Statistical Methodology for "Reproduction is Adapted to Survival Characteristics across Geographically Isolated Medfly Populations"

Hans-Georg Müller, Shuang Wu, Alexandros D. Diamantidis, Nikos T. Papadopoulos, James R. Carey

1. From lifetables to hazard rates

The survival schedule for each population is assessed by means of a lifetable, counting the number of deaths occurring daily among the flies in the cohort from a specific population. These lifetable data correspond to pairs (N_{j-1}, d_j), where N_{j-1} is the size of the cohort alive at the beginning of the jth day, N_0 is the size of the cohort, and d_j is the number of flies that die during the j-th day. The predictor trajectories for fertility are the hazard functions, also known as force of mortality and defined as $X(t) = \lim_{\Delta \to 0} Pr(T \in [t, t+\Delta] | T > t)/\Delta$. Here T is the survival time of an individual fly.

When estimating hazard rates from lifetable data, a classical estimate is the central death rate $\tilde{q}_c(t_j) = d_j/\{(N_{j-1} + N_j)\Delta/2\}$. However, one needs to contend with discretization bias since the lifetable data are aggregated while the hazard rate is a limit. Wang and Capra (1998) and Müller et al. (1997) proposed a transformation approach to address this bias by means of a transformation ψ of the central death rate $\tilde{q}_c$ such that $\psi(\tilde{q}_c(t)) = \int_{t-\Delta/2}^{t+\Delta/2} X(t)\, dt$. This leads to a reasonably close approximation of the underlying hazard rate $X(t)$ and yields

$$\psi(x) = \frac{1}{\Delta} \log \frac{2 + x\Delta}{2 - x\Delta}.$$

The predictor trajectories are sampled on a dense and regular grid of days t_j since eclosion,

$$X(t_j) = \log\{\psi(\tilde{q}_c(t_j))\}, \tag{1}$$

viewed as potentially noise-contaminated measurements of the predictor processes X.

2. Functional principal component analysis for reproductive schedules

The combined samples of survival-adjusted, respectively unadjusted, reproductive schedules are assumed to be independent realizations of an underlying random process Y, which is a square integrable smooth function defined in a bounded time interval $\mathcal{T}$, with mean function $\mathrm{E}Y(t) = \mu(t)$ and covariance function $\mathrm{cov}(Y(s), Y(t)) = G(s,t)$. Under very mild conditions, we may assume that the covariance function G has the following orthogonal expansion

$$G(s,t) = \sum_{k=1}^{\infty} \tau_k \psi_k(s) \psi_k(t), \tag{2}$$

and then the Karhunen-Loève representation (Ash and Gardner, 1975) of Y is

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} \zeta_k \psi_k(t), \tag{3}$$

where ψ_k are sequences of orthonormal eigenfunctions with non-increasing eigenvalues τ_k, $\sum_k \tau_k < \infty$. The coefficients ζ_k are referred to as functional principal components, which are uncorrelated random variables with zero means and variances $\mathrm{var}(\zeta_k) = \tau_k$. Since the functional principal components carry all the random variation present in the random trajectories Y, we can use the sequence of components $\{\zeta_1, \zeta_2, ...\}$ as proxy of Y.

We apply the functional principal component analysis through conditional expectation, proposed by Yao et al. (2005), to estimate the functional principal components for the ith fly

$$\hat{\zeta}_{ik} = \widehat{\mathrm{E}}(\zeta_{ik} | \widetilde{\boldsymbol{Y}}_i) = \hat{\tau}_k \hat{\boldsymbol{\psi}}_{ik}^T \widehat{\boldsymbol{\Sigma}}_i (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i), \tag{4}$$

where $\widetilde{\boldsymbol{Y}}_i = (Y_{i1}, ..., Y_{iN_i})^T$ is the vector of observations (egg-laying or transformed hazard rates), observed at time points $T_{i1}, ..., T_{iN_i}$; $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(T_{i1}), ..., \hat{\mu}(T_{iN_i}))^T$ and $\widehat{\boldsymbol{\Sigma}}_i = \text{cov}(\widetilde{\boldsymbol{Y}}_i, \widetilde{\boldsymbol{Y}}_i)$ are the mean and covariance estimates, respectively, both evaluated at the observed time points; $\hat{\tau}_k$ and $\hat{\boldsymbol{\psi}}_{ik} = (\hat{\psi}_k(T_{i1}), ..., \hat{\psi}_k(T_{iN_i}))^T$ are the estimated eigenvalues and eigenfunctions, obtained through spectral decomposition of discretized covariance surface estimation (Rice and Silverman, 1991; Capra and Müller, 1997). We refer to Yao et al. (2005) for more details about the estimation procedures related to model (3).

Since the functional principal components $\{\zeta_1, \zeta_2, ...\}$ form a countable but infinite sequence, we need to choose a number K such that $\{\zeta_1, ..., \zeta_K\}$ provide a reasonable approximation of the reproductive process Y, i.e.,

$$\hat{Y}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\zeta}_{ik}\hat{\psi}_k(t). \tag{5}$$

We select the number of principal components K by the fraction of variation explained (FVE) criterion, as described in Liu and Müller (2008). First the model (3) is fitted with a relatively large number of principal components, say M. The fraction of variation explained by the truncated expansion (5) is estimated as

$$V(K) = \frac{\sum_{k=1}^{K} \hat{\tau}_k}{\sum_{k=1}^{M} \hat{\tau}_k} \tag{6}$$

and the number K is selected such that the proportion of variation explained exceeds a preselected threshold; we choose this to be 0.8 or 80%. For alternative techniques to select the number of principal components, we refer to Yao et al. (2005).

3. Functional regression analysis for biodemography

The functional linear regression model relates square integrable random functions $X(t)$, $Y(t)$

with each other, where $t \in \mathcal{T}$ for a suitable interval $\mathcal{T}$. It is given by

$$\mathrm{E}(Y(t)|X) = \mu_{_Y}(t) + \int \beta(s,t)(X(s) - \mu_{_X}(s))\, ds, \tag{7}$$

where the bivariate regression coefficient function $\beta(s,t)$ is smooth and square integrable. In the present application of the model, there are only six different predictor levels. For each predictor level we have about 50 independent responses, as the flies collected from the same biotype have the same predictor function. The functional linear regression model (7) is still valid since we have independent response functions for each fly.

The predictor X and response Y are assumed to have Karhunen-Loève expansions:

$$X(s) = \mu_{_X}(s) + \sum_{j=1}^{\infty} \xi_j \phi_j(s), \tag{8}$$

$$Y(t) = \mu_{_Y}(t) + \sum_{k=1}^{\infty} \zeta_k \psi_k(t). \tag{9}$$

Under regularity conditions, the regression coefficient surface β then has the following basis representation

$$\beta(s,t) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\mathrm{E}(\xi_j \zeta_k)}{\mathrm{E}(\xi_j^2)} \phi_j(s)\psi_k(t) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \beta_{jk}\phi_j(s)\psi_k(t). \tag{10}$$

Observing that the functional principal components are uncorrelated random variables and (8), (9) and (10), model (7) can be decomposed into a series of simple linear regressions of the functional principal components of response processes ζ_k on those of predictor processes ξ_j,

$$\mathrm{E}(\zeta_k|\xi_j) = \beta_{jk}\xi_j. \tag{11}$$

Noting that the slopes β_{jk} and the regression coefficient surface $\beta(s,t)$ are uniquely determining each other then provides two ways to interpret the functional regression relation: through the regression coefficient surface $\beta(s,t)$ (as exemplified in Figure 5 of the article) or

through the doubly-indexed sequence of slope coefficients β_{jk} (as shown in Figure 4 of the article).

The strength of the linear relationship can be measured by the functional coefficient of determination R^2,

$$R^2 = \frac{\int \text{var}[\text{E}(Y(t)|X)]dt}{\int \text{var}(Y(t))dt} = \sum_{j=1}^{\infty} \frac{\sum_{k=1}^{\infty} R_{kj}^2 \tau_k}{\sum_{k=1}^{\infty} \tau_k},$$

where $\lambda_j = \text{E}(\xi_j^2)$, $\tau_k = \text{E}(\zeta_K^2)$ and

$$R_{kj}^2 = \frac{\text{E}[(\xi_j \zeta_k)]^2}{\lambda_j \tau_k} = \frac{\beta_{jk}^2 \lambda_j}{\tau_k}$$

are the coefficients of determination for the simple linear regressions (11), for $k, j = 1, 2, \ldots$.

In practice, the numbers of included functional principal components ξ_j and ζ_k need to be truncated at a certain number, say at J and K. We choose J and K manually, guided by the FVE (fraction of variation explained) criterion. Therefore, the estimated regression coefficient function is

$$\hat{\beta}(s,t) = \sum_{k=1}^{K} \sum_{j=1}^{J} \hat{\beta}_{jk} \hat{\phi}_j(s) \hat{\psi}_k(t), \tag{12}$$

where $\hat{\beta}_{jk}$ is the estimated slope of the simple linear regression of $\hat{\zeta}_k$ on $\hat{\xi}_j$, i.e. of the functional principal components of response processes on those of the predictor processes. These components are then estimated through conditional expectation, see Eq. (4). The functional coefficient of determination can be then estimated as

$$\hat{R}^2 = \sum_{j=1}^{J} \frac{\sum_{k=1}^{K} \hat{R}_{kj}^2 \hat{\tau}_k}{\sum_{k=1}^{K} \hat{\tau}_k}, \quad \text{with } \hat{R}_{kj}^2 = \frac{\hat{\beta}_{jk}^2 \hat{\lambda}_j}{\hat{\tau}_k}. \tag{13}$$

We use bootstrap inference to assess the overall significance of the functional regression (compare Müller et al., 2008). The bootstrap samples are constructed by resampling from the observed data X_{il} for predictor processes and Y_{im} for response processes separately,

$1 \leq i \leq n$, $1 \leq l, m \leq N_i$. Two sets of random samples of size n are generated with replacement from the fly index set $\{1, 2, .., n\}$. Denoting the selected indices by $\{i'\}$ and $\{i''\}$, respectively, the data $X_{i'l}$ and $Y_{i''m}$, using all selected n pairs of indices, then form one bootstrap sample. This bootstrap sample can be considered as a sample under the null hypothesis of no regression relationship, because predictors and responses are sampled independently. We apply the functional linear regression model (7) to B such null bootstrap samples and the estimated coefficients of determination $\hat{R}^2$ then provide a null distribution for R^2. The bootstrap p-value of the functional regression is defined as the empirical quantile of the estimated coefficient of determination $\hat{R}^2$ for the targeted regression model.

We also remark that the p-values reported in Table 1 are not obtained by bootstrapping but correspond to those of F-tests, based on Gaussian assumptions for the random vector comprised of the first two functional principal components. We found that there is no strong empirical evidence against the Gaussian distribution assumption.

4. Additional Figure for Section 3

Figure S1 in this Supplement displays the mean function of the fertility trajectories in the left panel and their first three eigenfunctions in the right panel, for the sample of flies that is used for the functional regression analysis reported in Section 3. These are all flies, while the flies used for the analysis in Section 2 have been selected from the entire sample, as described there, giving rise to Figure 2 in the main text. Figure S1 and Figure 2 differ only in minor aspects, so that Figure 2 can reasonably be used as proxy for the analysis in Section 3 of the main text.

References

Ash, R. B. and Gardner, M. F. 1975 *Topics in stochastic processes.* Academic Press [Harcourt Brace Jovanovich Publishers], New York.

Capra, W. B. and Müller, H.G. 1997 An accelerated-time model for response curves. *Journal of the American Statistical Association* **92**, 72–83.

Liu, B. and Müller, H.G. 2008 *Functional data analysis for sparse auction data in* W. Jank and G. Shmueli, eds. Statistical Methods in eCommerce Research. Wiley & Sons, Inc, New York.

Müller, H.G., Chiou, J.M., and Leng, X. 2008 Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics* **9**, 60.

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

Rice, J. and Silverman, B. W. 1991 Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53**, 233–243.

Wang, J.L., Müller, H.G. and Capra, W. B. 1998 Analysis of oldest-old mortality: Lifetables revisited. *The Annals of Statistics* **26**, 126–163.

Yao, F., Müller, H.G., and Wang, J.L. 2005 Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

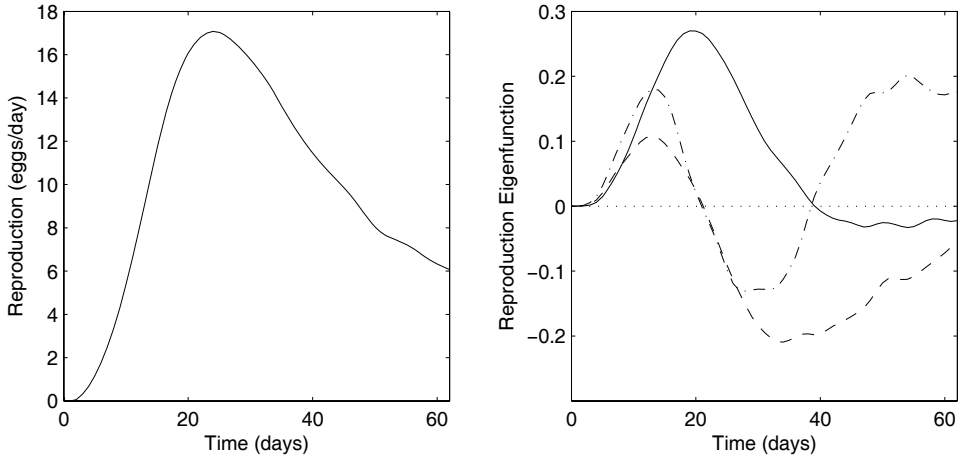

Figure S1: Estimated mean function (left panel) and the first three eigenfunctions (right panel) for individual reproductive trajectories as response functions. First (solid), second (dashed) and third (dash dot) eigenfunctions explain 48.85%, 29.71% and 10.93% of the total variation, respectively.