Statistica Sinica Preprint No: SS-2020-0398	
Title	Metric Learning via Cross-Validation
Manuscript ID	SS-2020-0398
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0398
Complete List of Authors	Linlin Dai, Kani Chen, Gang Li and Yuanyuan Lin
Corresponding Author	Linlin Dai
E-mail	ldaiab@swufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Statistica Sinica

Metric Learning via Cross-Validation

Linlin Dai, Kani Chen, Gang Li and Yuanyuan Lin

Southwestern University of Finance and Economics,

Hong Kong University of Science and Technology,

University of California, Los Angeles and

The Chinese University of Hong Kong

Abstract: In this paper, we propose a *cross-validation metric learning* approach to learn a distance metric for dimension reduction in the multiple-index model. We minimize a leave-one-out cross-validation-type loss function, where the unknown link function is approximated by a metric-based kernel-smoothing function. To the best of our knowledge, we are the first to reduce the dimensionality of multiple-index models in a framework of metric learning. The resulting metric contains crucial information on both the central mean subspace and the optimal kernel-smoothing bandwidth. Under weak assumptions on the design of the predictors, we establish asymptotic theories for the consistency and convergence rate of the estimated directions, as well as the optimal rate of the bandwidth. Furthermore, we develop a novel estimation procedure to determine the structural dimension of the central mean subspace. The proposed approach is relatively easy to implement numerically by employing fast gradient-based algorithms. Various empirical studies illustrate its advantages over other existing methods.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Key words and phrases: Multiple-index model, Sufficient dimension reduction; Nonparametric regression.

1. Introduction

The performance of many successful machine learning algorithms, such as the k-nearest neighbors (Cover and Hart, 1967) (KNN) and support vector machine (Cortes and Vapnik, 1995), rely heavily on the notion of a metric or a distance between pairs of inputs. Here, the Euclidean distance is a commonly used distance metric. However, it ignores how samples are distributed in the feature space, especially in high-dimensional settings. A great deal of effort has been devoted to learning a proper pseudometric or Mahalanobis distance in settings such as classification, regression, and clustering, among others. A comprehensive discussion may be found in Bellet et al. (2013). It is known that learning a Mahalanobis metric is equivalent to identifying a linear transformation of the feature vectors (or predictors), and applying the standard Euclidean metric to the transformed data (Xing et al., 2003). When the linear projection is of lower rank, the metric is particularly important for data visualization, dimension reduction, and algorithm efficiency. Specifically, for two entries $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$, the Mahalanobis metric

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') \equiv \sqrt{(\mathbf{x} - \mathbf{x}')^\top \mathbf{M} (\mathbf{x} - \mathbf{x}')} = \|\mathbf{A}^\top (\mathbf{x} - \mathbf{x}')\|,$$

for a $p \times p$ positive semi-definite matrix $\mathbf{M}$, where the second equality is the result of the decomposition that $\mathbf{M} = \mathbf{A}\mathbf{A}^\top$, for some $\mathbf{A}$ with $\text{rank}(\mathbf{A}) \leq p$. Goldberger et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

(2005) presented a neighborhood components analysis by maximizing a variant of the leave-one-out KNN score using gradient-based algorithms, which is conceptually appealing and effective for a low-rank $\mathbf{A}$. Nevertheless, its theoretical justifications are generally quite challenging. The large margin nearest neighbors algorithm (LMNN) by Weinberger et al. (2006) and Weinberger and Saul (2009) directly learns a metric $\mathbf{M}$ to determine the "target neighbors" in a KNN classification based on certain local pairs or triples conditions. For regression problems, Weinberger and Tesauro (2007) constructed a novel metric learning algorithm for a kernel regression, without any theoretical justification for the resulting metric; Noh et al. (2017) investigated an effective approach for reducing the bias and mean squared error in kernel regressions under Gaussian models. Though both works studied metric-learning for kernel regressions, they do not consider the problems for multiple-index models, which have received much attention and have been investigated intensively in many scientific fields.

In this study, we focus on dimension reduction for the multiple-index model in a framework of metric learning. Specifically, for a response $Y \in \mathbb{R}$ and a vector of predictors $\mathbf{X} \in \mathbb{R}^p$, we concentrate on reducing the dimensionality of the mean function $f(\mathbf{X}) = E(Y|\mathbf{X})$, leaving the rest of $Y|\mathbf{X}$ as "nuisance parameters." A reduced-rank structure of the regressors $f(\mathbf{x})$ leads to the popular multiple-index

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

model

$$Y = g(\mathbf{L}_0^\top \mathbf{X}) + \epsilon, \tag{1.1}$$

where $g : \mathbb{R}^{r_0} \to \mathbb{R}$ is an unknown link function ($r_0 \leq p$), $\mathbf{L}_0$ is a $p \times r_0$ column orthogonal matrix, and the noise ϵ satisfies $E(\epsilon|\mathbf{X}) = 0$, almost surely. The subspace spanned by the column vectors of $\mathbf{L}_0$ is referred to as the *central mean subspace* (CMS), as introduced by Cook and Li (2002), and is of major importance in the literature. It is well defined and is unique under mild conditions. We refer to r_0 as the structural dimension of the CMS, and to the column vectors of $\mathbf{L}_0$ as the directions in the CMS.

Note that there is a large body of literature in statistics on dimension reduction for model (1.1) and its variants. One of the most fundamental and powerful methods is the seminal sliced inverse regression (SIR), invented by Li (1991). It can be used to find vectors outside the CMS, but inside the central subspace, the smallest subspace capturing the complete dependence of Y on $\mathbf{X}$ (Cook and Li, 2002). Li (1991) also developed a sequential testing procedure to determine the dimension of the CMS when r_0 is unknown. Since then, many state-of-the-art inverse regression-based approaches have been developed, such as the sliced average variance estimation (SAVE) (Shao et al., 2007); see Bura and Cook (2001a), Bura (2003), Cook and Li (2004), Cook and Ni (2005), and Yin and Cook (2006), among many others. These methods are computationally simple, and thus widely applied in data mining.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

However, it is known that the inverse regression-based methods usually need strong assumptions on the design of $\mathbf{X}$, such as the elliptical symmetry condition, similar to the requirement in the principal Hessian directions method (pHd) (Li, 1992; Cook, 1998). As an important alternative, Xia et al. (2002) invented a novel minimum average variance estimation method (MAVE) based on local linear smoothing. Based on the MAVE, they also proposed a consistent estimate for the dimension of the CMS. Other related approaches, including the average derivative estimation (Härdle and Stoker, 1989), structure adaptive approach (SA) (Hristache et al., 2001), and outer products of gradients (OPG) (Samarov, 1993), are designed to estimate the derivative of the regressor $g(\mathbf{L}_0^\top \mathbf{x})$ pertaining to the CMS. More advancements can be found in Xia (2008), Wang and Xia (2008), Dalalyan et al. (2008), Chen et al. (2011), Alquier and Biau (2013), and Akritas (2016). Overall, compared with the inverse regression, direct regression methods are easy to implement and are superior in terms of their finite-sample performance (Hristache et al., 2001; Xia et al., 2002; Xia, 2007). With the bandwidth carefully chosen, direct regression methods report elegant results. Ma and Zhu (2012) provided a novel semiparametric approach to estimate the CMS by solving estimating equations, and later studied its efficiency issues (Ma and Zhu, 2014). Recently, an important discussion paper (Cannings and Samworth, 2017) introduced a general classifier for high-dimensional data using random projections.

In this paper, we propose the *cross-validation metric learning* (CVML) approach

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

to learn a distance metric that contains crucial information on the CMS and the nonparametric link function in model (1.1). For any fixed dimension r, such that $1 \leq r \leq p$, the CVML procedure minimizes a leave-one-out cross-validation-type sum of squared errors over matrix $\mathbf{A} \in \mathbb{R}^{p\times r}$, in which the link function is approximated by the Nadaraya-Watson kernel estimator. One can thus estimate the directions of the CMS and the bandwidth of the link function simultaneously using the singular value decomposition $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{A}}^\top = \hat{\mathbf{L}}_1\hat{\mathbf{H}}^{-2}\hat{\mathbf{L}}_1^\top$. When $r = r_0$, the CVML estimate for the directions of the CMS is shown to be consistent at a certain convergence rate. Furthermore, a sequential procedure is developed to determine the dimension r_0 of the CMS when it is unknown. The results of simulation studies show that the proposal outperforms other alternatives in terms of estimating the directions and dimension of the CMS. The CVML procedure is model-free, in the sense that its validity does not rely on any specific functional relation between the response variable and the predictors, making it practically appealing. Furthermore, unlike many other metric learning algorithms, such as the methods developed in Xing et al. (2003), Weinberger et al. (2006), and Weinberger and Saul (2009), the loss function of the proposed CVML is differentiable and free of local constraints. As a result, computation of the proposal is straightforward, with the help of gradient-based algorithms.

The rest of the paper is organized as follows. Section 2 describes the proposed CVML procedure for estimating the directions and the dimension of the CMS, and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

presents the theoretical results for the consistency, asymptotic expansion and convergence rate. The results of simulations and real-data applications are given in Sections 3 and 4, respectively. Assumptions and remarks are summarized in the Appendix. Technical proofs are provided in the Supplementary Material.

2. Cross-validation metric learning method

Suppose that $\{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$ are independent random copies of $(Y, \mathbf{X})$ taking values in $\mathbb{R} \times \mathbb{R}^p$, and ϵ_i are random errors such that

$$Y_i = g(\mathbf{L}_0^\top \mathbf{X}_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{X}_i$ is supported by a bounded set Ω.

Note that model (1.1) is not uniquely defined. This is because for any orthonormal transformation $\mathbf{Q} \in \mathbb{R}^{r_0 \times r_0}$,

$$f(\mathbf{x}) = g(\mathbf{L}_0^\top \mathbf{x}) = g(\mathbf{Q}^\top \mathbf{Q} \mathbf{L}_0^\top \mathbf{x}) \equiv g_1(\mathbf{L}_0^{*\top} \mathbf{x}),$$

where $g_1(\mathbf{u}) = g(\mathbf{Q}^\top \mathbf{u})$ and $\mathbf{L}_0^* = \mathbf{L}_0 \mathbf{Q}^\top$. Although $\mathbf{L}_0$ is not unique, the subspace spanned by the column vectors of $\mathbf{L}_0$, denoted by $\mathcal{S}(\mathbf{L}_0)$, is unique, with the projection matrix $\mathbf{L}_0 \mathbf{L}_0^\top$. In this paper, $\mathcal{S}(\mathbf{L}_0)$ is referred to as the CMS.

2.1. Estimating the directions in the CMS

Given the true projection matrix $\mathbf{L}_0$, the regression function can be written as

$$f(\mathbf{x}) = E(Y | \mathbf{X} - \mathbf{x} \in \mathbf{L}_0^\perp),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

where $\mathbf{L}_0^{\perp}$ denotes the space spanned by vectors perpendicular to $\mathcal{S}(\mathbf{L}_0)$. We estimate $f(\cdot)$ using the kernel smoothing method, as follows.

For any fixed $1 \leq r \leq p$, set $\mathbf{M} = \mathbf{L}_1\mathbf{H}^{-2}\mathbf{L}_1^{\top}$, where $\mathbf{L}_1$ is of size $p \times r$ satisfying $\mathbf{L}_1^{\top}\mathbf{L}_1 = \mathbf{I}_r$, and $\mathbf{H} = \text{diag}(h_1, \ldots, h_r)$ is the bandwidth matrix with $h_1 > 0, \ldots, h_r > 0$. Hence, the matrix $\mathbf{M}$ is positive semi-definite, and can be viewed as a distance metric between two samples. The kernel function based on the distance metric $\mathbf{M}$ is defined as

$$K_{\mathbf{M}}(\mathbf{t}) = \frac{1}{h_1 \cdots h_r} K(\mathbf{t}^{\top}\mathbf{M}\mathbf{t}) = \frac{1}{h_1 \cdots h_r} K(\mathbf{t}^{\top}\mathbf{L}_1\mathbf{H}^{-2}\mathbf{L}_1^{\top}\mathbf{t}), \ \mathbf{t} \in \mathbb{R}^p,$$

where $K(\cdot)$ is a univariate kernel function defined on $[0, \infty)$, with a bounded support satisfying $\int_{\mathbf{s}\in\mathbb{R}^r} K(\|\mathbf{s}\|^2)d\mathbf{s} = 1$.

Heuristically, the Nadaraya-Watson kernel-smoothing estimator of $f(\mathbf{x})$ is

$$\hat{f}_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_{\mathbf{M}}(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_{\mathbf{M}}(\mathbf{X}_i - \mathbf{x})}, \text{ for any } \ \mathbf{x} \in \mathbb{R}^p,$$

where $\mathbf{M}$ is unknown and to be estimated. Thus, we define

$$K^*_{j,i} = \frac{K_{\mathbf{M}}(\mathbf{X}_j - \mathbf{X}_i)}{\sum_{l\neq i} K_{\mathbf{M}}(\mathbf{X}_l - \mathbf{X}_i)}, \text{ for } \ j \neq i,$$

and $K^*_{i,i} = 0$. Note that $\sum_{j=1}^n K^*_{j,i} = 1$ and $K^*_{j,i} \neq K^*_{i,j}$, for any $i \neq j$. Let $\mathbb{S}_+^p$ be the cone of symmetric positive semi-definite $p \times p$ real-valued matrices. The proposed estimator of $\mathbf{M}$ is the minimizer of

$$\text{CM}_n(\mathbf{M}) = \frac{1}{n}\sum_{i=1}^n \{\hat{f}^{(-i)}(\mathbf{X}_i) - Y_i\}^2 w(\mathbf{X}_i), \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

over all $\mathbf{M} \in \mathbb{S}_+^p$, denoted by $\hat{\mathbf{M}}$, where

$$\hat{f}^{(-i)}(\mathbf{X}_i) = \sum_{j=1}^{n} Y_j K_{j,i}^* = \frac{\sum_{j\neq i} Y_j K_{\mathbf{M}}(\mathbf{X}_j - \mathbf{X}_i)}{\sum_{j\neq i} K_{\mathbf{M}}(\mathbf{X}_j - \mathbf{X}_i)}, \tag{2.3}$$

and $w(\cdot)$ is a bounded and positive weight function with support Ω° strictly inside Ω. The objective function (2.2) is essentially a leave-one-out cross-validation based on the squared errors, and thus the proposed procedure is called *cross-validation metric learning*. The weight function $w(\cdot)$ is introduced to handle the boundary effect by letting $w(\mathbf{x}) = 0$ if $\inf_{\mathbf{y}\in\partial\Omega} \|\mathbf{x} - \mathbf{y}\| < c$, for some constant $c > 0$, where $\partial\Omega$ is the boundary of Ω.

In practise, to obtain $\hat{\mathbf{M}}$, we remove the constraint $\mathbf{M} \in \mathbb{S}_+^p$ using the decomposition $\mathbf{M} = \mathbf{A}\mathbf{A}^\top$, for all $\mathbf{A} \in \mathbb{R}^{p\times r}$. Let $\mathrm{vec}(\mathbf{A})$ denote the vectorization of a matrix $\mathbf{A}$ by its column vectors, and let $\mathbf{A}_1 \otimes \mathbf{A}_2$ denote the Kronecker product of $\mathbf{A}_1$ and $\mathbf{A}_2$. Then, (2.3) can be written as

$$\hat{f}^{(-i)}(\mathbf{X}_i) = \sum_{j=1}^{n} Y_j K_{j,i}^* = \frac{\sum_{j\neq i} Y_j K(\|(\mathbf{I}_r \otimes \mathbf{X}_{ij}^\top)\mathrm{vec}(\mathbf{A})\|^2)}{\sum_{j\neq i} K(\|(\mathbf{I}_r \otimes \mathbf{X}_{ij}^\top)\mathrm{vec}(\mathbf{A})\|^2)},$$

where $\mathbf{X}_{ij} \equiv \mathbf{X}_j - \mathbf{X}_i$. Taking the derivative of (2.2) with respect to $\mathrm{vec}(\mathbf{A})$ yields a gradient rule:

$$\frac{\partial \mathrm{CM}_n(\mathbf{M})}{\partial \mathrm{vec}(\mathbf{A})} = -\frac{2}{n}\sum_{i=1}^{n} \frac{\partial \hat{f}^{(-i)}(\mathbf{X}_i)}{\partial \mathrm{vec}(\mathbf{A})}\{Y_i - \hat{f}^{(-i)}(\mathbf{X}_i)\}w(\mathbf{X}_i), \tag{2.4}$$

where

$$\frac{\partial \hat{f}^{(-i)}(\mathbf{X}_i)}{\partial \mathrm{vec}(\mathbf{A})} = \frac{2\sum_{j\neq i} \dot{K}(\|\mathbf{A}^\top\mathbf{X}_{ij}\|^2)\{Y_j - \hat{f}^{(-i)}(\mathbf{X}_i)\}(\mathbf{I}_r \otimes \mathbf{X}_{ij}\mathbf{X}_{ij}^\top)\mathrm{vec}(\mathbf{A})}{\sum_{j\neq i} K(\|\mathbf{A}^\top\mathbf{X}_{ij}\|^2)}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Here, $\dot{K}(\cdot)$ denotes the first derivative of the kernel function $K(\cdot)$. Therefore, the numerical computation of the proposed CVML approach can be carried out by employing gradient-based algorithms, such as the conjugate gradient or gradient descent algorithms. In particular, when $r \ll p$, the computation would be relatively efficient. To further improve the computational efficiency, one may also consider using algorithms such as the stochastic gradient decent. Once $\hat{\mathbf{M}}$ is obtained, the estimated bandwidth matrix $\hat{\mathbf{H}}$ and the directions $\hat{\mathbf{L}}_1$ can be calculated immediately using the singular value decomposition of $\hat{\mathbf{M}}$.

The detailed estimation procedure for the proposed CVML method is summarized in Algorithm 1. This procedure is free of local pairwise constraints that are required in many other metric learning methods. Moreover, the procedure to simultaneously estimate the effective directions and the bandwidths avoids the bias problem arising from using two separate cost functions to estimate the directions and the link function in many popular methods; see Hall (1989), Härdle and Stoker

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

(1989), and Carroll et al. (1997).

Algorithm 1: Estimation of directions and bandwidth

Data: $\mathbf{X}$, $\mathbf{y}$, r

Result: $\hat{\mathbf{L}}_1$, $\hat{\mathbf{H}}$

1 Get $\hat{\mathbf{A}}$ by minimizing (2.2) with the gradient (2.4);

2 Calculate $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{A}}^\top$;

3 Singular value decomposition $\hat{\mathbf{M}} = \hat{\mathbf{L}}_1\hat{\mathbf{\Lambda}}\hat{\mathbf{L}}_1^\top$;

4 $\hat{\mathbf{H}} = \hat{\mathbf{\Lambda}}^{-1/2}$;

Remark 1. Härdle et al. (1993) first applied the cross-validation technique to estimate the single-index model ($r_0 = 1$), and the estimator is shown to have good asymptotic properties. The proposed method is similar, but also substantially different from their method. The cross-validation method cannot be extended easily to multiple-index models, because it uses a grid search algorithm to estimate the directions and bandwidths, which is inefficient and costly in higher-dimensional settings. In addition, instead of estimating the bandwidths and directions separately, we simply regard the fusion matrix $\mathbf{M}$ as a Mahalanobis metric. A relevant work to Algorithm 1 is Weinberger and Tesauro (2007). Nonetheless, the bandwidth and directions in the CMS are not studied in their setup, and no theoretical justification for the properties of $\hat{\mathbf{M}}$ is established. They do not consider how to estimate the desired dimensionality when it is unknown. In contrast to Weinberger and Tesauro (2007) and Noh et al. (2017), we provide an in-depth study of the structure of $\hat{\mathbf{M}}$ (eigenval-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

ues and eigenvectors) in a statistical way and attempt to apply it to multiple-index models.

Remark 2. In contrast to some SIR-based methods (Li, 1991, 1992; Cook, 1998), the proposed method is free of the linearity condition and constant covariance condition; see condition (C1) in the Appendix. Methods such as the seminal MAVE and SA methods usually perform a nonparametric kernel estimation procedure to estimate the link function or its derivative, which involves selecting bandwidths to be used in estimating effective directions. This is not needed in the proposed CVML approach, because it directly obtains a data-driven bandwidth. For all large n, the estimated bandwidth is shown to be at the same rate as the theoretically optimal bandwidth in the sense of minimizing the mean weighted integrated squared errors

$$\int_{\mathbf{x}\in\mathbb{R}^p} E\{\hat{f}_n(\mathbf{x}) - f(\mathbf{x})\}^2 w(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \tag{2.5}$$

Here and after, $\|\mathbf{A}\|$ denotes the Frobenius norm for matrix $\mathbf{A}$, and $\dot{g}(\mathbf{x})$ and $\ddot{g}(\mathbf{x})$ denote the first and second derivatives of $g(\cdot)$ at $\mathbf{x}$, respectively.

Theorem 1 (Consistency)**.** *Suppose that the true dimension r_0 of the CMS is known and conditions (C1)–(C5) in the Appendix hold. Define $\mathbf{h} = (h_1, \ldots, h_{r_0})^\top$. If $\|\mathbf{h}\| \to 0$ and $h_1 h_2 \cdots h_{r_0} > n^{-\delta}$, for some $0 < \delta < 1$, then $\hat{\mathbf{L}}_1 \to \mathbf{L}^*$ in probability as $n \to \infty$, where $\mathcal{S}(\mathbf{L}^*) = \mathcal{S}(\mathbf{L}_0)$.*

Theorem 1 states that under certain conditions, the estimated direction $\hat{\mathbf{L}}_1$ con-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

verges to the directions in the true CMS. In other words, the CVML method is able to estimate the directions in the CMS consistently. To determine the convergence rate, we first present the asymptotic expansion of $\mathrm{CM}_n(\mathbf{M})$. Let

$$R_1(K)\mathbf{I}_{r_0} = \int_{\mathbf{s}\in\mathbb{R}^{r_0}} \mathbf{s}\mathbf{s}^\top K(\|\mathbf{s}\|^2)d\mathbf{s},\ R_2(K) = \int_{\mathbf{s}\in\mathbb{R}^{r_0}} K^2(\|\mathbf{s}\|^2)d\mathbf{s}.$$

Theorem 2 (Asymptotic expansion)**.** *Suppose that the true dimension r_0 of the CMS is known and conditions (C1)–(C5) in the Appendix hold. Let*

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{pmatrix}$$

be a $p\times p$ orthonormal matrix, where $\mathbf{L}_1 \in \mathbb{R}^{p\times r_0}$ and $\mathbf{L}_2$ is the augmented orthonormal basis in $\mathbb{R}^p$ satisfying $\|\mathbf{L}_0^\top \mathbf{L}_2\| \to 0$. Let $f_{r_0}(\cdot)$ be the density of $\mathbf{L}_0^\top \mathbf{X}$. Then, uniformly over $\{\mathbf{h} : \|\mathbf{h}\| \le \delta_n\}$ for any $\delta_n \to 0$, and $h_1 h_2 \cdots h_{r_0} > n^{-\delta}$ for some $0 < \delta < 1$,

$$\begin{aligned} \mathrm{CM}_n(\mathbf{M}) - \eta_0 &= \int_{\mathbf{t}\in\mathbb{R}^p} \{\psi(\mathbf{t},\mathbf{h},\mathbf{L}_1)\}^2 f_{\mathbf{X}}(\mathbf{t}) \frac{w(\mathbf{t})}{f_{r_0}^2(\mathbf{L}_0^\top \mathbf{t})} d\mathbf{t} + \frac{R_2(K)V_0}{nh_1\cdots h_{r_0}} \\ &\quad + o_p\left(\|\mathbf{L}_0^\top \mathbf{L}_2\|^2 + \|\mathbf{h}\|^4 + \frac{1}{nh_1\cdots h_{r_0}}\right), \end{aligned} \tag{2.6}$$

where $\eta_0 = n^{-1}\sum_{i=1}^n w(\mathbf{X}_i)\epsilon_i^2$,

$$\begin{aligned} \psi(\mathbf{t},\mathbf{h},\mathbf{L}_1) &= \dot{g}(\mathbf{L}_0^\top \mathbf{t})^\top \mathbf{L}_0^\top \mathbf{L}_2 \mathbf{b}(\mathbf{L}_0^\top \mathbf{t}) + R_1(K) tr\{\mathbf{H}\mathbf{L}_1^\top \mathbf{L}_0 \mathbf{A}(\mathbf{L}_0^\top \mathbf{t})\mathbf{L}_0^\top \mathbf{L}_1 \mathbf{H}\}, \\ \mathbf{A}(\mathbf{L}_0^\top \mathbf{t}) &= \frac{1}{2}\ddot{g}(\mathbf{L}_0^\top \mathbf{t}) f_{r_0}(\mathbf{L}_0^\top \mathbf{t}) + \dot{g}(\mathbf{L}_0^\top \mathbf{t})\dot{f}_{r_0}(\mathbf{L}_0^\top \mathbf{t})^\top, \quad \mathbf{t}\in\mathbb{R}^p, \\ \mathbf{b}(\mathbf{L}_0^\top \mathbf{t}) &= E_{\mathbf{u}_2|\mathbf{u}_1}(\mathbf{U}_2 - \mathbf{L}_0^{\perp\top}\mathbf{t} | \mathbf{U}_1 = \mathbf{L}_0^\top \mathbf{t}) f_{r_0}(\mathbf{L}_0^\top \mathbf{t}), \\ V_0 &= \int_{\mathbf{t}\in\mathbb{R}^p} \sigma^2(\mathbf{L}_0^\top \mathbf{t}) \frac{f_{\mathbf{X}}(\mathbf{t}) w(\mathbf{t})}{f_{r_0}(\mathbf{L}_0^\top \mathbf{t})} d\mathbf{t}. \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Remark 3. The asymptotic expansion in (2.6) offers some insight into the CVML method. The first term on the right-hand side of (2.6) is the bias term, and the second term is the variance term. For instance, when the identifiability condition (C4) in the Appendix is violated, there exists a unit vector $\boldsymbol{\ell}_1 \in \mathbb{R}^p$, such that $\boldsymbol{\ell}_1^\top \mathbf{L}_0 \dot{g}(\mathbf{L}_0^\top \mathbf{t}) = 0$, for all $\mathbf{t} \in \mathbb{R}^p$. Then, $\boldsymbol{\ell}_1^\top \mathbf{L}_0 \mathbf{A}(\mathbf{L}_0^\top \mathbf{t}) \mathbf{L}_0^\top \boldsymbol{\ell}_1 = 0$. The bandwidth along the direction $\boldsymbol{\ell}_1$ need not be small. In a special case that $g(\cdot)$ is constant, and thus $\mathbf{A}(\mathbf{L}_0^\top \mathbf{t}) = \mathbf{0}$, the bias term is irrelevant to the bandwidth $\mathbf{h}$, and only the variance term $R_2(K)V_0/(nh_1 \cdots h_{r_0})$ plays a role in $\mathrm{CM}_n(\mathbf{M}) - \eta_0$. As a result, the estimated bandwidth tends to be large and the estimate of the link function reduces to a constant.

The following corollary presents the rate of convergence of $\hat{\mathbf{L}}_1$. Define the distance between the subspaces spanned by $\mathbf{L}_0$ and $\hat{\mathbf{L}}_1$ as $m(\hat{\mathbf{L}}_1, \mathbf{L}_0) = \|\mathbf{L}_0^\top (\mathbf{I}_p - \hat{\mathbf{L}}_1 \hat{\mathbf{L}}_1^\top)\|$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix.

Corollary 1 (Rate of convergence)**.** *Suppose that the true dimension r_0 of the CMS is known and conditions (C1)–(C5) in the Appendix hold. Then,*

$$m(\hat{\mathbf{L}}_1, \mathbf{L}_0) = O_p(\|\mathbf{h}\|^2).$$

Moreover, the resulting bandwidth minimizing (2.6) is at the order of $n^{-1/(r_0+4)}$.

Recall that the theoretically optimal bandwidth for the nonparametric estimation in the sense of minimizing (2.5) is also at the order of $n^{-1/(r_0+4)}$. This implies that we

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

can simultaneously estimate the central mean subspace and the link function with the optimal rate of bandwidth.

On the other hand, it can be seen from the asymptotic expansion (2.6) in Theorem 2 that the esimated directions in the CMS are only relevant to the bias term. As a result, the convergence rate of $\hat{\mathbf{L}}_1$ is at the order of $\|\mathbf{h}\|^2$. Intuitively, a narrower bandwidth would result in a faster convergence rate. This finding induces the following correction method that allows a faster rate of convergence. Instead of minimizing $\mathrm{CM}_n(\mathbf{M})$, one can minimize

$$\mathrm{CM}_n(\mathbf{M}) - \frac{R_2(K)\hat{V}_0}{nh_1 \cdots h_{r_0}}$$

over $\mathbf{M}$, where $\hat{V}_0$ is an estimate of V_0. Recall that V_0 is related to the density $f_{\mathbf{X}}(\mathbf{x})$ and the variance function $\sigma^2(\mathbf{x})$, which are usually unknown. The density $f_{\mathbf{X}}(\mathbf{x})$ can be estimated using conventional density estimation methods. The variance function $\sigma^2(\cdot)$ can be estimated by referring to Härdle et al. (1993). The correction method improves the rate of convergence and is of theoretical interest.

The following remark provides greater insight into the convergence rate and the optimal rate of the bandwidth in Corollary 1.

Remark 4. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{r_0})^\top$, $h_1 = \alpha_1 n^{-1/(r_0+4)}, \cdots, h_{r_0} = \alpha_{r_0} n^{-1/(r_0+4)}$, $\mathbf{L}_0^\top \mathbf{L}_1 = \tilde{\mathbf{D}}_{r_0 \times r_0}$, which is an orthonormal matrix, and $\mathbf{L}_0^\top \mathbf{L}_2 = n^{-2/(r_0+4)} \mathbf{D}_{r_0 \times (p-r_0)}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

The leading term of the right-hand side of (2.6) is

$$n^{-\frac{4}{r_0+4}}\Bigg\{\int_{\mathbf{t}\in\mathbb{R}^p}\Big[\dot{g}(\mathbf{L}_0^\top\mathbf{t})^\top\mathbf{D}\mathbf{b}(\mathbf{L}_0^\top\mathbf{t})+R_1(K)\mathrm{tr}\big\{\mathrm{diag}(\boldsymbol{\alpha})\tilde{\mathbf{D}}^\top\mathbf{A}(\mathbf{L}_0^\top\mathbf{t})\tilde{\mathbf{D}}\mathrm{diag}(\boldsymbol{\alpha})\big\}\Big]^2 \times\frac{f_{\mathbf{X}}(\mathbf{t})w(\mathbf{t})}{f_{r_0}^2(\mathbf{L}_0^\top\mathbf{t})}d\mathbf{t}+\frac{R_2(K)V_0}{\alpha_1\cdots\alpha_{r_0}}\Bigg\}. \tag{2.7}$$

Denote the minimizer of (2.7) as $\boldsymbol{\alpha}_*,\mathbf{D}_*,\tilde{\mathbf{D}}_*$. As a result, the optimal bandwidth $\hat{\mathbf{h}}=n^{-1/(r_0+4)}\boldsymbol{\alpha}_*\{1+o(1)\}$ and $\hat{\mathbf{L}}_1=(\mathbf{L}_0-n^{-2/(r_0+4)}\hat{\mathbf{L}}_1^\perp\mathbf{D}_*^\top)\tilde{\mathbf{D}}_*^{-1}\{1+o(1)\}$.

2.2. Determining the dimension of the CMS

The true dimension r_0 is crucial to the estimation of the CMS, but it is often unknown in practice. Determining r_0 is also a nontrivial task. Many existing approaches used to determine the structural dimension of the CMS are inspired by the equivalence between dimension reduction and matrix eigen-decomposition. The sequential test methods (Li, 1991; Bura and Cook, 2001b; Cook and Ni, 2005) generally cannot give a consistent $\hat{r}$ owing to the type-I error. The bootstrapping methods (Ye and Weiss, 2003; Zhu and Zeng, 2006; Luo and Li, 2016) can determine the dimension in a data-driven manner, but are computationally burdensome. The BIC criterion (Zhu et al., 2006; Zhu and Zhu, 2007) and the ratio estimation methods (Luo et al., 2009; Xia et al., 2015; Zhu et al., 2019, 2020) are able to produce consistent estimations of r_0 and are computationally attractive. The sparse eigen-decomposition proposed by Zhu et al. (2010) can estimate directions and the structural dimension

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

of the CMS simultaneously. However, the aforementioned methods rely on a relevant kernel matrix, usually obtained by inverse regression-based estimation procedures, and thus the link function is lost. In a nonparametric regression framework, Xia et al. (2002) proposed determining r_0 using a leave-one-out cross-validation procedure based on MAVE estimated directions. Inspired by the novel ideas of Xia et al. (2002) and the ratio estimation approaches, we propose the CVML method for determining the dimension of the CMS.

Proposition 1. *Suppose that the conditions (C1)–(C5) in the Appendix hold. Under model (1.1), as $n \to \infty$, with probability tending to one,*

(i) $\mathrm{CM}_n(\hat{\mathbf{M}}_r)/\mathrm{CM}_n(\hat{\mathbf{M}}_{r_0}) > 1$, *for all* $1 \le r < r_0$;

(ii) $\mathrm{CM}_n(\hat{\mathbf{M}}_r)/\mathrm{CM}_n(\hat{\mathbf{M}}_{r_0}) \to 1$, *for all* $r_0 \le r \le p$.

Proposition 1 shows that $\mathrm{CM}_n(\hat{\mathbf{M}}_r) > \mathrm{CM}_n(\hat{\mathbf{M}}_{r_0})$, for all $r < r_0$, because of lack of fit. Intuitively, $\mathrm{CM}_n(\hat{\mathbf{M}}_r)$ would decrease as r increases until it arrived at r_0. Therefore, we attempt to track the first time that the ratio $\mathrm{CM}_n(\hat{\mathbf{M}}_r)/\mathrm{CM}_n(\hat{\mathbf{M}}_{r+1})$ hits one, and estimate the dimension of CMS as

$$\hat{r} \equiv \min_{0 \le r \le p-1} \left\{ r : \left| \frac{\mathrm{CM}_n(\hat{\mathbf{M}}_r)}{\mathrm{CM}_n(\hat{\mathbf{M}}_{r+1})} - 1 \right| < \tau_n \right\},$$

where τ_n is positive and converges to zero at a slow rate and $\mathrm{CM}_n(\hat{\mathbf{M}}_0) = n^{-1}(y_i - \bar{y})^2$. The choice of τ_n is given in Section 3. The estimation procedure is detailed in Algorithm 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Remark 5. The estimation procedure ranges r from 1 to p. The method in Xia et al. (2002) involves calculating cross-validation errors for all $r \in \{1, \ldots, p\}$. Our proposed approach stops at a certain $r < p$ and, thus, possibly avoids the computational burden caused by the calculation of $\hat{\mathbf{M}}_r$ for some large r. In practise, to ensure the estimation accuracy, one can require that the procedure stops only when two consecutive ratios are close to one or, equivalently, modify the stopping condition as $|\mathrm{CM}_n(\hat{\mathbf{M}}_r)/\mathrm{CM}_n(\hat{\mathbf{M}}_{r+1}) - 1| < \tau_n$ and $|\mathrm{CM}_n(\hat{\mathbf{M}}_{r+1})/\mathrm{CM}_n(\hat{\mathbf{M}}_{r+2}) - 1| < \tau_n$, for some fixed r.

The empirical performance of the proposed method in terms of determining the dimension of the CMS is shown in the next section.

3. Simulations

In this section, we examine the performance of the proposed CVML method in terms of estimating the directions in the CMS and determining the structural dimension of the CMS, respectively. We adopt the Gaussian-type kernel function $K(\mathbf{u}) = \exp(-\frac{1}{2}\mathbf{u}^\top \mathbf{M}\mathbf{u})$. For simplicity, the weight function $w(\mathbf{x})$ is set to be one, and thus all the observations have equal weights. The CVML approach is implemented with the help of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm in the *lbfgs* package in R. Let ϵ follow the standard normal distribution $N(0, 1)$. We generate models from the following cases:

Example 3.1. We generate X_i from the standard normal distribution $N(0, 1)$ inde-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Algorithm 2: Determining the dimension of the CMS

Data: $\mathbf{X}$, $\mathbf{y}$

Result: $\hat{r}$, $\hat{\mathbf{M}}_{\hat{r}}$

Initialization: τ_n, $r = 1$, $\Delta = 10$, $Err_0 = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$;

while $|\Delta - 1| > \tau_n$ **do**

  if *r=p+1* **then**

    print $\hat{r} = p$, **break**;

  else

    Calculate $\hat{\mathbf{M}}_r$ by Algorithm 1, $Err_1 = \mathrm{CM_n}(\hat{\mathbf{M}}_r)$;

    $\Delta = Err_0/Err_1$, $Err_0 = Err_1$;

    $r = r + 1$;

$\hat{r} = r - 2$.

pendently; we generate Y from

$$\text{Model 1: } Y = X_1/\{0.5 + (X_2 + 1.5)^2\} + 0.5\epsilon,$$

$$\text{Model 2: } Y = X_1(X_1 + X_2 + 1) + 0.5\epsilon,$$

where Model 1 and Model 2 follow those of, for example, Li (1991) and Xia et al. (2002). The sample size is set at $n = 200$ or $n = 400$ and $p = 10$ or $p = 30$. In each case, 100 replications are drawn. Let $\boldsymbol{\ell}_1 = (1, 0, \ldots, 0)^\top$, $\boldsymbol{\ell}_2 = (0, 1, \ldots, 0)^\top$, and $\mathbf{L}_0 = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Example 3.2. We generate X_i from the uniform distribution $U(0,1)$. The response variable Y is generated from

$$\text{Model 3: } Y = \sin(2\pi \boldsymbol{\ell}_1^\top \mathbf{X}) + 4(\boldsymbol{\ell}_2^\top \mathbf{X} - 0.5)^2 + \sigma\epsilon. \tag{3.1}$$

Let $\boldsymbol{\ell}_1 = (1, -1, 1, 0, \ldots, 0)^\top/\sqrt{3}$ and $\boldsymbol{\ell}_2 = (1, 1, 0, \ldots, 0)^\top/\sqrt{2}$.

Example 3.3. Consider the model:

$$\text{Model 4: } Y = 1 + 2(\boldsymbol{\ell}_1^\top \mathbf{X})(\boldsymbol{\ell}_2^\top \mathbf{X})^2 + 2.5\exp\{-(\boldsymbol{\ell}_3^\top \mathbf{X})^2\} + \sigma\epsilon,$$

where X_i is generated from $U(-1,1)$. Let $\boldsymbol{\ell}_1 = (1, -1, 1, 0, \ldots, 0)^\top/\sqrt{3}$, $\boldsymbol{\ell}_2 = (1, 1, 0, 1, 0, \ldots, 0)^\top/\sqrt{3}$, and $\boldsymbol{\ell}_3 = (-1, 0, 1, 1, 0, \ldots, 0)^\top/\sqrt{3}$.

With regard to estimating the directions in the CMS, we compare the results of the proposed methods with those of the MAVE, OPG, SIR, SAVE and pHd approaches. The means and standard deviations of the estimation error $m(\hat{\mathbf{L}}_1, \mathbf{L}_0)$ for Models 1–4 are presented in Tables 1 and 2. Table 3 summarizes the means of the estimated bandwidths for Model 3, with the sample size n varying from 200 to 1600. It is seen clearly from Table 1 that the estimation errors of the proposed CVML estimates are usually smaller than those of the alternatives for Models 1 and 2, especially when p is large and n is small. The results in Table 2 also indicate that the CVML method performs comparably with existing methods in terms of estimating the directions of the CMS. In addition, Table 3 shows that the estimated bandwidths obtained by the CVML method become smaller as the sample size increases.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

We also compare the performance of the proposed method in terms of determining the structural dimension with that of the MAVE-based method (MAVE, Xia et al. 2002), the ridge-type ratio estimation (RRE, Xia et al. 2015), and the BIC (BIC, Zhu et al. 2006). For those methods that involve a tuning parameter, we use the values recommended in the literature. In particular, we take the ridge value $c_n = \log(n)/(10\sqrt{n})$ for the RRE and the penalty value $\alpha_n = \sqrt{n}$ for the BIC. Based on our limited simulation experiments, we recommend $\tau_n = 2.5n^{-1/3}$ for the CVML method. The frequencies of the estimated dimensions for Models 1–4 are presented in Table 4. Figure 1 presents the box plots of the ratio $\mathrm{CM}_n(\mathbf{M}_r)/\mathrm{CM}_n(\mathbf{M}_{r+1})$ for Models 3 and 4, with red horizontal straight lines representing $y = 1$. It is seen from the results in Table 4 that the CVML performs comparably with the MAVE and outperforms the other competitors, especially for Model 4, where the true dimension of the CMS is three. Figure 1 verifies the feasibility of the proposed estimation procedure.

Overall, the simulation results support our theoretical results, and the proposed CVML works reasonably well in terms of dimension reduction, including estimating both the directions and the dimension of the CMS.

4. Real-data illustration

4.1. London air quality data set

Air pollution may cause diseases, allergies, and even death. It occurs when harm-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Table 1: Means and standard deviations (in parentheses) of $m(\hat{\mathbf{L}}_1, \mathbf{L}_0)$ for Model 1 and 2

Model	p	n	CVML	MAVE	OPG	SIR	pHd	SAVE
1	10	200	0.402	0.552	0.499	0.567	0.551	1.327
			(0.095)	(0.170)	(0.154)	(0.134)	(0.104)	(0.058)
		400	0.223	0.319	0.279	0.375	0.385	1.227
			(0.049)	(0.069)	(0.070)	(0.071)	(0.072)	(0.114)
	30	200	0.581	1.021	1.038	1.030	1.095	1.389
			(0.058)	(0.110)	(0.111)	(0.109)	(0.103)	(0.018)
		400	0.475	0.785	0.779	0.728	0.778	1.400
			(0.051)	(0.168)	(0.182)	(0.089)	(0.098)	(0.011)
2	10	200	0.365	0.390	0.363	0.740	0.811	1.058
			(0.102)	(0.122)	(0.123)	(0.194)	(0.186)	(0.086)
		400	0.240	0.242	0.210	0.484	0.628	0.961
			(0.061)	(0.065)	(0.050)	(0.127)	(0.217)	(0.104)
	30	200	0.511	0.770	0.761	1.165	1.101	1.372
			(0.065)	(0.130)	(0.121)	(0.113)	(0.038)	(0.038)
		400	0.475	0.785	0.779	0.728	0.778	1.400
			(0.051)	(0.168)	(0.182)	(0.089)	(0.098)	(0.011)

Table 2: Means and standard deviations (in parentheses) of $m(\hat{\mathbf{L}}_1, \mathbf{L}_0)$ for Model 3 and 4 with $\sigma = 0.2$.

Model	p	n	CVML	MAVE	OPG	SIR	pHd	SAVE
3	10	200	0.144	0.142	0.143	0.958	0.355	0.983
			(0.064)	(0.039)	(0.039)	(0.090)	(0.069)	(0.078)
		400	0.076	0.091	0.089	0.925	0.249	0.672
			(0.021)	(0.025)	(0.025)	(0.106)	(0.035)	(0.214)
	30	200	0.275	0.354	0.402	1.132	0.753	1.342
			(0.048)	(0.076)	(0.079)	(0.067)	(0.108)	(0.054)
		400	0.168	0.168	0.183	1.043	0.479	1.223
			(0.022)	(0.027)	(0.027)	(0.041)	(0.052)	(0.078)
4	10	200	0.417	0.657	0.619	1.269	1.082	1.183
			(0.125)	(0.273)	(0.298)	(0.099)	(0.127)	(0.115)
		400	0.209	0.222	0.202	1.241	0.999	1.066
			(0.047)	(0.047)	(0.039)	(0.102)	(0.120)	(0.101)
	30	200	0.598	0.968	0.969	1.614	1.224	1.537
			(0.119)	(0.104)	(0.103)	(0.054)	(0.052)	(0.075)
		400	0.499	1.034	1.026	1.475	1.287	1.448
			(0.099)	(0.141)	(0.188)	(0.050)	(0.075)	(0.072)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Table 3: Means and standard deviations (in parentheses) of estimated bandwidths for Model 3

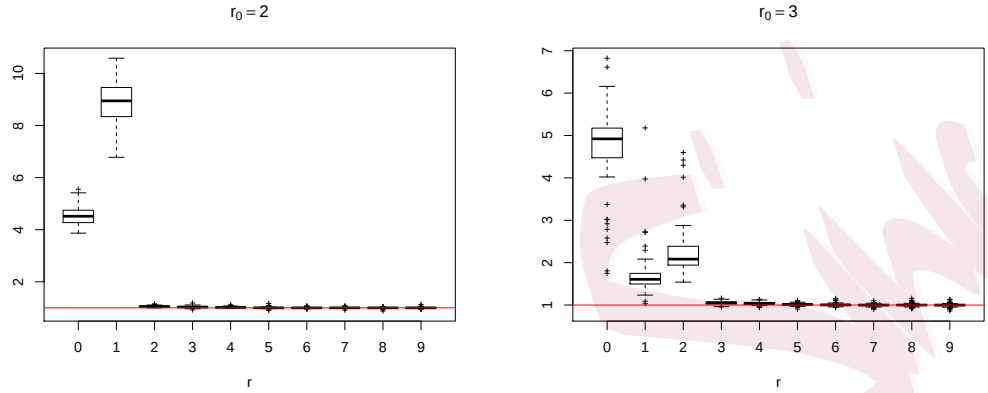
p	σ		$n = 200$	$n = 400$	$n = 800$	$n = 1600$
5	0.1	$\hat{h}_1 \times 10$	0.356	0.295	0.285	0.275
			(0.079)	(0.056)	(0.058)	(0.044)
		$\hat{h}_2 \times 10$	0.519	0.420	0.432	0.420
			(0.097)	(0.061)	(0.051)	(0.043)
	0.2	$\hat{h}_1 \times 10$	0.469	0.439	0.397	0.359
			(0.081)	(0.039)	(0.026)	(0.020)
		$\hat{h}_2 \times 10$	0.698	0.609	0.548	0.494
			(0.096)	(0.059)	(0.037)	(0.036)
10	0.1	$\hat{h}_1 \times 10$	0.313	0.291	0.274	0.266
			(0.056)	(0.053)	(0.052)	(0.050)
		$\hat{h}_2 \times 10$	0.452	0.415	0.414	0.404
			(0.072)	(0.048)	(0.047)	(0.040)
	0.2	$\hat{h}_1 \times 10$	0.327	0.323	0.300	0.288
			(0.073)	(0.055)	(0.053)	(0.052)
		$\hat{h}_2 \times 10$	0.483	0.471	0.460	0.443
			(0.099)	(0.056)	(0.043)	(0.034)

Table 4: Frequencies of estimated dimension for Models 1–4 with $p = 10$

Model		$n = 200$			$n = 400$		
		$\hat{r} < r_0$	$\hat{r} = r_0$	$\hat{r} > r_0$	$\hat{r} < r_0$	$\hat{r} = r_0$	$\hat{r} > r_0$
1	CVML	0.04	0.87	0.09	0	1	0
	MAVE	0	0.77	0.23	0	0.99	0.01
	BIC	0	0	1	0	0	1
	RRE	0.44	0.48	0.08	0.27	0.72	0.01
2	CVML	0.24	0.73	0.03	0.05	0.95	0
	MAVE	0.19	0.81	0	0.03	0.97	0
	BIC	0	0	1	0	0	1
	RRE	0.44	0.31	0.25	0.44	0.54	0.02
3	CVML	0	1	0	0	1	0
	MAVE	0	1	0	0	1	0
	BIC	0	0	1	0	0	1
	RRE	0.98	0.01	0.01	0.99	0.01	0
4	CVML	0.16	0.82	0.02	0	1	0
	MAVE	0.92	0.08	0	0.71	0.29	0.01
	BIC	0	0.56	0.44	0	0.15	0.85
	RRE	0.80	0.05	0.15	0.99	0.01	0

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Figure 1: Box plots of the ratio $\mathrm{CM}_n(\mathbf{M}_r)/\mathrm{CM}_n(\mathbf{M}_{r+1})$ for Model 3 (left) and Model 4 (right) with $n = 400$ and $p = 10$



ful substances, including particulates, liquid droplets, gases, and chemical molecules produced by human activity, are introduced into the atmosphere. Pollutants are classified mainly as primary and secondary substances. Primary pollutants are usually generated from a chemical process, such as the sulfur dioxide released from factories. Secondary pollutants form in the air when primary pollutants react or interact. Ground-level ozone (O_3) is a prominent example of a secondary pollutant.

In this study, we attempt to exploit the relationship between primary pollutants and the secondary pollutant O_3, based on the London air quality data in the R package *openair*. After deleting some missing data, the data set is collected from May 1, 1998, to September 30, 2004, with hourly updated records of wind speed (x_1), wind direction (x_2), oxides of nitrogen concentration NO_x (x_3), nitrogen dioxide concentration NO_2 (x_4), particulate PM10 in ug/m^3 (x_5), sulfur dioxide concentration SO_2

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

(x_6), carbon monoxide concentration CO (x_7), particulate PM2.5 in ug/m^3 (x_8), and Ozone concentration (Y). We convert the hourly level data to its daily average for all variables, and apply the CVML method to the treated daily level data set.

The structural dimension estimated by the CVML procedure is $\hat{r} = 1$. From Table 5, the direction estimated by the CVML approach indicates that NO_x and NO_2 have significant effects on O_3 concentration. This provides empirical evidence for the claim that secondary pollutants are usually products of the reactions of primary pollutants under certain environmental conditions. Nevertheless, the effects of wind speed, wind direction, and particulates seem not to be very significant.

Table 5: Estimated directions in CMS for London air quality data set

Direction	x_1 (ws)	x_2 (wd)	x_3 (NO_x)	x_4 (NO_2)	x_5 (PM10)	x_6 (SO_2)	x_7 (CO)	x_8 (PM2.5)
$\hat{\boldsymbol{\ell}}_1$	-0.045	0.006	0.948	-0.271	0.000	0.012	-0.161	-0.006

4.2. Beijing PM2.5 data set

Many cities experience hazy weather. The PM2.5—particulate matter less than 2.5 μm in diameter—is known to influence human health and the atmospheric climate. Epidemiological experts concluded that exposure to PM2.5 over a few hours to weeks can cause cardiovascular disease, and longer-term exposure increases the risk for cardiovascular mortality and even shortens the life span. In this real-data analysis, we investigate the factors that affect the PM2.5 concentration. We analyze the Beijing PM2.5 data set collected at the Aotizhongxin air-quality monitoring site. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

data set is downloadable from UCI database with the link https://archive.ics.uci.edu/ml/datasets/Beiji
Site+Air-Quality+Data. The PM2.5 (Y) data ranging from March 2013 to February 2017 are converted to daily averaged records, with potential affecting factors PM10 (x_1), SO_2 concentration (x_2), NO_2 concentration (x_3), CO concentration (x_4), O_3 concentration (x_5) in ug/m^3, temperature (x_6), pressure (x_7), dew point temperature (x_8), precipitation (x_9), and wind speed (x_{10}).

The structural dimension estimated by the CVML method is $\hat{r} = 3$. The three estimated directions are shown in Table 6. Note that the first three eigenvalues of $\hat{\mathbf{M}}$ are 154.447, 36.138, and 9.933. Therefore, the first direction is very important to reveal the relationship between the PM2.5 and the potential affecting factors. The first direction in Table 6 clearly indicates that the PM10, temperature, and dew point temperature are crucial variables associated with the PM2.5 concentration. The latter two factors were also identified by Zhang et al. (2017). The last two directions reveal that the PM2.5 concentration has weak relationships with pressure and wind speed, but is possibly related to NO_2 and CO, which are potential chemical components of the PM2.5.

Table 6: Estimated directions in CMS for Beijing PM2.5 data set

Direction	x_1 (PM10)	x_2 (SO_2)	x_3 (NO_2)	x_4 (CO)	x_5 (O_3)	x_6 (temp)	x_7 (pres)	x_8 (dewp)	x_9 (preci)	x_{10} (ws)
$\hat{\boldsymbol{\ell}}_1$	−0.540	0.002	0.079	−0.128	−0.029	0.410	−0.034	−0.717	0.025	0.003
$\hat{\boldsymbol{\ell}}_2$	−0.795	−0.139	−0.231	0.040	−0.109	−0.209	−0.034	0.446	−0.193	0.018
$\hat{\boldsymbol{\ell}}_3$	0.091	−0.504	−0.231	−0.492	−0.136	0.403	−0.159	0.250	0.397	−0.125

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

5. Discussion

In this study, we attempt to reduce the dimension of multiple-index models in the framework of metric learning. The proposed cross-validation-based metric learning method produces a metric that contains crucial information on both the central mean subspace and the unknown link function. The rate of convergence and the optimal order of the bandwidth are derived. A novel algorithm is proposed to determine the structural dimension of the CMS when it is unknown. For the purpose of prediction, we refer to the work of Conn and Li (2019), who show that the kernel estimate using a full bandwidth matrix achieves the optimal rate of convergence for a multiple-index model.

Appendix: Assumptions and remarks

Let $\mathbf{A}^{\perp}$ denotes the space orthogonal to that spanned by the column vectors of the matrix $\mathbf{A}$. The following regularity conditions are imposed.

(C1) [*Design of* $\mathbf{X}$.] The density function $f_{\mathbf{X}}(\mathbf{x})$ of $\mathbf{X}$ is positive, bounded and is continuously differentiable up to order two.

(C2) [*Link function.*] The second-order derivatives of $g(\cdot)$ exist and are bounded away from infinity.

(C3) [*Kernel function.*] The kernel function $K(\cdot)$ is a symmetric univariate density function with bounded derivatives.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

(C4) [*Identifiability.*] Let $\mathcal{F} = \{\mathbf{t} \in \mathbb{R}^p : \mathbf{t} \in \mathbf{L}_0^{\perp}\}$. For any $\mathbf{x} \in \mathbb{R}^p$, if $f(\mathbf{x}+c\mathbf{t}) = f(\mathbf{x})$ for all $c \in \mathbb{R}$, then it must have $\mathbf{t} \in \mathcal{F}$.

(C5) [*Moments of errors.*] The error satisfies $E(\epsilon_i|\mathbf{X}_i) = 0$, $E(\epsilon_i^2|\mathbf{X}_i) = \sigma^2(\mathbf{L}_0^{\top}\mathbf{X}_i) = \sigma_i^2$ almost surely and $\sup_i E(|\epsilon_i|^4) < \infty$ for all i, where $\sigma^2(\cdot)$ is bounded and continuous.

Condition (C1) is a relatively weaker assumption on the density of $\mathbf{X}$, compared with the linearity condition in many SIR-based methods. Conditions (C2)–(C3) are common conditions on the nonparametric link function and the kernel function, respectively. Condition (C3) is satisfied by many commonly-used kernel functions, such as the biweight kernel and the quadratic kernel. The subspace $\mathcal{F}$ in Condition (C4) indeed equals to the space orthogonal to $\mathcal{S}(\mathbf{L}_0)$. Hence, Condition (C4) indicates that the dimension r_0 cannot be further reduced and the regression function $f(\mathbf{x})$ remains constant when $\mathbf{x}$ varies in $\mathcal{F}$. For more insights into condition (C4), we consider a toy example in which $\mathbf{t} = (t_1, t_2)^{\top}$, $r_0 = 2$ and $f(\mathbf{x}) = (x_1 + x_2)^2$. By choosing $t_1 = -t_2$, we have $f(\mathbf{x} + c\mathbf{t}) = f(\mathbf{x})$ for all $c \in \mathbb{R}$ and $\mathbf{t} \in \mathcal{F}$ in this instance. The moment assumption up to the fourth order in condition (C5) is made for technical simplicity.

Supplementary Material

The technical proofs are provided in the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Acknowledgments

The authors are grateful to the editor, associate editor, and two anonymous reviewers for their insightful comments and suggestions that have substantially improved the content of this paper. Linlin Dai's research was supported by the Fundamental Research Funds for the Central Universities (Grant No. JBK1806002) of China and the National Natural Science Foundation of China (Grant No. 12001441). Kani Chen's research was supported by the Hong Kong Research Grant Council grants 16309816, 16300714, 600813, and 600612. The research of Gang Li was partly supported by the National Institute of Health Grants 1P01CA236585-01A1, 5P30CA-16042, UTR001881A, and NIH CA211015. Yuanyuan Lin's research was supported by the Hong Kong Research Grants Council (Grant No. 14306219 and 14306620), the National Natural Science Foundation of China (Grant No. 11961028), and Direct Grants for Research, The Chinese University of Hong Kong.

References

Akritas, M. G. (2016). Projection pursuit multi-index (PPMI) models. *Stat. Probabil. Lett.* **114**, 99–103.

Alquier, P. and Biau, G. (2013). Sparse single-index model. *J. Mach. Learn. Res.* **14**, 243–280.

Bellet, A., Habrard, A. and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.

Bura, E. (2003). Using linear smoothers to assess the structural dimension of regressions. *Statist. Sinica*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

13, 143–162.

Bura, E. and Cook, R. D. (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 393–410.

Bura, E. and Cook, R. D. (2001b). Extending sliced inverse regression: The weighted chi-squared test. *J. Am. Statist. Assoc.* **96**, 996-1003.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 959–1035.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477–489.

Chen, D., Hall, P. and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.* **39**, 1720–1747.

Conn, D. and Li, G. (2019). An oracle property of the Nadaraya-Watson kernel estimator for high dimensional nonparametric regression. *Scand. J. Statist.* **46**, 735–764.

Cook, R. D. (1998). Principal hessian directions revisited. *J. Amer. Statist. Assoc.* **93**, 84–100.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–474.

Cook, R. D. and Li, B. (2004). Determining the dimension of iterative Hessian transformation. *Ann. Statist.* **32**, 2501–2531.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

approach. *J. Amer. Statist. Assoc.* **100**, 410–428.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273–297.

Cover T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27.

Dalalyan, A. S., Juditsky, A. and Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.* **9**, 1648–1678.

Goldberger, J., Roweis, S., Hinton, G. and Salakhutdinov, R. (2005). Neighbourhood components analysis. In *Advances in NIPS* **17**, 513–520.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17**, 573–588.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986–995.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157–178.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29**, 1537–1566.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Amer. Statist. Assoc.* **86**, 316–342.

——— (1992). On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025–1039.

Luo, R., Wang, H. and Tsai, C. L. (2009). Contour projected dimension reduction. *Ann. Statist.* **37**, 3743–

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

3778.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–887.

Ma, Y. and Zhu, L. P. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107**, 168–179.

——— (2014). On estimation efficiency of the central mean subspace. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76**, 885–901.

Noh, Y. K., Sugiyama, M., Kim, K. E., Park, F. and Lee, D. D. (2017). Generative local metric learning for kernel regression. In *Advances in NIPS*, 2452–2462.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88**, 836–847.

Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94**, 285–296.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811–821.

Weinberger, K. Q., Blitzer, J. and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in NIPS* **18**, 1473–1480.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **10**, 207–244.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Weinberger, K. Q. and Tesauro, G. (2007). Metric learning for kernel regression. In *Proceedings of International Conference on AISTATS*, 612–619.

Xia, Q., Xu, W. and Zhu, L. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statist. Sinica* **25**, 1025–1044.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–2690.

——— (2008). A multiple-index model and dimension reduction. *J. Amer. Statist. Assoc.* **103**, 1631–1640.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space (with Discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 363–410.

Xing, E. P., Jordan, M. I., Russell, S. J. and Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in NIPS* **14**, 521–528.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–979.

Yin, X. and Cook, R. D. (2006). Dimension reduction via marginal high moments in regression. *Statistics & Probability Letters* **76**, 393–400.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proc. Math. Phys. Eng. Sci.* **473**, 20170457.

Zhu, L. P. and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *J. Mult. Anal.* **98**, 970–991.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Zhu, L. P., Yu, Z. and Zhu, L. X.(2010). A sparse eigen-decomposition estimation in semi-parametric regression. *Comp. Statist. Data Anal.* **54**, 976–986.

Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–643.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Statist. Assoc.* **101**, 1638-1651.

Zhu, X., Guo, X., Wang, T. and Zhu, L. X. (2020). Dimensionality determination: A thresholding double ridge ratio approach. *Comp. Statist. Data Anal.* **146**, 106910.

Zhu, X., Kang, Y. and Liu, J. (2019). Estimation of the number of endmembers via thresholding ridge ratio criterion. *IEEE Trans. Geosci. Remote Sens.* **58**, 637–649.

Linlin Dai

Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

E-mail: ldaiab@swufe.edu.cn

Kani Chen

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China.

E-mail: makchen@ust.hk

Gang Li

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0398

Department of Biostatistics, School of Public Health, University of California at Los Angeles, CA 90095-1772, USA

E-mail: vli@ucla.edu

Yuanyuan Lin

Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China

E-mail: ylin@sta.cuhk.edu.hk