Feature Selection under Multicollinearity & Causal Inference on Time Series

Indranil Bhattacharya

M.Sc(Engg.), CSA, IISc

Advisor : Dr. Arnab Bhattacharyya

November 20, 2017

Outline

High-dimensional Model

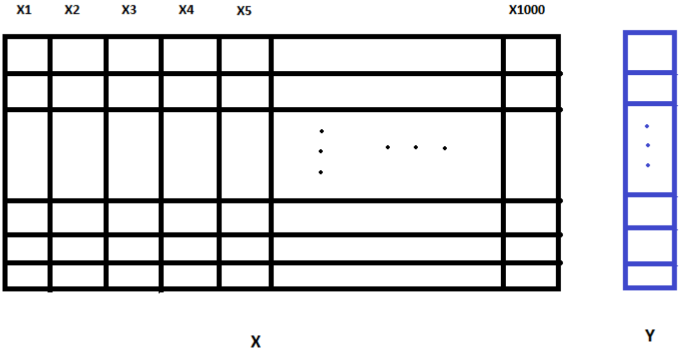


Figure: Model Estimation : $Y \approx f(X)$

Feature Selection ⇔ Sparse Estimation

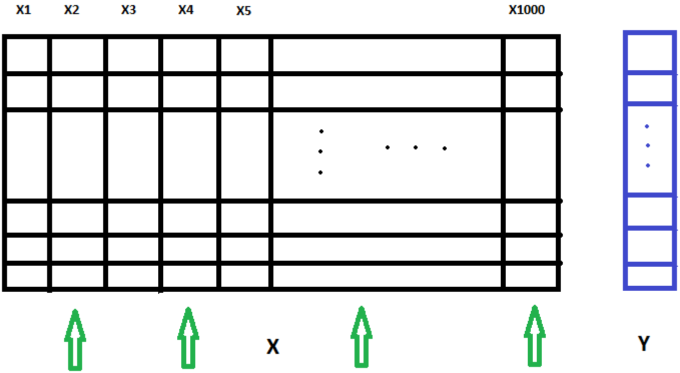


Figure: Sparse Estimation : $Y \approx f(\hat{X})$

Multicollinearity $\Leftrightarrow$ High Correlation

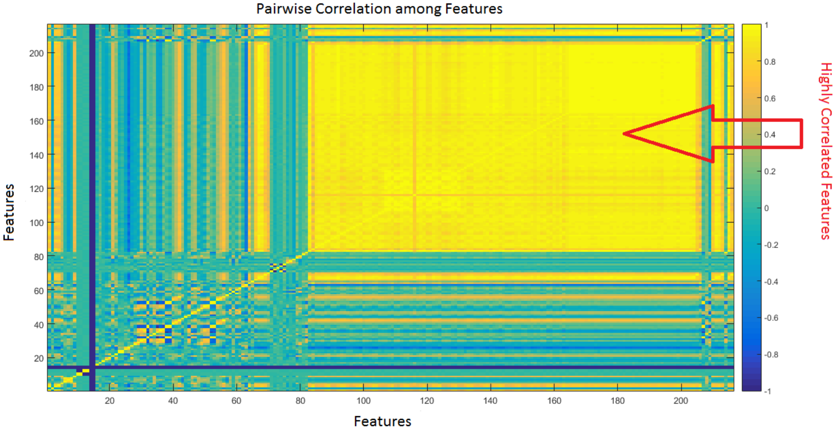


Figure: Shell Production Plant data heat map

Objective of Feature Selection

- Obtain a "sparse" and "significant" representation of output Y in terms of features from X
- Improve Prediction accuracy - balancing bias-variance trade-off
- Prevent model over-fitting issues
- Lower space and time usage
- Achieve significant savings in model training time
- Model interpretation - smaller subset shows the "big picture" - Occam's razor (Principle of Parsimony)

Challenges in Feature Selection

- Huge number of features (d) (e.g Text data)
- Lots of data points (n) too - Big data settings !
- X is a "fat matrix" ($d \gg n$)
- Strong Correlation among features - reality !
- Correlation affects consistent subset estimation
- Missing values for some data points
- Noisy observations also affects selection

Our Contribution

A new feature selection algorithm -
Bootstrap-enhanced **P**rojected **G**radient **D**escent, **BoPGD**

Our Contribution

A new feature selection algorithm -
Bootstrap-enhanced **P**rojected **G**radient **D**escent, **BoPGD**

- Offers scalability with dimensionality
- Consistent in true support recovery even when there is strong correlation among predictors
- Groups "strongly" correlated features together
- Uses re-sampling techniques to eliminate irrelevant, noisy features
- Provides a better interpretable model
- Orders of magnitude faster than existing algorithms ☺!

Outline

1. Feature Selection under Multicollinearity
   - Introduction
   - Problem Formulation
   - Existing algorithms for feature selection
   - Our algorithm - BoPGD
   - Experiments and Results
2. Causal Inference on Time Series
   - Introduction to Causality
   - Preliminaries and Model Assumptions
   - Existing approaches of Granger Causality
   - Our Concurrent Estimation Method
   - Experiments and Results

High-dimensional Sparse Estimation

- Given data points $X = [x_1, x_2, \ldots, x_n]^T$, each $x_i \in \mathcal{R}^d$
- A response vector $Y = [y_1, y_2, \ldots, y_n]^T$, each $y_i \in \mathcal{R}$
- Goal : Compute an s^*-sparse coefficient vector θ^* s.t,

$$\theta^* = \arg \min_{\theta : \|\theta\|_0 \leq s^*} f(\theta)$$

- $\|.\|_0$ is the L_0 norm function - Non-Convex in θ
- $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle)$ is the empirical risk function

High-dimensional Sparse Estimation

- Given data points $X = [x_1, x_2, \ldots, x_n]^T$, each $x_i \in \mathcal{R}^d$
- A response vector $Y = [y_1, y_2, \ldots, y_n]^T$, each $y_i \in \mathcal{R}$
- Goal : Compute an s^*-sparse coefficient vector θ^* s.t,

$$\theta^* = \arg \min_{\theta : \|\theta\|_0 \leq s^*} f(\theta)$$

- $\|.\|_0$ is the L_0 norm function - Non-Convex in θ
- $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle)$ is the empirical risk function
- Optimal estimation is infeasible due to NP-hardness ☹!

Our Model : Sparse Linear Regression

- Simple, yet useful high-dimensional linear model

$$Y = X\bar{\theta} + \zeta$$

- Loss function $f(\theta) = \frac{1}{n}\|Y - X\theta\|_2^2$
- ζ : an n-dimensional label noise vector where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$

Our Model : Sparse Linear Regression

- Simple, yet useful high-dimensional linear model

$$Y = X\bar{\theta} + \zeta$$

- Loss function $f(\theta) = \frac{1}{n} \|Y - X\theta\|_2^2$
- ζ : an n-dimensional label noise vector where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$
- Goal : Jointly minimize the empirical risk and sparsity of estimation

Outline

Feature Selection methods

Broadly four main themes -

- Filter methods
- Classical methods a.k.a Wrapper methods
- Shrinkage methods a.k.a Embedded methods
- Iterative Hard-thresholding techniques a.k.a Project Gradient Descent methods [which can also be categorized as an embedded method]

Filter methods

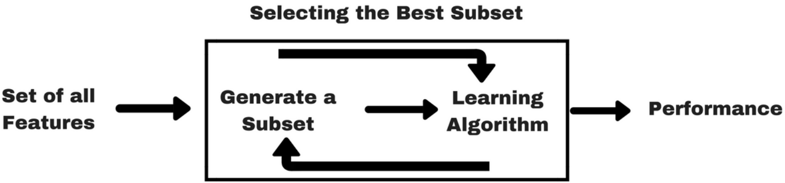
Set of all Features → **Selecting the Best Subset** → **Learning Algorithm** → **Performance**

- Feature selection done on the basis of scores in several statistical tests
- E.g.-
  - **Pearson's Correlation test** - quantifies linear dependence between two continuous-valued variables
  - **Chi-Square test** - evaluates the likelihood of correlation between two categorical variables
  - Others namely **LDA**, **ANOVA**, **Mutual Information** etc.

Drawbacks

- Determine the relevant feature subset without training a model on them
- Filter methods **do not** remove multicollinearity
- Selects all features correlated with the output
- Filter methods fail to find the best subset of features in many occasions

Classical methods

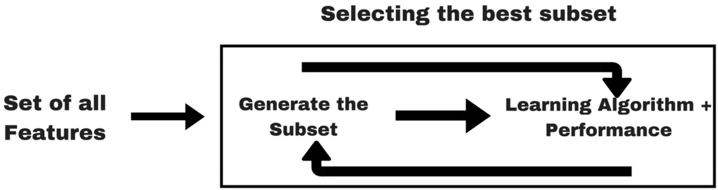


- **Best-Subset Selection** - Finds for each $k \in [d]$, the subset of size k having smallest residual sum of squares [FW00]
- **Forward or Backward Stepwise Selection** - Greedily adds (or removes) predictors one by one to the model
- **Forward Stage-wise Regression** - Iteratively identifies the variable most correlated with the current residual and adds it to the model

Drawbacks

- Infeasible when d is large ($d \approx 50$ usually works)
- "Slow-fitting" nature yields higher running time
- Sensitive to correlated variables
- Discrete process - variables either retained or discarded - often exhibits high variance
- Doesn't reduce the prediction error of the full model

Shrinkage methods



- Relax the non-convexity constraint of the L_0 norm by appropriate convex relaxations
- Promises to achieve global minima by optimizing over a larger constraint space which is convex
- Uses cross-validation to evaluate the performance of subset selection

Popular Shrinkage method based Algorithms

To name a few -

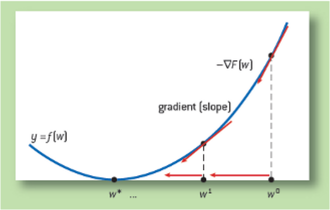
- Ridge [HK70] - L_2 norm relaxation of θ
- Lasso [Tib96] - L_1 norm relaxation of θ and its variants viz. Relaxed Lasso [Mei07], Adaptive Lasso [Zou06]
- Elastic Net [ZH05] - Combines L_1 and L_2 penalty
- OSCAR [BR08] - Combines L_1 and pairwise L_∞ norm penalization
- BoLasso [Bac08] - Bootstrap-enhanced Lasso support selection
- Cluster Representative Lasso (CRL) and Cluster Group Lasso [BRvdGZ13]

Drawbacks

- Not at all methods yield sparse solutions (e.g- ridge)
- Slower convergence rates - solves non-smooth optimization problems (e.g - Adaptive/Relaxed Lasso, OSCAR)
- Inconsistent estimation when X has high condition number - regular, sign, and pattern inconsistency (e.g - Lasso)
- Most of the methods do not take correlated structure of variables into account
- Even if some methods consider correlation (e.g CRL, CGL), they often overestimate $\hat{S}$ - lower precision and f1-score in true support recovery
- Higher running times when d is large

Off the Convex Path ...

Projected Gradient Descent (PGD) = Gradient Descent + Hard-thresholding



$+$

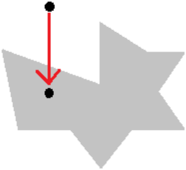


Figure: Empirical risk minimization with hard L_0 constraint

Iterative Hard-Thresholding techniques

- Iteratively do the following -
  - Gradient descent with gradient oracle
  - Hard threshold the gradient descent update onto the underlying non-convex set
- Projection can be performed efficiently for some interesting structures viz. sparsity, low rank
- Orders of magnitude faster than L_1 and greedy counterparts
- Achieves provable global guarantees when f is any arbitrary differentiable function satisfying RSC[1] and RSS[2] properties [JTK14]

[1] Restricted Strong Convexity
[2] Restricted Strong Smoothness

Outline

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

- Falls under the regime of Iterative Hard-thresholding

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

- Falls under the regime of Iterative Hard-thresholding
- Efficiently handles multicollinearity

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

- Falls under the regime of Iterative Hard-thresholding
- Efficiently handles multicollinearity
- Consistently estimates the true support - well empirically proven, No formal guarantees so far (work in progress)

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

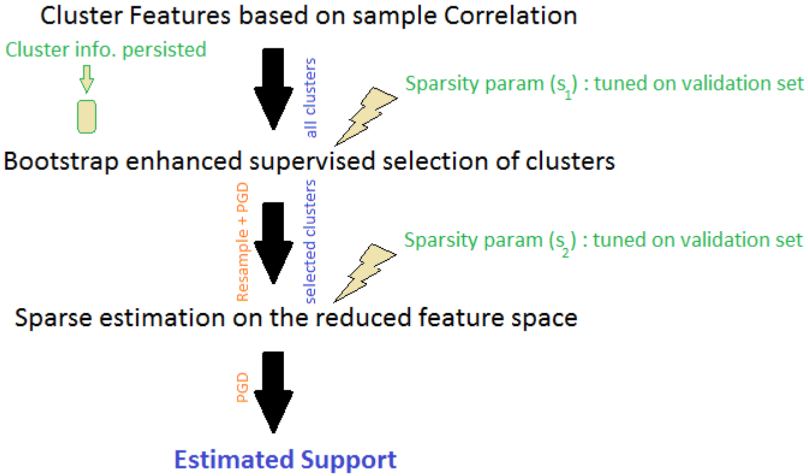
- Falls under the regime of Iterative Hard-thresholding
- Efficiently handles multicollinearity
- Consistently estimates the true support - well empirically proven, No formal guarantees so far (work in progress)
- Yields a "good" sparse model

Our algorithm - BoPGD

BoPGD - **Bo**otstrapped **P**rojected **G**radient **D**escent

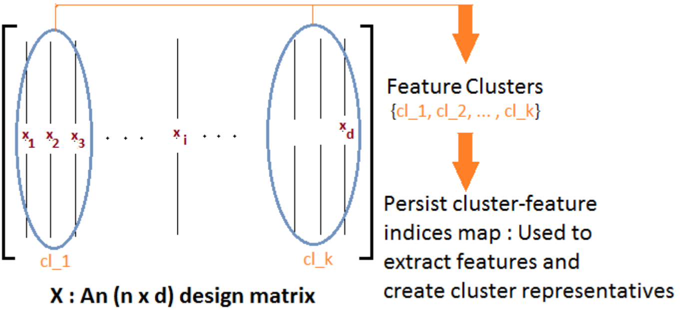
- Falls under the regime of Iterative Hard-thresholding
- Efficiently handles multicollinearity
- Consistently estimates the true support - well empirically proven, No formal guarantees so far (work in progress)
- Yields a "good" sparse model
- Scales well with the dimensionality of data

BoPGD - How it works



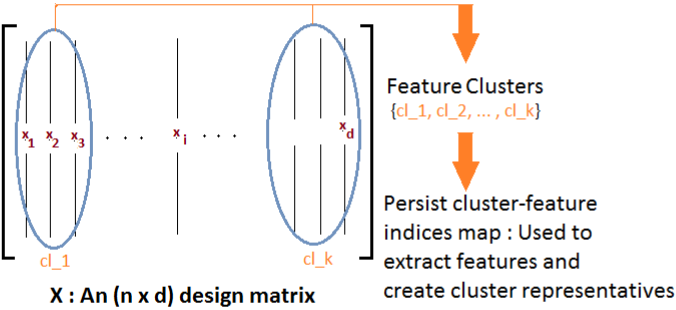
Clustering based on Inter-Feature Correlation

Objective : Take care of the strong correlation among the features - root cause of inconsistency in Lasso estimate !



Clustering based on Inter-Feature Correlation

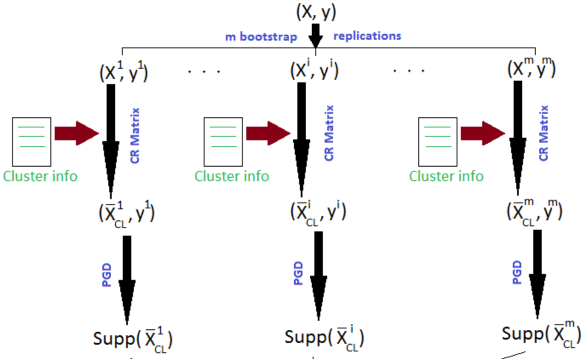
Objective : Take care of the strong correlation among the features - root cause of inconsistency in Lasso estimate !



Note : $n \approx \Omega(\log d)$ to correctly cluster covariates [BRvdGZ13]

Bootstrapped Supervised Cluster Selection

Objective : Extract the significant set of feature-clusters and eliminate nuisance clusters using bootstrap based re-sampling



$$S_{CL} = \bigcap_{i=1}^{m} \mathrm{Supp}(\bar{X}^i_{CL})$$

Sparse estimation on down-sampled features

Objective : Refine the over-estimated support using PGD - Lesser competition from noisy variables !

Sparse estimation on down-sampled features

Objective : Refine the over-estimated support using PGD - Lesser competition from noisy variables !

- $\hat{S}^u = \bigcup_{i \in S_{CL}} \textit{cls}(i)$

Sparse estimation on down-sampled features

Objective : Refine the over-estimated support using PGD - Lesser competition from noisy variables !

- $\hat{S}^u = \bigcup_{i \in S_{CL}} \mathit{cls}(i)$
- $\hat{\theta} = \textbf{PGD}(X_{\hat{S}^u},\, s_2)$

Sparse estimation on down-sampled features

Objective : Refine the over-estimated support using PGD - Lesser competition from noisy variables !

- $\hat{S}^u = \bigcup_{i \in S_{CL}} \mathit{cls}(i)$
- $\hat{\theta} = \mathbf{PGD}(X_{\hat{S}^u},\ s_2)$
- Sparsity parameter s_2 chosen via cross-validation - elbow point approach

Sparse estimation on down-sampled features

Objective : Refine the over-estimated support using PGD - Lesser competition from noisy variables !

- $\hat{S}^u = \bigcup_{i \in S_{CL}} \textit{cls}(i)$
- $\hat{\theta} = \textbf{PGD}(X_{\hat{S}^u}, s_2)$
- Sparsity parameter s_2 chosen via cross-validation - elbow point approach
- Report $\hat{S} = \{j : \hat{\theta}_j \neq 0\}$ as our "predicted" support

Outline

Evaluation Metrics

We used the following evaluation metrics :

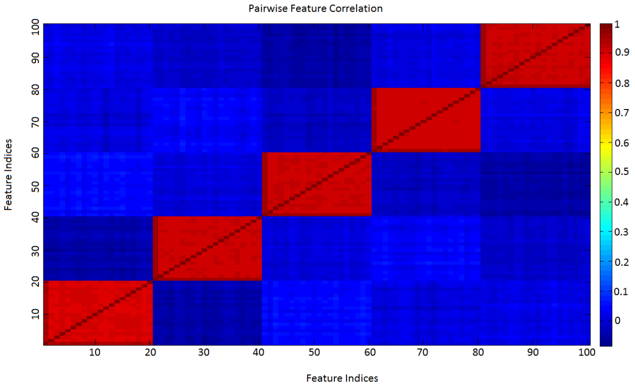
- Precision (P) $:= \frac{|\hat{S} \bigcap S^0|}{|\hat{S}|}$
- Recall (R) $:= \frac{|\hat{S} \bigcap S^0|}{|S^0|}$
- F-score $(F_1) := \frac{2PR}{P+R}$
- Mean Squared Error MSE $:= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(Y^i_{test} - \hat{Y}^i_{test})^2]$

where $\hat{S}$:= estimated support, S^0 := true support (if known)

Synthetic Experiments

Experiment 1. Single Block Model

- Data samples generated as $Z_i = (X_i, Y_i)$ where $X_i \sim \mathcal{N}(0, \Sigma)$ and $Y_i = \langle \theta, X_i \rangle + \zeta_i$ and $\zeta_i \sim \mathcal{N}(0, \sigma^2)$
- Σ block diagonal with high (≥ 0.85) intra-block and low (≤ 0.25) inter-block correlation



Comparison among different methods : Experiment 1

- The true support (S_0) is restricted to the first block
- $\mid S_0 \mid = 10$

Methods	**P**	**R**	**F$_1$**	**MSE**
Lasso	0.505	0.992	0.668	0.007
Elastic Net	0.464	0.996	0.631	0.006
BoLasso	0.756	0.996	0.859	0.005
CRL	0.500	1.000	0.667	0.005
PGD	0.913	0.988	0.947	0.006
BoPGD	**1.000**	0.996	**0.998**	0.006

Table: Results on Experiment 1 (averaged across 25 simulations)

Comparison among different methods : Experiment 2

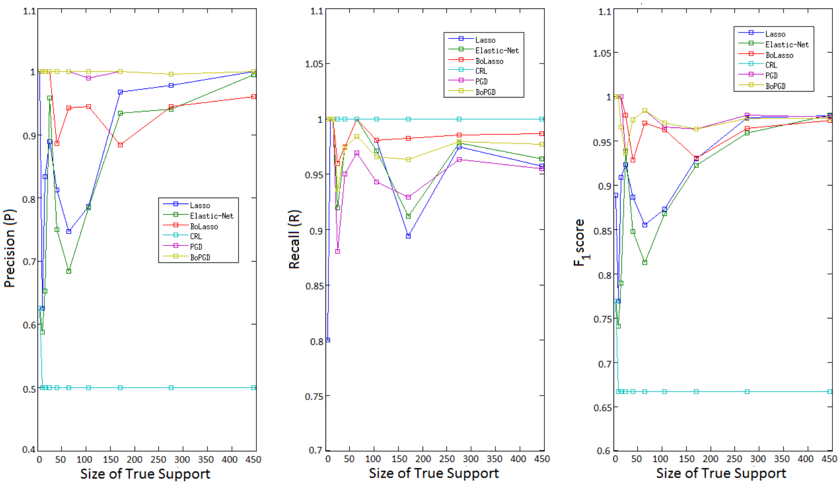
Experiment 2. Multi Block Model

- The true support is now spanned across all the five blocks
- From each block, half of the features are chosen randomly to be in support
- $\mid S_0 \mid = 50$

Methods	**P**	**R**	**F$_1$**	**MSE**
Lasso	0.710	0.941	0.809	0.013
Elastic Net	0.652	0.979	0.783	0.016
BoLasso	**0.934**	**0.921**	**0.927**	0.032
CRL	0.500	1.000	0.667	0.009
PGD	0.998	0.862	0.925	0.021
BoPGD	**0.978**	0.900	**0.938**	0.034

Table: Results on Experiment 2 (averaged across 25 simulations)

Performance of different algorithms as d varies



Experiments on Real data

Autompg data (UCI repository)

- The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multi-valued discrete and 5 continuous attributes
- There are 398 samples and 8 predictors viz. (2) cylinders, (3) displacement, (4) horsepower, (5) weight, (6) acceleration, (7) model year, (8) origin and (9) car name
- The target variable is (1) mpg (miles per gallon)
- Features $\{2, 3, 4, 5\}$ were strongly correlated ($\rho \geq 0.84$)
- Missing values (very less) were ignored

Results on Autompg data

Ftrs	Lasso	BoLasso	CRL	OSCAR	PGD	BoPGD
2	0	-0.716*	-0.648	-0.102	0	0
3	0	0.282*	1.890	-0.102	0.271	0.206
4	-0.002	0	-0.971	-0.102	0	0
5	-0.006	0.011	-5.285	-0.102*	-0.731	-0.661
6	0	0.082*	0	0	0	0
7	0.530	-0.007*	2.771	0.102*	0.353	0.349
8	0.387	0.160	1.174	0.061	0.163	0.166

- OSCAR does a “pretty decent” job of implicit clustering based on equality of coefficients

* marked coefficients are inconsistent (missed $\approx$ 5% of the cases)

Outline

Dependence vs Causation

"Stork deliver babies (p=0.008)" - [Mat00]

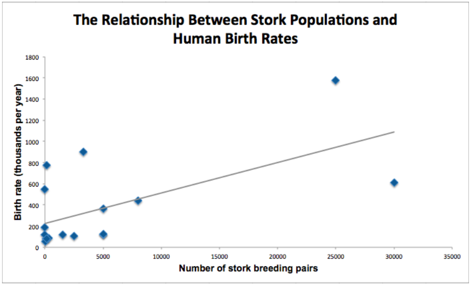


Figure: Human birth rate vs stork population across 17 European countries

Correlation $\not\Rightarrow$ Causation

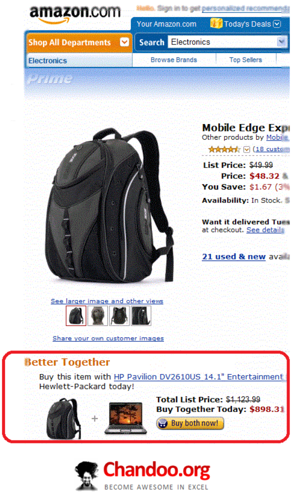


Figure: Snapshot from Amazon website (old)

Causal Inference on Time Series

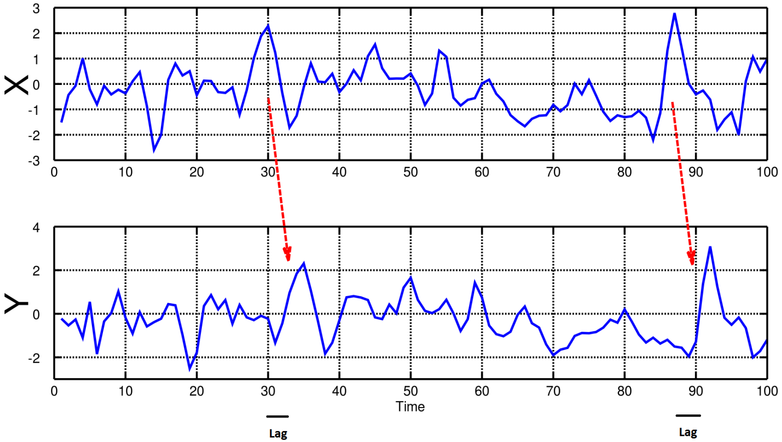


Figure: Time Series X influencing Y

From Data to Knowledge

Goal : Discover the underlying causal structure given as input large scale, high-dimensional time series data

From Data to Knowledge

Goal : Discover the underlying causal structure given as input large scale, high-dimensional time series data

- Climate Change data - output the temporal causal graph of climate enforcing agents (viz. CO, CO_2, CH_4 etc.)
- Gene regulatory Network discovery given gene expression time series data
- Social influence Analysis etc.

From Data to Knowledge

Goal : Discover the underlying causal structure given as input large scale, high-dimensional time series data

- Climate Change data - output the temporal causal graph of climate enforcing agents (viz. CO, CO_2, CH_4 etc.)
- Gene regulatory Network discovery given gene expression time series data
- Social influence Analysis etc.

For the rest of the talk we focus only on **time series** data.

Challenges

Challenges

- Output at a particular point in time in no longer a function of just the other variables at the same instant of time

Challenges

- Output at a particular point in time in no longer a function of just the other variables at the same instant of time
- Notion of time-lagged variables come into picture
- Effective number of features blows up with time-lagged variables

Challenges

- Output at a particular point in time in no longer a function of just the other variables at the same instant of time
- Notion of time-lagged variables come into picture
- Effective number of features blows up with time-lagged variables
- Much more computationally intensive process

Challenges

- Output at a particular point in time in no longer a function of just the other variables at the same instant of time
- Notion of time-lagged variables come into picture
- Effective number of features blows up with time-lagged variables
- Much more computationally intensive process
- Difficult when max lag L is unspecified or unknown

Our Contribution

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**
- Our algorithms estimate $\hat{L}$: a "best suited" value for maxlag

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**
- Our algorithms estimate $\hat{L}$: a "best suited" value for maxlag
- The selection is governed by Akaike Information Criterion (AIC) [Aka98]

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**
- Our algorithms estimate $\hat{L}$: a "best suited" value for maxlag
- The selection is governed by Akaike Information Criterion (AIC) [Aka98]
- Concurrently we are able to infer the "hypothesis" feature causal graph with "good" accuracy

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**
- Our algorithms estimate $\hat{L}$: a "best suited" value for maxlag
- The selection is governed by Akaike Information Criterion (AIC) [Aka98]
- Concurrently we are able to infer the "hypothesis" feature causal graph with "good" accuracy
- Both the algorithms scale well with dimensionality

Our Contribution

- We propose two algorithms - **Lasso Granger++** and **Group Lasso Granger++**
- Our algorithms estimate $\hat{L}$: a "best suited" value for maxlag
- The selection is governed by Akaike Information Criterion (AIC) [Aka98]
- Concurrently we are able to infer the "hypothesis" feature causal graph with "good" accuracy
- Both the algorithms scale well with dimensionality
- Data driven process - prior domain knowledge of the underlying physical data generating process not needed ☺!

Outline

Granger Causality [Gra69, Gra80]

- One of the most popular approaches to infer causal dependencies among time series variables
- Captures a statistical aspect of causality based on prediction
- X "granger causes" Y, denoted as $X \rightarrow Y$, if past values of X contain information that helps predict Y above and beyond that of contained in past values of Y alone
- Granger Causality principles -
  (a) The cause happens prior to the effect (Instantaneous causation is ignored !)
  (b) The cause makes unique changes in the effect

Linear Granger Causality tests - Bivariate data

- $X = \{x^t\}_{t=1}^T$, $Y = \{y^t\}_{t=1}^T$ time series variables (length T)
- $\mathbf{x}_t = [x^{(t-1)}, x^{(t-2)}, \dots, x^{(t-L)}]$, $\mathbf{y}_t = [y^{(t-1)}, y^{(t-2)}, \dots, y^{(t-L)}]$

Linear Granger Causality tests - Bivariate data

- $X = \{x^t\}_{t=1}^T$, $Y = \{y^t\}_{t=1}^T$ time series variables (length T)
- $\mathbf{x}_t = [x^{(t-1)}, x^{(t-2)}, \dots, x^{(t-L)}]$, $\mathbf{y}_t = [y^{(t-1)}, y^{(t-2)}, \dots, y^{(t-L)}]$
- Two different VAR models are fit to the Y as :

$$y^t \approx \langle \alpha, \mathbf{y}_t \rangle + \langle \beta, \mathbf{x}_t \rangle$$
$$y^t \approx \langle \gamma, \mathbf{y}_t \rangle$$

where $\alpha = [\alpha_1, \dots, \alpha_L]$, $\gamma = [\gamma_1, \dots, \gamma_L]$ and $\beta = [\beta_1, \dots, \beta_L]$

Linear Granger Causality tests - Bivariate data

- $X = \{x^t\}_{t=1}^T$, $Y = \{y^t\}_{t=1}^T$ time series variables (length T)
- $\mathbf{x}_t = [x^{(t-1)}, x^{(t-2)}, \dots, x^{(t-L)}]$, $\mathbf{y}_t = [y^{(t-1)}, y^{(t-2)}, \dots, y^{(t-L)}]$
- Two different VAR models are fit to the Y as :

$$y^t \approx \langle \alpha, \mathbf{y}_t \rangle + \langle \beta, \mathbf{x}_t \rangle$$
$$y^t \approx \langle \gamma, \mathbf{y}_t \rangle$$

  where $\alpha = [\alpha_1, \dots, \alpha_L]$, $\gamma = [\gamma_1, \dots, \gamma_L]$ and $\beta = [\beta_1, \dots, \beta_L]$
- Use any statistical significance tests (viz. F-statistic, χ^2 test) to ascertain whether the first model outperforms the second, and conclude X *"Granger causes"* Y

Linear Granger Causality tests - Multivariate data

- Given P time series $X_i = \{x_i^t\}_{t=1}^T, \forall i = \{1, 2, \ldots, P\}$

Linear Granger Causality tests - Multivariate data

- Given P time series $X_i = \{x_i^t\}_{t=1}^T$, $\forall i = \{1, 2, \ldots, P\}$
- A VAR model of order L is fit to X_i, $\forall t = L+1$ to T :

$$x_i^t = \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle + \epsilon_i^t$$

- $x_j^{(t,L)} = [x_j^{(t-1)}, \ldots, x_j^{(t-L)}]$ is the history of X_j up to time t
- $\beta_j^i = [\beta_j^i(1), \ldots, \beta_j^i(L)]$ is the coefficient vector
- ϵ_i^t is independent additive white noise
- No correlation among the residuals across time

Linear Granger Causality tests - Multivariate data

- Given P time series $X_i = \{x_i^t\}_{t=1}^T$, $\forall i = \{1, 2, \ldots, P\}$
- A VAR model of order L is fit to X_i, $\forall t = L+1$ to T :

$$x_i^t = \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle + \epsilon_i^t$$

- $x_j^{(t,L)} = [x_j^{(t-1)}, \ldots, x_j^{(t-L)}]$ is the history of X_j up to time t
- $\beta_j^i = [\beta_j^i(1), \ldots, \beta_j^i(L)]$ is the coefficient vector
- ϵ_i^t is independent additive white noise
- No correlation among the residuals across time
- X_j *"Granger causes"* X_i, if *at least one* value in β_j^i is *non-zero*

Outline

1. Feature Selection under Multicollinearity
   - Introduction
   - Problem Formulation
   - Existing algorithms for feature selection
   - Our algorithm - BoPGD
   - Experiments and Results
2. Causal Inference on Time Series
   - Introduction to Causality
   - Preliminaries and Model Assumptions
   - Existing approaches of Granger Causality
   - Our Concurrent Estimation Method
   - Experiments and Results

Existing methods

Existing methods

- Exhaustive Graphical Granger
    - $\mathcal{O}(P^2)$ Pairwise Linear Granger Causality tests

Existing methods

- Exhaustive Graphical Granger
  - $\mathcal{O}(P^2)$ Pairwise Linear Granger Causality tests
- Vector Auto Regressive Granger
  - Fit a VAR model of order L to all the P features

Existing methods

- Exhaustive Graphical Granger
  - $\mathcal{O}(P^2)$ Pairwise Linear Granger Causality tests
- Vector Auto Regressive Granger
  - Fit a VAR model of order L to all the P features
- Lasso Granger [ALA07]
  - Lasso type formulation identifies sparse neighborhood structure

Existing methods

- Exhaustive Graphical Granger
  - $\mathcal{O}(P^2)$ Pairwise Linear Granger Causality tests
- Vector Auto Regressive Granger
  - Fit a VAR model of order L to all the P features
- Lasso Granger [ALA07]
  - Lasso type formulation identifies sparse neighborhood structure
- Group Lasso Granger [LALR09]
  - Uses Group Lasso to leverage the group structure among the variables according to the time series they belong to

Lasso Granger Method [ALA07]

Lasso Granger Method [ALA07]

- Applies LASSO [Tib96] type formulation to the VAR model for each x_i, $i = \{1, 2, ..., P\}$
- Obtain a sparse estimate of the coefficient vector - corresponds to its neighborhood

Lasso Granger Method [ALA07]

- Applies LASSO [Tib96] type formulation to the VAR model for each x_i, $i = \{1, 2, ..., P\}$
- Obtain a sparse estimate of the coefficient vector - corresponds to its neighborhood
- Optimization problem of Lasso Granger :

$$\min_{\beta} \sum_{t=L+1}^{T} \left(x_i^t - \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle \right)^2 + \lambda \|\beta\|_1$$

- λ - regularization parameter ensures sparsity in β

Lasso Granger Method [ALA07]

- Applies LASSO [Tib96] type formulation to the VAR model for each x_i, $i = \{1, 2, ..., P\}$
- Obtain a sparse estimate of the coefficient vector - corresponds to its neighborhood
- Optimization problem of Lasso Granger :

$$\min_{\beta} \sum_{t=L+1}^{T} \left(x_i^t - \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle \right)^2 + \lambda \left\| \beta \right\|_1$$

- λ - regularization parameter ensures sparsity in β
- x_j *"Granger causes"* x_i, if *at least one* entry in β_j^i is *non-zero*

Group Lasso Granger Method [LALR09]

Group Lasso Granger Method [LALR09]

- Considers the natural *group structure* existing among the variables imposed by the respective time series they belong to
- Applies Group Lasso ([YL06]) type formulation

Group Lasso Granger Method [LALR09]

- Considers the natural *group structure* existing among the variables imposed by the respective time series they belong to
- Applies Group Lasso ([YL06]) type formulation
- Consider a partitioning of the set of predictors $\{x_1, \ldots, x_P\}$ into J groups
- Optimization problem of Group Lasso Granger :

$$\min_{\beta} \sum_{t=L+1}^{T} \left(x_i^t - \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle\right)^2 + \lambda \sum_{j=1}^{J} \sqrt{\rho_j}\, \left\|\beta_{\mathbb{G}_j}\right\|_2$$

- ρ_j accounts for the varying group size

Merits and Demerits

Merits and Demerits

Limitations :

- All the methods assume L is known beforehand
- Small values of L are considered in these models
- Pairwise tests instead of the full model yields spurious causality (e.g - exhaustive granger)
- Common Lag choice for all the features (e.g VAR granger) - unrealistic assumption !
- Do not scale under the high-dimensional setting (except Lasso Granger and Group Lasso Granger)

Merits and Demerits

Limitations :

- All the methods assume L is known beforehand
- Small values of L are considered in these models
- Pairwise tests instead of the full model yields spurious causality (e.g - exhaustive granger)
- Common Lag choice for all the features (e.g VAR granger) - unrealistic assumption !
- Do not scale under the high-dimensional setting (except Lasso Granger and Group Lasso Granger)

Advantages :

- Lasso and Group Lasso are consistent in sparse neighborhood selection [MB06]

Outline

Our Method

Our Method

- We propose a concurrent estimation technique
  - Outputs the hypothesis feature causal graph
  - Estimate a value of maxlag L_i for every feature X_i

Our Method

- We propose a concurrent estimation technique
  - Outputs the hypothesis feature causal graph
  - Estimate a value of maxlag L_i for every feature X_i
- For e.g.- Consider the VAR equations :

$$x(t) = a_1 * x(t-1) + a_2 * y(t-2) + \epsilon_1(t)$$
$$y(t) = b_1 * y(t-10) + \epsilon_2(t)$$

Our Method

- We propose a concurrent estimation technique
  - Outputs the hypothesis feature causal graph
  - Estimate a value of maxlag L_i for every feature X_i
- For e.g.- Consider the VAR equations :

$$x(t) = a_1 * x(t-1) + a_2 * y(t-2) + \epsilon_1(t)$$
$$y(t) = b_1 * y(t-10) + \epsilon_2(t)$$

- We present two algorithms in this context :
  - **Lasso Granger++ :** Adapting Lasso Granger method to unknown lag
  - **Group Lasso Granger++ :** Extension of Group Lasso Granger method with unknown lag

Our method with initial lag L_0

Our method with initial lag L_0

- We assume an upper bound M on the maxlag L
- Start with a small initial guess $L_0 = \ell$
- Run Lasso Granger for the target from $t = \ell + 1$ to T

Our method with initial lag L_0

- We assume an upper bound M on the maxlag L
- Start with a small initial guess $L_0 = \ell$
- Run Lasso Granger for the target from $t = \ell + 1$ to T

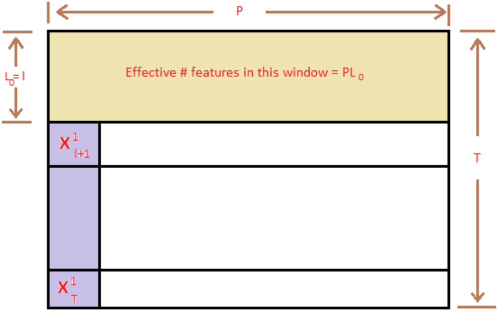


Figure: Lasso Granger++ with initial guess L_0

Our method with higher lags

Our method with higher lags

- We increment our maxlag estimate by ℓ i.e. $L_1 = 2\ell$
- Do not regress on all $2P\ell$ variables but a subset of them

Our method with higher lags

- We increment our maxlag estimate by ℓ i.e. $L_1 = 2\ell$
- Do not regress on all $2P\ell$ variables but a subset of them

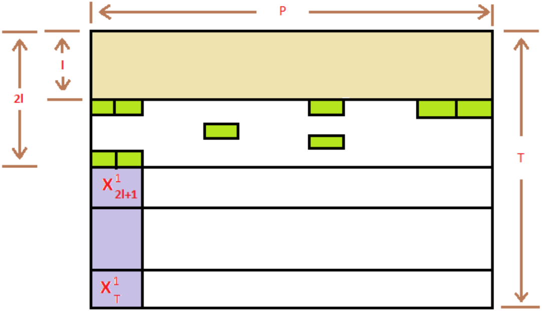


Figure: Lasso Granger++ with next guess L_1

Key points

- Choice of λ plays in key role in both the Lasso and Group Lasso routines - chosen via AIC(c)
- For each L_i we choose the "best" λ using AIC(c) [Aka98] score
- We select "best" maxlag $\hat{L}_i$ for X_i using minimum AIC(c) score

Space Complexity Analysis

Space complexity (κ) is :

$$\kappa = \mathcal{O}\Bigg(\sum_{j=1}^{\frac{T-2}{\ell}} (P\ell + s_{j-1})\Bigg), \quad where \quad 0 \le s_{j-1} \ll P(j-1)\ell$$

$$= \mathcal{O}\Bigg(T \max_j \big(P, \frac{s}{\ell}\big)\Bigg), \quad where \quad s = \max_j s_{j-1}$$

which is significantly smaller than $\frac{PT^2}{\ell}$ (brute force approach)

Time Complexity Analysis

Time complexity τ is :

$$\tau = \mathcal{O}\left(\sum_{j=1}^{\frac{T-2}{\ell}} (T - j\ell)(P\ell + s_{j-1})^2 \right), \quad where \quad 0 \leq s_{j-1} \ll P(j-1)\ell$$

$$= \mathcal{O}\left(T^2 \max_j \left(P^2\ell, \frac{s^2}{\ell}, Ps \right) \right), \quad where \quad s = \max_j s_{j-1}$$

which is significantly smaller than $\frac{P^2T^4}{\ell}$ (brute force approach)

Outline

Evaluation Criteria

- Let A denote the adjacency matrix of the source (original) feature causal graph
- Let $\hat{A}$ denote the same for the hypothesis (output) graph
- Precision (P) := $\frac{|\{(i,j) \in V \times V : \hat{A}(i,j) = A(i,j)\}|}{|\{(i,j) \in V \times V : \hat{A}(i,j) = 1\}|}$
- Recall (R) := $\frac{|\{(i,j) \in V \times V : \hat{A}(i,j) = A(i,j)\}|}{|\{(i,j) \in V \times V : A(i,j) = 1\}|}$
- F_1 score := $\frac{2PR}{P+R}$

Experiment 1

- We used one of the most cited benchmark datasets [Set10]
- 5 time series variables with small lags
- Ground truth is :

Variables	Causal Subset	Indiv. Lags	MaxLag
X_1	$\{X_1\}$	$\{1, 2\}$	2
X_2	$\{X_1\}$	$\{2\}$	2
X_3	$\{X_1\}$	$\{3\}$	3
X_4	$\{X_1, X_4, X_5\}$	$\{2, 1, 1\}$	2
X_5	$\{X_4, X_5\}$	$\{1, 1\}$	1

Results of Experiment 1

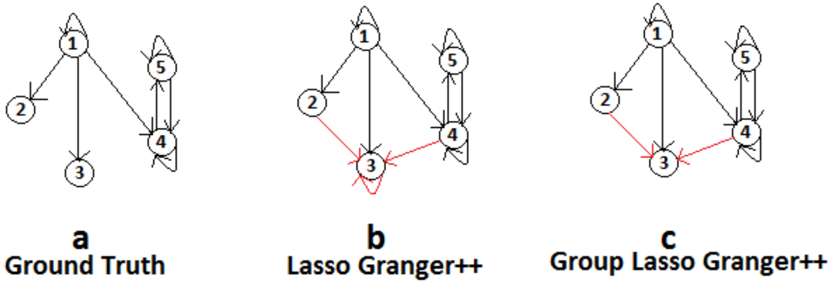


Figure: Ground Truth and the Hypothesis Causal Graphs inferred

Results of Experiment 1 - Contd ...

- Lasso Granger++ Lag prediction accuracy = 100%
- Group Lasso Granger++ Lag prediction accuracy = 100%

Methods	P	R	F_1
Lasso Granger++	0.727	1.000	0.842
Group Lasso Granger++	0.800	1.000	0.888
Standard Lasso Granger (L fixed)	0.701	1.000	0.821
Standard Group Lasso Granger (L fixed)	0.803	1.000	0.889

Table: Results on Experiment 1 (averaged across 10 simulations)

Experiment 2

- A star graph with 5 time series variables
- X_1 is the only target variable causally influenced by all other variables
- Variables $\{X_2, \ldots, X_5\}$ are independent noise processes
- Substantially different time lags along each edge from $X_j \to X_1$, $j \in \{2, \ldots, 5\}$
- Lag values chosen from $[1, 50]$ u.a.r (e.g. $L_2 = 46, L_3 = 7, L_4 = 46, L_5 = 32$)

Lasso Granger++ performance

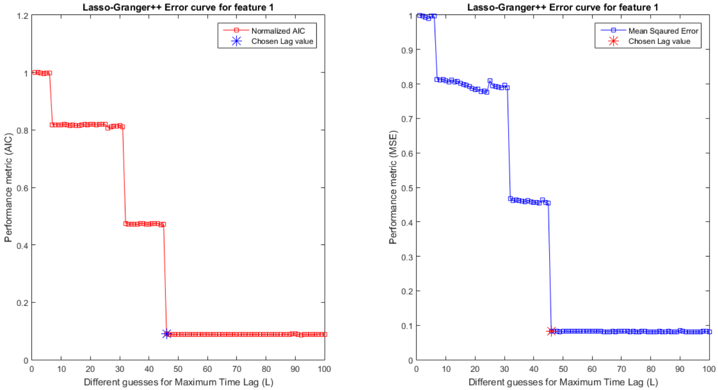
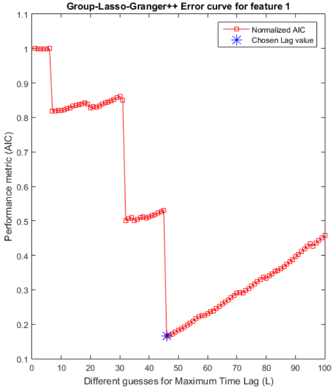


Figure: Lasso Granger++ step-wise error curve

Group Lasso Granger++ performance



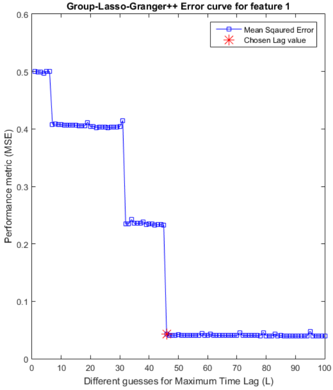
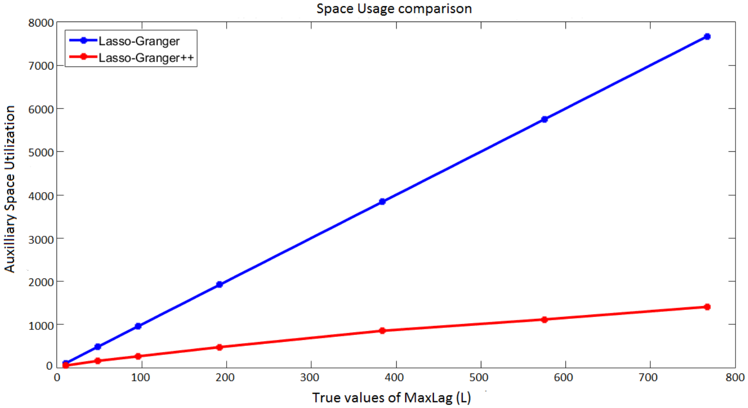


Figure: Group Lasso Granger++ step-wise error curve

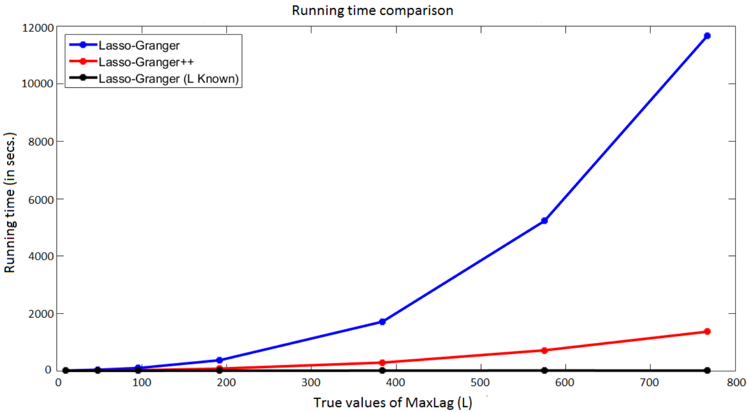
Performance tests of our Method I

1. Space Usage as a function of L



Performance tests of our Method II

2. Running time as a function of L



Publications based on the Thesis

Machine Learning and Statistical Analysis for Materials Science : Stability Analysis, Fingerprint Descriptors and Chemical Insights
- Published in The Journal of Chemistry of Materials, 2017
Authors : Praveen Pankajakshan, Suchismita Sanyal, Onno E. de Noord, Indranil Bhattacharya, Arnab Bhattacharyya, Umesh Waghmare.

Acknowledgment

Many thanks to Dr. Praveen Pankajakshan (Shell Technology Center, Bangalore) for his useful suggestions and valuable contributions.

Thank You ☺!

References I

- Hirotogu Akaike.
  Information theory and an extension of the maximum likelihood principle.
  In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.
- Andrew Arnold, Yan Liu, and Naoki Abe.
  Temporal causal modeling with graphical granger methods.
  In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM, 2007.
- Frank Arntzenius.
  Reichenbach's common cause principle.
  1999.

References II

Francis R Bach.
Bolasso: model consistent lasso estimation through the bootstrap.
In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

Howard D Bondell and Brian J Reich.
Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar.
Biometrics, 64(1):115–123, 2008.

References III

Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang.
Correlated variables in regression: clustering and sparse estimation.
Journal of Statistical Planning and Inference, 143(11):1835–1858, 2013.

George M Furnival and Robert W Wilson.
Regressions by leaps and bounds.
Technometrics, 42(1):69–79, 2000.

Clive WJ Granger.
Investigating causal relations by econometric models and cross-spectral methods.
Econometrica: Journal of the Econometric Society, pages 424–438, 1969.

References IV

Clive WJ Granger.
Testing for causality: a personal viewpoint.
Journal of Economic Dynamics and control, 2:329–352, 1980.

Arthur E Hoerl and Robert W Kennard.
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67, 1970.

Prateek Jain, Ambuj Tewari, and Purushottam Kar.
On iterative hard thresholding methods for high-dimensional m-estimation.
In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

References V

Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset.
Grouped graphical granger modeling for gene expression regulatory networks discovery.
Bioinformatics, 25(12):i110–i118, 2009.

Robert Matthews.
Storks deliver babies (p= 0.008).
Teaching Statistics, 22(2):36–38, 2000.

Nicolai Meinshausen and Peter Bühlmann.
High-dimensional graphs and variable selection with the lasso.
The annals of statistics, pages 1436–1462, 2006.

Nicolai Meinshausen.
Relaxed lasso.
Computational Statistics & Data Analysis, 52(1):374–393, 2007.

References VI

Anil K Seth.
A matlab toolbox for granger causal connectivity analysis.
Journal of neuroscience methods, 186(2):262–273, 2010.

Robert Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Ming Yuan and Yi Lin.
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

References VII

Hui Zou and Trevor Hastie.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

Hui Zou.
The adaptive lasso and its oracle properties.
Journal of the American statistical association, 101(476):1418–1429, 2006.