Advanced Bayesian Learning

Lecture 7 - Model comparison and evaluation

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Topic overview

- Bayesian model probabilities
- Model selection as a decision problem
- Predictive measures and Bayesian cross-validation
- Stacking and other approaches
- Bayesian variable selection and shrinkage

Likelihood ratios

- **Comparing models**:
  - M_1: $p_1(\mathrm{y}|\theta_1)$ against
  - M_2: $p_2(\mathrm{y}|\theta_2)$.
- **Likelihood ratio**

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

- $p_1(\mathrm{y}|\hat{\theta}_1) > p_2(\mathrm{y}|\hat{\theta}_2)$ if model M_1 is richer parametrized.
- **Hypothesis test**.
- **Non-nested models** are problematic.

Marginal likelihood and Bayes factor

- The **marginal likelihood** for model M_k with parameters θ_k

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

- Marginal likelihood is the **prior expected likelihood**

$$p_k(\boldsymbol{y}) = \mathbb{E}_{p_k(\theta_k)}\left[p_k(\boldsymbol{y}|\theta_k)\right]$$

- **Bayes factor**

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

- **Jeffreys' scale of evidence** for $B_{12}(\boldsymbol{y})$ (Kass-Raftery, JASA)
  - Barely worth mentioning: $1 - 3$
  - Positive: $3 - 20$
  - Strong: $20 - 150$
  - Very strong: > 150

Modeling perspectives

- $\mathcal{M}$-**closed** perspective
  - **Data generating process** $p_\star(\boldsymbol{y})$ is among the compared models.
  - Box: all models are false but some are useful.
- $\mathcal{M}$-**completed** perspective
  - $p_\star(\boldsymbol{y})$ is not among the compared models
  - A **subjective belief distribution** $p_u(\boldsymbol{y})$ exists.
- $\mathcal{M}$-**open** perspective
  - $p_\star(\boldsymbol{y})$ is not among the compared models
  - $p_u(\boldsymbol{y})$ is too complicated/costly to obtain.

Bayesian model probabilities

- $\mathcal{M}$**-closed** perspective, but often used also for $\mathcal{M}$-open.
- **Posterior model probabilities**

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

- **Variable selection**: p potential covariates. 2^p submodels M_k.
- **Prior over model space**, $\Pr(M_k)$, can be determined by
  - prior over the total number of effective covariates, p_{eff}.
  - uniform prior over subsets with p_{eff} effective covariates.

- A posterior distribution over model space is nice (mock-up):

	M_1	M_2	M_3	M_4
$\Pr(M_k)$	0.25	0.25	0.25	0.25
$\Pr(M_k\|\mathrm{y})$	0.05	0.81	0.10	0.04

Model choice in multivariate time series[1]

- **Multivariate time series**

$$x_t = \alpha\beta' z_t + \Phi_1 x_{t-1} + ...\Phi_k x_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

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

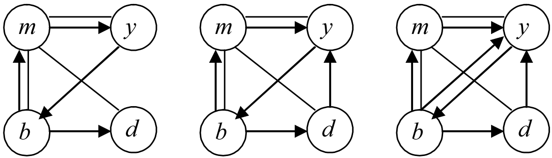
The most probable (k, r, s) combinations in the Danish monetary data.

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
$p(k, r, s \mid y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

[1] Corander and Villani (2004). Statistica Neerlandica.

Graphical models for multivariate time series[2]

- **Graphical models** for multivariate time series.
- Zero-restrictions on the effect from time series i on time series j, for all lags. (**Granger Causality**).
- Zero-restrictions on inverse covariance matrix of the errors. Contemporaneous conditional independence.



$p(G|\mathbf{X}) = 0.0033$ $p(G|\mathbf{X}) = 0.0028$ $p(G|\mathbf{X}) = 0.0025$

[2]Corander and Villani (2004). Journal of Time Series Analysis.

Properties of Bayesian model comparison

- **Coherent pair-wise comparisons**

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

- **Consistency** when $M_\star \in \mathcal{M} = \{M_1, ..., M_K\}$ (**$\mathcal{M}$-closed**)

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

- "KL-consistency" when $M_\star \notin \mathcal{M}$ (**$\mathcal{M}$-open**):

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

  $\tilde{M}$ minimizes **KL divergence** between $p_{\tilde{M}}(\mathrm{y})$ and $p_\star(\mathrm{y})$.

- KL-consistency may not be great in $\mathcal{M}$-open. More later.

Improper priors? Forget about it!

- **Improper priors cannot be used** for model comparison, not even as limits of proper priors.
- Prior $p_k(\theta) = c_k f_k(\theta_k)$ for some normalizing constant c_k.
- Posterior for θ_k: c_k cancels in the ratio

$$p_k(\theta_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta_k)p_k(\theta)}{\int p(\boldsymbol{y}|\theta_k)p_k(\theta)d\theta_k} = \frac{p(\boldsymbol{y}|\theta_k)f_k(\theta_k)}{\int p(\boldsymbol{y}|\theta_k)f_k(\theta_k)d\theta_k}$$

- Bayes factor: normalizing constants do not cancel

$$B_{kl} = \frac{\int p_k(\boldsymbol{y}|\theta_k)p_k(\theta_k)d\theta_k}{\int p_l(\boldsymbol{y}|\theta_l)p_l(\theta_l)d\theta_l} = \frac{c_k}{c_l} \cdot \frac{\int p_k(\boldsymbol{y}|\theta_k)f_k(\theta_k)d\theta_k}{\int p_l(\boldsymbol{y}|\theta_l)f_l(\theta_l)d\theta_l}$$

- Improper prior OK for parameters that appear in all models.
- Example: Error variance σ^2 in regression. But somewhat suspect, since interpretation of σ^2 depends on model.

Normal example

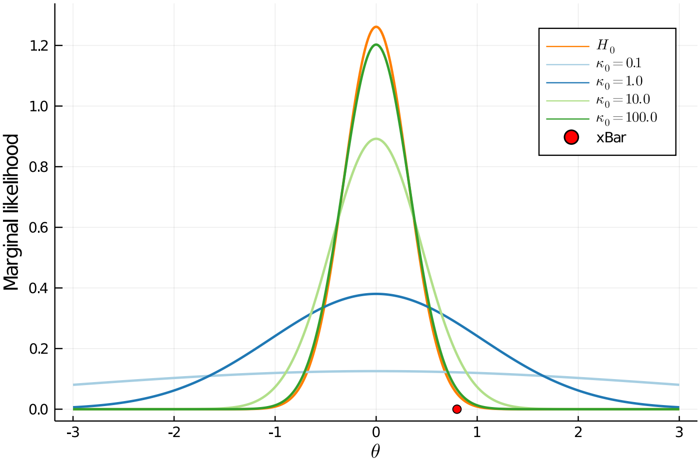
- **Model**: $x_1, \ldots, x_n \overset{iid}{\sim} N(\theta, \sigma^2)$, σ^2 known.
- **Prior**: $\theta \sim N(0, \sigma^2/\kappa_0)$.
- **Likelihood**: $\bar{x}$ is **sufficient** for θ and $\bar{x}|\theta \sim N(\theta, \sigma^2/n)$.
- **Marginal likelihood**: $p(\bar{x}|H_1) = N\left(0, \sigma^2(1/n + 1/\kappa_0)\right)$.
- Testing a **sharp null**: $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.

$$B_{01} = \frac{p(\bar{x}|H_0)}{p(\bar{x}|H_1)} = \frac{\sqrt{2\pi\sigma^2(1/n + 1/\kappa_0)}\exp\left(-\frac{1}{2\sigma^2(1/n)}(\bar{x}-0)^2\right)}{\sqrt{2\pi\sigma^2(1/n)}\exp\left(-\frac{1}{2\sigma^2(1/n+1/\kappa_0)}(\bar{x}-0)^2\right)}$$

$$\log \frac{p(\bar{x}|H_0)}{p(\bar{x}|H_1)} = -\frac{1}{2}\log\left(\frac{\kappa_0}{\kappa_0 + n}\right) - \frac{n\bar{x}^2}{2\sigma^2}\left(\frac{n}{\kappa_0 + n}\right)$$

- $\kappa_0 \to \infty$ then $B_{01} \to 1$ (prior under H_1 is a point mass at 0)
- $\kappa_0 \to 0$ then $B_{01} \to \infty$ ($p(\bar{x}|H_1)$ is average $p(\bar{x}|\theta)$ wrt prior)

Normal example



Marginal likelihood and predictive performance

- The marginal likelihood can be decomposed as

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

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

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

- **Prediction of y_1** is based on the prior of θ. Sensitive to prior.
- **Prediction of y_n** uses almost all the data to infer θ. Not sensitive to prior when n is not small.

Normal example

- **Model**: $y_1, ..., y_n|\theta \sim N(\theta, \sigma^2)$ with σ^2 known.
- **Prior**: $\theta \sim N(0, \sigma^2/\kappa_0)$.
- **Intermediate posterior** after observation i

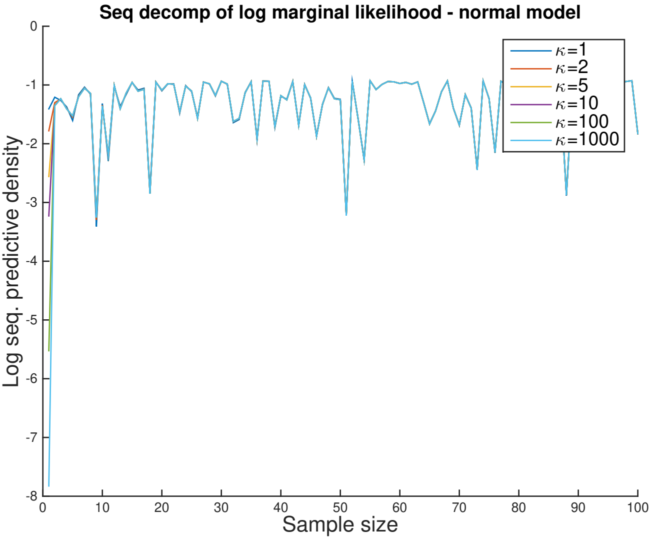
$$\theta|y_1, ..., y_i \sim N\left[w_i(\kappa_0) \cdot \bar{y}_i, \frac{\sigma^2}{i+\kappa_0}\right]$$

  where $w_i(\kappa) = \frac{i}{i+\kappa_0}$.
- **Intermediate predictive density** for y_{i+1}

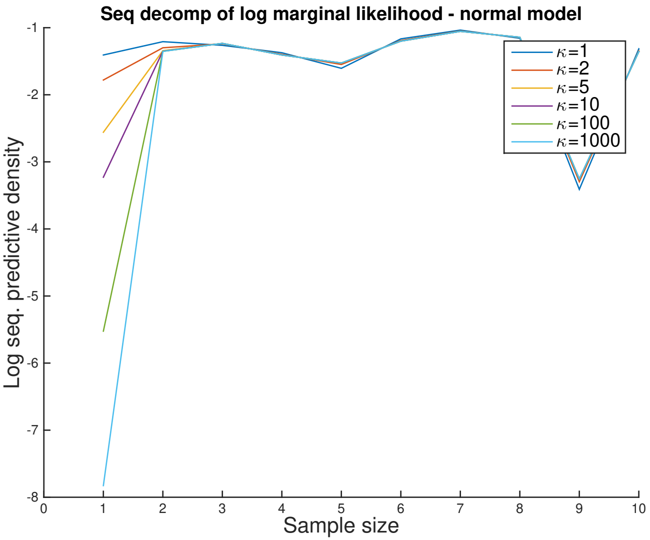
$$y_{i+1}|y_1, ..., y_i \sim N\left[w_i(\kappa_0) \cdot \bar{y}_i, \sigma^2\left(1+\frac{1}{i+\kappa_0}\right)\right]$$

- For $i = 1$: $y_1 \sim N\left[0, \sigma^2\left(1+\frac{1}{\kappa_0}\right)\right]$ can be very sensitive to κ_0.
- For $i = n$: $y_n|y_1, ..., y_{n-1} \overset{approx}{\sim} N\left(\bar{y}_{n-1}, \sigma^2\right)$, not sensitive to κ_0.

First observation is sensitive to $\kappa = 1/\sqrt{\kappa_0}$



First observation is sensitive to κ - zoomed



Log Predictive Score - LPS

- Reduce sensitivity to the prior: sacrifice n^* observations to train the prior into a posterior.
- **Predictive (Density) Score** (**PS**). Decompose $p(y_1, ..., y_n)$ as

$$\underbrace{p(y_1)p(y_2|y_1)\cdots p(y_{n^*}|y_{1:(n^*-1)})}_{\textit{training}}\ \underbrace{p(y_{n^*+1}|y_{1:n^*})\cdots p(y_n|y_{1:(n-1)})}_{\textit{test}}$$

- Usually report on log scale: **Log Predictive Score** (**LPS**).
- Time-series: obvious which data are used for training.
- Cross-sectional data: training-test split by **cross-validation**:

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Computing the marginal likelihood

- **Conjugate models**:

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

- Marginal likelihood is a prior expectation.

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

- **(Bad) Monte Carlo estimate**. Draw $\theta^{(i)} \overset{iid}{\sim} p(\theta)$ and

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

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

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

Computing the marginal likelihood

- **Chib's method** (1995, JASA). Great, but only **Gibbs sampling**.
- **Chib-Jeliazkov** (2001, JASA) generalizes to **MH algorithm** (good for IndepMH, terrible for RWM).
- **Reversible Jump MCMC** (RJMCMC) for model inference. (hard to design proposals, often slow convergence).
- **Bayesian nonparametrics** (e.g. Dirichlet process priors).
- **The Laplace approximation**:

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

  where p is the number of unrestricted parameters.
- **BIC approximation**: $J_{\hat{\theta},\mathrm{y}}$ behaves like $n \cdot I_p$ in large samples

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

$\Pr(M_k|y)$ can be overfident - macroeconomics[3]

Table: Posterior model probabilities - Smets-Wouters DSGE model

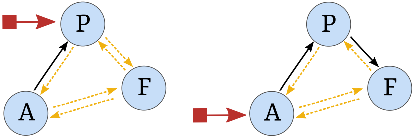
Base	M1	M2	M3	M4	M5	M6	M7	M8
0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00

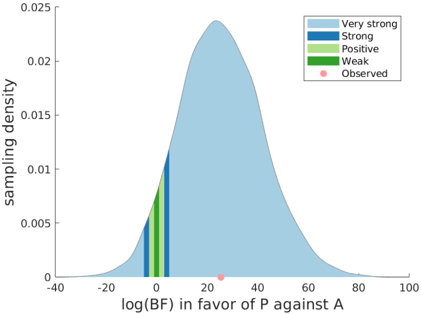
[3] Oelrich et al (2020). When are Bayesian model probabilities overconfident?

$\Pr(M_k|y)$ can be overfident - neuroscience[4]

Table: Posterior model probabilities - Dynamic Causal Models

A	F	P	AF	PA	PF	PAF
0.00	0.00	1.00	0.00	0.00	0.00	0.00





[4] Oelrich et al (2020). When are Bayesian model probabilities overconfident?

Model selection as a decision problem[5]

- **Utility**

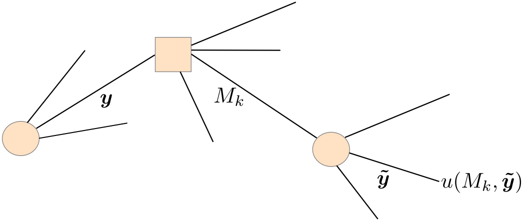
$$u(M_k, \tilde{\boldsymbol{y}})$$

- **Posterior expected utility**

$$\bar{u}(M_k|\boldsymbol{y}) = \int u(M_k, \tilde{\boldsymbol{y}}) p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) d\tilde{\boldsymbol{y}}$$

- **$\mathcal{M}$-closed**

$$p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{k=1}^{K} \Pr(M_k|\boldsymbol{y}) p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y})$$



[5] Bernardo and Smith (1994). Bayesian Theory, Wiley.

Scoring rules

- **Log score**

$$u(M_k, \tilde{\boldsymbol{y}}) = \log p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y})$$

- **Quadratic**

$$u(M_k, \tilde{\boldsymbol{y}}) = 1 - \int \left[p_k(\check{\boldsymbol{y}}|\boldsymbol{y}) - \delta_{\tilde{\boldsymbol{y}}}(\check{\boldsymbol{y}})\right]^2 d\check{\boldsymbol{y}} = 2p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y}) - \int p_k^2(\check{\boldsymbol{y}}|\boldsymbol{y}) d\check{\boldsymbol{y}}$$

- **Proper rule**: $\mathbb{E}_{p(\tilde{\boldsymbol{y}}|M_k)}\left[u(M, \tilde{\boldsymbol{y}})\right]$ is maximized for $M = M_k$.
- **Local rule**: $u(M_k, \tilde{\boldsymbol{y}})$ depends on $p(\boldsymbol{y}|M_k)$ only through the realized value $p(\tilde{\boldsymbol{y}}|M_k)$.
- The log score is the only local and proper scoring rule.
- Quadratic is proper, but not local.
- In **real problems** we may get utility from a model by
  - Predictive performance/profits etc
  - Computational and computer memory considerations.
  - Interpretation and communication abilities.

Choosing a model and an action

- Models are used for taking an action $a \in \mathcal{A} = \{a_1, \ldots, a_J\}$.
- **Utility**

$$u(M_k, a_j, \tilde{\boldsymbol{y}})$$

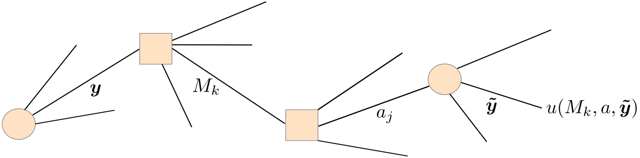
- **Expected utility** of model choice

$$\bar{u}(M_k|\boldsymbol{y}) = \int u(M_k, a^\star(\boldsymbol{y}), \tilde{\boldsymbol{y}}) p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) d\tilde{\boldsymbol{y}}$$

given **optimal action $a^\star(\boldsymbol{y})$ in M_k** obtained by maximizing

$$\bar{u}(a|M_k, \boldsymbol{y}) = \int u(M_k, a_j, \tilde{\boldsymbol{y}}) p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) d\tilde{\boldsymbol{y}}$$

- **Point prediction** $u(M_k, a_j, \tilde{y}) = -(a_j - \tilde{y})^2$ with solution $a_k^\star(\boldsymbol{y}) = \mathbb{E}(\tilde{y}|M_k, \boldsymbol{y})$.



Model averaging

- Not always a need for selecting one model.
- **Utility**

$$u(a_j, \tilde{\boldsymbol{y}})$$

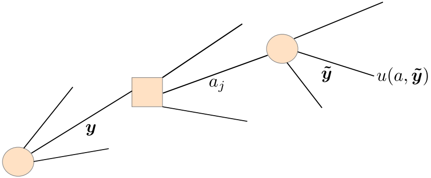
- **Expected utility** of action

$$\bar{u}(a_j|\boldsymbol{y}) = \int u(a_j, \tilde{\boldsymbol{y}}) p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) d\tilde{\boldsymbol{y}}$$

where $p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ is obtained by **model averaging**

$$p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{k=1}^{K} \Pr(M_k|\boldsymbol{y}) p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y})$$

- No model selection, but still **model comparison**: $\Pr(M_k|\boldsymbol{y})$.



Bayesian cross-validation

- $\mathcal{M}$-open: $p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ is not available $\Rightarrow$ Cross-validation [6]
- Generalization performance on new data $\tilde{\boldsymbol{y}} \sim p_\star(\tilde{\boldsymbol{y}})$.
- Here: focus on conditionally iid data $\tilde{y}_i|\theta$.
- Expected log pointwise predictive density for a new dataset of same size as training data $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$

$$\mathrm{elpd} = \sum_{i=1}^{n} \int \log p(\tilde{y}_i|\boldsymbol{y}) p_\star(\tilde{y}_i) d\tilde{y}_i$$

- Over-estimate of elpd from the training data

$$\mathrm{lpd} = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p(\theta|\boldsymbol{y}) d\theta$$

- Computing lpd by posterior simulation $\theta^{(s)} \sim p(\theta|\boldsymbol{y})$

$$\widehat{\mathrm{lpd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right)$$

[6] Bernardo and Smith (1994). Bayesian Theory, Wiley.

Leave-one-out (LOO) cross-validation

- Bayesian LOO estimate of out-of-sample performance

$$\mathrm{elpd}_{\mathrm{loo}} = \sum_{i=1}^{n} \log p(y_i | \boldsymbol{y}_{-i})$$

where

$$p(y_i | \boldsymbol{y}_{-i}) = \int p(y_i | \theta) p(\theta | \boldsymbol{y}_{-i}) d\theta$$

- Computationally costly: **need simulate from n posteriors**.
- **Importance sampling** with $p(\theta | \boldsymbol{y})$ as importance function.
- **Importance weights**

$$r_i^{(s)} = \frac{p(\theta^{(s)} | \boldsymbol{y}_{-i})}{p(\theta^{(s)} | \boldsymbol{y})} \propto \frac{1}{p(y_i | \theta^{(s)})}$$

- **LOO predictive distributions**

$$p(\tilde{y}_i | \boldsymbol{y}_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^{(s)} p(\tilde{y}_i | \theta^{(s)})}{\sum_{s=1}^{S} r_i^{(s)}}$$

Leave-one-out (LOO) cross-validation

- At actual test data $\tilde{y}_i = y_i$. **Harmonic mean** of $p(y_i|\theta^{(s)})$:

$$p(y_i|\boldsymbol{y}_{-i}) \approx \frac{1}{S^{-1}\sum_{s=1}^{S}\frac{1}{p(y_i|\theta^{(s)})}}.$$

- Large weights important for the variance.
- **PSIS-LOO**: Pareto Smoothed Importance Sampling.
  - **Fit generalized Pareto** to largest importance ratios. Get $\hat{k}$.
  - **Replace largest weights** with expected values of order statistics from generalized Pareto.
- **Pareto parameter** $\hat{k}$ can be used to assess the estimate:
  - $k < 1/2$ OK!
  - $1/2 \leq k \leq 1$ Warning!
  - $k > 1$. Red alert!
- Compute $p(y_i|\boldsymbol{y}_{-i})$ by sampling from $p(\theta|\boldsymbol{y}_{-i})$ when $\hat{k} > 0.7$.

WAIC

- Watanabe's Bayesian AIC (WAIC)

$$\widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{lpd}} - \hat{p}_{waic}$$

- p_{waic} is the effective number of parameters

$$p_{waic} = \sum_{i=1}^{n} \mathbb{V}_{p(\theta|\boldsymbol{y})} \left[\log p(y_i|\theta)\right]$$

- $\hat{p}_{waic}$ estimates p_{waic} from simulation $\theta^{(s)} \sim p(\theta|\boldsymbol{y})$.

- WAIC removes some of the bias in $\widehat{\text{lpd}}$ in estimating elpd.

Stacking - Optimal Prediction Pools

- **Model averaging**

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{k=1}^{K} \Pr(M_k|\boldsymbol{y}) p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y})$$

- **Stacking**: Optimize model selection loss wrt ω_k

$$g(\omega) = u\left(\sum_{k=1}^{K} \omega_k p_k(\tilde{\boldsymbol{y}}|\boldsymbol{y}), \tilde{\boldsymbol{y}}\right)$$

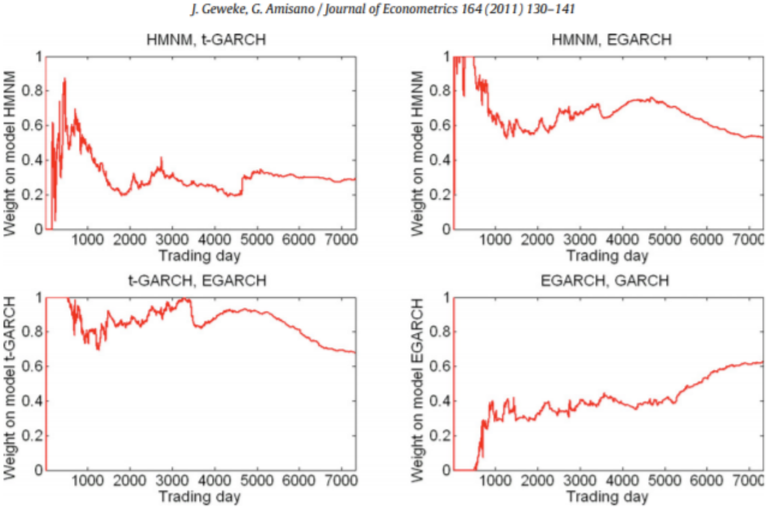
- Stacking uses **log score** out-of-sample (LOO)

$$g_{\text{ls}}(\omega) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \omega_k p_k(y_i|\boldsymbol{y}_{-i})$$

- **Stacking weights converge** as $n \to \infty$.
- Unlike $\Pr(M_k|\boldsymbol{y})$, **does not give zeros-one solutions**.[7]

[7] Geweke and Amisano (2011). Optimal Prediction Pools. Journal of Econometrics.

Stacking - Optimal Prediction Pools[8]



[8] Geweke and Amisano (2011). Optimal Prediction Pools. Journal of Econometrics.