

DEPARTMENT OF MATHEMATICS
TECHNICAL REPORT

Generalized Cross-Validation for Correlated Data (GCV_c)

Patrick S. Carmack

March 2011

No. 2011 – 1

UNIVERSITY OF CENTRAL ARKANSAS
Conway, AR 72035

GENERALIZED CROSS-VALIDATION FOR CORRELATED DATA (GCV_c)

PATRICK S. CARMACK[†]

*Department of Mathematics
University of Central Arkansas
201 Donaghey Avenue, Conway, AR 72035-5001, USA*

JEFFREY S. SPENCE

*Department of Internal Medicine, Epidemiology Division
University of Texas Southwestern Medical Center at Dallas
5323 Harry Hines Boulevard, Dallas, TX 75390-8874, USA*

WILLIAM R. SCHUCANY

*Department of Statistical Science
Southern Methodist University
P.O. Box 750332, Dallas, TX 75275-0332, USA*

ABSTRACT. Since its introduction by Stone (1974) and Geisser (1975), cross-validation has been studied and improved by several authors including Burman et al. (1994), Hart & Yi (1998), Racine (2000), Hart & Lee (2005), and Carmack et al. (2009). Perhaps the most widely used and best known is generalized cross-validation (GCV) (Craven & Wahba, 1979), which establishes a single-pass method that penalizes the fit by the trace of the smoother matrix assuming independent errors. We propose an extension to GCV in the context of correlated errors that has important implications about the definition for residual degrees of freedom, even in the independent case. The efficacy of the new method is demonstrated by simulation and application with concluding remarks about the heteroscedastic case and a potential maximum likelihood framework.

E-mail addresses: patrickc@uca.edu, jeffrey.spence@utsouthwestern.edu, schucany@smu.edu.

Date: March 28, 2011.

Key words and phrases. correlation, effective degrees of freedom, linear smoother, model selection.

[†]Corresponding author.

1. INTRODUCTION

Cross-validation has a rich history starting with ordinary cross-validation (Stone, 1974; Geisser, 1975). The original method works by withholding a single data point at a time while using the rest of the data to predict the withheld response. In the context of model selection, the model with the smallest cross-validated squared error is then declared to be the best one. Ordinary cross-validation is not consistent for model selection, but v -fold cross-validation addresses this issue by randomly partitioning the data into training and test sets where models are fit using training sets and assessed using test sets. Both of these methods assume independent errors.

Subsequent papers extended cross-validation for correlated data. One known as h -block cross-validation (Burman et al., 1994) did so by withholding blocks of data when estimating parameters and using the full dataset for model assessment. Racine (2000) combined h -block and v -fold cross-validation to arrive at a consistent method, hv -block cross-validation. Hart & Yi (1998) proposed one-sided cross-validation, which omits the data either to the left or right of the point of estimation, including the point, and then assesses squared error performance. While initially intended for independent errors, Hart & Lee (2005) demonstrated that the method is robust in the presence of low to moderately correlated errors. Finally, Carmack et al. (2009) proposed a method similar to h -block cross-validation known as far casting cross-validation (FCCV). Their method uses the full dataset to estimate model parameters while omitting certain neighbors for model assessment purposes.

Craven & Wahba (1979) proposed a single-pass consistent method for independent data known as generalized cross-validation (GCV). Their ingenious use of degrees of freedom makes this possible and is a concept that we extend to correlated data. We motivate the extension, generalized cross-validation for correlated data (GCV_c), from a nonparametric perspective, but conclude with some interesting connections to a parametric setting.

2. THEORY

Suppose $y_i = f(x_i) + \varepsilon_i$, $i = 1, \dots, n$, where $f(\cdot)$ is a function, and ε_i is stochastic with $E[\varepsilon_i] = 0$, $\text{Var}[\varepsilon_i] = \sigma^2 < \infty$, and $n \times n$ covariance matrix given by $(\Sigma)_{ij} = \sigma^2(C)_{ij} = \sigma^2 \text{cor}(\varepsilon_i, \varepsilon_j) = \sigma^2 r_{ij}$, $C \neq J$. We are interested in finding $\hat{f}(\cdot)$ to estimate $f(\cdot)$, which we assume takes the form of a linear smoother. That is, $\hat{f}(x) = \sum_{i=1}^n w_i y_i$, where the weights, w_1, \dots, w_n , are a function of \mathbf{x} and a vector of tuning parameters, $\boldsymbol{\theta}$, with $\sum_{i=1}^n w_i = 1$. Many cross-validation techniques are commonly used to estimate such tuning parameters by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \text{CV}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{k=1}^n \left(\hat{f}_{cv}(x_k | \boldsymbol{\theta}) - y_k \right)^2,$$

where $\hat{\boldsymbol{\theta}}$ is the vector of tuning parameters that minimizes the cross-validation error surface (assuming a unique minimum exists) and $\hat{f}_{cv}(x_k|\boldsymbol{\theta})$ is estimated on some portion of the data, which usually excludes (x_k, y_k) and possibly other data points. The final fit, $\hat{f}(\cdot|\hat{\boldsymbol{\theta}})$, is then given by using $\hat{\boldsymbol{\theta}}$ in conjunction with the full dataset. In the interest of compact notation, \hat{f}_{cv} and \hat{f} 's dependency on $\boldsymbol{\theta}$ will be omitted henceforth. Carmack et al. (2009) developed a cross-validation method that specifically deals with correlated errors. They derive the following expression for a single term in the CV(\cdot) function.

$$\begin{aligned} \mathbb{E} \left[\left(\hat{f}_{cv}(x_k) - y_k \right)^2 \right] &= \mathbb{E} \left[\left(\hat{f}(x_k) - f(x_k) \right)^2 \right] + \sigma^2 - \text{Var} \left[\hat{f}(x_k) \right] + \text{Var} \left[\hat{f}_{cv}(x_k) \right] \\ &\quad + \mathbb{E} \left[\hat{f}_{cv}(x_k) - \hat{f}(x_k) \right] \left(\mathbb{E} \left[\hat{f}_{cv}(x_k) + \hat{f}(x_k) \right] - 2f(x_k) \right) - 2\text{Cov} \left[\hat{f}_{cv}(x_k), y_k \right], \end{aligned} \quad (1)$$

where $\hat{f}_{cv}(x_k)$ is the hold-out estimate of $f(x_k)$, while $\hat{f}(x_k)$ is the estimate of $f(x_k)$ using all the data. If one allows $\hat{f}_{cv}(\cdot) = \hat{f}(\cdot)$, as is the case for GCV, this expression simplifies to

$$\mathbb{E} \left[\left(\hat{f}(x_k) - y_k \right)^2 \right] = \mathbb{E} \left[\left(\hat{f}(x_k) - f(x_k) \right)^2 \right] + \sigma^2 - 2\text{Cov} \left[\hat{f}(x_k), y_k \right], \quad (2)$$

which shows that the expectation of a single cross-validation term is the true squared error (a desirable property) along with the other two terms. Since σ^2 is a constant, it plays no role in minimizing the cross-validation error curve in expectation. However, the covariance term turns out to be crucial. Its role has been well recognized (Hastie et al., 2009). This is the primary reason that ordinary cross-validation performs well in independent data since the covariance term is identically zero in that case for (1). In such cases, GCV successfully accommodates the shift from (1) to (2) by penalizing the fit for the covariance term using a particular definition for residual degrees of freedom. However, once data are correlated, withholding (x_k, y_k) to estimate $f(x_k)$ is no longer sufficient to eliminate the covariance between the estimate, $\hat{f}_{cv}(x_k)$, and the data in ordinary cross-validation. The same is true of GCV since it only accounts for the covariance between $\hat{f}(x_k)$ and y_k under independence. Further expansion of (2) reveals how to properly handle the correlated case, specifically

$$\begin{aligned}
\mathbb{E} \left[\left(\hat{f}(x_k) - y_k \right)^2 \right] &= \mathbb{E} \left[\left(\hat{f}(x_k) - f(x_k) \right)^2 \right] + \sigma^2 - 2\text{Cov} \left[\hat{f}(x_k), y_k \right] \\
&= \left(\sum_{i=1}^n w_i f(x_i) - f(x_k) \right)^2 + \text{Var} \left[\hat{f}(x_k) \right] + \sigma^2 - 2\text{Cov} \left[\hat{f}(x_k), y_k \right] \\
&= \left(\sum_{i=1}^n w_i f(x_i) - f(x_k) \right)^2 + \sigma^2 + \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij} - 2 \sum_{i=1}^n w_i \sigma_{ik} \\
&= \left(\sum_{i=1}^n w_i f(x_i) - f(x_k) \right)^2 + \sigma^2 \left(1 + \sum_{i=1}^n \sum_{j=1}^n w_i w_j r_{ij} - 2 \sum_{i=1}^n w_i r_{ik} \right).
\end{aligned}$$

Letting $\hat{\mathbf{f}}(\mathbf{x}) = S\mathbf{y}$, where $(S)_{ij} = w_{ij}$, $\mu_k = \sum_{i=1}^n w_{ki}y_i - y_k$, and $R_{k\ell} = r_{k\ell} + \sum_{i=1}^n \sum_{j=1}^n w_{ki}w_{\ell j}r_{ij} - \sum_{i=1}^n w_{ki}r_{\ell i} - \sum_{i=1}^n w_{\ell i}r_{ki}$, one can now show that

$$\begin{aligned}
\mathbb{E} \left[\sum_{k=1}^n \left(\hat{f}(x_k) - y_k \right)^2 \right] &= \sum_{k=1}^n \mu_k^2 + \sigma^2 \sum_{k=1}^n R_{kk} \\
&= \sum_{k=1}^n \mu_k^2 + \sigma^2 \text{tr} [C + SCS' - 2SC].
\end{aligned}$$

Provided that $C = I$ and the smoother matrix S is symmetric and idempotent, as is the case for many linear fitting techniques, the trace term reduces to $n - \text{tr}[S]$, which is proportional to the familiar denominator in GCV.

Historically, there are three major contenders for defining residual degrees of freedom under independence in the context of linear smoothers (Buja et al., 1989). These are all equivalent for S idempotent and symmetric, namely

$$\text{tr} [I - (2S - SS')], \quad (3)$$

$$\text{tr} [I - S], \text{ and} \quad (4)$$

$$\text{tr} [I - SS']. \quad (5)$$

In light of the preceding discussion, we propose the following definition for residual degrees of freedom:

$$\text{tr} [C + SCS' - 2SC] = n - \text{tr} [2SC - SCS'], \quad (6)$$

which is the analogue of (3) when taking correlation into account. As shown by Buja et al. (1989) through eigen decomposition, $\text{tr}[I - (2S - SS')] \leq \text{tr}[I - S] \leq \text{tr}[I - SS']$. This implies that (6) is the most stringent of the three and differs from (4) employed by GCV even in the independent case. Adopting this definition of degrees of freedom leads us to define the generalized cross-validation for correlated data surface as

$$\text{GCV}_c^{(1)}(\boldsymbol{\theta}) = \frac{1}{n} \left(\frac{\sum_{k=1}^n y_k - \hat{f}(x_k)}{1 - \text{tr}[2SC - SC S'] / n} \right)^2, \quad (7)$$

with $\hat{\boldsymbol{\theta}}$ as its minimizer provided $0 \preceq S \prec I$.

3. SIMULATIONS

The simulation study here is similar to that of Carmack et al. (2009), where FCCV was shown to perform as well as or better than other cross-validation methods in correlated data. We are interested in assessing the performance of the included methods whose task is to select a global bandwidth for local linear regression (Fan, 1992) with serially correlated additive errors. The local linear regression estimate is given by

$$\begin{aligned} \hat{f}(x_k | h) &= \frac{\sum_{i=1}^n w_i(x_i | h) y_i}{\sum_{i=1}^n w_i(x_i | h)}, \text{ where} \\ w_i(x | h) &= K\left(\frac{x - x_i}{h}\right) (t_{n,2} - (x - x_i) t_{n,1}), \text{ and} \\ t_{n,j}(x | h) &= \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) (x - x_i)^j, j = 1, 2. \end{aligned}$$

The tricube kernel, $K(u) = 35/32(1 - u^2)^3$, $0 \leq u \leq 1$, was chosen for our purposes. The following four functions were selected for variety of structure (Fig. 1):

$$\begin{aligned} f_1(x) &= x^3(1 - x)^3, \\ f_2(x) &= (x/2)^3(1 - x/2)^2, \\ f_3(x) &= 1.741 \cdot \left[2x^{10}(1 - x)^2 + x^2(1 - x)^{10} \right], \text{ and} \\ f_4(x) &= \begin{cases} 0.0212 \cdot \exp(x - 1/3), & x < 1/3 \\ 0.0212 \cdot \exp(-2(x - 1/3)), & x \geq 1/3 \end{cases}. \end{aligned}$$

Each function, $f(\cdot) = f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, or $f_4(\cdot)$, was sampled at either $n = 75$ or $n = 150$ equally spaced design points in the interval $[0, 1]$ with $x_i = (i - 0.5)/n$, $i = 1, \dots, n$. A vector of serially correlated errors, ε_i , $i = 1, \dots, n$, was generated using `arma.sim` in R

([R Development Core Team, 2010](#)) from a first-order autoregressive process (AR(1)) with coefficient $\phi = 0.0, 0.3, 0.6$, or 0.9 , which ranges from independent to heavily correlated, and standard deviation $\sigma = 2^{-11}, 2^{-9}$, or 2^{-7} , which ranges from low to high variance. Each realization results in a dataset $(x_i, f(x_i) + \varepsilon_i), i = 1, \dots, n$, which was repeated 1,000 times for each combination of $f(\cdot), n, \phi$, and σ .

For comparison, we included FCCV along with the following:

$$\text{GCV}_c^{(2)}(\boldsymbol{\theta}) = \frac{1}{n} \left(\frac{\sum_{k=1}^n y_k - \hat{f}(x_k)}{1 - \text{tr}[SC]/n} \right)^2, \text{ and} \quad (8)$$

$$\text{GCV}_c^{(3)}(\boldsymbol{\theta}) = \frac{1}{n} \left(\frac{\sum_{k=1}^n y_k - \hat{f}(x_k)}{1 - \text{tr}[SCS']/n} \right)^2, \quad (9)$$

whose denominators are proportional to the correlated analogues of (4) and (5), respectively.

For each realization, one bandwidth was estimated as the minimizer of (7), (8), (9), or the FCCV error curve with the withholding neighborhood d set to the recommended value of $3/n$. The function `optimize` in R was used for all four methods with $0 \leq h \leq 1$. For the three GCV_c criteria, the correlation matrix C was either assumed known or estimated using the first five lags of the raw empirical semivariogram of the data fit using a nonparametric semivariogram estimator. The raw empirical semivariogram for a time series at lag k is $\tilde{\gamma}_k = \sum_{|i-j|=k} (y_i - y_j)^2 / 2(n-k)$ with the nonparametric fit given by $\hat{\gamma}(k) = \sum_{i=1}^m [1 - \Omega_\kappa(kt_i)] p_i$, where κ is the order of the basis set to 11 for our purposes, and $p_i, i = 1, \dots, m$, is the nonnegative least squares minimizer of $\sum_{k=1}^5 (\tilde{\gamma}_k - \hat{\gamma}(k))^2$. See [Cherry et al. \(1996\)](#) for further details concerning the nonparametric semivariogram estimator. One should note that $\tilde{\gamma}_k$ is a biased estimate of the semivariance at lag k since the data contains mean structure (i.e., $E[\tilde{\gamma}_k] = \sigma^2(1 - \phi^k)/(1 - \phi^2) + \sum_{|i-j|=k} [f(x_i) - f(x_j)]^2 / [2(n-k)]$), which is why the nonparametric semivariogram is fit using the first five lags of the empirical semivariogram where differencing hopefully removes much of the mean structure at lags close together. The nonparametric semivariogram is then used to estimate C as

$$\left(\hat{C}\right)_{ij} = 1 - \frac{\hat{\gamma}(|i-j|)}{\sum_{i=1}^m p_i},$$

where $\sum_{i=1}^m p_i$ is the sill estimate, which represents the variance of observations far apart.

Once each of the four methods yielded an estimate of bandwidth, \hat{h} , for each iteration, the average squared error was calculated as

$$\text{ASE}(\hat{h}) = \frac{1}{n} \sum_{i=1}^n \left(f(x_i) - \hat{f}(x_i | \hat{h}) \right)^2,$$

which will be our basis for comparison. Additionally, the bandwidth, h_0 , for each iteration and associated ASE, ASE_0 , using full knowledge of the underlying function was estimated to serve as a baseline.

As Figs. 3 and 4 (reviewers please see Figs. 7–12 in the Appendix) demonstrate, $\text{GCV}_c^{(1)}$ dominates both $\text{GCV}_c^{(2)}$ and $\text{GCV}_c^{(3)}$ provided the correlation structure is known in terms of mean ASE ratio. This is not the case when the correlation structure is estimated where $\text{GCV}_c^{(1)}$ exhibits considerably higher mean ASE ratios when the error variance is low. An inspection of the bandwidth estimates showed that $\text{GCV}_c^{(1)}$ tends to over smooth in these cases. Table 1 (reviewers please see Tables 2–4) contains the means of the ratios of each methods estimated bandwidth and h_0 for each iteration. Further investigation revealed that the bias in the empirical semivariogram is the cause since the magnitude of the function overwhelms that of the variance in the bias expression given above. This phenomenon causes the correlation to be overestimated, which $\text{GCV}_c^{(1)}$ punishes more heavily than either $\text{GCV}_c^{(2)}$ or $\text{GCV}_c^{(3)}$, especially at lower bandwidths (Fig. 5). However, $\text{GCV}_c^{(1)}$ again dominates the other two at medium and high variances when estimating C . Although the results for $f_2(\cdot)$, $f_3(\cdot)$, and $f_4(\cdot)$ are not shown, the preceding comments generally apply to them as well.

Further study of Table 1 (reviewers please see Tables 2–4) shows that $\text{GCV}_c^{(2)}$ appears to be less biased for h_0 than $\text{GCV}_c^{(1)}$ when the correlation structure is known with the latter selecting higher bandwidths on average. However, $\text{GCV}_c^{(2)}$'s reduced bias for h_0 appears to come at the price of higher average ASE relative to that of $\text{GCV}_c^{(1)}$. This is likely a case of bias-variance tradeoff since $\text{GCV}_c^{(1)}$ accounts for the total variance of $C + SCS' - SC - CS'$ when the errors are multivariate normal as the discussion section shows.

FCCV performed well, often besting $\text{GCV}_c^{(2)}$ and $\text{GCV}_c^{(3)}$ in terms of mean ASE. The same is true of $\text{GCV}_c^{(1)}$ when the variance is low and C is estimated. This is somewhat surprising given FCCV's simplistic approach of omitting neighborhoods about the point of estimation. In cases where the correlation structure proves difficult to estimate, FCCV is an excellent alternative to GCV_c .

4. APPLICATION

Although the covariance matrix for an empirical semivariogram is heteroscedastic, there is an intimate relationship between the covariance structure of the semivariogram and the

semivariogram itself that can be accommodated by our proposed method. Parametric semivariograms are frequently fit using weighted least squares (Cressie, 1985) by

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{\ell} \frac{N_{s_i}}{\hat{\gamma}(s_i | \boldsymbol{\theta})^2} (\tilde{\gamma}(s_i) - \hat{\gamma}(s_i | \boldsymbol{\theta}))^2 \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{\ell} N_{s_i} \left(\frac{\tilde{\gamma}(s_i)}{\hat{\gamma}(s_i | \boldsymbol{\theta})} - 1 \right)^2,\end{aligned}\tag{10}$$

since $\operatorname{Var}[\tilde{\gamma}(s)] \approx 2(\gamma(s))^2/N_s$, where $\gamma(\cdot)$ is the true semivariogram, $\tilde{\gamma}(\cdot)$ is the empirical semivariogram calculated from data, $\hat{\gamma}(\cdot|\boldsymbol{\theta})$ is a parametric semivariogram fit, N_s is the number of spatial locations s units apart, and ℓ is the number of lags used in the fitting process. Since each empirical semivariogram value has been divided by an estimate of its standard deviation, the covariance matrix of the ratios can now be treated as a correlation matrix. Furthermore, the semivariogram provides an estimate of the covariance of the data used to calculate the empirical semivariogram, which allows us to estimate the correlation matrix of the ratios $\tilde{\gamma}(s_i)/\hat{\gamma}(s_i|\boldsymbol{\theta})$, $i = 1, \dots, \ell$ (Genton, 1998).

In the application that follows, we fit a semiparametric semivariogram based on the method presented by Carmack et al. (2011) replacing their fitting criterion using a modified version of (10).

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{\ell} N_{s_i} \left(\frac{\frac{\tilde{\gamma}(s_i)}{\hat{\gamma}(s_i|\boldsymbol{\theta})} - 1}{1 - \operatorname{tr}[2S\hat{C} - S\hat{C}S']/\ell}} \right)^2,\tag{11}$$

where \hat{C} is the estimated correlation matrix of $\tilde{\gamma}(s_i)/\hat{\gamma}(s_i|\boldsymbol{\theta})$, $i = 1, \dots, \ell$, using a pilot estimate for $\hat{\gamma}(\cdot|\boldsymbol{\theta})$, and $\boldsymbol{\theta}' = [\kappa, \alpha]$. κ is the order of the basis with lower orders being more flexible, while α controls how the basis approaches the origin, which has an important impact on nugget estimation. The pilot fit used $\kappa = 11$, and $\alpha = 1$. In their paper, Carmack et al. (2011) estimated α using a custom fitting criterion, but fixed $\kappa = 11$ due to difficulties with establishing an objective function that could satisfactorily accommodate κ and α simultaneously, which is likely due to the strong correlation inherent in the empirical semivariogram.

Our primary interest is analyzing brain imaging data where we routinely use spatial modeling (kriging) for statistical inference (Spence et al., 2007). The particular example here deals with functional magnetic resonance imaging (fMRI) where a subject is placed in a magnet to perform an experiment. The magnet records changes in blood oxygen level dependent (BOLD) signals at thousands of locations across the brain with three dimensional volumes captured every few seconds. The temporal aspect of the data is usually removed

through a variety of statistical modeling methods (Lindquist, 2008) where practitioners are commonly interested in estimating the hemodynamic response function (HRF) to identify locations associated with the experimental protocol or in extracting features of the HRF at active locations.

The experiment in the example had the subject silently repeat nonsense words displayed on a monitor above their head for a total of 152 scans spaced 2 sec. apart. We then estimated a 13 parameter finite impulse response function (FIR) under a linear convolution invariance assumption with the parameters spaced 2 sec. apart to match the temporal resolution of the scans and the maximum duration of the HRF after a stimulus is applied (26 sec.). This was done at 1,557 spatial locations that comprise the left superior temporal gyrus, a portion of the brain thought to be associated with the protocol. For a healthy subject, the peak in the HRF generally occurs approximately 6 sec. after a stimulus. Hence, we will concern ourselves with the third FIR parameter at these 1,557 locations.

As Fig. 6 shows, the empirical semivariogram of the third FIR parameter exhibits spatial correlation to approximately 9 mm. Casual inspection of the empirical semivariogram suggests a linear approach to the origin is reasonable with the exponential or spherical parametric models being natural choices given their linear behavior towards the origin. A semiparametric fit was estimated using (11) in `optim` in R with boundary conditions $3 \leq \kappa \leq 25$, and $0 \leq \alpha \leq 1$. The lower bound on κ is necessary since this is a three dimensional spatial process, while the upper limit was set at 25 since the basis does not substantially change beyond that value. The bounds on α are established in Carmack et al. (2011). The optimization yielded $\hat{\kappa} = 20.6$ and $\hat{\alpha} = 0.709$ with the resulting semiparametric fit along with the parametric exponential and spherical fits shown in Fig. 6. The nugget estimate, which plays a critical role in kriging, is 15% of the estimated sill compared to the nugget estimated at 9% of the estimated sill in Carmack et al. (2011). The exponential and spherical fits produced nugget estimates of 0% and 20% of their respective estimated sills. The exponential is clearly a poor fit overshooting the early lags and failing to level out at later lags. While the spherical arguably does better at the early lags, it appears to overestimate the range.

5. DISCUSSION

As the theory section established, a natural definition for residual degrees of freedom is $\text{tr}[C + SCS' - 2SC]$, which suggests an extension of GCV for correlated data as defined by $GCV_c^{(1)}$. Historical consideration of three competing definitions for residual degrees of freedom and their correlated counterparts led us to define $GCV_c^{(2)}$ and $GCV_c^{(3)}$. As the simulation study showed, when the correlation structure is known, $GCV_c^{(1)}$ tends to dominate the other two. However, this is not universally the case when estimating correlation. When the variance is low, $GCV_c^{(1)}$ can lead to over smoothing since the bias due to the underlying function in the empirical semivariogram becomes large relative to the variance,

which leads to overinflated correlation estimates and over smoothing. But, for medium to high variances, it still tends to dominate the other two leading us to conclude that $\text{GCV}_c^{(1)}$ is the best choice for most situations. As such, we will refer to $\text{GCV}_c^{(1)}$ more simply as GCV_c for the rest of the discussion.

Given its surprisingly simplistic approach, FCCV performed admirably and should still be considered, particularly if the correlation structure is difficult to estimate. Finally, the fMRI application demonstrated how GCV_c can be used in a heteroscedastic setting where an intimate link exists between the function being estimated and the correlation in the context of spatial modeling.

It is important to note that we did not use a general covariance structure. In theory, one could develop a method similar to the one presented in the application for heteroscedastic errors by rescaling the data so that the resulting covariance matrix is a correlation matrix by using a criterion like (11). In practice, this presents a difficult challenge since σ_k^2 has to be estimated at each location in the presence of an unknown mean structure and correlated errors. This task is not to be taken lightly given the difficulty of doing so even when errors are independent. Even so, we intend to continue researching this difficult problem in the hopes of obtaining a viable solution in the future.

It is interesting to note that $\text{E}[\sum_{k=1}^n (\hat{f}(x_k) - y_k)^2] / \sigma^2$ matches the first moment of a non-central χ^2 on $\text{tr}[C + SCS' - 2SC]$ degrees of freedom. Given this observation, one might wonder why GCV_c divides by the square of the trace. Assuming that the errors are distributed multivariate normal, $\varepsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2 C)$, the variance is given by

$$\begin{aligned} \text{Var} \left[\sum_{k=1}^n (\hat{f}(x_k) - y_k)^2 \right] &= 4\sigma^2 \sum_{k=1}^n \sum_{\ell=1}^n \mu_k \mu_\ell R_{k\ell} + 2\sigma^4 \sum_{k=1}^n \sum_{\ell=1}^n R_{k\ell}^2 \\ &= 4\sigma^2 \sum_{k=1}^n \sum_{\ell=1}^n \mu_k \mu_\ell R_{k\ell} + 2\sigma^4 \text{tr} \left[(C + SCS' - SC - CS')^2 \right], \end{aligned}$$

which is generally not proportional to the variance of the aforementioned non-central χ^2 , unless $C + SCS' - SC - CS'$ happens to be idempotent. Furthermore, the trace term is not the divisor in GCV_c . Letting $V = C + SCS' - SC - CS'$, a closer inspection reveals that

$$(\text{tr}[V])^2 = \text{tr}[V \otimes V],$$

which is proportional to the trace of the covariance of the Wishart distribution whose matrix parameter is V . Thus, the divisor in GCV_c can be thought of as accounting for the total variance of V and not just that of the diagonal elements since $(\text{tr}[V])^2 = (\sum_{i=1}^n \lambda_i)^2$

and $\text{tr}[V^2] = \sum_{i=1}^n \lambda_i^2$, where $\lambda_i, i = 1, \dots, n$, are the eigenvalues of V .

Continuing with normality, the characteristic function of $\sum_{k=1}^n (\hat{f}(x_k) - y_k)^2$ when all the non-centrality parameters are zero is given by (Krishnaiah, 1961)

$$\psi(t) = \prod_{j=1}^n (1 - it2\sigma^2\lambda_j)^{-\frac{1}{2}}.$$

This is a convolution of gammas with common shape parameter $\alpha = -1/2$ and scale parameters $\beta_j = 2\sigma^2\lambda_j, j = 1, \dots, n$, which is a kind of generalization of the χ^2 . Several methods exist for computing the distribution of convolutions of gammas with truncated series or Monte Carlo methods, but a more simplistic approach in the spirit of Satterthwaite is an approximation by a single gamma with $\alpha = (\sum_{j=1}^n \lambda_j)^2 / (2 \sum_{j=1}^n \lambda_j^2)$ and $\beta = 2\sigma^2 \sum_{j=1}^n \lambda_j^2 / \sum_{j=1}^n \lambda_j$ (Stewart et al., 2007), which we found to work well with low to moderately correlated data with adequate sample sizes. In this framework, our proposed cross-validation criterion could be treated as estimating θ via maximum likelihood. We opted not to present this approach at this time since we are striving to keep the present method as nonparametric as possible, but this remains a promising avenue of research that intend to pursue.

REFERENCES

- Buja, A., Hastie, T., & Tibsirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2), 453–510.
- Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2), 351–358.
- Carmack, P., Spence, J., Schucany, W., Gunst, R., Lin, Q., & Haley, R. (2009). Far casting cross validation. *Journal of Computational and Graphical Statistics*, 18(4), 879–893.
- Carmack, P., Spence, J., Schucany, W., Gunst, R., Lin, Q., & Haley, R. (2011). A new class of semiparametric semivariogram and nugget estimators. *Computational Statistics and Data Analysis*, ?(?), ?–?
- Cherry, S., Banfield, J., & Quimby, W. (1996). An evaluation of a nonparametric method of estimating semivariograms of isotropic spatial processes. *Journal of Applied Statistics*, 23(4), 435–449.
- Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics*, 31, 377–403.
- Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17(5), 563–586.
- Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87, 998–1004.
- Geisser, S. (1975). A predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70, 320–328.

- Genton, M. (1998). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology*, 30(4), 323–345.
- Hart, J. & Lee, C. (2005). Robustness of one-sided cross-validation to autocorrelation. *Journal of Multivariate Analysis*, 92(1), 77–96.
- Hart, J. & Yi, S. (1998). One-sided cross-validation. *Journal of the American Statistical Association*, 93(442), 620–630.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer-Verlag.
- Krishnaiah, P. (1961). Remarks on a multivariate gamma distribution. *The American Mathematical Monthly*, 68(4), 342–346.
- Lindquist, M. (2008). The statistical analysis of fmri data. *Statistical Science*, 23(4), 439–464.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Racine, J. (2000). Consistent cross-validators model-selection for dependent data: *h_v*-block cross-validation. *Journal of Econometrics*, 99, 39–61.
- Spence, J., Carmack, P., Gunst, R., Schucany, W., Woodward, W., & Haley, R. (2007). Accounting for spatial dependence in the analysis of spect brain imaging data. *Journal of the American Statistical Association*, 102(478), 464–473.
- Stewart, T., Strijbosch, L., Moors, H., & Batenburg, P. V. (2007). A simple approximation to the convolution of gamma distributions. In *CentER Discussion Paper No. 2007-70*.
- Stone, M. (1974). Cross-validators choice and the assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, B* 36, 111–133.

TABLES

TABLE 1. $f_1(\cdot)$

ϕ	n	σ	$\overline{\hat{h}_1/h_0}$		$\overline{\hat{h}_2/h_0}$		$\overline{\hat{h}_3/h_0}$		$\overline{\hat{h}_{f_{ccv}}/h_0}$	
			C	\hat{C}	C	\hat{C}	C	\hat{C}		
0.0	75	2^{-11}	1.15	1.94	0.97	1.75	0.69	1.55	1.49	
		2^{-9}	1.18	1.53	1.02	1.40	0.76	1.23	1.33	
		2^{-7}	1.26	1.38	1.21	1.09	0.99	0.84	1.34	
	150	2^{-11}	1.12	1.69	0.98	1.56	0.67	1.42	1.20	
		2^{-9}	1.15	1.28	1.00	1.16	0.74	0.98	1.13	
		2^{-7}	1.24	1.31	1.08	1.16	0.85	0.95	1.16	
	0.3	75	2^{-11}	1.19	1.74	0.99	1.54	0.71	1.34	1.35
			2^{-9}	1.19	1.44	1.02	1.29	0.72	1.09	1.25
			2^{-7}	1.34	1.35	1.15	1.16	0.89	0.91	1.34
150		2^{-11}	1.15	1.53	0.99	1.40	0.71	1.23	1.10	
		2^{-9}	1.20	1.26	1.04	1.12	0.76	0.90	1.09	
		2^{-7}	1.28	1.26	1.12	1.10	0.88	0.86	1.16	
0.6		75	2^{-11}	1.20	1.51	0.99	1.29	0.78	1.06	1.15
			2^{-9}	1.27	1.32	1.06	1.14	0.81	0.89	1.06
			2^{-7}	1.46	1.18	1.21	0.95	0.91	0.68	1.20
	150	2^{-11}	1.18	1.28	1.00	1.13	0.77	0.91	0.90	
		2^{-9}	1.18	1.04	1.00	0.87	0.76	0.6	0.84	
		2^{-7}	1.32	1.03	1.13	0.83	0.87	0.55	0.97	
	0.9	75	2^{-11}	1.09	1.61	0.93	1.22	0.83	1.00	1.24
			2^{-9}	1.32	1.22	1.13	1.02	1.03	0.81	0.89
			2^{-7}	1.56	0.82	1.24	0.64	1.03	0.45	0.67
150		2^{-11}	1.21	1.14	1.06	0.96	0.98	0.77	0.73	
		2^{-9}	1.29	0.78	1.08	0.63	0.94	0.46	0.49	
		2^{-7}	1.41	0.46	1.16	0.33	0.93	0.21	0.34	

FIGURES

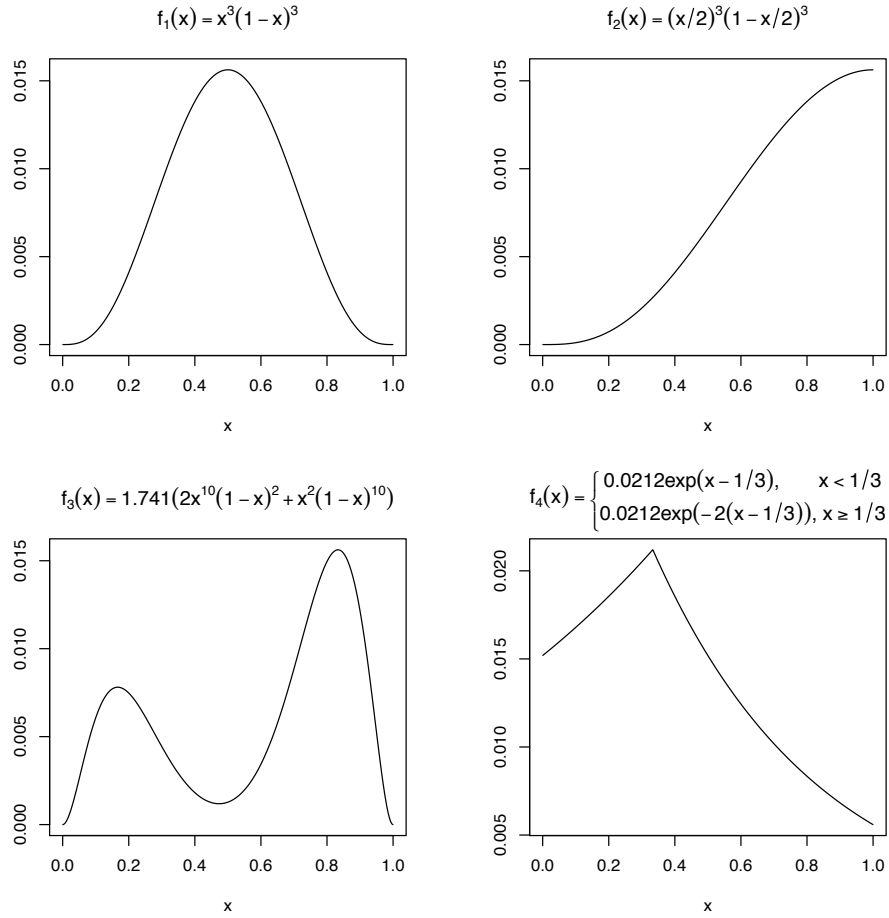


FIGURE 1. Plot of the four functions used in the simulation study. Additive errors of varying correlation and variance were added to each function at either $n = 75$ or $n = 150$ equally spaced design points in the interval $[0, 1]$. A global bandwidth for local linear regression was estimated by several competing cross-validation methods.

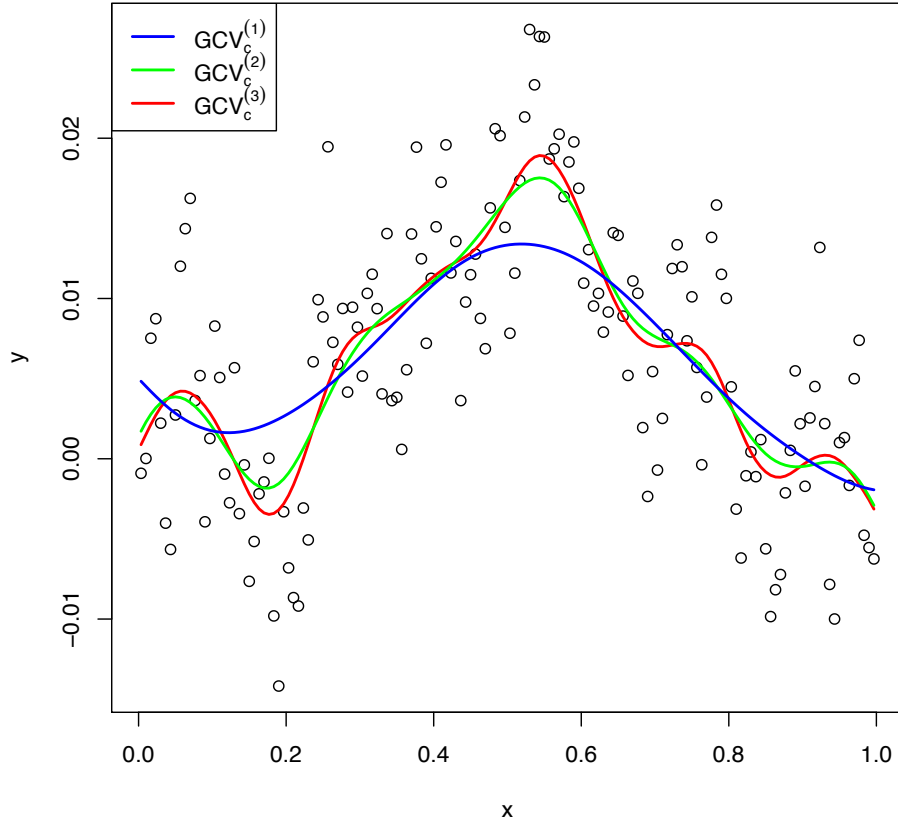


FIGURE 2. Three local linear regression fits with bandwidths estimated by the three GCV_c criteria using the known correlation structure. \hat{h} was estimated to be 0.337, 0.114, and 0.110 for $GCV_c^{(1)}$, $GCV_c^{(2)}$, and $GCV_c^{(3)}$, respectively. The sample was generated using $f(\cdot) = f_1(\cdot)$, $n = 150$, $\phi = 0.6$, and $\sigma = 1/128$.

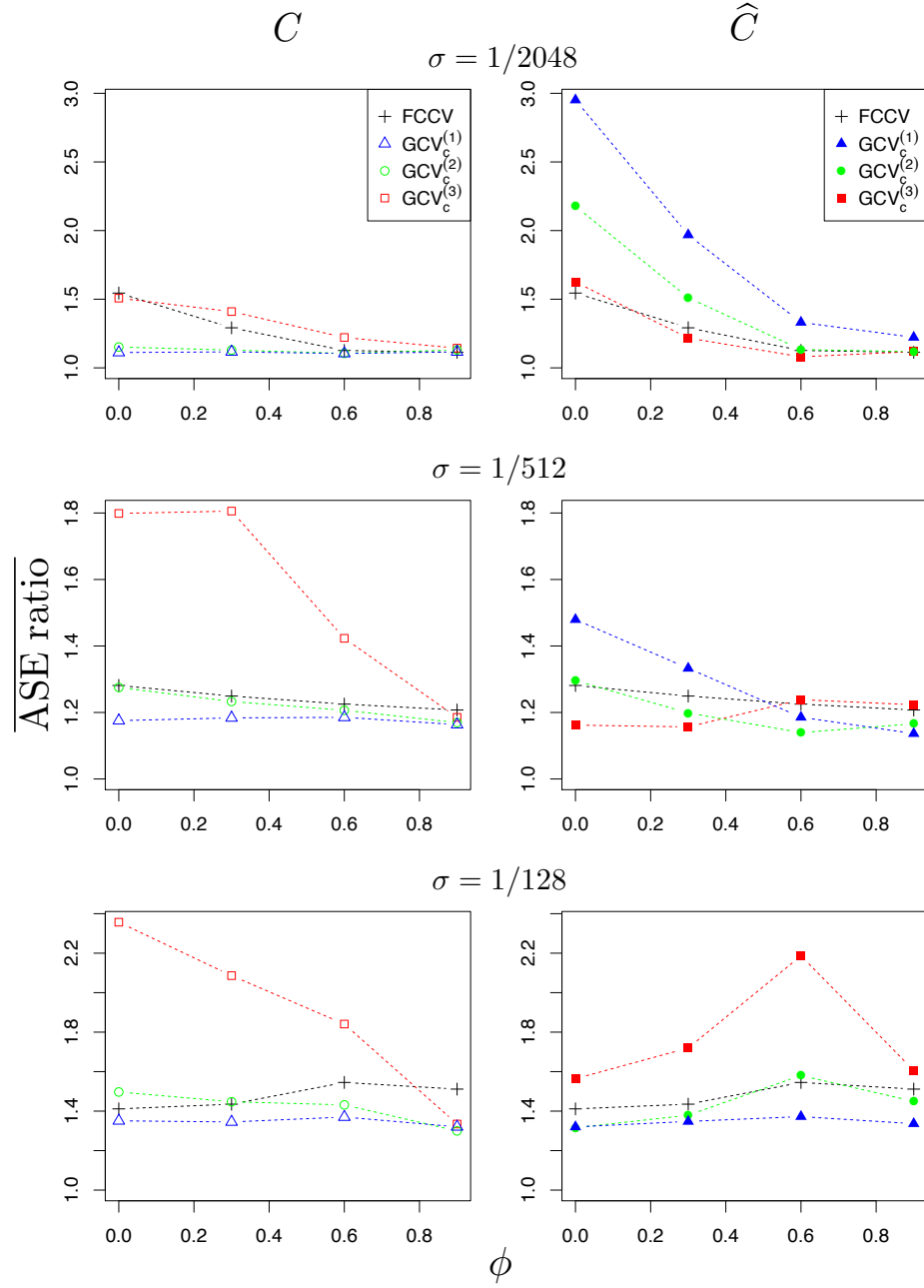


FIGURE 3. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_1(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

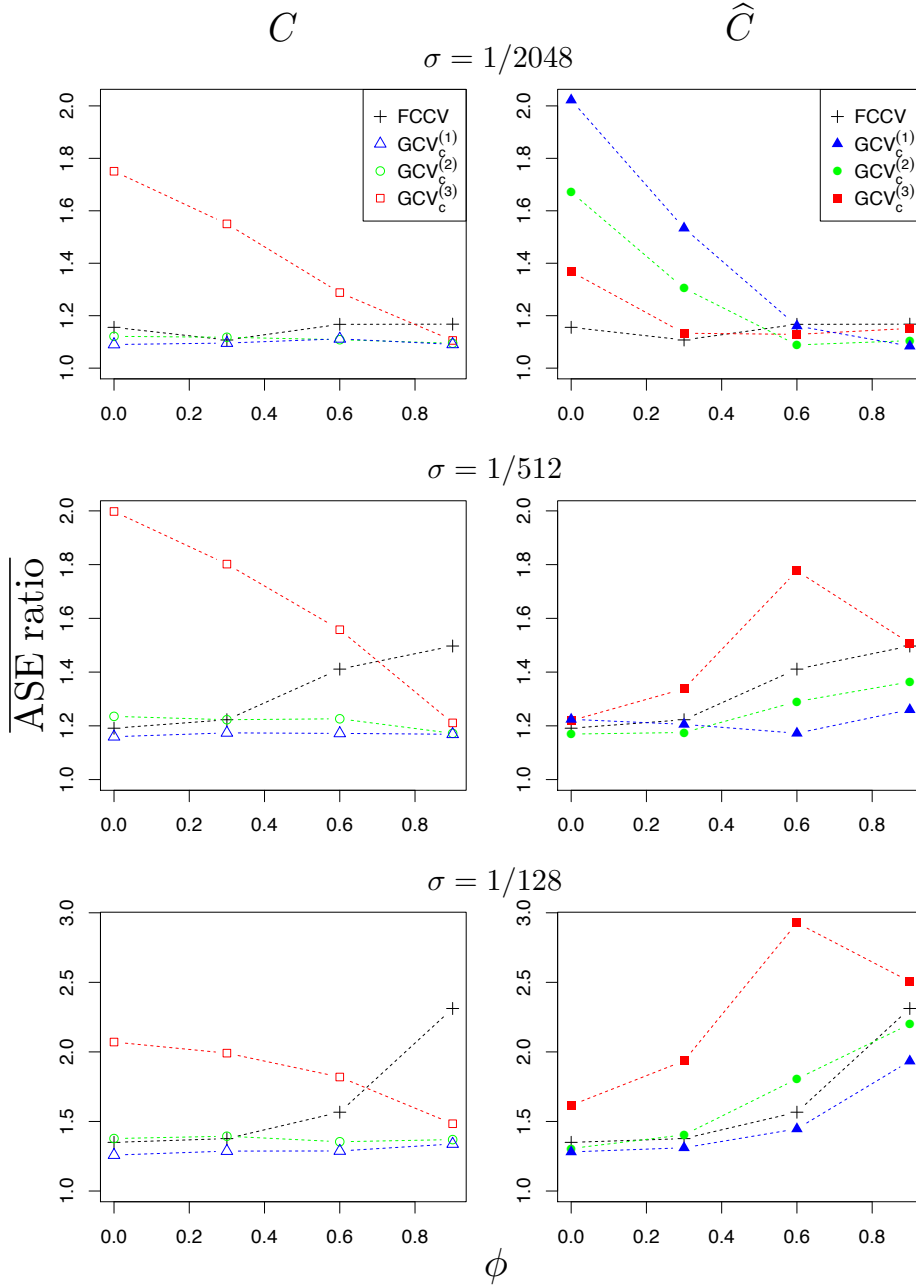


FIGURE 4. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_1(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

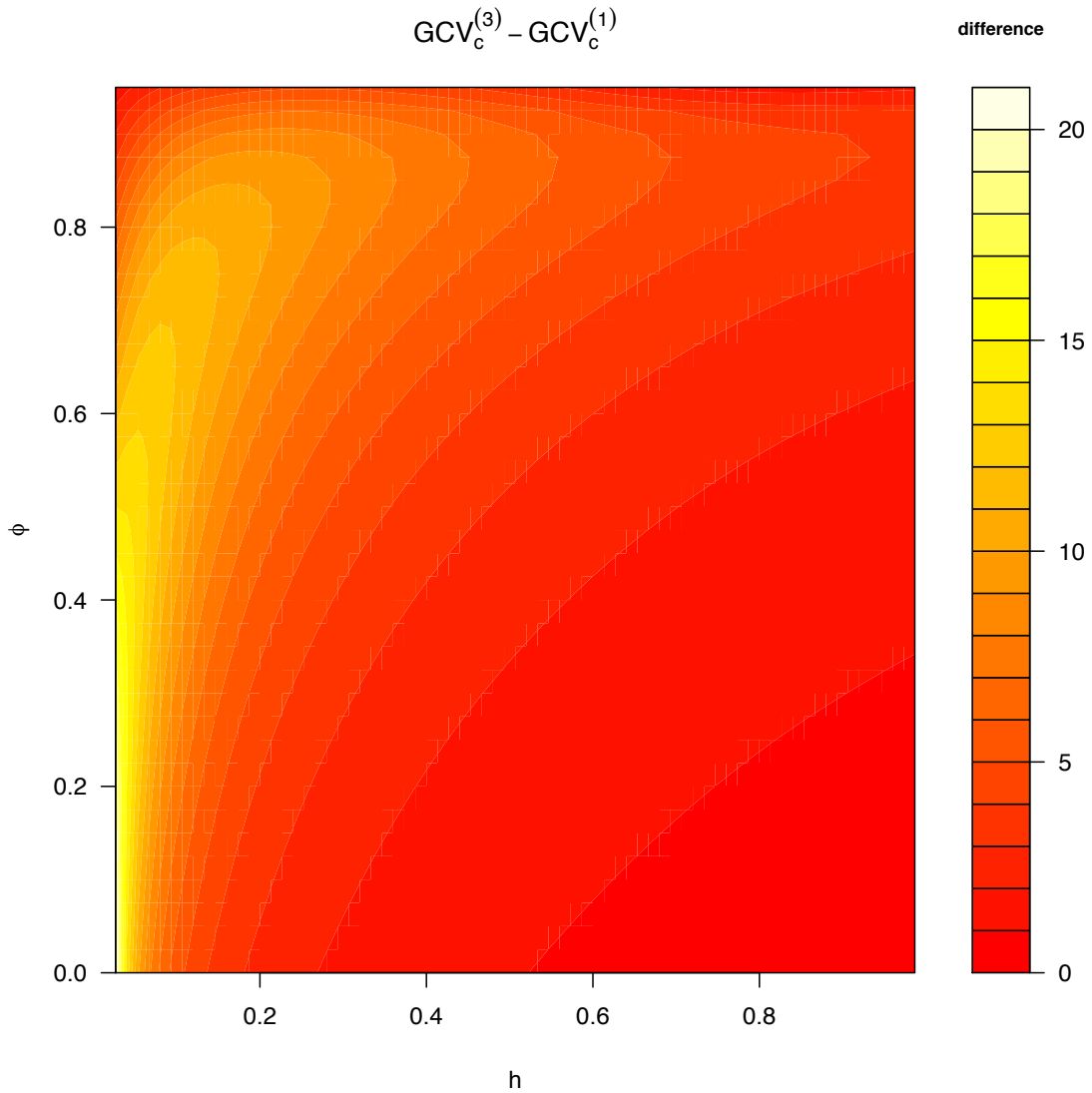


FIGURE 5. A filled contour plot of the difference in residual degrees of freedom as defined by $\text{GCV}_c^{(3)}$ and $\text{GCV}_c^{(1)}$ in the context of local linear regression with bandwidth h , $n = 75$ equally spaced design points, and known correlation structure given by an AR(1) process with coefficient ϕ . The largest differences occur for smaller bandwidths and/or higher correlation.

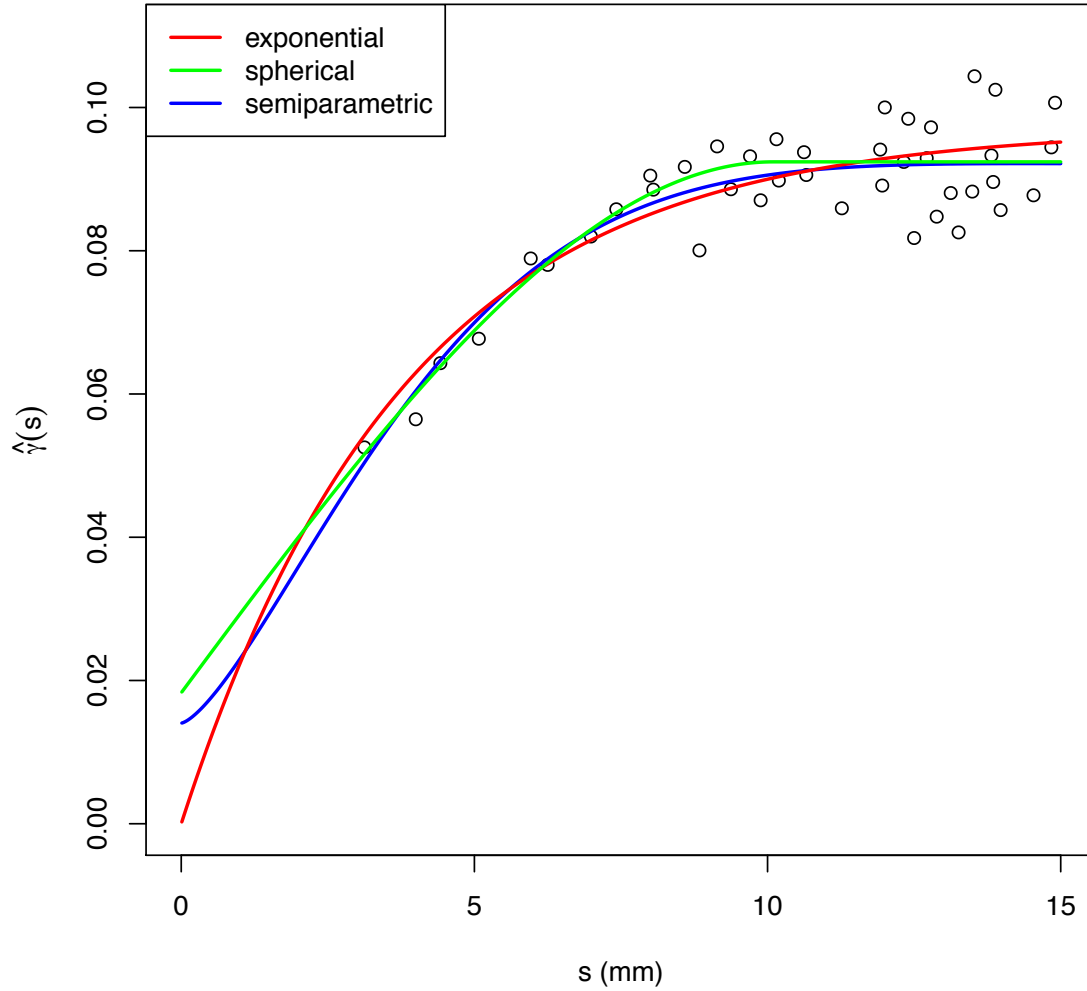


FIGURE 6. Plot of the empirical semivariogram of the third FIR parameter from the left superior temporal gyrus of a single subject in an fMRI experiment with three semivariogram fits. The exponential fit (red) overshoots the early lags and fails to level out at later lags. The spherical fit (green) overestimates the range of the correlation. The semiparametric fit (blue) was obtained using a modified form of $GCV_c^{(1)}$.

APPENDIX

TABLE 2. $f_2(\cdot)$

ϕ	n	σ	$\overline{\hat{h}_1/h_0}$		$\overline{\hat{h}_2/h_0}$		$\overline{\hat{h}_3/h_0}$		$\overline{\hat{h}_{fccv}/h_0}$
			C	\hat{C}	C	\hat{C}	C	\hat{C}	
0.0	75	2^{-11}	1.18	1.84	1.02	1.69	0.74	1.50	1.33
		2^{-9}	1.45	1.74	1.23	1.49	0.94	1.20	1.45
		2^{-7}	1.40	1.46	1.28	1.36	1.09	1.19	1.24
	150	2^{-11}	1.15	1.49	1.00	1.38	0.75	1.23	1.13
		2^{-9}	1.29	1.39	1.12	1.22	0.86	0.99	1.20
		2^{-7}	1.55	1.61	1.39	1.46	1.15	1.23	1.36
0.3	75	2^{-11}	1.19	1.64	1.02	1.48	0.75	1.28	1.23
		2^{-9}	1.65	1.79	1.38	1.51	1.03	1.16	1.49
		2^{-7}	1.37	1.34	1.25	1.22	1.03	1.00	1.18
	150	2^{-11}	1.15	1.37	1.00	1.24	0.76	1.06	1.06
		2^{-9}	1.39	1.38	1.18	1.18	0.91	0.91	1.21
		2^{-7}	1.45	1.41	1.31	1.28	1.09	1.06	1.25
0.6	75	2^{-11}	1.24	1.35	1.03	1.17	0.79	0.93	1.03
		2^{-9}	1.81	1.53	1.48	1.25	1.10	0.88	1.31
		2^{-7}	1.37	1.17	1.24	1.00	1.03	0.76	1.07
	150	2^{-11}	1.20	1.17	1.02	1.00	0.76	0.75	0.85
		2^{-9}	1.54	1.19	1.29	0.95	0.96	0.65	1.07
		2^{-7}	1.37	1.13	1.25	0.95	1.02	0.71	1.02
0.9	75	2^{-11}	1.34	1.23	1.15	1.02	1.05	0.81	0.89
		2^{-9}	2.17	0.92	1.53	0.72	1.18	0.51	0.72
		2^{-7}	1.61	0.63	1.41	0.46	1.09	0.28	0.60
	150	2^{-11}	1.27	0.76	1.07	0.61	0.93	0.45	0.48
		2^{-9}	1.78	0.51	1.37	0.37	1.01	0.23	0.35
		2^{-7}	1.57	0.41	1.36	0.27	1.07	0.16	0.39

TABLE 3. $f_3(\cdot)$

ϕ	n	σ	$\overline{\hat{h}_1/h_0}$	\widehat{h}_1/h_0	$\overline{\hat{h}_2/h_0}$	\widehat{h}_2/h_0	$\overline{\hat{h}_3/h_0}$	\widehat{h}_3/h_0	$\overline{\hat{h}_{f_{ccv}}/h_0}$	
			C	\widehat{C}	C	\widehat{C}	C	\widehat{C}		
0.0	75	2^{-11}	1.18	2.16	0.98	1.87	0.77	1.63	1.69	
		2^{-9}	1.16	1.75	0.98	1.57	0.69	1.38	1.45	
		2^{-7}	1.43	1.68	1.18	1.40	0.84	1.09	1.58	
	150	2^{-11}	1.16	1.95	0.99	1.77	0.63	1.59	1.37	
		2^{-9}	1.15	1.39	1.01	1.28	0.72	1.11	1.20	
		2^{-7}	1.26	1.36	1.08	1.19	0.80	0.93	1.20	
	0.3	75	2^{-11}	1.20	1.95	0.97	1.65	0.75	1.40	1.47
			2^{-9}	1.22	1.66	1.02	1.45	0.73	1.23	1.33
			2^{-7}	1.76	1.80	1.39	1.47	0.96	1.05	1.71
150		2^{-11}	1.17	1.69	0.98	1.51	0.66	1.32	1.20	
		2^{-9}	1.17	1.33	1.01	1.19	0.72	1.00	1.13	
		2^{-7}	1.35	1.36	1.14	1.14	0.83	0.84	1.19	
0.6		75	2^{-11}	1.24	1.87	0.99	1.50	0.81	1.23	1.32
			2^{-9}	1.24	1.48	1.02	1.26	0.79	1.01	1.11
			2^{-7}	2.06	1.49	1.55	1.14	1.04	0.76	1.49
	150	2^{-11}	1.21	1.45	1.00	1.24	0.78	1.02	0.98	
		2^{-9}	1.21	1.18	1.03	1.01	0.78	0.77	0.90	
		2^{-7}	1.56	1.14	1.23	0.88	0.87	0.55	1.00	
	0.9	75	2^{-11}	0.99	1.84	0.96	1.12	0.96	0.96	1.60
			2^{-9}	1.25	1.58	1.07	1.25	0.96	1.01	1.11
			2^{-7}	3.12	1.05	1.52	0.85	1.10	0.65	0.78
150		2^{-11}	1.22	1.65	1.02	1.27	0.90	1.01	1.14	
		2^{-9}	1.28	1.03	1.09	0.85	0.99	0.67	0.66	
		2^{-7}	1.99	0.62	1.28	0.47	0.98	0.32	0.41	

TABLE 4. $f_4(\cdot)$

ϕ	n	σ	$\overline{\hat{h}_1/h_0}$	$\overline{\hat{h}_2/h_0}$	$\overline{\hat{h}_3/h_0}$	$\overline{\hat{h}_{fccv}/h_0}$	C	\hat{C}	C	\hat{C}	C	\hat{C}	
0.0	75	2^{-11}	1.20	2.14	0.99	1.91	0.69	1.67	1.67				
		2^{-9}	1.21	1.45	1.03	1.29	0.74	1.07	1.43				
		2^{-7}	1.80	1.96	1.55	1.71	1.22	1.39	1.64				
	150	2^{-11}	1.19	1.72	1.01	1.57	0.68	1.38	1.31				
		2^{-9}	1.22	1.32	1.05	1.17	0.78	0.93	1.19				
		2^{-7}	1.52	1.63	1.29	1.40	0.99	1.10	1.38				
	0.3	75	2^{-11}	1.26	1.93	1.03	1.69	0.71	1.43	1.50			
			2^{-9}	1.29	1.43	1.09	1.25	0.79	1.00	1.38			
			2^{-7}	1.72	1.69	1.47	1.45	1.13	1.13	1.48			
150		2^{-11}	1.21	1.55	1.02	1.39	0.70	1.17	1.16				
		2^{-9}	1.26	1.28	1.08	1.10	0.79	0.84	1.16				
		2^{-7}	1.78	1.72	1.53	1.47	1.17	1.13	1.52				
0.6		75	2^{-11}	1.32	1.65	1.05	1.38	0.78	1.09	1.26			
			2^{-9}	1.38	1.27	1.14	1.06	0.85	0.77	1.19			
			2^{-7}	1.66	1.31	1.43	1.06	1.10	0.75	1.20			
	150	2^{-11}	1.26	1.35	1.05	1.15	0.77	0.90	0.94				
		2^{-9}	1.30	1.08	1.11	0.86	0.82	0.56	0.94				
		2^{-7}	1.82	1.36	1.54	1.10	1.18	0.76	1.28				
	0.9	75	2^{-11}	1.31	1.97	1.01	1.43	0.88	1.09	1.35			
			2^{-9}	1.60	1.15	1.27	0.92	1.09	0.70	0.84			
			2^{-7}	2.33	0.80	1.78	0.60	1.31	0.40	0.78			
150		2^{-11}	1.42	1.22	1.14	0.98	1.01	0.76	0.75				
		2^{-9}	1.45	0.67	1.15	0.52	0.93	0.37	0.42				
		2^{-7}	2.07	0.45	1.64	0.31	1.21	0.18	0.41				

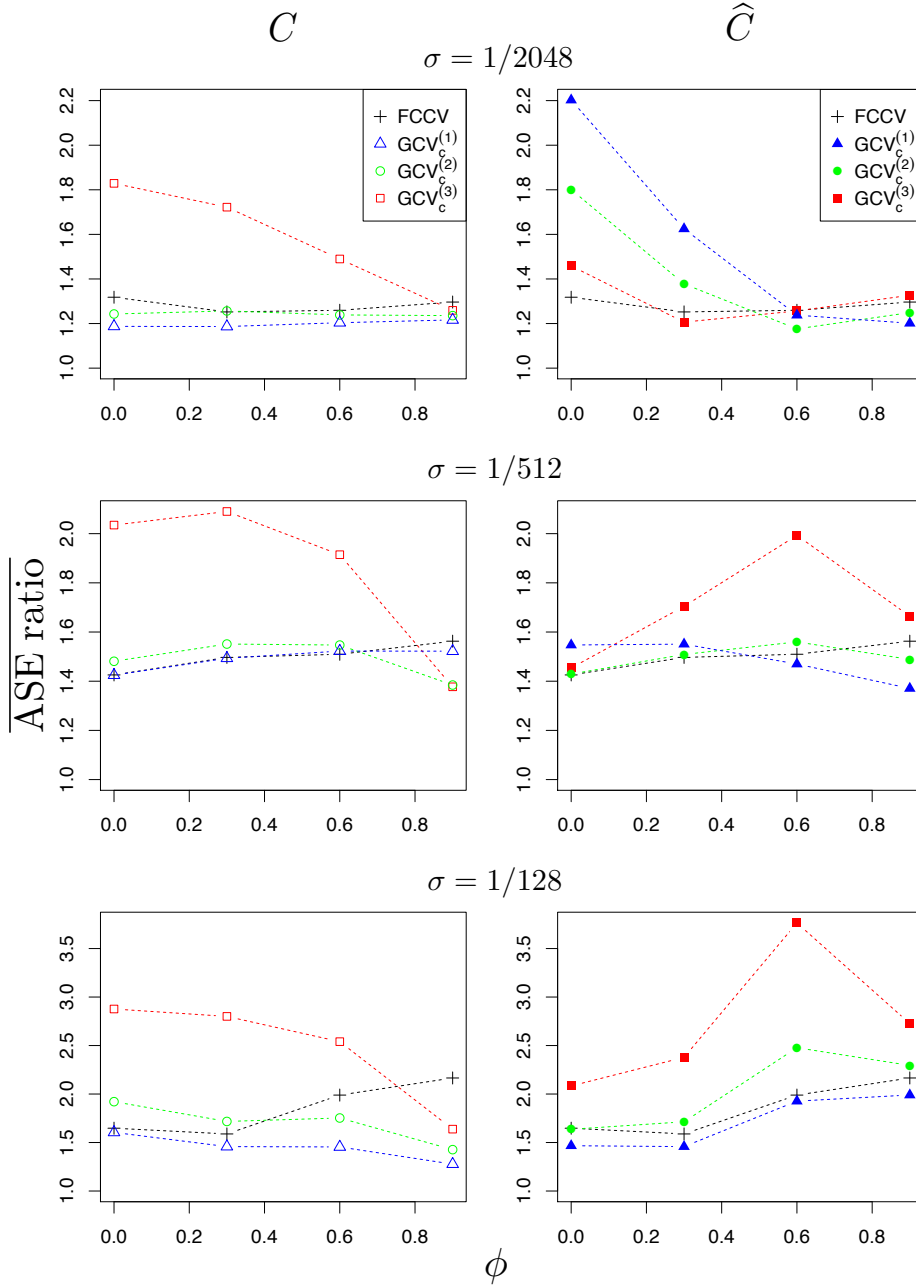


FIGURE 7. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_2(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

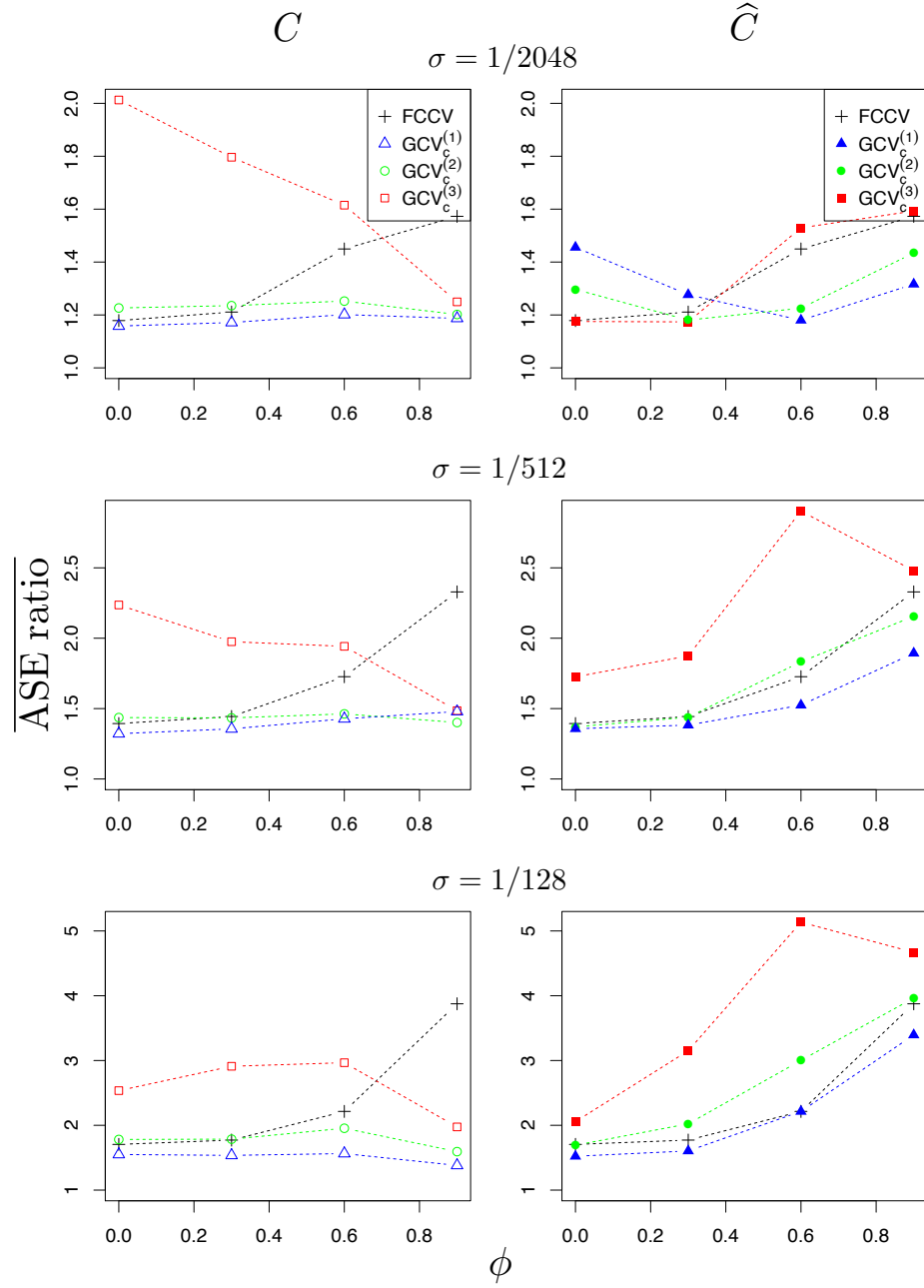


FIGURE 8. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_2(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

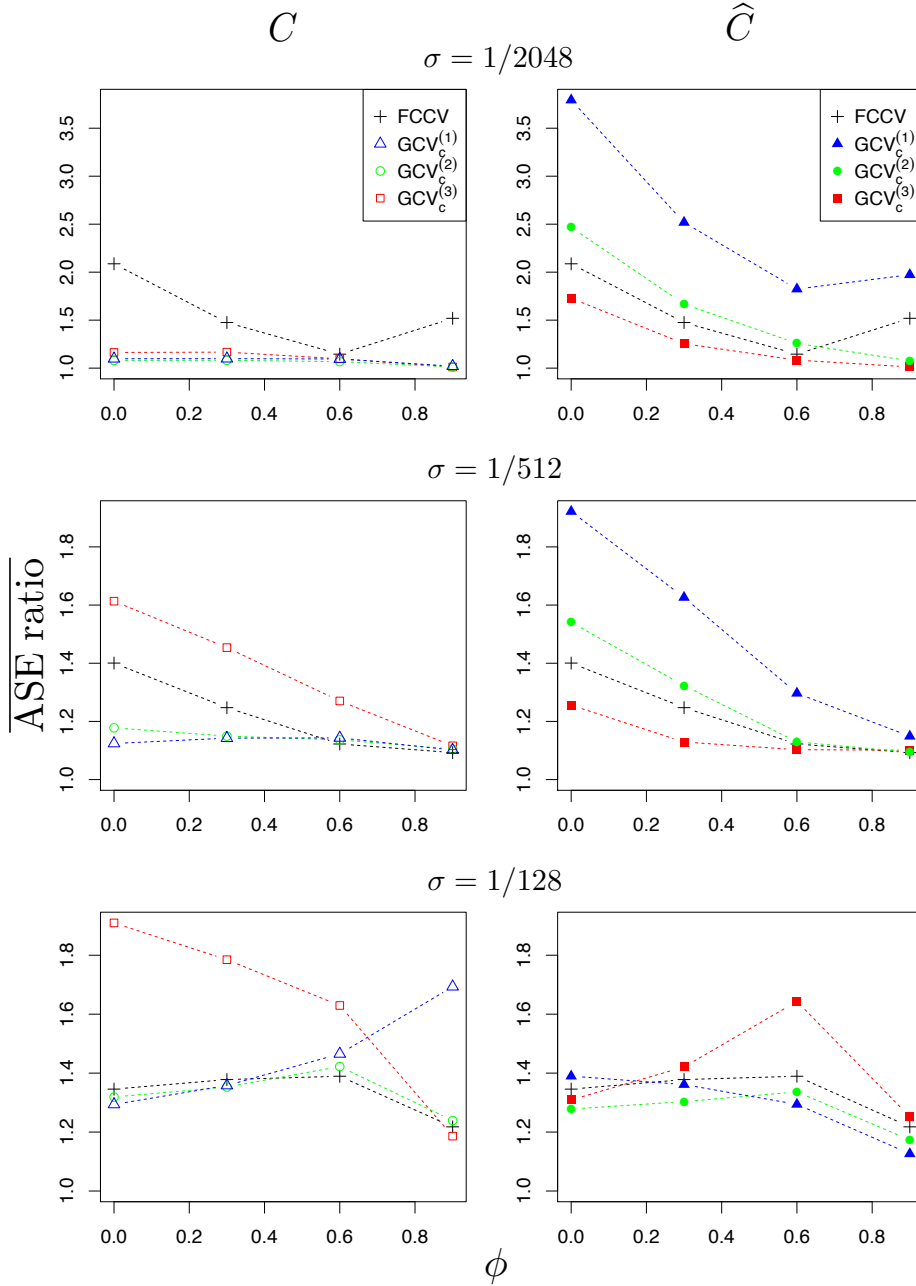


FIGURE 9. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_3(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

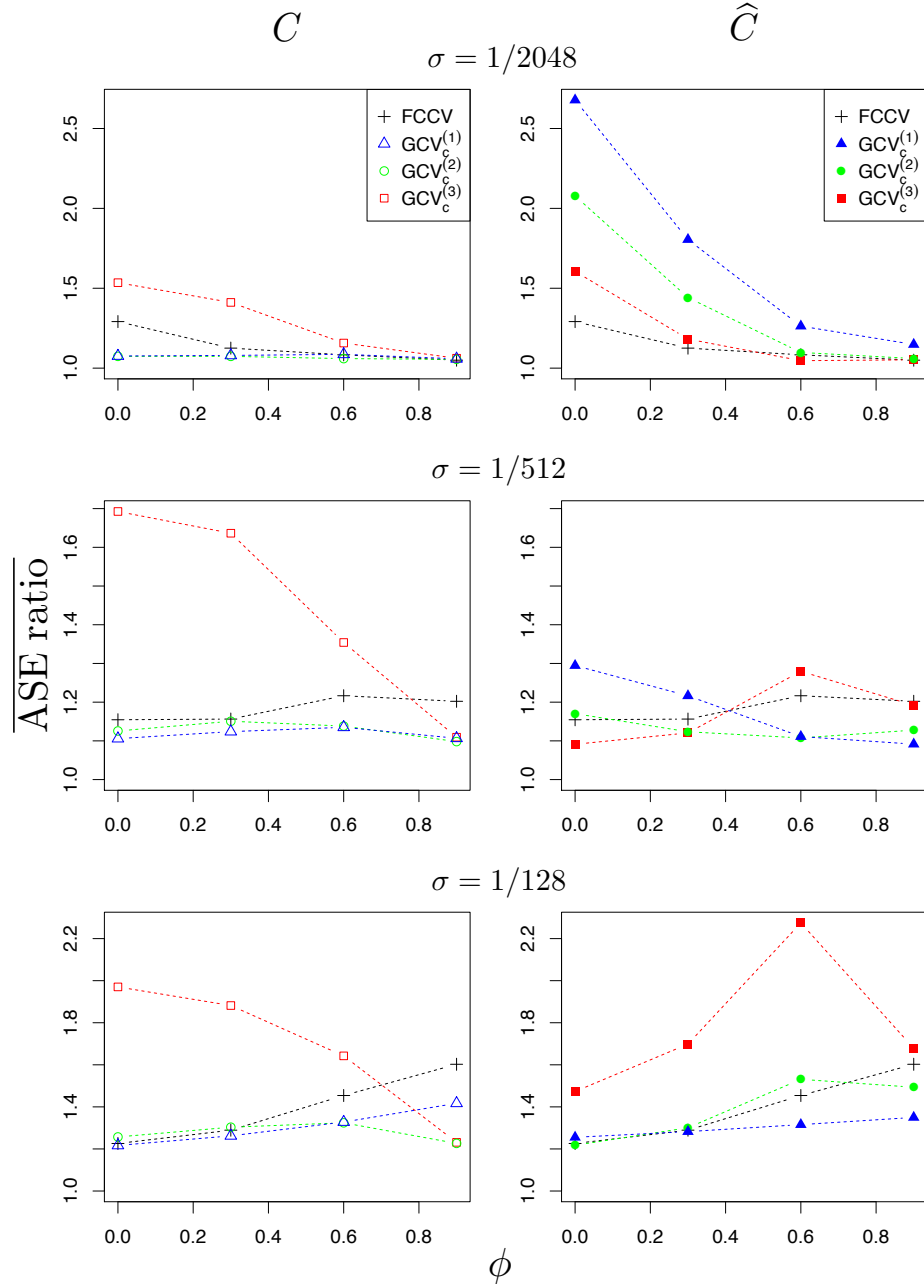


FIGURE 10. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_3(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

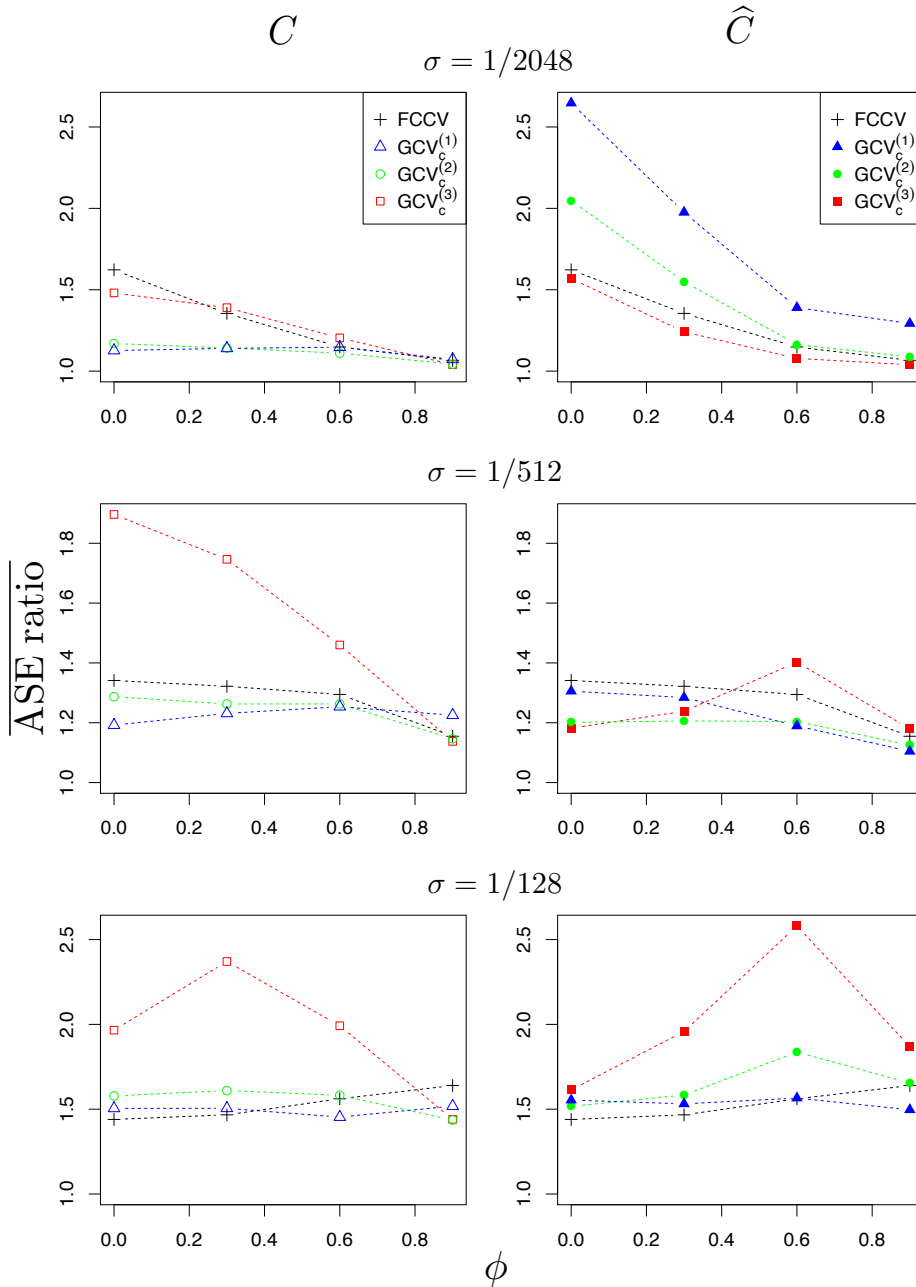


FIGURE 11. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_4(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .

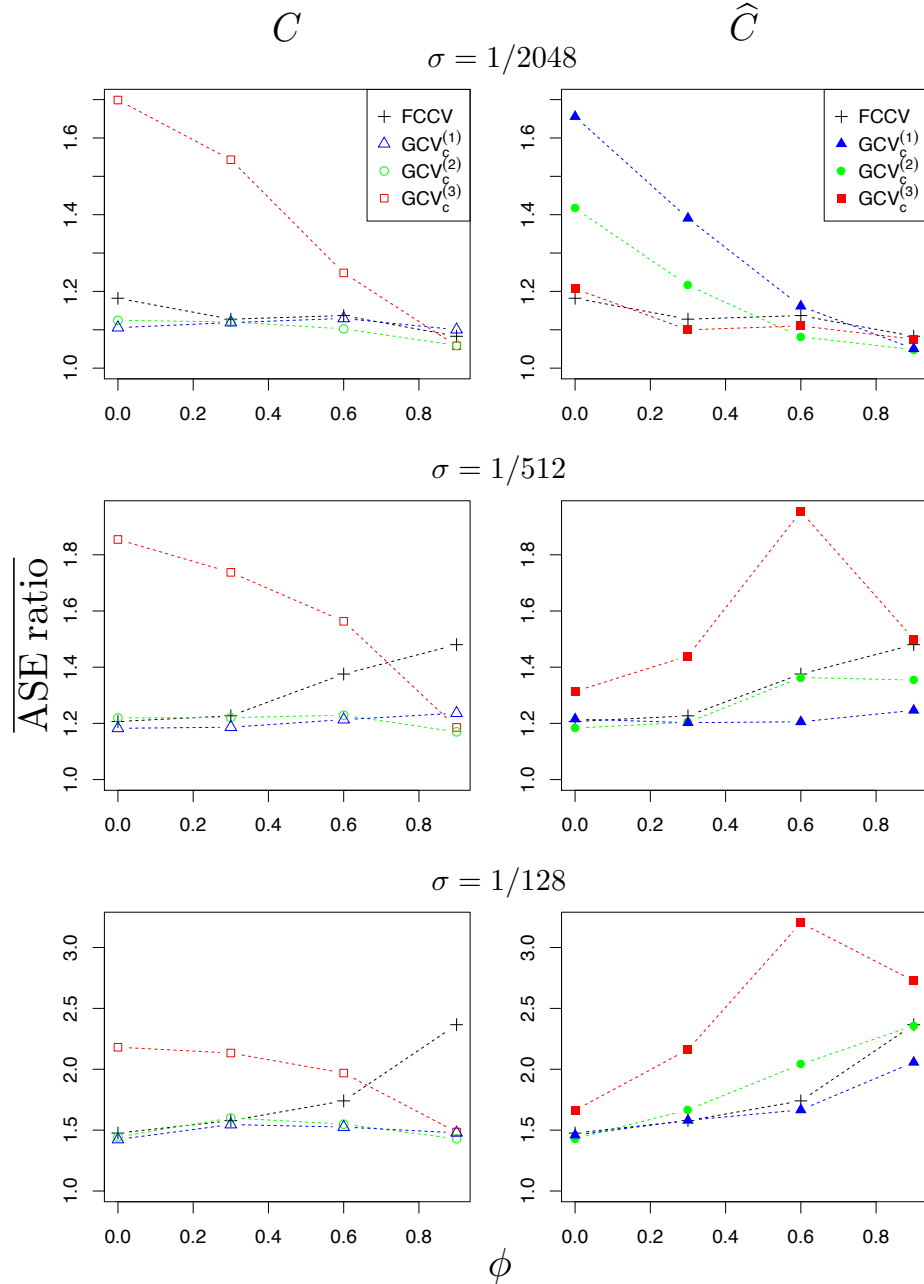


FIGURE 12. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ , for the simulations using $f_4(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 1,000 realizations. The rows are arranged by variance from low to high, top to bottom. The columns indicate whether the known correlation matrix, C , was used, or its nonparametric semivariogram estimate, \hat{C} .