Data-analysis and Inverse Methods in Astronomy

Gibbs sampling

Antti Penttilä

Homepage http://wiki.helsinki.fi/display/53834

Gibbs sampling

The algorithm "started" the ever-increasing popularity of MCMC methods when published by Geman and Geman at 1984. It was named after J. W. Gibbs who studied statistical physics.

Reasons behind popularity: Can solve (computationally) Bayesian problems that are analytically intractable; is (and was) fast enough for even early computers.

Idea in a nutshell – Joint distribution can be constructed using conditional distributions.

...Gibbs sampling

Suppose we want to know joint distribution $D(\boldsymbol{\theta})$, but we do not know it analytically. If we know all the conditional distributions

$$D(\theta_i|\boldsymbol{\theta}_{j\neq i})$$

The chain of simulated values for θ_i's from conditional distributions

$$\theta_1^{(1)} \sim D(\theta_1|\boldsymbol{\theta}_{j\neq 1})$$
$$\cdots$$
$$\theta_k^{(1)} \sim D(\theta_k|\boldsymbol{\theta}_{j\neq k})$$
$$\theta_1^{(2)} \sim D(\theta_1|\boldsymbol{\theta}_{j\neq 1})$$
$$\cdots$$

(when always using latest, updated values for $\boldsymbol{\theta}_{j\neq i}$) will converge to the (unknown) distribution $D(\boldsymbol{\theta})$

...Gibbs sampling

GS is a special case of Metropolis-Hastings, where update is done component-wise and proposed value is always accepted (therefore efficient).

The only drawback in GS is that conditionals must be known. In practice, hybrid methods are used. If conditionals are known for some block of parameters, pure GS is used there. For other blocks M-H could be used inside the GS step to approximate the unknown conditional.

Gibbs sampling in regression models

GS suits especially well to Bayesian linear or nonlinear regression, if certain choices are made with model assumptions and a priori distributions. In regression the model is

$$\boldsymbol{y} = f(\boldsymbol{X};\boldsymbol{\theta}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ are the observed data, f is the model, $\boldsymbol{X}$ the explanatory variables , $\boldsymbol{\theta}$ the model parameters, and $\boldsymbol{\varepsilon}$ the errors between model and data. Statistics come to play as we assume $\boldsymbol{\varepsilon}$ to be random having distribution $D(\mathbf{0},\boldsymbol{\Sigma})$, thus $\boldsymbol{y}$ has distribution $D(f(\boldsymbol{X};\boldsymbol{\theta}),\boldsymbol{\Sigma})$. We want to estimate distribution for $\boldsymbol{\theta}$, which is written in Bayesian regression (BR) as

$$D(\boldsymbol{\theta}, \Sigma|\boldsymbol{y}) = D_{pr}(\boldsymbol{\theta}, \Sigma)D(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma)$$

Steps towards GS in BR

Usually it is assumed that the residual variance $\boldsymbol{\Sigma}$ and model parameters $\boldsymbol{\theta}$ are independent, thus

$$D_{pr}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = D_{pr}(\boldsymbol{\theta}) D_{pr}(\boldsymbol{\Sigma})$$

and that observations are independent (and identical), thus

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I} \quad \text{(or } \sigma^2 \operatorname{diag}(w_1 \ldots w_n) \text{ if unequal errors)}$$

Collecting these gives

$$D(\boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}) = D_{pr}(\boldsymbol{\theta}) D_{pr}(\sigma^2) D(\boldsymbol{y} | \boldsymbol{\theta}, \sigma^2)$$

…GS in BR

Now, conditional distributions are

$$D(\boldsymbol{\theta}|\sigma^2, \boldsymbol{y}) \propto D_{pr}(\boldsymbol{\theta}) D(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)$$

$$D(\sigma^2|\boldsymbol{\theta}, \boldsymbol{y}) \propto D_{pr}(\sigma^2) D(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)$$

GS is efficient choice if the above distributions are known. Even if the choices above are suitable for only some of the parameters, hybrid GS combining 'pure GS' and M-H updates can be efficient.

...GS in BR

To my knowledge, there are at least the following cases where conditionals in previous slide are known. The error term $\boldsymbol{\varepsilon}$ is assumed Gaussian, the model linear, and

$$D_{pr}(\boldsymbol{\theta}) \sim 1 \text{ (inproper even dist.), shifted Exp}(\lambda_0), \text{ or } \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$D_{pr}(\sigma^2) \sim \text{ scaled-inverse-}\chi^2(\alpha, \beta), \text{ or } \mathcal{IG}(\alpha, \beta) \text{ (inverse-gamma)}$$

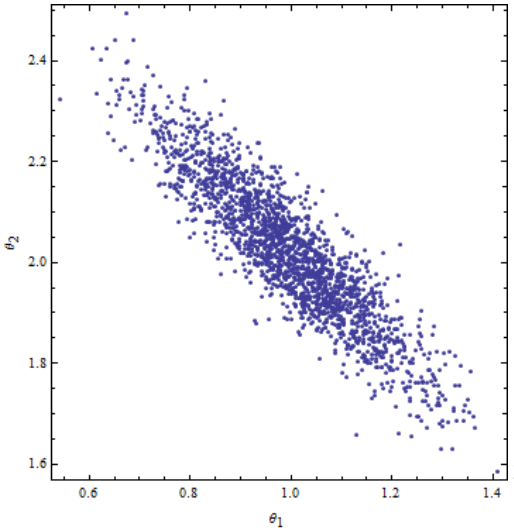
For example, with even distribution for $\boldsymbol{\theta}$ and inverse-gamma for σ^2 the conditionals for Gibbs sampling step are

$$D(\boldsymbol{\theta}|\sigma^2, \boldsymbol{y}) \sim \mathcal{N}(X^T X)^{-1} X^T \boldsymbol{y}, \sigma^2 (X^T X)^{-1})$$

$$D(\sigma^2|\boldsymbol{\theta}, \boldsymbol{y}) \sim \mathcal{IG}(\alpha + \frac{n}{2}, \beta + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\theta})^T (\boldsymbol{y} - X\boldsymbol{\theta}))$$

Computer task

- Use information on previous slide, and write a program that uses Gibbs sampling to estimate model fit and parameter distribution
  - Model is $y = \theta_1 x_1 + \theta_2 x_2$
  - Data can be found on course webpage
- Note: $(X^T X)^{-1} X^T \boldsymbol{y} = \hat{\boldsymbol{\theta}}$, a.k.a. the model fit
- $(\boldsymbol{y} - X\boldsymbol{\theta})^T (\boldsymbol{y} - X\boldsymbol{\theta})$ is the sum of squared residuals
- produce scatterplot of possible parameter values (see fig.) and estimates for parameter values
- (this example is a bit silly, because posterior is also known, and if noninformative priors are used, it does not offer much in addition to regular linear regression…)



Tips

- Random number from inverse gamma
  - If $x \sim \text{Gamma}(\alpha,1/\beta)$ then $1/x \sim \text{inv-Gamma}(\alpha,\beta)$
  - Suitable hyperparameters for noninformative inv-gamma would be $\alpha=\beta=0.001$
- In a nutshell
  - load data
  - Simulate parameters (θ_1,θ_2), then σ^2 over and over again using distributions in page 8. You will need a starting value for the first σ^2.
    - If you can simulate only one-dimensional normal distribution, see next slide
  - Throw off some of the first data, plot rest of (θ_1,θ_2)-pairs and compute estimates as means or medians of the simulated values.

...tips

If you can draw random numbers only from 1-dimensional normal distribution, you must alter the Gibbs chain accordingly:

- draw θ_1 from

$$\mathcal{N}(\hat{\theta}_1 + \frac{c_{12}(\theta_2 - \hat{\theta}_2)}{c_{22}}, \frac{(c_{11}c_{22} - c_{12}^2)\sigma^2}{c_{22}})$$

- and θ_2 from

$$\mathcal{N}(\hat{\theta}_2 + \frac{c_{12}(\theta_1 - \hat{\theta}_1)}{c_{11}}, \frac{(c_{11}c_{22} - c_{12}^2)\sigma^2}{c_{11}})$$

- and σ^2 from

$$\mathcal{IG}(\alpha + \frac{n}{2}, \beta + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}))$$

- where [see next slide], and iterate this

...tips

$$(\hat{\theta}_1, \hat{\theta}_2) = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix}$$