Learning from Aggregated Data

Joydeep Ghosh

Schlumberger Centennial Chair Professor,

Dept of Electrical and Computer Engineering

The University of Texas at Austin

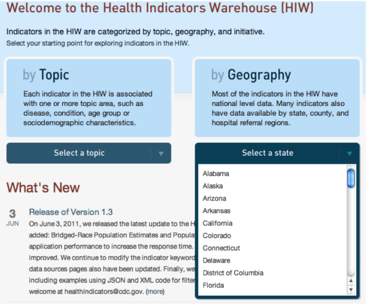
Co-authors

Avradeep Bhowmik
UT Austin

Oluwasanmi Koyejo
UIUC

Yubin Park
Accordion Health

Motivation: Privacy Aware Data Release

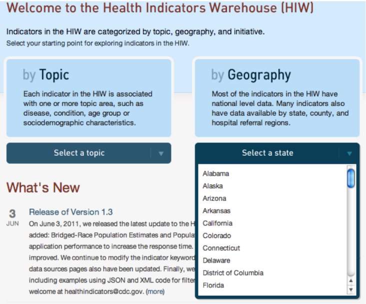


- Health Indicators Warehouse
  - http://healthindicators.gov
- CDC data and statistics
  - http://www.cdc.gov/DataStatistics
- Statehealthfacts.org – Kaiser Family Foundation
  - http://www.statehealthfacts.org

And many more ...

- Additional Drivers: Scale, Bandwidth, Robustness
  - Sensor Networks, IoT, etc.

Motivation: Privacy Aware Data Release



- Health Indicators Warehouse
  - http://healthindicators.gov
- CDC data and statistics
  - http://www.cdc.gov/DataStatistics
- Statehealthfacts.org – Kaiser Family Foundation
  - http://www.statehealthfacts.org

And many more ...

- Additional Drivers: Scale, Bandwidth, Robustness
  - Sensor Networks, IoT, etc.
- Can such variably aggregated, multi-source data be leveraged to improve predictive models at the individual level?
  - person $\longrightarrow$ hospital $\longrightarrow$ county $\longrightarrow$ HRR $\longrightarrow$ state
  - different time scales

This Talk: Learning from Aggregated Data

- Sensitive Variables are only available as group averages (KDD'14)
- Dependent variables are given only as histograms (AISTATS'15)
- All variables are group-wise aggregated (ICML'16)
- Some/All variables averaged over (different) time intervals (AISTATS'17)

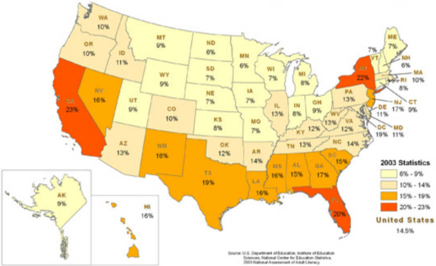
Ecological Fallacy

- Naive use of aggregated data leads to Ecological Fallacy
- Group-level attributes differ from individual level ground truth
  - Literacy rate vs. Proportion of Immigration[1]
  - State-level correlation: -0.53
  - Individual-level correlation: 0.12

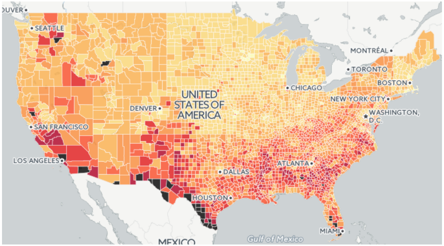
[1] Robinson, W. S., "Ecological correlations and the behavior of individuals", American Sociological Review (1950)

Ecological Fallacy

Real population data tends to be heterogeneous



State-Level Data



County-Level Data

Figure: Percentage of population lacking basic literacy (2003)

Images courtesy: missourieconomy.org, daniel carmody

Aggregated Sensitive Variables

Gender	Age	Diabetes	State
F	23	Neg.	TX
F	53	Neg.	CA
M	46	? (suppressed)	FL
F	63	? (suppressed)	FL
M	63	Neg.	CA
M	63	? (suppressed)	FL
	⋮		

State	Diabetes Rate
CA	8.5 %
FL	8.0 %
IL	6.3 %
NY	6.2 %
PA	9.2 %
TX	8.7 %
⋮	

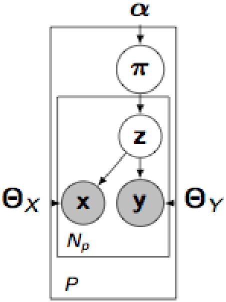
Y. Park, J. Ghosh, LUDIA: an aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data, KDD 2014

Key Considerations

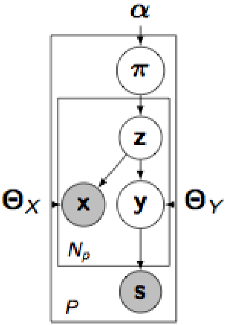
- Main Ideas:
  - Heterogeneous partitions with relatively homogeneous sub-populations
  - Aggregated statistics reflect different proportion of sub-populations
- Terminology
  - Individual level data vs. Aggregated data
- Objectives
  - Individual level inference using only Aggregated data
  - Avoid data reconstruction

Clustering Using features with DIfferent levels of Aggregation (CUDIA)

- **y** : sensitive feature (not observed)
- **s** : aggregated value of **y** over a partition
- P : a number of partitions
- N_p : a number of samples in partition p



(a) Complete Individual-level Data



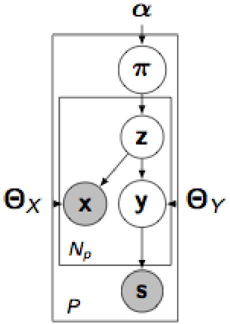
(b) Aggregated Sensitive Data

CUDIA - Probabilistic Formulation

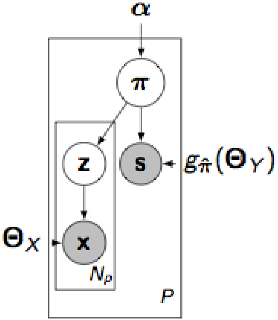
"Central Limit Theorem" for a mixture distribution:

$$\mathbf{s} = \text{Average}[\mathbf{y}] \sim \text{Normal}(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi^2), \quad \text{where} \quad \boldsymbol{\mu}_\pi = \sum_{k=1}^{K} \pi_k \boldsymbol{\theta}_{yk}$$

where $\boldsymbol{\theta}_k$ represents the mean of $p(\mathbf{y} \mid z_k = 1)$.[1]

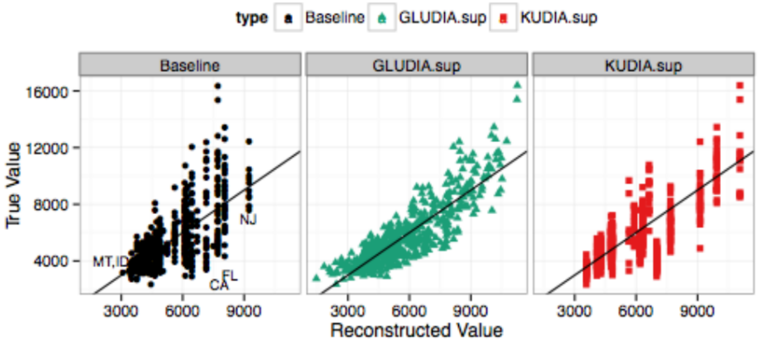


simplifies to



[1] In [Park and Ghosh, 2012], $\boldsymbol{\theta}_{yk}$ is a parameter set that contains the mean and variance.

County-level Medicare payments
Dartmouth Health Atlas



Using 2 aggregated and 1 county-level variable

KLUDIA Summary

- Aggregated and suppressed data can be effectively utilized in individual-inferential tasks
  - reconstruction, prediction, hypothesis testing,
- Can use multiple sources with different-levels of aggregation

Generalised Linear Modelling with Histogram Aggregated Data

- Individual Level Features
- Histogram-Aggregated Targets

- A Bhowmik, J Ghosh, O Koyejo, "Generalized Linear Models for Aggregated Data", Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, (AISTATS) 2015

Motivation

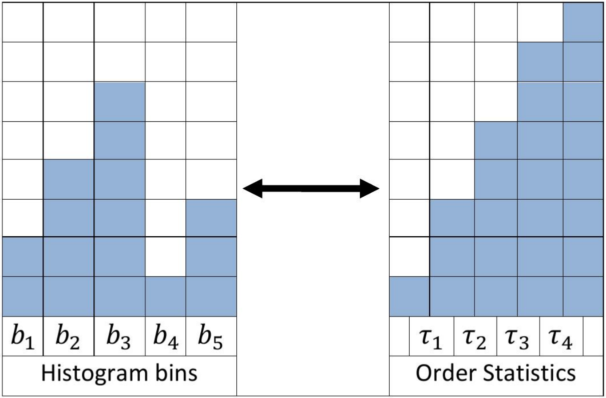
---

courtesy: Joydeep Ghosh and Yubin Park

Motivation

- Healthcare, sociological, etc. data can be of two kinds
  - Non-sensitive attributes (age, sex, etc.)
  - Privacy-sensitive information (income, health indicators, etc.)
- Sensitive data often summarised as mean, median, quartile, etc., or with histograms
- Focus of this work: order statistics and histograms

Order Statistics and Histograms



- We use histogram to mean a set of order statistics and vice versa

Problem Setup

- Standard GLM : Data obtained as (covariate, target) pairs $(\boldsymbol{x}_i, z_i)$

Problem Setup

- Standard GLM : Data obtained as (covariate, target) pairs $(\boldsymbol{x}_i, z_i)$
- This setting: z only known upto order statistics, correspondence between covariates and targets unknown

Problem Setup

- Standard GLM : Data obtained as (covariate, target) pairs $(\boldsymbol{x}_i, z_i)$
- This setting: z only known upto order statistics, correspondence between covariates and targets unknown
- Task is two-fold
  - Impute the targets $\boldsymbol{z}$ subject to order statistic constraints
  - Estimate parameter $\boldsymbol{\beta}$ using the imputed targets

Optimisation Problem

$$
\begin{aligned}
\min_{\boldsymbol{z},\boldsymbol{\beta}} \quad & D_{\boldsymbol{\phi}}\left(\boldsymbol{z} \| g_{\phi}(\boldsymbol{X}\beta)\right) \\
\text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_{\tau}
\end{aligned}
$$

- Use alternating minimisation to solve for β and $\boldsymbol{z}$ separately

Algorithm: Step I

$$\min_{\boldsymbol{\beta}} D_{\boldsymbol{\phi}}\left(\boldsymbol{z_{t-1}} \| g_\phi(\boldsymbol{X}\beta)\right)$$

- This is a standard GLM parameter estimation problem
- Can use any off-the-shelf package to solve for β

Algorithm: Step II

$$
\begin{aligned}
\min_{\boldsymbol{z}} \quad & D_\phi\left(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\beta)\right) \\
\text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_\tau
\end{aligned}
$$

- Constraint set looks non-convex, seems difficult to succinctly represent mathematically

Algorithm: Step II

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & D_{\phi}\left(\boldsymbol{z} \| g_{\phi}(\boldsymbol{X}\beta)\right) \\ \text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_{\tau} \end{aligned}$$

- Constraint set looks non-convex, seems difficult to succinctly represent mathematically
- But for GLMs, solution can be obtained in closed form!

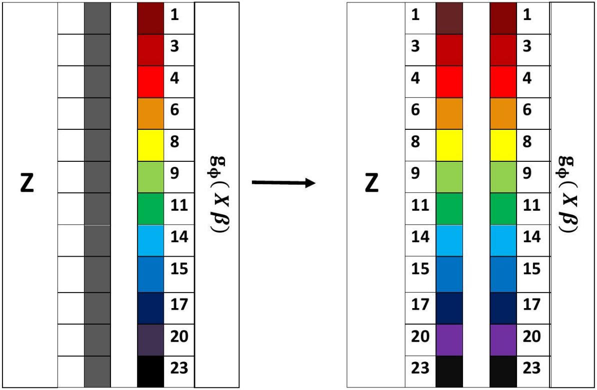
Algorithm: Step II

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & D_{\phi}\left(\boldsymbol{z} \| g_{\phi}(\boldsymbol{X}\beta)\right) \\ \text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_{\tau} \end{aligned}$$

- Constraint set looks non-convex, seems difficult to succinctly represent mathematically
- But for GLMs, solution can be obtained in closed form!
- Proof relies on fact that optimal z is isotonic with $g_{\phi}(\boldsymbol{X}\beta)$

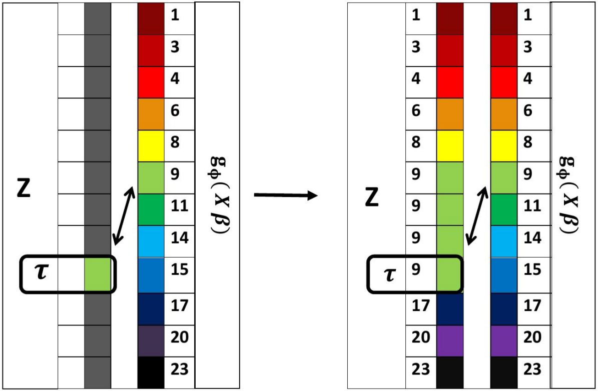
Without order statistic constraint : Direct Substitution

$$
\begin{aligned}
\boldsymbol{z}_t \; = \min_{\boldsymbol{z}} \quad & D_\phi(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\boldsymbol{\beta}_t)) \\
\text{s.t.} \quad & \boldsymbol{z} \in \mathbb{R}^n_{\downarrow}
\end{aligned}
$$



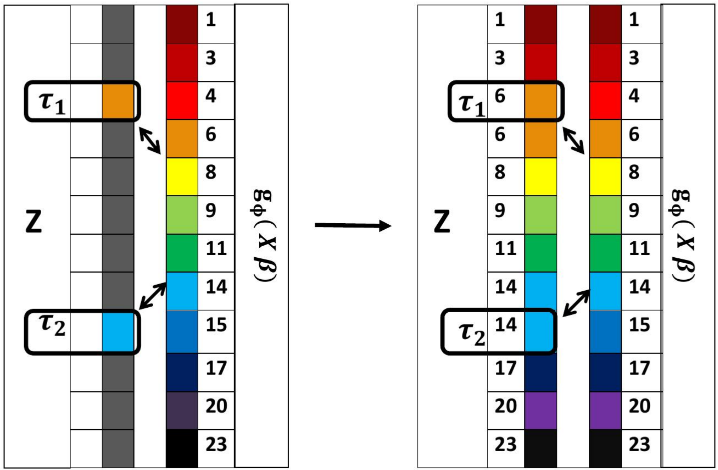
With order statistic constraint: Elementwise Thresholding

$$\begin{aligned} \boldsymbol{z}_t &= \min_{\boldsymbol{z}} \quad D_\phi(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\beta_t)) \\ \text{s.t.} & \quad \boldsymbol{z} \in \mathbb{R}^n_\downarrow, \ \boxed{z_\tau = s_\tau} \end{aligned}$$



Histogrammed Data : Multiple Order Statistics Constraints

Imputation of target variables - with histogram constraints



Summary of the Algorithm

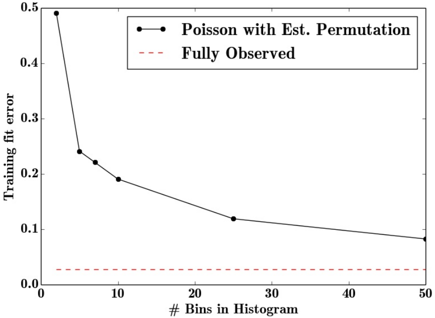
Two simple steps, repeated alternatingly till convergence:

- Estimation of GLM parameter β
  - Can use standard, off-the-shelf packages
- Imputation[2] of target variables z
  - Elementwise thresholding operation

[2] up to a re-permutation step

Experiments : TX Inpatient Discharge Dataset

- Fitting hospital charges on predictor variables age, race, sex, length of stay, etc.



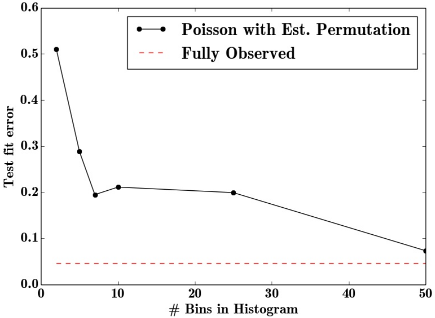


Figure: Accuracy on Training and Test set for TxID Dataset

Summary

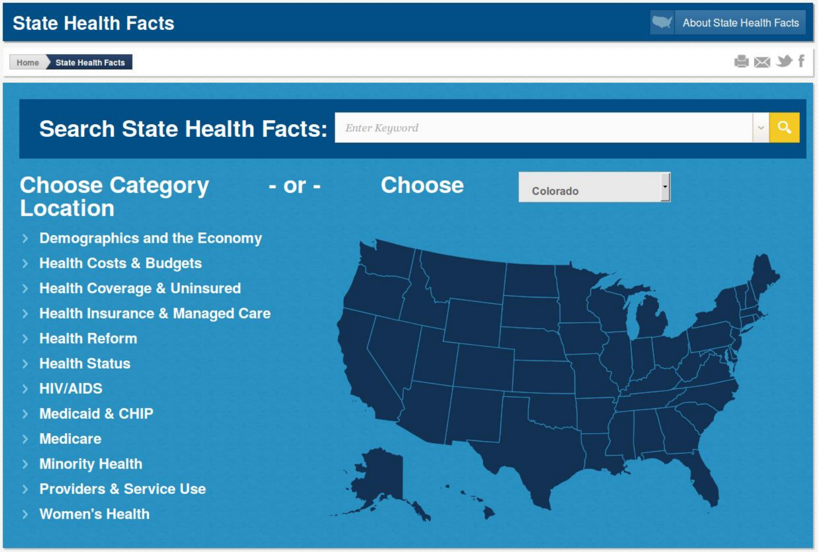
- Learning individual level regressors when targets are provided as histogram aggregates
- Simple, efficient algorithm, two alternating steps-
  - fit GLM model parameter
  - impute targets subject to constraints
- Estimation is reasonably effective given granular histograms
- Potential impact on aggregation as a privacy preserving technique

Sparse Linear Models for Group-wise Aggregated Data

- Features and Targets are Both Aggregated
- Recovery of Sparse Model parameter

- A Bhowmik, J Ghosh, O Koyejo, "Sparse Parameter Recovery from Aggregated Data", Proceedings of the 33rd International Conference on Machine Learning (ICML) 2016

Motivation


State Health Facts
About State Health Facts
Home
State Health Facts
Search State Health Facts:
Enter Keyword
Choose Category - or - Choose Location
Colorado
Demographics and the Economy
Health Costs & Budgets
Health Coverage & Uninsured
Health Insurance & Managed Care
Health Reform
Health Status
HIV/AIDS
Medicaid & CHIP
Medicare
Minority Health
Providers & Service Use
Women's Health

Setting

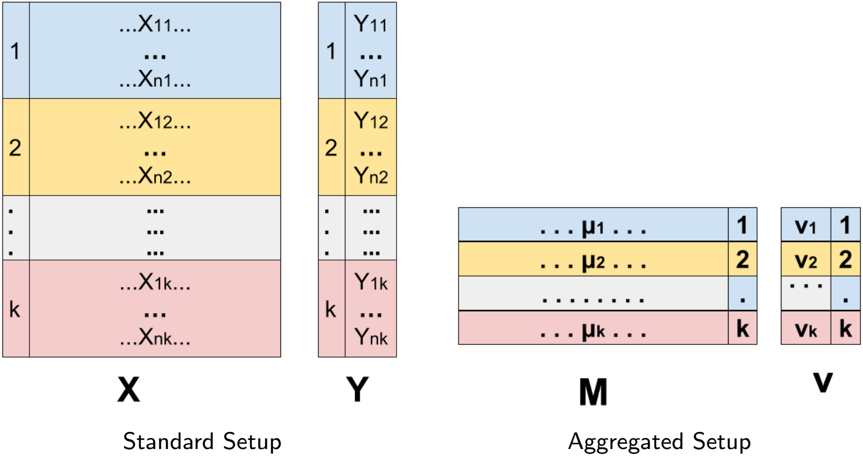


Figure: Learning Linear Models : Aggregated vs Non-Aggregated Setup

Summary of Main Results

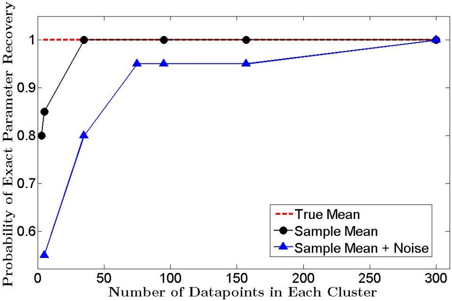
Assuming empirical estimates have been computed from sufficiently large number of samples, w.h.p.-

- Noise free aggregated data: exact parameter recovery
- Data with observation noise: recovery within arbitrarily small tolerance
- Histogram aggregation: approximate recovery upto tolerance specified by histogram granularity

Aggregated Data with Observation Noise

Theorem

Sparse parameter β^* can be recovered from noise aggregates within arbitrarily small tolerance with probability at least $1 - e^{-C_0 n} - e^{-C_1 n}$



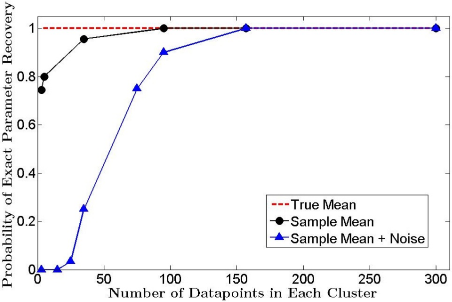


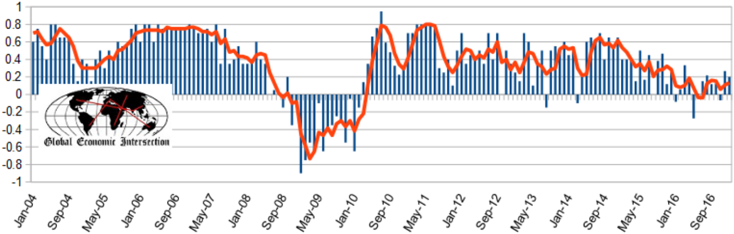
Figure: Probability of Exact Parameter Recovery for Gaussian and Bernoulli Models

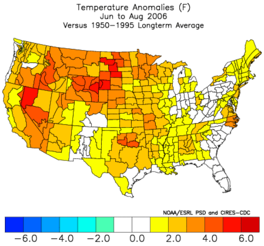
Frequency Domain Predictive Modelling with Aggregated Data

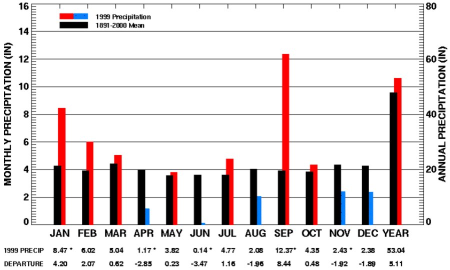
- Spatio-Temporally Correlated Data
- Non-uniform Aggregation

- A Bhowmik, J Ghosh, O Koyejo, "Frequency Domain Predictive Modelling with Aggregated Data", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017

Motivation







Images courtesy: Econintersect (BEA), NOAA, Blue Hill Observatory

Motivation

- Spatio-temporal data often released in aggregated form in practice (Burrell et al., 2004; Lozano et al., 2009; Davidson et al., 1978)
- Worse, sampling periods need not be aligned, aggregation periods need not be uniform[3]
  - ratio of government debt to GDP reported yearly
  - GDP growth rate reported quarterly
  - unemployment rate and inflation rate reported monthly
  - interest rate, stock market indices and currency exchange rates reported daily
- Challenges in mathematical representation
- Reconstruction is expensive and unreliable

[3] Bureau of Labor Statistics, Bureau of Economic Analysis

Idea: Transform Problem to Frequency Domain!

Idea: Transform Problem to Frequency Domain!

- bypasses local non-alignment by matching global properties

Idea: Transform Problem to Frequency Domain!

- bypasses local non-alignment by matching global properties
- avoids input data reconstruction

Idea: Transform Problem to Frequency Domain!

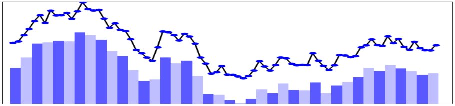
- bypasses local non-alignment by matching global properties
- avoids input data reconstruction
- achieves provably bounded generalization error

Problem Setup

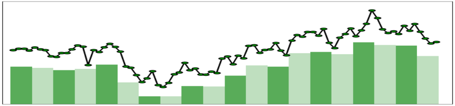
Features $\boldsymbol{x}(t) = [x_1(t), x_2(t) \cdots x_d(t)]$, targets $y(t)$

- Weak Stationarity+
  - Zero-mean, Finite variance
  - Autocorrelation function independent of time
- Performance measure
  - expected squared residual error $\mathcal{L}(\boldsymbol{\beta}) = E[|\boldsymbol{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$
  - Because of weak stationarity, $\mathcal{L}(\boldsymbol{\beta})$ is independent of t

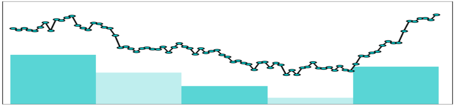
Data Aggregation in Time Series



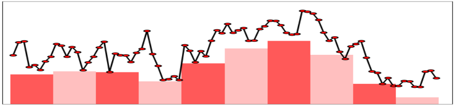
Non-Aggregated Feature $\boldsymbol{X}_1$
Aggregated Feature $\overline{X}_1$



Non-Aggregated Feature $\boldsymbol{X}_2$
Aggregated Feature $\overline{X}_2$



Non-Aggregated Feature $\boldsymbol{X}_3$
Aggregated Feature $\overline{X}_3$



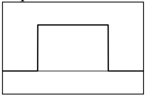
Non-Aggregated Target $\boldsymbol{Y}$
Aggregated Target $\overline{Y}$

Aggregation in Time and Frequency Domain

In time domain, convolution with square wave + sampling

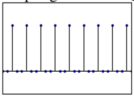
$z(t) \xrightarrow{\textit{convolution}}$

Square function u_T



$\xrightarrow{\textit{sampling}}$

Sampling Function δ_T

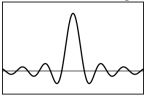


$\longrightarrow \bar{z}[k]$

In frequency domain, multiplication with sinc function + sampling

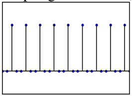
$Z(\omega) \xrightarrow{\textit{multiplication}}$

Sinc function U_T



$\xrightarrow{\textit{sampling}}$

Sampling Function δ_ω



$\longrightarrow \bar{Z}(\omega)$

Fourier Duality

Finite signal length, use T_0-restricted Fourier transforms -

$$Z_{T_0}(\omega) = \int_{-T_0}^{T_0} z(t) e^{-\iota \omega t} dt$$

Global Properties $\iff$ Local Properties

Bypass local mis-alignment by matching global properties

Algorithm: Step I

1. Input parameters T_0, ω_0, D, aggregated data samples $\overline{x}[k], \boldsymbol{y}[l]$

Algorithm: Step I

1. Input parameters T_0, ω_0, D, aggregated data samples $\overline{x}[k], \boldsymbol{y}[l]$

2. Sample D frequencies uniformly between $(-\omega_0, \omega_0)$

$$\Omega = \{\omega_1, \omega_2, \cdots \omega_D : \omega_i \in (-\omega_0 . \omega_0)\}$$

Algorithm: Step I

1. Input parameters T_0, ω_0, D, aggregated data samples $\overline{x}[k], \boldsymbol{y}[l]$

2. Sample D frequencies uniformly between $(-\omega_0, \omega_0)$

$$\Omega = \{\omega_1, \omega_2, \cdots \omega_D : \omega_i \in (-\omega_0 . \omega_0)\}$$

3. For each $\omega \in \Omega$, compute T_0-restricted Fourier Transforms $\overline{X}_{T_0}(\omega), \overline{Y}_{T_0}(\omega)$ from aggregated signals $\overline{x}[k], \overline{y}[l]$

Algorithm: Step II

Recall: U_T is Fourier transform of square wave

4. Estimate non-aggregated Fourier transforms

$$\widehat{X}_{i,T_0}(\omega) = \frac{\widehat{\boldsymbol{X}}_{i,T_0}(\omega)}{U_{T_i}(\omega)}, \;\; \widehat{Y}_{T_0}(\omega) = \frac{\overline{Y}_{T_0}(\omega)}{U_T(\omega)}$$

Algorithm: Step II

Recall: U_T is Fourier transform of square wave

4. Estimate non-aggregated Fourier transforms

$$\widehat{X}_{i,T_0}(\omega) = \frac{\widehat{\boldsymbol{X}}_{i,T_0}(\omega)}{U_{T_i}(\omega)}, \quad \widehat{Y}_{T_0}(\omega) = \frac{\overline{Y}_{T_0}(\omega)}{U_T(\omega)}$$

5. Estimate parameter $\widehat{\boldsymbol{\beta}}$ as:

$$\boxed{\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} E \| \widehat{\boldsymbol{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega) \|^2}$$

Main result I : Generalization Error

Generalisation error -
$\mathcal{L}^*$: best possible, $\hat{\mathcal{L}}$: frequency domain estimation

Theorem 1

For every small $\xi > 0$, $\exists$ corresponding T_0, D such that

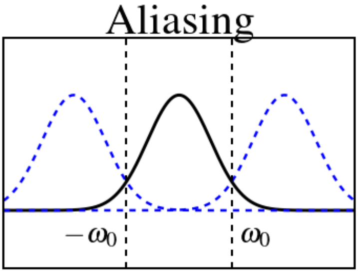
$$\hat{\mathcal{L}} < (1+\xi)\mathcal{L}^* + 2\xi$$

with probability at least $1 - e^{-O(D^2\xi^2)}$

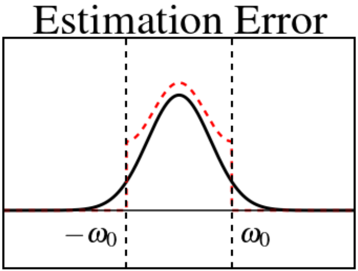
Thus, generalization error is bounded with sufficiently long signal (large T_0) and sufficient number of computed Fourier coefficients (large D)

Aliasing Effects, Non-uniform Sampling

- Signals not bandlimited $\Rightarrow$ Aliasing

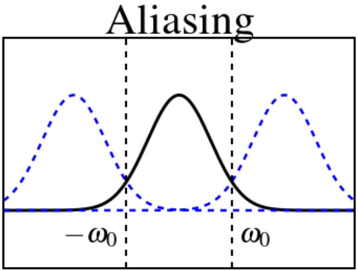


$\Longrightarrow$

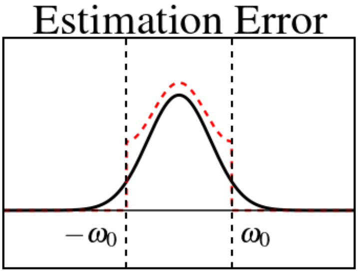


Aliasing Effects, Non-uniform Sampling

- Signals not bandlimited $\Rightarrow$ Aliasing



$\Longrightarrow$



- Assume autocorrelation function decays rapidly, e.g. a Schwartz function (Terzioğlu, 1969)
  - Then, most of the signal power concentrated between $(-\omega_0, \omega_0)$

Non-uniform aggregation, Finite samples

Generalisation error -
$\mathcal{L}^*$: best possible, $\hat{\mathcal{L}}$: frequency domain estimation

Theorem 2

With high probability, for any small $\xi > 0$, there exists T_0, D such that

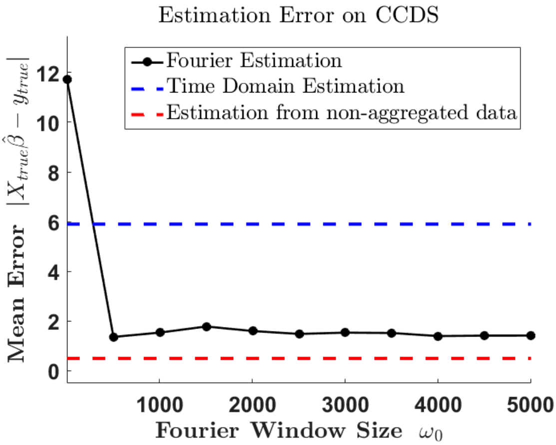
$$\hat{\mathcal{L}} < (1 + \xi)\mathcal{L}^* + 4\xi + \gamma$$

given sufficient samples, where $\gamma \sim e^{-O(\omega_s^2)}$.

Stronger performance guarantees for-

- sufficiently long signal
- large number of Fourier coefficients
- high sampling frequency

Comprehensive Climate Dataset (CCDS)



Regressing atmospheric vapour levels over continental United States vs readings of carbon dioxide levels, methane, cloud cover, and other extra-meteorological measurements

Additional Details

- More detailed analysis (not shown) allows for more precise error control
- Algorithm and analysis easily extend to sliding window aggregation schemata, and multi-dimensional settings e.g. spatio-temporal data using the multi-dimensional Fourier transform
- Extends to cases where aggregation and sampling period are non-overlapping.

Summary

- Spatio-temporal data is often aggregated, leading to significant challenges in learning and inference
- By converting the problem to frequency domain, we can bypass many of these problems using Fourier analysis
- Novel framework and estimation algorithm, provably bounded generalisation error
- Significant improvements vs reconstruction-based estimation.

Conclusion

- Data in many modern applications often released in aggregated form
- Summary of recent results
  - Learning generalised linear models with histogram-aggregated targets
  - Sparse learning for linear models with group-wise aggregated data
  - Frequency domain methods for aggregated spatio-temporal data
- Future work
  - Designing aggregation protocols for learning with privacy
  - Partially aggregated data and concept drift
  - Aggregation in matrix completion and recommendation systems

References

References I

Jenna Burrell, Tim Brooke, and Richard Beckwith. Vineyard computing: Sensor networks in agricultural production. *IEEE Pervasive computing*, 3(1): 38–45, 2004.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

James EH Davidson, David F Hendry, Frank Srba, and Stephen Yeo. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the united kingdom. *The Economic Journal*, pages 661–692, 1978.

David L Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

Simon Foucart. A note on guaranteed sparse recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009.

T Terzioğlu. On schwartz spaces. *Mathematische Annalen*, 182(3):236–242, 1969.

Additional Slides

KUDIA & LUDIA - Optimisation Formulation

$$\mathbf{D} = \mathbf{Z}\boldsymbol{\Theta} + \boldsymbol{\Xi} \qquad \text{(clustering)}$$

$$\mathbf{D} = \mathbf{U}\mathbf{V}^{\top} + \boldsymbol{\Xi} \qquad \text{(low-rank)}$$

$\mathbf{D}$: Complete data matrix $[\mathbf{X}\ \mathbf{Y}]$

$\mathbf{X}$: Non-sensitive feature matrix (observed)

$\mathbf{Y}$: Sensitive feature matrix (hidden)

$\mathbf{Z}$: Cluster assign matrix

$\boldsymbol{\Theta}$: Cluster parameter matrix

$\mathbf{U}$: Low-rank matrix for rows

$\mathbf{V}$: Low-rank matrix for columns

KUDIA & LUDIA - Optimisation Formulation

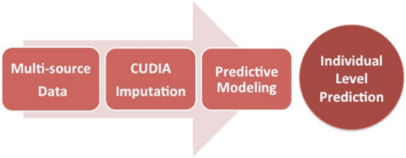
- KUDIA:
$$\min_{\mathbf{Y},\mathbf{Z},\boldsymbol{\Theta}} \| \begin{bmatrix} \mathbf{X} & \mathbf{Y} \end{bmatrix} - \mathbf{Z} \begin{bmatrix} \boldsymbol{\Theta}_X & \boldsymbol{\Theta}_Y \end{bmatrix} \|_2^2$$
$$\text{subject to} \quad \mathbf{AY} = \mathbf{S}$$

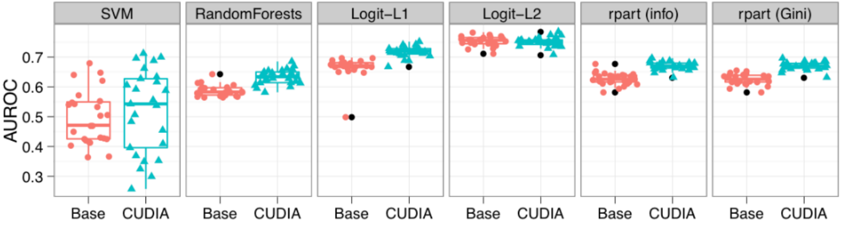
- LUDIA:
$$\min_{\mathbf{Y},\mathbf{U},\mathbf{V}} \| \begin{bmatrix} \mathbf{X} & \mathbf{Y} \end{bmatrix} - \mathbf{U} \begin{bmatrix} \mathbf{V}_X^\top & \mathbf{V}_Y^\top \end{bmatrix} \|_2^2$$
$$\text{subject to} \quad \mathbf{AY} = \mathbf{S}$$

$\mathbf{A}$: Aggregation matrix

$\mathbf{S}$: Aggregated sensitive feature matrix $\mathbf{S} = \mathbf{AY}$

Experiments: Individual-level Prediction





Predictive Performance: BRFSS (individual-level) + KFF (aggregated

$$(\text{Diabetes})_{\text{BRFSS}} \approx$$
$$(\text{Age})_{\text{BRFSS}} + (\text{BMI})_{\text{BRFSS}} + (\text{Avg. Fruit Consumption})_{\text{KFF}} + (\text{Avg. Heart Disease Rate})_{\text{KFF}}$$

- CUDIA uses cross-level imputed KFF features
- baseline model uses the average statistics as the individual-level estimates

Generalised Linear Modelling with Histogram Aggregated Data

- Individual Level Features
- Histogram-Aggregated Targets

- A Bhowmik, J Ghosh, O Koyejo, "Generalized Linear Models for Aggregated Data", Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, (AISTATS) 2015

Loss Function: Bregman Divergences

- Bregman Divergences are matching loss functions for GLMs

$$D_\phi(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\beta)) = \sum_i D_\phi(z_i \| g_\phi(\boldsymbol{x}_i^\top \beta))$$

- Generalisation of square loss, other examples are KL divergence, I-Divergence, etc.

- Convex in first argument, one-one correspondence with GLMs via ϕ

$$\boxed{\text{GLM link } g_\phi^{-1}(\cdot)} \quad \overset{\phi}{\Longleftrightarrow} \quad \boxed{D_\phi(\cdot \| \cdot) \text{ Bregman Divergence}}$$

Generalised Linear Models

- Covariates $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}]^\top$, Targets $z = [z_1, \cdots z_n]$, Parameter β
  - Mean of target variables related to a monotonically transformed linear function of covariates, that is, $E[z] = g_\phi(\boldsymbol{X}\beta)$
- Estimating the GLM parameter β is equivalent to solving

$$\min_{\boldsymbol{\beta}} D_{\boldsymbol{\phi}} \left(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\beta) \right)$$

  - $D_\phi(\cdot\|\cdot)$ is a Bregman Divergence- the matching loss functions for GLMs
  - E.g., square loss for Gaussian model (standard linear regression), I-divergence for Poisson model, etc.

Reformulation of Constraint Set

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & D_{\phi}\left(\boldsymbol{z} \| g_{\phi}(\boldsymbol{X}\beta)\right) \\ \text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_{\tau} \end{aligned}$$

Proposition : Say $\hat{\boldsymbol{z}}$ is the solution of the above problem. Then, the optimal $\hat{\boldsymbol{z}}$ is isotonic to $g_{\phi}(\boldsymbol{X}\beta)$

$$\hat{\boldsymbol{z}} \sim_{\downarrow} g_{\phi}(\boldsymbol{X}\beta)$$

Proof relies on the properties of identically separable Bregman Divergences and uses the fact that re-ordering a vector does not change its order statistics.

Reformulation of Constraint Set

$$
\begin{aligned}
\min_{\boldsymbol{z}} \quad & D_{\phi}\left(\boldsymbol{z} \| g_{\phi}(\boldsymbol{X}\beta)\right) \\
\text{s.t.} \quad & \tau^{th} \text{ order statistic of } \boldsymbol{z} = s_{\tau} \\
& \hat{\boldsymbol{z}} \sim_{\downarrow} g_{\phi}(\boldsymbol{X}\beta)
\end{aligned}
$$

- WLOG assume $\boldsymbol{z} \sim_{\downarrow} g_{\phi}(\boldsymbol{X}\beta) \in \mathbb{R}^n_{\downarrow}$ is ordered in descending order
- The, τ^{th} order statistic of $\boldsymbol{z}$ is simply $\boldsymbol{z}_{\tau}$

Reformulated Optimisation Problem

Putting it all together, we write the overall task as the following optimisation problem

$$
\begin{aligned}
\boldsymbol{z}_t \; = \min_{\boldsymbol{z}} \quad & D_\phi(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\boldsymbol{\beta}_t)) \\
\text{s.t.} \quad & \boldsymbol{z} \in \mathbb{R}^n_\downarrow, \\
& z_\tau = s_\tau
\end{aligned}
$$

- Solution can be obtained in closed form!

Histogram Constraints

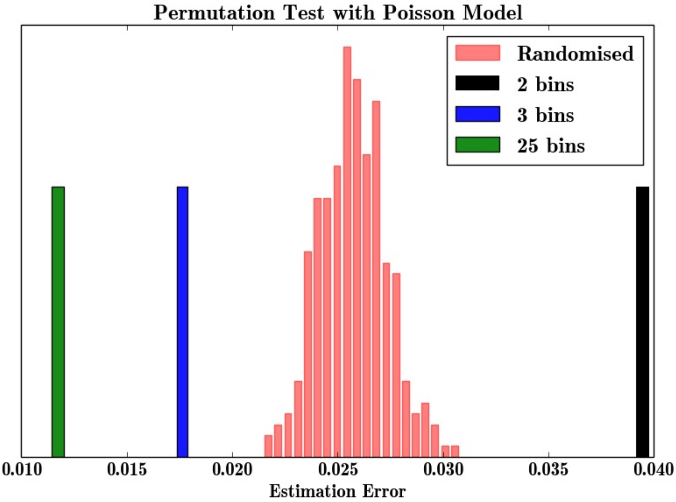
$$
\begin{aligned}
\boldsymbol{z}_t \; &= \min_{\boldsymbol{z}} \quad D_\phi(\boldsymbol{z} \| g_\phi(\boldsymbol{X}\boldsymbol{\beta}_t)) \\
&\quad \text{s.t.} \quad \boldsymbol{z} \in \mathbb{R}^n_\downarrow, \; z_{\tau_k} = s_{\tau_k}, \quad k = 1, 2, \cdots, h
\end{aligned}
$$

- Solution to this can be obtained in closed form: For all $1 < k < h$, and all $j \in \{1, 2, \cdots n\}$,

$$
\hat{z}^{(j)} = \begin{cases}
\min(g_\phi(\boldsymbol{X}\beta)^{(j)}, s_{\tau_1}) & j < \tau_1 \\
s_{\tau_k} & j = \tau_k \\
\min\left(s_{\tau_{k+1}}, \max(g_\phi(\boldsymbol{X}\beta)^{(j)}, s_{\tau_k})\right) & \tau_k \leq j \leq \tau_{k+1} \\
\max(g_\phi(\boldsymbol{X}\beta)^{(j)}, s_{\tau_h}) & j > \tau_h
\end{cases}
$$

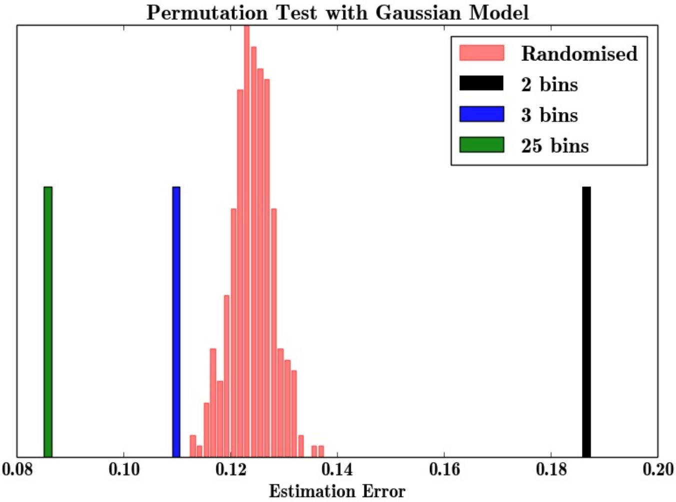
Experiments : Testing with Permutation

Results compared with estimation done on randomly permuted targets



Experiments : Testing with Permutation

Results compared with estimation done on randomly permuted targets



Experiments : Simulated Datasets

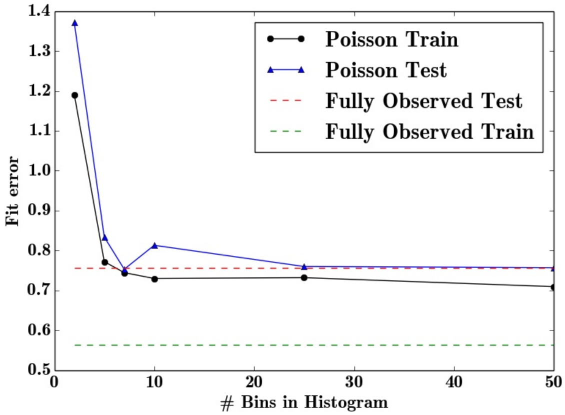


Figure: Accuracy on Training and Test Dataset with Poisson Model

Experiments : Simulated Datasets

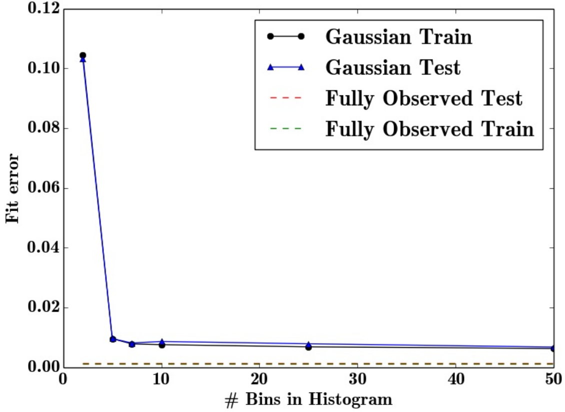
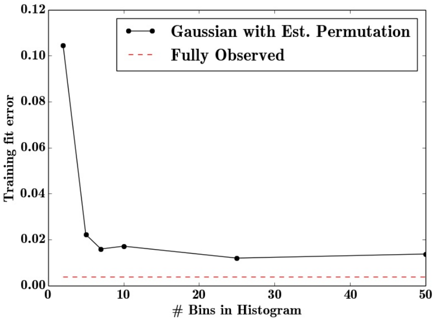


Figure: Accuracy on Training and Test Dataset with Gaussian Model

Experiments : DESynPUF Dataset

- Fitting primary payer reimbursement on predictor variables including age, race, rex, duration of coverage, presence/absence of certain diseases, etc.



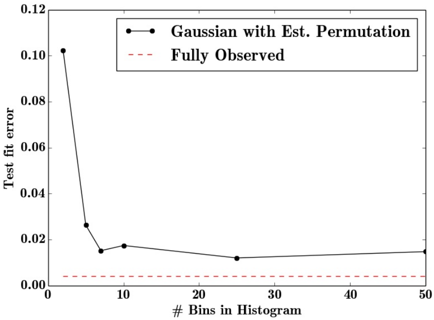


Figure: Accuracy on Training and Test set for DESynPUF Dataset

Sparse Linear Models for Group-wise Aggregated Data

- Features and Targets are Both Aggregated
- Recovery of Sparse Model parameter

- A Bhowmik, J Ghosh, O Koyejo, "Sparse Parameter Recovery from Aggregated Data", Proceedings of the 33rd International Conference on Machine Learning (ICML) 2016

Restricted Isometry Property

- $\boldsymbol{M}$ satisfies (s, δ_s)-RIP if for any s-sparse $\boldsymbol{z}$

$$(1-\delta_s)\|\boldsymbol{z}\|_2^2 \leq \|\boldsymbol{M}\boldsymbol{z}\|_2^2 \leq (1+\delta_s)\|\boldsymbol{z}\|_2^2$$

- Every small submatrix behaves approximately like an orthonormal system
- Satisfied by many random matrices including Gaussian ensembles, etc.

Parameter Estimation from True Means

- If $\boldsymbol{M}$ satisfies (s, δ_s)-RIP, given $(\boldsymbol{M}, \boldsymbol{\upsilon})$ a sparse β^* can be estimated[4] from an under-determined system

$$\boldsymbol{\upsilon} = \boldsymbol{M}\beta^* \ : \ \boldsymbol{M} \in \mathbb{R}^{k \times d}, \boldsymbol{\upsilon} \in \mathbb{R}^k, \boldsymbol{\beta}^* \in \mathbb{R}^d$$

- Assumption: True mean matrix $\boldsymbol{M}$ satisfies RIP, $\boldsymbol{\beta}^*$ is sparse
  - If exact $(\boldsymbol{M}, \boldsymbol{\upsilon})$ were known, we would be done.
- Real life: true $(\boldsymbol{M}, \boldsymbol{\upsilon})$ unknown, only empirical estimates available

[4] (Candes and Tao, 2006; Donoho, 2006; Candes, 2008; Foucart, 2010)

Noise-free Aggregated Data

Theorem 1

Given empirical aggregates , the true β^* can be recovered exactly with probability at least $1 - e^{-C_0 n}$.

Noise-free Aggregated Data

Theorem 1

Given empirical aggregates , the true β^* can be recovered exactly with probability at least $1 - e^{-C_0 n}$.

Here, the constant C_0 is such that

$$C_0 \sim O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1 + \delta_{2s_0})}\right)$$

- $\delta_{2s_0} < \Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$ be $2s_0$-restricted RIP constant for $\boldsymbol{M}$
- Each covariate has a sub-Gaussian distribution with parameter σ^2
- True β^* is κ_0-sparse, $\kappa_0 < s_0$

Noise-free Aggregated Data

Theorem 1

Given empirical aggregates , the true β^* can be recovered exactly with probability at least $1 - e^{-C_0 n}$.

Here, the constant C_0 is such that

$$C_0 \sim O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$$

Fewer samples required for-

- lower value of δ_{2s_0}, the RIP constant for true means $\boldsymbol{M}$
- smaller sub-Gaussian parameters for covariates σ^2

Aggregated Data with Observation Noise

- Each sample measurement corrupted by zero mean additive noise as

$$y = \boldsymbol{x}^\top \boldsymbol{\beta}^* + \epsilon$$

- Means $(\widehat{\boldsymbol{M}}_n, \widehat{Y}_\epsilon)$ computed from n noisy obs. for each group

$$\begin{aligned} \widehat{\boldsymbol{M}}_n = & \quad \boldsymbol{M} + \boldsymbol{\zeta}_{x,n} \\ \widehat{Y}_n = & \quad \boldsymbol{\upsilon} + \boldsymbol{\zeta}_{y,n} + \boldsymbol{\epsilon}_n \end{aligned} \tag{1}$$

- Two sources of error: aggregation and observation noise

Aggregated Data with Observation Noise

Theorem 2

Let $\xi > 0$ be any small positive real value. Given noisy empirical aggregates , true β^* can be recovered within an ℓ_2 distance of $O(\xi)$ with probability at least $1 - e^{-C_1 n} - \boldsymbol{e^{-C_2 n}}$.

Aggregated Data with Observation Noise

Theorem 2

Let $\xi > 0$ be any small positive real value. Given noisy empirical aggregates , true $\boldsymbol{\beta}^*$ can be recovered within an ℓ_2 distance of $O(\xi)$ with probability at least $1 - e^{-C_1 n} - \boldsymbol{e^{-C_2 n}}$.

Here, the constants C_1, C_2 are such that

$$C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right), \quad \boldsymbol{C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)}$$

- $\delta_{2s_0} < \Theta_1 = (\sqrt{2} - 1)$ be the $2s_0$-restricted RIP constant for $\boldsymbol{M}$
- Each covariate has a sub-Gaussian distribution with parameter σ^2
- True $\boldsymbol{\beta}^*$ is κ_0-sparse, $\kappa_0 < s_0$
- Zero-mean, sub-Gaussian noise with parameter ρ^2

Aggregated Data with Observation Noise

Theorem 2

Let $\xi > 0$ be any small positive real value. Given noisy empirical aggregates , true β^* can be recovered within an ℓ_2 distance of $O(\xi)$ with probability at least $1 - e^{-C_1 n} - \boldsymbol{e^{-C_2 n}}$.

Here, the constants C_1, C_2 are such that

$$C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1 + \delta_{2s_0})}\right), \quad \boldsymbol{C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)}$$

Fewer samples required for-

- lower value of δ_{2s_0}, the RIP constant for true means $\boldsymbol{M}$
- smaller sub-Gaussian parameters for covariates σ^2, & noise ρ^2
- Higher level of allowed tolerance ξ

Special Case: Histogram Aggregated Data

Theorem 3

Given data in histograms of bin size Δ, the true $\boldsymbol{\beta}^*$ can be recovered within an ℓ_2 distance of $O(\sqrt{k}\Delta)$ with probability at least $1 - e^{-C_1 n}$ where

$$C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$$

Recovery guarantees improve for-

- lower value of δ_{2s_0}, the RIP constant for true means $\boldsymbol{M}$
- lower value of sub-Gaussian parameter for covariates σ^2
- Finer granularity for histogram, i.e., lower bin size Δ

Histogram Aggregated Data

Theorem 2

With probability at least $1 - e^{-C_2 n}$, the sparse parameter β^* can be recovered approximately upto a tolerance specified by histogram granularity

Recovery guarantees improve if-

- true $\boldsymbol{M}$ satisfies RIP with higher fidelity
- data and noise distribution have a small "spread"
- histogram has fine granularity, i.e., lower bin size

Experiments on Synthetic Data: Gaussian Model

- Recovery Probability on Gaussian Ensemble

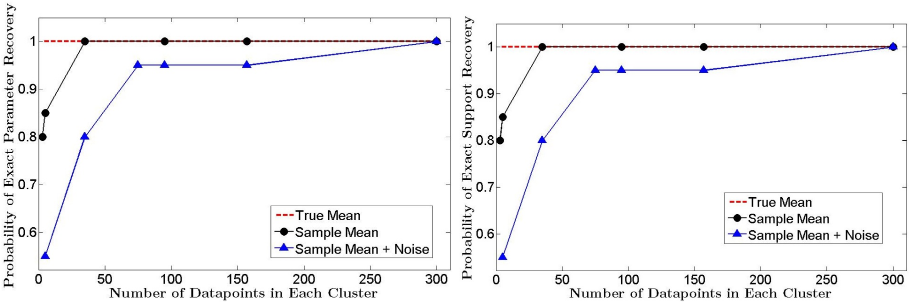


Figure: Probability of Exact Parameter Recovery and Exact Support Recovery for Gaussian Ensemble

Experiments on Synthetic Data: Bernoulli Model

- Recovery Probability on Bernoulli Ensemble

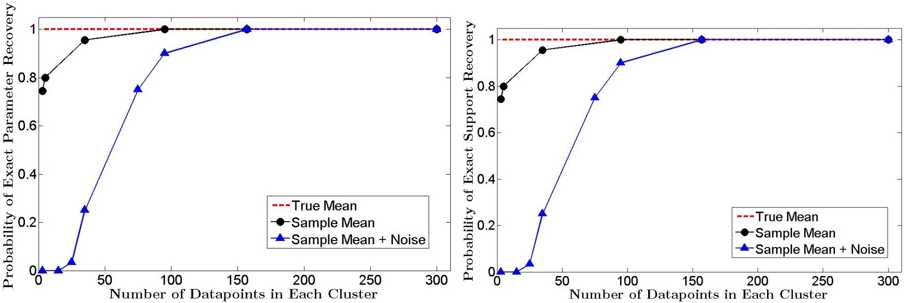


Figure: Probability of Exact Parameter Recovery and Exact Support Recovery for Bernoulli Ensemble

Experiments on Real Data: TxID dataset

- Modelling hospital charges using healthcare billing records in the Texas Inpatient Discharge Dataset (TxID) from the TX Dept. of State Health Services

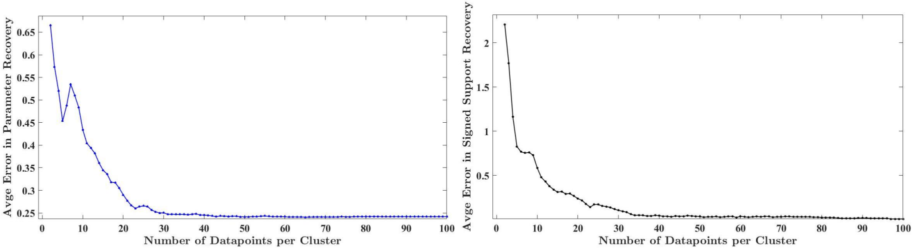


Figure: Parameter Recovery Error and Support Recovery Error for TxID Dataset

Frequency Domain Predictive Modelling with Aggregated Data

- Spatio-Temporally Correllated Data
- Non-uniform Aggregation

- A Bhowmik, J Ghosh, O Koyejo, "Frequency Domain Predictive Modelling with Aggregated Data", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017

Problem Setup

Features $\boldsymbol{x}(t) = [x_1(t), x_2(t) \cdots x_d(t)]$, targets $y(t)$

Weak Stationarity+

- Zero-mean $E[y(t)] = 0$.
- Finite variance $E[y(t)] < \infty$
- Autocorrelation function satisfies: $E[y(t)y(t')] = \rho(\|t - t'\|)$

same assumptions for $\boldsymbol{x}(t)$

Problem Setup

Features $\boldsymbol{x}(t) = [x_1(t), x_2(t) \cdots x_d(t)]$, targets $y(t)$

Weak Stationarity+

- Zero-mean $E[y(t)] = 0$.
- Finite variance $E[y(t)] < \infty$
- Autocorrelation function satisfies: $E[y(t)y(t')] = \rho(\|t - t'\|)$

same assumptions for $\boldsymbol{x}(t)$

Residual process

- Let $\varepsilon_\beta(t) = \boldsymbol{x}(t)^\top \boldsymbol{\beta} - y(t)$ be the residual error process of a linear model
- Observe that $\varepsilon_\beta(t)$ is weakly stationary

Main result I : Generalization Error

Theorem

For every small $\xi > 0$, $\exists$ *corresponding* T_0, D *such that:*

$$E\left[|\boldsymbol{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] < (1+\xi)\left(E\left[|\boldsymbol{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) + 2\xi$$

with probability at least $1 - e^{-O(D^2\xi^2)}$

Thus, generalization error bounded with sufficiently large T_0, D

Main result II : Generalisation Error

Non-uniform aggregation, Finite samples

Theorem

Let ω_i, ω_y *be the sampling rate for* $\mathbf{x}_i(t), y(t)$ *respectively. Let* $\omega_s = \min\{\omega_y, \omega_1, \omega_2, \cdots \omega_d\}$*. Then, for small* $\xi > 0, \exists$ *corresponding* T_0, D *such that:*

$$E\left[\left|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)\right|^2\right] < (1+\xi)\left(E\left[\left|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)\right|^2\right]\right) + 4\xi + 2e^{-O((\omega_s - 2\omega_0)^2)}$$

with probability at least $1 - e^{-O(D^2\xi^2)} - e^{-O(N^2\xi^2)}$

Generalization error can be made small if T_0, D are high, ω_0 is small, minimum sampling frequency ω_s is high

Synthetic Data

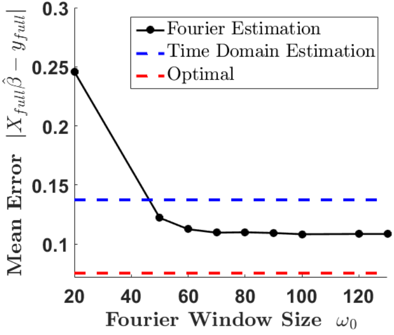


Fig 1(a): No Discrepancy

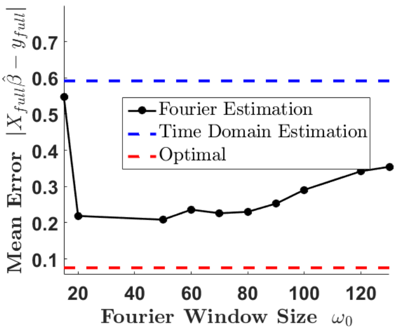


Fig 1(b): Low Discrepancy

- Performance on synthetic data with varying ω_0, and increasing sampling and aggregation discrepancy

Synthetic Data - II

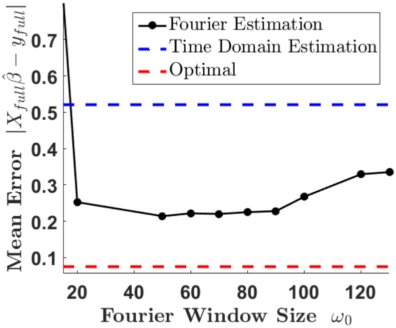


Fig 1(c): Medium Discrepancy

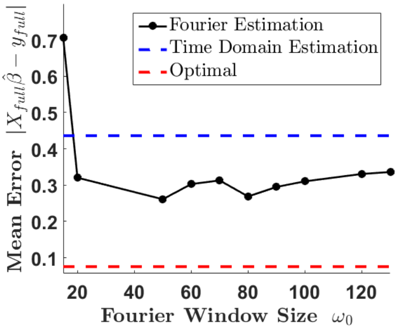
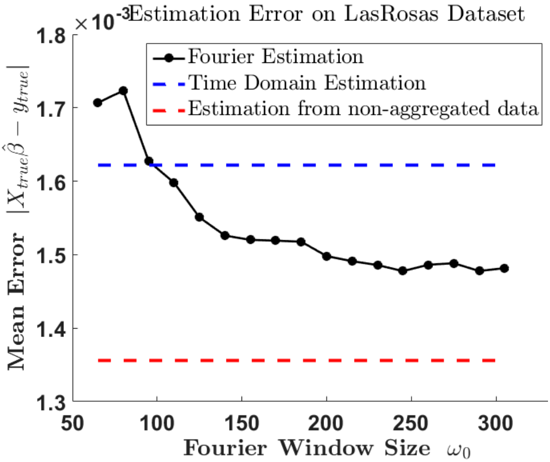


Fig 1(d): High Discrepancy

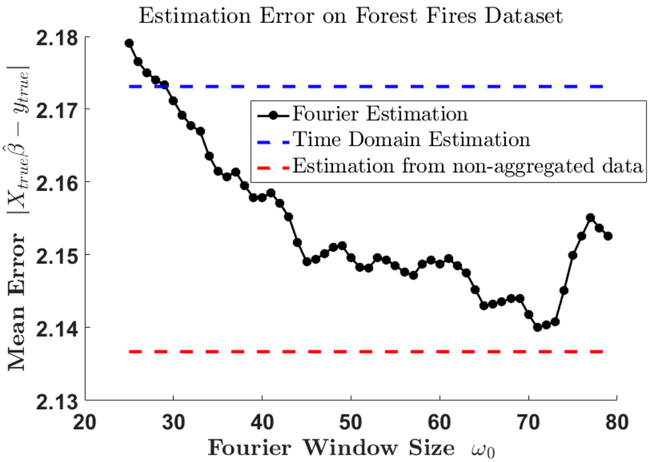
- Performance on synthetic data with varying ω_0, and increasing sampling and aggregation discrepancy

Las Rosas Dataset



Regressing corn yield against nitrogen levels, topographical properties, brightness value, etc.

UCI Forest Fires Dataset



Regressing burned acreage against meteorological features, relative humidity, ISI index, etc. on UCI Forest Fires Dataset