A flexible shrinkage operator for fussy grouped variable selection

By: Xiaoli Gao

Gao, X.L. (2016). A Flexible Shrinkage Operator for Fussy Grouped Variable Selection, *Statistical Papers*. DOI:10.1007/s00362-016-0799-y

The final publication is available at Springer via http://dx.doi.org/10.1007/s00362-016-0799-y

Abstract:

Existing grouped variable selection methods rely heavily on prior group information, thus they may not be reliable if an incorrect group assignment is used. In this paper, we propose a family of shrinkage variable selection operators by controlling the k-th largest norm (KAN). The proposed KAN method exhibits some flexible group-wise variable selection naturally even though no correct prior group information is available. We also construct a group KAN shrinkage operator using a composite of KAN constraints. Neither ignoring nor relying completely on prior group information, the group KAN method has the flexibility of controlling within group strength and therefore can reduce the effect caused by incorrect group information. Finally, we investigate an unbiased estimator of the degrees of freedom for (group) KAN estimates in the framework of Stein's unbiased risk estimation. Extensive simulation studies and real data analysis are performed to demonstrate the advantage of KAN and group KAN over the LASSO and group LASSO, respectively.

Keywords: Degrees of freedom | Group shrinkage | k-th largest norm | Shrinkage estimator | Variable selection

Article:

Consider a high-dimensional sparse linear regression model,

$y_i=\beta 0+x'_i\beta\beta+\varepsilon_i$, $1\leq i \leq n$, (1)

where $xi=(x_{i1},\ldots,x_{ip})'$ is a p− dimensional predictor, y_i is a univariate response variable, and εiεi's are independent and identically distributed random variables. Without loss of the generality, we assume both the response variable and predictors to be centered and standardized such that $\Sigma_{i=1}^{n} y_i = 0$, $\Sigma_{i=1}^{n} x_{ij}=0$ and $\Sigma_{i=1}^{n} X_{ij}^2=n$. Thus, $\beta_0=0$ is assumed in the true model. We are interested in estimating regression coefficients vector $\beta\beta=(\beta_1,\ldots,\beta p)'$. The true model in (1) is high-dimensional if the number of covariates *p* is much larger than the sample size *n*. The model is sparse since most elements in β are zero.

Variable selection is an important issue for such a high-dimensional sparse model. In the last twenty years, the least absolute shrinkage selection operator [LASSO, (Tibshirani 1996)] has

attracted much attention in generating sparse solutions in high-dimensional data analysis because of the simultaneous variable selection and estimation. If we denote the response vector as y=(y1,…,yn)′y=(y1,…,yn)′ and the covariate data matrix as X=(x′1,…,x′n)′, a LASSO estimator for model (1) minimizes

$\|y - X\beta\|_2^2$ subject to $\|\beta\|_1 \leq s$, (2)

or a penalizing loss function,

$\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ for some λ>0, (3)

where both s>0 and λ>0 are tuning parameters, $\|\cdot\|_2$ and $\|\cdot\|_1$ represent the ℓ_2 and ℓ_1 norms, respectively.

Because of the non-differentiability of the ℓ1ℓ1 norm, LASSO can provide exact zero estimates for some coefficients when λλ is large enough. However, LASSO does not encourage the group sparsity. The true model is assumed to have the group sparsity (i.e., sparsity at the group level) if a group of variables function together, that is, the entire group is either relevant or irrelevant to the response variable completely. For example, in a gene expression microarray data analysis, genes can be functional in terms of some known biological pathways such as the Kyoto encyclopedia of genes & genomes pathways (Kanehisa and Goto 2000). Another example is a multi-factor ANOVA analysis, where each factor with several levels can be expressed through a group of dummy variables. In both examples, all covariates can be clustered into certain groups and exhibit possible effects to the response variables at the group level. An ideal grouped variable selection method should be able to detect all important groups and shrink irrelevant ones to be 0. Yuan and Lin (2006) and Kim et al. (2006) extended the LASSO in (3) to a group LASSO (GLASSO) penalty for grouped variable selection. Asymptotic properties of group Lasso in generalized linear models was also studied in Wang et al. (2015).

Suppose that *p* covariates are pre-assigned into *G* non-overlapping groups. Then the covariates matrix $X=(X_1,\ldots,X_G)$ and the coefficients vector $\beta=(\beta'_1,\ldots,\beta'_G)'$, where Xg only includes covariates from group *g* and ββgββg is the corresponding coefficients sub-vector.

A GLASSO estimator minimizes the penalized objective function

$$\|y - x\beta\|_2^2 + \sum_{g=1}^{G} \lambda_g \left\|\beta_g\right\|_2, \text{ (4)}$$

where $\|\beta_g\|2=(\sum j\in\text{group } g\beta^2_j)^{1/2}$ and λg is a tuning parameter for group *g*.

Zhao et al. (2009) and Zou and Yuan (2008) also performed automatic factor selection in classification using a special composite of absolute penalty family (iCAP). An iCAP estimate is obtained by minimizing

$$\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \left\|\beta_g\right\|\infty, \text{ (5)}$$

where $\|\cdot\|\infty$ is the $\ell\infty$ norm and λ>0 is a tuning parameter.

One of the most important features of GLASSO and iCAP is the “all-in” or “all-out” property. Once an explicit group assignment is provided, GLASSO and iCAP select some important groups: all coefficients within these important groups are estimated to be nonzero (all-in) and all others are zero (all-out). Since both the GLASSO and iCAP do grouped variable selection based on an explicit group assignment and do not generate within-group sparsity, any fussy group information can be misleading. Especially, when the number of groups are under-estimated, both GLASSO and iCAP generate over-fitted models. To this end, a variable selection method with some robust property to the incorrect group assignment is needed.

In this paper, we propose a family of shrinkage variable selection operators by controlling the *k*-th largest norm (KAN), a special case of SLOPE studied in Bogdan et al. (2015) for the investigation of the false discovery rate. Different from the SLOPE, the proposed KAN method is designed to encourage some grouped variable selection without any group information. If some fussy group information is available, we are able to construct a group KAN shrinkage operator using a composite of KAN constraints. Neither ignoring nor relying completely on the prior group information, the group KAN method is able to improve the grouped variable selection efficiency using some correct group information, and at the same time, reduce the effect caused by the incorrect information. Furthermore, we investigate an unbiased estimator of the degrees of freedom of KAN estimates. The effect of the unbiased estimator is confirmed by some numerical studies. Such an unbiased estimate can be applied in any existing information criterion such as Akaike’s Information Criterion [AIC, (Akaike 1973)], Bayesian Information Criterion [BIC, (Schwarz 1978)] and extended BIC [EBIC, (Chen and Chen 2008)] to provide an optimal KAN or group KAN fit.

The remainder of the paper is structured as follows. In Sect. 2, we introduce both the KAN without group information and its extension to the group KAN with certain prior group information. In Sect. 3, we provide an unbiased estimator of the degrees of freedom of (group) KAN estimates. We discuss the computation of (group) KAN estimates in Sect. 4. In Sects. 5 and 6, we use extensive simulation studies and a real data example to demonstrate the performance of these methods over existing group shrinkage methods. Section 7 contains some discussion of our methods and some future directions. Some proofs are presented in the Appendix.

2 The K-largest norm shrinkage operator

For any coefficient vector β in (1), the *k-th largest norm*, denoted as the ℓ_k norm, is given by

$$\|\beta\|_{(k)} = \sum_{j=1}^{k} \left|\beta_{(j)}\right| , \ 1 \le k \le p, \quad (6)$$

where $|\beta_{(1)}| \ge \cdots \ge |\beta_{(p)}|$. Both $\ell 1$ and $\ell\infty$ norms are two special cases of the ℓ_k norm for k=p and k=1, respectively. The polyhedron generated from the *k*-th largest norm is convex.

Lemma 1

Let $K=\{\beta \in R^p : \|\beta\|_{(k)} < s\}$ for $1 \le k \le p$ and $s > 0$. Then K is convex.

Such a convexity is very important in both computation and theoretical investigation. The proof of Lemma 1 can be found in the Appendix.

2.1 KAN estimator

For the linear model (1) and the ℓ2ℓ2 loss function, a *k*-th largest norm shrinkage (KAN) estimator is defined by

$\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)'=\arg\min\{\|y-X\beta\|_2^2\}$ subject to $\|\beta\|_{(k)}\leq s$. (7)

Here ‖ββ‖(k)‖ββ‖(k) is the *k*-th largest norm defined in (6), s>0s>0 is a tuning parameter controlling the overall shrinkage strength, and 1≤k≤p1≤k≤p is the other tuning parameter adjusting the grouping strength on the shrinkage. At one extreme, KAN with k=pk=p reduces to the LASSO, encouraging individual sparsity among all covariates. At the other extreme, KAN with k=1k=1 reduces to iLASSO (Zhao et al. 2009), encouraging the entire group sparsity (all-in or all-out) among all covariates. In fact, the motivation of the KAN came from the study of LASSO and iLASSO, where the former does not encourage the group shrinkage, while the latter does not encourage the individual shrinkage.

The KAN in (7) is equivalent to minimizing the penalized least squares function,

$\|y-X\beta\|_2^2+\lambda\|\beta\|_{(k)}$, for some λ>0. (8)

Let β_0 be the true parameter value in (1) and $\beta_0=(\beta'_{10},\beta'_{20})'$, where β_{10} includes all *q* non-zero elements and β_{20} consists of all zeros. Suppose the design matrix satisfies certain regularity conditions as follows,

$C_n=n^{-1}\sum_{i=1}^{n}X_iX_i'\rightarrow$ C for some positive definite C > 0 (9)

and

$n-1\max_{1\leq i\leq n}x'_ix_i\rightarrow 0$. (10)

Then KAN has two important features given in the following theorem.

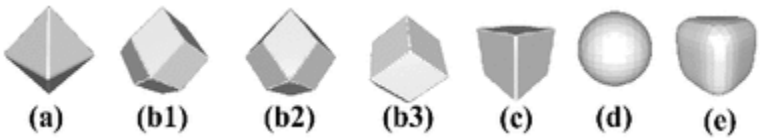

Fig. 1 The 3D contour plots of enforced constraints. **a** KAN with k=3 (LASSO), **b1**–**b3** KAN with k=2 from different directions, **c** KAN with k=1 (iLASSO), **d** bridge with γ=2, **e** bridge with γ=4

Theorem 1

If $\lambda/n\rightarrow 0$and $\lambda/\sqrt{n}\rightarrow\infty$, with probability tending to 1,

1. (i) $\hat{\beta}(\lambda,k)$ can include zero element when k > q;

2. (ii) $\hat{\beta}(\lambda,k)$ includes only non-zero element or all zero ones when $k \leq q$.

Theorem 1 tells us that KAN has two advantages compared to LASSO and iLASSO. LASSO only addresses the individual shrinkage and iLASSO only addresses the group shrinkage. However, the KAN has the ability of encouraging both, where *s* in (7) and λ in (8) are used to adjust the individual shrinkage, and *k* is used to control group shrinkage. We provide the proof of Theorem 1 in the Appendix.

In Fig. 1, we provide 3-D contour plots of different constraints for p=3p=3. The contour plot in (a) shows that the LASSO only generates the individual sparsity with no preference among covariates. The contour plot in (c) shows that the iLASSO only generates the group sparsity, but not the individual sparsity. (b1-b3) are the same contour plot of the KAN (k=2k=2) observed from different angles. As a comparison, we also provide the contour plots of constraints in the bridge penalty (Frank and Friedman 1993) with γ=2γ=2 (ridge) and γ=4γ=4 (over-ridge) in (d) and (e), where no individual sparsity is generated.

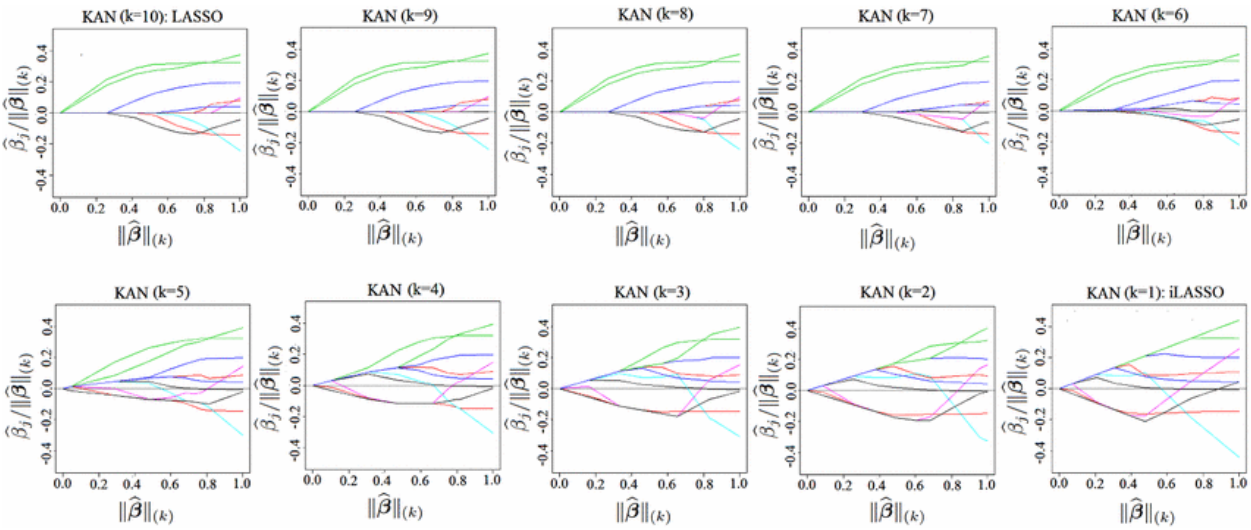

Fig. 2 Regularization path of KAN estimates for diabetes data from k=p=10 (LASSO) to k=1 (iLASSO). The horizontal and vertical axis contain the $\ell_{(k)}$-norm of the normalized coefficients and the normalized coefficients ($\hat{\beta}j/\|\hat{\beta}\|_{(k)}$) respectively. The *top* and *bottom* 5 panels are for k>p/2 and $k \leq p/2$, respectively

We take the classical diabetes data (Efron et al. 2004) as an example to demonstrate how KAN behaves under different *k*. There are p=10 potential predictors in this data. In Fig. 2, we plot the solution path of all 10 covariates from k=p=10 (LASSO) to 1 (iLASSO). The top and bottom five panels are for k>p/2 and k≤p/2, respectively. The top five panels show that the pool of non-zero estimates expands gradually as *k* approaches *p*, while the bottom five ones show that covariates turn to enter the model simultaneously when *k* approaches 1. From Theorem 1, it may suggest that there are at most 5 non-zero coefficients in the true model. In addition, we find no

overlap among all LASSO solution paths (k=p). However, when *k* becomes smaller, solution paths turn to overlap more until the *k*-th largest nonzero estimate becomes unique. Above observations shed some lights on the flexibility of the KAN family in terms of the grouped variable selection and also provide us some important information in finding an unbiased estimator of the degrees of freedom of the KAN estimate in Sect. 3.

It also explains why LASSO does not generate nonzero estimates with the same magnitude in general.

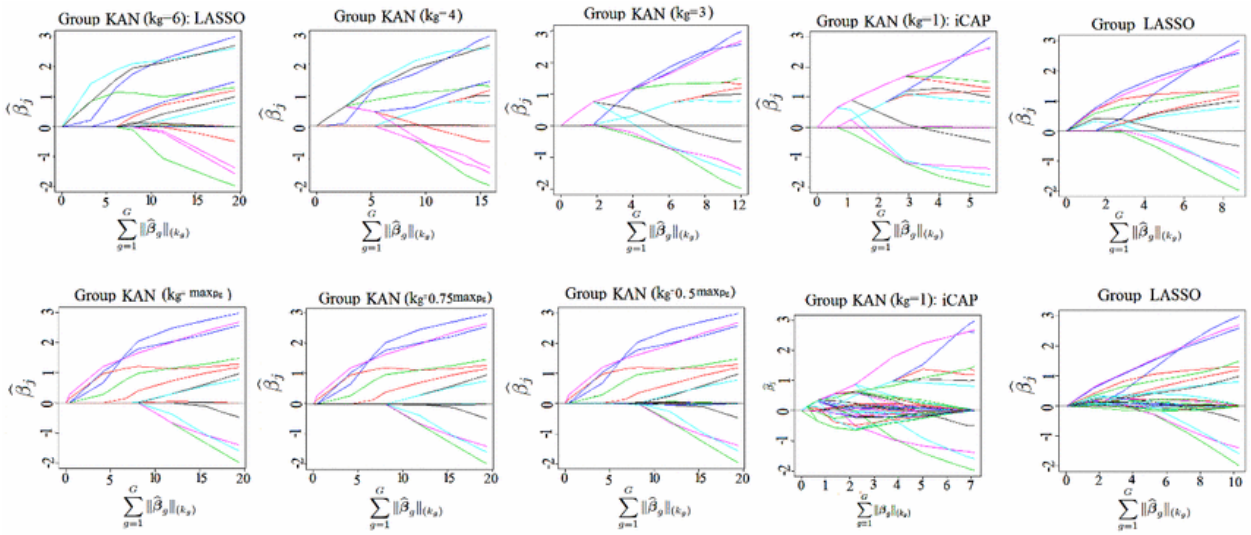

Fig. 3 Solution paths of all 48 coefficients for a simulated data generated from Example 3 in Sect. 5.2. The *top five panels* are produced using the correct group structure. The *bottom five panels* are produced using an incorrect group assignment. Panels from the *left* to the *right* are for GKAN with kg=pg,0.75pg,0.5pg,1 and the GLASSO, respectively

2.2 A composite of KAN constraints for group shrinkage

If the group assignment of the most but not all covariates are correct, the goal is to make use of the correct information to gain variable selection efficiency, while reduce the effect caused by the incorrect information. To reach this goal, we construct a group KAN model based on a composite of KAN constraints. First, we present the group KAN method based on the disjoint group assignments, and then extend it to dealing with hierarchical structured sparsity. Suppose *p* covariates are divided into *G* groups and the group *g* includes pgpg covariates for 1≤g≤G. A group KAN (GKAN) estimator is defined by,

$$\hat{\beta}^{GKAN} = \arg min_{\beta \in R^p} \{\|x - x\beta\|_2^2\} \text{ subject to } \sum_{g=1}^{G} \|\beta_g\|_{(kg)} < s. \quad (11)$$

Here s>0 is a shrinkage tuning parameter and $\|\beta g\|_{(kg)}$ for $1 \le k_g \le p_g$ is the kgkgth largest norm of coefficients in group *g*. A variety of kgkg's provide different group-wise shrinkage strength within group *g*. If $k_g=p_g$ for all *g*'s, then the GKAN reduces to the LASSO, enforcing no group-wise shrinkage among all covariates. If $k_g=1$ for all *g*'s, then the GKAN reduces to the iCAP, enforcing strongest group-wise shrinkage. In this case, variables within each group are either all-

in or all-out. One of the advantages of the GKAN is that the group shrinkage is affected by both a priori group assignment and the value of kg's. Thus if some covariates are incorrectly assigned in group *g*, a well-chosen group shrinkage factor $1 \le k_g \le p_g$ is expected to reduce the damage caused by the mis-classification within group *g*. It is worthwhile to point out that a special case of the GKAN for all kg's being 1 has also been adopted in other studies to encourage the grouping feature selection. However, to our knowledge, no previous work has investigated the grouping strength by the relaxation of kg=1 and pg to $1 \le k_g \le p_g$. It is novel to connect the variable selection at both the group level and individual level using additional factors, $k_1,\ldots,k_G$. In Sect. 4, we will give some details on how to choose kg's to save the computational cost.

In Fig. 3, we plot the solution path of GKAN estimates for a simulated data from Example 3 in Sect. 5.2. There are 2 relevant groups and 6 irrelevant ones. Each group includes pg=6pg=6covariates. The top five panels are produced adopting correct group information in both GKAN and GLASSO. The bottom five panels are produced based upon an incorrect group assignment, where all 48 covariates are clustered into only 4 groups using the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw 1990). Five panels from the left to the right are for KAN with kg=pg,0.75pg,0.5pg,1kg=pg,0.75pg,0.5pg,1 , and the GLASSO, respectively. The bottom last two panels show that both the iCAP and GLASSO are affected by the wrong group assignments severely.

In many cases, potential variables tend to affect the response variable hierarchically. For example, in an ANOVA model with both main effects and interactions, we often assume main effects to be important as long as its generated interactions are important. Similar to Zhao et al. (2009), we can simply use the overlap group information to realize such a nested hierarchical structure. Similarly, in the gene expression data analysis, if gene *j* functions with other genes from both two different pathways and other genes from those two pathways function independently, then gene *j* belongs to two groups simultaneously. If overlapping groups exist, some variables can be fuzzily assigned into more than one groups. Since GKAN has some robust properties to a fussy group assignment, one can still perform GKAN by either merging or separating all those overlapped groups. In Sect. 5.3, we will use two simulation examples to demonstrate the performance GKAN under both overlapping and hierarchical group structures.

3 The degrees of freedom

To compute the degrees of freedom, we should first bear in mind that for a given s>0s>0, the KAN estimates for any group strength factor $kg, 1 \le g \le Gkg, 1 \le g \le G$ is a modeling procedure, say MsMs, including both model selection and model fitting. The degrees of freedom measures the complexity of a modeling procedure, which can be computed by the sum of the sensitivity of the predicted values (Ye 1998). Let $\hat{\beta} \equiv \hat{\beta}$ (s,k)=($\hat{\beta}_1$(s,k),…, $\hat{\beta}$p(s,k))′ be a KAN estimate for some s>0 and $1 \le k \le p$. The predicted values are $\hat{\mu}_i$(y;s,k)= $\Sigma_{j=1}^{p} x_{ij}$ $\hat{\beta}$j(s,k) for $1 \le i \le n$. Then the degrees of freedom (df) of a KAN estimate $\hat{\beta}$ (s,k) is defined as

$$\mathrm{df}(\hat{\beta})=\sum_{i=1}^{n}\frac{\partial E[\hat{\mu}(y;s,k)]}{\partial \mu_i}=\lim_{\delta\to 0}E\,\frac{\hat{\mu}_i(y+\partial e_i;s,k)-\hat{\mu}_i(y;s,k)}{\delta},\quad(12)$$

where e_i is the $_i$th column of the n×n identity matrix.

3.1 Unbiased estimator

Let $A_{s,k}=\{1\le j\le p:\hat{\beta}_j(s,k)\neq 0\}$ be the active set of a KAN estimate $\hat{\beta}$ (s,k) with $1\le k\le p$ and s>0. Let $U_{s,k}=\{(j_1,\ldots,j_k):\Sigma_{i=1}^{k}=1|\hat{\beta}_{ji}|=\|\hat{\beta}\|_{(k)}$ and $1\le j_1\neq\cdots\neq jk\le p\}$. From Fig. 2 in Sect. 2, we have seen that solution paths of nonzero KAN estimates turn to overlap first, especially for relatively small *k*. Thus, for some 1≤k≤p1≤k≤p, there may exist some s2>s1>0s2>s1>0, such that $\|\hat{\beta}(s)\|_{(k)}$ and $U_{s,k}$ do not change for $s\in[s_1,s_2]$. Similar to the argument in Kato (2009), we can estimate the degrees of the freedom of the KAN estimates using $A_{s,k}$ and $U_{s,k}$. First, we obtain the absolutely continuity of the KAN solution as follows.

Lemma 2

For any s >0, the KAN fit with $1\le k\le p$, μ(y;s) is a uniformly Lipschitz function of y, and then absolutely continuous.

Both Lemma 1 and 2 help us to obtain an unbiased estimator of the degrees of freedom of a KAN estimate for any s>0 and $1\le k\le p$. We present the main result in Theorem 2 and postpone the corresponding proof to the Appendix.

Theorem 2

For any s > 0 and $1\le k\le p$, an unbiased estimator of the degrees of freedom of the KAN estimate $\hat{\beta}$ is given by

$$\widehat{\mathrm{df}}(\hat{\beta})=\begin{cases}|\mathcal{A}_{s,k}|-|\mathcal{U}_{s,k}| & \text{if } \|\hat{\beta}^0\|_{(k)}>s,\\ p & \text{if } \|\hat{\beta}^0\|_{(k)}\le s\end{cases}\quad(13)$$

where $\hat{\beta}^0$ is the least square estimate of β and |A| is the cardinal value of the set A.

For notation's convenience, we often omit $\hat{\beta}$ and let $\widehat{\mathrm{df}}=\widehat{\mathrm{df}}(\hat{\beta})$. The unbiased estimator in (13) agrees with existing results for the classical LASSO and iLASSO estimates.

If k=p, then $U_{s,p}$ only includes one element. Thus $\widehat{\mathrm{df}}=\#\{1\le j\le p:\hat{\beta}j\neq 0\}-1$ if $\|\beta\hat{}0\|1\le s\|\beta\beta\hat{}0\|1\le s$, This is consistent with the result in Kato (2009) on the $\widehat{\mathrm{df}}$ of the LASSO estimate in (2).

If k=, $U_{s,1}$ reduces to $\{1\le j\le p:\hat{\beta}_j=\|\hat{\beta}\|\infty\}\{1\le j\le p:\hat{\beta}_j=\|\hat{\beta}\|\infty\}$, and then an unbiased $\widehat{\mathrm{df}}$ for an iLASSO estimate is $\#\{1\le j\le p:0\neq|\hat{\beta}_j|<\|\hat{\beta}\|\infty\}$.

Remark 1

Zhao et al. (2009) justified that $1+\#\{1 \le j \le p: 0 \ne |\hat{\beta}_j| < \|\hat{\beta}\|\infty\}$ is an unbiased estimator of the df of an iLASSO estimate. There is a constant difference of "−1" between these two $\widehat{df}$'s. Kato (2009) already pointed out that the same difference between df^df^'s of LASSO solved by (2) and (3), respectively. The same argument can be used to justify the difference between two versions of $\widehat{df}$'s for iLASSO. In this paper, we solve KAN estimate by minimizing a loss function subject to a constraint, while iLASSO in Zhao et al. (2009) is solved by minimizing the penalized objective function, $\|y-X\beta\|22$ $\|y - X\beta\|_2^2 + \lambda\|\beta\|\infty$.

The above unbiased estimator can be also extended to the GKAN in the following theorem.

Theorem 3

For any s>0 and $1 \le k_g \le p_g$, $1 \le g \le G$, an unbiased estimator of the degrees of freedom of the GKAN estimate $\hat{\beta}$ is

$$\widehat{\mathrm{df}} = \begin{cases} |\mathcal{A}_s| - \sum_{g=1}^{G} |\mathcal{U}_s^g| & \text{if } \|\widehat{\boldsymbol{\beta}}^0\|_{(k)} > s, \\ p & \text{if } \|\widehat{\boldsymbol{\beta}}^0\|_{(k)} \le s \end{cases} \quad (14)$$

where $A_s=\{1 \le j \le p: \hat{\beta}\, j \ne 0\}$ is the active set of GKAN and $Ugs=\{(j_1,\ldots,j_{kg}) \in G_g: \Sigma_{i=1}^{kg} =1|\hat{\beta}_{ji}|=\|\hat{\beta}\|(kg)$ and $j_1 \ne \cdots \ne j_{kg}\}$, $\forall\ 1 \le g \le G$. The proof for Theorem 3 is similar to the one for Theorem 2, but much more complicated. Here, we skip the proof and only demonstrate the estimation results using numerical studies.

Remark 2

The above unbiased estimator in Theorem (2) and (3) are only defined under p<np<n.
When p>n the least square estimate $\hat{\beta}^0$ is not uniquely defined. One can analyze the degrees of freedom by representing $\hat{\mu}$ directly as a function of y since $\hat{\mu} = X\hat{\beta}$ is still unique. However, an explicit format of unbiased estimators of GKAN is not available any more. We refer Kato (2009) for some detailed argument.

3.2 Performance of the unbiased estimation

We simulate two hypothetical models constructed upon the diabetes data studied in Efron et al. (2004) and a simulated data from Example 3 in Sect. 5.2.

For the diabetes data, we obtain coefficient estimate using KAN in model (7) with *k* randomly chosen from 1 to 10. For the simulated data, we obtain coefficients estimate from the GKAN estimates in model (11) with kg=4kg=4 being fixed for all groups.

Let β^* and σ^* be corresponding coefficients and residual standard deviation estimation. We generate 500 Monto Carlo repetitions of responses from a hypothetical model,

$y^0_i = x'_i\beta* + \varepsilon^0_i$, i=1,…,n. (15)

For both data, we generate ε^0_i's independently from a normal distribution with center 0 and scale σ^*. Then for each s>0, both $\widehat{\mathrm{df}} = |A_s| - \sum_g |U^g_s|$ and df defined in (12) are computed for all hypothetical data sets.

The averages of df and $\widehat{\mathrm{df}}$ over 500 repetitions for KAN and GKAN are reported as "True DF" and "Estimated DF" in panel (a) and (c) of Fig. 4, respectively. Corresponding estimation biases with a 95 % confidence interval are also reported in panel (b) and (d). We observe that the reported "True DF" and "Estimated DF" are very close to each other, with estimation bias close to 0. Thus results established in both Theorem 2 and 3 are supported by our numerical studies.

4 Computation

From Lemma 1, any existing convex optimization algorithm can be used to trace the regularization path of (group) KAN estimates for a range of shrinkage tuning parameter *s* and within group strength factor $1 \le kg \le pg1 \le kg \le pg$. There are two types of parameters involved in model (11): *s* is to shrink all coefficients individually, while kgkg is to control the group strength within group *g* for $1 \le g \le G1 \le g \le G$. A well chosen kgkg is expected to control the group shrinkage strength of group *g* appropriately. To avoid the expensive computation of choosing all different combination of kgkg's, we let $kg = r \cdot pgkg = r \cdot pg$ for $1/pg \le r \le 11/pg \le r \le 1$. Thus, we only need to select an optimal (*r*, *s*) using existing model selection criteria.

For a given (*s*, *r*), AIC, BIC and EBIC are defined as,

$$\begin{aligned}
\mathrm{AIC}(s,r) &= \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(s,r)\|_2^2/\sigma^2 + 2\widehat{\mathrm{df}} \\
\mathrm{BIC}(s,r) &= \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(s,r)\|_2^2/\sigma^2 + \widehat{\mathrm{df}}\log n \\
\mathrm{EBIC}(s,r) &= \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(s,r)\|_2^2/\sigma^2 + \widehat{\mathrm{df}}\log(n) + 2\gamma\widehat{\mathrm{df}}\log(p)
\end{aligned} \tag{16}$$

where $\widehat{\mathrm{df}}$ is given in (14) (or (13) for G=1), $\sigma\sigma$ is the standard deviation of any y_i estimated from the data, and γ is a user-specified parameter. In our numerical studies, we choose the optimal s^o and r^o using BIC (when n>p) or EBIC (when n<p) with $\gamma=1$ instead of AIC since AIC turns to select an over-fitted model.

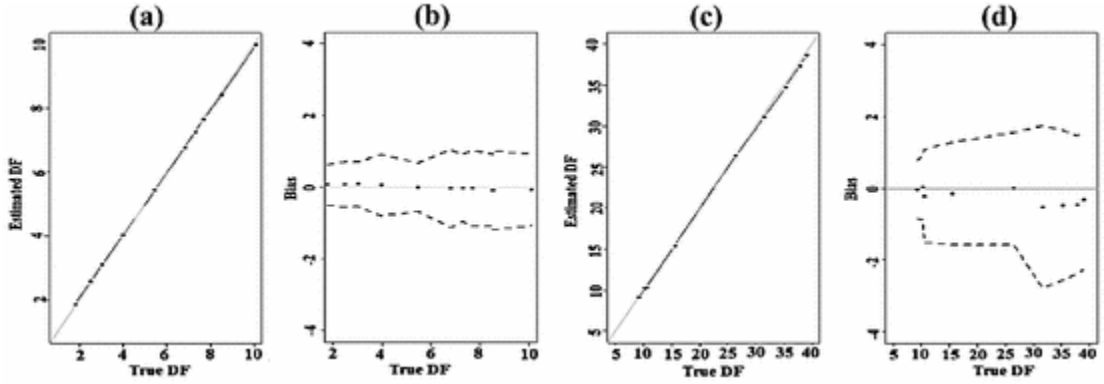

Fig. 4 The approximation of the true degrees of freedom and its estimation in (13) and (14) of a hypothetical model from Monto carlo simulations. **a** Compare $\widehat{\mathrm{df}}$ of KAN with the true degrees of freedom using the diabetes data, where *k* in (7) is uniformly generated from {1,…,p}; **b** the estimation bias from (**a**) and its point-wise 95 % confidence intervals are indicated by dashed lines; **c** compare $\widehat{\mathrm{df}}$ of the GKAN with the true degrees of freedome using a simulated data from Example 3 in Sect. 5.2, where k_g=4 in (11) is fixed for each group; **d** the estimation bias from **c** and its point-wise 95 % confidence intervals are indicated by *dashed lines*

A Matlab code for implementing the GKAN method can be also found from the author's website at https://www.sites.google.com/a/uncg.edu/xiaoli-gao/home/r-code.

5 Experiment

In this section, we use extensive simulation studies to illustrate the effect of KAN and GKAN estimates under different grouping structures. We first use some simulation examples to demonstrate the natural grouping effect of KAN estimates for different *k* in (7) without any prior group information. Then we compare the GKAN with GLASSO under some prior group information, where some covariates may be mis-classified. Finally we demonstrate the performance of GKAN and KAN under both overlapping and hierarchical group structures.

In each setting, 500 data sets are generated from the linear regression model (1), where ε_i's are independent and identically generated from the normal distribution $N(0,\sigma^2)$.

5.1 KAN versus LASSO and iLASSO

We first compare KAN with LASSO and iLASSO using two examples with the existence of both within group sparsity and between group sparsity among predictors. All predictors in X are generated from multivariate normal distribution with mean 0 and covariance matrix Σ, where Σ is chosen such that the within-group correlation is *a* and between group correlation is a·b, where a=0.8 or 0.5 and b=0.8 or 0.5. The random error has $\sigma^2=0.01\beta^T_0\Sigma\beta_0$.

Example 1

The sample size n=100, p=16 with G=4 equal-sized groups. The true model includes within group sparsity in one of two relevant groups,

$$\boldsymbol{\beta}_1^{(0)} = (1, 0.5, 0, 0)^T,\ \boldsymbol{\beta}_2^{(0)} = (1, -0.9, -1.3, -0.5)^T,\ \boldsymbol{\beta}_3^{(0)} = \boldsymbol{\beta}_4^{(0)} = (\underbrace{0, \ldots, 0}_{4})^T.$$

Example 2

The sample size n=100n=100 and p=40p=40 with G=5G=5 equal-sized groups. All relevant groups include within group sparsity and some small effects also exist in the true model,

$$\boldsymbol{\beta}_1^{(0)} = (1, 0.5, 0, -0.8, \underbrace{0, \ldots, 0}_{4})^T,\ \boldsymbol{\beta}_2^{(0)} = (1, -0.9, -1.3, -0.5, \underbrace{0, \ldots, 0}_{4})^T,$$

$$\boldsymbol{\beta}_3^{(0)} = (0, 0.1, -0.1, \underbrace{0, \ldots, 0}_{5})^T,\ \boldsymbol{\beta}_4^{(0)} = \boldsymbol{\beta}_5^{(0)} = (\underbrace{0, \ldots, 0}_{4})^T.$$

To evaluate the variable selection performance of each method, we compute the correctly fitted ratio (CFR), the over-fitted ratio (OFR) from all 500 iterations. We also evaluate the estimation performance using the relative error (RE),

$$\left\|\hat{\beta} - \beta_0\right\|_2 / \|\beta_0\|_2. \quad (17)$$

Simulation results from Example 1 and 2 are reported in Table 1 and 2, respectively. Among all scenarios, when *k* is close to 1 (k ≤ 6 for Example 1 and k < 10 for Example 2), KAN produces a over-fitted model almost surely because it enforces a very strong group strength among all covariates. It is consistent with the feature (ii) in Theorem 1 since q = 6 in Example 1 and q = 9 in Example 2. When *k* differs from 1 sufficiently, KAN performs better until the best performance is achieved (k=8 for Example 1 and k=13 for Example 2). Then KAN performs worse when *k* continues to grow. This is because KAN begins to lose strength for group shrinkage among all covariates when *k* is closer to *p*. Thus, without using any prior group information, the KAN family has the potential ability to encourage grouped variable selection even though both within and between group sparsity exist. We also observe similar patterns under mild correlation coefficients (a=0.2 or b=0.2, not reported).

Table 1. Results for Example 1

	(*a*, *b*)								
	(0.5, 0.5)			**(0.5, 0.8)**			**(0.8, 0.5)**		
	CFR (%)	**OFR (%)**	**RE (**103103**)**	**CFR (%)**	**OFR (%)**	**RE (**103103**)**	**CFR (%)**	**OFR (%)**	**RE (**103103**)**
iLASSO	0	100	3.99	0	100	8.18	0	100	7.41
k=2	0	100	3.99	0	100	8.18	0	100	7.41
k=3	0	100	3.98	0	100	8.13	0	100	7.41
k=4	0	100	3.95	0	100	8.19	0	100	7.44
k=5	0	100	4.04	0	100	8.16	0	100	7.40
k=6	0	100	4.06	0	100	8.24	0	100	7.44
k=7	98	2	1.87	96	4	4.92	98	2	6.30
k=8	98	2	1.25	100	0	2.67	98	2	3.13
k=9	96	4	1.20	96	4	2.15	90	10	2.53
k=10	94	6	1.23	92	8	2.18	84	16	2.34
k=11	90	10	1.26	86	14	2.25	84	16	2.34
k=12	88	12	1.28	86	14	2.25	84	16	2.34
k=13	88	12	1.28	86	14	2.25	84	16	2.34
k=14	88	12	1.28	86	14	2.25	82	18	2.37

k=15	88	12	1.28	86	14	2.25	82	18	2.37
LASSO	88	12	1.28	86	14	2.25	82	18	2.37

CFR is the ratio of selecting the exact model

OFR is the over-fitted ratio (the true model plus at least one additional noisy predictors)

RE (103103) is 103×103× the average of relative error computed in (17)

Table 2. Results for Example 2

	(*a*, *b*)								
	(0.5, 0.5)			**(0.5, 0.8)**			**(0.8, 0.5)**		
	CFR (%)	**OFR (%)**	**RE (**103103**)**	**CFR (%)**	**OFR (%)**	**RE (**103103**)**	**CFR (%)**	**OFR (%)**	**RE (**103103**)**
iLASSO	0	100	14.3	0	100	34.1	0	100	24.1
k=4	0	100	14.4	0	100	34.1	0	100	23.7
k=7	0	100	14.1	0	100	33.7	0	100	23.8
k=10	28	70	16.9	0	100	40.7	2	96	24.7
k=13	100	0	5.11	72	10	27.1	84	10	14.6
k=16	78	22	2.80	54	42	19.1	62	36	7.04
k=19	74	26	2.58	34	64	8.55	56	44	5.15
k=22	70	30	2.67	28	72	7.17	54	46	4.91
k=25	70	30	2.67	28	72	7.12	54	46	4.91
k=28	70	30	2.67	28	72	7.12	54	46	4.91
k=31	70	30	2.67	28	72	7.12	54	46	4.91
k=34	70	30	2.67	28	72	7.12	54	46	4.91
k=37	70	30	2.67	28	72	7.12	54	46	4.91
LASSO	70	30	2.67	28	72	7.12	54	46	4.91

CFR is the ratio of selecting the exact model

OFR is the over-fitted ratio

RE (103103) is 103×103× the average of relative error computed as (17)

5.2 GKAN versus GLASSO and iCAP

We now demonstrate the advantage of the GKAN over GLASSO and iCAP under two different types of fussy group information: p<np<n with equal-sized groups in Example 3 and p>np>n with non-equal sized groups in Example 4.

Example 3

Let n=100 and p=48 with G=8 equal-sized groups. The design matrix Σ is generated similar to Sect. 5.1 with a=0.5 and b=0.2. The true model includes two relevant groups without within-group sparsity,

$\beta_{0,1}=(1,1.2,-2.0,3.0,0.8,-1.4)^T$, $\beta_{0,3}=(-0.5,1.3,1.5,2.6,-1.6,2.7)^T$.

Example 4

Let n=100 and p=500 with G=35 groups. We consider different group sizes such that p_g=10, 15 and 20 for $1 \le g \le 10$, $11 \le g \le 30$ and $30 \le g \le 35$, respectively. The correlation matrix is block-wise diagonal such that the between group correlation is zero. The correlation sub-matrix within each group is power decay such that the correlation coefficient between $x^{(g)}_{j1}$ and $x^{(g)}_{j2}$ are $\rho(x^{(g)}_{j1},x^{(g)}_{j2})=0.5^{|j1-j2|}$ for $1 \le j1, j2 \le pg$ in group *g*. In the true model,

$$|\beta_{0j}| = \begin{cases} 0.8(1+0.9^{j-1}) & \text{if} \quad j = 1,\ldots,10 \\ 1+0.9^{j-101} & \text{if} \quad j = 101,\ldots,115 \\ 1.5(1+0.9^{j-401}) & \text{if} \quad j = 401,\ldots,420 \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$$

and then β_{0j} is randomly chosen to be either positive or negative.

We compare GKAN with iCAP and GLASSO under 3 different types input group information: (1) True group arrangements; (2) Incorrect grouping information obtained from the PAM algorithm. In particular, all *p* variables are grouped into 0.5*G* or 1.5*G* groups, where *G* is the number of the groups in the true model; (3) Manually controlled groups with 100(2r)% incorrect group assignments, where r=0.1,0.2 or 0.3, respectively. In particular, we assign 100r% variables from nonzero groups incorrectly into zero groups, and 100r% variables from zero groups incorrectly into nonzero groups.

To evaluate the performance of GKAN in terms of variable selection, besides CFR and OFR, we also report the under-fitting ratio (UFR) out of 500 simulations. Variable selection results from GKAN, iCAP, GLASSO and KAN (no grouping information is used) for both Examples 3 and 4 are reported in Table 3.

If correct group information is used as the above (1), all three methods incorporating prior group information (GLASSO, iCAP, GKAN) perform very well. When incorrect group assignment is adopted under the above (2) and (3), both GLASSO and iCAP almost lose their validity completely. However, GKAN is resistent to the incorrect group assignment especially when p<np<n in Example 3. When p>np>n, GKAN still shows certain abilities in variable selection as long as the majority of covariates stays within the right groups. In general, the validity of GLASSO is more severely affected by the incorrect grouping information than iCAP since the GLASSO tends to select incorrect models, while iCAP tends to choose larger models in most cases.

Table 3. Variable selection results for Example 3 and 4

Group	**GLASSO**	**iCAP**	**GKAN**	**KAN**
	CFR (OFR, UFR)	**CFR (OFR, UFR)**	**CFR (OFR, UFR)**	**CFR (OFR, UFR)**
Example 3				
True	100 % (0 %, 0 %)	94 % (6 %, 0 %)	100 % (0 %, 0 %)	82 % (14 %, 4 %)
$0.5G^0$	0 % (100 %, 0 %)	0 % (90 %, 0 %)	92 % (4 %, 4 %)	82 % (14 %, 4 %)
$1.5G^0$	0 % (40 %, 50 %)	0 % (100 %, 0 %)	84 % (8 %, 8 %)	82 % (14 %, 4 %)

0.1	0 % (0 %, 0 %)	0 % (92 %, 0 %)	58 % (12 %, 26 %)	82 % (14 %, 4 %)
0.2	0 % (98 %, 0 %)	0 % (100 %, 0 %)	68 % (12 %, 20 %)	82 % (14 %, 4 %)
0.3	0 % (100 %, 0 %)	0 % (100 %, 0 %)	64 % (14 %, 22 %)	82 % (14 %, 4 %)
Example 4				
True	84 % (8 %, 8 %)	100 % (0 %, 0 %)	96 % (4 %, 0 %)	82 % (0 %, 18 %)
$0.5G^0$	0 % (0 %, 0 %)	0 % (4 %, 0 %)	14 % (44 %, 16 %)	82 % (0 %, 18 %)
$1.5G^0$	0 % (0 %, 0 %)	0 % (86 %, 0 %)	40 % (38 %, 3 %)	82 % (0 %, 18 %)
0.1	0 % (0 %, 0 %)	0 % (52 %, 0 %)	78 % (18 %, 0 %)	82 % (0 %, 18 %)
0.2	0 % (0 %, 0 %)	0 % (70 %, 0 %)	42 % (38 %, 0 %)	82 % (0 %, 18 %)
0.3	0 % (0 %, 0 %)	0 % (88 %, 0 %)	64 % (68 %, 0 %)	82 % (0 %, 18 %)

CFR, OFR or CFR: correctly fitted model ratio, the over-fitted ratio, or the under-fitted ratio

TRUE, $0.5G^0$ or $1.5G^0$: true group information, incorrect group information with $G=0.5G^0$ or $1.5G^0$

0.1/0.2/0.3: 10/20/30%30% nonzeros (zeros) are incorrectly assigned into other zero (nonzero) groups

When comparing GKAN and KAN, we find that GKAN outperforms the KAN and benefits from correct or mostly correct grouping information. However, if a large number of important (unimportant) variables are incorrectly assigned into some non-important (important) groups, GKAN is affected and may perform worse than KAN. Overall, those simulation studies indicate that GKAN is still preferred if the majority of covariates are assigned with the correct group information. Otherwise, KAN is preferred as a more conservative group variable selection approach.

We also compute and compare REs in (17) from all four method for both Example 3 and 4. In most cases, GKAN and KAN outperform GLASSO and iCAP by producing significantly smaller estimation biases if incorrect group information is adopted. If correct group information is adopted, their estimation performances are comparable. See panel (a) of Fig. 5 for some partial boxplot output from Example 4. Results from iCAP shows a similar pattern to that under GLASSO and are omitted.

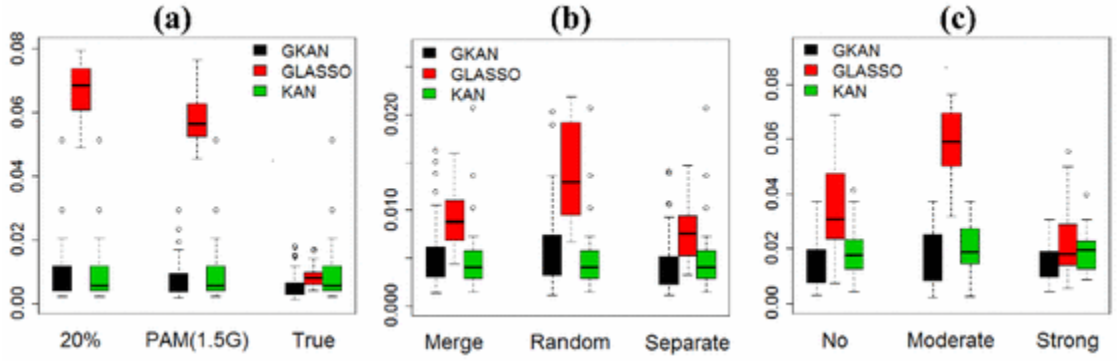

Fig. 5. Selected boxplots of REs in (17). **a** Example 3; **b** Example 5; **c** Example 6

5.3 Overlapping and hierarchical group structures

Example 5

(Overlapping group structure) Example 3 is modified such that group 1 and 2 are overlapped with j=7,8,9, and group 2 and 3 are overlapped with j=10,11,12. Thus 50 % elements from each of two nonzeros groups are overlapped with other zero groups.

As introduced in the last paragraph in Sect. 2.2, we run group variable selection in Example 5 by providing three different types of group information: (a) merge overlapped groups into one group; (b) separate two overlapped groups into three separate groups; or (c) randomly assign the overlapped set into either one of two overlapped groups.

The variable selection of Example 5 is reported in Table 4. It is observed that all four methods perform reasonably well if we separate overlapped groups. Among all four methods, GKAN still performs the best. If we merge or randomly assign overlapped groups, both GLASSO and iCAP lose their variable selection abilities completely. However, KAN still keeps its variable selection ability under all scenarios. GKAN performs even better than KAN under the Merge scenario, but it becomes worse than the KAN if those overlapped variables are randomly assigned. This suggests that GKAN is a preferred group variable selection tool when overlapping groups exist as long as those overlapped variables are not clustered with only one of two overlapped groups. The advantage of GKAN over the other methods on coefficients estimation for Example 5 can be observed from those boxplots in panel (b) in Fig. 5.

Example 6

(Hierarchical group structure) Each $x_i=[x_{iG1},x_{iG2},x_{iG3},x_{iG4}]$ for $1 \le i \le 121$, where $x_{iG1}=[z_{i1},z_{i2},z_{i3},z_{i1}z_{i2},z_{i1}z_{i3},z_{i2}z_{i3},z_{i1}z_{i2}z_{i3}]$, $x_{iG2}=[z_{i4},z_{i5},z_{i6},z_{i4}z_{i5},z_{i4}z_{i6},z_{i5}z_{i6}]$, $x_{iG3}=[z_{i7},z_{i8},z_{i7}z_{i8}]$, and $x_{iG4}=[z_{i9},z_{i10},z_{i9}z_{i10}]$. Here all z_{ij}s are identically and independently generated from $N(0, 1)$.

Let βj be the coefficients corresponding to the *j*th main effect (z_{ij}) and $\beta_{j,k}$ be the coefficients corresponding to the interaction term between the *j*th and *k*th main effect (zj×zk) and $\beta_{1,2,3}$ be the coefficients of $z_{i1}\times z_{i2}\times z_{i3}$. We set the first five main effect coefficients to be nonzero and the remaining to to be zero, that is, $\beta_1=7$, $\beta_2=2$, $\beta_3=1$, $\beta_4=2$, $\beta_5=1$ and $\beta_j=0$ for $6 \le j \le 10$. We consider three different types of hierarchical effects: (a) No interaction, $\beta_{j,k}=0$ for all *j* and *k*; (b) Moderate interaction, $\beta_{1,2}=2$, $\beta_{2,3}=2$, $\beta_{4,5}=1$; and (c) Strong interaction, $\beta_{1,2}=3$, $\beta_{1,3}=3$, $\beta_{2,3}=2$, $\beta_{1,2,3}=1$ and $\beta_{4,6}=1$. We choose $\sigma=3.7$

Table 4. Variable selection results for Example 5 and 6

Group	**GLASSO**	**iCAP**	**GKAN**	**KAN**
	CFR (OFR, UFR)	**CFR (OFR, UFR)**	**CFR (OFR, UFR)**	**CFR (OFR, UFR)**
Example 5				
Merge	0 % (100 %, 0 %)	0 % (100 %, 0 %)	82 % (10 %, 8 %)	82 % (8 %, 10 %)
Separate	88 % (12 %, 0 %)	92 % (8 %, 0 %)	94 % (4 %, 2 %)	82 % (10 %, 8 %)
Random	0 % (100 %, 0 %)	0 % (100 %, 0 %)	38 % (52 %, 8 %)	76 % (10 %, 8 %)
Example 6				
No	0 % (80 %, 0 %)	0 % (100 %, 0 %)	82 % (18 %, 1 %)	76 % (24 %, 0 %)
Moderate	0 % (98 %, 0 %)	0 % (100 %, 0 %)	80 % (6 %, 12 %)	74 % (2 %, 12 %)

Strong	0 % (98 %, 2 %)	0 % (100 %, 0 %)	56 % (4 %, 48 %)	34 % (4 %, 62 %)

The variable selection and estimation results of Example 6 are reported in Table 4 and panel (c) of Fig. 5, respectively. Not surprisingly, both GLASSO and iCAP over-select variables under hierarchical group structures, while GKAN and KAN still perform reasonably well, especially when the interaction effects are not too strong. It is also observed that GKAN performs better than KAN when hierarchical effect exists.

Table 5. Variable selection and coefficients estimation results for the real data example

Probes	G=4		G=9		G=100	
	GKAN	**GLASSO**	**GKAN**	**GLASSO**	**KAN**	**LASSO**
1368923_at	−0.004		−0.001	–	–	–
1369353_at	−0.016		−0.006	–	−0.148	–
1370222_at	−0.008	–	–	–	–	–
1370429_at	−0.005	–	−0.002	–	–	−0.073
1370551_a_at	−0.006	–	–	–	–	–
1370655_a_at	−0.004	–	−0.001	–	–	–
1371242_at	−0.013	–	−0.005	–	–	–
1375825_at	−0.014	–	−0.005	–	–	–
1376559_at	−0.010	–	−0.004	–	–	–
1385704_at	−0.007	–	–	–	–	–
1391303_at	−0.004	–	−0.001	–	–	–
1393979_at	−0.009	–	−0.003	–	–	–
1378425_at	−0.015	–	−0.005	–	–	–
1379971_at	0.246	–	0.247	0.092	0.312	0.014
1381744_at	−0.006	–	−0.002	–	–	–
1381787_at	−0.015	–	−0.005	–	–	–
1383749_at	−0.004	–	–	–	–	–
1386440_a_at	−0.009	–	−0.003	–	–	–
1390856_at	0.153	–	0.209	–	0.142	–
1391132_at	−0.005	–	−0.002	–	–	–
1391571_at	0.033	–	0.042	–	0.071	–
1391909_at	−0.014	–	−0.005	–	–	–
1398171_at	−0.011	–	−0.004	–	–	–
1374106_at	–	–	–	0.067	–	0.032
1379605_at	–	–	–	0.022	–	–
1389584_at	–	–	–	–	–	0.098
1383110_at	–	–	–	–	–	0.125
1383673_at	–	–	–	–	–	0.048
1386683_at	–	–	–	–	–	0.039
1385334_at	–	–	–	–	0.025	–

6 Real data example

In this section, we apply the group KAN to a real gene expression data set reported in Scheetz et al. (2006) and also analyzed by Huang et al. (2008). In this dataset, 120 twelve-week-old male offsprings of F1 animals were selected for tissue harvesting from the eyes for microarray analysis. The microarrays used to analyze the RNA from the eyes of these F2 animals contain over 31, 042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). For each probe set, gene expression intensity values were normalized using the robust multi-chip averaging method. Huang et al. (2008) studied a total of 18,976 probes including gene TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang et al. 2006), a genetically heterogeneous disease of multiple organ systems including the retina.

Our target is to find important probes that are most related to TRIM32 (Probe ID: 1389163_at) using a linear regression model. To improve the performance of the GKAN analysis, the data set is pre-processed as follows. First, all variables are standardized to have mean zero and standard deviation 1. Second, we keep 3,000 probes with the largest variances and select 100 from them with highest correlation with TRIM32. Finally, all those 100 probes are clustered into G groups using the PAM method if GKAN is used. The final data set consists of n=120 samples and p=100 probes with 5 different group assignments with G=4,5,7,9 or 100, respectively.

Both GKAN and GLASSO are applied based upon 3 different settings of grouping assignments. When G=100, the GKAN and GLASSO are reduced into KAN and LASSO, respectively. All tuning parameters are selected using EBIC introduced in Sect. 4 with γ=1. Similar estimation and variable selection results are found $4 \leq G \leq 7$. Due to the space limits, we only list results for G=4, 9 and 100 in Table 5. It is observed that GKAN shows more consistent variable selection results under different group arrangements. Important probes sets selected by GKAN turn to be nested under different group sizes. In particular, 4 probes (1369353_at, 1379971_at, 1390856_at and 1391571_at) are selected by both KAN and GKAN under all settings.

However, variable selection results from GLASSO exhibit more discrepancies under different group assignments. For example, no variable is detected using GLASSO for G=4.

We also compute the prediction errors from each method using cross validation following 300 random partitions of the data set. In each partition, the training set consists of 2/3 observations and the test set consists of the remaining 1/3 observations. The prediction errors for GKAN and GLASSO under G=4,5,7,9 and 100 are plotted in Fig. 6.

In all settings, GKAN produces relatively smaller and more robust predict errors than GLASSO.

7 Discussion

In this paper, we have proposed a KAN method to encourage natural group variable selection in a high-dimensional sparse model. When the covariates have complex group structures and true group information is unknown, the KAN family without any prior group information performs well in terms of the complex structured sparsity, including both the bi-level sparsity (both within group sparsity and between group sparsity) and hierarchical sparsity. It is because the KAN family relaxes both the $\ell 1 \ell 1$ norm from the LASSO and the $\ell \infty \ell \infty$ norm from the iLASSO to

the *k*-th largest norm and select *k* data-adaptively. We have provided a theoretical justification on the optimal KAN coefficients *k* and the number of nonzero coefficients *q* in the true model. Such a relationship between an optimal *k* and *q* help us to interpret the data better.

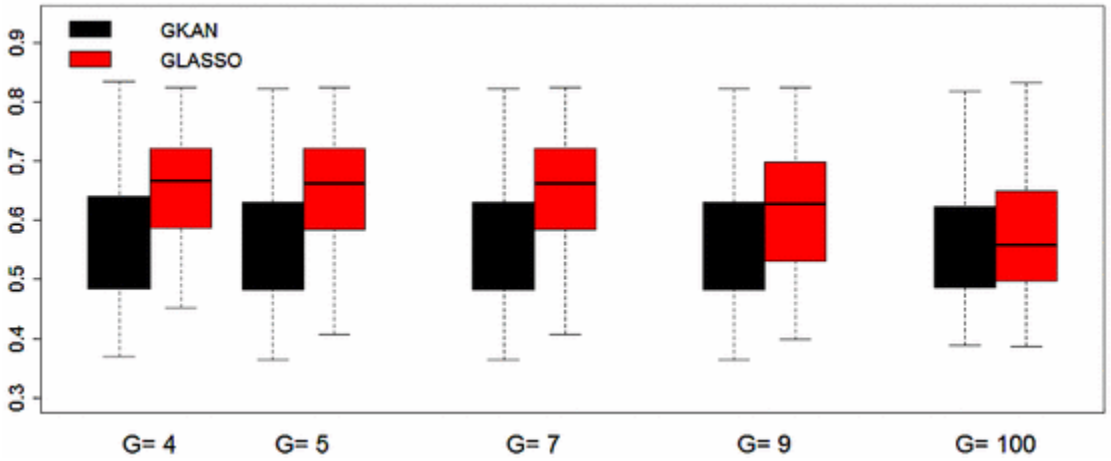

Fig. 6. Boxplots of prediction errors from 300 random partitions in real data analysis. Results for GKAN and GLASSO under G=100G=100 are actually KAN and LASSO analysis results

We also extend KAN to GKAN by taking into account some prior group information. When the majority of the covariates are assigned into true groups, GKAN can not only benefit from limited true group information but also be robust some incorrect group information. Unlike the group LASSO and iCAP, GKAN controls the within group strength adaptively from the data. This explains why GKAN has some robust properties when some covariates' group information are mis-specified.

We have also proposed an unbiased estimator of the degrees of freedom of GKAN estimates. For a KAN estimate, the number of unique nonzero estimates of is an "almost" unbiased estimator of df (with difference by "−1"). Such an unbiased estimator is useful in tuning parameters selection. Without using the cross validation, existing model selection criterion can be adopted to choose tuning parameters.

KAN estimates are very important in high-dimensional data analysis when the prior group information is unknown. It merits further investigation in many aspects. For example, an efficient computation is needed for dealing with extra high-dimensional data. In this paper, we only use the existing optimization technique, which limits the application of the study to a ultra-high dimensional data set. Another important issue is the choice of group strength factors $k_g, 1 \leq g \leq G$ for GKAN. We suggest choosing $k_g = r \cdot G$ from the data stochastically using some model selection criteria. Furthermore, we only consider KAN and GKAN for the linear regression model and the quadratic loss function. We are currently working on an extension for logistic regression also other models. Finally, the asymptotic properties of the GKLAN under $p>n$ also deserves further investigation.

Notes

Acknowledgments

The author wants to thank Sijian Wang and Yuan Wu for their valuable comments and Jonathan Rowell for his professional proofreading. She also would like to thank the reviewers for their helpful and constructive comments for improvement of the manuscript. The author gratefully acknowledges *Simons Foundation (#359337)* and *UNC Greensboro* (New Faculty Grant) for their support in this Project.

Appendix

Proof of Lemma 1

Lemma 1 is true for k=1 and *p*. We only need to verify the convexity for 1 < k < p.
Take p=3 and k=2 as an example. Suppose both $\alpha=(\alpha 1,\ldots,\alpha p)'$ and $\gamma=(\gamma 1,\ldots,\gamma p)'$ are vectors in K. We need to show that $\theta=c\alpha+(1-c)\gamma\in K$ for $0<c<1$. Without loss of generality, we assume $\alpha_{i1}\geq\alpha_{i2}\geq\alpha_{i3}$ and $\gamma_{j1}\geq\gamma_{j2}\geq\gamma_{j3}$, where $1\leq i_1\neq i_2\neq i_3\leq 3$ and $1\leq j_1\neq j_2\neq j_3\leq 3$. Then

$|\alpha i'1|+|\alpha i'2|\leq|\alpha i1|+|\alpha i2|<s,\ 1\leq i'1\neq i'2\leq 3$

and

$|\gamma j'1|+|\gamma j'2|\leq|\gamma j1|+|\gamma j2|<s, 1\leq j'1\neq j'2\leq 3.$

Suppose $\theta k1\geq\theta k2\geq\theta k3$, where $1\leq k1\neq k2\neq k3\leq 3$. Then

$$\begin{aligned}|\theta_{k_1}|+|\theta_{k_2}| &\leq c|\alpha_{k_1}|+(1-c)|\gamma_{k_1}|+c|\alpha_{k_2}|+(1-c)|\gamma_{k_2}|\\ &= c(|\alpha_{k_1}|+|\alpha_{k_2}|)+(1-c)(|\gamma_{k_1}|+|\gamma_{k_2}|)\\ &\leq c(|\alpha_1|+|\alpha_2|)+(1-c)(|\gamma_{j_1}|+|\gamma_{j_2}|)\leq s.\end{aligned}$$

Thus $\theta\in K$. The proof can be extended to p>3 and 1< k < p without any extra work. □

Proof of Theorem 1

Let

$Q(\beta)=\|y-X\beta\|_2^2+\lambda\|\beta\|_{(k)}$.

Due to the convex property of Q(β), from the standard results in Anderson and Gill (1982); Pollard (1991), we can verify that $\hat{\beta}-\beta_0=o_P(1)$ when *p* is fixed and $\lambda/n=o(1)$. The proof is the same as one for Theorem 1 in Knight and Fu (2000) and then omitted.

Let $\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2)$ satisfying $\|\hat{\beta}_1-\beta_{10}\|2=O_P(n^{-1/2})$ and $\|\hat{\beta}_2-\beta_{20}\|_2\leq cn^{-1/2}$ for some c>0. We want to verify when $n\to\infty$ whether with probability tending to 1

$$\frac{\partial Q(\beta)}{\partial\beta_j}\Big|_{\beta=\hat{\beta}}\begin{cases}>0 & for\quad 0<\beta_j<cn^{-1/2}\\ <0 & for\quad -cn^{-1/2}<\beta_j<0.\end{cases}\tag{18}$$

If (18) holds for some j=q+1,…,p, then $\hat{\beta}$ may include zero elements. Otherwise, $\hat{\beta}$ does not have any zero element, or all of them are zeros. By abuse of notation, we denote the *j*th column vector of X as xj. Then $x_j'\varepsilon/\sqrt{n}$ is sub-Gaussian, and therefore $|x_j'\varepsilon|=O_P(\sqrt{n})$. Then from the consistency of $\hat{\beta}$ and regularity conditions (9) and (10),

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} &= -\mathbf{x}_j'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})+\lambda\frac{\partial\|\boldsymbol{\beta}\|_{(k)}}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\\ &= -\mathbf{x}_j'\boldsymbol{\varepsilon}+\mathbf{x}_j'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\lambda\frac{\partial\|\boldsymbol{\beta}\|_{(k)}}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\\ &= \sqrt{n}\left(O_P(1)+(\lambda/\sqrt{n})\frac{\partial\|\boldsymbol{\beta}\|_{(k)}}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right),\end{aligned} \tag{19}$$

where

$$\frac{\partial(\|\boldsymbol{\beta}\|_{(k)})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}=\begin{cases}\operatorname{sgn}(\widehat{\beta}_j) & \text{if } |\widehat{\beta}_j|\geq|\widehat{\beta}_{(k)}|\\ 0 & \text{if } |\widehat{\beta}_j|<|\widehat{\beta}_{(k)}|.\end{cases} \tag{20}$$

From (19), (20) and λ/sqrtn→∞, the sign of the derivative in (18) is completely determined by that of $\hat{\beta}_j$ for $|\hat{\beta}^j|\geq|\hat{\beta}_{(k)}|$ and j = q + 1,…,p. (ii) holds since $|\hat{\beta}_j|<|\hat{\beta}_{(k)}|$ for j > q when k ≤ q . In addition, there exists j > q such that $|\hat{\beta}_j|>|\hat{\beta}_{(k)}|$ when k > q, for j > q. Thus (i) holds. □

Kato (2009) provided an unbiased estimator of the degrees of freedom defined in (13) of shrinkage estimators as in Lemma 3.

Lemma 3

Suppose K is a closed convex set with boundary ∂K with a disjoint partition $\partial K=D_1\cup\cdots\cup D_p$. If *f* is a Lipschitz function, let $\hat{\beta}_K=f(\hat{\beta}^0)$ be a shrinkage modeling procedure by shrinking $\hat{\beta}^0$ in terms of the the constraints K, or the orthogonal projection of $\hat{\beta}^0$ onto K such that $\|\hat{\beta}^0-\hat{\beta}_K\|_2^2=\min_{b\in K}\|\hat{\beta}^0-b\|_2^2$. Denote the normal cone of K at b as $N(K,b)=\{\beta-b:\beta_K=b\}$ for any $b\in\partial K$ and $D_m=\{b\in\partial K:\dim N(K,b)=m\}$. Denote $E_m=\{\hat{\beta}^0\in R^p\backslash K:\hat{\beta}_K\in D_m\}$ and E_m^0 as an interior point of E_m. If D_m is a (p−m)-dimensional C^2-manifold consisting of a finite number of relatively open connected components and $E_m\backslash E_m^0$ has zero lebesgue measure, then an unbiased estimator of the degrees of freedom of the shrinkage modeling procedure $\hat{\beta}_k$ is

$$\widehat{\mathrm{df}}(\hat{\beta}k)=\sum_{m=0}^{p}(p-m)I(\hat{\beta}^0\in E_m).$$

Proof of Theorem 2

Let K be the convex set in Lemma 1. Consider that k items among β_j's are involved in $\ell_{(k)}$ norm and β_j's have different possible signs. We can write K alternatively as, $\{\beta \in R^p: \alpha'_i\beta \le t, i=1,\ldots,C_{k,p}2^k\}$, where $\alpha_i=(\alpha_{1i},\ldots,\alpha_{pi})$ with a_{ji} being 1, −1, or 0 and $Ck,p=\binom{p}{k}$. Take p=3 and k=2 as an example. We have $C_{k,p}2^k=12$ and $\alpha_1=(1,1,0)$, $\alpha_2=(1,-1,0)$, $\alpha_3=(-1,-1,0)$, $\alpha_4=(-1,1,0)$, $\alpha_5=(1,0,1)$, $\alpha_6=(1,0,-1)$, $\alpha_7=(-1,0,-1)$, $\alpha_8=(-1,0,1)$, $\alpha_9=(0,1,1)$, $\alpha_{10}=(0,1,-1)$, $\alpha_{11}=(0,-1,-1)$, $\alpha_{12}=(0,-1,1)$. The open face of K_1 is of the form

$$\{\beta \in R^p: \alpha'i\beta = s, i \in L, \alpha_j\beta < s, j \in \{1,\ldots,C_{k,p}2^k\} \backslash L\}, \quad (21)$$

where L is a subset of $\{1,\ldots,C_{k,p}2^k\}$. Suppose a non-empty open face F of K is given by (21). Then the dimension of F is the p-rank(A), where $A=(\alpha i, i \in L)$.

Notice that if β_j changes from nonzero to zero, rank(A) increases by 1 since α_{ij} changing from 0 to 1 or −1 do not update the value of $\alpha'_i\beta$. Furthermore, if β_j changes from the (k−1)-th largest one to the same as the k-th largest, such that $\beta_j=\beta_{(k)}$, then then rank(A) also increases by 1 since $\alpha'_i\beta$ keeps the same if we replace $\alpha_{ij}=0$ by 1 or −1. Thus, $\text{rank}(A)=\#\{1\le j\le p: \beta_j=0\}+\#\{1\le j\le p: \beta_j=\beta(k)=s\}$, Following Lemma 1 and 3, we have

$$\widehat{df} = \begin{cases} \#\{1 \le j \le p: \widehat{\beta}_j \ne 0\} - \#\{\|\widehat{\boldsymbol{\beta}}\|_{(k)} = s\} = |\mathcal{A}| - |\mathcal{U}| & \text{if } \widehat{\boldsymbol{\beta}}^0 \notin \mathcal{K} \\ p & \text{if } \widehat{\boldsymbol{\beta}}^0 \in \mathcal{K} \end{cases}$$

Thus, results in Theorem 2 is proved under p<n. If p>n, then the LS estimation is not uniquely defined.

References

1. Akaike H (1973) Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60:255–265
2. Anderson PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120
3. Bogdan M, van den Berg E, Chiara S, Su W, Candes E (2015) SLOPE-adaptive variable selection via convex optimization. Ann ApplStat 9(3):1103–1140
4. Chen J, Chen Z (2008) Extended bayesian information criteria for model selection with large model spaces. Biometrika 95(3):759–771
5. Chiang AP, Beck JS, Yen HJ, Tayeh MK, Scheetz TE, Swiderski R, Nishimura D, Braun TA, Kim K, Huang J, Elbedour K, Carmi R, Slusarski DC, Casavant TL, Stone EM, Sheffield VC (2006) Homozygosity mapping with SNP arrays identifies a novel gene for Bardet-Biedl syndrome (BBS10). Proc Natl Acad Sci 103:6287–6292

6. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32:407–499

7. Frank I, Friedman J (1993) A statistical view of some chemometrics regression tools (with discussion). Technometrics 35:109–148

8. Huang J, Ma SG, Zhang C (2008) Adaptive lasso for sparse high-dimensional regression models. Stat Sin 18:1603–1618

9. Kanehisa M, Goto S (2000) Kyoto encyclopedia of genes and genomes. Nucl Acids Res 28:27–30

10. Kato K (2009) On the degrees of freedom in shrinkage estimation. J Multivar Anal 100:1338–1352

11. Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

12. Kim Y, Kim J, Kim Y (2006) Blockwise sparse regression. Stat Sin 16:375–390

13. Knight K, Fu W (2000) Asymptotics for lasso-type estimators. Ann Stat 28:1356–1378

14. Pollard D (1991) Asymptotics for least absolute deviation regression estimators. Econ Theory 7:186–199

15. Scheetz TE, Kim KYA, Swiderski RE, Philp AR, Braun TA, Knudtson KL, Dorrance AM, DiBona GF, Huang J, Casavant TL, Sheffield VC, Stone EM (2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proc Natl Acad Sci 103(39):14429–14434

16. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

17. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288

18. Wang L, You Y, Lian H (2015) Convergence and sparsity of lasso and group lasso in high-dimensional generalized linear models. Stat Pap 56:819–828

19. Ye J (1998) On measuring and correcting the effects of data mining and model selection. J Am Stat Assoc 93:120–131

20. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B 68:49–67

21. Zhao P, Rocha G, Yu B (2009) The composite absolute penalties family for grouped and hierarchical variable selection. Ann Stat 37(6A):3468–3497

22. Zou H, Yuan M (2008) The f-infinity-norm support vector machine. Stat Sin 18:379–398