The fused Kolmogorov filter: a nonparametric model-free screening method

Qing Mai* **and Hui Zou****

*Florida State University **University of Minnesota

1. INTRODUCTION

Data: $(\mathbf{X}^i, Y^i)_{i=1}^n$, where $\mathbf{X}^i$ is a p-dimensional predictor and Y^i is a one-dimensional response.
Challenge: $p \gg n$.
Sparsity assumption: define

$$\mathbf{D} = \{j : F(y \mid \mathbf{X}) \text{ functionally depends on } X_j \text{ for some } y\},$$

where $F(y \mid \mathbf{X})$ is the conditional cumulative probability function of Y. Then we assume that $|\mathbf{D}| \ll p$.
Variable selection: Aim to detect $\mathbf{D}$ exactly.

- Penalized methods, such as Lasso (Tibshirani (1996)) and SCAD (Fan and Li (2001)), among others.
- Numerically challenging.

2. VARIABLE SCREENING

Variable screening: Aim to detect a set $\mathbf{S}$ such that $|\mathbf{S}| \ll p$ and $\mathbf{D} \subset \mathbf{S}$.

- Rank the predictors marginally;
- Computationally efficient;
- Penalized methods can be applied to the reduced set $\mathbf{S}$ as a second-step analysis to identify $\mathbf{D}$ exactly.

SURE screening property: a variable screening method enjoys the SURE screening property if it can identify $\mathbf{S}$.

Our goal: develop a screening method that is:

- robust: it should require minimum distribution assumption so that it can handle heavy-tailed and/or skewed predictions that often emerge in practice;
- model-free: it should enjoy the SURE screening property without specifying a model between Y and $\mathbf{X}$ so one could apply any penalized method to detect $\mathbf{D}$ in the second step;
- unified: it can be applied to a wide range of problems including regression and classification problems;
- invariant: the screening results should remain the same if we transform the variables marginally.

3. EXISTING METHODS

- Many screening methods have been proposed for different models;
- Two most robust methods:
  - Distance correlation screening (DCS): ranks the importance of each predictor by its distance correlation (Székely, Rizzo, M. L. and Bakirov (2007)) with the response. The predictors with large distance correlations are kept.
  - The Kolmogorov filter (K-filter): for a binary classification problem, ranks the importance of each predictor by the Kolmogorov-Smirnov test statistic. The predictors with large test statistics are kept.

	Distance correlation screeing	The Kolmogorov filter
Model-free	Yes	Yes
Distribution-free	No	Yes
Unified	Yes	No
Invariant	No	Yes

4. OUR METHOD: The FUSED KOLMOGOROV FILTER

The fused Kolmogorov filter, involves slicing the response and two levels of fusion:

- **Step 1: Slicing**
  Define a partition

$$\mathbf{G} = \{[a_l, a_{l+1}) : a_l < a_{l+1}, l = 0, \cdots, G-1, \text{ and } \cup_{j=1}^{G-1} [a_l, a_{l+1}) \backslash \{a_0\} = \mathbb{R}\},$$

  where $a_0 = -\infty$ and $a_G = \infty$. Each $[a_l, a_{l+1})$ is called a slice. Then Define a random variable $H \in \{1, \cdots, G\}$ such that $H = l + 1$ if and only if Y is in the l'th slice.
- **Step 2: Fusion within a slicing scheme**
  Given a partition $\mathbf{G}$, we calculate

$$\hat{K}_j^{\mathbf{G}} = \max_{(l,m)} \sup_y |\hat{F}_j(x \mid H_j = l) - \hat{F}_j(x \mid H_j = m)|,$$

  where $\hat{F}_j(x \mid H = l) = \frac{1}{n_l} \sum_{H^i = l} 1(X_j^i \le x)$, and n_l is the sample size within the lth slice and $H^i = l$ if Y^i is in the lth slice.
- **Step 3: Fusion between slicing schemes**
  Repeat Steps 1–2 with N different slicing schemes, $\mathbf{G}_i$ for $i = 1, \ldots, N$. Then we let

$$\hat{K}_j = \sum_{i=1}^{N} \hat{K}_j^{\mathbf{G}_i}.$$

- **Step 4: Find $\hat{\mathbf{S}}$**
  Set $\hat{\mathbf{S}} = \{j : \hat{K}_j > \nu_n\}$, where ν_n is a pre-defined positive constant.

Remark 1 *1. On the population level, if X_j is independent of Y, then $K_j = 0$;*

2. On the population level, if (Y, X_j) is bivariate normal, then K_j is a monotone function of the Pearson correlation between Y and X_j.

3. It is recently observed that the fusion between different slicing schemes greatly improves the efficiency of a sufficient dimension reduction method (Cook and Zhang (2014)). But our method is the first one that applies this method for high-dimensional data.

4. The fused Kolmogorov filter is invariant under univariate monotone transformations.

Theorem 1 *Under mild assumptions, if $\log p = n^\xi$ for some $0 < \xi < 1$, then the fused Kolmogorov filter enjoys the SURE screening property with a probability tending to 1.*

Remark 2 *Our theorem does not impose any distribution assumption on Y or $\mathbf{X}$;*

5. SIMULATIONS

In all simulations, $n = 200, p = 5000$.
Model 1 (Linear transformation model) $\log Y = 2.8X_1 - 2.8X_2 + \epsilon$, where $X \sim N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = CS(0.7)$ and $\epsilon \sim N(0, 1)$.
Model 2 (Additive model): $Y = 4X_1 + 2\tan(\pi X_2/2) + 5X_3^2 + \epsilon$, where X_j's follow $\mathrm{Unif}(0, 1)$ independently and $\epsilon \sim N(0, 1)$ is independent of $\mathbf{X}$.
Model 3 (Heteroskedastic regression model): $Y = 2(X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5) + \exp(X_{20} + X_{21} + X_{22})\epsilon$, where $\epsilon \sim N(0, 1)$, and $\mathbf{X} \sim N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \mathrm{AR}(0.8)$.

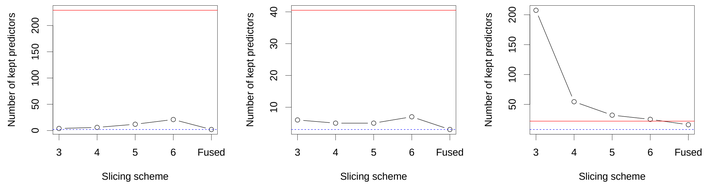

Figure 1: *Simulation results for Models 1, 2 & 3 (from left to right) based on 500 replicates. We report the minimum number of predictors we need to keep so that all the important predictors are kept. The blue dashed line is the truth. The black line denotes the Kolmogorov filter based on 3–6 slices and the fusion of all the slicing schemes. The red line is the results given by distance correlation screening.*

6. REAL DATA

- The Tecator dataset was collected by Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050 nm by the Near Infrared Transmission (NIT) principle. The predictors are 100 channel spectrum of absorbances. The response is the percentage of fat in finely chopped meat.
- After deleting 2 outliers, we randomly split the dataset to form training sets of size 215 and testing sets of size 41.
- In addition to the 100 predictors in the original dataset, we added 4900 independent noise variables following the Cauchy distribution.

We compare five successful screening methods in the literature with the fused Kolmogorov filter.

	Kolmogorov	DCS	NIS	SIS	QA		ELS
					$\alpha = 0.5$	$\alpha = 0.75$	
True predictors	99.6	75.4	77.3	11.7	45.4	42.2	6.24
	(0.06)	(0.44)	(0.28)	(0.27)	(0.56)	(0.43)	(0.14)

Table 2: *Comparison of the screening methods on the tecator dataset. We report the number of true predictors that are preserved after the screening step. The numbers are average over 100 replicates. Standard errors are in parentheses.*

We further combine the three methods that kept the most true predictors with random forest to compare the prediction accuracy.

	K-RF	DCS-RF	NIS-RF
Average MSE	0.097	0.102	0.103
	(0.009)	(0.010)	(0.010)

Table 3: *Comparison of the prediction performance on the tecator dataset. The numbers are average over 100 replicates. Standard errors are in parentheses. A paired t-test shows that K-RF is significantly better than DCS-RF and NIS-RF, with p-values less than 1×10^{-5}.*

7. MAIN REFERENCES

COOK, R. D. and ZHANG, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Amer. Statist. Assoc.* **109**, 815–827.
LI, R., ZHONG, W. and ZHU, L. P. (2012). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.*, **107**, 1129–1139.
MAI, Q. and ZOU, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, **100**, 229–234.