Introduction to Machine Learning Methods in Condensed Matter Physics

LECTURE 9, FALL 2021

Yi Zhang (张亿)

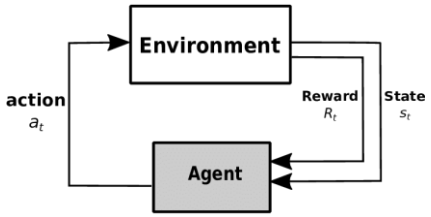
International Center for Quantum Materials, School of Physics
Peking University, Beijing, 100871, China

Email: frankzhangyi@pku.edu.cn

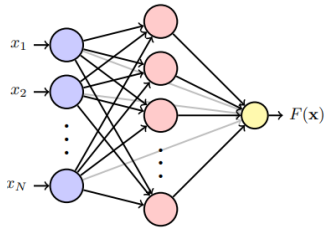
Unsupervised Machine learning

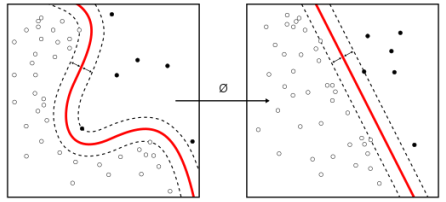
- Machine learning without data (with dynamical environment) → **(1) reinforcement learning**

(1)



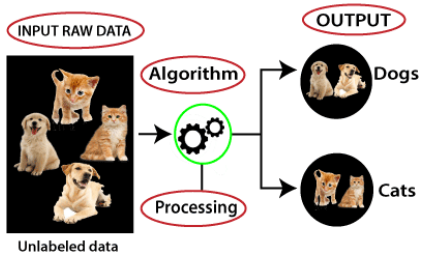
(2)



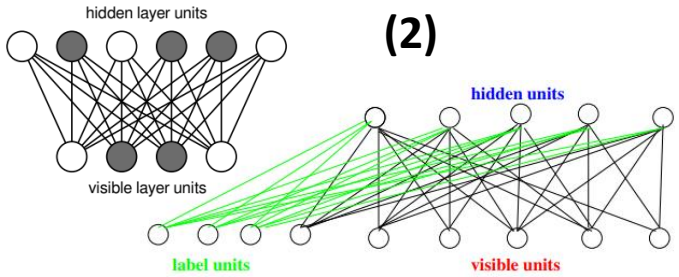


- Machine learning with data (sample input $\boldsymbol{x}$ with desired output label $\boldsymbol{y}$) → **(2) supervised machine learning**, e.g. artificial neural network, support vector machine, etc.
- Machine learning with data (sample input $\boldsymbol{x}$ only) → **(3) unsupervised machine learning**
- Restricted Boltzmann machine can be unsupervised (with no label) or supervised (with label).
- Why unsupervised machine learning?
  - Characterizations and labels unavailable
  - Learning on their own, closer to "real AI"
  - Group data according to similarities → find the underlying structure of dataset, new, useful insights

(3)



(2)



Unsupervised Machine learning

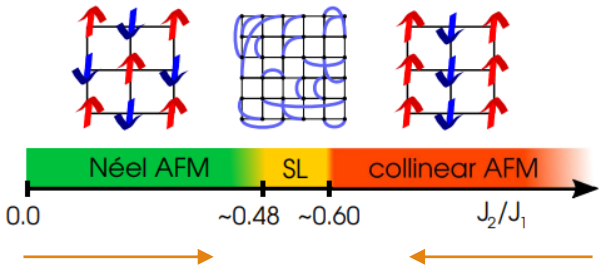
- One may potentially miss out new classifications in supervised machine learning:

Example:

$$\mathcal{H} = J_1 \sum_{\langle R,R' \rangle} \mathbf{S}_R \cdot \mathbf{S}_{R'} + J_2 \sum_{\langle\langle R,R' \rangle\rangle} \mathbf{S}_R \cdot \mathbf{S}_{R'}$$

Intermediate phase:

between two limits



sample

Cluster/group

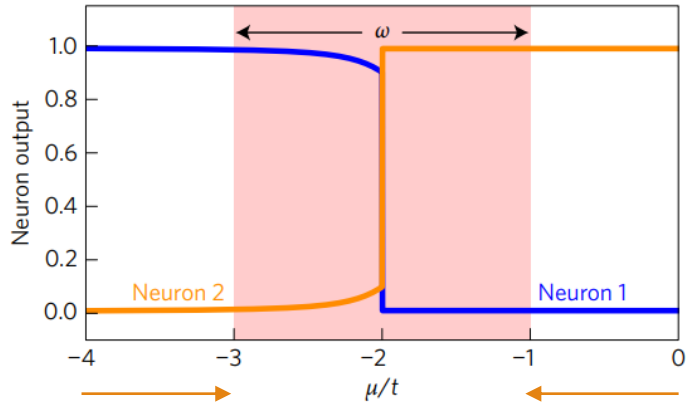
- Types of unsupervised machine learning:
  - Clustering: finding clusters (groups) according to structures or patterns if they exist in the data, e.g. Principal Component Analysis, K-means
  - Association: establish correlations amongst data objects inside large databases, e.g. RBM
- Disadvantages: intrinsically more difficult (less efficiency) and less controlled (less accuracy) than supervised machine learning

Learning by confusion

- A combination of supervised and unsupervised technique: tentatively separate the data into groups, then check the subsequent quality of supervised machine learning
- Example: Kitaev chain

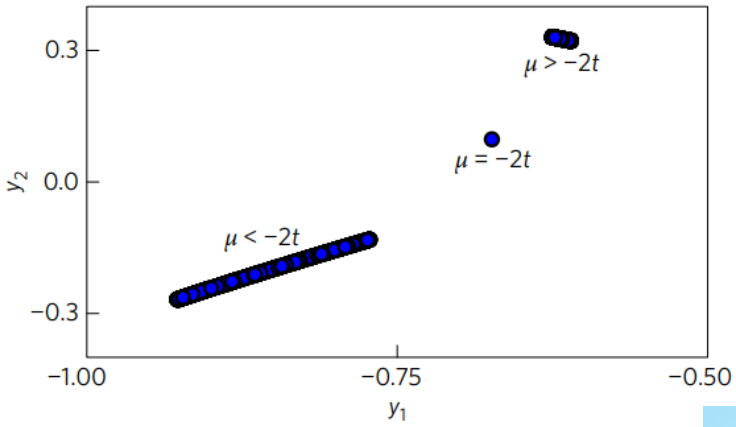
$$\hat{H} = -t \sum_{i=1}^{L} (\hat{c}_{i+1}^{\dagger} \hat{c}_i + \hat{c}_{i+1} \hat{c}_i + \text{h.c.}) - \mu \sum_{i=1}^{L} \hat{c}_i^{\dagger} \hat{c}_i$$

- Supervised Learning:



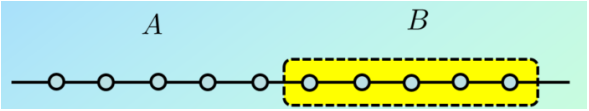
*E. P. L. van Nieuwenburg, Ye-Hua Liu, Sebastian D. Huber, Nature Physics **13**, 435–439 (2017).*

Unsupervised machine learning: (clustering)



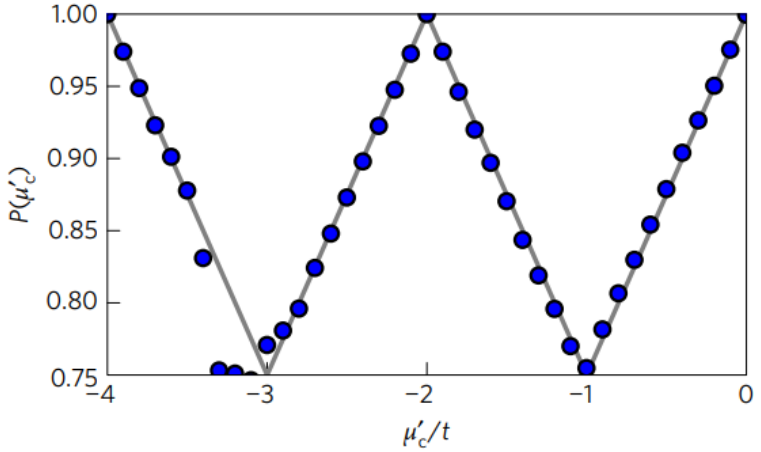
based on ground-state entanglement spectrum: the eigenvalues of the reduced density matrix

$$\rho_A = \mathrm{Tr}_B |\Psi\rangle\langle\Psi|$$

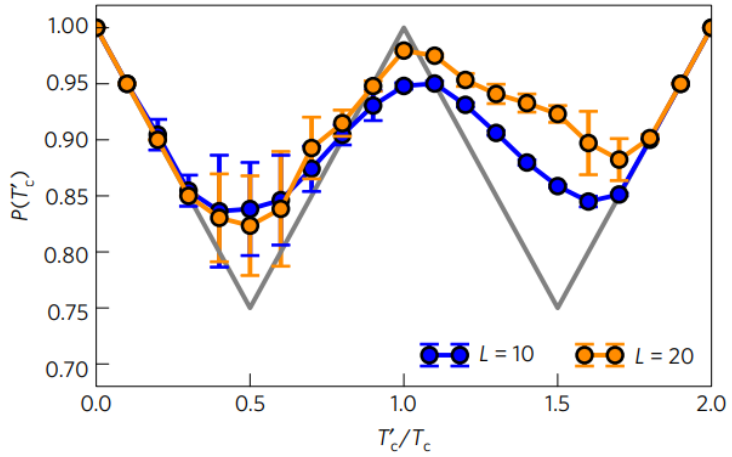


Learning by confusion

- ANN confusion as a signature: tentatively separate the data into groups, then check the subsequent quality of supervised machine learning
- Accuracy is higher when the proposed groups are closer to actual classifications:



Also work for phases of 2D Ising model with dataset of sampled spin configurations:



- Implicitly used knowledge: the phase diagram of a physical system behaves relatively well and simple in the model parameter space → vicinity of and distance between the data samples

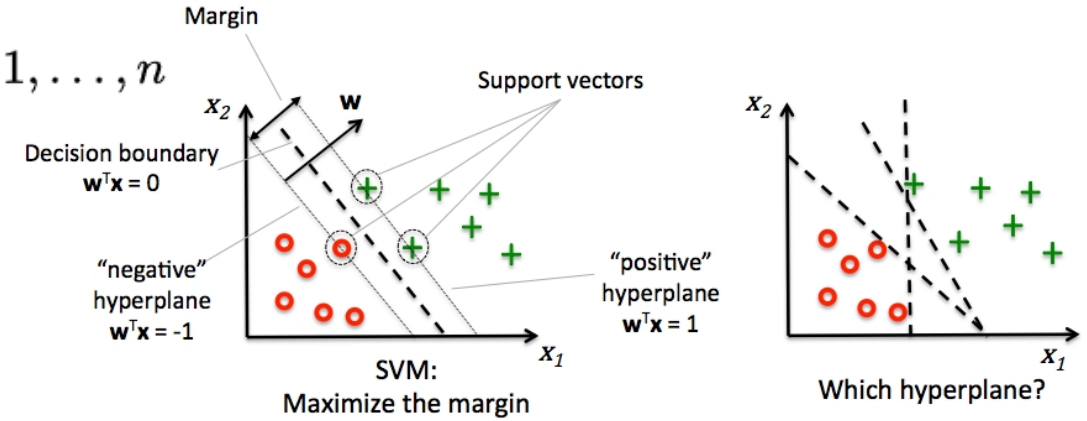
*E. P. L. van Nieuwenburg, Ye-Hua Liu, Sebastian D. Huber, Nature Physics **13**, 435–439 (2017).*

Support Vector Machine

- A supervised learning method, and also a stepping stone to the unsupervised machine learning
- At the linear level: maximizing the margin – the gap between the closest data points

➡ Minimize $\|\mathbf{w}\|$ subject to $y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1$ for $i = 1, \ldots, n$

For hyperplane defined by the slope $\boldsymbol{w}$ and intercept b: $\mathbf{w}^T\mathbf{x} - b = 0$



- Soft margin: (λ controls stiffness)

$$\lambda\|\mathbf{w}\|^2 + \left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b)\right)\right]$$

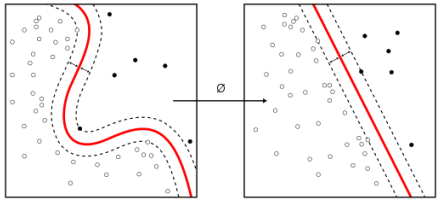
↖ tackle with gradient descent

- Advantages of SVM:
  - Deterministic
  - Convex optimization
  - Interpretability

*Similar to perceptron – ANN with no hidden layer

Kernel-based Support Vector Machine

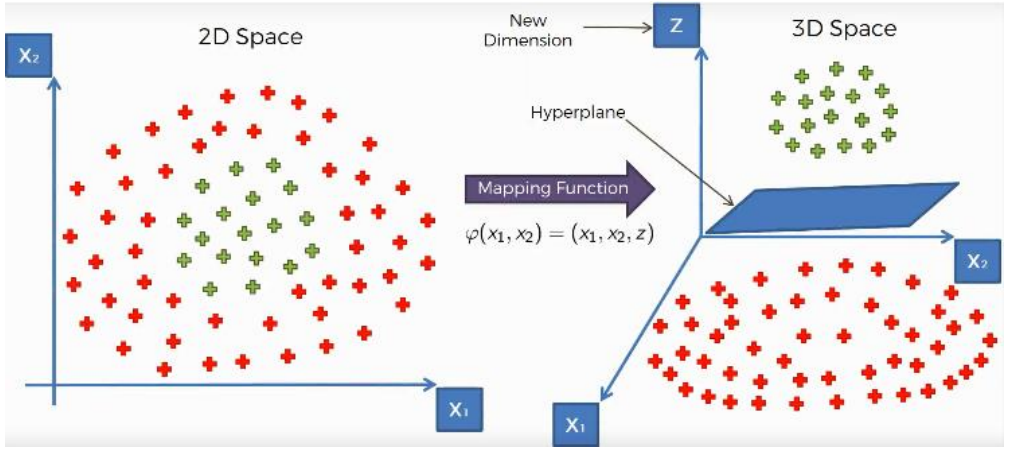
- Kernel trick: transforming data into higher-dimensional space where a linear separation is straightforward



- For nonlinear separations:

$$\mathbf{x}_i \Rightarrow \varphi(\mathbf{x}_i)$$

For example: $\varphi((a, b)) = (a, b, a^2 + b^2)$



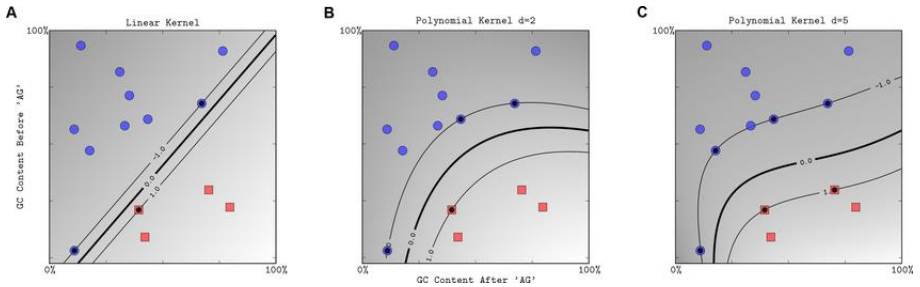
- Good kernel choices improve SVM performance.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$$

- Example: MINST data of 28x28 hand-written digits

- Accuracy: Linear SVM (91.4%)

RFB SVM (~98.6%), Polynomial SVM (~99%)



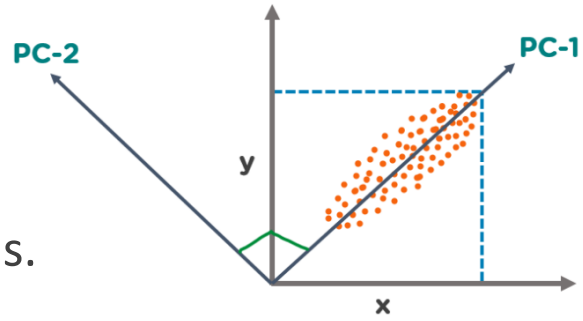
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2) \qquad K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}')^d$$

Principle component analysis

- Dimension reduction to find the most significant features in a dataset and better visualization.

  The vectors that capture the most variance of the data:

  Why? The larger the variance carried by a line, the larger the dispersion of the data points along it, the more the information it has.



- How to select the principle components? Orthogonal transformation!
- Step1: normalization

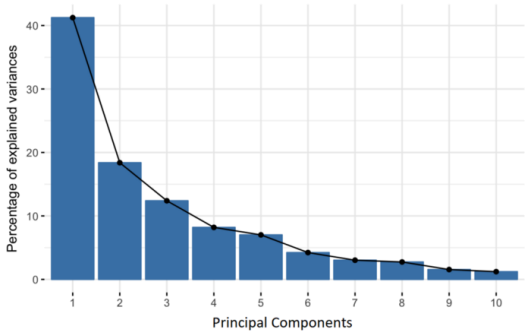
$$z = \frac{value - mean}{standard\ deviation}$$

- Step2: the co-variance matrix

Dimension of the data:

over all data samples

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$



- Step3: diagonalization and sort in descending order of eigenvalues

Principle component analysis

- Step 4. collapse (project) original data onto the new, featuring directions (eigenvectors):

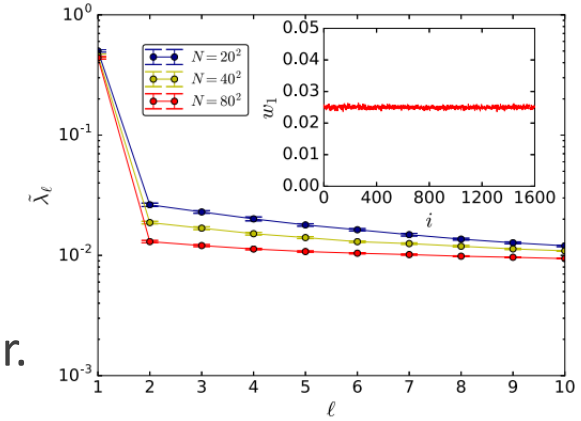
$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

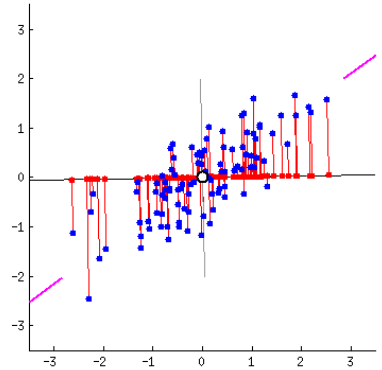
- Example: PCA on 2D classical Ising model

  variance of the principle components:

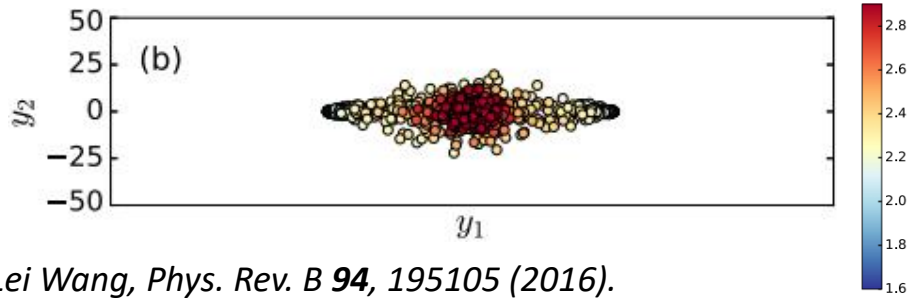
  weights of the first principle component:

  Average magnetization as order parameter.





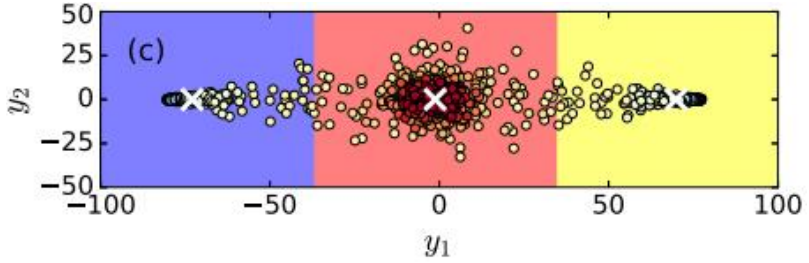
- Visualization of the data at different temperatures with the two principle components:



*Lei Wang, Phys. Rev. B **94**, 195105 (2016).*

Larger systems with K-means clustering

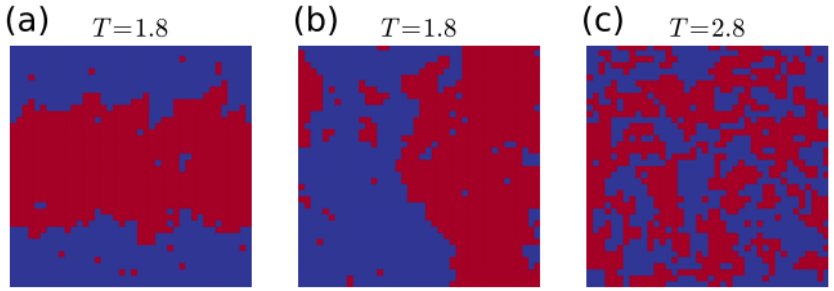
$T_c/J \approx 2.3$



Principle component analysis applications

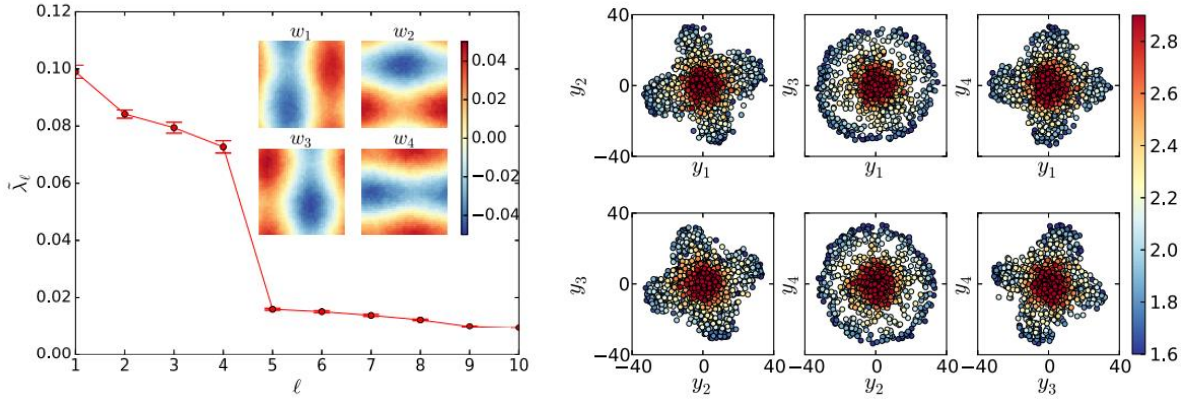
- More challenging example: classical Ising model with constraints (conserved order parameter)

$$H = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j \qquad \sum_i \sigma_i \equiv 0 \qquad T_c / J \approx 2.3$$



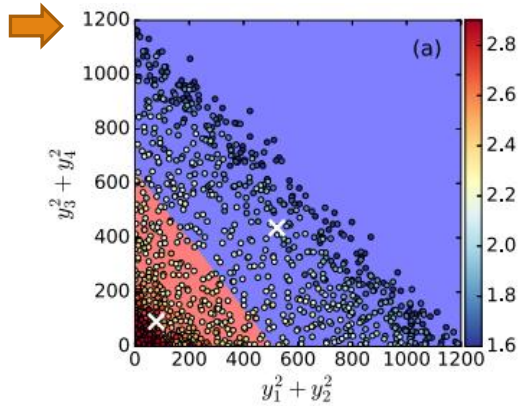
- Domains, yes. average magnetization, no.

- Dominance and weights of the leading four:



Data projected onto the four leading principle components

Reduction to two dimensions



PCA is a simple, linear method after all.

*Lei Wang, Phys. Rev. B **94**, 195105 (2016).*