This doc records some context about using Beijing housing price dataset and associated matlab code

Dataset

There are two CSV files need to be loaded as sourcing datasets

- 'new BJ house.csv':
    - this CSV file is the raw source dataset I got from Kaggle dataset. In the matlab code, only column 3 (long) and column 4 (lat) are necessary. The location information is used to determine whether the house falls into the manual crafted triangulation. Only valid data points are left
    - each column name indicates what variable it represents. For more details, you can find it from the source website: https://www.kaggle.com/datasets/ruiqurm/lianjia
- 'new BJ stepAIC design matrix std.csv':
    - this CSV file contains all variables including raw variables and interaction of raw variables. They are picked via AIC and stepwise selection method from the entire set consists of all main effects and interactions of all covariates. This step is implemented in R studio (but not added to the current folder). You can check the first row of this CSV file to know what it represents. The first column is called “intercept” and would always be 1. This dataset is already standardized
    - each column name indicates what variable or what interaction fact it represents

Matlab Code

You can check the file “CZ_github_BJ.m” and I add sections including loading both CSV files, design matrix generation process, and cross-validation code. I rerun the method Linear and UNPEN, the result is the same as the one we present in the paper. However, for the SCAD, I met with OOM issue on my current laptop

I copy and paste the result table from the paper as below

Method	R^2	MSEE	MSPE	BIC
Linear	0.6256	0.3744	0.3747	-0.980
UNPEN	0.8531	0.1469	0.1497	-1.837

SCAD	0.8495	0.1505	0.1524	-1,850