Bayesian Learning - Lecture 10

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

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

Using likelihood for model comparison

- Consider two models for the data $\mathbf{y} = (y_1, ..., y_n)$: M_1 and M_2.
- Let $p_i(\mathbf{y}|\theta_i)$ denote the data density under model M_i.
- If know θ_1 and θ_2, the **likelihood ratio** is useful

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

- The **likelihood ratio** with ML estimates plugged in:

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

- Bigger models always win in estimated likelihood ratio.
- **Hypothesis tests** are problematic for non-nested models. End results is not very useful for analysis.

Bayesian model comparison

- Just use your priors $p_1(\theta_1)$ och $p_2(\theta_2)$.
- The **marginal likelihood** for model M_k with parameters θ_k

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

- θ_k is removed by the prior. **Not a magic bullet**. **Priors matter!**
- The **Bayes factor**

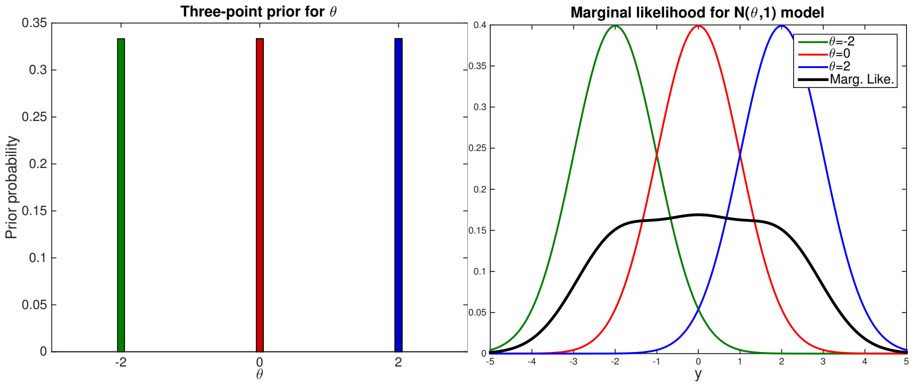
$$B_{12}(y) = \frac{p_1(y)}{p_2(y)}.$$

- **Posterior model probabilities**

$$\underbrace{\Pr(M_k|\mathbf{y})}_{\text{posterior model prob.}} \propto \underbrace{p(\mathbf{y}|M_k)}_{\text{marginal likelihood}} \cdot \underbrace{\Pr(M_k)}_{\text{prior model prob.}}$$

- Important: we have priors over the models $\Pr(M_k)$, but also priors for the parameters θ_k within model M_k.

Priors matter



Example: Geometric vs Poisson

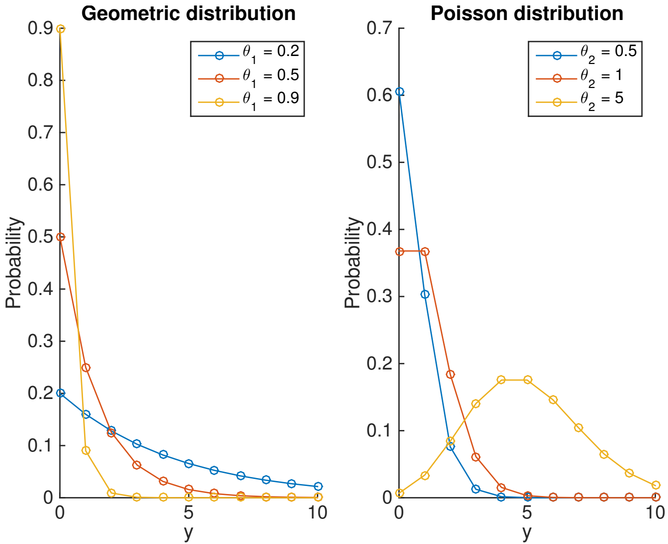
- Model 1 - **Geometric** with Beta prior:
  - $y_1, ..., y_n | \theta_1 \sim Geo(\theta_1)$
  - $\theta_1 \sim Beta(\alpha_1, \beta_1)$
- Model 2 - **Poisson** with Gamma prior:
  - $y_1, ..., y_n | \theta_2 \sim Poisson(\theta_2)$
  - $\theta_2 \sim Gamma(\alpha_2, \beta_2)$
- Marginal likelihood for M_1

$$p_1(y_1, ..., y_n) = \int p_1(y_1, ..., y_n | \theta_1) p(\theta_1) d\theta_1$$
$$= \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n + \alpha_1)\Gamma(n\bar{y} + \beta_1)}{\Gamma(n + n\bar{y} + \alpha_1 + \beta_1)}$$

- Marginal likelihood for M_2

$$p_2(y_1, ..., y_n) = \frac{\Gamma(n\bar{y} + \alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n + \beta_2)^{n\bar{y} + \alpha_2}} \frac{1}{\prod_{i=1}^n y_i!}$$

Geometric and Poisson



Geometric vs Poisson, cont.

- Priors match prior predictive means:

$$E(y_i|M_1) = E(y_i|M_2) \quad \Longleftrightarrow \quad \alpha_1\alpha_2 = \beta_1\beta_2$$

Geometric vs Poisson, cont.

- Priors match prior predictive means:

$$E(y_i|M_1) = E(y_i|M_2) \quad \Longleftrightarrow \quad \alpha_1\alpha_2 = \beta_1\beta_2$$

- **Data**: $y_1 = 0,\ y_2 = 0$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	1.5	4.54	5.87
$\Pr(M_1\|\mathbf{y})$	0.6	0.82	0.85
$\Pr(M_2\|\mathbf{y})$	0.4	0.18	0.15

Geometric vs Poisson, cont.

- Priors match prior predictive means:

$$E(y_i|M_1) = E(y_i|M_2) \quad \Longleftrightarrow \quad \alpha_1\alpha_2 = \beta_1\beta_2$$

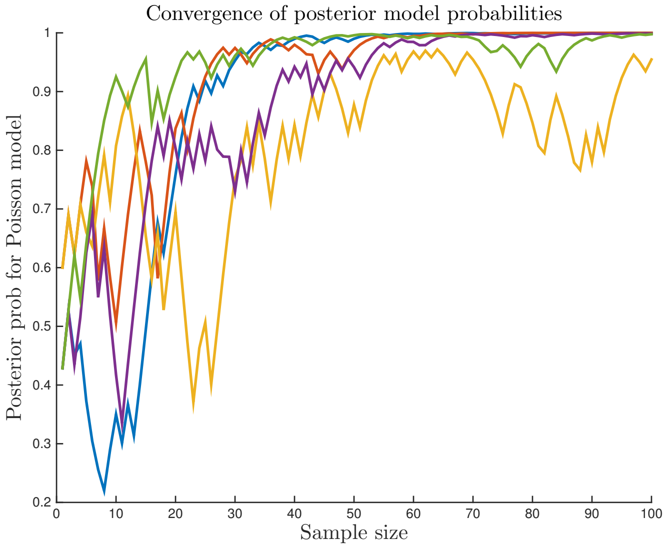
- **Data**: $y_1 = 0,\ y_2 = 0$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	1.5	4.54	5.87
$\Pr(M_1\|\mathbf{y})$	0.6	0.82	0.85
$\Pr(M_2\|\mathbf{y})$	0.4	0.18	0.15

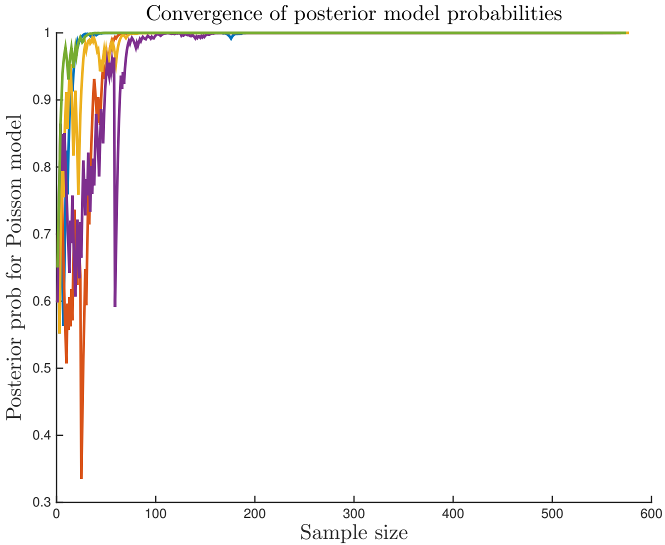
- **Data**: $y_1 = 3,\ y_2 = 3$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	0.26	0.29	0.30
$\Pr(M_1\|\mathbf{y})$	0.21	0.22	0.23
$\Pr(M_2\|\mathbf{y})$	0.79	0.78	0.77

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

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

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

$$\Pr(M = M^*|\mathbf{y}) \to 1 \quad \text{as} \quad n \to \infty$$

  where M^* is the model that minimizes Kullback-Leibler distance between $p_M(\mathbf{y})$ and $p_{TRUE}(\mathbf{y})$.
- Smaller models always win when priors are very vague.
- **Improper priors** can't be used for model comparison.

Model choice in multivariate time series

- Multivariate time series

$$\mathbf{x}_t = \alpha\beta'\mathbf{z}_t + \Phi_1\mathbf{x}_{t-1} + ...\Phi_k\mathbf{x}_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

- Need to choose:
  - **Lag length**, ($k = 1, 2.., 4$)
  - **Trend model** ($s = 1, 2, ..., 5$)
  - **Long-run (cointegration) relations** ($r = 0, 1, 2, 3, 4$).

The most probable (k, r, s) combinations in the Danish monetary data.

k	1	1	1	1	1	1	1	1	0	1
r	3	3	2	4	2	1	2	3	4	3
s	3	2	2	2	3	3	4	4	4	5
$p(k, r, s \vert y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

Bayesian hypothesis testing

- **Hypothesis testing** is just a special case of model selection:

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

$$M_1 : y_1, ..., y_n \overset{iid}{\sim} Bernoulli(\theta), \theta \sim Beta(\alpha, \beta)$$

$$p(y_1, ..., y_n | M_0) = \theta_0^s (1 - \theta_0)^f,$$

$$\begin{aligned} p(y_1, ..., y_n | M_1) &= \int_0^1 \theta^s (1-\theta)^f B(\alpha, \beta)^{-1} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= B(\alpha + s, \beta + f) / B(\alpha, \beta). \end{aligned}$$

- Posterior model probabilities

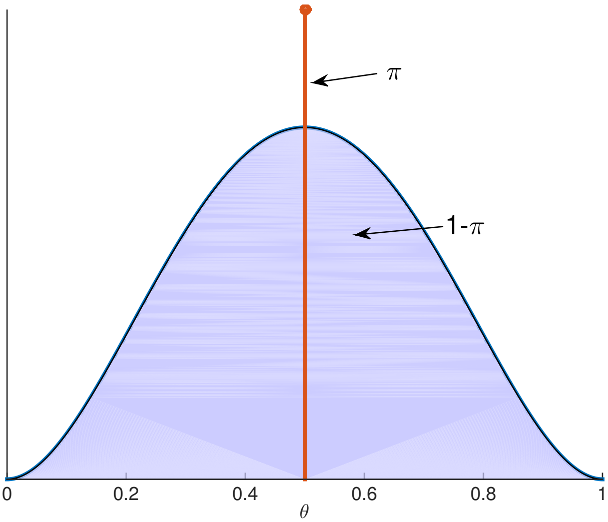
$$Pr(M_k | y_1, ..., y_n) \propto p(y_1, ..., y_n | M_k) Pr(M_k), \text{ for } k = 0, 1.$$

- Equivalent to using '**spike-and-slab**' prior:

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

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

Spike-and-slab prior



π

Marginal likelihood measures out-of-sample predictive performance

- The marginal likelihood can be decomposed as

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

- If we 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.
- The prediction of y_n uses almost all the data to infer θ. Very little influenced by the prior when n is not small.

Normal example

- **Model**: $y_1, ..., y_n|\theta \sim N(\theta, \sigma^2)$ with σ^2 known.
- **Prior**: $\theta|\sigma^2 \sim N(0, \kappa^2\sigma^2)$.
- Intermediate posterior at time $i-1$

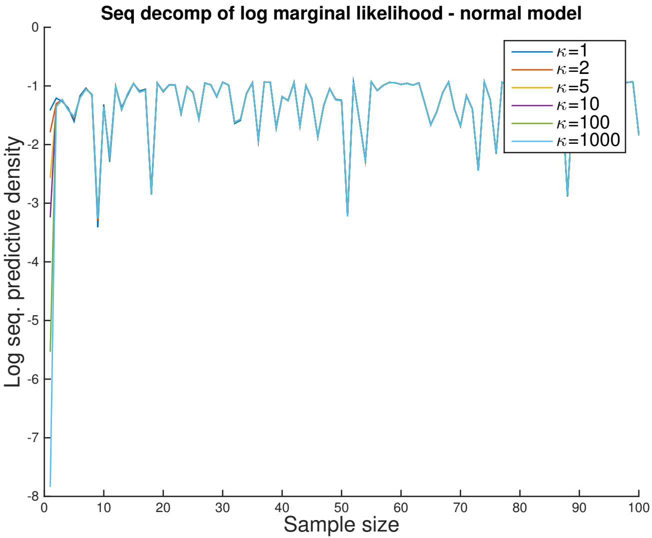
$$\theta|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \frac{\sigma^2}{i-1+\kappa^{-2}}\right]$$

  where $w_i(\kappa) = \frac{i-1}{i-1+\kappa^{-2}}$.
- Predictive density at time $i-1$

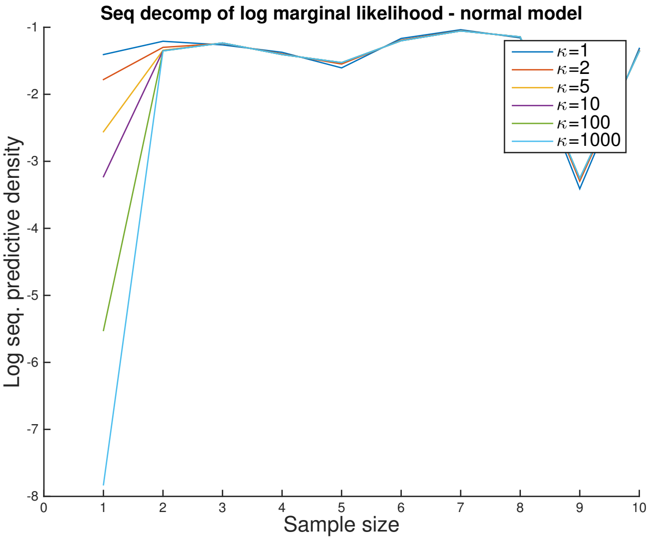
$$y_i|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \sigma^2\left(1 + \frac{1}{i-1+\kappa^{-2}}\right)\right]$$

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

First observation is sensitive to κ



First observation is sensitive to κ



Log Predictive Score - LPS

- To reduce sensitivity to the prior: sacrifice n^* observations to train the prior into a better posterior.
- Predictive density score: PS

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

- Usually report on log scale: **Log Predictive Score** (**LPS**).
- But which observations to train on (and which to test on)?
- Straightforward for time series.
- Cross-sectional data: **cross-validation**.

Model averaging

- Let γ be a quanitity with an interpretation which stays the same across the two models.
- Example: Prediction $\gamma = (y_{T+1}, ..., y_{T+h})'$.
- The marginal posterior distribution of γ reads

$$p(\gamma|\mathbf{y}) = p(M_1|\mathbf{y})p_1(\gamma|\mathbf{y}) + p(M_2|\mathbf{y})p_2(\gamma|\mathbf{y}),$$

  where $p_k(\gamma|\mathbf{y})$ is the marginal posterior of γ conditional on model k.
- Predictive distribution includes **three sources of uncertainty**:
  - **Future errors**/disturbances (e.g. the ε's in a regression)
  - **Parameter uncertainty** (the predictive distribution has the parameters integrated out by their posteriors)
  - **Model uncertainty** (by model averaging)

Marginal likelihood in conjugate models

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

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

- Bernoulli model example

$$\begin{aligned}
p(\theta) &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
p(y|\theta) &= \theta^{s}(1-\theta)^{f} \\
p(\theta|y) &= \frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}
\end{aligned}$$

- Marginal likelihood

$$p(y) = \frac{\theta^{s}(1-\theta)^{f}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}} = \frac{B(\alpha+s,\beta+f)}{B(\alpha,\beta)}$$

Computing the marginal likelihood

- Usually difficult to evaluate the integral

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

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

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

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

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

- **Modified Harmonic mean**: $g(\theta) = N(\tilde{\theta}, \tilde{\Sigma}) \cdot I_c(\theta)$, 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, cont.

- Rearrangement of Bayes' theorem: $p(\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\theta|\mathbf{y})$.
- We must know the posterior, **including** the normalization constant.
- But we only need to know $p(\theta|\mathbf{y})$ in a single point θ_0.
- **Kernel density estimator** to approximate $p(\theta_0|\mathbf{y})$. Unstable.
- Chib (1995, JASA) provide better solutions for **Gibbs sampling**.
- Chib-Jeliazkov (2001, JASA) generalizes to **MH algorithm** (good for IndepMH, terrible for RWM).
- Reversible Jump MCMC (RJMCMC) for model inference.
  - MCMC methods that moves in model space.
  - Proportion of iterations spent in model k estimates $\Pr(M_k|\mathbf{y})$.
  - Usually hard to find efficient proposals. Sloooow convergence.
- Bayesian nonparametrics (e.g. Dirichlet process priors).

Approximate marginal likelihoods

- Taylor approximation of the log posterior

$$\ln p(\mathbf{y}|\theta)p(\theta) \approx \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{1}{2}J_{\hat{\theta},\mathbf{y}}(\theta - \hat{\theta})^2,$$

$$p(\mathbf{y}|\theta)p(\theta) \approx p(\mathbf{y}|\hat{\theta})p(\hat{\theta}) \exp\left[-\frac{1}{2}J_{\hat{\theta},\mathbf{y}}(\theta - \hat{\theta})^2\right]$$

- **The Laplace approximation**:

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

  where p is the number of unrestricted parameters in the model.

- Note that $\hat{\theta}$ and $J_{\hat{\theta},\mathbf{y}}$ can be obtained with **numerical optimization** with BFGS update of Hessian.

- The **BIC approximation** is obtained if $J_{\hat{\theta},\mathbf{y}}$ behaves like $n \cdot I_p$ in large samples

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{p}{2}\ln n.$$