Supplementary materials for this article are available at http://pubs.amstat.org/toc/jcgs/18/4.

Far Casting Cross-Validation

Patrick S. CARMACK, William R. SCHUCANY, Jeffrey S. SPENCE, Richard F. GUNST, Qihua LIN, and Robert W. HALEY

Cross-validation has long been used for choosing tuning parameters and other model selection tasks. It generally performs well provided the data are independent, or nearly so. Improvements have been suggested which address ordinary cross-validation's (OCV) shortcomings in correlated data. Whereas these techniques have merit, they can still lead to poor model selection in correlated data or are not readily generalizable to high-dimensional data.

The proposed solution, far casting cross-validation (FCCV), addresses these problems. FCCV withholds correlated neighbors in every aspect of the cross-validation procedure. The result is a technique that stresses a fitted model's ability to extrapolate rather than interpolate. This generally leads to better model selection in correlated datasets.

Whereas FCCV is less than optimal in the independence case, our improvement of OCV applies more generally to higher dimensional error processes and to both parametric and nonparametric model selection problems. To facilitate introduction, we consider only one application, namely estimating global bandwidths for curve estimation with local linear regression. We provide theoretical motivation and report some comparative results from a simulation experiment and on a time series of annual global temperature deviations. For such data, FCCV generally has lower average squared error when disturbances are correlated.

Supplementary materials are available online.

Key Words: Dependent data; Optimistic error rates; Prediction; Temporal correlation; Tuning parameter.

1. INTRODUCTION

Cross-validation as described by Stone (1974) and Geisser (1975) is a well-established method for model selection and estimation of prediction error in the nonparametric regres-

Patrick S. Carmack is Assistant Professor, Department of Mathematics, University of Central Arkansas, 201 Donaghey Avenue, Conway, AR 72035 (E-mail: *patrickc@uca.edu*). Jeffrey S. Spence is Assistant Professor (E-mail: *Jeffrey.Spence@UTSouthwestern.edu*), Robert Haley is Professor (E-mail: *Robert.Haley@UTSouthwestern.edu*), and Qihua Lin is Assistant Professor (E-mail: *Catherine.Lin@UTSouthwestern.edu*), Department of Internal Medicine, Epidemiology Division, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-8874. William R. Schucany is Professor (E-mail: *schucany@smu.edu*) and Richard F. Gunst is Professor (E-mail: *rgunst@smu.edu*), Department of Statistical Science, Southern Methodist University, P.O. Box 750332, Dallas, TX 75275-0332.

Journal of Computational and Graphical Statistics, Volume 18, Number 4, Pages 879–893
DOI: 10.1198/jcgs.2009.07034

sion and classification settings. The original method starts by withholding one data point at a time, builds a model with the remainder of the data, and uses that model to predict the withheld data point. In a model selection setting, the model with the lowest average squared prediction error is declared the best. In a modification of ordinary cross-validation (OCV), sometimes known as v-fold cross-validation, the data are randomly partitioned into v training sets and test sets for building the model and assessing its prediction error, respectively. As in ordinary cross-validation, the model with the smallest average squared prediction error is chosen as the best one. Both these approaches work well when the data are independent, but exhibit overly *optimistic* prediction error rates when the data are correlated. This can lead to biased model selection. Hastie, Tibshirani, and Friedman (2001, chapter 7) had an excellent discussion on model selection and error assessment as it relates to optimism.

Burman, Chow, and Nolan (1994) addressed this dependence issue with what they called h-block cross-validation. Instead of withholding one data point at a time or dividing the data into training and test sets, h-block cross-validation obtains a parameter estimate at each data point, which is the minimizer of weighted least squares where the weights omit the point and its neighbors within h units. This set of estimates is then used to arrive at the corrected h-block cross-validated estimate. Racine (2000) provided a consistent modification called hv-block cross-validation that essentially combines h-block and v-fold cross-validation. Hart and Yi (1998) studied a method known as one-sided cross-validation (OSCV) that approaches the problem by omitting the data from either the left or right of the point of prediction and using the remaining data for both model estimation and prediction. All of these methods produce less optimistic error estimates and thereby improve model selection. The focus of this article is to introduce a conceptually similar, yet fundamentally different, improvement to cross-validation in correlated data, which we have dubbed far casting cross-validation (FCCV).

2. FAR CASTING CROSS-VALIDATION

The new method, FCCV, withholds a block of neighbors from the entire cross-validation process. This approach assesses the candidate model's ability to predict across deleted neighborhoods, or extrapolate, as opposed to its ability to interpolate when the full dataset is used. Basically, OSCV approaches the problem in a similar fashion by only using data to the left or right of the point of prediction to estimate the model and obtain a prediction. Unfortunately, the notion of left and right does not have a unique extension in higher dimensions.

A parameter, denoted by d, specifies the radial distance for neighbors to exclude from the prediction process with $d = 0$ being equivalent to ordinary cross-validation. In this context, extrapolation is used to refer to predictions obtained within deleted neighborhoods. As shown in the theory section, the optimal selection of d primarily depends on the error covariance structure. The greater the positive correlation among neighboring points, the larger d should be for better performance.

To simplify discussion, consider the special case of global bandwidth estimation in local linear regression. Let $y_1, \ldots, y_n$ be from the regression model

$$y_i = r(x_i) + \varepsilon_i, \tag{2.1}$$

where r is a smooth mean function, $x_1 < x_2 < \cdots < x_n$ are fixed design points, and ε_i are uncorrelated random disturbances such that $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$. Suppose one wishes to estimate r using a local linear estimator (Fan 1992). Consider the criterion function

$$\sum_{i=1}^{n} (y_i - a - b(x_i - x))^2 K\left(\frac{x - x_i}{h}\right), \tag{2.2}$$

where a is the local linear estimate of $r(x)$, $h > 0$ is the bandwidth of the unimodal kernel, K, which is symmetric about 0 and has finite variance. Minimizing (2.2) with respect to a and b for a fixed h leads to

$$\hat{r}(x) = \hat{a} = \frac{\sum_{i=1}^{n} y_i \cdot w_i(x)}{\sum_{i=1}^{n} w_i(x)}, \tag{2.3}$$

where

$$w_i(x) = K\left(\frac{x - x_i}{h}\right)(t_{n,2} - (x - x_i)t_{n,1}) \tag{2.4}$$

and

$$t_{n,j} = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(x - x_i)^j, \qquad j = 1, 2. \tag{2.5}$$

In practice, the global bandwidth, h, must be estimated from the data. Values of h that are too small will result in a fit that is excessively variable whereas values of h that are too large oversmooth the curve. In FCCV, the smoothing parameter h is chosen as the minimizer of cross-validation error defined as

$$\mathrm{CV}(h, d) = \frac{1}{n} \sum_{i=1}^{n} (\hat{r}_{h,d}(x_i) - y_i)^2, \tag{2.6}$$

where $\hat{r}_{h,d}(x_i)$ denotes the local linear estimate at x_i obtained by deleting $\{(x_j, y_j) : |x_i - x_j| \le d\}$. This means that FCCV does not use these neighboring points to minimize (2.2) or to evaluate (2.6). This is in contrast to h-block, which omits them in (2.2), but not in its analog to (2.6). When $d = 0$, (2.6) reduces to ordinary cross-validation. OSCV makes a similar modification by omitting either $\{(x_j, y_j) : x_i \ge x_j\}$ or $\{(x_j, y_j) : x_i \le x_j\}$ during both fitting and prediction. The fundamental difference between FCCV and OSCV is in the definition of deleted neighborhood. For the remainder of the article, we will dispense with the assumption that the disturbances $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated.

2.1 THEORETICAL MOTIVATION

The rationale behind using ordinary cross-validation is that $\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_h(x_i)-y_i)^2$ serves as a (biased) approximation of the average squared error, $\mathrm{ASE}(h)=\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_h(x_i)-r(x_i))^2$. Taking the expectation of a single term, $(\hat{r}_{h,d}(x_{i_0})-y_{i_0})^2$, exposes the problem when errors are correlated. In the interest of compact notation, y, r, $\hat{r}_h$, and $\hat{r}_{h,d}$ will be used in place of y_{i_0}, $r(x_{i_0})$, $\hat{r}_h(x_{i_0})$, and $\hat{r}_{h,d}(x_{i_0})$, respectively. For fixed h and d,

$$\begin{aligned}
\mathrm{E}[(\hat{r}_{h,d}-y)^2] &= \mathrm{E}[\hat{r}_{h,d}^2]-2\mathrm{E}[\hat{r}_{h,d}\cdot y]+\mathrm{E}[y^2]\\
&= \mathrm{E}[\hat{r}_{h,d}^2]-2\,\mathrm{Cov}[\hat{r}_{h,d},y]-2r\mathrm{E}[\hat{r}_{h,d}]+\sigma^2+r^2\\
&= \mathrm{E}[(\hat{r}_{h,d}-r)^2]-2\,\mathrm{Cov}[\hat{r}_{h,d},y]+\sigma^2\\
&= \mathrm{E}\big[((\hat{r}_h-r)+(\hat{r}_{h,d}-\hat{r}_h))^2\big]-2\,\mathrm{Cov}[\hat{r}_{h,d},y]+\sigma^2\\
&= \mathrm{E}[(\hat{r}_h-r)^2]+2\mathrm{E}[(\hat{r}_h-r)(\hat{r}_{h,d}-\hat{r}_h)]+\mathrm{E}[(\hat{r}_{h,d}-\hat{r}_h)^2]\\
&\quad -2\,\mathrm{Cov}[\hat{r}_{h,d},y]+\sigma^2\\
&= \mathrm{E}[(\hat{r}_h-r)^2]+\sigma^2-\mathrm{Var}[\hat{r}_h]+\mathrm{Var}[\hat{r}_{h,d}]\\
&\quad +\mathrm{E}[\hat{r}_{h,d}-\hat{r}_h](\mathrm{E}[\hat{r}_{h,d}+\hat{r}_h]-2r)-2\,\mathrm{Cov}[\hat{r}_{h,d},y].
\end{aligned}\tag{2.7}$$

The leading term to the right of the final equality is the expected value of a single ASE term whereas the remaining terms are the bias. The second term, σ^2, does not depend on either d or h and does not influence where the minimum occurs with respect to h. The third term relies on h alone whereas the remainder are affected by both d and h. To minimize the effect of bias terms, d should be selected so that

$$\mathrm{Var}[\hat{r}_{h,d}]-\mathrm{Var}[\hat{r}_h]+\mathrm{E}[\hat{r}_{h,d}-\hat{r}_h](\mathrm{E}[\hat{r}_{h,d}+\hat{r}_h]-2r)-2\,\mathrm{Cov}[\hat{r}_{h,d},y]\tag{2.8}$$

is approximately constant as a function of h because this does not change where the minimum occurs for the cross-validation curve. Assuming r admits two derivatives of bounded variation, we apply a second-order approximation to obtain

$$\mathrm{E}[\hat{r}_h]=\frac{\sum_i w_i r(x_i)}{\sum_i w_i}\tag{2.9}$$

$$\approx\frac{\sum_i w_i(r(x_{i_0})+(x_i-x_{i_0})r'(x_{i_0})+(x_i-x_{i_0})^2r''(x_{i_0})/2)}{\sum_i w_i}\tag{2.10}$$

$$=r(x_{i_0})+r''(x_{i_0})\frac{\sum_i w_i(x_i-x_{i_0})^2}{2\sum_i w_i}.\tag{2.11}$$

It follows that

$$\begin{aligned}
&\mathrm{E}[\hat{r}_{h,d}-\hat{r}_h](\mathrm{E}[\hat{r}_{h,d}+\hat{r}_h]-2r)\\
&\quad\approx r''(x_{i_0})^2\left(\left(\frac{\sum_{|i-i_0|>d} w_i(x_i-x_{i_0})^2}{2\sum_{|i-i_0|>d} w_i}\right)^2-\left(\frac{\sum_i w_i(x_i-x_{i_0})^2}{2\sum_i w_i}\right)^2\right)
\end{aligned}\tag{2.12}$$

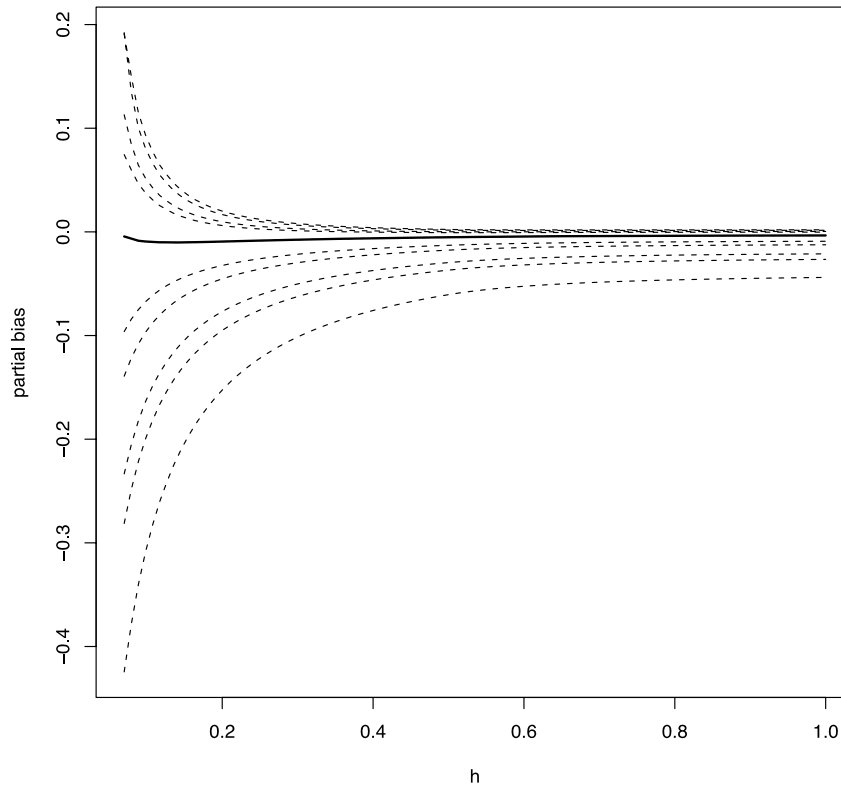

Figure 1. Plot of the partial bias (variance and covariance terms only) versus bandwidth, h, for an interior point expressed as a proportion of σ^2 for an AR(1) error process with $\phi = 0.6$ and $n = 150$. The bottom curve is ordinary cross-validation ($d = 0$), which exhibits heavy negative partial bias especially for smaller values of h. The other curves are for $d = 1/150, \ldots, 9/150$ with $d = 5/150$ shown as a solid line. The cross-validation curve associated with $d = 5/150$ has the lowest integrated squared bias in expectation. Values of d above $5/150$ have positive partial bias especially for smaller values of h.

and

$$\left| r''(x_{i_0})^2 \left(\left(\frac{\sum_{|i-i_0|>d} w_i (x_i - x_{i_0})^2}{2 \sum_{|i-i_0|>d} w_i} \right)^2 - \left(\frac{\sum_i w_i (x_i - x_{i_0})^2}{2 \sum_i w_i} \right)^2 \right) \right| < c_{h,d} r''(x_{i_0})^2. \tag{2.13}$$

This implies when r is sufficiently smooth, the variance and covariance terms will dominate the part of the bias controlled by the withholding neighborhood, d. When possible, d should be selected so that the partial bias, $\mathrm{Var}[\hat{r}_{h,d}] - \mathrm{Var}[\hat{r}_h] - 2\,\mathrm{Cov}[\hat{r}_{h,d}, y]$, is approximately constant as a function of h (Figure 1). In general, FCCV is tailored to the case where $\mathrm{Cov}[\hat{r}_{h,d}, y] > 0$ and is decreasing in d. In such cases, a radial withholding neighborhood about the point of prediction can reduce the absolute bias due to the (co)variance terms. One should note that this may not hold for more complex correlation patterns.

In the independence case, the covariance term vanishes, which leaves the two variance terms whose difference is minimized for $d = 0$ (i.e., ordinary cross-validation); however, the difference may become small enough for the curvature term to play a role in regions where r'' is large in magnitude. In such cases, OSCV will have a smaller bias approach-

ing a peak or valley from one side than OCV. It is interesting to note that the seemingly paradoxical observation by previous authors (Marron 1986; van Es 1992) that ordinary cross-validation performs worse when the error variance is low may be due to the curvature term dominating the bias. An additional insight from (2.8) is that spurious runs of positive correlation in independent errors may explain why ordinary cross-validation can sometimes fail because their covariances can downwardly bias the estimated error, especially for small values of h.

2.2 ESTIMATING d

In theory, d can be determined for an arbitrary error process provided the covariance structure is known. In practice, this is a difficult task because r is unknown. For illustration, consider the special case of a first-order autoregressive (AR(1)) process. One could estimate the semivariance at lag k by

$$\hat{\gamma}_k = \sum_{|i-j|=k} \frac{(y_i - y_j)^2}{2(n-k)}. \tag{2.14}$$

This estimator will be biased because the underlying mean function has not been removed with $\mathrm{E}[\hat{\gamma}_k] = \frac{\sigma^2}{1-\phi^2}(1-\phi^k) + \sum_{|i-j|=k}(r_i - r_j)^2/(2(n-k))$. Provided the variance of the AR(1) process is large relative to the bias, a simple estimate of ϕ is

$$\hat{\phi} = \hat{\gamma}_2/\hat{\gamma}_1 - 1. \tag{2.15}$$

$\hat{\phi}$ can then be used to numerically estimate d as the minimizer of the estimated squared partial bias,

$$\hat{d} = \operatorname*{argmin}_{d} \int_{d_{max}}^{1} (\widehat{\mathrm{Var}}[\hat{r}_{h,d}] - \widehat{\mathrm{Var}}[\hat{r}_h] - 2\,\widehat{\mathrm{Cov}}[\hat{r}_{h,d}, y])^2\, dh, \tag{2.16}$$

where d_{max} is the largest withholding neighborhood to be considered. More sophisticated procedures for general covariance structures can in principle be similarly derived.

3. BANDWIDTH SELECTION SIMULATIONS

All the methods used in the simulation experiment employed the compactly supported Epanechnikov kernel $K(u) = 0.75(1-u^2)$, $-1 \le u \le 1$ (Epanechnikov 1969). This is a second-order kernel because $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$, and $0 < \int |u^2 K(u)|\,du < \infty$. Other popular kernels such as the biweight and tricube were not used because the available software for one of the competing methodologies uses the Epanechnikov. For the sake of comparison, the simulation study largely follows that of Hart and Yi (1998). The following four functions with $0 \le x \le 1$ were considered:

$$r_1(x) = x^3(1-x)^3, \tag{3.1}$$

$$r_2(x) = (x/2)^3(1-x/2)^2, \tag{3.2}$$

$$r_3(x) = 1.741 \cdot [2x^{10}(1-x)^2 + x^2(1-x)^{10}], \tag{3.3}$$

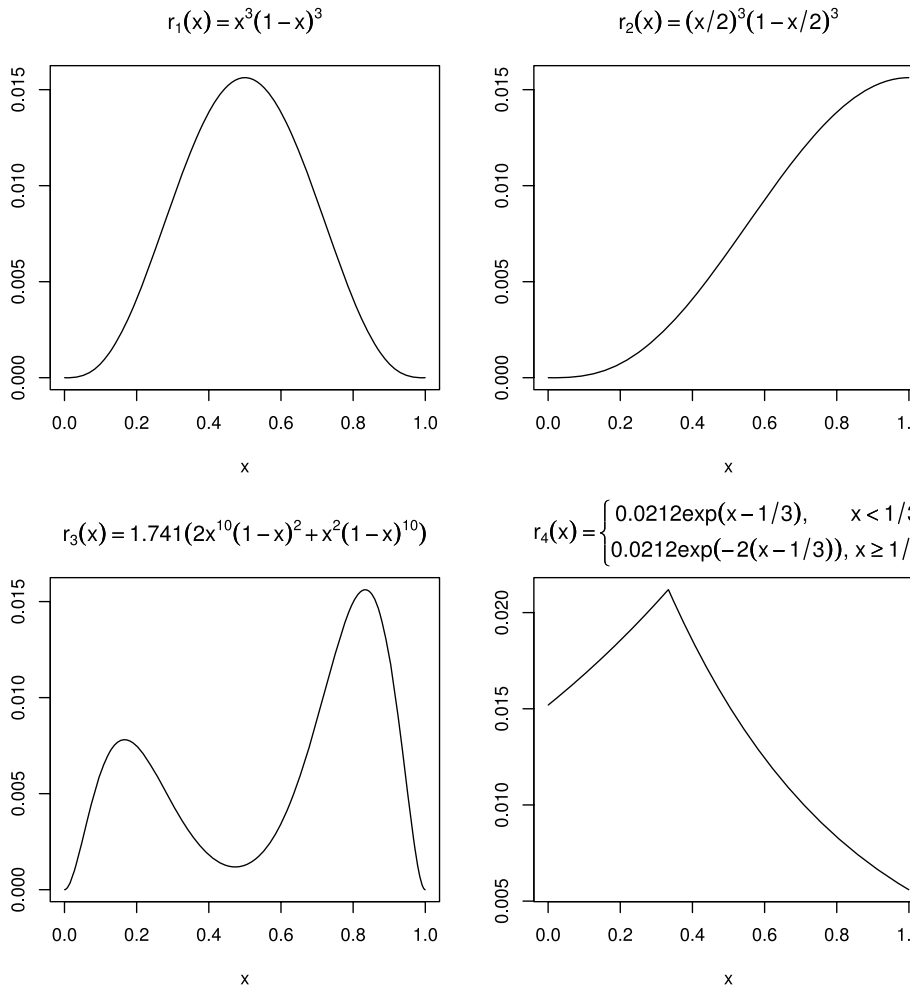

Figure 2. Plots of the four underlying functions to be estimated after adding error of varying correlation and variance. The particular functions were chosen to represent a variety of behaviors.

and

$$r_4(x) = \begin{cases} 0.0212 \cdot \exp(x - 1/3), & x < 1/3 \\ 0.0212 \cdot \exp(-2(x - 1/3)), & x \geq 1/3. \end{cases} \tag{3.4}$$

Figure 2 shows that these four functions represent a wide variety of shapes. Even though the second derivative of r_4 does not exist at $x = 1/3$, it was included to assess performance when theoretical assumptions are violated. The additive errors were taken from an AR(1) error process with coefficient ϕ and standard deviation σ. That is, $\varepsilon_i = \phi \cdot \varepsilon_{i-1} + \delta_i$, where $\delta_i \sim \mathrm{N}(0, \sigma^2/(1-\phi^2))$. Regardless of what value ϕ takes, the variance of the error process remains the same because $\mathrm{Var}(\varepsilon_i) = \{\sigma^2/(1-\phi^2)\}(1-\phi^2) = \sigma^2$ for an AR(1) process. The realizations were obtained using `arima.sim` in *R* which provides for proper correlation structure burn-in.

Four levels of ϕ were included, from 0.0 to 0.9 in 0.3 increments, which include no, minimal, moderate, and strong correlation. The equally spaced design points for both $n = 75$ and $n = 150$ were set at $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$. Finally, the amount of noise was either low ($\sigma = 2^{-11}$), medium ($\sigma = 2^{-9}$), or high ($\sigma = 2^{-7}$). Errors for each combination of r, ϕ, n, and σ were independently generated 1000 times.

Even though plug-in bandwidth selection methods are not cross-validation techniques, they are popular competitors and worthy of consideration. Several such methods exist, but the approach developed by Gasser, Kneip, and Köhler (1991) and implemented in the `lokern` library in *R* was used here and referred to as PI. OSCV requires the user to select the data either to the left or to the right of the point of estimation. Without loss of generality, the simulations always used the data to the left. For these runs we set the FCCV parameter d so that the point of prediction and three adjacent neighbors on either side were excluded for prediction purposes. This selection is not optimal for all cases, but theoretical calculations and extensive simulations show it to be a good choice for the various combinations of parameters included in the simulations. Altogether, Figure 3 shows relative ASE for estimating r_4 as a function of both withholding neighborhood d and error correlation ϕ. Direct calculations of the partial bias for various values of ϕ and the figure suggest $d = 3/n$ is a generally reasonable choice. Four bandwidth selection algorithms were investigated, OCV, OSCV, PI, and FCCV. Once each method produced an estimate of h for a given realization, its performance was compared on the basis of average squared error:

$$\mathrm{ASE}(h) = \frac{1}{n}\sum_{i=1}^{n}(r(x_i) - \hat{r}_h(x_i))^2, \tag{3.5}$$

using all (x_i, y_i) for $\hat{r}_h$.

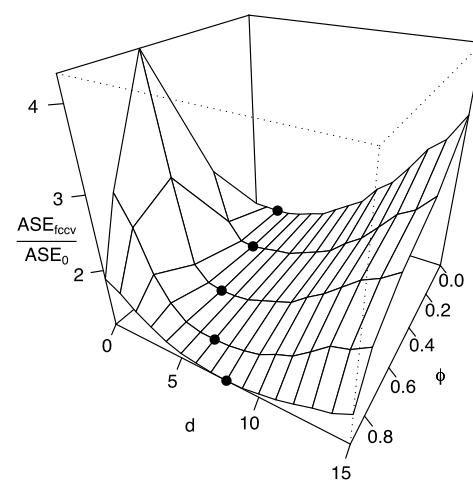

Figure 3. The ratio of the FCCV average squared error (ASE) to the optimal ASE as a function of the size of the withholding neighborhood, d, and AR(1) correlation coefficient ϕ. The points indicate the minimum ASE for each level of ϕ. The optimal neighborhood size is increasing with respect to ϕ. Each point is the result of 1000 simulations using the function r_4, $\sigma = 2^{-9}$, and $n = 150$.

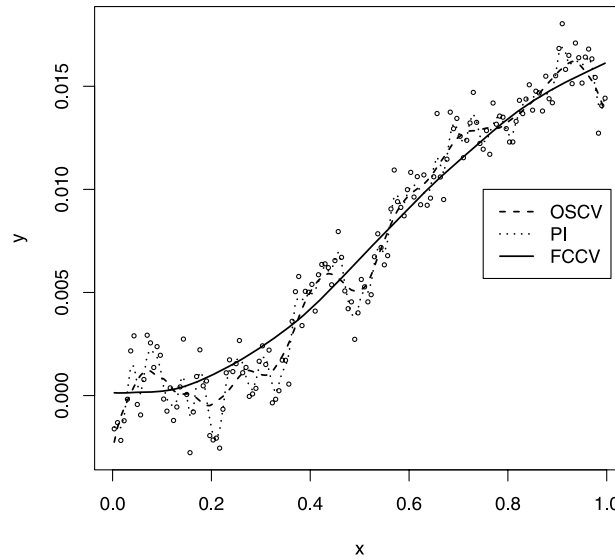

Figure 4. Example of the OSCV, PI, and FCCV methods applied to the function r_2 with $\phi = 0.6$ AR(1) error structure, $\sigma = 2^{-9}$, and $n = 150$. The circles are the sample points. The dash, dot, and solid curves are the smoothed estimates produced using OSCV, PI, and FCCV, respectively. A withholding neighborhood of size $d = 3/150$ was used for FCCV.

Before discussing the results of the simulations, a look at a particular realization would be informative. Figure 4 shows r_2 with additive error from an AR(1), where $\phi = 0.6$, $n = 150$, and medium variance as circles. The nonparametric estimates based on the bandwidths selected by OSCV, PI, and FCCV are shown as dash, dot, and solid curves, respectively. As is well known, plug-in methods do not work well in such moderately correlated error cases because the crucial step of second-derivative estimation performs poorly. Thus, the PI estimate is excessively variable. Even though OSCV was not originally developed with correlated errors in mind, it does better than PI, as shown by the more stable estimate (Hart and Lee 2005). Finally, FCCV does the best with the solid curve closely matching r_2 shown in Figure 2. The OCV estimate is not shown because it is very similar to the most erratic PI curve.

Figure 5 illustrates how FCCV affects the error curve estimates in selecting bandwidth for a less correlated sample from r_1. The optimal bandwidth, estimated by minimizing ASE, is indicated by the vertical line. Even in uncorrelated data, OCV is known to select small bandwidths and the problem is exacerbated in the correlated error case. This is demonstrated by the dot–dash CV error curve and selected bandwidth indicated by the solid dot on the curve. OSCV improves the CV curve by omitting the correlated neighbors to the right of the points of estimation, but it still selects too small a bandwidth as indicated by its solid circle. Finally, FCCV comes very close to selecting the estimated optimal bandwidth as indicated by the solid dot. It omits three neighbors on either side of the points of prediction.

Figure 6 and online Appendix Figures 8–10 display Monte Carlo density estimates for the ASE for various combinations of ϕ, and σ with $n = 150$. Our gold standard is the

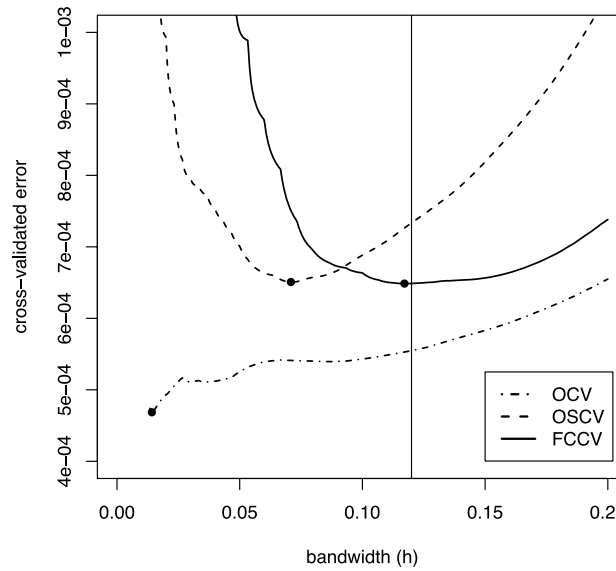

Figure 5. Comparison of cross-validation error curves. OCV is shown as the dot–dash curve whereas OSCV is dash and FCCV is solid. The minimum cross-validated error for each technique is indicated by the solid dots. The vertical line indicates the optimal bandwidth that minimizes ASE for this particular sample which was generated using r_1, $n = 150$, $\phi = 0.3$, and $\sigma = 2^{-9}$. A withholding neighborhood of $d = 3/150$ was used for FCCV.

estimated ASE density for the ASE optimal bandwidth shown in green. The top row corresponds to independent errors ($\phi = 0.0$) where OCV, OSCV, and PI generally compare favorably to the optimal error. In the interest of conserving space, only one of the four sets of density estimates is shown; however, the following comments relate to all four choices of r. Depending on the configuration, FCCV either compares favorably or is only slightly worse, but not dramatically so. The modest loss makes sense in the context of discarding neighbors when the errors are not correlated. For the other rows, FCCV generally dominates the other methods, particularly OCV and PI when the correlation is either low ($\phi = 0.3$) or medium ($\phi = 0.6$). OSCV exhibits more resilience than these two methods when faced with correlated errors, but FCCV still generally delivers smaller ASE, most notably in the moderately ($\phi = 0.6$) to highly correlated ($\phi = 0.9$) noise combinations coupled with moderate to large variability. An interesting feature is the seemingly paradoxical increase in performance for OCV and PI in the highly correlated cases relative to their poor showings under low and medium correlation. This may be due to the difficulty separating the underlying function from long runs of correlated error. That is, no bandwidth may be able to satisfactorily smooth out the correlated errors while preserving the underlying structure of r.

An examination of the bandwidths selected by the four methods relative to the ASE optimal bandwidth provides insight into the ASE density estimates. Table 1 and online Appendix Tables 2–4 summarize the means of the ratios of the global bandwidths estimated by the four techniques to the estimated ASE optimal global bandwidth for r_3 and various combinations of ϕ, n, and σ. The table, unlike the figure, includes $n = 75$. For clarity, $\hat{h}_{ocv}$, $\hat{h}_{oscv}$, $\hat{h}_{pi}$, $\hat{h}_{fccv}$, and $\hat{h}_0$ are used to distinguish the estimates of h, with $\hat{h}_0$

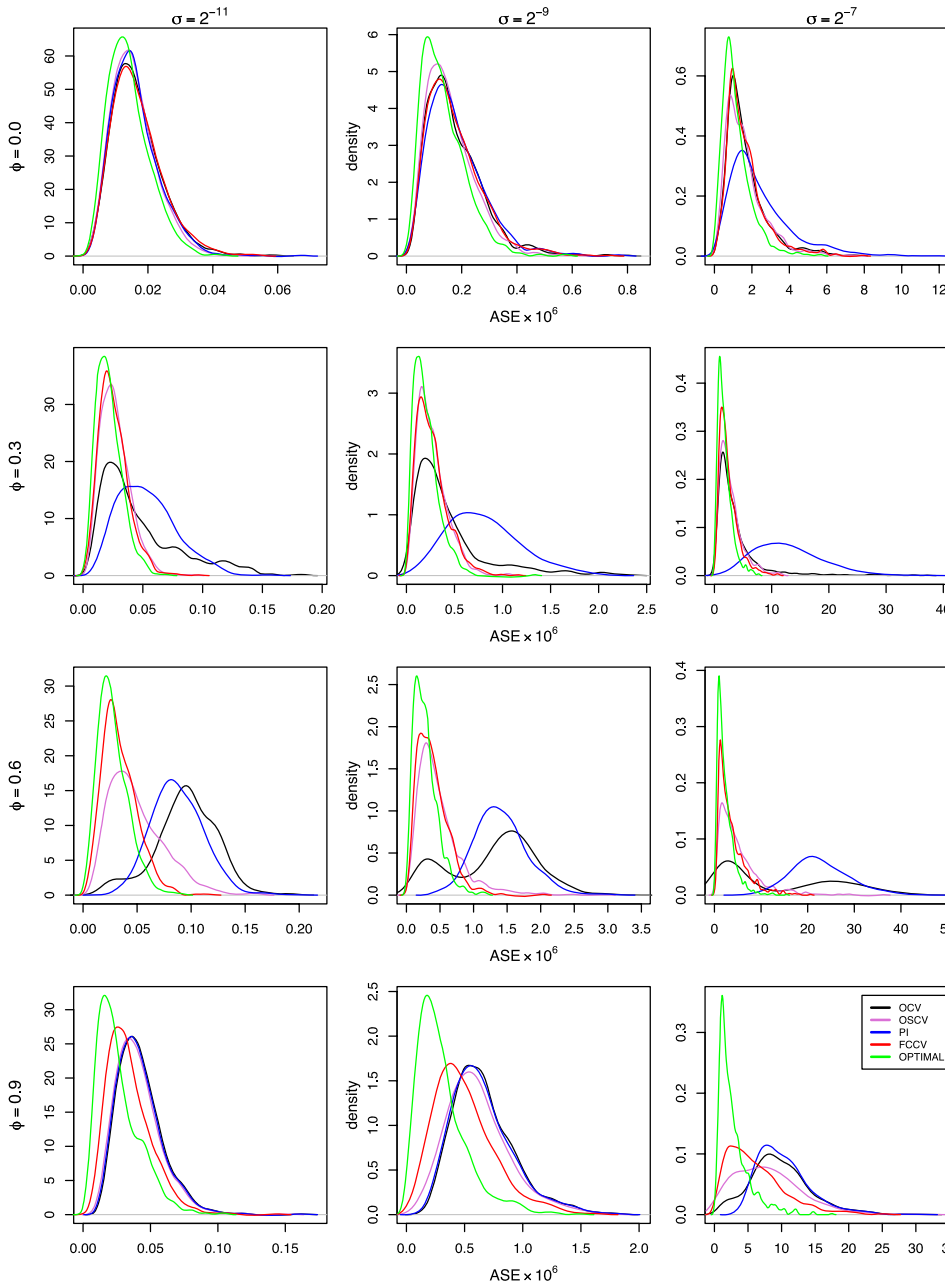

Figure 6. Comparison of the distribution of ASE for r_2 with $n = 150$ of the different methods using different combinations of correlation and noise variance. The rows are indexed by the AR(1) parameter ϕ and the columns by variance parameter σ. The optimal ASE is shown in green, whereas OCV, OSCV, PI, and FCCV are shown in black, purple, blue, and red, respectively. Each distribution was estimated using 1000 realizations.

Table 1. $r_3 = 1.741 \cdot [2x^{10}(1-x)^2 + x^2(1-x)^{10}]$.

			mean[$\hat{h}/\hat{h}_0$](sd[$\hat{h}$])			
ϕ	n	σ	$\hat{h}_{ocv}$	$\hat{h}_{oscv}$	$\hat{h}_{pi}$	$\hat{h}_{fccv}$
0.0	75	2^{-11}	1.06 (0.008)	1.00 (0.003)	1.00 (0.004)	1.99 (0.009)
		2^{-9}	1.05 (0.017)	0.97 (0.007)	0.95 (0.011)	1.60 (0.015)
		2^{-7}	1.28 (0.083)	1.26 (0.085)	0.93 (0.033)	2.62 (0.310)
	150	2^{-11}	1.02 (0.005)	0.97 (0.002)	1.03 (0.003)	1.50 (0.005)
		2^{-9}	1.01 (0.012)	0.95 (0.005)	0.98 (0.007)	1.27 (0.011)
		2^{-7}	1.09 (0.033)	1.00 (0.024)	0.95 (0.022)	1.25 (0.038)
0.3	75	2^{-11}	0.77 (0.007)	0.82 (0.004)	0.78 (0.004)	1.78 (0.009)
		2^{-9}	0.68 (0.020)	0.80 (0.010)	0.61 (0.012)	1.46 (0.017)
		2^{-7}	0.83 (0.110)	1.06 (0.095)	0.39 (0.027)	2.74 (0.336)
	150	2^{-11}	0.58 (0.007)	0.78 (0.003)	0.68 (0.004)	1.32 (0.006)
		2^{-9}	0.52 (0.018)	0.79 (0.009)	0.49 (0.011)	1.18 (0.014)
		2^{-7}	0.67 (0.050)	0.87 (0.040)	0.30 (0.019)	1.27 (0.084)
0.6	75	2^{-11}	0.70 (0.002)	0.66 (0.003)	0.67 (0.002)	1.76 (0.004)
		2^{-9}	0.41 (0.009)	0.55 (0.010)	0.41 (0.006)	1.25 (0.016)
		2^{-7}	0.38 (0.102)	0.73 (0.094)	0.20 (0.010)	2.46 (0.341)
	150	2^{-11}	0.38 (0.002)	0.53 (0.003)	0.45 (0.002)	1.09 (0.006)
		2^{-9}	0.22 (0.005)	0.50 (0.012)	0.26 (0.004)	0.98 (0.017)
		2^{-7}	0.22 (0.043)	0.65 (0.055)	0.13 (0.004)	1.18 (0.126)
0.9	75	2^{-11}	1.39 (0.000)	0.87 (0.002)	0.80 (0.002)	3.47 (0.000)
		2^{-9}	0.60 (0.002)	0.55 (0.003)	0.56 (0.002)	1.51 (0.004)
		2^{-7}	0.26 (0.007)	0.33 (0.021)	0.26 (0.004)	1.42 (0.230)
	150	2^{-11}	0.64 (0.000)	0.54 (0.001)	0.62 (0.001)	1.60 (0.000)
		2^{-9}	0.26 (0.001)	0.31 (0.003)	0.29 (0.002)	0.75 (0.006)
		2^{-7}	0.12 (0.002)	0.19 (0.028)	0.13 (0.002)	0.50 (0.061)

referring to the ASE optimal bandwidth. The last column shows that FCCV selects too large a bandwidth when the errors are independent. In contrast, the other three methods come much closer to matching the ASE optimal bandwidth with OSCV generally coming closest. In the correlated errors cases, FCCV generally comes closer to the ASE optimal bandwidth with the other three approaches generally undersmoothing the data. This is particularly true for PI, which tends to grossly undersmooth when the errors are correlated and have large variance. As a final comment, FCCV is a data-hungry method as the $n = 75$ table entries demonstrate. The sampling rate needs to be high enough to retain sufficient information about the mean function when deleting neighbors.

4. APPLICATION

Figure 7 shows global temperature deviations spanning the years 1880–1987 as circles. As pointed out by Woodward, Bottone, and Gray (1997), the residuals for these data from standard parametric regression procedures are correlated. This is an ideal example in the present context. The same figure also shows three nonparametric regression estimates with

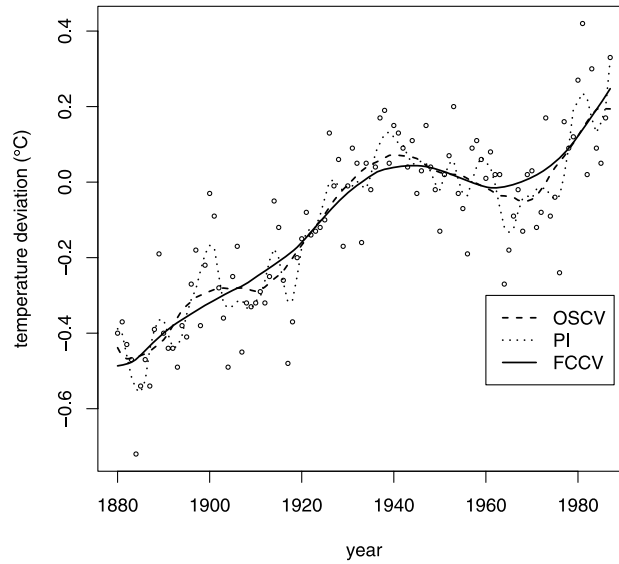

Figure 7. Sample application of OSCV, PI, and FCCV to 108 years of deviations of mean global temperatures from 1880 to 1987 (Hansen and Lebedeff 1987). The dash, dot, and solid curves are smoothed estimates produced by OSCV, PI, and FCCV, respectively. PI produced the smallest bandwidth estimate at 0.026 (2.8 yr) whereas OSCV was in the middle at 0.093 (10.0 yr). FCCV, using a withholding neighborhood of $d = 4/108$, yielded the largest bandwidth estimate of 0.168 (16.8 yr).

global bandwidths selected by OSCV, PI, and FCCV. Similarly to the simulation results, the PI method selects the smallest bandwidth of 2.8 yr, which is not surprising given the serially correlated errors. The PI fit, shown as a dotted curve, is excessively variable tracking small-scale features. OSCV selects a larger bandwidth of 10.0 yr and the corresponding fit reflects this as the much smoother dashed curve. This parallels the simulation results where OSCV performed well in the minimally correlated cases. Using the procedure described earlier for estimating d, $\hat{\phi} = 0.38$ which leads to a estimated withholding neighborhood of $\hat{d} = 4/108$ that minimizes the squared partial bias. Using $\hat{d} = 4/108$, FCCV provides the largest bandwidth of 16.8 yr and the smoothest fit, shown as a solid curve. The code and data for this example are available at the *JCGS* website.

5. CONCLUSIONS

Tuning parameter estimation and model selection when data are correlated present several difficulties. Even grossly misspecified models can appear to perform well in terms of CV error under ordinary cross-validation because highly correlated neighbors retained in the prediction phase essentially impute the withheld data point, unbeknownst to the user. FCCV works well in large, strongly correlated datasets. This occurs because model estimates change very little with a few withheld observations. Misspecified models robbed of correlated neighbors are unlikely to yield good estimates. By assessing the model's ability to extrapolate, FCCV reduces overly optimistic model performance measures.

In the context of global bandwidth selection, FCCV outperforms OCV and PI when errors are correlated. Even though OSCV can perform similarly to FCCV in one dimension when errors are minimally correlated, OSCV does not readily extend to two or higher dimensions. FCCV overcomes this limitation by removing radially defined neighborhoods, which readily adapts to any dimension. Our motivation for FCCV was precisely this difficulty. We found it necessary to find an alternative to existing methods while investigating recursive partitioning of two-dimensional kriging models (Carmack 2004).

Although this new method performs well in correlated datasets, it is less than optimal with uncorrelated data. Also, when errors are correlated, the sampling rate needs to be high enough to capture the underlying functional form when neighbors are discarded. The simulation studies presented in this article excluded fixed-size neighborhoods, which is not necessarily optimal. Provided the covariance structure is known or can be well estimated, better choices of d may be produced. In the special case of an AR(1) error structure, a simple procedure for estimating d was outlined and applied to an example. Further research is necessary to establish a more general data-driven solution and related robustness properties.

SUPPLEMENTAL MATERIALS

Data Sets and Computer Code: The supplemental materials contain the global temperature deviations spanning the years 1880–1987 (global_temperature_deviations.dat), R code for applying FCCV to one dimensional data sets (fccv.R), and script for applying FCCV to the global temperature deviations data (global_temperature_deviations.R). The appendix (fccv_supplemental.pdf) contains additional tables and figures. (supplements.zip)

ACKNOWLEDGMENTS

This study was supported by the VA IDIQ contract number VA549-P-0027 awarded and administered by the Dallas, TX VA Medical Center. This study was also supported by the U.S. Army Medical Research and Material Command cooperative agreement numbers DAMD17-97-2-7025 and DAMD17-01-1-0741 through a consortium agreement with the University of Texas Southwestern Medical Center at Dallas. The content of this article does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred.

[Received March 2007. Revised July 2008.]

REFERENCES

Burman, P., Chow, E., and Nolan, D. (1994), "A Cross-Validatory Method for Dependent Data," *Biometrika*, 81 (2), 351–358.

Carmack, P. (2004), "Recursive Partioning in Spatially Correlated Data," Ph.D. thesis, Dept. of Statistical Science, Southern Methodist University.

Epanechnikov, V. (1969), "Nonparametric Estimates of a Multivariate Probability Density," *Theory of Probability and Applications*, 14, 153–158.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Gasser, T., Kneip, A., and Köhler, W. (1991), "A Flexible and Fast Method for Automatic Smoothing," *Journal of the American Statistical Association*, 86, 643–652.

Geisser, S. (1975), "A Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

Hansen, J., and Lebedeff, S. (1987), "Global Trends of Measured Surface Air Temperature," *Journal of Geophysical Research*, 92, 13345–13372.

Hart, J., and Lee, C. (2005), "Robustness of One-Sided Cross-Validation to Autocorrelation," *Journal of Multivariate Analysis*, 92 (1), 77–96.

Hart, J., and Yi, S. (1998), "One-Sided Cross-Validation," *Journal of the American Statistical Association*, 93 (442), 620–630.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Marron, J. (1986), "Will the Art of Smoothing Ever Become a Science?" in *Function Estimates. Contemporary Mathematics*, Vol. 59, Providence, RI: American Mathematical Society, pp. 169–178.

Racine, J. (2000), "Consistent Cross-Validatory Model-Selection for Dependent Data: *hv*-Block Cross-Validation," *Journal of Econometrics*, 99, 39–61.

Stone, M. (1974), "Cross-Validatory Choice and the Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–133.

van Es, B. (1992), "Asymptotics for Least Squares Cross-Validation Bandwidths in Nonsmooth Cases," *The Annals of Statistics*, 20, 1131–1146.

Woodward, W., Bottone, S., and Gray, H. (1997), "Improved Tests for Trend in Time Series Data," *Journal of Agricultural, Biological, and Environmental Statistics*, 2 (4), 403–416.