Learning with Aggregated Data; A Tale of Two Approaches

Sanmi Koyejo

University of Illinois at Urbana-Champaign

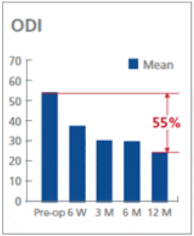
Joint work with

Avradeep Bhowmik

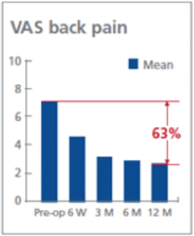
Joydeep Ghosh

@University of Texas at Austin

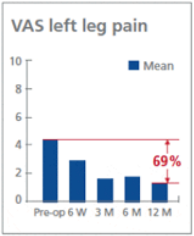
Healthcare data often released in aggregated form



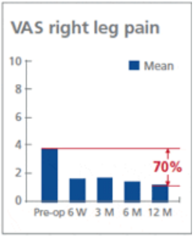
Improvement of ODI of 55% one year post-op.



Improvement of VAS of 63% one year post-op.



Improvement of VAS of 69% one year post-op.

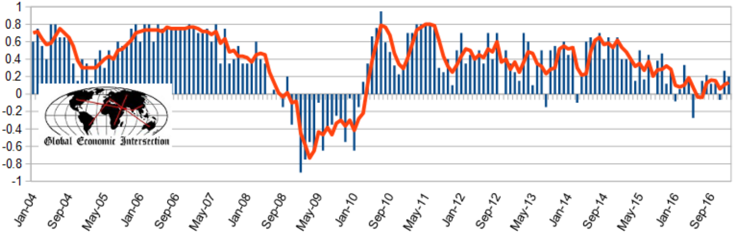


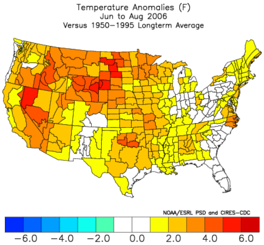
Improvement of VAS of 70% one year post-op.

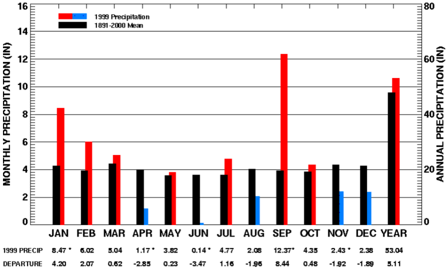
Images courtesy: Paradigm Spine

Aggregation sometimes used to satisfy privacy concerns

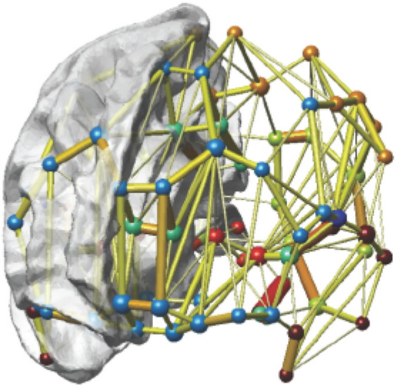
Images courtesy: Joydeep Ghosh and Yubin Park







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



- **Brain Imaging Data:** Observations are aggregated over both space (i.e. voxels) and time

- Data often released in aggregated form in practice (Burrell et al., 2004; Lozano et al., 2009; Davidson et al., 1978)
- Naive fitting of aggregated data may result in ecological fallacy (Freedman et al., 1991; Robinson, 2009)
- Reconstruction (before model fitting) is expensive and unreliable

Motivating question:

Is it possible to learn accurate individual level models from aggregated data?

Yes! In at least two cases:

- high dimensional linear model with group-wise IID data, compressed sensing will recover sparse model[a]
- spatiotemporal data with a linear model estimator, proposed procedure achieves strong generalization error guarantees[a]

[a] under certain conditions...

Motivating question:

Is it possible to learn accurate individual level models from aggregated data?

Yes! In at least two cases:

- high dimensional linear model with group-wise IID data, compressed sensing will recover sparse model[a]
- spatiotemporal data with a linear model estimator, proposed procedure achieves strong generalization error guarantees[a]

[a] under certain conditions...

Motivating question:

Is it possible to learn accurate individual level models from aggregated data?

Yes! In at least two cases:

- high dimensional linear model with group-wise IID data, compressed sensing will recover sparse model[a]
- spatiotemporal data with a linear model estimator, proposed procedure achieves strong generalization error guarantees[a]

[a] under certain conditions...

Motivating question:

Is it possible to learn accurate individual level models from aggregated data?

Yes! In at least two cases:

- high dimensional linear model with group-wise IID data, compressed sensing will recover sparse model[a]
- spatiotemporal data with a linear model estimator, proposed procedure achieves strong generalization error guarantees[a]

[a] under certain conditions...

Related work in statistics

- known as *ecological regression* (Goodman, 1953; Freedman et al., 1991)
- often considered a reasonable technique for anonymizing data (Armstrong et al., 1999)

Related work in machine learning

- most popular in classification, known as learning from label proportions (Quadrianto et al., 2009; Patrini et al., 2014)
- particularly relevant for big data with high label acquisition costs

Other related work

- sensor network / internet of things data may be aggregated to reduce communication overhead (Li et al., 2013; Wagner, 2004; Zhao et al., 2003)

Related work in statistics

- known as *ecological regression* (Goodman, 1953; Freedman et al., 1991)
- often considered a reasonable technique for anonymizing data (Armstrong et al., 1999)

Related work in machine learning

- most popular in classification, known as learning from label proportions (Quadrianto et al., 2009; Patrini et al., 2014)
- particularly relevant for big data with high label acquisition costs

Other related work

- sensor network / internet of things data may be aggregated to reduce communication overhead (Li et al., 2013; Wagner, 2004; Zhao et al., 2003)

Related work in statistics

- known as *ecological regression* (Goodman, 1953; Freedman et al., 1991)
- often considered a reasonable technique for anonymizing data (Armstrong et al., 1999)

Related work in machine learning

- most popular in classification, known as learning from label proportions (Quadrianto et al., 2009; Patrini et al., 2014)
- particularly relevant for big data with high label acquisition costs

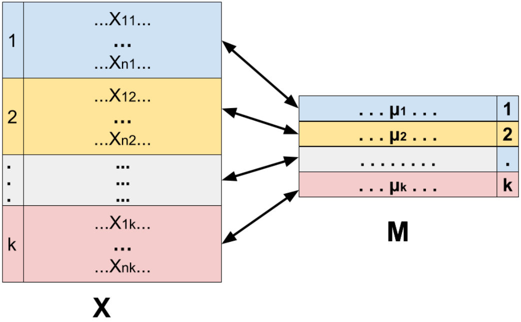
Other related work

- sensor network / internet of things data may be aggregated to reduce communication overhead (Li et al., 2013; Wagner, 2004; Zhao et al., 2003)

Part 1:

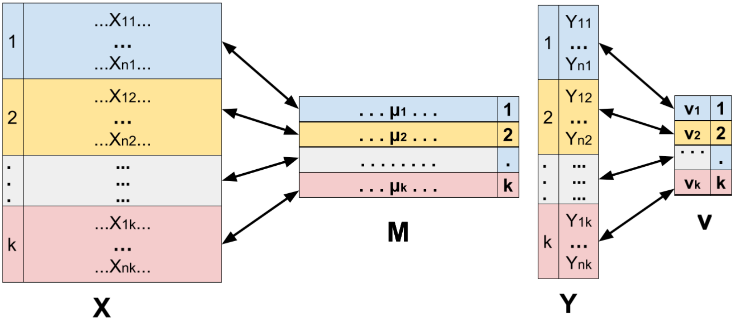
Learning a Sparse Linear model from Group-Wise Aggregated Data

Group-wise data aggregation



(a) Features / Covariates

Group-wise data aggregation

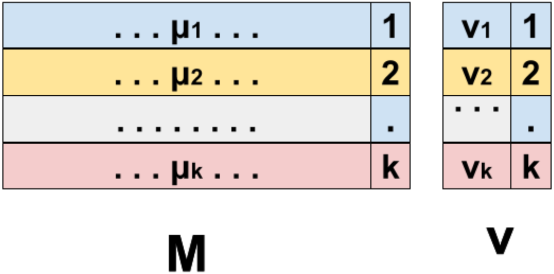


(a) Features / Covariates (b) Targets

Observed training data

- group-wise averages from k population sub-groups

$$\mathbb{D}_{agg} = \left\{ \boldsymbol{\mu}_j = \hat{E}_j[\mathbf{x}], \nu_j = \hat{E}_j[y] \,|\, j = 1, 2, \cdots k \right\}.$$



Population statistics

- for each group $j \in [k]$,

$$\boldsymbol{\mu}_j = E_j[\mathbf{x}],\ \nu_j = E_j[y].$$

With a linear model

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

- if $E[\epsilon] = 0$,

$$E[\mathbf{y}] = E[\mathbf{X}]\boldsymbol{\beta}^* \iff \boldsymbol{\upsilon} = \mathbf{M}\boldsymbol{\beta}^*.$$

---

where expectation is wrt. each group-wise distribution

Population statistics

- for each group $j \in [k]$,

$$\boldsymbol{\mu}_j = E_j[\mathbf{x}],\, \nu_j = E_j[y].$$

With a linear model

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

- if $E[\epsilon] = 0$,

$$E[\mathbf{y}] = E[\mathbf{X}]\boldsymbol{\beta}^* \iff \boldsymbol{\upsilon} = \mathbf{M}\boldsymbol{\beta}^*.$$

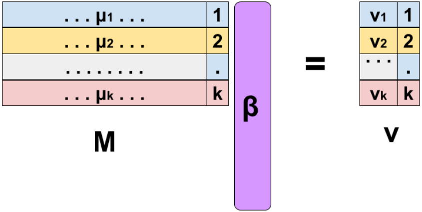
---

where expectation is wrt. each group-wise distribution

Group-wise expectation preserves linear model

- if $k \geq d$, straightforward to estimate $\beta^* \in \mathbf{R}^d$ by solving the linear system

$$\boldsymbol{v} = \mathbf{M}\boldsymbol{\beta}^* \text{ where, } \mathbf{M} \in \mathbf{R}^{k \times d}, \boldsymbol{v} \in \mathbf{R}^k$$

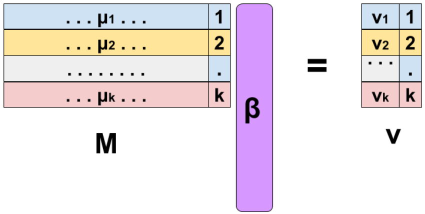


- if $k \ll d$ i.e. under-determined system, recovery is no longer possible without additional assumptions

Group-wise expectation preserves linear model

- if $k \geq d$, straightforward to estimate $\boldsymbol{\beta}^* \in \mathbf{R}^d$ by solving the linear system

$$\boldsymbol{v} = \mathbf{M}\boldsymbol{\beta}^* \text{ where, } \mathbf{M} \in \mathbf{R}^{k \times d}, \boldsymbol{v} \in \mathbf{R}^k$$



- if $k \ll d$ i.e. under-determined system, recovery is no longer possible without additional assumptions

Sparse parameter estimation from true group means

Restricted Isometry Property

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

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

- Informally, every small submatrix behaves approximately like an orthonormal system

Informal Lemma (Recovery with population means)

Suppose $\mathbf{M}$ satisfies (s, δ_s)-RIP, given $(\mathbf{M}, \boldsymbol{\upsilon})$, a sparse β^* can be estimated using standard compressed sensing techniques[a]

[a] Donoho (2006); Candes et al. (2006); Foucart (2010)

Sparse parameter estimation from true group means

Restricted Isometry Property

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

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

- Informally, every small submatrix behaves approximately like an orthonormal system

Informal Lemma (Recovery with population means)

Suppose $\mathbf{M}$ satisfies (s, δ_s)-RIP, given $(\mathbf{M}, \boldsymbol{\upsilon})$, a sparse $\boldsymbol{\beta}^*$ can be estimated using standard compressed sensing techniques[a]

[a] Donoho (2006); Candes et al. (2006); Foucart (2010)

Empirical aggregation error

- however, $(\mathbf{M}, \boldsymbol{\upsilon})$ unknown in practice, instead use estimates:

$$\widehat{\mathbf{M}}_n[j] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,j}, \ \ \hat{\boldsymbol{\upsilon}}_n[j] = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}.$$

- results in additional empirical error:

$$\widehat{\mathbf{M}}_n = \mathbf{M} + \zeta_{x,n}, \ \ \hat{\boldsymbol{\upsilon}}_n = \boldsymbol{\upsilon} + \zeta_{y,n}.$$

- **Key Insight:** aggregation is a linear procedure, thus:

$$\hat{\boldsymbol{\upsilon}}_n = \widehat{\mathbf{M}}_n^\top \beta^* \text{ and } \zeta_{y,n} = \zeta_{x,n}^\top \beta^*.$$

Empirical aggregation error

- however, $(\mathbf{M}, \boldsymbol{\upsilon})$ unknown in practice, instead use estimates:

$$\widehat{\mathbf{M}}_n[j] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,j}, \;\; \hat{\boldsymbol{\upsilon}}_n[j] = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}.$$

- results in additional empirical error:

$$\widehat{\mathbf{M}}_n = \mathbf{M} + \boldsymbol{\zeta}_{x,n}, \;\; \hat{\boldsymbol{\upsilon}}_n = \boldsymbol{\upsilon} + \boldsymbol{\zeta}_{y,n}.$$

- **Key Insight:** aggregation is a linear procedure, thus:

$$\hat{\boldsymbol{\upsilon}}_n = \widehat{\mathbf{M}}_n^\top \boldsymbol{\beta}^* \text{ and } \boldsymbol{\zeta}_{y,n} = \boldsymbol{\zeta}_{x,n}^\top \boldsymbol{\beta}^*.$$

Empirical aggregation error

- however, $(\mathbf{M}, \boldsymbol{\upsilon})$ unknown in practice, instead use estimates:

$$\widehat{\mathbf{M}}_n[j] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,j}, \;\; \hat{\boldsymbol{\upsilon}}_n[j] = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}.$$

- results in additional empirical error:

$$\widehat{\mathbf{M}}_n = \mathbf{M} + \boldsymbol{\zeta}_{x,n}, \;\; \hat{\boldsymbol{\upsilon}}_n = \boldsymbol{\upsilon} + \boldsymbol{\zeta}_{y,n}.$$

- **Key Insight:** aggregation is a linear procedure, thus:

$$\hat{\boldsymbol{\upsilon}}_n = \widehat{\mathbf{M}}_n^\top \boldsymbol{\beta}^* \text{ and } \boldsymbol{\zeta}_{y,n} = \boldsymbol{\zeta}_{x,n}^\top \boldsymbol{\beta}^*.$$

Main Results

Additive noise-free aggregated data

$$\text{Solve } \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \widehat{\mathbf{M}}_n \boldsymbol{\beta} = \hat{\boldsymbol{v}}_n.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\boldsymbol{\beta}^*$ is recovered exactly with probability at least $1 - e^{-C_0 n}$,

Additive noise-free aggregated data

$$\text{Solve } \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \widehat{\mathbf{M}}_n\boldsymbol{\beta} = \hat{\boldsymbol{\upsilon}}_n.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\boldsymbol{\beta}^*$ is recovered exactly with probability at least $1 - e^{-C_0 n}$,

where:

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

- $\boldsymbol{\beta}^*$ is κ_0-sparse, $\kappa_0 < s_0$
- $\delta_{2s_0} < \Theta_0 \approx 0.465$ is $2s_0$-restricted RIP constant for $\mathbf{M}$
- X is sub-Gaussian with parameter σ^2

Additive noise-free aggregated data

$$\text{Solve } \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \widehat{\mathbf{M}}_n \boldsymbol{\beta} = \hat{\boldsymbol{v}}_n.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\boldsymbol{\beta}^*$ is recovered exactly with probability at least $1 - e^{-C_0 n}$,

where:

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

Observe that fewer samples required for estimating $\widehat{\mathbf{M}}_n$ when:

- smaller RIP constant for true means $\mathbf{M}$ i.e. δ_{2s_0}
- thinner tails i.e. smaller σ^2

Additive noise-free aggregated data

$$\text{Solve } \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \widehat{\mathbf{M}}_n \boldsymbol{\beta} = \hat{\boldsymbol{v}}_n.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\boldsymbol{\beta}^*$ is recovered exactly with probability at least $1 - e^{-C_0 n}$,

where:

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

- contrast with prior work that assume error in measurement matrix and/or targets, but only provide approximate recovery (Herman and Strohmer, 2010; Zhao and Yu, 2006; Rudelson and Zhou, 2015)

Aggregated data with observation noise

- each sample measurement corrupted by zero mean additive noise as
$$y = \mathbf{x}^\top \boldsymbol{\beta}^* + \epsilon.$$
- means $(\widehat{\mathbf{M}}_n, \hat{\boldsymbol{\upsilon}}_\epsilon)$ computed from noisy obs. for each group
$$\widehat{\mathbf{M}}_n = \mathbf{M} + \boldsymbol{\zeta}_{x,n}, \quad \hat{\boldsymbol{\upsilon}}_n = \boldsymbol{\upsilon} + \boldsymbol{\zeta}_{y,n} + \boldsymbol{\epsilon}_n.$$

Aggregated data with observation noise - II

$$\text{Solve } \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \|\widehat{\mathbf{M}}_n \boldsymbol{\beta} - \hat{\boldsymbol{v}}_\epsilon\|_2 < \xi.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\left\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\right\| \leq O(\xi)$ with probability at least $1 - e^{-C_1 n} - e^{-C_2 n}$.

Aggregated data with observation noise - II

$$\text{Solve } \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \|\widehat{\mathbf{M}}_n \boldsymbol{\beta} - \hat{\boldsymbol{v}}_\epsilon\|_2 < \xi.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\left\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\right\| \le O(\xi)$ with probability at least $1 - e^{-C_1 n} - e^{-C_2 n}$.

where:

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

- $\boldsymbol{\beta}^*$ is κ_0-sparse, $\kappa_0 < s_0$
- $\delta_{2s_0} < \Theta_1 = (\sqrt{2} - 1)$ is $2s_0$-restricted RIP constant for $\mathbf{M}$
- (X, ϵ) sub-Gaussian with parameters (σ^2, ρ^2) respectively

Aggregated data with observation noise - II

$$\text{Solve } \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \|\widehat{\mathbf{M}}_n \boldsymbol{\beta} - \hat{\boldsymbol{v}}_\epsilon\|_2 < \xi.$$

Theorem (Bhowmik, Ghosh, and Koyejo (2016))

$\left\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\right\| \leq O(\xi)$ with probability at least $1 - e^{-C_1 n} - e^{-C_2 n}$.

where:

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

Observe that fewer samples required for estimating $\widehat{\mathbf{M}}_n$ when:

- smaller RIP constant for true means $\mathbf{M}$ i.e. δ_{2s_0}
- thinner tails i.e. smaller σ^2, ρ^2
- looser tolerance ξ

Empirical Evaluation

Synthetic data

$d = 150, k = 45, \sigma^2 = 0.1, s = 30$

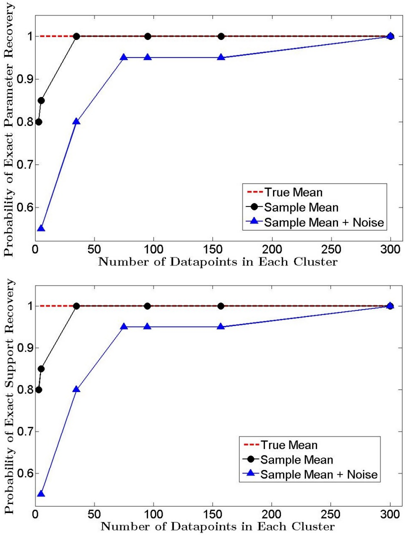


Figure: Probability of exact parameter recovery and exact support recovery for Gaussian ensemble

Synthetic data - II

$d = 150, k = 45, \sigma^2 = 0.1, s = 30$

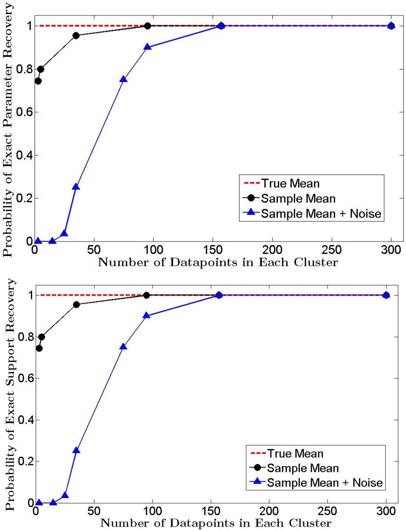


Figure: Probability of exact parameter recovery and exact support recovery for Bernoulli ensemble

Annual outpatient reimbursement (Lousiana, 2008)

- dataset from the Centers for Medicare and Medicaid Services
- predictor variables include duration of coverage, chronic conditions, etc. $(d = 24, k = 12)$

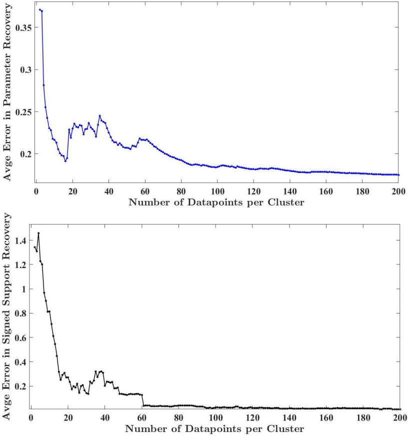


Figure: Parameter Recovery and Support Recovery vs. Lasso

Healthcare charges (Texas, 4^{th} quarter of 2006)

- dataset from Texas Department of State Health Services
- predictor variables include demographic information, length of hospital stay, etc. ($d = 213, k = 15$)

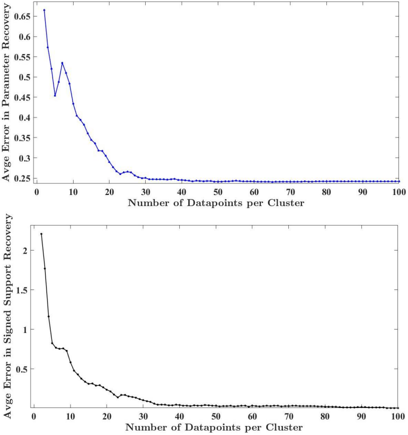


Figure: Parameter Recovery and Support Recovery vs. Lasso

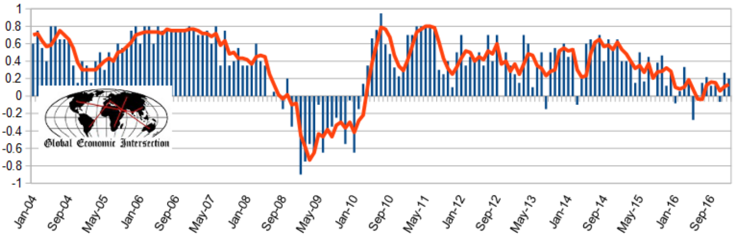
Summary

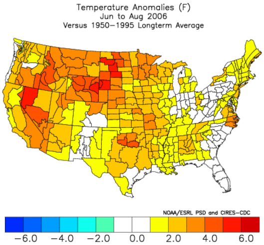
Part 1

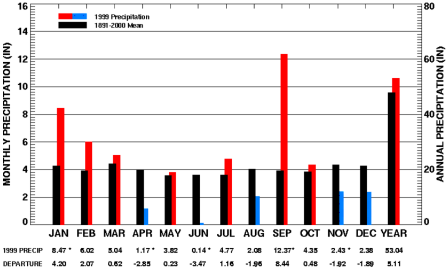
- Presented an analysis of sparse parameter recovery from aggregated data, subject to:
  - empirical aggregation errors
  - additive noise
- Application to healthcare
  - predictive modeling of CMS Medicare reimbursements
  - estimation of Texas state hospital charges
- Manuscript includes additional discussion:
  - higher order moments
  - data aggregated as histograms

Part 2:

Learning a Linear model with Aggregated Spatio-temporal Data







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

Motivation

- Aggregation often applied to time series, spatial data, spatio-temporal data, ...
- Worse, aggregation periods may not be aligned or uniform[1]
  - 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

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

Main Contribution

Model estimation procedure in the frequency domain

- avoids input data reconstruction
- achieves provably bounded generalization error.

Problem Setup

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

Main Contribution

Model estimation procedure in the frequency domain

- avoids input data reconstruction
- achieves provably bounded generalization error.

Problem Setup

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

Residual process

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

Performance Evaluation

- performance measure is the expected squared residual error

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\varepsilon_\beta(t)|^2] = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

- which is optimized as:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

Residual process

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

Performance Evaluation

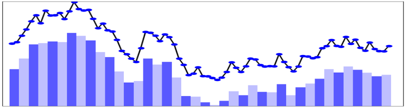
- performance measure is the expected squared residual error

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\varepsilon_\beta(t)|^2] = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

- which is optimized as:

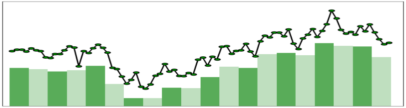
$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

Data aggregation in time series



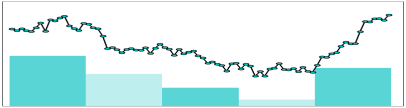
non-aggregated feature $\mathbf{X}_1$

aggregated feature $\overline{\mathbf{X}}_1$



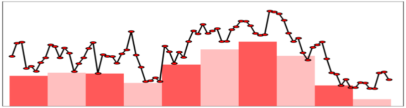
non-aggregated feature $\mathbf{X}_2$

aggregated feature $\overline{\mathbf{X}}_2$



non-aggregated feature $\mathbf{X}_3$

aggregated feature $\overline{\mathbf{X}}_3$



non-aggregated target $\mathbf{Y}$

aggregated target $\overline{\mathbf{Y}}$

Data aggregation in time series - II

- each coordinate of the feature set is aggregated

$$\overline{\mathbf{x}}_i[l] = \frac{1}{T_i} \int\limits_{(l-1)T_i/2}^{lT_i/2} x_i(\tau)d\tau$$

- similarly, the targets are aggregated

$$\overline{\mathbf{y}}[k] = \frac{1}{T} \int\limits_{(k-1)T/2}^{kT/2} y(\tau)d\tau$$

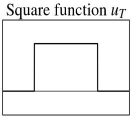
for $k, l \in \mathbb{Z} = \{\cdots - 1, 0, 1, \cdots\}$.

Aggregation: time and frequency domain

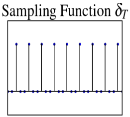
Fourier space captures global properties of the signal

In time domain, convolution with square wave + sampling

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



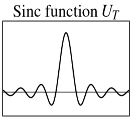
$\xrightarrow{\textit{sampling}}$



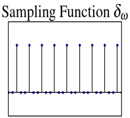
$\longrightarrow \overline{z}[k]$

In frequency domain, multiplication with sinc function + sampling

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



$\xrightarrow{\textit{sampling}}$



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

Restricted Fourier transform

For signal $z(t)$, T-**restricted Fourier Transform** defined as:

$$Z_T(\omega) = \mathcal{F}_T[z](\omega) = \int_{-T}^{T} z(t) e^{-\iota\omega t} dt$$

- equivalent to a full Fourier Transform if the signal is time-limited within $(-T, T)$
- always exists finitely if the signal $z(t)$ is finite

Time-limited data

- infinite time series data are not available, instead assume data available between time intervals $(-T_0, T_0)$
- we apply T_0-restricted Fourier transforms computed from time-limited data
- assume time-restricted Fourier transform decay rapidly with frequency e.g. autocorrelation function is a Schwartz function (TerzioĞglu, 1969)
- thus, most of the signal power between frequencies $(-\omega_0, \omega_0)$

Proposed Algorithm

Step I

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

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 $\overline{\mathbf{X}}_{T_0}(\omega), \mathbf{Y}_{T_0}(\omega)$ from aggregated signals $\overline{\mathbf{x}}[k], \mathbf{y}[l]$

Step I

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

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 $\overline{\mathbf{X}}_{T_0}(\omega), \mathbf{Y}_{T_0}(\omega)$ from aggregated signals $\overline{\mathbf{x}}[k], \mathbf{y}[l]$

Step I

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

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 $\overline{\mathbf{X}}_{T_0}(\omega), \mathbf{Y}_{T_0}(\omega)$ from aggregated signals $\overline{\mathbf{x}}[k], \mathbf{y}[l]$

Step II

Recall: U_T is Fourier transform of square wave

4. estimate non-aggregated Fourier transforms

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

5. estimate parameter $\hat{\beta}$ as:

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

Step II

Recall: U_T is Fourier transform of square wave

4. estimate non-aggregated Fourier transforms

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

5. estimate parameter $\hat{\boldsymbol{\beta}}$ as:

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

Generalization Analysis

Main result I

Theorem (Bhowmik, Ghosh, and Koyejo (2017))

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

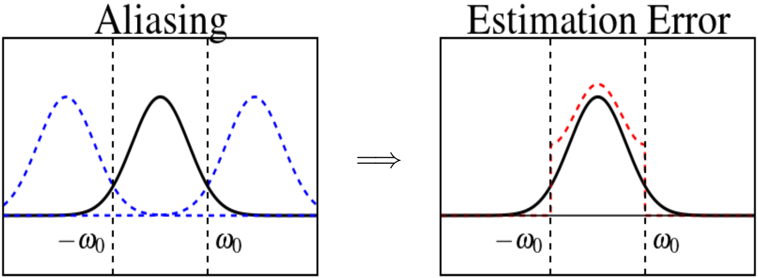
$$E\left[|\mathbf{x}(t)^\top \hat{\boldsymbol{\beta}} - y(t)|^2\right] < (1+\xi)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) + 2\xi$$

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

Thus, generalization error bounded with sufficiently large T_0, D

Aliasing effects, non-uniform sampling

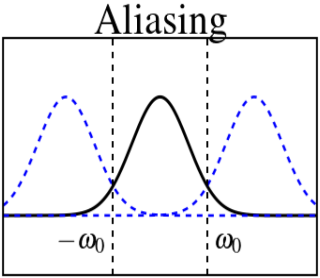
- signals not bandlimited $\Rightarrow$ Aliasing
- errors minimum for frequencies around 0



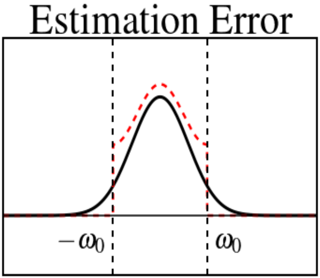
- non-uniform sampling leads to further error
- performance will depend on rapid decay of power spectral density

Aliasing effects, non-uniform sampling

- signals not bandlimited $\Rightarrow$ Aliasing
- errors minimum for frequencies around 0



$\Longrightarrow$



- non-uniform sampling leads to further error
- performance will depend on rapid decay of power spectral density

Main result II

Non-uniform aggregation, finite samples

Theorem (Bhowmik, Ghosh, and Koyejo (2017))

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:*

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

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

Additional details

- more detailed analysis (not shown) allows for more precise error control
- algorithm and analysis easily extend to multi-dimensional indexes e.g. spatio-temporal data using the multi-dimensional Fourier transform
  - number of frequency samples may depend exponentially on index dimension (typically < 4)
- extends to cases where aggregation and sampling period are non-overlapping.
- extends to sliding windows, weighted smoothing

Empirical Evaluation

Synthetic Data

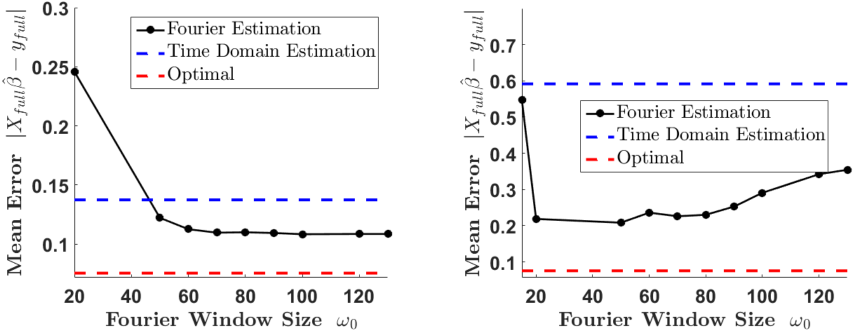


Fig 1(a): No discrepancy

Fig 1(b): Low discrepancy

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

Synthetic Data - II

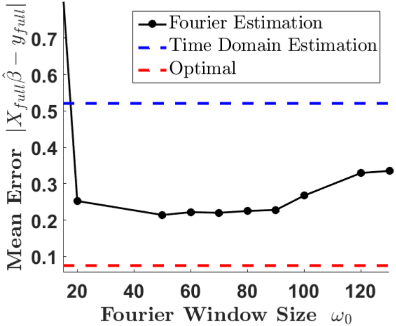


Fig 1(c): Medium discrepancy

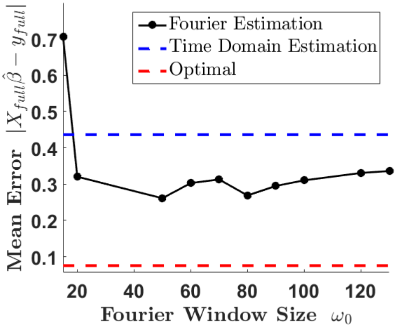
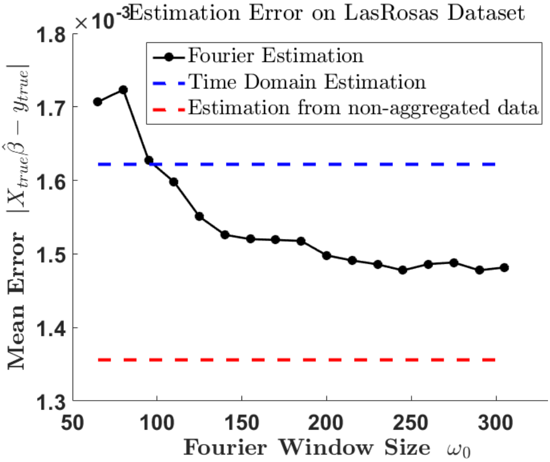


Fig 1(d): High discrepancy

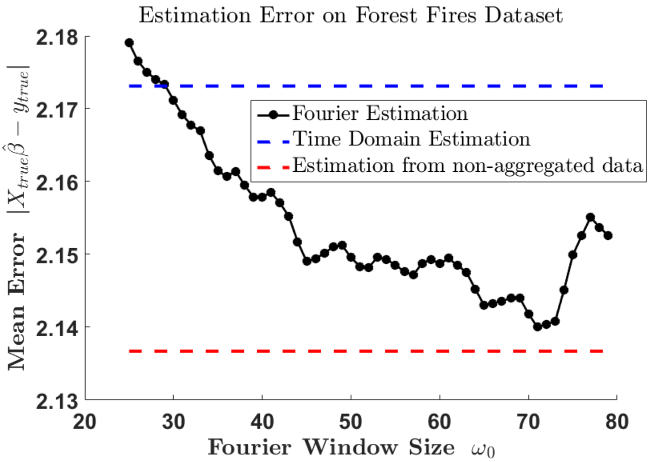
- performance on synthetic data with varying ω_0, and increasing sampling and aggregation discrepancy

Las Rosas dataset



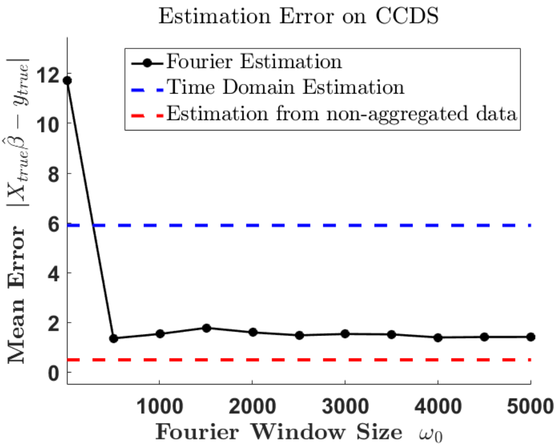
Regressing corn yield against nitrogen levels, topographical properties, brightness value, etc.

UCI forest fires dataset



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

Comprehensive climate dataset (CCDS)



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

Conclusion

Part 2

- proposed a novel procedure with bounded generalization error for learning with aggregated data
- significant improvements vs reconstruction-based estimation.

Future work:

- exploit frequency domain structure e.g. sparse spectrum to improve estimates.
- exploit generative structure e.g. sparse models to improve estimates.

Conclusion

Part 2

- proposed a novel procedure with bounded generalization error for learning with aggregated data
- significant improvements vs reconstruction-based estimation.

Future work:

- exploit frequency domain structure e.g. sparse spectrum to improve estimates.
- exploit generative structure e.g. sparse models to improve estimates.

Overall conclusion

It possible to learn provably accurate individual level models from aggregated data in at least two cases

- high dimensional linear model with group-wise IID data, compressed sensing will recover sparse model[a]
- spatiotemporal data with a linear model estimator, freq-domain regression achieves strong generalization error guarantees[a]

[a] under certain conditions...

Future work

- Can we learn from richer aggregate information?
  c.f. distribution regression (Szabó et al., 2016; Bhowmik et al., 2015)
- What can we say about non-linear models?
- Can we design aggregation that makes learning *easier*?
  Related to sufficient statistics, sketching
- Can we design aggregation that makes learning *harder*?
  Related to preserving privacy

Future work

- Can we learn from richer aggregate information?
  c.f. distribution regression (Szabó et al., 2016; Bhowmik et al., 2015)
- What can we say about non-linear models?
- Can we design aggregation that makes learning *easier*?
  Related to sufficient statistics, sketching
- Can we design aggregation that makes learning *harder*?
  Related to preserving privacy

Future work

- Can we learn from richer aggregate information?
  c.f. distribution regression (Szabó et al., 2016; Bhowmik et al., 2015)
- What can we say about non-linear models?
- Can we design aggregation that makes learning *easier*?
  Related to sufficient statistics, sketching
- Can we design aggregation that makes learning *harder*?
  Related to preserving privacy

Future work

- Can we learn from richer aggregate information?
  c.f. distribution regression (Szabó et al., 2016; Bhowmik et al., 2015)
- What can we say about non-linear models?
- Can we design aggregation that makes learning *easier*?
  Related to sufficient statistics, sketching
- Can we design aggregation that makes learning *harder*?
  Related to preserving privacy

Thank You!

Bhowmik, A., Ghosh, J. and Koyejo, O. *Frequency Domain Predictive Modeling with Aggregated Data*, AISTATS 2017.
Bhowmik, A., Ghosh, J. and Koyejo, O. *Sparse Parameter Recovery from Aggregated Data*, ICML 2016.
Bhowmik, A., Ghosh, J. and Koyejo, O. *Generalized linear models for aggregated data*. AISTATS 2015.

sanmi@illinois.edu

References

References I

Marc P Armstrong, Gerard Rushton, and Dale L Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5):497–525, 1999.

Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Generalized Linear Models for Aggregated Data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 93–101, 2015.

Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Sparse parameter recovery from aggregated data. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1090–1099, 2016.

Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Frequency domain predictive modelling with aggregated data. In *Proceedings of the 20th International conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

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, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 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.

David A Freedman, Stephen P Klein, Jerome Sacks, Charles A Smyth, and Charles G Everett. Ecological regression and voting rights. *Evaluation Review*, 15(6):673–711, 1991.

Leo A Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

References II

Matthew A Herman and Thomas Strohmer. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, 2010.

Shancang Li, Li Da Xu, and Xinheng Wang. Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Transactions on Industrial Informatics*, 9(4):2177–2186, 2013.

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.

Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374, 2009.

William S Robinson. Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2):337–341, 2009.

Mark Rudelson and Shuheng Zhou. High dimensional errors-in-variables models with dependent measurements. *arXiv preprint arXiv:1502.02355*, 2015.

Zoltán Szabó, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

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

David Wagner. Resilient aggregation in sensor networks. In *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*, pages 78–87. ACM, 2004.

Jerry Zhao, Ramesh Govindan, and Deborah Estrin. Computing aggregates for monitoring wireless sensor networks. In *Sensor Network Protocols and Applications, 2003*, pages 139–148. IEEE, 2003.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.