Biometrics 69, 396–404
June 2013
DOI: 10.1111/biom.12011

Recovering Gradients from Sparsely Observed Functional Data

Sara López-Pintado* **and Ian W. McKeague****

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, 6th Floor, New York, New York 10032, U.S.A.

*email: sl2929@columbia.edu

**email: im2131@columbia.edu

Summary. The recovery of gradients of sparsely observed functional data is a challenging ill-posed inverse problem. Given observations of smooth curves (e.g., growth curves) at isolated time points, the aim is to provide estimates of the underlying gradients (or growth velocities). To address this problem, we develop a Bayesian inversion approach that models the gradient in the gaps between the observation times by a tied-down Brownian motion, conditionally on its values at the observation times. The posterior mean and covariance kernel of the growth velocities are then found to have explicit and computationally tractable representations in terms of quadratic splines. The hyperparameters in the prior are specified via nonparametric empirical Bayes, with the prior precision matrix at the observation times estimated by constrained ℓ_1 minimization. The infinitessimal variance of the Brownian motion prior is selected by cross-validation. The approach is illustrated using both simulated and real data examples.

Key words: Functional data analysis; Growth trajectories; Ill-posed inverse problem; Nonparametric empirical Bayes; Tied-down Brownian motion.

1. Introduction

The extensive development of functional data analysis over the last decade has led to many useful techniques for studying samples of trajectories (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). Typically, a crucial first step is needed before such analyses are possible: the trajectories need to be reconstructed on a fine grid of equally spaced time points (if they are not already in such a form). Methods for reconstructing trajectories in this way have been studied using kernel smoothing (Ferraty and Vieu, 2006), smoothing splines (Ramsay and Silverman, 2005), local linear smoothing (Hall, Müller, and Wang, 2006), mixed effects models (James, Hastie, and Sugar, 2000; Rice and Wu, 2000), and principal components analysis through conditional expectations (Yao, Müller, and Wang, 2005 a,b).

In this article we study the problem of reconstructing *gradients* of trajectories on the basis of sparse (or widely separated) observations. The proposed approach is specifically designed for reconstructing growth velocities from longitudinal observation of childhood developmental indices, e.g., height, weight, BMI, head circumference, or measures of brain maturity obtained via fMRI (Dosenbach et al., 2010). Growth velocities based on such indices play a central role in life course epidemiology, often providing fundamental indicators of prenatal or childhood development that are related to adult health outcomes (Barker et al., 2005).

Repeated measurements of childhood developmental indices may be available on most subjects in a study, but usually only *sparse* temporal sampling is feasible (McKeague et al., 2011). It can thus be challenging to gain a detailed understanding of growth patterns. Moreover, the problem is exacerbated by the presence of large fluctuations in growth velocity during early infancy, and high variability between subjects. In addition, although the patterns of examination times vary among children, they tend to cluster around "nominal" ages (e.g., birthdays), so there can be large gaps without data. We call this *regular* sparsity, in contrast to *irregular* sparsity in which the observation times for each individual are widely separated but become dense when merged over all subjects.

The problem of reconstructing gradients under irregular sparsity (even with only one observation time per trajectory) has recently been studied by Liu and Müller (2009). In their approach, the best linear predictor of the gradient is estimated (assuming Gaussian trajectories) in terms of estimated functional principal component scores. The accuracy of the reconstruction depends on how well each individual gradient can be represented in terms of a small number of estimated principal component functions. This in turn requires an accurate estimate of the covariance kernel of the trajectories, which is not possible in the case of regular sparsity.

Numerical analysis methods can be used to reconstruct gradients using only individual level data in the case of regular sparsity. For example, difference quotients between observation times provide simple approximate gradients, but these estimates are piecewise constant and would not be suitable for use in functional data analysis unless the observation times are dense. Spline smoothing to approximate the gradient of the trajectory over a fine grid is recommended by Ramsay and

© 2013, The International Biometric Society

Silverman (2005). More generally, methods of numerical differentiation, including spline smoothing, are an integral part of the extensive literature on ill-posed inverse problems for linear operator equations. In this literature, the observation times are usually viewed as becoming dense (for the purpose of showing convergence), see Kirsch (1996); in particular, the assumption of asymptotically dense observation times plays a key role in the study of penalized least squares estimation and cross-validation (Nashed and Wahba, 1974; Wahba, 1977).

In this article we develop a flexible Bayesian approach to reconstructing gradients, focusing on the regular sparsity case. The prior gradient is specified by a general multivariate normal distribution at n fixed observation times, and a (conditional) tied-down Brownian motion between the observation times. This leads to a simple and explicit representation of the posterior distribution of the gradient in terms of the prior mean and the prior precision matrix at the observation times. Based on a sample of subjects, the nonparametric empirical Bayes method along with cross-validation is used to specify the hyperparameters in the prior, with the prior precision matrix estimated by constrained ℓ_1 minimization (Cai, Liu, and Luo, 2011). An important aspect of the proposed approach is that the reconstructed gradients can be computed rapidly over a fine grid, and then used directly as input into existing software, without the need for sophisticated smoothing techniques. In addition, our approach furnishes ways of assessing the errors in the reconstruction (using credible intervals around the posterior mean) and of assessing uncertainties in the conclusions of standard functional data analyses that use the reconstructed gradients as predictors (e.g., using repeated draws from the posterior distribution in a sensitivity analysis).

For background and an introduction to empirical Bayes methods we refer the interested reader to Efron (2010). Previous work on the use of such methods in the setting of growth curve modeling include reconstructing individual growth velocity curves from parametric growth models (Shohoji et al., 1991), and nonparametric testing for differences in growth patterns between groups of individuals (Barry, 1995). A nonparametric hierarchical-Bayesian growth curve model for reconstructing individual growth curves has also been developed (Arjas, Liu, and Maglaperidze, 1997), but requires the use of computationally intensive Markov chain Monte Carlo methods (MCMC) and may not be suitable for exploratory analyses.

As already mentioned, our motivation for developing the proposed Bayesian reconstruction method comes from the problem of carrying out functional data analysis for growth velocities given measurements of some developmental index at various ages. The first panel of Figure 1 shows cubic-spline-interpolated growth curves for 10 children based on their height (or length) measurements at birth, 4, 8, and 12 months, and 3, 4, and 7 years. There are no data to fill in the gaps between these observation times. The corresponding growth velocities, obtained by differentiating the cubic splines (Ramsay and Silverman, 2005), are displayed in the second panel. Unfortunately, however, such growth rate curves are unsuitable surrogates for the actual growth rates because artifacts of the spline interpolation emerge as the dominant features, and there is no justification for ignoring all random variation between observation times. Moreover, it is not easy to see how to quantify the error involved in such reconstructions.

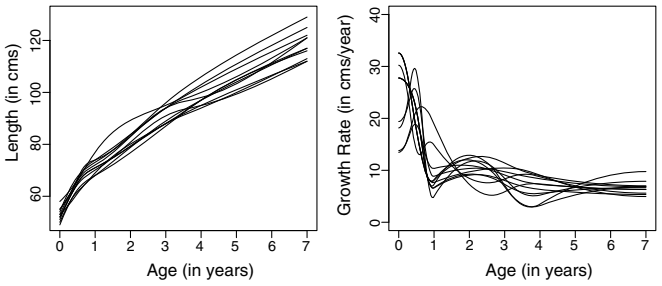

Figure 1. Growth curves based on natural cubic spline interpolation between the observation times (left panel), and the corresponding growth velocities (right panel).

Our proposed reconstruction method, as developed in Section 2. provides a way around this problem. Simulation and real growth data examples are used to illustrate the performance of the proposed method in Sections 3 and 4. In Section 5 we compare our approach with the popular method of analyzing growth trajectories via latent variable models. Proofs of the main results are provided in the Appendix. We have developed an R package `growthrate` implementing the proposed reconstruction method (López-Pintado and McKeague, 2011); this package is available on the CRAN archive, and includes the real data set used in Section 4.

2. Gradients of Sparsely Observed Trajectories

In this section we develop the proposed Bayesian approach to recovering gradients. Explicit formulae for the posterior mean and covariance kernel of the gradients are provided. An empirical Bayes approach to estimating the hyperparameters in the prior is also developed.

2.1 *Posterior Gradients*

We first consider in detail how to reconstruct the gradient for a single subject. In the growth velocity context, the observation times will typically vary slightly across the sample, but will be clustered around certain nominal ages. Let the observation times for the specific individual be $0 = t_1 < t_2 < \cdots < t_n = T$, and assume that the endpoints of the time interval over which the reconstruction is needed are included.

The statistical problem is to estimate the whole growth velocity curve $X = \{X(t), 0 \le t \le T\}$ from data on its integral (i.e., growth) over the gaps between the observation times. Equivalently, we observe

$$y_i = \frac{1}{\Delta_i} \int_{t_i}^{t_{i+1}} X(s)\, ds,\ i = 1, \ldots, n-1, \tag{1}$$

where the $\Delta_i = t_{i+1} - t_i$ are the lengths of the intervals between the observation times. Here y_i is the one-sided difference quotient estimate of $X(t)$ for $t \in [t_i, t_{i+1}]$, as commonly used in numerical differentiation. Reconstructing X based on such data is an ill-posed inverse problem in the sense that no unique solution exists, so some type of regularization is needed to produce a unique solution. When the observation times are equally spaced, the one-sided difference quotient can

be derived as the X minimizing $\|X\|_{L^2}$ under the constraint (1), see Kirsch (1996, p. 97).

A more sophisticated approach is to take into account the proximity of t_i to neighboring observation times t_{i-1} and t_{i+1}, and estimate $X(t_i)$ by the second-order difference quotient

$$\overline{y}_i = w_i y_{i-1} + (1 - w_i) y_i,$$

where $w_i = \Delta_i/(\Delta_{i-1} + \Delta_i)$, for $i = 2, \ldots, n-1$. As we shall see, the building blocks needed to construct the proposed Bayes estimator consist of the y_i and the n-vector

$$\mathbf{y} = (y_1, \overline{y}_2, \ldots, \overline{y}_{n-1}, y_{n-1})^T.$$

The central problem is to specify a flexible class of prior distributions for X in such a way that is tractable to find its posterior distribution. An unusual feature of our problem, however, is that a *direct* approach via Bayes formula does not work: the conditional distribution of y_i given X is degenerate, so there is no common dominating measure for all values of X (i.e., there is no full likelihood). Our way around this difficulty is to first find the marginal posterior distribution of X *restricted to the observation times*, for which the usual Bayes formula applies, and then show that this leads to a full posterior distribution. This approach turns out to be tractable when we specify the prior using the following hierarchical model: X has a multivariate normal distribution at the observation times, and is a tied-down Brownian motion in the gaps between observation times (conditional on X at those time points). This provides a fully coherent prior distribution of X. In some cases the prior of X can be specified unconditionally (as with the shifted Brownian motion discussed in Section 2.4), but in general the hierarchical specification we use is simpler and provides greater flexibility.

More precisely, we specify the prior on X as follows:

1. At the observation times:
$$\mathbf{X} \equiv (X(t_1), \ldots, X(t_n))^T \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$
where $\boldsymbol{\Sigma}_0$ is nonsingular.
2. The conditional distribution of X given $\mathbf{X}$ is a tied-down Brownian motion with given infinitessimal variance $\sigma^2 > 0$.

In the growth velocity setting, the observation times are typically chosen to concentrate data collection in periods of high variability (e.g., the first year of life), so it is natural that the prior should reflect such information. Moreover, allowing an arbitrary (multivariate normal) prior at the observation times provides flexibility that would not be possible using a Brownian motion prior for the whole of X. In addition, the availability of data at these time points makes is possible to specify the hyperparameters in the multivariate normal (as we see later), which is crucial for practical implementation of our approach.

We now state our main result giving the posterior distribution of X. In particular, the result shows that the posterior mean takes the computationally tractable form of a quadratic spline with knots at the observation times. The posterior mean is the best linear predictor of X, providing the optimal reconstruction of X in the sense of mean-squared error (MSE).

THEOREM 1. *The posterior distribution of X is Gaussian with mean*

$$\widehat{\mu}(t) = \widehat{\mu}_i + [\widehat{\mu}_{i+1} - \widehat{\mu}_i](t - t_i)/\Delta_i + 6(t - t_i)(t_{i+1} - t)\left[y_i - (\widehat{\mu}_i + \widehat{\mu}_{i+1})/2\right]/\Delta_i^2$$

for $t \in [t_i, t_{i+1}]$, where

$$\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_i) = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{D}\mathbf{y})$$

is the posterior mean of $\mathbf{X}$, and

$$\mathbf{Q} = \frac{3}{\sigma^2}\begin{pmatrix} \frac{1}{\Delta_1} & \frac{1}{\Delta_1} & 0 & \cdots & 0 \\ \frac{1}{\Delta_1} & \frac{1}{\Delta_1} + \frac{1}{\Delta_2} & \frac{1}{\Delta_2} & \ddots & \vdots \\ 0 & \frac{1}{\Delta_2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{\Delta_{n-1}} \\ 0 & \cdots & 0 & \frac{1}{\Delta_{n-1}} & \frac{1}{\Delta_{n-1}} \end{pmatrix},$$

$$\mathbf{D} = \frac{6}{\sigma^2}\mathrm{diag}\left(\frac{1}{\Delta_1}, \ldots, \frac{1}{\Delta_{i-1}} + \frac{1}{\Delta_i}, \ldots, \frac{1}{\Delta_{n-1}}\right).$$

The posterior covariance kernel of X is $\widehat{K} = \sigma^2\widetilde{K} + K^*$, where

$$\widetilde{K}(s,t) = (s \wedge t - t_i) - (s - t_i)(t - t_i)/\Delta_i - 3(s - t_i)(t - t_i)(t_{i+1} - s)(t_{i+1} - t)/\Delta_i^3$$

for $s, t \in [t_i, t_{i+1})$, with $\widetilde{K}(s,t) = 0$ if s and t are in disjoint intervals;

$$K^*(s,t) = (a_k(s), b_k(s))\begin{bmatrix} \widehat{\Sigma}_{kl} & \widehat{\Sigma}_{k,l+1} \\ \widehat{\Sigma}_{k+1,l} & \widehat{\Sigma}_{k+1,l+1} \end{bmatrix}(a_l(t), b_l(t))^T$$

for $s \in [t_k, t_{k+1})$ and $t \in [t_l, t_{l+1})$, with $k, l = 1, \ldots, n-1$, where

$$a_i(t) = 1 - (t - t_i)/\Delta_i - 3(t - t_i)(t_{i+1} - t)/\Delta_i^2,$$

$$b_i(t) = (t - t_i)/\Delta_i - 3(t - t_i)(t_{i+1} - t)/\Delta_i^2,$$

for $t \in [t_i, t_{i+1})$, $i = 1, \ldots, n-1$, and

$$\widehat{\boldsymbol{\Sigma}} = (\widehat{\Sigma}_{ij}) = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q})^{-1}$$

is the posterior covariance matrix of $\mathbf{X}$.

Remarks:

1. The posterior mean of X provided by Theorem 1 has a simple and explicit form that can be computed rapidly once the hyperparameters in the prior (namely $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, and σ^2) are provided. We discuss various ways of specifying the hyperparameters in the next section.
2. The infinitessimal variance σ^2 can be regarded as a smoothing parameter, and plays the role of a time-scale, cf. the adaptive Bayesian estimation procedure of van der Vaart and van Zanten (2009) based on Gaussian random field priors with an unknown time-scale. When $\sigma^2 \to \infty$, the posterior distribution of $\mathbf{X}$ converges to its prior distribution, and between observation times the posterior variance of X tends to infinity.

3. The matrix $\mathbf{Q}$ represents the posterior precision (at the observation times) corresponding to a noninformative (improper) prior, and it is singular, reflecting the fact that we are dealing with n parameters to be estimated from $n-1$ observations $y_1, \ldots, y_{n-1}$. The problem is ill-posed, but an "informative" prior provides regularization: the posterior precision matrix $\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q}$ is nonsingular.
4. In the special case that the prior distribution of $\mathbf{X}$ is a Gaussian Markov random field, i.e., nonneighboring components are conditionally independent given the rest (Rue and Held, 2005), its posterior distribution is also a Gaussian Markov random field. That is, since $\mathbf{Q}$ is tridiagonal, whenever the prior precision matrix $\boldsymbol{\Sigma}_0^{-1}$ is tridiagonal, so is the posterior precision matrix $\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q}$. Tridiagonal matrices often arise in inverse problems, and efficient algorithms for computation of their inverses and eigenvalues are available.
5. In typical Bayesian settings, the information in the data tends to swamp the prior information as the sample size increases. It can often be shown in such settings that the posterior contracts to the true parameter given that it is in the support of the prior distribution (Ghosal and van der Vaart, 2007; Knapik, van der Vaart, and van Zanten, 2011), but such results rely on the existence of a full likelihood and are not applicable in our setting.

2.2 *Specifying the Hyperparameters*

Suppose that we are given data on $\mathbf{y}$ for a sample of N individuals, each having the same fixed set of observation times $t_1, \ldots, t_n$. How can we use such data to specify the hyperparameters?

The hyperparameters are not identifiable in general, even if the prior distribution is correctly specified. To see this, note that the (Gaussian) marginal distribution of $\mathbf{y}$ is determined by $p = n - 1 + n(n-1)/2$ means and covariances, but there are $n + n(n+1)/2 + 1 > p$ hyperparameters in the prior. Although it is possible to identify these hyperparameters by imposing extra structure on $\boldsymbol{\Sigma}_0$ and using a *parametric* empirical Bayes approach (as we show in the next section), the presence of prior misspecification would be a serious issue for applications. Another possibility would be to define a flexible higher level prior for the hyperparameters, but this would again require the use of computationally intensive methods (MCMC).

Instead we adopt the following nonparametric empirical Bayes approach. The prior mean $\boldsymbol{\mu}_0$ is naturally specified by the sample mean of $\mathbf{y}$, and this does not require the prior to be correctly specified. The corresponding sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_N$, however, is singular [having rank $\min(N, n-1) < n$] and hence unstable for estimating $\boldsymbol{\Sigma}_0$, and cannot specify the posterior distribution which depends on the prior precision matrix $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0^{-1}$. We use the constrained ℓ_1 minimization method of sparse precision matrix estimation (CLIME) recently developed by Cai et al. (2011): $\boldsymbol{\Omega}_0$ is specified as the (appropriately symmetrized) solution of

$$\min ||\boldsymbol{\Omega}||_1 \quad \text{subject to} \quad |\widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Omega} - \boldsymbol{I}|_\infty \le \lambda_N, \quad \boldsymbol{\Omega} \in \mathbb{R}^{n \times n},$$

where the tuning parameter λ_N is selected by fivefold cross validation using the likelihood loss function.

The infinitessimal variance σ^2, or equivalently σ, is selected by a form of cross validation introduced by Wahba (1977). The prediction error based on leaving-out an interior observation time (t_{i+1}, for some fixed $i = 1, \ldots, n-2$) is given by

$$\mathrm{CV}(\sigma) = \frac{1}{N} \sum_{j=1}^{N} E_{ij} \left[y_{ij} - \frac{1}{\Delta_i} \int_{t_i}^{t_{i+1}} \widehat{X}_j^{-(i+1)}(s)\, ds \right]^2,$$

where j indexes the subjects, the expectation E_{ij} is over draws $\widehat{X}_j^{-(i+1)}(\cdot)$ from the posterior distribution of $X_j(\cdot)$ based on the reduced data with t_{i+1} removed; here $\boldsymbol{\mu}_0$ and $\boldsymbol{\Omega}_0$ are specified as above, except the $(i+1)$th component is not used. This expression can be written explicitly using the bias-variance decomposition as

$$\begin{aligned}\mathrm{CV}(\sigma) = \frac{1}{N} \sum_{j=1}^{N} &\left[y_{ij} - \frac{1}{\Delta_i} \int_{t_i}^{t_{i+1}} \widehat{\mu}_j^{-(i+1)}(s)\, ds \right]^2 \\ &+ \frac{1}{\Delta_i^2} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \widehat{K}^{-(i+1)}(s,u)\, ds\, du,\end{aligned}$$

in terms of the mean and covariance kernel of $\widehat{X}_j^{-(i+1)}(\cdot)$ that are available from Theorem 1; note that the covariance kernel only depends on the observation times so it is not indexed by j.

2.3 *Variations in Observation Times among Subjects*

In this section we discuss how our approach to specifying the hyperparameters can be adapted to handle situations in which the observation times vary among subjects. In the applications we have in mind, most observation times tend to be close to "nominal" observation times $\{t_i, i = 1, \ldots, n\}$, so it is reasonable to use these as a first approximation. That is, in terms of the nominal observation times, and using the procedure described above, we find initial estimates $\widehat{\mu}_j(\cdot)$ for all subjects, and a value of σ. Then, for the purpose of specifying $\boldsymbol{\mu}_0$ and $\boldsymbol{\Omega}_0$ in a way that is tailored to the observation times of the kth subject, we adjust the data on the other subjects to become

$$y_{ij}^{(k)} = \frac{1}{\Delta_i^{(k)}} \int_{t_{ik}}^{t_{i+1,k}} \widehat{\mu}_j(s)\, ds, \;\; i = 1, \ldots, n_k, \quad j \ne k,$$

where $\{t_{ik}, i = 1, \ldots, n_k\}$ are the actual observation times for the kth subject and $\Delta_i^{(k)} = t_{i+1,k} - t_{ik}$. The final estimate $\widehat{\mu}_k(\cdot)$ is then calculated by applying our formula for $\widehat{\mu}(\cdot)$ to the data on the kth subject, with the hyperparameters estimated from the adjusted data (thus borrowing strength from the whole sample). The adjustment has no effect when the observation times agree with their nominal values, and small perturbations around the nominal values would have little effect on the reconstructed gradients.

A computationally simpler approach, that is essentially equivalent to what we just described, is to restrict the posterior covariance kernel and mean based on the nominal observation times to the actual observation times for each given subject, thus directly obtaining suitable hyperparameters across the whole sample that adjust for any changes from the nominal observation times; this is the approach implemented in the `growthrate` package.

2.4 *Example: Shifted Brownian Motion Prior*

Shifted Brownian motion priors have been used in various nonparametric Bayesian settings in recent years (van der Vaart and van Zanten, 2008a, b) and provide a simple illustration of Theorem 2.1. Suppose the prior distribution of X at the observation times (i.e., the prior of $\mathbf{X}$) is specified so that

$$X(t_i) = \mu_i + \sigma_1 Z + \gamma B(t_i), \quad i = 1, \ldots, n, \tag{2}$$

where $Z \sim N(0,1)$, B is an independent standard Brownian motion, $\sigma_1 > 0$, $\gamma > 0$, and the prior mean $\boldsymbol{\mu}_0 = (\mu_1, \ldots, \mu_n)^T$ as before. In particular, when $\gamma = \sigma$ the prior distribution for the entire trajectory X takes the form of a shifted Brownian motion.

Under (2), the prior covariance matrix at the observation times has (i,j)th entry

$$(\boldsymbol{\Sigma}_0)_{ij} = \sigma_1^2 + \gamma^2 \min(t_i, t_j). \tag{3}$$

The prior precision matrix has a simple (tridiagonal) form similar to $\mathbf{Q}$, namely

$$\boldsymbol{\Sigma}_0^{-1} = \frac{1}{\gamma^2} \begin{pmatrix} \frac{\gamma^2}{\sigma_1^2} + \frac{1}{\Delta_1} & -\frac{1}{\Delta_1} & 0 & \cdots & 0 \\ -\frac{1}{\Delta_1} & \frac{1}{\Delta_1} + \frac{1}{\Delta_2} & -\frac{1}{\Delta_2} & \ddots & \vdots \\ 0 & -\frac{1}{\Delta_2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{1}{\Delta_{n-1}} \\ 0 & \cdots & 0 & -\frac{1}{\Delta_{n-1}} & \frac{1}{\Delta_{n-1}} \end{pmatrix},$$

cf. Rue and Held (2005, p. 99). In the special case that $\gamma^2 = \sigma^2/3$ the posterior covariance matrix becomes diagonal:

$$\widehat{\boldsymbol{\Sigma}} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q}\right)^{-1} = \frac{\sigma^2}{6} \mathrm{diag}\left(\Delta_{i-1} w_i, i = 1, \ldots, n\right),$$

where $w_1 \equiv \Delta_1/(\Delta_0 + \Delta_1)$, $\Delta_0 = 6\sigma_1^2/\sigma^2$, $w_n \equiv 1$, and the other w_i are defined as in Section 2.1. The components of the posterior mean at the observation times are then given by

$$\widehat{\mu}_i = \begin{cases} (1-w_1)y_1 + w_1(\mu_1 + \mu_2)/2, & i = 1; \\ \overline{y}_i + [\mu_i - (w_i \mu_{i-1} & i = 2, \ldots, n-1; \\ \quad + (1-w_i)\mu_{i+1})]/2, & \\ y_{n-1} + (\mu_n - \mu_{n-1})/2, & i = n. \end{cases}$$

It can then be seen that $\widehat{\mu}$ provides a uniformly consistent estimator of X in the numerical analysis sense: if $\mu_i = \mu(t_i)$, where $\mu(\cdot)$ is a fixed continuous function, and $\max_{i=1,\ldots,n-1} \Delta_i \to 0$, then $\widehat{\mu}(t) \to X(t)$ uniformly in t for any continuous X.

A parametric empirical Bayes approach to specifying the hyperparameters can be developed in this setting. Conditioning on $X(t_i)$ and $X(t_{i+1})$ yields $2(n-1)$ estimating equations involving means and second moments:

$$\begin{aligned} Ey_i &= (\mu_i + \mu_{i+1})/2, \\ Ey_i^2 &= (\mu_i + \mu_{i+1})^2/4 + \sigma^2 \Delta_i/12 + (\sigma_i^2 + \sigma_{i+1}^2 + 2\sigma_{i,i+1})/4, \end{aligned}$$

$i = 1, \ldots, n-1$, where $\sigma_{i,i+1}$ is the prior covariance of $X(t_i)$ and $X(t_{i+1})$, and σ_i^2 is the prior variance of $X(t_i)$. Under the shifted Brownian motion model (2), the prior distribution has only $n+3$ parameters ($\mu_1, \ldots, \mu_n$, σ_1^2, γ^2 and σ^2), and the second-moment estimating equation simplifies to

$$Ey_i^2 = (\mu_i + \mu_{i+1})^2/4 + \sigma^2 \Delta_i/12 + \sigma_1^2 + \gamma^2(t_i + \Delta_i/4).$$

Another $(n-1)(n-2)/2$ estimating equations are obtained from the covariances of the y_i:

$$Ey_i y_j = (\mu_i + \mu_{i+1})(\mu_j + \mu_{j+1})/4 + \sigma_1^2 + \gamma^2(t_i + \Delta_i/2),$$

for $i = 1, \ldots, n-2, j = i+1, \ldots, n-1$, provided $n \geq 3$. The marginal distribution of the data is Gaussian with mean and covariance only depending on the prior means μ_i through the sums $\mu_i + \mu_{i+1}, i = 1, \ldots, n-1$, so these means are not identifiable unless one of them (say μ_1) is known. Once μ_1 is given (or specified say using the sample mean of y_1), all the other parameters in the shifted Brownian motion prior are identifiable.

In view of known convergence-rate results for the CLIME estimator (Cai et al., 2011), it may be possible to extend the consistency result shown above to the general setting, to the effect that the empirical Bayes version of each $\widehat{\mu}(t)$ converges uniformly to $X(t)$ as $\max_{i=1,\ldots,n-1} \Delta_i \to 0$ and $N \to \infty$. However, $\widehat{\mu}(t)$ is is not analytically tractable in general (only in the special case discussed in this section), so this would be a challenging problem.

3. Simulation Study

In this section we report the results of a simulation study designed to assess the performance of $\widehat{\mu}(t)$ as a method of estimating X. In order to calibrate $\widehat{\mu}(t)$, the prior mean and precision matrix are specified from the data as in Section 2.2. We also examine the performance of the cross validation method for choosing σ.

The simulation model for generating the underlying X is defined in a parallel fashion to the prior:

1. At the observation times, $\mathbf{X} \equiv (X(t_1), \ldots, X(t_n))^T$ is a zero-mean stationary Gaussian Markov random field with covariance matrix having (i,j)th entry $e^{-\alpha|t_i - t_j|}$, where $\alpha > 0$.
2. The conditional distribution of X given $\mathbf{X}$ is a tied-down fractional Brownian motion (fBm) with Hurst exponent $0 < H < 1$.

The tied-down fBm used here is represented (conditionally on $\mathbf{X}$) as in (A.3), except that $B_i^0(t) = B_i(t) - tB_i(1)$, $t \in [0,1]$, where the B_i are independent standard fBms, and $\sigma = 1$. The Hurst exponent H is a measure of the smoothness of the sample paths between the observation times: $H = 1/2$ gives standard tied-down Brownian motion and agrees with the prior (provided $\sigma = 1$); we also consider the cases $H = 0.7$ and 0.9 to give examples with much smoother sample path behavior than Brownian motion. We considered two values of the simulation model parameter: $\alpha = 3$ and 6, representing "high" and "low" levels of correlation in $\mathbf{X}$, respectively. The sample size is fixed at $N = 100$, and we consider $n = 5$ and $n = 10$ equispaced observation times on the interval $[0,1]$ (that is, $T = 1$).

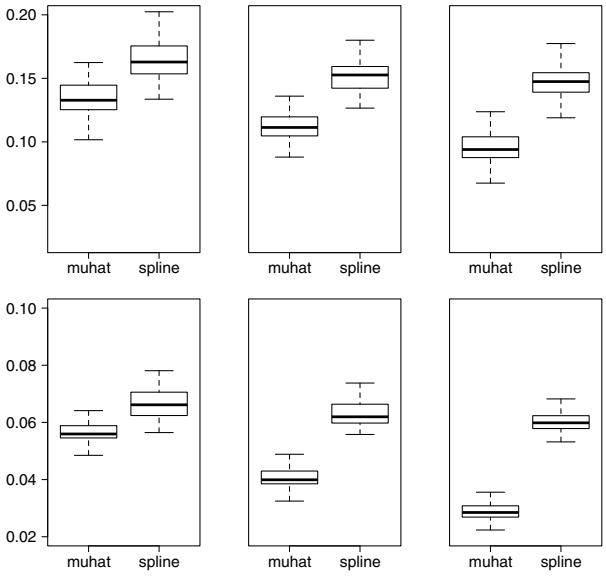

Figure 2. Simulation model with $\alpha = 3$. Boxplots comparing the MSE of the proposed estimator $\widehat{\mu}(\cdot)$ and the spline estimator of X at the observation times; $n = 5$ (first row), $n = 10$ (second row), $H = 0.5$ (first column), $H = 0.7$ (second column), $H = 0.9$ (third column).

Figure 2 shows boxplots comparing the MSE of our approach with the MSE of the spline interpolation approach described in the Introduction. The boxplots are based on 50 independent samples, and setting $\alpha = 3$ in the simulation model. Web Figure 1 shows the corresponding boxplots for $\alpha = 6$, and the results are very similar. Here the MSE of $\widehat{\mu}(\cdot)$ is defined by

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_j(t_i) - X_j(t_i))^2,$$

where $\widehat{\mu}_j(\cdot)$ is the calibrated posterior mean of $X_j(\cdot)$ with $\sigma = 1$. In all cases, $\widehat{\mu}(\cdot)$ has a smaller median MSE than the spline estimator, and in some cases the reduction is more than 50% (namely for $n = 10$ and $H = 0.9$). Note that the improvement of $\widehat{\mu}(\cdot)$ over the spline method increases with H (and also with n). This indicates that $\widehat{\mu}(\cdot)$ is robust to departures from the prior that involve smoother trajectories X. Similar results (not shown) can be obtained in terms of the mean absolute deviation.

We applied the cross validation method for choosing σ to a single sample generated by the above simulation model for $\alpha = 3$, $H = 1/2$ and $n = 10$. Figure 3 shows plots of cross validation error $\text{CV}(\sigma)$ based on removing three of the interior observation times, as well as averaging $\text{CV}(\sigma)$ with each interior point successively removed. In all cases, the minimum is located close to the true value of $\sigma = 1$, so we calibrated $\widehat{\mu}(\cdot)$ using $\sigma = 1$ in all the simulations reported above.

We have also done extensive simulations (not shown) based on shifted Brownian motion priors, and found that the two competing approaches have comparable MSE; this is not surprising, because, as can be seen from the explicit form in Section 2.4, $\widehat{\mu}(\cdot)$ is very close to the spline estimator in this case. In addition, we found that measurement error in the observations $\mathbf{y}$ has little affect on the accuracy of the reconstructions.

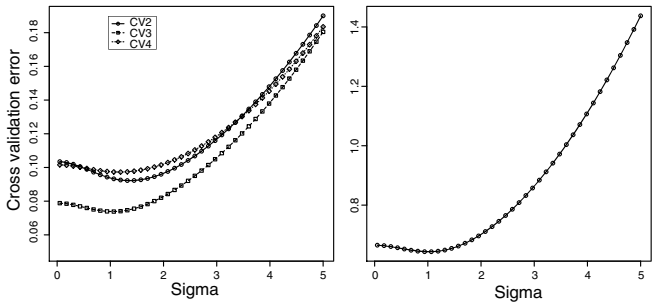

Figure 3. Simulation model with $\alpha = 3$, $n = 10$, $H = 0.5$. The cross validation error $\text{CV}(\sigma)$ over a fine grid of values of σ, based on removing each of the first three interior observation times (left panel), and based on averaging $\text{CV}(\sigma)$ with all interior points successively removed (right panel).

4. Growth Velocity Curves

In this section we illustrate our approach using data from the Collaborative Perinatal Project (CPP). This was an NIH study of prenatal and familial antecedents of childhood growth and development conducted during 1959–1974 at 12 medical centers across the United States. There were approximately 58,000 study pregnancies, mothers being examined during pregnancy, labor and delivery. The children were given neonatal examinations and follow-up examinations at 4, 8, and 12 months, and 3, 4, and 7 years. We restrict attention to the subsample of girls having birthweight 1500–4000 g, gestational age 37–42 weeks, nonbreast-fed, maternal age 20–40 years, the mother did not smoke during pregnancy, and for whom complete data on height (at these ages) and all the covariates are available. This gave a sample of size $N = 532$. The data are included in the `growthrate` package, and additional discussion can be found in McKeague et al. (2011).

Figure 4 shows the cross validation error based on removing each of the five interior observation times (1/3, 2/3, 1, 3, 4 years) in turn. Note that, due to the nonequispaced observation times, the various curves are ordered according to which time point is removed. The curves suggest that a choice of σ in the range 1–3 is reasonable, although, since cross validation tends to overfit, the lower end of this range might be preferable.

Figure 5 gives the reconstructed growth velocity curves for two of these subjects, and for three choices of σ. The choice $\sigma = 1$ produces very tight bands, which may be unrealistic because the growth rate is unlikely to have sharp bends at the observation times; the more conservative choices $\sigma = 2$ and 3 allow enough flexibility in this regard and appear to be more reasonable. Notice that the $\sigma = 2$ and $\sigma = 3$ bands bulge between observation times (and this is especially noticeable in the last observation time interval), which is a desirable feature since we would expect greater precision in the estimates close to the observation times. Plots of the posterior mean growth velocity curves for a random subset of 200 subjects based on $\sigma = 1, 2, 3$ are provided in Web Figure 2.

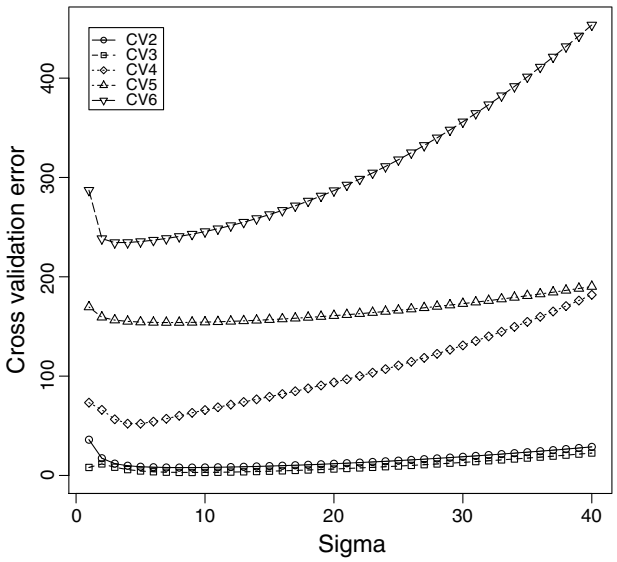

Figure 4. The cross validation error $\mathrm{CV}(\sigma)$ over a fine grid of values of σ in the growth velocity example based on removing each of the five interior observation times (1/3, 2/3, 1, 3, 4 years), CV2, . . . , CV6, respectively.

5. Discussion

The standard approach to the longitudinal data analysis of growth trajectories is via mixed-effects or latent variable models (Bollen and Curran, 2006; Wu and Zhang, 2006), and in this section we compare it with the proposed approach.

The hierarchical prior we use for growth velocity can be seen as a flexible nonparametric model for an *infinite*-dimensional latent process. This contrasts with the standard mixed model approach of representing a trajectory by a polynomial, allowing additive uncorrelated random disturbances at the observation times, and treating some or all of the coefficients as random, i.e., a *finite*-dimensional latent structure. In contrast, our approach does not model within-subject (within-curve) and between-subject (between-curves) variations *separately*, as in mixed-effects models. Rather, the between-curve variation is represented by the prior—a random effect specified in terms of a tied-down Brownian motion process; this prior also suffices to provide the within-subject variation that in mixed models is typically provided by the uncorrelated disturbance terms, or white noise. At the infinitessimal scale, Brownian motion is white noise, so in a sense the two approaches are parallel in their handling of within-subject variation, but the advantage of using a single prior is that the full power of the Bayesian approach comes into play. In particular, this allows a closed-form calculation of the estimated growth velocity curves, without the need for sophisticated numerical methods that play a role in fitting complex mixed models. In summary, although mixed models provide an array of effective techniques for understanding trajectories, and longitudinal data more generally, in the context of growth velocity reconstruction we believe that the proposed approach can offer some advantages: greater flexibility and computational efficiency.

A referee raised the question of whether the proposed approach is sensitive to outliers. The sample mean of $\mathbf{y}$ could indeed be a poor estimate of the prior mean $\boldsymbol{\mu}_0$ if there are outliers. The same issue could be raised about the CLIME estimator we use for $\boldsymbol{\Sigma}_0^{-1}$ (as with any method based on a sample mean or sample covariance). One recourse would be to use robust estimators for $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$, but, as our reconstruction is linear in the data $\mathbf{y}$, it would still be sensitive to outliers. A better recourse would be to carefully prescreen

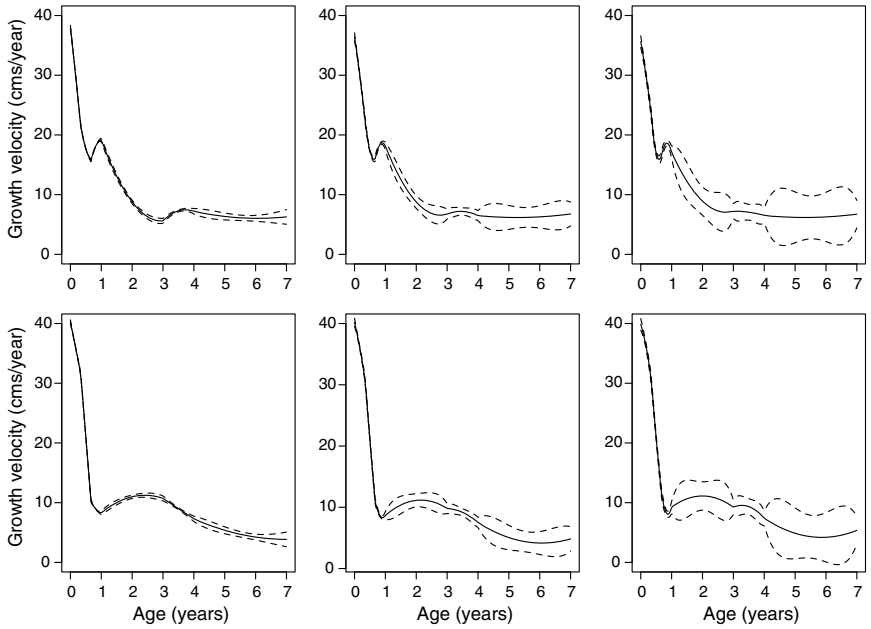

Figure 5. Reconstructed growth velocity curves for two subjects; posterior mean (solid line), pointwise 95% credible intervals (dashed lines) based on $\sigma = 1, 2, 3$ for the first, second, and third plots in each row, respectively; for one subject in the first row, and a second subject in the second row.

the data for outliers before using the method. In the real data examples (of childhood growth curves) that we have studied, it has not been necessary to remove outliers.

6. Supplementary Materials

Web Figures referenced in Sections 3 and 4 and R code implementing the simulation study in Section 3 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The work of Ian McKeague was supported by NIH Grant R01GM095722-01. The authors thank Russell Millar for his helpful suggestions.

References

Arjas, E., Liu, L., and Maglaperidze, N. (1997). Prediction of growth: A hierarchical Bayesian approach. *Biometrical Journal* **39,** 741–759.

Barker, D. J. P., Osmond, C., Forsén, T. J., Kajantie, E., and Eriksson, J. G. (2005). Trajectories of growth among children who have coronary events as adults. *The New England Journal of Medicine* **353,** 1802–1809.

Barry, D. (1995). A Bayesian model for growth curve analysis. *Biometrics* **51,** 639–655.

Bollen, K. A. and Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley Series in Probability and Statistics. Hoboken: Wiley Interscience.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l_l minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106,** 594–607.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R., Barch, D. M., Petersen, S. E., and Schlaggar, B. L. (2010). Prediction of individual brain maturity using fMRI. *Science* **329,** 1358–1361.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge: Cambridge University Press.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *The Annals of Statistics* **35,** 192–223.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal components methods for functional and longitudinal data analysis. *The Annals of Statistics* **34,** 1493–1517.

James, G., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87,** 587–602.

Kirsch, A. (1996). *An Introduction to the Mathematical Theory of Inverse Problems*. New York: Springer.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics* **39,** 2626–2657.

Liu, B. and Müller, H. G. (2009). Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *Journal of the American Statistical Association* **104,** 704–717.

López-Pintado, S. and McKeague, I. W. (2011). Growthrate: Bayesian reconstruction of growth velocity. R package version 1.0, `http://cran.r-project.org/package=growthrate`.

McKeague, I. W., López-Pintado, S., Hallin, M., and Šiman, M. (2011). Analyzing growth trajectories. *Journal of Developmental Origins of Health and Disease* **2,** 322–329.

Nashed, M. Z. and Wahba, G. (1974). Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. *Mathematics of Computation* **28,** 69–80.

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

Rice, J. and Wu, C. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57,** 253–259.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Boca Raton: Chapman & Hall/CRC.

Shohoji, T., Kanefuji, K., Sumiya, T., and Qin, T. (1991). A prediction of individual growth of height according to an empirical Bayesian approach. *Annals of the Institute of Statistical Mathematics* **43,** 607–619.

van der Vaart, A. W. and van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36,** 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, vol. **3**, 200–222. Beachwood, OH: Institute of Mathematical Statistics.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics* **37,** 2655–2675.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis* **14,** 651–667.

Wu, H. and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. Hoboken: Wiley Interscience.

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

Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33,** 2873–2903.

Received August 2011. *Revised October* 2012. *Accepted October* 2012.

Appendix

Proof of Theorem 1. The first part of the proof is to determine the posterior distribution of $\mathbf{X}$. In terms of its prior density $p(\mathbf{x})$, the posterior density of $\mathbf{X}$ is given by Bayes formula as

$$p(\mathbf{x}|y_1,\ldots,y_{n-1}) \propto p(y_1,\ldots,y_{n-1}|\mathbf{x})p(\mathbf{x}) \propto \left[\prod_{i=1}^{n-1} p(y_i|x_i,x_{i+1})\right] p(\mathbf{x}), \tag{A.1}$$

where $\mathbf{x} = (x_1,\ldots,x_n)^T$, the observation in the ith time interval is y_i, and we have used the independent increments property of Brownian motion to separate the terms. Also, using standard properties of Brownian motion, it can be shown that $p(y_i|x_i,x_{i+1})$ is normal with mean $(x_i + x_{i+1})/2$ and variance $\sigma^2\Delta_i/12$. Apart from the addition of a constant, the log-likelihood term above is given (as a function of $\mathbf{x}$ for fixed

$y_1, \ldots, y_{n-1}$) by

$$\begin{aligned}
&\log[p(y_1, \ldots, y_{n-1}|\mathbf{x})] \\
&= -\frac{6}{\sigma^2} \sum_{i=1}^{n-1} \left[\left(x_i^2 + x_{i+1}^2 + 2x_i x_{i+1} \right)/4 - y_i \left(x_i + x_{i+1} \right) \right] / \Delta_i \\
&= -\frac{1}{2} \frac{3}{\sigma^2} \left[\frac{x_1^2}{\Delta_1} + \sum_{i=2}^{n-1} \left(\frac{1}{\Delta_{i-1}} + \frac{1}{\Delta_i} \right) x_i^2 + \frac{x_n^2}{\Delta_{n-1}} \right. \\
&\qquad \left. + 2 \sum_{i=1}^{n-1} \frac{x_i x_{i+1}}{\Delta_i} \right] \\
&\quad + \frac{6}{\sigma^2} \left[\frac{y_1}{\Delta_1} x_1 + \sum_{i=2}^{n-1} \left(\frac{y_{i-1}}{\Delta_{i-1}} + \frac{y_i}{\Delta_i} \right) x_i + \frac{y_{n-1}}{\Delta_{n-1}} x_n \right] \qquad \text{(A.2)} \\
&= -\frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{b}^T \mathbf{x},
\end{aligned}$$

where $\mathbf{Q}$ is defined in the statement of the theorem and

$$\mathbf{b} = \frac{6}{\sigma^2} \left(\frac{y_1}{\Delta_1}, \ldots, \frac{y_{i-1}}{\Delta_{i-1}} + \frac{y_i}{\Delta_i}, \ldots, \frac{y_{n-1}}{\Delta_{n-1}} \right)^T .$$

Writing the prior density of $\mathbf{X}$ in the form

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{Q}_0 \mathbf{x} + {\mathbf{b}_0}^T \mathbf{x} \right),$$

where $\mathbf{Q}_0 = \boldsymbol{\Sigma}_0^{-1}$ and $\mathbf{b}_0 = \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0$ (see Rue and Held, 2005, p. 27) and using (A.1) and (A.2), we obtain

$$p(\mathbf{x}|y_1, \ldots, y_{n-1}) \propto \exp\left(-\frac{1}{2} \mathbf{x}^T \widehat{\mathbf{Q}} \mathbf{x} + \widehat{\mathbf{b}}^T \mathbf{x} \right),$$

where $\widehat{\mathbf{Q}} = \boldsymbol{\Sigma}_0^{-1} + \mathbf{Q}$ and $\widehat{\mathbf{b}} = \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{b}$. This implies that the posterior distribution of $\mathbf{X}$ is Gaussian with covariance matrix $\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{Q})^{-1}$ and mean

$$\widehat{\boldsymbol{\mu}} = \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{b}} = \widehat{\boldsymbol{\Sigma}} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{b}) = \widehat{\boldsymbol{\Sigma}} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{D}\mathbf{y}).$$

The next part of the proof is to determine the conditional distribution of X given $\mathbf{X}$ and the data. From the structure of the prior distribution of X between observation times, the conditional distribution of X given $\mathbf{X}$ and the data coincides with the distribution of the process

$$\begin{aligned}
X(t) = {} & \sigma \Delta_i^{1/2} B_i^0((t - t_i)/\Delta_i) + X(t_i) \\
& + [X(t_{i+1}) - X(t_i)](t - t_i)/\Delta_i \qquad \text{(A.3)}
\end{aligned}$$

for $t \in [t_i, t_{i+1})$, where B_i^0, $i = 1, \ldots, n-1$ are independent standard Brownian bridges subject to the constraint (1) imposed on X by the data, i.e.,

$$\sigma \int_0^1 B_i^0(s)\, ds + (X(t_i) + X(t_{i+1}))/2 = y_i .$$

From Lemma 1 that follows the proof, providing the conditional distribution of B_i^0 given $\int_0^1 B_i^0(s)\, ds$, it is then seen that the conditional distribution of X given $\mathbf{X}$ and the data is Gaussian with mean

$$\begin{aligned}
E(X(t)|\mathbf{X}, y_1, \ldots, y_{n-1}) = {} & X(t_i) + [X(t_{i+1}) - X(t_i)](t - t_i)/\Delta_i \\
& + 6(t - t_i)(t_{i+1} - t) \left[y_i - (X(t_i) \right. \\
& \left. + \; X(t_{i+1}))/2 \right] / \Delta_i^2
\end{aligned}$$

for $t \in [t_i, t_{i+1})$, and covariance kernel $\sigma^2 \widetilde{K}$, where $\widetilde{K}$ is defined in the statement of the theorem. Setting $z_i = X(t_i) - \widehat{\mu}_i$ and rearranging terms in the above display then gives

$$E(X(t)|\mathbf{X}, y_1, \ldots, y_{n-1}) = \widehat{\mu}(t) + Z(t),$$

where

$$\begin{aligned}
Z(t) &= z_i + (z_{i+1} - z_i)(t - t_i)/\Delta_i \\
&\quad - 3(t - t_i)(t_{i+1} - t)(z_i + z_{i+1})/\Delta_i^2 \\
&= a_i(t) z_i + b_i(t) z_{i+1}
\end{aligned}$$

for $t \in [t_i, t_{i+1})$, $i = 1, \ldots, n-1$. Here $a_i(t)$ and $b_i(t)$ are defined in the statement of the theorem.

The final step of the proof is to remove the conditioning on $\mathbf{X}$. From the first part of the proof, we have that the posterior distribution of $(z_1, \ldots, z_n)^T = \mathbf{X} - \widehat{\boldsymbol{\mu}}$ is Gaussian with mean zero and covariance matrix $\widehat{\boldsymbol{\Sigma}}$. It follows immediately that the the posterior distribution of the process Z is Gaussian with mean zero and covariance kernel K^*, as defined in terms of $\widehat{\boldsymbol{\Sigma}}$ in the statement of the theorem. As we showed above, the conditional distribution of X given Z (or $\mathbf{X}$) and the data are Gaussian with mean function $\widehat{\mu} + Z$ and covariance kernel $\sigma^2 \widetilde{K}$. Since $\widehat{\mu}$ and $\widetilde{K}$ do not depend on Z, it follows using the convolution formula that the posterior distribution of X is Gaussian with mean function $\widehat{\mu}$ and covariance kernel $\widehat{K} = \sigma^2 \widetilde{K} + K^*$.

LEMMA 1: *Let B^0 be a standard Brownian bridge. The conditional distribution of B^0 given $\int_0^1 B^0(s)\, ds$ is Gaussian with mean $\mu^0(t) = 6t(1-t) \int_0^1 B^0(s)\, ds$ and covariance kernel $K^0(s,t) = s \wedge t - st - 3ts(1-t)(1-s)$, for $s, t \in [0,1]$.*

Proof. Note that $\boldsymbol{W} = (B^0(s), B^0(t), \int_0^1 B^0(u)\, du)^T$ is a zero-mean Gaussian random vector. Partition its covariance matrix as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are the covariance matrices of $\boldsymbol{W}^{(1)} = (B^0(s), B^0(t))^T$, and $\boldsymbol{W}^{(2)} = \int_0^1 B^0(u)\, du$, respectively. Then, from the covariance of B^0,

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} s(1-s) & s \wedge t - st \\ s \wedge t - st & t(1-t) \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T = (s(1-s)/2, t(1-t)/2)^T ,$$

and $\boldsymbol{\Sigma}_{22} = 1/12$. The conditional distribution of $\boldsymbol{W}^{(1)}$ given $\boldsymbol{W}^{(2)}$ is Gaussian with mean $\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{W}^{(2)} - \boldsymbol{\mu}^{(2)})$, where $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ are their respective means, and covariance $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$. The result now follows since the finite-dimensional distributions of a Gaussian process are determined by its mean and covariance kernel.