Brazilian Journal of Probability and Statistics
2009, Vol. 23, No. 1, 1–17
DOI: 10.1214/08-BJPS004

Nonparametric maximum likelihood estimation of randomly time-transformed curves

Birgitte B. Rønn[a] **and Ib M. Skovgaard**[b]

[a]*Genmab a/s*
[b]*University of Copenhagen*

Abstract. Alignment of curves by nonparametric maximum likelihood estimation can be done when the individual transformations of the time axis is assumed to be of a parametric form, known up to some individual unobserved random parameters. We suggest a fast algorithm, based on a Laplace approximation, to find the nonparametric maximum likelihood estimator (NPMLE) for the shape function. We find smooth estimates for the shape functions without choosing any smoothing parameters or kernel function and we estimate realizations of the unobserved transformation parameters that align the curves to satisfy the eye. The method is applied to two data examples of electrophoretic spectra on feta cheese samples and on wheat samples, respectively. A small simulation study indicates reasonable robustness against assumptions regarding the error covariance function.

1 Introduction

When data from a process evolving in time is recorded the variation between replications is both in amplitude and phase. Here amplitude refers to the (vertical) variation in the response variable, whereas the variation in phase is meant to cover the differences between the individual timing of the curves. The individual transformed time may correspond to biological time, physical time or some artificial time depending on circumstances of the experiment that might vary beyond the control of the researcher. An example of replicated process data is shown in the left panel of Figure 1, where nine electrophoretic spectra of the same feta cheese are plotted versus time. The technique used is capillary electrophoresis. Thus, the x-axis represents the time of migration through the capillary and the y-axis represents intensity, reflecting the concentration of the substance with the given migration time. There are 412 observations for each samples. The data are seen in the left panel of Figure 1.

Since the 9 samples were highly variable in their level of intensity, background intensity was initially removed before plotting and further analysis. This was done using the same method as in Glasbey, Vali and Gustafsson (1995). The data were

Key words and phrases. Curve alignment, Laplace approximation, nonparametric maximum likelihood estimation, self-modeling regression, semiparametric model, warping.
Received October 2007; accepted February 2008.

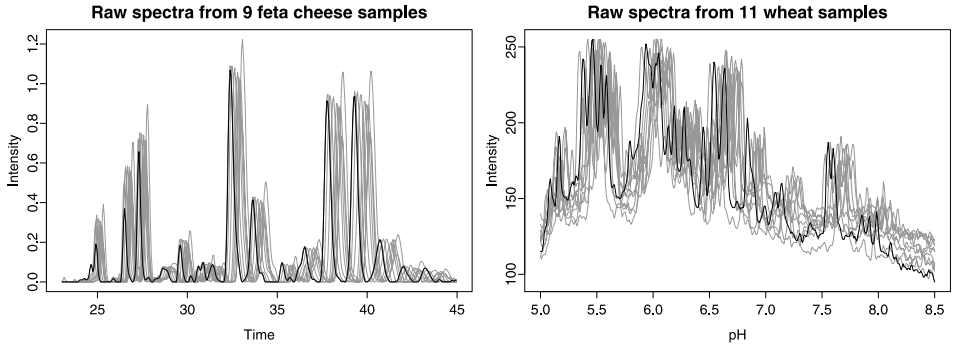

Figure 1 *Two examples of observed electrophoretic spectra: capillary electrophoresis with nine samples of feta cheese (left) and isoelectric focusing with 11 samples of wheat (right). One sample is enhanced in each plot.*

collected to obtain information on the protein profile of feta cheese and replications were made to determine which features were real and which were artifacts occurring only in some replicates due to disturbances from other sources than the cheese. Here horizontal variation between replications is unavoidable due to the nature of experiment. The variation in phase seems to be by far the most substantial and the cross-sectional mean clearly is a poor estimator of the shape function of the protein profile, since it will be less peaky than any of the nine individual curves. The feta cheese experiment is described in detail in Wium, Kristiansen and Qvist (1998).

Another sample of curves was recorded to characterize variety of wheat. Electrophoretic spectra of ten different varieties of wheat were made by isoelectric focusing on eleven different plates. The eleven spectra of variety number one are shown in the right panel of Figure 1. Here the x-axis represents the isoelectric pH-value for the substance, not time.

It is difficult to see much in the figure due to the substantial variation in phase between the individual curves. The features of the eleven curves differ more than the features of the nine feta cheese curves, and the variation in amplitude, the "measurement" error on the signal, is of significant size in the wheat data. Both types of variation must be accounted for in order to obtain a meaningful estimator for the shape function of interest. For details on the wheat experiment and data see Jensen et al. (1995).

Data of the type above can be modeled by a smooth function evolving in time, subject to individual transformations of the time axis, and measured with noise,

$$Y_{ij} = m\{g_i^{-1}(t_{ij})\} + e_{ij}, \tag{1.1}$$

where the jth observation of curve i at time point t_{ij} is denoted Y_{ij}, the common shape function is denoted m, the individual time transformation is denoted g_i and the error term is denoted e_{ij}.

Several structural assumptions on the time transformation might be reasonable for practical applications. Estimation of the transformations with few assumptions can be done nonparametrically, for example, by so-called dynamic time warping, considered in Wang and Gasser (1997), Ramsay and Li (1998), Ramsay and Silverman (1997) and more recently by Brumback and Lindstrom (2004) and Liu and Müller (2004). The above papers study statistical aspects of fitting the individual time transformations by time warping, an idea originally introduced within the engineering literature; see the above papers for references. When both the shape function, m, and the individual time transformations, g_i, $i = 1, \ldots, n$, are estimated nonparametrically the horizontal variation must be separated from the vertical. Otherwise, features of the shape function can be absorbed in the time transformations and visa versa. In some applications the focus might be on transforming observed curves to match a reference curve, hence nonparametric estimation of the transformation is clearly a method allowing for a flexible class of transformations. In other applications, as the two examples introduced above, the focus is on estimation of the common shape function, hence the transformations of the time axes are merely experimental noise and might therefore be modeled parametrically with the parameters considered as nuisance parameters. This can be done within the framework of self-modeling regression, which spans a wide class of regression models. The models similar to model (1.1), when the time transformation is assumed to be parametric, $g_i(t) = g_{\theta_i}(t)$, are models of this type. Here the time transformation is assumed to be the same, up to some parameter θ_i, for all individuals. Several suggestions have been made on how to estimate the parameters in self-modeling regression; see, for example, Lawton, Sylvestre and Maggio (1972), Stützle et al. (1980), Kneip and Gasser (1992) and Kneip and Engel (1995). The very simple special case of model (1.1), where the individual time axes are assumed to be the observed time axis subject to individual rigid shifts, $g_i^{-1}(t) = t - \theta_i$, has been studied in Rønn (2001). The interpretation of the time transformations as experimental variation has been adopted and the shift parameters have been modeled as unobserved nuisance parameters with some known distribution. A nonparametric maximum likelihood approach has been followed, inspired by the applications of this approach within event history analysis; see, for example, Gill (1989), Fernholz (1983), Groeneboom and Wellner (1992), Murphy (1995) and Parner (1998). This approach has been generalized to transformations involving more parameters in Decker, Rønn and Jørgensen (2000) and Gervini and Gasser (2005) where the nonparametric score equation for the shape function has been shown to lead to a fairly simple equation, immediately suggesting an iterative algorithm for estimation. The calculations involved are, however, substantial and while Gervini and Gasser (2005) suggest a simulation based estimation procedure, we show in the present paper that simple and accurate approximations lead to a relatively fast and reliable estimation procedure.

The approach from the latter three above-mentioned references is based on a nonparametric maximum likelihood estimation (NPMLE) for model (1.1) with a

parametric time transformation, $g_i(t) = g_{\theta_i}(t)$, where g is a known function of the p-dimensional transformation parameter θ_i. The transformation parameters are considered to be unobserved random variables, and their distribution enters the likelihood as a penalty on extreme transformations. To approximate the critical terms in the algorithm, a Laplace approximation to the crucial integral is suggested and the expressions leading to an approximate NPMLE of the shape function are worked out. We then apply the algorithm to the feta cheese data and the wheat data and show in a small simulation study that the method also works well in situations where the assumptions from the working model are not met. The present paper deals with estimation of the shape function, but the proposed algorithm also provides estimates for the parameters of the time transformation and the set of smooth, aligned curves. Hence, the method can serve as pre-processing of data prior to functional analysis by principal component analysis or following approaches described in, for example, Brumback and Rice (1998), Ramsay and Silverman (2002) or Anselmo, Dias and Garcia (2005). The latter uses methods from Ramsay and Li (1998) as a step in the functional data analysis process. The wheat profiles presented represent 1 out of 10 varieties and the aim of the study was to classify the profiles. Good alignment of the profiles was essential to achieve successful classification. Similarly, protein profiles of plant seed oil have been successfully classified by the method proposed in the present paper; see Decker, Rønn and Jørgensen (2000). Note that, as in Rønn (2001), an advantage of the method is that the degree of smoothing is controlled by the data through the probability that a given data point contributes to the function value in question. The present work was introduced in the unpublished thesis Rønn (1998).

2 A model for randomly time-transformed curves

The shape invariant model corresponding to randomly time-transformed curves is given by

$$m_i(s) = m\{g_{\theta_i}^{-1}(s)\},$$

where m_i is the ith curve, θ_i is the ith transformation parameter, g_{θ_i} is the transformation function and m is the shared shape function, defined on some interval, J, on which also the transformations, g_θ, are defined. The transformation parameter θ_i is assumed to follow a distribution with continuous density function, f_θ, and compact support $\text{supp}(f_\theta) = \Theta \subset \mathbb{R}^p$. The mean of the distribution is assumed to correspond to no transformation, $g_\xi(t) = t$ for all $t \in \mathbb{R}$, where $\xi = \text{E}(\theta_i)$. Each transformation, g_θ, is assumed to be strictly increasing in t,

$$\frac{\partial}{\partial s} g_\theta(s) > 0 \qquad \text{for all } s \in J \subset \mathbb{R}, \theta \in \Theta,$$

and hence a strictly increasing inverse, g_θ^{-1}, exists. In some settings other constraints on the transformation may be natural. If, for example, each curve represents the growth curve of an individual, the observed time axis represents the age of the individual. The biological age, which corresponds to the transformed observed age, must be 0 in the same point as the observed age for biological age to be meaningful. Thus, depending on the situation, boundary conditions on g, such as $g_\theta(t_0) = t_0$ for all θ and some fixed point t_0, may be natural.

Observations from the above model obtained at deterministic time points may be subject to further random variation modeled as the error term e_{ij} in the equation

$$Y_{ij} = m\{g_{\theta_i}^{-1}(t_{ij})\} + e_{ij}, \tag{2.1}$$

where the observation of the ith individual at the jth point in time denoted Y_{ij} and the corresponding time point is denoted t_{ij}, $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, N\}$. The error terms, e_{ij}, are assumed to be independent, normally distributed with mean 0 and variance σ^2. This assumption is rarely realistic, but since it is used to derive the estimator of the mean, essentially by introducing a penalty on the residual vertical deviations, it is not as prohibitive as it may seem.

3 Nonparametric maximum likelihood estimation

The infinite-dimensional parameter we want to estimate is the shared shape function. The NPMLE for the shape function is the shape function that maximizes the likelihood function. The log-likelihood function, as a function of the parameter of interest, m, is

$$\begin{aligned} l(m) &= \sum_{i=1}^{n} \log\left\{ \int_\Theta f_{Y_i|\theta}(u)\, f_\theta(u)\, du \right\} \\ &= \sum_{i=1}^{n} \log\left(\int_\Theta (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2} \|Y_i - m\{g_u^{-1}(t_i)\}\|^2 \right] \cdot f_\theta(u)\, du \right), \end{aligned}$$

where $f_\theta(u)$ and $f_{Y_i|\theta}(u)$ denote the density functions for the distribution of the transformation parameter and the density function of the observations from the ith curve given the transformation parameter, both evaluated in the p-dimensional vector u. The N-dimensional vectors Y_i, t_i, $g_u^{-1}(t_i)$ and $m\{g_u^{-1}(t_i)\}$ have elements Y_{ij}, t_{ij}, $g_u^{-1}(t_{ij})$ and $m\{g_u^{-1}(t_{ij})\}$, respectively. Furthermore, the usual norm of an N-dimensional vector $v \in \mathbb{R}^N$ is denoted $\|v\| = (\sum_j v_j^2)^{1/2}$, and the integration is with respect to the usual Lebesgue measure on $\mathbb{R}^p$.

In Decker, Rønn and Jørgensen (2000) and Gervini and Gasser (2005) the infinite-dimensional score function of m has been derived and shown to be zero at $m = \hat{m}$ if and only if

$$\hat{m}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^N Y_{ij}\, \hat{f}_{Z_{ij}|Y_i}(t)}{\sum_{i=1}^n \sum_{j=1}^N \hat{f}_{Z_{ij}|Y_i}(t)}, \tag{3.1}$$

where

$$Z_{ij} = g_U^{-1}(t_{ij})$$

denotes the time-point corresponding to t_{ij} by the back-transformation given by U. Existence of a solution, $\hat{m}$, to the equation above has been shown only for compact parameter space, essentially ruling out an infinite-dimensional model. However, an approximate solution may well exist and be found by the algorithm.

Noting that the posterior densities on the right-hand side of the equation, $\hat{f}_{Z_{ij}|Y_i}$, are estimates depending on the shape function $\hat{m}$ itself, an obvious choice is to iterate the computation of the right-hand side of the equation, plugging in the current estimate of m. This is what is done in Rønn (2001) for the case of rigid shifts of the curves, and in Decker, Rønn and Jørgensen (2000) and Gervini and Gasser (2005) in the present case. More precisely the algorithm consists of the following:

A. *Initialization:*

- Decide on a grid of time points on which the score equation shall be fulfilled, for example, the grid of observation points, $t_1, \ldots, t_N$.
- Calculate an initial estimate, $\hat{m}_0$, of the smooth function m, for example, as the cross-sectional mean.

B. *Iteration:*

- Calculate estimates for the weight functions, $\hat{w}_{ij}(\hat{m}_{k-1})(t)$, by plugging in the estimate of the smooth function from the previous step $\hat{m}_{k-1}$.
- Find values of a new estimate, $\hat{m}_k$, of the smooth function on the grid by the weighted mean of observations obtained from the score equation,

$$\hat{m}_k(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N} Y_{ij}\hat{w}_{ij}(\hat{m}_{k-1})(t)}{\sum_{i=1}^{n}\sum_{j=1}^{N} \hat{w}_{ij}(\hat{m}_{k-1})(t)}. \tag{3.2}$$

  Values of the new estimate in any time point can then be found by for example cubic spline interpolation between the grid points.

The estimator $\hat{m}(t)$ is a Nadaraya–Watson type kernel estimator, where any observation Y_{ij} contributes according to the likelihood of the corresponding time point, t_{ij}, to be transformed into t, measured in the empirical posterior distribution of the transformation parameters given data. Hence the width of the kernel is determined by certainty in data of the time transformations.

4 Approximation to the empirical posterior transformation density

In the present section a Laplace-type approximation is introduced to approximate the crucial quantities needed in equation (3.1) such that its solution becomes computationally feasible. An even simpler approximation is also given, based on a

normal approximation to the empirical posterior distribution of the transformation parameter for each curve. The two approximations are given in equations (4.1) and (4.3), respectively.

In each step of the algorithm we need to estimate the weight functions for a given shape function, given as

$$w_{ij}(t) = f_{Z_{ij}|Y_i}(t).$$

However, the calculation of this empirical posterior density of the transformed time given data is highly demanding. The empirical posterior density of the transformation parameter given data is

$$\begin{aligned} f_{\theta|Y_i}(u) &= \frac{f_{Y_i|\theta=u} f_\theta(u)}{\int_\Theta f_{Y_i|\theta=v} f_\theta(v)\,dv} = \frac{(2\pi\sigma^2)^{-N/2}\exp\{b_i(u)\}}{\int_\Theta (2\pi\sigma^2)^{-N/2}\exp\{b_i(v)\}\,dv} \\ &= c\exp(-b_i(u)), \end{aligned}$$

where c is the normalizing constant and

$$b_i(u) = -\frac{1}{2\sigma^2}\sum_{k=1}^{N}[Y_{ik} - m\{g_u^{-1}(t_{ik})\}]^2 + \log\{f_\theta(u)\}.$$

For this situation, with a posterior parameter-density of the form above, Tierney, Kass and Kadane (1989) give a Laplace-type approximation to a posterior density of a (smooth) function, g say, of the parameter; in our case

$$g_{ij}(u) = g_u^{-1}(t_{ij}).$$

Their approximation, calculated at the estimated function $\hat{m}$, reads

$$\hat{f}_{Z_{ij}|Y_i}(t) \approx (2\pi)^{-1/2}\beta_{ij}(t)\exp\{-b_i(\tilde{u}_{ij}(t)) + b_i(\hat{u}_i)\}, \tag{4.1}$$

where $\tilde{u}_{ij}(t)$ minimizes $b_i(u)$ subject to the condition $g_{ij}(u) = t$ while $\hat{u}_i$ is the unconstrained minimum, and where β_{ij}, G_{ij} and λ_{ij} are defined by

$$\begin{aligned} \beta_{ij}(t) &= |G_{ij}(\tilde{u}_{ij}(t))|^{-1/2}|g'_{ij}(\tilde{u}_{ij}(t))G_{ij}(\tilde{u}_{ij}(t))^{-1}g'_{ij}(\tilde{u}_{ij}(t))^T|^{-1/2} \\ &\quad \times |b_i''(\hat{u}_i)|^{1/2}, \\ G_{ij}(\tilde{u}_{ij}(t)) &= b_i''(\tilde{u}_{ij}(t)) - \lambda(t)g''_{ij}(\tilde{u}_{ij}(t)), \\ \lambda_{ij}(t) &= \frac{g'_{ij}(\tilde{u}_{ij}(t))b'_i(\tilde{u}_{ij}(t))^T}{g'_{ij}(\tilde{u}_{ij}(t))g'_{ij}(\tilde{u}_{ij}(t))^T}. \end{aligned}$$

Here and throughout this section we use a prime to indicate a (vector) derivative with respect to the argument of the function.

Notice the interpretation that $\hat{u}_i$ is the empirical posterior mode estimate of the transformation parameter for curve i while $\tilde{u}_{ij}(t)$ similarly estimates this parameter given that t_{ij} is transformed to t.

Tierney, Kass and Kadane (1989) prove their formula to give an asymptotic approximation of high accuracy when the exponent, $b_i(u)$, is proportional to some n going to infinity. The crucial point here is that the exponent is well approximated by a quadratic in the neighborhood of the minimum where the contribution to the integral matters. In our experience from examples like the ones presented in this paper, this is a very good approximation.

There is one remaining computational problem: the minimum must be found for each curve, i, for each combination of estimation point, t, and observation point, t_{ij}. However, for the large majority of these terms the contribution to the sum will be negligible because the empirical posterior probability may essentially rule out a transformation mapping t to the vicinity t_{ij}. Thus, as a first part, we truncate the sum over j to those values of t_{ij} that are in the neighborhood of

$$\hat{t}_{ij} = g_{ij}(\hat{u}_i),$$

the unconstrained estimate of the transformed value of t.

Second, we may for each j use a Taylor series expansion of $\tilde{u}_{ij}(t)$ as a function of t around $\hat{t}_{ij}$. The resulting Taylor series expansion to first order is

$$\tilde{u}_{ij}(t) \sim \hat{u}_i + \frac{\partial}{\partial t}\tilde{u}_{ij}(t)\Big|_{t=\hat{t}_{ij}} (t - \hat{t}_{ij}),$$

where the derivative may be shown to be

$$\frac{\partial \tilde{u}_{ij}(t)}{\partial t}\Big|_{t=\hat{t}_{ij}} = \frac{\{b''(\hat{u}_i)\}^{-1} g'_{ij}(\hat{u}_i)^T}{g'_{ij}(\hat{u}_i)\{b''(\hat{u}_i)\}^{-1} g'_{ij}(\hat{u}_i)^T}. \tag{4.2}$$

The only minimizations required for this method are one for each curve, leading to the estimated transformation $\hat{u}_i$.

Although the Laplace approximation is generally fast and accurate, the large number of data points often available in the warping problems calls for an even simpler and faster candidate approximation. Approximating the empirical posterior density of Z_{ij} given Y_i by a normal distribution with mean $\hat{u}_i$ and inverse variance $b''_i(\hat{u}_i)$, we get

$$\tilde{f}_{Z_{ij}|Y_i} = (2\pi)^{-1/2}|D|^{1/2} \exp\left\{-\frac{1}{2}D(t - \hat{t}_{ij})^2\right\}, \tag{4.3}$$

where

$$D = \{g'_{ij}(\hat{u}_i)(b''(\hat{u}_i))^{-1} g'_{ij}(\hat{u}_i)^T\}^{-1}.$$

We recommend that in each particular problem both approximations are tried for speed and accuracy, although the latter (simpler) approximation has been excellent in the examples we have tried.

Linear back-transformations including a shift

Simplifications for $\beta(t)$ arise for certain models. Consider back-transforms of the type

$$g_\theta^{-1}(t) = \theta_1 + \theta_2 a_2(t) + \cdots + \theta_p a_p(t),$$

where $a_k(t)$ for $k = 1, \ldots, p$ are known functions. Polynomial back-transformations with $a_k(t) = t^{k-1}$, as we shall use in the examples, are of this form.

Define the p-vector

$$a(t) = (1, a_2(t), \ldots, a_p(t)).$$

Then $g_{ij}(u)$ is linear in u with derivative $a(t_{ij})$ so that

$$\beta(t) = (a(t_{ij})\{b''(\tilde{u}_{ij})\}^{-1} a(t_{ij})^T)^{-1/2} |b''(\tilde{u}_{ij})|^{-1/2} |b''(\hat{u}_i)|^{1/2}, \tag{4.4}$$

and $G_{ij} = b''$.

Estimation of the variance and covariance parameters

We have, so far, neglected the estimation of the transformation parameters. In many applications we might only fix the distributions up to the residual variance, σ^2, and the covariance matrix of the transformation parameters, Σ. Approximate maximum likelihood estimates for these variance parameters can be calculated in each step of the above algorithm and used in the calculations of the weight functions. Differentiation of the log-likelihood function with respect to the error variance in combination with a Laplace approximation similar to those used in the previous section leads to the usual estimate of the error variance as the mean squared residual

$$\hat{\sigma}^2 = \frac{1}{nN} \sum_{i=1}^{n} \| Y_i - \hat{m}\{g_{\hat{\theta}_i}(t_i)\} \|^2.$$

By a similar argument we arrive at the plug-in type estimate of the covariance matrix for the transformation parameters,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - \xi)(\hat{u}_i - \xi)^T.$$

Whether the sum should rather be divided by $n - 1$ seems less important with data as the two examples, where the many features of the individual curves ensure a very peaky empirical posterior distribution.

Prediction of the transformation parameters

The unobserved random variables, corresponding to the individuals considered in the study, are estimated by their conditional mean given data,

$$\hat{\theta}_i = \mathrm{E}(\theta | Y_i).$$

In order to calculate an explicit expression for the above conditional mean we will have to calculate integrals similar to the integrals in the weight function. Hence, again we need Laplace-type approximations, and we find that

$$\hat{\theta}_i \approx \arg\min_u b_i(u),$$

or, in other words, we use the conditional (or empirical posterior) mode. Thus, the realization of the unobserved random transformation variables for the curve i, is estimated by $\hat{u}_i$, a quantity we need to calculate anyway to obtain an estimator for the shape function.

5 Examples

Various parametric time transformations may be considered in practical data analysis. The nine curves in the feta cheese example introduced above can be modeled, with the simplest possible individual transformation, namely individual shifts,

$$g_\theta^{-1}(s) = s - \theta.$$

In Rønn (2001), the shape function for the protein profile was estimated, when the shifts were assumed to follow a slightly truncated normal distribution. The analysis indicated that the transformation with individual shifts is too simple to fit the data perfectly. The resulting alignment is seen in the left panel of Figure 2.

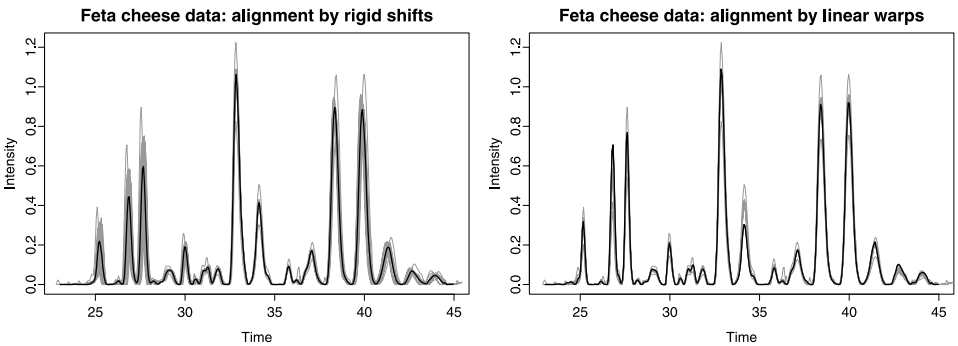

Figure 2 *The nine aligned profiles for the feta cheese example (grey) together with the estimated shape function (enhanced). The transformation alignment used is a rigid shift (left) and linear (right).*

A transformation that allows for stretching as well as shifting might be preferred. This two-dimensional transformation can be parameterized as follows:

$$g_\theta^{-1}(s) = \theta_1 + \theta_2(s - \bar{s}), \tag{5.1}$$

where the time scale has been centered to avoid too strong dependence between the two transformation parameters. The transformation parameters are assumed to follow a two-dimensional normal distribution

$$\theta_i \sim N_2\left\{\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right\}$$

in principle truncated to a (large) set giving monotone transformations. The estimates of the residual standard deviation and the standard deviations of θ_1 and θ_2 are $\hat{\sigma} = 0.046$, $\sqrt{\hat{\Sigma}_{11}} = 0.030$, $\sqrt{\hat{\Sigma}_{22}} = 0.016$, while the estimate of the correlation between the two transformation parameters is 0.0988. These results were obtained using the Laplace approximation (4.1), but virtually identical results were obtained using the even simpler and faster normal approximation (4.3). Thus, the deviations in the standard relative deviations given above were all less than 1 percent, and the gain in speed was about a factor five.

The estimated protein profile is shown, together with the aligned data profiles, in the right panel of Figure 2. Comparison of the two plots shows that the alignment as well as the estimated profile are clearly sharper with the linear alignment. The estimated individual linear time transformation is seen to align each of the nine protein profiles almost perfectly. Thus, the linear transformation model seems to fit the feta cheese data extremely well. It is also seen that the estimator is a smooth function that contains all the features present in all the nine curves. False peaks, occurring only in one curve each, are not visible in the estimator. Furthermore, the height of the peaks is close to the average height of the nine individual peaks, which is also a desirable property of an estimator.

For the wheat data, inspection of the profiles quickly reveals that rigid shifts are not sufficient to align the profiles (not shown). In the left panel of Figure 3 we see the aligned spectra together with the estimated shape function, using the linear transformation (5.1).

The linear transformation is seen to align the major features in the center of the observed interval very well, whereas the smaller bumps near the end of the observed interval are out of order. Using instead a quadratic transformation model given by

$$g_\theta^{-1}(s) = \theta_1 + \theta_2(s - \bar{s}) + \theta_3(s - \bar{s})^2, \tag{5.2}$$

we obtain the aligned spectra shown in the right panel of Figure 3. The quadratic transformation gives an almost perfect alignment of the data, although local fine-tuning on a very small scale might be desirable in a few instances.

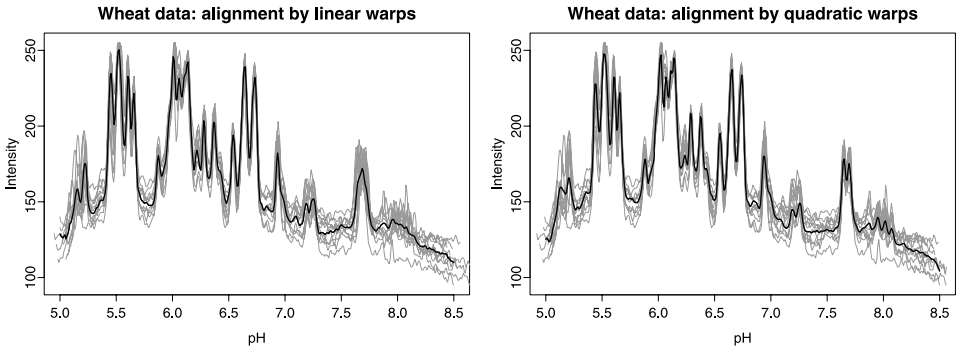

Figure 3 *The 11 aligned profiles for the wheat example (grey) together with the estimated shape function (enhanced). The transformation alignment used is linear (left) and quadratic (right).*

Thus, it is seen that in these realistic examples the model and the algorithm together provide a very satisfactory solution to the alignment of the curves and to the estimation of the common function.

The time per iteration for the two examples were between 2 and 4 seconds per iteration, using the (faster) normal approximation (4.3), and the convergence criteria were met after 33 iterations for the wheat data and 30 iterations for the feta cheese data. The algorithm was implemented in the statistical software R on a standard laptop. Visually, convergence was obtained after less than 10 iterations, but the stopping criteria were rather strict, demanding a relative change less than 0.0001 for all parameters including the transformation parameters and the function, m, evaluated on a fine grid. Finally, the hardest criterion to meet was an absolute change less than 0.01 in twice the negative log-likelihood, which involves the sum of squared errors for the entire set of observations and had a magnitude in the order 10000.

For comparison the we tried the function `register.fd` from the R-package `fda`, Ramsay (2007), which is based on methods from Ramsay and Li (1998) and used by Anselmo, Dias and Garcia (2005). Memory problems arose when a fine-scale representation matching the number of observations were attempted. With a representation of 100 base-vectors, alignment of one curve against another took 10 minutes and did not align even the major peaks correctly in the cases we tried. A reason is probably the large flexibility in the nonparametric class of time transformations. More knowledge of the function might well have helped as other representations and user choices may be made, but the method seems to be more directed towards smoother functions than the examples presented here.

6 Simulation study

The parametric transformation of the x-axis seems to align the profiles from the above examples very well. However, the working model assumptions of indepen-

dent errors are clearly not satisfied and also the assumption on normally distributed transformation parameters is an idealization, of course. Since the result is an estimation method, and not a statistical analysis involving probability calculations, the assumptions behind the likelihood function do not necessarily lead to poor estimation, but, at worst, to inefficiency compared to an ideal estimation method. To get some information of the sensitivity of the NPMLE to the distributional assumptions we made a small simulation study, with serial correlation, exponentially distributed error or exponentially distributed transformation parameters.

The feta cheese data were used as inspiration for the simulation and the estimated shape function under the model with random shifts was applied as the true shape function m_0. Data from the following four models for the errors and the random shifts were simulated:

1. The working model: independent normally distributed errors with standard deviation $\sigma = 0.031$ and independent normally distributed random shifts with standard deviation $\omega = 0.26$.
2. Autoregressive process of order 1 with autocorrelation $\rho = 0.89$ and marginal standard deviation $\sigma = 0.031$ for the errors and independent normally distributed random shifts with standard deviation $\omega = 0.26$.
3. Independent exponentially distributed errors with standard deviation 0.031 and independent normally distributed random shifts with standard deviation $\omega = 0.26$.
4. Independent normally distributed errors with standard deviation $\sigma = 0.031$ and independent exponentially distributed random shifts with standard deviation 0.26.

These cases were chosen to reflect various kinds of discrepancies from the working model without altering the signal-to-noise ratio.

For each model 100 data sets consisting of 10 shifted profiles were simulated. The NPMLE assuming independent normally distributed errors and shifts were calculated for each simulated data set as described in the present paper. The estimation procedure converged for all data sets and the 100 estimated NPMLEs and their mean are plotted together with the true profile m_0 in Figure 4. As the NPMLEs are only well defined up to a shift, the 100 NPMLEs were aligned by rigid shifts before plotting.

It is seen that the profile is remarkably well estimated in all cases; in particular the negligible bias in the estimation of the peaks is noteworthy. This latter conclusion follows since the dashed white line, representing the true curve, virtually falls on top of the mean of the 100 estimates, represented by the black line.

Plots of the 100 estimated shifts were made against the 100 true shifts together with identity lines for all four simulation scenarios in Figure 5. The estimated and the true shifts showed almost perfect agrement, suggesting that the algorithm is robust towards misspecification of the distribution of the transformation parameters.

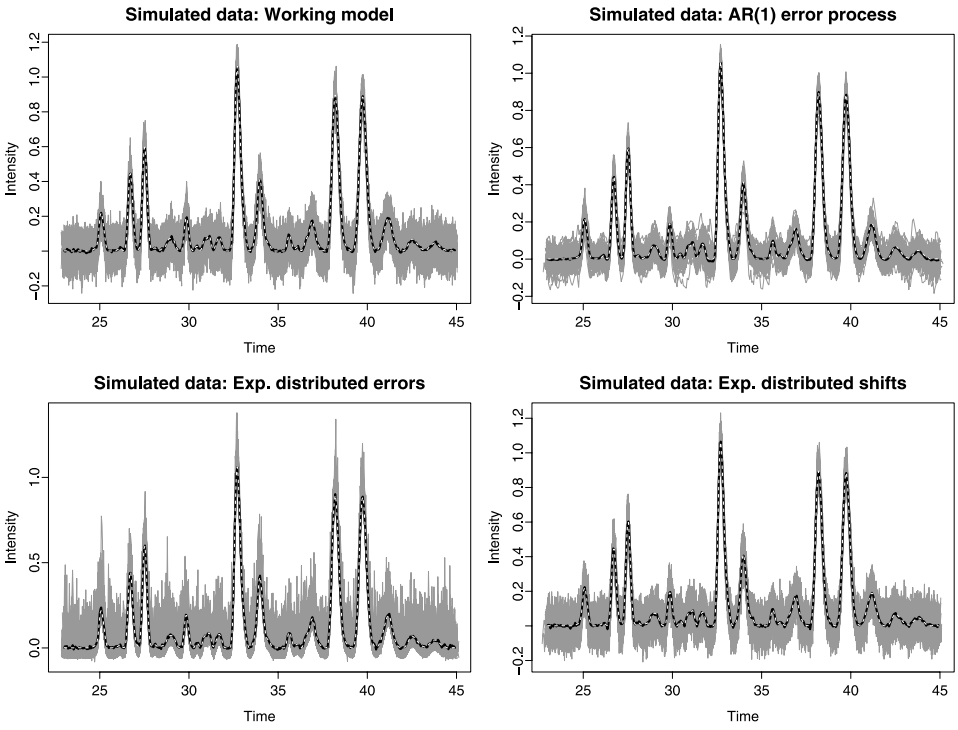

Figure 4 *The 100 estimated functions (grey) together with their mean (thick black line) and the true function (dashed white line) for each of the four simulation models.*

Figure 5 *The 100 estimated shifts versus the true shifts together with the identity line for four simulation scenarios. The vertical line is placed at zero, and the four cases have been shifted vertically together with the identity line drawn for each case.*

Closer inspection reveals reflections of the choice of error structure of the simulations in the four cases, but the important message here is that this does not interfere with the alignment of the curves or with the estimation of the shape function.

7 Discussion

The algorithm used here to obtain the estimator of the shape function generalizes the proposal for rigid shifts from Rønn (2001), and starts from the same working model and score equation, (3.1), as in Gervini and Gasser (2005), but their method of computation differs from ours by using simulations to compute the conditional means needed for the weights in the iterative calculation of (3.1). Optimization based on simulations is highly time consuming and it is hardly feasible to obtain the precision achieved in the examples. Our claim, originating from Rønn (1998) and Decker, Rønn and Jørgensen (2000), is that a Laplace approximation solves the problem accurately and efficiently in many problems among which are the feta cheese and wheat data. In fact, for these examples the normal approximation to the empirical posterior distribution solves the problem equally well and even more efficiently. Computations may still be a challenge, however, because of the complexity of the problem, and numerical problems may arise when data are very sharply peeked, as with the feta cheese data. Therefore one should be somewhat liberal with the convergence criteria, at the same time allowing for the possibility that an exact solution may not exist.

The two successful applications indicate that a good estimator for the shape function is found and that the alignment of the replicated curves is done in a satisfactory manner. The working assumption of independent normal errors is still in conflict with data obtained from real experiment, but the simulations suggest that the alignment works well anyway, and the estimated shape function is equally sharp as when the working model holds true. Generalization to more models with dependence is an obvious future reseach challenge. However, the reduction of the score equation to the iterated Nadaraya–Watson type kernel smoothing scheme (3.2) is not possible without the independence and entirely different algorithms should probably then be used.

The examples furthermore show that simple parametric transformations are sufficient for accurate alignment in these realistic examples, but also that it is important to consider other transformations than rigid shifts. When the alignment is part of a classification procedure it is important that the class of transformations used is not so rich that different shape functions, from different varieties, for example, easily align to look similar. This is avoided by use of low-dimensional parametric transformations.

The aligned profiles look even better if the profiles were preprocessed by removal of background. This was done in the feta cheese example but was avoided

in the wheat example to show that the alignment method is largely insensitive to such pre-processing. Thus, both examples were tried with and without background removal, with similar results.

Gervini and Gasser (2005) show that when the space of shape functions is compact, the estimate of the shape function converges at the rate of $1/\sqrt{n}$ towards the true shape function, when the sup-norm is used. The compactness assumption, however, essentially rules out a nonparametric model, because a compact set is either of finite dimension or is very fragmented by being everywhere thin. That is, for no function in the model is any neighborhood of this function contained in the model. We conjecture, on the basis of results in Stone (1980), that when the parameter space is infinite-dimensional, the asymptotic rate will be slower in the same way as for nonparametric regression.

Acknowledgments

We are grateful to Ib Søndergaard and to Kirsten Jensen for providing the wheat data and to Karsten B. Qvist for providing the feta cheese data. Also, we thank two anonymous reviewers for their constructive and useful comments. The research was supported by The Danish Research and Development Programme for Food technology through the Center for Advanced Food Studies.

References

Anselmo, C. A. F., Dias, R. and Garcia, N. L. (2005). Adaptive basis selection for functional data analysis via stochastic penalization. *Computational & Applied Mathematics* **24** 209–229. MR2186845

Brumback, B. A. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93** 961–976. MR1649194

Brumback, L. C. and Lindstrom, M. J. (2004). Self modelling with flexible, random time transformations. *Biometrics* **60** 461–470. MR2066281

Decker, M., Rønn, B. B. and Jørgensen, S. S. (2000). Thermally assisted in-line methylation and gas chromatography with statistical data analysis for determination of fatty acid distribution and fingerprinting of plant seeds and oils. *European Food Research and Technology* **211** 366–373.

Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals*. Springer, New York. MR0713611

Gervini, D. and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820. MR2234187

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scandinavian Journal of Statistics. Theory and Applications* **16** 97–128. MR1028971

Glasbey, C., Vali, L. and Gustafsson, J. (1995). A statistical model for unwarping of 1-D electrophoresis gels. *Electrophoresis* **26** 4237–4242.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparamteric Maximum Likelihood Estimation*. Birkhauser, Basel. MR1180321

Jensen, K., Søndergaard, I., Skovgaard, I. and Nielsen, H. (1995). From image processing to classification: I. Modelling disturbance of isoelectric focusing patterns. *Electrophoresis* **16** 921–926.

Kneip, A. and Engel, J. (1995). Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics* **23** 551–570. MR1332581

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics* **20** 1266–1305. MR1186250

Lawton, W., Sylvestre, E. and Maggio, M. (1972). Self-modeling nonlinear regression. *Technometrics* **14** 513–532.

Liu, X. and Müller, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* **99** 687–699. MR2090903

Murphy, S. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics* **23** 182–198. MR1331663

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26** 183–214. MR1611788

Ramsay, J. (2007). R-package `fda`: Functional data analysis. R Development Core Team. Available at http://cran.r-project.org/.

Ramsay, J. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society. Series B* **60** 351–363. MR1616045

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer, New York. MR2168993

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York. MR1910407

Rønn, B. B. (1998). Analyses of functional data. Ph.D. thesis, Department of Mathematics and Physics, The Royal Veterinary and Agricultural University.

Rønn, B. B. (2001). Non-parametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society. Series B* **63** 243–259. MR1841413

Stone, C. J. (1980). Optimal rate of convergence for nonparametric estimators. *The Annals of Statistics* **8** 1348–1360. MR0594650

Stützle, W., Gasser, T., Molinari, L., Largo, R., Prader, A. and Huber, P. (1980). Shape-invariant modeling of human growth. *Annals of Human Biology* **5** 1–24.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions (Corrigenda **78** 233–234). *Biometrika* **76** 425–433. MR1040637

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics* **25** 1251–1276. MR1447750

Wium, H., Kristiansen, K. R. and Qvist, K. B. (1998). Proteolysis and its role in relation to texture of Feta cheese made from ultrafiltered milk with different amounts of rennet. *Journal of Dairy Research* **65** 665–674.

Biometrics, Genmab a/s
Copenhagen
Denmark
E-mail: brn@genmab.com

Department of Natural Sciences
Faculty of Life Sciences
University of Copenhagen
Denmark
E-mail: ims@life.ku.dk