Want $P(\theta, \sigma^2 \mid X)$ and we have

$$P(\theta \mid X, \sigma^2) = N\left(\frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},\ \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right) \quad \text{and}$$

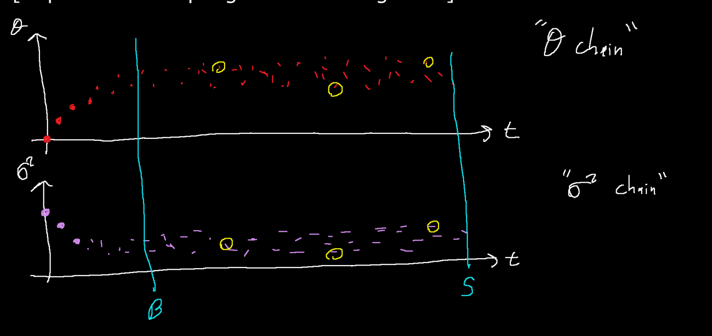
$$P(\sigma^2 \mid X, \theta) = \text{InvGamma}\left(\frac{n_0 + n}{2},\ \frac{n_0 \sigma_0^2 + n \hat{\sigma}^2_{MLE}}{2}\right)$$

Can we use these two uni-dimensional non-marginal posteriors to sample from the full posterior? Note that:

$$P(\theta, \sigma^2 \mid X) = P(\theta \mid X, \sigma^2)\, P(\sigma^2 \mid X) = P(\sigma^2 \mid X, \theta)\, P(\theta \mid X)$$

Consider the following numerical sampling algorithm:

(1) Begin at θ_0 = some reasonable value.
(2) Draw σ^2_1 from P(σ^2 | X, θ = θ_0) via rinvgamma.
(3) Draw θ_1 from P(θ | X, σ^2 = σ^2_1) via rnorm.
(4) Draw σ^2_2 from P(σ^2 | X, θ = θ_1) via rinvgamma.
(5) Draw θ_2 from P(θ | X, σ^2 = σ^2_2) via rnorm.
.....
[Repeat this sampling until "convergence"]



"B" is the iteration number that refers to the "burn-in" i.e. the number of samples of this algorithm it takes for the algorithm to be converged. Once it's converged, it's considered [almost] ready to provide samples for your inference. There is another problem... these samples are dependent!! Why θ_98 is very related to θ_97 which is dependent on θ_96, etc... It seems like it is impossible to get iid samples from the posterior! That is technically true... but... we can remove ourselves by enough iterations that this dependence is negiligible. How do we assess how many iterations? And negligibility?

Let's go back to basic stats for a minute. Consider two r.v.'s X_1, X_2. Then

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

$$\rho := \text{Corr}[X_1, X_2] := \frac{\sigma_{12}}{\sigma_1 \sigma_2} \in [-1, +1] \quad \text{proved in 368.}$$

We have estimators/estimates for these parameters:

$$\sigma_{12} \approx s_{12} := \frac{1}{n-1} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$$

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

Now consider X_1, X_2, ..., X_S to be r.v.'s from an iterative process where X_1 is the first iteration, X_2 is the second iteration, etc where each iteration has a dependence on the previous iteration. We define "autocorrelation" which is correlation with a previous iteration. First, we have autocorrelation with the iteration directly before:

$$r_{a1} := \frac{\sum_{t=2}^{S} (X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=1}^{S} (X_t - \bar{X})^2}, \quad \bar{X} = \frac{1}{S} \sum_{t=1}^{S} X_t,$$

$$r_{a2} := \frac{\sum_{t=3}^{S} (X_t - \bar{X})(X_{t-2} - \bar{X})}{\sum_{t=1}^{S} (X_t - \bar{X})^2}$$

⋮

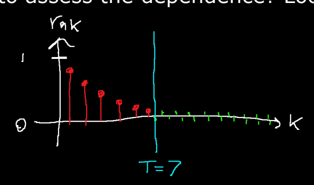
etc

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

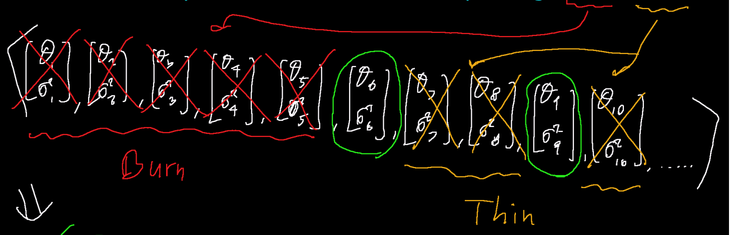
$$r_{ak}^{\theta} := \frac{\sum_{t=B+1+k}^{S} (\theta_t - \bar{\theta})(\theta_{t-k} - \bar{\theta})}{\sum_{t=B+1}^{S} (\theta_t - \bar{\theta})^2}, \quad \bar{\theta} = \frac{1}{S-B} \sum_{t=B+1}^{S} \theta_t$$

$k = 1, 2, \ldots,$ the "lag"

How to assess the dependence? Look at the autocorrelation plot:



Once the autocorrelation appears to be not significantly different from zero, you select that first k and call it "T" for "thinning" distance. Then, you "thin" the chain of samples e.g. B = 5, T = 3:



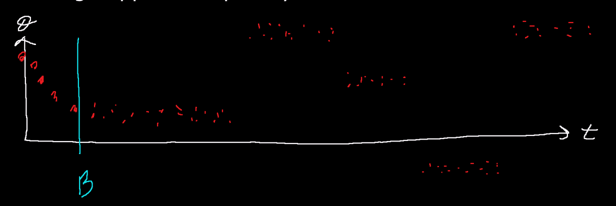
$$\Downarrow$$

$$\mathcal{M} := \left\{ \begin{bmatrix} \theta_6 \\ \sigma^2_6 \end{bmatrix}, \begin{bmatrix} \theta_9 \\ \sigma^2_9 \end{bmatrix}, \begin{bmatrix} \theta_{12} \\ \sigma^2_{12} \end{bmatrix}, \ldots \right\},$$

the burned and thinned chain ready for running approximatie inference!

These are considered iid samples from the posterior P(θ, σ | X)

There are still problems with this approach. For instance, the following happens frequently:



The problem here is the posterior has many pockets of density that may not be fully explored in this sampling scheme. Partial solution: (1) let the chain run a very long time and (2) begin the chain from different locations and then union the multiple chains together.

Another practical problem is that you either need to know all conditional distributions or be able to grid sample them accurately and without undue computational burden.

The algorithm we just discussed is quite famous and is called the "Systematic Sweep Gibbs Sampler". Here is the general algorithm below to sample from P(θ_1, θ_2, ..., θ_p | X):

(1) Pick $\vec{\theta}_0 = \langle \theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,p} \rangle$

(2-1) Sample $\theta_{1,1}$ from $P(\theta_1 \mid \theta_2 = \theta_{0,2}, \ldots, \theta_p = \theta_{0,p})$

(2-2) Sample $\theta_{1,2}$ from $P(\theta_2 \mid \theta_1 = \theta_{1,1}, \theta_3 = \theta_{0,3}, \ldots, \theta_{0,p})$

the p conditional distributions

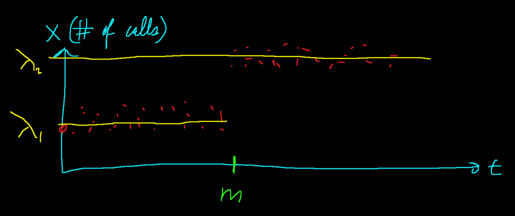
...
(2-p) Sample $\theta_{1,p}$ from $P(\theta_p \mid \theta_1 = \theta_{1,1}, \ldots, \theta_{p-1} = \theta_{1,p-1})$

(3) Record $\vec{\theta}_1 = \langle \theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,p} \rangle$ i.e. the result of step 2

(4) Repeat step 2 and 3 many times.

(5) Burn all the chains at the highest B value across all p chains.
(6) Thin the chains at the highest T value across all p chains.

A real-world example finally! 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 which is diffferent than λ_1.



The goal is inference for m (an unknown parameter that is the change time). There are two unknown nuisance parameters λ_1 and λ_2 which you don't need to infer. Thus p = 3 dimensions and the full posterior will be P(m, λ_1, λ_2 | x). We will build a Gibbs sampler that will provide us inference next class.