Lecture 4
State-Space Models

Dennis Sun
Stats 253

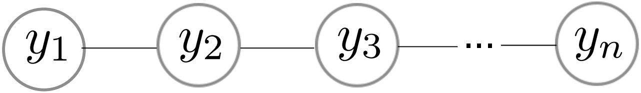
July 2, 2014

Outline of Lecture

1. Last Class
2. Measurement Error Model
3. State-space models
4. Wrapping Up

Where are we?

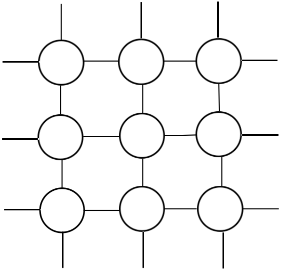
Autoregressive Processes



In time, there are two equivalent formulations:

- $y_t - \mu_t = \phi(y_{t-1} - \mu_{t-1}) + \epsilon_t$
- $y_t \mid y_1, ..., y_{t-1} \sim N(\mu_t + \phi(y_{t-1} - \mu_{t-1}), \sigma^2)$

Autoregressive Processes



- In space, these two formulations are different:

$$y_i - \mu_i = \phi \sum_j w_{ij}(y_j - \mu_j) + \epsilon_i$$

$$y_i \,\big|\, (y_j, j \neq i) \sim N\left(\mu_i + \phi \sum_j w_{ij}(y_j - \mu_j), \sigma^2\right)$$

- Can solve by writing in vector form.
- When is the conditional formulation well-defined?

When is CAR well-defined?

- **Main question:** When do conditional distributions $p(y_i|y_j, j \neq i)$ specify a valid joint distribution $p(y_1, ..., y_n)$?
- **Brook's Lemma (Besag 1974):**

$$\frac{p(\mathbf{y})}{p(\mathbf{0})} = \prod_{i=1}^{n} \frac{p(y_i|y_1, ..., y_{i-1}, 0, ..., 0)}{p(0|y_1, ..., y_{i-1}, 0, ..., 0)}$$

- Allows us to calculate $p(\mathbf{y})$ up to a normalizing constant, i.e.,

$$p(\mathbf{y}) = \frac{\frac{\mathbf{p(y)}}{\mathbf{p(0)}}}{\sum_{\mathbf{y}} \frac{\mathbf{p(y)}}{\mathbf{p(0)}}}$$

- *What this tells us:* One necessary condition for $p(\mathbf{y})$ to be well-defined is that

$$\sum_{\mathbf{y}} \frac{p(\mathbf{y})}{p(\mathbf{0})} = \sum_{\mathbf{y}} \prod_{i=1}^{n} \frac{p(y_i|y_1, ..., y_{i-1}, 0, ..., 0)}{p(0|y_1, ..., y_{i-1}, 0, ..., 0)} < \infty$$

When is CAR well-defined?

- There are also symmetry constraints. To see this,

$$\frac{p(\mathbf{y})}{p(\mathbf{0})} = \left(\frac{p(\mathbf{0})}{p(\mathbf{y})}\right)^{-1} = \left(\prod_{i=1}^{n} \frac{p(0|0,...,0,y_{i+1},...,y_n)}{p(y_i|0,...,0,y_{i+1},...,y_n)}\right)^{-1}$$
$$= \prod_{i=1}^{n} \frac{p(y_i|0,...,0,y_{i+1},...,y_n)}{p(0|0,...,0,y_{i+1},...,y_n)}.$$

  But on the previous slide we had

$$\frac{p(\mathbf{y})}{p(\mathbf{0})} = \prod_{i=1}^{n} \frac{p(y_i|y_1,...,y_{i-1},0,...,0)}{p(0|y_1,...,y_{i-1},0,...,0)}.$$

  These must be equal!
- The symmetry constraints will be satisfied provided that $w_{ij} = w_{ji}$.

Tradeoffs between SAR and CAR

- SAR: no constraints on w_{ij}, but $\mathrm{Cov}(\epsilon_i, y_j) \neq 0$.
- CAR: generalizes to non-Gaussian data, requires $w_{ij} = w_{ji}$.

A Word about Homework 1

- Play around with SAR and CAR models on an image data set.
- "Big data": a 64×64 image is 4096 nodes. Processing will take some time!
- **Tips:** Form adjacency matrix `W` and call `mat2listw(W)`.

Where are we?

Problems with the Autoregressive Model

$$y_i = \phi \sum_j w_{ij} y_j + \epsilon_i$$

- Does not allow for measurement error.
- The fitted values are

$$\hat{y}_i = \hat{\phi} \sum_j w_{ij} y_j$$

The Measurement Error Model

$$\textbf{State model:} \quad y_i - \mu_i = \phi \sum_j w_{ij}(y_j - \mu_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$\textbf{Data model:} \quad z_i = y_i + \delta_i, \quad \delta_i \sim N(0, \tau^2)$$

There are now two inferential questions:

- As before, we might wish to estimate parameters: ϕ, σ^2, and now τ^2.
- We might also wish to estimate the underlying state y_i from the noisy observations z_i.

Parameter Estimation in the Measurement Error Model

Write in vector form:

State model:	$\boldsymbol{y} = \boldsymbol{\mu} + (I - \phi W)^{-1}\boldsymbol{\epsilon}$
Data model:	$\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\delta}$

- We want the likelihood $L(\phi, \sigma^2, \tau^2) = p(\boldsymbol{z}|\phi, \sigma^2, \tau^2)$.
- (Note that the model is defined in terms of $p(\boldsymbol{y}|\phi, \sigma^2)p(\boldsymbol{z}|\boldsymbol{y}, \tau^2)$.)
- $\boldsymbol{z} = \boldsymbol{\mu} + (I - \phi W)^{-1}\boldsymbol{\epsilon} + \boldsymbol{\delta}$ so

$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \sigma^2(I - \phi W)^{-1}(I - \phi W^T)^{-1} + \tau^2 I)$$

- Since we now know $p(\boldsymbol{z}|\phi, \sigma^2, \tau^2)$, we can now calculate the MLE.

State Estimation in the Measurement Error Model

- How should we optimally estimate y_i, having observed z_i?
- **Criterion**: Find $f(\boldsymbol{z})$ to minimize the mean squared error:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \operatorname{E}||\boldsymbol{y} - f(\boldsymbol{z})||^2.$$

- **Solution**: $\hat{f}(\boldsymbol{z}) = \operatorname{E}(\boldsymbol{y}|\boldsymbol{z})$, the minimum mean squared error (MMSE) estimator.
- **Question**: How do we actually calculate $\operatorname{E}(\boldsymbol{y}|\boldsymbol{z})$?
- We will use the fact that $\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}$ is a multivariate normal.

A Detour into Multivariate Normal Theory

- We say $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ if

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

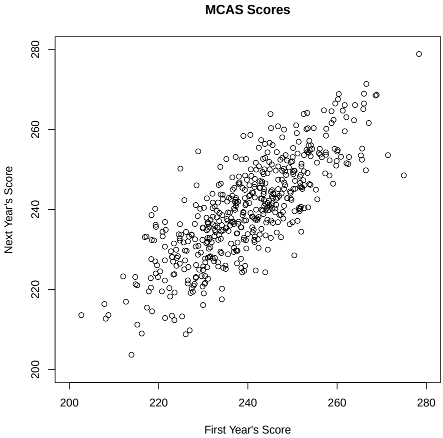
- Can show that $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{x}) = \Sigma$.
- If $\boldsymbol{x}$ is partitioned so that

$$\begin{pmatrix}\boldsymbol{x}_1 \\ \boldsymbol{x}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right),$$

  then we have

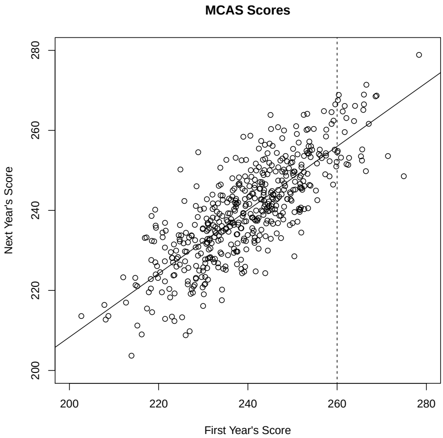
$$\begin{aligned}\mathrm{E}(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ \mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

2D Picture of the Conditional Distribution



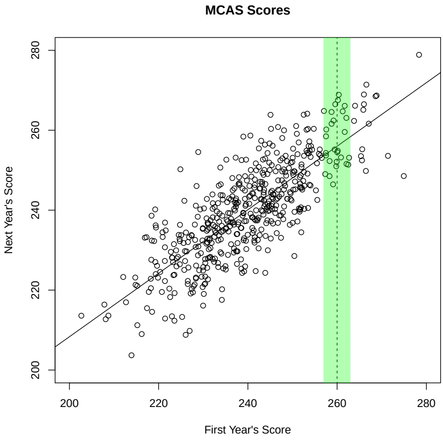
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

2D Picture of the Conditional Distribution



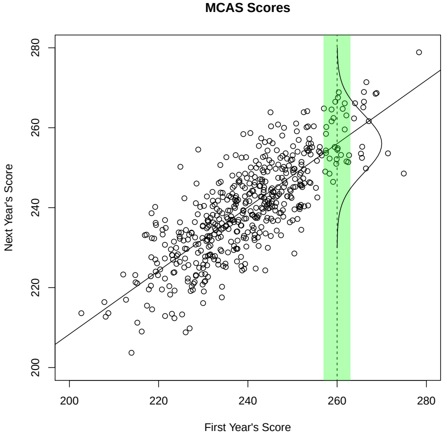
What is $x_1 | x_2 = 260$?

2D Picture of the Conditional Distribution



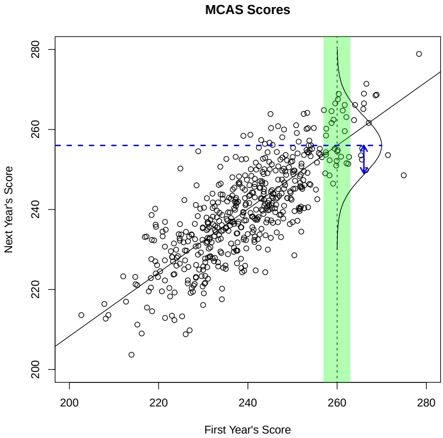
What is $x_1 | x_2 = 260$?

2D Picture of the Conditional Distribution



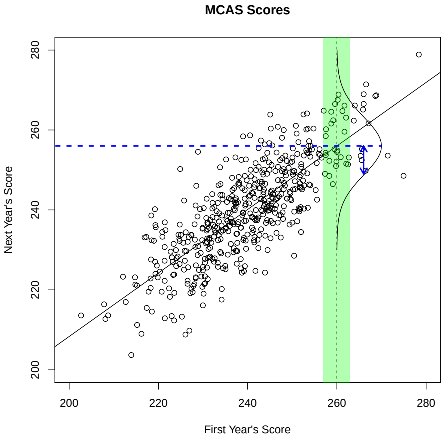
What is $x_1 | x_2 = 260$?

2D Picture of the Conditional Distribution



$$\mathrm{E}(x_1|x_2) = \mu_1 + \underbrace{\rho\sigma_1\sigma_2}_{\Sigma_{12}} \underbrace{\sigma_2^{-2}}_{\Sigma_{22}^{-1}}(x_2 - \mu_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$$

2D Picture of the Conditional Distribution



$$\mathrm{Var}(x_2|x_1) = \sigma_2^2 - (\rho\sigma_1\sigma_2)\sigma_1^{-2}(\rho\sigma_1\sigma_2) = \sigma_2^2(1-\rho^2)$$

Let's apply the theory

Recall the **measurement error model**:

State model: $\boldsymbol{y} = \boldsymbol{\mu} + (I - \phi W)^{-1}\boldsymbol{\epsilon}$

Data model: $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\delta}$

What is $\mathrm{E}(\boldsymbol{y}|\boldsymbol{z})$?

- First need to determine joint distribution

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_z} \end{pmatrix}, \begin{pmatrix} \Sigma_{\boldsymbol{yy}} & \Sigma_{\boldsymbol{yz}} \\ \Sigma_{\boldsymbol{zy}} & \Sigma_{\boldsymbol{zz}} \end{pmatrix} \right).$$

- Then apply formula $\mathrm{E}(\boldsymbol{y}|\boldsymbol{z}) = \boldsymbol{\mu_y} + \Sigma_{\boldsymbol{yz}}\Sigma_{\boldsymbol{zz}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu_z})$.

Summary

- We introduced the **measurement error model**, which assumes the observations contain noise.
- There are two problems:
  - Parameter estimation: essentially the same as with the AR model.
  - State estimation: calculate the conditional expectation $\mathrm{E}(\boldsymbol{y}|\boldsymbol{z})$.

Where are we?

State-space model

The **state-space model** is a (multivariate) generalization of the measurement error model.

State model: $$\boldsymbol{y}_i - \boldsymbol{\mu}_i = \sum_j w_{ij} \Phi(\boldsymbol{y}_j - \boldsymbol{\mu}_j) + \boldsymbol{\epsilon}_i$$

Data model: $$\boldsymbol{z}_i = A\boldsymbol{y}_i + \boldsymbol{\delta}_i$$

Since each observation is now a vector, if we aggregate all the observations together, we get a matrix, i.e.,

$$Y = \begin{bmatrix} | & & | \\ \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_n \\ | & & | \end{bmatrix}$$

so the state model is something like

$$Y - M = \Phi(Y - M)W^T + E.$$

State-space model

State model: $$Y - M = \Phi(Y - M)W^T + E$$

Data model: $$Z = AY + D$$

- Still possible to write down $p(Z|\Phi, \sigma^2, A, \tau^2)$. Requires use of vec operator and tensor products $\otimes$.
- Hence, principles of inference remain the same:
  - parameter estimation by MLE
  - state estimation by MMSE (i.e., $\mathrm{E}(\boldsymbol{y}|\boldsymbol{z}) = \mu_{\boldsymbol{y}} + \Sigma_{\boldsymbol{yz}}\Sigma_{\boldsymbol{zz}}^{-1}(\boldsymbol{z} - \mu_{\boldsymbol{z}})$)
- Dimensions become enormous!

Bottom line: major pain

The Kalman Filter

When working with temporal data, there is a cleverer way to do state estimation.

$$\textbf{State model:} \quad \boldsymbol{y}_t - \boldsymbol{\mu}_t = \Phi(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_{t-1}) + \boldsymbol{\epsilon}_t$$

$$\textbf{Data model:} \quad \boldsymbol{z}_t = A\boldsymbol{y}_t + \boldsymbol{\delta}_t$$

Goals: Calculate $\boldsymbol{y}_{t|t} = \mathrm{E}(\boldsymbol{y}_t|\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ and $P_{t|t} = \mathrm{E}(\boldsymbol{y}_t - \boldsymbol{y}_{t|t})(\boldsymbol{y}_t - \boldsymbol{y}_{t|t})^T$.

The Kalman Filter

- $\boldsymbol{y}_{t|t}$ and $P_{t|t}$ can be calculated in terms of $\boldsymbol{y}_{t|t-1}$ and $P_{t|t-1}$.
- Likewise, $\boldsymbol{y}_{t+1|t}$ and $P_{t+1|t}$ can be calculated in terms of $\boldsymbol{y}_{t|t}$ and $P_{t|t}$.
- Alternating between filtering and one-step-ahead prediction gives the Kalman filter.

Filtering

$$\boldsymbol{y}_{t|t} = \boldsymbol{y}_{t|t-1} + K_t(\boldsymbol{z}_t - A\boldsymbol{y}_{t|t-1})$$
$$P_{t|t} = (I - K_t A)P_{t|t-1}$$

where $K_t = P_{t|t-1}A^T(AP_{t|t-1}A^T + \tau^2 I)^{-1}$.

One-Step Ahead Prediction

$$\boldsymbol{y}_{t+1|t} = \boldsymbol{\mu}_{t+1} + \Phi(\boldsymbol{y}_{t|t} - \boldsymbol{\mu}_t)$$
$$P_{t+1|t} = \Phi P_{t|t}\Phi^T + \sigma^2 I$$

Kalman Filter Summary

- The Kalman filter provides an efficient algorithm for state estimation in a state-space model.
- It turns out that having an efficient algorithm for state estimation also simplifies parameter estimation. (See Shumway and Stoffer Ch. 6.)
- To my knowledge, no analog of Kalman filter in space.
- State-space models in space: modeling and efficient algorithms. (Project idea?)

Where are we?

Summary of Today's Lecture

- Extended AR model to allow for measurement noise.
- Measurement noise model is a special case of a state-space model.
- State estimation requires some knowledge about multivariate normals.
- The Kalman filter is a custom algorithm to do state estimation for time series.

Administrivia

- Homework 1 due Monday.
- Workshops start tomorrow.
  - Edgar: Thursdays 2:30-4:30pm in 380-380F
  - Jingshu: Fridays 1-3pm in Sequoia 200
- Please use Piazza!