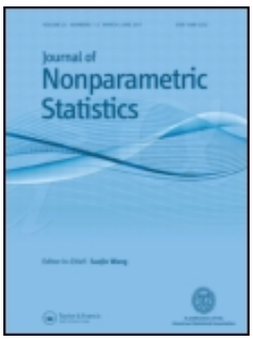

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Constrained spline regression in the presence of AR(p) errors

Huan Wang, Mary C. Meyer & Jean D. Opsomer

To cite this article: Huan Wang, Mary C. Meyer & Jean D. Opsomer (2013) Constrained spline regression in the presence of AR(p) errors, Journal of Nonparametric Statistics, 25:4, 809-827, DOI: 10.1080/10485252.2013.804075

To link to this article: http://dx.doi.org/10.1080/10485252.2013.804075

Journal of Nonparametric Statistics, 2013
Vol. 25, No. 4, 809–827, http://dx.doi.org/10.1080/10485252.2013.804075

Constrained spline regression in the presence of AR(p) errors

Huan Wang*, Mary C. Meyer and Jean D. Opsomer

Department of Statistics, Colorado State University, Fort Collins, CO, USA

(*Received 10 December 2012; accepted 6 May 2013*)

Extracting the trend from the pattern of observations is always difficult, especially when the trend is obscured by correlated errors. Often, prior knowledge of the trend does not include a parametric family, and instead the valid assumptions are vague, such as 'smooth' or 'monotone increasing'. Incorrectly specifying the trend as some simple parametric form can lead to overestimation of the correlation. The proposed method uses spline regression with shape constraints, such as monotonicity or convexity, for estimation and inference in the presence of stationary AR(p) errors. Standard criteria for selection of penalty parameter, such as Akaike information criterion (AIC), cross-validation and generalised cross-validation, have been shown to behave badly when the errors are correlated and in the absence of shape constraints. In this article, correlation structure and penalty parameter are selected simultaneously using a correlation-adjusted AIC. The asymptotic properties of unpenalised spline regression in the presence of correlation are investigated. It is proved that even if the estimation of the correlation is inconsistent, the corresponding projection estimation of the regression function can still be consistent and have the optimal asymptotic rate, under appropriate conditions. The constrained spline fit attains the convergence rate of unconstrained spline fit in the presence of AR(p) errors. Simulation results show that the constrained estimator typically behaves better than the unconstrained version if the true trend satisfies the constraints.

Keywords: shape-restricted inference; penalised spline; AR(p); correlation-adjusted AIC

AMS Subject Classifications: 62G05; 62G08; 62G20

1. Literature review and introduction

Regression splines are a popular nonparametric function estimator method, but they are known to be sensitive to knot number and placement. However, if there is more information about the shape of the regression function, like monotonicity or convexity, the shape-restricted splines are robust to knot choices.

For shape-restricted regression, Brunk (1955, 1958) proposed unsmoothed monotone regression estimation and studied its asymptotic behaviour. See Robertson, Wright, and Dykstra (1988) for details about estimation inference. Ramsay (1998) proposed a device to estimate a smooth strictly monotone function of arbitrary flexibility. Tantiyaswasdikul and Woodroofe (1994) proposed the monotone smoothing splines with penalty on the integrated first derivative. Mammen and Thomas-Agnan (1999) showed that the monotone smoothing splines have an optimal $n^{-p/(2p+1)}$ convergence rate, where $p = \max\{k, r\}$, k is the order of spline and r is the order of derivative. Hall and Huang (2001) developed a biased-bootstrap method for monotonising general linear,

*Corresponding author. Email: wangh@stat.colostate.edu

kernel-typed estimators. Meyer (2008) proposed an algorithm for the cubic monotone case, and also extended the method to convex constraints and variants such as increasing–concave.

Penalised splines, introduced by Eilers and Marx (1996), use a large number of knots compared to regression splines, but fewer than in smoothing splines, and hence are less computationally cumbersome. The penalisation shrinks the coefficients towards zero, constraining their influence and resulting in a less variable fit than regression splines. Penalised splines are increasingly popular in handling a wide range of nonparametric and semi-parametric problems. Ruppert, Wand, and Carroll (2003) provided details of this method. Hall and Opsomer (2005) used a white-noise process representation of the penalised spline estimator to obtain the mean squared error and consistency of the estimator. This representation treats the data as being generated from a continuously varying set of basis functions, subject to a penalty, so the complicating effect of the finite set of basis functions is removed. This enabled them to explore the role of the penalty and its relationship with the sample size in ways that are not possible in the discrete-data, finite-basis setting. Li and Ruppert (2008) showed that penalised splines behave similarly to Nadaraya–Watson kernel estimators with equivalent kernels. By this equivalent kernel representation, they developed an asymptotic theory of penalised splines for the cases of piecewise-constant or linear splines, with a first- or second-order difference penalty. Claeskens, Krivobokova, and Opsomer (2009) developed a general theory of the asymptotic properties of penalised spline estimators for any order of spline and general penalty. They demonstrated that the theoretical properties of penalised spline estimators are either similar to those of regression splines or to those of smoothing splines, with a clear breakpoint distinguishing the cases. Kauermann, Krivobokova, and Fahrmeir (2009) used a Bayesian viewpoint by imposing a priori distribution on all parameters and coefficients, arguing that with the postulated rate at which the spine basis dimension increases with the sample size the posterior distribution of the spline coefficients is approximately normal.

Nonparametric regression estimators are often sensitive to the presence of correlation in the errors. Most of the data-driven smoothing parameter selection methods, such as cross-validation, general cross-validation and Akaike information criterion (AIC), will break down if the correlation is ignored. Diggle and Hutchinson (1989) presented an extension of generalised cross-validation (GCV) which accommodates a known correlation matrix for the errors. Altman (1990) suggested two methods, a direct method and an indirect method, for correcting the selection criteria when the correlation function is known. Hart (1991) used a risk estimation procedure to select the bandwidth in the kernel regression with correlated errors. Hart (1994) proposed time series cross-validation to estimate the bandwidth and gave a time series model for the errors simultaneously. Wang (1998) extended the generalised maximum likelihood, GCV and unbiased risk methods to estimate the smoothing parameters and the correlation parameters simultaneously, when the correlation matrix is assumed to depend on a parsimonious set of parameters. Opsomer, Wang, and Yang (2001) gave a general review of the literature in kernel regression, smoothing splines and wavelet regression under correlation. Hall and Keilegom (2003) used difference-based methods to construct estimators of error variance and autoregressive parameters in nonparametric regression with time series errors. They proved that the difference-based estimators can be used to produce a simplified version of time series cross-validation. Francisco-Fernandez and Opsomer (2005) proposed to adjust the GCV criterion for the spatial correlation and showed that it leads to improved smoothing parameter selection results even when the covariance model is misspecified. Kim, Park, Moon, and Kim (2009) investigated a bandwidth selector based on the use of a bimodal kernel for nonparametric regression with fixed design and proved that the proposed selector is quite effective when the errors are severely correlated.

In this article, we propose a constrained spline estimator for data with stationary AR(p) errors with unknown order and unknown correlation parameters. Because of the popularity of the penalised splines, we include a penalty into our estimator. A new correlation-adjusted AIC is given for the selection of the penalty parameter and autoregressive parameters simultaneously.

We prove the asymptotic properties of the constrained unpenalised spline estimator in the presence of stationary AR(p) errors. The asymptotic properties of constrained penalised splines regression, due to the complexity of proofs for penalised splines, are still being studied and not included in this paper.

The proposed estimator and the method to select the order of correlation and the penalty parameter are presented in Section 2. In Section 3, the convergence rate of the estimator in the presence of correlation is derived in a general setting of both parametric and nonparametric regression, and also the specific application of constrained spline regression. The comparison of the convergence rate of the constrained spline regression and the unconstrained spline regression is also discussed in Section 3. Simulations evaluating the selection method of the order of AR(p) process and comparing the proposed method with the other two alternatives are conducted in Section 4. In Section 5, we analyse the global temperature data with the proposed method and compare with other methods.

2. Model setup and proposed estimator

Assume that the observed data $\{(x_i, y_i)\}$, for $1 \le i \le n$, are generated by the model

$$y_i = f(x_i) + \sigma \varepsilon_i,$$

where f is a smooth function. Suppose that $x_i \in [0, 1]$ and equally spaced. The errors $\varepsilon_1, \ldots, \varepsilon_n$ come from a segment of a mean zero autoregressive process with order p, i.e. an AR(p) process. Specifically, for some integer $p \ge 1$,

$$\varepsilon_i = \sum_{j=1}^{p} \theta_j \varepsilon_{i-j} + e_i,$$

where e_i are independent standard normal random variables.

The function f is approximated by a linear combination of spline basis functions. Given a set of knots $0 = t_1 < \cdots < t_k = 1$, a set of $m = k + d - 1$ basis functions $b_1(x), \ldots, b_m(x)$ are defined, where $d = 2$ for quadratic splines and $d = 3$ for cubic splines. The standard B-spline basis is used in this article, but another basis spanning the same space can be used instead. Let $\mathbf{b}_1, \ldots, \mathbf{b}_m$ be basis vectors, where $b_{ij} = b_j(x_i)$, so that the basis functions span the space of smooth piecewise polynomial regression functions with the given knots, and the basis vectors span an m-dimensional subspace of $\mathbb{R}^n$.

For the independent-error case, the penalised sum of squares of Eilers and Marx (1996) is

$$\sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} \alpha_j b_j(x_i) \right]^2 + \lambda \sum_{j=q+1}^{m} (\Delta^q \alpha_j)^2,$$

where $\Delta^1 \alpha_j = \alpha_j - \alpha_{j-1}$ and $\Delta^q \alpha_j = \Delta^{q-1} \Delta \alpha_j$ for $q > 1$. Let $\mathbf{B}$ be the $n \times m$ matrix with the $\mathbf{b}_j$ vectors as columns, let $\mathbf{D}$ be the qth order difference matrix and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$. The penalty parameter $\lambda \ge 0$ controls the smoothness. Minimising the penalised sum of squares is equivalent to minimising the vector expression:

$$\psi(\boldsymbol{\alpha}; \mathbf{y}) = \boldsymbol{\alpha}'(\mathbf{B}'\mathbf{B} + \lambda \mathbf{D}'\mathbf{D})\boldsymbol{\alpha} - 2\mathbf{y}'\mathbf{B}\boldsymbol{\alpha}.$$

For the monotone case, we use quadratic splines and define the $k \times m$ matrix $\mathbf{T}$ of the slopes at the knots by $T_{ij} = b_j'(t_i)$. Then, the linear combination $\sum_{j=1}^{m} \alpha_j b_j(x)$ is non-decreasing if and

only if the coefficient vector is in the set

$$\mathcal{C} = \{\boldsymbol{\alpha} : \mathbf{T}\boldsymbol{\alpha} \geq \mathbf{0}\} \subseteq \mathbb{R}^m.$$

For the convex case, we use cubic splines and $T_{ij} = b_j''(t_i)$; then, the linear combination is convex if and only if $\mathbf{T}\boldsymbol{\alpha} \geq \mathbf{0}$.

It is straightforward to find the appropriate spline degree and a constraint matrix $\mathbf{T}$ for constraint such as increasing and concave, or sigmoidal (convex or concave) with known inflation point.

When errors are correlated, let $cor(\boldsymbol{\varepsilon}) = \mathbf{R}$, and first suppose $\mathbf{R}$ is known. Let $\mathbf{R} = \mathbf{L}\mathbf{L}'$ be the Cholesky decomposition, and use the weighted least-squares method to estimate the coefficients. This is equivalent to transforming $\tilde{\mathbf{y}} = \mathbf{L}^{-1}\mathbf{y}, \tilde{\mathbf{B}} = \mathbf{L}^{-1}\mathbf{B}, \tilde{\boldsymbol{\varepsilon}} = \mathbf{L}^{-1}\boldsymbol{\varepsilon}$, which has correlation matrix $\mathbf{I}$. The weighted least-squares criterion is

$$\psi(\boldsymbol{\alpha}; \tilde{\mathbf{y}}) = \boldsymbol{\alpha}'(\tilde{\mathbf{B}}'\tilde{\mathbf{B}} + \lambda\mathbf{D}'\mathbf{D})\boldsymbol{\alpha} - 2\tilde{\mathbf{y}}'\tilde{\mathbf{B}}\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ is again restricted to $\mathcal{C}$.

Let $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' = (\tilde{\mathbf{B}}'\tilde{\mathbf{B}} + \lambda\mathbf{D}'\mathbf{D})$, then $\boldsymbol{\phi} = \tilde{\mathbf{L}}'\boldsymbol{\alpha}, \mathbf{z} = \tilde{\mathbf{L}}^{-1}\tilde{\mathbf{B}}'\tilde{\mathbf{y}}$, then

$$\psi(\boldsymbol{\alpha}; \tilde{\mathbf{y}}) = \psi(\boldsymbol{\phi}; \mathbf{z}) = \|\boldsymbol{\phi} - \mathbf{z}\|^2,$$

where $\boldsymbol{\phi}$ is restricted to $\tilde{\mathcal{C}} = \{\boldsymbol{\phi} : \mathbf{A}\boldsymbol{\phi} \geq \mathbf{0}\} \subseteq \mathbb{R}^m$, a polyhedral cone, where the $k \times m\mathbf{A} = \mathbf{T}(\tilde{\mathbf{L}}')^{-1}$ is full row-rank. Referring to the setup in Meyer (2012b), let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{m-k}$ span the null space $\mathcal{V}$ of $\mathbf{A}$, and let $\tilde{\mathbf{A}}$ be the square, nonsingular matrix with the rows of $\mathbf{A}$ as first k rows and $\boldsymbol{\nu}$ vectors as the last rows. The first k columns of $\tilde{\mathbf{A}}^{-1}$ are the edges $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$ of the cone, therefore the cone can be written as

$$\tilde{\mathcal{C}} = \left\{ \boldsymbol{\phi} : \boldsymbol{\phi} = \boldsymbol{\nu} + \sum_{j=1}^{k} \beta_j \boldsymbol{\delta}_j, \quad \boldsymbol{\nu} \in \mathcal{V}, \quad \beta_j \geq 0, j = 1, \ldots, k \right\}.$$

The minimiser $\hat{\boldsymbol{\phi}}$ is the projection of $\mathbf{z}$ onto the cone $\tilde{\mathcal{C}}$ and lands on a face of the cone. The 2^k faces, which partition $\tilde{\mathcal{C}}$, are indexed by the collection of sets $J \subseteq \{1, \ldots, m\}$, and are defined by

$$\mathcal{F}_J = \left\{ \boldsymbol{\phi} : \boldsymbol{\phi} = \boldsymbol{\nu} + \sum_{j \in J} \beta_j \boldsymbol{\delta}_j, \quad \boldsymbol{\nu} \in \mathcal{V}, \quad \beta_j > 0, j \in J \right\}.$$

The interior of the cone is a face with $J = \{1, \ldots, m\}$, and the origin is the face with $J = \emptyset$. We use the hinge algorithm from Meyer (2012b) to determine the face $\mathcal{F}_J$ on which the projection falls, so that the estimate coincides with the ordinary least-squares projection onto the linear space spanned by the edges of the chosen face. Let Δ_J be the matrix whose columns are those edges indexed by J, where $J \subseteq \{1, \ldots, m\}$. The projection is $\hat{\boldsymbol{\phi}} = \Delta_J(\Delta_J'\Delta_J)^{-1}\Delta_J'\mathbf{z}$, and the estimated coefficient vector is

$$\hat{\boldsymbol{\alpha}}_c = (\tilde{\mathbf{L}}')^{-1}\hat{\boldsymbol{\phi}} = (\tilde{\mathbf{L}}')^{-1}\Delta_J(\Delta_J'\Delta_J)^{-1}\Delta_J'\tilde{\mathbf{L}}^{-1}\tilde{\mathbf{B}}'\tilde{\mathbf{y}}.$$

For $\boldsymbol{\mu} \in \mathbb{R}^n$, where $\mu_i = f(x_i)$, the constrained estimated mean with the known $\mathbf{R}$ is $\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c = \mathbf{B}\hat{\boldsymbol{\alpha}}_c$. The matrix

$$\mathbf{P}_{\mathbf{R}}^c = \mathbf{B}(\tilde{\mathbf{L}}')^{-1}\Delta_J(\Delta_J'\Delta_J)^{-1}\Delta_J'\tilde{\mathbf{L}}^{-1}\mathbf{B}'\mathbf{R}^{-1},$$

such that $\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c = \mathbf{P}_{\mathbf{R}}^c\mathbf{y}$, is used to calculate effective degrees of freedom, i.e. edf $= \mathrm{tr}(\mathbf{P}_{\mathbf{R}}^c)$.

If $J = \{1, \ldots, m\}$, that is, all edges are used, then $\Delta_J(\Delta_J'\Delta_J)^{-1}\Delta_J' = \mathbf{I}$, and the unconstrained spline satisfies the constraints and is identical to the constrained fit. The unconstrained estimated coefficient vector is

$$\hat{\boldsymbol{\alpha}}_u = (\tilde{\mathbf{L}}')^{-1}(\tilde{\mathbf{L}})^{-1}\tilde{\mathbf{B}}'\tilde{\mathbf{y}} = (\mathbf{B}'\mathbf{R}^{-1}\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{R}^{-1}\mathbf{y}.$$

and the unconstrained estimated mean with known $\mathbf{R}$ is $\hat{\boldsymbol{\mu}}^u_{\mathbf{R}} = \mathbf{B}\hat{\boldsymbol{\alpha}}_u$. The trace of $\mathbf{P}^u_{\mathbf{R}} = \mathbf{B}(\mathbf{B}'\mathbf{R}^{-1}\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{R}^{-1}$ is the unconstrained edf.

The edf for the constrained fit is a random quantity with $m + 1$ possible values, the largest of which is that of the unconstrained version. Meyer (2012a) discussed this for the independent-error case.

However, typically $cor(\boldsymbol{\varepsilon}) = \mathbf{R}$ is unknown. Here, we assume AR(p) and use Cochrane–Orcutt type iterations to estimate the matrix $\mathbf{R}$. Altman (1992) introduces a similar iteration procedure for kernel regression in the presence of correlation and also discuss the behaviours of the procedure for different types of kernels.

For our method, given p and λ, the iteration procedure for either constrained or unconstrained trend estimation is

(1) Pilot fit: ignoring the correlation, obtain $\hat{\boldsymbol{\mu}}^c_{\mathbf{I}}$ and residuals $\hat{\epsilon}_i = y_i - \hat{\mu}^c_{\mathbf{I}i}$.
(2) Use the Yule-Walker method in Chapter 8 of Brockwell and Davis (2009) and residual vector $\hat{\boldsymbol{\epsilon}}$ to estimate coefficients $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and the error variance. If $\hat{\gamma}(0)$ and $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$ are the estimates of correlation function values; then obtain $\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\gamma}}_p$;
(3) Use Cholesky decomposition $\hat{\mathbf{R}} = \hat{\boldsymbol{\Sigma}}/\hat{\gamma}(0) = \mathbf{L}\mathbf{L}'$, to transform data and basis into $\tilde{\mathbf{y}} = \mathbf{L}^{-1}\mathbf{y}, \tilde{\mathbf{B}} = \mathbf{L}^{-1}\mathbf{B}$. Using data $\tilde{\mathbf{y}}$ and spline basis matrix $\tilde{\mathbf{B}}$, obtain adjusted estimators $\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}}^c_{\hat{\mathbf{R}}}$.
(4) Iterate (2)–(3) twice more, obtaining the final estimators $\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}}^c_{\hat{\mathbf{R}}}$.

By the result obtained from the iteration procedure, we can compute the correlation-adjusted AIC

$$\text{AIC} = n\log(\hat{\sigma}^2) + 2(p + \text{edf}). \tag{1}$$

We use this criterion to choose p and λ simultaneously, that is, we compute fits for a grid of p and λ values and choose the pair to minimise the AIC.

Most commonly used data-driven selection methods for tuning parameter, such as GCV and AIC, have been developed under the assumption of independent observations. When the regression is attempted in the presence of correlated errors, those automated methods will break down if the correlation is ignored. They tend to select a small tuning parameter and the fits become progressively more under-smoothed as the correlation increases. Opsomer et al. (2001) gave an overview of these problems. We will see that these problems are alleviated if the trend is constrained to be monotone or convex.

3. Large sample theory

3.1. *Rates of convergence in the presence of correlation*

We derive several general theorems on the convergence rate of projection estimation for unconstrained and unpenalised regression in the presence of stationary AR(p) errors. By using those results, the theorems on the convergence rate for constrained unpenalised spline regression are presented.

Without loss of generality, assume $\sigma = 1$. We model the regression function f as being a member of some linear function space $\mathbf{H}$, which is a subspace of all square-integrable, real-valued functions on $[0, 1]$. The least-squares estimation is a projection onto a finite-dimensional approximating subspace $\mathbf{G}_n$, which will be defined explicitly in Condition 3.2. If $\mathbf{H}$ is finite-dimensional, then we can choose $\mathbf{G}_n = \mathbf{H}$, leading to classical linear regression. Let $\hat{\boldsymbol{\mu}}_\mathbf{I}$ be the ordinary least-squares estimator of $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}_\mathbf{R}$ be the weighted least-squares estimator of $\boldsymbol{\mu}$, when $\mathbf{R}$ is known. If $\mathbf{R}$ is unknown, suppose there is an $n \times n$ symmetric positive-definite matrix $\mathbf{S}$, which can be used as an estimator of $\mathbf{R}$; let $\hat{\boldsymbol{\mu}}_\mathbf{S}$ be the weighted least-squares estimator with the given matrix $\mathbf{S}$.

It is well known that $\hat{\boldsymbol{\mu}}_\mathbf{R}$ is superior to $\hat{\boldsymbol{\mu}}_\mathbf{I}$ in that the variance of any linear contrast $\boldsymbol{\lambda}'\hat{\boldsymbol{\mu}}_\mathbf{R}$ is no larger than the variance of the corresponding linear contrast of $\boldsymbol{\lambda}'\hat{\boldsymbol{\mu}}_\mathbf{I}$. However, the construction of $\hat{\boldsymbol{\mu}}_\mathbf{R}$ requires the knowledge of $\mathbf{R}$ and generally $\mathbf{R}$ is not known. In fact, one may wish to estimate the mean function prior to investigating the covariance structure of the errors. Therefore, the properties of the ordinary least-squares estimator $\hat{\boldsymbol{\mu}}_\mathbf{I}$ are of interest. Furthermore, if the mean function is estimated with an arbitrary positive-definite symmetric non-random matrix $\mathbf{S}$, it is of interest to check whether this $\hat{\boldsymbol{\mu}}_\mathbf{S}$ can still attain the same rate of convergence under some appropriate conditions.

Huang (1998) developed a general theory on rates of convergence for independent observations in a more general setting in which the predictor variable can be random or fixed. We will extend Huang's theory to correlated observations in the case of equally spaced x_i.

For $\boldsymbol{\mu} \in \mathbb{R}^n$, define the norm as $\|\boldsymbol{\mu}\|^2 = 1/n\langle \boldsymbol{\mu}, \boldsymbol{\mu}\rangle$, where $\langle \mathbf{a}, \mathbf{b}\rangle = \sum_{i=1}^n a_i b_i$. Let $\mathbf{P}$ be the orthogonal projection matrix onto $\mathbf{G}_n$. Let $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y}$ and $\tilde{\boldsymbol{\mu}} = \mathbf{P}\boldsymbol{\mu}$, which is called the best approximation in $\mathbf{G}_n$ to $\boldsymbol{\mu}$. The total error can be decomposed as

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}) + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}).$$

We refer $\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}$ as the estimation error, and $\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}$ as the approximation error.

By the triangle inequality,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\| + \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|.$$

Therefore, we can examine separately the contributions of the two parts in this decomposition to the integrated squared error. The contribution to the integrated squared error from the first part is bounded in probability by N_n/n, where N_n is the dimension of $\mathbf{G}_n$, while the contribution from the second part is governed by ρ_n, the approximation power of $\mathbf{G}_n$. The convergence rates for the two parts equal the corresponding rates for the independent scenario under some conditions.

First, we state the conditions for the main results. The first two conditions, following Huang (1998), are on the approximating spaces. The first condition requires that the approximating space satisfies a stability constraint. This condition is satisfied by polynomials, trigonometric polynomials and splines. The second condition says that the approximating space must grow so that its distance from any function in $\mathbf{H}$ approaches zero.

For any function f on $[0, 1]$, set $\|f\|_\infty = \max_{x\in[0,1]} |f(x)|$.

Condition 3.1 There are positive constants A_n such that, $\|f\|_\infty \le A_n\|f\|$ for all $f \in \mathbf{G}_n$ and $\lim_n A_n^2 N_n/n = 0$.

Condition 3.2 There are nonnegative numbers $\rho_n = \rho_n(\mathbf{G}_n)$ such that for $\boldsymbol{\mu} \in \mathbf{H}$,

$$\inf_{\mathbf{g}\in\mathbf{G}_n} \|\mathbf{g} - \boldsymbol{\mu}\|_\infty \le \rho_n \to 0 \quad \text{as} \quad n \to \infty.$$

and $\limsup_n A_n\rho_n < \infty$.

If $\mathbf{H}$ is finite-dimensional, then we choose $\mathbf{G}_n = \mathbf{H}$, for all n. Condition 3.1 is automatically satisfied with A_n independent of n, and Condition 3.2 is satisfied with $\rho_n = 0$.

For the third condition, we require short-term dependence of errors.

Condition 3.3 Let $\gamma_{|i-j|} = E\varepsilon_i\varepsilon_j$, then there is a positive constant $M \in \mathbb{R}^1$, such that $\sum_{i=1}^{\infty} |\gamma_i| \le M$.

This condition implies that the row or column sum of correlation matrix $\mathbf{R}$ is bounded by a constant.

THEOREM 3.4 *Let* $\mathbf{P_I}$ *be the projection matrix of the ordinary least-squares estimation, then* $\hat{\boldsymbol{\mu}}_\mathbf{I} = \mathbf{P_I}\mathbf{y}$ *and* $\tilde{\boldsymbol{\mu}}_\mathbf{I} = \mathbf{P_I}\boldsymbol{\mu}$. *If Conditions* 3.1–3.3 *hold, then*

$$\|\hat{\boldsymbol{\mu}}_\mathbf{I} - \tilde{\boldsymbol{\mu}}_\mathbf{I}\|^2 = O_p\left(\frac{N_n}{n}\right), \quad \|\tilde{\boldsymbol{\mu}}_\mathbf{I} - \boldsymbol{\mu}\|^2 = O(\rho_n^2).$$

Consequently, $\|\hat{\boldsymbol{\mu}}_\mathbf{I} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$.

Huang (1998) derived the convergence rate of the least-squares estimate for independent observations in this general setting of both classical regression and nonparametric regression. Chapter 9 in Fuller (2009) derives the convergence rate for the least-squares estimate for correlated observations in linear regression. We propose a proof for this general setting with AR(p) errors.

Proof Let $\{\boldsymbol{\psi}_j, 1 \le j \le N_n\}$ be an orthonormal basis of $\mathbf{G}_n$.

$$\hat{\boldsymbol{\mu}}_\mathbf{I} - \tilde{\boldsymbol{\mu}}_\mathbf{I} = \sum_j \langle \hat{\boldsymbol{\mu}}_\mathbf{I} - \tilde{\boldsymbol{\mu}}_\mathbf{I}, \boldsymbol{\psi}_j\rangle \boldsymbol{\psi}_j = \sum_j \langle \mathbf{y} - \boldsymbol{\mu}, \boldsymbol{\psi}_j\rangle \boldsymbol{\psi}_j = \sum_j \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}_j\rangle \boldsymbol{\psi}_j.$$

Then, $\|\hat{\boldsymbol{\mu}}_\mathbf{I} - \tilde{\boldsymbol{\mu}}_\mathbf{I}\|^2 = (1/n)\sum_j \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}_j\rangle^2$, and

$$\begin{aligned}
E\|\hat{\boldsymbol{\mu}}_\mathbf{I} - \tilde{\boldsymbol{\mu}}_\mathbf{I}\|^2 &= \frac{1}{n}\sum_{j=1}^{N_n} E\langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}_j\rangle^2 \\
&= \frac{1}{n}\sum_{j=1}^{N_n} \boldsymbol{\psi}_j' \mathbf{R} \boldsymbol{\psi}_j \\
&= \frac{1}{n}\sum_{j=1}^{N_n}\sum_{l=1}^{n}\sum_{k=1}^{n} R_{lk}\psi_{lj}\psi_{kj} \\
&= \frac{1}{n}\sum_{j=1}^{N_n}\left[\sum_{l=1}^{n} R_{ll}\psi_{lj}^2 + 2\sum_{l=1}^{n}\sum_{k>l} R_{lk}\psi_{lj}\psi_{kj}\right] \\
&\le \frac{1}{n}\sum_{j=1}^{N_n}\left[\sum_{l=1}^{n} R_{ll}\psi_{lj}^2 + \sum_{l=1}^{n}\sum_{k>l} R_{lk}(\psi_{lj}^2 + \psi_{kj}^2)\right] \\
&= \frac{1}{n}\sum_{j=1}^{N_n}\left(\sum_{l=1}^{n}\sum_{k=l} R_{kl}\psi_{lj}^2 + \sum_{l=1}^{n}\sum_{k>l} R_{lk}\psi_{lj}^2 + \sum_{l=1}^{n}\sum_{k<l} R_{lk}\psi_{lj}^2\right)
\end{aligned}$$

$$= \frac{1}{n}\sum_{j=1}^{N_n}\sum_{l=1}^{n}\left(\sum_{k=1}^{n}R_{kl}\right)\psi_{lj}^2$$
$$\leq \frac{1}{n}\sum_{j=1}^{N_n}\sum_{l=1}^{n}M\psi_{lj}^2$$
$$= \frac{N_n}{n}M,$$

where R_{ij} is the i,jth element of $\mathbf{R}$, for $i,j = 1,\ldots,n$. So, $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = O_p(N_n/n)$. That $\|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O(\rho_n^2)$ is proved by Huang (1998). From Condition 3.2, we can find $\mathbf{g} \in \mathbf{G}_n$ such that $\|\boldsymbol{\mu} - \mathbf{g}\|_\infty \leq 2\rho_n$ and hence $\|\boldsymbol{\mu} - \mathbf{g}\| \leq 2\rho_n$. Then, we have that $\|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \mathbf{g}\|^2 = \|\mathbf{P}(\boldsymbol{\mu} - \mathbf{g})\|^2 \leq \|\boldsymbol{\mu} - \mathbf{g}\|^2$. Hence, by the triangle inequality,

$$\|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 \leq 2\|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \mathbf{g}\|^2 + 2\|\boldsymbol{\mu} - \mathbf{g}\|^2 \leq 4\|\boldsymbol{\mu} - \mathbf{g}\| = O(\rho_n^2).$$

Then, we have $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$. ■

We need another condition to prove the next results.

Condition 3.5 The error vector $\boldsymbol{\varepsilon}$ comes from a stationary AR(p) process, for an integer $p \geq 1$.

THEOREM 3.6 *Let* $\mathbf{P}_{\mathbf{R}}$ *be the projection matrix of the weighted least-squares estimation with the known correlation matrix* $\mathbf{R}$, *then* $\hat{\boldsymbol{\mu}}_{\mathbf{R}} = \mathbf{P}_{\mathbf{R}}\mathbf{y}$ *and* $\tilde{\boldsymbol{\mu}}_{\mathbf{R}} = \mathbf{P}_{\mathbf{R}}\boldsymbol{\mu}$. *If Conditions* 3.1–3.3 *and* 3.5 *hold, then*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 = O_p\left(\frac{N_n}{n}\right), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O(\rho_n^2).$$

Consequently, $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$.

Proof Let $\mathbf{L}$ be the Cholesky decomposition of $\mathbf{R}$, then $\mathbf{R} = \mathbf{L}\mathbf{L}'$. Let $\mathbf{y}^* = \mathbf{L}^{-1}\mathbf{y}, \boldsymbol{\mu}^* = \mathbf{L}^{-1}\boldsymbol{\mu}, \boldsymbol{\varepsilon}^* = \mathbf{L}^{-1}\boldsymbol{\varepsilon}$, then the model can be transformed into $\mathbf{y}^* = \boldsymbol{\mu}^* + \boldsymbol{\varepsilon}^*$ and $E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}) = \mathbf{I}$.

Let $\mathbf{G}_n^*$ be the transformed approximating subspace, spanned by $\mathbf{L}^{-1}\boldsymbol{\Psi}$, where the columns of $\boldsymbol{\Psi}$ span $\mathbf{G}_n$. Let $\hat{\boldsymbol{\mu}}^*$ be the orthogonal projection of $\mathbf{y}^*$ onto $\mathbf{G}_n^*$. Let $\tilde{\boldsymbol{\mu}}^*$ be the projection of $\boldsymbol{\mu}^*$ onto $\mathbf{G}_n^*$. By Theorem 2.1 in Huang (1998), we have $\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 = O_p(N_n/n)$, and $\|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2 = O(\rho_n^2)$. Then $\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 = \|\mathbf{L}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})\|^2 = 1/n(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})$. Since $\mathbf{R}^{-1}$ is a Hermitian matrix, its eigenvalues are all real. By the Rayleigh–Ritz theorem, the Rayleigh–Ritz ratio is bounded by the largest and smallest eigenvalues of $\mathbf{R}^{-1}$,

$$\lambda_{\min} \leq \frac{(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})}{(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})'(\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}})} \leq \lambda_{\max},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbf{R}^{-1}$. For $\mathbf{R}$ positive definite, it is easy to prove that $\mathbf{R}^{-1}$ is also positive definite. So, there exist two constant sequences m_n and M_n, where $0 < m_n \leq M_n < \infty$, for each specific n, such that, $m_n \leq \lambda_{\min} \leq \lambda_{\max} \leq M_n$, for each n. By Proposition 4.5.3 in Brockwell and Davis (2009), for a stationary AR(p) process, the eigenvalues of its covariance matrix are bounded away from zero and ∞ uniformly in n. Hence, for any n, there exist two constants M and m, such that $m \leq \lambda_{\min} \leq \lambda_{\max} \leq M$, for each n. Then, $1/M\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 \leq 1/m\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2$. Therefore, $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 = O_p(N_n/n)$.

By the same method used in the proof of $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 = O_p(N_n/n)$, we can prove $1/M \|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2 \leq \|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 \leq 1/m\|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2$. Therefore, $\|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(\rho_n^2)$. So, we have $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$. ■

For the case that $\mathbf{R}$ is unknown, we show that the estimation of trend is consistent and attains the same asymptotic rate as $\hat{\boldsymbol{\mu}}_{\mathbf{R}}$ for any suitable fixed $n \times n$ matrix $\mathbf{S}$ and argue that therefore the trend is estimated consistently with an estimator of $\mathbf{R}$ based on data.

THEOREM 3.7 *If the correlation matrix* $\mathbf{R}$ *is unknown, and choose a sequence of matrices* $\mathbf{S}$ *satisfying the following conditions*:

A1: $\mathbf{S}$ *is symmetric and positive-definite*;
A2: *All the eigenvalues of* $\mathbf{S}$ *are bounded from zero and* ∞, *uniformly in* n;
A3: *Let* $\mathbf{L_S}$ *be the Cholesky decomposition of* $\mathbf{S}$, *then* $\mathbf{L_S}^{-1}\mathbf{R}(\mathbf{L_S}^{-1})'$ *satisfies Condition* 3.3, *which means the sum of the absolute value of its first row is bounded by a constant*;

Let $\mathbf{P_S}$ *be the projection matrix of the weighted least-squares estimation with the substitute correlation matrix* $\mathbf{S}$, *then* $\hat{\boldsymbol{\mu}}_{\mathbf{S}} = \mathbf{P_S}\mathbf{y}$ *and* $\tilde{\boldsymbol{\mu}}_{\mathbf{S}} = \mathbf{P_S}\boldsymbol{\mu}$. *If Conditions* 3.1 *and* 3.2 *hold*, *then*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}\|^2 = O_p\left(\frac{N_n}{n}\right), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O(\rho_n^2).$$

Consequently, $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$.

Proof Let $\mathbf{y}_{\mathbf{S}}^* = \mathbf{L}_{\mathbf{S}}^{-1}\mathbf{y}, \boldsymbol{\mu}_{\mathbf{S}}^* = \mathbf{L}_{\mathbf{S}}^{-1}\boldsymbol{\mu}, \boldsymbol{\varepsilon}_{\mathbf{S}}^* = \mathbf{L}_{\mathbf{S}}^{-1}\boldsymbol{\varepsilon}$, then the model can be transformed into $\mathbf{y}_{\mathbf{S}}^* = \boldsymbol{\mu}_{\mathbf{S}}^* + \boldsymbol{\varepsilon}_{\mathbf{S}}^*, E(\boldsymbol{\varepsilon}_{\mathbf{S}}^*\boldsymbol{\varepsilon}_{\mathbf{S}}^{*\prime}) = \mathbf{L}_{\mathbf{S}}^{-1}\mathbf{R}(\mathbf{L}_{\mathbf{S}}^{-1})'$. Let $\mathbf{G}_n^{\mathbf{S}}$ be the transformed approximating subspace. Let $\hat{\boldsymbol{\mu}}_{\mathbf{S}}^*$ be the projection of $\mathbf{y}_{\mathbf{S}}^*$ onto $\mathbf{G}_n^{\mathbf{S}}$. Let $\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*$ be the projection of $\boldsymbol{\mu}_{\mathbf{S}}^*$ onto $\mathbf{G}_n^{\mathbf{S}}$. For $\mathbf{L}_{\mathbf{S}}^{-1}\mathbf{R}(\mathbf{L}_{\mathbf{S}}^{-1})'$ satisfying the Condition A3, then we have $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^* - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*\|^2 = O_p(N_n/n)$, and $\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^* - \boldsymbol{\mu}_{\mathbf{S}}^*\|^2 = O(\rho_n^2)$, by Theorem 3.4. Therefore, $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^* - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*\|^2 = \|\mathbf{L}_{\mathbf{S}}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})\|^2 = 1/n(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})'\mathbf{S}_n^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})$.

Then, by Rayleigh–Ritz Theorem, we have,

$$\lambda_{\min}^{\mathbf{S}} \leq \frac{(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})'\mathbf{S}_n^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})}{(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})'(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})} \leq \lambda_{\max}^{\mathbf{S}},$$

where $\lambda_{\min}^{\mathbf{S}}$ and $\lambda_{\max}^{\mathbf{S}}$ are the smallest and largest eigenvalues of $\mathbf{S}_n^{-1}$. By the condition A2 that the eigenvalues of $\mathbf{S}$ are bounded away from zero and ∞ uniformly in n, the eigenvalues of $\mathbf{S}^{-1}$ are also bounded from zero and ∞, which means that there exist two constants m and M, where $0 < m \leq M < \infty$, such that, $m \leq \lambda_{\min}^{\mathbf{S}} \leq \lambda_{\max}^{\mathbf{S}} \leq M$, for each n. Then, $1/M\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^* - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*\|^2 \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}\|^2 \leq 1/m\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^* - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*\|^2$. Therefore, $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}\|^2 = O_p(N_n/n)$. As in the proof of Theorem 3.6, we have $1/M\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^* - \boldsymbol{\mu}_{\mathbf{S}}^*\|^2 \leq \|\tilde{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 \leq 1/m\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^* - \boldsymbol{\mu}_{\mathbf{S}}^*\|^2$. Thus, we have $\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(\rho_n^2)$, and $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$. ■

Remark 1 Theorems 3.4, 3.6 and 3.7 can readily be applied to classical linear regression. Let $\mathbf{G}_n$ be the linear space spanned by the columns of $\mathbf{X}$, where $\mathbf{X}$ is an $n \times p$ full row-rank matrix with fixed values, so that $\mathbf{G}_n = \mathbf{H}$, $N_n = p$ and $\rho_n = 0$. The three estimators of $\boldsymbol{\mu}$ achieve the same convergence rate of $n^{-1/2}$.

The next theorem proves the consistency of the estimates of the correlation function.

THEOREM 3.8 *Let $\hat{\boldsymbol{\mu}}$ be a consistent estimator of $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$, $\hat{\gamma}_{\hat{\varepsilon}}(h) = (1/n)\sum_{i=1}^{n-h} \hat{\varepsilon}_i\hat{\varepsilon}_{i+h}$, $\hat{\gamma}_{\boldsymbol{\varepsilon}}(h) = (1/n)\sum_{i=1}^{n-h} \varepsilon_i\varepsilon_{i+h}$, $\gamma(h) = \mathbf{E}\varepsilon_i\varepsilon_{i+h}$, where $i = 1, \ldots, n-h; h = 0, 1, \ldots, n-1$. Let $\boldsymbol{\gamma}_{\hat{\varepsilon}} = (\gamma_{\hat{\varepsilon}}(1), \ldots, \gamma_{\hat{\varepsilon}}(n-1))'$ and $\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}} = (\gamma_{\boldsymbol{\varepsilon}}(1), \ldots, \gamma_{\boldsymbol{\varepsilon}}(n-1))'$, then*

$$\|\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}\|^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2), \quad \|\hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}} - \boldsymbol{\gamma}\|^2 = O_p\left(\frac{1}{n}\right).$$

Consequently, $\|\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}} - \boldsymbol{\gamma}\|^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 + 1/n)$.

Proof

$$\begin{aligned}
&\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}}(h) - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}(h) \\
&= \frac{1}{n}\sum_{i=1}^{n-h}(\hat{\varepsilon}_i\hat{\varepsilon}_{i+h} - \varepsilon_i\varepsilon_{i+h}) \\
&= \frac{1}{n}\sum_{i=1}^{n-h}[\varepsilon_i(\mu_{i+h} - \hat{\mu}_{i+h}) + \varepsilon_{i+h}(\mu_i - \hat{\mu}_i) + (\mu_i - \hat{\mu}_i)(\mu_{i+h} - \hat{\mu}_{i+h})] \\
&\le \left(\frac{1}{n}\sum_{i=1}^{n-h}\varepsilon_i^2\right)^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n-h}(\mu_{i+h} - \hat{\mu}_{i+h})^2\right]^{1/2} + \left(\frac{1}{n}\sum_{i=1}^{n-h}\varepsilon_{i+h}^2\right)^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n-h}(\mu_i - \hat{\mu}_i)^2\right]^{1/2} \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^{n-h}(\mu_{i+h} - \hat{\mu}_{i+h})^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n-h}(\mu_i - \hat{\mu}_i)^2\right]^{1/2} \\
&= O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| + \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2).
\end{aligned}$$

If $\hat{\boldsymbol{\mu}}$ is consistent, $\|\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}\|^2 = (1/(n-1))\sum_{h=1}^{n-1}[\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}}(h) - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}(h)]^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2)$. By Theorem 7.2.1 in Brockwell and Davis (2009), $\sqrt{n}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}(h))$ is asymptotic normally distributed, for $h = 0, 1, \ldots, n-1$. Thus, $\|\hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}} - \boldsymbol{\gamma}\|^2 = O_p(1/n)$. Therefore, we have $\|\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}} - \boldsymbol{\gamma}\|^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 + 1/n)$. ■

In the proposed iteration procedure, the pilot fit is the estimation ignoring correlation. By Theorem 3.8, the estimated correlation based on the pilot fit is consistent. Therefore, it satisfies Conditions A1–A3. By Theorem 3.7, the renewed estimation of trend based on this consistent estimator of correlation is consistent and attains the optimal asymptotic rate.

Remark 2 For classical linear regression, $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} = O_p(n^{-1/2})$, by Theorem 9.3.1 in Fuller (2009), $\hat{\boldsymbol{\gamma}}_{\hat{\varepsilon}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}} = O_p(1/n)$.

3.2. *Fixed-knot unpenalised unconstrained spline regression*

Theorems 3.4, 3.6 and 3.7 can also be applied to fixed-knot spline estimates when the knot positions are prespecified but the number of knots is allowed to increase with the sample size. In this section, we investigate the large sample theory for only unpenalised situation, i.e. $\lambda = 0$. Suppose f is *p-smooth* for a specified positive number p, that is, f is $[p]$ times continuously differentiable on $\mathbf{H}$, where $[p]$ is the greatest integer less than p, and all the $[p]$th-*order* mixed partial derivatives of f satisfy a Hölder condition with exponent $p - [p]$, referring to Huang (1998). Let $\mathbf{G}_n$ be the linear space of regression splines with degree $d \ge p - 1$. Suppose the knots have bounded mesh ratio, that is, the ratio of the largest inter-knot interval to the smallest is bounded from zero and infinity,

uniformly in n. Let a_n denote the smallest distance between two consecutive knots. For the two sequences of positive numbers a_{1n} and a_{2n}, let $a_{1n} \asymp a_{2n}$ mean that the ratio a_{1n}/a_{2n} is bounded away from zero and ∞. Then, we have $N_n \asymp 1/a_n$ and $\rho_n \asymp a_n^p \asymp N_n^{-p}$. Hence, the convergence rate for the three estimators, i.e. $\hat{\boldsymbol{\mu}}_{\mathbf{I}}$, $\hat{\boldsymbol{\mu}}_{\mathbf{R}}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{S}}$, is $O_p((a_n)/n + a_n^{2p})$. In order to let the rate of convergence be optimal, which means no estimate has a faster rate of convergence uniformly over the class of p-smooth functions, referring to Stone (1982), choose $a_n \asymp n^{-1/(2p+1)}$. This balances the estimation error and the approximation error, that is, $a_n/n \asymp a_n^{2p}$. Applying Theorems 3.4, 3.6 and 3.7 to this setting, we obtain the following results.

COROLLARY 3.9 *Suppose Conditions* 1–3 *hold and the knots have bounded mesh ratio. If we choose* $a_n \asymp n^{-1/(2p+1)}$, *then*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = O_p(n^{-2p/(2p+1)}), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O(n^{-2p/(2p+1)});$$
$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = O_p(n^{-2p/(2p+1)}), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O(n^{-2p/(2p+1)});$$
$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = O_p(n^{-2p/(2p+1)}), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O(n^{-2p/(2p+1)}).$$

Consequently, $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)})$; $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)})$; $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)})$.

3.3. *Fixed-knot constrained unpenalised spline regression*

In Theorems 3.4, 3.6 and 3.7, we derived the convergence rate in a general setting for both classical regression and nonparametric regression. The next theorem will compare the convergence rate of constrained estimator and the corresponding unconstrained estimator in spline regression in the presence of correlated errors. Let $\mathbf{G}_n^u = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{Bb}\}$, which is a finite-dimensional approximating subspace to $\mathbf{H}$ spanned by spline basis. Assume $f \in \mathbf{H}_c$, a subset of all square-integrable, real-valued, constrained functions on $\mathcal{X}$; $\boldsymbol{\mu} \in \mathbb{R}^n$, where $\mu_i = f(x_i)$. Let $\mathbf{G}_n^c = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{Bb}, \mathbf{Tb} \geq 0\}$, which is a finite-dimensional approximating subset of $\mathbf{H}_c$.

As before, we consider three kinds of estimators: ordinary least-squares, the weighted least-squares with known $\mathbf{R}$ and the weighted least-squares using a given matrix $\mathbf{S}$ as an substitute of correlation, for both constrained spline regression and unconstrained spline regression, and compare their convergence rates. Let $\mathbf{P}_{\mathbf{I}}^c$ be the projection matrix of the ordinary least-squares estimator in the constrained spline regression. It is a random matrix and depends on J, the index of the face identified by cone projection algorithm for a specific $\mathbf{y}$. Let $\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c = \mathbf{P}_{\mathbf{I}}^c\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^c = \mathbf{P}_{\mathbf{I}}^c\boldsymbol{\mu}$. Let $\mathbf{P}_{\mathbf{I}}^u$ be the projection matrix of the ordinary least-squares estimator in the unconstrained spline estimator. It is a fixed matrix, and corresponds to $\mathbf{P}_{\mathbf{I}}^c$ with $J = \{1, \ldots, m\}$. Let $\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u = \mathbf{P}_{\mathbf{I}}^u\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^u = \mathbf{P}_{\mathbf{I}}^u\boldsymbol{\mu}$. Let $\mathbf{P}_{\mathbf{R}}^c$ be the projection matrix of the weighted least-squares estimator in the constrained spline regression with the known $\mathbf{R}$, then $\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c = \mathbf{P}_{\mathbf{R}}^c\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}^c = \mathbf{P}_{\mathbf{R}}^c\boldsymbol{\mu}$. Let $\mathbf{P}_{\mathbf{R}}^u$ be the projection matrix of the weighted least-squares estimator in the unconstrained spline estimator, then $\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u = \mathbf{P}_{\mathbf{R}}^u\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}^u = \mathbf{P}_{\mathbf{R}}^u\boldsymbol{\mu}$. Let $\mathbf{P}_{\mathbf{S}}^c$ be the projection matrix of the weighted least-squares estimator in the constrained spline regression with the given matrix $\mathbf{S}$ as an estimator of the unknown $\mathbf{R}$, then $\hat{\boldsymbol{\mu}}_{\mathbf{S}}^c = \mathbf{P}_{\mathbf{S}}^c\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^c = \mathbf{P}_{\mathbf{S}}^c\boldsymbol{\mu}$. Let $\mathbf{P}_{\mathbf{S}}^u$ be the projection matrix of the weighted least-squares estimator in the unconstrained spline estimator with the given matrix $\mathbf{S}$ as an estimator of the unknown $\mathbf{R}$, then $\hat{\boldsymbol{\mu}}_{\mathbf{S}}^u = \mathbf{P}_{\mathbf{S}}^u\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^u = \mathbf{P}_{\mathbf{S}}^u\boldsymbol{\mu}$.

Assume that $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^u \in \mathbf{G}_n^c$, so that the shape restrictions hold; otherwise, $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^u$ cannot be consistent for $\boldsymbol{\mu}$, which is assumed to follow the given shape restrictions. Under this assumption, it is easy to prove that $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^u = \tilde{\boldsymbol{\mu}}_{\mathbf{I}}^c$. The same assumption and reasoning also apply to the other two estimators, therefore $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}^u = \tilde{\boldsymbol{\mu}}_{\mathbf{R}}^c$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^u = \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^c$. In this context, we use $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}$ instead of $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^u$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}^c$. The same treatment is used for $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{S}}$. Therefore, the approximation error for the constrained

estimators and unconstrained estimators in the same setting are the same, and the comparison of the total error is reduced to the comparison of the estimation error.

THEOREM 3.10 *Let the knots $t_1, \ldots, t_k$ have bounded mesh ratio, then*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \boldsymbol{\mu}\|^2 \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \boldsymbol{\mu}\|^2.$$

Hence, the convergence rate of the ordinary least-squares estimator in constrained spline regression attains that of the corresponding unconstrained spline regression, in the presence of correlation.

Proof The decomposition of errors is $\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) + (\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu})$ and $\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) + (\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu})$. So we only need to prove $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2$. We have

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 &= \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 + \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c\|^2 + 2(\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c)^t(\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) \\ &= \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 + \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c\|^2 \\ &\quad - 2(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^u)^t(\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) + 2(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c)^t(\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}).\end{aligned}$$

The Karush–Kuhn–Tucker conditions (see Silvapulle and Sen 2004, Appendix 1) imply, $\langle \mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c, \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c \rangle = 0$ and $\langle \mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathbf{I}}^c, \tilde{\boldsymbol{\mu}}_{\mathbf{I}}^c \rangle \leq 0$. Therefore, $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^u - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 \geq \|\hat{\boldsymbol{\mu}}_{\mathbf{I}}^c - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2$. ■

THEOREM 3.11 *Let the knots $t_1, \ldots, t_k$ have bounded mesh ratio. Then, there exists a constant $C \in \mathbb{R}^1$, bounded away from zero and ∞, such that,*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \boldsymbol{\mu}\|^2 \leq C\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \boldsymbol{\mu}\|^2.$$

Hence, the convergence rate of the weighted least-squares estimator with known correlation in constrained spline regression attains that of the corresponding unconstrained spline regression.

Proof Let $\mathbf{L}$ be the Cholesky decomposition of $\mathbf{R}$, then $\mathbf{R} = \mathbf{L}\mathbf{L}'$. Let $\mathbf{y}^* = \mathbf{L}^{-1}\mathbf{y}$, $\boldsymbol{\mu}^* = \mathbf{L}^{-1}\boldsymbol{\mu}$, and $\boldsymbol{\varepsilon}^* = \mathbf{L}^{-1}\boldsymbol{\varepsilon}$, then the model can be transformed into $\mathbf{y}^* = \boldsymbol{\mu}^* + \boldsymbol{\varepsilon}^*$ and $E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*'}) = \mathbf{I}$. Using the result in Theorem 3.10, we have $\|\hat{\boldsymbol{\mu}}_c^* - \tilde{\boldsymbol{\mu}}^*\|^2 \leq \|\hat{\boldsymbol{\mu}}_u^* - \tilde{\boldsymbol{\mu}}^*\|^2$, and when transformed back, we get $\|\mathbf{L}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \mathbf{L}^{-1}\tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 \leq \|\mathbf{L}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \mathbf{L}^{-1}\tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2$, that is $(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}}) \leq (\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})$. Since $\mathbf{R}^{-1}$ is Hermitian matrix, its eigenvalues are all real. By the Rayleigh–Ritz Theorem, the Rayleigh–Ritz ratio is bounded by the largest and smallest eigenvalues of $\mathbf{R}^{-1}$. Then we have,

$$\frac{(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}})}{(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}})} \geq \lambda_{\min} \quad \text{and} \quad \frac{(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})'\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})}{(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}})} \leq \lambda_{\max};$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbf{R}^{-1}$. There exist two constant sequences m_n and M_n, where $0 < m_n \leq M_n < \infty$, for each specific n, such that, $m_n \leq \lambda_{\min} \leq \lambda_{\max} \leq M_n$, for each n. As in the proof of Theorem 3.6 for any n, there exist two constants M and n, such that $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \tilde{\boldsymbol{\mu}}\|^2 \leq \frac{M}{m}\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \tilde{\boldsymbol{\mu}}\|^2$. So, there exist a constant $C \in \mathbb{R}^1$, bounded away from zero and ∞, such that, $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^c - \boldsymbol{\mu}\|^2 \leq C\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^u - \boldsymbol{\mu}\|^2$. ■

THEOREM 3.12 *Let the knots $t_1, \ldots, t_k$ have 'bounded mesh ratio'. The correlation matrix $\mathbf{R}$ is unknown, so choose any matrix $\mathbf{S}$ satisfying Conditions* A1–A3. *Then, there exists a constant $C \in \mathbb{R}^1$, bounded away from zero and ∞, such that,*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^c - \boldsymbol{\mu}\|^2 \leq C\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^u - \boldsymbol{\mu}\|^2, \quad \textit{with probability approaching one.}$$

Hence, the convergence rate of the weighted least-squares estimator with the given matrix as an substitute of correlation in constrained spline regression attains that in the corresponding unconstrained spline regression.

Using Theorems 3.10 and 3.11, the proof is entirely analogous to that of Theorem 3.7, and is omitted here.

4. Simulation

4.1. *Data Scenarios*

Simulations were carried out to examine the performance of the constrained penalised spline estimator, and to compare it with the unconstrained penalised spline estimator and the classical linear regression estimator. Data with different scenarios of trend and noise were generated. For the trend, linear, sigmoid and truncated cubic are used. For the noise, a series of AR(p) errors, where $p = 1, 2, 3, 4$, with gradually increasing correlation were generated, then $y_i = f(i/n) + \epsilon_i$. For the mean function $f(x)$, we used

(1) linear: $f(x) = x$,
(2) sigmoid: $f(x) = \mathrm{e}^{10x-5}/(1 + \mathrm{e}^{10x-5})$,
(3) truncated cubic: $f(x) = 4(x - 1/2)^3 I_{x>1/2}$.

Two series of noise were used in simulations:

- $\epsilon_i = \theta\epsilon_{i-1} + z_i, \theta = 0.1, 0.3, 0.5, 0.7$,
- $\epsilon_i = 0.3\epsilon_{i-p} + z_i, p = 1, 2, 3, 4$,

where $z_i's$ are independent and identically normally distributed with mean zero and standard deviation 0.2. The sample size for the simulation is 250 and the number of replications is 1000.

4.2. *The selection of order p and penalty parameter by AIC*

Many authors have studied the effects of correlation on the selection of the smoothing parameter and derived correlation-adjusted selection methods, see Diggle and Hutchinson (1989), Altman (1990) and Wang (1998). None of these selects the order of correlation and the smoothing parameter simultaneously for penalised spline regression. In this article, we use a correlation-adjusted AIC (1) criterion to select the penalty parameter and the order p simultaneously.

For each simulated data set, we compute 60 AIC values using $p = (0, 1, 2, 3, 4, 5)$ and 10 values of λ as candidates. The edf for both constrained and unconstrained estimators with unknown correlation is random. We choose the candidate λ by letting the corresponding edf for unconstrained penalised spline estimator for independent data be $(4, 5, 6, 8, 10, 12, 16, 20, 25, 30)$. We choose p and λ as the joint minimiser of $\mathrm{AIC}_{p,\lambda}$. We repeat this procedure for $N = 1000$ times, and calculate the fraction of times that AIC chooses the true p.

In Table 1, for truncated cubic data and sigmoid data, the values in first and third columns are always greater than the values in the corresponding second and fourth columns, which is just as we expected when the assumptions on the shape are correct. Linear regression behaves poorly for truncated cubic data and sigmoid data when the correlation is zero or small because it is more likely to choose a larger p. When the data are generated by a linear trend, the linear regression does the best job but the behaviour of the proposed estimator is still acceptable. We also conducted the simulations with larger p and higher degree of correlated errors, such as, AR(4) with $\theta = (0.4, 0.3, 0.15, 0.1)$. If the correlation is large enough, it can cause the failure of the AIC to select the true p for all three methods.

Table 1. The proportion of data sets for which the correlation-adjusted AIC criterion selects the true p, for the proposed estimator, the unconstrained penalised estimator and the classical linear regression estimator.

		AR(1)				AR(p)		
f	$\boldsymbol{\theta}$	Constrained	Unconstrained	Linear	p	Constrained	Unconstrained	Linear
Linear	0	0.617	0.619	0.717	0	0.617	0.619	0.717
	0.3	0.617	0.581	0.740	2	0.631	0.599	0.756
	0.5	0.663	0.632	0.760	3	0.645	0.599	0.799
	0.7	0.677	0.620	0.761	4	0.695	0.613	0.834
Cubic	0	0.673	0.592	0.078	0	0.673	0.592	0.078
	0.3	0.686	0.601	0.541	2	0.709	0.561	0.595
	0.5	0.697	0.587	0.700	3	0.748	0.562	0.691
	0.7	0.704	0.595	0.750	4	0.785	0.593	0.797
Sigmoid	0	0.575	0.532	0.019	0	0.575	0.532	0.019
	0.3	0.605	0.546	0.426	2	0.625	0.549	0.521
	0.5	0.658	0.611	0.669	3	0.627	0.520	0.649
	0.7	0.663	0.599	0.727	4	0.720	0.582	0.744

Note: The simulated data are generated by three different mean functions, linear, truncated cubic and sigmoid, with AR(1) errors, where $\theta = 0, 0.1, 0.3, 0.5, 0.7$ and AR(p) errors, where $\epsilon_i = 0.3\epsilon_{i-p} + z_i, p = 0, 2, 3, 4$.

4.3. *Three performance measures*

To compare the performance of the proposed estimator with the unconstrained penalised spline estimator and classical linear regression estimator, the following three measures are constructed. The first and the second measures are used to compare the estimations of the correlation. The third one is used to compare the estimations of the trend.

$$\Delta_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{\mathrm{K}}(\hat{\theta}_{i\hat{p}k} - \theta_k)^2}; \quad \Delta_{\boldsymbol{\gamma}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{h=1}^{20}(\hat{\gamma}_{i\hat{p}h} - \gamma_h)^2};$$

and

$$\Delta_{\mathbf{f}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{\sum_{l=1}^{n}(\hat{f}_{\hat{p}i}(x_l) - f(x_l))^2}{n}}.$$

Here, K is the largest length of $\theta, \theta = (\theta_1, \ldots, \theta_K), \hat{f}_{\hat{p}i}(x_l)$ is the estimated mean at a specific value x_l under selected order $\hat{p}$ in ith repetition, and $\hat{\gamma}_{i\hat{p}h}$ is the estimator of γ_h with $\hat{p}$ in ith repetition.

In Tables 2 and 3, the values in the first, third and fifth columns are all positive, except for the linear trend with independent error for measure 2. These results show that the constrained penalised spline estimator behaves better than the unconstrained penalised spline estimator in the estimation of both the trend and correlation. For linear data, comparing with unconstrained spline estimator, the proposed estimator still improves around 5–10% for measure 1, 3–19% for measure 2 and 1–13% for measure 3. For both cubic data and sigmoid data, the superiority of the proposed estimator in the estimation of both the trend and correlation is quite evident, where the improvement is around 26–57% for cubic data and 11–31% for sigmoid data. The improvements have an increasing trend with the increase of the correlation, because the constrained estimator is less sensitive than the unconstrained estimator to the increase of correlation.

For the linear data, all the values in the first, third and fifth columns in both Tables 2 and 3 are negative. We cannot expect a nonparametric method to perform better than the correct parametric method. For cubic data and sigmoid data, the improvement of proposed method is noteworthy for small amount of correlation. But for the estimation of $\boldsymbol{\theta}$, the proposed method performs about as

Table 2. The simulated percentage of the proposed estimator's relative improvement in the three measures, comparing with the unconstrained spline estimator and the classical linear regression estimator.

		Δ_{θ}		Δ_{γ}		Δ_{f}	
f	θ	Unconstrained	Linear	Unconstrained	Linear	Unconstrained	Linear
Linear	0	5.2	−41.4	−0.41	−26.5	1.0	−40.8
	0.3	8.8	−26.2	8.06	−34.5	5.0	−40.3
	0.5	5.8	−26.2	9.18	−34.0	5.9	−40.8
	0.7	13.3	−25.5	19.34	−33.8	13.8	−36.8
Cubic	0	60.6	284.6	25.5	192.1	25.5	224.0
	0.3	29.1	37.4	34.7	82.7	30.1	145.0
	0.5	29.7	1.5	46.6	24.9	30.2	89.1
	0.7	35.9	−15.5	57.2	−12.7	35.4	37.4
Sigmoid	0	27.8	204.4	10.5	143.8	15.0	195.1
	0.3	15.8	30.0	19.9	91.3	13.2	115.4
	0.5	14.9	−4.7	21.1	20.2	15.4	71.0
	0.7	18.1	−21.3	29.9	−20.6	18.9	25.1

Notes: Each value is calculated as a ratio. The numerator is the difference of measures, i.e. measure of constrained estimator minus that of the unconstrained estimator or the linear estimator. The denominator is the corresponding measure of the constrained estimator. The data sets are generated by three different kinds of true mean functions, linear, truncated cubic and sigmoid, with AR(1) errors, where $\theta = 0, 0.3, 0.5, 0.7$.

Table 3. The simulated percentage of the proposed estimator's relative improvement in the three measures, comparing with the unconstrained spline estimator and the classical linear regression estimator.

		Δ_{θ}		Δ_{γ}		Δ_{f}	
f	p	Unconstrained	Linear	Unconstrained	Linear	Unconstrained	Linear
Linear	0	5.2	−41.4	−0.4	−26.5	1.0	−40.8
	1	8.8	−26.2	8.1	−34.5	5.0	−40.3
	2	7.8	−6.1	3.7	−28.2	4.9	−43.2
	3	10.2	−31.6	3.9	−30.0	8.2	−46.0
	4	16.2	−34.0	6.6	−24.8	12.2	−46.6
Cubic	0	60.6	284.6	25.5	192.1	25.5	224.0
	1	29.1	37.4	34.7	82.7	30.1	145.0
	2	38.6	21.6	37.4	63.7	36.2	141.5
	3	46.2	16.8	34.4	71.8	42.9	151.6
	4	58.8	8.5	32.2	60.4	46.1	145.7
Sigmoid	0	27.8	204.4	10.5	143.8	15.0	195.1
	1	15.8	30.0	19.9	91.2	13.2	115.4
	2	17.1	12.8	14.2	67.3	16.1	118.1
	3	24.6	0.8	16.1	66.2	18.1	111.4
	4	30.9	−5.8	13.6	67.5	20.9	113.5

Notes: Each value is calculated as a ratio. The numerator is the difference of measures, i.e. measure of constrained estimator minus that of the unconstrained estimator or the linear estimator. The denominator is the corresponding measure of the constrained estimator. The data sets are generated by three different kinds of true mean functions, linear, truncated cubic and sigmoid, with AR(p) errors, where $\epsilon_i = 0.3\epsilon_{i-p} + z_i, p = 0, 1, 2, 3, 4$.

well as the linear model for large correlation. There are some extreme large positive values for cubic and sigmoid data with independent observations in the first, third and fifth columns in both Tables 2 and 3. These values demonstrate that incorrectly assuming a parametric form would cost a great deviation when there is no prior information of the parametric family of the trend. The deviation is evident when there is no correlation, and would be obscured when the correlation is increasing.

Table 4. The simulated percentage of increase in measure 3 by ignoring the correlation, comparing with the estimation by proposed iteration.

f	$\boldsymbol{\theta}$	Constrained	Unconstrained	Linear
Linear	0.3	30.23	37.24	−2.54
	0.5	33.59	56.92	−0.12
	0.7	29.67	65.60	4.45
Cubic	0.3	4.00	20.78	0.05
	0.5	11.38	38.06	0.02
	0.7	11.27	43.19	−0.40
Sigmoid	0.3	8.11	13.94	−0.01
	0.5	12.21	23.94	−0.08
	0.7	14.70	37.25	−1.12

Notes: Each value is calculated as a ratio. The numerator is the difference of measures, i.e. measure of fit ignoring correlation minus the measure of fit obtaining from proposed iteration. The denominator is the corresponding measure of the proposed iteration estimator. The data sets are generated by three different kinds of true mean functions, linear, truncated cubic and sigmoid, with AR(1) errors, where $\theta = 0.3, 0.5, 0.7$. Sample size is 500.

Finally, we investigate whether incorporating the correlation actually improves the estimation of the mean function, relative to the simpler estimator that completely ignores the correlation. Incorporating the correlation into the estimation procedure improves the estimation of trend, for both constrained spline regression and unconstrained spline regression. Table 4 shows the differences of effects whether estimating the correlation by the proposed iteration or ignoring the correlation for those three regression models. The increase of constrained estimator is smaller than that of the unconstrained estimator. Because of incorporation of qualitative knowledge of trend, the estimation of trend is more robust to the estimation of correlation. The increment of simple linear regression is almost zero and sometimes even negative.

5. Global temperature data

There has been much interest in the research of the global temperature change. Hansen et al. (2006) have a discussion on the pattern of global warming. In this article, we use the 'Global Annual Mean Surface Air Temperature Change Data' from 1882 to 2008 to demonstrate the behaviour of the proposed estimator. The data set comes from http://data.giss.nasa.gov/gistemp/graphs_v3/Fig.A2.txt. Assume that the global annual temperature is a stationary auto-regressive process with a monotone increasing tendency during the 1882 to 2008. We fit the data with the monotone constrained penalised spline regression and compare the performance with the unconstrained penalised spline estimator and the classical linear regression estimator. The correlated-adjusted AIC in this paper would be used to select the penalty parameter and the order p. We fit this data with 20 knots and 35 knots for a comparison. The results are in Figures 1 and 2.

For the situation of 20 knots, $\hat{p} = 2$ and $\hat{\boldsymbol{\theta}} = (0.278, -0.138)$ for constrained penalised spline estimator. This looks more reasonable than the results of the unconstrained penalised estimator, where $\hat{p} = 5$ and $\hat{\boldsymbol{\theta}} = (0.251, -0.160, -0.133, 0.114, -0.224)$. The penalty parameters for both the constrained and unconstrained regression are 0.024, so that the corresponding effective degree of freedom of constrained estimator is 8 and that of the unconstrained estimator is 14. Unconstrained spline regression tends to select the smallest value by AIC among all the candidate λ, which easily leads to overfitting and generates a wiggly curve. The linear regression is inclined to overestimate the correlation as expected, selecting $\hat{p} = 4$ and $\hat{\boldsymbol{\theta}} = (0.507, 0.003, 0.050, 0.222)$. Woodward and Gray (1993) point out that statistical tests based on simple linear model have little or no ability

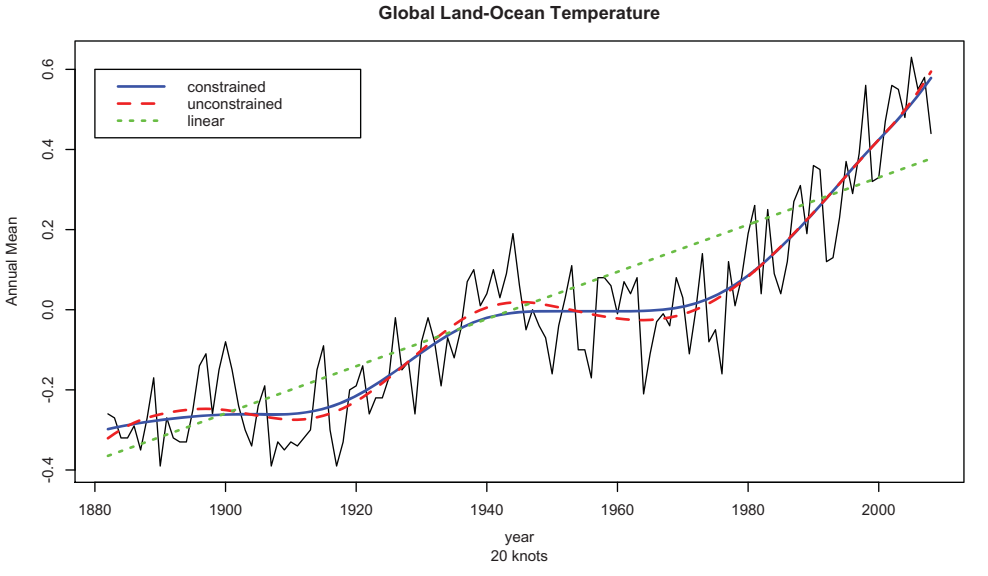

Figure 1. Estimated global temperature trends, using constrained penalised spline estimators and unconstrained penalised spline estimator both with 20 knots and penalty and correlation order selected with AIC, and linear regression fit.

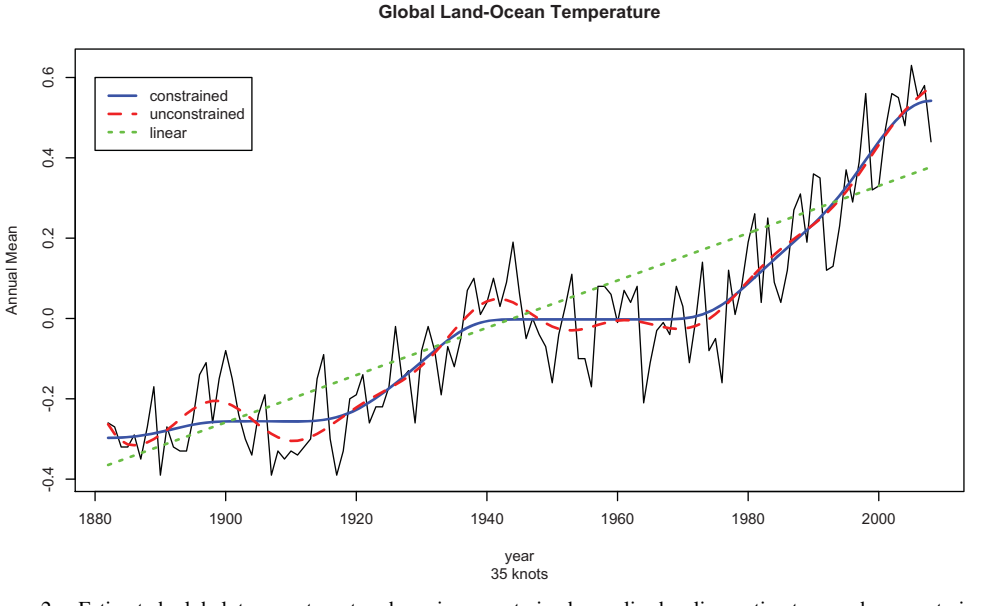

Figure 2. Estimated global temperature trends, using constrained penalised spline estimators and unconstrained penalised spline estimator both with 35 knots and penalty and correlation order selected with AIC, and linear regression fit.

to distinguish the realisations from the ARMA model with high correlation and those from the linear model. If the number of knots is increased to 35, the constrained penalised spline estimator still estimate $\hat{p} = 2$, and $\hat{\boldsymbol{\theta}} = (0.256, -0.158)$, which is quite similar to the case with 20 knots. But the behaviour of the unconstrained penalised spline estimator becomes very unstable with $\hat{p} = 9$ and $\hat{\boldsymbol{\theta}} = (0.0247, -0.334, -0.333, -0.116, -0.364, -0.176, -0.154, -0.111, -0.186)$. From Figure 2, the unconstrained fit becomes more wiggly, but the constrained fit has little change.

We also compared the proposed estimation with constrained spline estimation with ignoring the correlation on the Global Temperature Data. But the two curves are almost identical to each other. The difference of the fits is quite small if we use the constrained method.

In conclusion, this example illustrates the robustness to knot choice of the constrained spline regression, and the ability of the proposed AIC criterion to select suitable values for the penalty and the order of the correlation automatically. Meyer (2012a) has previously found the robustness of constrained spline regression to penalty choice for independent data, so this confirms these good practical properties for situations with correlated data.

6. Conclusion

The asymptotic rate for constrained penalised spline estimators with estimation of correlation and ignoring the correlation have been proved to be the same. Even if we have an inconsistent estimator of correlation, as long as it satisfies appropriate conditions, the estimation of trend based on this estimator is still consistent and attains the optimal rate. However, as illustrated in Table 4, estimation of the trend is substantially improved for moderate-sized samples under proposed iteration method. Further, the asymptotic variances of the three estimators are different. In ongoing research, we are studying the hypothesis tests of the trend, such as, constant vs. monotone, in the presence of AR(p) errors. The asymptotic distribution of the test statistic depends on consistent estimation of the correlation.

The test statistic has an approximating χ^2-distribution for known σ^2 and approximating *Beta*-distribution for unknown σ^2 only when the estimation of correlation is consistent. If the correlation is ignored, the test size will be greatly expanded.

References

Altman, N. (1992), 'An Iterated Cochrane-Orcutt Procedure for Nonparametric Regression', *Journal of Statistical Computation and Simulation*, 40(1–2), 93–108.

Altman, N.S. (1990), 'Kernel Smoothing of Data with Correlated Errors', *Journal of the American Statistical Association*, 85 (411), 749–759.

Brockwell, P., and Davis, R. (2009). *Time Series: Theory and Methods*, Springer Series in Statistics, New York: Springer.

Brunk, H.D. (1955), 'Maximum Likelihood Estimates of Monotone Parameters', *The Annals of Mathematical Statistics*, 26(4), 607–616.

Brunk, H.D. (1958), 'On the Estimation of Parameters Restricted by Inequalities', *The Annals of Mathematical Statistics*, 29, 437–454.

Claeskens, G., Krivobokova, T., and Opsomer, J.D. (2009), 'Asymptotic Properties of Penalized Spline Estimators', *Biometrika*, 96(3), 529–544.

Diggle, P.J., and Hutchinson, M.F. (1989), 'On Spline Smoothing with Autocorrelated Errors', *Australian Journal of Statistics*, 31(1), 166–182.

Eilers, P.H.C., and Marx, B.D. (1996), 'Flexible Smoothing with B-Splines and Penalties', *Statistical Science*, 11(2), 89–102.

Francisco-Fernandez, M., and Opsomer, J. (2005), 'Smoothing Parameter Selection Methods for Nonparametric Regression with Spatially Correlated Errors', *Canadian Journal of Statistics*, 33(2), 279–295.

Fuller, W. (1996). *Introduction to Statistical Time Series*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley.

Hall, P., and Huang, L.-S. (2001), 'Nonparametric Kernel Regression Subject to Monotonicity Constraints', *The Annals of Statistics*, 29(3), 624–647.

Hall, P. and Keilegom, I. (2003), 'Using Difference-Based Methods for Inference in Nonparametric Regression with Time Series Errors', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65 (2), 443–456.

Hall, P., and Opsomer, J.D. (2005), 'Theory for Penalised Spline Regression', *Biometrika*, 92(1), 105–118.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D.W., and Medina-Elizade, M. (2006), 'Global Temperature Change', *Proceedings of the National Academy of Sciences*, 103(39), 14288–14293.

Hart, J.D. (1991), 'Kernel Regression Estimation with Time Series Errors', *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1), 173–187.

Hart, J.D. (1994), 'Automated Kernel Smoothing of Dependent Data by Using Time Series Cross- Validation', *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3), 529–542.

Huang, J.Z. (1998), 'Projection Estimation in Multiple Regression with Application to Functional ANOVA Models', *The Annals of Statistics*, 26(1), 242–272.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009), 'Some Asymptotic Results on Generalized Penalized Spline Smoothing', *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 71(2), 487–503.

Kim, T., Park, B., Moon, M., and Kim, C. (2009), 'Using Bimodal Kernel for Inference in Nonparametric Regression with Correlated Errors', *Journal of Multivariate Analysis*, 100(7), 1487–1497.

Li, Y., and Ruppert, D. (2008), 'On the Asymptotics of Penalized Splines', *Biometrika*, 95(2), 415–436.

Mammen, E., and Thomas-Agnan, C. (1999), 'Smoothing Splines and Shape Restrictions', *Scandinavian Journal of Statistics*, 26, 239–252.

Meyer, M.C. (2008), 'Inference Using Shape-Restricted Regression Splines', *The Annals of Applied Statistics*, 2(3), 1013–1033.

Meyer, M.C. (2012a), 'Constrained Penalized Splines', *Canadian Journal of Statistics*, 40(1), 190–206.

Meyer, M.C. (2012b), 'A Simple New Algorithm for Quadratic Programming with Applications in Statistics', *Communications in Statistics-Simulation and Computation*, 42(5), 1126–1139.

Opsomer, J., Wang, Y., and Yang, Y. (2001), 'Nonparametric Regression with Correlated Errors', *Statistical Science*, 16(2), 134–153.

Ramsay, J.O. (1998), 'Estimating Smooth Monotone Functions', *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 60, 365–375.

Robertson, T., Wright, F., and Dykstra, R. (1988), *Order Restricted Statistical Inference*, Probability and Statistics Series, Canada: John Wiley & Sons.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Silvapulle, M., and Sen, P. (2004). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley.

Stone, C.J. (1982), 'Optimal Global Rates of Convergence for Nonparametric Regression', *The Annals of Statistics*, 10(4), 1040–1053.

Tantiyaswasdikul, C., and Woodroofe, M.B. (1994), 'Isotonic Smoothing Splines Under Sequential Designs', *Journal of Statistical Planning and Inference*, 38, 75–87.

Wang, Y. (1998), 'Smoothing Spline Models with Correlated Random Errors', *Journal of the American Statistical Association*, 93(441), 341–348.

Woodward, W., and Gray, H. (1993), 'Global Warming and the Problem of Testing for Trend in Time Series Data', *Journal of Climate*, 6(5), 953–962.