STAT 431 — Applied Bayesian Analysis — Course Notes

Additional Topics

Spring 2019

Time Series

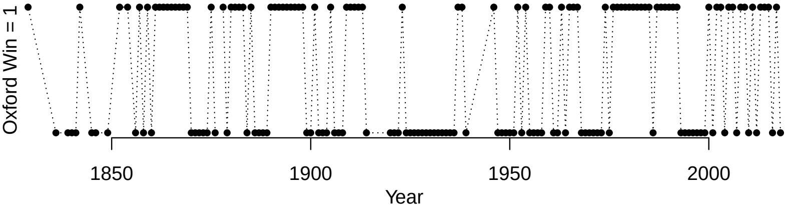
When observations occur over time, they may be correlated.

Temporal correlation structure is more complicated than for a simple random-effects model:

Observations closer in time are usually more correlated than those further apart (like successive values from a Markov chain).

Example: Oxford/Cambridge Men's Boat Race

Consider the sequence of annual outcomes for Oxford:



Wins and losses sometimes seem to come in "streaks" — an outcome one year is often the same as the previous year.

Is there evidence for “streaks” of wins/losses?

Might they influence predictions of the winner?

Maybe we can use a GLM (with binary response), but we need a way to model possible dependence: correlations that depend on time lag.

We will use a *latent variable* model: Win probabilities will depend on an underlying *continuous* stochastic process.

Gaussian processes are convenient candidates ...

Classical *time-series analysis* models consider a sequence

$$Z_1,\ Z_2,\ Z_3,\ \ldots$$

of *normally-distributed* random variables.

For simplicity, take them to be identically distributed with mean zero:

$$Z_t \quad \sim \quad \mathrm{N}(0, \sigma^2)$$

They are not assumed independent.

Their correlation structure generally follows a model ...

The AR(1) Model

A simple model is the **autoregressive model of order 1 (AR(1))**:

$$Z_t \;=\; \rho Z_{t-1} \,+\, \varepsilon_t, \qquad\qquad t = 2, 3, 4, \ldots$$

where

$$\varepsilon_t \;\sim\; \text{i.i.d. N}(0, \sigma_\varepsilon^2)$$

and ρ and $\sigma_\varepsilon^2 > 0$ are parameters.

$$Z_t \;=\; \rho Z_{t-1} \,+\, \varepsilon_t$$

Consider what happens when

- $\rho > 0$
- $\rho < 0$
- $\rho = 0$

We must also specify

$$Z_1 \quad \sim \quad \mathrm{N}(0, \sigma^2)$$

and that Z_1 is independent of $\{ \varepsilon_2, \varepsilon_3, \varepsilon_4, \ldots \}$.

We must also specify

$$Z_1 \sim \mathrm{N}(0, \sigma^2)$$

and that Z_1 is independent of $\{ \varepsilon_2, \varepsilon_3, \varepsilon_4, \ldots \}$.

Since all Zs have the same marginal distribution,

$$\begin{aligned} \sigma^2 = \operatorname{Var}(Z_2) &= \operatorname{Var}(\rho Z_1 + \varepsilon_2) \\ &= \rho^2 \operatorname{Var}(Z_1) + \operatorname{Var}(\varepsilon_2) \\ &= \rho^2 \sigma^2 + \sigma_\varepsilon^2 \end{aligned}$$

which implies

$$-1 < \rho < 1 \qquad\qquad \sigma_\varepsilon^2 = (1 - \rho^2)\sigma^2$$

Overall, we have

$$\left.\begin{array}{rcl} Z_1 \mid \sigma^2 & \sim & \mathrm{N}(0, \sigma^2) \\[2ex] \varepsilon_t \mid \rho, \sigma^2 & \sim & \text{i.i.d. } \mathrm{N}\big(0, (1-\rho^2)\sigma^2\big) \end{array}\right\} \text{ conditionally independent}$$

$$Z_t \;=\; \rho Z_{t-1} + \varepsilon_t$$

$$t = 2, 3, 4, \ldots$$

Overall, we have

$$\left.\begin{array}{rcl} Z_1 \mid \sigma^2 & \sim & \mathrm{N}(0, \sigma^2) \\[1em] \varepsilon_t \mid \rho, \sigma^2 & \sim & \text{i.i.d. } \mathrm{N}\big(0, (1-\rho^2)\sigma^2\big) \end{array}\right\} \text{ conditionally independent}$$

$$Z_t \;=\; \rho Z_{t-1} + \varepsilon_t$$

$$t = 2, 3, 4, \ldots$$

Notice that Z_{t-1} is a function of $\varepsilon_{t-1}, \ldots, \varepsilon_2$ and Z_1.

So ε_t is independent of Z_{t-1}.

Thus,

$$Z_t \mid Z_{t-1}, \rho, \sigma^2 \;\sim\; \mathrm{N}\big(\rho Z_{t-1},\, (1-\rho^2)\sigma^2\big)$$

An alternative way to express the AR(1) model:

$$\left.\begin{array}{rcl} Z_1 \mid \sigma^2 & \sim & \mathrm{N}(0, \sigma^2) \\[2ex] Z_2 \mid Z_1, \rho, \sigma^2 & \sim & \mathrm{N}\big(\rho Z_1,\ (1-\rho^2)\sigma^2\big) \\[2ex] \vdots & & \vdots \\[2ex] Z_t \mid Z_{t-1}, \rho, \sigma^2 & \sim & \mathrm{N}\big(\rho Z_{t-1},\ (1-\rho^2)\sigma^2\big) \\[2ex] \vdots & & \vdots \end{array}\right\} \begin{array}{c}\text{all}\\ \text{cond'l.}\\ \text{indep.}\end{array}$$

By the way, autocorrelations in the AR(1) model are

$$\mathrm{Corr}(Z_s, Z_t) \quad = \quad \rho^{|s-t|}$$

So, for example, two successive values have correlation ρ. This is the *lag-1 autocorrelation*.

Latent Time-Series Regression

As in the boat race example, suppose we observe a Bernoulli time series

$$Y_t \mid \pi_t \;\sim\; \text{Bernoulli}(\pi_t) \qquad t = 1, 2, \ldots$$

We will incorporate all time dependence into the probabilities π_t, so that the Y_t variables are conditionally independent given these probabilities:

$$Y_t \mid \pi_t \;\sim\; \text{indep. Bernoulli}(\pi_t)$$

How can we make the probabilities π_t time-series dependent?

First, transform the probabilities to an unbounded scale, then use a Gaussian time series model for the transformed probabilities.

How can we make the probabilities π_t time-series dependent?

First, transform the probabilities to an unbounded scale, then use a Gaussian time series model for the transformed probabilities.

For example, how about

$$\text{logit}(\pi_t) \;\; = \;\; Z_t$$

where $Z_1, Z_2, \ldots$ follow an AR(1) model?

Then $\rho > 0$ would tend to give "streaks" of wins/losses. (Why?)

If the time series $Z_1, Z_2, \ldots$ has zero mean, we might want to add a possibly nonzero mean:

$$\text{logit}(\pi_t) \;\; = \;\; \beta_0 \, + \, Z_t$$

(This allows the average probability to be something other than $\text{logit}^{-1}(0) = 0.5$.)

If the time series $Z_1, Z_2, \ldots$ has zero mean, we might want to add a possibly nonzero mean:

$$\text{logit}(\pi_t) \;\; = \;\; \beta_0 \, + \, Z_t$$

(This allows the average probability to be something other than $\text{logit}^{-1}(0) = 0.5$.)

We can even add a predictor variable X:

$$\text{logit}(\pi_t) \;\; = \;\; \beta_0 \, + \, \beta_1(x_t - \bar{x}) \, + \, Z_t$$

(a kind of *logistic-normal* mixed regression model)

For example, for the Boat Race data, to model an overall time trend, take

$$x_t \quad = \quad \text{year for time index } t$$

$(x_1 = 1829,\ x_2 = 1830,\ \ldots)$

Assigning priors to β_0, β_1, ρ, and σ^2 requires caution because

- logistic regression coefficients with flat priors can lead to improper posteriors
- when $\rho \approx 0$, σ^2 is nearly unidentifiable

We must make sure the priors on β_0, β_1, and σ^2 are not too vague.

We will try independent (hyper)priors

$$\begin{aligned}
\beta_0 &\sim t(0, 100, 1) \\
\beta_1 &\sim t(0, 2, 1) \\
\rho &\sim \mathrm{uniform}(-1, 1) \\
\sigma^2 &\sim \mathrm{IG}(2.5, 1.5)
\end{aligned}$$

after *standardizing* the predictor X.

Important: JAGS code uses a *reciprocal* scale parameter for the t distribution (like using precision instead of variance).

Note:

The Boat Race data has some gaps (years with no race or with a tie).

For simplicity, we fill in these years with a `NA` response.

JAGS will simulate these missing responses from their posterior predictive distribution.

R/JAGS Extra Example 1:

Latent Logistic-Normal Time-Series Regression

Outliers

What if data suggest that a stochastic value in a model is unusual, relative to what is typical for the model?

It could be a data value y_i, or it could be a parameter θ_j in a hierarchical model.

In either case, it could be called an **outlier**.

Possible reasons: mistakes in the data, unused explanatory variables, mis-specified distributions

Typically, outliers in a quantitative variable are extreme (very large or small) compared to what the posterior would predict.

Models that use the normal distribution are especially susceptible to outliers because the normal density has **light (short) tails**:

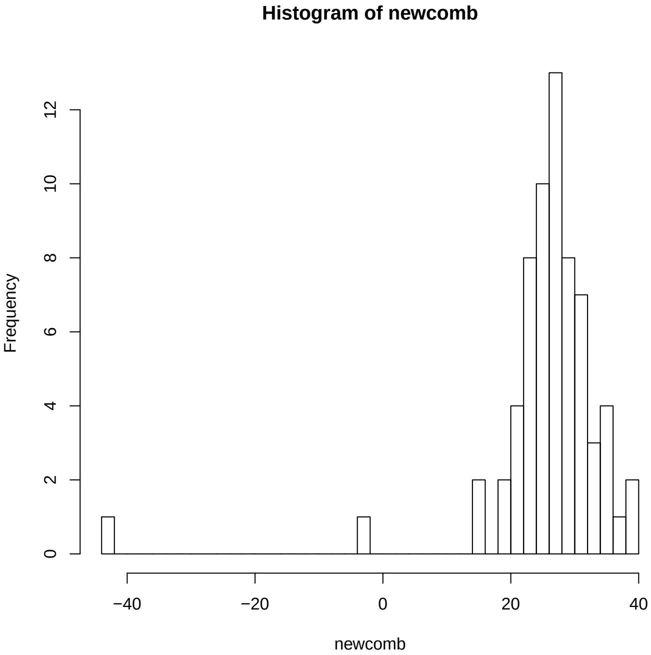
k	Prob. a $\mathrm{N}(\mu, \sigma^2)$ variable is $> k\sigma$ from μ
1	0.3173105
2	0.0455003
3	0.0026998
4	0.0000633
5	0.0000006

Newcomb Data

The `newcomb` data set comes from an early (1882) experiment to determine the speed of light. The 66 measurements are shifted and scaled times for light to travel the same known distance (in air).

Continuous measurements like these would typically be modeled with a normal distribution.

But the `newcomb` data has obvious outliers, relative to what would be expected under a normal distribution ...


Histogram of newcomb
Frequency
0
2
4
6
8
10
12
−40
−20
0
20
40
newcomb

Robustness

For models with standard distributions (e.g. normal), outliers can be highly influential — removing them disproportionately affects the inference.

We seek models that are more **robust**: less sensitive to extreme values of a few observations (or parameters).

A robust model can be used as an alternative to a standard model, either in sensitivity analysis, or for improved inference.

t Distribution

Recall the (location-scale) t distribution $t(\mu, \sigma^2, \nu)$, with continuous density

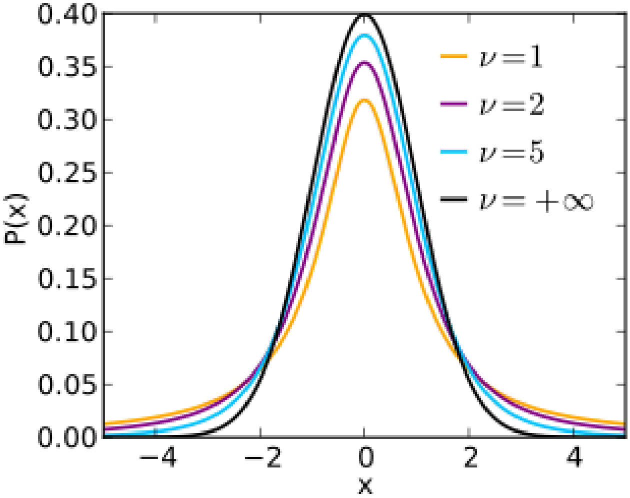
$$p(x) \quad \propto \quad \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \qquad \text{for all } x$$

The mean is μ when $\nu > 1$. Otherwise, the mean is undefined, but μ is still the median.

The variance is

$$\frac{\nu}{\nu - 2}\,\sigma^2$$

if $\nu > 2$, and undefined otherwise.



From: Student's t-distribution. (2017, November 30). In *Wikipedia, The Free Encyclopedia*. Retrieved November 30, 2017, from `https://en.wikipedia.org/w/index.php?title=Student%27s_t-distribution&oldid=812931585`

For small ν, the t distribution has **heavy (long) tails**, which are thicker than those of a normal.

For example, consider $\nu = 2$:

k	Prob. a $t(\mu, \sigma^2, 2)$ variable is $> k\sigma$ from μ
1	0.4226
2	0.1835
3	0.0955
4	0.0572
5	0.0377

For larger ν, the tails become lighter, and the distribution becomes more like a normal. In fact,

$$t(\mu, \sigma^2, \nu) \quad \underset{\nu \to \infty}{\longrightarrow} \quad \mathrm{N}(\mu, \sigma^2)$$

So a t distribution could have heavy or light tails, as controlled by ν.

Note: ν need not be an integer — it can be any positive value.

Heavy Tails and Outliers

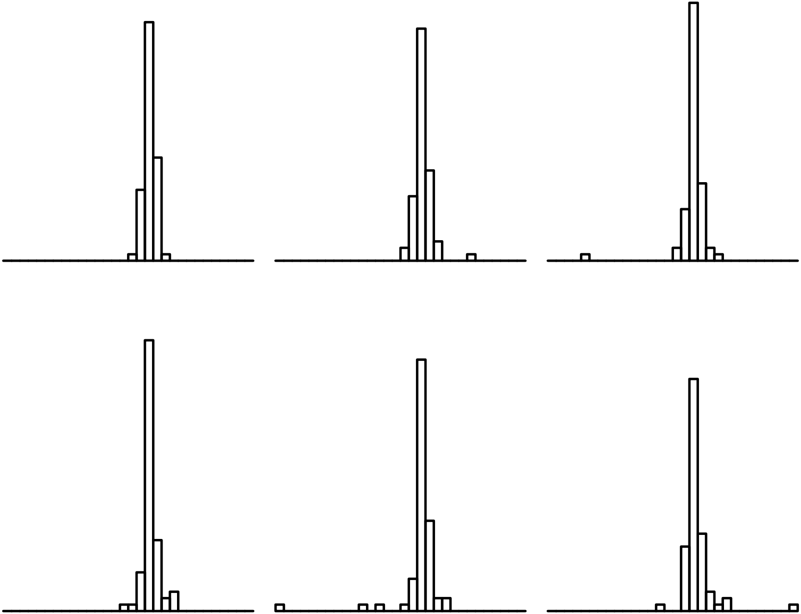
A continuous distribution with heavy tails tends to produce samples that have outliers, relative to a normal distribution.

For example, let's simulate 6 samples of size 66 from $t(0, 1, 2)$:

```
> n <- 66

> tsamp <- matrix(rt(6*n, 2), n, 6)
```

Now make a histogram of each column ...



Modeling with the t Distribution

When the ideal model would be a normal distribution, but the data have outliers, a (location-scale) t distribution is a viable alternative.

Using the t requires a way to handle the degrees of freedom parameter $\nu > 0$. Smaller values give heavy tails, larger values give light tails.

A Bayesian model for a t-distributed sample:

$$\begin{aligned} Y_1, \ldots, Y_n \mid \mu, \sigma^2 &\sim \text{i.i.d. } t(\mu, \sigma^2, \nu) \\ \mu &\sim 1\, d\mu & -\infty < \mu < \infty \\ \sigma^2 &\sim (\sigma^2)^{-1}\, d\sigma^2 & \sigma^2 > 0 \end{aligned}$$

ν is chosen arbitrarily, in this case.

Note: Using same improper joint prior as for a two-parameter normal sample.

Provided $\nu \geq 1$ and all $n > 1$ values y_i are distinct, the posterior is proper.

(Otherwise, be careful!)

Generally, we consider only $\nu > 1$, to retain interpretation of the mean (and to avoid possible issues with propriety and computation).

Random ν

To avoid specifying ν, we can give it a prior.

Because t converges to normal as $\nu \to \infty$, instead consider the reciprocal $1/\nu$:

$$0 \ < \ 1/\nu \ < \ 1$$

The lower bound represents the normal and the upper bound the Cauchy.

One textbook suggests

$$1/\nu \ \sim \ \text{uniform}(0, 1)$$

The resulting Bayesian model for a t-distributed sample:

$$\begin{aligned}
Y_1, \ldots, Y_n \mid \mu, \sigma^2, \nu &\sim \text{i.i.d. } t(\mu, \sigma^2, \nu) \\
\mu &\sim 1\, d\mu & -\infty < \mu < \infty \\
\sigma^2 &\sim (\sigma^2)^{-1}\, d\sigma^2 & \sigma^2 > 0 \\
1/\nu &\sim \text{uniform}(0, 1)
\end{aligned}$$

A Trick for Semi-Conjugacy

What if we want to use *proper* priors, instead?

The t distribution has no obvious conjugate or semi-conjugate priors.

However, we can recover semi-conjugacy for μ and σ^2 by using a special representation of the t ...

Recall: If

$$\begin{aligned} X \mid W = w &\sim \mathrm{N}(\mu_0, w/\kappa) \\ W &\sim \mathrm{IG}(\alpha, \beta) \end{aligned}$$

then

$$X \sim t\bigl(\mu_0,\, \beta/(\alpha\kappa),\, 2\alpha\bigr)$$

Now take

$$\begin{array}{ccc} X = Y_i & W = W_i & \mu_0 = \mu \\ \kappa = 1 & \alpha = \nu/2 & \beta = \nu\sigma^2/2 \end{array}$$

to get that ...

$$Y_i \mid \mu, \sigma^2, \nu \;\; \sim \;\; t(\mu, \sigma^2, \nu)$$

if

$$\begin{aligned} Y_i \mid W_i = w_i, \mu \;\; &\sim \;\; \mathrm{N}(\mu, w_i) \\ W_i \mid \sigma^2, \nu \;\; &\sim \;\; \mathrm{IG}(\nu/2,\, \nu\sigma^2/2) \end{aligned}$$

This represents the t distribution of Y_i as a *scale mixture of normals*, by introducing a new latent variable W_i.

In this new hierarchical representation, μ is just a normal mean. It can be shown that

$$\mu \;\; \sim \;\; \mathrm{N}(\mu_0, \sigma_0^2)$$

is semi-conjugate.

What about σ^2?

The hierarchy involves σ^2 only as a factor in the second ("β") parameter of an inverse gamma distribution.

It can be shown that the *gamma* distribution is semi-conjugate for that parameter, and hence (by scaling) for σ^2:

$$\sigma^2 \quad \sim \quad \mathrm{gamma}(\alpha_0, \beta_0)$$

Unfortunately, there is no natural continuous semi-conjugate prior for ν.

Overall, the (partially) semi-conjugate hierarchy becomes

$$
\begin{aligned}
Y_i \mid W_i = w_i, \mu &\sim \text{indep. } \mathrm{N}(\mu, w_i) \\
W_i \mid \sigma^2, \nu &\sim \text{i.i.d. } \mathrm{IG}(\nu/2,\, \nu\sigma^2/2) \\
\left.\begin{aligned}
\mu &\sim \mathrm{N}(\mu_0, \sigma_0^2) \\
\sigma^2 &\sim \mathrm{gamma}(\alpha_0, \beta_0) \\
1/\nu &\sim \mathrm{uniform}(0,1)
\end{aligned}\quad\right\} &\ \text{independent}
\end{aligned}
$$

JAGS does not require us to use this hierarchy — the t distribution can be specified directly — but it motivates a reasonable choice of prior.

Example: Newcomb Data

The data are in the `newcomb` data set of R package `MASS`:

$$Y_i \;=\; i\text{th shifted, scaled speed of light measurement}$$

$$i \;=\; 1, \ldots, 66$$

$$\bar{y} \;\approx\; 26.212 \qquad \text{median} \;=\; 27$$

Without the two outliers:

$$\bar{y} \;=\; 27.75 \qquad \text{median} \;=\; 27.5$$

A JAGS model for the case of random ν:

```
model {
  for(i in 1:length(y)) {
    y[i] ~ dt(mu, 1/sigmasq, 1/nuinv)
    yrep[i] ~ dt(mu, 1/sigmasq, 1/nuinv)
  }

  mu ~ dnorm(0, 0.00000001)
  sigmasq ~ dgamma(0.00001, 0.00001)
  nuinv ~ dunif(0, 1)
}
```

Note inclusion of a replicate data vector `yrep` for posterior predictive assessment.

R/JAGS Extra Example 2:

Robust t Location-Scale Analysis

Remarks:

- Can extend to linear regression: Replace normally-distributed observations with t-distributed observations.
- Can use the t distribution in the prior portion of a hierarchy (e.g., if some random effects are outliers).