Frequency Domain Predictive Modeling with Aggregated Data

Sanmi Koyejo

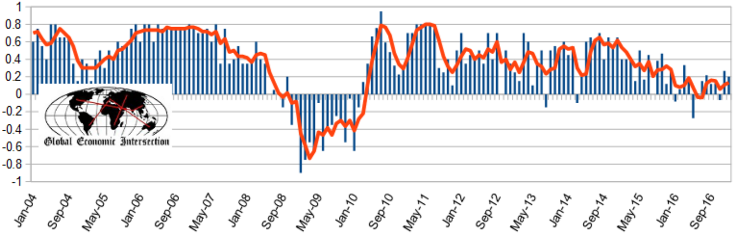
University of Illinois at Urbana-Champaign

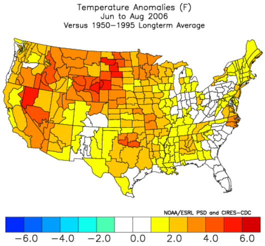
Joint work with

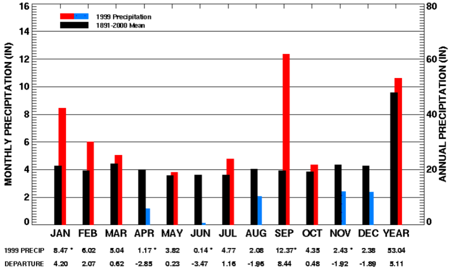
Avradeep Bhowmik

Joydeep Ghosh

@University of Texas at Austin







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

Motivation

- 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[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

Motivation - II

- Naive fitting of aggregated data may result in ecological fallacy (Freedman et al., 1991; Robinson, 2009)
- Reconstruction (before model fitting) is expensive and unreliable

Main Contribution

Model estimation procedure in the frequency domain

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

Motivation - II

- Naive fitting of aggregated data may result in ecological fallacy (Freedman et al., 1991; Robinson, 2009)
- Reconstruction (before model fitting) is expensive and unreliable

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

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

Problem Setup - II

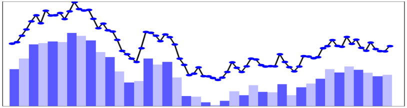
- Performance measure is the expected squared residual error

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

- which is optimized as:

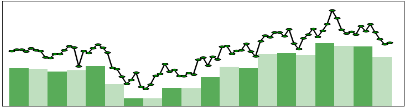
$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

Data Aggregation in Time Series



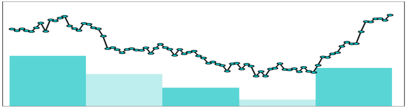
Non-Aggregated Feature $\mathbf{X}_1$

Aggregated Feature $\overline{\mathbf{X}}_1$



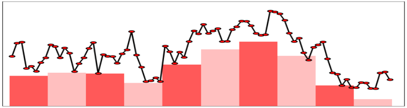
Non-Aggregated Feature $\mathbf{X}_2$

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

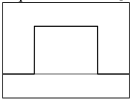
Aggregation in Time and Frequency Domain

Fourier captures global properties of the signal

In time domain, convolution with square wave + sampling

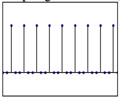
$z(t) \xrightarrow{\textit{convolution}}$

Square function u_T



$\xrightarrow{\textit{sampling}}$

Sampling Function δ_T

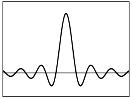


$\longrightarrow \quad \overline{z}[k]$

In frequency domain, multiplication with sinc function + sampling

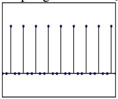
$Z(\omega) \xrightarrow{\textit{multiplication}}$

Sinc function U_T



$\xrightarrow{\textit{sampling}}$

Sampling Function δ_ω



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

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 (TerzioǦglu, 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

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

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

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

Step II

Recall: U_T is Fourier transform of square wave

4. Estimate non-aggregated Fourier transforms

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

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

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

Generalization Analysis

Main result I

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

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

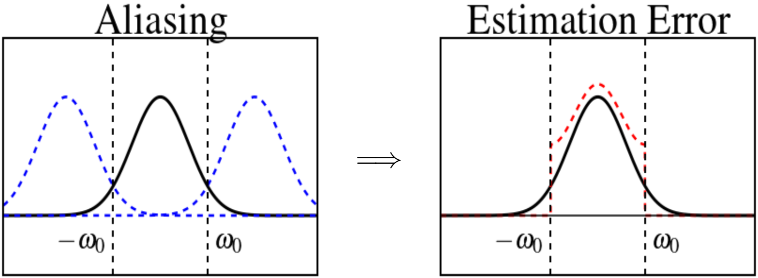
$$E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] < (1+\xi)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) + 2\xi$$

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

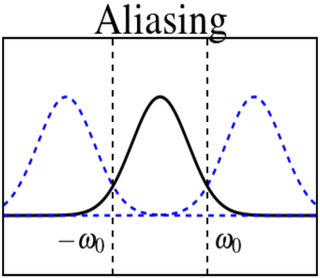
- Signals not bandlimited $\Rightarrow$ Aliasing
- Errors minimum for frequencies around 0



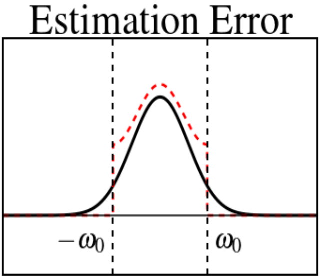
- Non-uniform sampling leads to further error
- Performance will depend on rapid decay of power spectral density

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:*

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

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

Empirical Evaluation

Synthetic Data

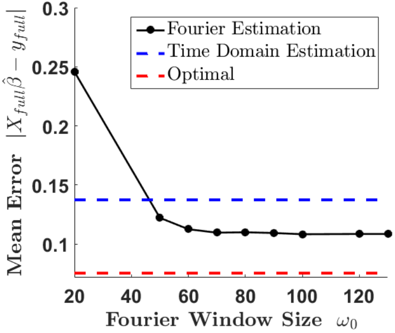


Fig 1(a): No Discrepancy

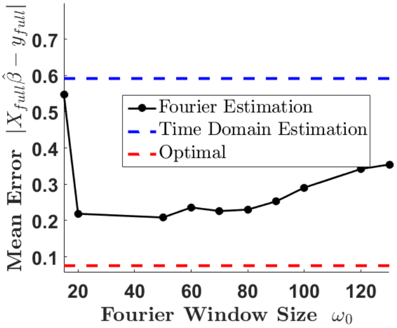


Fig 1(b): Low Discrepancy

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

Synthetic Data - II

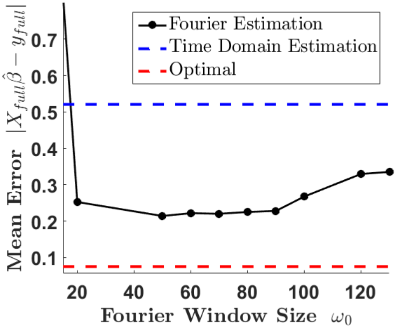


Fig 1(c): Medium Discrepancy

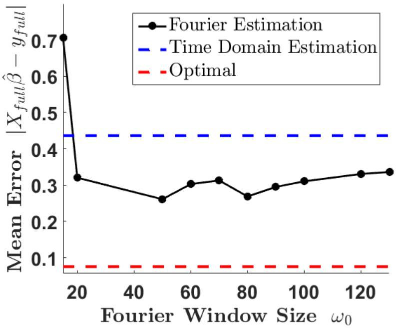
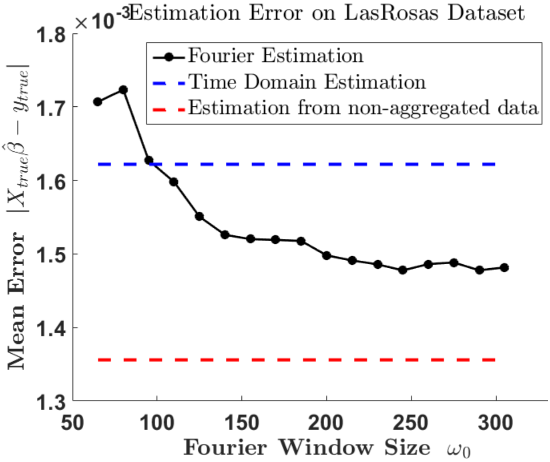


Fig 1(d): High Discrepancy

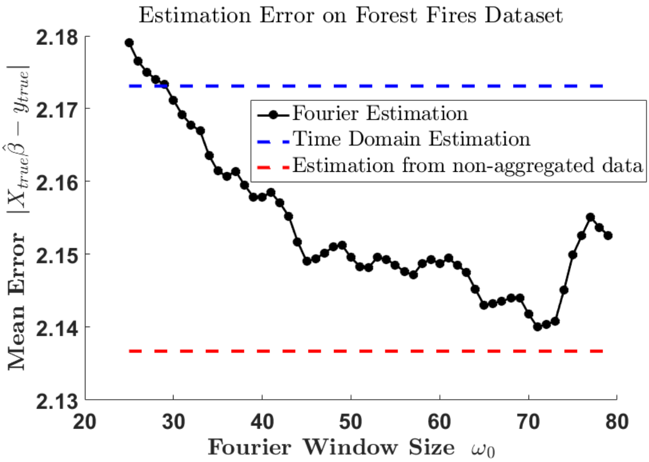
- Performance on synthetic data with varying ω_0, and increasing sampling and aggregation discrepancy

Las Rosas Dataset



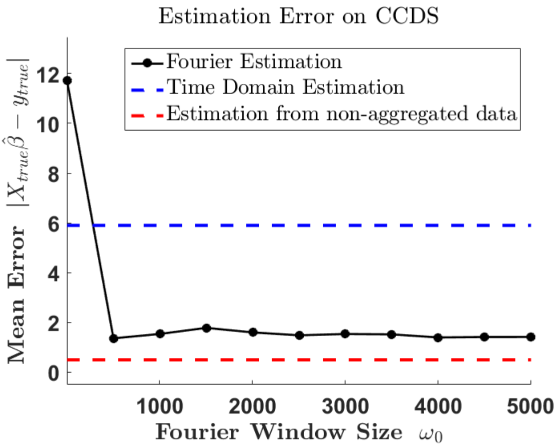
Regressing corn yield against nitrogen levels, topographical properties, brightness value, etc.

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

Conclusion

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.

Conclusion and Future Work

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

Future Work:

- Exploit other frequency domain structure e.g. sparse spectrum to improve estimates.
- Extensions to non-linear estimators

Conclusion and Future Work

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

Future Work:

- Exploit other frequency domain structure e.g. sparse spectrum to improve estimates.
- Extensions to non-linear estimators

Questions?

For more details:

Bhowmik, A., Ghosh, J. and Koyejo, O., 2017. *Frequency Domain Predictive Modeling with Aggregated Data*. In Proceedings of the 20th International conference on Artificial Intelligence and Statistics (AISTATS).

sanmi@illinois.edu

References

References I

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.

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 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.

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.

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

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