Data Analysis: Statistical Modeling and Computation in Applications

Time Series Analysis: Model Selection, Linear Processes, LLR

Overview

- Determining model order for AR models
  - Partial Autocorrelation
  - Akaike Information Criterion
  - Cross-validation
- Linear Processes
- Subseasonal weather forecasting and Local linear regression

Recap: Time series models

- **White Noise:** $X_t = W_t$;
  W_t independent, mean zero, same variance σ_w^2
- **Autoregressive AR(p):**
  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + W_t$
- **Moving Average MA(q):**
  $X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q}$
- **ARMA / ARIMA:**
  $X_t = \phi_1 X_{t-1} + \ldots \phi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q}$
  (ARIMA: ARMA after differencing)

Fitting a time series: Overview (Part I)

1. transform to make it stationary
   - log-transform
   - remove trends / seasonality
   - differentiate successively
2. check for white noise (ACF)
3. if stationary: plot autocorrelation. If finite lag, fit MA (ACF gives order), otherwise AR.

Fitting AR(p):

1. compute PACF *to get order*
2. estimate coefficients ϕ_k and noise variance σ_w^2 via Yule-Walker equations
3. compute residuals, test for white noise

Model order for AR(p)

- MA(q) model: autocorrelation reveals order q

Model order for AR(p)

- MA(q) model: autocorrelation reveals order q
- But not for AR(p) model!

Model order for AR(p)

- MA(q) model: autocorrelation reveals order q
- But not for AR(p) model!
- **Example** AR(1): $X_t = \phi X_{t-1} + W_t = \phi^2 X_{t-2} + \phi W_{t-1} + W_t$ ✓

$$\mathrm{corr}(X_t, X_{t-2}) =$$

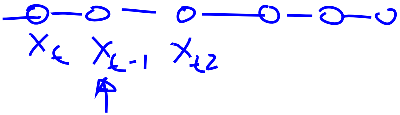
$$\mathrm{corr}(\phi^2 X_{t-2} + \phi W_{t-1} + W_t, X_{t-2})$$

$$= \phi^2 \gamma(0) > 0$$

Model order for AR(p)

- MA(q) model: autocorrelation reveals order q
- But not for AR(p) model!
- **Example** AR(1): $X_t = \phi X_{t-1} + W_t$

  $\mathrm{corr}(X_t, X_{t-2}) = \phi^2 \gamma(0)$
- Correlation via X_{t-1}.



Model order for AR(p)

- MA(q) model: autocorrelation reveals order q
- But not for AR(p) model!
- **Example** AR(1): $X_t = \phi X_{t-1} + W_t$

  $\text{corr}(X_t, X_{t-2}) = \phi^2\gamma(0)$
- Correlation via X_{t-1}.
- Idea: if we *remove linear effect of X_{t-1}*, then X_t, X_{t-2} should be uncorrelated.

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
    - regress X on Z; Y on Z

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \mathrm{corr}(X - \hat{X}, Y - \hat{Y})$.

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \mathrm{corr}(X - \hat{X}, Y - \hat{Y})$.
- for time series: **Partial Autocorrelation** of X_t, X_{t-h}.
  Remove the effect of all variables "in between", $X_{t-1}, \ldots X_{t-h+1}$.

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \text{corr}(X - \hat{X}, Y - \hat{Y})$.
- for time series: **Partial Autocorrelation** of X_t, X_{t-h}.
  Remove the effect of all variables "in between", $X_{t-1}, \ldots X_{t-h+1}$.
- equivalent computation:

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \mathrm{corr}(X - \hat{X}, Y - \hat{Y})$.

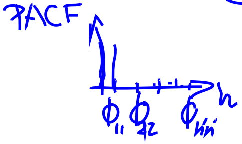
- for time series: **Partial Autocorrelation** of X_t, X_{t-h}.
  Remove the effect of all variables "in between", $X_{t-1}, \ldots X_{t-h+1}$.

- equivalent computation:
  - fit AR(h) model to obtain $\hat{\phi}_{h1}, \ldots \hat{\phi}_{hh}$

$\hat{X}_t = \underbrace{\hat{\phi}_{h1} X_{t-1} \cdots}\ + \hat{\phi}_{hh} X_{t-h}$

Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \text{corr}(X - \hat{X}, Y - \hat{Y})$.
- for time series: **Partial Autocorrelation** of X_t, X_{t-h}.
  Remove the effect of all variables "in between", $X_{t-1}, \ldots X_{t-h+1}$.
- equivalent computation:
  - fit AR(h) model to obtain $\hat{\phi}_{h1}, \ldots \hat{\phi}_{hh}$
  - (estimated) partial autocorrelation is last coefficient $\hat{\phi}_{hh}$



Partial Autocorrelation (PACF)

- **Partial Correlation** of X, Y given Z:
  - regress X on Z; Y on Z
  - $\rho_{XY|Z} = \text{corr}(X - \hat{X}, Y - \hat{Y})$.
- for time series: **Partial Autocorrelation** of X_t, X_{t-h}.
  Remove the effect of all variables "in between", $X_{t-1}, \ldots X_{t-h+1}$.
- equivalent computation:
  - fit AR(h) model to obtain $\hat{\phi}_{h1}, \ldots \hat{\phi}_{hh}$
  - (estimated) partial autocorrelation is last coefficient $\hat{\phi}_{hh}$
- for AR(p): $\phi_{hh} = 0$ for $h > p$.

Autocorrelation (ACF) and PACF for AR(2)

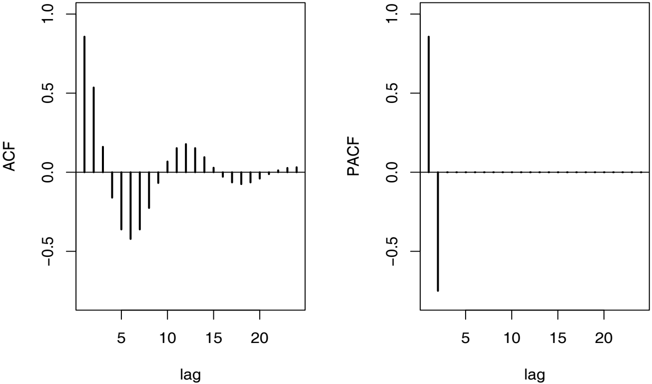
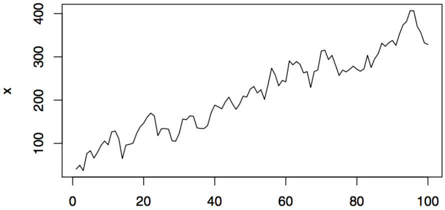


Fig. 3.4. The ACF and PACF of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

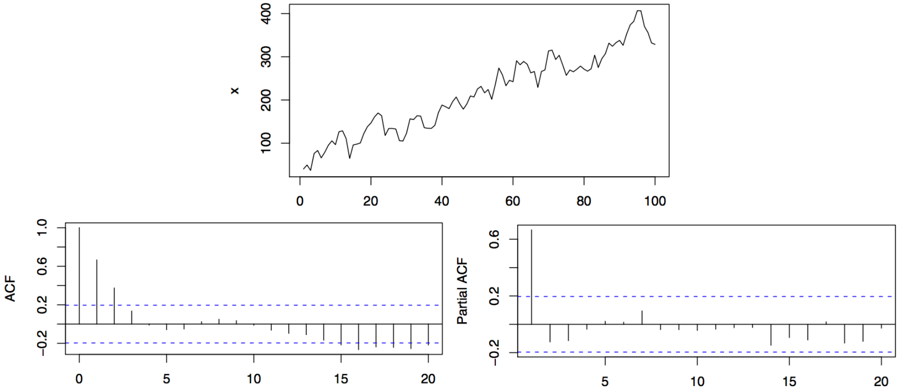
Partial ACF as a diagnostic tool: example from last lecture

Example: $X_t = T_t + Y_t$, sum of
linear trend ($T_t = 50 + 3t$) and AR(1) ($Y_t = 0.8Y_{t-1} + W_t$, $\sigma_W = 20$).



Partial ACF as a diagnostic tool: example from last lecture

Example: $X_t = T_t + Y_t$, sum of
linear trend ($T_t = 50 + 3t$) and AR(1) ($Y_t = 0.8Y_{t-1} + W_t$, $\sigma_W = 20$).



top: series; bottom: autocorrelation and partial autocorrelation of residuals after fitting only a linear model.

Overview: ACF, PACF and model order

	ACF	PACF
AR(p)	decays	zero for $h > p$
MA(q)	zero for $h > q$	decays
ARMA(p,q)	decays	decays

Overview

- Determining model order for AR models
  - Partial Autocorrelation
  - Akaike Information Criterion
  - Cross-validation
- Linear Processes
- Subseasonal weather forecasting and Local linear regression

Akaike Information Criterion (AIC)

- Main idea: tradeoff between data fit and model complexity.

Akaike Information Criterion (AIC)

- Main idea: tradeoff between data fit and model complexity.
- For a model with k parameters:

$$AIC(k) = \underbrace{-2\text{log-likelihood}}_{\text{model fit}} + \underbrace{2k}_{\text{complexity}}$$

Cross-validation for time series

Cross-validation for time series

- "usual" cross-validation: randomly partition into k folds.
  Repeatedly train on $k-1$ folds, test on remaining.

Cross-validation for time series

- "usual" cross-validation: randomly partition into k folds. Repeatedly train on $k-1$ folds, test on remaining.
- cross-validation works in special cases for AR models *(Bergmeier, Hyndman, Koo 2015)*

Cross-validation for time series

- "usual" cross-validation: randomly partition into k folds. Repeatedly train on $k-1$ folds, test on remaining.
- cross-validation works in special cases for AR models *(Bergmeier, Hyndman, Koo 2015)*
- evaluation on a rolling forecasting origin

Overview

- Determining model order for AR models
  - Partial Autocorrelation
  - Akaike Information Criterion
  - Cross-validation
- Linear Processes
- Subseasonal weather forecasting and Local linear regression

Relating MA and AR?

- **Autoregressive AR(p):**
  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots \phi_p X_{t-p} + W_t$

- **Moving Average MA(q):**
  $X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \dots + \theta_q W_{t-q}$

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

- $\mathbb{E}[X_t] = 0$

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(t, t+h) = \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \sigma_w^2 = \gamma_X(h)$

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(t, t+h) = \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \sigma_w^2 = \gamma_X(h)$
  Linear Process is stationary!

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(t, t+h) = \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \sigma_w^2 = \gamma_X(h)$
  Linear Process is stationary!

- MA(q) is linear process, causal

Relating MA and AR: Linear Process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \qquad \sum_j |\psi_j| < \infty$$

- **causal:** $\psi_j = 0$ whenever $j < 0$
  (function of the past)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(t, t+h) = \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \sigma_w^2 = \gamma_X(h)$
  Linear Process is stationary!

- MA(q) is linear process, causal
- *What about AR?*

AR(1) as Linear Process

$$X_t = W_t + \phi_1 X_{t-1}$$

AR(1) as Linear Process

$$
\begin{aligned}
X_t &= W_t + \phi_1 X_{t-1} \\
&= \sum_{j=0}^{\infty} \phi_1^j W_{t-j}
\end{aligned}
$$

AR(1) as Linear Process

$$\begin{aligned} X_t &= W_t + \phi_1 X_{t-1} \\ &= \sum_{j=0}^{\infty} \phi_1^j W_{t-j} \end{aligned}$$

- converges if $|\phi_1| < 1$.
  Then: AR(1) is causal and stationary.

AR(1) as Linear Process

$$
\begin{aligned}
X_t &= W_t + \phi_1 X_{t-1} \\
&= \sum_{j=0}^{\infty} \phi_1^j W_{t-j}
\end{aligned}
$$

- converges if $|\phi_1| < 1$.
  Then: AR(1) is causal and stationary.
- general result: AR(p) is stationary & causal if linear process converges.

AR(1) as Linear Process

$$
\begin{aligned}
X_t &= W_t + \phi_1 X_{t-1} \\
&= \sum_{j=0}^{\infty} \phi_1^j W_{t-j}
\end{aligned}
$$

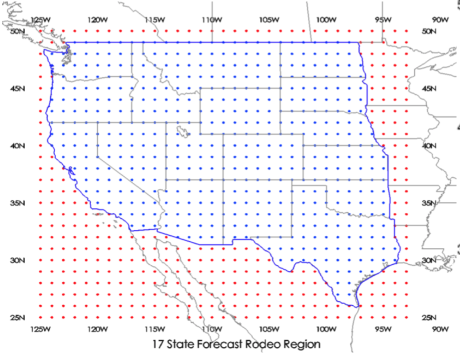
- converges if $|\phi_1| < 1$.
  Then: AR(1) is causal and stationary.
- general result: AR(p) is stationary & causal if linear process converges.
- *Similarly:* can write MA(1) as an infinite AR process (under similar convergence conditions) in the form $W_t = X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots$. If converges: *invertible*.

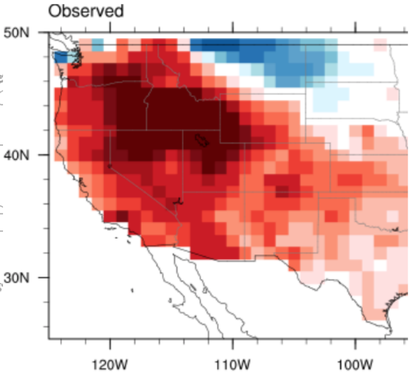
Overview

- Determining model order for AR models
  - Partial Autocorrelation
  - Akaike Information Criterion
  - Cross-validation
- Linear Processes
- Subseasonal weather forecasting and Local linear regression

J. Hwang, P. Orenstein, K. Pfeiffer, J. Cohen, L. Mackey. Improving Subseasonal Forecasting in the Western U.S. with Machine Learning. KDD 2019.

Subseasonal Rodeo





images: Lester Mackey

Subseasonal Rodeo: measurements

multiple time series $Y_{t,g}$, indexed by grid points g

- temperature
- precipitation
- sea surface temperature & sea ice concentration
- ENSO index (pressure, wind, SST, temperature, cloudiness)
- Madden-Julian oscillation
- relative humidity / pressure
- North Americal Model Ensemble

Subseasonal Rodeo: measurements

multiple time series $Y_{t,g}$, indexed by grid points g

- temperature
- precipitation
- sea surface temperature & sea ice concentration
- ENSO index (pressure, wind, SST, temperature, cloudiness)
- Madden-Julian oscillation
- relative humidity / pressure
- North Americal Model Ensemble

- use **anomalies**

$$a_t = y_t - c_{\text{monthday(t)}}$$

- **accuracy measure**:

$$\text{skill}(\hat{a}_t, a_t) = \cos(\hat{a}_t, a_t) = \frac{\langle \hat{a}_t, a_t \rangle}{\|\hat{a}_t\| \|a_t\|}$$

Questions

- What model should we fit?
- Which variables should we use for prediction?

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure
- locally well approximated by a linear function
- **Idea**: To predict at $\mathbf{x}_0$, locally fit a linear model around $\mathbf{x}_0$

$$\mathcal{D} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq h\}$$

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure
- locally well approximated by a linear function
- **Idea**: To predict at $\mathbf{x}_0$, locally fit a linear model around $\mathbf{x}_0$

$$\mathcal{D} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq h\}$$

- Obtain one $\boldsymbol{\beta}$ for each $\mathbf{x}_0$:

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}_0} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathbf{x}_i \in \mathcal{D}} w_i (y_i - \boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_0))^2$$

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure
- locally well approximated by a linear function
- **Idea**: To predict at $\mathbf{x}_0$, locally fit a linear model around $\mathbf{x}_0$

$$\mathcal{D} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq h\}$$

- Obtain one $\boldsymbol{\beta}$ for each $\mathbf{x}_0$:

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}_0} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathbf{x}_i \in \mathcal{D}} w_i (y_i - \boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_0))^2$$

- larger h: smoother function

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure
- locally well approximated by a linear function
- **Idea**: To predict at $\mathbf{x}_0$, locally fit a linear model around $\mathbf{x}_0$

$$\mathcal{D} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq h\}$$

- Obtain one $\boldsymbol{\beta}$ for each $\mathbf{x}_0$:

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}_0} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathbf{x}_i \in \mathcal{D}} w_i (y_i - \boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_0))^2$$

- larger h: smoother function
- weighting: can use a *kernel* $w_i = K(\frac{\mathbf{x}_i - \mathbf{x}_0}{h})$

Local Linear Regression (no time series)

- data from a nonlinear function, but we don't know exact structure
- locally well approximated by a linear function
- **Idea**: To predict at $\mathbf{x}_0$, locally fit a linear model around $\mathbf{x}_0$

$$\mathcal{D} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq h\}$$

- Obtain one $\boldsymbol{\beta}$ for each $\mathbf{x}_0$:

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}_0} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathbf{x}_i \in \mathcal{D}} w_i (y_i - \boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_0))^2$$

- larger h: smoother function
- weighting: can use a *kernel* $w_i = K(\frac{\mathbf{x}_i - \mathbf{x}_0}{h})$
  ($K(u) = 0$ for $|u| > 1$, $K(u) = K(-u)$, $K(u) > 0$ for $|u| < 1$)

Local Linear Regression for Time Series

- one model for each day of the year (shared across years) and grid point g.
- $\mathcal{D} = \pm 56$-day span around target day of the year

Local Linear Regression for Time Series

- one model for each day of the year (shared across years) and grid point g.
- $\mathcal{D} = \pm 56$-day span around target day of the year
- for each day of the year:

$$\hat{\boldsymbol{\beta}}_g = \arg\min_{\boldsymbol{\beta}} \sum_{t \in \mathcal{D}} w_{t,g} (y_{t,g} - b_{t,g} - \boldsymbol{\beta}^\top \mathbf{x}_{t,g})^2$$

Local Linear Regression for Time Series

- one model for each day of the year (shared across years) and grid point g.
- $\mathcal{D} = \pm 56$-day span around target day of the year
- for each day of the year:

$$\hat{\boldsymbol{\beta}}_g = \arg\min_{\boldsymbol{\beta}} \sum_{t \in \mathcal{D}} w_{t,g}(y_{t,g} - b_{t,g} - \boldsymbol{\beta}^\top \mathbf{x}_{t,g})^2$$

- weighting: e.g. $w_{t,g} = 1$ or $w_{t,g} = 1/\operatorname{var}(a_t)$

Local Linear Regression for Time Series

- one model for each day of the year (shared across years) and grid point g.
- $\mathcal{D} = \pm 56$-day span around target day of the year
- for each day of the year:

$$\hat{\boldsymbol{\beta}}_g = \arg\min_{\boldsymbol{\beta}} \sum_{t \in \mathcal{D}} w_{t,g}(y_{t,g} - b_{t,g} - \boldsymbol{\beta}^\top \mathbf{x}_{t,g})^2$$

- weighting: e.g. $w_{t,g} = 1$ or $w_{t,g} = 1/\operatorname{var}(a_t)$
- offsets: $b_{t,g} = 0$ or $b_{t,g} = c_{\text{monthday}(t)}$

Model 1: external regressors with multitask feature selection

- **Candidate features:** measurements from each data source at different lags, and constant feature

Model 1: external regressors with multitask feature selection

- **Candidate features:** measurements from each data source at different lags, and constant feature
- allow different features for different days of the year
- but same features across all grid points for a given day

Model 1: external regressors with multitask feature selection

- **Candidate features:** measurements from each data source at different lags, and constant feature
- allow different features for different days of the year
- but same features across all grid points for a given day

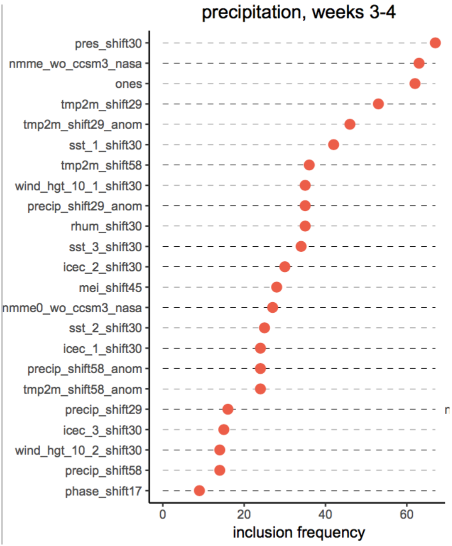
- **Backward regression:** Start with all features, prune features one by one.

Model 1: external regressors with multitask feature selection

- **Candidate features:** measurements from each data source at different lags, and constant feature
- allow different features for different days of the year
- but same features across all grid points for a given day

- **Backward regression:** Start with all features, prune features one by one.
- Drop feature that reduces the prediction accuracy the least, if below a tolerance threshold
- Accuracy: Leave-one-year out cross-validation with "skill"

Selected features (example)



Selected features (example)

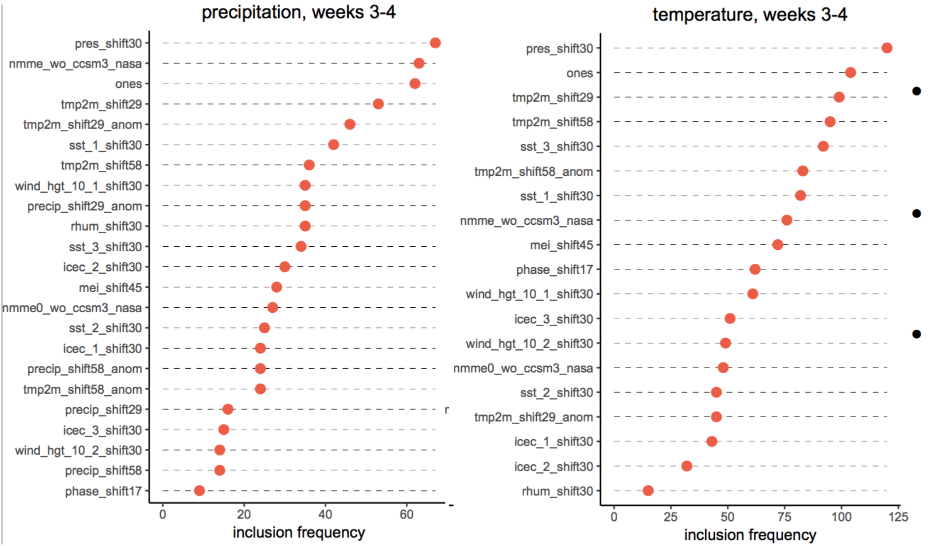


image source: (Hwang et al 2019)

Summary

- Determining model order for AR models:
  - Partial Autocorrelation
  - Akaike Information Criterion
  - Cross-validation
- Relating AR and MA modesl: Linear Processes
- Nonlinear model: Local linear regression and application example

References and Further Reading

- Paul S.P. Cowpertwait and Andrew V. Metcalfe. Introductory Time Series with R. Springer, 2009.
  Chapter 4.5, 4.6
- R.H. Shumway, D.S. Stoffer. Time Series Analysis and its Applications, with Examples in R. Springer, 2011.
  Chapter 3.1, 3.3.

- R. Carmona. Statistical Analysis of Financial Data in R. Springer, 2014.
  Chapter 6.
- J. Hwang, P. Orenstein, K. Pfeiffer, J. Cohen, L. Mackey. Improving Subseasonal Forecasting in the Western U.S. with Machine Learning. KDD 2019.