Common Substructure Learning of Multiple Graphical Gaussian Models

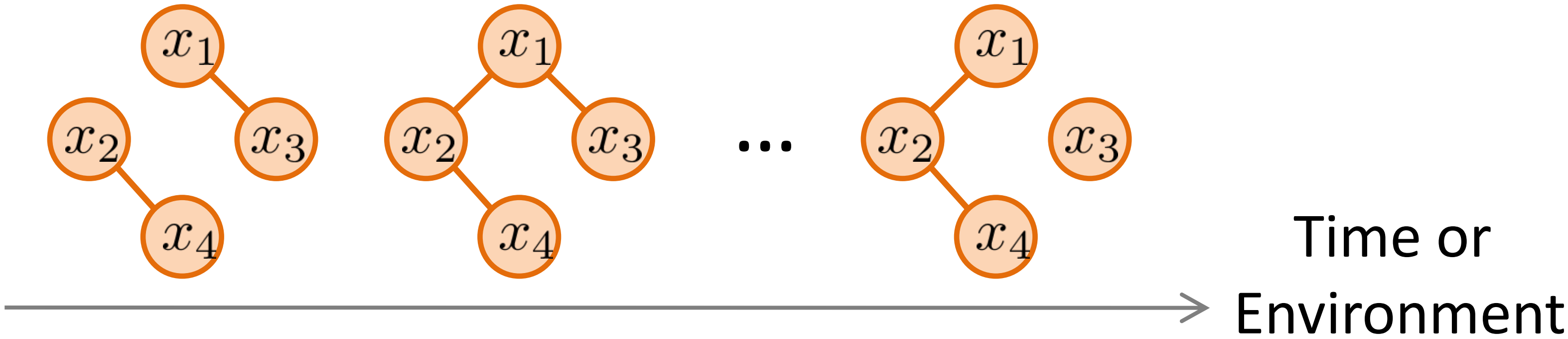
Satoshi Hara, Takashi Washio

The Institute of Scientific and Industrial Research,
Osaka University, Japan

ECML PKDD 2011@Athens, 07/09/2011

Dynamics of Graphical Model

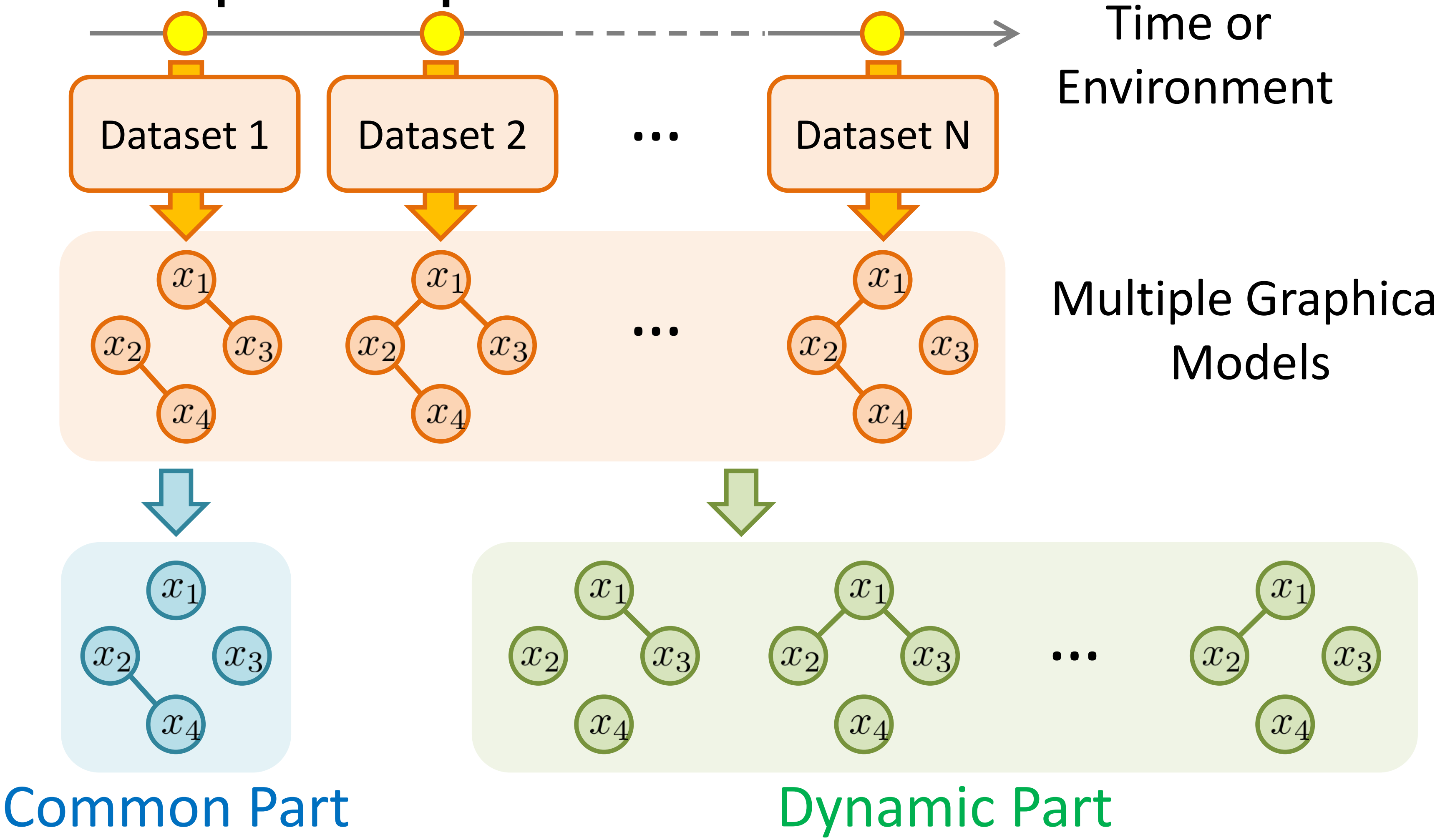
- Evolution of a Data Generating Mechanism
  - e.g., Non-stationarity or Change of Environments
  - The dependency structure may also change.



- Structure changes entirely, or only partially?
  - The change may occur **only partially** when e.g.
    - System Error : fault in subsystems
    - Short Term Changes : natural assumption

Goal of the Research

- Identifying a **Common Substructure** of Multiple Graphical Models



Contents

Background: Graphical Gaussian Model (GGM)

- If a random variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$ is generated from Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1})$,
  - Variables x_j and $x_{j'}$ are conditionally independent.

    $$\Leftrightarrow \quad \Lambda_{jj'} = 0$$

  Λ : Precision Matrix (Inverse of Covariance Σ)

- Structure Learning of GGM

  $\Leftrightarrow$ Identification of zero pattern in Λ
  - Ordinary MLE gives only dense estimate of Λ.
  - Use of sparse methods.
    - ℓ_1 -regularization and its variants

Related Work: Structure Learning of GGM

- ℓ_1-regularized Maximum Likelihood

(Yuan et al., Biometrika 2007, Banerjee et al. JMLR 2008)

$$\max_{\Lambda} \ \ell(\Lambda; \hat{\Sigma}) - \rho \|\Lambda\|_1 \quad \text{s.t.} \quad \Lambda \succ 0$$

  - $\rho > 0$, $\ell(\Lambda; \hat{\Sigma})$ is a log likelihood of Gaussian :

$$\ell(\Lambda; \hat{\Sigma}) = \log\det\Lambda - \mathrm{Tr}\left(\hat{\Sigma}\Lambda\right)$$

  - Convex Optimization, GLasso Algorithm

(Friedman et al., Biostatistics 2008)

- Multi-task Structure Learning (Honorio et al., ICML 2010)
  - Learn GGMs $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$

$$\max_{\{\Lambda_i;\Lambda_i \succ 0\}_{i=1}^{N}} \ \sum_{i=1}^{N} t_i \ell(\Lambda_i; \hat{\Sigma}_i) - \rho \sum_{j \neq j'} \max_{1 \leq i \leq N} |\Lambda_{i,jj'}|$$

Regularization on Joint Structure

Our Proposal: Common Substructure of GGMs

- The common substructure of multiple GGMs (with $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$) is expressed by an adjacency matrix Θ defined by

$$\Theta_{jj'} = \begin{cases} \Lambda_{1,jj'} \;, & \text{if } \Lambda_{1,jj'} = \Lambda_{2,jj'} = \ldots = \Lambda_{N,jj'} \\ 0 \;, & \text{otherwise} \end{cases}$$

  - weak stationarity on partial covariance

- (j, j') th element is common.

Maximal variation is zero.

$$\Leftrightarrow \quad \max_{1 \le i, i' \le N} \left| \Lambda_{i,jj'} - \Lambda_{i',jj'} \right| = 0$$

Our Proposal: Problem Formulation

- Use of 2 Regularizations
  - Regularization on Joint Structure (Honorio et al., ICML2010)
  - Regularization on Maximal Variation (Our Proposal)

$$\max_{\{\Lambda_1\}_{i=1}^{N}} \sum_{i=1}^{N} t_i \ell(\Lambda_i; \hat{\Sigma}_i) - \sum_{j \neq j'} \left(\rho \max_{1 \leq i \leq N} |\Lambda_{i,jj'}| + \gamma \max_{1 \leq i, i' \leq N} |\Lambda_{i,jj'} - \Lambda_{i',jj'}| \right)$$

$$\text{s.t.} \quad \Lambda_1, \Lambda_2, \ldots, \Lambda_N \succ 0$$

Regularization on Joint Structure

Regularization on Maximal Variation

  - $\rho, \gamma > 0$, non-negative weights $\sum_{i=1}^{N} t_i = 1$
  - Convex Optimization Problem

Our Proposal: Relation to The Existing Work

- Structural Changes between two datasets (Zhang et al., UAI 2010)
  - Lasso type approach (Meinshausen et al., Ann. Statist. 2006) + Fused Lasso type regularization

- Connection to the current problem

	Proposed	Zhang et al.
Objective Function	Regularized MLE of Gaussians	Fused Lasso Type (Approximation)
# of Datasets N	$N \geq 2$	$N = 2$ only
Algorithm	$N \geq 2$	$N = 2$ only

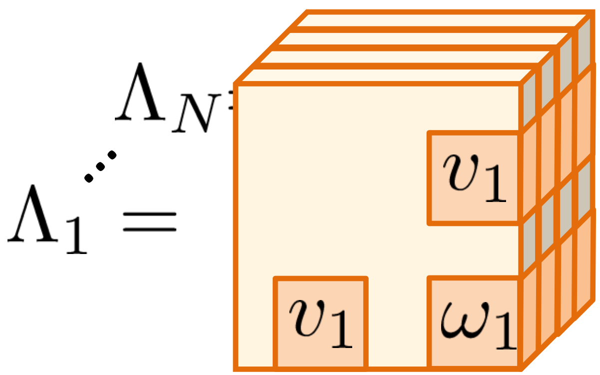
More General Framework

Contents

Block Coordinate Descent

- Iteratively update each elements of matrices.
  - Solve subproblems for each (j, j')th elements of precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$.
  - Different sub-problems for diagonal elements $\boldsymbol{\omega}$ and non-diagonal elements $\boldsymbol{v}$.

vector of (j, j') th elements



$$\boldsymbol{v} = (v_1, v_2, \ldots, v_N)^\top$$

$$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)^\top$$

- Convergence to the global optimum is guaranteed.

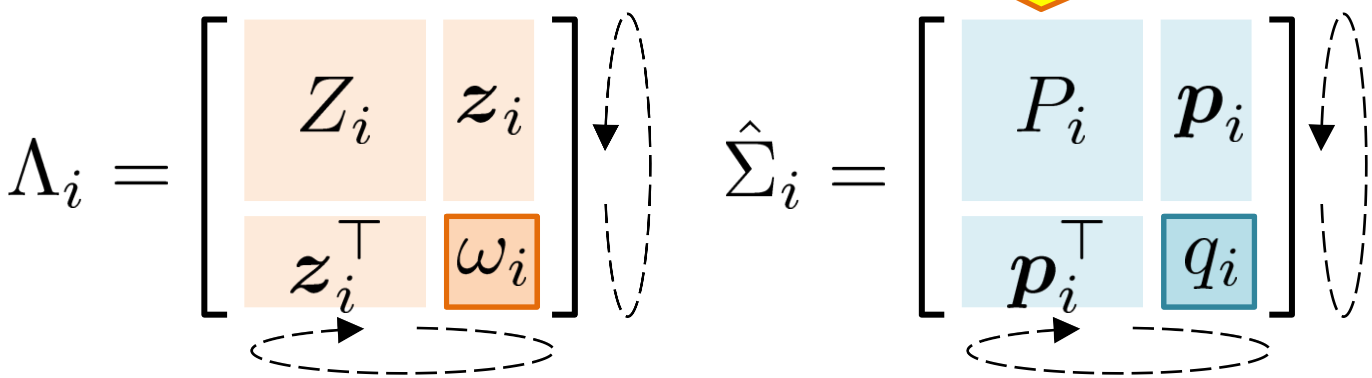
(Tseng, JOTA 2001)

Optimization of Diagonal Entries

- Analytic Solution

$$\omega_i = \boldsymbol{z}_i^\top Z_i^{-1} \boldsymbol{z}_i + q_i^{-1}$$

1. Permute row and column of matrices.
2. Divide into (j, j) th elements and remainings.



- Positive Definiteness
  - If $Z_i \succ 0$, then $\Lambda_i \succ 0$ always holds.
  - Positive definiteness is preserved at each updating step of the block coordinate descent.

Optimization of Non-diagonal Entries

Dual Problem

$$\min_{\boldsymbol{\xi}} \ \frac{1}{2}(\boldsymbol{b}-\boldsymbol{\xi})^\top \mathrm{diag}(\boldsymbol{a})^{-1}(\boldsymbol{b}-\boldsymbol{\xi})$$
$$\text{s.t.} \ |\mathbf{1}_N^\top \boldsymbol{\xi}| \le \rho, \quad \|\boldsymbol{\xi}\|_1 \le \rho + 2\gamma$$

Primal (Non-Diagonals)

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_N)^\top$$

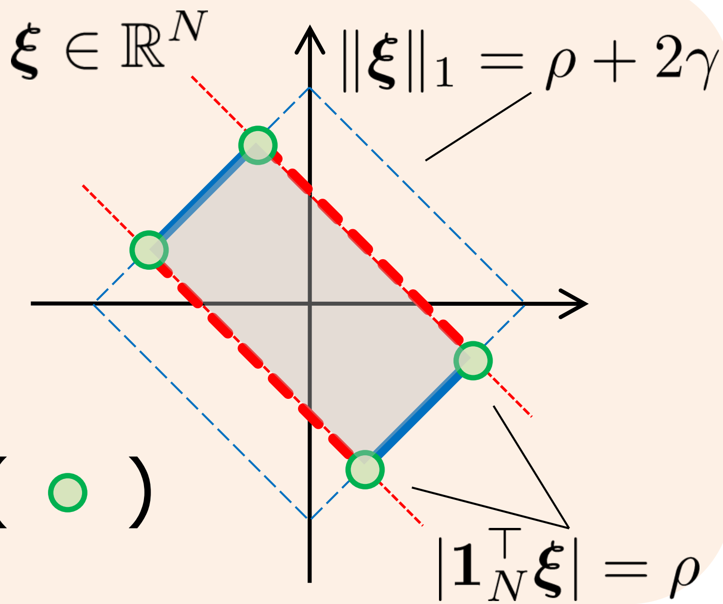
Dual Variable

$$\boldsymbol{\xi} = \boldsymbol{b} - \mathrm{diag}(\boldsymbol{a})\boldsymbol{v}$$

- $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$: defined from remaining parameters, $\hat{\Sigma}_i$

4 Types of Solutions

- $\boldsymbol{\xi} = \boldsymbol{b}$ ( )
- $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$ ( )
- $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$ ( )
- $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$, $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$ ( )



Solution to Each Case

1) $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$ (——)

- Continuous Quadratic Knapsack Problem

$$\min_{\boldsymbol{y}} \sum_{i=1}^{N} \frac{1}{2a_i}(|b_i| - y_i)^2$$
$$\text{s.t. } \boldsymbol{y} \geq 0,\ \mathbf{1}_N^\top \boldsymbol{y} = \rho + 2\gamma$$

$(\xi_i = \mathrm{sgn}(b_i) y_i)$

2) $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$ (- - -)

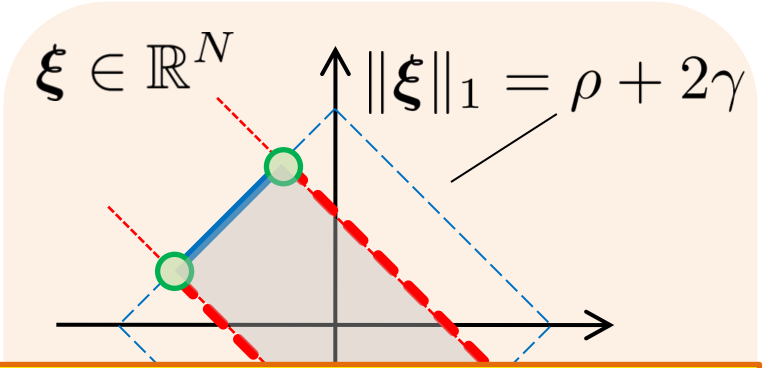
- Analytic Solution

$$v_0 = \frac{\mathbf{1}_N^\top \boldsymbol{b} - \rho\, \mathrm{sgn}(\mathbf{1}_N^\top \boldsymbol{b})}{\mathbf{1}_N^\top \boldsymbol{a}}$$

$(\boldsymbol{\xi} = \boldsymbol{b} - v_0 \boldsymbol{a})$

3) $\begin{cases} \|\boldsymbol{\xi}\|_1 = \rho + 2\gamma \\ |\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho \end{cases}$ (○)

- Continuous Quadratic Knapsack Problem



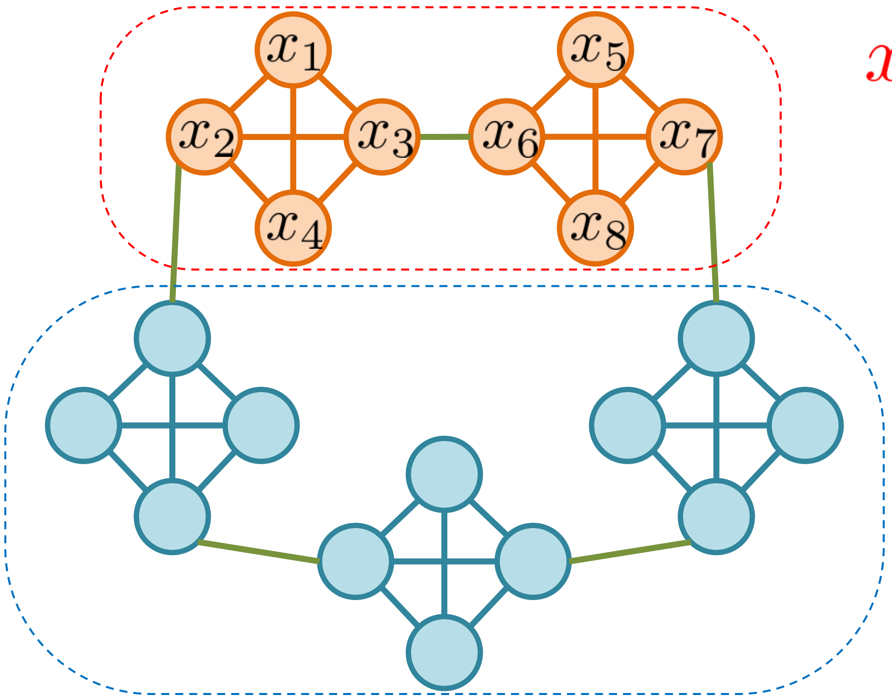
One of these 3 cases or $\xi = b$ is the solution.

Contents

Simulation Setup

GGM with Common Substructure

- Dim. $d = 20$, # of Datasets $N = 5$
- Λ_i : Diagonals$= 1$, Non-zeros $\sim [-0.8, -0.1] \cup [0.1, 0.8]$
- 100 data points from each Gaussian $\mathcal{N}(\mathbf{0}, \Lambda_i^{-1})$



$x_1 \sim x_8$
Common Substructure
(Structure, weights are common.)

$x_9 \sim x_{20}$
Individual Substructure
(Structure, weights changes.)

Baseline Methods

- Naïve Way to Learn Common Substructure
  1: Estimate $\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_N$ with existing methods
    - GLasso (Friedman et al., Biostatistics 2008)
    - Multi-task Structure Learning (Honorio et al., ICML 2010)

  2: Find seemingly common parts

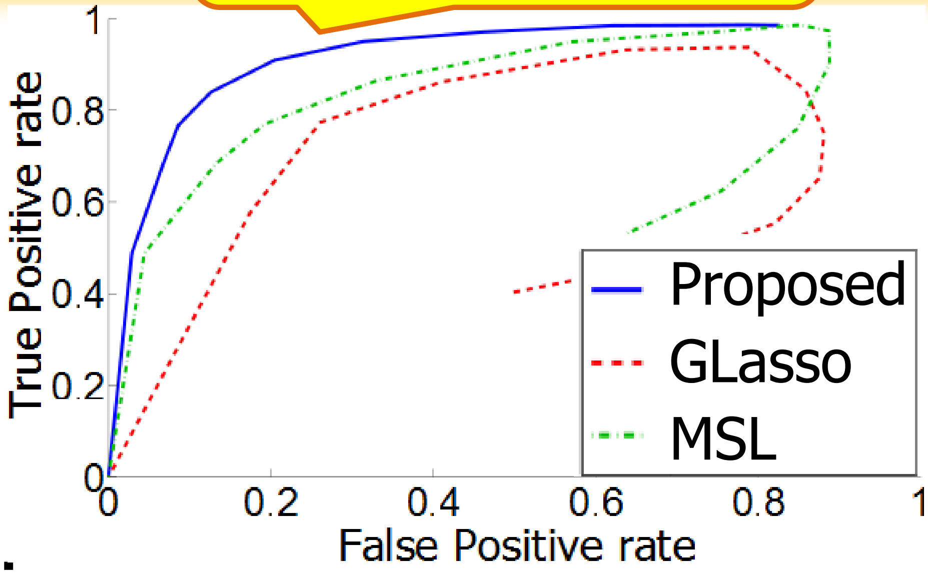
- Seemingly Common Substructure

$$\hat{\Theta}_{jj'} = \begin{cases} \hat{\theta}_{jj'} \ , & \text{if } \max_{i,i'} |\hat{\Lambda}_{i,jj'} - \hat{\Lambda}_{i',jj'}| < \epsilon \\ 0 \ , & \text{otherwise} \end{cases}$$

- $\hat{\theta}_{jj'} = 0$ if $\hat{\Lambda}_{i,jj'} = 0 \ , \ ^{\forall}i$, $\hat{\theta}_{jj'} = 1$ otherwise

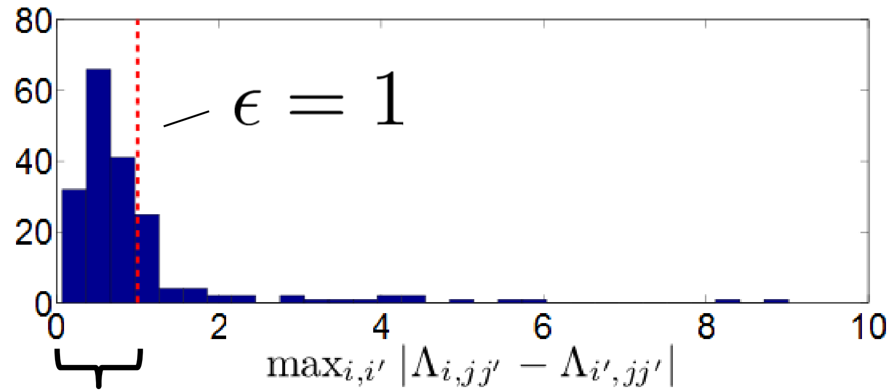
Result

- ROC by varying ρ
  - Average of 100 run
  - $\epsilon = 1$
  - γ by a heuristic



- $\epsilon = 1$ is quite optimistic.
  - 62% of true common substructure have a variation more than 1.
  - The proposed method avoids this estimation variance problem.

GLasso ($\rho = 0.0032$)



74% of non-zeros are under the threshold.

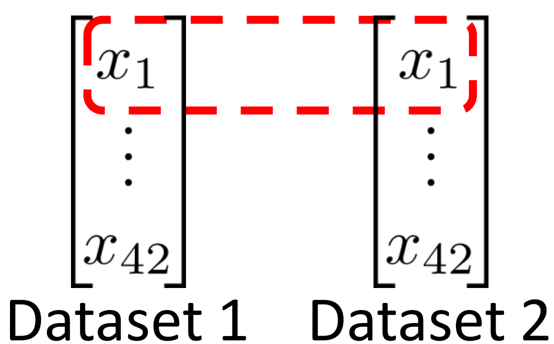
Contents

Application to Anomaly Detection

- **Automobile Sensor Error Data** (Ide et al., SDM 2009)
  - 42 sensor values from a real car
  - 79 datasets from normal states and 20 from faulty
  - Fault : miswiring of 24th and 25th sensors

One covariance for each dataset

- **Detection of Correlation Anomaly** (Ide et al., SDM 2009)
  - Capture the dependency structure by GGM
  - **Anomaly Score:** KL-divergence between conditional distributions for each pair of variables



$$a_j = \max(d_j^{12}, d_j^{21})$$

$$d_j^{12} = \int D_{\mathrm{KL}}[p_1(x_j|\boldsymbol{x}_{\backslash j})||p_2(x_j|\boldsymbol{x}_{\backslash j})]p_1(\boldsymbol{x}_{\backslash j})d\boldsymbol{x}_{\backslash j}$$

Simulation Setting

- Use 25 datasets (20 normal, 5 faulty)

1. Estimate 25 Precision Matrices

Base lines
- Individual estimation by GLasso (Friedman et al., 2008)
- Multi-task Structure Learning (Honorio et al., 2010)
- Common Substructure Learning

Weights are chosen to balance two states.

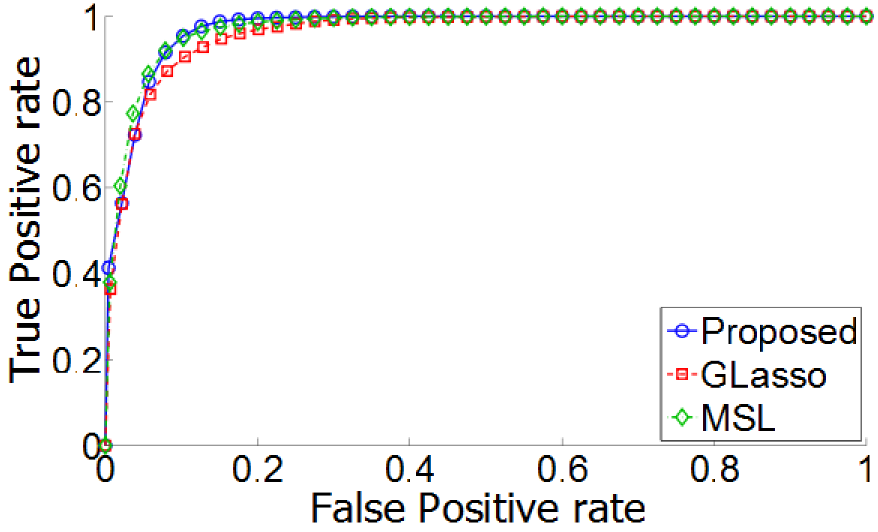
2. Calculate Anomaly Scores
- Average scores for all 20×5 pairs.
- Detect anomaly sensors by thresholding.

Result (Detection Performance)

- Randomly pickup 25 datasets for 100 times.
  - Regularization parameter ρ is in $0.05 \sim 0.30$.
  - The parameter γ is chosen by a heuristic.

- Draw best ROC by changing the threshold.

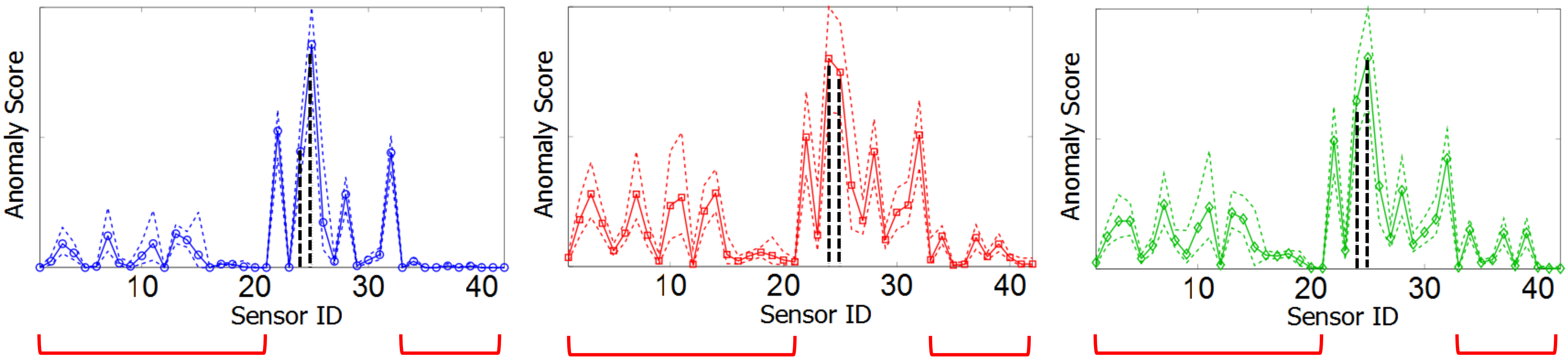
	Best AUC	ρ
Proposed	**0.97**	0.05
GLasso	0.96	0.20
MSL	**0.97**	0.05



Result (Anomaly Score)

- Normal-Faulty states (median, 25/75% of 100 run)

Proposed GLasso MSL



- The proposed method captures the dependency among healthy sensors as common and shows **lower scores**.
- The variation of scores are also low.

→ **More stable than other two**

Summary & Conclusion

- Common Substructure Learning
  - Identifying common parts of dynamical dependency structure
  - Optimization by block-coordinate descent
  - Factorization of subproblem to 4 cases
- Numerical Evaluation
  - Validity of the proposed method are observed both on synthetic and real world data.
  - Naïve approaches tend to fail detecting common substructure due to the estimation variance.

Supplemental Materials

Learning GGM (Covariance Selection)

- Maximum Likelihood Estimator : $\hat{\Lambda} = \hat{\Sigma}^{-1}$
  - $\hat{\Lambda}$ is usually dense. ($\hat{\Sigma}$: MLE of Σ)
  - GGM is a complete graph, and the true dependency structure is masked.

- ℓ_1-regularized Maximum Likelihood

(Yuan et al., Biometrika 2007, Banerjee et al. JMLR 2008)

$$\max_{\Lambda} \ \ell(\Lambda; \hat{\Sigma}) - \rho \|\Lambda\|_1 \quad \text{s.t.} \quad \Lambda \succ 0$$

  - $\rho > 0$, $\ell(\Lambda; \hat{\Sigma})$ is a log likelihood of Gaussian :

$$\ell(\Lambda; \hat{\Sigma}) = \log \det \Lambda - \mathrm{Tr}\left(\hat{\Sigma}\Lambda\right)$$

  - Convex Optimization, **GLasso Algorithm**

(Friedman et al., Biostatistics 2008)

Joint Estimation of GGMs

- Multi-task Structure Learning (Honorio et al., ICML 2010)
  - Learn GGMs from covariances $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_N$.
  - Assumption: All GGMs have same edge patterns.

$$\max_{\{\Lambda_i;\Lambda_i \succ 0\}_{i=1}^N} \sum_{i=1}^{N} t_i \ell(\Lambda_i; \hat{\Sigma}_i) - \rho \sum_{j \neq j'} \max_{1 \leq i \leq N} |\Lambda_{i,jj'}|$$

- Joint structure is sparse.

$$\tilde{\Lambda}_{jj'} \equiv \max_{1 \leq i \leq N} |\Lambda_{i,jj'}| = 0 \quad \Leftrightarrow \quad \Lambda_{i,jj'} = 0, \quad {}^{\forall} i$$

- Share edge pattern information and improve the result.

Algorithm (Block Coordinate Descent)

- Input : Covariance Matrices $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_N$
  Regularization Parameters $\rho, \gamma > 0$
  Weights $t_1, t_2, \ldots, t_N \geq 0, \quad \sum_{i=1}^{N} t_i = 1$
- Output : Precision Matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$

- Initialize $\Lambda_i \leftarrow \hat{\Sigma}_i^{-1} \; (1 \leq i \leq N)$
- Repeat until convergence

  For $j = 1$ to d , $j' = 1$ to d

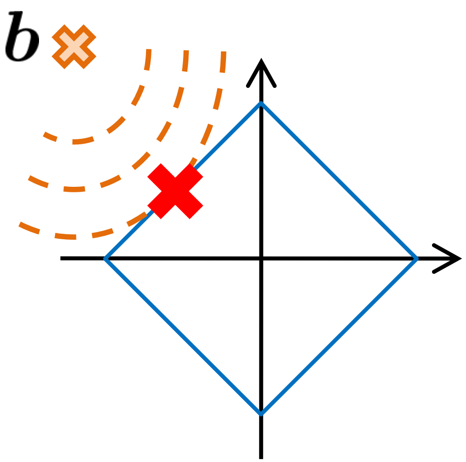
  Update (j, j')th elements of $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$

  End For

Treat remaining elements as constants.

Solution to the Dual Problem 1/3

- Case1: The solution is on $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$.



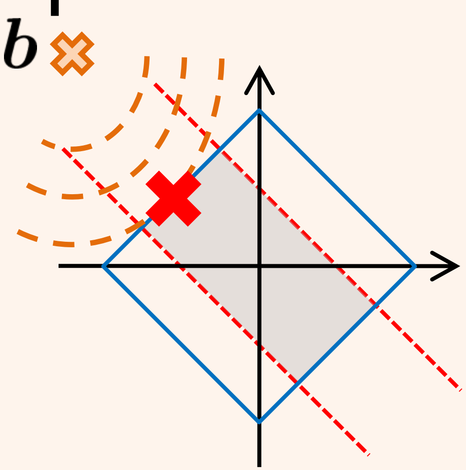
Continuous Quadratic Knapsack Problem

$(\xi_i = \mathrm{sgn}(b_i) y_i)$

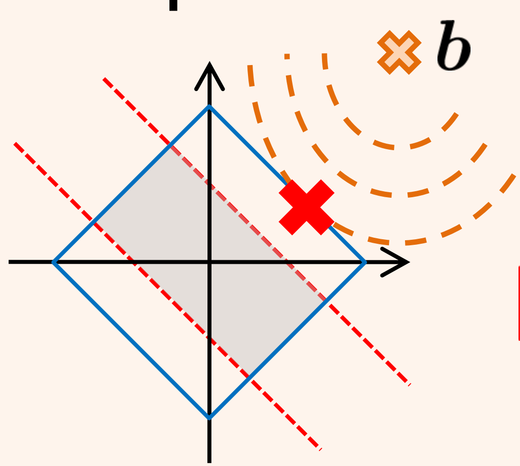
$$\min_{\boldsymbol{y}} \sum_{i=1}^{N} \frac{1}{2a_i} (|b_i| - y_i)^2$$
$$\text{s.t.} \quad \boldsymbol{y} \geq 0, \; \mathbf{1}_N^\top \boldsymbol{y} = \rho + 2\gamma$$

Efficient algorithm exists.
(Honorio et al., ICML2010)

- Optimal



- Not Optimal

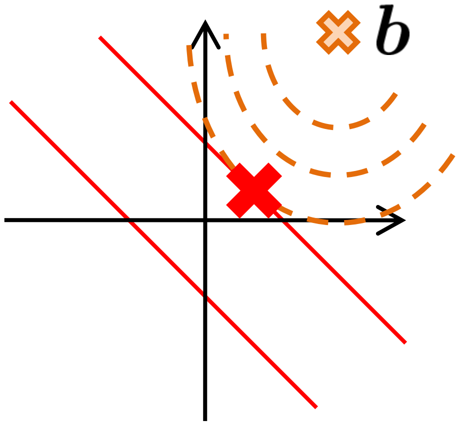


$|\mathbf{1}_N^\top \boldsymbol{\xi}| \geq \rho$

→ Case2

Solution to the Dual Problem 2/3

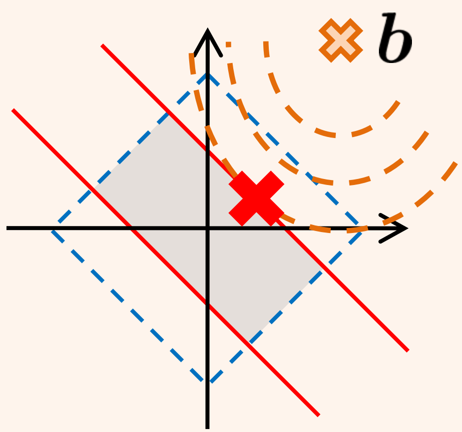
- Case2: The solution is on $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$.



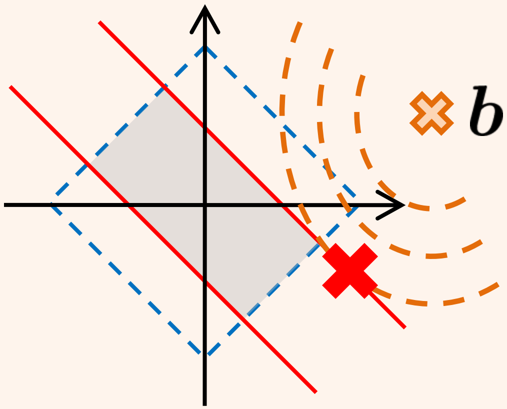
Analytic Solution $(\boldsymbol{\xi} = \boldsymbol{b} - v_0 \boldsymbol{a})$

$$v_0 = \frac{\mathbf{1}_N^\top \boldsymbol{b} - \rho \,\mathrm{sgn}(\mathbf{1}_N^\top \boldsymbol{b})}{\mathbf{1}_N^\top \boldsymbol{a}}$$

- Optimal



- Not Optimal

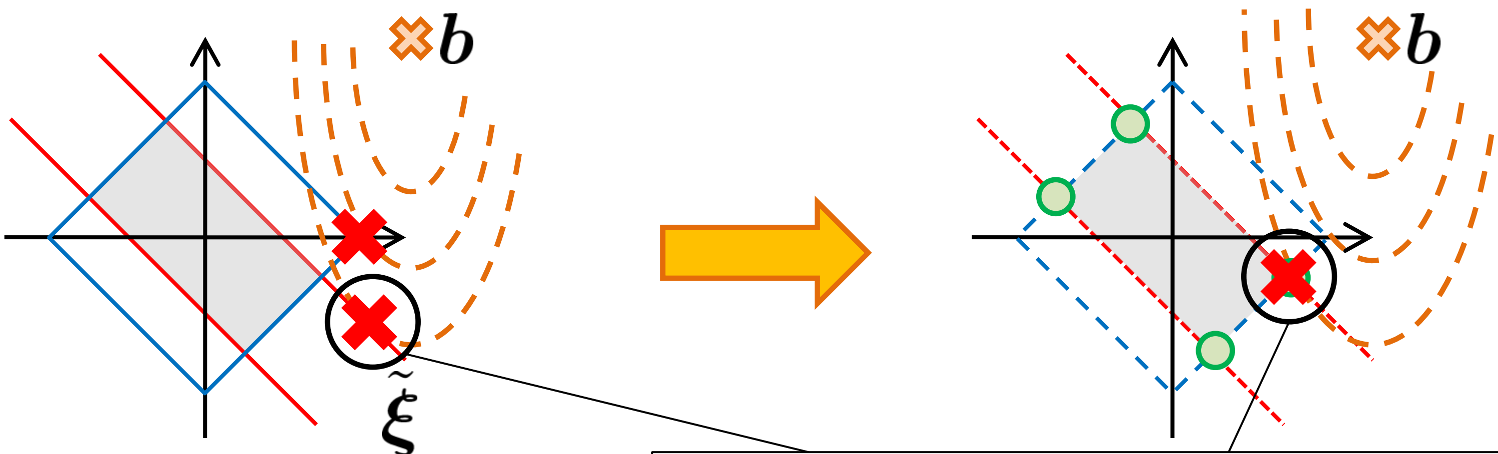


$\|\boldsymbol{\xi}\|_1 \geq \rho + 2\gamma$

→ Case3

Solution to the Dual Problem 3/3

- Case3: The solution is on $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$, $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$.



When both Case 1, 2 are not optimal

Solutions to Case 2, 3 have the same sign.

$\tilde{\boldsymbol{\xi}}$: Solution to Case 2

- Problems for each signs, $\tilde{\xi}_i \geq 0$ and $\tilde{\xi}_i < 0$
  - Two Continuous Quadratic Knapsack Problems

Solution to the Dual Problem 3/3 (Cont.)

- Target Problem

$$\min_{\boldsymbol{\xi}} \ \frac{1}{2}(\boldsymbol{b}-\boldsymbol{\xi})^\top \mathrm{diag}(\boldsymbol{a})^{-1}(\boldsymbol{b}-\boldsymbol{\xi}) \ \text{ s.t. } \ |\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho, \ \ \|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$$

- Equivalent Two Distinct Problems
  - Continuous Quadratic Knapsack Problems

$$\min_{\boldsymbol{y}^+} \sum_{\tilde{\xi}_i \geq 0} \frac{1}{2a_i}(y_i^+ - b_i)^2 \ \text{ s.t. } \ \boldsymbol{y}^+ \geq 0, \ \mathbf{1}_N^\top \boldsymbol{y}^+ = \alpha^+$$

$$\min_{\boldsymbol{y}^-} \sum_{\tilde{\xi}_i < 0} \frac{1}{2a_i}(y_i^- - b_i)^2 \ \text{ s.t. } \ \boldsymbol{y}^- \geq 0, \ \mathbf{1}_N^\top \boldsymbol{y}^- = \alpha^-$$

  - $\xi_i = y_i^+ \ (\tilde{\xi}_i \geq 0)$ and $\xi_i = -y_i^- \ (\tilde{\xi}_i < 0)$
  - $(\alpha^+, \alpha^-) = (\rho+\gamma, \gamma)$ or $(\alpha^+, \alpha^-) = (\gamma, \rho+\gamma)$

Solution to Continuous Quadratic Knapsack Problem

Continuous Quadratic Knapsack Problem

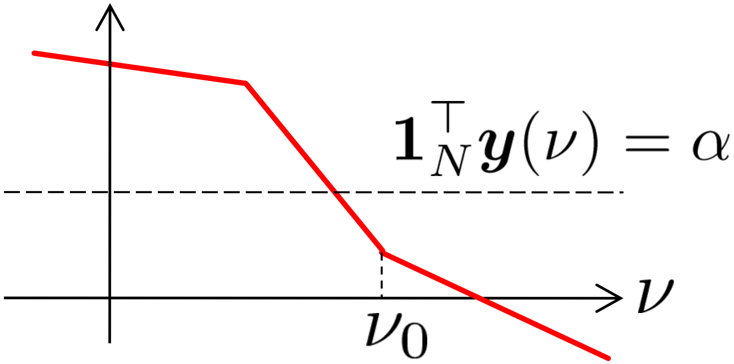
$$\min_{\boldsymbol{y}} \sum_{i=1}^{N} \frac{1}{2c_i}(y_i - d_i)^2 \ \text{ s.t. } \ \boldsymbol{y} \geq 0, \ \mathbf{1}_N^\top \boldsymbol{y} = \alpha$$

- Solution: $y_i(\nu) = \max(d_i - \nu c_i, 0)$
- ν is what satisfies $\mathbf{1}_N^\top \boldsymbol{y}(\nu) = \alpha$.

Search of Optimal ν

- $\mathbf{1}_N^\top \boldsymbol{y}(\nu)$ is decreasing and piece-wise linear with breakpoints $\{d_i/c_i\}_{i=1}^N$.

$$\nu = \frac{\sum_{d_i - \nu_0 c_i \geq 0} d_i - \alpha}{\sum_{d_i - \nu_0 c_i \geq 0} c_i}$$



Regularization Parameters

- ρ : Regularization of the Joint Structure
- γ : Regularization of the Maximal Variation
- Bivariate Case : $\hat{\Sigma}_i = \begin{bmatrix} a_i & r_i \\ r_i & b_i \end{bmatrix}, \ \Lambda_i = \begin{bmatrix} u_i & z_i \\ z_i & v_i \end{bmatrix}$

$$|r_i| \leq \rho + 2\gamma \ \text{ and } \ \left| \sum_{i=1}^{N} t_i r_i \right| \leq \rho \quad \Rightarrow z_i = 0$$

  - ρ : Threshold to round small covariances
  - γ : Difference of characteristic scalings between r_i and $\tilde{r} = \sum_{i=1}^{N} t_i r_i$

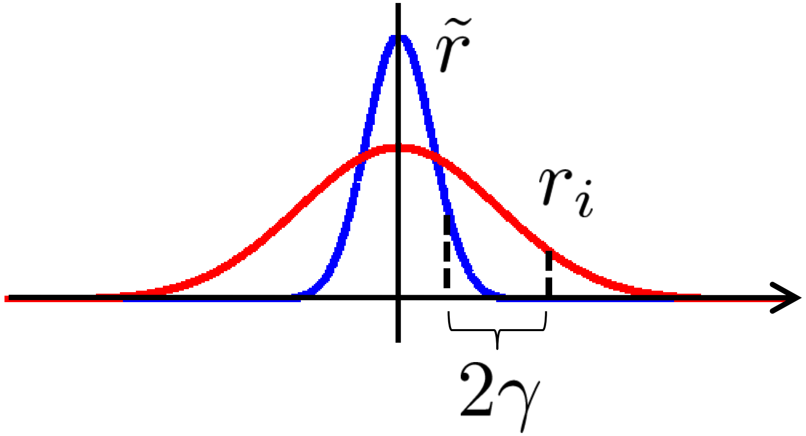
Choice of Parameter γ

Intuition on γ

- Difference of characteristic scalings between r_i and $\tilde{r} = \sum_{i=1}^{N} t_i r_i$

Heuristic Choice

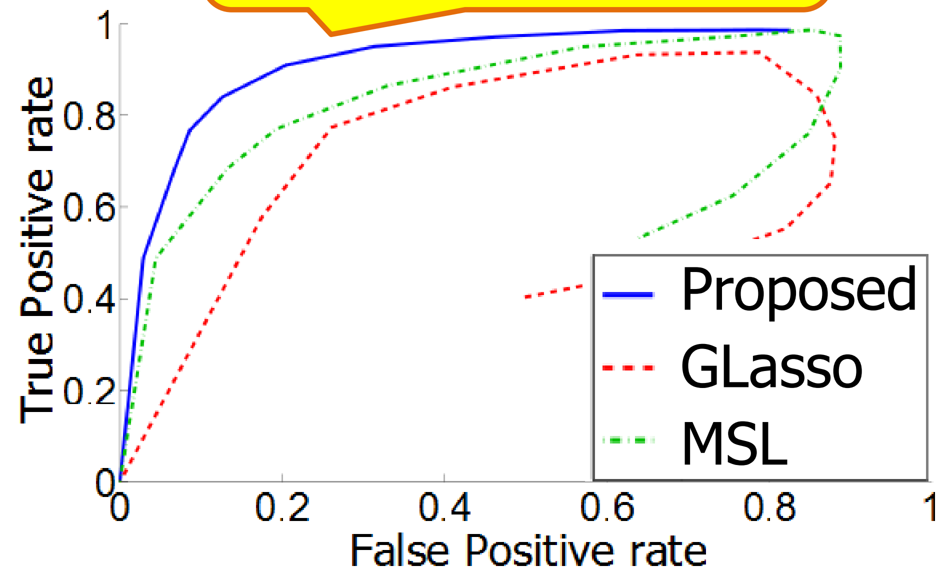
- Approximation : r_i , $\tilde{r}$ are Gaussian.
- Adopt $100(1-\alpha)\%$ points as their characteristic scalings



Result (1)

Proposed method is the best.

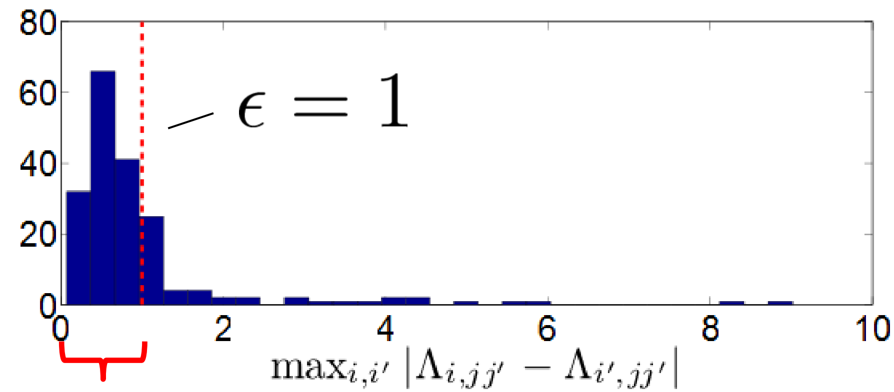
- ROC by varying ρ
  - Average of 100 run
  - $\epsilon = 1$
  - γ by a heuristic



- $\epsilon = 1$ is quite optimistic.
  - 62% of true common substructure have a variation more than 1

(Estimation Variance)

GLasso ($\rho = 0.0032$)

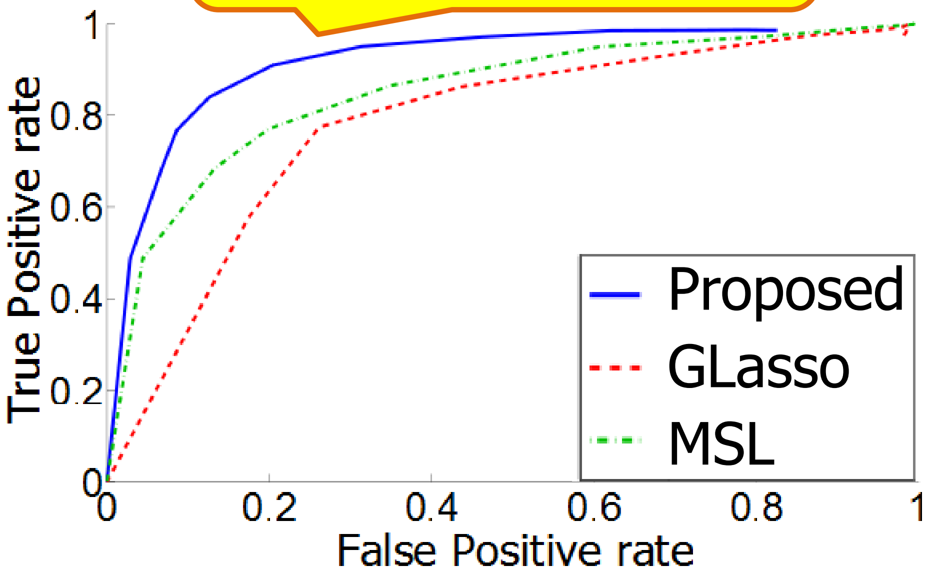


74% of non-zeros are under threshold.

Result (2)

- ROC by varying ρ
  - Average of 100 run
  - $\epsilon = 10$
  - Naïve approaches treat almost all parts as common.

Proposed method is the best.



- Ordinary GGM estimation have high variances.
  - Common substructure is masked and naïve approaches fail.
  - The proposed method could avoid this problem.

Application to Anomaly Detection

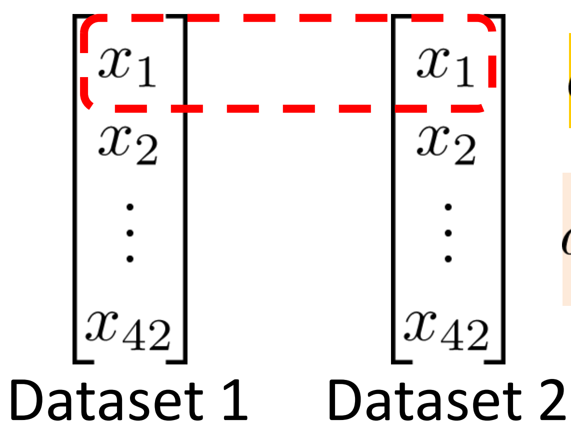
- Anomaly Detection Task
  - Identify contributions of each variable to the difference between two datasets.
- Correlation Anomaly (Ide et al., SDM 2009)
  - Use sparse GGM estimation for suppressing pseudo correlation in noisy situations.
- Use of Common Substructure Learning
  - If fault occurs only in some subsystems, other healthy parts will show common dependency.

Dataset Description

- Automobile Sensor Error Data (Ide et al., SDM 2009)
  - 42 sensor values from a real car
  - 79 datasets from normal states and 20 from faulty
  - Fault : miswiring of 24th and 25th sensors

One covariance for each dataset

- Anomaly Score (Ide et al., SDM 2009)
  - KL-divergence between conditional distributions calculated for each pair of variables

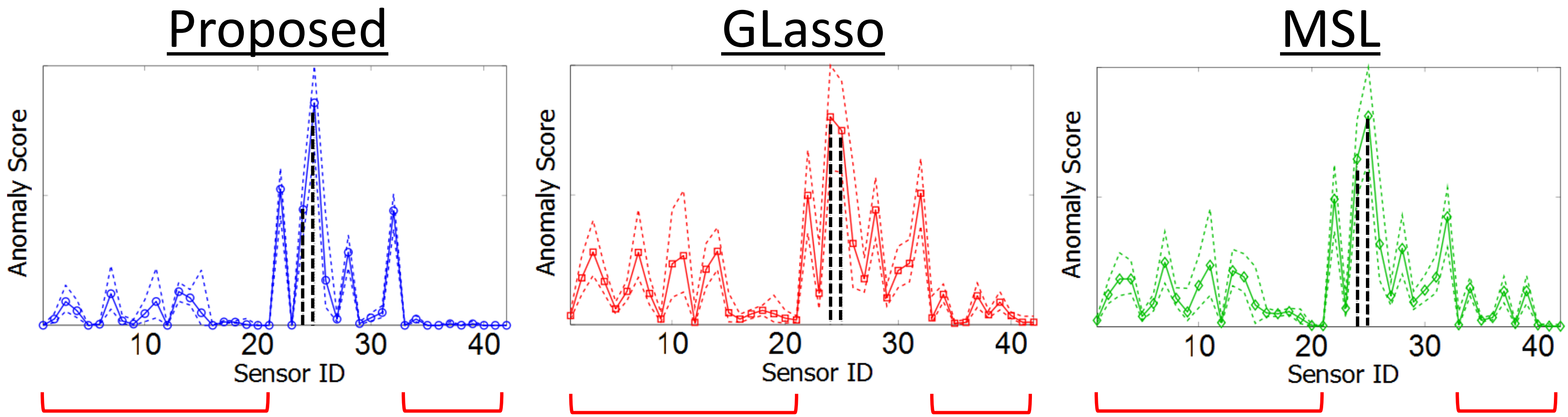


$$a_j = \max(d_j^{12}, d_j^{21})$$

$$d_j^{12} = \int D_{\mathrm{KL}}[p_1(x_j|\boldsymbol{x}_{\backslash j})||p_2(x_j|\boldsymbol{x}_{\backslash j})]\,p_1(\boldsymbol{x}_{\backslash j})\,d\boldsymbol{x}_{\backslash j}$$

Result (Anomaly Score)

- Normal-Faulty states (median, 25/75% of 100 run)

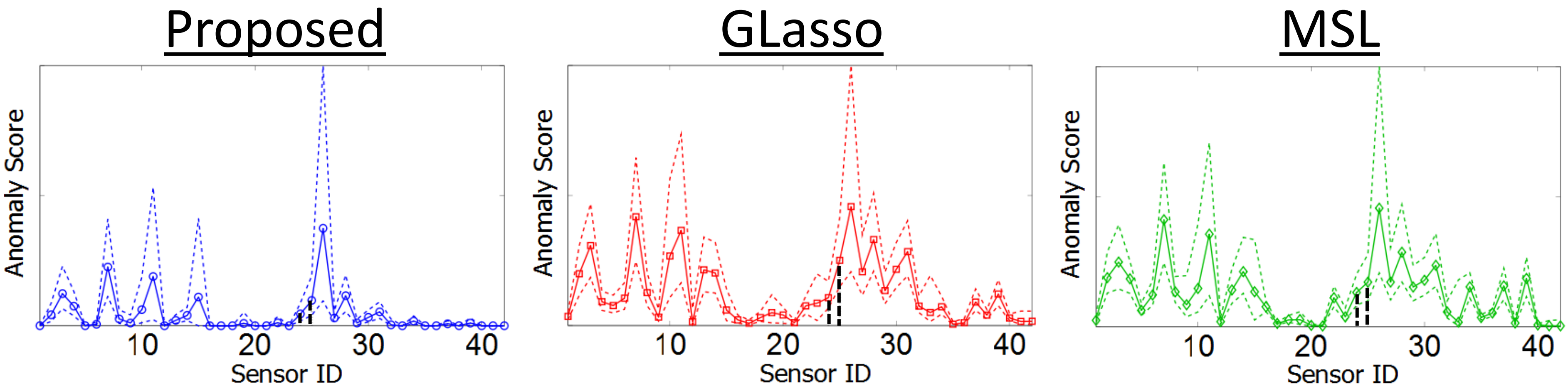


- The proposed method shows **lower scores** at healthy sensors.
- The variation of scores are also low.

  → **More stable than other two**

Result (Anomaly Score 2)

- Normal-Normal states (median, 25/75% of 100 run)



- Same tendency as Normal-Fault
  - Lower score, Lower variation
- Ideally, "score=0" for Normal-Normal states
  - Some sensor are quite noisy.
  - Contrasting with Normal-Fault gives additional info.

Result (Anomaly Score)