Biometrics 74, 1417–1426
December 2018

DOI: 10.1111/biom.1290

Varying-Coefficient Semiparametric Model Averaging Prediction

Jialiang Li,[1,2,3,*] Xiaochao Xia,[1,4,] Weng Kee Wong,[5,***] and David Nott[1,****]**

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore
[2]Duke University-NUS Graduate Medical School, Singapore
[3]Singapore Eye Research Institute, Singapore
[4]College of Science, Huazhong Agricultural University, Wuhan, China
[5]Department of Biostatistics, University of California, Los Angeles, U.S.A.
[*]*email:* stalj@nus.edu.sg
[**]*email:* xia_xiao_chao@126.com
[***]*email:* wkwong@ucla.edu
[****]*email:* standj@nus.edu.sg

Summary. Forecasting and predictive inference are fundamental data analysis tasks. Most studies employ parametric approaches making strong assumptions about the data generating process. On the other hand, while nonparametric models are applied, it is sometimes found in situations involving low signal to noise ratios or large numbers of covariates that their performance is unsatisfactory. We propose a new varying-coefficient semiparametric model averaging prediction (VC-SMAP) approach to analyze large data sets with abundant covariates. Performance of the procedure is investigated with numerical examples. Even though model averaging has been extensively investigated in the literature, very few authors have considered averaging a set of semiparametric models. Our proposed model averaging approach provides more flexibility than parametric methods, while being more stable and easily implemented than fully multivariate nonparametric varying-coefficient models. We supply numerical evidence to justify the effectiveness of our methodology.

Key words: Model averaging; Prediction error; Quadratic programming; Semiparametric estimation; Varying coefficient model.

1. Introduction

Forecasting and predictive inference are fundamental tasks for economic and medical data analysis (Clements and Hendry, 1998; Chatfield, 2001). Parametric methods which make strong assumptions dominate practical applications, but there is no reason why real life data generating mechanisms should obey common parametric assumptions such as linearity. In contrast, nonparametric and semiparametric models may acknowledge the existence of more complex and realistic functional covariate effects (Wu and Zhang, 2004). To incorporate multiple predictor variables, many multivariate nonparametric models are available (Matzner et al., 1998; De Gooijer and Gannoun, 2000; Fan and Yao, 2003). However, empirical studies indicate the predictive performance of multi-dimensional nonparametric models may not be satisfactory in applications involving low signal to noise ratios and large numbers of covariates. In this article, we propose a new semiparametric model average prediction (SMAP) approach to analyze large data sets with abundant covariates and investigate its predictive performance in numerical examples. This approach involves fitting individual partly linear varying-coefficient nonparametric models and combining them using a linear weighting structure. The procedure may provide more flexible predictive inference than a parametric model while being more stable than a fully nonparametric approach.

When numerous candidate models are available for prediction, one must consider appropriate approaches for dealing with uncertainty about the model. One popular approach is to employ a *model selector* and identify a single optimal model from all candidates. Traditional approaches for model selection include subset selection, regularization (Fan and Li, 2001) and dimension reduction (Jolliffe, 1986; Ma and Zhu, 2013). These approaches have been extended to incorporate high-dimensional covariates in the recent literature, e.g., Chen et al. (2010) and Fan et al. (2011). Since a model selector yields only one final model, useful information may be lost when variables absent from the final model are also relevant to predicting the outcome.

As an attractive alternative, *model averaging* may include a set of models and make prediction via a weighted average from all the models (Buckland et al., 1997; Yang, 2001, 2003; Hansen, 2007, 2008). Most early work on model averaging was done in a Bayesian framework (Hoeting et al., 1999), and provided good solutions to many practical problems given appropriate prior choices and computational methodology. Hjort and Claeskens (2003) systematically discussed the advantages of weighting estimators across models, proposed a general framework for frequentist model averaging. Following their work, model averaging has been investigated in, for example, semiparametric models (Claeskens and Carrol, 2007), generalized partially linear additive models (Zhang and Liang, 2011), mixing partially linear regression models (Liu and Yang, 2013), high dimensional factor-augmented linear regression (Chen and Hansen, 2015), and ultra-high dimensional nonparametric additive models for time series (Chen et al., 2017). Some theoreticians have argued that model averaging shares all the good properties of model selection under general settings (Giraud, 2015).

Almost all previous authors focused on averaging a set of parameterized models such as linear regression models. Such simple models are easy to interpret and widely accepted by scientific researchers and business analysts. However, to provide accurate characterization of the relationship between the response and the predictors, it may be more sensible to consider nonparametric models with less structural restrictions. Li et al. (2015) developed a flexible prediction approach by averaging a set of nonparametric models obtained from local constant smoothing. Their numerical works suggest that the nonparametric averaging approach may perform better than the traditional parametric averaging approach. Our current development extends Li et al. (2015) who addressed a series of univariate sub-models.

We aim at predicting the response variable Y by constructing a varying-coefficient submodel-based prediction from complicated data. There is a large literature concerning varying-coefficient models such as Fan and Huang (2005) among others. Fitting a model with multiple varying coefficients is as difficult as fitting other multivariate nonparametric models. In particular, we need to select the bandwidth for functional estimates when adopting the familiar local polynomial regression. Different covariates may actually require different degrees of smoothness. Yet most existing programs to fit such models allow only a single bandwidth for all the estimated varying coefficients.

In this article, we propose varying-coefficient semiparametric model averaging prediction (VC-SMAP) which works much more satisfactorily than the ordinary varying coefficient model. In particular, now, we consider only one varying coefficient in each submodel and thus the bandwidth selected for such a sub-model is itself optimal. Each submodel only involves one nonparametric component and thus can be easily fitted using univariate smoothing. In each model, we also adjust all other covariates linearly to approximate the conditional relationship more accurately. The overall prediction is to average individual submodel predictions. The SMAP procedure is motivated by approximating the Gibbs mixing of estimators (Giraud, 2015). We stress that there is no new model introduced in this article. *All models are wrong.* However, by combining useful submodels in an effective manner, we may achieve improved prediction accuracy. We carry out extensive simulations to investigate the proposed methods in this article. Two biomedical data sets are analyzed to further illustrate our methodology.

2. Method and Estimation

Suppose that we have sample data $\{(U_i, \mathbf{X}_i, Y_i); i = 1, \ldots, n\}$, consisting of n independent copies of $(U, \mathbf{X}, Y)$, where $\mathbf{X} = (X_1, \ldots, X_p)^T$ is a p-vector of covariates, U is an index variable and Y is the response variable. We assume $X_1 = 1$ in the following presentation. We write $\mathbf{Z} = (\mathbf{X}^T, U)^T = (Z_1, \ldots, Z_{p+1})^T$. It is well known that, when the dimension of $\mathbf{Z}$ (or $\mathbf{X}$) is high, modeling the conditional mean function $m(\mathbf{Z}) = E(Y|\mathbf{Z})$ by purely multivariate nonparametric methods without any structure specification is not practical due to the curse of dimensionality. There is a large literature concerned with approximating the regression function, $m(\mathbf{Z})$, by an affine combination of low dimensional semiparametric regression functions. Earlier authors considered additive models, partly linear additive models, varying-coefficient models (Cai et al., 2000) and partly linear varying-coefficient models (Fan and Huang, 2005), among many other choices. In practice, using a specified model with fixed regression structure may lead to very poor prediction because of the risk of misspecification. To rectify this problem, we adopt the model averaging principle (Hansen, 2007) in this article.

Most authors use parametric models when applying the model averaging method. Li et al. (2015) first proposed to approximate $m(\mathbf{Z})$ by a weighted average of nonparametric regression models. Although their resulting semiparametric model average prediction (SMAP) on the response allows nonlinear structure for predictors, each of the models is marginal and completely ignores the presence of other factors. This kind of prediction may be insufficient since it ignores the potentially strong confounding effects among predictors. In addition, there might be interaction between the predictors, which is not uncommon in practice, especially for high dimensional data.

In this manuscript, we propose an alternative strategy to approximate $m(\mathbf{Z})$ by a class of semiparametric regression functions, which covers varying-coefficient regression, in the framework of model averaging. Specifically, we seek weights to minimize the following

$$E\left\{\left(Y - \sum_{j=1}^{p} w_j m_j\right)^2\right\}, \tag{1}$$

where $m_j = \alpha_j(U)X_j + \sum_{k\neq j}^{p} \beta_{jk}X_k$, $j = 1, \ldots, p$, $\mathbf{w} = (w_1, \ldots, w_p)^\tau \in H$ and $H = \left\{\mathbf{w} : \sum_{k=1}^{p} w_k = 1, w_k \geq 0\right\}$. In this case $\alpha_j(U)$ is the varying coefficient for the jth covariate X_j while the β_{jk}s are constant coefficients for X_k, $k \neq j$. In fact, the jth sub-model is equivalent to fitting the following regression problem

$$Y = \alpha_j(U)X_j + \sum_{k\neq j}^{p} \beta_{jk}X_k + \varepsilon, \tag{2}$$

where ε is a random error. Such a model allows discrete as well as continuous covariates to be considered while only continuous terms are allowed under Li et al. (2015)s approach. On the other hand, the overall combined model can be rewritten as a fully varying-coefficient model

$$m(\mathbf{Z}) = \sum_{j=1}^{p} a_j(U)X_j, \tag{3}$$

where $a_j(U) = w_j\alpha_j(U) + \sum_{k\neq j} w_k\beta_{kj}$. However, directly fitting such a model to obtain the estimates of $\alpha_j(\cdot)$ and β_{jk}

is infeasible since (3) is not identifiable without additional conditions. For the purpose of predicting Y, perhaps attaining good estimates for unstructured a_j would be sufficient. However, such a standard varying-coefficient model may not work as well as our VC-SMAP. There are two possible reasons. First, some X_j may have constant coefficient and should be modeled in this way. When forcing all coefficients to be functionals, the usual varying-coefficient model may overfit, as will be seen in our simulation results. The weights in the proposed VC-SMAP adjust the relative importance of the varying vs. constant coefficients for the same predictor and hence the final prediction is more robust against model misspecification. Secondly, we are smoothing each varying-coefficient function separately in our procedure and thus do not require all the a_j's to be smoothed with the same bandwidth. These considerations naturally improve the prediction performance of VC-SMAP over existing semiparametric single-model-based prediction.

In this study, what we are most interested in is to predict the response. It is necessary to accurately estimate m_j and the model average weights. Without loss of generality, we assume that the index variable U has been scaled to $[0,1]$ and X_j, $j=2,\ldots,p$, are all standardized to be of mean 0 and variance 1. In the first step, for a fixed j, we estimate m_j by the profile least squares method in Fan and Huang (2005). To simplify the presentation, we use matrix notation and write $\mathcal{Y}=(Y_1,\ldots,Y_n)^T$, $\mathcal{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^T$, $\mathbf{X}_i=(X_{i1},\ldots,X_{ip})^T$ and $\mathbf{Z}_i=(\mathbf{X}_i^T,U_i)^T$, $i=1,\ldots,n$. Let $\mathcal{X}_{(j)}$ be a sub-matrix of $\mathcal{X}$ without the jth column. Write $\mathcal{W}(u)=\mathrm{diag}\{\mathrm{K_h(U_1-u)},\ldots,\mathrm{K_h(U_n-u)}\}$, where $K_h(\cdot)=K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth,

$$\mathbf{D}_j(u)=\begin{pmatrix} X_{1j} & X_{1j}(U_1-u)/h \\ \vdots & \vdots \\ X_{nj} & X_{nj}(U_n-u)/h \end{pmatrix}$$

and

$$\mathbf{S}_j=\begin{pmatrix} (X_{1j},0)\{\mathbf{D}_j^T(U_1)\mathbf{W}(U_1)\mathbf{D}_j(U_1)\}^{-1}\mathbf{D}_j^T(U_1)\mathcal{W}(U_1) \\ \vdots \\ (X_{nj},0)\{\mathbf{D}_j^T(U_n)\mathcal{W}(U_n)\mathbf{D}_j(U_n)\}^{-1}\mathbf{D}_j^T(U_n)\mathcal{W}(U_n) \end{pmatrix}.$$

Let $\boldsymbol{\beta}_{(j)}=(\beta_{j1},\ldots,\beta_{j(j-1)},\beta_{j(j+1)},\ldots,\beta_{jp})^T$. Using the above notation, the profile least squares estimate of $\boldsymbol{\beta}_{(j)}$ is given by

$$\widehat{\boldsymbol{\beta}}_{(j)}=\{\mathcal{X}_{(j)}^T(\mathbf{I}-\mathbf{S}_j)^T(\mathbf{I}-\mathbf{S}_j)\mathcal{X}_{(j)}\}^{-1}\mathcal{X}_{(j)}^T(\mathbf{I}-\mathbf{S}_j)^T\mathcal{Y} \quad (4)$$

and the local linear estimate of $\alpha_j(u)$ is given by

$$\widehat{\alpha}_j(u)=(1,0)\{\mathbf{D}_j^T(u)\mathcal{W}(u)\mathbf{D}_j(u)\}^{-1}\mathbf{D}_j^T(u)\mathcal{W}(u)\{\mathcal{Y}-\mathcal{X}_{(j)}\widehat{\boldsymbol{\beta}}_{(j)}\}. \quad (5)$$

Therefore, we obtain the jth model-based prediction $\mathcal{M}_j=(m_j(\mathbf{Z}_1),\ldots,m_j(\mathbf{Z}_n))^T$ for the n samples as $\widehat{\mathcal{M}}_j=\mathbf{S}_j\{\mathcal{Y}-\mathcal{X}_{(j)}\widehat{\boldsymbol{\beta}}_{(j)}\}+\mathcal{X}_{(j)}\widehat{\boldsymbol{\beta}}_{(j)}=\mathcal{A}_j\mathcal{Y}$, where $\mathcal{A}_j=\mathbf{S}_j+(\mathbf{I}-\mathbf{S}_j)\mathcal{X}_{(j)}\{\mathcal{X}_{(j)}^T(\mathbf{I}-\mathbf{S}_j)^T(\mathbf{I}-\mathbf{S}_j)\mathcal{X}_{(j)}\}^{-1}\mathcal{X}_{(j)}^T(\mathbf{I}-\mathbf{S}_j)^T$.

In the second step, we estimate the weights $\mathbf{w}$. We first consider estimation without restricting the weights to be in space H. The optimal weight estimator can be obtained via minimizing the least squares function $Q(\mathbf{w})=\|\mathcal{Y}-\sum_{k=1}^p w_k\widehat{\mathcal{M}}_k\|^2$. Let $\widehat{\mathcal{M}}=(\widehat{\mathcal{M}}_1,\ldots,\widehat{\mathcal{M}}_p)$. We have a closed-form solution $\widehat{\mathbf{w}}=(\widehat{\mathcal{M}}^T\widehat{\mathcal{M}})^{-1}\widehat{\mathcal{M}}^T\mathcal{Y}$.

For constrained estimation, we have to use a quadratic programming technique to obtain $\widehat{\mathbf{w}}$. For the optimization problem under the constraint H, one may apply the commonly used "interior-point-convex" algorithm (Anna and Gondzio, 1999), which is implemented in familiar software: for example, the `quadprog` package in `R`, the `quadprog` command in `MATLAB` and the `qprog` command in `GAUSS`. Finally, we output the VC-SMAP for the mean response predicted at a future observation $\mathbf{z}=(\mathbf{x}^T,u)^T$ as

$$\widehat{m}(\mathbf{z})=\sum_{j=1}^p\widehat{w}_j\{x_j\widehat{\alpha}_j(u)+\mathbf{x}_{(j)}^T\widehat{\boldsymbol{\beta}}_{(j)}\}. \quad (6)$$

2.1. *Theoretical Issues*

For varying-coefficient partly linear models, Fan and Huang (2005) developed the asymptotic properties of the parametric and nonparametric components. When the assumed model is indeed the same as the underlying data generating mechanism, their results can be directly applied. However, we do not impose any *true* model in this manuscript. Therefore, some theoretical issues must be clarified.

Suppose the true functional relationship is $m_0(\mathbf{Z})=E(Y|\mathbf{Z})$. The prediction error for a function m is usually framed as the population risk under the L_2 loss

$$R(m)=E\{Y-m(\mathbf{Z})\}^2$$

and the minimizer is $m=m_0$. Let the corresponding empirical risk be

$$R_n(m)=n^{-1}\sum_{i=1}^n\{Y_i-m(\mathbf{Z}_i)\}^2$$

and we have $R(m)=ER_n(m)$ where the expectation is conditional given $\mathbf{Z}_i$, $i=1,\ldots,n$. The excess risk $R(m)-R(m_0)$ is equal to $\|m-m_0\|^2$ where $\|\cdot\|$ is the L_2 norm.

Now we define the function space where we search for m. If we assume a varying-coefficient for the jth covariate, the functional space is $F_j=\{m:m(\mathbf{Z})=\alpha_j(U)X_j+\sum_{k\neq j}^p\beta_{jk}X_k\}$. Let $\widehat{m}_j=\arg\min_{m\in F_j}R_n(m)$ and let $m_{j*}=\arg\min_{m\in F_j}R(m)$. We can easily show that

$$\|\widehat{m}_j-m_0\|^2=\|\widehat{m}_j-m_{j*}\|^2+\|m_{j*}-m_0\|^2,$$

where the first term is the estimation error and the second term is the approximation error. When the model is misspecified, that is, $m_0\notin F_j$, the excess risk is always positive. Thus, a theoretical condition for a valid varying-coefficient modeling is to require $\delta_j = \|m_{j*} - m_0\|^2$ to be at a negligible order. The so-called *margin condition* may be a useful technical tool to guarantee this requirement. We refer to Lemma 2.1 in van de Geer (2007) which may be used to provide a probability bound on $ER(\widehat{m}_j) - R(m_0)$.

Another theoretical issue of interest is the asymptotic behavior of the estimated weights $\widehat{\mathbf{w}}$. Li et al. (2015) considered such an issue and derived the asymptotic normality for the weight distribution under regularity conditions. It is not hard to modify their derivation to show the asymptotic distribution for the estimated weights in this article. Similar to their development, the large sample theory for the weight estimates is unaffected by model misspecification. An extension from local constant smoother to local polynomial smoother is available in Huang and Li (2018). However, such a large sample result is obtained without acknowledging the fact that $\mathbf{w} \in H$. For constrained minimizers resulted from quadratic programming the technical argument for consistency and asymptotic distribution may be more complicated. In the numerical studies of this article, we carry out a bootstrap procedure to estimate the variability of the estimated weights and thus facilitate the inference. The detailed bootstrap procedure is given in Web Appendix of Supplementary Materials.

2.2. *Computational Issues*

We have provided the estimation procedures in earlier section. A few technical details warrant readers' attention and we carry out a discussion in this subsection.

First, the semiparametric estimation for varying-coefficient models is a straight forward computation well developed for applications. We choose Epanechnikov kernel in this manuscript for its relative efficiency over other kernel functions. The bandwidth is selected by the cross-validation method. We note that bandwidth selection is an important step for nonparametric function estimation. When fitting a model with multiple nonparametric functions, one usually has to assume the same bandwidth for all the function estimates to facilitate computation. Such a convenient choice may not be the most appropriate when different functions require different degrees of smoothing. Allowing different function estimates to have different bandwidth is not practically feasible since that can greatly increase the computation burden and also cause unstable estimation. In contrast, in our model averaging procedure, every sub-model only contains one nonparametric function and thus only involves selection of one bandwidth. The final SMAP combines these individual model estimates and effectively provide the most appropriate smoothing for different functional effect estimation without seriously increasing the computational cost. Our numerical work indicates that this method works very well in a wide range of scenarios.

Secondly, the estimation of optimal weights is carried out using a least squares approach. The default least squares estimation does not acknowledge the fact that the weights must be non-negative and the sum of the weights must exactly equal one. In order to satisfy these constraints on parameter estimation, we may adopt the well-known quadratic programming techniques available in all sorts of statistical packages such as `R`. Quadratic programming is a special type of mathematical optimization problem—specifically, the problem of optimizing (minimizing or maximizing) a quadratic function of several variables subject to linear constraints on these variables. It is a particular type of nonlinear programming. For general problems a variety of methods are commonly used, including interior point, active set, augmented Lagrangian, conjugate gradient, gradient projection, and extensions of the simplex algorithm. In all numerical studies of this article, we consider two versions of VC-SMAP: one without constraint and one with constraint. For the constrained estimation, we use the `R` function `solve.QP` from the package `quadprog` developed by Berwin Turlach and Andreas Weingessel. This routine basically implements the dual method of Goldfarb and Idnani (1982, 1983).

One very relevant question is whether we can incorporate more than one index variables in the VC-SMAP procedure. In principle for q index variables, we may compute q sets of sub-models using the Fan–Huang estimators and then evaluate the average of all $q \times p$ models. Our methods can thus be straight forwardly extended. This procedure might face a computational challenge when q and p are large. Some regularization steps must be included in that case. In the numerical examples considered in this article, we confine our investigation to a single index variable which is usually determined from experience and agreed by data analysts to be the most likely to be interacting with other variables.

3. Simulation Studies

In this section, we examine the performance of VC-SMAP via numerical examples. To compare with the existing model-based prediction methods, we investigate the following methods in all simulated data sets: (i) VC-SMAP without constraints, (ii) VC-SMAP with weight constraints (denoted as VC-SMAPc), (iii) the model averaging method by Li et al. (2015) (denoted as SMAP), (iv) SMAP with weight constraints (denoted as SMAPc), (v) the prediction method using varying-coefficient model (denoted as VCP). Moreover, we apply these approaches under two cases with model misspecification, where (vi) we use an incorrect index variable (denoted as VC-SMAP(MI), VC-SMAPc(MI), and VCP(MI)); and (vii) we exclude some relevant variables in the sub-models (denoted as VC-SMAP(MC), VC-SMAPc(MC), SMAP(MC), SMAPc(MC), and VCP(MC)). We compare these approaches to the oracle method with the true model form known.

To evaluate the prediction performance for proposed procedures, we calculate the bias $\sum_{i\in\mathcal{I}}(\widehat{Y}_i - Y_i)/|\mathcal{I}|$, the mean absolute prediction error (MAPE) $\sum_{i\in\mathcal{I}}|Y_i - \widehat{Y}_i|/|\mathcal{I}|$, and the mean squared prediction error (MSPE) $\sum_{i\in\mathcal{I}}(Y_i - \widehat{Y}_i)^2/|\mathcal{I}|$, where $\mathcal{I}$ represents the index set of either the training sample or the test sample. We report the means and the standard deviations of bias, MAPE and MSPE over 1000 simulation.

EXAMPLE 1. *We generate data from* $Y = g_1(U)X_1 + g_2(U)X_2 + g_3(U)X_3 + 4X_4 + 4X_5 + 4X_6 + \varepsilon$, *where* $g_1(x) = \cos(2\pi x)$, $g_2(x) = (2 + x^2)/(1 + x^2)$ *and* $g_3(x) = \{2\exp(-0.5x^2)\}/\{\exp(-0.5x^2) + 1\}$, $\mathbf{X} = (X_1, \ldots, X_6)^T \sim N(0, \Sigma)$ *with* $\Sigma = (0.5^{|j-k|})_{j,k=1}^6$, $U \sim \mathrm{Unif}(0, 1)$. *We simulate* *the noise from five different distributions: $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 1, 2,$ and 4, respectively, $t(2)/5$ as well as a mixture normal distribution $0.4N(-3, 1) + 0.6N(2, 1)$. Each sample is of size $n = n_{est} + n_{grid}$, consisting of a train set of size n_{est} and a test set of size $n_{grid} = 50$. We compare the performance with sample sizes $n = 100$ and 200. In VC-SMAP, we use U as the index variable and all X_j, $j = 1, \ldots, 6$ to construct the submodels. In misspecification case (vi), we consider using X_1 as the index variable for all the submodels. In misspecification case (vii), we build all the models without X_6. The results are displayed in Table 1. For space consideration, the results regarding misspecification case (vii) and two noise cases: $N(0, 2^2)$ and $N(0, 4^2)$ are retained in Tables S1 and S2 in Web Appendix A of Supplementary Materials.*

Table 1 shows that the in-sample performance of the VCP method is the best in most simulation situations. However, for the out-of-sample performance, we can see that the proposed VC-SMAPc method is better than all other methods as it attains the smallest prediction errors, and thus is closest to the oracle prediction method. In fact VCP forces all coefficients to be functions and thus leads to overfitting for the training set. Our VC-SMAP and VC-SMAPc perform much better than SMAP and SMAPc in all cases, including the two misspecified settings. This is not surprising because the underlying data generation mechanism is not in an additive form. When there is nonlinear interaction terms as in this example, the VC-SMAP procedures may provide more accurate results than the ordinary SMAP procedures.

We have implemented extensive simulation studies under many other data generating settings and the numerical results are in Web Appendix A of Supplementary Materials. In summary, our proposed methods always perform very well, relative to other existing methods. We may attempt to provide an empirical answer on when and why our proposed prediction method would work and outperform the existing approaches. In practice, when the true data generating mechanism is very complicated and the model form cannot be easily decided without prior experience or preliminary numerical studies, one usually has to adopt working models to make the prediction. VC-SMAP approaches will be relatively more robust against model mis-specification. One can see easily that this method does not require a rigid designation of a true joint model but integrates a number of possible sub-models. The plausibility of each sub-model is then evaluated by its weight in the averaging step. This flexible approach avoids making a fixed parametric or nonparametric model assumption and thus can yield prediction results closer to the real data, as witnessed in our numerical studies.

4. Applications

4.1. *New Zealand Workforce Study*

We apply our proposal to a cross-sectional data set of a workforce company, plus another health survey, in New Zealand during the 1990s. The data were collected from a confidential self-administered questionnaire for a large observational study conducted during 1992–1993 and included physical, lifestyle, and psychological variables. More details of the study can be found in MacMahon et al. (1995). The data set `xs.nz` is available in R package `VGAMdata`. Our primary research aim is to construct accurate predictions of the response variable, body mass index (BMI), defined as the weight (kg) divided by the squared height (m) (Yee, 2015). Before employing our prediction methods, we first clean the data by removing observations with missing values and extreme outliers (Iglewics and Hoaglin, 1993). We further exclude variables that do not seem to affect BMI based on a preliminary exploratory analysis.

After removing missing cases, we retain 3765 observations with 12 covariates for the following analysis, of which there are 7 continuous variables ("age," "sbp," "dbp," "cholest," "drinkmaxday," "feethour," "sleep") and 5 binary variables ("sex," "diabetes," "hypertension," "acne," "nervous"). The marginal relationships between BMI and 12 predictors are plotted in Figure S1 of Web Appendix B in the Supplementary Web Materials. It is clear that the dependence pattern may not be linear from eyeballing the plots. A direct application of a linear model to this data suggest that all the estimated coefficients except that for "sbp" are significant at level 0.05.

Because age is well-known for its interactive effects with other variables, we choose it to be the index variable U. All of the remaining predictors are included as covariates $\mathbf{X}$ for VC-SMAP, VC-SMAPc, and VCP. BMI is log transformed and all the continuous covariates are standardized to have mean zero and variance one. To evaluate the predictive performance of various methods, we randomly split the data set into two equally sized sets for training and validation. We report the in-sample performance and the out-of-sample performance in terms of MSPE and MAPE, as well as their standard deviations over 100 random partitions. Corresponding results are summarized in Table 2. For in-sample performance, we can see that VCP performs best followed by VC-SMAP, VC-SMAPc, linear model prediction (LMP), and SMAPc. For out-of-sample performance, on the other hand, one can see clearly that VC-SMAP, VC-SMAPc and LMP are uniformly better than the others (SMAP, SMAPc and VCP). Our proposal VC-SMAPc performs better than LMP, although there is no large difference between VC-SMAP, VC-SMAPc and LMP. The plot of predicted BMI by VC-SMAP and VC-SMAPc approaches versus the VCP approach is showed in Figure 1.

To examine the estimated weights, we compare four model averaging methods in Table 3. The weights for the constrained prediction methods (VC-SMAPc and SMAPc) are relatively sparse with much smaller standard deviation. We note that the interpretation of weight coefficients in VC-SMAP is quite distinct from that of ordinary regression coefficients. For example, in the constrained VC-SMAP, only the weights for sub-models 1, 5, and 6 are positive. This would still suggest all variables be used to predict the BMI response; however, it is best that we choose nonparametric function for the effects of age and varying-coefficients for drinkmaxday and feethour, and use constant coefficients for all other variables. In other words, our VC-SMAP method may serve as a new method to identify the nonparametric and parametric components of the predictive model.

The estimated varying-coefficients multiplied by the corresponding weights from VC-SMAP and VC-SMAPc are displayed in Figure 2. One may notice that the curves are of different degrees of smoothness (with different bandwidths) and this is achieved with our procedure easily. The functions

Table 1

Simulation results for Example 1 with three noise distributions: (a1) $\varepsilon \sim N(0,1)$, (a2) $\varepsilon \sim t(2)/5$ and (a3) $\varepsilon \sim 0.4N(-3,1)+0.6N(2,1)$, where the Bias, MSPE, MAPE and their standard deviations (in parenthesis) are computed over 1000 replications

			In-sample error			Out-of-sample error		
ε	n_{est}	Method	Bias	MSPE	MAPE	Bias	MSPE	MAPE
(a1)	100	Oracle	0.001 (0.09)	0.79 (0.14)	0.70 (0.07)	−0.005 (0.16)	1.23 (0.29)	0.88 (0.10)
		VC-SMAP	−0.001 (0.09)	0.83 (0.13)	0.73 (0.06)	−0.002 (0.16)	1.27 (0.32)	0.89 (0.11)
		VC-SMAPc	0.000 (0.10)	0.89 (0.13)	0.75 (0.06)	−0.001 (0.16)	**1.19** (0.28)	**0.87** (0.11)
		SMAP	−0.012 (0.87)	3.88 (1.37)	1.54 (0.27)	0.002 (1.0)	6.02 (2.85)	1.87 (0.40)
		SMAPc	−0.052 (1.89)	8.33 (5.60)	2.25 (0.80)	−0.039 (2.0)	9.85 (6.52)	2.44 (0.83)
		VCP	−0.001 (0.08)	**0.69** (0.13)	**0.65** (0.07)	−0.001 (0.17)	1.40 (0.37)	0.93 (0.12)
		VC-SMAP(MI)	−0.001 (0.05)	1.09 (0.21)	0.83 (0.08)	0.001 (0.29)	3.36 (5.47)	1.15 (0.23)
		VC-SMAPc(MI)	0.000 (0.05)	1.14 (0.20)	0.85 (0.08)	0.003 (0.24)	2.27 (2.64)	1.07 (0.16)
		VCP(MI)	−0.001 (0.05)	0.88 (0.17)	0.73 (0.08)	0.009 (0.34)	4.13 (4.05)	1.28 (0.26)
	200	Oracle	0.000 (0.07)	0.88 (0.10)	0.75 (0.05)	0.005 (0.15)	**1.11** (0.23)	**0.84** (0.09)
		VC-SMAP	−0.002 (0.07)	0.93 (0.10)	0.77 (0.05)	0.003 (0.16)	1.14 (0.24)	0.85 (0.09)
		VC-SMAPc	−0.001 (0.07)	0.97 (0.11)	0.78 (0.05)	0.001 (0.16)	**1.11** (0.22)	**0.84** (0.09)
		SMAP	0.005 (0.73)	3.12 (0.95)	1.38 (0.20)	0.003 (0.84)	4.08 (1.62)	1.55 (0.29)
		SMAPc	0.008 (1.67)	9.47 (4.76)	2.36 (0.64)	−0.005 (1.73)	9.63 (5.14)	2.39 (0.67)
		VCP	−0.001 (0.07)	**0.83** (0.11)	**0.72** (0.05)	0.006 (0.16)	1.19 (0.26)	0.87 (0.10)
		VC-SMAP(MI)	0.000 (0.04)	1.26 (0.15)	0.89 (0.06)	0.013 (0.25)	2.98 (9.59)	1.07 (0.21)
		VC-SMAPc(MI)	0.000 (0.04)	1.29 (0.15)	0.90 (0.06)	0.010 (0.22)	2.11 (5.22)	1.03 (0.15)
		VCP(MI)	0.001 (0.04)	1.08 (0.13)	0.82 (0.05)	−0.009 (0.26)	3.24 (4.25)	1.16 (0.20)
(a2)	100	Oracle	0.002 (0.07)	0.50 (6.23)	0.28 (0.15)	0.000 (0.10)	**0.49** (1.12)	**0.38** (0.13)
		VC-SMAP	−0.001 (0.07)	0.51 (3.30)	0.33 (0.19)	0.005 (0.15)	1.08 (17.10)	0.40 (0.21)
		VC-SMAPc	−0.001 (0.08)	0.57 (3.66)	0.34 (0.17)	0.003 (0.15)	1.01 (17.04)	0.39 (0.18)
		SMAP	0.001 (0.88)	3.73 (9.42)	1.42 (0.36)	0.000 (1.06)	5.39 (3.13)	1.73 (0.44)
		SMAPc	−0.013 (1.88)	8.21 (10.98)	2.16 (0.87)	−0.020 (1.97)	9.16 (6.69)	2.32 (0.87)
		VCP	−0.001 (0.06)	**0.42** (2.96)	**0.26** (0.16)	0.005 (0.15)	1.17 (17.21)	0.43 (0.21)
		VC-SMAP(MI)	0.000 (0.04)	0.75 (5.59)	0.53 (0.20)	−0.007 (0.27)	3.07 (12.12)	0.80 (0.29)
		VC-SMAPc(MI)	0.001 (0.04)	0.80 (6.02)	0.54 (0.16)	−0.008 (0.23)	1.92 (7.43)	0.73 (0.22)
		VCP(MI)	0.000 (0.04)	0.49 (1.09)	0.45 (0.10)	0.010 (0.27)	3.82 (30.29)	0.89 (0.36)
	200	Oracle	0.001 (0.05)	**0.55** (4.90)	0.29 (0.09)	−0.001 (0.10)	**0.46** (1.23)	**0.34** (0.10)
		VC-SMAP	0.001 (0.06)	0.66 (6.35)	0.33 (0.14)	−0.007 (0.10)	0.49 (1.66)	0.36 (0.12)
		VC-SMAPc	0.001 (0.06)	0.69 (6.45)	0.34 (0.12)	−0.008 (0.09)	0.48 (1.65)	0.36 (0.11)
		SMAP	0.001 (0.71)	2.64 (1.84)	1.20 (0.23)	−0.004 (0.81)	3.35 (2.34)	1.35 (0.30)
		SMAPc	0.018 (1.60)	9.07 (4.94)	2.26 (0.67)	0.028 (1.64)	8.94 (5.38)	2.27 (0.68)
		VCP	0.001 (0.06)	0.60 (6.27)	**0.29** (0.11)	−0.006 (0.10)	0.51 (1.73)	0.36 (0.11)
		VC-SMAP(MI)	0.001 (0.03)	0.85 (2.46)	0.58 (0.10)	0.003 (0.24)	2.42 (16.42)	0.74 (0.21)
		VC-SMAPc(MI)	0.001 (0.03)	0.88 (2.64)	0.59 (0.08)	0.004 (0.20)	1.87 (14.04)	0.70 (0.17)
		VCP(MI)	0.000 (0.03)	0.65 (1.44)	0.51 (0.07)	0.005 (0.23)	2.17 (2.89)	0.78 (0.19)
(a3)	100	Oracle	−0.011 (0.25)	5.68 (0.65)	2.07 (0.15)	0.013 (0.42)	8.36 (1.57)	2.50 (0.22)
		VC-SMAP	0.012 (0.24)	5.66 (0.73)	2.05 (0.17)	0.011 (0.42)	8.79 (2.37)	2.54 (0.23)
		VC-SMAPc	0.014 (0.25)	5.96 (0.68)	2.13 (0.15)	0.008 (0.40)	**8.10** (1.55)	**2.48** (0.21)
		SMAP	−0.034 (0.89)	8.70 (1.70)	2.43 (0.23)	−0.058 (1.19)	12.71 (3.62)	2.90 (0.37)
		SMAPc	−0.078 (1.89)	12.97 (5.89)	2.88 (0.61)	−0.100 (2.12)	16.94 (8.12)	3.29 (0.72)
		VCP	0.010 (0.22)	4.94 (0.61)	1.90 (0.14)	0.014 (0.43)	9.44 (2.10)	2.60 (0.26)
		VC-SMAP(MI)	−0.004 (0.12)	5.82 (0.74)	2.06 (0.17)	0.002 (0.58)	14.13 (22.96)	2.71 (0.43)
		VC-SMAPc(MI)	−0.002 (0.12)	6.03 (0.71)	2.11 (0.16)	0.013 (0.49)	9.90 (9.10)	2.56 (0.28)
		VCP(MI)	0.001 (0.12)	**4.92** (0.65)	**1.84** (0.16)	−0.014 (0.69)	16.40 (11.90)	2.94 (0.47)
	200	Oracle	−0.001 (0.19)	6.29 (0.45)	2.22 (0.10)	−0.012 (0.41)	7.66 (1.02)	2.46 (0.18)
		VC-SMAP	−0.007 (0.18)	6.33 (0.48)	2.22 (0.11)	−0.013 (0.39)	7.77 (1.09)	2.46 (0.18)
		VC-SMAPc	−0.006 (0.18)	6.51 (0.44)	2.28 (0.09)	−0.012 (0.39)	**7.49** (0.96)	**2.44** (0.17)
		SMAP	0.014 (0.75)	8.35 (1.07)	2.42 (0.15)	0.039 (1.0)	10.37 (2.36)	2.68 (0.29)
		SMAPc	0.049 (1.65)	14.70 (4.99)	3.04 (0.48)	0.068 (1.8)	16.19 (6.00)	3.22 (0.57)
		VCP	−0.006 (0.17)	**5.83** (0.43)	2.12 (0.09)	−0.005 (0.40)	8.04 (1.19)	2.48 (0.19)
		VC-SMAP(MI)	−0.003 (0.09)	6.57 (0.52)	2.23 (0.11)	0.021 (0.51)	10.81 (25.65)	2.57 (0.32)
		VC-SMAPc(MI)	−0.003 (0.09)	6.69 (0.50)	2.26 (0.11)	0.020 (0.48)	9.15 (17.47)	2.51 (0.25)
		VCP(MI)	−0.002 (0.09)	**5.83** (0.46)	**2.07** (0.10)	0.017 (0.56)	12.79 (10.80)	2.72 (0.35)

Table 2
Prediction results for New Zealand workforce study, where the values in parenthesis are standard deviations

	MSPE	MAPE
	In-sample performance	
VC-SMAP	0.0144 (0.0003)	0.0954 (0.0011)
VC-SMAPc	0.0145 (0.0003	0.0958 (0.0011)
SMAP	0.0162 (0.0003)	0.1017 (0.0012)
SMAPc	0.0164 (0.0003)	0.1026 (0.0011)
VCP	0.0135 (0.0003)	0.0926 (0.0011)
LMP	0.0149 (0.0003)	0.0970 (0.0012)
	Out-of-sample performance	
VC-SMAP	0.0151 (0.0003)	0.0975 (0.0011)
VC-SMAPc	0.0149 (0.0003)	0.0971 (0.0011)
SMAP	0.0171 (0.0014)	0.1031 (0.0014)
SMAPc	0.0167 (0.0004)	0.1035 (0.0012)
VCP	0.0157 (0.0010)	0.0987 (0.0014)
LMP	0.0151 (0.0003)	0.0973 (0.0012)

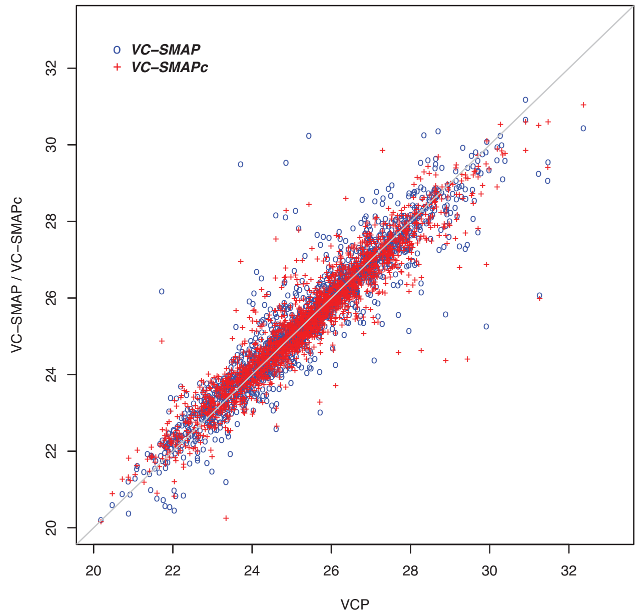

Figure 1. Predicted BMI by VC-SMAP and VC-SMAPc versus VCP for the training sample of New Zealand workforce study.

with less variability are reduced to a constant under the constrained estimation. Besides offering accurate prediction, the constrained VC-SMAP may be more interpretable and thus more appealing to practitioners.

4.2. *Bovine Collagen Trial Study*

We next consider some clinical data, which was previously studied by Li and Wong (2009) and originally from a 3-year NIH-sponsored randomized Bovine Collagen Trial for Scleroderma patients conducted at 12 centers in the USA (Postlethwaite et al., 2008). The raw data set consists of 831 observations. The response variable is the Modified Rodnan Skin Score (mrss) that uses skin thickness to measure disease severity. It is a continuous variable with an aggregated score from 17 areas of the body with a score of 0–3 from each area. The maximum value of mrss is 51 and a larger value means more disease severity. In our analysis, we log transformed the skin score and include nine continuous predictors and two discrete predictors in our analysis. The covariates include patient global assessment, physician global assessment, health assessment questionnaire, a pain score, several pulmonary function measures, and age. The two discrete covariates are ethnic and sex. Further information of the study and the variables are in Postlethwaite et al. (2008).

To apply our method, we select age as the index variable U since it is generally accepted that patient's age interacts with all other predictors. The results for MSPE and MAPE are summarized in Table 4. We observe that the training data

Table 3
*Results for estimated weights for New Zealand workforce study, where the values in the brackets are *standard deviations computed based on 400 bootstrap samples. The weights w_js correspond to different submodels with nonparametric component corresponding to "age" (intercept), "sbp," "dbp," "cholest," "drinkmaxday," "feethour," "sleep," "sex," "diabetes," "hypertension," "acne," and "nervous," respectively*

	SMAPc	SMAP		VC-SMAPc	VC-SMAP
Weight	Estimate	Estimate	Weight	Estimate	Estimate
w_1	0.038 (0.000)	0.201 (0.127)	w_1	0.528 (0.223)	0.750 (0.218)
w_2	0.081 (0.026)	0.195 (0.051)	w_2	0.000 (0.084)	0.187 (0.324)
w_3	0.649 (0.030)	0.651 (0.054)	w_3	0.000 (0.086)	0.201 (0.412)
w_4	0.232 (0.017)	0.353 (0.089)	w_4	0.000 (0.101)	0.239 (0.242)
w_5	0.000 (0.000)	0.059 (0.960)	w_5	0.073 (0.092)	0.467 (0.210)
w_6	0.000 (0.000)	−0.459 (0.532)	w_6	0.400 (0.196)	0.745 (0.144)
w_7	0.000 (0.000)	0.000 (0.653)	w_7	0.000 (0.101)	−0.112 (0.495)
			w_8	0.000 (0.151)	−0.068 (0.253)
			w_9	0.000 (0.081)	0.240 (0.546)
			w_{10}	0.000 (0.072)	−1.510 (0.500)
			w_{11}	0.000 (0.066)	0.151 (0.593)
			w_{12}	0.000 (0.078)	−0.290 (0.491)

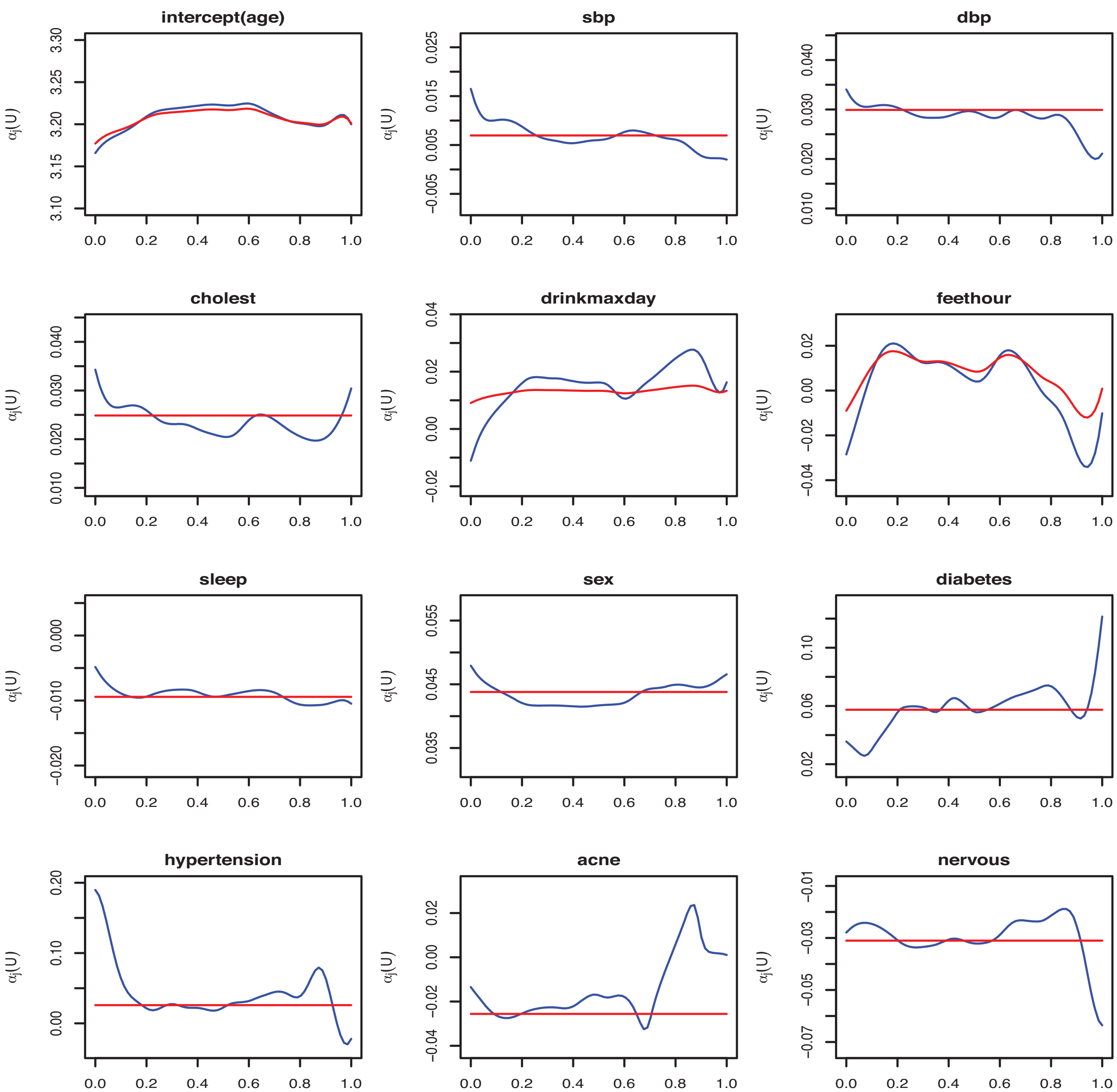

Figure 2. Estimated coefficient curves by two proposed methods: VC-SMAP and VC-SMAPc, where blue line in each subplot corresponds to the VC-SMAP method and red line represents the VC-SMAPc method.

performance of our proposed VC-SMAP is very close to that of VCP which achieves the lowest in-sample prediction error. When applying to the test sets, the VC-SMAPc outperforms other methods for the out-of-sample prediction.

This clinical data is much smaller than the New Zealand workforce study. Using moderate training sample size, we examine how the training size affects the prediction accuracy. When training sample size is adjusted to be four times that of the test sample, a similar conclusion can be obtained from Table 4. Other training sample sizes are also investigated in our analysis and the results are very similar. More analysis results can be found in Web Appendix B of the Supplementary Web Materials.

5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3 and 4, along with the *R* and *Matlab* code, are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Li's work is partly supported by Academic Research Funding R-155-000-174-114, R-155-000-195-114. Xia's work is partly supported by Hubei Provincial Natural Science Foundation of China (Grant No.2017CFB141). Wong's work is partly supported by the National Institute of General Medical

Table 4

Prediction results for Bovine Collagen data, where the values in parenthesis are standard deviations

	Size ratio of train and test sets 1:1		Size ratio of train and test sets 4:1	
	MSPE	MAPE	MSPE	MAPE
		In-sample performance		
VC-SMAP	0.0943 (0.008)	0.2262 (0.009)	0.1043 (0.004)	0.2382 (0.005)
VC-SMAPc	0.1033 (0.008)	0.2371 (0.008)	0.1083 (0.004)	0.2423 (0.004)
SMAP	0.1312 (0.011)	0.2674 (0.010)	0.1335 (0.005)	0.2687 (0.005)
SMAPc	0.1373 (0.011)	0.2737 (0.009)	0.1381 (0.005)	0.2739 (0.005)
VCP	0.0944 (0.011)	0.2221 (0.015)	0.0906 (0.011)	0.2153 (0.015)
LMP	0.1344 (0.009)	0.2730 (0.009)	0.1350 (0.005)	0.2720 (0.005)
		Out-of-sample performance		
VC-SMAP	0.1591 (0.122)	0.2754 (0.019)	0.1734 (0.499)	0.2609 (0.043)
VC-SMAPc	0.1296 (0.016)	0.2659 (0.009)	0.1237 (0.061)	0.2561 (0.020)
SMAP	0.1507 (0.022)	0.2837 (0.010)	0.1414 (0.022)	0.2753 (0.018)
SMAPc	0.1445 (0.012)	0.2812 (0.008)	0.1429 (0.021)	0.2771 (0.017)
VCP	0.1696 (0.036)	0.2930 (0.016)	0.1400 (0.035)	0.2696 (0.021)
LMP	0.1408 (0.010)	0.2791 (0.008)	0.1429 (0.021)	0.2792 (0.017)

Sciences of the National Institutes of Health under Award Number R01GM107639.

References

Anna, A. and Gondzio, J. (1999). Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. *Optimization Methods and Software* **11**, 275–302.

Buckland, S. T, Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603–618.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**, 941–956.

Chatfield, C. (2001). *Time-Series Forecasting*. Boca Raton: Chapman and Hall/CRC.

Chen, X. and Hansen, B. (2015). Forecasting with factor-augmented regression: A freguentist model averaging approach. *Journal of Econometrics* **186**, 280-293.

Chen, J., Li, D., Linton, O., and Lu, Z. (2017). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association* https://doi.org/10.1080/01621459.2017.1302339

Chen, X., Zou, C., and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**, 3696–3723.

Cheng, M. Y., Zhang, W., and Chen, L. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association* **104**, 1179–1191.

Claeskens, G. and Carrol, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249–265.

Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

De Gooijer, J. G. and Gannoun, A. (2000). Nonparametric conditional predictive regions for time series. *Computational Statistics & Data Analysis* **33**, 259–275.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. New York: Chapman & Hall, CRC Press.

Goldfarb, D. and Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*, J. P. Hennart (eds), volume 909, 226–239, Berlin: Springer-Verlag.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* **27**, 1–33.

Hansen, B. E. (2007) Least squares model averaging. *Econometrica* **75**, 1175–1189.

Hansen, B. E. (2008). Least squares forecast averaging. *Journal of Econometrics* **146**, 342–350.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with Discussion). *Statistical Science*, **14**, 382–401. [Corrected version.] Correction: **15**, 193–195.

Huang, T. and Li, J. (2018). Semiparametric model average prediction in panel data analysis. *Journal of Nonparametric Statistics* https://doi.org/10.1080/10485252.2017.1404061

Iglewics, B. and Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. Milwaukee, WI: American Socialty for Quanlity Control.

Jolliffe, I. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Li, D., Linton, O., and Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187**, 345–357.

Li, J. and Wong, W. K. (2009) A semiparametric analysis for identifying Scleroderma patients responsive to an antifibrotic agent. *Contemporary Clinical Trials* **30**, 105–113.

Liu, S. and Yang, Y. H. (2013). Mixing partially linear regression models. *Sankhya A* **75**, 74–95.

Ma, Y., and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review* **81**, 134–150.

MacMahon, S., Norton, R., Jackson, R., Mackie, M. J., Cheng, A., Vander Hoorn, S., et al. (1995). Fletcher challenge-university of auckland heart & health study: Design and baseline findings. *New Zealand Medical Journal* **108**, 499–502.

Matzner-Lober, E., Gannoun, A., and De Gooijer, J. G. (1998). Nonparametric forecasting: A comparison of three kernel-based methods. *Communications in Statistics-Theory and Methods* **27**, 1593–1617.

Postlethwaite, A. E., Wong, W. K., Clements, P., Chatterjee, S., Fessler, B. J., Kang, A. H., et al. (2008). A multicenter, randomized, double-blind, placebo-controlled trial of oral type I collagen treatment in patients with diffuse cutaneous systemic sclerosis: I. Oral type I collagen does not improve skin in all patients, but may improve skin in late-phase disease. *Arthritis Rheumatology* **58**, 1810–1822.

van de Geer, S. A. (2007). *Oracle inequalities and regularization.* In *Lectures on Empirical Processes*, E. Del Barrio, P. Deheuvels, and S. Van De Geer (eds), EMS Series of Lectures in Mathematics, European Mathematical Society (EMS), Zürich.

Wu, H. and Zhang, J. T. (2004). *Nonparametric Regression Methods for Longitudinal Data Analysis.* New York: Wiley.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–586.

Yang, Y. (2003). Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* **13**, 783–809.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models.* New York: Springer.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annanls of Statistics* **39**, 174–200.

Received August 2017. *Revised March* 2018.
Accepted April 2018.