Statistical properties of the quantile normalization method for density curve alignment
Santiago Gallón, Jean-Michel Loubes, Elie Maza

To cite this version:

HAL Id: hal-01135106
https://hal.archives-ouvertes.fr/hal-01135106
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:**

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, bld. 13, 650010 Medellín, Colombia
b Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
c Genomic & Biotechnology of the Fruit Laboratory, 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, . . . , m samples of j = 1, . . . , n i.i.d random variables Xij, the mean distribution is achieved by projecting the jth empirical vector of sample quantiles, qj = (q1j, . . . , qmj), onto the vector d = (1/√m, . . . , 1/√m). This gives projd qj = (m−1 Xj1 + · · · + qmj, . . . , m−1 Xjm + · · · + qmj), which is such that if all m data vectors Xj, j = 1, . . ., m, share the same distribution, then the plot of the quantiles gives a straight line along the line 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...
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 \( X_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, n_i \) be a sample of \( m \) independent real-valued random variables of size \( n_i \) with density function \( f_i : \mathbb{R} \rightarrow [0, 1] \) and distribution function \( F_i : \mathbb{R} \rightarrow [0, 1] \). We assume without loss of generality that \( n_i = n \) for all units \( i = 1, \ldots, m \). 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 \( F : \mathbb{R} \rightarrow [0, 1] \) by an invertible and differentiable warping function \( H_i \), of the following manner:

\[
F_i(t) = \Pr(X_{ij} \leq t) = F \circ H_i^{-1}(t)
\]

where \( H_i \) is random, in the sense that \( (H_i, \ldots, H_m) \) is an i.i.d random sample from a (non parametric) warping stochastic process \( \mathcal{H} : \Omega \rightarrow \mathcal{C}(\mathbb{R}) \) defined on an unknown probability space \( (\Omega, \mathcal{A}, \mathbb{P}) \), while \( \mathcal{C}(\mathbb{R}) \) denotes the space of all continuous functions defined on \( \mathbb{R} \). Define \( \phi \) its mean and let \( \sigma \) 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

\[
q_{\text{SE}}(x) := F_{\text{SE}}^{-1}(x) = \phi \circ F^{-1}(x), \quad 0 \leq x \leq 1.
\]

Inverting Eq. (1) leads to

\[
q_i(x) = F_i^{-1}(x) = H_i \circ F^{-1}(x), \quad 0 \leq x \leq 1,
\]

where \( q_i(x) \) is the population quantile function (the left continuous distribution function) and \( F_i \) is the restriction of \( F_i \) to the interval \( [0, 1] \) and \( F_{\text{SE}} \) is the distribution function of the quantile

\[
\begin{align*}
\mathbb{E} \left[ |H_i(x) - E[H_i(x)] - (H_i(\beta) - E[H_i(\beta)])|^2 \right] & \leq C_1|x - \beta|^2. \\
\mathbb{E} \left[ |F_i^{-1}(x) - F^{-1}(x)|^2 \right] & \leq C_2|x - \beta|^2.
\end{align*}
\]

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

**Theorem 1.** The estimator \( \overline{q}_m(x) \) is consistent in the sense that

\[
\left\| \overline{q}_m(x) - \mathbb{E} \left[ q_m(x) \right] \right\|_\infty = \frac{\| q_{\text{SE}}(x) - q_m(x) \|_\infty}{\sqrt{m}} \rightarrow 0, \quad m \rightarrow \infty.
\]

Moreover, under Assumptions A1 and A2, the estimator is asymptotically Gaussian, for any \( K \in \mathbb{N} \) and fixed \( (x_1, \ldots, x_K) \in [0, 1]^K \),

\[
\sqrt{m} \left[ \begin{array}{c}
q_m(x_1) - q_{\text{SE}}(x_1) \\
q_m(x_2) - q_{\text{SE}}(x_2) \\
\vdots \\
q_m(x_K) - q_{\text{SE}}(x_K)
\end{array} \right] \xrightarrow{d} N_K(0, \Sigma),
\]

where \( \Sigma \) is the asymptotic variance–covariance matrix whose \((k, k')\)-element is given by

\[
\Sigma_{k,k'} = \frac{1}{n} q_{\text{SE}}(x_k), \quad q_{\text{SE}}(x_k), \quad \text{for all } (x_k, x_k') \in [0, 1]^2 \text{ with } x_k < x_{k'}.
\]

3. Quantile estimation and the quantile normalization method

The distribution function is not observed and only random samples \( X_{1,i}, \ldots, X_{n,i} \) from \( F_i(x) \) for \( i = 1, \ldots, m \) 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 \( X_{1,i} \leq X_{2,i} \leq \cdots \leq X_{n,i} \), hence the estimation of the quantile functions, \( q_i(x) \), is obtained by
Also sample quantiles, by mean the projection of the empirical quantile vector of the $q_{ij}$ consists in finding a distribution function $F$ respectively $p_{ij}$.

Finally, the estimator of the structural quantile is given by

$$\overline{q}_j = \frac{1}{m} \sum_{i=1}^{m} q_{ij}, \quad j = 1, \ldots, n.$$  

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 $j$th sample quantiles, $q_i = (q_{i1}, \ldots, q_{in})^T$, onto the $m$-vector $d = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^T$, given by $proj_d(q_i) = (m^{-1} \sum_{j=1}^{m} q_{ij}, \ldots, m^{-1} \sum_{j=1}^{m} q_{ij})$. 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 $d$.

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

**Theorem 2.** The quantile normalization estimator $\overline{q}_j$ is strongly consistent, $\overline{q}_j \overset{a.s.}{\longrightarrow} q_{\mathbb{R}^k}(x_j)$, $j = 1, \ldots, n$. Also, under the assumptions of compactly central data, $|X_{ij} - \mathbb{E}[X_{ij}]| \leq L < \infty$ for all $i$ and $j$, and $\sqrt{m}/n \to 0$, it is asymptotically Gaussian. Actually, for any $K \in \mathbb{N}$ and fixed $(x_1, \ldots, x_K) \in [0, 1]^K$,

$$\sqrt{m} \left[ \begin{array}{c} \overline{q}_1 - q_{\mathbb{R}^k}(x_1) \\ \vdots \\ \overline{q}_K - q_{\mathbb{R}^k}(x_K) \end{array} \right] \overset{d}{\longrightarrow} N(K\mathbf{0}, \Sigma),$$

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

This theorem relies on the asymptotic behavior of the quantile estimator, $q_{\mathbb{R}^k}(x)$, given by the following proposition.

**Proposition 1.** Assume $F_{ij}$ is continuously differentiable at the $j$th population quantile $q_j(x)$ which is the unique solution of $F_{ij}(q_j(x)) - x < z \leq F_{ij}(q_j(x))$, and $f_j(q_j(x)) > 0$ for a fixed $0 < x < 1$. Also assume $n^{-1/2}/(n - z) = o(1)$. Then, for $i = 1, \ldots, m$, the estimator $q_{\mathbb{R}^k}(x)$ is strongly consistent, $q_{\mathbb{R}^k}(x) \overset{a.s.}{\longrightarrow} q_j(x)$; and asymptotically Gaussian

$$\sqrt{n} [X_{ij} - H \circ q_j(z)] \overset{d}{\longrightarrow} N \left( 0, \frac{2(1-x)}{(f \circ H^{-1}(H \circ q_j(z)))' \left( H' \circ H^{-1}(H \circ q_j(z)) \right)^{-1}} \right),$$

where $H^{-1}(z) = \left( H \circ q_j(z) - dz = \left( H' \circ H^{-1}(H \circ q_j(z)) \right)^{-1}.$

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 $F_j$'s the distribution functions and $\mu_j$'s the distributions of the i.i.d. sample $X_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, the problem consists in finding a distribution function $F$ and a probability $\mu$ which plays the role of a mean function but close enough to the data. This corresponds to the usual registration problem of the $F_j$'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 $\mu$ which minimizes

$$\mu \mapsto \frac{1}{m} \sum_{j=1}^{m} W^2_2(\mu, \mu_j).$$  

where $W_2^2$ stands for the 2-Wasserstein distance

$$W^2_2[\mu, \mu_j] = \int |F^{-1}_j (x) - F^{-1}_i (x)|^2 dx.$$

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 $\mu_j$'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 $F_j$'s and the $\mu_j$'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

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

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 $d_{\text{g}}$, an approximation of the geodesic distance, is provided using an Isomap-type graph approximation, following [25]. This gives rise to the criterion

$$F \mapsto \frac{1}{m} \sum_{j=1}^{m} d_{\text{g}} \left( F^{-1}_j, F^{-1}_i \right).$$

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 $X_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ random variables. In order to mimic the geodesic distance between the inverse of the distribution functions, we will directly estimate $F^{-1}_j(x)$, for $j \neq i < n < z < j/n$ by the corresponding order statistics $X_{ij}$. Hence, we sort the observations for each sample $i$, and denote by $X_{ij}$ the sorted vector $X_{1j}, \ldots, X_{nj}$ and thus we obtain an array of sorted observations $(X_{1j}, \ldots, X_{nj})_i$. We then consider $d_{\text{g}}$ an approximation of the geodesic distance between the vectors $X_{ij}$ and define the corresponding geodesic mean as the minimizer over all the observation vectors $x \in (X_{ij}, i = 1, \ldots, m)$ of the criterion

$$x \mapsto \frac{1}{m} \sum_{j=1}^{m} d_{\text{g}}(x, X_{ij}).$$
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 $\phi_i(x; \mu_i, \sigma_i), i = 1, 2, 3,$

$$f_i(x) = \sum_{i=1}^{3} \omega_i \phi_i(x; \mu_i, \sigma_i), \quad i = 1, \ldots, m,$$

where $\omega_i \in [0, 1]$ are probability weights which satisfy $\sum_{i=1}^{3} \omega_i = 1, i = 1, \ldots, m$. 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 $\mu_1 = 1, \mu_2 = 4$ and $\mu_3 = 7$; to the scale parameters $\sigma_1 = 0.7, \sigma_2 = 0.8$, and $\sigma_3 = 0.9$; and to the probability weights $\omega_1 = 0.4, \omega_2 = 0.3$, and $\omega_3 = 0.3$. 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).**

$$U(\mu_1 - 0.15, \mu_2 + 0.15) \text{ for } i = 1, 2, 3.$$  

**Case 2 (Scale variations).**

$$U(\sigma_1 - 0.35, \sigma_2 + 0.35) \text{ for } i = 1, 2, \text{ and } U(\sigma_3 - 0.5, \sigma_3 + 0.5) \text{ for } l = 3.$$  

**Case 3 (Location and scale variations).**

Cases 1 and 2 together.

**Case 4 (Probability weight variations).**

For $l = 1, 2$, $U(\omega_l - 0.1, \omega_l + 0.1)$.

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, $\omega_l = \omega_{i1}, l = 1, 2, 3$ for $i = 1, \ldots, m$. 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 $M \subset \mathbb{R}^d$ 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-ENSAAT), 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 \texttt{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.](image-url)
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 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,
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

![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.)](image)
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.
Important features on genes can be found by means of the scatterplot between the average of log₂ 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 log₂ 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 these 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

### Table 8

<table>
<thead>
<tr>
<th>Gene</th>
<th>Bud</th>
<th>Manifold</th>
<th>qRT-PCR</th>
<th>Anthesis</th>
<th>Manifold</th>
<th>qRT-PCR</th>
<th>Post-anthesis</th>
<th>Manifold</th>
<th>qRT-PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>10318</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes (88)</td>
<td>yes (12)</td>
<td>yes (11)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>10617</td>
<td>yes (4)</td>
<td>yes (6)</td>
<td>no</td>
<td>yes (51)</td>
<td>yes (51)</td>
<td>yes</td>
<td>yes</td>
<td>yes (61)</td>
<td>yes (77)</td>
</tr>
<tr>
<td>10260</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes (145)</td>
<td>yes (137)</td>
<td>yes</td>
<td>yes (48)</td>
<td>yes (45)</td>
<td>yes</td>
</tr>
<tr>
<td>12722</td>
<td>yes (13)</td>
<td>yes (13)</td>
<td>yes</td>
<td>yes (103)</td>
<td>yes (77)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2266</td>
<td>yes (1)</td>
<td>yes (1)</td>
<td>yes</td>
<td>yes (29)</td>
<td>yes (29)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>3228</td>
<td>yes (2)</td>
<td>yes (2)</td>
<td>yes</td>
<td>yes (275)</td>
<td>yes (268)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>4209</td>
<td>yes (9)</td>
<td>yes (9)</td>
<td>yes</td>
<td>yes (9)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>7468</td>
<td>yes (85)</td>
<td>yes (75)</td>
<td>yes</td>
<td>yes (303)</td>
<td>yes (299)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>7392</td>
<td>yes (303)</td>
<td>yes (299)</td>
<td>yes</td>
<td>yes (147)</td>
<td>yes (171)</td>
<td>yes</td>
<td>yes (81)</td>
<td>yes (79)</td>
<td>yes</td>
</tr>
<tr>
<td>3161</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes (158)</td>
<td>yes (155)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes (109)</td>
<td>yes (103)</td>
<td>yes</td>
<td>yes (36)</td>
<td>yes (36)</td>
<td>yes</td>
</tr>
<tr>
<td>5626</td>
<td>yes (33)</td>
<td>yes (31)</td>
<td>yes</td>
<td>yes (44)</td>
<td>yes (40)</td>
<td>yes</td>
<td>yes (83)</td>
<td>yes (74)</td>
<td>yes</td>
</tr>
<tr>
<td>2238</td>
<td>yes (33)</td>
<td>yes (31)</td>
<td>yes</td>
<td>yes (83)</td>
<td>yes (80)</td>
<td>yes</td>
<td>yes (36)</td>
<td>yes (36)</td>
<td>yes</td>
</tr>
<tr>
<td>10258</td>
<td>yes (138)</td>
<td>yes (134)</td>
<td>yes</td>
<td>yes (109)</td>
<td>yes (103)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>3043</td>
<td>yes (138)</td>
<td>yes (134)</td>
<td>yes</td>
<td>yes (109)</td>
<td>yes (103)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>11189</td>
<td>yes (138)</td>
<td>yes (134)</td>
<td>yes</td>
<td>yes (109)</td>
<td>yes (103)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2402</td>
<td>yes (138)</td>
<td>yes (134)</td>
<td>yes</td>
<td>yes (109)</td>
<td>yes (103)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>8241</td>
<td>yes (28)</td>
<td>yes (26)</td>
<td>no</td>
<td>yes (28)</td>
<td>yes (26)</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12911</td>
<td>yes (20)</td>
<td>yes (25)</td>
<td>yes</td>
<td>yes (35)</td>
<td>yes (39)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>10932</td>
<td>yes (115)</td>
<td>yes (110)</td>
<td>yes</td>
<td>yes (27)</td>
<td>yes (28)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
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).

### 6.1.4. 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 stages 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 indicate 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 first 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/lmmaGUI/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.

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 differentially 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
warping model for functional data by Dupuy et al. [9]. The asymptotic properties of quantile normalization are established 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 proce-
dure, 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 quantile normalization identified 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 $B$. 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

$$
E(q_m(x)) = E(F_x^{-1}(q(x))) = E(H_i \circ F_x^{-1}(x)) = E(H_i) \circ F_x^{-1}(x) = F_x^{-1}(x) = q(x),
$$

where $q(x) = F_x^{-1}(x) = \inf \{x \in R : F(x) \geq x\}$, $0 \leq x \leq 1$, thus we have

$$
\overline{q}_m(x) - E(q_m(x)) = \frac{1}{m} \sum_{i=1}^{m} (H_i - \phi) \circ F_x^{-1}(x) = \frac{1}{m} \sum_{i=1}^{m} (H_i - \phi) \circ q(x).
$$

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

The asymptotic normality of $\overline{q}_m(x)$ is now obtained applying the multivariate central limit theorem. For any $K \in N$, and fixed $(x_1, \ldots, x_K) \in [0, 1]^K$,

$$
\sqrt{m} \left[ \overline{q}_m(x_1) - E(q_m(x_1)) : \ldots : \overline{q}_m(x_K) - E(q_m(x_K)) \right] \xrightarrow{m \to \infty} N_m(0, \Sigma),
$$

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

---

Table 10

<table>
<thead>
<tr>
<th>Gene</th>
<th>$M$-value</th>
<th>Moderated $t$</th>
<th>Adj. $p$-value</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2961</td>
<td>-2.553</td>
<td>-16.684</td>
<td>0.0033</td>
<td>6.415</td>
</tr>
<tr>
<td>3723</td>
<td>-2.072</td>
<td>-15.647</td>
<td>0.0033</td>
<td>6.079</td>
</tr>
<tr>
<td>1611</td>
<td>-2.082</td>
<td>-15.471</td>
<td>0.0033</td>
<td>6.018</td>
</tr>
<tr>
<td>3721</td>
<td>-2.133</td>
<td>-14.282</td>
<td>0.0038</td>
<td>5.580</td>
</tr>
<tr>
<td>1609</td>
<td>-2.233</td>
<td>-14.008</td>
<td>0.0038</td>
<td>5.472</td>
</tr>
<tr>
<td>8295</td>
<td>1.291</td>
<td>13.442</td>
<td>0.0038</td>
<td>5.237</td>
</tr>
<tr>
<td>319</td>
<td>-1.243</td>
<td>-12.275</td>
<td>0.0038</td>
<td>5.165</td>
</tr>
<tr>
<td>515</td>
<td>1.246</td>
<td>13.080</td>
<td>0.0038</td>
<td>5.080</td>
</tr>
<tr>
<td>7602</td>
<td>1.124</td>
<td>12.738</td>
<td>0.0039</td>
<td>4.925</td>
</tr>
<tr>
<td>3790</td>
<td>1.168</td>
<td>12.598</td>
<td>0.0039</td>
<td>4.860</td>
</tr>
<tr>
<td>157</td>
<td>-1.701</td>
<td>-12.216</td>
<td>0.0039</td>
<td>4.678</td>
</tr>
<tr>
<td>5931</td>
<td>-1.066</td>
<td>-12.105</td>
<td>0.0039</td>
<td>4.624</td>
</tr>
<tr>
<td>7307</td>
<td>1.147</td>
<td>11.987</td>
<td>0.0039</td>
<td>4.565</td>
</tr>
<tr>
<td>7491</td>
<td>1.282</td>
<td>11.740</td>
<td>0.0039</td>
<td>4.440</td>
</tr>
<tr>
<td>1697</td>
<td>1.057</td>
<td>11.726</td>
<td>0.0039</td>
<td>4.433</td>
</tr>
<tr>
<td>3726</td>
<td>-1.238</td>
<td>-11.698</td>
<td>0.0039</td>
<td>4.419</td>
</tr>
<tr>
<td>683</td>
<td>1.290</td>
<td>11.608</td>
<td>0.0039</td>
<td>4.372</td>
</tr>
<tr>
<td>7036</td>
<td>1.295</td>
<td>11.549</td>
<td>0.0039</td>
<td>4.341</td>
</tr>
<tr>
<td>5084</td>
<td>-1.049</td>
<td>-11.118</td>
<td>0.0044</td>
<td>4.110</td>
</tr>
<tr>
<td>5075</td>
<td>1.223</td>
<td>11.110</td>
<td>0.0044</td>
<td>4.106</td>
</tr>
<tr>
<td>4188</td>
<td>-1.206</td>
<td>-11.044</td>
<td>0.0044</td>
<td>4.069</td>
</tr>
<tr>
<td>4380</td>
<td>1.181</td>
<td>10.965</td>
<td>0.0044</td>
<td>4.025</td>
</tr>
<tr>
<td>7542</td>
<td>1.072</td>
<td>10.890</td>
<td>0.0044</td>
<td>3.983</td>
</tr>
<tr>
<td>2276</td>
<td>1.104</td>
<td>10.861</td>
<td>0.0044</td>
<td>3.967</td>
</tr>
<tr>
<td>4032</td>
<td>1.206</td>
<td>10.808</td>
<td>0.0044</td>
<td>3.937</td>
</tr>
<tr>
<td>6903</td>
<td>-1.185</td>
<td>-10.436</td>
<td>0.0053</td>
<td>3.720</td>
</tr>
<tr>
<td>4017</td>
<td>-1.042</td>
<td>-10.202</td>
<td>0.0061</td>
<td>3.579</td>
</tr>
<tr>
<td>6023</td>
<td>0.933</td>
<td>10.100</td>
<td>0.0061</td>
<td>3.516</td>
</tr>
<tr>
<td>3695</td>
<td>0.998</td>
<td>10.031</td>
<td>0.0062</td>
<td>3.474</td>
</tr>
<tr>
<td>4546</td>
<td>1.189</td>
<td>9.744</td>
<td>0.0071</td>
<td>3.292</td>
</tr>
</tbody>
</table>

---

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.)
The Taylor series expansion of the relation $\mathcal{H}_i = X_i(n)H_i \circ q(x_i) \circ \mathcal{H}_i \circ q(x_i)$

\[ \begin{align*}
\text{Cov}(\mathcal{H}_i, \mathcal{H}_j) &= \text{Cov}\left( \frac{1}{m} \sum_{i=1}^{m} \mathcal{H}_i \circ q(x_i), \frac{1}{m} \sum_{i=1}^{m} \mathcal{H}_i \circ q(x_i) \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \text{Cov}(\mathcal{H}_i \circ q(x_i), \mathcal{H}_i \circ q(x_i)) \\
&= \frac{1}{m} \theta(q(x_i), q(x_i)),
\end{align*} \]

where $\theta(q(x_i), q(x_i))$ is the autocovariance function of $H_i$ when $i = 1, \ldots, m$.

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

\[ E\left[ \sqrt{m} \left( \mathcal{H}_i - E(\mathcal{H}_i) \right)^2 \right] = E\left[ \sqrt{m} \left( \mathcal{H}_i - E(\mathcal{H}_i) \right)^2 \right] \]

Theorem. (Strong consistency of quantile estimator) If the $i$th population quantile, $q(x)$, is the unique solution of $F(x) - x = q(x) \circ F(x)$, then $\mathcal{H}_i = \mathcal{H}_i(x)$.

Therefore $\mathcal{H}_i(x) \rightarrow F^{-1}(x)$ as $n \rightarrow \infty$.

Theorem. (Asymptotic normality of order statistics) For a fixed $0 < x < 1$, assume $F$ is continuously differentiable at the $i$th population quantile, $q(x)$, and $F'(q(x)) > 0$, and $n^{-1/2}(1/n - x) = O(1)$. Then

\[ \sqrt{n}(X_{(i)}, - q(x)) \sim N(0, \mathcal{H}_i - F^{-1}(x))^2, \]

where $X_{(i)}$ is the $i$th sample quantile, and $[x]$ denotes the greatest integer less or equal than $x$.

Furthermore, we have

\[ \sqrt{n}(X_{(i)} - q(x)) \sim \mathcal{N}(0, \mathcal{H}_i - F^{-1}(x))^2, \]

that conditioned to a fixed $H_i$, we have

\[ \sqrt{n}(X_{(i)} - H_i \circ q(x)) \sim \mathcal{N}(0, \mathcal{H}_i - H_i \circ q(x))^2, \]

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

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 $X_{(i)} = (H_i^{-1}(U_{(i)}))$ around the point $E(U_{(i)}) = \mathcal{H}_i = f^{-1}(q_i)$, where $U_{(i)}$ denotes the $i$th 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])

\[ E(X_{(i)}|H_i) = \mathcal{H}_i + \frac{2^i}{(n+2)!} \partial^i \mathcal{H}_i \sum_{j=1}^{i} \frac{1}{j!} \mathcal{H}_i \circ q(x)^{(j)}(H_i \circ q(x)) \]

\[ \text{Var}(X_{(i)}|H_i) = \frac{2^{i}}{(n+2)!} \partial^i \mathcal{H}_i \sum_{j=1}^{i} \frac{1}{j!} \mathcal{H}_i \circ q(x)^{(j)}(H_i \circ q(x)) \]

\[ \text{Cov}(X_{(i)}|H_i, X_{(j)}|H_j) = \frac{2^{i+j}}{(n+2)!} \partial^{i+j} \mathcal{H}_i \sum_{k=1}^{i+j} \frac{1}{k!} \mathcal{H}_i \circ q(x)^{(k)}(H_i \circ q(x)) \]

with $i < j$.

While through Eq. (A.3), the covariance between $\overline{q}_i$ and $\overline{q}_j$ for $k < k' < 1, \ldots, K$ is given by

\[ \text{Cov}(\overline{q}_i, \overline{q}_j) = \frac{2^{i+j}}{(n+2)!} \partial^{i+j} \mathcal{H}_i \sum_{k=1}^{i+j} \frac{1}{k!} \mathcal{H}_i \circ q(x)^{(k)}(H_i \circ q(x)) \]

where $\mathcal{H}_i = F^{-1}(q_i)$ with $F^{-1}(q_i) > C$.}

While through Eq. (A.3), the covariance between $\overline{q}_i$ and $\overline{q}_j$ for $k < k' < 1, \ldots, K$ is given by
and Proposition 1, respectively.

Proof. From above equations we have that

\[ q_{ij} = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - \frac{n}{2(n+2)} \right) q_{ij} + O \left( \frac{1}{n^2} \right) \]

and

\[ \text{Cov} \left( q_{ij}, q_{ij} \right) = \frac{1}{m} \sum_{i=1}^{m} \left( H_i(q_{ij}) + \frac{1}{8(n+2)} - \frac{M}{c^2} \right) + O \left( \frac{1}{n^2} \right) \]

for all \((q_{ij}, q_{ij}) \in [0, 1]^2\) with \(q_{ij} < q_{ij}^\ast\).

From above equations we have that

\[ \mathbf{E} \left( \tilde{q}_i \right) \xrightarrow{m} q_{ij}(x) \quad (A.4) \]

and

\[ \text{Cov} \left( \tilde{q}_i, \tilde{q}_i \right) \xrightarrow{m} \frac{1}{m} \mathbf{V}(q_{ij}(x), q_{ij}(x)) \quad (A.5) \]

Proof. Proof of Theorem 2.

The almost sure convergence of \(\tilde{q}_i\) is established applying the results of strong consistency of \(q_{ij}(x)\) and \(q_{ij}(z)\) from Theorem 1 and Proposition 1, respectively.

The asymptotic normality of \(\tilde{q}_i\) is obtained as follows

\[ \sqrt{m} \left( \tilde{q}_i - q_{ij}(z) \right) \xrightarrow{\mathbf{V}} \mathcal{N}(0, \sigma^2) \quad (A.6) \]

where \((q_{ij}, q_{ij}) \in [0, 1]^2\) and \((k, k')\)-element of \(\Sigma\) is given by

\[ \Sigma_{kk'} = \sigma^2(q_{ik}(x), q_{ij}(x)) \]

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

\[ \sum_{i=1}^{m} \mathbf{E} \left( X_{ij} - \mathbf{E} \left( X_{ij} \right) \right)^2 = \frac{1}{m} \sum_{i=1}^{m} \mathbf{V} \left( X_{ij} \right) \xrightarrow{m \to \infty} 0 \]

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

\[ \frac{1}{m} \sum_{i=1}^{m} \mathbf{E} \left( X_{ij} - \mathbf{E} \left( X_{ij} \right) \right)^2 = \frac{L}{m} \sum_{i=1}^{m} \mathbf{V} \left( X_{ij} \right) \xrightarrow{m \to \infty} 0 \]

given that \(\mathbf{V} \left( X_{ij} \right) \xrightarrow{m \to \infty} 0\).

Therefore the Lyapunov’s condition is satisfied. \(\square\)

References


