under some technical conditions on the μ_i's in [1]. However, for 1 dimensional distributions, an explicit solution can be given, which corresponds to the *structural expectation* defined in [9]. Here, the F_i's and the μ_i's are not observed and only their empirical version are available. The estimation counterpart is considered in Section 3.

As pointed here, Wasserstein distance appears as a natural way to model distance between distribution functions which are warped one from another. Nevertheless, other criterion than (6) can be investigated. Indeed, for any distance d on the inverse of distribution functions, we can define a criterion to be minimized

$$F \mapsto \frac{1}{m}\sum_{i=1}^{m}d\left(F^{-1}, F_i^{-1}\right).$$

Each choice of d implies different properties for the minimizers. Recall that the choice of the L^2 loss corresponds to the Wasserstein distance between the distributions. Another choice, when dealing with warping problems, is to consider that the functional data belong to a non euclidean set, and to look for the most suitable corresponding distance. Hence, a natural framework is given by considering that the functions belong to a manifold using a manifold embedding and, in this context, the geodesic distance provides a natural way to compare two objects. This point of view has been developed in [7] where $\hat{d}_g$, an approximation of the geodesic distance, is provided using an Isomap-type graph approximation, following [25]. This gives rise to the criterion

$$F \mapsto \frac{1}{m}\sum_{i=1}^{m}\hat{d}_g\left(F^{-1}, F_i^{-1}\right).$$

Only the approximation of the distribution function remains.

A theoretical study of this framework is difficult, mainly due to the problems of both choosing a good manifold embedding and then approximating the geodesic distance.

Many authors have considered this issue but results on the consistency of minimizers of such criterion are very scarce. Hence, we provide here a feasible algorithm to compute it and compare the performances of the corresponding estimator. For this, recall that we observe $X_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ random variables. In order to mimic the geodesic distance between the inverse of the distribution functions, we will directly estimate $F_i^{-1}(\alpha)$, for $j - 1/n < \alpha \leqslant j/n$ by the corresponding order statistics $X_{i,j:n}$. Hence, we sort the observations for each sample i, and denote by $X_{(i).}$ the sorted vector $X_{i,1:n}, \ldots, X_{i,n:n}$ and thus we obtain an array of sorted observations $(X_{(1).}, \ldots, X_{(m).})$. We then consider $\hat{d}_g$ an approximation of the geodesic distance between the vectors $X_{(i).}$ and define the corresponding geodesic mean as the minimizer over all the observation vectors $x \in \{X_{(i).}, \ i = 1, \ldots, m\}$ of the criterion

$$x \mapsto \frac{1}{m}\sum_{i=1}^{m}\hat{d}_g(x, X_{(i).}).$$

Even if the theoretical properties of this estimate are hard to understand due to the difficulties inherent to the graph-type geodesic approximation, its practical properties for density normalization will be studied in the next section. The software used for this estimation is available upon request.

5. Simulation study

In this section, we illustrate by mean of simulated data the cases in which the quantile normalization method by Bolstad et al. [4] works and the situation in which it has problems to represent properly the behavior of the sample of density curves.

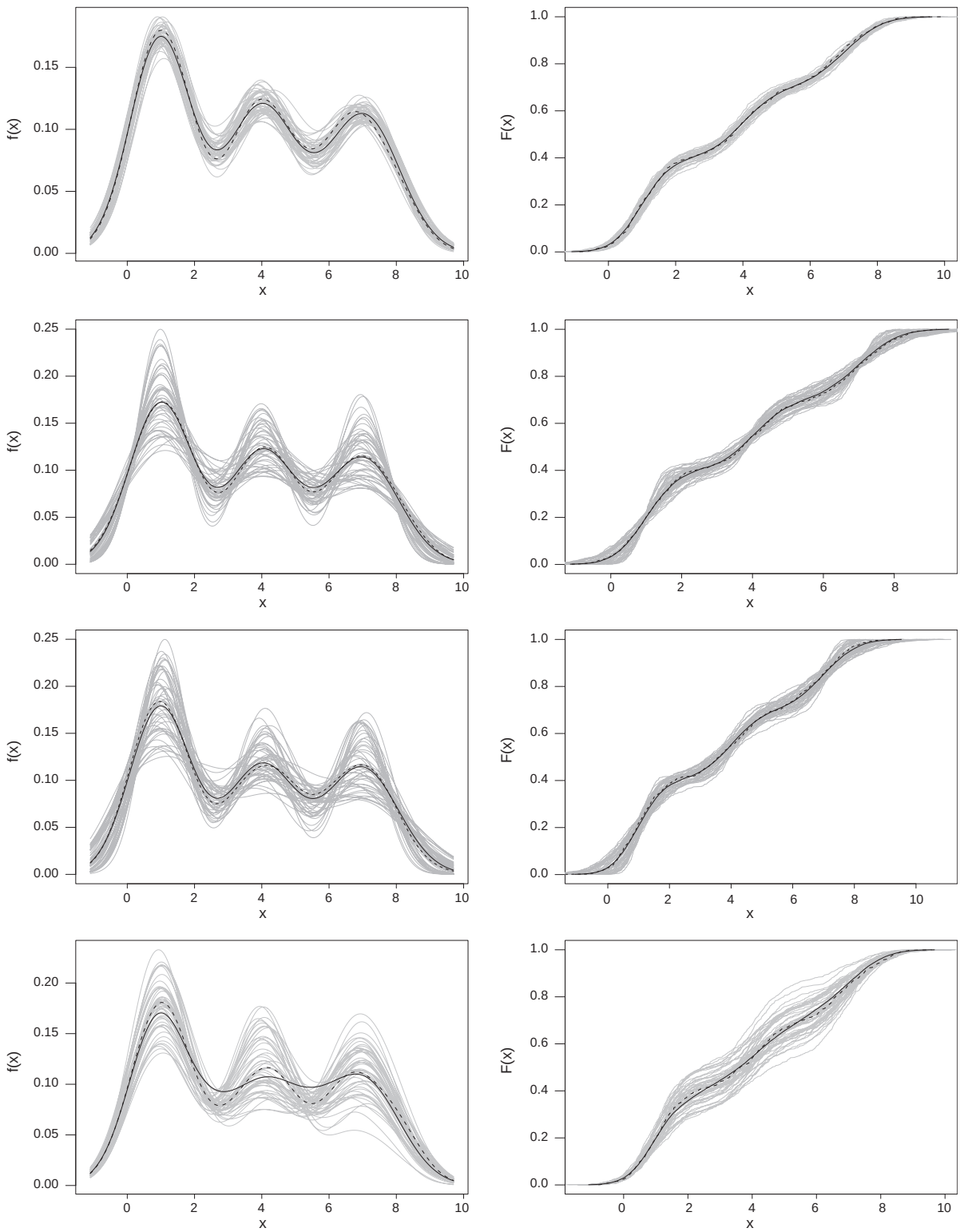

Fig. 2. Simulated density (left side) and distribution (right side) functions. Quantile (bold solid) and manifold (dash) normalizations. Cases 1–4 from the top to bottom.

We simulated a sample of m mixture density functions as linear combinations of three Gaussian probability density functions $\phi_{il}(x;\mu_{il},\sigma_{il}),\ l=1,2,3,$

$$f_i(x)=\sum_{l=1}^{3}\omega_{il}\phi_{il}(x;\mu_{il},\sigma_{il}),\quad i=1,\ldots,m,$$

where $\omega_{il}\in[0,1]$ are probability weights which satisfy $\sum_{l=1}^{3}\omega_{il}=1,\ i=1,\ldots,m.$

The simulated sample of mixture density functions were generated following the next procedure:

1. For each $i=1,\ldots,m$ three samples of size n of random observations are drawn from a Gaussian distribution.
2. A sampling (with replacement) of size n is carried out on the three samples based on the probability weights for obtaining the elements for each i.
3. Finally, for each i a kernel density estimate is obtained.

The values assumed to the location parameters were $\mu_{i1}=1,\ \mu_{i2}=4$ and $\mu_{i3}=7$; to the scale parameters $\sigma_{i1}=0.7,\ \sigma_{i2}=0.8$, and $\sigma_{i3}=0.9$; and to the probability weights $\omega_{i1}=0.4,\ \omega_{i2}=0.3$, and $\omega_{i3}=0.3$. The number of simulated curves and observations assumed were $m=50$ and $n=1000$ respectively. The variability for the sample of curves was generated according to the next cases:

Case 1 (*Location variations*).

$U(\mu_{il}-0.15,\mu_{il}+0.15)$ for $l=1,2,3.$

Case 2 (*Scale variations*).

$U(\sigma_{il}-0.35,\sigma_{il}+0.35)$ for $l=1,2,$ and $U(\sigma_{i3}-0.5,\sigma_{i3}+0.5)$ for $l=3.$

Case 3 (*Location and scale variations*).

Cases 1 and 2 together.

Case 4 (*Probability weight variations*).

For $l=1,2,\ U(\omega_{il}-0.1,\omega_{il}+0.1).$

where U is a uniformly distributed random variable.

Fig. 2 shows the simulated density and distribution functions for each case. The estimated "mean" density and distribution functions using the quantile and manifold normalization methods corresponds to the solid and dash lines, respectively. From the graphs, we can see that the quantile normalization estimate represents the variability among the density curves for the Cases 1–3, i.e when the probability weights do not vary among the densities, $\omega_{il}=\omega_{i'l},\ l=1,2,3$ for $i,\ i'=1,\ldots,m$. In Case 4, on the contrary, there are large differences between quantile and manifold normalization methods, where the based quantile method does not capture the structural characteristics across the set of densities.

To overcome the drawback corresponding to Case 4, we propose to apply the manifold embedding approach to estimate the structural mean pattern f based on an approximation of the induced geodesic distance on an unknown connected and geodesically complete Riemannian manifold $\mathcal{M}\subset\mathbb{R}^n$ by [7]. As we can see in Fig. 2, the estimation of the "mean" density f through the manifold normalization improves the normalization of the sample of densities for the case of variations in probability weights (Case 4) capturing properly the structural mean behavior of sample of curves.

6. Application to identification of differentially expressed genes

In this section, we apply the quantile and manifold methods to normalize two-channel (two-color) spotted microarrays in order to remove, from the expression measures, the systematic variations which arise from the microarray technology rather than from the differences between the probes, retaining the biological signals. For a description on two-channel spotted microarrays see [32,30]. We also evaluate the new manifold normalization method with respect to its ability to identify differentially expressed genes. For this we use two data sets of Tomato and Swirl experiments.

6.1. Tomato data set

The two-channel spotted microarray expression data comes from an experiment carried out by Wang et al. [27] in the Génomique et Biotechnologie des Fruits (GBF) laboratory at the Institut National Polytechnique-Ecole Nationale Superieure Agronomique de Toulouse (INP-ENSAT), which studies the underlying molecular mechanisms of the process of fruit set (i.e. the transition from flower-to-fruit) of tomato plants (*Solanum lycopersicum*). The data are provided by the experiment E-MEXP-1617 downloaded from the ArrayExpress database of functional genomic experiments at the European Bioinformatics Institute (EBI).

The data set contains 11,860 spots (probes) and 18 arrays. The Bioconductor `limma` package (http://www.bioconductor.org/) based on the `R` programming language was used to read and carry out the quality assessment of the intensity data [24,22]. Fig. 3 shows the density plots for individual-channel intensities of two-color microarrays. Dotted and solid lines correspond to densities of "green" and "red" color intensities for each array, respectively.

We normalize the two-channel microarray data applying the single-channel normalization method by Yang and Thorne [32], which removes the systematic intensity bias from the red and green channels separately, both within and between arrays. The method proceeds in two stages: a within-array normalization followed by a between-array (between all channels from multiple arrays) normalization. The first stage normalizes the expression log-ratios (M-values, $M=\log_2(R/G)$, where R and G are the red and green intensities, respectively) from two-color arrays such that these average to zero within each array separately. The advantage of using the log-ratios for measuring relative gene expression within two samples on the same slide rather than log-intensity values is due to these are considered to be more stable than the absolute intensities across slides [32]. The second stage normalizes the log intensities across arrays ensuring that these have the same empirical distribution across arrays and across channels. Procedures for

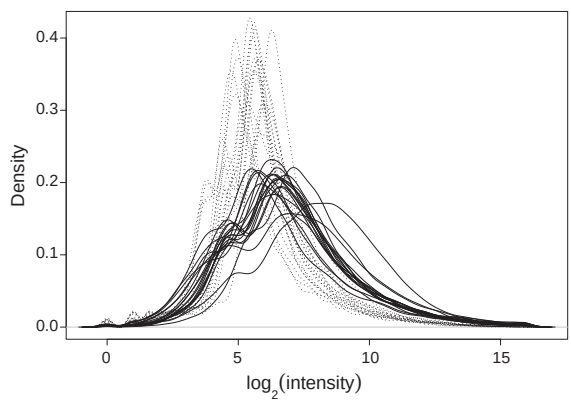

Fig. 3. Densities for individual-channel intensities for two-color microarray data. Dotted and solid lines correspond to the "green" and "red" color arrays, respectively.

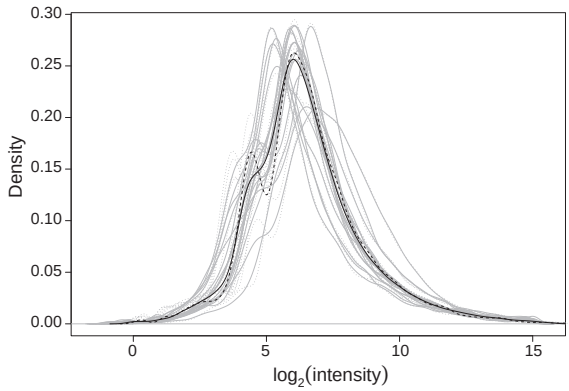

Fig. 4. Densities for individual-channel intensities for two-color microarray data after loess normalization within arrays. Solid and dashed lines correspond the normalization between arrays applying the quantile and manifold normalization, respectively.

Table 1
Number of differentially expressed genes identified for each stage of tomato fruit assuming an adjusted *p*-value less than 0.05.

	Tomato stage		
	Bud	Anthesis	Post-anthesis
Quantile	93	1291	262
Manifold	68	1274	254
Quantile ∩ Manifold	68	1250	252
Quantile − Manifold	25	41	10
Manifold − Quantile	0	24	2

$A - B$ denotes the difference set between sets A and B.
$A \cap B$ denotes the intersection between sets A and B.

Table 2
Bud stage: top 30 differentially expressed genes identified from Tomato data (quantile normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
3733	2.860	29.874	0.0001	7.958
9737	2.484	13.668	0.0110	5.246
12795	−0.897	−11.891	0.0172	4.533
10960	0.896	11.437	0.0172	4.324
10905	0.876	11.270	0.0172	4.244
5812	0.910	10.270	0.0227	3.725
8334	1.046	10.211	0.0227	3.692
6768	0.881	10.118	0.0227	3.640
8712	0.906	9.849	0.0244	3.485
6848	0.903	9.387	0.0305	3.205
9173	0.765	9.010	0.0308	2.964
7338	0.665	8.783	0.0308	2.811
2266	0.706	8.700	0.0308	2.755
12214	−0.814	−8.611	0.0308	2.693
2489	0.775	8.600	0.0308	2.686
3786	0.786	8.584	0.0308	2.674
4603	0.669	8.525	0.0308	2.633
3426	−0.752	−8.509	0.0308	2.622
7180	1.040	8.498	0.0308	2.614
4646	0.627	8.302	0.0308	2.473
12787	1.358	8.265	0.0308	2.447
4192	0.863	8.250	0.0308	2.435
7859	−0.842	−8.077	0.0308	2.308
7948	0.770	8.044	0.0308	2.283
6826	0.632	7.995	0.0308	2.245
2432	0.586	7.992	0.0308	2.243
11454	0.639	7.929	0.0308	2.195
6474	0.591	7.822	0.0308	2.113
87	0.691	7.822	0.0308	2.112
11019	−0.630	−7.796	0.0308	2.092

Table 3
Bud stage: top 30 differentially expressed genes identified from Tomato data (manifold normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
3733	2.854	18.769	0.0024	6.488
9737	2.371	13.394	0.0133	5.056
12795	−0.866	−11.721	0.0227	4.382
10905	0.815	10.681	0.0272	3.883
8334	0.989	10.566	0.0272	3.823
10960	0.860	10.336	0.0272	3.701
3786	0.769	9.184	0.0342	3.028
6768	0.828	9.166	0.0342	3.017
7338	0.662	8.999	0.0342	2.909
8712	0.878	8.933	0.0342	2.867
7859	−0.847	−8.846	0.0342	2.809
7180	1.057	8.748	0.0342	2.744
2266	0.678	8.742	0.0342	2.740
4646	0.622	8.547	0.0342	2.607
12214	−0.790	−8.542	0.0342	2.604
4603	0.649	8.517	0.0342	2.587
5812	0.886	8.401	0.0342	2.505
12787	1.317	8.387	0.0342	2.495
7948	0.757	8.102	0.0342	2.289
2489	0.750	8.079	0.0342	2.273
6826	0.616	8.077	0.0342	2.270
3426	−0.755	−8.068	0.0342	2.264
4192	0.817	8.051	0.0342	2.252
2432	0.572	8.048	0.0342	2.249
8254	0.882	7.972	0.0342	2.192
12181	−0.613	−7.910	0.0342	2.146
6848	0.842	7.847	0.0342	2.098
87	0.676	7.844	0.0342	2.096
9173	0.731	7.831	0.0342	2.085
6474	0.562	7.743	0.0342	2.018

within-array and between-array normalizations are implemented in the `normalizeWithinArrays` and `normalizeBetweenArrays` functions from the `limma` package, respectively.

The log-ratios within arrays were normalized using the loess method (see [24,30]). Fig. 4 plots the densities for each array after loess normalization. The normalization between arrays applying the quantile and manifold normalization are plotted in the same figure in solid and dashed lines. As we can see, the manifold normalization captures better the structural characteristics of the densities, in particular those that corresponding to the inflection points present in the individual arrays.

Now we evaluate the usefulness of the manifold normalization to identify differentially expressed genes. One of the aims of the tomato experiment in [27] is to identify gene expression in the (MicroTom) tomato lines downregulated in the expression of the Indole Acetic Acid 9 gene (AS-IAA9) and the wild type at three developmental stages during fruit set: flower bud, anthesis (i.e. the period during which the flower is fully open and functional), and post-anthesis. Thus, there are three experiments of identification of genes taking each tomato fruit stage separately. Hence, the experimental designs were based on six arrays for each corresponding stage, in two dye-swap pairs.

The statistical tool used for the identification of differentially expression genes in designed microarray experiments was the procedure based on the fit of gene-wise linear models and the application of empirical Bayes methods developed by Smyth [21]. The method relies on the "moderated" *t*-statistic across genes, a classical *t*-statistic improved by moderation of the standard errors, i.e., posterior estimators that shrunk the standard errors towards a common prior value using a Bayesian model (see [21,23] for details). The tables for each stage show the top 30 differentially expression genes identified using the expression intensities normalized with the quantile and manifold normalizations methods, respectively. In the subsequent tables (Table 2–10 are included, for each identified gene, the *M*-value, the moderate *t*-statistic,

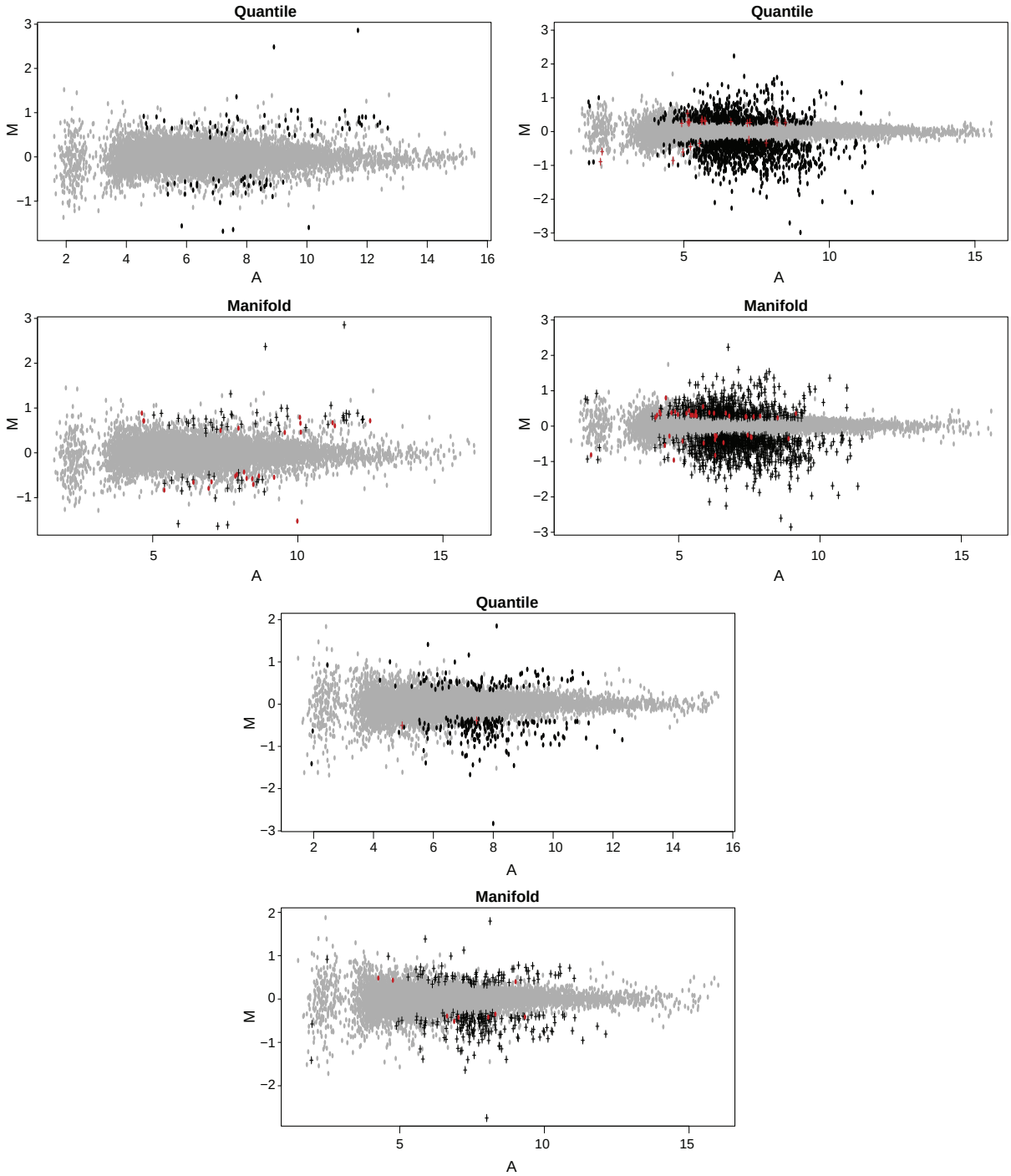

Fig. 5. Graphical illustration of the differentially expressed genes identified for bud (top-left), anthesis (top-right) and post-anthesis (bottom) stages using the normalized expressions with the quantile and manifold methods. Black points and pluses correspond to the detected genes with a assuming an adjusted *p*-value less than 0.05. The red points (pluses) symbols correspond to the genes identified with the quantile (manifold) normalization but not with the manifold (quantile) method (see Table 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the adjusted *p*-value and the *B*-statistic (log-odds that the gene is differentially expressed). The ranking of genes with significant differential expression are reported in order of increasing *B*-values. To adjust the *p*-values for multiple testing the Benjamini–Hochberg's method was used to control the expected false discovery rate (FDR) (see [23]). The number of differentially expressed genes detected, for each stage of the tomato fruit, by the use of normalized log-ratios through the two normalization methods are reported in Table 1, according to an assumed threshold value of 0.05 for adjusted *p*-values. In the same table are also reported the number of common genes shared by both methods, and the number of genes identified with the quantile (manifold) normalization but not with the manifold (quantile) method.

6.1.1. Bud stage

For the bud stage of tomato fruit, the number of differentially expressed genes identified employing the normalized expression log-ratios through quantile and manifold normalization methods were 93 and 68, respectively, with a common number of genes of 68. The top 30 of differentially expressed genes detected are

Table 4
Anthesis stage: top 30 differentially expressed genes identified from Tomato data (quantile normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
9306	−2.100	−35.992	0.0000	11.171
10855	1.390	20.624	0.0004	8.521
4051	1.600	19.425	0.0004	8.169
7075	1.435	19.289	0.0004	8.127
7	−1.488	−18.515	0.0004	7.880
8180	−1.474	−17.630	0.0004	7.578
7825	−2.703	−17.582	0.0004	7.561
6884	1.124	17.529	0.0004	7.542
12861	1.178	17.365	0.0004	7.483
8334	1.172	17.304	0.0004	7.461
7904	1.004	17.185	0.0004	7.418
10617	1.193	16.704	0.0004	7.238
9963	1.537	16.465	0.0004	7.146
1127	−1.801	−16.448	0.0004	7.139
4040	−2.265	−16.429	0.0004	7.132
12795	−2.985	−15.828	0.0005	6.891
7686	−0.972	−15.777	0.0005	6.871
6218	−0.992	−15.596	0.0005	6.795
9582	1.177	15.176	0.0005	6.617
12911	−1.804	−15.118	0.0005	6.592
9457	0.970	15.052	0.0005	6.563
132	−1.235	−15.045	0.0005	6.560
4312	1.156	14.946	0.0005	6.516
11406	0.952	14.742	0.0005	6.425
3117	−1.153	−14.714	0.0005	6.412
7164	0.948	14.659	0.0005	6.387
7118	0.841	14.604	0.0005	6.363
8241	−1.421	−14.426	0.0005	6.281
9876	−1.735	−14.260	0.0006	6.204
2339	−1.219	−14.234	0.0006	6.192

Table 5
Anthesis stage: top 30 differentially expressed genes identified from Tomato data (manifold normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
9306	−2.137	−34.987	0.0000	10.919
10855	1.334	21.035	0.0004	8.541
7075	1.383	19.027	0.0004	7.962
4051	1.537	18.535	0.0004	7.806
8180	−1.423	−17.529	0.0004	7.467
7	−1.424	−17.453	0.0004	7.440
6884	1.098	17.425	0.0004	7.430
7904	0.986	17.383	0.0004	7.415
12861	1.149	17.349	0.0004	7.403
8334	1.150	17.313	0.0004	7.390
10617	1.170	16.994	0.0004	7.275
7825	−2.604	−16.838	0.0004	7.217
9963	1.487	16.104	0.0006	6.936
4040	−2.253	−15.820	0.0006	6.822
12795	−2.852	−15.751	0.0006	6.794
7686	−0.952	−15.565	0.0006	6.718
6218	−0.964	−15.241	0.0006	6.582
4209	1.406	15.144	0.0006	6.541
7164	0.908	14.723	0.0006	6.357
11406	0.940	14.714	0.0006	6.353
1127	−1.700	−14.673	0.0006	6.335
9457	0.936	14.651	0.0006	6.325
7118	0.826	14.631	0.0006	6.316
3117	−1.126	−14.627	0.0006	6.314
12911	−1.752	−14.600	0.0006	6.302
8241	−1.395	−14.433	0.0006	6.226
2339	−1.158	−14.370	0.0006	6.198
132	−1.193	−14.346	0.0006	6.186
9876	−1.694	−14.251	0.0006	6.142
4312	1.118	14.241	0.0006	6.138

Table 6
Post-anthesis stage: top 30 differentially expressed genes identified from Tomato data (quantile normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
3161	−1.669	−21.432	0.0013	7.380
7953	−2.826	−19.875	0.0013	7.073
5626	−1.119	−15.296	0.0051	5.851
8339	−1.329	−14.746	0.0051	5.662
980	−1.439	−14.143	0.0054	5.441
4789	−0.850	−13.435	0.0064	5.163
11217	−1.455	−12.734	0.0064	4.866
914	−0.843	−12.533	0.0064	4.776
8590	−0.814	−12.436	0.0064	4.732
6927	−0.872	−12.339	0.0064	4.688
9637	−0.752	−12.265	0.0064	4.653
8912	−0.918	−11.934	0.0064	4.496
3211	−1.005	−11.833	0.0064	4.447
6253	−0.962	−11.809	0.0064	4.435
7038	−1.100	−11.629	0.0064	4.346
4040	−1.187	−11.549	0.0064	4.306
5405	−0.712	−11.475	0.0064	4.269
985	−0.723	−11.407	0.0064	4.234
2649	−0.753	−11.245	0.0067	4.150
5672	0.813	11.005	0.0073	4.022
12787	1.850	10.709	0.0080	3.860
7342	−0.808	−10.449	0.0082	3.713
6785	1.816	12.964	0.0079	3.711
2282	−0.800	−10.405	0.0082	3.688
6992	0.607	10.361	0.0082	3.662
11474	−0.799	−10.290	0.0082	3.621
10032	−1.138	−10.265	0.0082	3.606
8456	−0.838	−10.228	0.0082	3.584
6234	−0.965	−10.199	0.0082	3.566
6806	−0.805	−10.176	0.0082	3.553

Table 7
Post-anthesis stage: top 30 differentially expressed genes identified from Tomato data (manifold normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
3161	−1.637	−21.262	0.0012	7.305
7953	−2.744	−20.204	0.0012	7.099
5626	−1.080	−15.746	0.0045	5.961
8339	−1.295	−14.250	0.0054	5.447
980	−1.399	−14.218	0.0054	5.435
4789	−0.822	−13.017	0.0062	4.957
914	−0.806	−12.947	0.0062	4.927
11217	−1.393	−12.568	0.0062	4.761
8590	−0.796	−12.442	0.0062	4.704
7038	−1.151	−12.424	0.0062	4.696
6927	−0.841	−12.081	0.0062	4.537
9637	−0.733	−12.062	0.0062	4.528
3211	−0.986	−11.957	0.0062	4.477
6253	−0.934	−11.726	0.0063	4.365
8912	−0.880	−11.601	0.0063	4.304
985	−0.705	−11.536	0.0063	4.271
4040	−1.149	−11.394	0.0063	4.199
5405	−0.684	−11.337	0.0063	4.169
6785	1.731	14.589	0.0062	4.164
2649	−0.742	−11.270	0.0063	4.134
12787	1.796	11.231	0.0063	4.114
6234	−0.921	−10.715	0.0081	3.836
5672	0.764	10.585	0.0081	3.764
10240	0.572	10.413	0.0081	3.666
7342	−0.766	−10.407	0.0081	3.662
6806	−0.781	−10.350	0.0081	3.629
6992	0.580	10.311	0.0081	3.607
10032	−1.091	−10.311	0.0081	3.607
2282	−0.783	−10.300	0.0081	3.601
6497	−0.745	−10.187	0.0084	3.534

shown in Tables 2 and 3. The ordering of genes of first 30 genes is more or less parallel between both normalization methods. Some common genes have a quite different position, e.g. genes 5812, 6848, 9173, 3786, 7180, 4646 and 7859. Mostly of these top genes are common, except the genes 11,454 and 11,019 in the quantile normalization, and genes 8254 and 12181 in the manifold method.

Table 8
Validation by qRT-PCR of Tomato experiment.

Gene	Bud			Anthesis			Post-anthesis		
	Bolstad	Manifold	qRT-PCR	Bolstad	Manifold	qRT-PCR	Bolstad	Manifold	qRT-PCR
10318	no	no	yes	yes (88)	yes (96)	yes			
10617				yes (12)	yes (11)	yes			
10960	yes (4)	yes (6)	no						
12580	no	no	yes	yes (51)	yes (51)	yes			
12722				no	no	yes			
2266	yes (13)	yes (13)	yes	yes (103)	yes (77)	yes			
6848									
3733	yes (1)	yes (1)	yes	yes (145)	yes (137)	yes	yes (48)	yes (45)	yes
9737	yes (2)	yes (2)	yes						
9876				yes (29)	yes (29)	yes			
11243				yes (275)	yes (268)	yes			
12861				yes (9)	yes (9)	yes			
7468				yes (85)	yes (75)	yes			
7392				yes (303)	yes (299)	yes			
6299	no	no	yes	no	no	yes			
3161							yes (1)	yes (1)	yes
2	no	no	no	yes (147)	yes (171)	yes			
2020							yes (81)	yes (79)	yes
5626							yes (3)	yes (3)	no
2238							yes (158)	yes (155)	yes
12413	no	no	yes	yes (161)	yes (172)	yes			
10258				yes (44)	yes (40)	yes	yes (83)	yes (74)	yes
3043	yes (33)	yes (31)	yes	yes (83)	yes (80)	yes	yes (36)	yes (36)	yes
11189				yes (138)	yes (134)	yes			
2402				yes (109)	yes (103)	yes			
8241				yes (28)	yes (26)	no			
12911				yes (20)	yes (25)	yes	yes (35)	yes (39)	yes
10032				yes (115)	yes (110)	yes	yes (27)	yes (28)	yes

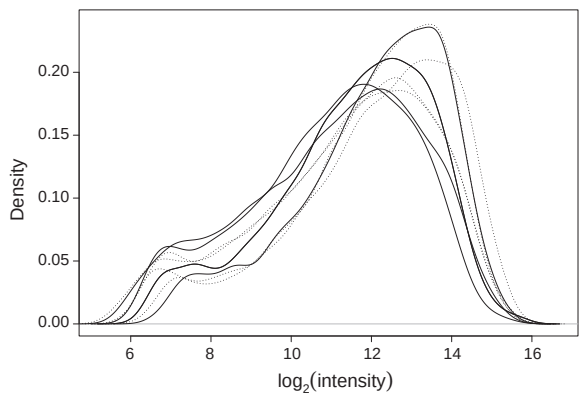

Fig. 6. Densities for individual-channel intensities for two-color microarray data. Dotted and solid lines correspond to the "green" and "red" color arrays, respectively.

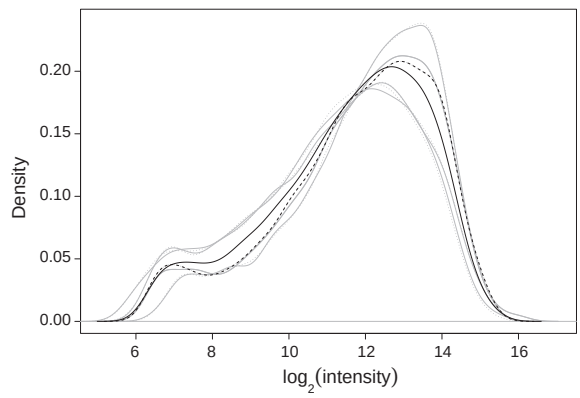

Fig. 7. Densities for individual-channel intensities for two-color microarray data after print-tip loess normalization within arrays. Solid and dashed lines correspond the normalization between arrays applying the quantile and manifold normalization, respectively.

Important features on genes can be found by means of the scatterplot between the average of $\log_2$ fold changes against the average of log-intensity $A = \log_2\sqrt{R \times G}$ for each probe over all arrays in the experiment (MA-plot). There are other plots over which the identification can be contrasted, e.g., scatterplots between the moderated t-statistics and the average of log-intensity A, between the B-statistics against average of $\log_2$ fold changes (volcano plot), and quantile–quantile plots of moderated t-statistics [31]. Although we choose the MA-plots to save space, the results were practically the same for these graphs. The MA-plot for the respective normalization method are in Fig. 5. The black symbols correspond to differentially expressed genes with adjusted p-value less than 0.05. From the MA-plots is clear that these symbols are well separated from the clouds such that the corresponding genes are likely to be differentially expressed [31]. The genes detected with the normalized expressions by the quantile normalization that are not identified with the manifold method are signed in red points for the manifold MA-plot (on the bottom) for comparison. The number of theses genes are reported in Table 1.

6.1.2. Anthesis stage

The number of differentially expressed genes detected in the anthesis stage with the expression intensities normalized with the quantile and manifold methods were 1291 and 1274, respectively, with 1250 genes in common. As is illustrated in Tables 4 and 5, the first 30 genes identified with both normalization methods are almost the same, except the gene 9582 for the quantile normalization and the gene 4209 for the manifold normalization. As in the bud stage, the position of genes in this ranking is fairly parallel for both methods, where only the position of genes 7825, 1127, 132, 4312 and 7164 is slightly different.

Although there are not big differences in the MA-plots between both normalization methods shown in Fig. 5, the identification

with the normalized intensities with the manifold method is a little bit sparser with respect to the quantile method, identifying 17 genes less. In the plots, the genes detected with the quantile normalization that are not identified with the manifold method are signed in red points for the manifold MA-plot (on the bottom), and the genes identified with the manifold normalization but not detected by the quantile method are represented by red pluses for the quantile MA-plot (on the top).

6.1.3. Post-anthesis stage

For the post-anthesis stage, 262 and 254 differentially expressed genes were identified with the quantile and manifold normalization, respectively, where the number of genes shared by both methods was 252. The top 30 genes are reported in Tables 6 and 7. As in the two previous stages, the position of these genes in both tables is parallel, especially for the top 10 genes. After, the position changes a little, specially for genes 7038, 6785, 2282, 6234 and 6806. There exist only four no common genes in the first 30 identified genes in both tables (gene 11,474 and 8456 for the quantile method and 10,240 and 6497 for the manifold normalization). The MA-plots for both methods are shown in Fig. 5. As in the bud and anthesis stages, mostly of detected genes are relatively far from the to zero line on the *M*-axis.

6.1.4. Validation by qRT-PCR data

Finally, in order to validate the microarray analysis of tomato data set in terms of the accuracy to detect differentially expressed genes during the fruit set using the normalized log ratios thought the quantile and manifold normalization methods, the results of a quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) analysis over 28 genes carried out by Wang et al. [27] were employed. In Table 8, the qRT-PCR column, for each of developmental stage of fruit set, indicates whether the corresponding analyzed gene was validated (categorized as "yes") or not in the analysis by Wang et al. [27]. The columns of the quantile and manifold methods indicates whether the respective gene was detected as differentially expressed in each stage. On the same Table, the numbers within parenthesis indicate the position of the identified gene in the ranking of genes.

The comparison between the identification results with the normalization methods and the validation results by the qRT-PCR shows that the detection of genes using normalized expressions with both of normalization methods have a good accuracy. In general, for all stages of fruit set, there is a proportion of 75.5% of favorable cases (i.e. identified gene matching with validated gene or not identified gene matching with not validated gene). Additionally, the manifold method seems to have a higher significance, identifying first the validated gene with respect to the quantile method; 33 cases of validated genes are identified fist by the manifold normalization (82.5%).

6.2. Swirl zebrafish data set

The same exercise of identification of differentially expressed genes was carried out with the popular Swirl data set, which can be downloaded from http://bioinf.wehi.edu.au/limmaGUI/DataSets.html. This experiment was conducted using zebrafish (*Brachydanio rerio*) as a model organism to study early development in vertebrates. Swirl is a point mutation in the BMP2 gene that affects the dorsal–ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and the notochord are expanded. One of the goals of the experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. See [8,31,21] for detailed information about this experiment. A total of four arrays were performed in two dye-swap pairs with 8448 probes. Smyth [21] normalized the expression of log-ratios within-arrays using the print-tip loess normalization with a window span of 0.3 and three robustifying iterations. We follow his method, but instead of between arrays scale normalization of log intensities, here, of course, the quantile and manifold normalization are applied to compare both methods. The Figs. 6 and 7 show the densities of unnormalized individual-channel intensities for two-color microarrays and its corresponding print-tip loess normalization within arrays, respectively. The solid and dashed lines in Fig. 7 correspond to the densities after normalization between arrays applying the quantile and manifold normalization, respectively.

Table 9
Top 30 differentially expressed genes identified from Swirl data (quantile normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
2961	−2.633	−17.198	0.0020	6.966
3723	−2.185	−16.415	0.0020	6.713
1611	−2.186	−15.736	0.0020	6.479
3721	−2.198	−14.329	0.0024	5.939
1609	−2.325	−13.710	0.0024	5.677
7602	1.210	13.121	0.0024	5.412
8295	1.306	13.070	0.0024	5.388
319	−1.265	−13.050	0.0024	5.378
515	1.308	12.835	0.0024	5.277
5075	1.373	12.795	0.0024	5.258
3790	1.187	12.356	0.0024	5.042
157	−1.792	−12.301	0.0024	5.014
7307	1.228	12.253	0.0024	4.989
7036	1.376	12.018	0.0024	4.869
2276	1.253	11.978	0.0024	4.848
7491	1.353	11.907	0.0024	4.810
3726	−1.280	−11.873	0.0024	4.792
5931	−1.091	−11.857	0.0024	4.784
683	1.350	11.657	0.0026	4.676
1697	1.119	11.534	0.0026	4.609
4380	1.265	11.415	0.0027	4.543
7542	1.141	11.287	0.0028	4.471
4032	1.341	10.884	0.0034	4.239
4188	−1.220	−10.827	0.0034	4.205
5084	−1.072	−10.731	0.0035	4.147
6903	−1.251	−10.585	0.0036	4.059
6023	1.012	10.238	0.0044	3.843
3695	1.057	10.167	0.0045	3.798
4546	1.269	10.012	0.0048	3.697
2679	−1.233	−9.750	0.0052	3.524

The top 30 differentially expressed genes based on the quantile and manifold normalizations are reported in Tables 9 and 10, respectively. With a threshold value of 0.05 for adjusted *p*-values, the number of differentially expressed genes identified with the quantile and manifold methods were 168 and 150, respectively. The 150 genes detected using the manifold method are also identified with the quantile normalization. The additional 18 genes detected with the quantile method are signed in the MA-plot for the manifold case in red points. As in the Tomato data, comparing with the quantile normalization, the detection of differential expressed genes based on the intensities normalized with the manifold method restricts a bit more the number of genes, being a more conservative (sparse) method. The MA-plots are shown in Fig. 8.

Unfortunately, for the swirl zebrafish experiment there is not exist qRT-PCR data to validate the results of identification of differentially expressed genes founded when the expression intensities normalized with the quantile and manifold normalizations methods are used.

7. Conclusions

Motivated by the density estimation problem when the data are a sample of density curves and by the popularity of the quantile normalization method by Bolstad et al. [4], we relate both issues

Table 10
Top 30 differentially expressed genes identified from Swirl data (manifold normalization).

Gene	*M*-value	Moderated *t*	Adj. *p*-value	*B*
2961	−2.553	−16.684	0.0033	6.415
3723	−2.072	−15.647	0.0033	6.079
1611	−2.082	−15.471	0.0033	6.018
3721	−2.133	−14.282	0.0038	5.580
1609	−2.233	−14.008	0.0038	5.472
8295	1.291	13.442	0.0038	5.237
319	−1.243	−13.275	0.0038	5.165
515	1.246	13.080	0.0038	5.080
7602	1.124	12.738	0.0039	4.925
3790	1.168	12.598	0.0039	4.860
157	−1.701	−12.216	0.0039	4.678
5931	−1.066	−12.105	0.0039	4.624
7307	1.147	11.987	0.0039	4.565
7491	1.282	11.740	0.0039	4.440
1697	1.057	11.726	0.0039	4.433
3726	−1.238	−11.698	0.0039	4.419
683	1.290	11.608	0.0039	4.372
7036	1.295	11.549	0.0039	4.341
5084	−1.049	−11.118	0.0044	4.110
5075	1.223	11.110	0.0044	4.106
4188	−1.206	−11.044	0.0044	4.069
4380	1.181	10.965	0.0044	4.025
7542	1.072	10.890	0.0044	3.983
2276	1.104	10.861	0.0044	3.967
4032	1.206	10.808	0.0044	3.937
6903	−1.185	−10.436	0.0053	3.720
4017	−1.042	−10.202	0.0059	3.579
6023	0.933	10.100	0.0061	3.516
3695	0.998	10.031	0.0062	3.474
4546	1.189	9.744	0.0071	3.292

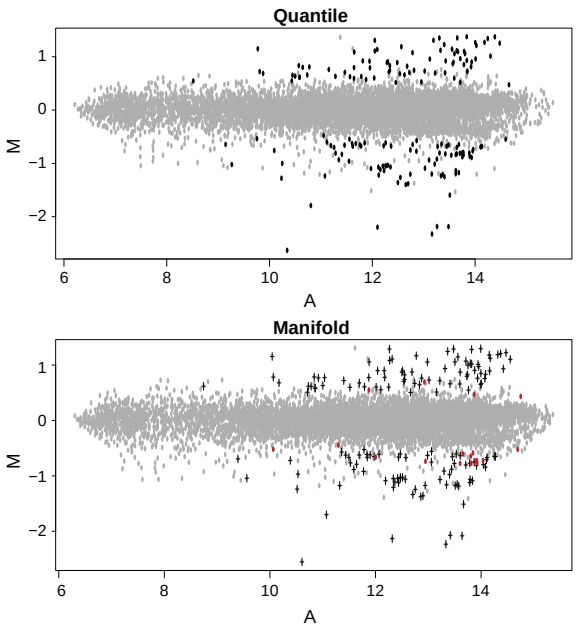

Fig. 8. Graphical illustration of differentially expressed genes identified using the normalized expressions with the quantile and manifold methods. The red symbols correspond to the genes identified with the quantile normalization but not with the manifold method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

considering the quantile normalization as a particular case of the structural mean estimation procedure based on a non parametric warping model for functional data by Dupuy et al. [9]. The asymptotic statistical properties of quantile normalization are established based on this connection. In addition, a new normalization procedure, using a manifold embedding framework, is proposed to improve the situation where the quantile method does not capture the structural features across the sample of density curves as pointed out by a simulation study. Both normalization approaches are applied on two data sets of two-channel spotted microarrays, and their utility in terms of detection of differentially expressed genes is studied. Both methods have similar results for identifying differential expressed genes, with the slight difference that the identification using the intensities normalized with the manifold method, is a little bit sparser with respect to the quantile normalization. Therefore, our manifold method is an alternative methodology for normalization. It achieves the same performance than the usual quantile normalization method in many cases as was shown for the studied real data sets. Moreover when the expression data densities have a large variability as illustrated in case four in the simulations, the proposed estimate density is more accurate than the estimator considered using the quantile normalization.

Acknowledgments

We are grateful to the anonymous reviewers and the editor for their valuable comments and constructive suggestions.

Appendix A

Proof. Proof of Theorem 1. To prove the almost sure convergence the next corollary by [16] is needed.

Corollary 1 [16, Corollary 7.10].
Let W be a Borel random variable with values in a separable Banach space $\mathcal{B}$. Then $S_m/m \to 0$ strongly as $m \to \infty$ if and only if $E\|W\| < \infty$ and $EW = 0$.

First note from Eq. (3) that

$$\begin{aligned}\mathbf{E}(q_i(\alpha)) &= \mathbf{E}\left(F_i^{-1}(\alpha)\right) = \mathbf{E}\left(H_i \circ F^{-1}(\alpha)\right) = \mathbf{E}(H_i) \circ F^{-1}(\alpha) \\ &= \phi \circ F^{-1}(\alpha) = F_{SE}^{-1}(\alpha) = \phi \circ q(\alpha) = q_{SE}(\alpha),\end{aligned}$$

where $q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geqslant \alpha\}$, $0 \leqslant \alpha \leqslant 1$, thus we have

$$\begin{aligned}\overline{q_m(\alpha)} - \mathbf{E}\left(\overline{q_m(\alpha)}\right) &= \frac{1}{m}\sum_{i=1}^{m} H_i \circ F^{-1}(\alpha) - \phi \circ F^{-1}(\alpha) \\ &= \frac{1}{m}\sum_{i=1}^{m}(H_i - \phi) \circ F^{-1}(\alpha) = \frac{1}{m}\sum_{i=1}^{m}(H_i - \phi) \circ q(\alpha).\end{aligned}$$

Setting $S_m = \sum_{i=1}^{m} W_i$ where $W_i = (H_i - \phi) \circ q(\alpha)$ is a sequence of independent and identically distributed random variables in a separable Banach space $\mathcal{B} = \mathcal{C}([0,1])$, and applying the above Corollary, the almost sure convergence of $\overline{q_m(\alpha)}$ is guaranteed.

The asymptotic normality of $\overline{q_m(\alpha)}$ is now obtained applying the multivariate central limit theorem. For any $K \in \mathbb{N}$, and fixed $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$,

$$\sqrt{m}\begin{bmatrix}\overline{q_m(\alpha_1)} - \mathbf{E}\overline{q_m(\alpha_1)} \\ \vdots \\ \overline{q_m(\alpha_K)} - \mathbf{E}\overline{q_m(\alpha_K)}\end{bmatrix} = \sqrt{m}\begin{bmatrix}\frac{1}{m}\sum_{i=1}^{m}(H_i - \phi) \circ q(\alpha_1) \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{m}(H_i - \phi) \circ q(\alpha_K)\end{bmatrix} \xrightarrow[m\to\infty]{\mathcal{D}} \mathcal{N}_K(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the asymptotic variance–covariance matrix whose (k,k')-element is given by $\Sigma_{k,k'} = \vartheta(q(\alpha_k), q(\alpha_{k'}))$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$, which is obtained as

$$\mathbf{Cov}\left(\overline{q_m(\alpha_k)}, \overline{q_m(\alpha_{k'})}\right) = \mathbf{Cov}\left(\frac{1}{m}\sum_{i=1}^{m} q_i(\alpha_k), \frac{1}{m}\sum_{i=1}^{m} q_i(\alpha_{k'})\right)$$
$$= \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Cov}(H_i \circ q(\alpha_k), H_i \circ q(\alpha_{k'}))$$
$$= \frac{1}{m}\vartheta(q(\alpha_k), q(\alpha_{k'})),$$

where $\vartheta(q(\alpha_k), q(\alpha_{k'}))$ is the autocovariance function of $H_i,\ i = 1, \ldots, m$.

Following [26], the tightness moment condition to weak convergence is given by

$$\mathbf{E}\left[\left|\sqrt{m}\left(\overline{q_m(\alpha)} - \mathbf{E}\overline{q_m(\alpha)}\right) - \sqrt{m}\left(\overline{q_m(\beta)} - \mathbf{E}\overline{q_m(\beta)}\right)\right|^2\right]$$
$$= \mathbf{E}\left[\left|\sqrt{m}\left(\left(\overline{q_m(\alpha)} - \mathbf{E}\overline{q_m(\alpha)}\right) - \left(\overline{q_m(\beta)} - \mathbf{E}\overline{q_m(\beta)}\right)\right)\right|^2\right]$$
$$= \mathbf{E}\left[m\left|\left(\frac{1}{m}\sum_{i=1}^{m} H_i \circ q(\alpha) - \phi \circ q(\alpha)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} H_i \circ q(\beta) - \phi \circ q(\beta)\right)\right|^2\right]$$
$$= \mathbf{E}\left[m\left|\frac{1}{m}\sum_{i=1}^{m}(H_i - \phi) \circ (q(\alpha) - q(\beta))\right|^2\right] \leqslant C_1 C_2 |\alpha - \beta|^2,$$

if the Assumptions A1 and A2 are satisfied. □

Proof. Proof of Proposition 1.

The proof is a direct application of the following theorems of strong consistency and asymptotic normality for quantile estimators. See [19,6] for its proofs.

Theorem. (**Strong consistency of quantile estimator**) *If the αth population quantile, $q(\alpha)$, is the unique solution of $F(x-) \leqslant \alpha \leqslant F(x)$, then $\hat{q}_n(\alpha) \underset{n\to\infty}{\overset{a.s.}{\to}} q(\alpha)$.*

Therefore $\hat{q}_{i,n}(\alpha) \underset{n\to\infty}{\overset{a.s.}{\to}} q_i(\alpha)$ for $i = 1, \ldots, m$.

Theorem. (**Asymptotic normality of order statistics**) *For a fixed $0 < \alpha < 1$, assume F is continuously differentiable at the αth population quantile, $q(\alpha), f(q(\alpha)) > 0$, and $n^{-1/2}(j/n - \alpha) = o(1)$. Then $\sqrt{n}(X_{j:n} - q(\alpha)) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}(0, \alpha(1-\alpha)/f^2(q(\alpha)))$, where $X_{j:n} = X_{[\alpha n]+1}$ is the jth sample quantile, and $[\alpha n]$ denotes the greatest integer less or equal than αn.*

In consequence for $i = 1, \ldots, m$ we have

$$\sqrt{n}(X_{i,j:n} - q_i(\alpha)) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f_i^2(q_i(\alpha))}\right),$$

that conditioned to a fixed H_i implies

$$\sqrt{n}(X_{i,j:n} - H_i \circ q(\alpha)) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{\left(f \circ H_i^{-1}(H_i \circ q(\alpha)) \cdot \left(H_i^{-1}\right)'(H_i \circ q(\alpha))\right)^2}\right),$$

where $\left(H_i^{-1}\right)'(z) = dH_i^{-1}(z)/dz = \left\{H_i' \circ H_i^{-1}(z)\right\}^{-1}$. □

The moments of order statistics are hard to compute for many distributions so these can be approximated reasonably using a linear Taylor series expansion of the relation $X_{i,j:n} \overset{d}{=} F_i^{-1}(U_{i,j:n})$ around the point $\mathbf{E}(U_{i,j:n}) = \alpha_j = j/(n+1)$, where $U_{i,j:n}$ denotes the jth order statistic in a sample of size n from the uniform $(0,1)$ distribution. The approximated means, variances and covariances of order statistics for $i = 1, \ldots, m$ are given by (see, for example, [6,2])

$$\mathbf{E}(X_{i,j:n}|H_i) = q_{i,j} + \frac{\alpha_j(1-\alpha_j)}{2(n+2)} q_{i,j}'' + \frac{\alpha_j(1-\alpha_j)}{(n+2)^2}\left[\frac{1}{3}((1-\alpha_j) - \alpha_j) q_{i,j}'' + \frac{1}{8}\alpha_j(1-\alpha_j) q_{i,j}^{(4)}\right] + O\left(\frac{1}{n^2}\right), \tag{A.1}$$

$$\mathbf{Var}(X_{i,j:n}|H_i) = \frac{\alpha_j(1-\alpha_j)}{n+2} q_{i,j}'^2 + \frac{\alpha_j(1-\alpha_j)}{(n+2)^2} \times \left[2((1-\alpha_j) - \alpha_j) q_{i,j}' q_{i,j}'' + \alpha_j(1-\alpha_j)\left(q_{i,j}' q_{i,j}''' + \frac{1}{2} q_{i,j}''^2\right)\right] + O\left(\frac{1}{n^2}\right) \tag{A.2}$$

and

$$\mathbf{Cov}(X_{i,j:n}, X_{i,s:n}|H_i) = \frac{\alpha_j(1-\alpha_s)}{n+2} q_{i,j}' q_{i,s}' + \frac{\alpha_j(1-\alpha_s)}{(n+2)^2}\left[((1-\alpha_j) - \alpha_j) q_{i,j}'' q_{i,s}' + ((1-\alpha_s) - \alpha_s) q_{i,j}' q_{i,s}'' + \frac{1}{2}\alpha_j(1-\alpha_j) q_{i,j}''' q_{i,s}' + \frac{1}{2}\alpha_s(1-\alpha_s) q_{i,j}' q_{i,s}''' + \frac{1}{2}\alpha_j(1-\alpha_s) q_{i,j}'' q_{i,s}''\right] + O\left(\frac{1}{n^2}\right), \tag{A.3}$$

where, since $\alpha_j = F_i(q_{i,j})$, we have

$$q_{i,j}' = \frac{dq_{i,j}}{d\alpha_j} = \frac{1}{f_i(q_{i,j})} < \infty,$$

$$q_{i,j}'' = -\frac{f_i'(q_{i,j})}{f_i^2(q_{i,j})} = -\frac{df_i(q_{i,j})}{dq_{i,j}} \frac{1}{f_i^3(q_{i,j})} < \infty \quad \text{and so on,}$$

where $f_i(q_{i,j}) > C$, with $C > 0$ is the density-quantile function of X evaluated at $q_{i,j} = q_i(\alpha_j) = H_i \circ F^{-1}(\alpha_j)$ with $\alpha_j = j/(n+1),\ j = 1, \ldots, n$. $|f_i'| < M, |f_i''| < M$, and $|f_i'''| < M$.

This approximation method is due to David and Johnson [5] where they derived approximations of order $(n+2)^{-3}$. Additionally, note that the asymptotic means, variances, and covariances correspond to the first terms of Eqs. (A.1)–(A.3), respectively [6].

Using the approximation in Eq. (A.1), the mean of $\bar{\tilde{q}}_j$ is calculated as

$$\mathbf{E}\left(\bar{\tilde{q}}_j\right) = \mathbf{E}\left[\mathbf{E}\left(\bar{\tilde{q}}_j|H_i\right)\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left[\mathbf{E}(X_{i,j:n}|H_i)\right]$$
$$= \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left[q_{i,j} + \frac{\alpha_j(1-\alpha_j)}{2(n+2)} q_{i,j}'' + O\left(\frac{1}{n^2}\right)\right]$$
$$= \frac{1}{m}\sum_{i=1}^{m}\left[q_{SE}(\alpha_j) + \frac{\alpha_j(1-\alpha_j)}{2(n+2)}\mathbf{E}\left(\frac{-df_i(q_{i,j})}{dq_{i,j}}\frac{1}{f_i^3(q_{i,j})}\right) + O\left(\frac{1}{n^2}\right)\right]$$
$$= \frac{1}{m}\sum_{i=1}^{m}\left[q_{SE}(\alpha_j) + \frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{n^2}\right)\right]$$
$$= q_{SE}(\alpha_j) + \frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{n^2}\right),$$

where $|df_i(q_{i,j})/dq_{i,j}| < M$ and $f_i^3(q_{i,j}) > C$.

While through Eq. (A.3), the covariance between of $\bar{\tilde{q}}_{j_k}$ and $\bar{\tilde{q}}_{j_{k'}}$ for $k \neq k'$ $k = 1, \ldots, K$ is given by

$$
\begin{aligned}
\mathbf{Cov}\left(\bar{\hat{q}}_{j_k}, \bar{\hat{q}}_{j_{k'}}\right) &= \mathbf{Cov}\left[\frac{1}{m}\sum_{i=1}^{m} X_{i,j_k:n}, \frac{1}{m}\sum_{i=1}^{m} X_{i,j_{k'}:n}\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Cov}\left(X_{i,j_k:n}, X_{i,j_{k'}:n}\right) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\left\{\mathbf{E}\left[\mathbf{Cov}\left(X_{i,j_k:n}, X_{i,j_{k'}:n}|H_i\right)\right]\right. \\
&\quad \left. + \mathbf{Cov}\left[\mathbf{E}(X_{i,j_k:n}|H_i), \mathbf{E}\left(X_{i,j_{k'}:n}|H_i\right)\right]\right\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\left\{\mathbf{E}\left[\frac{\alpha_{j_k}(1-\alpha_{j_{k'}})}{n+2} q'_{i,j_k} q'_{i,j_{k'}} + O\left(\frac{1}{n^2}\right)\right]\right. \\
&\quad + \mathbf{Cov}\left[q_{i,j_k} + \frac{\alpha_{j_k}(1-\alpha_{j_k})}{2(n+2)} q''_{i,j_k} + O\left(\frac{1}{n^2}\right), q_{i,j_{k'}}\right. \\
&\quad \left.\left. + \frac{\alpha_{j_{k'}}(1-\alpha_{j_{k'}})}{2(n+2)} q''_{i,j_{k'}} + O\left(\frac{1}{n^2}\right)\right]\right\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\left\{\mathbf{E}\left[\frac{1}{4(n+2)}\frac{1}{C^2}\frac{1}{C^2} + O\left(\frac{1}{n^2}\right)\right]\right. \\
&\quad + \mathbf{Cov}\left[H_i(q(\alpha_{j_k})) + \frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right)\right. \\
&\quad \left. + O\left(\frac{1}{n^2}\right), H_i(q(\alpha_{j_{k'}})) + \frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right)\right. \\
&\quad \left.\left. + O\left(\frac{1}{n^2}\right)\right]\right\} \\
&= \frac{1}{m}\left[\frac{1}{4(n+2)}\frac{1}{C^4} + O\left(\frac{1}{n^2}\right)\right] \\
&\quad + \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Cov}\left[H_i(q(\alpha_{j_k})), H_i(q(\alpha_{j_{k'}}))\right] \\
&= \frac{1}{m}\left[\frac{1}{4(n+2)}\frac{1}{C^4} + O\left(\frac{1}{n^2}\right)\right] + \frac{1}{m}\vartheta\left(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\right)
\end{aligned}
$$

for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$.

From above equations we have that

$$
\mathbf{E}\left(\bar{\hat{q}}_j\right) \underset{n\to\infty}{\to} q_{SE}(\alpha_j) \tag{A.4}
$$

and

$$
\mathbf{Cov}\left(\bar{\hat{q}}_{j_k}, \bar{\hat{q}}_{j_{k'}}\right) \underset{n\to\infty}{\to} \frac{1}{m}\vartheta\left(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\right). \tag{A.5}
$$

Proof. Proof of Theorem 2.

The almost sure convergence of $\bar{\hat{q}}_j$ is established applying the results of strong consistency of $\overline{q_m(\alpha)}$ and $\hat{q}_{i,n}(\alpha)$ from Theorem 1 and Proposition 1, respectively.

The asymptotic normality of $\bar{\hat{q}}_j$ is obtained as follows

$$
\begin{aligned}
\sqrt{m}\frac{\left(\bar{\hat{q}}_j - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta(q(\alpha_j))}} &= \sqrt{m}\frac{\left(\frac{1}{m}\sum_{i=1}^{m} X_{i,j:n} - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta(q(\alpha_j))}} \\
&= \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\left(X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right)\right)}{\sqrt{\vartheta(q(\alpha_j))}} + \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}(X_{i,j:n}) - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta(q(\alpha_j))}} \\
&= \frac{\left(\sum_{i=1}^{m}\left(X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right)\right)\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}}{\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}\sqrt{\vartheta(q(\alpha_j))}} \\
&\quad + \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}(X_{i,j:n}) - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta(q(\alpha_j))}} \\
&= \frac{\left(\sum_{i=1}^{m} X_{i,j:n} - \sum_{i=1}^{m}\mathbf{E}(X_{i,j:n})\right)}{\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}}\frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}}{\sqrt{\vartheta(q(\alpha_j))}} \\
&\quad + \frac{\sqrt{m}\left(\frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{n^2}\right)\right)}{\sqrt{\vartheta(q(\alpha_j))}}.
\end{aligned}
$$

Given that $\mathbf{Var}(X_{i,j:n}) \underset{n\to\infty}{\to} \vartheta(q(\alpha_j))$, and under the assumption $\sqrt{m}/n \to 0$ we obtain, by the Lindeberg–Feller's central limit theorem for independent but not identically distributed random variables to the independent random variables $X_{1,j:n}, \ldots, X_{m,j:n}$, that

$$
\sqrt{m}\frac{\left(\bar{\hat{q}}_j - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta(q(\alpha_j))}} \underset{m,n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}(0,1),
$$

that in multivariate terms is expressed as

$$
\sqrt{m}\begin{bmatrix} \bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1) \\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K) \end{bmatrix} \underset{m,n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}_K(\mathbf{0}, \mathbf{\Sigma}),
$$

where $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$ and the (k,k')-element of $\mathbf{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\left(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\right)$.

The Lindeberg–Feller's central limit theorem holds if the Lyapunov's condition

$$
\frac{\sum_{i=1}^{m}\mathbf{E}\left|X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right|^{2+\delta}}{\left(\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}\right)^{2+\delta}} \underset{m,n\to\infty}{\to} 0
$$

is satisfied for some $\delta > 0$. Indeed for $\delta = 1$ and under the compactly central data hypothesis, $\left|X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right| \leqslant L < \infty$ for all i and j, we have

$$
\begin{aligned}
&\frac{1}{\left(\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}\right)^{2+1}}\sum_{i=1}^{m}\mathbf{E}\left|X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right|^{2+1} \\
&\quad \leqslant \frac{L}{\left(\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}\right)^{2+1}}\sum_{i=1}^{m}\mathbf{E}\left|X_{i,j:n} - \mathbf{E}(X_{i,j:n})\right|^{2} \\
&\quad = \frac{L}{\left(\sqrt{\sum_{i=1}^{m}\mathbf{Var}(X_{i,j:n})}\right)} \underset{m,n\to\infty}{\to} 0
\end{aligned}
$$

given that $\mathbf{Var}(X_{i,j:n}) \underset{n\to\infty}{\to} \vartheta(q(\alpha_j))$.

Therefore the Lyapunov's condition is satisfied. □

References

[1] M. Agueh, G. Carlier, Barycenters in the Wasserstein space, SIAM J. Math. Anal. 43 (2011) 904–924.

[2] B. Arnold, N. Balakrishnan, H. Nagaraja, A First Course in Order Statistics, Classics in Applied Mathematics, vol. 54, SIAM, Philadelphia, 2008.

[3] E. Boissard, T. Le Gouic, J.-M. Loubes. Distribution's template estimate with Wasserstein metrics, ArXiv e-prints (2011). http://arxiv.org/pdf/1111.5927v1.pdf.

[4] B.M. Bolstad, R.A. Irizarry, M. Åstrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.

[5] F.N. David, N.L. Johnson, Statistical treatment of censored data. Part I: Fundamental formulae, Biometrika 41 (1954) 228–240.

[6] H.A. David, H.N. Nagaraja, Order Statistics, third ed., Wiley, New Jersey, 2003.

[7] C. Dimeglio, S. Gallón, J.M. Loubes, E. Maza, Manifold embedding for curve registration, HAL: hal-00580792 (2012) submitted for publication.

[8] S. Dudoit, Y.H. Yang, Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data, in: G. Parmigiani, E.S. Garrett, R.A.

Irizarry, S.L. Zeger (Eds.), The Analysis of Gene Expression Data: Methods and Software, Statistics for Biology and Health, Springer, New York, pp. 73–101.
[9] J. Dupuy, J.M. Loubes, E. Maza, Non parametric estimation of the structural expectation of a stochastic increasing function, Statist. Comput. 21 (2011) 121–136.
[10] F. Gamboa, J.M. Loubes, E. Maza, Semi-parametric estimation of shits, Electron. J. Stat. 1 (2007) 616–640.
[11] T. Gasser, A. Kneip, Searching for structure in curve sample, J. Am. Statist. Assoc. 90 (1995) 1179–1188.
[12] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, T. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2003) 249–264.
[13] G. James, Curve alignment by moments, Ann. Appl. Statist. 1 (2007) 480–501.
[14] A. Kneip, T. Gasser, Statistical tools to analyze data representing a sample of curves, Ann. Statist. 20 (1992) 1266–1305.
[15] A. Kneip, J. Ramsay, Combining registration and fitting for functional models, J. Am. Statist. Assoc. 103 (2008) 1155–1165.
[16] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, Ergebnisse der Mathematik und ihrer Grenzgebiete 3, Folge. A Series of Modern Surveys in Mathematics, vol. 23, Springer, Berlin, 1991.
[17] X. Liu, H. Müller, Functional convex averaging and synchronization for time-warped random curves, J. Am. Statist. Assoc. 99 (2004) 687–699.
[18] J.O. Ramsay, X. Li, Curve registration, J. R. Statist. Soc. Ser. B Statist. Methodol. 60 (1998) 351–363.
[19] R. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.
[20] B.W. Silverman, Incorporating parametric effects into functional principal components analysis, J. R. Statist. Soc. Ser. B Statist. Methodol. 57 (1995) 673–689.
[21] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Statist. Appl. Genet. Mol. Biol. 3 (2004).
[22] G.K. Smyth, `limma`: linear models for microarray data, in: R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, 2005, pp. 397–420.
[23] G.K. Smyth, M. Ritchie, N. Thorne, J. Wettenhall, W. Shi, `limma`: linear models for microarray data user's guide, 2012, Software manual available from <http://www.bioconductor.org/>.
[24] G.K. Smyth, T.P. Speed, Normalization of cDNA microarray data, Methods 31 (2003) 265–273.
[25] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[26] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[27] H. Wang, N. Schauer, B. Usadel, P. Frasse, M. Zouine, M. Hernould, A. Latché, J.C. Pech, A. Fernie, M. Bouzayen, Regulatory features underlying pollination-dependent and -independent tomato fruit set revealed by transcript and primary metabolite profiling, Plant Cell 21 (2009) 1428–1452.
[28] K. Wang, T. Gasser, Alignment of curves by dynamic time warping, Ann. Statist. 25 (1997) 1251–1276.
[29] K. Wang, T. Gasser, Synchronizing sample curves nonparametrically, Ann. Statist. 27 (1999) 439–460.
[30] Y.H. Yang, A.C. Paquet, Preprocessing two-color spotted arrays, in: R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, 2005, pp. 49–69.
[31] Y.H. Yang, T.P. Speed, Design and analysis of comparative microarray experiments, in: T.P. Speed (Ed.), Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC, Boca Raton, 2003, pp. 35–91.
[32] Y.H. Yang, N.P. Thorne, Normalization for two-color cDNA microarray data, in: D.R. Goldstein (Ed.), Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes, vol. 40, New York, 2003, pp. 403–418.