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
ODI

- Mean

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

VAS back pain

- Mean

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

VAS left leg pain

- Mean

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

VAS right leg pain

- Mean

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

Images courtesy: Paradigm Spine
Aggregation sometimes used to satisfy privacy concerns
Visualizing data aggregation

(a) Features / Covariates

<table>
<thead>
<tr>
<th></th>
<th>...X_{11}...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...X_{n1}...</td>
</tr>
<tr>
<td>2</td>
<td>...X_{12}...</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...X_{n2}...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>k</td>
<td>...X_{1k}...</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...X_{nk}...</td>
</tr>
</tbody>
</table>

\[ X \]

\[ (a) \text{ Features / Covariates} \]

\[ M \]

\[ \begin{array}{c|c}
\ldots & 1 \\
\ldots & 2 \\
\ldots & k \\
\end{array} \]
Visualizing data aggregation

(a) Features / Covariates

(b) Targets
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 anatomizing 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 anatomizing 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 anatomizing 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 High-Dimensional Aggregated Data
Observed training data

- group-wise averages from \( k \) population sub-groups

\[
\mathcal{D}_{agg} = \left\{ \mu_j = \hat{E}_j[x], \nu_j = \hat{E}_j[y] \mid j = 1, 2, \ldots k \right\}.
\]
Population statistics

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

\[ \mu_j = E_j[x], \quad \nu_j = E_j[y]. \]

With a linear model

\[ y = x^\top \beta^* + \epsilon. \]

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

\[ E[y] = E[X] \beta^* \iff \nu = M\beta^*. \]

where expectation is wrt. each group-wise distribution.
Population statistics

- for each group \( j \in [k] \),

\[
\mu_j = E_j[x], \ \nu_j = E_j[y].
\]

With a linear model

\[
y = x^\top \beta^* + \epsilon.
\]

- if \( E[\epsilon] = 0 \),

\[
E[y] = E[X]\beta^* \iff \nu = M\beta^*.
\]

where expectation is wrt. each group-wise distribution
Group-wise expectation preserves linear model

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

$$\nu = M\beta^* \quad \text{where,} \quad M \in \mathbb{R}^{k \times d}, \nu \in \mathbb{R}^k$$

- 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 \( \beta^* \in \mathbb{R}^d \) by solving the linear system

\[
u = M\beta^* \quad \text{where,} \quad M \in \mathbb{R}^{k \times d}, \quad \nu \in \mathbb{R}^k
\]

- 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

- $M$ satisfies $(s, \delta_s)$-RIP if for any $s$-sparse $z$

\[(1 - \delta_s)\|z\|_2^2 \leq \|Mz\|_2^2 \leq (1 + \delta_s)\|z\|_2^2\]

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

Informal Lemma (Recovery with population means)

Suppose $M$ satisfies $(s, \delta_s)$-RIP, given $(M, \nu)$, a sparse $\beta^*$ can be estimated using standard compressed sensing techniques\(^\text{a}\)

\[^{\text{a}}\text{Donoho (2006); Candes et al. (2006); Foucart (2010)}\]
Sparse parameter estimation from true group means

Restricted Isometry Property

- $M$ satisfies $(s, \delta_s)$-RIP if for any $s$-sparse $z$

\[(1 - \delta_s) \|z\|_2^2 \leq \|Mz\|_2^2 \leq (1 + \delta_s) \|z\|_2^2\]

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

Informal Lemma (Recovery with population means)

Suppose $M$ satisfies $(s, \delta_s)$-RIP, given $(M, \nu)$, a sparse $\beta^*$ can be estimated using standard compressed sensing techniques\(^a\)

\(^{\text{aDonoho (2006); Candes et al. (2006); Foucart (2010)}}\)
Empirical aggregation error

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

\[
\hat{\mathbf{M}}_n[j] = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}, \quad \hat{\nu}_n[j] = \frac{1}{n} \sum_{i=1}^{n} y_{i,j}.
\]

- results in additional empirical error:

\[
\hat{\mathbf{M}}_n = \mathbf{M} + \zeta_{x,n}, \quad \hat{\nu}_n = \nu + \zeta_{y,n}.
\]

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

\[
\hat{\nu}_n = \hat{\mathbf{M}}_n^\top \beta^* \text{ and } \zeta_{y,n} = \zeta_{x,n}^\top \beta^*.
\]
Empirical aggregation error

- however, \((M, \nu)\) unknown in practice, instead use estimates:

\[
\hat{M}_n[j] = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}, \quad \hat{\nu}_n[j] = \frac{1}{n} \sum_{i=1}^{n} y_{i,j}.
\]

- results in additional empirical error:

\[
\hat{M}_n = M + \zeta_{x,n}, \quad \hat{\nu}_n = \nu + \zeta_{y,n}.
\]

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

\[
\hat{\nu}_n = \hat{M}_n^\top \beta^* \quad \text{and} \quad \zeta_{y,n} = \zeta_{x,n}^\top \beta^*.
\]
Empirical aggregation error

- however, \((M, \nu)\) unknown in practice, instead use estimates:

\[
\hat{M}_n[j] = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}, \quad \hat{\nu}_n[j] = \frac{1}{n} \sum_{i=1}^{n} y_{i,j}.
\]

- results in additional empirical error:

\[
\hat{M}_n = M + \zeta_{x,n}, \quad \hat{\nu}_n = \nu + \zeta_{y,n}.
\]

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

\[
\hat{\nu}_n = \hat{M}_n^\top \beta^* \text{ and } \zeta_{y,n} = \zeta_{x,n}^\top \beta^*.
\]
Main Results
Additive noise-free aggregated data

Solve \( \min_{\beta} \|\beta\|_1 \) s.t. \( \hat{M}_n \beta = \hat{v}_n \).

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

\( \beta^* \) is recovered exactly with probability at least \( 1 - e^{-C_0 n} \),
Additive noise-free aggregated data

Solve
\[
\min_{\beta} \|\beta\|_1 \quad \text{s.t.} \quad \hat{M}_n \beta = \hat{\upsilon}_n.
\]

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

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

where:

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

- $\beta^*$ is $\kappa_0$-sparse, $\kappa_0 < s_0$
- $\delta_{2s_0} < \Theta_0 \approx 0.465$ is $2s_0$-restricted RIP constant for $M$
- $X$ is sub-Gaussian with parameter $\sigma^2$
Additive noise-free aggregated data

Solve \( \min_{\beta} \|\beta\|_1 \) s.t. \( \widehat{M}_n \beta = \widehat{\upsilon}_n \).

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

\( \beta^* \) 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 \( \widehat{M}_n \) when:

- smaller RIP constant for true means \( M \) i.e. \( \delta_{2s_0} \)
- thinner tails i.e. smaller \( \sigma^2 \)
Additive noise-free aggregated data

\[
\text{Solve } \min_{\beta} \|\beta\|_1 \quad \text{s.t. } \hat{M}_n \beta = \hat{\upsilon}_n.
\]

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

\( \beta^* \) 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
  \[ y = \mathbf{x}^\top \beta^* + \epsilon. \]
- means \((\hat{M}_n, \hat{\nu}_n)\) computed from noisy obs. for each group
  \[ \hat{M}_n = \mathbf{M} + \zeta_{x,n}, \quad \hat{\nu}_n = \nu + \zeta_{y,n} + \epsilon_n. \]
Aggregated data with observation noise - II

\[
\hat{\beta} = \underset{\beta}{\text{arg min}} \| \beta \|_1 \quad \text{s.t.} \quad \| \hat{M}_n \beta - \hat{\nu}_\epsilon \|_2 < \xi.
\]

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

\[
\| \beta^* - \hat{\beta} \| \leq O(\xi) \text{ with probability at least } 1 - e^{-C_1 n} - e^{-C_2 n}.
\]
Aggregated data with observation noise - II

Solve $\hat{\beta} = \arg\min_{\beta} \|\beta\|_1$ s.t. $\|\hat{M}_n \beta - \hat{\nu}_\epsilon\|_2 < \xi$.

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

$\|\beta^* - \hat{\beta}\| \leq O(\xi)$ with probability at least $1 - e^{-C_1n} - e^{-C_2n}$.

where:

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

- $\beta^*$ is $\kappa_0$-sparse, $\kappa_0 < s_0$
- $\delta_{2s_0} < \Theta_1 = (\sqrt{2} - 1)$ is $2s_0$-restricted RIP constant for $M$
- $(X, \epsilon)$ sub-Gaussian with parameters $(\sigma^2, \rho^2)$ respectively
Aggregated data with observation noise - II

Solve \( \hat{\beta} = \arg \min_{\beta} \| \beta \|_1 \) s.t. \( \| \hat{M}_n \beta - \hat{\nu}_\epsilon \|_2 < \xi \).

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

\[ \| \beta^* - \hat{\beta} \| \leq O(\xi) \] 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 \( \hat{M}_n \) when:

- smaller RIP constant for true means \( M \) i.e. \( \delta_{2s_0} \)
- thinner tails i.e. smaller \( \sigma^2, \rho^2 \)
- looser tolerance \( \xi \)
Histogram aggregated data (bin-size $\Delta$)

Estimate the mean of histogram approx. distribution

$$\text{Solve } \hat{\beta} = \arg \min_{\beta} \| \beta \|_1 \text{ s.t. } \| \hat{M}_n \beta - \hat{\nu}_\epsilon \|_2 < \xi_\Delta.$$  

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

$$\| \beta^* - \hat{\beta} \| \leq O(\sqrt{k} \Delta) \text{ with probability at least } 1 - e^{-C_1 n}.$$
Histogram aggregated data (bin-size $\Delta$)

Estimate the mean of histogram approx. distribution

Solve $\hat{\beta} = \arg \min_{\beta} \|\beta\|_1 \text{ s.t. } \|\hat{M}_n\beta - \hat{\nu}_\epsilon\|_2 < \xi_\Delta$.

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

$\|\beta^* - \hat{\beta}\| \leq O(\sqrt{k}\Delta)$ with probability at least $1 - e^{-C_1n}$.

where:

$C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1 + \delta_{2s_0})}\right)$,  \hspace{1cm} \xi_\Delta = \frac{\sqrt{k}\Delta}{2}$.

- $\beta^*$ is $\kappa_0$-sparse, $\kappa_0 < s_0$
- $\delta_{2s_0} < \Theta_1 = (\sqrt{2} - 1)$ is $2s_0$-restricted RIP constant for $M$
- $X$ sub-Gaussian with parameter $\sigma^2$
Histogram aggregated data (bin-size \( \Delta \))

Estimate the mean of histogram approx. distribution

\[
\text{Solve } \hat{\beta} = \arg \min_{\beta} \| \beta \|_1 \text{ s.t. } \| \widehat{M}_n \beta - \hat{\nu}_\epsilon \|_2 < \xi_\Delta.
\]

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

\[
\| \beta^* - \hat{\beta} \| \leq O(\sqrt{k\Delta}) \text{ with probability at least } 1 - e^{-C_1n}.
\]

where:

\[
C_1 \sim O \left( \frac{(\Theta_1 - \delta_2s_0)^2}{kd\sigma^2(1 + \delta_2s_0)} \right), \quad \xi_\Delta = \frac{\sqrt{k\Delta}}{2}.
\]

Observe that fewer samples required for estimating \( \widehat{M}_n \) when:

- smaller RIP constant for true means \( M \) i.e. \( \delta_2s_0 \)
- thinner tails i.e. smaller \( \sigma^2 \)
Histogram aggregated data (bin-size $\Delta$)

Estimate the mean of histogram approx. distribution

\[
\hat{\beta} = \arg \min_{\beta} \|\beta\|_1 \text{ s.t. } \|\hat{M}_n\beta - \hat{\nu}_\epsilon\|_2 < \xi_{\Delta}.
\]

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

\[\|\beta^* - \hat{\beta}\| \leq O(\sqrt{k}\Delta) \text{ with probability at least } 1 - e^{-C_1n}.\]

where:

\[C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1 + \delta_{2s_0})}\right), \quad \xi_{\Delta} = \frac{\sqrt{k}\Delta}{2}.\]

- note: $\Delta$ affects accuracy, but not sample complexity
- exact recovery as $\Delta \to 0$
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

- vs. random \( M \) which achieves exact recovery
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

- vs. random $M$ which achieves exact recovery
Annual outpatient reimbursement (Louisiana, 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, 4th 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
  - data collected as histograms

- Application to healthcare
  - predictive modeling of CMS Medicare reimbursements
  - estimation of Texas state hospital charges

- Manuscript includes additional discussion:
  - higher order moments
Part 2: Learning a Linear model with Aggregated Spatio-temporal Data
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 annually
  - 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 $x(t) = [x_1(t), x_2(t) \cdots x_d(t)]$, 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 $x(t)$
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 $\varepsilon_\beta(t) = x(t)^\top \beta - y(t)$ 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
  \[
  \mathcal{L}(\beta) = E[|\varepsilon_\beta(t)|^2] = E[|x(t)^\top \beta - y(t)|^2]
  \]
- which is optimized as:
  \[
  \beta^* = \arg\min_\beta \mathcal{L}(\beta)
  \]
Residual process

- let \( \varepsilon_{\beta}(t) = \mathbf{x}(t)^{\top} \beta - y(t) \) 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

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

- which is optimized as:

\[
\beta^* = \arg\min_{\beta} \mathcal{L}(\beta)
\]
Data aggregation in time series

non-aggregated feature $X_1$
aggregated feature $\overline{X}_1$

non-aggregated feature $X_2$
aggregated feature $\overline{X}_2$

non-aggregated feature $X_3$
aggregated feature $\overline{X}_3$

non-aggregated target $Y$
aggregated target $\overline{Y}$
each coordinate of the feature set is aggregated

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

similarly, the targets are aggregated

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

for \( k, l \in \mathbb{Z} = \{ \cdots - 1, 0, 1, \cdots \} \).
Aggregation in time and frequency domain
Fourier space captures global properties of the signal

In time domain, convolution with square wave + sampling

\[ z(t) \xrightarrow{\text{convolution}} \text{Square function } u_T \xrightarrow{\text{sampling}} \overline{z}[k] \]

In frequency domain, multiplication with sinc function + sampling

\[ Z(\omega) \xrightarrow{\text{multiplication}} \text{Sinc function } U_T \xrightarrow{\text{sampling}} \overline{Z}(\omega) \]
Restricted Fourier transform

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

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

- 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 (Terzioglu, 1969)

- thus, most of the signal power between frequencies $(-\omega_0, \omega_0)$
Proposed Algorithm
Step 1

1. input parameters $T_0, \omega_0, D$, aggregated data samples $\bar{x}[k], y[l]$

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

   \[ \Omega = \{\omega_1, \omega_2, \ldots, \omega_D : \omega_i \in (-\omega_0, \omega_0)\} \]

3. for each $\omega \in \Omega$, compute $T_0$-restricted Fourier Transforms $\bar{X}_{T_0}(\omega), Y_{T_0}(\omega)$ from aggregated signals $\bar{x}[k], y[l]$
Step 1

1. input parameters $T_0, \omega_0, D$, aggregated data samples $\overline{x}[k], y[l]$.

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 $\overline{X}_{T_0}(\omega), Y_{T_0}(\omega)$ from aggregated signals $\overline{x}[k], y[l]$. 
Step 1

1. input parameters $T_0, \omega_0, D$, aggregated data samples $\overline{x}[k], y[l]$

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 $\overline{X}_{T_0}(\omega), Y_{T_0}(\omega)$ from aggregated signals $\overline{x}[k], y[l]$
Recall: $U_T$ is Fourier transform of square wave

- estimate non-aggregated Fourier transforms
  
  $$
  \hat{X}_{i,T_0}(\omega) = \frac{\hat{X}_{i,T_0}(\omega)}{U_{T_i}(\omega)}, \quad \hat{v}_{T_0}(\omega) = \frac{Y_{T_0}(\omega)}{U_T(\omega)}
  $$

- estimate parameter $\hat{\beta}$ as:

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

Recall: $U_T$ is Fourier transform of square wave

4. estimate non-aggregated Fourier transforms

$$\hat{X}_{i,T_0}(\omega) = \frac{\hat{X}_{i,T_0}(\omega)}{U_{Ti}(\omega)}, \quad \hat{v}_{T_0}(\omega) = \frac{\overline{Y}_{T_0}(\omega)}{U_T(\omega)}$$

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

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} E \left\| \hat{X}_{T_0}(\omega)^\top \beta - \hat{v}_{T_0}(\omega) \right\|^2$$
Generalization Analysis
Main result I

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

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

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

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 ⇒ Aliasing
- errors minimum for frequencies around 0

non-uniform sampling leads to further error
performance will depend on rapid decay of power spectral density
Theorem (Bhowmik, Ghosh, and Koyejo (2017))

Let \( \omega_i, \omega_y \) be the sampling rate for \( x_i(t), y(t) \) 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:

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

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, \( \omega_0 \) is small, minimum sampling frequency \( \omega_s \) 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 $\omega_0$, and increasing sampling and aggregation discrepancy
Synthetic Data - II

Fig 1(c): Medium discrepancy

- performance on synthetic data with varying $\omega_0$, and increasing sampling and aggregation discrepancy

Fig 1(d): High discrepancy
Las Rosas dataset

Regressing corn yield against nitrogen levels, topographical properties, brightness value, etc.
Regressing burned acreage against meteorological features, relative humidity, ISI index, etc. on UCI Forest Fires Dataset

![Estimation Error on Forest Fires Dataset](image)
Regressing atmospheric vapour 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!


sanmi@illinois.edu
References


David L Donoho. For most large underdetermined systems of linear equations the minimal $\ell_1$-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.


