There are no true statistical models.

1 Model Search, Selection, and Averaging.

Although some model selection procedures boil down to testing hypotheses about parameters and choosing the best parameter or a subset of parameters, model selection is a broader inferential task. It can be non-parametric, for example. Model selection sometimes can be interpreted as an estimation problem. If the competing models are indexed by $i \in \{1, 2, \ldots, m\}$, getting the posterior distribution of an index i would lead to the choice of best model, for example, the model that maximizes posterior probability of i.

1.1 When you hear hoofbeats, think horses, not zebras.

Ockham's razor is a logical principle attributed to the medieval philosopher and Franciscan monk William of Ockham. The principle states that one should not make more assumptions than the minimum needed. This principle is often called the *principle of parsimony.* It is essential for all scientific modeling and theory building.

Figure 1: Ockhams Razor: *Pluralitas non est ponenda sine necessitate.* Franciscan monk William of Ockham (ca. 1285-1349)

As Jefferys and Berger (1991) pointed out, the idea of measuring complexity and connecting the notions of complexity and prior probability goes back to Sir Harold Jeffreys' pioneering work on statistical inference in the 1920s. On page 47 of his classical work [11], Jeffreys says:

> Precise statement of the prior probabilities of the laws in accordance with the condition of convergence requires that they should actually be put in an order of decreasing prior probability. But this corresponds to actual scientific procedure. A physicist would test first whether the whole variation is random against the existence of a linear trend; than a linear law against a quadratic one, then proceeding in order of increasing complexity. All we have to say is that simpler laws have the greater prior probabilities. This is what Wrinch and I called the simplicity postulate. To make the order definite, however, requires a numerical rule for assessing the complexity law. In the case of laws expressible by differential equations this is easy. We would define the complexity of a differential equation, cleared of roots and fractions, by the sum of order, the degree, and the absolute values of the coefficients. Thus $s = a$ would be written

as $ds/dt = 0$ with complexity $1 + 1 + 1 = 3$. $s = a + ut + \frac{1}{2}gt^2$ would become $d^2s/dt^2 = 0$ with complexity $2 + 1 + 1 = 4$. Prior probability 2^{-m} of $6/\pi^2 m^2$ could be attached to the disjunction of all laws of complexity m and distributed uniformly among them.

In the spirit of Jeffreys' ideas, and building on work of Wallace and Boulton, Akaike, Dawid, Good, Kolmogorov, and others, Rissanen (1978) proposed the Minimum Description Length Principle (MDLP) as a paradigm in statistical inference. Informally, the MDLP can be stated as follows:

The preferred $\mathcal{M}$ for explaining observed data D is one that minimizes:

- the length of the description of the theory (Ockham's razor principle)
- the length of the description of the data with the help of the chosen theory.

Let C represent *some* measure of complexity. Then the above may be expressed, again informally, as:

Prefer the model $\mathcal{M}$, for which $C(\mathcal{M}) + C(D|\mathcal{M})$ is minimal.

In the above sentence we emphasized the word "some." Aside from the formal algorithmic definitions of complexity, which lack recursiveness, one can define a complexity measure by other means. The following example gives one way.

It is interesting that Bayes rule implies MDLP in the following way. For a Bayesian, the model $\mathcal{M}$ for which

$$P(\mathcal{M}|D) = \frac{P(D|\mathcal{M})P(\mathcal{M})}{P(D)} \tag{1}$$

is maximal, is preferred. Taking negative logarithms on both sides and considering the negative logarithm of probability as a measure of complexity, we get

$$\begin{aligned} -\log P(\mathcal{M}|D) &= -\log P(D|\mathcal{M}) - \log P(\mathcal{M}) + \log P(D) \\ &= C(D|\mathcal{M}) + C(\mathcal{M}) + \text{Const.} \end{aligned} \tag{2}$$

The Maximum Likelihood Principle (MLP) can also be interpreted as a special case of Rissanen's MDL principle. The ML principle says that, given the data, one should prefer a model that maximizes $P(D|\mathcal{M})$, or that minimizes complexity of the data under the model $-\log P(D|\mathcal{M})$, the first term in the right hand side of (2).

If the complexities of the models are constant, i.e., if their descriptions have the same length, then the MDL principle becomes the method of maximum likelihood (ML). From the MDLP standpoint, the ML is subjective, viewing all models to be of the same complexity.

Bayesian interpretation of the algorithmic complexity criterion (Barron-Cover, 1989). Let $X_1, X_2, \ldots, X_n$ be observed random variables from an unknown probability density we want to estimate. The class of candidates Γ is enumerable, and to each density f in the class Γ, the prior probability $\pi(f)$ is assigned. The "complexity" $C(f)$ of a particular density f is $-\log \pi(f)$.

The minimum over Γ of

$$C(f) + \log \frac{1}{\prod_k f(X_k)}$$

is equivalent to the maximum of $\pi(f) \prod_k f(X_k)$, which as a function of f, is proportional to the Bayes posterior probability of f given $X_1, \ldots, X_n$.

Remark: There is a connection between the Bayesian and the coding interpretations in that if π is a prior on Γ then $\log \frac{1}{\pi(f)}$ is the length (rounded up to integer) of the Shannon code for $f \in \Gamma$ based on the prior π. Conversely, if $C(f)$ is a codelength for a uniquely decodable code for f, then $\pi(f) = 2^{-C(f)}/D$ defines a proper prior probability ($D = \sum_{f\in\Gamma} 2^{-C(f)} \leq 1$ is the normalizing constant).

Let $\hat{f}_n$ be a minimum complexity density estimator. If the true density f is on the list Γ, then

$$(\exists n_0)(\forall n \geq n_0)\hat{f}_n = f.$$

Unfortunately, the number n_0 is not effective, i.e. given Γ that contains the true density and $X_1, \ldots, X_n$, we do not know if $\hat{f}_n$ is equal to f or not. Even when the true density f is not on the list Γ, we have the consistency of $\hat{f}_n$. Let $\bar{\Gamma}$ denote *the information closure* of Γ, i.e. the class of all densities f for which $\inf_{g\in\Gamma} D(f||g) = 0$, where $D(f||g)$ is the Kullback-Leibler distance between f and g. The following result holds [2]: If $\sum_{g\in\Gamma} 2^{-C(g)}$ is finite, and the true density is in $\bar{\Gamma}$, then

$$\lim_n \hat{P}_n(S) = P(S)$$

holds with probability 1, for all Borel sets S.

Wallace and Freeman (1987) propose a criterion similar to the Barron-Cover criterion for the case when Γ is a parametric class of densities.

Let $X_1, X_2, \ldots X_n$ be a sample from the population with density $f(x|\theta)$. Let $\pi(\theta)$ be a prior on θ.

The Minimum Message Length (MML) estimate is defined as

$$arg\min_\theta[-\log\pi(\theta) - \log\prod_{i=1}^n f(x_i|\theta) + \frac{1}{2}\log|\mathcal{I}(\theta)|]. \tag{3}$$

where $\mathcal{I}(\theta)$ is the appropriate information matrix. Note that this is equivalent to maximizing

$$\frac{\pi(\theta)\prod_{i=1}^n f(x_i|\theta)}{|\mathcal{I}(\theta)|^{1/2}}. \tag{4}$$

Interestingly, if the prior on θ is chosen to be the noninformative Jeffreys' prior, then the MML estimator reduces to ML estimator. Another nice property of the MML estimator is its invariance under 1-1 transformations.

Dividing by $|\mathcal{I}(\theta)|^{1/2}$ in (4) may not be what a Bayesian would do. In this case, instead of choosing the highest posterior mode, the MML estimator chooses the local posterior mode with the highest probability content, if it exists.

Example: Suppose a Bernoulli experiment gives m successes and $n - m$ failures. Assume the $\mathcal{B}eta(a, b)$ prior on θ. Then, $\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}$.

The MML estimator is a value that maximizes $\theta^{a+m-1/2}(1-\theta)^{b+n-m-1/2}$, i.e.

$$\theta' = \frac{a + m - \frac{1}{2}}{a + b + n - 1}. \tag{5}$$

Note that the Bayes estimator $\hat{\theta}_B = \frac{a+m}{a+b+n}$ slightly differs from the MML estimator.

Example: Another example of the application of MML criteria is a simple model selection procedure.

Let $\mathcal{P}_\mu = \{N(\mu, \sigma^2), \sigma^2 \text{ known}\}$. Select one of the hypotheses: $H_0 : \mu = \mu_0$, and $H_1 : \mu \neq \mu_0$, in light of data $x = (x_1, \ldots, x_n)$.

H_0 is parameter-free, the message length is $-\log f(x|\mu)$.

Let $\mu \sim \mathcal{U}(L\sigma, U\sigma)$. Then, assuming equal prior probabilities for H_0 and H_1, the hypothesis H_0 is preferred to H_1 if

$$z = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| < \sqrt{\log \frac{n\, e\, (U-L)^2}{12}}. \tag{6}$$

This is in contrast with the usual frequentist significance test in which the right-hand side of (6) has the constant $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

In the case of vague prior information on μ $(U - L \to \infty)$, the above criterion leads to a strong favoring of the simple hypothesis H_0, in the spirit of Jeffreys (1939).

Remark: O'Hagan (1987) proposed a modification of the MML estimator as follows: Estimate θ by the value $\hat{\theta}$ that maximizes

$$\frac{\pi(\theta|x)}{H(\theta, x)^{1/2}} \tag{7}$$

where $H(\theta, x) = -\frac{\partial^2}{\partial\theta^2} \log \pi(\theta|x)$. O'Hagan's modification is more in the Bayesian spirit, since everything depends only on the posterior. But the maximizing $\hat{\theta}$ may not be at any posterior mode, and in addition, the invariance property of the MML estimator is lost.

1.2 Bayes Factors Again.

Suppose that after observing data D, one wants to compare m competing models, $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$, which are specified by different parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$. By Bayes theorem, probability of model $\mathcal{M}_i$ is

$$p(\mathcal{M}_i|D) = \frac{p(D|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{j=1}^m p(D|\mathcal{M}_j)p(\mathcal{M}_j)}. \tag{8}$$

In (8) the value $p(D|\mathcal{M}_i)$ is obtained not by maximizing with respect to $\boldsymbol{\theta}_i$, but by averaging,

$$p(D|\mathcal{M}_i) = \int_{\Theta_i} p(D|\boldsymbol{\theta}_i, \mathcal{M}_i)p(\boldsymbol{\theta}_i|\mathcal{M}_i)d\boldsymbol{\theta}_i,$$

where $p(D|\boldsymbol{\theta}_i, \mathcal{M}_i)$ is the likelihood of $\boldsymbol{\theta}_i$ given the model $\mathcal{M}_i$.

Suppose that we compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The Posterior Odds are equal to Bayes Factor $\times$ Prior Odds,

$$\frac{p(\mathcal{M}_2|D)}{p(\mathcal{M}_1|D)} = \frac{p(D|\mathcal{M}_2)}{p(D|\mathcal{M}_1)} \times \frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)}.$$

The Bayes Factor in favor of model 2 compared to model 1 is

$$B_{21} = \frac{p(D|\mathcal{M}_2)}{p(D|\mathcal{M}_1)}.$$

1.2.1 BIC as an Approximation to Bayes Factor

When selecting a model, the use of Maximum Likelihood only leads to choosing the model of highest possible dimension. Akaike (1974) proposed subtracting the dimension of the model from the log likelihood, which is known as the AIC.[1] The AIC tends to overestimate the true model order. The first Bayesian model selection was proposed by Kashyap (1977), but what is now known as BIC is officially credited to Schwarz (1978).

Next we show that BIC can approximate Bayes factors using the large-sample heuristic. For details see Kass and Raftery (1995).

Consider $g(\boldsymbol{\theta}) = \log p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Then the Taylor expansion in the neighborhood of the point $\boldsymbol{\theta}^*$ is

$$g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)' g'(\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)' g''(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(||\boldsymbol{\theta} - \boldsymbol{\theta}^*||^2).$$

If $\boldsymbol{\theta}^*$ is the posterior mode, $g'(\boldsymbol{\theta}^*) = 0$, and

$$g(\boldsymbol{\theta}) \approx g(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)' g''(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

Now,

$$\begin{aligned} p(D) &= \int_\Theta p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int_\Theta \exp\{g(\boldsymbol{\theta})\}d\boldsymbol{\theta} \approx \exp\{g(\boldsymbol{\theta}^*)\} \int_\Theta \exp\left\{\frac{1}{2}\left((\boldsymbol{\theta} - \boldsymbol{\theta}^*)' g''(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right)\right\} d\boldsymbol{\theta}. \end{aligned}$$

The last expression is proportional to multivariate normal density and

$$p(D) \approx \exp\{g(\boldsymbol{\theta}^*)\}(2/\pi)^{p/2}\,|A|^{-1}\,,$$

where p is dimension of parameter space and $A = -g''(\boldsymbol{\theta}^*)$ and the error is of order $O(n^{-1})$.

Thus,

$$\log p(D) = \log p(D|\boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) + p/2\ \log(2\pi) - 1/2\ \log|A| + O(n^{-1}).$$

Now we switch to MLE's. In large samples $\boldsymbol{\theta}^* \approx \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is an MLE estimator, and $A \approx nI$ where I is the expected Fisher information matrix for a single observation,

$$-E\left[\frac{\partial^2 p(y_1|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right].$$

Since $|A| \approx n^p|I|$,

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + p/2\ \log(2\pi) - p/2\ \log n - 1/2\ \log|I| + O(n^{-1/2}).$$

Finally,

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) - p/2\ \log n + O(1), \tag{9}$$

[1] It is interesting that originally AIC stand for Information Criterion and letter A was added to the name IC since Fortran Language in the 1970's would not allow (without declaration) for the name of a real variable to start with any of the letters: I, J, K, L, M, N. Today, the acronym AIC evolved to Akaike Information Criterion.

where the error $O(1)$ can be improved to $O(n^{-1/2})$ if the prior is multivariate normal with mean at the MLE and covariance matrix equal to inverse Fisher information matrix. In the light of a selected model $\mathcal{M}$, the Log Posterior is approximately Penalized Log Likelihood at the MLE,

$$\log p(D|\mathcal{M}) \approx \log p(D|\hat{\boldsymbol{\theta}}, \mathcal{M}) - p/2 \ \log n.$$

The Schwarz BIC criteria is based on (10).

Consider B_{21}.

$$\log B_{21} \approx \log p(D|\hat{\boldsymbol{\theta}}_2, \mathcal{M}_2) - \log p(D|\hat{\boldsymbol{\theta}}_1, \mathcal{M}_1) - \frac{p_2 - p_1}{2} \log n,$$

i.e.,

$$2 \log B_{21} \approx \chi^2_{21} - (p_2 - p_1) \log n,$$

where χ^2_{21} is the ML test[2] for testing $\mathcal{M}_1$ against $\mathcal{M}_2$ and $p_2 - p_1$ is the number of degrees of freedom.

Let $\mathcal{M}_S$ be the full (saturated) model, i.e., the model that fits the data exactly. The deviance χ^2_{Sk} is the likelihood ratio statistics for $\mathcal{M}_k$ when compared to the saturated model $\mathcal{M}_S$. The BIC_k is defined as

$$BIC_k = \chi^2_{Sk} - df_k \times \log n,$$

Since $BIC_k \approx 2 \log B_{Sk}$ and $B_{jk} = \frac{B_{Sk}}{B_{Sj}}$, it follows that

$$2 \log B_{jk} \approx BIC_k - BIC_j.$$

Thus, when comparing two models $\mathcal{M}_j$ and $\mathcal{M}_k$ the difference of BIC values approximates twice the log Bayes factor.

1.2.2 DIC

In 1974 Dempster suggested examining the posterior distribution of classical deviance,

$$D(\theta) = -2 \log f(y|\theta) - 2 \log g(y),$$

where g is a (standardizing) function depending only on data.

Spiegelhalter et al. (2002) utilized Dempster's proposal and developed a criterion, now known as the DIC, deviance information criterion. DIC is defined as,

$$DIC = \bar{D} + p_d,$$

where $\bar{D}$ is the posterior expectation of deviance,

$$\bar{D} = \mathbb{E}^{\theta|y} D(\theta) = \mathbb{E}^{\theta|y} \left\{ -2 \log f(y|\theta) \right\}.$$

The term p_D is called *effective number of parameters*, and measures the complexity of the model. The effective number of parameters p_D is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the Bayes rule, $\hat{\theta}_B$, i.e.,

$$p_D = \bar{D} - D(\hat{\theta}_B) = \mathbb{E}^{\theta|y} D(\theta) - D(\mathbb{E}^{\theta|y}(\theta)) = \mathbb{E}^{\theta|y} \left\{ -2 \log f(y|\theta) \right\} - 2 \log f(y|\hat{\theta}_B).$$

[2] Let L_1 be the maximum value of the likelihood for an unrestricted set parameters with maximum likelihood estimates substituted for these parameters. Let L_0 be the maximum value of the likelihood when the parameters are restricted (and reduced in number). Assume that k parameters were lost (i.e., L_0 has k parameters less than L_1). The ratio $\lambda = L_0/L_1$ is always between 0 and 1 and the less likely the restricted model is, the smaller λ will be. This is quantified by $\chi^2 = -2 \log \lambda$ which has an approximate χ^2_k distribution. Critical for restricted model are large values of χ^2.

1.2.3 Bayes Factors from MCMC Traces

An approach to approximate Bayes factors from MCMC traces was developed by Chib (1995). This approach is applicabile in a wide range of settings (Han and Carlin 2000).

1.3 SSVS - Stochastic Search Variable Selection of George and McCulloch.

One of the key issues in model building and model selection is what variables should be included. Suppose that the adopted model is linear in predictors

$$X_1, X_2, \ldots, X_p,$$

and that a possible selected model is

$$y = \beta_1^* X_1^* + \cdots + \beta_q^* X_q^*,$$

where $\{X_1^*, \ldots, X_q^*\}$ is a subset from $\{X_1, \ldots, X_p\}$. Typical classical solutions are based on criteria such as AIC, BIC, Mallows' C_p, and their numerous variants. When p is small, typically less than 10, the above model choice is feasible, but when p is large (hundreds, or thousands) the alternative is to consider stepwise procedures.

Stochastic Search Variable Selection (SSVS) by George and McCulloch (1993) finds a promising subset of predictors in a Bayesian fashion.

SSVS Method puts the probability distribution on the set of all models such that the most appropriate models are given the highest posterior probability. This approach is effective even if p is large (order of thousands) and n - number of observation is small ($n \ll p$). The number of possible models on the set of p predictors is the cardinality of the partition set of $\{1, 2, \ldots, p\}$, i.e., 2^p.

1.3.1 Model

Assume that

$$[Y] \sim \mathcal{MVN}_n(X\beta, \sigma^2 I_n),$$

where n is the number of observations, $Y = (y_1, \ldots, y_n)'$, $X = (X_1, \ldots, X_p)$ is a $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)'$ is the vector of parameters (coefficients), σ^2 is a scalar, and I_n is the $n \times n$ identity matrix. The components β_i are modeled as

$$\begin{aligned} \pi(\beta_i|\gamma_i) &= (1-\gamma_i)\mathcal{N}(0, \tau_i^2) + \gamma_i \mathcal{N}(0, c_i^2 \tau_i^2), \\ P(\gamma_i = 1) &= 1 - P(\gamma_i = 0) = \omega_i, \end{aligned}$$

where $\tau_i^2 \ll c_i^2 \tau_i^2$. Thus, the component $\mathcal{N}(0, \tau_i^2)$ is concentrated about 0, while the component $\mathcal{N}(0, c_i^2 \tau_i^2)$ is diffuse. When $\gamma_i = 0$, the prior is concentrated, reflecting that the coefficient is close to zero, with small variation about 0, and when $\gamma_i = 1$, the prior accommodates non-zero coefficients. The variable/model selection is achieved by selecting X_i for which the corresponding γ_i have maximal posterior probability of being 1.

What are some conditionals in the model? Given γ, β is normal,

$$[\beta|\gamma] \sim \mathcal{MVN}_p(0, D_\gamma^2),$$

where D_γ is a diagonal matrix with its ith diagonal element equal to

$$d_{ii} = (1-\gamma_i)\tau_i + \gamma_i c_i \tau_i,$$

so that β_i are independent, given γ. An efficient prior on γ is the independent Bernoulli prior,

$$[\gamma] \sim \prod_{i=1}^{p} \omega_i^{\gamma_i}(1-\omega_i)^{1-\gamma_i}. \tag{10}$$

Especially important is $\pi(\gamma) = \frac{1}{2^p}$ which is in fact (10) for $\omega_i = 1/2$.

Furthermore,

$$[\sigma^2] \sim \mathcal{G}amma\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right),$$

so that the prior expectation of σ^{-2} is $1/\lambda$ and the prior variance is $2/(\nu\lambda^2)$. Selecting ν small corresponds to vague information about σ^{-2}.

The full conditionals are as follows:

$$\begin{aligned}
[\beta|\sigma^2,\gamma,Y] &\sim \mathcal{MVN}_n\left((X'X+\sigma^2 D_\gamma^{-2})^{-1}X'Y, \sigma^2(X'X+\sigma^2 D_\gamma^{-2})^{-1}\right), \\
\left[\sigma^2|\beta,Y\right] &\sim \mathcal{G}amma\left(\frac{n+\nu}{2}, \frac{||Y-X\beta||^2+\nu\lambda}{2}\right), \\
[\gamma_i|\beta,\gamma_{\neq i}] &\sim \mathcal{B}er\left(\frac{a}{a+b}\right), \text{ where} \\
& \quad a = \pi(\beta|\gamma_{\neq i},\gamma_i=1)\pi(\gamma_{\neq i},\gamma_i=1), \text{ and} \\
& \quad b = \pi(\beta|\gamma_{\neq i},\gamma_i=0)\pi(\gamma_{\neq i},\gamma_i=0).
\end{aligned}$$

Under Bernoulli's prior (10)

$$[\gamma_i|\beta_i] \sim \mathcal{B}er\left(\frac{a}{a+b}\right),$$

where

$$a = \pi(\beta_i|\gamma_i=1)\omega_i, \quad \text{and } b = \pi(\beta_i|\gamma_i=0)(1-\omega_i).$$

Thus, γ_i are generated one step at a time.

The simulated sequence $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(m)}, \ldots$ after a burn-in period simulates draws from $\pi(\gamma|Y)$. Since a change in γ_i corresponds to inclusion/exclusion of variable X_i, generating the MCMC sequence of γ's corresponds to stochastic search.

The SSVS can be adapted to GLM's and selection between exchangeable regressions.

Example 1. In the BUGS program that follows the problem is selecting the best subset of variables. Only 10 observations are generated from the model

$$y_i = -2.2X_{i,3} + 4X_{i,5} + 2.46X_{i,7} + \epsilon_i, \; i = 1, \ldots, 10$$

where $\epsilon_i \sim \mathcal{N}(0,1)$. The design matrix of size 10×7 is fixed,

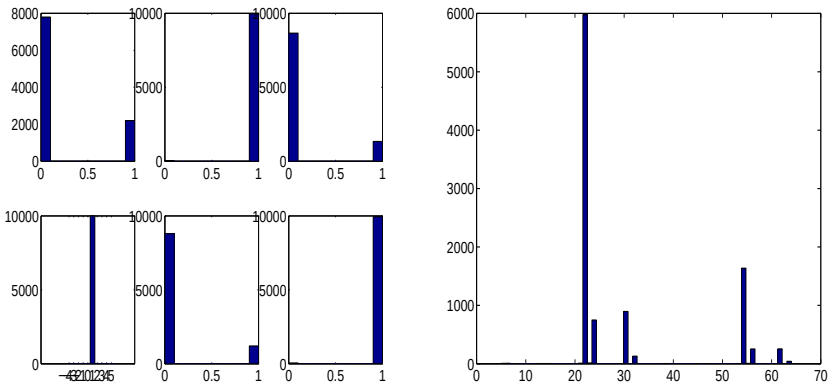
$$X = (X_{i,j}) = \begin{pmatrix} 5 & 1 & 3 & 7 & 9 & 0 & 3 \\ 4 & 4 & 3 & 4 & 4 & 6 & 4 \\ 3 & 1 & 1 & 1 & 2 & 7 & 7 \\ 0 & 7 & 0 & 6 & 9 & 2 & 1 \\ 4 & 6 & 3 & 8 & 9 & 7 & 5 \\ 7 & 4 & 8 & 1 & 5 & 2 & 2 \\ 1 & 4 & 0 & 6 & 8 & 4 & 0 \\ 0 & 5 & 2 & 1 & 1 & 8 & 2 \\ 3 & 5 & 9 & 6 & 3 & 8 & 6 \\ 0 & 5 & 2 & 8 & 6 & 1 & 2 \end{pmatrix}$$

and the response is $Y = (37.7594, 18.9744, 22.4716, 38.3637, 40.3193, 6.5916, 33.8860, 1.5786, 7.9400, 23.3282)'$, We start with full model

$$y_i = \sum_{j=1}^{7} \beta_j X_{i,j} + \epsilon_i,$$

Although 7 variables would give $2^7 = 128$ different models, we restrict attention to models not containing variable 1 (An oracle information by a birdie). That leaves 64 models and we give equal prior probability of 1/64 to each.

The rest of the set-up follows the SSVS theory – see the BUGS file. (File `ssvs.odc` is attached on the course web page). Figure 2 gives histograms of γ_2-γ_7 and posterior visits to different models.



(a) (b)

Figure 2: (a) Posterior simulations of γ_2-γ_7 (γ_1 excluded). Note that histograms of γ_3, γ_5, and γ_7 are concentrated at 1. In fact all simulated γ_5's are 1; (b) Number of visits to different models. Model number 22 [$\gamma_1 = 0, \gamma_2 = 0, \gamma_2 = 1, \gamma_4 = 0, \gamma_5 = 1, \gamma_6 = 0, \gamma_1 = 1,$] is the right model. The true model has most aposteriori visits.

The table below gives posterior estimators of the parameters $\beta_i,\ i = 1, \ldots, 7$.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
beta[1]	-0.04885	0.267	0.003264	-0.5551	-0.05317	0.4917	1001	10000
beta[2]	-0.299	0.3087	0.004812	-0.955	-0.2904	0.2736	1001	10000
beta[3]	-2.062	0.2762	0.004363	-2.584	-2.063	-1.551	1001	10000
beta[4]	0.0919	0.2706	0.002707	-0.4091	0.08264	0.6209	1001	10000
beta[5]	4.118	0.291	0.003671	3.525	4.131	4.658	1001	10000
beta[6]	0.02058	0.2292	0.003009	-0.4216	0.01738	0.466	1001	10000
beta[7]	2.295	0.3798	0.006879	1.533	2.306	2.992	1001	10000

```
SSVS of George and McCulloch
        after  Francesca's SSVSsim.b
% data generation (matlab )
% the model includes 3rd, 5th and 7th variable.
% beta_3 = -2.2; beta_5=4; beta_7 = 2.46
% tau = 1;
clear all;
rand('seed',1)
randn('seed',1)
  x=floor(10 * rand(10,7) );
  Y = -2.2 * x(:,3) + 4.0 * x(:,5) + 2.46 * x(:,7) +  randn(10,1);

 #Model Begin
model {
        for ( i  in 1:nn ) {     #nn is sample size
            Y[i] ~ dnorm(mu[i],tau)
            mu[i] <- beta[1]*x[i,1]+beta[2]*x[i,2]+beta[3]*x[i,3]+
           beta[4]*x[i,4]+beta[5]*x[i,5]+beta[6]*x[i,6] +beta[7]*x[i,7]
             }
                   #------------
                   for (j in 1:7) {  #j is the  index of variables
                        gamma[j] <- g[j,k]    #k is model number
                        beta[j] ~ dnorm(0,tau.beta[j])
                        tau.beta[j] <- equals(gamma[j],1)*tau.1+equals(gamma[j],0)*tau.2
                                        }
#Priors:
 tau ~ dgamma(0.001,0.001)
# discrete prior on set of all N models
  for (n in 1:N) {
                   prior[n] <- 1/N}
                   k ~ dcat(prior[]);
# alternating precision parameters  in tau.beta
  tau.1 <- 0.1
  tau.2 <- 10
#  Specification of  models by gamma. The decision maker believes that variable #1 is NOT
#  in the model, however the rest 6 variables give 2^6=64 plausible models that are apriori
#  assumed equiprobable
g[1,1]<-0; g[2,1]<-0; g[3,1]<-0; g[4,1]<-0; g[5,1]<-0; g[6,1]<-0; g[7,1]<-0; #0000000
g[1,2]<-0; g[2,2]<-0; g[3,2]<-0; g[4,2]<-0; g[5,2]<-0; g[6,2]<-0; g[7,2]<-1; #0000001
g[1,3]<-0; g[2,3]<-0; g[3,3]<-0; g[4,3]<-0; g[5,3]<-0; g[6,3]<-1; g[7,3]<-0; #0000010
...
g[1,21]<-0; g[2,21]<-0; g[3,21]<-1; g[4,21]<-0; g[5,21]<-1; g[6,21]<-0; g[7,21]<-0; #0010100
g[1,22]<-0; g[2,22]<-0; g[3,22]<-1; g[4,22]<-0; g[5,22]<-1; g[6,22]<-0; g[7,22]<-1; #0010101*****
g[1,23]<-0; g[2,23]<-0; g[3,23]<-1; g[4,23]<-0; g[5,23]<-1; g[6,23]<-1; g[7,23]<-0; #0010110
...
g[1,63]<-0; g[2,63]<-1; g[3,63]<-1; g[4,63]<-1; g[5,63]<-1; g[6,63]<-1; g[7,63]<-0; #0111110
g[1,64]<-0; g[2,64]<-1; g[3,64]<-1; g[4,64]<-1; g[5,64]<-1; g[6,64]<-1; g[7,64]<-1; #0111111
}
```

```
  #Model End


  #Data Begin
list(
nn=10, #sample size
N=64,  #number of apriori plausible models
Y=c(       37.7594,
           18.9744,
           22.4716,
           38.3637,
           40.3193,
            6.5916,
           33.8860,
            1.5786,
            7.9400,
           23.3282),
x=structure(.Data=c(
     5,     1,     3,    7,     9,     0,     3,
     4,     4,     3,    4,     4,     6,     4,
     3,     1,     1,    1,     2,     7,     7,
     0,     7,     0,    6,     9,     2,     1,
     4 ,    6,     3,    8,     9,     7,     5,
     7,     4 ,    8,    1,     5,     2,     2,
     1,     4 ,    0,    6,     8,     4,     0,
     0,     5,     2,    1,     1,     8,     2,
     3,     5 ,    9,    6,     3,     8,     6,
     0,     5,     2,    8,     6,     1,     2),
.Dim=c(10,7)))
   #Data End


    #Inits Begin
list(
tau=0.4,
k=4)
    #Inits End
```

Example 2. The wavelet shrinkage approach used by Clyde, Parmigiani and Vidakovic (1998) is based on a limiting form of the conjugate SSVS prior in George and McCulloch (1994). Clyde *et al.* (1996) consider a prior for $\underset{\sim}{\theta}$ which is a mixture of a point mass at 0 if the variable is excluded from the wavelet regression and a normal distribution if it is included,

$$[\theta|\gamma_j, \sigma^2] \sim N(0, (1-\gamma_j) + \gamma_j c_j \sigma^2). \tag{11}$$

The γ_j are indicator variables that specify which basis element or column of W should be selected. As before, the subscript j points to the level to which θ belongs. The set of all possible vectors $\underset{\sim}{\gamma}$ will be referred to as the subset space. The prior distribution for σ^2 is inverse χ^2 i.e.

$$[\lambda\nu/\sigma^2] \sim \chi^2_\nu,$$

where λ and ν are fixed hyperparameters and the γ_j's are independently distributed as Bernoulli (p_j) random variables.

The posterior mean of $\underset{\sim}{\theta}|\underset{\sim}{\gamma}$ is

$$E(\underset{\sim}{\theta}|\underset{\sim}{d},\underset{\sim}{\gamma}) = \Gamma(I_n + C^{-1})^{-1}\underset{\sim}{d} \tag{12}$$

where Γ and C are diagonal matrices with γ_{jk} and c_{jk} respectively on the diagonal and 0 elsewhere. For a particular subset determined by the ones in $\underset{\sim}{\gamma}$ (12) corresponds to thresholding with linear shrinkage.

The posterior mean is obtained by averaging over all models. Model averaging leads to a multiple shrinkage estimator of $\underset{\sim}{\theta}$:

$$E(\underset{\sim}{\theta}|\underset{\sim}{d}) = \sum_{\underset{\sim}{\gamma}} \pi(\underset{\sim}{\gamma}|\underset{\sim}{d})\Gamma\left(I_n + C^{-1}\right)^{-1}\underset{\sim}{d}, \tag{13}$$

where $\pi(\underset{\sim}{\gamma}|\underset{\sim}{d})$ is the posterior probability of a particular subset $\underset{\sim}{\gamma}$.

An additional nonlinear shrinkage of the coefficients to 0 results from the uncertainty in which subsets should be selected.

Calculating the posterior probabilities of $\underset{\sim}{\gamma}$ and the mixture estimates for the posterior mean of $\underset{\sim}{\theta}$ above involve sums over all 2^n values of $\underset{\sim}{\gamma}$. The calculational complexity of the mixing is prohibitive even for problems of moderate size, and either approximations or stochastic methods for selecting subsets $\underset{\sim}{\gamma}$ possessing high posterior probability must be used.

In the orthogonal case, Clyde, DeSimone, and Parmigiani (1996) obtain an approximation to the posterior probability of $\underset{\sim}{\gamma}$ which is adapted to the wavelet setting in [5]. The approximation can be achieved by either conditioning on σ (plug-in approach) or by assuming independence of the elements in $\underset{\sim}{\gamma}$.

The approximate model probabilities, for the conditional case, are functions of the data through the regression sum of squares and are given by:

$$\begin{aligned} \pi(\underset{\sim}{\gamma}|\underset{\sim}{d}) \approx \tilde{\pi}(\underset{\sim}{\gamma}|y) &= \prod_{j,k} \rho_{jk}^{\gamma_{jk}}(1-\rho_{jk})^{1-\gamma_{jk}} \\ \rho_{jk}(\underset{\sim}{d},\sigma) &= \frac{a_{jk}(\underset{\sim}{d},\sigma)}{1+a_{jk}(\underset{\sim}{d},\sigma)} \end{aligned} \tag{14}$$

where

$$\begin{aligned} a_{jk}(\underset{\sim}{d},\sigma) &= \frac{p_{jk}}{1-p_{jk}}(1+c_{jk})^{-1/2}\cdot\exp\left\{\frac{1}{2}\frac{S_{jk}^2}{\sigma^2}\right\} \\ S_{jk}^2 &= d_{jk}^2/(1+c_{jk}^{-1}). \end{aligned}$$

The p_{jk} can be used to obtain a direct approximation to the multiple shrinkage Bayes rule. The independence assumption leads to more involved formulas. Thus, the posterior mean for θ_{jk} is approximately

$$\rho_{jk}(1+c_{jk}^{-1})^{-1}d_{jk}. \tag{15}$$

Equation (15) can be viewed as a level dependent wavelet shrinkage rule, generating a variety of nonlinear rules. Depending on the choice of prior hyperparameters shrinkage may be monotonic, if there are no level dependent hyperparameters, or non-monotonic; see Figure 3. Authors report good MSE performance of approximation rules.

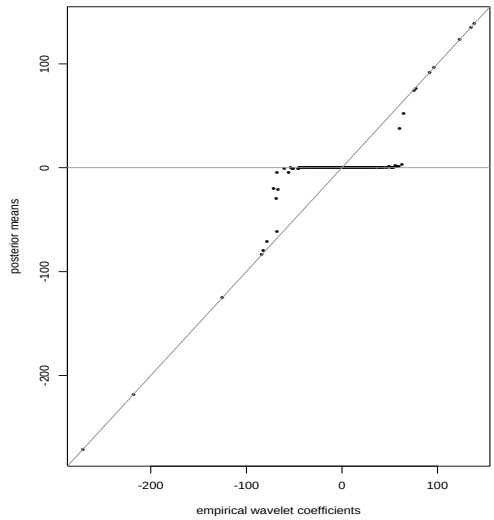


Figure 3: Shrinkage rule from [5] based on independence approximation (15)

1.4 Bayesian Model Averaging.

Epicurus Principle of Multiple Explanations: *Keep all models that are consistent with the data.*

Bayesian Model Averaging (BMA) has for its goal to account for model uncertainty, the same way diversification of an investment portfolio has for its goal to account for the stock market uncertainties. BMA uses Bayes Theorem and averages the models by their posterior probabilities. The BMA improves predictive performance, and is theoretically elegant, but could be computationally costly.

A simplistic explanation for the predictive success of BMA follows from the observation that BMA predictions are weighted averages of predictions coming from various models. If the individual predictions are approximately (or exactly) unbiased estimates of the same quantity, then averaging will tend to reduce unwanted variance.

BMA weights each single model prediction by its corresponding posterior model probability. Thus, BMA uses the data to adaptively increase the influence of those predictions whose models are more supported by the data.

Suppose that Δ is a quantity of interest (size of an effect, the future predictive observation, the precision of an estimator, the utility of a course of an action, etc.) The posterior distribution of Δ, given the data D is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\delta|\mathcal{M}_k, D) \times p(\mathcal{M}_k|D).$$

This is an average of posterior distributions under each model considered. Here we consider the set of K models, $cM_1, cM_2, \ldots, cM_K$.

The model weights are posterior probabilities of the models,

$$p(\mathcal{M}_k|D) = \frac{p(D|\mathcal{M}_k) \times p(\mathcal{M}_k)}{\sum_{i=1}^{K} p(D|\mathcal{M}_i) \times p(\mathcal{M}_i)},$$

where

$$p(D|\mathcal{M}_k) = \int p(D|\theta_k, \mathcal{M}_k) p(\theta_k|\mathcal{M}_k) d\theta_k$$

Example: To be added.

1.5 Models of Varying Dimensions. Reversible Jump MCMC.

To be added.

References

[1] Akaike, H. (1974) A New Look at the Statistical Identification Model, *IEEE Trans. Auto Control,* Vol. AC-19, pp. 716–723.

[2] BARRON, A. R. and COVER, T. M. (1989). Minimum complexity density estimation. *University of Illinois, Statistical Department, Technical Report #28.*

[3] Chib, S. (1995). Marginal likelihood from the gibbs output. Journal of the American Statistical Association 90 (432), 1313-1321.

[4] Clyde, M., DeSimone, H. and Parmigiani,G. (1996). Prediction via Orthogonalized Model Mixing. *Journal of the American Statistical Association*, **91**, 1197–1208.

[5] Clyde, M., Parmigiani, G., Vidakovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets, *Biometrika*,**85**, 391–402.

[6] George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.*, **88**, 881–889.

[7] Good, I. J. (1968). Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *British J. Philos. Sci.* **19** *123-143.*

[8] Han, C. and B. Carlin (2000). MCMC methods for computing Bayes factors: A comparative review.

[9] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky (1999). Bayesian Model Averaging: A Tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors *Statist. Sci.*,**14**, 382-417.

[10] Jefferys, W. H. and Berger, J. O. (1991). Sharpening Ockham's razor on Bayesian strop. *Technical Report #91-44C, Statistical Department, Purdue University*

[11] Jeffreys, H. (1939). *Theory of probability* Clarendon Press. Oxford 1985.

[12] Kashyap, R. L. (1977). A Bayesian Comparison of Different Classes of Dynamic Models Using Empirical Data, *IEEE Trans. Auto Control,* Vol. AC-22, No. 5, pp. 715–727.

[13] Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

[14] O'Hagan, A. (1987). Discussion of the papers by Dr Rissanen and Professors Wallace and Freeman. *J. R. Statist. Soc. B.* **49** *3, 256–257.*

[15] Rissanen, J. (1978). Modeling by the shortest data description. *Automatica* **14** *465–471.*

[16] Rissanen, J. (1982). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** *416–431.*

[17] Schwarz, G. (1978). Estimating the Dimension of a Model,*The Annals of Statistics,* Vol. 5, No. 2, pp 461–464.

[18] Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **49** *240–265.*

[19] Wallace, C. S. and Boulton, D. M. (1975). An invariant Bayes method for point estimation. *Classification Soc. Bull.* **3** *11–34.*