Statistical Properties of the Quantile Normalization Method for Density Curve Alignment

Santiago Gallón, Jean-Michel Loubes, Elie Maza

Statistical Properties of the Quantile Normalization Method for Density Curve Alignment

S. Gallón[*], J-M. Loubes[†]and E. Maza[‡]

May 16, 2011

Abstract

We present a proof for the quantile normalization method proposed by Bolstad et al. [2] 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 an application to density curve registration of the new method proposed in Dupuy et al. [6], the structural expectation. Moreover, when this method fails in some case of mixture, we propose a new methodology to cope with this issue.

Keywords: Quantile normalization; Structural expectation; Curve registration; Density curve alignment; Order statistics.
Subject Class. MSC-2010: 62G05, 62G30, 62G20.

1 Introduction

The outcome of a statistical process is often a sample of curves $\{f_i,\ i = 1, \ldots, m\}$ showing an unknown common structural pattern, f, which characterizes the

[*] Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia; and Institut de Mathématiques de Toulouse, Université Paul Sabatier Toulouse 3. E-mail: santiagog@udea.edu.co

[†] Institut de Mathématiques de Toulouse, Université Paul Sabatier Toulouse 3. E-mail: jean-michel.loubes@math.univ-toulouse.fr

[‡] École Nationale Supérieure Agronomique de Toulouse. Genomic & Biotechnology of the Fruit Laboratory. UMR 990 INRA/INP-ENSAT. E-mail: elie.maza@ensat.fr

common behavior of the sample. Examples are numerous, among others growth curves analysis in biology and medicine, quantitative analysis of microarrays in molecular biology and genetics, speech signals recognition in engineering, study of expenditure and income curves in economics. Hence, in recent decades, there has been a growing interest to develop statistical methodologies which enables to recover from the observation functions a single "mean curve" that conveys all the information of the data.

A major difficulty comes from the fact that there are systematic variations among the sample of curves associated to both amplitude (variation in the y-axis) or phase (variation in the $x-$axis) variations, like is illustrated in Figure 1 for a sample of three simulated curves, which prevent any direct extraction of classical statistics such as the mean, median, correlations or any other standard statistical multivariate procedure such as principal component and canonical correlation analysis. See Kneip and Gasser [11] or Ramsay and Silverman [17] and references therein for more details. Indeed the classical cross-sectional mean does not provide a consistent estimate of the function of interest f when the phase variations are ignored since it fails to capture the structural characteristics in the sample of curves as is quoted in Ramsay and Li [16] and is also illustrated in Figure 1. Hence curve registration (also called curve alignment, structural averaging, and time warping) methods have been proposed in the statistical literature. We refer to Silverman [20], Gasser and Kneip [8], Wang and Gasser [25], Kneip et al. [13] Rønn [18], Liu and Müller [15], Gamboa et al. [7], James [10], Kneip and Ramsay [12], and Dupuy et al. [6] just to name a few.

The same kind of problem occurs when dealing with a sample of density curves with variations between curves which are not correlated to the phenomena which is studied and thus need to be removed. In bioinformatics and computational biology, a popular method to reduce this kind of variability is known as normalization and it is widely applied in high density oligonucleotide array data in biomedical research. It is fully described in Bolstad et al. [2]. Among the many normalization methods there is the popular quantile normalization method proposed by Bolstad et al. [2] which uses the quantile-quantile plot extended to m dimensions. The procedure consists in assuming that there is an underlying common distribution followed by the curves and obtaining a mean distribution through the projection of the jth empirical quantile vector of sample quantiles, $\hat{\boldsymbol{q}}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{m,j})^\top$, onto the vector $\boldsymbol{d} = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$, given by $\mathrm{proj}_{\boldsymbol{d}} \hat{\boldsymbol{q}}_j = (\frac{1}{m}\sum_{i=1}^m \hat{q}_{i,j}, \ldots, \frac{1}{m}\sum_{i=1}^m \hat{q}_{i,j})^\top$, 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 $\boldsymbol{d}$. See Bolstad et al. [2] and Irizarry et al. [9] for applications of this method.

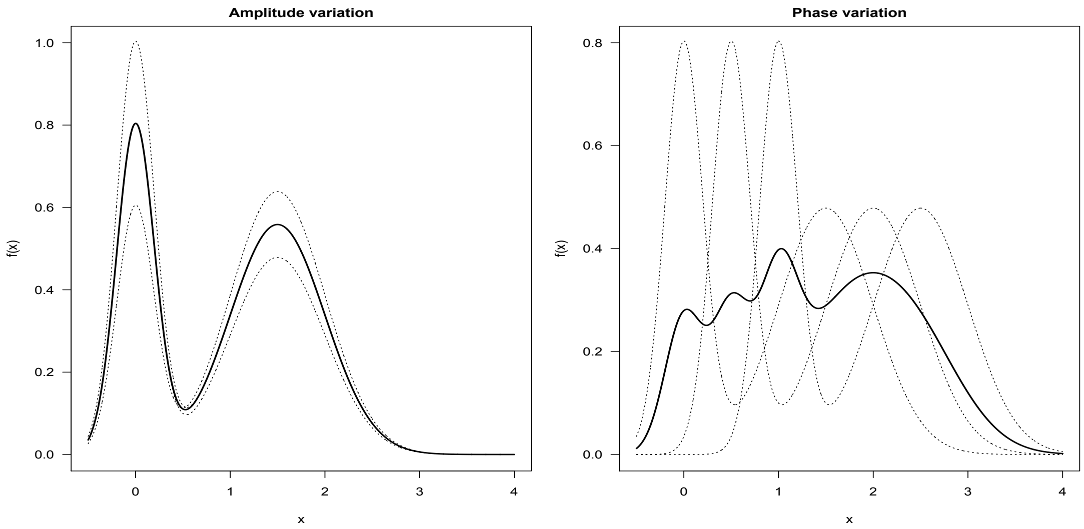

Figure 1: Examples of amplitude (on the left side) and phase (on the right side) variations. The solid line corresponds to the cross-sectional mean function.

The Figure 2, for example, plots the densities of a sample of 18 two-color microarrays after normalization within arrays. The dot-dashed and solid lines through densities corresponds to cross-sectional mean and quantile normalization, respectively. The quantile normalization method has the advantages to be simple and quick with respect to others normalization procedures. However its statistical properties have not been derived yet up to our knowledge.

In this paper we point out that the quantile normalization can be seen as a particular case of the structural median procedure, described in Dupuy et al. [6]. We study the large sample properties of the quantile normalization method and prove its consistency. In addition, when this procedure fails, we propose a variation of this method to still recover a mean density and thus improve one 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 method and derive the asymptotic properties of the quantile normalization method. Section 4 is devoted to present the manifold type pattern extraction. The results of a simulation study showing the situation in which the quantile normalization method does not work properly to represent the behavior of a sample of density curves are reported in Section 5. Finally, in Section 6 we apply the methods to normalize two-channel spotted microarray densities. All

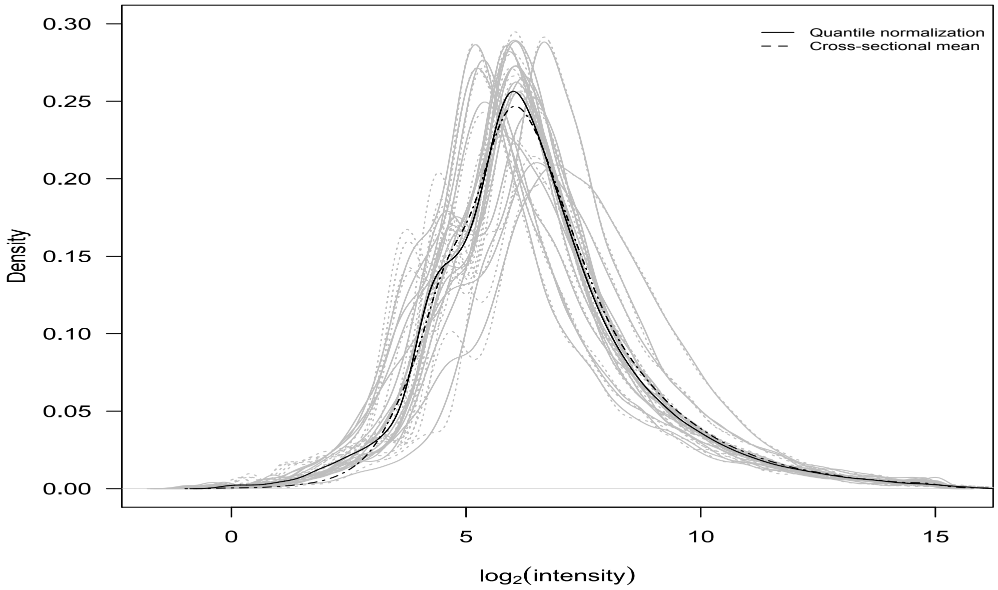

Figure 2: Densities for individual-channel intensities for two-color microarray data after normalization within arrays. Dotted and solid gray lines correspond to the "green" and "red" color arrays, respectively.

the proofs are gathered in Section 7.

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 $f_i\colon \mathbb{R} \to [0,+\infty)$ and distribution function $F_i\colon \mathbb{R} \to [0,1]$. 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 $F\colon \mathbb{R} \to [0,1]$ by an invertible and differentiable warping function H_i, of the following manner

$$F_i(t) = \Pr(X_{ij} \leqslant t) = F \circ H_i^{-1}(t), \qquad i = 1,\ldots,m,\ j = 1,\ldots,n \qquad (1)$$

where H_i is random in the sense that $H_1,\ldots,H_m$ is an i.i.d random sample from a (non parametric) warping stochastic process $\mathcal{H}\colon \Omega \to \mathcal{C}(\mathbb{R})$ defined on a 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 Gamboa et al. [7], and Dupuy et al. [6].

Since the model (1) to estimate the function f is not identifiable (see Dupuy et al. [6]), we consider the *structural expectation (SE)* of the quantile function to overcome this problem as

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

Inverting equation (1) leads to

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

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

$$q_i(\alpha) = F_i^{-1}(\alpha) = \inf\left\{x_{ij} \in \mathbb{R}\colon F_i(x_{ij}) \geqslant \alpha\right\}, \qquad 0 \leqslant \alpha \leqslant 1. \tag{4}$$

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

$$\overline{q_m(\alpha)} = \frac{1}{m}\sum_{i=1}^{m} q_i(\alpha), \qquad 0 \leqslant \alpha \leqslant 1. \tag{5}$$

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

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

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

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

The following theorem deals with the asymptotic behavior of the estimator (5).

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

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

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

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

where the (k,k')-element of the asymptotic variance-covariance matrix $\mathbf{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ 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 $X_{i,1}, \ldots, X_{i,n}$ from $F_i(x)$ for $i = 1, \ldots, m$ are observed. The i-th 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 $X_{i,1:n} \leqslant X_{i,2:n} \leqslant, \ldots, \leqslant X_{i,n:n}$, hence the estimation of the quantile functions, $q_i(\alpha)$, is obtained by

$$\begin{aligned} \hat{q}_{i,n}(\alpha) &= \mathbb{F}_{i,n}^{-1}(\alpha) = \inf\{x_{ij} \in \mathbb{R} \colon \mathbb{F}_{i,n}(x_{ij}) \geqslant \alpha\} \\ &= X_{i,j:n} \quad \text{for} \quad \frac{j-1}{n} < \alpha \leqslant \frac{j}{n}, \qquad j = 1, \ldots, n. \end{aligned} \tag{6}$$

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

Finally, the estimator of the structural quantile is given by

$$\bar{\hat{q}}_j = \frac{1}{m}\sum_{i=1}^m \hat{q}_{i,j} = \frac{1}{m}\sum_{i=1}^m X_{i,j:n}, \qquad j = 1, \ldots, n. \tag{7}$$

Note that, this procedure corresponds to the so-called quantile normalization method proposed by Bolstad et al. [2].

Based on sample quantiles we can obtain a "mean" distribution through the projection of the empirical quantile vector of the j-th sample quantiles, $\hat{\boldsymbol{q}}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{m,j})^\top$, onto the vector $\boldsymbol{d} = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$, given by

$\mathrm{proj}_{\boldsymbol{d}}\hat{\boldsymbol{q}}_j = (\frac{1}{m}\sum_{i=1}^m \hat{q}_{i,j}, \ldots, \frac{1}{m}\sum_{i=1}^m \hat{q}_{i,j})^\top$. 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 $\boldsymbol{d}$.

The asymptotic behavior of the quantile normalization estimator (7) is established by the next theorem.

Theorem 2. *The quantile normalization estimator $\bar{\hat{q}}_j$ is strongly consistent*

$$\bar{\hat{q}}_j \xrightarrow[m,n\to\infty]{a.s} q_{SE}(\alpha_j), \qquad j = 1, \ldots, n,$$

and under the assumptions of compactly central data, $|X_{i,j:n} - \mathbf{E}\left(X_{i,j:n}\right)| \leqslant L < \infty$ for all i and j, and $\frac{\sqrt{m}}{n} \to 0$, it is asymptotically Gaussian. Actually, for any $K \in \mathbb{N}$ and fixed $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$,

$$\sqrt{m}\begin{bmatrix} \bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1) \\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K) \end{bmatrix} \xrightarrow[m,n\to\infty]{\mathcal{D}} \mathcal{N}_K\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)$$

where the (k,k')-element of the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ 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 $f_i\big(q_i(\alpha)\big) > 0$ 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,*

$$\hat{q}_{i,n}(\alpha) \xrightarrow[n\to\infty]{a.s.} q_i(\alpha)$$

and asymptotically Gaussian

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

where $\big(H_i^{-1}\big)'(z) = \frac{dH_i^{-1}(z)}{dz} = \frac{1}{H_i' \circ H_i^{-1}(z)}$.

4 Using manifold registration for density alignment

One of the major issue in registration problems is to find the fitting criterion which enables to give a sense to the notion of mean of a sample of points. Hence, it seems natural to consider that the data belongs to a non euclidean set and to look for the most suitable corresponding distance. A natural framework is given by a manifold embedding where the geodesic distance provides a natural way to compare two objects from this manifold. This point of view has been developed in Dimeglio et al. [5].

Density registration lies in the field of applications of these technics. For this, recall that we observe X_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$ random variables. First, we sort the observations for each sample i, and denote by $X_{(i).}$ the sorted vector $X_{i(1)}, \ldots, X_{(i)n}$ and thus we consider the array of observations $(X_{(1).}, \ldots, X_{(m).})$. All previous methods aim at finding a good representative for these vectors. The mean of the sorted vectors gives the structural quantile which corresponds to the Boldstad's normalization method described in Section 3, but the manifold point of view gives an alternative framework to define this mean pattern. Actually, consider that the vectors $\{X_{(i).}, i = 1, \ldots, m\}$ are embedded into a manifold $\mathcal{M}$ with geodesic distance d_g. The natural mean is defined as

$$\hat{X}_m = \arg\min_{x \in \mathcal{M}} \sum_{i=1}^{m} d_g(x, X_{(i).}),$$

which is estimated by the estimator defined in Dimeglio et al. [5], by approximating the geodesic distance using an ISOMAP-type graph approximation, following [23]. Even if the theoretical properties of this estimate are difficult to understand due to the difficulties inherent to the graph-type geodesic approximation, its practical properties for quantile normalization will be studied in the next sections.

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. [2] works and the situation in which it has problems to represent properly the behavior of the sample of density curves.

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

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

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

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

1. For each $i = 1, \dots, m$ 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 (variation in location): $U(\mu_{il} - 0.15, \mu_{il} + 0.15)$, $l = 1, 2, 3$.

Case 2 (variation in scale): $U(\sigma_{il} - 0.35, \sigma_{il} + 0.35)$, $l = 1, 2$ and $U(\sigma_{i3} - 0.5, \sigma_{i3} + 0.5)$.

Case 3 (variation in probability weight): $U(\omega_{il} - 0.1, \omega_{il} + 0.1)$, $l = 1, 2$.

where U is a uniformly distributed random variable.

The Figure 3 shows the simulated density and distribution functions to each case. The estimated "mean" density and distribution functions using the quantile and manifold (described below) 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 and 2, i.e when the probability weights do not vary among the densities, $\omega_{il} = \omega_{i'l}$, $l = 1, 2, 3$ for $i, i' = 1, \dots, m$. whereas it fails in the case 3.

To overcome the drawback corresponding to case 3, 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 Dimeglio et al. [5]. As we can see in the Figure 3, the estimation of the "mean" density f through the manifold normalization method improves the normalization of the sample of densities for the case of variations in probability weight (case 3) capturing properly the structural mean behavior of sample of curves.

6 Application

In this section, we apply and compare the quantile and manifold methods to normalize two-channel (also two-color) spotted microarrays in order to identify and remove systematic variations retaining the biological signals. For a detailed description on two-channel spotted microarrays see Yang and Thorne [27] and Yang and Paquet [26].

The two-channel spotted microarray data were provided by the Toulouse School of Agronomy (ENSAT). The data base contains 13056 rows corresponding to the spots (probes) and 18 columns corresponding to the intensities for the arrays. We used the `limma` Bioconductor software package based on the R statistical programming language, to read and carry out the quality assessment of the intensity data (Smyth and Speed [22] and Smyth [21]). The Figure 4 shows the density plots for individual-channel intensities of two-color microarray data. Dotted and solid lines correspond to the "green" and "red" color arrays, respectively.

For two-channel microarrays the normalization between arrays usually occurs after normalization within arrays to remove from the expression measures any systematic trends which arise from the microarray technology rather than from differences between the probes. This also make intensities consistent within each array. Smyth and Speed [22] review the normalization methods within arrays. The expression measures for each two-color microarray were normalized using the loess method (see Smyth and Speed [22] and Yang and Paquet [26]). The Figure 5 plots the densities for each two-color microarray after loess normalization. The normalization between arrays applying the quantile and manifold normalization are plotted in the same Figure (bold solid and dashed lines) in order to ensure that the intensities have the same empirical distribution across arrays and across channels (Yang and Thorne [27]). As we can see, the manifold normalization captures the structural characteristics of the densities, in particular those that corresponding to the inflection points present in the individual arrays.

7 Appendix

Proof. Proof of Theorem 1.

First note from equation (3) that

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

where $q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} \colon F(x) \geqslant \alpha\}$, $0 \leqslant \alpha \leqslant 1$, 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 i.i.d. random variables in a separable Banach space $\mathcal{B} = \mathcal{C}([0,1])$ and applying the next corollary, the almost sure convergence of $\overline{q_m(\alpha)}$ is guaranteed.

Corollary (Corollary 7.10, Ledoux and Talagrand [14]). Let W be a Borel random variable with values in a separable Banach space $\mathcal{B}$. Then $S_m/m \xrightarrow[m\to\infty]{a.s.} 0$ if and only if $\mathbf{E}\|W\| < \infty$ and $\mathbf{E}W = 0$.

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 $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$,

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

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

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

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

Following van der Vaart and Wellner [24], the tightness moment condition to ensure the weak convergence is given by

$$\begin{aligned}
&\mathbf{E}\left[\left|\sqrt{m}\left(\overline{q_m(\alpha)}-\mathbf{E}\overline{q_m(\alpha)}\right)-\sqrt{m}\left(\overline{q_m(\beta)}-\mathbf{E}\overline{q_m(\beta)}\right)\right|^2\right]\\
&=\mathbf{E}\left[\left|\sqrt{m}\Big(\left(\overline{q_m(\alpha)}-\mathbf{E}\overline{q_m(\alpha)}\right)-\left(\overline{q_m(\beta)}-\mathbf{E}\overline{q_m(\beta)}\right)\Big)\right|^2\right]\\
&=\mathbf{E}\left[m\left|\left(\frac{1}{m}\sum_{i=1}^{m}H_i\circ q(\alpha)-\phi\circ q(\alpha)\right)\right.\right.\\
&\qquad\qquad\left.\left.-\left(\frac{1}{m}\sum_{i=1}^{m}H_i\circ q(\beta)-\phi\circ q(\beta)\right)\right|^2\right]\\
&=\mathbf{E}\left[m\left|\frac{1}{m}\sum_{i=1}^{m}(H_i-\phi)\circ\big(q(\alpha)-q(\beta)\big)\right|^2\right]\leqslant C_1C_2\left|\alpha-\beta\right|^2,
\end{aligned}$$

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 Serfling [19] or David and Nagaraja [4] 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 $\hat{q}_n(\alpha)\xrightarrow[n\to\infty]{a.s.}q(\alpha)$.

Therefore $\hat{q}_{i,n}(\alpha)\xrightarrow[n\to\infty]{a.s.}q_i(\alpha)$ for $i=1,\dots,m$.

Theorem (Asymptotic normality of order statistics). For a fixed $0<\alpha<1$, assume F is continuously differentiable at the αth population quantile, $q(\alpha)$, $f\big(q(\alpha)\big)>0$, and $n^{-1/2}(j/n-\alpha)=o(1)$. Then $\sqrt{n}\big(X_{j:n}-q(\alpha)\big)\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}\left(0,\frac{\alpha(1-\alpha)}{f^2(q(\alpha))}\right)$, 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 $i=1,\dots,m$ we have

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

that conditioned to a fixed H_i implies

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

where $\big(H_i^{-1}\big)^{'}(z) = \frac{dH_i^{-1}(z)}{dz} = \frac{1}{H_i^{'} \circ H_i^{-1}(z)}$. ■

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 $i = 1, \dots, m$ are given by (see, for example, David and Nagaraja [4] or Arnold et al. [1])

$$\begin{aligned}\mathbf{E}(X_{i,j:n}) = q_{i,j} + \frac{\alpha_j(1-\alpha_j)}{2(n+2)} q_{i,j}^{''} + \frac{\alpha_j(1-\alpha_j)}{(n+2)^2}\Bigg[\frac{1}{3}\big((1-\alpha_j) - \alpha_j\big) q_{i,j}^{'''} \\ + \frac{1}{8}\alpha_j(1-\alpha_j) q_{i,j}^{(4)}\Bigg] + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbf{Var}(X_{i,j:n}) = \frac{\alpha_j(1-\alpha_j)}{n+2} q_{i,j}^{'2} + \frac{\alpha_j(1-\alpha_j)}{(n+2)^2}\Bigg[2\big((1-\alpha_j) - \alpha_j\big) q_{i,j}^{'} q_{i,j}^{''} \\ + \alpha_j(1-\alpha_j)\left(q_{i,j}^{'} q_{i,j}^{'''} + \frac{1}{2} q_{i,j}^{''2}\right)\Bigg] + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{9}$$

$$\begin{aligned}\mathbf{Cov}(X_{i,j:n}, X_{i,s:n}) = \frac{\alpha_j(1-\alpha_s)}{n+2} q_{i,j}^{'} q_{i,s}^{'} + \frac{\alpha_j(1-\alpha_s)}{(n+2)^2}\Bigg[\big((1-\alpha_j) - \alpha_j\big) q_{i,j}^{''} q_{i,s}^{'} \\ + \big((1-\alpha_s) - \alpha_s\big) q_{i,j}^{'} q_{i,s}^{''} + \frac{1}{2}\alpha_j(1-\alpha_j) q_{i,j}^{'''} q_{i,s}^{'} \\ + \frac{1}{2}\alpha_s(1-\alpha_s) q_{i,j}^{'} q_{i,s}^{'''} + \frac{1}{2}\alpha_j(1-\alpha_s) q_{i,j}^{''} q_{i,s}^{''}\Bigg] + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{10}$$

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

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

and

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

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

This approximation method is due to David and Johnson [3] where they derived approximations of order $(n+2)^{-3}$. Additionally, note that the asymptotic means, variances, and covariances correspond to the first terms of equations (8), (9), and (10), respectively (David and Nagaraja [4]).

Using the approximation in equation (8), the mean of $\bar{\hat{q}}_j$ is calculated as

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

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

While through equation (10), the covariance between of $\bar{\hat{q}}_{j_k}$ and $\bar{\hat{q}}_{j_{k'}}$ for $k \neq k'$

$k = 1, \dots, K$ is given by

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

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

From above equations we have that

$$\mathbf{E}\big(\bar{\hat{q}}_j\big) \xrightarrow[n\to\infty]{} q_{SE}(\alpha_j) \tag{11}$$

and

$$\mathbf{Cov}\Big(\bar{\hat{q}}_{j_k}, \bar{\hat{q}}_{j_{k'}}\Big) \xrightarrow[n\to\infty]{} \frac{1}{m}\vartheta\big(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\big). \tag{12}$$

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

$$\sqrt{m}\frac{\left(\bar{\hat{q}}_j - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} = \sqrt{m}\frac{\left(\frac{1}{m}\sum_{i=1}^{m} X_{i,j:n} - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}$$

$$= \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\left(X_{i,j:n} - \mathbf{E}\left(X_{i,j:n}\right)\right)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} + \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left(X_{i,j:n}\right) - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}$$

$$= \frac{\left(\sum_{i=1}^{m}\left(X_{i,j:n} - \mathbf{E}\left(X_{i,j:n}\right)\right)\right)\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathbf{Var}\left(X_{i,j:n}\right)}}{\sqrt{\sum_{i=1}^{m}\mathbf{Var}\left(X_{i,j:n}\right)}\sqrt{\vartheta\big(q(\alpha_j)\big)}} + \frac{\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left(X_{i,j:n}\right) - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}$$

$$= \frac{\left(\sum_{i=1}^{m} X_{i,j:n} - \sum_{i=1}^{m}\mathbf{E}\left(X_{i,j:n}\right)\right)}{\sqrt{\sum_{i=1}^{m}\mathbf{Var}\left(X_{i,j:n}\right)}}\frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathbf{Var}\left(X_{i,j:n}\right)}}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} + \frac{\sqrt{m}\left(\frac{1}{8(n+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{n^2}\right)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}$$

Given that $\mathbf{Var}\left(X_{i,j:n}\right) \xrightarrow[n\to\infty]{} \vartheta\big(q(\alpha_j)\big)$, and under the assumption $\frac{\sqrt{m}}{n} \to 0$ we obtain, by the Lindeberg-Feller's Central Limit Theorem −CLT− for independent but not identically distributed random variables to the independent random variables $X_{1,j:n}, \ldots, X_{m,j:n}$, that

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

that in multivariate terms is expressed as

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

where $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$ and the (k,k')-element of $\mathbf{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\big)$. The Lindeberg-Feller's Central Limit Theorem holds if the

Lyapunov's condition

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

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

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

given that $\mathbf{Var}\left(X_{i,j:n}\right)\xrightarrow[n\to\infty]{}\vartheta\big(q(\alpha_j)\big)$.

Therefore the Lyapunov's condition is satisfied and consequently the CLT is verified. ■

References

[1] B. Arnold, N. Balakrishnan, and H. Nagaraja. *A First Course in Order Statistics*, volume 54. Classics in Applied Mathematics, SIAM, 2008. Philadelphia.

[2] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[3] F. N. David and N. L. Johnson. Statistical treatment of censored data part I. Fundamental formulae. *Biometrika*, 41(1/2):228–240, 1954.

[4] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, 3rd edition, 2003. New Jersey.

[5] C. Dimeglio, J.-M. Loubes, and M. E. Manifold embedding for curve registration. *Submitted to The Annals of Applied Statistics*, 2011.

[6] J. Dupuy, J.-M. Loubes, and E. Maza. Non parametric estimation of the structural expectation of a stochastic increasing function. *Statistics and Computing*, 21:121–136, 2011.

[7] F. Gamboa, J.-M. Loubes, and E. Maza. Semi-parametric estimation of shits. *Electronic Journal of Statistics*, 1:616–640, 2007.

[8] T. Gasser and A. Kneip. Searching for structure in curve sample. *Journal of the American Statistical Association*, 90:1179–1188, 1995.

[9] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[10] G. James. Curve alignment by moments. *The Annals of Applied Statistics*, 1(2):480–501, 2007.

[11] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305, 1992.

[12] A. Kneip and J. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008.

[13] A. Kneip, X. Li, X. MacGibbon, and J. Ramsay. Curve registration by local regression. *Canadian Journal of Statistics*, 28(1):19–29, 2000.

[14] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete 3. Folge. A Series of Modern Surveys in Mathematics.* Springer-Verlag, 1991. Berlin.

[15] X. Liu and H. Müller. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699, 2004.

[16] J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society*, 60(2):351–363, 1998.

[17] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, New York, 2nd edition, 2005.

[18] B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society. Series B*, 63(2):243–259, 2001.

[19] R. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, 1980. New York.

[20] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society. Series B*, 57 (1):673–689, 1995.

[21] G. K. Smyth. `limma`: linear models for microarray data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420, New York, 2005. Springer.

[22] G. K. Smyth and T. P. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, 2003.

[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[24] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, 1996. New York.

[25] K. Wang and T. Gasser. Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27(2):439–460, 1999.

[26] Y. H. Yang and A. C. Paquet. Preprocessing two-color spotted arrays. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 49–69, New York, 2005. Springer.

[27] Y. H. Yang and N. P. Thorne. Normalization for two-color cDNA microarray data. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40, pages 403–418, New York, 2003. IMS Lecture Notes - Monograph Series.

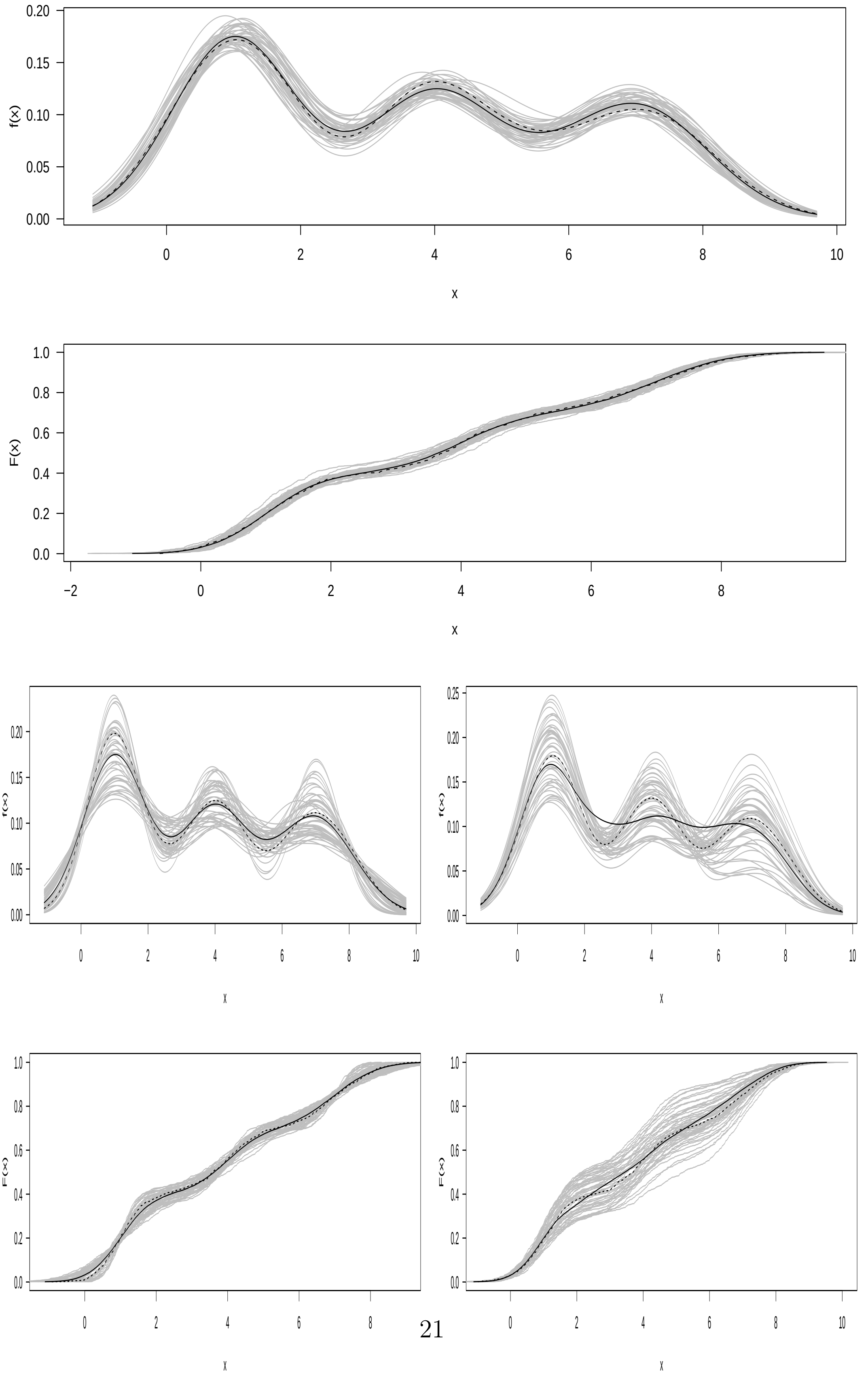

Figure 3: Simulation results. Bold solid and dashed lines correspond to quantile and manifold normalization methods, respectively.

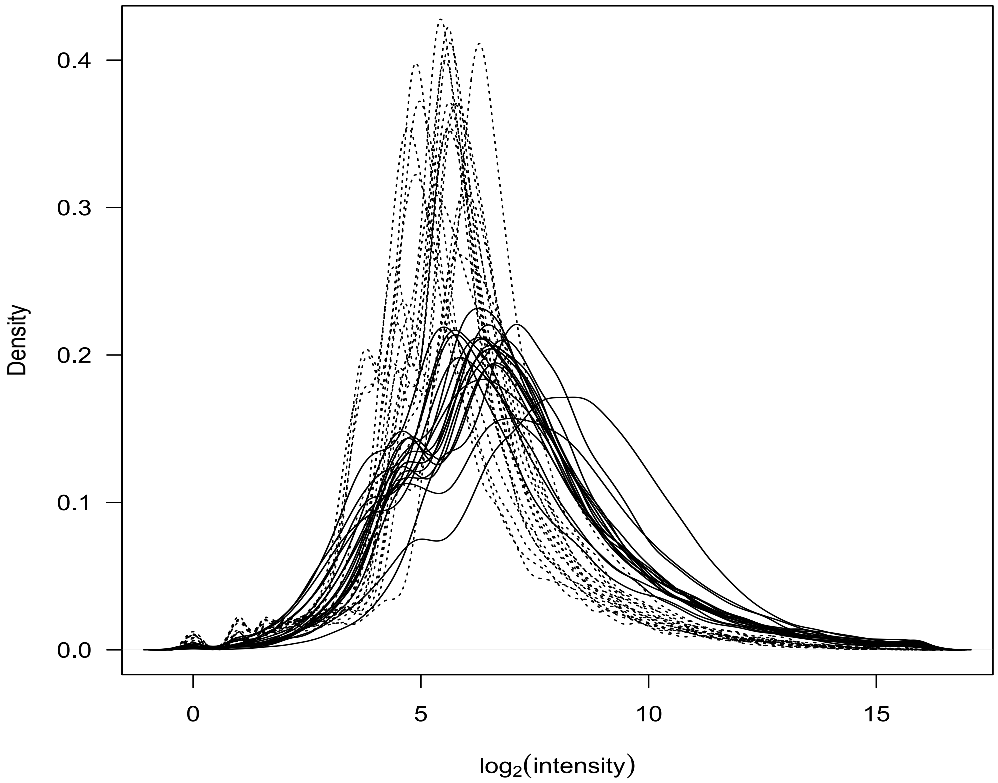

Figure 4: Densities for individual-channel intensities for two-color microarray data. Dotted and solid lines correspond to the “green” and “red” color arrays, respectively.

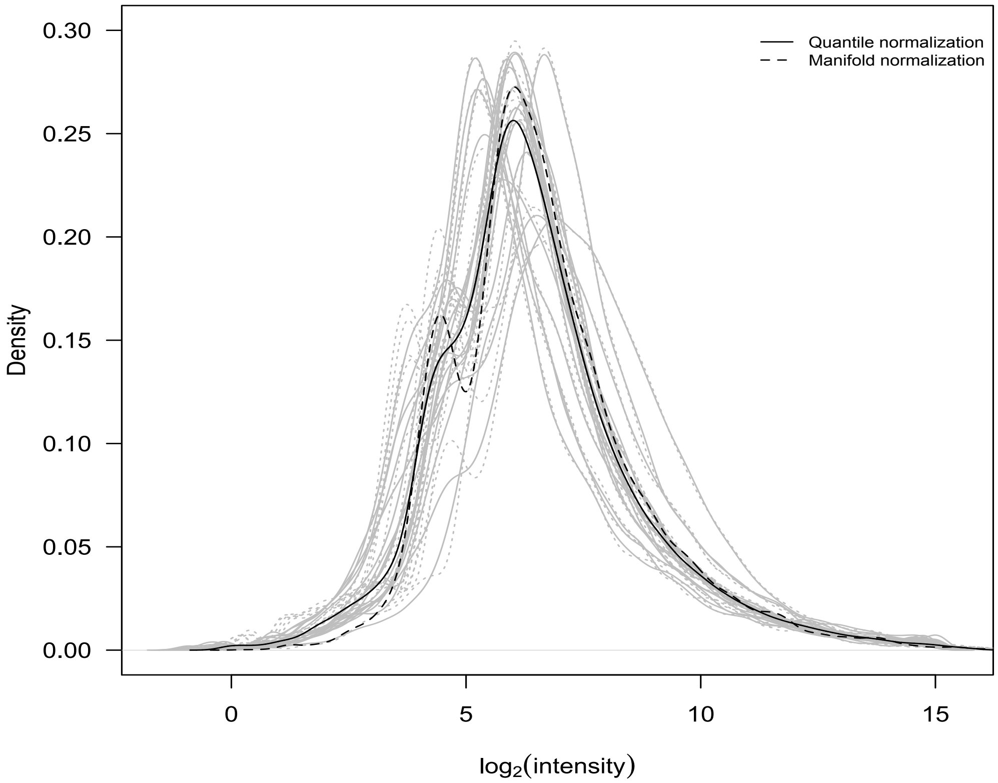

Figure 5: Densities for individual-channel intensities for two-color microarray data after loess normalization within arrays. Dotted and solid lines correspond to the "green" and "red" color arrays, respectively.