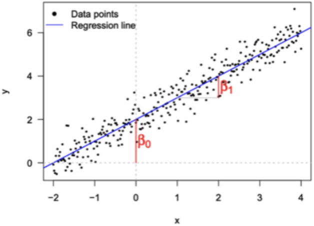
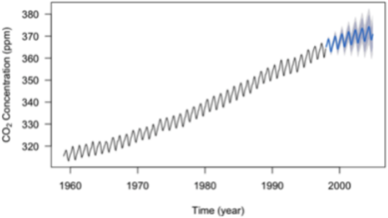
# Lecture 16

## Additional Topics in Regression and Time Series Analysis

*MATH 4070: Regression and Time-Series Analysis*




Whitney Huang
Clemson University

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

LASSO

# Moving Away From Linear Regression







- We have mainly focused on linear regression so far

  **Model:** $y = \boldsymbol{x}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$

  **Data:** $\boldsymbol{y}$ (response vector); $\boldsymbol{X}$ (design matrix)

  - $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$; $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}}_{\boldsymbol{H}\text{: “Hat” matrix}}\boldsymbol{y}$
  - $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$

- Non-parametric regression modeling

  **Model:** $y = f(\boldsymbol{x}) + \varepsilon \Rightarrow \mathrm{E}[y|\boldsymbol{x}] = f(\boldsymbol{x})$

  - The (smooth) function $f(\boldsymbol{x})$ must be represented somehow
  - The degree of smoothness of $f(\boldsymbol{x})$ must be made controllable
  - Some means for estimating the most appropriate degree of smoothness from data is required

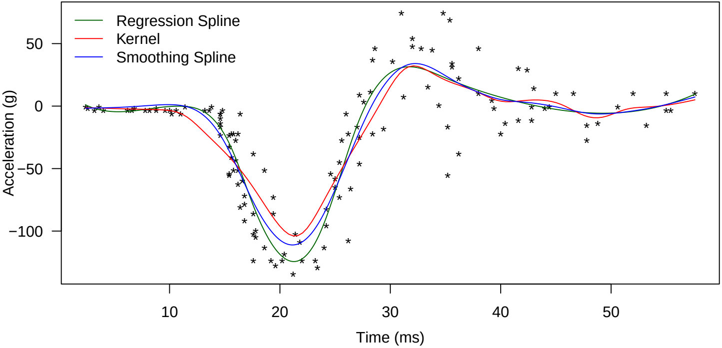
# Examples of Nonparametric Regression Fits



Regression Spline: 10 degrees of freedom quantile knot

Smoothing Spline: the amount of smoothness is estimated from the data by GCV

Kernel Regression: $K$: Epanechnikov kernel and $h = 5$



# Generalized Additive Models for Multiple Predictors

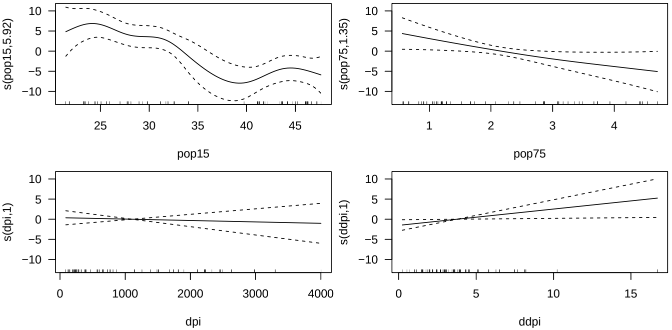
General non-parametric regression models

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

suffer from the "curse of dimensionality"

**Generalized Additive Models:**

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$$



# Shrinkage Methods





$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

$x_1, x_2, \cdots, x_{p-1}$ are the predictors.

**Question**: What if we have too many predictors (i.e., $p$ is "large")?

- Explanation can be difficult due to collinearity
- Can lead to overfitting by using too many predictors

Two methods, namely Ridge regression and LASSO, allow us to "shrink" the information contained in all the predictors into a more useful form

# Ridge Regression [Hoerl & Kennard, 1970]

Ridge regression assumes that the regression coefficients (after normalization) should not be very large

- The ridge regression estimate chooses the $\boldsymbol{\beta}$ that minimizes:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p-1}\beta_j^2,$$

  where $\lambda \geq 0$ is a **tuning parameter** to be determined via cross-validation

- The ridge regression estimates:

$$\hat{\beta}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

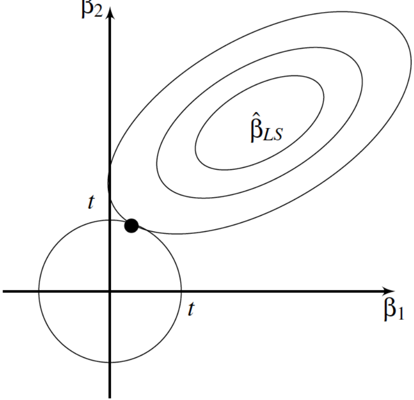
- Ridge regression is particularly effective when the model matrix is collinear

## Graphical Illustration of Ridge Regression

Estimation of ridge regression can also be solved by choosing $\beta$ to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t^2$




**Source:** p. 175, Fig. 11.9 *Linear Models with R*, Faraway, 2014

# Least Absolute Shrinkage and Selection Operator (LASSO)
## Tibshirani, 1996







LASSO assumes the effects are **sparse** in that the response can be explained by a small number of predictors with the rest having no effect

- LASSO choose $\hat{\beta}$ to minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p-1}|\beta_j|$$

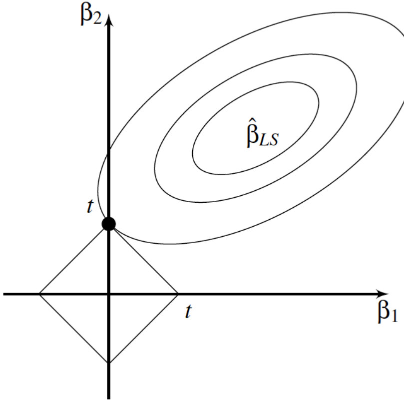
- No explicit solution to this minimization problem
- The penalty term has the effect of forcing some of the coefficient estimates to be zero when the tuning parameter $\lambda$ is "large" $\Rightarrow$ performs shrinkage and variable selection

# Graphical Illustration of LASSO

Estimation of LASSO can also be solved by choosing $\beta$ to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$




**Source:** p. 175, Fig. 11.9 *Linear Models with R*, Faraway, 2014

# Generalized Linear Model

- **Gaussian Linear Model:**

$$y \sim \mathrm{N}(\mu, \sigma^2), \quad \mu = \boldsymbol{x}^T \boldsymbol{\beta}$$

- **Bernoulli Linear Model:**

$$y \sim \mathrm{Bernoulli}(\pi), \quad \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^T \boldsymbol{\beta}$$

- **Poisson Linear Regression:**

$$y \sim \mathrm{Poisson}(\lambda), \quad \log \lambda = \boldsymbol{x}^T \boldsymbol{\beta}$$

These models fall into the family of generalized linear models [Nelder and Wedderburn (1972); McCullagh and Nelder (1989)] with the **distributional assumptions** (normal, Bernoulli, Poisson) and the **link functions** (identity, logit, and log)

# Frequency Domain Time Series Analysis







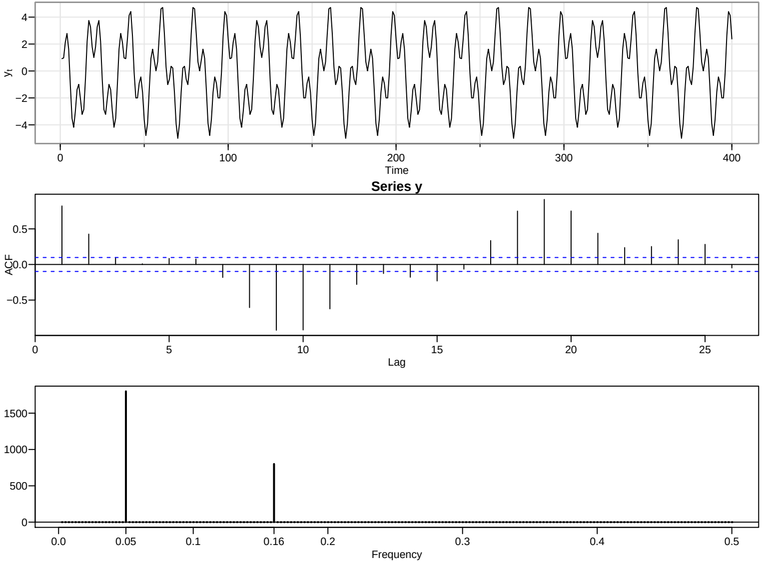
- Time domain methods [Box and Jenkins, 1970]:
  - Regress present on past

    **Example**: $Y_t = \phi Y_{t-1} + Z_t, \quad |\phi| < 1, \{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$
  - Capture dynamics in terms of "velocity", "acceleration", etc
- Frequency domain methods [Priestley, 1981]:
  - Regress present on periodic sines and cosines

    **Example**: $Y_t = \alpha_0 + \sum_{j=1}^{p} [\alpha_{1j} \cos(2\pi\omega_j t) + \alpha_{2j} \sin(2\pi\omega_j t)]$
  - Capture dynamics in terms of resonant frequencies

# Searching Hidden Periodicities

$$y_t = 3\cos\left(2\pi(\frac{10}{200})t\right) + 2\cos\left(2\pi(\frac{32}{200}t + 0.3)\right)$$



# Spectral Density and Covariance Functions

## Spectral density $\Longleftrightarrow$ Covariance function

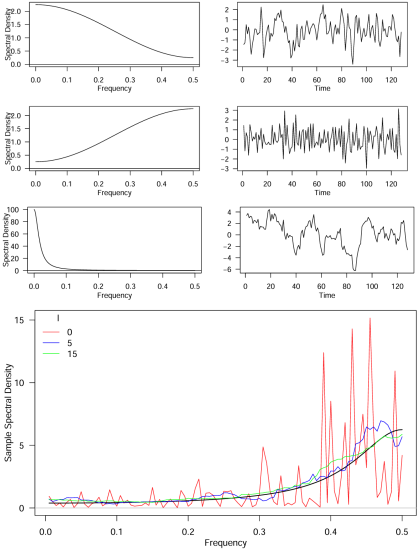
If $\{Y_t\}$ has $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, then its spectral density is

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}$$

for $-\infty < \omega < \infty$. We have

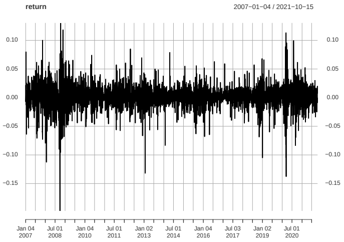
$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega)\, d\omega$$

Smoothing techniques, like those in **nonparametric regression**, are needed to estimate $f(\omega)$ well



# GARCH Models for Volatility

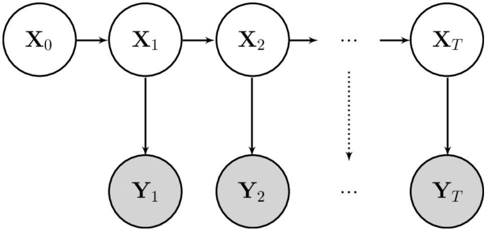
Log-returns, $r_t = \log(\frac{y_t}{y_{t-1}})$, are often modeled instead of daily stock prices, $y_t$, in financial time series analysis


Generalized Autoregressive Conditional Heteroskedasticity (GARCH) is commonly used to model the dynamics of fluctuations in log-returns to capture volatility clustering.

$$r_t = \mu_t + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$$

## State-Space Model




State: $\boldsymbol{X}_t = M_t \boldsymbol{X}_{t-1} + \boldsymbol{V}_t, \quad \boldsymbol{V}_t \overset{i.i.d.}{\sim} \mathrm{WN}(\mathbf{0}, Q_t), \quad t = 1, 2, \cdots$

Observation: $\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t, \quad \boldsymbol{W}_t \overset{i.i.d.}{\sim} \mathrm{WN}(\mathbf{0}, R_t), \quad t = 1, 2, \cdots$

- $\boldsymbol{X_t} \in \mathbb{R}^p$ and $\boldsymbol{Y_t} \in \mathbb{R}^q$ are the state vector and the observation vector at time $t$
- $\boldsymbol{M}_t$ is the $p \times p$ transition matrix, and $\boldsymbol{H}_t$ is the $q \times p$ observation matrix
- $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ are the state and observation noises

# Forecasting, Filtering, and Smoothing

**Goal**: To estimate the underlying unobserved signal $\boldsymbol{X}_t$, given the data $\boldsymbol{Y}_{1:s} = \boldsymbol{y}_{1:s} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_s\}$:

- When $s < t$, the problem is called forecasting or prediction
- When $s = t$, the problem is called filtering
- When $s > t$, the problem is called smoothing

In addition to these estimates, we would also want to measure their precision. The solution to these problems is accomplished via the Kalman filter and Kalman smoother

# The Kalman Filter: General Results



Assume the filtering distribution at time $t-1$ is

$$[\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{1:t-1}] \sim \mathrm{N}\left(\boldsymbol{\mu}^a_{t-1}, \Sigma^a_{t-1}\right)$$

- **Forecast Step:** Gives the forecast distribution at time $t$:

  $$[\boldsymbol{X}_t|\boldsymbol{Y}_{1:t-1}] \sim \mathrm{N}\left(\boldsymbol{\mu}^f_t, \Sigma^f_t\right),$$

  where $\boldsymbol{\mu}^f_t = M_t\boldsymbol{\mu}^a_{t-1}$, and $\Sigma^f_t = M_t\Sigma^a_{t-1}M^T_t + Q_t$.
- **Update Step:** updates the forecast distribution using new data $\boldsymbol{Y}_t$

  $$[\boldsymbol{X}_t|\boldsymbol{Y}_{1:t}] \sim \mathrm{N}\left(\boldsymbol{\mu}^a_t, \Sigma^a_t\right),$$

  where $\boldsymbol{\mu}^a_t = \boldsymbol{\mu}^f_t + K_t\left(\boldsymbol{Y}_t - H_t\boldsymbol{\mu}^f_t\right)$, and $\Sigma^a_t = \left(I - K_tH^T_t\right)\Sigma^f_t$, and

  $$K_t = \Sigma^f_t H^T_t\left(H_t\Sigma^f_t H^T_t + R_t\right)^{-1}$$

  is the Kalman gain matrix

# Multivariate Time Series Analysis

All the methods presented for univariate time series also apply to multivariate processes

$$\{\boldsymbol{Y}_t \in \mathbb{R}^p\}$$

- The theory becomes more involved as we generalize to the cross-covariance:

  $$\mathrm{Cov}(\boldsymbol{Y}_s, \boldsymbol{Y}_t) = \boldsymbol{C}(s, t),$$

  where $\boldsymbol{C}(\cdot,\cdot)$ is the $p \times p$ matrix-valued cross-covariance function (CCVF)

- Similarly, in the frequency domain approach, the cross-spectrum is given by:

  $$f_{XY}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{XY}(h) e^{-2\pi i \omega h}$$

# Vector Autoregressive (VAR) Models



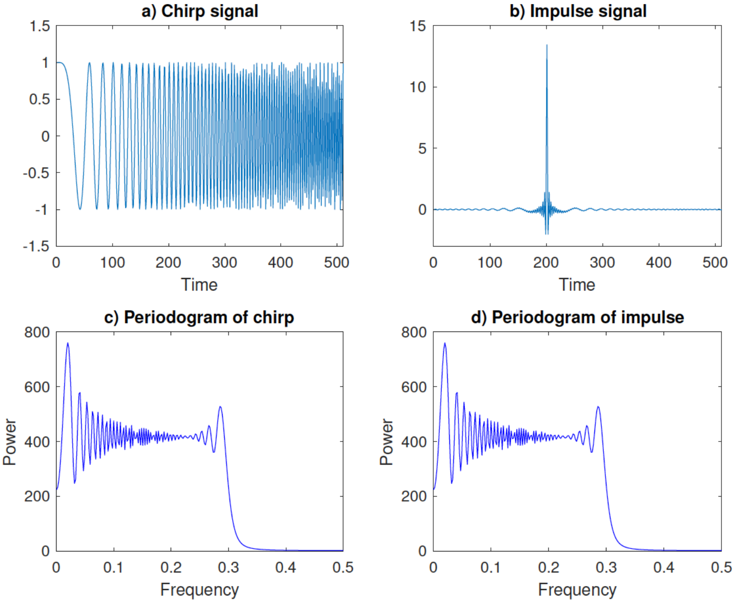
VAR($p$) model:

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + A_1 \boldsymbol{Y}_{t-1} + \cdots + A_p \boldsymbol{Y}_{t-p} + \boldsymbol{W}_t, \quad t = 0, 1, 2, \cdots,$$

where

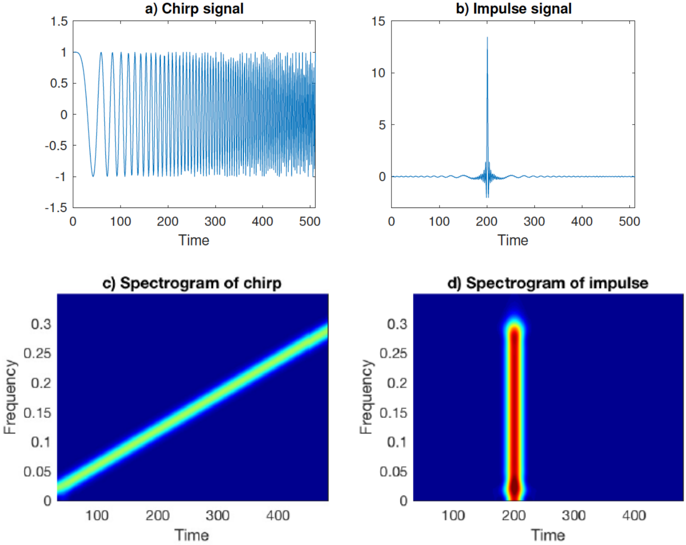
- $\boldsymbol{Y}_t = (Y_{1t}, \cdots, Y_{pt})^T$ is a $(p \times 1)$ random vector
- $A_i$ are $(p \times p)$ coefficient matrices
- $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^T$ is the intercept vector
- $\boldsymbol{W}_t = (W_{1t}, \cdots, W_{pt})^T$ is a p-dimensional white noise, i.e., $\mathrm{E}[\boldsymbol{W}_t] = \mathbf{0}$, $\mathrm{E}[\boldsymbol{W}_t \boldsymbol{W}_t^T] = \Sigma_{\boldsymbol{W}}$ and $\mathrm{E}[\boldsymbol{W}_s \boldsymbol{W}_t^T] = \mathbf{0}$ for $s \neq t$.

# Time-Frequency Analysis: A Motivation Example



# Time-Frequency Analysis: Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time



# Non-Gaussian Time Series Methods







Some selected references:

- Regression models for time series analysis, Kedem and Fokianos, 2002
- Handbook of discrete-valued time series, edited by Davis, Holan, Lund, Ravishanker, 2016
- Bayesian Dynamic Generalized Linear Models, Gamerman *et. al*, 2016
- Count Time Series: A Methodological Review, Davis *et. al.*, 2021