Pseudo value method for ultra high-dimensional semiparametric models with life-time data

Tony Sit, Yue Xing, Yongze Xu and Minggao Gu

Department of Statistics, The Chinese University of Hong Kong

This set of supplementary notes include further details that are discussed in the manuscript. In particular, it covers the algorithm for Step 2 of our proposal for the transformation model, further discussion and illustration of the computation time needed as well as performance of SIS on Cox's proportional hazards model. Proofs for theorems presented in Section 2.3 will also be presented.

A1. Algorithm for Step 2 of PVM for the general transformation models.

Choose a positive integer λ, m, κ_2 and a sequence $\nu_k \downarrow 0$. We repeat (a) and (b) for κ_2 times:

(a) For a fixed k, set $U_0^{(k)} = U_m^{(k-1)}$. For $i = 1, \ldots, m$, generate $U_i^{(k)}$ from the transition probability $\Pi_{Y^{(k-1)}}\{U_{i-1}^{(k)}\}$.

(b) Update the estimate $\hat{Y}$ iteratively via

$$\mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \nu_k \Delta \mathbf{Y}^{(k)},$$

where

$$\begin{aligned}
\Gamma^{(k)} &= \Gamma^{(k-1)} + \nu_k[\bar{I}_0\{\mathbf{Y}^{(k-1)}, U^{(k)}\} + a_\lambda\{\mathbf{Y}^{(k-1)}\} - \Gamma^{(k-1)}] \\
\begin{bmatrix} \Delta \mathbf{Y}^{(k)} \\ \omega^{(k)+} \end{bmatrix} &= \begin{bmatrix} -\Gamma^{(k)} & (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})^\top \\ \mathbf{I} - \mathbf{H}_{\mathcal{Z}} & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} -\bar{H}\{\hat{\mathbf{Y}}^{(k-1)}, U^{(k)}\} + b_\lambda\{\mathbf{Y}^{(k-1)}\} \\ -(\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\hat{\mathbf{Y}}^{(k-1)} \end{bmatrix}
\end{aligned}$$

with

$$a_\lambda\{Y^{(k)}\} = T_{\mathbf{Z}}^\top \Sigma_{\mathbf{Z}}(T_{\mathbf{Z}}\mathbf{Y})T_{\mathbf{Z}}, \qquad b_\lambda\{Y^{(k)}\} = a_\lambda\{Y^{(k)}\}Y,$$

$$T_Z = (\mathbf{Z}\mathbf{Z}^\top)^+\mathbf{Z}, \qquad \Sigma_\lambda(\boldsymbol{\beta}) = \mathrm{diag}\{p'_\lambda(|\boldsymbol{\beta}_1|)/|\boldsymbol{\beta}_1|, \ldots, p'_\lambda(|\boldsymbol{\beta}_p|)/|\boldsymbol{\beta}_p|\}.$$

At the end of this stage, we can obtain $\hat{\mathbf{Y}}$ as the average of the last 10% of the sequence $\{\hat{\mathbf{Y}}^{(k)}\}_{k=1,\ldots,\kappa_2}$.

A2. Additional details on Computing time

In this subsection, we present more details regarding the computation burden of PVM, especially for models that requires MCMC-SA procedure needed in the maximum likelihood step. The following table considers only the computing times needed for various steps in order to perform variable selection under the PVM framework. Note that final estimation under low-dimensional setting is not included as the time needed is negligible.

To illustrate, we provide the following three examples:

1. AFT model with $n = 300, p = 100$: If 200 iterations are performed for stage 1, 800 times for stage 2, the whole algorithm will need about $0.02 + 200 \times 10^{-3} + 800 \times 0.009 + 0.11 = 7.53$ seconds.

2. Cox model with $n = 200, p = 1000$: If 500 iterations are performed for stage 1, 1,000 times for stage 2 with $m = 50$, the whole algorithm will need about $0.35+1500\times 50\times 200\times 5.1\times 10^{-5}+500\times 10^{-3}+1000\times 0.017+1.06 = 783.91$ seconds.

3. Probit model with $n = 400, p = 5000$: If 500 iterations are performed for stage 1, 1,500 times for stage 2 with $m = 100$, the whole algorithm will need about $54.68+2000\times 100\times 400\times 10^{-4}+500\times 10^{-3}+1500\times 0.42+10.4 = 8,695.6$ seconds which is about 145 minutes.

A3. Additional results on SIS for Cox's proportional hazards model

We summarise our results here for Fan et al. (2010) SIS on Cox's model. Readers may compare the results here with Panel (a) of Table 1 to see the edge that PVM offers.

In summary, Fan et al. (2010)'s SIS method tends to over-select variables; in some cases, about 30% of cases select more than 10 irrelevant variables. PVM, on the contrary, offers a more reasonable choice of the active set.

A4. Technical proofs for Theorems 1 and 2

Proof of Theorem 1 To begin, we first introduce the following lemmas:

Lemma 1. *Under [**C1**] to [**C6**], λ satisfies*

$$P(\|\tilde{\beta}-\beta\|_1 > \lambda) = o(1)$$

with $\lambda = O(\sqrt{\frac{\log p_n}{n}})$, then we have for some constant $C > 0$,

$$P\left(\max_{j=1,\dots,p_n} |\sum_{t=1}^{n} \varepsilon_t z_{tj}| \geq C(n\log p_n)^{1/2}\right) = o(1). \tag{S0.1}$$

And for any $C > 0$,

$$P\left(\max_{j=1,\dots,p_n} (\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_{tj})^2 \geq Cn^{-\gamma}\right) = o(1) \tag{S0.2}$$

and

$$P\left(\sum_{t=1}^{n} \varepsilon_t^2 \geq nC\right) = o(1) \tag{S0.3}$$

if n is large enough.

Proof. Through [**C3**], equation (1) holds since

$$P\left(\max_{j=1,\dots,p_n} |\sum_{t=1}^{n} \varepsilon_t z_{tj}| \geq C(n\log p_n)^{1/2}\right) \;\leq\; P\left(n\|\tilde{\beta}-\beta\|_1 C_{\max}^2 \geq C(n\log p_n)^{1/2}\right),$$

therefore such a $C > 0$ exists. For equation (S0.2), by **[C5]**, it is automatically satisfied. For equation (S0.3), using Markov inequality, we can get

$$P\left(\sum_{t=1}^{n} \varepsilon_t^2 \geq nC\right) \leq P(\|\tilde{\beta} - \beta\|_1^2 C_{\max}^2 \geq C) = o(1)$$

since $\lambda = O\left(\sqrt{\frac{\log p_n}{n}}\right) = o(1)$. $\square$

Remark 1. The original proof of Ing and Lai (2011) uses the independent property of the noise ϵ. Hence under some mild conditions, the results in Lemma 1 can be achieved. In our framework, we need to consider the error on the estimated parameters so as to bound Lemma 1 through bounding $\sum_{t=1}^{n} \varepsilon_t^2$ or $\max_{i,j} z_{ij}$, since no simple bound of $\sum_{t=1}^{n} \varepsilon_t$ is available.

Lemma 2 (Modified result of (3.8) in Ing and Lai (2011))**.** *Under **[C1]** to **[C4]**, there exists a positive constant s, independent of $1 \leq m \leq K_n$ and n, such that*

$$\lim_{n\to\infty} P(A_n^c(K_n)) = 0,$$

where

$$A_n(m) = \left\{ \max_{(J,i):\#(J)\leq m-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \leq s(\log p_n/n)^{1/2} \right\}$$

with

$$\mu_{J,i} = \sum_{j\notin J} \beta_j \mathbb{E}[(z_j - z_j^{(J)})z_i], \qquad \hat{\mu}_{J,i} = \frac{1}{n} \frac{\sum_{t=1}^{n}(y_t - \hat{y}_{t,J})x_{ti}}{(n^{-1}\sum_{t=1}^{n} x_{ti}^2)^{1/2}}.$$

Proof. It follows by the definition of $\hat{\mu}$ and μ that

$$\hat{\mu}_{J,i} - \mu_{J,i} = \frac{\sum_{t=1}^{n} \epsilon_t \hat{z}_{ti;J}^{\perp}}{\sqrt{n}(\sum_{t=1}^{n} z_{ti}^2)^{1/2}} + \sum_{j \notin J} \beta_j \left\{ \frac{n^{-1}\sum_{t=1}^{n} z_{tj}\hat{z}_{ti;J}^{\perp}}{(n^{-1}\sum_{t=1}^{n} z_{ti}^2)^{1/2}} - \mathbb{E}(z_j z_{i:j}^{\perp}) \right\} \quad \text{(S0.4)}$$

where $\hat{z}_{ti;J}^{\perp}$ represents the attribute after regression on the attributes in J. The parts which are independent to ϵ follows the proof of Ing and Lai (2011), and it suffices to show that for some $d > 0$,

$$P\left(\max_{\#(J) \le K_n - 1, i \notin J} |\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \hat{z}_{ti;J}^{\perp}| > d(\log p_n/n)^{1/2} \right) = o(1).$$

Follow the idea of Ing and Lai (2011), we split $\max_{\#(J) \le K_n - 1, i \notin J} |\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \hat{z}_{ti;J}^{\perp}|$ into three parts:

$$\begin{aligned}
\max_{\#(J) \le K_n - 1, i \notin J} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \hat{z}_{ti;J}^{\perp} \right| \quad &\le \quad \max_{1 \le i \le p_n} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right| \\
&\quad + \max_{\#(J) \le K_n - 1, i \notin J} \left| (\frac{1}{n}\sum_{t=1}^{n} z_{ti;J}^{\perp} z_t(J))^T \hat{\Gamma}^{-1}(J)(\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_t(J)) \right| \\
&\quad + \max_{\#(J) \le K_n - 1, i \notin J} \left| g_i^T(J)\Gamma^{-1}(J)(\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_t(J)) \right| \\
&:= \quad S_{1,n} + S_{2,n} + S_{3,n}.
\end{aligned}$$

Note that

$$S_{3,n} \le \max_{1 \le i \le p_n} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right| \max_{1 \le \#(J) \le K_n - 1, i \notin J} \|\Gamma^{-1}(J) g_i(J)\|_1 \le M \max_{1 \le i \le p_n} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right|$$

with probability converging to 1.

Using **[C6]** and the inequalities in Lemma 1, for some $d > 0$, if n is large enough, we have

$$
\begin{aligned}
& P(S_{1,n}+S_{3,n}>d\Big(\frac{\log p_n}{n})^{1/2}\Big) \\
\leq\quad & P\Big((M+1)\max_{1\leq i\leq p_n}\Big|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t z_{ti}\Big|>d(\frac{\log p_n}{n})^{1/2}\Big)=o(1).
\end{aligned}
$$

Then follow the idea of Ing and Lai (2011), we have $P(S_{2,n}>d'(\log p_n/n)^{1/2})=o(1)$ similarly. $\square$

Proof. Based on Lemma 2, we can get the result through following the proof of Ing and Lai (2011) directly. $\square$

Proof of Theorem 2

Proof. Follow the idea of Ing and Lai (2011), we will show that $P(\hat{k}<\tilde{k})=o(1)$.

Define

$$
\begin{aligned}
\hat{A}_n &= \frac{1}{n}\mathbf{Z}^T_{\hat{j}_{\tilde{k}}}(\mathbf{I}-\mathbf{H}_{\hat{J}_{\tilde{k}}})\mathbf{Z}_{\hat{j}_{\tilde{k}}}, \\
\hat{B}_n &= \frac{1}{n}\mathbf{Z}^T_{\hat{j}_{\tilde{k}}}(\mathbf{I}-\mathbf{H}_{\hat{J}_{\tilde{k}}})\varepsilon, \\
\hat{C}_n &= \frac{1}{n}\varepsilon^T(\mathbf{I}-\mathbf{H}_{\hat{J}_{\tilde{k}}})\varepsilon,
\end{aligned}
$$

and $v_n=\min_{1\leq\#(J)\leq K_n}\lambda_{min}(\Gamma(J))$. In the case that $\hat{k}<\tilde{k}_n$, we have

$$
\beta^2_{\hat{j}_{\tilde{k}}}\hat{A}_n+2\beta_{\hat{j}_{\tilde{k}}}\hat{B}_n+\hat{A}_n^{-1}\hat{B}_n^2\leq\frac{1}{n}w_n(\log p_n)(\tilde{k}-\hat{k})(\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_{t;\hat{J}_{\tilde{k}}})^2) \tag{S0.5}
$$

which implies that

$$
\beta^2_{\hat{j}_{\tilde{k}}}\hat{A}_n+2\beta_{\hat{j}_{\tilde{k}}}\hat{B}_n\leq\frac{1}{n}w_n\log p_n\lfloor an^{\gamma}\rfloor|\hat{C}_n|.
$$

In the next lemma, we show that for any $\theta > 0$,

$$P(\hat{A} \le v_n/2, \mathcal{D}_n) + P(|\hat{B}_n| \ge \theta n^{-\gamma/2}, \mathcal{D}_n) + P(w_n(\log p_n)|\hat{C}_n| \ge \theta n^{1-2\gamma}, \mathcal{D}_n) = o(1)$$

where $\mathcal{D}_n = \{N_n \subset \hat{J}_{\lfloor an^\gamma \rfloor}\} = \{\tilde{k} \le an^\gamma\}$ for sufficiently large a. Combining together with **[C5]** and Theorem 1, we have that the order of LHS of (S0.5) is larger than RHS while both are positive, hence $P(\hat{k} < \tilde{k}) = o(1)$. Follow the proof of Ing and Lai (2011), we finally get $P(N_n \subset \hat{N}_n) = 1$.

$\square$

Lemma 3 (Modifies result of (4.15) in Ing and Lai (2011))**.** *Under* ***[C1]*** *to* ***[C6]***,

$$P(\hat{A} \le v_n/2, \mathcal{D}_n) + P(|\hat{B}_n| \ge \theta n^{-\gamma/2}, \mathcal{D}_n) + P(w_n(\log p_n)|\hat{C}_n| \ge \theta n^{1-2\gamma}, \mathcal{D}_n) = o(1).$$

Proof. Following the proof of Ing and Lai (2011), we can directly get $P(\hat{A} \le v_n/2, \mathcal{D}_n) = o(1)$. Denote $m_0 = \lfloor an^\gamma \rfloor$. For $\hat{B}_n$, we have

$$P(|\hat{B}_n| \ge \theta n^{-\gamma/2}, \mathcal{D}_n) \quad \le \quad P\left(\max_{\#(J) \le m_0 - 1, i \notin J} \left| \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t \hat{z}_{t;J}^{\perp} \right| \ge \theta n^{-\gamma/2} \right) = o(1)$$

using a procedure similar with Lemma 2.

The proof of $\hat{C}_n$ is similar with $\hat{B}_n$:

$$\begin{aligned}
& P\left(w_n(\log p_n) \left| \hat{C}_n \right| \ge \theta n^{1-2\gamma}, \mathcal{D}_n \right) \\
\le \quad & P\left(\left| \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t^2 \right| \ge \theta/2 \right) \\
& + P\left\{ \max_{1 \le \#(J) \le m_0} \|\hat{\Gamma}^{-1}(J)\| m_0 \max_{1 \le j \le p_n} \left(\frac{1}{n} \sum_{t=1}^{n} \varepsilon_t z_{tj} \right)^2 \ge \theta/2 \right\},
\end{aligned}$$

where

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2\right| \geq \theta/2\right) = o(1)$$

and

$$P\left\{\max_{1\leq\#(J)\leq m_0}\|\hat{\Gamma}^{-1}(J)\|m_0 \max_{1\leq j\leq p_n}\left(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t z_{tj}\right)^2 \geq \theta/2 - o(1)\right\} = o(1).$$

Therefore for large enough n, $P(w_n(\log p_n)|\hat{C}_n| \geq \theta n^{1-2\gamma}, \mathcal{D}_n) = o(1)$. $\square$

Remark 2. The convergence rate for step one and two depends on the specific penalty term and model. On the other hand, in large sample case, most of them will converge. Therefore we introduce a new set of conditions:

(C1*) $p_n = o(n^{1/2}/\log n)$ and is small enough to satisfy $\sqrt{p_n/n}$-consistency.

(C2*) $\tilde{\beta}$ is $\sqrt{p_n/n}$-consistent.

(C5*) There exists $0 \leq \gamma < 1 - \log_n p_n$ such that $n^\gamma = o(p_n \wedge (n/\log p_n)^{1/2})$ and

$$\liminf_{n\to\infty} n^\gamma \min_{1\leq j\leq p_n;\beta_j\neq 0}\beta_j^2\sigma_j^2 > 0.$$

Condition (C5) is stronger than (C5*) since it takes effort in bounding the error terms. Under $\sqrt{p_n/n}$-consistency, we have faster bounds for error than high-dimensional case, which also enables us to consider some parameters which may converge to 0 with a slower rate than ε. Under these conditions, Theorem 2 still holds as long as (C1) to (C6) are satisfied. In Ing and Lai (2011), it also holds that $P(\hat{k} > \tilde{k}) = o(1)$.

References

Fan, J., Yang, F. & Wu, Y. (2010), 'High-dimensional variable selection for cox's proportional hazards model', *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown* **6**, 70–86.

Ing, C.-K. & Lai, T. (2011), 'A stepwise regression method and consistent model selection for high-dimensional sparse linear models', *Statisica Sinica* **21**, 1473–1513.

Table 1: Summary of the computation time needed for different steps need in the PVM procedure. The notation K, m and n denotes respectively, the number of iterations, the number of samplings carried out for estimating the expectation and the sample size, respectively; see also the algorithm presented on Page 15 in Section 2.1.

Stage	Step	Main time-consuming	Model	Time(s)
0	Initialization	Get Matrix $T_{\mathcal{Z}}$	All	CT_0
1 or 2	Generate U	MCMC estimation	Cox	$K \times m \times n \times 5.1 \times 10^{-5}$
1 or 2	Generate U	MCMC estimation	Probit	$K \times m \times n \times 1.0 \times 10^{-4}$
1 or 2	Generate U	MCMC estimation	PO	$K \times m \times n \times 4.8 \times 10^{-5}$
1	Update pseudo value		All	$K \times 10^{-3}$
2	Update pseudo value	Second derivative calculation	All	$K \times CT_1$
3	Variable selection	OGA algorithm	All	CT_2

where

CT_0	n	p	Time(s)	CT_1	n	p	Time(s)	CT_2	n	p	Time(s)
	200	100	< 0.01		200	100	0.007		200	100	0.1
	300	100	0.02		300	100	0.009		300	100	0.11
	400	100	0.03		400	100	0.010		400	100	0.11
	200	1000	0.35		200	1000	0.017		200	1000	1.06
	300	1000	0.36		300	1000	0.026		300	1000	1.08
	400	1000	0.37		400	1000	0.036		400	1000	1.11
	200	5000	53.8		200	5000	0.260		200	5000	10.2
	300	5000	54.2		300	5000	0.360		300	5000	10.2
	400	5000	54.68		400	5000	0.420		400	5000	10.4

Table 2: Performance of SIS on Cox proportional hazards model under ultra high-dimensional settings, *i.e.* $n \ll p$. Frequency, in 100 simulations, of including all relevant variables (Correct), of selecting exactly the relevant variable (E), of selecting all relevant variables and i irrelevant variables ($E+i$), and of selecting some relevant variables with i relevant ones omitted ($E-i$). The column "Correct" specifies the number of cases where all the relevant variables are selected.

n	150	200	400	200	200
p	1000	1000	1000	5000	10000
E	18	12	0	47	45
$E+1$	32	21	1	33	26
$E+2$	25	26	3	11	9
$E+3$	13	17	6	3	1
$E+4$	5	14	9	1	0
$E+5$	0	5	8	0	0
$E+6$	1	5	13	0	0
$E+7$	0	0	8	0	0
$E+8$	1	0	11	0	0
$E+9$	0	0	10	0	0
$E+10^{+}$	0	0	31	1	3
Correct	95	100	100	96	84
$E-1$	0	0	0	0	0
$E-2$	2	0	0	0	1
$E-3$	0	0	0	0	1
$E-4$	0	0	0	0	4
$E-5$	0	0	0	4	6
$E-3+1$	0	0	0	0	1
$E-3+2$	1	0	0	0	1
$E-3+6$	1	0	0	0	0
$E-4+1$	1	0	0	0	1
$E-5+2$	0	0	0	0	1