Trace Pursuit: A General Framework for Model-Free Variable Selection

Lixing Zhu

Jointly with Zhou Yu and Yuexiao Dong

Hong Kong Baptist University

October 16, 2014

Outline

Outline

Variable selection

- Classical variable selection: Nonnegative garrotte (Breiman, 1995), LASSO, adaptive LASSO and group LASSO(Tibshirani, 1996; Zou, 2006; Yuan and Lin, 2006), SCAD (Fan and Li, 2001), Dantzig selector (Candés and Tao, 2007), and MCP (Zhang, 2010), etc.
- Model-free selection: Let $\mathbf{X} = (x_1, \cdots, x_p)^T$ be the vector of predictors and Y be the scalar response. Define $\mathcal{A}^c$ the complement of $\mathcal{A}$ in the index set $\mathcal{I} = \{1, \cdots, p\}$, $\mathbf{X}_{\mathcal{A}} = \{x_i : i \in \mathcal{A}\}$ is the set of active predictors, and $\mathbf{X}_{\mathcal{A}^c} = \{x_i : i \in \mathcal{A}^c\}$ the set of inactive predictors.
- Why model-free selection? Hard to get idea about data structure in high-dimensional scenario.

Variable selection

- Consider the nonparametric dimension reduction sparse model

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}} \mathbf{X}, \text{ and } Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}, \tag{1}$$

  Here where $\perp\!\!\!\perp$ stands for independence, $P_{(\cdot)}$ denotes the q-dimensional projection operator with respect to the standard inner product.

- A special model is $Y = G(\beta_1^T X, \cdots, \beta_q^T X, \epsilon)$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)$ is an $p \times q$ orthonormal matrix, where q is unknown.

Variable selection: $p < n$

- Sufficient dimension reduction-based methods: regularized SIR (Li and Yin, 2008); coordinate-independent sparse dimension reduction (CISE) (Chen et al., 2010); ($p < n$)
- Sequential test-based methods: Cook (2004), Shao et al. (2007), and Li et al. (2005).
- All these methods rely on an initial estimator of the central space $S_{Y|\mathbf{X}}$. Then $p > n$? not possible.
- Methods with no such an initial estimator in demand for ultra-high dimensional paradigms.

Variable selection: $p > n$

- Correlation pursuit (Zhong et al. (2012))
- It is based on SIR, and then would miss significant variables linked to Y in a quadratic function or interaction.
- SIRI (Jiang and Liu (2013)): SIRI is also based on SIR, but can deal with a model with interaction terms.
- Both involve estimating the structural dimension q of $S_{Y|X}$, which makes the problem very challenging when $p > n$.

Outline

Trace pursuit: the basic procedure

- Design method-specific (SIR, SAVE, or DR) trace tests;
- Basic algorithm: Stepwise trace pursuit (STP) algorithm. Like stepwise regression selecting predictors into model forward and backward. Time consuming when p is large!
- A more efficient two-stage hybrid trace pursuit (HTP) algorithm: 1). Forward trace pursuit (FTP) algorithm forward selection with a modified BIC criterion to get a chosen model containing all active predictors as a screening step. 2). STP is followed for the refined variable selection.

Trace pursuit: An example of the basic procedure

- Assume $E(\mathbf{X}) = \mathbf{0}$ and $E(Y) = 0$.
- The principle of the SIR-based trace pursuit. For working index set $\mathcal{F}$ and index $j \in \mathcal{F}^c$, check whether or not

$$Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}} \tag{2}$$

- For any index set $\mathcal{F}$, denote $\mathbf{X}_{\mathcal{F}} = \{x_i : i \in \mathcal{F}\}$, $\mathrm{var}(\mathbf{X}_{\mathcal{F}}) = \mathbf{\Sigma}_{\mathcal{F}}$, and $\mathbf{U}_{\mathcal{F},h} = E(\mathbf{X}_{\mathcal{F}} | Y \in J_h)$. Define the SIR matrix $\mathbf{M}_{\mathcal{F}}^{sir}$ as

$$\mathbf{M}_{\mathcal{F}}^{sir} = \mathbf{\Sigma}_{\mathcal{F}}^{-1/2} \left(\sum_{h=1}^{H} p_h \mathbf{U}_{\mathcal{F},h} \mathbf{U}_{\mathcal{F},h}^T \right) \mathbf{\Sigma}_{\mathcal{F}}^{-1/2}. \tag{3}$$

Trace pursuit: An example of the basic procedure

- Denote $\mathcal{F} \cup j$ as the index set of j together with all the indices in $\mathcal{F}$. Given that $\mathbf{X}_{\mathcal{F}}$ is already in the model, **check the contribution of the additional variable x_j to Y by** $tr(\mathbf{M}^{sir}_{\mathcal{F}\cup j}) - tr(\mathbf{M}^{sir}_{\mathcal{F}})(\geq 0)$.
- Recall that $\mathcal{A}$ denotes the active index set satisfying $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$, and $\mathcal{I} = \{1, \ldots, p\}$ denotes the full index set.
- **Proposition 1**: For any index set $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, we have $tr(\mathbf{M}^{sir}_{\mathcal{A}}) = tr(\mathbf{M}^{sir}_{\mathcal{F}}) = tr(\mathbf{M}^{sir}_{\mathcal{I}})$. Then, $tr(\mathbf{M}^{sir}_{\mathcal{F}\cup j}) - tr(\mathbf{M}^{sir}_{\mathcal{F}}) = 0$ given that $\mathcal{A} \subseteq \mathcal{F}$.
- It means that when $\mathcal{A} \subseteq \mathcal{F}$, any additional variable brings no more information on Y.

Trace pursuit: An example of the basic procedure

- Suppose $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$ is the basis of $S_{Y|\mathbf{X}}$ where q is the unknown structural dimension. Recall $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,p})^T$ for $i = 1, \ldots, q$, and let $\beta_{\min} = \min_{j\in\mathcal{A}} \sqrt{\sum_{i=1}^{q} \beta_{i,j}^2}$.
- **Proposition 2:** For any $\mathcal{F}$ such that $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$, we have

$$\max_{j\in\mathcal{F}^c\cap\mathcal{A}} \{tr(\mathbf{M}^{sir}_{\mathcal{F}\cup j}) - tr(\mathbf{M}^{sir}_{\mathcal{F}})\} \geq \lambda_q \lambda^{-1}_{\max}(\boldsymbol{\Sigma}) \lambda^2_{\min}(\boldsymbol{\Sigma}) \boldsymbol{\beta}^2_{\min} > 0,$$

where $\lambda_{\max}(\boldsymbol{\Sigma})$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ are the largest and the smallest eigenvalues of $\boldsymbol{\Sigma}$ respectively.

- It means that when $\mathcal{F}$ does not contain $\mathcal{A}$, there is at least one element in $\mathcal{F}^c$ contains information on Y.

Trace pursuit: An example of the basic procedure

- When the SIR matrix $\mathbf{M}_{\mathcal{F}}^{sir}$ is replaced by its empirical version $\hat{\mathbf{M}}_{\mathcal{F}}^{sir}$, we can define a test statistic and have the following property.
- **Proposition 3:** When $Y \perp\!\!\!\perp x_j | \mathbf{X}_{\mathcal{F}}$, $j \in \mathcal{F}^c$, the test statistic

$$T_{j|\mathcal{F}}^{sir} \longrightarrow \sum_{k=1}^{H} \omega_{j|\mathcal{F},k}^{sir} \chi_1^2, \text{ where } T_{j|\mathcal{F}}^{sir} = n \left\{ tr(\hat{\mathbf{M}}_{\mathcal{F}\cup j}^{sir}) - tr(\hat{\mathbf{M}}_{\mathcal{F}}^{sir}) \right\}.$$

  Here $\omega_{j|\mathcal{F},1}^{sir} \geq \ldots \geq \omega_{j|\mathcal{F},H}^{sir}$ are the eigenvalues of a matrix $\mathbf{\Omega}_{j|\mathcal{F}}^{sir}$ that is related to the eigenvectors of the SIR matrix (see Yu, Dong and Zhu (2014) for details.)
- *Similar results can be derived when SAVE or DR is used.*

Trace pursuit: An example of the basic procedure

- **Proposition 4:** Assume that there exist $0 < c < C$ and $0 < \xi_{\min} < 1/2$ such that

$$\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing} \max_{j\in\mathcal{F}^c\cap\mathcal{A}} \{tr(\mathbf{M}^{sir}_{\mathcal{F}\cup j}) - tr(\mathbf{M}^{sir}_{\mathcal{F}})\} > \varsigma n^{-\xi_{\min}}. \tag{4}$$

- If we set $0 < \bar{c}^{sir} < cn^{1-\xi_{\min}}$, then as $n \to \infty$,

$$Pr(\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing} \max_{j\in\mathcal{F}^c\cap\mathcal{A}} T^{\text{SIR}}_{j|\mathcal{F}} > \bar{c}^{sir}) \to 1.$$

- If we set $\underline{c}^{\text{SIR}} > Cn^{1-\xi_{\min}}$, then as $n \to \infty$,

$$Pr(\max_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}=\varnothing} \min_{j\in\mathcal{F}} T^{\text{SIR}}_{j|\{\mathcal{F}\setminus j\}} < \underline{c}^{\text{SIR}}) \to 1.$$

Trace pursuit: A comparison with COP

- (**The COP procedure(Zhong et al 2012):**) Given the structural dimension q, denote the largest q eigenvalues of $\mathbf{M}^{sir}_{\mathcal{F}\cup j}$ as $\lambda^{(k)}_{\mathcal{F}\cup j}$, and the largest q eigenvalues of $\mathbf{M}^{sir}_{\mathcal{F}}$ as $\lambda^{(k)}_{\mathcal{F}}$, $k = 1, \ldots, q$.
- COP is based on the key quantity $\sum_{k=1}^{q}(1-\lambda^{(k)}_{\mathcal{F}\cup j})^{-1}(\lambda^{(k)}_{\mathcal{F}\cup j} - \lambda^{(k)}_{\mathcal{F}})$.
- The COP test reduces to the trace test with our quantity $tr(\mathbf{M}^{sir}_{\mathcal{F}\cup j}) - tr(\mathbf{M}^{sir}_{\mathcal{F}})$ if we drop the scaling factor $(1-\lambda^{(k)}_{\mathcal{F}\cup j})^{-1}$ and assume both $\mathbf{M}^{sir}_{\mathcal{F}\cup j}$ and $\mathbf{M}^{sir}_{\mathcal{F}}$ have rank q.
- Compared with COP, the SIR-based trace pursuit does not involve estimating q, which is a challenge task when p is large.

The stepwise trace pursuit algorithm

(a) *Initialization*. Set the initial working set to be $\mathcal{F} = \varnothing$.

(b) *Forward addition*. Find index $a_{\mathcal{F}}$ such that

$$a_{\mathcal{F}} = \arg\max_{j \in \mathcal{F}^c} tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}\cup j}\big). \tag{5}$$

If $T^{\text{sir}}_{a_{\mathcal{F}}|\mathcal{F}} = n\{tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}\cup a_{\mathcal{F}}}\big) - tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}}\big)\} > \bar{c}^{sir}$, update $\mathcal{F}$ to be $\mathcal{F}\cup a_{\mathcal{F}}$.

(c) *Backward deletion*. Find index $d_{\mathcal{F}}$ such that

$$d_{\mathcal{F}} = \arg\max_{j \in \mathcal{F}} tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}\setminus j}\big). \tag{6}$$

If $T^{\text{sir}}_{d_{\mathcal{F}}|\{\mathcal{F}\setminus d_{\mathcal{F}}\}} = n\{tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}}\big) - tr\big(\hat{\mathbf{M}}^{sir}_{\mathcal{F}\setminus d_{\mathcal{F}}}\big)\} < \underline{c}^{\text{SIR}}$, update $\mathcal{F}$ to be $\mathcal{F}\setminus d_{\mathcal{F}}$.

(d) Repeat steps (b) and (c) until no predictors can be added or deleted.

The stepwise trace pursuit algorithm

- To determine only one index $a_{\mathcal{F}}$, STP needs to go over all possible candidate indices in $\mathcal{F}^c$ and compare a total of $p - |\mathcal{F}|$ test statistics: overwhelming computation burden when p is large!

The forward trace pursuit algorithm

An initial screening algorithm:

(a) *Initialization*. Set $\mathcal{S}^{(0)} = \varnothing$.

(b) *Forward addition*. For $k \geq 1$, $\mathcal{S}^{(k-1)}$ is given at the beginning of the kth iteration. For every $j \in \mathcal{I} \backslash \mathcal{S}^{(k-1)}$, compute $tr\big(\hat{\mathbf{M}}^{\text{SIR}}_{\mathcal{S}^{(k-1)} \cup j}\big)$, and find a_k such that

$$a_k = \underset{j \in \mathcal{I} \backslash \mathcal{S}^{(k-1)}}{\arg\max}\ tr\big(\hat{\mathbf{M}}^{\text{SIR}}_{\mathcal{S}^{(k-1)} \cup j}\big).$$

(c) *Solution path*. Repeat step (b) n times, to get a sequence of n nested candidate models. Denote the solution path as $\mathcal{S} = \{\mathcal{S}^{(k)} : 1 \leq k \leq n\}$, where $\mathcal{S}^{(k)} = \{a_1, \ldots, a_k\}$.

The forward trace pursuit algorithm

- Use the modified BIC criterion (idea from Chen and Chen 2008)

$$\text{BIC}(\mathcal{F}) = -\log\left\{tr(\hat{\mathbf{M}}_{\mathcal{F}}^{sir})\right\} + n^{-1}|\mathcal{F}|\left(\log n + 2\log p\right). \qquad (7)$$

  The candidate model $\mathcal{S}^{(\hat{m})}$ is selected with $\hat{m} = \arg\min_{1\le k\le n} \text{BIC}(\mathcal{S}^{(k)})$.

- **Proposition 5:** Under certain regularity conditions, as $n \to \infty$ and $p \to \infty$, $Pr\left(\mathcal{A} \subset \mathcal{S}^{(\hat{m})}\right) \to 1$.

The hybrid trace pursuit (HTP) algorithm

- Combines FTP as the initial screening step with STP as the refined selection step:

(a) Perform FTP to get solution path $\mathcal{S} = \{\mathcal{S}^{(k)} : 1 \leq k \leq n\}$.

(b) Based on BIC of (7), select $\mathcal{S}^{(\hat{m})}$ with $\hat{m} = \arg\min_{1 \leq k \leq n} \text{BIC}(\mathcal{S}^{(k)})$.

(c) Perform STP with C being the $1 - 0.1/p$ ($\alpha = 0.1/p$) quantile of the weighted chi-square distribution as the critical value in both forward addition and backward delection step. (Jiang and Liu 2013 used cross validation.), where the full index set $\mathcal{I} = \{1, \ldots, p\}$ is updated to the screened index set $\mathcal{S}^{(\hat{m})}$.

Outline

Simulations

- We consider the following models:

$$\text{I}: \quad Y = \text{sgn}(x_1 + x_p)\exp(x_2 + x_{p-1}) + \epsilon, (asymmetric)$$

$$\text{II}: \quad Y = 2x_1^2 x_p^2 - 2x_2^2 x_{p-1}^2 + \epsilon, (symmetric)$$

$$\text{III}: \quad Y = x_1^4 - x_p^4 + 3\exp(.8x_2 + .6x_{p-1}) + \epsilon.(mixture)$$

- $\mathbf{X} = (x_1, \ldots, x_p)^T$ is multivariate normal with $E(\mathbf{X}) = \mathbf{0}$ and $\text{var}(\mathbf{X}) = \mathbf{\Sigma}$, and $\epsilon \sim N(0, \sigma^2)$ is independent of $\mathbf{X}$.
- The structural dimensions for Models I to III are respectively $q = 2$, 4 and 3. The index set of significant predictors is $\mathcal{A} = \{1, 2, p-1, p\}$.

Simulations

- Set $\sigma = .2$, the sample size $n = 300$, the number of slices $H = 4$. Consider three settings of p: $p = 10$ for small dimensionality, $p = 100$ for moderate dimensionality, and $p = 1000$ for high dimensionality. Denote the (i, j)th entry of $\mathbf{\Sigma}$ as $\rho^{|i-j|}$, and $\rho = 0$ is with uncorrelated predictors and $\rho = .5$ with correlated predictors.
- Use Bentler and Xie (2000)'s approach to approximate the αth upper quantile of a weighted χ^2 distribution.

Simulations

- Based on the $N = 100$ repetitions, report the underfitted count (UF), the correctly fitted count (CF), the overfitted count (OF), and the average model size (MS):
- Let $\widehat{\mathcal{A}}_{(i)}$ be the estimated active set in the ith repetition and define

$$UF = \sum_{i=1}^{N} I(\mathcal{A} \not\subseteq \widehat{\mathcal{A}}_{(i)}), CF = \sum_{i=1}^{N} I(\mathcal{A} = \widehat{\mathcal{A}}_{(i)}),$$

$$OF = \sum_{i=1}^{N} I(\mathcal{A} \subset \widehat{\mathcal{A}}_{(i)}), \text{ and } MS = N^{-1} \sum_{i=1}^{N} |\widehat{\mathcal{A}}_{(i)}|.$$

Table 1

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
I	HTP-SIR	10	0	100	0	4.00	0	100	0	4.00
		100	0	100	0	4.00	0	100	0	4.00
		1000	0	100	0	4.00	0	100	0	4.00
	HTP-SAVE	10	9	59	32	4.31	4	39	57	4.00
		100	32	0	68	20.53	46	1	53	18.14
		1000	90	0	10	18.93	91	0	9	15.59
	HTP-DR	10	0	98	2	4.02	0	99	1	4.01
		100	0	95	5	4.07	0	93	7	4.08
		1000	0	96	4	4.04	0	94	6	4.08

Table 2

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
II	HTP-SIR	10	100	0	0	.31	100	0	0	.24
		100	100	0	0	.13	100	0	0	.08
		1000	100	0	0	.03	100	0	0	.03
	HTP-SAVE	10	2	97	1	3.99	2	94	4	4.02
		100	3	53	44	4.63	3	50	47	4.71
		1000	3	48	49	4.79	7	41	52	4.95
	HTP-DR	10	3	95	2	3.99	3	93	4	4.01
		100	2	56	42	4.70	3	46	51	4.77
		1000	7	44	49	4.76	7	45	48	4.91

Table 3

			$\rho = 0$				$\rho = .5$			
Model	Method	p	UF	CF	OF	MS	UF	CF	OF	MS
III	HTP-SIR	10	100	0	0	2.07	100	0	0	2.23
		100	100	0	0	2.58	100	0	0	3.56
		1000	100	0	0	6.34	100	0	0	6.56
	HTP-SAVE	10	4	33	63	5.14	7	45	48	4.61
		100	47	8	45	12.42	36	6	58	8.43
		1000	86	0	14	21.76	78	2	20	16.86
	HTP-DR	10	0	91	9	4.11	0	98	2	4.02
		100	3	83	14	4.13	4	79	17	4.16
		1000	4	88	8	4.06	5	61	34	4.39

Table 4

Table: Comparison between COP, SIRI and HTP-DR. Selection performances based on $p = 1000$ and $N = 100$ repetitions are reported.

		$\rho = 0$				$\rho = .5$			
Model	Method	UF	CF	OF	MS	UF	CF	OF	MS
I	COP	0	86	14	4.14	0	85	15	4.16
	SIRI	0	66	34	4.46	0	86	14	4.19
	HTP-DR	0	96	4	4.04	0	94	6	4.08
II	COP	100	0	0	4.00	100	0	0	4.00
	SIRI	52	38	10	3.79	36	45	19	4.05
	HTP-DR	7	44	49	4.76	7	45	48	4.91
III	COP	100	0	0	3.09	100	0	0	3.15
	SIRI	1	99	0	3.99	2	98	0	3.98
	HTP-DR	4	88	8	4.06	5	61	34	4.39

Table 5

Table: Comparison between SIS, DC-SIS and FTP algorithms for screening. Frequencies of cases including all active predictors are reported based on $p = 2000$ and $N = 100$ repetitions.

	$\rho = 0$			$\rho = .5$		
Method	Model I	Model II	Model III	Model I	Model II	Model III
DC-SIS	100	100	100	100	100	100
FTP-SIR	100	0	0	100	0	0
FTP-SAVE	12	97	7	10	98	31
FTP-DR	100	97	98	100	98	97

We also found that DC-SIS tends to choose much more variables.

THANK YOU!