Estimating Structured Vector Autoregressive Models

Igor Melnyk and Arindam Banerjee

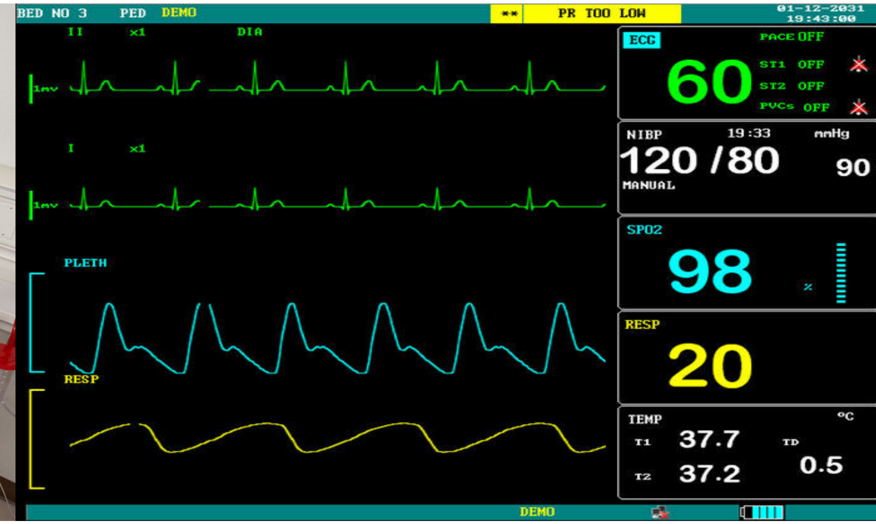
Department of Computer Science & Engineering

University of Minnesota, Twin Cities

International Conference on Machine Learning
New York, NY

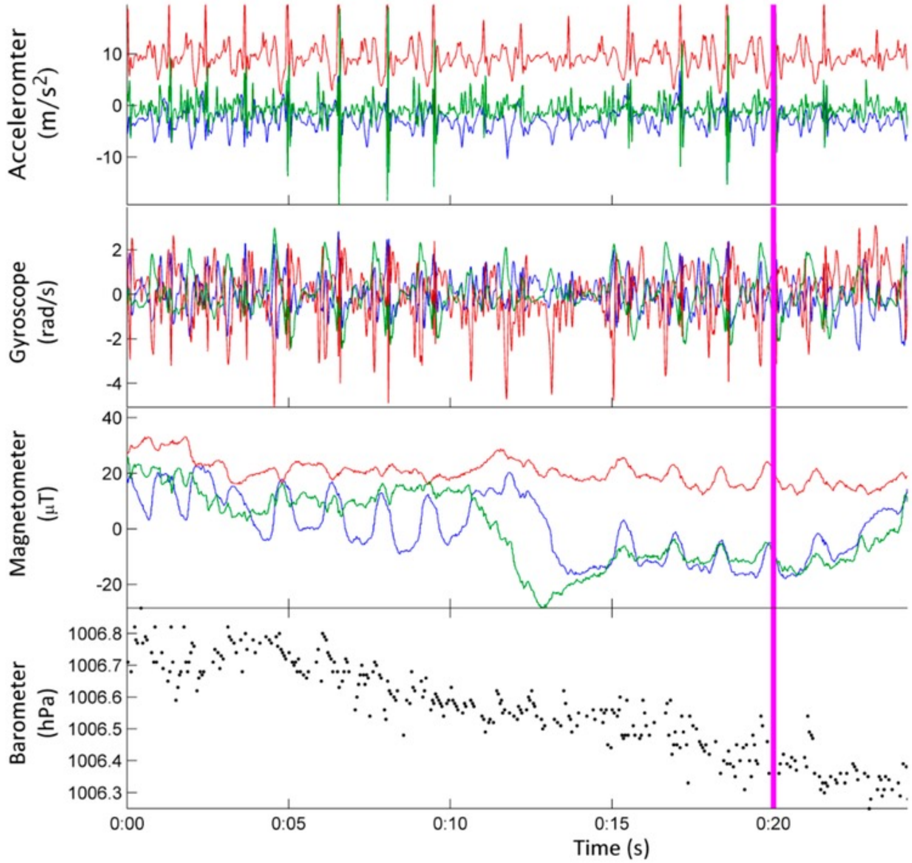
June 21, 2016

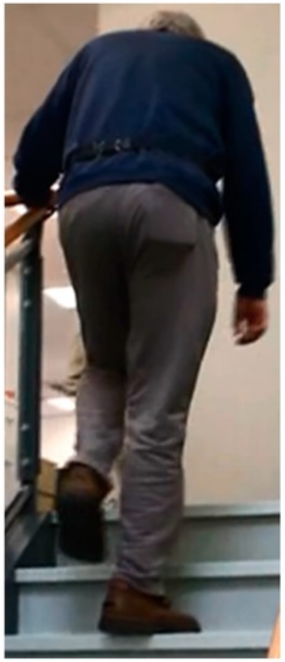
Healthcare



- Data
  - Multiple records of vital signs: blood pressure, temperature, pulse
- Objective
  - Monitor patient's health

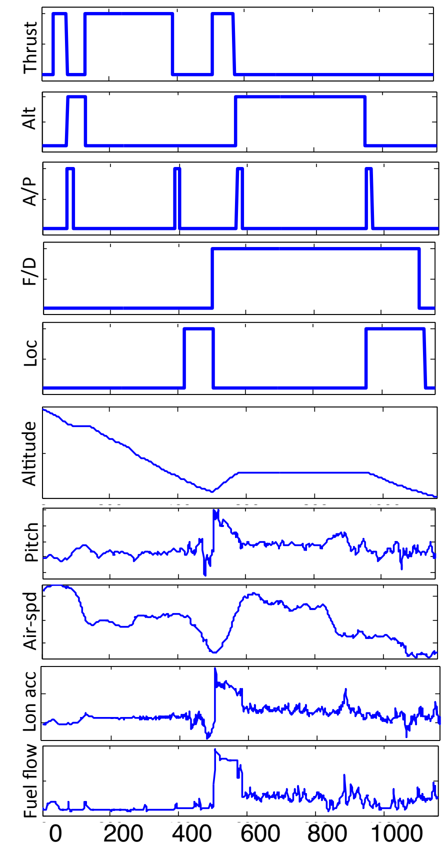
Activity Recognition





- Data
  - Multiple wearable sensors: accelerometer, gyroscope, barometer
- Objective
  - Track person's activity

Aviation Systems



- Data
  - Flight sensors, pilot commands, weather information
- Objective
  - Monitor flight, detect anomalous activity

Data Modeling

- Data
  - Dynamic, multivariate
- Objective
  - Monitor activity, make predictions

- Vector AutoRegressive model (VAR) *[Lutkepohl '07]*

$$x_t = A_1 x_{t-1} + \cdots + A_d x_{t-d} + \epsilon_t$$

  - $x_t \in \mathbb{R}^p$ - multivariate time series
  - $A_k \in \mathbb{R}^{p \times p}$ - model parameters, $d \geq 1$ - order of the model
  - $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ - Gaussian noise: $\mathbb{E}(\epsilon_t \epsilon_t^T) = \Sigma$, $\mathbb{E}(\epsilon_t \epsilon_{t+\tau}^T) = 0$, $\tau \neq 0$

Estimation Problem

- Estimate A_k's
- Let $(x_0, x_1, \ldots, x_T)$ be VAR output across $T+1$ steps

$$x_d = A_1 x_{d-1} + \cdots + A_d x_0 + \epsilon_d$$
$$\vdots$$
$$x_T = A_1 x_{T-1} + \cdots + A_d x_{T-d} + \epsilon_T$$

$$\underbrace{\begin{bmatrix} x_d^T \\ \vdots \\ x_T^T \end{bmatrix}}_{Y \in \mathbb{R}^{N \times p}} = \underbrace{\begin{bmatrix} x_{d-1}^T & \cdots & x_0^T \\ \vdots & \ddots & \vdots \\ x_{T-1}^T & \cdots & x_{T-d}^T \end{bmatrix}}_{X \in \mathbb{R}^{N \times dp}} \underbrace{\begin{bmatrix} A_1^T \\ \vdots \\ A_d^T \end{bmatrix}}_{B \in \mathbb{R}^{dp \times p}} + \underbrace{\begin{bmatrix} \epsilon_d^T \\ \vdots \\ \epsilon_T^T \end{bmatrix}}_{E \in \mathbb{R}^{N \times p}}$$

$$\boxed{N = T - d + 1}$$

$$Y = XB + E$$

Estimation Problem

- Estimate A_k's

$$Y = XB + E$$

vectorize

$$\underbrace{\mathrm{vec}(Y)}_{\mathbf{y} \in \mathbb{R}^{Np}} = \underbrace{(I_{p \times p} \otimes X)}_{Z \in \mathbb{R}^{Np \times dp^2}} \underbrace{\mathrm{vec}(B)}_{\boldsymbol{\beta} \in \mathbb{R}^{dp^2}} + \underbrace{\mathrm{vec}(E)}_{\boldsymbol{\epsilon} \in \mathbb{R}^{Np}}$$

$$\mathbf{y} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Regularized estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{dp^2}}{\operatorname{argmin}} \frac{1}{2N} ||\mathbf{y} - Z\boldsymbol{\beta}||_2^2 + \lambda_N R(\boldsymbol{\beta})$$

$R(\cdot)$ - regularization norm $\lambda_N > 0$ - regularization parameter

Regularized Estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{dp^2}}{\operatorname{argmin}} \frac{1}{2N} ||\mathbf{y} - Z\boldsymbol{\beta}||_2^2 + \lambda_N R(\boldsymbol{\beta})$$

- Examples of regularizations
  - $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{dp} |\beta_i|$ - Lasso
  - $\|\boldsymbol{\beta}\|_{GL} = \sum_{k=1}^{K} \|\beta_{G_k}\|_2$ - Group Lasso
  - $\|\boldsymbol{\beta}\|_{SGL} = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_{GL}$ - Sparse Group Lasso
  - $\|\boldsymbol{\beta}\|_{OWL} = \sum_{i=1}^{dp} c_i|\beta|_{(i)}$ for $c_1 \geq \ldots \geq c_{dp} \geq 0$ - Order Weighted Lasso (OWL)
- Main properties
  - Samples $\{y_i, z_i\}$ are correlated
  - $R(\cdot)$ - general regularization norm
- Questions
  - How many samples $\{y_i, z_i\}$ needed to get accurate estimate $\hat{\beta}$?
  - How to select λ_N ?

Related Work

- Linear Regression
  - Main assumption: data is i.i.d.
  - *[Wainwright '09, Meinshausen et al. '09, Bickel et al. '09]* $R(\cdot)$ is L_1
  - *[Negahban et al. '12]* $R(\cdot)$ is any decomposable norm
  - *[Banerjee et al. '14]* $R(\cdot)$ is any norm

- VAR
  - Most work is focused on L_1 regularization
  - *[Loh et al. '11]* $R(\cdot)$ is L_1; considered only first-order VAR
  - *[Song & Bickel '13]* $R(\cdot)$ is L_1 and group L_1; assumptions on data dependency
  - *[Han & Liu '13]* L_1- based formulation under Gaussian noise
  - *[Kock & Callot '15]* $R(\cdot)$ is L_1; exploited martingale property o data
  - *[Basu et al. '15]* $R(\cdot)$ is L_1; any-order VAR; spectral analysis of VAR

Our Work

- Establish estimation guarantees for VAR under general $R(\cdot)$

- Our approach is based on
  - Error analysis framework *[Chandrasekaran ’12, Amelunxen ’13, Banerjee ’14]*
    - Restricted eigenvalue condition
    - Regularization of parameter characterization
  - Generic chaining argument *[Talagrand ’06, Mendelson ’07]*
    - Notion of Gaussian width
  - VAR spectral analysis *[Basu et. al. ’15]*
    - Characterize correlation structure of the data
  - Martingale properties of data *[Lutkepohl ’07, Shamir ’11]*
    - Bound sequential dependencies in the data

Error Analysis Framework

- Return back to our estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{dp^2}}{\operatorname{argmin}} \frac{1}{2N}||\mathbf{y} - Z\boldsymbol{\beta}||_2^2 + \lambda_N R(\boldsymbol{\beta})$$

- Denote error between true and estimated parameter

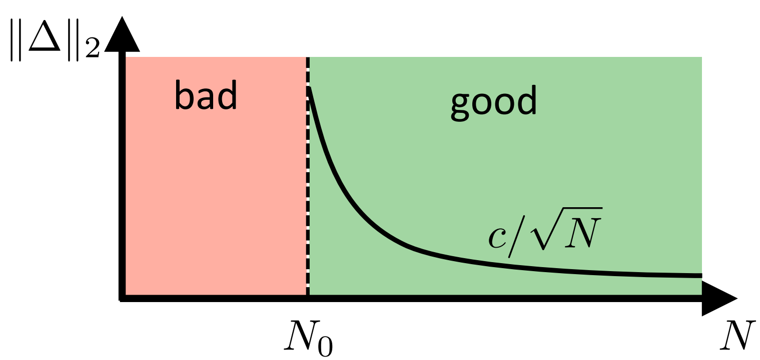
$$\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$$

- Our task
  - Establish conditions on
    - N (sample size)
    - λ_N (regularization parameter)
  - Bound the error

$$\|\Delta\|_2 \leq \delta, \delta > 0$$

Results

- Select number of data samples such that $N \geq \mathcal{O}(w^2(\Theta))$
  - $w(\Theta)$ - Gaussian width of an error set
- Choose regularization parameter such that $\lambda_N \geq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}}\right)$
  - $w(\Omega_R)$- Gaussian width of unit norm ball
- Norm of estimation error is then bounded by $\|\Delta\|_2 \leq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}}\right)\Psi$
  - High probability statement
  - Norm compatibility constant: $\Psi = \sup_{U \in \text{cone}(\Omega_E)} \frac{R(U)}{||U||_2}$



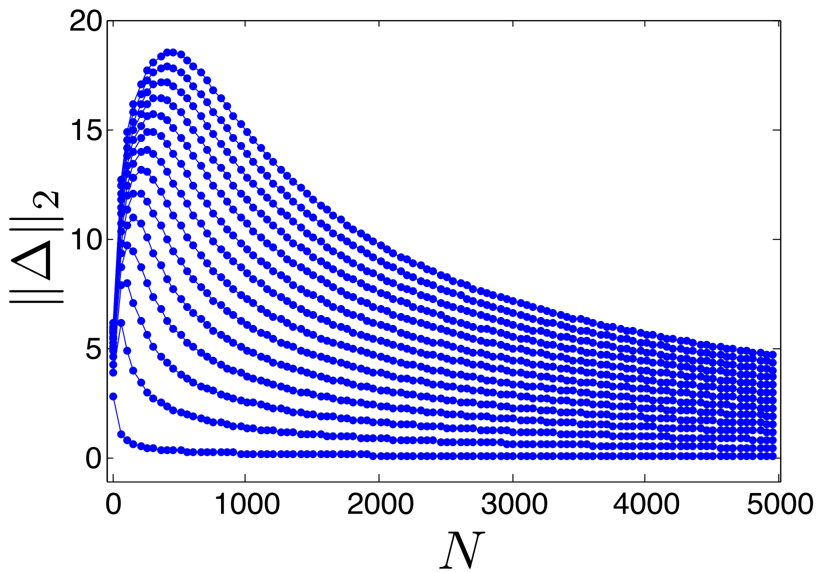
Special Cases

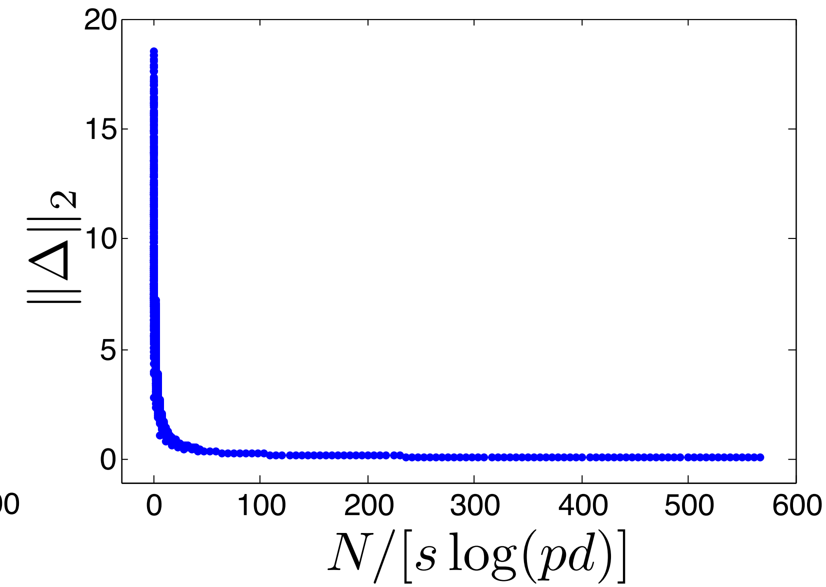
- Examples (VAR regularized estimation)
  - $\|\Delta\|_2 \leq \mathcal{O}\left(\sqrt{\frac{s\log(dp)}{N}}\right)$ - Lasso
  - $\|\Delta\|_2 \leq \mathcal{O}\left(\sqrt{\frac{s_G(m+\log(K))}{N}}\right)$ - Group Lasso
  - $\|\Delta\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\alpha s\log(dp)+(1-\alpha)s_G(m+\log(K))}{N}}\right)$ - Sparse Group Lasso
  - $\|\Delta\|_2 \leq \mathcal{O}\left(\frac{2c_1}{\bar{c}}\sqrt{\frac{s\log(dp)}{\bar{c}N}}\right)$ - Order Weighted Lasso

s - sparsity
s_G - group sparsity
K - number of groups
m - size of largest group
$\bar{c} = \frac{1}{n}\sum_{i=1}^{dp} c_i$
$\alpha \in [0,1]$

Experiments: Synthetic Data

- Investigate scaling of errors and lambda
  - Simulated first-order VAR
  - Parameters: $p \in [10, 600],\ s \in [4, 260],\ N \in [10, 5000]$
- Lasso

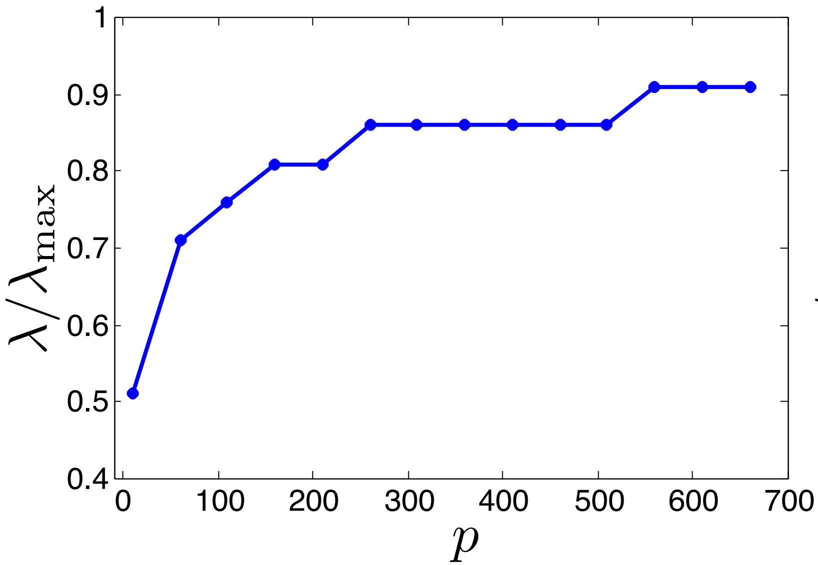


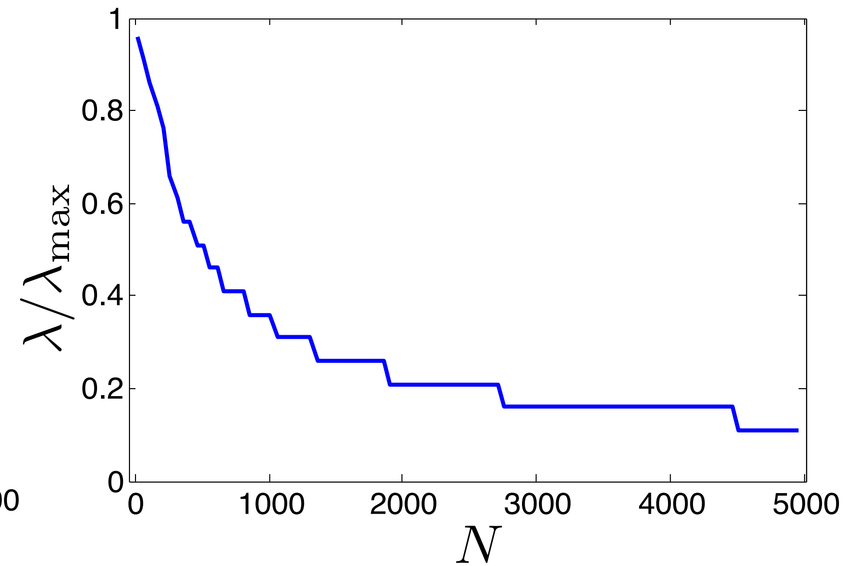


$$\|\Delta\|_2 = \mathcal{O}\left(\sqrt{\frac{s\log(dp)}{N}}\right)$$

Experiments: Synthetic Data

- Investigate scaling of errors and lambda
  - Simulated first-order VAR
  - Parameters: $p \in [10, 600],\ s \in [4, 260],\ N \in [10, 5000]$
- Lasso

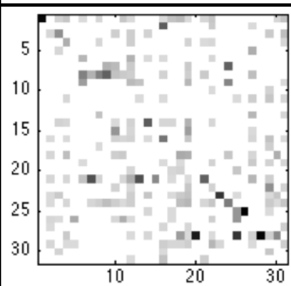
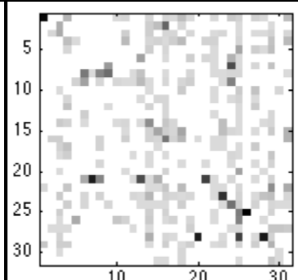
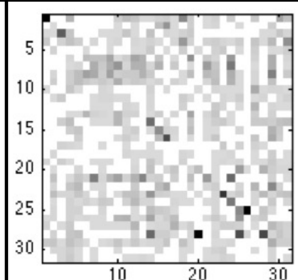
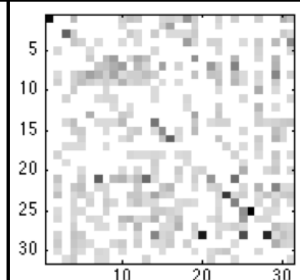
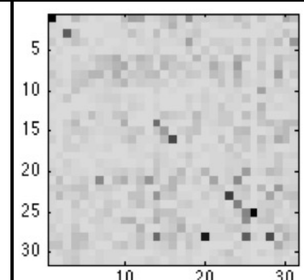




$$\lambda_N = \mathcal{O}\left(\sqrt{\frac{\log(dp)}{N}}\right)$$

Experiments: Aviation Data

- Compare different VAR regularizations
  - First-order VAR
  - Norms: Lasso, Group Lasso, Sparse Group Lasso, OWL, Ridge
- NASA flight dataset
  - Selected 300 flights, 31 parameters, sampled at 1Hz; landing part of flight

MSE	32.2	32.2	32.7	32.2	33.5
Sparsity	32.7	44.5	75.3	38.4	99.9
Sparsity Pattern					
Regularization Norm	Lasso	OWL	Group Lasso	Sparse Group Lasso	Ridge

Thank you!