IBM **Research**

L_0-constrained Gaussian Graphical Model for Anomaly Localization

Dzung Phan, Tsuyoshi Ide, Jayant Kalagnanam
IBM Research

Matt Menickelly
Argonne National Laboratory

Katya Scheinberg
Lehigh University

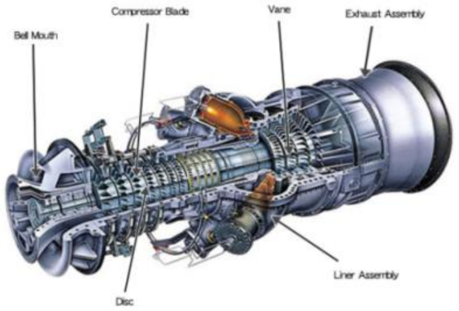
Outline

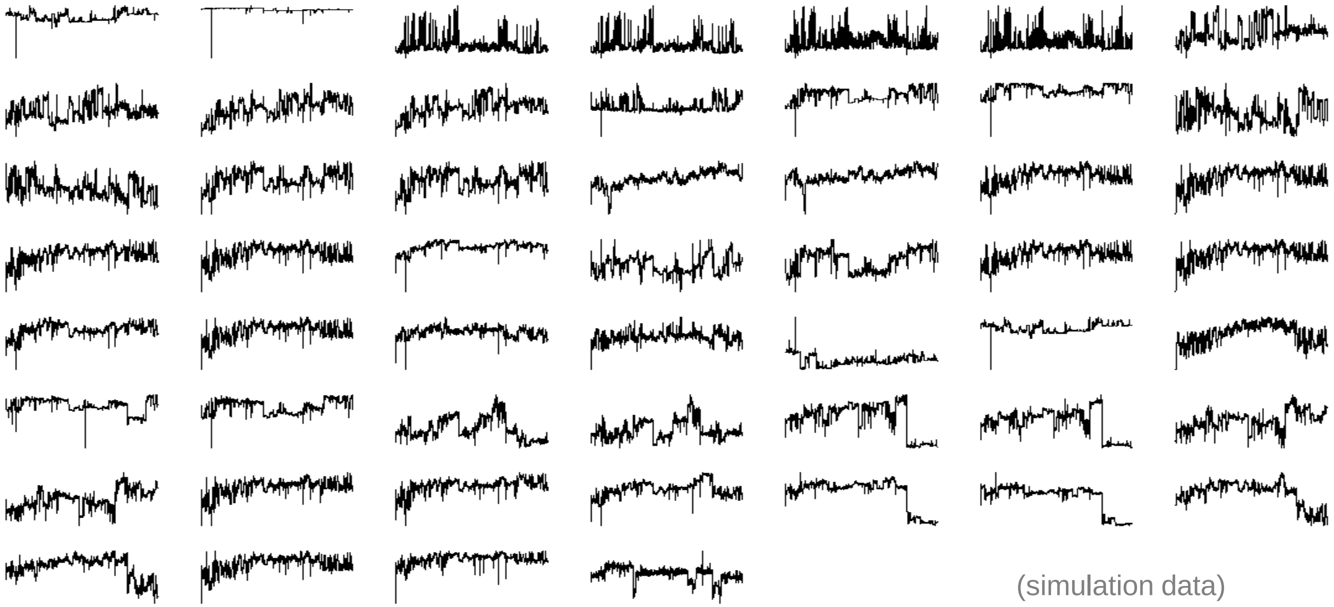
- Introduction to Anomaly Localization
- L_0-constrained Gaussian Graphical Model
- Optimization Algorithm for ℓ_0 Sparse Models
- Experiments

Detecting anomalies from noisy multivariate sensor data is hard even to experienced engineers

- Example: sensor data of a compressor of oil production system
  - Data taken under a normal operational condition
  - Noisy, nonstationary, heterogeneous, high-dimensional …
  - Hard to recognize useful patterns by human eye

- Data mining algorithms help capture major patterns embedded in the data





(simulation data)

Anomaly Localization

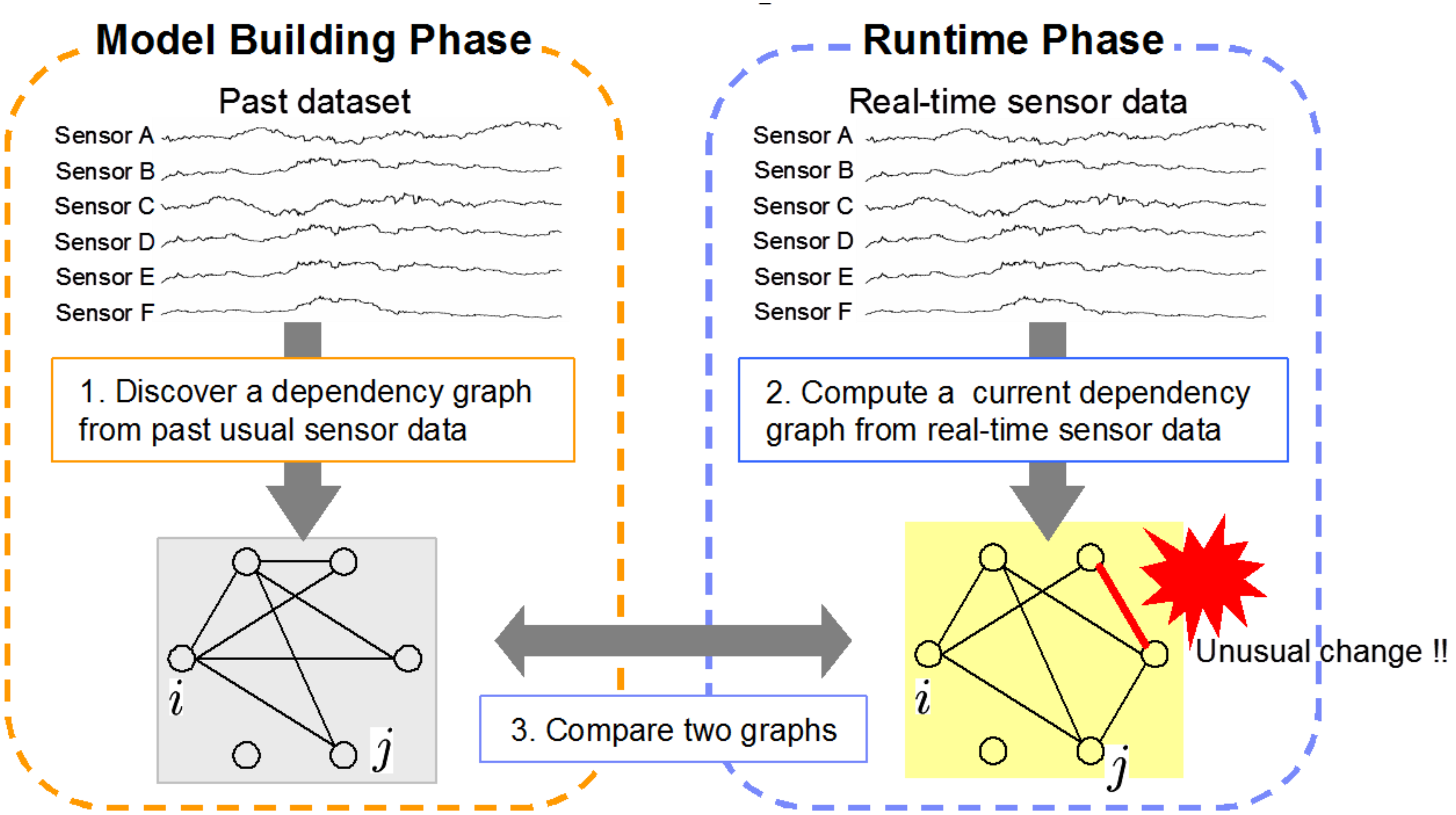
- The traditional **anomaly detection** is to compute the degree of anomalousness for a multivariate measurement, giving an overall anomaly score.

- **Anomaly localization** focuses on a variable-wise anomaly score, and two main lines of research have been proposed

  - sparse principal components analysis (PCA) to identify a set of variables that have nonzero weights in a subspace

  - graph-based anomaly localization approach, where two separate dependency graphs are inferred from training and testing data and an anomaly scoring method is used

Anomaly Localization for Multivariate Noisy Sensor Data



Detecting anomalies amongst sensors in real-world situations helps operators decide when and where maintenance is required

Outline

- Introduction to Anomaly Localization
- **L_0-constrained Gaussian Graphical Model**
- Optimization Algorithm for ℓ_0 Sparse Models
- Experiments

Sparsity-Constrained Gaussian Graphical Model

- **Problem**: Given an empirical covariance matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T$$

  find a sparse inverse covariance matrix **X** to represent the data

- **Classical convex approach:** Minimize the objective function

$$\min_{\mathbf{X} \succ \mathbf{0}} \; F(\mathbf{X}) + \lambda \|\mathbf{X}\|_1, \quad F(\mathbf{X}) = \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X})$$

$F(\mathbf{X})$ is the negative log likelihood function and the ℓ_1 term is a sparsity promoting regularizer.

Sparsity-Constrained Gaussian Graphical Model

- **Convex approach:** The ℓ_1 model minimizes

$$\min_{\mathbf{X}\succ\mathbf{0}} \quad \mathrm{tr}(\mathbf{S}\mathbf{X}) - \log\det(\mathbf{X}) + \lambda\|\mathbf{X}\|_1$$

- **Novel nonconvex approach:** We directly constrain sparsity. The ℓ_0 model minimizes

$$\min_{\mathbf{X}\succ\mathbf{0}} \quad f(\mathbf{X}) = \mathrm{tr}(\mathbf{S}\mathbf{X}) - \log\det(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2$$

$$\boxed{\text{s.t. } \|\mathbf{X}\|_0 \leq \kappa}$$

$$X_{i,j} = 0 \;\; \forall (i,j) \in \mathfrak{I}$$

κ : the maximally allowable number of nonzeros

$\mathfrak{I}$: the set of known conditionally independent variables

- It is a very challenging optimization problem: highly nonlinear, nonconvex

L1 Model versus l0 Model

$$\min_{\mathbf{X} \succ \mathbf{0}} \ \mathrm{tr}(\mathbf{S}\mathbf{X}) - \log\det(\mathbf{X}) + \lambda \|\mathbf{X}\|_1$$

$$\begin{aligned} \min_{\mathbf{X} \succ \mathbf{0}} \ & f(\mathbf{X}) = \mathrm{tr}(\mathbf{S}\mathbf{X}) - \log\det(\mathbf{X}) + \tfrac{\lambda}{2}\|\mathbf{X}\|_F^2 \\ \text{s.t. } & \|\mathbf{X}\|_0 \leq \kappa \\ & X_{i,j} = 0 \ \forall (i,j) \in \mathfrak{I} \end{aligned}$$

- ℓ_0 based models often recover sparsity pattern better than its ℓ_1 counterpart since ℓ_1 norm is just a relaxation of ℓ_0 norm
- The ℓ_0-constraint guarantees that the solution will admit a certain level of sparsity
- The ℓ_2-regularization term keeps the magnitude of all entries uniformly similar and encourages the capacity of selecting groups in the presence of highly correlated variables
- **Theorem:** The solution set of ℓ_0 model is bounded.

Outline

- Introduction to Anomaly Localization
- L_0-constrained Gaussian Graphical Model
- **Optimization Algorithm for ℓ_0 Sparse Models**
- Experiments

Notations

- Define the constraint set as

$$\Omega \triangleq \{\mathbf{X} \in \mathbb{R}^{n\times n} : \|\mathbf{X}\|_0 \le \kappa,\ \ X_{i,j} = 0,\ \forall (i,j) \in \mathfrak{I}\}$$

- The projection operator is

$$P_\Omega(\mathbf{X}) \triangleq \arg\min_{\mathbf{Y}\in\Omega} \|\mathbf{X}-\mathbf{Y}\|_F$$

➡ It is a low-cost operator.

- The gradient is

$$\nabla f(\mathbf{X}) = \mathbf{S} - \mathbf{X}^{-1} + \lambda \mathbf{X}$$

Gradient-projection Algorithm

- Consider

$$\min\{f(\mathbf{X}) : \mathbf{X} \in \mathbb{R}^{n\times n}, \|\mathbf{X}\|_0 \leq \kappa, X_{i,j} = 0,\ \forall (i,j) \in \mathfrak{I}, \mathbf{X} \succ 0, \mathbf{X} = \mathbf{X}^T\}$$

- Main idea:
  - Feasibility w.r.t membership in $\Omega = \{\|\mathbf{X}\|_0 \leq \kappa,\ X_{i,j} = 0, \forall (i,j) \in \mathfrak{I}\}$ is handled via projection
  - Symmetric positive-definiteness $\{\mathbf{X} \succ 0, \mathbf{X} = \mathbf{X}^T\}$ is ensured through a line-search procedure

Algorithm 1: Gradient projection - $GP(\mathbf{X}^0, \Omega, \lambda)$

Given parameters $\delta > 0, \sigma \in (0,1), [\alpha_{min}, \alpha_{max}] \subset (0, \infty)$. Set $k = 0$.
while *some stopping criteria not satisfied* **do**
  Step a: Initialize step size
    Choose $\alpha_0 \in [\alpha_{min}, \alpha_{max}]$
  Step b: Line search along projection arc
    Set $\alpha = \sigma^j \alpha_0$, where $j \geq 0$ is the smallest integer such that
    $f(\mathbf{X}^{k+1}) \leq f(\mathbf{X}^k) - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2$ and $\mathbf{X}^{k+1} \succ 0$, where
    $\mathbf{X}^{k+1} \leftarrow P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$
  Step c: $k \leftarrow k+1$ **and go to Step a.**

Convergence Analysis

- **Theorem:** Assume $\mathbf{X}^0$ is a feasible solution, let $\{\mathbf{X}^k\}$ be the sequence generated by Algorithm 1. Suppose $\mathbf{X}^*$ is an accumulation point of $\{\mathbf{X}^k\}$ Then the following hold.

  (i) The sequence $\{f(\mathbf{X}^k)\}$ admits an accumulation point.

  (ii) $\mathbf{X}^*$ is a strictly local minimizer.

Outline

- Introduction to Anomaly Localization
- L_0-constrained Gaussian Graphical Model
- Optimization Algorithm for ℓ_0 Sparse Models
- **Experiments**

Graph learning: Execution Time and Accuracy Comparison

			Proposed ℓ_0 model			Convex ℓ_1 model		
Test	n	κ	time	TNR	TPR	time	TNR	TPR
Rand	1000	16848	15.55	0.9985	0.9149	29.13	0.9961	0.7706
Rand	1500	25324	46.79	0.9992	0.9269	109.29	0.9972	0.7600
Rand	2000	34160	108.51	0.9993	0.9180	231.07	0.9978	0.7522
AR2	1000	4994	16.72	1	1	68.09	0.9983	0.6700
AR2	1500	7494	54.28	1	1	147.59	0.9989	0.6685
AR2	2000	9994	112.52	1	1	455.39	0.9992	0.6692
AR3	1000	4994	19.99	1	1	61.48	0.9980	0.7149
AR3	1500	7494	56.13	1	1	216.92	0.9987	0.7145
AR3	2000	9994	123.01	1	1	433.73	0.9990	0.7145

true negative rate (specificity) TNR :

$$\frac{\mathrm{TN}}{\mathrm{TN+FP}} = \frac{|\{(i,j) : X_{i,j}=0, \hat{S}_{i,j}=0\}|}{|\{(i,j) : \hat{S}_{i,j}=0\}|}$$

true positive rate (sensitivity) TPR :

$$\frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{|\{(i,j) : X_{i,j}\neq 0, \hat{S}_{i,j}\neq 0\}|}{|\{(i,j) : \hat{S}_{i,j}\neq 0\}|}$$

$\hat{\mathbf{S}}$ true covariance matrix

X estimated inverse covariance matrix

Anomaly Localization: Sparsely Supervised with L_0

	$\ell_0+\ell_2$+KL	$\ell_0+\ell_2$+QP	$\ell_0+\ell_2$+SNN	ℓ_1+KL	ℓ_1+QP	ℓ_1+SNN	ℓ_0+KL	ℓ_0+QP	ℓ_0+SNN	$\ell_1+\ell_2$+KL	$\ell_1+\ell_2$+QP	$\ell_1+\ell_2$+SNN
Sensor Error data												
mean	**0.9767**	0.9412	0.9637	0.9631	0.8334	0.9388	0.9448	0.8606	0.9551	0.9627	0.9536	0.9454
std	**0.0545**	0.1318	0.0606	0.1043	0.1933	0.0747	0.1039	0.1399	0.0650	0.1055	0.1095	0.0683
Sensor Error data with added noise												
mean	**0.9071**	0.8670	0.7635	0.8168	0.8418	0.6857	0.7371	0.6560	0.7350	0.8646	0.7952	0.6682
std	**0.0768**	0.1221	0.1156	0.1537	0.1128	0.1442	0.1786	0.2062	0.1390	0.1561	0.1517	0.1437
Sun Spot Sensor data												
mean	0.8849	**0.8917**	0.7914	0.7472	0.7744	0.5777	0.7976	0.7846	0.7744	0.7471	0.7913	0.5810
std	0.1537	**0.1461**	0.1682	0.2896	0.2372	0.2718	0.2057	0.2170	0.1721	0.2895	0.2214	0.2658
Sun Spot Sensor data with added noise												
mean	**0.8515**	0.8502	0.7336	0.7018	0.7211	0.5748	0.6638	0.7375	0.6015	0.7040	0.7278	0.5997
std	0.1691	**0.1677**	0.1832	0.2915	0.2739	0.2658	0.3128	0.2577	0.2899	0.2877	0.2538	0.3023

The mean and standard deviation for AUC values

Thank you!