Sparse Coding for Feature Selection on Genome-wide Association Data

Ingrid Brænne

Institute for Neuro- and Bioinformatics
Department of Internal Medicine II,
Graduate School for Computing in Medicine and Life Sciences
University of Lübeck, Germany

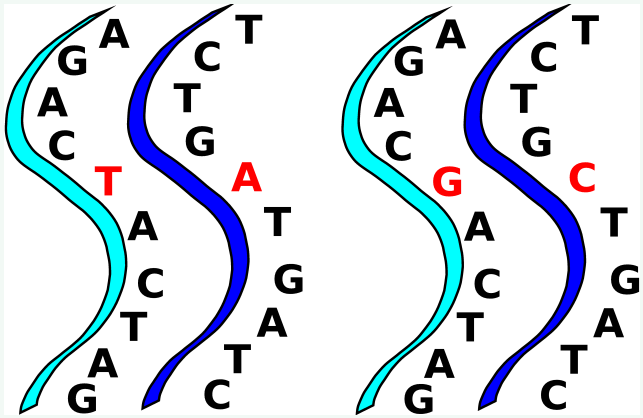
ICANN 2010, Thessaloniki, Greece, 15.09.2010

Graduate School for Computing in Medicine and Life Sciences

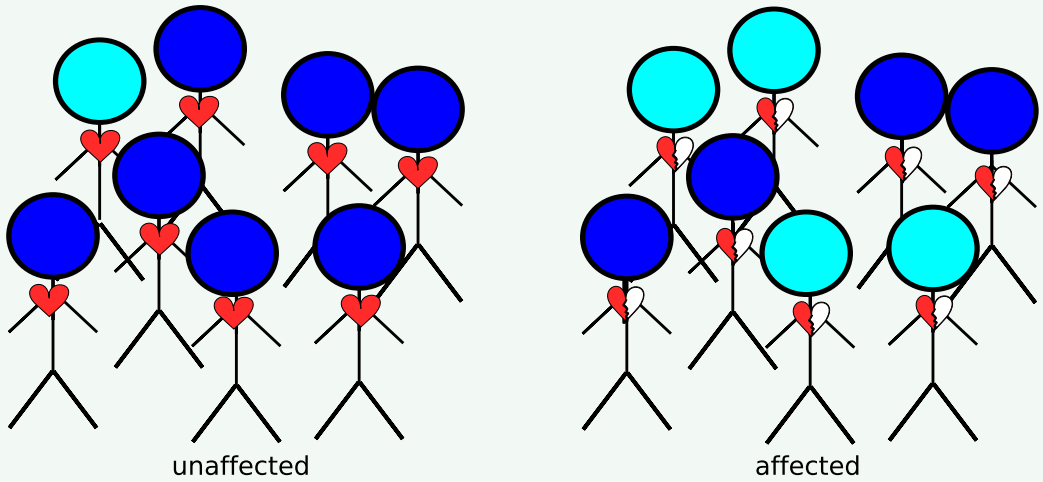
Genome Data

Single Nucleotide Polymorphism (SNP)

- Different bases at one locus in the genome
- Occur in at least 1% of the population
- Three possible genotypes (-1, 0, 1)
- Used as genetic markers



Genome-wide Association (GWA) studies



- Large numbers of SNPs
- Identify genetic loci
- Reveal genetic mechanisms
- Risk prediction

What makes it difficult?

- High-dimensional data (hundred thousand SNPs)
- Few data points (a few thousand)
- Discrete data (1,-1,0)
- Noisy data

Hypothesis

What do we expect?

- Disease effect through interaction
- No uniform pattern among all cases
- Disease pattern consisting of SNPs without individual effects
- Disease-unspecific patterns

Standard Approaches

- Single association test (p-values) [1]
- Support Vector Machine (SVM) [2]
- Principal Component Analysis (PCA) [3]

[1] Sekar Kathiresan, N Engl J Med,1240-1249, 2008
[2] Zhi Wei, PLoS Genetics,5(10):e1000678+,2009
[3] Peristera Pashou, PLoS genetics, 3(9) 1672-1686,2007

Standard Approaches

Association test

- Chi-squared distribution
- Feature selection by the p-values

SVM

- Linear hard-margin SVM
- Feature selection by the influence on the margin

Principal Component Analysis

Aim: cover as much of the variability of the data as possible

- Separate analysis for cases and controls
- M principal components

$$V^{class} = (\mathbf{v}_1^{class}, \ldots, \mathbf{v}_M^{class}) \tag{1}$$

- Feature selection by:

$$r_j = |\max_i |(\mathbf{v}_i^1)_j| - \max_i |(\mathbf{v}_i^{-1})_j|| \tag{2}$$

Sparse Coding

Aim: represent data $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $x_i \in \mathbb{R}^D$, as a sparse linear combination of a dictionary C and coefficient vectors a_i, $\|\mathbf{a}\|_0 \leq k$.

$$\mathbf{x}_i = C\mathbf{a}_i$$

Sparse Coding

Goal: Minimize the reconstruction error: $\frac{1}{L}\sum_{i=1}^{L}\|\mathbf{x}_i - C\mathbf{a}_i\|_2^2$

- Problem 1: find optimal C
  –> Solution: Bag of Pursuits and Neural Gas [1]

- Problem 2: for given C find optimal a_i
  –> Solution: Bag of Pursuits [2]

$$\mathbf{x}_i^{\text{opt}} = C\mathbf{a}_i \text{ with } \mathbf{a}_i = \arg\min_{\mathbf{a}} \|\mathbf{x}_i - C\mathbf{a}\| \,, \|\mathbf{a}\|_0 \leq k$$

[1] K. Labusch, E. Barth, T. Martinetz, Proc. COMPSTAT, 327-336,2010
[2] K. Labusch, T. Martinetz, Proc. ESANN, 241-246, 2010

Sparse Coding

- separate analysis for cases and controls

$$C^{class} = (\mathbf{c}_1^{class}, \ldots, \mathbf{c}_M^{class})$$

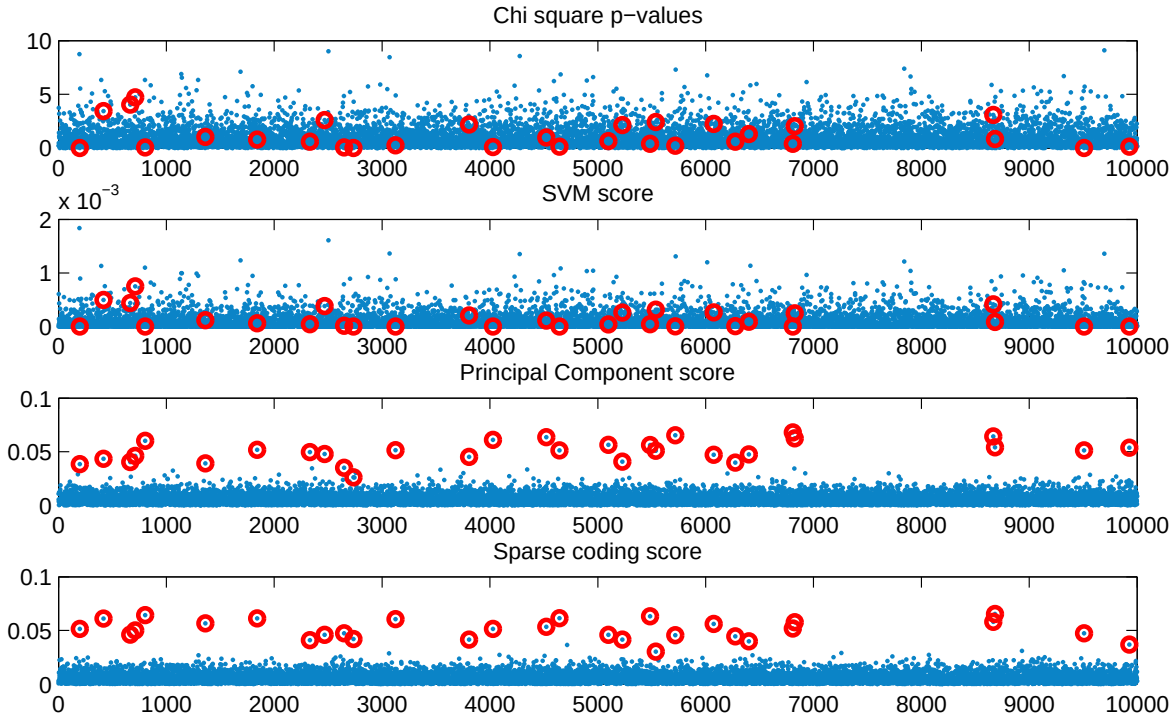
- feature selection by:

$$r_j = |\max_i |(\mathbf{c}_i^1)_j| - \max_i |(\mathbf{c}_i^{-1})_j||$$

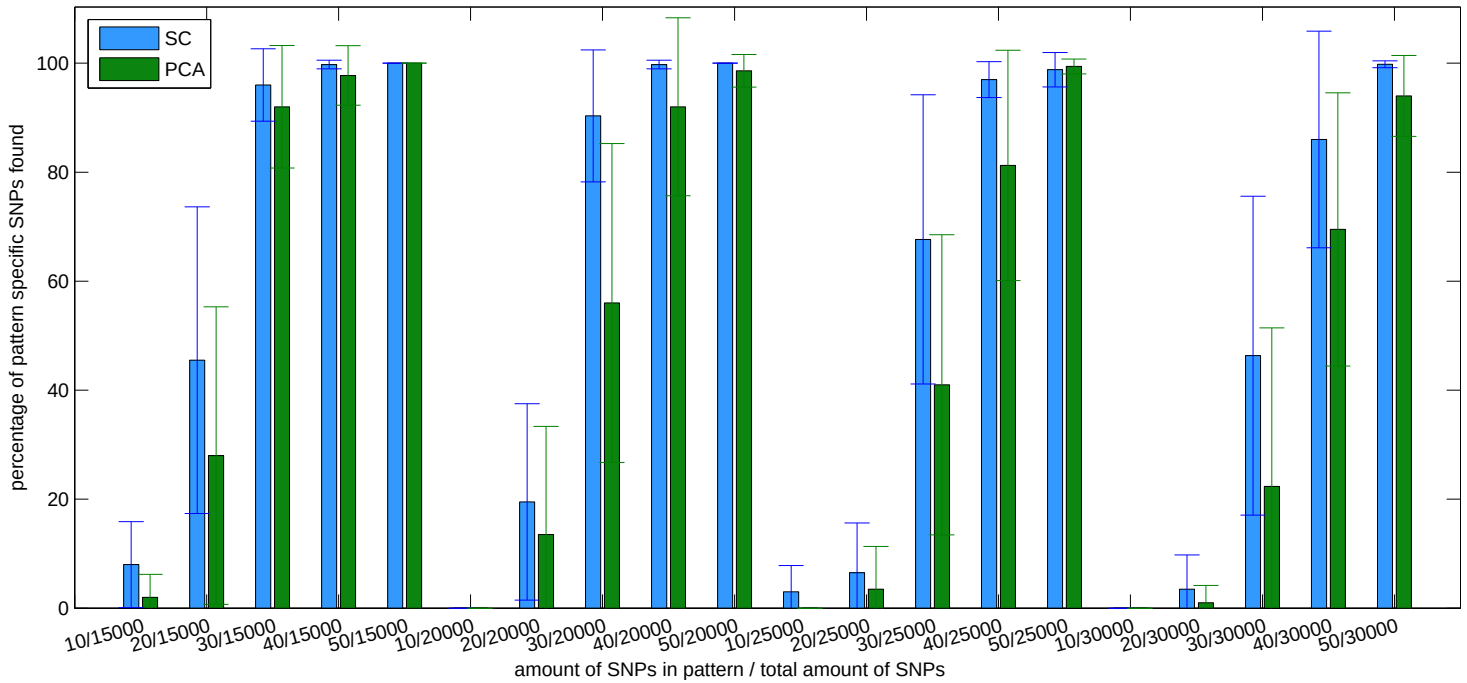
Simulated Data

- 2 types of Data sets
  - 1 disease pattern 1 unspecific pattern
  - 1 disease pattern 5 unspecific patterns
- Pattern size: SNPs: $10, 20, ..., 50$, individuals: 100
- SNPs: $15000, 20000, ..., 30000$
- Individuals: 500 cases and 500 controls

Comparing feature selection approaches

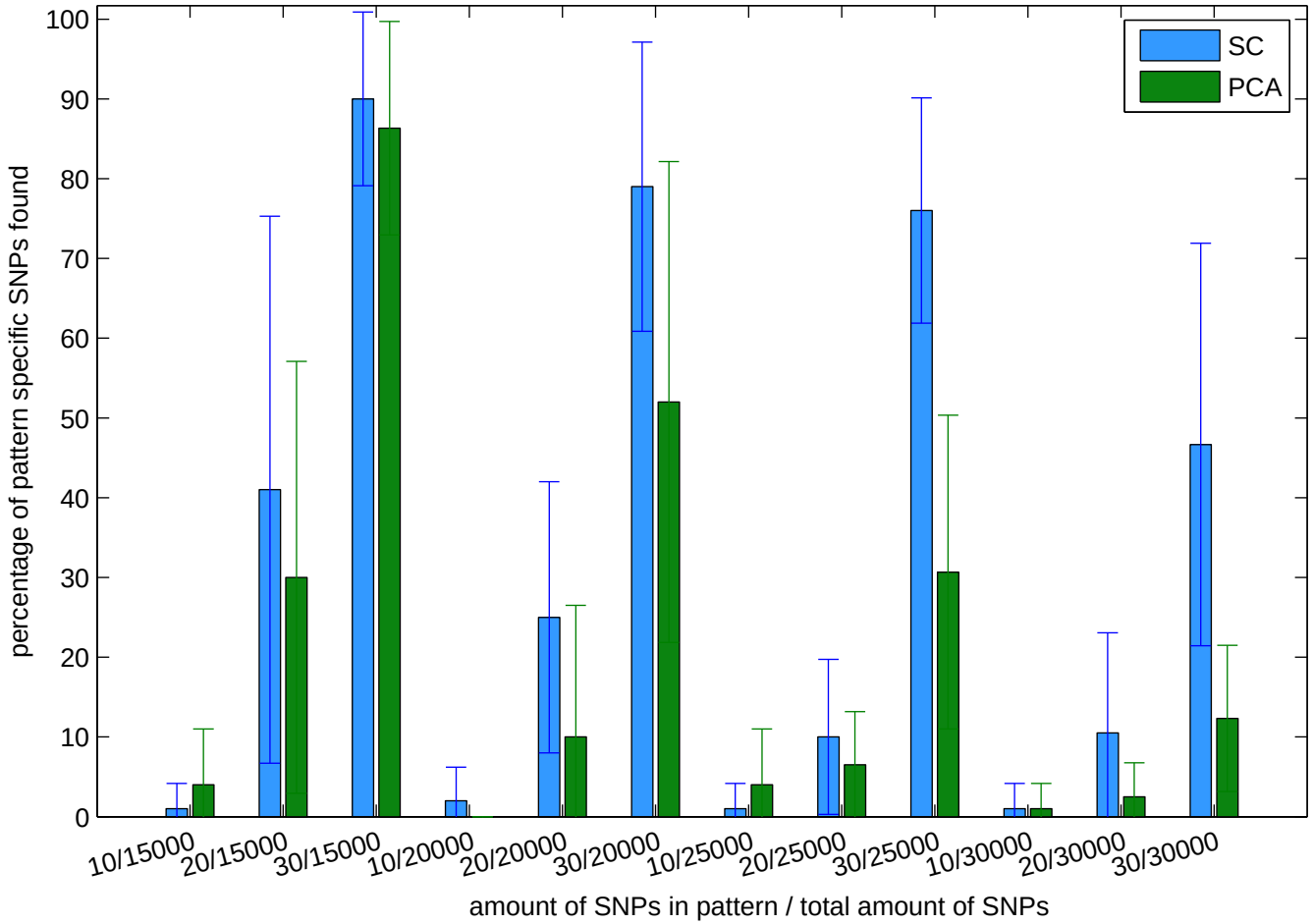


Sparse Coding versus PCA



1 disease-specific pattern, 1 unspecific pattern

Sparse Coding versus PCA



1 disease-specific pattern, 5 unspecific patterns

Conclusion

- p-values or SVM are not suitable for feature selection
- PCA and Sparse Coding suitable for the task
- Sparse Coding more robust to noise pattern
- Sparse Coding promising for real data

Thank you for your attention