Constrained Spline Regression in the Presence of Correlated Errors

Huan Wang, Mary C. Meyer, Jean D. Opsomer
Department of Statistics, Colorado State University, Fort Collins, CO

Abstract

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 assumption 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 for estimation and inference in the presence of correlated errors. Standard criteria for selection of penalty parameter, such as Akaike information criterion (AIC), cross-validation and generalized 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 unpenalized 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. The constrained spline fit attains the convergence rate of unconstrained spline fit in the presence of correlated observations. Simulation results show that the constrained estimator typically behaves better than the unconstrained version if the true trend satisfies the constraints.

1 Literature Review

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) propose the unsmoothed monotone regression estimation and studied its asymptotic behavior. See Robertson et al. (1988) for details about estimation inference. Ramsay (1998) propose a device to estimate a smooth strictly monotone function of arbitrary flexibility. Tantiyaswasdikul and Woodroofe (1994) propose the monotone smoothing splines with penalty on the integrated first derivative. Mammen and Thomas-Agnan (1999) show 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) develope a biased-bootstrap method for monotonizing general linear, kernel-typed estimators. Meyer (2008) propose an algorithm for the cubic monotone case, and also extended the method to convex constraints and variants such as increasing-concave.

Extending an original idea of O'Sullivan (1986), Eilers and Marx (1996) introduces penalized splines. Penalized splines use a large number of knots compared to regression splines, but less than in smoothing splines, and hence are less computationally cumbersome. The penalization shrinks the coefficients towards zero, constraining their influence and resulting in a less variable fit than regression splines. Penalized splines are increasingly popular in handling a wide range of nonparametric and semiparametric problems. Ruppert et al. (2003) provides details of this method. Hall and Opsomer (2005) use a white-noise process representation of the penalized 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 enables 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) show that penalized splines behave similarly to Nadaraya-Watson kernel estimators with equivalent kernels. By this equivalent kernel representation, they developed an asymptotic theory of penalized splines for the cases of piecewise-constant or linear splines, with a first- or second-order difference penalty. Claeskens et al. (2009) develop a general theory of the asymptotic properties of penalized spline estimators for any order of spline and general penalty. They demonstrated that the theoretical properties of penalized spline estimators are either similar to those of regression splines or to those of smoothing splines, with a clear breakpoint distinguishing the cases. Kauermann et al. (2009) use 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 AIC, will break down if the correlation is ignored. Diggle and Hutchinson (1989) present an extension of generalized cross-validation which accommodates a known correlation matrix for the errors. Altman (1990) suggests two methods, a direct method and an indirect method, for correcting the selection criteria when the correlation function is known. Hart (1991) uses a risk estimation procedure to select the bandwidth in the kernel regression with correlated errors. Hart (1994) proposes a time series cross-validation to estimate the bandwidth and give a time series model for the errors simultaneously. Wang (1998) extends the generalized maximum likelihood, generalized cross-validation 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 et al. (2001) give a general review of the literature in kernel regression, smoothing splines and wavelet regression under correlation. Hall and Keilegom (2003) use difference-based methods to construct estimators of error variance and autoregressive parameters in nonparametric regression with time series errors. They also prove that the difference-based estimators can be used to produce a simplified version of time series cross-validation. Francisco-Fernandez and Opsomer (2005) propose to adjust the generalized cross-validation (GCV) criterion for the spatial correlation and show that it leads to improved smoothing parameter selection results even when the covariance model is misspecified. Kim et al. (2009) investigate a bandwidth selector based on the use of a bimodal kernel for nonparametric regression with fixed design and prove that the proposed selector is quite effective when the errors are severely correlated.

In this article, we propose a constrained penalized spline estimator for the correlated observations. The correlation type is restricted to an autoregressive time series process with unknown order and unknown correlation parameters. A new correlation-adjusted AIC is given for the selection of penalty parameter and autoregressive parameter simultaneously. We prove the asymptotic properties of the constrained spline estimator in the presence of correlation.

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 spine 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 analyze 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 $1 \leq i \leq n$, are generated by the model

$$y_i = f(x_i) + \sigma\varepsilon_i, \tag{1}$$

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

$$\varepsilon_i = \sum_{j=1}^{p} \theta_j \varepsilon_{i-j} + e_i, \tag{2}$$

where e_i are independent standard normal random variables.

The function f is estimated by a linear combination of spline basis functions. Specify a set of knots $0 = t_1 < \cdots < t_k = 1$. A set of $m = k + d - 1$ basis functions $b_1(x), \cdots, b_m(x)$ 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 $\mathbf{b}_1, \cdots, \mathbf{b}_m$ 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 penalized sum of squares of Eilers and Marx (1996) is:

$$\sum_{i=1}^{n} [y_i - \sum_{j=1}^{m} \alpha_j b_j(x_i)]^2 + \lambda \sum_{j=q+1}^{m} (\Delta^q \alpha_j)^2, \tag{3}$$

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 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)^{'}$. The penalty parameter $\lambda \geq 0$ controls the smoothness. Minimizing the penalized sum of squares is equivalent to minimizing the vector expression:

$$\psi(\boldsymbol{\alpha}; \mathbf{y}) = \boldsymbol{\alpha}^{'}(\mathbf{B}^{'}\mathbf{B} + \lambda \mathbf{D}^{'}\mathbf{D})\boldsymbol{\alpha} - 2\mathbf{y}^{'}\mathbf{B}\boldsymbol{\alpha}, \tag{4}$$

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

$$\mathcal{C} = \{\boldsymbol{\alpha} : \mathbf{T}\boldsymbol{\alpha} \geq \mathbf{0}\} \subseteq \mathbb{R}^m. \tag{5}$$

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 $\mathbf{T}\boldsymbol{\alpha} \geq \mathbf{0}$.

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 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 corresponding to (2.4) is

$$\psi(\boldsymbol{\alpha}; \tilde{\mathbf{y}}) = \boldsymbol{\alpha}^{'}(\tilde{\mathbf{B}}^{'}\tilde{\mathbf{B}} + \lambda \mathbf{D}^{'}\mathbf{D})\boldsymbol{\alpha} - 2\tilde{\mathbf{y}}^{'}\tilde{\mathbf{B}}\boldsymbol{\alpha}, \tag{6}$$

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

$$\psi(\boldsymbol{\alpha}; \tilde{\mathbf{y}}) = \psi(\boldsymbol{\phi}; \mathbf{z}) = \| \boldsymbol{\phi} - \mathbf{z} \|^2, \tag{7}$$

where $\boldsymbol{\phi}$ is restricted to $\tilde{\mathcal{C}} = \{\boldsymbol{\phi} : \mathbf{A}\boldsymbol{\phi} \geq \mathbf{0}\} \subseteq \mathbb{R}^m$, which is 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 (2013), let $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_{m-k}$ 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 $\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_k$ of the cone, therefore the cone can be written as

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

The minimizer $\hat{\boldsymbol{\phi}}$ 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

$$\mathcal{F}_J = \left\{ \boldsymbol{\phi} : \boldsymbol{\phi} = \sum_{i=1}^{m-k} \beta_i \boldsymbol{\nu}_i + \sum_{j \in J} \beta_j \boldsymbol{\delta}_j, \quad \boldsymbol{\nu}_i \in \mathcal{V}, \quad \beta_j > 0, j \in J \right\}.$$

The interior of the cone is a face with $J = \{1, \ldots, m\}$, and the origin is the face with $J = \emptyset$. Here, we use the Hinge Algorithm from Meyer (2013) to determine the face $\mathcal{F}_J$ on which the projection falls, so that the estimation 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. $edf = tr(\mathbf{P}_{\mathbf{R}}^c)$.

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 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 the known $\mathbf{R}$ is $\hat{\boldsymbol{\mu}}^u_{\mathbf{R}} = \mathbf{B}\hat{\boldsymbol{\alpha}}_u$. Hence 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 constrained fit is a random quantity with $m + 1$ possible values, the largest of which is that of the unconstrained version. Meyer (2012) discusses this for the independent error case.

However, typically $cor(\boldsymbol{\varepsilon}) = \mathbf{R}$ is unknown. Here, we assume AR(p) and use Cochran-Orcutt iterations to estimate the matrix $\mathbf{R}$. Given p and λ, the Cochran-Orcutt 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. Estimate covariance function:
   - Use the Yule-Walker method in Chapter 8 of Brockwell and Davis (2009) and residual vector $\hat{\boldsymbol{\epsilon}}$ to estimate coefficients $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ and the error variance $\gamma(0)$ and $\boldsymbol{\gamma}_p = (\gamma_1, \cdots, \gamma_p)^{'}$; then obtain $\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\theta}}^{'}\hat{\boldsymbol{\gamma}}_p$;
   - Use Cholesky decompositon $\hat{\mathbf{R}} = \hat{\boldsymbol{\Sigma}}/\hat{\gamma}(0) = \mathbf{L}\mathbf{L}^{'}$, to transform data and basis into $\tilde{\mathbf{y}} = \mathbf{L}^{-1}\mathbf{y}, \tilde{\mathbf{B}} = \mathbf{L}^{-1}\mathbf{B}$.
3. 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), until convergence, obtaining the final estimators $\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}}^c_{\hat{\mathbf{R}}}$.

By the result obtained from the Cochran-Orcutt iteration procedure, we can compute the correlation-adjusted Akaike information criterion (AIC)

$$AIC = n\log(\hat{\sigma}^2) + 2(p + edf). \tag{8}$$

We use this criterion to choose p and λ simultaneously.

Most commonly used data-driven selection methods for tuning parameter such as generalized cross-validation (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 unpenalized regression in the presence of correlation. Then using those results, the theorem on the convergence rate for unpenalized constrained spline regression estimator is 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 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 $\|\boldsymbol{\mu}\|^2 = \frac{1}{n}\langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$, 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,

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}) + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}). \tag{9}$$

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,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\| + \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|. \tag{10}$$

Therefore, we can examine separately the contributions to the integrated squared error from the two parts in the decomposition. 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, coming from 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 1. *There are positive constants A_n such that, $\|f\|_\infty \leq A_n\|f\|$ for all $f \in \mathbf{G}_n$ and $\lim_n A_n^2 N_n/n = 0$.*

Condition 2. *There are nonnegative numbers $\rho_n = \rho_n(\mathbf{G}_n)$ such that, for $\boldsymbol{\mu} \in \mathbf{H}$,*

$$\inf_{\mathbf{g}\in\mathbf{G}_n} \|\mathbf{g} - \boldsymbol{\mu}\|_\infty \leq \rho_n \to 0 \quad as \quad n \to \infty. \tag{11}$$

and $\lim sup_n A_n\rho_n < \infty$.

If $\mathbf{H}$ is finite-dimensional, then we choose $\mathbf{G}_n = \mathbf{H}$, for all n. Condition 1 is automatically satisfied with A_n independent of n, and Condition 2 is satisfied with $\rho_n = 0$.

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

Condition 3. *Let $\gamma_{|i-j|} = E\varepsilon_i\varepsilon_j$, then there is a positive constant $M \in \mathbb{R}^1$, such that $\sum_{i=1}^{\infty} |\gamma_i| \leq M$.*

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

Now, we state the main results.

Theorem 1. *Let $\mathbf{P_I}$ be the projection matrix of the ordinary least squares estimation, then $\hat{\boldsymbol{\mu}}_\mathbf{I} = \mathbf{P_I y}$ and $\tilde{\boldsymbol{\mu}}_\mathbf{I} = \mathbf{P_I}\boldsymbol{\mu}$. If conditions 1, 2 and 3 hold, then*

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

Consequently,

$$\|\hat{\boldsymbol{\mu}}_\mathbf{I} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2). \tag{12}$$

Huang (1998) derives 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, $\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = \frac{1}{n}\sum_j \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}_j \rangle^2$, and

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

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 2, we can find $\mathbf{g} \in \mathbf{G}_n$ such that $\|\boldsymbol{\mu} - \mathbf{g}\|_\infty \le 2\rho_n$ and hence $\|\boldsymbol{\mu} - \mathbf{g}\| \le 2\rho_n$. Then we have that

$$\|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \mathbf{g}\|^2 = \|\mathbf{P}(\boldsymbol{\mu} - \mathbf{g})\|^2 \le \|\boldsymbol{\mu} - \mathbf{g}\|^2.$$

Hence, by the triangle inequality,

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

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

Theorem 2. *Let $\mathbf{P_R}$ be the matrix in Equation* **??** *of the weighted least squares estimation with the known correlation matrix $\mathbf{R}$, then $\hat{\boldsymbol{\mu}}_\mathbf{R} = \mathbf{P_R y}$ and $\tilde{\boldsymbol{\mu}}_\mathbf{R} = \mathbf{P_R}\boldsymbol{\mu}$. If Conditions 1, 2, 3 and 4 hold, then*

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

Consequently,

$$\|\hat{\boldsymbol{\mu}}_\mathbf{R} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2). \tag{13}$$

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

$$\mathbf{y}^* = \boldsymbol{\mu}^* + \boldsymbol{\varepsilon}^*, \quad E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*'}) = \mathbf{I}.$$

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 the Theorem 2.1 in Huang (1998), we have

$$\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 = O_p(N_n/n), \quad \|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2 = O(\rho_n^2).$$

Then

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 &= \|\mathbf{L}^{-1}(\hat{\boldsymbol{\mu}}_\mathbf{R} - \tilde{\boldsymbol{\mu}}_\mathbf{R})\|^2 \\ &= \frac{1}{n}(\hat{\boldsymbol{\mu}}_\mathbf{R} - \tilde{\boldsymbol{\mu}}_\mathbf{R})^{'}\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_\mathbf{R} - \tilde{\boldsymbol{\mu}}_\mathbf{R}). \end{aligned}$$

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 λ_{min} and λ_{max} are the smallest and largest eigenvalues of $\mathbf{R}^{-1}$. For $\mathbf{R}$ is 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 the Proposition 4.5.3 in Brockwell and Davis (2009), for a stationary AR(p) process, the eigenvalues of its covariance matrix are bounded 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

$$\frac{1}{M}\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2 \leq \|\hat{\boldsymbol{\mu}}_\mathbf{R} - \tilde{\boldsymbol{\mu}}_\mathbf{R}\|^2 \leq \frac{1}{m}\|\hat{\boldsymbol{\mu}}^* - \tilde{\boldsymbol{\mu}}^*\|^2.$$

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

$$\frac{1}{M}\|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2 \leq \|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 \leq \frac{1}{m}\|\tilde{\boldsymbol{\mu}}^* - \boldsymbol{\mu}^*\|^2.$$

Therefore, $\|\tilde{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(\rho_n^2)$. So, we have $\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2)$. □

Theorem 3. *If the correlation matrix* $\mathbf{R}$ *is unknown, and there is an estimator matrix* $\mathbf{S}$ *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* $\mathbf{L_S^{-1}R(L_S^{-1})'}$ *satisfies Condition 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 matrix* $\mathbf{S}$ *an estimator of the correlation matrix, then* $\hat{\boldsymbol{\mu}}_{\mathbf{S}} = \mathbf{P_S y}$ *and* $\tilde{\boldsymbol{\mu}}_{\mathbf{S}} = \mathbf{P_S}\boldsymbol{\mu}$*. If conditions 1 and 2 hold, then*

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

Consequently,

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(N_n/n + \rho_n^2). \tag{14}$$

Proof. Let $\mathbf{y}_{\mathbf{S}}^* = \mathbf{L_S^{-1}y}, \boldsymbol{\mu}_{\mathbf{S}}^* = \mathbf{L_S^{-1}}\boldsymbol{\mu}, \boldsymbol{\varepsilon}_{\mathbf{S}}^* = \mathbf{L_S^{-1}}\boldsymbol{\varepsilon}$, then the model can be transformed into

$$\mathbf{y}_{\mathbf{S}}^* = \boldsymbol{\mu}_{\mathbf{S}}^* + \boldsymbol{\varepsilon}_{\mathbf{S}}^*, \quad E(\boldsymbol{\varepsilon}_{\mathbf{S}}^*\boldsymbol{\varepsilon}_{\mathbf{S}}^{*'}) = \mathbf{L_S^{-1}R(L_S^{-1})'}.$$

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 $\mathbf{L_S^{-1}R(L_S^{-1})'}$ satisfying the Condition A3, then we have

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

by Theorem 1. Therefore,

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^* - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^*\|^2 &= \|\mathbf{L_S^{-1}}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})\|^2 \\ &= \frac{1}{n}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}})'\mathbf{S}_n^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}). \end{aligned}$$

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 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

$$\frac{1}{M}\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^{*} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^{*}\|^2 \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}\|^2 \leq \frac{1}{m}\|\hat{\boldsymbol{\mu}}_{\mathbf{S}}^{*} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}^{*}\|^2.$$

Therefore, $\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \tilde{\boldsymbol{\mu}}_{\mathbf{S}}\|^2 = O_p(N_n/n)$. By a similar proof, we have

$$\frac{1}{M}\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^{*} - \boldsymbol{\mu}_{\mathbf{S}}^{*}\|^2 \leq \|\tilde{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 \leq \frac{1}{m}\|\tilde{\boldsymbol{\mu}}_{\mathbf{S}}^{*} - \boldsymbol{\mu}_{\mathbf{S}}^{*}\|^2.$$

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 1, 2 and 3 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 $\rho_n = 0$. Assume $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters. If $\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{S}^{-1}\mathbf{X}$ are all nonsingular, we have*

$$\hat{\boldsymbol{\mu}}_{\mathbf{I}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \tilde{\boldsymbol{\mu}}_{\mathbf{I}} = \boldsymbol{\mu};$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{R}} = \mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}, \quad \tilde{\boldsymbol{\mu}}_{\mathbf{R}} = \boldsymbol{\mu};$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{S}} = \mathbf{X}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}\mathbf{y}, \quad \tilde{\boldsymbol{\mu}}_{\mathbf{S}} = \boldsymbol{\mu};$$

Applying the Theorems 1, 2 and 3 to this setting, we have the following results:

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\| = O_p(n^{-1/2});$$

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\| = O_p(n^{-1/2});$$

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\| = O_p(n^{-1/2}).$$

Theorem 4. *Let $\hat{\boldsymbol{\mu}}$ be a consistent estimator of $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$, $\hat{\gamma}_{\hat{\boldsymbol{\varepsilon}}}(h) = \frac{1}{n}\sum_{i=1}^{n-h}\hat{\varepsilon}_i\hat{\varepsilon}_{i+h}$, $\hat{\gamma}_{\boldsymbol{\varepsilon}}(h) = \frac{1}{n}\sum_{i=1}^{n-h}\varepsilon_i\varepsilon_{i+h}$, $\gamma(h) = \mathbf{E}\varepsilon_i\varepsilon_{i+h}$, where $i = 1, \ldots, n-h; h = 0, 1, \ldots, n-1$. Let $\boldsymbol{\gamma}_{\hat{\boldsymbol{\varepsilon}}} = (\gamma_{\hat{\boldsymbol{\varepsilon}}}(1), \ldots, \gamma_{\hat{\boldsymbol{\varepsilon}}}(n-1))'$ and $\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}} = (\gamma_{\boldsymbol{\varepsilon}}(1), \ldots, \gamma_{\boldsymbol{\varepsilon}}(n-1))'$, then*

$$\|\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\varepsilon}}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}\|^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2), \quad \|\hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}} - \boldsymbol{\gamma}\|^2 = O_p(1/n). \tag{15}$$

Consequently,

$$\|\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\varepsilon}}} - \boldsymbol{\gamma}\|^2 = O_p\left(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 + \frac{1}{n}\right). \tag{16}$$

Proof.

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

If $\hat{\boldsymbol{\mu}}$ is consistent,

$$\|\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\varepsilon}}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}\|^2 = \frac{1}{n-1}\sum_{h=1}^{n-1}[\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\varepsilon}}}(h) - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\varepsilon}}(h)]^2 = O_p(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2).$$

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

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

□

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

3.2 Fixed-Knot Unpenalized Unconstrained Spline Regression

Theorems 1, 2 and 3 can also be applied to fixed-knot spline estimates when the knot positions are pre-specified but the number of knots is allowed to increase with the sample size. In this section, we investigate the large sample theory for only unpenalized situation, i.e. $\lambda = 0$. Let $\mathbf{G}_n$ be the linear space of regression splines with degree $q \geq p-1$. 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 form zero and ∞. Then we have $N_n \asymp 1/a_n$ and $\rho_n \asymp a_n^p \asymp N_n^{-p}$. So, 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 $O_p(\frac{a_n}{n} + a_n^{2p})$. 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, $\frac{a_n}{n} \asymp a_n^{2p}$. Applying the theorems 1, 2 and 3 to this setting, we obtain the following results.

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

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 = O_p(n^{-2p/(2p+1)}), \quad \|\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O(n^{-2p/(2p+1)}). \tag{17}$$

Consequently,

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)}); \tag{18}$$

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)}); \tag{19}$$

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{S}} - \boldsymbol{\mu}\|^2 = O_p(n^{-2p/(2p+1)}). \tag{20}$$

3.3 Fixed Knot Constrained Unpenalized Spline Regression

In the theorems 1, 2 and 3, 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 $\mathbf{H}_c$.

Consider three kinds of estimators: ordinary least squares estimator, the weighted least squares estimators with known $\mathbf{R}$ and the weighted least square estimator using a given matrix $\mathbf{S}$ as an estimate 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 $J = \{1, \cdots, m\}$. 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 $\hat{\boldsymbol{\mu}}^u_{\mathbf{S}} = \mathbf{P}^u_{\mathbf{S}}\mathbf{y}$ and $\tilde{\boldsymbol{\mu}}^u_{\mathbf{S}} = \mathbf{P}^u_{\mathbf{S}}\boldsymbol{\mu}$.

Assume that $\tilde{\boldsymbol{\mu}}^u_{\mathbf{I}} \in \mathbf{G}^c_n$, so that the shape restrictions hold; otherwise, $\tilde{\boldsymbol{\mu}}^u_{\mathbf{I}}$ is not be consistent to $\boldsymbol{\mu}$, the value of which is assumed to be monotone. Under this assumption, it is easy to prove that $\tilde{\boldsymbol{\mu}}^u_{\mathbf{I}} = \hat{\boldsymbol{\mu}}^c_{\mathbf{I}}$. The same assumption and equation are also applied in the other two estimators, therefore $\tilde{\boldsymbol{\mu}}^u_{\mathbf{R}} = \hat{\boldsymbol{\mu}}^c_{\mathbf{R}}$ and $\tilde{\boldsymbol{\mu}}^u_{\mathbf{S}_n} = \hat{\boldsymbol{\mu}}^c_{\mathbf{S}_n}$. In this context, we use $\tilde{\boldsymbol{\mu}}_{\mathbf{I}}$ instead of $\tilde{\boldsymbol{\mu}}^u_{\mathbf{I}}$ and $\tilde{\boldsymbol{\mu}}^c_{\mathbf{I}}$. The same treatment is used for $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{R}}$. 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 5. *Let $f \in \mathbf{C}^{d+1}[0,1]$, and the knots $t_1, \ldots, t_k$ have bounded mesh ratio, then*

$$\|\hat{\boldsymbol{\mu}}^c_{\mathbf{I}} - \boldsymbol{\mu}\|^2 \leq \|\hat{\boldsymbol{\mu}}^u_{\mathbf{I}} - \boldsymbol{\mu}\|^2. \tag{21}$$

It means that 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

$$\hat{\boldsymbol{\mu}}^c_{\mathbf{I}} - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}}^c_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) + (\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu})$$

and

$$\hat{\boldsymbol{\mu}}^u_{\mathbf{I}} - \boldsymbol{\mu} = (\hat{\boldsymbol{\mu}}^u_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}) + (\tilde{\boldsymbol{\mu}}_{\mathbf{I}} - \boldsymbol{\mu}).$$

So we only need to prove $\|\hat{\boldsymbol{\mu}}^c_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 \leq \|\hat{\boldsymbol{\mu}}^u_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2$.

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

The Karush-Kuhn-Tucker conditions (see Silvapulle and Sen (2004) Appendix 1) imply,

$$\langle \mathbf{y} - \hat{\boldsymbol{\mu}}^c_{\mathbf{I}}, \hat{\boldsymbol{\mu}}^c_{\mathbf{I}} \rangle = 0 \quad \text{and} \quad \langle \mathbf{y} - \hat{\boldsymbol{\mu}}^c_{\mathbf{I}}, \tilde{\boldsymbol{\mu}}^c_{\mathbf{I}} \rangle \leq 0.$$

Therefore,

$$\|\hat{\boldsymbol{\mu}}^u_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2 \geq \|\hat{\boldsymbol{\mu}}^c_{\mathbf{I}} - \tilde{\boldsymbol{\mu}}_{\mathbf{I}}\|^2.$$

□

Theorem 6. *Let $f \in \mathbf{C}^{d+1}[0,1]$, and 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,*

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{c} - \boldsymbol{\mu}\|^2 \leq C\|\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{u} - \boldsymbol{\mu}\|^2. \tag{22}$$

It means that the convergence rate of the weighted least squares estimator with known correlation in constrained spline regression attains that of the corresponding unconstrained spline regression, under some appropriate conditions.

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

$$\mathbf{y}^* = \boldsymbol{\mu}^* + \boldsymbol{\varepsilon}^*, \quad E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*'}) = \mathbf{I}.$$

Using the result in Theorem 5, 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

$$\|\mathbf{L}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{c} - \mathbf{L}^{-1}\tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2 \leq \|\mathbf{L}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{u} - \mathbf{L}^{-1}\tilde{\boldsymbol{\mu}}_{\mathbf{R}}\|^2;$$

$$(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{c} - \tilde{\boldsymbol{\mu}})^{'}\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{c} - \tilde{\boldsymbol{\mu}}) \leq (\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{u} - \tilde{\boldsymbol{\mu}})^{'}\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{R}}^{u} - \tilde{\boldsymbol{\mu}}).$$

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, λ_{min} and λ_{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. 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 $\mathbf{R}$ are bounded from zero and ∞, uniformly in n, then the eigenvalues of $\mathbf{R}^{-1}$ are also bounded from zero and ∞, uniformly in n, which means that 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 7. *Let $f \in \mathbf{C}^{d+1}[0,1]$, and the knots $t_1, \ldots, t_k$ have "bounded mesh ratio". The correlation matrix $\mathbf{R}$ is unknown, and there is an estimator matrix $\mathbf{S}$, satisfying the conditions A1, A2 and A3. Then there exists a constant $C \in \mathbb{R}^1$, bounded away from zero and ∞, such that,*

$$\|\hat{\boldsymbol{\mu}}^c_{\mathbf{S}} - \boldsymbol{\mu}\|^2 \leq C\|\hat{\boldsymbol{\mu}}^u_{\mathbf{S}} - \boldsymbol{\mu}\|^2, \textit{with probability approaching one.} \tag{23}$$

It means that the convergence rate of the weighted least squares estimator with the given matrix as an estimator of correlation in constrained spline regression attains that in the corresponding unconstrained spline regression, under some appropriate conditions.

By Theorems 6 and 7, and the same technique used in the proof of Theorem 3, it is easy to prove this result.

4 Simulation

4.1 Data Set

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

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

Two series of noise will be 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 normal distributed with mean zero and standard deviation 0.2. The sample size for the simulation is 250 and the number of replications is 1000.

f	$\boldsymbol{\theta}$	AR(1)			p	AR(p)		
		constrained	unconstrained	linear		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

Table 1: The proportion of datasets for which the correlation-adjusted AIC criteria selects the true p, for the proposed estimator, the unconstrained penalized estimator and the classical linear regression estimator. The simulated data are generated by the 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 $\epsilon_i = 0.3\epsilon_{i-p} + z_i, p = 0, 2, 3, 4$.

4.2 The selection of order p and penalty parameter by AIC

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

$$AIC_{p,\lambda} = nlog(\hat{\sigma}^2) + 2(p + edf).$$

For each simulated data set, we compute 60 AIC values using $p = (0, 1, 2, 3, 4, 5)$ and ten values of λ as candidates. The effective degrees of freedom for both constrained and unconstrained estimators with unknown correlation are random. We choose the candidate λ by letting the corresponding effective degrees of freedom for unconstrained penalized spline estimator for independent data be $(4, 5, 6, 8, 10, 12, 16, 20, 25, 30)$. We choose p and λ as the joint minimizer of $AIC_{p,\lambda}$. 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 proportion for the constrained penalized estimator always greater than the corresponding unconstrained penalized estimator, which is just

as we expected that if the prior information about the shape is correct. The shape restricted regression can separate the noise and trend better than the corresponding regression without using this information. Linear regression behaves poorly for truncated cubic data and sigmoid data when the correlation is zero or small; it is more likely to choose a larger p. When the data are generated by linear trend, the linear regression does the best job but the behavior of the proposed estimator is still reasonable. We also conducted the simulation with larger p and more 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.

4.3 Three Performance Measures

To compare the performance of the proposed estimator with the unconstrained penalized spline estimator and classical linear regression estimator, the following three measures are constructed. The first measure is the average Euclidean distance of the $\hat{\theta}$ and the true θ. The second one is the average Euclidean distance of $\hat{\gamma}$ and the true γ. The third one is the average Euclidean distance of estimated mean and true mean. The first and the second measures are used to compare the estimation of the correlation. The third one is used to compare the estimation of the trend.

$$\Delta_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{\mathtt{K}}(\hat{\theta}_{i\hat{p}k} - \theta_k)^2};$$

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

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

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

In Tables 2 and 3, the values in the first column of each measure is the simulated improvement percentage of the proposed estimator, comparing to the unconstrained penalized spline estimator. While the values in second column of each measure is the simulated improvement percentage of the simple linear regression estimator, comparing to the unconstrained penalized spline estimator. Positive values mean that the proposed estimator did better than the corresponding estimators and

f	$\boldsymbol{\theta}$	$\Delta_{\boldsymbol{\theta}}$		$\Delta_{\boldsymbol{\gamma}}$		$\Delta_{\mathbf{f}}$	
		uncon	linear	uncon	linear	uncon	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

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. 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 datasets are generated by the three different kinds of true mean functions, linear, truncated cubic and sigmoid, with AR(1) errors, where $\theta = 0; 0.3; 0.5; 0.7$.

f	p	$\Delta_{\boldsymbol{\theta}}$		$\Delta_{\boldsymbol{\gamma}}$		$\Delta_{\mathbf{f}}$	
		uncon	linear	uncon	linear	uncon	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

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. 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 datasets are generated by the three different kinds of true mean functions, linear, truncated cubic and sigmoid, with AR(p) errors, where $\epsilon_i = 0.3\epsilon_{i-p} + z_i, p = 0, 1, 2, 3, 4$.

negative values mean that the unconstrained penalized spline estimator did better than the proposed estimator.

Comparing performance of constrained penalized spline estimator and unconstrained penalized spline estimator, the values in the first, third and fifth columns in both Tables 2 and 3 are all positive, except for the linear trend with independent error for measure 2. This results show that the constrained penalized spline estimator behaves better than the unconstrained penalized spline estimator in the estimation of both the trend and correlation when the observations are correlated. 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 linear data. The only negative values, but not significantly different from zero, in those three columns in both tables is for the linear data with independent observations. In this situation, both the spline estimators follow the data and are not forced to be linear, but the unconstrained estimator with a greater degree of freedom can approach the true linear trend a little bit better than the constrained estimator. 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. In all those measures, 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 in all those measures. That is 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 the linear trend with correlated errors, classical linear regression did better than both the spline estimators as we expected. For cubic data and sigmoid data, improvement of proposed method is great for small amount of correlation. But for the estimation of $\boldsymbol{\theta}$, the proposed method performs about as 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. That is why the improvement has a decreasing tendency with the increase of the correlation in the first, third and fifth columns in both Tables 2 and 3. Especially for measure 3, the estimation of trend for linear regression is stable to the increase of the correlation, while the constrained spline estimator behaves worse with the increase of correlation, so that the improvement is decreasing.

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 behavior 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 penalized spline regression and compare the performance with the unconstrained penalized 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 penalized spline estimator. This looks more reasonable than the results of the unconstrained penalized 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. The data choose the smallest value of all the candidates of penalty parameters for unconstrained regression, which easily to lead a wiggly curve of unconstrained regression. The linear regression obtain a greater correlation as we expected, where $\hat{p} = 4$ 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 to distinguish the realizations from the ARMA model with high correlation and those from the linear model. If the number of knots is increased to 35, the constrained penalized 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 behavior of the unconstrained penalized 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 the Figure 2, the unconstrained fit becomes more wiggly, but the constrained fit has little change.

In conclusion, with the constraint, the penalized spline regression are more robust to the choice of number of knots on both the estimation of the trend and the correlation. We have proved that the constrained penalized spline estimator attains the convergence rate of the unconstrained penalized spline estimator. Meyer (2012) also state that the constrained penalized spline fits are robust to penalty parameter choice for the independent observations, and we see here that this robustness carries over to cases with correlated errors.

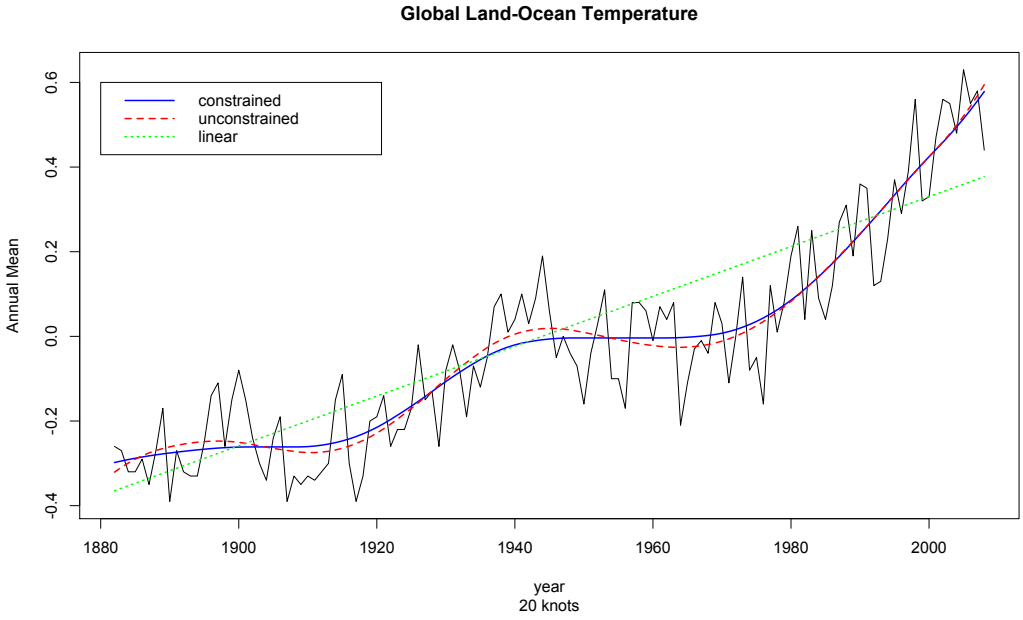

Figure 1: Estimated global temperature trends, using constrained penalized spline estimators and unconstrained penalized spline estimator both with 20 knots; also comparing them with the classical linear regression.

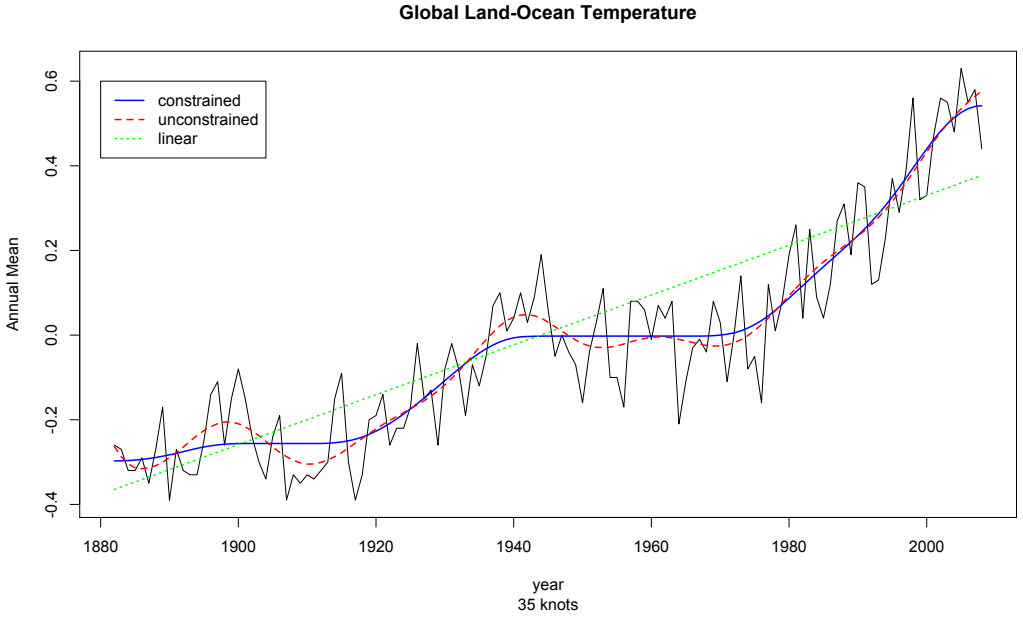

Figure 2: Estimated global temperature trends, using constrained penalized spline estimators and unconstrained penalized spline estimator both with 35 knots; also comparing them with the classical linear regression.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association 85*(411), pp. 749–759.

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

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics 26*(4), pp. 607–616.

Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. *Ann. Math. Statist. 29*, 437–454.

Claeskens, G., T. Krivobokova, and J. D. Opsomer (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Diggle, P. J. and M. F. Hutchinson (1989). On spline smoothing with autocorrelated errors. *Australian Journal of Statistics 31*(1), 166–182.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), pp. 89–102.

Francisco-Fernandez, M. and J. Opsomer (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics 33*(2), 279–295.

Fuller, W. A. (2009). *Introduction to Statistical Time Series*. John Wiley & Sons, Inc.

Hall, P. and L.-S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics 29*(3), 624–647.

Hall, P. and I. Keilegom (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 J. D. Opsomer (2005). Theory for penalised spline regression. *Biometrika 92*(1), 105–118.

Hansen, J., M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade (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), pp. 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), pp. 529–542.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics 26*(1), pp. 242–272.

Kauermann, G., T. Krivobokova, and L. Fahrmeir (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., B. Park, M. Moon, and C. Kim (2009). Using bimodal kernel for inference in nonparametric regression with correlated errors. *Journal of Multivariate Analysis 100*(7), 1487–1497.

Li, Y. and D. Ruppert (2008). On the asymptotics of penalized splines. *Biometrika 95*(2), 415–436.

Mammen, E. and C. Thomas-Agnan (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), pp. 1013–1033.

Meyer, M. C. (2012). Constrained penalized splines. *Canadian Journal of Statistics 40*(1), pp. 190–206.

Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*.

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical Science 16*(2), pp. 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., F. T. Wright, and R. Dykstra (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Inc.

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

Silvapulle, M. J. and P. K. Sen (2004, November). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions* (1 ed.). Wiley-Interscience.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), pp. 1040–1053.

Tantiyaswasdikul, C. and M. B. Woodroofe (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), pp. 341–348.

Woodward, W. and H. Gray (1993). Global warming and the problem of testing for trend in time series data. *Journal of Climate 6*(5), 953–962.