Tackling Data Synthesis Using a Multitaper Spectrum Estimation Technique

Joshua Pohlkamp-Hartt & David Riegert

Queen's University

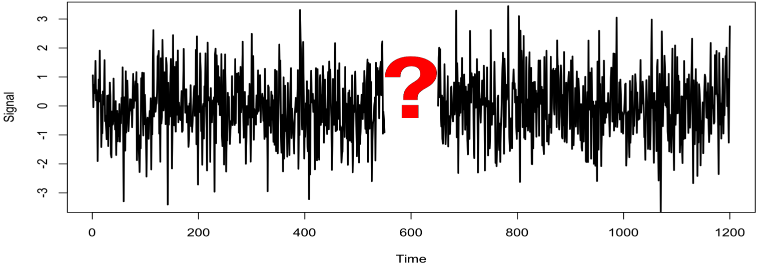
June 17, 2015

"Prediction is very difficult, especially if it's about the future."

— Niels Bohr

Motivating Example

Here we have 1200 samples from a set of 7 sinusoids in noise, with the middle 100 points missing. The signal-to-noise ratio for all sinusoids is below .30.



Overview

First demonstrated by Dr. David Thomson in 1990, under ideal conditions (Stationarity and White Noise), we can reconstruct the periodic components found within a time series.

The steps are:

1. Multitaper Spectrum Estimation (mean of multiple windowed Fourier transforms)
2. F-test & Complex Mean Values (regression in the frequency domain)
3. Inverse Fourier Transform of Complex Mean Values (transform model back to time domain)

The Multitaper Method (MTM)

- It is the primary tool for spectral estimation that balances the variance and bias of the estimated spectrum.

$$\overline{S}(f) = \frac{1}{K}\sum_{k=0}^{K-1} |Y_k(f)|^2, \tag{1}$$

$$Y_k(f) = \sum_{t=0}^{N-1} \upsilon_t^{(k)} e^{-2i\pi ft} x_t, \tag{2}$$

  where Y_k are the eigenspectra and $\upsilon_t^{(k)}$ are the Slepian sequences in the time domain.

- The Slepian sequences are defined for a choice of NW, with the parameters of NW and K being user selected.

F-test for line components

We model the eigenspectra found when performing the MTM by windowed sinusoids in the frequency domain at each frequency. We then perform an F-test to determine if each model is significant. Significant models indicate the presence of signals.

- We assume a model of

$$Y_k(f) = \mu(f)V_k(0) + \epsilon(f) \tag{3}$$

- Then we test $H_0 : \mu(f) = 0$ with the statistic,

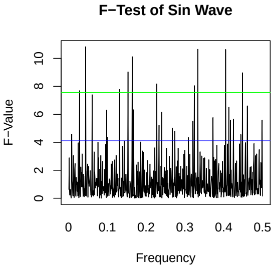
$$F(f) = (K-1)\frac{|\hat{\mu}(f)|^2 \sum_{k=0}^{K-1} |V_k(0)|^2}{\sum_{k=0}^{K-1} |(Y_k(f) - \hat{\mu}(f)V_k(0))^2}, \tag{4}$$

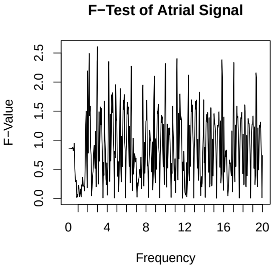
$$\hat{\mu}(f) = \frac{\sum_{k=0}^{K-1} V_k(0)Y_k(f)}{\sum_{k=0}^{K-1} |V_k(0)|^2}. \tag{5}$$

- The F-statistic will follow an $F(2, 2K-2, \alpha)$ distribution if H_0 is true.

Identifying Key Frequencies

- When we reject H_0 for a frequency at a given significance level, α, we conclude that the data has a sinusoidal component at that frequency.
- The complex mean value, $\hat{\mu}(f)$, associated with each significant frequency determines the magnitude of the sinusoidal component.
- Depending on α there is a change in the set of signals found in the data.





Inverse Fourier Transform

- After identifying the significant frequencies, we set the complex mean values of all non-significant frequencies to zero.

$$\hat{\mu}_\alpha(f) = \begin{cases} \hat{\mu}(f), & \hat{F}(f) > F_{(2nw-1,2,\alpha)}. \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

- We now can take the inverse fourier transform of the complex mean values to produce the periodic reconstruction of the time series.

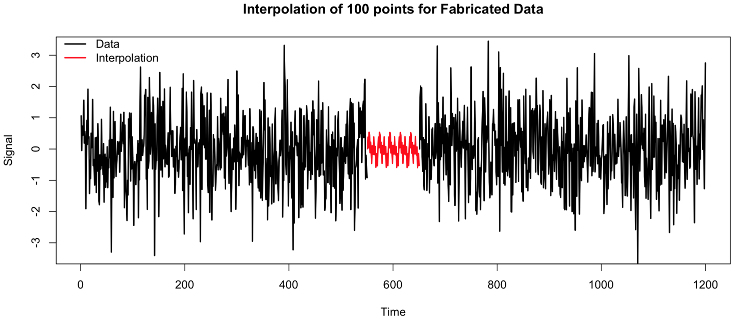
$$\hat{x}_t = \frac{\sum_{k=0}^{K-1} |V_k(0)|^2}{\sum_{k=0}^{K-1} V_k^\star(0)\nu_t^{(k)}} \mathcal{F}^{-1}(\hat{\mu})(t). \tag{7}$$

Application to Interpolation and Prediction

Thomson's periodic reconstruction method can be directly applied to the problems of interpolation and prediction.

- For interpolating a gap in the data, we simply fill the gap with a simple interpolation (linear or mean value works well) and perform Thomson's method. This will produce a periodic reconstruction across the gap. Be warned that the choice of starting interpolation is important and can affect the estimates.
- As for prediction, we zero-pad (add zeros after last data point) the windowed sets of data and perform Thomson's method. This will result in predictions for the points we added as zeros and has the added bonus of improving the frequency resolution.

Interpolation Example: Gap Filling



Potential Areas of Improvement

Within Thomson's method we have identified 3 areas that we feel can be addressed to improve the reconstruction of a time series.

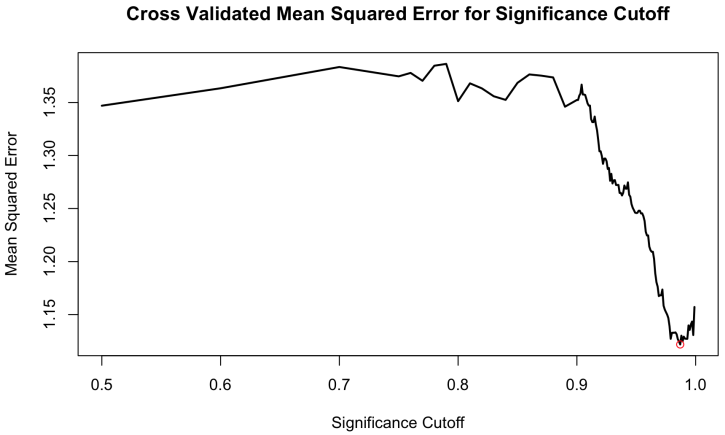
- Choice of significance level, α. (Cross Validation)
- Improve the estimate of the periodic reconstruction. (Gradient Boosting)
- Find confidence intervals for reconstruction. (Bootstrapping)

Choosing α

We would like an unsupervised method for identifying the best α value for interpolation or prediction. Using cross-validation as a framework for our decision making we should be able to find an optimal choice.

1. Split the data into bins.
2. Remove one bin and reconstruct the data for a set α.
3. Find the mean squared reconstruction error for the removed bin.
4. Perform steps $1-3$ across all bins and calculate the mean of the mean squared errors.
5. Repeat $1-4$ for all α values in our potential set.
6. Select the α with the minimum mean-mean squared error.

Interpolation Example: Cutoffs

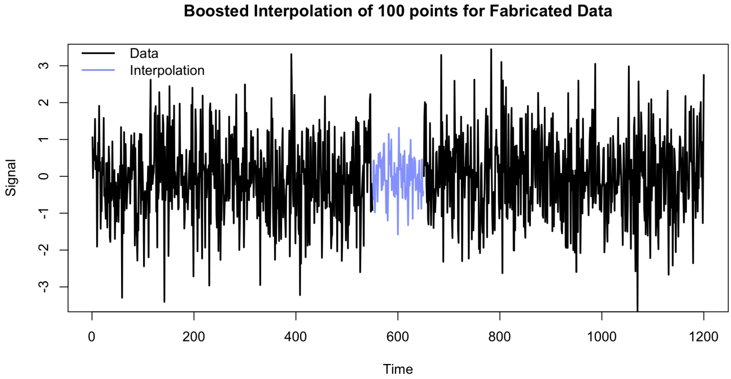


Improving Reconstruction

There may still be additional signals left over after the periodic reconstruction. We will use a gradient boosting approach on the residuals to determine if any more signals are present.

1. Treat the residuals, r_t, as a new time series and find the periodic reconstruction.
2. Now make a greedy model by finding $\underset{\gamma}{\operatorname{argmin}} \sum_{t \in T} (y_t - (\hat{y}_t + \gamma \hat{r}_t))^2$.
3. Define the new reconstruction as $\hat{y}_t' = \hat{y}_t + \gamma \hat{r}_t$ and test for significance with an F-test.
4. If the model is significant we repeat steps $1 - 3$ with the updated reconstruction's residuals, $y_t - \hat{y}_t'$, as our new series.
5. Continue repeating steps $1 - 3$ with the new residuals until the updated reconstruction is not considered significant under the F-test. Then consider the reconstruction from the previous iteration as the final boosted periodic reconstruction.

Interpolation Example: Boosting

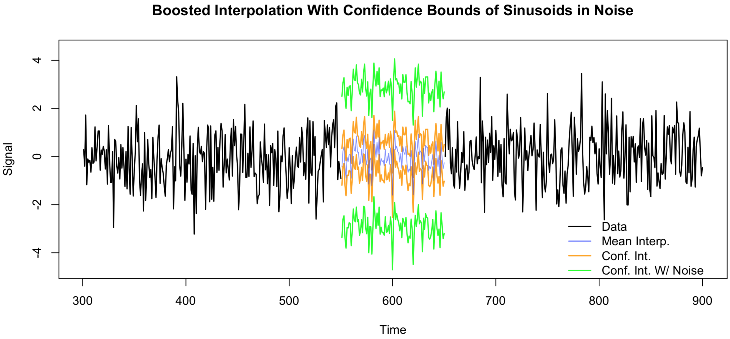


Confidence Intervals

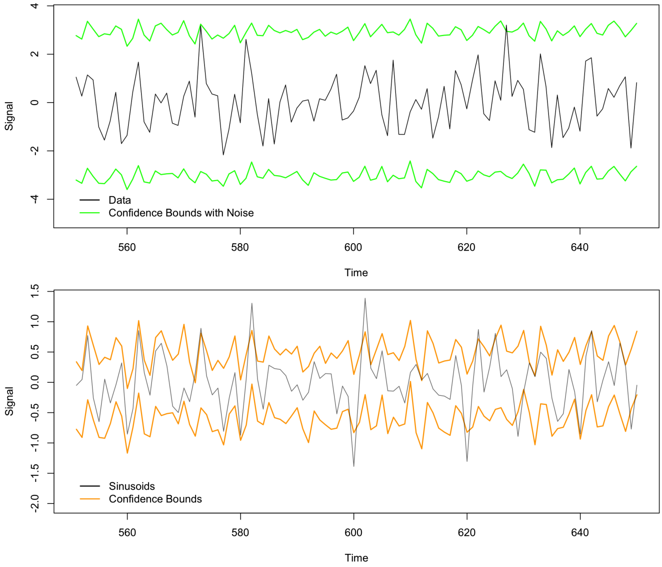
Confidence intervals on our reconstructions give us a better understanding of what we may expect the true could be. To determine the confidence intervals we employ a bootstrapping procedure.

1. After creating a periodic reconstruction, sample with replacement the residuals and create a new simple interpolation or zero-padding with the sampled residuals added to the previously used values.
2. Find a periodic reconstruction with the added residuals series.
3. Repeat steps 1 and 2, n times ($n > 60$).
4. The resulting set of reconstructions allows us to estimate the location and deviation at each point. This estimation can be parametric or not, depending on your assumptions.
5. We can also obtain estimates of the distribution of the noise by using the residuals from each periodic reconstructions. This will give us confidence intervals on the noise.
6. For overall confidence bounds we add these two bounds together.

Interpolation Example: Confidence Intervals



Interpolation Example: How did we do?



Acknowledgments

This research is supported by Queen's University and my supervisors Dr. Takahara and Dr. Thomson.

References

- D. Freedman, *Bootstrapping regression models*, The Annals of Statistics **9** (1981), no. 6, 1218–1228.
- S. Lahiri, *Resampling methods for dependent data*, vol. 14, Springer, 2003.
- D. Slepian and H.O. Pollack, *Prolate spheroidal wave functions, fourier analysis and uncertainty - I*, Bell System Technical Journal **40** (1961), no. 1, 43–64.
- R. Tibshirani T. Hastie and J. Friedman, *The elements of statistical learning*, vol. 1, Springer, 2001.
- D.J. Thomson, *Spectrum estimation and harmonic analysis*, Proceedings of the IEEE **70** (1982), no. 09, 1055–1096.
- D.J. Thomson, *Quadratic-inverse spectrum estimates: applications to paleoclimatology*, Philosophical Transactions: Physical Sciences and Engineering **332** (1990), no. 1627, 536–597.