Robust graphical modeling under data contamination

Christian Capezza, Antonio Lepore, Biagio Palumbo

Statistics for Engineering Research group (SFERe) www.sfere.unina.it
Department of Industrial Engineering, University of Naples Federico II

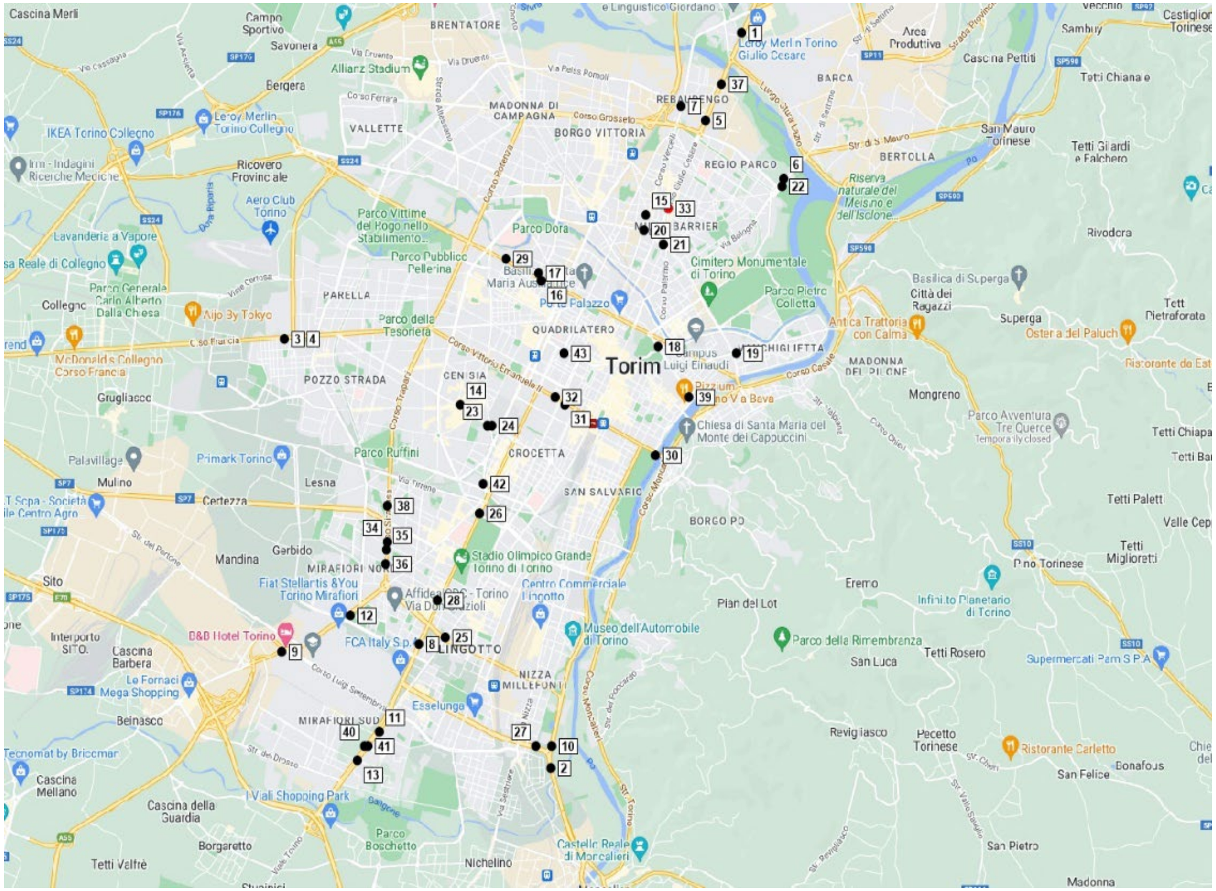
Monday, 7 September 2026, 12:40pm
Session «Statistical / Stochastic Modelling and Statistical Computing»

Outline

- ❑ **A Motivating Example**
- ❑ Introduction
- ❑ Literature Review
- ❑ Methodology
- ❑ Simulation Study
- ❑ Case Study
- ❑ Conclusions

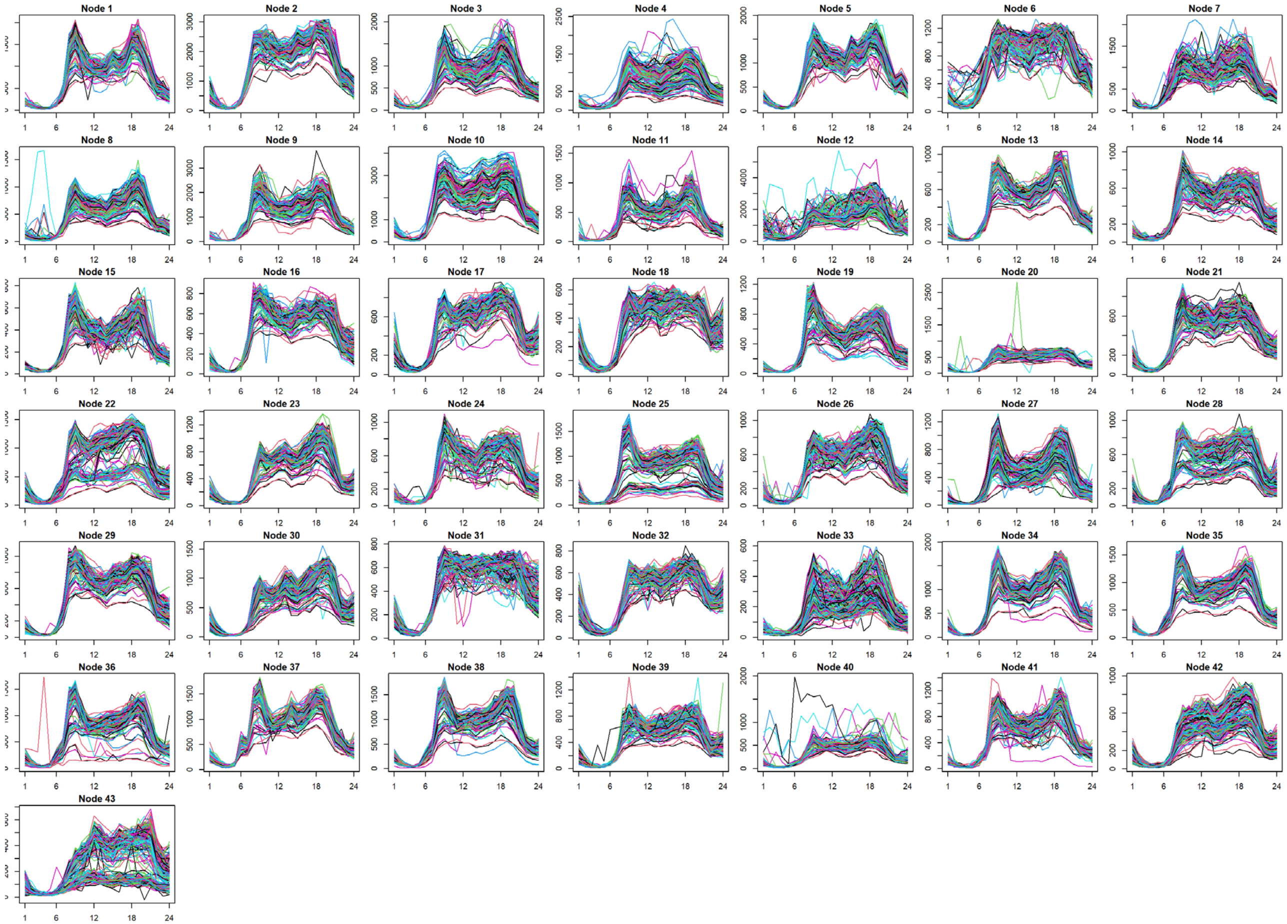
A Motivating Example: Monitoring Traffic Network

Map of Turin, Italy



A Motivating Example: Monitoring Traffic Network

Daily profiles for the 43 traffic nodes



A Motivating Example: Monitoring Traffic Network

1. **Network Inference**
   - ❑ How to properly model and interpret dependence among sensors?

2. **Monitoring**
   - ❑ How to use the network to detect anomalies in future observations?

Outline

❑ A Motivating Example

❑ **Introduction**

❑ Literature Review

❑ Methodology

❑ Simulation Study

❑ Case Study

❑ Conclusions

Introduction

- Multivariate data are observed simultaneously

$$\boldsymbol{X} = \left(X_1, \dots, X_p\right)^\top,\ \boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- The mean describes central behaviour of data
  - However, focusing only on averages ignores how variables interact

- The *covariance matrix* provides information about the relationships between variables
  - However, by itself, it is not sufficient to understand the true dependency structure
    - It does **not distinguish direct from indirect dependencies**

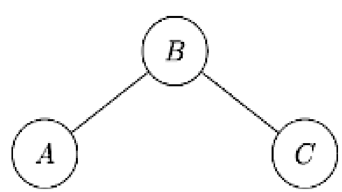
Introduction

Gaussian Graphical Models

- Key concept → **Precision matrix:** $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$
  - Conditional independence between X_i and X_j, given all other variables, corresponds to a zero entry in the precision matrix

$$\Theta_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid \textit{all other variables}$$

- Represent relationships using an *undirected* graph $G(V, E)$
  - *Nodes* V → variables
  - *Edges* E → $\Theta_{ij} \neq 0$



$A \perp\!\!\!\perp C \mid B$

- Particularly useful in high-dimensional settings

Introduction

Obtaining the Precision matrix

- $\boldsymbol{\Theta}$ can be estimated via penalized maximum-likelihood estimation

$$\arg\max_{\Theta \succ 0} \left\{ \underbrace{\log\det\Theta - \mathrm{tr}(S\Theta)}_{log-likelihood} - \underbrace{\lambda \sum_{j \neq l} |\Theta_{jl}|}_{penalty} \right\}$$

Where:
S is the sample covariance matrix
$\lambda > 0$ is the penalty parameter

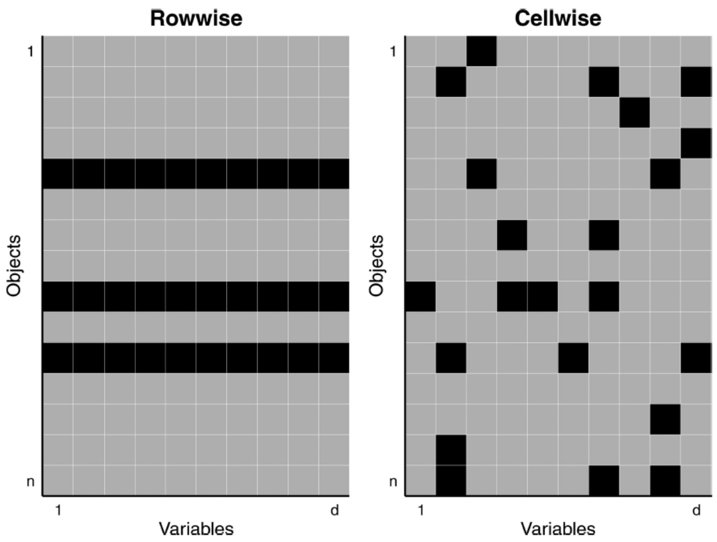
- This results in
  - a sparse precision matrix
  - a **sparse Graphical Model** which is easy to interpret and focuses on the most relevant relationships

- The solution of this problem is known as the **Graphical LASSO (glasso)** *(Friedman et al., 2008).*

Introduction

The Problem of Outliers

- ❑ Real-world data often contain outliers due to
  - ○ Measurement errors
  - ○ Anomalous events
- ❑ Outliers can be classified as
  - ○ **Rowwise**: entire observations are contaminated
  - ○ **Cellwise**: only individual entries are contaminated
    - • Harder to detect and handle



(Rousseeuw and Bossche, 2018)

- ❑ Classical estimators assume clean data and are *highly sensitive* to outliers
  - ○ Outliers can lead to unrealistic and unstable graph structures

Introduction

Robust Graphical Models

- **Robust statistics** aims to produce reliable estimates under data contamination
  - *Robust* = outlier resistant
  - Contamination is inevitable in high-dimensional data (often cellwise)
    - Outliers cannot be simply removed (loss of information)
  - Anomalies are handled, not discarded

- **Robust graphical models** preserve the notion of conditional dependence while reducing the influence of anomalous observations

Outline

- A Motivating Example
- Introduction
- **Literature Review**
- Methodology
- Simulation Study
- Case Study
- Conclusions

Literature Review

Exisisting robust approaches

- **Distribution-based methods**
  - *t-distribution based models (tlasso)* (Finegold & Drton, 2011)
    - Heavy-tailed distributions reduce the influence of extreme observations
  - *γ-lasso* (Hirose et al., 2017)
    - Replaces standard log-likelihood with a divergence measure to reduce the influence of outliers

- **Covariance-based methods**
  - *Robust covariance-based graphical models* (Ollerer & Croux, 2015)
    - Replace the sample covariance matrix with a robust estimator

- **Data-cleaning approaches**
  - *Trimmed Graphical Lasso* (Yang & Lozano, 2015)
    - A fraction of the most extreme observations is removed during estimation
  - *Winsorized Graphical Lasso* (Lafit et al., 2022)
    - Extreme values are shrunk toward less extreme ones

Outline

- ❑ A Motivating Example
- ❑ Introduction
- ❑ Literature Review
- ❑ **Methodology**
- ❑ Simulation Study
- ❑ Case Study
- ❑ Conclusions

Methodology: Key Steps

Data Matrix $\mathbf{X}$
↓
① Robustification
↓
Robust Covariance $\mathbf{S}$
↓
② Adaptive Graphical Lasso
↓
Precision Matrix $\widehat{\boldsymbol{\Theta}}$
↓
③ Model Selection
↓
Final Graphical Model

① **Detect Deviating Cells (DDC)** *(Rousseeuw and Bossche, 2018)*
- Predicts each cell using correlated variables
- Flag large cellwise residuals
- Replaces only corrupted entries

② **Adaptive Graphical Lasso** *(Zou, 2006; Fan et al., 2009)*
- Data-driven penalties to control sparsity vs bias
  - Strong edges → less penalized
  - Weak edges → more penalized

③ **Stability Selection** *(Meinshausen & Buhlmann, 2010)*
- Fit the model on random subsamples
- Keep edges with selection probability > 90%
- Reduce overfitting

Outline

Simulation Study

- **Goal**
  - Evaluate the ability of the proposed methods to **recover the true graph structure** under different contamination scenarios

Data Generating Process

- **Baseline model**
  - $\boldsymbol{X} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with Fixed dimensions $n = 300$ and $p = 50$
  - $(\mathrm{AR}(1))$ covariance structure: $\Sigma_{ij} = \rho^{|i-j|}$ with $\boldsymbol{\rho} \in \{0.3, 0.5, 0.9\}$

- **Contamination mechanism**
  - Contamination scenarios: rowwise or cellwise
  - Contamination proportion $\boldsymbol{\epsilon} \in \{0.05, 0.10, 0.20\}$ (here we focus on 0.10 only)
  - Contamination intensity $\boldsymbol{\delta} \in \{0,1,2,3\}$

Simulation Study: Essential Results ($\rho = 0.5$)

F1-Score

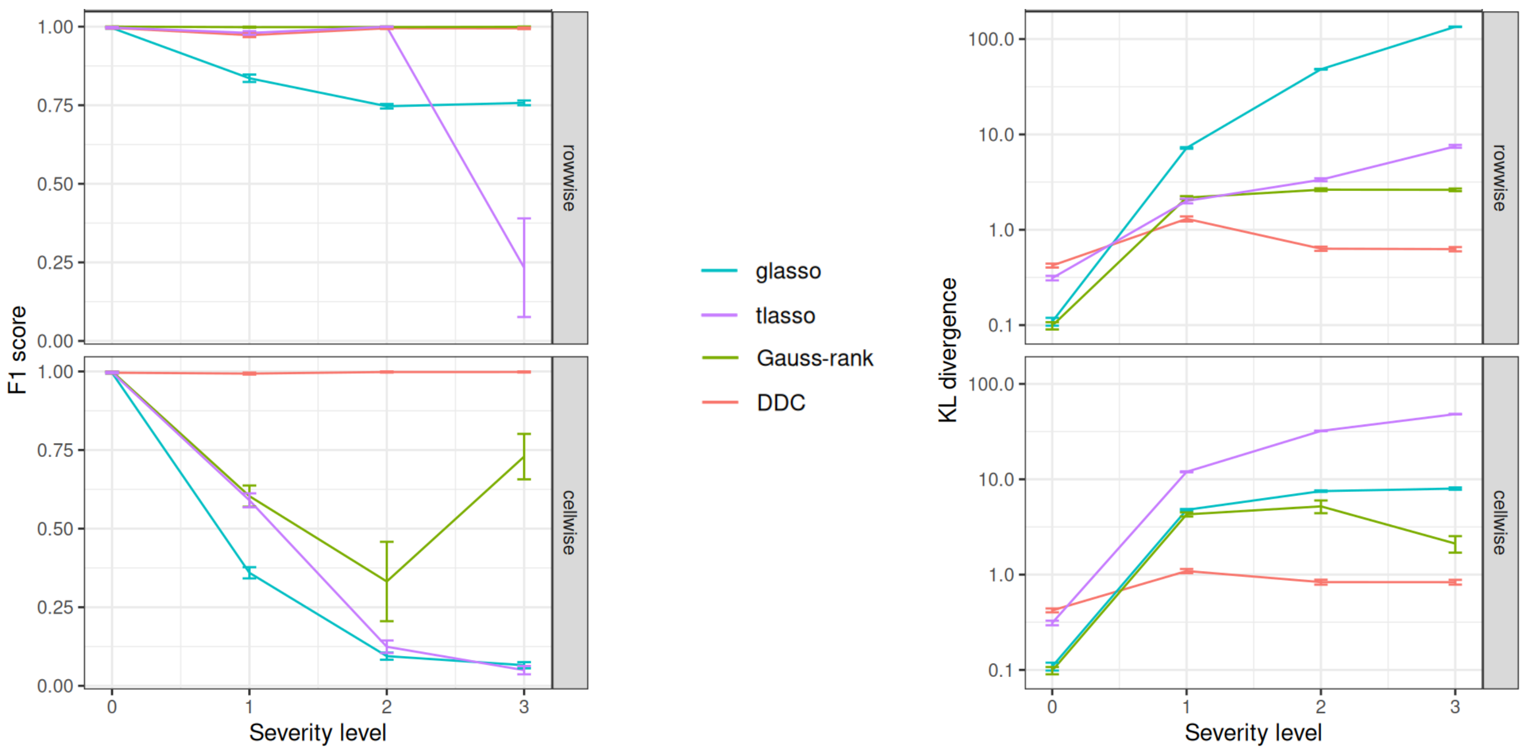
$$\mathrm{F1} = 2\frac{precision \cdot recall}{precision + recall}$$

The *higher* the better

KL Divergence

$$KL(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = \mathrm{tr}(\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{\Theta}}) - log\ det(\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{\Theta}}) - p$$

The *lower* the better



Outline

- ❑ A Motivating Example
- ❑ Introduction
- ❑ Literature Review
- ❑ Methodology
- ❑ Simulation Study
- ❑ **Case Study**
- ❑ Conclusions

Case Study: Traffic Flow Data

- **Objectives**
  1. Characterize dependencies between sensors
  2. Use the network for *process monitoring*

- **Data description**
  - $p = 43$ sensor nodes (Turin, Italy)
  - Data collected between the years 2018 - 2019

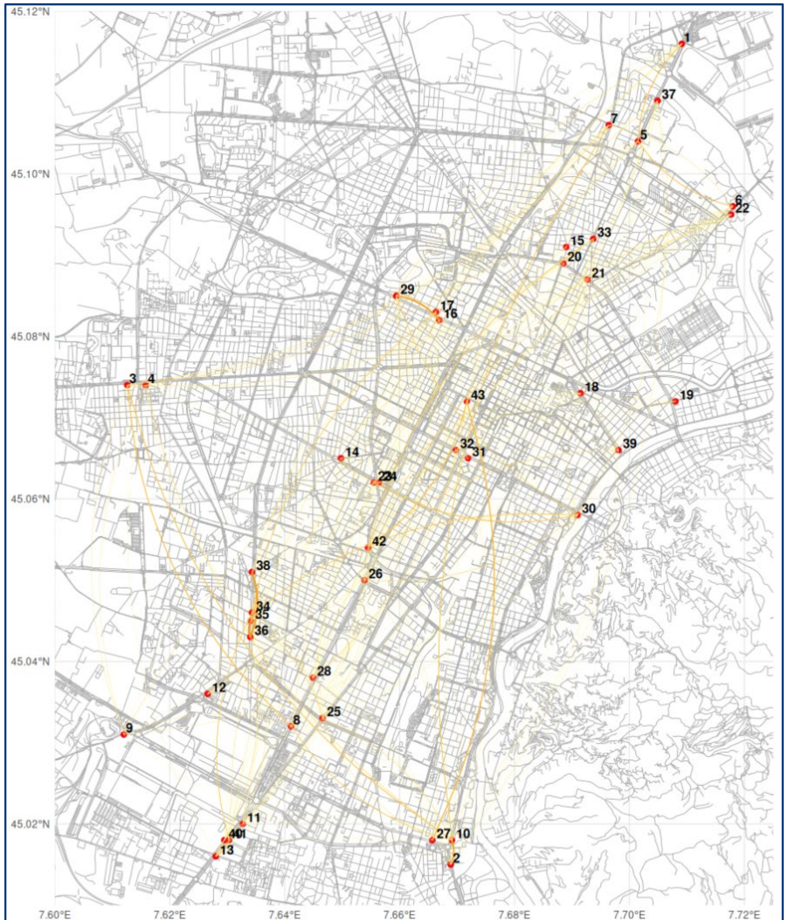
- **Data structure**
  - *24 hourly* graphical models for each dataset
    - Working days $(298x43)$
    - Saturdays $(69x43)$
    - Holidays $(79x43)$

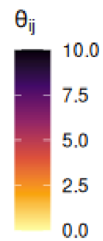
- **Methods**
  - Classical approach (glasso)
  - *Proposed framework (DDC)*

Case Study: Traffic Flow Network Analysis

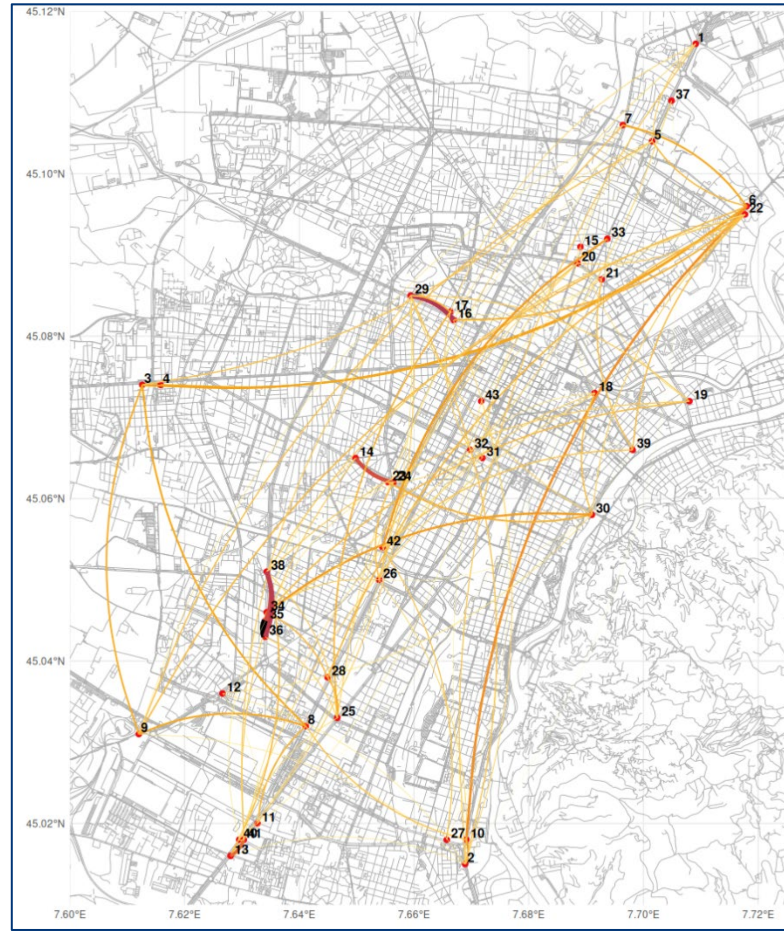
Working days, hour 14

Classic Approach (glasso)





Proposed Framework (DDC)



Case Study: Monitoring Traffic Flow

- Monitoring performed using *statistical process monitoring*
  - **Phase I**: estimate at each hour of the day h the in-control behaviour of the system
    - mean vector $\boldsymbol{\mu}_h$
    - precision matrix $\boldsymbol{\Theta}_h$
  - **Phase II**: monitor new observations (at day d, hour h) with the Hotelling $T^2_{d,h}$ statistic:

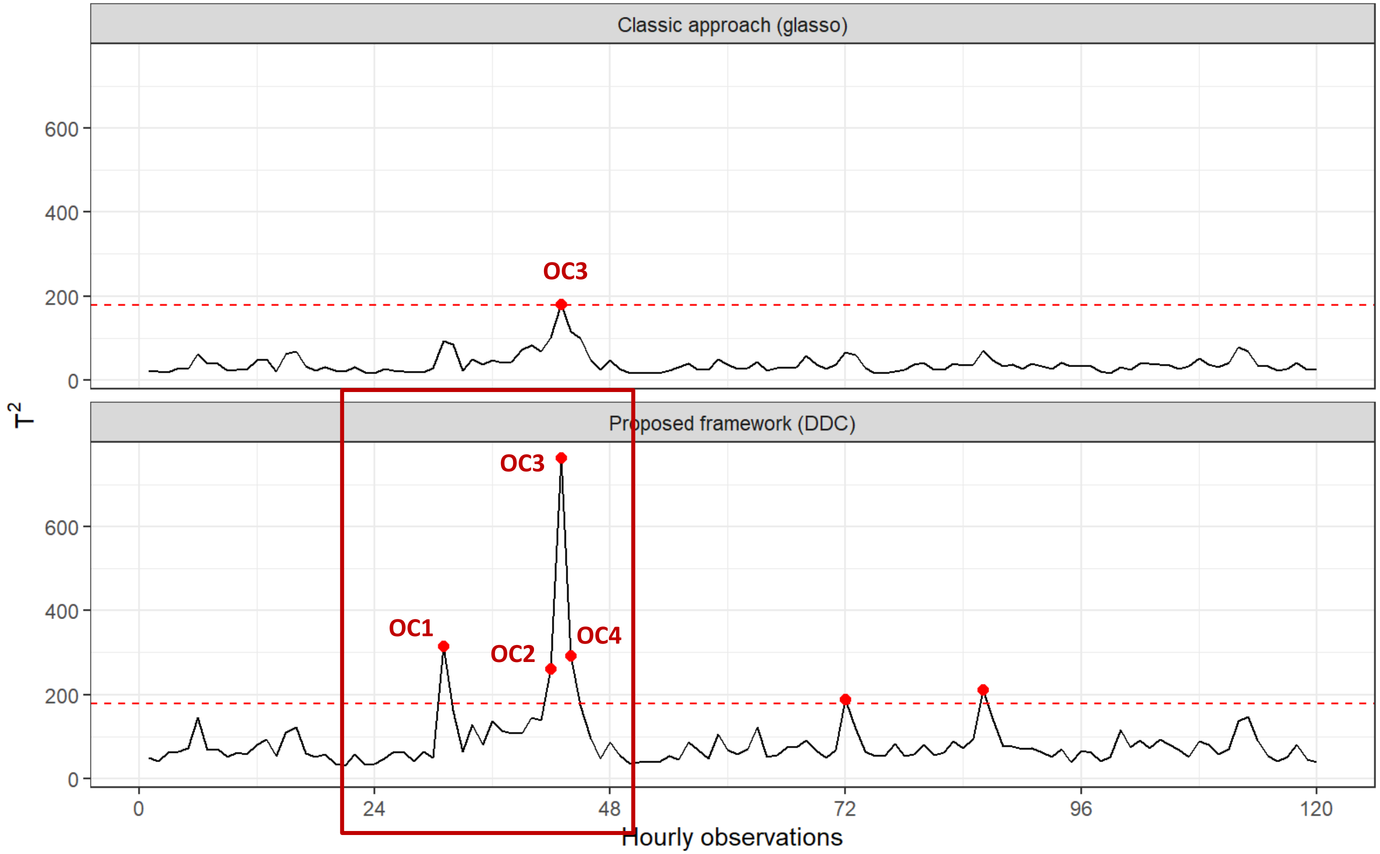
$$T^2_{d,h} = (\mathbf{x}_{d,h} - \widehat{\boldsymbol{\mu}}_h)^\top \widehat{\boldsymbol{\Theta}}_h (\mathbf{x}_{d,h} - \widehat{\boldsymbol{\mu}}_h)$$

- If $T^2_{d,h}$ exceeds the Upper Control Limit (UCL) → alarm

$$\mathrm{UCL} = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha},$$

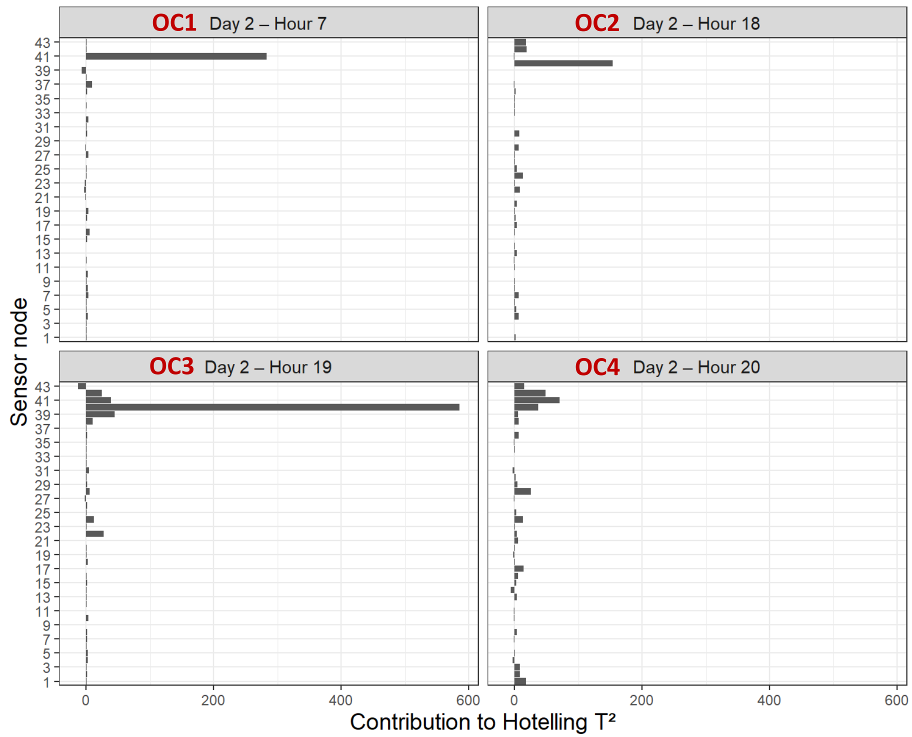
Case Study: Monitoring Traffic Flow

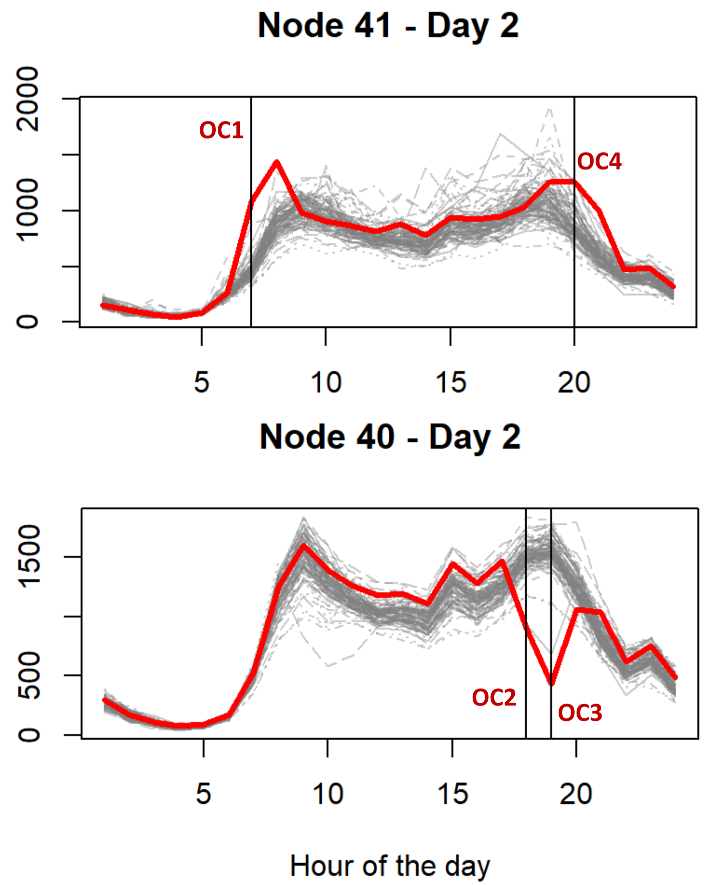
Phase II Control Charts



Case Study: Monitoring Traffic Flow

Detecting OC Alarms





Outline

- ❑ A Motivating Example
- ❑ Introduction
- ❑ Literature review
- ❑ Methodology
- ❑ Simulation Study
- ❑ Case Study
- ❑ **Conclusions**

Conclusions

- **Impact of outliers**
  - Classical graphical model estimation is highly sensitive to anomalous observations
  - Robust approaches improve structural recovery and stability

- **Simulation results**
  - DDC-based framework consistently outperforms competitors

- **Real-world application**
  - The robust framework provides more interpretable and stable traffic networks
  - Improved anomaly detection, revealing events masked by classical approach

- **Future work**
  - Extension to dynamic graphical models to capture time-varying dependencies

Acknowledgements

The research work of C. Capezza was funded by the Department of Industrial Engineering at the University of Naples Federico II, within the project SMADI - Statistical Methods for Anomaly Detection in Industry 4.0.

References

Finegold, M., and M. Drton. "Robust Graphical Modeling of Gene Networks using Classical and Alternative T-distributions." *The Annals of Applied Statistics*, Vol. 5, No. 2A, 2011

Friedman, J., T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, Vol. 9, No. 3, 2008

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Second Edition, Stanford, California, 2008

Hirose, K., Fujisawa, H., and Sese, J. "Robust sparse gaussian graphical modeling." *Journal of Multivariate Analysis*, 172–190,2017

Lafit, G., Nogales, J., Ruiz, M., and Zamar, R. "Robust and Sparse Estimation of Graphical Models Based on Multivariate Winsorization." *Springer International Publishing*, 2023

Liu, H., Roeder, K., and Wasserman, L. "Stability approach to regularization selection (stars) for high dimensional graphical models." *Advances in neural information processing systems*, 23, 2010

Meinshausen, N., and P. Bühlmann. "High-dimensional Graphs and Variable Selection with the Lasso." *The Annals of Statistics*, Vol. 34, No. 3, 2006

Meinshausen, N., and P. Bühlmann. "Stability selection." *Journal of the Royal Statistical Society* Series B: Statistical Methodology 72.4, 2010

Montgomery, D. C. *Introduction to Statistical Quality Control*. John Wiley & Sons, Hoboken, NJ, 8th edition, 2019

Öllerer, V., and C. Croux. "Robust high-dimensional precision matrix estimation." *Modern nonparametric, robust and multivariate methods*, 2015

Rousseeuw, P. J. and Bossche, W. V. D. "Detecting deviating data cells." *Technometrics*, 60(2):135–145, 2018

Yang, E., and A. Lozano. "Robust Gaussian graphical modeling with the trimmed graphical lasso." *Advances in neural information processing systems* 28, 2015

Zou, H. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association*, Vol. 101, No. 476, 2006

Robust graphical modeling under data contamination

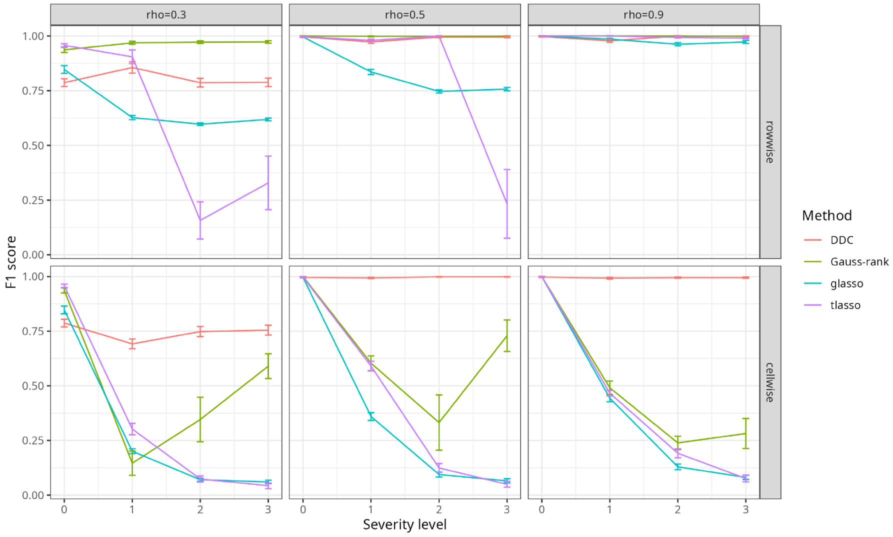
Christian Capezza, Antonio Lepore, Biagio Palumbo

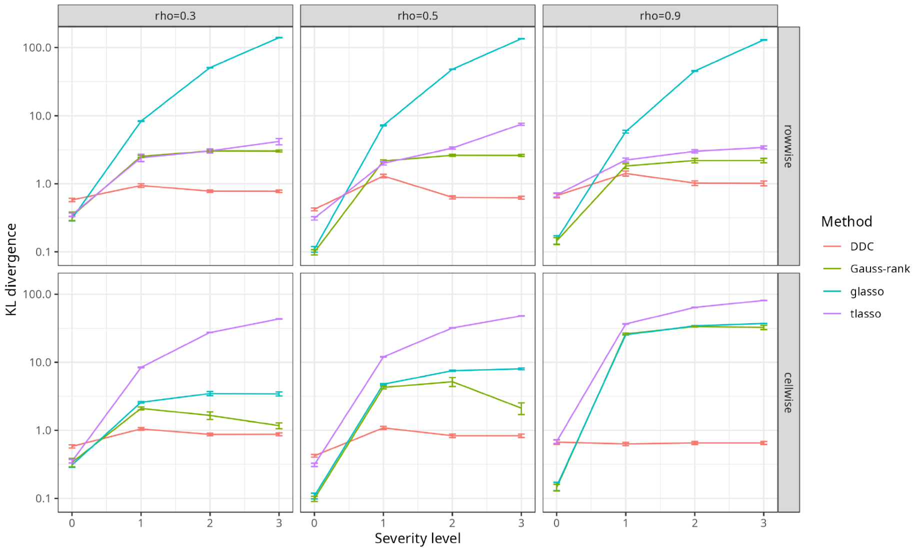
Department of Industrial Engineering, University of Naples Federico II

Monday, 7 September 2026, 12:40pm
Session «Statistical / Stochastic Modelling and Statistical Computing»

Additional results

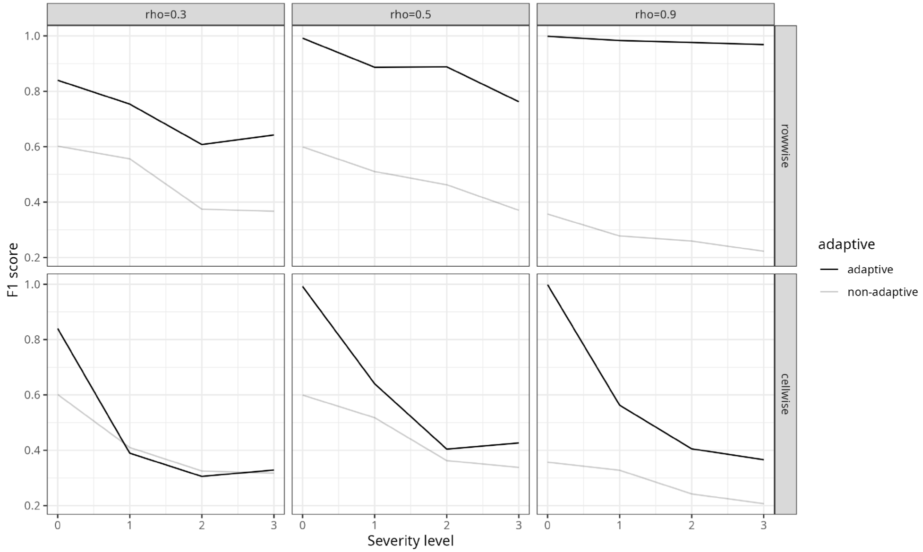
More complete results (with varying correlation parameter ρ, with contamination probability $\epsilon = 0.10$)

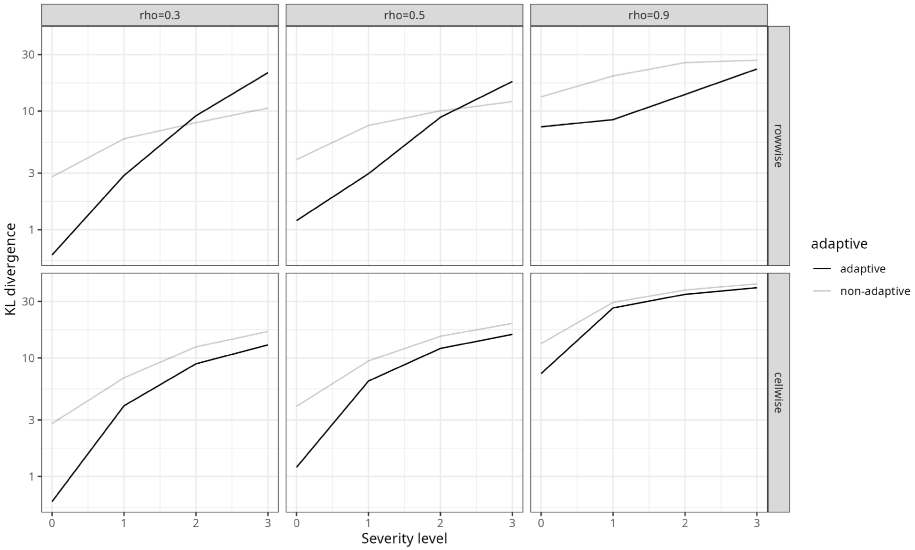




Additional results

Effect of adaptive glasso vs non-adaptive





Additional results

Effect of model selection strategy