Advanced Bayesian Learning

Lecture 8 - Bayesian Variable Selection and Shrinkage

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Topic overview

- **Bayesian cross-validation**
- **Bayesian variable selection**: Spike-and-slab priors.
- **Shrinkage**: Ridge, Lasso, **Horseshoe**.
- **Stacking and other approaches**

Bayesian cross-validation

- $\mathcal{M}$-open: $p_u(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ is not available $\Rightarrow$ Cross-validation [1]
- 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 $\boldsymbol{y} = (y_1, \dots, y_n)^\top$

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

- 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 $\theta^{(s)} \sim p(\theta|\boldsymbol{y})$

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

[1] 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 $p(y_i|\theta^{(s)})$:

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

- 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 $p(y_i|\mathbf{y}_{-i})$ by sampling from $p(\theta|\mathbf{y}_{-i})$ 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

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

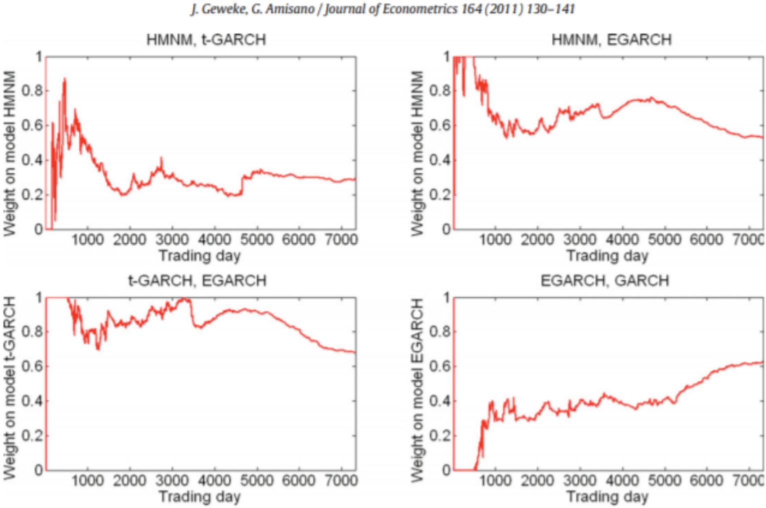
- Stacking uses **log score** out-of-sample (LOO)

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

- **Stacking weights converge** as $n \to \infty$.
- Unlike $\Pr(M_k|\boldsymbol{y})$, **does not give zeros-one solutions**.[2]

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

Stacking - Optimal Prediction Pools[3]



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

Bayesian variable selection

- **Linear regression**:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

- **Binary variable selection indicators** $\mathcal{I} = (I_1, ..., I_p)$

$$I_j = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0 \end{cases}$$

- Example: $\mathcal{I} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so x_3 drops out of the model.

- Usually assume that the intercept is always in the model.

Bayesian variable selection

- Model inference, just crank the Bayesian machine:

$$p(\mathcal{I}|\mathrm{y},\mathrm{X}) \propto p(\mathrm{y}|\mathrm{X},\mathcal{I}) \cdot p(\mathcal{I})$$

- The prior $p(\mathcal{I})$ is typically taken to be

$$I_1, ..., I_p|\theta \overset{iid}{\sim} \mathrm{Bernoulli}(\theta)$$

- θ is the **prior inclusion probability**.
- Hierarchical prior

$$\theta \sim \mathrm{Beta}(a, b).$$

- Marginal **marginal likelihood** for each model ($\mathcal{I}$)

$$p(\mathrm{y}|\mathrm{X},\mathcal{I}) = \int p(\mathrm{y}|\mathrm{X},\mathcal{I},\beta)p(\beta|\mathrm{X},\mathcal{I})d\beta$$

Bayesian variable selection

- Let $\beta_{\mathcal{I}}$ denote the **non-zero** coefficients under $\mathcal{I}$.
- Prior:

$$\beta_{\mathcal{I}}|\sigma^2 \sim N\left(0, \sigma^2\Omega_{\mathcal{I},0}^{-1}\right)$$
$$\sigma^2 \sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Marginal likelihood**

$$p(\mathrm{y}|\mathrm{X}, \mathcal{I}) \propto \left|\mathrm{X}_{\mathcal{I}}'\mathrm{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}^{-1}\right|^{-1/2} \left|\Omega_{\mathcal{I},0}\right|^{1/2} \left(\nu_0\sigma_0^2 + RSS_{\mathcal{I}}\right)^{-(\nu_0+n-1)/2}$$

where $\mathrm{X}_{\mathcal{I}}$ is the covariate matrix for the subset selected by $\mathcal{I}$.

- $RSS_{\mathcal{I}}$ - **Bayesian residual sum of squares** for model with $\mathcal{I}$

$$RSS_{\mathcal{I}} = \mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{X}_{\mathcal{I}}\left(\mathrm{X}_{\mathcal{I}}'\mathrm{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}\right)^{-1}\mathrm{X}_{\mathcal{I}}'\mathrm{y}$$

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! *Ouch*!
- ... but most have essentially zero posterior probability. *Phew*!
- **Simulate** from the joint posterior distribution:

$$p(\beta, \sigma^2, \mathcal{I}|\mathrm{y},\mathrm{X}) = p(\beta, \sigma^2|\mathcal{I}, \mathrm{y},\mathrm{X})p(\mathcal{I}|\mathrm{y}, \mathrm{X}).$$

- Simulate from $p(\mathcal{I}|\mathrm{y}, \mathrm{X})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathrm{y}, \mathrm{X}$
  - Draw $I_2|\mathcal{I}_{-2},\mathrm{y}, \mathrm{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathrm{y}, \mathrm{X}$
- Note that: $Pr(I_i = 0|\mathcal{I}_{-i}, \mathrm{y}, \mathrm{X}) \propto Pr(I_i = 0, \mathcal{I}_{-i}|\mathrm{y}, \mathrm{X})$.
- Compute $p(\mathcal{I}|\mathrm{y}, \mathrm{X}) \propto p(\mathrm{y}|\mathrm{X}, \mathcal{I}) \cdot p(\mathcal{I})$ for $I_i = 0$ and for $I_i = 1$.
- **Model averaging** in a single simulation run.
- If needed, simulate from $p(\beta, \sigma^2|\mathcal{I}, \mathrm{y},\mathrm{X})$ for each draw of $\mathcal{I}$.

Simple general Bayesian variable selection

- The previous algorithm only works when we can compute

$$p(\mathcal{I}|\mathrm{y},\mathrm{X}) = \int p(\beta,\sigma^2,\mathcal{I}|\mathrm{y},\mathrm{X})d\beta d\sigma$$

- **MH** - **propose** β and $\mathcal{I}$ jointly from the proposal distribution

$$q(\beta_p|\beta_c,\mathcal{I}_p)q(\mathcal{I}_p|\mathcal{I}_c)$$

- Main difficulty: how to propose the non-zero elements in β_p?
- Simple approach:
  - Approximate posterior with **all** variables in the model:

$$\beta|\mathrm{y},\mathrm{X} \overset{approx}{\sim} N\left[\hat{\beta}, J_{\mathrm{y}}^{-1}(\hat{\beta})\right]$$

  - Propose β_p from $N\left[\hat{\beta}, J_{\mathrm{y}}^{-1}(\hat{\beta})\right]$, conditional on the zero restrictions implied by $\mathcal{I}_p$. (See GP topic for formulas).

Variable selection - eBay

Table 3. Poisson regression with the number of bids as the response variable

Coeff	Covariate	Mean	SD	Incl. prob.	IF
λ	Const	1.056	0.023	1.000	1.694
	Power	−0.031	0.037	0.010	–
	ID	−0.401	0.093	0.997	1.304
	Sealed	0.444	0.049	1.000	1.379
	MinBlem	−0.027	0.055	0.005	–
	MajBlem	−0.235	0.090	0.111	1.615
	NegScore	0.085	0.056	0.011	–
	LBook_d	−0.113	0.028	0.973	1.416
	MinBidShare_d	−1.894	0.074	1.000	2.797

Variable selection - electricity expenditure

TABLE 6. Posterior means and inclusion probabilities in the one-component split-t model for the electricity expenditure data.

Variable	β_μ	$\mathcal{I}_\mu$	β_ϕ	$\mathcal{I}_\phi$	β_ν	$\mathcal{I}_\nu$	β_λ	$\mathcal{I}_\lambda$
Intercept	256.62	–	3.82	–	2.83	–	1.34	–
log(rooms)	**49.47**	0.90	−0.65	0.43	−0.05	0.04	**0.97**	1.00
log(income)	2.71	0.48	**−0.36**	1.00	−0.05	0.02	**0.55**	1.00
log(people)	**40.62**	1.00	−0.20	0.22	0.06	0.03	**0.34**	1.00
mhtgel	**27.28**	1.00	0.07	0.12	−0.18	0.03	0.13	0.15
sheonly	10.11	0.72	0.01	0.04	**2.10**	0.99	0.04	0.05
whtgel	17.74	0.68	−0.23	0.18	0.33	0.04	**0.82**	0.99
cookel	**27.80**	0.99	−0.19	0.14	0.01	0.04	**0.39**	1.00
poolfilt	−6.50	0.50	−0.11	0.23	1.62	0.07	0.32	0.76
airrev	**14.06**	0.91	0.06	0.07	−0.03	0.03	0.12	0.16
aircond	5.58	0.46	0.03	0.11	0.01	0.03	**0.29**	0.96
mwave	8.08	0.75	−0.38	0.49	−0.39	0.05	0.43	0.49
dish	12.96	0.66	0.08	0.05	1.16	0.04	0.11	0.07
dryer	**19.64**	0.99	0.06	0.12	−0.29	0.05	**0.20**	0.90

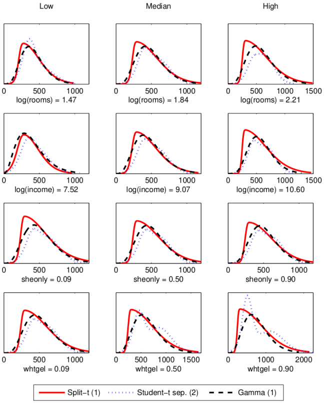


FIGURE 3. Conditional predictive densities for different values of the most important covariates. All other covariates are held fixed at their mean.

Finite step Newton variable selection

- **General regression** model

$$p(y|\boldsymbol{X}, \beta) = \prod_{i=1}^{n} p(y_i|\phi_i),$$

parameters ϕ_i depend on covariates through a **link function**

$$k(\phi_i) = \boldsymbol{x}_i^T \beta.$$

- **GLM** examples:
  - Linear regression. $\phi_i = \mu_i$, $k(\cdot)$ is the identity link.
  - Poisson regression. $\phi_i = \mu_i$, $k(\cdot)$ is the log link.
- Metropolis-Hastings with **Newton proposal**:
  - get posterior mode of β and the Hessian at the mode.
  - Multivariate student-t distribution proposal.
  - Only need first two derivatives of $\log p(y_i|\phi_i)$ wrt ϕ_i.

Finite step Newton proposal for variable selection [4]

- More general regression

$$p(\boldsymbol{y}|\boldsymbol{X}, \beta, \gamma) = \prod_{i=1}^{n} p(y_i|\phi_i, \psi_i)$$

$$k(\phi_i) = \boldsymbol{x}_i^T \beta \quad \text{and} \quad h(\psi_i) = \boldsymbol{x}_i^T \gamma.$$

- Example: Regression with heteroscedastic errors: $\psi_i = \sigma_i^2$.
- Sample from full conditionals $p(\beta|\gamma, \boldsymbol{y}, \boldsymbol{X})$ and $p(\gamma|\beta, \boldsymbol{y}, \boldsymbol{X})$:
  - Newtons method in each updating step. Time-consuming.
  - **Finite-step Newton**. Take only 1-3 steps toward the mode.
- Sample β and selection indicators $\mathcal{I}$ jointly with MCMC.
- How to propose β conditional on $\mathcal{I}$?
  - Finite-step Newton with variable dimension.
  - $k(\phi_i) = \boldsymbol{x}_i^T \beta$ has always the same dimension
  - $k(\phi_{ic}) = \boldsymbol{x}_i^T \beta_c$ and $k(\phi_{ip}) = \boldsymbol{x}_i^T \beta_p$ are expected to be close.

[4] Villani et al (2012). Generalized Smooth Finite Mixtures. Journal of Econometrics.

Variable selection - electricity expenditure

TABLE 6. Posterior means and inclusion probabilities in the one-component split-t model for the electricity expenditure data.

Variable	β_μ	$\mathcal{I}_\mu$	β_ϕ	$\mathcal{I}_\phi$	β_ν	$\mathcal{I}_\nu$	β_λ	$\mathcal{I}_\lambda$
Intercept	256.62	–	3.82	–	2.83	–	1.34	–
log(rooms)	**49.47**	0.90	−0.65	0.43	−0.05	0.04	**0.97**	1.00
log(income)	2.71	0.48	**−0.36**	1.00	−0.05	0.02	**0.55**	1.00
log(people)	**40.62**	1.00	−0.20	0.22	0.06	0.03	**0.34**	1.00
mhtgel	**27.28**	1.00	0.07	0.12	−0.18	0.03	0.13	0.15
sheonly	10.11	0.72	0.01	0.04	**2.10**	0.99	0.04	0.05
whtgel	17.74	0.68	−0.23	0.18	0.33	0.04	**0.82**	0.99
cookel	**27.80**	0.99	−0.19	0.14	0.01	0.04	**0.39**	1.00
poolfilt	−6.50	0.50	−0.11	0.23	1.62	0.07	0.32	0.76
airrev	**14.06**	0.91	0.06	0.07	−0.03	0.03	0.12	0.16
aircond	5.58	0.46	0.03	0.11	0.01	0.03	**0.29**	0.96
mwave	8.08	0.75	−0.38	0.49	−0.39	0.05	0.43	0.49
dish	12.96	0.66	0.08	0.05	1.16	0.04	0.11	0.07
dryer	**19.64**	0.99	0.06	0.12	−0.29	0.05	**0.20**	0.90

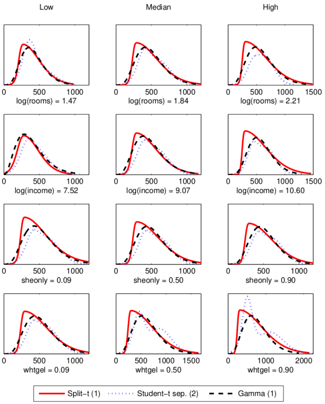


FIGURE 3. Conditional predictive densities for different values of the most important covariates. All other covariates are held fixed at their mean.

Shrinkage priors

- Shrinkage regression priors:
  - $\beta|\sigma^2 \sim N\left(0, \lambda^{-1}\sigma I_p\right)$ gives **Ridge regression**.
  - $\beta|\sigma^2 \sim \text{Laplace}\left(0, \lambda^{-1}\sigma I_p\right)$ + Posterior mode = **Lasso**.
- **Posterior mode Ridge** (when $X^T X = I_p$)

$$\hat{\beta}_{\text{RR}} = \kappa\hat{\beta}_{\text{LS}}, \quad \kappa = \frac{1}{1+\lambda}.$$

- Both Ridge and Lasso are applying **global** shrinkage.
- **Horseshoe prior**: Shrink globally, Act locally.

$$\boldsymbol{y} \sim N\left(\boldsymbol{x}_i^T\beta, \sigma^2 I_n\right)$$

$$\beta_j|\lambda_j, \tau \overset{indep}{\sim} N(0, \lambda_j\tau)$$

$$\lambda_j \overset{iid}{\sim} \text{Cauchy}^+(0,1)$$

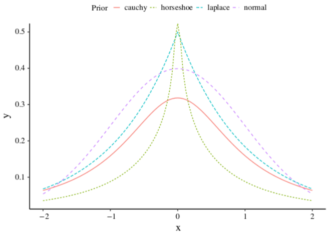
$$\tau \sim \text{Cauchy}^+(0,1)$$

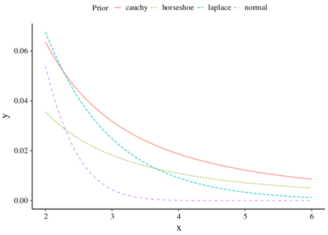
Horseshoe

- **Horseshoe** prior

$$\beta_j|\lambda_j,\tau \overset{indep}{\sim} N(0,\lambda_j\tau) \quad \lambda_j \overset{iid}{\sim} \text{Cauchy}^+(0,1) \quad \tau \sim \text{Cauchy}^+(0,1)$$

- **Marginal prior** $p(\beta_j|\tau)$ not tractable. Tight lower bound.





- Infinite spike at zero $\Longrightarrow$ **sparsity**.
- Heavy but integrable tails $\Rightarrow$ **signals untouched**.
- Horseshoe **posterior mean estimate**: $\hat{\beta}_{\text{HS},j} = \mathbb{E}_{\text{post}}(\kappa_j)\hat{\beta}_{\text{LS},j}$.
- **Shrinkage factor a priori**: $\kappa_j \sim \text{Beta}(1/2, 1/2)$ (*U*-**shaped**).

Model projections

- TBD ...