Bayesian Variable Selection

Practical Workshop

Miquel Torrens

Barcelona Graduate School of Economics

January 13th, 2017

Basics

Notation:

- $\mathbf{t} \in \mathbb{R}^n$: outcome variable for n observations
- $\mathbf{\Phi} \in \mathbb{R}^{n \times p}$: design matrix with p input variables
- M_j: model $j \in \{1, \ldots, 2^p\}$ out of 2^p possible models. Each model has associated one (and only one) binary vector of predictor inclusion $\boldsymbol{\gamma}_j \in \{0,1\}^p$. All models lie in the model space $\mathcal{M}_p$, composed of linear models with at most p active predictors
- $\mathbf{\Phi}_j$: design matrix with only the active predictors under M_j
- $\boldsymbol{\theta}$: set of parameters of the model, for standard linear models this is equivalent to regression coefficients and residual variance, i.e. $\boldsymbol{\theta} \equiv \{\mathbf{w}, q\} \in \mathbb{R}^{p+1}$

Bayesian model selection

For any given model $M_k \in \mathcal{M}_p$, BMS relies on computing:

$$\begin{aligned}
\mathbb{P}(M_k \mid \mathbf{t}) &= \frac{\mathbb{P}(\mathbf{t} \mid M_k)\mathbb{P}(M_k)}{\mathbb{P}(\mathbf{t})} \qquad \text{(Bayes' Theorem)} \\
&= \frac{\mathbb{P}(\mathbf{t} \mid M_k)\mathbb{P}(M_k)}{\sum_{j=1}^{2^p} \mathbb{P}(\mathbf{t} \mid M_j)\mathbb{P}(M_j)} \\
&= \frac{\int \mathbb{P}(\mathbf{t} \mid \boldsymbol{\theta}, M_k)\mathbb{P}(\boldsymbol{\theta} \mid M_k)\mathrm{d}\boldsymbol{\theta}\mathbb{P}(M_k)}{\sum_{j=1}^{2^p} \mathbb{P}(\mathbf{t} \mid M_j)\mathbb{P}(M_j)}
\end{aligned}$$

This calculation poses (at least) **four** delicate issues to solve.

Computing posterior probabilities

Each part of the expression needs to be carefully specified:

$$\mathbb{P}(M_k \mid \mathbf{t}) = \frac{\overbrace{\int \mathbb{P}(\mathbf{t} \mid \boldsymbol{\theta}, M_k)\, \overbrace{\mathbb{P}(\boldsymbol{\theta} \mid M_k)}^{1}\, \mathrm{d}\boldsymbol{\theta}}^{3}\, \overbrace{\mathbb{P}(M_k)}^{2}}{\underbrace{\sum_{j=1}^{2^p} \mathbb{P}(M_j)}_{4}\, \mathbb{P}(\mathbf{t} \mid M_j)}$$

Concerns:

1. Setting a prior on the parameter space: $\mathbb{P}(\boldsymbol{\theta} \mid M_k)$
2. Setting a prior on the model space: $\mathbb{P}(M_k)$
3. Computing the integrated likelihood, i.e. $p+1$ ugly integrals
4. Fully exploring $\mathcal{M}_p$, which has 2^p elements

Priors on parameters (I)

Consider the linear model with $\boldsymbol{\theta} \equiv \{\mathbf{w}, q\}$. Then:

$$\mathbf{t} \mid \mathbf{w}, q, M_k \sim \mathcal{N}\left(\boldsymbol{\Phi}_k \mathbf{w}_k, q\mathbf{I}\right)$$

Usually, we seek conjugacy to compute the integrated likelihood easily:

$$\mathbb{P}(\mathbf{w}, q \mid M_k) = \underbrace{\mathbb{P}(\mathbf{w} \mid q, M_k)}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, q\mathbf{D})} \quad \underbrace{\mathbb{P}(q \mid M_k)}_{q^{-1} \sim \mathrm{Gam}\left(\frac{a_q}{2}, \frac{b_q}{2}\right)}$$

With *conjugate priors*, the integral is in closed form[1]. And so:

- This is nice with 2^p models to deal with...
- ...but it requires to set hyperparameters: $\{a_q, b_q, \mathbf{D}\}$

[1] See Bishop 2.3.7

Priors on parameters (II)

Generally, $\{a_q, b_q\}$ are not very influential for moderate n:

- Set them “small”, but positive (to be proper)
- Say, $a_q = b_q = 10^{-3}$

The battle is with **D**.

[2] See lecture on prior modelling

[3] Main advantage: they favour nested (smaller) models even as n is not “large”, whenever they fit better

Priors on parameters (II)

Generally, $\{a_q, b_q\}$ are not very influential for moderate n:

- Set them "small", but positive (to be proper)
- Say, $a_q = b_q = 10^{-3}$

The battle is with $\mathbf{D}$.

Two families of priors:

1. Local priors
    - Zellner's g-prior[2]
    - Unit information prior: Zellner's with $g = n$
    - ...
2. Non-local priors[3]
    - MOM, eMOM, iMOM (all same idea, changing shape, tails)

Here we will work with Unit Information and iMOM priors as illustration.

[2] See lecture on prior modelling

[3] Main advantage: they favour nested (smaller) models even as n is not "large", whenever they fit better

Priors on models

Let:

- $d_k := \|\boldsymbol{\gamma}_k\|$ (number of active predictors in M_k)
- $\omega := \mathbb{P}(w_i \neq 0)$, for $i = 1, \ldots, p$ uniformly (though not necessarily)

Common specifications on model priors:

1. Uniform for all models: $\mathbb{P}(M_k) = 1/2^p$
2. Binomial: $d_k \sim \text{Binom}(p, \omega)$
    - Models of size $\approx \omega p$ have higher probability
3. Beta-Binomial$(1, 1)$: as if $d_k \sim \text{Binom}(p, \omega)$ and $\omega \sim \text{Unif}(0, 1)$
    - This gives equal probability to models of any size
    - If Beta-Binomial$(1, p)$, then one would favour smaller models
4. Same but with other distributions on d_k: Beta-binomial, Poisson...

A large model space $\mathcal{M}_p$

When p grows, the set of models to explore gets **huge**:

$$\begin{aligned} 1,024 &= 2^{10} \\ 1,073,741,824 &= 2^{30} \\ 1,152,921,504,606,846,976 &= 2^{60} \\ 1,267,650,600,228,229,401,496,703,205,376 &= 2^{100} \\ &\dots \end{aligned}$$

To gain intuition: $2^{265} \approx \#$ particles in the observable Universe.

Therefore:

- For **small** p: enumerate all models and compute $\mathbb{P}(M_j \mid \mathbf{t}), \forall j$
- For **large** p: stochastic search within $\mathcal{M}_p$, i.e. MCMC, Gibbs...

Choosing a model

Comparing two models $\{k, l\}$ is easy, regardless (!) of the size of $\mathcal{M}_p$:

$$\frac{\mathbb{P}(M_k \mid \mathbf{t})}{\mathbb{P}(M_l \mid \mathbf{t})} = \frac{\mathbb{P}(\mathbf{t} \mid M_k)\mathbb{P}(M_k)/\mathbb{P}(\mathbf{t})}{\mathbb{P}(\mathbf{t} \mid M_l)\mathbb{P}(M_l)/\mathbb{P}(\mathbf{t})} = \frac{\mathbb{P}(\mathbf{t} \mid M_k)}{\mathbb{P}(\mathbf{t} \mid M_l)} \frac{\mathbb{P}(M_k)}{\mathbb{P}(M_l)}$$

And even evaluating a set of them[4]:

$$\left\{ \frac{\mathbb{P}(M_{j_1} \mid \mathbf{t})}{\mathbb{P}(M_1 \mid \mathbf{t})}, \frac{\mathbb{P}(M_{j_2} \mid \mathbf{t})}{\mathbb{P}(M_1 \mid \mathbf{t})}, \dots, \frac{\mathbb{P}(M_k \mid \mathbf{t})}{\mathbb{P}(M_1 \mid \mathbf{t})} \right\}$$

But to actually choose one, we need to either (a) compute them all, or (b) at least the most relevant candidates.

Common selection strategies for **explanatory** models:

1. Highest probability model (HPM)
2. Thresholding: fix $\alpha \in (0, 1)$ and pick all $i : \mathbb{P}(\gamma_i = 1 \mid \mathbf{t}) > \alpha$
3. Bayesian false discovery rate (FDR)

[4] Say model M_1 is the null model, with no variables

Bayesian model averaging

For **predictive** models, BMA is the most common choice.

Idea: take the weighted average with our posterior model probs.:

$$\begin{aligned}
\mathbb{E}(t_{n+1} \mid \mathbf{t}) &= \boldsymbol{\phi}_{n+1}^{\mathrm{T}} \mathbb{E}(\mathbf{w} \mid \mathbf{t}) \\
&= \boldsymbol{\phi}_{n+1}^{\mathrm{T}} \left(\sum_{j=1}^{2^p} \mathbb{E}(\mathbf{w} \mid M_j, \mathbf{t}) \mathbb{P}(M_j \mid \mathbf{t}) \right) \\
&= \boldsymbol{\phi}_{n+1}^{\mathrm{T}} \left(\sum_{j=1}^{2^p} \left((\boldsymbol{\Phi}_j^{\mathrm{T}} \boldsymbol{\Phi}_j)^{-1} \boldsymbol{\Phi}_j^{\mathrm{T}} \mathbf{t} \right) \mathbb{P}(M_j \mid \mathbf{t}) \right)
\end{aligned}$$

This prediction minimises error under ℓ_2 loss.

Because this is computationally costly, some approximations exist:

- Median probability model
- Model with closest prediction to BMA on average

Heuristics

A few methods on avoiding (sometimes utopic) computations:

1. Stepwise methods
   - Forward
   - Backward
   - Hybrid
2. Model space restriction
   - LASSO, SCAD, other penalised methods
   - DECO (Wang, Dunson and Leng, 2016)
   - Block-search (Papaspiliopoulos and Rossell, 2016)
3. Pre-screening variables
   - SIS (Fan and Lv, 2008)
   - HOLP (Wang and Leng, 2015)

(Switch to code now)

- Main code: `bvsw.R`
- Auxiliary functions: `bvsf.R`