Supplemental Materials for "Performance Assessment of High-dimensional Variable Identification"

Yanjia Yu*, Yi Yang†, Yuhong Yang ‡

In this supplemental document, we provide technical proofs for the theorems in "Performance Assessment of High-dimensional Variable Identification" with additional remarks, and give further numerical results, including one on sensitivity of the complexity parameter ψ and one on the impact of the candidate models.

1. Proof of Theorem 1

Part I: F-measure

Proof. Denote by ∇ the symmetric difference between two sets. Estimated F-measure can be rewritten as

$$\widehat{F}(\mathcal{A}^0) = \sum_k w_k F(\mathcal{A}^0; \mathcal{A}^k), \qquad F(\mathcal{A}^0; \mathcal{A}^k) = \frac{|\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k|}{|\mathcal{A}^0| + |\mathcal{A}^k|}.$$

*Corresponding author, School of Statistics, University of Minnesota, USA. (E-mail: yuxxx748@umn.edu)
†Department of Mathematics and Statistics, McGill University, Canada (E-mail: yi.yang6@mcgill.ca)
‡School of Statistics, University of Minnesota, USA (E-mail: yangx374@umn.edu)

We have

$$
\begin{aligned}
&|\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| = \left|\sum_k w_k F(\mathcal{A}^0;\mathcal{A}^k) - F(\mathcal{A}^0)\right| \\
&= \left|\sum_k w_k (F(\mathcal{A}^0;\mathcal{A}^k) - F(\mathcal{A}^0))\right| \le \sum_k w_k |F(\mathcal{A}^0;\mathcal{A}^k) - F(\mathcal{A}^0)| \\
&= \sum_k w_k \left|1 - \frac{|\mathcal{A}^0 \nabla \mathcal{A}^k|}{|\mathcal{A}^0| + |\mathcal{A}^k|} - 1 + \frac{|\mathcal{A}^0 \nabla \mathcal{A}^*|}{|\mathcal{A}^0| + |\mathcal{A}^*|}\right| \\
&= \sum_k w_k \left|\frac{|\mathcal{A}^0| \cdot (|\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k|) + |\mathcal{A}^k| \cdot |\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^*| \cdot |\mathcal{A}^0 \nabla \mathcal{A}^k|}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)}\right| \\
&\le \underbrace{\sum_k w_k \frac{|\mathcal{A}^0| \cdot ||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)}}_{A} + \underbrace{\sum_k w_k \frac{|\mathcal{A}^k| \cdot ||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)}}_{B} \\
&+ \underbrace{\sum_k w_k \frac{||\mathcal{A}^k| - |\mathcal{A}^*|| \cdot |\mathcal{A}^0 \nabla \mathcal{A}^k|}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)}}_{C}.
\end{aligned}
$$

For ease of notation, we divide the right-most hand side of the above inequality into three parts and denote them by A, B, and C respectively. Note that since $\big||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k|\big| \le |\mathcal{A}^* \nabla \mathcal{A}^k|$, we have

$$
A \le \sum_k w_k \frac{|\mathcal{A}^0| \cdot |\mathcal{A}^* \nabla \mathcal{A}^k|}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)} \le \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.
$$

Similarly, it can be shown that

$$
B \le \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.
$$

Let us now prove a similar bound also holds for C. Specifically, we have

$$
\begin{aligned}
C &= \sum_k w_k \frac{||\mathcal{A}^k| - |\mathcal{A}^*|| \cdot |\mathcal{A}^0 \nabla \mathcal{A}^k|}{(|\mathcal{A}^0| + |\mathcal{A}^k|)(|\mathcal{A}^0| + |\mathcal{A}^*|)} \le \sum_k w_k \frac{\big||\mathcal{A}^k| - |\mathcal{A}^*|\big|}{|\mathcal{A}^0| + |\mathcal{A}^*|} \\
&= \sum_k w_k \frac{\big|(|\mathcal{A}^k \backslash \mathcal{A}^*| + |\mathcal{A}^k \cap \mathcal{A}^*|) - (|\mathcal{A}^* \backslash \mathcal{A}^k| + |\mathcal{A}^k \cap \mathcal{A}^*|)\big|}{|\mathcal{A}^0| + |\mathcal{A}^*|} \\
&= \sum_k w_k \frac{\big||\mathcal{A}^k \backslash \mathcal{A}^*| - |\mathcal{A}^* \backslash \mathcal{A}^k|\big|}{|\mathcal{A}^0| + |\mathcal{A}^*|} \le \sum_k w_k \frac{|\mathcal{A}^k \backslash \mathcal{A}^*| + |\mathcal{A}^* \backslash \mathcal{A}^k|}{|\mathcal{A}^0| + |\mathcal{A}^*|} \\
&= \sum_k w_k \frac{|\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^0| + |\mathcal{A}^*|} \le \sum_k w_k \frac{|\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|}.
\end{aligned}
$$

It follows that for any $\mathcal{A}^0$ in $\mathbb{C}$

$$|\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \leq A + B + C \leq 3 \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.$$

Therefore,

$$\sup_{\mathcal{A}^0 \in \mathbb{C}} |\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \leq 3 \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.$$

Now under the assumption that the model weighting w is weakly consistent,

$$\sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|} \xrightarrow{p} 0.$$

We have proved $\sup_{\mathcal{A}^0 \in \mathbb{C}} |\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \xrightarrow{p} 0$. $\square$

Part II: G-measure

Proof. For a given $\mathcal{A}^0$ in $\mathbb{C}$, the estimated G-measure can be rewritten as

$$\widehat{G}(\mathcal{A}^0) = \sum_k w_k G(\mathcal{A}^0; \mathcal{A}^k), \qquad G(\mathcal{A}^0; \mathcal{A}^k) = \frac{|\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k|}{2\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}}.$$

Suppose $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)|$ does not converge to 0 in probability uniformly over $\mathbb{C}$, then there exist some subsequence $n_1, n_2, \cdots$, $\epsilon_1 > 0, \delta > 0$, $\mathcal{A}^0_{n_j} \in \mathbb{C}$, and sets $\mathcal{S}_{n_j}$, s.t. $P(\mathcal{S}_{n_j}) \geq \delta$ and $|\widehat{G}(\mathcal{A}^0_{n_j}) - G(\mathcal{A}^0_{n_j})| > \epsilon_1$ on $\mathcal{S}_{n_j}$. For ease of notation, we denote $\mathcal{A}^0_{n_j}$ as $\mathcal{A}^0$ in the following proof.

With the above, we first prove that we must have $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ as $n_j \to \infty$. If not, then there exist $\epsilon_2 > 0$, a subsequence n_{j_l} and sets $\mathcal{N}_{n_{j_l}}$ such that on $\mathcal{N}_{n_{j_l}}$ we have $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$. Then we can actually prove $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{N}_{n_{j_l}}$ as follows.

By definition of $\widehat{G}$ and G, and $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$ on $\mathcal{N}_{n_{j_l}}$, we have

$$
\begin{aligned}
|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| &= |\sum_k w_k G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)| \le \sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)| \\
&= \sum_k w_k \left| \frac{|\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k|}{2\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}} - \frac{|\mathcal{A}^0| + |\mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^*|}{2\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^*|}} \right| \\
&\le \sum_k w_k \frac{|\sqrt{|\mathcal{A}^*|} - \sqrt{|\mathcal{A}^k|}| \cdot ||\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0| \cdot |\mathcal{A}^k|}} \\
&\quad + \sum_k w_k \frac{\sqrt{|\mathcal{A}^k|} \cdot ||\mathcal{A}^k| - |\mathcal{A}^*| + |\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0| \cdot |\mathcal{A}^k|}} \\
&\le \underbrace{\sum_k w_k \frac{|\sqrt{|\mathcal{A}^*|} - \sqrt{|\mathcal{A}^k|}| \cdot ||\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0| \cdot |\mathcal{A}^k|}}}_{A} \\
&\quad + \underbrace{\sum_k w_k \frac{||\mathcal{A}^k| - |\mathcal{A}^*||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0|}}}_{B} + \underbrace{\sum_k w_k \frac{||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0|}}}_{C}.
\end{aligned}
$$

For notational convenience, we divide the right-most-hand side of the above inequality into three parts and denote them by A, B, and C respectively. For part A, because $|\mathcal{A}^0| + |\mathcal{A}^k| - |\mathcal{A}^0 \nabla \mathcal{A}^k| = 2|\mathcal{A}^0 \cap \mathcal{A}^k|$ and $\big||\mathcal{A}^*| - |\mathcal{A}^k|\big| \le |\mathcal{A}^* \nabla \mathcal{A}^k|$, together with $|\mathcal{A}^0 \cap \mathcal{A}^k| \le \sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}$, we have

$$
A = \sum_k w_k \frac{\big||\mathcal{A}^*| - |\mathcal{A}^k|\big| \cdot |\mathcal{A}^0 \cap \mathcal{A}^k|}{\left(\sqrt{|\mathcal{A}^*|} + \sqrt{|\mathcal{A}^k|}\right) \sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0| \cdot |\mathcal{A}^k|}} \le \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.
$$

For part B, since $||\mathcal{A}^k| - |\mathcal{A}^*|| \le |\mathcal{A}^k \nabla \mathcal{A}^*|$ and $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$ on $\mathcal{N}_{n_{j_l}}$, we have

$$
B = \sum_k w_k \frac{\big||\mathcal{A}^k| - |\mathcal{A}^*|\big|}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0|}} \le \frac{1}{2\sqrt{\epsilon_2}} \sum_k w_k \frac{|\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|}.
$$

For part C, it follows from the facts that $||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k|| \le |\mathcal{A}^* \nabla \mathcal{A}^k|$ and that $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$ on $\mathcal{N}_{n_{j_l}}$, we have

$$
C = \sum_k w_k \frac{||\mathcal{A}^0 \nabla \mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0|}} \le \frac{1}{2\sqrt{\epsilon_2}} \frac{\sum_k w_k |\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.
$$

Consequently, we have that on $\mathcal{N}_{n_{j_l}}$,

$$|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \leq A + B + C \leq (1 + \frac{1}{\sqrt{\epsilon_2}}) \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}.$$

Under the assumption that the model weighting w is weakly consistent,

$$\sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|} \xrightarrow{p} 0,$$

we must have $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{N}_{n_{j_l}}$. This contradicts with the statement that $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| > \epsilon_1 > 0$ on $\mathcal{S}_{n_j}$. Therefore, we have proved that $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ under the beginning supposition.

Next, we prove actually we must have $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ as $n_j \to \infty$. Because $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$, we can set $\delta_n = \sqrt{\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|}}$, then $\delta_n \xrightarrow{p} 0$ and $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*| \cdot \delta_n} = \delta_n \xrightarrow{p} 0$. Then

$$|G(\mathcal{A}^0)| = \frac{||\mathcal{A}^0| + |\mathcal{A}^*| - |\mathcal{A}^0 \nabla \mathcal{A}^*||}{2\sqrt{|\mathcal{A}^*| \cdot |\mathcal{A}^0|}} = \frac{|\mathcal{A}^0 \cap \mathcal{A}^*|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^*|}} \leq \sqrt{\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|}} \xrightarrow{p} 0,$$

that is, $G(\mathcal{A}^0) \xrightarrow{p} 0$. Now we prove that we also have $\widehat{G}(\mathcal{A}^0) \xrightarrow{p} 0$ as follows. Observe on $\mathcal{S}_{n_j}$

$$\begin{aligned}
\widehat{G}(\mathcal{A}^0) &= \sum_k I(|\mathcal{A}^k| \leq |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}} + \sum_k I(|\mathcal{A}^k| > |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}} \\
&\leq \sum_k I(|\mathcal{A}^k| \leq |\mathcal{A}^*| \cdot \delta_n) \cdot w_k + \sum_k I(|\mathcal{A}^k| > |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}}.
\end{aligned}$$

Then because $\sum_k w_k \frac{|\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ and

$$\begin{aligned}
&\sum_k w_k \frac{|\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|} \geq \sum_k w_k \frac{||\mathcal{A}^*| - |\mathcal{A}^k||}{|\mathcal{A}^*|} \\
&\geq \sum_k w_k \frac{||\mathcal{A}^*| - |\mathcal{A}^k||}{|\mathcal{A}^*|} \cdot I(|\mathcal{A}^k| \leq |\mathcal{A}^*| \cdot \delta_n) \\
&\geq \frac{1}{2} \sum_k w_k \cdot I(|\mathcal{A}^k| \leq |\mathcal{A}^*| \cdot \delta_n),
\end{aligned}$$

we know $\sum_k I(|\mathcal{A}^k| \le |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \xrightarrow{p} 0$. On $\mathcal{S}_{n_j}$, we also have

$$\begin{aligned}
&\sum_k I(|\mathcal{A}^k| > |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}} \\
&\le \sum_k I(|\mathcal{A}^k| > |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \sqrt{\frac{|\mathcal{A}^0|}{|\mathcal{A}^k|}} \\
&\le \sum_k I(|\mathcal{A}^k| > |\mathcal{A}^*| \cdot \delta_n) \cdot w_k \sqrt{\frac{|\mathcal{A}^0|}{|\mathcal{A}^*| \cdot \delta_n}} \\
&\xrightarrow{p} 0,
\end{aligned}$$

since $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*| \cdot \delta_n} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$. Therefore, we have shown $\widehat{G}(\mathcal{A}^0) \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$.

Now since we have proved that on $\mathcal{S}_{n_j}$, $G(\mathcal{A}^0) \xrightarrow{p} 0$ and $\widehat{G}(\mathcal{A}^0) \xrightarrow{p} 0$, so $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$, which contradicts with the beginning supposition that $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| > \epsilon_1 > 0$ on $\mathcal{S}_{n_j}$. Therefore the supposition does not hold, and we have proved the $|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)|$ does converge to 0 in probability uniformly over $\mathbb{C}$. □

2. Proof of Theorem 2

Part I: standard deviation of F-measure

Proof. For any $\mathcal{A}^0$ in $\mathbb{C}$, by definition of the standard deviation of F-measure, we have

$$\begin{aligned}
\mathrm{sd}\big(\widehat{F}(\mathcal{A}^0)\big) &\equiv \sqrt{\sum_k w_k \big(F(\mathcal{A}^0; \mathcal{A}^k) - \widehat{F}(\mathcal{A}^0)\big)^2} \\
&\le \sqrt{\sum_k w_k |F(\mathcal{A}^0; \mathcal{A}^k) - \widehat{F}(\mathcal{A}^0)|} \\
&\le \sqrt{\sum_k w_k |F(\mathcal{A}^0; \mathcal{A}^k) - F(\mathcal{A}^0)| + |F(\mathcal{A}^0) - \widehat{F}(\mathcal{A}^0)|}.
\end{aligned}$$

Using the facts proved in the proof for Theorem 1,

$$|\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \le \sum_k w_k |F(\mathcal{A}^0; \mathcal{A}^k) - F(\mathcal{A}^0)| \le 3 \sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|},$$

we know

$$\mathrm{sd}(\widehat{F}(\mathcal{A}^0)) \leq \sqrt{6\sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}},$$

and

$$\sup_{\mathcal{A}^0 \in \mathbb{C}} \mathrm{sd}(\widehat{F}(\mathcal{A}^0)) \leq \sqrt{6\sum_k w_k \frac{|\mathcal{A}^* \nabla \mathcal{A}^k|}{|\mathcal{A}^*|}} \xrightarrow{p} 0$$

under the assumption that the model weighting w is weakly consistent. □

Part II: standard deviation of G-measure

Proof. For any $\mathcal{A}^0$ in $\mathbb{C}$, by definition of the standard deviation of G-measure, we have

$$\begin{aligned}
\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big) &\equiv \sqrt{\sum_k w_k \big(G(\mathcal{A}^0; \mathcal{A}^k) - \widehat{G}(\mathcal{A}^0)\big)^2} \\
&\leq \sqrt{\sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - \widehat{G}(\mathcal{A}^0)|} \\
&\leq \sqrt{\sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)| + |G(\mathcal{A}^0) - \widehat{G}(\mathcal{A}^0)|}.
\end{aligned}$$

Using the facts in Theorem 1, we have

$$|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0.$$

So it suffices to show $\sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)| \xrightarrow{p} 0$. The arguments are similar to those in the proof of Theorem 1. For completeness, the full proof is given below.

Suppose $\sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)|$ does not converge to 0 in probability uniformly over $\mathbb{C}$, then there exist some subsequence $n_1, n_2, \cdots, \epsilon_1 > 0, \delta > 0, \mathcal{A}^0_{n_j} \in \mathbb{C}$, and sets $\mathcal{S}_{n_j}$, s.t. $P(\mathcal{S}_{n_j}) \geq \delta$ and $\sum_k w_k |G(\mathcal{A}^0_{n_j}; \mathcal{A}^k) - G(\mathcal{A}^0_{n_j})| > \epsilon_1$ on $\mathcal{S}_{n_j}$. For ease of notation, we denote $\mathcal{A}^0_{n_j}$ as $\mathcal{A}^0$. We first prove that we must have $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ as $n_j \to \infty$. If not, then there exist $\epsilon_2 > 0$, a subsequence n_{j_l} and sets $\mathcal{N}_{n_{j_l}}$ such that on $\mathcal{N}_{n_{j_l}}$ we have $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$. Then we can actually prove $\sum_k w_k |G(\mathcal{A}^0; \mathcal{A}^k) - G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{N}_{n_{j_l}}$ as follows. On $\mathcal{N}_{n_{j_l}}$, since $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} > \epsilon_2 > 0$, we have that

$$
\begin{aligned}
&\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)| \\
&\le \underbrace{\sum_k w_k \frac{|\sqrt{|\mathcal{A}^*|}-\sqrt{|\mathcal{A}^k|}|\cdot||\mathcal{A}^0|+|\mathcal{A}^k|-|\mathcal{A}^0\nabla\mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*|\cdot|\mathcal{A}^0|\cdot|\mathcal{A}^k|}}}_{A} \\
&\quad + \underbrace{\sum_k w_k \frac{||\mathcal{A}^k|-|\mathcal{A}^*||}{2\sqrt{|\mathcal{A}^*|\cdot|\mathcal{A}^0|}}}_{B} + \underbrace{\sum_k w_k \frac{||\mathcal{A}^0\nabla\mathcal{A}^*|-|\mathcal{A}^0\nabla\mathcal{A}^k||}{2\sqrt{|\mathcal{A}^*|\cdot|\mathcal{A}^0|}}}_{C} \\
&\le (1+\frac{1}{\sqrt{\epsilon_2}})\sum_k w_k\frac{|\mathcal{A}^*\nabla\mathcal{A}^k|}{|\mathcal{A}^*|}.
\end{aligned}
$$

Under the assumption that the model weighting w is weakly consistent,

$$\sum_k w_k\frac{|\mathcal{A}^*\nabla\mathcal{A}^k|}{|\mathcal{A}^*|} \xrightarrow{p} 0,$$

we must have $\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{N}_{n_{j_l}}$. This contradicts with the statement that $\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)| > \epsilon_1 > 0$ on $\mathcal{S}_{n_j}$. Therefore, we have proved that $\frac{|\mathcal{A}^0|}{|\mathcal{A}^*|} \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ under the beginning supposition.

Next, we prove actually we must have $\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)| \xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$ as $n_j\to\infty$. Similar to the proof in Theorem 1, we can prove that $G(\mathcal{A}^0)\xrightarrow{p} 0$ and $\widehat{G}(\mathcal{A}^0)\xrightarrow{p} 0$ on $\mathcal{S}_{n_j}$. We then have

$$\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)| \le \sum_k w_k G(\mathcal{A}^0;\mathcal{A}^k)+G(\mathcal{A}^0) = \widehat{G}(\mathcal{A}^0)+G(\mathcal{A}^0) \xrightarrow{p} 0$$

on $\mathcal{S}_{n_j}$, which contradicts with the beginning supposition that $\sum_k w_k|G(\mathcal{A}^0_{n_j};\mathcal{A}^k)-G(\mathcal{A}^0_{n_j})| > \epsilon_1 > 0$ on $\mathcal{S}_{n_j}$. Therefore the supposition does not hold, and we have proved the $\sum_k w_k|G(\mathcal{A}^0_{n_j};\mathcal{A}^k) - G(\mathcal{A}^0_{n_j})|$ does converge to 0 in probability uniformly over $\mathbb{C}$. Since we have $\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big) \le \sqrt{\sum_k w_k|G(\mathcal{A}^0;\mathcal{A}^k)-G(\mathcal{A}^0)|+|G(\mathcal{A}^0)-\widehat{G}(\mathcal{A}^0)|} \xrightarrow{p} 0$ for any $\mathcal{A}^0\in\mathbb{C}$, we have proved

$$\sup_{\mathcal{A}^0\in\mathbb{C}} |\mathrm{sd}\big(\widehat{G}(\mathcal{A}^0)\big)| \xrightarrow{p} 0 \qquad \text{as } n\to\infty.$$

$\square$

3. Proof of Theorem 3

Proof. When a model screening is used to obtain the reduced candidate model list $\mathbb{S}$, the weights of the models in $\mathbb{S}$ are renormalized as $\widetilde{w}_k = w_k/w_{\mathbb{S}}$, where $w_{\mathbb{S}} = \sum_{k\in\mathbb{S}} w_k$. We next show that this renormalized weighting, though random, is still weakly consistent (in spite of possibly missing the true model in $\mathbb{S}$). Indeed,

$$\sum_{k\in\mathbb{S}} \widetilde{w}_k \frac{|\mathcal{A}^*\nabla\mathcal{A}^k|}{|\mathcal{A}^*|} = \left(\sum_{k\in\mathbb{S}} w_k \frac{|\mathcal{A}^*\nabla\mathcal{A}^k|}{|\mathcal{A}^*|}\right)/w_{\mathbb{S}} \le \left(\sum_{k\in\mathbb{C}} w_k \frac{|\mathcal{A}^*\nabla\mathcal{A}^k|}{|\mathcal{A}^*|}\right)/w_{\mathbb{S}},$$

which clearly converges to zero in probability under the weak consistency of w and the weak inclusion property of $\mathbb{S}$. Then the arguments for the convergence of $\widehat{F}$ and $\widehat{G}$ in the proofs of Theorems 1 and 2 continue to work. Thus we know that the conclusions of Theorems 1 and 2 still hold. $\square$

4. Remarks on Theorem 3

Theorem 3 relies on a good quality of the set of candidate models obtained from a model screening step. The weak inclusion property demands $\mathbb{S}$ to contain some (good) models with non-vanishing cumulated weight, but does not require $\mathcal{A}^*$ to be in $\mathbb{S}$ with high-probability. If the true model is really strong, it is not very likely to be missed by $\mathbb{S}$. In contrast, if there are very weak true coefficients, the true model may not be included in $\mathbb{S}$. Fortunately, in this case, as long as the number of small effects is asymptotically negligible compared to the true model size, some models close to $\mathcal{A}^*$ are most likely to be included in $\mathbb{S}$, and the weak inclusion property may hold. For example, suppose the true model size is of order $\log n$ and there are no more than $(\log n)^{1/2}$ small coefficients. Then the models without some of the small-effect variables are likely to receive comparable or even higher weights than the true model. Then, even if the true model is missed in $\mathbb{S}$, the weak inclusion property holds.

In particular, if $\mathbb{S}$ is obtained as the solution path of a penalized method and has the weak inclusion property, the method is said to be *weakly path-inclusive* or *weakly path-consistent*. Note that for a consistent weighting, our definition here on $\mathbb{S}$ is weaker than the path-consistency that requires the true model to be included on the solution path with probability going to 1.

In the high-dimensional case, we can set $\mathbb{S}$ as a large collection of the models obtained from the solution paths of multiple penalized methods, such as (adaptive) Lasso, SCAD and MCP. Specifically, we can obtain the models $\mathbb{S}_{\text{Lasso}}$, $\mathbb{S}_{\text{SCAD}}$, $\mathbb{S}_{\text{MCP}}$ for (adaptive) Lasso, SCAD and MCP respectively on the solution paths $\{\widehat{\boldsymbol{\beta}}^{\lambda_1}, \ldots, \widehat{\boldsymbol{\beta}}^{\lambda_L}\}$ for decreasing sequences of tuning parameters $\{\lambda_1, \ldots, \lambda_L\}$. These models are then combined together as a union of candidate models $\mathbb{S} = \{\mathbb{S}_{\text{Lasso}}, \mathbb{S}_{\text{SCAD}}, \mathbb{S}_{\text{MCP}}\}$. These penalized methods are good choices, since according to existing theories (Tibshirani, 1996; Zou, 2006; Fan and Li, 2001; Zhang, 2010), $\mathbb{S}$ produced by the solution paths of these methods ensure path-consistency under certain regularity conditions. In fact, in order to get Theorem 3, only one of $\mathbb{S}_{\text{Lasso}}$, $\mathbb{S}_{\text{SCAD}}$ and $\mathbb{S}_{\text{MCP}}$ needs to be weakly path-consistent. Of course, users are not limited to these options, they can add models obtained from any other weakly path-consistent variable selection methods into $\mathbb{S}$ to further enhance the chance of capturing the true/best model. More details about candidate models are discussed in Section 4.1 of the main paper.

5. Additional Simulation Results

Table A1: *Classification case (Example 2).*

	F	G	d_F	d_G
		Lasso		
True	0.631 (0.008)	0.680 (0.006)		
ARM	0.697 (0.007)	0.734 (0.006)	0.066 (0.002)	0.054 (0.002)
BIC-p	0.639 (0.008)	0.686 (0.006)	0.008 (0.001)	0.006 (0.001)
		AdLasso		
True	0.989 (0.004)	0.989 (0.004)		
ARM	0.929 (0.002)	0.935 (0.002)	0.067 (0.002)	0.062 (0.002)
BIC-p	0.987 (0.003)	0.988 (0.002)	0.009 (0.001)	0.008 (0.001)
		MCP		
True	0.964 (0.008)	0.967 (0.008)		
ARM	0.922 (0.004)	0.929 (0.004)	0.065 (0.002)	0.059 (0.002)
BIC-p	0.965 (0.008)	0.968 (0.007)	0.009 (0.001)	0.008 (0.001)
		SCAD		
True	0.955 (0.010)	0.960 (0.009)		
ARM	0.919 (0.005)	0.926 (0.004)	0.065 (0.002)	0.059 (0.002)
BIC-p	0.956 (0.009)	0.961 (0.008)	0.009 (0.001)	0.008 (0.001)

Table A2: *Classification case (Example 3).*

	F	G	d_F	d_G
		Lasso		
True	0.154 (0.011)	0.278 (0.010)		
ARM	0.129 (0.009)	0.251 (0.009)	0.025 (0.002)	0.028 (0.002)
BIC-p	0.159 (0.011)	0.283 (0.010)	0.010 (0.002)	0.010 (0.002)
		AdLasso		
True	0.712 (0.021)	0.751 (0.018)		
ARM	0.627 (0.020)	0.682 (0.016)	0.091 (0.006)	0.076 (0.005)
BIC-p	0.716 (0.021)	0.754 (0.017)	0.030 (0.006)	0.026 (0.005)
		MCP		
True	0.498 (0.015)	0.576 (0.012)		
ARM	0.433 (0.015)	0.523 (0.012)	0.067 (0.004)	0.056 (0.003)
BIC-p	0.511 (0.015)	0.586 (0.012)	0.026 (0.005)	0.020 (0.004)
		SCAD		
True	0.214 (0.006)	0.344 (0.005)		
ARM	0.183 (0.006)	0.312 (0.006)	0.032 (0.002)	0.033 (0.002)
BIC-p	0.225 (0.007)	0.352 (0.006)	0.017 (0.004)	0.014 (0.003)

Table A3: *Classification case (Example 4).*

	F	G	d_F	d_G
		Lasso		
True	0.720 (0.005)	0.734 (0.005)		
ARM	0.493 (0.006)	0.572 (0.004)	0.227 (0.007)	0.163 (0.006)
BIC-p	0.616 (0.006)	0.667 (0.004)	0.109 (0.005)	0.077 (0.005)
		AdLasso		
True	0.794 (0.005)	0.800 (0.005)		
ARM	0.722 (0.006)	0.755 (0.005)	0.081 (0.006)	0.059 (0.005)
BIC-p	0.876 (0.006)	0.883 (0.005)	0.096 (0.006)	0.094 (0.006)
		MCP		
True	0.751 (0.005)	0.770 (0.005)		
ARM	0.793 (0.004)	0.813 (0.004)	0.063 (0.005)	0.056 (0.004)
BIC-p	0.932 (0.005)	0.934 (0.005)	0.182 (0.006)	0.164 (0.005)
		SCAD		
True	0.778 (0.006)	0.789 (0.006)		
ARM	0.755 (0.005)	0.781 (0.004)	0.064 (0.006)	0.055 (0.005)
BIC-p	0.913 (0.006)	0.916 (0.005)	0.141 (0.007)	0.132 (0.006)

Table A4: *Classification case (Example 5).*

	F	G	d_F	d_G
		Lasso		
True	0.386 (0.006)	0.440 (0.005)		
ARM	0.223 (0.004)	0.348 (0.004)	0.163 (0.006)	0.093 (0.005)
BIC-p	0.359 (0.006)	0.465 (0.005)	0.039 (0.004)	0.043 (0.003)
		AdLasso		
True	0.726 (0.005)	0.735 (0.005)		
ARM	0.616 (0.008)	0.669 (0.006)	0.118 (0.007)	0.079 (0.005)
BIC-p	0.859 (0.008)	0.865 (0.008)	0.137 (0.007)	0.133 (0.006)
		MCP		
True	0.683 (0.008)	0.695 (0.008)		
ARM	0.639 (0.009)	0.687 (0.007)	0.079 (0.006)	0.063 (0.005)
BIC-p	0.868 (0.008)	0.871 (0.008)	0.186 (0.006)	0.177 (0.006)
		SCAD		
True	0.634 (0.008)	0.637 (0.008)		
ARM	0.506 (0.010)	0.580 (0.008)	0.131 (0.007)	0.072 (0.005)
BIC-p	0.743 (0.009)	0.766 (0.008)	0.110 (0.006)	0.130 (0.006)

6. Sensitivity Analysis of ψ

In this simulation, we study how the choices of the prior weight parameter ψ impact the estimation performance of PAVI. We only present results for the regression case, since we found that the classification case gives similar results. We adopt the simulation setting of Example 3 defined in Section 5.1, except that we let $\sigma^2 = 1$, $n = 100$ and we vary $p = \{200, 2000\}$. We compare $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ with the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ under nine different values of ψ, that is, $\psi \in \{0, 0.5, 1, 2, 4, 6, 8, 10, 20\}$.

All simulation cases are repeated for 100 times and the corresponding values are computed and averaged. The results are shown in Figure A5 for $p = 200$ case and A6 for $p = 2000$ case. We can see that by using either the ARM or BIC-p weighting with $\psi = 1$ or 2, the estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ can better reflect the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for all four different variable selection methods under evaluation. We observed similar results in other simulation settings. We conclude that overall, under $\psi = 1$ or 2 setting, PAVI is stably reliable in our simulation, while either a too large or too small value of ψ leads to poor estimation performance.

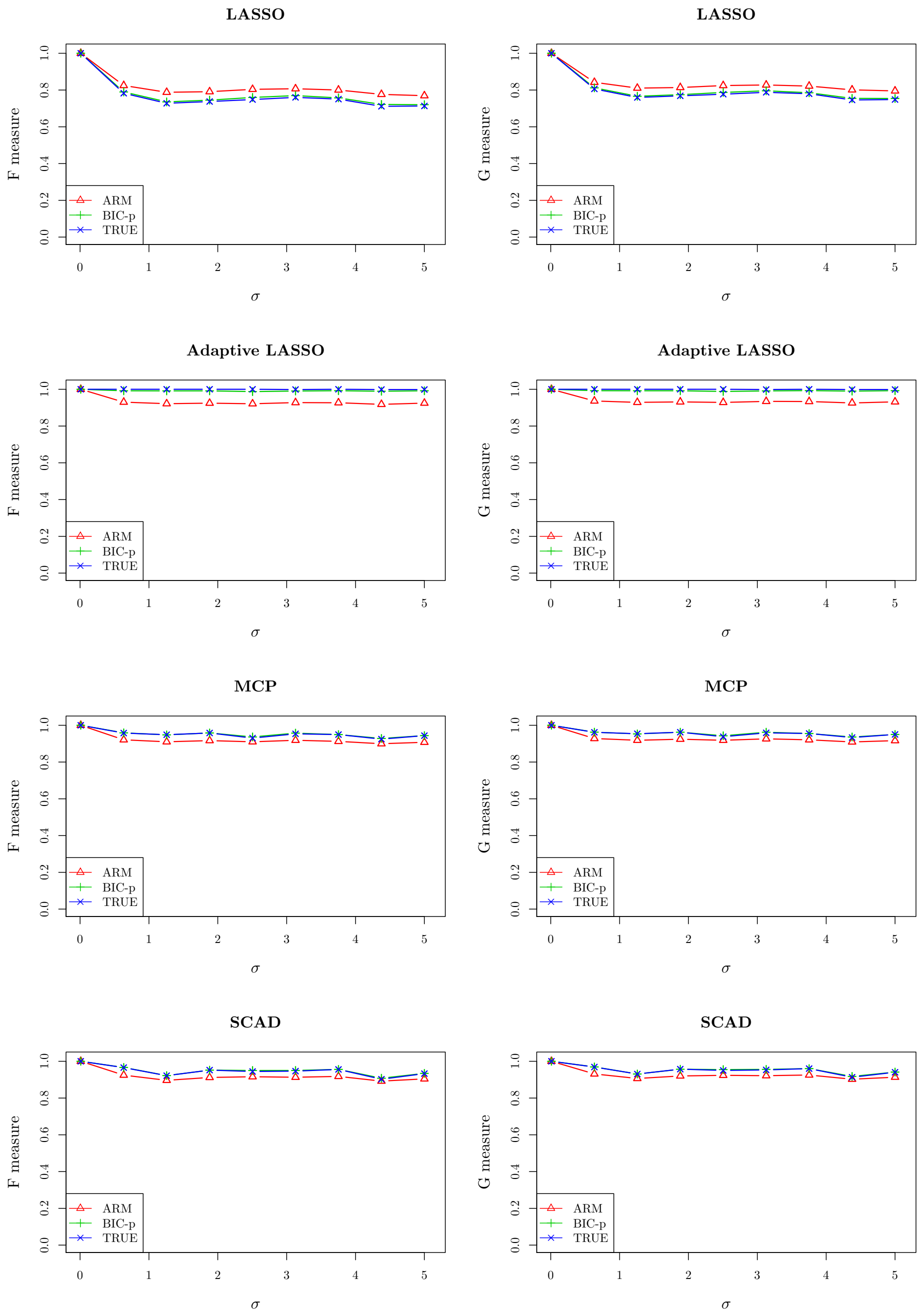

Figure A1: *Regression case (Example 2).*

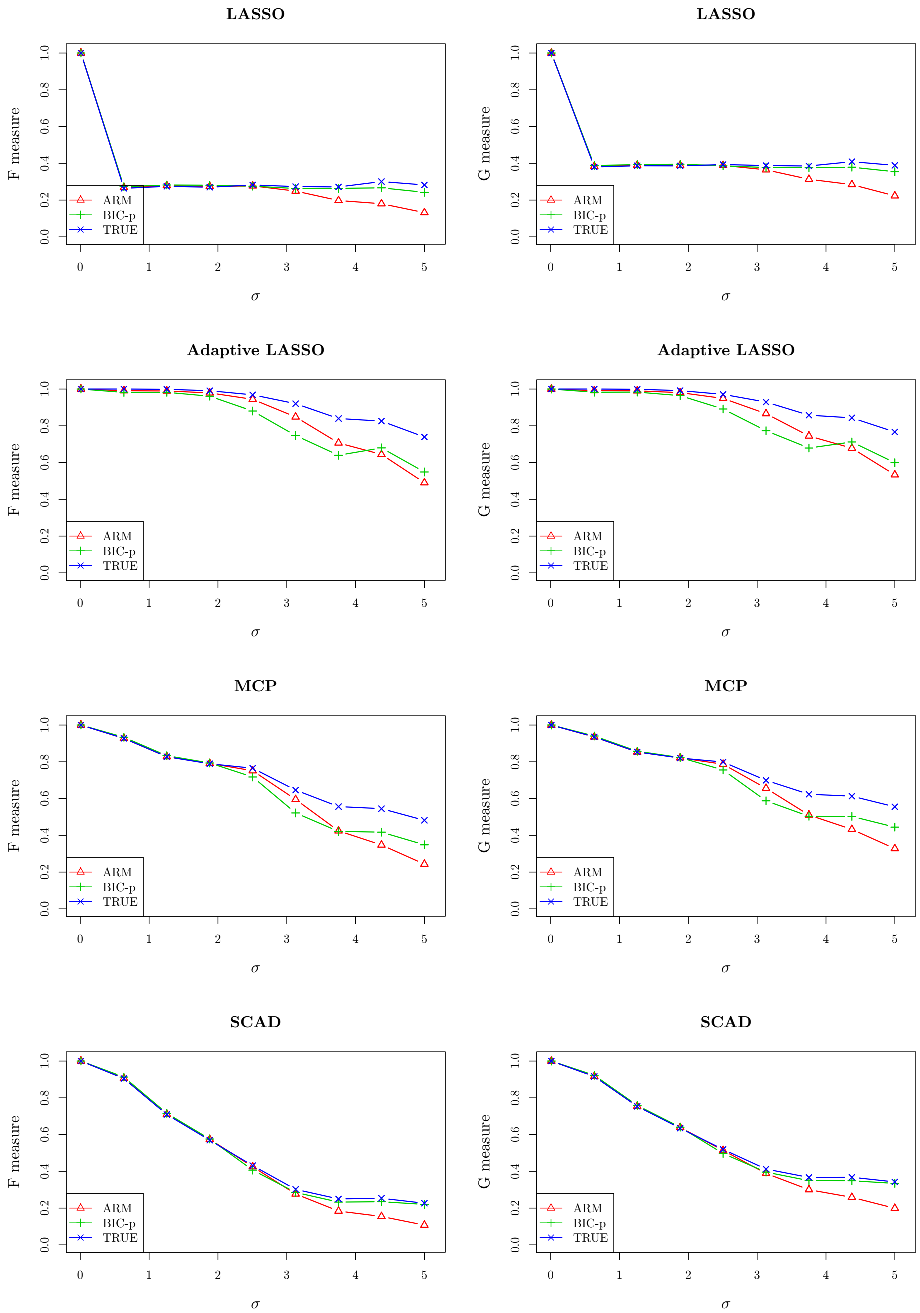

Figure A2: *Regression case (Example 3)*

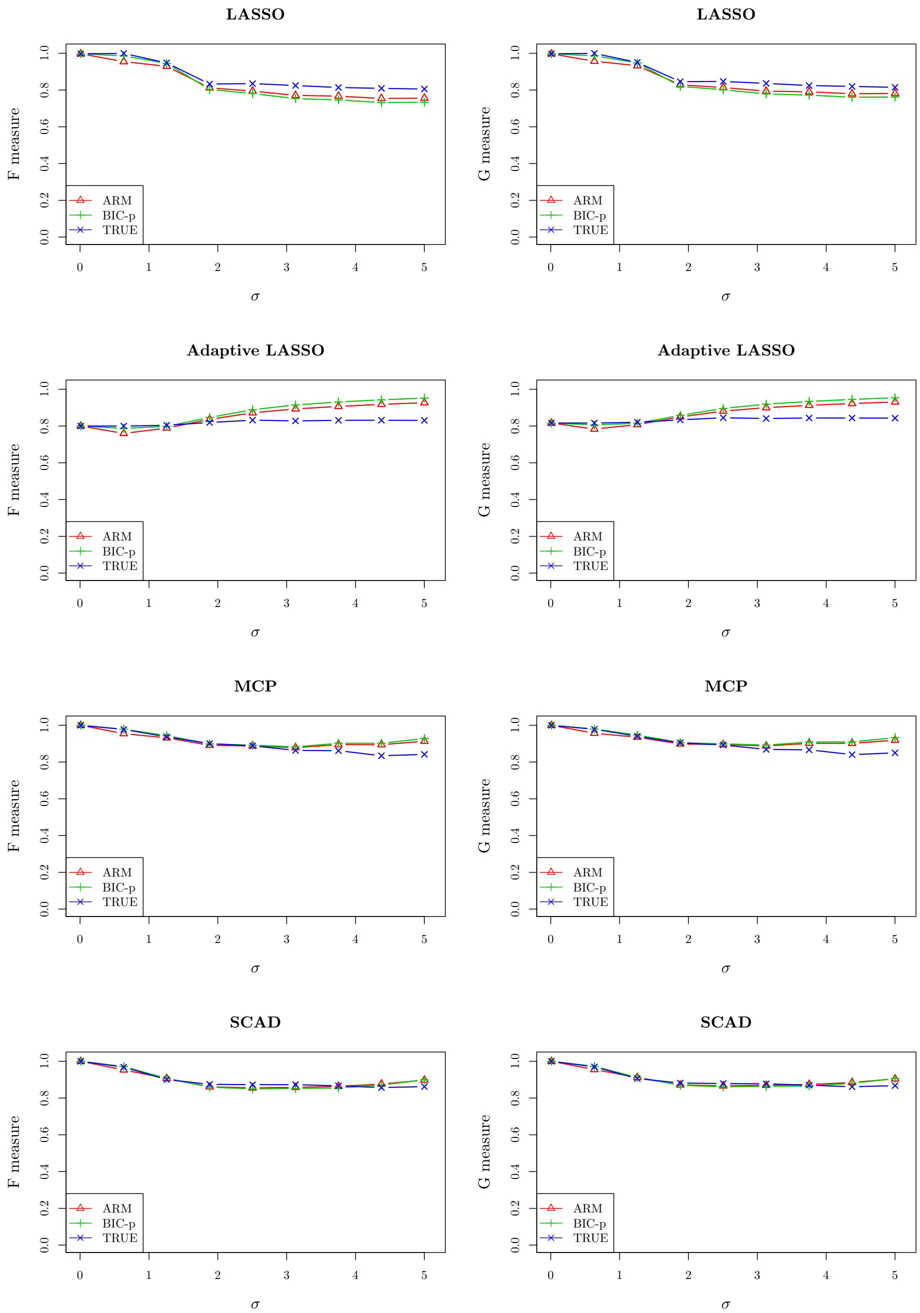

Figure A3: *Regression case (Example 4).*

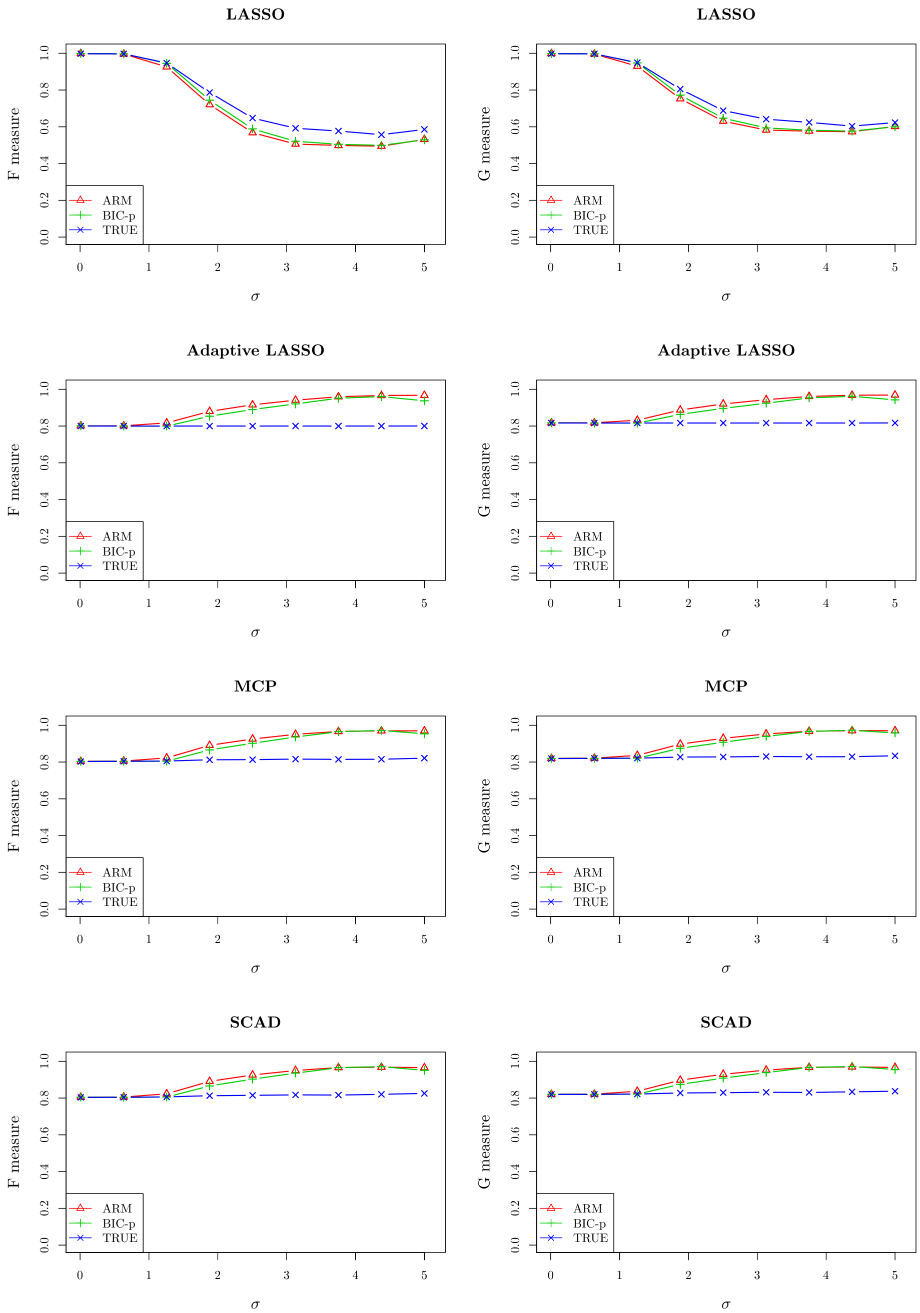

Figure A4: *Regression case (Example 5).*

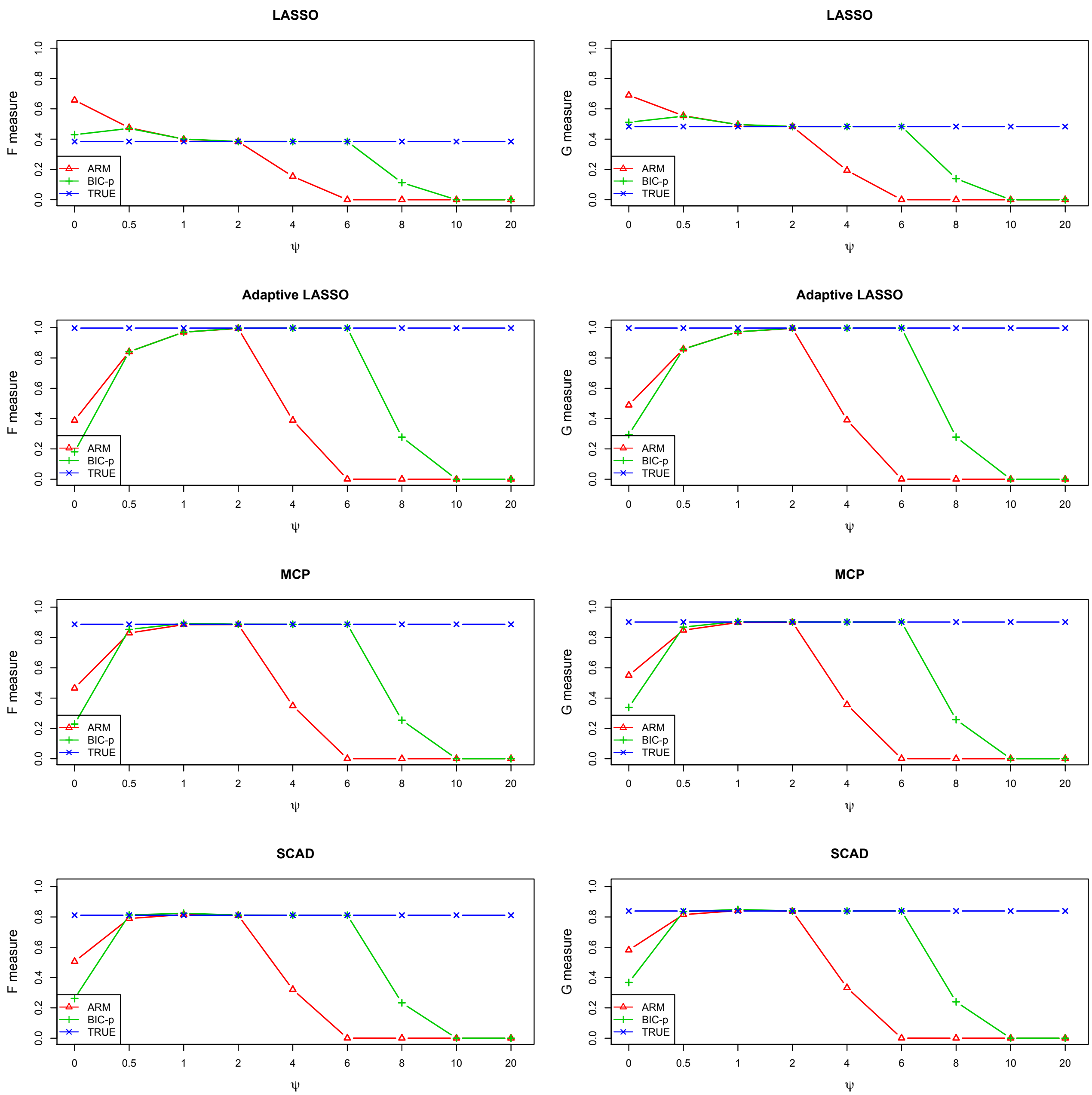

Figure A5: *Sensitivity analysis of ψ. Regression case, $n = 100$ and $p = 200$.*

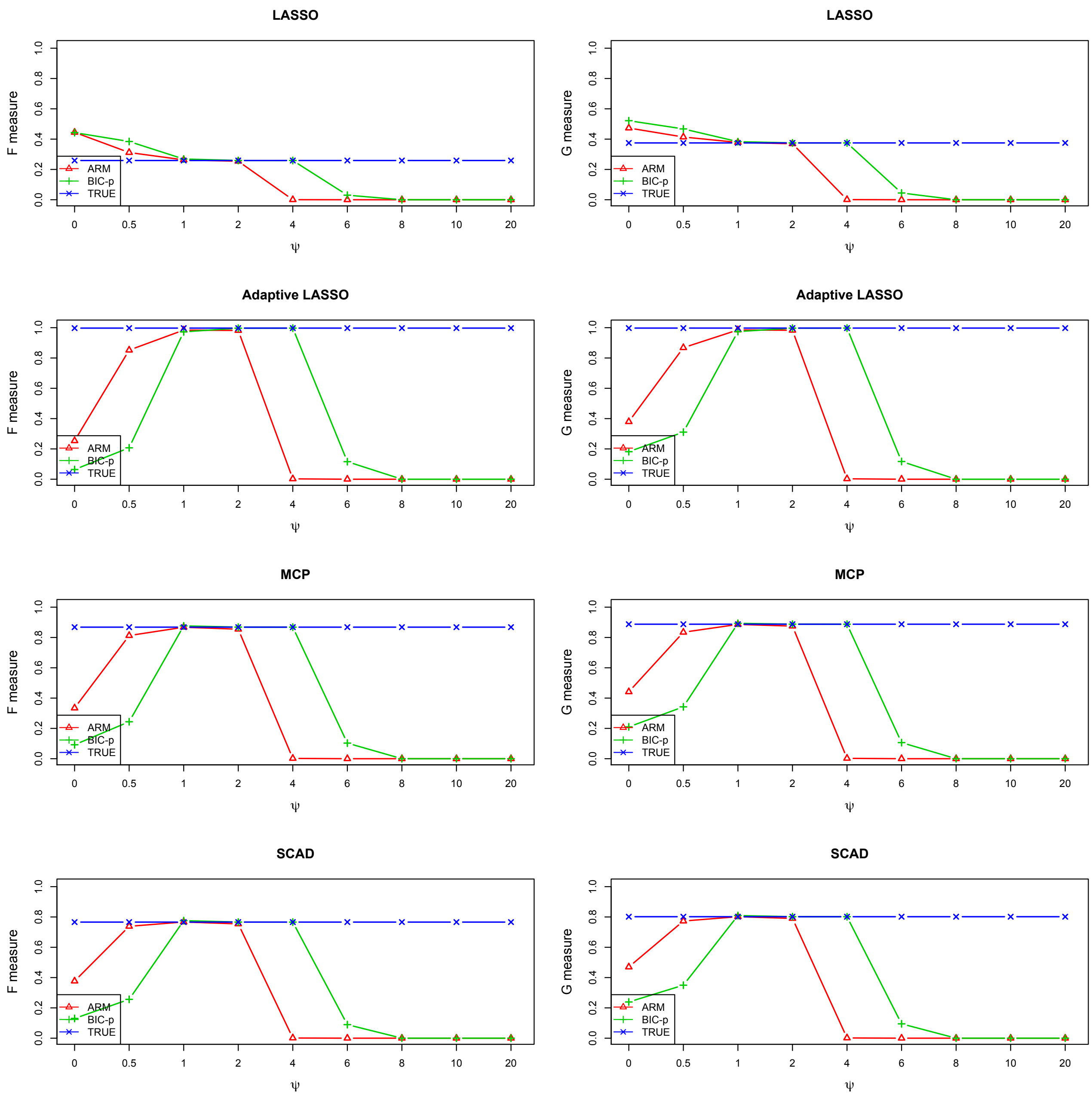

Figure A6: *Sensitivity analysis of* ψ. *Regression case,* $n = 100$ *and* $p = 2000$.

7. Impact of Candidate Models

In this simulation study, we investigate how the quality of the candidate models impacts the estimation performance of PAVI:

- How heterogeneity of the candidate model $\mathbb{S}$ affects the estimation performance.
- How it affects estimation performance when $\mathbb{S}$ contains/not contain the true model.

We only present the results from the regression case. The data are generated using the setting described in Example 3 of Section 5.1, under eight different noise levels σ ranging from 0.01 to 4. We set $n = 50$ and $p = 100$. The true model is represented by the vector $\mathcal{A}^* = (1, 1, 1, 0, 0, 0, \dots, 0)$ with $|\mathcal{A}^*| = 3$, i.e. only the first three variables are nonzero, the remaining 97 are noise variables. Suppose that a given MCP model $\mathcal{A}^0$ is evaluated by using the estimated F-measure $\widehat{F}(\mathcal{A}^0)$ obtained from the BIC-p (the modified BIC) weighting with prior adjustment $\psi = 1$. The sets of candidate models used in estimation of $\widehat{F}(\mathcal{A}^0)$ are generated under the following two settings:

Setting I ($\mathcal{A}^*$ is not included in $\mathbb{S}$.) We use a union of 100 models as the set of candidate models $\mathbb{S} = \{\mathcal{A}^k\}_{k=1}^{100}$. Each $\mathcal{A}^k$ is a contaminated version of the true model $\mathcal{A}^*$ with a pre-specified contamination level $r \in (0, 1)$. Specifically, each $\mathcal{A}^k$ is generated in the following way: we take $\mathcal{A}^*$, randomly select $100r\%$ of its elements and flip their values, i.e. switch to 1 if the original value is 0, and to 0 if the original value is 1. Thus r controls heterogeneity of $\mathbb{S}$: the smaller r becomes, the closer the candidate model gets to the true model.

Setting II ($\mathcal{A}^*$ is included in $\mathbb{S}$.) The set of candidate models $\mathbb{S} = \{\mathcal{A}^k\}_{k=1}^{100}$ is also generated using Setting I, except that one of $\mathcal{A}^k$'s is replaced by $\mathcal{A}^*$.

We compare estimation performances of $\widehat{F}(\mathcal{A}^0)$ under Setting I and II with varying contamination levels $r = \{0.01, 0.03, 0.05, 0.1, 0.2\}$. All simulation cases are repeated for 100 times and the corresponding values are computed and averaged. The results are shown in Figure A7: (1) The left panel shows the results under Setting I. We find that less heterogeneity in $\mathbb{S}$ leads to better estimation performance of $\widehat{F}(\mathcal{A}^0)$ when $\mathcal{A}^* \notin \mathbb{S}$. This indicates that, if the true model is not included in the candidate models, it leads to better performance when $\mathbb{S}$ has most of its models being close to the true model; (2) However, from the results under Setting II shown in the right panel, we can see that if the true model is included in $\mathbb{S}$, then heterogeneity of $\mathbb{S}$ becomes not much influential on the estimation performance.

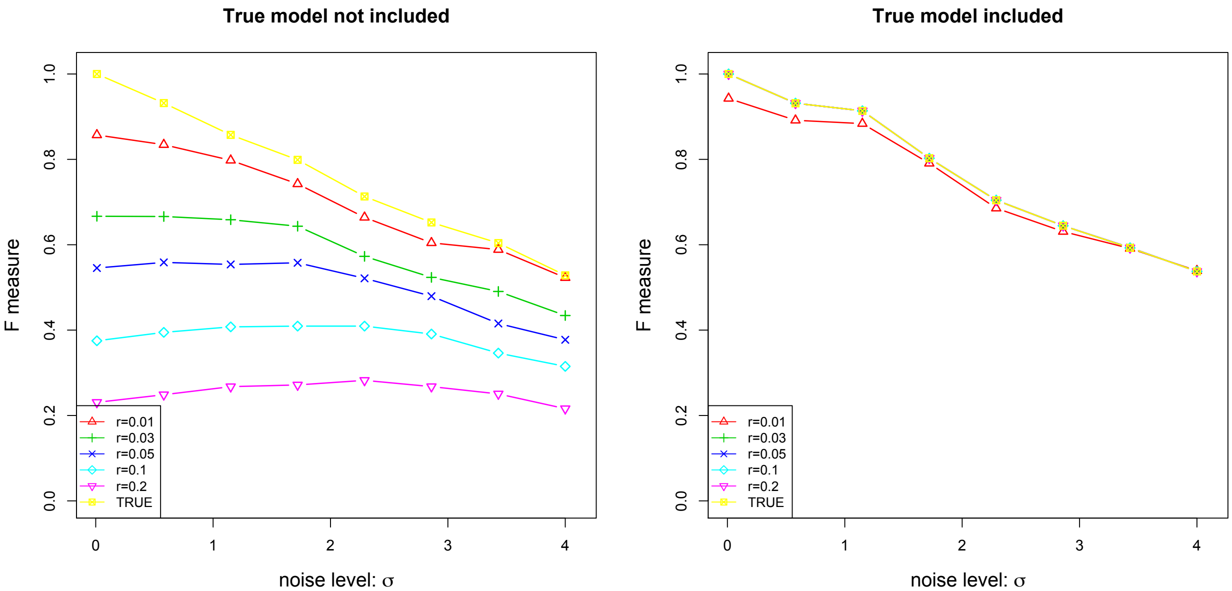

Figure A7: *Impact of candidate models on estimation performance of F-measures in the regression case, $n = 50$ and $p = 100$, under **Setting I:** $\mathcal{A}^*$ is not included in $\mathbb{S}$ (left panel); **Setting II:** $\mathcal{A}^*$ is included in $\mathbb{S}$ (right panel) with varying contamination levels* $r = \{0.01, 0.03, 0.05, 0.1, 0.2\}$.

8. Additional Real Data Examples

Table A5: *Estimated F- and G-measures and standard deviations for* `Prostate`*. L10 has numerically zero $\widehat{F}$ and $\widehat{G}$ values (bolded in the Table).*

	ARM				BIC-p			
	F	$sd.F$	G	$sd.G$	F	$sd.F$	G	$sd.G$
Lasso	0.064	0.004	0.181	0.005	0.064	0.003	0.181	0.004
AdLasso	0.190	0.011	0.323	0.009	0.189	0.008	0.323	0.007
MCP	0.018	0.019	0.027	0.022	0.018	0.012	0.027	0.014
SCAD	0.097	0.006	0.225	0.007	0.096	0.005	0.225	0.005
ImpS	0.333	0.011	0.447	0.008	0.333	0.012	0.447	0.009
S12	0.395	0.037	0.494	0.047	0.400	0.003	0.500	0.007
L10	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

Table A6: *Labels of selected genes for* `Colon`*.*

	Labels of selected genes
Lasso	{66, 249, 377, 493, 765, 1325, 1346, 1423, 1582, 1644, 1772, 1870}
AdLasso	{249, 377, 765, 1582, 1772, 1870}
MCP	{249, 377, 1644, 1772, 1870}
SCAD	{377, 617, 765, 1024, 1325, 1346, 1482, 1504, 1582, 1644, 1772, 1870}
ImpS	{249, 1772}
L11	{249, 286, 765, 1058, 1485, 1671, 1771, 1836}
Y10	{14, 161, 249, 377, 492, 493, 576, 792, 822, 1042, 1210, 1346, 1400, 1423, 1549, 1635, 1772, 1843, 1924}
C11	{249, 399, 513, 515, 780, 1042, 1325, 1582, 1771, 1772}
L10	**{732, 994, 1473, 1763, 1794, 1843}**

Table A7: *Labels of selected genes for* `Leukemia`*.*

	Labels of selected genes
Lasso	{804, 1239, 1674, 1745, 1779, 1796, 1834, 1882, 1928, 1933, 1941, 2121, 2288, 3847, 4196, 4328, 4847, 4951, 4973, 5002, 5107, 5335, 5766, 6055, 6169, 6539, 6855}
AdLasso	{1779, 1834, 4328, 4847, 4951}
MCP	{804, 1941, 3837, 4714, 4847, 4951, 6539}
SCAD	{804, 1674, 1745, 1779, 1834, 1882, 1928, 1941, 2288, 3847, 4196, 4328, 4847, 4951, 4973, 5002, 5766, 5772, 6169, 6225, 6281, 6539, 6855}
ImpS	{1239, 4847, 4951}
J11[1]	{1376, 1394, 1674, 1882, 2186, 2402, 6200, 6201, 6803}
J11[2]	{1394, 1674, 1882, 2186, 5976, 6200, 6201, 6806}
Y10	{760, 804, 1745, 1829, 1834, 1882, 2354, 3320, 4052, 4211, 4377, 4535, 4847, 5039, 6041, 6218, 6376, 6540}
L10	{220, 1086, 1834, 2020}

Table A8: *Labels of selected genes for* `Prostate`.

	Labels of selected genes
Lasso	{1107, 3617, 4282, 4438, 4525, 4636, 5661, 5838, 5890, 6145, 6185, 6838, 7375, 7428, 7539, 7623, 7915, 8123, 8965, 9034, 9093, 9816, 9850, 10234, 10537, 10956, 11858, 11871, 12153, 12462}
AdLasso	{5661, 5890, 6185, 7539, 7623, 8965, 9034, 9093, 10234, 11858}
MCP	{7623, 7924, 8965, 9034, 9816, 10234, 11858}
SCAD	{1107, 3540, 4636, 5661, 5838, 5890, 6185, 7623, 8603, 8965, 9034, 9093, 9816, 10234, 10956, 11858, 11871, 12153}
ImpS	{8965, 9034, 10234, 11858}
S12	{4377, 6185, 6390, 6915}
L10	{4743, 6096, 8475, 9575, 9927, 12331}

References

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 58*(1), 267–288.

Zhang, C.-H. (2010, 04). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.