Application of Graphical Lasso to Change Detection for Diabetes

Y.K. Kim*, Y. Yun, M. Yoon & H. Nakayama
Graduate School of Science and Engineering
Kansai University, Japan

Contents

- Background and Purpose
- Basic Analysis
- Graphical Lasso and Structure Analysis
- Importance for Factors by SVM
- 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.

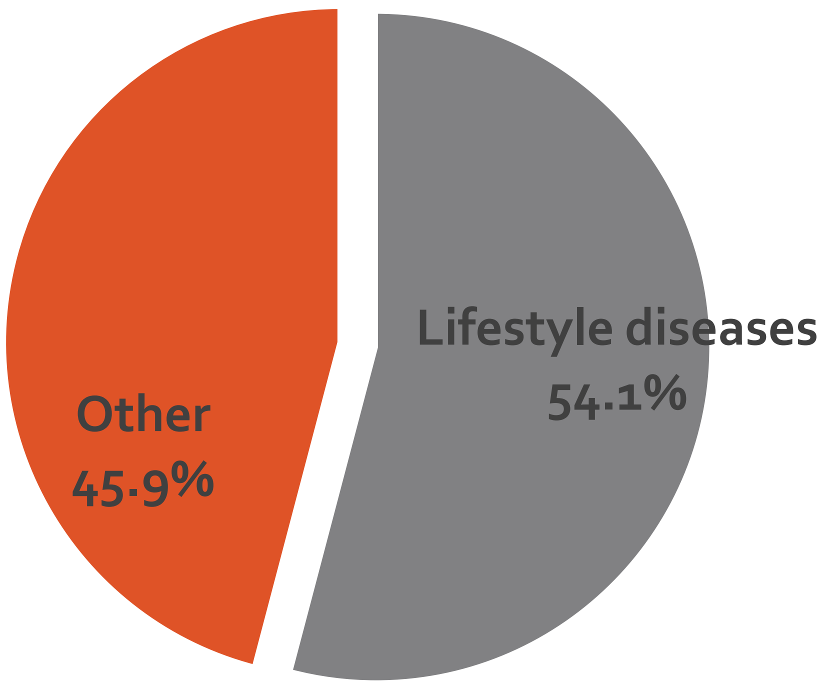


Fig.1 : The demographic statistics of Japan, 2015
(source : Demographic Statistics, Ministry of Health, 2015

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 $D = \left\{ \boldsymbol{x}^{(i)} \;\middle|\; i = 1, \dots, l \right\}, \boldsymbol{x} \in \mathbb{R}^m$
- $m -$ 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 the maximum likelihood method with an L_1 regularization.

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

S : covariance matrix for given data

$\rho > 0$: given regularization parameter

to make a sparse learning by varying ρ

Structure analysis by graphical lasso

- Precision matrix $\Lambda = (\lambda_{ij})$ $\Rightarrow$ **adjacency matrix** which yields a direct correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$

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

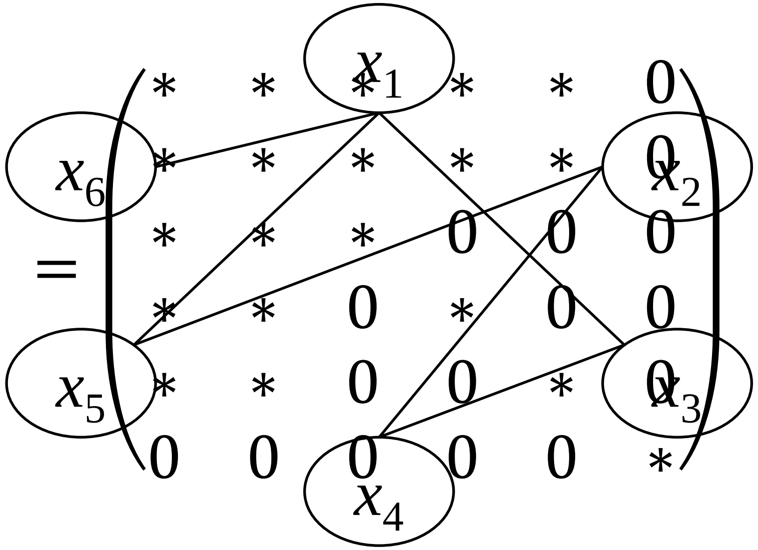


Fig.3 : Correlation for a dataset 1

vs.

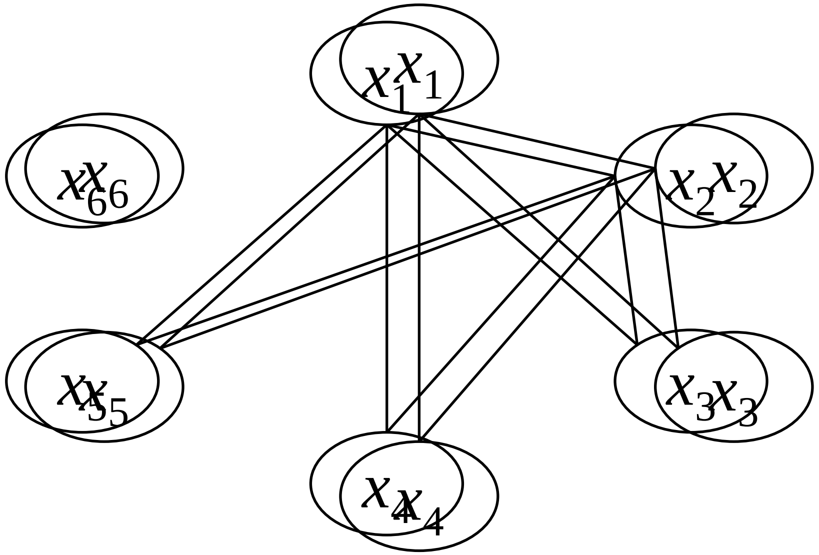


Fig.4 : Correlation for a dataset 2

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

Overview of data

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

- Factors

① Pregnancies ② Glucose ③ Blood Pressure ④ Skin Thickness

⑤ Insulin ⑥ BMI ⑦ Diabetes Pedigree Function ⑧ Age

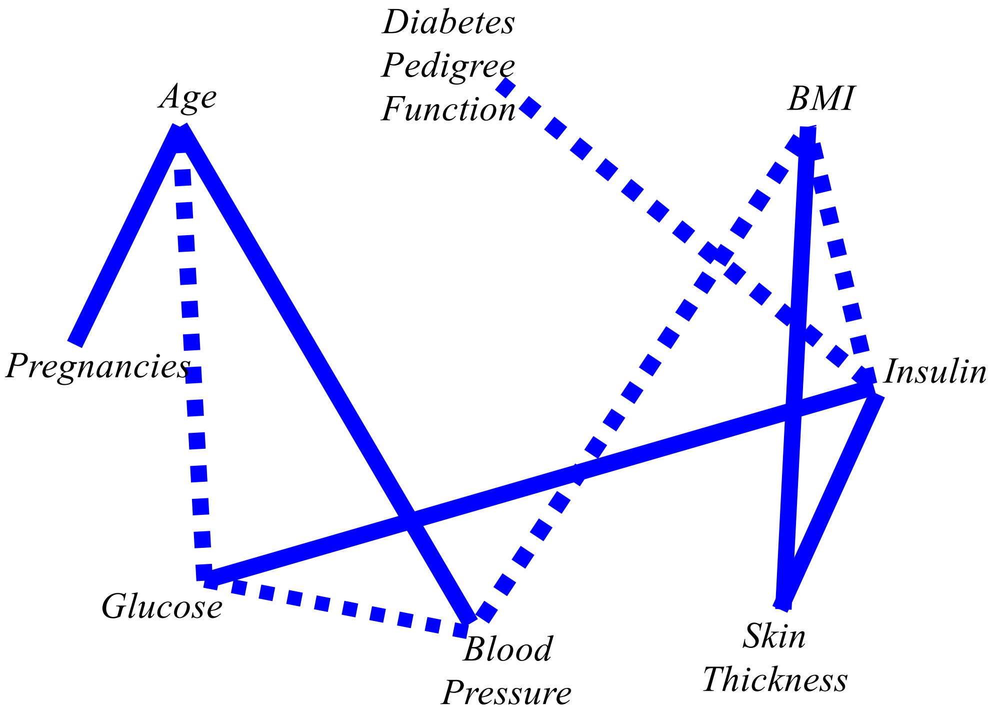
Comparison (1) for correlation matrices

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Pregnancies	-	0.06	0.22	-0.09	-0.16	0.04	-0.06	**0.62**
Glucose	-0.16	-	0.23	0.03	0.35	0.18	0.13	0.20
Blood Pressure	-0.02	0.14	-	0.19	0.07	0.32	0.02	0.28
Skin Thickness	-0.12	0.14	0.14	-	0.39	**0.48**	0.06	-0.08
Insulin	-0.16	0.30	0.02	**0.49**	-	0.29	0.28	-0.12
BMI	-0.13	0.12	0.15	0.30	0.06	-	0.11	0.10
Diabetes Pedigree Function	-0.10	0.04	0.08	0.34	0.09	0.24	-	0.06
Age	0.37	0.07	0.21	-0.02	0.03	-0.20	-0.15	-

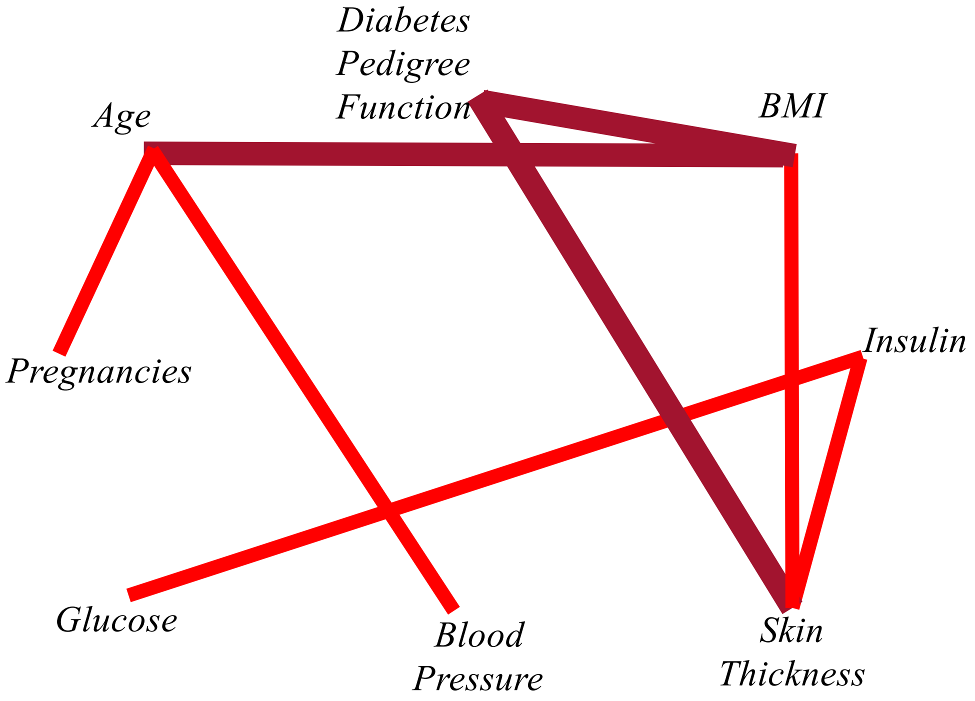
For non-diabetics

For diabetics

Comparison (2) for structures by graphical lasso



(a) Non-diabetics



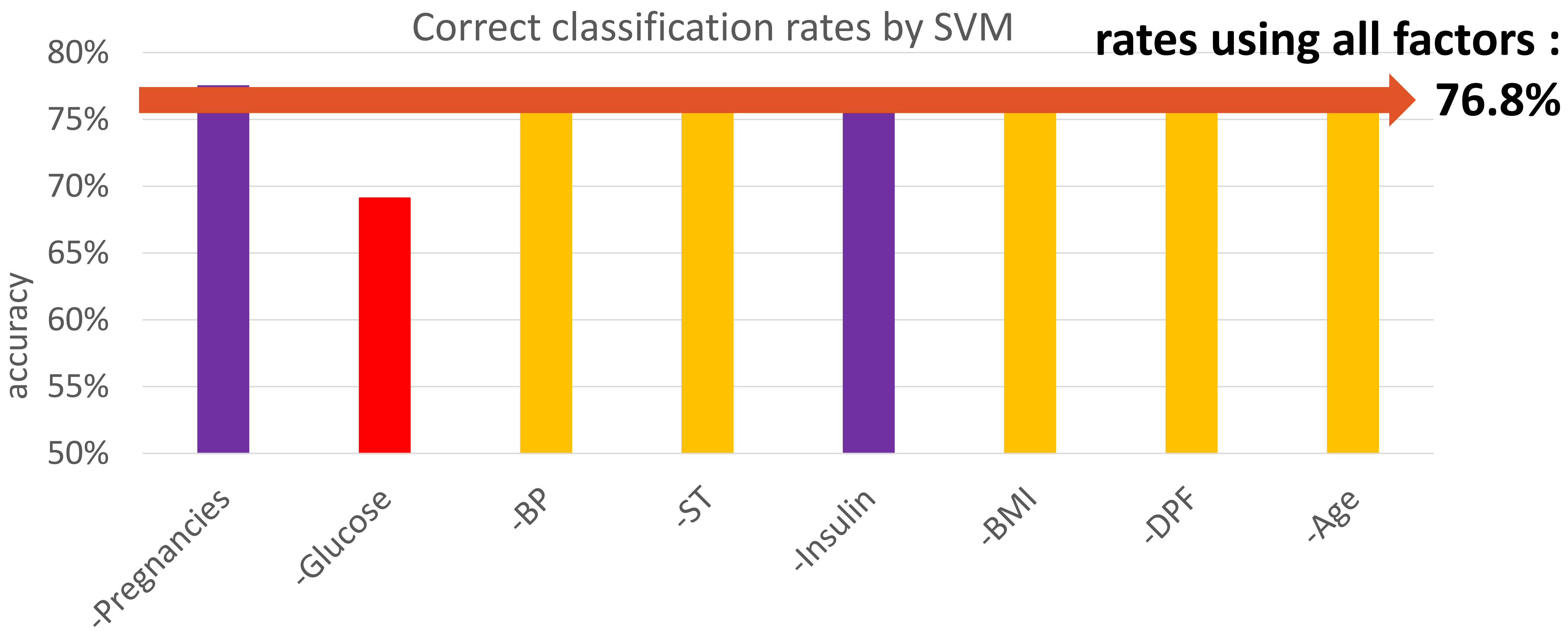
(b) Diabetics

Fig. 5 : Structures based on adjacency matrices

: the relation not appearing in diabetics

: the new relation not appearing in non-diabetics

Importance for each factor in discrimination by SVM



Conclusion

- The results suggest that Glucose is one of the most influential factors for Pima Indians.
- 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/