

STOCHASTIC MODELING OF ENVIRONMENTAL TIME SERIES

Richard W. Katz

LECTURE 3

- (1) High Frequency Approach to Time Series Modeling**
- (2) Low Frequency Approach to Overdispersed Time Series**
- (3) Low Frequency Approach: Observed Mixture/Covariate**
- (4) Low Frequency Approach: Hidden Mixture/Variable**

(1) High Frequency Approach to Time Series Modeling

Basic Idea: Apparent overdispersion actually attributable to inadequate model for high frequency variations

- **High Frequency Approach to Modeling Precipitation**
- **Example: Chico Daily Precipitation**
- **Methodological Issues: Presence of Overdispersion**

High Frequency Approach to Modeling Precipitation

- Chain-Dependent process

$\{X_t: t = 1, 2, \dots, T\}$ denotes amount of precipitation on t th day

-- Occurrence $\{J_t: t = 1, 2, \dots, T\}$ modeled as two-state, first-order Markov chain ($J_t = 1$ if prec. occurs on t th day $J_t = 0$ otherwise)

(parameters P_{01}, P_{11} or π, d)

Number of wet days: $N(T) = J_1 + J_2 + \dots + J_T$

-- Intensity (i.e., amount X_t given $J_t = 1$) assumed conditional i.i.d. given $\{J_t\}$

(intensity mean μ and variance σ^2)

-- Variance of total precipitation ($S_T = X_1 + X_2 + \dots + X_T$)

$$\begin{aligned}\text{Var}(S_T) &= E[N(T)] \text{Var}[X_t | J_t = 1] + \text{Var}[N(T)] \{E[X_t | J_t = 1]\}^2 \\ &\approx T\{\pi\sigma^2 + \pi(1 - \pi) [(1 + d)/(1 - d)] \mu^2\}\end{aligned}$$

• Extensions of Chain-Dependent process

-- High-order Markov Chain (e.g., order 2, 3, ...)

-- Conditionally dependent intensities [AR(1) process for transformed intensities]

-- Conditionally non-identically distributed intensities (distribution depends on J_{t-1})

Example: Chico Daily Precipitation

- **Chico, CA, USA**

- **January time series of daily precipitation amount (mm)**

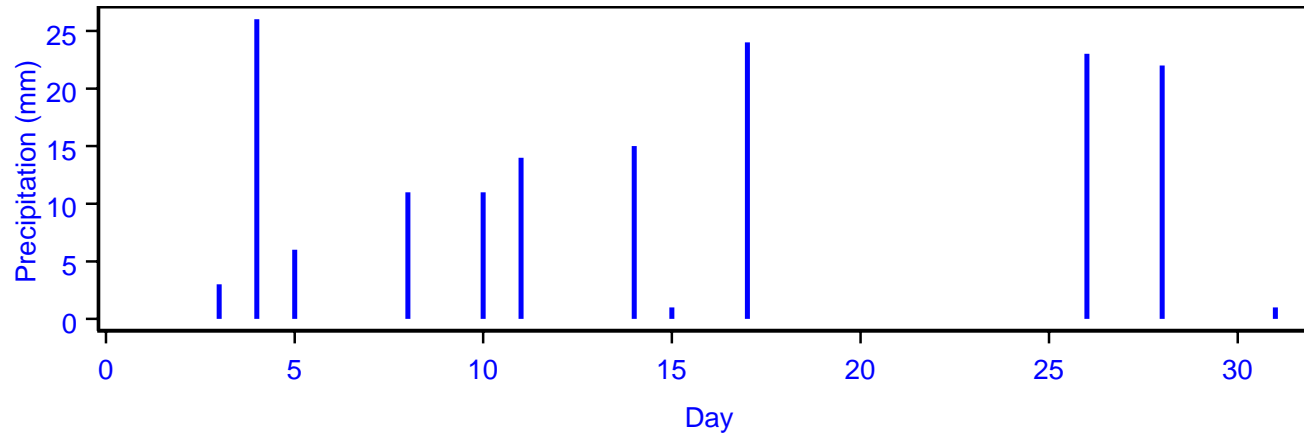
- **78 total years during time period 1907-1998 (4 years eliminated because of missing observations)**

- **Climate**

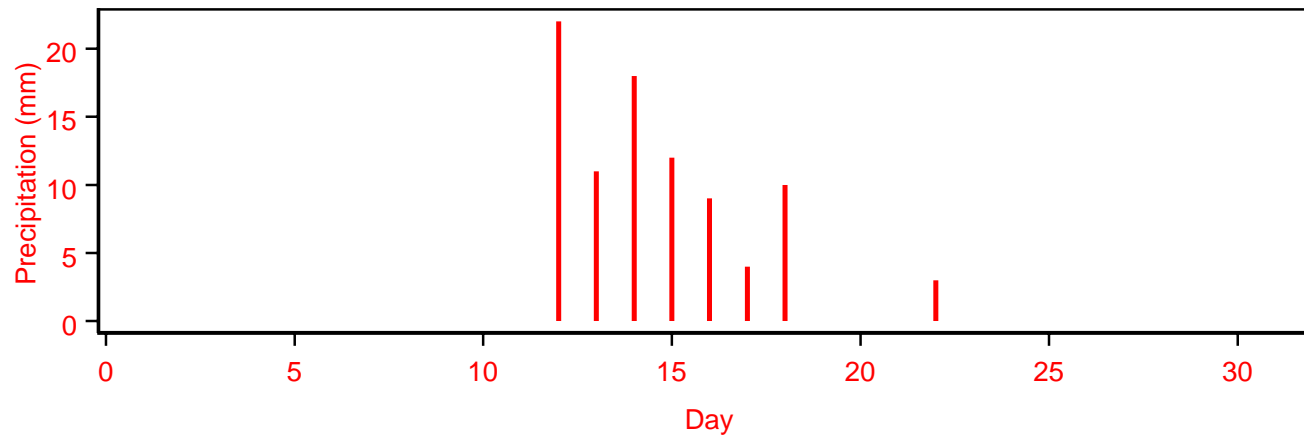
- **Winter is wet season (large annual cycle with summer dry)**

- **Relatively persistent climate (precipitation governed by pressure patterns over Pacific Ocean adjacent to California coast)**

Chico, CA, USA daily precipitation amount: January 1907



Chico, CA, USA daily precipitation amount: January 1913



- **Results for Chico**

Markov chain order	Intensity correlation	Identical intensity dist.	Number param.	$[\text{Var}(S_T)]^{1/2}$
1	No	Yes	4	68.7 mm
2	No	Yes	6	71.1 mm
2	Yes	Yes	7	74.5 mm
2	Yes	No	9	76.4 mm
3	Yes	No	13	79.5 mm
4	Yes	No	21	80.6 mm
<i>Observed:</i>				88.6 mm

Methodological Issues: Presence of Overdispersion

- **Model selection criteria**
 - How do **LRT**, **AIC**, or **BIC** perform when overdispersion ignored?
 - Tend to overfit (because low-frequency overdispersion can induce apparent high-frequency effects)
 - Difficult to protect against
 - Effects on analysis of variance estimates (biased estimates of “potential predictability”?)
 - Compelled to explicitly model overdispersion

(2) Low Frequency Approach to Overdispersed Time Series

Basic idea

-- Two climate regimes between which parameters of daily precipitation time series randomly shift from year to year

-- Regimes may be treated as either *observed* or *hidden*

- Description of Low Frequency Model

- Properties of Low Frequency Model

Description of Low Frequency Model

- **Conditioning variable**

- Two-state conditioning variable I with $w = \Pr\{I = 1\}$
- Low frequency in that I only varies from year to year, *not* day to day

- **Conditional chain-dependent process**

- Conditional on $I = i$

Markov chain for daily occurrence has transition probs. $P_{01}(i), P_{11}(i)$ (or π_i, d_i)

Daily intensities i.i.d. with mean μ_i , variance σ_i^2

Properties of Low Frequency Model

- Variance of total precipitation

$$\begin{aligned}\text{Var}(S_T) &= (1 - w) \text{Var}(S_T | I = 0) + w \text{Var}(S_T | I = 1) \\ &+ w(1 - w) \{E(S_T | I = 1) - E(S_T | I = 0)\}^2\end{aligned}$$

Here

$$E(S_T | I = i) = T\pi_i \mu_i$$

$$\text{Var}(S_T | I = i) \approx T\{\pi_i \sigma_i^2 + \pi_i(1 - \pi_i) [(1 + d_i)/(1 - d_i)] \mu_i^2\}$$

- **Interpretation**

-- **Autocorrelation function for single chain-dependent process:**

$$\text{Corr}(X_t, X_{t+l}) = [\pi(1 - \pi) d^l \mu^2] / \text{Var}(X_t)$$

where

$$\text{Var}(X_t) = \pi\sigma^2 + \pi(1 - \pi)\mu^2$$

-- **Autocovariance function for mixture of two chain-dependent processes:**

$$\begin{aligned} \text{Cov}(X_t, X_{t+l}) = & (1 - w) \pi_0(1 - \pi_0) d_0^l \mu_0^2 + w \pi_1(1 - \pi_1) d_1^l \mu_1^2 \\ & + w(1 - w) (\pi_1 \mu_1 - \pi_0 \mu_0)^2 \end{aligned}$$

Resembles second-order Markov chain, along with another term for shift in conditional mean of daily precipitation

(3) Low Frequency Approach: Observed Mixture/Covariate

Parameter estimation straightforward for observed covariate, but useful in formulating **EM** algorithm for hidden variable case

- **Parameter Estimation & Model Selection**
- **Application to Chico Precipitation**
 - Also 13 locations across California
- **Overdispersion Results**

Parameter Estimation and Model Selection

- **Parameter estimation**

- **Occurrences**

MLE's for transition probs. of Markov chain are based on transition counts:

e.g., $n_{01}/(n_{00} + n_{01})$ is MLE for P_{01} (n_{jk} no. times state j is followed by state k)

- **Intensities**

Assume i.i.d. with power transform distribution:

$X_t^* = X_t^s$ normally distributed with mean μ^* and variance $(\sigma^*)^2$ (e.g., $s = 1/4$)

MLE's just sample means and variances of transformed intensities

Log likelihood function for chain-dependent process:

$$\log L[P_{01}, P_{11}, \mu^*, (\sigma^*)^2] = \sum_j [n_{j0} \log(1 - P_{j1}) + n_{j1} \log P_{j1}] \\ - (n_{\cdot 1}/2) \log[2\pi(\sigma^*)^2] - \{1/[2(\sigma^*)^2]\} \sum_t (x_t^* - \mu^*)^2$$

where $n_{\cdot 1} = n_{01} + n_{11}$

\sum_j is over $j = 0, 1$

\sum_t is over $n_{\cdot 1}$ terms for which $x_t > 0$

- **Application to Chico January Precipitation**

- Observed covariate**

January mean sea level pressure at grid point off coast of California:

$I = 1$ if greater than sample mean, $I = 0$ otherwise

- Model fitting for Chico**

Conditioning	π	d	μ (mm)	σ (mm)	μ^*	σ^*
None	0.329	0.359	13.36	14.68	1.699	0.521
$I = 0$	0.413	0.334	15.16	15.50	1.777	0.515
$I = 1$	0.253	0.349	10.75	12.99	1.585	0.509

-- Model selection for Chico

Issue of how to choose “sample size” n for **BIC**:

Take $n = 78$ years (not $n = 78 \cdot 31 = 2418$ days)

More parsimonious to parameterize in terms of π, d (instead of P_{01}, P_{11})

AIC & BIC both select model in which π & μ^* varied with circulation index I

January precipitation at 13 sites across California (including Chico), still conditioning on same covariate: no. sites for which vary parameter with I

	π	d	μ^*	σ^*
AIC	12	4	9	1
BIC	9	1	4	0

- **Overdispersion Results**

- **Chico Jan. precipitation (St. dev. of monthly total)**

Single process **69.2 mm** (**Observed 88.6 mm**)

Mixture **86.8 mm** (**Observed 88.6 mm**)

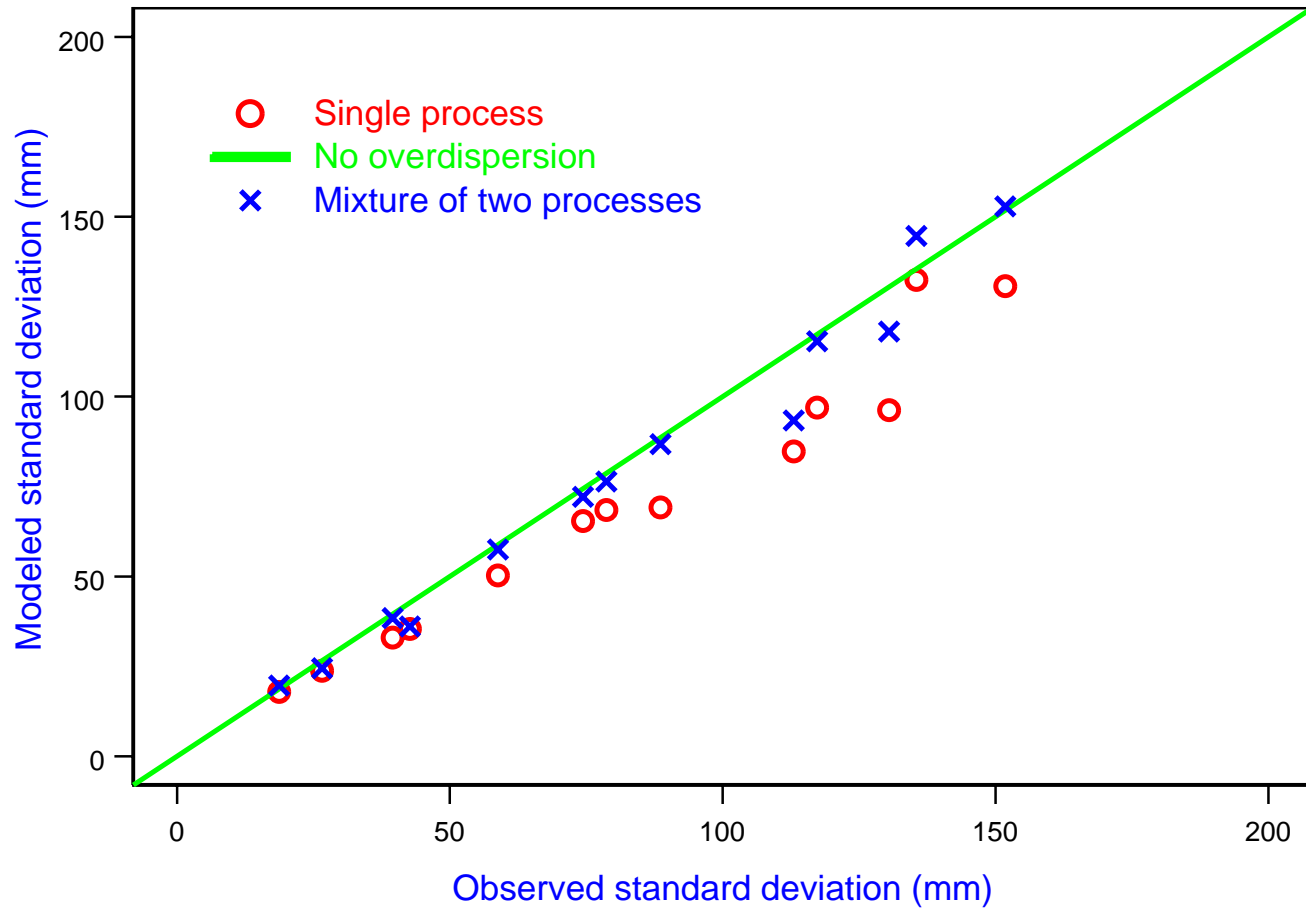
Given $I = 0$ **80.8 mm** (**Observed 82.4 mm**)

Given $I = 1$ **52.2 mm** (**Observed 58.9 mm**)

- **All 13 sites in California**

Apparent reduction in extent of overdispersion for majority of sites

Standard deviation of total precipitation in January (13 sites)



(4) Low Frequency Approach: Hidden Mixture/Variable

Treat conditioning variable as hidden (ignore any observed covariates)

- **Formulation of EM algorithm**
- **Model Fitting & Selection**
- **Overdispersion Results**

Formulation of EM Algorithm

- Complete-data likelihood function

Observed data:

$\{x_t(m): t = 1, 2, \dots, T; m = 1, 2, \dots, M\}$ prec. amount on t th day of m th yr.

Unobservable data:

$\{i(m), m = 1, 2, \dots, M\}$, $i(m) = 0, 1$ denoting index state for m th year

Vector of parameters: $\Theta = (w, \Theta_0, \Theta_1)$

where $\Theta_i = [P_{01}(i), P_{11}(i), \mu_i^*, (\sigma_i^*)^2]$, $i = 0, 1$

Log-likelihood function (complete data):

$$\log L_C(\Theta) = \sum_m \{ [1 - i(m)] \log[(1 - w) L_m(\Theta_0)] + i(m) \log[w L_m(\Theta_1)] \}$$

where $L_m(\Theta_i)$ denotes likelihood function for chain-dependent process given index state i evaluated for daily precipitation time series in m th year

MLEs for model parameters (complete data):

$$w: \quad [\sum_m i(m)] / M$$

$$P_{j1}(1), j = 0, 1: \quad [\sum_m i(m) n_{j1}(m)] / [\sum_m i(m) n_{j.}(m)]$$

$$\mu_1^*: \quad [\sum_m i(m) s_1(m)] / [\sum_m i(m) n_{.1}(m)]$$

$$(\sigma_1^*)^2: \quad [\sum_m i(m) s_2(m)] / [\sum_m i(m) n_{.1}(m)] - (\mu_1^*)^2$$

Here $s_1(m) = \sum_t x_t(m), s_2(m) = \sum_t [x_t(m)]^2$

EM algorithm:

(1) E-step

-- Estimate posterior probabilities by

$$\Pr\{I(m) = 1\} = w L_m(\Theta_1) / [(1 - w) L_m(\Theta_0) + w L_m(\Theta_1)],$$

$$m = 1, 2, \dots, M$$

(2) M-step

-- Replace $i(m)$ with estimated posterior probability in expressions of MLE's for complete data case

Model Fitting and Selection

- **Results for Chico**

- **Fitted models**

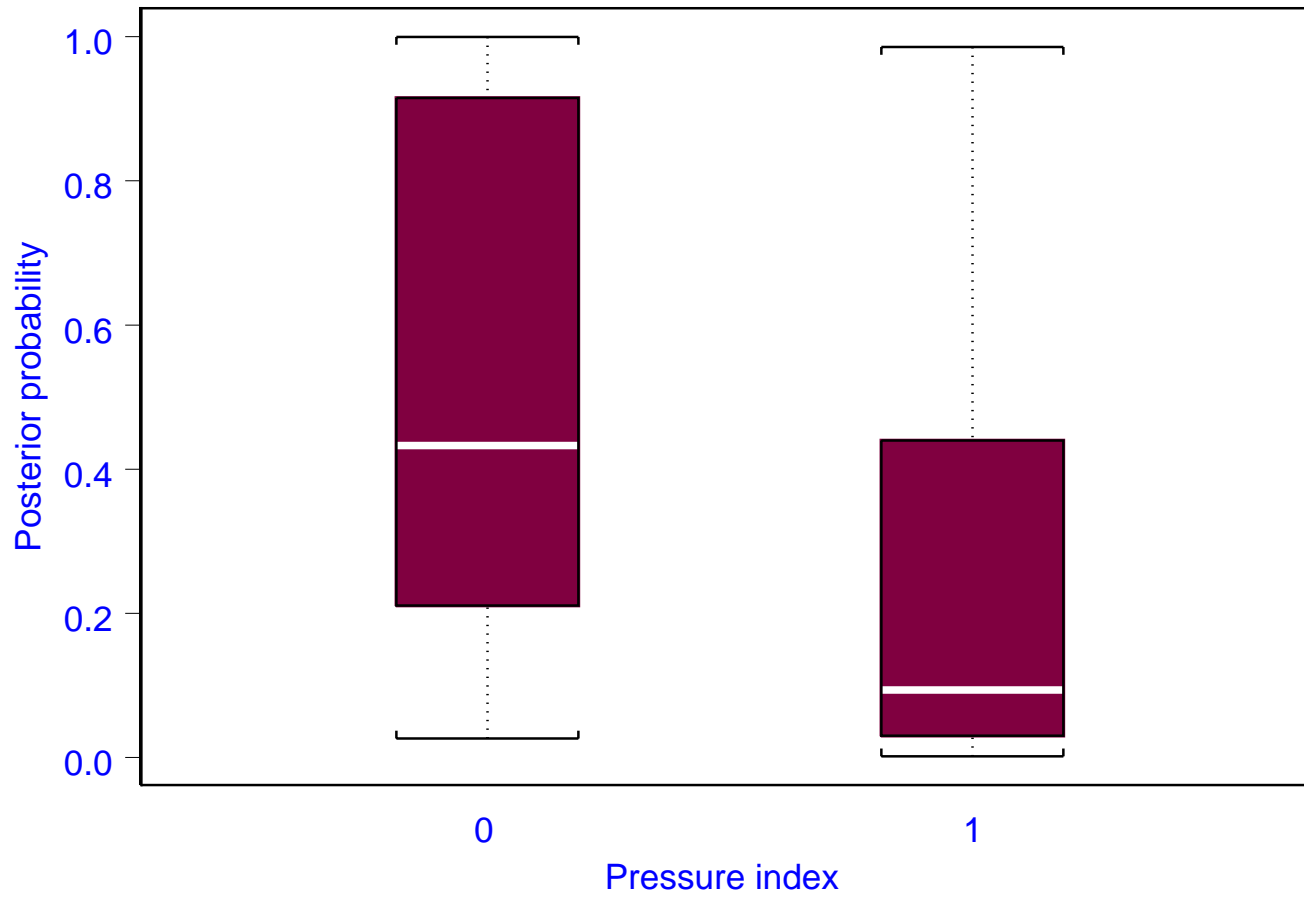
Model	w	$P_{01}(i),$ $i = 0, 1$	$P_{11}(i),$ $i = 0, 1$	$\mu_i^*,$ $i = 0, 1$	$\sigma_i^*,$ $i = 0, 1$
Completely constrained	-----	0.211	0.571	1.702	0.521
		0.211	0.571	1.702	0.521
$P_{01}(0) = P_{01}(1)$ $\sigma_0^* = \sigma_1^*$	0.371	0.211	0.505	1.585	0.503
		0.211	0.660	1.859	0.503
Completely unconstrained	0.378	0.214	0.505	1.587	0.504
		0.205	0.662	1.861	0.502

-- Model selection

Again taking $n = 78$ years

Model	Number parameters	Log likelihood	AIC	BIC
Completely constrained	4	-1924.246	3856.49	3865.92
$P_{01}(0) = P_{01}(1)$ $\sigma_0^* = \sigma_1^*$	7	-1912.393	3838.79	3855.28
Completely unconstrained	9	-1912.386	3842.77	3863.98

Chico Jan. Prec.: Posterior probability of hidden state 1



Overdispersion Results

- **Estimated variance of Chico January total precipitation:**

Completely constrained **70.4 mm**

“Optimal model” **89.4 mm**

Completely unconstrained **88.9 mm**

Observed **88.6 mm**