CPSC 535 - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

30th January 2007

1.1– Outline

- Bayesian linear model and model selection.
- Bayesian linear model and variable selection.
- Extensions.

1.2– Summary of Last Lecture

- Assume you have some data and two statistical models.

Under H_0, $\theta_0 \in \Theta_0$, the prior is $\pi_0(\theta_0)$ and the likelihood is $f_0(x|\theta_0)$.
Under H_1, $\theta_1 \in \Theta_1$, the prior is $\pi_1(\theta_1)$ and the likelihood is $f_1(x|\theta_1)$ then

$$B_{10}^{\pi} = \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\,d\theta_1}{\int f_0(x|\theta_0)\pi_0(\theta_0)\,d\theta_0}$$

- One can have $\Theta_0 = \mathbb{R}$ and $\Theta_1 = \mathbb{R}^{1000}$.

- In this case you have a parameter space $\Theta = \Theta_0 \cup \Theta_1$ which is the union two subspaces of different dimensions.

1.2– Summary of Last Lecture

- A simple way to think about it is to say that we put a prior on $\{H_0\} \times \Theta_0 \cup \{H_1\} \times \Theta_1$ where

$$\pi(H,\theta) = \pi(H)\,\pi(\theta|\,H) = \begin{cases} \pi(H_0)\,\pi_0(\theta) & \text{if } H = H_0 \\ \pi(H_1)\,\pi_1(\theta) & \text{if } H = H_1 \end{cases}$$

- The likelihood is given by

$$f(x|\,H,\theta) = \begin{cases} f_0(x|\,\theta) & \text{if } H = H_0 \\ f_1(x|\,\theta) & \text{if } H = H_1 \end{cases}.$$

1.2– Summary of Last Lecture

- The posterior is defined by

$$\pi\left(H,\theta\right|x) = \frac{\pi\left(H,\theta\right) f\left(x\right| H,\theta)}{\int_{\Theta_0} \pi\left(H_0\right)\pi_0\left(\theta\right) f_0\left(x\right|\theta) \, d\theta + \int_{\Theta_1} \pi\left(H_1\right)\pi_1\left(\theta\right) f_1\left(x\right|\theta) \, d\theta}.$$

- Clearly by marginalization, we obtain

$$\pi\left(H_1\right|x) = \int_{\Theta_1} \pi\left(H_1,\theta\right|x) \, d\theta$$

and

$$\pi\left(\theta\right| x, H_1).$$

- **Be careful**: conditional expectations of the following form

$$\mathbb{E}\left[\theta\right|x]$$

can have no physical meaning; i.e. you can average variables having different meanings for H_0 and H_1.

2.1– Bayesian Polynomial Regression Example

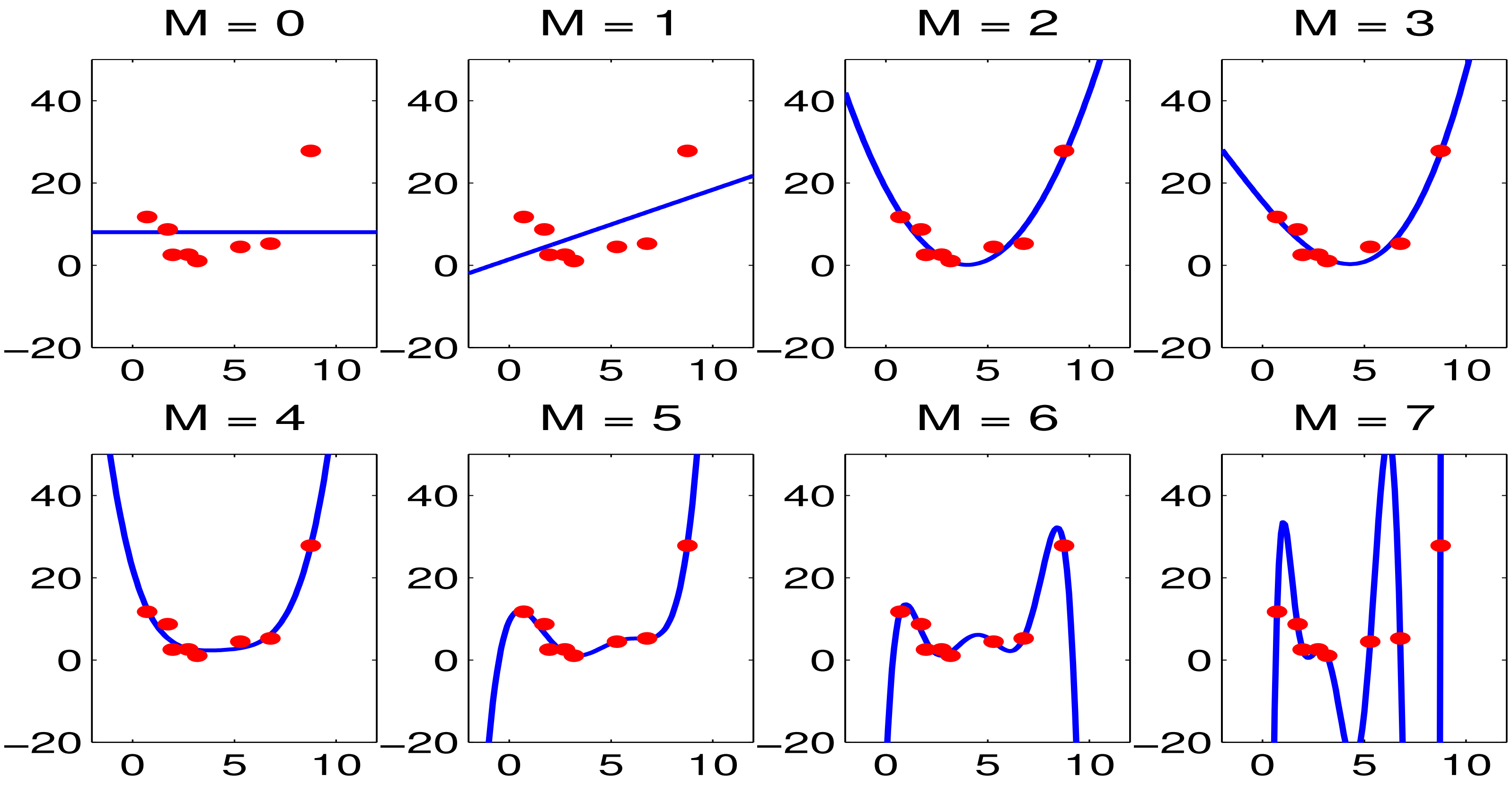
- In practice, you might have more than 2 potential models/hypothesis for your data.

- Consider the following polynomial regression problem where $D = \{x_i, y_i\}_{i=1}^{n}$ where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$.

$$
\begin{aligned}
Y &= \sum_{k=0}^{M} \beta_i X^k + \sigma V \text{ where } V \sim \mathcal{N}(0,1) \\
&= \beta_{0:M}^{\mathrm{T}} f_M(X) + \sigma V
\end{aligned}
$$

- Here the problem is that if M is too large then there will be overfitting.

As M increases, the model overfits.



2.1– Bayesian Polynomial Regression Example

- Candidate Bayesian models H_M for $M \in \{0, ..., M_{\max}\}$.

- For the model H_M, we take the prior $\pi_M\left(\beta_{0:M}, \sigma^2\right)$

$$\begin{aligned} \pi_M\left(\beta_{0:M}, \sigma^2\right) &= \pi_M\left(\beta_{0:M} \middle| \sigma^2\right) \pi_M\left(\sigma^2\right) \\ &= \mathcal{N}\left(\beta_{0:M}; 0, \delta^2 \sigma^2 I_{M+1}\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- We have the following Gaussian likelihood

$$f\left(D \middle| \beta_{0:M}, \sigma^2\right) = \prod_{i=1}^{n} \mathcal{N}\left(y_i; \beta_{0:M}^{\mathrm{T}} f_M\left(x_i\right), \sigma^2\right)$$

- Standard calculations (see assignement) yield

$$
\begin{aligned}
\pi_M\left(\beta_{0:M},\sigma^2\middle| D\right) &= \mathcal{N}\left(\beta_{0:M};\mu_M,\sigma^2\Sigma_M\right) \\
&\quad \times\mathcal{IG}\left(\sigma^2;\frac{\nu_0+n}{2},\frac{\gamma_0+\sum_{i=1}^n y_i^2-\mu_M^{\mathrm{T}}\Sigma_M^{-1}\mu_M}{2}\right)
\end{aligned}
$$

where

$$\mu_M=\Sigma_M\left(\sum_{i=1}^n y_i f_M\left(x_i\right)\right),\ \Sigma_M^{-1}=\delta^{-2}I_{M+1}+\sum_{i=1}^n f_M\left(x_i\right)f_M^{\mathrm{T}}\left(x_i\right)$$

knowing that

$$\mathcal{IG}\left(\sigma^2;\alpha,\beta\right)=\frac{\beta^\alpha}{\Gamma\left(\beta\right)}\frac{1}{\left(\sigma^2\right)^{\alpha+1}}\exp\left(-\frac{\beta}{\sigma^2}\right).$$

- The marginal likelihood/evidence is given by

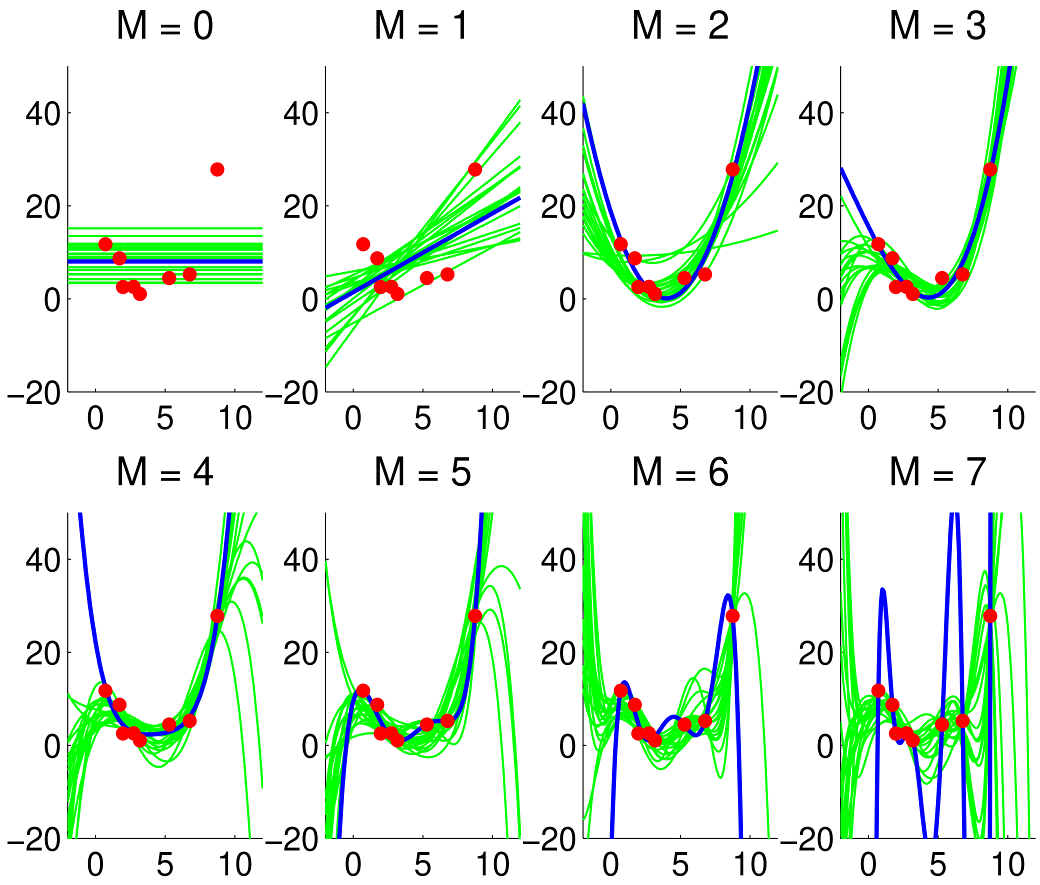
$$\pi\left(D\middle|\,H_M\right)=\int f\left(D\middle|\,\beta_{0:M},\sigma^2\right)\pi_M\left(\beta_{0:M},\sigma^2\right)d\beta_{0:M}d\sigma^2$$

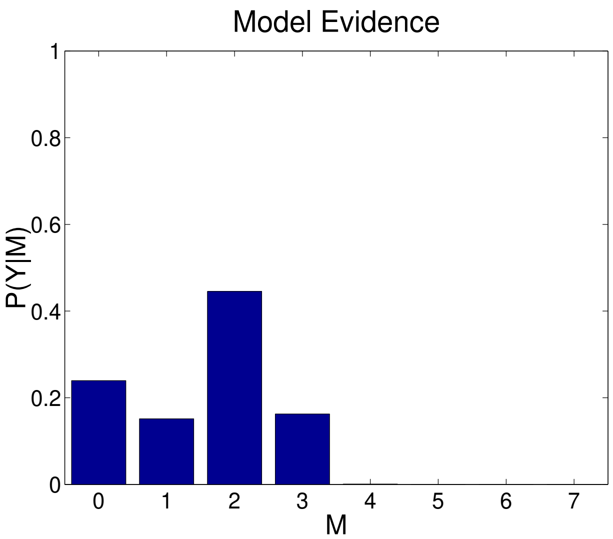
$$=\Gamma\left(\frac{\nu_0+n}{2}+1\right)\delta^{-(M+1)}\left|\Sigma_M\right|^{1/2}\left(\frac{\gamma_0+\sum_{i=1}^n y_i^2-\mu_M^{\mathrm{T}}\Sigma_M^{-1}\mu_M}{2}\right)^{-\left(\frac{\nu_0+n}{2}+1\right)}$$

- We can also compute

$$\pi\left(H_M\middle|\,D\right)=\frac{\pi\left(D\middle|\,H_M\right)\pi\left(H_M\right)}{\sum_{i=0}^{M_{\max}}\pi\left(D\middle|\,H_i\right)\pi\left(H_i\right)}$$

2.1– Bayesian Polynomial Regression Example





- We have assumed here that δ^2 was fixed and set to $\delta^2 = 1$.

- As $\delta^2 \to \infty$, the prior on $\beta_{0:M}$ is getting vague but then

$$\lim_{\delta^2 \to \infty} \pi\left(H_0 \middle| D \right) = 1$$

as for $M \geq 1$

$$\frac{\pi\left(D \middle| H_0 \right)}{\pi\left(D \middle| H_M \right)} = \frac{\delta^{-1} \left|\Sigma_0\right|^{1/2} \left(\frac{\gamma_0 + \sum_{i=1}^n y_i^2 - \mu_0^{\mathrm{T}} \Sigma_0^{-1} \mu_0}{2} \right)^{-\left(\frac{\nu_0+n}{2}+1\right)}}{\delta^{-(M+1)} \left|\Sigma_M\right|^{1/2} \left(\frac{\gamma_0 + \sum_{i=1}^n y_i^2 - \mu_M^{\mathrm{T}} \Sigma_M^{-1} \mu_M}{2} \right)^{-\left(\frac{\nu_0+n}{2}+1\right)}} \underset{\delta^2 \to \infty}{\longrightarrow} \infty$$

- **Do not use vague priors for model selection**!!!

- For a robust model, select a random δ^2 and estimate it from the data. However, numerical methods are then necessary.

• In practice, you might have models of different natures for your data $x = (x_1, ..., x_T)$.

• $\mathcal{M}_1$: Gaussian white noise $X_n \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_{WN}^2\right)$.

• $\mathcal{M}_2$: An AR process of order k_{AR}, k_{AR} being fixed, excited by white Gaussian noise $V_n \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_{_{AR}}^2\right)$,

$$X_n = \sum_{i=1}^{k_{AR}} a_i X_{n-i} + V_n.$$

• $\mathcal{M}_3$: $k_{\sin}$ sinusoids, $k_{\sin}$ being fixed, embedded in a white Gaussian noise sequence $V_n \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_{\sin}^2\right)$,

$$X_n = \sum_{j=1}^{k_{\sin}} \left(a_{c_j} \cos\left[\omega_j n\right] + a_{s_j} \sin\left[\omega_j n\right]\right) + V_n.$$

2.3– Bayesian Model for Model Choice

- Generally speaking you have a countable collection of models $\{\mathcal{M}_i\}$.

- For each model $\mathcal{M}_i$, you have a prior $\pi_i(\theta_i)$ on Θ_i and a likelihood function $f_i(x|\theta_i)$.

- You attribute a prior probability $\pi(i)$ to each model $\mathcal{M}_i$.

- The parameter space is $\bigcup_i \{i\} \times \Theta_i$ and the prior on this space is

$$\pi(i, \theta_i) = \pi(i)\, \pi_i(\theta_i).$$

• In the polynomial regression example

$$\Theta =_{i=0}^{M_{\max}} \underbrace{\{i\}}_{\text{model indicator}} \times \underbrace{\mathbb{R}^{i+1}}_{\text{regression parameters } \beta_{0:i}} \times \underbrace{\mathbb{R}^{+}}_{\text{noise variance}} .$$

• Remark: In all models, you have a noise variance to estimate. This parameter has a different interpretation for each model.

• In the non-nested example $\Theta = \{1\} \times \Theta_1 \cup \{2\} \times \Theta_2 \cup \{3\} \times \Theta_3$ where

$$
\begin{aligned}
\theta_1 &= \sigma^2_{WN} \text{ and } \Theta_1 = \mathbb{R}^+, \\
\theta_2 &= \left(a_1, ..., a_{k_{AR}}, \sigma^2_{AR}\right) \text{ and } \Theta_2 = \mathbb{R}^{k_{AR}} \times \mathbb{R}^+, \\
\theta_3 &= \left(a_{c_1}, a_{s_1}, \omega_1, \ldots, a_{c_{k_{\sin}}}, a_{s_{k_{\sin}}}, \omega_{k_{\sin}}, \sigma^2_{WN}\right), \ \Theta_3 = \mathbb{R}^{2k_{\sin}} \times [0, \pi]^{k_{\sin}} \times \mathbb{R}^+.
\end{aligned}
$$

• Remark: In all models, you have a noise variance to estimate. This parameter has a different interpretation for each model.

• Be careful, we don't select here $\Theta = \{1, 2, 3\} \times \Theta_1 \times \Theta_2 \times \Theta_3$.

2.3– Bayesian Model for Model Choice

- The posterior is given by Bayes' rule

$$\pi\left(k, \theta_k \middle| x\right) = \frac{\pi\left(k\right) \pi_k\left(\theta_k\right) f_k\left(x \middle| \theta_k\right)}{\sum_k \pi\left(k\right) \int_{\Theta_k} \pi_k\left(\theta_k\right) f_k\left(x \middle| \theta_k\right) d\theta_k}.$$

- We can obtain the posterior model probabilities through

$$\pi\left(k \middle| x\right) = \int_{\Theta_k} \pi\left(k, \theta_k \middle| x\right) d\theta_k.$$

- Once more, it is conceptually simple but it requires the calculation of many/an infinite number of integrals.

2.4– Bayesian Model Averaging

- Assume you are doing some prediction of say $Y \sim g_k(y|\theta_k)$. Then in light of x, we have

$$
\begin{aligned}
g(y|x) &= \sum_k \int_{\Theta_k} g_k(y|\theta_k)\,\pi(k,\theta_k|x)\,d\theta_k \\
&= \sum_k \underbrace{\pi(k|x)}_{\text{posterior proba of model } k} \underbrace{\int_{\Theta_k} g_k(y|\theta_k)\,\pi(\theta_k|x,k)\,d\theta_k}_{\text{Prediction from model } k}
\end{aligned}
$$

- This is called Bayesian model averaging. All the models are taken into account to perform the prediction.

2.4– Bayesian Model Averaging

- An alternative way to make prediction consists of selecting the best "model"; say the model which has the highest posterior proba.

- The prediction is performed according to

$$\int_{\Theta_{k_{\text{best}}}} g_{k_{\text{best}}}\left(y|\,\theta_{k_{\text{best}}}\right)\pi\left(\theta_{k_{\text{best}}}\right|x, k_{\text{best}}\right)d\theta_{k_{\text{best}}}$$

- This is computationally much simpler and cheaper. This can also be very misleading.

2.5– Bayesian Variable Selection Example

- Consider the standard linear regression problem

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

- Often you might have too many predictors, so this model will be inefficient.

- A standard Bayesian treatment of this problem consists of selecting only a subset of explanatory variables.

- This is nothing but a model selection problem with 2^p possible models.

2.6– Bayesian Variable Selection Example

- A standard way to write the model is

$$Y = \sum_{i=1}^{p} \gamma_i \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where $\gamma_i = 1$ if X_i is included or $\gamma_i = 0$ otherwise. However this suggests that β_i is defined even when $\gamma_i = 0$.

- A neater way to write such models is to write

$$H_\gamma \colon Y = \sum_{\{i:\gamma_i=1\}} \beta_i X_i + \sigma V = \beta_\gamma^{\mathrm{T}} X_\gamma + \sigma V$$

where, for a vector $\gamma = (\gamma_1, ..., \gamma_p)$, $\beta_\gamma = \{\beta_i : \gamma_i = 1\}$, $X_\gamma = \{X_i : \gamma_i = 1\}$ and $n_\gamma = \sum_{i=1}^{p} \gamma_i$.

- Prior distributions

$$\pi_\gamma\left(\beta_\gamma, \sigma^2\right) = \mathcal{N}\left(\beta_\gamma; 0, \delta^2\sigma^2 I_{n_\gamma}\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$$

and $\pi(H_\gamma) = \pi(\gamma) = \prod_{i=1}^{p} \pi(\gamma_i) = 2^{-p}$.

• For a fixed model γ and n observations $D = \{x_i, y_i\}_{i=1}^n$ then we can determine the marginal likelihood and the posterior analytically

$$\pi_\gamma\left(D|\,\beta_\gamma,\sigma^2\right) = \Gamma\left(\frac{\nu_0+n}{2}+1\right)\delta^{-n_\gamma}\left|\Sigma_\gamma\right|^{1/2}\left(\frac{\gamma_0+\sum_{i=1}^n y_i^2-\mu_\gamma^{\mathrm{T}}\Sigma_\gamma^{-1}\mu_\gamma}{2}\right)^{-\left(\frac{\nu_0+n}{2}+1\right)}$$

and

$$\begin{aligned}\pi_\gamma\left(\beta_\gamma,\sigma^2\middle|\,D\right) &= \mathcal{N}\left(\beta_\gamma;\mu_\gamma,\sigma^2\Sigma_\gamma\right)\\ &\quad\times\mathcal{IG}\left(\sigma^2;\frac{\nu_0+n}{2},\frac{\gamma_0+\sum_{i=1}^n y_i^2-\mu_\gamma^{\mathrm{T}}\Sigma_\gamma^{-1}\mu_\gamma}{2}\right)\end{aligned}$$

where

$$\mu_\gamma = \Sigma_\gamma\left(\sum_{i=1}^n y_i x_{\gamma,i}\right),\ \Sigma_\gamma^{-1} = \delta^{-2}I_{n_\gamma}+\sum_{i=1}^n x_{\gamma,i}x_{\gamma,i}^{\mathrm{T}}.$$

2.7– Bayesian Variable Selection Example

- Popular alternative Bayesian models include

$$\begin{aligned}\gamma_i &\sim \mathcal{B}(\lambda) \text{ where } \lambda \sim \mathcal{U}[0,1],\\ \gamma_i &\sim \mathcal{B}(\lambda_i) \text{ where } \lambda_i \sim \mathcal{B}e(\alpha,\beta).\end{aligned}$$

- g-prior (Zellner)

$$\beta_\gamma | \sigma^2 \sim \mathcal{N}\left(\beta_\gamma; 0, \delta^2\sigma^2 \left(X_\gamma^{\mathrm{T}} X_\gamma\right)^{-1}\right).$$

- Robust models where additionally one has

$$\delta^2 \sim \mathcal{IG}\left(\frac{a_0}{2}, \frac{b_0}{2}\right).$$

- Such variations are very important and can modify dramatically the performance of the Bayesian model.

2.8– Bayesian Variable Selection Example

- Caterpillar dataset: 1973 study to assess the influence of some forest settlement characteristics on the development of catepillar colonies.

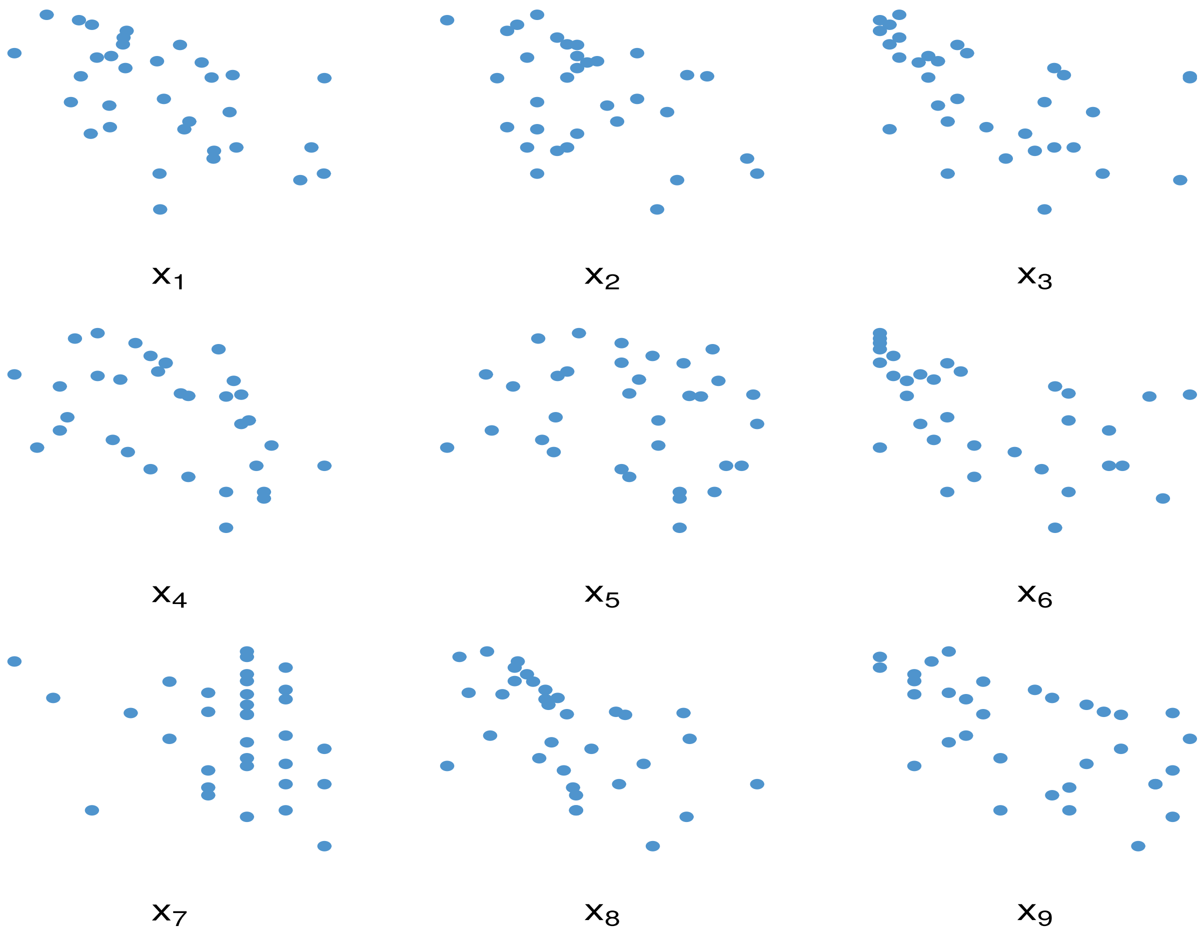
- The response variable is the log of the average number of nests of caterpillars per tree on an area of 500 square meters.

- We have $n = 33$ data and 10 explanatory variables

2.9– Bayesian Variable Selection Example

- x_1 is the altitude (in meters),
- x_2 is the slope (in degrees),
- x_3 is the number of pines in the square,
- x_4 is the height (in meters) of the tree sampled at the center of the square,
- x_5 is the diameter of the tree sampled at the center of the square,
- x_6 is the index of the settlement density,
- x_7 is the orientation of the square (from 1 if southbound to 2 otherwise),
- x_8 is the height (in meters) of the dominant tree,
- x_9 is the number of vegetation strata,
- x_{10} is the mix settlement index (from 1 if not mixed to 2 if mixed).

2.9– Bayesian Variable Selection Example



2.9– Bayesian Variable Selection Example

- Top five most likely models

$\pi(\gamma\|x)$ (Ridge $\delta^2 = 10$)	$\pi(\gamma\|x)$ (g-p $\delta^2 = 10$)	$\pi(\gamma\|x)$ (g-p, δ^2 estimated)
0,1,2,4,5/0.1946	0,1,2,4,5/0.2316	0,1,2,4,5/0.0929
0,1,2,4,5,9/0.0321	0,1,2,4,5,9/0.0374	0,1,2,4,5,9/0.0325
0,12,4,5,10/0.0327	0,1,9/0.0344	0,1,2,4,5,10/0.0295
0,1,2,4,5,7/0.0306	0,1,2,4,5,10/0.0328	0,1,2,4,5,7/0.0231
0,1,2,4,5,8/0.0251	0,1,4,5/0.0306	0,1,2,4,5,8/0.0228

2.10– Extension to Bayesian Neural Networks

• Bayesian variable selection can be directly used to 'learn' Bayesian neural networks. In this case, given some observations $\{x_i, y_i\}_{i=1}^{p}$, we select a regression model of the form

$$H_\gamma : \; Y = \sum_{\{i:\gamma_i=1\}} \beta_i \Psi \left(\|X - x_i\| \right) + \sigma V$$

where $\Psi \left(\|X - x_i\| \right)$ is a radial basis function, e.g. Gaussian.

2.11– Conclusion

- Bayesian model selection is a simple and principled way to do model selection.

- Bayesian model selection appears in numerous applications.

- Vague/Improper priors have to be banned in the model selection context!!!!

- Bayesian model selection only allows us to “compare” models. It does not tell you if any of the candidate models makes sense.

- Except for simple problems, it is impossible to perform calculations in closed-form.