# Semiparametric shape-invariant models for periodic data

Holger Hürtgen
University of Wisconsin–Milwaukee

Daniel Gervini[1]
University of Wisconsin–Milwaukee

October 7, 2008

[1]Corresponding author. Email: gervini@uwm.edu. Address: Department of Mathematical Sciences, P.O. Box 413, Milwaukee, WI 53201, USA.

**Abstract**

This article presents a novel shape-invariant modeling approach to quasi-periodic data. We propose a dynamic semiparametric method that estimates the common cycle shape in a nonparametric way and the individual phase and amplitude variability in a parametric way. An efficient algorithm to compute the estimators is proposed. The behavior of the estimators is studied by simulation and by a real-data example.

*Key Words*: Circadian Rhythms; Nonparametric Regression; Spline Smoothing.

# 1 Introduction

Many natural phenomena occur in periodic or cyclical patterns. For example: in meteorology, temperature follows an annual cycle; in physiology, blood pressure follows a daily (circadian) cycle. Many other examples could be given (see e.g. Refinetti, 2005). These cyclical patterns, however, are never perfectly regular; there is always some variability in amplitude and length of the cycles. The daily temperature in Melbourne, for example (Fig. 1), shows an annual cyclical pattern with regular cycle lengths but very variable amplitudes.

There are different approaches to the statistical analysis of periodic data. The classical time series approach is based on Fourier expansions (see e.g. chapter 12 of Huettel et al., 2004). A nonparametric method is proposed in Ramsay and Silverman (2005): the original series is split into individual cycles and each cycle is treated as an independent curve. A more efficient method is semiparametric shape-invariant modeling (Lawton and Sylvestre, 1971; Lawton et al., 1972; Lindstrom, 1995; Wang and Brown, 1996).

Shape-invariant modeling has a number of advantages. One is the higher efficiency obtained by assuming a common shape function, which is then estimated "borrowing strength" from the whole dataset, rather than using only the data for each cycle. The departure of each cycle from the common shape is modeled in a parametric way. These parameters are usually interpretable, for example as shift or scale parameters, which is another advantage of shape-invariant modeling. Interpretability is important because the parameters can sometimes be modeled as functions of explanatory variables. For example, Wang and Brown (1996) were able to study the association between ACT hormone level and core body temperature in this way. The classical time series approach, in contrast, may provide a good fit to the data, but the model parameters (the Fourier coefficients) are hardly interpretable and difficult to relate to explanatory variables. The classical time-series approach, however, does have an advantage: it treats the whole sequence of observations as a single series. The other methods treat each cycle as an independent curve, which may be appropriate if a single cycle is observed per individual

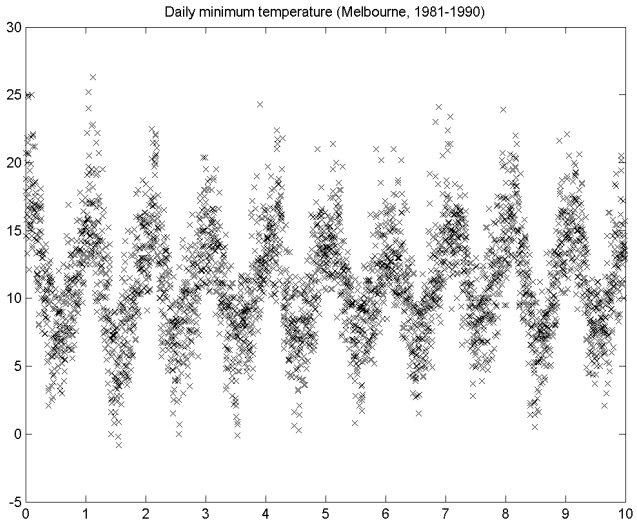


Figure 1: Daily minimum temperature (Melbourne, 1981-1990)

(as in Wang and Brown, 1996), but it is not appropriate for time series such as the temperature data in Fig. 1.

The presence of phase variability complicates both the classical time-series approach and the semiparametric shape-invariant modeling. The effect of phase variability on the classical time-series approach is to inflate the number of terms in the Fourier expansion. The shape-invariant model of Wang and Brown (1996), while including a time-shift parameter, is not adequate for consecutive cycles of varying length; more complex models are needed.

We think it is possible to retain the simplicity and interpretability of shape-invariant models while addressing the problem of phase variability. What we propose in this article is to model the periodic series as a sequence of contiguous cycles of varying length and amplitude but common shape. The shape function is estimated nonparametrically, while amplitude and phase variability are accounted for by interpretable parameters. These parameters can subsequently be used, for example, to study the relationships between the dynamics of the process and the

covariates. The models we propose are introduced in Section 2, studied by simulation in Sections 3 and applied to real data in Section 4. Technical material is left to the Appendix.

# 2 Periodic shape-invariant models

Suppose we have $n$ observations $\{y_i\}$ of a random process measured at certain points $\{t_i\}$ in a time interval $[a, b]$, such that

$$y_i = f(t_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $f$ is a (quasi) periodic function and $\{\varepsilon_i\}$ are random errors. We assume $c$ cycles of $f$ occur in $[a, b]$, with each cycle varying in amplitude and length but not in shape.

Specifically, we propose the following model:

$$f(t) = \mu + \sum_{j=1}^{c} \alpha_j g\left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j}\right), \tag{2}$$

where $\mu$ is the overall mean, $\{\alpha_j\}$ are positive amplitude parameters, $\{\tau_j\}$ are the cycle endpoints (with $\tau_1 = a$ and $\tau_{c+1} = b$), and $g : \mathbb{R} \to \mathbb{R}$ is the common shape function (with $g(x) = 0$ for $x \notin [0, 1]$). For identifiability, we assume $\bar{\alpha} = 1$ and $g(0) = g(1) = 0$. The typical regression function $f(t)$ under model (2) will look as in Fig. 2.

Model (2) is identifiable if the shape function $g$ has a finite number of zeros (this is proved in the Appendix). It is not clear if this assumption is really necessary or just technical; in any case, most shape functions satisfy this condition in practice. The number of cycles is assumed to be fixed and known (it is easy to see that if $c$ were allowed to vary, model (2) would not be identifiable).

The shape function $g$ will be modeled as a piecewise polynomial or spline function (Eubank, 1999). Splines are flexible enough to fit a large number of

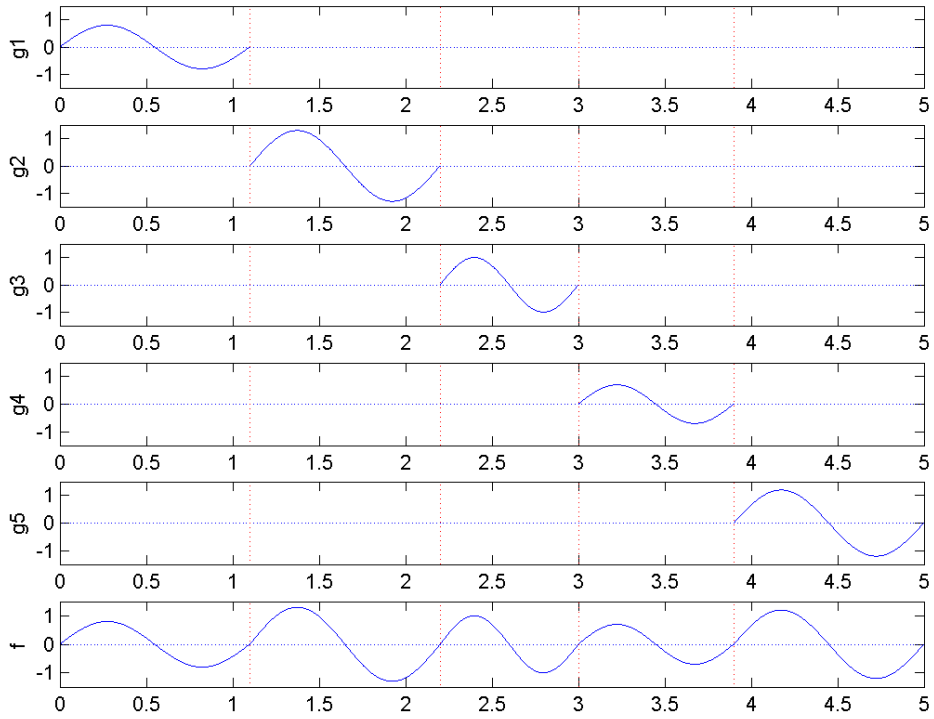


Figure 2: Regression function $f$

cycle shapes, while retaining the computational simplicity of polynomial regression. Given a knot sequence $\xi_1, ..., \xi_p$ in $(0, 1)$, a spline function of order $r$ is a linear combination of the basis functions $\{1, x, ..., x^{r-1}, (x - \xi_1)_+^{r-1}, ..., (x - \xi_p)_+^{r-1}\}$. For numerical stability, however, we use the equivalent B-spline basis (see Eubank, 1999, for the definition). This basis is denoted by $b_1(x), ..., b_m(x)$, where $m = p + r$ and we set $b_j(x) = 0$ for $x \notin [0, 1]$. Then

$$g(x) = \sum_{k=1}^{m} \beta_k b_k(x). \tag{3}$$

## 2.1 Model estimation

The parameters of model (1)–(3) can be estimated by least squares. But in order to obtain a smooth estimate of the shape function $g$, it is necessary to add a roughness penalty to the least squares criterion. This gives us the flexibility to use a large number of knots, since the regularity of $\hat{g}$ will be essentially determined by the penalization parameter (Ruppert, 2002).

The objective function is then

$$\begin{aligned} S(\mu, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\beta}, \lambda, \mathbf{l}) \;=\; & \\ & \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \mu - \sum_{j=1}^{c} \alpha_j g\left( \frac{t_i - \tau_j}{\tau_{j+1} - \tau_j} \right) \right\}^2 + \\ & \lambda \int_0^1 \{g''(x)\}^2 dx + l_1 (\sum_{j=1}^{c} \alpha_j - c) + l_2 g(0) + l_3 g(1). \end{aligned} \tag{4}$$

Note that the identifiability restriction $\bar{\alpha} = 1$ and the conditions $g(0) = g(1) = 0$ were introduced in (4) via Lagrange multipliers. The roughness penalty term has a more explicitly expression,

$$\lambda \int_0^1 \{g''(x)\}^2 dx = \lambda \boldsymbol{\beta}^T \Omega \boldsymbol{\beta},$$

where $\Omega_{ij} = \int_0^1 b_i''(x) b_j''(x) dx$. Also, if $\mathbf{b}(x)$ denotes the vector of B-spline basis

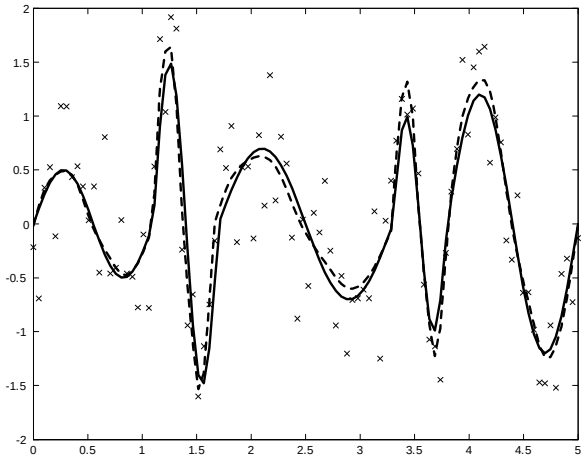


Figure 3: Simulated sample with true regression function (solid line) and estimated regression function (dashed line).

functions, $g(0) = \mathbf{b}(0)^T\boldsymbol{\beta}$ and $g(1) = \mathbf{b}(1)^T\boldsymbol{\beta}$.

Model (2) is nonlinear in the parameters $\tau$ and then the objective function $S(\mu, \alpha, \tau, \beta, \lambda, l)$ is nonconvex. But the following alternating algorithm will converge to a local minimum for a given starting point: *(i)* fix $g$ (i.e. $\beta$) and $\tau$, and minimize (4) with respect to $\mu$ and $\alpha$; *(ii)* fix $g$, $\mu$ and $\alpha$, and minimize (4) with respect to $\tau$; *(iii)* fix $\mu$, $\alpha$ and $\tau$, and minimize (4) with respect to $g$ (i.e. with respect to $\beta$); iterate *(i)–(iii)* until convergence. The penalization parameter $\lambda$ is kept fixed throughout the iterations; the optimal $\lambda$ is then found by cross-validation. See details of implementation in the Appendix. As initial values of $\tau$ we use equispaced points in $[a, b]$, but occasionally this can lead to a local suboptimal minimizer, especially when the cycle lenghts are very irregular; in such cases, it is recommendable to try alternative (for intance, random) starting points.

# 3 Simulations

In this section we study by simulation the behavior of our estimators for small and moderate sample sizes. We generated data from model (2) plus independent $N(0, \sigma^2)$ errors with $\sigma = 0.5$ and $\sigma = 1$. As shape function we took $g(x) = \sin(2\pi x)$ and considered $c = 5$ cycles. The model parameters were $\mu = 0$, $\boldsymbol{\alpha} = (0.5, 1.5, 0.7, 1, 1.2)^T$, and $\boldsymbol{\tau} = (0, 1.1, 1.7, 3.3, 3.8, 5)^T$. The input points $\{t_i\}$ were equidistant in $[0, 5]$. We simulated three different sample sizes: 100, 250 and 500, giving an average of 20, 50 and 100 observations per cycle, respectively. Each model was replicated 500 times. As estimator of $g$ we used the B-spline model (3) with 10 equidistant knots and order 4 (cubic splines). The penalization parameter $\lambda$ was chosen by generalized cross-validation.

A typical sample, with $n = 100$ and $\sigma = 0.5$, is depicted in Fig. 3 together with the true and the estimated regression functions. We see that the fit is very good, considering the small number of observations per cycle (20 on average) and the low signal-to-noise ratio. Any method that attempted to fit each cycle independently (like Ramsay and Silverman, 2005) would not be able to simultaneously attain a smooth fit and a low bias, since it would have to use only the 20 or so observations available for each cycle. Our method, on the other hand, can "borrow strength" by using all 100 observations to estimate the underlying shape function. Therefore, it can simultaneously attain lower bias and variance than alternative methods.

To assess the consistency of the estimators as $n$ increases, we used the mean absolute error as criteria for $\hat{\mu}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\tau}}$, and $E\{\|\hat{\mathbf{f}} - \mathbf{f}\|^2/n\}^{1/2}$ where $\mathbf{f} = (f(t_1), \ldots, f(t_n))^T$ for the estimated regression function. The simulated errors are given in Table 1. We see that, as expected, the estimation errors decrease as the sample size increases. Although $\mathrm{MAE}(\hat{f})$ will not converge to zero because of the asymptotic bias (due to the fact that $g$ is not a spline function), it is clear from Table 1 that $\mathrm{MAE}(\hat{f})$ decreases very fast as $n$ goes to infinity.

| | | $n$ | | |
|---|---|---|---|---|
| $\sigma$ | MAE | 100 | 250 | 500 |
| 0.5 | $\hat{\mu}$ | .082 (.003) | .049 (.002) | .036 (.001) |
| | $\hat{\boldsymbol{\alpha}}$ | .357 (.005)† | .242 (.004) | .170 (.003) |
| | $\hat{\boldsymbol{\tau}}$ | .071 (.002) | .038 (.0007) | .026 (.0005) |
| | $\hat{f}$ | .203 (.002) | .123 (.001) | .085 (.0008) |
| 1.0 | $\hat{\mu}$ | .187 (.007) | .102 (.004) | .064 (.002) |
| | $\hat{\boldsymbol{\alpha}}$ | 2.155 (.126)† | .555 (.016)† | .352 (.014) |
| | $\hat{\boldsymbol{\tau}}$ | .258 (.007)† | .098 (.003)† | .061 (.002) |
| | $\hat{f}$ | .493 (.006) | .262 (.003) | .174 (.002) |

Table 1: Simulated mean absolute errors (Monte Carlo standard errors in parenthesis).
†Due to outliers, 10% trimmed mean and standard error are given.

# 4 Example

As an illustration of the applicability of our method, we analyzed the series of minimum daily temperatures in Melbourne, Australia, for the ten-year period between 1981 and 1990 (Fig. 1). The series starts on January 1st, which is rather inconvenient for our model (that works best when $\mu$ is close to $\bar{y}$). So we trimmed the first 100 observations of the series, and consequently the last 265 observations, and ended up with a nine-year series (alternatively, we could have taken $\tau_1$ and $\tau_{c+1}$ as estimable parameters in model (2), rather than fixed endpoints; but those observations outside $[\hat{\tau}_1, \hat{\tau}_{c+1}]$ would have been discarded in any case).

We fitted a cubic-spline model with 20 equispaced knots for the shape function $g$. The resulting regression function $\hat{f}$ is shown in Fig. 4. Overall, the fit is good considering the simplicity of our model. The estimates of the parameters are: $\hat{\mu} = 11.29$, $\hat{\boldsymbol{\alpha}} = (1.03, 1.23, 0.93, 0.97, 0.96, 0.78, 0.99, 0.96, 1.14)^T$, and $\hat{\boldsymbol{\tau}} = (0, 1.02, 2.01, 2.98, 4.04, 5.01, 6.02, 7.02, 8.02, 9)^T$. We see that the amplitude parameters $\hat{\alpha}_j$ show considerable deviations from the mean $\bar{\alpha} = 1$, indicating that there are mild but noticeable fluctuations in annual temperature range from year

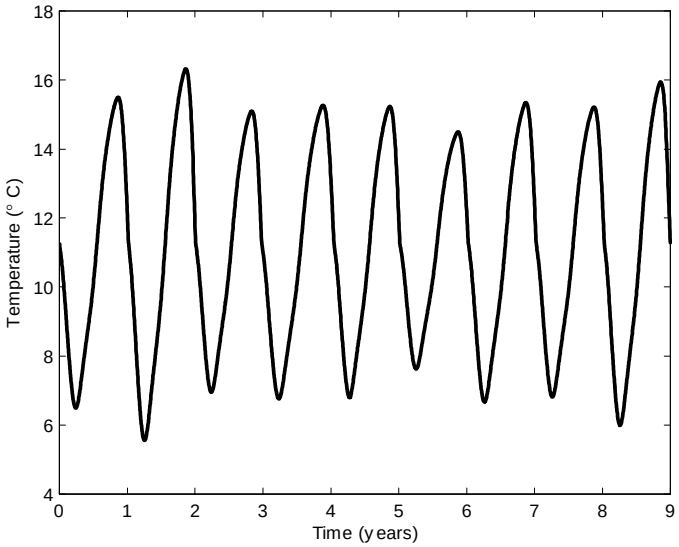


Figure 4: Daily minimum temperature in Melbourne: shape-invariant fit for the period April 1981 – April 1989

to year. In contrast, there is little variation in the duration of the annual cycles, a few days at most (note that 0.02 in the scale of the $\tau$s represents approximately 7 days).

Even though model (2) fits the data reasonably well, it has certain limitations. Figure 5 shows a scatter plot of residuals versus fitted values for these data. We see a mild "bow tie" shape, with smaller residual variance for predicted values near $\hat{\mu}$ and larger variances at both extremes. This is partly due to the tendency of all smoothing techniques to underestimate the peaks and overestimate the valleys, but it also suggests that a better fit may be provided by a model that accounts for independent amplitude variability in summer and winter. The general form of such model would be

$$f(t) = \mu + \sum_{j=1}^{c} g_0\left(\frac{t-\tau_j}{\tau_{j+1}-\tau_j}\right) + \sum_{\ell=1}^{q}\sum_{j=1}^{c} \alpha_{j\ell}\cdot g_\ell\left(\frac{t-\tau_j}{\tau_{j+1}-\tau_j}\right),$$

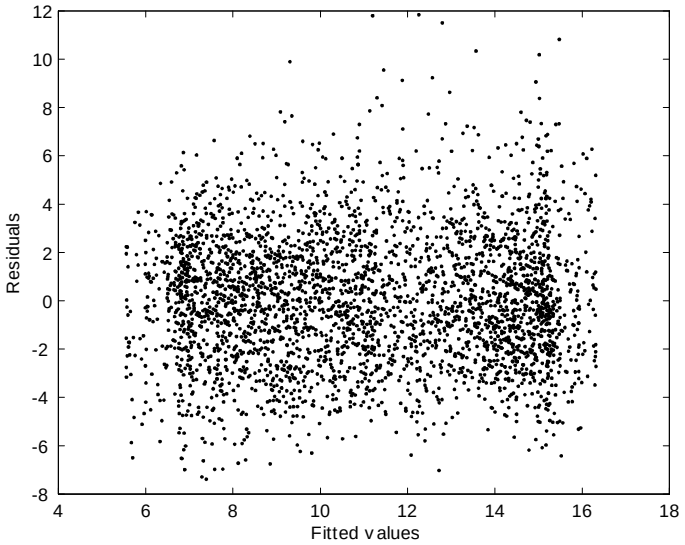


Figure 5: Daily minimum temperature in Melbourne: residuals-vs-fitted-values plot.

which accommodates $q$ independent forms of amplitude variability for each cycle. But this is a topic for future research.

Other possible extensions of model (2) include a nonconstant trend, such as $\mu(t) = \alpha + \beta t$ or some other parametric model; this may be useful for fitting economic data. And even though we have made no assumptions about the correlation structure of the random errors, the penalized least squares criterion (4) is best adapted to i.i.d. errors; in presence of dependence or heteroskedasticity, alternative estimators like weighted least squares or maximum likelihood are worth considering.

# A Appendix

## A.1 Proof of identifiability of model (2)

Suppose there were two sets of parameters satisfying (2); that is,

$$\mu + \sum_{j=1}^{c} \alpha_j g\left(\frac{t-\tau_j}{\tau_{j+1}-\tau_j}\right) = \tilde{\mu} + \sum_{j=1}^{c} \tilde{\alpha}_j \tilde{g}\left(\frac{t-\tilde{\tau}_j}{\tilde{\tau}_{j+1}-\tilde{\tau}_j}\right), \text{ for all } t \in [a,b].$$

Since $f(a) = \mu = \tilde{\mu}$, it is clear that the parameter $\mu$ is identifiable.

Now let $\mathcal{Z}(g)$ be the set of zeros of the function $g$, excluding $x = 1$. By hypothesis $\mathcal{Z}(g)$ is finite and contains at least one point, $x = 0$; let $N(g)$ be the number of elements in $\mathcal{Z}(g)$. It is easy to see that, as a result, the set $\{t \in [a,b] : f(t) = \mu\}$ is also finite and has exactly $N(g)c + 1$ elements. This implies that $N(g) = N(\tilde{g})$.

Now suppose $\tau_2 < \tilde{\tau}_2$ (we already know that $\tau_1 = \tilde{\tau}_1 = a$ by hypothesis). If $\mathcal{Z}(g) = \{x_1, \ldots, x_{N(g)}\}$, on the one hand we have

$$f\{(\tau_2 - a)x_i + a\} = \mu + \alpha_1 g(x_i) = \mu$$

and on the other hand

$$f\{(\tau_2 - a)x_i + a\} = \tilde{\mu} + \tilde{\alpha}_1 \tilde{g}\left\{\left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right) x_i\right\}.$$

This implies

$$\tilde{g}\left\{\left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right) x_i\right\} = 0 \text{ for all } i = 1, \ldots, N(g),$$

and then

$$\left\{\left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right) x_1, \ldots, \left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right) x_{N(g)}\right\} \subseteq \mathcal{Z}(\tilde{g}).$$

But

$$f(\tau_2) = \mu = \tilde{\mu} + \tilde{\alpha}_1 \tilde{g}\left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right),$$

so

$$\tilde{g}\left(\frac{\tau_2 - a}{\tilde{\tau}_2 - a}\right) = 0$$

and then, since $(\tau_2 - a)/(\tilde{\tau}_2 - a) < 1$, the set $\mathcal{Z}(\tilde{g})$ would contain at least $N(g) + 1$ points, which is impossible. Therefore, $\tau_2$ cannot be strictly less than $\tilde{\tau}_2$. With a similar argument, we can also rule out the possibility that $\tilde{\tau}_2 < \tau_2$, hence $\tau_2 = \tilde{\tau}_2$. Proceeding in this way, it is possible to show that $\tau_j = \tilde{\tau}_j$ for all $j$.

Since we have already proved that $\mu = \tilde{\mu}$ and $\boldsymbol{\tau} = \tilde{\boldsymbol{\tau}}$, it follows that

$$\alpha_j g\left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j}\right) = \tilde{\alpha}_j \tilde{g}\left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j}\right), \text{ for all } t \in [\tau_j, \tau_{j+1}],\ j = 1, \ldots, c. \tag{5}$$

Then

$$\begin{aligned}
\alpha_j \int_{\tau_j}^{\tau_{j+1}} \left| g\left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j}\right)\right| \, dt &= \alpha_j(\tau_{j+1} - \tau_j) \int_0^1 |g(x)| \, dx \\
&= \tilde{\alpha}_j \int_{\tau_j}^{\tau_{j+1}} \left|\tilde{g}\left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j}\right)\right| \, dt \\
&= \tilde{\alpha}_j(\tau_{j+1} - \tau_j) \int_0^1 |\tilde{g}(x)| \, dx,
\end{aligned}$$

so

$$\alpha_j = \tilde{\alpha}_j(\textstyle\int |g| / \int |\tilde{g}|), \text{ for } j = 1, \ldots, c.$$

Since $\sum_{j=1}^c \alpha_j = \sum_{j=1}^c \tilde{\alpha}_j = c$ by assumption, it follows that $\alpha_j = \tilde{\alpha}_j$ for all $j$ and then, from (5), we also have $g(x) = \tilde{g}(x)$ for all $x \in [0, 1]$, which completes the proof.

## A.2 Estimation algorithm

First, note that the observational model for $\mathbf{y} = (y_1, ..., y_n)$ can be expressed in matrix form either as

$$\mathbf{y} = \mu \mathbf{1}_n + G\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

with $G = G(\boldsymbol{\tau}, \boldsymbol{\beta})$ given by

$$G_{ij} = g\left(\frac{t_i - \tau_j}{\tau_{j+1} - \tau_j}\right), \tag{6}$$

or as

$$\mathbf{y} = \mu \mathbf{1}_n + C\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{7}$$

with $C = C(\boldsymbol{\alpha}, \boldsymbol{\tau})$ given by

$$C_{ik} = \sum_{j=1}^{c} \alpha_j b_k\left(\frac{t_i - \tau_j}{\tau_{j+1} - \tau_j}\right). \tag{8}$$

The estimation algorithm we propose, for each $\lambda$, is the following:

1. *Initialization*. Take $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}_m$, $\hat{\mu}^{(0)} = \bar{y}$, $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{1}_c$ and $\hat{\tau}_j^{(0)} = a + (b-a)(j-1)/c$, for $j = 1, \ldots, c+1$.

2. *Iterations*. Given $\hat{\boldsymbol{\beta}}^{(h)}$, $\hat{\mu}^{(h)}$, $\hat{\boldsymbol{\alpha}}^{(h)}$ and $\hat{\boldsymbol{\tau}}^{(h)}$:

   (a) Compute $G$ as in (6) and define updates $\hat{\mu}^{(h+1)}$ as in (9) and $\hat{\boldsymbol{\alpha}}^{(h+1)}$ as in (10).

   (b) Compute $\nabla F(\hat{\boldsymbol{\tau}}^{(h)})$ and $H(\hat{\boldsymbol{\tau}}^{(h)})$ using $\hat{\mu}^{(h+1)}$ and $\hat{\boldsymbol{\alpha}}^{(h+1)}$, and define update $\hat{\boldsymbol{\tau}}^{(h+1)}$ as in (12).

   (c) Compute $C$ as in (8) using $\hat{\mu}^{(h+1)}$, $\hat{\boldsymbol{\alpha}}^{(h+1)}$ and $\hat{\boldsymbol{\tau}}^{(h+1)}$, and define the update $\hat{\boldsymbol{\beta}}^{(h+1)}$ as in (14).

   (d) Repeat (a)–(c) until convergence.

Each step is explained next. The choice of smoothing parameter is also explained below.

### A.2.1 Step a

If the shape function $g$ and the shift parameters $\boldsymbol{\tau}$ are fixed, minimizing (4) with respect to $\mu$ and $\boldsymbol{\alpha}$ is equivalent to minimizing

$$\|\mathbf{y} - \mu\mathbf{1}_n - G\boldsymbol{\alpha}\|^2 + l_1\left(\mathbf{1}_c^T\boldsymbol{\alpha} - c\right),$$

with $G$ as in (6). This is a simple quadratic minimization problem in $\mu$ and $\boldsymbol{\alpha}$. It is easy to see that the minimizers satisfy the equations

$$\hat{\mu} = \frac{1}{n}\mathbf{1}_n^T(\mathbf{y} - G\hat{\boldsymbol{\alpha}}), \tag{9}$$

$$\hat{\boldsymbol{\alpha}} = (G^TG)^{-1}\{G^T(\mathbf{y} - \hat{\mu}\mathbf{1}_n) - \frac{\hat{l}_1}{2}\mathbf{1}_c\}, \tag{10}$$

with

$$\frac{\hat{l}_1}{2} = \frac{\mathbf{1}_c^T(G^TG)^{-1}G^T(\mathbf{y} - \hat{\mu}\mathbf{1}_n) - c}{\mathbf{1}_c^T(G^TG)^{-1}\mathbf{1}_c}.$$

### A.2.2 Step b

If the shape function $g$ and the parameters $\mu$ and $\boldsymbol{\alpha}$ are fixed, minimizing (4) with respect to $\boldsymbol{\tau}$ is equivalent to minimizing

$$F(\boldsymbol{\tau}) := \sum_{i=1}^{n} r_i^2(\boldsymbol{\tau}) \tag{11}$$

with

$$r_i(\boldsymbol{\tau}) = y_i - \mu - \sum_{j=1}^{c} \alpha_j g\left(\frac{t_i - \tau_j}{\tau_{j+1} - \tau_j}\right).$$

This is now a non-linear least squares problem, that can be iteratively solved using, for instance, a Fisher-scoring algorithm. Note that

$$\begin{aligned}\nabla F(\boldsymbol{\tau}) &= 2\sum_{i=1}^{n} r_i(\boldsymbol{\tau})\,\nabla r_i(\boldsymbol{\tau}),\\ \nabla^2 F(\boldsymbol{\tau}) &= 2\sum_{i=1}^{n} \nabla r_i(\boldsymbol{\tau})\{\nabla r_i(\boldsymbol{\tau})\}^T + 2\sum_{i=1}^{n} r_i(\boldsymbol{\tau})\,\nabla^2 r_i(\boldsymbol{\tau}),\end{aligned}$$

with

$$\frac{\partial r_i(\boldsymbol{\tau})}{\partial \tau_j} = \alpha_j g'\left(\frac{t_i-\tau_j}{\tau_{j+1}-\tau_j}\right)\frac{(t_i-\tau_{j+1})}{(\tau_{j+1}-\tau_j)^2} + \alpha_{j-1} g'\left(\frac{t_i-\tau_{j-1}}{\tau_j-\tau_{j-1}}\right)\frac{(t_i-\tau_{j-1})}{(\tau_j-\tau_{j-1})^2}$$

for $j = 2,\ldots,c$ (remember that $\tau_1 = a$ and $\tau_{c+1} = b$ are fixed). The Fisher-scoring algorithm approximates $\nabla^2 F(\boldsymbol{\tau})$ with

$$H(\boldsymbol{\tau}) := 2\sum_{i=1}^{n} \nabla r_i(\boldsymbol{\tau})\{\nabla r_i(\boldsymbol{\tau})\}^T,$$

avoiding the use of second derivatives. The minimizer of (11) is found in an iterative way: given a current estimate $\hat{\boldsymbol{\tau}}^{(h)}$, an update is defined as

$$\hat{\boldsymbol{\tau}}^{(h+1)} := \hat{\boldsymbol{\tau}}^{(h)} + H(\hat{\boldsymbol{\tau}}^{(h)})^{-1}\nabla F(\hat{\boldsymbol{\tau}}^{(h)}), \tag{12}$$

and the procedure is iterated until convergence. In practice it is not necessary to fully iterate (12) for each current value of $\hat{g}$, $\hat{\mu}$ and $\hat{\boldsymbol{\alpha}}$; it is sufficient to perform a one-step update.

### A.2.3 Step c

To estimate $g$ when $\mu$, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ are fixed, note that minimizing (4) with respect to $\boldsymbol{\beta}$ is equivalent to minimizing

$$\|\mathbf{y} - \mu\mathbf{1}_n - C\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\Omega\boldsymbol{\beta} + l_2\mathbf{b}(0)^T\boldsymbol{\beta} + l_3\mathbf{b}(1)^T\boldsymbol{\beta},$$

with $C$ given by (8). This is again a simple quadratic minimization problem, whose solution satisfies

$$A \begin{pmatrix} \boldsymbol{\beta} \\ \frac{1}{2}l_2 \\ \frac{1}{2}l_3 \end{pmatrix} = \begin{pmatrix} C^T(\mathbf{y} - \mu\mathbf{1}_n) \\ 0 \\ 0 \end{pmatrix}$$

with

$$A = \begin{pmatrix} C^TC + \lambda\Omega & \mathbf{b}(0) & \mathbf{b}(1) \\ \mathbf{b}(0)^T & 0 & 0 \\ \mathbf{b}(1)^T & 0 & 0 \end{pmatrix}. \tag{13}$$

Then

$$\hat{\boldsymbol{\beta}} = (A^{-1})_{11}C^T(\mathbf{y} - \mu\mathbf{1}_n), \tag{14}$$

where $(A^{-1})_{11}$ is the $m \times m$ submatrix on the top left corner of $A^{-1}$.

### A.2.4 Choosing the smoothing parameter

The smoothing parameter $\lambda$ that controls the trade-off between goodness-of-fit and roughness of $\hat{g}$ can be chosen by cross-validation (CV) or generalized cross-validation (GCV). Cross-validation minimizes the leave-one-out sum of squares of the residuals,

$$CV := \sum_{i=1}^{n}(y_i - \hat{y}_{[i]})^2,$$

where $\hat{y}_{[i]}$ is the predicted value of $y_i$ obtained from the reduced sample $\{(t_i, y_i)\}_{i \neq n}$. It is known (see e.g. Hastie, Tibshirani and Friedman, 2001) that if the fitted values can be expressed as $\hat{\mathbf{y}} = S_\lambda\mathbf{y}$ for some $n \times n$ matrix $S_\lambda$ (known as the smoothing matrix) then the cross-validation criterion can be expressed as

$$CV = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\{1 - (S_\lambda)_{ii}\}^2}. \tag{15}$$

A more stable criterion, called generalized cross-validation (GCV), is obtained by replacing $(S_\lambda)_{ii}$ in (15) by its average value:

$$GCV := \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\{1 - \mathrm{tr}(S_\lambda)/n\}^2}. \tag{16}$$

Here $\mathrm{tr}(S_\lambda)$ can be interpreted as the effective dimension of the model, so (16) penalizes model complexity as well as lack of fit.

We can define an analogous of (16) for our estimators, although the derivation is less straightforward due to the non-linearity of the model. If $\mu$, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ were fixed, it would follow from (7) and (14) that

$$\hat{\mathbf{y}}_0 = C(A^{-1})_{11}C^T\mathbf{y}_0,$$

where $\mathbf{y}_0 = \mathbf{y} - \mu\mathbf{1}_n$, and then $C(A^{-1})_{11}C^T$ would play the role of smoothing matrix. Then we define

$$S_\lambda := \hat{C}(\hat{A}^{-1})_{11}\hat{C}^T,$$

where $\hat{A}$ is as in (13) and $\hat{C}$ as in (8) (computed with $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\tau}}$). But since $S_\lambda$ was derived under the assumption that $\mu$, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ were fixed, to properly account for all the model parameters we add $1 + 2(c-1)$ to $\mathrm{tr}(S_\lambda)$ in the denominator of (16). Then the criterion we minimize to obtain the optimal $\lambda$ is

$$GCV = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{[1 - \{\mathrm{tr}(S_\lambda) + 2c - 1\}/n]^2}. \tag{17}$$

Typically, the minimization is carried out on a logarithmically-spaced grid of possible values of $\lambda$. The estimator $\hat{g}$ changes little as $\lambda$ varies in a neighborhood of the optimal value, so in practice it is not necessary to find the exact minimizer of (17).

# References


Eubank, R. (1999). *Nonparametric Smoothing and Spline Regression*. Marcel Dekker, New York.

Hastie, T., Tibshirani R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Huettel, S. A., Song, A. W. and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sinauer Associates, Sunderland.

Lawton, W. H. and Sylvestre, E. A. (1971). Self-modeling curve resolution. *Technometrics*, 12, 617–633.

Lawton, W. H., Sylvestre, E. A. and Maggio, M. S. (1972). Self-modeling linear regression. *Technometrics*, 13, 513–532.

Lindstrom, M. J. (1995). Self modeling with random shift and scale Parameters and free-knot spline shape function. *Statistics in Medicine*, 14, 2009–2021.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis (Second Edition)*. Springer, New York.

Refinetti, R. (2005). *Circadian Physiology: Second Edition*. CRC Press, Boca Raton.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statististics*, 11, 735–757.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics*, 52, 588–596.