WANT $P(\theta, \sigma^2 | X)$, could not analytically derive with prior from last class

we have $p(\theta | X, \sigma^2) = N\left(\dfrac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \dfrac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)$ (I) and

$$P(\sigma^2 | X, \theta) = \text{InvGamma}\left(\frac{n_0 + n}{2}, \frac{n_0 \sigma_0^2 + n \hat{\sigma}^2_{MLE}}{2} \right) \text{(II)}$$

Can we use these two unidimensional non-marginal posteriors to sample from the full posterior?

$$P(\theta, \sigma^2 | X) = P(\theta | X, \sigma^2) P(\sigma^2 | X) \leftarrow \text{kernel for this was too difficult}$$
$$= P(\sigma^2 | X, \theta) P(\theta | X) \leftarrow \text{same here, would result into grid sampling again.}$$

Cannot use (I) and (II) together to get full posterior

Consider the following Alg (numerical sampling):

1) Begin at θ_0 = some reasonable value.
2) Draw σ_1^2 from $P(\sigma^2 | X, \theta_0)$ via ~~rnorm~~ rinvgamma
3) Draw θ_1 from $P(\theta | X, \sigma_1^2)$ via ~~rinvgamma~~ rnorm
4) Draw σ_2^2 from $P(\sigma^2 | X, \theta_1)$ via ~~rn~~ rinvgamma
5) Draw θ_2 from $P(\theta | X, \sigma_2^2)$ via rnorm

[repeat sampling until "convergence"]

B = iteration number which refers to "burn-in" i.e. # of samples of this alg it takes for it to converge its warm-up.

once its converged it is (almost) ready to be sampled from, for your inference.

These samples are dependent, θ_{98} is related to θ_{97} which is dependent on θ_{96} etc. So it looks like no iid samples. This is technically true. However, . . .

we can remove ourselves by enough iterations that this dependence is negligible. How do we assess how many iterations are needed?

Back to Basic stats....... Consider two r.v.'s X_1, X_2 Then.

$$\sigma_{12} := \mathrm{Cov}[X_1, X_2] := E[(X_1-\mu_1)(X_2-\mu_2)]$$

$$\rho := \mathrm{Corr}[X_1, X_2] := \frac{\sigma_{12}}{\sigma_1\sigma_2} \in [-1, 1]$$

we have estimators / estimates for these parameters

$$\sigma_{12} \approx s_{12} := \frac{1}{n-1}\sum_i (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)$$

$$\rho \approx r := \frac{s_{12}}{s_1 s_2} = \frac{\sum (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sqrt{\sum (x_{1i}-\bar{x}_1)^2 \sum (x_{2i}-\bar{x}_2)^2}}$$

Now consider $X_1, X_2, \ldots$ to X_S be r.v.'s from an iterative process where X_1 is the first iteration etc where each iteration has a dependence on the previous. We define autocorrelation which is correlation with a previous iteration.

First autocorrelation with iteration directly before call this!

$$r_{a_1} := \frac{\sum_{t=2}^{S}(x_t-\bar{x})(x_{t-1}-\bar{x})}{\sum_{t=1}^{S}(x_t-\bar{x})^2} \qquad \bar{x} = \frac{1}{S}\sum_{t=1}^{S} x_t$$

$$r_{a_2} := \frac{\sum_{t=3}^{S}(x_t-\bar{x})(x_{t-2}-\bar{x})}{\sum_{t=1}^{S}(x_t-\bar{x})^2}$$

⋮

etc.

In our sampling idea, we assess autocorrelations for both θ and σ^2.

for θ

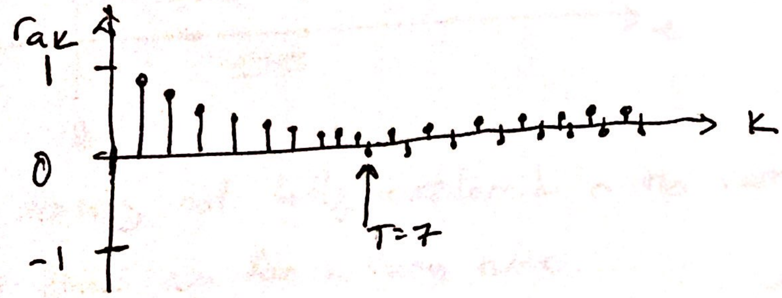
$$r_{\theta k} := \frac{\sum_{t=B+1+k}^{S} (\theta_t - \bar{\theta})(\theta_{t-k} - \bar{\theta})}{\sum_{t=B+1}^{S} (\theta_t - \bar{\theta})^2} \qquad \bar{\theta} = \frac{1}{S-B} \sum_{t=B+1}^{S} \theta_t$$

$k = 1, 2, 3, \ldots$

$\uparrow$

called the lag.

How to assess the dependence? Look at autocorrelation plot:



Once the autocorrelation appears to be not significantly diff from zero, you select the first k and call it "T" for "thinning" distance.

Then, you "thin" the chain of samples $B=5$ $T=3$: $\leftarrow$ retain every 3^{rd} val

$$\left[\cancel{\begin{bmatrix}\theta_1 \\ \sigma_1^2\end{bmatrix}}, \cancel{\begin{bmatrix}\theta_2 \\ \sigma_2^2\end{bmatrix}}, \cancel{\begin{bmatrix}\theta_3 \\ \sigma_3^2\end{bmatrix}}, \ldots, \cancel{\begin{bmatrix}\theta_5 \\ \sigma_5^2\end{bmatrix}}, \begin{bmatrix}\theta_6 \\ \sigma_6^2\end{bmatrix}, \begin{bmatrix}\theta_9 \\ \sigma_9^2\end{bmatrix}, \cancel{\begin{bmatrix}\theta_{10} \\ \sigma_{10}^2\end{bmatrix}}, \cancel{\begin{bmatrix}\theta_{11} \\ \sigma_{11}^2\end{bmatrix}}, \begin{bmatrix}\theta_{12} \\ \sigma_{12}^2\end{bmatrix}, \ldots \right]$$

delete all these

B=5 Burn

keep

keep

keep

throw away $\begin{bmatrix}\theta_7 \\ \sigma_7^2\end{bmatrix}$ $\begin{bmatrix}\theta_8 \\ \sigma_8^2\end{bmatrix}$

MATH 341 Lec 22 5/10

$$\left\{ \begin{bmatrix}\theta_6\\ \sigma_6^2\end{bmatrix}, \begin{bmatrix}\theta_9\\ \sigma_9^2\end{bmatrix}, \begin{bmatrix}\theta_{12}\\ \sigma_{12}^2\end{bmatrix}, \ldots \right\} =: M$$

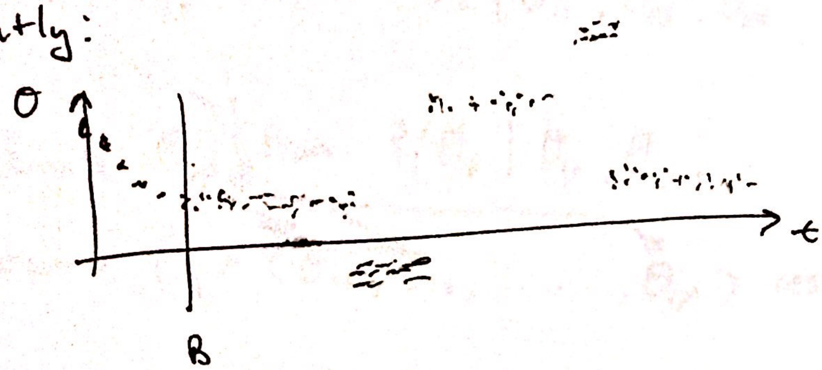
"Burned and thinned" chain, ready for inference

these are iid samples from the posterior. $P(\theta, \sigma^2 | X)$

Use the same procedures from grid approximation on M.

pval, h-test, point estimates, CRs etc.

There are still problems with this method. The following happens frequently:



pockets of density not fully explored in the sampling scheme

Soln 1) Let chain run for a long time

2) Begin chain from diff locations and ~~add~~ union multiple chains together.

Another problem: need to know all conditional distributions or be able to src sample them accurately with undue computational burden.

The algorithm we just discussed is called "Systematic Sweep Gibbs Sampler". General Algorithm.

To sample from $P(\theta_1, \theta_2, \theta_3, \dots, \theta_p | X)$

General Algorithm:

1) Pick $\vec{\theta} = [\theta_{01}, \theta_{02}, \theta_{03}, \dots, \theta_{0p}]$ reasonable guesses for each

2-1) Sample θ_{11} from $P(\theta_1 | \theta_2 = \theta_{02} \dots \theta_{0p})$ ← sampling mixing from conditional distributions

2-2) Sample θ_{12} from $P(\theta_2 | \theta_{11}, \theta_{03}, \theta_{04}, \dots \theta_{0p})$

⋮

2-p) Sample θ_{1p} from $P(\theta_p | \theta_{11}, \theta_{12}, \dots \theta_{1,p-1})$

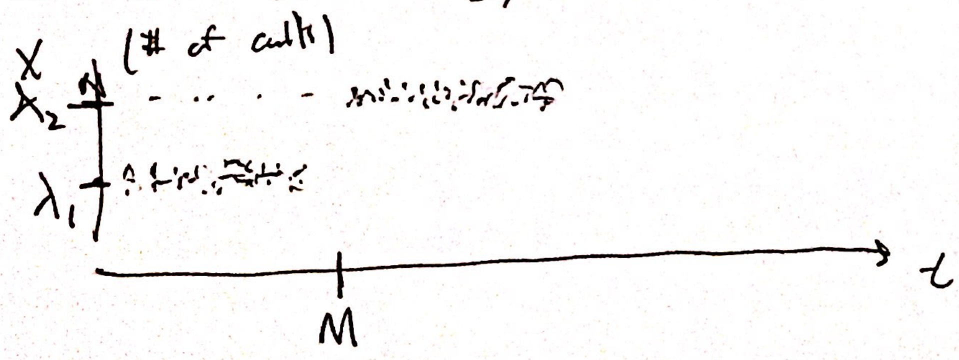
3) Record $\vec{\theta}_1 = \langle \theta_{11}, \theta_{12}, \dots, \theta_{1p} \rangle$ results of steps (2-1)→(2-p)

4) repeat steps 2, 3 many times.

5) Burn all chains at the highest B value across all p chains.

6) Thin the chains at the highest T value across all p chains.

Real World Example. Change-point modeling. Assume there is a poisson number of phone calls with mean λ_1 and then at some point in time it changes to a poisson number of calls with mean λ_2, diff from λ_1.



Goal: inference for M a ~~...~~ unknown parameter that is a change time

There are two unknown nuisance parameters λ_1 and λ_2 which we don't need to infer. Thus $p = 3$ dimensions and the full posterior will be $P(m, \lambda_1, \lambda_2 \mid X)$. We will built a Gibbs sampler that will provide us inference next class.

<u>Review Sunday B4 EXAM</u>: 6PM Night B4.