Marginal Functional Regression Models for Analyzing the Feeding Behavior of Pigs

Jan GERTHEISS, Verena MAIER, Engel F. HESSEL, and Ana-Maria STAICU

We observe a group of pigs over a period of about 100 days. Using high frequency radio frequency identification, it is recorded when each pig is feeding, leading to very dense binary functional data for each pig and day. One aim of the data analysis is to find pig-specific feeding profiles showing us the typical feeding pattern of each pig. For modeling the data, we use a marginal functional logistic regression approach, allowing us to model the densely observed binary measurements by assuming an underlying smooth subject-specific profile. The method also allows to incorporate additional covariates such as temperature and humidity that may influence the pigs' behavior. To account for correlation of measurements, we use robust standard errors and corresponding pointwise confidence intervals. Before analyzing the feeding behavior of pigs, the method employed is evaluated in simulation studies. As our approach is rather general, it may also be applied to other types of generalized functional data with similar characteristics as the pig data.

Key Words: Animal husbandry; Binary functional data; Generalized additive models; Penalized splines; Pig fattening.

1. INTRODUCTION

We observe a group of 127 pigs over one fattening period of about 100 days. On a very dense grid of time points, it is recorded when each pig is feeding. Data have been obtained from the PIGWISE project funded by the European Union within the ICT-AGRI 2010 call for transnational research projects. The objective of the project is to "optimize the performance and welfare of fattening pigs using high frequent radio frequency identification (HF RFID) and synergistic control on individual level."

HF RFID is used to record feeding times of the pigs. More precisely, HF RFID antennas were installed above the troughs to identify feeding pigs fitted with passive RFID tags on their ears (see Fig. 1). The HF RFID system at the trough registered the presence of the tags when they came within range of the antenna (Maselyne et al. 2014). This leads to binary functional data for each pig and day, as for each time point it is recorded whether the pig is

Jan Gertheiss (✉) and Engel F. Hessel Department of Animal Sciences, Georg-August-University of Göttingen, Göttingen, Germany (E-mail: *jgerthe@gwdg.de*). Verena Maier Department of Statistics, Ludwig-Maximilians-University of Munich, Munich, Germany. Ana-Maria Staicu Department of Statistics, North Carolina State University, Raleigh, USA.

Journal of Agricultural, Biological, and Environmental Statistics
DOI: 10.1007/s13253-015-0212-7

Published online: 07 July 2015

Figure 1. HF RFID tag attached to the pig's ear and antenna installed above the trough.

present at the trough or not. A reasonable assumption, however, is that pigs are only coming to the trough for feeding. So let $y_{ij}(t) = 1$ if pig i is feeding at time t at day j, and 0 otherwise.

Movements of the pig during feeding, however, can move the ear tag in and out of the range of the antenna. For these reasons, consecutive RFID registrations of an ear tag will display irregular time gaps between readings (Maselyne et al. 2014). Therefore, the data were downsampled to consecutive sampling intervals of 10 s . As sometimes a pig is just passing by the trough but not feeding, a pig was considered as feeding when the respective tag was registered at least twice within a 10 s interval. So on a very dense and regular grid of time points $t_1, t_2, t_3, \ldots, t_{8640}$ binary observations $y_{ij}(t_r) \in \{0, 1\}$ are available for pig i across day j, saying whether the pig is feeding ($y = 1$) or not ($y = 0$).

In addition to the feeding data, there are measurements such as temperature and humidity available that may influence the pigs' behavior. One important objective of the data analysis is to find pig-specific feeding profiles telling us when a certain pig is typically feeding. First, these profiles can be used for summarizing and illustrating the data observed. Second, they are a potential basis for further data analysis and smart usage of the HF RFID technology. On the one hand, this technology is intended for monitoring pigs and identifying pigs showing unusual feeding behavior since this may indicate problems such as sickness. For identifying "unusual" behavior, however, we first need to know the usual feeding behavior of a pig. On the other hand, once the feeding profiles are obtained, we can use them as a basis for further data analysis, for example, for clustering pigs or comparing groups of pigs.

An important feature of the dataset described is that there are over 800,000 observations available for each of the 127 pigs, making the dataset quite big; this has to be considered when thinking about methods for modeling these data. The aim of this paper is to present and discuss a feasible approach for modeling the data and estimating the pig-specific feeding profiles needed. For doing so, we propose a marginal functional logistic regression approach allowing us to model binary-valued functional observations. The main idea is to assume that there is an underlying smooth subject-specific profile that yields the binary functional observations. The method also allows to incorporate additional covariates. Although our methodology is motivated by the "big pig data" application, the proposed methods are general and applicable to other types of binary functional data with similar characteristics too.

The remainder of the paper is organized as follows. Section 2 reviews the existing literature on modeling binary functional data and presents our approach for modeling the pig data. In Sect. 3, we investigate the performance of the proposed approach numerically in simulations studies. Section 4 discusses the application of the methods to the pig data as well as subsequent data analysis, and Sect. 5 concludes.

2. MODELING RFID DATA

Modeling binary functional data have attracted great interest lately, see e.g., Hall et al. (2008), Serban et al. (2013), Goldsmith et al. (2015). Specifically, Hall et al. (2008) extended functional principal component analysis to generalized responses through a latent Gaussian process, and a monotone increasing known link function. The methods are designed for estimation and prediction but do not accommodate covariate effects, nor are developed for inference of the model parameters. More recently Serban et al. (2013) extended these ideas to the case where there are multiple binary-valued curves per subject and discussed adjustments when the probability of event is rare. The latter approach refers to the case considered in this paper, where we observe multiple binary-valued profiles for each pig. The approach by Serban et al. (2013) models the latent process as the sum of a smooth mean function, a pig-specific smooth random curve, and a random deviation from the pig's mean curve. The proposed methods focus on estimation of the population effects and are not developed to do prediction of the subject effects, nor accommodate additional covariate effects. However, prediction/estimation of the pig-specific latent trajectory as well as accommodating other covariates is the main objective in our project presented here. These research objectives have been considered in an very recent paper by Goldsmith et al. (2015), which propose estimation, inference, and prediction in a multilevel generalized functional model using a spline-based methodology and a Bayesian framework. Unfortunately these methods are recognized by the authors to be computationally intensive; in particular they would be computationally prohibitive for our data comprising 127 subjects, with about 100 binary-valued curves per subject and nearly 9000 observations per curve. We will therefore use an alternative and computationally feasible frequentist fixed effects approach, which is also more suitable for the research questions investigated with our data.

2.1. A MARGINAL FUNCTIONAL REGRESSION MODEL

In this section we introduce the proposed modeling framework for our data. Recall that $Y_{ij}(t)$ is the 0/1 valued response at time t of the jth day corresponding to the ith pig in the study. Denote by $E(Y_{ij}(t)) = \pi_{ij}(t)$ the probability that pig i is feeding at time t at day j. Assume that $Y_{ij}(t)$ is related to the (unknown) smooth underlying profile $\alpha_i(t)$ of pig i through the logit link, i.e.,

$$\pi_{ij}(t) = \frac{\exp\{\eta_{ij}(t)\}}{1 + \exp\{\eta_{ij}(t)\}}, \text{ with } \eta_{ij}(t) = \alpha_i(t). \tag{1}$$

In other words, we make the assumption that each pig i has its specific but latent feeding profile $\alpha_i(t)$ which determines at what time of the day the pig is typically feeding. Given pig i, we hence consider the observed curves $y_{ij}(t)$ as independent replicates across days $j = 1, \ldots, J$, with $J = 102$.

An important advantage of this model is that it also allows to incorporate additional covariates by extending $\eta_{ij}(t)$. Let $x_{1j}(t)$ and $x_{2j}(t)$ be two real-valued covariates observed at time t for the jth day; for example, $x_{1j}(\cdot)$ and $x_{2j}(\cdot)$ could be temperature and humidity curves corresponding to the jth day. The influence of these covariates on feeding probabilities can be taken into account via $\eta_{ij}(t) = \alpha_i(t) + \beta_{1i}x_{1j}(t) + \beta_{2i}x_{2j}(t)$. Here β_{1i} and β_{2i} are time-invariant subject-specific parameters capturing the effect of the corresponding covariates on the expected response. Furthermore the model can be extended to include interaction effects of the functional covariates. Specifically,

$$\eta_{ij}(t) = \alpha_i(t) + \beta_{1i}x_{1j}(t) + \beta_{2i}x_{2j}(t) + \beta_{12i}\{x_{1j}(t)x_{2j}(t)\}. \tag{2}$$

As, for example, the effect of temperature may vary with the value of humidity, an interaction model as described in (2) may be preferable to the main effects model above. Interaction models for functional regression with continuous response are, for example, investigated by Usset et al. (2013). The pig-specific coefficients β_{1i}, β_{2i}, and β_{12i} at (2) capture pig-specific reactions to changes in temperature and humidity. To separate the profile indicating the pig's typical feeding times from the effects of temperature and humidity, we can rewrite model (2) as $\eta_{ij}(t) = \alpha_i(t) + \beta_{0i} + \beta_{1i}x_{1j}(t) + \beta_{2i}x_{2j}(t) + \beta_{12i}\{x_{1j}(t)x_{2j}(t)\}$ with centered profile such that $\int_{0\text{h}}^{24\text{h}} \alpha_i(t)\,\mathrm{d}t = 0$. Note, in contrast to Hall et al. (2008) and Serban et al. (2013), the profiles $\alpha_i(t)$'s and effects $\beta_{\cdot i}$'s are specified as fixed effects, because we are particularly interested in the specific feeding behavior of pig i in the specific environment defined by the pigsty and pen the pig is raised in.

As a further generalization, regression coefficients $\beta_{\cdot i}$s could be specified as time-varying in terms of $\beta_{\cdot i}(t)$. However, while it is well known that the pigs' feeding can be influenced by temperature and humidity, there have been no studies in animal sciences indicating that effect sizes $\beta_{\cdot i}$s are varying across the day. Therefore, we will use the model with time-invariant β-parameters here.

2.2. ESTIMATION OF PROFILES AND REGRESSION PARAMETERS

The observations made at time points t_r are now used to estimate the pig-specific profiles and effects of temperature and humidity. Although these observations are correlated within curves, even when conditioned on the latent profiles, we can use generalized estimation equations (GEE; Liang and Zeger 1986; Zeger and Liang 1986) with a working independence assumption to estimate model parameters $\alpha_i(t)$ and $\beta_{\cdot i}$. This approach is justified by our interest in the marginal effects of the latent profiles (and other covariates) on feeding behavior at each time t; see (1) and (2) above. A working independence assumption is often used to obtain point estimates in longitudinal data analysis (see e.g., Wang et al. 2008).

Our primary interest is in the subject-specific profiles $\alpha_i(t)$, which are assumed to be smooth. For estimating them and guaranteeing some smoothness, we use penalized GEEs

in the lines of Chen et al. (2013). However, the approach by Chen et al. (2013) cannot be applied directly because the profiles $\alpha_i(t)$ show some specific behavior that has to be taken into account: Each profile describes the typical feeding behavior of a pig over the day. Of course the day begins, at time $t = 0$ h/12 am, when the day before ends, at time $t = 24$ h/12 pm. Therefore, feeding behavior is cyclic by nature and $\alpha_i(t)$ has to be cyclic too. For this reason, we use a *cyclic* cubic regression spline for $\alpha_i(t)$, that is, a penalized cubic regression spline where the ends of the smooth function match and have the same values up to the second derivative. It follows that, when $\alpha_i(t)$ is represented by a cubic spline function with m knots $k_1, \ldots, k_m$, we have $\alpha_i(k_1) = \alpha_i(k_m)$, $\alpha_i'(k_1) = \alpha_i'(k_m)$, $\alpha_i''(k_1) = \alpha_i''(k_m)$. So the boundary knots k_1 and k_m act like another interior knot connecting the two ends of the function. With appropriate basis functions $b_1(t), \ldots, b_{m-1}(t)$, the profiles can also be represented by $\alpha_i(t) = \sum_{l=1}^{m-1} b_l(t)\gamma_{il}$. For details on cyclic cubic regression splines, see Wood (2006).

To make sure that the estimated profiles can be interpreted, they have to be reasonably smooth. A simple approach to guaranty smoothness is to use a small number of (smooth) basis functions only, limiting, however, the types of profiles that can be fitted. An alternative and more flexible approach, advocated by, for example, Marx and Eilers (1999) and Ramsay and Silverman (2005), uses a relatively large number of basis functions/knots to be able to fit various shapes, but penalizes roughness of the profile. Specifically, we choose a rich basis with 30 equidistant knots and penalize curvature $\int_{0\text{h}}^{24\text{h}} \{\alpha_i''(t)\}^2 \, \mathrm{d}t$ in terms of penalized GEEs (as mentioned above). That means in case of model (1), instead of the "quasi score" equations $s(\gamma_i) = \sum_{j=1}^{J} B^\top (y_{ij} - \pi_{ij}) = 0$, we have to solve the penalized version

$$s_p(\gamma_i) = s(\gamma_i) - \lambda_i \Omega \gamma_i = 0, \tag{3}$$

where γ_i is the vector of basis coefficients of pig i corresponding to $\alpha_i(t)$, $(B)_{rl} = b_l(t_r)$ is the matrix of basis functions, λ_i is the (pig-specific) penalty parameter, Ω is the penalty matrix with entries $(\Omega)_{kl} = \int b_k''(t) b_l''(t) \mathrm{d}t$, $y_{ij} = (y_{ij}(t_1), y_{ij}(t_2), \ldots)^\top$ is the vector of observed values, and $\pi_{ij} = (\pi_{ij}(t_1), \pi_{ij}(t_2), \ldots)^\top$, with, according to (1), $\pi_{ij}(t) = \exp\{\alpha_i(t)\}/(1 + \exp\{\alpha_i(t)\})$. As in common penalized regression, parameter λ_i determines smoothness of the estimated function $\hat{\alpha}_i(t)$. Since feeding profiles may be very different for different pigs, penalty parameter λ_i is pig specific. Selection of smoothing parameters is discussed in Sect. 2.4.

In the model with additional covariates, the design matrix B varies across days. We therefore have $s_p(\gamma_i) = \sum_j B_j^\top (y_{ij} - \pi_{ij}) - \lambda_i \Omega \gamma_i$, where B_j is the (design) matrix of basis functions plus additional covariates $x_{1j}(t_r), x_{2j}(t_r)$, etc., and $\pi_{ij}(t) = \exp\{\eta_{ij}(t)\}/(1 + \exp\{\eta_{ij}(t)\})$, with $\eta_{ij}(t)$ as given at (2). Penalty matrix Ω is the same as before, but with some zeros for non-penalized regression coefficients.

As (3) is equivalent to penalized (binomial) likelihood estimation with (conditionally) independent data, estimation can then be carried out easily using the statistical program R (Core Team 2014) and add-on package `mgcv` (Wood 2006). As all parameters are pig specific, we can treat each pig separately and estimation can be done in parallel.

2.3. STATISTICAL INFERENCE

2.3.1. Robust Standard Errors and Pointwise Confidence Intervals

R package `mgcv` also provides estimated standard errors and (pointwise) confidence intervals for the profiles and further regression parameters. In our case, however, these estimates are terribly biased downwards, since they are based on the working independence assumption. Although a working independence assumption can be used for obtaining point estimates of regression parameters, when computing measures of uncertainty (such as standard errors) it has to be taken into account that observations are actually correlated.

In analogy to robust standard errors in non-functional marginal models (Liang and Zeger 1986; Fahrmeir and Tutz 2001), we can compute robust standard errors for the basis and regression coefficients. In the simple model (1) without further covariates, the formulas given by Chen et al. (2013) simplify considerably. For a specific pig i with smoothing parameter λ_i, and given an estimate $\widehat{S}_i$ of the covariance matrix S_i of $y_i = (y_i(t_1), y_i(t_2), \ldots)^\top$, we obtain the sandwich formula

$$\widehat{V}_i = \widehat{\mathrm{Cov}}(\hat{\gamma}_i) = \frac{1}{J}\left(B^\top \widehat{W}_i B + \frac{\lambda_i}{J}\Omega\right)^{-1} B^\top \widehat{S}_i B \left(B^\top \widehat{W}_i B + \frac{\lambda_i}{J}\Omega\right)^{-1}, \tag{4}$$

where $\hat{\gamma}_i$ is the vector of estimated basis coefficients of pig i, $(B)_{rl} = b_l(t_r)$ is the matrix of basis functions, $\widehat{W}_i = \mathrm{diag}\{\exp(\hat{\alpha}_i(t_r))/(1+\exp(\hat{\alpha}_i(t_r)))^2\}$ is a (diagonal) matrix with dimension corresponding to the number of measurement points, and $(\Omega)_{kl} = \int b_k''(t) b_l''(t)\mathrm{d}t$ is the penalty matrix. In general, J is the number of curves available. In our application, it is the number (102) of days each pig is observed. Equation (4) nicely shows how estimation accuracy depends on this number. For pointwise standard errors of the estimated profile $\hat{\alpha}_i = (\hat{\alpha}_i(t_1), \hat{\alpha}_i(t_2), \ldots)^\top$, we have $\widehat{\mathrm{Cov}}(\hat{\alpha}_i) = B\widehat{V}_i B^\top$. The covariance matrix S_i of y_i needed at (4) can, for example, be estimated via the empirical covariance matrix, or a smoothed version of it. Approximate, pointwise 95 % confidence intervals are obtained by adding and subtracting 2 estimated (pointwise) standard errors.

In the model with additional covariates, the design matrix B varies across days, and analogously to Chen et al. (2013), we use

$$\widehat{V}_i = \widehat{\mathrm{Cov}}(\hat{\gamma}_i) = F_i^{-1}\left\{\sum_j B_j^\top (y_{ij} - \hat{\pi}_{ij})(y_{ij} - \hat{\pi}_{ij})^\top B_j\right\} F_i^{-1}, \tag{5}$$

where $\hat{\gamma}_i$ is the vector of estimated basis and regression coefficients of pig i, $F_i = \{\sum_j B_j^\top \widehat{W}_{ij} B_j\} + \lambda_i \Omega$, B_j is the (design) matrix of basis functions plus additional covariates $x_{1j}(t_r)$, $x_{2j}(t_r)$, etc., $\widehat{W}_{ij} = \mathrm{diag}\{\exp(\hat{\eta}_{ij}(t_r))/(1+\exp(\hat{\eta}_{ij}(t_r)))^2\}$. Ω is the penalty matrix as before, but with some zeros for non-penalized coefficients, $y_{ij} = (y_{ij}(t_1), y_{ij}(t_2), \ldots)^\top$ and $\hat{\pi}_{ij} = (\hat{\pi}_{ij}(t_1), \hat{\pi}_{ij}(t_2), \ldots)^\top$. In Sect. 3 we will further evaluate these standard errors in simulation studies before using them for our RFID data.

2.3.2. Pointwise Versus Simultaneous Confidence Intervals

The approximate $100(1-\vartheta)\,\%$ pointwise confidence intervals for $\alpha_i(\cdot)$, as given above, are derived from pointwise normality approximation of $\widehat{\alpha}_i(t)$ at each time point t: $\hat{\alpha}_i(\mathfrak{t}) \pm z_{1-\vartheta}\sqrt{\text{diag}\{\text{Cov}(\hat{\alpha}_i(\mathfrak{t}) - \alpha_i(\mathfrak{t}))\}}$, where $\mathfrak{t} = (t_1, t_2, \ldots)^\top$ is the grid of measurement points, $z_{1-\vartheta}$ is the $(1-\vartheta)$ quantile of the standard normal distribution, and $\text{Cov}(\hat{\alpha}_i(\mathfrak{t}) - \alpha_i(\mathfrak{t}))$ is replaced by the robust estimate $\widehat{\text{Cov}}(\hat{\alpha}_i)$. Although these intervals are commonly reported as a measure of estimation accuracy of smooth terms, their usage is limited, mainly because they are based on the pointwise variance of the estimator and thus ignore the dependence structure. In particular, if it is of interest to study whether the smooth effect, as a function, is significantly different from zero, joint confidence intervals are more appropriate. Joint confidence intervals are constructed roughly using the same structure but with a critical value that accounts for the covariance structure of the estimator. Degras (2011), for example, discussed simultaneous confidence intervals in nonparametric functional regression when using local linear estimators.

In our setting with spline functions, however, it is more appropriate to follow Ruppert et al. (2003), Crainiceanu et al. (2012) and Goldsmith et al. (2013) and use $\hat{\alpha}_i(\mathfrak{t}) \pm m_{1-\vartheta}\sqrt{\text{diag}\{\text{Cov}(\hat{\alpha}_i(\mathfrak{t}) - \alpha_i(\mathfrak{t}))\}}$, where $m_{1-\vartheta}$ is the $(1-\vartheta)$ quantile of the random variable $M = \max_t \left| \frac{\hat{\alpha}_i(t) - \alpha_i(t)}{\sqrt{\{\text{Var}(\hat{\alpha}_i(t) - \alpha_i(t))\}}} \right|$. Using that $\hat{\alpha}_i(t)$ is approximately unbiased and the vector of estimated basis coefficients $\hat{\gamma}_i$ is approximately normal with mean zero and (estimated) covariance matrix $\widehat{V}_i$, we can draw a sample of M and estimate $m_{1-\vartheta}$ by the empirical quantile. As before, $\text{Cov}(\hat{\alpha}_i(\mathfrak{t}) - \alpha_i(\mathfrak{t}))$ is replaced by $\widehat{\text{Cov}}(\hat{\alpha}_i)$.

As pointed out above, one of the biggest advantages of simultaneous intervals over pointwise ones is that they can be used for hypothesis testing, in particular, if the function is differing from a constant or straight line. In our case, however, the functions we are estimating are latent profiles determining the typical feeding times of a pig. It is well known in animal sciences (see e.g., Young and Lawrence 1994) that pigs do have their preferred feeding times; at least they are usually not feeding at night but during the day (see e.g., Hyun et al. 1997). Therefore, testing such hypotheses is not of interest in this paper. Nevertheless, more generally, hypothesis testing in the framework of marginal functional regression models may be valuable; we leave this topic to future research here as it exceeds the scope of this paper. The main focus in our application is in estimating the individual profiles and furthermore determining the specific time (points) at which the profile estimates have to be taken with caution. Specifically, the estimated pointwise standard errors are interpreted relatively to one another and are informative with respect to times or intervals of time where the profile estimates are not reliable. So in our specific application pointwise intervals provide more useful information.

2.4. SELECTION OF PENALTY PARAMETERS

Several very elegant methods are available for determining penalty parameters, such as AIC or REML, which are also implemented in `mgcv`. As above with standard errors, however, these approaches rely on (conditional) independence of observations, which is not the case with our data. As a consequence of the very large number of data points, an

independence assumption leads to far too small penalty parameters and hence drastic undersmoothing. We therefore use pig-specific K-fold cross-validation on a daily basis. More precisely, we use folds of entire days, and for each day j in the test set the integrated Brier score (IBS)

$$\int_{0\text{h}}^{24\text{h}} (y_{ij}(t) - \hat{\pi}_{ij}(t))^2 \, \mathrm{d}t, \tag{6}$$

is calculated, with $\hat{\pi}_{ij}(t)$ being the estimated (marginal) probability that pig i is feeding at time t (given in seconds since 0 h) at day j. For each pig minimization of the IBS is done separately, producing pig-specific smoothing parameters.

2.5. MARGINAL MODELS VERSUS FUNCTIONAL RANDOM EFFECTS

When marginal models are used, there is usually a conditional model incorporating random effects that could be employed alternatively, with the latter being preferred by some authors, see e.g., Lee and Nelder (2004). In case of our functional data, the appropriate multilevel model would take the form $\eta_{ij}(t) = \theta_i(t) + u_{ij}(t)$, where $\theta_i(t)$ is, in some sense, comparable to $\alpha_i(t)$ at (1), and $u_{ij}(t)$ is a (pig- and) day-specific functional random effect modeled as a Gaussian process with mean zero and some covariance matrix $E\{u_{ij}(t)u_{ij}(s)\} = \Sigma(t, s)$ (Hall et al. 2008; Serban et al. 2013). The latter is typically expanded in eigenfunctions, see e.g., Di et al. (2009) and Greven et al. (2010). For simplicity, we only consider the model without further covariates here. Given the random effect $u_{ij}(t)$, observations made at points $t_t, t_2, \ldots$ are assumed as independent.

There are mainly two reasons why we are not using the model with functional random effects above. First, for two adjacent measurement points t_r and t_{r-1} association is rather strong, because given pig i is not feeding at time t_{r-1} it is quite likely that it is not feeding at time t_r either, and if it is feeding at time t_{r-1} it is also feeding at time t_r with relatively large probability. To be able to explicitly model these dependencies, $u_{ij}(t)$ has to make a few very large and abrupt jumps, while otherwise being almost constant at small (i.e., negative) or large values, respectively. A behavior like this, however, can hardly be described by a Gaussian process. Second, and very important in applied research like ours, even though $\theta_i(t)$ seems to play the role of $\alpha_i(t)$, it is not generating pig-specific marginal feeding probabilities like $\alpha_i(t)$ at (1). This would be the case in a (functional) linear model (with respect to the marginal mean), but in the logit model it is not. Therefore, $\alpha_i(t)$ is preferred in our application from the viewpoint of interpretation.

3. SIMULATION STUDIES

We evaluate the proposed methods using a simulation experiment design to mimic the design of the data application. We consider two main scenarios: (1) when the setting contains no covariates and (2) when the setting involves two functional covariates without measurement error. The corresponding covariates in our application are temperature and humidity, for which it seems reasonable to assume that they are observed without error or with negligible measurement error. In case of noisy, irregularly sampled or sparse data, smooth-

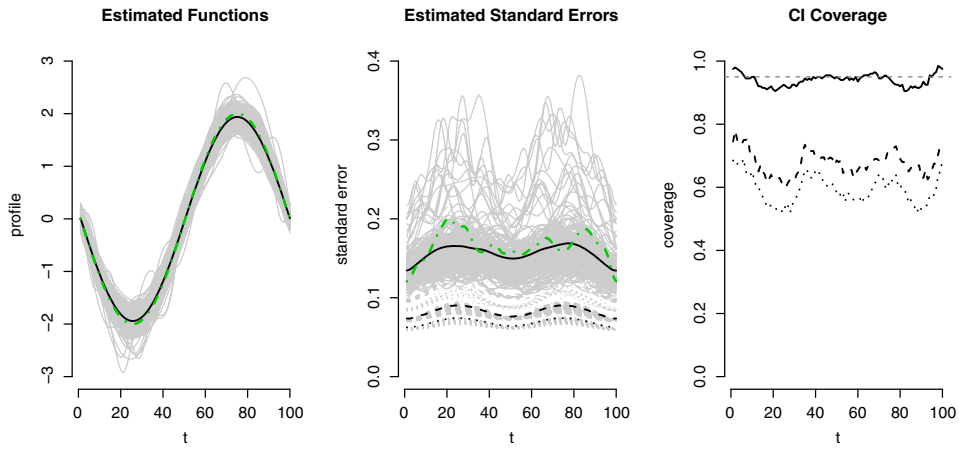

Figure 2. Simulation scenario (i) with correlation matrix (7). *Left* Estimated profiles with mean (*solid black*) and true profile (*dashed*/*dotted*). *Middle* Estimated naive standard errors assuming independence ("frequentist"/*dotted*, "bayesian"/*dashed*), robust standard errors (*solid*) as well as "true" ones as standard deviations of estimates in the left panel (*dashed*/*dotted*); *black curves* refer to mean estimates. *Right* Coverage (across 200 simulation runs) of 95 % (pointwise) confidence intervals using standard errors from the middle panel (Color figure online).

ing/reconstructing the data, for example, using functional principal components analysis, is recommended before fitting a model such as (2).

We consider equidistant measurement points $t_1, \ldots, t_{100}$ between 0 and 100 with marginal (feeding) probabilities defined by a sine-like profile $\alpha(t)$ as shown by the green (dashed/dotted) line in Fig. 2 (left). This profile can be interpreted as a pig that does not like feeding at night and in the morning but during the rest of the day. For the correlation matrix of observations $y(t_1), \ldots, y(t_{100})$ we use a band matrix Σ, since this correlation structure is typical for functional data as the RFID records. Specifically, in the first simulation scenario (i), we use a rather narrow band with off-diagonals

$$(\Sigma)_{t \neq s} = \begin{cases} 0.7(|t-s|-10)^2/81 & \text{if } |t-s| \leq 10, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We generate binary data $y_j(t_1), \ldots, y_j(t_{100})$, $j = 1, \ldots, 40$, which can be interpreted as feeding records of a specific pig at 40 days, with the pig having the latent green (dashed/dotted) feeding profile from Fig. 2 (left). For actually generating the data, we use R package `bindata` (Leisch et al. 2012) and function `rmvbin()`. This function creates correlated multivariate binary random variables with given marginal probabilities by thresholding a multivariate normal distribution with an adequately chosen covariance matrix. Data generation, estimation of profiles, and standard errors are repeated 200 times. Figure 2 (left) shows the estimated profiles (gray) and the corresponding mean (black). We see that estimation works very well. Due to penalization, estimates are slightly biased towards zero, but the bias is only visible around the true profile's minimum and maximum. The central panel of Fig. 2 shows the estimated standard errors for each simulation run (gray) and the corresponding mean values (black). The solid lines refer to the robust standard errors using (4), the dashed and dotted lines show the naive standard errors obtained from `mgcv`

when independence is assumed. The dotted lines refer to the "frequentist" approach, and the dashed ones to the "Bayesian" covariance matrix. The green dashed/dotted line shows the "true" standard errors as the standard deviations of the (gray) estimates in the left panel. We see that these standard errors are dramatically underestimated by the naive approaches assuming independence, whereas the robust approach works very well. This is also seen when looking at the coverage of (approximate) 95 % confidence intervals created by adding and subtracting 2 estimated standard deviations (Fig. 2, right). The confidence intervals using the robust standard errors (solid) show coverage very close to the nominal level for almost the entire domain, whereas naive standard errors produce intervals that are virtually useless. As before, the dotted line refers to the "frequentist" approach and the dashed line to the "Bayesian." For instance, the solid line is obtained as the pointwise coverage across simulation runs when in each run the corresponding solid (gray) line from the middle panel is used as the standard errors. Only in the regions where estimates of profiles and standard errors are biased (see Fig. 2, left/middle), coverage of robust confidence intervals is a little bit lower than 95 %.

In the second scenario (ii), we consider lower correlations but a much wider band than in the first scenario (7). The off-diagonals of the correlation matrix are now chosen as

$$(\Sigma)_{t\neq s} = \begin{cases} 0.3(|t-s|-50)^2/2401 & \text{if } |t-s| \leq 50, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The results are very similar to those obtained before. Estimates of profiles are very good and only slightly biased (Fig. 3, left). Estimated robust standard errors are very close to the "true" ones (Fig. 3, middle), whereas naive standard errors are much too small. The coverage of robust confidence intervals is even better than before and very close to the nominal level of 95 % (Fig. 3, right). When the sample size is increased, e.g., to $j = 1, \ldots, 100$, results are qualitatively similar, but standard errors become smaller (not shown).

To investigate the performance when estimating profile $\alpha(t)$ and parameters β_1, β_2, and β_{12} from model (2), we generate a set of predictor curves $x_{1j}(t)$, $j = 1, \ldots, 40$, using (similar to Tutz and Gertheiss (2010))

$$x_{1j}(t) = \frac{1}{10}\left\{\sum_{k=1}^{5}(b_{jk}\sin(t\pi(5-b_{jk})/50) - m_{jk}) + 15\right\}, \tag{9}$$

where $t \in (0, 100)$, $b_{jk} \sim U(0, 5)$, $m_{jk} \sim U(0, 2\pi)$, with $U(a, b)$ denoting the uniform distribution on $[a, b]$. The corresponding regression coefficient β_1 is set to 1. In addition to $x_{1j}(t)$, we generate noise variables $x_{2j}(t)$ (also using (9)), that is, $\beta_2 = \beta_{12} = 0$. Profile $\alpha(t)$ and measurement points $t_1, \ldots, t_{100}$ remain the same as before. Using the corresponding marginal probabilities and correlation matrix (8) again, we generate binary observations $y_j(t_1), \ldots, y_j(t_{100})$, $j = 1, \ldots, 40$, estimate regression parameters and standard errors, and repeat this procedure 200 times. Results for profile $\alpha(t)$ (not shown) are very similar to the simulation studies before. Table 1 shows the resulting mean estimates for both our robust standard errors and the naive ones using the "bayesian" or the "frequentist" covariance matrix from `mgcv`. As before, "true" refers to the observed standard deviation of the

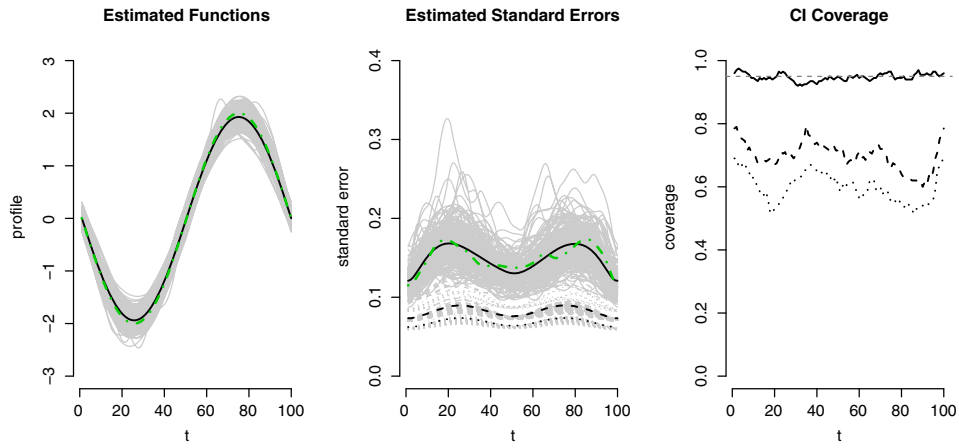

Figure 3. Simulation scenario (ii) with correlation matrix (8). *Left* Estimated profiles with mean (*solid black*) and true profile (*dashed*/*dotted*). *Middle* Estimated naive standard errors assuming independence ("frequentist"/*dotted*, "bayesian"/*dashed*), robust standard errors (*solid*) as well as "true" ones as standard deviations of estimates in the left panel (*dashed*/*dotted*); *black curves* refer to mean estimates. *Right* Coverage (across 200 simulation runs) of 95 % (pointwise) confidence intervals using standard errors from the middle panel (Color figure online).

Table 1. Mean estimates of regression coefficients and standard errors plus confidence interval coverage across 200 simulation runs when using robust, naive "bayesian" (b) or "frequentist" (f) standard errors; "true" refers to the observed standard deviation of the estimated regression coefficients.

	Regression coefficient	Standard error/CI coverage			
		Robust	Naive (b)	Naive (f)	True
β_1	0.962	0.178/0.935	0.075/0.565	0.074/0.560	0.187/0.975
β_2	0.010	0.160/0.940	0.063/0.475	0.063/0.475	0.178/0.970
β_{12}	−0.006	0.235/0.955	0.106/0.595	0.105/0.595	0.238/0.940

estimated regression coefficients. Approximate 95 % confidence intervals were generated in each simulation run by adding and subtracting 2 estimated standard errors to/from the estimated regression coefficient. For "true" we used the same standard errors (the "true" ones) in each run. We see that the robust standard errors give good estimates of the true standard deviations, and corresponding confidence intervals are close to the nominal level, making them also useful for testing. Naive standard errors, by contrast, are extremely biased, making them rather useless in practice.

4. APPLICATION TO RFID DATA

We now discuss the application of the proposed methods to the pig data study. Recall that there are 127 pigs, and each pig is observed over 102 consecutive days. The response is the indicator whether the respective pig is feeding at time t, $t \in [0, 24)$. Additionally the data contain temperature and humidity curves for the corresponding day.

In Sect. 4.1 we show and discuss the pig-specific profiles and covariate effects from model (2), which were estimated using penalized GEEs as described in Sect. 2. In Sect. 4.2 we use the results from Sect. 4.1 for subsequent data analysis studying the relationship between the estimated profiles and the pigs' weight. By doing so we examine potential effects of animal husbandry and trough design on the pigs' weight, which is an important factor with respect to the farmer's economic success.

4.1. PROFILES AND COVARIATE EFFECTS

Figure 4 (top/bottom) exemplarily shows the estimated (centered) profiles for four pigs (solid black), plus/minus two estimated standard errors. The naive standard errors obtained from `mgcv` (via the "bayesian" covariance matrix) are displayed as shaded regions, the robust errors according to (5) are given as dashed red lines. The naive standard errors indicate very precise estimation, which is extremely over-optimistic (compare Sect. 3). The robust errors are much wider, but still indicate that the profiles' most prominent features are reliable. We see, for example, that pig A (top left) and pig B (top right) show two clear peaks in the morning and afternoon/evening, with the second peak being much sharper for pig B. Such a two-peak feeding behavior is believed to be somewhat natural feeding behavior of pigs. Also pig C shows two peaks but the first one is much less prominent. Pig D typically feeds during the day, primarily in the afternoon, but we cannot identify any specific peaks.

A possible explanation and interpretation of these finding is as follows. The space at the trough is limited; more precisely, each pen contains 16 pigs, but the trough installed only provides space for two pigs at a time (see also Fig. 1). As a consequence, only the stronger pigs (the pigs rather on top of the hierarchy) can afford to follow the "natural" two-peak profile. The weaker pigs, such as pig D, by contrast, have to take the opportunity whenever the trough is vacant. Therefore, specific peaks are hardly identified. Other pigs, such as C, choose a third strategy: they focus on the peak in the afternoon/evening, when the stronger pigs already have fed.

The effects of the covariates temperature and humidity across all pigs are summarized in Fig. 4 (bottom). For most pigs main effects of temperature and humidity are clearly negative. However, when taking a closer look at the interaction terms, although seemingly small in Fig. 4, the overall effects of the covariates temperature and humidity on feeding probabilities as found in our data become rather small. A possible explanation for that is that the variation of temperature and humidity in the pigsty is quite low. Specifically, the inter-quartile range (IQR) of temperature is just 2°C, with median/mean temperature of about 25°C; for humidity the IQR is just 10 % with the median/mean being about 58/59 %. Although the presumably optimal temperature for pigs is somewhat lower than 25°C the variation around 25°C found in the data is too small to cause large changes in feeding probabilities. Therefore, it is hard to see any effects of temperature and humidity, at least for the majority of pigs. Nevertheless, when taking standard errors into account (not shown), regression coefficients appear to be significantly nonzero for most of the pigs. Therefore, we decided to include the covariates in the model to correct for potential, albeit small, effects when estimating the feeding profiles.

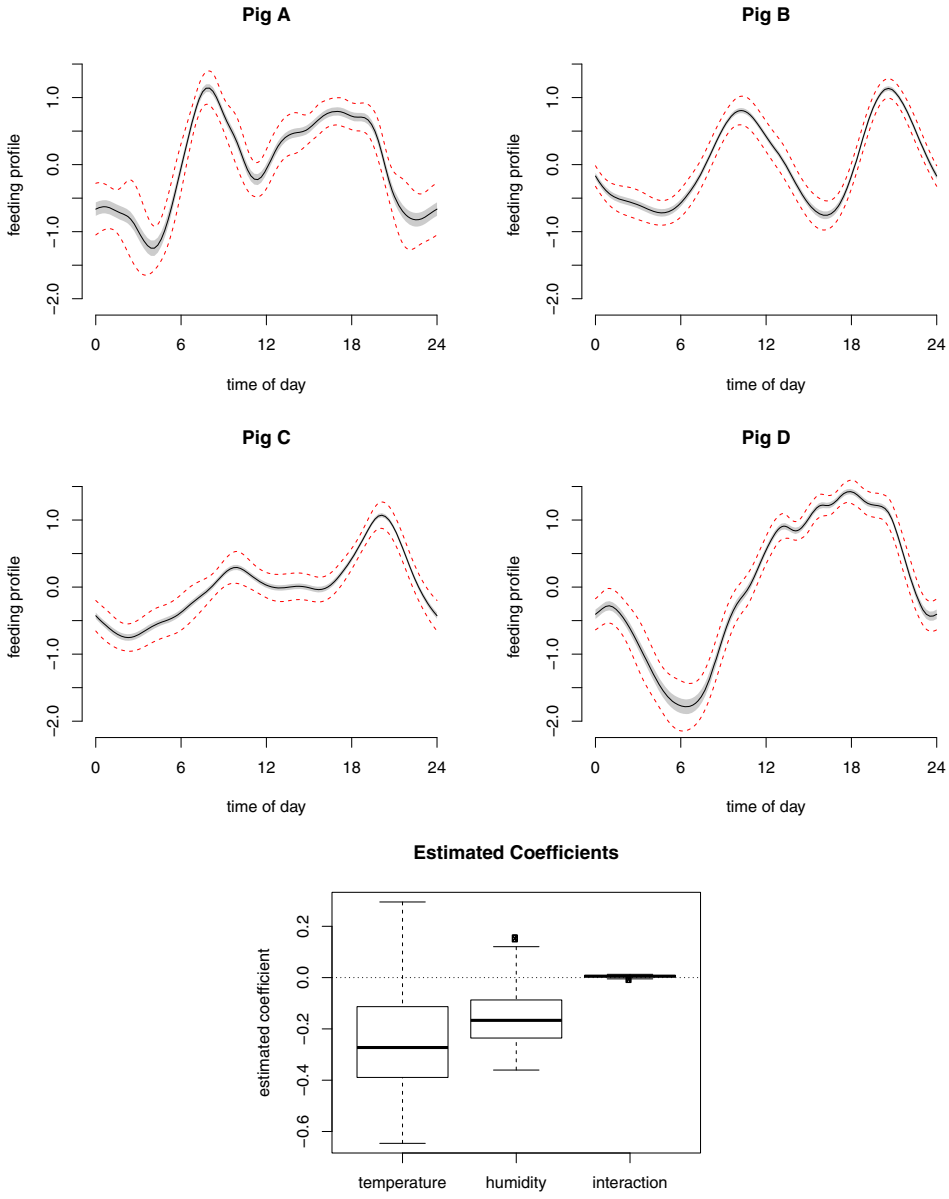

Figure 4. Estimated feeding profiles for four pigs (*top*/*middle*), ± 2 estimated standard errors (naive *shaded*; robust *dashed red*), and boxplots of the estimated regression coefficients across all pigs (*bottom*) (Color figure online).

4.2. THE RELATIONSHIP BETWEEN FEEDING PROFILES AND WEIGHT

Once the feeding profiles are obtained, they can be used in a secondary data analysis. Based on the interpretation given above, we are particularly interested whether the pigs' weight and the profiles are related; more precisely, whether supposedly stronger pigs with a two-peak profile gain more weight than the weaker pigs. For answering this question we will follow two approaches: (1) We will first cluster pigs according to their profiles and then

compare the weight (gain) in the different clusters. (2) We will directly regress the weight (gain) on the profiles using a functional regression model.

4.2.1. Cluster Analysis

Figure 5 (top) shows the dendrogram for an agglomerative hierarchical clustering with complete linkage. For measuring the distance between two estimated profiles/pigs i and l we used $\int_{6\text{h}}^{24\text{h}} \{\hat{\alpha}_i(t) - \hat{\alpha}_l(t)\}^2 \, \mathrm{d}t$. Note that, as the pigs rarely feed at night, we exclude the time between midnight and 6 am from the calculation of the distance. From the dendrogram we identify three large- and three small/medium-sized clusters. Figure 5 (bottom left) shows the cluster-specific mean profiles (the gray lines correspond to the two smallest clusters with less than ten observations). Cluster 1, for instance, shows a clear two-peak profile, whereas cluster 3 primarily contains, supposedly weak, pigs like D from Fig. 4 that tend to feed in the afternoon but without sharp peaks.

Figure 5 also shows the weight of the pigs at the end of the fattening period (bottom middle) and the weight gain across the entire period (bottom right) separately for the six clusters. Indeed, it seems that pigs from (two-peak) cluster 1 gained slightly more weight than pigs from cluster 3. Differences observed in the boxplots, however, are not significant (one-way ANOVA p values 0.21 and 0.17 for weight and weight gain, respectively).

4.2.2. Regression Analysis

Besides the cluster analysis above, we directly regress the weight (gain) of the pigs on the obtained profiles $\hat{\alpha}_i(t)$. In addition to the profiles, we also consider covariate z_i as the value of the parametric term $\hat{\beta}_{0i} + \hat{\beta}_{1i} x_{1j}(t) + \hat{\beta}_{2i} x_{2j}(t) + \hat{\beta}_{12i} \{x_{1j}(t) x_{2j}(t)\}$ at temperature 25°C and humidity 59 %, which is the mean temperature and humidity in the pigsty over the fattening period (see above). So z_i indicates the overall level of registrations of pig i at the trough. The regression model we are using is a functional linear model, sometimes also called signal regression, $y_i = \gamma_0 + \gamma_1 z_i + \int_{0\text{h}}^{24\text{h}} \delta(t) \hat{\alpha}_i(t) \, \mathrm{d}t + \epsilon_i$, with functional predictor $\hat{\alpha}_i(t)$, scalar covariate z_i, and response y_i being the weight of pig i at the end of the fattening period or its weight gain across the fattening period, respectively. Errors ϵ_i are assumed as iid normal with mean zero and constant variance. For details on the functional linear model, see e.g., Ramsay and Silverman (2005). For estimating model parameters γ_0, γ_1, and $\delta(t)$, we again use R package `mgcv`. As the influence of the feeding profile right before and right after midnight can be assumed to be very similar, coefficient function $\delta(t)$ is modeled as a cyclic cubic spline, with smoothing parameter estimated by REML (as, for example, suggested in Goldsmith et al. 2011). Figure 6 shows the results for both response weight at the end of the fattening period (maximum weight) and weight gain across the fattening period. The negative values of the coefficient functions in the afternoon/evening indicate that pigs with large feeding probabilities in the afternoon only have/gain less weight than the pigs that also have peaks in the morning, as peaks in the morning have positive effect on the weight. As with the cluster analysis above, however, these results have to be taken with caution due to large uncertainty as seen from the wide (pointwise 95 %) confidence intervals (shaded).

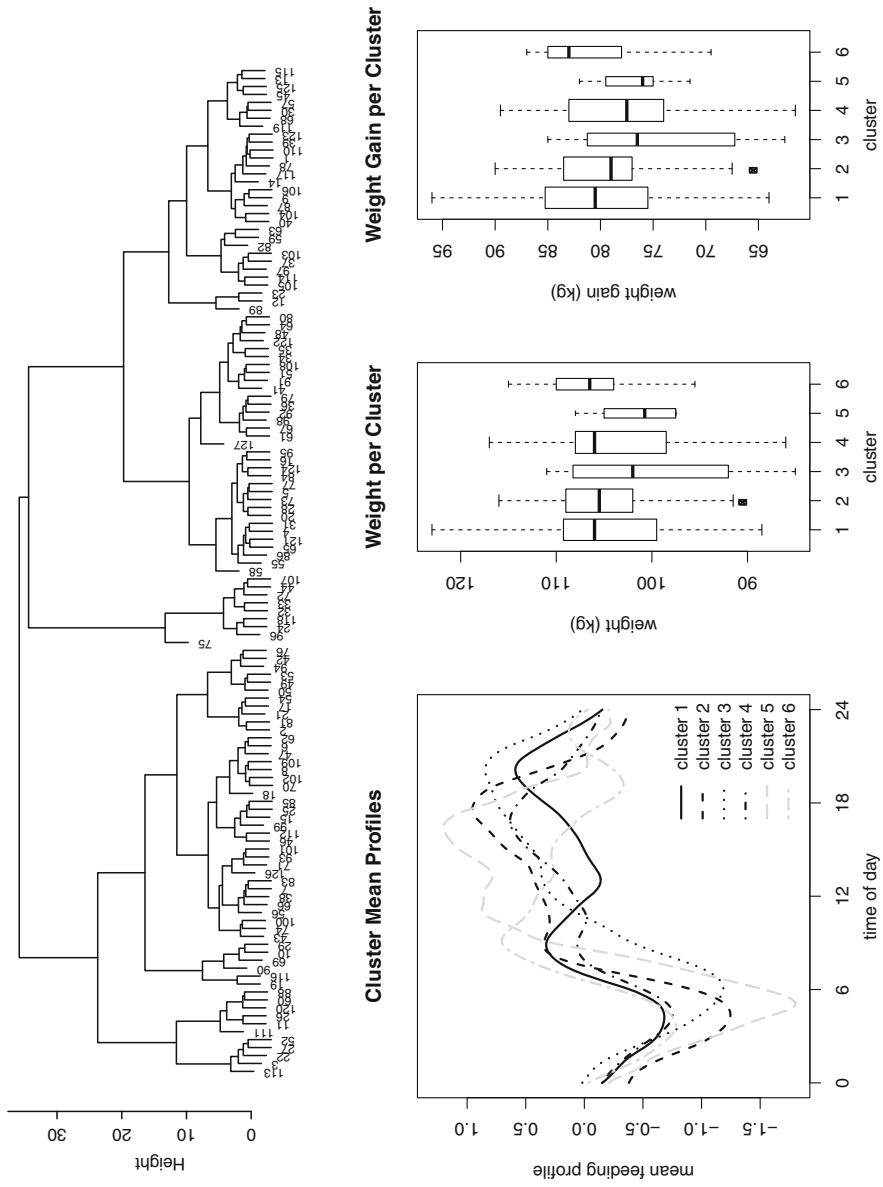

Figure 5. Dendrogram of the pigs in the study when clustering them according to their feeding profiles (*top*), cluster-specific mean feeding profiles (*bottom left*), and weight (gain) per cluster (*bottom middle/right*); box widths are proportional to the square-roots of the number of pigs in the respective cluster.

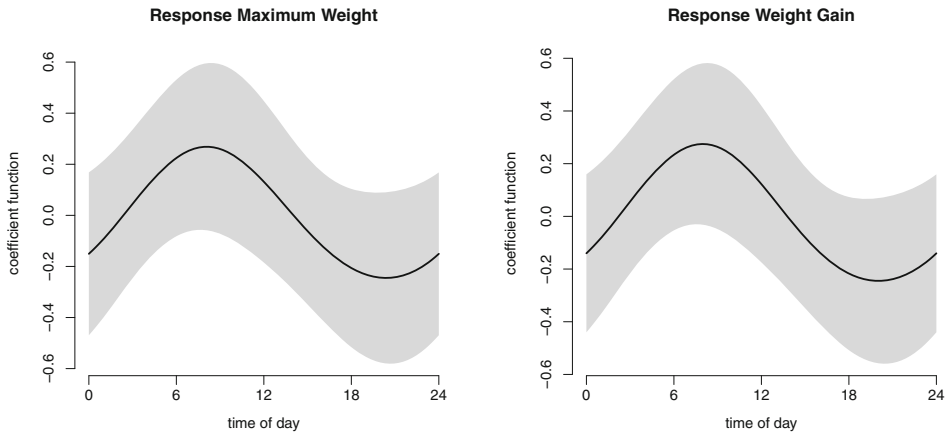

Figure 6. Estimated coefficient functions in a scalar-on-function regression model with response weight of the pigs at the end of the fattening period (*left*), or weight gain across fattening period (*right*).

The influence of covariate z on weight is positive, but only for weight gain it is (weakly) significant (on the 10 % level). Intuitively, a pig that is registered frequently at the trough is supposed to feed much more—and hence gain more weight—than a pig that is registered only sporadically. Although this positive effect is observed in the data, it is weaker than what might have been expected. This indicates that registration is not only influenced by feeding but also highly dependent on factors which are not associated with feeding behavior, such as the position of the tag on the pig's ear. Similar findings have also been made by Maselyne et al. (2014) when comparing video surveillance and RFID registration. But note, though some tags are harder to register than others, the pig/tag-specific profiles (as shown in Fig. 4) still give valuable information about the feeding behavior of the respective pig. In particular, the peaks indicate the typical feeding times.

5. SUMMARY AND DISCUSSION

We analyzed binary but functional HF RFID registrations of fattening pigs at the trough. We assumed that smooth pig-specific latent feeding profiles are generating the binary outcomes observed. For estimating these profiles we used a marginal functional logistic regression model, which also allows to correct for additional covariate effects. The obtained profiles indicate when a specific pig is typically feeding. For quantifying uncertainty with respect to the estimated profiles, we used robust standard errors in analogy to non-functional marginal models. As shown in simulation studies, this approach gives reasonable estimates of uncertainty also for finite samples.

Our analysis illustrated that pigs can show very different feeding behavior, with several pigs deviating from the "natural" two-peak profile. Having our trough design with limited space in mind, our hypothesis is that those pigs are the weaker ones which can only feed when the stronger pigs allow them to do so. Cluster and regression analysis indeed indicated that pigs with larger feeding profile in the afternoon/evening only tend to gain less weight than pigs also having peaks in the morning. Although these findings were not significant, it

might be worth thinking about a pen/trough design which makes sure that also the weaker pigs have the feeding times they need. On the one hand, such a modified trough design would require more space and financial investment. On the other hand, however, the more weight the pigs have at the end of the fattening period the better for the farmer.

ACKNOWLEDGEMENTS

The results presented are generated in the framework of the ICT-AGRI era-net project PIGWISE "Optimizing performance and welfare of fattening pigs using HF RFID and synergistic control on individual level" (Call for transnational research projects 2010). The German contribution was funded by the German Federal Office for Agriculture and Food (BLE). Furthermore, we would like to thank two anonymous Reviewers for their helpful comments and constructive criticism which have led to a much improved manuscript.

[Received October 2014. Accepted June 2015.]

REFERENCES

Chen, H., Wang, Y., Paik, M. C., and Choi, H. A. (2013), A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data, *Journal of the American Statistical Association*, 108, 1216–1229.

Crainiceanu, C. M., Staicu, A.-M., Ray, S., and Punjabi, N. (2012), Bootstrap-based inference on the difference in the means of two correlated functional processes, *Statistics in Medicine*, 31, 3223–3240.

Degras, D. A. (2011), Simultaneous confidence bands for nonparametric regression with functional data, *Statistica Sinica*, 21, 1735–1765.

Di, C.-Z., Crainiceanu, C., Caffo, B. S., and Punjabi, N. M. (2009) Multilevel functional principal components analysis, *The Annals of Applied Statistics*, 3, 458–488.

Fahrmeir, L., and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed., Springer, New York.

Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B., and Reich, D. (2011), Penalized functional regression, *Journal of Computational and Graphical Statistics*, 20, 830–851.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013), Corrected confidence bands for functional data using principal components, *Biometrics*, 69, 41–51.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015), Generalized multilevel functional-on-scalar regression and principal components analysis, *Biometrics*, 71, 344–353.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010), Longitudinal functional principal components analysis, *Electronic Journal of Statistics*, 4, 1022–1054.

Hall, P., Müller, H.-G., and Yao, F. (2008), Modelling sparse generalized longitudinal observations with latent Gaussian processes, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 70, 703–723.

Hyun, Y., Ellis, M., McKeith, F. K., and Wilson, E. R. (1997), Feed intake pattern of group-housed growing-finishing pigs monitored using a computerized feed intake recording system, *Journal of Animal Science*, 75, 1443–1451.

Lee, Y. and Nelder, J. A. (2004), Conditional and marginal models: Another view, *Statistical Science*, 19, 219–238.

Leisch, F., Weingessel, A., and Hornik, K. (2012), *bindata: Generation of Artificial Binary Data*. R package version 0.9-19.

Liang, K. Y. and Zeger, S. L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, 73, 13–22.

Marx, B. D. and Eilers, P. H. C. (1999), Generalized linear regression on sampled signals and curves: A p-spline approach, *Technometrics*, 41, 1–13.

Maselyne, J., Saeys, W., De Ketelaere, B., Mertens, K., Vangeyte, J., Hessel, E. F., Millet, S., and Van Nuffel, A. (2014), Validation of a High Frequency Radio Frequency Identification (HF RFID) system for registering feeding patterns of growing-finishing pigs, *Computers and Electronics in Agriculture*, 102, 10–18.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge.

Serban, N., Staicu, A.-M., and Carroll, R. J. (2013), Multilevel cross-dependent binary longitudinal data, *Biometrics*, 69, 903–913.

Tutz, G. and Gertheiss, J. (2010), Feature extraction in signal regression: A boosting technique for functional data regression, *Journal of Computational and Graphical Statistics*, 19, 154–174.

Usset, J., Staicu, A.-M., and Maity, A. (2013), *Interaction models for functional regression*. *Preprint*.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, *Journal of the American Statistical Association*, 103, 1556–1569.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC, London.

Young, R. J. and Lawrence, A. B. (1994), Feeding behaviour of pigs in groups monitored by a computerized feeding system, *Animal Science*, 58, 145–152.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, 42, 121–130.