Model Selection/Averaging

- Suppose we have several believable models but are not sure which one is correct. The models may depend on parameters (different parameters for each model, of course)
- We supply (in general different) priors for each parameter set for each model
- We also supply a prior over the models
- Then we can calculate the posterior probability of the *models* by integrating the joint posterior probability over the entire parameter set for each model leaving only the model indicator

Model Selection/Averaging

- Thus, consider models $M_1, M_2, \ldots, M_N$ with parameter sets $\theta^1, \theta^2, \ldots, \theta^N$ respectively (each θ^k is a vector, of possibly differing dimensions, possibly unnested). We set priors on the parameters $p(\theta^1|M_1), p(\theta^2|M_2), \ldots, p(\theta^N|M_N)$ and priors on the models $p(M_1), p(M_2), \ldots, p(M_N)$.
- The likelihoods are also model-dependent: $p(x|\theta^1,M_1)$, $p(x|\theta^2,M_2), \ldots, p(x|\theta^N,M_N)$
- The joint posterior probability is

$$p(\theta^k, M_k \mid x) \propto p(x \mid \theta^k, M_k)\, p(\theta^k \mid M_k)\, p(M_k)$$

- The marginals with respect to θ^k are the posterior probabilities of the models:

$$p(M_k \mid x) = \int p(\theta^k, M_k \mid x) d\theta^k$$

Model Selection/Averaging

- Model averaging is done similarly, except here the models are usually empirical (e.g., models based on polynomial approximations, truncated Fourier series, or other basis functions, of variable size). Thus, we do not "believe" that any of the models is actually correct, but are using them as proxies for some unknown underlying model
- There may be some parameter, say $\theta_1^k=\rho$ that is common to all the models. This could, for example, be the parallax of a star. We are interested in the posterior distribution of that parameter.

Model Selection/Averaging

- In model averaging we marginalize differently. Having computed the joint posterior probability

$$p(\rho = \theta_1^k, \theta_{j\neq 1}^k, M_k \mid x)$$

where I have broken out the common parameter, we marginalize over the nuisance variables $\theta_{j\neq 1}^k$ and the *models* M_k, to obtain

$$p(\rho \mid x) = \sum_k \int p(\rho, \theta_{j\neq 1}^k, M_k \mid x) d\theta_2 d\theta_3 \ldots d\theta_{l_k}$$

- Thus in effect we are regarding the model index M_k itself to be a nuisance variable in addition to the $\theta_{j\neq 1}^k$; it's just that it is a discrete rather than a continuous variable

Model Selection/Averaging

- So we see that Bayesian model selection and model averaging is just an obvious application of the usual Bayesian mantra

  posterior ∝ prior × likelihood

  together with appropriate marginalization that depends only upon what we consider to be the interesting parameters (model index; common parameter)

Approximations

- A "quick and dirty" approach to model selection can be obtained using Schwarz' BIC (Bayesian Information Criterion), or the the better-known AIC (Akaike Information Criterion)
- The BIC compares two models (but can obviously be used to pick the best of k models). The data are x, of dimension n, and the vector of parameters of model j are θ_j. As we already saw, we want to compare the marginals of the two models:

$$m_j(x) = \int \underbrace{p(x \mid \theta_j, M_j)}_{\text{Likelihood}} \underbrace{p(\theta_j \mid M_j) p(M_j)}_{\text{Prior}} d\theta_j$$

Approximations

- The BIC is obtained by expanding the log of the posterior distribution to obtain the Laplace approximation (Tom Loredo will already have discussed this) and looking at the asymptotic expression in the limit of large n.
- Integrating out the parameters, one obtains

$$\text{BIC} = -2\log(L_1 / L_2) + (p_1 - p_2)\log n \approx -2\log(m_1(x)/m_2(x))$$

- The BIC penalizes models with more parameters even more severely than the similar Akaike information criterion (AIC)

$$\text{AIC} = -2\log(L_1 / L_2) + 2(p_1 - p_2)$$

Approximations

- The Deviance information criterion (DIC) can be used when comparing models for which MCMC simulations have been obtained for each model separately.
- Define the *deviance* as the quantity

$$D(\theta) = -2\log(L(\theta \mid y)) + C$$

  where C is a constant that cancels out an does not need to be known.
- The expectation $\overline{D} = E^{\theta}[L(\theta \mid y)]$ measures how well the model explains the data.
- The *effective number of parameters* is calculated as

$$p_D = \overline{D} - D(\overline{\theta})$$

- Then $DIC = \overline{D} + p_D$, analogously to AIC

Approximations

- As with AIC and BIC, DIC is only valid asymptotically as the sample gets large, and depends upon the posterior distribution (or likelihood) being approximately multivariate normal.
- The constant C cancels out whenever we compare models using DIC, since we subtract the DIC for the two models to obtain the comparison criterion
- The advantage of DIC is that it can be calculated simply from the MCMC sample, whereas AIC and BIC require evaluating at the maximum of the likelihood, which is not readily available from the sample.

Model Selection/Averaging

- The difficulties in fully Bayesian model selection or averaging are in the details. We need to consider
  - What priors to put on the parameters?
    - » There will be an automatic Ockham's razor that which will favor models with a smaller number of parameters over those with a larger number of parameters
    - » However, the priors on the parameters in higher-dimensional models must not be too "spread out", or they will have no chance to compete
    - » There are suggestions for good priors on the parameters, but it is best to discuss this with an experienced statistician

Model Selection/Averaging

- The difficulties in fully Bayesian model selection or averaging are in the details. We need to consider
  - How do we do the marginalization?
    - » In the case of multivariate linear normal models, the marginalization can mostly be done exactly, although there are some difficulties
    - » More generally, though, we'll have to run a complicated MCMC simulation that moves not only in parameter space but also in model space. This turns out to be a little tricky

MCMC Simulation

- We can calculate results using simulation (useful for when an exact solution is unavailable)
- We simulate a random walk in both model space $\{H_0, H_1\}$ and in parameter (θ) space. Thus we are sampling on both discrete and continuous parameters
- The key is to allow ourselves to propose jumps between models. This will in general be a M-H step
- Since the models may have differing numbers of parameters we will have to propose parameters and models simultaneously
- I will describe a technique known as *reversible jump MCMC* which is very effective

MCMC Simulation

- The best introduction to the reversible jump MCMC technique that I have found is "On Bayesian Model and Variable Selection Using MCMC," by Petros Dellaportas, Jonathan Forster and Ioannis Ntzoufras. It has been published in *Statistics and Computing*. A copy may be downloaded from my website:

  http://bayesrules.net/courses/stat295.2005/DellaportasModelSelect.pdf

- I will discuss only a basic but widely applicable version of the reversible jump MCMC sampler, the "independence sampler". The full reversible jump algorithm is more flexible, but more involved (see the paper)

MCMC Simulation

- Suppose we are in in state (H_m,θ_m) and wish to jump to another state (H_n,θ_n)
  - Propose to jump to state (H_n,θ_n) with probability $q(H_n,\theta_n)$ independent of (H_m,θ_m) (independence sampler)
  - Compute (the log of) the Metropolis-Hastings factor:

$$\alpha = \frac{p(X \mid H_n,\theta_n)p(H_n,\theta_n)q(H_m,\theta_m)}{p(X \mid H_m,\theta_m)p(H_m,\theta_m)q(H_n,\theta_n)}$$

  - Generate u, (the log of) a $U(0,1)$ random variable and accept the step if $u<\alpha$ ($\log u < \log \alpha$), otherwise stay where you are
  - The resulting Markov chain samples the models in proportion to their posterior probability; marginalize out θ by ignoring it

MCMC Simulation

- We see that α is the ratio of two quantities of the form

$$\beta_{n|m} = \frac{p(X \mid H_n,\theta_n)p(H_n,\theta_n)}{q(H_n,\theta_n)}$$

  where m and n refer to the two states
- Thus

$$\alpha = \frac{\beta_{n|m}}{\beta_{m|n}}$$

MCMC Simulation

- And, we might factorize both p and q:

$$\beta_{n|m} = \frac{p(X \mid H_n,\theta_n)p(\theta_n \mid H_n)p(H_n)}{q(\theta_n \mid H_n)q(H_n)}$$

  - Example: Coin tosses. We observe h heads and t tails and wish to determine whether the coin is fair, given this data
  - Choose prior on H_n, for example, $p(H_0)= p(H_1)= 1/2$
  - Choose a proposal, for example $q(H_0)= q(H_1)= 1/2$
  - Under H_0, $\theta\equiv 1/2$, whereas under H_1 there is a distribution $\pi(\theta \mid H_1)$ on θ
  - We need also to consider the proposals $q(\theta \mid H_n)$ for n=0,1

MCMC Simulation

- And, we might factorize both p and q:

$$\beta_{n|m} = \frac{p(X \mid H_n, \theta_n) p(\theta_n \mid H_n) p(H_n)}{q(\theta_n \mid H_n) q(H_n)}$$

  - An excellent choice of $q(\theta_n \mid H_n)$ (if possible) would be to make it proportional to the posterior $p(\theta_n|X,H_n)$! For then we would get

$$\beta_{n|m} \propto \frac{p(H_n)}{q(H_n)}$$

    which is a constant. Indeed, if we can also arrange things so that $\beta_{n|m}=1$ we would get an exact Gibbs sampler!
  - This can be accomplished approximately by using a small training sample and picking $q(H_n)$ using the results

MCMC Simulation

- And, we might factorize both p and q:

$$\beta_{n|m} = \frac{p(X \mid H_n, \theta_n) p(\theta_n \mid H_n) p(H_n)}{q(\theta_n \mid H_n) q(H_n)}$$

  - Usually the excellent strategy is not possible; if one can approximate the posterior by a distribution that you can sample from, that is a good strategy.
  - A simple-minded strategy would simply be to choose $q(\theta|H_1)$ flat on (0,1), whence

$$\beta_{1|m} = \frac{p(X, \theta_1 \mid H_1) p(H_1)}{q(H_1)}$$

    - » Works if the posterior is not too sharp (not much data). Not so good if there is a very sharp posterior

MCMC Simulation

- We now look at an R simulation for the coin-tossing problem. We presume 60 heads and 40 tails, which is almost rejected by classical p-value tests at a traditional 0.5 alpha-level test. (two-sided p-value is 0.057, whereas for 61 heads and 39 tails it is 0.035)
- We try in turn first the simple-minded strategy; then an approximate strategy; and finally an exact Gibbs strategy.

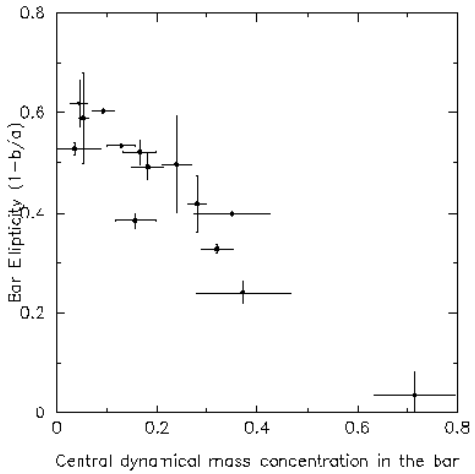
MCMC Simulation

- We note that the posterior probability (on equal prior probabilities on the models and a uniform prior on the parameter under model 2) actually slightly favors the simpler model 1, even though the p-value is nearly small enough to reject model 1
- This is one case where Bayesian and frequentist calculations give very different results.
- It is an example of the Bayesian Ockham's Razor

A Galaxy Problem

- A graduate student at Maryland, Mousumi Das, asked me to assist her on a problem involving galaxy data. She wanted to know if her data showed clear evidence of correlation, and if so, what the correlation was and how strong was the evidence for it.
- But there is an unusual feature of her problem: she knew that the data were imperfect, and for each data point had an error bar in both *x* and *y* (the "Errors-in-variables" case). Standard treatments of correlation do not address this situation.

The Data



Likelihood

- We know the distribution of the data x_i and y_i conditional on ξ_i and η_i, the (unknown) true parameters for each galaxy. Their joint distribution is given by

$$p(x_i, y_i \mid \xi_i, \eta_i, s_i, t_i) \propto \exp\left[-\frac{(x_i - \xi_i)^2}{2s_i^2}\right] \exp\left[-\frac{(y_i - \eta_i)^2}{2t_i^2}\right]$$

- Here s_i and t_i are the standard deviations of the data points, assumed known perfectly for this analysis (these are the basis of the error bars I showed earlier...).
- Since we do not know ξ_i and η_i for each galaxy but instead only the observed values x_i and y_i, we introduced the ξ_i and η_i for each galaxy as *latent variables*. These are parameters to be estimated.

Likelihood

- We can write down the full likelihood, the joint probability of observing the data, conditional on the latent parameters:

$$p(X, Y \mid \Xi, \mathrm{H}, S, T) \propto \prod_i p(x_i, y_i \mid \xi_i, \eta_i, s_i, t_i)$$

where X, Y, Ξ, H, S, and T are vectors whose components are x_i, y_i, ξ_i, η_i,s_i, and t_i, respectively.

Priors

- Assume that the underlying "true" (but unknown) galaxy parameters ξ_i and η_i (corresponding to the observed x_i and y_i) are distributed as a bivariate normal distribution

$$p(\xi_i,\eta_i \mid \rho,a,b,\sigma_\xi,\sigma_\eta) \propto \frac{1}{\sigma_\xi\sigma_\eta\sqrt{(1-\rho^2)}}$$

$$\times\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(\xi_i-a)^2}{\sigma_\xi^2}+\frac{(\eta_i-b)^2}{\sigma_\eta^2}-2\rho\frac{(\xi_i-a)(\eta_i-b)}{\sigma_\xi\sigma_\eta}\right)\right]$$

Priors

- This expression may be regarded as our prior on the latent variables ξ_i and η_i. It depends on other parameters ("hyperparameters") $\{a, b, \rho, \sigma_\xi, \sigma_\eta\}$.
- Here a and b give the true center of the distribution; ρ is the true correlation coefficient, and σ_ξ and σ_η are the true standard deviations of the distribution. None of these quantities are known. They are also parameters which must be estimated.
- The joint prior on all the latent variables ξ_i and η_i can be written as a product:

$$p(\Xi,\mathrm{H} \mid \rho,a,b,\sigma_\xi,\sigma_\eta) \propto \prod_i p(\xi_i,\eta_i \mid \rho,a,b,\sigma_\xi,\sigma_\eta)$$

Priors

- The next step is to assign priors for each of the remaining parameters, including the latent variables. Lacking special information, we chose conventional priors for all but Ξ and H. Thus, we assign
  - Improper constant flat priors on a and b.
  - Improper hierarchical independence Jeffreys priors $1/(a_\xi+\sigma_\xi)$ and $1/(a_\eta+\sigma_\eta)$ on σ_ξ and σ_η.
  - Priors on Ξ and H were displayed earlier

Bayesian Model Selection/Averaging

- This is a model selection problem (is there correlation or not?)
- Given models M_i, each of which depends on a *vector* of parameters ϑ_M, and given data Y, Bayes' theorem tells us that

$$p(\vartheta_M,M_i \mid Y) \propto p(Y \mid \vartheta_M,M_i)\,p(\vartheta_M \mid M_i)\,p(M_i),$$

- The probabilities $p\,(\vartheta_M \mid M_i)$ and $p\,(M_i)$ are the prior probabilities of the parameters given the model and of the model, respectively; $p\,(Y \mid \vartheta_M, M_i)$ is the likelihood function, and $p\,(\vartheta_M, M_i \mid Y)$ is the joint posterior probability distribution of the parameters and models, given the data.

Priors

- We must assign prior probabilities to each model, and also prior probability on ρ under each model
  - We have two models, one with correlation (M_1) and one without (M_0). We assigned $p(M_1)= p(M_0)=1/2$
  - We will compare M_1 and M_0 by computing their posterior probabilities. I chose the prior $p(\rho|M_1)$ on ρ to be flat and normalized on [–1,1] and zero elsewhere; we chose a delta-function prior $p(\rho|M_0)= \delta(\rho–0)$ under M_0

Posterior Distribution

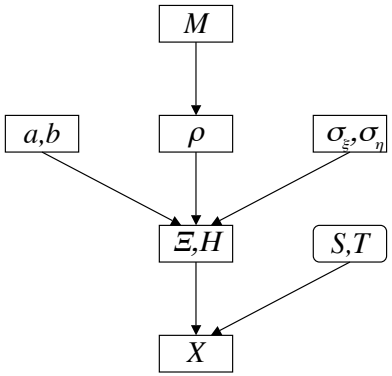
- The posterior distribution is proportional to the prior times the likelihood, as Bayes instructs us

$$p(\rho,a,b,\sigma_\xi,\sigma_\eta,\Xi,\Pi,M \mid X,Y,S,T)$$

$$\propto \frac{p(\rho \mid M)p(M)}{\sigma_\xi\sigma_\eta} \times p(\Xi,\mathrm{H} \mid \rho,a,b,\sigma_\xi,\sigma_\eta,M) \times p(X,Y \mid \Xi,\mathrm{H},S,T)$$

Posterior Distribution

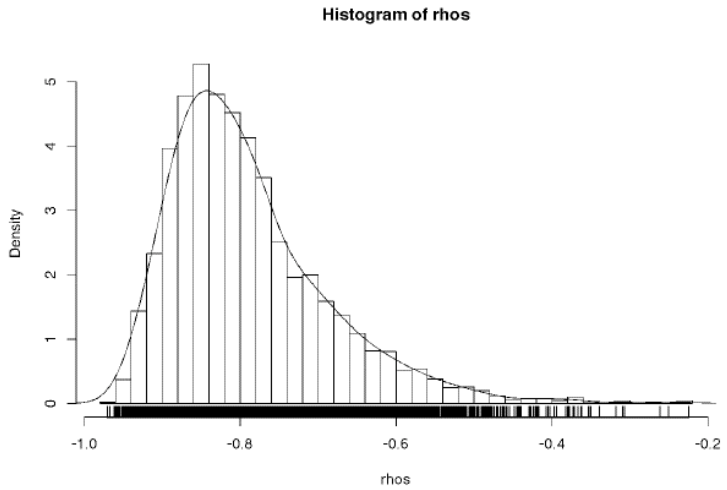
- Here is the DAG corresponding to the hierarchical model



Results

- For the full data set, we obtained
  - Odds on model with correlation = 207 (assumes prior odds equal to 1)
    - » [Without taking EIV into account, this would have been about 10 times larger…showing the importance of doing a proper EIV analysis]
  - Median rho = -0.81
  - Mean rho = -0.79 ± 0.10

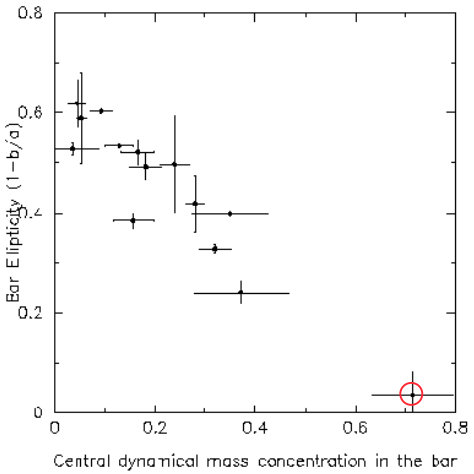
Posterior distribution of ρ



Comments on Data

- The student was concerned about how the lowest point affected any correlation. What would happen, she wondered, if they were not included in the sample?

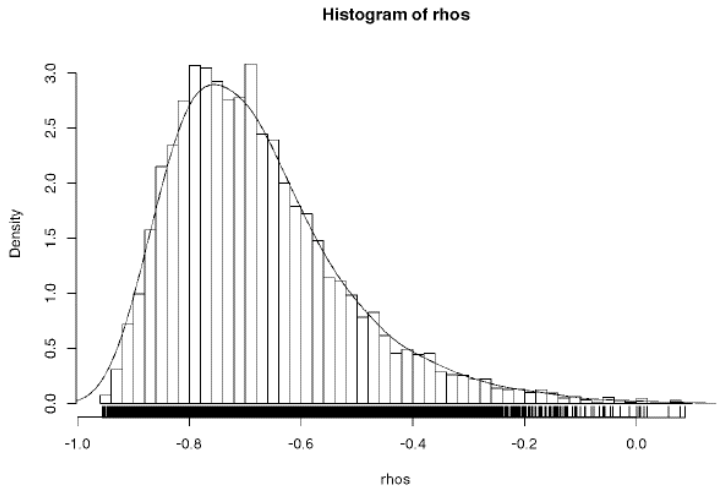
The Data



Results

- For the data set including the circled point, we obtained
  - Odds on model with correlation = 207 (assumes prior odds equal to 1)
  - Median rho = -0.81
  - Mean rho = -0.79 ± 0.10
- For the data set without the circled point we obtained
  - Odds on model with correlation = 9.9
  - Median rho = -0.70
  - Mean rho = -0.68 ± 0.16

Posterior distribution of ρ (Excluding 1 point)



Sampling Strategy for Our Problem

- To summarize:
  - We sampled the a, b, ξ_j, η_j in Gibbs steps (a and b appear in the posterior distribution as a bivariate normal distribution, as do the ξ_j, η_j).
  - We sampled σ_ξ, σ_η with M-H steps using symmetric uniform proposals centered on the current point, adjusting the maximum step for good mixing
  - We sampled ρ and M in a simultaneous reversible-jump M-H step, using a beta proposal on ρ with parameters tuned by experiment for efficiency and good mixing under the complex model, and with a proposal on M that also was chosen by experiment with an eye to getting an accurate estimate of the posterior odds on M.