Model Population Analysis

(模型集群分析)

Hong-Dong Li and Yi-Zeng Liang

lhdcsu@gmail.com

Aug., 2011

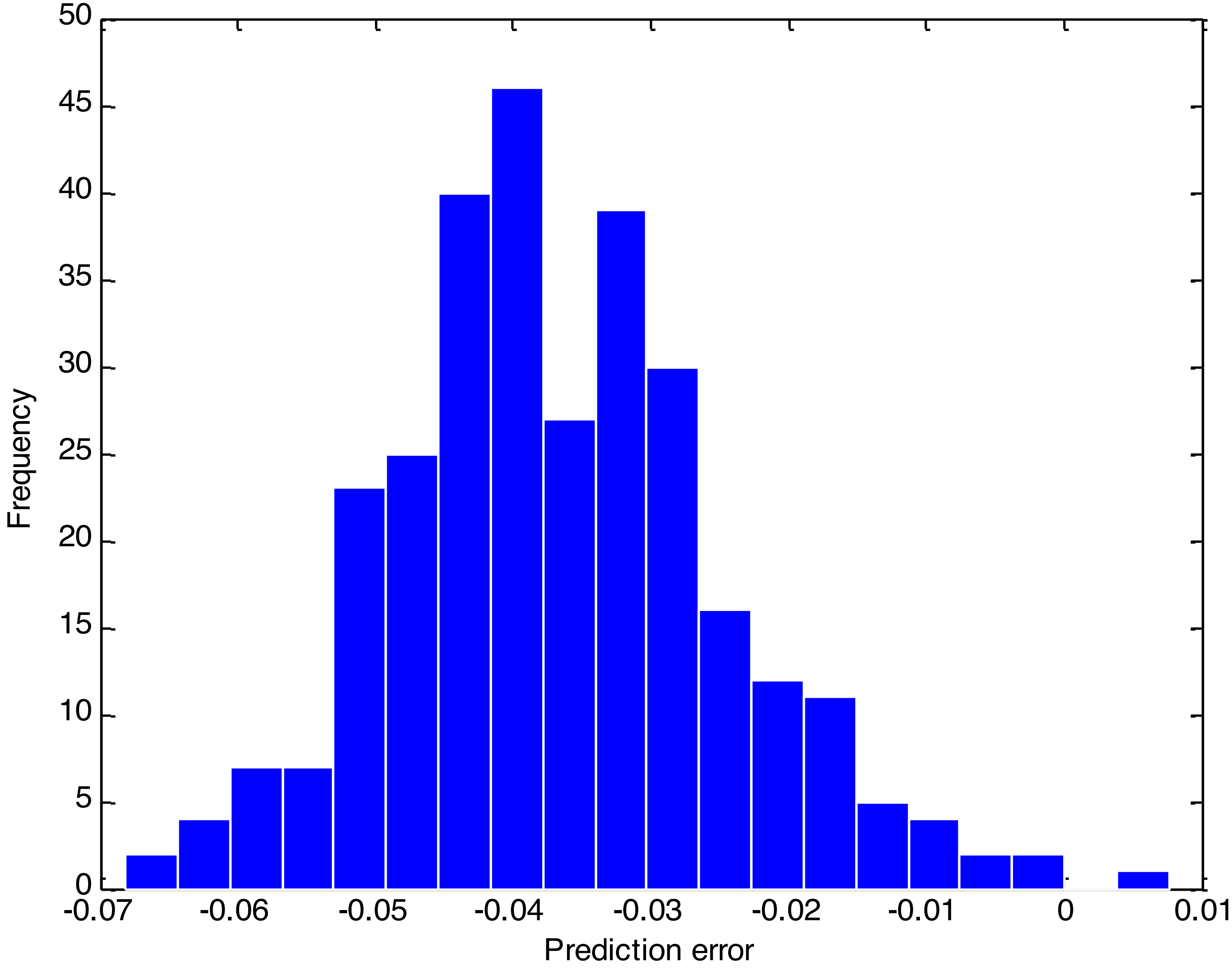
Outline

- Context
- Model population analysis (MPA)
- Variable assessment using MPA

Context

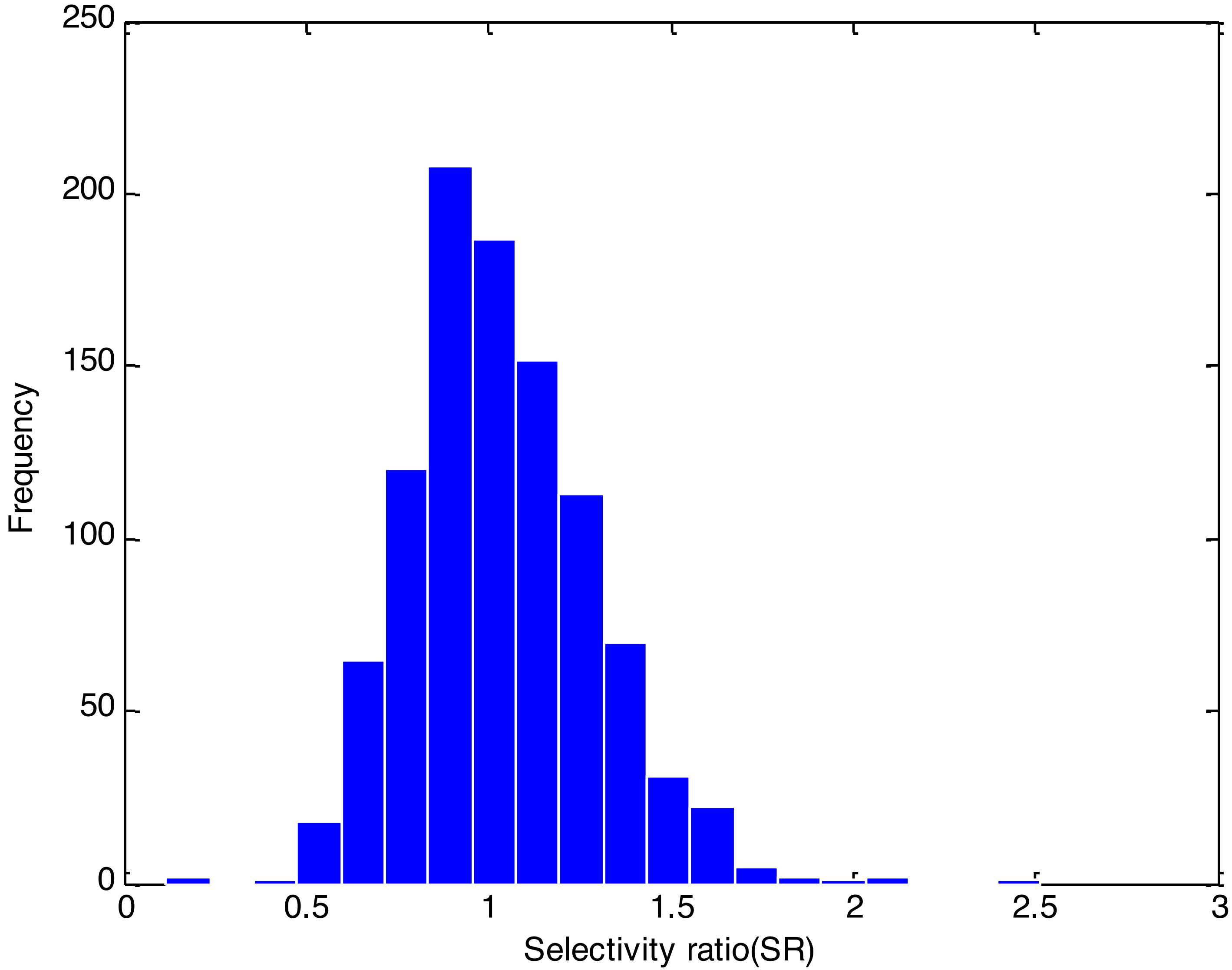
- Outlier detection
- Variable assessment
- Model performance
- Ensemble learning

Outlier



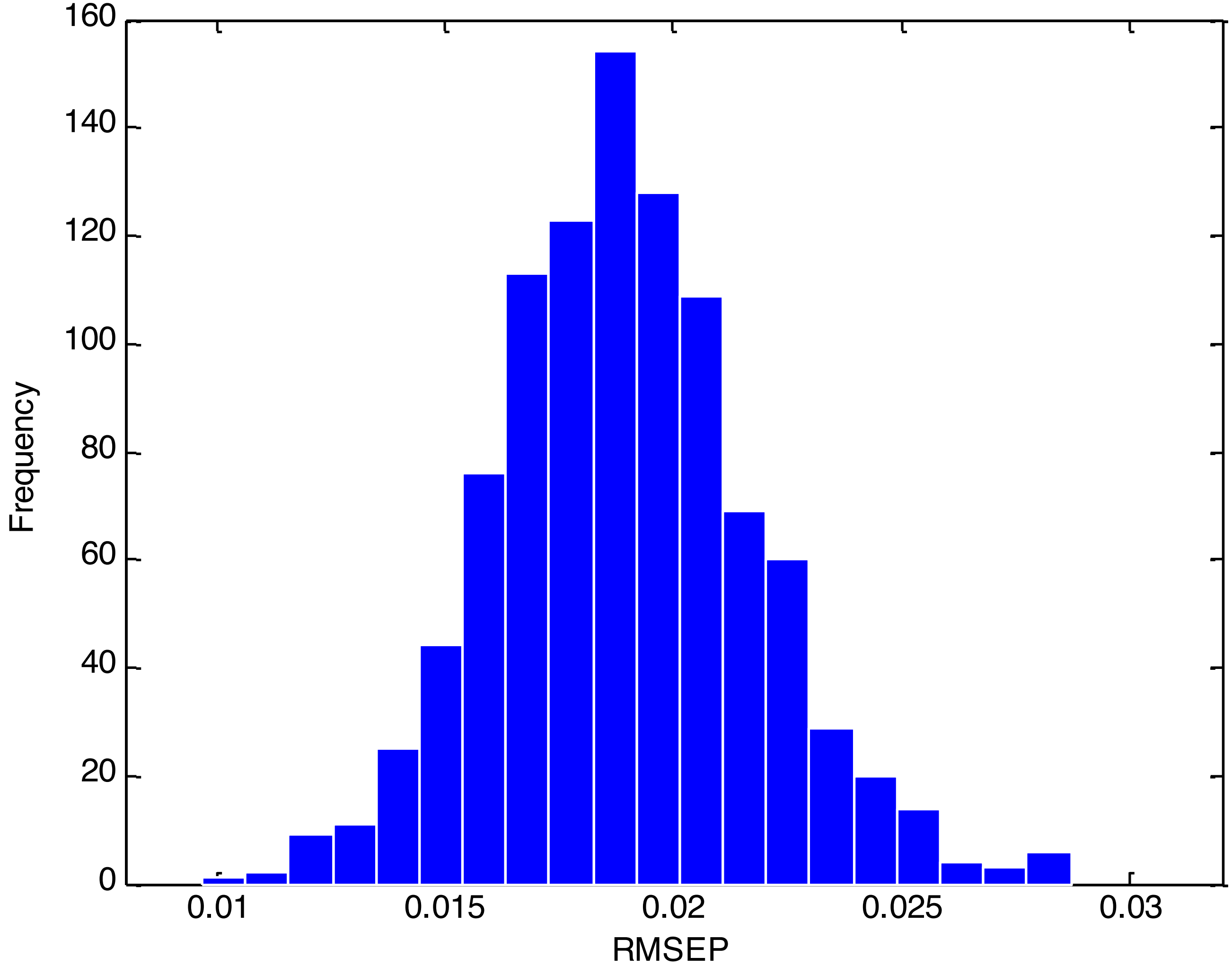
Corn data: Prediction errors of a test sample **from 303 models**

Variable importance



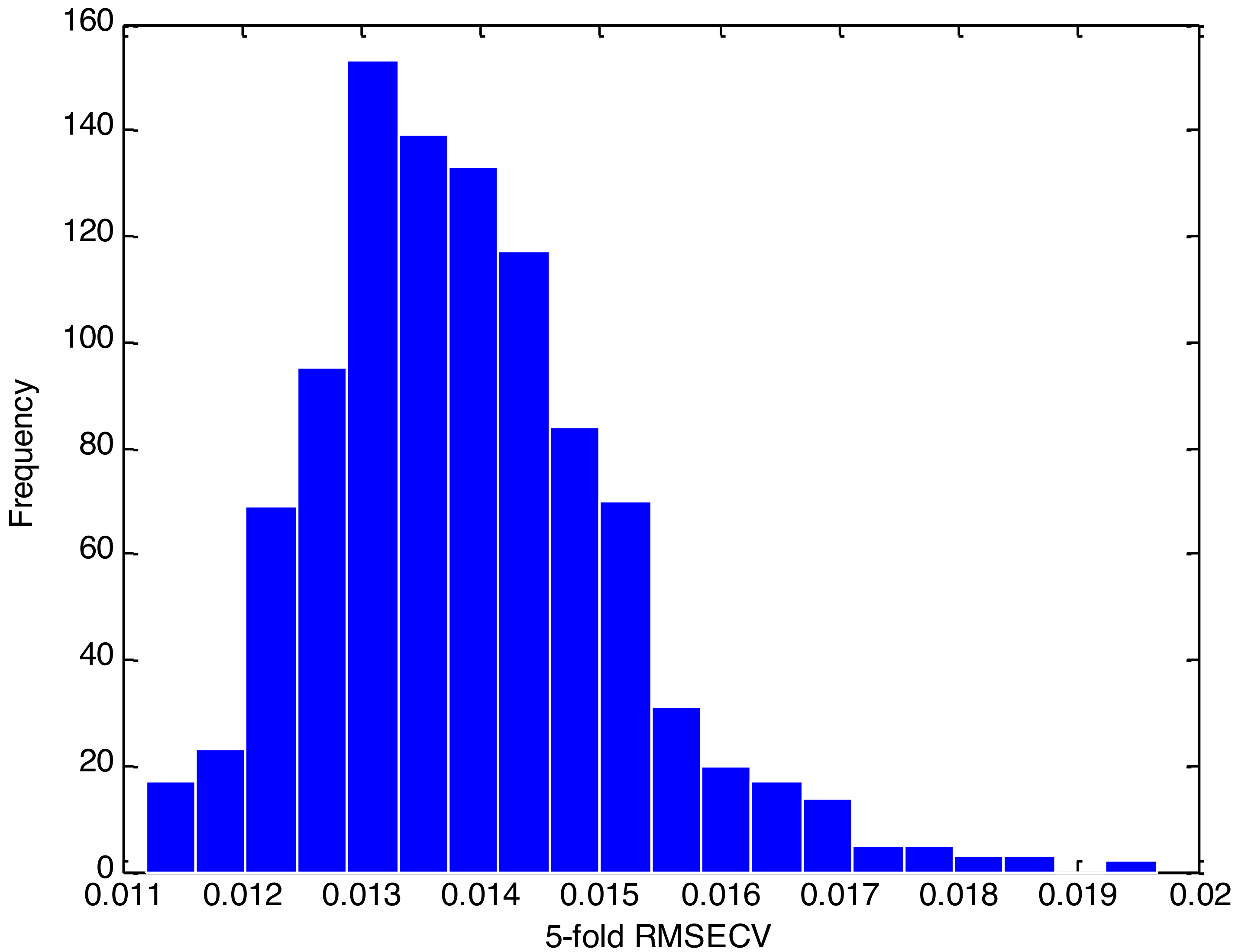
T2DM data (n=90, p=21), 70% samples, **from 1000 models**

Model performance: test set



Data: corn m5 moisture. **From1000 models**

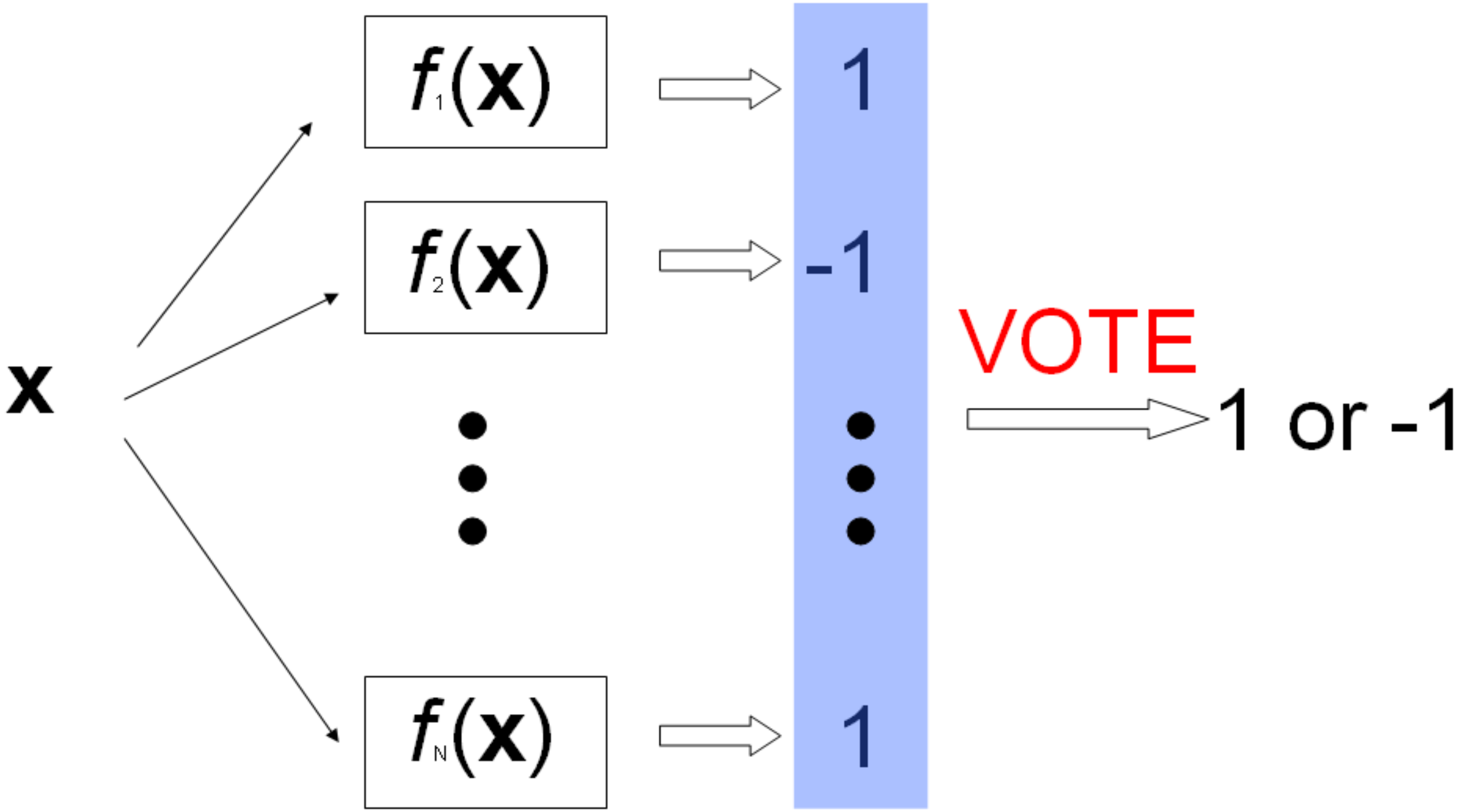
Model performance: cross validation



Data: corn m5 moisture. **From 1000 models**

Ensemble learning

Bagging, Boosting and Random forest



A population of N models

Conclusions

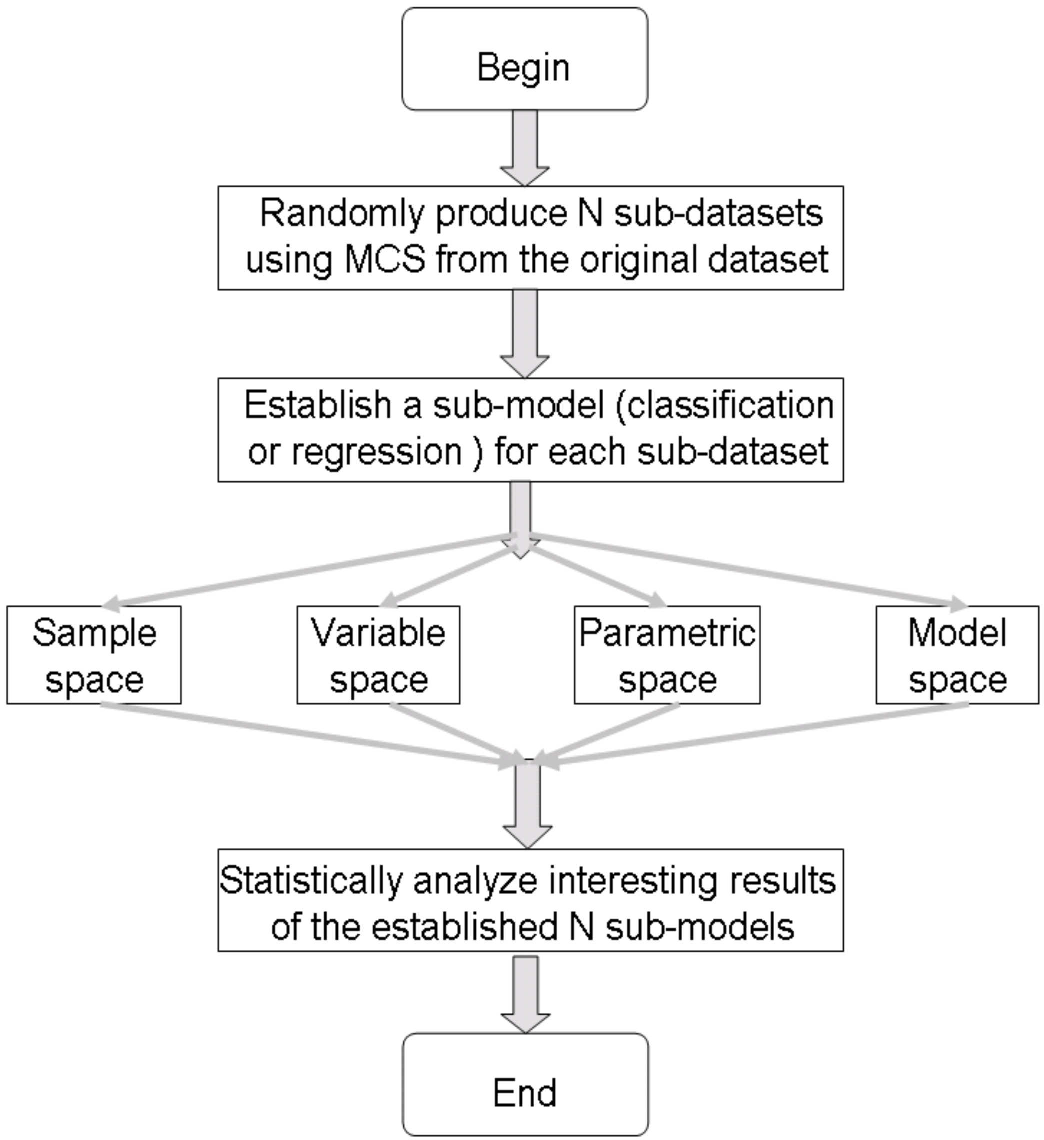
- ◆ Prediction errors or variable importance or model performance is **data-dependent**
- ◆A **single number** is not sufficient to characterize…
- ◆Hence we suggest to use the **distribution** of …

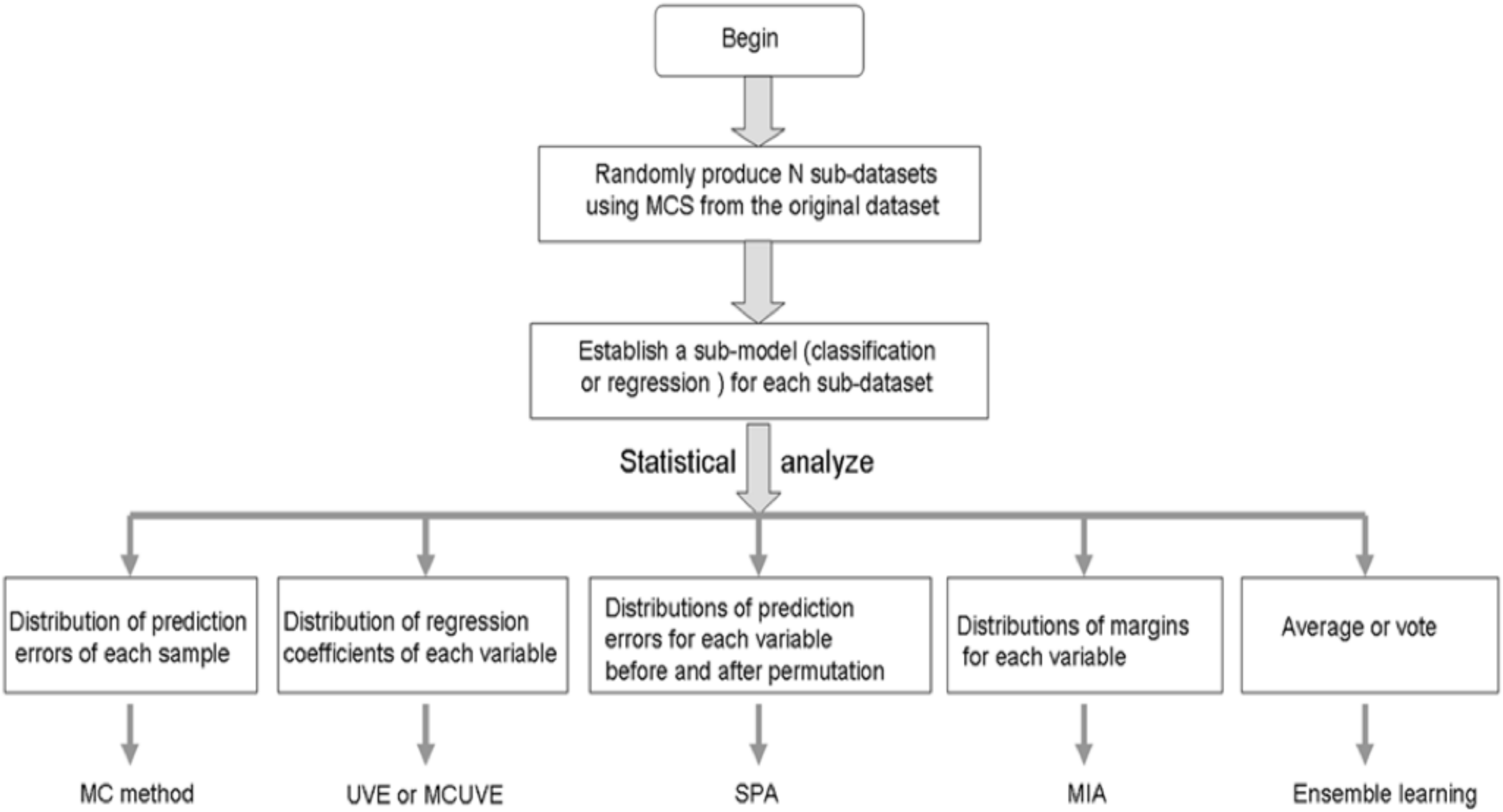
A new concept
Model Population Analysis

Hong-Dong Li, Yi-Zeng Liang, Qing-Song Xu, Dong-Sheng Cao, model population analysis for variable selection, *Journal of Chemometrics* **2009,** 24, (7-8), 418-423

What is Model Population Analysis?

A general framework for developing data analysis methods




Begin
Randomly produce N sub-datasets using MCS from the original dataset
Establish a sub-model (classification or regression) for each sub-dataset
Statistical analyze
Distribution of prediction errors of each sample
Distribution of regression coefficients of each variable
Distributions of prediction errors for each variable before and after permutation
Distributions of margins for each variable
Average or vote
MC method
UVE or MCUVE
SPA
MIA
Ensemble learning

Our work on model population analysis

[1]. Li, H.-D., Liang, Y.-Z., Xu, Q.-S. & Cao, D.-S. Model population analysis for variable selection. Journal of Chemometrics 24, 418-423 (2009).

[2]. Cao, D.S., Liang, Y.Z., Xu, Q.S., Li, H.D. & Chen, X. A New Strategy of Outlier Detection for QSAR/QSPR. J. Comput. Chem. 31, 592-602 (2010).

[3]. Li, H.-D. et al. Recipe for revealing informative metabolites based on model population analysis. Metabolomics 6, 353-361 (2010).

[4]. Li, H.-D. et al. Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis, http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2011).

[5]. Wang, Q., Li, H.-D., Xu, Q.-S. & Liang, Y.-Z. Noise incorporated subwindow permutation analysis for informative gene selection using support vector machines. Analyst 136, 1456-1463 (2011).

[6]. Li, H.-D., Liang, Y.-Z &. Xu, Q.-S, Model population analysis and its applications in chemical and biological modeling, *under review*

[7]. Li, H.-D., Liang Y.-Z&Xu, Q.-S, Variable complementary network: a novel approach for identifying disease related variables and their mutual associations, in preparation

[8]. Li, H.-D., Liang Y.-Z&Xu, Q.-S, statistical model comparison via model population analysis, an invited book chapter, *under review*

How to implement MPA?

1.Monte Carlo Sampling to obtain **sub-datasets**

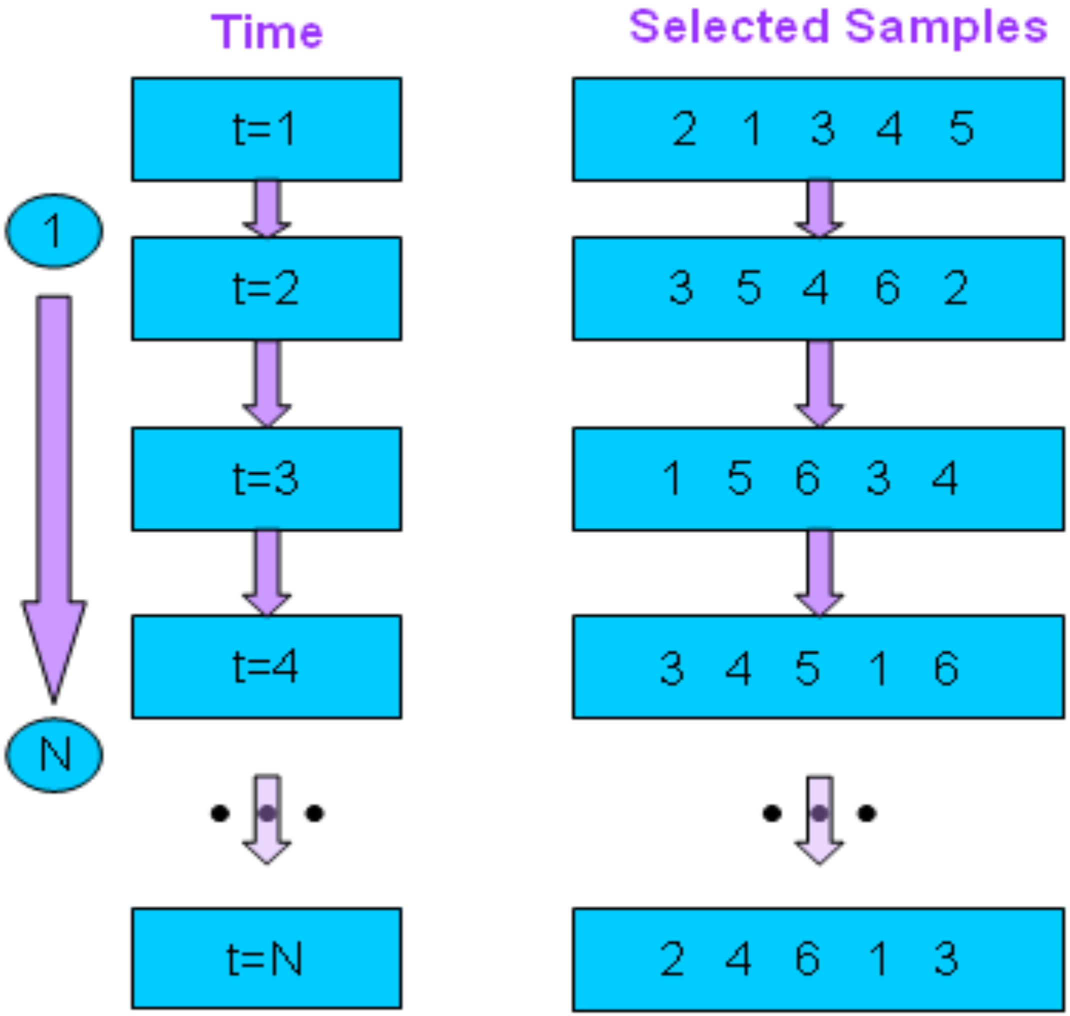
Monte Carlo Sampling

- Jacknife
- Bootstrap

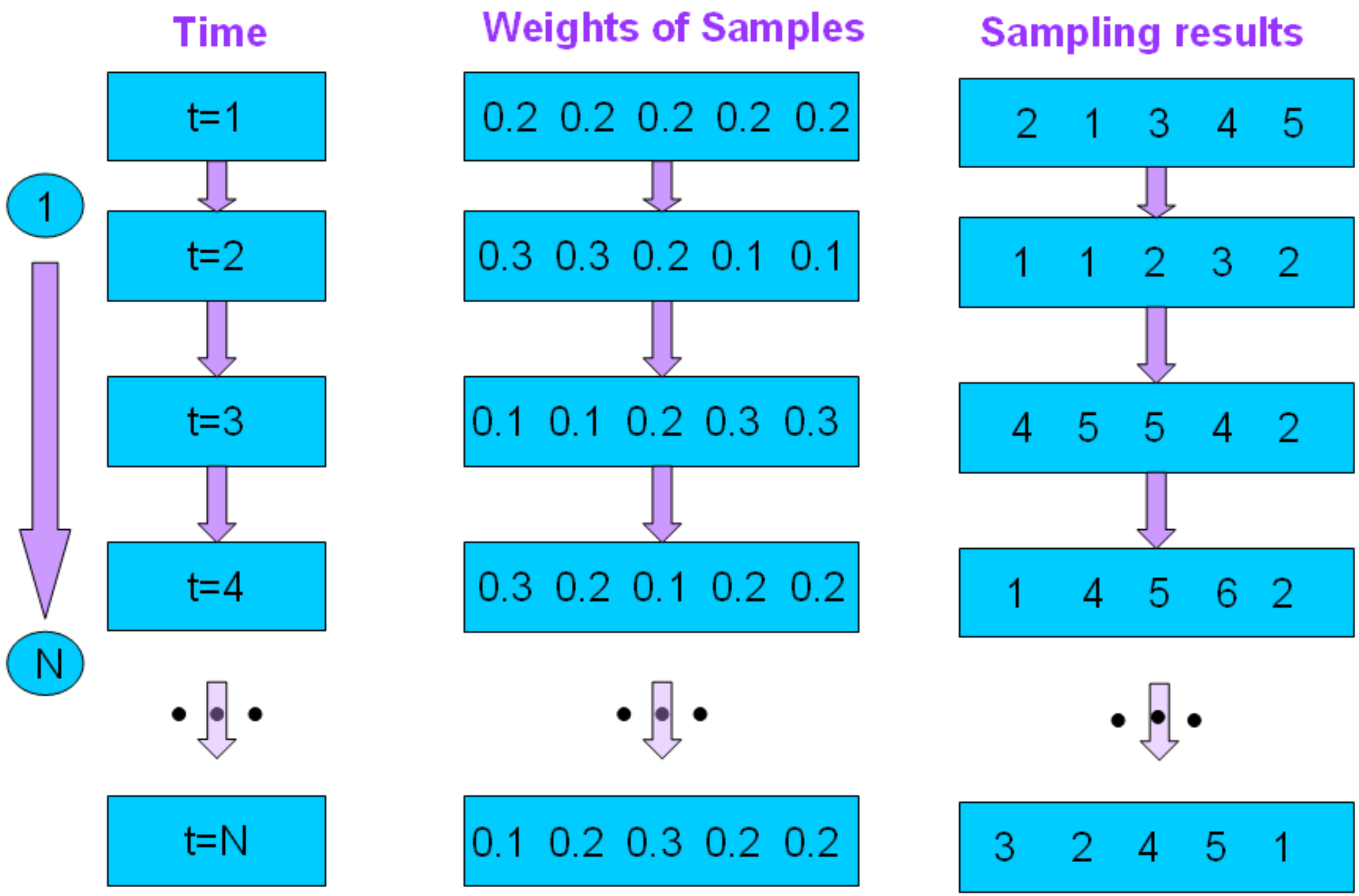
$\left.\begin{matrix}\\ \\ \end{matrix}\right\}$ Reweighted version

Jacknife

---Suppose we have 6 samples，denoted by 1, 2, 3, 4, 5 and 6



Weighted sampling



Suppose we have 5 samples, denoted by 1, 2, 3, 4 and 5

2. Build N sub-model for all **N sub-datasets**

Partial least squares

Support vector machines

Classification And Regression Trees

…

3. Statistical analysis of an **interesting output** of all the N sub-models

- Prediction residual of a sample
- Regression coefficient of a variable
- Variable importance
- Model-related parameter
- …

Model Population Analysis for **variable assessment**

Three new algorithms based on MPA:

SPA: Subwindow Permutation Analysis
MIA: Margin Influence Analysis
CIMPA: Condtional importance

To illustrate that:

Different kinds of designs for statistical analysis of some interesting parameters will result in different algorithms.

Subwindow Permutation Analysis

Motivated by:

- **Random forest**
- **Model Population Analysis**
- **Detecting synergistic effect**

HD Li, MM Zeng, BB Tan, YZ Liang, QS Xu, DS Cao, Recipe for revealing informative metabolites based on model population analysis, *Metabolomics* **2010,** 6, (3), 353-361.

What is permutation?

ID	normal	permuted	permuted
1	0.75	0.47	0.53
2	0.67	0.02	0.20
3	0.20	0.45	0.85
4	0.93	0.85	0.02
5	0.53	0.93	0.67
6	0.42	0.67	0.93
7	0.45	0.75	0.42
8	0.85	0.42	0.47
9	0.47	0.20	0.45
10	0.02	0.53	0.75

Lindgren, F., Hansen, B., & Karcher, W. (1996). Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics*, *10*, 521–532.

Variable importance in Random Forest (RF)

$Error\_normal = RF(Xtest)$

$Error\_permuted = RF(Xtest_j)$

$Variable\ importance_j = Error\_permuted - Error\_normal$

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

SPA is developed

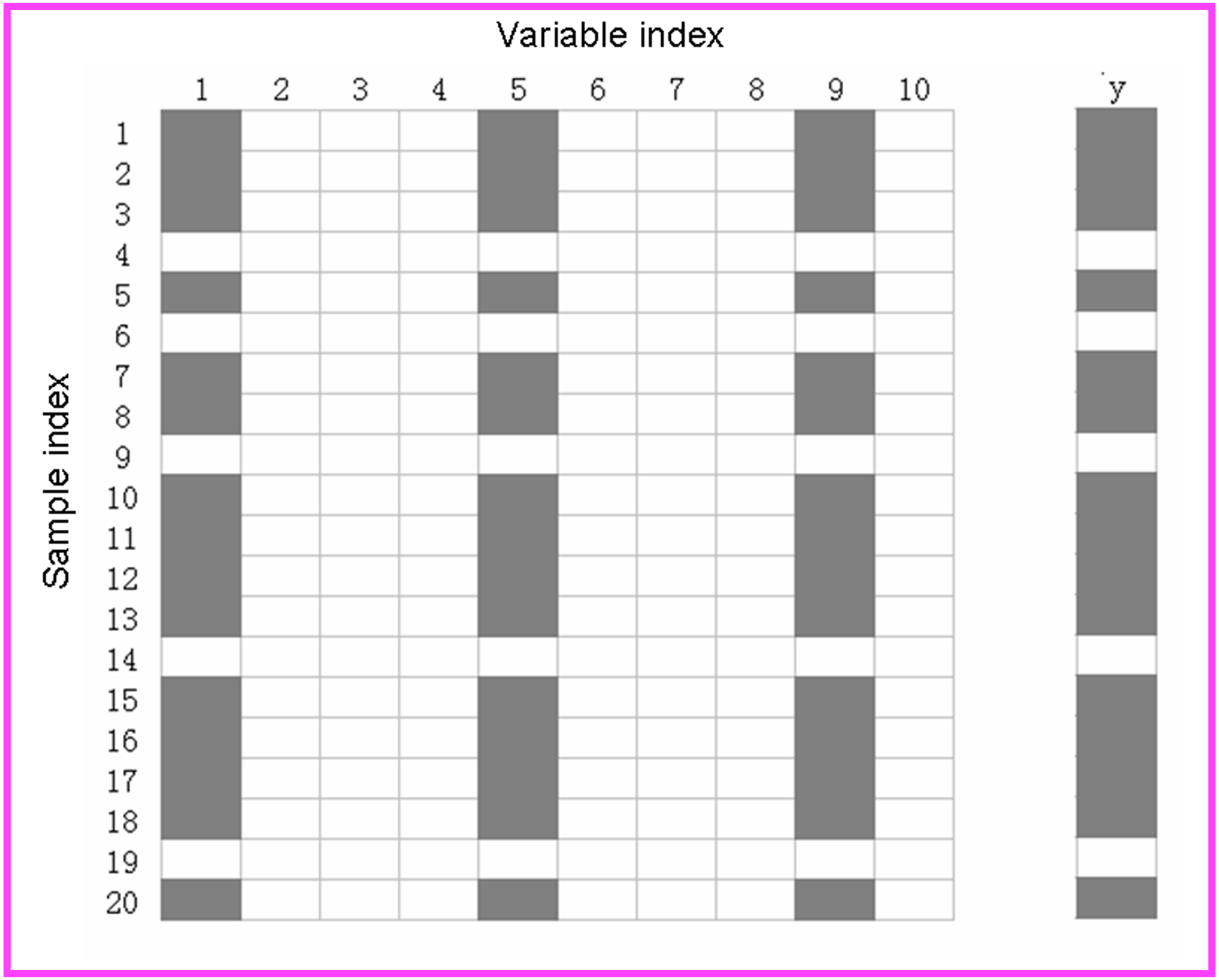
by exactly following the three elements of MPA

(1) sub-dataset sampling (N)

(2) sub-model building (N)

(3) statistical analysis of the interesting parameters of all the N models.

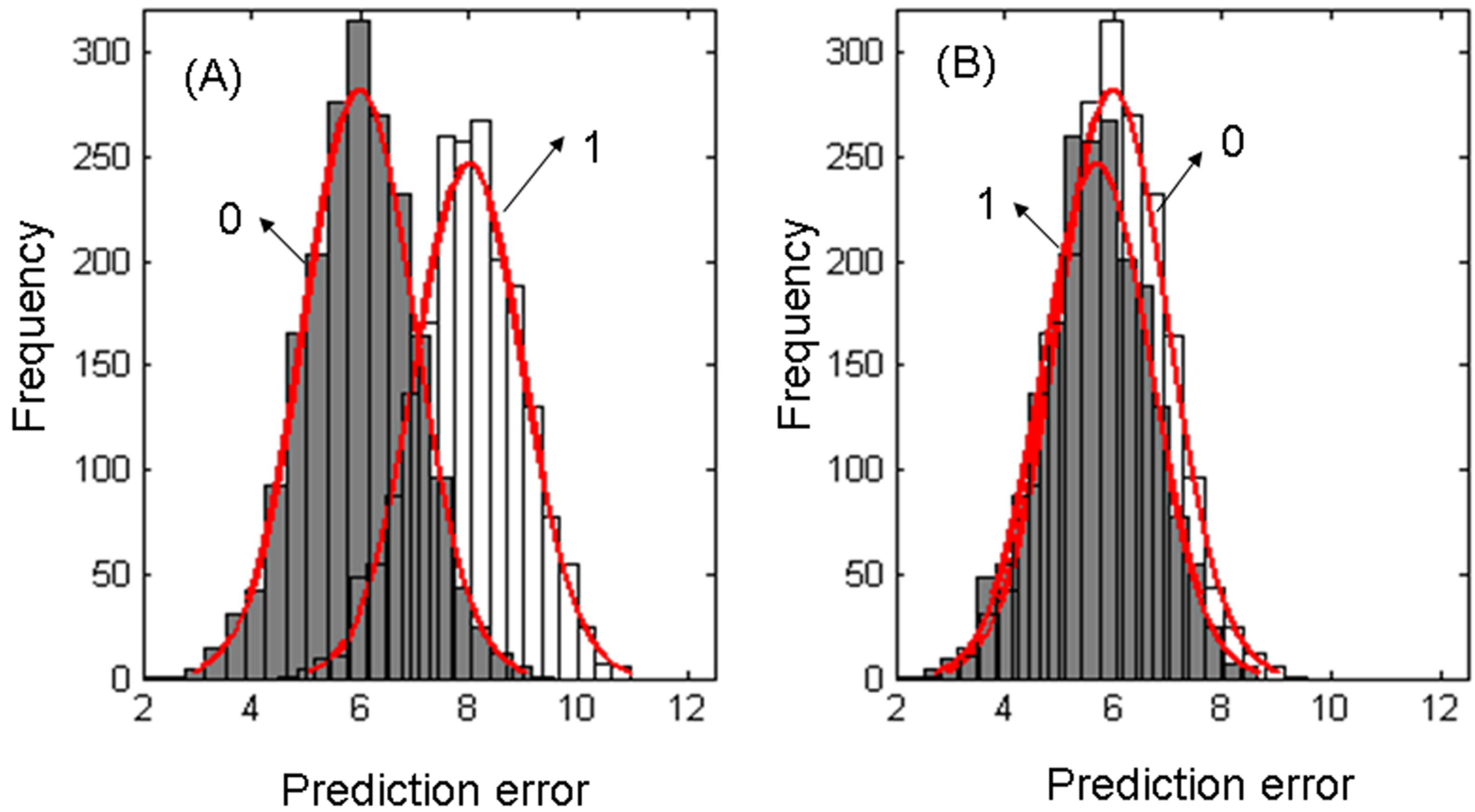
1. Sub-dataset sampling



2. Build N sub-models

ID	variable			Model	Test set	Prediction on test sets			
1	1	3	5	M1	Xtest1	NPE	PPE1	PPE3	PPE5
2	7	2	6	M2	Xtest2	NPE	PPE7	PPE2	PPE6
3	9	1	10	M3	Xtest3	NPE	PPE9	PPE1	PPE10
.	.	.	.	.	.	......			
N	6	4	8	MN	XtestN	NPE	PPE6	PPE4	PPE8

3. Statistical analysis of the prediction errors of the N sub-models



Peak 1: Permuted prediction errors (PPEs)

Peak 0: Normal prediction errors (NPEs)

How to compare the paired distributions?

We use the nonparametric Mann-Whitney U test

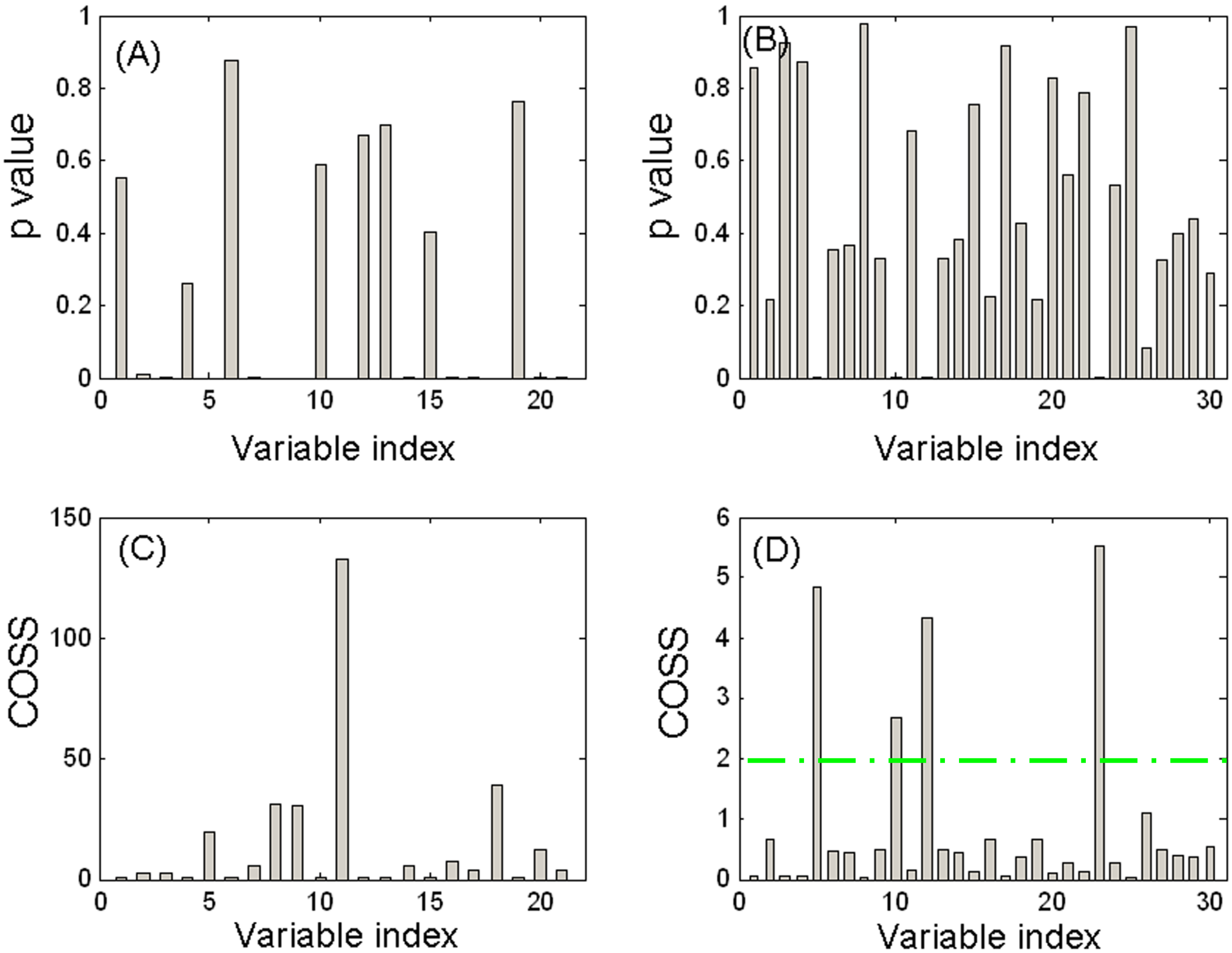
Lead to a COnditional Synergistic Score: COSS

$$\boxed{\mathrm{COSS} = -\mathrm{Log}_{10}(\mathrm{p})}$$

Applications of SPA to

- Type 2 diabetes mellitus data
- Childhood overweight data

Source codes **in MATLAB and R can be freely available at** **http://code.google.com/p/spa2010**



SPA-based Conditional P-value and the COSS score

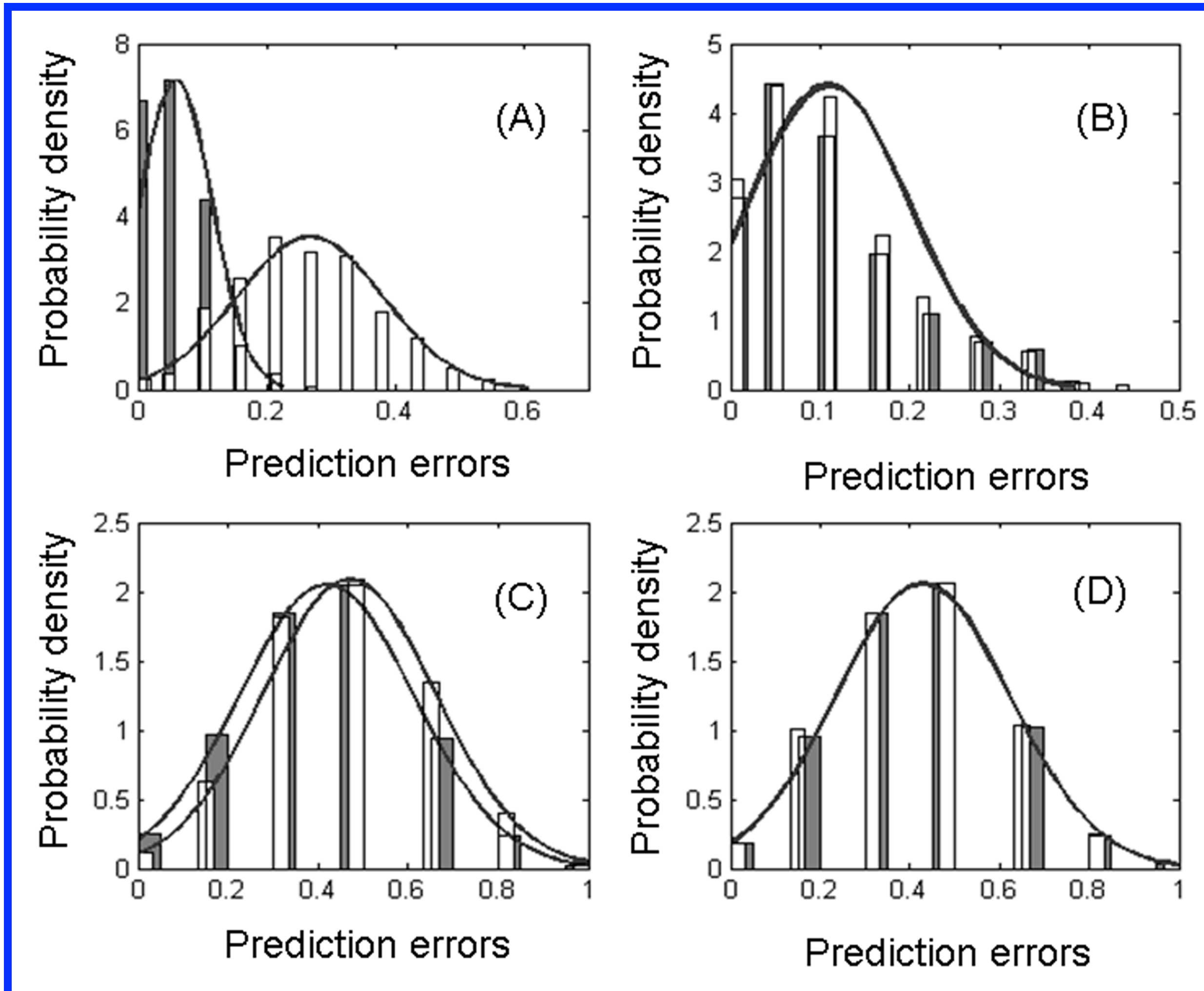


Fig. 4 Plot **A** and **B** shows the distributions of normal prediction errors (*grey bar*) and permuted prediction errors (*white bar*) of an informative metabolite (C18:1n-9, $p = 0$) and an uninformative one (C16:1n-7, $p = 0.8791$) for T2DM data, respectively. By analogy, such kind of distributions of an informative metabolite (Palmitic acid, $p = 3 \times 10^{-6}$) and an uninformative one (Leucine, $p = 0.9791$) for the childhood overweight data are shown in Plot **C** and **D**, respectively

Compare two distributions

Unsupervised

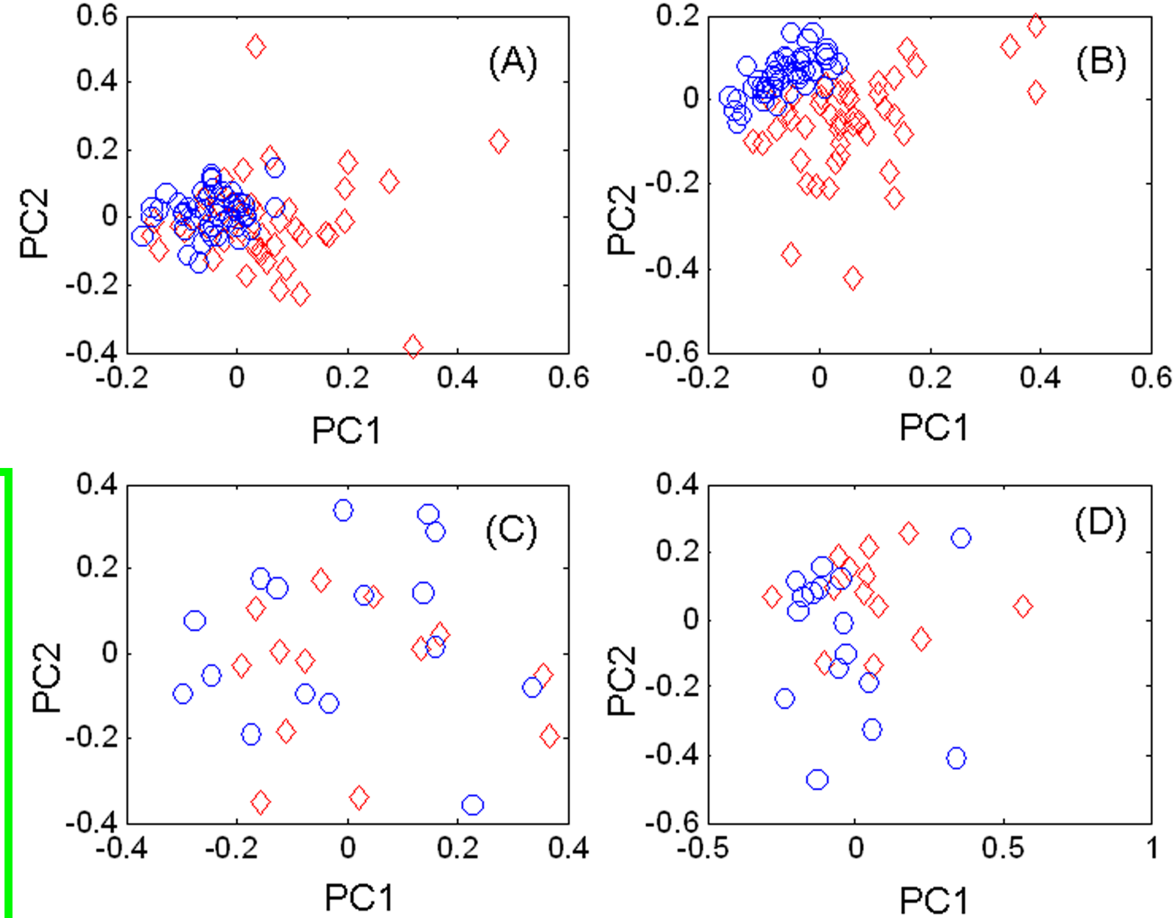


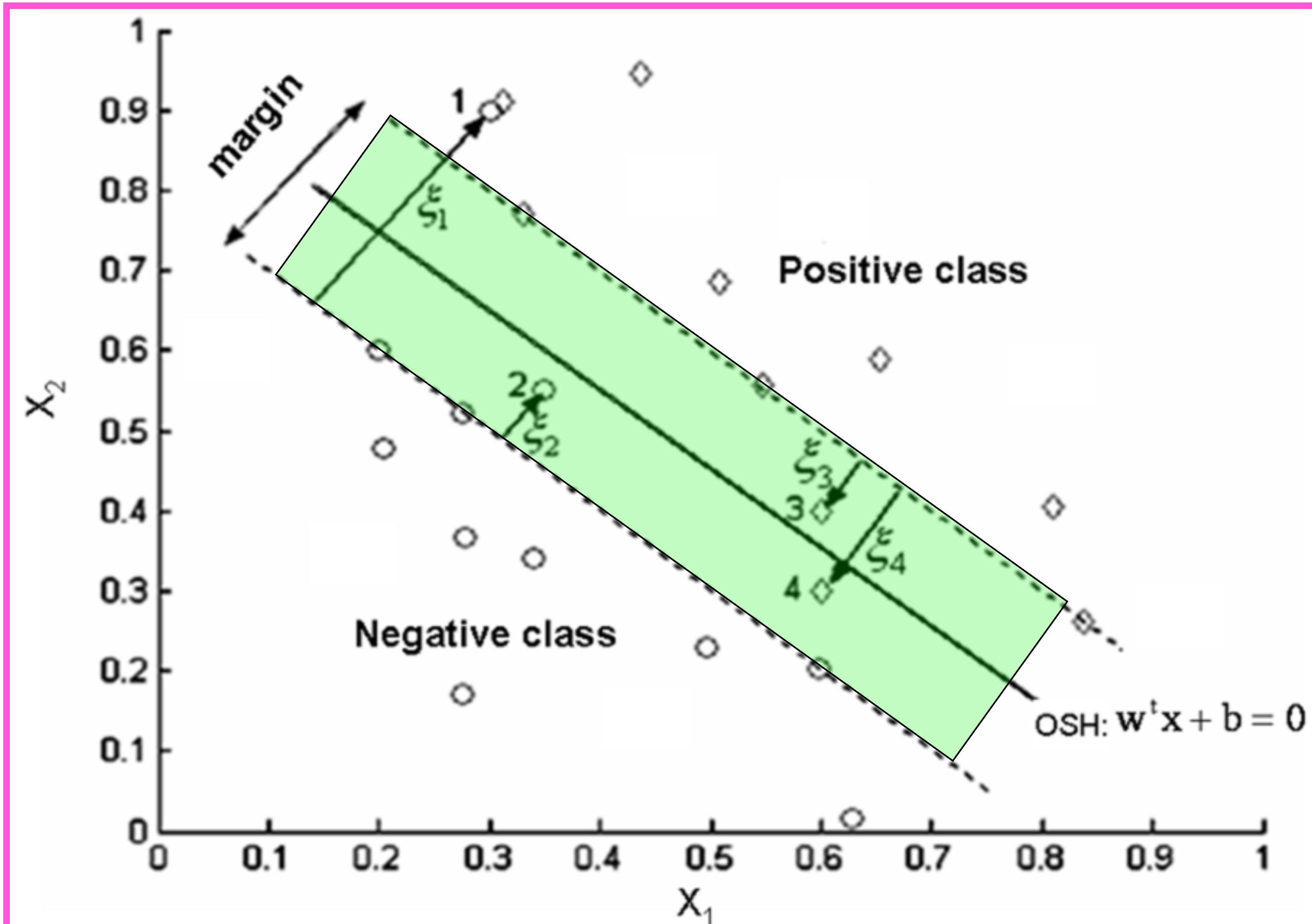
Fig. 5 Plot **A** and **B** display the PCA projected samples (*circle*: normal, *diamond*: patients) of the T2DM data using all the 23 metabolites and the selected three metabolites by SPA, respectively. Analogously, Plot **C** and **D** display the PCA projected samples (*circle*: normal, *diamond*: overweight) of the childhood overweight data using all the 30 metabolites and the selected three metabolites by SPA, respectively

Better separation

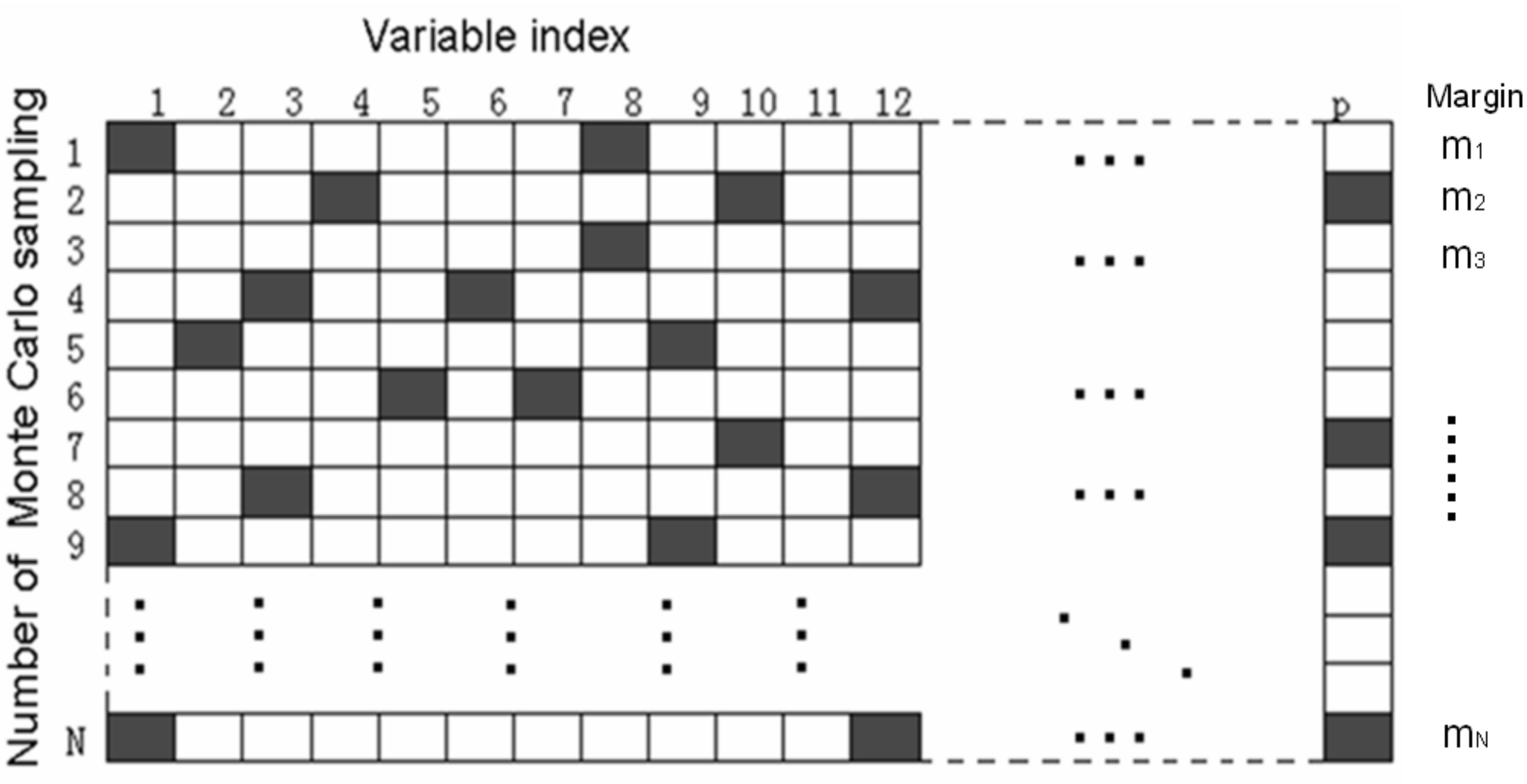
Margin Influence Analysis

Hong-Dong Li, Yi-Zeng Liang*, Qing-Song Xu et al, Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36*

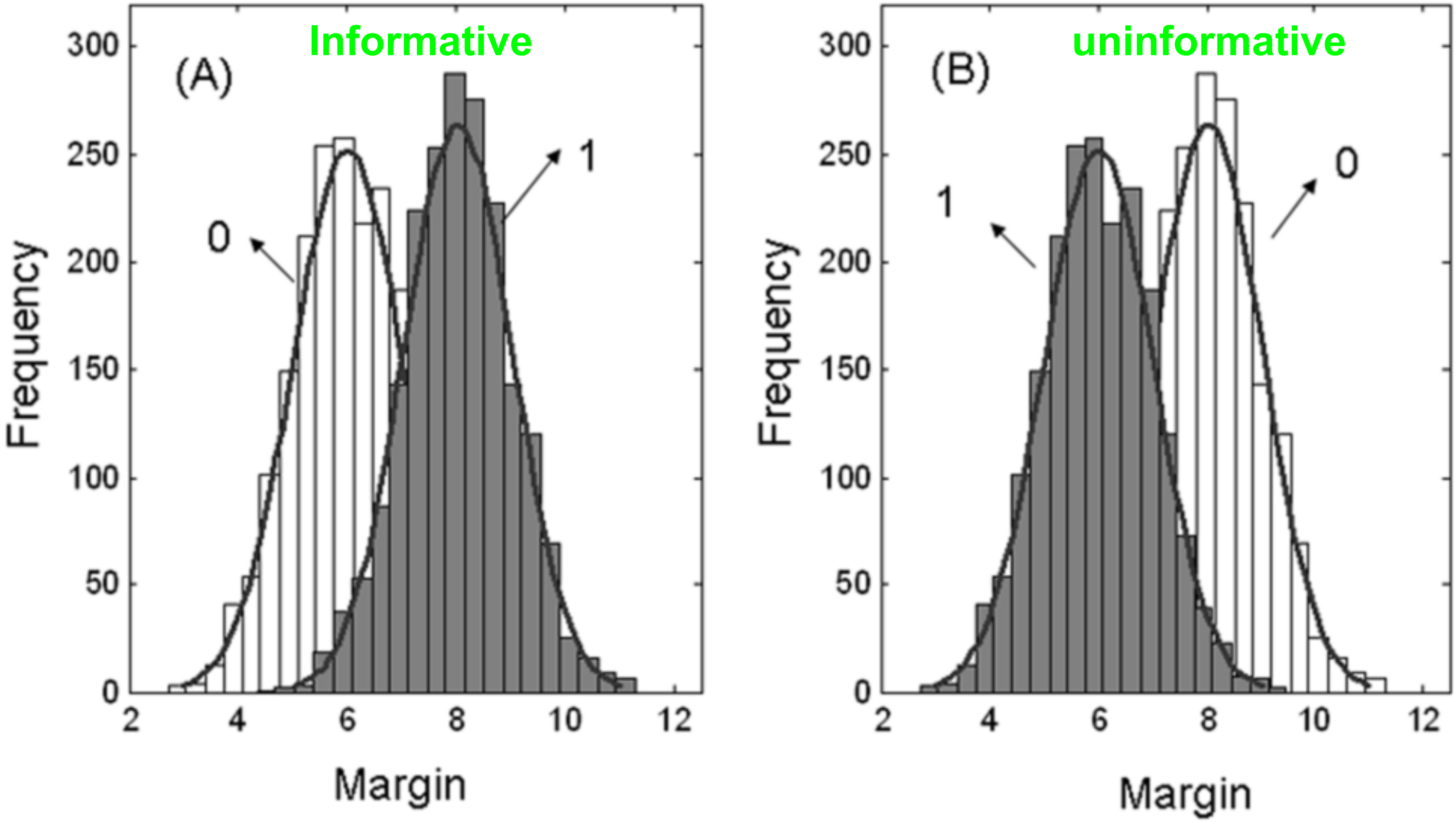
The margin of a SVM model



1 Sub-dataset sampling in variable space
2 Build N SVM models



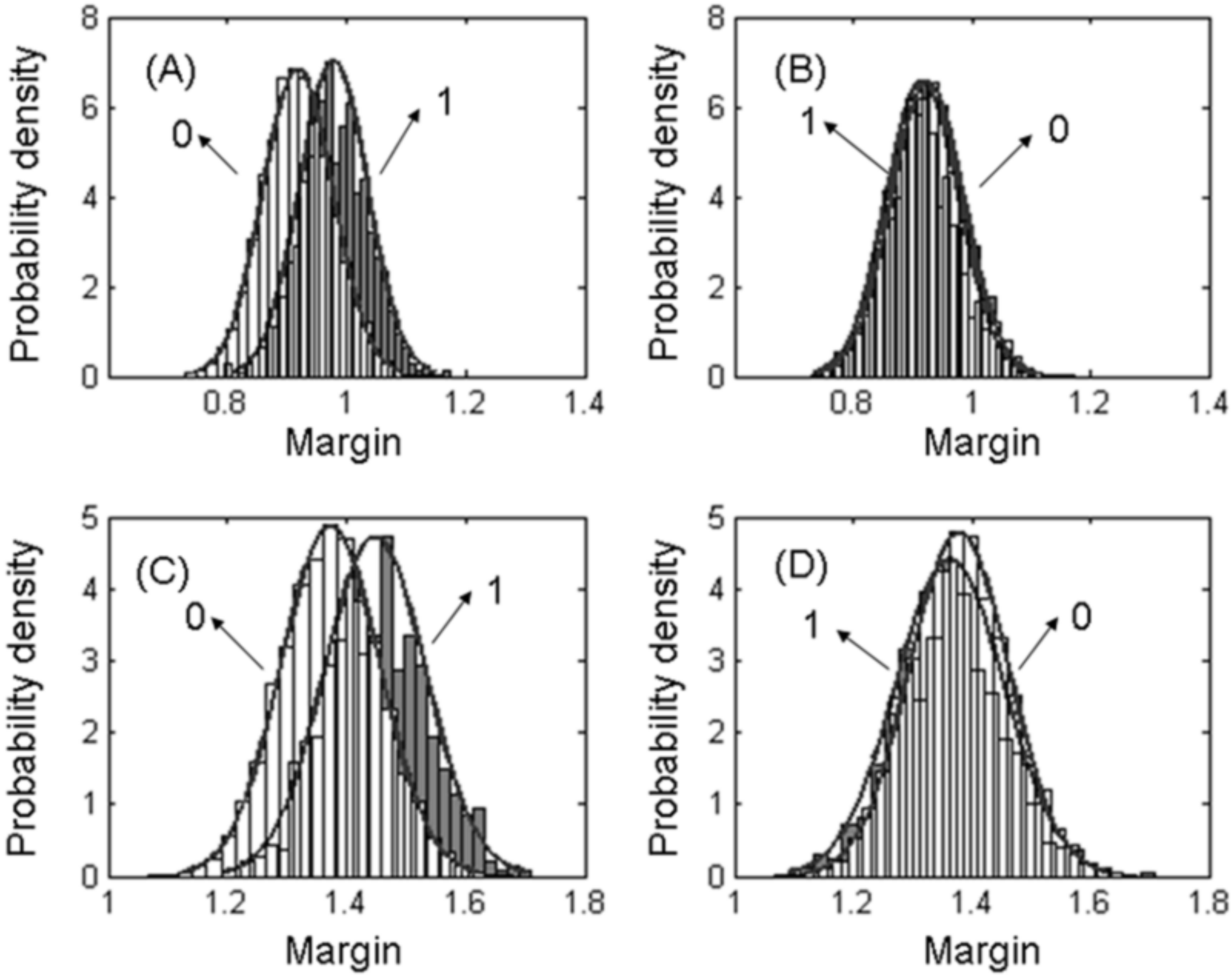
3 Statistical analysis of the margin's distribution



Peak 1: Margins of the Models with the variable included
Peak 0: Margins of the Models without the variable included

Applications

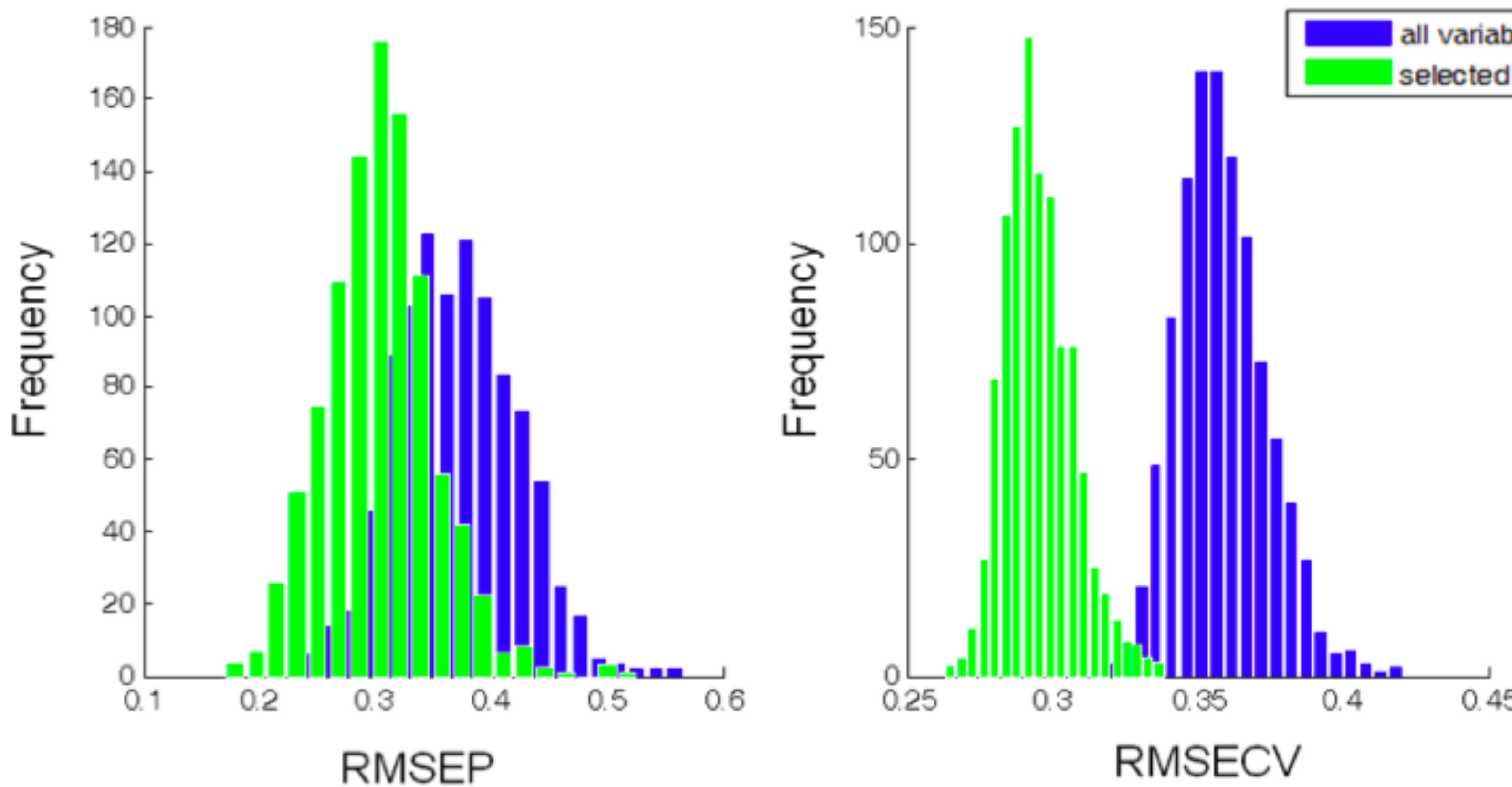
- **Colon data: 62 x 2000**
- **Estrogen data: 49 x 3333**



(A)
(B)
(C)
(D)
0
1
Probability density
Margin
0.8
1
1.2
1.4
1.6
1.8

Model Assessment

Model assessment



Li, H.-D., Liang Y.-Z&Xu, Q.-S, statistical model comparison via model population analysis, **an invited book chapter, *under review***

Features of MPA-based methods:

- The computing process is **random**

- The final output is **stable**

For discussion?

- **Posterior distribution** from Bayesian analysis

- Theoretical analysis of MPA if possible?

Thank you very much