# Lecture 12
# Dynamic (AR) Spatio-Temporal Models

Dennis Sun
Stats 253

July 30, 2014

# Outline of Lecture

1. The Model
2. Application to the Wind Data
3. Computational Issues
4. Extensions
5. Wrapping Up

# Where are we?

# Last Lecture

- Last lecture, we discussed kriging for spatio-temporal data, which is not very different from kriging in three dimensions.
- It is difficult to construct models that don't have full symmetry:

$$C(\boldsymbol{r}, h) = C(\boldsymbol{r}, -h)$$

# Last Lecture

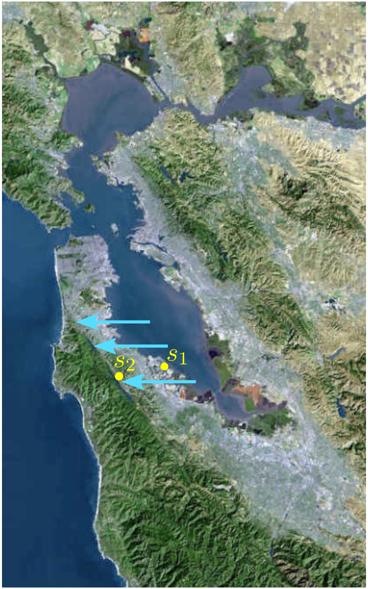
- Last lecture, we discussed kriging for spatio-temporal data, which is not very different from kriging in three dimensions.
- It is difficult to construct models that don't have full symmetry:

$$C(\boldsymbol{r}, h) = C(\boldsymbol{r}, -h)$$

# Last Lecture

- Last lecture, we discussed kriging for spatio-temporal data, which is not very different from kriging in three dimensions.
- It is difficult to construct models that don't have full symmetry:

$$C(\boldsymbol{r}, h) = C(\boldsymbol{r}, -h)$$



# Dynamic AR Models

- **Goal:** a model that makes evolution over time explicit.
- At time $t$, observe vector $\boldsymbol{y}_t = \begin{bmatrix} y(\boldsymbol{s}_1, t) & \cdots & y(\boldsymbol{s}_n, t) \end{bmatrix}^T$.
- (Multivariate) AR model, also called dynamical model:

$$\boldsymbol{y}_t = \Phi \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma).$$

(Can estimate and subtract out mean trend first: $\boldsymbol{y}_t \leftarrow \boldsymbol{y}_t - \boldsymbol{\mu}_t$.)

- Many many parameters: $\Phi$ and $\Sigma$.
- Parametrize $\Phi = \Phi(\boldsymbol{\alpha})$ and $\Sigma = \Sigma(\boldsymbol{\theta})$, e.g.,

$$\Phi_{ij}(\boldsymbol{\alpha}) = \alpha_1 e^{-\boldsymbol{\alpha}_2^T(\boldsymbol{s}_i - \boldsymbol{s}_j)} \qquad\qquad \Sigma_{ij} = \theta_1 e^{-\theta_2 ||\boldsymbol{s}_i - \boldsymbol{s}_j||}$$

- How would you interpret these parameters?

## Likelihood Equations

$$\boldsymbol{y}_t = \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \Sigma(\boldsymbol{\theta})).$$

- Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ by maximum likelihood.
- The likelihood is

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{t=1}^{T} p_{\boldsymbol{\alpha},\boldsymbol{\theta}}(\boldsymbol{y}_t | \boldsymbol{y}_0, ..., \boldsymbol{y}_{t-1}) = \prod_{t=1}^{T} p_{\boldsymbol{\alpha},\boldsymbol{\theta}}(\boldsymbol{y}_t | \boldsymbol{y}_{t-1})$$

  (We typically assume $\boldsymbol{y}_0 = \boldsymbol{0}$.)
- What is $p_{\boldsymbol{\alpha},\boldsymbol{\theta}}(\boldsymbol{y}_t | \boldsymbol{y}_{t-1})$?

$$p(\boldsymbol{y}_t | \boldsymbol{y}_{t-1}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma(\boldsymbol{\theta})|}} e^{-\frac{1}{2}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T \Sigma(\boldsymbol{\theta})^{-1} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})}$$

- So putting everything together, we have

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \frac{\text{const.}}{|\Sigma(\boldsymbol{\theta})|^{T/2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T \Sigma(\boldsymbol{\theta})^{-1} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})}$$

# Likelihood Equations

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \frac{\text{const.}}{|\Sigma(\boldsymbol{\theta})|^{T/2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T \Sigma(\boldsymbol{\theta})^{-1}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})}$$

- The log-likelihood is

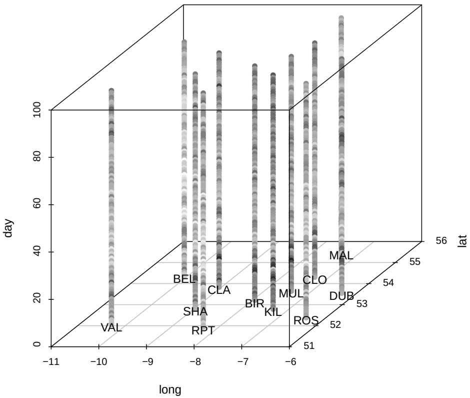
$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \text{const.} - \frac{T}{2}\log|\Sigma(\boldsymbol{\theta})| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T \Sigma(\boldsymbol{\theta})^{-1}(\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})$$

- Pass function into a numerical optimizer (e.g., R's `optim`).

# Where are we?

# Irish Wind Data

$$\boldsymbol{y}_t = \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma(\boldsymbol{\theta})).$$




Let's fit this model to the Irish wind data.

- What does $\Phi_{ij}$ capture? $\Sigma_{ij}$?
- Recall that

$$\Phi_{ij}(\boldsymbol{\alpha}) = \alpha_1 e^{-\boldsymbol{\alpha}_2^T(\boldsymbol{s}_i - \boldsymbol{s}_j)}.$$

What do $\alpha_1$ and $\boldsymbol{\alpha}_2$ mean?

Let's go into R.

# Where are we?

# Avoiding Matrix Inversion

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \text{const.} - \frac{T}{2} \log |\Sigma(\boldsymbol{\theta})| - \frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T \Sigma(\boldsymbol{\theta})^{-1} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})$$

- Can avoid matrix inversion by modeling precision matrix ($K = \Sigma^{-1}$).
- Then log-likelihood becomes

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \text{const.} + \frac{T}{2} \log |K(\boldsymbol{\theta})| - \frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})^T K(\boldsymbol{\theta}) (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1})$$

- Now most demanding task is computing the determinant.

# Higher-Order Information

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \text{const.} + \frac{T}{2} \log |K(\boldsymbol{\theta})| \\ - \frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha}) \boldsymbol{y}_{t-1})^T K(\boldsymbol{\theta}) (\boldsymbol{y}_t - \Phi(\boldsymbol{\alpha}) \boldsymbol{y}_{t-1})$$

- Notice that we simply passed $-\ell$ to `optim`, and it minimized it. How is this possible?
- `optim` uses a heuristic called the Nelder-Mead method.
- This will fail when there are more parameters.
- More reliable methods require the gradient (first derivative) or Hessian (second derivative).

# Where are we?

# Measurement Noise

$$\begin{aligned} \boldsymbol{y}_t &= \Phi(\boldsymbol{\alpha})\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, & \boldsymbol{\epsilon}_t &\sim N(\mathbf{0}, \Sigma(\boldsymbol{\theta})) \\ \boldsymbol{z}_t &= \boldsymbol{y}_t + \boldsymbol{\delta}_t, & \boldsymbol{\delta}_t &\sim N(\mathbf{0}, \sigma^2 I) \end{aligned}$$

If we only observe $\boldsymbol{z}_t$, then we have two problems:

- Estimate true states $\boldsymbol{y}_t$. $\Rightarrow$ Kalman filter
- Estimate parameters $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$. $\Rightarrow$ EM algorithm

# A Word about the EM Algorithm

- The EM algorithm is designed for maximum likelihood estimation when there is missing data (e.g., $\boldsymbol{y}_t$).
- The idea is simple. Start with a guess of the parameters and iterate between:
  1. **Expectation:** "filling in" $\boldsymbol{y}_t$ given current parameters.
  2. **Maximization:** updating parameters by maximum likelihood.
- Where might Kalman filter help with EM algorithm, for our problem?
- EM algorithm is a special case of the majorization-minimization algorithm (Homework 4a), where majorizing function is:

$$-\ell_{\boldsymbol{z}}(\boldsymbol{\theta}) \leq E_{\boldsymbol{\theta}^{(m)}}(-\ell_{\boldsymbol{y},\boldsymbol{z}}(\boldsymbol{\theta})|\boldsymbol{z})$$

- EM calculations tend to be extremely messy. (Details for Kalman filter in Shumway / Stoffer.)

# Inconsistent Spatial Information

What if we observe different spatial locations at each time frame?

1. Take set of all locations that are ever observed. If you don't observe a location at a given time, treat that as "missing data." (There is a missing data extension for the Kalman filter, see Shumway / Stoffer.)
2. If we have a parametric model:

$$\Phi_{ij} = \alpha_1 e^{-\alpha_2 ||\boldsymbol{s}_i - \boldsymbol{s}_j||},\ i = 1, ..., n$$

   we can simply allow $\Phi$ to be time varying:

$$\Phi_{ij}^{(t)} = \alpha_1 e^{-\alpha_2 ||\boldsymbol{s}_i^{(t)} - \boldsymbol{s}_j^{(t-1)}||},\ i = 1, ..., n_t,\ j = 1, ..., n_{t-1}$$

# Where are we?

# Summary

- Dynamic AR models provide an attractive way of modeling spatial dependence and evolution over time.
- Estimation by maximum likelihood. (Requires iterative methods.)
- Measurement noise can be incorporated. The Kalman filter makes these computations efficient.

# This Week (Week 6)

- We are done covering the "core" material of this class.
- I will be leaving town tomorrow, returning next Wednesday. I will be available by e-mail.
- Edgar and Jingshu will be here until the end of the week and hold their scheduled workshops.

# Next Week (Week 7)

- No lectures Monday and Wednesday. (All of us will be away at the Joint Statistical Meetings in Boston.)
- Project presentations Thursday and Friday:
  - Thursday 2:30-4:30 in 380-380W
  - Friday 1-3 in Sequoia 200
- I will send out a link to a sign-up form tonight.
- We'll try to group the presentations by topic.

# Week 8

- Homeworks 4a and/or 4b due Monday!
- The most requested topic in the midquarter surveys was hierarchical models.
- We'll cover Bayesian hierarchical models in the last week.
- Some of the other requested topics will be covered by student projects: specific applications, discrete cosine transform, etc.
- Come to the presentations next Thursday and Friday!