Functional Density Synchronization

Zhen Zhang
Abbott Vascular Inc., Santa Clara
Hans-Georg Müller
University of California, Davis

November 2010

ABSTRACT. Estimating an overall density function from repeated observations on each of a sample of independent subjects or experimental units is of interest. An example is provided by biodemographic studies, where one observes age-at-death for several cohorts of flies. Cohorts are kept in separate cages, which form the experimental units. Time-variation then is likely to exist between the cohort densities and hazard rates due to cage effects on aging. Given the densities of age-at-death for the individual cohorts, one aims to obtain an estimate for the underlying overall density and hazard rate. In microarray gene expression experiments, similar problems arise when addressing the need for normalization of probe-level data from different arrays. Conventional methods, such as the cross-sectional average density, ignore time variation and hence are often not representative for such data. We view densities as functional data and model individual densities as warped versions of an underlying overall density, where the observed densities are assumed to be realizations of an underlying stochastic process. Quantile-synchronized distribution functions are obtained from an inverse warping mapping, based on quantile synchronization, leading to quantile-synchronized density and hazard functions. Kernel type smoothing methods with plug-in bandwidth selection can be used for estimating the components of the model. Asymptotic properties of the synchronized density estimates are derived. Simulation results show that functional density synchronization is often advantageous when compared to conventional density averaging or simple time-shift warping. Our approach complements previous quantile normalization methods used for microarray expression data and is illustrated with both longevity data obtained for 54 cohorts of mexflies (Mexican fruit flies) and gene expression data of the Ts1Cje mouse study for Down syndrome.

Key words: Age-at-death; density estimation; curve registration; hazard rate; functional average; functional data analysis; gene expression; kernel smoothing; mean density; mortality; normalization; quantile function; warping

---

Correspondence: Hans-Georg Müller, *Email*: mueller@wald.ucdavis.edu, *Address*: Department of Statistics, One Shields Ave., University of California, Davis, CA 95616, USA.

1 Introduction

In biomedical studies, density and hazard functions of age-at-death and density functions of gene expressions are often of primary interest. In economics, densities of income distributions across various populations are studied, and in this context, Kneip and Utikal (2001) extended functional data analysis to the case of samples of density functions by utilizing functional principal component analysis (Rice and Silverman 1991). In this setting, one observes many densities and these densities are then considered as functional data, i.e., as a sample of realizations of an underlying stochastic process.

In recent years, statistical methodology for densities as functional data was developed by various authors. These include Delicado (2007) and Nerini and Ghattas (2007), who studied functional ANOVA and functional classification, respectively, when data are densities, and Delicado (2011), who investigated dimension reduction for functional data, including functional principal components and representations based on multidimensional scaling. Related work on functional hazard analysis can be found in Quintela-del-Río (2008), Ferraty, Rahbi and Vieu (2008) and Chiou and Müller (2009), and on functional classification in Cuevas, Febrero and Fraiman (2007). Manté et al. (2005) consider functional classification for the case of a sample of distribution functions within a biological data analysis framework. Further related descriptions and references can be found in Ferraty and Vieu (2006), and a general introduction to functional data analysis is given in Ramsay and Silverman (2005). While in some of these approaches density, hazard and distribution functions are viewed as functional data, time warping is usually not considered to be a central issue for these functional objects. The purpose of this paper is to demonstrate that including time warping components in functional density and hazard analysis is often quite beneficial and may lead to a better understanding of the data. We introduce *Quantile Synchronization*, a simple and straightforward density warping approach in order to implement this objective.

A motivating example is provided by data that were generated in a recent biodemographic study, which is described in more detail in Section 6.1. In this experiment, age-at-death data for large cohorts of Mexican fruit flies were obtained to study longevity and survival of these flies. The data consist of lifetables for 54 cages of male flies, whose survival was monitored over a 174 day period. For each cage, the number of deaths per day was recorded. As the flies in a cage interact with each other, the natural unit of analysis is cage (cohort). Each cohort, corresponding to a sampling unit, possesses an associated density or hazard function, for the age-at-death distribution of the flies in the cohort. One may then construct a density or hazard function from the observed age-at-death (lifetime) data for the flies in a cohort, separately for each of the cohorts.

Figure 1 displays the observed hazard rates for the first 10 cohorts, obtained with the method

described in Müller et al. (1997). The hazard rates are seen to share common features and there is also evidence for significant time-variation between cohorts. Adopting the viewpoint that each density is a random function that may include time variation, a basic statistical question arises as to how to arrive at a reasonable overall density or hazard rate for this sample of random densities. A solution of this problem for densities will imply a solution for hazard rates as well, due to the 1-to-1 relationship between densities and hazard rates. In biodemographic applications, one seeks representative population *trajectories of mortality*, which need to be assembled from many observed cohorts (Vaupel et al., 1998). The variation inherent in various cohort trajectories is likely to be composed of both temporal and amplitude variation. We propose here to reduce the temporal variation by time-synchronizing the cohort trajectories prior to an averaging step by using quantile synchronization, which emerges as the most natural way to approach this problem.

Another example is provided by microarray gene expression experiments, where the expression levels recorded in oligonucleotide arrays are often of interest and a problem analogous to time-variation arises when multiple arrays are involved. The problem here is to adjust for the variation across densities obtained from different arrays. Such variation is usually due to non-biological differences between arrays and is considered a nuisance. At best this variation is of minor interest in the study of gene expressions, where the objective is to detect over/under-expression of genes (for a detailed discussion of sources of array variation, see Hartemink et al., 2001 or Irizarry et al., 2003). Therefore, a preprocessing step that eliminates undesirable array density variation is often required. Such preprocessing is also called *normalization* in microarray data analysis, and so-called quantile normalization has been proposed for the purpose of normalizing the various observed densities (Bolstad et al., 2003). This provides another instance in which quantile synchronization is highly useful.

Conventional methods to find a representative density, such as the cross-sectional mean of a sample of densities, are often inadequate, as they ignore the differences in time scale between the individual densities. Our goal is to obtain a reasonable overall ("normalized") density that summarizes the densities that are observed for each individual cohort and that may serve as a representative of the population of cohorts. In order to identify the average density in the presence of random time scale variation, we propose to apply a version of curve synchronization or warping, also known as curve registration (Sakoe and Chiba, 1978; Ramsay and Li, 1998), or alignment (Kneip and Gasser, 1992).

In time-dynamic biological systems, time-variation between individual trajectories is a common feature. A typical example is growth curve. It is well-known that in human growth, different individuals tend to have different growth schedules. Events shared by all subjects, such as growth spurts, typically occur at different individual ages or times (Gasser et al., 1984). In such situations, subject-specific time scales adequately reflect internal dynamics of the process, as each individual or cohort progresses

according to its own time scale. To obtain meaningful overall densities, it is then appropriate to warp individual time scales to a common time scale, the "synchronized time". A different rationale for the same objective is the need for normalization in gene expression arrays, as discussed above, where "time" corresponds to gene expression level.

Several synchronization methods have been proposed. The "structural average", a version of the landmark method, was studied in Kneip and Gasser (1992), with more recent extensions in Gervini and Gasser (2005). In this approach, individual curves are aligned or shifted towards the average of common structural locations. These structural locations correspond to features that exist across all curves and can be determined from estimated curves and derivatives. Further statistical analysis, such as averaging across the sample, is done subsequently, after the curves have been aligned. Wang and Gasser (1998) derived confidence intervals for the "structural average" and a bootstrap estimation procedure. Silverman (1995) proposed a simple "time-shift" model, see Leng and Müller (2006) for a recent application. Ramsay and Li (1998) provided a flexible family of time transformations, leading to smooth monotone curve registration functions based on splines, aligning towards an overall mean, referred to as Procrustes method. Self-modeling warping functions have been developed in Gervini and Gasser (2004) and several basic ideas for warping and curve alignment are due to Wang and Gasser (1997, 1999).

In a study on warping methods by Liu and Müller (2004), a general method for estimating a population mean function when curve data contain random time transformations, called functional convex averaging, was introduced and an implementation based on area-under-the-curve synchronization was proposed. This approach and the need to analyze time-warped biodemographic cohort densities and hazard rates motivate the development of area-under-the-curve density synchronization. Additional motivation is derived from the problem of gene expression normalization, where researchers perform "quantile normalization", a variant of area-under-the-curve synchronization. Specifically, Bolstad et al. (2003) proposed three normalization methods for microarray data analysis and compared their performance with conventional methods. In these comparisons, quantile normalization emerged as the best and fastest algorithm for erasing unwanted variation in expression level densities between arrays. Our study places this method into the formal context of density synchronization, which to the best of our knowledge has not yet been addressed formally. We will explore the application to gene expression further in Section 6.2. Our approach borrows ideas from functional convex averaging and area-under-the-curve synchronization and extends these to samples of densities by establishing a connection to quantile functions and quantile-density functions (Parzen, 1979).

The article is organized as follows. The functional density synchronization model and quantile synchronization are introduced in Section 2. Estimation of model components, derivation of synchronized

hazard rates, and preliminary facts regarding kernel smoothing are the theme of Section 3. In Section 4, we focus on asymptotic properties of the quantile-synchronized density estimate, and provide a key theorem. Simulation results are discussed in Section 5, demonstrating that the quantile synchronization approach overcomes the drawbacks of conventional averaging. We then demonstrate the application of our methods to gene expression and mexfly data (Section 6). Concluding remarks are given in Section 7, and auxiliary results and brief proofs in an Appendix.

2 Quantile synchronization

Our starting point is a sample of density functions $f_k,\ k = 1, \ldots, n$, which we interpret as realizations of a stochastic process $\mathcal{F}$ that produces trajectories which are densities. For a given large $T > 0$, let $f_k(x),\ x \in [0, T]$, be the random density corresponding to the k-th subject, where $f_k \geq 0$ and $\int_0^T f_k(x)dx = 1$. We assume the support of f_k to be an interval $[0, T]$, and f to be an arbitrary random density generated by the process $\mathcal{F}$. In reality, the densities f_k are generally not directly observed, but rather we observe data that are generated from each density f_k. Estimation of f_k then requires a preliminary estimation step via smoothing methods, such as kernel smoothing (Silverman, 1986). We remark here that for technical reasons, our results require that the densities have bounded support. In practice and also in simulations, we work with densities with unbounded support as well. In such cases, the targets are densities which are truncated to a suitably large finite interval.

In analogy to Liu and Müller (2004), we consider *warped density functions*

$$\{f(x),\ x \in [0, T], f \text{ is a density function},\ f \in \mathcal{F}\},$$

which are realizations of the density-valued process $\mathcal{F}$. Our proposed approach to align such random densities is area-under-the-curve warping, i.e., to postulate that the underlying latent warping mapping corresponds to the inverse of the cumulative distribution function (cdf) F, i.e., the random quantile function associated with each f, given by

$$F^{-1}(t) = \inf\{x : F(x) \geq t\},\ t \in [0, 1]. \tag{1}$$

The quantile function then maps a postulated synchronized time $t \in [0, 1]$ to warped time $x \in [0, T]$. As an example, we plot the cdfs for simulated data in Figure 2, where the time-variation across individual densities is contained in these cdfs. This time-variation is equivalently reflected in the subject/cohort quantile functions (Figure 3). The use of quantile functions in statistics has been investigated by many authors, including Parzen (1979), Sheather and Marron (1990), and Jones (1992), among others.

A key step in the modeling is to assume that each random quantile function F^{-1} can be expressed as sum of a fixed smooth quantile function F_0^{-1}, representing common features shared by all subjects/cohorts, and a random smooth deviation function δ, representing specific features of an individual subject/cohort. Define stochastic quantile processes

$$F^{-1}(t) = F_0^{-1}(t) + \delta(t), \quad t \in [0,1],$$

where δ has mean zero and bounded variation, $F_0^{-1}(\cdot) + \delta(\cdot) \in [0,T]$, $F_0^{-1}(\cdot) + \delta(\cdot)$ is invertible, and δ is such that $\delta(0) = \delta(1) = 0$, so that $F_0^{-1}(0) + \delta(0) = 0$, $F_0^{-1}(1) + \delta(1) = T$. We refer to $t \in [0,1]$ as the *synchronized time*. Note that the argument does not necessarily correspond to a physical time, depending on the application.

Density processes $\{f(x), x \in [0,T]\}$ are then generated from the random quantile functions $\{(F^{-1}(t), t \in [0,1]\}$ through a density warping map $\varphi : F^{-1} \mapsto f$, where synchronized time t is mapped by φ to warped time x by $x = F^{-1}(t)$, thus providing an area-under-the-curve time transformation mapping for each random density f. The inverse mapping $\varphi^{-1} : f \mapsto F^{-1}$ produces the corresponding quantile function and maps warped time x to synchronized time t. The invertibility of the mapping φ is guaranteed by the continuity of the cdf F (see Appendix for details).

The primary target of our analysis is the overall quantile function,

$$E(F^{-1}(t)) = E(F_0^{-1}(t) + \delta(t)) = F_0^{-1}(t), \quad t \in [0,1], \tag{2}$$

in synchronized time t. The *quantile-synchronized overall density* is then

$$f_\oplus(x) = \varphi\{F_0^{-1}\}(x) = \frac{d}{dx} F_0(x), \quad x \in [0,T]. \tag{3}$$

The fact that $f_\oplus$ is a density follows immediately from the above definition, as F_0 is a cdf, the *quantile-synchronized distribution function.*

For a sample of densities f_k, $k = 1, \ldots, n$, the synchronized overall density $f_\oplus$ is targeted by

$$\bar{f}_\oplus(x) = \varphi\{\frac{1}{n}\sum_{k=1}^{n} F_k^{-1}\}, \quad x \in [0,T], \tag{4}$$

applying the warping mapping φ to the average quantile function. Note that the quantile-synchronized density estimate $\bar{f}_\oplus$ differs substantially from the usual cross-sectional average density $\bar{f} = \frac{1}{n}\sum_{k=1}^{n} f_k$, which does not incorporate time-warping and is linear in the sample densities f_k, while $\bar{f}_\oplus$ is nonlinear. Since for each fixed n, $\frac{1}{n}\sum_{k=1}^{n} F_k^{-1}$ is a quantile function, $\bar{f}_\oplus$ is always a density function.

Comparing this approach with the model in Liu and Müller (2004), we note that there, functional data are assumed to be generated by a bivariate stochastic process $(X(t), Y(t))$, $t \in [0,T]$, such that

the observed process is $\tilde{Y}(x) = Y(X^{-1}(x))$, $x \in [0,T]$. Transposing this concept to the present situation, this bivariate process model would be applied to the random distribution function process with realizations F, where $X(t) = F^{-1}(t)$, $Y(t) = t$ and then $\tilde{Y}(x) = Y(X^{-1}(x)) = F(x)$. A key feature of the present situation is that $Y(t) = t$ is known, which resolves the basic non-identifiability issue that was discussed in Liu and Müller (2004) and also otherwise simplifies the approach. As before, writing $\mu_X(t) = EX(t) = F_0^{-1}(t)$, $\mu_Y(t) = EY(t) = t$, the "functional convex mean curve" emerges here as $\mu_Y(\mu_X^{-1}(x)) = F_0(x)$, which, upon differentiation, leads to the quantile-synchronized density $f_\oplus$.

3 Estimating the components for quantile synchronization

A multitude of smoothing methods have been devised for nonparametric density estimation; we use here kernel density and kernel quantile smoothing. Assume that for the k-th subject/cohort, $k = 1, \ldots, n$, a sample of m_k scalar data $X_{k1}, \ldots, X_{km_k} \sim F_k$ are observed, and that these data are conditionally independent, conditioning on the random distribution function F_k, which randomly varies across experimental units, subjects or cohorts, and is a realization of a distribution function process. To ensure that the numbers of measurement asymptotically increases in the same way across the subjects/cohorts, we assume that there exists a sequence $m = m(n)$ with $m \to \infty$, as $n \to \infty$, such that

$$\frac{m_k}{m} \to \tau_k, \tag{5}$$

for positive constants τ_k, such that $0 < c_0 \leq \inf_{1 \leq k \leq n} \tau_k \leq \sup_{1 \leq k \leq n} \tau_k < C_0 < \infty$, $k = 1, \ldots, n$, as $n \to \infty$.

For the distribution function F_k of the k-th cohort, $k = 1, \ldots, n$, the quantile function $F_k^{-1}(t)$ is defined by $F_k^{-1}(t) = \inf\{x : \ F_k(x) \geq t\}$, $t \in (0,1)$. Given the sample $X_{k1}, \ldots, X_{km_k}$, a natural estimator of the quantile function is the empirical quantile function

$$\hat{F}_k^{-1}(t) = \inf\{x : \ \hat{F}_k(x) \geq t\}, \ t \in (0,1), \tag{6}$$

where $\hat{F}_k(x) = m_k^{-1} \sum_{i=1}^{m_k} I(X_{ki} \leq x)$ is the empirical distribution function. Cheng and Parzen (1997) proposed a unified kernel quantile estimator, given by

$$\tilde{F}_k^{-1}(t) = \int_0^1 \hat{F}_k^{-1}(v) \alpha_b(t - v)\, dv, \tag{7}$$

where α is a kernel function, chosen as a probability density function that is symmetric around 0, and $\alpha_b(\cdot) = \alpha(\cdot/b)/b$.

Whether one proceeds to smooth the empirical quantile functions first, followed by averaging these smoothed quantile functions, as in (8) below, or to average empirical quantile functions first, followed by smoothing, makes no difference, as the smoothing step is linear. The motivation for including

a smoothing step is twofold: First, smoothing the quantile function eliminates the awkwardness in defining arbitrary quantiles from the empirical quantile function, as the smoothed quantile function typically will be strictly monotone on most of the range. Second, smoothing improves second order efficiency by alleviating the so-called relative deficiency, as has been shown in Falk (1984).

The quantile-synchronized overall density estimate is obtained as follows. First, the synchronization mapping functions $\tilde{F}_k^{-1},\ k = 1, ..., n$, for each subject/cohort are estimated by the smoothing step (7), using a Gaussian kernel α and adaptive global plug-in bandwidth selectors. More specifically, our implementation follows the description in Gasser, Kneip and Köhler (1991) and Brockmann, Gasser and Hermann (1993), and we adopt their publicly available R package `lokern` with global bandwidth choice. This leads to the estimate

$$\hat{F}_0^{-1}(t) = \frac{1}{n}\sum_{k=1}^{n} \tilde{F}_k^{-1}(t) \tag{8}$$

for F_0^{-1}. Variants of this approach are possible, some of which may be advantageous in certain situations. For example, one can easily substitute the Gaussian kernel with other kernel functions such as the Epanechnikov kernel. Furthermore, other smoothing methods might be used, for example splines or local linear fitting when the quantile function is available on a discrete grid of points.

By numerical inversion, one obtains the distribution function estimate $\hat{F}_0$ corresponding to $\hat{F}_0^{-1}$, and finally an estimate of $\bar{f}_\oplus$ (4), targeting $f_\oplus$ (3), by convolving $\hat{F}_0$ with a derivative kernel,

$$\hat{f}_\oplus(x) = \int_0^T \hat{F}_0(v)\beta_h^{(1)}(x - v)\, dv. \tag{9}$$

Here $h = h(n)$ is a bandwidth, with $h \to 0$ as $n \to \infty$, $nh^2 \to \infty$, and $\beta_h^{(1)}(\cdot) = \frac{1}{h^2}\beta^{(1)}(\frac{\cdot}{h})$ where $\beta^{(1)}$ is the derivative of a kernel function β with properties outlined in (H3) in the Appendix. The required numerical differentiation of the estimated quantile-synchronized distribution $\hat{F}_0$ in (9) is thus implemented by convolving with a derivative kernel.

For kernels β we chose the standard Gaussian density, which worked well in practice. As for kernel α, again alternative kernel choices can be easily substituted, such as the Epanechnikov kernel, as well as alternative approaches for differentiation, e.g. based on smoothing splines. The bandwidth h or equivalent smoothing parameter typically is chosen very small. To control biases near endpoints when applying kernel methods for functions with domains on a bounded interval, one often benefits from applying specially designed boundary kernels (see, e.g., Müller, 1991, 1993) or local polynomial approximations (Zhang and Karunamuni, 1998), but such adjustments do not matter much here, due to the very small bandwidths. Another alternative to implement numerical differentiation would be to use difference quotients, but this approach requires the choice of a step size and has some drawbacks (see e.g. Gasser and Müller, 1984, for a discussion of these issues).

Note that the computation of the quantile-synchronized overall density estimate $\hat{f}_\oplus$ does not actually require to obtain subject/cohort densities $\tilde{f}_k$, as it is based on the cohort quantile functions $\tilde{F}_k^{-1}$. In applications it is nevertheless often helpful to compute and inspect the estimates $\tilde{f}_k$ of the individual cohort densities, as this leads to better understanding of the characteristics of subject/cohort behavior and is important for interpretation of the data and the detection of outliers. For estimating the densities $\tilde{f}_k$, one may apply histograms, smoothed histograms, kernel density estimators or other nonparametric density estimation schemes (see Silverman, 1986).

From the one-to-one relationship between density functions and hazard rates, we obtain the *quantile-synchronized overall hazard rate* from the quantile-synchronized overall density estimate by

$$\hat{\lambda}_\oplus(x) = \frac{\hat{f}_\oplus(x)}{1 - \hat{F}_0(x)}. \tag{10}$$

4 Asymptotic results

We denote weak convergence in the function space $\mathcal{C}$ of continuous functions by $\Rightarrow$. As discussed above, the subject/cohort densities f_k and the corresponding quantile functions F_k^{-1}, $k = 1, \ldots, n$, are typically not directly observed but must be estimated from samples generated by these distributions. Starting with $\hat{F}_0^{-1}$ in (8), which is based on the estimates $\tilde{F}_k^{-1}$ of quantile functions F_k^{-1}, one obtains the warped cdf $\hat{F}_0$ by numerical inversion and then the quantile-synchronized overall density estimate $\hat{f}_\oplus$ (9). In the following, we discuss the asymptotic convergence of $\hat{F}_0$ and $\hat{f}_\oplus$. Proofs and auxiliary results for the following results are given in the Appendix. Recall $F_0^{-1} = E(F^{-1})$, according to (2).

Theorem 1. Under conditions (B2)-(B3) and (H1)-(H3) in the Appendix, as $m \to \infty$ and $n \to \infty$, the warped cdf estimate $\hat{F}_0$ derived from (8) satisfies

$$\sqrt{n}(\hat{F}_0 - F_0) \Rightarrow \Xi_F \text{ on } [0, T],$$

where Ξ_F is a centered Gaussian process defined in (19) in the Appendix.

This result provides the asymptotic distribution of the estimated quantile-synchronized cdf $\hat{F}_0$. This is the key quantity in the proposed quantile synchronization method. All other overall sample characteristics such as the overall density, but also overall mean, median or other quantities are derived from F_0, and corresponding estimates are derived from $\hat{F}_0$.

Theorem 2. Under the assumptions of Theorem 1 and condition (H4) in the Appendix, for bandwidths h as defined in (9), it holds that

$$\sup_{x \in [0,T]} \left| \hat{f}_\oplus(x) - f_\oplus(x) \right| = O_p\left(\frac{1}{\sqrt{n}h} \right) + O(h^2).$$

We have seen that $f_\oplus$ is always a density, according to (3). Theorem 2 implies that also the estimates $\hat{f}_\oplus(x)$ are arbitrarily close to a density for large samples. In finite samples, this is not necessarily the case. One may wish to apply an adjustment to $\hat{f}_\oplus$ by projecting onto the space of density functions, e.g., by truncating the function when it is negative at 0 and normalizing, so that the integral becomes 1 (see, e.g., Gajek, 1986).

Theorem 2 provides a theoretical basis for the application of quantile synchronization in large samples. For example, quantile normalization in microarray analysis can be studied within our framework, and Theorem 2 guarantees the validity of this approach for large samples and identifies $f_\oplus$ as the appropriate target. Choosing the rate $h \sim n^{-1/6}$, one finds that Theorem 2 establishes a convergence rate of $n^{-1/3}$ in the sup norm. As long as $m/n \to \infty$, i.e., the average number of observations per cohort (see (5)) grows asymptotically faster than the number of cohorts in the data sample, the convergence is determined by the number of available cohorts n that enter into the synchronization step.

For the synchronized overall hazard rate estimator $\hat{\lambda}_\oplus$ (10), the target function is the synchronized overall hazard rate function $\lambda_\oplus(x) = \frac{f_\oplus(x)}{1-F_0(x)}$. Theorems 1 and 2 immediately imply the consistency result

$$\sup_{x\in[0,T]} \left|\hat{\lambda}_\oplus(x) - \lambda_\oplus(x)\right| = O_p\left(\frac{1}{\sqrt{n}h}\right) + O(h^2).$$

5 Illustrations with simulated data

In order to illustrate the effectiveness of our method, we generated three sets of simulated data. In additional simulations, not reported here, we studied the effect of the selection of bandwidths b in eq. (7) on the resulting quantile-synchronized overall density estimate $\hat{f}_\oplus$, and found the influence to be relatively small, so that this choice appears not very crucial.

Simulation 1. In the first example with simulated data, the starting point was the quantile function Q_0 of the Gaussian density with mean 4 and standard deviation 1. Warping functions were generated as $\delta(t) = A\sin(\pi t)$, where $A \sim$Uniform$(-1, 1)$, leading to warping processes $Q(t)$, where $n = 60$ copies of

$$Q_k(t) = Q_0(t) + \delta_k(t), \quad k = 1, \ldots, 60,$$

were generated. Then $m = 100$ observations were randomly drawn from each distribution $F_k = Q_k^{-1}$. Individual density functions were estimated from each sample using kernel estimators.

In Figure 4, the corresponding sample of density estimates $\tilde{f}_k$ is displayed. Also shown are (a) the proposed quantile-synchronized overall density estimate $\hat{f}_\oplus$, as defined in (9), and implemented as

described in the previous section; (b) the conventional overall cross-sectional density

$$\hat{f}_{CS}(x) = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_k(x),\ k = 1, \ldots, n,\ x \in [0, T]; \tag{11}$$

and (c) the shift-registered overall cross-sectional density

$$\hat{f}_{SRCS}(x) = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_{SRk}(x),\ k = 1, \ldots, n,\ x \in [0, T], \tag{12}$$

where $\tilde{f}_{SRk}(x)$ are the shift-registered individual density estimates, simply obtained by shifting $\tilde{f}_k$ to the center of the samples $\frac{1}{nm}\sum_{k=1}^{n}\sum_{j=1}^{m} x_{kj}$.

Here the cross-sectional overall density $\hat{f}_{CS}$ serves as a simple-minded baseline, and would be used conventionally as representing the sample of densities. The shift-registered overall cross-sectional density $\hat{f}_{SRCS}$ provides a second, somewhat more sophisticated baseline, combining a simple shift-warping approach with cross-sectional averaging. This provides a very simple to implement improvement over the cross-sectional estimate $\hat{f}_{CS}$. The results clearly demonstrate that neither the cross-sectional density estimate $\hat{f}_{CS}$ nor the shift-registered cross-sectional density estimate $\hat{f}_{SRCS}$ comes very close to the underlying Gaussian density, however $\hat{f}_{SRCS}$ performs considerably better than $\hat{f}_{CS}$. The quantile-synchronized overall density estimate $\hat{f}_{\oplus}$ provides the best approximation to the true underlying density and outperforms the other methods.

Simulation 2. A second simulation was carried out in a similar way, but for an underlying Gamma(κ, θ) density, with shape parameter $\kappa = 5$ and scale parameter $\theta = 1$. Warping functions were generated with similar shapes but more variation compared to the first simulation, by $\delta(t) = A\sin(\pi t)$, where $A \sim$Uniform$(-2, 2)$. From the resulting warping processes $Q(t)$, $n = 60$ copies were generated as

$$Q_k(t) = Q_0(t) + \delta_k(t), \quad k = 1, \ldots, 60.$$

Then $m = 500$ observations were randomly drawn from each distribution $F_k = Q_k^{-1}$.

The comparison of the performance of the proposed quantile-synchronized overall density estimates $\hat{f}_{\oplus}$, with the conventional cross-sectional estimate $\hat{f}_{CS}$ and the improved shift-warped cross-sectional estimate $\hat{f}_{SRCS}$ is illustrated in Figure 5. The quantile-synchronized overall density is seen to provide the best summary estimate, and adequately reflects the true underlying shape. In contrast, the cross-sectional estimates are both less representative of the underlying Gamma distribution.

Simulation 3. A more complex structure was studied in a third example, where we consider the situation of a bimodal underlying density. The quantile function Q_0 in this case was that of a Gaussian mixture of two normal densities ϕ_1, ϕ_2 with means 3 and 4 and standard deviations 0.3 and 0.4 respectively, with $f_{\oplus} = 0.5\phi_1(y) + 0.5\phi_2(y)$. Warping functions were generated as $\delta(t) = A\sin(2\pi t)$, where

$A \sim$Uniform$(-0.2, 0.2)$, leading to warping processes $Q(t)$, where $n = 60$ copies of

$$Q_k(t) = Q_0(t) + \delta_k(t), \quad k = 1, \ldots, 60,$$

were generated. Then $m = 500$ observations were randomly drawn from each distribution $F_k = Q_k^{-1}$.

The generated densities and the results obtained from quantile synchronization $\hat{f}_{\oplus}$, the conventional cross-sectional estimate $\hat{f}_{CS}$ and the shift-warped cross-sectional estimate $\hat{f}_{SRCS}$ are in Figure 6. Viewing the sample densities, the warping structure is seen to be quite complex in this example. The representative density should be close to the underlying Gaussian mixture. This is the case for the density estimate obtained by quantile synchronization, while the conventional and the shift-registered cross-sectional averaging provide similar and almost indistinguishable estimates that are far away from the target, clearly demonstrating that quantile synchronization provides a more useful representation in this case.

The corresponding comparisons for the hazard rates are shown in Figure 7, where we compare the synchronized hazard rate estimate $\hat{\lambda}_{\oplus}$ with the alternative estimates from the cross-sectional averaging methods, replacing $\hat{f}_{\oplus}$ in (10) by $\hat{f}_{CS}$ or $\hat{f}_{SRCS}$ and $\hat{F}_0$ by $\hat{F}_{CS}$ or $\hat{F}_{SRCS}$, obtained by averaging the cohort-specific empirical distribution functions or shift-registered cohort-specific empirical distribution, respectively. The quantile-synchronized estimate correctly reflects the wave pattern of the underlying hazard rate in the middle of the domain, while both conventional and shift-warped cross-sectional estimates are clearly inferior.

6 Data Applications

6.1 Mexican fruit fly study

Due to genetic differences, chance and cohort effects, different individuals or cohorts of the same species often have different developmental schedules (Finch, 1994). When studying age-at-death in a biodemographic experiment that involves samples of life tables derived from individual cohorts, it is of interest to determine the common density and hazard functions associated with the observed lifetimes (age-at-death) across all cohorts. The proposed synchronization method provides a natural approach for this purpose, removing distortions of survival in individual cohorts that may arise through various cohort effects. Flies can easily be raised in large cohorts and their entire lifespan is observable, so there are no censoring issues. A biodemographic experiment was conducted with Mexican fruit flies (mexflies) with the goal to study mortality across a fairly large number of cohorts (see Carey et al., 2005, for background).

The mexfly data consist of life tables for 54 cages of male flies observed over a 174 day-period, where each cage contained more than 1000 flies. Observations of age-at-death of flies raised within the same cage must be assumed to be correlated, as these flies share the same living environment, food source, and temperature, which are thought to slightly vary between cages, due to small random variations. An additional unavoidable source of correlations are interactions among individuals raised jointly within the same cage (cohort). Hence entire cohorts constitute the natural independent units for data analysis. The available data consist of a life table per cohort, i.e., the daily number of deaths, which was recorded separately for each cage. Time-variation is indeed present when viewing the cohort hazard rates (Figure 1), motivating the need to find an overall representative hazard.

We describe now some specific issues when implementing the proposed quantile synchronization method for biodemographic lifetime data. Histograms provide a natural first estimate for the cohort-specific densities of age-at-death. A histogram density estimate is defined as follows: Given a random sample $x_1, \ldots, x_m \in (a, b)$ that is sampled from a density function f, define a partition of (a, b) by $(t_0, \ldots, t_P)$, where $a = t_0 < t_1 < \ldots < t_P = b$. The bin width Δ_j of bin $B_j = [t_{j-1}, t_j),\ j = 1, \ldots, P$, is $\Delta_j = t_j - t_{j-1}$. Denote the count of sample points falling within bin B_j by $y_j,\ j = 1, \ldots, P$, so that $\sum_{j=1}^{P} y_j = m$. The histogram density estimate $\hat{f}_H$ of f is given by

$$\hat{f}_H(x) = \frac{y_j}{m\Delta_j},\ x \in B_j,\ j = 1, \ldots, P. \tag{13}$$

As alternative, one might consider histosplines (Wahba, 1976).

The mexfly data is naturally binned by days, as the counts of deaths for all cohorts are recorded at days $1, \ldots, M$, where M is the maximum age (in days) observed in the study. Thus, in this application, $\Delta_j = \Delta = 1$ day, and the bins B_j are one-day intervals. We denote the observations for the k-th cohort by $(x_j,\ y_{kj})$, where $k = 1, \ldots, n,\ x_j = j - 1/2,\ j = 1, \ldots, M$, denote the midpoints of the j-th day, n is the total number of cohorts (here $n = 54$), and y_{kj} corresponds to the number of deaths observed on day j for cohort k. The histogram density estimate $\hat{f}_{Hk}$ defined in (13) for the underlying density f_k of age-at-death for cohort k is

$$\hat{f}_{Hk}(x_j) = \frac{y_{kj}}{\Delta \sum_{j=1}^{M} y_{kj}},\ k = 1, \ldots, n,\ j = 1, \ldots, M. \tag{14}$$

The empirical distribution function $\hat{F}_{Hk}$ can then be defined as

$$\hat{F}_{Hk}(x_j) = \Delta \left\{ \sum_{v<x_j} \hat{f}_{Hk}(v) + \frac{1}{2}\hat{f}_{Hk}(x_j) \right\},\ k = 1, \ldots, n,\ j = 1, \ldots, M. \tag{15}$$

We will therefore replace $\hat{F}_k^{-1}$ in (6) by the empirical quantile plot $(\hat{F}_{Hk}(x_j), x_j)$, which then is entered into the kernel quantile estimator (7).

The resulting individual warping function estimates $\tilde{F}_k^{-1}(t)$ (7), $k = 1, 2, ..., 54$, and the overall time-synchronizing function $\hat{F}_0^{-1}(t)$ (8) are displayed in Figure 8. Given the synchronized overall density estimate $\hat{f}_\oplus$ (9), the synchronized hazard rate $\hat{\lambda}_\oplus$ is obtained by equation (10). In Figure 9, we compare the synchronized hazard rate estimate $\hat{\lambda}_\oplus$ with the alternative estimates that are obtained from the cross-sectional averaging methods. The synchronized hazard function estimate $\hat{\lambda}_\oplus$ shows a"plateau", i.e., a flattening of the hazard rate towards older ages. A few cohort trajectories but not all of them show this phenomenon as well (see Figure 1). Such plateaus are commonly found in biodemographic studies (Carey et al., 1992; Wachter, 1999). Their presence implies that the flies' vulnerability does not always increase as they age and plateaus may be interpreted as a slowing of the aging process for very old individuals. The hazard function estimates derived from the cross-sectional methods do not show a plateau. Since the variability of hazard rate estimates is quite high in the right tail, these findings must be considered suggestive but not conclusive.

6.2 Gene expression analysis of Ts1Cje mouse for Down syndrome

Down syndrome (DS) is the most frequent chromosomal cause of mental retardation and is caused by trisomy 21 (HSA21). DS is present in approximately one out of every 700 live-born infants. The distal end of mouse chromosome 16 (MMU16) is orthologous to most of HSA21. The Ts1Cje mouse model of DS with segmental trisomy of MMU16 is one of the many mouse models developed for DS and the learning and behavioral abnormalities displayed by the Ts1Cje mouse are similar to the mental retardation associated with DS.

Amano et al. (2004) investigated global gene expression profiles in whole brains of six Ts1Cje mice and six normal littermate (2N) mice at postnatal day 0, using DNA microarrays. Probe-level data reveal that the sets of expressed genes are almost the same between the Ts1Cje mouse and a normal mouse. However, most genes in the trisomic region are over-expressed in Ts1Cje mice while the expression levels for other chromosomes are almost the same between Ts1Cje and normal mice. Amano et al. concluded that the over-expression of genes in the trisomic region has a bearing on the pathogenesis of DS. The raw, unfiltered, probe-level data for both Ts1Cje and normal mice are available from Gene Expression Omnibus (GEO) at NCBI.

The GeneChip hybridization used in Amano et al. (2004) was based on Affymetrix Murine Genome U74A and U74B GeneChips (24935 probe sets) (Affymetrix, Inc., Santa Clara, CA, USA). There are two types of probes for Affymetrix GeneChips. *Perfect match* (PM) represents the reference probes that match a target sequence. Each gene is typically represented by a set of 11-20 pairs of probes and the expression intensity for a gene is obtained through a summary of the measurements from the probe set.

In the commonly encountered case where multiple arrays are involved, the construction of an overall

summary density is not trivial and the simple cross-sectional average density is not effective, due to random array effects which effectively correspond to expression level warping. This problem is recognized in genomics and leads to the need for array normalization as a preprocessing step. We use the density of log(PM) to illustrate how quantile-synchronization can be used to advantage. Here the experimental units are the arrays, for each of which one obtains expression data that give rise to a density function corresponding to the gene expression level distribution, as measured by the specific array.

Array level-variation effects are evident from the densities of expressions for normal 2N mice using $n = 6$ arrays (Figure 10) and Ts1Cje mice, also using $n = 6$ arrays (Figure 11), based on U74B GeneChips. Adjustment for the presence of expression level warping is clearly necessary in order to obtain a useful overall summary density. Applying our method to the probe-level data and comparing $\hat{f}_{\oplus}$ with the cross-sectional densities $\hat{f}_{CS}$ and $\hat{f}_{SRCS}$ shows that the undesirable level variation between arrays is removed when applying the proposed estimate $\hat{f}_{\oplus}$ (Figures 10 and 11). In contrast, cross-sectional estimates $\hat{f}_{CS}$ and $\hat{f}_{SRCS}$ are distorted, owing to the array level variation and are inferior to capture the structure of the essential features of the expression distribution. Further microarray examples where the very similar method of quantile normalization has been applied can be found in Bolstad et al. (2003).

7 Discussion and concluding remarks

Viewing cohort densities as functional data, we propose a simple yet efficient time-synchronization estimation method to obtain a reasonable underlying overall density that represents the essential features of the sample. In addition to the quantile-synchronized overall density, the procedure also provides synchronized hazard rate estimation. The proposed approach is motivated by a general model for warping in functional data analysis and uses quantile synchronization, aligning locations in the domains of the densities that correspond to the same quantile. The method proves useful for applications and has good asymptotic properties.

Quantile synchronization is particularly useful when significant warping effects exist between individual densities, as is often the case in biomedical studies. Examples based on both simulated and real data illustrate that conventional methods may fail to provide a useful overall density and hazard rate estimate in such situations. Boundary effects can be a problem that may affect estimation in the tails of the density and in some instances (especially when choosing larger bandwidths) need to be carefully addressed. Our method is related to "quantile normalization", a well established pre-processing method for gene expression arrays. The theoretical results we provide to justify quantile synchronization also extend to this quantile normalization method.

One of the anonymous reviewers of this article provided an example for discussion, where (in a slightly modified version) f_0 is a standard Gaussian density and the observed sample of densities is drawn from

$$f_i(x) = f_0((x - \mu_i)/\sigma_i), \quad \mu_i \sim U([-1/2, 1/2]),\ \sigma_i = 1 - \mu_i^2,\ i = 1, \ldots, n.$$

In this case, the quantile-synchronized overall density that represents the sample does not belong to the manifold in which these densities lie. This shows that the representing density is not necessarily close to the densities in the sample. The density in the above example which represents the average location and also belongs to the sample is the standard Gaussian; however, this density is not representative as it does not correctly represent the variances of these densities, which are much smaller than 1 on average. The proposed quantile-synchronized representative is a density that is symmetric around 0 but with variance smaller than 1, and this makes sense for a representative density. It can be characterized as an extrinsic rather than intrinsic mean of the manifold, and is the Fréchet mean in a suitable metric. Such a metric can be derived from the metric given in Liu and Müller (2004), making use of the discussion after eq. (4) in Section 2.

There are several open questions. We only discussed density and hazard rate estimation for the case of no censoring. An extension to the case of censored data will be useful for various biomedical and reliability applications where some event times are censored. While bootstrap can be applied, a deeper investigation of inference procedures such as confidence intervals for estimates based on quantile synchronization in relation to the target densities or corresponding tests remain an open problem. While we provided a basic framework for a particularly natural version of quantile synchronization, the development of other approaches of warping or curve registration for samples of densities will be of interest.

APPENDIX: PROOFS AND AUXILIARY RESULTS

For simplicity of notation, set $Q_k = F_k^{-1}$ for the quantile functions $F_k^{-1}(t),\ t \in [0,1],\ k = 1, \ldots, n$, defined in (1). Let Q be a generic process distributed as Q_k, and let S denote the support of f_k. If Q_k is differentiable, the derivative $q_k(t) = Q_k'(t),\ t \in (0,1)$, is the quantile density function (compare Parzen, 1979, and Cheng and Parzen, 1997).

We require the following assumptions. The first assumption is needed to ensure that asymptotic errors are dominated by the error that stems from the finite number of n of available densities, rendering the error associated with the estimation of the individual densities negligible in comparison.

(B1) The number n of sampled densities satisfies $n \to \infty$, and the number of observations m available per density satisfies $\frac{m}{n} \to \infty$.

(B2) There exists a constant $0 < L < \infty$ such that

$$E|Q(t) - EQ(t) - (Q(s) - EQ(s))|^2 \leq L|t-s|^2,\ t,s \in [0,1],\ \text{and}\ \sup_t \text{var}(Q(t)) < L.$$

(B3) The function $Q_0 = EQ = F_0^{-1}$ (see (2)) is continuously differentiable in t and

$$\inf_{t\in[0,1]} |q_0(t)| > 0,\ \text{for}\ q_0 = Q_0'. \tag{16}$$

We remark that condition (B2) cannot be satisfied for quantile functions on unbounded intervals, and therefore the domain of the f_k needs to be finite. Finite support is a standard assumption in quantile estimation problems.

Let $G_Q(t),\ t \in [0,1]$ be a Gaussian process with mean function $EG_Q(t) = 0$ and covariances

$$\text{cov}(G_Q(s), G_Q(t)) = \text{cov}(Q(s), Q(t)),\ s,t \in [0,1],$$

and define

$$\bar{Q}_0 = \frac{1}{n}\sum_{k=1}^{n} Q_k, \qquad \bar{F}_0 = \bar{Q}_0^{-1}. \tag{17}$$

Lemma 1. Under assumptions (B2) and (B3),

$$\sqrt{n}(\bar{F}_0 - F_0) \Rightarrow \Xi_F,\ \text{on}\ [0,T], \tag{18}$$

where Ξ_F is the Gaussian process given by

$$\Xi_F(x) = \frac{1}{q_0 \circ F_0(x)}\ G_Q \circ F_0(x). \tag{19}$$

Here q_0 and F_0 are defined in (16) and (3), and "$\circ$" denotes the composition of two functions.

Proof of Lemma 1. Under conditions (B2) and (B3), (18) follows from Theorem 3 in Liu and Müller (2004).

Lemma 1 provides the asymptotic distribution of the synchronized cdf $\bar{F}_0$, based on the assumption that the cohort quantile functions are known. However, in applied settings, these functions are not observed directly, but rather need to be estimated from available data. Therefore, we require an extension that applies to estimated quantile functions $\tilde{Q}_k = \tilde{F}_k^{-1},\ k = 1,\ldots,n$ (7).

For this purpose, we invoke the following additional assumptions.

(H1) The quantile density functions $q_k,\ k = 1,\ldots,n$, are twice continuously differentiable in (0,1), and for a constant $c_0 > 0$ satisfy $\inf_{t\in[0,1]} q_k(t) \geq c_0 > 0$, for all k. There is a $\gamma > 0$ such that $\sup_{u\in[0,1]} u(1-u)|J_k(u)| \leq \gamma$ for all k, where $J_k(u) = [d \log q_k(u)/du]$ is the score function.

(H2) There exists $0 < L_0 < \infty$, such that $\sup_{u\in[0,1]} \left|\int_0^1 q_k(t)\alpha_b(u-t)dt\right| \leq L_0$ for all k.

(H3) (1) The kernel functions α and β are probability density functions which are symmetric around 0, and β is differentiable.

(2) For any function g that is at least twice continuously differentiable in $(0,1)$, for a $\rho > 1/2$ it holds that $\limsup_{m\to\infty} b^2 m^\rho < \infty$, and

$$\sup_{u\in[a,b]} \left| g(u) - \int_0^1 g(t)\alpha_b\left(u-t\right)dt \right| = O(m^{-\rho}),$$

where $[a,b] \subset (0,1)$ and m is as defined in (5).

(H4) The overall cdf F_0 (3) is three times continuously differentiable on [0,T], with $\sup_{x\in[0,T]} |F_0^{(3)}(x)| < \infty$.

Lemma 2. Under the conditions (H1), (H2), and (H3), we have

$$\sup_k \sup_{t\in[0,1]} \left|\tilde{Q}_k(t) - Q_k(t)\right| = O_p(m^{-1/2}),\ k = 1,\ldots,n. \tag{20}$$

Proof. By Theorem 2.1 (2) in Cheng and Parzen (1997),

$$\sup_{t\in[0,1]} \left|\tilde{Q}_k(t) - Q_k(t)\right| = O_p(m^{-1/2} + m^{-\rho}) = O_p(m^{-1/2}), \tag{21}$$

for each k, as $\rho > 1/2$. Note that (H1) implies conditions (Q1)-(Q3) in Cheng and Parzen (1997), and conditions (K1)-(K3) in Cheng and Parzen (1997) are satisfied if the kernel function α satisfies (H3). Furthermore, the extension from $[a,b] \subset (0,1)$ to $[0,1]$ is made possible by (H1). Finally, assumption (H2) and the fact that the bound in equation (2.7) in Cheng and Parzen (1997) is a universal bound which does not depend on k allows the extension from (21) to (20).

Let $\hat{Q}_0 = \frac{1}{n}\sum_{k=1}^n \tilde{Q}_k = \hat{F}_0^{-1}$ in (8), and $\hat{F}_0 = \hat{Q}_0^{-1}$. We need the following result,

Lemma 3. Under assumptions (H1)-(H3),

$$\sup_{t\in[0,1]} \sqrt{n}\left|\hat{Q}_0(t) - \bar{Q}_0(t)\right| = O_p((\frac{n}{m})^{1/2}).$$

Proof. According to Lemma 2,

$$\begin{aligned}
\sup_{t\in[0,1]} \sqrt{n}\left|\hat{Q}_0(t) - \bar{Q}_0(t)\right| &= \sup_{t\in[0,1]} \sqrt{n}\left|\frac{1}{n}\sum_{k=1}^n \tilde{Q}_k(t) - \frac{1}{n}\sum_{k=1}^n Q_k(t)\right| \\
&\leq \sqrt{n}\frac{1}{n}\sum_{k=1}^n \sup_{1\leq k\leq n}\sup_{t\in[0,1]}\left|\tilde{Q}_k(t) - Q_k(t)\right| \\
&= O_p((\frac{n}{m})^{1/2}).
\end{aligned}$$

The following result is analogous to Lemma A.1 in Liu and Müller (2004).

Lemma 4. Under conditions (B2), (B3),

$$\sup_{t\in[0,1]} \sqrt{n} \left|\hat{Q}_0(t) - \bar{Q}_0(t)\right| = O_p((\frac{n}{m})^{1/2}),\ t \in [0,1],$$

implies

$$\sup_{x\in[0,T]} \sqrt{n} \left|\hat{F}_0(x) - \bar{F}_0(x)\right| = O_p((\frac{n}{m})^{1/2}),\ x \in [0,T].$$

Proof of Theorem 1. Note that

$$\begin{aligned} \sqrt{n}(\hat{F}_0 - F_0) &= \sqrt{n}(\hat{F}_0 - \bar{F}_0) + \sqrt{n}(\bar{F}_0 - F_0) \\ &= I + II \end{aligned}$$

By Lemma 3 and Lemma 4, $\sup_{t\in[0,1]} |I| = O_p((\frac{n}{m})^{1/2})$ and by Lemma 1 (18), $II \Rightarrow \Xi_F$. In view of (B1), this completes the proof of Theorem 1.

Proof of Theorem 2. The synchronized overall density $\hat{f}_\oplus$ (9) is an estimate of the derivative of the warped cdf $\hat{F}_0$, which in turn is an estimate of the unobserved $\bar{F}_0$ (17). For $\hat{f}_\oplus$, we have

$$\begin{aligned} & \sup_{x\in[0,T]} \left|\hat{f}_\oplus(x) - f_\oplus(x)\right| \\ = & \sup_{x\in[0,T]} \left|\int_0^T \hat{F}_0(v)\beta_h^{(1)}(x-v)\,dv - f_\oplus(x)\right| \\ \le & \sup_{x\in[0,T]} \left|\int_0^T \left[\hat{F}_0(v) - F_0(v)\right]\beta_h^{(1)}(x-v)dv\right| + \sup_{x\in[0,T]} \left|\int_0^T F_0(v)\beta_h^{(1)}(x-v)\,dv - f_\oplus(x)\right| \\ = & III + IV. \end{aligned}$$

By (H3), (H4) and Lemma 1 in Bhattacharya (1967), $IV = O(h^2)$. By substituting $u = (x-v)/h$,

$$III \le \sup_{x\in[0,T]} \int_0^T \sup_{u\in[\frac{x-T}{h},\frac{x}{h}]} \left|\hat{F}_0(x-uh) - F_0(x-uh)\right| \left|\frac{1}{h}\beta^{(1)}(u)\right| du = O_p\left(\frac{1}{\sqrt{n}h}\right),$$

using

$$\sup_{v\in[0,T]} \left|\hat{F}_0(v) - F_0(v)\right| = O_p\left(\frac{1}{\sqrt{n}}\right),$$

which follows from Theorem 1. This completes the proof of Theorem 2.

Acknowledgements

We express our sincere thanks to three anonymous referees and the Associate Editor, whose comments led to major improvements of the paper. Especially the careful reading and detailed comments of one referee proved extremely helpful. We are grateful to Professor James Carey for access to the mexfly data. This research was supported in part by NSF grant DMS08-06199.

References

Amano, K., Sago, H., Uchikawa, C., Suzuki, T., Kotliarova, S. E., Nukina, N., Estein, C. J., Yamakawa, K., 2004. Dosage-dependent over-expression of genes in the trisomic region of Ts1Cje mouse model for Down syndrome. Human Molecular Genetics, 13, 1333-1340.

Bhattacharya, P.K., 1967. Estimation of a probability density function and its derivatives. Sankhyā Series A, 29, 373-382.

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

Brockmann, M., Gasser, T., Herrmann, E., 1993. Locally adaptive bandwidth choice for kernel regression estimators. Journal of the American Statistical Association, 88, 1302-1309.

Carey, J.R., Liedo, P., Müller, H.G., Wang, J.L., Sentürk, D., Harshman, L., 2005. Biodemography of a long-lived tephritid: Reproduction and longevity in a large cohort of female Mexican fruit flies, Anastrepha ludens. Experimental Gerontology, 40, 793-800.

Carey, J.R., Liedo, P., Orozco, D., Vaupel, J.W., 1992. Slowing of mortality rates at older ages in large medfly Cohorts. Science, 258, 457-461.

Cheng, C., Parzen, E., 1997. Unified estimators of smooth quantile and quantile density functions. Journal of Statistical Planning and Inference, 59, 291-307.

Chiou, J., Müller, H.G., 2009. Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. Journal of the American Statistical Association, 104, 572-585.

Cuevas, A., Febrero, M., Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics, 22, 481-496.

Delicado, P., 2007. Functional k-sample problem when data are density functions. Computational Statistics, 22, 391-410.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Computational Statistics and Data Analysis, 55, 401-420.

Falk, M., 1984. Relative deficiency of kernel type estimators of quantiles. Annals of Statistics, 12, 261-268.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer, New York

Ferraty, F., Rahbi, A., Vieu, P., 2008. Estimation non-paramétrique de la fonction de hasard avec variable explicative fonctionnelle. Revue Roumaine de Mathématiques Pures et Appliquées, 53, 1-18.

Finch, C.E., 1994. Longevity, Senescence and the Genome. University of Chicago Press.

Gajek, G., 1986. On improving density estimators which are not bona fide functions. Annals of Statistics, 14, 1612-1618.

Gasser, T., Kneip, A., Köhler, W., 1991. A flexible and fast method for automatic smoothing. Journal of the American Statistical Association, 86, 643-652.

Gasser, T., Müller, H.G., 1984. Estimating regression functions and their derivatives by the kernel method. Scandinavian Journal of Statistics, 11, 171-184.

Gasser, T., Müller, H.G., Köhler, W., Molinari, L., Prader, A., 1984. Nonparametric regression analysis of growth curves. The Annals of Statistics, 12, 210-229.

Gervini, D., Gasser, T., 2004. Self-modelling warping functions. Journal of the Royal Statistical Society, Series B, 66, 959-971.

Gervini, D., Gasser, T., 2005. Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. Biometrika, 92, 801-820.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2001. Maximum likelihood estimation of optimal scaling factors for expression array normalization. SPIE BIOS 2001.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K., Scherf, U., Speed, T.P., 2003. Exploration, normalization and summaries of high density oligonucleotide array probe lever data. Biostatistics, 4, 249-264.

Jones, M.C., 1992. Estimating densities, quantiles, quantile densities and density quantiles. Annals of the Institute of Statistical Mathematics, 44, 721-727.

Kneip, A., Gasser, T., 1992. Statistical tools to analyze data representing a sample of curves. Annals of Statistics, 20, 1266-1305.

Kneip, A., Utikal, K.J., 2001. Inference for density families using functional principal component analysis. Journal of the American Statistical Association, 96, 519-532.

Leng, X. and Müller, H.G., 2006. Time ordering of gene co-expression. Biostatistics, 7, 569-584.

Liu, X.L. and Müller, H.G., 2004. Functional convex averaging and synchronization for time-warped random curves. Journal of the American Statistical Association, 99, 687-699.

Manté, C., Durbec, J.P., Dauvin, J.C., 2005. A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (Western English Channel). Journal of Applied Statistics, 32, 831-840.

Müller, H.G., 1991. Smooth optimum kernel estimators near endpoints. Biometrika, 78, 521-530.

Müller, H.G., 1993. On the boundary kernel method for nonparametric curve estimation near endpoints. Scandinavian Journal of Statistics, 20, 313-328.

Müller, H.G., Wang, J.L., Capra, W.B., 1997. From lifetables to hazard rates: The transformation approach. Biometrika, 84, 881-892.

Nerini, D., Ghattas, B., 2007. Classifying densities using functional regression trees: Applications in oceanology. Computational Statistics and Data Analysis, 51, 4984-4993.

Parzen, E., 1979. Nonparametric statistical data modeling. Journal of the American Statistical Association, 74, 105-121.

Quintela-del-Río, A., 2008. Hazard function given a functional variable: Non-parametric estimation under strong mixing conditions. Journal of Nonparametric Statistics, 20, 413-430.

Ramsay, J.O., Li, X., 1998. Curve registration. Journal of the Royal Statistical Society, Series B, 60, 351-363.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.

Rice, J., Silverman, B.W., 1991. Estimating the mean and covariance Structure nonparametrically when the data are curves. Journal of the Royal Statistical Society, Series B, 53, 233-243.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 26, 43-49.

Sheather, S.J., Marron, J.S., 1990. Kernel quantile estimators. Journal of the American Statistical Association, 85, 410-416.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Silverman, B.W., 1995. Incorporating Parametric Effects Into Functional Principal Components Analysis. Journal of the Royal Statistical Society, Series B, 57, 673-689.

Vaupel, J.W., Carey, J.R., Christensen, K., Johnson T.E., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng,Y., Manton, K.G., Curtsinger, J.W., 1998. Biodemographic trajectories of longevity. Science, 280, 855-860.

Wachter, K.W., 1999. Evolutionary demographic models for mortality plateaus. Proceedings of the National Academy of Sciences, 96, 10544-10547.

Wahba, G., 1976. Histosplines with knots which are order statistics. Journal of the Royal Statistical Society B, 38, 140-151.

Wang, K., Gasser, T., 1997. Alignment of curves by dynamic time warping. Annals of Statistics, 25, 1251-1276.

——— 1998. Asymptotic and bootstrap confidence bounds for the structural average of curves. Annals of Statistics, 26, 972-991.

——— 1999. Synchronizing sample curves nonparametrically. Annals of Statistics, 27, 439-460.

Zhang, S., Karunamuni, R.J., 1998. On kernel density estimation near endpoints. Journal of Statistical Planning and Inference, 70, 301-316.

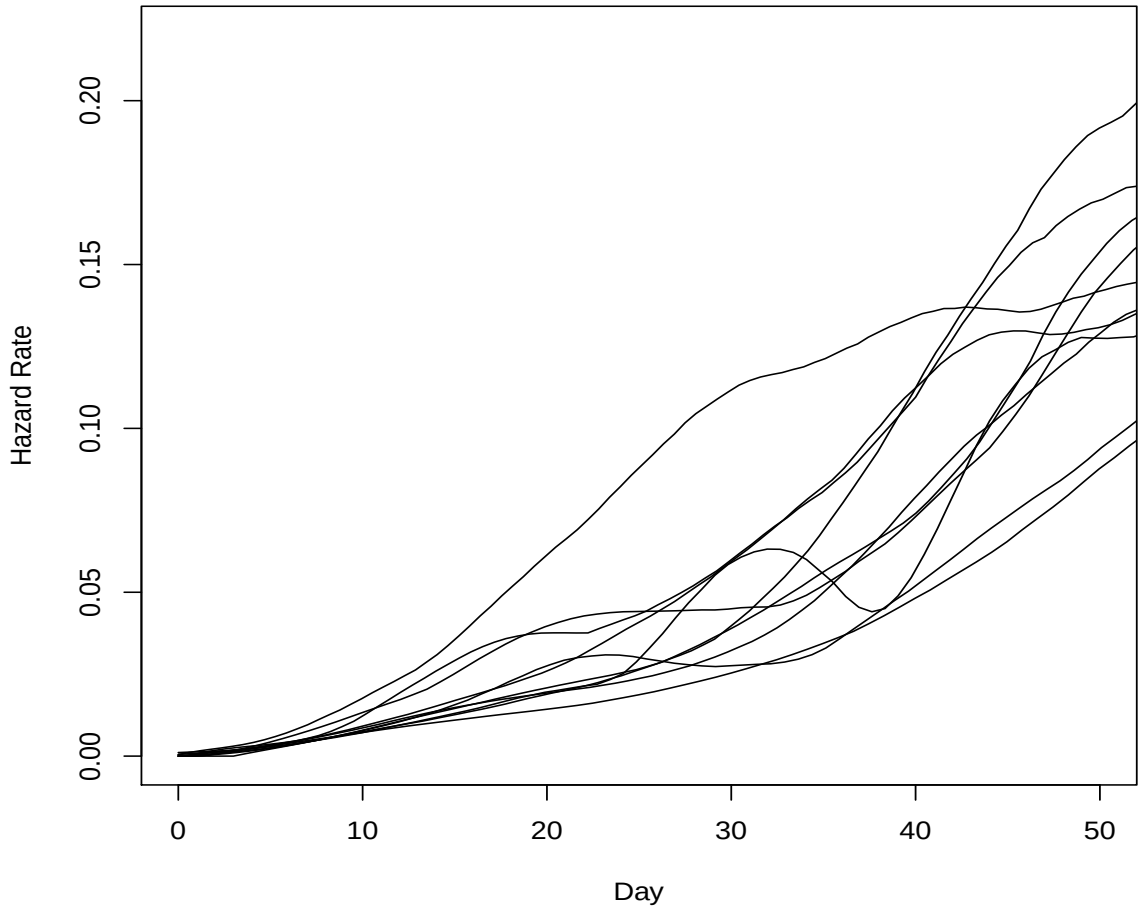

Figure 1: Hazard rate estimates for the age-at-death distribution for the first 10 cohorts of the mexfly data.

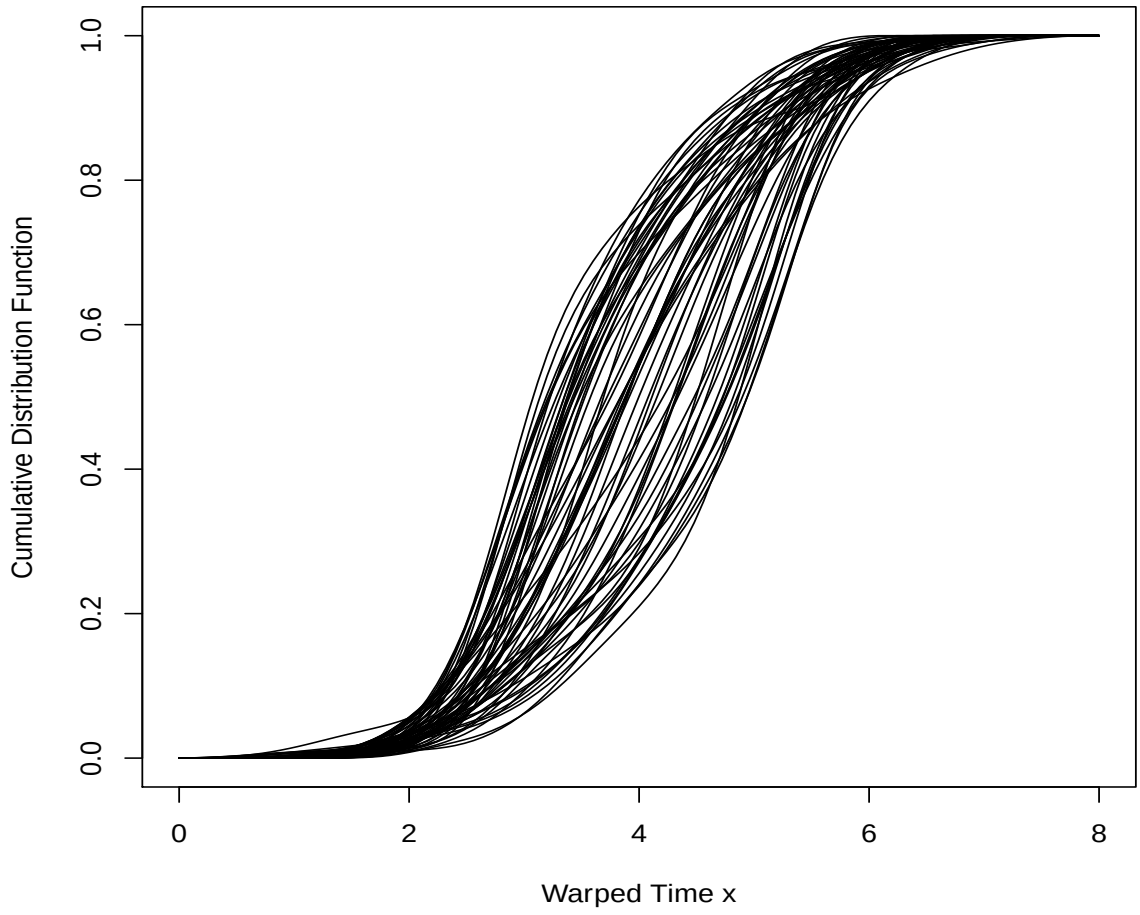

Figure 2: Cumulative distribution function (cdf) estimates $\tilde{F}_k,\ k = 1, \ldots, 60$, for simulated data (from Simulation 1, described in Section 5).

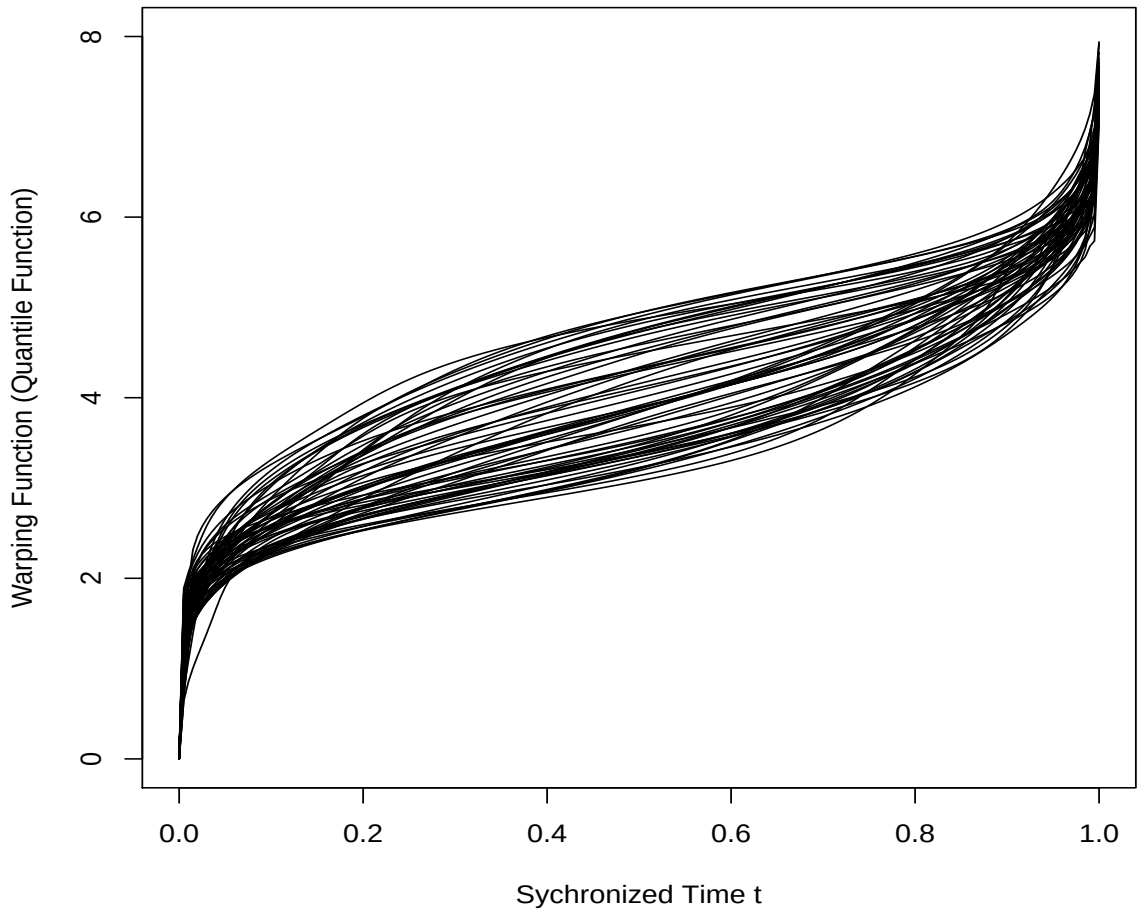

Figure 3: Synchronization function (quantile function) estimates $\tilde{F}_k^{-1}$ (7), $k = 1, \ldots, 60$, corresponding to the simulated data in Figure 2.

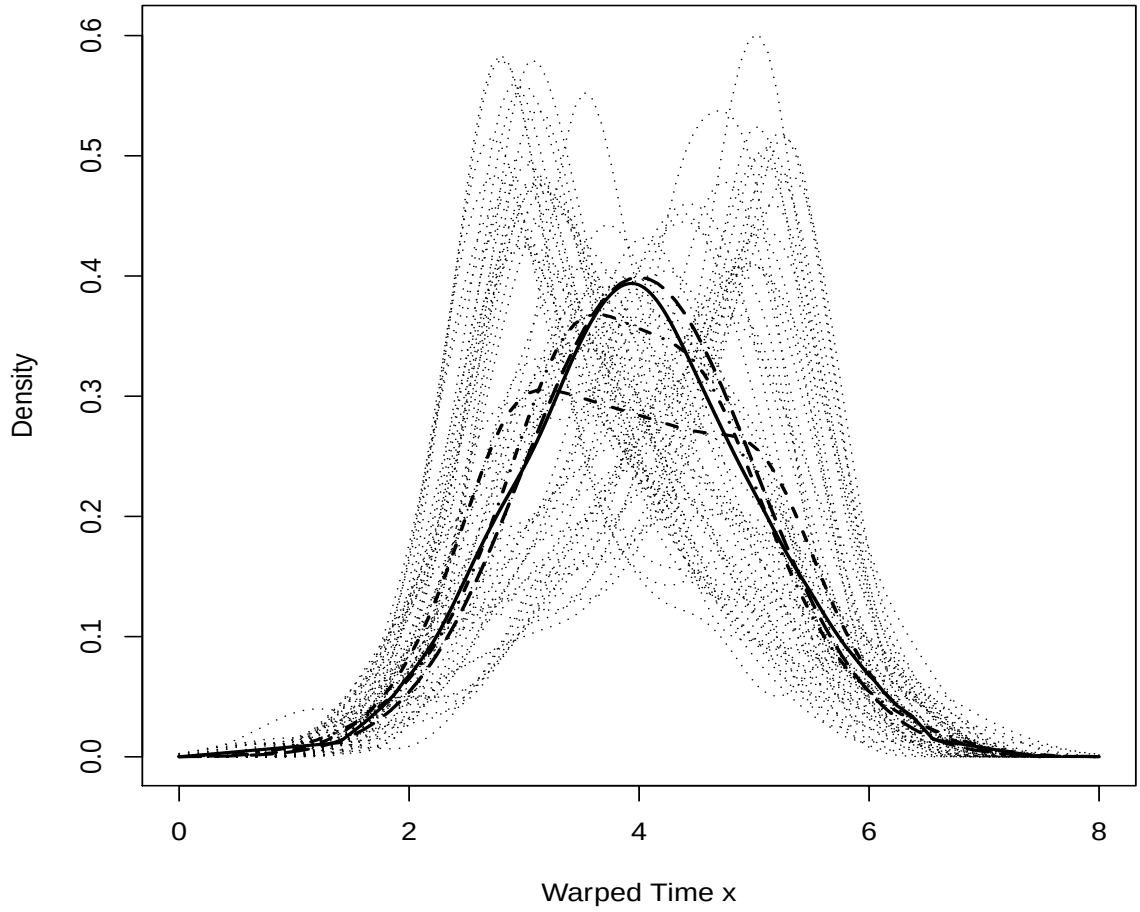

Figure 4: Results for Simulation 1: Individual kernel-smoothed densities $\tilde{f}_k$ (dotted), $k = 1, \ldots, 60$, target density $f_{\oplus}$ (3) (long-dash), proposed estimated synchronized overall density $\hat{f}_{\oplus}$ (9) (solid), cross-sectional estimate $\hat{f}_{CS}(x)$ (11) (dashed) and shift-averaged estimate $\hat{f}_{SRCS}(x)$ (12) (dot-dash).

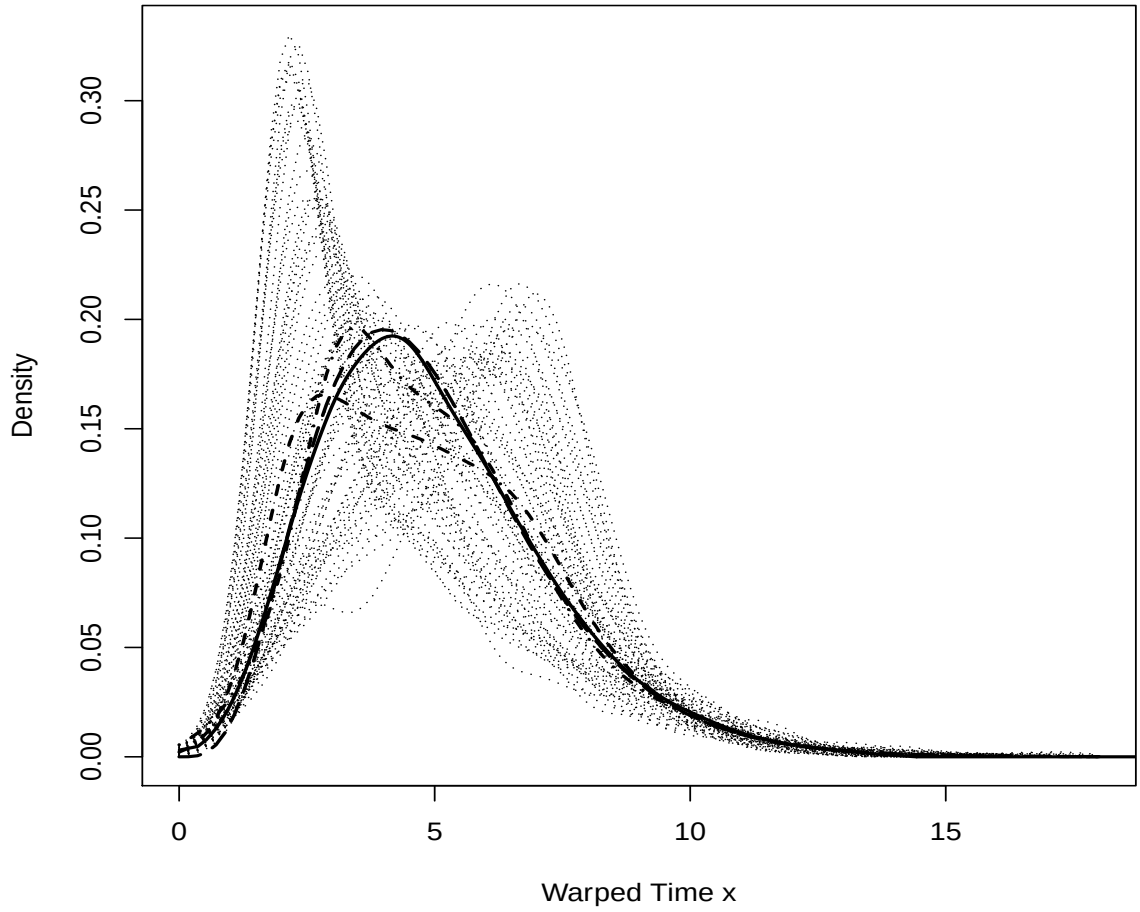

Figure 5: Results for Simulation 2: Individual kernel-smoothed densities $\tilde{f}_k$ (dotted), $k = 1, \ldots, 60$, target density $f_{\oplus}$ (3) (long-dash), proposed estimated synchronized overall density $\hat{f}_{\oplus}$ (9) (solid), cross-sectional estimate $\hat{f}_{CS}(x)$ (11) (dashed) and shift-averaged estimate $\hat{f}_{SRCS}(x)$ (12) (dot-dash).

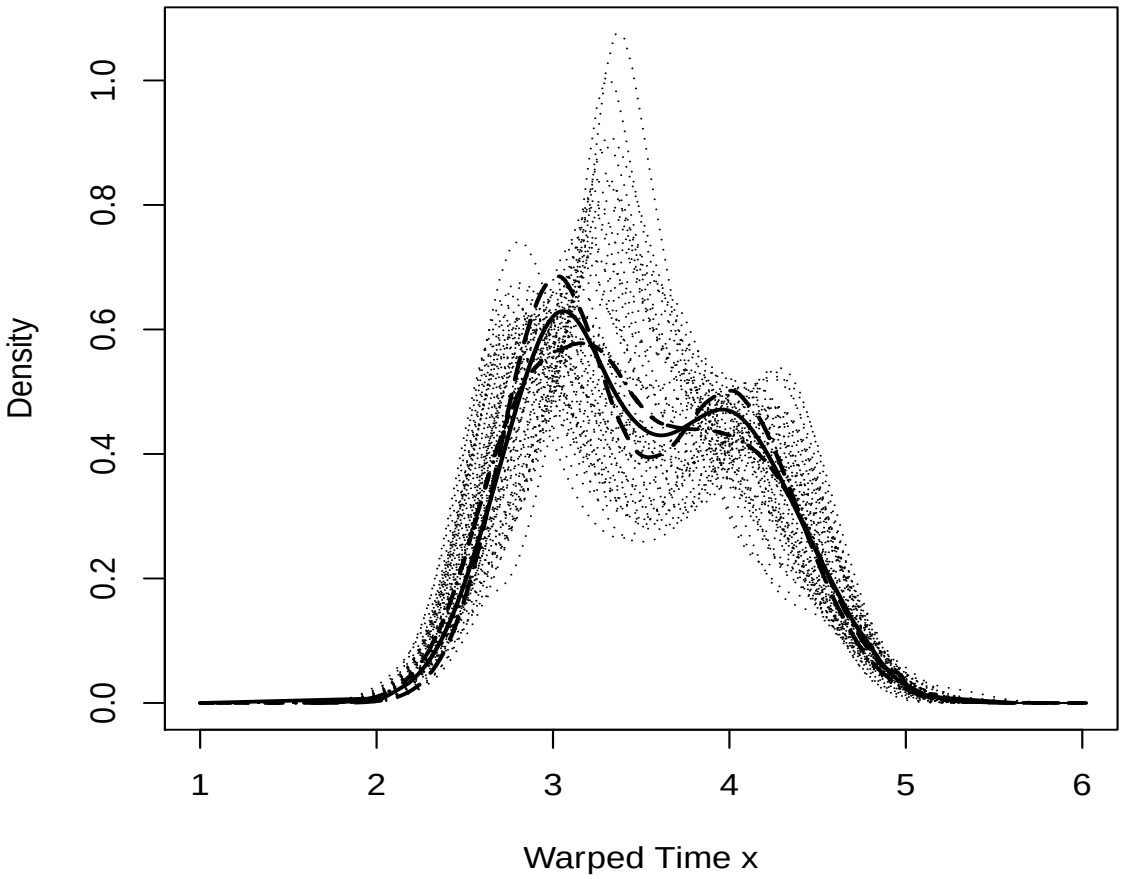

Figure 6: Results for Simulation 3: Individual kernel-smoothed densities $\tilde{f}_k$ (dotted), $k = 1, \ldots, 60$, target density $f_{\oplus}$ (3) (long-dash), proposed estimated synchronized overall density $\hat{f}_{\oplus}$ (9) (solid), cross-sectional estimate $\hat{f}_{CS}(x)$ (11) (dashed) and shift-averaged estimate $\hat{f}_{SRCS}(x)$ (12) (dot-dash).

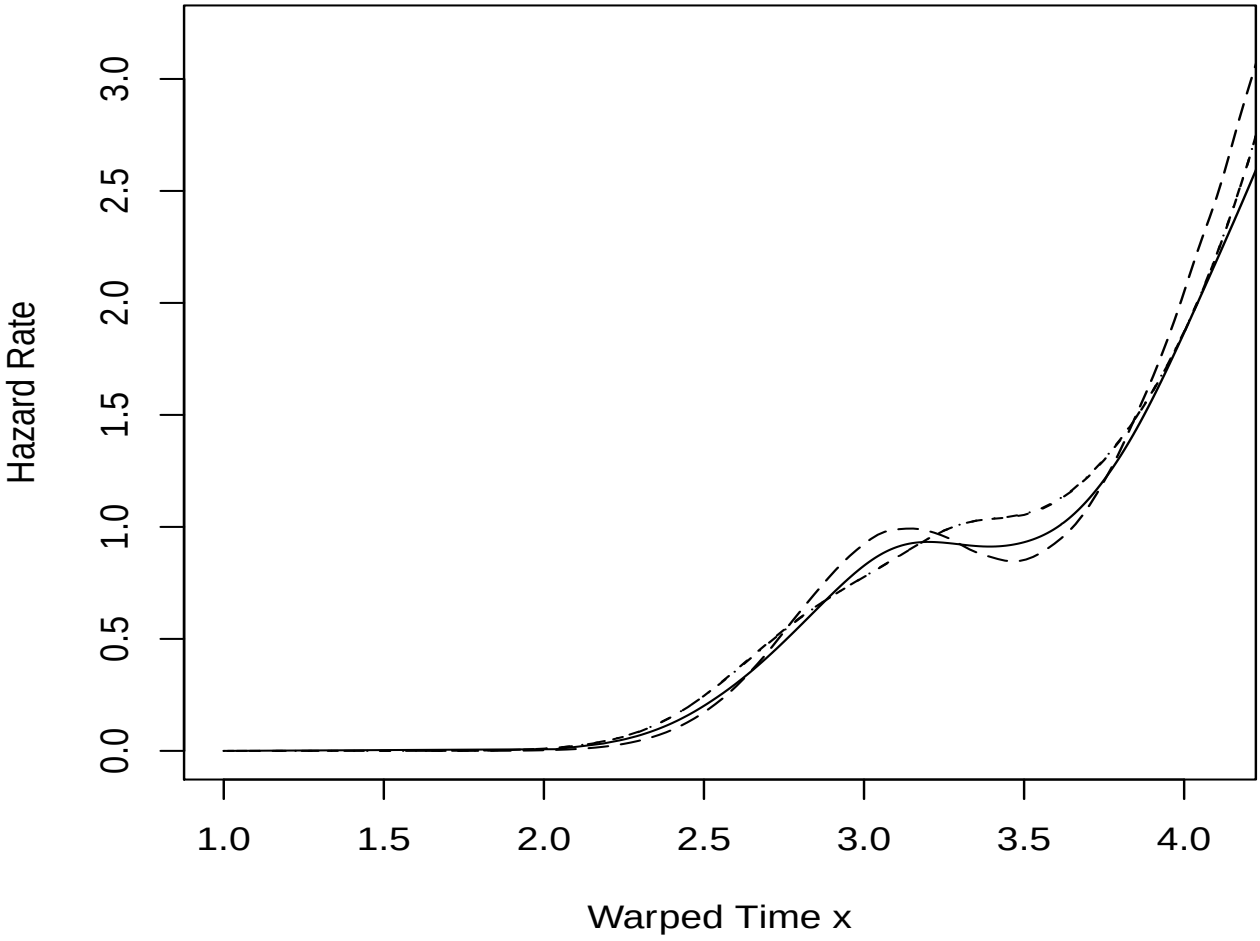

Figure 7: Results for Simulation 3: Estimated hazard functions with target hazard function $\lambda_{\oplus}$ (long dash), comparing the estimates based on quantile synchronization (defined in eq. (10), solid), cross-sectional averaging (dashed) and shift-averaging (dot-dash).

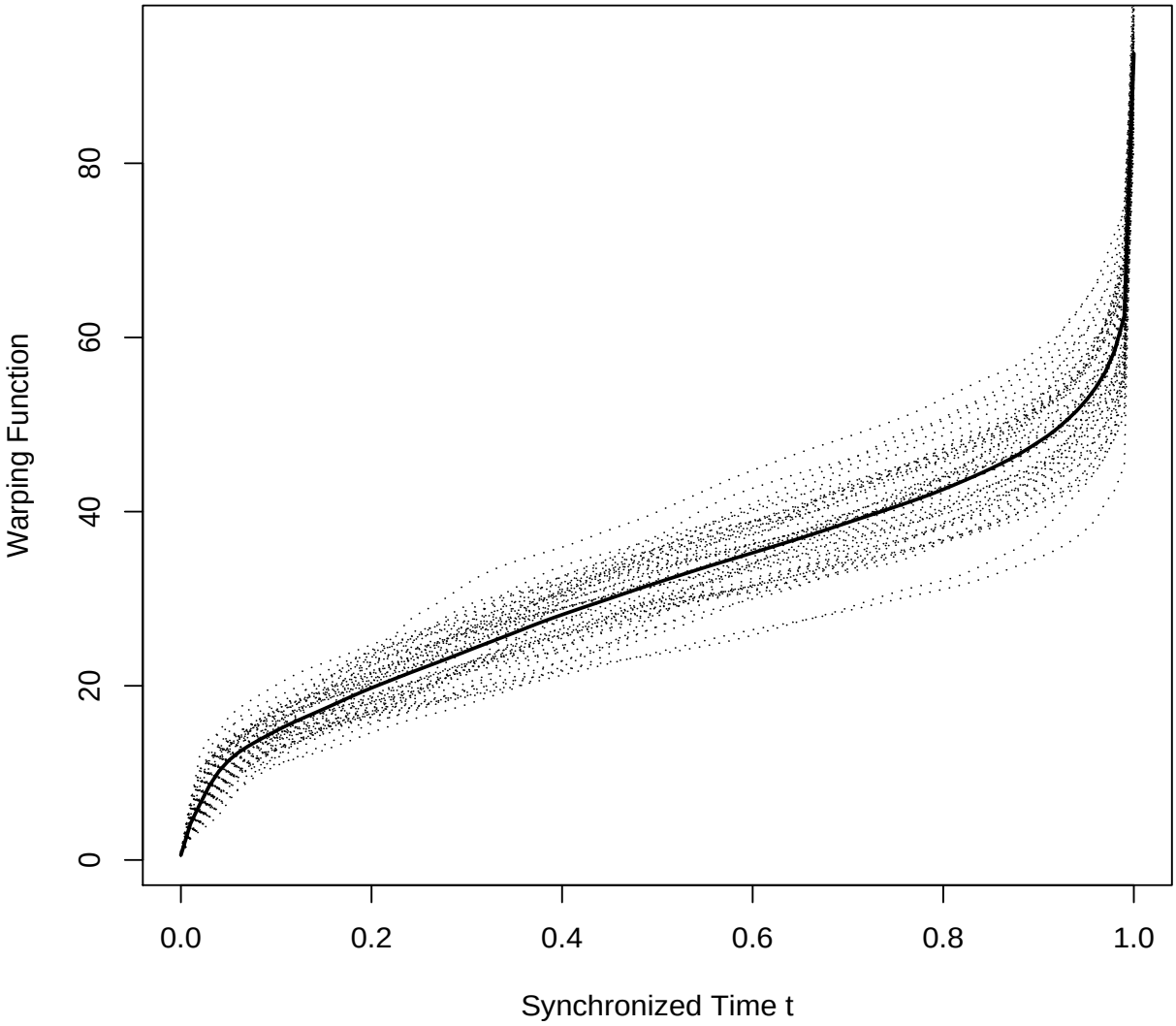

Figure 8: Estimated synchronization functions for each individual cohort $\tilde{F}_k^{-1}(t)$ (7) (dotted), $k = 1, 2, ..., 54,\ t \in [0, 1]$, and the estimated overall synchronization mapping $\hat{F}_0^{-1}(t)$ (8) (solid), for the mexfly data.

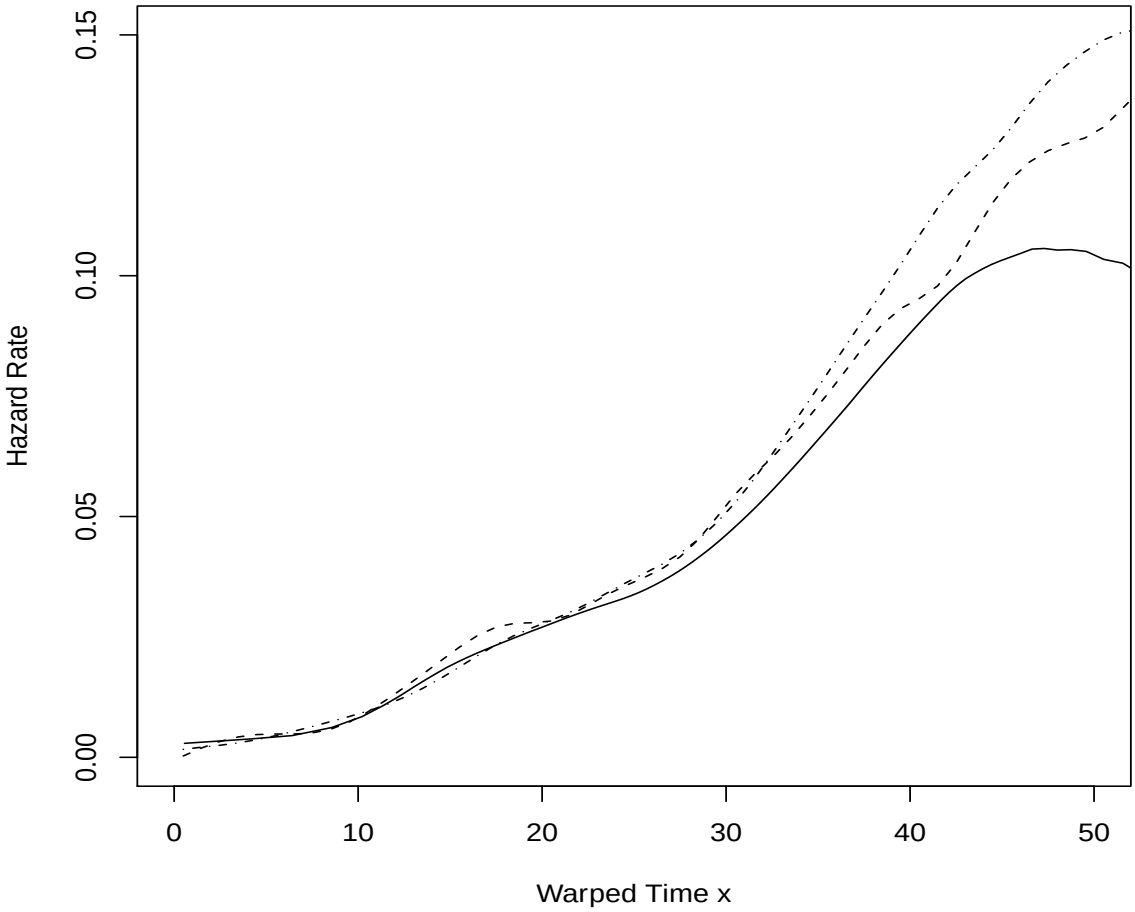

Figure 9: Hazard function estimate $\hat{\lambda}_{\oplus}$ (10) (solid), derived from the synchronized overall density estimate $\hat{f}_{\oplus}$ (9), hazard rate estimate derived from the cross-sectional density estimate $\hat{f}_{CS}(x)$ (11) (dashed), and from the shift-registered cross-sectional average density $\hat{f}_{SRCS}(x)$ (12) (dot-dash), for the mexfly data.

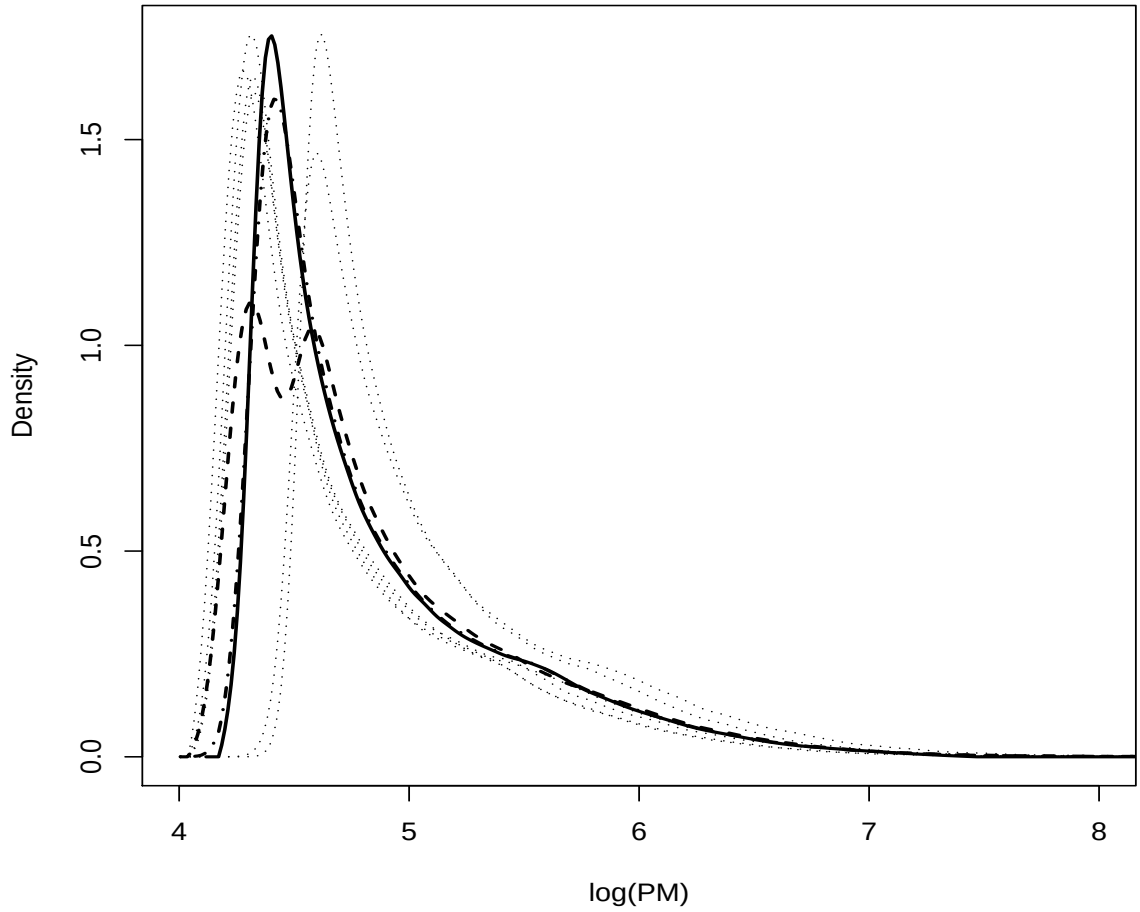

Figure 10: Kernel-smoothed densities of log(PM) for each array $\tilde{f}_k$ (dotted), $k = 1, \ldots, 6$, for the normal (2N) mouse data: Synchronized overall density estimate $\hat{f}_{\oplus}$ (9) (solid), cross-sectional density estimate $\hat{f}_{CS}(x)$ (11) (dashed) and shift-registered cross-sectional density estimate $\hat{f}_{SRCS}(x)$ (12) (dot-dash).

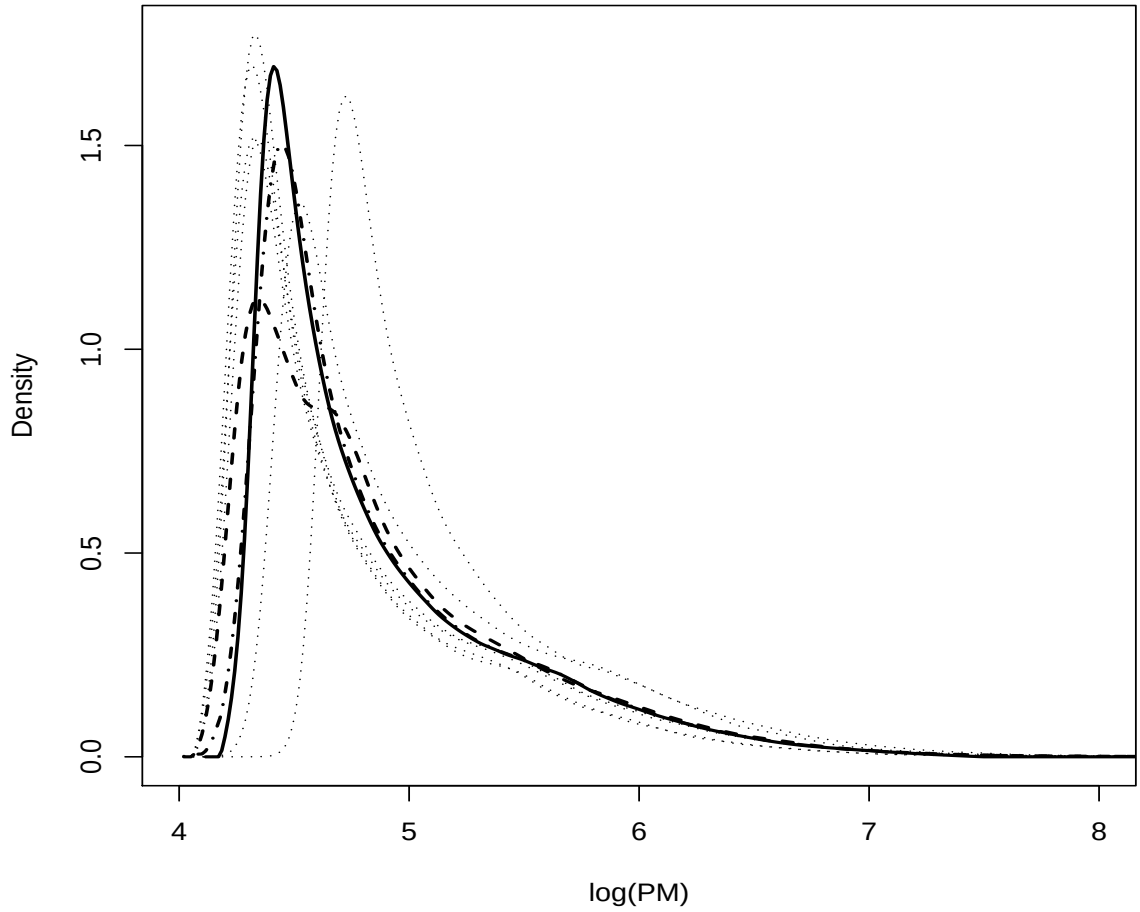

Figure 11: Kernel-smoothed densities of log(PM) for each array $\tilde{f}_k$ (dotted), $k = 1, \ldots, 6$, for the Ts1Cje mouse data: Synchronized overall density estimate $\hat{f}_{\oplus}$ (9) (solid), cross-sectional density estimate $\hat{f}_{CS}(x)$ (11) (dashed) and shift-registered cross-sectional density estimate $\hat{f}_{SRCS}(x)$ (12) (dot-dash).