STOCHASTIC MODELING OF ENVIRONMENTAL TIME SERIES

Richard W. Katz

LECTURE 4

(1) Hidden Markov Models: Properties

(2) Hidden Markov Models: Motivational Example

(3) EM Algorithm for Hidden Markov Model

(1) Hidden Markov Models: Properties

- **Definition of Hidden Markov Model**

-- Finite-state observed variable

- **Properties of Hidden Markov Model**

-- Binary hidden Markov model

Definition of Hidden Markov Model

- **General case**

-- **Observed sequence of variables** $\{X_t\text{: } t = 1, 2, \ldots, T\}$

-- **Finite-state, first-order Markov chain** $\{J_t\text{: } t = 1, 2, \ldots, T\}$

Transition probabilities $P_{ij} = \Pr\{J_{t+1} = j \mid J_t = i\},\ i,\ j = 0, 1, \ldots$

-- $\{X_t\}$ **conditionally i.i.d. given** $\{J_t\}$

Conditional distribution function $F_i(x) = \Pr\{X_t \leq x \mid J_t = i\},\ i = 0, 1, \ldots$

[conditional mean $\mu_i = \mathrm{E}(X_t \mid J_t = i)$**, variance** $\sigma_i^2 = \mathrm{Var}(X_t \mid J_t = i)$**]**

- **Finite-state observed variable**

-- This case receives most attention

$$\Pr\{X_t = x \mid J_t = i\} = p_i(x), \quad x = 0, 1, \ldots; \quad i = 0, 1, \ldots$$

-- For example, binary variable ($x = 0, 1$)

Set $p_i = p_i(1) = \Pr\{X_t = 1 \mid J_t = i\}, \quad i = 0, 1$

Conditional mean $\mu_i = p_i, \quad i = 0, 1, \ldots$

Conditional variance $\sigma_i^2 = p_i(1 - p_i), \quad i = 0, 1, \ldots$

Properties of Hidden Markov Model

- **Properties of Markov chain (two-state, first-order)**

-- Variables: $\{J_t: t = 1, 2, \ldots, T\}, \; J_t = 0, 1$

-- Transition probs.: $P_{i1} = \Pr\{J_{t+1} = 1 \mid J_t = i\}, \; i = 0, 1$

-- Initial prob.: $\delta = \delta(1) = \Pr\{J_1 = 1\}, \; \delta(0) = 1 - \delta$

-- Stationary process: Set $\delta = \pi = P_{01} / (P_{10} + P_{01})$

-- Moments (stationary case): $E(J_t) = \pi, \; Var(J_t) = \pi(1 - \pi),$

$\mathrm{Corr}(J_t, J_{t+l}) = d^l$, where $d = P_{11} - P_{01}$

-- *l*-step transition probs.:

Define $P_{i1}^{(l)} = \Pr\{J_{t+l} = 1 \mid J_t = i\}, \; i = 0, 1, \; l = 1, 2, \ldots$

Then $P_{01}^{(l)} = \pi\,(1 - d^{\,l}), \; P_{11}^{(l)} = \pi + (1 - \pi)\, d^{\,l},$

-- Conditional independence

$$\Pr\{ J_{t+1} = j \mid J_t = i\} = \Pr\{ J_{t+1} = j \mid J_t = i, J_{t-1}, \ldots, J_1\}$$

But $\Pr\{ J_{t+1} = j \mid J_{t-1} = i\}$ still depends on i

-- Run length (geometric distribution)

$$\Pr(\text{run of state } i \text{ of length } l\} = P_{ii}^{\,l-1} P_{i,\,1-i}\,, \; l = 1, 2, \ldots; \; i = 0, 1$$

- **Hidden two-state Markov chain**

General observed variable (i.e., not necessarily discrete)

-- Moments

$E(X_t) = (1 - \pi)\mu_0 + \pi\mu_1,$

$\mathrm{Var}(X_t) = (1 - \pi)\sigma_0^2 + \pi\sigma_1^2 + \pi(1 - \pi)(\mu_1 - \mu_0)^2$

$\mathrm{Corr}(X_t, X_{t+l}) =$

$$\{[\pi(1 - \pi)(\mu_1 - \mu_0)^2] / [(1 - \pi)\sigma_0^2 + \pi\sigma_1^2 + \pi(1 - \pi)(\mu_1 - \mu_0)^2]\} d^l$$

-- Reduces to independent two-component mixture model if $d = 0$

- **Binary Hidden Two-State Markov Model**

Assume two-state observations, as well as two-state Markov chain

-- Moments

$$E(X_t) = \Pr\{X_t = 1\} = (1 - \pi)p_0 + \pi p_1,$$

$$\mathrm{Var}(X_t) = (1 - \pi)p_0(1 - p_0) + \pi p_1(1 - p_1) + \pi(1 - \pi)(p_1 - p_0)^2$$

$$\mathrm{Corr}(X_t, X_{t+l}) =$$

$$\{[\pi(1 - \pi)(p_1 - p_0)^2] / [(1 - \pi)p_0(1 - p_0) + \pi p_1(1 - p_1) + \pi(1 - \pi)(p_1 - p_0)^2]\} d^{\,l}$$

-- Run length

Can be mixture of two geometric distributions

Capable of producing longer or shorter mean run length than comparable Markov chain model (i. e., $\Pr\{X_t = 1\}$ same for both models)

Examples: See MacDonald & Zucchini text (pp. 88-89)

-- Reduces to Markov chain (e. g., if $p_0 = 0$, $p_1 = 1$)

-- Hidden Markov model is not necessarily Markov chain

Examples: See Guttorp text (p. 105) & MacDonald & Zucchini text (pp. 82-83)

(2) Hidden Markov Models: Motivational Example

- **Stochastic "weather generator"**

  -- Markov chain not necessarily "hidden" (relax conditional independence condition)

  -- Definition of model

  $\{(X_t, J_t)$: $t = 1, 2, \ldots, T\}$ where X_t denotes temperature (e.g., minimum or maximum) and J_t denotes precipitation occurrence on tth day at same location

  $\{J_t\}$ is assumed first-order, two-state Markov chain (parameters P_{01}, P_{11} or π, d)

Assume $X_t \mid J_t = i$ is distributed as $N(\mu_i, \sigma_i^2)$, $i = 0, 1$

Given $J_t = i$, set $Z_t = (X_t - \mu_i) / \sigma_i$ (random standardization)

Conditional on $\{J_t\}$, assume $\{Z_t: t = 1, 2, \ldots\}$ as AR(1) process with conditional first-order autocorrelation coefficient φ; that is,

$$Z_{t+1} = \varphi Z_t + \varepsilon_t, \text{ where } \varepsilon_t \text{ is } N(0, 1 - \varphi^2),$$

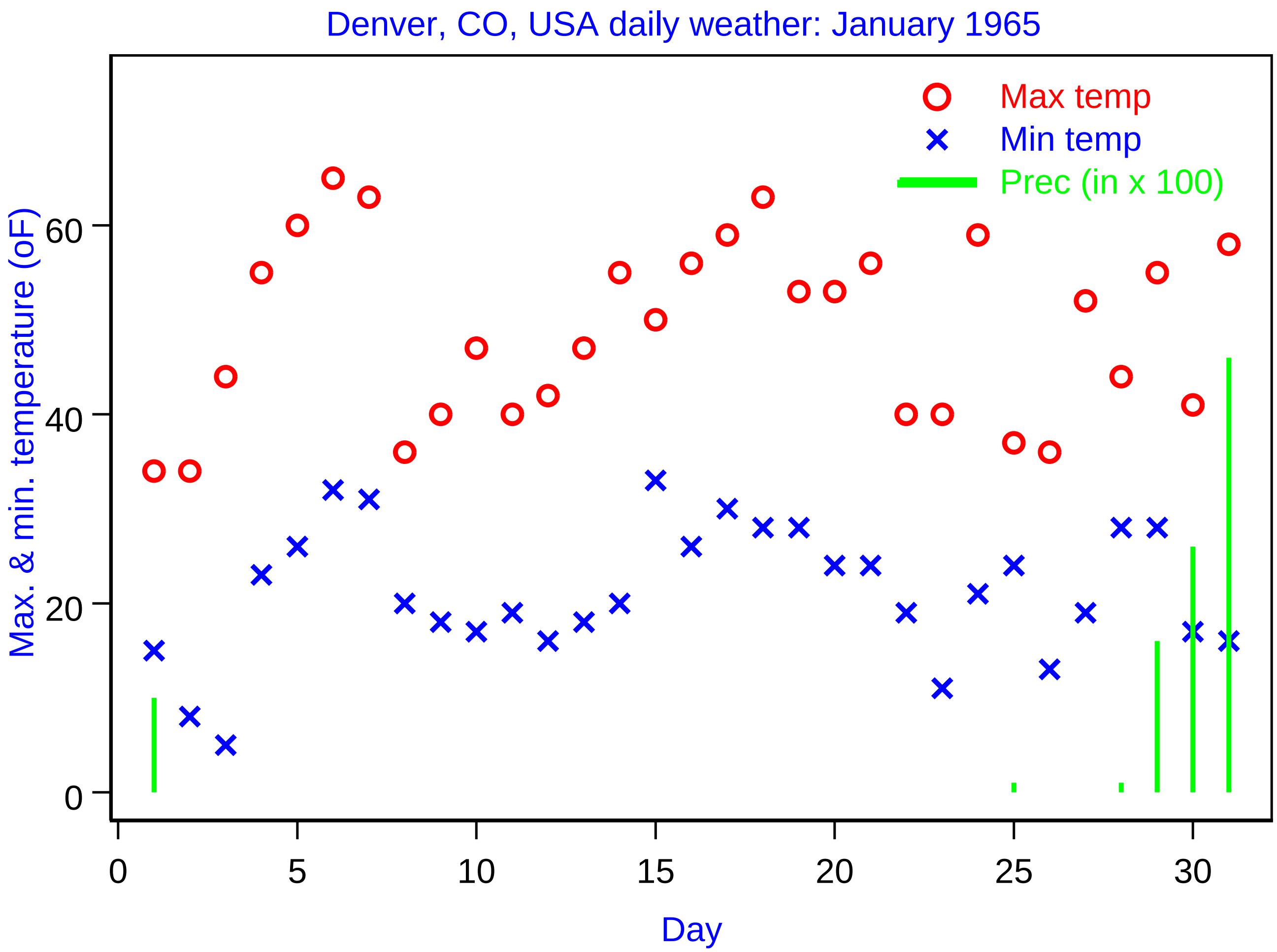
-- Properties of model

$$E(X_t) = (1 - \pi)\mu_0 + \pi\mu_1,$$

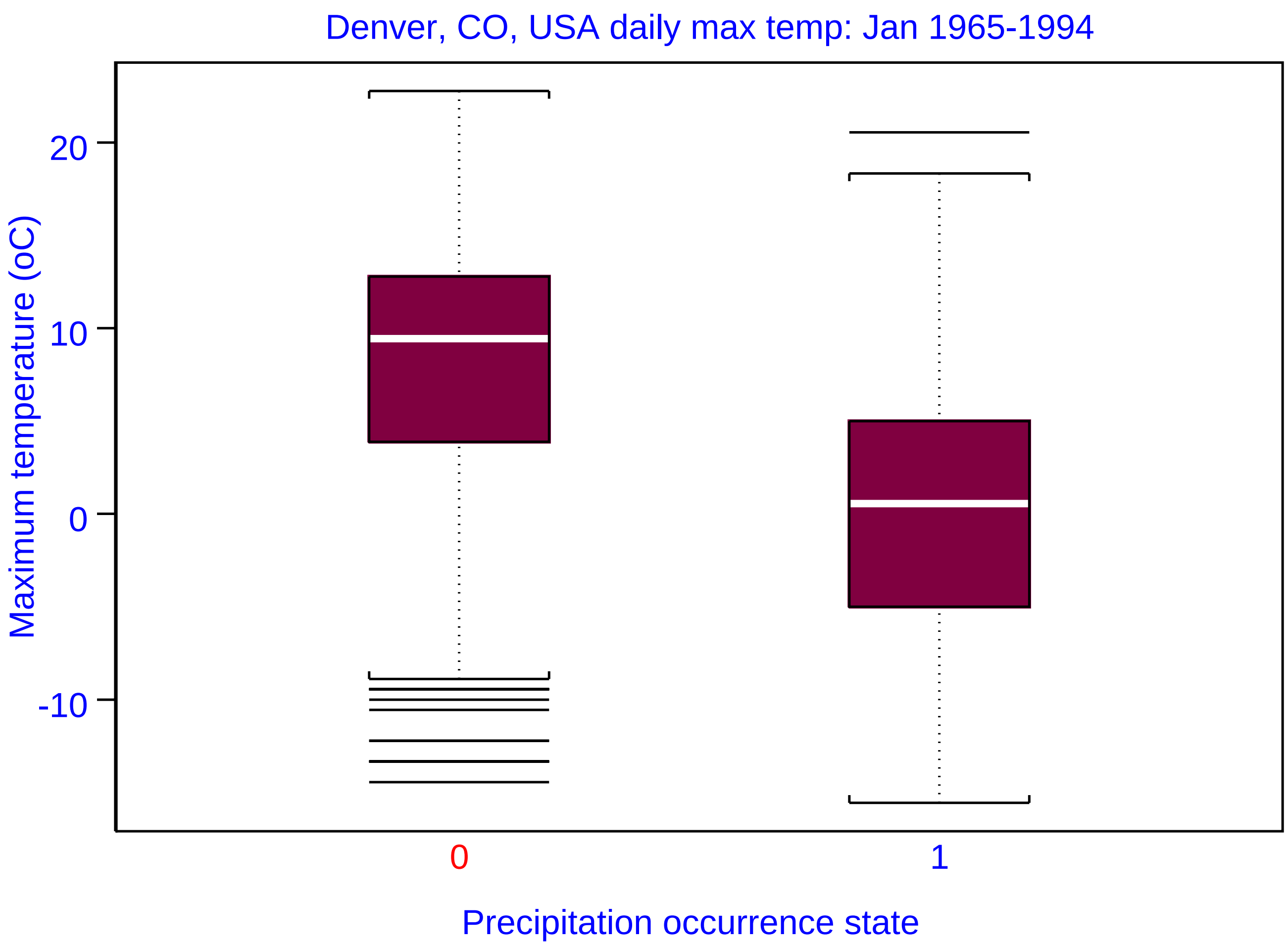
$$Var(X_t) = (1 - \pi)\sigma_0^2 + \pi\sigma_1^2 + \pi(1 - \pi)(\mu_1 - \mu_0)^2$$

$$Cov(X_t, X_{t+l}) =$$

$$\varphi^l [(1 - \pi)\sigma_0^2 + \pi\sigma_1^2] + \pi(1 - \pi)[d^l (\mu_1 - \mu_0)^2 - \varphi^l (1 - d^l)(\sigma_1 - \sigma_0)^2]$$



Denver, CO, USA daily weather: January 1965
Max temp
Min temp
Prec (in x 100)
Max. & min. temperature (oF)
0
20
40
60
0
5
10
15
20
25
30
Day


Denver, CO, USA daily max temp: Jan 1965-1994
Maximum temperature (oC)
20
10
0
-10
0
1
Precipitation occurrence state

- **Application**

-- **Denver, Colorado, USA January daily maximum temperature (°C) and precipitation occurrence, 1965-1994**

Precipitation occurrence: $\pi = 0.178$, $d = 0.219$

Sample statistics	Unconditional	Given $J_t = 0$	Given $J_t = 1$
Mean (°C)	6.5	8.0	0.0
Std. dev. (°C)	7.4	6.7	7.0
Autocorrelation	0.62**	0.55*	0.55*

* **Constrained to be same for $J_t = 0$ & $J_t = 1$**

** **Model calculation yields 0.49**

(3) EM Algorithm for Hidden Markov Model

- **Software**

  -- Speech recognition orientation

- **Tutorial**

  -- Companion to software

- **Formulation of EM Algorithm**

Software for Fitting Hidden Markov Models

- **Download from WWW**

  -- Web address: ***www.cfar.umd.edu/~kanungo/software/software.html***

  -- Author: Tapas Kanungo, Center for Automation Research, Univ. of Maryland

  -- Software written in C

  -- Treats finite-state observed variable with finite-state, first-order hidden Markov chain

Tutorial on Hidden Markov Models

- **PDF File (*hmmtut.pdf*)**

  -- Posted on EPFL site for this course

  -- Notation

  Makes use of quite different notation

  (from engineering/speech processing literature)

- **Speech Recognition**

-- Utterance

Observed sequence is acoustic signal corresponding to “utterance”

-- Words

Assume utterance comes known vocabulary of “words”

-- State of hidden Markov chain represents configuration of vocal tract

-- Training Problem

For each word in vocabulary, fit hidden Markov chain to utterance

-- Classification Problem

Use training model to compute probability of each word (choose word with highest probability)

-- "Localized decoding"

Determine state of Markov chain at time (i.e., J_t) with highest posterior probability

-- "Global decoding"

Determine sequence of states of Markov chain (i.e., $J_1, J_2, \ldots, J_T$) with highest joint posterior probability

Formulation of EM Algorithm for Hidden Markov Model

Binary observed variable with two-state hidden Markov chain

- **Data**

  -- **Observed data** $\{x_t: t = 1, 2, \ldots, T\}$, $x_t = 0, 1$

  -- **Unobservable data** $\{j_t: t = 1, 2, \ldots, T\}$

  Here $j_t = 0$ **or** 1 **(state of hidden Markov chain** J_t**)**

  -- **Notation:** $n_{ij}(t) = 1$ **if** $j_t = i$ **and** $j_{t+1} = j$ **[**$n_{ij}(t) = 0$ **otherwise]**, $t = 1, 2, \ldots, T - 1$

  $n_i(t) = 1$ **if** $j_t = i$ **[**$n_i(t) = 0$ **otherwise]**, $t = 1, 2, \ldots, T$

- **Complete Data Log-Likelihood Function**

$$\log L_C(\delta, P_{01}, P_{11}, p_0, p_1) = \log \delta(j_1) + \sum_{i,j,t} n_{ij}(t) \log P_{ij} + \sum_{i,t} n_i(t) \log p_i(x_t)$$

- **Incomplete Data Likelihood Function**

$$L_T = \Pr\{X_1 = x_1, \ldots, X_T = x_T; \delta, P_{01}, P_{11}, p_0, p_1\}$$

-- Computing L_T

$$L_T = a_T(0) + a_T(1)$$

where

$$a_t(i) = \Pr\{X_1 = x_1, \ldots, X_t = x_t, J_t = i\},$$

$$i = 0, 1;\ t = 1, 2, \ldots, T$$

-- "Forward recursion" for $\{a_t(i)\}$

(1) Initial step

$$a_1(i) = \delta(i)\, p_i(x_1), \quad i = 0, 1$$

(2) Iterative step

$$a_{t+1}(i) = p_i(x_{t+1})\, [a_t(0)\, P_{0i} + a_t(1)\, P_{1i}],$$

$$i = 0, 1; \quad t = 1, 2, \ldots, T - 1$$

[in practice, issue of avoiding underflow by rescaling, etc.]

- **E-Step of EM Algorithm**

-- Need to determine terms of form $\Pr\{J_t = i \mid x_1, x_2, \ldots, x_T\}$

Define

$$b_t(i) = \Pr\{X_{t+1} = x_{t+1}, \ldots, X_T = x_T \mid J_t = i\}$$

$$i = 0, 1; \quad t = 1, 2, \ldots, T$$

with convention that $b_T(i) = 1, i = 0, 1$

Then $\Pr\{J_t = i \mid x_1, x_2, \ldots, x_T\} = [a_t(i)\, b_t(i)] / L_T$

$i = 0, 1; \quad t = 1, 2, \ldots, T$

-- “Backward recursion” for $\{b_t(i)\}$

(1) Initial step

$$b_T(i) = 1,\ i = 0, 1$$

(2) Iterative step

$$b_t(i) = p_0(x_{t+1})\ b_{t+1}(0)\ P_{i0} + p_1(x_{t+1})\ b_{t+1}(1)\ P_{i1},$$

$$i = 0, 1;\quad t = T - 1, T - 2, \ldots, 1$$

[in practice, issue of avoiding underflow by rescaling, etc.]

- **M-Step**

-- Parameter estimates can be obtained via output of forward & backward recursions

(1) Estimate of δ

$$j_1^{(k)} = [a_1(1)\, b_1(1)] \,/\, L_T$$

(estimate of $\Pr\{J_1 = 1 \mid x_1, x_2, \ldots, x_T\}$)

(2) Estimates of p_i ($i = 0, 1$)

$$\left[\sum_t n_i(t)^{(k)} x_t\right] / \sum_t n_i(t)^{(k)}$$

where $n_i(t)^{(k)} = [a_t(i)\, b_t(i)] / L_T,\ \ i = 0, 1;\ \ t = 1, 2, \ldots, T$

($n_i(t)^{(k)}$ **is estimate of** $\Pr\{J_t = i \mid x_1, x_2, \ldots, x_T\}$**)**

(3) Estimates of P_{i1} ($i = 0, 1$)

$$\sum_t n_{i1}(t)^{(k)} / \sum_t \left[n_{i0}(t)^{(k)} + n_{i1}(t)^{(k)}\right]$$

(summations over $t = 1, 2, \ldots, T - 1$**)**

where $n_{ij}(t)^{(k)} = \left[P_{ij}\, p_j(x_{t+1})\, a_t(i)\, b_{t+1}(j)\right] / L_T,\ \ i = 0, 1;\ \ j = 0, 1;\ \ t = 1, 2, \ldots, T - 1$

($n_{ij}(t)^{(k)}$ **is estimate of** $\Pr\{J_t = i, J_{t+1} = j \mid x_1, x_2, \ldots, x_T\}$**)**

- **Other Issues**

-- Standard errors

More complex than for independent mixtures

-- Model selection

Same issues arise as with independent mixtures

-- Assumptions about hidden Markov chain

Distribution of initial state: fixed, arbitrary or stationary?

-- Alternative of direct maximization

Requires efficient calculation of incomplete data likelihood function

-- Posterior probabilities

Localized decoding: Already computed as part of EM algorithm

Global decoding: Can use Viterbi Algorithm (a form of dynamic programming)