

Optimal Smoothing with Correlated Data

Chong Gu and Chun Han

Purdue University

Abstract

Penalized likelihood method offers versatile smoothing techniques in a variety of stochastic settings, and the proper selection of the smoothing parameters and other tuning parameters is crucial to the practical performance of penalized likelihood estimates. In this article, we study the selection of the smoothing parameters and the correlation parameters in penalized likelihood regression with correlated Gaussian data. We propose a simple modification of Mallows' C_L to accommodate the correlation parameters, and derive a profiled version for use with unknown variance. The proposed methods are shown to be optimal in a certain sense through asymptotic analysis and numerical simulations. Real-data example is also presented and related issues discussed.

AMS 2000 subject classifications. Primary 62G08; secondary 62G05, 62G20, 62M10, 41A15.

Key words and phrases: correlated data, cross-validation, penalized likelihood, regression.

1 Introduction

Consider regression models of the form $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \dots, n$, where $\eta(x)$ is smooth on a generic domain in the sense that $J(\eta) \leq \rho$ for some roughness functional $J(\eta)$ and $\rho > 0$, and ϵ_i 's are normal errors with mean 0 and covariance $E[\epsilon\epsilon^T] = \sigma^2 W^{-1}$. The penalized likelihood method estimates η via the minimization of

$$(\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) + n\lambda J(\eta), \quad (1.1)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\eta} = (\eta(x_1), \dots, \eta(x_n))^T$ and the smoothing parameter λ controls the tradeoff between the goodness-of-fit and the smoothness of η ; the minimizer η_λ of (1.1) is the maximum likelihood estimate under constraint $J(\eta) \leq \rho$ for some $\rho > 0$, with λ being the Lagrange multiplier. When ϵ_i are independent with $W = I$, (1.1) reduces to penalized least squares estimation, which has been extensively studied in the literature; see, e.g., Wahba (1990) and Gu (2002) for comprehensive treatments of penalized least squares regression.

The practical performance of penalized likelihood estimates depends heavily on the selection of the smoothing parameter λ ; too large a λ forces a parametric model in the null space of $J(\eta)$ and too

small a λ yields interpolation. When W is known, the generalized cross-validation (GCV) of Craven and Wahba (1979) can be used to select λ , which was shown by Li (1986) to be asymptotically optimal, in a sense to be specified shortly. For W unknown but modeled to depend on a few parameters, say γ , one also needs to select the correlation parameters γ . The purpose of this article is to devise methods for the joint selection of (λ, γ) , and to establish their optimality through asymptotic analysis and empirical simulations.

A few examples of correlation models are given below. An example of the roughness functional $J(\eta)$ will be given in Section 2.

Example 1.1 (AR(1) model) Consider a stationary first order autoregressive (AR(1)) model for ϵ_i , $\epsilon_i = \gamma\epsilon_{i-1} + a_i$, where $a_i \sim N(0, \sigma^2)$ are independent and $|\gamma| < 1$. One has

$$W^{-1} = \frac{1}{1-\gamma^2} \begin{pmatrix} 1 & \gamma & \gamma^2 & \dots & \gamma^{n-1} \\ \gamma & 1 & \gamma & \dots & \gamma^{n-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \gamma^{n-1} & \gamma^{n-2} & \gamma^{n-3} & \dots & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 1 & -\gamma & 0 & \dots & 0 \\ -\gamma & 1+\gamma^2 & -\gamma & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \quad (1.2)$$

Note that W here is tridiagonal. \square

Example 1.2 (MA(1) model) Consider an invertible first order moving average (MA(1)) model for ϵ_i , $\epsilon_i = a_i - \gamma a_{i-1}$, where $a_i \sim N(0, \sigma^2)$ are independent and $|\gamma| < 1$. One has

$$W^{-1} = \begin{pmatrix} 1+\gamma^2 & -\gamma & 0 & \dots & 0 \\ -\gamma & 1+\gamma^2 & -\gamma & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1+\gamma^2 \end{pmatrix}. \quad (1.3)$$

W is not available in closed forms here, but asymptotically it should be well approximated by the expression of W^{-1} in (1.2) for the AR(1) model. \square

Example 1.3 (Longitudinal model) Consider longitudinal observations $Y_{sj} = \eta(x_{sj}) + \epsilon_{sj}$, $j = 1, \dots, m_s$, from subjects $s \in \{1, \dots, p\}$; $n = \sum_{s=1}^p m_s$. A simple model for intra-subject correlation is $\epsilon_{sj} = a_{sj} + b_s$, where $a_{sj} \sim N(0, \sigma^2)$ and $b_s \sim N(0, \sigma^2\gamma)$ are independent. In alternative notation which we will use, one may write $Y_i = \eta(x_i) + \epsilon_i$ and $\epsilon_i = a_i + b_{s_i}$, where s_i is the subject identification of the arbitrarily labeled i th observation. Ordering the observations by subject such that $s_i \leq s_j$ for $i < j$, one has

$$W^{-1} = I + \gamma \text{diag}(\mathbf{1}_{m_1} \mathbf{1}_{m_1}^T, \dots, \mathbf{1}_{m_p} \mathbf{1}_{m_p}^T), \quad W = I - \text{diag}(\alpha_{m_1} \mathbf{1}_{m_1} \mathbf{1}_{m_1}^T, \dots, \alpha_{m_p} \mathbf{1}_{m_p} \mathbf{1}_{m_p}^T), \quad (1.4)$$

where $\mathbf{1}_m$ denotes vector of 1's of length m and $\alpha_m = \gamma/(1 + m\gamma)$. \square

Smoothing parameter selection with correlated data was studied by Wang (1998), who illustrated the middling performance of various versions of cross-validation, in contrast to the more reliable performance of the generalized maximum likelihood (GML) method of Wahba (1985) derived under the Bayes model of smoothing splines; see also Opsomer, Wang, and Yang (2001). In this article, we propose a modification of Mallows' C_L to accommodate the correlation parameters in the selection process for use with a known σ^2 , and derive a profiled version for use with an unknown σ^2 ; Mallows' C_L was shown by Li (1986) to be asymptotically optimal for use in penalized least squares regression with independent data. We will establish similar asymptotic optimality of the proposed methods for use in (1.1) with correlated data, and illustrate the empirical performances of the methods in simulation studies. Real-data example will also be shown and numerous related issues will be discussed.

The simple longitudinal model of Example 1.3 is a special case of the general variance component models, $\epsilon_i = a_i + \mathbf{z}_i^T \mathbf{b}$, where $a_i \sim N(0, \sigma^2)$ and $\mathbf{b} \sim N(\mathbf{0}, \sigma^2 B)$ are independent; for the simple longitudinal model, $\mathbf{z}_i^T \mathbf{b} = \mathbf{e}_{s_i}^T \mathbf{b} = b_{s_i}$ and $B = \gamma I_p$, where \mathbf{e}_s denotes the s th unit vector and I_p the identity matrix of size $p \times p$. In general, one has $W^{-1} = I + ZBZ^T$, where $Z = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$, and $W = I - ZD^{-1}Z^T$, where $D = Z^T Z + B^{-1}$; the correlation parameters γ are hidden in B . Optimal smoothing under the variance component models was studied in Gu and Ma (2004), where the term $\mathbf{z}_i^T \mathbf{b}$ was treated as a mean component and the error term reduced to the independent a_i , and the GCV method of Craven and Wahba (1979) was used to select tuning parameters in a doubly penalized (for η and \mathbf{b}) least squares procedure, which is different from the procedure in (1.1) that we will use in this article. The two procedures cover different domains of applications, as will be noted later in the article.

The rest of the article is organized as follows. In Section 2, penalized likelihood estimation is formulated, an example presented, and preliminary analysis conducted. Section 3 derives the proposed methods and discusses related issues. The asymptotic optimality of the proposed methods is established in Section 4, followed by empirical simulations in Section 5. Section 6 presents a real-data example and Section 7 collects a few remarks.

2 Penalized Likelihood Estimation

We shall now fill in some details concerning the penalized likelihood functional in (1.1). With $\mathbf{Y} \sim N(\boldsymbol{\eta}, \sigma^2 W^{-1})$, the minus log likelihood is seen to be

$$\frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) - \frac{1}{2} \log |W| + \frac{n}{2} \log \sigma^2 + \frac{n}{2} \log 2\pi. \quad (2.1)$$

Adding an roughness penalty $J(\eta)$, dropping terms not involving η , and absorbing σ^2 into the smoothing parameter λ , one gets (1.1).

The minimization of (1.1) shall be performed in a space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square seminorm. The evaluation functional $[x]\eta = \eta(x)$ appears in the log likelihood term, and is assumed to be continuous in \mathcal{H} . A space \mathcal{H} in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot, \cdot)$, a non-negative definite function satisfying $R_x(\cdot) = R(x, \cdot) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, and $\langle R(x, \cdot), f(\cdot) \rangle = f(x)$, $\forall f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in \mathcal{H} . Typically, $\langle \cdot, \cdot \rangle = J(\cdot, \cdot) + \tilde{J}(\cdot, \cdot)$, where $J(\cdot, \cdot)$ is the semi inner product associated with $J(\cdot)$ and $\tilde{J}(\cdot, \cdot)$ is an inner product in the null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ of $J(\eta)$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where the space \mathcal{H}_J has $J(\eta)$ as its square norm and an RK R_J satisfying $J(R_J(x, \cdot), f(\cdot)) = f(x)$, $\forall f \in \mathcal{H}_J$. See, e.g., Gu (2002, §2.1).

Example 2.1 (Cubic Spline) For $x \in [0, 1]$, a choice of $J(\eta)$ is $\int_0^1 \ddot{\eta}^2 dx$, which yields the popular cubic splines. A choice of $\tilde{J}(f, g)$ is $(\int_0^1 f dx)(\int_0^1 g dx) + (\int_0^1 \dot{f} dx)(\int_0^1 \dot{g} dx)$, yielding $\mathcal{H}_J = \{\eta : \int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0, J(\eta) < \infty\}$ and the RK $R_J(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(x_1 - x_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. The null space \mathcal{N}_J has a basis $\{1, k_1(x)\}$, where $k_1(x) = x - 0.5$. See, e.g., Gu (2002, §2.3.3). \square

It is known that the minimizer of (1.1) in \mathcal{H} resides in the space $\mathcal{N}_J \oplus \text{span}\{R_J(x_i, \cdot), i = 1, \dots, n\}$. For independent data with $W = I$, Gu and Kim (2002) considered the space $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \dots, q\}$, where $\{z_j\}$ is a random subset of $\{x_i\}$, and showed that the minimizer of (1.1) in \mathcal{H}_q shares the same asymptotic convergence rates as that in \mathcal{H} , for $q \rightarrow \infty$ at rates much slower than n ; for cubic splines, a rate of $q \asymp n^{2/9}$ may suffice. Convergence rates are yet to be established for $W \neq I$, however.

Without loss of generality, one may substitute an expression $\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{j=1}^q c_j \xi_j(x)$ for $\eta(x)$ in (1.1), where $\{\phi_\nu\}$ is a basis of \mathcal{N}_J and $\xi_j(x) = R_J(z_j, x)$. This yields

$$(\mathbf{Y} - S\mathbf{d} - R\mathbf{c})^T W (\mathbf{Y} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c},$$

where S is $n \times m$ with the (i, ν) th entry $\phi_\nu(x_i)$, R is $n \times q$ with the (i, j) th entry $\xi_j(x_i) = R_J(z_j, x_i)$, Q is $q \times q$ with the (j, k) th entry $R_J(z_j, z_k)$, and \mathbf{d} and \mathbf{c} are vectors of coefficients. Writing $\hat{\mathbf{Y}} = (\eta_\lambda(x_1), \dots, \eta_\lambda(x_n))^T$, one has $W^{1/2} \hat{\mathbf{Y}} = A(\lambda, \gamma) W^{1/2} \mathbf{Y}$, where

$$A(\lambda, \gamma) = W^{1/2} (S, R) \begin{pmatrix} S^T W S & S^T W R \\ R^T W S & R^T W R + n\lambda Q \end{pmatrix}^{-1} \begin{pmatrix} S^T \\ R^T \end{pmatrix} W^{1/2} \quad (2.2)$$

is known as the smoothing matrix; note that W depends on the correlation parameters γ . The smoothing matrix $A(\lambda, \gamma)$ plays an important role in the selection procedures for the tuning parameters (λ, γ) .

3 Tuning Parameter Selection

We now devise tools for the selection of (λ, γ) , where λ may also contain smoothing parameters hidden in $J(\eta)$, such as those for tensor product splines; see, e.g., Gu (2002, Section 2.4) for tensor product splines. Denote by $\boldsymbol{\eta}_0$ the true mean of \mathbf{Y} and $\sigma^2 W_0^{-1}$ the true covariance. A good selection should yield a small Kullback-Leibler loss

$$\begin{aligned}\tilde{L}(\lambda, \gamma) &= E_0 \left[\frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\eta})^T W_\gamma (\mathbf{Y} - \boldsymbol{\eta}) - \frac{1}{2} \log |W_\gamma| - \frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\eta}_0)^T W_0 (\mathbf{Y} - \boldsymbol{\eta}_0) + \frac{1}{2} \log |W_0| \right] \\ &= \frac{1}{2\sigma^2} (\boldsymbol{\eta}_0 - \boldsymbol{\eta})^T W_\gamma (\boldsymbol{\eta}_0 - \boldsymbol{\eta}) + \frac{1}{2} \text{tr}(W_\gamma W_0^{-1} - I) - \frac{1}{2} \log |W_\gamma W_0^{-1}|,\end{aligned}\quad (3.1)$$

where W_γ spells out the dependence of W on γ . For $W = I$, (3.1) reduces to the standard mean square error, $\sigma^2 L(\lambda) = (2\sigma^2/n)\tilde{L}(\lambda) = \sum_{i=1}^n (\eta_\lambda(x_i) - \eta_0(x_i))^2/n$.

First consider the case with a known σ^2 . Given W for fixed γ , one may select λ through the minimization of Mallows' C_L ,

$$\tilde{U}(\lambda) = \frac{1}{n} \mathbf{Y}^T W^{1/2} (I - A(\lambda))^2 W^{1/2} \mathbf{Y} + 2 \frac{\sigma^2}{n} \text{tr} A(\lambda), \quad (3.2)$$

where the known γ is removed from the notation $A(\lambda)$. $\tilde{U}(\lambda)$ was shown by Li (1986) to be asymptotically optimal for $W = I$, in the sense that $\tilde{U}(\lambda) - \sigma^2 L(\lambda) - \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}/n = o_p(L(\lambda))$; the result readily extends to cases with general W that is known. It is noted that $(n/2\sigma^2)\tilde{U}(\lambda)$ consists of the minus log likelihood plus a penalty term $\text{tr} A(\lambda)$; terms in (2.1) that do not depend on λ are dropped. Adding the term $-(1/2) \log |W|$ from the minus log likelihood back into $(n/2\sigma^2)\tilde{U}(\lambda)$ and scaling proportionally, one may minimize

$$U(\lambda, \gamma) = \frac{1}{n\sigma^2} \mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma))^2 W_\gamma^{1/2} \mathbf{Y} - \frac{1}{n} \log |W_\gamma| + \frac{2}{n} \text{tr} A(\lambda, \gamma) \quad (3.3)$$

to jointly select (λ, γ) .

For σ^2 unknown, one may add the term $\log \sigma^2$ from the scaled minus log likelihood back into (3.3) and profile out σ^2 via $\hat{\sigma}^2 = \{\mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma))^2 W_\gamma^{1/2} \mathbf{Y}\}/n$, yielding

$$V(\lambda, \gamma) = \log \left\{ n^{-1} \mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma))^2 W_\gamma^{1/2} \mathbf{Y} \right\} - \frac{1}{n} \log |W_\gamma| + \frac{2}{n} \text{tr} A(\lambda, \gamma). \quad (3.4)$$

For $W = I$ and $\mu = \text{tr} A(\lambda)/n = o(1)$, (3.4) reduces to

$$\log \left\{ (n^{-1} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y}) e^{2\mu} \right\} = \log \left\{ \frac{n^{-1} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y}}{(1 - \mu)^2} (1 + O(\mu^2)) \right\} = \log \left\{ \tilde{V}_1(\lambda) (1 + O(\mu^2)) \right\},$$

where

$$\tilde{V}_\alpha(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))^2\mathbf{Y}}{\{n^{-1}\text{tr}(I - \alpha A(\lambda))\}^2} \quad (3.5)$$

for $\alpha = 1$ is the GCV score of Craven and Wahba (1979); it was also shown by Li (1986) that $\tilde{V}_1(\lambda) - \sigma^2 L(\lambda) - \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} / n = o_p(L(\lambda))$. An obvious drawback of (3.4) is that the third term is bounded from above since $I - A \geq 0$, while the first term will go to $-\infty$ as A approaches I , favoring interpolation. To guard against this, we shall use a modified version

$$V_*(\lambda, \gamma) = \log \left\{ \mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma))^2 W_\gamma^{1/2} \mathbf{Y} / n \right\} - \frac{1}{n} \log |W_\gamma| + \frac{2 \text{tr} A(\lambda, \gamma)}{n - \text{tr} A(\lambda, \gamma)}; \quad (3.6)$$

when $\mu = \text{tr} A(\lambda, \gamma) / n = o(1)$, $V_*(\lambda, \gamma) - V(\lambda, \gamma) = 2\mu^2(1 + o(1))$.

4 Asymptotic Optimality

We now establish the asymptotic optimality of $U(\lambda, \gamma)$, $V(\lambda, \gamma)$, and $V_*(\lambda, \gamma)$. We first present and prove the general theory under regularity conditions, then verify the conditions for specific correlation models. To keep things simple, the dependence of quantities on (λ, γ) is often dropped from the notation where no confusion may result.

4.1 General Theory

Rescaling \tilde{L} in (3.1), we define a loss function

$$L = \frac{1}{n\sigma^2} (\boldsymbol{\eta}_0 - \boldsymbol{\eta})^T W (\boldsymbol{\eta}_0 - \boldsymbol{\eta}) + \frac{1}{n} \text{tr}(W W_0^{-1} - I) - \frac{1}{n} \log |W W_0^{-1}|, \quad (4.1)$$

where $\boldsymbol{\eta} = W^{-1/2} A W^{1/2} \mathbf{Y} = \tilde{A} \mathbf{Y} = \tilde{A}(\boldsymbol{\eta}_0 + \boldsymbol{\epsilon})$ with $\tilde{A} = W^{-1/2} A W^{1/2}$. Taking expectations in (4.1), one has the risk function

$$R = \frac{1}{n\sigma^2} \boldsymbol{\eta}_0^T (I - \tilde{A})^T W (I - \tilde{A}) \boldsymbol{\eta}_0 + \frac{1}{n\sigma^2} \text{tr}(\tilde{A}^T W \tilde{A} W_0^{-1}) + \frac{1}{n} \text{tr}(W W_0^{-1} - I) - \frac{1}{n} \log |W W_0^{-1}|. \quad (4.2)$$

Noting that

$$\begin{aligned} U &= \frac{1}{n\sigma^2} (\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) - \frac{1}{n} \log |W| + \frac{2}{n} \text{tr} A \\ &= \frac{1}{n\sigma^2} (\boldsymbol{\eta}_0 - \boldsymbol{\eta})^T W (\boldsymbol{\eta}_0 - \boldsymbol{\eta}) + \frac{2}{n\sigma^2} (\boldsymbol{\eta}_0 - \boldsymbol{\eta})^T W \boldsymbol{\epsilon} + \frac{1}{n\sigma^2} \boldsymbol{\epsilon}^T W \boldsymbol{\epsilon} - \frac{1}{n} \log |W| + \frac{2}{n} \text{tr} A, \end{aligned} \quad (4.3)$$

some algebra yields

$$\begin{aligned}
U - L - \frac{1}{n\sigma^2}\boldsymbol{\epsilon}^T W_0 \boldsymbol{\epsilon} + \frac{1}{n} \log |W_0| &= \frac{2}{n\sigma^2} \boldsymbol{\eta}_0^T (I - \tilde{A})^T W \boldsymbol{\epsilon} \\
&\quad - \frac{2}{n} \left\{ \frac{1}{\sigma^2} \boldsymbol{\epsilon}^T \tilde{A}^T W \boldsymbol{\epsilon} - \text{tr} A \right\} \\
&\quad + \frac{1}{n} \left\{ \frac{1}{\sigma^2} \boldsymbol{\epsilon}^T (W - W_0) \boldsymbol{\epsilon} - \text{tr}(W W_0^{-1} - I) \right\}. \tag{4.4}
\end{aligned}$$

We shall establish the asymptotic optimality of $U(\lambda, \gamma)$ under the following conditions. The conditions will be verified in a few settings in later sections.

Condition C.1 For W_γ in a shrinking neighborhood of W_0 , as $\lambda \rightarrow 0$ and $n\lambda \rightarrow \infty$, $R(\lambda, \gamma) \rightarrow 0$ and $nR(\lambda, \gamma) \rightarrow \infty$.

The meaning of shrinking neighborhood will be made clear later. Condition C.1 assures that the estimates are consistent, but concedes that the typical parametric convergence rates of $O(n^{-1})$ are not achievable.

Condition C.2 $\pm(W_\gamma^{1/2} W_0^{-1} W_\gamma^{1/2} - I) \leq \rho_n I$ for some positive $\rho_n = O(R^{1/2}(\lambda, \gamma))$.

Condition C.2 requires W_γ to converge to W_0 at a certain rate so that the largest absolute eigenvalue of $W_\gamma W_0^{-1} - I$ is of the order $O(R^{1/2}(\lambda, \gamma))$.

Condition C.3 $\{n^{-1} \text{tr} A(\lambda, \gamma)\}^2 / \{n^{-1} \text{tr} A^2(\lambda, \gamma)\} \rightarrow 0$.

Condition C.3 holds in settings where $\text{tr} A \asymp \text{tr} A^2 = o(n)$. Note that under C.2,

$$I \leq (1 - \rho_n)^{-1} W^{1/2} W_0^{-1} W^{1/2},$$

which, combined with C.1, yields

$$n^{-1} \text{tr} A^2 \leq n^{-1} (1 - \rho_n)^{-1} \text{tr}(A^2 W^{1/2} W_0^{-1} W^{1/2}) = n^{-1} \text{tr}(\tilde{A}^T W \tilde{A} W_0^{-1})(1 + o(1)) = O(R).$$

Condition C.3 also implies that $\mu = n^{-1} \text{tr} A \rightarrow 0$, as $n^{-1} \text{tr} A^2 \leq n^{-1} \text{tr} A \leq 1$.

Theorem 4.1 *Under Conditions C.1–C.3, as $\lambda \rightarrow 0$ and $n\lambda \rightarrow \infty$, one has*

$$U(\lambda, \gamma) - L(\lambda, \gamma) - \frac{1}{n\sigma^2} \boldsymbol{\epsilon}^T W_0 \boldsymbol{\epsilon} + \frac{1}{n} \log |W_0| = o_p(L(\lambda, \gamma)).$$

Proof. It suffices to show that $L - R$ and the three terms on the right-hand side of (4.4) are all of the order $o_p(R)$. It is easy to see that

$$L - R = -\frac{2}{n\sigma^2} \boldsymbol{\eta}_0^T (I - \tilde{A})^T W \tilde{A} \boldsymbol{\epsilon} + \frac{1}{n} \left\{ \frac{1}{\sigma^2} \boldsymbol{\epsilon}^T \tilde{A}^T W \tilde{A} \boldsymbol{\epsilon} - \text{tr}(\tilde{A}^T W \tilde{A} W_0^{-1}) \right\}; \tag{4.5}$$

these terms are similar to the first two terms of (4.4) and shall be bounded using similar techniques. For the first term of (4.4), one has

$$\begin{aligned}\text{Var}\left[\frac{1}{n\sigma^2}\boldsymbol{\eta}_0^T(I-\tilde{A})^TW\boldsymbol{\epsilon}\right] &= \frac{1}{n^2\sigma^2}\boldsymbol{\eta}_0^T(I-\tilde{A})^TWW_0^{-1}W(I-\tilde{A})\boldsymbol{\eta}_0 \\ &\leq \frac{1+\rho_n}{n^2\sigma^2}\{\boldsymbol{\eta}_0^T(I-\tilde{A})^TW(I-\tilde{A})\boldsymbol{\eta}_0\} = O(R^2/nR) = o(R^2),\end{aligned}$$

where $W^{1/2}W_0^{-1}W^{1/2} \leq (1+\rho_n)I$ is used; the first term of (4.5) is similarly bounded,

$$\begin{aligned}\text{Var}\left[\frac{1}{n\sigma^2}\boldsymbol{\eta}_0^T(I-\tilde{A})^TW\tilde{A}\boldsymbol{\epsilon}\right] &= \frac{1}{n^2\sigma^2}\boldsymbol{\eta}_0^T(I-\tilde{A})^TW^{1/2}AW^{1/2}W_0^{-1}W^{1/2}AW^{1/2}(I-\tilde{A})\boldsymbol{\eta}_0 \\ &\leq \frac{1+\rho_n}{n^2\sigma^2}\{\boldsymbol{\eta}_0^T(I-\tilde{A})^TW^{1/2}A^2W^{1/2}(I-\tilde{A})\boldsymbol{\eta}_0\} = o(R^2),\end{aligned}\quad (4.6)$$

as $A^2 \leq I$. For the second term of (4.4), noting that $E[\boldsymbol{\epsilon}^TC\boldsymbol{\epsilon}] = \sigma^2\text{tr}(CW_0^{-1})$ and $\text{Var}[\boldsymbol{\epsilon}^TC\boldsymbol{\epsilon}] = 2\sigma^4\text{tr}(CW_0^{-1}C^TW_0^{-1})$ for any deterministic matrix C , one has

$$\begin{aligned}E\left[\frac{1}{n\sigma^2}\boldsymbol{\epsilon}^T\tilde{A}^TW\boldsymbol{\epsilon} - \frac{1}{n}\text{tr}A\right]^2 &= \frac{2}{n^2}\text{tr}(\tilde{A}^TWW_0^{-1}W\tilde{A}W_0^{-1}) + \frac{1}{n^2}(\text{tr}(\tilde{A}^TWW_0^{-1}) - \text{tr}A)^2 \\ &\leq \frac{2(1+\rho_n)}{n^2}\text{tr}(\tilde{A}^TW\tilde{A}W_0^{-1}) + \frac{\rho_n^2}{n^2}(\text{tr}A)^2 = O(R/n) + (\text{tr}A/n)^2O(R) = o(R^2),\end{aligned}$$

where $|\text{tr}(A(W^{1/2}W_0^{-1}W^{1/2} - I))| \leq \rho_n\text{tr}A$ is used; the second term of (4.5) is similar to the variance term above but with a couple of A 's inserted as in (4.6). For the third term of (4.4), one simply has

$$\text{Var}\left[\frac{1}{n\sigma^2}\boldsymbol{\epsilon}^T(W - W_0)\boldsymbol{\epsilon}\right] = \frac{1}{n^2}\text{tr}(WW_0^{-1} - I)^2 \leq \frac{\rho_n^2}{n} = o(R^2).$$

This completes the proof. \square

To establish similar results for $V(\lambda, \gamma)$ and $V_*(\lambda, \gamma)$, one needs an additional condition.

Condition C.4 $n^{-1}\text{tr}(W_\gamma W_0^{-1} - I) = o(R^{1/2}(\lambda, \gamma))$.

Note that Condition C.2 only guarantees that $n^{-1}\text{tr}(W_\gamma W_0^{-1} - I) = O(R^{1/2}(\lambda, \gamma))$.

Theorem 4.2 *Under Conditions C.1–C.4, as $\lambda \rightarrow 0$ and $n\lambda \rightarrow \infty$, one has*

$$V(\lambda, \gamma) - L(\lambda, \gamma) - K = o_p(L(\lambda, \gamma)),$$

$$V_*(\lambda, \gamma) - L(\lambda, \gamma) - K = o_p(L(\lambda, \gamma)),$$

where $K = (n\sigma^2)^{-1}\boldsymbol{\epsilon}^TW_0\boldsymbol{\epsilon} - n^{-1}\log|W_0| + \log\sigma^2 - 1$ does not depend on (λ, γ) .

Proof: One only needs to prove the first equation; the second equation follows the fact that $V_* - V = 2\mu^2(1 + o(1))$, where $\mu = n^{-1}\text{tr}A = o(R^{1/2})$ by Condition C.3.

Define as a function of σ^2 ,

$$U_*(\sigma^2) = U + \log \sigma^2 = \frac{1}{n\sigma^2}(\mathbf{Y} - \boldsymbol{\eta})^T W(\mathbf{Y} - \boldsymbol{\eta}) - \frac{1}{n} \log |W| + \frac{2}{n} \text{tr} A + \log \sigma^2.$$

One has $V = U_*(\hat{\sigma}^2) - 1$, where $\hat{\sigma}^2 = (\mathbf{Y} - \boldsymbol{\eta})^T W(\mathbf{Y} - \boldsymbol{\eta})/n$, and $(d^2 U_*/d(\sigma^2)^2|_{\hat{\sigma}^2}) = 1/\hat{\sigma}^4$. It follows that

$$U + \log \sigma^2 - V - 1 = U_*(\sigma^2) - U_*(\hat{\sigma}^2) = (1/2\hat{\sigma}^4)(\sigma^2 - \hat{\sigma}^2)^2(1 + o(1)). \quad (4.7)$$

It will be shown below that $(\sigma^2 - \hat{\sigma}^2)^2 = o_p(R)$, which, combined with (4.7) and Theorem 4.1, yields the theorem.

We now show that $(\hat{\sigma}^2 - \sigma^2)^2 = o_p(R)$. Simple algebra leads to

$$\begin{aligned} \hat{\sigma}^2 - \sigma^2 &= \frac{1}{n} \boldsymbol{\eta}_0^T (I - \tilde{A})^T W (I - \tilde{A}) \boldsymbol{\eta}_0 + \frac{2}{n} \boldsymbol{\eta}_0^T (I - \tilde{A})^T W (I - \tilde{A}) \boldsymbol{\epsilon} \\ &\quad + \frac{1}{n} \boldsymbol{\epsilon}^T \tilde{A}^T W (\tilde{A} - 2I) \boldsymbol{\epsilon} + \left\{ \frac{1}{n} \boldsymbol{\epsilon}^T W \boldsymbol{\epsilon} - \sigma^2 \right\} \\ &= O(R) + o_p(R) + \{O_p(\mu) + o_p(R)\} + o_p(R^{1/2}) = o_p(R^{1/2}), \end{aligned} \quad (4.8)$$

where the orders of terms are mainly by the arguments in the proof of Theorem 4.1, except that Condition C.4 is used through

$$\begin{aligned} E \left[\frac{1}{n} \boldsymbol{\epsilon}^T W \boldsymbol{\epsilon} - \sigma^2 \right]^2 &= \frac{2\sigma^4}{n^2} \text{tr}(W W_0^{-1} W W_0^{-1}) + \frac{\sigma^4}{n^2} \{ \text{tr}(W W_0^{-1} - I) \}^2 \\ &\leq \frac{2\sigma^4(1 + \rho_n)^2}{n} + o(R) = o(R). \end{aligned}$$

This completes the proof. \square

4.2 AR(1) Model

We now verify Conditions C.1–C.4 for the AR(1) model of Example 1.1, with the covariance given in (1.2). It is assumed that $\gamma_0^2 \neq 1$ and that γ be bounded away from ± 1 .

From (1.2), it is easy to verify that $(1 - |\gamma|)^2 I \leq W_\gamma \leq (1 + |\gamma|)^2 I$, so the eigenvalues of W_γ are bounded from below and above. The following lemmas play central roles in the verification of the conditions.

Lemma 4.1 *If $l_\gamma I \leq W_\gamma^{-1} \leq u_\gamma I$, where l_γ and u_γ are positive constants, then*

$$\text{tr} \check{A}(u_\gamma \lambda) \leq \text{tr} A(\lambda, \gamma) \leq \text{tr} \check{A}(l_\gamma \lambda), \quad (4.9)$$

$$\text{tr} \check{A}^2(u_\gamma \lambda) \leq \text{tr} A^2(\lambda, \gamma) \leq \text{tr} \check{A}^2(l_\gamma \lambda), \quad (4.10)$$

where $\check{A}(\lambda)$ is the smoothing matrix of (2.2) with $W = I$.

Lemma 4.2 *If $l_\gamma I \leq W_\gamma^{-1} \leq u_\gamma I$, where l_γ and u_γ are positive constants, then*

$$C_\gamma^{-1}(I - \check{A}(u_\gamma \lambda))^2 \leq (I - \check{A}(\lambda, \gamma))^T W_\gamma (I - \check{A}(\lambda, \gamma)) \leq C_\gamma (I - \check{A}(u_\gamma \lambda))^2, \quad (4.11)$$

where $C_\gamma = \exp\{u_\gamma((l_\gamma^{-1} - u_\gamma^{-1})^2 + (l_\gamma^{-1} - u_\gamma^{-1}) + 1)\}$.

The proofs of the lemmas are given at the end of this section. The lemmas allow one to relate quantities with those in the independent data case, which are well studied in the literature. In particular, we summarize the most relevant results in the following assumption.

Assumption A For some $r > 1$, as $\lambda \rightarrow 0$ and $n\lambda^{1/r} \rightarrow \infty$, $n^{-1}\boldsymbol{\eta}_0^T(I - \check{A}(\lambda))^2\boldsymbol{\eta}_0 = O(\lambda)$ and $\text{tr}\check{A}(\lambda) \asymp \text{tr}\check{A}^2(\lambda) \asymp \lambda^{-1/r}$.

Assumption A is assumed through the rest of the article; it holds in most settings of interest; see, e.g., Craven and Wahba (1979), Wahba (1985), Li (1986), and Gu (2002, §4.2.3). For Example 2.1, $r = 4$.

We first verify Condition C.1. Recall $R(\lambda, \gamma)$ from (4.2); only the first two terms are present for the independent data case with $W = I$. By Lemma 4.1, $\text{tr}A^2(\lambda, \gamma) \geq \text{tr}\check{A}^2(u_\gamma \lambda) \asymp \lambda^{-1/r} \rightarrow \infty$ as $\lambda \rightarrow 0$, so $nR(\lambda, \gamma) \rightarrow \infty$. By Lemmas 4.1 and 4.2, the first two terms of $R(\lambda, \gamma)$ are of the order $O(\lambda + n^{-1}\lambda^{-1/r})$, and from (1.2), it is easy to calculate

$$\frac{1}{n}\text{tr}(WW_0^{-1} - I) = \frac{(\gamma - \gamma_0)^2}{1 - \gamma_0^2} - \frac{2}{n} \frac{\gamma(\gamma - \gamma_0)}{1 - \gamma_0^2} \quad (4.12)$$

and $\log|WW_0^{-1}| = \log\{(1 - \gamma^2)/(1 - \gamma_0^2)\}$, so $R(\lambda, \gamma) \rightarrow 0$ as $\lambda \rightarrow 0$, $n\lambda^{1/r} \rightarrow \infty$, and $\gamma \rightarrow \gamma_0$. Theorems 4.1 and 4.2 are only relevant when the γ minimizing $U(\lambda, \gamma)$, $V(\lambda, \gamma)$, or $V_*(\lambda, \gamma)$ converges to γ_0 , which we shall verify later in the section.

Since $(1 + |\gamma|)^{-2}I \leq W_\gamma^{-1} \leq (1 - |\gamma|)^{-2}I$ and $2(1 + |\gamma|)I - (dW/d\gamma) \geq 0$, one has

$$\begin{aligned} |\mathbf{x}^T(W_0^{-1/2}W_\gamma W_0^{-1/2} - I)\mathbf{x}| &= |\gamma - \gamma_0| |\mathbf{x}^T W_0^{-1/2} (dW/d\gamma|_{\gamma^*}) W_0^{-1/2} \mathbf{x}| \\ &\leq |\gamma - \gamma_0| 2(1 + |\gamma^*|) |\mathbf{x}^T W_0^{-1} \mathbf{x}| \leq |\gamma - \gamma_0| 2(1 + |\gamma^*|) (1 - |\gamma_0|)^{-2} = O(|\gamma - \gamma_0|), \end{aligned}$$

where $\mathbf{x}^T \mathbf{x} = 1$ but otherwise arbitrary, γ^* is between γ and γ_0 , and the equality is by the mean value theorem. It is clear that the largest absolute eigenvalue of $(WW_0^{-1} - I)$ is of the order $O(|\gamma - \gamma_0|) = O(R^{1/2})$, so Condition C.2 holds; see (4.12).

By Lemma 4.1, $\text{tr}A \asymp \text{tr}A^2 \asymp \lambda^{-1/r}$, so Condition C.3 holds.

By the definition of $R(\lambda, \gamma)$ and Condition C.1, $n^{-1}\text{tr}(WW_0^{-1} - I) = O(R) = o(R^{1/2})$, so Condition C.4 holds.

We shall now establish the consistency of the γ estimates through the minimization of $U(\lambda, \gamma)$, $V(\lambda, \gamma)$, or $V_*(\lambda, \gamma)$, without assuming Conditions C.1–C.4. Simple algebra yields

$$\begin{aligned} \frac{\partial}{\partial \gamma} U(\lambda, \gamma) &= \frac{1}{n\sigma^2} (\boldsymbol{\eta}_0 + \boldsymbol{\epsilon})^T (I - \tilde{A})^T (-\dot{W}\tilde{A} - \tilde{A}^T \dot{W} + \dot{W})(I - \tilde{A})(\boldsymbol{\eta}_0 + \boldsymbol{\epsilon}) \\ &\quad - \frac{1}{n} \frac{d}{d\gamma} \log |W| + \frac{1}{n} \text{tr}((I - A)AW^{-1/2}\dot{W}W^{-1/2}), \end{aligned} \quad (4.13)$$

where $\pm \dot{W} = \pm dW/d\gamma \leq 2(1 + |\gamma|)I$. For the first term in (4.13), collect

$$\begin{aligned} Q(\lambda, \gamma) &= \frac{1}{n} (\boldsymbol{\eta}_0 + \boldsymbol{\epsilon})^T (I - \tilde{A})^T (-\dot{W}\tilde{A} - \tilde{A}^T \dot{W} + \dot{W})(I - \tilde{A})(\boldsymbol{\eta}_0 + \boldsymbol{\epsilon}) - \frac{1}{n} \boldsymbol{\epsilon}^T \dot{W} \boldsymbol{\epsilon} \\ &= \frac{1}{n} \boldsymbol{\eta}_0^T (I - \tilde{A})^T (-2\dot{W}\tilde{A} + \dot{W})(I - \tilde{A})\boldsymbol{\eta}_0 + \frac{2}{n} \boldsymbol{\eta}_0^T (I - \tilde{A})^T (-\dot{W}\tilde{A} - \tilde{A}^T \dot{W} + \dot{W})(I - \tilde{A})\boldsymbol{\epsilon} \\ &\quad + \frac{1}{n} \boldsymbol{\epsilon}^T (3\tilde{A}^T \dot{W} \tilde{A} + 2\dot{W}\tilde{A}\tilde{A} - 4\dot{W}\tilde{A} - 2\tilde{A}^T \dot{W} \tilde{A}\tilde{A})\boldsymbol{\epsilon}. \end{aligned} \quad (4.14)$$

It can be shown that $Q(\lambda, \gamma) = O_p(\lambda + n^{-1}\lambda^{-1/r})$ for $\lambda \rightarrow 0$, $n\lambda^{1/r} \rightarrow \infty$, and all γ , by the boundedness of the eigenvalues of A , W , W^{-1} , and $\pm \dot{W}$, the Cauchy-Schwarz inequality, and Lemmas 4.1 and 4.2; see the appendix. By Lemma 4.1 and the boundedness of the eigenvalues of $W^{-1/2}\dot{W}W^{-1/2}$, the third term in (4.13) is of the order $O(n^{-1}\lambda^{-1/r})$. Summing up, one has for $\lambda \rightarrow 0$, $n\lambda^{1/r} \rightarrow \infty$, and all γ ,

$$\begin{aligned} \frac{\partial}{\partial \gamma} U(\lambda, \gamma) &= \frac{1}{n} \left(\frac{1}{\sigma^2} \boldsymbol{\epsilon}^T \dot{W} \boldsymbol{\epsilon} - \frac{d}{d\gamma} \log |W| \right) + O_p(\lambda + n^{-1}\lambda^{-1/r}) \\ &= \frac{d}{d\gamma} \left(\frac{1}{n} \text{tr}(WW_0^{-1}) - \frac{1}{n} \log |W| \right) + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}) \\ &= \frac{2}{1 - \gamma_0^2} (\gamma - \gamma_0) + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}), \end{aligned} \quad (4.15)$$

where $(n\sigma^2)^{-1} \boldsymbol{\epsilon}^T \dot{W} \boldsymbol{\epsilon} = n^{-1} \text{tr}(\dot{W}W_0^{-1}) + O_p(n^{-1/2})$ as $\text{Var}[\boldsymbol{\epsilon}^T \dot{W} \boldsymbol{\epsilon}] = 2\sigma^4 \text{tr}(\dot{W}W_0^{-1}\dot{W}W_0^{-1}) = O(n)$. Hence, $\hat{\gamma}_u - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$ for the minimizer $\hat{\gamma}_u$ of $U(\lambda, \gamma)$ as $\lambda \rightarrow 0$ and $n\lambda^{1/r} \rightarrow \infty$. Similarly, since $n^{-1} d \log |W| / d\gamma = O(n^{-1})$,

$$\frac{\partial}{\partial \gamma} V(\lambda, \gamma) = \frac{1}{\hat{\sigma}^2} \frac{1}{n} \boldsymbol{\epsilon}^T \dot{W} \boldsymbol{\epsilon} + O_p(\lambda + n^{-1}\lambda^{-1/r}) = \frac{2\sigma^2/\hat{\sigma}^2}{1 - \gamma_0^2} (\gamma - \gamma_0) + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}), \quad (4.16)$$

so $\hat{\gamma}_v - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$ for the minimizer $\hat{\gamma}_v$ of $V(\lambda, \gamma)$. Parallel result for the minimizer of $V_*(\lambda, \gamma)$ follows trivial modification of the formulas.

The rest of the section collects the proofs of Lemmas 4.1 and 4.2.

Proof of Lemma 4.1: Write $K(\alpha) = \alpha a I + (1 - \alpha)W$ for $\alpha \in (0, 1)$ and some positive constant a . Define $H(\alpha)$ by replacing W in (2.2) with $K(\alpha)$ and $\tilde{H} = K^{-1/2} H K^{1/2}$; $\tilde{H}(0) = \tilde{A}(\lambda, \gamma)$ and $\tilde{H}(1) = \check{A}(\lambda/a)$. Take $f(\alpha) = \text{tr} H(\alpha)$ and differentiate with respect to α , noting that $dM^{-1}/d\gamma =$

$-M^{-1}(dM/d\gamma)M^{-1}$ for nonsingular matrix M , some algebra yields

$$f'(\alpha) = \text{tr}(K^{-1/2}(H - H^2)K^{-1/2}(aI - W))$$

Setting $a = u_\gamma^{-1}$, $f'(\alpha) \leq 0$, so $f(1) \leq f(0)$, yielding the first inequality in (4.9); setting $a = l_\gamma^{-1}$, $f'(\alpha) \geq 0$, so $f(1) \geq f(0)$, yielding the second inequality in (4.9). Parallel calculations with $g(\alpha) = \text{tr}H^2(\alpha)$ yield (4.10). \square

Proof of Lemma 4.2: Write $K(\alpha) = \alpha u_\gamma^{-1}I + (1 - \alpha)W$ for $\alpha \in (0, 1)$; $u_\gamma^{-1}I \leq K(\alpha) \leq l_\gamma^{-1}I$. Define $H(\alpha)$ by replacing W in (2.2) with $K(\alpha)$ and $\tilde{H} = K^{-1/2}HK^{1/2}$; $\tilde{H}(0) = \tilde{A}(\lambda, \gamma)$ and $\tilde{H}(1) = \tilde{A}(u_\gamma\lambda)$.

For an arbitrary vector $\mathbf{x} \neq \mathbf{0}$, define $f(\alpha) = \mathbf{x}^T(I - \tilde{H})^TK(I - \tilde{H})\mathbf{x}$. Straightforward algebra yields

$$f'(\alpha) = \mathbf{x}^T(I - \tilde{H})^T(-\dot{K}\tilde{H} - \tilde{H}^T\dot{K} + \dot{K})(I - \tilde{H})\mathbf{x},$$

where $\dot{K} = dK/d\alpha = u_\gamma^{-1}I - W$ is nonpositive definite satisfying $-\dot{K} \leq (l_\gamma^{-1} - u_\gamma^{-1})I$. Noting that $I \leq u_\gamma K$ and $\tilde{H}^T\tilde{H} \leq u_\gamma\tilde{H}^TK\tilde{H} \leq u_\gamma K$, one has

$$\begin{aligned} |f'(\alpha)| &\leq \mathbf{x}^T(I - \tilde{H})^T(\dot{K}^2 + \tilde{H}^T\tilde{H} - \dot{K})(I - \tilde{H})\mathbf{x} \\ &\leq u_\gamma((l_\gamma^{-1} - u_\gamma^{-1})^2 + (l_\gamma^{-1} - u_\gamma^{-1}) + 1)\mathbf{x}^T(I - \tilde{H})^TK(I - \tilde{H})\mathbf{x} = f(\alpha) \log C_\gamma, \end{aligned}$$

or $|f'(\alpha)/f(\alpha)| \leq \log C_\gamma$. It follows that $|\log f(0) - \log f(1)| \leq \log C_\gamma$, which yields (4.11). \square

4.3 Block AR(1) Models

Suppose $\boldsymbol{\epsilon}^T = (\boldsymbol{\epsilon}_1^T, \dots, \boldsymbol{\epsilon}_p^T)$ is decomposed into p independent blocks of sizes m_j , $\sum_{j=1}^p m_j = n$, and AR(1) models hold within each block with parameters γ_j but a common σ^2 . Denote $\bar{\gamma} = \max_j |\gamma_j|$, which is assumed bounded away from 1. Also assume a fixed p and $m_j/n \rightarrow r_j$, where r_j are fixed proportions satisfying $\sum_{j=1}^p r_j = 1$; other configurations are possible but the analysis could be more delicate.

It is easy to see that $(1 - \bar{\gamma})^2I \leq W \leq (1 + \bar{\gamma})^2I$, so Lemmas 4.1 and 4.2 hold. Also,

$$\frac{1}{n}\text{tr}(WW_0^{-1} - I) = \sum_{j=1}^p \frac{m_j}{n} \left\{ \frac{(\gamma_j - \gamma_{j,0})^2}{1 - \gamma_{j,0}^2} - \frac{2}{m_j} \frac{\gamma_j(\gamma_j - \gamma_{j,0})}{1 - \gamma_{j,0}^2} \right\},$$

and $\log |WW_0^{-1}| = \sum_{j=1}^p \log\{(1 - \gamma_j^2)/(1 - \gamma_{j,0}^2)\}$, where $\gamma_{j,0}$ are the true correlation parameters. Using the same arguments as for the AR(1) model, Condition C.1 holds for $\gamma_j \rightarrow \gamma_{j,0}$, so do Conditions C.2–C.4. The consistency of $\hat{\gamma}_u$ and $\hat{\gamma}_v$, now vectors of length p , also follows readily.

If one assumes a common correlation parameter $\gamma = \gamma_1 = \dots = \gamma_p$, then things also hold for $p \rightarrow \infty$ and $p/n \rightarrow 0$.

4.4 MA(1) Models

Let us now consider the MA(1) model of Example 1.2, with the covariance given in (1.3). It is assumed that γ is bounded away from ± 1 .

From (1.3), $(1 - |\gamma|)^2 I \leq W^{-1} \leq (1 + |\gamma|)^2 I$, so Lemmas 4.1 and 4.2 hold. Write

$$W^{-1} = \begin{pmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ -\gamma & 1 + \gamma^2 & -\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} + \begin{pmatrix} \gamma & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \gamma \end{pmatrix} \begin{pmatrix} \gamma & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \gamma \end{pmatrix} = D + BB^T.$$

One has $W = D^{-1} - D^{-1}B(I + B^T D^{-1}B)^{-1}B^T D^{-1}$, where

$$B^T D^{-1} = \frac{\gamma}{1 - \gamma^2} \begin{pmatrix} 1 & \gamma & \cdots & \gamma^{n-1} \\ \gamma^{n-1} & \gamma^{n-2} & \cdots & 1 \end{pmatrix}, \quad (I + B^T D^{-1}B)^{-1} = \frac{1 - \gamma^2}{1 - \gamma^{2n+2}} \begin{pmatrix} 1 & -\gamma^{n+1} \\ -\gamma^{n+1} & 1 \end{pmatrix},$$

and $|W^{-1}| = |D| |I + B^T D^{-1}B| = (1 - \gamma^{2n+2})/(1 - \gamma^2)$. Some algebra yields

$$\frac{1}{n} \text{tr}(WW_0^{-1} - I) = \frac{(\gamma - \gamma_0)^2}{1 - \gamma^2} + \frac{2}{n} \frac{\gamma\gamma_0}{1 - \gamma^2} - \frac{2}{n} \frac{\gamma^2(1 + \gamma_0^2)}{(1 - \gamma^2)^2} + \frac{4}{n} \frac{\gamma^3\gamma_0}{(1 - \gamma^2)^2} + o\left(\frac{1}{n}\right) = \frac{(\gamma - \gamma_0)^2}{1 - \gamma^2} + O\left(\frac{1}{n}\right).$$

The same arguments used for AR(1) lead to Conditions C.1, C.3, and C.4 for $\gamma \rightarrow \gamma_0$. For Condition C.2, one may use the technique in §4.2 to bound the eigenvalues δ_i of $(W_0^{1/2}W^{-1}W_0^{1/2} - I)$; the eigenvalues of $(W_0^{-1/2}WW_0^{-1/2} - I)$ are simply $-\delta_i/(1 + \delta_i)$.

For the consistency of $\hat{\gamma}_u$ and $\hat{\gamma}_v$, one needs to bound the eigenvalues of $\pm \dot{W}$. Note that $\dot{W} = -W(dW^{-1}/d\gamma)W$, $\pm dW^{-1}/d\gamma \leq 2(1 + |\gamma|)I$, and $W \leq (1 - |\gamma|)^{-1}I$, so the eigenvalues of $\pm \dot{W}$ are bounded. Since

$$\frac{d}{d\gamma} \frac{1}{n} \text{tr}(WW_0^{-1}) = \frac{d}{d\gamma} \frac{(\gamma - \gamma_0)^2}{1 - \gamma^2} + O(n^{-1}) = \frac{2(1 - \gamma_0^2) + 2(\gamma - \gamma_0)^2}{(1 - \gamma^2)^2} (\gamma - \gamma_0) + O(n^{-1})$$

and $n^{-1}d \log |W|/d\gamma = O(n^{-1})$, adaptations of (4.15) and (4.16) yield $\hat{\gamma}_u - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$ and $\hat{\gamma}_v - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$ as $\lambda \rightarrow 0$ and $n\lambda^{1/r} \rightarrow \infty$.

Very much like the block AR(1) models, things also hold for block MA(1) models; details are straightforward, and are omitted here.

4.5 Longitudinal Model

Now consider the longitudinal model of Example 1.3, with the covariance given in (1.4). Assume that $\gamma \in [0, \bar{\gamma}]$ for some $\bar{\gamma} < \infty$. Also assume an upper bound on $\bar{m} = \max_j m_j$. It is necessary that $p \asymp n$, which invalidates the theory of Gu and Ma (2004), where p can not exceed $O(n^{1/2})$.

It is easy to see that $(1 + \bar{m}\bar{\gamma})^{-1}I \leq W \leq I$, so Lemmas 4.1 and 4.2 hold. Also,

$$\begin{aligned} \frac{1}{n}\text{tr}(WW_0^{-1} - I) - \frac{1}{n}\log|WW_0^{-1}| &= \frac{\gamma_0 - \gamma}{n} \sum_{j=1}^p \frac{m_j}{1 + m_j\gamma} + \frac{1}{n} \sum_{j=1}^p \log \frac{1 + m_j\gamma}{1 + m_j\gamma_0} \\ &= (\gamma - \gamma_0) \frac{1}{n} \sum_{j=1}^p \frac{m_j^2(\gamma^* - \gamma_0)}{(1 + m_j\gamma^*)^2} = (\gamma - \gamma_0)^2 C(\gamma), \end{aligned} \quad (4.17)$$

where $\gamma^* = \gamma_0 + \alpha(\gamma - \gamma_0)$ for some $\alpha \in (0, 1)$, the second equality is by the mean value theorem, and the constant $C(\gamma) \leq \bar{m}$. Condition C.1 is seen to hold for $\gamma \rightarrow \gamma_0$, so does Condition C.3. The p nonzero eigenvalues of $(W_0^{-1/2}WW_0^{-1/2} - I)$ are $(\gamma_0 - \gamma)m_j/(1 + m_j\gamma)$, so Condition C.2 holds if $C(\gamma)$ is bounded away from 0 when $\gamma \rightarrow \gamma_0$. Using a second order Taylor expansion for the left-hand side function in (4.17), it is easy to verify that

$$C(\gamma) = \frac{1}{2n} \sum_{j=1}^p \frac{m_j^2}{(1 + m_j\gamma^{**})^3} (1 + m_j\gamma_0 - m_j(\gamma^{**} - \gamma_0)),$$

where γ^{**} is between γ and γ_0 , thus $C(\gamma) \geq \{4(1 + \bar{m}\bar{\gamma})^3\}^{-1}$ for $\gamma - \gamma_0 \leq 1/(2\bar{m})$. Condition C.4 does not hold for all $\gamma \rightarrow \gamma_0$, but does for $\gamma - \gamma_0 = o(R^{1/2})$. We will show below that $\hat{\gamma}_v - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}) = o_p(R^{1/2})$ in meaningful settings, so Theorem 4.2 remains relevant.

The nonzero eigenvalues of \dot{W} are $-m_j/(1 + m_j\gamma)^2$, so $\pm\dot{W} \leq \bar{m}I$. Since

$$\frac{d}{d\gamma} \left(\frac{1}{n}\text{tr}(WW_0^{-1}) - \frac{1}{n}\log|W| \right) = \left(\frac{1}{n} \sum_{j=1}^p \frac{m_j^2}{(1 + m_j\gamma)^2} \right) (\gamma - \gamma_0),$$

adaptation of (4.15) leads to $\hat{\gamma}_u - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$.

The analysis of $\hat{\gamma}_v$ is more involved, however, as $n^{-1}d\log|W|/d\gamma$ is no longer of the order $O(n^{-1})$ as for the AR(1) and MA(1) models. Following the lines of (4.15) and (4.16), one has

$$\begin{aligned} \frac{\partial}{\partial\gamma} V(\lambda, \gamma) &= \frac{1}{n} \left(\frac{1}{\hat{\sigma}^2} \epsilon^T \dot{W} \epsilon - \frac{d}{d\gamma} \log|W| \right) + O_p(\lambda + n^{-1}\lambda^{-1/r}) \\ &= \frac{1}{n} \text{tr}(\dot{W}W_0^{-1}) - \frac{1}{n} \frac{d}{d\gamma} \log|W| - \frac{1}{n} \text{tr}(\dot{W}W_0^{-1}) \frac{1}{\hat{\sigma}^2} (\hat{\sigma}^2 - \sigma^2) + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}) \\ &= \frac{1}{n} \text{tr}(\dot{W}W_0^{-1}) - \frac{1}{n} \frac{d}{d\gamma} \log|W| - \frac{\sigma^2}{\hat{\sigma}^2} \frac{1}{n} \text{tr}(\dot{W}W_0^{-1}) \frac{1}{n} \text{tr}(WW_0^{-1} - I) \\ &\quad + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}) \end{aligned}$$

where all but the term $n^{-1}\epsilon^T W \epsilon - \sigma^2 = n^{-1}\sigma^2 \text{tr}(WW_0^{-1} - I) + O_p(n^{-1/2})$ of (4.8) were bounded

by Lemmas 4.1 and 4.2; see the appendix for similar arguments. Simple calculations yield

$$\begin{aligned} & \frac{\partial}{\partial \gamma} V(\lambda, \gamma) + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}) \\ &= \left\{ \frac{1}{n} \sum_{j=1}^p \frac{m_j^2}{(1 + m_j \gamma)^2} - \frac{\sigma^2}{\hat{\sigma}^2} \frac{1}{n} \sum_{j=1}^p \frac{m_j}{1 + m_j \gamma} \frac{1}{n} \sum_{j=1}^p \frac{m_j(1 + m_j \gamma_0)}{(1 + m_j \gamma)^2} \right\} (\gamma - \gamma_0) = (\gamma - \gamma_0) D(\gamma), \end{aligned}$$

so $\hat{\gamma}_v - \gamma_0 = O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2})$ if $D(\gamma)$ is bounded away from 0 in a neighborhood of γ_0 .

Assume that $p/n \leq 1 - 2\beta$ for some fixed $\beta \in (0, 1/2)$; this is true unless $\#\{j : m_j > 1\} = o(n)$. From (4.8), one has

$$|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2 |\gamma - \gamma_0| + O_p(\lambda + n^{-1}\lambda^{-1/r} + n^{-1/2}),$$

where the fact $|n^{-1}\text{tr}(WW_0^{-1} - I)| \leq |\gamma - \gamma_0|$ was used, so $\sigma^2/\hat{\sigma}^2 \leq (1 - |\gamma - \gamma_0| + o_p(1))^{-1} \leq (1 - \delta - |\gamma - \gamma_0|)^{-1}$ for $|\gamma - \gamma_0| \leq 1/2$ and a small constant δ . Also, $(1 + m_j \gamma_0)/(1 + m_j \gamma) \leq 1 + \bar{m}|\gamma - \gamma_0|$. it follows that for γ in a small but fixed neighborhood of γ_0 ,

$$\frac{p}{n} \frac{\sigma^2}{\hat{\sigma}^2} \frac{1 + m_j \gamma_0}{1 + m_j \gamma} \leq (1 - 2\beta) \frac{1 + \bar{m}|\gamma - \gamma_0|}{1 - \delta - |\gamma - \gamma_0|} \leq 1 - \beta;$$

note that δ can be made arbitrarily small. For γ in this neighborhood,

$$D(\gamma) \geq \frac{p}{n} \left\{ \frac{1}{p} \sum_{j=1}^p \frac{m_j^2}{(1 + m_j \gamma)^2} - (1 - \beta) \left(\frac{1}{p} \sum_{j=1}^p \frac{m_j}{1 + m_j \gamma} \right)^2 \right\} \geq \frac{\beta}{\bar{m}(1 + \bar{m}\bar{\gamma})^2}.$$

5 Empirical Performance

We now present some simulations to illustrate the empirical performance of $V_*(\lambda, \gamma)$ and variations thereof. $U(\lambda, \gamma)$ and $V(\lambda, \gamma)$ are not as useful in practice, with the former assuming a known σ^2 and the latter having a global minimum at $\lambda = 0$.

For independent data, it is well known that Mallows' C_L and the GCV of Craven and Wahba (1979) may yield severe undersmoothing in up to 10% of the cases, despite their otherwise adequate performance and the asymptotic optimality established by Li (1986). A simple multiplicative fudge factor $\alpha > 1$ applied to the term $\text{tr}A$ as shown in (3.5) proves to be very effective in curbing the occasional undersmoothing without affecting the general effectiveness; see, e.g., Kim and Gu (2004). Adopting the same strategy, we consider

$$V_\alpha(\lambda, \gamma) = \log \left\{ \mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma))^2 W_\gamma^{1/2} \mathbf{Y} / n \right\} - \frac{1}{n} \log |W_\gamma| + \alpha \frac{2 \text{tr}A(\lambda, \gamma)}{n - \text{tr}A(\lambda, \gamma)}, \quad (5.1)$$

for some $\alpha \geq 1$. Values of α in the range $1.2 \sim 1.4$ appeared to be good default choices in simulations

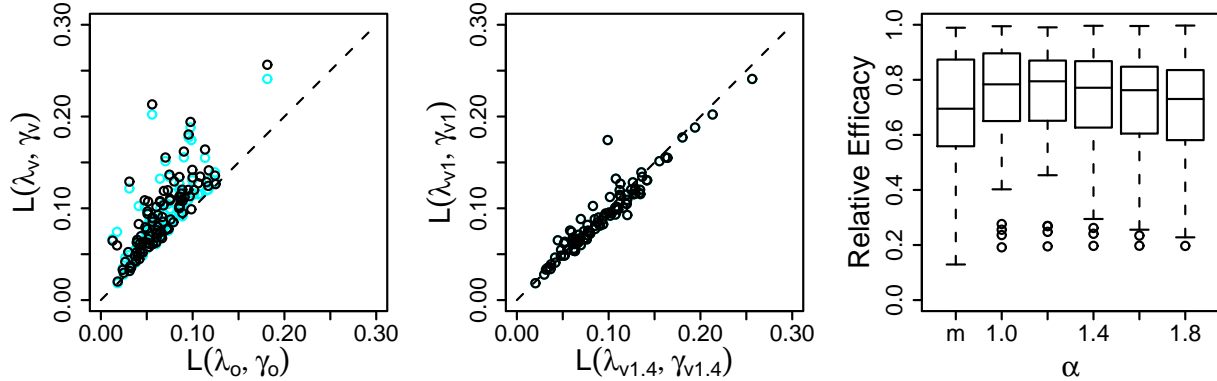


Figure 5.1: AR(1) Simulation with $\gamma = 0.6$ and $n = 100$. Left: Performance of $V_\alpha(\lambda, \gamma)$ with $\alpha = 1$ (faded) and $\alpha = 1.4$ (normal) versus the optimal. Center: Comparison of $V_1(\lambda, \gamma)$ versus $V_{1.4}(\lambda, \gamma)$. Right: Relative efficacy of $M(\lambda, \gamma)$ and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8$.

in other problem settings; see, e.g., Gu and Wang (2003) and Kim and Gu (2004). We also evaluate in parallel the performance of the GML method, the primary viable alternative, which minimizes

$$M(\lambda, \gamma) = \frac{\mathbf{Y}^T W_\gamma^{1/2} (I - A(\lambda, \gamma)) W_\gamma^{1/2} \mathbf{Y}}{|W_\gamma^{1/2} (I - A(\lambda, \gamma)) W_\gamma^{1/2}|_+^{1/(n-m)}}, \quad (5.2)$$

where $|\cdot|_+$ denotes the product of the $n - m$ nonzero eigenvalues and m is the dimension of the null space \mathcal{N}_J (cf. Section 2); see, e.g., Wahba (1985), Opsomer et al. (2001), and Kim and Gu (2004) for the derivation of $M(\lambda, \gamma)$.

5.1 AR(1) Model

Data were generated from $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \dots, n$, for $n = 100, 200$, where $\eta(x) = 5 + 3 \sin 2\pi x$, $x_i \sim U(0, 1)$, and $\epsilon \sim N(\mathbf{0}, 0.5^2 W^{-1}(\gamma))$ with $W(\gamma)$ as given in (1.2). Parallel experiments were conducted with seven different values of the correlation parameter, $\gamma = 0, \pm 0.3, \pm 0.6, \pm 0.9$. For each of the fourteen simulation settings, one hundred replicates were generated, and cubic splines of Example 2.1 were calculated for each replicate with (λ, γ) minimizing (i) the loss $L(\lambda, \gamma)$ of (4.1), (ii) the proposed score $V_\alpha(\lambda, \gamma)$ of (5.1) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, and (iii) the GML score $M(\lambda, \gamma)$ of (5.2).

Shown in Figure 5.1 are some of the simulation results for $\gamma = 0.6$ and $n = 100$. In the left frame, the loss achieved by the minimizer (λ_v, γ_v) of $V_\alpha(\lambda, \gamma)$ for $\alpha = 1$ (faded) and $\alpha = 1.4$ (normal) are plotted against the minimum possible loss $L(\lambda_o, \gamma_o)$. In the middle frame, $V_\alpha(\lambda, \gamma)$ with $\alpha = 1$ and $\alpha = 1.4$ are compared against each other. In the right frame, boxplots are drawn to illustrate the relative efficacy of $V_\alpha(\lambda, \gamma)$ for various α along with that of $M(\lambda, \gamma)$, where relative efficacy is defined by the ratio of the minimum possible loss over the loss achieved by the respective

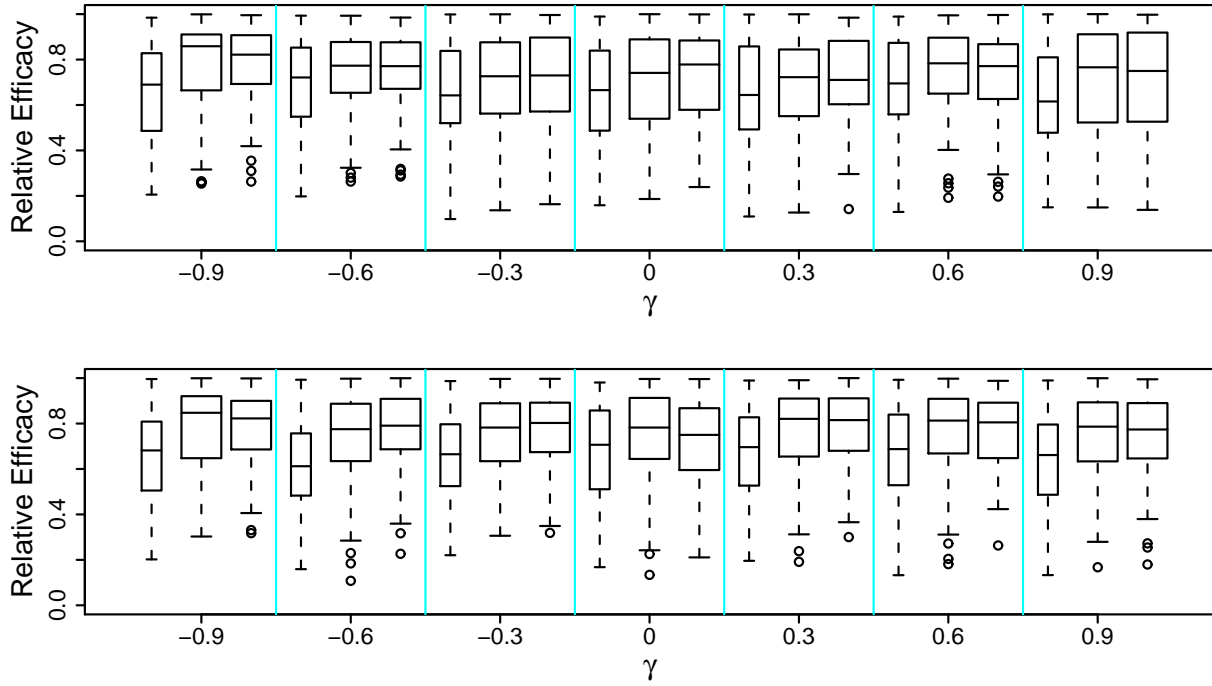


Figure 5.2: AR(1) Simulation Summary. Relative efficacy of $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Top: $n = 100$. Bottom: $n = 200$.

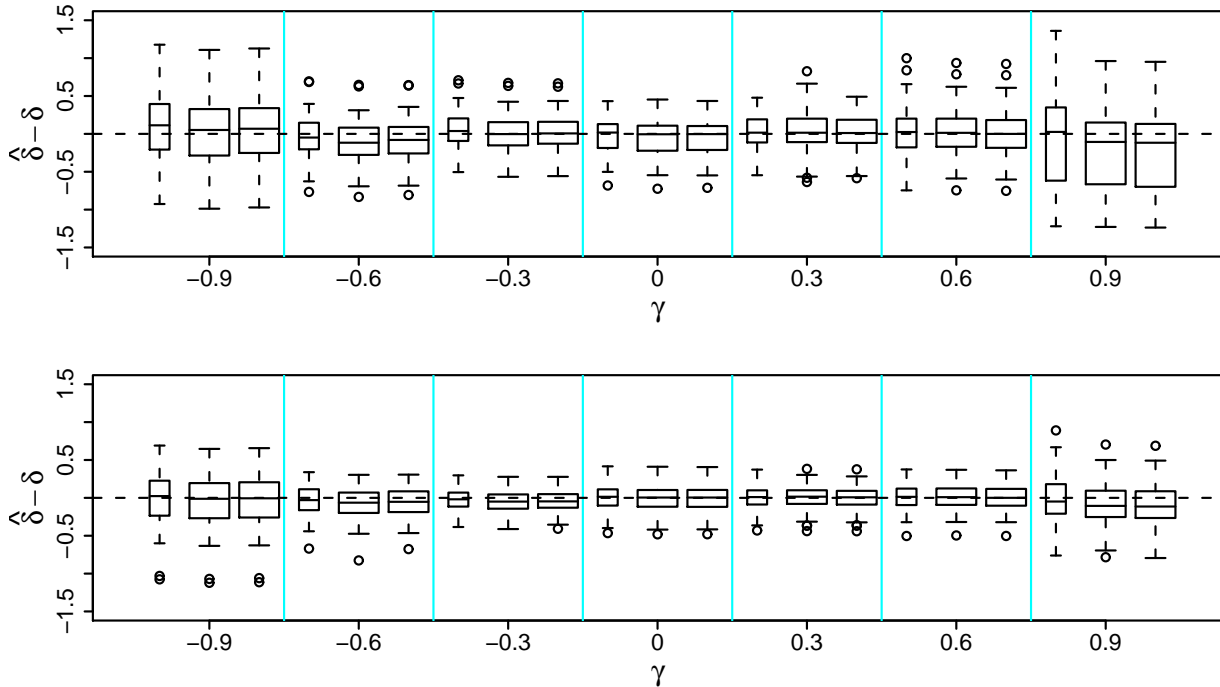


Figure 5.3: AR(1) Simulation Summary. Estimation precision of $\delta = \log\{(1 + \gamma)/(1 - \gamma)\}$ using $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Top: $n = 100$. Bottom: $n = 200$.

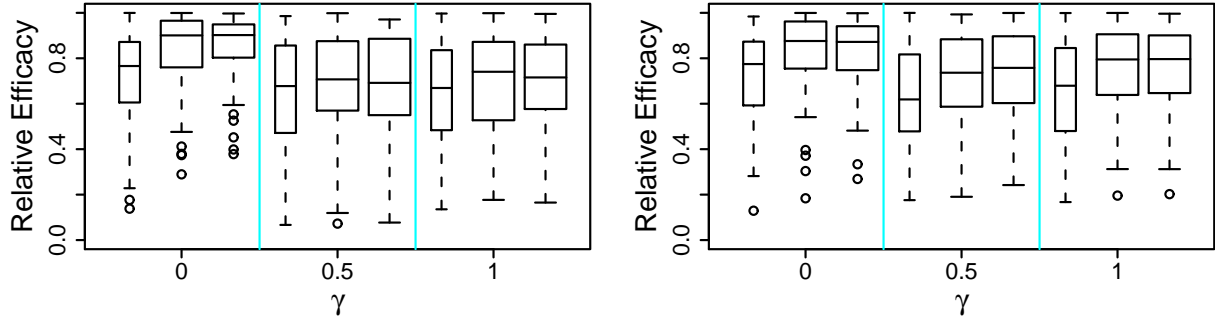


Figure 5.4: Longitudinal Simulation Summary. Relative efficacy of $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Left: $n = 100$. Right: $n = 200$.

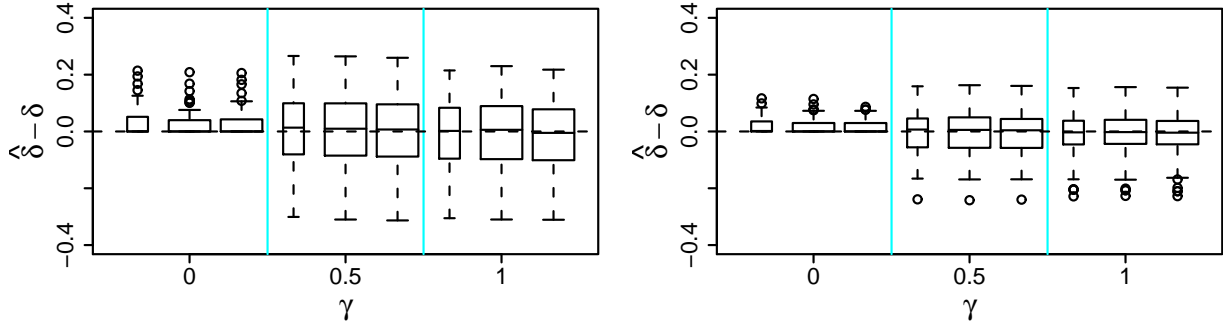


Figure 5.5: Longitudinal Simulation Summary. Estimation precision of $\delta = \gamma/(1+\gamma)$ using $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Left: $n = 100$. Right: $n = 200$.

methods. Parallel plots for other settings look similar. The relative efficacy of $M(\lambda, \gamma)$ and $V_\alpha(\lambda, \gamma)$, $\alpha = 1, 1.4$, in all fourteen settings are summarized in Figure 5.2. It is clear that $V_\alpha(\lambda, \gamma)$ generally outperform $M(\lambda, \gamma)$. Figure 5.3 illustrate the estimation of γ through $M(\lambda, \gamma)$ and $V_\alpha(\lambda, \gamma)$, where the boxplots depict $\log\{(1 + \hat{\gamma})/(1 - \hat{\gamma})\} - \log\{(1 + \gamma)/(1 - \gamma)\}$.

5.2 Longitudinal Model

Data were generated from $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \dots, n$, for $n = 100, 200$, where $\eta(x) = 5 + 3 \sin 2\pi x$, $x_i \sim U(0, 1)$, and $\epsilon \sim N(\mathbf{0}, 0.5^2 W^{-1}(\gamma))$ with $W(\gamma)$ as given in (1.4) for $m_j = m = 5$ and $p = n/m = 20, 40$. Parallel experiments were conducted with three different values of the correlation parameter, $\gamma = 0, 0.5, 1$. For each of the six simulation settings, one hundred replicates were generated, and cubic splines of Example 2.1 were calculated for each replicate with (λ, γ) minimizing (i) the loss $L(\lambda, \gamma)$ of (4.1), (ii) the proposed score $V_\alpha(\lambda, \gamma)$ of (5.1) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, and (iii) the GML score $M(\lambda, \gamma)$ of (5.2). The results are summarized in Figures 5.4 and 5.5 in parallel to Figures 5.2 and 5.3. The counterparts of Figure 5.1, not shown here, look similar.

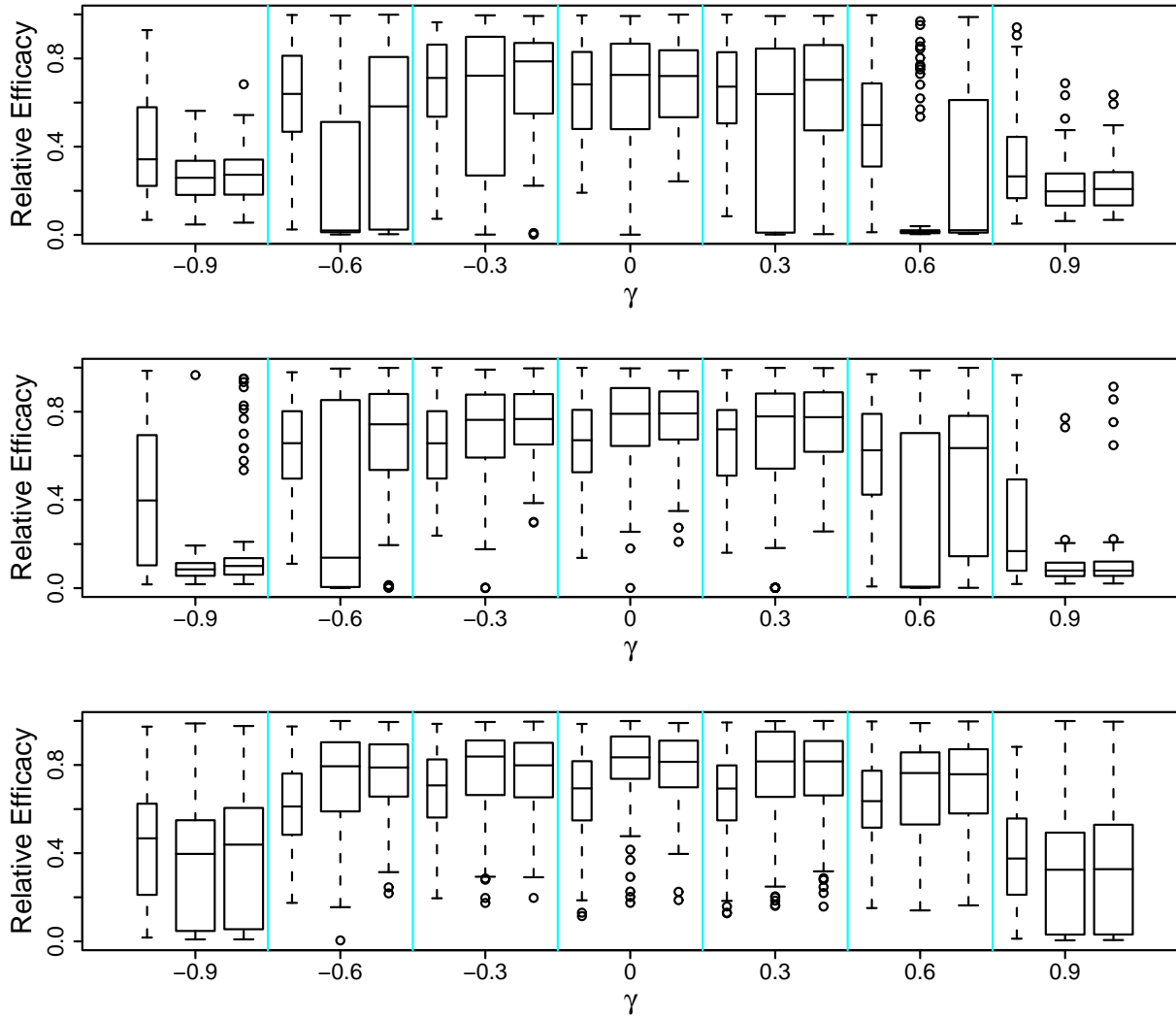


Figure 5.6: Ma(1) Simulation Summary. Relative efficacy of $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Top: $n = 100$. Middle: $n = 200$. Bottom: $n = 400$.

5.3 MA(1) Model

Data were generated from $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \dots, n$, for $n = 100, 200, 400$, where $\eta(x) = 5 + 3 \sin 2\pi x$, $x_i \sim U(0, 1)$, and $\epsilon \sim N(\mathbf{0}, 0.5^2 W^{-1}(\gamma))$ with $W(\gamma)$ as given in (1.3). Parallel experiments were conducted with seven different values of the correlation parameter, $\gamma = 0, \pm 0.3, \pm 0.6, \pm 0.9$. For each of the twenty one simulation settings, one hundred replicates were generated, and cubic splines of Example 2.1 were calculated for each replicate with (λ, γ) minimizing (i) the loss $L(\lambda, \gamma)$ of (4.1), (ii) the proposed score $V_\alpha(\lambda, \gamma)$ of (5.1) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, and (iii) the GML score $M(\lambda, \gamma)$ of (5.2). The results are summarized in Figures 5.6 and 5.7 in parallel to Figures 5.2 and 5.3; some of the boxplots in Figure 5.7 are out of bounds due to the delimited vertical range made

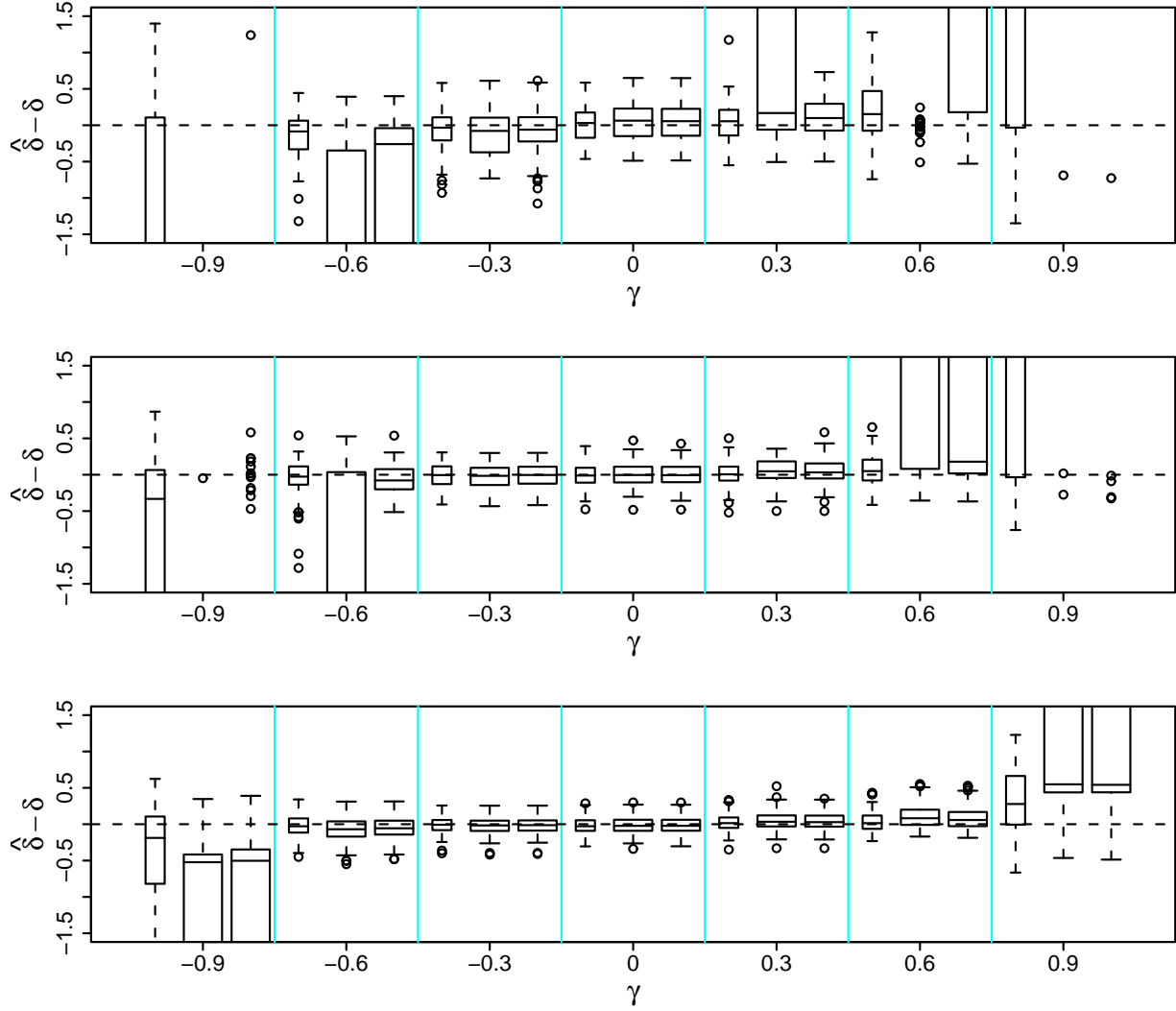


Figure 5.7: MA(1) Simulation Summary. Estimation precision of $\delta = \log\{(1 + \gamma)/(1 - \gamma)\}$ using $M(\lambda, \gamma)$ (thinner boxes) and $V_\alpha(\lambda, \gamma)$ with $\alpha = 1, 1.4$, in order. Top: $n = 100$. Middle: $n = 200$. Bottom: $n = 400$.

to be comparable to Figure 5.3.

In contrast to the parallel results in the AR(1) and longitudinal model simulations, the overall performance of the methods in the MA(1) model settings appear less favorable. A careful look at the plots reveals the following: (i) for $n = 100$, $V_1(\lambda, \gamma)$ works only for $\gamma = 0$ while $V_{1.4}(\lambda, \gamma)$ works for $\gamma = 0, \pm 0.3$, and both are a toss-up against $M(\lambda, \gamma)$; (ii) for $n = 200$, the methods do work for $\gamma = 0, \pm 0.3$, and the comparisons of $V_\alpha(\lambda, \gamma)$ and $M(\lambda, \gamma)$ follow the same pattern as seen in the AR(1) and longitudinal model simulations; (iii) for $n = 400$, the methods work for $\gamma = 0, \pm 0.3, \pm 0.6$ and the comparisons of $V_\alpha(\lambda, \gamma)$, and $M(\lambda, \gamma)$ again follow the same pattern. Also, when the methods do work, the estimation precisions of γ are similar to those in the AR(1) model

settings as seen in Figure 5.3. It appears that the asymptotic optimality does take over eventually, but the sample size needed for it to kick in depends on the γ at work. In search for a possible explanation for the phenomenon, we observe the fixed upper bound of $W_\gamma \leq (1 + |\gamma|)^2 I \leq 4I$ for the AR(1) model and $W_\gamma \leq I$ for the longitudinal model, but the “floating” upper bound of $W_\gamma \leq (1 - |\gamma|)^{-2} I$ for the MA(1) model.

It is also evident from Figure 5.7 that in the majority of the failed cases the estimation of γ errs towards the extremes, i.e., towards -1 for $\gamma < 0$ and towards 1 for $\gamma > 0$. It was also observed, though not shown here, that the selected λ in the failed cases was towards the smaller end, yielding undersmoothing or interpolation.

6 Example

We now apply the techniques to a data set listed as Series A in Box, Jenkins, and Reinsel (1994, page 541) to illustrate further aspects. The series consists of 197 chemical process concentration readings taken at every two hours in an industrial setting. A “pure-noise” AR(2) model

$$Y_t - \mu = \gamma_1(Y_{t-1} - \mu) + \gamma_2(Y_{t-2} - \mu) + a_t,$$

with $\mu = 17.06$, $\gamma_1 = 0.4245$, and $\gamma_2 = 0.2531$, fits the data well.

We shall consider here a “signal-plus-noise” model of the form $Y_t = \eta(t) + \epsilon_t$, with ϵ_t following the AR(1) model of Example 1.1. Shown in Figure 6.1 are four cubic spline fits of $\eta(t)$ along with the data, where two were fitted by penalized least squares assuming independence of ϵ_t with λ selected via (3.5) for $\alpha = 1, 1.4$, and two by penalized likelihood with (λ, γ) selected via (5.1) for $\alpha = 1, 1.4$. The four sets of tuning parameters are $(\log_{10}(n\lambda), \gamma) = (-5.61, 0), (-4.43, 0), (-4.68, 0.08)$, and $(-3.07, 0.25)$, in order; the second and the third fits are nearly identical. The four sets of residuals were also collected and AR(1) models were fitted to the residuals, yielding the estimates $\gamma = -0.06, 0.10, 0.08$, and 0.26 , in order, with standard errors all rounding to 0.07 . Inspections of the autocorrelation functions indicate that the first three sets of residuals are pretty much white noise and an AR(1) model fits the fourth set well. A fifth fit was also calculated with (λ, γ) selected by $M(\lambda, \gamma)$, which virtually coincided with the fourth fit through $V_{1.4}(\lambda, \gamma)$.

In this example, the time t is used both as the covariate in the regression function $\eta(t)$ and as the index defining the correlation structure of the error ϵ_t , contrasting the separation of the two in the simulation settings of Section 5. When the covariate is also used to define the correlation structure of the error, the identifiability of multiple models can be problematic, and indeed the four cubic spline fits discussed here plus the “pure-noise” AR(2) fit all appear reasonable for the data.

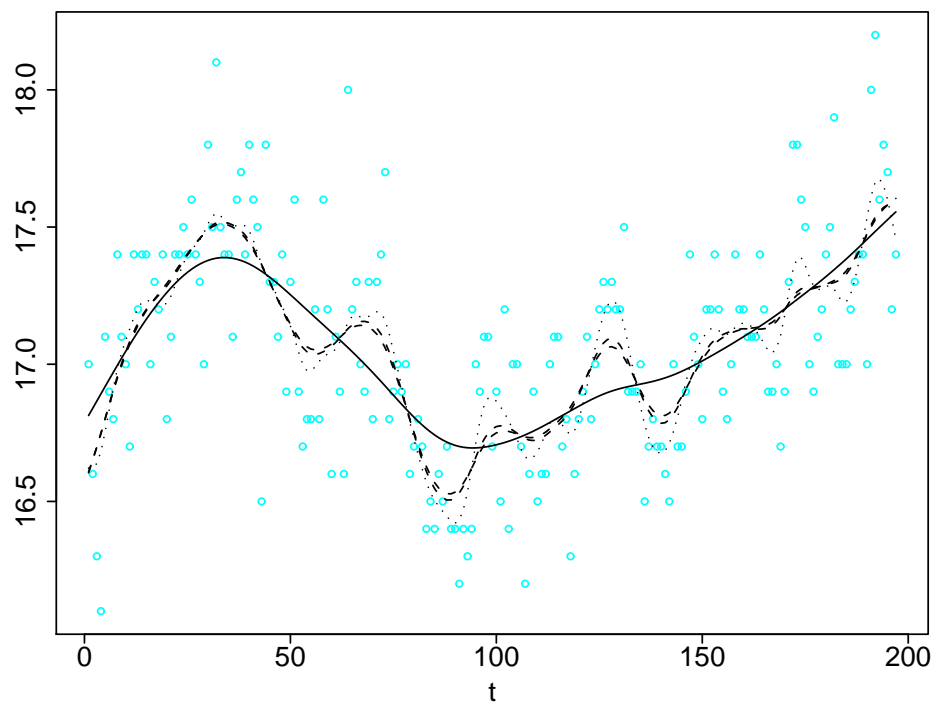


Figure 6.1: Chemical Concentration. Fit with $(\log_{10}(n\lambda), \gamma) = (-3.07, 0.25)$ is in solid line, that with $(-5.61, 0)$ is in dotted line, and those with $(-4.43, 0)$ and $(-4.68, 0.08)$ are in dashed lines. The data are in faded circles.

7 Discussion

In this article, we proposed an effective approach to tuning parameter selection in penalized likelihood estimation with correlated Gaussian data. The asymptotic optimality of the method has been established theoretically and illustrated empirically for a few commonly used correlation models.

While the general theory is verified only in a few simple correlation models where key quantities are analytically tractable, its validity could hold in more involved settings. In the light of the illuminating results from the MA(1) model simulations, the upper bound on the eigenvalues of the matrix W appears to be a deciding factor. Remember that $E[\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}^T] = \sigma^2 W^{-1}$, so a tight upper bound on the eigenvalues of W effectively keeps the correlation matrix away from singularity. In particular, correlation models of the form $W^{-1} = \delta I + V(\gamma)$ with $\delta > 0$ and $V(\gamma) \geq 0$ are likely safe domains for the proposed method to work effectively.

Caution must be exercised in the practical applications of the developed technique. While the method generally works well with correctly specified correlation structures, robustness against misspecifications remain unknown. Model identification also remains largely a judgment call based on the application setting rather than a data analytical exploration. When the covariate is involved in the correlation model specification as in the example of Section 6, model identifiability remains an issue for the interpretation of data analysis.

Appendix

In this appendix, we calculate the order of $Q(\lambda, \gamma)$ defined in (4.14).

Note that $\dot{W}^k \leq c_1 W$, $k = 1, 2$, for some constant c_1 as the eigenvalues of \dot{W} and W^{-1} are bounded from above, so $n^{-1} \boldsymbol{\eta}_0^T (I - \tilde{A})^T \dot{W}^k (I - \tilde{A}) \boldsymbol{\eta}_0 = O(\lambda)$ by Lemma 4.2 and Assumption A. Similarly, $\tilde{A}^T \tilde{A} = W^{1/2} A W^{-1} A W^{1/2}$ and $A W^{-1} A \leq c_2 I$ for some c_2 , so $n^{-1} \boldsymbol{\eta}_0^T (I - \tilde{A})^T \tilde{A}^T \tilde{A} (I - \tilde{A}) \boldsymbol{\eta}_0 = O(\lambda)$. Combining these and applying the Cauchy-Schwarz inequality, the first term of (4.14) is of the order $O(\lambda)$. Similar arguments, though a bit messier, yield the order $O_p(\lambda)$ for the second term.

We now bound $\boldsymbol{\epsilon}^T \dot{W} \tilde{A} \boldsymbol{\epsilon}$ in the third term. As $\dot{W} \tilde{A} = \dot{W} W^{-1/2} A W^{1/2}$, by the Cauchy-Schwarz inequality,

$$|\boldsymbol{\epsilon}^T \dot{W} \tilde{A} \boldsymbol{\epsilon}| \leq (\boldsymbol{\epsilon}^T \dot{W} W^{-1/2} A W^{-1/2} \dot{W} \boldsymbol{\epsilon})^{1/2} (\boldsymbol{\epsilon}^T W^{1/2} A W^{1/2} \boldsymbol{\epsilon})^{1/2}.$$

Now

$$\begin{aligned} E[\boldsymbol{\epsilon}^T \dot{W} W^{-1/2} A W^{-1/2} \dot{W} \boldsymbol{\epsilon}] &= \text{tr}(A W^{-1/2} \dot{W} W_0^{-1} \dot{W} W^{-1/2}) = O(\lambda^{-1/r}), \\ E[\boldsymbol{\epsilon}^T W^{1/2} A W^{1/2} \boldsymbol{\epsilon}] &= \text{tr}(A W^{1/2} W_0^{-1} W^{1/2}) = O(\lambda^{-1/r}), \end{aligned}$$

by Lemma 4.1 and Assumption A, as $W^{-1/2} \dot{W} W_0^{-1} \dot{W} W^{-1/2} \leq c_3 I$ and $W^{1/2} W_0^{-1} W^{1/2} \leq c_4 I$ for

some c_3 and c_4 , so $\epsilon^T \tilde{W} \tilde{A} \epsilon = O_p(\lambda^{-1/r})$. Other terms can be bounded similarly but the arguments are slightly messier. The third term of (4.14) is thus of the order $O_p(n^{-1} \lambda^{-1/r})$.

Acknowledgements

This research was supported by NIH under Grant R33HL68515.

References

- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis* (Third ed.). Englewood Cliffs, NJ: Prentice Hall.
- Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* 31, 377–403.
- Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.
- Gu, C. and Y.-J. Kim (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist.* 30, 619–628.
- Gu, C. and P. Ma (2004). Optimal smoothing in nonparametric mixed-effect models. Revised for *Ann. Statist.*
- Gu, C. and J. Wang (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation. *Statist. Sin.* 13, 811–826.
- Kim, Y.-J. and C. Gu (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B* 66, 337–356.
- Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing. *Ann. Statist.* 14, 1101–1112.
- Opsomer, J. D., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statist. Sci.* 16, 134–153.
- Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* 13, 1378–1402.
- Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.
- Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* 93, 341–348.