Nonparametric Variable Selection via Sufficient Dimension Reduction

Lexin Li

NC State University
DEPARTMENT OF STATISTICS

Workshop on Current Trends and Challenges in Model Selection and Related Areas
Vienna, Austria July 24, 2008

Outline

- Introduction to model free variable selection
- Introduction to sufficient dimension reduction (SDR)
- Regularizatied SDR for variable selection
- Simulation study and real data analysis
- Concluding remarks and discussions

Joint work with Dr. Bondell at NCSU

Introduction to Model Free Variable Selection

Model based variable selection:

- most existing variable selection approaches are model based, i.e., we assume the underlying true model is known up to a finite dimensional parameter, or the imposed working model is usefully close to the true model

Potential limitations:

- the true model is unknown, and model formulation can be complex
- accessing the goodness of model fitting can be difficult when interweaved with model building and selection

Introduction to Model Free Variable Selection

Model free variable selection: selection that does *not* require any traditional model

A "disclaimer": model free variable selection in the exploratory stage of the analysis; refined by model based variable selection approaches

Sufficient dimension reduction: model free variable selection is to be achieved through the framework of SDR (Li, 1991, 2000, Cook, 1998)

Introduction to SDR

General framework of SDR:

- study conditional distribution of $Y \in \mathbb{R}^r$ given $X \in \mathbb{R}^p$
- find a $p \times d$ matrix $\eta = (\eta_1, \ldots, \eta_d)$, $d \leq p$, such that

$$Y|X \overset{\mathcal{D}}{=} Y|\eta^{\mathsf{T}}X \qquad \Leftrightarrow \qquad Y \perp\!\!\!\perp X|\eta^{\mathsf{T}}X$$

- replace X with $\eta^{\mathsf{T}}X$ without loss of information on regression $Y|X$

Key concept: central subspace $\mathcal{S}_{Y|X}$

- $Y \perp\!\!\!\perp X|\eta^{\mathsf{T}}X \quad \Rightarrow \quad \mathcal{S}_{DRS} = \mathsf{Span}(\eta) \quad \Rightarrow \quad \mathcal{S}_{Y|X} = \cap\, \mathcal{S}_{DRS}$
- $\mathcal{S}_{Y|X}$ is a parsimonious population parameter that captures all regression information of $Y|X$; main object of interest in SDR

Introduction to SDR

Known regression models:

- single / multi-index model: $Y = f_1(\eta_1^{\mathsf{T}} X) + \ldots + f_d(\eta_d^{\mathsf{T}} X) + \varepsilon$
- heteroscedastic model: $Y = f(\eta_1^{\mathsf{T}} X) + g(\eta_2^{\mathsf{T}} X)\varepsilon$
- logit model: $\log\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right) = f(\eta_1^{\mathsf{T}} X) \quad \ldots$

Existing SDR estimation methods:

- sliced inverse regression (SIR), sliced average variance estimation (SAVE), principal Hessian directions (PHD), ...
- inverse regression estimation (IRE), covariance inverse regression estimation (CIRE), ...

A Simple Motivating Example

Consider a response model:

$$Y = \exp(-0.5\eta_1^{\mathsf{T}} X) + 0.5\varepsilon$$

- all predictors X and error ε are independent standard normal
- $\mathcal{S}_{Y|X} = \text{Span}(\eta_1)$, where $\eta_1 = (1, -1, 0, \ldots, 0)^{\mathsf{T}}/\sqrt{2}$
- CIRE estimate ($n = 100, p = 6$):
  $(0.659, -0.734, -0.128, 0.097, -0.015, 0.030)^{\mathsf{T}}$

Observations:

- produce linear combinations of *all* the predictors
- interpretation can be difficult; no variable selection

Goal: do variable selection by obtaining sparse SDR estimate

Key Ideas

Regularized SDR estimation:

- observe that the majority of SDR estimators can be formulated as a genearlized spectral decomposition problem
- transform the spectral decomposition into an equivalent least squares formulation
- add L_1 penalty to the least squares

Focus of our work:

- demonstrate that the resulting model free variable selection can achieve the usual selection consistency under the usual conditions as in the case of model based variable selection (say, e.g., the multiple linear regression model)

Minimum Discrepancy Approach

Generalized spectral decomposition formulation:

$$\Omega\beta_j = \lambda_j \Sigma \beta_j, \quad j = 1, \ldots, p$$

where Ω is a $p \times p$ positive semi-definite symmetric matrix, $\Sigma = \mathrm{cov}(X)$. For instance, $\Omega_{\mathsf{SIR}} = \mathrm{cov}[\mathrm{E}\{X - \mathrm{E}(X)|Y\}]$, $\Omega_{\mathsf{SAVE}} = \Sigma - \mathrm{cov}(X|Y)$, $\Omega_{\mathsf{PHD}} = \mathrm{E}[\{Y - \mathrm{E}(Y)\}\{X - \mathrm{E}(X)\}\{X - \mathrm{E}(X)\}^{\mathsf{T}}]$. It satisfies that

$$\Sigma^{-1}\mathsf{Span}(\Omega) \subseteq \mathcal{S}_{Y|X}$$

Assumptions:

- above SDR methods impose assumptions on the marginal distribution of X, instead of the conditional distribution of $Y|X$
- model free

Minimum Discrepancy Approach

An equivalent least squares optimization formulation: consider

$$\min_{\eta,\gamma} L(\eta,\gamma) = \min_{\eta^{p\times d},\gamma^{d\times h}} \sum_{j=1}^{h} (\theta_j - \eta\gamma_j)^{\mathsf{T}} \Sigma (\theta_j - \eta\gamma_j),$$

subject to $\eta^{\mathsf{T}}\Sigma\eta = I_d$, Let $(\tilde{\eta},\tilde{\gamma}) = \arg\min_{\eta,\gamma} L(\eta,\gamma)$. Then $\tilde{\eta}$ consists of the first d eigenvectors $(\beta_1,\ldots,\beta_d)$ from the eigen decomposition

$$\Omega\beta_j = \lambda_j \Sigma\beta_j, \quad j = 1,\ldots,p,$$

where $\Omega = \Sigma\left(\sum_{j=1}^{h}\theta_j\theta_j^{\mathsf{T}}\right)\Sigma$.

In matrix form:

$$L(\eta,\gamma) = \{\mathrm{vec}(\theta) - \mathrm{vec}(\eta\gamma)\}^{\mathsf{T}} V \{\mathrm{vec}(\theta) - \mathrm{vec}(\eta\gamma)\}$$

where $V = I_h \otimes \Sigma$.

Minimum Discrepancy Approach

A minimum discrepancy formulation:

- start with the construction of a $p \times h$ matrix $\theta = (\theta_1, \ldots, \theta_h)$, such that, Span$(\theta) \subseteq \mathcal{S}_{Y|X}$; given data, construct a $\sqrt{n}$-consistent estimator $\hat{\theta}$ of θ
- construct a positive definite matrix $V^{ph \times ph}$, and a $\sqrt{n}$-consistent estimator $\hat{V}$ of V
- estimate (η, γ) by minimizing a quadratic discrepancy function:

$$(\hat{\eta}, \hat{\gamma}) = \arg \min_{\eta^{p \times d}, \gamma^{d \times h}} \left\{ \mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\eta\gamma) \right\}^{\mathsf{T}} \hat{V} \left\{ \mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\eta\gamma) \right\}$$

- Span$\{\hat{\eta}\}$ forms a consistent *inverse regression estimator* of $\mathcal{S}_{Y|X}$
- Cook and Ni (2005)

Minimum Discrepancy Approach

A whole class of estimators: its individual member is determined by the choice of the pair (θ, V) and $(\hat{\theta}, \hat{V})$; for instance,

- for sliced inverse regression (SIR):

$$\begin{aligned} \theta_s &= f_s \Sigma^{-1} \{ \mathrm{E}(X|J_s = 1) - \mathrm{E}(X) \}, \\ V &= \mathrm{diag}(f_s^{-1}) \otimes \Sigma \end{aligned}$$

- for covariance inverse regression estimation (CIRE):

$$\begin{aligned} \theta_s &= \Sigma^{-1} \mathrm{cov}(Y J_s, X), \\ V &= \Gamma^{-1}, \end{aligned}$$

where Γ is the asymptotic covariance of $n^{1/2}\{\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta)\}$

Regularized Minimum Discrepancy Approach

Proposed regularization solution:

- let $\alpha = (\alpha_1, \ldots, \alpha_p)^{\mathsf{T}}$ denote a $p \times 1$ shrinkage vector, given $(\hat{\theta}, \hat{\eta}, \hat{\gamma})$

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\mathrm{diag}(\alpha)\hat{\eta}\hat{\gamma}) \right\}^{\mathsf{T}} \hat{V} \left\{ \mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\mathrm{diag}(\alpha)\hat{\eta}\hat{\gamma}) \right\},$$

$$\text{subject to } \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tau \geq 0.$$

- $\mathrm{Span}\{\mathrm{diag}(\hat{\alpha})\hat{\eta}\}$ is called the *shrinkage inverse regression estimator* of $\mathcal{S}_{Y|X}$.
- note that:
  - when $\tau \geq p$, $\hat{\alpha}_j = 1$ for all j's
  - when τ decreases, some $\hat{\alpha}_j$'s are shrunken to zero, which in turn shrinking the entire rows of η

Regularized Minimum Discrepancy Approach

Additional notes:

- generalized the shrinkage SIR estimator of Ni, Cook, and Tsai (2005)
- closely related to nonnegative garrote (Breiman, 1995)
- $Pr(\hat{\alpha}_j \geq 0) \rightarrow 1$ for all j's
- use an information-type criterion to select the tuning parameter τ
- achieve simultaneous dimension reduction and variable selection

Regularized Minimum Discrepancy Approach

Optimization:

$$\arg\min_{\alpha} n \left\{ \mathrm{vec}(\hat{\theta}) - \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix} \alpha \right\}^{\mathsf{T}} \hat{V} \left\{ \mathrm{vec}(\hat{\theta}) - \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix} \alpha \right\}.$$

It becomes a "standard" lasso problem, with the "response" U^{ph}, and the "predictors" $W^{ph\times p}$:

$$U = \sqrt{n}\hat{V}^{1/2}\mathrm{vec}(\hat{\theta}), \qquad W = \sqrt{n}\hat{V}^{1/2} \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix},$$

The optimization is easy.

Variable Selection without a Model

Goal: to seek the smallest subset of the predictors $X_{\mathcal{A}}$, with partition $X = (X_{\mathcal{A}}^{\mathsf{T}}, X_{\mathcal{A}^c}^{\mathsf{T}})^{\mathsf{T}}$, such that

$$Y \perp\!\!\!\perp X_{\mathcal{A}^c} | X_{\mathcal{A}}$$

Here $\mathcal{A}$ denotes a subset of indices of $\{1, \ldots, p\}$ corresponding to the relevant predictor set $X_{\mathcal{A}}$, and $\mathcal{A}^c$ is the compliment of $\mathcal{A}$.

Existence and uniqueness: Given the existence of the central subspace $\mathcal{S}_{Y|X}$, $\mathcal{A}$ uniquely exists.

Variable Selection without a Model

Relation between $\mathcal{A}$ and basis of $\mathcal{S}_{Y|X}$: (Cook, 2004, Proposition 1)

$$\eta^{p\times d} = \begin{pmatrix} \eta_{\mathcal{A}} \\ \eta_{\mathcal{A}^c} \end{pmatrix}, \eta_{\mathcal{A}} \in \mathbb{R}^{(p-p_0)\times d}, \eta_{\mathcal{A}^c} \in \mathbb{R}^{p_0\times d}.$$

The rows of a basis of the central subspace corresponding to $X_{\mathcal{A}^c}$, i.e., $\eta_{\mathcal{A}^c}$, are all zero vectors; and all the predictors whose corresponding rows of the $\mathcal{S}_{Y|X}$ basis equal zero belong to $X_{\mathcal{A}^c}$.

Variable Selection without a Model

Partition:

$$\theta^{p\times h} = \begin{pmatrix} \theta_{\mathcal{A}} \\ \theta_{\mathcal{A}^c} \end{pmatrix}, \theta_{\mathcal{A}} \in \mathbb{R}^{(p-p_0)\times h}, \theta_{\mathcal{A}^c} \in \mathbb{R}^{p_0\times h},$$

Re-describe $\mathcal{A}$ as: $\mathcal{A} = \{j : \theta_{jk} \neq 0 \text{ for some } k, 1 \leq j \leq p, 1 \leq k \leq h\}$

For the shrinkage inverse regression estimator: define
$\hat{\mathcal{A}} = \{j : \tilde{\theta}_{jk} \neq 0 \text{ for some } k,\ 1 \leq j \leq p, 1 \leq k \leq h\}$, where
$\tilde{\theta} = \text{diag}(\hat{\alpha})\hat{\eta}\hat{\gamma}$

Asymptotic Properties

Theorem:

1. Assume that the initial estimator satisfies that $\sqrt{n}$ $\left\{\text{vec}(\hat{\theta}) - \text{vec}(\theta)\right\} \to N(0, \Gamma)$, for some $\Gamma > 0$, and that $\hat{V}^{1/2} = V^{1/2} + o(1/\sqrt{n})$.
2. Assume that $\lambda \to \infty$ and $\lambda/\sqrt{n} \to 0$

Then the shrinkage inverse regression estimator satisfies that:

1. Consistency in variable selection: $Pr(\hat{\mathcal{A}} = \mathcal{A}) \to 1$
2. Asymptotic normality: $\sqrt{n}\left\{\text{vec}(\tilde{\theta}_{\mathcal{A}}) - \text{vec}(\theta_{\mathcal{A}})\right\} \to N(0, \Lambda)$, for some $\Lambda > 0$

A Simulation Study

Motivating example revisited:

- $\mathcal{S}_{Y|X} = \text{Span}(\eta_1)$, where $\eta_1 = (0.707, -0.707, 0, \ldots, 0)^\top$
- CIRE estimate: $(0.659, -0.734, -0.128, 0.097, -0.015, 0.030)^\top$
- Shrinkage CIRE estimate: $(0.663, -0.748, 0, 0, 0, 0)^\top$

Another simulation example:

$$Y = \text{sign}(\eta_1^\top X)\log(|\eta_2^\top X + 5|) + 0.2\varepsilon,$$

- all predictors X and error ε are independent standard normal
- $\eta_1 = (1, \ldots, 1_q, 0, \ldots, 0)^\top$, $\eta_2 = (0, \ldots, 0, 1, \ldots, 1_q)^\top$, $q = 1, 5, 10$, $\mathcal{S}_{Y|X} = \text{Span}(\eta_1, \eta_2)$
- $n = 200/400$, $p = 20$

A Simulation Study

Table 1: Finite sample performance of the shrinkage CIRE estimator.

		# actives		positive rate		vector correlation	
		true	est	true	false	S-CIRE	CIRE
$q = 1$	$n = 200$	2	3.31	1.000	0.073	0.989	0.879
	$n = 400$	2	2.49	1.000	0.027	0.999	0.951
$q = 5$	$n = 200$	10	11.19	0.997	0.122	0.934	0.909
	$n = 400$	10	10.40	1.000	0.040	0.979	0.961
$q = 10$	$n = 200$	20	18.91	0.946	–	0.794	0.884
	$n = 400$	20	19.96	0.998	–	0.932	0.953

A Real Data Analysis

Automobile data:

- response: car price in log scale
- predictors: wheelbase, length, width, height, curb weight, engine size, bore, stroke, compression ratio, horsepower, peak rpm, city mpg, highway mpg; $p = 13$
- $n = 195$ observations

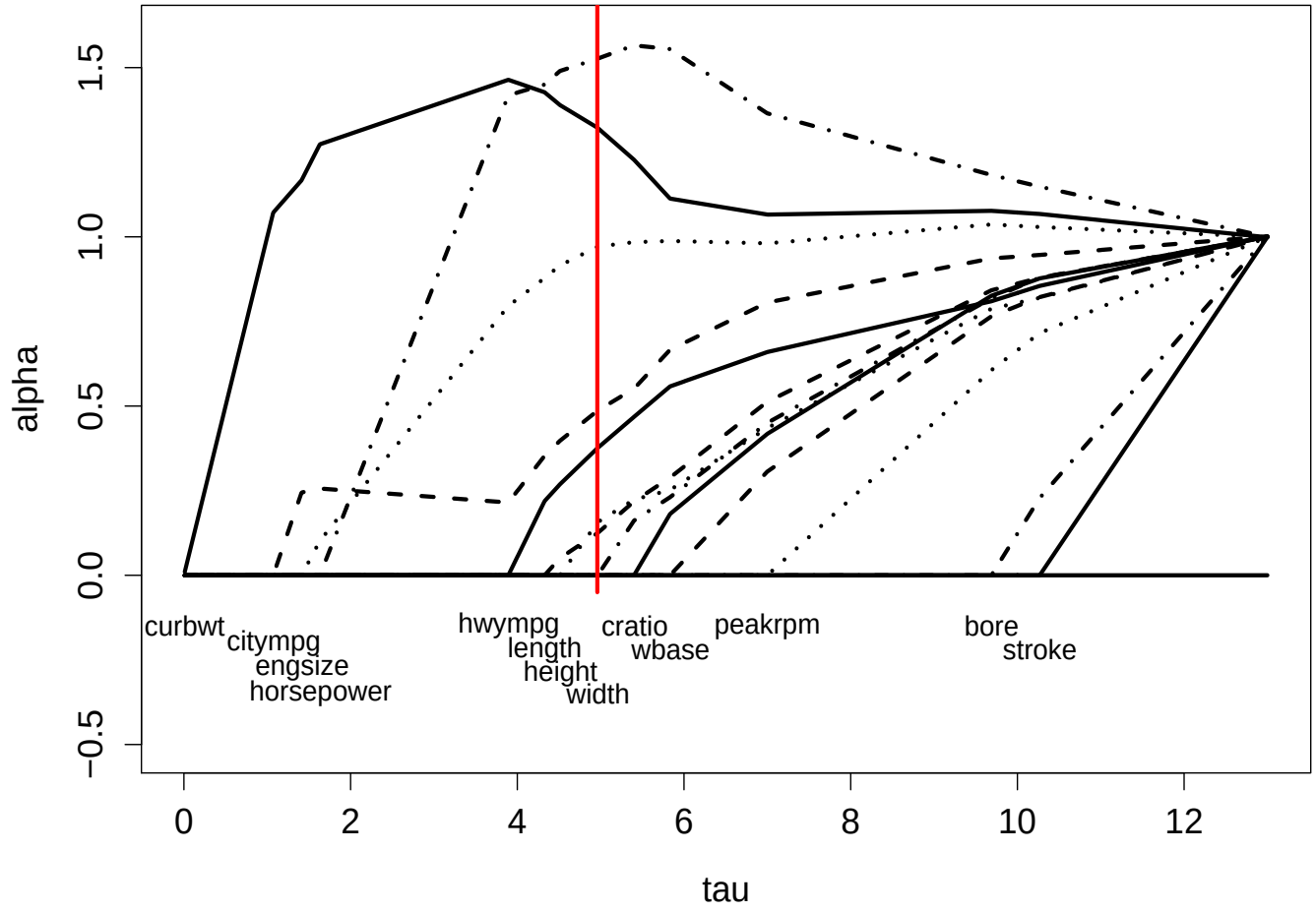

Figure 1: Solution paths for the automobile data.

Discussions

Summary:

- select relevant predictors consistently, without assuming a model
- the basis estimation corresponding to the relevant predictors is $\sqrt{n}$-consistent
- apply to a wide class of SDR methods

Reference:

Bondell, H.D., and Li, L. (2008). Shrinkage inverse regression estimation for model free variable selection. *Journal of the Royal Statistical Society, Series B.*, accepted.

NSF grant DMS 0706919

Thank You!