Hands-On Compressed Sensing:

SISSO (Sure Independence Screening plus Sparsifying Operator)

Prepared by Zhong-Kang Han, Debalaya Sarker, and Sergey V. Levchenko
Advanced Materials Modeling
Skoltech, May 23, 2020

A quick summary of the exercises

A guide through the tutorial

This tutorial aims to give a basic introduction of the state of the art compressed sensing method SISSO (Sure Independence Screening plus Sparsifying Operator). SISSO enables us to identify the best low-dimensional descriptor for material properties in an immensity of offered candidates.

Part I: Installation of the SISSO code

Part II: Input and output file description

Part III: Water adsorption on different transition metal surfaces

Part I: Installation

This program is written in Fortran 90, and a MPI Fortran compiler (from Intel or GNU) is needed for the installation.
E.g.: go to the folder "~/softs/zhongkang/SISSO.2.3/src" and do: mpiifort -O2 var_global.f90 libsisso.f90 DI.f90 FC.f90 SISSO.f90 -o ~/bin/ executable_file_name
or mpigfortran -O2 var_global.f90 libsisso.f90 DI.f90 FC.f90 SISSO.f90 -o ~/bin/executable_file_name
where executable_file_name is the executable name of your choice (e.g., sisso.exe).
(The code compiled using mpiifort was found ~ 1.5X faster than that using mpigfortran according to our tests)

The modules:
- var_global.f90 module for declaring global variables
- libsisso.f90 module of mathematical functions
- DI.f90 module for descriptor identification
- FC.f90 module for feature construction

Part II: Input and output file description

Running SISSO

Input Files: "SISSO.in" (Figure 1) and "train.dat" (Figure 2)

(see folder "~/softs/zhongkang/SISSO.2.3/input_template" for the templates; As a toy example, just run SISSO with the templates without any changes)

Important keywords in the input file "SISSO.in"

Keyword: desc_dim = integer, dimension of the descriptor. The descriptors and corresponding models for lower dimensions will be computed automatically.
Keyword: nsample = integer, total number of samples (data points) in your data set.
Keyword: nsf = integer, total number of scalar *primary* features.
Keyword: rung = integer, rung of the feature space to be constructed. In practice, a huge pool of candidate descriptors is constructed iteratively by combining user-defined primary features with a set of mathematical operators. The number of times (the value of rung) the operators are applied determines the complexity of the resulting descriptors. In this tutorial we suggest to test up to two levels of complexity (feature spaces): Φ_1 and Φ_2. Note that a given feature space Φ_n also contains all of the lower rung (i.e. n-1) feature spaces.
Keyword: opset = string, mathematical operator set for feature construction.
Keyword: ndimtype = integer, number of dimension types (physically distinct units) of the primary features. A dimensional analysis will be performed, which ensures that only physically meaningful combinations are retained (e.g., only primary features with the same unit are added or subtracted).
Keyword: subs_sis = integer, SISSO selected subspace size. The sparsifying ℓ_0 (number of non-zero coefficients in the linear model of the target property) constraint is applied to a smaller feature subspace selected by a screening procedure (sure independence screening (SIS)), where the size of the subspace is equal to a user-defined SIS value set by this keyword times the dimension of the descriptor.
Keyword: metric = string, metric for model selection. Both root mean square error RMSE (set string to LS_RMSE) and maximum absolute error MaxAE (set string to LS_MaxAE) are available.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

$$MaxAE = \max_{i=1.....N} |Predicted_i - Actual_i|$$

Keyword: nmoutput = integer, number of the best models to be output.

Output:
- File SISSO.out (Figure 3): all the information regarding parameter setting, feature space, the best descriptors/models, and the associate fitting errors
- Folder models: the top ranked candidate descriptors/models
- Folder desc_dat: the values of the target properties and the best descriptors/models predicted properties

- Folder residual: residual data generated at each iteration
- Folder feature_space: SIS-selected subspaces (feature data and names)

```
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
task='FCDI'          !'FCDI': iterative FC and DI; 'FC': one-off FC; 'DI': one-off DI
restart=.false.      ! for 'FCDI' and when the file 'CONTINUE' is present
desc_dim=2           ! dimension of the descriptor (<=2 for 'quali')
ptype='quali'        ! property type: 'quanti'(quantitative property),'quali'(qualitative property)
nprop=1              ! number of properties (multi-task learning if nprop >1)
nsample=(2,3)        ! number of samples for each property (seperate them by comma for nprop>1)
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
nsf= 3               ! number of scalar features (one number for one material)
rung=2               ! rung (0,1,2,3) of the feature space to be constructed
maxcomplexity=10     ! max feature complexity (number of operators in a feature)
opset='(+)(-)(*)(/)(exp)(log)(^-1)(^2)(^3)(sqrt)(cbrt)(|-|)'  ! ONE operator set for feature transformation
ndimtype=2           ! number of dimension type (or unit) of the primary features
dimclass=(1:2)(3:3)  ! group features according to their dimension/unit; those not in any (): dimensionless
maxfval_lb=1e-3      ! features having the max. abs. data value <maxfval_lb will not be selected
maxfval_ub=1e5       ! features having the max. abs. data value >maxfval_ub will not be selected
subs_sis=20          ! SIS-selected (single) subspace size

!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
! parameters for descriptor identification
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
method='L0'          ! sparsification operator: 'L1L0' or 'L0'
fs_size_DI=          ! to be set by SISSO automatically if task='FCDI': feature-space size
fs_size_L0=          ! to be set by SISSO automatically if (task='FCDI' and method ='L0'): size for L0
metric='LS_RMSE'     ! for ptype='quanti'; metric for model selection: LS_RMSE,LS_MaxAE
nm_output=100        ! number of the best models to output
```

Figure 1: An example SISSO.in file.

```
material      property  feature1  feature2  feature3
sample1        3.0862    0.8626    0.7043    0.6312
sample2        2.8854    0.7260    0.7818    0.6119
sample3        0.6907    0.4943    0.0044    0.4420
sample4        0.9902    0.0106    0.0399    0.9877
sample5        0.7242    0.0970    0.3199    0.5504
```

Figure 2: An example train.dat file.

```
iteration:   1
--------------------------------------------------------------------------------
FC starts ...
Standard Deviation (SD) of property 001:    1.07688
Total number of features in the space phi00:              3
Total number of features in the space phi01:             37
Total number of features in the space phi02:           2286
Size of the SIS-selected subspace from phi02:         400
Wall-clock time (second) for this FC:            0.01
FC done!

DI starts ...
parameter "fs_size_DI" set to        400
parameter "fs_size_L0" set to        400
L0 starts ..., space size:         400

Model/descriptor for generating residual:
================================================
  1D descriptor (model):
Total LS_RMSE,LS_MaxAE:   0.104833  0.178032
@@@descriptor:
                    1:[((feature1*feature2)/log(feature3))]
       coefficients_001:    -0.18030E+01
          Intercept_001:     0.75055E+00
       LSrmse,maxAE_001:     0.10483E+00    0.17803E+00
================================================
Wall-clock time (second) for this DI:            0.00
DI done!
```

Figure 3: An example SISSO.out file.

Part III: Water adsorption on different transition metal surfaces

A quick start

Copy the folder ~/softs/zhongkang/SISSO.2.3/water-adsorption to your own folder.

Step I: Training data accumulation

In this exercise, we use the lowest adsorption energy of water at different transition metal surfaces by considering different possible adsorption sites (Figure 4) calculated by DFT as data input (totally 45 data were included). All the energies can be found in the file ~/softs/zhongkang/SISSO.2.3/water-adsorption/train.dat

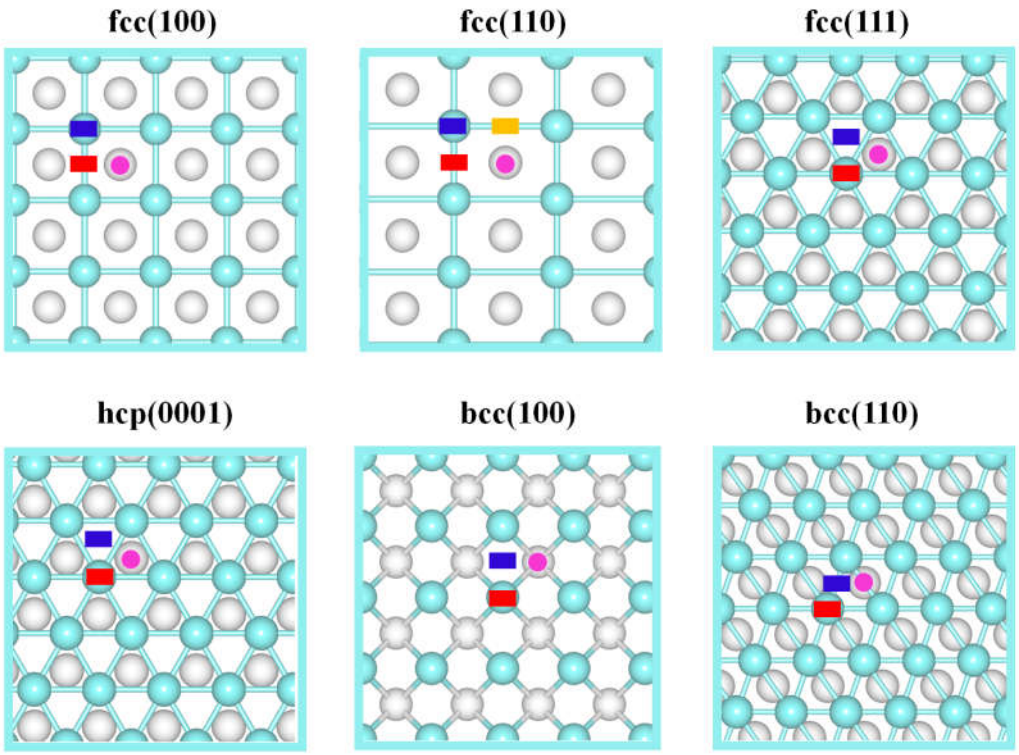

Figure 4: The considered different possible adsorption sites and surface cuts.

Step II: Feature space construction and selection

For constructing the Φ_1 and Φ_2 feature spaces we suggest to use in this tutorial the following set of algebraic/functional operators:

$$\hat{H}^{(m)} \equiv \{+, -, \cdot, /, \ln, \exp, \exp-, ^{-1}, ^{2}, ^{3}, \sqrt{}, ^{3}\sqrt{}, |-|\}. \quad (1)$$

The superscript *m* indicates that when applying $\hat{H}^{(m)}$ to primary features φ_1 and φ_2 a dimensional analysis is performed, which ensures that only physically meaningful combinations are retained (only primary features with the same unit are added or subtracted). Figure 5 shows the primary features used in this tutorial. The values of the primary features for the training data sets can be found in the file "~/softs/zhongkang/SISSO.2.3/water-adsorption/train.dat". Note that the primary features do not include any properties of adsorbed water molecule (otherwise there would remain nothing to predict).

The sparsifying ℓ_0 (number of non-zero coefficients in the linear model of the property) constraint is applied to a smaller feature subspace selected by a screening procedure (sure independence screening (SIS)), where the size of the subspace is equal to *a user-defined SIS value* times the dimension of the descriptor. The SIS value is not an ordinary hyperparameter and its optimization through a validation data set is not straightforward. Ideally, one would want to search the entire feature space for the

optimal descriptor. However, this is not computationally tractable since the computational cost of the sparsifying ℓ_0 constraint grows exponentially with the size of the searched feature space. Instead, the SIS value should be chosen as large as computationally possible. The reasonable SIS values (15 was found to be sufficient in this exercise) are chosen based on the convergence of the training error.

Primary features

Class	Name	Abbreviation
Atomic	Atom radius	R
	Electronegativity	E
	HOMO	H
	LUMO	L
	Ionization energy	I
Bulk	*d* band center	DB
	Fermi energy	F
Surface	*d* band center	DS
	Chemical potential	C
	Coordination number	CN
	Effective coordination number	ECN

Figure 5: Primary features used in this exercise.

Step III: *k*-fold cross-validation.

To test the predictive power of obtained models, we employ 20-fold cross-validation (CV20). In this validation approach, the dataset is first split into 20 subsets, and the descriptor identification along with the linear model training (fitting) is performed using 19 subsets. Then the errors in predicting properties of the systems in the remaining subset is evaluated with the obtained model. The CV20 error is defined as the average value of the test errors obtained for each of the 20 subsets:

$$CV = \frac{1}{N}\sum_{k=1}^{N} RMSE(test_i),$$

where N is the number of tests (20 for 20-fold CV). For the k^{th} test, the model is fitted with k-1 parts of the data, and the CV error is calculated based on the prediction RMSE for the k^{th} subset.

In SISSO over-fitting may occur with increasing *dimensionality* of the descriptor (i.e., the number of complex features that are used in construction of the linear model). The CV20 error first decreases with the dimension of the descriptor, since more

parameters become available to fit the data. However, as the dimension (the number of model parameters) increases, the model becomes too sensitive to a particular selection of the training data. The descriptor dimension at which the CV20 error starts increasing identifies the optimal dimensionality of the descriptor. Then this dimension is used to find the best model based on the whole training data set.

The python script that can be used to randomly split the data set into *k* subsets with the similar size can be found in the folder "~/softs/zhongkang/SISSO.2.3/water-adsorption". If the data set cannot be divided by *k* evenly, the remaining data are randomly distributed among the *k* subsets.

Homework assignment:

1) Perform 20-fold cross-validation (CV) study for descriptor dimensions (hyperparameter value) 1-7. It is sufficient to request in the input file the maximum dimension (7), the descriptors and corresponding models for lower dimensions will be computed automatically.

2) Report CV error for each dimension. Identify the best dimension (for which CV error is lowest).

3) Identify best descriptor and the corresponding model for the adsorption energy of H_2O on metal surfaces with the best hyperparameter value (dimension).

4) Plot SISSO-predicted versus DFT calculated values of adsorption energies for all surfaces. Do the same for the *d*-band center model. The values of the *d*-band center for each surfaces (DS) can be found in the file "~/softs/zhongkang/SISSO.2.3/water-adsorption/train.dat"

5) Report RMSE and MaxAE for the best SISSO model. Do the same for the *d*-band center model.

6) Discuss the results. Discuss what you can learn from the 1- and

2-dimensional descriptors.

If you have any questions, please email Dr. Zhong Kang Han (h.zhongkang@skoltech.ru) with a copy to Sergey Levchenko (s.levchenko@skoltech.ru).