ANALYSIS FOR DIABETES USING GRAPHICAL LASSO

Y.K. Kim*, Y. Yun, M. Yoon & H. Nakayama

Graduate School of Science and Engineering

Kansai University, Japan

Contents

- Background and Purpose
- Basic Analysis
  - T-test
  - correlation analysis
- Graphical Lasso and Structure Analysis
- Importance in Support Vector Machines
- Conclusion

Background

- Recently, approximately 60% of total deaths are caused by lifestyle diseases.
- Especially, diabetes may affect serious illnesses, for example, cerebral infarction and myocardial infarction.

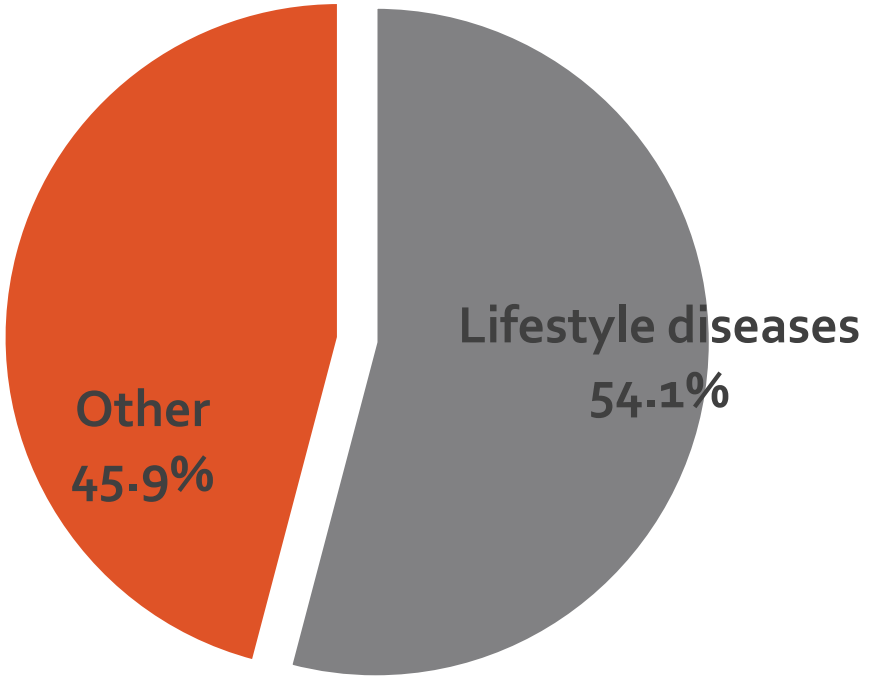


Fig.1 : The demographic statistics of Japan, 2015
(source : Demographic Statistics, Ministry of Health, 2015

Related works

- *Predictive modeling and analytics for diabetes using a machine learning approach*

  V. Kumari and H. Kaur

  Applied Computing and Informatics, 2008

- *A method for classification using machine learning technique for diabetes*

  R. Aishwarya, P. Gayathri, et al.

  International Journal of Engineering and Technology, 2013

- *Prediction of diabetes using classification algorithm*

  D. Sisodia and D. S. Sisodia

  Procedia Computer Science, 2018

Purpose

For Pima Indians diabetes data,

- to investigate whether a structure change exists between data for diabetics and for non-diabetics by **using graphical lasso**
  - to compare a **direct correlation** between factors
  - to evaluate a **change score** between non-diabetics and diabetics
- to investigate **importance for each factor** for detecting diabetes **by using SVM**

⇨ *It will be expected for increasing the effectiveness in disease prevention and health promotion*

Graphical Lasso

- Data set $D = \left\{ \boldsymbol{x}^{(i)} \;\middle|\; i = 1, \ldots, l \right\}, \boldsymbol{x} \in \mathbb{R}^m$
- $m-$ dimensional multivariate normal distribution

$$\mathrm{N}(\boldsymbol{x} \mid 0, \Lambda^{-1}) = \frac{(\det \Lambda)^{1/2}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} \boldsymbol{x}^{\boldsymbol{T}} \Lambda \boldsymbol{x} \right)$$

In the graphical lasso, the precision matrix $\Lambda := \Sigma^{-1}$ *is estimated*

*by the following t*he maximum likelihood method with an L_1 regularization.

$$\Rightarrow \Lambda^* = \arg\max_{\Lambda} \left(\ln \det \Lambda - \mathrm{tr}(S\Lambda) - \boldsymbol{\rho} ||\boldsymbol{\Lambda}||_{\boldsymbol{1}} \right)$$

making a sparse structure depending on ρ

S : covariance matrix for given data

$\rho > 0$: given regularization parameter

Structure analysis by graphical lasso

- Precision matrix $\Lambda = (\lambda_{ij})$ $\Rightarrow$ **adjacency matrix**

  which yields a direct correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$

$$\Lambda = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$

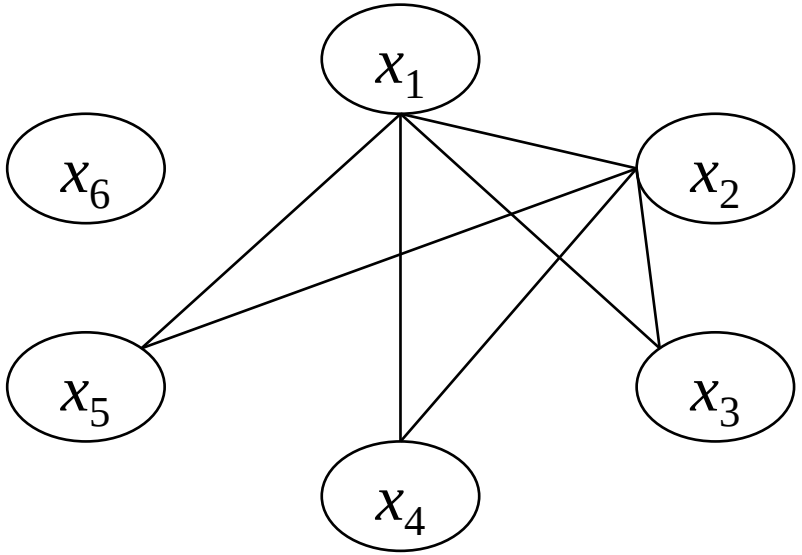


Fig. 2 : Correlation graph between x_i and x_j based on an adjacency matrix

Structural difference for two data sets

$$\Lambda = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$

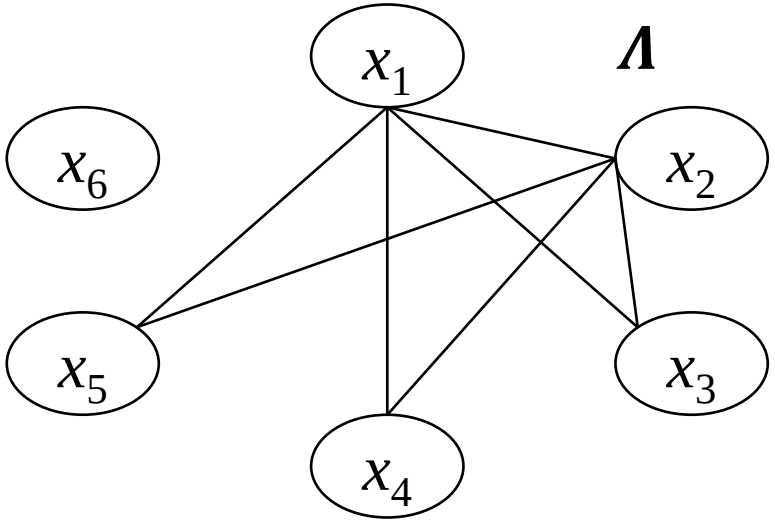


Fig.3 : Correlation for a dataset 1

VS.

$$\Lambda' = \begin{pmatrix} * & 0 & * & 0 & * & * \\ 0 & * & 0 & * & * & 0 \\ * & 0 & * & * & 0 & 0 \\ 0 & * & * & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & 0 & * \end{pmatrix}.$$

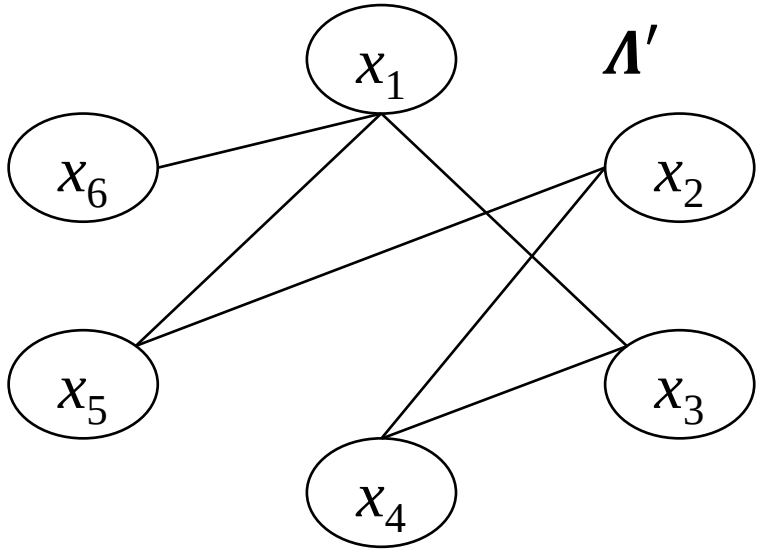


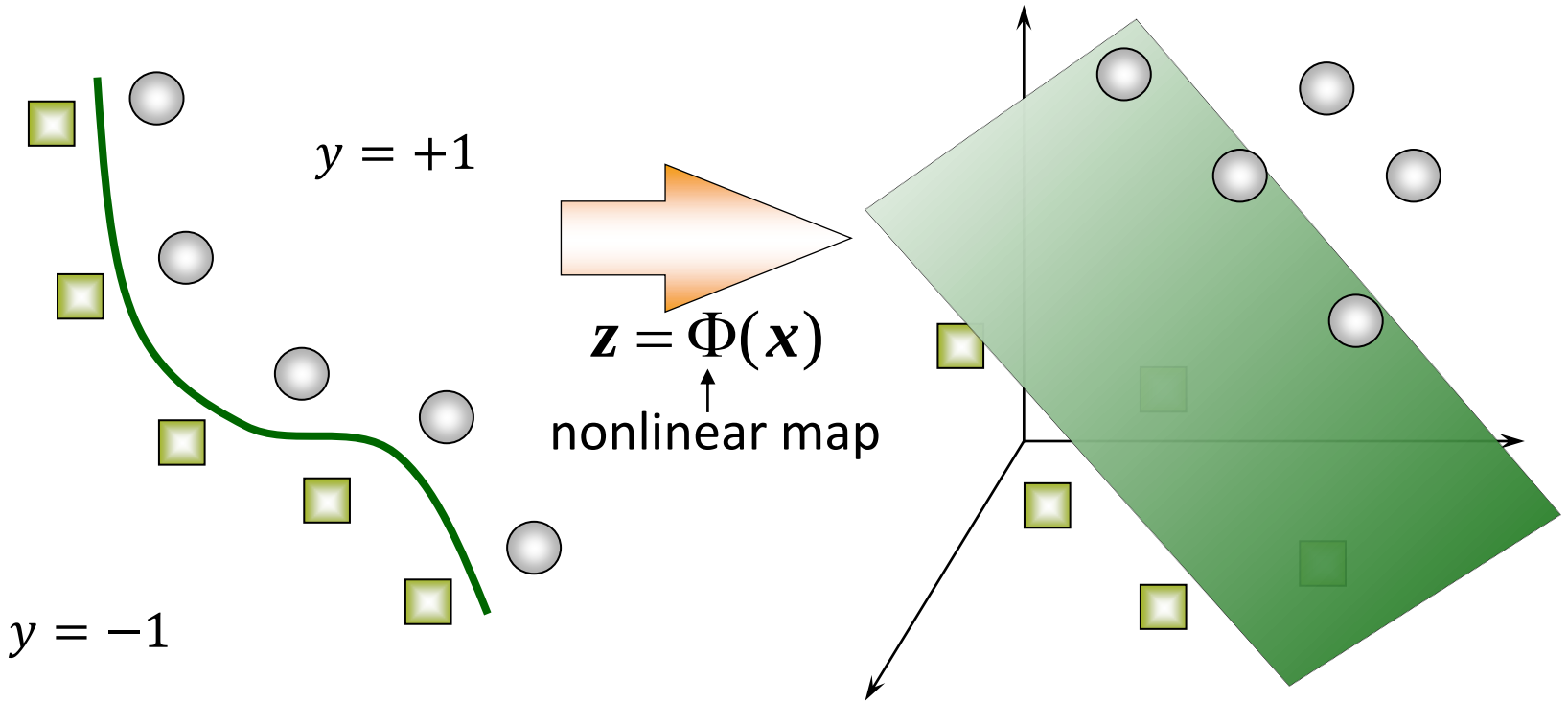
Fig.4 : Correlation for a dataset 2

Support Vector Machines (SVM)

- Powerful tool for machine learning
- Proposed by Vapnik and Chervonenkis in statistical learning theory
  - classification (Cortes & Vapnik, 1995)
  - regression (Vapnik et al., 1996)
- Main features
  - evaluation of generalization ability by VC-dimension
  - maximal margin linear classifier on the feature space
  - kernel representation

Support Vector Machines (SVM)

- Mapping data from input space to feature space



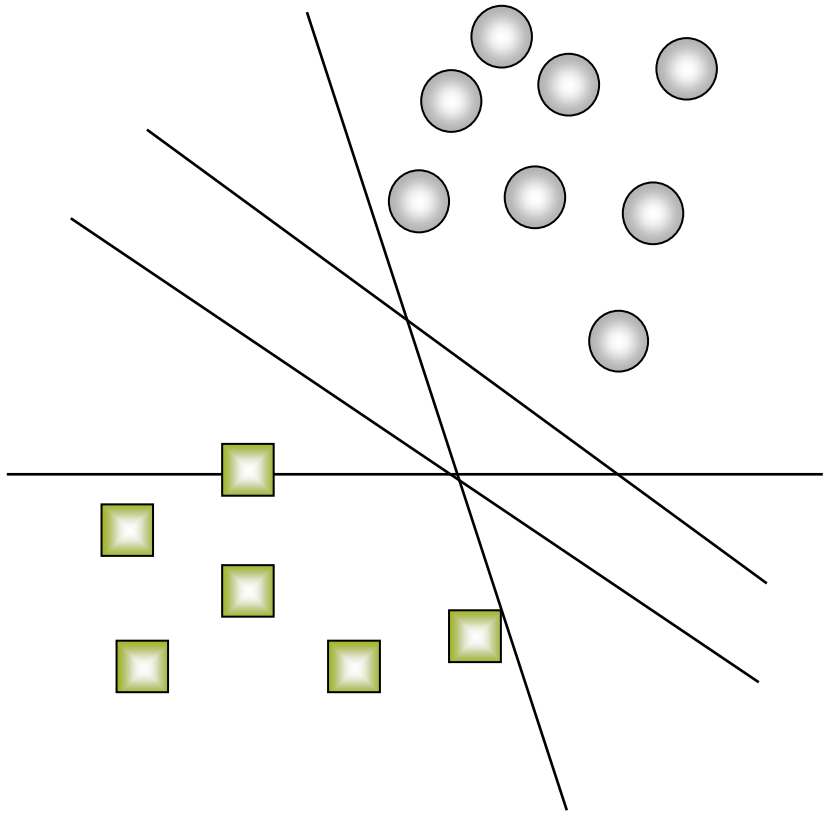
X : original input space

＜linearly non-separable＞

Z : high dimensional feature space

＜linearly separable＞

Support Vector Machines (SVM)



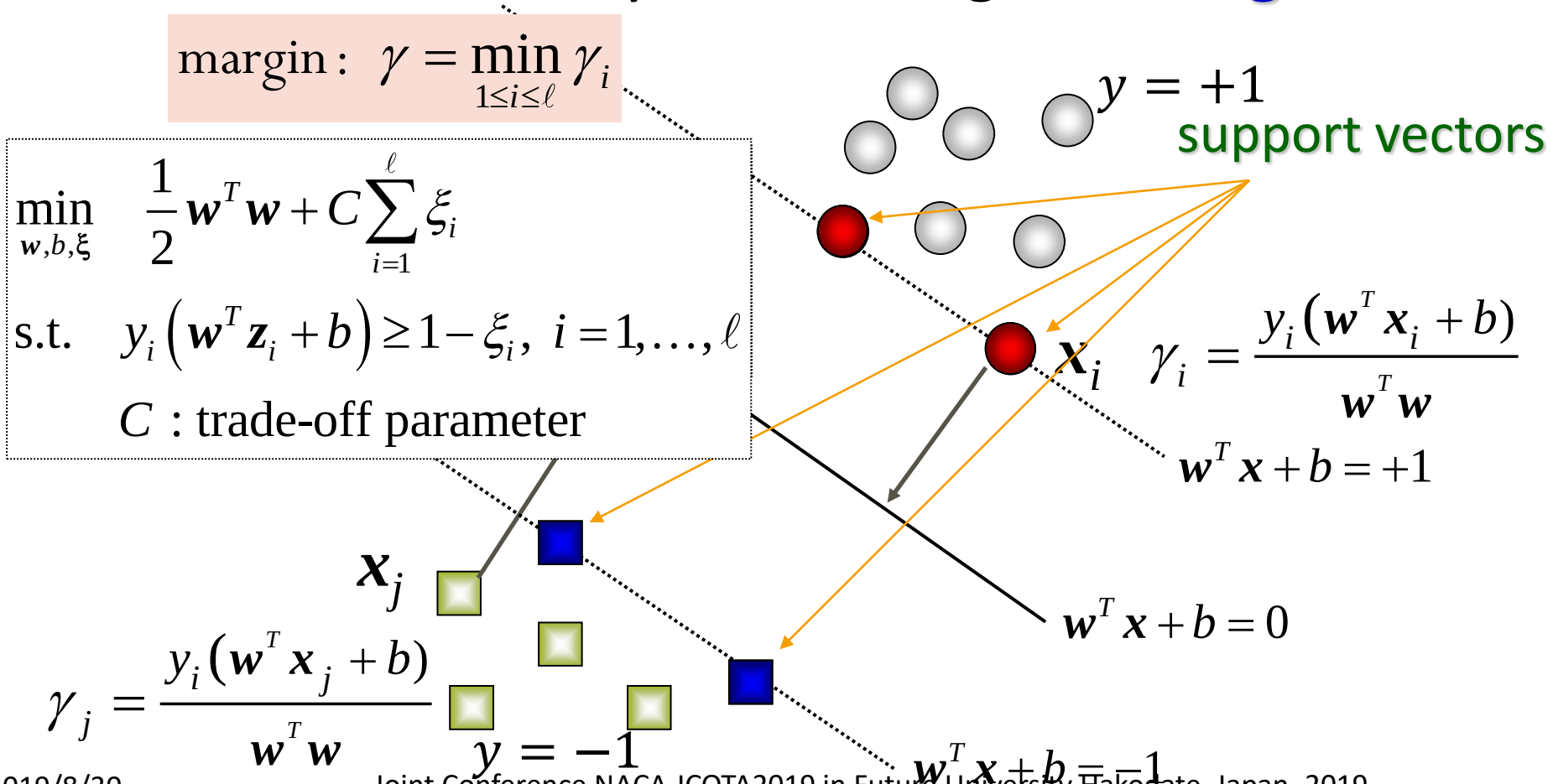
There exist many hyperplanes to separate two classes

Which one should we choose?

How can we find an optimal one?

Margin in SVM

- SVM finds a separating hyperplane far away as possible from two classes by maximizing the **margin**.



Overview of data

- Pima Indians diabetes (1990) downloaded https://www.kaggle.com/
  - 768 women with 8 factors
  - # non-diabetics = 500, # diabetics = 268

factor	description	Value
Pregnancies	Number of times pregnant	-
Glucose	Plasma glucose concentration in a 2 hours in oral glucose tolerance test	mg/dl
Blood Pressure	Diastolic blood pressure	mm Hg
Skin Thickness	Triceps skin fold thickness	mm
Insulin	2-hour serum insulin	mu U/ml
BMI	Body mass index	kg/m^2
Diabetes Pedigree Function	Diabetes Pedigree Function	-
Age	Years	-

Flow of the analysis

T-test and Correlation analysis

Analysis between factors by Graphical Lasso

Evaluation for importance
of each factor by SVM

Comparison (1) for averages by T-test

factor	T-value	P-value
Pregnancies	-5.91	0.00
Glucose	-13.75	0.00
Blood Pressure	-1.71	0.09
Skin Thickness	-1.97	0.05
Insulin	-3.30	0.00
BMI	-8.62	0.00
Diabetes Pedigree Function	-4.58	0.00
Age	-6.92	0.00

Comparison (2) for correlation matrices

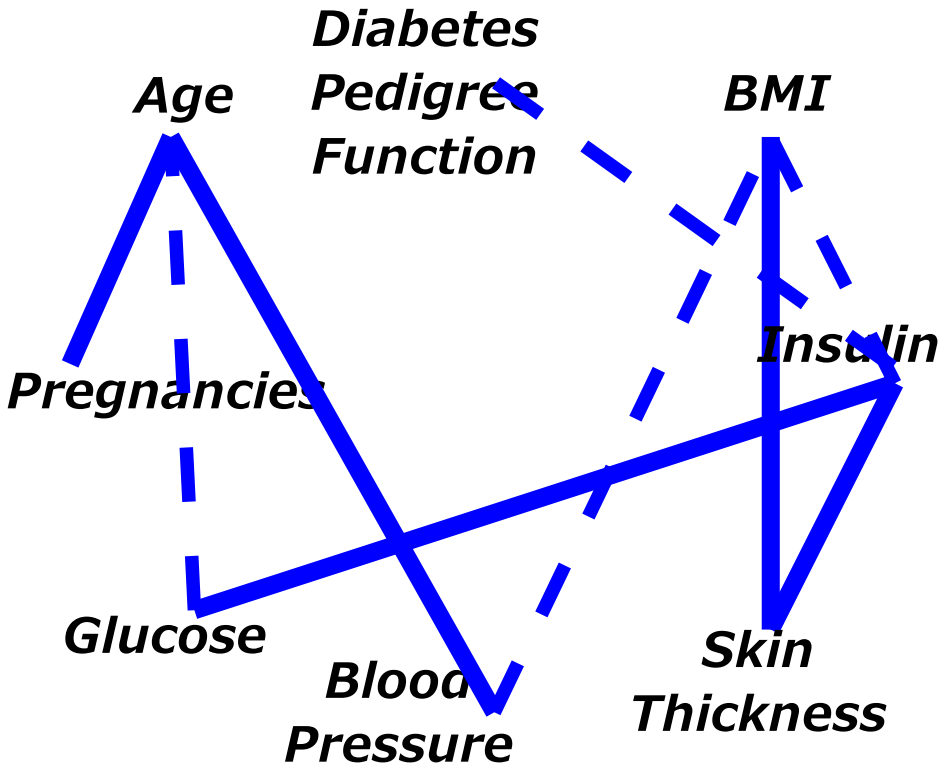
⇒ no indirect correlation between factors

	Pregnancies	Glucos					Function	
Pregnancies	-	0.10	0.13	-0.12	-0.13	0.02	-0.08	**0.57**
Glucose	-0.05	-	0.19	0.02	0.35	0.13	0.10	0.23
Blood Pressure	0.13	0.07	-	0.19	0.07	0.36	0.03	0.21
Skin Thickness	-0.08	0.04	0.23	-	0.41	0.44	0.10	-0.16
Insulin	-0.08	0.26	0.09	**0.46**	-	0.25	0.23	-0.15
BMI	-0.16	0.05	0.13	0.31	0.06	-	0.07	0.04
Diabetes Pedigree Function	-0.07	0.03	0.03	0.27	0.10	0.14	-	0.04
Age	**0.44**	0.10	0.26	-0.09	0.02	-0.19	-0.09	-

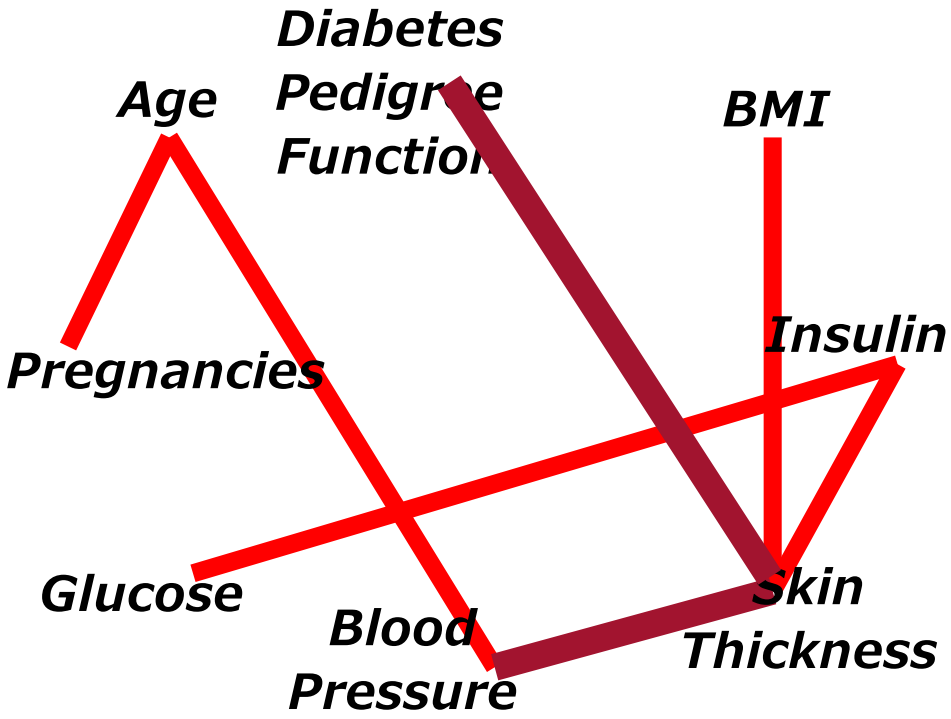
For Non-diabetics

For Diabetics

Comparison (3) for structures by graphical lasso



(a) Non-diabetics



(b) Diabetics

Fig. 5 : Structures based on adjacency matrices

: the relation not appearing in diabetics

: the new relation not appearing in non-diabetics

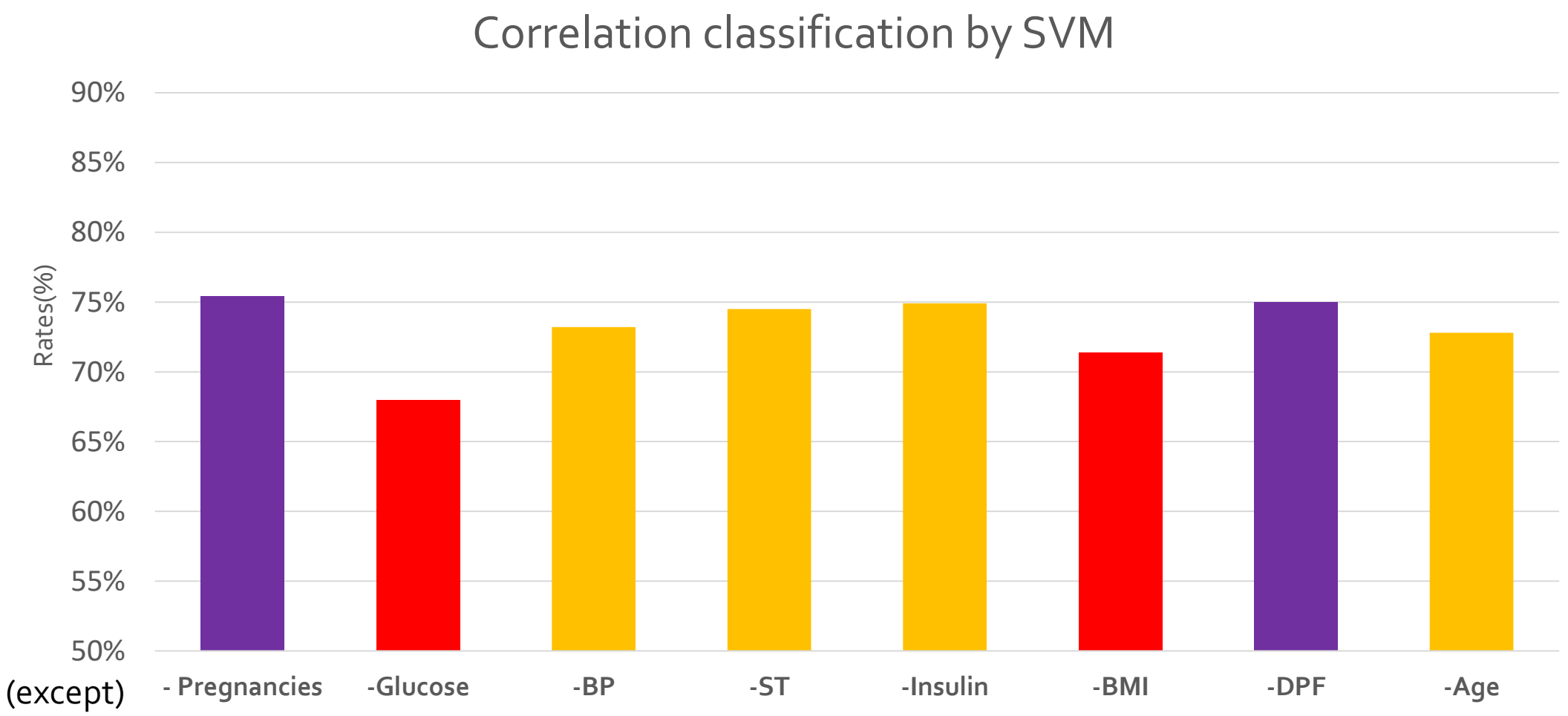
Change Score*

defined by $a_i := \frac{1}{2}\ln\frac{\Lambda_{ii}}{\Lambda'_{ii}} - \frac{1}{2}\left\{\frac{[\Lambda S\Lambda]_{ii}}{\Lambda_{ii}} - \frac{[\Lambda' S\Lambda']_{ii}}{\Lambda'_{ii}}\right\}$

Source :
Change detection from heterogeneous data sources, Tsuyoshi ide, 2014

factor	score
Pregnancies	0.039
Glucose	0.023
Blood Pressure	0.023
Skin Thickness	0.015
Insulin	0.014
BMI	**0.063**
Diabetes Pedigree Function	0.011
Age	0.037

Importance for each factor in SVM



Summary

- The change score in **BMI** is the highest(change score = **0.063**) among all factors between data.
- On the other hand, **Diabetes Pedigree Function** shows that change score between non-diabetics and diabetics is not high.
- Also, Pregnancies (correlation value = **0.44**) is due to a **high correlation value** with age, change score is high.
- The results of t-test and SVM show that **BMI and Glucose** are **important factor** for diabetes diagnosis.
- On the other hand, **pregnancies and diabetes pedigree function** are **not so important factor** in the diagnosis of diabetes.

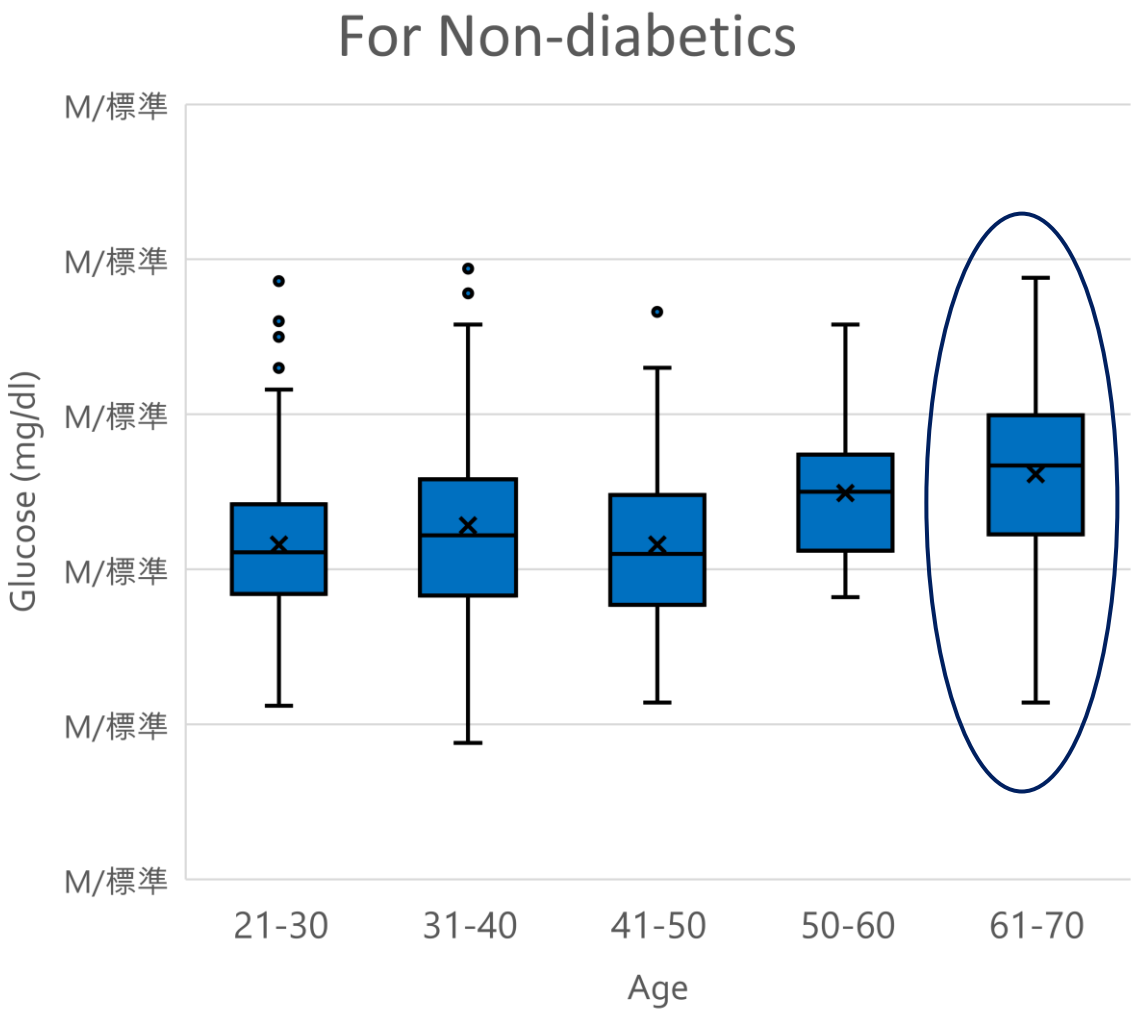
Conclusion

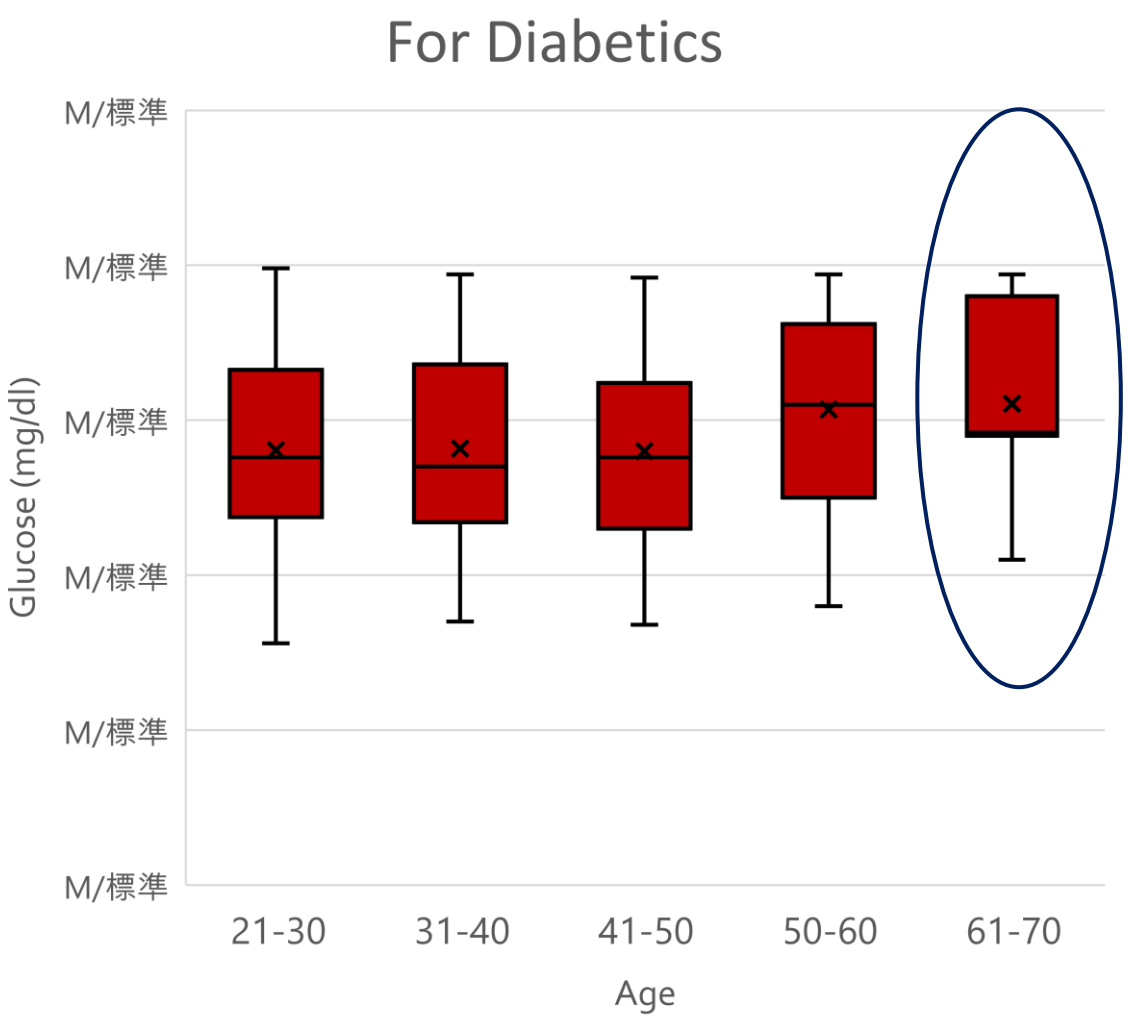
- The results suggest that BMI is one of the most influential factors for diabetes diagnosis.
- In the future, we will consider not only the graphical lasso but also other methods to compare the important factors of diabetes diagnosis.
- In addition, we are going to apply graphical lasso to feature selection in SVM.

Reference

- Demographic statics, Ministry of Health, 2015.
- Data Health, Ministry of Health, 2017
- Anomaly Detection and Change Detection, Tshuyoshi Ide, 2009
- Sparse gaussian markov random field mixtures for anomaly detection, Tshuyoshi ide, 2016
- Pima Indians Diabetes Database, Kaggle, 2007 https://www.kaggle.com/

Age vs. Glucose





Skin Thickness vs. Diabetes Pedigree Function