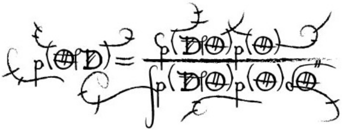


Bayesian Learning 732A46: Lecture 10

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

May 2016

Lecture overview

- Bayesian model comparison
- Computing marginal likelihoods
- Bayesian model averaging

Using the likelihood for model comparison

- Consider two models for the data $y = (y_1, ..., y_n)$: M_1 and M_2.
- Let $p_k(y|\theta_k)$ denote the **data density** (fixed θ_k) under model M_k.
- If we know θ_1 and θ_2, the **likelihood ratio** is useful

$$\frac{p_1(y|\theta_1)}{p_2(y|\theta_2)}.$$

- But often we **do not know** θ_1 and θ_2.
- **Frequentist**: The **likelihood ratio** with the MLE plugged in:

$$\frac{p_1(y|\hat{\theta}_1)}{p_2(y|\hat{\theta}_2)}.$$

- **Bigger models** always win with estimated likelihood ratio.
- **Hypothesis tests** become problematic for non-nested models.

Bayesian model comparison

- Use your priors $p_1(\theta_1)$ and $p_2(\theta_2)$ to get rid (**average over**) of θ.
- The **marginal likelihood** for model M_k with parameters θ_k

$$p_k(y) = \int p_k(y|\theta_k)p_k(\theta_k)d\theta_k.$$

- Recall **Bayes' theorem** in the simple case of $\theta = \{H, H^c\}$

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}, \quad \Pr(E) = \Pr(E|H)\Pr(H) + \Pr(E|H^c)\Pr(H^c)$$

The marginal likelihood in words

The **marginal likelihood** $\Pr(E)$ is a **weighted average** of the probability of the evidence under the different hypothesis. The weights **are given by the prior probabilities**.

- θ_k (or H, H^c) is removed (**averaged out**) by the prior. **Priors matter!**

Bayesian model comparison, cont.

- The **Bayes factor**

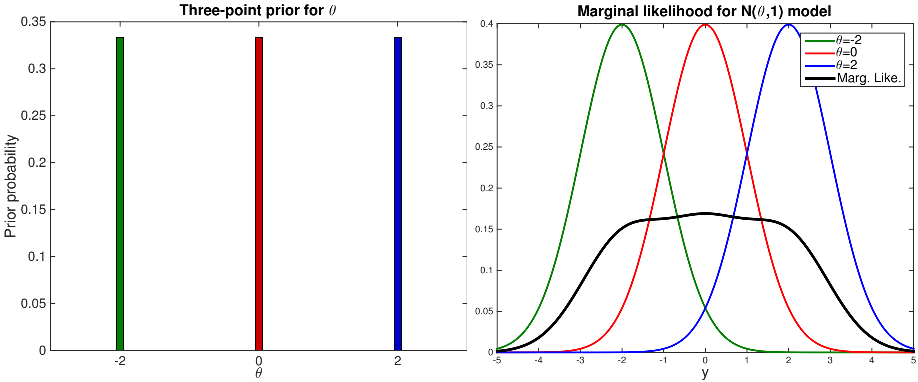
$$B_{12}(y) = \frac{p_1(y)}{p_2(y)}.$$

- **Bayesian machinery**: Posterior model probabilities

$$\underbrace{\Pr(M_k|y)}_{\text{Posterior model prob.}} \quad \propto \quad \underbrace{p(y|M_k)}_{\text{marginal likelihood } [=p_k(y)]} \quad \cdot \quad \underbrace{\Pr(M_k)}_{\text{prior model prob.}}$$

- **Important**: Two sets of priors
  1. Prior for **the parameters** θ_k within model M_k (**"the usual" prior**)
  2. Prior for **the models** $\Pr(M_k)$.

Priors matter



Example: Geometric vs Poisson

- Model 1 - **Geometric** with **Beta** prior:
  - $y_1, ..., y_n | \theta_1 \sim \text{Geometric}(\theta_1)$,
    $$p(y_i|\theta_1) = (1-\theta_1)^{y_i}\theta_1 \quad y_i \in \{0, 1, 2, \dots\}, 0 \leq \theta_1 \leq 1.$$
  - $\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$,
    $$p(\theta_1) = \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}.$$
- Model 2 - **Poisson** with **Gamma** prior:
  - $y_1, ..., y_n | \theta_2 \sim \text{Poisson}(\theta_2)$,
    $$p(y_i|\theta_2) = \frac{\theta_2^{y_i}\exp(-\theta_2)}{y_i!} \quad y_i \in \{0, 1, 2, \dots\}, \theta_2 > 0.$$
  - $\theta_2 \sim \text{Gamma}(\alpha_2, \beta_2)$,
    $$p(\theta_2) = \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}\theta_2^{\alpha_2-1}\exp(-\beta_2\theta_2).$$

Geometric vs Poisson: $p(y)$ for Geometric (M_1)

- **Marginal likelihood** for M_1 $[y = (y_1, \ldots, y_n)]$

$$\begin{aligned} p_1(y) &= \int p_1(y|\theta_1)p(\theta_1)d\theta_1 \\ &= \int \left(\prod_{i=1}^{n} p(y_i|\theta_1)\right) p(\theta_1)d\theta_1 \\ &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \int (1-\theta_1)^{\sum_{i=1}^{n} y_i}\theta_1^n \times \theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}d\theta_1 \\ &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \int \theta_1^{n+\alpha_1-1}(1-\theta_1)^{n\bar{y}+\beta_1-1}d\theta_1 \end{aligned}$$

- The **beta function**

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt, \quad a,b > 0.$$

- **Nice property** of the beta function

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Geometric vs Poisson: $p(y)$ for Geometric (M_1), cont

- **Thus**

$$\begin{aligned} p_1(y) &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\int \theta_1^{n+\alpha_1-1}(1-\theta_1)^{n\bar{y}+\beta_1-1}d\theta_1 \\ &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}B(n+\alpha_1, n\bar{y}+\beta_1) \\ &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\frac{\Gamma(n+\alpha_1)\Gamma(n\bar{y}+\beta_1)}{\Gamma(n+\alpha_1+n\bar{y}+\beta_1)}. \end{aligned}$$

- **Note**: It **does not** depend on θ_1. θ_1 has been averaged out!

Geometric vs Poisson: $p(y)$ for Poisson (M_2)

- **Marginal likelihood** for M_2 $[y = (y_1, \ldots, y_n)]$

$$
\begin{aligned}
p_2(y) &= \int p_2(y|\theta_2) p(\theta_2) d\theta_2 \\
&= \int \left(\prod_{i=1}^{n} p(y_i|\theta_2) \right) p(\theta_2) d\theta_2 \\
&= \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \int \frac{\theta_2^{\sum y_i}}{\prod_{i=1}^{n} y_i} \exp(-n\theta_2) \times \theta_2^{\alpha_2 - 1} \exp(-\beta_2 \theta_2) d\theta_2 \\
&= \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2) \prod_{i=1}^{n} y_i} \int \theta_2^{n\bar{y} + \alpha_2 - 1} \exp(-(n + \beta_2)\theta_2) d\theta_2
\end{aligned}
$$

- The **gamma function**

$$
\Gamma(c) = \int_0^\infty t^{c-1} \exp(-t) dt, \quad c > 0.
$$

- ... rewritten to fit **our form above** (simple change of variables) ...

$$
\frac{1}{(n + \beta_2)^c} \Gamma(c) = \int_0^\infty t^{c-1} \exp(-(n + \beta_2)t) dt, \quad c > 0.
$$

Geometric vs Poisson: $p(y)$ for Poisson (M_2), cont.

- **Thus**

$$p_2(y) = \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)\prod_{i=1}^n y_i}\int \theta_2^{n\bar{y}+\alpha_2-1}\exp(-(n+\beta_2)\theta_2)d\theta_2$$
$$= \frac{\beta_2^{\alpha_2}\Gamma(n\bar{y}+\alpha_2)}{\Gamma(\alpha_2)(n+\beta_2)^{n\bar{y}+\alpha_2}\prod_{i=1}^n y_i}.$$

- **Note** (**again!**): It **does not** depend on θ_2. θ_2 has been averaged out!

Geometric vs Poisson, cont.

- **Before** comparing the results we need to set the hyper-parameters in **some suitable way**.

- Set **hyper-parameters** so that the prior predictive means match

$$E(y_i|M_1) = E(y_i|M_2) \quad \Longrightarrow \quad (\alpha_1 - 1)\alpha_2 = \beta_1\beta_2$$

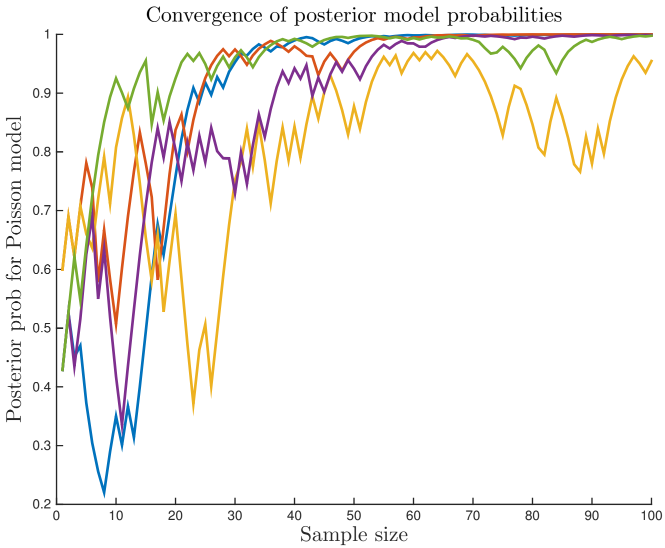
- The **prior predictive mean** computed by the **tower property**

$$E(y_i|M_k) = E_\theta\left(E_{y_i|\theta}(y_i|\theta, M_k)\right), \quad \text{for } k = 1, 2,$$

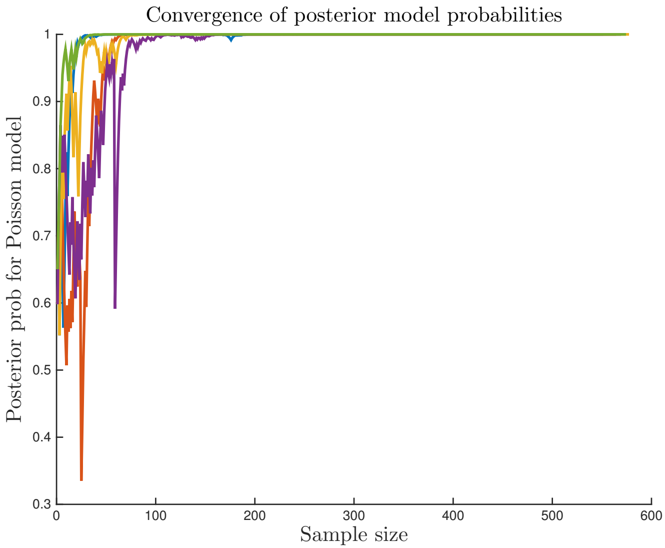
and

$$E_{y_i|\theta}(y_i|\theta, M_k) = \begin{cases} \frac{\theta_1}{1-\theta_1}, & \text{if } k = 1 \\ \theta_2, & \text{if } k = 2. \end{cases}$$

Geometric vs Poisson for Pois(1) data



Geometric vs Poisson for Pois(1) data



Properties of Bayesian model comparison

- **Coherence** of pair-wise comparisons

$$B_{12} = B_{13} \cdot B_{32}.$$

- **Consistency** when true model is in $\mathcal{M} = \{M_1, ..., M_K\}$

$$\Pr\left(M = M_{TRUE}|y\right) \to 1 \quad \text{as} \quad n \to \infty.$$

- **"KL-consistency"** when $M_{TRUE} \notin \mathcal{M}$

$$\Pr\left(M = M^{\star}|y\right) \to 1 \quad \text{as} \quad n \to \infty,$$

where $M^{\star}$ is the model that minimizes Kullback-Leibler distance

$$D_{KL}(p_{TRUE}, p_M) = \int p_{TRUE}(y) \log\left(\frac{p_M(y)}{p_{TRUE}(y)}\right) dy$$

between $p_M(y)$ and $p_{TRUE}(y)$.

Some warnings

- Smaller models **always win** when priors are very vague.
- **Improper priors** **can't be used** for model comparison.
- **Bayes factors** are **relative measures**! **Does not** say anything about a single model's adequacy.

Bayesian hypothesis testing

- **Hypothesis testing** is a **model selection** problem.
- **Example**: Bernoulli model with prior $\theta \sim \mathrm{Beta}(\alpha, \beta)$

$$M_0 : y_1, ..., y_n|\theta_0 \overset{iid}{\sim} \mathrm{Bernoulli}(\theta_0)$$
$$M_1 : y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Bernoulli}(\theta).$$

- **Likelihood**: $p(y|\theta) = \theta^s(1-\theta)^f$ $(y = (y_1, \ldots, y_n)$, $s = \sum y_i$, $f = n - s)$.
- **Marginal likelihoods**
  - For model M_1

$$p(y|M_1) = \theta^s(1-\theta)^f.$$

  - For model M_2

$$p(y|M_2) = \int \theta^s(1-\theta)^f \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$
$$= \frac{\Gamma(\alpha+\beta)\Gamma(s+\alpha)\Gamma(f+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}.$$

Bayesian hypothesis testing, cont.

- Reject (or accept) based on the **posterior model probabilities**

$$Pr(M_k|y) \propto p(y|M_k)Pr(M_k), \text{ for } k = 0, 1.$$

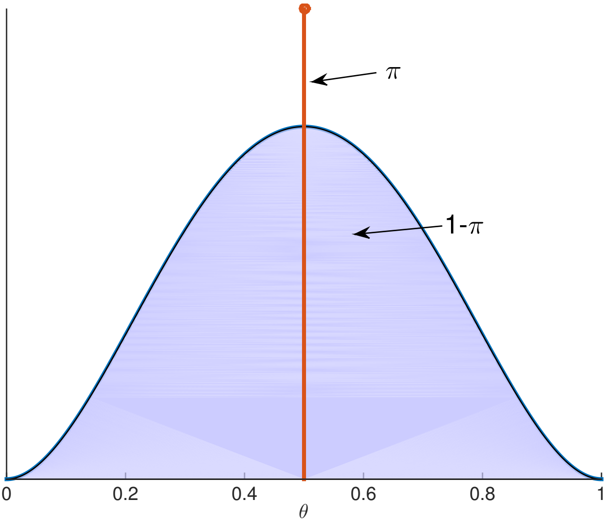
- A **"sharp null"** hypothesis is equivalent to using '**spike-and-slab**' prior:

$$p(\theta) = \pi\delta_{\theta_0}(\theta) + (1-\pi)\mathrm{Beta}(\alpha, \beta).$$

- Think about the **shrinkage mechanism**!

- **Note**: data can now **support** a null hypothesis (not only reject it).

Spike-and-slab prior [with $\theta_0 = 0.5$]



π

Marginal likelihood - a measure of out-of-sample predictive performance

- **The marginal likelihood** can be decomposed as

$$p(y_1, ..., y_n) = p(y_1)p(y_2|y_1)\cdots p(y_n|y_1, y_2, ..., y_{n-1}).$$

- Assume that y_i is **independent** of $y_1, ..., y_{i-1}$ **conditional** on θ:

$$p(y_i|y_1, ..., y_{i-1}) = \int p(y_i|\theta)p(\theta|y_1, ..., y_{i-1})d\theta$$

- **The prediction** of y_1 is **based on the prior** of θ, and is therefore **sensitive to the prior**.

- In contrast, **the prediction** of y_n **uses almost all the data** to infer θ. If n is large **influence of prior is negligible** for y_n.

- **Summary**: "Early" out-of-sample predictions are more influenced by $p(\theta)$.

Illustrating the sensitivity to the prior for early obs

- **Model**: $y_1, ..., y_n|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 **known**.
- **Prior**: $\theta \sim \mathcal{N}(0, \kappa^2\sigma^2)$ [for simplified expressions].
- **Partial posterior** up to observation $i-1$ ($\mu_0 = 0$)

$$\theta|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \frac{\sigma^2}{i-1+\kappa^{-2}}\right]$$

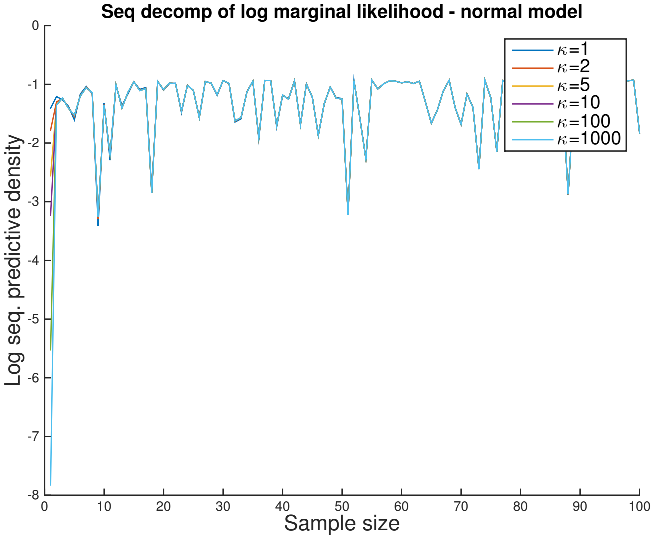
where $w_i(\kappa) = \frac{i-1}{i-1+\kappa^{-2}}$ [*the usual weighted average story*].

- **Predictive density** for obs $i-1$

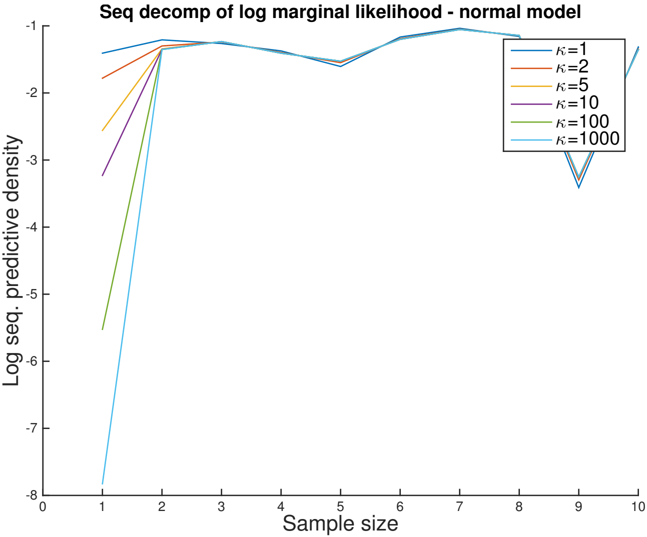
$$y_i|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \sigma^2\left(1 + \frac{1}{i-1+\kappa^{-2}}\right)\right].$$

- **Terms with i large**: $y_i|y_1, ..., y_{i-1} \overset{approx}{\sim} \mathcal{N}\left(\bar{y}_{i-1}, \sigma^2\right)$, **not sensitive** to κ
- For $i = 1$, $y_1 \sim \mathcal{N}\left[0, \sigma^2\left(1 + \frac{1}{\kappa^{-2}}\right)\right]$ can be **very sensitive** to κ.

First observation is sensitive to κ



First observation is sensitive to κ



Log Predictive Score - LPS: a way to reduce the sensitivity

- **Simple idea**: a measure similar to the marginal likelihood but where **the first observation is less sensitive** to the prior.
- **Sacrifice** $n^\star$ observations to train/update the prior.
- **Predictive density score**: PS

$$PS(n^\star) = p(y_{n^\star+1}|y_1, ..., y_{n^\star}) \cdots p(y_n|y_1, ..., y_{n-1}).$$

- **Compare** PS to $p(y)$ in factorized form.
- Usually report on log scale: **Log Predictive Score** (**LPS**).
- Which observations to **train/update** with (and which to predict)?
- **Split the data**: *Training* and *test* data
  - Straightforward for **time series**.
  - **Cross-sectional data**: cross-validation is useful.

Computing the marginal likelihood: Conjugate models

- Computing the **marginal likelihood** requires integration w.r.t. θ.
- **Short cut** for **conjugate models** by rearrangement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

- By conjugacy $p(\theta|y)$ is **analytically available**.
- Insert everything and **work out the algebra.**

Computing the marginal likelihood: Simulation methods

- Usually difficult (or **impossible**) to analytically derive

$$p(y) = \int p(y|\theta)p(\theta)d\theta = E_\theta[p(y|\theta)].$$

- Draw from the prior $\theta^{(1)}, ..., \theta^{(N)}$ and use the usual **Monte Carlo estimate**

$$\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|\theta^{(i)}).$$

- **Unstable** (huge variance) if the likelihood is somewhat different from the prior.
- **Importance sampling**. Let $\theta^{(1)}, ..., \theta^{(N)}$ be iid draws from $g(\theta)$.

$$\int p(y|\theta)p(\theta)d\theta = \int \frac{p(y|\theta)p(\theta)}{g(\theta)}g(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(y|\theta^{(i)})p(\theta^{(i)})}{g(\theta^{(i)})}.$$

- **Modified Harmonic mean**: $g(\theta) = \mathcal{N}(\tilde{\theta}, \tilde{\Sigma}) \cdot I_c(\theta)$, where $\tilde{\theta}$ and $\tilde{\Sigma}$ is the posterior mean and covariance matrix estimated from an MCMC chain, and $I_c(\theta) = 1$ if $(\theta - \tilde{\theta})'\tilde{\Sigma}^{-1}(\theta - \tilde{\theta}) \leq c$.

Computing the marginal likelihood: Simulation methods, cont.

- Rearrangement of **Bayes' theorem** (again!): $p(y) = p(y|\theta)p(\theta)/p(\theta|y)$.
- **Note 1**: Need the full expression for the posterior, **including** the constants ind of θ.
- **Note 2**: LHS is **independent** of θ. RHS **depends** on θ...
- ... any θ must cancel. Enough to evaluate in a single point θ_0.
- **Kernel density estimator** to approximate $p(\theta_0|y)$. Unstable.
- Chib (1995) provide better solutions for **Gibbs sampling**.
- Chib and Jeliazkov (2001) generalizes to **MH algorithm** (good for Independence MH, not so good for RWM).

Computing the marginal likelihood: Approximation

- By **normal approximation** of the posterior distribution (**Lecture 6**).
- **Recall**: for large n

$$\begin{aligned} p(\theta|y) &\approx \mathcal{N}_p(\theta^\star, \Sigma_{\theta^\star} = J_{\theta^\star,y}^{-1}) \\ &= (2\pi)^{-p/2}|J_{\theta^\star,y}^{-1}|^{-1/2}\exp\left(-\frac{1}{2}(\theta-\theta^\star)' J_{\theta^\star,y}(\theta-\theta^\star)\right). \end{aligned}$$

- **The Laplace approximation**: Use rearranged Bayes' theorem with $\theta = \theta^\star$

$$\log \hat{p}(y) = \log p(y|\theta^\star) + \log p(\theta^\star) + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\left|J_{\theta^\star,y}^{-1}\right|.$$

- **As usual**: $\theta^\star$ and $J_{\theta^\star,y}$ $[-H_{\theta^\star}]$ are obtained via a numerical optimization (e.g. `optim` in `R`).

Bayesian model averaging

- Let γ have the **same interpretation** across the model space

$$\mathcal{M} = \{M_1, \dots, M_K\}.$$

Let $\theta = \{\theta_1, \dots, \theta_K\}$ be the corresponding set of parameters.

- The **marginal posterior** (marginalized over $\mathcal{M}$) of γ

$$p(\gamma|y) = \int p(\gamma, \mathcal{M}|y) d\mathcal{M} = \sum_{k=1}^{K} p(\gamma|M_k, y) p(M_k|y),$$

where $p(\gamma|M_k, y)$ is the **marginal posterior** (marginalized over θ_k) of γ conditional on model k,

$$p(\gamma|M_k, y) = \int p(\gamma|\theta_k, y) p(\theta_k|y) d\theta_k.$$

- Note the **two layers** of averaging... **Bayes is all about averaging out (marginalize) unknown quantities!**

Bayesian model averaging, cont.

- **Example**: h-step ahead prediction for time series: $\gamma = (y_{T+1}, ..., y_{T+h})$,

$$\begin{aligned} p(\gamma|M_k, y) &= p_k(y_{T+1}, ..., y_{T+h}|y) \quad \text{[Posterior predictive for } M_k\text{]} \\ p(M_k|y) &\propto p(y|M_k)p(M_k), \quad [p(y|M_k) \text{ - Marg. likelihood for } M_k] \end{aligned}$$

- $p(y_{T+1}, ..., y_{T+h}|y)$ includes **three sources of uncertainty**:
  - **Future errors**/disturbances. **Simpler analogy**: σ^2 (assume known) in

$$\begin{aligned} y_1, \ldots, y_n|\theta &\overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2), \quad \text{and } p(\theta) \propto c \text{ gives} \\ p(\theta|y) &= \mathcal{N}(\bar{y}, \sigma^2/n) \\ p(\tilde{y}|y) &= \mathcal{N}\left(\bar{y}, \sigma^2 + \frac{\sigma^2}{n}\right) \quad \text{[Posterior predictive for future } \tilde{y}\text{]}. \end{aligned}$$

  - **Parameter uncertainty** (Posterior predictive averaged over posterior of θ).
  - **Model uncertainty** (by model averaging).
- **Any painful integrals**? Compute by simulation!

References

Chib, S., (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313-1321

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the MetropolisHastings output. *Journal of the American Statistical Association*, 96(453):270-281.

Lavine, M. and Schervish, M.J., (1999). Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119-122.