extra notes (15 Feb) ①

the big picture

ingredients in a problem involving uncertainty:

① θ, something unknown & of interest to you (think of a vector of real numbers of length k)

② D, a dataset you judge relevant to decreasing your uncertainty about θ (think of a vector of real numbers of length n)

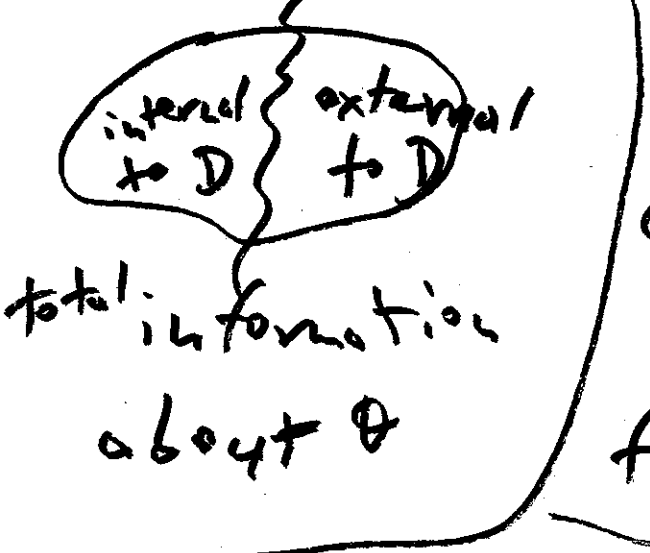
③ B, a (finite) collection of true-false propositions, all regarded by you as true, (B_1 and B_2 and ... and B_ℓ) capturing relevant background info.

probability (RT Cox) | $P(B|A)$ | T/F propositions

quantifies your information about the truth of B given that you regard A as T

Can extend this

to cumulative distribution functions (CDFs) and densities for quantifying uncertainty about quantities θ living continuously on the real line, by considering $F_{(\theta|A)}(q) = P(\theta \le q | A)$ (CDF) and $\frac{d}{dq} F_{(\theta|A)}(q) = p_{(\theta|A)}(q)$ (density); in small abuse of notation people write $F(\theta|A)$ (CDF) and $p(\theta|A)$ (density)



you have to specify 4 ingredients: ③

④ $p(\theta|B)$, your info. about θ external to D (your prior dist. for θ)

④ $p(D|\theta B)$, your info. about θ internal to D (your sampling dist. for θ)

these 2 are enough for inference & prediction

for decision, need to specify 2 more ingredients:

④ a = your action space of possible choices

④ $U(a, \theta)$, your utility function quantifying costs & benefits if you choose a & unknown = θ

call $\{p(\theta|B), p(D|\theta B), a, U(a,\theta)\}$ your model for your uncertainty about θ (given $D \& B$)

(usually you're uncertain about θ but you're also uncertain about how to specify your uncertainty about θ through the model: call this second type of uncertainty model uncertainty) ← (we have so far pretended that this type of uncertainty does not exist)

payoff: with these 4 ingredients

specified, the optimal method for yielding logically-internally-consistent answers is as follows:

① (inference) (a) write $\ell(\theta|DB) = c\, p(D|\theta B)$ (for arbitrary positive c) (b) compute:

$$p(\theta|DB) = c\, p(\theta|B)\, \ell(\theta|DB)$$

posterior dist. (full-info. dist)

② (prediction) future data D^*

$$p(D^*|DB) = \int_{\text{set of possible values of } \theta} p(D^*|\theta B)\, p(\theta|DB)\, d\theta$$

③ (decision) find a^* that maximizes

$E\ U(a, \theta)$ (expected utility)

$(\theta | DB)$ " $\int U(a, \theta)\, p(\theta | DB)\, d\theta$

④ challenges:

Ⅰ the specification burden

Ⅱ computation: $p(\theta | DB) = c\, p(\theta | B)\, \ell(\theta | DB)$

$\theta = (\theta_1, \ldots, \theta_k)$

each of Bayes is a k-dimensional probability dist.

hard to visualize & work with directly

so: (A) $p(\theta_1 | DB) = \int \cdots \int p(\theta_1, \theta_2, \ldots, \theta_k | DB)\, d\theta_2 \cdots d\theta_k$

$(k-1)$-d.m. integral

marginal post. dist. for θ_1 given DB

& there are (k) of these (k-1) dim. integrals (7)

(8) (prediction) $p(D^*|DB) = \int p(D^*|\theta, DB)\, p(\theta|DB)\, d\theta$

this is another k-dimensional integral (9) (inference)

the normalizing constant c in $p(\theta|DB) = c\, p(\theta|B)\, p(D|\theta B)$

is another k-dimensional integral (10) the expected

utility $\int u(a,\theta)\, p(\theta|DB)\, d\theta$ is another

k-dimensional integral, and if a is large you'll

have to compute or approximate this integral

a lot of times (ex. variable selection in regression.)

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i$ ← IID $N(0, \sigma^2)$ (8)

$\in \mathbb{R}$

which subset of $(x_1, \ldots, x_m)$ is "best" for predicting y?

there are 2^m subsets, so a lot 2^m actions to compare (ex: $m = 100$; $2^m = 10^{30}$ actions to choose among)

when k is small, sometimes all these problems go away, if you can find a conjugate prior to the sampling dist. in your problem ex. $\ell(\theta | \underline{y}) = \theta^s (1-\theta)^{n-s}$, $s = \sum_{i=1}^{n} y_i$

AMI (10) $\underline{y} = (y_1, \ldots, y_n)$

$p(\theta|B) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\alpha,\beta>0$ (R.v. θ family of Beta densities) has property that

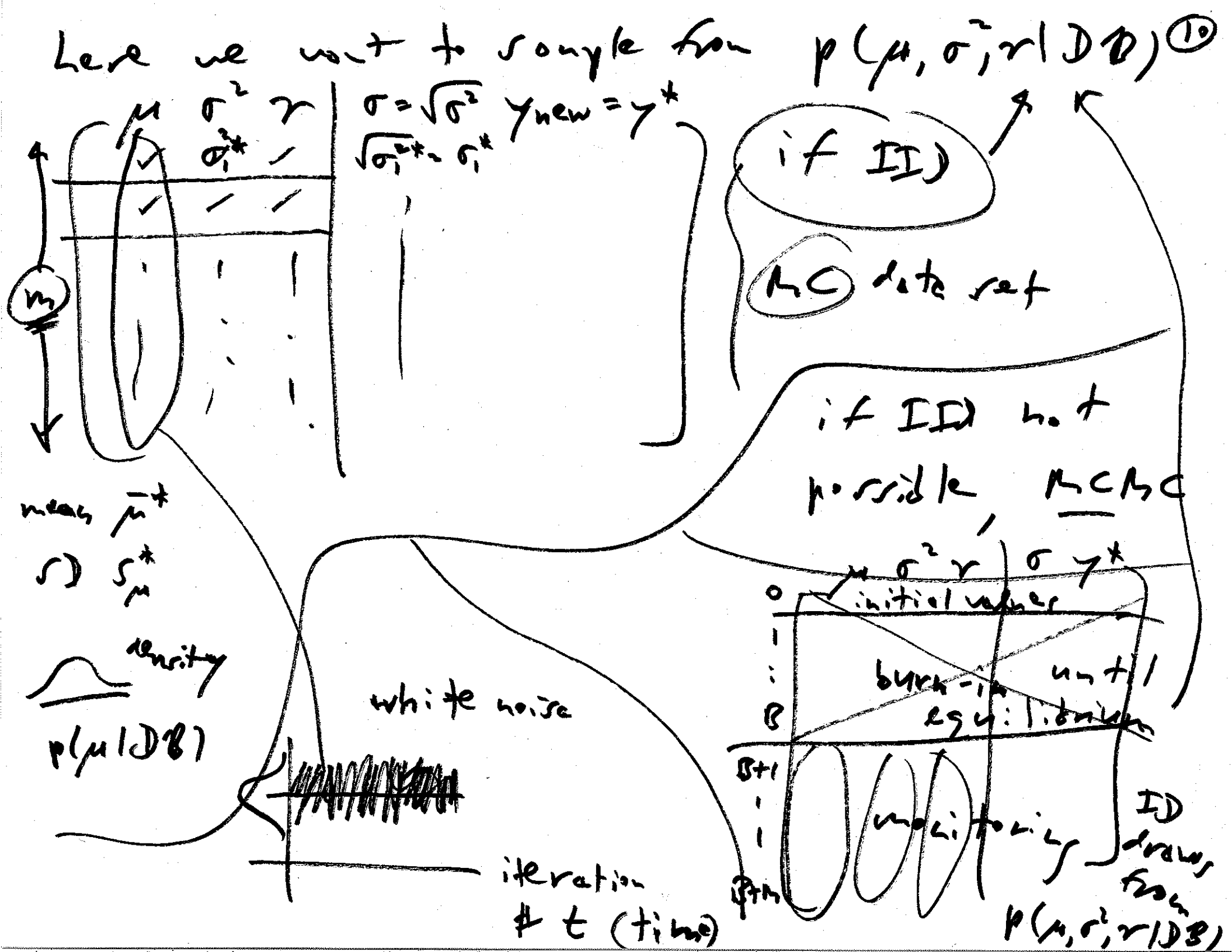
$p(\theta|DB) = c\,p(\theta|B)\,\ell(\theta|DB)$

Beta Beta · Beta

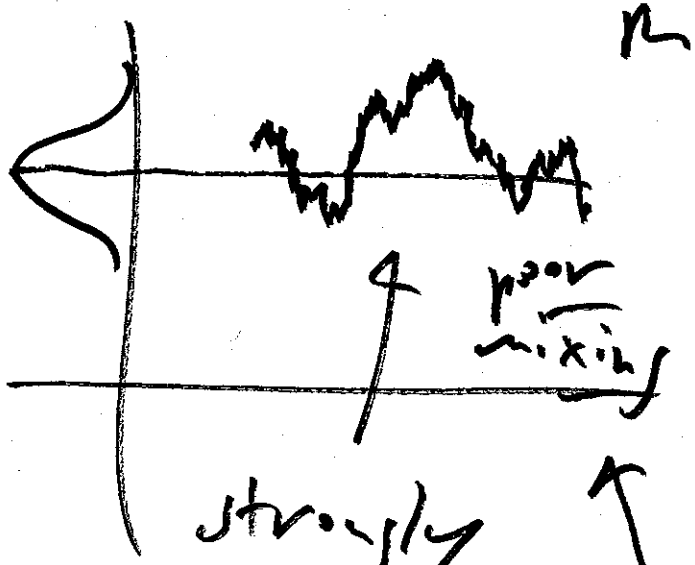
but conjugate priors only exist for a small subset of {real problems with small k}; need another more general computing method: MCMC

ex. $(y_i|\mu,\sigma^2,\tau,B) \overset{iid}{\sim} t_\tau(\mu,\sigma^2)$ $(i=1,\dots,n)$ $k=3$

NB to data $(\mu,\sigma,\tau|B) \sim p(\mu,\sigma^2,\tau|B)$


Lets we want to sample from $p(\mu, \sigma^2, r | DB)$ (10)
μ σ^2 r
$\sigma = \sqrt{\sigma^2}$ $y_{new} = y^*$
$\sqrt{\sigma_1^{2*}} \sim \sigma_1^*$
if II)
MC data set
if II) not possible, MCMC
μ σ^2 r σ y^*
0
initial values
burn-in
until
B
B+1
ergodicity
B+m
II) draws from $p(\mu, \sigma^2, r | DB)$
mean $\bar{\mu}^*$
s_μ^*
density
$p(\mu | DB)$
white noise
iteration
t (time)

MCMC time series plot



iter #t

strongly positively autocorrelated time series

$\hat{\rho}_1$ large & positive (near +1)

(lag 1 autocorrelation)

$$\begin{bmatrix} \mu^*_{B+1} & \mu^*_{B+2} \\ \vdots & \vdots \\ \mu^*_{B+m-1} & \mu^*_{B+m} \\ \mu^*_{B+m} & X \end{bmatrix} \quad \begin{matrix} \uparrow \\ m-1 \\ + \end{matrix} \quad \begin{matrix} \hat{\rho}_1 \\ = \end{matrix}$$

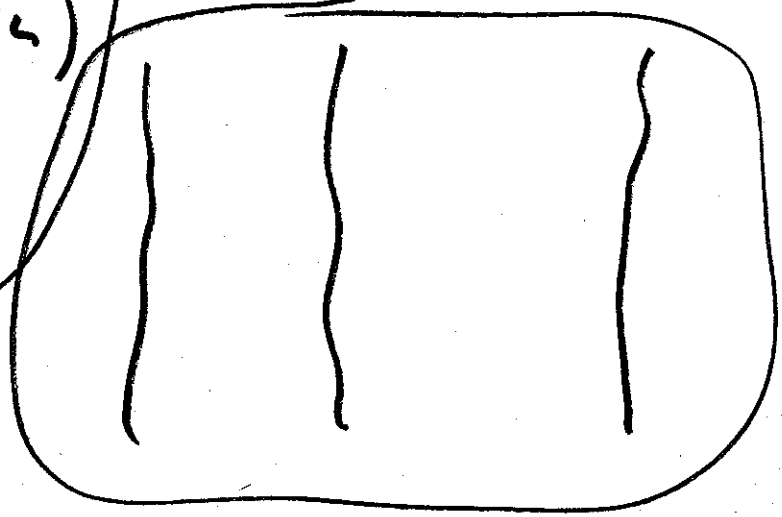
brute-force parallelizing MCMC

node 1 | node 2 | .. | node N

random # seed 1 2 .. N

same init. values (or different)

same dgm is

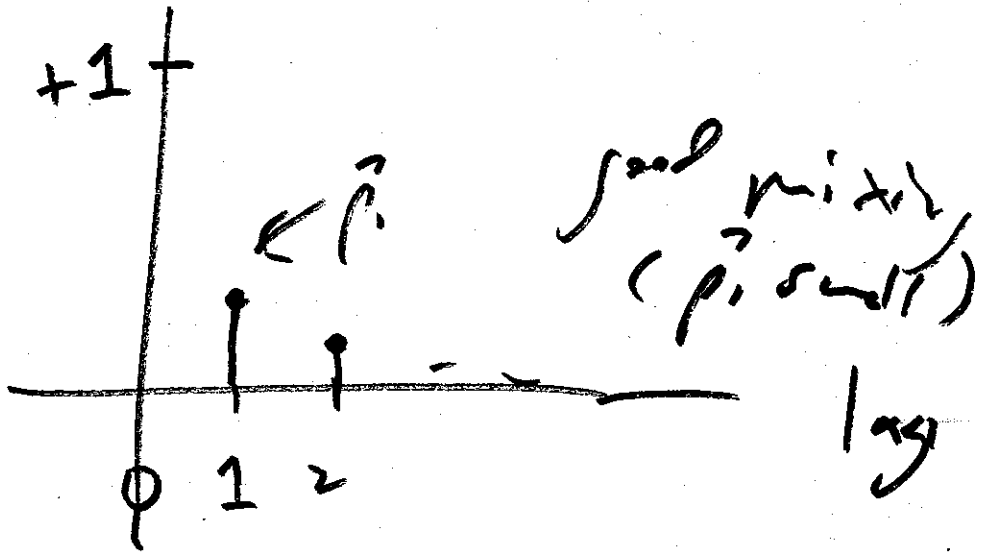


$\hat{\rho}(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_k)$ ACF

sample autocorrelation function

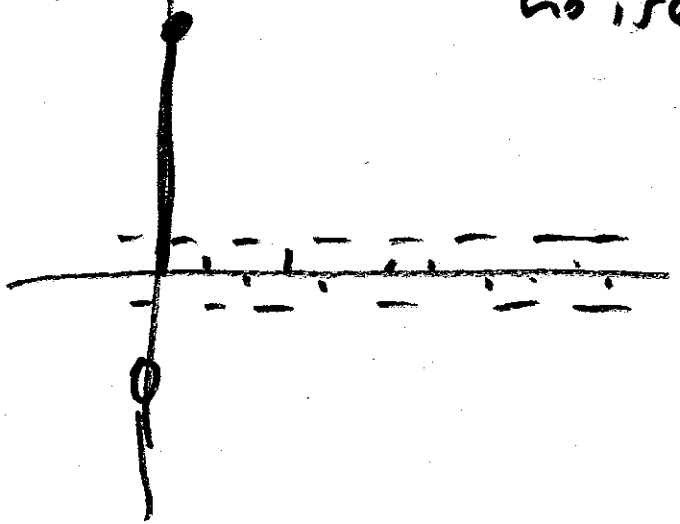
$\hat{\rho}_1$

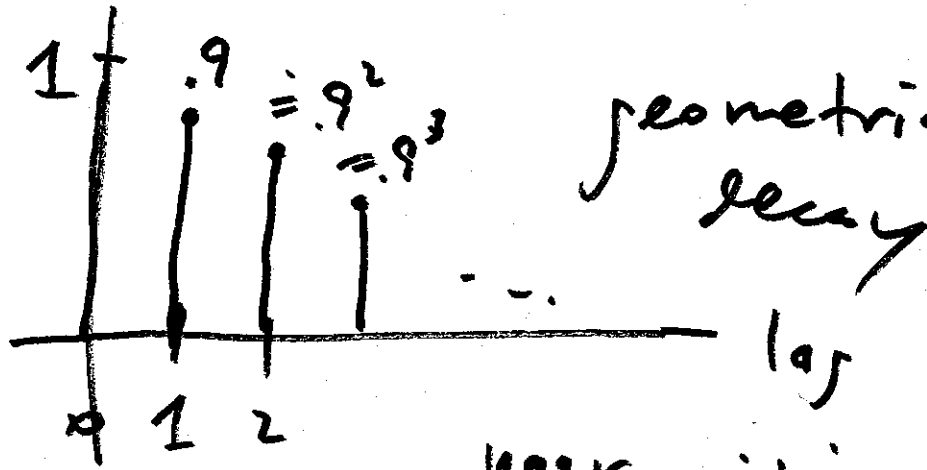
$\hat{\rho}_2$



good mixing ($\hat{\rho}_i$ small)

ACF white noise





geometric decay

poor mixing with large $\hat{\rho}_i$

common class of time series models (Box-Jenkins, time domain) is called autoregressive (AR) models of order (m):

$$\theta_t = \alpha_1 \theta_{t-1} + \alpha_2 \theta_{t-2} + \cdots + \alpha_m \theta_{t-m} + e_t$$

$e_t \overset{iid}{\sim} N(0, \sigma^2)$

---

fact many columns in mcmc datasets look like AR_1 (AR, order 1) time series with 1st order autocorrelation ρ_1

---

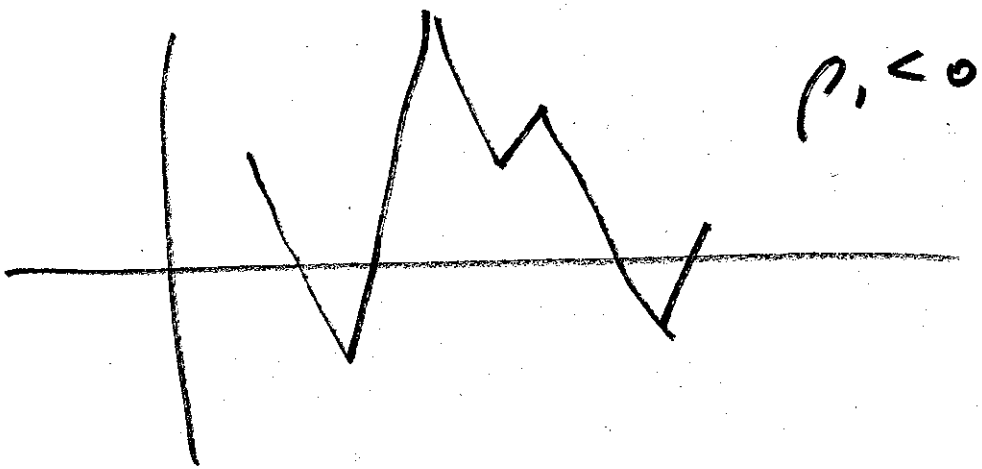
if $\rho_1 = 0 \rightarrow$ white noise (IID) (AR_0)

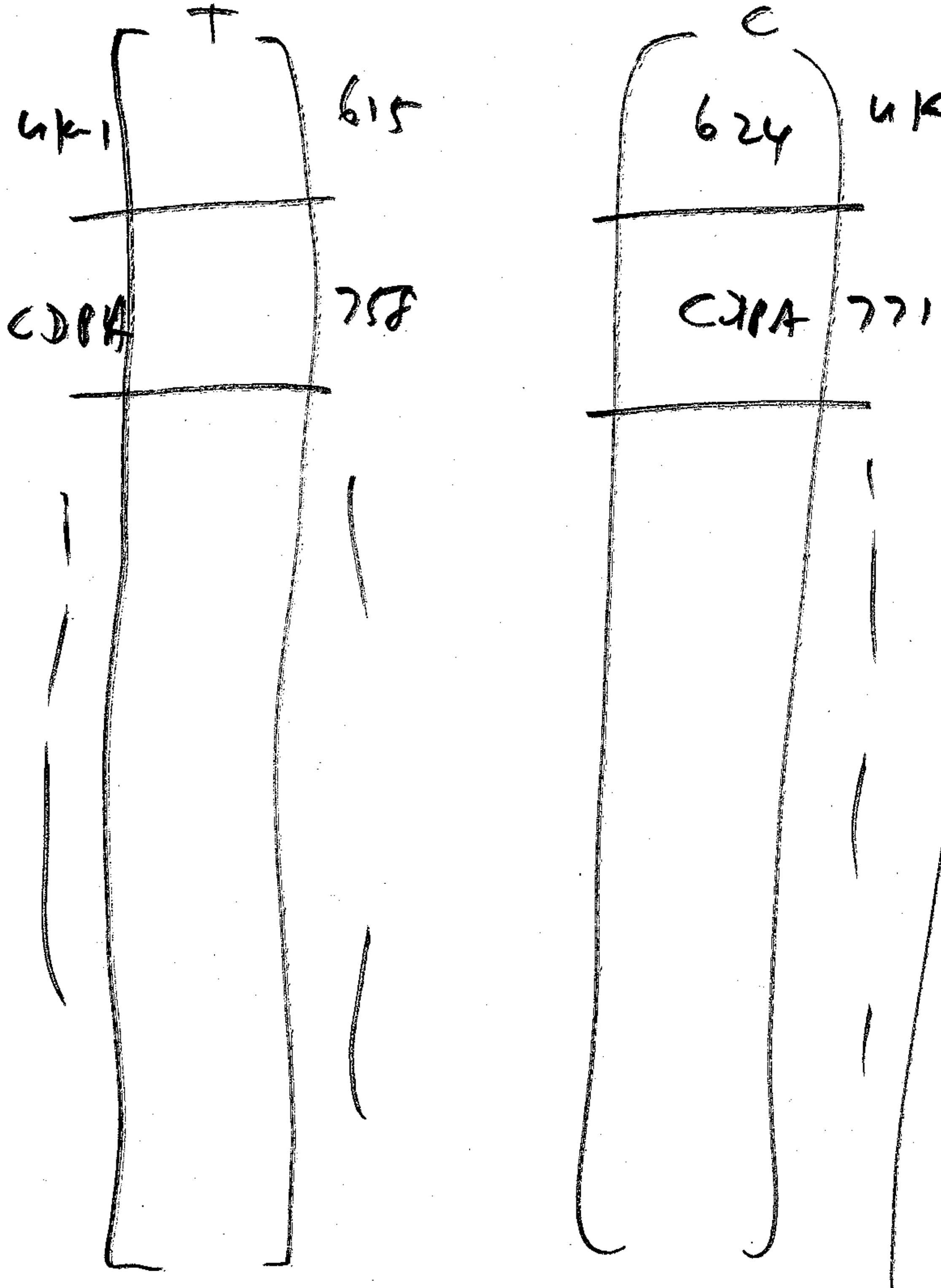
Gibbs sampling in the NBIO t model

is $p(\mu, \sigma^2, r \mid DB)$; full conditionals are

$$\left\{ \begin{array}{l} p(\mu \mid \sigma^2 r DB) \\ p(\sigma^2 \mid \mu r DB) \\ p(r \mid \mu \sigma^2 DB) \end{array} \right\}$$

ex. of methods called stochastic relaxation





naive pooling (fixed effects analysis)

(terrible idea)

in my own study i:

$\widehat{SE}(\bar{T}_i) = \widehat{SE}(\hat{p}_{T_i}) = \sqrt{\frac{\hat{p}_{T_i}(1-\hat{p}_{T_i})}{n_{T_i}}}$

$\frac{1}{n_{T_i}} \sum Y_{ij} = \hat{p}_{T_i}$

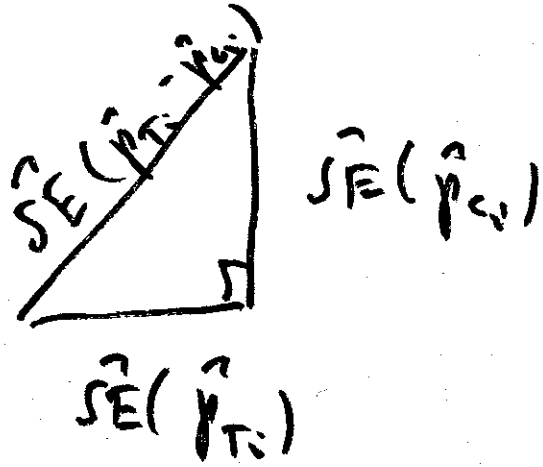
(1/0 → Y_{ij}; study → i; patient → j)

$\widehat{SE}(\bar{C}_i) = \widehat{SE}(\hat{p}_{C_i}) = \sqrt{\frac{\hat{p}_{C_i}(1-\hat{p}_{C_i})}{n_{C_i}}}$

$$\widehat{SE}(\bar{T}_i - \bar{C}_i) = \widehat{SE}(\hat{p}_{T_i} - \hat{p}_{C_i}) =$$

$\Downarrow$

indep.



$$= \sqrt{\left(\widehat{SE}(\hat{p}_{T_i})\right)^2 + \left(\widehat{SE}(\hat{p}_{C_i})\right)}$$

$$= \sqrt{\frac{\hat{p}_{T_i}(1-\hat{p}_{T_i})}{n_{T_i}} + \frac{\hat{p}_{C_i}(1-\hat{p}_{C_i})}{n_{C_i}}}$$

$$= \sqrt{V_i} = SE \text{ of diff } (\%)$$

a straightforward way to make operational the judgment of exchangeability of unknown quantities $\theta_1, \ldots, \theta_k$ is to regard them as (conditionally) IID draws from the same dist: eg.

$(\theta_j \mid \mu, \sigma^2) \overset{iid}{\sim} N(\mu, \sigma^2)$

data have no info about some quantity → likelihood-flat → post = prior for that quantity

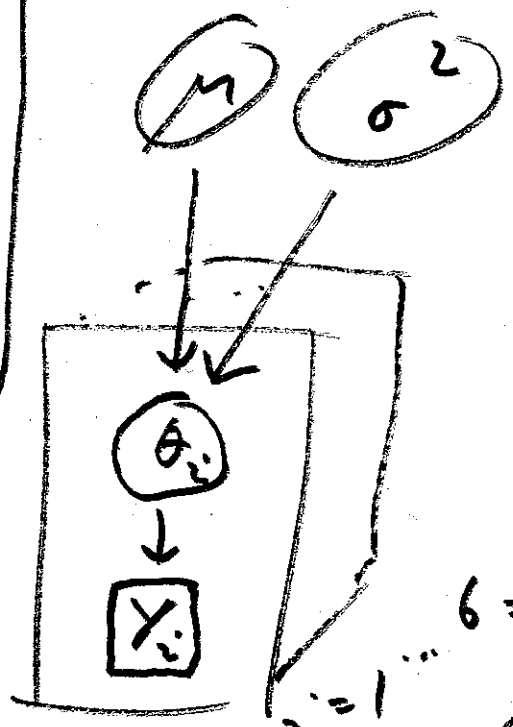
graphical representation of model (1) & (2) of part 4 of PDF notes

$(\mu, \sigma^2) \sim p(\mu, \sigma^2)$

$(\theta_i \mid \mu, \sigma^2) \overset{iid}{\sim} N(\mu, \sigma^2)$

$(Y_i \mid \theta_i) \overset{ind}{\sim} N(\theta_i, V_i)$

known



μ, σ^2 parents of θ_i

θ_i parent of Y_i

1 sheet (plate) for each θ_i

○ unknown

□ known

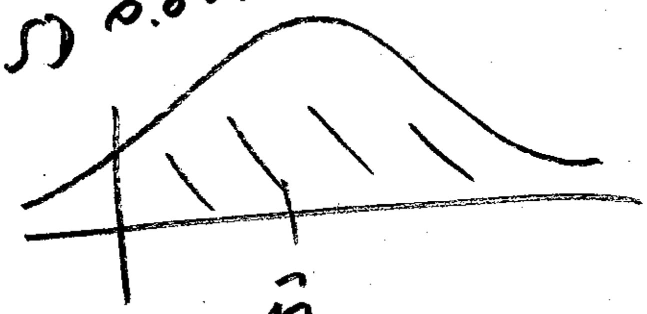
$\theta = (\theta_1, \ldots, \theta_k)$

$Y = (Y_1, \ldots, Y_k)$

(Y_i ⊥ (μ, σ^2) conditionally indep. given θ)

$$p(\mu, \sigma^2, \theta, Y) = p(\mu)\, p(\sigma^2 \mid \mu)\, p(\theta \mid \mu, \sigma^2)\, p(Y \mid \theta, \cancel{\mu, \sigma^2})$$

SD 0.809



approx

$p(\mu \mid D\,B)$

$p(\mu > 0 \mid D\,B)$ ← model (1) is correct

$\doteq 0.96$

$$\frac{0 - 1.45}{0.809} = -1.79$$

$p($ aspirin beneficial $\mid D\,B)$

μ^* positive effect $= \begin{cases} 1 & \text{if } \mu^* > 0 \\ 0 & \text{else} \end{cases}$ in pop

$$\left[\begin{array}{c} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{array} \right]$$

mean $\hat{p}(\mu > 0 \mid D\,B)$

$\frac{96}{4} = 24$ to 1 post. odds of that aspirin works