GENERALIZED CORRELATED CROSS-VALIDATION (GCCV)

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, which is motivated by a natural definition for residual degrees of freedom. The efficacy of the new method is investigated with a simulation experiment on a kernel smoother with bandwidth selection in local linear regression. Next, the winning methodology is illustrated by application to spatial modeling of fMRI data using a nonparametric semivariogram. We conclude with remarks about the heteroscedastic case and a potential maximum likelihood framework for Gaussian random processes.

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

Date: May 10, 2011.

Key words and phrases. effective degrees of freedom, fMRI, model selection, nonparametric, semivariogram, supervised learning, tuning parameter.

[†]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 correlated cross-validation (GCCV), from a nonparametric perspective, but conclude with some interesting connections to a parametric setting.

2. Theoretical Foundation of Cross-Validation

Suppose $y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, where $f(\cdot)$ is a function, and ε_i is stochastic with $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 < \infty$, and $n \times n$ covariance matrix given by $(\Sigma)_{ij} = \sigma^2 (C)_{ij} = \sigma^2 \mathrm{cor}(\varepsilon_i, \varepsilon_j) = \sigma^2 r_{ij}$, $C \neq J$. 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, $w_1, \ldots, w_n$, 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

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

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.

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

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

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

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

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

Letting $\hat{\mathbf{f}}(\mathbf{x}) = S\mathbf{y}$, where $(S)_{ij} = w_{ij}$, $\mu_k = \sum_{i=1}^n w_{ki} f(x_i) - f(x_k)$, 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

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

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 $n - \mathrm{tr}[S]$, which is proportional to the familiar denominator in GCV.

3. Proposed Methodology

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

$$\mathrm{tr}\left[I - \left(2S - SS'\right)\right], \tag{3}$$

$$\mathrm{tr}\left[I - S\right], \text{ and} \tag{4}$$

$$\mathrm{tr}\left[I - SS'\right]. \tag{5}$$

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

$$\mathrm{tr}\left[C + SCS' - 2SC\right] = n - \mathrm{tr}\left[2SC - SCS'\right], \tag{6}$$

which is the analogue of (3) when taking correlation into account. One can show that $0 \leq \mathrm{tr}[I - (2S - SS')] \leq \mathrm{tr}[I - S] \leq \mathrm{tr}[I - SS'] \leq n$ provided $0 \leq \lambda_i \leq 1$ using von Neumann's trace inequality (Mirsky, 1975) to show that $\mathrm{tr}[SS'] \leq \mathrm{tr}[S]$, where λ_i, $i = 1, \ldots, n$, are the eigenvalues of S. Similarly, (6) is the most stringent of the correlated analogues of (3), (4), and (5) since one can show that $\mathrm{tr}[SCS'] \leq \mathrm{tr}[SC]$ again using von Neumann's inequality in conjunction with the eigenvalues of S and SC. It is interesting to note that (6) is equivalent to (3) in the independent case, and so differs from (4) employed by GCV. Adopting (6) as our definition of degrees of freedom leads us to define the generalized cross-validation for correlated data surface as

$$\mathrm{GCCV}_1\left(\boldsymbol{\theta}\right) = \frac{1}{n}\left(\frac{\sum_{k=1}^n y_k - \hat{f}\left(x_k\right)}{1 - \mathrm{tr}\left[2SC - SCS'\right]/n}\right)^2, \tag{7}$$

with $\hat{\boldsymbol{\theta}}$ as its minimizer. An application with a two-dimensional tuning parameter appears in Section 5.

4. Simulations of Kernel Smoothers

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 methods such as ordinary cross-validation, one-sided cross-validation, and plugin, in correlated data. We are interested in assessing the performance of these proposed methods whose task is to select a global bandwidth, one-dimensional $\theta = h$, for local linear regression (Fan, 1992) with serially correlated additive errors. The local linear regression estimate is given by

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

The tricube kernel, $K(u) = 35/32(1 - |u|^3)^3$, $|u| \leq 1$, was used because it performs well in a variety of settings. The following four functions were selected for the variety of structure (Fig. 1):

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

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 $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$. A vector of serially correlated errors, ε_i, $i = 1, \ldots, n$, was generated using `arima.sim` in R (R Development Core Team, 2011) 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 $(x_i, f(x_i) + \varepsilon_i)$, $i = 1, \ldots, n$, which was repeated 10,000 times for each combination of $f(\cdot)$, n, ϕ, and σ. We designed this simulation with fMRI data in mind. Hence our primary interest is in $n = 150$.

For comparison, we include FCCV along with the following:

$$\mathrm{GCCV}_2(\boldsymbol{\theta}) = \frac{1}{n}\left(\frac{\sum_{k=1}^{n} y_k - \hat{f}(x_k)}{1 - \mathrm{tr}[SC]/n}\right)^2, \text{ and} \tag{8}$$

$$\mathrm{GCCV}_3(\boldsymbol{\theta}) = \frac{1}{n}\left(\frac{\sum_{k=1}^{n} y_k - \hat{f}(x_k)}{1 - \mathrm{tr}[SCS']/n}\right)^2, \tag{9}$$

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 GCCV criteria, the correlation matrix C was either assumed known or estimated using the first five lags of the empirical semivariogram of the detrended data fit using a nonparametric semivariogram estimator. Detrending was accomplished using `loess` in `R` with a span of 0.75 for $n = 75$, and 0.75/2 for $n = 150$ to remove gross mean trend. This value of 0.75 is the default value in `loess`, while 0.75/2 was selected for $n = 150$ since the sampling rate is twice that of $n = 75$ in the unit interval. The empirical semivariogram for a time series at lag k is $\tilde{\gamma}_k = \sum_{|i-j|=k}(r_i - r_j)^2/2(n-k)$ with the nonparametric fit given by $\hat{\gamma}(k) = \sum_{i=1}^{m}[1 - \Omega_\kappa(kt_i)]p_i$, where $r_i = y_i - \hat{y}_i$ is the i^{th} residual where $\hat{y}_i$ is the LOESS estimate of $f(x_i)$, κ is the order of the basis set to 11 for our purposes, and p_i, $i = 1, \ldots, m$, 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 residuals likely contain mean structure, which is why the nonparametric semivariogram is fit using the first five lags of the empirical semivariogram. The nonparametric semivariogram is then used to estimate C as

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

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

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

which is 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. Fig. 2 shows a sample realization along with the fits produced by the three GCCV criteria. Although not included below, we also recorded the results using ordinary and generalized cross-validation, which are known to perform poorly when data are correlated. Their results were omitted from the following summaries since their ASE was often several times higher than the other included methods. One should also remember that GCV is equivalent to GCCV$_2$ when $C = I$.

Figs. 3 and 4 for function $f_1(\cdot)$ (reviewers please see Figs. 6–11 for functions $f_2(\cdot) - f_4(\cdot)$) in the Appendix), which have been Bonferroni corrected at the 0.01 level of significance on a per figure basis, demonstrate that GCCV$_1$ generally dominates both GCCV$_2$ and GCCV$_3$ in terms of mean ASE ratio relative to ASE$_0$. One can visualize GCCV$_1$ as the winner in all six combinations in Fig. 3 ($n = 150$) since the blue triangles are the smallest values of ASE, often statistically so. The non-overlapping plotting symbols are significantly different at the corrected level within each figure.

Only when the correlation structure is estimated, the error variance is low, and using the smaller sample size did GCCV$_2$ and GCCV$_3$ outperform GCCV$_1$. An investigation revealed that positive bias due to the LOESS residuals in the empirical semivariogram was the culprit. This leads to overestimating the correlation structure, which GCCV$_1$ punishes more heavily than the other two (reviewers please see Fig. 16 in the Appendix). This in turn leads to GCCV$_1$ over smoothing (reviewers please see Figs. 12-15 in the Appendix) the data resulting in significantly higher mean ASE ratios.

Figs. 3 and 4 (reviewers please see Figs. 6–11 in the Appendix) also suggest that the three GCCV methods generally perform similarly when $\phi = 0.9$. As $\phi \to 1$, $C \to J$, which implies that all three definitions of residual degrees of freedom approach 0 regardless of S. Hence, their similarity at $\phi = 0.9$ is not surprising. However, the differences in their performances for lower values of ϕ merely indicates that the three definitions approach 0 at different rates. For example,

$$\frac{1 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij} - 2\sum_{i=1}^{n} w_i r_{ik}}{1 - \sum_{i=1}^{n} w_i r_{ik}} = 1 + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij} - \sum_{i=1}^{n} w_i r_{ik}}{1 - \sum_{i=1}^{n} w_i r_{ik}},$$

is approximately 1 provided $\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij} \approx \sum_{i=1}^{n} w_i r_{ik}$ (i.e., the weighted mean of C is approximately the weighted mean of one of its rows). Similarly,

$$\frac{1 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij} - 2\sum_{i=1}^{n} w_i r_{ik}}{1 - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}} = 1 + \frac{2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij} - \sum_{i=1}^{n} w_i r_{ik}\right)}{1 - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}},$$

which is again approximately 1 under the same conditions. In the context of local linear regression, this means that GCCV$_1$, GCCV$_2$, and GCCV$_3$ are roughly equivalent where the approximation generally holds for higher bandwidths and/or lower values of ϕ, but not at lower bandwidths and/or higher values of ϕ.

FCCV performed well, often besting GCCV_2 and GCCV_3 in terms of mean ASE ratio. The same is true of GCCV_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 GCCV. Similar conclusions may be reached from the results for functions $f_2(\cdot)$, $f_3(\cdot)$, and $f_4(\cdot)$.

5. Application to Spatial Semivariograms

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

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

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 $\tilde{\gamma}(s_i)/\hat{\gamma}(s_i|\boldsymbol{\theta})$, $i = 1, \ldots, \ell$ (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).

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

where $\widehat{C}$ is the estimated correlation matrix of $\tilde{\gamma}(s_i)/\hat{\gamma}(s_i|\boldsymbol{\theta})$, $i = 1, \ldots, \ell$, 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 practioners 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 this application had the subject silently repeat nonsense words displayed on a monitor above their head for a total of 152 scans spaced 2 seconds 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.). These were fit 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. 5 shows, the empirical semivariogram of the third FIR parameter exhibits spatial correlation to approximately 9 mm. Our experienced view 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. This semiparametric fit is estimated using (11) with `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 is 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 yields $\hat{\kappa} = 20.6$ and $\hat{\alpha} = 0.709$ with the resulting semiparametric fit along with the parametric exponential and spherical fits shown in Fig. 5. 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.

6. Discussion

As the theory section established, a natural definition for residual degrees of freedom is $\mathrm{tr}[C + SCS' - 2SC]$, which suggests an extension of GCV for correlated data as defined by GCCV_1. Historical consideration of three competing definitions for residual degrees of freedom and their correlated counterparts led us to define GCCV_2 and GCCV_3. As the simulation study showed, GCCV_1 tends to dominate the other two. However, this is not universally the case when estimating correlation at smaller sample sizes with low variance. In that case, GCCV_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 all the other cases, the first still tends to dominate the other two leading us to conclude that GCCV_1 is the best choice for most situations. As such, we will refer to GCCV_1 more simply as GCCV 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 GCCV 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 do 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 $\mathrm{E}[\sum_{k=1}^{n}(\hat{f}(x_k) - y_k)^2]/\sigma^2$ matches the first moment of a non-central χ^2 on $\mathrm{tr}[C + SCS' - 2SC]$ degrees of freedom. Given this observation, one might wonder why GCCV divides by the square of the trace. Assuming that the errors are distributed multivariate normal, $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 C)$, the variance is given by

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

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 GCCV. Letting $V = C + SCS' - SC - CS'$, a closer inspection reveals that

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

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

Continuing with the multivariate 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)

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

This is a convolution of gammas with common shape parameter $\alpha = -1/2$ and scale parameters $\beta_j = 2\sigma^2\lambda_j$, $j = 1, \ldots, n$, which is a 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 may be viewed as estimating $\boldsymbol{\theta}$ via maximum likelihood. We opted not to present this approach at this time since we desire to make the present method as nonparametric as possible, but this remains a promising avenue of research that we 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. *Matematical 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* (2$^{\text{nd}}$ 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.

Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte für Mathematik*, *79*(4), 303–306.

R Development Core Team (2011). *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-validatory model-selection for dependent data: *hv*-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-validatory choice and the assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, *B 36*, 111–133.

Figures

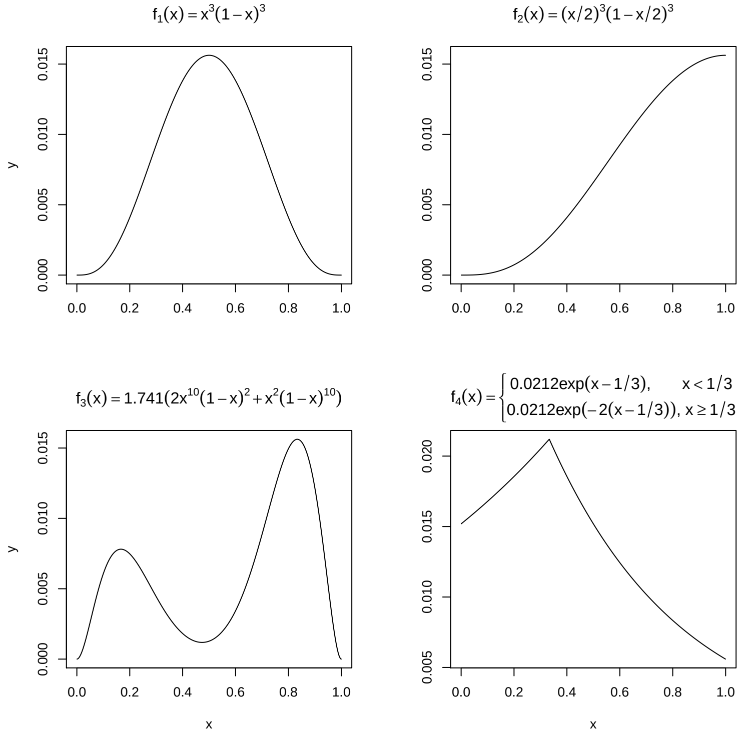

Figure 1. Plot of the four functions used in the simulation study. Additive errors of varying correlation and variance were added to each function at either $n = 75$ or $n = 150$ equally spaced design points in the interval $[0, 1]$. A global bandwidth for local linear regression was estimated by several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

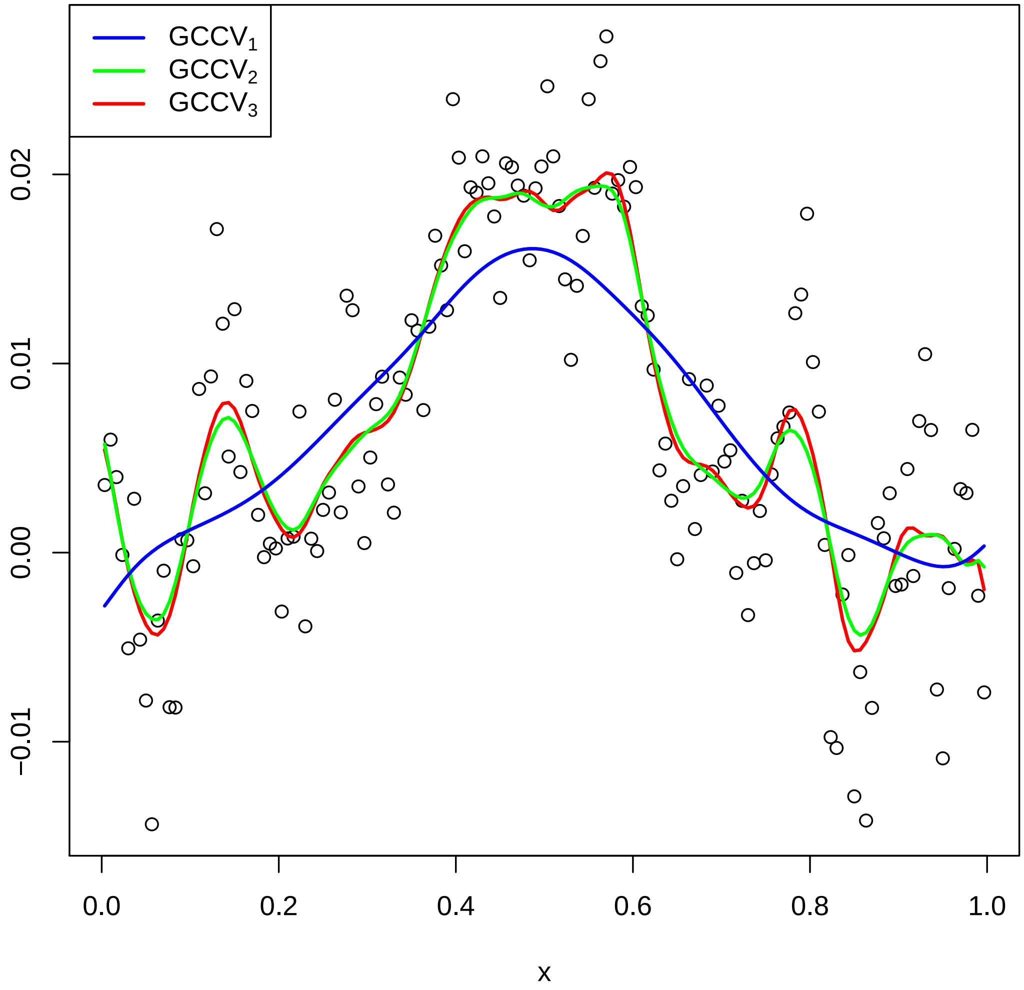

FIGURE 2. Three local linear regression fits with bandwidths estimated by the three GCCV criteria using the known correlation structure. $\hat{h}$ was estimated to be 0.251, 0.067, and 0.058 for GCCV$_1$, GCCV$_2$, and GCCV$_3$, respectively. The sample was generated using $f(\cdot) = f_1(\cdot)$, $n = 150$, $\phi = 0.6$, and $\sigma = 1/128$.

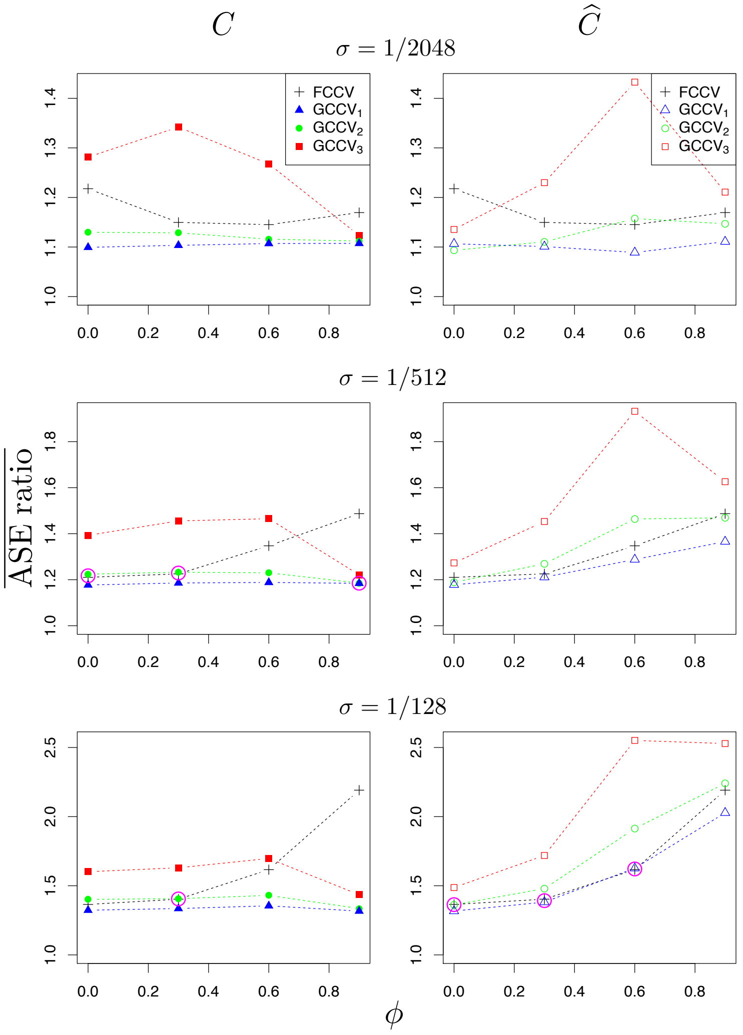

FIGURE 3. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_1(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

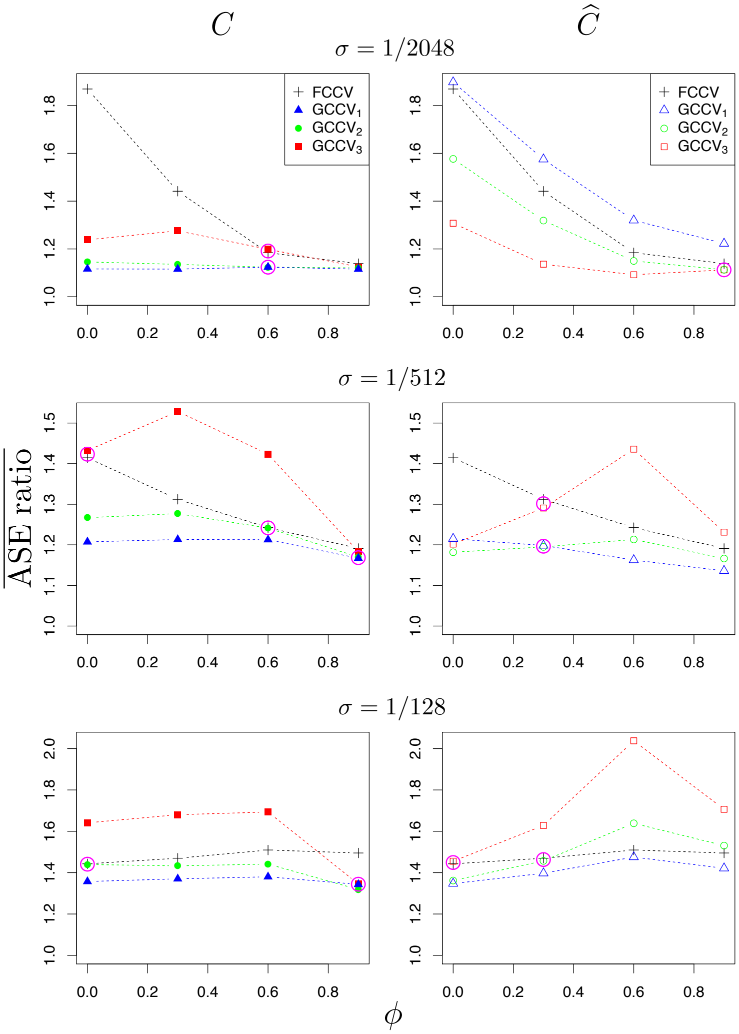

FIGURE 4. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_1(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

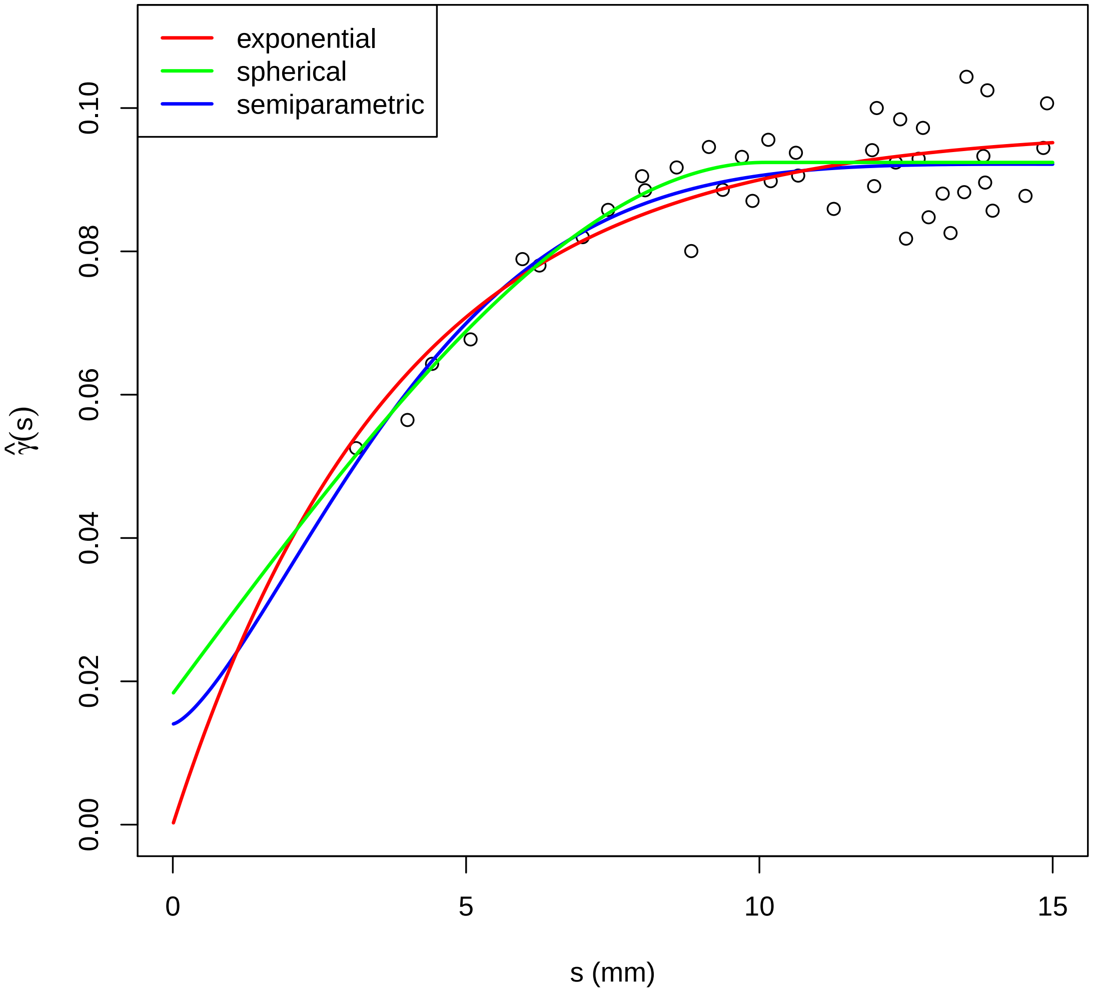

FIGURE 5. Plot of the empirical semivariogram of the third FIR parameter from the left superior temporal gyrus of a single subject in an fMRI experiment with three semivariogram fits. The exponential fit (red) overshoots the early lags and fails to level out at later lags. The spherical fit (green) overestimates the range of the correlation. The semiparametric fit (blue) was obtained using a modified form of GCCV_1.

Appendix

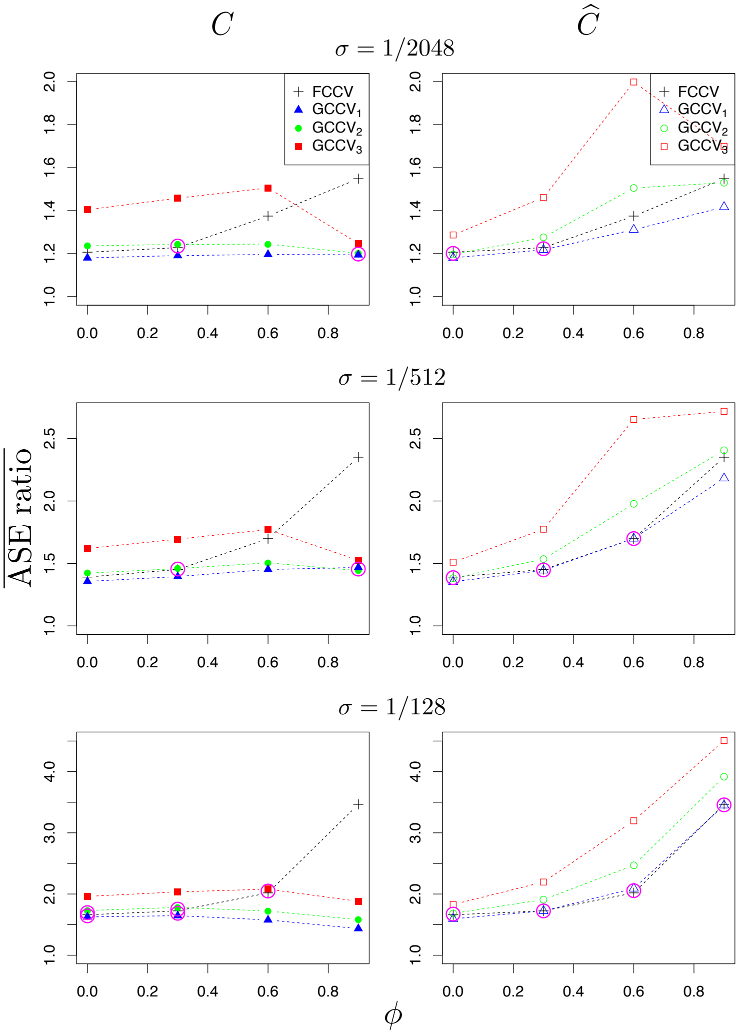

FIGURE 6. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_2(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

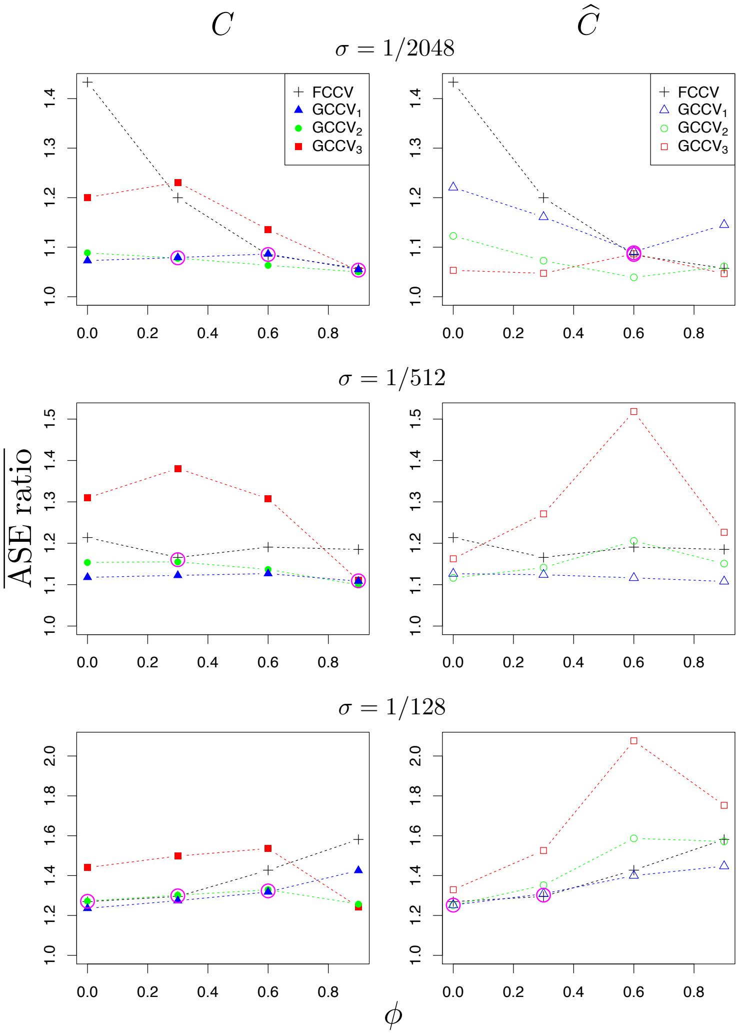

FIGURE 7. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_3(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

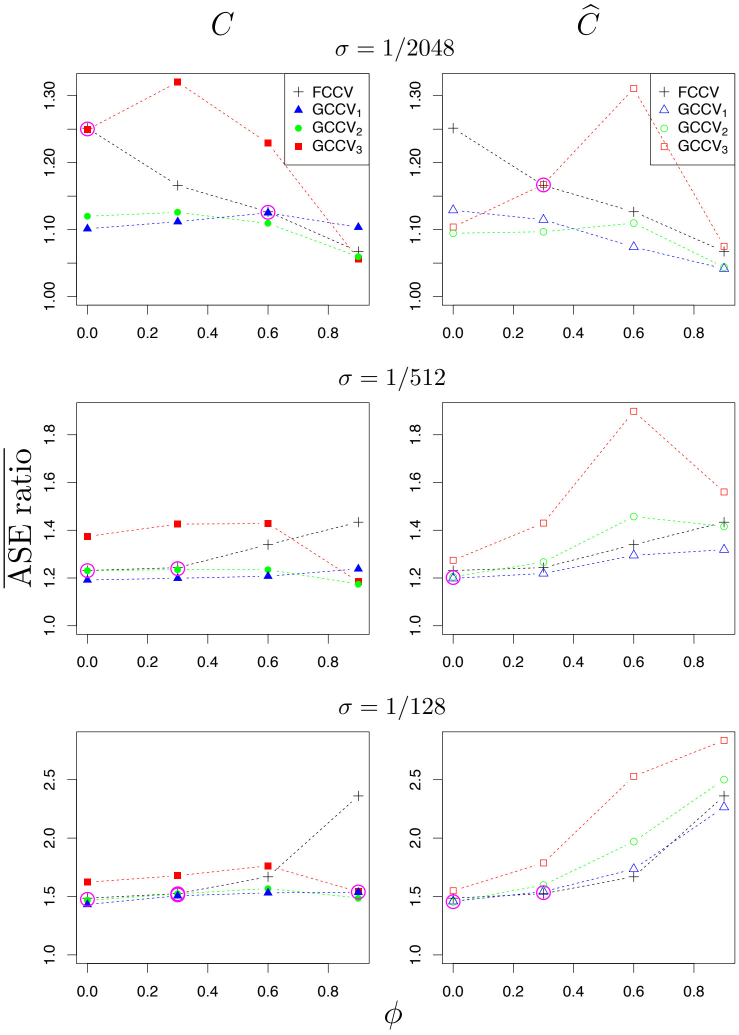

FIGURE 8. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_4(\cdot)$ and $n = 150$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

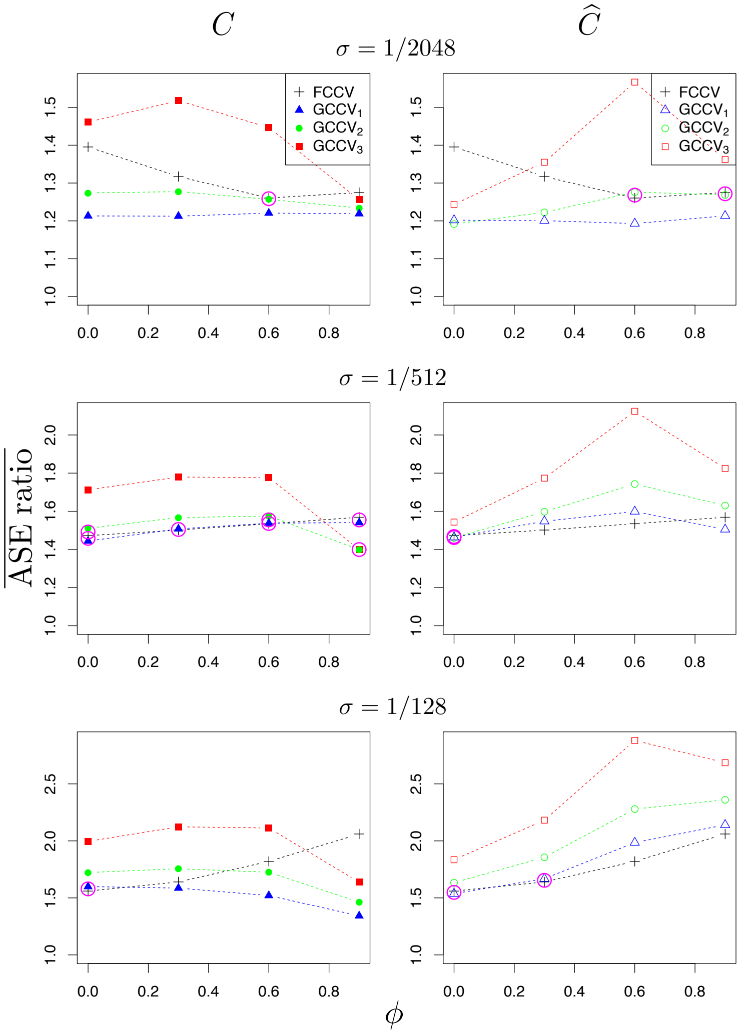

Figure 9. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_2(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

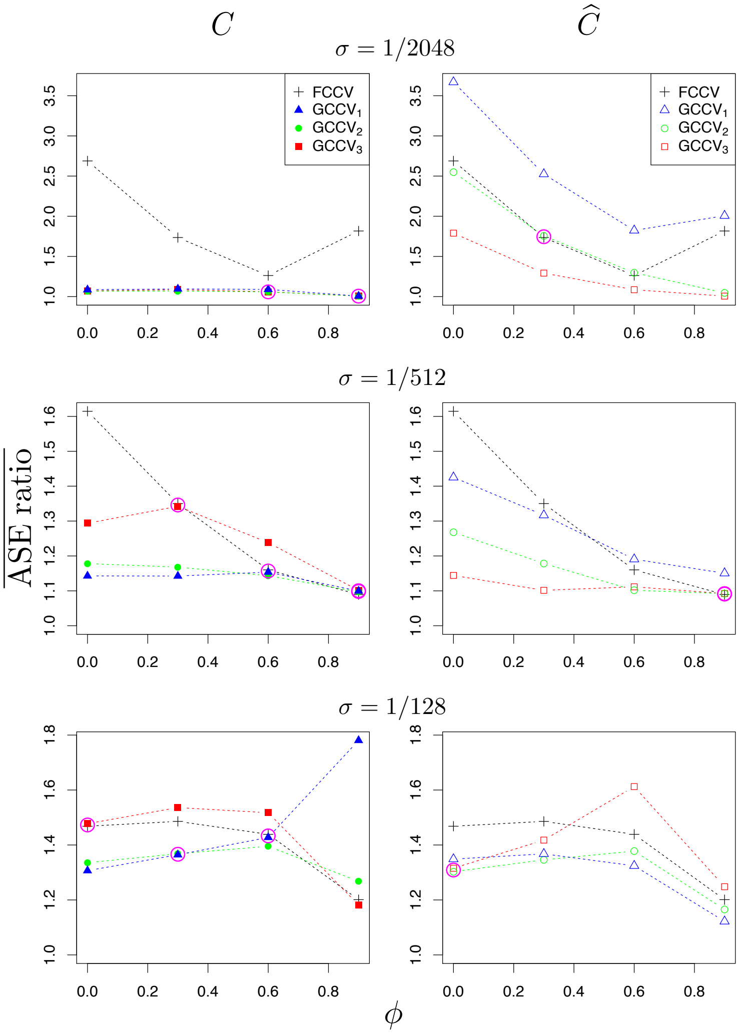

Figure 10. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_3(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

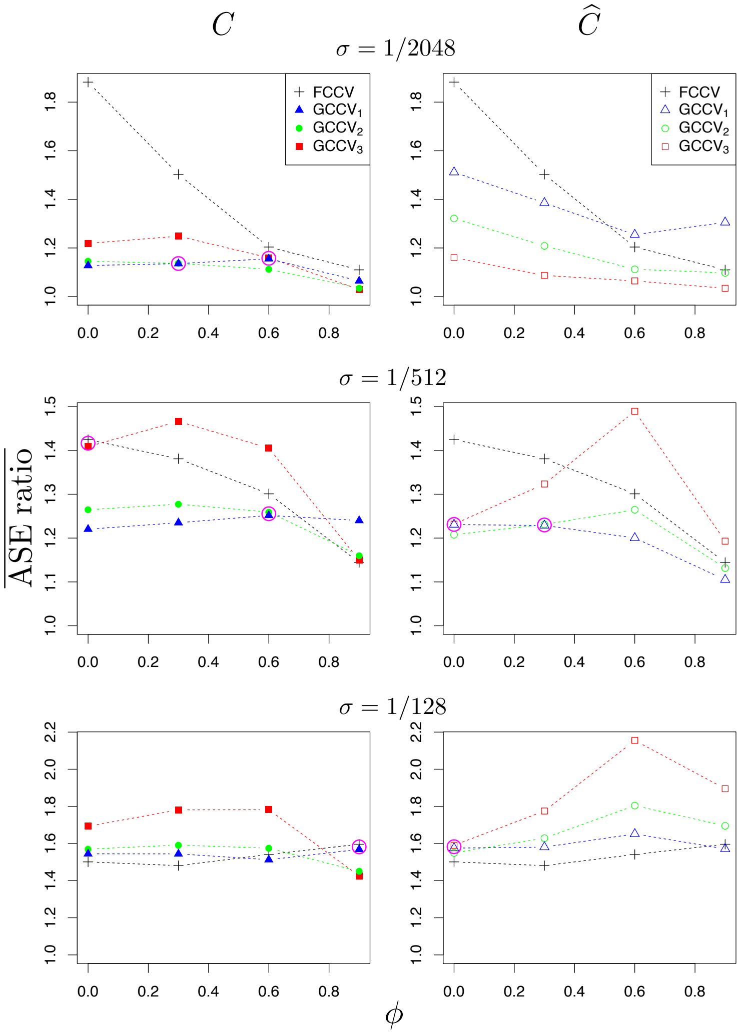

FIGURE 11. Plots of the mean ASE ratios versus the AR(1) coefficient, ϕ, for the simulations using $f_4(\cdot)$ and $n = 75$ for the methods indicated in the legends. The means were formed by taking the average of the ratio of each method's ASE over the optimal ASE for each of the 10,000 realizations. 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, $\widehat{C}$. The magenta circles indicate pairs that are not significantly different at 0.01 level of significance Bonferroni corrected for the 144 comparisons in the figure.

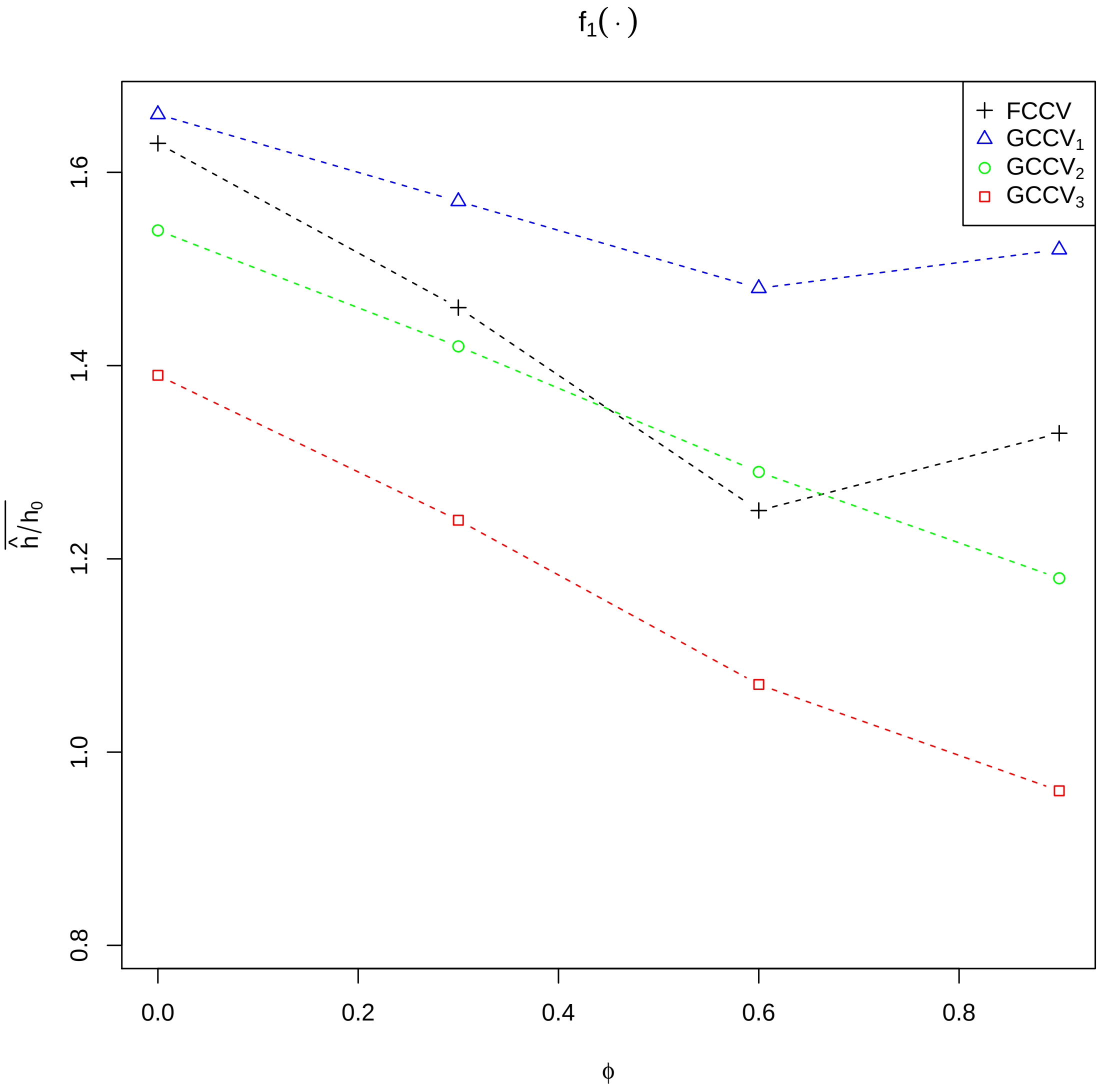

FIGURE 12. Plot of mean bandwidth ratios, $\overline{\hat{h}/h_0}$, for the four methods for $f_1(\cdot)$, $n = 75$, $\phi = 1/2048$, and $\widehat{C}$. This corresponds to the upper right panel of Fig. 4.

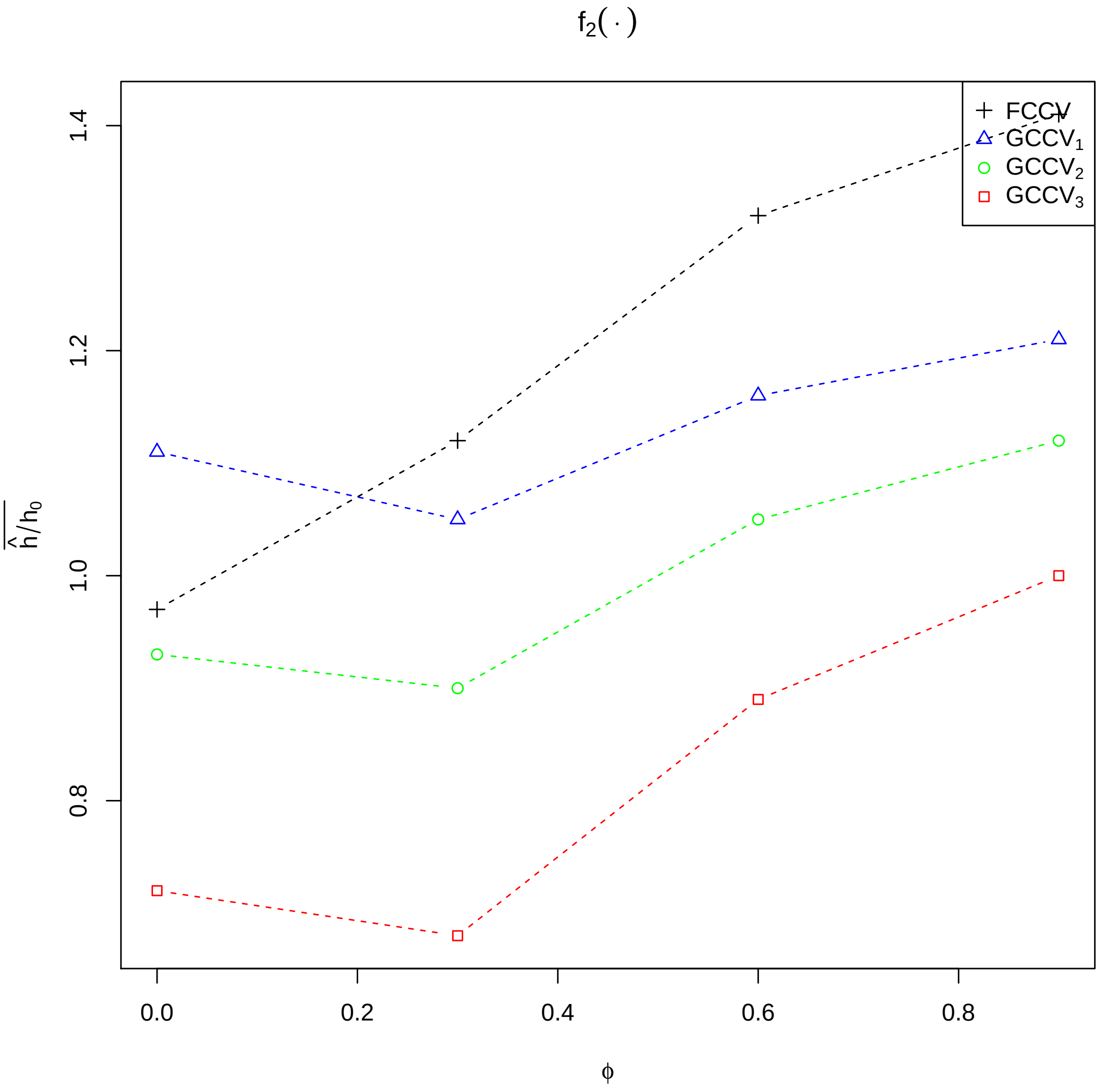

Figure 13. Plot of mean bandwidth ratios, $\overline{\hat{h}/h_0}$, for the four methods for $f_2(\cdot)$, $n = 75$, $\phi = 1/2048$, and $\widehat{C}$. This corresponds to the upper right panel of Fig. 9.

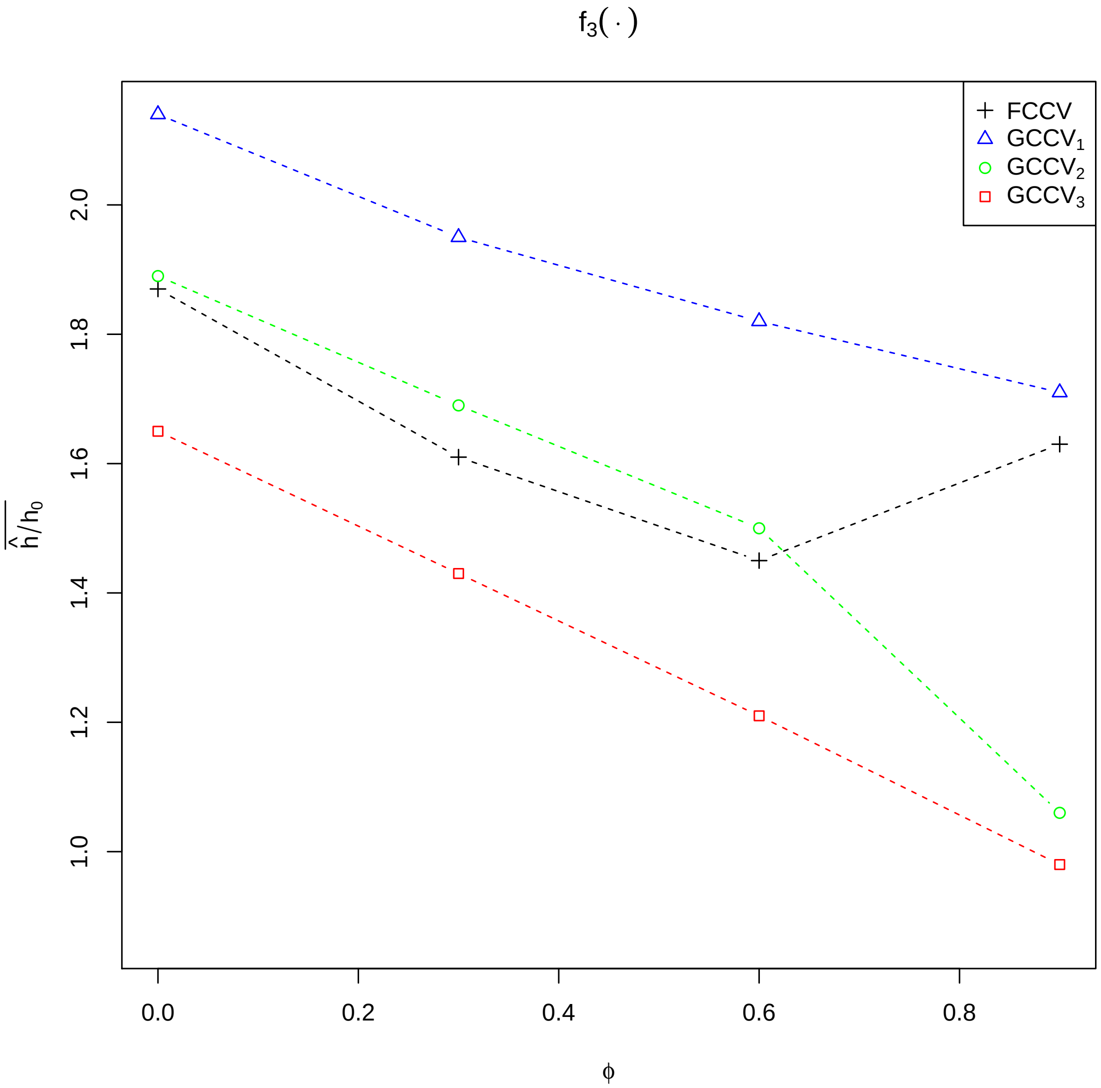

FIGURE 14. Plot of mean bandwidth ratios, $\overline{\hat{h}/h_0}$, for the four methods for $f_3(\cdot)$, $n = 75$, $\phi = 1/2048$, and $\widehat{C}$. This corresponds to the upper right panel of Fig. 10.

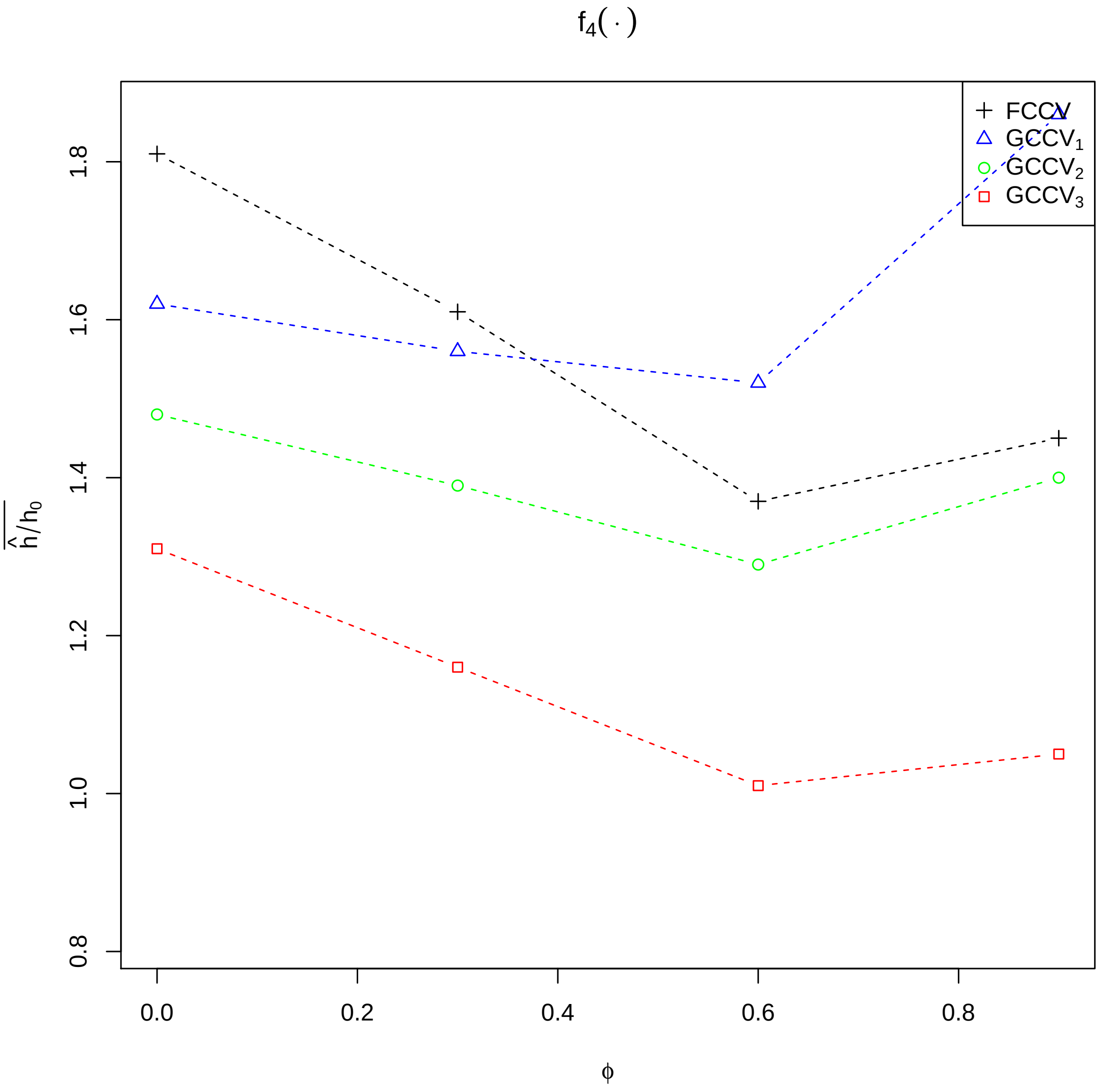

FIGURE 15. Plot of mean bandwidth ratios, $\overline{\hat{h}/h_0}$, for the four methods for $f_4(\cdot)$, $n = 75$, $\phi = 1/2048$, and $\widehat{C}$. This corresponds to the upper right panel of Fig. 11.

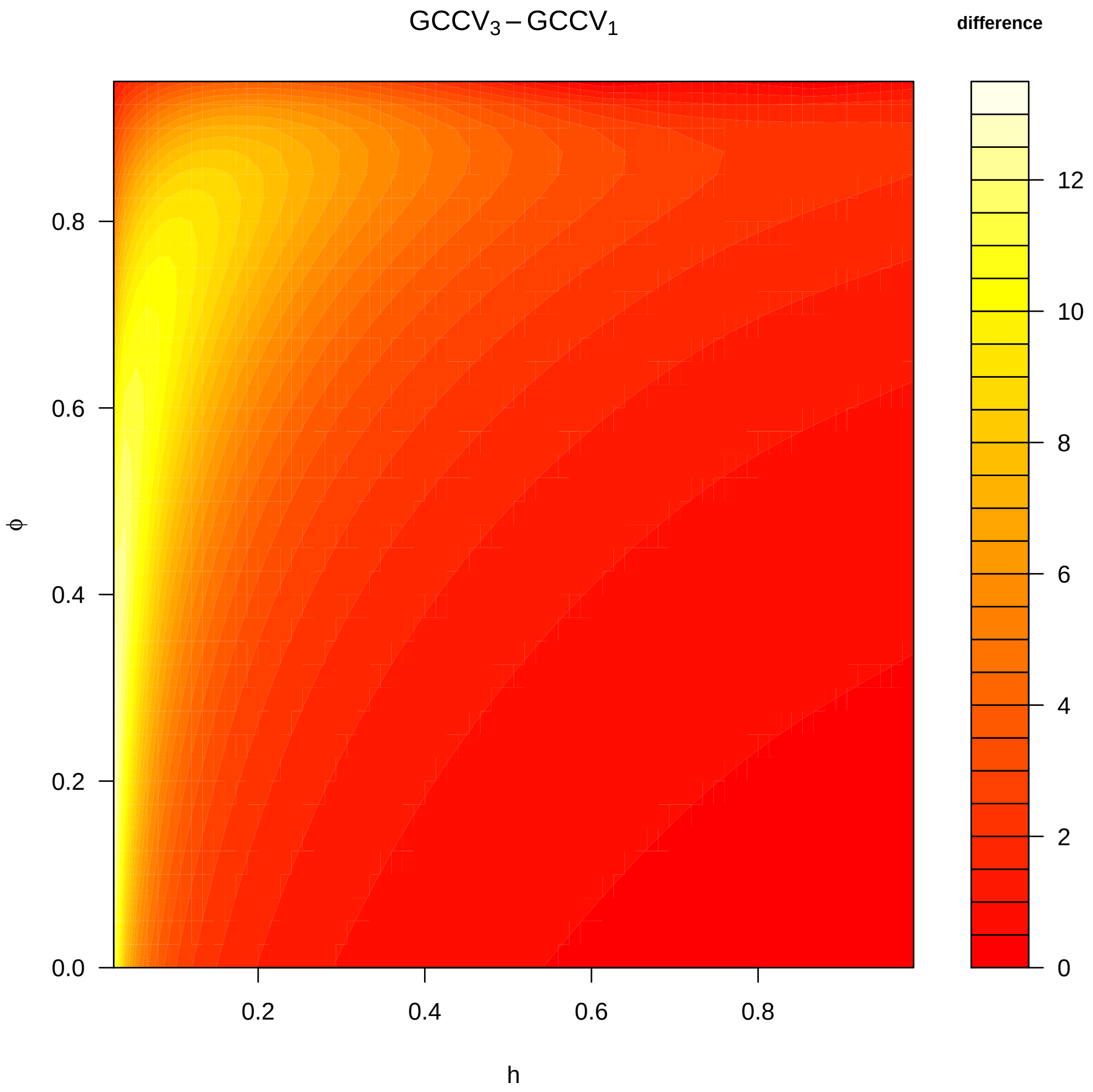

FIGURE 16. A filled contour plot of the difference in residual degrees of freedom as defined by GCCV_3 and GCCV_1 in the context of local linear regression with bandwidth h, $n = 75$ equally spaced design points, and known correlation structure given by an AR(1) process with coefficient ϕ. The largest differences occur for smaller bandwidths and/or higher correlation.