

# BAYESIAN ANALYSIS FOR THE SOCIAL SCIENCES

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 4, 2012

# Introduction to Bayesian Inference

- Bayesian inference relies exclusively on Bayes Theorem:

$$p(\theta|\text{data}) \propto p(\theta) p(\text{data}|\theta)$$

- $\theta$  is usually a parameter (but could also be a data point, a model, a hypothesis)
- $p$  are **probability densities** (or probability mass functions in the case of discrete  $\theta$  and/or discrete data)
- $p(\theta)$  a **prior** density;  $p(\text{data}|\theta)$  the **likelihood** or **conditional density** of the data given  $\theta$
- $p(\theta|\text{data})$  is the **posterior** density for  $\theta$  given the data.
- Gives rise to the ***Bayesian mantra***:

*a posterior density is proportional to the prior times the likelihood*

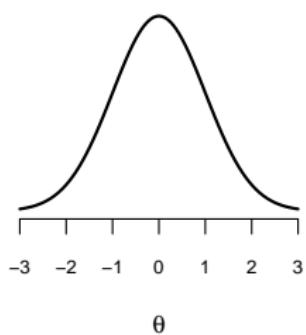
# Probability Densities as Representations of Beliefs

## Definition (Probability Density Function (informal))

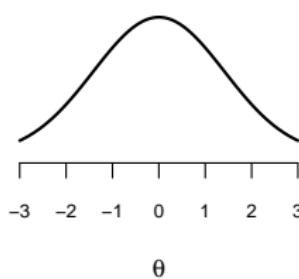
Let  $\theta$  be a unknown quantity,  $\theta \in \Theta \subseteq \mathbb{R}$ . A function  $p(\theta)$  is a proper probability density function if

- ①  $p(\theta) \geq 0 \forall \theta$ .
- ②  $\int_{\Theta} p(\theta) d\theta = 1$ .

$N(0,1)$



$N(0,2)$



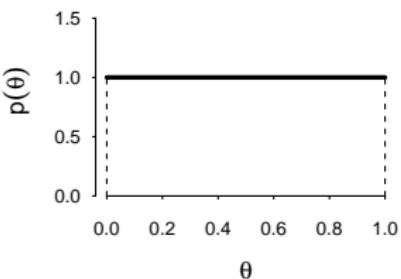
# Probability Densities as Representations of Beliefs

## Definition (Probability Density Function (informal))

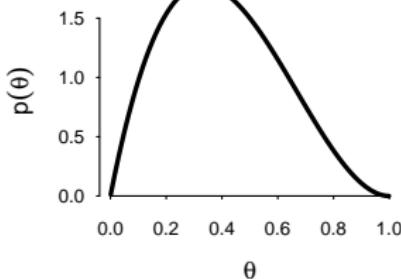
Let  $\theta$  be a unknown quantity,  $\theta \in \Theta \subseteq \mathbb{R}$ . A function  $p(\theta)$  is a proper probability density function if

- ①  $p(\theta) \geq 0 \forall \theta$ .
- ②  $\int_{\Theta} p(\theta) d\theta = 1$ .

Unif(0,1)



Beta(2,3)

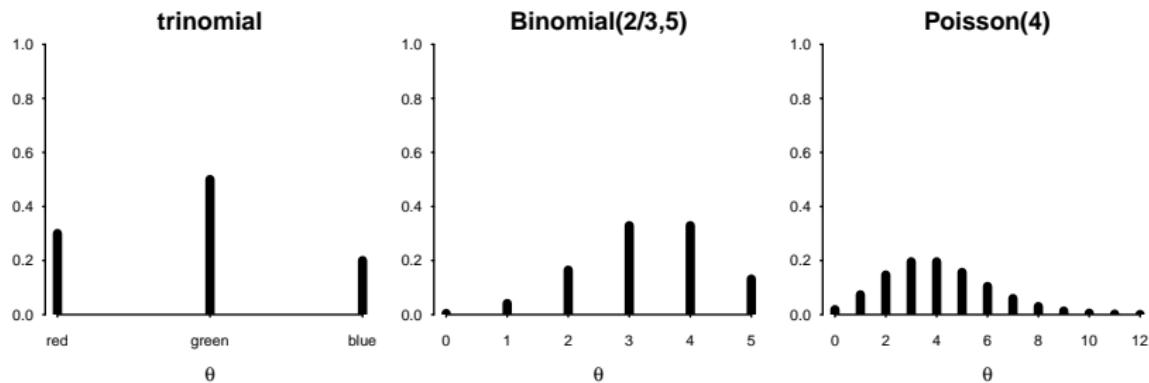


# Probability Mass Function

## Definition (Probability Mass Function)

If  $\theta$  is a discrete random variable, taking values in a countable space  $\Theta \subset \mathbb{R}$ , then a function  $p : \Theta \mapsto [0, 1]$  is a probability mass function if

- ①  $p(\theta) = 0 \forall \theta \in \mathbb{R} \setminus \Theta$
- ②  $\sum_{\theta \in \Theta} p(\theta) = 1$



# Introduction to Bayesian Inference

$$p(\theta|\text{data}) \propto p(\theta) p(\text{data}|\theta)$$

- Bayesian inference involves **computing, summarizing and communicating** summaries of the **posterior density**  $p(\theta|\text{data})$ .
- How to do this is **what this class is about**.
- Depending on the problem, doing all this is easy or hard; we solve “hard” with computing power.
- We’re working with densities (or sometimes, mass functions).
- Bayesian point estimates are a single number summary of a posterior density
- Uncertainty assessed/communicated in various ways: e.g., the standard deviation of the posterior, width of interval spanning 2.5th to 97.5th percentiles of the posterior, etc.
- Sometimes, can just draw a picture; details, examples coming.

# Introduction to Bayesian Inference

$$p(\theta|\text{data}) \propto p(\theta) p(\text{data}|\theta)$$

- Bayes Theorem tells us how to *update* beliefs about  $\theta$  in light of evidence (“data”)
- a general method for *induction* or for “learning from data”:  
prior → data → posterior
- Bayes Theorem is itself uncontroversial: follows from widely accepted axioms of probability theory (e.g., Kolmogorov) and the definition of conditional probability

# Why Be Bayesian?

- **conceptual simplicity:** “say what you mean” and “mean what you say” (subjective probability)
- a foundation for inference that does not rest on the thought experiment of repeated sampling
- **uniformity of application:** no special tweeks for this or that data analysis. Apply Bayes Rule.
- **modern computing** makes Bayesian inference easy and nearly universally applicable

# Conceptual Simplicity

$$p(\theta|\text{data}) \propto p(\theta) p(\text{data}|\theta)$$

- the posterior density (or mass function)  $p(\theta|\text{data})$  is a complete characterization of beliefs after looking at data
- as such it contains everything we need for making inferences
- Examples:
  - the posterior probability that a regression coefficient is positive, negative or lies in a particular interval;
  - the posterior probability that a subject belongs to a particular latent class;
  - the posterior probability that a hypothesis is true; or,
  - the posterior probabilities that a particular statistical model is true model among a family of statistical models.

# Contrast Frequentist Inference

- Model for data:  $y \sim f(\theta)$ .
- Estimate  $\theta$ : e.g., least squares, MLE, etc, to yield  $\hat{\theta} \equiv \hat{\theta}(y)$
- null hypothesis e.g.,  $H_0 : \theta_{H_0} = 0$
- Inference via the *sampling distribution* of  $\hat{\theta}$  conditional on  $H_0$ : e.g.,

assuming  $H_0$ , over repeated applications of the sampling process, *how frequently* would we observe a result *at least as extreme* as the one we obtained?

- “At least as extreme”? Assessed via a test statistic, e.g.,

$$t(y) = (\theta_{H_0} - \hat{\theta}) / \sqrt{\text{var}(\hat{\theta} | \theta = \theta_{H_0})}$$

- “how frequently”? The *p*-value, relative frequency with which we see  $|t| > t(y)$  in *repeated applications of the sampling process*. Often  $t(y) \xrightarrow{d} N(0, 1)$ .

# Contrast Frequentist Inference

- null hypothesis e.g.,  $H_0 : \theta_{H_0} = 0$
- test-statistic:

$$t(y) = (\theta_{H_0} - \hat{\theta}) / \sqrt{\text{var}(\hat{\theta} | \theta = \theta_{H_0})}$$

- Often  $t(y) \xrightarrow{d} N(0, 1)$ .
- $p$ -value is a statement about the plausibility of the statistic  $\hat{\theta}$  relative to what we might have observed in random sampling assuming  $H_0 : \theta_{H_0} = 0$
- one more step need to reject/fail-to-reject  $H_0$ . Is  $p$  sufficiently small?
- frequentist  $p$ -value is a summary of the distribution of  $\hat{\theta}$  under  $H_0$

# Contrast Frequentist Inference

- n.b., frequentist inference treats  $\hat{\theta}$  as a random variable
- $\theta$  is a fixed but unknown feature of the population from which data is being (randomly) sampled
- Bayesian inference:  $\hat{\theta}$  is fixed, a function of the data available for analysis
- Bayesian inference:  $\theta$  is a random variable, subject to (subjective) uncertainty

	Bayesian	Frequentist
$\theta$	random	fixed but unknown
$\hat{\theta}$	fixed	random
“random-ness”	subjective	sampling
distribution of interest	posterior $p(\theta y)$	sampling distribution $p(\hat{\theta}(y) \theta = \theta_{H_0})$

# Subjective Uncertainty

- how do we do statistical inference in situations where repeated sampling is infeasible?
- inference when we have the entire population and hence no uncertainty due to sampling: e.g., parts of comparative political economy.
- Bayesians rely on a notion of *subjective* uncertainty
- e.g.,  $\theta$  is a random variable because we don't know its value
- Bayes Theorem tells us how to manage that uncertainty, how to update beliefs about  $\theta$  in light of data
- Contrast objectivist notion of probability: probability as a property of the object under study (e.g., coins, decks of cards, roulette wheels, people, groups, societies).

# Subjective Uncertainty

Many Bayesians regard objectivist probability as **metaphysical nonsense**.  
de Finetti:

## *PROBABILITY DOES NOT EXIST*

*The abandonment of superstitious beliefs about...Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is not less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs. In investigating the reasonableness of our own modes of thought and behaviour under uncertainty, all we require, and all that we are reasonably entitled to, is consistency among these beliefs, and their reasonable relation to any kind of relevant objective data (“relevant” in as much as subjectively deemed to be so). This is Probability Theory.*

# Subjective Uncertainty

- Bayesian probability statements are thus about states of mind over states of the world, and not about states of the world *per sé*.
- Borel: one can guess the outcome of a coin toss while the coin is still in the air and its movement is perfectly determined, or even after the coin has landed but before one reviews the result.
- i.e., subjective uncertainty obtains irrespective of “objective uncertainty (however conceived)”
- not just any subjective uncertainty: beliefs must conform to the rules of probability: e.g.,  $p(\theta)$  should be *proper*: i.e.,  $\int_{\Theta} p(\theta)d\theta = 1$ ,  $p(\theta) \geq 0 \forall \theta \in \Theta$ .

# Bayes Theorem

- Conditional probability: Let  $A$  and  $B$  be events with  $P(B) > 0$ . Then the conditional probability of  $A$  given  $B$  is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}.$$

- Multiplication rule:

$$P(A \cap B) = P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Law of Total Probability:

$$P(B) = P(A \cap B) + P(\sim A \cap B) = P(B|A)P(A) + P(B|\sim A)P(\sim A)$$

- Bayes Theorem: If  $A$  and  $B$  are events with  $P(B) > 0$ , then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Bayes Theorem, Example case, drug-testing

- Prior work suggests that about 3% of the subject pool (elite athletes) uses a particular prohibited drug.
- $H_U$ : test subject uses the prohibited substance.
- $p(H_U) = .03$ .
- $E$  (evidence) is a positive test result.
- Test has a false negative rate of .05; i.e.,  
 $P(\sim E|H_U) = .05 \Rightarrow P(E|H_U) = .95$ .
- Test has a false positive rate of .10: i.e.,  $P(E|H_{\sim U}) = .10$ .
- Bayes Theorem:

$$\begin{aligned}P(H_U|E) &= \frac{P(H_U)P(E|H_U)}{\sum_{i \in \{U, \sim U\}} P(H_i)P(E|H_i)} \\&= \frac{.03 \times .95}{(.03 \times .95) + (.97 \times .10)} \\&= \frac{.0285}{.0285 + .097} \\&= .23\end{aligned}$$

# Bayes Theorem, Continuous Parameter

- Bayes Theorem:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}$$

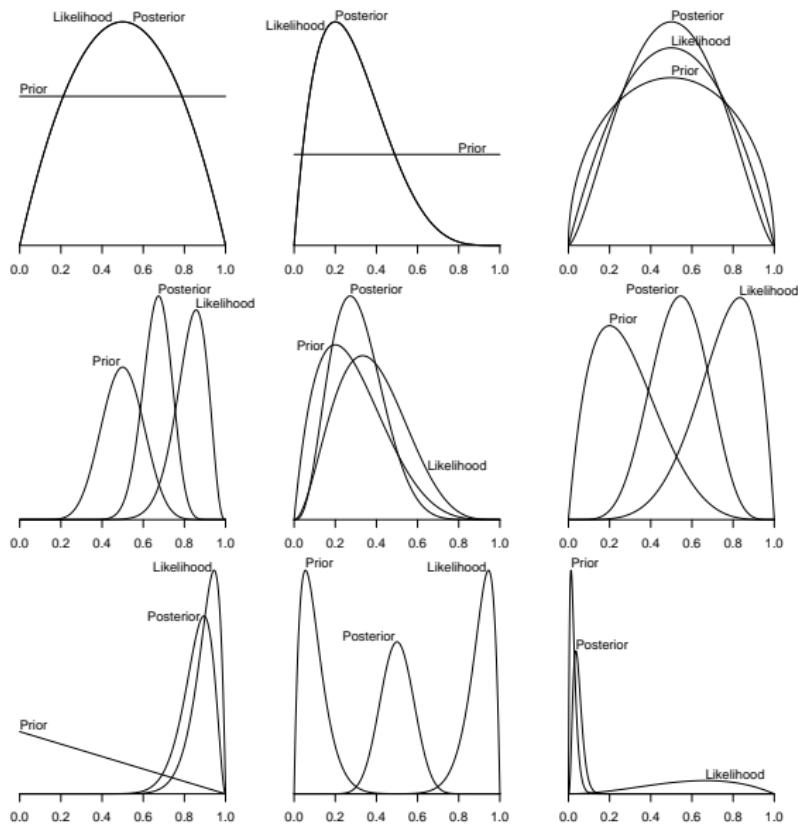
- Proof: by the definition of conditional probability

$$p(\theta, \mathbf{y}) = p(\theta|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta), \quad (1)$$

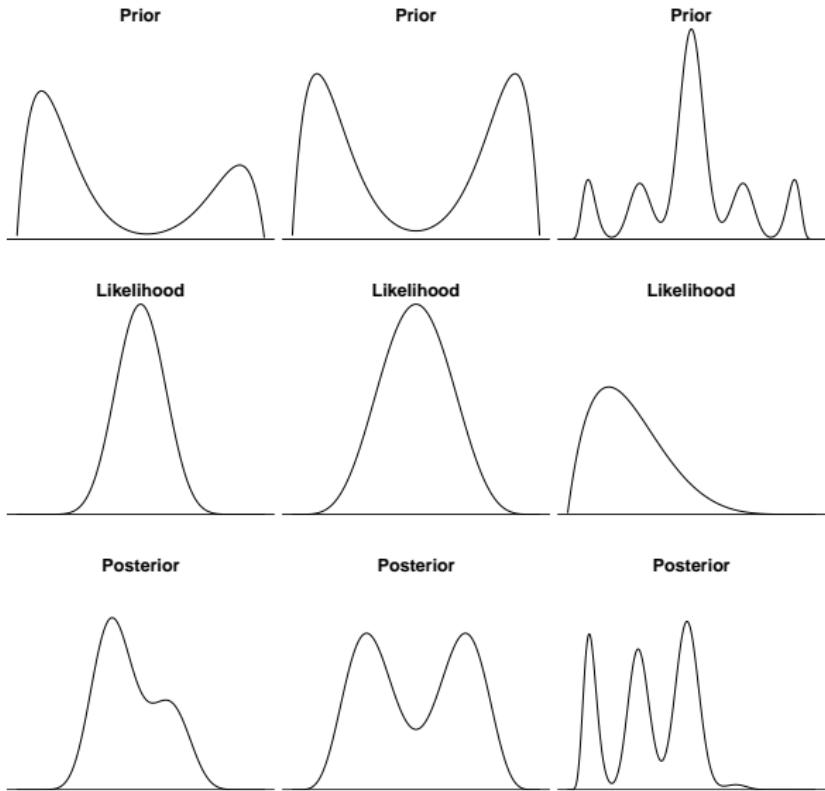
where all these densities are assumed to exist and have the properties  $p(z) > 0$  and  $\int p(z)dz = 1$  (i.e., are *proper* probability densities).

The result follows by re-arranging the quantities in equation equation 1 and noting that  $p(\mathbf{y}) = \int p(\mathbf{y}, \theta)d\theta = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ .

# Prior and Posterior Densities, Continuous Parameter



# Prior, Likelihood and Posteriors: less standard cases

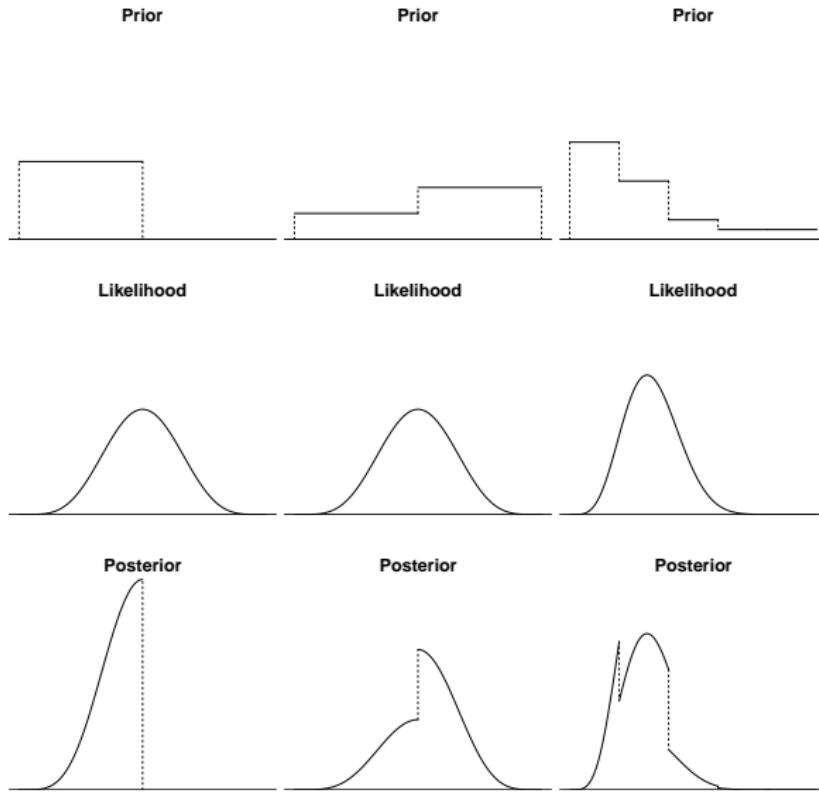


# Cromwell's Rule: the dangers of dogmatism

$$p(\theta|\text{data}) \propto p(\theta) p(\text{data}|\theta)$$

- $p(\theta|\text{data}) = 0 \forall \theta \text{ s.t. } p(\theta) = 0.$
- Cromwell's Rule: After the English deposed, tried and executed Charles I in 1649, the Scots invited Charles' son, Charles II, to become king. The English regarded this as a hostile act, and Oliver Cromwell led an army north. Prior to the outbreak of hostilities, Cromwell wrote to the synod of the Church of Scotland, "I beseech you, in the bowels of Christ, consider it possible that you are mistaken".
- a dogmatic prior that assigns zero probability to a hypothesis can never be revised
- likewise, a hypothesis with prior weight of 1.0 can never be refuted.

# Cromwell's Rule



# Bayesian Point Estimates

- **Bayes estimates:** single number summary of a posterior density
- but which one?: e.g., mode, median, mean, some quantile(s)?
- different loss functions rationalize different point estimate
- Loss: Let  $\Theta$  be a set of possible states of nature  $\theta$ , and let  $a \in \mathcal{A}$  be actions available to the researcher. Then define  $l(\theta, a)$  as the *loss* to the researcher from taking action  $a$  when the state of nature is  $\theta$ .
- Posterior expected loss: Given a posterior distribution for  $\theta$ ,  $p(\theta|y)$ , the posterior expected loss of an action  $a$  is  
 $v(p(\theta|y), a) = \int_{\Theta} l(\theta, a)p(\theta|y)d\theta.$

# Posterior Mean as Bayes Estimator Under Quadratic Loss

- quadratic loss: If  $\theta \in \Theta$  is a parameter of interest, and  $\tilde{\theta}$  is an estimate of  $\theta$ , then  $l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$  is the quadratic loss arising from the use of the estimate  $\tilde{\theta}$  instead of  $\theta$ .
- Posterior Mean as Bayes Estimate Under Quadratic Loss:

$$E(\theta|\mathbf{y}) = \tilde{\theta} = \int_{\Theta} \theta p(\theta|\mathbf{y}) d\theta.$$

- Proof: Quadratic loss implies that the posterior expected loss is

$$v(\theta, \tilde{\theta}) = \int_{\Theta} (\theta - \tilde{\theta})^2 p(\theta|\mathbf{y}) d\theta.$$

Expanding the quadratic yields

$v(\theta, \tilde{\theta}) = \int_{\Theta} \theta^2 p(\theta|\mathbf{y}) d\theta + \tilde{\theta}^2 - 2\tilde{\theta} E(\theta|\mathbf{y}).$  Differentiate with respect to  $\tilde{\theta}$ , noting that the first term does not involve  $\tilde{\theta}$ . Solve the 1st order condition for  $\tilde{\theta}$  and the result follows.

# Bayes Estimates

- Quadratic Loss: mean of the posterior density,

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta$$

- Symmetric Linear Loss: median of the posterior density, n.b., only well-defined for  $\theta \in \Theta \subseteq \mathbb{R}$ , in which case  $\tilde{\theta}$  is defined such that

$$\int_{-\infty}^{\tilde{\theta}} .5$$

- All-or-nothing Loss: mode of the posterior density

$$\tilde{\theta} = \operatorname{argmax}_{\theta \in \Theta} p(\theta|y)$$

# Credible Region; HPD region

## Definition (Credible Region)

A region  $C \subseteq \Omega$  such that  $\int_C p(\theta) d\theta = 1 - \alpha$ ,  $0 \leq \alpha \leq 1$  is a  $100(1 - \alpha)\%$  credible region for  $\theta$ .

For single-parameter problems (i.e.,  $\Omega \subseteq \mathbb{R}$ ), if  $C$  is not a set of disjoint intervals, then  $C$  is a credible interval.

If  $p(\theta)$  is a (prior/posterior) density, then  $C$  is a (prior/posterior) credible region.

## Definition (Highest Probability Density Region)

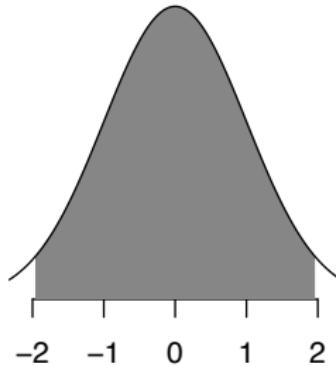
A region  $C \subseteq \Omega$  is a  $100(1 - \alpha)\%$  highest probability density region for  $\theta$  under  $p(\theta)$  if

- ①  $P(\theta \in C) = 1 - \alpha$
- ②  $P(\theta_1) \geq P(\theta_2), \forall \theta_1 \in C, \theta_2 \notin C$

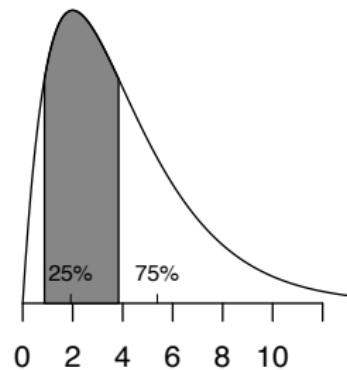
# HPD intervals

- A  $100(1 - \alpha)\%$  HPD region for a symmetric, unimodal density is unique and symmetric around the mode; e.g., a normal density.
- Cf skewed distributions; a HPD differs from simply reading off the quantiles.

$N(0,1)$

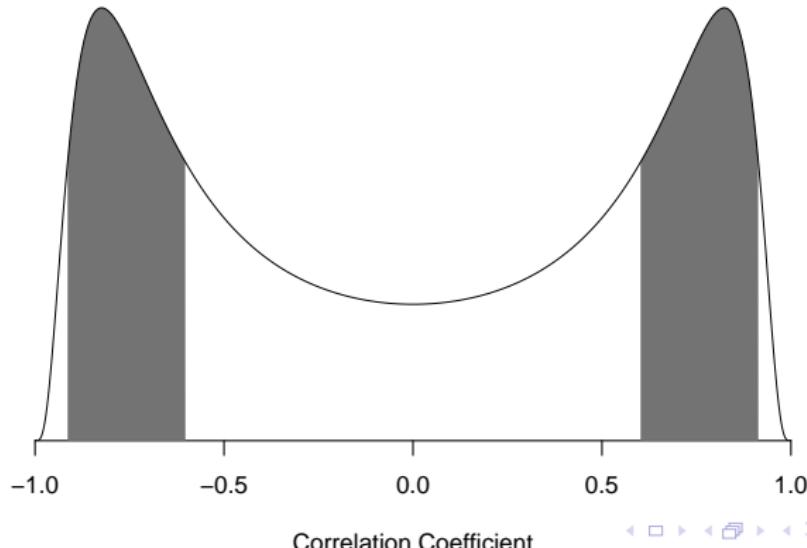


$\chi^2$  4 df



# HPD intervals

- HPDs can be a series of disjoint intervals, e.g., a bimodal density
- these are uncommon; but in such a circumstance, presenting a picture of the density might be the reasonable thing to do.
- See Example 1.7, p28:  $\mathbf{y}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ , subject to extreme missingness.  
The posterior density of  $\rho(\boldsymbol{\Sigma}) = \sigma_{12} / \sqrt{\sigma_{11}\sigma_{22}}$ :

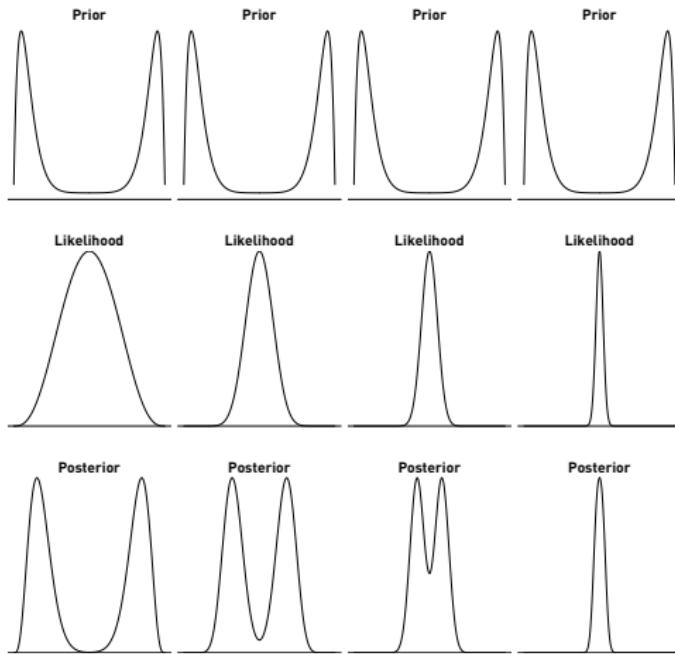


# Bayesian Consistency

- for anything other than a dogmatic/degenerate prior (see the earlier discussion of Cromwell's Rule), more and more data will overwhelm the prior.
- Bayesian asymptotics: with an arbitrarily large amount of sample information relative to prior information, the posterior density tends to the likelihood (normalized to be a density over  $\theta$ ).
- central limit arguments: since likelihoods are usually approximately normal in large samples, then so too are posterior densities.

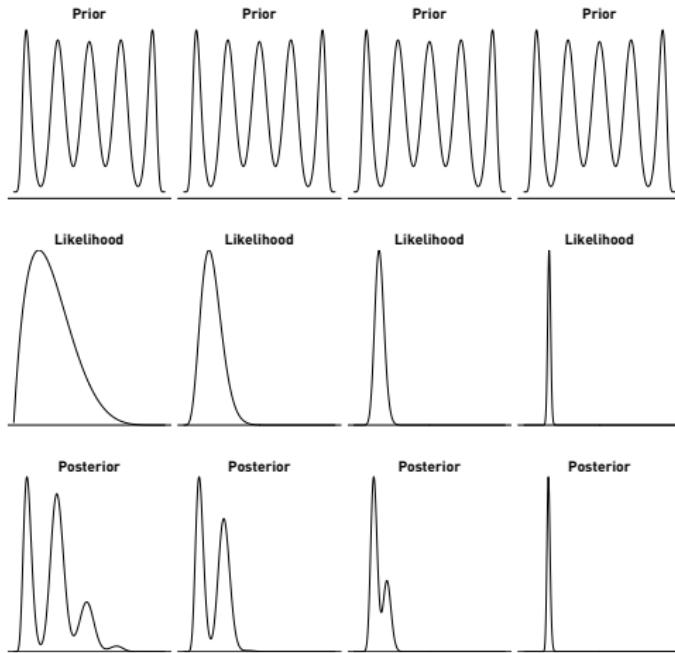
# Bayesian Consistency

The prior remains fixed across the sequence, as sample size increases and  $\theta^*$  is held constant. In this example,  $n = 6, 30, 90, 450$  across the four columns.



# Bayesian Consistency

The prior remains fixed across the sequence, as sample size increases and  $\theta^*$  is held constant. In this example,  $n = 6, 30, 150, 1500$  across the four columns.



# Other topics from Chapter One

- §1.8. Bayesian hypothesis testing.
- §1.9. Exchangeability. de Finetti's Representation Theorem.

# BAYESIAN INFERENCE FOR SIMPLE PROBLEMS

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 11, 2012

# Conjugacy

- Bayes Rule says  $p(\theta|y) \propto p(\theta)p(y|\theta)$
- Mantra: “Posterior is proportional to prior times likelihood”
- This math easy to do when we use prior densities that are **conjugate** with respect to the likelihood  $p(y|\theta)$ .

# Conjugacy

**Definition 1.2 (p15 BASS):** Suppose a prior density  $p(\theta)$  belongs to a class of parametric densities,  $\mathcal{F}$ . Then the prior density is said to be conjugate with respect to a likelihood  $p(y|\theta)$  if the posterior density  $p(\theta|y)$  is also in  $\mathcal{F}$ .

- Up until the Markov-chain Monte Carlo revolution in the 1990s, Bayesian inference was almost all done with
  - conjugate priors
  - simple problems; e.g., rates and proportions (Bernoulli trials, coin-flipping), counts (Poisson), means, variances and regression (normal data).
- Bayes estimates such as the mean of the posterior density  $E(\theta|y)$  have a simple mathematical form; can be computed “by hand”; are **“precision-weighted averages”** of estimates based on the prior and on the data.

## Example: coin flipping (p50)

- $y_i \in \{0, 1\}$ , exchangeable
- unknown success probability  $\theta \in [0, 1]$ ,
- data:  $y_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta), i = 1, \dots, n$
- $r \sim \text{Binomial}(\theta; n), r = \sum y_i$ .
- binomial likelihood  $p(r|\theta)$ :

$$\binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

- What prior density  $p(\theta)$  is conjugate wrt the likelihood  $p(r|\theta)$ ?
- That is, what form does  $p(\theta)$  have to be such that  $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$  is of the same form?

## Example: coin flipping (p50)

- $y_i \in \{0, 1\}$ , exchangeable
- unknown success probability  $\theta \in [0, 1]$ ,
- data:  $y_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta), i = 1, \dots, n$
- $r \sim \text{Binomial}(\theta; n), r = \sum y_i$ .
- binomial likelihood  $p(r|\theta)$ :

$$\binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

- What prior density  $p(\theta)$  is conjugate wrt the likelihood  $p(r|\theta)$ ?
- That is, what form does  $p(\theta)$  have to be such that  $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$  is of the same form?
- Answer: the **Beta** density is conjugate wrt a binomial likelihood.

# Beta density

- A prior density for  $\theta \in [0, 1]$  must have the properties:
  - 1  $p(\theta) \geq 0, \theta \in [0, 1].$
  - 2  $\int_0^1 p(\theta) d\theta = 1.$
- A conjugate prior must also have the property that  $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$  is of the same form as  $p(\theta).$

## Definition

### Beta density:

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1},$$

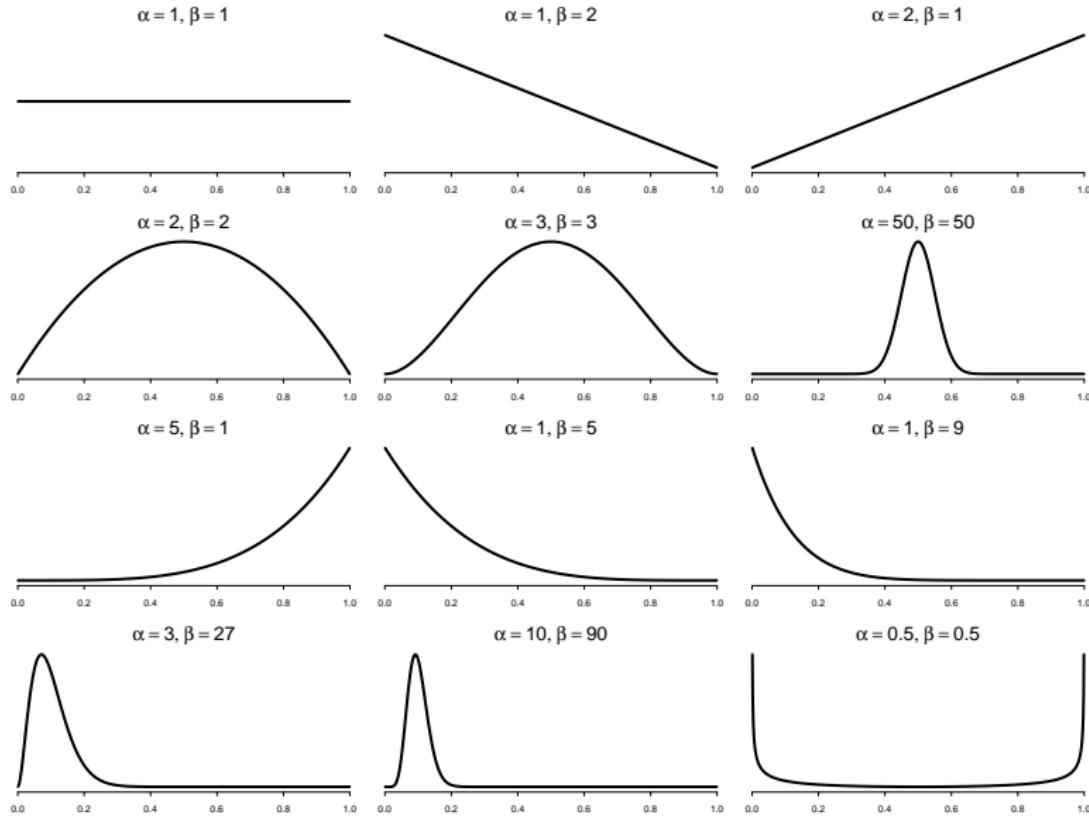
where  $\theta \in [0, 1]$ ,  $\alpha, \beta > 0$  and  $\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t) dt$  is the Gamma function (Definition B.19).

# Beta density

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

- Note that the leading terms involving the Gamma functions do not involve  $\theta$ :  $p(\theta; \alpha, \beta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}$ .
- A uniform density on  $[0, 1]$  is a special case of the Beta density, arising when  $\alpha = \beta = 1$ .
- Symmetric densities with a mode/mean/median at .5 are generated when  $\alpha = \beta$  for  $\alpha, \beta > 1$ .
- the mean,  $E(\theta) = \frac{\alpha}{\alpha + \beta}$
- the mode:  $\frac{\alpha - 1}{\alpha + \beta - 2}$
- the variance:  $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

# Beta density



# Conjugacy of the Beta wrt the Binomial Likelihood

## Theorem

**Conjugacy of Beta Prior, Binomial Data.** Given a binomial likelihood over  $r$  successes in  $n$  Bernoulli trials, each independent conditional on an unknown success parameter  $\theta \in [0, 1]$ , i.e.,

$$\mathcal{L}(\theta; r, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

then the prior density  $p(\theta) = \text{Beta}(\alpha, \beta)$  is conjugate with respect to the binomial likelihood, generating the posterior density  
 $p(\theta|r, n) = \text{Beta}(\alpha + r, \beta + n - r)$ .

# Conjugacy of the Beta wrt the Binomial Likelihood

## Theorem

**Conjugacy of Beta Prior, Binomial Data.** Given a binomial likelihood over  $r$  successes in  $n$  Bernoulli trials, each independent conditional on an unknown success parameter  $\theta \in [0, 1]$ , i.e.,

$$\mathcal{L}(\theta; r, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

then the prior density  $p(\theta) = \text{Beta}(\alpha, \beta)$  is conjugate with respect to the binomial likelihood, generating the posterior density  
 $p(\theta|r, n) = \text{Beta}(\alpha + r, \beta + n - r).$

Shorthand:

$$\theta \sim \text{Beta}(\alpha, \beta), r \sim \text{Binomial}(\theta, n) \Rightarrow \theta|r, n \sim \text{Beta}(\alpha + r, \beta + n - r).$$

# Conjugacy of the Beta wrt the Binomial Likelihood

**Proof of Proposition: Conjugacy of Beta Prior, Binomial Data.**

By Bayes Rule,

$$p(\theta|r, n) = \frac{\mathcal{L}(\theta; r, n)p(\theta)}{\int_0^1 \mathcal{L}(\theta; r, n)p(\theta)d\theta} \propto \mathcal{L}(\theta; r, n)p(\theta)$$

Ignoring terms that do not depend on  $\theta$ ,

$$\begin{aligned} p(\theta|r, n) &\propto \underbrace{\theta^r(1-\theta)^{n-r}}_{\text{likelihood}} \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \\ &= \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1} \end{aligned}$$

which is the *kernel* of a Beta density. That is,

$p(\theta|r, n) = c\theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$  where  $c$  is the normalizing constant

$$\frac{\Gamma(n + \alpha + \beta)}{\Gamma(r + \alpha)\Gamma(n - r + \beta)},$$

In other words,  $\theta|r, n \sim \text{Beta}(\alpha + r, \beta + n - r)$ .

# Interpretation of Conjugacy in Data-Equivalent Terms

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$r \sim \text{Binomial}(\theta, n)$$

$$\theta|r, n \sim \text{Beta}(\alpha + r, \beta + n - r)$$

- *as if* our prior distribution represents the information in a sample of  $\alpha + \beta - 2$  independent Bernoulli trials, in which we observed  $\alpha - 1$  “successes” or  $y_i = 1$ .
- $p(\theta) \equiv \text{Unif}(0, 1) \equiv \text{Beta}(1, 1)$  has the “data equivalent” interpretation of  $\alpha + \beta - 2 = 0$  prior observations.

## Example: Florida Polling

- Florida poll, March 2000, voting intentions for the November 2000 presidential election.
- $n = 621$ . Bush 45% ( $n = 279$ ), Gore 37% (230), Buchanan 3% (19) and undecided 15% (93).
- For simplicity, we ignore the undecided and Buchanan vote share, leaving Bush with 55% of the two-party vote intentions, and Gore with 45%, and  $n = 509$  respondents expressing a preference for the two major party candidates.
- We assume that the survey responses are independent, and (perhaps unrealistically) that the sample is a random sample of Floridian voters.
- Thus, the binomial likelihood is (ignoring constants that do not depend on  $\theta$ ),

$$p(r|\theta, n) \propto \theta^{279} (1 - \theta)^{509 - 279}.$$

The maximum likelihood estimate of  $\theta$  is

$$\hat{\theta}_{MLE} = r/n = 279/509 = .548$$

## Example: Florida Polling

- Prior information from previous elections.
- Forecasting model produces a forecast of Bush vote share of 49.1%, with a standard error of 2.2 percentage points.
- Convert this to a Beta density: we seek values for  $\alpha$  and  $\beta$  such that

$$E(\theta; \alpha, \beta) = \alpha / (\alpha + \beta) = .491$$

$$V(\theta; \alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = .022^2$$

which yields a system of equations in two unknowns.

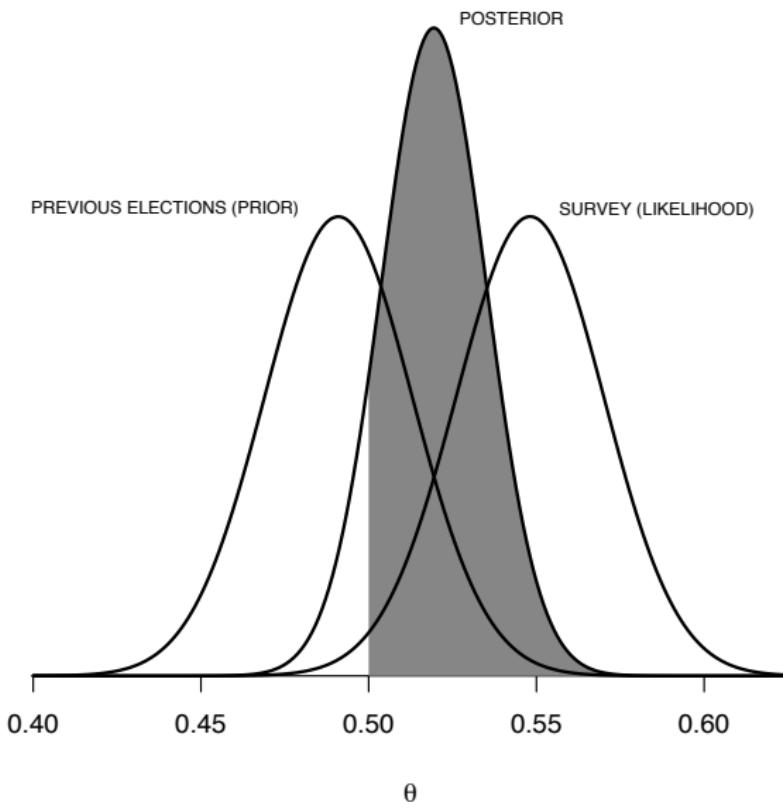
- Solving yields

$$\alpha = 515.36 \times .491 = 253.04,$$

$$\beta = 515.36 \times (1 - .491) = 262.32.$$

- information in the previous elections is equivalent to having ran another poll with  $n \approx 515$  in which  $r \approx 253$  respondents said they would vote for the Republican presidential candidate.

# Example: Florida Polling



# Bayes Estimate as a Convex Combination of Prior and Data

Consider a Bayes estimate such as the posterior mean:

$$E(\theta|r, n) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha + r}{\alpha + \beta + n} = \frac{n_0\theta_0 + n\hat{\theta}}{n_0 + n}$$

where  $\hat{\theta} = r/n$  is the maximum likelihood estimate of  $\theta$ ,  $n_0 = \alpha + \beta$  and  $\theta_0 = \alpha/(\alpha + \beta) = E(\theta)$  is the mean of the prior density for  $\theta$ .

Alternatively,

$$E(\theta|r, n) = \gamma\theta_0 + (1 - \gamma)\hat{\theta}$$

where  $\gamma = n_0/(n_0 + n)$ , and since  $n_0, n > 0$ ,  $\gamma \in [0, 1]$ . Alternatively,

$$E(\theta|r, n) = \hat{\theta} + \gamma(\theta_0 - \hat{\theta}).$$

That is, a Bayes estimate of  $\theta$  --- is a *weighted average* of the prior mean  $\theta_0$  and the maximum likelihood estimate  $\hat{\theta}$ .

# Conjugate analysis of normal data

We first consider the simple case of normal data, with unknown mean  $\mu$  and known variance  $\sigma^2$ :

$$y_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$$

Likelihood:

$$p(y_1, \dots, y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ \frac{-(y_i - \mu)^2}{2\sigma^2} \right]$$

Conjugacy: we seek a prior for  $\mu$ ,  $p(\mu)$ , s.t. the posterior

$$p(\mu|y_1, \dots, y_n) \propto p(y_1, \dots, y_n|\mu, \sigma^2)p(\mu)$$

is in the same class as  $p(\mu)$ .

# Conjugate analysis of normal data

For normal data with unknown mean  $\mu$ , a normal prior for  $\mu$  is conjugate:

## Theorem

Let  $y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ , with  $\sigma^2$  known, and  $\mathbf{y} = (y_1, \dots, y_n)'$ . If  $\mu \sim N(\mu_0, \sigma_0^{-2})$  is the prior density for  $\mu$ , then  $\mu$  has posterior density

$$\mu | \mathbf{y} \sim N \left( \frac{\mu_0 \sigma_0^{-2} + \bar{y} \frac{n}{\sigma^2}}{\sigma_0^{-2} + \frac{n}{\sigma^2}}, \left( \sigma_0^{-2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

Note the precision-weighted average form of the mean of the posterior density:

- precision = 1/variance
- the prior has precision  $\sigma_0^{-2}$
- the MLE of  $\mu$ ,  $\bar{y}$  has precision  $n/\sigma^2$ .
- the posterior precision  $\sigma_0^{-2} + n/\sigma^2$  is the sum of the prior precision and the data precision.

# Conjugate analysis, normal data, mean and variance unknown

- Same model as in previous section:  $y_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ .
- Likelihood:

$$p(y_1, \dots, y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

- Parameters: a vector  $\Theta = (\mu, \sigma^2)'$ ,  $\mu \in \mathbb{R}$ ,  $\sigma^2 \in \mathbb{R}^+$ ; i.e.,  $\Theta \in \Theta = \mathbb{R} \times \mathbb{R}^+$ .
- Prior: it is easier to obtain a conjugate prior if we factor the joint density over  $\Theta$  into the product of a conditional density for  $\mu$  given  $\sigma^2$  and a marginal density for  $\sigma^2$ ; i.e.,  $p(\Theta) = p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ .

# Conjugate analysis, normal data, mean and variance unknown

Prior:  $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$  where

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0)$$

$$\sigma^2 \sim \text{Inverse-Gamma}(v_0/2, v_0 \sigma_0^2/2)$$

where

- $\mu_0 = E(\mu|\sigma^2) = E(\mu)$  is the mean of the prior density for  $\mu$
- $\sigma^2/n_0$  is the variance of the prior density for  $\mu$ , conditional on  $\sigma^2$ , with  $n_0$  interpretable as a “prior sample size”
- $v_0 > 0$  is a prior “degrees of freedom” parameter
- $v_0 \sigma_0^2$  is equivalent to the sum of squares one obtains from a (previously observed) data set of size  $v_0$

# inverse-Gamma density

## Definition

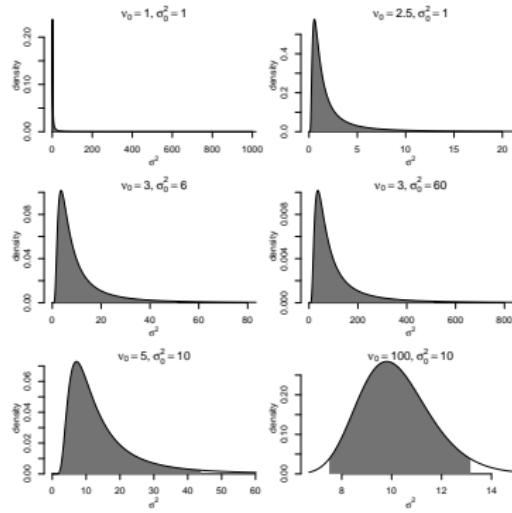
If  $x > 0$  follows an inverse-Gamma density with shape parameter  $a > 0$  and scale parameter  $b > 0$ , conventionally written as  $x \sim \text{inverse-Gamma}(a, b)$ , then

$$p(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(\frac{-b}{x}\right),$$

- $E(x) = \frac{b}{a-1}$  if  $a > 1$ .
- $V(x) = \frac{b^2}{(a-1)^2(a-2)}$  if  $a > 2$ .
- $p(x)$  has a mode at  $b/(a+1)$ .
- If  $x \sim \text{inverse-Gamma}(a, b)$  then  $1/x \sim \text{Gamma}(a, b)$ .
- Typical use is  $\sigma^2 \sim \text{inverse-Gamma}(v_0/2, v_0\sigma_0^2/2)$ . An improper density  $p(\sigma^2) \propto 1/\sigma^2$  results with  $v_0 = 0$ .

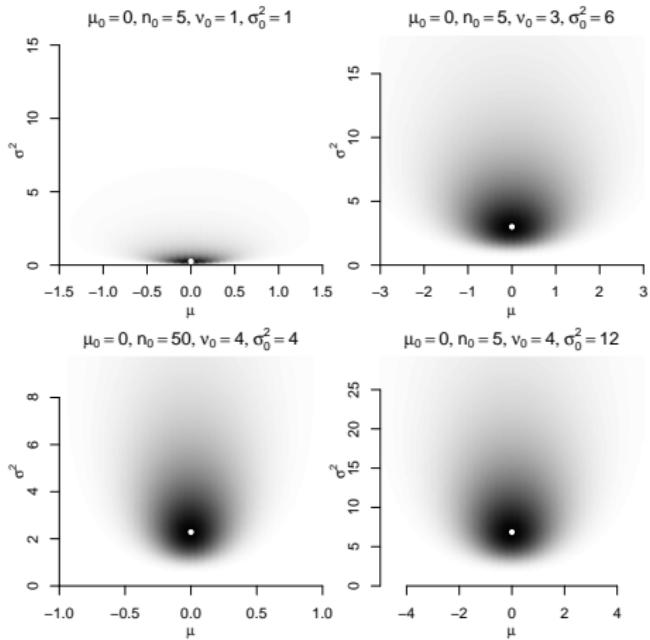
$$\sigma^2 \sim \text{inverse-Gamma}(v_0/2, v_0\sigma_0^2/2)$$

- The mean,  $E(\sigma^2)$ , is  $v_0\sigma_0^2/(v_0 - 2)$ , provided  $v_0 > 2$ , otherwise the mean is undefined, and the mode occurs at  $v_0\sigma_0^2/(v_0 + 2)$ .
- The mean and the mode tend to coincide as  $v_0 \rightarrow \infty$ ; i.e., the inverse-Gamma density tends to a (symmetric) normal density as  $v_0 \rightarrow \infty$ , but otherwise is skewed right.



# normal/inverse-Gamma density

A density for  $\boldsymbol{\theta} = (\mu, \sigma^2)' \in \Theta = \mathbb{R} \times \mathbb{R}^+$ , indexed by four parameters,  $\mu_0, n_0, v_0, \sigma_0^2$ .



# Conjugacy of the normal/inverse-Gamma prior

## Theorem (Proposition 2.5, BASS)

Let  $y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ . If

$\Theta = (\mu, \sigma^2)' \sim \text{normal/inverse-Gamma}(\mu_0, n_0, v_0, \sigma_0^2)$ , then  
 $\Theta | \mathbf{y} \sim \text{normal/inverse-Gamma}(\mu_1, n_1, v_1, \sigma_1^2)$ , where

$$\begin{aligned}\mu_1 &= \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n} \\ n_1 &= n_0 + n, \quad v_1 = v_0 + n \\ v_1 \sigma_1^2 &= v_0 \sigma_0^2 + S + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{y})^2\end{aligned}$$

and where  $S = \sum_{i=1}^n (y_i - \bar{y})^2$ . That is,

$$\mu | \sigma^2, \mathbf{y} \sim N(\mu_1, \sigma^2/n_1)$$

$$\sigma^2 | \mathbf{y} \sim \text{inverse-Gamma} \left( \frac{v_1}{2}, \frac{v_1 \sigma_1^2}{2} \right)$$

# Marginal Posterior Density, Normal Mean, Conjugacy

- The conditional posterior density for  $\mu$ ,  $p(\mu|\mathbf{y}, \sigma^2)$  is a normal density in which  $\sigma^2$  appears in the expression for the variance of the conditional posterior density.
- But if we're interested in  $\mu$ , we want its marginal posterior density  $p(\mu|\mathbf{y})$
- We “integrate out” or “average over” uncertainty with respect to  $\sigma^2$ ; i.e.,

$$p(\mu|\mathbf{y}) = \int_0^\infty p(\mu|\sigma^2, \mathbf{y})p(\sigma^2|\mathbf{y})d\sigma^2$$

where the limits of integration follow from the fact that variances are strictly positive.

- Resulting marginal density for  $\mu$  is a student- $t$  density.

# Marginal Posterior Density, Normal Mean, Conjugacy

## Theorem (Proposition 2.6, BASS)

Assume conditions of the previous theorem. Then the marginal posterior density of  $\mu$  is a student-t density (Definition B.37), with location parameter  $\mu_1$ , scale parameter  $\sqrt{\sigma_1^2/n_1}$  and  $v_1$  degrees of freedom, where  $n_1 = n_0 + n$ ,

$$\mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_1},$$

$$\sigma_1^2 = S_1/v_1, S_1 = v_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_1}(\bar{y} - \mu_0)^2, v_1 = v_0 + n,$$
$$s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \text{ and } \bar{y} = n^{-1} \sum_{i=1}^n y_i.$$

## Proof.

Proposition C.7, BASS.



# Posterior Predictive Density, normal data, conjugacy

- Consider making a prediction for a future observation,  $y^*$ .
- This quantity also has a posterior density, called a *posterior predictive density*:

$$\begin{aligned} p(y^* | \mathbf{y}) &= \int_{\mathbb{R}^+} \int_{\mathbb{R}} p(y^* | \mu, \sigma^2, \mathbf{y}) p(\mu, \sigma^2 | \mathbf{y}) d\mu d\sigma^2 \\ &= \int_{\mathbb{R}^+} \int_{\mathbb{R}} p(y^* | \mu, \sigma^2, \mathbf{y}) p(\mu | \sigma^2, \mathbf{y}) p(\sigma^2 | \mathbf{y}) d\mu d\sigma^2 \end{aligned}$$

- The prediction should not simply condition on particular values of the parameters  $\mu$  and  $\sigma^2$ ; we're uncertain about the prediction because we're uncertain about the parameters  $\mu$  and  $\sigma^2$ .
- The double integral might look a little formidable, but the posterior predictive density has a familiar form...

# Posterior Predictive Density, normal data, conjugacy

## Theorem (Proposition 2.7, BASS)

Let  $y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ,  $(\mu, \sigma^2) \sim \text{normal/inverse-Gamma}(\mu_0, n_0, v_0, \sigma_0^2)$ . Then the posterior predictive density for a future observation  $y^*$ ,  $p(y^* | \mathbf{y})$ , is a student-t density, with location parameter

$$E(y^* | \mathbf{y}) = E(\mu | \mathbf{y}) = \mu_1 = \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n},$$

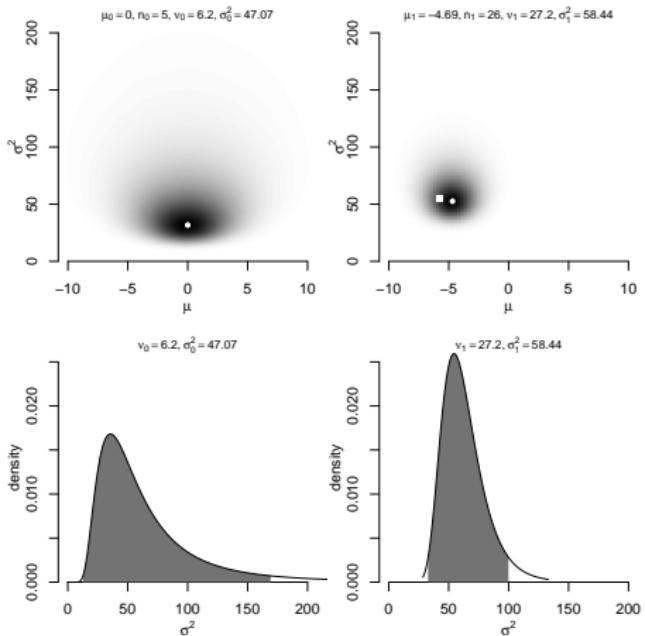
scale parameter  $\sigma_1 \sqrt{(n_1 + 1)/n_1}$  and  $v_1 = n + v_0$  degrees of freedom, where  $n_1 = n_0 + n$ ,  $\sigma_1^2 = S_1/v_1$  and  $S_1 = v_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{y})^2$ .

## Proof.

Proposition C.8, BASS.



# Example 2.13, Suspected voter fraud in Pennsylvania



Prior and Posterior normal/inverse-Gamma Densities. Prior densities on the left; posterior densities on the right. Normal/inverse-Gamma densities for  $(\mu, \sigma^2)$  in the upper panels, with darker colors indicating regions of higher density, the circle indicating the mode, and for the posterior density, the square indicating the location of the maximum likelihood estimates. Marginal inverse-Gamma densities for  $\sigma^2$  appear in the lower panels, with the shaded area corresponding to a 95% highest density region.

# Regression

## Theorem (Conjugate Prior Normal Regression Model)

$$\begin{aligned}y_i | \mathbf{x}_i &\stackrel{iid}{\sim} N(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2) \\ \boldsymbol{\beta} | \sigma^2 &\sim N(\mathbf{b}_0, \sigma^2 \mathbf{B}_0) \\ \sigma^2 &\sim \text{inverse-Gamma}(v_0/2, v_0 \sigma_0^2/2)\end{aligned}$$

then

$$\begin{aligned}\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X} &\sim N(\mathbf{b}_1, \sigma^2 \mathbf{B}_1), \\ \sigma^2 | \mathbf{y}, \mathbf{X} &\sim \text{inverse-Gamma}(v_1/2, v_1 \sigma_1^2/2) \\ \mathbf{b}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}' \mathbf{X})^{-1} (\mathbf{B}_0^{-1} \mathbf{b}_0 + \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}) \\ \mathbf{B}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}' \mathbf{X})^{-1} \\ v_1 &= v_0 + n \quad \text{and} \\ v_1 \sigma_1^2 &= v_0 \sigma_0^2 + S + r.\end{aligned}$$

# Conjugacy Summary

Prior	Data/Likelihood	Posterior
$\theta \sim \text{Beta}$	$r \sim \text{Binomial}(\theta; n)$	$\theta r, n \sim \text{Beta}$
$\mu \sigma^2 \sim N$	$y \sim N(\mu, \sigma^2)$	$\mu y, \sigma^2 \sim N$ $\mu y \sim t$
$\sigma^2 \sim \text{inverse-Gamma}$	$y \sim N(\mu, \sigma^2)$	$\sigma^2 y \sim \text{inverse-Gamma}$
$\theta \sim \text{Gamma}$	$y \sim \text{Poisson}(\theta)$	$\theta y \sim \text{Gamma}$
$\Sigma \sim \text{inverse-Wishart}$	$\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\boldsymbol{\Sigma} \mathbf{y} \sim \text{inverse-Wishart}$
$\alpha \sim \text{Dirichlet}$	$\mathbf{r} \sim \text{Multinomial}(\boldsymbol{\alpha}; n)$	$\boldsymbol{\alpha} \mathbf{r}, n \sim \text{Dirichlet}$

# Limitations of Conjugacy

- seemingly small set of problems amenable to conjugate Bayesian analysis
- e.g., no logit/probit regression!
- prior to MCMC revolution, Bayesian ideas interesting, perhaps even “right”, but extremely difficult (impossible) to implement outside small set of problems
- MCMC changed this, circa 1990. Avalanche of Bayesian applications in statistics, econometrics. Now standard.
- Bayes Theorem to 1990: a long time!
- still make use of conjugacy (or conditional conjugacy) in implementing a MCMC algorithm known as the Gibbs sampler.

# MONTE CARLO METHODS

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 3, 2012

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

- Example: compute  $E(\theta), \theta \in \Theta \subseteq \mathbb{R}$ .
- Analytically:  $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$ .

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

- Example: compute  $E(\theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}$ .
- Analytically:  $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$ . This math might be “too hard”.

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

- Example: compute  $E(\theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}$ .
- Analytically:  $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$ . This math might be “too hard”.
- Monte Carlo estimate:
  - sample  $\theta$  from  $p(\theta)$ ; call these  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ ;  $T$  large.
  - compute the following *Monte Carlo estimate* of  $E(\theta)$ :

$$\widehat{E(\theta)}_T = \sum_{t=1}^T \theta^{(t)} / T$$

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

- Example: compute  $E(\theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}$ .
- Analytically:  $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$ . This math might be “too hard”.
- Monte Carlo estimate:
  - sample  $\theta$  from  $p(\theta)$ ; call these  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ ;  $T$  large.
  - compute the following *Monte Carlo estimate* of  $E(\theta)$ :

$$\widehat{E(\theta)}_T = \sum_{t=1}^T \theta^{(t)} / T$$

- accuracy of estimate  $\widehat{E(\theta)}_T$  improves as  $T \rightarrow \infty$ .

# The Monte Carlo principle

*anything we want to know about a random variable  $\theta$  can be learned by sampling many times from  $p(\theta)$ , the density of  $\theta$ .*

- Example: compute  $E(\theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}$ .
- Analytically:  $E(\theta) = \int_{\Theta} \theta p(\theta) d\theta$ . This math might be “too hard”.
- Monte Carlo estimate:
  - sample  $\theta$  from  $p(\theta)$ ; call these  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ ;  $T$  large.
  - compute the following *Monte Carlo estimate* of  $E(\theta)$ :

$$\widehat{E(\theta)}_T = \sum_{t=1}^T \theta^{(t)} / T$$

- accuracy of estimate  $\widehat{E(\theta)}_T$  improves as  $T \rightarrow \infty$ .
- all this generalizes to (a) vectors  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)'$ ; (b) estimates of functionals of  $\boldsymbol{\theta}$ ,  $h(\boldsymbol{\theta})$ .

# Brief history of Monte Carlo methods

- Buffon's needle, estimating  $\pi$ ; see `simpi` in my `pscl` package for R
- Lord Kelvin
- Enrico Fermi
- Metropolis and Ulam; Manhattan Project, Los Alamos

# Brief history of Monte Carlo methods

## JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247

SEPTEMBER 1949

Volume 44

### THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULM

*Los Alamos Laboratory*

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

# Simulation Consistency

## Theorem (Simulation Consistency, Independent Draws, Part 1)

Suppose  $\{\boldsymbol{\Theta}^{(t)}\}$  is a sequence of independent draws from the density  $f(\boldsymbol{\Theta})$ , with  $\boldsymbol{\Theta} \in \mathbb{R}^k$  and  $h : \mathbb{R}^k \rightarrow \mathbb{R}$ . Then

$$\bar{h}^{(T)} = T^{-1} \sum_{t=1}^T h(\boldsymbol{\Theta}^{(t)}) \xrightarrow{a.s.} E[h(\boldsymbol{\Theta})]$$

## Proof.

The claim is a restatement of the strong law of large numbers (e.g., Proposition B.10; BASS).



# Simulation Consistency

## Theorem (Simulation Consistency, Independent Draws, Part 2)

Further, suppose that for  $p \in (0, 1)$ , there is a unique  $q_p$  such that  $\Pr[h(\boldsymbol{\Theta}) \leq q_p] \geq p$  and  $\Pr[h(\boldsymbol{\Theta}) \geq q_p] \geq 1 - p$  are both true. Consider  $q_p^{(T)} \in \mathbb{R}$  such that

$$T^{-1} \sum_{t=1}^T \mathcal{I}(-\infty < h(\boldsymbol{\Theta}^{(t)}) < q_p^{(T)}) \geq p$$

where  $p \in (0, 1)$  and  $\mathcal{I}(\cdot)$  is a binary indicator function, equal to 1 if its argument is true, and zero otherwise. Then  $q_p^{(T)} \xrightarrow{a.s.} q_p$ .

## Proof.

Geweke (2005, Theorem 4.1.1); Rao (1973, 423); van der Vaart (1998, 305). □

# Learning about a uniform random variable

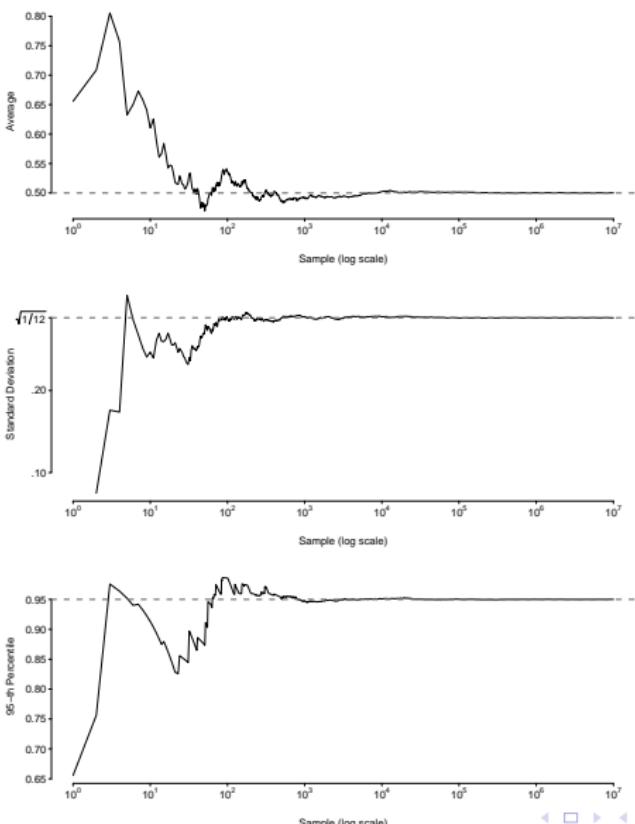
- Suppose  $\theta \sim \text{Unif}(0, 1)$
  - But we have forgotten that  $E(\theta) = .5$ ,  $\text{sd}(\theta) = \sqrt{1/12}$ ,  $\Pr(\theta \leq q) = q$ ,  $q \in (0, 1)$  etc.
  - Monte Carlo methods? Use `rnorm` in R.
- ① the average of the sampled values,  $\bar{\theta}^{(T)} = T^{-1} \sum_{t=1}^T \theta_t$ , will be very close to  $E(\theta) = .5$ . Importantly,  $\bar{\theta}^{(T)}$  gets arbitrarily close to .5 as  $T \rightarrow \infty$ .
- ② likewise, the standard deviation of the sampled values

$$\text{sd}(\theta)^{(T)} = \left[ (T - 1)^{-1} \sum_{t=1}^T (\theta_t - \bar{\theta}^{(T)})^2 \right]^{1/2}$$

will get arbitrarily close to  $\sqrt{1/12}$  as  $T \rightarrow \infty$ .

- ③ the  $p$ -th quantile of the sampled values,  $q_p^{(T)}$ ,  $p \in (0, 1)$ ;  $q_p^{(T)} \xrightarrow{a.s.} p$ .

# Learning about a uniform random variable (Figure 3.1)



# Monte carlo integration/marginalization (method of composition)

- $\boldsymbol{\theta} = (\theta_1, \theta_2)$ , with  $\theta_j \in \Theta_j \subseteq \mathbb{R}, j = 1, 2$ .
- Posterior density:  $p(\boldsymbol{\theta}|\mathbf{y})$ .
- But interest centers on the marginal posterior density of  $\theta_1$ ,

$$p(\theta_1|\mathbf{y}) = \int_{\Theta_2} p(\theta_1, \theta_2|\mathbf{y}) d\theta_2 = \int_{\Theta_2} p(\theta_1|\theta_2, \mathbf{y}) p(\theta_2|\mathbf{y}) d\theta_2.$$

- Method of composition:
  - 1: **for**  $t = 1$  to  $T$  **do**
  - 2:   sample  $\theta_2^{(t)}$  from  $p(\theta_2|\mathbf{y})$
  - 3:   sample  $\theta_1^{(t)}$  from  $p(\theta_1|\theta_2^{(t)}, \mathbf{y})$ .
  - 4: **end for**
- $\theta_1^{(t)} \sim p(\theta_1|\mathbf{y})$ , as desired.

# Method of Composition to sample from a $t$ density

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &= \int_0^{\infty} p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})p(\sigma^2|\mathbf{y}, \mathbf{X})d\sigma^2 \\ p(\sigma^2|\mathbf{y}, \mathbf{X}) &\equiv \text{inverse-Gamma}(v/2, vs^2/2) \\ p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) &\equiv N(\mathbf{b}, \sigma^2 \mathbf{B}) \end{aligned}$$

- 1: **for**  $t = 1$  to  $T$  **do**
- 2:   sample  $\sigma^{2(t)}$  from  $p(\sigma^2) \equiv \text{inverse-Gamma}(v/2, vs^2/2)$
- 3:   sample  $\boldsymbol{\beta}^{(t)}$  from  $p(\boldsymbol{\beta}|\sigma^{2(t)}) \equiv N(\mathbf{b}, \sigma^{2(t)} \mathbf{B})$
- 4: **end for**

i.e.,  $\boldsymbol{\beta}^{(t)} \sim p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \equiv \text{student-}t$ .

# Monte carlo inference, functions of parameters

- 2-by-2 table:

		$x_i$		
		0	1	
$y_i$	0	$n_0 - r_0$	$n_1 - r_1$	
	1	$r_0$	$r_1$	
		$n_0$	$n_1$	$n$

- Two success probabilities:  $0 \leq \theta_0, \theta_1 \leq 1$ ;  $\Pr(y_i = 1|x_i = j) = \theta_j$ ,  $j \in \{0, 1\}$ .
- MLEs:  $\hat{\theta}_j = r_j/n_j$ ,  $j \in \{0, 1\}$ .
- Bayesian analysis: independent, conjugate Beta priors over each  $\theta_j$ ,  $\theta_j \sim \text{Beta}(\alpha_j, \beta_j)$ . posterior densities are independent Beta densities  $\theta_j|r_j, n_j \sim \text{Beta}(\alpha_j + r_j, \beta_j + n_j - r_j)$ .

# Inference for Difference of Two Binomial Proportions

- $q = \theta_1 - \theta_0$ ; difference of two binomial proportions
- $p(\theta_j | r_j, n_j) \equiv \text{Beta}$ .
- But what is  $p(q | r_0, r_1, n_0, n_1)$ ?
- Until recently (Pham-Gia and Turkkan 1993), the density of the difference of two Betas was unknown (unavailable in closed form)
- Characterize this density by Monte Carlo methods.

## Algorithm:

- 1 sample  $\theta_0^{(t)}$  from  $\text{Beta}(\alpha_0 + r_0, \beta_0 + n_0 - r_0)$
- 2 sample  $\theta_1^{(t)}$  from  $\text{Beta}(\alpha_1 + r_1, \beta_1 + n_1 - r_1)$
- 3 compute  $q^{(t)} = \theta_1^{(t)} - \theta_0^{(t)}$

Repeat many times,  $t = 1, \dots, T$ ; sampled  $q^{(t)}$  are a sample from the posterior density of  $q$ ; summarize numerically, make histogram, etc.

## Example 3.2; War and revolution in Latin America

Sekhon (2005); Geddes (1990); Skocpol (1979):

	Revolution	No Revolution
Defeated & Invaded/Lost Territory	1	7
Not Defeated within 20 years	2	74

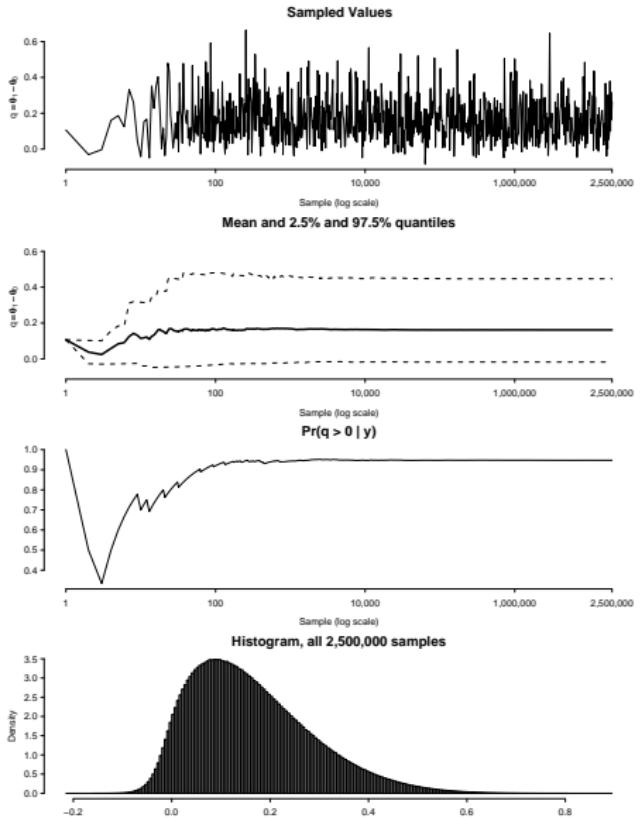
- Each observation spans 20 years for each Latin American country
- The sole observation in the top left of the cross-tabulation is Bolivia: it suffered a military defeat in 1935, and a social revolution in 1952.
- The two observations in the lower left of the table are Mexico (revolution in 1910) and Nicaragua (revolution in 1979).
- The MLEs are  $\hat{\theta}_0 = 2/(2 + 74) = .026$  and  $\hat{\theta}_1 = 1/(7 + 1) = .125$ , suggesting that revolutions are much more likely conditional on military defeat than conditional on not having experienced a military defeat.
- Uniform priors  $\theta_j \sim \text{Beta}(1, 1), j = 0, 1$ ; posterior densities  $\theta_0|r_0, n_0 \sim \text{Beta}(3, 75)$  and  $\theta_1|r_1, n_1 \sim \text{Beta}(2, 8)$ .

## Example 3.2; War and revolution in Latin America

We seek the posterior density of  $q = \theta_1 - \theta_0$ ; we use Monte Carlo methods:

- 1: **for**  $t = 1$  to  $T$  **do**
- 2:   sample  $\theta_1^{(t)}$  from  $p(\theta_1|\mathbf{y}) \equiv \text{Beta}(2, 8)$
- 3:   sample  $\theta_0^{(t)}$  from  $p(\theta_0|\mathbf{y}) \equiv \text{Beta}(3, 75)$
- 4:    $q^{(t)} \leftarrow \theta_1^{(t)} - \theta_0^{(t)}$
- 5: **end for**

# Example 3.2; War and revolution in Latin America



## Example 3.2; War and revolution in Latin America

R code is trivial:

```
nsims <- 1e6
theta1 <- rbeta(nsims,2,8)
theta0 <- rbeta(nsims,3,75)
q <- theta1 - theta0
summary(q)
mean(q>0)
```

# In JAGS

```
model{  
    ## model for the data  
    for(i in 1:2){  
        r[i] ~ dbin(theta[i],n[i])  
    }  
  
    ## priors  
    for(i in 1:2){  
        theta[i] ~ dbeta(1,1)  
    }  
  
    ## quantity of interest  
    q <- theta[2] - theta[1]  
}
```

# Sampling algorithms

Suppose  $\theta \sim p$ ? How to sample from  $p$ ?

- ① inverse-CDF method
- ② importance sampling
- ③ rejection sampling
- ④ slice sampler

# Inverse-CDF method

- $\theta \sim p, \theta \in \Theta \subseteq \mathbb{R}$ .
- $F(q) = \Pr(\theta \leq q) = \int_{-\infty}^q p(\theta)d\theta$ ; n.b.,  $F : \Theta \mapsto (0, 1)$ .
- Suppose  $F^{-1}$  exists, is computable;  $F^{-1} : (0, 1) \mapsto \Theta$
- Inverse-CDF algorithm:
  - 1: **for**  $t = 1$  to  $T$  **do**
  - 2:   sample  $p^{(t)} \sim \text{Unif}(0, 1)$
  - 3:    $\theta^{(t)} \leftarrow F^{-1}(p^{(t)})$
  - 4: **end for**
- Reasonably rare that an inverse-CDF exists. E.g., not available in closed form for the normal, but good approximations exist; e.g., Wichura (1988), used in the `pnorm` function in R).

# Importance Sampling

- we can evaluate the target density at any given point in its support; i.e., we can compute  $p(\theta) \forall \theta \in \Theta$ .
- no algorithm for direct sampling from  $p(\theta)$ .
- we *can* sample from a density  $s(\theta)$ , where  $s(\theta)$  has the property that  $p(\theta) > 0 \Rightarrow s(\theta) > 0, \forall \theta \in \Theta$ .
- Exploit the following identity: consider some  $h(\theta)$ , then

$$E[h(\theta)] = \int_{\Theta} h(\theta)p(\theta)d\theta = \int_{\Theta} h(\theta)p(\theta)/s(\theta)s(\theta)d\theta.$$

- Importance sampling algorithm:

```
1: for  $t = 1$  to  $T$  do
2:   sample  $\theta^{(t)} \sim s(\theta)$ .
3:    $w^{(t)} \leftarrow p(\theta^{(t)})/s(\theta^{(t)})$ 
4: end for
5:  $\bar{h}^{(T)} \leftarrow T^{-1} \sum_{t=1}^T h(\theta^{(t)})w^{(t)}$ 
```

# Accept-Reject Sampling (von Neumann 1951)

- $\theta \sim p(\theta)$ , but can't sample from this density
- can sample from a *majorizing function*  $g(\theta)$ , that is, where  $g(\theta) > p(\theta), \forall \theta$ .
- can trivially find a majorizing function by rescaling a *proposal* density:  $g(\theta) = cm(\theta)$ :

```
1: for  $t = 1$  to  $T$  do
2:   sample  $z \sim m(\theta)$ 
3:   sample  $u \sim \text{Unif}(0, 1)$ 
4:    $r \leftarrow p(z)/cm(z)$ 
5:   if  $u \leq r$  then
6:      $\theta^{(t)} \leftarrow z$  {"accept"}
7:   else
8:     go to 2 {"reject"}
9:   end if
10: end for
```

# Accept-Reject Sampling

- The target density  $p$  need only be known up to a factor of proportionality. All that matters is that we can sample from a function that majorizes the target density, and we can control that through the scaling constant,  $c$ .
- Useful in Bayesian analysis, where it is often the case that posterior densities are only known up to an (unknown) proportionality constant.
- An accept-reject algorithm produces potentially many draws that are rejected, and hence the algorithm can be computationally inefficient.
- More efficient if the majorizing function  $g$  closely approximates the target density,  $p$ ; e.g.,  $r = p/g \approx 1$ .

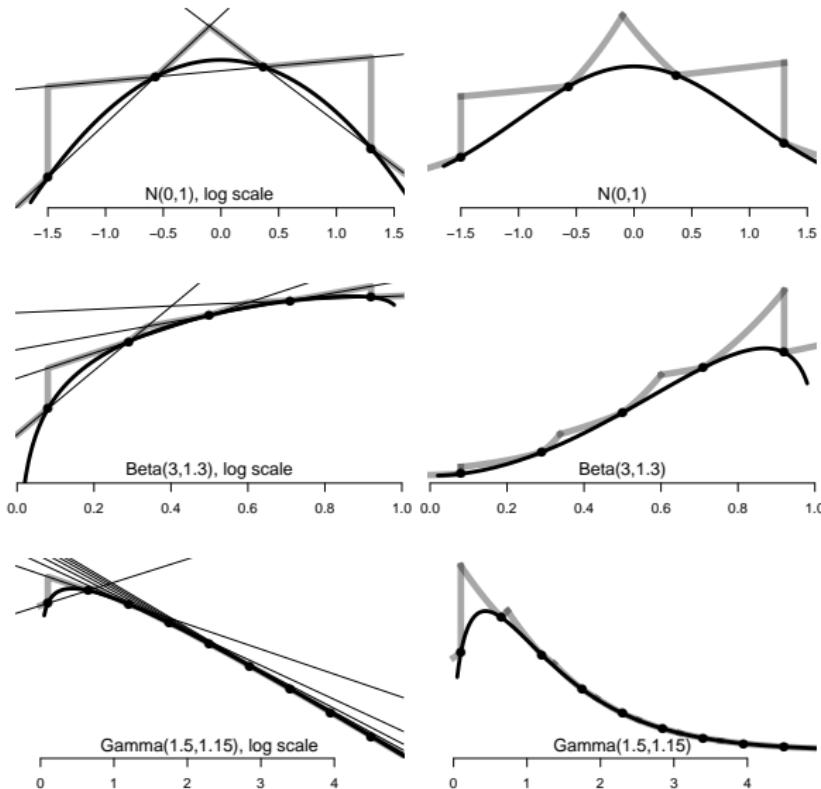
# Adaptive Rejection Sampling

- hard to find a good proposal density
- but if target density  $p(\theta)$  is log-concave and continuously differentiable, then use adaptive rejection sampling
- build a proposal density as a set of piecewise exponential densities bracketing the target density
- A density  $p(\Theta)$ ,  $\Theta \in \mathbb{R}^k$ , is log-concave if the determinant of

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log p}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log p}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 \log p}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log p}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log p}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log p}{\partial \theta_k \partial \theta_k} \end{pmatrix}$$

is non-positive.

# Adaptive Rejection Sampling



# Adaptive Rejection Sampling

- algorithm is *adaptive*: successfully sampled points are added to the set of evaluation points.
- initialization/adaptation phase consists of getting a good set of evaluation points
- ARS was a critical step in developing a general purpose computer program for simulation-based Bayesian statistical analysis; e.g., Gilks and Wild (1992).
- Extension to non-log-concave densities (e.g., Gilks, Best and Tan 1995)

## Slice Sampling §5.2.7

- $\theta \sim p(\theta), \theta \in \Theta \subseteq \mathbb{R}$ , restricting ourselves to the one-dimensional case for the time being.
- This is equivalent to sampling the pair  $(\theta, U)$  uniformly from the set  $\mathcal{J} = \{(\theta, u) : 0 < u < p(\theta)\}$ .
- i.e., let  $\tilde{\Theta} = \Theta \times [0, m]$ , where  $p(\theta) \leq m \forall \theta \in \Theta$ .
- Now pick a random point  $(\theta^*, U^*) \in \tilde{\Theta}$ .
- if  $0 < U^* < p(\theta^*)$ , then accept the draw.
- sampling over the  $u$  dimension is equivalent to marginalizing  $u$  out of  $f(\theta, u)$ , i.e.,

$$p(\theta) = \int_0^{p(\theta)} f(u) du = \int_0^{p(\theta)} du, \quad 0 < u < p(\theta),$$

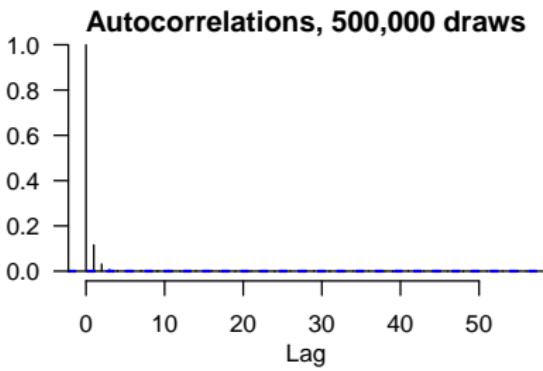
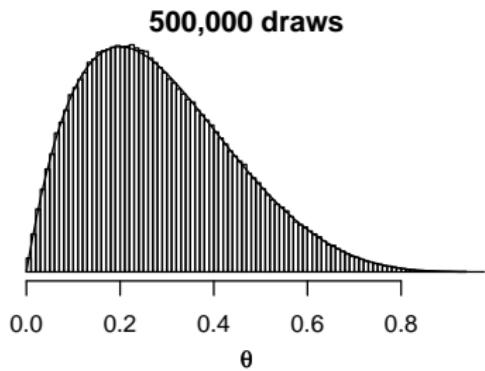
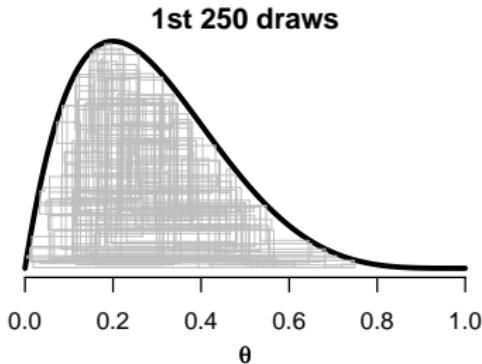
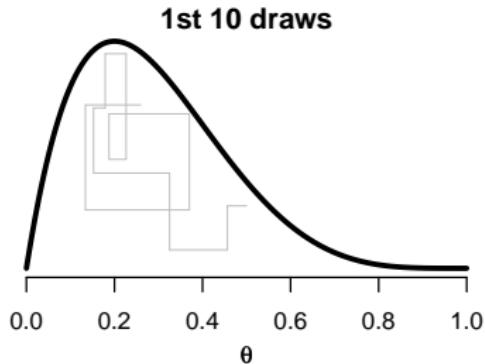
since  $f(u)$  here is a constant.

## Slice Sampling §5.2.7

- sequentially sample from  $g(U|\theta)$  and  $g(\theta|U)$
- Given  $(\theta^{(t-1)}, U^{(t-1)})$ :
  - 1: sample  $U^{(t)} \sim \text{Unif}(0, p(\theta^{(t-1)}))$
  - 2: sample  $\theta^{(t)} \sim \text{Unif}(\mathcal{A}^{(t)})$  where  $\mathcal{A}^{(t)} = \{\theta : p(\theta) \geq U^{(t)}\}$ .
- special case of the Gibbs sampler

# Slice sampling from a Beta density, Example 5.12

$p(\theta) \equiv \text{Beta}(2, 5), \theta \in \Theta \equiv [0, 1]$



# References

- Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis* 2:131--150.
- Geweke, John. 2005. *Contemporary Bayesian Econometrics and Statistics*. Hoboken, New Jersey: Wiley.
- Gilks, W. R., N. G. Best and K.K.C. Tan. 1995. "Adaptive rejection Metropolis sampling withing Gibbs sampling." *Applied Statistics* 44:455--472.
- Gilks, W. R. and P. Wild. 1992. "Adaptive rejection sampling for Gibbs sampling." *Applied Statistics* 41:337--348.
- Pham-Gia, Thu and Noyan Turkkan. 1993. "Bayesian analysis of the difference of two proportions." *Communications in Statistics --- Theory and Methods* 22:1755--1771.
- Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and Its Applications*. Second ed. New York: Wiley.
- Sekhon, Jasjeet S. 2005. "Making Inference from  $2 \times 2$  Tables: The Inadequacy of the Fisher Exact Test for Observational Data and a Bayesian Alternative." Typescript. Survey Research Center, University of California, Berkeley.
- Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.
- van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press.
- von Neumann, J. 1951. "Various techniques used in connection with random digits." *National Bureau of Standards Applied Mathematics Series* 12:36--38.
- Wichura, Michael J. 1988. "Algorithm AS 241: The Percentage Points of the Normal Distribution." *Applied Statistics* 37:477--484. <http://links.jstor.org/sici?doi=0035-9254>



# **Non-informative Priors Multiparameter Models**

Statistics 220

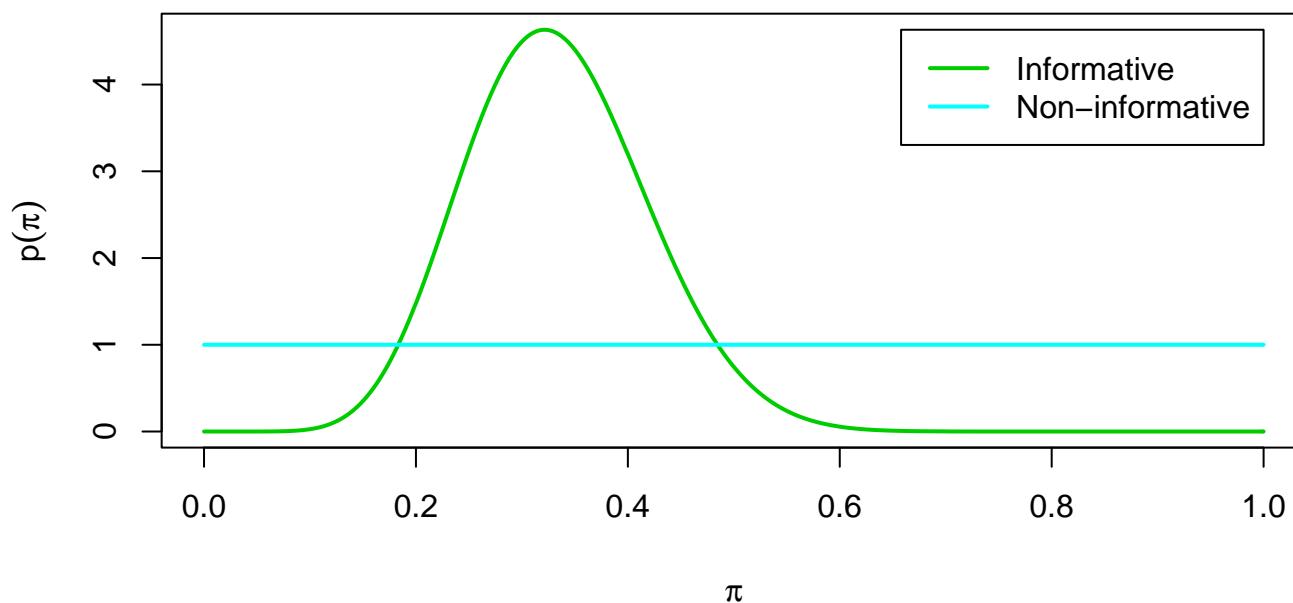
Spring 2005



# Prior Types

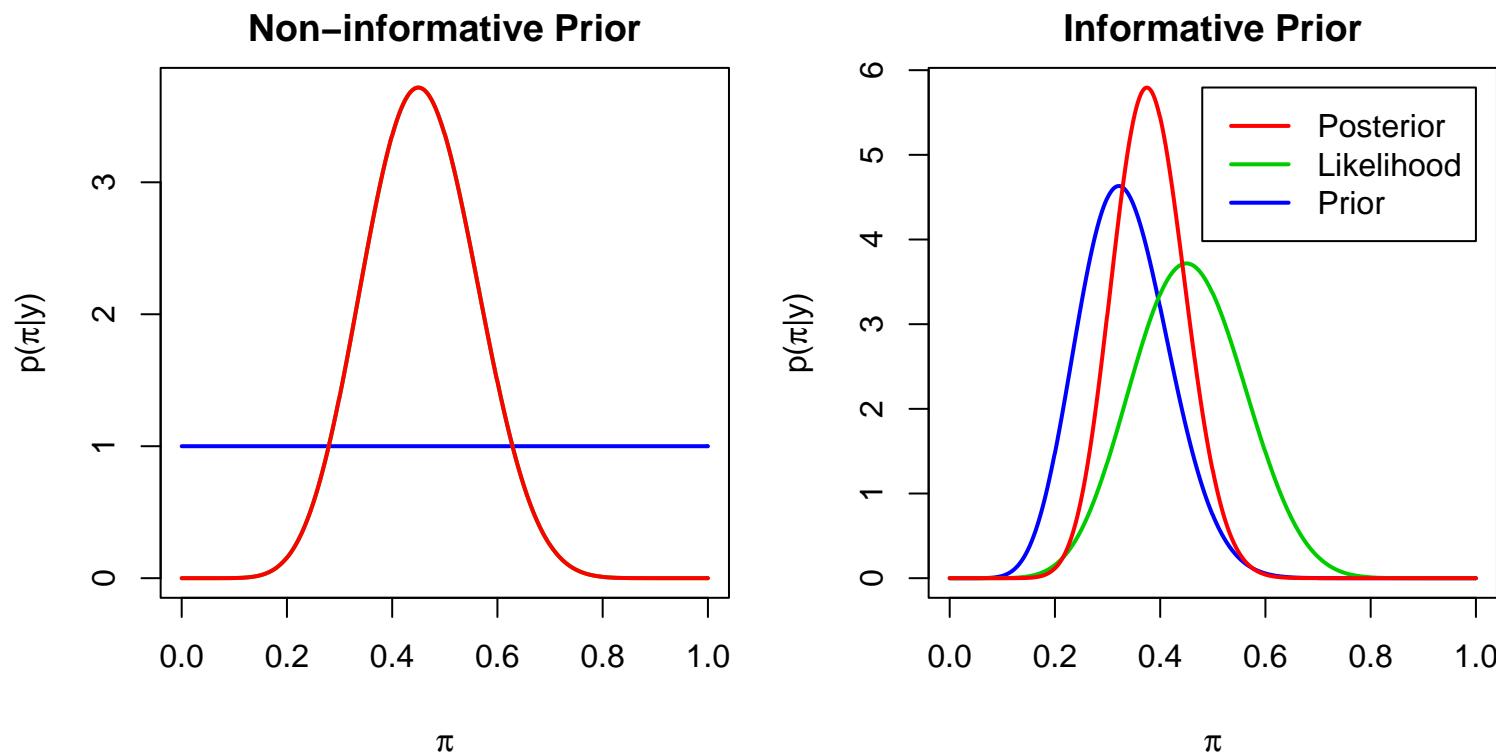
- Informative vs Non-informative

There has been a desire for prior distributions that play a minimal role in the posterior distribution. These are sometimes referred to as non-informative or reference priors.



These priors are often described as vague, flat, or diffuse.

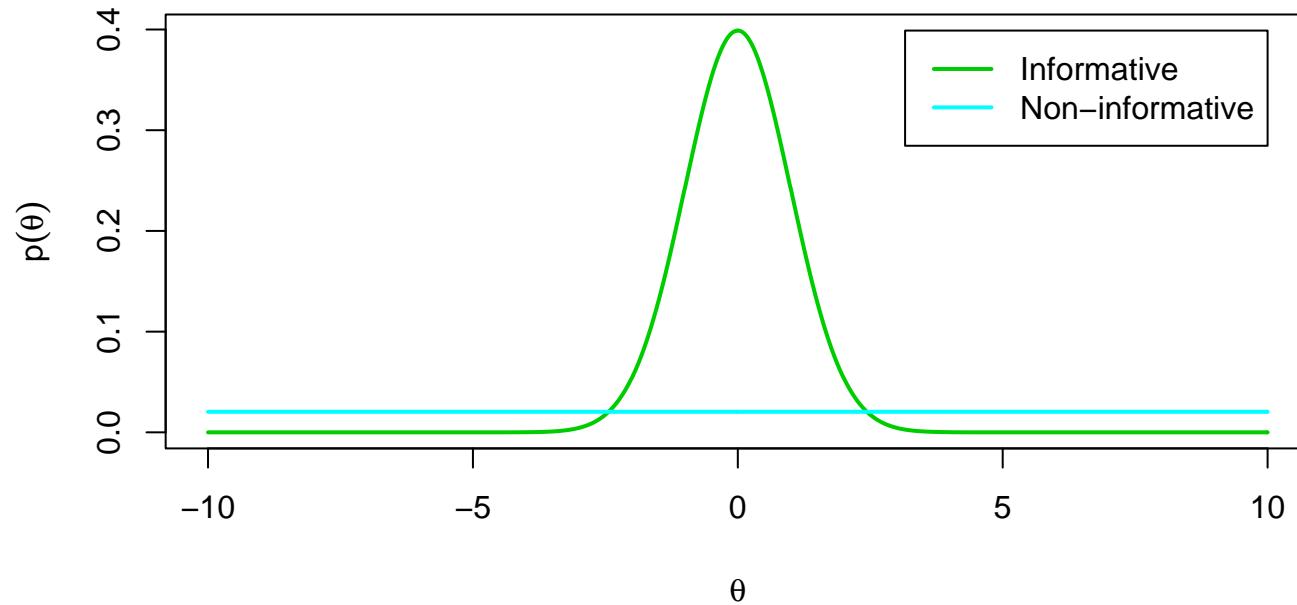
In the case when the parameter of interest exists on a bounded interval (e.g. binomial success probability  $\pi$ ), the uniform distribution is an “obvious” non-informative prior.



For this example, with the non-informative prior, Posterior = Likelihood

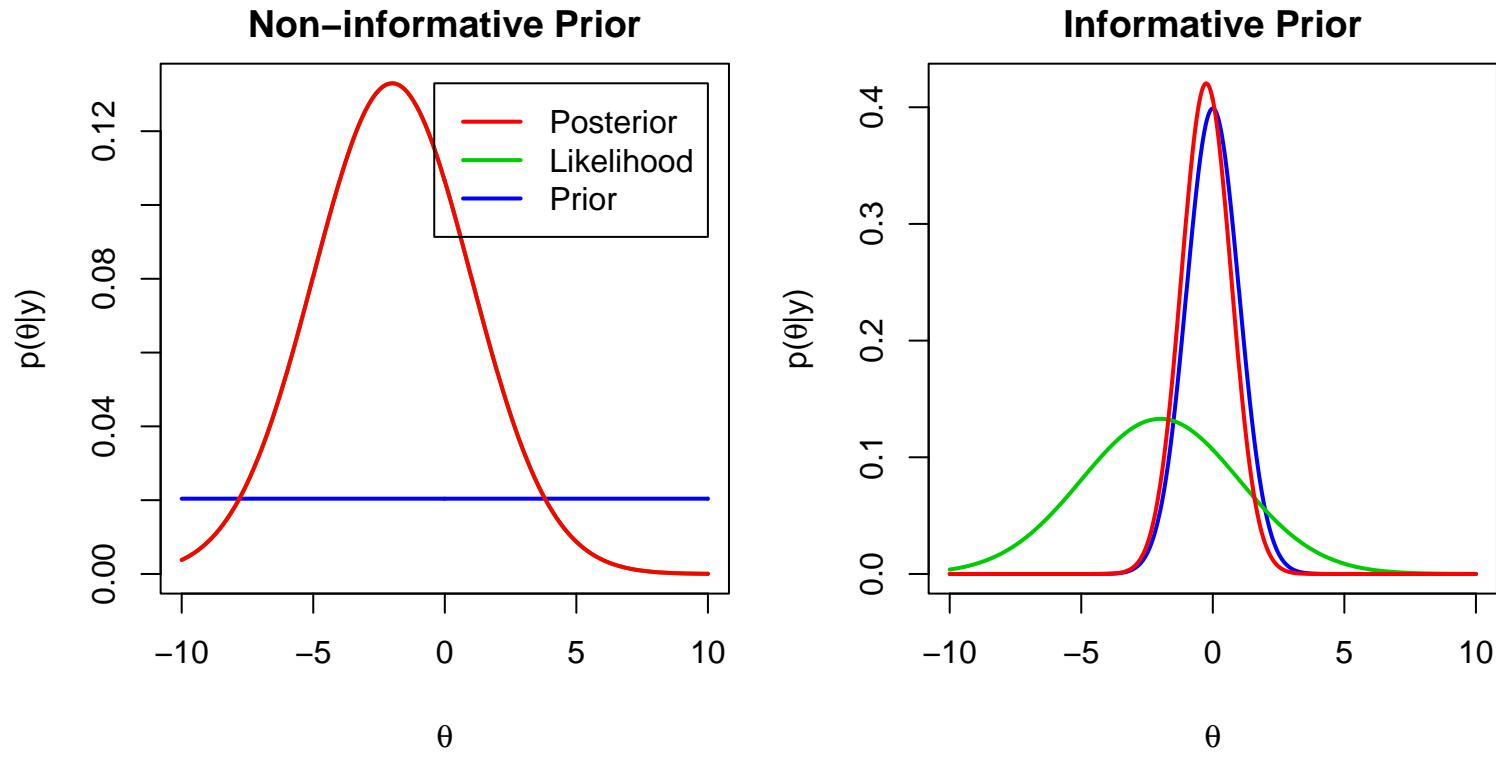
However for a parameter that occurs on an infinite interval (e.g. a normal mean  $\theta$ ), using a uniform prior on  $\theta$  is problematic.

For the normal mean example, lets use the conjugate prior  $N(\mu_0, \tau_0^2)$ , but with a very big variance  $\tau_0^2$



The posterior mean and precision are

$$\mu_n = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$



So if we let  $\tau_0^2 \rightarrow \infty$ , then

$$\mu_n \rightarrow \bar{y} \quad \text{and} \quad \frac{1}{\tau_n^2} \rightarrow \frac{n}{\sigma^2}$$

This equivalent to the posterior being proportional to the likelihood, which is what we get if  $p(\theta) \propto 1$  (e.g. uniform).

This does not describe a valid probability density as

$$\int_{-\infty}^{\infty} d\theta = \infty$$

- Proper vs Improper

A prior is called proper if it is a valid probability distribution

$$p(\theta) \geq 0, \quad \forall \theta \in \Theta \quad \text{and} \quad \int_{\Theta} p(\theta) d\theta = 1$$

(Actually all that is needed is a finite integral. Priors only need to be defined up to normalization constants.)

A prior is called improper if

$$p(\theta) \geq 0, \quad \forall \theta \in \Theta \quad \text{and} \quad \int_{\Theta} p(\theta) d\theta = \infty$$

If a prior is proper, so must the posterior.

If a prior is improper, the posterior often is, i.e.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

is a proper distribution for all  $y$ . Note that an improper prior may lead to an improper prior. For many common problems, popular improper reference priors will usually lead to proper posteriors, assuming there is enough data.

For example

$$\begin{aligned} y_1, \dots, y_n | \theta &\stackrel{iid}{\sim} N(\theta, \sigma^2) \\ p(\theta) &\propto 1 \end{aligned}$$

will have a proper posterior as long  $n$  is at least 1.

## Non-informative Priors

While it may seem that picking a non-informative prior distribution might be easy, (e.g. just use a uniform), its not quite that straight forward.

Example: Normal observations with known mean, but unknown variance

$$\begin{aligned} y_1, \dots, y_n | \sigma &\stackrel{iid}{\sim} N(\theta, \sigma^2) \\ p(\sigma) &\propto 1 \end{aligned}$$

What is the equivalent prior on  $\sigma^2$

**Aside:** Let  $\theta$  be a random variable with density  $p(\theta)$  and let  $\phi = h(\theta)$  be a one-one transformation. Then the density of  $\phi$  satisfies

$$f(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta) |h'(\theta)|^{-1} \quad \text{where } \theta = h^{-1}(\phi)$$

If  $h(\sigma) = \sigma^2$ ,  $h'(\sigma) = 2\sigma$ , then a uniform prior on  $\sigma$  leads to

$$p(\sigma^2) = \frac{1}{2\sigma}$$

which clearly isn't uniform. This implies that our prior belief is that the variance should be small

Similarly, if there is a uniform prior on  $\sigma^2$ , the equivalent prior on  $\sigma$  is

$$p(\sigma) = 2\sigma$$

This implies that we believe sigma to be large.

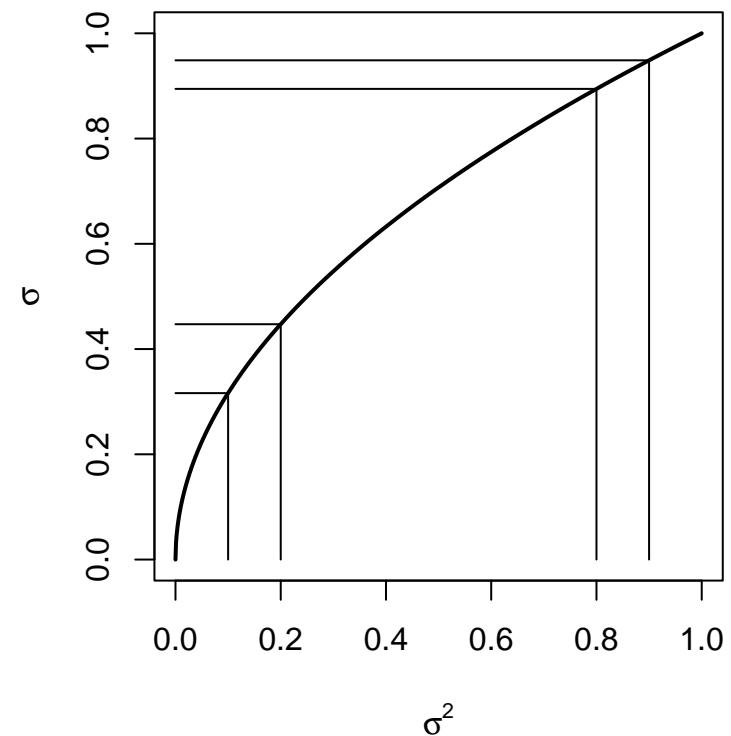
One way to think about what is happening is to look at what happens to intervals of equal measure.

In the case  $\sigma^2$  being uniform, an interval  $[a, a + 0.1]$  must have the same prior measure as the interval  $[0.1, 0.2]$ .

When we transform to  $\sigma$ , the prior measure on it must have intervals  $[\sqrt{a}, \sqrt{a + 0.1}]$  having equal measure.

But note that the length of the interval  $[\sqrt{a}, \sqrt{a + 0.1}]$  is a decreasing function of  $a$ , which agrees with the increasing density in  $\sigma$ .

So when talking about non-informative priors you need to think about on what scale.



## Jeffreys' Priors

Can we pick a prior where the scale the parameter is measured in doesn't matter.

Jeffreys' principle states that any rule for determining the prior density  $p(\theta)$  should yield an equivalent result if applied to the transformed parameter.

That is applying

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta) |h'(\theta)|^{-1} \quad \text{where } \theta = h^{-1}(\phi)$$

should give the same answer as dealing directly with the transformed model

$$p(y, \phi) = p(\phi)p(y|\phi)$$

Applying this principle gives

$$p(\theta) = [J(\theta)]^{1/2}$$

where  $J(\theta)$  is the *Fisher information* for  $\theta$

$$J(\theta) = E \left[ \left( \frac{d \log p(y|\theta)}{d\theta} \right)^2 | \theta \right] = -E \left[ \frac{d^2 \log p(y|\theta)}{d\theta^2} | \theta \right]$$

Why does this work?

It can be shown that (see page 63)

$$J(\phi) = J(\theta) \left| \frac{d\theta}{d\phi} \right|^2$$

so

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right|$$

For example, for the normal example with unknown variance, the Jeffreys' prior for the standard deviation  $\sigma$  is

$$p(\sigma) \propto \frac{1}{\sigma}$$

Alternative descriptions under different parameterizations for the variability are

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$p(\log \sigma^2) \propto p(\log \sigma) \propto 1$$

For exponential data ( $y_i \stackrel{iid}{\sim} Exp(\theta); \theta = \frac{1}{E[y|\theta]}$ ), the Jeffreys' prior is

$$p(\theta) = \frac{1}{\theta}$$

If you wish to parameterize in terms of the mean ( $\lambda = \frac{1}{\theta}$ ), the Jeffreys' prior is

$$p(\lambda) = \frac{1}{\lambda}$$

For parameters with infinite parameter spaces (like a normal mean or variance), the Jeffrey's prior is often improper under the usual parameterizations.

As we have seen, different approaches may lead to different non-informative priors.

# Pivotal Quantities

There are some situations where the common approaches give the same non-informative distributions.

- Location Parameter

Suppose that the density of  $p(y - \theta | \theta)$  is a function that is free of  $\theta$ , call it  $f(u)$ . For example, if  $y \sim N(\mu, 1)$ ,

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Then  $y - \theta$  is known as a pivotal quantity and  $\theta$  is known as a pure location parameter.

In this situation, a reasonable approach would assume that a non-informative prior would give  $f(y - \theta)$  as the posterior density of  $y - \theta | y$ .

This gives

$$p(y - \theta|y) \propto p(\theta)p(y - \theta|\theta)$$

which implies  $p(\theta) \propto 1$  (i.e.  $\theta$  is uniform)

- Scale parameters

Suppose that the density of  $p(y/\theta|\theta)$  is a function that is free of  $\theta$ , call it  $g(u)$ . For example, if  $y \sim N(0, \sigma^2)$ ,

$$f(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$

In this case  $y/\theta$  is also a pivotal quantity and  $\theta$  is known as a pure scale parameter.

If we follow the same approach as to above to where  $g(y/\theta)$  as the posterior, this gives

$$p(\theta|y) = \frac{y}{\theta} p(y|\theta)$$

which implies  $p(\theta) \propto \frac{1}{\theta}$

The standard deviation from a normal distribution and the mean of an exponential distribution are scale parameters.

Using the earlier result for the standard deviation, it implies that in some sense, the “right” scale for a scale parameter  $\theta$  is  $\log \theta$  as

$$p(\theta) \propto \frac{1}{\theta}$$

$$p(\theta^2) \propto \frac{1}{\theta^2}$$

$$p(\log \theta) \propto 1$$

Note that pivotal quantities also come into standard frequentist inference. Examples involving  $y_1, \dots, y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$  are

$$\sqrt{n} \frac{\bar{y} - \mu}{s} \sim t_{n-1} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

The standard confidence intervals and hypothesis tests use the fact that these are pivotal quantities.

# Multiparameter Models

Most analyzes we wish to perform involve multiple parameters

- $y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$
- Multiple Regression:  $y_i|x_i \stackrel{ind}{\sim} N(x_i^t \beta, \sigma^2)$
- Logistic Regression:  $y_i|x_i \stackrel{ind}{\sim} Bern(p_i)$  where  $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$

In these cases we want to assume all of the parameters are unknown and want to perform inference on some or all of them.

An example of the case, where only some of them may be of interest is multiple regression. Usually only the regression parameters  $\beta$  are of interest. The measurement variance  $\sigma^2$  is often considered as a *nuisance parameter*.

Lets consider the case with two parameters  $\theta_1$  and  $\theta_2$  and that only  $\theta_1$  is of interest. An example of this would be  $N(\mu, \sigma^2)$  data where  $\theta_1 = \mu$  and  $\theta_2 = \sigma^2$ .

Want to base our inference on  $p(\theta_1|y)$ . We can get at this a couple of ways. First we can start with the joint posterior

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

This gives

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2$$

We can also get it by

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2$$

This implies that distribution of  $\theta_1$  can be considered a mixture of the conditional distributions, averaged over the nuisance parameter.

Note that this marginal conditional distribution is often difficult to determine explicitly. Normally it needs to be examined by Monte Carlo methods.

### Example: Normal Data

$$y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

For a prior, lets assume that  $\mu$  and  $\sigma^2$  are independent and use the standard non-informative priors

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \propto \frac{1}{\sigma^2}$$

So the joint posterior satisfies

$$\begin{aligned} p(\mu, \sigma^2) &\propto \frac{1}{\sigma^2} \prod_{i=1}^n \frac{1}{\sigma} \exp \left( -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^{n+2}} \exp \left( -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right] \right) \\ &= \frac{1}{\sigma^{n+2}} \exp \left( -\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{y} - \mu)^2] \right) \end{aligned}$$

where  $s^2$  is the sample variance of the  $y_i$ 's. Note that the sufficient statistics are  $\bar{y}$  and  $s^2$ .

- The conditional distribution  $p(\mu|\sigma, y)$

Note that we have already derived this as this is just the fixed and known variance case. So

$$\mu|\sigma, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

We can also get it by looking at the joint posterior. The only part that contains  $\mu$  looks like

$$p(\mu|\sigma, y) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right)$$

which is proportional to a  $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$  density.

- The marginal posterior distribution  $p(\sigma^2|y)$

To get this, we must integrate  $\mu$  out of the joint posterior.

$$\begin{aligned}
p(\sigma^2|y) &\propto \int \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu \\
&= \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right) d\mu
\end{aligned}$$

The piece left inside the integral is  $\sqrt{2\pi\sigma^2/n}$  times the  $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$  density which gives

$$\begin{aligned}
p(\sigma^2|y) &\propto \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\
&\propto \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right)
\end{aligned}$$

Which is a scaled inverse- $\chi^2$  density

$$\sigma^2|y \sim Inv - \chi^2(n - 1, s^2)$$

A random variable  $\theta \sim Inv - \chi^2(n - 1, s^2)$  if

$$\frac{(n - 1)s^2}{\theta} \sim \chi_{n-1}^2$$

Note that this result agrees with the standard frequentist result on the sample variance. However this shouldn't be surprising using the results on non-informative priors, particularly the result involving pivotal quantities.

- The marginal posterior distribution  $p(\sigma^2|y)$

Now that we have  $p(\mu|\sigma^2, y)$  and  $p(\sigma^2|y)$ , inference on  $\mu$  isn't difficult.

One method is to use the Monte Carlo approach discussed earlier

1. Sample  $\sigma_i^2$  from  $p(\sigma^2|y)$
2. Sample  $\mu_i$  from  $p(\mu|\sigma_i^2, y)$

Then  $\mu_1, \dots, \mu_m$  is a sample from  $p(\mu|y)$ .

Note that in this case, it is actually possible to derive the exact density of  $p(\mu|y)$ .

In this case

$$p(\mu|y) = \int p(\mu, \sigma^2|y) d\sigma^2$$

is tractable. With the substitution  $z = \frac{A}{2\sigma^2}$  where  $A = (n-1)s^2 + n(\bar{y} - \mu)^2$ , leaves a integral involving the gamma density (see the book, page 76).

Cranking though this leaves

$$p(\mu|y) \propto \frac{1}{\left[1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}\right]^{n/2}}$$

a  $t_{n-1}(\bar{y}, \frac{s^2}{n})$  density.

Or

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} | y \sim t_{n-1}$$

which corresponds to the standard result used for inference on a population mean

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} | \mu \sim t_{n-1}$$

# MARKOV CHAIN MONTE CARLO

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 8, 2012

# Markov chain Monte Carlo

- Two MCs
- Monte Carlo: we talked about yesterday
- Markov chains: Chapter 4, BASS.
- In general, we simply can't sample directly from the posterior density  $p(\theta|data)$ : e.g.,
  - $\theta$  is a big object (many parameters).
  - $p(\theta|data)$  is a nasty function, difficult to sample from.
- Sampling from  $p(\theta|data)$  in these cases usually require us to ***give up independence*** in the series of sampled values.
- That is, the resulting sequence of sampled values  $\{\theta^{(t)}\}$  are “serially dependent”.

# Results from Markov chain theory

- Simulation consistency results hold even when we don't have independent samples from  $p(\boldsymbol{\theta})$ .
- Proof relies on results from Markov chain theory
- A Markov chain is a stochastic process: a useful, physical analogy is a particle moving randomly in some space.
- In the context of Bayesian statistics, we have  $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ ; i.e., the “particle” is  $\boldsymbol{\theta}^{(t)}$  and the state space of the Markov chain is  $\Theta$ .
- Markov chain on  $\Theta$ :  $\{\boldsymbol{\theta}^{(t)}\} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots\}$ .
- **Ergodic theorem:** *how often* the Markov chain  $\{\boldsymbol{\theta}^{(t)}\}$  visits site  $\mathcal{A} \in \Theta$  is a simulation-consistent estimate of  $\Pr(\boldsymbol{\theta} \in \mathcal{A})$ .

# Markov chains

- Discrete state space: possible locations/states  $\Theta$  is a finite set, say with cardinality  $D$ .  $\mathbf{p}^{(t)}$  is a  $D$ -by-1 vector, with  $p_d^{(t)} = \Pr(\Theta^{(t)} = d), d \in \Theta$ .
- Continuous state space: we will consider the probability of a move from a point  $\Theta^{(t)}$  to a point  $\Theta^{(t+1)}$  in a *region*  $\mathcal{A} \subseteq \Theta$ .
- the move from  $\Theta^{(t)}$  to  $\Theta^{(t+1)}$  is governed by the Markov chain's **transition kernel**.
- for a chain on a discrete space:  $\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} \mathbf{K}$ , where  $\mathbf{K}$  is a *transition matrix*.
- for a chain on a continuous space we have a function, a *transition kernel*,  $K(\Theta^{(t)}, \cdot)$ , and  $p(\Theta)$  a density over  $\Theta$ .

$$p^{(t+1)}(\Theta) = \int_{\Theta} K(\Theta^{(t)}, \cdot) p^{(t)}(\Theta) d\Theta^{(t)}$$

# Stationary distribution of a Markov chain

- Discrete case:

$$\mathbf{p} = \mathbf{pK} \Rightarrow \mathbf{p}(\mathbf{I} - \mathbf{K}) = \mathbf{0} \Rightarrow (\mathbf{I} - \mathbf{K})' \mathbf{p}' = \mathbf{0}$$

i.e., an eigenvector of  $\mathbf{K}$  gives us the stationary distribution (up to a normalizing factor).

- Continuous case:

$$p(\boldsymbol{\Theta}^{(t+1)}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}^{(t)}) K(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Theta}^{(t+1)}) d\boldsymbol{\Theta}^{(t)}$$

i.e., we have the same density over  $p$  over  $\boldsymbol{\Theta}$  irrespective of the value of  $t$ .

## Theorem (Pointwise Ergodic Theorem; Law of Large Numbers for Markov chains)

Let  $\{\Theta^{(t)}\}$  be a Harris recurrent Markov chain on  $\Theta$  with a  $\sigma$ -finite invariant measure  $p$ . Consider a  $p$ -measurable function  $h$  s.t.  $\int_{\Theta} |h(\Theta)| dp(\Theta) < \infty$ . Then

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T h(\Theta^{(t)}) = \int_{\Theta} h(\Theta) dp(\Theta) \equiv E_p h(\Theta).$$

# Implications of the Ergodic Theorem for MCMC

- If we can construct a Markov chain the “right way”, then:
- the Markov chain will have a unique, limiting distribution, a posterior density that we happen to be interested in,  $p \equiv p(\boldsymbol{\theta}|\text{data})$
- no matter where we start the Markov chain, if we let it run long enough, it will eventually wind up generating a random tour of the parameter space, visiting sites in the parameter space  $\mathcal{A} \in \Theta$  with relative frequency proportional to  $\int_{\mathcal{A}} p(\boldsymbol{\theta}|\text{data}) d\boldsymbol{\theta}$
- the Ergodic Theorem means that averages  $\bar{h} = T^{-1} \sum_{t=1}^T h(\boldsymbol{\theta}^{(t)})$  taken over the Markov chain output are simulation-consistent estimates of

$$E[h(\boldsymbol{\theta})|\text{data}] = \int_{\Theta} h(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\text{data}) d\boldsymbol{\theta}.$$

- $T$  might have to be big, even “massive”...

# Conditions Needed for Ergodicity

- **Harris recurrence** (Definition 4.6, BASS).
  - **irreducibility** (Definition 4.10), BASS: the Markov chain can (eventually) get from regions  $\mathcal{A}$  to  $\mathcal{B}$ ,  $\forall \mathcal{A}, \mathcal{B} \in \Theta$ .
  - **uniqueness of invariant distribution**: the Markov chain has a kernel  $K$  such that
$$p(\boldsymbol{\theta}^{(t+1)}) = \int_{\Theta} p(\boldsymbol{\theta}^{(t)}) K(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) d\boldsymbol{\theta}^{(t)}$$
i.e., iterating the chain doesn't change  $p$ .
- almost every Markov chain we encounter in MCMC has these properties

## Simulation Inefficiency, §4.4.1

- We pay a price for not having independent draws from the posterior density  $p(\boldsymbol{\theta}|\mathbf{y})$ .
- estimand  $h(\boldsymbol{\theta})$ ; we estimate  $E(h(\boldsymbol{\theta})|\mathbf{y})$  --- the mean of the posterior density of  $h(\boldsymbol{\theta})$  --- with the average  $\bar{h}_T = T^{-1} \sum_{t=1}^T h(\boldsymbol{\theta}^{(t)})$
- Ergodic theorem says we have a simulation consistent estimator
- But the rate at which  $\bar{h}_T$  converges on  $E(h(\boldsymbol{\theta})|\mathbf{y})$  --- the rate at which the Monte Carlo error of  $\bar{h}_T$  approaches zero --- is not as fast as the  $\sqrt{T}$  rate we get from an independence sampler.
- Formalizations of this “simulation inefficiency”

# Simulation Inefficiency, §4.4.1

## Definition (Integrated Correlation Time)

Let  $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \dots, \boldsymbol{\Theta}^{(T)}$  be realizations from  $p$ , the stationary distribution of the Markov chain  $\{\boldsymbol{\Theta}^{(t)}\}$ , and let  $h(\boldsymbol{\Theta})$  be some (scalar) quantity of interest. If  $\rho_j$  is the lag- $j$  autocorrelation of the sequence  $\{h(\boldsymbol{\Theta}^{(t)})\}$  then

$$\tau_{\text{int}}[h(\boldsymbol{\Theta})] = \frac{1}{2} + \sum_{j=1}^{\infty} \rho_j.$$

is the integrated autocorrelation time of the chain.

- n.b., for an independence sampler  $\rho_j \approx 0 \forall j \Rightarrow \tau_{\text{int}}[h(\boldsymbol{\Theta})] \approx 1/2$ .

# “Effective sample size” of an ergodic average

- estimand  $h(\boldsymbol{\Theta})$ ; Markov chain  $\{h(\boldsymbol{\Theta}^{(t)})\}$ , stationary distribution  $p$ .
- estimated with ergodic average  $\bar{h}_T = T^{-1} \sum_{t=1}^T h(\boldsymbol{\Theta}^{(t)})$ .
- $\text{var}_p(h_t) = \sigma^2$ .
- But  $\text{var}(\bar{h}_T) = \frac{\sigma^2}{T} \times 2 \times \tau_{\text{int}}[h(\boldsymbol{\Theta})]$ .
- The factor  $2 \times \tau_{\text{int}}[h(\boldsymbol{\Theta})]$  is a measure of how the dependency inherent in the Markovian exploration of  $p(\boldsymbol{\Theta}|\mathbf{y})$  is degrading the precision of the summary statistic  $\bar{h}$ .
- Large and slowly decaying autocorrelations make  $\tau_{\text{int}}[h(\boldsymbol{\Theta})]$  large.
- for an independence sampler

$$2 \times \tau_{\text{int}}[h(\boldsymbol{\Theta})] \approx 2 \times \frac{1}{2} = 1 \Rightarrow \text{var}(\bar{h}_T) \approx \sigma^2 / T$$

## “Effective sample size” of an ergodic average

- See function `effectiveSize` in R package `coda`
- Suppose we have a 1st order Markov chain:

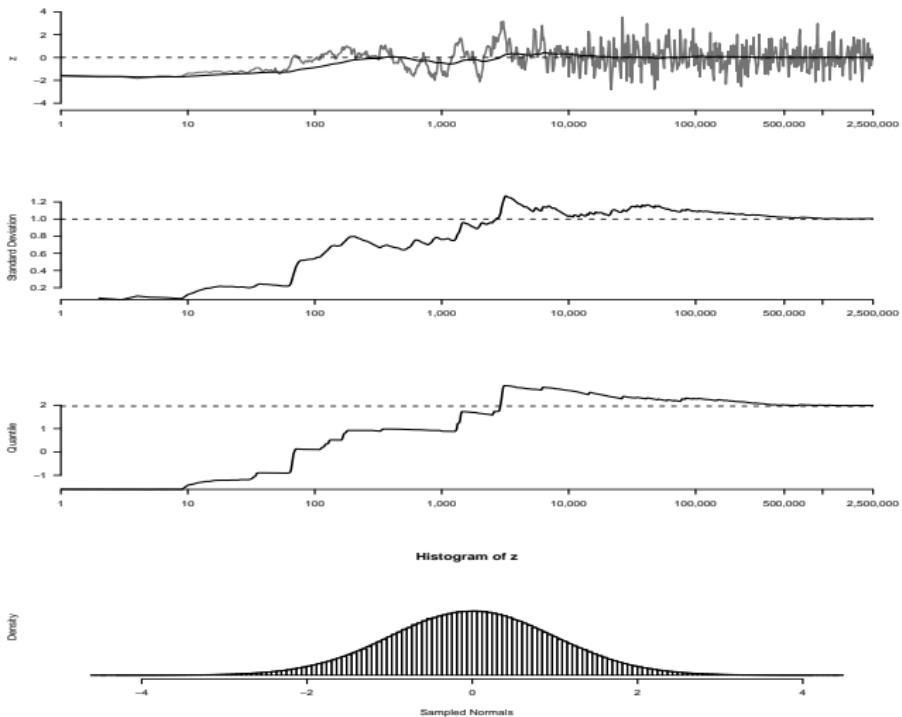
$$E(h_t | h_{t-1}) = \rho h_{t-1}, |\rho| < 1, \text{var}(h_t) = \sigma^2$$

then  $\text{var}(\bar{h}_T) = \frac{\sigma^2}{T} \frac{1 + \rho}{1 - \rho}$ .

- $\rho \rightarrow 0$ , we tend to the independence sampler
- $\rho \rightarrow 1$ , the dependency increases,  $(1 + \rho)/(1 - \rho) \rightarrow \infty$ .
- e.g.,  $\rho = .9$  and we seek a given level of Monte Carlo error in ergodic average  $\bar{h}_T$ .
- $(1 + .9)/(1 - .9) = 1.9/.1 = 19$  or we require  $\sqrt{19} \approx 4.36$  as many iterations of the Markov chain to get the same level of Monte Carlo error as we would if we were using an independence sampler.

# Example 4.12, highly dependent, stationary series

$$z_t = \rho z_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, \omega^2), \omega^2 = 1 - \rho^2, \rho = .995.$$



# Sampling algorithms used in MCMC

- Metropolis-Hastings algorithm; §5.1
- Gibbs sampler; §5.2

# Metropolis-Hastings algorithm

1: sample  $\boldsymbol{\theta}^*$  from a “proposal” or “jumping” distribution  $J_t(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$ .

2:

$$r \leftarrow \frac{p(\boldsymbol{\theta}^* | \mathbf{y}) J_t(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)} | \mathbf{y}) J_t(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)}, \quad (1)$$

3:  $\alpha \leftarrow \min(r, 1)$

4: sample  $U \sim \text{Unif}(0, 1)$

5: if  $U \leq \alpha$  then

6:    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^*$

7: else

8:    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$

9: end if

## Theory for the Metropolis sampler §5.1.1

- Transition kernel  $K(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$  generates a *reversible* Markov chain.
- Reversibility implies  $p(\boldsymbol{\theta}|\mathbf{y})$  is the stationary distribution of the Markov chain.
- Ergodicity follows if we can establish irreducibility and aperiodicity. Sufficiently permissive  $J_t$  accomplishes this.
- e.g.,  $J_t(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) > 0 \forall \boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}$ .
- Aperiodicity follows if  $\Pr(\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}) > 0$ .

# Metropolis-Hastings algorithm

- proposal density  $J_t$  is key to the algorithm
- $t$  subscript for proposal density indicates that the proposal density can evolve, “tuning” the algorithm for an efficient exploration of  $p(\boldsymbol{\theta}|\mathbf{y})$
- original paper is Metropolis et al. (1953) with  $r_M = p(\boldsymbol{\theta}^*|\mathbf{y})/p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})$ .
- modification by Hastings (1970) to give the acceptance ratio given on previous slide
- **Random walk M-H:** select a candidate point  $\boldsymbol{\theta}^*$  by taking a random perturbation around the current point  $\boldsymbol{\theta}^{(t)}$ , i.e.,  $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)} + \boldsymbol{\epsilon}$ . e.g.,
  - $\boldsymbol{\epsilon}_j \sim \text{Unif}(-\delta_j, \delta_j)$ ,  $j = 1, \dots, J$  dimensions of  $\boldsymbol{\theta}$ .
  - $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Omega})$ . Here the key parameter is  $\boldsymbol{\Omega}$ .
- **Independence M-H:** e.g.,  $J = N(\hat{\boldsymbol{\theta}}, c \cdot V(\hat{\boldsymbol{\theta}}))$ ,  $c$  a tuning parameter.

# Random Walk Metropolis, Example 5.1

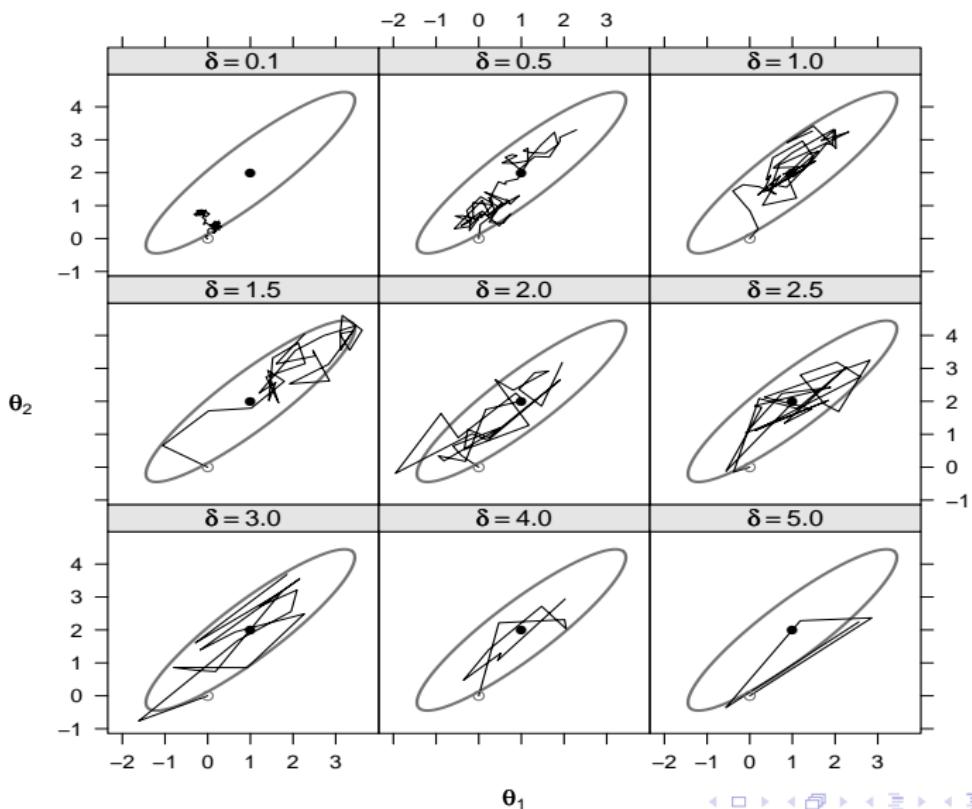
- $\theta \sim N$
- random-walk Metropolis, but what distribution for  $\epsilon$ ?
- Example 5.1:  $\theta \sim N(\mu, \Sigma)$ , where

$$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix},$$

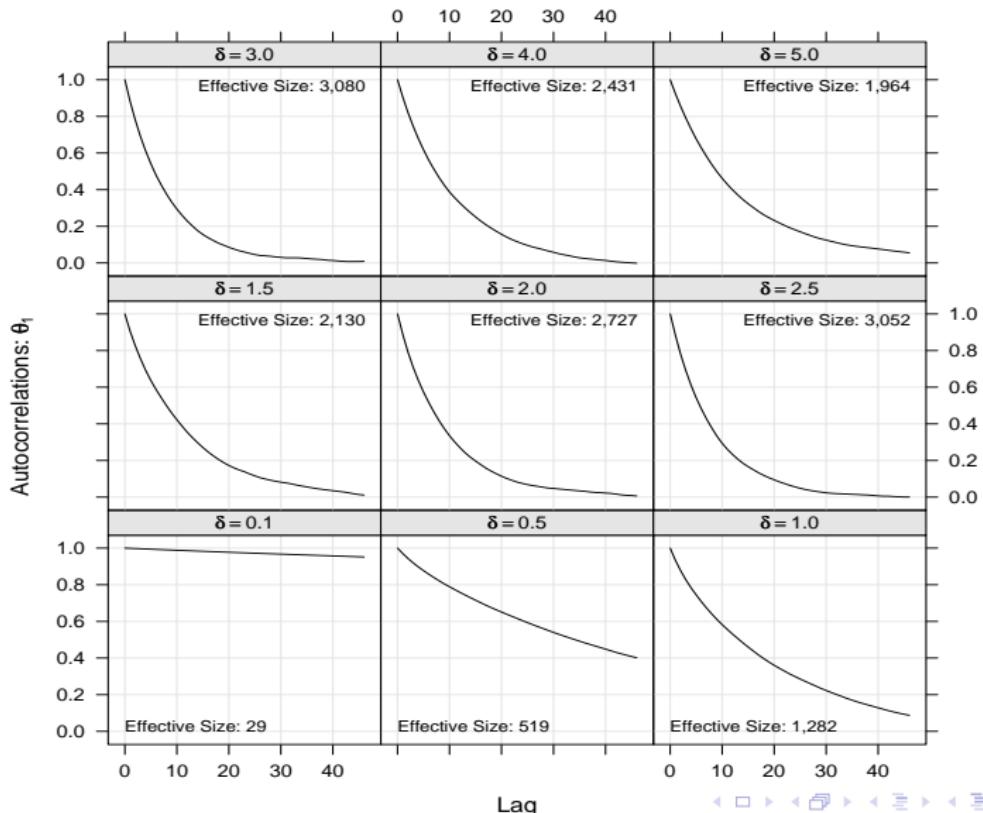
and where  $\epsilon_j \sim \text{Unif}(-\delta, \delta), j = 1, 2$ .

- Consider different choices of  $\delta$ .

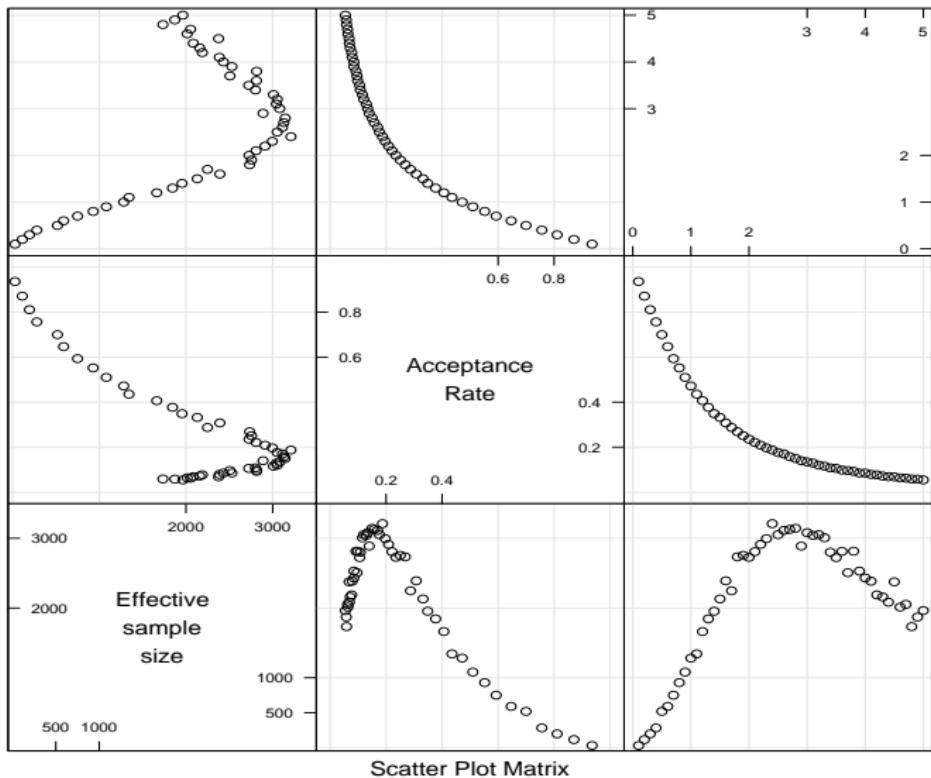
# Random Walk Metropolis, Example 5.1



# Random Walk Metropolis, Example 5.1



# Random Walk Metropolis, Example 5.1



# Random Walk Metropolis, Ex 5.2, Poisson regression

- $y_i | \mathbf{x}_i, \boldsymbol{\beta} \sim \text{Poisson}(\lambda_i), \lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$
- $y_i \in \{0, 1, 2, \dots\}$ ,  $\mathbf{x}_i$  is a vector of covariates,  $\boldsymbol{\beta}$  is a vector of  $k$  unknown coefficients and  $i = 1, \dots, n$  indexes observations.
- likelihood:  $f(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$
- Data: 915 biochemistry graduate students, article counts over last 3 years of PhD studies. Gender differences key.
- Modal number of article counts is zero (30%); 95%-ile is 5, max is 19.
- No conjugate prior for  $\boldsymbol{\beta}$ ; usually just express the posterior in the form it comes from Bayes Rule

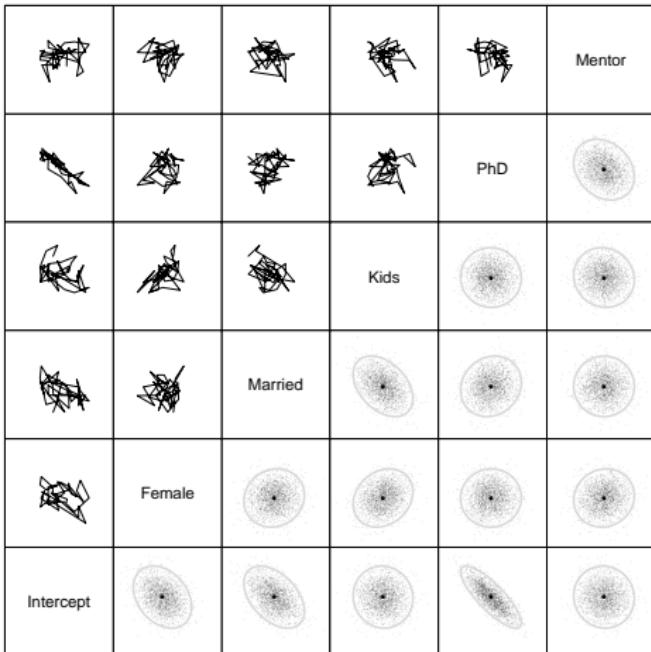
$$p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}) f(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta})$$

# Random Walk Metropolis, Ex 5.2, Poisson regression

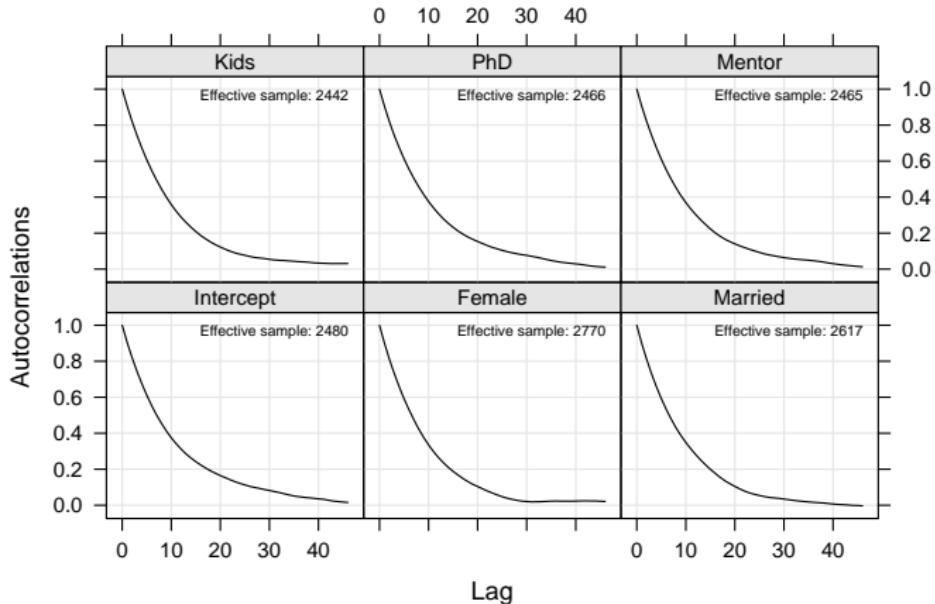
- implementation in R package MCMCpack with the function MCMCpoisson
- multivariate normal prior for  $\beta$ ,  $\beta \sim N(\mathbf{b}_0, \mathbf{B}_0^{-1})$ .
- Metropolis proposal density is  $\beta^* \sim N(\beta^{(t)}, \mathbf{P})$  where  $\mathbf{P} = \mathbf{T}(\mathbf{B}_0 + \mathbf{V}^{-1})^{-1}\mathbf{T}$ , with  $\mathbf{T}$  a  $k$ -by- $k$  diagonal, positive definite matrix containing tuning parameters and  $\mathbf{V}$  is the large-sample approximation to the frequentist sampling covariance matrix of the maximum likelihood estimates  $\hat{\beta}$ .
- default is  $\mathbf{T} = 1.1 \cdot \mathbf{I}$  and to initialize the random-walk Metropolis algorithm at the MLEs  $\hat{\beta}$ .
- we use vague priors, with  $\mathbf{b}_0 = \mathbf{0}$  and  $\mathbf{B}_0 = 10^4 \cdot \mathbf{I}$ .

# Random Walk Metropolis, Ex 5.2, Poisson regression

50,000 iterations of Metropolis algorithm: upper panels show a trace plot of the algorithm in two dimensions, for the first 250 iterations of the algorithm; lower panels summarize the full 50,000 iterations, plotting the algorithm's history at each of 2,500 evenly-spaced iterations over the full 50,000 iterations.



# Random Walk Metropolis, Ex 5.2, Poisson regression



# Random Walk Metropolis, Ex 5.2, Poisson regression

	Bayes	MLE
Intercept	0.30 [0.088, 0.50]	0.30 [0.10, 0.51]
Female	-0.22 [-0.33, -0.12]	-0.22 [-0.33, -0.12]
Married	0.16 [0.044, 0.28]	0.16 [0.035, 0.28]
Kids < 5	-0.18 [-0.27, -0.11]	-0.19 [-0.26, -0.11]
PhD Prestige	0.013 [-0.040, 0.065]	0.014 [-0.039, 0.065]
Mentor Articles	0.026 [0.022, 0.030]	0.026 [0.022, 0.029]

# Gibbs sampler

Suppose  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)'$ .

- 1: **for**  $t = 1$  to  $T$  **do**
- 2:   sample  $\boldsymbol{\theta}_1^{(t+1)}$  from  $g_1(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \dots, \boldsymbol{\theta}_d^{(t)}, \mathbf{y})$ .
- 3:   sample  $\boldsymbol{\theta}_2^{(t+1)}$  from  $g_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \dots, \boldsymbol{\theta}_d^{(t)}, \mathbf{y})$ .
- 4:   ...
- 5:   sample  $\boldsymbol{\theta}_d^{(t+1)}$  from  $g_d(\boldsymbol{\theta}_d \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \dots, \boldsymbol{\theta}_{d-1}^{(t+1)}, \mathbf{y})$ .
- 6:    $\boldsymbol{\theta}^{(t+1)} \leftarrow (\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \dots, \boldsymbol{\theta}_d^{(t+1)})'$ .
- 7: **end for**

# Gibbs sampler

- “divide and conquer”
- sample from lower dimensional conditional densities, given other elements of  $\theta$ .
- it works! See theoretical discussion at §5.2.1.
- joint probability densities completely characterized by component conditional densities

## Example 5.3, Gibbs sampler for bivariate normal

- $\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \dots, n$
- $\boldsymbol{\mu} = (\mu_1, \mu_2)'$  and  $\boldsymbol{\Sigma}$  is a known 2-by-2 covariance matrix
- unknown parameters here are  $\boldsymbol{\theta} = \boldsymbol{\mu} = (\mu_1, \mu_2)'$ .
- Our goal is to compute the posterior density  $p(\boldsymbol{\theta} | \mathbf{y})$ .
- Independent, conjugate prior densities for each element of  $\boldsymbol{\mu}$ , say,  $\boldsymbol{\mu} = (\mu_1, \mu_2)' \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)'$  with

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \sigma_{01}^2 & 0 \\ 0 & \sigma_{02}^2 \end{bmatrix},$$

- Posterior density for  $\boldsymbol{\theta}$  is known in this case; it is bivariate normal.
- But we explore with Gibbs sampler (quite unnecessary for this problem, but helpful for exposition).

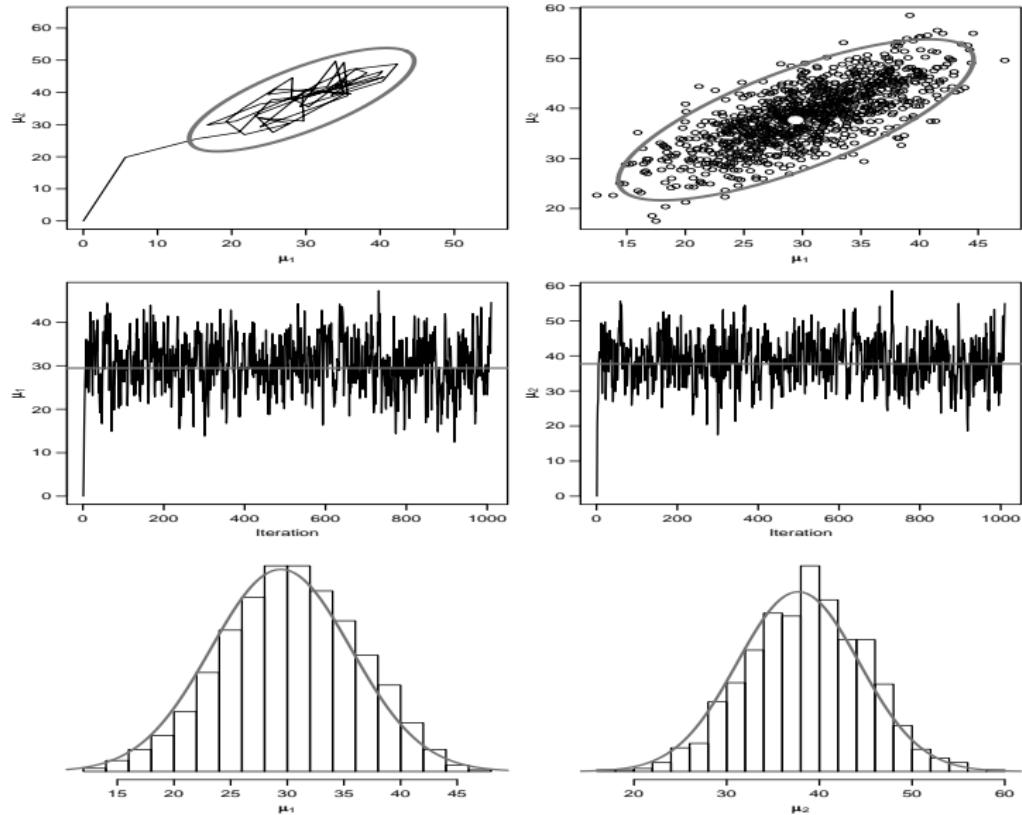
## Example 5.3, Gibbs sampler for bivariate normal

At iteration  $t$ ,

- ① sample  $\mu_1^{*(t)}$  from its conditional distribution  $g_1(\mu_1^* | \mu_2^{*(t-1)}, \Sigma^*, \mathbf{y})$ , a normal density with mean  $\mu_1^* + \frac{\sigma_{12}^*}{\sigma_{22}^*} (\mu_2^{*(t-1)} - \mu_2^*)$  and variance  $\sigma_{11}^* - \sigma_{12}^{*2}/\sigma_{22}^*$
- ② sample  $\mu_2^{*(t)}$  from its conditional distribution  $g_2(\mu_2^* | \mu_1^{*(t)}, \Sigma^*, \mathbf{y})$ , a normal density with mean  $\mu_2^* + \frac{\sigma_{12}^*}{\sigma_{11}^*} (\mu_1^{*(t)} - \mu_1^*)$  and variance  $\sigma_{22}^* - \sigma_{12}^{*2}/\sigma_{11}^*$ .

Note that we condition on  $\mu_2^{*(t-1)}$  when sampling  $\mu_1^{*(t)}$ ; then, given the sampled value  $\mu_1^{*(t)}$ , we condition on it when sampling  $\mu_2^{*(t)}$ .

# Example 5.3, Gibbs sampler for bivariate normal



## Example 5.3, Gibbs sampler for bivariate normal

	Analytic	1 000 iterations	50 000 iterations
$E(\mu_1 \mathbf{y})$	29.44	30.34	29.38
$E(\mu_2 \mathbf{y})$	37.72	38.63	37.65
$V(\mu_1 \mathbf{y})$	38.41	35.54	38.91
$V(\mu_2 \mathbf{y})$	43.17	40.12	43.84
$C(\mu_1, \mu_2 \mathbf{y})$	30.59	28.07	31.12

# Conditional distributions for the Gibbs sampler

## Theorem

If a statistical model can be expressed as a directed acyclic graph (a DAG)  $\mathcal{G}$ , then the conditional density of node  $\theta_j$  in the graph is

$$f(\theta_j | \mathcal{G} \setminus \theta_j) \propto f(\theta_j | \text{parents}[\theta_j]) \times \prod_{w \in \text{children}[\theta_j]} f(w | \text{parents}[w]), \quad (2)$$

where  $\mathcal{G} \setminus \theta_j$  stands for all nodes in  $\mathcal{G}$  other than  $\theta_j$ .

## Proof.

See Spiegelhalter and Lauritzen (1990). □

## Example 5.6, 2-level hierarchical model

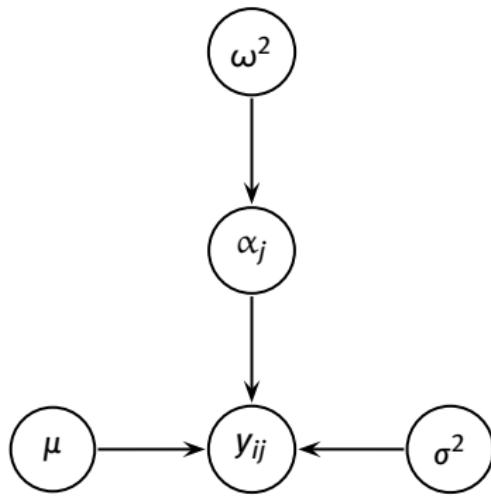
- We have multiple observations  $i = 1, \dots, n$  for each of  $j = 1, \dots, J$  units (e.g., students indexed by  $i$  in schools indexed by  $j$ ) on a real-valued variable  $y_{ij}$ .
- Model:

$$\begin{aligned}y_{ij} | \mu, \alpha_j, \sigma^2 &\sim N(\mu + \alpha_j, \sigma^2) \\ \alpha_j &\sim N(0, \omega^2)\end{aligned}$$

- Likelihood:  $f(\mathbf{Y} | \mu, \alpha, \sigma^2) = \prod_{i=1}^n \prod_{j=1}^J \phi\left(\frac{y_{ij} - \mu - \alpha_j}{\sigma}\right)$
- The hyper-parameter  $\omega^2$  is referred to as the *between* unit variance, while  $\sigma^2$  is the *within* unit variance.
- Priors on  $\mu$ ,  $\omega^2$  and  $\sigma^2$ : *a priori* independence for these parameters,  $p(\mu, \omega^2, \sigma^2) = p(\mu)p(\omega^2)p(\sigma^2)$ .
- $\boldsymbol{\theta} = (\mu, \alpha, \sigma^2, \omega^2)'$

## Example 5.6; Figure 5.11

- $y_{ij}$  is a child of  $\mu, \alpha_j, \sigma^2$ ;  $\mu$  etc are the parents of  $y_{ij}$  etc.
- $\alpha_j$  is a child of  $\omega^2$ .
- $y_{ij}$  is conditionally independent of  $\omega^2$  given  $\alpha_j$ ; i.e.,  
 $\{y_{ij} \perp\!\!\!\perp \omega^2\} | \alpha_j, \forall i, j$ .
- $\{\alpha_j \perp\!\!\!\perp \alpha_k\} | \omega^2, \forall j \neq k$ .



## Example 5.6; Figure 5.11; Gibbs sampler

- ① sample  $\sigma^{2(t)}$  from  $g(\sigma^2 | \mathcal{G}_{-\sigma^2}) = g(\sigma^2 | \mathbf{Y}, \mu, \alpha) \propto f(\mathbf{Y} | \mu, \alpha, \sigma^2) p(\sigma^2)$
- ② sample  $\omega^{2(t)}$  from  $g(\omega^2 | \mathcal{G}_{-\omega^2}) = g(\omega^2 | \alpha)$ , noting that  $\{\omega^2 \perp (\mathbf{Y}, \mu, \sigma^2)\} | \alpha$ . Thus,  $g(\omega^2 | \alpha) \propto f(\alpha | \omega^2) p(\omega^2)$ .
- ③ for  $j = 1, \dots, J$ , sample  $\alpha_j^{(t)}$  from  $g(\alpha_j | \mathcal{G}_{-\alpha_j}) = g(\alpha_j | \mathbf{y}_j, \sigma^2, \omega^2, \mu)$ , where  $\mathbf{y}_j$  is a vector of the observations from unit  $j$ , and noting that  $\{\alpha_j \perp \alpha_k\} | \omega^2 \forall j \neq k$ ; i.e.,

$$\begin{aligned} g(\alpha_j | \mathbf{y}_j, \sigma^2, \omega^2, \mu) &\propto f(\mathbf{y}_j | \mu, \alpha_j, \sigma^2) p(\alpha_j | \omega^2) \\ &= \prod_{i=1}^n \phi\left(\frac{y_{ij} - \mu - \alpha_j}{\sigma}\right) \cdot \phi\left(\frac{\alpha_j}{\omega}\right) \end{aligned}$$

- ④ sample  $\mu^{(t)}$  from  $g(\mu | \mathcal{G}_{-\mu}) = g(\mu | \mathbf{Y}, \alpha, \sigma^2)$ , since  $\{\mu \perp \omega^2\} | \alpha$ ; i.e.,

$$g(\mu | \mathbf{Y}, \alpha, \sigma^2) \propto f(\mathbf{Y} | \mu, \alpha, \sigma^2) p(\mu) = \prod_{i=1}^n \prod_{j=1}^J \phi\left(\frac{y_{ij} - \mu - \alpha_j}{\sigma}\right) \cdot p(\mu)$$

# Implementation in JAGS

JAGS code

```
model{  
    ## loop over data frame  
    for(i in 1:N){  
        ## expression for E(y[i])  
        ## note double-subscript on alpha  
        ymu[i] <- mu + alpha[j[i]]  
  
        ## sampling model for y[i]  
        y[i] ~ dnorm(ymu[i],tau.sigma)  
    }  
  
    ## hierarchical model for alphas  
    for(i in 1:J){  
        alpha[i] ~ dnorm(0,tau.omega)  
    }  
  
    ## predictions for a future election?  
    for(i in 1:J){  
        muFuture[i] <- mu + alpha[i]  
        yFuture[i] ~ dnorm(muFuture[i],tau.sigma)  
    }  
  
    ## prior for mu  
    mu ~ dnorm(0,.01)  
  
    ## prior for standard deviations (not variances!)  
    sigma ~ dunif(0,10)  
    omega ~ dunif(0,10)  
    tau.sigma <- pow(sigma,-2) ## precision!  
    tau.omega <- pow(omega,-2) ## precision!  
}
```



# References

- Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains, and their applications." *Biometrika* 57:97--109.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. "Equations of state calculations by fast computing machines." *Journal of Chemical Physics* 21:1087--91.
- Spiegelhalter, David J. and S. L. Lauritzen. 1990. "Sequential updating of conditional probabilities on directed graphical structures." *Networks* 20:579--605.

# HIERARCHICAL MODELS

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 11, 2012

# Hierarchical Models

- data span multiple groups or time periods
- “context matters”
- group-specific statistical structure to data

$y_j|\theta_j \sim p(y_j|\theta_j)$  (model for the data in group  $j = 1, \dots, J$ )

$\theta_j|v \sim p(\theta_j|v)$  (between-group model or “prior” for the parameters  $\theta_j$ )

$v \sim p(v)$  (prior for the *hyperparameter*  $v$ ),

# Example: one-way analysis of variance

$$y_{ij} | \alpha_j, \sigma^2 \sim N(\alpha_j, \sigma^2) \quad (1a)$$

$$\alpha_j | \mu, \omega^2 \sim N(\mu, \omega^2). \quad (1b)$$

- $i$  indexes observations;  $j$  indexes  $J$  groups
- $\alpha_j$ : mean of  $y$  in group  $j$
- equation 1b is a model for how  $\alpha_j$  varies across groups.
- $\mu$  is the mean of the distribution of the group means (the “grand mean”)
- variance  $\omega^2$ , also known as the *between* variance;
- $\sigma^2$  is known as the *within* variance for group  $j$ ; constant across groups here, could relax this and have  $\sigma_j^2$  (group-wise heteroskedasticity)
- $\omega^2 / (\omega^2 + \sigma^2)$  is known as the *intra-class correlation* and is a measure of the “relative similarity” of observations in each group
- Bayesian analysis: need priors for  $\mu$ ,  $\sigma^2$  and  $\omega^2$ .

## Example: multilevel regression

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma_j^2) \quad (2a)$$

$$\boldsymbol{\beta}_{jk} \sim N(\mathbf{z}_j\boldsymbol{\gamma}_k, \omega_k^2) \quad (2b)$$

- $i$  indexes observations;  $j$  indexes  $J$  groups
- $k$  indexes  $K$  covariates; i.e.,  $\mathbf{x}_{ij}\boldsymbol{\beta}_j = x_{ij1}\beta_{j1} + \dots + x_{ijk}\beta_{jk}$
- need priors for  $\boldsymbol{\gamma}_k$  and  $\omega_k^2$ ,  $k = 1, \dots, K$ ; priors for  $\sigma_j^2$ ,  $j = 1, \dots, J$ .

# Representation as a Mixed Model

$$\mathbf{y} | \mu, \alpha, \sigma^2, \omega^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}) \quad (3)$$

- $\mathbf{X}$  is a  $n$ -by- $k$  matrix of predictors that pick up fixed effects  $\boldsymbol{\beta}$  (a  $k$ -by-1 vector of coefficients)
- $\mathbf{Z}$  is a  $n$ -by- $p$  matrix of predictors that pick up random effects  $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Omega})$ , where  $\boldsymbol{\Omega}$  is a  $p$ -by- $p$  covariance matrix.
- $\boldsymbol{\Sigma}$  is a  $n$ -by- $n$  covariance matrix.
- i.e., the effect of  $\mathbf{x}$  in group  $j$  is  $\beta + b_j$ .
- $\text{cov}(b_j, \varepsilon_{ij} | \mu) = 0 \forall i, j$ .

# Variance Components Representation

- One-way ANOVA decomposes variation in  $\mathbf{y}$  around  $\mu$  into two components: the “within-group” variance  $\sigma^2$  and the “between-group” variance  $\omega^2$ .

$$\text{var}(\mathbf{y}) = \text{var}(\mu) + \underbrace{\mathbf{Z}\text{var}(\mathbf{b})\mathbf{Z}'}_{\text{“between variance”}} + \underbrace{\text{var}(\boldsymbol{\epsilon})}_{\text{“within variance”}} \quad (4a)$$

$$= \omega^2 \mathbf{Z}\mathbf{I}_J\mathbf{Z}' + \sigma^2 \mathbf{I}_n \quad (4b)$$

$$= \omega^2 \mathbf{F} + \sigma^2 \mathbf{I}_n, \quad (4c)$$

- $\mathbf{F} = \mathbf{Z}\mathbf{Z}'$  is a block diagonal matrix with blocks  $\mathbf{1}_{n_j} \mathbf{1}_{n_j}'$  (a square  $n_j$ -by- $n_j$  matrix of ones),  $j = 1, \dots, J$ .

# Variance Components Representation

$$\text{var}(\mathbf{y}) = \text{var}(\mu) + \underbrace{\mathbf{Z}\text{var}(\mathbf{b})\mathbf{Z}'}_{\text{"between variance"}} + \underbrace{\text{var}(\boldsymbol{\epsilon})}_{\text{"within variance"}}$$

- for group  $j$ , we have

$$\text{var}(\mathbf{y}_j) = \omega^2 \mathbf{l}_{n_j} \mathbf{l}'_{n_j} + \sigma^2 \mathbf{I}_{n_j} = \begin{bmatrix} \sigma^2 + \omega^2 & \omega^2 & \dots & \omega^2 \\ \omega^2 & \sigma^2 + \omega^2 & \dots & \omega^2 \\ \vdots & \vdots & \ddots & \vdots \\ \omega^2 & \omega^2 & \dots & \sigma^2 + \omega^2 \end{bmatrix},$$

- non-zero covariance across observations *within* groups; these observations share the group-specific term  $b_j \sim N(0, \omega^2)$ .
- within-cluster covariance is explicit
- Contrast classical approaches that treat the “clustered” nature of the data as a nuisance; inference for the fixed effects with “cluster robust” standard errors.

# Exchangeable parameters generate hierarchical models

- introduced exchangeability earlier
- extend concept from data to parameters
- generates models for parameters
- may require covariates if not unconditionally exchangeable; i.e., parameters in hierarchical model induce conditional exchangeability

# Example: Exchangeability and hierarchical models for polling data

- Suppose we have data from a survey conducted in  $J$  counties in the United States. Individual level responses, support for border protection:  $y_{ij} = 1$  if respondent  $i$  wants more spending on border protection and 0 otherwise.
- No individual-level predictors with which to model the responses, and so exchangeability is a reasonable assumption at the micro-level. Thus,  $r_j \sim \text{Binomial}(\theta_j, n_j)$ ,  $r_j = \sum_{i=1}^n y_{ij}$
- $d_j$  is distance of county  $j$  from US border

$$\begin{aligned} r_j &\sim \text{Binomial}(\theta_j, n_j) \\ \log\left(\frac{\theta_j}{1 - \theta_j}\right) &\sim N(\beta_0 + \beta_1 d_j, \omega^2), \\ (\beta_0, \beta_1)' &\sim N(\mathbf{b}, \Sigma) \end{aligned}$$

# Hierarchical models “borrow strength” across units

$$\begin{aligned}y_{ij} | \alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j | \mu, \omega^2 &\sim N(\mu, \omega^2).\end{aligned}$$

- inferences for the group-level parameters  $\alpha_j$  reflect not just the information about  $\alpha_j$  in group  $j$ , but, via the hierarchical model, will also draw on relevant information in the other groups.
- information about the  $\alpha_j$  flows “up” the hierarchy to inform inferences about the distribution of the  $\alpha_j$  across groups
- i.e., data informative for  $\mu$  and  $\omega^2$  too
- data from group  $j$  helps shape the posterior density over  $\alpha_k$  ( $\forall k \neq j$ ) via contribution to inferences for the hyperparameters  $\mu$  and  $\omega^2$ .
- “sharing” or “borrowing” information across groups follows from exchangeability/hierarchical models

# Hierarchical model as “semi-pooling”

$$\begin{aligned}y_{ij} | \alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j | \mu, \omega^2 &\sim N(\mu, \omega^2).\end{aligned}$$

- Compare two other “extreme” models:
- **No pooling:**  $y_{ij} | \alpha_j, \sigma^2 \sim N(\alpha_j, \sigma^2)$ , dropping the hierarchical component of the model.
- Equivalent to setting  $\omega^2 = \infty$  in  $\alpha_j \sim N(\mu, \omega^2)$ .
- **Complete pooling:** grouping in the data is irrelevant, and we impose the restriction that  $\alpha_j = \mu, \forall j$ , generating the model  $y_{ij} \sim N(\mu, \sigma^2)$ .
- Equivalent to setting  $\omega^2 = 0$  in  $\alpha_j \sim N(\mu, \omega^2)$ .
- **Hierarchical model** lies between these two extreme cases

# Hierarchical Model as a “Shrinkage” Estimator

$$\begin{aligned}y_{ij} | \alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j | \mu, \omega^2 &\sim N(\mu, \omega^2).\end{aligned}$$

- Bayes estimates of  $\alpha_j$  are:
  - ① shifted away from the group-level mean  $\bar{y}_j$  in the direction of  $\mu$ , the mean of the distribution of the group level parameters.
  - ② are more precise than inferences based on an analysis of group  $j$  in isolation from the other groups.
- “shrinkage”: the  $\alpha_j$  are pulled towards the grand mean  $\mu$ , relative to the distribution of group level parameters we obtain with no pooling.
- Stein’s (1955) result: The Bayesian “semi-pooled” or “shrinkage” estimator dominates both the “no pooling” and “complete pooling” estimators with respect to total mean square error.

## Theorem (Bayes estimates, one-way ANOVA as hierarchical model)

If  $y_{ij} \sim N(\alpha_j, \sigma_j^2)$ ,  $\alpha_j \sim N(\mu, \omega^2)$  then  $\alpha_j | \mathbf{y}_j, \sigma_j^2, \mu, \omega^2 \sim N(\tilde{\mu}_j, V_j)$  where

$$\tilde{\mu}_j = \frac{\mu\omega^{-2} + \bar{y}_j \frac{n_j}{\sigma_j^2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} \quad \text{and} \quad V_j = \left( \omega^{-2} + \frac{n_j}{\sigma_j^2} \right)^{-1}.$$

Equivalently,

$$\begin{aligned}\tilde{\mu}_j &= \lambda_j \mu + (1 - \lambda_j) \bar{y}_j \\ \lambda_j &= \frac{\omega^{-2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} = \frac{\frac{\sigma_j^2}{n_j}}{\omega^2 + \frac{\sigma_j^2}{n_j}} = \frac{V(\bar{y}_j)}{V(\bar{y}_j) + \omega^2}\end{aligned}$$

and  $\lambda_j$  is a measure of how much  $\alpha_j$  is “shrunk” away from  $\bar{y}_j$  towards  $\mu$ .

# Bayes estimate, one-way ANOVA as hierarchical model

$$\tilde{\mu}_j = \lambda_j \mu + (1 - \lambda_j) \bar{y}_j$$

where

$$\lambda_j = \frac{\omega^{-2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} = \frac{\frac{\sigma_j^2}{n_j}}{\omega^2 + \frac{\sigma_j^2}{n_j}} = \frac{V(\bar{y}_j)}{V(\bar{y}_j) + \omega^2}$$

- familiar “precision-weighted” average form
- when group  $j$  provides little information about  $\alpha_j$  in group  $j$  --- e.g.,  $n_j$  is small, and so  $V(\bar{y}_j)$  is large, relative to the between variance  $\omega^2$  --- then the shrinkage factor  $\lambda_j$  grows, and the Bayes estimate of  $\alpha_j$  is pulled towards the grand mean  $\mu$ .
- if the between variance  $\omega^2$  is relatively large --- e.g., groups are quite heterogeneous -- then the Bayes estimate of  $\alpha_j$  will display less shrinkage, relying more on information in group  $j$  and less “borrowing strength” from other groups.

# Computation via Markov chain Monte Carlo

- Easy under conjugacy: normal data, normal prior for group-specific  $\alpha_j$ , normal for hyperparameter  $\mu$ , inverse-Gamma for within-variance  $\sigma^2$  and between-variance  $\omega^2$ .
- not necessary: e.g.,  $\sigma_j \sim \text{Unif}(0, k)$
- DAG structure makes hierarchical models well suited for general-purpose solutions like BUGS/JAGS
- Many other programs too: MLWin, HLM, lme4 package in R

# One way ANOVA, conditionally conjugate hierarchical model

$$y_{ij} | \alpha_j, \sigma^2 \sim N(\alpha_j, \sigma^2)$$

$$\alpha_j | \mu_0, \omega^2 \sim N(\mu_0, \omega^2)$$

$$\mu_0 \sim N(b_0, B_0)$$

$$\sigma^2 \sim \text{inverse-Gamma}(v_0/2, v_0\sigma_0^2/2)$$

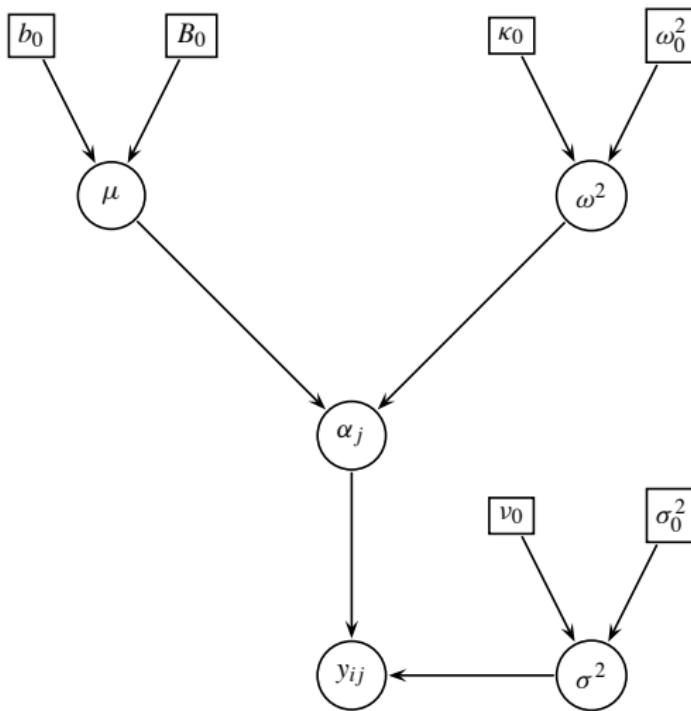
$$\omega^2 \sim \text{inverse-Gamma}(\kappa_0/2, \kappa_0\omega_0^2/2)$$

- $\boldsymbol{\Theta} = (\alpha_1, \dots, \alpha_J, \mu_0, \sigma^2, \omega^2)$

- prior:

$$\begin{aligned} p(\boldsymbol{\Theta}) &= p(\alpha_1, \dots, \alpha_J, \mu_0, \sigma^2, \omega^2) \\ &= p(\alpha_1, \dots, \alpha_J, | \mu_0, \omega^2) p(\mu_0) p(\sigma^2) p(\omega^2) \\ &= \prod^J p(\alpha_j | \mu_0, \omega^2) p(\mu_0) p(\sigma^2) p(\omega^2) \end{aligned}$$

# DAG for one way ANOVA as hierarchical model



# Conditional distributions, Gibbs sampler

- $p(\alpha_j | \mathcal{G} \setminus \alpha_j), j = 1, \dots, J$ : parents of each  $\alpha_j$  are  $\mu_0$  and  $\omega^2$ ; the children of  $\alpha_j$  are the data in unit  $j$ ,  $\mathbf{y}_j = (y_1, \dots, y_{n_j})'$  and the parents of  $\mathbf{y}_j$  are  $\alpha_j$  and  $\sigma^2$ .

$$\alpha_j | (\mathcal{G} \setminus \alpha_j) \sim N \left( \frac{\mu_0 \omega^{-2} + \bar{y}_j \frac{n_j}{\sigma^2}}{\omega^{-2} + \frac{n_j}{\sigma^2}}, \left( \omega^{-2} + \frac{n_j}{\sigma^2} \right)^{-1} \right),$$

- $p(\mu_0 | \mathcal{G} \setminus \mu_0)$ . Parents of  $\mu_0$  are just its prior hyperparameters, the prior mean and variance  $b_0$  and  $B_0$  respectively. The children of  $\mu_0$  are  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)'$ . The  $\boldsymbol{\alpha}$  have two parents,  $\mu_0$  and  $\omega^2$ .

$$\mu_0 | (\mathcal{G} \setminus \mu_0) \sim N \left( \frac{b_0 B_0^{-1} + \bar{\mu} \frac{J}{\omega^2}}{B_0^{-1} + \frac{J}{\omega^2}}, \left( B_0^{-1} + \frac{J}{\omega^2} \right)^{-1} \right),$$

where  $\bar{\mu} = J^{-1} \sum_{j=1}^J \alpha_j$ .

# Conditional distributions, Gibbs sampler

- $p(\omega^2 | \mathcal{G} \setminus \omega^2)$ . The parents of  $\omega^2$  are just its prior hyperparameters,  $\kappa_0$  and  $\omega_0^2$ . The children of  $\omega^2$  are the  $\alpha_j$ ; the parents of  $\alpha_j$  are  $\omega^2$  and  $\mu_0$ .

$$\omega^2 | (\mathcal{G} \setminus \omega^2) \sim \text{inverse-Gamma} \left( \frac{\kappa_0 + J}{2}, \frac{\kappa_0 \omega_0^2 + S_\mu}{2} \right)$$

where  $S_\mu = \sum_{j=1}^J (\alpha_j - \mu_0)^2$ .

- $p(\sigma^2 | \mathcal{G} \setminus \sigma^2)$ . The parents of  $\sigma^2$  are just its prior hyperparameters,  $v_0$  and  $\sigma_0^2$ . The children of  $\sigma^2$  are the  $y_{ij}$ ; the parents of the  $y_{ij}$  are the  $\alpha_j$  and  $\sigma^2$ .

$$\sigma^2 | \mathcal{G} \setminus \sigma^2 \sim \text{inverse-Gamma} \left( \frac{v_0 + n}{2}, \frac{v_0 \sigma_0^2 + S_Y}{2} \right)$$

where  $n = \sum_{j=1}^J n_j$  is the total number of observations and  $S_Y = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \alpha_j)^2$  is the total sum-of-squares of Y.

# Example 7.6, one way ANOVA , HSB

---

JAGS code

---

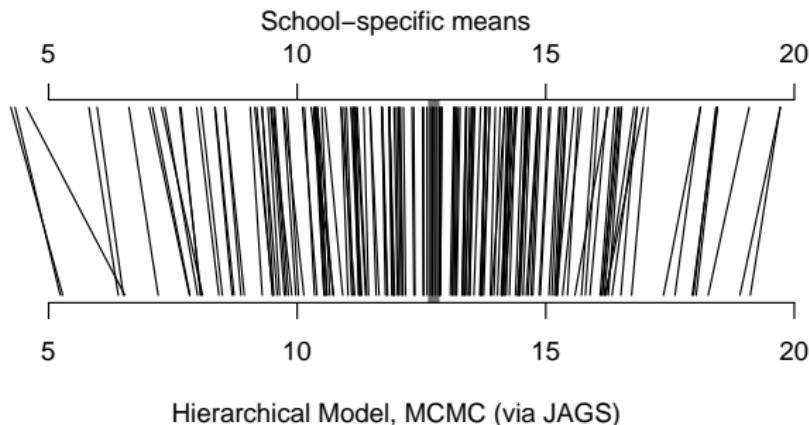
```
model{
    for(i in 1:N) {
        mu.y[i] <- mu[j[i]]
        math[i] ~ dnorm(mu.y[i],tau[1])
    }

    for(p in 1:J) {
        mu[p] ~ dnorm(mu0,tau[2])
    }

    mu0 ~ dnorm(0,.0001)
    for(p in 1:2) {
        tau[p] <- pow(sigma[p],-2)
        sigma[p] ~ dunif(0,10)
    }
}
```

## Example 7.6, one way ANOVA, HSB

- The Bayes estimates of  $\alpha_j$  --- means of the marginal posterior densities of the  $\alpha_j$ , or  $E(\alpha_j | \mathbf{y}, \sigma^2 \omega^2)$  --- are “shrunk” towards the grand mean by the hierarchical model relative to the MLEs.
- The MLEs are simply the sample means, i.e.,  $\hat{\alpha}_j^{(MLE)} = \bar{y}_j$ .



## 2-way ANOVA, state and year effects in presidential elections data, Example 7.7

- Two-levels of grouping: state  $i = 1, \dots, 50$  and year  $t = 1984, 1988, \dots, 2004$ .
- Model:

$$\begin{aligned}y_{it} &\sim N(\mu + \alpha_i + \delta_t, \sigma^2) \\ \mu &\sim N(50, 15^2) \\ \alpha_i &\sim N(0, \sigma_\alpha^2) \\ \delta_t &\sim N(0, \sigma_\delta^2) \\ \sigma_\alpha &\sim \text{Unif}(0, 15) \\ \sigma_\delta &\sim \text{Unif}(0, 15)\end{aligned}$$

- Slow-mixing MCMC algorithm

## Example 7.7: slow-mixing for $\mu$

50,000 iterations, thinned by 5:

Parameter	Geweke	Heidelberger-Welch	Raftery-Lewis	
	$z$	$p$	$N$	$I$
$\mu$	0.74	0.82	256665	68.50
$\sigma$	-0.53	0.82	18705	4.99
$\sigma_\alpha$	-0.61	0.99	18550	4.95
$\sigma_\delta$	-1.83	0.89	19170	5.12

## Example 7.7: over-parameterization for better mixing

$$\begin{aligned}y_{it} &\sim N(\mu + \alpha_i + \delta_t, \sigma^2) \\ \mu &\sim N(0, 100^2) \\ \alpha_i &\sim N(\mu_\alpha, \sigma_\alpha^2) \\ \delta_t &\sim N(\mu_\delta, \sigma_\delta^2) \\ \mu_\alpha &\sim N(0, 100^2) \\ \mu_\delta &\sim N(0, 100^2)\end{aligned}$$

- $\mu, \mu_\alpha$  and  $\mu_\delta$  not identified.
- Map back to identified parameters by imposing the restrictions

$$\sum_{i=1}^n \alpha_i = 0 \Rightarrow \bar{\alpha} = 0 \quad \sum_{t=1}^T \delta_t = 0 \Rightarrow \bar{\delta} = 0$$

- apply these identifying restrictions in JAGS or by *post-processing* the MCMC output in R

## Example 7.7: over-parameterization for better mixing

- At iteration  $m$ , define

$$\begin{aligned}\alpha_i^{*(m)} &= \alpha_i^{(m)} - \bar{\alpha}^{(m)}, \quad i = 1, \dots, n \\ \delta_t^{*(m)} &= \delta_t^{(m)} - \bar{\delta}^{(m)}, \quad t = 1, \dots, T \\ \mu^{*(m)} &= \mu^{(m)} + \bar{\alpha}^{(m)} + \bar{\delta}^{(m)}.\end{aligned}$$

- n.b., simply a re-parameterization; we get the same likelihood contributions either way, since

$$\begin{aligned}\mu^* + \alpha_i^* + \delta_i^* &= \mu + \bar{\alpha} + \bar{\delta} + \alpha_i - \bar{\alpha} + \delta_i - \bar{\delta} \\ &= \mu + \alpha_i + \delta_i.\end{aligned}$$

# Example 7.7: over-parameterization for better mixing

JAGS code

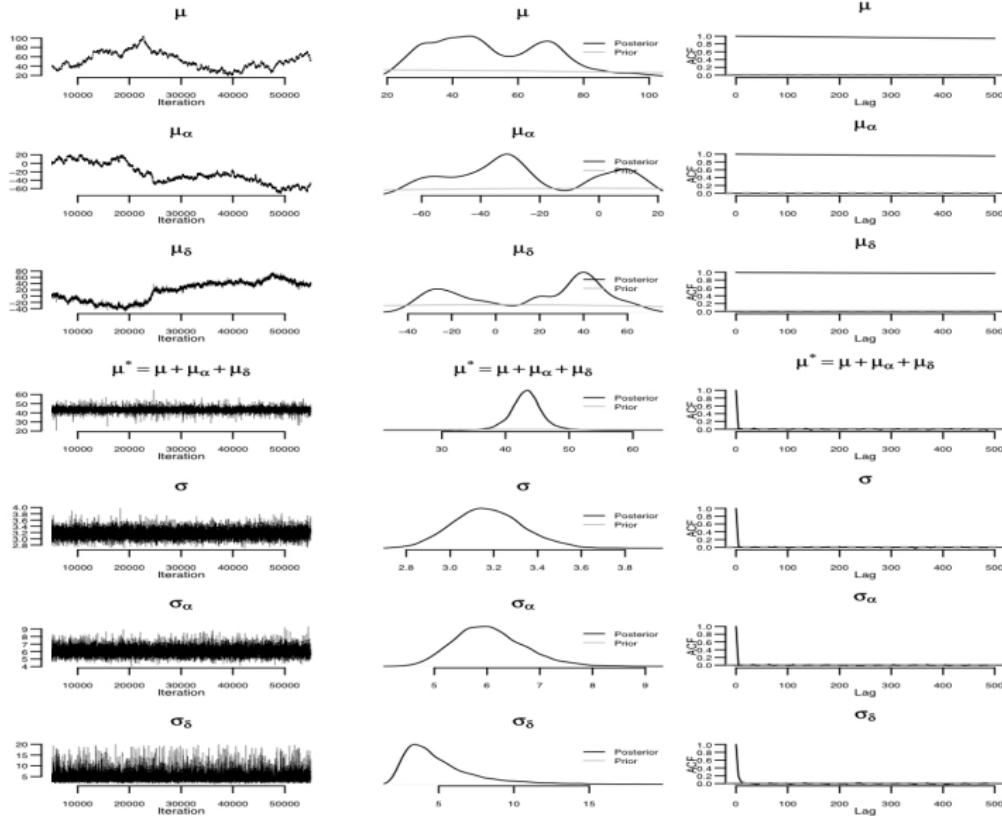
```
model{  
    for(i in 1:n){  
        mu.y[i] <- mu[1] + alpha[s[i]] + delta[j[i]]  
        demVote[i] ~ dnorm(mu.y[i],tau[1])  
    }  
    sigma[1] ~ dunif(0,20)  
    sigma[2] ~ dunif(0,20)  
    sigma[3] ~ dunif(0,20)  
  
    for(i in 1:50){  
        alpha[i] ~ dnorm(mu[2],tau[2])  
    }  
    for(i in 1:nyear){  
        delta[i] ~ dnorm(mu[3],tau[3])  
    }  
    for(i in 1:3){  
        tau[i] <- pow(sigma[i],-2)  
    }  
    for(i in 1:3){  
        mu[i] ~ dnorm(0,1E-4)  
    }  
  
    ## transformations for identified parameters  
    mustar <- mu[1] + mean(alpha[]) + mean(delta[])  
    for(i in 1:50){  
        alphastar[i] <- alpha[i] - mean(alpha[])  
    }  
    for(i in 1:nyear){  
        deltastar[i] <- delta[i] - mean(delta[])  
    }  
}
```



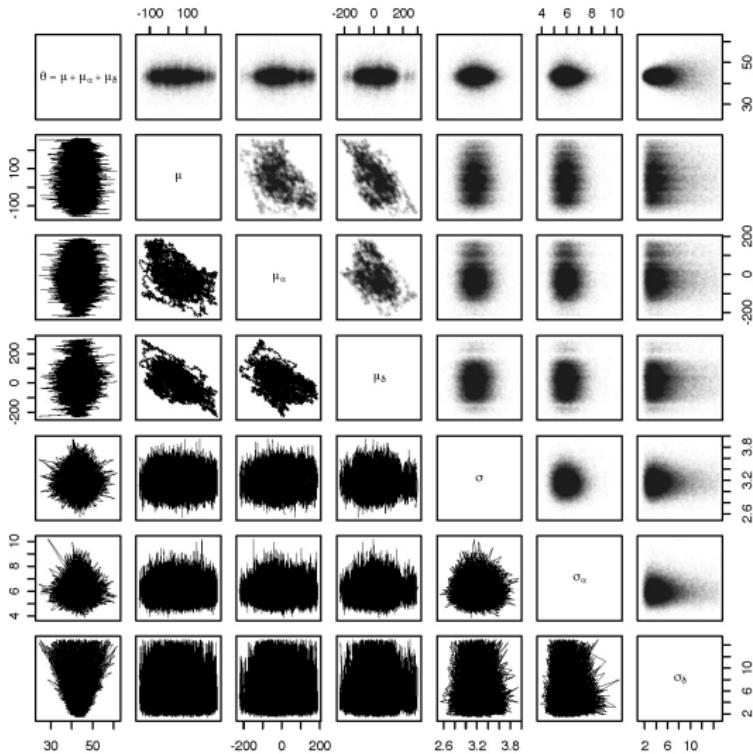
## Example 7.7: over-parameterization for better mixing

Parameter	Geweke <i>z</i>	Heidelberger-Welch <i>p</i>	Raftery-Lewis <i>N</i>	<i>I</i>
$\mu^* = \mu + \bar{\alpha} + \bar{\delta}$	-0.58	0.91	19645	5.24
$\sigma$	-0.34	0.53	18855	5.03
$\sigma_\alpha$	-0.22	0.21	19010	5.07
$\sigma_\delta$	1.25	0.75	18705	4.99

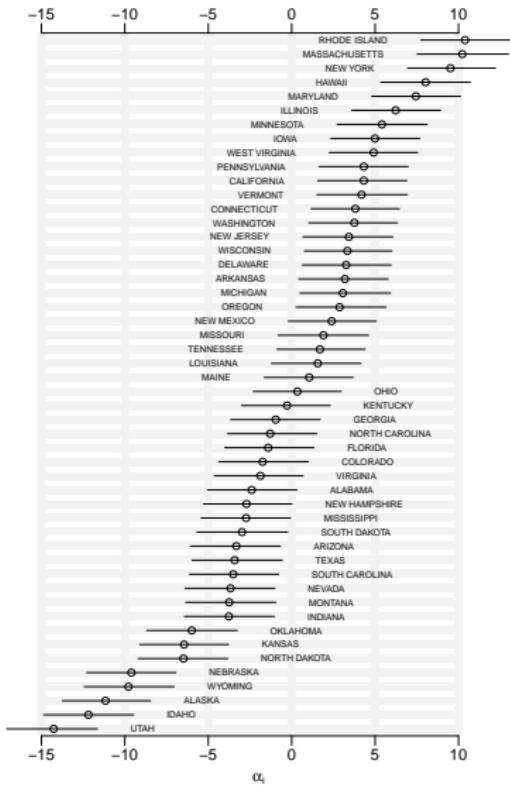
# Example 7.7: over-parameterization for better mixing



# Ex 7.7: over-parameterization for better mixing



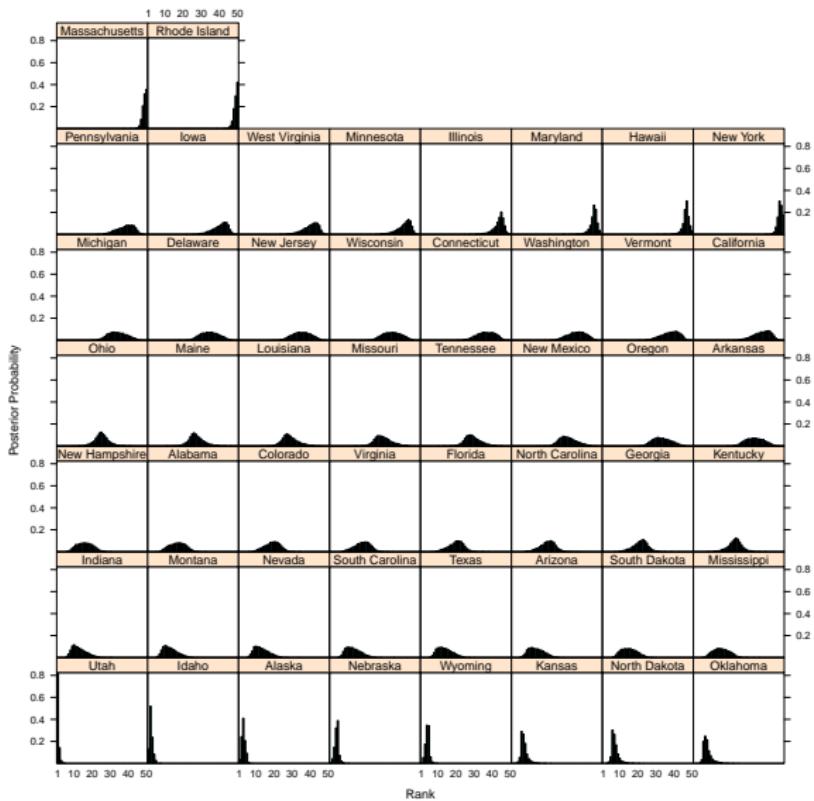
# Ex 7.7: marginal posterior densities, state effects $\alpha_i$



## Extension of Ex 7.7: inducing a posterior mass function over ranks of $\alpha_i$

- At each iteration of the Gibbs sampler we have  $\boldsymbol{\alpha}^{(t)} = (\alpha_1^{(t)}, \dots, \alpha_n^{(t)})'$
- Compute ranks: to produce  $\mathbf{r}^{(t)} = (r_1^{(t)}, \dots, r_n^{(t)})'$ ,  $r_i \in \{1, 2, \dots, n\} \forall i$ .
- A simulation consistent estimate of the posterior probability that  $\alpha_i$  occupies rank  $p$  is simply the proportion of times we see  $r_i^{(t)} = p$  over many iterations of the Gibbs sampler,  $t = 1, \dots, T$ .
- Demo with code in `alphaSort.R`

# Posterior Mass Function over ranks



## Other Examples, do in “slow-motion” in R

- multi-level regression, HSB
- Green and Vavreck, “Rock The Vote” cluster-randomized field experiment on voter turnout: hierarchical model for treatment effects in binomial model.
- show superior out-of-sample performance of hierarchical model with linear growth curves; e.g., presidential elections data, rat growth, etc.
- Exercises from Ch 7 in book
- hierarchical models also appear in Ch 8 (e.g., hierarchical model for interviewer effects); Ch 9 (e.g., modeling latent variables as a function of observables).

# **Model Checking and Improvement**

Statistics 220

Spring 2005



# Model Checking

“All models are wrong but some models are useful”

– George E. P. Box

So far we have looked at a number of models and examined them with example data sets. Do the models used accurately describe the data used?

In standard analyses, we will often check model assumptions. For example, in standard regression we will check for

- Correct form of the regression function (e.g. linear vs quadratic)
- Constant variance of the residuals
- Independence of the residuals
- Normality of the residuals

Basic question: How sensitive are our posterior inferences to our modelling assumptions?

Rat Example: Will the following models give significantly different answers about tumor rates in each group

### 1. Original model

- Data model:  $y_i$  = number of tumors in group  $i$

$$y_i | \theta_i \stackrel{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \dots, 71$$

- Process model:  $\theta_i$  = tumor rate in group  $i$

$$\theta_i | \alpha, \beta \stackrel{ind}{\sim} Beta(\alpha, \beta)$$

- Parameter model:

$$p(\alpha, \beta) \propto \frac{1}{(\alpha + \beta)^{5/2}}$$

## 2. Alternative model 1

- Data model:  $y_i = \text{number of tumors in group } i$

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Bin}(n_i, \theta_i) \quad i = 1, \dots, 71$$

- Process model:  $\theta_i = \text{tumor rate in group } i$

$$\text{logit}(\theta_i) | \mu, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mu, \sigma^2)$$

- Parameter model:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

### 3. Alternative model 2

- Data model:  $y_i = \text{number of tumors in group } i$

$$y_i | \alpha_i, \beta_i \stackrel{\text{ind}}{\sim} \text{Beta-bin}(n_i, \alpha_i, \beta_i) \quad i = 1, \dots, 71$$

- Process model:  $(\alpha_i, \beta_i) = \text{tumor rate parameters in group } i$

$$\alpha_i, \beta_i | \gamma_\alpha, \delta_\alpha, \gamma_\beta, \delta_\beta \stackrel{\text{ind}}{\sim} \text{Gamma}(\alpha_i | \gamma_\alpha, \delta_\alpha) \text{Gamma}(\beta_i | \gamma_\beta, \delta_\beta)$$

The tumor rate for group  $i$  is

$$E \left[ \frac{y_i}{n_i} | \alpha_i, \beta_i \right] = \frac{\alpha_i}{\alpha_i + \beta_i}$$

- Parameter model:

$$p(\gamma_\alpha, \delta_\alpha, \gamma_\beta, \delta_\beta) \propto 1$$

Note that we will not be trying to answer the question of whether our model is correct or not. Its not (see Box). We are interested in whether the inaccuracies matter.

Examples you may have seen in the past where deviations from assumptions don't hurt much (at least in big samples):

- $t$ -test of  $H_0 : \mu = \mu_0$  vs  $H_A : \mu \neq \mu_0$

Normality often isn't important, though large skewness can hurt.

- Linear Regression:  $Y = X\beta + \epsilon$

- $\hat{\beta} = (X^T X)^{-1} X^T Y$  is unbiased if  $E[\epsilon] = 0$
- $\hat{\beta}$  is minimum variance unbiased estimator if  $E[\epsilon] = 0$  and constant variance. (Gauss-Markov theorem)

Neither of these results require normality of  $\epsilon$ .

There are cases where assumptions can matter. For example consider the  $F$ -test for examining  $H_0 : \sigma_1^2 = \sigma_2^2$  vs  $H_A : \sigma_1^2 \neq \sigma_2^2$ . The results of this test can be highly dependent on the iid normal assumptions for each group.

One approach to build a super-model that contains all of our models of interest as special cases. This approach usually isn't taken as it is usually difficult to build this super-model and computation is usually infeasible, assuming you can build the model.

Instead we will base these checks on the posterior predictive distribution. Does our data look like our fitted model says it should.

This can either be done by

- *External validation:* future data is compared with the posterior predictive distribution.
- *Internal validation:* observed data is compared with the posterior predictive distribution.

# Posterior Predictive Checking

Idea: If the model fits, replicated data generated under the model should look similar to the observed data.

If we see some discrepancy, is it due to model misspecification or due to chance.

Approach: Generate  $L$  datasets,  $y_1^{rep}, \dots, y_L^{rep}$  from the posterior predictive distribution  $p(y^{rep}|y)$ .  $y^{rep}$  corresponds to replicated data. So if there are any covariates that are conditioned on in the original data.

For example, in the rat tumor example, we need to use the same group sample sizes as in the original data set.

$\tilde{y}$  represents any future outcome whereas  $y^{rep}$  indicates a replication exactly like the observed  $y$ .  $\tilde{y}$  does not need to have the same covariate structure as the original data.

The approach has a similar feel to hypothesis testing, where a test statistic  $T(y, \theta)$  needs to be defined to measure the discrepancy between the data and the predictive simulations.

Note that the test statistic can depend on the data  $y$  and the parameters and hyperparameters  $\theta$ , which is different from standard hypothesis testing where the test statistic only depends on the data, but not the parameters.

## Tail-area probabilities

The lack of fit of the data as compared to the posterior predictive distribution can be compared by a tail-area probability (e.g. *p*-value) of the test statistic  $T(y, \theta)$ . To calculate this probability we will use the replicates sampled from  $p(y^{rep}|y)$ .

- Classical *p*-value

$$p_C = P[T(y^{rep}) \geq T(y)|\theta]$$

where the probability is calculated over the distribution of  $y^{rep}$  given a fixed  $\theta$ . In the classical testing setting  $\theta$  would correspond to the null hypothesis value. It could also be a point estimate (say the MLE).

- Posterior predictive  $p$ -values

To evaluate the fit of a Bayesian model, we need to consider what possible data sets are plausible under the model. When doing this we need to consider not only the observations  $y$ , but also the parameter values  $\theta$ . Thus the  $p$ -value of interest is

$$\begin{aligned} p_B &= P[T(y^{rep}, \theta) \geq T(y, \theta) | y] \\ &= \int \int I(T(y^{rep}, \theta) \geq T(y, \theta)) p(y^{rep} | \theta) p(\theta | y) dy^{rep} d\theta \end{aligned}$$

Usually we can't calculate the Bayesian  $p$ -value exactly, but can do it by simulation. Suppose that we have  $L$  simulations of  $\theta(\theta^1, \dots, \theta^L)$  from the posterior distribution  $p(\theta | y)$ . Then for each of these  $\theta$  samples, generate one sample  $y^{rep^l}$  from  $p(y^{rep} | \theta^l)$ .

We want to compare each of the  $T(y^{rep^l}, \theta^l)$  with  $T(y, \theta^l)$

Then

$$\hat{p}_B = \frac{1}{L} \sum_{l=1}^L I(T(y^{rep^l}, \theta^l) \geq T(y, \theta^l))$$

(i.e. the proportion of samples where  $T(y^{rep^l}, \theta^l) \geq T(y, \theta^l)$ ) is an estimate of  $p_B$ .

Note that the test statistic  $T(y, \theta)$  needs to be chosen to investigate deviations of interest. This is similar to choosing a powerful test statistic when conducting a hypothesis test

For example, in the analysis of Newcomb's speed of light experiment discussed in the text, a worry was the effect of outliers. Thus  $T(y, \theta)$  needs to be chosen to focus on this issue.

In the book  $T(y, \theta) = \min y_i$  was used (they were worried about low outliers). Another possibility would be  $T(y, \theta) = \max |y_i - \mu|$  (e.g. the biggest residual in magnitude). This would be appropriate if the worry was either big positive or big negative residuals. This might occur if

$$y_i | \mu, \sigma^2 \sim t_\nu(\mu, \sigma^2)$$

instead of

$$y_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

as used in the analysis.

While the approach is a bit more focused on test statistics, this has a similar feel to residual analysis in regression.

- Is there any pattern in the residual plot ( $e_i$  vs  $\hat{y}_i$ )
- Plotting  $e_i$  vs  $e_{i-1}$  or the Durbin-Watson test to examine whether residuals are correlated over time
- Normal scores plot or Anderson-Darling test for normality of residuals

For example, if we see some curvature (but constant variance) in the residual plot, it suggests we might be missing a  $x^2$  term in the model.

If there is some curvature and non-constant variance in the residual plot, maybe we need to transform  $y$ .

Examples:

For the two random effects model examples (detergent filling and sodium level in beer), two concerns might be

1. Conditional normality of the observations (e.g.  $y_{ij} \sim N(\theta_j, \sigma^2)$ )
2. Constant variance of observations within each group

Note that these are probably of limited concern in both of these examples, as the total sample size is fairly large and there are equal numbers of observations in each group for both data sets.

Possible test statistics to evaluate these are

1. Normality: Let  $e_{ij} = y_{ij} - \theta_j$  and  $e_{(1)} \leq e_{(2)} \leq \dots \leq e_{(n)}$  be the ordered residuals. Let

$$T(y, \theta) = \text{Corr} \left( e_{(i)}, \Phi^{-1} \left( \frac{i}{n+1} \right) \right)$$

(e.g. Correlation of points in a normal scores plot). If the data is conditionally normal, this correlation should be close to one. Otherwise the normal scores plot will have some non-linearity, which will pull this correlation down from one.

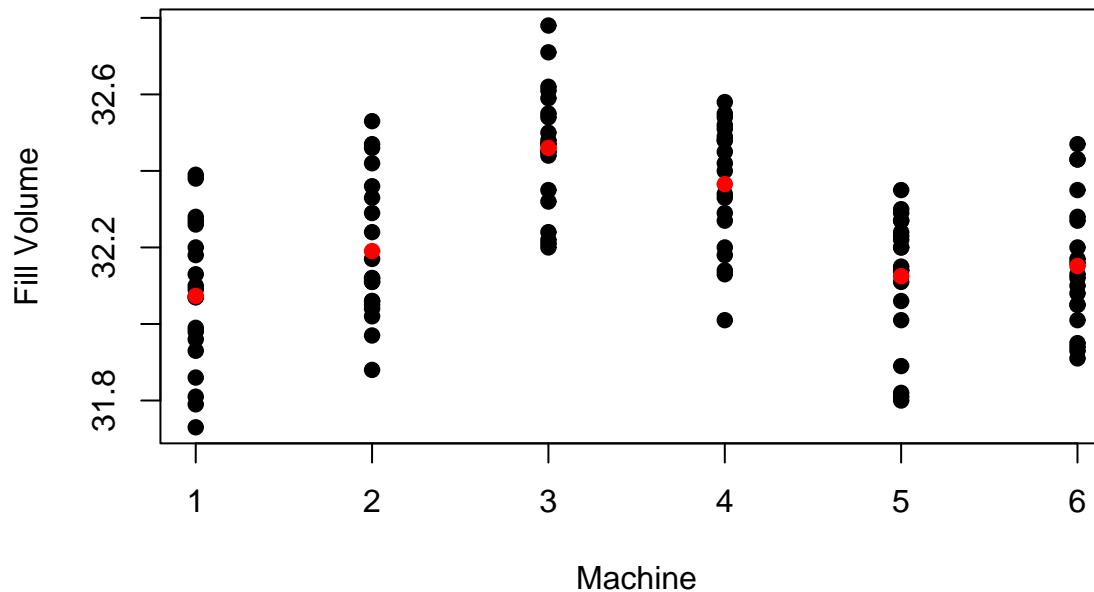
(This test statistic has the feel of the Shapiro-Wilks normality test.)

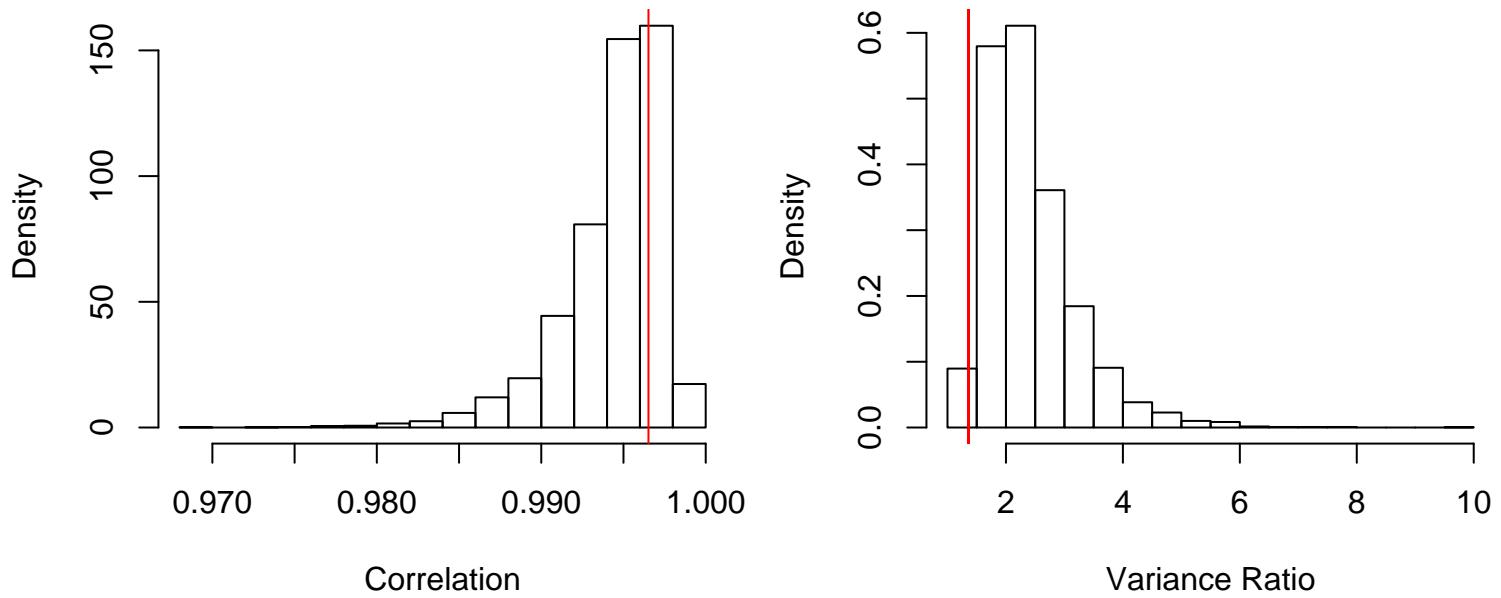
2. Equal variance: Let  $s_i^2$  be the sample variance of the observation in group  $i$ . If the constant variance assumption is reasonable

$$T(y, \theta) = \frac{\max s_i^2}{\min s_i^2}$$

should not be much bigger than one.

## Detergent example:

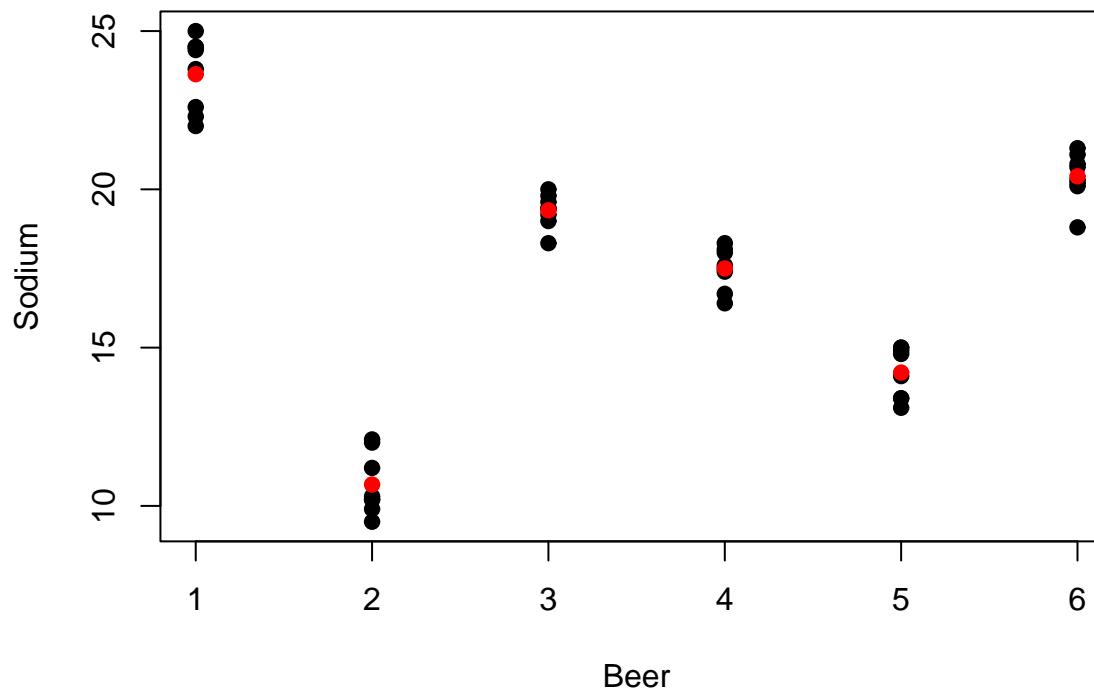


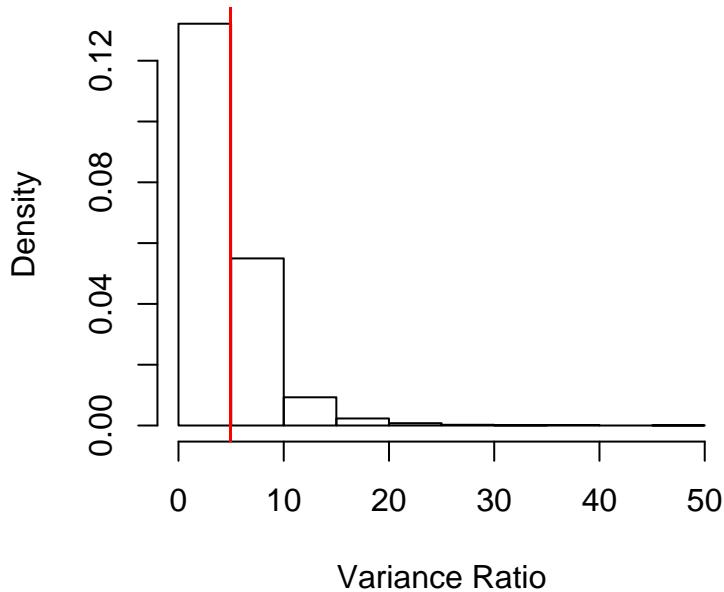
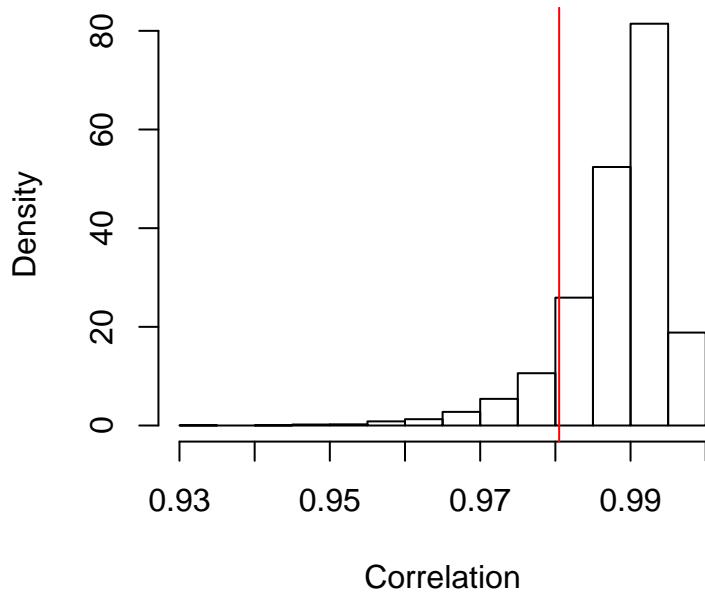


Normality test:  $\hat{p}_B = 0.4886$

Equal variance test:  $\hat{p}_B = 0.9866$

Beer example:





Normality test:  $\hat{p}_B = 0.5270$

Equal variance test:  $\hat{p}_B = 0.3520$

# **Model Checking and Improvement II**

Statistics 220

Spring 2005



# Graphical Checks

As in standard frequentist analyses, graphical summaries are also useful to examine the fit of a model. There are three common types of plots

1. Displaying data
2. Displaying data summaries or parameter inferences
3. Graphs of residuals or other discrepancy measures

## Choosing $T(y, \theta)$

As mentioned last time,  $T(y, \theta)$  should be chosen to examine possible deviations of interest and examining more than one at a time is reasonable.

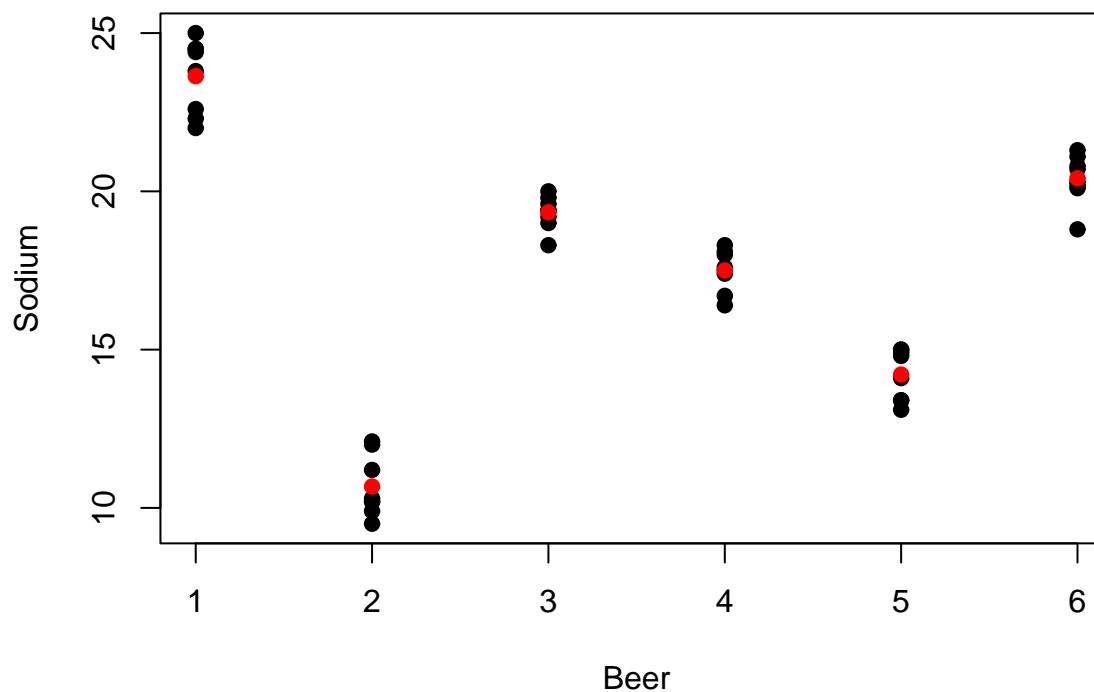
For example, in the examples last time, two statistics were studied, one investigating the normality of deviations from the means in a one-way ANOVA and the other investigating homoscedacity of the deviations.

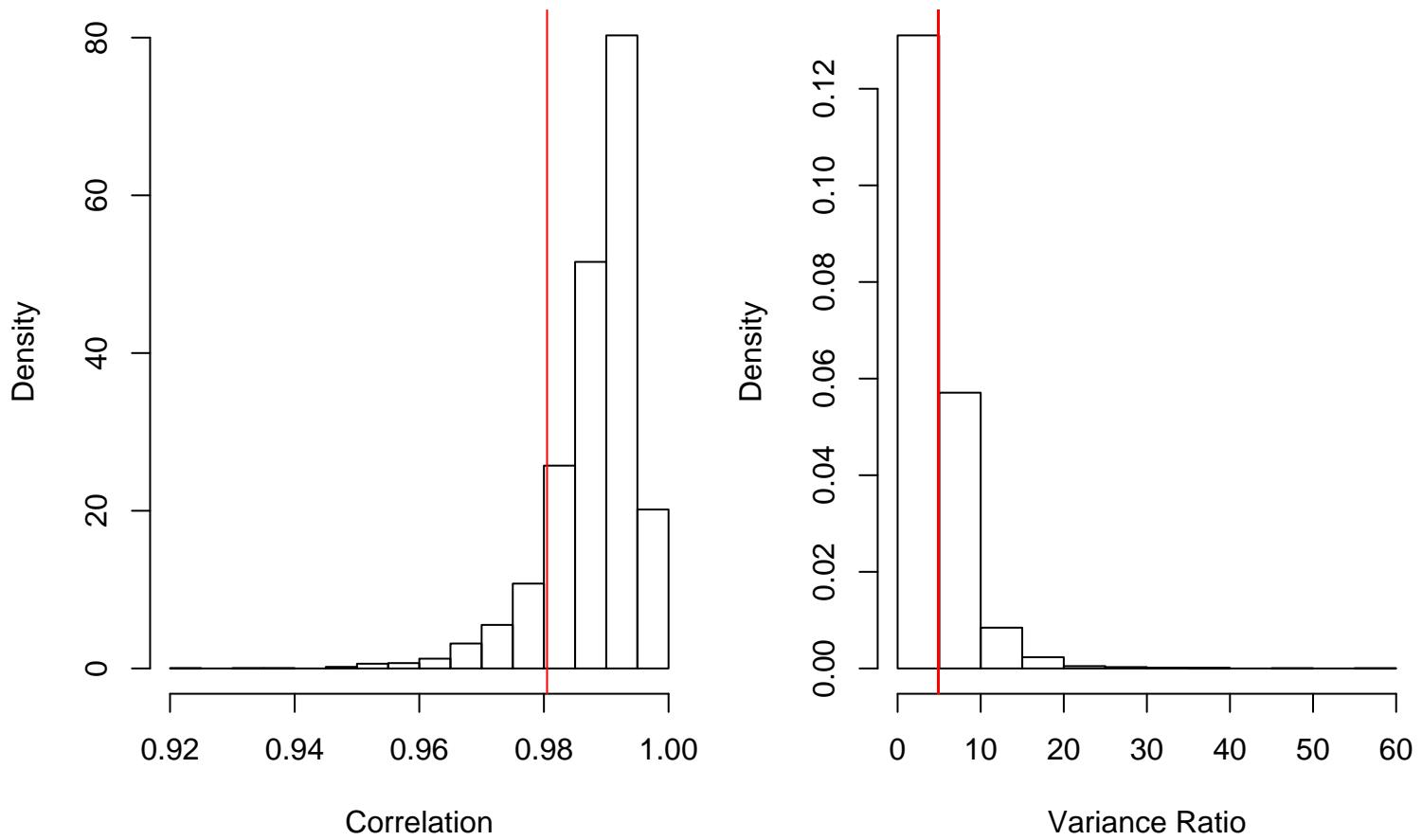
For each  $T(y, \theta)$ , an estimate of the Bayesian  $p$ -value

$$\hat{p}_B = \frac{1}{L} \sum_{l=1}^L I(T(y^{rep^l}, \theta^l) \geq T(y, \theta^l))$$

was calculated.

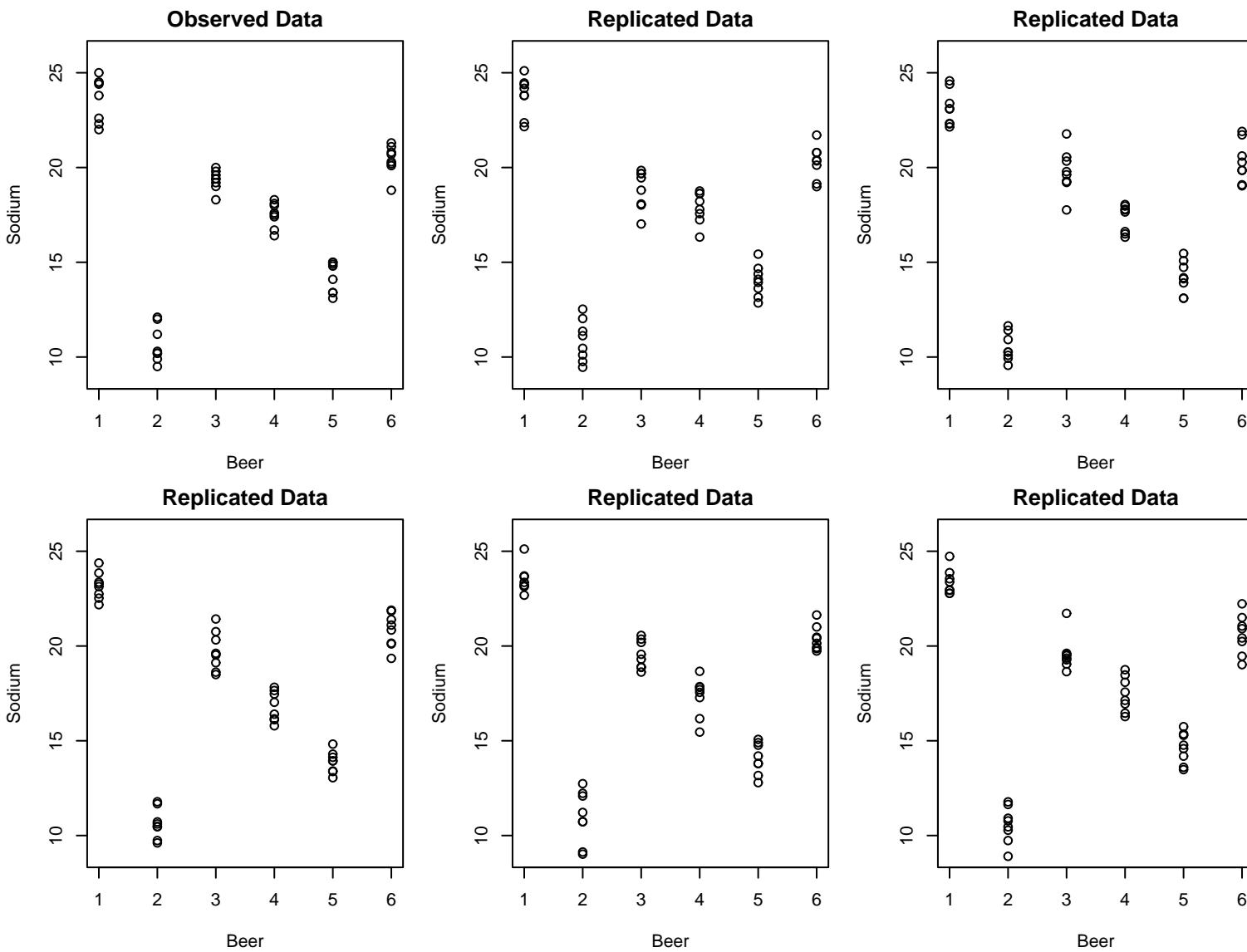
Beer example:





Normality test:  $\hat{p}_B = 0.5420$

Equal variance test:  $\hat{p}_B = 0.3408$



Note that the Bayesian  $p$ -values only tell part of the story.

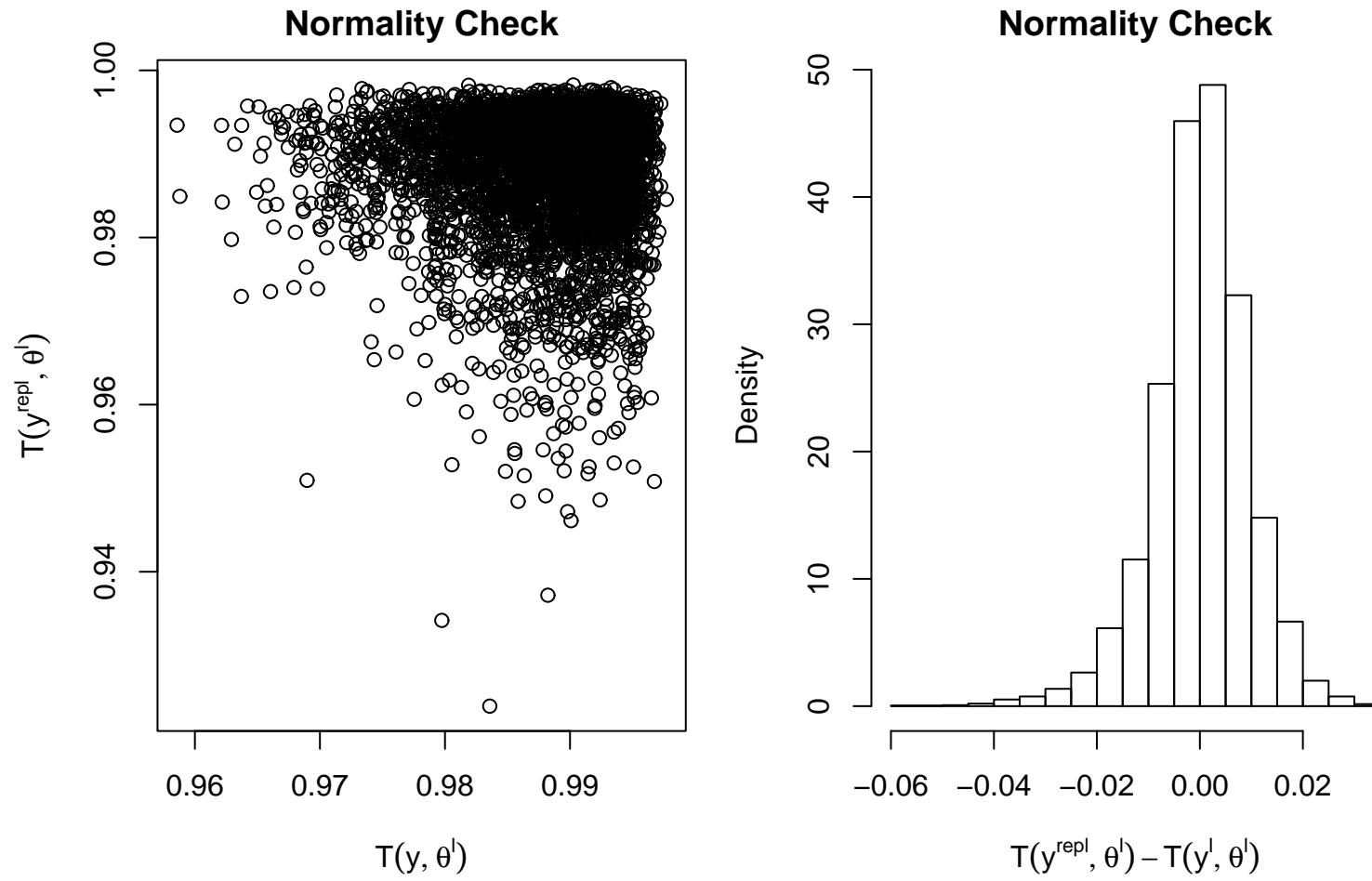
It is also useful to look at the relationship between  $T(y^{rep^l}, \theta^l)$  and  $T(y, \theta^l)$ .

This could be done as in the previous graphs or by

- plotting  $T(y^{rep^l}, \theta^l)$  versus  $T(y, \theta^l)$
- a histogram of  $T(y^{rep^l}, \theta^l) - T(y, \theta^l)$
- a histogram of  $\frac{T(y^{rep^l}, \theta^l)}{T(y, \theta^l)}$  or of  $\log T(y^{rep^l}, \theta^l) - \log T(y, \theta^l)$

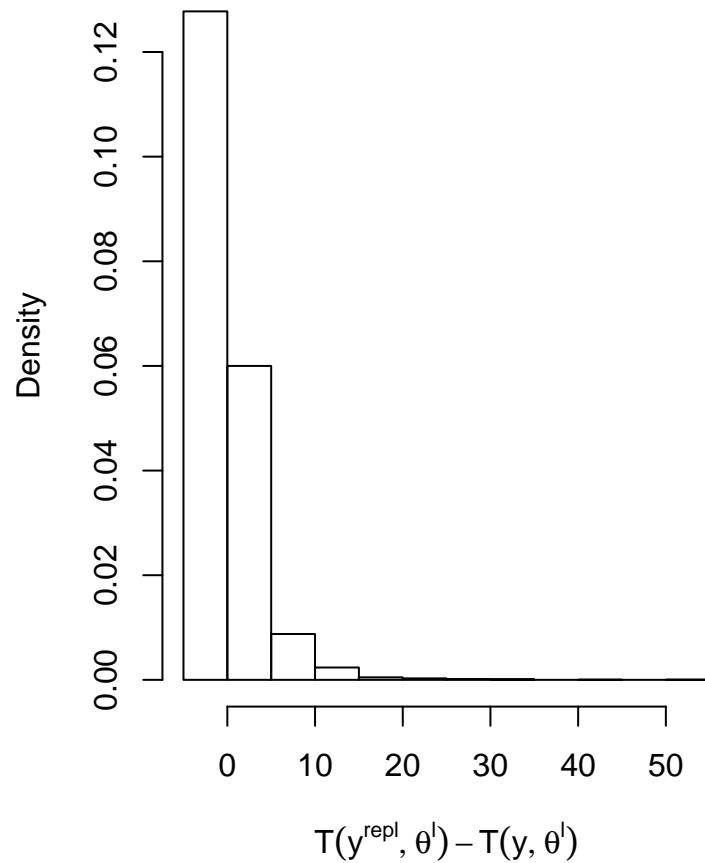
These are preferable to what I had last time for the normality check plot as the summary of the data fit on the plot (the red line) ignored the uncertainty in  $\theta$ .

(The red line was the correlation in the normal scores plot using the residuals from a standard one-way ANOVA.)

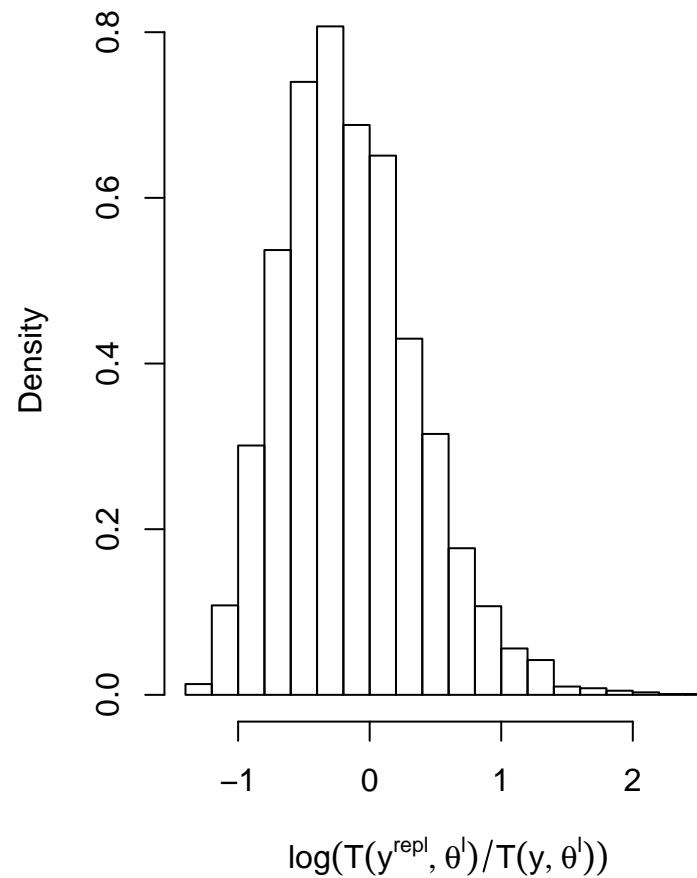


The histogram being centered at approximately 0 suggests that the fit of the observed data is roughly in the middle of what would be expected based on the posterior predictive distribution.

**Variance Check**



**Variance Check**



## Multiple Comparisons:

In many situations with multiple test statistics you want to adjust for multiple looks at the data. For example, a Bonferroni correction (involving  $k$   $p$ -values) suggests that the  $p$ -values should be compared with  $\frac{\alpha}{k}$  instead of  $\alpha$ .

This is not recommending in this setting. There is no worry about “Type I error” rates here. We are not using the  $p$ -values to accept or reject a model but as summaries to investigate limits of the model in realistic replications.

Aside: While strictly not doing hypothesis tests, if the values of  $T(y^{rep^l} | \theta^l)$  indicate how the model can be improved, the model should be abandoned in favour of a better one. So in one sense, you are sort of acting like rejecting one model in favour of another one.

## Omnibus tests:

In addition to focused test statistics, there are which more general measures of fit. The most common one is the  $\chi^2$  discrepancy

$$T(y, \theta) = \sum_i \frac{(y_i - E[y_i|\theta_i])^2}{\text{Var}(y_i|\theta_i)}$$

If  $\theta$  is known, this is similar to the classical  $\chi^2$  goodness of fit statistic.

An alternative to this is  $T(y, \theta) = -2 \log p(y|\theta)$ , the deviance.

In the the classical setting,  $\theta$  must be specified. This might be by

- $\theta = \theta_{null}$
- $\theta = \theta_{mle}$
- $\theta = \arg \min_{\theta} T(y, \theta)$

In the Bayesian approach, we average over  $\theta$  and the sampling distribution is automatically calculated by the posterior predictive simulations.

## Example: Rat Tumors

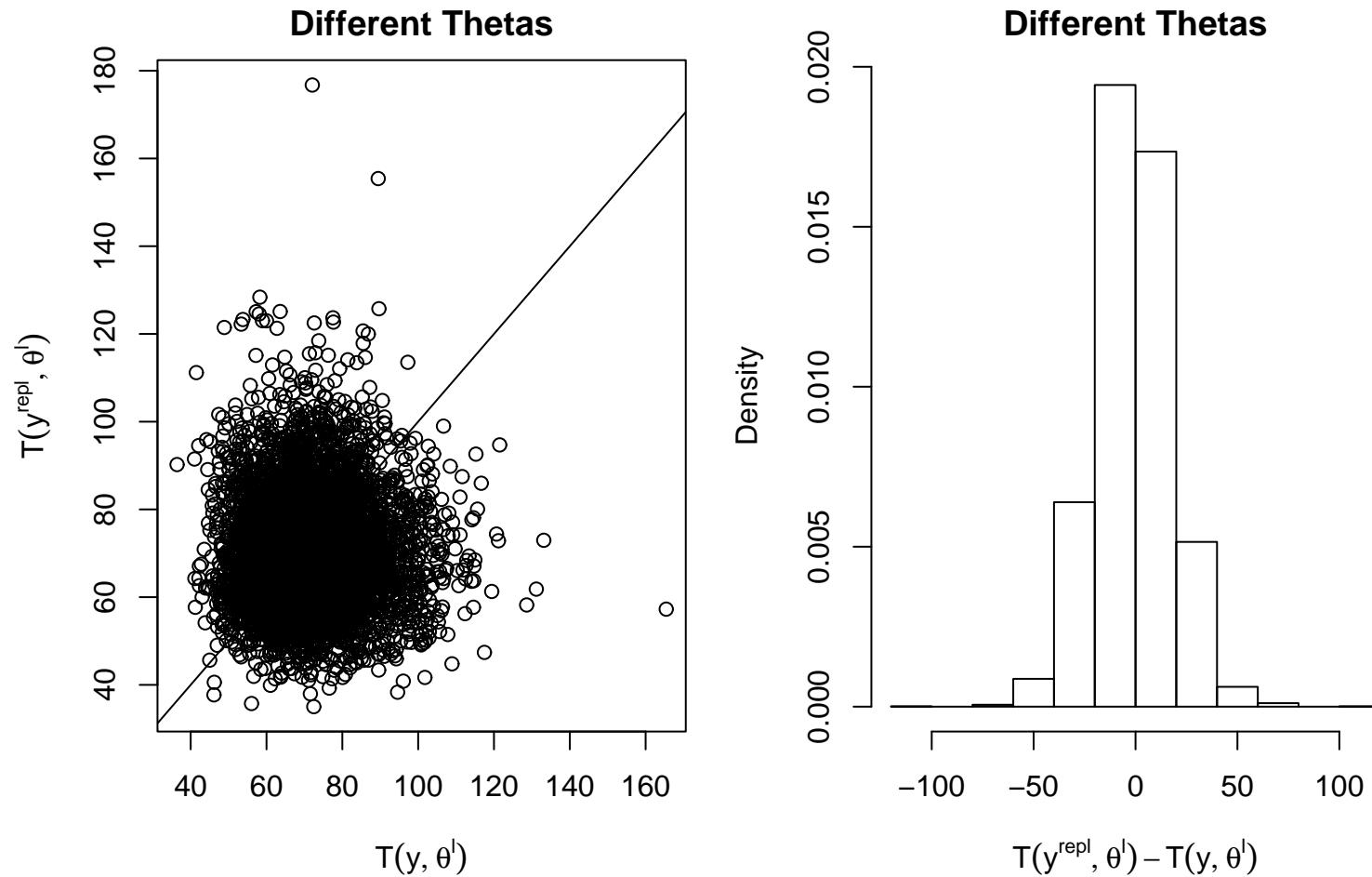
Compare two models:

### 1. Variable tumor rates

$$\begin{aligned} y_i | \theta_i &\stackrel{\text{iid}}{\sim} \text{Bin}(n_i, \theta_i) \\ \theta_i &\stackrel{\text{iid}}{\sim} \text{Beta}(\alpha, \beta) \\ p(\alpha, \beta) &\propto \frac{1}{(\alpha + \beta)^{5/2}} \end{aligned}$$

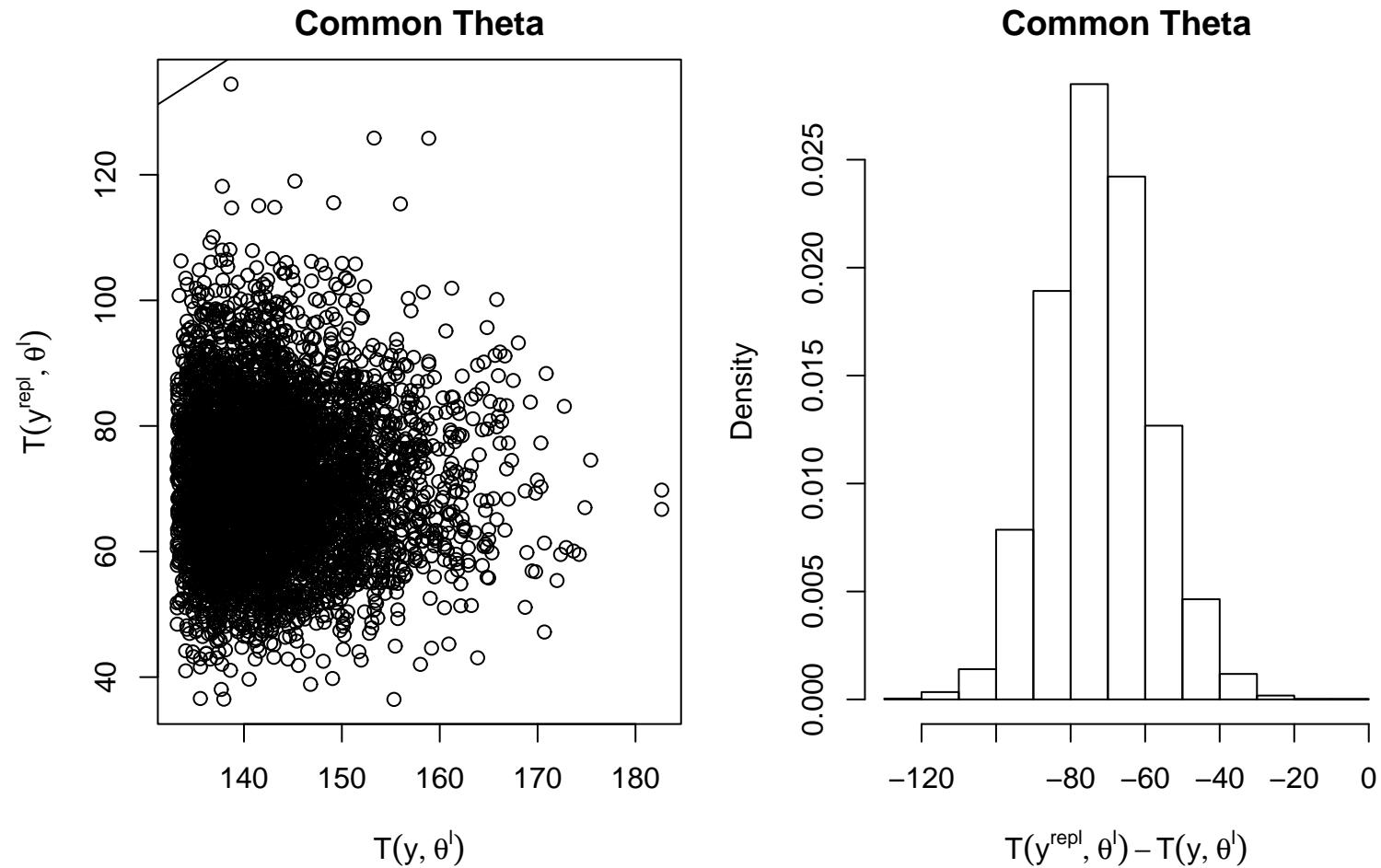
### 2. Common tumor rates

$$\begin{aligned} y_i | \theta &\stackrel{\text{iid}}{\sim} \text{Bin}(n_i, \theta) \\ \theta &\sim \text{Beta}(\alpha, \beta) \\ p(\alpha, \beta) &\propto \frac{1}{(\alpha + \beta)^{5/2}} \end{aligned}$$



$$\hat{p}_B = 0.4752$$

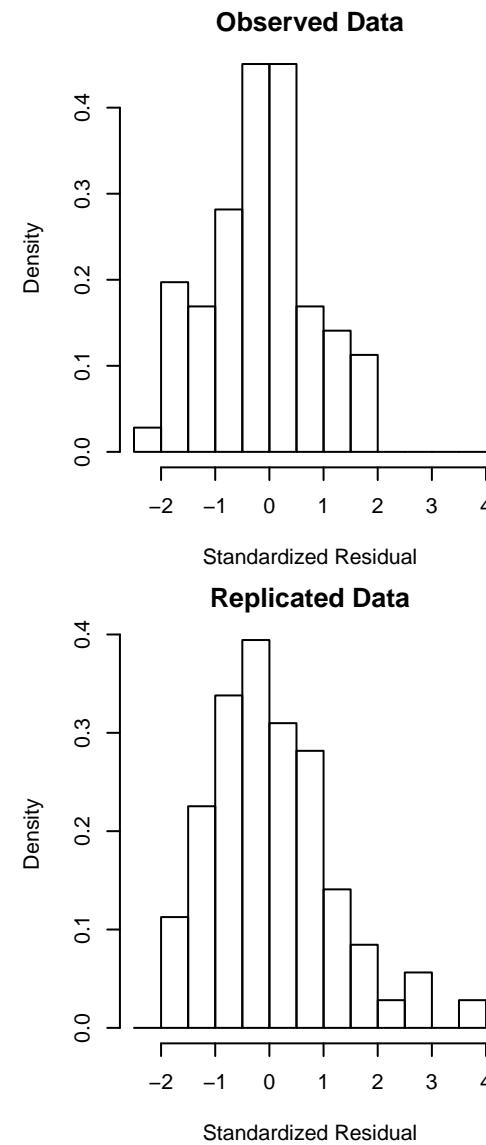
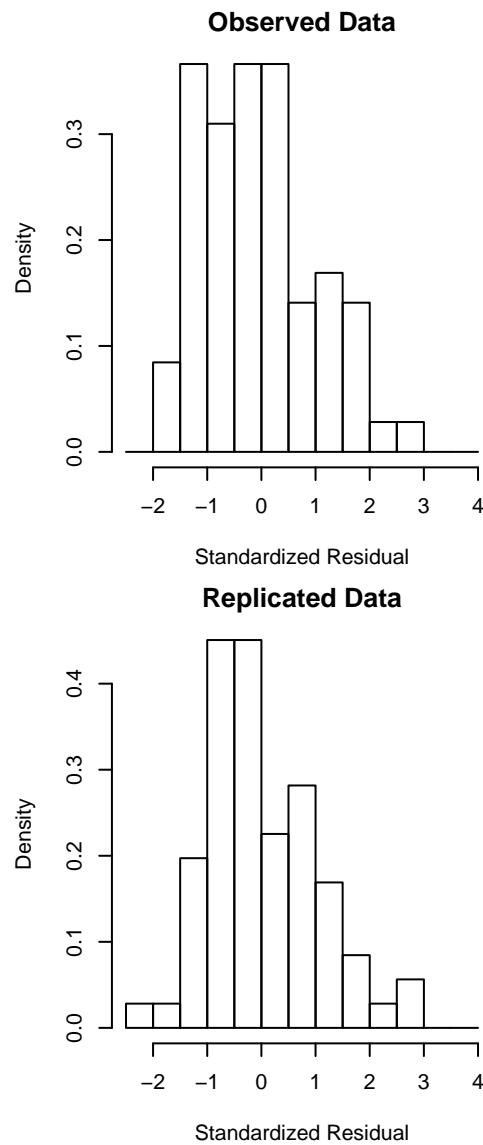
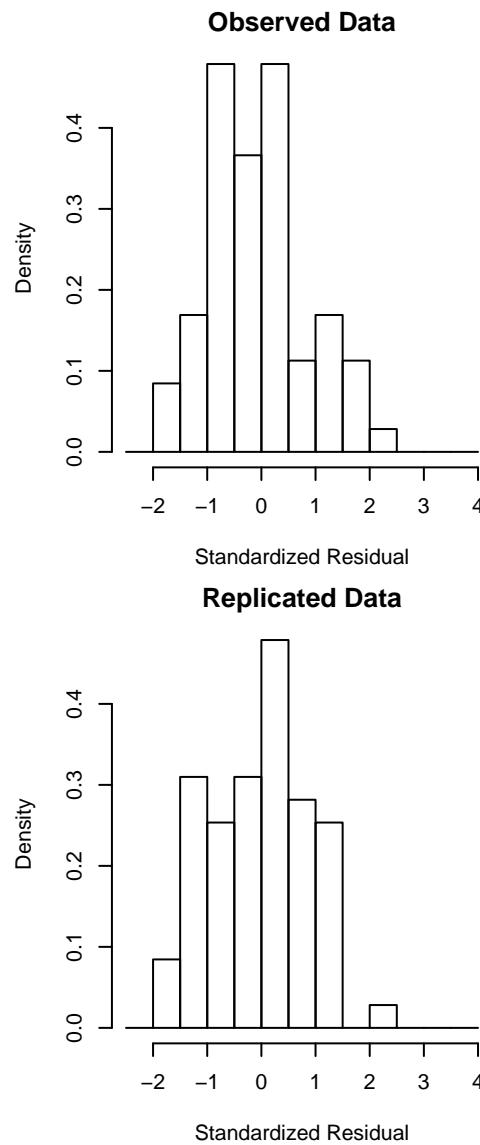
$$\hat{E}[y^{\text{rep}^l}, \theta^l] = 71.28 \quad \hat{E}[y, \theta^l] = 72.09$$



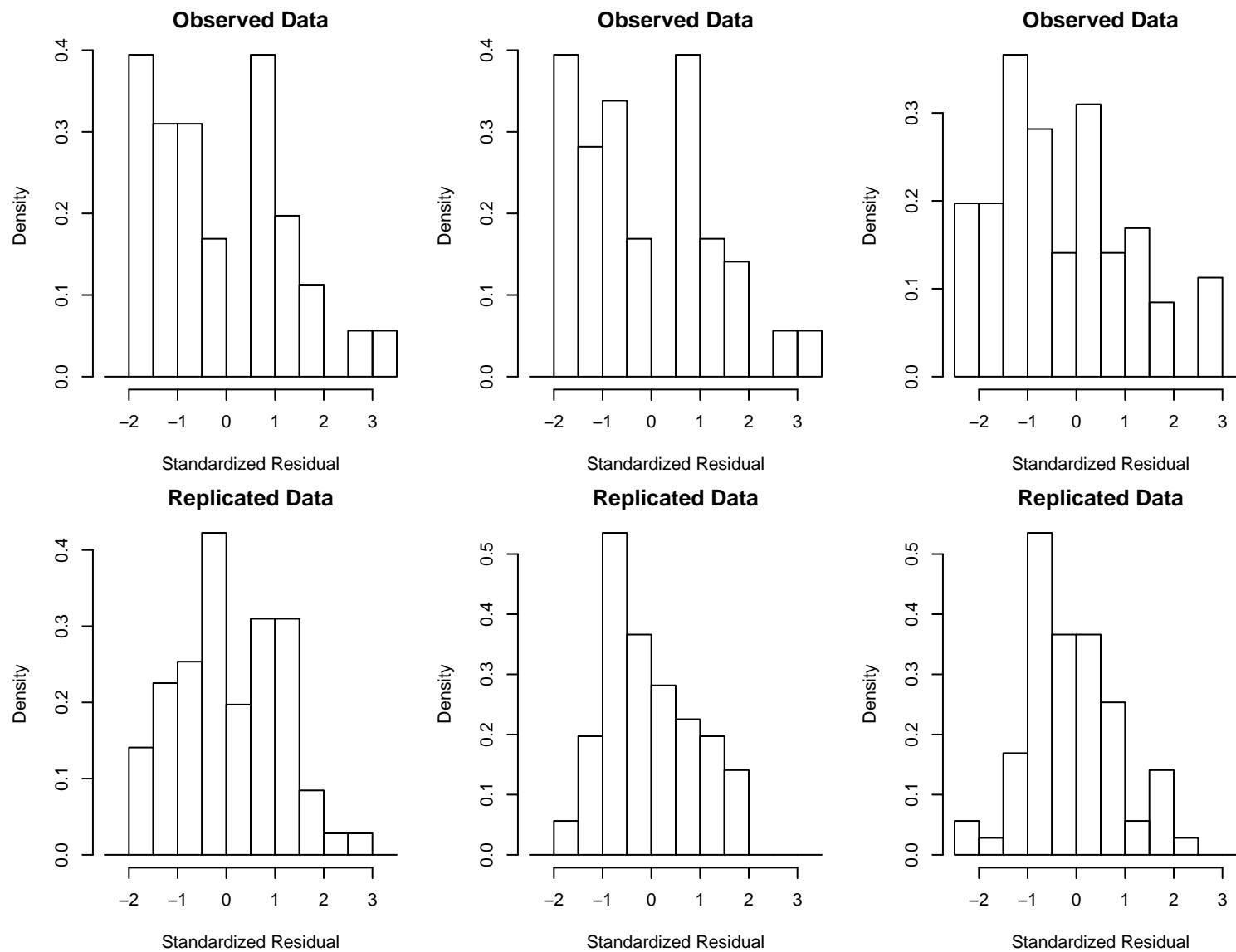
$$\hat{p}_B = 0$$

$$\hat{E}[y^{\text{rep}^l}, \theta^l] = 70.89 \quad \hat{E}[y, \theta^l] = 143.09$$

## Different Thetas



# Common Theta



## Interpreting Posterior Predictive $p$ -values

An extreme  $p$ -value for a test statistic  $T(y, \theta)$  (e.g. near 0 or 1) indicates that the observed pattern in the data would be unlikely if the model were true.

While it is a probability, it is not  $P[\text{model is true} \mid \text{data}]$ . As we have seen before, it is

$$P[T(y^{rep}, \theta) \geq T(y, \theta) \mid y]$$

a statement about probabilities of data sets, not models.

If a  $p$ -value is extreme, it usually doesn't matter how extreme. For example a  $p$ -value of 0.00001 is effectively no stronger than a  $p$ -value of 0.001.

As with normal  $p$ -values, these measure "statistical significance" not "practical significance". Small changes to the model can make large changes in the  $p$ -value.

# Model Expansion

## Adding Parameters to a Model

While there are many ways of coming up with new models when the data doesn't seem to fit, adding parameters to a model is a common approach.

1. To deal with lack of fit or missing prior knowledge about the data, process, or parameters.
2. To get around questionable modelling assumption or ones with no justification.
3. If two (or more) possible models are under consideration, it may be possible to consider them as special cases or a more general model.

For example, for the two normal based models

$$y_{ij} | \theta_j \stackrel{ind}{\sim} N(\theta_j, \sigma^2)$$

$$\theta_j \stackrel{iid}{\sim} N(\mu, 10)$$

$$p(\mu) \propto 1$$

and

$$y_{ij} | \mu \stackrel{ind}{\sim} N(\mu, \sigma^2)$$

$$p(\mu) \propto 1$$

are special cases of the model

$$\begin{aligned}y_{ij} | \theta_j &\stackrel{\text{ind}}{\sim} N(\theta_j, \sigma^2) \\ \theta_j &\stackrel{\text{iid}}{\sim} N(\mu, \tau^2) \\ p(\mu) &\propto 1 \\ \tau &\sim p(\tau)\end{aligned}$$

The first is  $\tau^2 = 10$  and the second is  $\tau^2 = 0$ .

4. A model can be expanded to include additional data. One way of handling this is fitting the current piece into a hierarchical model.

Or the model can be modified to handle different data structures. For example, allowing for missing data.

In these cases, the current model  $p(y, \theta)$  is embedded into the model  $p(y, \theta, \phi)$  or  $p(y, y^*, \theta, \phi)$

This approach also gives a way to examine sensitivity to the choice prior. We could examine  $p(\theta|y)$  directly under different fixed priors on  $\theta$  (e.g. for different values of  $\phi$  plugged into  $p(\theta|\phi)$ ). Or we could put a prior on  $\phi$  and look at

1. conditional posterior  $p(\theta|y, \phi)$  for different choices of  $\phi$
2. marginal posterior  $p(\theta|y)$

$$p(\theta|y) \propto \int p(\theta|y, \phi)p(\phi|y)d\phi$$

3. marginal posterior  $p(\phi|y)$

# **Model Checking and Improvement III**

Statistics 220

Spring 2005



# Model Comparison

There are two situations where comparing models may be reasonable

1. Nested models:

Example:

Model 1

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Model 2

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

The first model is a special case of the second model ( $\beta_2 = 0$ ).

In most cases the larger model will fit the data better but can be more difficult to interpret and compute. Two questions of interest in this comparison are

- (a) Is the improvement in fit large enough to justify the additional difficulty in interpretation and computation?
- (b) Is the prior distribution on the additional parameters reasonable?

Note: This second question is why I noted that the larger model will usually fit better, instead of always. A bad prior may bias the fits for small sample sizes. In a likelihood based analysis, the larger model must always give a better fit (assuming deviance is the measure of fit).

Standard hypothesis testing methods address the first question in frequentist analyses. For example, the use of an  $F$ -test to compare two linear regression models.

The approach we will take here is one that measures the distance of the data to each of the models.

- $\theta$ : parameters in first model
- $\phi$ : additional parameters in second model

So we want to compare  $p(\theta|y)$  and  $p(y^{rep}|y)$  with  $p(\theta, \phi|y)$  and  $p(y^{rep}|y)$

## 2. Non-nested models:

Example:

Model 1

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Model 2

$$y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i$$

In this case, you can't make one model a special case of the following.

Comparisons of this sort can be useful, if for example, you wish to compare two regression with different sets predictors to see which one does better. Maybe for cost considerations you can only afford one predictor, so you need to figure out which is the best.

# Expected Deviance

Last class we discussed the use of

$$T(y, \theta) = \sum_i \frac{(y_i - E[y_i|\theta_i])^2}{\text{Var}(y_i|\theta_i)}$$

as measure of model fit.

Another option, which tends to work better for our purposes is the deviance

$$D(y, \theta) = -2 \log p(y|\theta)$$

This has ties to the Kullback-Leibler information

$$\begin{aligned} H(\theta) &= \int \log \left( \frac{f(y)}{p(y|\theta)} \right) f(y) dy \\ &= \int \log f(y) f(y) dy - \int \log p(y|\theta) f(y) dy \end{aligned}$$

So

$$2H(\theta) = \int D(y, \theta) f(y) dy + 2 \int \log f(y) f(y) dy$$

which implies

$$E[D(y, \theta)] = 2H(\theta) - 2 \int \log f(y) f(y) dy$$

Thus the expected deviance (averaged over the true sampling distribution  $f(y)$ ) is twice  $H(\theta)$  minus a factor that doesn't depend on  $\theta$ .

As we have discussed before, as the sample size goes to infinity, our posterior inference goes to the model with the smallest value of  $H(\theta)$ . So this suggests using an estimate of the expected deviance as a measure of overall model fit.

In using the deviance to measure model fit there are two approaches.

1. Plug in estimate of  $\theta$

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

where  $\hat{\theta}(y)$  is an estimate of  $\theta$  based on the data  $y$ , say, for example, the posterior mean of  $\theta$ .

2. Average over the posterior realizations of  $\theta$

$$D_{avg}(y) = E[D(y, \theta)|y]$$

which can be estimated using posterior simulations  $\theta^1, \dots, \theta^L$  by

$$\hat{D}_{avg}(y) = \frac{1}{L} \sum_{l=1}^L D(y, \theta^l)$$

$\hat{D}_{avg}(y)$  is a better measure of model error since it averages over our uncertainty about  $\theta$ .

$D_{\hat{\theta}}(y)$  tends to indicate a better fit than we really have as it calculates the discrepancy under a more probable  $\theta$

# Counting Parameters and Model Complexity

While  $D_{\hat{\theta}}(y)$  is not a good measure of model fit, it is an interesting descriptor as

$$p_D^{(1)} = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

is a measure of the effective number of parameters in the Bayesian Model.

An alternative measure is

$$p_D^{(2)} = \frac{1}{2} \widehat{\text{Var}}(D(y, \theta) | y) = \frac{1}{2L-1} \sum_{l=1}^L (D(y, \theta^l) - \hat{D}_{avg}(y))^2$$

Both of these measures are based on the distribution of  $D(y, \theta)$ , relative to its minimum, having an asymptotic  $\chi^2$  distribution. (Note  $E[\chi_\nu^2] = \nu$  and  $\text{Var}(\chi_\nu^2) = 2\nu$ .)

The second of these ( $p_D^{(2)}$ ) is the preferred as the asymptotics should work a bit better (less worry about bias).

The way to think of  $p_D$  is as the number of 'unconstrained' parameters in the model. A parameter will count as

- 1 if it is completely unconstrained (no information about it in the prior)
- 0 if it is completely constrained (all information about it is in the prior)
- intermediate if information about it comes from the data and the prior

## Deviance Information Criterion

A useful measure for model selection is based on

$$D_{avg}^{pred}(y) = E[D(y^{rep}, \hat{\theta}(y))]$$

where  $D(y^{rep}, \hat{\theta}(y)) = -2 \log p(y^{rep}|\theta)$

This can be estimated by the *Deviance Information Criterion* (DIC)

$$DIC = \hat{D}_{avg}^{pred}(y) = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

This can be thought of as

$$DIC = \hat{D}_{avg}(y) + p_D^{(1)}$$

the average deviance plus a penalty term, giving it the same flavour as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) used in likelihood analysis

$$AIC = -2 \log p(y|\hat{\theta}) + 2K$$

$$BIC = -2 \log p(y|\hat{\theta}) + K \log n$$

where  $K$  is the number of parameters in the model

Note that DIC is different than many of the other measures discussed in that it uses an estimated value for  $\theta$  instead of averaging the posterior distribution of  $\theta|y$ .

The goal of DIC is use it to predict a model with the best out-of-sample predictive power. The book motivates DIC by starting with

$$D_{avg}^{pred}(y) = E \left[ \frac{1}{n} \sum_{i=1}^n (y_i^{rep} - E[y_i^{rep}|y])^2 \right]$$

(This is for normal likelihoods, though the book doesn't mention it.)

Example: Rat Tumor Rates

Model	$D_{\hat{\theta}}$	$\hat{D}_{avg}$	$p_D^{(2)}$	DIC
Shrinkage	167.9	253.0	85.1	338.1
Common Theta	343.8	344.8	1	345.8

(Calculated by WinBugs)

So this suggests that the Shrinkage model is a preferable model to the Common Theta model (as we have already seen).

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff	
theta	0.2	0.0	0.1	0.1	0.2	0.2	0.2	1	5000	
a	139.7	5659.1	0.0	0.0	0.3	1.1	58.2	1	3400	
b	733.1	29395.6	0.0	0.1	0.4	3.0	328.5	1	3100	
u	0.4	0.2	0.1	0.2	0.3	0.5	0.9	1	5000	
v	5.2	28.6	0.1	0.5	1.1	2.8	28.1	1	3100	
deviance	344.8		1.4	343.8	343.9	344.3	345.1	348.8	1	5000
pD = 1 and DIC = 345.8 (using the rule, pD = var(deviance)/2)										

## Comment on WinBugs output and Table 6.2 in the text

In WinBugs, as noted in the output, the reported  $p_D = p_D^{(2)}$  and  $DIC = \hat{D}_{avg} + p_D^{(2)}$ , not  $DIC = \hat{D}_{avg} + p_D^{(1)}$  as you get if you follow the development of the math in the text.

In the table on the previous page,  $D_{\hat{\theta}} = \hat{D}_{avg} - p_D^{(2)}$  which doesn't quite match the books development, which has  $D_{\hat{\theta}} = \hat{D}_{avg} - p_D^{(1)}$ .

I suspect that Table 6.2 in the text, which is comparison of 3 Normal random effects models (SAT coaching example) was calculated by WinBugs as well as the same relationships between  $D_{\hat{\theta}}$ ,  $\hat{D}_{avg}$ ,  $p_D^{(2)}$ , and DIC hold.

Note that these deviations from the books development shouldn't be very important as sample sizes increase  $p_D^{(1)}$  and  $p_D^{(2)}$  should approach each other.

### **Comment of $p_D$**

As mentioned earlier,  $p_D$  can be thought of as a measure of the effective number of unconstrained parameters in a model. Note that estimates of this quantity don't have to be less than the number of actual parameters in the model.

For example, for the rat tumor example with the shrinkage model,  $p_D^{(2)} = 85.1$ . However the actually number of parameters in the model is 73 (71  $\theta$ s,  $\alpha$ , and  $\beta$ ). I suspect (though I'm not sure) that this difference is due to randomness in the data.

An analogue would trying to estimate the degrees of freedom in a  $\chi^2$  test based on the observed test statistics and an assumption that the data was really generated under the null hypothesis.

## Bayes Factors

As discussed earlier, Bayes theorem can be thought of in terms of posterior odds, which satisfies

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Likelihood Ratio}$$

Thus the posterior odds of two models  $H_1$  and  $H_2$  is given by

$$\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \times \frac{p(y|H_2)}{p(y|H_1)}$$

where

$$p(y|H_i) = \int p(\theta_i|H_i)p(y|\theta_i, H_i)d\theta_i$$

The ratio of the marginal likelihoods

$$BF(H_2; H_1) = \frac{p(y|H_2)}{p(y|H_1)}$$

is known as the Bayes Factor. Its a measure of how much the data favours model  $H_2$  over  $H_1$

Note that the Bayes factor is only defined when the marginal likelihood of each model is proper.

Consider the model

$$y|\theta \sim N(\theta, 1)$$

$$p(\theta) \propto 1$$

Then

$$p(y) \propto \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \theta)^2\right) d\theta = 1$$

Thus marginally  $y$  is uniform on  $(-\infty, \infty)$  and thus has an improper distribution.

(Aside: even though the marginal density on  $y$  is improper, the posterior  $p(\theta|y)$  is proper as we have seen. It happens that things cancel properly when going to the posterior.)

Note that the Bayes factor only really make sense when there are a discrete number of models being investigated.

A good example of where Bayes factor have been used is in genetic counselling, such as was done for diseases like Huntington's (an autosomal dominant trait with the gene on chromosome 4) in the late 80's and early 90's.

$H_1$  = subject is not at risk for disease (not a gene carrier)

$H_2$  = subject is at risk (gene carrier)

$y$  = marker data on subject and relatives and disease status in relatives (subject is currently unaffected).

Of interest is  $P[H_2|y]$  or equivalently  $\frac{p(H_2|y)}{p(H_1|y)}$ .

As testing was usually done in people with one affected parent, lets assume that  $\frac{p(H_2)}{p(H_1)} = 1$ . It will be different if other affection pattern in the family led to the test. Both parents being affected would raise the prior odds to 3.

The likelihoods under the two situations was calculated using a peeling algorithm.

With these prior odds, the posterior odds is actually the Bayes factor. I believe with the standard test of the time, the largest the Bayes factor could be is about 33 and the smallest is about  $\frac{1}{33}$  (the marker used was about 3cM from the gene I think).

Note current technology eliminates the needs for these sorts of calculations for many diseases. Though not completely. Even though the Huntington gene (IT-15) has been found, sequenced, and is somewhat understood, the recent test for this condition isn't perfect. Thus this sort of analysis (though) with different sorts of calculations due to the different types of data.

While Bayes factors can be calculated for models described by continuous parameter, then tend not to be useful.

For example consider the set of normal based models  $H_\tau$

$$\begin{aligned}
 y_{ij} | \theta &\stackrel{ind}{\sim} N(\theta_j, \sigma^2) \\
 \theta_j &\stackrel{iid}{\sim} N(\mu, \tau^2) \\
 \mu &\sim p(\mu) \quad (\text{Proper})
 \end{aligned}$$

While the Bayes factor is well defined in this case

$$BF(H_\tau; H_0) = \frac{p(y|\tau)}{p(y|\tau = 0)}$$

it is not particularly helpful for picking which  $H_\tau$  is the best model.

Also in these situations, the Bayes factor can break down and become unstable, particularly when  $p(\mu)$  is pushed to the limit to approximate an improper prior.

# BAYESIAN ANALYSIS OF CHOICE DATA

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 3, 2012

# Discrete Choice

- binary (e.g., probit model; we looked at with data augmentation)

# Discrete Choice

- binary (e.g., probit model; we looked at with data augmentation)
- ordinal (ordinal logit or probit)

# Discrete Choice

- binary (e.g., probit model; we looked at with data augmentation)
- ordinal (ordinal logit or probit)
- multinomial models for unordered choices: e.g., multinomial logit (MNL), multinomial probit (MNP).

# Discrete Choice

- binary (e.g., probit model; we looked at with data augmentation)
- ordinal (ordinal logit or probit)
- multinomial models for unordered choices: e.g., multinomial logit (MNL), multinomial probit (MNP). We won't consider models for “tree-like” choice structures (nested logit, GEV, etc).

# Binary Choices: logit or probit

- for “standard” models (e.g., no “fancy” hierarchical structure, no concerns re missing data etc), other avenues besides BUGS/JAGS
- e.g., MCMCpack
- implementations in BUGS/JAGS: don’t use data augmentation *a la* Albert & Chib (1991).
- dbern or dbin and sample from the conditional distributions using Metropolis-within-Gibbs, slice sampling

# Binary Choices: logit or probit

## Voter turnout example.

JAGS code

```
model{  
  for (i in 1:N){          ## loop over observations  
    y[i] ~ dbern(p[i])      ## binary outcome  
    logit(p[i]) <- ystar[i]  ## logit link  
    ystar[i] <- beta[1]       ## regression structure for covariates  
    + beta[2]*educ[i]  
    + beta[3]*(educ[i]*educ[i])  
    + beta[4]*age[i]  
    + beta[5]*(age[i]*age[i])  
    + beta[6]*south[i]  
    + beta[7]*govelec[i]  
    + beta[8]*closing[i]  
    + beta[9]*(closing[i]*educ[i])  
    + beta[10]*(educ[i]*educ[i]*closing[i])  
  }  
  
  ## priors  
  beta[1:10] ~ dmnorm(mu[] , B[ , ])      # diffuse multivariate Normal prior  
                                              # see data file  
}
```

## Binary Data Is Binomial Data when Grouped (§8.1.4)

- big, micro-level data sets with binary data (e.g., CPS)
- MCMC gets slow
- collapse the data into *covariate classes*, treat as *binomial* data; much smaller data set, much shorter run-times
- $y_i | \mathbf{x}_i \sim \text{Bernoulli}(F[\mathbf{x}_i \boldsymbol{\beta}])$ , where  $\mathbf{x}_i$  is a vector of covariates.
- *Covariate classes*: a set  $\mathcal{C} = \{i : \mathbf{x}_i = \mathbf{x}_{\mathcal{C}}\}$  i.e., the set of respondents who have covariate vector  $\mathbf{x}_{\mathcal{C}}$ .
- probability assignments over  $y_i \forall i \in \mathcal{C}$  are conditionally exchangeable given their common  $\mathbf{x}_i$  and  $\boldsymbol{\beta}$ .
- binomial model  $r_{\mathcal{C}} \sim \text{Binomial}(p_{\mathcal{C}}; n_{\mathcal{C}})$ , where  $p_{\mathcal{C}} = F(\mathbf{x}_{\mathcal{C}} \boldsymbol{\beta})$ ,  $r_{\mathcal{C}} = \sum_{i \in \mathcal{C}} y_i$  is the number of “successes” in  $\mathcal{C}$  and  $n_{\mathcal{C}}$  is the cardinality of  $\mathcal{C}$ .

# Example 8.5; binomial model for grouped binary data

Form covariate classes, and groupedData object; original data set  
 $n \approx 99000$ ; only 636 unique covariates classes.

R code

---

```
## collapse by covariate classes
X <- cbind(nagler$age,nagler$educYrs)
X <- apply(X,1,paste,collapse=":")
covClasses <- match(X,unique(X))
covX <- matrix(unlist(strsplit(unique(X) , ":")), ncol=2,byrow=TRUE)
r <- tapply(nagler$turnout,covClasses,sum)
n <- tapply(nagler$turnout,covClasses,length)
groupedData <- list(n=n, r=r,
                      age=as.numeric(covX[,1]),
                      educYrs=as.numeric(covX[,2]),
                      NOBS=length(n))
```

## Example 8.5; binomial model for grouped binary data

We can then pass the groupedData data frame to JAGS. We specify the binomial model  $r_i \sim \text{Binomial}(p_i; n_i)$  with  $p_i = F(\mathbf{x}_i\boldsymbol{\beta})$  and vague normal priors on  $\boldsymbol{\beta}$  with the following code:

---

JAGS code

---

```
model{
  for (i in 1:NOBS){
    logit(p[i]) <- beta[1] + age[i]*beta[2]
    + pow(age[i],2)*beta[3]
    + educYrs[i]*beta[4]
    + pow(educYrs[i],2)*beta[5]
    r[i] ~ dbin(p[i],n[i])  ## binomial model for each covariate class
  }

  beta[1:5] ~ dmnorm(b0[],B0[,])
}
```

# Ordinal Responses

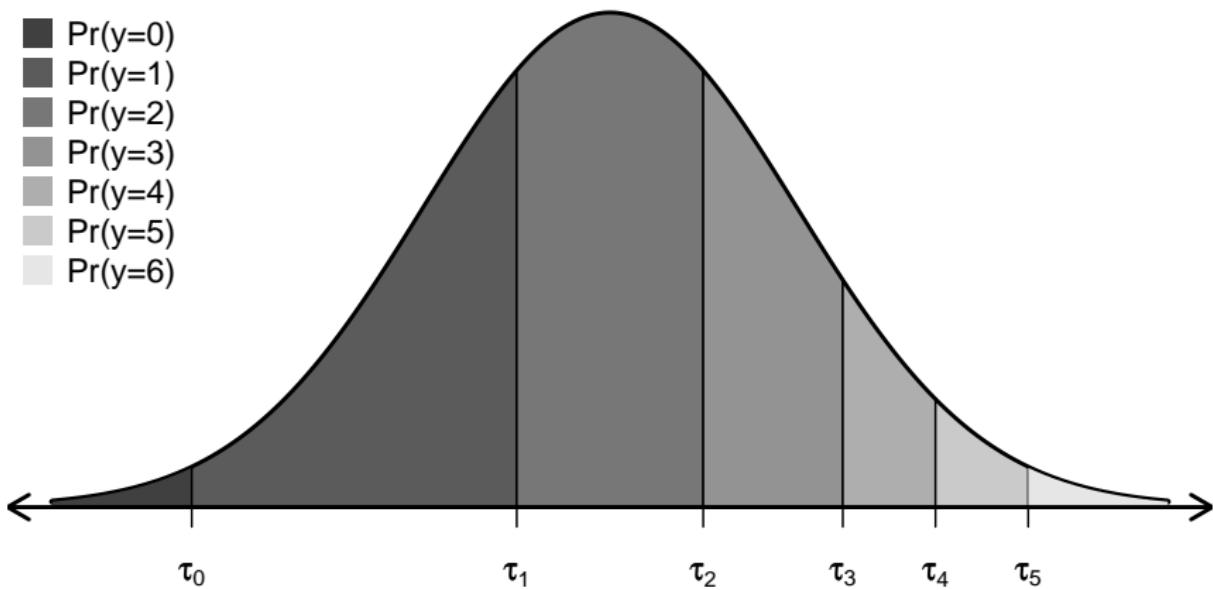
- e.g., 7-point scale when measuring party identification in the U.S., assigning the numerals  $y_i \in \{0, \dots, 6\}$  to the categories {"Strong Republican", "Weak Republican", ..., "Strong Democrat"}.
- Censored, latent variable representation:

$$\begin{aligned}y_i^* &= \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \\y_i = 0 &\iff y_i^* < \tau_1 \\y_i = j &\iff \tau_j < y_i^* \leq \tau_{j+1}, \quad j = 1, \dots, J-1 \\y_i = J &\iff y_i^* > \tau_J\end{aligned}$$

- threshold parameters obey the ordering constraint  
 $\tau_1 < \tau_2 < \dots < \tau_J$ .
- The assumption of normality for  $\varepsilon_i$  generates the probit version of the model; a logistic density generates the ordinal logistic model.
- Bayesian analysis: we want  $p(\boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}) p(\boldsymbol{\beta}, \boldsymbol{\tau})$ .

Ordinal responses,  $y_i^* \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$

$$\Pr[y_i = j] = \Phi[(\tau_{j+1} - \mathbf{x}_i\boldsymbol{\beta})/\sigma] - \Phi[(\tau_j - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$$



# Identification

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n.$$
$$y_i = 0 \iff y_i^* < \tau_1$$
$$y_i = j \iff \tau_j < y_i^* \leq \tau_{j+1}, \quad j = 1, \dots, J-1$$
$$y_i = J \iff y_i^* > \tau_J$$

- Model needs identification constraints
- Set one of the  $\tau$  to a point (zero); set  $\sigma$  to a constant (one)
- Drop the intercept and fix  $\sigma$
- Fix two of the  $\tau$  parameters.

# Priors on thresholds

- $\tau_j \sim N(0, 10^2)$ , subject to ordering constraint  $\tau_j > \tau_{j-1}, \forall j = 2, \dots, J$ .  
In JAGS only, use nifty sort function:

---

JAGS code

---

```
1 for(j in 1:4){  
2     tau0[j] ~ dnorm(0,.01)  
3 }  
4 tau[1:4] <- sort(tau0)    ## JAGS only, not in WinBUGS!
```

- BUGS:

$$\tau_1 \sim N(t_1, T_1)$$

$$\delta_j \sim \text{Exponential}(d), \quad j = 2, \dots, J,$$

$$\tau_j = \tau_{j-1} + \delta_j, \quad j = 2, \dots, J,$$

---

BUGS code

---

```
1 tau[1] ~ dnorm(0,.01)  
2 for(j in 1:3){  
3     delta[j] ~ dexp(2)  
4     tau[j+1] <- tau[j] + delta[j]  
5 }
```

## Example 8.6, interviewer ratings of respondents

- 5 point rating scale used by interviewers in assessing respondents' levels of political information
- In 2000 ANES:

Label	y	n	%
Very Low	0	105	6
Fairly Low	1	334	19
Average	2	586	33
Fairly High	3	450	25
Very High	4	325	18

- covariates: education, gender, age, home-owner, public sector employment

# Ordinal Logistic Model

JAGS code

```
model{  
    for(i in 1:N){ ## loop over observations  
        ## form the linear predictor (no intercept)  
        mu[i] <- x[i,1]*beta[1] +  
            x[i,2]*beta[2] +  
            x[i,3]*beta[3] +  
            x[i,4]*beta[4] +  
            x[i,5]*beta[5] +  
            x[i,6]*beta[6]  
  
        ## cumulative logistic probabilities  
        logit(Q[i,1]) <- tau[1]-mu[i]  
        p[i,1] <- Q[i,1]  
        for(j in 2:4){  
            logit(Q[i,j]) <- tau[j]-mu[i]  
            ## trick to get slice of the cdf we need  
            p[i,j] <- Q[i,j] - Q[i,j-1]  
        }  
        p[i,5] <- 1 - Q[i,4]  
        y[i] ~ dcat(p[i,1:5]) ## p[i,] sums to 1 for each i  
    }  
  
    ## priors over betas  
    beta[1:6] ~ dmmnorm(b0[],B0[,])  
  
    ## thresholds  
    for(j in 1:4){  
        tau0[j] ~ dnorm(0, .01)  
    }  
    tau[1:4] <- sort(tau0) ## JAGS only not in BUGS!  
}
```



# Redundant Parameterization

- exploit lack of identification
- run the MCMC algorithm deployed in the space of *unidentified parameters*
- *post-processing*: map MCMC output back mixes better than the MCMC algorithm in the space of the *identified* parameters
- get a better mixing Markov chain
- in ordinal model case, exploit lack of identification between thresholds and intercept parameters
- take care!

# Interviewer heterogeneity in scale-use

- Different interviewers use the rating scale differently: e.g., interviewer  $k$  is a tougher grader than interviewer  $k'$ .
- We tap this with a set of interviewer terms, varying over interviewers  $k = 1, \dots, K$
- We augment the usual ordinal model as follows:

$$\Pr(y_i \geq j) = F(\tau_j - \mu_i), \quad j = 0, \dots, J - 1$$

$$\Pr(y_i = J) = 1 - F(\tau_{J-1} - \mu_i)$$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta} + \eta_k$$

$$\eta_k \sim N(0, \sigma^2) \quad k = 1, \dots, K$$

- A positive  $\eta_k$  is equivalent to the thresholds being shifted down (i.e., interviewer  $k$  is an easier-than-average grader).
- Zero-mean restriction on  $\eta_k$ : why?
- Alternative model: each interviewer gets their own set of thresholds, perhaps fit these hierarchically.

# JAGS code for hierarchical model

---

## JAGS code

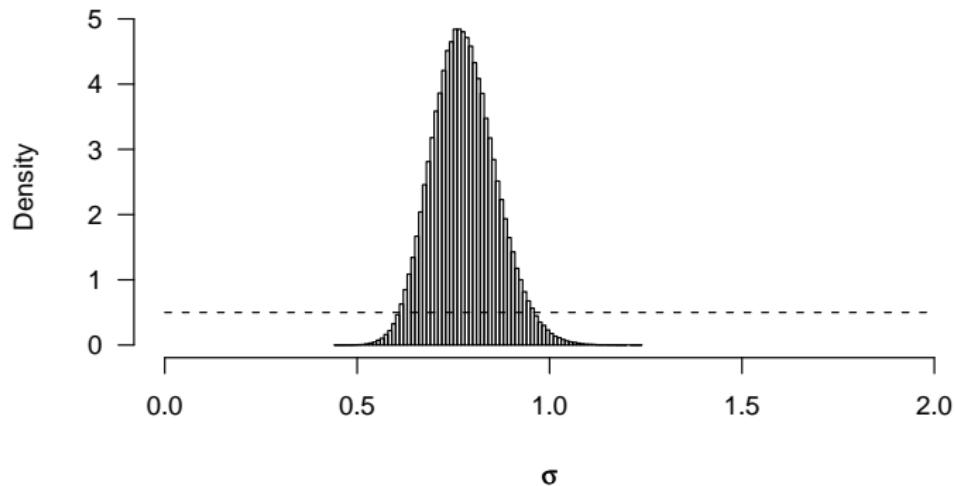
---

```
model{
  for(i in 1:N){ ## loop over observations
    ## form the linear predictor
    mu[i] <- x[i,1]*beta[1] + x[i,2]*beta[2] + x[i,3]*beta[3] +
      x[i,4]*beta[4] + x[i,5]*beta[5] + x[i,6]*beta[6] + eta[id[i]]

    ## cumulative logistic probabilities
    logit(Q[i,1]) <- tau[1]-mu[i]
    p[i,1] <- Q[i,1]
    for(j in 2:4){
      logit(Q[i,j]) <- tau[j]-mu[i]
      p[i,j] <- Q[i,j] - Q[i,j-1]
    }
    p[i,5] <- 1 - Q[i,4]
    y[i] ~ dcat(p[i,1:5]) ## p[i,] sums to 1 for each i
  }
  ## priors over betas
  beta[1:6] ~ dmmnorm(b0[],B0[,])
  ## hierarchical model over etas, note zero mean restriction
  for(k in 1:NID){
    eta[k] ~ dnorm(0.0,eta.tau)
  }
  eta.tau <- 1/pow(sigma,2) ## convert stddev to precision
  sigma ~ dunif(0,2)
  ## priors over thresholds
  for(j in 1:4){
    tau0[j] ~ dnorm(0,.01)
  }
  tau[1:4] <- sort(tau0) ## JAGS only, not in WinBUGS!
}
```



# $\sigma$ , prior and posterior densities



- since  $\eta_k \sim N(0, \sigma^2)$ , if we set  $\sigma$  to its posterior mean of .77, then half of the interviewer effects will lie more than  $1.35 \sigma \approx 1.04$  “logits” away from zero.

# Tabular summary of results

	Non-Hierarchical	Hierarchical
College Degree	1.46 (.10)	1.61 (.10)
Female	-.66 (.09)	-.76 (.09)
log(Age)	.47 (.12)	.42 (.13)
Home Owner	.45 (.10)	.48 (.10)
Government Employee	.17 (.14)	.16 (.14)
log(Interview Length)	1.13 (.15)	1.45 (.18)
$\sigma$	0	.77 (.08)
Threshold parameters:		
$\tau_0$	3.85 (.67)	4.69 (.75)
$\tau_1$	5.66 (.67)	6.60 (.75)
$\tau_2$	7.37 (.68)	8.46 (.76)
$\tau_3$	8.83 (.69)	10.08 (.77)

# Models for Multinomial Choices, §8.3

- multinomial logit (MNL), §8.3.1
- multinomial probit (MNP) §8.3.2

## Multinomial logit (MNL), §8.3.1

- Random utility rationale: utility to decision-maker  $i$  of choice  $j$  is linear in some predictors, plus a random component,

$$U_{ij} = \mathbf{x}_i \boldsymbol{\beta}_j + \varepsilon_{ij}, j = 0, \dots, J$$

- $\varepsilon_{ij}$  are drawn from a distribution whose cumulative distribution function is a Type-1 extreme value distribution with functional form  $F(\varepsilon_{ij}) = \exp[-\exp(-\varepsilon_{ij})]$  and hence  $\varepsilon_{ij}$  has density

$$p(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \exp[-\exp(-\varepsilon_{ij})].$$

- Decision-maker  $i$  chooses option  $j$  with probability

$$\pi_{ij} = \Pr(y_i = j) = \Pr[U_{ij} > U_{ik}], \quad \forall k \neq j.$$

## Multinomial logit (MNL), §8.3.1

- Consider a choice set with 3 elements, {"0", "1", "2"}.
- Suppose we observe  $y_i = 2$ :

$$\begin{aligned}\Pr(y_i = 2) &= \Pr(U_{i2} > U_{i1}, U_{i2} > U_{i0}) \\&= \Pr[\mathbf{x}_i\beta_2 + \varepsilon_{i2} > \mathbf{x}_i\beta_1 + \varepsilon_{i1}, \mathbf{x}_i\beta_2 + \varepsilon_{i2} > \mathbf{x}_i\beta_0 + \varepsilon_{i0}], \\&= \Pr[\varepsilon_{i2} + \mathbf{x}_i\beta_2 - \mathbf{x}_i\beta_1 > \varepsilon_{i1}, \varepsilon_{i2} + \mathbf{x}_i\beta_2 - \mathbf{x}_i\beta_0 > \varepsilon_{i0}], \\&= \int_{-\infty}^{\infty} f(\varepsilon_2) \left[ \int_{-\infty}^{\varepsilon_{i2} + \mathbf{x}_i\beta_2 - \mathbf{x}_i\beta_1} f(\varepsilon_1) d\varepsilon_1 \cdot \int_{-\infty}^{\varepsilon_{i2} + \mathbf{x}_i\beta_2 - \mathbf{x}_i\beta_0} f(\varepsilon_0) d\varepsilon_0 \right] d\varepsilon_2, \\&= \int_{-\infty}^{\infty} f(\varepsilon_2) \times \exp[-\exp(-\varepsilon_{i2} - \mathbf{x}_i\beta_2 + \mathbf{x}_i\beta_1)] \times \exp[-\exp(-\varepsilon_{i2} - \mathbf{x}_i\beta_2 + \mathbf{x}_i\beta_0)] \\&= \frac{\exp(\mathbf{x}_i\beta_2)}{\exp(\mathbf{x}_i\beta_0) + \exp(\mathbf{x}_i\beta_1) + \exp(\mathbf{x}_i\beta_2)}.\end{aligned}$$

- Thus:

$$\pi_{ij} = \Pr(y_i = j) = \frac{\exp(\mathbf{x}_i\beta_j)}{\sum_{k=0}^J \exp(\mathbf{x}_i\beta_k)}.$$

## Multinomial logit (MNL), §8.3.1

$$\pi_{ij} = \Pr(y_i = j) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{\sum_{k=0}^J \exp(\mathbf{x}_i \boldsymbol{\beta}_k)}.$$

- Identification by normalizing on a “baseline outcome”, e.g.,  $\boldsymbol{\beta}_0 = \mathbf{0}$ .
- Independence of irrelevant alternatives §8.3.2

## Example 8.7

- Vote choice in the 1992 U.S. Presidential election
- ANES data; choices are Clinton, George H.W. Bush, Perot.  $n = 909$ .
- Original analysis by Alvarez and Nagler (1995), who used MNP.
- Predictors: dummies for Dem or Rep party-id, dummy for gender, retrospective evaluations of the national economy (-1, 0, 1), and  $z_{ij}$ , square of the distance of respondent  $i$  from candidate  $j$ .
- 

$$\begin{aligned}\Pr(U_{ij} > U_{ik}) &= \Pr(\mathbf{x}_i\boldsymbol{\beta}_j + z_{ij}\gamma + \varepsilon_{ij} - \mathbf{x}_i\boldsymbol{\beta}_k - z_{ik}\gamma - \varepsilon_{ik} > 0) \\ &= \Pr(\mathbf{x}_i[\boldsymbol{\beta}_j - \boldsymbol{\beta}_k] + [z_{ij} - z_{ik}]\gamma > \varepsilon_{ik} - \varepsilon_{ij}).\end{aligned}$$

## Example 8.7, using dcat

JAGS code

```
model{  
    for(i in 1:NOBS){  
        for(j in 1:3){  ## loop over choices  
            mu[i,j] <- beta[j,1]  
            + beta[j,2]*dem[i]  
            + beta[j,3]*ind[i]  
            + beta[j,4]*rep[i]  
            + beta[j,5]*female[i]  
            + beta[j,6]*natlecon[i]  
            + gamma*dist[i,j]  
            emu[i,j] <- exp(mu[i,j])  
            p[i,j] <- emu[i,j]/sum(emu[i,1:3])  
        }  
        y[i] ~ dcat(p[i,1:3])  
    }  
  
    ## priors  
    for(k in 1:6){  
        beta[1,k] <- 0  ## identifying restriction  
    }  
    for(j in 2:3){  
        beta[j,1:6] ~ dmmnorm(b0,B0)  ## b0, B0 passed as data from R  
    }  
    gamma ~ dnorm(0,.01)  
  
    ## plus code for mapping to identified parameters, see book  
}
```

# Multinomial Probit, §8.4

- same random utility rationale:

$$U_{ij} = \mathbf{r}_{ij}\boldsymbol{\beta} + v_{ij}, \quad j = 0, 1, \dots, J; i = 1, \dots, n$$

- MNP for MVN model for un-modelled sources of utility:

$$\mathbf{v}_i = (v_{i1}, \dots, v_{iJ})' \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{V})$$

where  $\mathbf{V}$  is a  $(J + 1)$ -by- $(J + 1)$  covariance matrix.

- But probabilities are difficult to compute:

$$\begin{aligned}\pi_{ij} &= \Pr(y_i = j) = \Pr(U_{ij} > U_{ik}), \quad \forall k \neq j \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{U_{ij}} \cdots \int_{-\infty}^{U_{ij}} f(U_0, U_1, \dots, U_J) dU_0 dU_1 \dots dU_j\end{aligned}$$

# Multinomial Probit, MCMC via data augmentation

- if choice  $j$  is observed for person  $i$ , we know that  $U_{ij} - U_{ik} > 0 \forall j \neq k$ .
- Without loss of generality choose a “baseline” outcome,  $j = 0$ , and define the utility differences  $\mathbf{w}_i = (w_{i1}, \dots, w_{iJ})'$  with  $w_{ij} = U_{ij} - U_{i0}, j = 1, \dots, J$ :

$$w_{ij} = (\mathbf{r}_{ij} - \mathbf{r}_{i0})\boldsymbol{\beta} + v_{ij} - v_{i0} = \mathbf{x}_{ij}\boldsymbol{\beta} + \varepsilon_{ij}.$$

where  $\boldsymbol{\varepsilon}_i \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$

- mapping from latent variables to observed choices:

$$y_i = h(\mathbf{w}_i) \equiv \begin{cases} 0 & \text{if } \max(\mathbf{w}_i) < 0 \\ j & \text{if } \max(\mathbf{w}_i) = w_{ij} > 0 \end{cases}$$

- Identification: the distribution of  $\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}$  is the same as the distribution of  $\mathbf{y}|\mathbf{X}, c\boldsymbol{\beta}, c^2\boldsymbol{\Sigma}$
- solution: set  $\sigma_{11} = 1$ .

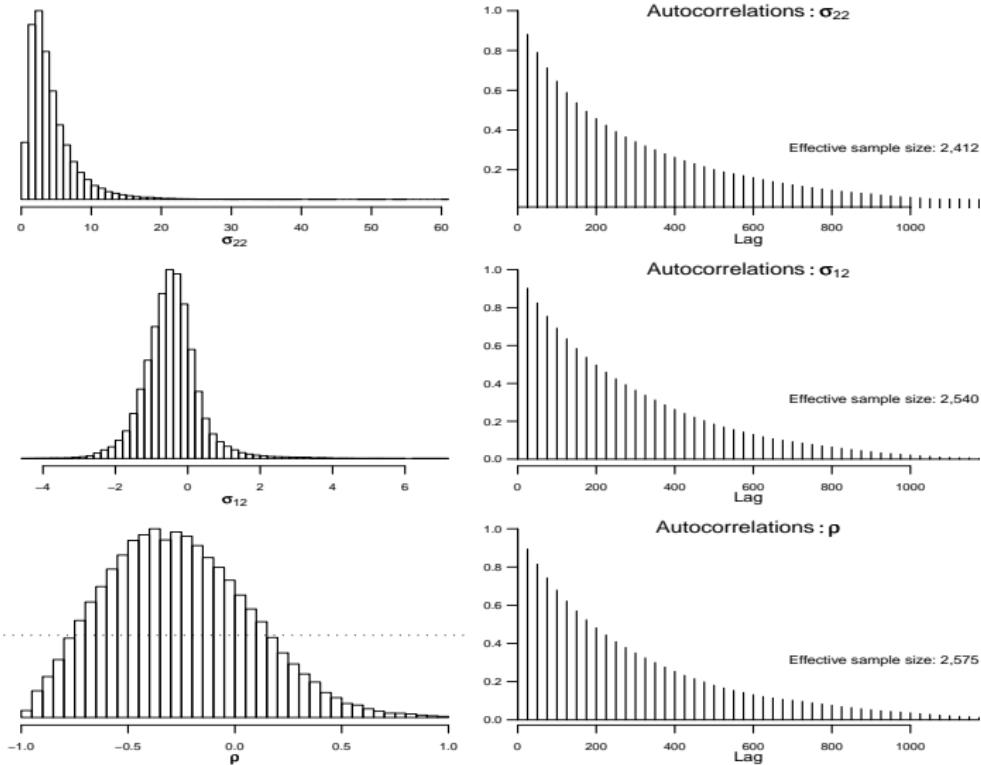
# Multinomial Probit, MCMC via data augmentation

- posterior density  $p(\beta, \Sigma | \mathbf{y}, \mathbf{X})$ 
  - 1 sample  $\mathbf{w}_i^{(t)}$  from  $p(\mathbf{w}_i | \beta^{(t-1)}, \Sigma^{(t-1)}, \mathbf{y}, \mathbf{X}), i = 1, \dots, n$ , the data-augmentation step
  - 2 sample  $\beta^{(t)}$  from  $p(\beta | \Sigma^{(t-1)}, \mathbf{W}^{(t)}, \mathbf{y}, \mathbf{X})$ .
  - 3 sample  $\Sigma^{(t)}$  from  $p(\Sigma | \beta^{(t)}, \mathbf{W}^{(t)}, \mathbf{y}, \mathbf{X})$ .
- Conditional on the latent  $\mathbf{w}_i$ , we have a very simple multivariate normal regression (McCulloch and Rossi 1994; Chib and Greenberg 1997; McCulloch, Polson and Rossi 1998).
- For step 3, the prior and the conditional distribution for  $\Sigma$  is complicated by the identifying constraint  $\sigma_{11} = 1$ .
- Implemented in MNP package in R (Imai and van Dyk 2005).

## Example 8.8, 1992 U.S. Presidential election

- $n = 909, j \in \{\text{Perot, Bush, Clinton}\}$
- mix of individual (party-id, gender, evaluations of the economy) and choice-specific covariates (squared ideological distance from candidates)
- MNP in R, 1.5M iterations, extremely inefficient exploration of the posterior densities for some parameters

# Example 8.8, 1992 U.S. Presidential election



# Example 8.8, 1992 U.S. Presidential election

1.5 million iterations:

	<i>z</i>	<i>p</i>	$\rho$	<i>N</i>	<i>I</i>	EffSamp
$\beta_{11}$ , Intercept, Perot	-0.82	0.62	0.37	303,375	81.00	5,908
$\beta_{21}$ , Intercept, Clinton	-0.36	0.36	0.46	547,875	146.00	6,211
$\beta_{12}$ , Dem Id, Perot	-1.12	0.49	0.48	320,025	85.40	4,076
$\beta_{22}$ , Dem Id, Clinton	-0.27	0.69	0.73	860,850	230.00	3,175
$\beta_{13}$ , Repub Id, Perot	0.66	0.72	0.34	659,850	176.00	6,599
$\beta_{23}$ , Repub Id, Clinton	0.42	0.72	0.75	958,400	256.00	2,763
$\beta_{14}$ , Female, Perot	-0.07	0.32	0.01	94,025	25.10	58,544
$\beta_{24}$ , Female, Clinton	1.04	0.51	0.02	99,850	26.70	50,304
$\beta_{15}$ , Econ Retro, Perot	0.51	0.97	0.18	198,950	53.10	11,272
$\beta_{25}$ , Econ Retro, Clinton	0.15	0.57	0.59	607,750	162.00	3,709
$\gamma$ , Ideological Distance	0.21	0.57	0.72	967,950	258.00	2,778
$\sigma_{12}$	-1.46	0.09	0.90	975,375	260.00	2,540
$\sigma_{22}$	-0.78	0.98	0.88	902,500	241.00	2,412
$\rho$	-1.16	0.33	0.89	1,159,350	309.00	2,575

# References

- Alvarez, R. Michael and Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714--44.
- Chib, Siddhartha and Edward Greenberg. 1997. "Analysis of Multivariate Probit Models." *Biometrika* 85:347--361.
- Imai, Kosuke and David A. van Dyk. 2005. "MNP: R Package for Fitting the Multinomial Probit Model." *Journal of Statistical Software* 14:1--32.
- McCulloch, Robert E., Nicholas G. Polson and Peter E. Rossi. 1998. "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters." Typescript. Graduate School of Business, University of Chicago.
- McCulloch, Robert E. and Peter E. Rossi. 1994. "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics* 64:207--40.

# *Ordered Probit*

Econ 674

Purdue University

March 9, 2009

In some cases, the variable to be modeled has a natural *ordinal* interpretation.

Some examples include:

- ① Education, measured categorically, (e.g. 1 = < HS, 2 = HS, 3 = Some college, etc.).
- ② Income, also measured categorically.
- ③ Survey responses, coded as a degree of opinion (e.g. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree).
- ④ These are in contrast to other choices such as type of insurance or selected mode of transportation, for example, that are not ordered.
- In this lecture we discuss *ordinal choice* models, and focus on the *ordered probit* in particular.

## The Ordered Probit Model

Suppose that the variable to be modeled,  $y$  takes on  $J$  different values, which are naturally ordered:

$$y_i = \begin{cases} 1 \\ 2 \\ \vdots \\ J \end{cases}, \quad i = 1, 2, \dots, n.$$

As with the probit model, we assume that the observed  $y$  is generated by a latent variable  $y^*$ , where



The link between the latent and observed data is given as follows:



## The Ordered Probit Model

The  $\alpha_j$  are called *cutpoints* or *threshold* parameters. They are estimated by the data and help to match the probabilities associated with each discrete outcome.

Without any additional structure, the model is not identified. In particular, there are too many cutpoints and some restrictions are required. The most common way to achieve identification is to set:

- 

and retain an intercept parameter in the model.

What happens, for example when  $J = 2$  and these restrictions are imposed?

## The Ordered Probit Model

The likelihood for the ordered probit is simply the product of the probabilities associated with each discrete outcome:

$$\bar{L}(\beta, \alpha) = \prod_{i=1}^n \Pr(y_i = j|x_i),$$

where

$$\alpha = [\alpha_3 \ \alpha_4 \ \cdots \ \alpha_J].$$

The  $i^{th}$  observation's contribution to the likelihood is



## The Ordered Probit Model

Therefore,

$$\bar{L}(\beta, \alpha) = \prod_{i=1}^n \Phi(\alpha_{y_i+1} - x_i\beta) - \Phi(\alpha_{y_i} - x_i\beta)$$

and



For purposes of computing the MLE, it can be useful to define

$$Z_{ij} = I(y_i = j).$$

Thus, we can write:

$$L(\beta, \alpha) = \sum_{i=1}^n \sum_{j=1}^J z_{ij} (\log [\Phi(\alpha_{j+1} - x_i\beta) - \Phi(\alpha_j - x_i\beta)]).$$

(Some textbooks present the material this way, though we will not make use of this here).

## The Ordered Probit Model

This yields the score for the parameter vector  $\beta$ :

- 

and likewise, we obtain the FOC for  $\alpha_k$ ,  $k = 3, 4, \dots, J$ :

$$\begin{aligned} L_{\alpha_k}(\beta, \alpha) &= \sum_{i:y_i=k} -\frac{\phi(\alpha_k - x_i\beta)}{\Phi(\alpha_{k+1} - x_i\beta) - \Phi(\alpha_k - x_i\beta)} \\ &+ \sum_{i:y_i=k-1} \frac{\phi(\alpha_k - x_i\beta)}{\Phi(\alpha_k - x_i\beta) - \Phi(\alpha_{k-1} - x_i\beta)} \end{aligned}$$

We do not report the Hessian here, as the expressions are rather lengthy. Nonetheless, standard MLE can be applied.

## Marginal Effects

To fix ideas, consider the case of an ordered probit model with  $J = 3$ , in which case we have:

- 

From these, we obtain the category-specific *marginal effects*:

-

## Marginal Effects

What do we learn from this simple model?

- ① Like the probit, the marginal effects depend on  $x$ . We can evaluate these at sample means, or take a sample average of the marginal effects.
- ② Unlike the probit, *the signs of the “interior” marginal effects are unknown* and not completely determined by the sign of  $\beta_k$ .
- ③ We can, however, sign the effects of the lowest and highest categories based on  $\beta_k$ . The others, however, can not be known by the reader simply by looking at a table of point estimates.

## Interpretation

Continue to consider the case with  $J = 3$  and suppose there are no covariates and only an intercept parameter is included.

In this case we have

$$\Pr(y_i = 1) = 1 - \Phi(\beta)$$

$$\Pr(y_i = 2) = \Phi(\alpha - \beta) - [1 - \Phi(\beta)] = \Phi(\alpha - \beta) - \Phi(-\beta)$$

$$\Pr(y_i = 3) = 1 - \Phi(\alpha - \beta)$$

What do you think will happen in terms of the MLE's?

The likelihood is:

$$\bar{L}(\alpha, \beta) = \prod_{i:y_i=1} [1 - \Phi(\beta)] \prod_{i:y_i=2} [\Phi(\alpha - \beta) - \Phi(-\beta)] \prod_{i:y_i=3} [1 - \Phi(\alpha - \beta)]$$

which reduces to



## Interpretation

In the last slide, we have defined  $n_j$  as the number of observations for which  $y_i = j$  and also note that  $n_1 + n_2 + n_3 = n$ .  
Thus we obtain the log-likelihood

$$L(\alpha, \beta) = n_1 \log[1 - \Phi(\beta)] + n_2 \log[\Phi(\alpha - \beta) - \Phi(-\beta)] + n_3 \log[1 - \Phi(\alpha - \beta)].$$

The  $\alpha$  FOC gives:



with  $\hat{P}_j$  denoting the fitted probability for category  $j$ .

## Interpretation

Likewise, we get an FOC for  $\beta$ :

$$-n_1 \frac{\phi(\hat{\beta})}{1 - \Phi(\hat{\beta})} + n_2 \frac{\phi(\hat{\beta}) - \phi(\hat{\alpha} - \hat{\beta})}{\Phi(\hat{\alpha} - \hat{\beta}) - \Phi(-\hat{\beta})} + n_3 \frac{\phi(\hat{\alpha} - \hat{\beta})}{1 - \Phi(\hat{\alpha} - \hat{\beta})} = 0.$$

Grouping terms, and using our  $\alpha$  FOC, the  $\beta$  FOC can be shown to imply:

$$\frac{n_2}{\hat{P}_2} = \frac{n_1}{\hat{P}_1}.$$

## interpretation

Noting that

$$\hat{P}_1 + \hat{P}_2 + \hat{P}_3 = 1$$

and

$$n_1 + n_2 + n_3 = n,$$

these two FOC's can be manipulated to yield:



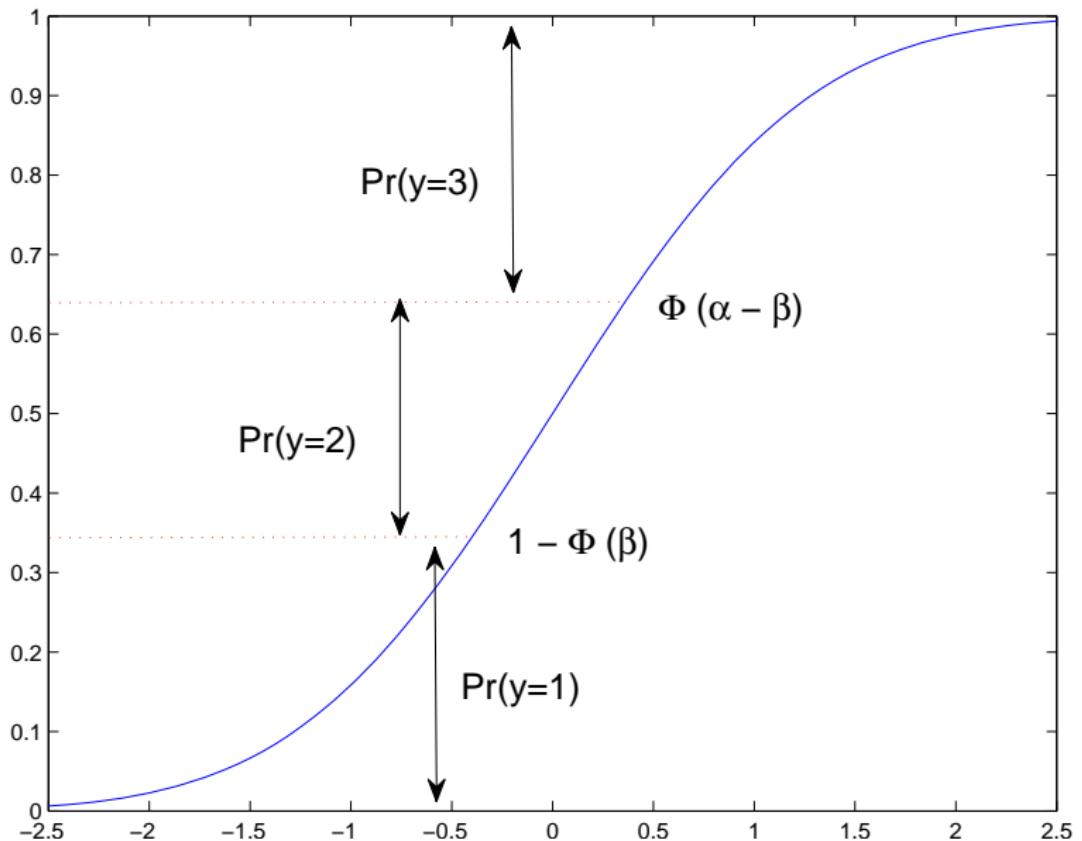
*That is, the parameters will be selected so that the fitted values of each category exactly match the observed frequencies of outcomes in that category.* Note that this generalizes to any  $J$  and any link function!

For  $\hat{\beta}$ , for example, we obtain:

$$1 - \Phi(\hat{\beta}) = n_1/n$$

or

$$\hat{\beta} = \Phi^{-1}\left(\frac{n - n_1}{n}\right).$$



## Ordered Probit and the EM Algorithm

Suppose that  $J = 3$  and consider the following model:

$$y_i^* = x_i\beta + \epsilon_i, \quad \epsilon_i | x_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$$

and

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq 0 \\ 2 & \text{if } 0 < y_i^* \leq \alpha \\ 3 & \text{if } y_i^* > \alpha. \end{cases}$$

Let



## Ordered Probit and the EM Algorithm

This reparameterization defines an equivalent model:

- 

where

- 

Note that, in this representation, there are *no unknown cutpoints*.

# *Ordered Probit and the EM Algorithm*

## **Step 1: E-Step**

Note that



and thus

$$L(\delta, \sigma^2; z^*) = \text{constant} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (z^* - X\delta)'(z^* - X\delta).$$

The E-step is completed by taking expectations over  $z^* | \theta = \theta_t, y$ :

$$E[L(\delta, \sigma^2; z^*)] \equiv \text{constant} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} E_{z^* | \theta = \theta_t, y} (z^* - X\delta)'(z^* - X\delta).$$

## Ordered Probit and the EM Algorithm

### **Step 2: M-Step:**

To implement the  $M-$  step, we must evaluate this expectation and then maximize over  $\delta$  and  $\sigma^2$ .

You will probably recognize the  $\delta$ -part of this exercise. It will follow similarly to the probit, where:

- 

with

-

To evaluate this mean, suppose



and we seek



That is,



## *Ordered Probit and the EM Algorithm*

Applying this result, we can evaluate  $\mu(\delta_t, \sigma_t^2, y)$

$$\mu(\delta_t, \sigma_t^2, y_i) = \begin{cases} x_i \delta_t - \sigma \frac{\phi(x_i \delta_t / \sigma_t)}{\Phi(-x_i \delta_t / \sigma_t)} & \text{if } y_i = 1 \\ x_i \delta_t + \sigma \frac{\phi(x_i \delta_t / \sigma_t) - \phi([1-x_i \delta_t] / \sigma_t)}{\Phi([1-x_i \delta_t] / \sigma_t) - \Phi(-x_i \delta_t / \sigma_t)} & \text{if } y_i = 2 \\ x_i \delta_t + \sigma \frac{\phi([1-x_i \delta_t] / \sigma_t)}{1 - \Phi([1-x_i \delta_t] / \sigma_t)} & \text{if } y_i = 3 \end{cases}$$

Thus, the parameters  $\delta$  are easily updated.

## *Ordered Probit and the EM Algorithm*

As for the updating of  $\sigma^2$ , note that it will be obtained as:

$$\sigma_{t+1}^2 = \frac{1}{n} \sum_i E_{z_i^* | \delta = \delta_t, \sigma^2 = \sigma_t^2, y_i} (z_i^* - x_i \delta_{t+1})^2.$$

Expanding this out, we obtain (dropping the subscript on the expectation for simplicity):

$$\sigma_{t+1}^2 = \frac{1}{n} \sum_i [E([z_i^*]^2) - 2\mu(\delta_t, \sigma_t^2, y_i)x_i \delta_{t+1} + (x_i \delta_{t+1})^2].$$

Only the first term in the summation above requires further evaluation. We first note that

$$E([z_i^*]^2) = (x_i \delta_t)^2 + 2x_i \delta_t E(v_i | \delta_t, \sigma_t^2, y_i) + E(v_i^2 | \delta_t, \sigma_t^2, y_i).$$

We first recognize that

$$E(v_i | \theta_t, y_i) = \mu(\delta_t, \sigma_t^2, y_i) - x_i \delta_t.$$

## *Ordered Probit and the EM Algorithm*

As for the  $E(v_i^2|\delta_t, \sigma_t^2, y_i)$  term, a little work gives:

$$E(v_i^2|\delta_t, \sigma_t^2, y_i) = \sigma^2 \left[ 1 + \frac{-x_i\delta_t/\sigma_t \phi(x_i\delta_t/\sigma_t) - [1-x_i\delta_t]/\sigma_t \phi([1-x_i\delta_t]/\sigma_t)}{\Phi([1-x_i\delta_t]/\sigma_t) - \Phi(-x_i\delta_t/\sigma_t)} \right].$$

## *Ordered Probit and the EM Algorithm*

So, we have everything we need to implement the EM algorithm. It would proceed as follows:

- ① Pick some starting values.
- ② Calculate  $\mu(\delta_t, \sigma_t^2, y_i)$  using the formula provided. Use this to update  $\delta_t$  to  $\delta_{t+1}$ .
- ③ Calculate  $E(v_i^2 | \delta_t, \sigma_t^2, y_i) \forall i$  using the formula provided, and use it [together with  $\delta_{t+1}$  and  $\mu(\delta_t, \sigma_t^2, y_i)$ ] to update  $\sigma_t^2$  to  $\sigma_{t+1}^2$ .
- ④ Iterate to convergence.
- ⑤ Transform back by setting

$$\hat{\alpha} = \hat{\sigma}^{-1}, \quad \hat{\beta} = \hat{\delta}\hat{\alpha}.$$

## *Ordered Probit and the EM Algorithm*

We use  $n = 139$  law school applications from 1985.

The dependent variable is the rank of the law school with  $y = 1$  if the rank is less than or equal to 25,  $y = 2$  if the rank is between 25 and 50 and  $y = 3$  if the rank exceeds 50.

The independent variables include the applicant's LSAT score, GPA and student/faculty ratio (the latter is rather questionable).

In the following slides, we present the EM ordered probit estimates (which matched STATA's EXACTLY and were obtained faster!) We report some statistics evaluated at the sample mean of the x's and also setting *LSAT* and *GPA* to their maximum sample values.

## *Ordered Probit and the EM Algorithm*

$$\hat{\beta} = [49.1 \quad - .24 \quad - 2.73 \quad - .01], \quad \hat{\alpha} = 1.00.$$

Category	Fitted Probability		Marginal Effect		
	xbar	xmax	LSAT/xbar	LSAT/ xmax	GPA/ xbar
$y = 1$	.04	.99	.02	.003	.25
$y = 2$	.20	.01	.05	-.003	.59
$y = 3$	.76	.00	-.08	.000	-.85

# BAYESIAN INFERENCE FOR LATENT STATES

SIMON JACKMAN

Stanford University  
<http://jackman.stanford.edu/BASS>

February 3, 2012

# Chapter 9 of *BASS*

- factor analysis
- item-response theory (IRT) models
- dynamic linear model

# Inference for Latent States

- latent quantities  $\xi$
- observed quantities  $\mathbf{Y}$
- unobserved parameters  $\beta$ , indexing the functional relationship between  $\mathbf{Y}$  and  $\xi$
- $\boldsymbol{\theta} = (\xi, \beta)'$
- Bayesian analysis:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

# Factor Analysis

- factor analysis typically presented as a model for *covariance structure*:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

where  $\boldsymbol{\Lambda}$  is a  $p$ -by- $k$  matrix of factor loadings,  $\boldsymbol{\Phi} = \mathbf{I}_k$  and  $\boldsymbol{\Psi}$  is a diagonal  $p$ -by- $p$  matrix with “uniquenesses” on the diagonal.

- obscures the fact that factor analysis is a model for observables conditional on unobservables
- at the level of the indicators, a Gaussian measurement model is

$$y_{ij} \sim N(\gamma_{j0} + \gamma_{j1}\xi_i, \omega_j^2)$$

where  $i$  indexes observations,  $j$  indexes  $p$  indicators,  $\gamma_{j0}$  and  $\gamma_{j1}$  are intercept and slope parameters,  $\omega^2$  is a measurement error variance and  $\xi_i$  are latent states.

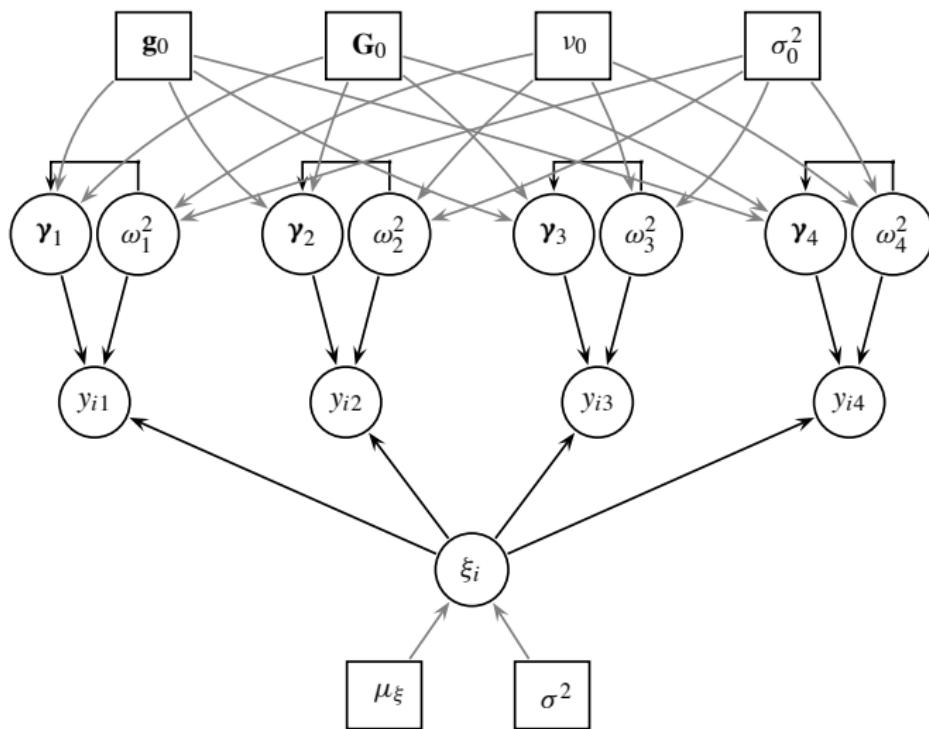
# Conjugate prior densities for Gaussian factor analysis model

$$\begin{aligned}\xi_i &\stackrel{\text{iid}}{\sim} N(\mu_\xi, \sigma^2), \quad i = 1, \dots, n, \\ \gamma_j | \omega_j^2 &\sim N(\mathbf{g}_{j0}, \omega_j^2 \mathbf{G}_{j0}), \quad j = 1, \dots, p, \\ \omega_j^2 &\sim \text{inverse-Gamma}(v_{j0}/2, v_{j0}\omega_{j0}^2/2), \quad j = 1, \dots, p,\end{aligned}$$

where  $\mu_\xi, \sigma^2, \mathbf{g}_{j0}, \mathbf{G}_{j0}, v_{j0}$  and  $\omega_{j0}^2, j = 1, \dots, p$  are user-specified hyper-parameters.

# Factor Analysis in Terms of Latent Variables

DAG, four indicator model, suggests Gibbs sampling scheme etc



# Identification

- model parameters not identified
- location shifts in  $\xi_i$  can be offset by shifts in intercepts  $\gamma_{j0}$ .
- scale shifts in  $\xi_i$  can be offset by rescaling slopes  $\gamma_{j1}$ .
- scale shifts in  $\omega_j$  can be offset with re-scalings of  $\gamma_{j0}, \gamma_{j1}, \xi_i$ .
- lack of identification not a formal problem for Bayesian analysis
- nonetheless, we deal with by imposing a location/scale restriction (“normalization”) on the  $\xi_i$ : mean zero, standard deviation one.

# Posterior inference via Gibbs sampling

$$\xi_i | \mu_\xi, \sigma^2, \boldsymbol{\Gamma}, \boldsymbol{\Psi}, \mathbf{y}_i \sim N(\mu_\xi^*, \sigma^{2*})$$

where

$$\mu_\xi^* = \frac{\frac{\mu_\xi}{\sigma^2} + \frac{\hat{\xi}_i}{V(\hat{\xi}_i)}}{\frac{1}{\sigma^2} + \frac{1}{V(\hat{\xi}_i)}} \quad \text{and} \quad \sigma^{2*} = \frac{\omega_j^2}{\frac{1}{\sigma^2} + \frac{1}{V(\hat{\xi}_i)}}$$

and where

$$\hat{\xi}_i = (\mathbf{Y}'_1 \mathbf{Y}_1)^{-1} \mathbf{w}'_i \mathbf{Y}_1 = \sum_{j=1}^p \gamma_{j1}^2 / \sum_{j=1}^p w_{ij} \gamma_{j1} \quad \text{and}$$

$$V(\hat{\xi}_i) = \omega_j^2 (\mathbf{Y}'_1 \mathbf{Y}_1)^{-1} = \frac{\omega_j^2}{\sum_{j=1}^p \gamma_{j1}^2},$$

# Posterior inference via Gibbs sampling

$$\boldsymbol{\gamma}_j | \mathbf{g}_{j0}, \mathbf{G}_{j0}, \boldsymbol{\xi}, \omega_j^2, \mathbf{y}_j \sim N(\mathbf{g}_{j1}, \omega_j^2 \mathbf{G}_{j1}),$$

where

$$\mathbf{g}_{j1} = (\mathbf{G}_{j0}^{-1} + \mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{G}_{j0}^{-1}\mathbf{g}_{j0} + \mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\gamma}}_j),$$

$$\mathbf{G}_{j1} = (\mathbf{G}_{j0}^{-1} + \mathbf{Z}'\mathbf{Z})^{-1},$$

$$\mathbf{Z} = [\mathbf{l} \ \boldsymbol{\xi}] \text{ and}$$

$$\hat{\boldsymbol{\gamma}}_j = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

That is,  $\mathbf{Z}$  is the  $n$ -by-2 matrix formed with a unit vector  $\mathbf{l}$  in the first column and the  $\boldsymbol{\xi}$  in the second column (a regressor matrix for the purposes of inference for  $\boldsymbol{\gamma}_j$ ).

# Posterior inference via Gibbs sampling

$\omega_j^2 | v_{j0}, \sigma_{j0}^2, \mathbf{Y}_j, \boldsymbol{\xi}, \mathbf{y}_j \sim \text{inverse-Gamma}(v_1/2, v_1\sigma_1^2/2),$

where

$$v_1 = v_0 + n,$$

$$v_1\sigma_1^2 = v_0\sigma_0^2 + S_j + r_j,$$

$$S_j = (\mathbf{y} - \mathbf{Z}\hat{\mathbf{Y}}_j)'(\mathbf{y} - \mathbf{Z}\hat{\mathbf{Y}}_j) \quad \text{and}$$

$$r_j = (\mathbf{g}_{j0} - \hat{\mathbf{Y}}_j)'(\mathbf{G}_{j0} + (\mathbf{Z}'\mathbf{Z})^{-1})^{-1}(\mathbf{g}_{j0} - \hat{\mathbf{Y}}_j).$$

## References



# BAYESIAN TIME SERIES

*A (hugely selective) introductory overview*

*- contacting current research frontiers -*

Mike West

Institute of Statistics & Decision Sciences

Duke University

*June 5th 2002, Valencia VII - Tenerife*

## Topics

### Dynamic linear models (state space models)

- Sequential context, Bayesian framework
- Standard classes of models, model decompositions

### Models and methods in physical science applications

- Time series decompositions, latent structure
- Neurophysiology - climatology - speech processing

### Multivariate time series:

- Financial applications - Latent structure, volatility models

### Simulation-Based Computation

- MCMC - Sequential simulation methodology

## Standard Dynamic Models

### Dynamic Linear Models

$$y_t = x_t + \nu_t$$

$$x_t = \mathbf{F}'_t \boldsymbol{\theta}_t$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \omega_t$$

### Linear State Space Models

- signal  $x_t$ , state vector  $\boldsymbol{\theta}_t = (\theta_{t1}, \dots, \theta_{td})'$
- regression vector  $\mathbf{F}_t$  and state matrix  $\mathbf{G}_t$
- zero mean measurement errors  $\nu_t$  and state innovations  $\omega_t$ 
  - often zero-mean and normally distributed

## Examples

“Slowly varying” level observed with noise:

$$y_t = x_t + \nu_t \quad x_t = x_{t-1} + \omega_t$$

Dynamic linear regression:

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

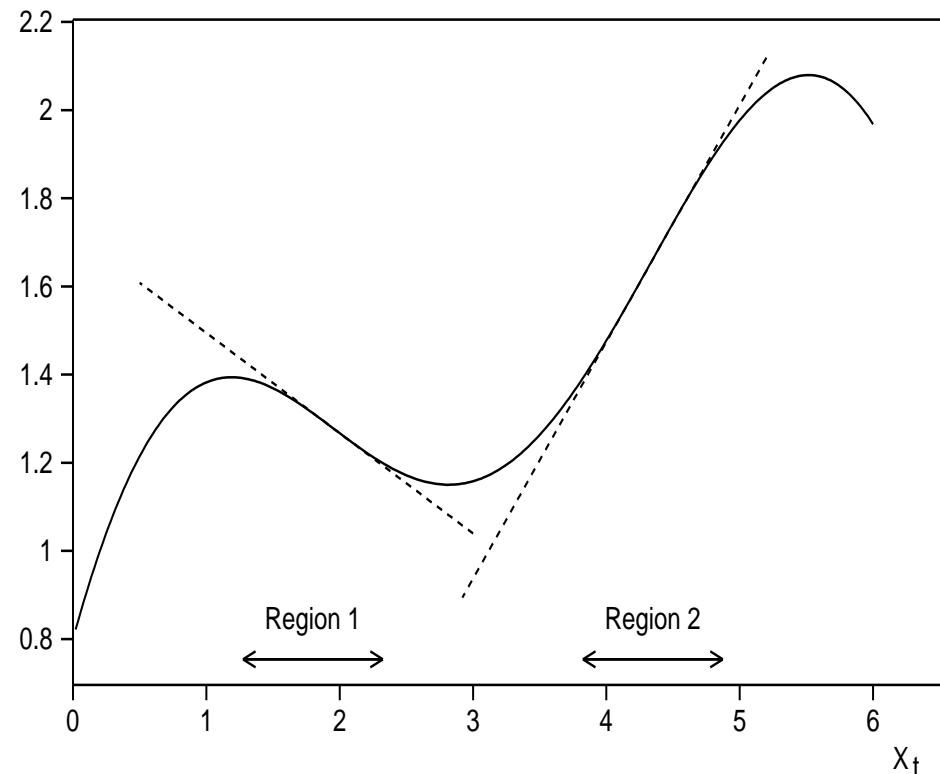
Error models for  $\nu_t, \omega_t$

- normal distributions
- mixtures of normals: outliers and abrupt “structural” changes

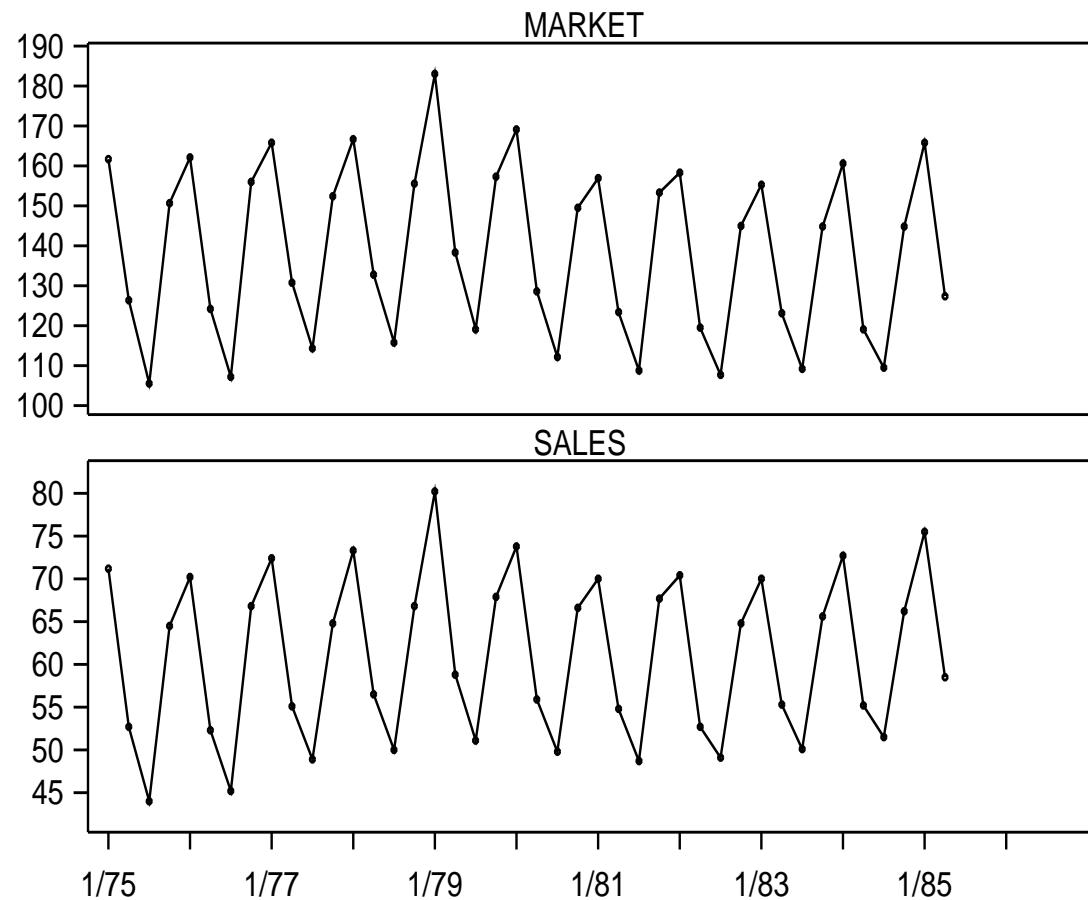
## Simple regression example

$$y_t = x_t + \nu_t \quad x_t = a_t + b_t X_t$$

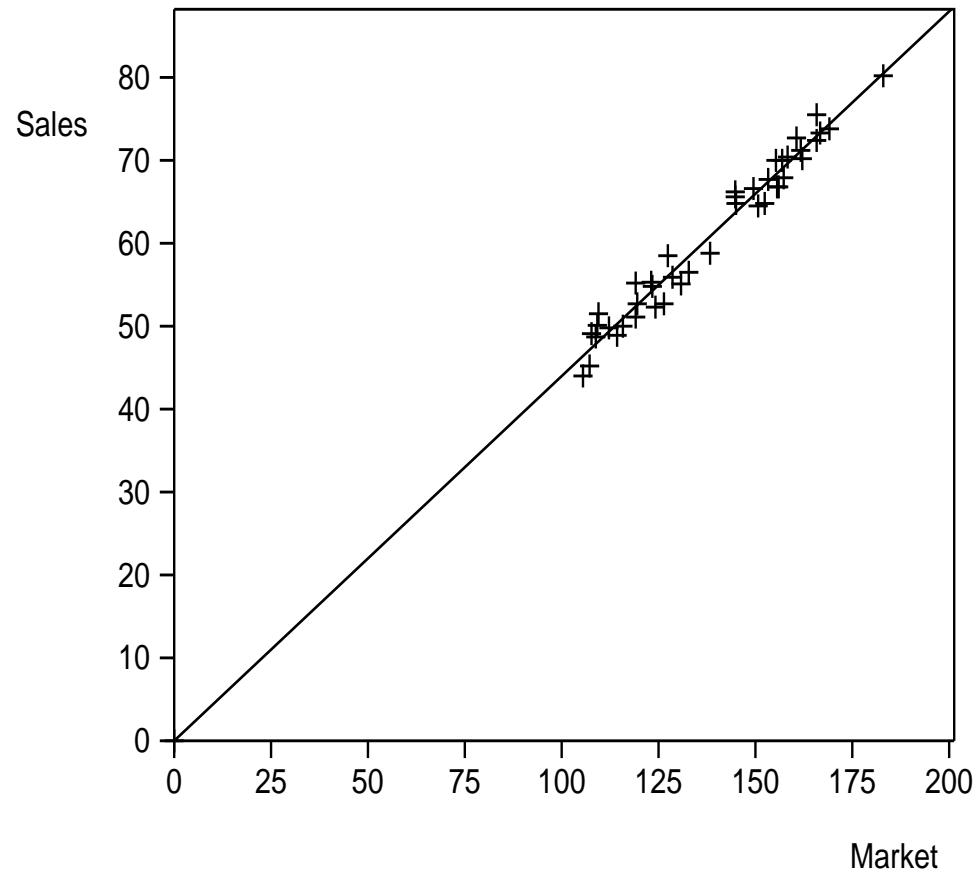
$\mathbf{F}_t = (1, X_t)'$  and  $\boldsymbol{\theta}_t = (a_t, b_t)'$  “wanders” through time



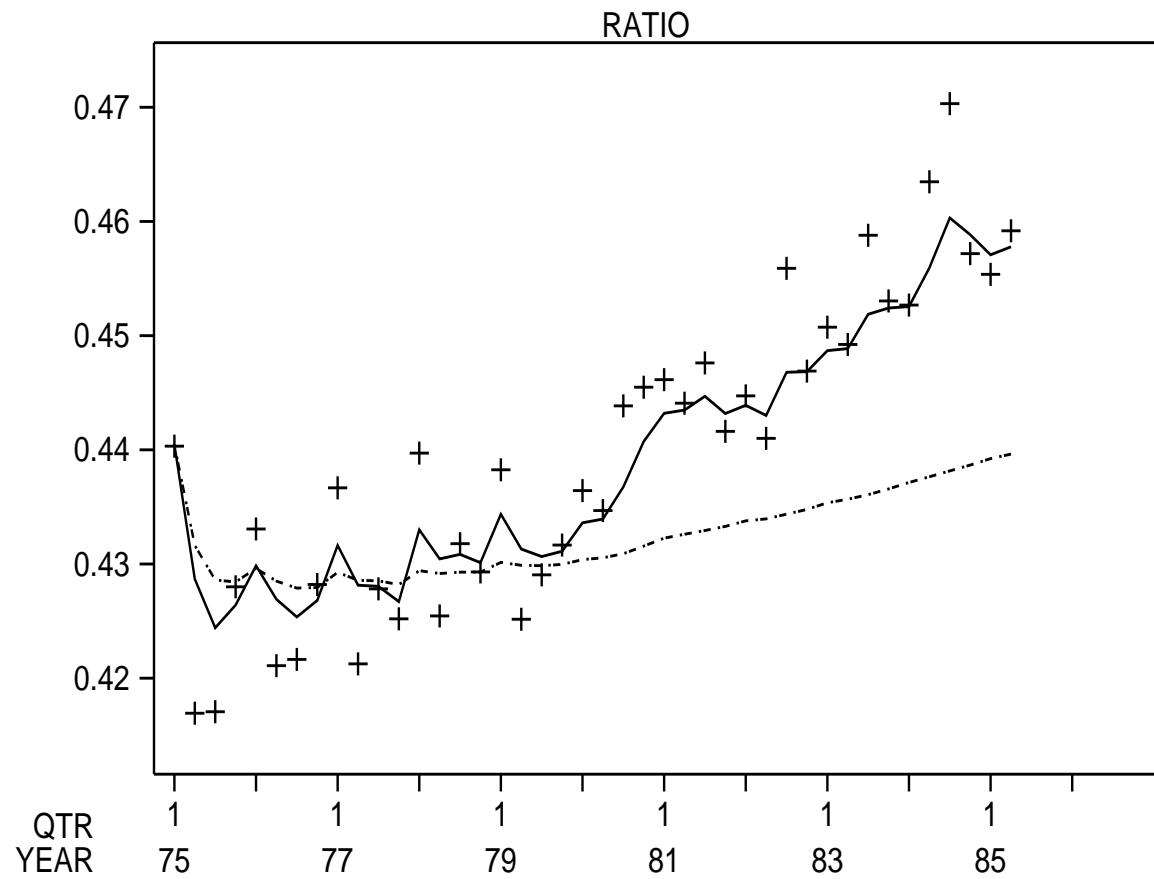
## Sales Data Example



## Sales Data Example



## Sales Data Example



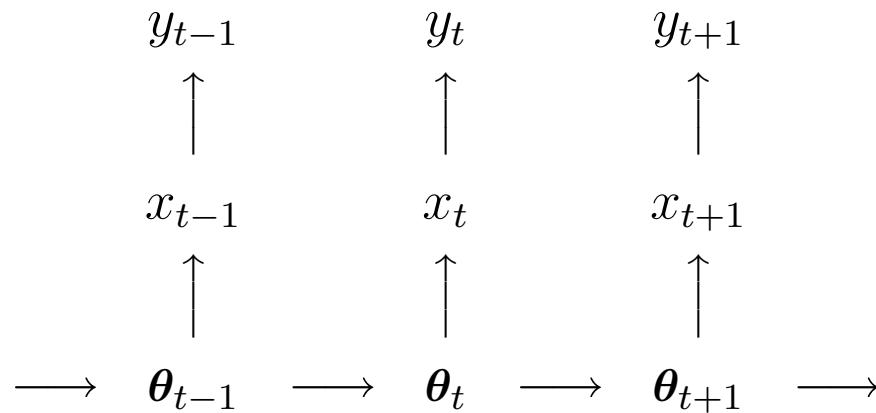
## Simple Regression Example

Relative to “static” model, dynamic regression delivers:

- improved estimation via adaptation for “local” regression parameters
- and increased (honest) uncertainty about regression parameters
- adaptability to (small) changes → improved point forecasts
- partitions variation: *parameter* vs *observation error*  
→ increased precision of stated forecasts  
i.e., improved prediction: point forecasts AND precision

## General Dynamic Linear Model

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \omega_t$$



- Sequential model definition : Markov evolution structure
- CI structure :  $\boldsymbol{\theta}_t$  sufficient for “future” at time  $t$

# Bayesian Forecasting

## *Key concepts:*

- Bayesian: modelling & learning is probabilistic
- Time-varying parameter models: often non-stationary
- Sequential view, sequential model definitions
  - encourages interaction, intervention

## *Statistical framework:*

- Forecasting: “What might happen?” and “What if?”
- Data processing and statistical learning from observations
- Updating of models and probabilistic summaries of belief
- Time series analysis ... *Retrospection*: “What happened?”

## Bayesian Machinery

- **Inferences** based on information  $D_t = \{(y_1, \dots, y_t), I_t\}$

Find and summarise

- $p(\boldsymbol{\theta}_t | D_t)$       and       $p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_t | D_t)$
- and update as  $t \rightarrow t + 1 \rightarrow \dots$

- **Forecasts:**

- $p(y_{t+1}, \dots, y_{t+k} | D_t)$
- and update as  $t \rightarrow t + 1 \rightarrow \dots$

- **Implementation & computations:**

- Linear/normal models: neat theory, Kalman filtering
- Extend to infer variance components, non-normal errors ..
  - \* need approximations, simulation methods, MCMC

## Commercial Applications

- Short term forecasting of consumer sales and demand
- Monitoring: stocks and inventories of consumer products
- Many items or sectors: Aggregation and multi-level models

### Standard models for commercial applications:

$$\begin{array}{ccccccccc} \text{Data} & = & \text{Trend} & + & \text{Seasonal} & + & \text{Regression} & + & \text{Error} \\ & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ y_t & = & x_{1t} & + & x_{2t} & + & x_{3t} & + & \nu_t \end{array}$$

## Models in Commercial Applications

Component signals  $x_{jt}$  follow individual dynamic models

- Trends: e.g., “locally linear” trend

$$x_{1t} = x_{1,t-1} + \beta_t + \partial x_{1t}, \quad \beta_t = \beta_{t-1} + \partial \beta_t$$

- Seasonals:

$$x_{2t} = \sum_j (a_{j,t} \cos(2 * \pi j / p) + b_{j,t} \sin(2 * \pi j / p))$$

where  $(a_{j,t}, b_{j,t})$  wander through time

**Key concept:** Model (De)Composition

- Modelling, prior specification, interventions: **component-wise**
- Posterior inference: detrending, deseasonalisation, etc

## Special Classes of DLMs

### Time Series DLMs – constant $\mathbf{F}, \mathbf{G}$

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'\boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \omega_t$$

- Includes all “standard” point-forecasting methods  
(exponential smoothing, variants, ... )
- Polynomial trend and seasonal components in commercial models
- Includes all practically useful ARIMA models

### *Multiple representations:*

$$\phi_t = \mathbf{E}\boldsymbol{\theta}_t \leftrightarrow \mathbf{G} \rightarrow \mathbf{E}\mathbf{G}\mathbf{E}^{-1}$$

**General class:** Time-varying  $\mathbf{F}_t, \mathbf{G}_t$

Includes non-stationary models, time-varying ARIMA models, etc., ...

## Simulation-Based Computation

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \omega_t$$

- Normal error models  $p(\nu_t), p(\omega_t)$
- Fixed time window  $t = 1, \dots, n$
- State vector set:  $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$
- **Require:** full posterior sample  $\boldsymbol{\Theta}^*$  from  $p(\boldsymbol{\Theta}|D_n)$

Available via “*Forward-filtering: Backward-sampling*” algorithm

Carter and Kohn (1994) *Biometrika*

Frühwirth-Schnatter (1994) *J Time Series Anal*

West and Harrison 1997

## FFBS Algorithm

### *Forward-filtering:*

- Standard normal/linear analysis: Kalman filter
- delivers normal  $p(\boldsymbol{\theta}_t|D_t)$  at each  $t = 1, \dots, n$

### *Backward-sampling:*

- at  $t = n$  : sample  $\boldsymbol{\theta}_n^*$  from  $p(\boldsymbol{\theta}_n|D_n)$
- for  $t = n-1, n-2, \dots, 1$  : sample  $\boldsymbol{\theta}_t^*$  from normal distribution  $p(\boldsymbol{\theta}_t|D_t, \boldsymbol{\theta}_{t+1}^*)$

Builds up  $\boldsymbol{\Theta}^*$  as  $\boldsymbol{\theta}_n^*, \boldsymbol{\theta}_{n-1}^*, \dots$

Exploits Markovian/CI model structure

## MCMC in DLMs

*DLM parameters:* e.g., constant model

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'\boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \omega_t$$

*Parameters:*

- Variances (variance matrices) of  $\nu_t, \omega_t$
- Elements of  $\mathbf{F}, \mathbf{G}$
- Indicators in normal mixture models for errors

*Add parameters to analysis: MCMC utilising FFBS*

## MCMC in DLMs

Parameters  $\Phi$  (may depend on sample size)

*Gibbs sampling:*  $p(\Theta, \Phi | D_n)$  iteratively resampled via

- Apply FFBS algorithm to draw  $\Theta^*$  from  $p(\Theta | \Phi^*, D_n)$
- Draw new value of  $\Phi^*$  from  $p(\Phi | \Theta^*, D_n)$
- Iterate

“Standard” Gibbs sampling: MCMC

May need “creativity” in sampling  $\Phi^*$ : Metropolis-Hastings, etc

Often “easy”: as in Autoregressive DLM

## MCMC: Autoregressive Model Example

*Data:*  $y_t = x_t + \nu_t$

*State AR(d)*  $x_t = \sum_{j=1}^d \phi_j x_{t-j} + \epsilon_t$

DLM for  $x_t$ :  $x_t = (1, 0, \dots, 0)\mathbf{x}_t, \quad \mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \omega_t$

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{d-1} & \phi_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \quad \omega_t = \begin{pmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

- **Parameters:**  $\{\phi_1, \dots, \phi_d; V(\nu_t), V(\epsilon_t)\}$
- Conditional posteriors standard: linear regression parameters

## Models in Scientific Applications

Historical interest in biomedical monitoring (change-points), engineering applications (control, tracking), environmental monitoring, ...

### Goals:

- Exploratory “discovery” of interpretable latent processes
- Nonstationary time series: “hidden” quasi-periodicities
- Changes over time at different time scales
- Time:frequency structure (in time domain)

### State-space models:

- Stationary and/or nonstationary, time-varying parameters
- General decomposition theory for state space-space models
- DLM autoregressions and time-varying autoregressions

## Autoregressive DLM

*Latent AR( $d$ ) process:*  $x_t = \sum_{j=1}^d \phi_j x_{t-j} + \epsilon_t$

*Time series:*  $y_t = x_{0t} + x_t + \nu_t$  with trend, etc in  $x_{0t}$

DLM for  $x_t$ :  $x_t = (1, 0, \dots, 0)\mathbf{x}_t$ ,  $\mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \omega_t$

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{d-1} & \phi_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \quad \omega_t = \begin{pmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

- $x_t$  latent, unobserved
- $\mathbf{G} = \mathbf{G}(\phi)$  with  $\phi = (\phi_1, \dots, \phi_d)'$  to be estimated

## Time Series Decomposition

*Eigenstructure:*  $\mathbf{G} = \mathbf{E}^{-1} \mathbf{A} \mathbf{E}$

Reparametrise to “diagonal” model:  $\mathbf{x}_t \rightarrow \mathbf{E}\mathbf{x}_t$

Transforms to

$$x_t = \sum_{j=1}^{d_z} z_{t,j} + \sum_{j=1}^{d_a} a_{t,j}$$

- $z$  terms: one for each pair of complex conjugate eigenvalues
- $a$  terms: one for each real eigenvalue
- underlying *latent* processes  $z_{t,j}$  and  $a_{t,j}$  follow “simple” models
  - $a_{t,j}$  is AR(1) process - short-term correlations
  - $z_{t,j}$  is *quasi-periodic* ARMA(2,1) - noisy sine wave with randomly time-varying amplitude & phase, fixed frequency

## Time-varying Autoregression

**TV-AR( $d$ ) model:**  $x_t = \sum_{j=1}^d \phi_{t,j} x_{t-j} + \epsilon_t$

- *AR parameter:*  $\phi_t = (\phi_{t,1}, \dots, \phi_{t,d})'$  “wanders” through time:

$$\phi_t = \phi_{t-1} + \partial\phi_t$$

- stochastic “shocks”  $\partial\phi_t$
- *Innovations:*  $\epsilon_t \sim N(0, \sigma_t^2)$  – time-varying variance  $\sigma_t^2$

### Flexible representations:

- non-stationary process, time-varying spectral properties
- latent component structure
- other evolutions for  $\phi_t$  (e.g., Godsill *et al* on speech processing)

## DLM Form of TV-AR( $d$ ):

$$x_t = (1, 0, \dots, 0)' \mathbf{x}_t$$

$$\mathbf{x}_t = \mathbf{G}(\phi_t) \mathbf{x}_{t-1} + (\epsilon_t, 0, \dots, 0)'$$

$$\phi_t = \phi_{t-1} + \partial\phi_t$$

with

$$\mathbf{G}(\phi_t) = \begin{pmatrix} \phi_{t,1} & \phi_{t,2} & \cdots & \phi_{t,d-1} & \phi_{t,d} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

*Analysis:* Posterior distributions for  $\{\phi_t, \sigma_t : \forall t\}$

*Component models:*  $y_t = x_{0t} + x_{1t} + \nu_t$

→ infer latent TV-AR processes too:  $\{x_{0t}, x_t : \forall t\}$

## TVAR Decomposition

$$x_t = \sum_{j=1}^{d_z} z_{t,j} + \sum_{j=1}^{d_a} a_{t,j}$$

- $z$  terms: one for each pair of complex conjugate eigenvalues
- $a$  terms: one for each real eigenvalue
- underlying *latent* processes:
  - $a_{t,j}$  is **TV-AR(1)** process - short-term correlations  
+ **time-varying correlation**
  - $z_{t,j}$  is **TV-ARMA(2,1)** - noisy sine wave with  
randomly time-varying amplitude & phase  
+ **time-varying frequency** ( $2\pi/\text{wavelength}$ )

Number of complex/real eigenvalues can vary over time too!

## Time:Frequency Analysis

- Fit flexible (high order?) AR or TV-AR models
- Estimate latent components and their frequencies, amplitudes over time
- Time domain representation of spectral structure
- Often, some  $z_{t,j}$  physically meaningful, some (high frequency) represent noise, model approximation
- $a_{t,j}$  – noise, model approximation and and (possibly) low frequency “trend”

## Paleoclimatology Example

- deep ocean cores: relative abundance of  $\delta^{18}\text{O}$
- $\delta^{18}\text{O} \downarrow$  as global temperatures  $\uparrow$  (smaller ice mass)
- *reverse sign*: higher recent global temperatures
- “well known” periodicities: earth orbital dynamics  $\rightarrow$  impact on solar insolation – Milankovitch; Shackleton *et al* since 1976
  - eccentricity*: 95-120 kyear
  - obliquity*: 40-42 kyear
  - precession*: 19-25 kyear (1 or 2?)

## Oxygen Isotope Data

- Form of time variation in individual cycles ?
- Timing/nature of onset of “ice-age” cycle  $\leftrightarrow$  eccentricity component  $\sim 1000$  kyears ago ?
- *Time scale: errors, interpolation, ... measurement, sampling error, etc*

**Models:** High order TV-AR,  $p = 20$ , plus smooth trend (outliers?)

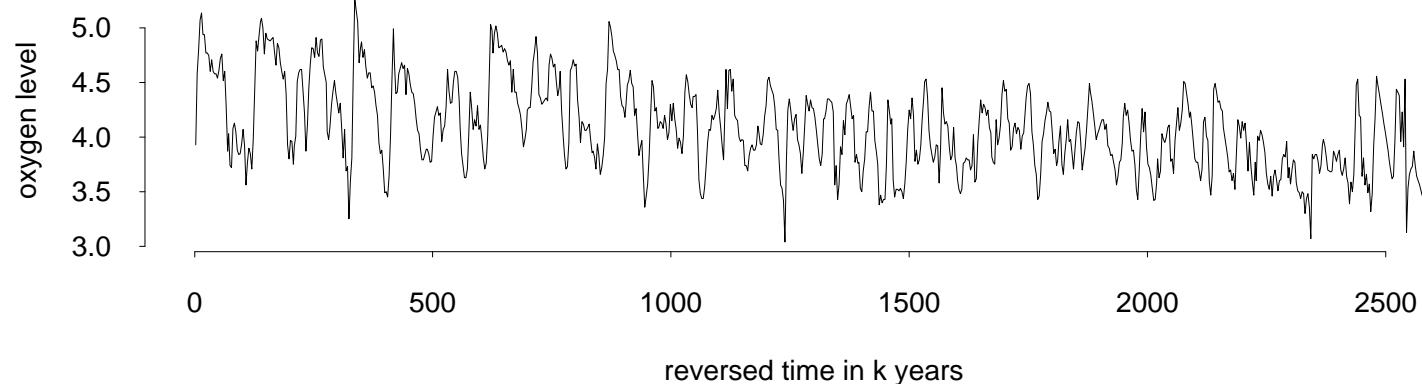
Variance components estimated: changing AR parameters

**Decomposition:** Posterior mean of  $x_t, \phi_{t,j}$  at each  $t$

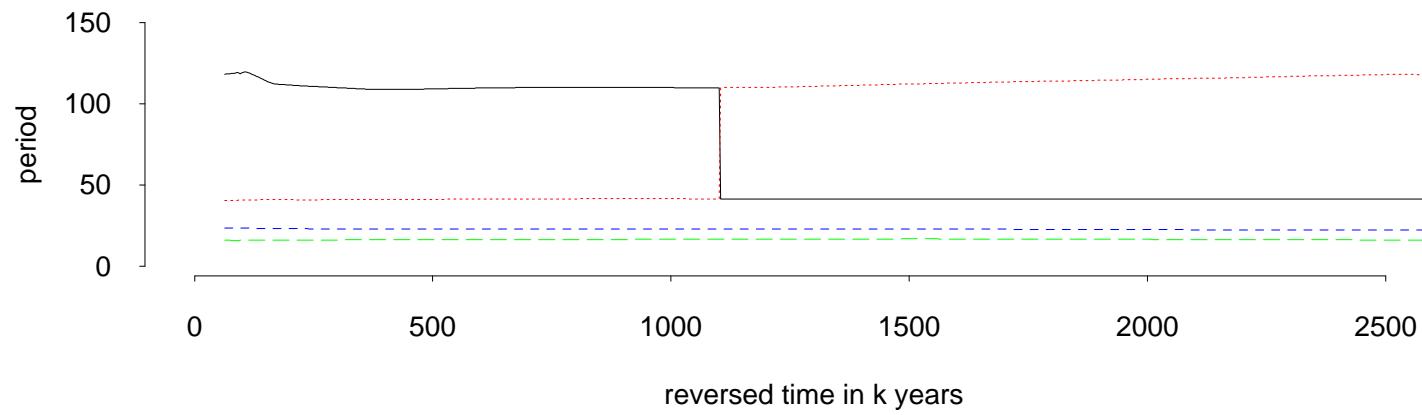
4 dominant quasi-periodic components: order by estimated *amplitude* (innovation variance)

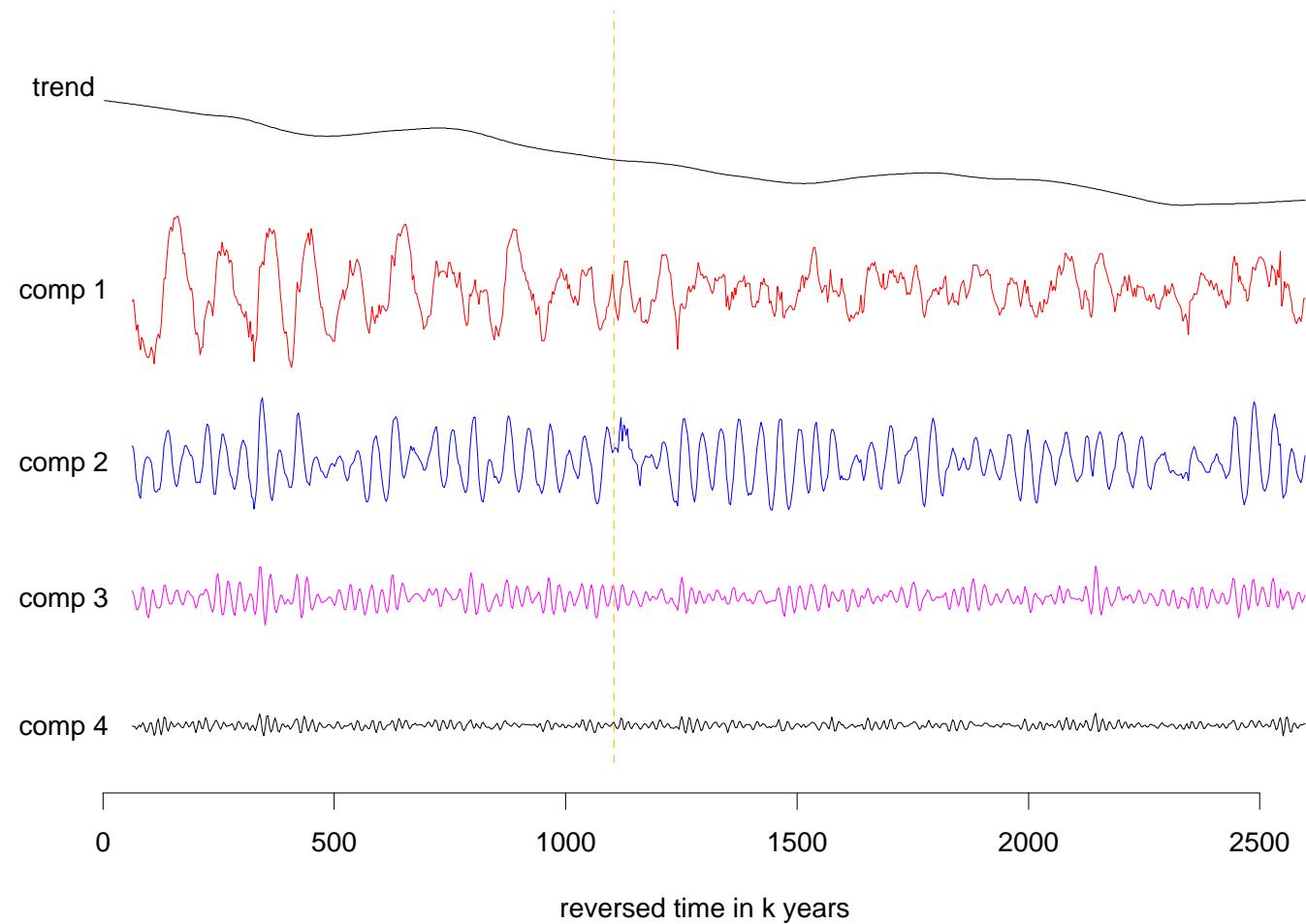
**Others:** residual structure &/or contaminations

### oxygen isotope series



### trajectories of time-varying periods of components





## Oxygen

- Wavelengths vary only modestly
- Estimated periods/wavelengths consistent with geological determinations
  - .... 108–120, peak 110
  - .... 40.8–41.6, peak 41.5
  - .... 22.2–23, peak 22.8
- “Switch” due to order of estimated amplitude
  - Geological interpretation? Structural climate change  $\sim 1.1\text{m yrs}^2$ ?

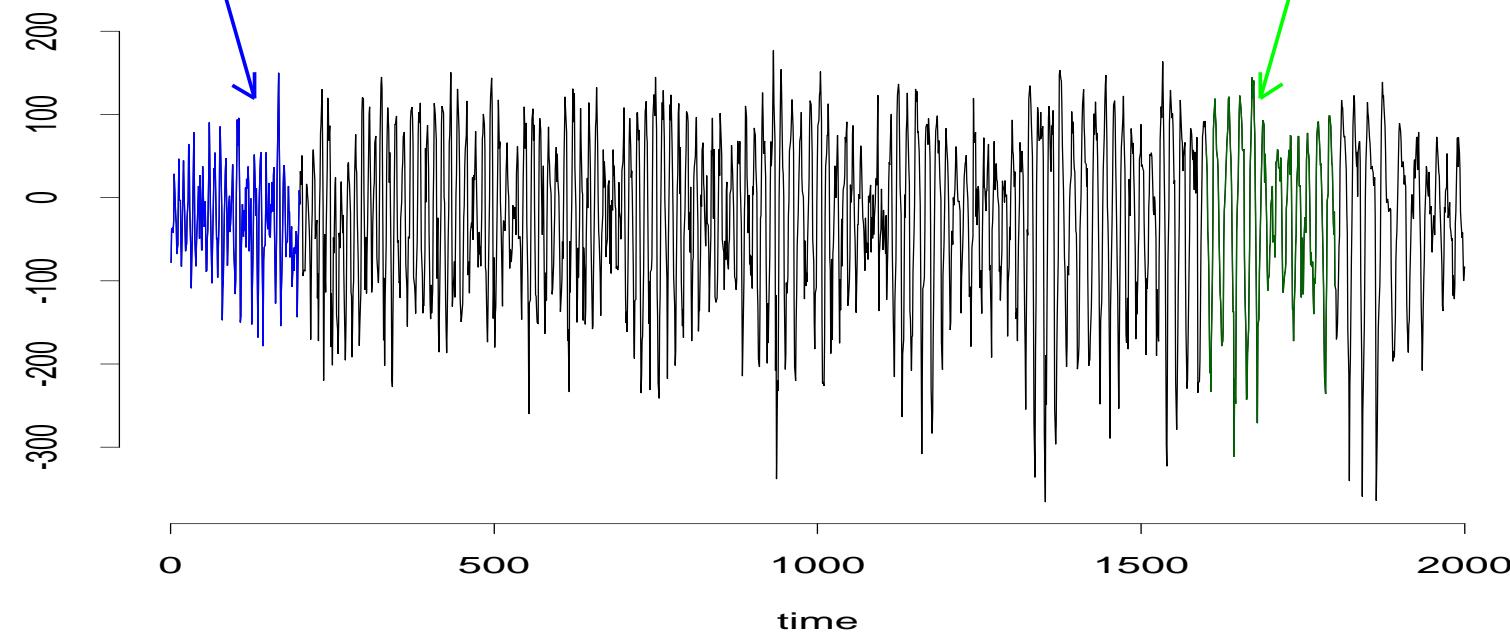
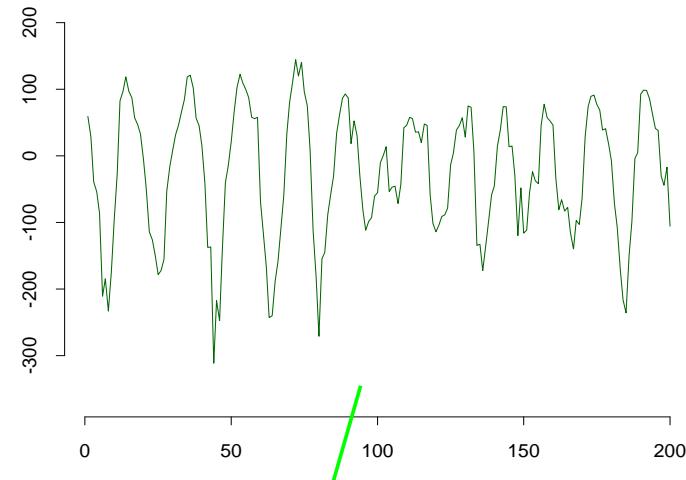
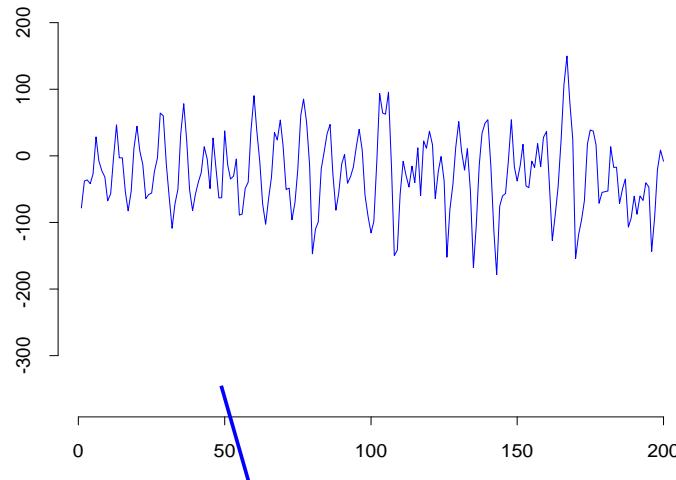
## Example: EEG Study

- Clinical uses of electroconvulsive therapy
- Measure seizure treatment outcomes via long, multiple EEG (electroencephalogram) series – electrical potential fluctuations on scalp
- Many multiple series: one seizure – 19 channels, 256/sec

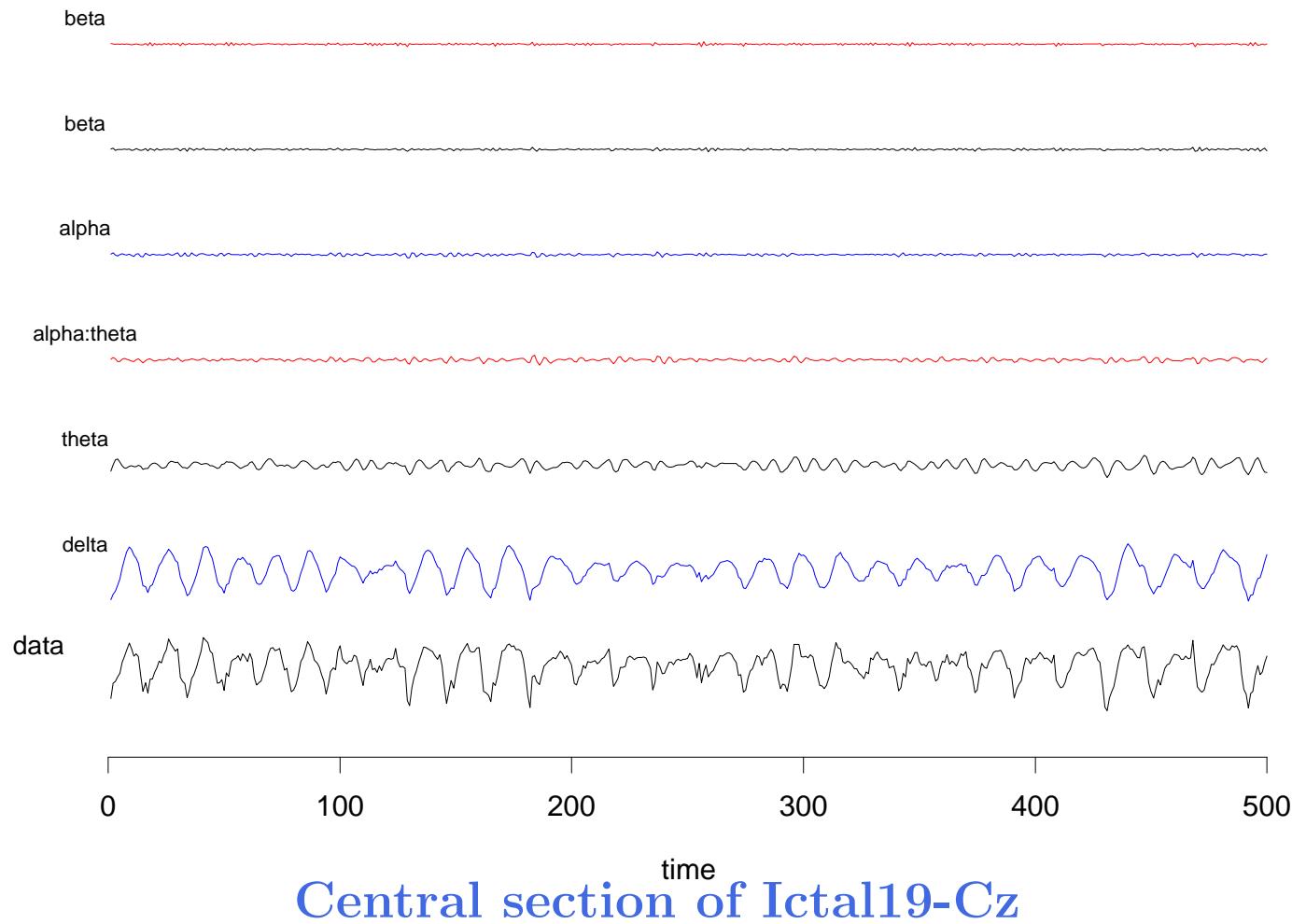
### Models to:

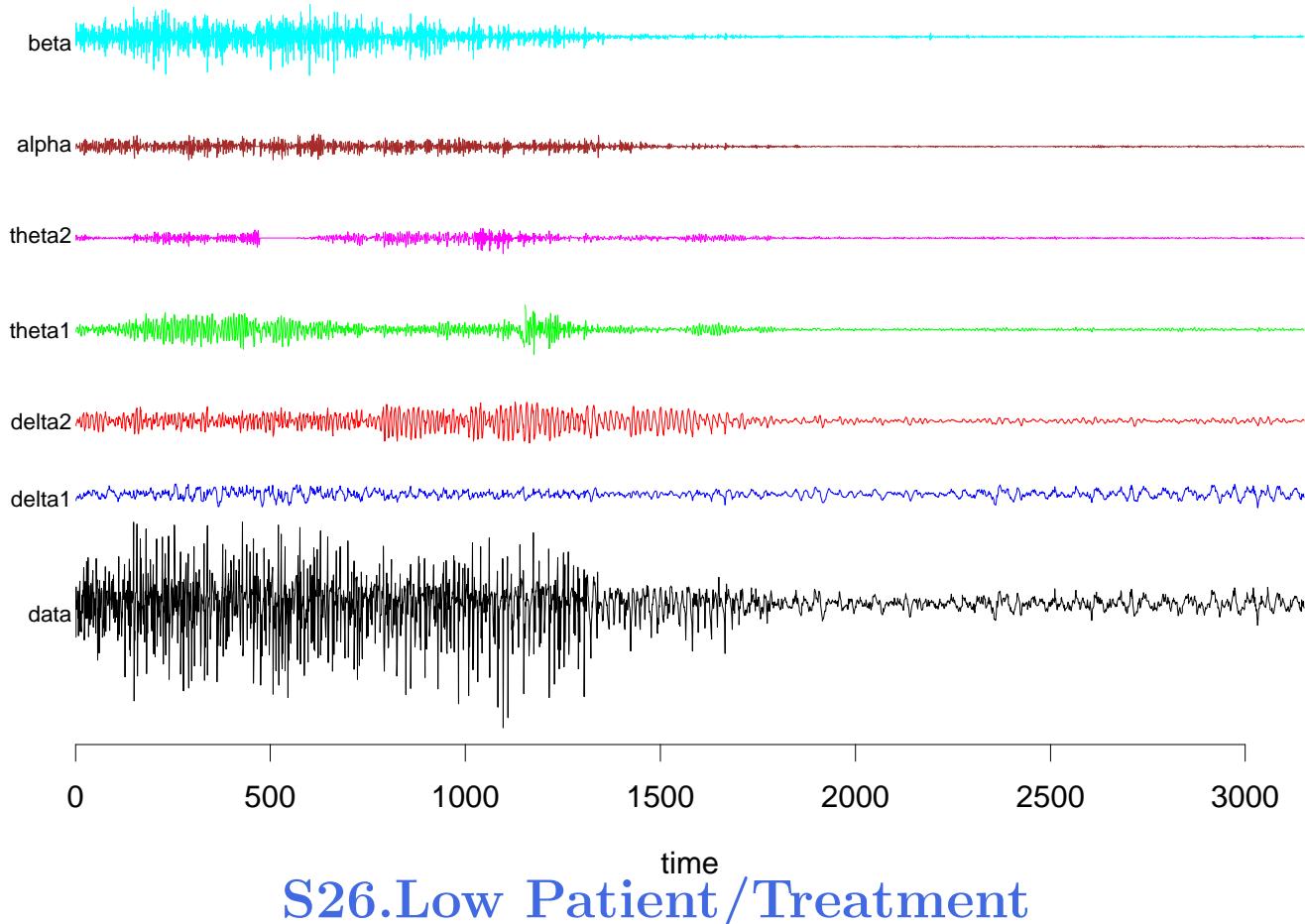
- Characterise seizure “waveforms” ... time varying amplitudes at ranges of frequencies (alpha waves, etc)
- Superimposed on “normal” waveform, noise, ...
- Identify/extract latent components: infer **seizure effects**
- Spatial connectivities: related multiple series

## Why TVAR Models?



## EEG Decomposition Examples



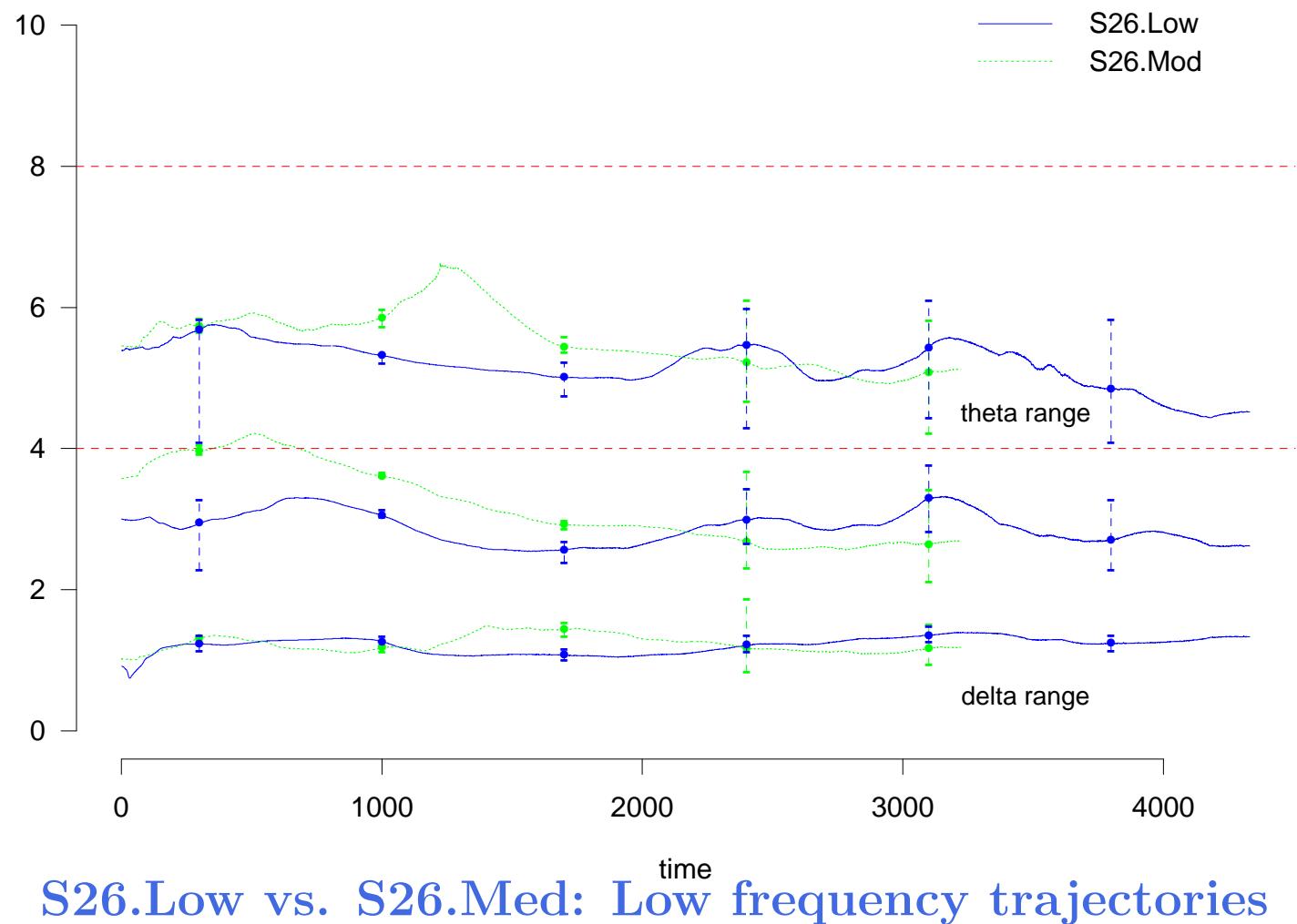


## EEG Treatment Comparison

**Two EEG series on one individual:** S26.Low cf S26.Mod

- Repeat seizures with varying **ECT treatment**
- TVAR(20) with time-varying  $\sigma_t^2$
- Evident high frequency structure and “spiky” traces  
→ higher order models
- Several frequency bands influenced by seizure – several components
- **Identification issues**

## EEG Comparisons



S26.Low vs. S26.Mod:  $\frac{\text{time}}{\text{frequency}}$  trajectories

# Sequential Simulation Analysis

*Time t :*

- States  $\boldsymbol{\Theta}_t = \{\theta_{t-k}, \dots, \theta_t\}$  of interest
- Summarised via set of posterior samples:  $p(\boldsymbol{\Theta}_t | D_t)$

*Time t + 1 :*

- Observe  $y_t \sim p(y_t | \boldsymbol{\theta}_t)$
- Require updated summary, posterior samples:  $p(\boldsymbol{\Theta}_{t+1} | D_t)$

*Issues:*

- Expanding/Changing state space and dimension
- Simulation-based summaries: discrete approximations
- Inference on parameters as well as state vectors
- New data may “conflict” with prior/predictions

## Particle Filtering

**Key Goal:** sequentially update posteriors

$$\cdots \rightarrow p(\boldsymbol{\theta}_t | D_t) \rightarrow p(\boldsymbol{\theta}_{t+1} | D_{t+1}) \rightarrow \cdots$$

**Numerical Approximations (points/weights):**

$$\{\boldsymbol{\theta}_t^{(j)}, \omega_t^{(j)} : j = 1, \dots, N_t\}$$

**Theoretical update:**

$$p(\boldsymbol{\theta}_{t+1} | D_{t+1}) \propto p(\mathbf{y}_{t+1} | \boldsymbol{\theta}_{t+1}) p(\boldsymbol{\theta}_{t+1} | D_t)$$

**MC approximation to “prior”:**

$$p(\boldsymbol{\theta}_{t+1} | D_t) \approx \sum_{k=1}^{N_t} \omega_t^{(k)} p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t^{(k)})$$

- Mixture prior: sample and accept/reject ideas natural

## Example: Auxilliary Particle Filter

APF state update from  $t \rightarrow t + 1$ :

- for each  $k$ ,
  - “estimates”  $\boldsymbol{\mu}_{t+1}^{(k)} = E(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t^{(k)})$
  - and weights  $g_{t+1}^{(k)} \propto \omega_t^{(k)} p(\mathbf{y}_{t+1} | \boldsymbol{\mu}_{t+1}^{(k)})$
- sample (aux) indicators  $j$  with probs  $g_{t+1}^{(j)}$
- **time  $t + 1$  samples:**  $\boldsymbol{\theta}_{t+1}^{(j)} \sim p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t^{(j)})$
- **and weights:**

$$\omega_{t+1}^{(j)} = \frac{p(\mathbf{y}_{t+1} | \boldsymbol{\theta}_{t+1}^{(j)})}{p(\mathbf{y}_{t+1} | \boldsymbol{\mu}_{t+1}^{(j)})}$$

## Multivariate Models in Finance

(*Quintana et al invited talk, Valencia VII*)

- Futures markets, exchange rates, portfolio selection
- Multiple time series: time-varying covariance patterns
- Econometric/dynamic regressions/hierarchical models
- Latent factors in hierarchical, dynamic models
- Common time-varying structure in multiple series
- Bayesian multivariate stochastic volatility

$$\mathbf{y}_t = (y_{1t}, \dots, y_{pt})'$$

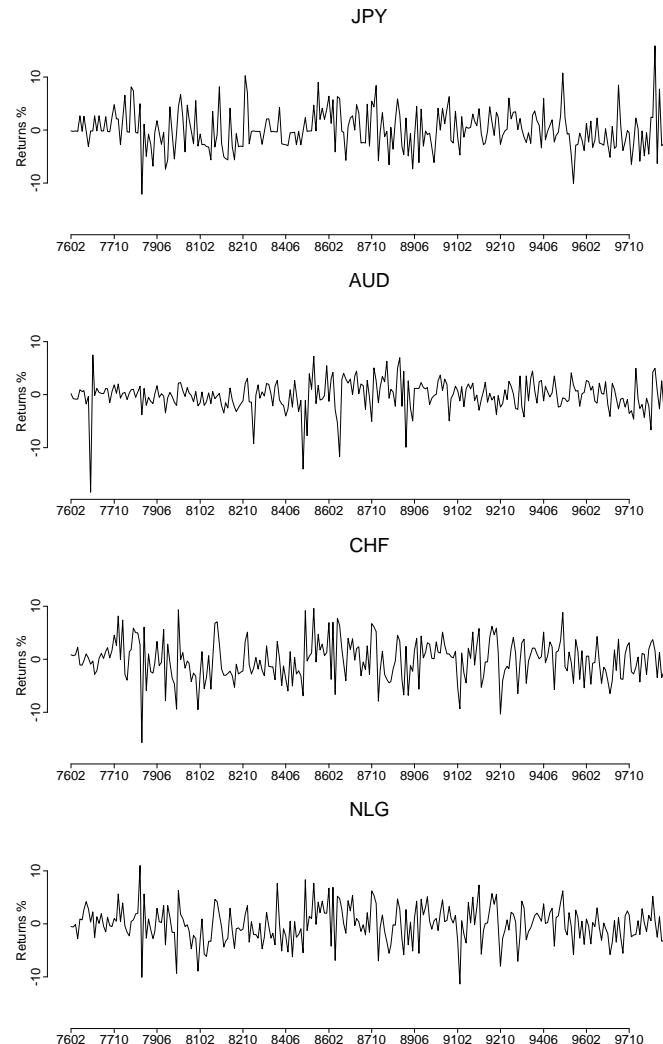
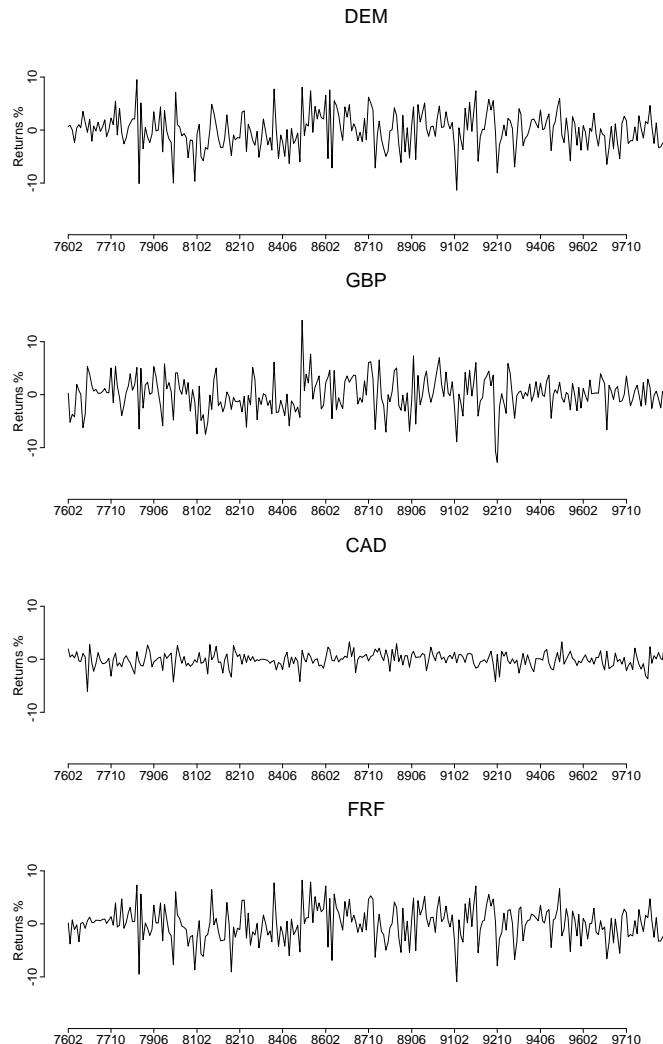
e.g.,  $y_{it}$  is  $p$ -vector of *returns* on investment  $i$   
(exchange rate futures, etc.)

# Dynamic Factor Models

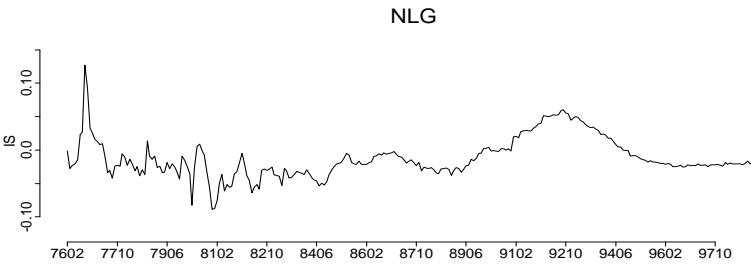
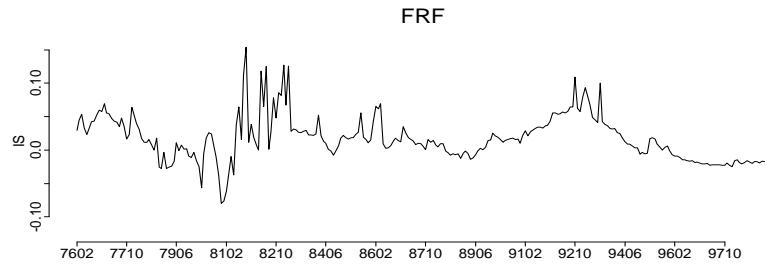
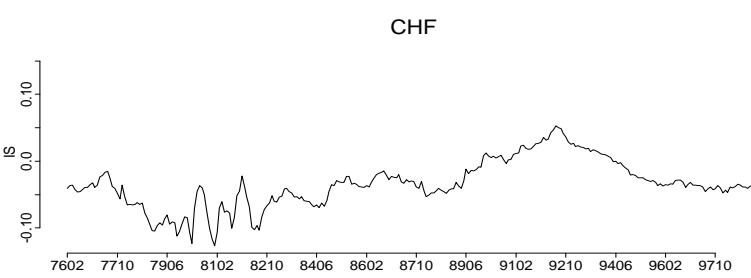
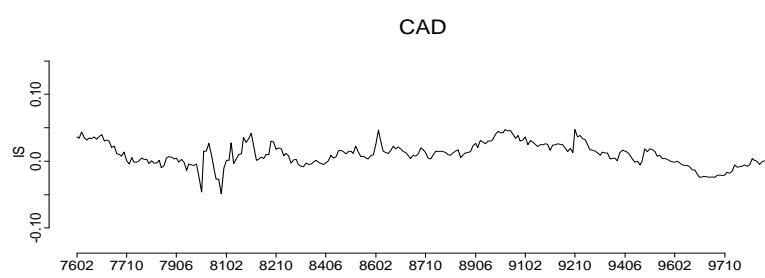
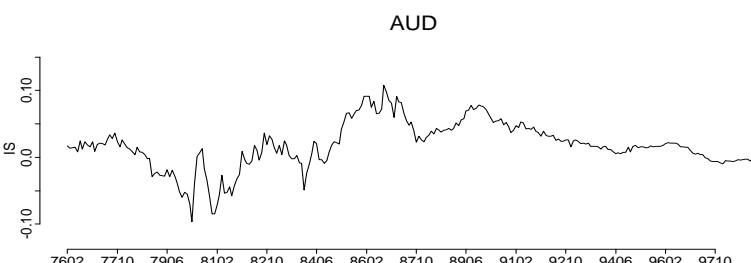
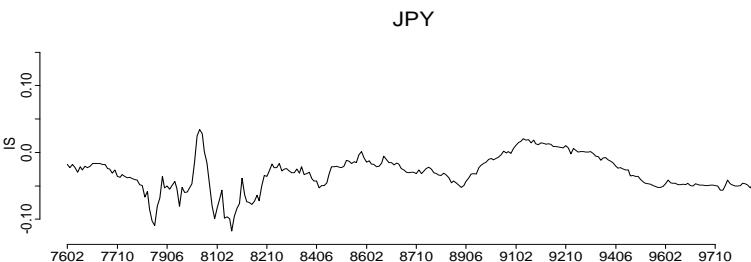
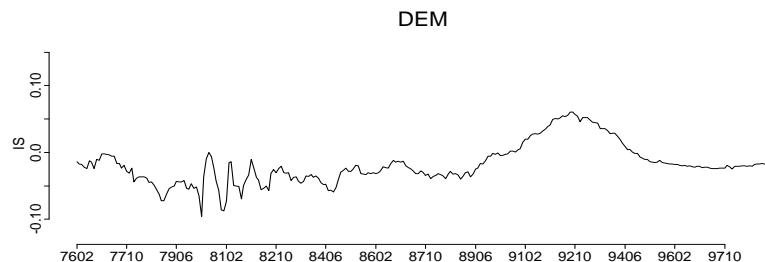
## Exchange rate modelling for dynamic asset allocation

- Monthly (daily) currency exchange rates
- Dynamic regression/econometric predictors
- Residual structure and residual stochastic volatility
  - Time-varying variances and covariances
  - Dynamic factor models
- Dynamic asset allocation & risk management: portfolio studies
- Bayesian analysis: model fitting, sequential analysis, forecasting, decision analysis

## Excess Returns on Exchange Rates (monthly)



## Short Interest Rates (monthly)



## Stochastic Volatility Factor Models

$$\mathbf{y}_t = \text{dynamic regression}_t + \text{residual}_t \quad - \quad \text{residual}_t \sim N(\mathbf{0}, \Sigma_t)$$

residual <sub>t</sub>	=	$\mathbf{X}_t \mathbf{f}_t + \mathbf{e}_t$
$\mathbf{f}_t$	~	$N(\mathbf{f}_t   \mathbf{0}, \mathbf{F}_t)$
$\mathbf{e}_t$	~	$N(\mathbf{e}_t   \mathbf{0}, \mathbf{E}_t)$

$\mathbf{f}_t$  ...  $k$ -vector of latent factors

$$\mathbf{F}_t = \text{diag}(\exp(\lambda_{1,t}), \dots, \exp(\lambda_{k,t}))$$

$\mathbf{e}_t$  ...  $q$ -vector of “idiosyncracies”

$$\mathbf{E}_t = \text{diag}(\exp(\lambda_{k+1,t}), \dots, \exp(\lambda_{k+q,t}))$$

$$\Sigma_t = \mathbf{X}_t \mathbf{F}_t \mathbf{X}'_t + \mathbf{E}_t$$

Factor and idiosyncratic latent log volatilities:  $\lambda_t = (\lambda_{1,t}, \dots, \lambda_{k+q,t})'$

## Dynamic Factor Model: Factor Loadings Structure

- Constant factor loadings matrix  $\mathbf{X}_t = \mathbf{X}$   
– or “slowly varying” over time –
- Identification constraints on  $\mathbf{X}$ :

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ x_{2,1} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ x_{k,1} & x_{k,2} & x_{k,3} & \cdots & 1 \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \cdots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{q,1} & x_{q,2} & x_{q,3} & \cdots & x_{q,k} \end{pmatrix}$$

- Series order *defines* interpretation of factors

# Stochastic Volatility Models: Factors & Idiosyncracies

## Multivariate SV models

- vector AR(1) model for latent log volatilities  $\lambda_t$
- volatility persistence: AR(1) coefficients  $\Phi$  (diagonal)
- marginal volatility levels: time-varying  $\mu_t$
- dependent innovations: shocks to volatilities related across series  
global/sector “effects” represented through correlations

$$\lambda_t = \mu_t + \Phi(\lambda_{t-1} - \mu_{t-1}) + \omega_t$$

$$\omega_t \sim N(\mathbf{0}, \mathbf{U})$$

$$\mu_t \sim \text{random walk}$$

## Dynamic Regressions and Shrinkage Models

- several predictors (e.g., interest rates)
- regression coefficients time-varying
- shrinkage models: coefficients related across series

e.g., sensitivity to short-term interest rates “similar” across currencies

Series  $j$ , time  $t$  :

$$\beta_{jt} = \phi_{jt}\beta_{j,t-1} + (1 - \phi_{jt})\gamma_t + \text{innovation}_{jt}$$

- global or sector “average”  $\gamma_t$
- time-varying degrees of “shrinkage”  $\phi_{jt}$
- multiple series, several predictors: state-space model formulations

## Model Fitting & Analysis: MCMC

Inference based on a fixed sample over  $t = 1, \dots, T$

- Bayesian analysis via posterior simulations
- Monte Carlo samples from *joint posterior* for
  - model parameters, dynamic regression parameters, AND
  - *latent processes: factors & volatilities*

$$\{ \mathbf{f}_t, \boldsymbol{\lambda}_t : t = 1, \dots, T \}$$

- examples of highly structured, hierarchical models with many latent variables and parameters
- custom MCMC built of standard “modules”
- simple MC evaluation of forecast/predictive distributions

## Model Fitting & Analysis: Sequential Analyses

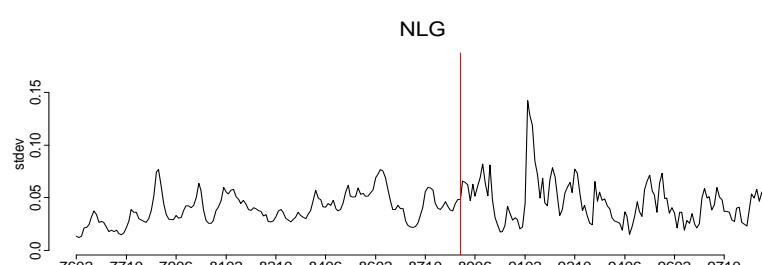
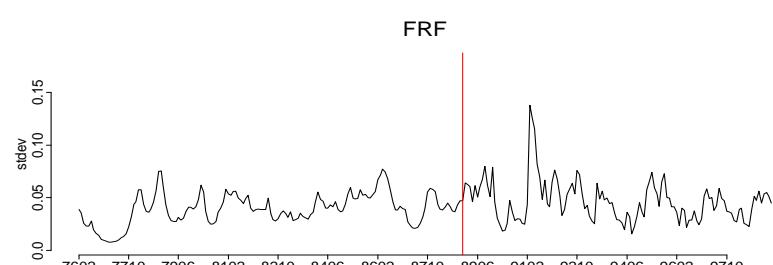
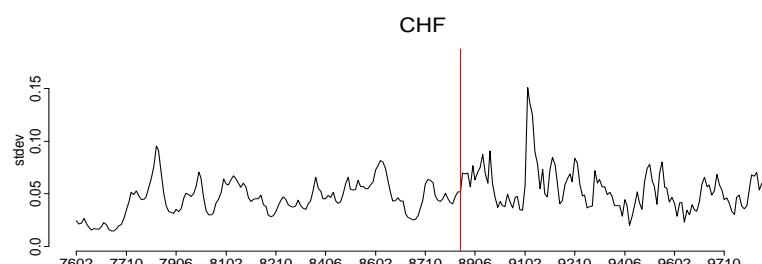
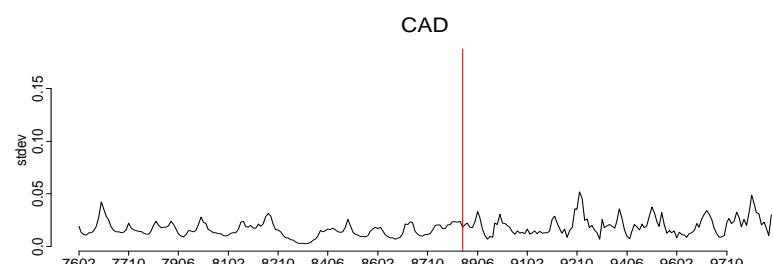
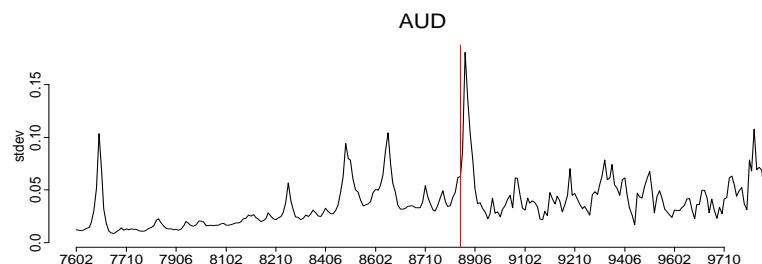
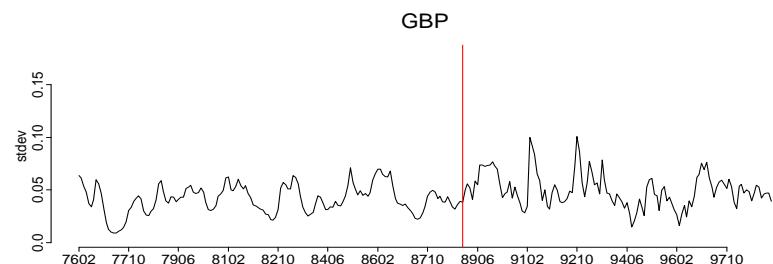
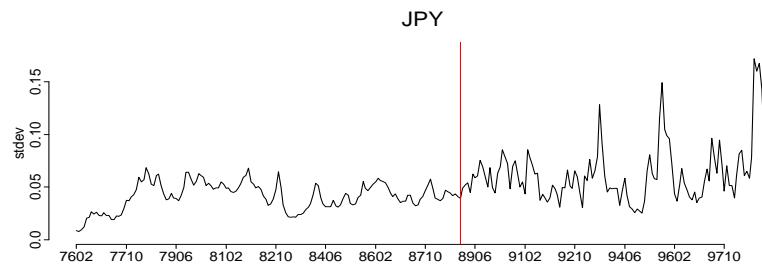
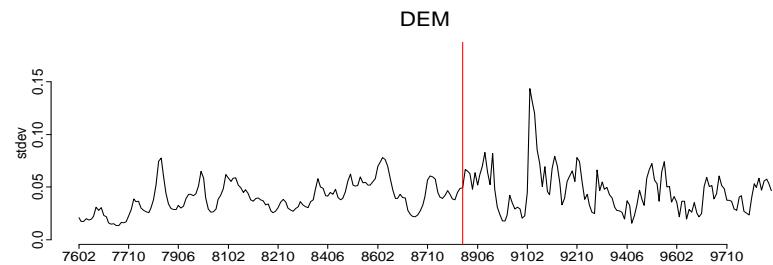
Sequential particle filtering over  $t = T + 1, T + 2, \dots$

- “particulate” approximation to posterior distributions
- **prior**: cloud of particles, associated importance weights
- **posterior**: sampling/importance resampling to *revise* posterior samples

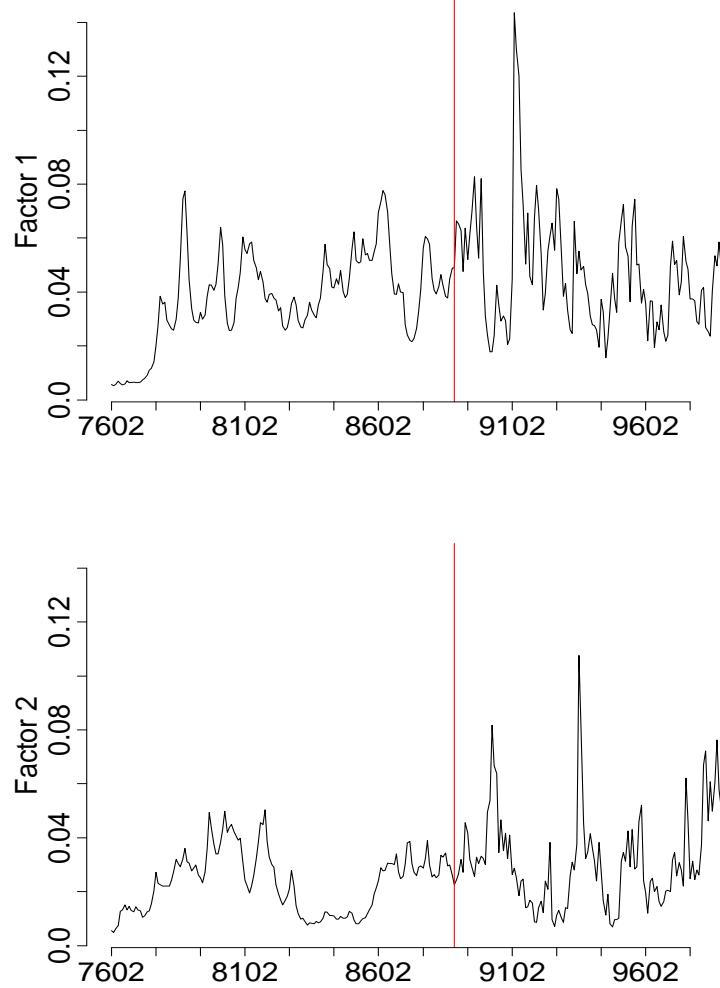
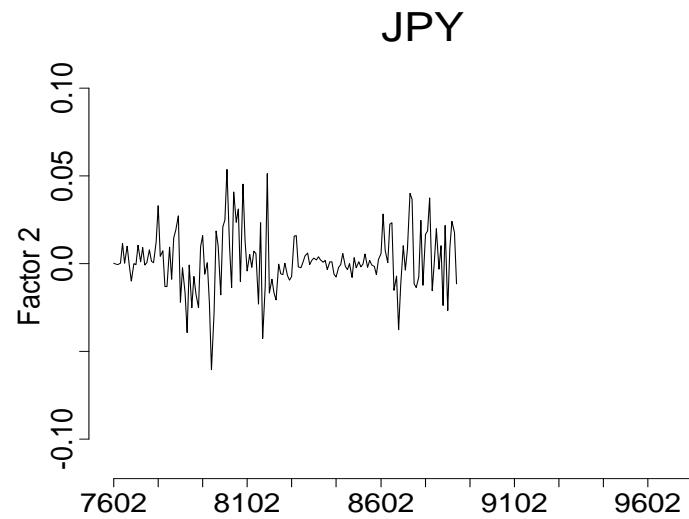
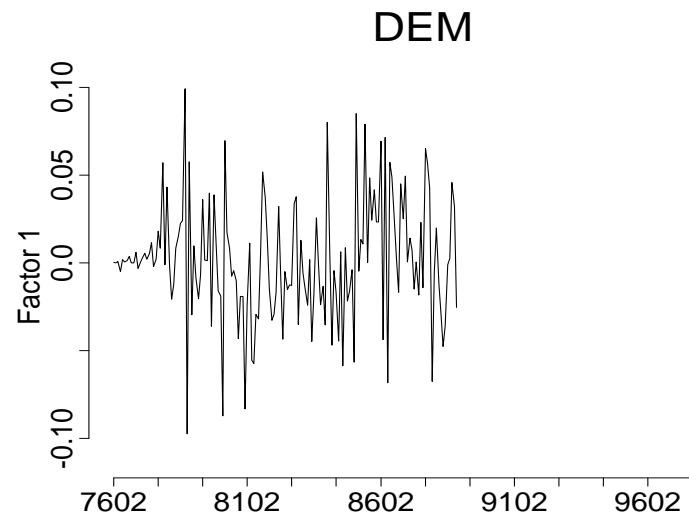
$$\cdots \rightarrow p(\cdot | D_{t-1}) \rightarrow p(\cdot | D_t) \rightarrow \cdots$$

- sample “regeneration” by local smoothing of particulate prior
- time-varying latent variables *and* model parameters

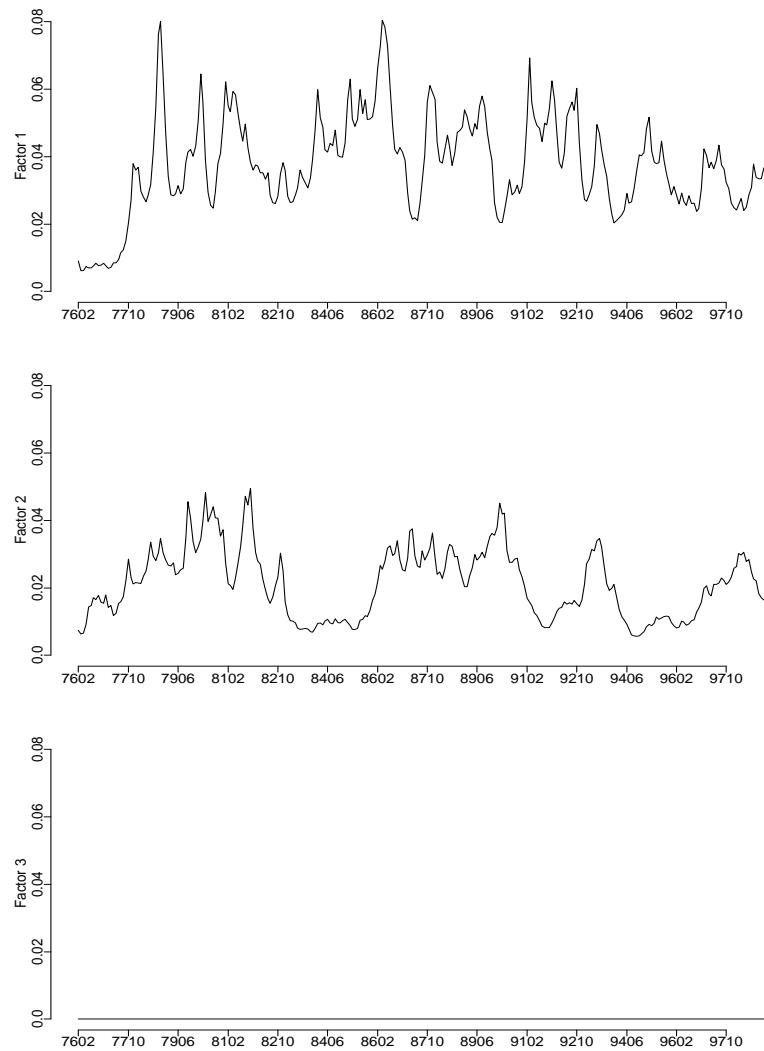
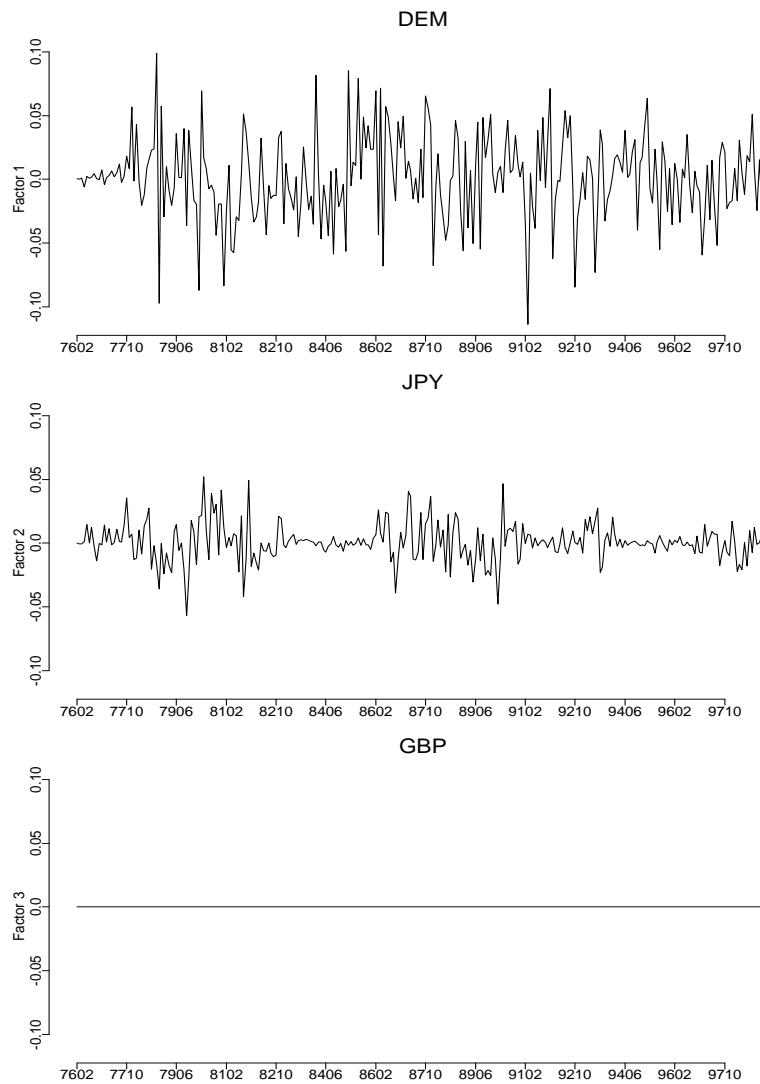
## Volatilities (SD) of Currency Returns (monthly)



## 2 Latent Factors and their Volatilities (SD) (monthly)



## 3 Latent Factors and their Volatilities (SD) (monthly)



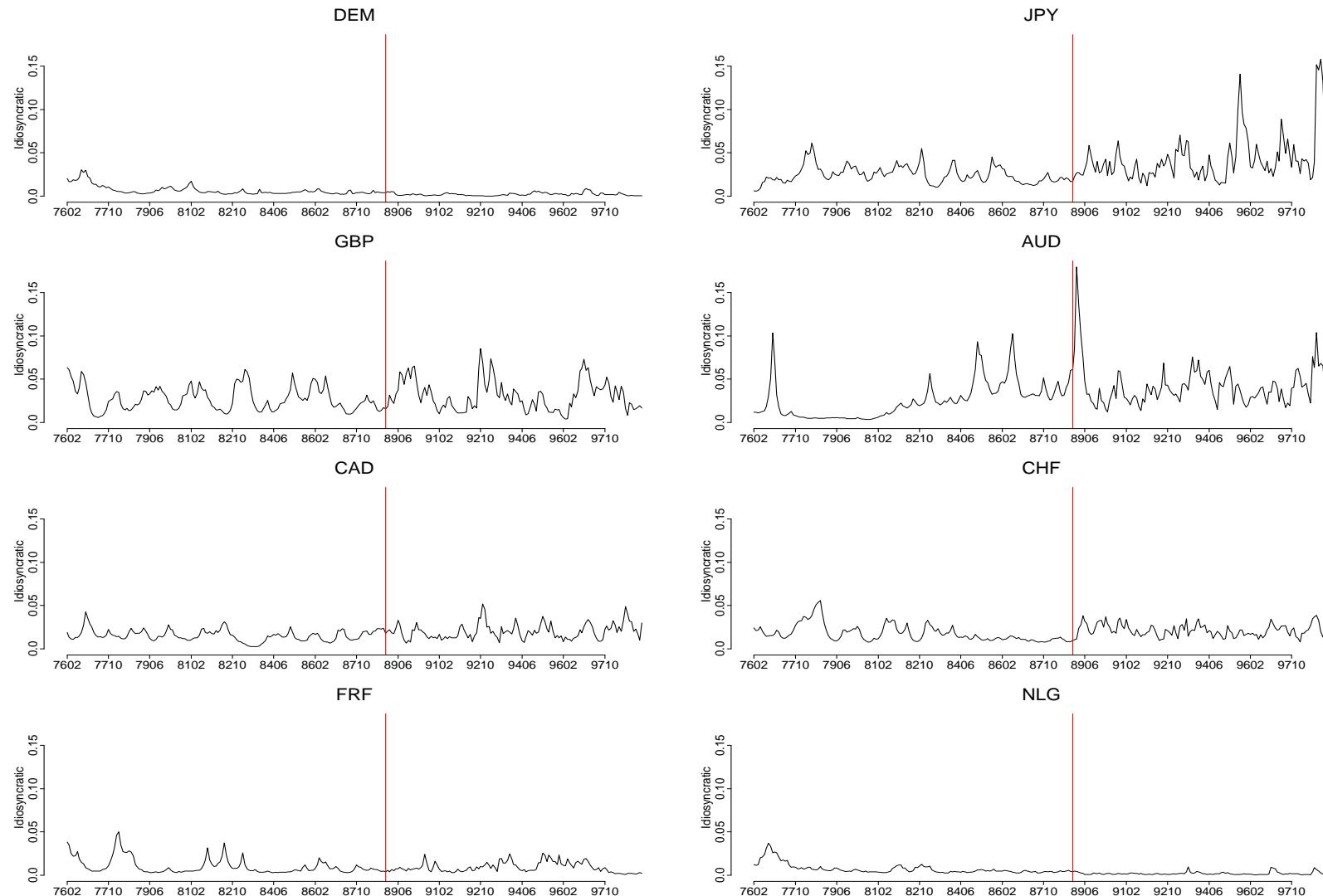
## Factor Loading Matrix X (monthly)

	Factor 1	Factor 2
DEM	1.00 (0.00)	0.00 (0.00)
JPY	0.55 (0.06)	1.00 (0.00)
GBP	0.68 (0.05)	0.48 (0.14)
AUD	0.23 (0.03)	0.35 (0.07)
CAD	0.04 (0.02)	0.03 (0.08)
CHF	1.04 (0.03)	0.23 (0.07)
FRF	0.96 (0.01)	-0.01 (0.02)
ESP	0.99 (0.01)	-0.01 (0.01)

## Factor Loading Matrix X (monthly)

	Factor 1	Factor 2
DEM	1	.
FRF	1	.
ESP	1	.
CHF	1	0.2
GBP	0.7	0.5
JPY	0.5	1
AUD	0.2	0.3
CAD	.	.

## Idiosyncratic Volatilities (SD) (monthly)



## Financial Time Series

- Models/forecasts feed into portfolio decision management
  - *Quintana et al invited talk, Valencia VII*
- Live/Operational assessments of dynamic factor models
- Improved sequential particle filtering: parameters
- Time-varying loadings matrices  $\mathbf{X}_t$ , parameters
- Uncertainty about number of factors

## Bayesian Time Series, Currently

### *Applied aspects:*

- Financial modelling and forecasting
- Natural/engineering sciences: signal processing
- Spatial time series: epidemiology, environment, ecology

### *Models and methods:*

- Highly structured multiple time series
- Spatial time series
- Computational methods: Sequential simulation methods

## Links & Materials

- Books and papers: [www.isds.duke.edu/~mw](http://www.isds.duke.edu/~mw)
  - *Copious references to broad literatures*
- 1997 Tutorial – extensive and historical – at website
- Teaching materials, notes, from past time series courses
- Software: [www.isds.duke.edu/~mw](http://www.isds.duke.edu/~mw)
  - *TVAR (matlab, fortran), AR component models, BATS, links*
- Encyclopedia of Statistical Sciences (1998): *Bayesian Forecasting* (eds: S. Kotz, C.B. Read, and D.L. Banks), Wiley.
- 1997 2nd Edition: West & Harrison/Springer book  
*Bayesian Forecasting & Dynamic Models*
- Aguilar, Prado, Huerta & West (1999), Valencia 6 invited paper

## Key Recent (since 1995) & Current Authors

– many/most are here! –

Omar Aguilar (Lehman Bros)

Sid Chib (Washington Univ/St Louis)

Simon Godsill & Co (Cambridge)

Genshiro Kitagawa (ISM Tokyo)

Jane Liu (UBS NY)

Viridiana Lourdes (ITAM Mexico)

Michael Pitt (Warwick)

Raquel Prado (Santa Cruz)

Peter Rossi (Chicago)

Chris Carter (Hong Kong)

Sylvia Frühwirth-Schnatter (Vienna)

Gabriel Huerta (Univ New Mexico)

Robert Kohn (Sydney)

Hedibert Lopes (Rio)

Giovanni Petris (Univ Arkansas)

Nicolas Polson (Chicago)

José M Quintana (Nikko NY)

Neil Shephard (Oxford)

**... and, of course, ...**

## Valencia VII Invited Papers

- Quintana, Lourdes, Aguilar & Liu
  - Global gambling: multivariate financial time series
- Davy & Godsill
  - Signal processing, latent structure, MCMC

**... plus a number of contributed talks and posters**

## BAYESIAN TIME SERIES ANALYSIS AND FORECASTING

Mike West

Institute of Statistics & Decision Sciences  
Duke University

*AISTATS-97 Tutorial*

Fort Lauderdale, January 4th 1997

- Introductory comments
- Dynamic linear models (state space models)
  - Sequential context, Bayesian framework
  - Standard classes of models (commercial applications)
  - Sequential monitoring and intervention
- Varieties of mixture models in time series
- Multivariate time series
- Models and methods in physical science applications
  - Time series decompositions
  - More multivariate models and latent structure
- Computation: Simulation methodology, MCMC
- Other topics: Current research frontiers

## KEY SOURCES AND ENTRY POINTS:

- West and Harrison (1989, 1997 February 2nd Edn.)  
*Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York.
- Pole, West and Harrison (1994) *Applied Bayesian Forecasting and Time Series Analysis.* Chapman-Hall, New York.
- West (to appear) Bayesian Forecasting. In *Encyclopedia of Statistical Sciences*, Wiley. Includes many references as well as overview of the field (*circa mid-1995*)
- Above article and many others at <http://www.stat.duke.edu> under *Discussion Papers*

## 5 DECADES OF TIME SERIES:

- 1946-1956
  - 1946 M Bartlett, on autocorrelations (RSS symposium)
  - 1946 N Levinson, on Wiener filtering and prediction
  - 1949 N Wiener, book on time series in engineering
- 1956-1966
  - 1957 C C Holt, on exponential moving averages
  - 1957-1966 P J Harrison, exponential smoothing methods
  - 1960-1963 R E Kalman & R S Bucy, linear filtering
  - 1963 P Whittle, *Prediction and Regulation*
  - 1965 J W Cooley & J W Tukey, FFT for spectral analysis

- 1966-1976
  - P J Harrison & C Stevens, Bayesian Forecasting
  - G E P Box & G Jenkins, ARMA Modelling
  - H Akaike, Model selection and state space models
- 1976-1986
  - “Big picture” Bayesian framework:  
monitoring and intervention; non-Gaussian and non-linear  
models; mixture models; multiple time series; ...
  - R B Litterman, Bayesian vector AR modelling
- 1986-1996
  - Wide ranging developments in applications
  - Simulation revolution: Numerical methods, MCMC

## STANDARD DYNAMIC MODELS

*Dynamic Linear Models*

*Linear State Space Models*

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

- signal  $x_t$ , state vector  $\boldsymbol{\theta}_t = (\theta_{t1}, \dots, \theta_{td})'$
- regression vector  $\mathbf{F}_t$  and state matrix  $\mathbf{G}_t$
- zero mean measurement errors  $\nu_t$  and state innovations  $\boldsymbol{\omega}_t$ 
  - often zero-mean and normally distributed

Examples:

\* “Slowly varying” level observed with noise:

$$y_t = x_t + \nu_t \quad x_t = x_{t-1} + \omega_t$$

\* Dynamic linear regression:

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \omega_t$$

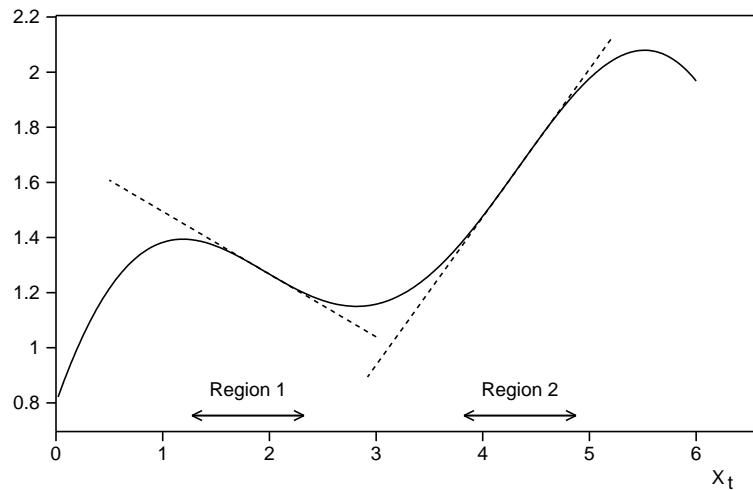
Error models for  $\nu_t, \omega_t$

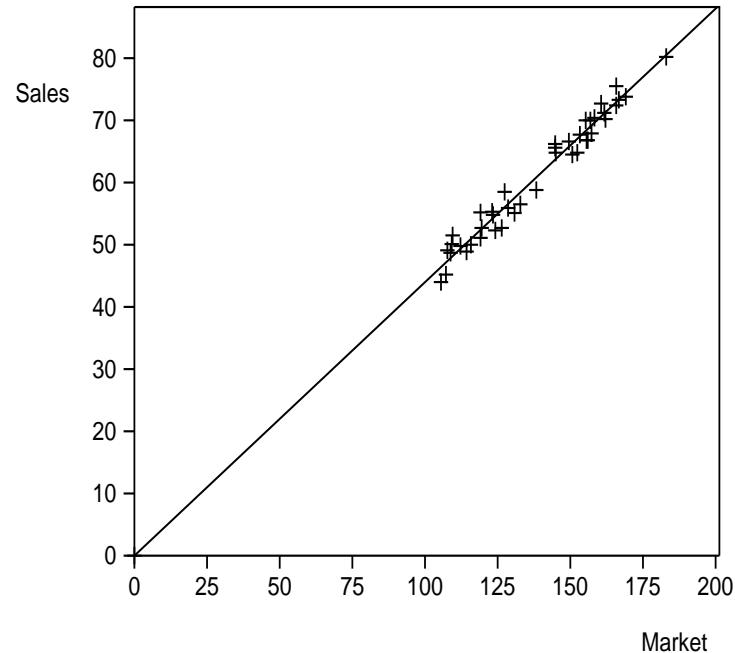
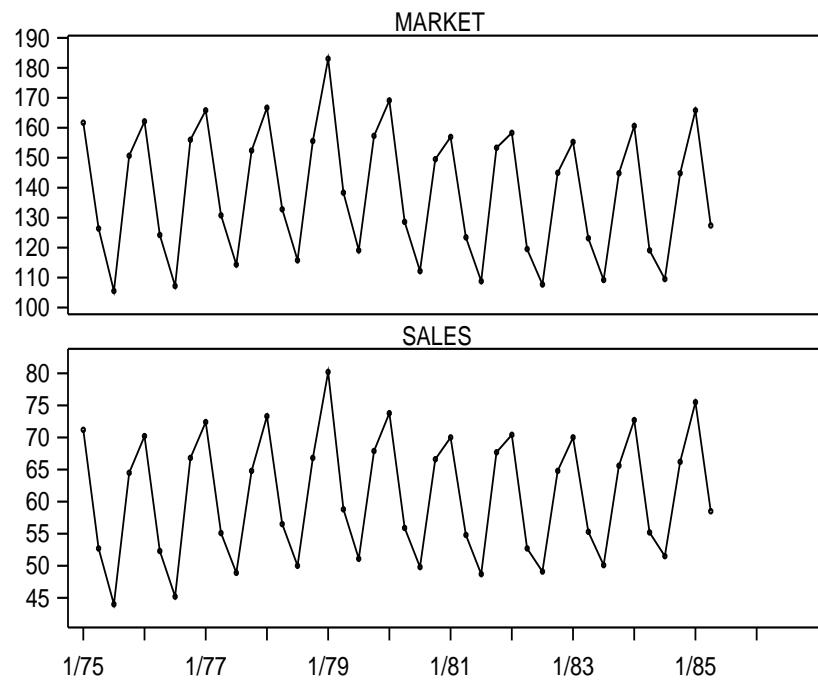
- normal distributions
- mixtures of normals: outliers  
abrupt “structural” changes

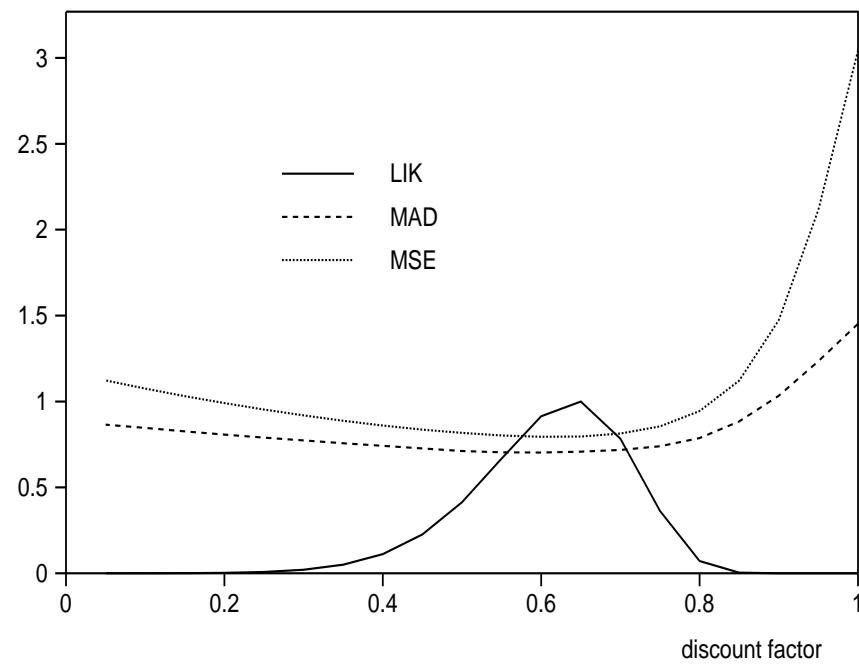
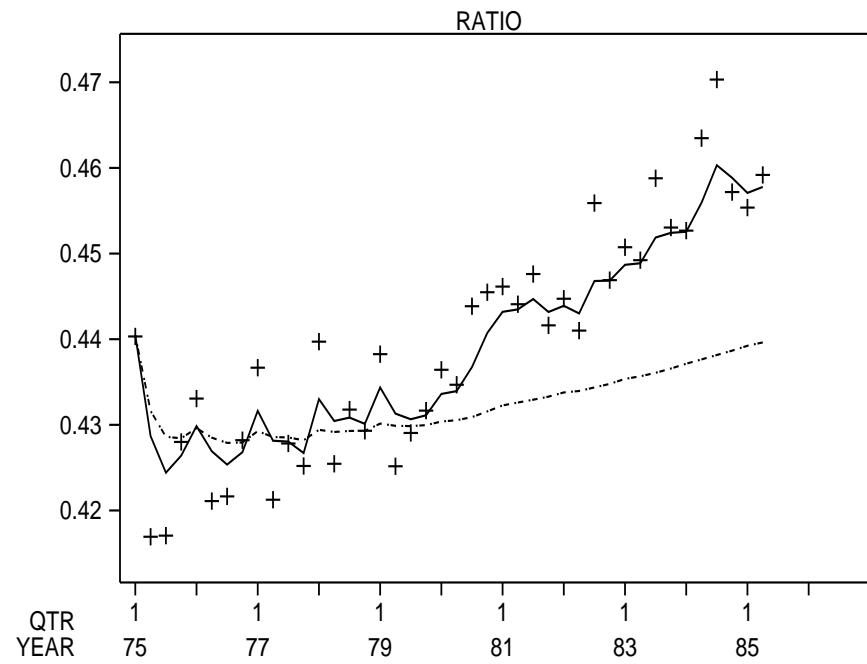
Simple regression example:

$$y_t = x_t + \nu_t \quad x_t = a_t + b_t X_t$$

$\mathbf{F}_t = (1, X_t)'$  and  $\boldsymbol{\theta}_t = (a_t, b_t)'$  “wanders” through time







*Simple regression example:*

Relative to “static” model, dynamic regression delivers

- improved estimation via adaptation for “local” regression parameters
- and increased (honest) uncertainty about regression parameters
- adaptability to (small) changes → improved point forecasts
- partitions variation: *parameter* vs *observation error*  
→ increased precision of stated forecasts  
i.e., improved prediction: point forecasts AND precision

*Regression example: Sales data analyses*

*Dynamic regression:*

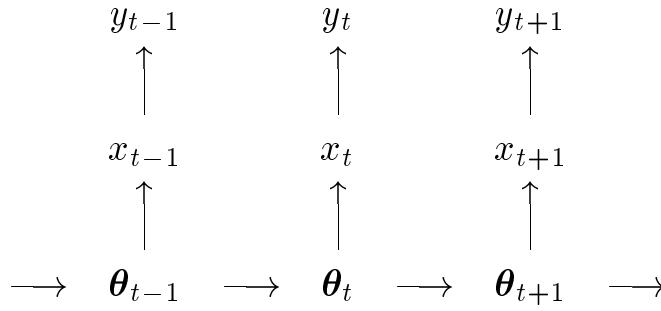
t	Forecast error	Forecast SD	Observation SD
40	-0.45	1.07	0.77
41	-0.58	1.10	0.77
42	0.27	0.94	0.76

*Static regression:*

t	Forecast error	Forecast SD	Observation SD
40	2.76	1.62	1.63
41	2.77	1.66	1.66
42	2.54	1.68	1.69

General class of DLMs:

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$



- Sequential model definition : Markov evolution structure
- CI structure :  $\boldsymbol{\theta}_t$  sufficient for “future” at time  $t$

Key Concepts:

- Bayesian: modelling & learning is probabilistic
- Time-varying parameter models: often non-stationary
- Sequential view, sequential model definitions
  - encourages interaction, intervention

Statistical Framework:

- Forecasting: “What might happen?” and “What if?”
- Data processing and statistical learning from observations
- Updating of models and probabilistic summaries of belief
- Time series analysis ... Retrospection: “What happened?”

- *Inferences* based on information  $D_t = \{(y_1, \dots, y_t), I_t\}$

Find and summarise

- $p(\boldsymbol{\theta}_t | D_t)$  and  $p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_t | D_t)$
- and update as  $t \rightarrow t + 1 \rightarrow \dots$

- *Forecasts*:

- $p(y_{t+1}, \dots, y_{t+k} | D_t)$
- and update as  $t \rightarrow t + 1 \rightarrow \dots$

- *Implementation & Computations*:

- Linear/normal models: neat theory, Kalman filtering
- Extend to infer variance components, non-normal errors ..  
\* need approximations, simulation methods, MCMC

*Kalman Filter and all that:*

DLM: Normal/linear structure plus conditional independencies

$p(\boldsymbol{\theta}_t | D_t)$  normal with moments  $(m_t, M_t)$

*Kalman Filter*:  $(m_t, M_t) = f(m_{t-1}, M_{t-1}, y_t)$

*Other things*:

- Complete probabilistic framework
- Present (updating), Past (filtering), Future (forecasting)
- *Open models*: Intervention
- General framework: Non-linear/non-normal models

## COMMERCIAL APPLICATIONS

- Short term forecasting of consumer sales and demand
- Monitoring: stocks and inventories of consumer products
- Inventory and quality control
- Many items or sectors: Aggregation and multi-level models

### *Simple Model:*

$$y_t = x_t + \nu_t \quad x_t = x_{t-1} + \omega_t$$

- *smoothing*: estimate signal/level series  $x_t$
- *monitoring and control*: detect and adapt to “big” changes
- *intervention*: changes in model, priors, “on-line”
- error models for  $\nu_t, \omega_t$ 
  - normal distributions: related to exponential smoothing
  - mixtures of normals: outliers and abrupt level changes

### *Broader Context:*

Decision analysis → control schemes, “cusums”, etc.

“Standard” Models for Commercial Applications:

$$\begin{array}{ccccccc} \text{Data} & = & \text{Trend} & + & \text{Seasonal} & + & \text{Regression} & + & \text{Error} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ y_t & = & x_{1t} & + & x_{2t} & + & x_{3t} & + & \nu_t \end{array}$$

- component signals  $x_{jt}$  follow individual dynamic models
  - time-varying trend: “local polynomial”
  - time-varying seasonal factors or sin/cosine coefficients
  - time-varying regression parameters
- combine to give overall dynamic linear model for  $y_t$

*TRENDS:* e.g., “locally linear” trend

$$x_{1t} = x_{1,t-1} + \beta_t + \partial x_{1t}, \quad \beta_t = \beta_{t-1} + \partial \beta_t$$

*SEASONALS:*

$$x_{2t} = \sum_j (a_t \cos(2 * \pi t / p) + b_t \sin(2 * \pi t / p))$$

where  $(a_t, b_t)$  wander through time

Various other (equivalent) representations

## Key Concept: MODEL (DE)COMPOSITION

- prior modelling: component-wise
- intervention: changes to components
- posterior inference: detrending, deseasonalisation, etc
- compare: typical pre-model detrending, deseasonalisation, etc
  
- overlay monitoring, intervention, adaptive mixture modelling and other techniques and methods for “abrupt” changes in component parameters

## MODEL SUPERPOSITION: DLMs in “block” form

$$y_t = \sum_{i=1}^h x_{it} + \nu_t$$

$$(\forall i) : \quad x_{it} = \mathbf{F}'_{it} \boldsymbol{\theta}_{it}, \quad \boldsymbol{\theta}_{it} = \mathbf{G}_{it} \boldsymbol{\theta}_{i,t-1} + \boldsymbol{\omega}_{it}$$

MODELLING CHANGE: each  $V(\boldsymbol{\omega}_{it})$  defined by single (or few)  
*discount factors*

- measures of “rates of decay of information”
- parsimonious
- interprets traditional *ad-hoc* smoothing methods
- sensitivity to choices, inference on discount factors

**SPECIAL CLASS: TIME SERIES DLMs – constant  $\mathbf{F}_i, \mathbf{G}_i$** 

e.g., Trend and seasonal components in commercial models

- Includes all “standard” point-forecasting methods  
(exponential smoothing, variants, ... )
- Includes all practically useful ARIMA models
  - in terms of equivalent point forecasts
  - under limiting assumptions (e.g., no interventions)
  - see also state-space autoregressions to come later

**GENERAL CLASS:**

Includes non-stationary models, time-varying ARIMA models, etc.,

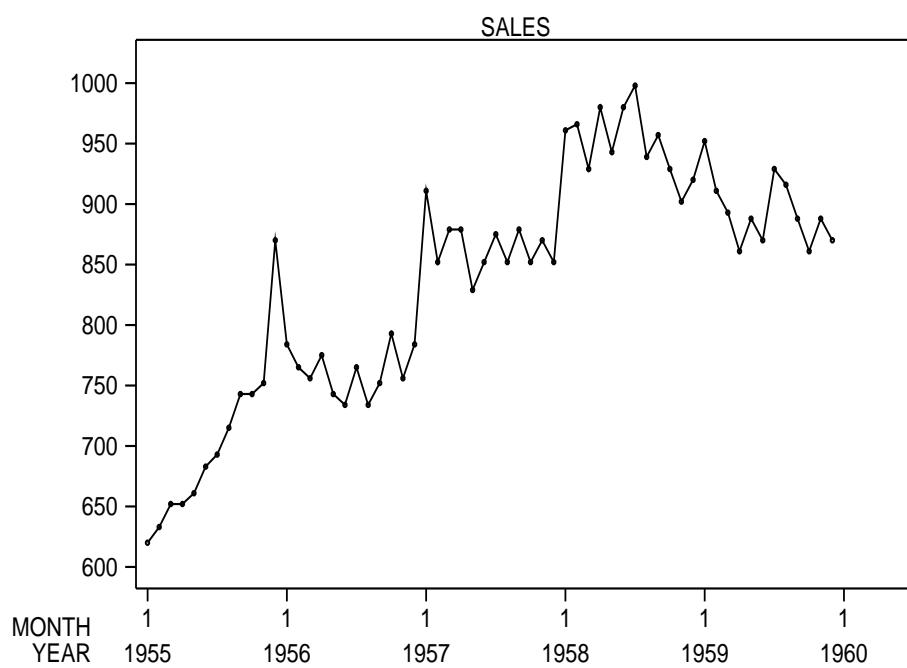
...

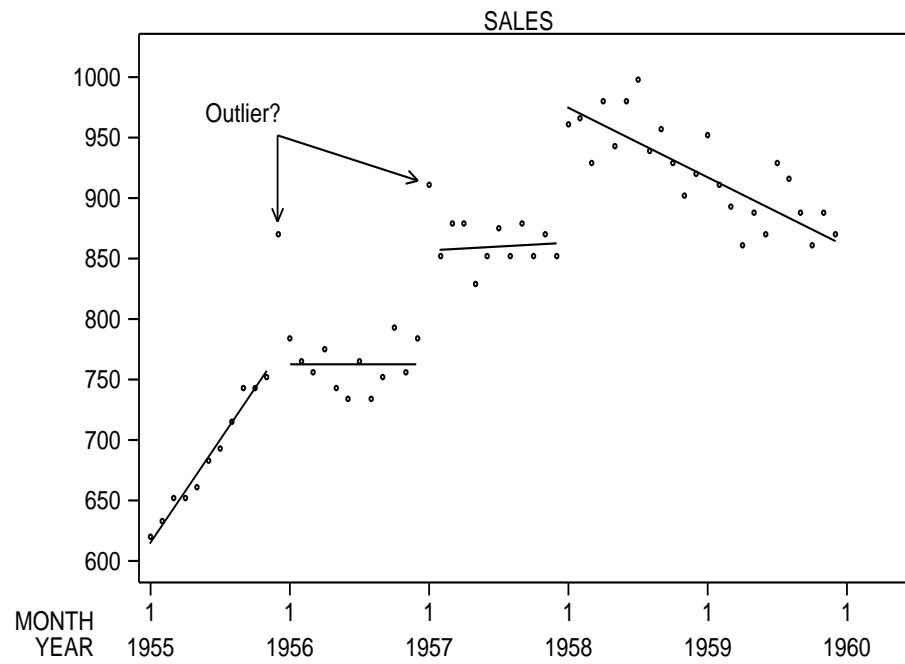
**SEQUENTIAL MONITORING & INTERVENTION**

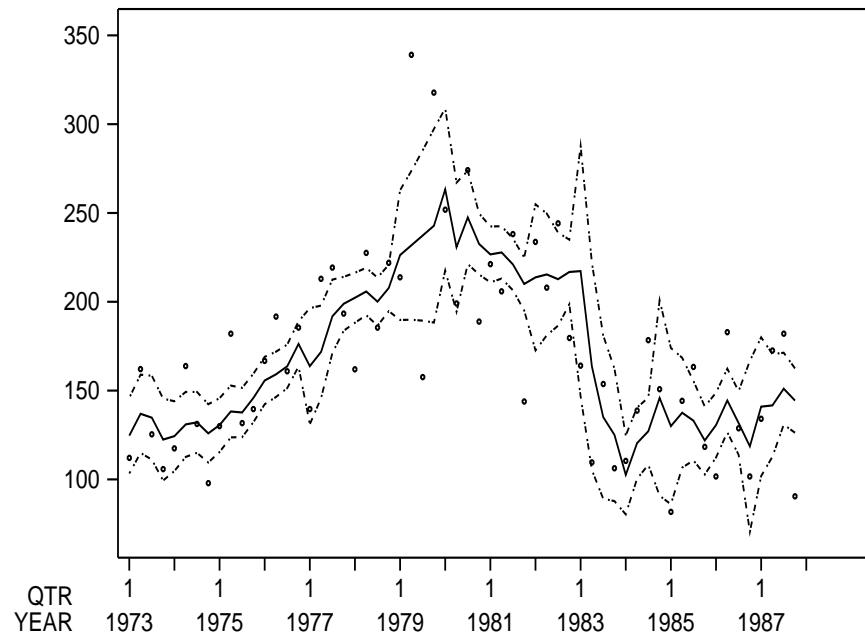
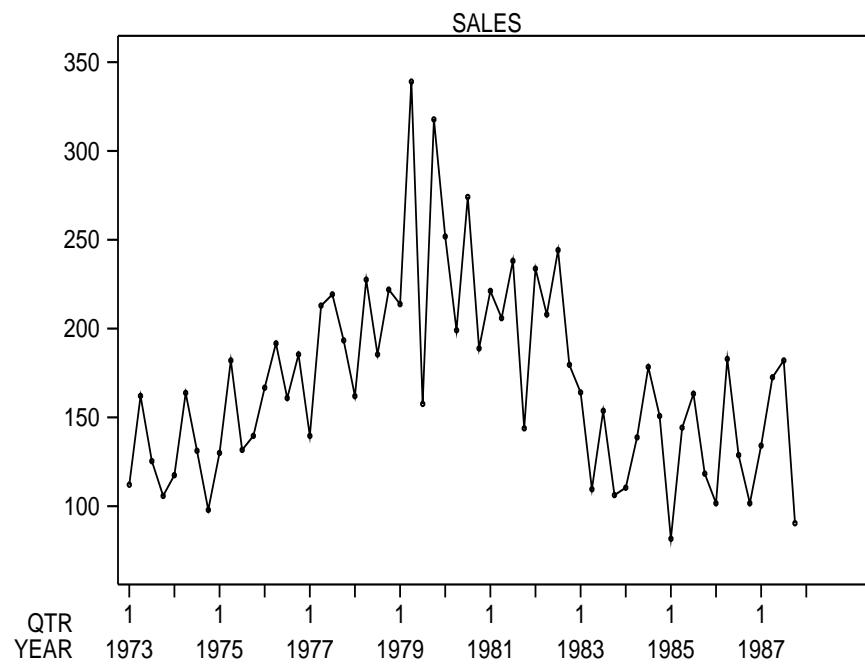
- “Routine” learning, forecasting, most of the time
- “Exceptions” signaled by monitoring methods or models
- “Management by exception”
- *Interventions:* feedforward and feedback

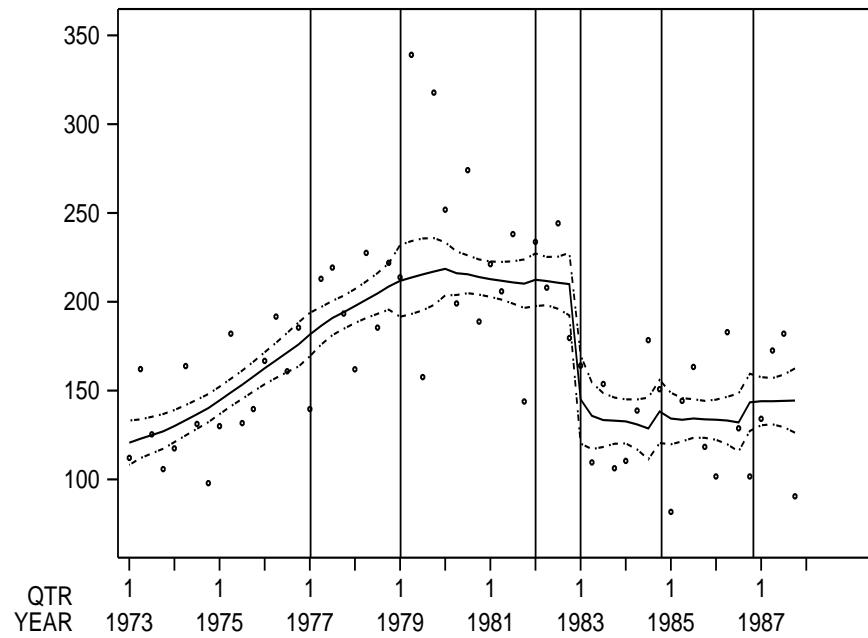
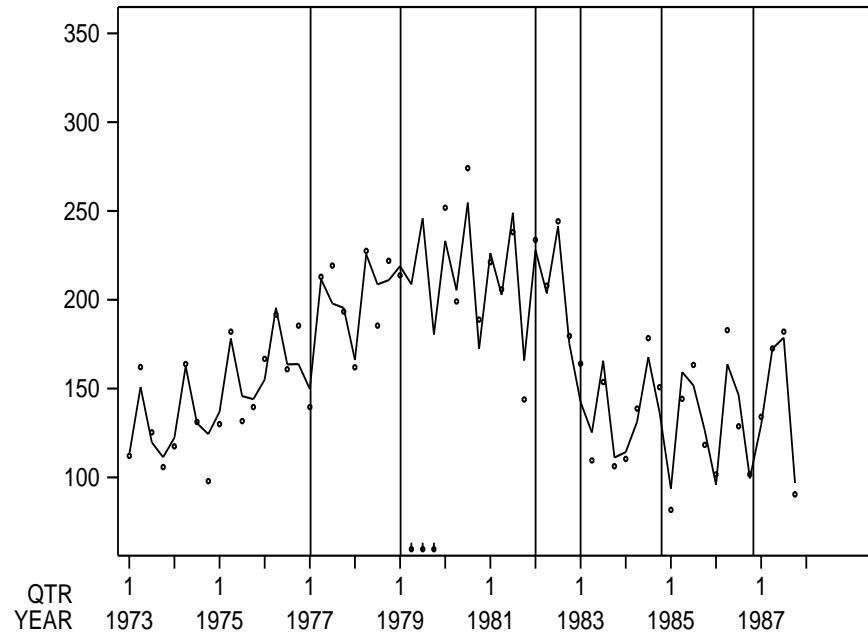
## INTERVENTION TOOLS

- Feedforward: adjust prior distributions for future states
- Feedback: reactive, based on monitoring and assessment
- Increase uncertainty about parameters: “spread” of priors
- Component-wise/selective interventions
  - e.g. Level more likely to change than seasonal factor
- Assess/monitor intervention effects









*MODEL MONITORING:* Sequential “tracking schemes”

Assessing model “predictive performance”:

Model forecast  $p(y_t|D_{t-1})$

“Alternative”  $p_A(y_t|D_{t-1})$

BF (Bayes’ factor)  $H_t = p(y_t|D_{t-1})/p_A(y_t|D_{t-1})$

Cumulative BF:

$$H_t(k) = H_t H_{t-1} \cdots H_{t-k+1}$$

$$= p(y_t, \dots, y_{t-k+1}|D_{t-k})/p_A(y_t, \dots, y_{t-k+1}|D_{t-k})$$

*MODEL MONITORING:* via Bayes’ factors

Alternatives:

- Specific “model failures” (scale inflation, level shift, ...)
- Neutral: e.g., “diffuse” version of  $p(y_t|D_{t-1})$
- Monitor one or several in parallel

Assessments:

- Low  $H_t \rightarrow$  red flag (outlier? change-point? .. ? )
- Low  $H_t(k)$  for  $k > 1$  : localise time of “breakdown”
- Neat theory to update monitor:  $L_t = \min_k H_t(k)$  as  $t \rightarrow t + 1$   
 : Bayesian “cusum” monitors and SPRT (BATS)  
 : Decision theory: West and Harrison, 2nd Edn.

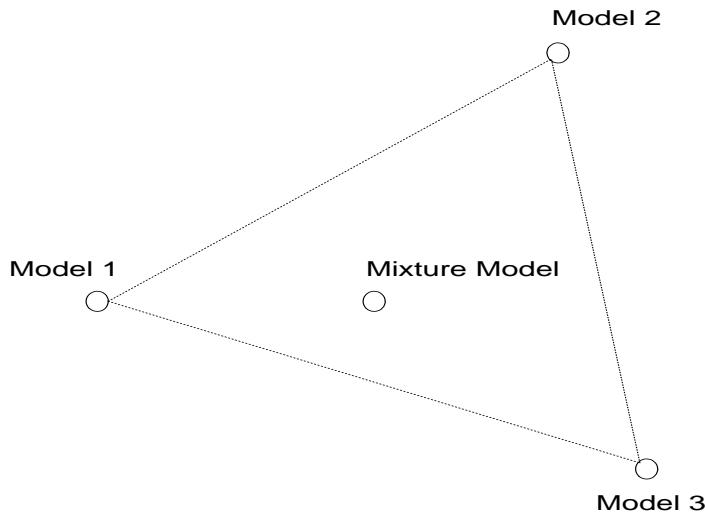
## MIXTURE MODELS IN BAYESIAN TIME SERIES

Varieties:

- Models in parallel: model uncertainty, comparison, averaging
- Mixtures as analytic approximations: Non-linear models
- *Anything can happen!*
  - Multi-process mixtures
  - Regime switching (hidden Markov models)

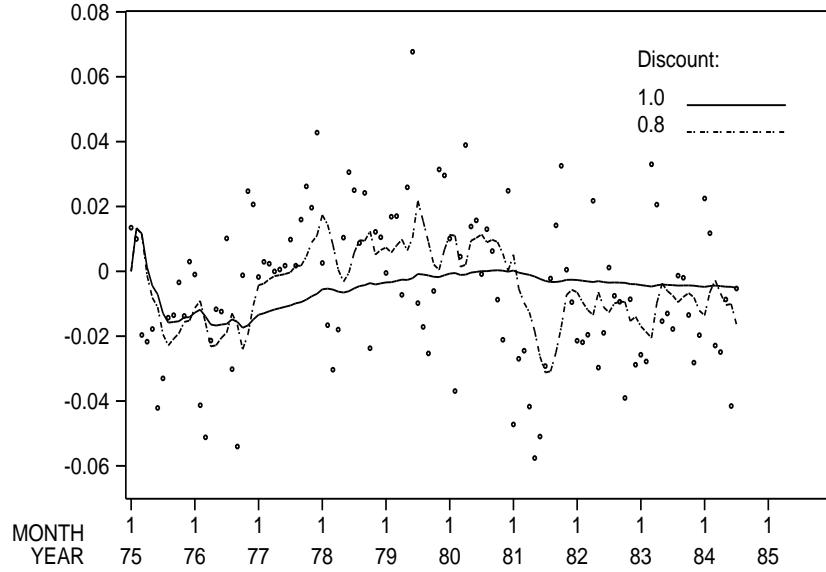
Mixtures of DLMs in application since mid-1960s: commercial, biomedical time series, engineering (“Gaussian sums”), ...

*MODELS IN PARALLEL:*  $M_i = \text{Model } i \text{ from a set } i = 1, \dots, h$



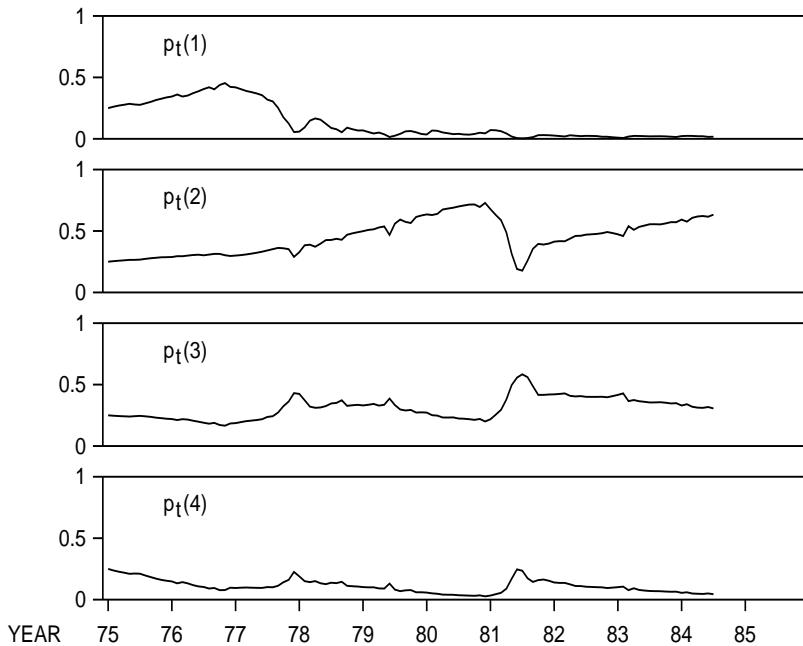
*There is no “TRUE” model*

Simple example: Exchange rate series  
 $M_i$  model different degrees of change in average level  
 via differing discount factors

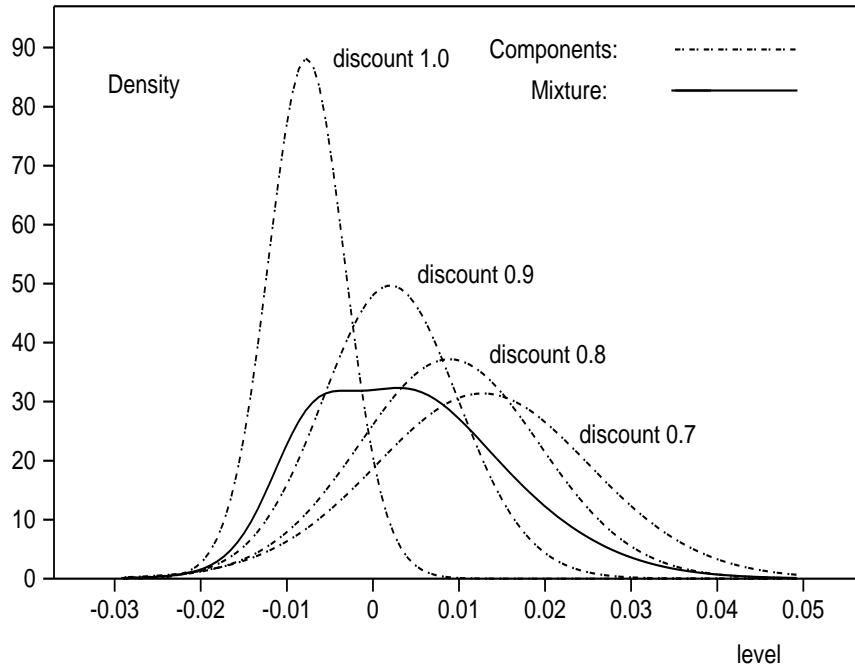


Models in parallel:  $M_i = \text{Model } i \text{ from a set } i = 1, \dots, h$

Sequential analysis  $\rightarrow Pr(M_i | D_t) = p_t(i)$  over time  $t$



Models mixing/averaging:



### MULTI-PROCESS MIXTURES

$$Y_t = x_t + \nu_t, \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t, \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

Mixture error models:

$$\nu_t \sim N(0, V_t), \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \mathbf{W}_t)$$

where  $\{V_t, \mathbf{W}_t\}$  take values in a discrete set of size  $h$

$\rightarrow$  models  $M_i$  ( $i = 1, \dots, h$ )

Normal DLM under each model  $M_i$  at each time point

Times  $t = 1, \dots, T$ :  $h^T$  combinations of models

*Historically:* Approximations to “prune” and “collapse” mixtures

*Recently:* MCMC methods and others

### MULTI-PROCESS MIXTURES: simple examples

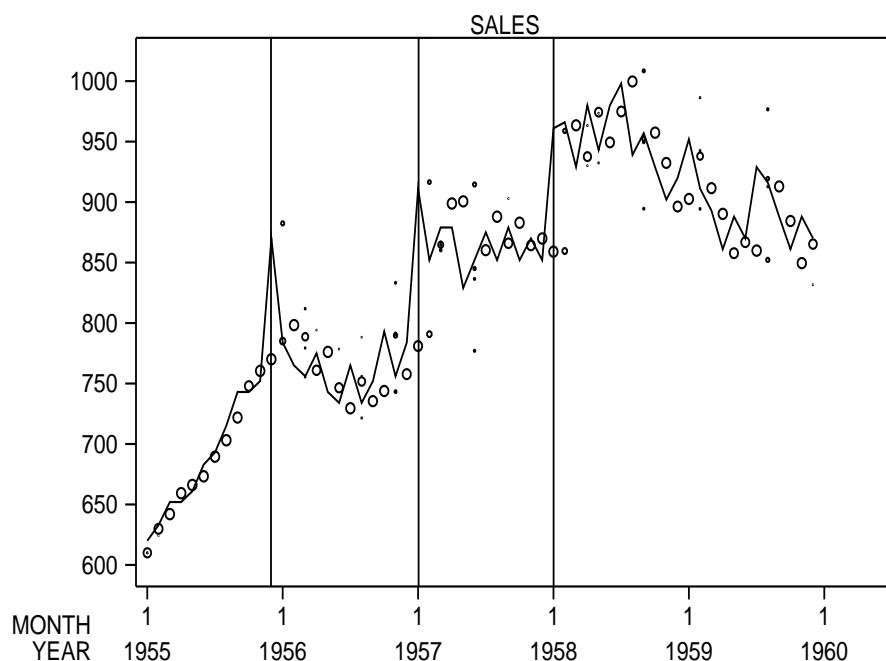
Outliers and change-points (West and Harrison 1989,97):

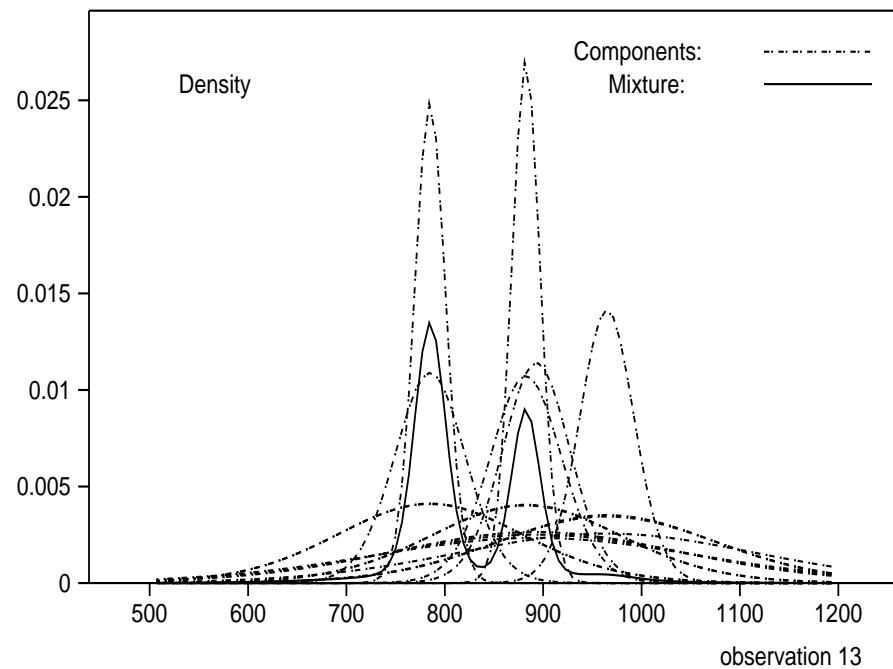
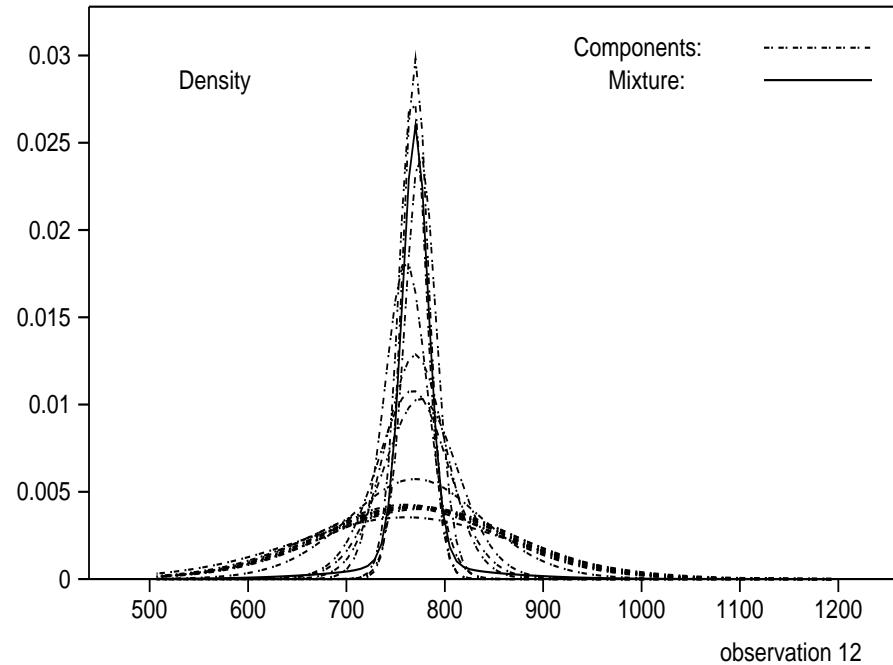
- $M_1$  : local linear trend model
- $M_2$  : outlier model: large observation variance
- $M_3$  : level change at  $t$  : large variance on level change in evolution equation
- $M_4$  : slope change at  $t$  : large variance on slope change in evolution equation

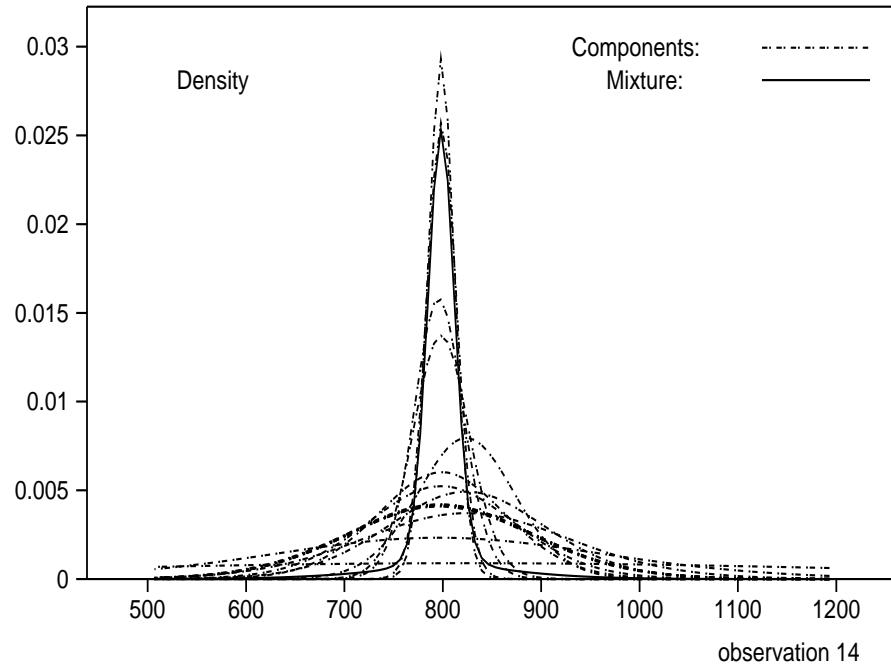
$M_i$  applies independently at each  $t$ , priors  $Pr(M_i \text{ at } t | D_0)$

Various commercial and biomedical applications 1970-present

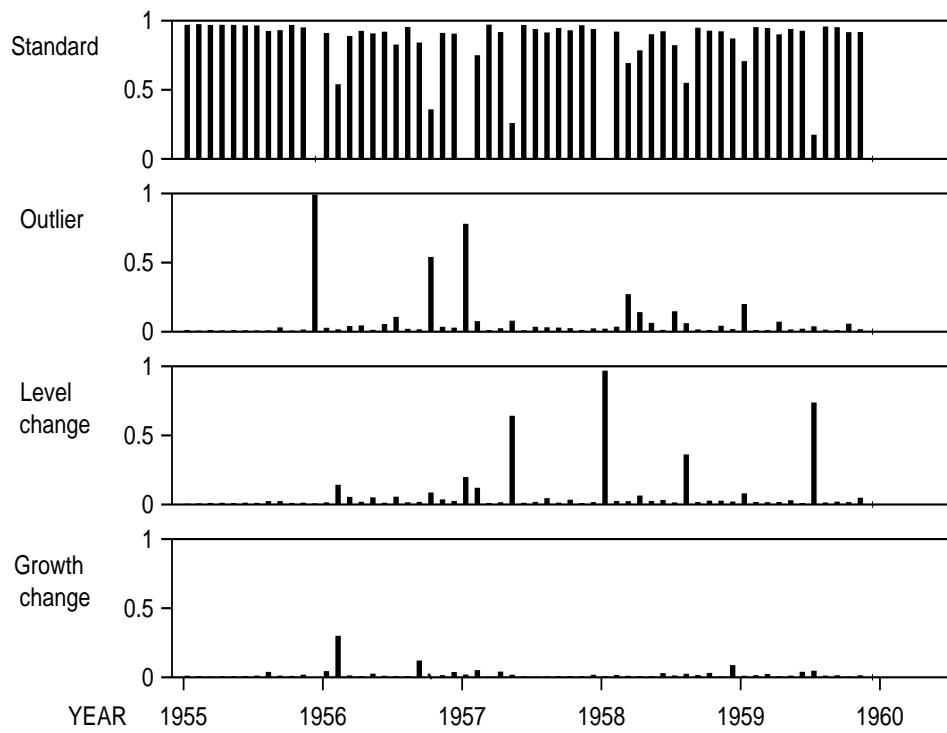
Variants: Markov switching models: Models for  $Pr(M_i \text{ at } t | D_{t-1})$







$Pr(M_i \text{ at } t-1 | D_t)$  over time:



MIXTURES AS APPROXIMATIONS:  
to non-linear/non-normal models

(a) Flexible “curve-fitting” approximations:

Non-linear autoregressive models: Conditional density estimation

(P Müller, M West, S McEachern 1995)

(b) Direct analytic approximations:

Example:  $Y_t = \exp(x_t)\chi_2^2$

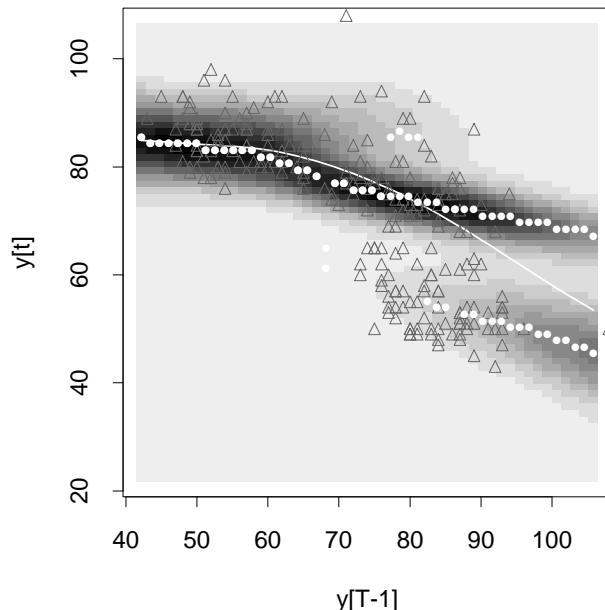
$$x_t = \mathbf{F}'_t \boldsymbol{\theta}_t, \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

So  $y_t = \log(Y_t)$  follows DLM  $y_t = x_t + \nu_t$  with non-normal error

Analytic approximation:  $p(\nu_t) \approx$  mixture of 5 normals (known)

Applications in: Stochastic volatility models in finance and ...

$$y_t = x_t, \quad p(x_t | x_{t-1}) = \sum_j w_j(x_{t-1}) N(a_j + b_j x_{t-1}, v_j)$$



... and in *Non-parametric Bayesian spectral analysis*:

– assumedly stationary time series –

$Y_1, Y_2, \dots$  : sample periodogram ordinates

$x_1, x_2, \dots$  : multiple of “true” spectral density function

Model:  $x_t = \mathbf{F}'_t \boldsymbol{\theta}_t, \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$

- “smoothness priors” for log spectrum
- one-dimensional Markov random field priors
- C Carter and R Kohn (1996)
- G Petris and M West (1996) ... smoothness priors plus components for long-range dependence in time series

## MULTIVARIATE SERIES

### MODELS IN FINANCIAL APPLICATIONS

- Futures markets, exchange rates, portfolio selection
- Multiple time series: time-varying covariance patterns
- Econometric/dynamic regressions/hierarchical models
- Bayesian multivariate stochastic volatility

$$\mathbf{y}_t = (y_{1t}, \dots, y_{pt})'$$

e.g.,  $y_{it}$  is  $p$ -vector of returns on investment  $i$   
(exchange rate futures, etc.)

## A DLM CLASSES FOR MULTIVARIATE SERIES:

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{F}_t[\boldsymbol{\theta}_{1t}, \dots, \boldsymbol{\theta}_{pt}] + \boldsymbol{\nu}_t, \quad \boldsymbol{\theta}_{it} = \mathbf{G}_t \boldsymbol{\theta}_{i,t-1} + \boldsymbol{\omega}_{it}$$

- $\mathbf{x}_t$  is  $p$ -vector of DLMs for trend, regression, etc.
- $\boldsymbol{\theta}_{it}$  series-specific state vectors
- $V(\boldsymbol{\nu}_t) = \boldsymbol{\Sigma}_t$  ... time-varying covariance patterns:  
stochastic “volatility”
- $\boldsymbol{\omega}_{it}, \dots, \boldsymbol{\omega}_{pt} \sim \text{matrix normal distribution}$ 
  - within-series covariances  $V(\boldsymbol{\omega}_{it})$
  - AND correlated across series  $i$  via patterns in  $\boldsymbol{\Sigma}_t$
- Random shocks impact across series in similar ways in both state and observation equations

Example: *One-step ahead portfolio choice*:

- at  $t - 1$ , predict  $\mathbf{y}_t$  and hence portfolio  $r_t = \mathbf{a}'_t \mathbf{y}_t$  for given distribution of assets  $\mathbf{a}_t$
- $r_t \sim N(\cdot, q_t + s_t \mathbf{a}'_t \boldsymbol{\Sigma}_t \mathbf{a}_t)$
- choose  $\mathbf{a}_t$  to maximise expected return subject to risk balance:  
Bayesian decision analysis
- expected loss functions  $L(\mathbf{a}_t | D_{t-1})$  critically depend on correlation patterns
- $\boldsymbol{\Sigma}_t$  ... time-varying covariance patterns: require sensitive tracking, adaptation ... Issue: Model and infer changes in  $\boldsymbol{\Sigma}_t$
- J M Quintana, early work in Quintana & West (1986)

*Stochastic models:* “Random walk evolution” for  $\Sigma_t$

- $\Sigma_t = f(\Sigma_{t-1}, \mathbf{U}_t)$  for random matrix  $\mathbf{U}_t$
- Results in “exponential smoothing” for estimates of  $\Sigma_t$
- Adapts to changing patterns: portfolio responds

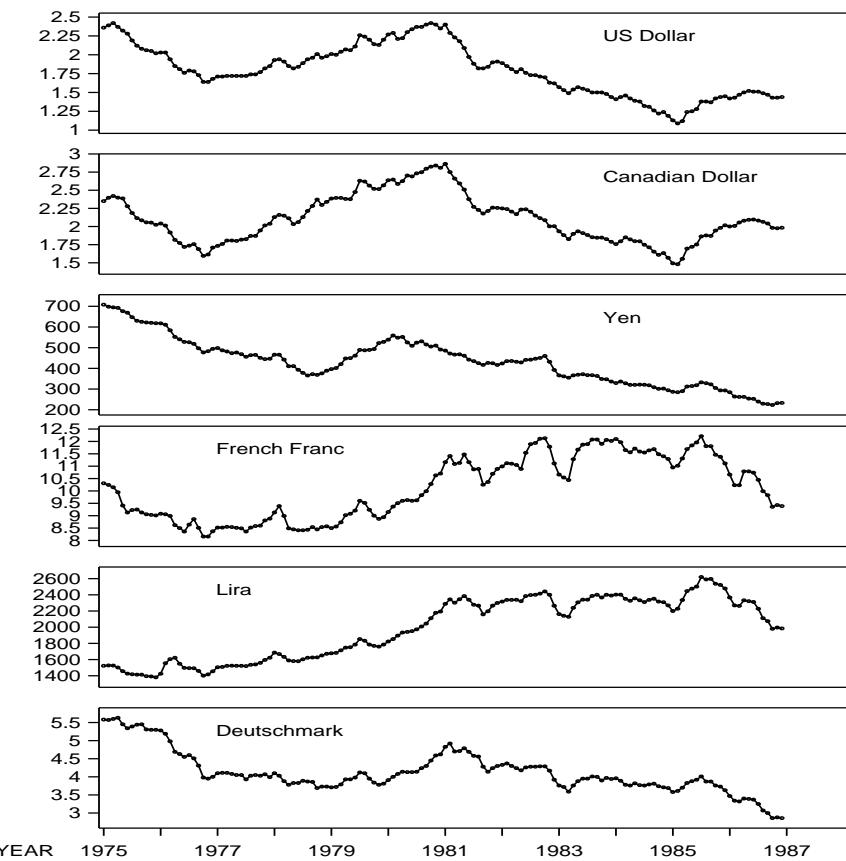
Documented applied success: Chase Investors, Banker’s Trust, ...

*Principal Component/Factor structure:*

$$\nu_t = \sum_i z_{it} \mathbf{e}_{it} \quad \text{and} \quad \Sigma_t = \sum_i b_{it} \mathbf{e}_{it} \mathbf{e}'_{it}$$

- latent factor  $z_{it}$  with variance  $b_{it}$
- orthogonal vectors  $\mathbf{e}_{it}$

*Example:* International exchange rate time series (vs. Sterling)



Weights of currencies in first three factors (based on  $E(\Sigma_t | Data)$ )

	December 1983			December 1986		
	$e_1$	$e_2$	$e_3$	$e_1$	$e_2$	$e_3$
\$ U.S.A.	0.3	-0.6	-0.1	0.4	-0.5	0.3
\$ Canada	0.3	-0.6	.	0.3	-0.5	0.3
Yen	0.5	.	0.9	0.4	0.7	0.6
Franc	0.5	0.3	-0.3	0.4	.	-0.5
Lira	0.4	0.2	-0.2	0.5	.	-0.2
DMark	0.4	0.3	-0.3	0.4	.	-0.4
% Variation	73	18	6	71	20	9

## MODELS IN SCIENTIFIC APPLICATIONS

Historical interest in biomedical monitoring (change-points), engineering applications (control, tracking), environmental monitoring, ...

Recent work in *biomedical and geophysical areas*

- time series decompositions: retrospective analysis
- evaluation of quasi-periodic patterns in LATENT subseries
- non-stationary, time-varying “spectral” characteristics
- changes over time at different time scales
- issues of uncertainty about time scales and timing of data

*DLM autoregressions and time-varying autoregressions*

**AUTOREGRESSIVE COMPONENT DLM:**

Latent AR( $d$ ) process:  $x_t = \sum_{j=1}^d \phi_j x_{t-j} + \epsilon_t$

Time series:  $y_t = x_{0t} + x_t + \nu_t$  with trend, etc in  $x_{0t}$

DLM for  $x_t$ :  $x_t = (1, 0, \dots, 0)\mathbf{x}_t$ ,  $\mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \boldsymbol{\omega}_t$

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{d-1} & \phi_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \quad \boldsymbol{\omega}_t = \begin{pmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

–  $x_t$  latent, unobserved

–  $\mathbf{G} = \mathbf{G}(\boldsymbol{\phi})$  with  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_d)'$  to be estimated

**TIME-VARYING AUTOREGRESSION:**

TV-AR( $d$ ) model:  $x_t = \sum_{j=1}^d \phi_{t,j} x_{t-j} + \epsilon_t$

- AR parameter:  $\boldsymbol{\phi}_t = (\phi_{t,1}, \dots, \phi_{t,d})'$  “wanders” through time:

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_{t-1} + \partial\boldsymbol{\phi}_t$$

- stochastic “shocks”  $\partial\boldsymbol{\phi}_t$
- Innovations:  $\epsilon_t \sim N(0, \sigma_t^2)$  – time-varying variance  $\sigma_t^2$

Flexible representations:

- non-stationary process, time-varying spectral properties
- latent component structure: (see below)

DLM TV-AR( $d$ ):

$$x_t = (1, 0, \dots, 0)' \mathbf{x}_t$$

$$\mathbf{x}_t = \mathbf{G}(\phi_t) \mathbf{x}_{t-1} + (\epsilon_t, 0, \dots, 0)'$$

$$\phi_t = \phi_{t-1} + \partial\phi_t$$

with

$$\mathbf{G}(\phi_t) = \begin{pmatrix} \phi_{t,1} & \phi_{t,2} & \cdots & \phi_{t,d-1} & \phi_{t,d} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

*Analysis:* Posterior distributions for  $\{\phi_t, \sigma_t : \forall t\}$

*Component models:*  $y_t = x_{0t} + x_{1t} + \nu_t$

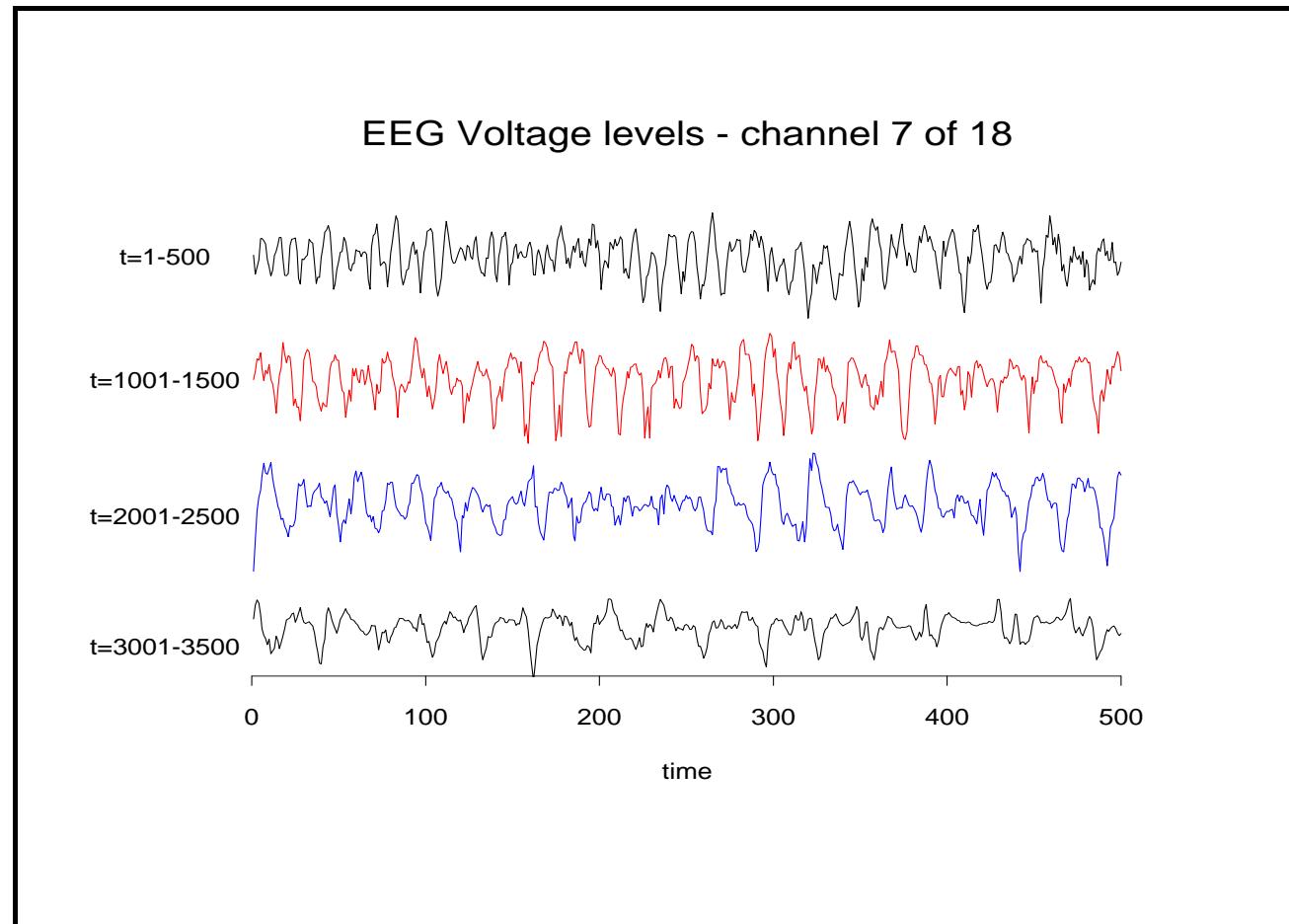
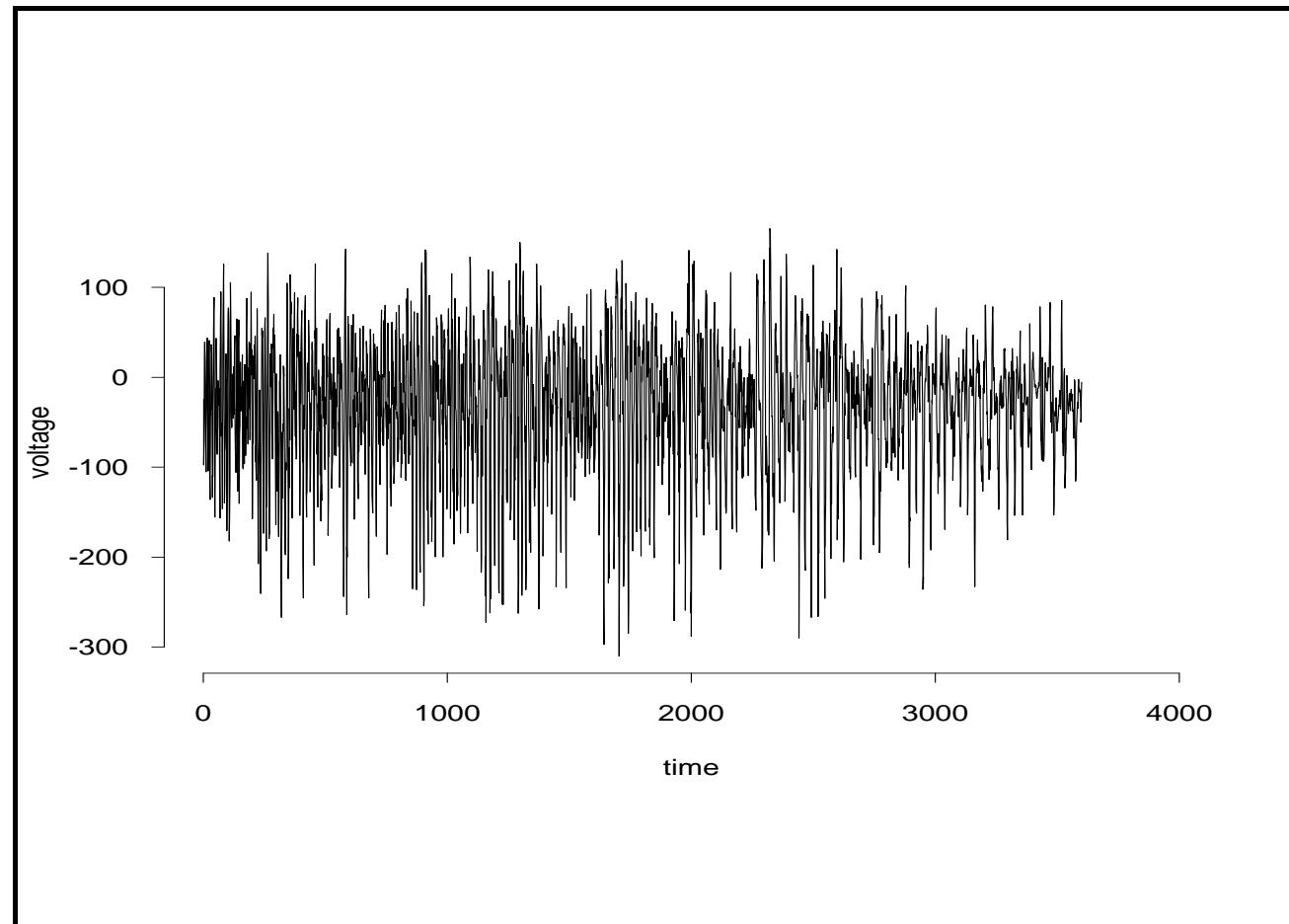
→ infer latent TV-AR processes too:  $\{x_{0t}, x_t : \forall t\}$

*EXAMPLE: EEG study (+ R Prado & A Krystal of Duke)*

- many multiple series: ECT therapy, clinical & basic neuroscience
- one seizure: 18 channels, 256/sec (subsample 25,000 obsns)
  - Comparisons across channels – *redundancies? placement design? varying seizure effects “spatially”?*
  - Clinical issues: – *treatment effects on seizure waveform?*

*Issues:* Characterise seizure episodes ... *models*

- Seizure “waveform” ... time varying amplitudes, frequencies
- Superimposed on “normal” waveform, noise, ...



**TIME SERIES DECOMPOSITIONS:** In general DLM framework

$$\text{DLM: } x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

*Eigenstructure:*  $p \times p$  state matrix  $\mathbf{G}_t$  distinct eigenvalues  
 $c$  complex conjugate pairs,  $r$  real eigenvalues

*Decomposition:*

$$x_t \equiv \sum_{j=1}^c z_{t,j} + \sum_{j=1}^r a_{t,j}$$

- $z_{t,j} \sim$  quasi-periodic TV-ARMA(2,1)  
*i.e. “sinusoid” with randomly time-varying amplitude and phase  
AND time-varying frequency/wavelength*
- $a_{t,j} \sim$  TV-AR(1)  
*(usually) short-term correlated noise components  
(less often) low frequency “trends”*

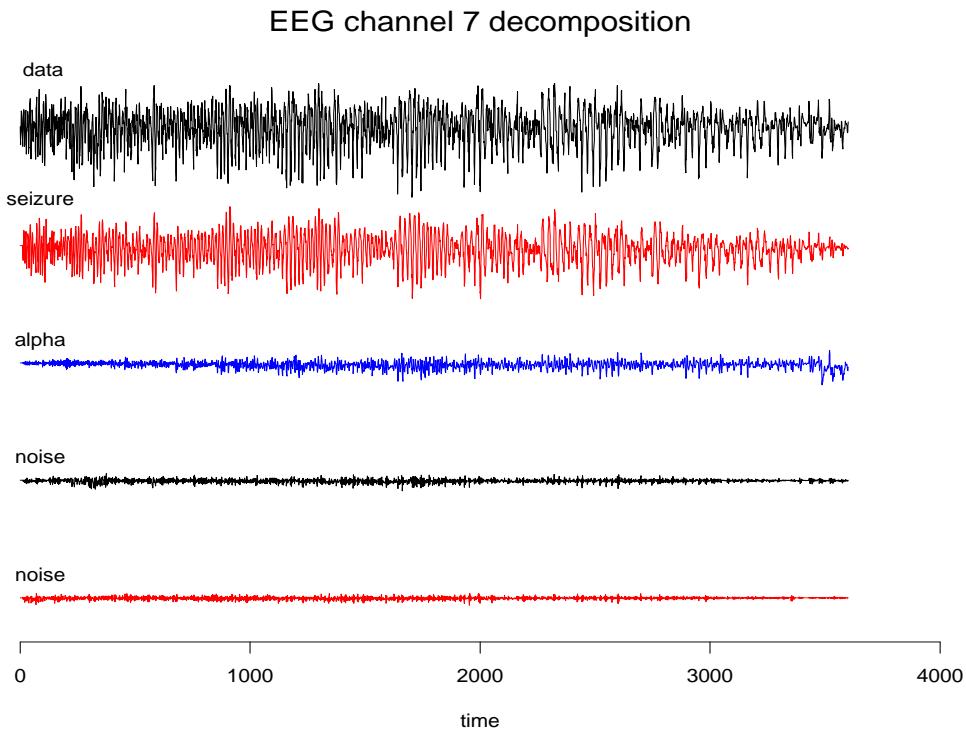
- *Constructive Decomposition Result:* Given  $\{\mathbf{F}_t, \mathbf{G}_t, x_t \quad \forall t\}$   
**compute** latent processes  $z_{t,j}$  and  $a_{t,j}$  over time window of data
- Physically interpretable components?
- Time constant  $\mathbf{F}, \mathbf{G} \rightarrow$ 
  - latent ARMA(2,1)  $z_t$  processes  
*(sinusoids with randomly time-varying amplitude and phase, FIXED frequency/wavelengths)*
  - and AR(1)  $a_t$  processes
- *Maths:* Eigenstructure of  $\mathbf{G}_t$  : Similarity transforms, Similar models (“standard” DLM/linear systems theory)
- $y_t = x_t + \dots + \nu_t$  : analysis to infer/estimate latent  $x_t$  process

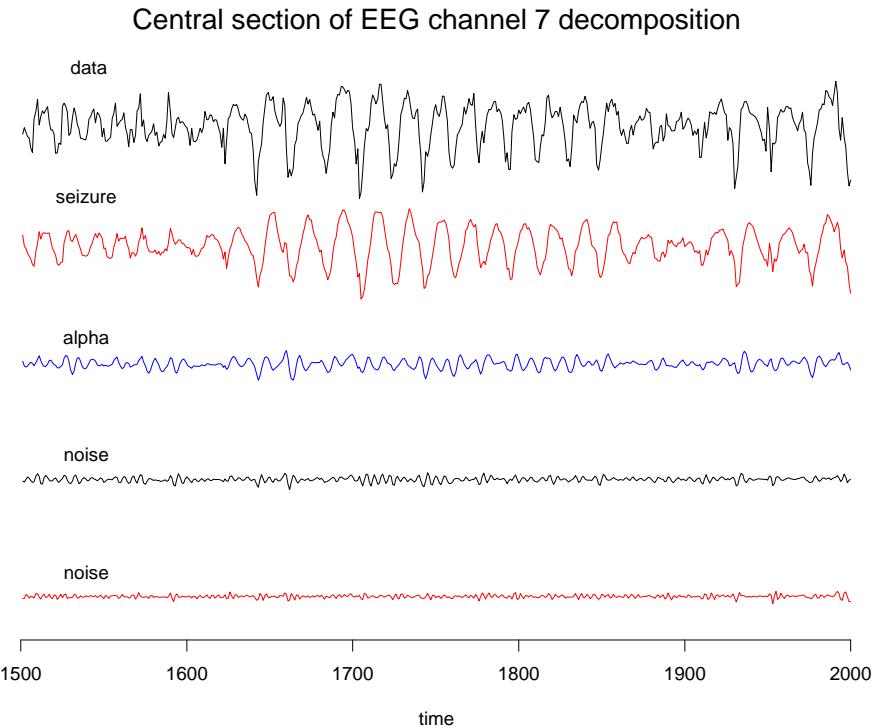
*EEG series – One channel:*

Data  $x_t \sim \text{TV-AR}(12)$

*Decomposition:* Posterior mean of  $\phi_t$  at each  $t$  :  $E(\phi_t | \text{Data})$

- seizure “*slow wave*” (2-4 cps):  
trajectories of frequency and amplitude → characterise seizure
- normal “*alpha wave*” (5-7 cps):  
slowly-varying frequency, amplitude – seizure effects
- residual components: *noise*





*EEG series – One channel (continued)*

*Higher order model:*

TV-AR(12) delivers two basic processes  $\approx$  TV-AR(4)

*Rationale:*

- *Unobserved signal process:*  $x_t \sim \text{TV-AR}(4)$
- *Uncorrelated error/noise process:*  $\nu_t$
- Data  $y_t = x_t + \nu_t$
- $\rightarrow y_t \sim \text{TV-AR}(\infty) \approx \text{TV-AR}(d), d \gg 4$

Suggests basic TV-AR(4) with two key  $z_t$  processes, plus noise

Verifiable within analysis (functions of  $\phi_t$ )

### *EEG series – One channel (continued)*

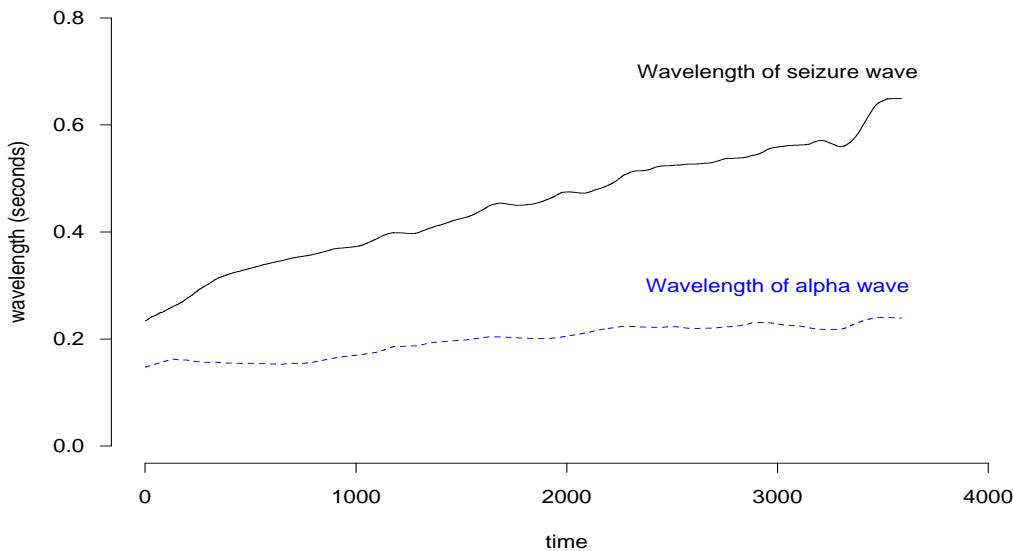
TV-ARMA(2,1)  $z_t$  processes:

Time-varying frequencies/wavelengths, amplitudes, phases

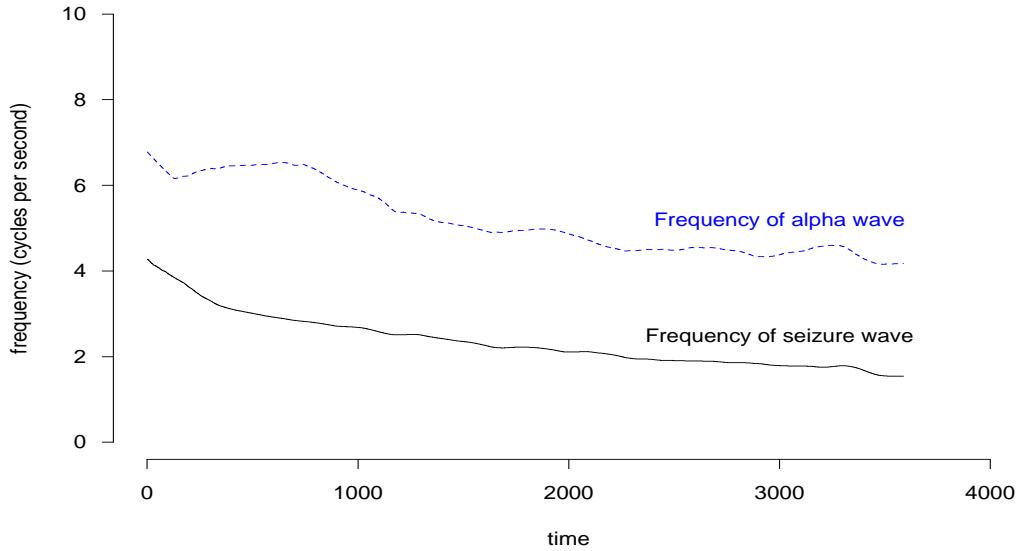
- frequencies/wavelength computable functions of  $\phi_t$
- posterior for  $\{\phi_t, \forall t\}$
- → infer “instantaneous” frequency (spectral) characteristics
- non-stationary spectral analysis (in the time domain)

cf., Kitagawa and Gersch 1996 (Springer-Verlag): estimate spectrum over time

Time-variation in wavelengths of seizure and alpha waves



... and frequencies = 1/wavelengths



### EXAMPLE: Oxygen isotope series (one of several)

- deep ocean cores: relative abundance of  $\delta^{18}\text{O}$
- $\delta^{18}\text{O} \downarrow$  as global temperatures  $\uparrow$  (smaller ice mass)
- Shackleton and Hall (1989), J Parks (1992)
- *reverse sign*: higher recent global temperatures
- “well known” periodicities: earth orbital dynamics  $\rightarrow$  impact on solar insolation – Milankovitch; Shackleton *et al* since 1976
  - eccentricity*: 95-120 kyear
  - obliquity*: 40-42 kyear
  - precession*: 19-25 kyear

- Form of time variation in individual cycles ?
- Timing/nature of onset of “ice-age” cycle  $\leftrightarrow$  eccentricity component  $\sim 1000$  kyears ago ?
- Time scale: errors, interpolation, ... measurement, sampling error, etc

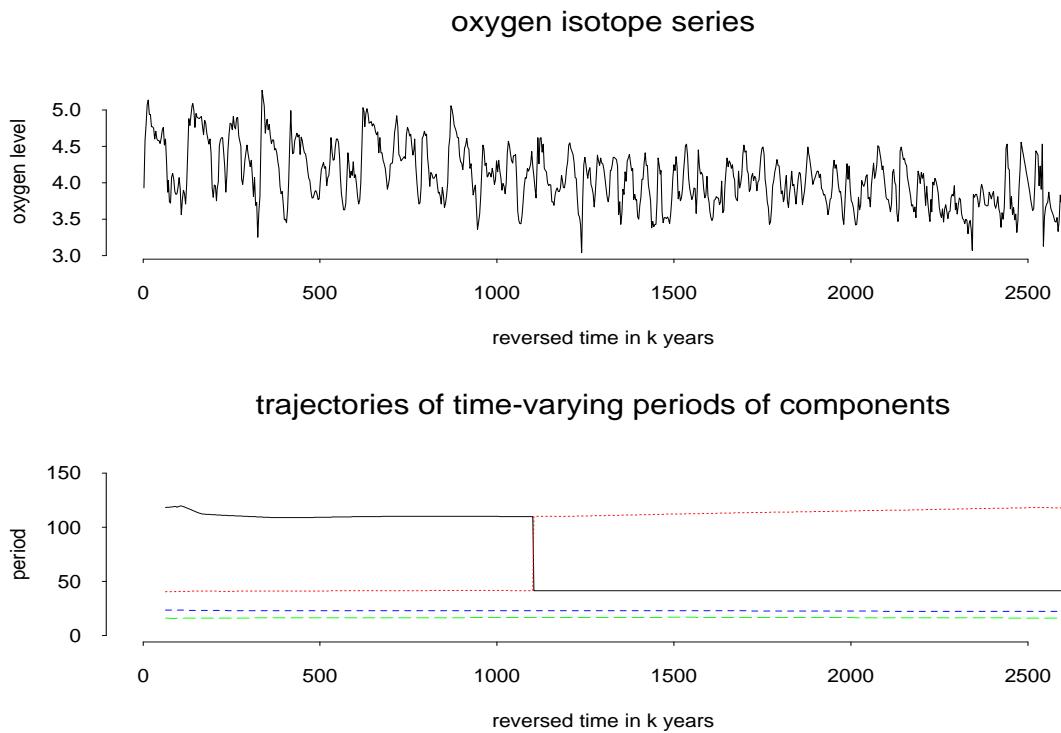
*Models:* High order TV-AR,  $p = 20$ , plus smooth trend (outliers?)

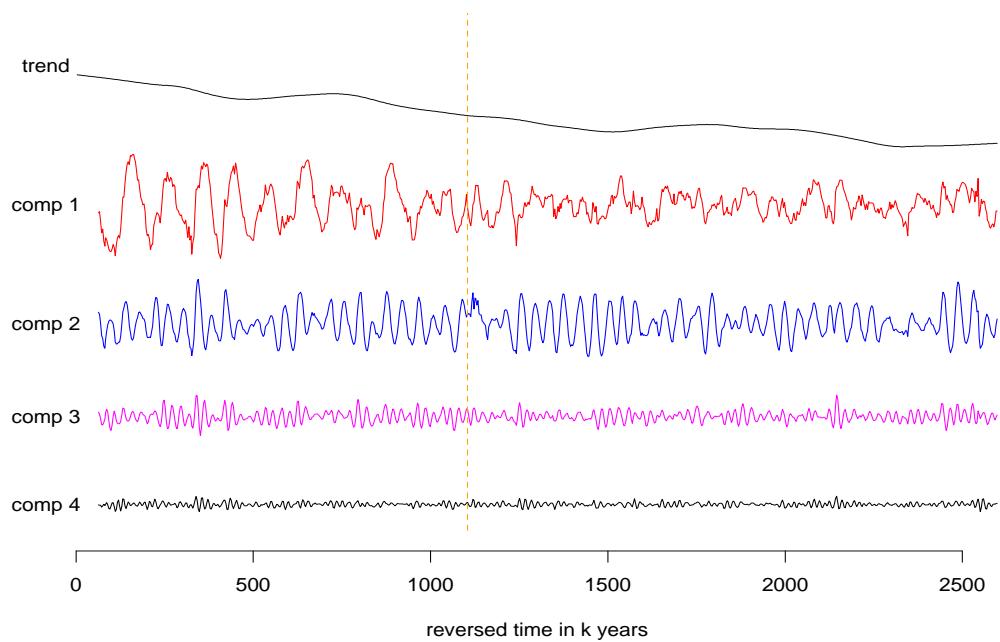
Variance components estimated: changing AR parameters

*Decomposition:* Posterior mean of  $x_t, \phi_{t,j}$  at each  $t$

3 dominant quasi-periodic components: order by estimated amplitude (innovation variance)

*Others:* residual structure &/or contaminations





*Inference on periods:*

<i>eccentricity:</i>	95-120 kyear	posterior: 108–120, peak 110
<i>obliquity:</i>	40-42 kyear	posterior: 40.8–41.6, peak 41.5
<i>precession:</i>	19-25 kyear	posterior: 22.2–23, peak 22.8

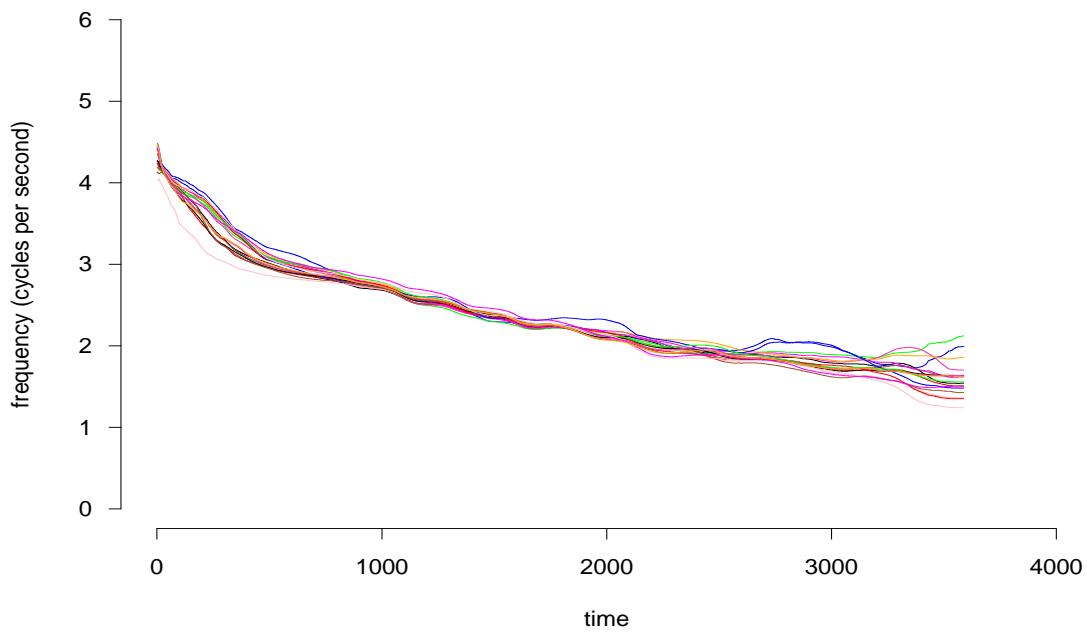
## MORE MULTIVARIATE MODELS

Multiple latent processes and factor structure

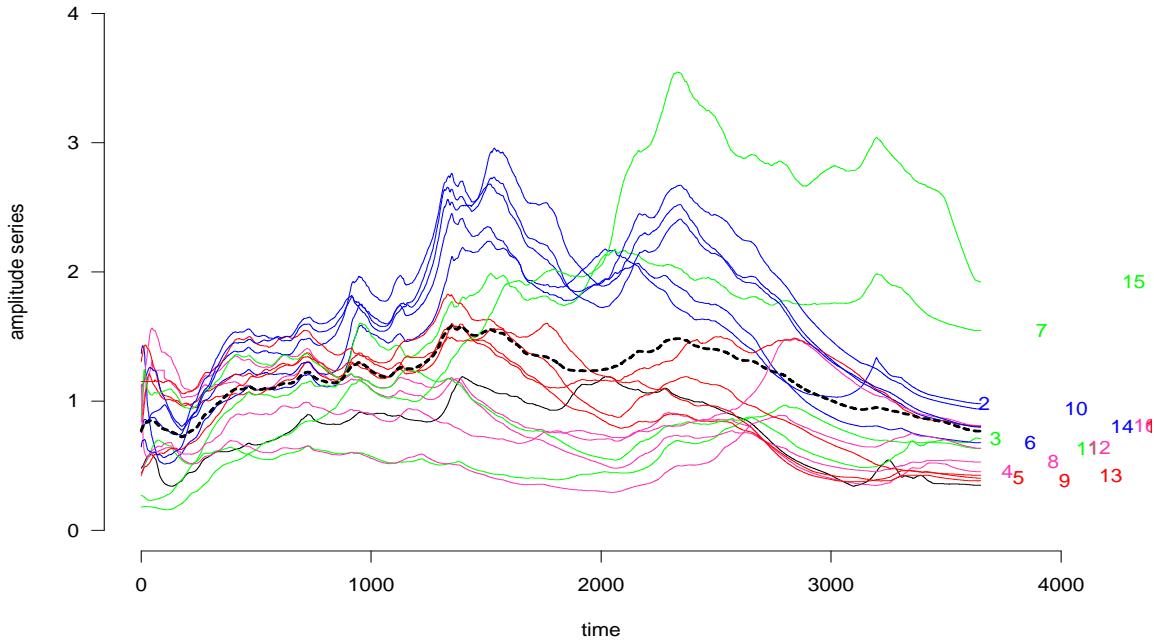
*EXAMPLE: Multiple EEG channels,  $m = 18$  series*

Similar models, decompositions

- common underlying seizure waveform plus alpha rhythm
- differing amplitude characteristics, noise properties



EEG channels: trajectories of amplitudes of dominant components



### Multivariate Structure: Dynamic Latent Factor Models

*Latent seizure process:*  $x_{t,1} \dots$  TV-AR(2)

*Latent "alpha" rhythm process:*  $x_{t,2} \dots$  TV-AR(2)

*EEG channel  $i$ :*

$$y_{t,i} = a_{i,1}x_{t,1} + a_{i,2}x_{t,2} + e_{t,i}$$

*Vector model:* all  $m = 18$  channels  $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,m})'$

$$\mathbf{y}_t = \mathbf{a}_1 x_{t,1} + \mathbf{a}_2 x_{t,2} + \mathbf{e}_t$$

Infer latent processes  $x_{t,i}$  and factor weight vectors  $\mathbf{a}_i$ , etc.

### Dynamic “Factor” models

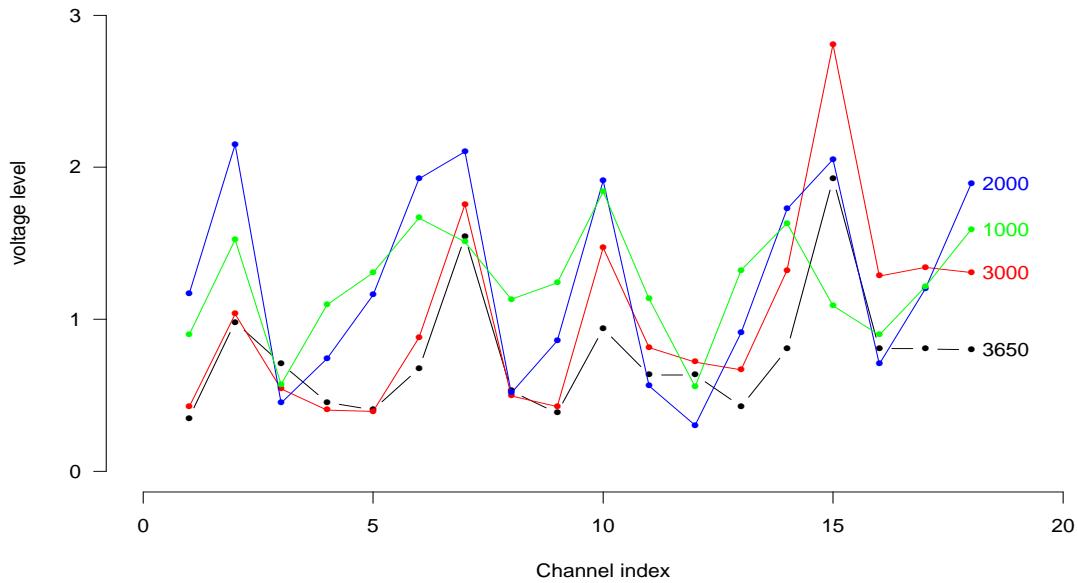
$$\mathbf{y}_t = \sum_{i=1}^k \mathbf{a}_i x_{t,i} + \mathbf{e}_t$$

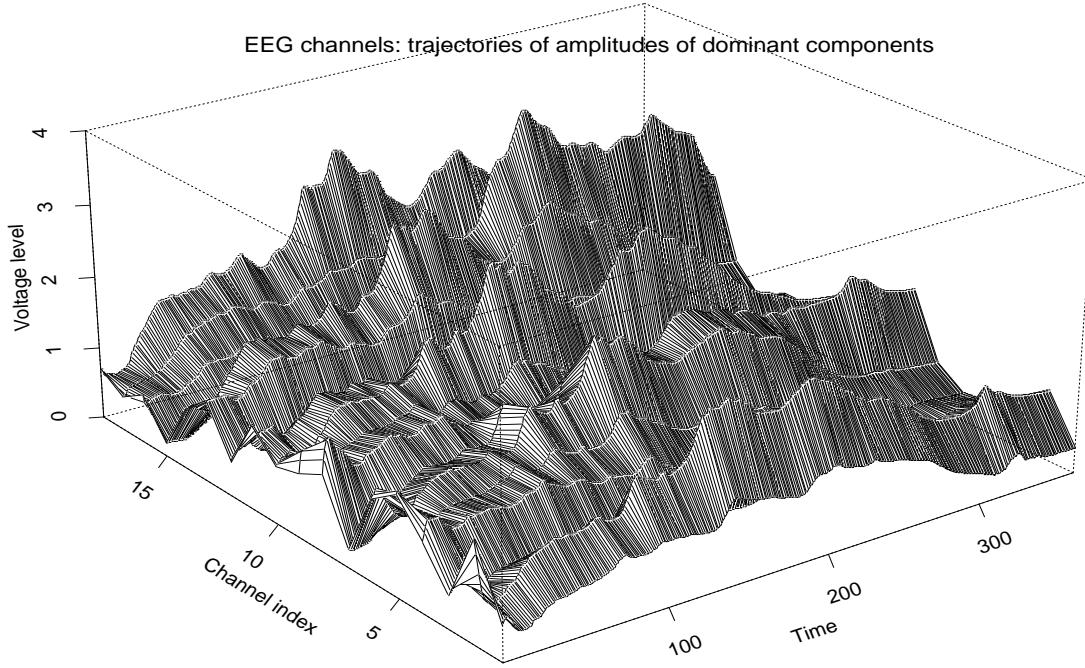
- Physical (latent) univariate processes  $x_{t,i}$
- e.g.  $\dim(y_t) = 18$ , etc., large; # factors  $k = 2$  or 3
- $x_{t,i} \sim$  state space model; possibly dependent innovations
- Possible lag structures: e.g.  $x_{t,2} = x_{t-1,1}$
- Dynamic “principal components” – structured

Priors/Models for factor weights  $\mathbf{a}_i$  :

EEG: “spatial” connectivities

EEG channels: spatial relationship in channel amplitudes





### *Applications:*

EEG context: Many multiple series, patients, treatments

Multiple geological series: climatic change

### *Time-varying factor weights:*

$$\mathbf{y}_t = \sum_{i=1}^k \mathbf{a}_{t,i} x_{t,i} + \mathbf{e}_t$$

Financial time series: Exchange rates, portfolio analysis

Models for short-term *prediction* of multivariate stochastic volatility:  $x_{t,i} \sim$  “simple” DLMs;  $\mathbf{a}_{t,i} \sim ?$

### *Implementations?*

Simulation-based: MCMC, current research frontiers

## ADVANCED TOPIC: COMPUTATION

*SIMULATION IN DLM:*

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

Normal error models  $p(\nu_t), p(\boldsymbol{\omega}_t)$

*Fixed time window*  $t = 1, \dots, n$ :

State vector set:  $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$

Require: a full posterior sample  $\boldsymbol{\Theta}^*$  from  $p(\boldsymbol{\Theta}|D_n)$

Available via “Forward-filtering: Backward-sampling” algorithm

Carter and Kohn (1994) *Biometrika*

Frühwirth-Schnatter (1994) *J Time Series Anal*

West and Harrison 1997

*Forward-filtering: Backward-sampling (FFBS):*

- *Forward-filtering:*

- Standard normal/linear analysis: Kalman filter
- delivers normal  $p(\boldsymbol{\theta}_t|D_t)$  at each  $t = 1, \dots, n$

- *Backward-sampling:*

- at  $t = n$ : sample  $\boldsymbol{\theta}_t^*$  from  $p(\boldsymbol{\theta}_n|D_n)$
- for  $t = n - 1, n - 2, \dots, 1$ : sample  $\boldsymbol{\theta}_t^*$  from normal distribution  $p(\boldsymbol{\theta}_t|D_t, \boldsymbol{\theta}_{t+1}^*)$

Builds up  $\boldsymbol{\Theta}^*$  as  $\boldsymbol{\theta}_n^*, \boldsymbol{\theta}_{n-1}^*, \dots$

Exploits Markovian/CI model structure

**DLM PARAMETERS:**

$$y_t = x_t + \nu_t \quad x_t = \mathbf{F}'_t \boldsymbol{\theta}_t \quad \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$$

**Parameters:**

- Variances (variance matrices) of  $\nu_t, \boldsymbol{\omega}_t$
- Elements of  $\mathbf{F}_t, \mathbf{G}_t$
- Indicators in normal mixture models:
  - e.g.,  $\nu_t \sim N(0, V_t(1 + 99z_t))$
  - $z_t = 0$  or  $1$ :  $Pr(z_t = 1) = 0.05$  ... outlier model

**EXAMPLE:** Autoregressive component DLM:

Latent AR( $d$ ) process:  $x_t = \sum_{j=1}^d \phi_j x_{t-j} + \epsilon_t$

Time series:  $y_t = x_t + \nu_t$

DLM for  $x_t$ :  $x_t = (1, 0, \dots, 0)\mathbf{x}_t, \quad \mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \boldsymbol{\omega}_t$

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{d-1} & \phi_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \quad \boldsymbol{\omega}_t = \begin{pmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Parameters:  $\boldsymbol{\Phi} = \{V(\nu_t); V(\epsilon_t); \phi_1, \dots, \phi_d\}$

Generally, “parameter”  $\boldsymbol{\Phi}$  (may depend on sample size)  
such that the FFBS algorithm applies to sample  $p(\boldsymbol{\Theta}|\boldsymbol{\Phi}, D_n)$

*Gibbs Sampling:*  $p(\boldsymbol{\Theta}, \boldsymbol{\Phi}|D_n)$  iteratively resampled via

- Apply FFBS algorithm to draw  $\boldsymbol{\Theta}^*$  from  $p(\boldsymbol{\Theta}|\boldsymbol{\Phi}^*, D_n)$
- Draw new value of  $\boldsymbol{\Phi}^*$  from  $p(\boldsymbol{\Phi}|\boldsymbol{\Theta}^*, D_n)$
- Iterate

“Standard” Gibbs sampling: MCMC

May need “creativity” in sampling  $\boldsymbol{\Phi}^*$ : Metropolis-Hastings, etc

Often “easy”: as in Autoregressive DLM

*Applications and Examples:*

Carter and Kohn (several 1993-1996)

- Gaussian mixture error models
- non-parametric function estimation, e.g., spectral methods

West and coauthors (1992-1996)

- Time-varying auto-regressions:  
Geological and Biomedical contexts
- Non-Gaussian error models
- Latent structure: “Hidden” quasi-periodicities
- Multiple time series: Factor models

**NON-LINEAR DYNAMIC MODELS:**

Data equation:  $y_t \sim p(y_t | \boldsymbol{\theta}_t)$

State evolution:  $\boldsymbol{\theta}_t \sim p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$

e.g., discrete data models (multinomial time series; proportions)

non-linear stochastic volatility models (finance)

*Computational complications:*

- Exploit conditionally linear/normal structure
- “Side-by-side” simulations:  $(\boldsymbol{\theta}_t | \text{neighbours, data})$

*Carlin, Polson and Stoffer (1992) JASA*

*Scipione and Berliner (1993) ASA Proceedings (SBSS)*

*Shephard and coauthors (1996) Biometrika*

*Cargnoni, West and Müller (1997) JASA*

**SIMULATION: SEQUENTIAL ANALYSIS AND UPDATING**

Hard and unsolved problems

*TIME t :*

- States  $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_{t-k}, \dots, \boldsymbol{\theta}_t\}$  of interest
- Summarised via set of posterior samples:  $p(\boldsymbol{\Theta}_t | D_t)$

*TIME t + 1 :*

- Observe  $y_t \sim p(y_t | \boldsymbol{\theta}_t)$
- Require updated summary, posterior samples:  $p(\boldsymbol{\Theta}_{t+1} | D_t)$

*Issues:*

- Expanding/Changing “parameter” space and dimension
- Simulation-based summaries: discrete approximations
- New data may “conflict” with prior/predictions

*Needs:* Efficient ADAPTIVE and general algorithms

*First approaches:*

- Adaptive importance sampling and mixtures  
(West, Interface 92, JRSSB 93)
- Adaptive resampling  
(Lui and Wong, JASA 96; Berzuini, Gilks et al 97?)

## ADVANCED TOPIC: TIMING ISSUES

*Errors/uncertainties in timing of observations*

\* e.g. geological time series analysis ....  
geochemical measurements ← climate change  
ice-core isotopes, lake sediments, etc

*Times uncertain:*

measured *depths* in ice cores, sediment  
radio-carbon dating; synchronisation

truncation/rounding errors in times .. nearest year, etc

*TS data:*  $Y = \{y_i = y(t_i), i = 1, \dots, n\}$

*Timings:*  $T = \{t_1, \dots, t_n\}$

– *Include times T in analysis: Impute as latent variables*

“Standard” time series model:  $p(Y|T, \Theta)$

$\Theta$  = all state vectors and model parameters

Plus prior/timing models (perhaps complex):  $p(T)$

*Analysis:* accessible via MCMC methods

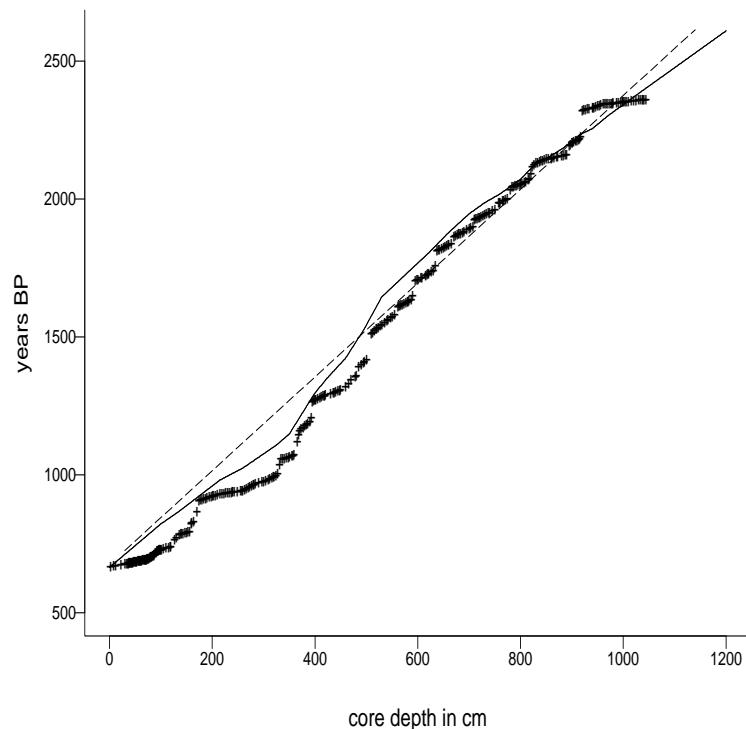
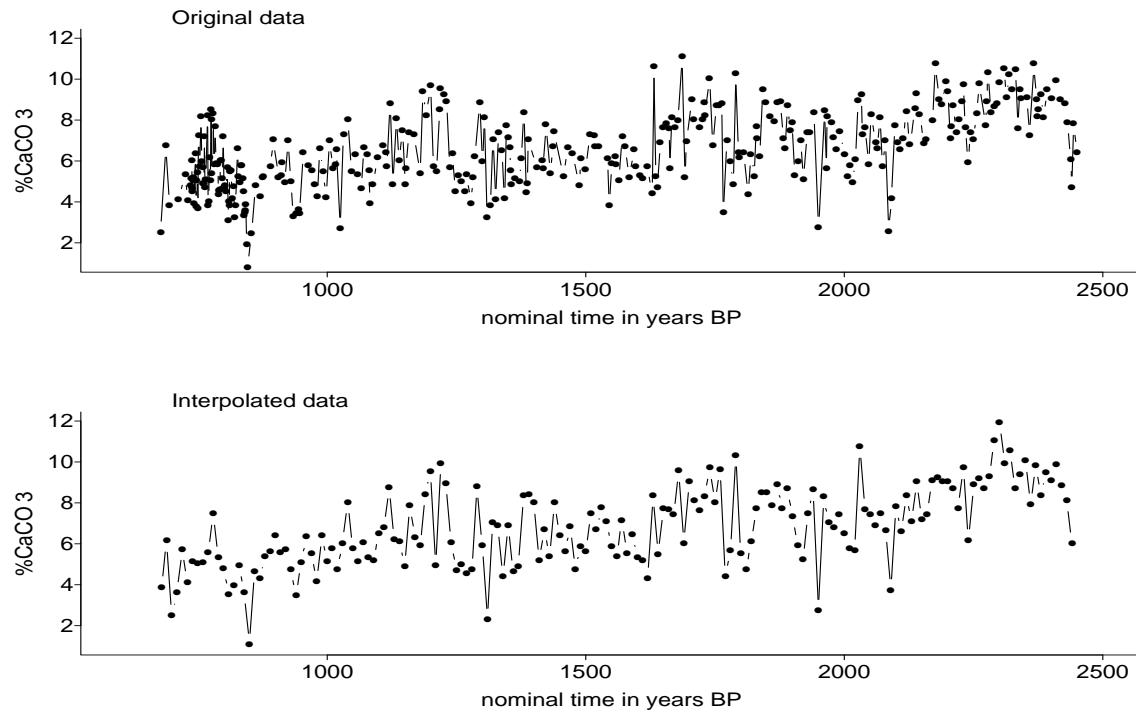
- $p(\Theta|Y, T)$  — “standard” analyses GIVEN times  $T$   
Gibbs simulations, samples for  $\Theta$
- $p(T|Y, \Theta)$  — non-standard:  
Metropolis-Hastings simulations of sequences of times

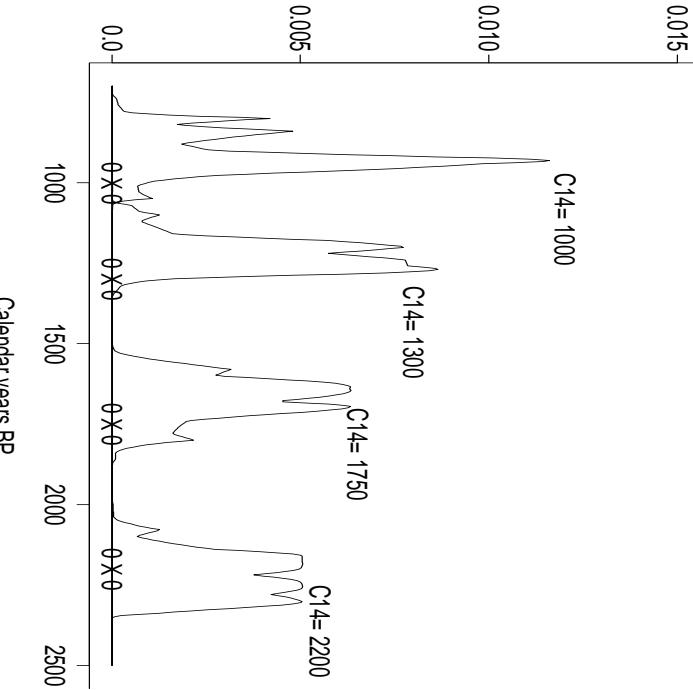
*Example:*

Carbonate records in deep lake sediments: Climatic change

West, *Bayesian Statistics 5* (1996)

- periodicities in Palaeoclimate signals: 2-3kyears
- trend+AR+noise component DLMs
- complicated calibration model  $p(T) : \text{core-depth} \rightarrow C_{14} \rightarrow T$
- limitations of  $C_{14}$  dating





... requires  $p(Y|T, \Theta)$  defined for all  $T$

- Underlying fine time scale,  $t_i$  in multiples of base unit  
(standard DLM “trick”; West and Harrison 1989,97)
- Continuous time models/SDEs: Observe in discrete time

General “time deformation” models

- stochastic “distortions” of time scale to “latent time”
- linear model in latent time: non-linear in real time
- model fitting/inference as above: MCMC in continuous/discrete framework

## CURRENT AND NEAR FUTURE

- Flavour of current directions evident in application areas
- Bayesians in FINANCE research and applications
  - portfolio, futures & options
  - econometric components
- Bayesians and BIOMEDICAL sciences
  - medical monitoring, sequential analysis
  - neurosciences (EEG, ECOG, networks), time series and imaging
- Growing connections with spatial statistics (environment, time-evolving imaging)

## CURRENT AND NEAR FUTURE (continued)

- *Models and Methods:*
  - Highly structured multiple time series: factor models, hierarchical models
    - possibly spatial connections in hierarchy
  - Large data sets: long, multiple time series: data reduction and summary
  - *Continuous time models, stochastic timing models, non-linear models, ...*
  - Computational methods: simulation & MCMC
    - non-linear dynamic models: MRF
    - Sequential updating of MCMC analyses

# Bayesian Analysis of Random Coefficient Logit Models Using Aggregate Data

Peter Rossi  
University of Chicago, Booth  
Jan 2009

Joint with Renna Jiang and Puneet Manchanda

# Introduction

- Often only aggregate sales/share data are available
- Berry, Levinsohn & Pakes (1995)
  - Introduce an aggregate logit model of products' market shares
    - Integrate over individual choice probabilities to obtain market shares
    - Need an aggregate error term to avoid a deterministic system after aggregation
  - Can handle endogeneity problems
  - Generalized Method of Moments (GMM)
    - Key step: invert shares to obtain mean utilities
- We conduct Bayesian inference for this model

# GMM

- Advantages
  - Fewer distributional assumptions
  - Easier to implement
- Disadvantages
  - Inefficiency
  - Inference for functions of model parameters(e.g., price elasticity, price-cost margin) of parameter estimates is difficult or computationally intensive
  - Numerical problems | multiple local maxima etc ...

# Bayes

- Advantages
  - Could be more efficient
  - Stochastic search can handle irregular criterion functions
  - Finite sample inference w/o resort to asymptotic approximations for all functions of model parms
- Disadvantages
  - Requires one additional distributional assumption (so what?)
  - Derivation of the likelihood
  - Reliable MCMC algorithm

# Other Bayesian approaches

- Chen and Yang (2003)
  - No aggregate demand shocks
- Musalem, Bradlow & Raju (2006)
  - Apply to situations where there is a fixed, known, and relatively small set of consumers over which aggregate demand is formed.
- Romeo (2007)
  - Use GMM criterion as the basis of a pseudo-likelihood

# Model

- Consumer  $i$ , product  $j$  ( $0, 1, \dots, J$ ), time  $t$

$$U_{ijt} = X_{jt}\theta_i + \eta_{jt} + \varepsilon_{ijt}$$

where  $X_{jt}$  is an observed product attribute

$\theta_i$  is a consumer specific coefficient

$\eta_{jt}$  unobservable (to researchers) aggregate demand shock

$\varepsilon_{ijt}$  iid Type I (maximum) Extreme Value (0,1)

- Assumptions
  - Consumers maximize current period utility
  - $\theta_i \sim MVN(\bar{\theta}, \Sigma)$
  - $\eta_{jt} \sim N(0, \tau^2)$

# Market Share

- Utility:  $U_{ijt} = X_{jt}\theta_i + \eta_{jt} + \varepsilon_{ijt}$

- Individual choice probability

$$s_{ijt} = \frac{\exp(X_{jt}\theta_i + \eta_{jt})}{1 + \sum_{k=1}^J \exp(X_{kt}\theta_i + \eta_{kt})}$$

- Market share

$$\begin{aligned}s_{jt} &= \int s_{ijt} d\Phi(\theta_i | \bar{\theta}, \Sigma) \\ &= \textcolor{magenta}{h}(\eta_t | \bar{\theta}, \Sigma, X_t)\end{aligned}$$

- $s_t$  inherits randomness from  $\eta_t = (\eta_{1t}, \dots, \eta_{Jt})'$

# Likelihood

- Can derive density of  $s_t$  from density of  $\eta_t$  using Change-of-Variable Theorem
- Density of  $\eta_t$  is

$$\phi\left(\textcolor{magenta}{h}^{-1}\left(s_t \mid \bar{\theta}, \Sigma, X_t\right) \mid \tau\right)$$

Normal pdf,  $\eta_{jt} \sim N(0, \tau^2)$

- Therefore, density of  $s_t$  is

$$\begin{aligned}\pi\left(s_t \mid \bar{\theta}, \Sigma, \tau, X_t\right) &= \phi\left(\textcolor{magenta}{h}^{-1}\left(s_t \mid \bar{\theta}, \Sigma, X_t\right) \mid \tau\right) J_{(\eta_t \rightarrow s_t)} \\ &= \phi\left(\textcolor{magenta}{h}^{-1}\left(s_t \mid \bar{\theta}, \Sigma, X_t\right) \mid \tau\right) \left(J_{(s_t \rightarrow \eta_t)}\right)^{-1}\end{aligned}$$

# Computing $h^{-1}$

- $h$  is defined by a system of  $J$  non-linear equations

$$s_{1t} = \int \frac{\exp(X_{1t}\theta_i + \eta_{1t})}{1 + \sum_{k=1}^J \exp(X_{kt}\theta_i + \eta_{kt})} d\Phi(\theta_i | \bar{\theta}, \Sigma)$$

⋮

$$s_{Jt} = \int \frac{\exp(X_{Jt}\theta_i + \eta_{Jt})}{1 + \sum_{k=1}^J \exp(X_{kt}\theta_i + \eta_{kt})} d\Phi(\theta_i | \bar{\theta}, \Sigma)$$

- BLP propose an iterative procedure which they prove to be a contraction mapping:

$$\eta_{jt}^{new} = \eta_{jt}^{old} + \ln(s_{jt}) - \ln(h(\eta_{1t}^{old}, \dots, \eta_{Jt}^{old} | \bar{\theta}, \Sigma))$$

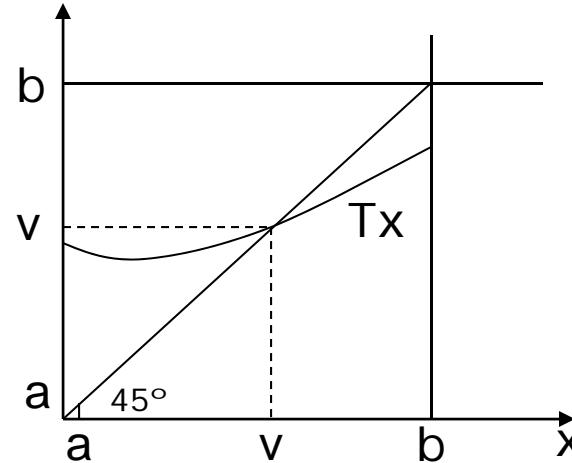
iterate until  $\eta_{jt}^{new}$  and  $\eta_{jt}^{old}$  are close “enough”

# Review of contraction mapping

DEFINITION Let  $T : S \rightarrow S$  be a function mapping  $S$  into itself.  
 $T$  is a **contraction mapping** if  $\|Tx - Ty\| < \|x - y\|$ , for all  $x, y \in S$ .

## CONTRACTION MAPPING THEOREM

If  $T$  is a contraction mapping, then  $T$  has **exactly one fixed point**  
 $v$  in  $S$  such that  $Tv = v$ .



In our model,  $T : R^J \rightarrow R^J$ . BLP prove that  $\|T\eta^{\text{new}} - T\eta^{\text{old}}\| < \|\eta^{\text{new}} - \eta^{\text{old}}\|$ ,  
i.e., we are guaranteed for a fixed point  $T\eta = \eta$ .

# Jacobian does not depend on $\bar{\theta}$ or $\tau^2$

- Jacobian:  $J_{(s_t \rightarrow \eta_t)} = \|\nabla_{\eta_t} s_t\|$

where  $\partial s_{jt} / \partial \eta_{kt} = \begin{cases} \int -s_{ijt} s_{ikt} d\Phi(\theta_i | \bar{\theta}, \Sigma) & \text{if } k \neq j \\ \int s_{ijt} (1 - s_{ijt}) d\Phi(\theta_i | \bar{\theta}, \Sigma) & \text{if } k = j \end{cases}$

- Re-write utility:  $U_{ijt} = X_{jt} \theta_i + \eta_{jt} + \varepsilon_{ijt}$   
 $= \underbrace{\left( X_{jt} \bar{\theta} + \eta_{jt} \right)}_{\mu_{jt}} + X_{jt} v_i + \varepsilon_{ijt}, \quad v_i \sim N(\mathbf{0}, \Sigma)$

- Elements in Jacobian:

$$\int f(s_{igt}) d\Phi(\theta_i | \bar{\theta}, \Sigma) = \int f \left( \frac{\exp(\mu_{jt} + X_{jt} v_i)}{1 + \sum_{k=1}^J \exp(\mu_{kt} + X_{kt} v_i)} \right) d\Phi(v_i | \mathbf{0}, \Sigma) = f(\mu_t, \Sigma | X_t)$$

- But given shares (and covariates),  $\mu_t$  is determined by  $\Sigma$
- Thus, given shares (and covariates), Jacobian is only a function of  $\Sigma$

# MCMC: overview

- Recall parameters are:  $\Sigma, \bar{\theta}, \tau^2$
- Re-parameterize  $\Sigma$  in terms of  $r$  (why?) :

$$\Sigma = U'U, \quad U = \begin{bmatrix} \exp(r_{11}) & r_{12} & \cdots & r_{1K} \\ 0 & \exp(r_{22}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{K-1,K} \\ 0 & \cdots & 0 & \exp(r_{KK}) \end{bmatrix}$$

- Priors

$$\bar{\theta} \sim MVN(\bar{\theta}_0, V_{\bar{\theta}})$$

$$r_{jj} \sim N(0, \sigma_j^2), \quad r_{jk} \sim N(0, \sigma_{off}^2), \quad j \neq k$$

$$\tau^2 \sim \nu_0 s_0^2 / \chi_{\nu_0}^2$$

- Posterior

$$\propto \underbrace{\prod_t \pi(s_t | \bar{\theta}, r, \tau, X_t)}_{\text{likelihood}} \times Prior(\bar{\theta}, r, \tau)$$

# More on the re-parameterization

$$\Sigma = U'U = \begin{bmatrix} \exp(2r_{11}) & r_{12} \exp(r_{11}) & r_{13} \exp(r_{11}) & r_{14} \exp(r_{11}) \\ & r_{12}^2 + \exp(2r_{22}) & r_{12}r_{13} + r_{23} \exp(r_{22}) & r_{12}r_{14} + r_{24} \exp(r_{22}) \\ symmetric & & r_{13}^2 + r_{23}^2 + \exp(2r_{33}) & r_{13}r_{14} + r_{23}r_{24} + r_{34} \exp(r_{33}) \\ & & & r_{14}^2 + r_{24}^2 + r_{34}^2 + \exp(2r_{44}) \end{bmatrix}$$

- Priors:  $r_{jj} \sim N(0, \sigma_j^2)$ ,  $r_{jk} \sim N(0, \sigma_{off}^2)$ ,  $j \neq k$
- Implied prior variance and mean on diagonals of  $\Sigma$  are ( $k=1,2,3,4$ ):

$$\begin{aligned} \text{var}[\Sigma_{kk}] &= 2(k-1)\sigma_{off}^4 + \exp(8\sigma_k^2) - \exp(4\sigma_k^2) \\ E[\Sigma_{kk}] &= (k-1)\sigma_{off}^2 + \exp(2\sigma_k^2) \end{aligned}$$

- Implied prior variance and mean on off-diagonals of  $\Sigma$  are ( $k=1,2,3$ ):

$$\begin{aligned} \text{var}[\Sigma_{k,(k+1)}] &= (k-1)\sigma_{off}^4 + \sigma_{off}^2 \exp(2\sigma_k^2) \\ E[\Sigma_{k,(k+1)}] &= 0 \end{aligned}$$

- Goal: let diagonals of  $\Sigma$  have same prior variance

So, set  $\sigma_1^2 = 0.50666596$ ,  $\sigma_{off}^2 = 1.00001292477193$ ,  $\sigma_2^2 = 0.5019265$ ,  $\sigma_3^2 = 0.4969934$ ,  $\sigma_4^2 = 0.4918498$

$$\Rightarrow \text{var}[\Sigma_{11}] = \text{var}[\Sigma_{22}] = \text{var}[\Sigma_{33}] = \text{var}[\Sigma_{44}] = 50$$

$$\text{var}[\Sigma_{12}] = 2.7548, \text{var}[\Sigma_{23}] = 3.7288, \text{var}[\Sigma_{34}] = 4.7021$$

$$E[\Sigma_{11}] = 2.7548, E[\Sigma_{22}] = 3.7288, E[\Sigma_{33}] = 4.7020, E[\Sigma_{44}] = 5.6744$$

$$E[\Sigma_{12}] = E[\Sigma_{23}] = E[\Sigma_{34}] = 0$$

# MCMC algorithm

Two sets of conditional draws

$$\begin{aligned} r | \bar{\theta}, \tau^2, \{s_t, X_t\}_{t=1}^T, \sigma_{r\_diag}^2, \sigma_{r\_off}^2 \\ \bar{\theta}, \tau^2 | r, \{s_t, X_t\}_{t=1}^T, \bar{\theta}_0, V_{\bar{\theta}}, \nu_0, s_0^2 \end{aligned}$$

1. Random-Walk Metropolis Chain to propose for  $r$

$$r^{new} = r^{old} + MVN(\mathbf{0}, \sigma^2 D_r)$$

2. Gibbs sampler for  $\bar{\theta}$  and  $\tau^2$

- A univariate Bayes regression:

$$\mu_{jt} = X_{jt} \bar{\theta} + \eta_{jt}, \quad \eta_{jt} \sim N(0, \tau^2)$$

# MCMC part of the code...

```
# ----- (1) Gibbs Sampler for thetabar and taosq -----
output=runiregG(y=mu,X=X,XpX=XpX,Xpy=crossprod(X,mu),sigmasq=taosq,
                  A=Athetabar,betabar=thetabar0,nu=nu0,ssq=s0sq)

thetabar=output$betadraw
taosq=output$sigmasqdraw

# ----- (2) Metropolis for r -----
# Random-Walk Chain
rN=r+rnorm(1,rep(0,(K*(K+1)/2)),varn_r)

ON=Loglhd(rN,mu,thetabar,taosq)
prior_old=sum(-r[1:K]^2/2/sigmasqR_DIAG)+sum(-r[(K+1):(K*(K+1)/2)]^2/2/sigmasqR_off)
prior_new=sum(-rN[1:K]^2/2/sigmasqR_DIAG)+sum(-rN[(K+1):(K*(K+1)/2)]^2/2/sigmasqR_off)

# Evaluate old r (mu) at new (thetabar,taosq)
eta=mu-X%*%thetabar
llhd_old=sum(log(dnorm(eta,sd=sqrt(taosq))))+OO$sumlogjacob

ratio=exp(ON$llhd+prior_new-llhd_old-prior_old)
alphaS=min(1,ratio) # S stands for Sigma
if (runif(1)<=alphaS) {
  r=rN; OO=ON; ns=ns+1; mu=OO$mu
}
```

# Brute-Force log-likelihood code...

```

LogLhd_slow = function(theta_bar,r,taosq,mu){
  # Purpose: Evaluate log likelihood. Sigma is re-parameterized as r.

  # (1). Transform r to L, where Sigma=LL'
  L=diag(exp(r[1:K]))
  L[lower.tri(L)]=r[(K+1):(K*(K+1)/2)]

  # (2). At given L, do inversion to get mu. Then compute eta
  temp=invert_slow(L,mu,v,crit,T,H,J,lnactS,indTHJ,indJTH)
  mu = temp$mu; prob = temp$prob; niter = temp$niter
  eta=mu-X%*%theta_bar

  # (3). Jacobian
  # Form J diagonal elements at each time t
  diagonal=rowMeans(prob*(1-prob)) # TJ by 1 vector

  # Form the off diagonal elements
  dd=-prob%*%t(prob)/H # TJ by TJ
  cc=aaa*dd+diag(diagonal)#TJ by TJ matrix: block diagonal

  cct= $\nabla_{\eta_t} s_t$ 
  for (t in 1:T){
    cct=cc[((t-1)*J+1):(t*J),((t-1)*J+1):(t*J)] #(t)th block of cc
    logjacob[t]=-log(abs(det(cct)))
  }

  # (4). Form Log Likelihood
  sumlogjacob=sum(logjacob)
  llhd=sum(log(dnorm(eta,sd=sqrt(taosq))))+sumlogjacob

  list(llhd=llhd,mu=mu,niter=niter,sumlogjacob=sumlogjacob)
}

```

# Slow inversion code

```

invert_slow =
function(L,mu,v,crit,T,H,J,lnactS,indTHJ,indJTH){

# Purpose: Invert observed shares S at give L to get mean utility mu's.

niter=0 # number of iterations taken for the inversion
munew=mu
# starting value
muold=munew/2
upart=X%*%L%*%v
while (max(abs((muold-munew)/munew))>crit){

  muold=munew
  num=exp(upart+ muold) # JT by H numerator
  den1=matrix(double(T*H),nrow=T)
  for (t in 1:T){
    den1[t,]=1+colSums(num[((t-1)*J+1):(t*J),]) #T by H
  }
  den=matrix(rep(den1,each=J),ncol=H) #replc each t J times, JT by H
  prob=num/den # JT by H
  sh=t(matrix(rowMeans(prob), nrow=J)) # T by J predicted share
  munew=t(matrix(muold,nrow=J))+log(S)-log(sh) # T by J
  munew=as.vector(t(munew)) # length JT vector
  niter=niter+1
}
List(mu=munew,prob=prob,niter=niter)
}

```

$$1 + \sum_{k=1}^J \exp(X_{kt} \theta_h + \eta_{kt})$$

# profile it (on 1000 iterations)...

\$by.total

	total.time	total.pct	self.time	self.pct
<b>rRCLogit_slow</b>	<b>663.68</b>	<b>100.0</b>	<b>0.08</b>	<b>0.0</b>
<b>Log1hd_slow</b>	<b>660.68</b>	<b>99.5</b>	<b>7.96</b>	<b>1.2</b>
<b>invert_slow</b>	<b>513.98</b>	<b>77.4</b>	<b>46.46</b>	<b>7.0</b>
colSums	262.24	39.5	127.60	19.2
exp	98.34	14.8	98.34	14.8
is.data.frame	81.74	12.3	3.58	0.5
inherits	78.16	11.8	69.46	10.5
matrix	74.30	11.2	22.68	3.4
as.vector	52.18	7.9	43.52	6.6
diag	49.26	7.4	8.74	1.3
array	40.30	6.1	40.26	6.1
%*%	30.04	4.5	30.04	4.5
+	29.24	4.4	29.24	4.4
prod	28.52	4.3	26.04	3.9
log	25.24	3.8	7.30	1.1
/	24.14	3.6	24.14	3.6
<b>det</b>	<b>17.48</b>	<b>2.6</b>	<b>2.02</b>	<b>0.3</b>
*	14.14	2.1	14.14	2.1
determinant	11.62	1.8	1.98	0.3
t	11.36	1.7	1.68	0.3
determinant.matrix	9.64	1.5	2.74	0.4
rowMeans	7.70	1.2	6.42	1.0
:	7.56	1.1	7.56	1.1
-	6.50	1.0	6.50	1.0
<	5.62	0.8	5.62	0.8
>	4.56	0.7	4.56	0.7
\$	3.34	0.5	3.34	0.5
storage.mode<--	2.96	0.4	0.84	0.1
.Call	2.30	0.3	2.30	0.3
storage.mode	2.04	0.3	1.14	0.2
length	1.78	0.3	1.78	0.3

# Make it faster: Vectorize it!

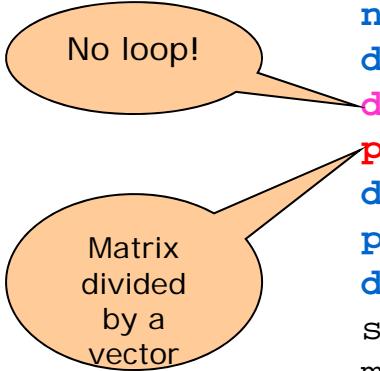
```
invert =
function(L,mu,v,crit,T,H,J,lnactS,indTHJ,indJTH){

# Purpose: Invert observed shares S at give L to get mean utility mu's.

niter=0                                # number of iterations taken
munew=mu                                 # starting value
muold=munew/2
upart=X%*%L%*%v
while (max(abs((muold-munew)/munew))>crit){

    muold=munew
    num=exp(upart+ muold)                 # num is JT x H
    dim(num)=NULL
    num=num[indTHJ]                      # convert num to JTH vector
    dim(num)=c(T*H,J)                  # convert num to THJ vector
    den=1+rowSums(num)                   # convert num to TH * J matrix
    # TH vector
    prob=num/den                         # TH * J matrix
    dim(prob)=NULL
    prob=prob[indJTH]                    # convert prob to THJ vector
    dim(prob)=c(J*T,H)                # convert prob to JTH vector
    # convert prob to JT * H matrix
    sh=rowMeans(prob)                   # JT vector
    munew=muold+lnactS-log(sh)          # JT vector
    niter=niter+1

}
list(mu=munew,prob=prob,niter=niter)
}
```



# More on the re-indexing function

```
JTH_THJ=function(J,H,T){  
#  
# function to convert and index a vector ordered j by t by h (i.e. j  
# varies faster than t than h) into a vector ordered t by h by j  
#  
ind=double(J*H*T)  
cnt=1  
for (j in 1:J){  
  for (h in 1:H) {  
    for (t in 1:T) {  
      ind[cnt]=(t-1)*J+(h-1)*(T*J)+j  
      cnt=cnt+1  
    }  
  }  
}  
return(ind)  
}
```

- Similarly, THJ\_JTH is a function that converts and indexes a vector ordered t by h by j (i.e. t varies faster than h than j), into a vector ordered j by t by h.
- Pre-compute the two indices:

```
indTHJ=JTH_THJ(J,H,T)  
indJTH=THJ_JTH(J,H,T)
```

# Eliminate Determinants

- Work on algebra to eliminate “det”:

$$cct = U'U$$

$$|cct| = |U| |U| = \left( \prod_{j=1}^J \text{diag}(U) \right)^2$$

$$\begin{aligned} \log(|cct|^{-1}) &= -\log \left( \prod_{j=1}^J \text{diag}(U) \right)^2 \\ &= -2 \sum_{j=1}^J \log(\text{diag}(U)) \end{aligned}$$

- New code:

```
for (t in 1:T){  
    cct=cc[((t-1)*J+1):(t*J),((t-1)*J+1):(t*J)] #(t)th block of cc  
    logjacob[t]=-2*sum(log(diag(chol(cct))))  
  
    # old code:  
    # logjacob[t]=-log(abs(det(cct)))  
}
```

# profile it again (on 1000 iterations)...

\$by.total

	total.time	total.pct	self.time	self.pct
<b>rRCLogit</b>	<b>282.08</b>	<b>100.0</b>	<b>0.04</b>	<b>0.0</b>
<b>Loglhd</b>	<b>279.32</b>	<b>99.0</b>	<b>14.36</b>	<b>5.1</b>
<b>invert</b>	<b>170.66</b>	<b>60.5</b>	<b>47.84</b>	<b>17.0</b>
exp	91.66	32.5	91.66	32.5
/	27.22	9.6	27.22	9.6
diag	23.56	8.4	10.76	3.8
log	20.88	7.4	3.96	1.4
crossprod	19.98	7.1	19.30	6.8
sum	19.38	6.9	0.92	0.3
*	12.74	4.5	12.74	4.5
chol	11.12	3.9	2.50	0.9
-	10.46	3.7	10.46	3.7
rowSums	8.88	3.1	8.42	3.0
diag<-	7.12	2.5	7.02	2.5
rowMeans	6.04	2.1	4.82	1.7
as.matrix	5.26	1.9	4.58	1.6
+	4.72	1.7	4.72	1.7
.Call	1.74	0.6	1.74	0.6
nrow	0.96	0.3	0.86	0.3
is.data.frame	0.96	0.3	0.04	0.0
inherits	0.94	0.3	0.24	0.1
t	0.76	0.3	0.14	0.0
as.matrix.default	0.66	0.2	0.58	0.2
t.default	0.62	0.2	0.62	0.2
mvrnorm	0.62	0.2	0.04	0.0
%*%	0.58	0.2	0.58	0.2
min	0.56	0.2	0.52	0.2
:	0.52	0.2	0.52	0.2
JTH_THJ	0.52	0.2	0.44	0.2
THJ_JTH	0.48	0.2	0.42	0.1
runiregG	0.46	0.2	0.02	0.0

Savings  
of  
57.5%!

# GMM

- Berry, Levinsohn & Pakes (1995):

$$\mu_{jt} = X_{jt}\bar{\theta} + \eta_{jt}$$

- Theoretical moments :

$$E[Z_t' \eta_t] = 0$$

- Sample analog:

$$\hat{m}_T(\bar{\theta}, \Sigma) = \frac{1}{T} \sum_{t=1}^T Z_t' (\hat{\mu}_t(\Sigma) - X_t \bar{\theta})$$

- GMM objective:

$$\min_{\bar{\theta}, \Sigma} \hat{m}_T(\bar{\theta}, \Sigma)' A^{-1} \hat{m}_T(\bar{\theta}, \Sigma)$$

- GMM search can be limited to only  $\Sigma$ , because  $\bar{\theta}$  can be concentrated out as follows,

$$\hat{\bar{\theta}}(\Sigma) = (X' Z A^{-1} Z' X)^{-1} X' Z A^{-1} Z' \hat{\mu}(\Sigma)$$

# Implementing GMM

- How to form  $Z$ ?
  - Total # of parameters for GMM:  $\text{dim}(\theta) + \text{dim}(r)$ .
  - Form  $Z$  by expanding exogenous variables in  $X$  or other instruments into polynomials, exponentials, logarithms, and interact with brand intercepts
- How to form  $A$ ? Two-step GMM
  - Step 1: Let  $A = \frac{1}{T} \sum_{t=1}^T Z_t' Z_t$   
Minimize the GMM objective=>obtain the residuals  $\hat{\eta}_{jt}^{(1)}$ , (1) stands for Step 1.
  - Step 2: Construct a new  $A = \frac{1}{T^2} \sum_{t=1}^T Z_t' \hat{\eta}_t^{(1)} \hat{\eta}_t^{(1)'} Z_t$   
Minimize the GMM objective again.  
After convergence, re-start the optimization routine from the converged estimates to ensure that the optimization converged.
- Take the final converged results=>  $\hat{\Sigma}_{GMM}, \hat{\theta}_{GMM}, \hat{\eta}_{ji}^{(2)}$ , and  $\hat{\tau}_{GMM} = sd(\hat{\eta}^{(2)})$

## GMM asymptotic standard errors

- Let  $\psi = (\bar{\theta}, r)$ , then

$$Var(\hat{\psi}_{GMM}) = \frac{1}{T} \left( D' \hat{V}^{-1} D \right)^{-1} \Big|_{\psi = \hat{\psi}_{GMM}}$$

- where  $D$  is a matrix of derivatives of the GMM criterion,  $g(\cdot)$ , w.r.t. to the parameters.  $\hat{V}$  is a consistent estimate of the variance of  $\hat{m}_T$ .

$$D = \begin{bmatrix} \frac{\partial m_T}{\partial \bar{\theta}} & \frac{\partial m_T}{\partial r} \end{bmatrix}$$

where

$$\frac{\partial m_T}{\partial \bar{\theta}} = -\frac{1}{T} \sum_t Z_t' X_t$$

and

$$\frac{\partial m_T}{\partial r} = \frac{1}{T} \sum_t Z_t' \left( \frac{\partial \hat{\mu}_t}{\partial r} \right)$$

The derivatives of mean utility w.r.t. to  $r$  are computed numerically.

# A Sampling Experiment

- $J = 3$  brands, one outside good
- $X$  contains 3 brand intercepts and a uniform[0,1] distributed attribute
- $T = 300$  time periods
- Base parameters

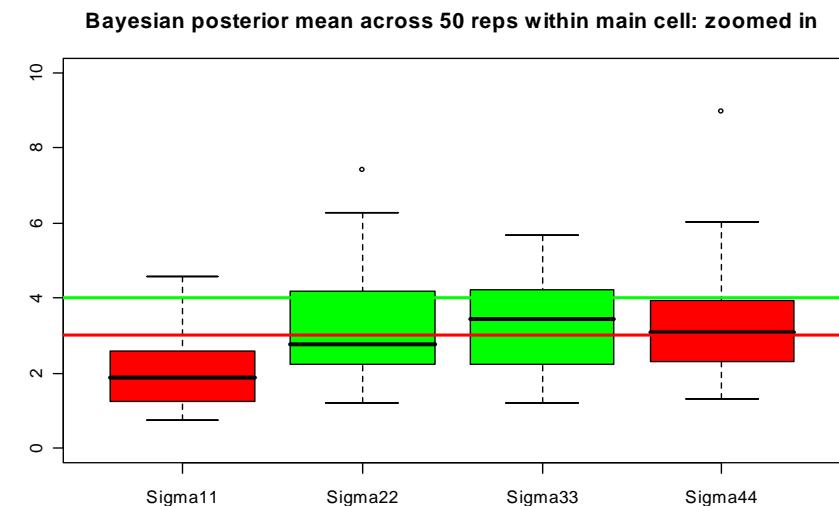
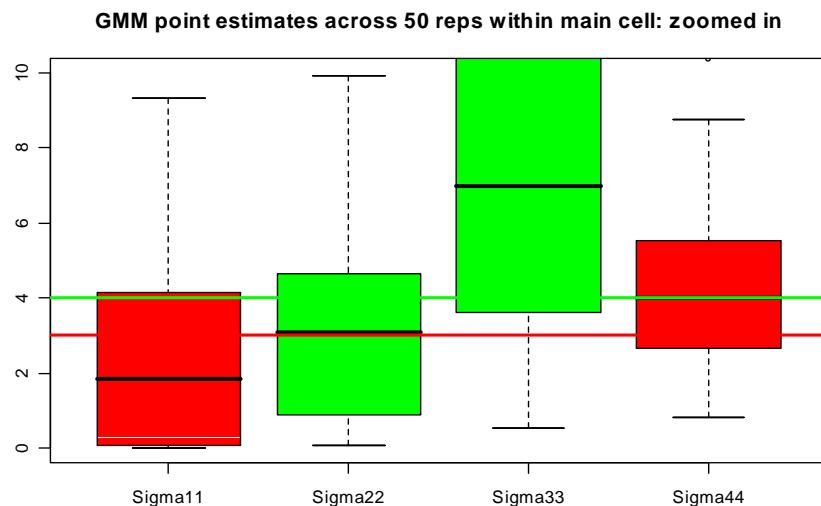
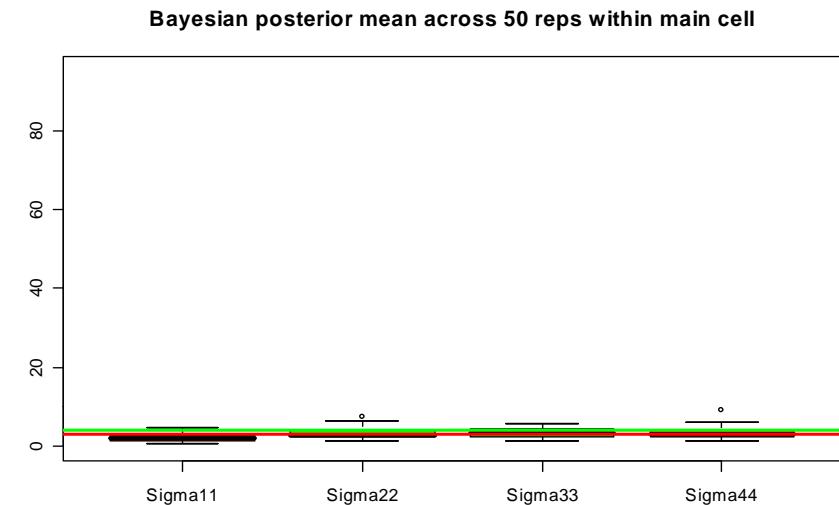
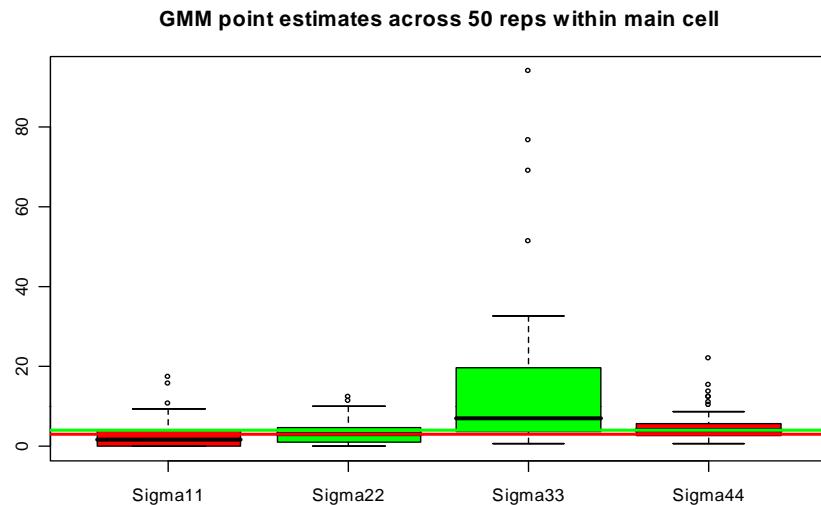
$$\bar{\theta}_{base} = (-2, -3, -4, -5)$$

$$\Sigma_{base} = \begin{bmatrix} 3 & 2 & 1.5 & 1 \\ 4 & -1 & 1.5 & \\ 4 & -0.5 & & \\ & & 3 & \end{bmatrix}$$

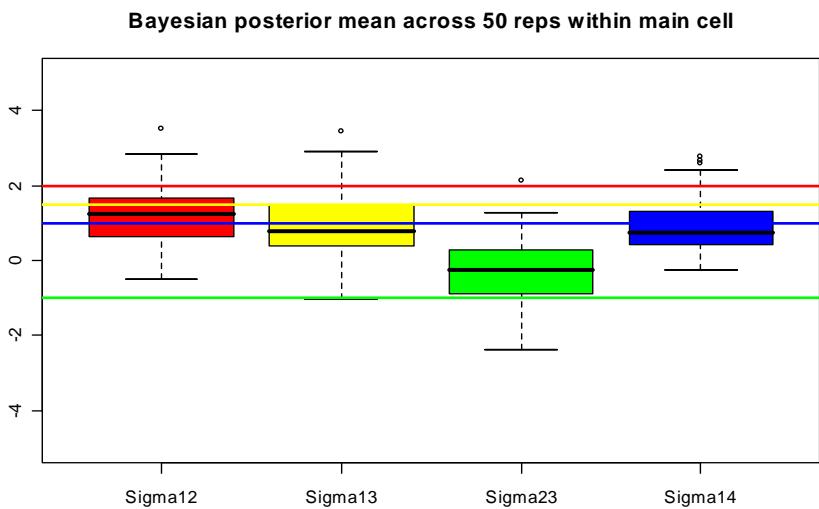
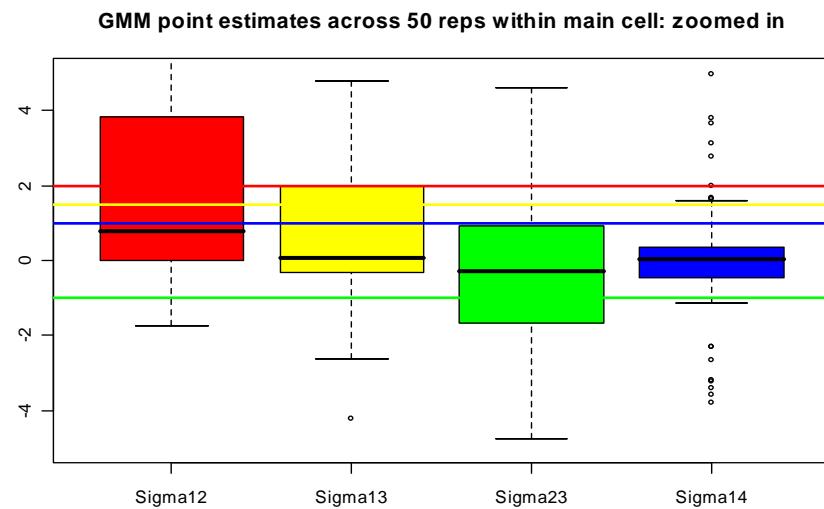
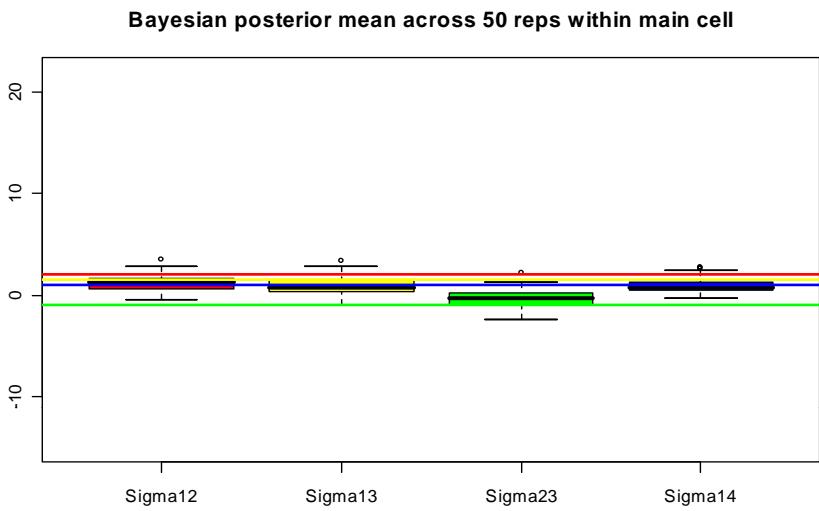
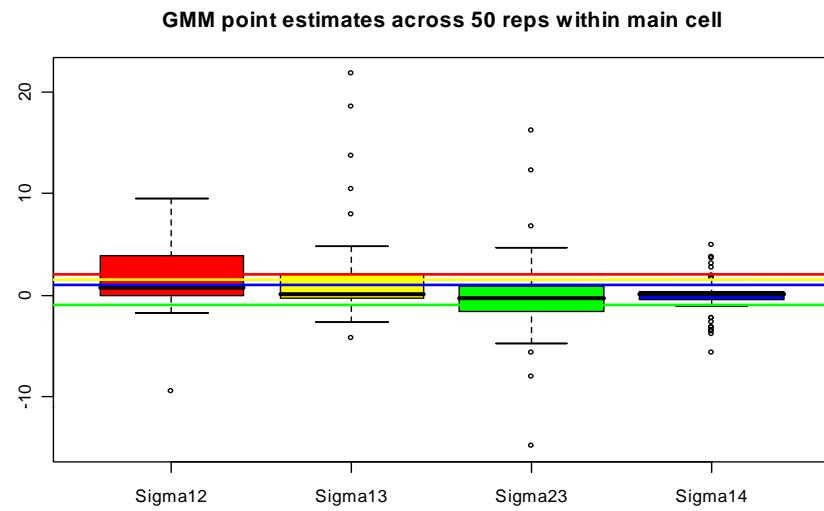
$$\tau_{base}^2 = 1$$

- Generate 50 datasets
  - E.g, in one dataset, implied average shares (stdev) over time for each brand:  
8.0% (8.5%), 5.1% (4.8%), 1.7% (2.6%)

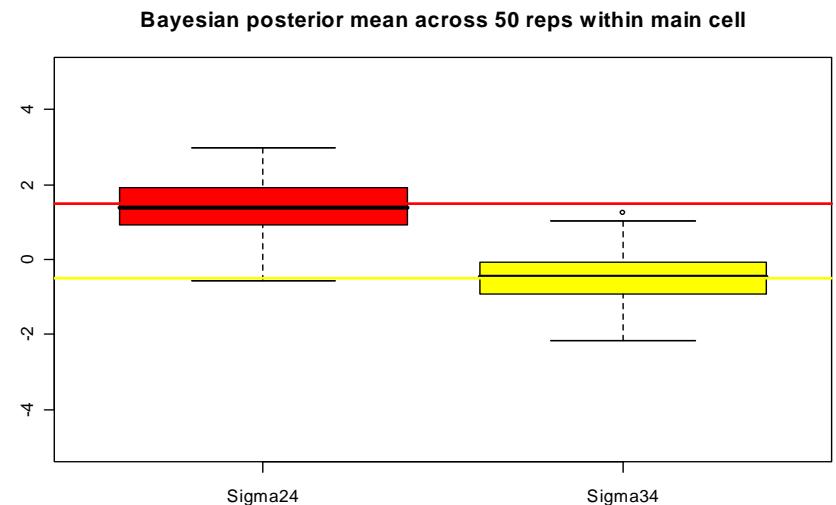
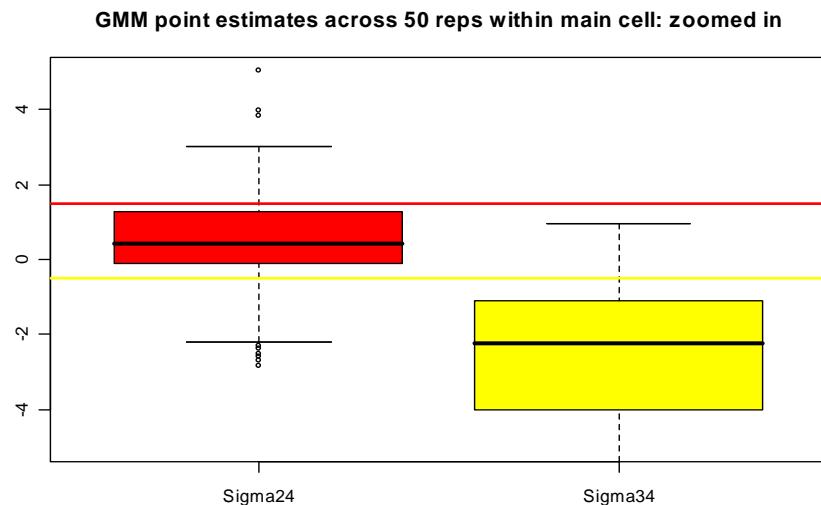
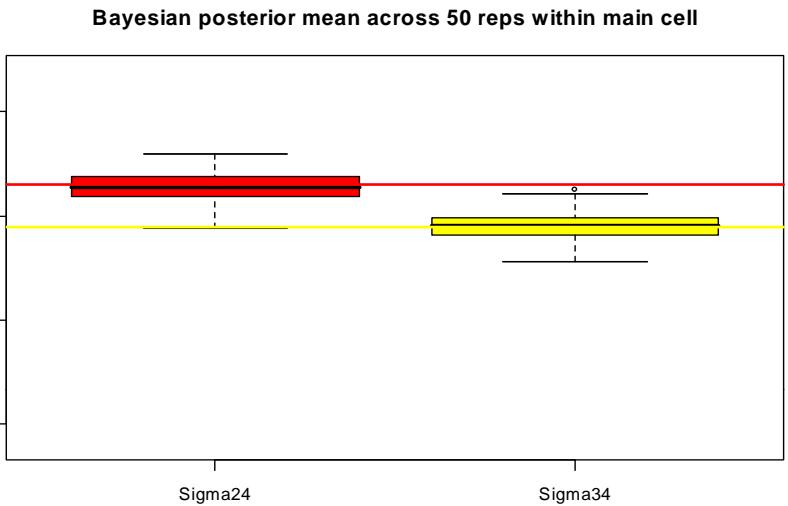
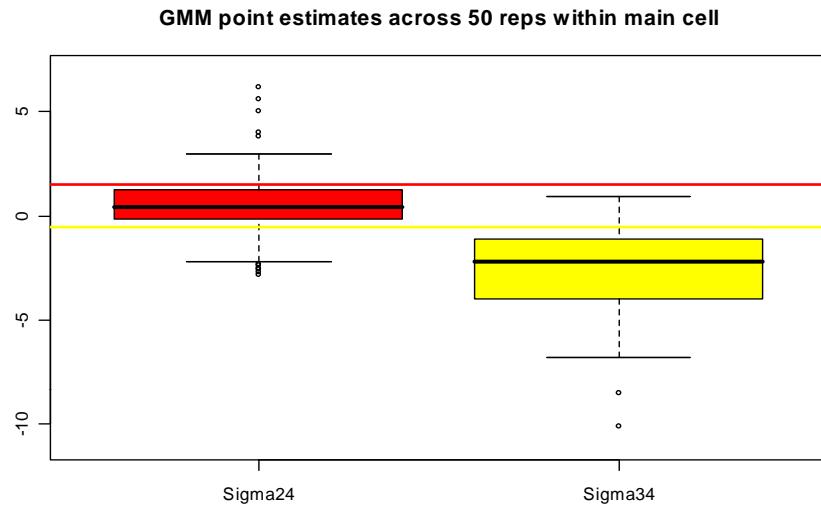
# Results: diagonals of $\Sigma$



# Results: off-diagonals of $\Sigma$

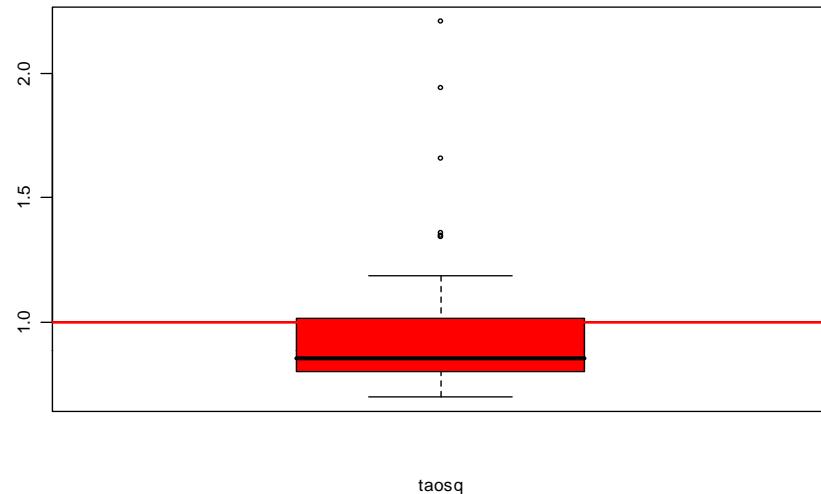


# Results: off-diagonals of $\Sigma$

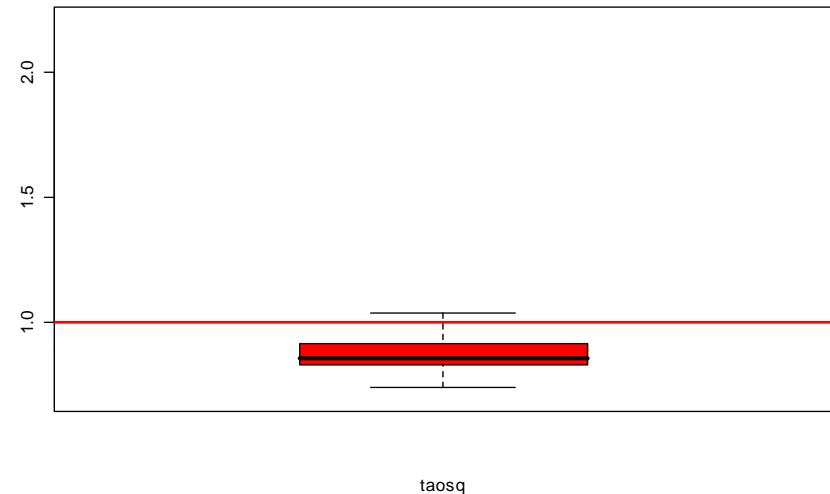


# Results: $\tau^2, \bar{\theta}$

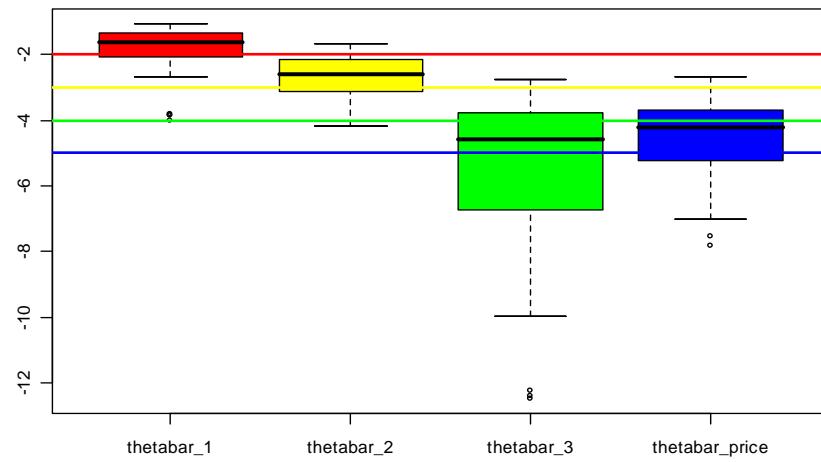
GMM point estimates across 50 reps within main cell



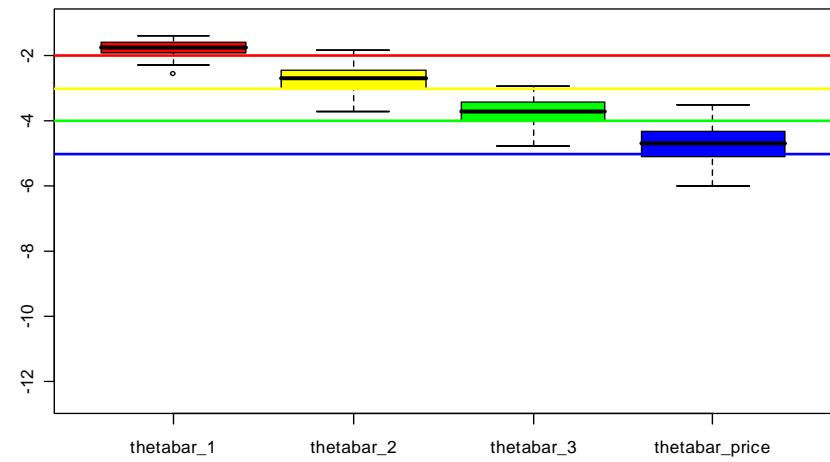
Bayesian posterior mean across 50 reps within main cell



GMM point estimates across 50 reps within main cell



Bayesian posterior mean across 50 reps within main cell



# MSE and bias

	MSE		Bias	
	Bayes	GMM	Bayes	GMM
$\tau^2$	0.02	0.09	-0.13	-0.03
$\bar{\theta}_1$	0.11	0.54	0.22	0.13
$\bar{\theta}_2$	0.26	0.54	0.25	0.29
$\bar{\theta}_3$	0.25	8.51	0.27	-1.51
$\bar{\theta}_{price}$	0.41	1.71	0.28	0.47
$\Sigma_{11}$	1.94	14.89	-1.04	0.13
$\Sigma_{22}$	2.63	9.52	-0.70	-0.46
$\Sigma_{33}$	1.95	498.86	-0.62	10.68
$\Sigma_{44}$	2.23	21.73	0.45	2.19
$\Sigma_{12}$	1.24	10.02	-0.75	-0.29
$\Sigma_{13}$	1.15	23.49	-0.56	0.35
$\Sigma_{23}$	1.41	19.77	0.67	0.86
$\Sigma_{14}$	0.54	5.13	-0.06	-1.09
$\Sigma_{24}$	0.63	5.19	-0.08	-0.91
$\Sigma_{34}$	0.52	10.05	0.02	-2.22

GMM: 0.7 hr/rep, Bayes: 2.2 hr/rep (Pentium 4, CPU 3 GHz, 1GB RAM)

# MSE and bias

(increase number of heterogeneity draws H=200)

	MSE				Bias			
	Bayes		GMM		Bayes		GMM	
	H=50	H=200	H=50	H=200	H=50	H=200	H=50	H=200
$\tau^2$	0.02	0.016	0.09	0.082	-0.13	-0.109	-0.03	-0.042
$\bar{\theta}_1$	0.11	0.10	0.54	0.42	0.22	0.13	0.13	0.1
$\bar{\theta}_2$	0.26	0.12	0.54	0.65	0.25	0.13	0.29	0.2
$\bar{\theta}_3$	0.25	0.28	8.51	3.04	0.27	0.33	-1.51	-0.67
$\bar{\theta}_{price}$	0.41	0.30	1.71	1.97	0.28	0.26	0.47	0.78
$\Sigma_{11}$	1.94	1.58	14.89	12.3	-1.04	-0.82	0.13	-0.12
$\Sigma_{22}$	2.63	1.58	9.52	17.12	-0.70	-0.72	-0.46	-0.18
$\Sigma_{33}$	1.95	2.47	498.86	66.53	-0.62	-0.98	10.68	3.34
$\Sigma_{44}$	2.23	0.73	21.73	15.37	0.45	-0.01	2.19	1.37
$\Sigma_{12}$	1.24	0.72	10.02	10.76	-0.75	-0.49	-0.29	0.15
$\Sigma_{13}$	1.15	0.89	23.49	9.61	-0.56	-0.63	0.35	-0.01
$\Sigma_{23}$	1.41	1.11	19.77	13.63	0.67	0.71	0.86	1.67
$\Sigma_{14}$	0.54	0.28	5.13	5.18	-0.06	-0.13	-1.09	-1.52
$\Sigma_{24}$	0.63	0.50	5.19	6.06	-0.08	-0.14	-0.91	-1.47
$\Sigma_{34}$	0.52	0.40	10.05	11.49	0.02	-0.05	-2.22	-2.09

# Address mis-specification concerns

- So far, we first generate iid  $\eta_{jt}$  from  $N(0, 1)$ , then conduct inferences assuming  $\eta_{jt} \sim N(0, \tau^2)$
- Now let's investigate the performance of our estimator in situations where the model is mis-specified.
- Specifically, we generate  $\eta_{jt}$  from as follows, then fit the model assuming  $\eta_{jt} \sim N(0, \tau^2)$ 
  - Conditional Heteroskedasticity
  - AR(1): 0.9
  - Asymmetric Beta distribution
  - Symmetric Beta distribution

# Mis-specification: Heteroskedasticity

- Instead of homoskedastic shocks, generate shocks from

$$\eta_{jt} \sim N(0, V_{jt} \equiv f(X_{jt}))$$

- To keep things comparable, we require  $E[V_{jt}] = 1$
- So,

$$1 = E[V_{jt}] = E[\exp(a + bP_{jt})] = \frac{\exp(a)}{b}(\exp(b) - 1)$$

where  $P_{jt} \sim Unif[0,1]$

- Set  $b=1$ , then  $a = -\log(\exp(1) - 1) \approx -0.5413$

## Mis-specification: AR(1)

- Instead of iid, generate shocks for product  $j$  ( $j=1,2,3$ ) according to

$$\eta_{j,t+1} = \rho\eta_{j,t} + u_{j,t+1} \quad u_{j,t} \stackrel{iid}{\sim} N(0, \sigma_u^2)$$

- We require  $\text{var}[\eta_{j,t}] = 1$

- So,

$$1 = \text{var}[\eta_{j,t}] = \frac{\sigma_u^2}{1 - \rho^2}$$

- Set  $\rho = 0.9$ , we get  $\sigma_u \approx 0.4359$

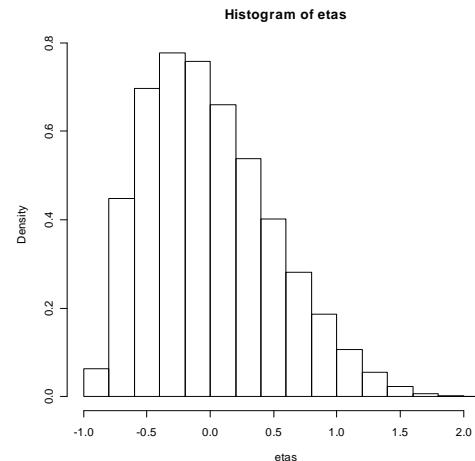
# Mis-specification: Asymmetric Beta

- Instead of Normal shocks, generate

$$\eta_{jt} \sim d \cdot \text{Beta}[\alpha, \beta] - c$$

- Set  $\alpha = 2, \beta = 5$  to get a logNormal-shaped asymmetric Beta distribution
- We require that  $\eta$  be centered at zero and  $sd[\eta_{jt}] = 0.5$
- So,  $0.5 = sd[\eta_{jt}] = d \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = d \sqrt{\frac{2 \times 5}{(2+5)^2(2+5+1)}}$   $\Rightarrow d \approx 3.1305$
- Finally, calculate the de-mean factor  $c$  to keep  $\eta$  centered at zero:

$$c = d \cdot \frac{\alpha}{\alpha + \beta} \approx 0.8944$$



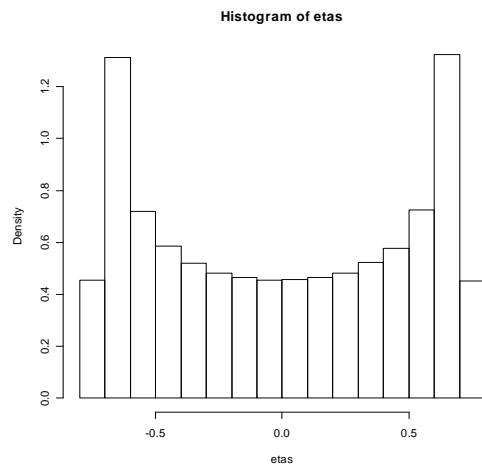
# Mis-specification: Symmetric Beta

- Instead of Normal shocks, generate

$$\eta_{jt} \sim d \cdot \text{Beta}[\alpha, \beta] - c$$

- Set  $\alpha = 0.5, \beta = 0.5$  to get a U-shaped symmetric Beta distribution
- We require that  $\eta$  centered at zero and  $sd[\eta_{jt}] = 0.5$
- Follow the same procedure as asymmetric Beta, we get

$$d \approx 1.4142, c \approx 0.7071$$



# MSE and bias in mis-specified cells

		MSE		Bias	
		Bayes	GMM	Bayes	GMM
$\Sigma_{11}$	iid N	1.94	14.89	-1.04	0.13
	Hetro	11.07	25.81	1.85	0.23
	AR1	3.91	35.43	-0.38	0.32
	AsyBeta	2.17	66.28	-1.22	1.20
$\Sigma_{22}$	SymBeta	2.14	8.49	-1.19	-1.05
	iid N	2.63	9.52	-0.70	-0.46
	Hetro	15.73	26.65	2.78	-0.33
	AR1	5.3	181.16	0.13	0.83
$\Sigma_{33}$	AsyBeta	4.00	87.09	-1.71	1.96
	SymBeta	2.34	38.38	-0.92	0.46
	iid N	1.95	498.86	-0.62	10.68
	Hetro	40.2	566.35	3.6	12.35
$\Sigma_{44}$	AR1	8.08	1927.91	0.50	15.95
	AsyBeta	3.47	601.03	-1.33	8.83
	SymBeta	3.04	163.88	-0.42	6.04
	iid N	2.23	21.73	0.45	2.19
$\Sigma_{12}$	Hetro	5.12	23.11	1.16	2.33
	AR1	5.41	24.05	1.44	2.40
	AsyBeta	0.71	64.72	-0.16	2.73
	SymBeta	2.42	21.58	0.50	1.91

# MSE and bias in mis-specified cells

		MSE		Bias	
		Bayes	GMM	Bayes	GMM
$\Sigma_{13}$	iid N	1.15	23.49	-0.56	0.35
	Hetro	3.85	53.12	0.40	1.00
	AR1	2.09	49.7	-0.63	0.50
	AsyBeta	0.50	10.28	-0.54	-0.77
	SymBeta	0.77	8.08	-0.37	-0.82
$\Sigma_{23}$	iid N	1.41	19.77	0.67	0.86
	Hetro	3.85	49.15	0.45	2.13
	AR1	3.86	173.38	0.83	4.04
	AsyBeta	1.50	15.91	0.99	1.15
	SymBeta	1.43	17.75	0.61	1.95
$\Sigma_{14}$	iid N	0.54	5.13	-0.06	-1.09
	Hetro	4.05	17.47	-1.34	-1.70
	AR1	1.13	8.85	0.16	-0.83
	AsyBeta	0.41	12.42	-0.16	-0.53
	SymBeta	0.78	4.95	0.19	-0.07
$\Sigma_{24}$	iid N	0.63	5.19	-0.08	-0.91
	Hetro	3.48	17.44	-1.02	-1.72
	AR1	1.42	9.47	0.08	-1.13
	AsyBeta	0.45	16.02	-0.30	-0.53
	SymBeta	0.73	8.15	0.22	0.10
$\Sigma_{34}$	iid N	0.52	10.05	0.02	-2.22
	Hetro	3.01	31.25	-1.08	-3.02
	AR1	1.16	28.4	0.07	-1.66
	AsyBeta	0.3	27.72	-0.03	-0.44
	SymBeta	0.74	10.43	0.09	-0.25

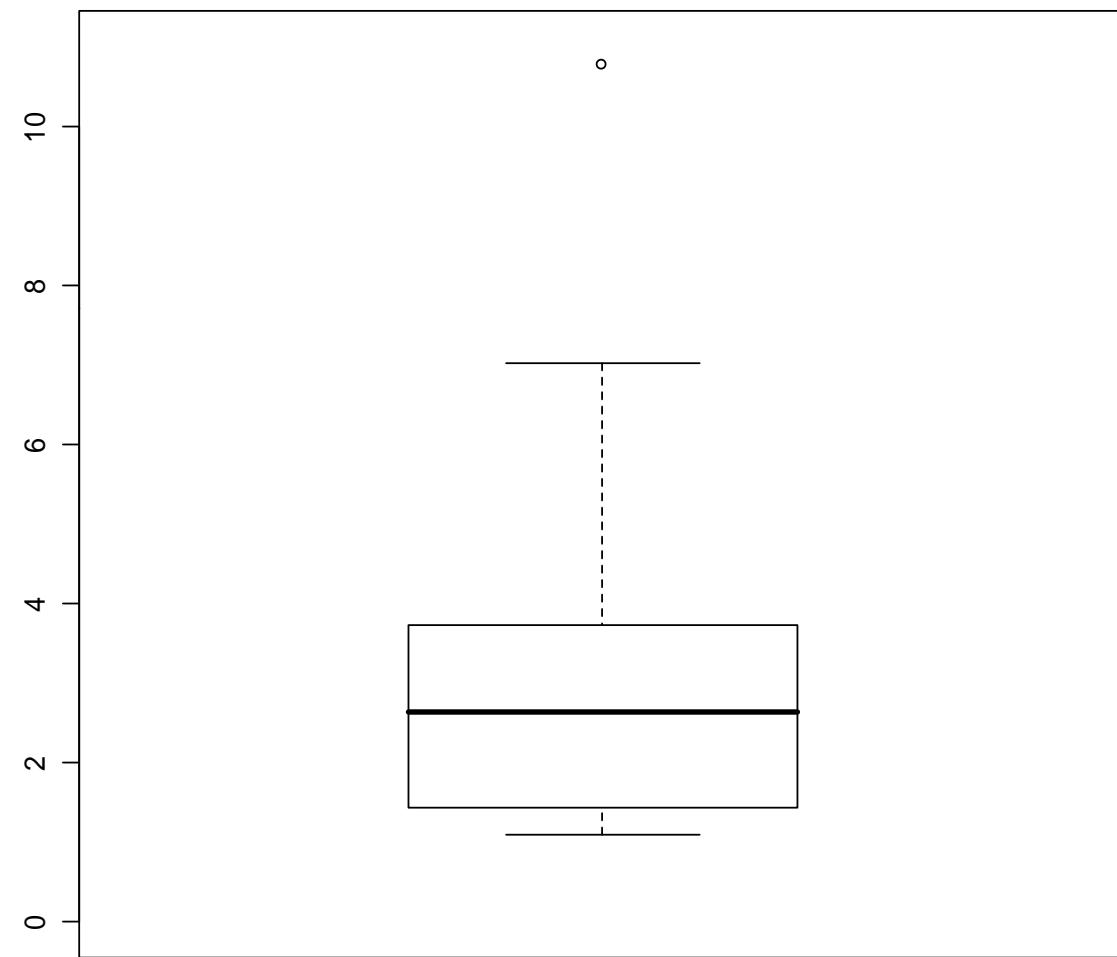
# MSE and bias in mis-specified cells

		MSE		Bias	
		Bayes	GMM	Bayes	GMM
$\tau^2$	iid N	0.02	0.09	-0.13	-0.03
	Hetro	0.009	0.134	-0.021	-0.011
	AR1	0.049	0.227	-0.172	-0.058
	AsyBeta	0.002	0.007	-0.044	-0.002
$\bar{\theta}_1$	SymBeta	0.001	0.006	-0.03	-0.01
	iid N	0.11	0.54	0.22	0.13
	Hetro	0.53	0.43	-0.87	0.18
	AR1	0.22	0.55	0.24	0.16
$\bar{\theta}_2$	AsyBeta	0.12	0.5	0.23	0.02
	SymBeta	0.17	0.29	0.31	0.33
	iid N	0.26	0.54	0.25	0.29
	Hetro	0.87	1.04	-0.52	0.25
$\bar{\theta}_3$	AR1	0.39	1.7	0.22	0.33
	AsyBeta	0.29	2.04	0.45	-0.14
	SymBeta	0.25	1.52	0.33	0.10
	iid N	0.25	8.51	0.27	-1.51
$\bar{\theta}_{price}$	Hetro	2.00	12.08	-0.93	-2.02
	AR1	0.84	10.92	0.14	-1.46
	AsyBeta	0.41	9.39	0.50	-1.11
	SymBeta	0.38	5.01	0.32	-1.04
	iid N	0.41	1.71	0.28	0.47
	Hetro	0.85	2.16	0.62	0.67
	AR1	0.59	2.39	-0.10	0.33
	AsyBeta	0.51	2.27	0.6	0.37
	SymBeta	0.34	2.48	0.23	0.29

# Summary of the five cells

- In base cell (iid N), Bayes estimator outperforms GMM estimator
  - GMM produces large MSE values for elements in  $\Sigma$
  - Even for the regression parameters, Bayes has an MSE  $\frac{1}{2}$  to  $\frac{1}{4}$  of GMM
- Performance of Bayes relative to GMM
  - declines somewhat for the Hetero cell,
  - but does not change much for any other mis-specified cells: AR(1), U-shaped Beta, Asymmetric Beta
- GMM's performance degrades in the presence of Hetero and non-Normality
- Bayes exhibits same or less level of bias as GMM
- GMM estimates of off-diagonal elements of Sigma tend to be biased toward zero, hence attenuate the correlation in the random coefficient distribution

# Ratio of sample standard deviation to median asymptotic standard errors (base cell)



# Coverage of confidence intervals

Frequency (out of 50 Reps in base cell) of true parameter values covered by the 95% interval:

	GMM +/- 1.96*Asymp. Standard Error	Bayes post. Mean +/- 1.96*Standard Deviation
$\bar{\theta}_1$	11	31
$\bar{\theta}_2$	21	35
$\bar{\theta}_3$	31	43
$\bar{\theta}_{price}$	25	41
r1	30	30
r2	31	43
r3	34	48
r4	39	45
r5	31	36
r6	41	43
r7	39	44
r8	35	41
r9	35	44
r10	37	45
<b>Total (Freq)</b>	<b>440</b>	<b>569</b>
<b>Total (Percentage)</b>	<b>440/(50*14)=62.86%</b>	<b>569/(50*14)=81.29%</b>

# Inclusion of instrumental variables

- Denote  $X_{jt} = \{W_{jt}, P_{jt}\}$ , where price  $P_{jt}$  is endogenous:

$$P_{jt} = Z_{jt}\delta + \xi_{jt} \quad \begin{pmatrix} \xi_{jt} \\ \eta_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega \equiv \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{pmatrix}\right)$$

- Joint density of price and share at time  $t$  is

$$\begin{aligned} \pi(P_t, s_t | \bar{\theta}, r, \delta, \Omega) &= \pi(\xi_t, \eta_t | \bar{\theta}, r, \delta, \Omega) J_{(P_t, s_t \rightarrow \xi_t, \eta_t)} \\ &= \pi(\xi_t, \eta_t | \bar{\theta}, r, \delta, \Omega) \left( J_{(P_t, s_t \rightarrow \xi_t, \eta_t)} \right)^{-1} \end{aligned}$$

where the Jacobian:

$$J_{(P_t, s_t \rightarrow \xi_t, \eta_t)} = \begin{vmatrix} \nabla_{\xi_t} P_t & \nabla_{\eta_t} P_t \\ \nabla_{\xi_t} s_t & \nabla_{\eta_t} s_t \end{vmatrix} = \begin{vmatrix} I & \mathbf{0} \\ \nabla_{\xi_t} s_t & \nabla_{\eta_t} s_t \end{vmatrix} = \|\nabla_{\eta_t} s_t\| = J_{(s_t \rightarrow \eta_t)}$$

- Two sets of conditionals

$$\begin{aligned} \bar{\theta}, \delta, \Omega | r, \{s_t, X_t, Z_t\}_{t=1}^T, \bar{\theta}_0, V_{\bar{\theta}}, \bar{\delta}, V_{\delta}, \nu_0, V_{\Omega} \\ r | \bar{\theta}, \delta, \Omega, \{s_t, X_t, Z_t\}_{t=1}^T, \sigma_{r\_diag}^2, \sigma_{r\_off}^2 \end{aligned}$$

# MSE and bias in IV cell

	MSE		Bias	
	Bayes	GMM	Bayes	GMM
$\bar{\theta}_1$	0.50	9.89	0.49	-0.93
$\bar{\theta}_2$	0.44	13.46	0.51	-1.28
$\bar{\theta}_3$	0.41	34.11	0.41	-2.16
$\bar{\theta}_{price}$	0.28	10	0.33	-0.02
$\Sigma_{11}$	3.82	315.49	-1.59	6.86
$\Sigma_{22}$	3.11	383.2	-1.51	8.74
$\Sigma_{33}$	3.68	6301.31	-1.30	19.09
$\Sigma_{44}$	0.75	104.68	-0.06	4.02
$\Sigma_{12}$	2.33	117.63	-1.24	1.59
$\Sigma_{13}$	1.64	82.45	-1.00	1.20
$\Sigma_{23}$	1.92	139.48	0.78	2.65
$\Sigma_{14}$	0.36	38.25	-0.25	-1.42
$\Sigma_{24}$	0.56	24.03	-0.32	-1.05
$\Sigma_{34}$	0.20	24.87	0.10	-1.89
$\delta_1$	0.002	0.002	0.003	0.001
$\delta_2$	0.002	0.002	0.002	0.001
$\delta_3$	0.002	0.002	-0.002	-0.003
$\delta_4$	0.004	0.004	-0.005	-0.002
$\Omega_{11}$	0.0002	0.0002	0.0012	-0.0003
$\Omega_{12}$	0.002	0.003	-0.02	-0.01
$\Omega_{22}$	0.03	1.28	-0.16	0.39
$Corr_{\Omega}$	0.003	0.008	-0.002	-0.062

# Summary of Sampling Experiments

Bayes outperforms GMM in all following sampling experiments (cells):

- Base cell :  $\eta_{jt} \sim N(0, \tau^2)$ 
  - Number of heterogeneity draws H=50 vs. 200
- Cells with mis-specified  $\eta$ :
  - Heteroskedasticity:  $\eta_{jt} \sim N(0, V_{jt} \equiv f(X_{jt}))$
  - AR(1): 0.9
  - Asymmetric Beta[2,5]: log-normal shaped
  - Symmetric Beta[0.5,0.5]: U-shaped
- Instrumental Variable cell:
  - corr(price shock, demand shock)=0.46,  
instruments explain 32% of total price variation

# Conclusions

- The impression that aggregate share models are hard to estimate is partly because of the method used, not the intrinsic property of the model
- Bayesian inference is dramatically more efficient than GMM
- Distribution of aggregate shocks
  - There is concern that Bayes makes more distributional assumptions than GMM. We found that it doesn't matter much
  - Straightforward to extend to a Mixture of Normals. Probably requires more data to identify the shape of this distribution

# **Multiparameter Models - Normal Data, Poisson Regression**

Statistics 220

Spring 2005



## Normal Inference Models - Semiconjugate Prior

Another popular prior is

$$\begin{aligned}\mu | \sigma^2 &\sim N(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

In this case,  $\mu$  and  $\sigma^2$  are independent apriori. This prior is useful when the prior information on  $\mu$  isn't thought of in terms of a number of prior measurements.

Note that this isn't a conjugate prior. The posterior is not the product of normal and Inv- $\chi^2$  densities. In fact the posterior is not particularly nice, in that parts of it do not reduce to standard densities.

$$p(\mu|\sigma^2, y):$$

Given that  $\sigma^2$  is fixed, this is a case we have already seen

$$\mu|\sigma^2, y \sim N(\mu_n, \tau_n^2)$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

This gives an idea where the term semi-conjugate comes from. If we consider the posterior distribution of one parameter conditional on the other parameters, the posterior is of the same form as the prior.

$$p(\sigma^2 | \mu, y):$$

Similarly

$$\sigma^2 | \mu, y \sim \text{Inv-}\chi^2 \left( \nu_0 + n, \frac{\nu_0 \sigma_0^2 + nv}{\nu_0 + n} \right)$$

where

$$v = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2$$

$$p(\sigma^2|y):$$

Here is where the nice distributional results breakdown.

$$\begin{aligned} \sigma^2|y &\propto \int \text{Inv}-\chi^2(\sigma^2|\nu_0, \sigma_0^2) N(\mu|\mu_0, \tau_0^2) \prod_{i=1}^n N(y_i|\mu, \sigma^2) d\mu \\ &\propto \text{Inv}-\chi^2\left(\sigma^2|\nu_0 + n, \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n}\right) \int N(\mu|\mu_0, \tau_0^2) N\left(\bar{y}|\mu, \frac{\sigma^2}{n}\right) d\mu \end{aligned}$$

Since the part inside the integral is proportional to a normal density, the density  $p(\sigma^2|y)$  can be calculated in closed form.

Unfortunately this isn't a standard density. However we can get a handle on it based on the fact

$$p(\sigma^2|y) = \frac{p(\mu, \sigma^2|y)}{p(\mu|\sigma^2, y)}$$

This comes directly from

$$p(\mu, \sigma^2 | y) = p(\mu | \sigma^2, y) p(\sigma^2 | y)$$

So

$$p(\sigma^2 | y) \propto \frac{N(\mu | \mu_0, \tau_0^2) \text{Inv} - \chi^2(\sigma^2 | \nu_0, \sigma_0^2) \prod_{i=1}^n N(y_i | \mu, \sigma^2)}{N(\mu | \mu_n, \tau_n^2)}$$

While it appears that this depends on  $\mu$ , it actually doesn't so we can pick any value of  $\mu$  to make computation as easy as possible. A good choice is to evaluate this at  $\mu = \mu_n$ , giving

$$p(\sigma^2 | y) \propto \tau_n N(\mu_n | \mu_0, \tau_0^2) \text{Inv} - \chi^2(\sigma^2 | \nu_0, \sigma_0^2) \prod_{i=1}^n N(y_i | \mu_n, \sigma^2)$$

$$p(\mu|y):$$

This is even uglier. While it appears that a closed form solution to the integral

$$p(\mu|y) \propto \int p(\mu, \sigma^2|y) d\sigma^2$$

is possible (the integrand is proportion to an inverse gamma density), this is usually handled by simulation. One approach is the two stage simulation approach mentioned before

1. Simulate  $\sigma_1^2, \dots, \sigma_m^2 \stackrel{iid}{\sim} \sigma^2|y$
2. Simulate  $\mu_i \sim \mu|\sigma_i^2, y = N(\mu_i, \tau_i^2)$

Step 1 could be done by an acceptance-rejection method or by the grid simulation approach discussed in the text.

An alternative approach would be to use a Gibbs sampler for both  $\mu$  and  $\sigma^2$ .

This approach will be taken for the example.

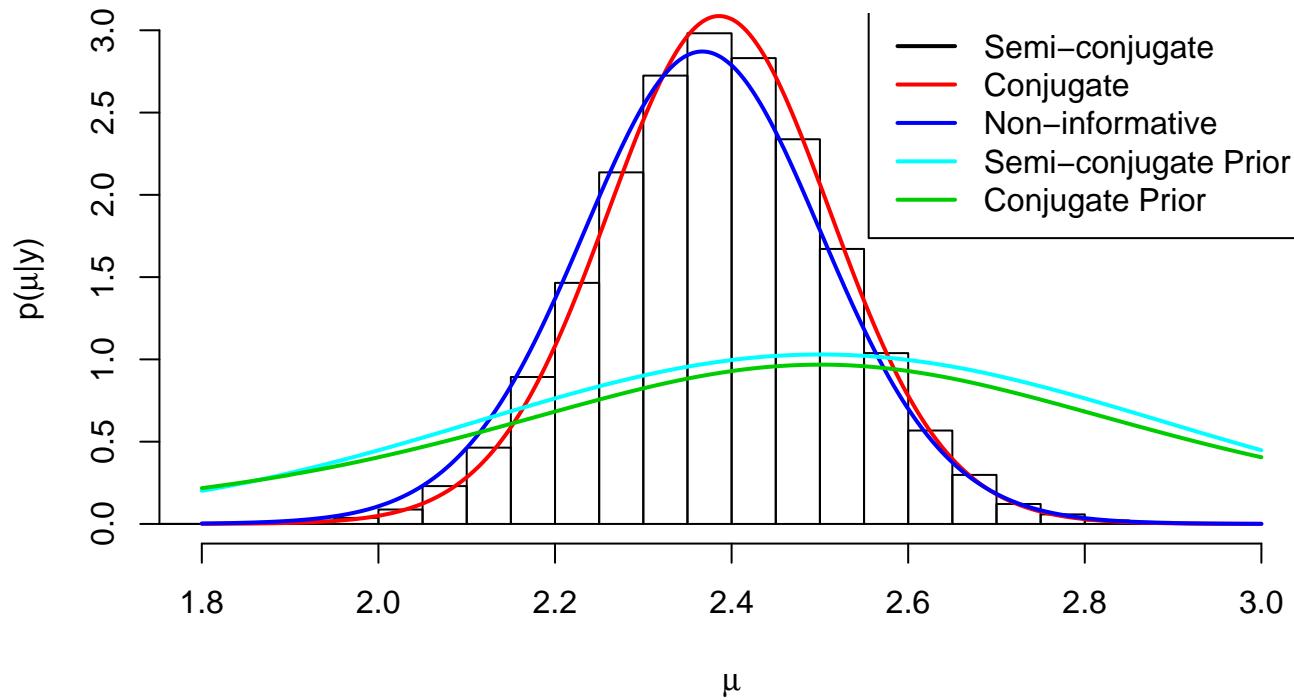
The following prior was chosen, trying to match the conjugate prior used last class

$$\mu|\sigma^2 \sim N\left(2.5, 0.15 = \frac{0.75}{5}\right) \quad \sigma^2 \sim \text{Inv}-\chi^2(4, 0.75)$$

In this case  $\mu_0$  was set to  $\mu_0$  and  $\tau_0^2$  was set to  $\frac{\sigma_0^2}{\kappa_0}$  as used in the conjugate prior. The prior on  $\sigma^2$  was the same in each case.

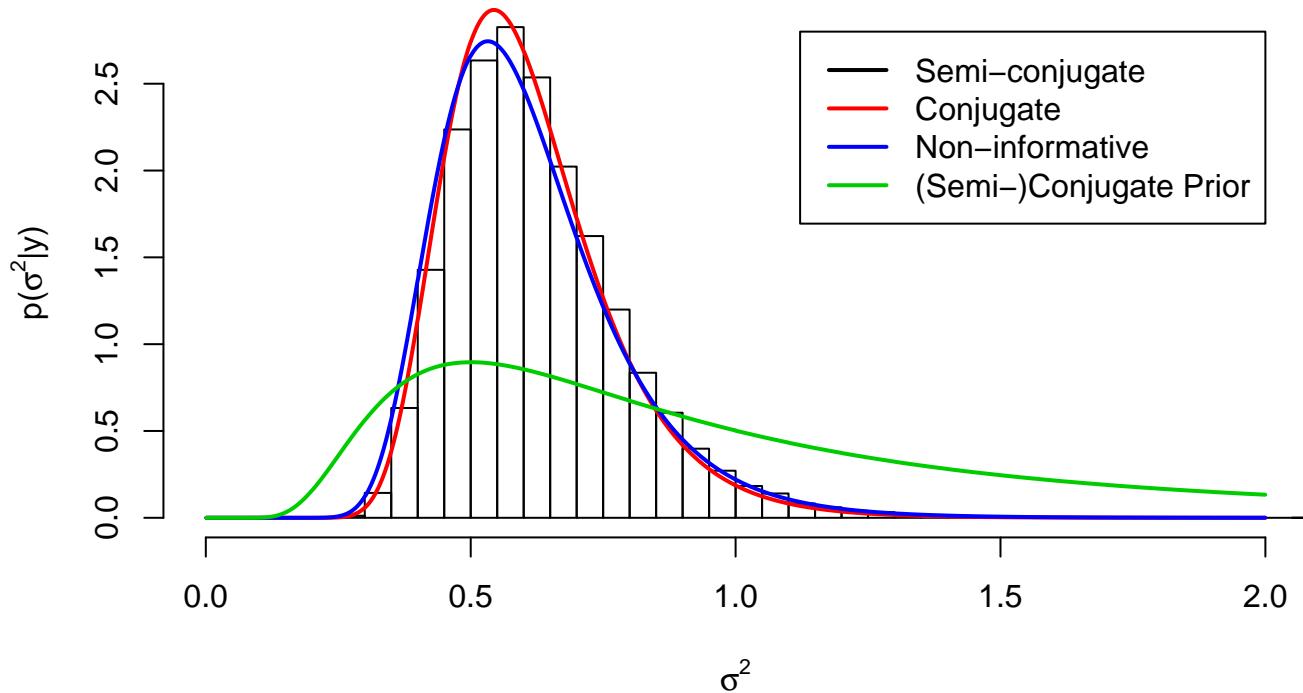
The results are based on 500,000 imputations.

$\mu|y:$



Prior	$E[\mu y]$	$SD(\mu y)$	95% Cred. Int.
Non-informative	2.367	0.143	(2.085, 2.649)
Conjugate	2.386	0.132	(2.125, 2.647)
Semi-conjugate	2.383	0.135	(2.119, 2.650)

$$\sigma^2|y$$



Prior	$E[\sigma^2 y]$	$SD(\sigma^2 y)$	95% Cred. Int.
Non-informative	0.6114	0.1729	(0.3610, 1.0287)
Conjugate	0.6120	0.1580	(0.3768, 0.9887)
Semi-conjugate	0.6270	0.1641	(0.3835, 1.0189)

In this case, the two informative priors give similar answers, though the semi-conjugate prior seems to give slightly larger answers for  $\sigma^2$ . This isn't particularly surprising as the form of the  $N-\text{Inv}-\chi^2$  distribution should lead to small values of  $\mu$  pulling down  $\sigma^2$ . In this case, the data suggests that  $\mu$  should be a bit lower than the prior specified.

Though one surprising result is that the posterior correlation between  $\mu$  and  $\sigma^2$  seems larger in the semi-conjugate case ( $r = 0.0253$ ) than in the conjugate case ( $r = 0.0014$ ).

However there is a suggestion that there might be a problem with the simulation calculating these (particularly in the conjugate case) so this might be taken with a grain of salt).

## Multivariate Normal Models

$y$  is a vector of length  $d$  with mean vector  $\mu$  (also of length  $d$ ) and  $d \times d$  variance matrix  $\Sigma$ , ( $y|\mu, \Sigma \sim N_d(\mu, \Sigma)$ ). The density of a single observation is

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu)\right)$$

where  $|\Sigma|$  is the determinant of the matrix  $\Sigma$ .

The likelihood of  $n$  iid observations is

$$\begin{aligned} p(y_1, \dots, y_n | \mu, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right) \\ &= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \text{tr}(\Sigma^{-1} S_0)\right) \end{aligned}$$

where  $\text{tr}(A)$  is the trace of the matrix  $A$  (the sum of the diagonal entries) and

$$S_0 = \sum_{i=1}^n (y_i - \mu)(y_i - \mu)^T$$

So the density and likelihood look like what we get in the univariate case, but with matrix and vectors instead.

Note that most of the inference in this model is a direct analogue to the univariate case. However we need a multivariate analogue to the  $\chi^2$  and  $\text{Inv-}\chi^2$  distributions.

# Wishart and Inverse Wishart Distributions

- Wishart distribution ( $\text{Wishart}_\nu(\Lambda)$ )

Multivariate analogue of a scaled  $\chi^2$  distribution

If  $z_1, \dots, z_\nu \stackrel{iid}{\sim} N_d(0, \Lambda)$  then

$$\Sigma = \sum_{i=1}^{\nu} z_i z_i^T \sim \text{Wishart}_\nu(\Lambda)$$

like  $z_1, \dots, z_\nu \stackrel{iid}{\sim} N(0, \tau^2)$  then

$$S = \sum_{i=1}^{\nu} z_i^2 \sim \tau^2 \chi_\nu^2$$

- Inverse Wishart distribution ( $\text{Inv-Wishart}_\nu(\Lambda^{-1})$ )

Multivariate analogue of a scaled  $\text{Inv-}\chi^2$  distribution

If  $\Sigma \sim \text{Wishart}_\nu(\Lambda)$  then

$$\Sigma^{-1} \sim \text{Inv-Wishart}_\nu(\Lambda^{-1})$$

# Common Multivariate Normal Models

- Unknown mean but known variance

$$\mu | \Sigma \sim N(\mu_0, \Lambda_0)$$

$$\mu | \Sigma, y \sim N(\mu_n, \Lambda_n)$$

where

$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

Like the univariate case, the posterior mean is a weighted average of the prior mean and the sample average and the posterior precision matrix is the prior ‘precision matrix + data precision matrix’.

- Unknown mean and variance - conjugate prior

$$\begin{aligned}\Sigma &\sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1}) \\ \mu|\Sigma &\sim N(\mu_0, \Sigma/\kappa_0)\end{aligned}$$

The posterior distribution satisfies

$$\begin{aligned}\Sigma|y &\sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1}) \\ \mu|\Sigma, y &\sim N(\mu_n, \Sigma/\kappa_n)\end{aligned}$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)(\bar{y} - \mu_0)^T \\
S &= \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T
\end{aligned}$$

In addition, it is possible to integrate out the variance matrix showing that

$$\begin{aligned}
\mu | y &\sim t_{\nu_n - d + 1}(\mu_n, \Lambda_n / (\kappa_n(\nu_n - d + 1))) \\
&\text{(i.e. multivariate } t \text{ with } \nu_n - d + 1 \text{ degrees of freedom)}
\end{aligned}$$

- Unknown mean and variance - non-informative prior

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

which is the Jeffreys' prior and is the limit of the conjugate prior as  $\kappa_0 \rightarrow 0$ ,  $\nu_0 \rightarrow -1$ , and  $|\Sigma_0| \rightarrow 0$ .

The posterior in this case satisfies

$$\begin{aligned}\Sigma | y &\sim \text{Inv-Wishart}_{n-1}(S) \\ \mu | \Sigma, y &\sim N(\bar{y}, \Sigma/n)\end{aligned}$$

Similarly to the univariate case,

$$\mu | y \sim t_{n-d}(\bar{y}, S/(n(n-d)))$$

# Poisson Regression

Example: Geriatric study

A researcher in geriatrics designed a 6 month prospective study on  $n = 100$  subjects to investigate the effects of two interventions on the frequency of falls. We will examine the effect of the intervention along with one of the covariates (Strength index) believed to be associated with the number of falls.

Data model: ( $y_i$  = number of falls during study,  $z_i$  = Intervention,  $x_i$  = Balance index)

$$y_i | \lambda_i \stackrel{ind}{\sim} Pois(\lambda_i)$$
$$\log \lambda_i = a + bx_i + cz_i$$

Prior:

Assume  $a$ ,  $b$ , and  $c$  are independent with

$$a \sim N(0, 100)$$

$$b \sim N(0, 100)$$

$$c \sim N(0, 100)$$

This is intended to be a fairly non-informative prior and clearly isn't a conjugate. The posterior distribution is of the form

$$p(a, b, c|y) \propto e^{-a^2/200} e^{-b^2/200} e^{-c^2/200} \prod_{i=1}^n e^{(a+bx_i+cz_i)y_i} e^{-e^{a+bx_i+cz_i}}$$

Given the form of this posterior, it will need to be examined by simulation. 5000 samples will be generated by the Gibbs sampler.

## Questions of interest:

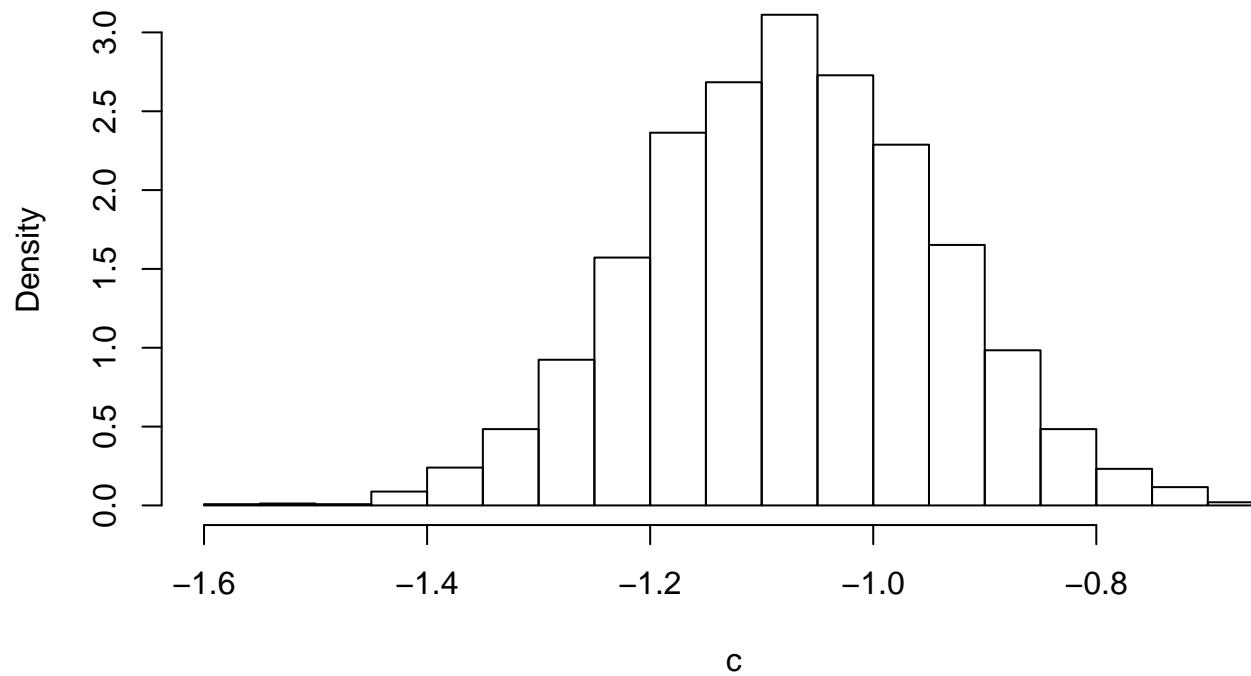
- $p(c|y)$
- $p(b|y)$
- $p(b, c|y)$
- $p(a, b|y)$
- $P[c < 0|y]$  ( $c < 0$  indicates intervention works)
- $p(e^c|y)$  ( $e^c$  gives the rate of change in the expected number of falls)

$$\lambda = e^{cz}e^{a+bx}$$

- $\lambda$  when  $x = 40$  under no intervention and intervention

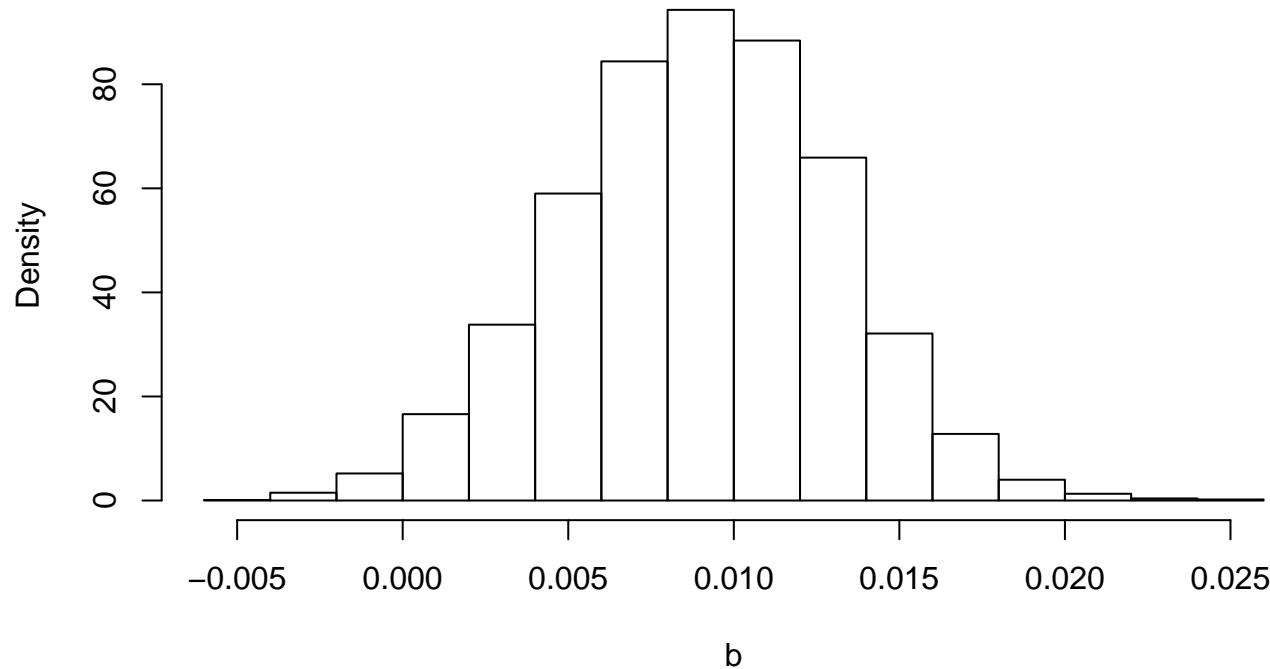
Answers:

- $p(c|y)$



$$E[c|y] = -1.074; \quad \text{SD}(c|y) = 0.131$$

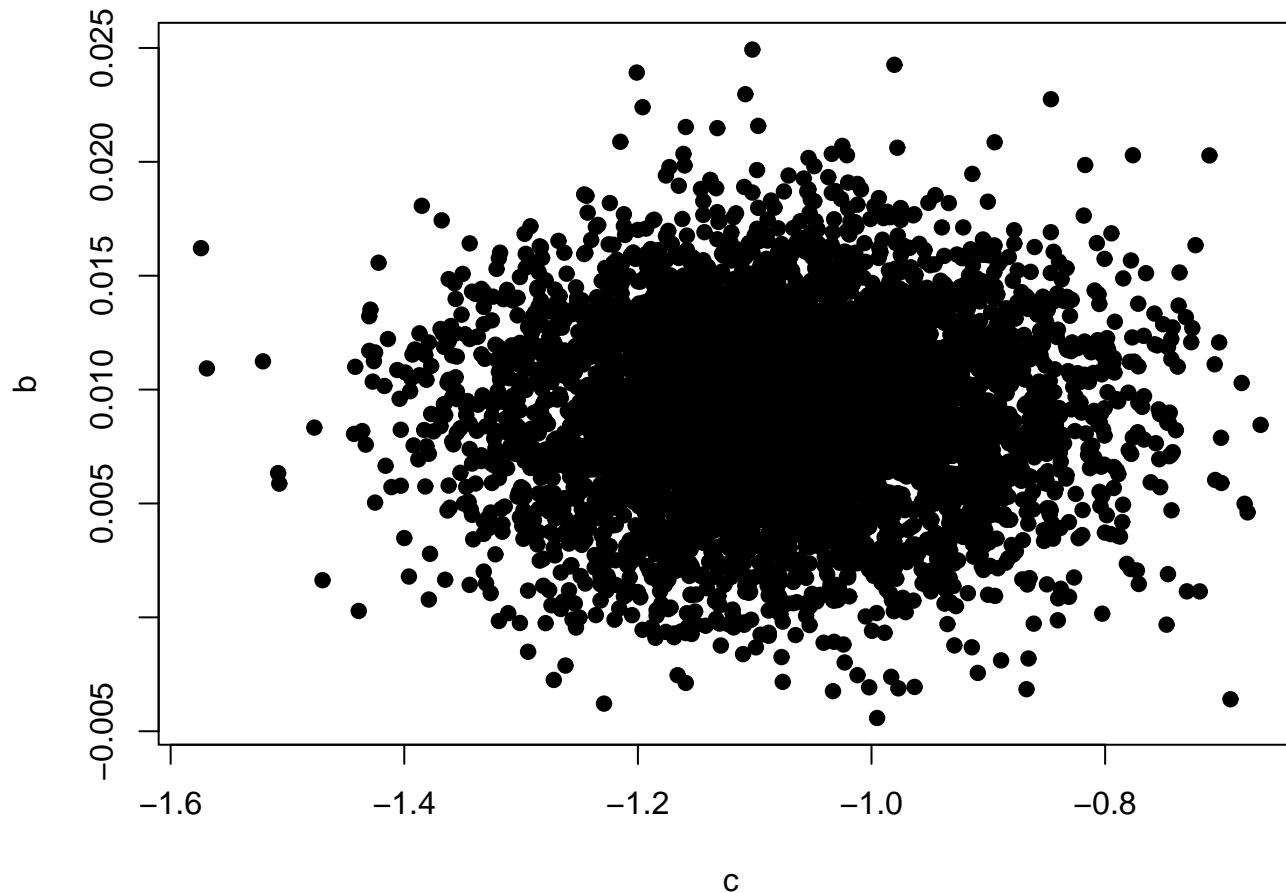
- $p(b|y)$



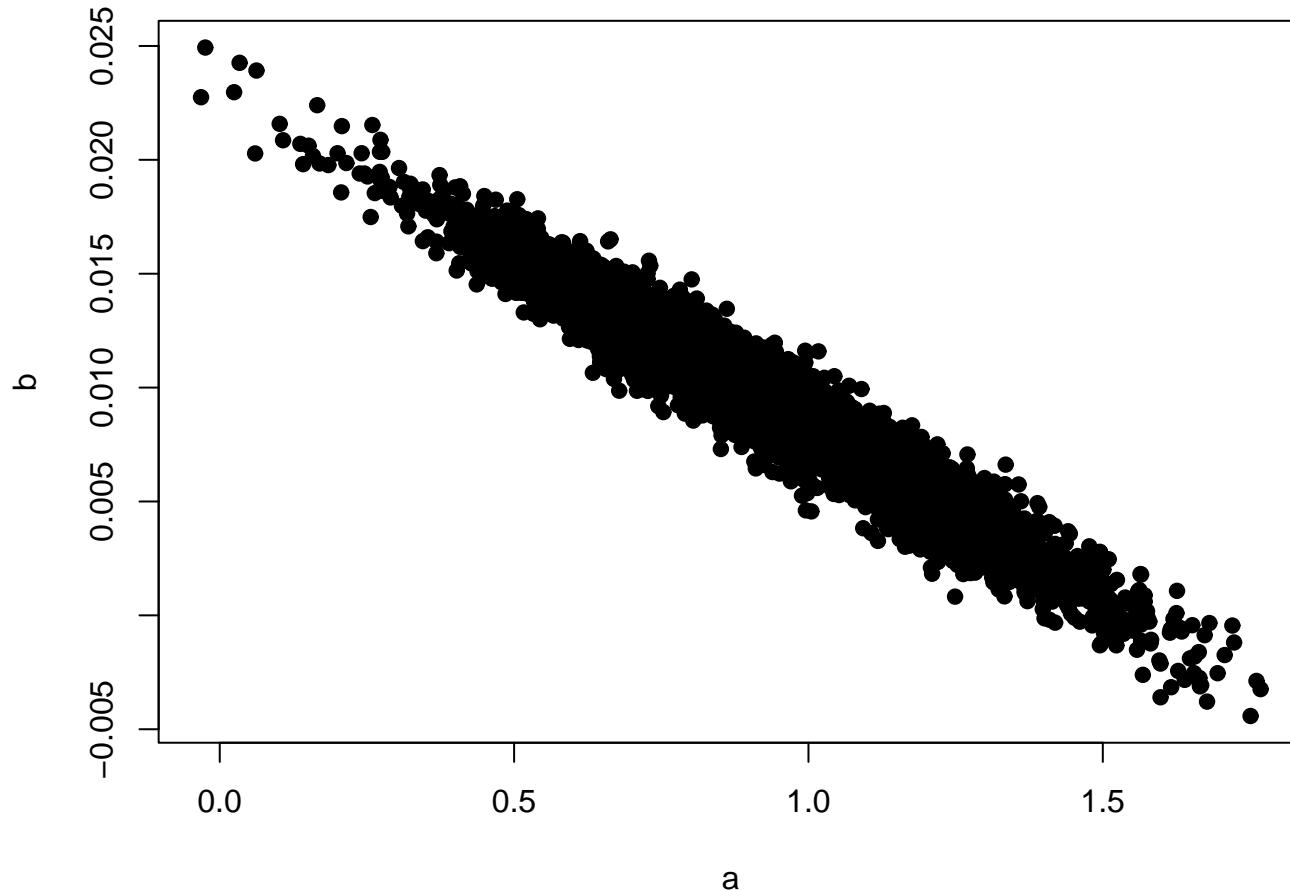
$$E[b|y] = 0.00898; \quad \text{SD}(b|y) = 0.00406$$

Note that this result is a bit surprising, since an increased strength index is expected to lead to fewer falls.

- $p(b, c|y)$



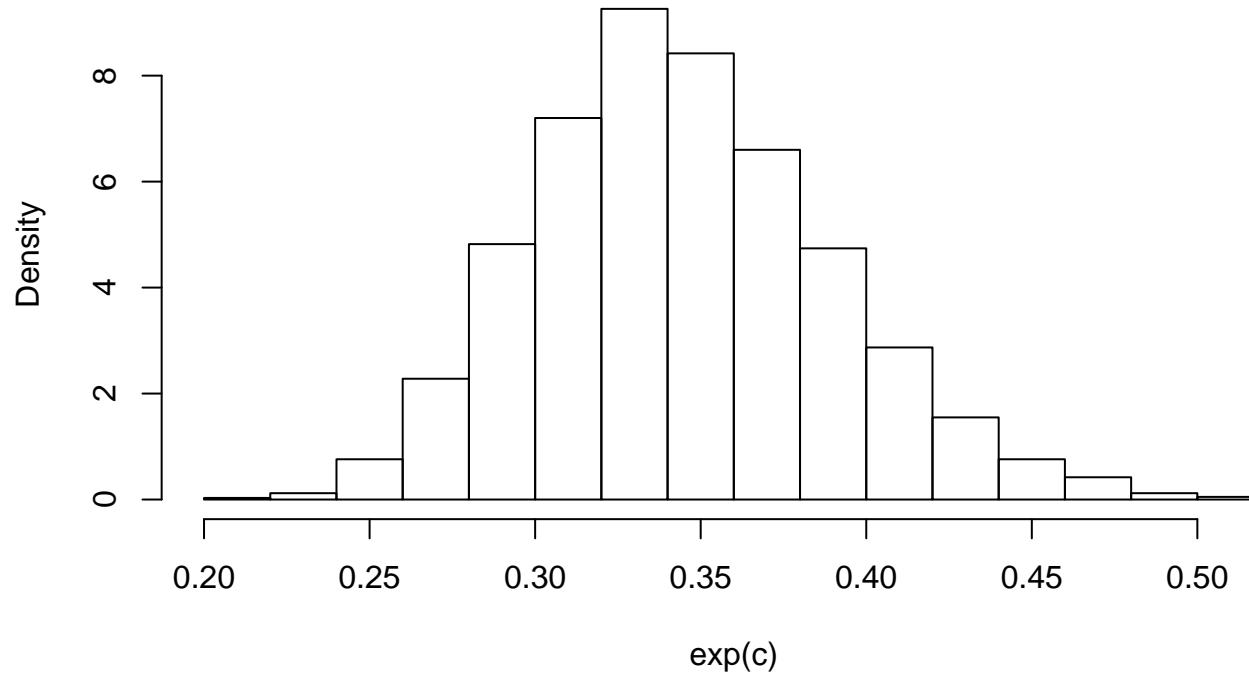
- $p(a, b|y)$



- $P[c < 0|y]$

$$P[c < 0|y] \approx \frac{1}{m} \sum_{i=1}^m I(c_i < 0) = 1$$

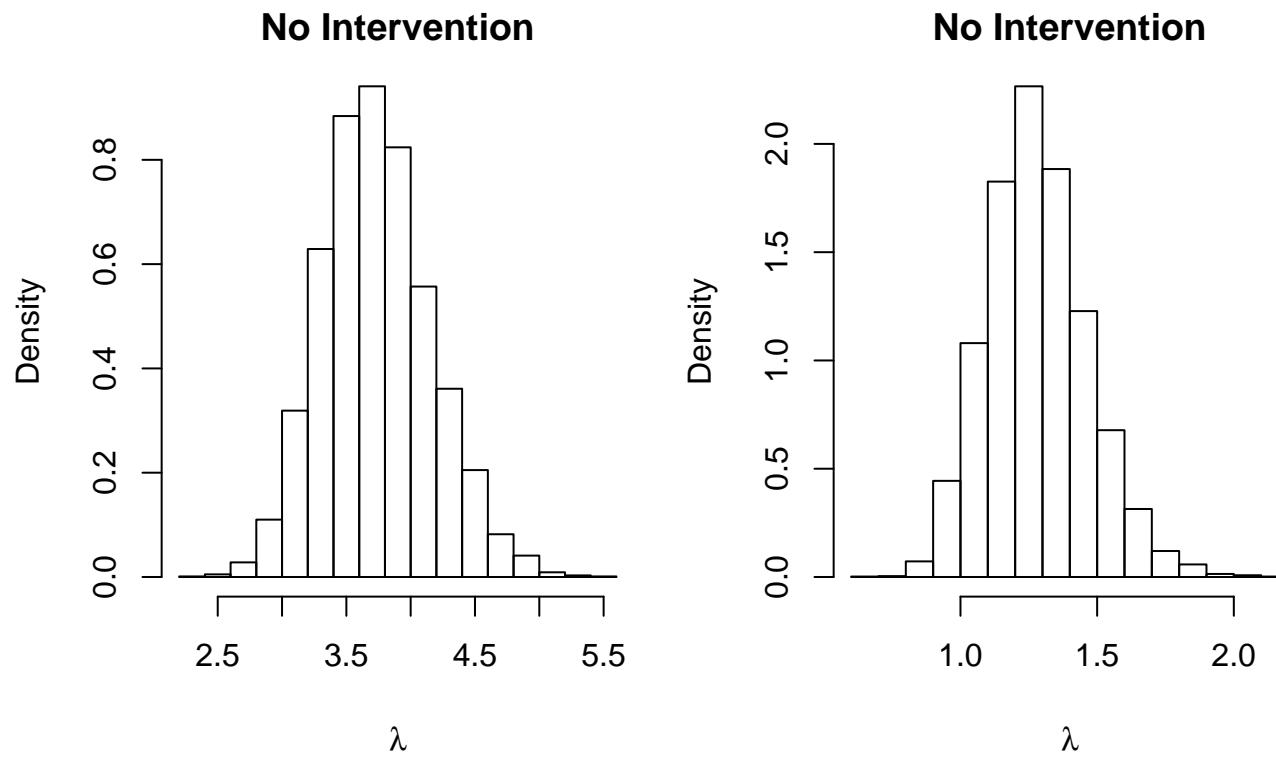
- $p(e^c|y)$



$$E[e^c|y] = 0.345; \quad \text{SD}(e^c|y) = 0.0451$$

This implies that the effect of the intervention should lead to people having less than half a many falls. The best guess is about a third as many.

- $\lambda$  when  $x = 40$  under no intervention and intervention



Intervention	$E[\lambda y]$	$SD(\lambda y)$
No	3.735	0.421
Yes	1.281	0.184

# Unit 10: Survival analysis and messy data

# Course Units

- ▶ Introduction to Bayesian Statistics
- ▶ Prior Distributions
- ▶ Simple Models
- ▶ Hierarchical Modeling
- ▶ Bayesian Computation
- ▶ Model Assessment and Comparison
- ▶ Regression Modeling
- ▶ Clinical Trials
- ▶ Bayesian Nonparametrics
- ▶ Survival Analysis and Messy Data

# Outline of the Unit

## 1. Survival Analysis

- ▶ Ibrahim et al. (2001) Bayesian Survival Analysis.

## 2. Messy Data:

- ▶ GCSR Chs. 7 and 21.

# Overview

Probability of surviving to time  $t$  is

$$P(T > t) = S(t) = 1 - F(t) = 1 - \int_0^t f(u)du$$

Hazard function is the instantaneous probability of failure at time  $t$

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Properties of the hazard function:

$$\begin{aligned} h(t) &\geq 0 \\ \int h(t)dt &= \infty \end{aligned}$$

Other relationships:

$$\begin{aligned} h(t) &= -\frac{d}{dt} \log(S(t)) \\ S(t) &= \exp\left(-\int_0^t h(u)du\right) \end{aligned}$$

# Cox (proportional hazards) model

The Cox model is

$$h(t|x) = h_0(t) \exp(f(x))$$

and commonly one assumes a multiplicative effect of covariates, namely a linear relationship in the exponential

$$h(t|x) = h_0(t) \exp(x^T \beta)$$

which implies

$$\log RR = \log \frac{h(t|x_i)}{h(t|x_j)} = x_i^T \beta - x_j^T \beta$$

Under right-censoring one has the following likelihood

$$\begin{aligned} P(y|\beta, h_0(t)) &\propto \prod_i f(y_i|\beta, h_0(y_i))^{\nu_i} S(y_i|\beta, h_0(y_i))^{1-\nu_i} \\ &= \prod_i (h_0(y_i) \exp(x_i^T \beta))^{\nu_i} \cdot \\ &\quad \exp \left( - \sum_i \exp(x_i^T \beta) \int_0^{y_i} h_0(u) du \right) \end{aligned}$$

where  $y_i = \min(t_i, c_i)$  where  $t_i$  is the failure time and  $c_i$  is the censoring time, with  $\nu_i = 1$  if  $t_i \leq c_i$  and  $\nu_i = 0$  if censored.

# Cox's partial likelihood

The PL avoids having to specify the baseline hazard,  $h_0(t)$ . One maximizes the following partial likelihood with respect to  $\beta$ .

$$PL(y|\beta) = \prod_{j=1}^d \frac{\exp(x_{(j)}^T \beta)}{\sum_{I \in R_j} \exp(x_I^T \beta)}$$

where  $d$  is the number of observed failure times and the risk set,  $R_j$ , is the set of individuals at risk at time  $y_{(j)}$  (i.e., all those not yet failed, not yet censored, plus the person who fails at that time).

Question: What is the dilemma for the Bayesian?

# Parametric models

- ▶ Exponential model (constant hazard over time):

$$h_0(t) = 1$$

$$h(t_i) = \exp(x_i^T \beta)$$

$$P(y, v | \beta) = \prod_i (\exp(x_i^T \beta))^{\nu_i} \exp \left( - \sum_i \exp(x_i^T \beta) y_i \right)$$

Inference is possible using MCMC: BUGS uses ARS.

- ▶ Weibull model

$$h(t_i) = \alpha t_i^{\alpha-1} \exp(x_i^T \beta); \alpha > 0$$

$$P(y, v | \beta, \alpha) = \prod_i (\alpha y_i^{\alpha-1} \exp(x_i^T \beta))^{\nu_i}.$$

$$\exp \left( - \sum_i \exp(x_i^T \beta) y_i^\alpha \right)$$

- ▶ What does the baseline hazard look like here?
- ▶ Again one can do MCMC in a straightforward way (BUGS uses ARS).
- ▶ Standard prior for  $\alpha$  is a gamma.
- ▶ Log transformation may be useful: this allows one to distinguish  $\alpha < 0$  from  $\alpha = 0$  (constant hazard) from  $\alpha > 0$ .

# Semiparametric models (1)

- ▶ Basic idea is to use the parametric proportional hazards assumption with a nonparametric specification of the baseline hazard,  $h_0(t)$ .
- ▶ Piecewise constant baseline hazard:  
Let  $0 < s_1 < s_2 < \dots < s_J$  with  $s_J > \max y_i$  and take

$$h_0(t) = \lambda_j \text{ for } t \in (s_{j-1}, s_j]$$

See Ibrahim et al. for the likelihood.

- ▶ A common prior is  $\lambda_j \stackrel{\text{iid}}{\sim} \mathcal{G}(\alpha_j, \gamma_j)$ .
- ▶ A more satisfying generalization is to have the piecewise baseline hazards be correlated, e.g., with a AR type structure.

# Semiparametric models (2)

- ▶ The most commonly used approach is the so-called gamma process.
- ▶ Gamma processes are distributions over functions whose sample functions (realizations, paths) are increasing functions.
  - ▶ So we would use the gamma process on the cumulative baseline hazard,  $H_0(t) = \int_0^t h_0(u)du$
- ▶ We say  $Z(t) \sim \mathcal{GP}(c\alpha(t), c)$  if

$$Z(0) = 0$$

$Z(t)$  has indep. increments in disjoint intervals

$$Z(t) - Z(s) \sim \mathcal{G}(c(\alpha(t) - \alpha(s)), c) \text{ for } t > s$$

- ▶  $\alpha(t)$  is like the base measure in the Dirichlet process: the process is centered around the function but with additional variability whose magnitude depends on the hyperparameter ( $c$  here).
- ▶ To implement for the survival case, we take  $H_0(t) \sim \mathcal{GP}(cH_0^*(t), c)$ 
  - ▶ One would generally specify  $H_0^*(t)$  to be smooth.
  - ▶ We might take a Weibull form,  $H_0^*(t) = \eta t^\kappa$ .

# Gamma Process

- ▶ For computation, see Ibrahim et al.
  - ▶ For a discretization of time, we have that
$$\lambda_j \sim \mathcal{G}(c(H_0^*(s_j) - H_0^*(s_{j-1})), c).$$
  - ▶ Compared to the piecewise model, this ties together the hazards through  $H_0^*(t)$ .
- ▶ BUGS 'leuk' example appears to be of this form with  $H_0^*(t)$  based on the exponential model (i.e.,  $h_0(t) = 1$ ).
- ▶ Also note that Cox's partial likelihood is a limiting case of the marginal posterior for  $\beta$  under a gamma process prior.
- ▶ A good way to proceed might be to start with a simple parametric Weibull model and then consider a Gamma process model with a baseline Weibull as a sensitivity analysis.

# Frailty models

- ▶ Frailty models are the extension of survival analysis to include random effects.
- ▶ Basic model is

$$h(y_{ij} | w_i, x_{ij}) = h_0(y_{ij}) w_i \exp(x_{ij}^T \beta)$$

where the  $w_i$  is the effect for the  $i$ th cluster.

- ▶ The most common distribution for  $w$  is  $w_i \sim \mathcal{G}(\kappa^{-1}, \kappa^{-1})$  where  $\kappa$  is the variance of the random effects.
  - ▶ Under the Weibull baseline hazard, one can marginalize over the  $w_i$ .
  - ▶ Under the nonparametric baseline hazard models, one generally needs to sample the  $w_i$ .

# Measurement error

- ▶ Classical measurement error problem is

$$\begin{aligned} y_i &\sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \\ w_i &\sim \mathcal{N}(x_i, \sigma_x^2) \end{aligned}$$

where we observe  $w$  instead of  $x$ .

- ▶ To the Bayesian, this is simple in principle - just treat  $x_i$  values as parameters.
  - ▶ We need information about  $\sigma_x^2$  to proceed. This could be from fixing  $\sigma_x^2$  as a sensitivity analysis, a subjective prior, or information from a validation study.
  - ▶  $x_i$  are nuisance parameters.
  - ▶ Hopefully we can integrate over them to simplify computation of the posterior for the parameters of interest, but start the problem by specifying the full model.
- ▶ One might think of this as a type of missing data problem.

# Missing data: missing observations

- ▶ A general principle is to think of whether data are missing or not as another piece of data.
  - ▶ Complete data likelihood

$$P(y, I | \theta, \phi) = P(y | \theta)P(I | y, \phi)$$

where  $I$  is a vector indicating which observations are missing.

- ▶ However, we only observe  $y_{obs}$  where  $y = (y_{obs}, y_{mis})$  so the likelihood we work with is the observed data likelihood,

$$P(y_{obs}, I | \theta, \phi) = \int P(y, I | \theta, \phi) dy_{mis}$$

- ▶ In the Bayesian approach, one may want to treat  $y_{mis}$  as parameters in the model (i.e., data augmentation) or one may integrate it out of the model as done to get the observed data likelihood.
- ▶ If one samples  $y_{mis}$  in an MCMC, this is Bayesian multiple imputation.

# Ignorable missingness

- ▶ Whether missingness is ignorable comes down to whether the information about whether an observation is missing is involved in the posterior for the parameters of interest.
- ▶ If  $P(\theta|x, y_{obs}, I) = P(\theta|x, y_{obs})$  then the missingness is ignorable and we don't need to model the missingness process (i.e., just use  $P(y_{obs}|\theta, x)$ )
- ▶ There are two conditions for ignorability

$$P(I|x, y, \phi) = P(I|x, y_{obs}, \phi)$$

and

$$P(\phi|x, \theta) = P(\phi|x)$$

- ▶ The latter condition is that any parameters related to the missingness process must not be related to the parameters in the rest of the model.
- ▶ The best thing to do if confused is to write out a full model for  $y_{obs}, y_{mis}, I$ , simplify based on your assumptions about conditional independences and see what things simplify to.

# Missing data: missing covariates

- ▶ We're interested in

$$P(\theta|y, X_{obs}) = \int P(\theta, X_{mis}|X_{obs}, y) dX_{mis}$$

- ▶ Bayesian treats  $X_{mis}$  as parameters.

- ▶ One needs a prior for  $X_{mis}$ :

$$\begin{aligned} P(\theta, X_{mis}|X_{obs}, y) &\propto P(y|X_{mis}, X_{obs}, \theta)P(\theta, X_{mis}|X_{obs}) \\ &= P(y|X_{mis}, X_{obs}, \theta)P(\theta)P(X_{mis}|X_{obs}) \end{aligned}$$

But the prior is likely to take the form of a model

$$P(X_{mis}|X_{obs}) = \int P(X_{mis}|X_{obs}, \phi)P(\phi)d\phi$$

- ▶ So the full posterior is likely of the form

$$P(\theta, \phi, X_{mis}|X_{obs}, y).$$

- ▶ One marginalizes over  $X_{mis}$  and  $\phi$ , e.g., using the MCMC output.

# Censoring

- ▶ Suppose observations are censored. E.g., concentrations may only be measurable above some threshold and values below this are reported as below the limit of detection (LOD).
- ▶ A Bayesian approach is to treat the missing concentrations as unobserved latent variables and include them as random variables (i.e., parameters) in the model.

$$\begin{aligned}c_i = 0, \quad & y_i > L \\c_i = 1, \quad & y_i \leq L\end{aligned}$$

- ▶ So one would include  $y_i$  in the MCMC for all the observations for which  $y_i \leq L$ .  
The sampling would be based on the following likelihood accounting for censoring

$$P(y_i|\theta) \cdot I(y_i \leq L)$$

So any proposals should only propose allowable values of  $y_i$ .

- ▶ For computation, one may be able to integrate over the latent  $y_i$  values.
- ▶ Similar reasoning can be used for interval censoring.

# Order constraints

Suppose you have a model in which you believe there is a particular ordering of parameters:  $\theta_1 < \theta_2 < \theta_3$ . How enforce this in the Bayesian paradigm?

1. Reparameterize such that the constraint is forced to be true.
2. Add  $I(\theta_1 < \theta_2 < \theta_3)$  to your prior and only propose values of  $\theta$  that satisfy the constraint (other orderings have prior density of 0).
3. If you are working with samples from the posterior, ignore the constraint until after getting the samples, and then throw out all iterations that do not satisfy the constraint, i.e., explicitly estimating  $P(\theta|y, \theta_1 < \theta_2 < \theta_3)$ .

# What to do if you're confused (1)

If I'm confused about what the model should be, I tend to do the following in terms of writing out the model.

- ▶ Go back to first principles.
- ▶ Write down the posterior:  $P(\text{unknown}|\text{data}, \text{known})$  and work out in terms of conditionals and marginals (i.e., likelihood and prior terms).
- ▶ If parts of the data or the 'known' (i.e., covariates) are actually not known, step back to  $P(\text{unknown}, \text{data}, \text{known})$  and start writing out as marginal and conditionals, breaking up any pieces into observed and unobserved.
- ▶ Decide what is conditionally independent of what else and how that simplifies the model.

# What to do if you're confused (2)

## ▶ Marginalization:

For computations, write out the full model,  $P(\theta_1, \dots, \theta_k | y)$ . Then see what you can marginalize out of the model, e.g.,

$$P(\theta_1 | y) = \int P(\theta_1, \theta_2 | y) d\theta_2$$

- ▶ Set up a sampling approach for  $\theta_1$ . Then if you need to do inference for  $\theta_2$  (i.e., it's not a nuisance parameter), sample from

$$P(\theta_2 | \theta_1, y)$$

- ▶ Note that if you were able to do the integral in closed form to get the marginal, this is likely to be a known closed-form conditional from which you can directly sample.

## ▶ Predictive sampling:

- ▶ If  $z$  is new (unobserved) data or  $z$  is related to  $\theta$  but not involved in the likelihood for  $y$  then:
- ▶ We can factorize  $P(\theta, z | y) = P(\theta | y)P(z | \theta)$  so we can sample  $z$  separately from the main sampling. Also, if  $z$  is not of scientific interest, it does not need to be included in the model.

# Normal Random Effects Model

Statistics 220

Spring 2005



## Normal-Normal Hierarchical Model

Have  $J$  independent groups, with known variance  $\sigma^2$

$$y_{ij} | \theta_j \stackrel{ind}{\sim} N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j; \quad j = 1, \dots, J$$

Except for the fixed measurement variance, this is the basis for the 1-way ANOVA model. So following the analysis for this model, the sample mean for each group be

$$\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

Its sampling variance is

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

So

$$\bar{y}_{.j} | \theta_j \stackrel{\text{ind}}{\sim} N(\theta_j, \sigma_j^2)$$

For what follows, we are going to base it on the above normal model, independent observations with (potentially) different but known variances.

Note: In most situations, the assumption of known measurement error variances is dubious. However it is not always. The book discusses two examples where assuming that these variances are effective known is reasonable. Both involve situations where the data to be analyzed comes from summary measures from analyzes of large data sets.

If this assumption is not reasonable, we can put a prior distribution on  $\sigma^2$ . In this case, the analysis isn't quite as nice as what follows, but is tractable. We'll come back to it in Chapter 15.

We now need a model for  $\theta_1, \dots, \theta_J$ . A popular choice is

$$\theta_j | \mu, \tau^2 \stackrel{iid}{\sim} N(\mu, \tau^2)$$

When combined with the original data model, this gives us the standard normal random effects model used in ANOVA.

Next we need to put a prior on  $\mu$  and  $\tau^2$ . While we could put an informative prior on these, say by following semi-conjugate ideas discussed earlier, lets follow the text and use a non-informative prior. For many problems fitting into this framework, the data swamps the prior in the analysis.

One reasonable choice is to have  $\mu$  and  $\tau^2$  independent ( $p(\mu, \tau^2) = p(\mu)p(\tau^2)$ ). With this, the obvious prior on  $\mu$  is

$$p(\mu) \propto 1$$

i.e. uniform.

For  $\tau^2$ , one valid choice is

$$p(\tau) \propto 1$$

i.e. again uniform.

Note that the Jeffreys' prior for  $\tau$  ( $p(\log \tau) \propto 1, p(\tau) \propto \frac{1}{\tau}$ ) won't work as it leads to an improper posterior distribution.

- Joint posterior distribution

$$\begin{aligned} p(\theta, \mu, \tau | y) &\propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\ &\propto p(\mu, \tau) \prod_{j=1}^J N(\theta_j | \mu, \tau^2) \prod_{j=1}^J N(\theta_j | \mu, \tau^2) \end{aligned}$$

- Conditional posterior distribution of the normal means  $\theta_j$

Given the structure of the problem (independence of  $\theta_j$ 's given  $\mu$  and  $\tau$  and the independence of the  $\bar{y}_{.j}$ 's given the  $\theta_j$ 's), the conditional posterior  $p(\theta|\mu, \tau, y)$  factors into  $J$  independent pieces.

Notice that for each  $\theta$ , this is similar to the case of a single normal mean with the conjugate prior.

$$\begin{aligned} p(\theta_j|\mu, \tau, y) &\propto p(\theta_j|\mu, \tau^2)p(\bar{y}_{.j}|\theta_j, \sigma^2) \\ &\propto N(\theta_j|\mu, \tau^2)N(\theta_j|\mu, \tau^2) \\ &= N(\theta_j|\hat{\theta}_j, V_j) \end{aligned}$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

- Marginal posterior distribution of the posterior distribution of the hyperparameters  $\mu$  and  $\tau$

$$p(\mu, \tau | y) \propto p(\mu, \tau) p(y | \mu, \tau)$$

As the book mentions, this decomposition isn't usually helpful as  $p(y|\mu, \tau)$  usually doesn't have a nice form. However for normal-normal model this can be determined as the integral

$$\begin{aligned} p(y|\mu, \tau) &= \int p(y, \theta|\mu, \tau) d\theta \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \theta)^2\right) \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right) d\theta \end{aligned}$$

can be calculated and seen to be nice. Given the quadratic structure of the exponential piece, it must be a normal distribution. The integration can be done by completing the square for  $\theta$  (giving a normal density to integrate out) or by getting the mean and variance of  $y|\mu, \tau$  by

$$\begin{aligned}
E[y] &= E[E[y|\theta]] = E[\theta] = \mu \\
\text{Var}(y) &= \text{Var}(E[y|\theta]) + E[\text{Var}(y|\theta)] \\
&= \text{Var}(\theta) + E[\sigma^2] = \tau^2 + \sigma^2
\end{aligned}$$

So

$$p(\mu, \tau | y) \propto p(\mu, \tau) \prod_{j=1}^J N(\theta_j | \mu, \sigma_j^2 + \tau^2)$$

Note: In the general situation, let  $\phi$  be the hyperparameter. While the use of conjugate priors will often give nice forms for  $p(y|\phi)$ , they don't combine well with the prior. For example, in the rat tumor example

$$p(y|\alpha, \beta) = \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + y)\Gamma(\beta + n - y)}{\Gamma(\alpha + \beta + n)}$$

(i.e. the Beta-Binomial distribution). The posterior density can be calculated (as we did last class), but there isn't a nice conjugate density to this distribution which allows for easy calculation in the future steps.

This sort of situation is commonly the case. The reason why things work nicely for the normal-normal model is that is the conjugate to itself.

Now lets use the fact  $p(\mu, \tau) \propto 1$

Similarly to before

$$\mu | \tau, y \sim N(\hat{\mu}, V_\mu)$$

where

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2} \bar{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}$$

The marginal posterior of  $\tau|y$  isn't quite as nice, though a useful form for the density can be found, based on the idea

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} \\ &\propto \frac{p(\tau) \prod_{j=1}^J N(\bar{y}_{.j}|\mu, \sigma^2 + \tau^2)}{N(\mu|\hat{\mu}, V_\mu)} \end{aligned}$$

As noted before, this must hold for any choice of  $\mu$ , so pick one to make this easy to work with. In this case evaluate at  $\mu = \hat{\mu}$  giving,

$$\begin{aligned}
p(\tau|y) &\propto \frac{p(\tau) \prod_{j=1}^J N(\bar{y}_{.j}|\hat{\mu}, \sigma^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\
&\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^J \frac{1}{\sqrt{\sigma^2 + \tau^2}} \exp\left(-\frac{(\bar{y}_{.j} - \hat{\mu})^2}{2(\sigma^2 + \tau^2)}\right) \\
&= V_\mu^{1/2} \prod_{j=1}^J \frac{1}{\sqrt{\sigma^2 + \tau^2}} \exp\left(-\frac{(\bar{y}_{.j} - \hat{\mu})^2}{2(\sigma^2 + \tau^2)}\right)
\end{aligned}$$

Comment on prior  $p(\tau)$ : As mentioned earlier, the Jeffreys' prior ( $p(\tau) \propto \tau$ ) leads to an improper posterior. To show this, you can integrate the density and show that it is infinite. Effectively what is happening is that there are few degrees of freedom for estimating  $\tau$ . The Jeffreys' prior puts too much weight on larger  $\tau$ s, which leads to the integral to blowup.

Computation:

As  $p(\tau|y)$  doesn't correspond to a standard distribution, analyzing the joint posterior is usually done by the following simulation scheme

1. Sample  $\tau_k$  from  $p(\tau|y)$
2. Sample  $\mu_k$  from  $p(\mu_k|\tau_k, y) = N(\mu_k|\hat{\mu}_k, V_{\mu_k})$  where

$$\hat{\mu}_k = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau_k^2} \bar{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau_k^2}} \quad \text{and} \quad V_{\mu_k}^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau_k^2}$$

3. Sample  $\theta_k$  from  $p(\theta_k | \mu_k, \tau_k, y)$ . In this case, the individual components are conditionally independent given  $\mu_k, \tau_k$ , and  $y$  giving

$$\theta_{j,k} \sim N(\hat{\theta}_{j,k}, V_{j,k})$$

where

$$\hat{\theta}_{j,k} = \frac{\frac{1}{\sigma_j^2} \bar{y}_{.j} + \frac{1}{\tau_k^2} \mu_k}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}} \quad V_{j,k} = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}}$$

Note the conditional independence of the  $\theta_j$ s holds in many hierarchical model. For example, it also held the rat tumor example. It also holds for many of the homework problems (e.g. Chapter 5, # 11,12). This situation will be found to be useful when we get to Gibbs sampling for doing the calculations.

Posterior predictive distributions:

There are two situations where the posterior predictive distribution may need to be calculated. These can be fit into the simulations already done

1.  $\tilde{y}$  from a group  $j$  already observed.

Sample  $\tilde{y}_{j,k}$  from  $N(\theta_{j,k}, \sigma^2)$

If  $m$  observations are needed, draw  $m$  values of  $\tilde{y}$  from the above distribution.

2.  $\tilde{y}$  from a new group  $\tilde{j}$

Sample  $\theta_{\tilde{j},k}$  from  $N(\theta|\mu_k, \tau_k)$  (draw from prior for  $\theta$ , not the posterior)

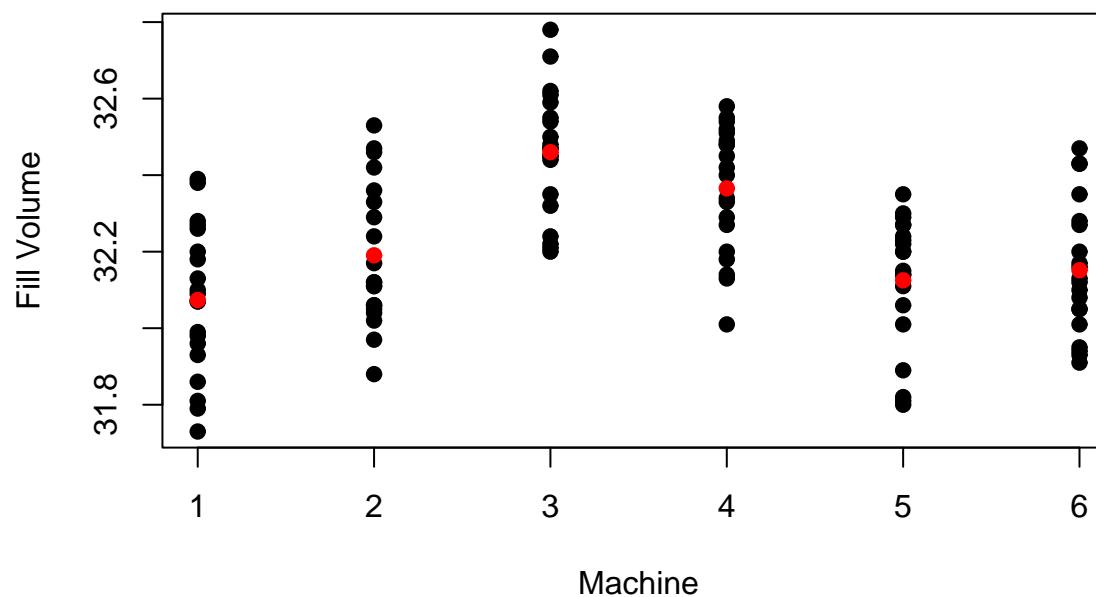
Sample  $\tilde{y}_{\tilde{j},k}$  from  $N(\theta_{\tilde{j},k}, \sigma^2)$ . Similarly to above if  $m$  samples are needed.

The key difference is do we need to draw a new  $\theta$  or use one we already have. The second situation will lead to more variable samples as there is less information about the corresponding  $\theta$  in this case.

# Examples

## Example 1: Detergent Filling Machines

Six filling machines of the same make and model were examined to see whether they put the same amount of detergent into a box. 20 observations from each machine were taken. The nominal amount that should be in a box is 32 ounces.



Note that for this example,  $\sigma_j^2$  is unknown, but can be estimated based the MSE from the 1-way ANOVA. We will proceed with this value ( $\sigma_j^2 = \sigma^2 = 0.00244$ ) is assumed known.

Calculating  $p(\tau|y)$ :

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2} \bar{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}} = \bar{y}_{..} = 32.228$$

$$V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}$$

$$V_\mu = \frac{\sigma^2 + \tau^2}{6} = \frac{0.00244 + \tau^2}{6}$$

By plugging these values into

$$\begin{aligned} p(\tau|y) &\propto \frac{p(\tau) \prod_{j=1}^J N(\bar{y}_{.j}|\hat{\mu}, \sigma^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^J \frac{1}{\sqrt{\sigma^2 + \tau^2}} \exp\left(-\frac{(\bar{y}_{.j} - \hat{\mu})^2}{2(\sigma^2 + \tau^2)}\right) \\ &= V_\mu^{1/2} \prod_{j=1}^J \frac{1}{\sqrt{\sigma^2 + \tau^2}} \exp\left(-\frac{(\bar{y}_{.j} - \hat{\mu})^2}{2(\sigma^2 + \tau^2)}\right) \end{aligned}$$

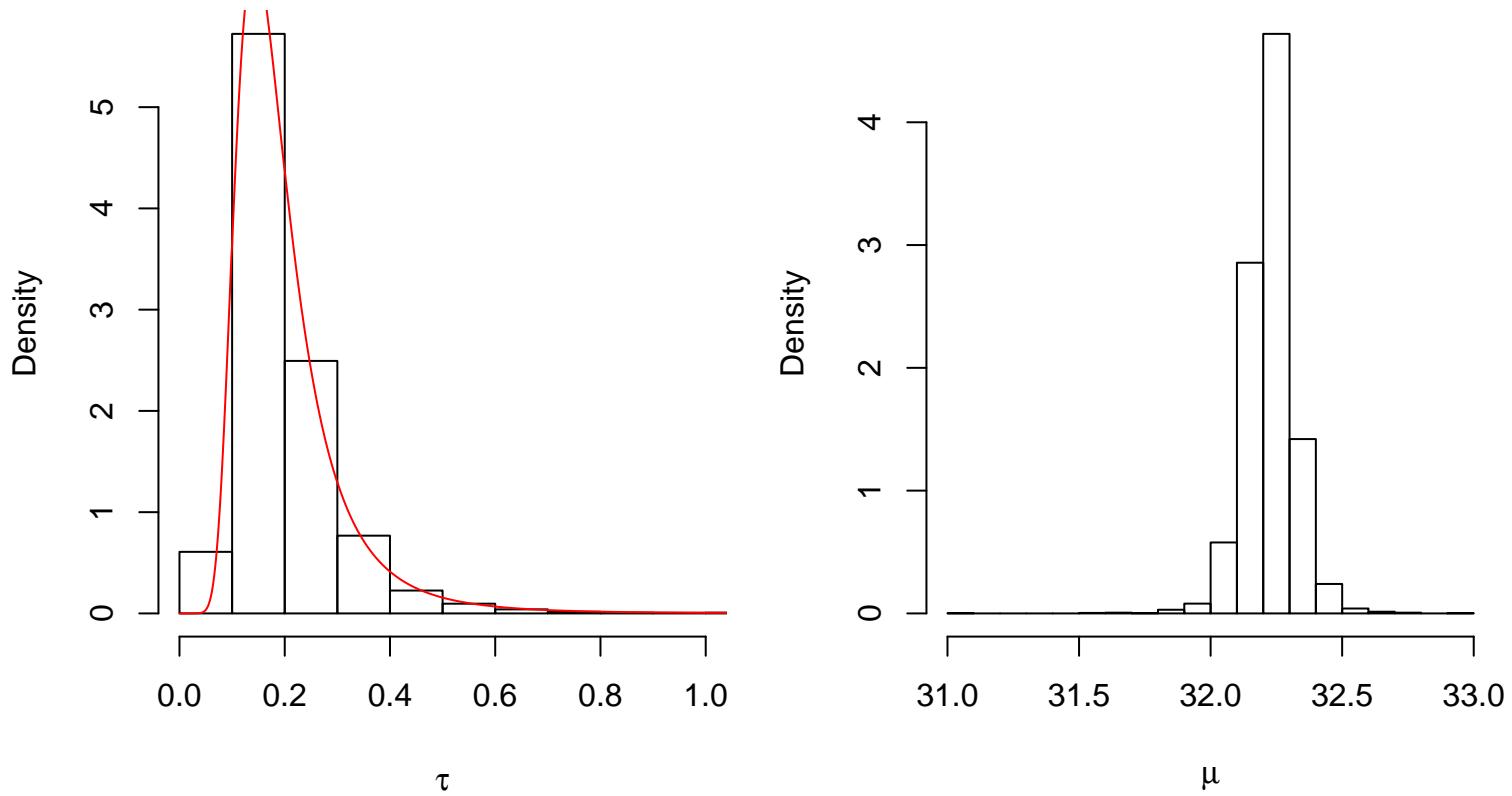
gives  $p(\tau|y)$ .

Lets simulate  $\tau_1, \dots, \tau_{5000}$  based on this unnormalized density.

Then  $p(\mu_k | \tau_k, y) = N(\mu_k | \hat{\mu}_k, V_{\mu_k})$  is calculated by

$$\begin{aligned}\hat{\mu}_k &= \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau_k^2} \bar{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau_k^2}} = \bar{y}_{..} = 32.228 \\ V_{\mu} &= \frac{\sigma^2 + \tau_k^2}{6} = \frac{0.00244 + \tau_k^2}{6}\end{aligned}$$

Now sample  $\mu_1, \dots, \mu_{5000}$  based on this conditional distributions.



$$E[\tau|y] = 0.2020 \quad E[\tau^2|y] = 0.0518$$

$$\text{Mode}(\tau|y) = 0.143 \quad \text{Mode}(\tau^2|y) = 0.0204$$

$$E[\mu|y] = 32.228 \quad P[\mu > 32|y] = 0.9878$$

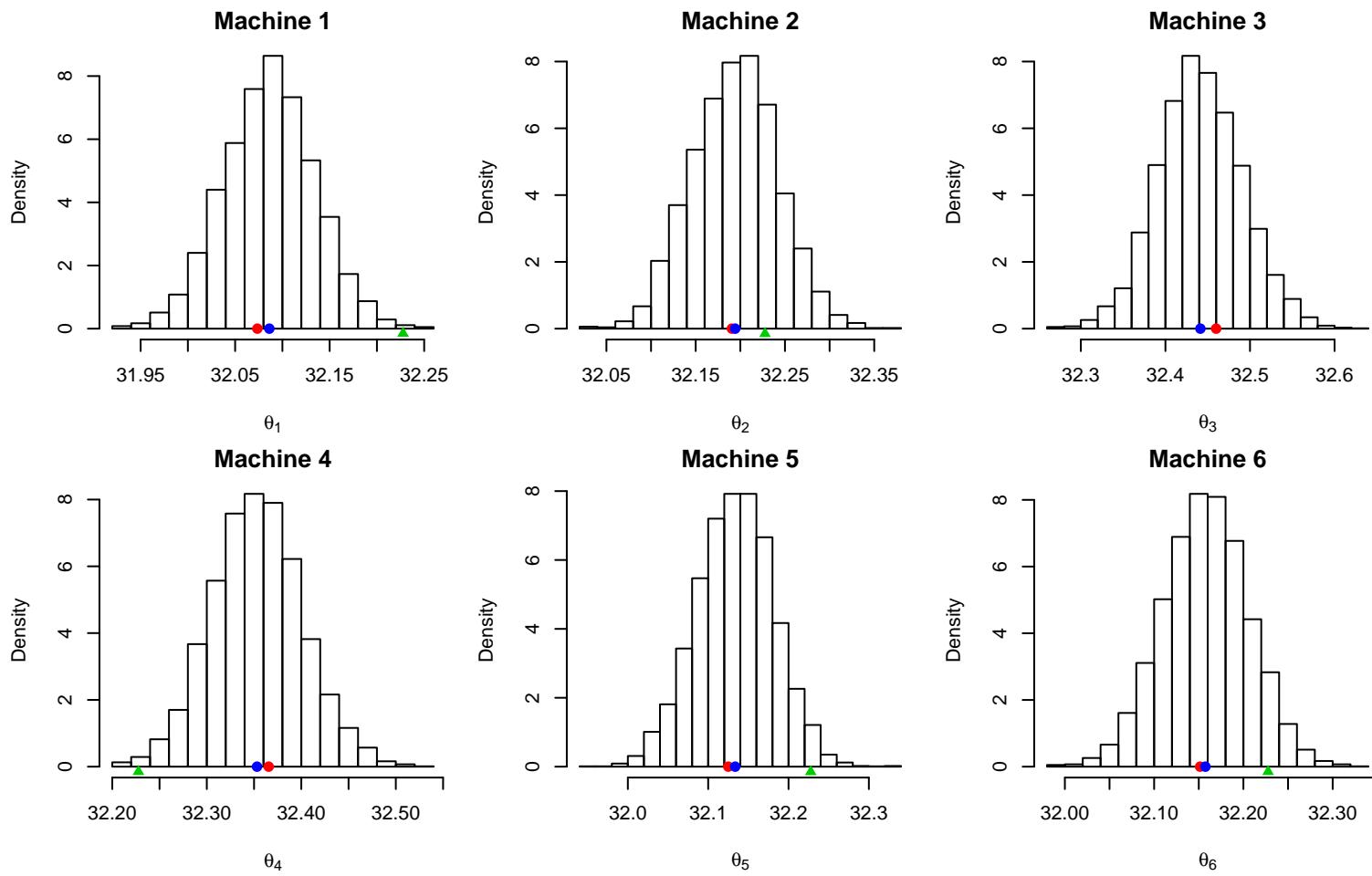
Now lets sample  $\theta_{j,k}$  from

$$\theta_{j,k} \sim N(\hat{\theta}_{j,k}, V_{j,k})$$

where

$$\hat{\theta}_{j,k} = \frac{\frac{1}{\sigma_j^2} \bar{y}_{.j} + \frac{1}{\tau_k^2} \mu_k}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}} \quad V_{j,k} = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}}$$

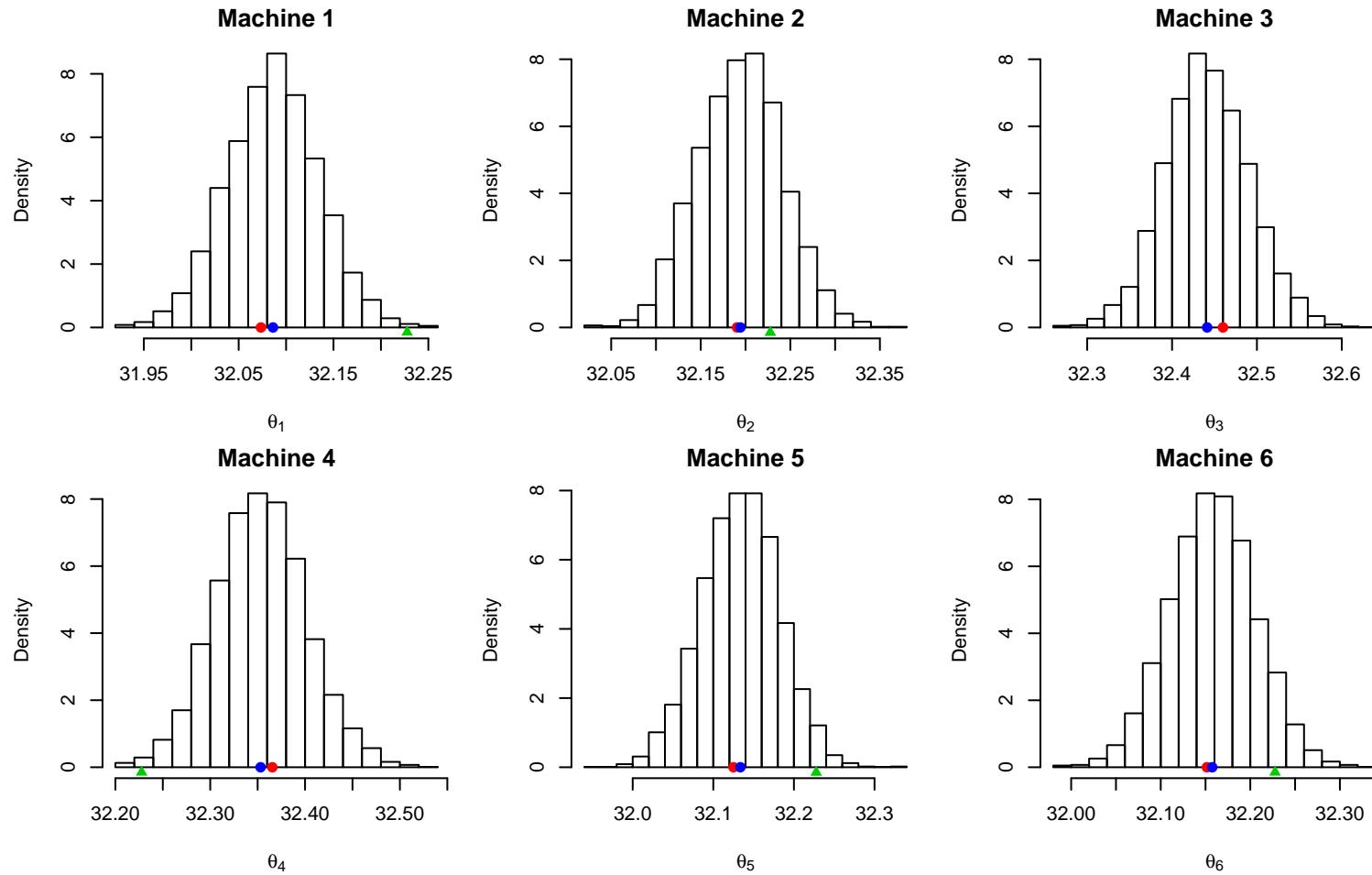
The histograms of these samples are



This plot suggests that we get some shrinkage in the estimate of the machine mean fills (posterior means are blue dots) from the sample averages (red dots). Note that the amount of shrinkage varies from machine to machine.

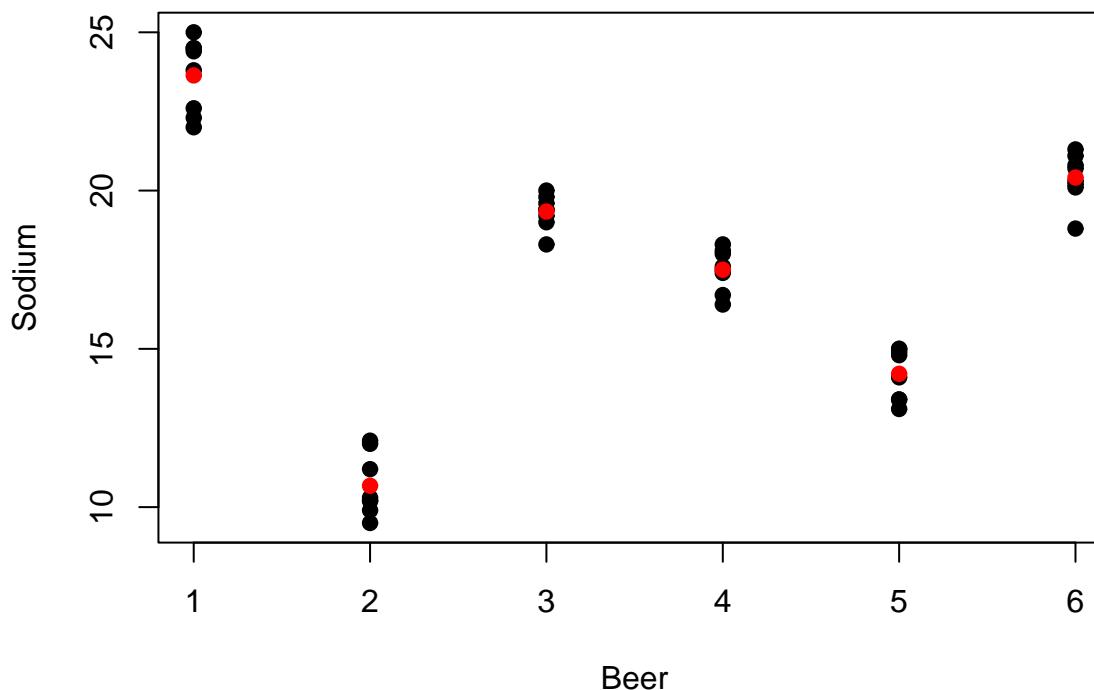
Also of interest is which machines have different fill levels. We can answer this by looking at  $P[\theta_i < \theta_j | y]$  for different pairs of machines.

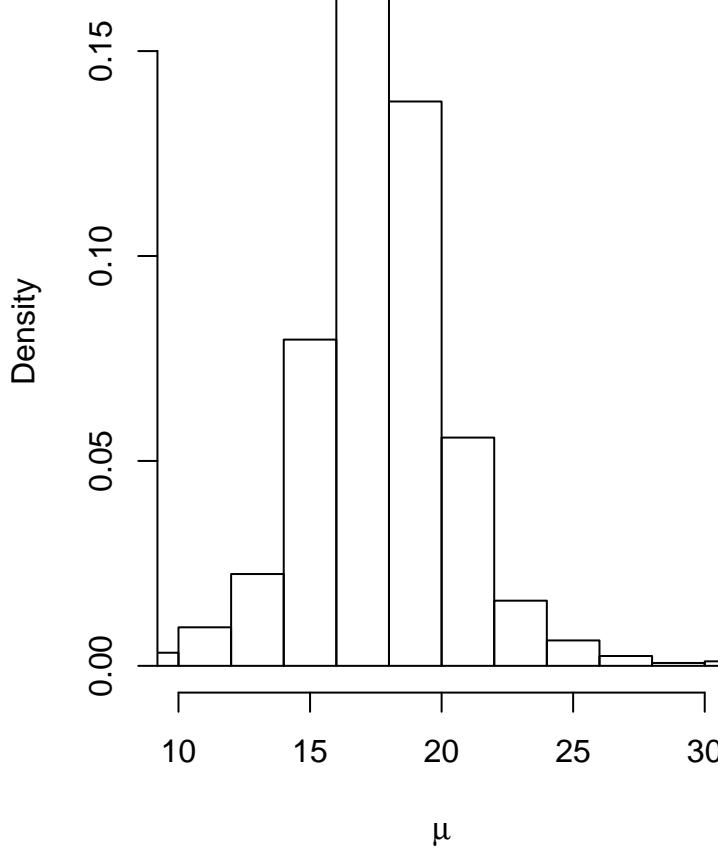
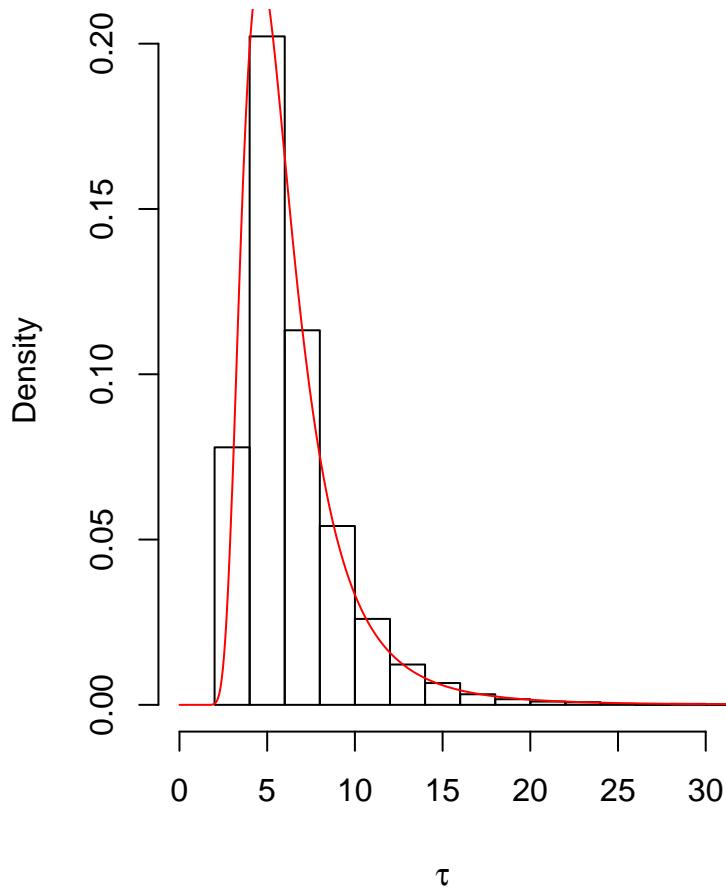
For example  $P[\theta_1 < \theta_3 | y] = 1$ , whereas  $P[\theta_1 < \theta_5 | y] = 0.7508$



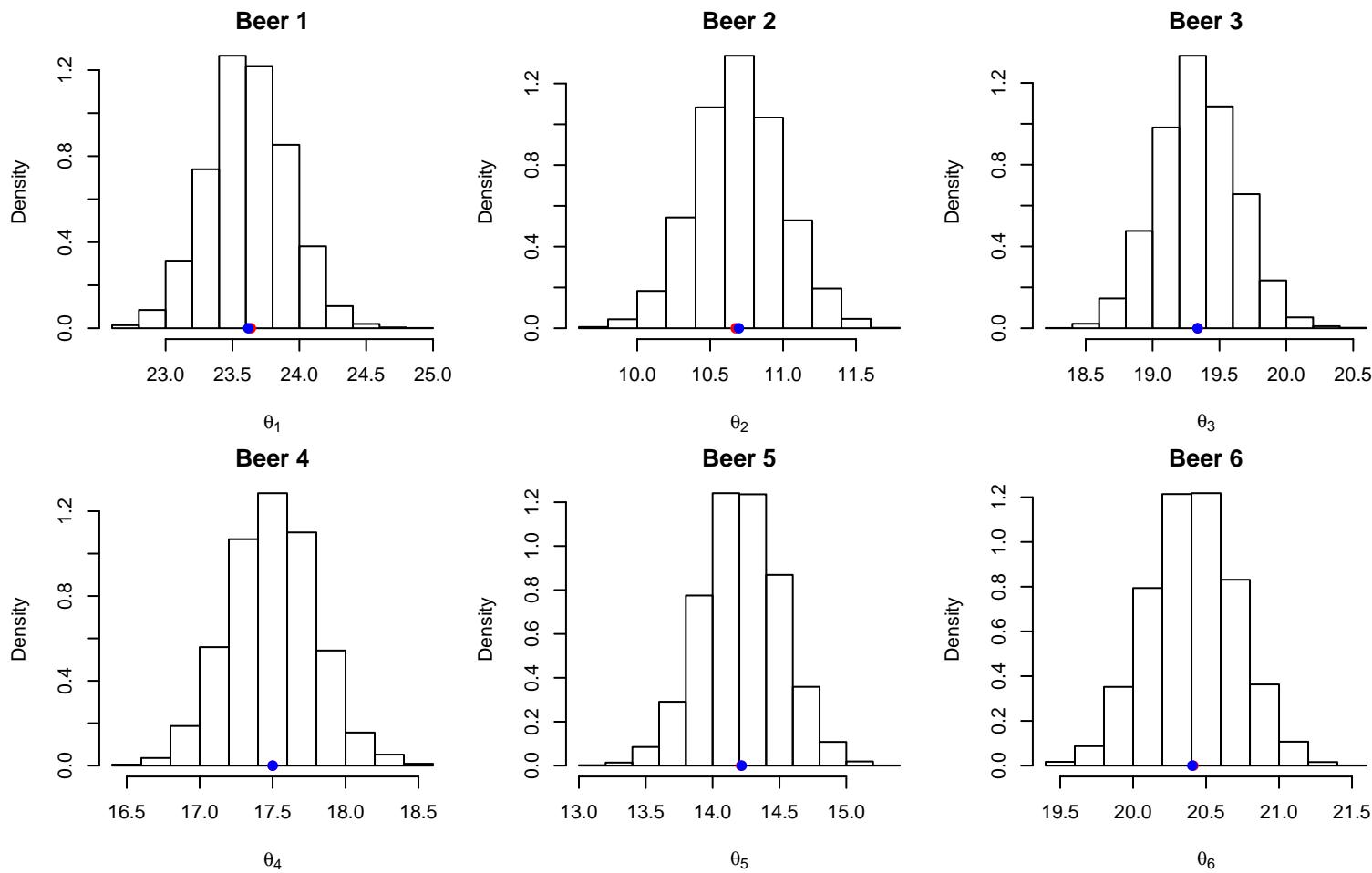
## Sodium Content in Beer

A study was done to investigate the sodium content of 6 randomly chosen brands of U.S. and Canadian beer. For each brand, 8 randomly chosen bottles or cans were analyzed to measure the sodium content (in mg) of each bottle or can. For this analysis,  $\sigma_j^2 = 0.0895$ , which again is based on the MSE from the 1-way ANOVA.

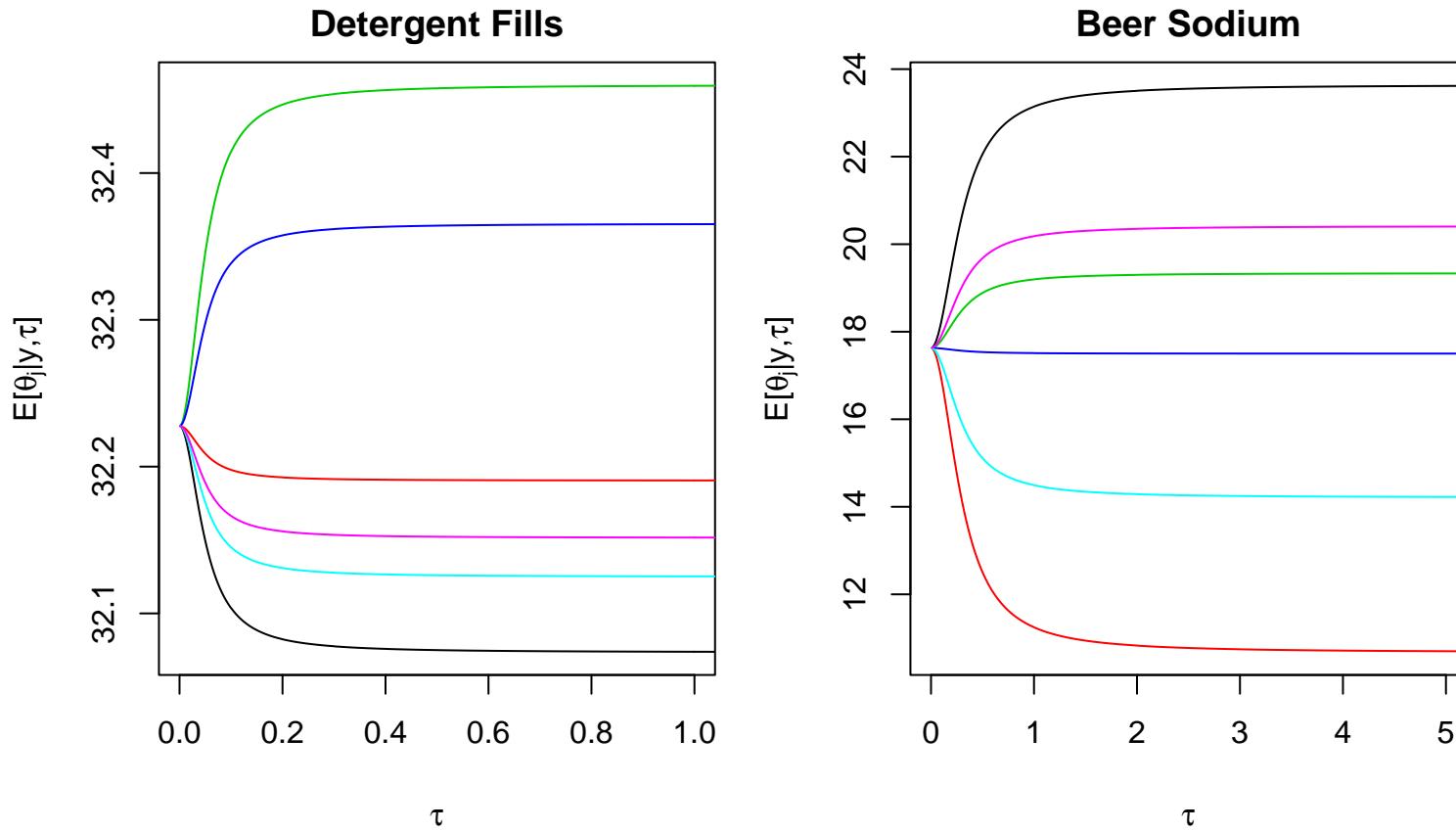




$$\begin{array}{ll}
 E[\tau|y] = 6.448 & E[\tau^2|y] = 50.847 \\
 \text{Mode}(\tau|y) = 4.61 & \text{Mode}(\tau^2|y) = 21.25 \\
 E[\mu|y] = 17.67 & SD(\mu|y) = 2.928
 \end{array}$$



There is much less shrinkage in this example. This is not surprising since  $\tau$  appears to be much bigger relative to  $\sigma_j^2$  in this example.



The relationship between the amount of shrinkage and  $\sigma_j^2$  and  $\tau^2$  can be seen by

$$E[\theta_i|\mu, \tau, y] = \frac{\tau^2}{\sigma_j^2 + \tau^2} \bar{y}_{.j} + \frac{\sigma_j^2}{\sigma_j^2 + \tau^2} \mu$$

# **Approximations based on Posterior Modes**

Statistics 220

Spring 2005



## Posterior Modes

As we have seen earlier, often as  $n \rightarrow \infty$

$$p(\theta|y) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where  $\hat{\theta}$  is the the posterior mode and  $I(\theta)$  is the observed information

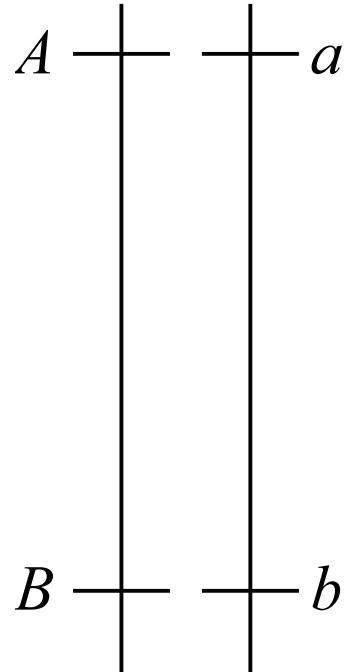
$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

While the asymptotics are interesting, the approximate normality is helpful in other situations.

1. Crude estimates as starting points for approximations
2. Normal (or related) mixture approximations to the posterior
3. Separate approximations for different marginal and conditional posterior distributions
4. Approximating distributions for Monte Carlo methods (e.g. a proposal in Metropolis-Hastings or in Importance Sampling)

To find the posterior mode and information, numerical methods often need to be used as closed form solutions usually aren't available.

## Example: Linkage Analysis (Rao, 1973, pp 268-269)



Two genes on a chromosome are separated by a recombination fraction  $\theta \leq \frac{1}{2}$

For an organism with joint haplotype  $AB|ab$ , there are 4 possible haplotypes that can be passed to its offspring

Haplotype	Probability
$AB$	$\frac{1-\theta}{2}$
$Ab$	$\frac{\theta}{2}$
$aB$	$\frac{\theta}{2}$
$ab$	$\frac{1-\theta}{2}$

An experiment was performed to estimate  $\theta$ . The breeding experiment used  $AB|ab \times AB|ab$  crosses and recorded the observed phenotypes. In this experiment, 2 dominant traits were observed ( $A$  dominant to  $a$  and  $B$  dominant to  $b$ ).

While there are 16 different possible joint haplotypes in the offspring (4 from the father times 4 from the mother), there are only 4 possible phenotypes.

Phenotype	Probability	Counts
$AB$	$\frac{3-2\theta+\theta^2}{4}$	125
$Ab$	$\frac{2\theta-\theta^2}{4}$	18
$aB$	$\frac{2\theta-\theta^2}{4}$	20
$ab$	$\frac{1-2\theta+\theta^2}{4}$	34

So the likelihood function is

$$p(y|\theta) = (3 - 2\theta + \theta^2)^{125} (2\theta - \theta^2)^{18+20} (1 - 2\theta + \theta^2)^{34}$$

Now lets put a truncated Beta prior on  $\theta$

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}I(\theta \leq 0.5)$$

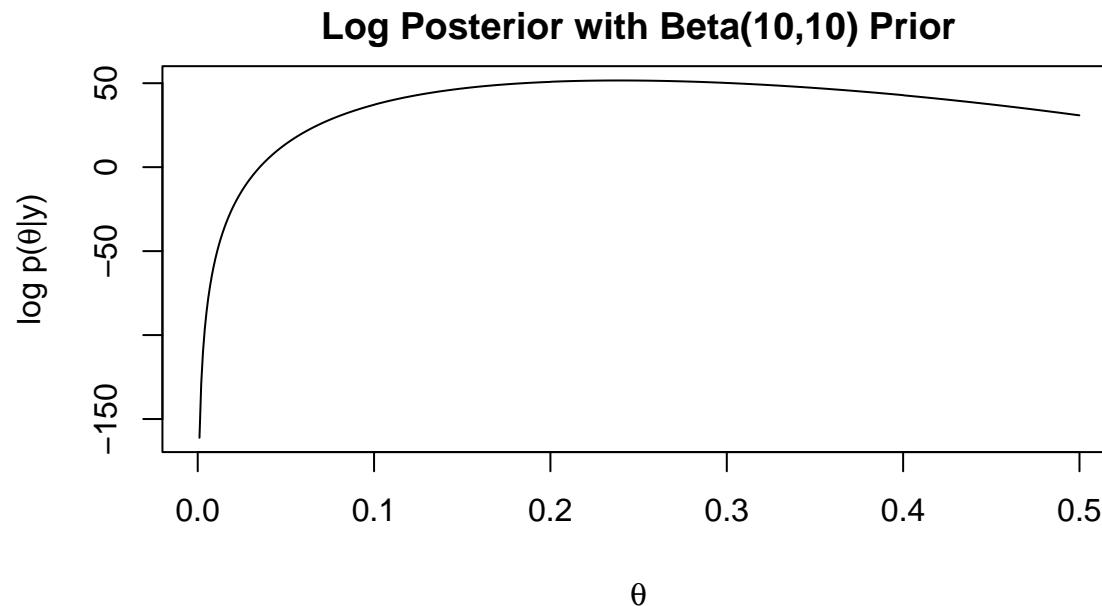
Thus the log posterior is

$$\begin{aligned}\log p(\theta|y) &= 125 \log(3 - 2\theta + \theta^2) + 38 \log(2\theta - \theta^2) + 34 \log(1 - 2\theta + \theta^2) \\ &\quad + (\alpha - 1) \log \theta + (\beta - 1) \log(1 - \theta) \\ &= 125 \log(3 - 2\theta + \theta^2) + 38 \log(2 - \theta) + \\ &\quad (\alpha + 37) \log \theta + (\beta + 67) \log(1 - \theta)\end{aligned}$$

Following the usual approach solving  $\frac{d}{d\theta} \log p(\theta|y) = 0$  to optimize gives

$$\frac{d}{d\theta} \log p(\theta|y) = \frac{125(2\theta - 2)}{3 - 2\theta + \theta^2} - \frac{38}{2 - \theta} + \frac{\alpha + 37}{\theta} - \frac{\beta + 67}{1 - \theta} = 0$$

This does not have an obvious closed form solution so we need to find another approach to maximizing the posterior (or log posterior).



There are a wide array of numerical approaches for optimizing functions. I want to discuss two

1. Newton-Raphson (and approximations)
2. EM algorithm

## Newton-Raphson

Following the text, let

$$L(\theta) = \log p(\theta|y)$$

Note that this can be an unnormalized density as

$$L_c(\theta) = \log cp(\theta|y) = L(\theta) + \log c$$

as the same optima ( $c$  can't be a function of  $\theta$ ). So we can also use

$$L(\theta) = \log p(y|\theta)p(\theta) = \log p(y|\theta) + \log p(\theta)$$

as the function to optimize.

So we want to solve the function

$$L'(\theta) = 0$$

where  $L'(\theta)$  is the vector of first partial derivatives (i.e. the gradient).

For Newton-Raphson, we also need  $L''(\theta)$ , the matrix of second partial derivatives.

The the Newton-Raphson algorithm is

1. Choose a starting value,  $\theta^0$
2. For  $t = 1, 2, 3, \dots$ 
  - (a) Compute  $L'(\theta^{t-1})$  and  $L''(\theta^{t-1})$ . The Newton method step at time  $t$  is based on the quadratic approximation to  $L(\theta)$  centered at  $\theta^{t-1}$ .
  - (b) Set the new iterate,  $\theta^t$ , to maximize the quadratic approximation

$$\theta^t = \theta^{t-1} - [L''(\theta^{t-1})]^{-1} L'(\theta^{t-1})$$

So for the example

$$L'(\theta) = \frac{250(\theta - 1)}{3 - 2\theta + \theta^2} - \frac{38}{2 - \theta} + \frac{\alpha + 37}{\theta} - \frac{\beta + 67}{1 - \theta}$$

$$L''(\theta) = \frac{250}{(3 - 2\theta + \theta^2)} - \frac{500(\theta - 1)^2}{(3 - 2\theta + \theta^2)^2} - \frac{38}{(2 - \theta)^2} - \frac{\alpha + 37}{\theta^2} - \frac{\beta + 67}{(1 - \theta)^2}$$

Note that Newton-Raphson is not guaranteed to converge. The starting point  $\theta^0$  can be very important, particularly when  $-L''$  is not positive definite.

One advantage to Newton-Raphson is that once you get close to the solution, the convergence is very fast (quadratic convergence). Also if the sequence won't converge, it is usually obvious quickly.

There is another advantage to Newton-Raphson in the Bayesian (or likelihood) framework. The update formula can be written as

$$\theta^t = \theta^{t-1} - [L''(\theta^{t-1})]^{-1} L'(\theta^{t-1}) = \theta^{t-1} + [I(\theta^{t-1})]^{-1} L'(\theta^{t-1})$$

Thus as we are determining the mode, we are also calculating the information matrix, and depending on how the update is done, we are also getting the asymptotic posterior variance matrix  $[I(\theta)]^{-1}$ .

Note that when implementing this, you usually do not want to invert  $I(\theta^{t-1})$ , but instead solve the system

$$L''(\theta^{t-1})\Delta\theta = L'(\theta^{t-1}) \quad \text{or} \quad I(\theta^{t-1})\Delta\theta^* = L'(\theta^{t-1})$$

and update with

$$\theta^t = \theta^{t-1} - \Delta\theta \quad \text{or} \quad \theta^t = \theta^{t-1} + \Delta\theta^*$$

as it is faster and more numerically stable.

## Approximations to Newton-Raphson

Note as described, Newton-Raphson requires the calculations of derivatives. However it is easy to approximate derivatives numerically. One approach is approximate the derivatives with

$$L'_i(\theta) = \frac{dL}{d\theta_i} \approx \frac{L(\theta + \delta_i e_i) - L(\theta - \delta_i e_i)}{2\delta_i}$$

and

$$\begin{aligned} L''_{ij}(\theta) &= \frac{d^2L}{d\theta_i d\theta_j} = \frac{d}{d\theta_j} \frac{dL}{d\theta_i} \\ &\approx \frac{L'_i(\theta + \delta_j e_j) - L'_i(\theta - \delta_j e_j)}{2\delta_j} \\ &\approx [L(\theta + \delta_i e_i + \delta_j e_j) - L(\theta - \delta_i e_i + \delta_j e_j) \\ &\quad - L(\theta - \delta_i e_i + \delta_j e_j) + L(\theta - \delta_i e_i - \delta_j e_j)] / (4\delta_i \delta_j) \end{aligned}$$

where  $e_i$  is the unit vector corresponding to the  $i$ th component of  $\theta$ .

$\delta_i$ , the size of the deviation to take along direction  $e_i$  depends on the scale of the problem, but should be small.

You don't want it too big as curvature of  $L$  will make this a poor approximation.

But you don't want it too small as to avoid round off error.

Often a value such as 0.0001 is reasonable.

# EM Algorithm

Dempster, Laird, and Rubin (1977)

EM = Expectation – Maximization

An approach for finding MLEs and posterior modes.

In the likelihood situation, it is often based on decomposing data  $X = (Y, Z)$  into observed ( $Y$ ) and missing parts ( $Z$ ). Want to maximize

$$p(y|\theta) = \int p(y, z|\theta) dz$$

In the Bayesian situation, its based on splitting  $\theta = (\phi, \gamma)$ , where you want to maximize over  $\phi$  after average over  $\gamma$ .

$$p(\phi|y) = \int p(\phi, \gamma|y) d\gamma$$

Want posterior mode of  $p(\phi|y)$  instead of  $p(\phi, \gamma|y)$ .

I will present thing in terms of the Bayesian solution. For the implementation in the likelihood situation, see the 221 notes on the course web site.

EM in this setting is based on the relationship

$$p(\phi|y) = \frac{p(\phi, \gamma|y)}{p(\gamma|\phi, y)}$$

Now lets take logs, giving

$$\log p(\phi|y) = \log p(\phi, \gamma|y) - \log p(\gamma|\phi, y)$$

Lets take expectation of both sides, with respect to the density  $p(\gamma|\phi^{\text{old}}, y)$ , where  $\phi^{\text{old}}$  is a current (old) guess of  $\phi$

$$\log p(\phi|y) = E_{\text{old}}[\log p(\phi, \gamma|y)] - E_{\text{old}}[\log p(\gamma|\phi, y)]$$

Let

$$Q(\phi|\phi^{\text{old}}) = E_{\text{old}}[\log p(\phi, \gamma|y)]$$

and

$$H(\phi|\phi^{\text{old}}) = E_{\text{old}}[\log p(\gamma|\phi, y)]$$

So

$$\log p(\phi|y) = Q(\phi|\phi^{\text{old}}) - H(\phi|\phi^{\text{old}})$$

It is possible to show that  $H(\phi|\phi^{\text{old}})$ , treated as a function of  $\phi$ , is maximized at  $\phi^{\text{old}}$ , i.e.

$$H(\phi|\phi^{\text{old}}) \leq H(\phi^{\text{old}}|\phi^{\text{old}}) \quad \text{for all } \phi$$

Now let  $\phi^{\text{new}}$  be any value of  $\phi$  such that

$$Q(\phi^{\text{new}}|\phi^{\text{old}}) \geq Q(\phi^{\text{old}}|\phi^{\text{old}})$$

Thus

$$\begin{aligned}\log p(\phi^{\text{new}}|y) &= Q(\phi^{\text{new}}|\phi^{\text{old}}) - H(\phi^{\text{new}}|\phi^{\text{old}}) \\ &\geq Q(\phi^{\text{old}}|\phi^{\text{old}}) - H(\phi^{\text{old}}|\phi^{\text{old}}) = \log p(\phi^{\text{old}}|y)\end{aligned}$$

This relationship is the main idea behind the Generalized EM (GEM) algorithm.

At each step, finding a  $\phi$  which leads to an increase of  $Q(\phi|\phi^{\text{old}})$  must lead to an increase in the marginal log posterior  $\log p(\phi|y)$ .

# Implementing the EM Algorithm

1. Start with a estimate of the parameter  $\phi^0$ .
2. For  $t = 1, 2, 3, \dots$ 
  - (a) E-step: Determine the expected log posterior density function

$$Q(\phi|\phi^{t-1}) = E_t[\log p(\phi, \gamma|y)] = \int \log p(\phi, \gamma|y)p(\gamma|\phi^{t-1}, y)d\gamma$$

- (b) M-step: Maximize the expected log posterior density

$$\phi^t = \arg \sup Q(\phi|\phi^{t-1})$$

For a GEM algorithm,  $\phi^{t-1}$  doesn't have to maximize  $Q(\phi|\phi^{t-1})$  but only satisfy  $Q(\phi^t|\phi^{t-1}) \geq Q(\phi^{t-1}|\phi^{t-1})$ .

## Example: Fatal Airline Accidents

$$\begin{aligned}y_i &\stackrel{iid}{\sim} Poisson(\lambda) \\ \lambda &\sim Exp(\mu) \\ \mu &\sim Gamma(\alpha, \beta)\end{aligned}$$

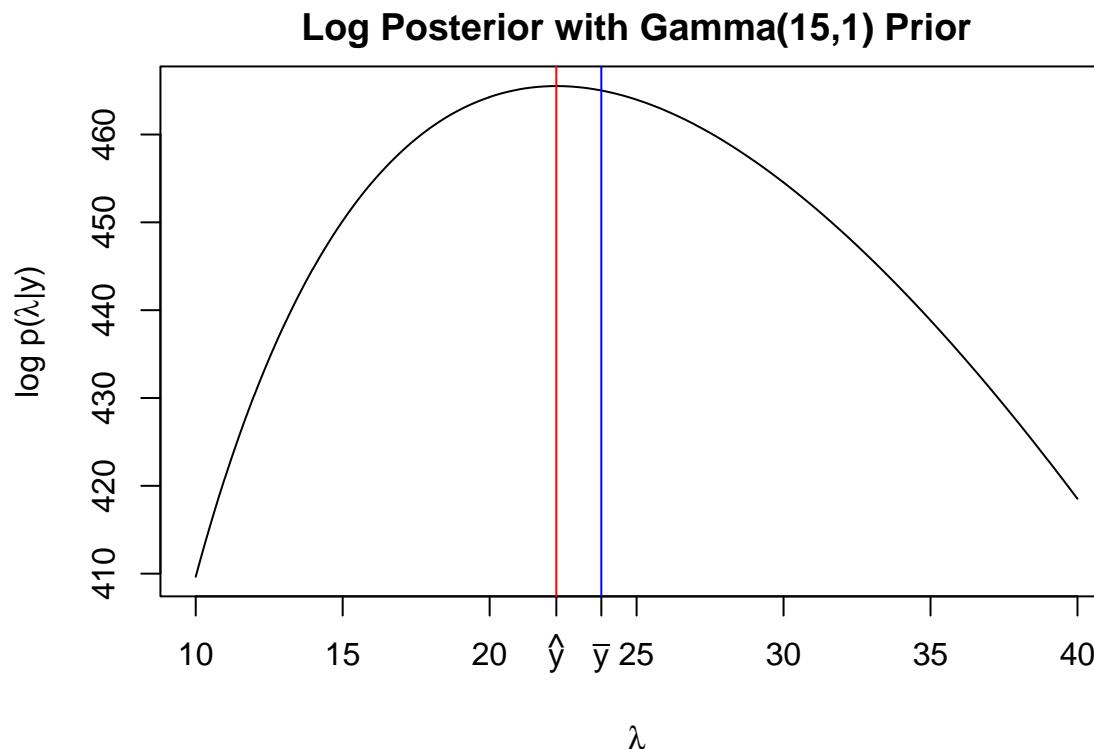
Want to maximize  $p(\lambda|y)$ . Note that

$$p(\lambda|y) \propto \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{(\lambda + \beta)^{\alpha+1}}$$

The mode is a solution to the quadratic equation

$$n\lambda^2 + (\alpha + 1 + n\beta - n\bar{y})\lambda - n\beta = 0$$

For the data in Table 2.2,  $n = 10$  and  $\bar{y} = 23.8$ . If  $\alpha = 15$  and  $\beta = 1$



Need for EM algorithm:

1.  $p(\lambda, \mu | y)$

$$p(\lambda, \mu | y) \propto \lambda^{\sum y_i} e^{-n\lambda} \mu e^{-\mu\lambda} \mu^{\alpha-1} \beta^\alpha \frac{e^{-\mu\beta}}{\Gamma(\alpha)}$$

Note that the sufficient statistic here is  $\bar{y} = 23.8$  and  $n = 10$

2.  $p(\mu | \lambda, y)$

$$\mu | \lambda, y \sim Gamma(\alpha + 1, \lambda + \beta)$$

### 3. E-step: find $Q(\lambda|\lambda^{t-1})$

$$\log p(\lambda, \mu|y) = \alpha \log \mu - \mu(\lambda + \beta) - n\lambda + n\bar{y} \log \lambda + c$$

$$\begin{aligned} Q(\lambda|\lambda^{t-1}) &= E_{\lambda^{t-1}}[\mu\lambda - n\lambda + n\bar{y} \log \lambda + c] \\ &= \lambda E_{\lambda^{t-1}}[\mu] + n\lambda + n\bar{y} \log \lambda + c \\ &= \lambda(E_{\lambda^{t-1}}[\mu] + n) + n\bar{y} \log \lambda + c \end{aligned}$$

since  $\mu|\lambda, y$  is  $Gamma(\alpha + 1, \lambda + \beta)$

$$E_{\lambda^{t-1}}[\mu] = \frac{\alpha + 1}{\lambda^{t-1} + \beta}$$

then

$$Q(\lambda|\lambda^{t-1}) = \lambda \left( \frac{\alpha + 1}{\lambda^{t-1} + \beta} + n \right) + n\bar{y} \log \lambda + c$$

#### 4. M-step:

$$\begin{aligned}\lambda^t &= \arg \sup Q(\lambda | \lambda^{t-1}) \\ &= \frac{n\bar{y}}{\frac{\alpha+1}{\lambda^{t-1}+\beta} + n} \\ &= \frac{n\bar{y}(\lambda^{t-1} + \beta)}{\alpha + 1 + n(\lambda^{t-1} + \beta)}\end{aligned}$$

Note that it can be shown that this sequence will converge to a root of

$$n\lambda^2 + (\alpha + 1 + n\beta - n\bar{y})\lambda - n\beta = 0$$

which is the same equation derived for  $\log p(\lambda|y)$  directly.

If  $\alpha = 15$  and  $\beta = 1$  and  $\lambda^0 = 15$ , the sequence of updates is

$t$	$\lambda^t$
0	15.00000
1	21.63636
2	22.22881
3	22.26630
4	22.26861
5	22.26875
6	22.26876
7	22.26876

Under some regularity conditions, for any GEM, the sequence  $\phi^1, \phi^2, \phi^3, \dots$  converges to a local mode of the posterior density.

Note that the proof of this result in Dempster, Laird, and Rubin (1979) wasn't quite right. Wu (1983) found valid conditions to indicate when this sequence would converge to a local mode.

**Theorem.** *Under some regularity conditions, for any GEM sequence  $\{\phi^t\}$ ,*

$$\log p(\phi^t | y) > \log p(\phi^{t-1} | y)$$

*if*

$$\phi^{t-1} \notin \Phi = \left\{ \phi : \frac{d}{d\phi} \log p(\phi | y) = 0 \right\}$$

Thus you will continue to increase the posterior density until you hit a local mode.

The proof of the theorem depends on the fact that

$$\frac{d}{d\phi} \log p(\phi|y) = \frac{d}{d\phi} Q(\phi|\phi)$$

when the derivative of  $Q$  is taken with respect to the first  $\phi$ . Thus if you are at a mode,  $Q$  must have a 0 derivative, implying you can't increase  $Q$ .

The EM algorithm has linear convergence, thus it tends to converge slower than algorithms like Newton-Raphson. However it does have the advantage that it is guaranteed to converge, unlike Newton-Raphson.

For this algorithm to be feasible, it must be possible to maximize  $Q$  easily, or at least find a value which increases it. Thus the EM algorithm isn't feasible for all problems. However there are a number of extensions that can make some of the more difficult problems feasible.

# Generalized Linear Models

Statistics 220

Spring 2005



# Generalized Linear Models

For many problems, standard linear regression approaches don't work. Sometimes, transformations will help, but not always. Generalized Linear Models are an extension to linear models which allow for regression in more complex situations.

As before let  $y$  be the response variable and  $X$  be the predictor variables. We want to determine  $p(y|X)$ , which will depend on some parameters  $\beta$  and  $\phi$ .

- Non-linearity, multiplicative effects and errors
- Bounded responses
- Discrete responses

Generalized linear model involve the following 4 pieces.

1. Linear predictor:  $\eta = X\beta$
2. Link function  $g(\cdot)$ : Relates the linear predictor to the mean of the outcome variable

$$g(\mu) = \eta = X\beta \quad \mu = g^{-1}(\eta) = g^{-1}(X\beta)$$

3. Distribution: What is the distribution of the response variable  $y$ . These are usually a member of the exponential family which includes, normal, lognormal, poisson, binomial, gamma, hypergeometric.
4. Dispersion parameter  $\phi$ : Some distributions have an additional parameter dealing with the spread of the distribution. The form of this usually depends on the relationship between the mean and the variance. With some distributions, this is fixed (e.g. Poisson or binomial), while with others it is an additional parameter to be modelled and estimated (e.g. normal or gamma).

Normal linear regression is a special case of a generalized linear model where

1. Linear predictor:  $\eta = X\beta$
2. Link function:  $g(\mu) = \mu = X\beta$  (Identity Link)
3. Distribution:  $y_i|x_i, \beta, \sigma^2 \sim N(x_i^T\beta, \sigma^2)$
4. Dispersion parameter:  $\sigma^2$

We have seen other GLIMs in class this term.

- Poisson regression:

The usual form for Poisson regression uses the log link

$$\log \mu = X\beta \quad \mu = \exp(X\beta)$$

Another example used the identity link

$$\mu = X\beta$$

This is less common as the mean of a Poisson random variable must be positive which the identity link doesn't guarantee. The log link however does, which is one reason it is very popular.

As mentioned earlier the dispersion parameter  $\phi = 1$  is fixed as  $\text{Var}(y) = \mu$  for the Poisson distribution.

- Logistic regression:

This form is slightly different as we work with the mean of the sample proportion  $\frac{y_i}{n_i}$  instead the mean of  $y_i$ .

Logistic regression is based on  $y_i|x_i \sim Bin(n_i, \mu_i)$  where  $\mu_i$  is a function of  $x_i$ . The link function is

$$g(\mu) = \log \frac{\mu}{1 - \mu}$$

i.e. the log odds ratio.

The inverse link function gives

$$\mu = g^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

Thus the likelihood is

$$p(y|\beta) = \prod_{i=1}^n \binom{n_i}{y_i} \left( \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \right)^{y_i} \left( \frac{1}{1 + e^{X_i\beta}} \right)^{n_i - y_i}$$

The dispersion parameter  $\phi = 1$

## Link Functions

Changing the link functions allows for different relationships between the response and predictor variables. The choice of link function  $g(\cdot)$  should be made so that the relationship between the transformed mean and the predictor variables is linear.

Note transforming the mean via the link function is different from transforming the data

For example consider the two models

1.  $\log y_i | X_i, \beta \sim N(X_i\beta, \sigma^2)$  or equivalently  $y_i | X_i, \beta \sim \text{log}N(X_i\beta, \sigma^2)$

$$E[y_i | X_i, \beta] = \exp\left(X_i\beta + \frac{\sigma^2}{2}\right)$$

and

$$\text{Var}(y_i | X_i, \beta) = \exp(2(X_i\beta + \sigma^2))(\exp(\sigma^2) - 1)$$

2.  $y_i|X_i, \beta \sim N(\mu_i, \sigma^2)$  where  $\log \mu_i = X_i\beta$ ,  $\mu_i = \exp(X_i\beta)$  (normal model with log link)

The first model has a different mean and the variability depends on  $X$  whereas the variability in the second model does not depend on  $X$ .

When choosing a link function, you often need to consider the plausible values of the mean of the distribution.

For example, with binomial data, the success probability must be in  $[0,1]$ . However  $X\beta$  can take values on  $(-\infty, \infty)$ .

Thus you can get into trouble with binomial data with the model  $\mu = X\beta$  (identity link).

Possible choices include

- Logit link:

$$g(\mu) = \log \frac{\mu}{1 - \mu}$$

- Probit link:

$$g(\mu) = \Phi^{-1}(\mu) \quad (\text{Standard Normal Inverse CDF})$$

- Complementary Log-Log link

$$g(\mu) = \log(-\log(\mu))$$

All of these happen to be quantile functions for different distributions.

Thus the inverse link functions are CDFs

- Logit link:

$$g^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta} \quad (\text{Standard Logistic})$$

- Probit link:

$$g^{-1}(\eta) = \Phi(\eta) \quad (N(0, 1))$$

- Complementary Log-Log link:

$$g^{-1}(\eta) = e^{-e^\eta} \quad (\text{Gumbel})$$

Thus in this case any distribution defined on  $(-\infty, \infty)$  could be the basis for a link function, but these are the popular ones. One other choice that is used are based on  $t_\nu$  distributions as they have some robustness properties.

Note that a link function doesn't have to have the property of mapping the range of the mean to  $(-\infty, \infty)$ .

We've seen the identity link ( $g(\mu) = \mu$ ) in Poisson regression and it is also used in binomial problem.

In the binomial case, it can be reasonable if the success probabilities lie in the range  $(0.2, 0.8)$ .

Similarly, an inverse link function doesn't have to have to map  $X\beta$  back to the whole range of the mean for a distribution.

For example, the log link will only give positive means ( $\mu = e^\eta$ ). This can be an useful model with normal data, even though in general a normal mean can take any value.

# Common Link Functions

The following are common link function choices for different distributions

- Normal
  - Identity:  $g(\mu) = \mu$
  - Log:  $g(\mu) = \log \mu$
  - Inverse:  $g(\mu) = \frac{1}{\mu}$
- Binomial
  - Logit:  $g(\mu) = \log \frac{\mu}{1-\mu}$
  - Probit:  $g(\mu) = \Phi^{-1}(\mu)$
  - Complementary Log-Log link:  $g(\mu) = \log(-\log(1-\mu))$
  - Log:  $g(\mu) = \log \mu$

- Poisson

- Log:  $g(\mu) = \log \mu$
- Identity:  $g(\mu) = \mu$
- Square root:  $g(\mu) = \sqrt{\mu}$

- Gamma

- Inverse:  $g(\mu) = \frac{1}{\mu}$
- Log:  $g(\mu) = \log \mu$
- Identity:  $g(\mu) = \mu$

- Inv-Normal

- Inverse squared:  $g(\mu) = \frac{1}{\mu^2}$
- Inverse:  $g(\mu) = \frac{1}{\mu}$
- Log:  $g(\mu) = \log \mu$
- Identity:  $g(\mu) = \mu$

The first link function mentioned for each distribution is the canonical link which is based on the writing the density of each distribution in the exponential family form.

$$p(y|\theta) = f(y)g(\theta) \exp(\phi(\theta)^T u(y))$$

# Dispersion Parameter

So far we have only discussed the mean function. However we also need to consider the variability of the data as well. For any distribution, we can consider the variance to be a function of the mean ( $V(\mu)$ ) and a dispersion parameter ( $\phi$ )

$$\text{Var}(y) = \phi V(\mu)$$

The variance functions and dispersion parameters for the common distributions are

Distribution	$N(\mu, \sigma^2)$	$Pois(\mu)$	$Bin(n, \mu)$	$Gamma(\alpha, \nu)$
$V(\mu)$	1	$\mu$	$\mu(1 - \mu)$	$\mu$
$\phi$	$\sigma^2$	1	$\frac{1}{n}$	$\frac{1}{\nu}$

Note for the Gamma distribution, the form of these can depend on how the distribution is parameterized. (McCullagh and Nelder have different formulas due to this.)

So when building models we need models for dealing with the dispersion in the data. Exactly how you want to do this will depend on the problem.

## Overdispersion

Often data will have more variability than might be expected.

For example, consider Poisson like data and consider a subset of data which has the same levels of the predictor variables (call it  $y_1, y_2, \dots, y_m$ ).

If the data is Poisson, the sample variance should be approximately the sample mean

$$s_m^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 \approx \bar{y}$$

If  $s_m^2 > \bar{y}$ , this suggests that there is more variability than can be explained by the explanatory variables.

This extra variability can be handled in a number of ways.

One approach is to add in some additional variability into the means.

$$\begin{aligned} y_i | \mu_i &\stackrel{\text{ind}}{\sim} \text{Pois}(\mu_i) \\ \mu_i | X_i, \beta, \sigma^2 &\stackrel{\text{ind}}{\sim} N(X_i \beta, \sigma^2) \end{aligned}$$

In this approach every observation with the same level of the explanatory variables will have a different mean, which will lead to more variability in the  $y$ s.

$$\begin{aligned} \text{Var}(y_i) &= E[\text{Var}(y_i | \mu_i)] + \text{Var}(E[y_i | \mu_i]) \\ &= E[\mu_i] + \text{Var}(\mu_i) \\ &= X_i \beta + \sigma^2 \geq X_i \beta = E[y_i] \end{aligned}$$

Note that normally you probably model  $\log \mu_i \sim N(X_i \beta, \sigma^2)$ , but showing the math was easier with the identity link instead of the log link.

# Bayesian Approach to Generalized Linear Models

So far there really hasn't been any Bayes in the discussion so far. Lets now look at the complete model with priors added assuming no overdispersion.

$$\begin{aligned} y_i | \mu_i, \phi &\stackrel{ind}{\sim} p(y_i | \mu_i, \phi) \\ g(\mu_i) &= X_i \beta \quad \text{or } \mu_i = g^{-1}(X_i \beta) \\ \beta &\sim p(\beta) \\ \phi &\sim p(\phi) \end{aligned}$$

The likelihood piece of the model may not correspond to the usual parameterization of the model, and thus the usual parameters will have to be calculated from  $\mu_i$  and  $\phi$ .

The model for  $\beta$  is general and could have an additional hierarchical structure following the ideas from chapter 15.

For example, if two of the explanatory variable are categorical factors, the  $\beta$  will need to have an ANOVA like structure to account for these.

In addition, the dispersion parameter in this framework is the same for all observations, but it could be allowed to vary from observation to observation by having it depend on  $X_i$ . The classical approach to GLIMs usually doesn't do this.

To allow for overdispersion a distribution can be put on  $\mu_i$ , or what is usually easier on  $g(\mu_i)$

$$g(\mu_i) \sim p(\mu_i | X_i \beta)$$

Note that  $g(\mu_i) = X_i \beta$  is a special case of this.

## Choice of Priors

There are a number of approaches used for putting priors on the regression and dispersion parameters. The usual approach factorizes the prior as

$$p(\beta, \phi) = p(\beta|\phi)p(\phi)$$

Often  $\beta$  and  $\phi$  will have independent priors.

Then the prior on  $\beta$  is commonly done in one of the three ways

- Noninformative: A flat improper prior can be put on  $\beta$ , which yields the classical analysis of the GLIM. Thus the MLE is the posterior mode. This can be determined with standard GLIM software (such as `gls` in R). Approximate inference can be obtained by a normal approximation to the likelihood.

- Conjugate: A conjugate prior can be implemented by  $n_0$  idealized observations  $y_0$  with covariate matrix  $X_0$ . Inference is carried out by performing analysis on the augmented response variable  $\begin{pmatrix} y \\ y_0 \end{pmatrix}$  with augmented covariate matrix  $\begin{pmatrix} X \\ X_0 \end{pmatrix}$  and a flat prior on  $\beta$ . Then inference is performed as in the noninformative prior case.
- Non-conjugate: The most common approach is to express prior information directly on  $\beta$ . A common approach is to use

$$\beta \sim N(\beta_0, \Sigma_\beta)$$

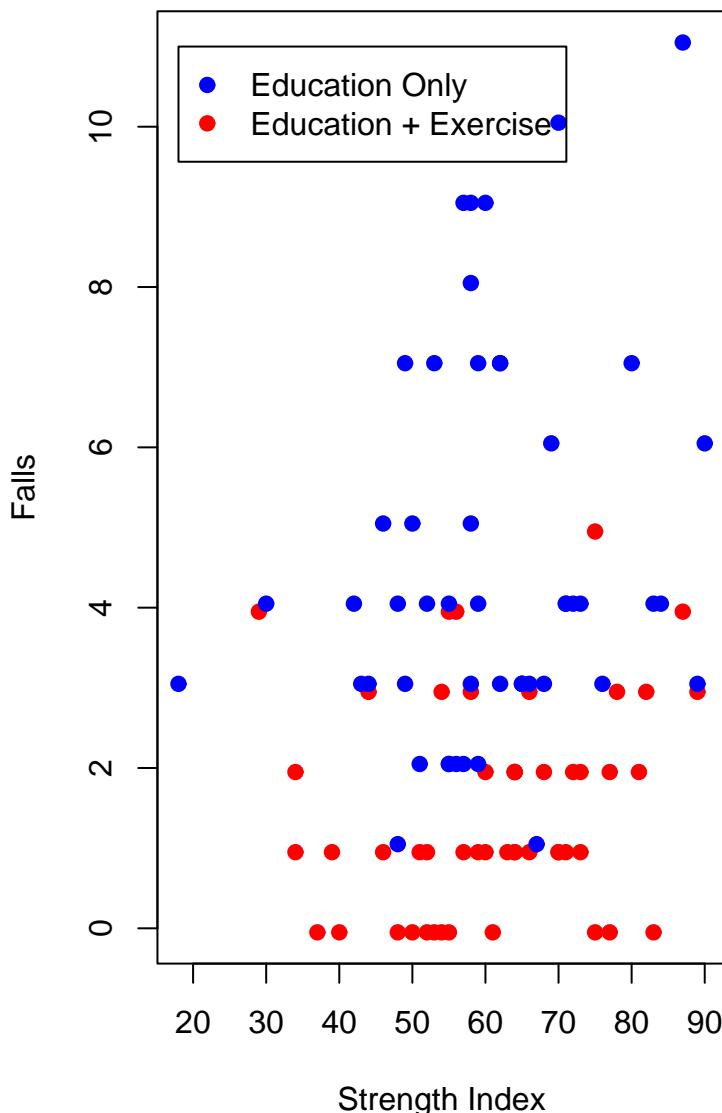
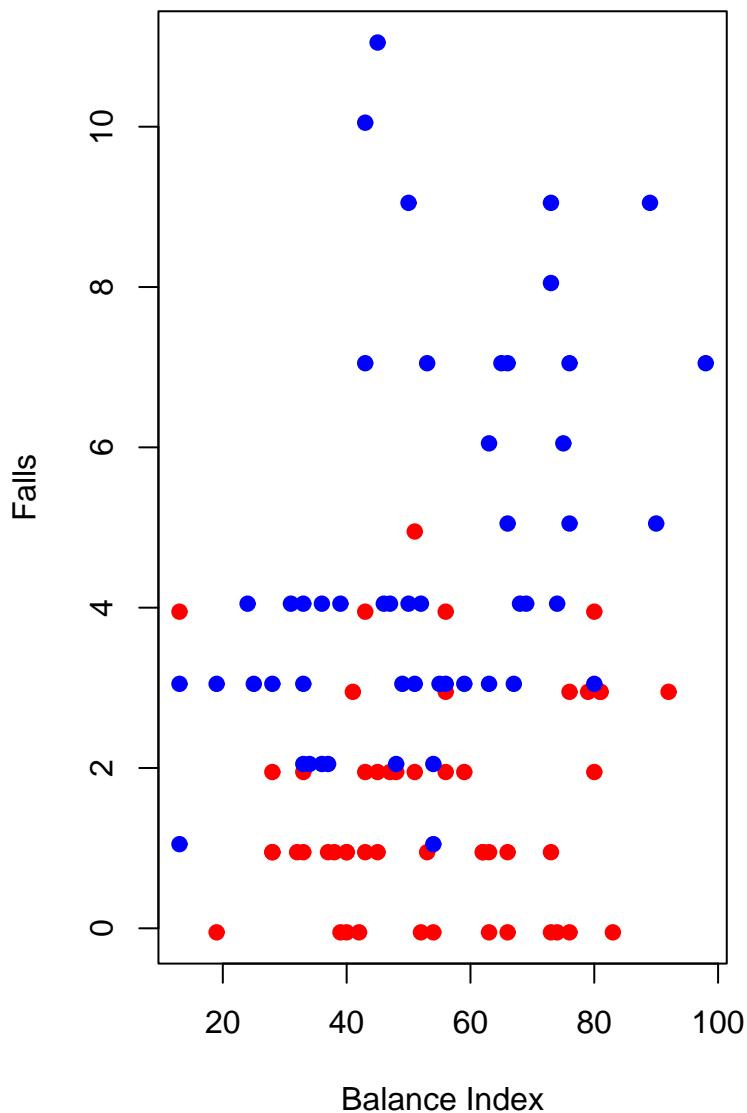
While the normal is common, other distributions can be used. The normal is convenient when approximate inference based on normal approximations are used. It also has the advantage that it fits into the way people often think of uncertainty, in terms of means and standard deviations (or variances).

## Example

### Geriatric Study to Reduce Falls

100 subject were studied to investigate two treatments to which is better to reduce falls.

- $y$ : number of falls during 6 months of study (self-reported)
- $x_1$ : Treatment - 0 = education only, 1 = education + aerobic exercise
- $x_2$ : Gender - 0 = female, 1 = male
- $x_3$ : Balance Index (bigger is better)
- $x_4$ : Strength Index (bigger is better)



## 1. Overdispersion

$$\begin{aligned}y_i | \mu_i &\stackrel{ind}{\sim} Pois(\mu_i) \\ \log \mu_i | \beta, \sigma^2 &\stackrel{ind}{\sim} N(X_i \beta, \sigma^2) \\ \beta | \sigma_\beta^2 &\sim N(0, \sigma_\beta^2 I) \\ \sigma_\beta^2 &\sim \text{Inv-Gamma}(0.001, 0.001) \\ \sigma^2 &\sim U(0, 1000)\end{aligned}$$

## 2. No Overdispersion

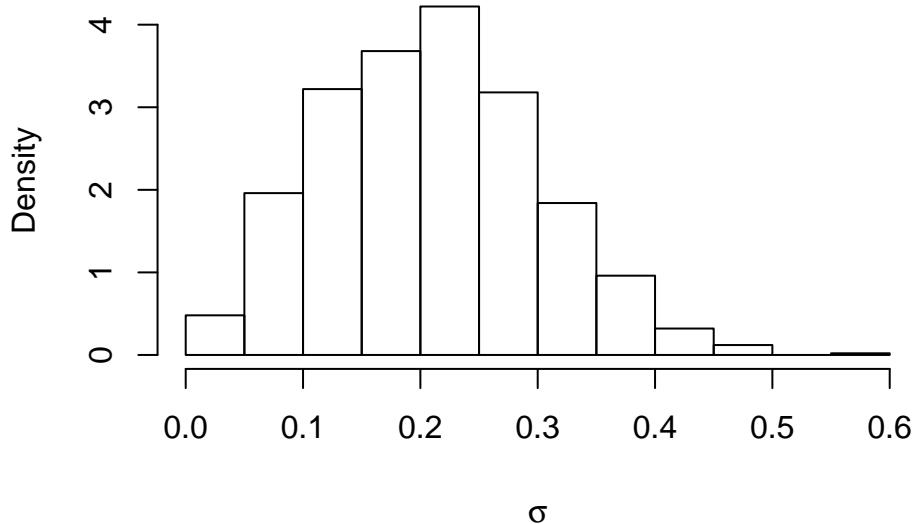
$$\begin{aligned}y_i | \mu_i &\stackrel{ind}{\sim} Pois(\mu_i) \\ \log \mu_i &= X_i \beta \\ \beta | \sigma_\beta^2 &\sim N(0, \sigma_\beta^2 I) \\ \sigma_\beta^2 &\sim \text{Inv-Gamma}(0.001, 0.001)\end{aligned}$$

First lets examine whether we need the overdispersion parameter  $\sigma^2$

First lets look at its posterior distribution

$$E[\sigma|y] = 0.21$$

$$\text{SD}(\sigma|y) = 0.09$$



This suggests that if there is any overdispersion, it must be small.

In addition lets look at the DICs

Model	$DIC$	$p_D$
Overdispersion	402.9	41
No Overdispersion	378.9	6.1

This implies we are getting a better fit without the overdispersion.

Now lets examine whether the treatment has any effect based on the No Overdispersion model

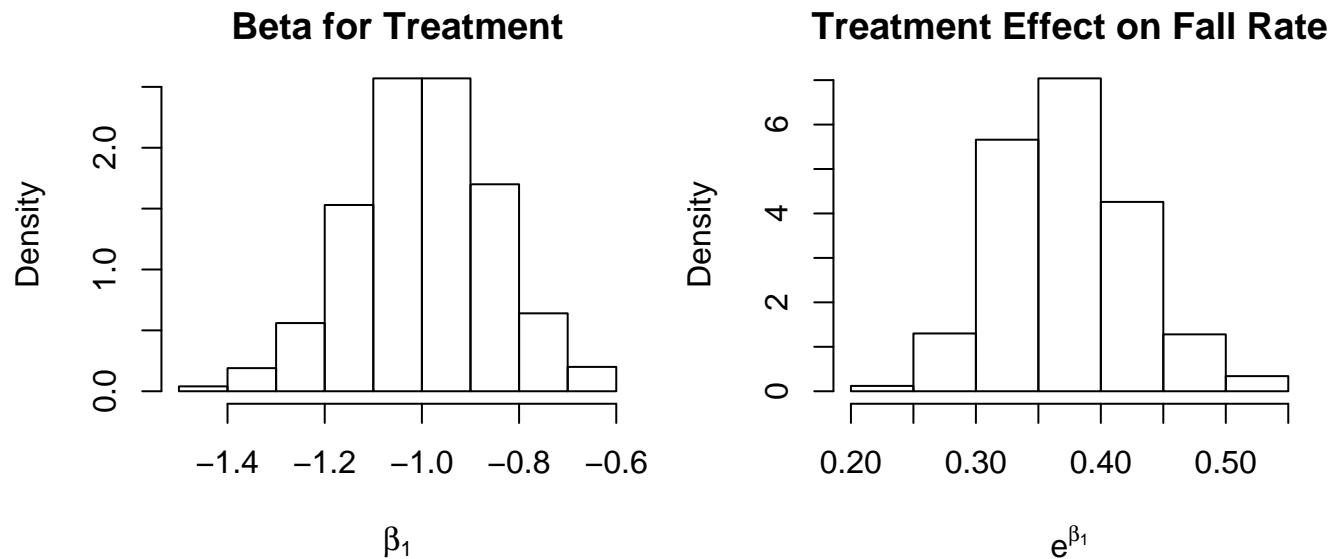
Parameter	$E[\beta_j y]$	$SD(\beta_j y)$
Intercept	0.443	0.338
Treatment	-1.010	0.139
Gender	-0.050	0.122
Balance	0.010	0.003
Strength	0.009	0.004

There is strong evidence that the aerobic exercise helps as  $E[\beta_1|y] < 0$  ( $P[\beta_1 < 0|y] = 1$ ).

As we are using a log link here

$$\mu_i = e^{\beta_1 x_{1i}} \exp(\beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i})$$

If everything else is kept the same, exercise should lower the fall rate by a factor of  $0.36 = e^{-1.010}$ . So the exercise should lower the fall rate to about a third of what it would be with education only.



Based on the posterior distribution of  $e^{\beta_1}$  we have strong evidence that the fall rate should at least halve.

Parameter	$E[\beta_j y]$	$SD(\beta_j y)$
Intercept	0.443	0.338
Treatment	-1.010	0.139
Gender	-0.050	0.122
Balance	0.010	0.003
Strength	0.009	0.004

It appears that there is no significant gender effect as  $\beta_2$  appears to be close to zero.

There is at first look a slight surprising result for  $\beta_3$  and  $\beta_4$ . At first thought you might think that people that have better balance and strength might fall less. However these parameter estimates suggest that they fall more. One possible explanation is that the stronger people are more active, meaning they might have more opportunity to fall.

## Scope

- Sequential analysis of dynamic models
- Fixed parameters & time-varying state variables
- Discrete numerical posterior approximations
  - Historical perspectives
  - Simulation methods
  - Smoothing & regenerating parameter samples
  - Kernel methods, evolution methods
- Combined particle filtering algorithms (APF+)
- Multiple factor models in finance
  - e.g., 30+ parameters, 3+ states
- Open questions, directions

## Models, Notation, Goals

- Data  $\mathbf{y}_t$  ( $t = 1, 2, \dots$ )
- Information at time  $t$  :  $D_t = \{D_{t-1}, \mathbf{y}_t\}$
- State vector/variables  $\mathbf{x}_t$
- Fixed model parameters  $\boldsymbol{\theta}$
- Observation model

$$p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\theta})$$

- Markov evolution model

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \boldsymbol{\theta})$$

**Goals:** sequentially update posteriors

$$\cdots \rightarrow p(\mathbf{x}_t, \boldsymbol{\theta} | D_t) \rightarrow p(\mathbf{x}_{t+1}, \boldsymbol{\theta} | D_{t+1}) \rightarrow \cdots$$

**Numerical Approximations (points/weights):**

$$\{\mathbf{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)} : j = 1, \dots, N_t\}$$

and

$$\{\omega_t^{(j)} : j = 1, \dots, N_t\}$$

## Some History in Statistics

### Mixture models:

- Discrete grids of parameters
- Mixtures of normals/mixtures of Kalman filters
- Harrison (71+), Alspach & Sorenson (71+)
- Multi-processes (West & Harrison, 97 chap. 12)

### Adaptive Quadrature methods:

- Pole and West (88, 89, 90)
- Combined parameters and state variables
- **Parameter grids evolve/adapt in time**
- **Exploit Markov evolution model to generate new points & interpolate**
- Efficient numerical integration
- ... but, limited to small dimensions

### Simulation methods in 90s:

- Longer history in engineering
- New directions in particle filtering (bootstraps, SIR, APF, ...)

## APF (Auxiliary Particle Filter) for States

(Pitt & Shephard 99)

Theoretical update:

$$p(\mathbf{x}_{t+1}|D_{t+1}) \propto p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})p(\mathbf{x}_{t+1}|D_t)$$

MC approximation to “prior”:

$$p(\mathbf{x}_{t+1}|D_t) \approx \sum_{k=1}^N \omega_t^{(k)} p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(k)})$$

APF state update from  $t \rightarrow t + 1$ :

- for each  $k$ ,
  - “estimates”  $\boldsymbol{\mu}_{t+1}^{(k)} = E(\mathbf{x}_{t+1}|\mathbf{x}_t^{(k)})$
  - and weights  $g_{t+1}^{(k)} \propto \omega_t^{(k)} p(\mathbf{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(k)})$
- sample (aux) indicators  $j$  with probs  $g_{t+1}^{(j)}$
- **time  $t + 1$  samples:**  $\mathbf{x}_{t+1}^{(j)} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(j)})$
- **and weights:**

$$\omega_{t+1}^{(j)} = \frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^{(j)})}{p(\mathbf{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)})}$$

## Filtering for Parameters & States

### Theoretical update:

$$p(\mathbf{x}_{t+1}, \boldsymbol{\theta} | D_{t+1}) \propto p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}, \boldsymbol{\theta}) p(\mathbf{x}_{t+1}, \boldsymbol{\theta} | D_t)$$

with

$$p(\mathbf{x}_{t+1}, \boldsymbol{\theta} | D_t) = p(\boldsymbol{\theta} | D_t) \int p(\mathbf{x}_{t+1} | \mathbf{x}_t, \boldsymbol{\theta}) p(\mathbf{x}_t | \boldsymbol{\theta}, D_t) d\mathbf{x}_t$$

where  $p(\boldsymbol{\theta} | D_t)p(\mathbf{x}_t | \boldsymbol{\theta}, D_t)$  “available” as sample points/weights

### Particle filtering on parameters:

- include  $\boldsymbol{\theta}$  in state, but
- ? degenerating weights on fixed  $\boldsymbol{\theta}$  points
- ? explicit need for density function  $p(\boldsymbol{\theta} | D_t)$ 
  - unavailable in strict sequential context –

### Needs:

- Regeneration/new points
- Smoothing/interpolation of prior samples & weights

## Treatment of Parameters: AE

“Artificial Evolution” – Gordon et al 93:

Add “noise” to parameters between  $t$  and  $t + 1$

– parameter becomes state variable –

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\nu}_{t+1} \quad \text{with} \quad \boldsymbol{\nu}_{t+1} \sim \text{indep } N(\mathbf{0}, \mathbf{W}_{t+1})$$

resulting in

$$p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t) = N(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \mathbf{W}_{t+1})$$

Incorporate parameter in particle filtering ...

e.g., efficient APF, easy computationally

But:

- Adds uncertainty to parameters each time point
- “Loss of parameter information”
- Degrades performance, over-diffuse posteriors, can “drift” radically
- Modern times: many parameters – lots of “information lost”

## Treatment of Parameters: KS

**Kernel Smoothing** - West 93 (90 tech. report):

- Smooth/interpolate parameter samples → continuous distribution
- Use in *adaptive importance sampling*

**Time  $t$  sample/weights:**  $\{\boldsymbol{\theta}_t^{(j)}, \omega_t^{(j)}\}$

**MC mean/variance matrix:**  $\mathbf{m}_t, \mathbf{V}_t$

**Kernel density:**

$$p(\boldsymbol{\theta}|D_t) \approx \sum_{j=1}^N \omega_t^{(j)} N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2 \mathbf{V}_t)$$

– smoothing parameter  $h$  –

- Regenerate samples
- Evaluate density (in importance sampling)
- Choice of locations  $\mathbf{m}_t$ ? and smoothing factor  $h$ ?

## Treatment of Parameters: KS (cont.)

Kernel with Shrinkage (West 93a,b):

- Standard kernel locations  $\boldsymbol{m}_t^{(j)} = \boldsymbol{\theta}_t^{(j)}$   
results in oversmoothing:  $V(\boldsymbol{\theta}|D_t) > \mathbf{V}_t$
- Corrected using **kernel location shrinkage**:

$$\boldsymbol{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1 - a)\boldsymbol{m}_t$$

where  $a^2 = 1 - h^2$  (is near 1)

- Kernel/mixture has correct MC mean and variance matrix:

$$E(\boldsymbol{\theta}|D_t) = \boldsymbol{m}_t$$

and

$$V(\boldsymbol{\theta}|D_t) = \mathbf{V}_t$$

## Modifying the AE Method

- AE:  $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\nu}_{t+1}$
- where  $\boldsymbol{\nu}_{t+1} \sim N(\mathbf{0}, \mathbf{W}_{t+1})$
- is now **correlated**  $\boldsymbol{\theta}_t$  via

$$C(\boldsymbol{\nu}_{t+1}, \boldsymbol{\theta}_t | D_t) = -\mathbf{W}_{t+1}/2$$

**Result:**

$$E(\boldsymbol{\theta}_{t+1} | D_t) = E(\boldsymbol{\theta}_t | D_t) = \mathbf{m}_t$$

and

$$V(\boldsymbol{\theta}_{t+1} | D_t) = V(\boldsymbol{\theta}_t | D_t) = \mathbf{V}_t$$

– no “information loss” in AE –

**Choice/specification of  $\mathbf{W}_{t+1}$ :**

- Standard discount factor method (West and Harrison 97)
- $\mathbf{W}_{t+1} = \epsilon \mathbf{V}_t$  for small  $\epsilon$  (e.g., 0.01)
- Other, more general versions

## Modified AE = KS

Modified AE implies

$$p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t) \sim N(\boldsymbol{\theta}_{t+1} | a\boldsymbol{\theta}_t + (1 - a)\boldsymbol{m}_t, h^2 \mathbf{V}_t)$$

with

- $a = 1 - \epsilon/2$
- $h^2 = 1 - a^2$

Same as KS with shrinkage:

Implied prior at  $t + 1$  is

$$p(\cdot | D_t) \approx \sum_{j=1}^N \omega_t^{(j)} N(\cdot | \boldsymbol{m}_t^{(j)}, h^2 \mathbf{V}_t)$$

with  $\boldsymbol{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1 - a)\boldsymbol{m}_t$

And easy: single discount factor determines  $h$  and  $a$

## Combined Filtering

### Parameter Extended APF (APF+)

Time  $t$  MC summary:

$$\{\boldsymbol{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)} ; \omega_t^{(j)} : j = 1, \dots, N\}$$

APF+:

- Given  $D_t$ , for every sample  $k$  find estimates:

$$\begin{cases} \boldsymbol{\mu}_{t+1}^{(k)} = E(\boldsymbol{x}_{t+1} | \boldsymbol{x}_t^{(k)}, \boldsymbol{\theta}_t^{(k)}) & \text{of } \boldsymbol{x}_{t+1} \\ \boldsymbol{m}_t^{(k)} & \text{of } \boldsymbol{\theta} \end{cases}$$

- and weights  $g_{t+1}^{(k)} \propto \omega_t^{(k)} p(\boldsymbol{y}_{t+1} | \boldsymbol{\mu}_{t+1}^{(k)}, \boldsymbol{m}_t^{(k)})$
- Sample (aux) indicators  $j$  with probs  $g_{t+1}^{(j)}$
- and then **time  $t + 1$  samples**:
  - $\boldsymbol{\theta}_{t+1}^{(j)} \sim N(\cdot | \boldsymbol{m}_t^{(j)}, h^2 \mathbf{V}_t)$
  - $\boldsymbol{x}_{t+1}^{(j)} \sim p(\cdot | \boldsymbol{x}_t^{(j)}, \boldsymbol{\theta}_{t+1}^{(j)})$
- and weights**:

$$\omega_{t+1}^{(j)} = \frac{p(\boldsymbol{y}_{t+1} | \boldsymbol{x}_{t+1}^{(j)}, \boldsymbol{\theta}_{t+1}^{(j)})}{p(\boldsymbol{y}_{t+1} | \boldsymbol{\mu}_{t+1}^{(j)}, \boldsymbol{m}_t^{(j)})}$$

## Simple AR(1) Example

– no state variables, one parameter –

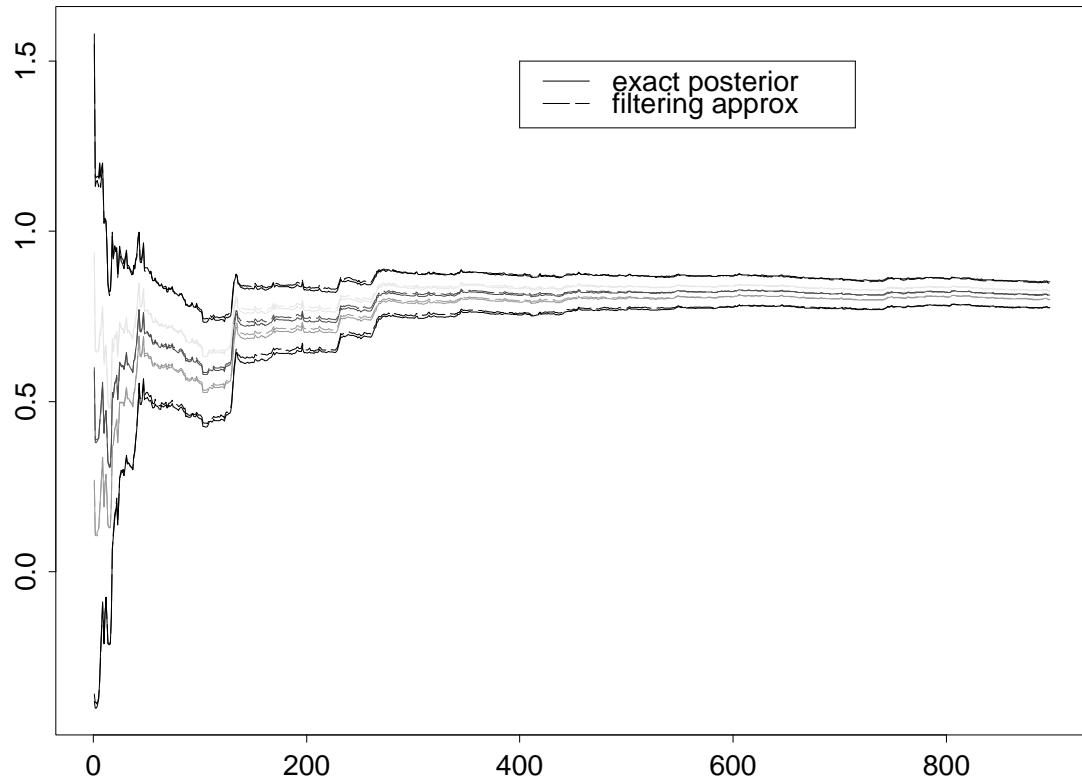


Figure 1: Posterior quantiles for AR(1) parameter  $\phi$

	2.5%	25%	50%	75%	97.5%
exact:	0.776	0.801	0.814	0.827	0.852
approx:	0.776	0.800	0.812	0.824	0.849

Table 1: Quantiles at  $t = 897$

## Dynamic Factor Models

– serious test case: 30 parameters, 3 state variables –

### Factor model:

- $\mathbf{y}_t$  is  $q$ -vector of exchange rate returns (daily)
- $\mathbf{f}_t$  is  $k$ -vector of *latent factors*
- $\mathbf{X}$  is  $q \times k$  matrix of *factor loading coefficients*

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{X}\mathbf{f}_t + \boldsymbol{\epsilon}_t$$

- $\boldsymbol{\epsilon}_t \sim \text{indep } N(\mathbf{0}, \boldsymbol{\Psi})$  with  $\boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_q)$
- Latent factors  $\mathbf{f}_t$  zero-mean normal with  
with  $V(\mathbf{f}_t) = \text{diag}(\exp(x_{t1}), \dots, \exp(x_{tk}))$

### State evolution model:

- State vector  $\mathbf{x}_t = (x_{t1}, \dots, x_{tk})'$
- Vector AR(1) model

$$\mathbf{x}_t = \boldsymbol{\mu} + \Phi(\mathbf{x}_{t-1} - \boldsymbol{\mu}) + \text{indep } N(\mathbf{0}, \mathbf{U})$$

with diagonal  $\Phi$

## Dynamic Factor Models

Observation model:

$$p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\theta}) = N(\mathbf{y}_t | \mathbf{0}, \mathbf{X} \exp(\mathbf{x}_t) \mathbf{X}' + \boldsymbol{\Psi})$$

where

$$\exp(\mathbf{x}_t) = \text{diag}(\exp(x_{t1}), \dots, \exp(x_{tk}))$$

State evolution model:

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \boldsymbol{\theta}) = N(\mathbf{x}_t | \boldsymbol{\mu} + \boldsymbol{\Phi}(\mathbf{x}_{t-1} - \boldsymbol{\mu}), \mathbf{U})$$

Parameters:

$$\boldsymbol{\theta} = \{\mathbf{X}, \boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Phi}, \mathbf{U}\}$$

$q = 6, k = 3 : \rightarrow 3$  states, 36 parameters

## Dynamic Factor Model Analysis

- *Shephard and Pitt 98*: Model variations, MCMC analyses, APF on states
- *Aguilar and West 98*: MCMC, forecasting & portfolio construction uses APF on states

### Analysis:

- Daily exchange rates in \\$s of  $q = 6$  currencies
- MCMC over 2.5 years to “origin”  $t = 914$
- Combined filtering via APF+ over  $t = 915, \dots$
- Comparisons of APF+ results with full MCMC analysis at  $t = 924, 964$

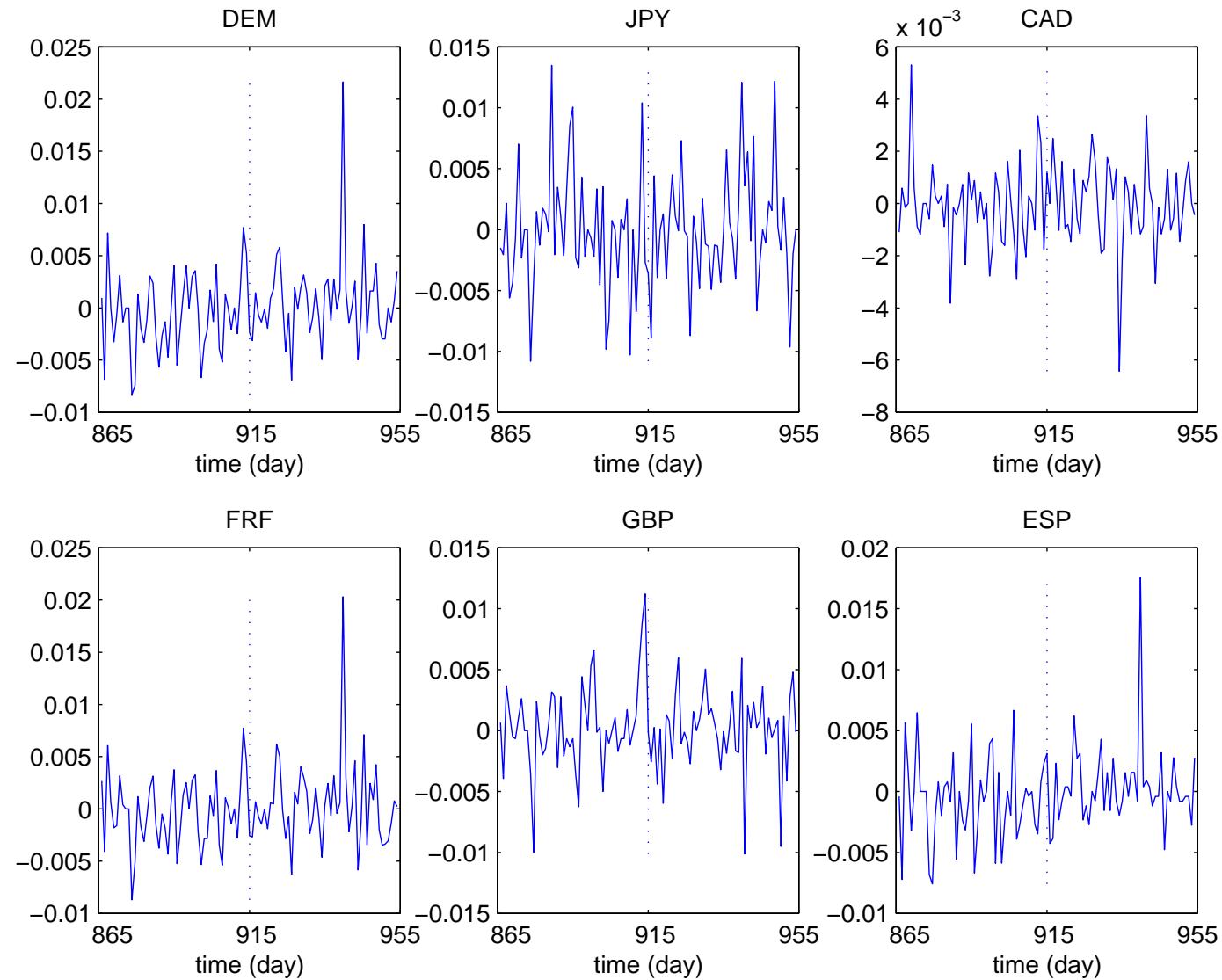
**Wrinkle:** APF+ fixes  $\mathbf{U}$  at  $E(\mathbf{U}|D_{914}) =$

$$0.01 \begin{pmatrix} 1.71(0.26) & 0.27(0.20) & 0.09(0.18) \\ & 1.94(0.30) & 0.13(0.20) \\ & & 1.74(0.27) \end{pmatrix}$$

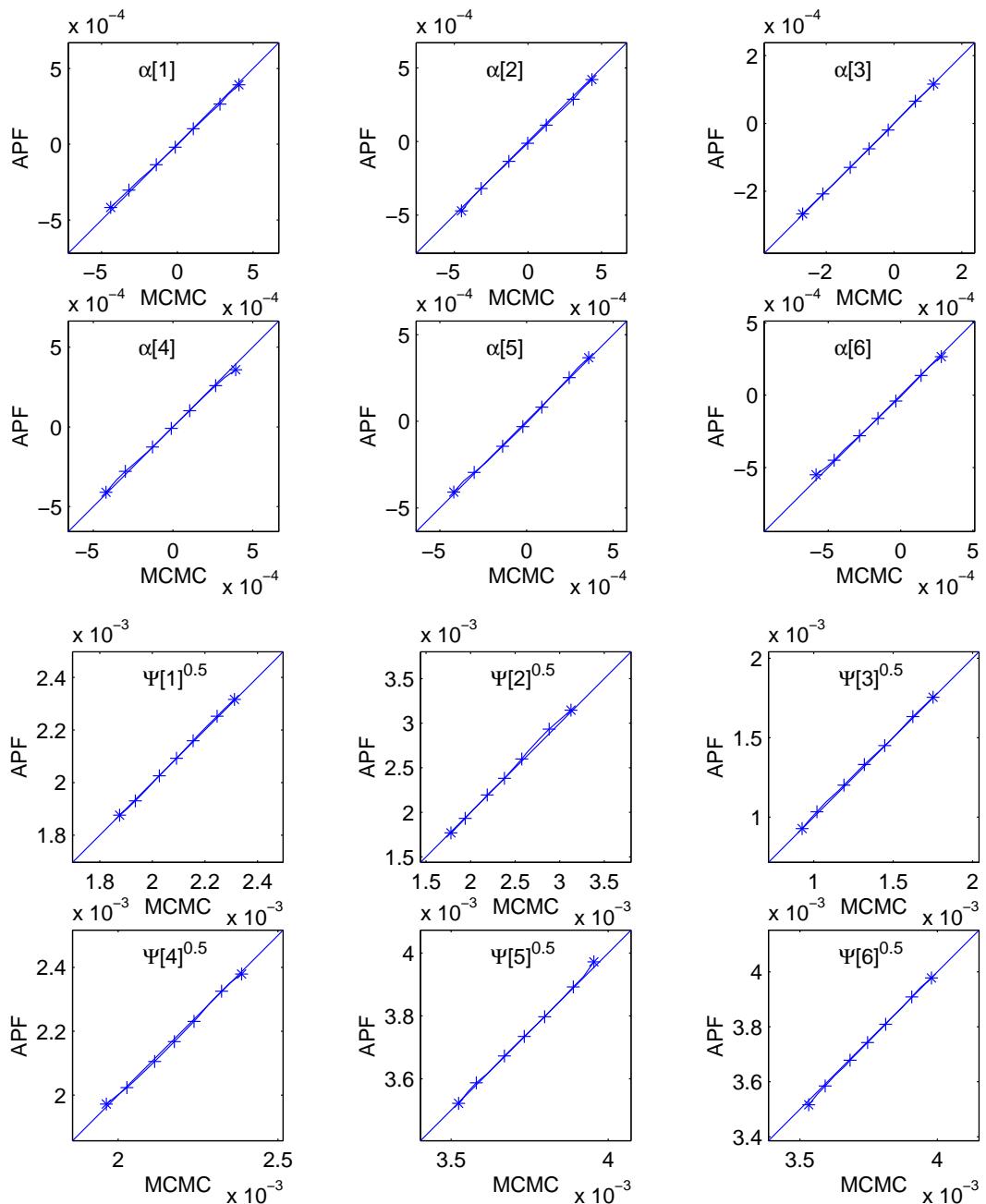
Why? Normal kernels!

36 → 30 parameters ...

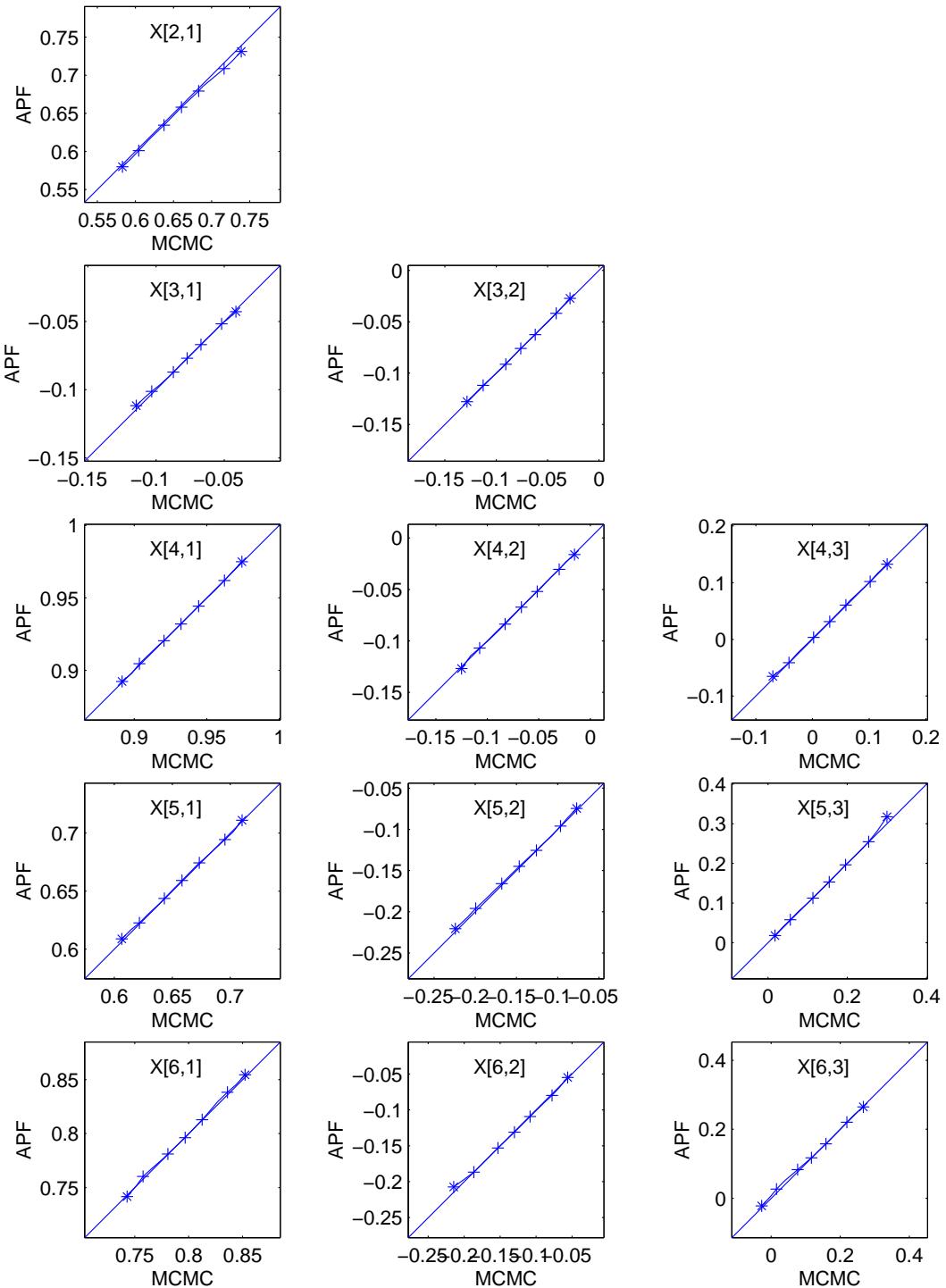
# Exchange Rate Time Series



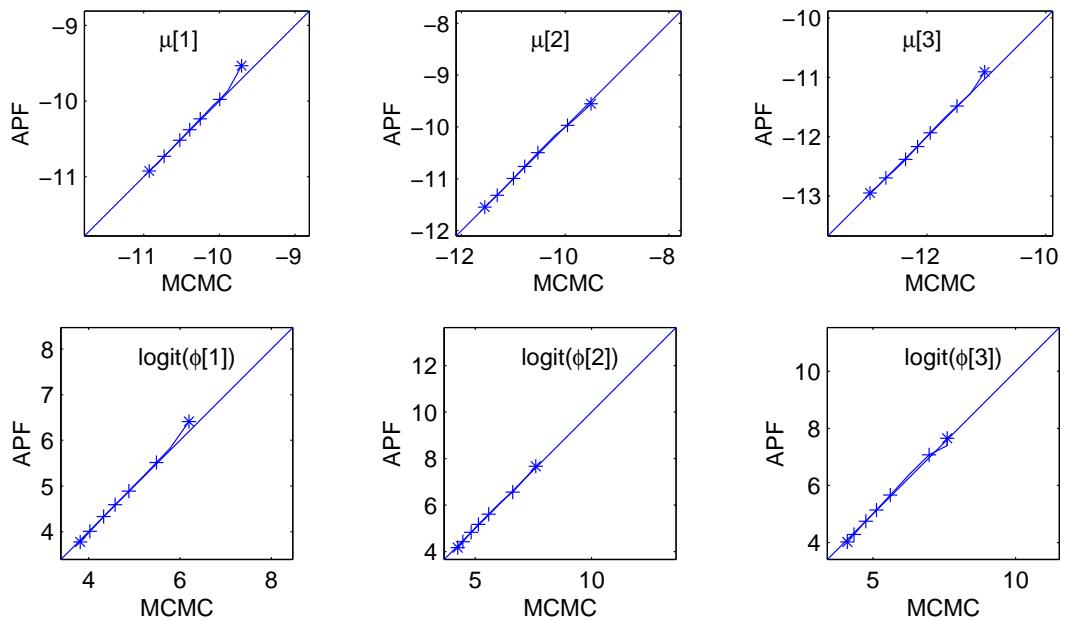
## **DFM results at 10 steps ( $\alpha, \Psi$ )**



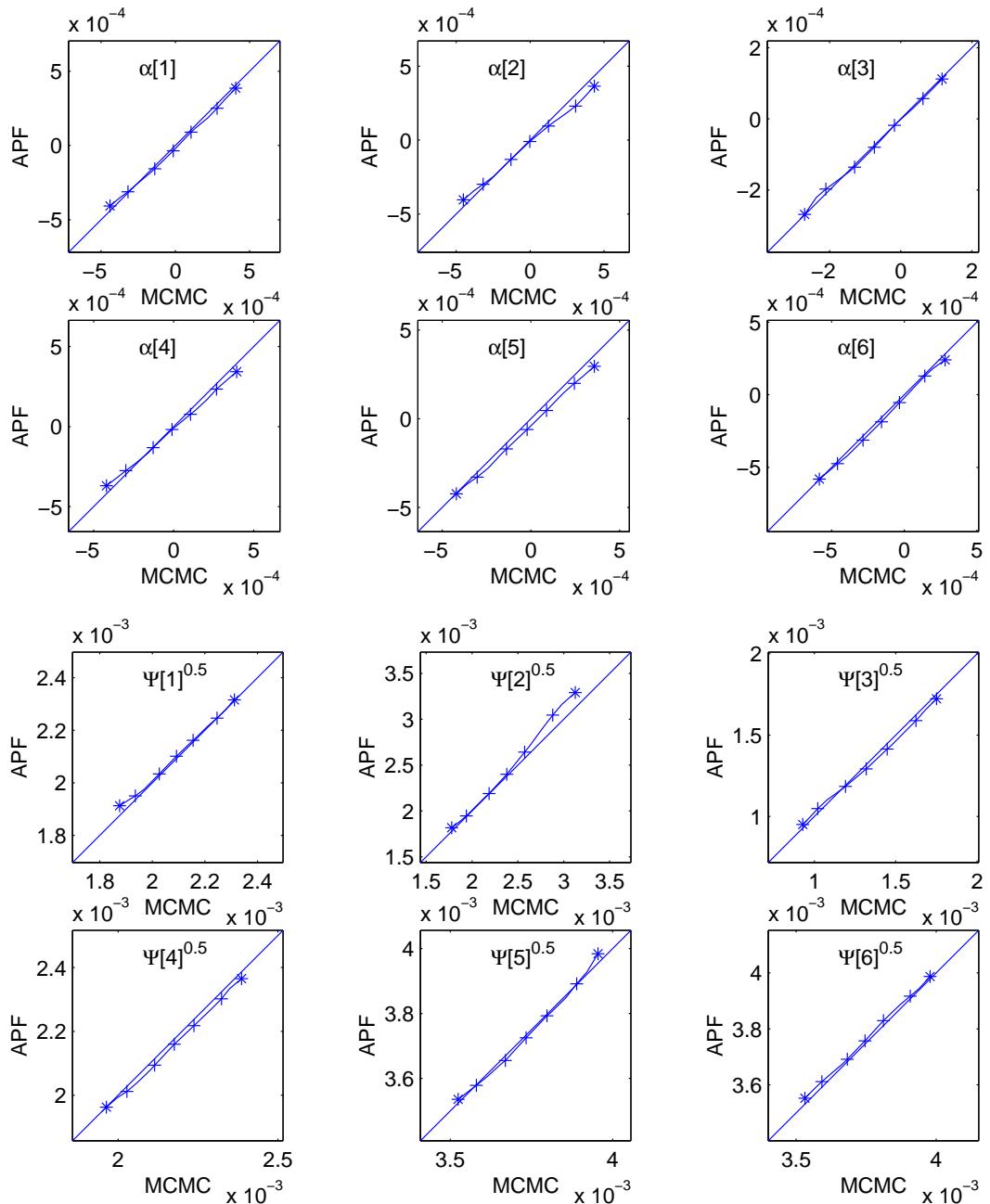
## **DFM results at 10 steps ( $X$ )**



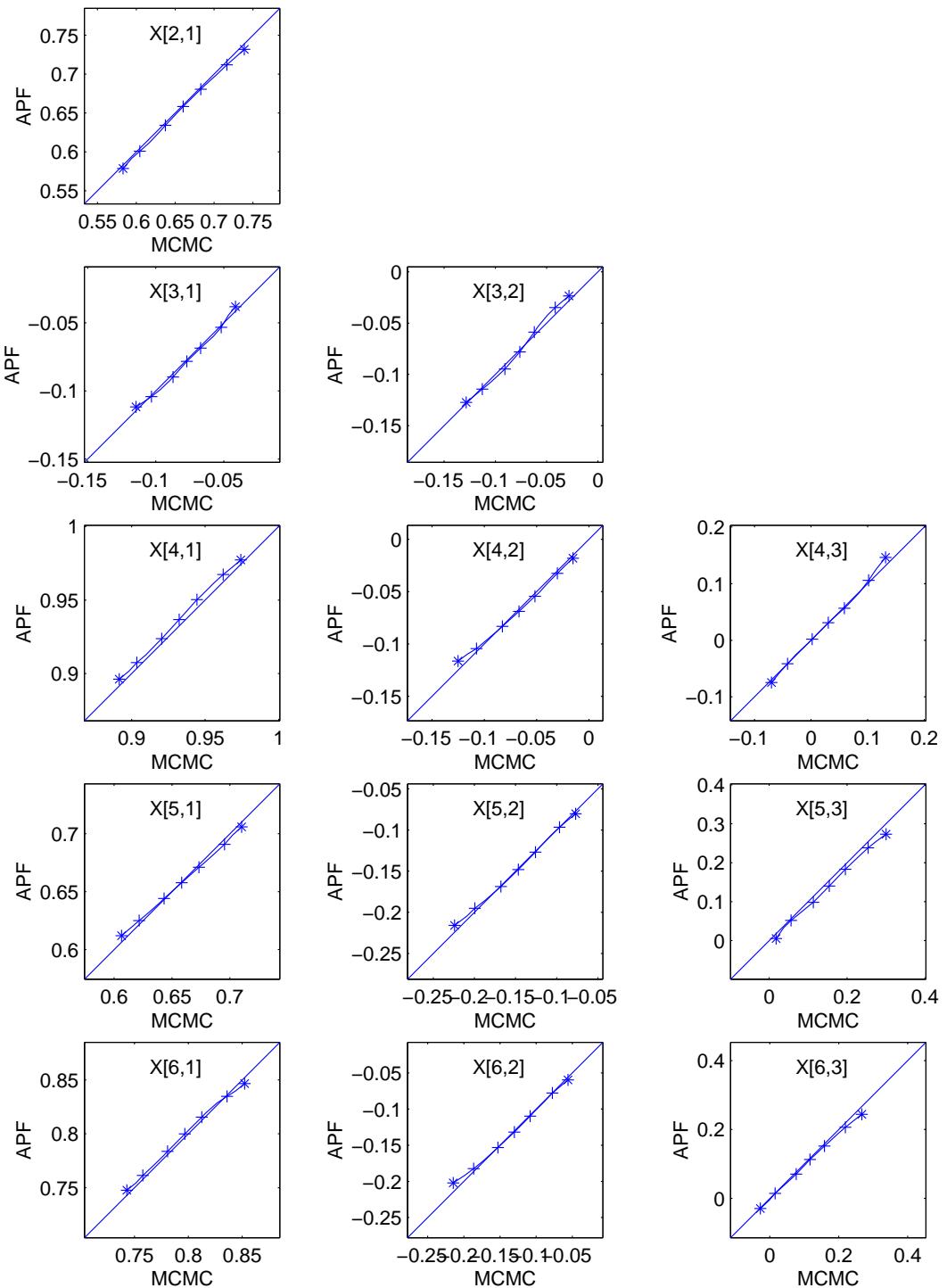
## **DFM results at 10 steps ( $\mu, \Phi$ )**



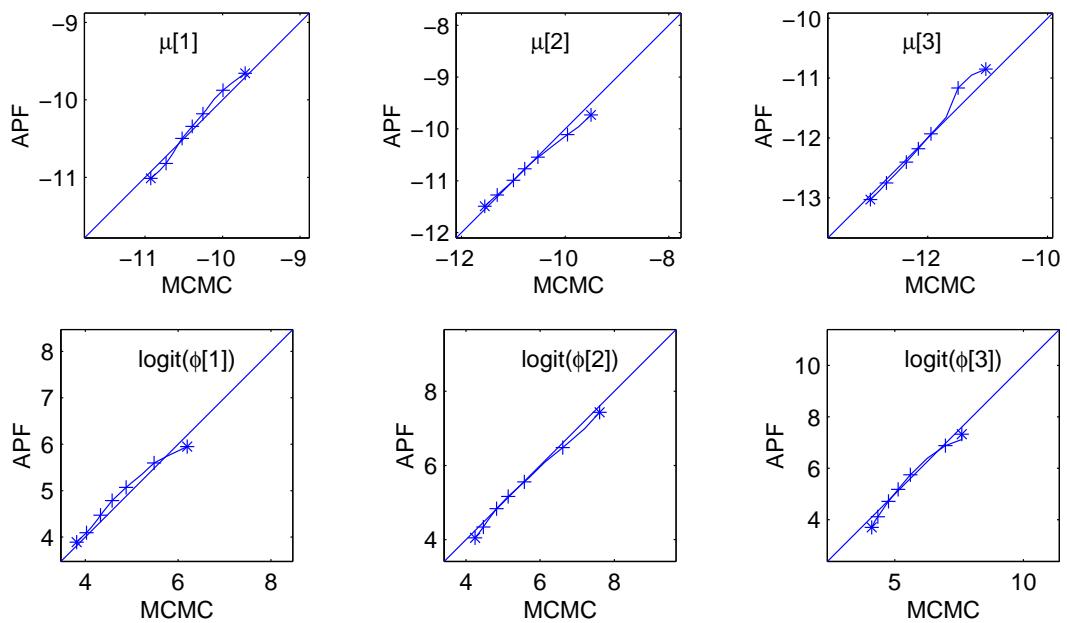
## ***DFM results at 50 steps ( $\alpha, \Psi$ )***



## **DFM results at 50 steps ( $X$ )**



## **DFM results at 50 steps ( $\mu, \Phi$ )**



## **Comments, issues, future**

### **DFM Analysis:**

- Small differences between MCMC and APF+ at larger time steps
- APF+ analysis ignores uncertainty in  $\mathbf{U}$ 
  - Correlated most with  $\boldsymbol{\mu}, \Phi$
  - Biggest differences on these parameters
- Build-up of errors in any sequential analysis
- Reality check: Context and usage
  - Fast sequential analysis, parallel models
  - Daily updates – MCMC “at weekends”
- Pessimism on “black boxes”

## **Comments, issues, future**

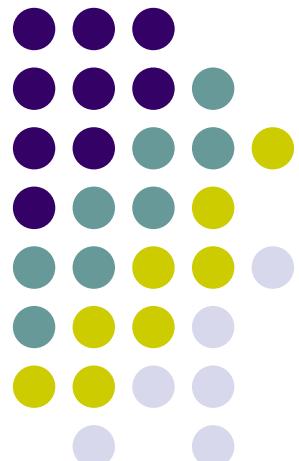
### **Modifications/extensions: APF++?**

- “Interesting” posteriors - varying tail-weight, skewness, multiple modes, ...
- Common kernel shape/scale inefficient
- Multiple modes: avoid shrinkage to global mean
- Variable/Local kernels?  
(West 93; Givens & Raftery 96)
- Discussion, ideas (“theory”) in Liu and West 99

**References:** see Liu and West 99

# MCMC and Particle Filtering

- Single-move MCMC;
- Block-move MCMC;
- Bootstrap filter;
- Auxiliary Particle Filter;
- APS + parameter estimation



## Stochastic Volatility Models



# Stochastic volatility models

$$y_t \sim N(0, \exp(\lambda_t))$$

$$\lambda_t = \alpha + \phi \lambda_{t-1} + \omega_t \quad \omega_t \sim N(0, \sigma^2)$$

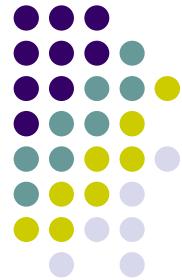
## Priors

$$\lambda_1 \sim N\left(\frac{\alpha}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right)$$

$$\alpha \sim N(a_\alpha, b_\alpha)$$

$$\phi \sim TN(a_\phi, b_\phi)$$

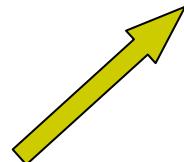
$$\sigma^2 \sim IG(a_\sigma, b_\sigma)$$



# Single Move MCMC (Jacquier et al. 1994)

- Sampling one state at the time:

$$\begin{aligned} p(\lambda_t | \lambda_{(-t)}, \Theta) &= p(\lambda_t | \lambda_{t-1}, \lambda_{t+1}, \Theta) \\ &\propto p(y_t | \lambda_t) p(\lambda_t | \lambda_{t-1}) p(\lambda_{t+1} | \lambda_t) \end{aligned}$$



Density does not have standard form...

... accept/reject step (or possibly MH)

Any other complication?



# FFBS (Kim, Shephard, Chib 98)

- Problem: How to filter forward?
  - Solution: Approximation through mixtures

$$\log(y_t^2) = \lambda_t + \log(\nu_t^2)$$

$$\lambda_t = \alpha + \phi \lambda_{t-1} + \omega_t$$

$$\log(\nu_t^2) \sim \log(\chi^2) \approx \sum_{i=1}^7 q_i N(a_i, b_i)$$



## FFBS (Kim, Shephard, Chib 98)

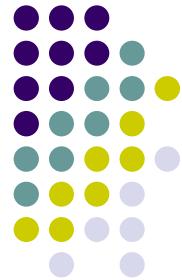
- Sample the indicator variable

$$P(k_t = j | \log(y_t^2), \lambda_t, \Theta) \propto q_j N(\log(y_t^2) | a_j + \lambda_t, b_j)$$

- Forward-Filtering Backward Sampling (as usual).

$$p(\lambda_1, \dots, \lambda_T | D_T, \Theta) = p(\lambda_T | D_T, \Theta) \prod_{t=1}^{T-1} p(\lambda_t | \lambda_{t+1}, D_t, \Theta)$$

- Details in notes from STA214



# Particle Filtering

- Observational model

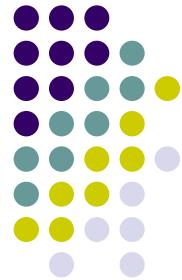
$$p(y_t | x_t, \Theta)$$

- Markov evolution model

$$p(x_t | x_{t-1}, \Theta)$$

- Goal: sequentially update posteriors

$$\dots \rightarrow p(x_t, \Theta | D_t) \rightarrow p(x_{t+1}, \Theta | D_{t+1}) \rightarrow \dots$$



# Particle Filtering

- Example

$$y_t = \frac{x_t^2}{20} + \nu_t$$

$$x_t = \frac{1}{2}x_{t-1} + 25\frac{x_{t-1}}{(1 + x_{t-1}^2)} + 8\cos(1.2t) + \omega_t$$



# Particle Filtering

- Prior

$$p(x_t | D_{t-1}, \Theta) = \int p(x_{t-1} | D_{t-1}, \Theta) p(x_t | x_{t-1}, \Theta) dx_{t-1}$$

- Prediction

$$p(y_t | D_{t-1}, \Theta) = \int p(y_t | x_t, D_{t-1}, \Theta) p(x_t | D_{t-1}, \Theta) dx_t$$

- Update

$$p(x_t | D_t, \Theta) \propto p(x_t | D_{t-1}, \Theta) p(y_t | x_t, D_{t-1}, \Theta)$$



# Particle Filtering

- Possible solutions:
  - Extended Kalman-filters
  - Grid-based methods for integration
  - Piecewise linear approximations
  - Sequential importance sampling (**particle filters**)



# Particle Filtering

- Numerical approximations based on “**particles**” and corresponding weights

$$\{x_t^{(j)} : j = 1, \dots, N\} \quad \{w_t^{(j)} : j = 1, \dots, N\}$$

- Prior and posterior can be approximated by the following mixtures:

$$\hat{p}(x_{t+1}|D_t, \Theta) = \sum_{j=1}^N p(x_{t+1}|x_t^{(j)}, \Theta) w_t(j)$$

$$\hat{p}(x_{t+1}|D_{t+1}, \Theta) \propto p(y_{t+1}|x_{t+1}, D_t) \sum_{j=1}^N p(x_{t+1}|x_t^{(j)}, \Theta) w_t^{(j)}$$



# Bayesian Bootstrap Filter (Gordon et al. 93)

- At time t, suppose we have a set of random samples

$$\{x_t(j) : j = 1, \dots, N\} \sim p(x_t | D_t, \Theta)$$

- We can **evolve** the particles through the system to obtain samples from the prior

$$\{x_{t+1}^*(j) : j = 1, \dots, N\} \sim p(x_{t+1} | D_t, \Theta)$$



# Bayesian Bootstrap Filter

- Using the prior as a importance density, the set of samples...

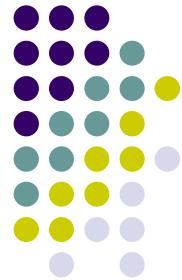
$$\{x_{t+1}^*(j) : j = 1, \dots, N\}$$

- ...with corresponding weights...

$$q_j \propto p(y_{t+1} | x_{t+1}^*(j), D_t, \Theta)$$

- ...form a weighted sample from the posterior

$$p(x_{t+1} | D_{t+1}, \Theta)$$



# Bayesian Bootstrap Filter

- Why? Sampling Importance Re-sampling (SIR)...

$$q_j = \frac{p(x_{t+1}^*(j) | D_{t+1}, \Theta)}{p(x_{t+1}^*(j) | D_t, \Theta)}$$

$$\propto \frac{p(y_{t+1} | x_{t+1}^*(j), D_t, \Theta) p(x_{t+1}^* | D_t, \Theta)}{p(x_{t+1}^* | D_t, \Theta)}$$

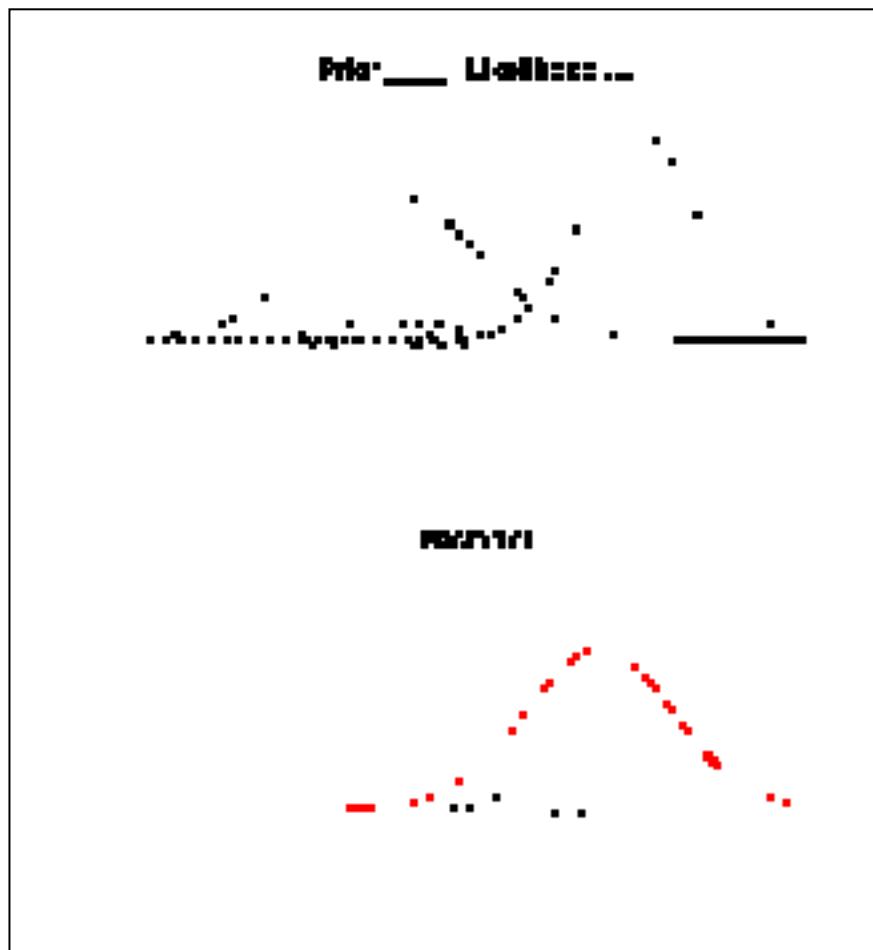
$$= p(y_{t+1} | x_{t+1}^*(j), D_t, \Theta)$$

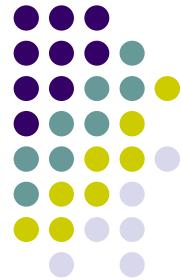
  
**Key cancellation**



# Bayesian Bootstrap Filter

- Problem: degeneration of the filter





# Auxiliary Particle Filter (Pitt & Shephard 99)

- The idea is to use the mixture approximation to facilitate computations while improving the importance function. The update step will be done by sampling from the following “auxiliary” posterior

$$p(x_{t+1}, k | D_{t+1}) \propto p(y_{t+1} | x_{t+1}, D_t) p(x_{t+1} | x_t^{(k)})$$

$$k = 1, \dots, N$$

- Drawing from the above joint density and discarding the index  $k$ , produce a sample from the approximate posterior density. Again, SIR is used.



# Auxiliary Particle Filter

- At time  $t$ , suppose we have a set of random samples and weights

$$\{x_t^{(k)}, w_t^{(k)} : k = 1, \dots, N\}$$

- For each  $k$ , set the “estimates” and weights

$$\mu_{t+1}^{(k)} = E(x_{t+1} | x_t^{(k)})$$

$$g_{t+1}^{(k)} \propto w_t^{(k)} p(y_{t+1} | \mu_{t+1}^{(k)})$$

- Sample the auxiliary variable  $j$  with probability given by  $g_{t+1}^{(j)}$  followed by

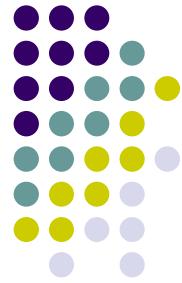
$$x_{t+1}^{(j)} \sim p(x_{t+1} | x_t^{(j)})$$



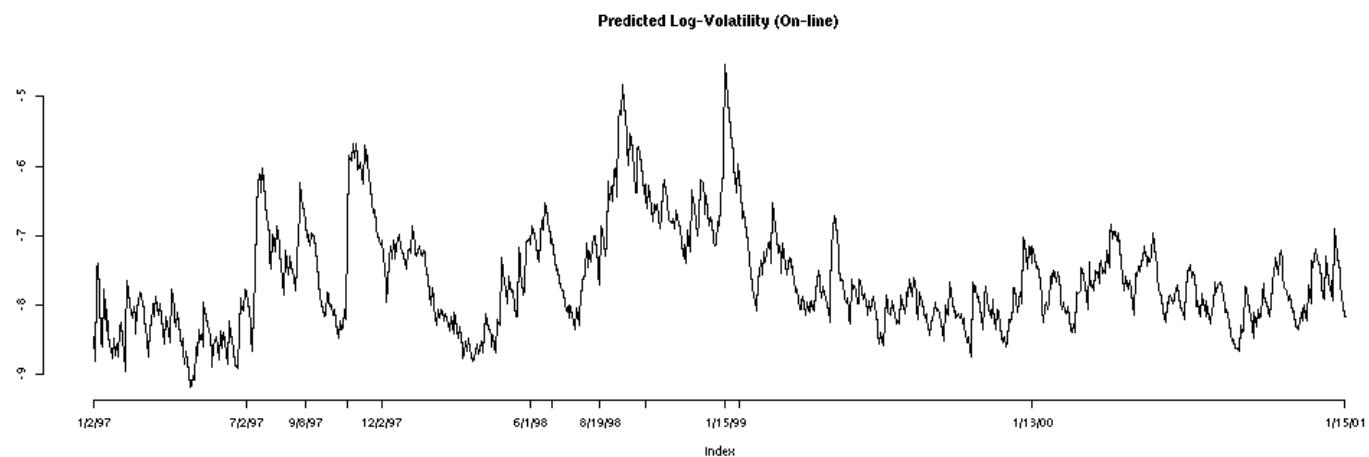
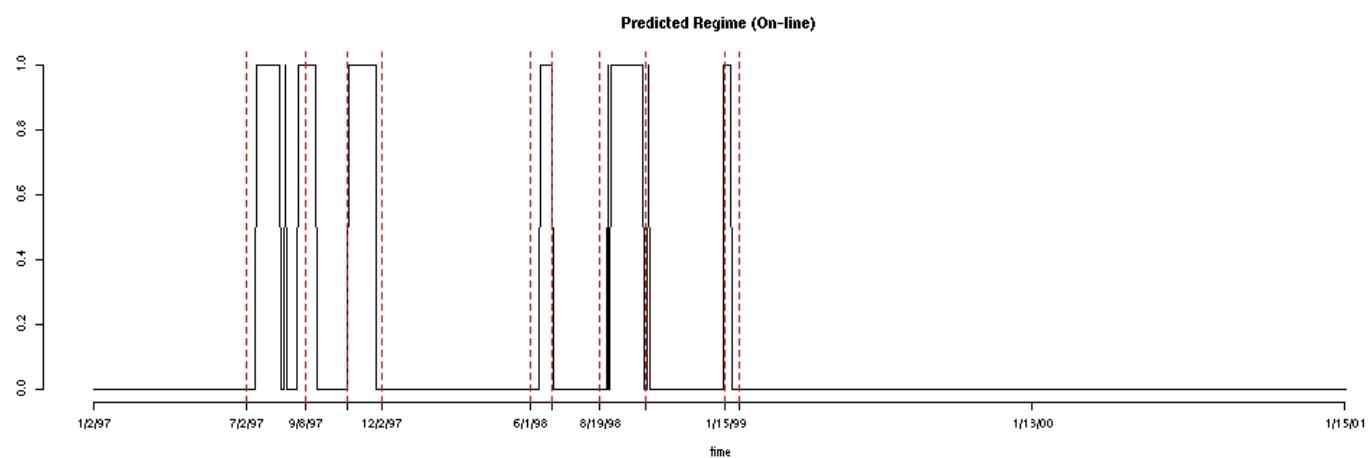
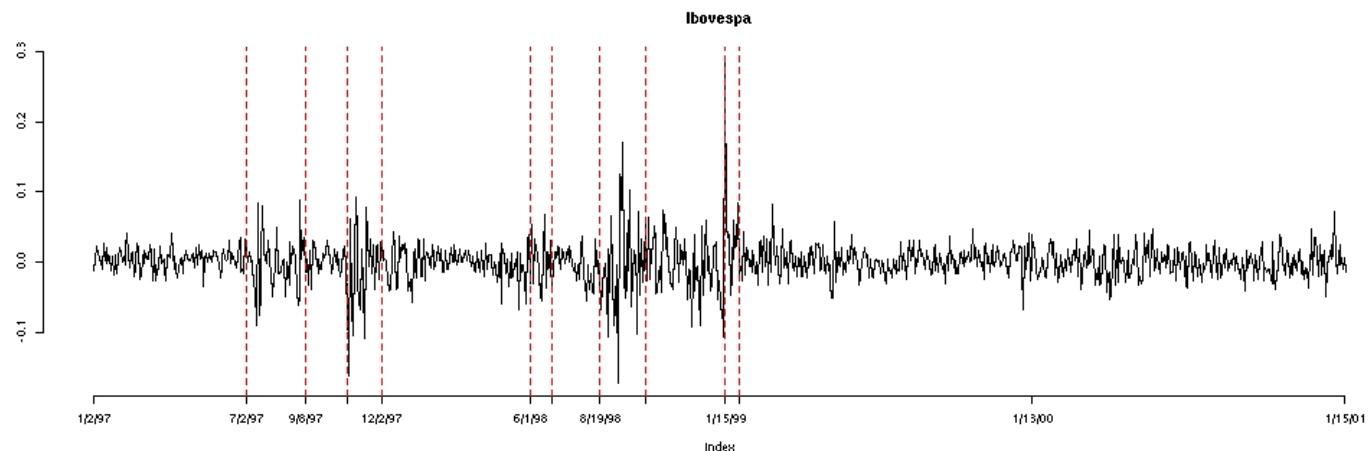
# Auxiliary Particle Filter

- Compute the new weights

$$w_{t+1}^{(j)} \propto \frac{p(y_{t+1} | x_{t+1}^{(j)})}{p(y_{t+1} | \mu_{t+1}^{(j)})}$$

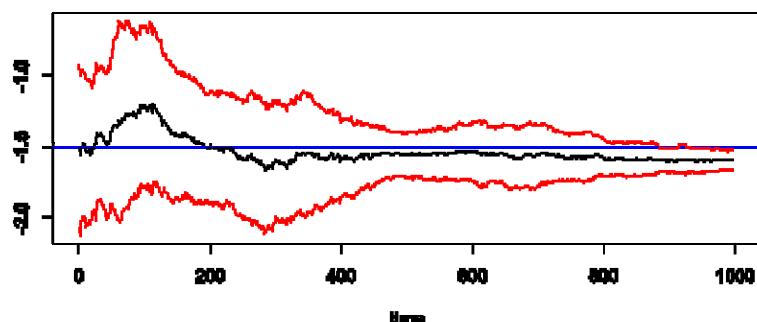


# Back to SVM

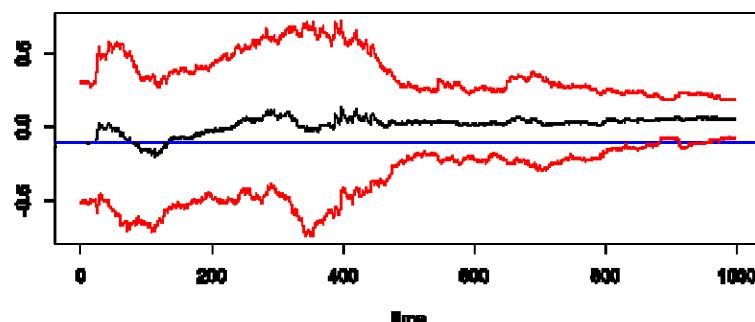




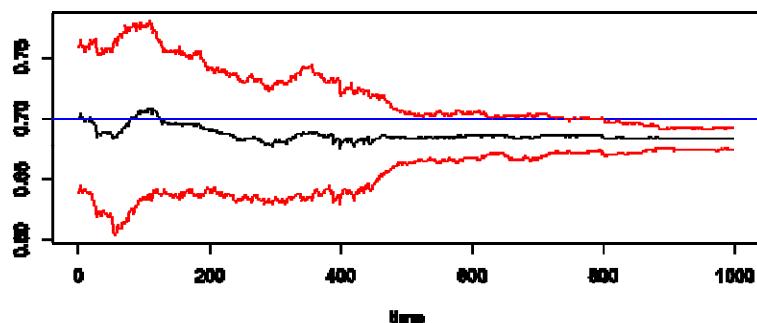
alpha1



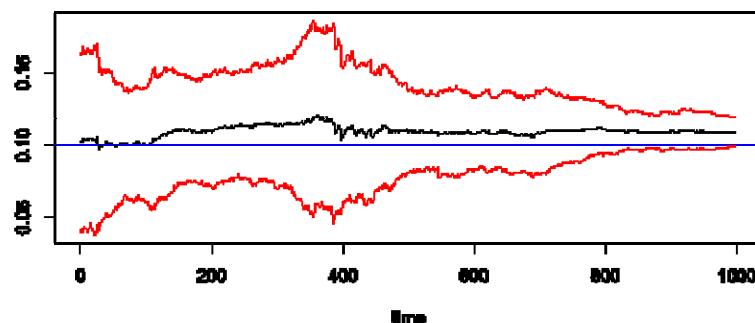
gamma1



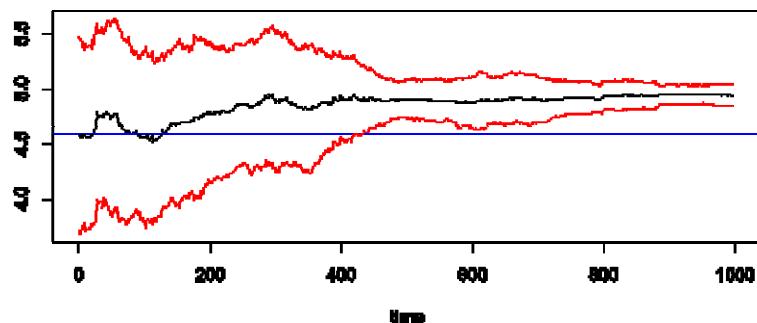
phi



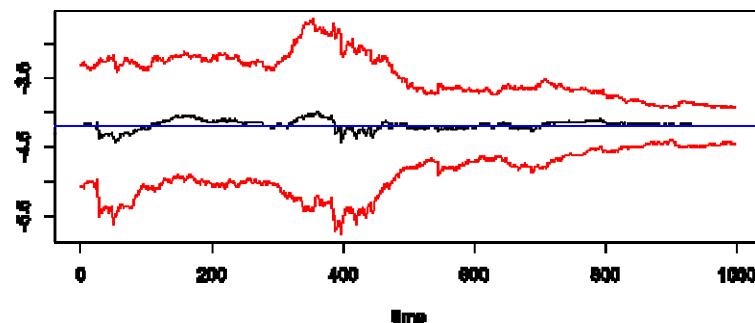
sigma2



logit(p11)



logit(p12)



# **Bayesian Dynamic Factor Models with Shrinkage in Asset Allocation**

Omar Aguilar

Merrill Lynch Quantitative Research.

Jose Quintana

CDC Investment Management Corporation.

Mike West

Institute of Statistics & Decision Sciences

Duke University

<http://www.isds.duke.edu/>

## **Outline**

- International Exchange Rates.
- Dynamic Bayesian **Partial Shrinkage** Models for the Expected Returns.
- Bayesian **Dynamic Factor Models**.
- Multivariate **Stochastic Volatility** Components.
- Dynamic Asset Allocation.
- Extensions and Future Directions.

## International Exchange Rates

### DATA:

- Monthly spot exchange rates ( $s_{it}$ ), with respect to the US Dollar.

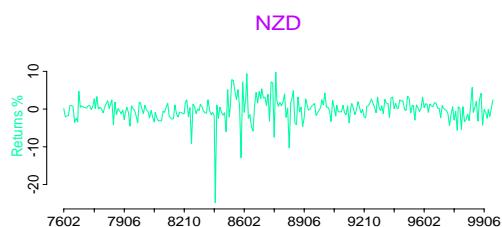
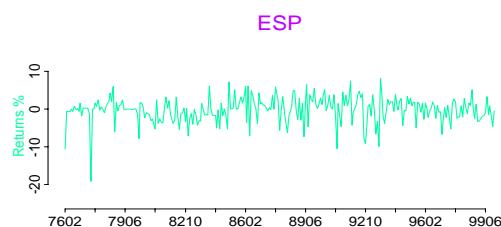
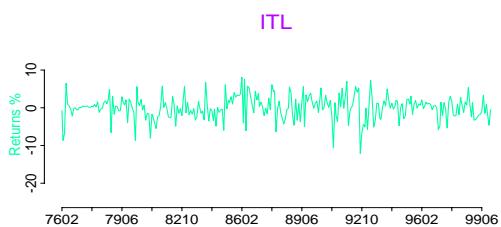
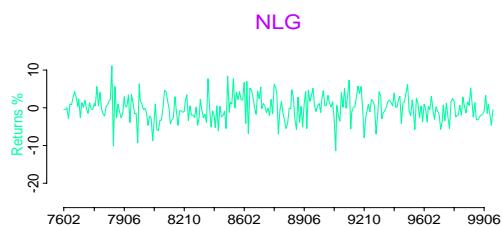
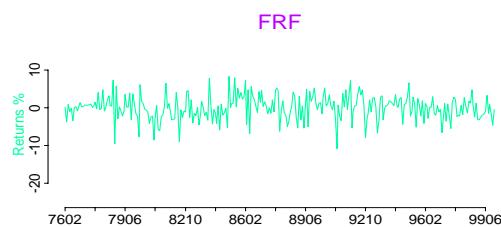
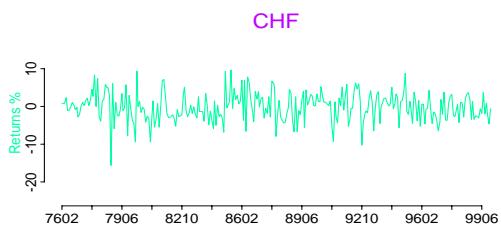
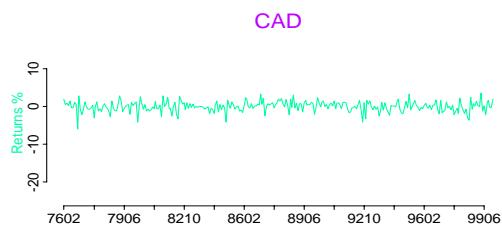
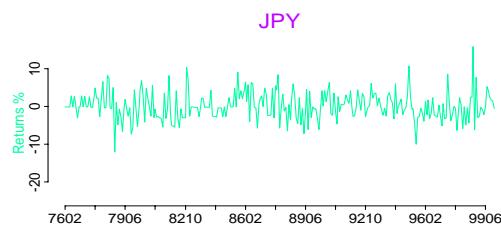
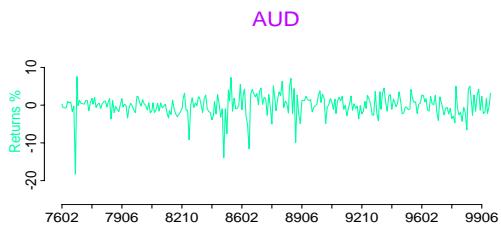
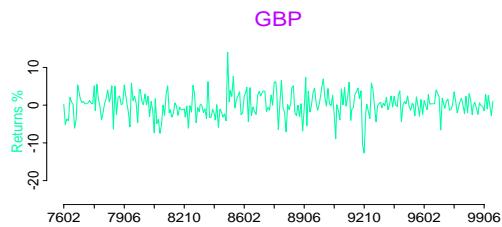
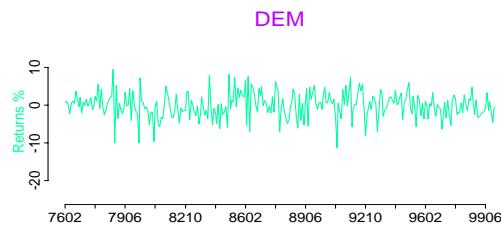
Country	Currency	Code	Country	Currency	Code
Germany	Mark	DEM	France	Franc	FRF
Great Britain	Pound	GBP	Netherlands	Guilder	NLG
Australia	Dollar	AUD	Italy	Lira	ITL
Japan	Yen	JPY	Spain	Peseta	ESP
Canada	Dollar	CAD	New Zealand	Dollar	NZD
Switzerland	Franc	CHF	United States	Dollar	USD

- One-month-ahead excess returns, from 02/76 to 12/99.

$$y_{jt} = \log(s_{jt}/s_{j,t-1}) - [\log(1 + i_{USA,t-1}) - \log(1 + i_{j,t-1})].$$

- Conditional independent and Normally distributed excess returns,  
 $\mathbf{y}_t \sim N(\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t)$ .

# International Exchange Rates Returns

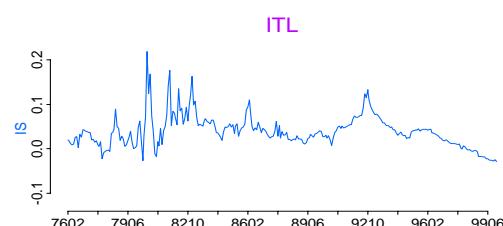
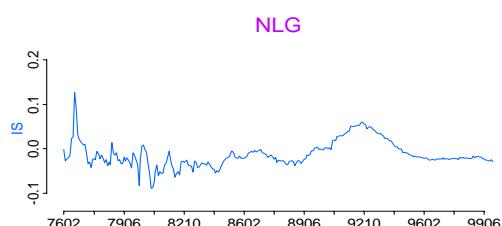
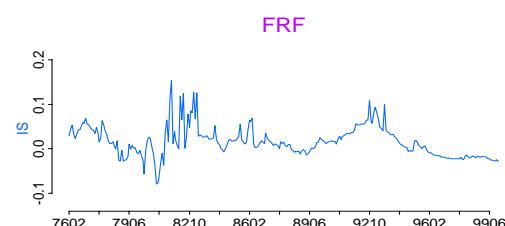
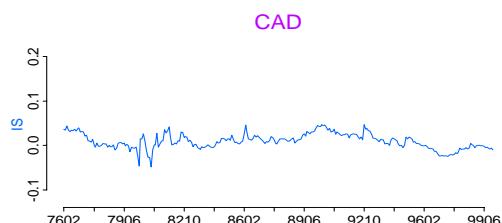
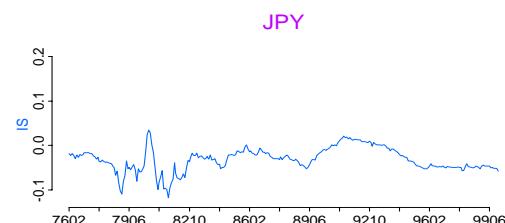
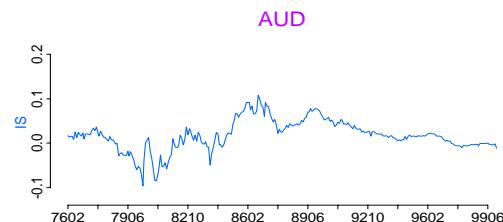
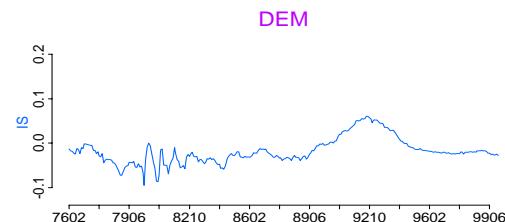


## ***International Exchange Rates***

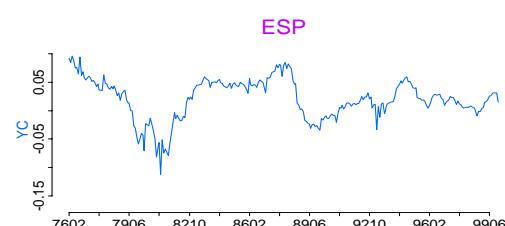
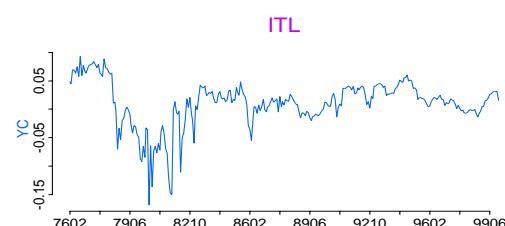
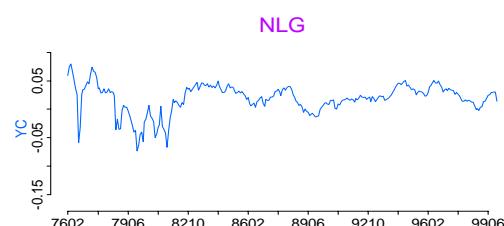
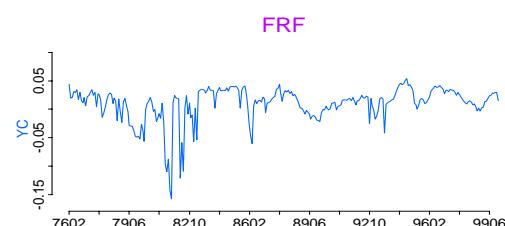
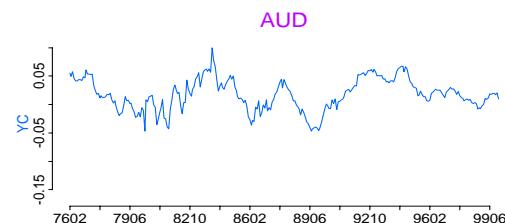
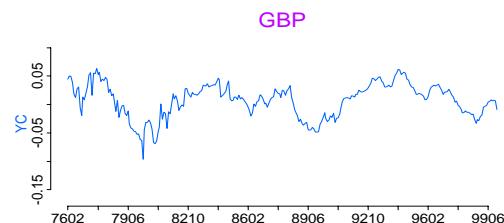
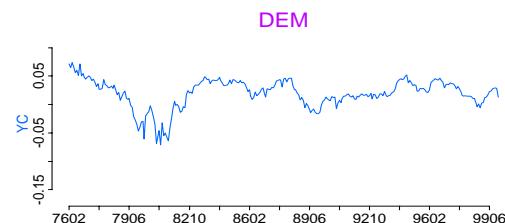
### **GOALS:**

- Dynamic regressions with economic predictors to estimate and predict the expected returns  $\theta_t$ .
- Explore patterns of variability and residual structure over time via modeling  $\Sigma_t$ .
- Find latent processes driving the changes in variances and correlations.
- Improve short-term forecasts of means  $\theta_t$  and variances  $\Sigma_t$ .
- Dynamic asset allocation in portfolio construction via sequential updating.

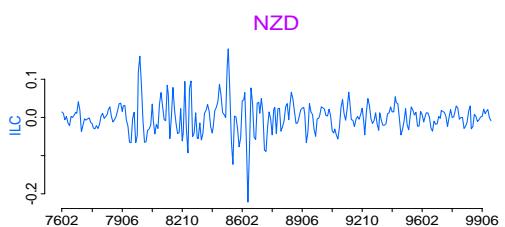
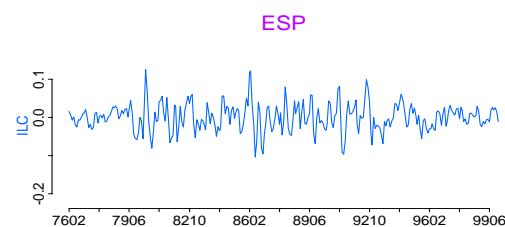
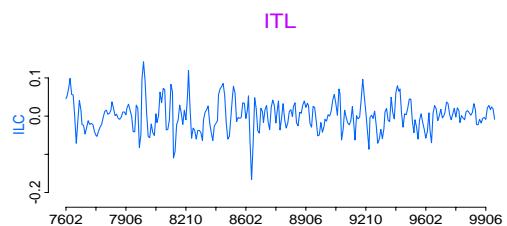
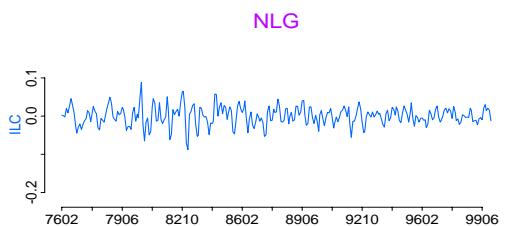
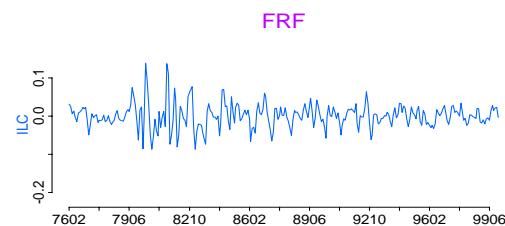
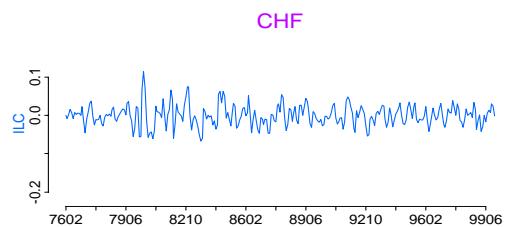
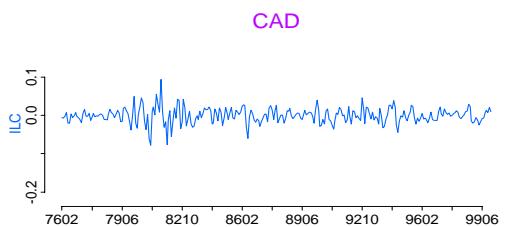
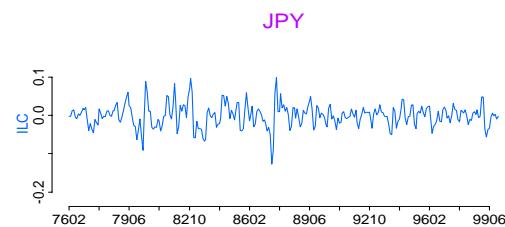
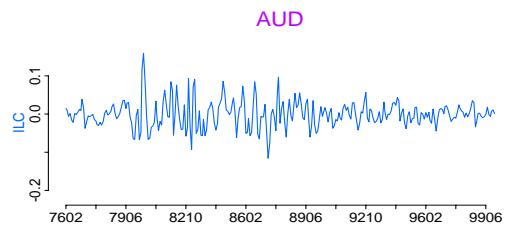
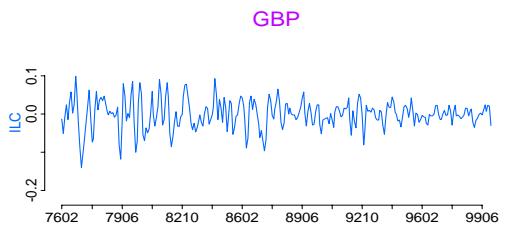
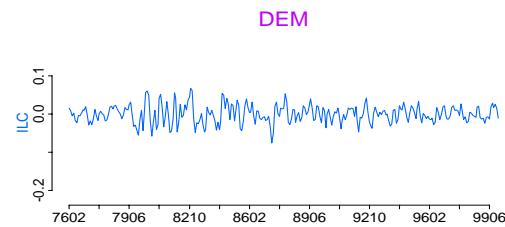
# Short Interest Rates



# **Yield Curve**



# Interest Rate Acceleration



## Bayesian Partial Shrinkage Model for $\theta_t$

### Dynamic SUR Model with Shrinkage:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\theta}_t + \boldsymbol{\nu}_t, & \boldsymbol{\nu}_t &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_t), \\ \boldsymbol{\theta}_t &= \mathbf{Z}_t \boldsymbol{\beta}_t, \\ \boldsymbol{\beta}_t &= \mathbf{L}_t \boldsymbol{\alpha}_t + \mathbf{d}_t, & \mathbf{d}_t &\sim N(\mathbf{0}, \mathbf{U}_t), \\ \boldsymbol{\alpha}_t &= \mathbf{G}_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t &\sim N(\mathbf{0}, \mathbf{W}_t), \\ \boldsymbol{\alpha}_{t-1} &\sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}). \end{aligned}$$

- The linkage matrix  $\mathbf{L}_t$  determines the relationships between the beta coefficients.
- References include Quintana, Chopra and Putnam (1995) and Zellner, Hong and Min (1991).

## Bayesian Partial Shrinkage Model for $\theta_t$

Assume there is only one predictor  $z_{jt}$ , for  $j = 1, \dots, q$ .

$$\begin{aligned} y_{jt} &= \theta_{jt} + \nu_{jt}, & \nu_{jt} &\sim N(0, \sigma_{jt}^2), \\ \theta_{jt} &= z_{jt}\beta_{jt}, \\ \beta_{jt} &= \mu_t + \epsilon_{jt}, \\ \mu_t &= \mu_{t-1} + \xi_t, & \xi_t &\sim N(0, u), \\ \epsilon_{jt} &= \phi_j \epsilon_{j,t-1} + \eta_{jt}, & \eta_t &\sim N(0, \tau_j). \end{aligned}$$

- These relationships imply that

$$\beta_{jt} = \phi_j \beta_{j,t-1} + (1 - \phi_j) \mu_{t-1} + \eta_{jt}.$$

- Shrinkage parameter  $0 \leq \phi_j \leq 1$  where  $\phi_j = \begin{cases} 0 & \text{full shrinkage,} \\ 1 & \text{no shrinkage.} \end{cases}$

## Dynamic Factor Models for $\Sigma_t$

Aguilar and West (2000) dynamic factor model for  $\mathbf{y}_t \sim N(\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t)$ ,

$$\begin{aligned}\mathbf{y}_t &= \boldsymbol{\theta}_t + \mathbf{X}_t \mathbf{f}_t + \boldsymbol{\epsilon}_t, \\ \mathbf{f}_t &\sim N(\mathbf{f}_t | \mathbf{0}, \mathbf{H}_t), \\ \boldsymbol{\epsilon}_t &\sim N(\boldsymbol{\epsilon}_t | \mathbf{0}, \boldsymbol{\Psi}_t).\end{aligned}$$

$$\boldsymbol{\Sigma}_t = \mathbf{X}_t \mathbf{H}_t \mathbf{X}_t' + \boldsymbol{\Psi}_t.$$

- $\mathbf{X}_t$  is the  $q \times k$  factor loadings matrix with  $q \gg k$ ,
- $\mathbf{f}_t$  is a  $k \times 1$  vector of conditionally independent latent common factors,
- $\mathbf{H}_t = \text{diag}(h_{1t}, \dots, h_{kt})$  instantaneous factor variances,
- $\boldsymbol{\epsilon}_t$  is a  $q \times 1$  vector of series-specific quantities,
- $\boldsymbol{\Psi}_t = \text{diag}(\psi_{1t}, \dots, \psi_{qt})$  instantaneous, “idiosyncratic” variances and
- $\boldsymbol{\epsilon}_t$  and  $\mathbf{f}_s$  are mutually independent for all  $t, s$ .

## Stochastic Volatility Components

- Multivariate SV models for the latent factor processes  $\mathbf{f}_t \sim N(\mathbf{0}, \mathbf{H}_t)$ .
  - For  $z_{it} = \log(f_{it}^2) = \lambda_{it} + \nu_{it}$  and  $\lambda_{it} = \log(h_{it})$  we assume **latent VAR(1) models**,

$$\begin{aligned}\mathbf{z}_t &= \boldsymbol{\lambda}_t + \boldsymbol{\nu}_t, \\ \boldsymbol{\lambda}_t &= \boldsymbol{\mu}_t + \boldsymbol{\gamma}_t,\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\eta}_t, \\ \boldsymbol{\gamma}_t &= \boldsymbol{\Phi}\boldsymbol{\gamma}_{t-1} + \boldsymbol{\omega}_t.\end{aligned}$$

with correlated innovations across factors  $\boldsymbol{\omega}_t \sim N(\boldsymbol{\omega}_t | \mathbf{0}, \mathbf{U})$ .

- These relationships imply

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu}_t + \boldsymbol{\Phi}(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\mu}_{t-1}) + \boldsymbol{\omega}_t.$$

## **Stochastic Volatility Components**

### **COMMENTS:**

- **Mean reversion:** marginal/overall volatility levels  $\mu_{jt}$ .
- **Volatility Persistence:**  $\Phi$  high.
- **Global Risk Parameter:** Common level of volatility across factors when  $\mu_{jt} = \mu_t, \forall j$ .
- **Time-varying factor effects:** Series  $\mu_t$  determine the relative importance of each one of the factors over time.
- Independent univariate SV models for each one of the  $q$  idiosyncratic variances  $\psi_{it}$ .

## Specifications

### IDENTIFICATION:

- Constant or "slowly varying" loadings matrix  $\mathbf{X}_t = \mathbf{X}$ ,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ x_{2,1} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ x_{k,1} & x_{k,2} & x_{k,3} & \cdots & 1 \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \cdots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{q,1} & x_{q,2} & x_{q,3} & \cdots & x_{q,k} \end{pmatrix}.$$

- Series order defines interpretation of factors.
- Informative priors on SVM innovations variances.

## **Model Fitting and Bayesian Analysis**

Inference based on a fixed sample over  $t = 1, \dots, T$

- Bayesian analysis via posterior simulations (**GIBBS-Metropolis**).
- **MCMC** samples from the *joint posterior* for,
  - shrinkage model parameters, dynamic factor model parameters and
  - *latent processes: factors and volatilities*

$$\{ \mathbf{f}_t, \boldsymbol{\lambda}_t, \psi_{it} : t = 1, \dots, T \}$$

Sequential Analysis over  $t = T + 1, T + 2, \dots$

- **Sequential Particle Filtering** to update posterior samples

$$\cdots \rightarrow p(\cdot | D_{t-1}) \rightarrow p(\cdot | D_t) \rightarrow \cdots$$

- and compute/revise predictive distributions

$$\cdots \rightarrow p(\mathbf{y}_t | D_{t-1}) \rightarrow p(\mathbf{y}_{t+1} | D_t) \rightarrow \cdots$$

## References

### Stochastic Volatility Models:

- N Shephard (1994, 96), Harvey, Ruiz and Shephard (1994).
- E Jacquier, NG Polson and PE Rossi (1994, 95).
- S Kim, N Shephard and S Chib (1998).

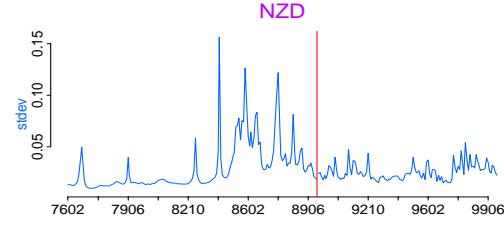
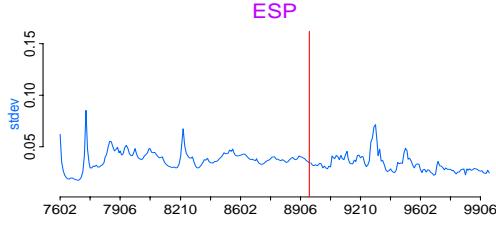
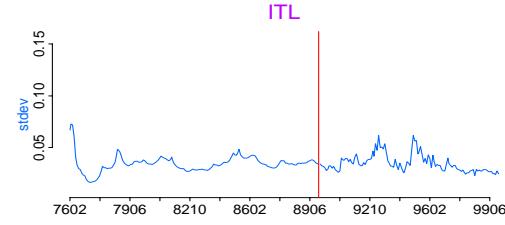
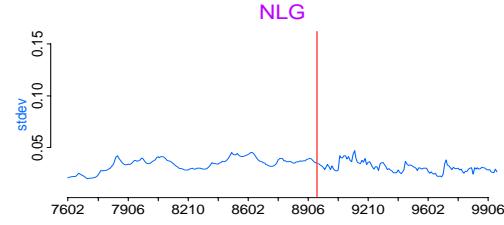
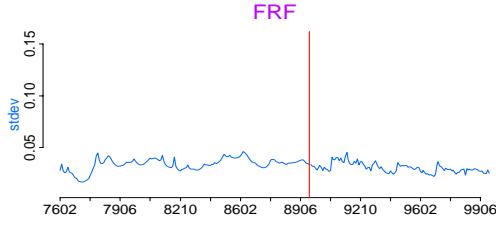
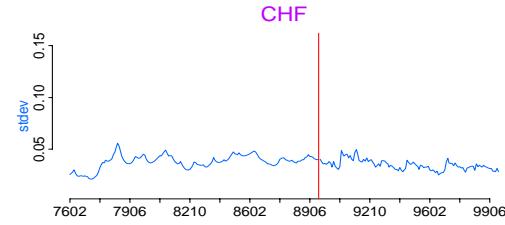
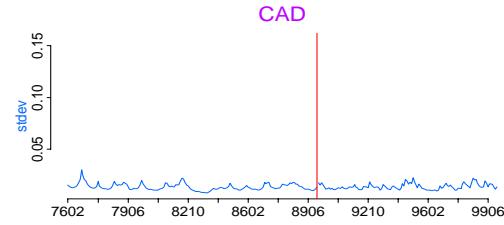
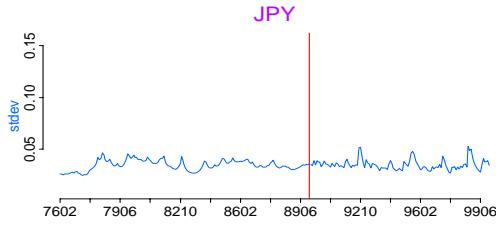
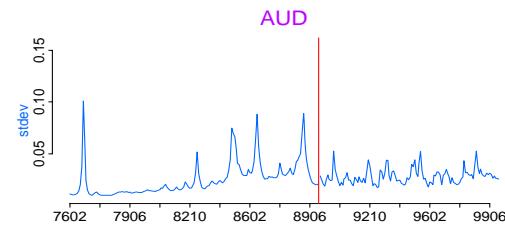
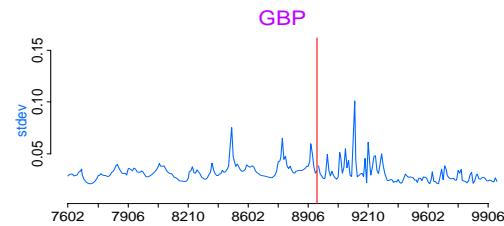
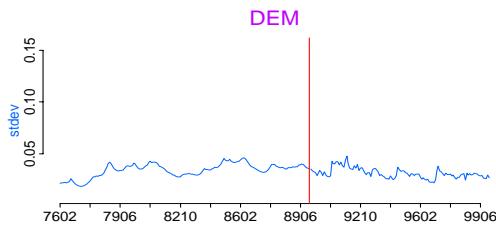
### Dynamic Factor Models:

- N Shephard and M Pitt (1999), O Aguilar and M West (2000).

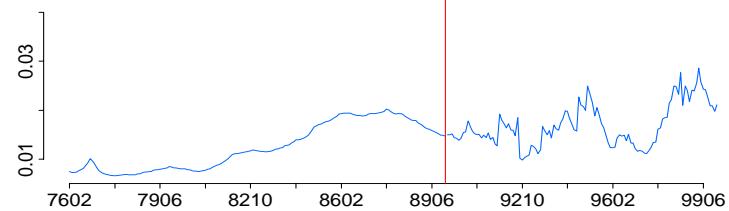
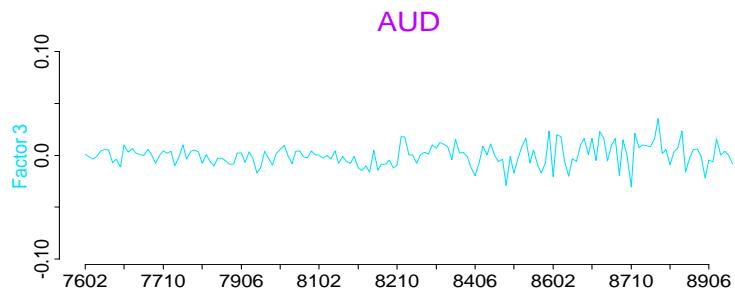
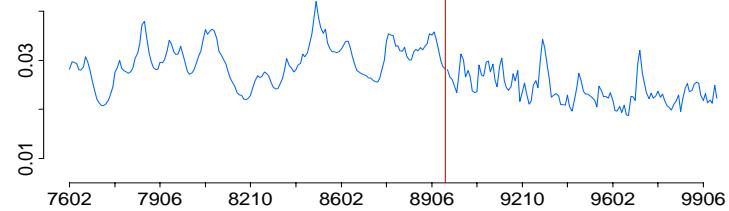
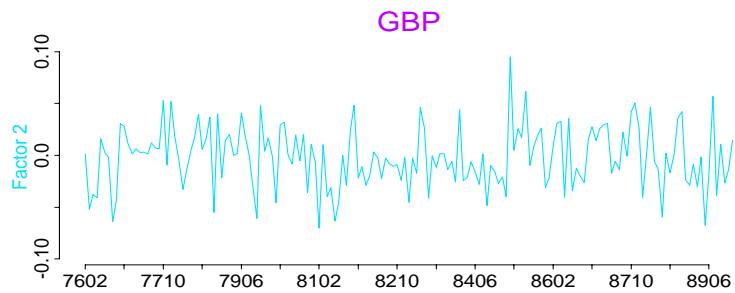
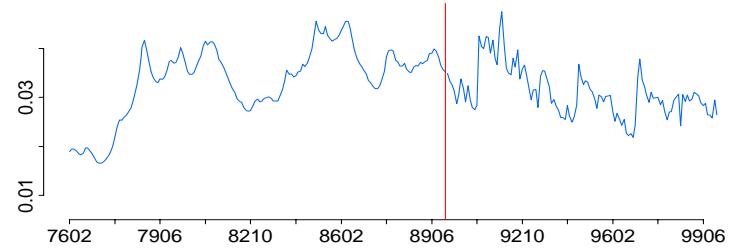
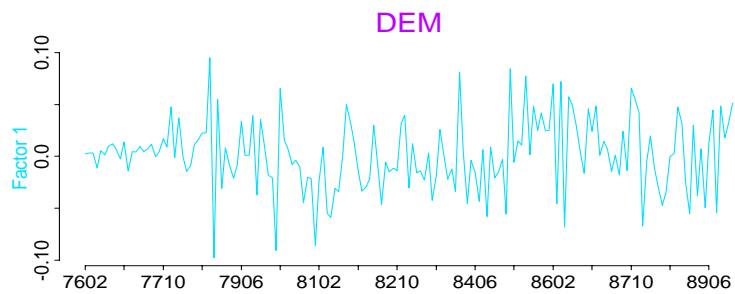
### Particle Filtering:

- M Pitt and N Shephard (1999), O Aguilar and M West (2000).
- J Liu and M West (1999).
- M West (1993), NJ Gordon, DJ Salmond and AFM Smith (1993).

# Volatilities of Currency Returns



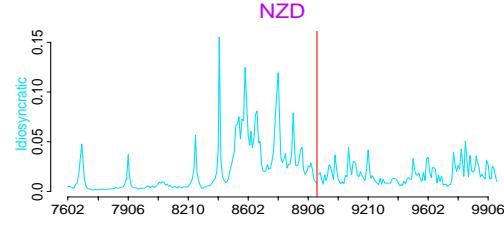
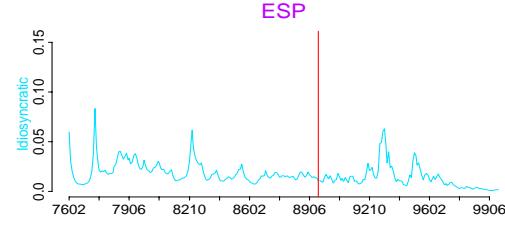
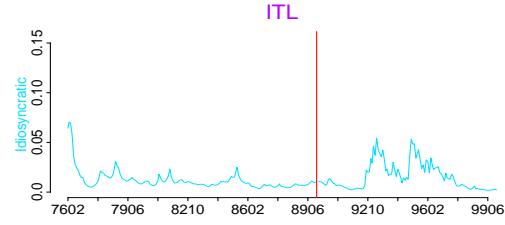
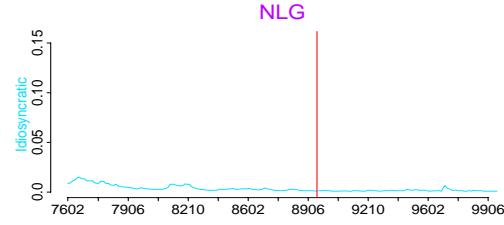
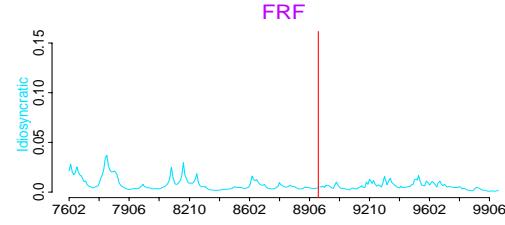
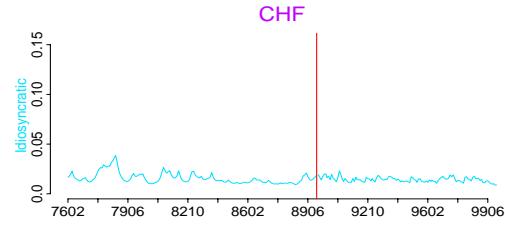
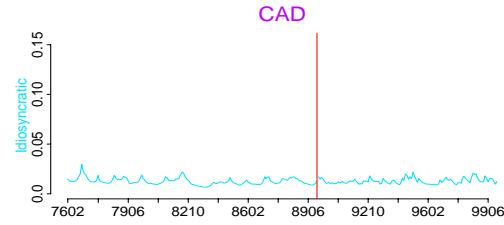
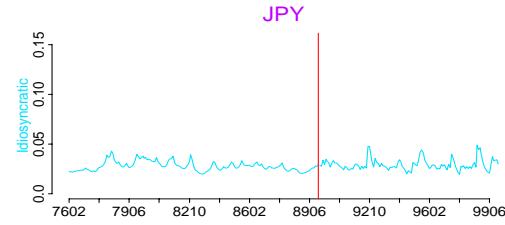
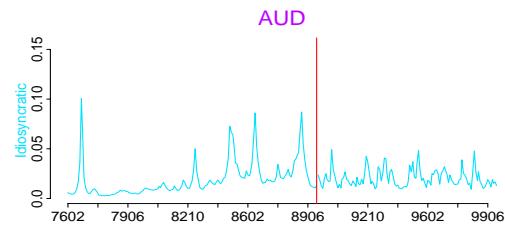
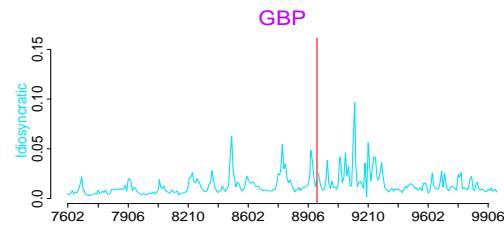
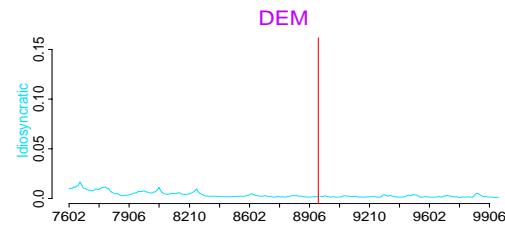
## ***Latent Factors and Their Volatilities***



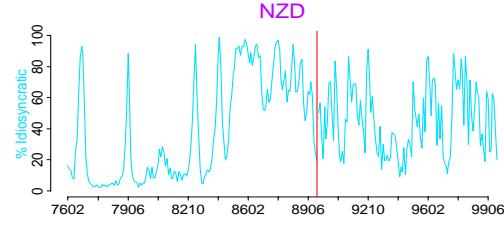
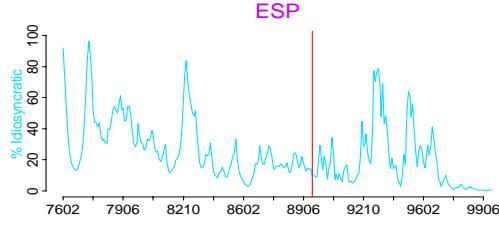
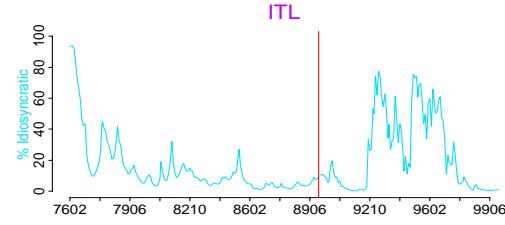
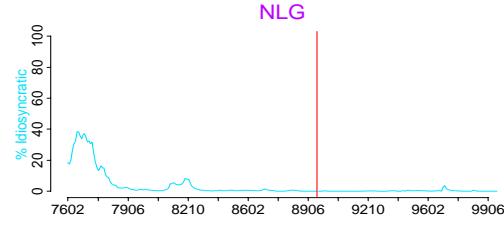
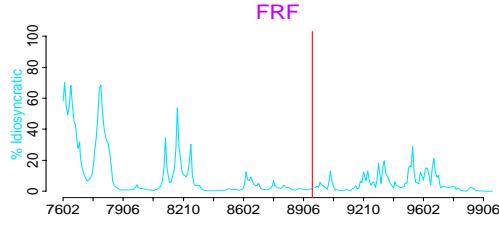
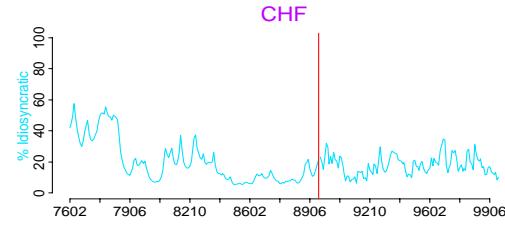
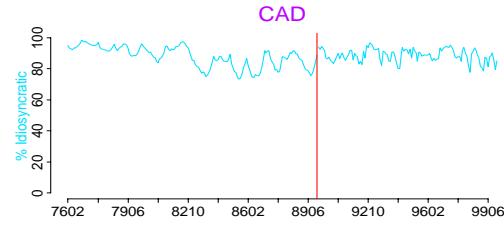
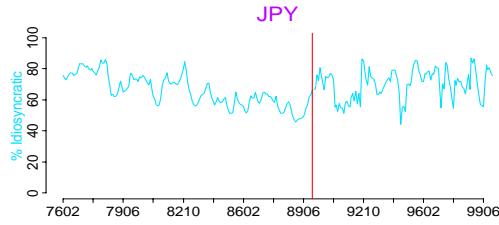
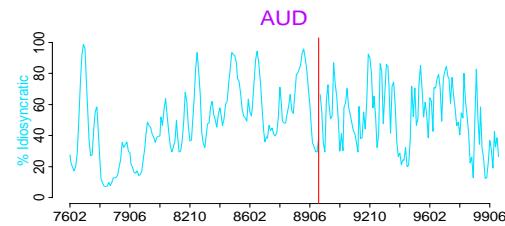
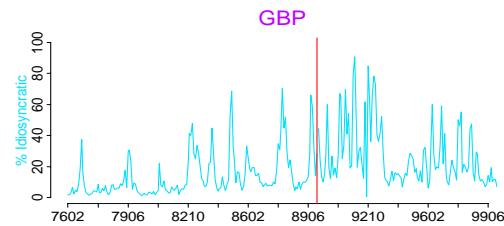
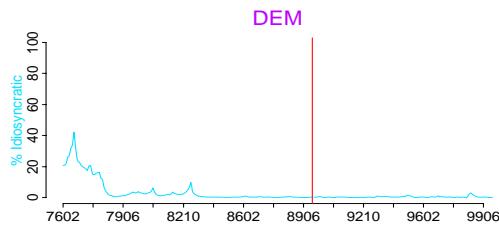
## **Factor Loadings Matrix X**

	Factor 1	Factor 2	Factor 3
<b>DEM</b>	<b>1.00 (0.00)</b>	<b>0.00 (0.00)</b>	<b>0.00 (0.00)</b>
GBP	0.08 (0.01)	1.00 (0.00)	0.00 (0.00)
AUD	0.15 (0.02)	0.16 (0.05)	1.00 (0.00)
JPY	0.52 (0.03)	0.23 (0.03)	0.40 (0.04)
CAD	<b>0.02 (0.00)</b>	<b>0.09 (0.01)</b>	<b>0.22 (0.02)</b>
CHF	<b>1.00 (0.03)</b>	<b>0.08 (0.01)</b>	<b>0.01 (0.00)</b>
FRF	<b>0.95 (0.01)</b>	<b>0.03 (0.02)</b>	-0.01 (0.01)
NLG	<b>0.99 (0.01)</b>	<b>0.01 (0.02)</b>	0.00 (0.01)
ITL	<b>0.92 (0.01)</b>	<b>0.02 (0.02)</b>	-0.02 (0.01)
ESP	<b>0.93 (0.01)</b>	<b>0.05 (0.02)</b>	-0.01 (0.01)
NZD	0.17 (0.01)	0.34 (0.01)	0.84 (0.01)

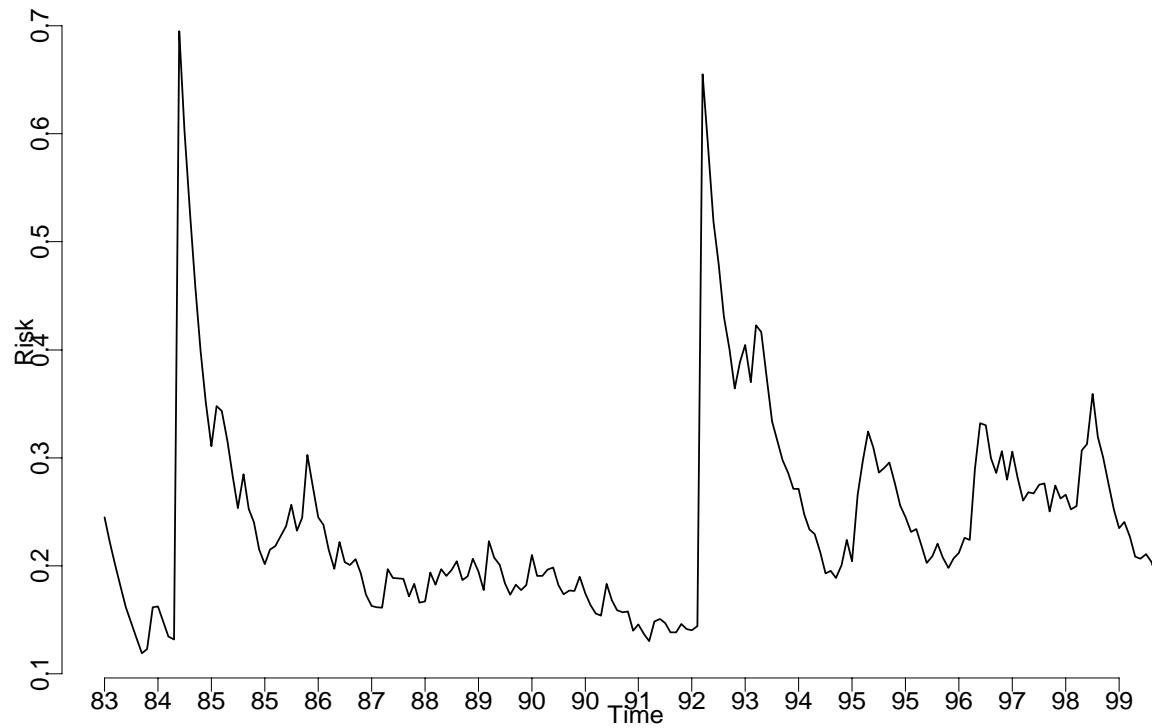
# *Idiosyncratic Volatilities*



# ***Idiosyncratic Volatilities as % of Total***



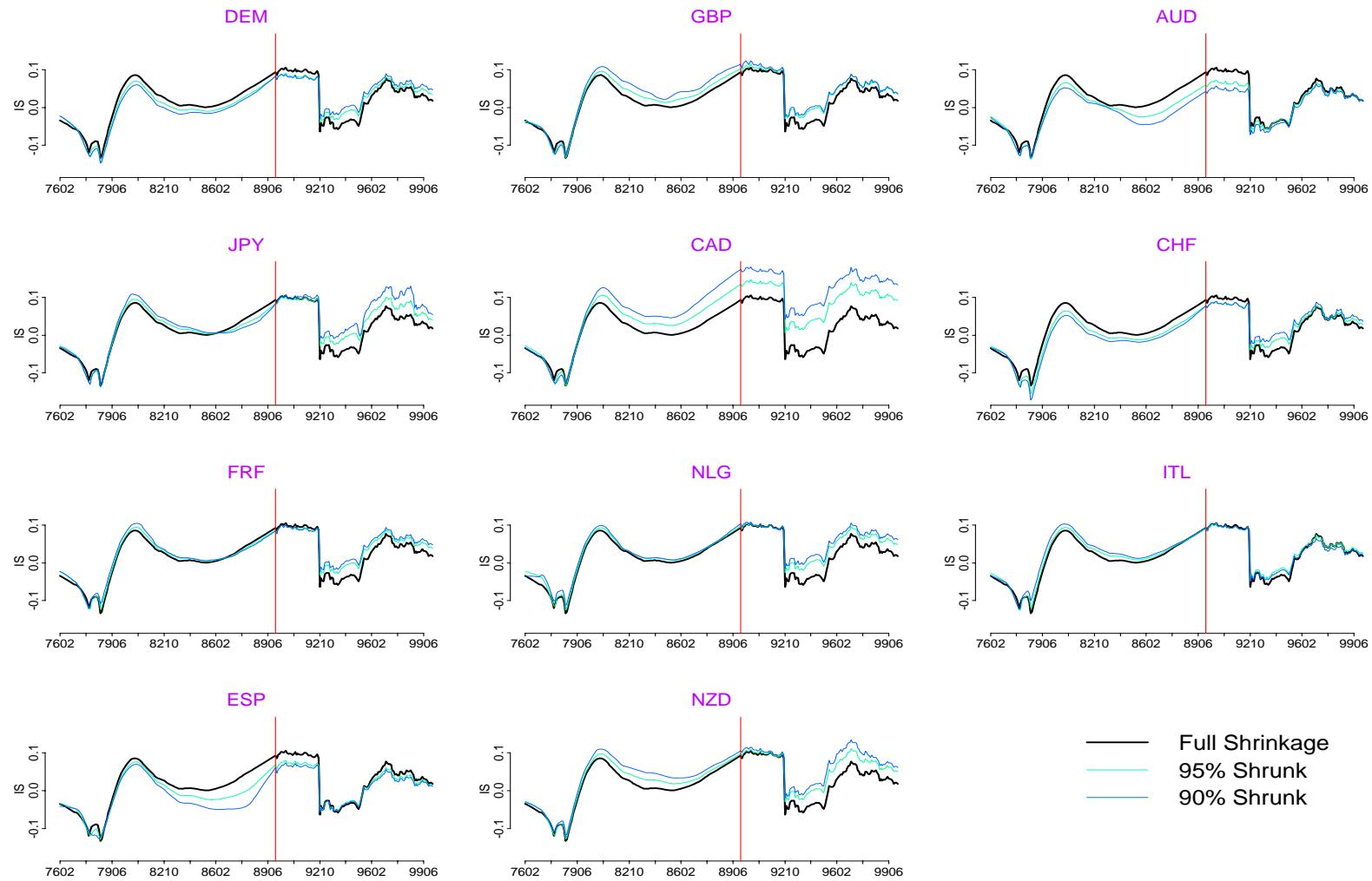
## *Global Risk Parameter*



# Regression Parameter for Short Interest Rates



# Regression Parameter for Short Interest Rates

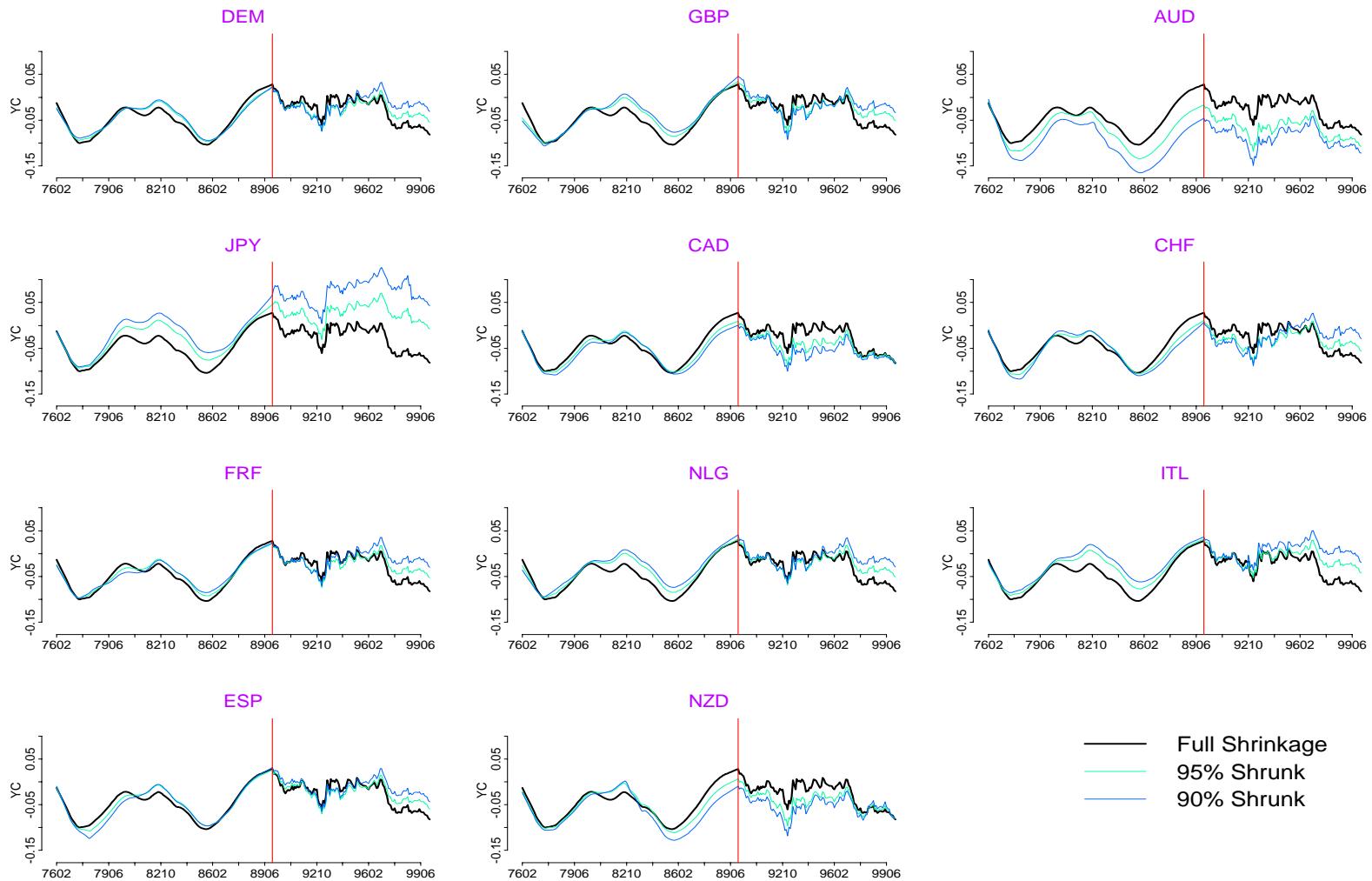


— Full Shrinkage  
— 95% Shrunk  
— 90% Shrunk

# Regression Parameter for Yield Curve



# Regression Parameter for Yield Curve

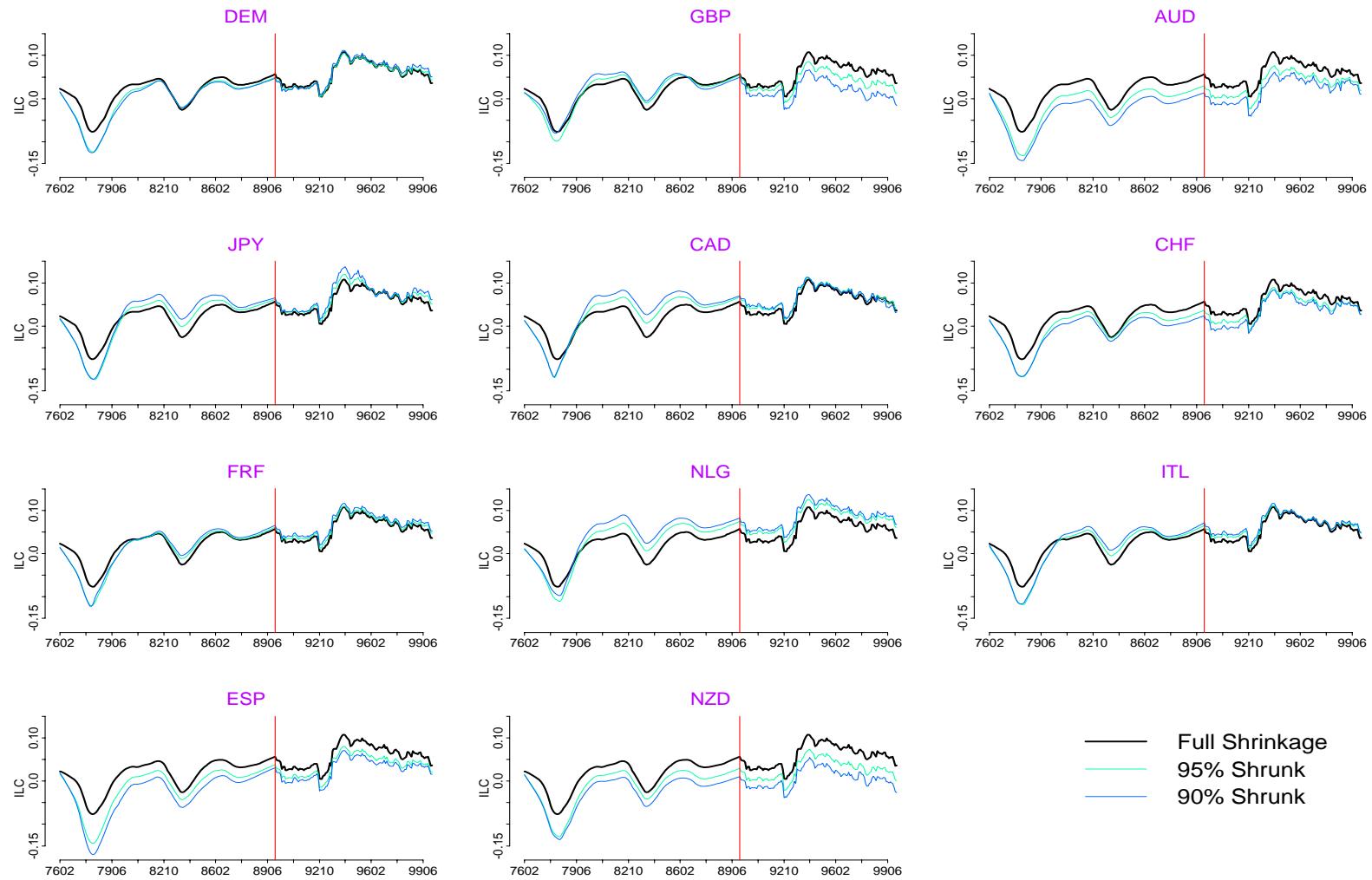


— Full Shrinkage  
— 95% Shrunk  
— 90% Shrunk

# Regression Parameter for Interest Rate Acceleration

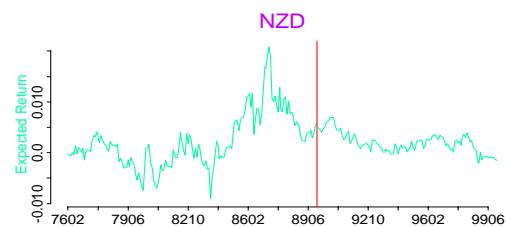
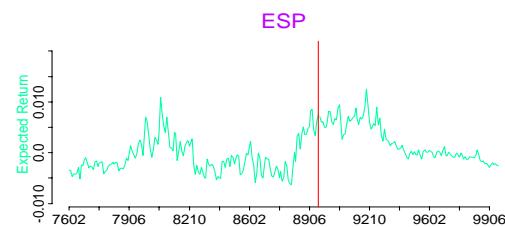
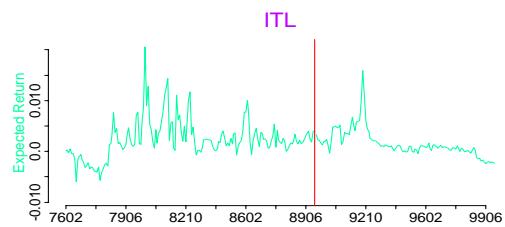
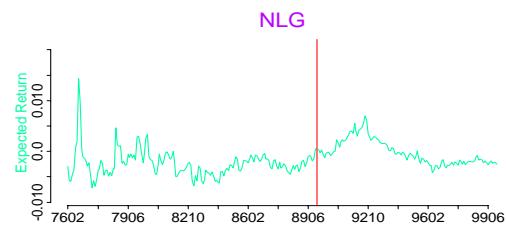
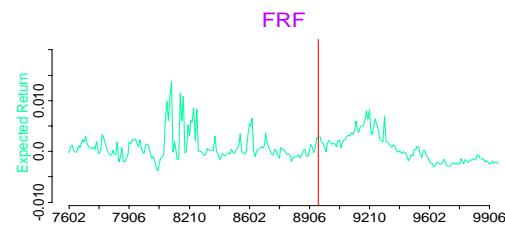
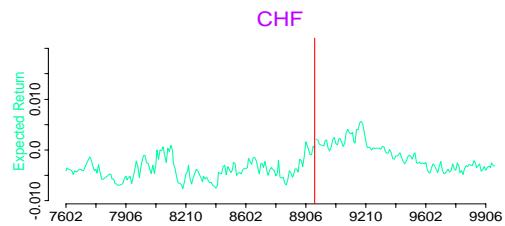
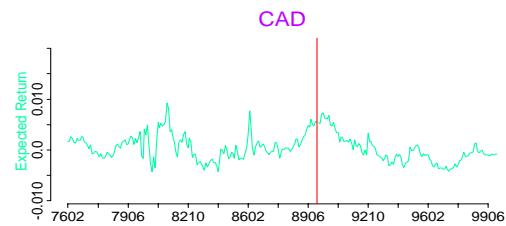
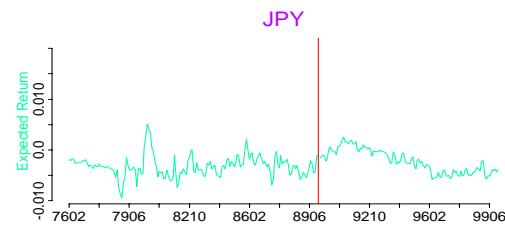
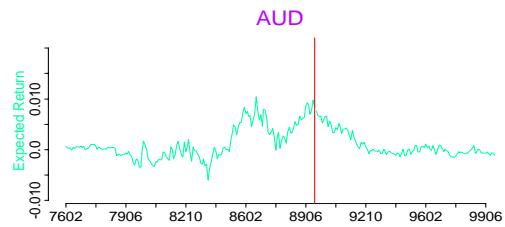
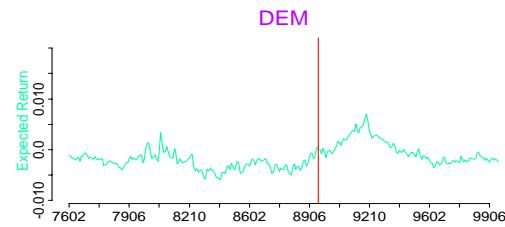


# Regression Parameter for Interest Rate Acceleration



— Full Shrinkage  
— 95% Shrunk  
— 90% Shrunk

# Expected Return Estimates



## Dynamic Asset Allocation

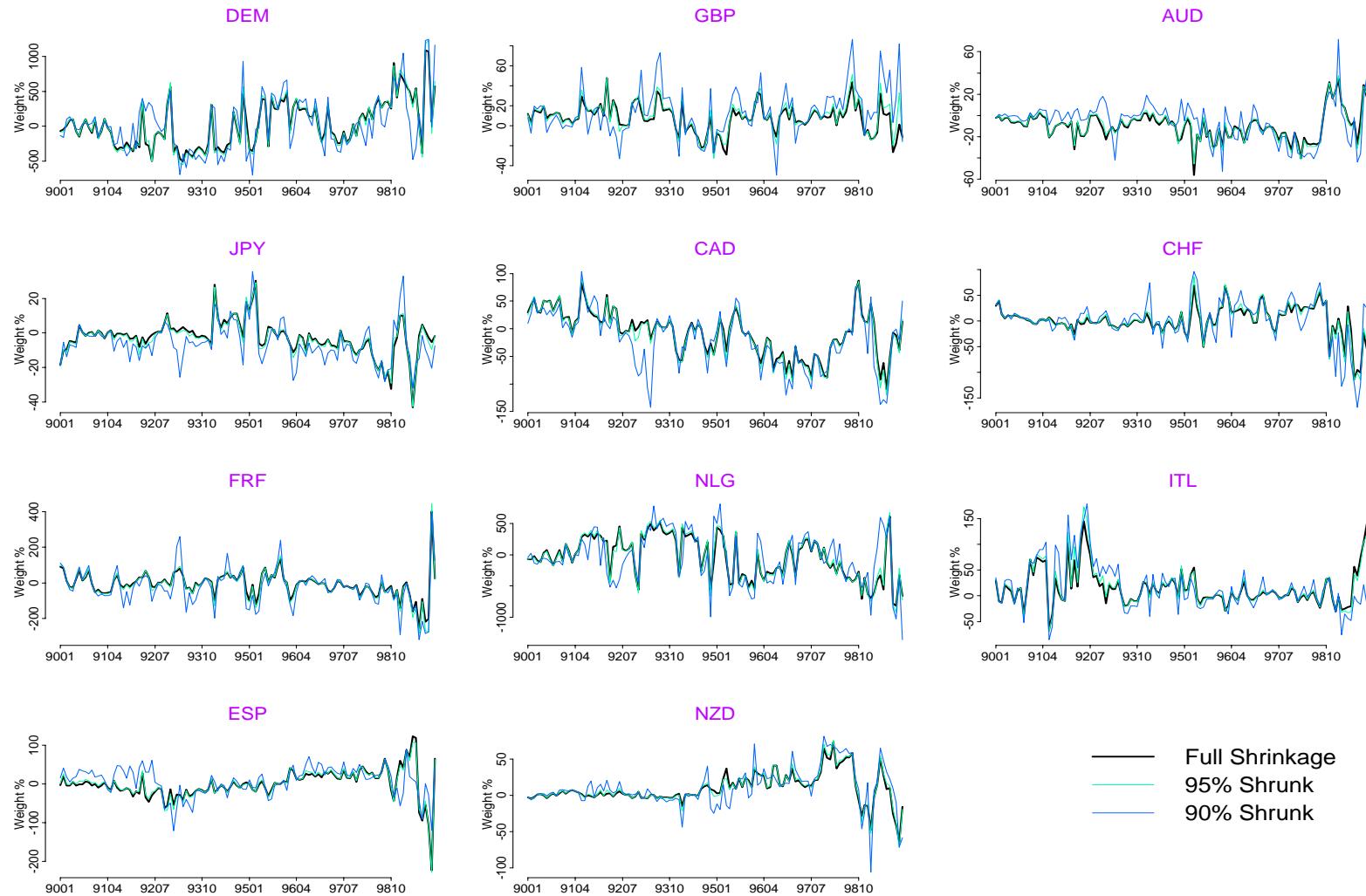
Let  $\mathbf{a}_t = (a_{1t}, a_{2t}, \dots, a_{qt})'$  be the one-step ahead portfolio where  $a_{it}$  is the proportion of wealth invested in the  $i$ -th currency.

- Portfolio return at time  $t$ ,  $r_t = \mathbf{a}_t' \mathbf{y}_t$ .
- **Forecast:** predictive distributions  $p(\mathbf{y}_t | D_{t-1})$  with  $\mathbf{g}_t$  and  $\mathbf{G}_t$  denoting the corresponding predictive mean and covariance matrix.

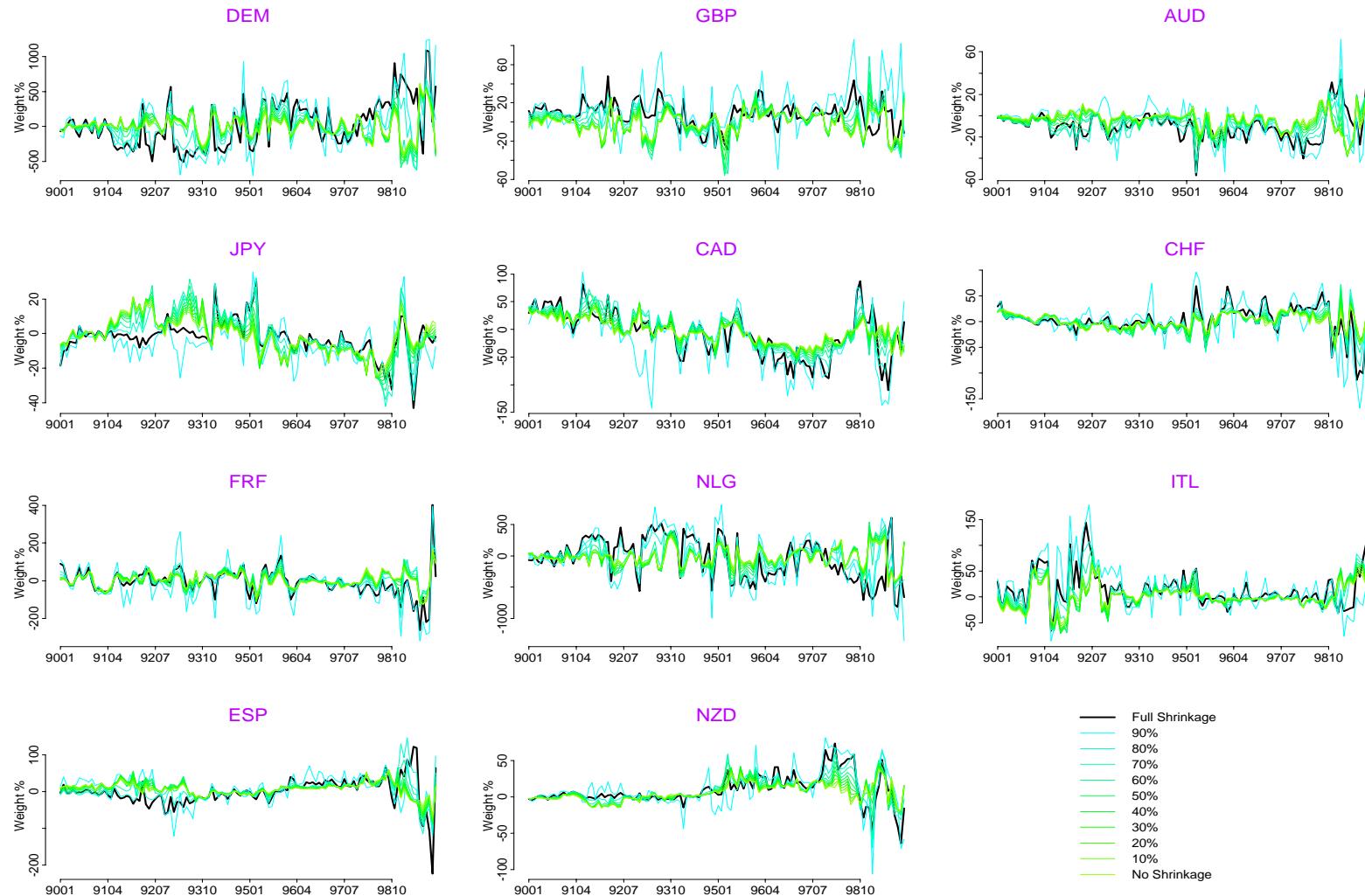
### Optimization:

- minimise risk:  $\mathbf{a}_t' \mathbf{G}_t \mathbf{a}_t$ ,
- for a specified "target" return:  $\mathbf{a}_t' \mathbf{g}_t = m$ ,
- unconstrained, transaction costs efficient, liquidity and turnover.

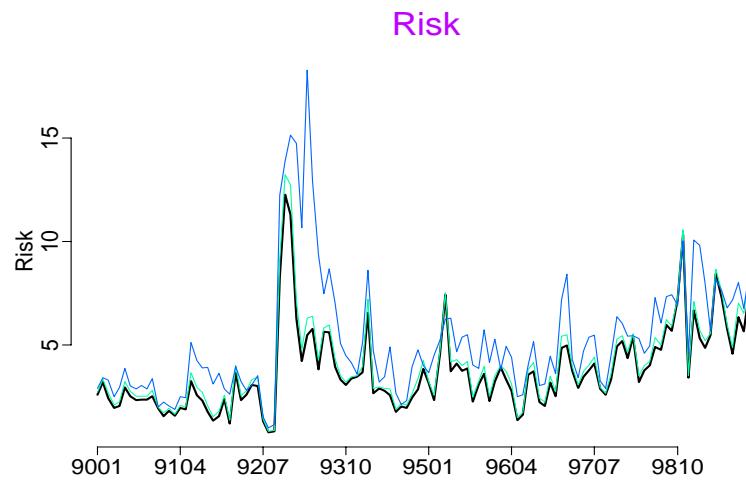
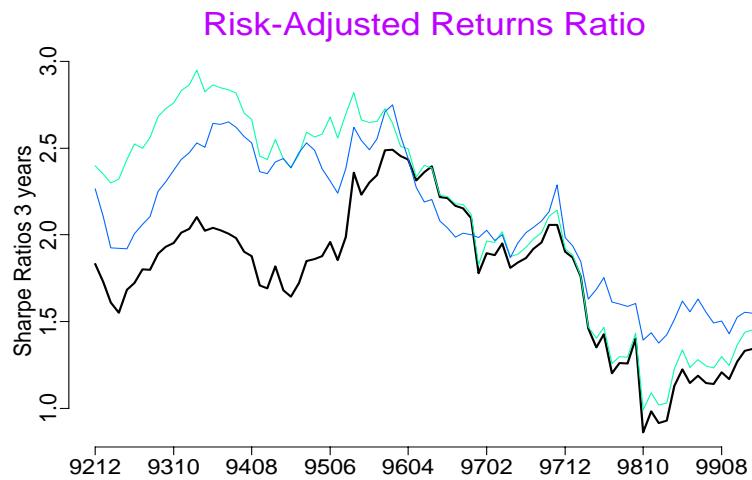
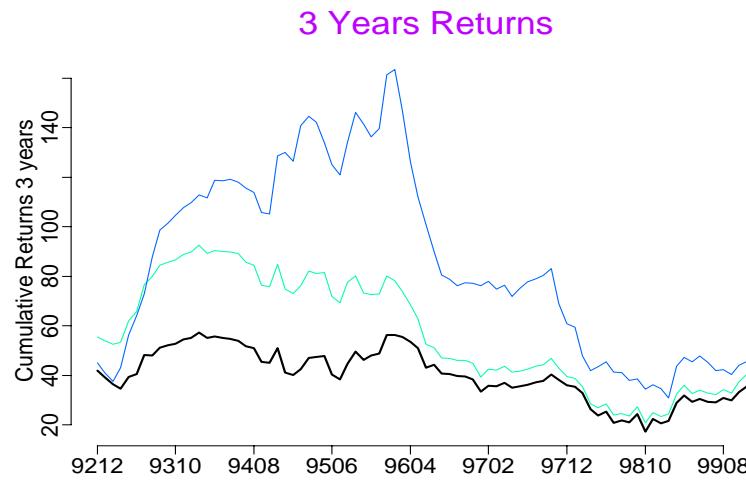
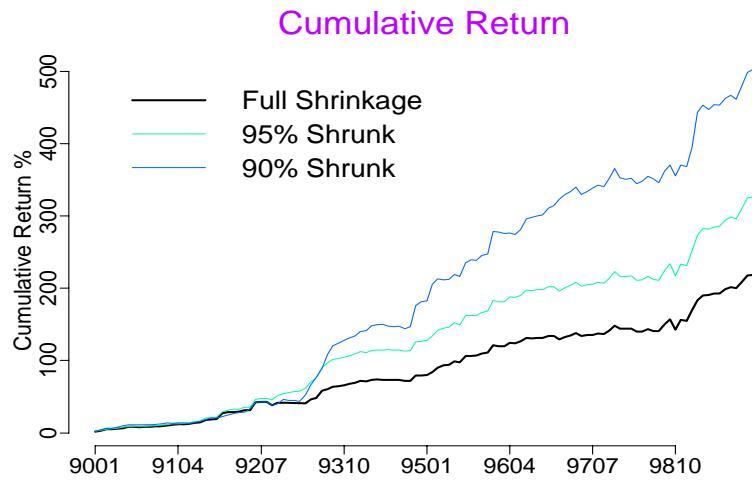
## $a_t$ – Unconstrained Portafolio with $m = 0.005$



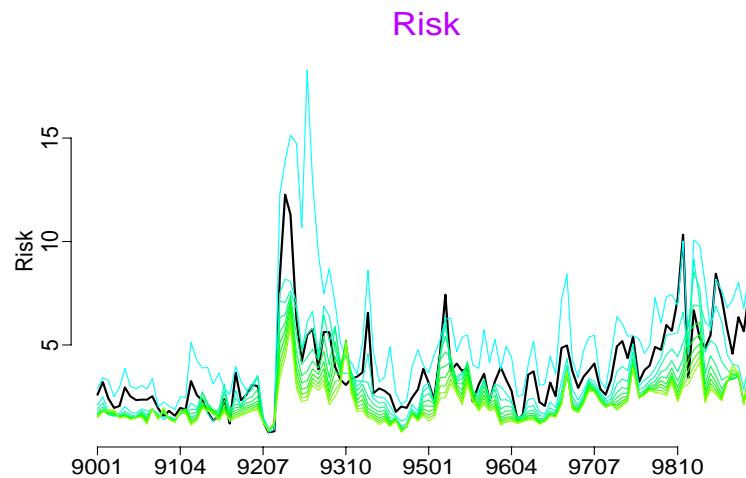
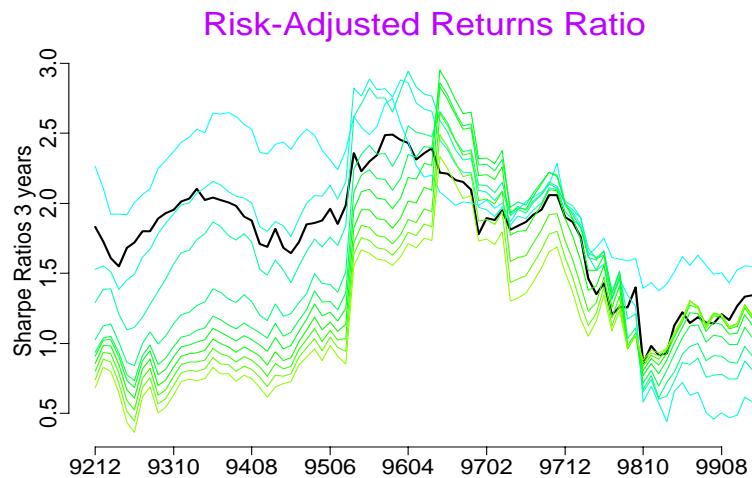
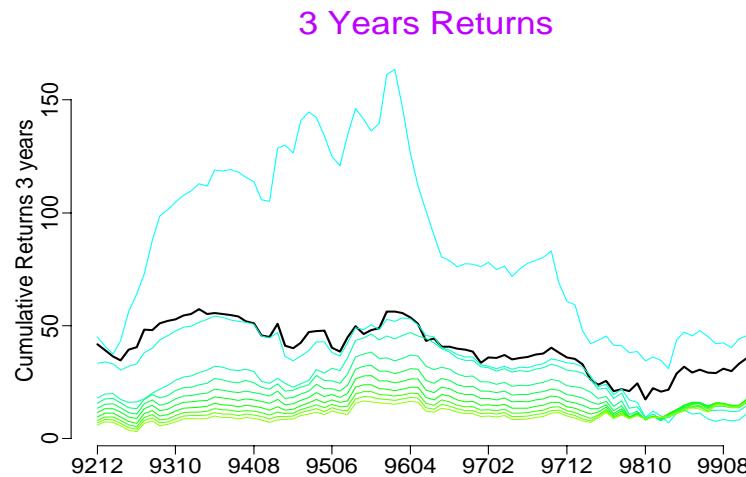
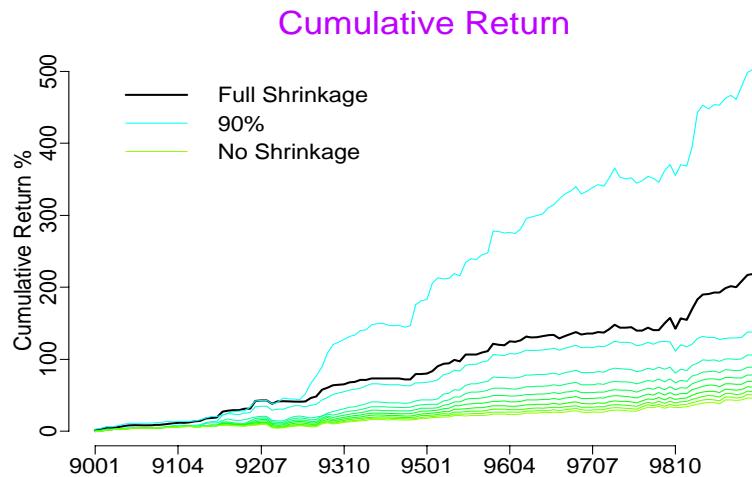
## $a_t$ – Unconstrained Portafolio with $m = 0.005$



# Cumulative Returns and Performance



# Cumulative Returns and Performance



## **Summary and Future Research**

- Peaks in the volatilities are consistent with positive correlations in volatility across the factors (EU (DEM), AUD+NZD and JPY+GBP).
  - Non-negligible over-all idiosyncratic variations.
  - Improvements in short-term forecasting and decision making through partial shrinkage models.
  - Novel and customized MCMC algorithms for fitting and computation of posterior and predictive analyses.
- 
- Improved sequential particle filtering.
  - Change points:
    - Intervention analysis.
    - Heavy tailed components.
  - Model uncertainty: number of factors, order of factors.

# Time-Frequency Decompositions

—

# State-Space Model-Based Approaches

Mike West

Institute of Statistics & Decision Sciences

Duke University

<http://www.stat.duke.edu>

## **KEY THEME : Latent structure in time series**

“Finding” latent processes/signals in data

- Nonstationary time series: “hidden” quasi-periodicities
- Patterns of change over time in latent processes
- Time:frequency structure (in time domain)

“Building” latent structure in multiple time series

- Common underlying structure in multiple series
- Latent factor processes

# MODELS AND APPLIED CONTEXTS

State-space models:

- Stationary and/or nonstationary, time-varying parameters
- General decomposition theory for state space-space models
- Time series as latent components of complex models

Applied interests in:

- *EEG traces in clinical psychiatry studies*
- *Climatological and geochemical indicators/climatic change*
- ...

## A GENERAL TIME SERIES MODEL CLASS

Dynamic model for univariate time series  $y_t$ ,  $t = 1, 2, \dots$ :

$$Datum : \quad y_t = x_t + \nu_t$$

Signal :  $x_t$  = first element of vector  $\theta_t$

$$State : \quad \theta_t = G(\phi_t)\theta_{t-1} + \omega_t$$

- state vector  $\theta_t = (\theta_{t,1}, \dots, \theta_{t,d})'$
- state matrix  $G(\cdot)$  based on parameters  $\phi_t$
- time-varying parameters: e.g.,  $\phi_t = \phi_{t-1} + \partial\phi_t$
- measurement errors  $\nu_t$ , state innovations  $\omega_t$ 
  - perhaps zero-mean, normal or heavier-tailed, ...

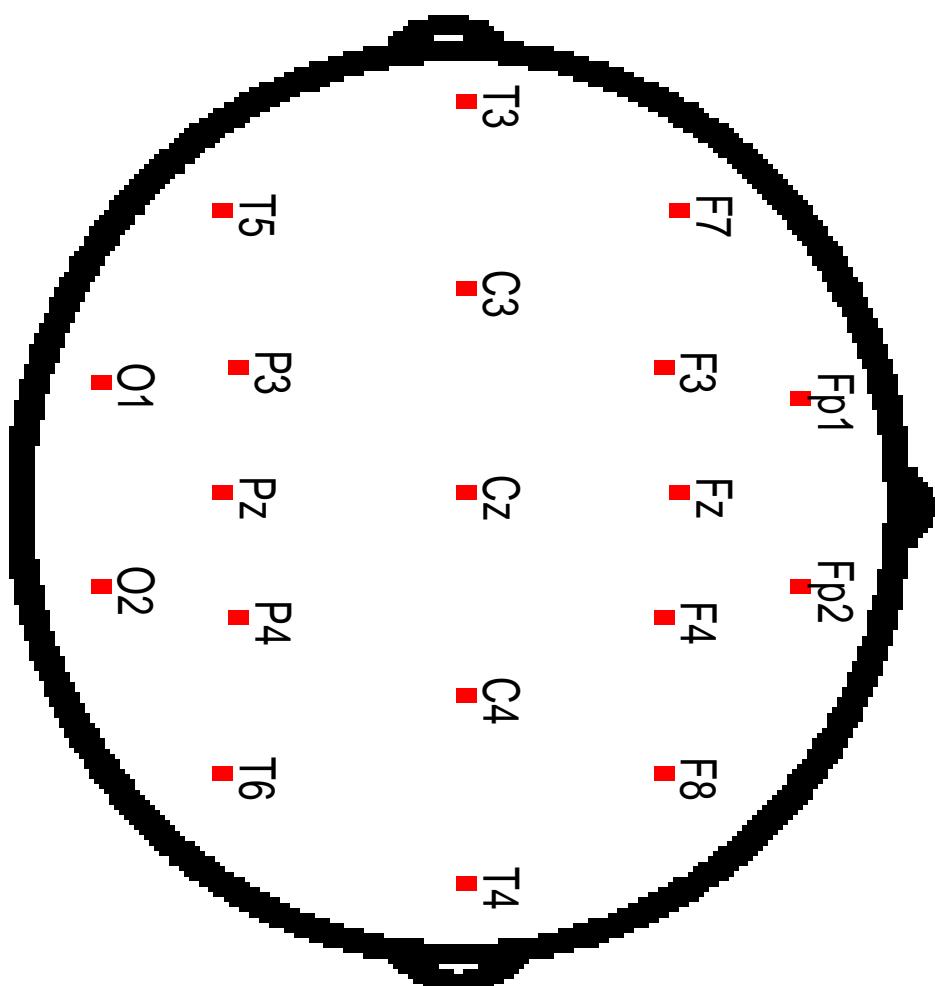
## KEY EXAMPLE: Time-Varying Autoregressions

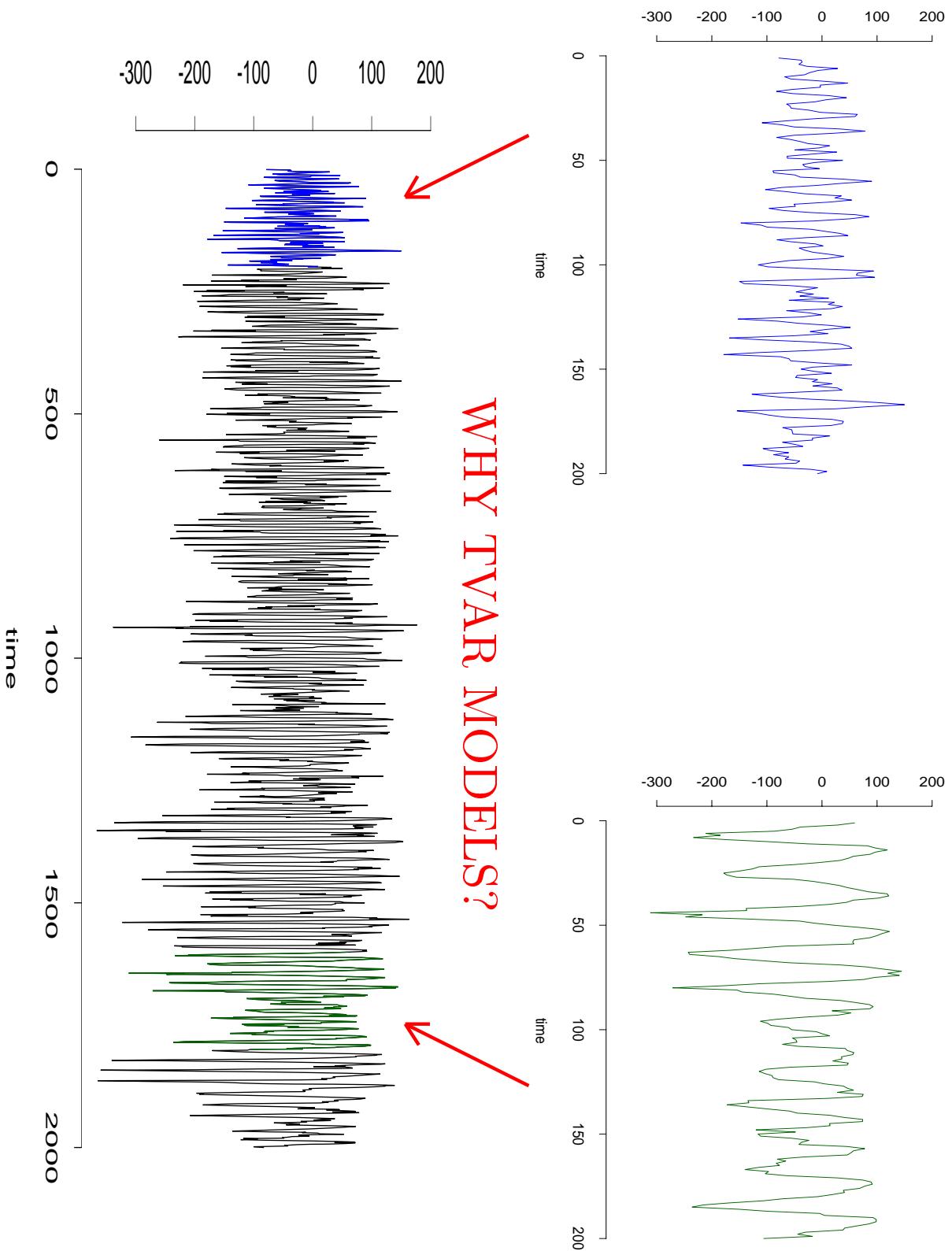
- TVAR( $d$ ) model:  $x_t = \sum_{j=1}^d \phi_{t,j} x_{t-j} + \epsilon_t$
- $\theta_t = (x_t, x_{t-1}, \dots, x_{t-d+1})'$
- time-varying AR parameters  $\phi_t = (\phi_{t,1}, \dots, \phi_{t,d})'$  and
$$G(\phi_t) = \begin{pmatrix} \phi_{t,1} & \phi_{t,2} & \cdots & \phi_{t,d-1} & \phi_{t,d} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$
- $\omega_t = \epsilon_t(1, 0, \dots, 0)'$  with  $\epsilon_t \sim N(0, \sigma_t^2)$
- Flexible model class: Nonstationary (nonlinear) models

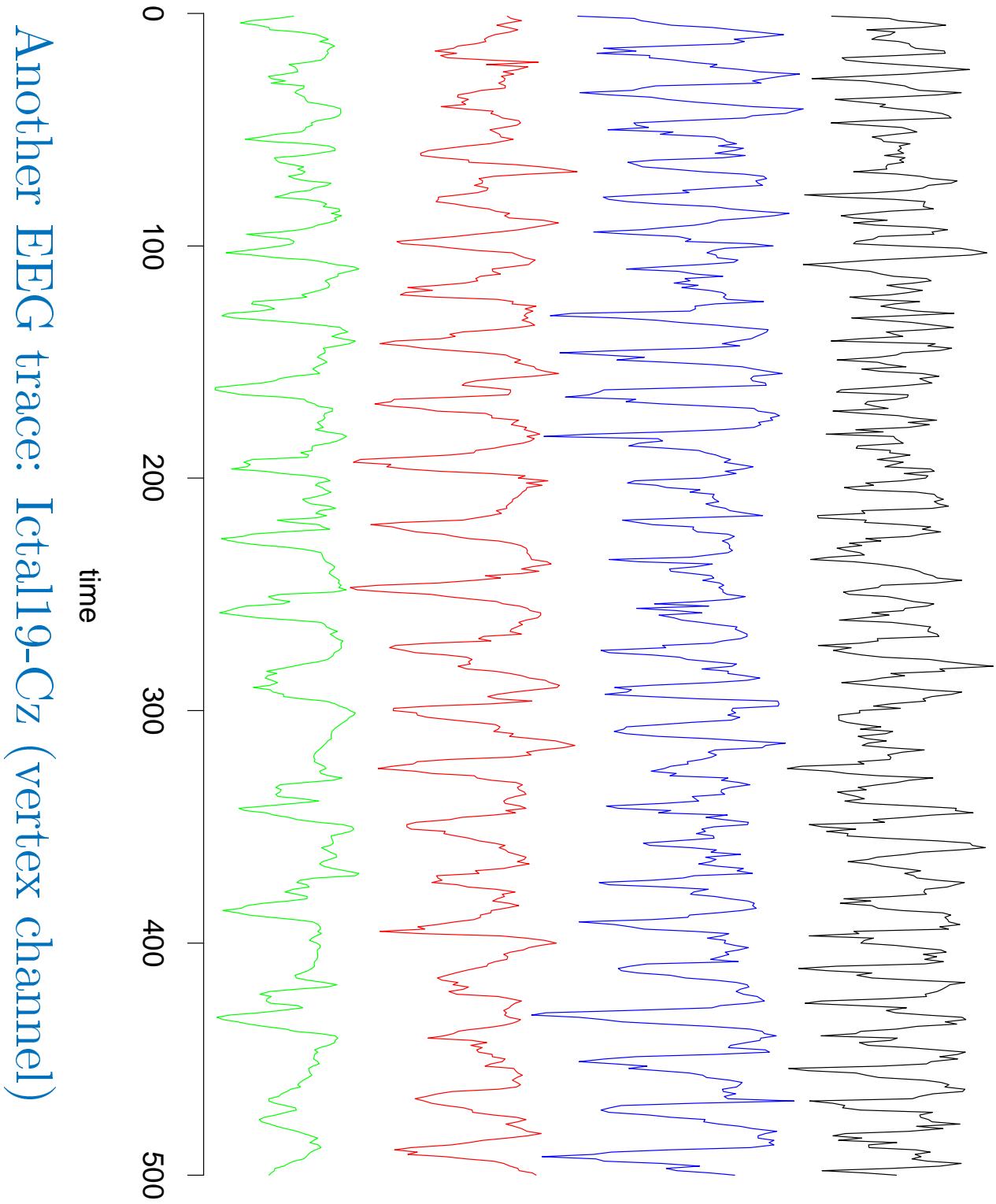
## **EEG STUDY**

(Raquel Prado & Andrew Krystal)

- Clinical uses of electroconvulsive therapy
  - Measure seizure treatment outcomes via long, multiple EEG (electroencephalogram) series – electrical potential fluctuations on scalp
  - Many multiple series: one seizure – 19 channels, 256/sec
- Models to:
- Characterise seizure “waveforms” ... time varying amplitudes at ranges of frequencies (alpha waves, etc)
  - Identify/extract latent components: infer **seizure effects**
  - Spatial connectivities: related multiple series







Another EEG trace: Ictal19-Cz (vertex channel)

## TIME SERIES DECOMPOSITION

Model “determined” by eigenvalues of state matrices  $G(\phi_t)$

With  $d_z$  complex pairs and  $d_a$  real eigenvalues:

$$x_t = \sum_{j=1}^{d_z} z_{t,j} + \sum_{j=1}^{d_a} a_{t,j}$$

- underlying latent processes  $z_{t,j}$  and  $a_{t,j}$  follow “simple” models, implied by the  $G(\cdot)$  sequence
- new uses of old theory of *Similar models*
- General: any  $G(\cdot)$ ,  $\omega_t$  – includes “nonlinear” models
- estimate model → estimate latent processes . . . Interpret?

## DECOMPOSITION THEORY

Eigenstructure:  $G(\phi_t) = E_t A_t E_t^{-1}$

Reparametrise model to diagonalise  $G(\phi_t)$ :  $\beta_t = H_t \theta_t$

Transforms to

$$x_t = (1, 1, \dots, 1)' \beta_t \quad \text{and} \quad \beta_t = A_t \mathbf{K}_t \beta_{t-1} + H_t \omega_t$$

$K_t$  depends on  $E_t, E_{t-1}$

- $K_t = I$  if  $G(\cdot)$  constant
- $K_t \approx I$  in “slowly varying” (unless “close” eigenvalues)
- If  $K_t = I$  then  $\beta_{t,i}$  are AR(1) processes, some complex

CONSTANT CASE:  $G$  has e-vals  $r_j \exp(\pm i\omega_j)$

$$x_t = \sum_{i=1}^d \beta_{t,i} = \sum_{j=1}^{d_z} z_{t,j} + \sum_{j=1}^{d_a} a_{t,j}$$

- real e-vals:  $a_{t,j} \equiv \beta_{t,j}$  is real AR(1) process

$$a_{t,j} = r_j a_{t-1,j} + \rho_{t,j}$$

- complex e-vals:  $z_{t,j} = \text{sum of conjugates } \beta_{t,j}, \beta_{t,j'}$

$$z_{t,j} = (2r_j \cos(\omega_j)) z_{t-1,j} - r_j^2 z_{t-2,j} + \rho_{t,j}^*$$

- quasi-periodic ARMA(2,1) with frequency  $\omega_j$
- sinusoid with randomly time-varying amplitude & phase

e.g., AR( $d$ ) models

## GENERAL TIME-VARYING CASE:

$G(\phi_t)$  has e-vals  $r_{t,j} \exp(\pm i\omega_{t,j})$

- $a_{t,j} \approx \text{TVAR}(1)$

$$a_{t,j} = r_{t,j} a_{t-1,j} + \rho_{t,j}$$

- $z_{t,j} \approx \text{TVARMA}(2,1)$

$$z_{t,j} = (2r_{t,j} \cos(\omega_{t,j})) z_{t-1,j} - r_{t,j}^2 z_{t-2,j} + \rho_{t,j}^*$$

- Time-varying quasi-periodic process with randomly time-varying amplitude, phase & frequency

- $K_t$  : Latent processes “mix” between  $t-1$  and  $t$
- EEG studies:  $K_t \approx I$  to within  $10^{-5}$  elementwise

## TIME: FREQUENCY ANALYSIS

- Fit flexible TVAR models: infer  $\phi_t$  over time
- Estimate latent components and their frequencies, amplitudes over time
- Time domain representation of time-varying spectral structure
- $z_{t,j}$  process: “instantaneous” spectral peak at  $\omega_{t,j}$  “characteristic” frequency  $\omega_{t,j}$
- Often, some  $z_{t,j}$  physically meaningful, some (high frequency) represent noise, model approximation
- $a_{t,j}$  – noise, model approximation and (possibly) low frequency “trend”

## MODEL FITTING

- Standard **dynamic linear model** analysis
  - filtering and smoothing for  $\phi_t$  (linear)
  - also for  $\sigma_t^2$  (nonlinear)
  - closed form posterior analysis
- Time-variation in  $\phi_t$  and  $\sigma_t^2$ 
  - determined by **discount factors**
- Discount factors and model order specified
  - inspection of marginal likelihood function
- West and Harrison (1997 2nd Edition)
- West, Prado & Krystal (JASA, to appear)

MODEL EXTENSIONS in which  $x_t$  is unobserved (latent)

$$y_t = F'_t \mu_t + x_t + \nu_t \quad x_t = \text{TVAR, as above}$$

$\mu_t$  = state space model

e.g.,  $\mu_t$  contains “local” level, growth parameters for smooth but slowly-varying trends

**Model Fitting:** simulation via Markov Chain Monte Carlo

Iteratively simulate from posterior of

- all  $\mu_t, x_t$  from state space model
- all parameters  $\phi_t, \sigma_t^2$ , etc

**Extensive literature:** Frühwirth-Schnatter (94); Carter & Kohn (94-6); Shephard and de Jong (95); West (95-7); + ...

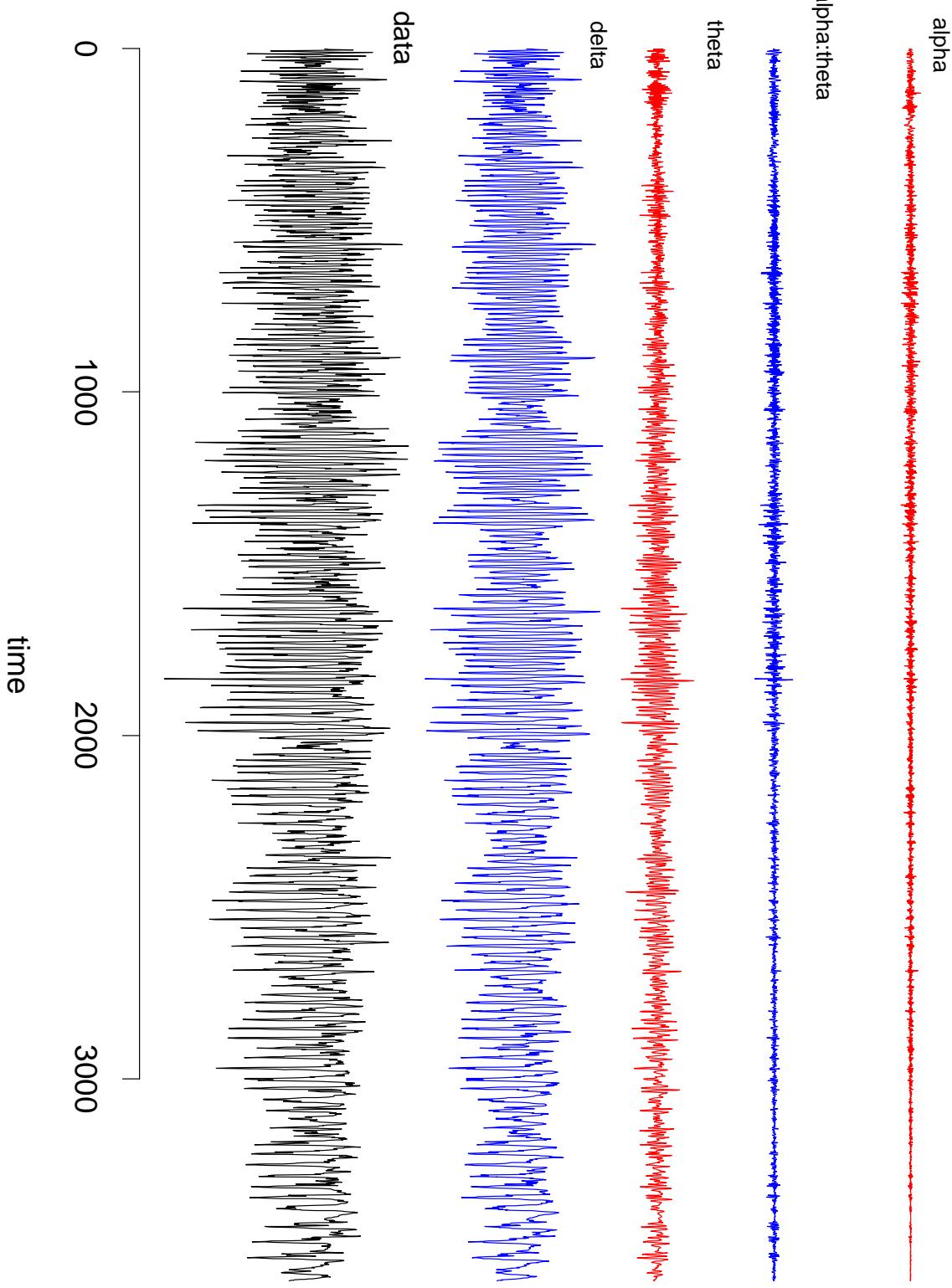
## EEG series: Ictal19-Cz

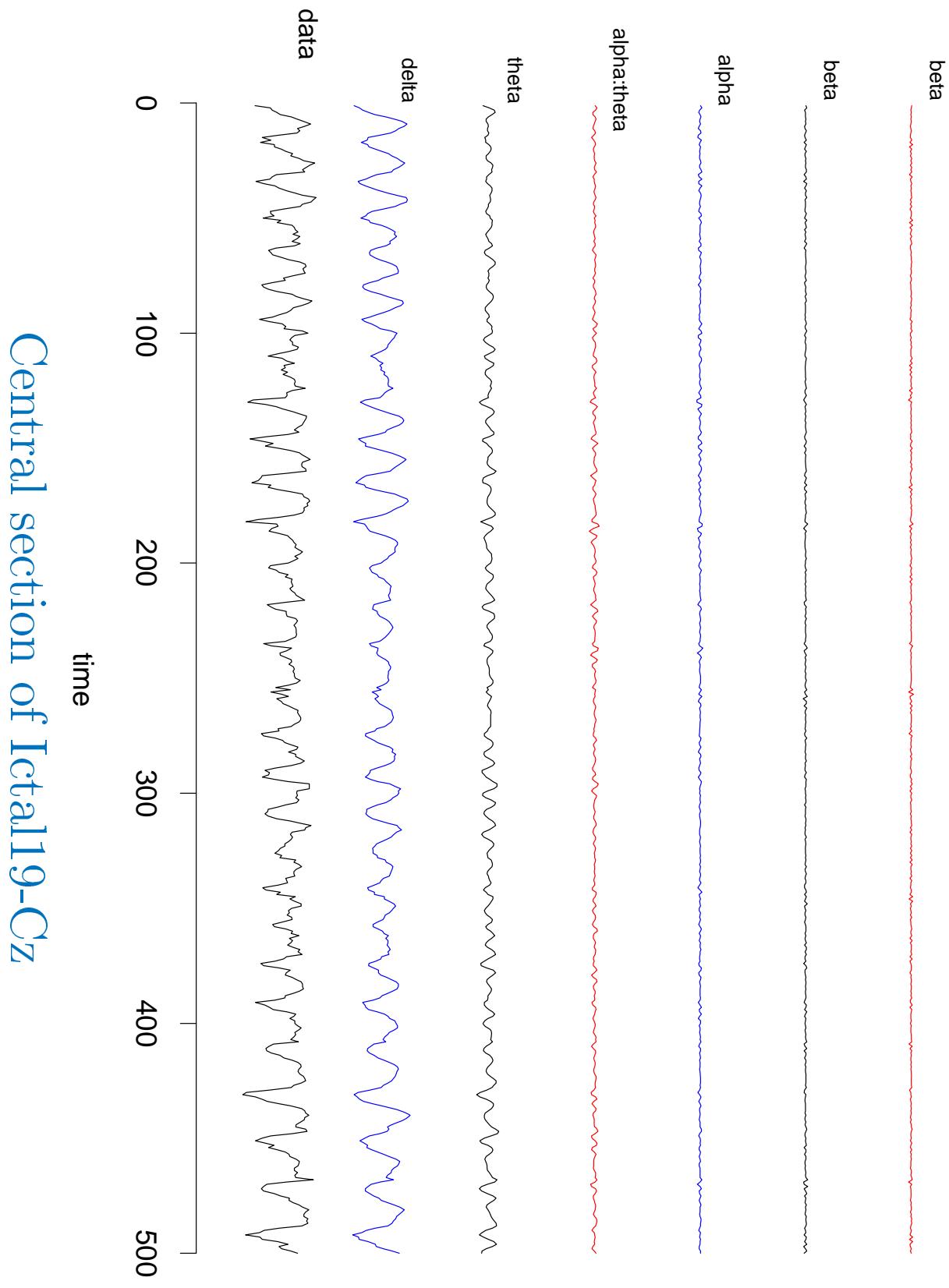
- TV-AR(12), time-varying innovation variance  $\sigma_t^2$
- *Decomposition*: Posterior mean of  $\phi_{t,j}$  at each  $t$

## More EEG: Series S26.Low cf S26.Mod

- Repeat seizures with varying ECT treatment  
treatment comparison study
- TVAR(20) with time-varying  $\sigma_t^2$
- More evident high frequency structure and “spiky” traces  
 $\rightarrow$  higher order models
- More latent components
- Identification issues

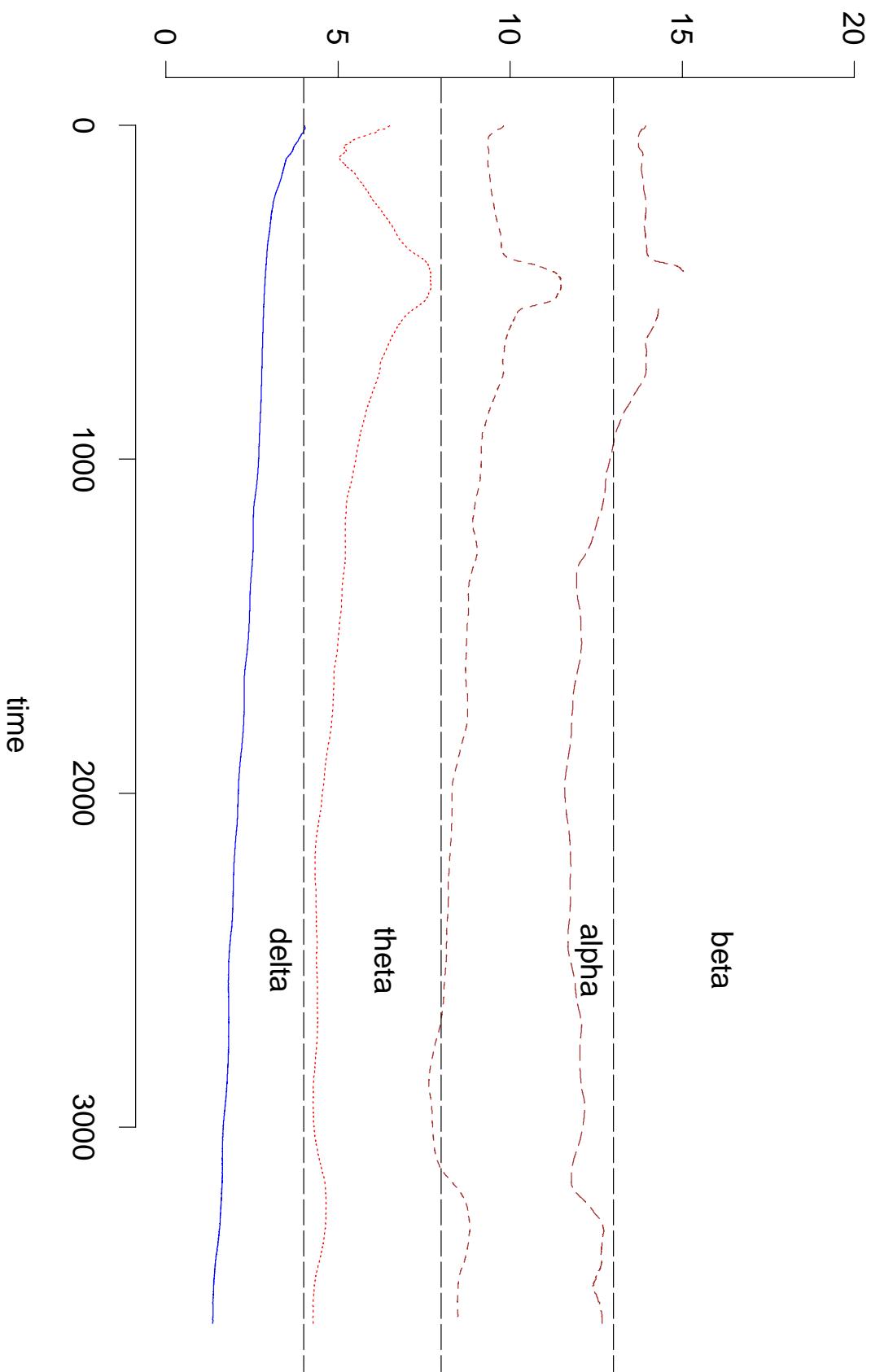
# Decomposition of Ictal19-CZ



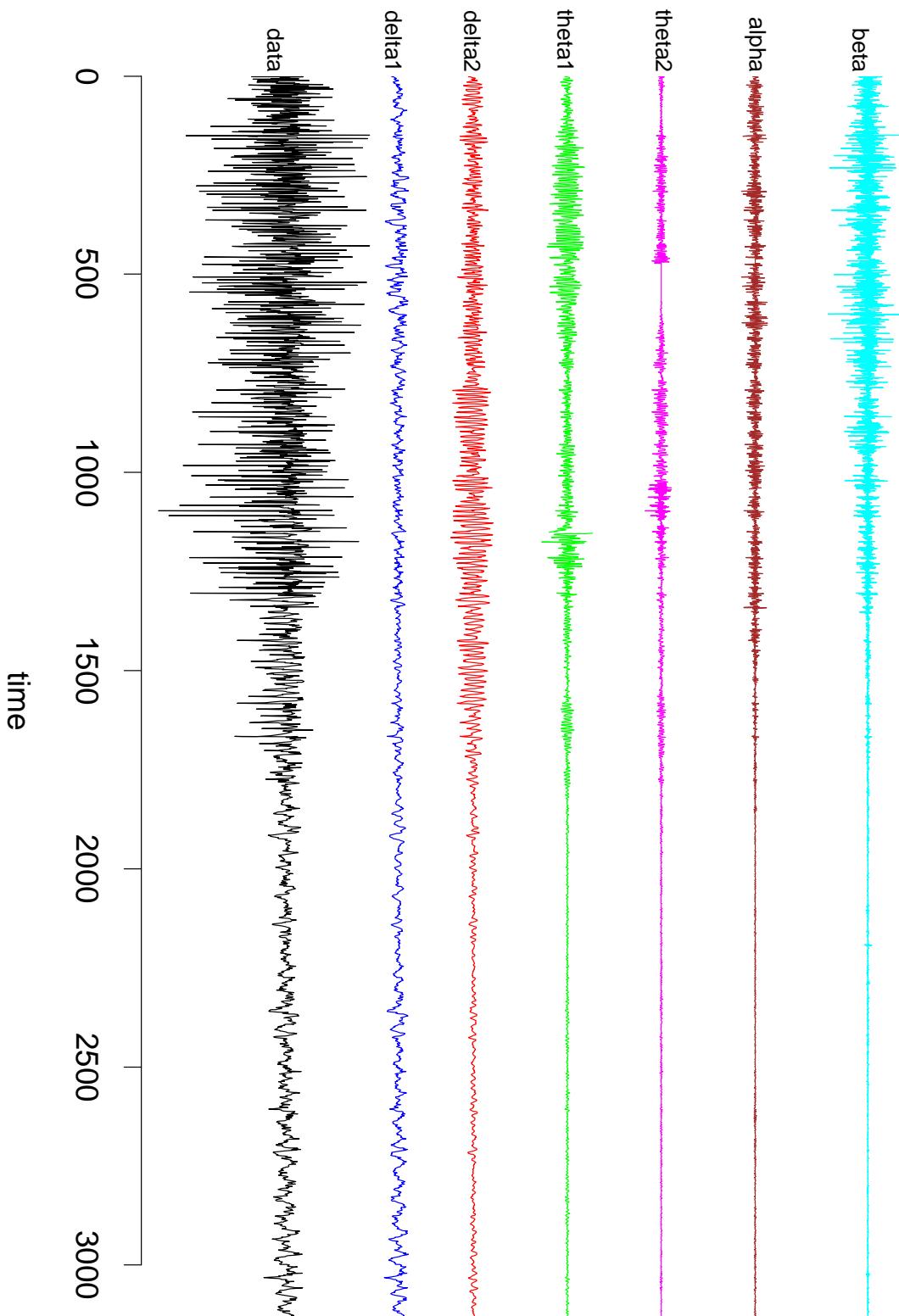


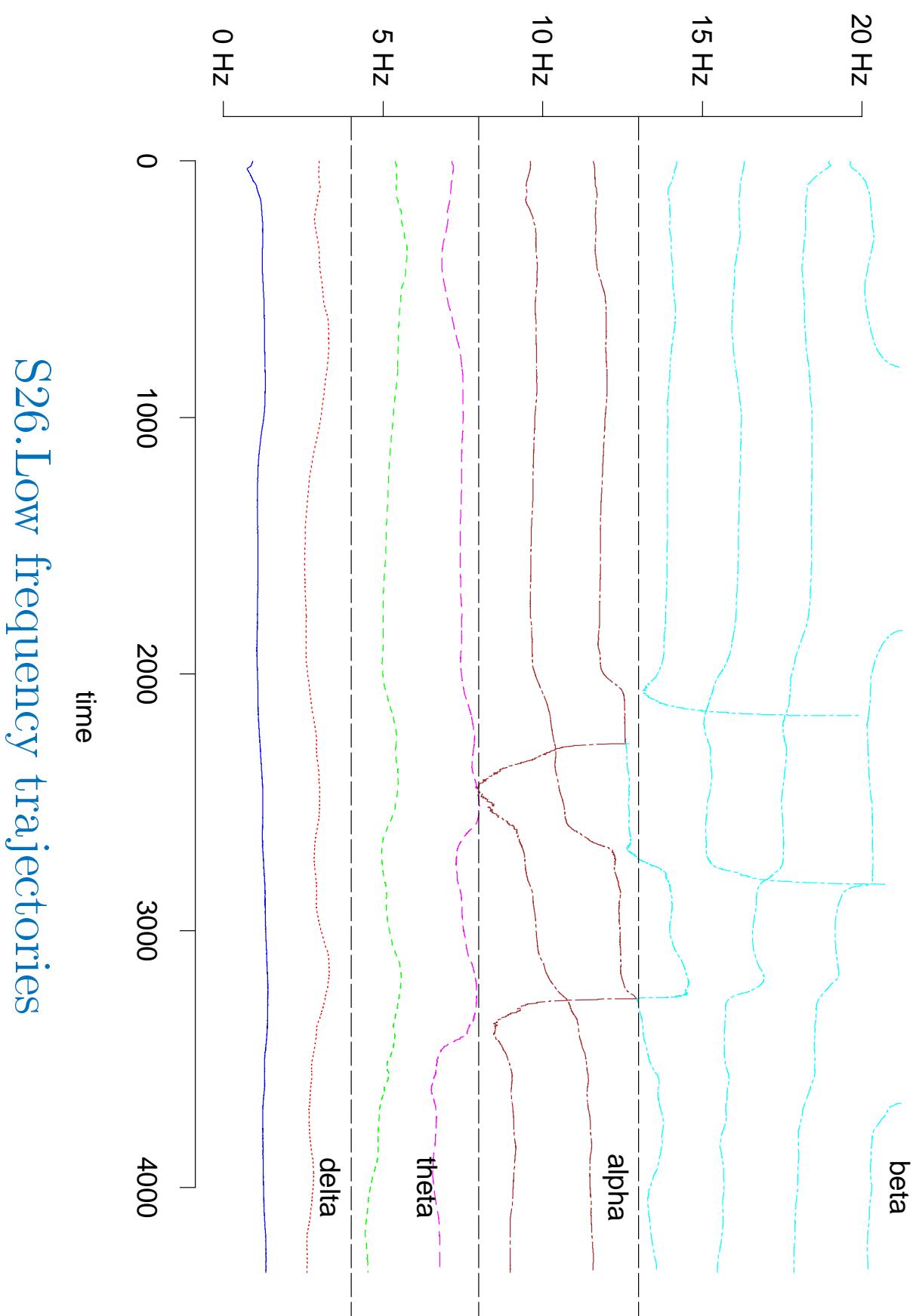
Central section of Ictal19-Cz

# Ictal19-Cz frequency trajectories

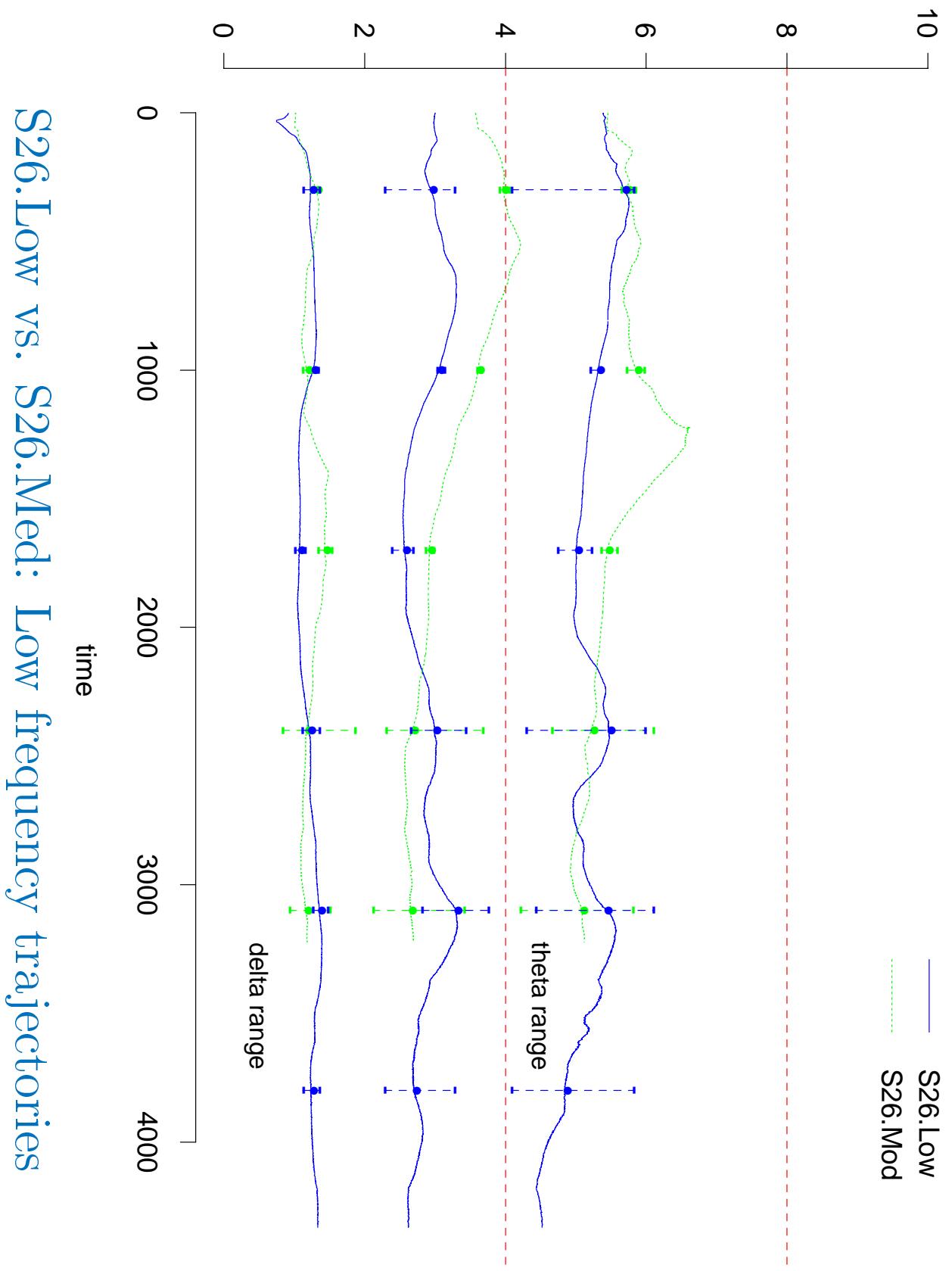


## S26. Low Patient /Treatment





S26.Low frequency trajectories



S26.Low vs. S26.Med: Low frequency trajectories

## OXYGEN DATA AND CLIMATE CHANGE

- deep ocean cores: relative abundance of  $\delta^{18}\text{O}$  to  $\delta^{16}\text{O}$
- $\delta^{18}\text{O} \downarrow$  as global temperatures  $\uparrow$  (smaller ice mass)
  - plotted with *reverse sign*: higher recent global temperatures
- periodicities: earth orbital dynamics  $\rightarrow$  impact on solar isolation – Milankovitch; Shackleton *et al* since 1976 (Shackleton and Hall 1989, Parks 1992)

*eccentricity* : 95 – 120 kyear

*obliquity* : 40 – 42 kyear

*precession* : 22 – 24 kyear

( $\pm 19$  kyear?)

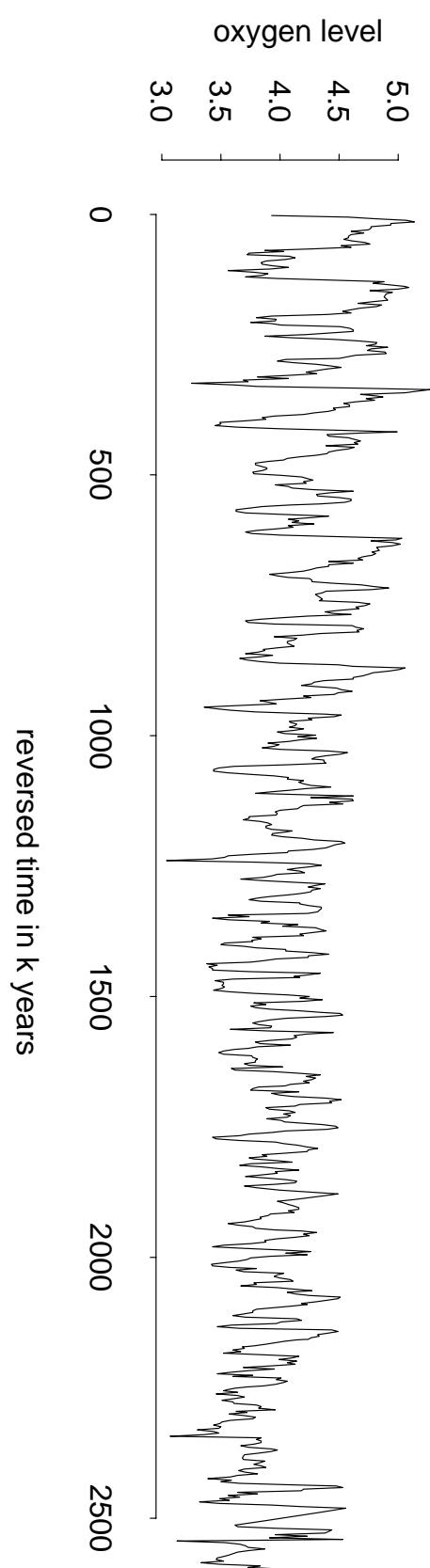
- Form of time variation in individual cycles ?
- Timing/nature of onset of “ice-age” cycle  $\leftrightarrow$  eccentricity component  $\sim 1000$  kyears ago ?
- *Time scale: errors, interpolation, ... measurement, sampling error, etc*

*Models:* High-order TVAR,  $d = 20$ , plus smooth trend

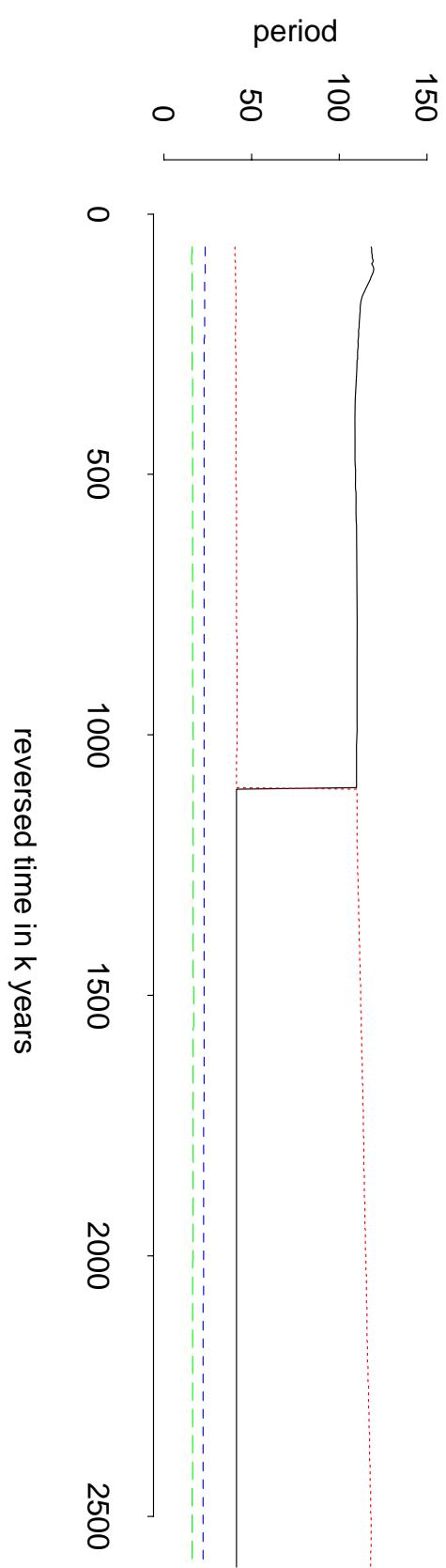
*Decomposition:*

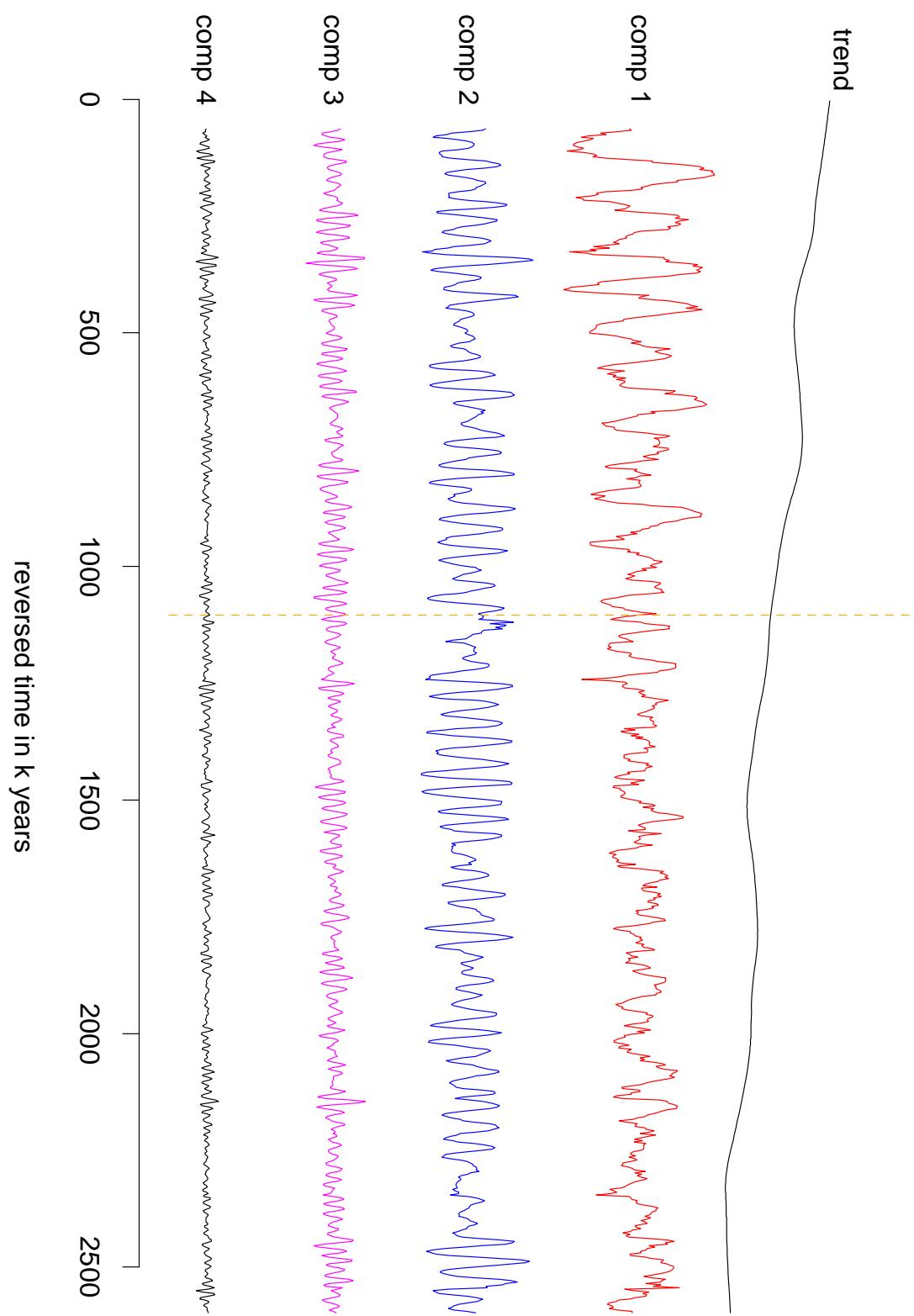
- Posterior mean of  $x_t, \phi_t$  at each  $t$
- 3 dominant quasi-periodic components:
  - *ordered by estimated “instantaneous” amplitudes*
- plus higher-frequency residual structure &/or contaminations

## oxygen isotope series



## trajectories of time-varying periods of components





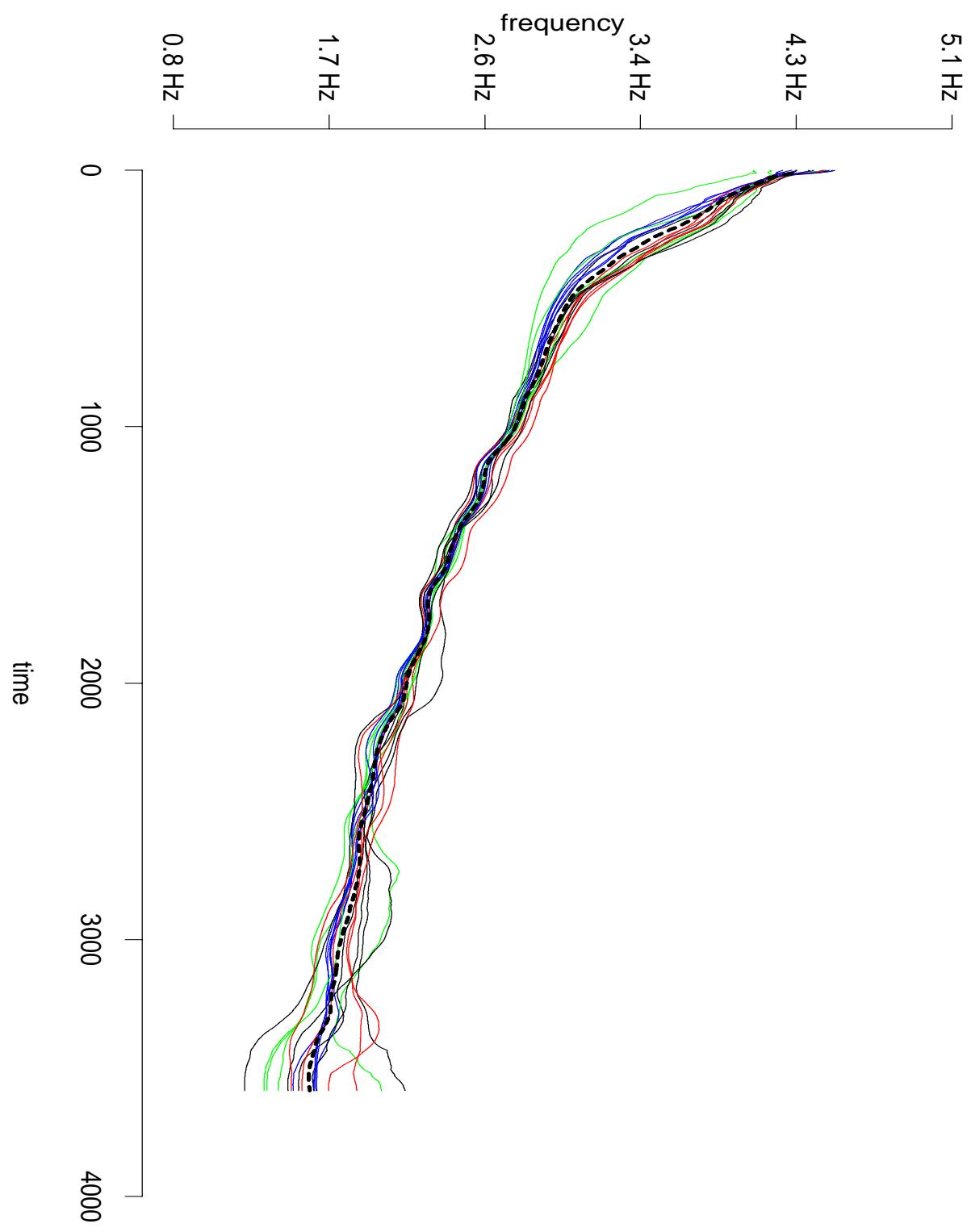
- ..... 108 – 120, peak 110
- ..... 40.8 – 41.6, peak 41.5
- ..... 22.2 – 23, peak 22.8
- .....  $\sim 19 \pm kyear$

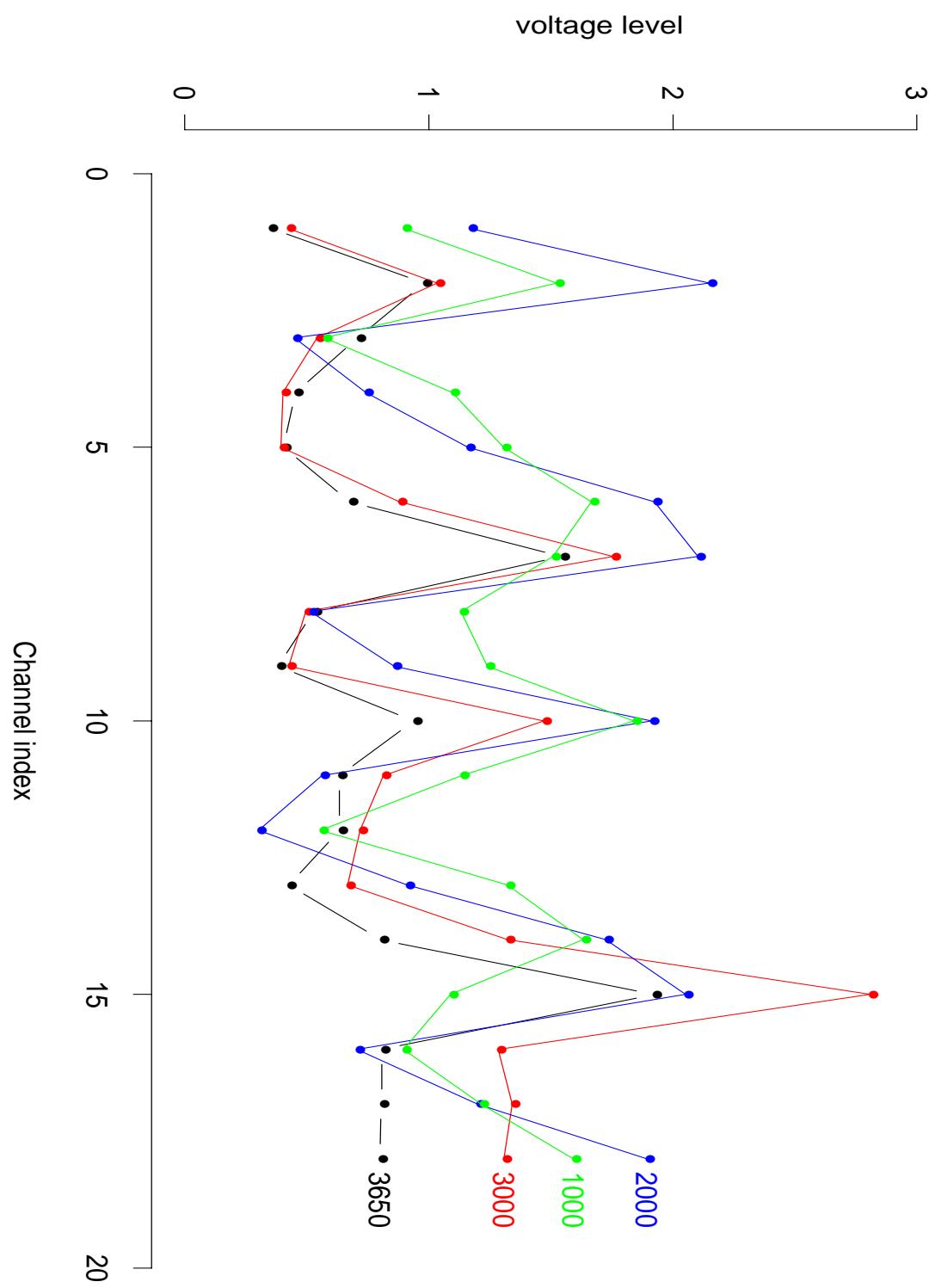
## MULTIPLE TIME SERIES

- Exploring/comparing decompositions across series
- EEG series:
  - Time variation in frequencies of delta waves
  - Amplitude associations across series (EEG channels)
- Common latent factor structure:

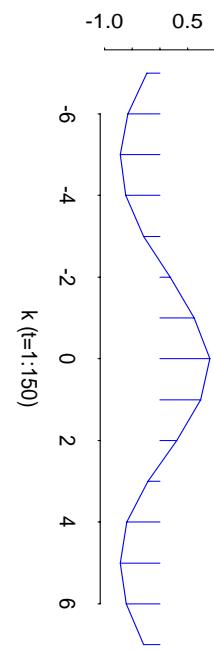
Multivariate models?

Datum at time  $t$ :  $\mathbf{x}_t$

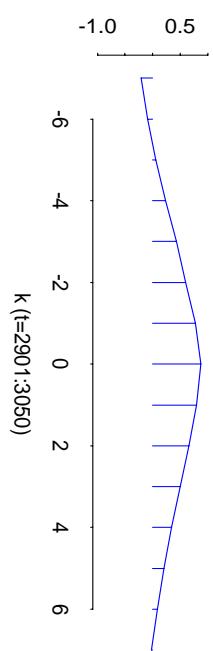
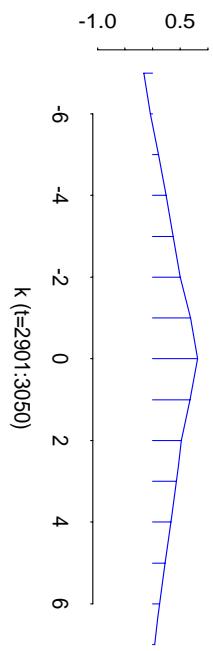
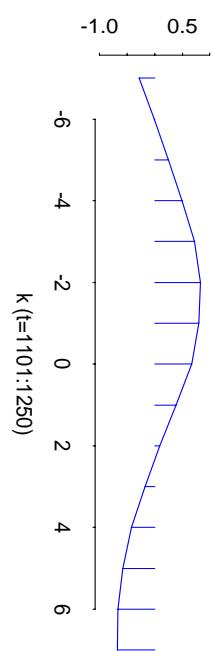
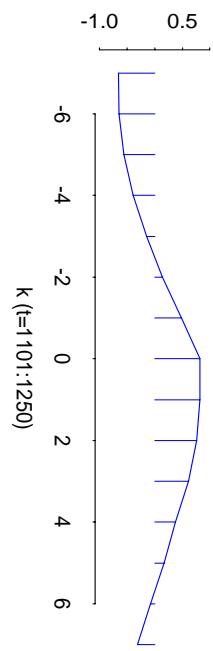
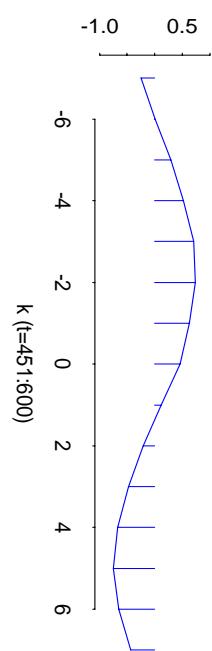
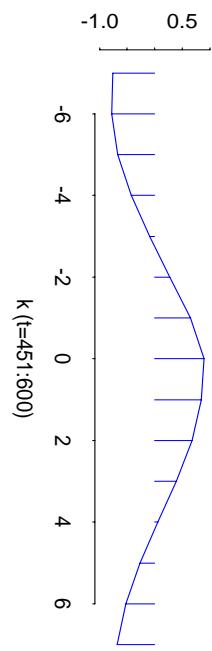
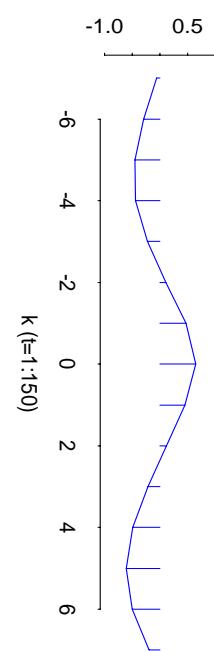




channel Fp 2



channel O 1

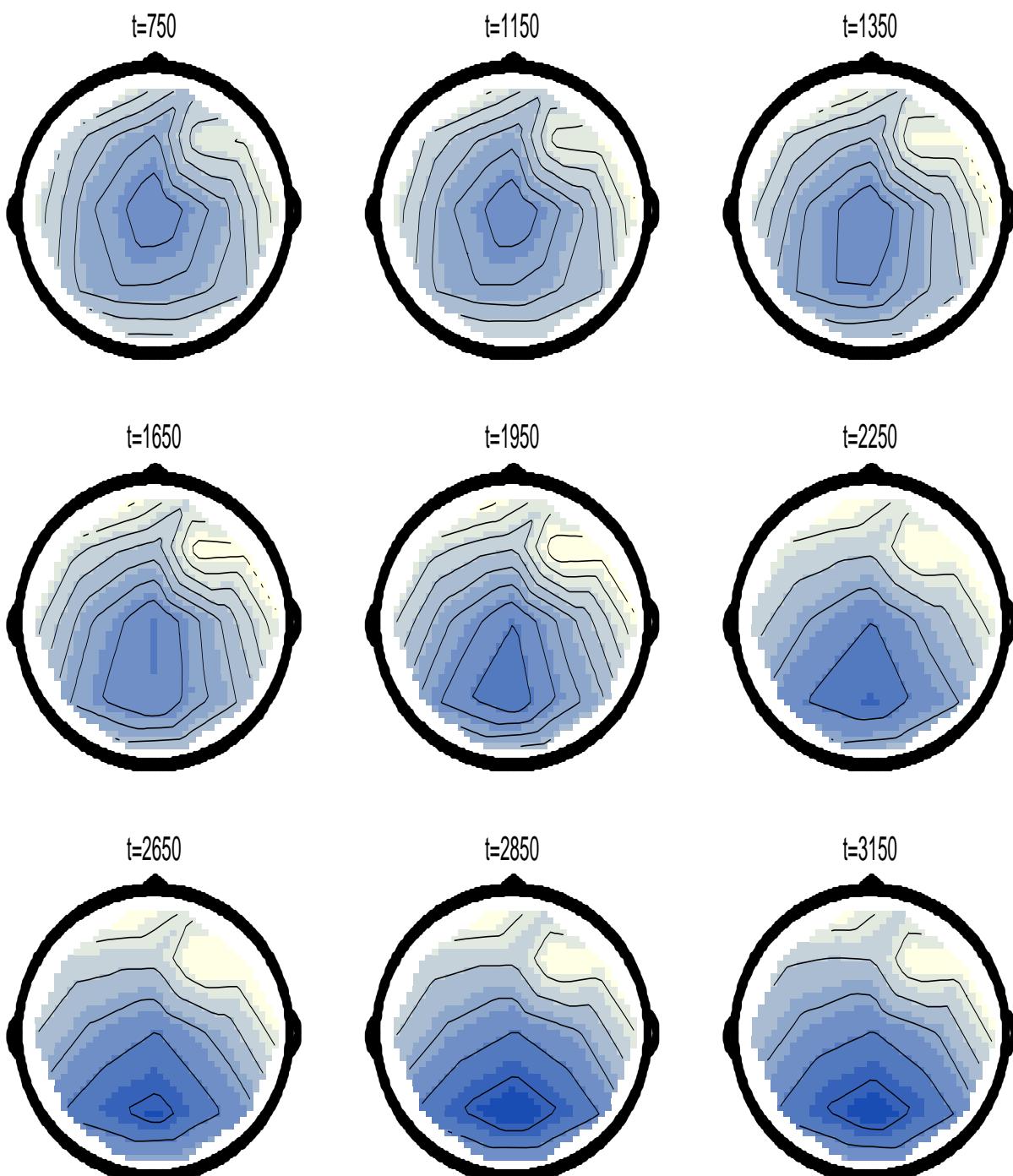


Cross-correlations with Cz at various time 'windows'

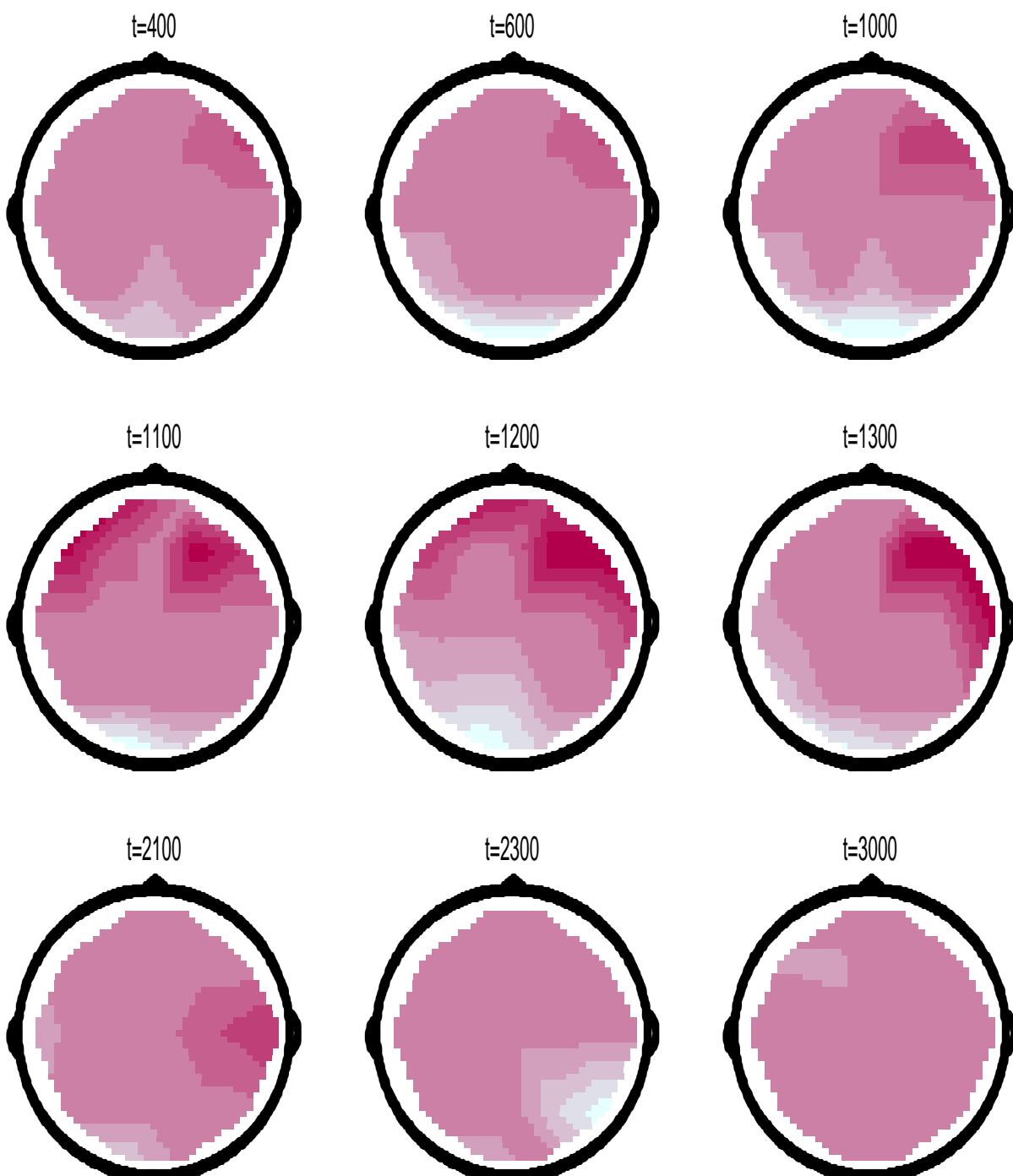
## EXPLORATORY CONDITIONAL MODELLING

Regress  $\mathbf{x}_t$  on several lagged/lead values of one series

- EEG example: regress on vertex channel Cz
- time-varying regression: time-varying lag/lead structure
- elucidates patterns of spatial dependencies and their changes over time (in lags and amplitudes)
- $$x_{i,t} = \sum_{r=-\alpha}^{\alpha} (b_{i,t} p_{i,r,t}) x_{Cz,t-r} + \nu_{i,t}$$
- $\sum_r p_{i,r,t} = 1$
- Explore patterns of “overall” regression coefficient for channel  $i$  relative to Cz:  $b_{i,t}$



0.4    0.6    0.8    1.0    1.2



-2      -1      0      1      2

## “DIRECT” FACTOR MODELLING

Multiple series:

$$\mathbf{x}_t = \mathbf{B}_t \mathbf{z}_t + \boldsymbol{\nu}_t$$

- dynamic models for latent factors  $\mathbf{z}_t$
- generalisation of Peña and Box (1987)
  - constant  $\mathbf{B}_t$  and ARMA models for  $\mathbf{z}_t$  –
- **complex framework:**
  - identification and overspecification
  - time-varying structures for factors
  - time-varying responses  $\mathbf{B}_t$
  - time-varying effects of *lagged* factors
  - ...

## MULTIVARIATE MODELS & THEORY

Vector autoregressions (VAR)

$$\mathbf{x}_t = \sum_{j=1}^d \Phi_j \mathbf{x}_{t-j} + \omega_t$$

and TVVAR models

$$\mathbf{x}_t = \sum_{j=1}^d \Phi_{t,j} \mathbf{x}_{t-j} + \omega_t$$

NEW DECOMPOSITION THEORY: (R Prado, ISDS PhD )

- elements of  $y_t$  “driven” by common factor processes
- idiosyncratic time-varying lags, phases, amplitudes ...

## CURRENT DIRECTIONS: in VAR models

- Parametric constraints in VAR models
- Constrain (many) parameters in  $\Phi_j$ 
  - reduce dimension, impose structure
- **Reduced rank** VAR models:
  - VAR arises from an underlying **factor representation**
- Theory of latent structure under such constraints?
- Model specification and fitting?
- Time-varying versions?

## SOME CURRENT PROJECTS

- Understanding latent factor structure in multiple time series models
  - ‘Reduced rank’ vector AR models; Time-varying VAR
- Nonlinear models in time-varying “linear” framework
- Continuous time models
- Structured priors in time-varying parameter models
- Methods and computation in dynamic factor models with *multivariate stochastic volatility*
  - multiple financial series:  
*Short-term forecasting & adaptive portfolio allocation*
- sequential computation

# Unit 9: Nonparametrics

# Course Units

- ▶ Introduction to Bayesian Statistics
- ▶ Prior Distributions
- ▶ Simple Models
- ▶ Hierarchical Modeling
- ▶ Bayesian Computation
- ▶ Model Assessment and Comparison
- ▶ Regression Modeling
- ▶ Clinical Trials
- ▶ Bayesian Nonparametrics
- ▶ Biostatistical Methods

# Outline of the Unit

## Overview

## Function estimation

Basis functions and splines

Gaussian processes

## Density estimation

Dirichlet processes

Mixture models

# Background

- ▶ Definitions:
  - ▶ Practical: model specifications in which the functional form of the density or regression function is not a simple parametric function of a few parameters.
    - ▶ Flexible Bayesian inference comparable to nonparametric techniques such as kernel density estimation and scatterplot smoothing.
  - ▶ Technical definition: probability models with infinitely many parameters, or models on function spaces.
- ▶ Function estimation (avoid linearity)
  - ▶ Basis functions and splines
  - ▶ Gaussian processes
- ▶ Density estimation (avoid normality)
  - ▶ Dirichlet processes
  - ▶ Mixture models
- ▶ References: Ruppert, Wand and Carroll (2003) Semiparametric Regression; Muller and Quintana (2004) Statistical Science, David Dunson (NIH/NIEHS, Duke)

# Basic model

$$y_i = f(x_i) + \epsilon_i$$

We might have  $x_i$  being

- ▶ a covariate
- ▶ a set of covariates
- ▶ time
- ▶ spatial location

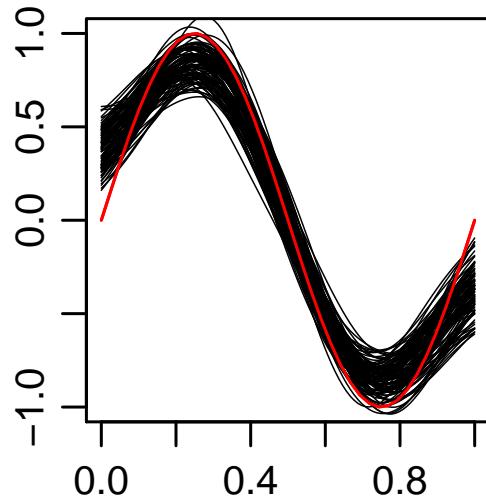
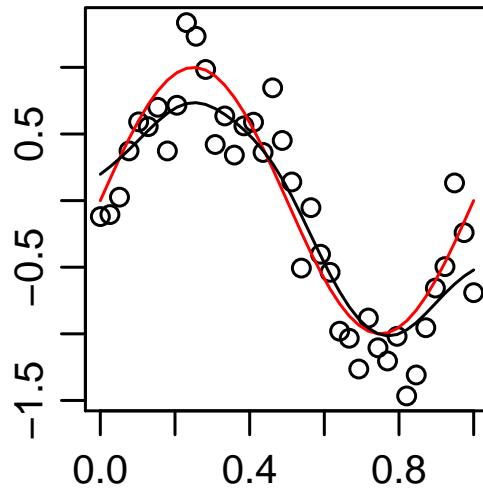
Idea is to avoid *simple* parametric assumptions or the ugliness of polynomial models.

Basic approach addresses the need to avoid overfitting:

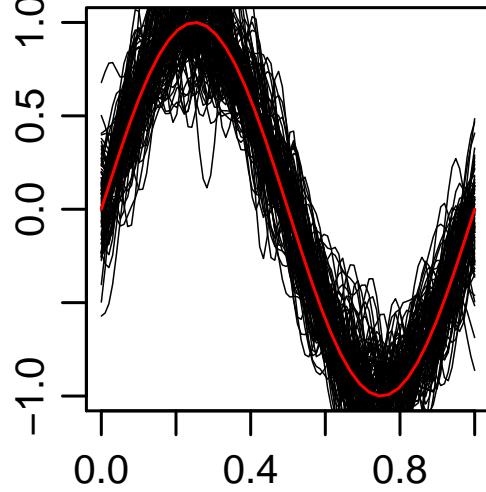
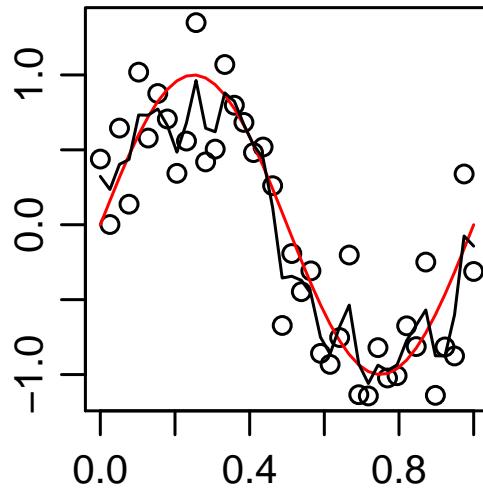
- ▶ Non-Bayesian approaches generally penalize the complexity of  $f$ , such as  $\int (f''(x))^2 dx$ , with the penalty often estimated by various types of cross-validation.
- ▶ Bayesian approaches put a prior on  $f$  that estimates complexity based on the data, usually via a variance component parameter.
  - ▶ We'll see that such priors generally introduce a natural complexity penalty (nice!) which is often not recognized.

# Bias-variance tradeoff

**oversmoothing**



**undersmoothing**



# Models for the function

- ▶ Basis functions
  - ▶ Regression splines
  - ▶ Penalized splines
- ▶ Gaussian processes
- ▶ Differencing priors

# Bases

- ▶ A basis is a set of simple functions that can be added together in a weighted fashion to form more complicated functions.
- ▶ A simple basis is the polynomial basis. Consider a polynomial regression model:

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \epsilon_i.$$

- ▶ The basis is the set of power functions from which one can construct the regression function:  
 $\{x, x^2, x^3, x^4, x^5, x^6, \dots\}$
- ▶ With a sufficiently large number of basis functions, one essentially has a nonparametric function estimator.

# Power basis

- ▶ In the regression model,

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \epsilon_i,$$

the function is represented using four basis functions:  $f_i = \sum_{j=1}^4 B_j(x_i) b_j$ , where  $B_j(x) = x^j$ , and the weights are the coefficients,  $b_j$ .

- ▶ We can express the function for all the observations jointly as  $f = Bb$ , where the matrix  $B$  contains the basis functions,  $B_j(x_i)$ , evaluated for each of the observations, e.g. for 4 coefficients:

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} 1 & B_1(x_1) & B_2(x_1) & B_3(x_1) \\ 1 & B_1(x_2) & B_2(x_2) & B_3(x_2) \\ 1 & \vdots & \vdots & \vdots \\ 1 & B_1(x_n) & B_2(x_n) & B_3(x_n) \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

- ▶ The polynomial basis has some problems. In particular, it is unstable at the boundaries and can swing wildly, because the estimate of each coefficient is influenced by all the data.
- ▶ Instead we'll consider two types of spline models: regression splines and penalized splines.

# Regression splines

The simplest spline basis is the linear spline basis:

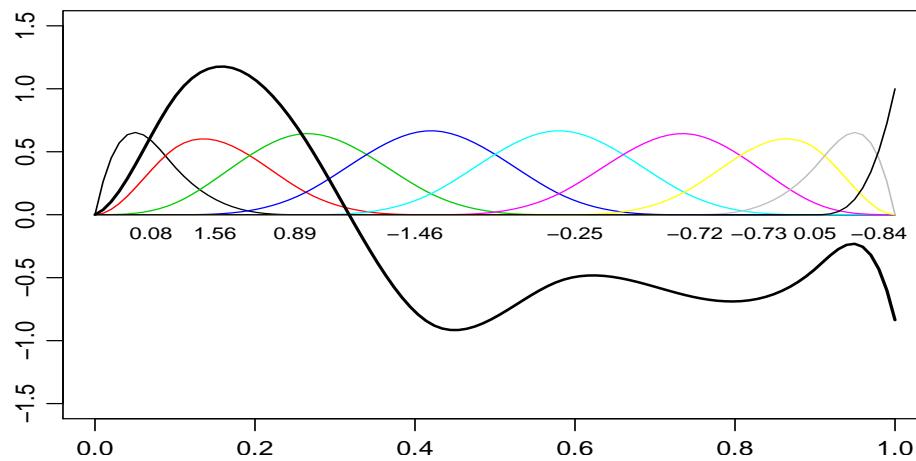
$$f(x) = \beta_0 + \beta_1 x + \sum_k b_k (x - \kappa_k)_+$$

The  $\kappa_k$  are knots.

A basis that gives smoother functions is the cubic spline basis

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_k b_k (x - \kappa_k)_+^3$$

A related basis is the b-spline basis, which doesn't have as simple a form but has some nice numerical features:



- ▶ One can think of knot number and location as a model selection problem and use strategies such as BIC, BF, and RJMCMC. (see DiMatteo et al. 2001 *Biometrika*)

# Penalized regression splines

- ▶ Instead of having to do model selection, choose a sufficiently large number of basis functions to approximate a function of more complexity than you expect and penalize complexity.
- ▶ The result is a flexible model, computationally-feasible, that avoids some issues with knot placement and the number of knots.
- ▶ The non-Bayesian optimization problem is to minimize

$$\sum_i (y_i - B_i^T b)^2 + \lambda b^T S b$$

where the matrix,  $S$ , is constructed using the spline basis chosen, and  $B$  is the basis matrix, as previously. Here the penalty makes sure the coefficients don't get too big, so the function can't swing too wildly.

# Mixed model formulation

- ▶ As a related alternative, David Ruppert, Matt Wand, and Ray Carroll have constructed penalized regression splines in a mixed model form

$$y_i = f(x_i) + \epsilon_i = x_i^T \beta + B_i^T b + \epsilon_i$$

where  $x_i^T \beta$  is a linear term and represents large-scale trends and  $B_i^T b$  represents smaller-scale features.

- ▶ [show plots of knots and fit to LIDAR in unit9code.r]

# Priors for the coefficients

Two schools of thought:

- ▶ Ruppert, Wand, Carroll
  - ▶ independence priors: rely on the basis matrix to do the smoothing:  $b \sim \mathcal{N}(0, \sigma_b^2 I)$
- ▶ Eilers & Marx
  - ▶ for one-dimension, use priors that penalize differences between adjacent coefficients
  - ▶ in two dimensions, use other priors that encourage nearby coefficients to be similar
- ▶ The Eilers & Marx approach with respect to the knots is similar to the ice example in BUGS where the differencing is done directly on coefficients for different age groups.
  - ▶ Second differences (corresponding to second derivatives) are penalized via a prior.
  - ▶ This corresponds to  $b \sim \mathcal{N}(0, \sigma_b^2 D)$  for a choice of  $D$  appropriate for the differencing.
- ▶ Something of an open question in whether there are any 'fixed effects' (apart from an intercept) or all coefficients are penalized.

# Natural complexity penalty: Take 1

- ▶ Suppose the data can be fit reasonably well via a very smooth regression function, corresponding to  $\sigma_b^2$  small. Let's compare

$$\log P(y|\theta) + \log P(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (y_i - (Bb)_i)^2 - \frac{K}{2} \log \sigma_b^2 - \frac{1}{2\sigma_b^2} \sum b_k^2$$

for a fitted model with small and large values of  $\sigma_b^2$ .

- ▶ Note: This is not so different from a LMM in which the random effects don't get estimated to be too large because of the normalization constant involving the random effects variance component. The model favors variability being attributed to effects (i.e., what we think of as fixed effects) in the model that are not constrained by their priors.

# Definition of GPs

- ▶ A linear model with a prior over  $\beta$  defines a prior over regression functions. Nonlinear functions have a prior weight of 0.
- ▶ The basis function approach extends this; the prior on coefficients defines a prior over functions.
- ▶ We can instead think of putting a prior on the regression function

$$f(\cdot) \sim \mathcal{H}(\theta)$$

- ▶ A common prior is a Gaussian process prior, in which the functions are Gaussian processes, namely

$$f(\cdot) \sim \mathcal{GP}(\mu(\cdot), C(\cdot, \cdot))$$

- ▶  $\mu(\cdot)$  is a mean function and  $C(\cdot, \cdot)$  is a covariance function that defines the covariance between the function evaluated at any pair of values.
- ▶ For a discrete set of covariate values,  $x$ , this says that  $f$  is MVN

$$f \sim \mathcal{N}(\mu_1, C)$$

where  $C_{ij} = C(x_i, x_j)$ .

- ▶ Thus the prior over functions reduces to a MVN prior over a vector.

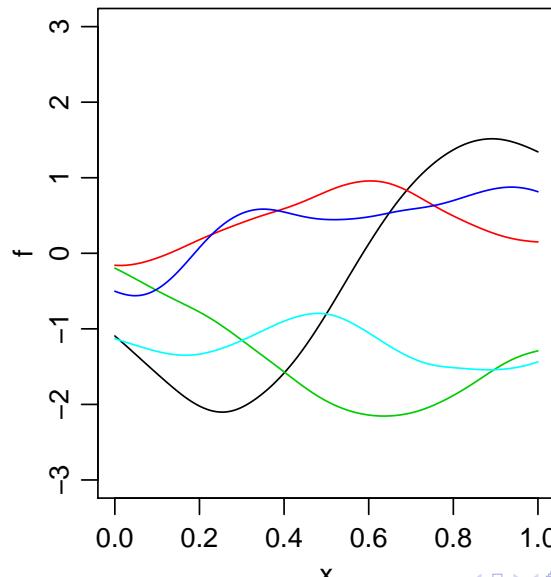
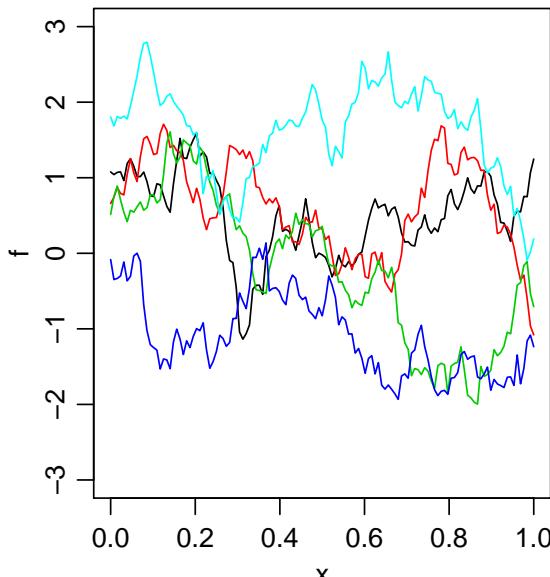
# Sample realizations

- ▶ In general, we use a simple parametric form for  $\mu(\cdot)$  and a covariance function that gives decreasing covariance with increasing distance between the covariate values.
  - ▶ Suppose we take  $\mu(\cdot) = m$  and  $C(x_i, x_j) = \sigma^2 \exp\left(-\frac{|x_i - x_j|}{\rho}\right)$ , the exponential covariance.
  - ▶ We can draw a discretized sample realization (a random function from the prior over the functions) as

$$f = m + Lu$$

$$C = LL^T$$

$$u \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$$



# Computation

Computation is straightforward:

- ▶ Define a model with likelihood a function of  $f(x_i)$ , e.g.,  $y \sim \mathcal{N}(\mu 1 + f, \sigma^2)$
- ▶ Place your prior on  $f(x_i)$  evaluated at the observed covariate values as a MVN, e.g.,  $f \sim \mathcal{N}(0, \tau^2 R(\rho))$
- ▶ Place priors on the GP hyperparameters:  $\tau^2$  and  $\rho$ .
- ▶ Proceed.
- ▶ To evaluate the function at unobserved covariate values (for prediction or plotting posterior samples of the function on a fine grid), use the conditional normal calculations.
  - ▶ Let  $f_1$  be the function at the observed covariate values and  $f_2$  at the prediction covariate values,  $C_{11}$  be the covariance matrix for the pairs of observed covariates,  $C_{22}$  for the pairs of prediction covariate values, and  $C_{21}$  the covariance between prediction and observed covariates, then

$$f_2|f_1, y \sim \mathcal{N}(\mu 1 + C_{21} C_{11}^{-1} (f_1 - \mu 1), V)$$

$$V = C_{22} - C_{21} C_{11}^{-1} C_{21}^T$$

# Structure in the mean or structure in the covariance?

- ▶ In random effects models, we saw that we could express clustering through a random effect for each cluster or through a particular covariance structure:
  - ▶ Mean model:

$$\begin{aligned}y_{ij} &\sim \mathcal{N}(\mu + b_j, \sigma^2) \\b_j &\sim \mathcal{N}(0, \tau^2)\end{aligned}$$

- ▶ Covariance model:

$$y_{ij} \sim \mathcal{N}(\mu 1, \Sigma)$$

where  $\Sigma_{kk} = \sigma^2 + \tau^2$ , off-diagonals for observations within a cluster are  $\tau^2$  and for observations not in the same cluster are 0.

- ▶ Gaussian process model

- ▶ Mean model:

$$\begin{aligned}y_i &\sim \mathcal{N}(\mu + f(x_i), \sigma^2) \\f &\sim \mathcal{N}(0, \tau^2 R(\theta))\end{aligned}$$

- ▶ Covariance model:

$$y \sim \mathcal{N}(\mu 1, \sigma^2 I + \tau^2 R(\theta))$$

# Choice of covariance?

- ▶ AR(1) models are essentially equivalent to Gaussian processes on continuous domains with an exponential covariance.
- ▶ Note that exponential covariances produce continuous but not differentiable functions.
- ▶ There are other covariances, e.g., the Matern common in spatial modeling, that produce more smooth (i.e., differentiable) functions.
- ▶ Spline and other basis models with normal priors on the coefficients corresponds to Gaussian processes with funny covariance functions:

$$f = Bb \sim \mathcal{N}(0, B\text{Var}(b)B^T)$$

# Natural complexity penalty: Take 2

- ▶ Now consider a GP model with small variance and large range (smooth functions) compared to large variance and small range (wiggly functions).
- ▶ Suppose the data can be fit reasonably well via a very smooth function. Let's compare

$$\log P(y|\theta) + \log P(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (y_i - f(x_i))^2 - \frac{1}{2} \log |\Sigma| - \frac{1}{2} f^T \Sigma^{-1} f$$

for a fitted model with  $\Sigma$  having large diagonal and off-diagonal elements and small ones.

- ▶ Large diagonal elements reduce the log posterior (think of a diagonal matrix). Large off-diagonal elements also reduce the log posterior [go back to 2x2 example]
- ▶ Once again, complexity is only favored to the extent that the likelihood increases in a substantial way.

# Bayesian penalty on complexity

- ▶ We've seen with BF that comparing a model with a sharp prior and one with a vague prior, the BF favors the simpler model (sharp prior) because the complicated one spreads prior mass over parameter values that give poor values of the likelihood.
- ▶ Similarly, a Bayesian nonparametric model naturally favors a simple model if it sufficiently well explains the data compared to a more complicated model. A complicated model spreads its predictive mass  $P(y|M)$  over a larger space of possible data than does a simple model. So if the data can be fit well by the simple model, the complicated model is penalized by the poor fit for many of the possible functions under the prior (the prior distribution over possible functions - the complicated model puts less mass on simple functions).
- ▶ As with Gelman's reasoning about model comparison in general, I favor defining flexible models that can give simple or complicated functions depending on a key parameter (e.g.,  $\sigma_b^2$ ) without explicit model choice.
  - ▶ Penalized splines rather than knot selection with regression splines is one example.

# Overview

Why do we need nonparametric density models?

- ▶ The main idea here is that densities in a model may not follow simple parametric forms.
- ▶ Examples include
  - ▶ complicated residual distributions:
    - ▶ suppose the distribution of the residuals is not normal
    - ▶ or perhaps the distribution is a function of covariates
  - ▶ random effects distributions
    - ▶ latent variables might have long tails or exhibit bimodality because of some sort of clustering
- ▶ Approaches:
  - ▶ Dirichlet process priors
  - ▶ Mixture distributions

# Distributions over distributions

- ▶ When we say the residuals or random effects are distributed  $\mathcal{N}(0, \tau^2)$ , we are putting a prior over distribution functions. This prior happens to put all its mass on normal distributions, with different variances.
  - ▶ But recognize that the prior over  $\tau^2$  induces a prior over distribution functions.
- ▶ Instead, we can relax the parametric assumption and say  $\phi_i \sim G$  where  $G$  is an unspecified distribution.
  - ▶ So we need a distribution over distributions!
  - ▶ We can say  $G \sim \mathcal{G}(\theta)$  where  $\mathcal{G}$  is a prior over distribution functions, parameterized by  $\theta$ .
  - ▶ So now what do we do for  $\mathcal{G}$ ?

# Dirichlet Process Model

- ▶ The basic DP model is

$$G \sim \mathcal{DP}(\alpha, G_0)$$

- ▶  $G_0$  is the 'base measure', usually a parametric distribution.
- ▶  $\alpha$  is a precision parameter that says how far away  $G$  is from  $G_0$ . For large  $\alpha$ ,  $G \approx G_0$ .
- ▶  $G$  is a discrete distribution, which is not always desirable. It is an infinite mixture of point masses.
- ▶ A simple choice is  $G_0 = \mathcal{N}(\mu, \tau^2)$ .

# Sampling from the prior

- ▶ The model for values drawn from  $G$  is

$$\begin{aligned}\phi_i &\stackrel{\text{iid}}{\sim} G \\ G &\sim \mathcal{DP}(\alpha, G_0)\end{aligned}$$

- ▶ Since realizations of  $G$  are actually discrete distributions on countably infinite support points, it difficult to draw from  $\mathcal{DP}(\alpha, G_0)$ .
- ▶ However, we can draw  $\phi$  in a way that marginalizes over the uncertainty in  $G$ :
  - ▶ Draw  $\phi_1$  from  $G_0$ .
  - ▶ Next draw  $\phi_i$  with probability  $\frac{1}{i-1+\alpha}$  as a random sample from  $\phi_1, \dots, \phi_{i-1}$  and with probability  $\frac{\alpha}{i-1+\alpha}$  from  $G_0$ .
- ▶ Note that we induce clusters and that the larger the clusters get, the more likely we are to add another value to the cluster, but that the locations of the clusters is random.
- ▶ The unique values of  $\phi$  are called 'atoms'.
- ▶ When we condition on data, the locations of the clusters will depend on the data.
- ▶ Note that the discreteness becomes clear when we think about letting  $i \rightarrow \infty$ . We get infinite mixtures of point masses.

# The Chinese Restaurant Process

- ▶ Suppose you have a set of tables, some occupied and some not. The next customer comes in and either sits at an occupied table, with probability proportional to the number of people already at the table or chooses a new table. The value of  $\phi$  assigned to a person is the same for everyone at a given table.
  - ▶ This generates samples of the random variables with the DP distribution, marginalizing over  $G$ .
  - ▶ The full DP results from having an infinite number of tables and infinite number of customers, and each table able to accomodate an infinite number of people!
  - ▶ [I believe computer scientists came up with this analogy...]

# Why 'Dirichlet'?

## Formal definition:

A DP distribution is one for which for any partition,  $A_1, \dots, A_k$  of the sample space and any  $k$ , the random probabilities follow the distribution

$$G(A_1), \dots, G(A_k) \sim \mathcal{D}(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$$

- ▶ Basically, the probabilities of events are centered around the probabilities under the base measure, but with some variability, whose magnitude is determined by  $\alpha$ .
- ▶ So we have overdispersion; marginally, the draws from  $P(\phi_i | G_0, \alpha) = \int P(\phi_i | G)P(G | G_0, \alpha) dG$  are overdispersed compared to  $\phi_i \sim G_0$ .
- ▶ Useful exercise to work out marginal mean and variance of  $\phi_i$ , marginalizing over  $G$ . How?

# Simple density estimation

Suppose we have data,  $y_1, \dots, y_n$ , and we assume the likelihood  $y_i \stackrel{\text{iid}}{\sim} G$  with DP prior,  $G \sim \mathcal{DP}(\alpha, G_0)$ . Then let  $\delta_x(\cdot)$  be a point mass at  $x$ . The posterior for  $G$  is

$$G|y \sim \mathcal{DP}(\alpha + n, G_1)$$

where the new base measure is  $G_1 \propto G_0 + \sum_i \delta_{y_i}$ . What happens as one gets more and more data?

# Stick-breaking process

- ▶ We can express  $G \sim \mathcal{DP}(\alpha, G_0)$  in another way:  
$$G = \sum_k w_h \delta_{\theta_h}$$
 where  $\theta_h \stackrel{\text{iid}}{\sim} G_0$  and  
$$w_h = u_h \prod_{j < h} (1 - u_j)$$
 where  $u_h \sim \text{Beta}(1, \alpha)$ .
- ▶ The components here are the weights and the atom values.
- ▶ Weights: The  $w_h$  behave as if you take a stick, break it, and then continue by next breaking the smaller of the two pieces that result.
- ▶ Atom values: The atoms come from the base measure.

# MCMC

- ▶ One of the key reasons for the use of the DP approach is that it allows for Gibbs sampling of the  $\phi$ 's.
- ▶ Basically the standard MCMC integrates over the random distribution  $G$ , so we cannot make inference about  $G$  (recall that it is an infinite set of point masses), only about  $\phi_i$ ,  $\alpha$ , and any parameters in the base measure.
- ▶ To sample from  $P(\phi_i|\phi_{-i}, y, \alpha, G_0)$  we have the conditional prior and the likelihood.

$$\phi_i|\phi_{-i}, \alpha, G_0 \sim \frac{\alpha}{\alpha + n - 1} G_0 + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\phi_j}$$

$$y|\phi_i \sim F(\phi_i)$$

So if  $G_0$  is conjugate for  $F$  then we would have a mixture distribution that is easy to sample from after we work out the conditional posterior distribution. As in the exam, you'd have to work out a normalizing constant to get the weights on the different pieces of the mixture.

- ▶ Full MCMC would also sample  $\alpha$  and parameters in  $G_0$ .

# Discreteness and the DP

- ▶ The primary drawback of the DP model is that the random distributions are discrete.
- ▶  $G \sim \mathcal{DP}(\alpha, G_0)$  gives discrete realizations  $G$ .
- ▶ This means that  $\phi_i|G$  are drawn from a discrete distribution.
- ▶ However, marginally, the  $\phi$ 's are drawn from an overdispersed  $G_0$  which is continuous. (I think!)
- ▶ Tricky to get one's head around.

# Simple mixture models

We can define a flexible continuous density as a mixture of simple parametric densities

$$G(\phi) = \sum_{k=1}^K \pi_k P_k(\phi)$$

or more simply using the same functional form for each component, but with different parameters:

$$G(\phi) = \sum_{k=1}^K \pi_k P(\phi|\theta_k)$$

What is a natural prior for  $\pi$ ? Remember that  $\pi_k \in (0, 1)$  and  $\sum_k \pi_k = 1$ .

Computation is often done by introducing additional parameters: a membership indicator for each observation from the mixture density.

Be careful about label switching: one may want to impose constraints such as nondecreasing means for the components.

# Dirichlet mixture

$$\begin{aligned}\phi_i &\sim P(\phi|\theta_{c_i}) \\ c &\sim \text{Mult}(\pi_1, \dots, \pi_K) \\ \theta_i &\sim G_0 \\ \pi &\sim \mathcal{D}(\alpha_1, \dots, \alpha_K)\end{aligned}$$

Choices: might take  $G_0 = \mathcal{N}(\mu, \tau^2)$  and  $\alpha_i = \alpha_0 \frac{1}{K}$ .

Example:  $P$  could be a normal - in which case we'd need  $\theta_i = (m_i, \sigma_i^2)$ . This would require  $G_0$  to be a bivariate distribution, e.g., normal-inverse-gamma.

# Number of components?

We've already seen some examples of mixture models in which we integrate over a continuous range of mixing parameters (analogous to the  $\pi$ 's above)

- ▶  $t$  as scale mixture of normals
- ▶ beta-binomial as mixture of binomials, etc.

Concerns with our discrete mixture: how do we choose the number of components?

- ▶ model selection techniques
- ▶ RJMCMC
- ▶ DP mixture model

# DP Mixtures (DPM)

- ▶ Recall: A DP distribution is a distribution over discrete distributions. Use the DP prior for the parameters of parametric components

$$\begin{aligned} G(\phi) &= \int P(\phi|\theta)dG(\theta) \\ &= \sum_i P(\phi|\theta_i)Pr(\theta_i) \end{aligned}$$

with  $G \sim \mathcal{DP}(\alpha, G_0)$ .

- ▶ This extends a Dirichlet prior over  $\pi$  to allow for a random number of components where we estimate the number of components based on the data.
  - ▶ Recall the natural complexity parameter - we hope that this will help us here as well.
- ▶ Computation: If  $P$  and  $G_0$  are a conjugate pair, one can work out a Gibbs sampler for  $[\theta_i | \theta_{-i}, \text{rest, data}]$
- ▶ Result is that we can embed the mixture model with a random number of components naturally within a single model rather than as a model selection or RJMCMC problem.