Bayesian Linear Models
and
Bayesian Variable Selection

Natalie Packham
Berlin School of Economics & Law

Bayesian Statistics Explorations

17 December 2019

Contents

Overview

- Bayesian linear models: Bayesian approach to coefficient estimation in linear models
- Bayesian variable selection: Instead of assiging non-zero coefficients to all independent variables, select the most relevant variables
- Special cases: Ridge Regression ($L2$-regularisation), Lasso ($L1$-regularisation)
- Exposition here uses material mainly from (Fahrmeir et al., 2009) and (Koop, 2003).

Contents

- Overview
- Bayesian linear models
  - VW example
- Bayesian variable selection
  - Bayesian model comparison
  - Bayesian variable selection
  - Connection to Ridge Regression and Lasso
  - VW example

Bayesian linear models

- We consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon.$$

- Under the classical (strong) assumptions,

$$\begin{aligned}\boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}) \\ \hat{\beta} &\sim \mathrm{N}(\beta, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1})\end{aligned}$$

- In a Bayesian setting – see Section 3.5 of (Fahrmeir et al., 2009) – we assume

$$\boldsymbol{y}|\beta, \sigma^2 \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}),$$

with β and σ^2 stochastic.

Conjugate prior

- A conjugate prior is

$$\beta|\sigma^2 \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{M})$$
$$\sigma^2 \sim \mathsf{IG}(a, b),$$

where $\mathsf{IG}(a, b)$ denotes the inverse gamma distribution with parameters a, b.[1]

[1] *Let Y follow a Gamma distribution with parameters a, b, i.e., $Y \sim G(a, b)$. Then $X = 1/Y$ is* inverse gamma *distributed, $X \sim IG(a, b)$. The density of X is*

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp(-b/x), \quad x > 0.$$

Conjugate prior

- Alternative formulation: pair $(\boldsymbol{\beta}, \sigma^2)$ follows NIG distribution (NIG = normal inverse gamma) with parameters $(\boldsymbol{m}, \boldsymbol{M}, a, b)$.[2]

The random variables $(\boldsymbol{\beta}, \sigma^2)$ *follow a* normal inverse gamma distribution *with parameters* $\boldsymbol{m}, \boldsymbol{M}, a, b$ *if*

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{m}, \boldsymbol{M} \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{M})$$
$$\sigma^2 | a, b \sim IG(a, b).$$

The density of $(\boldsymbol{\beta}, \sigma^2) \sim NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ *is*

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2) p(\sigma^2)$$
$$= \frac{1}{(2\pi)^{p/2} (\sigma^2)^{p/2} |\boldsymbol{M}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) \right) \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2} \right).$$

Marginal and posterior distributions

- The marginal distribution of $\boldsymbol{\beta}$ is a multivariate t-distribution:

$$\boldsymbol{\beta} \sim t_{2a}\left(\boldsymbol{m}, \frac{b}{a}\boldsymbol{M}\right).$$

- The *posterior distributions* are given as

$$\begin{aligned} \boldsymbol{\beta}|\cdot &\sim \mathrm{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right) \\ \sigma^2|\cdot &\sim \mathrm{IG}(a', b') \\ (\boldsymbol{\beta}, \sigma^2)|\boldsymbol{y} &\sim \mathrm{NIG}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{M}}, \tilde{a}, \tilde{b}), \end{aligned}$$

Marginal and posterior distributions

- where

$$
\begin{aligned}
\boldsymbol{\Sigma}_\beta &= \left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{M}^{-1}\right)\\
\boldsymbol{\mu}_\beta &= \boldsymbol{\Sigma}_\beta\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{y} + \frac{1}{\sigma^2}\boldsymbol{M}^{-1}\boldsymbol{m}\right)\\
a' &= a + \frac{n}{2} + \frac{p}{2}\\
b' &= b + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}'\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{m}\right)'\boldsymbol{M}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{m}\right)\\
\tilde{\boldsymbol{M}} &= \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1}\right)^{-1}\\
\tilde{\boldsymbol{m}} &= \tilde{\boldsymbol{M}}\left(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y}\right)\\
\tilde{a} &= a + \frac{n}{2}\\
\tilde{b} &= b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} - \tilde{\boldsymbol{m}}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}}\right).
\end{aligned}
$$

- To sample, one can use a Gibbs sampler that draws iteratively from $\boldsymbol{\beta}|\cdot$ and $\sigma^2|\cdot$.

Inference

- A point estimate for $\boldsymbol{\beta}$ is given as

$$\hat{\boldsymbol{\beta}}^B = \mathbb{E}(\boldsymbol{\beta}|\boldsymbol{y}) = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1}\right)^{-1}\left(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y}\right).$$

- Writing $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{X}$, one kann write the Bayesian point estimate as a weighted average of the prior expectation $\boldsymbol{m}$ and the OLS estimate $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}^B = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{m} + \boldsymbol{A}\hat{\boldsymbol{\beta}}.$$

Contents

Overview

Bayesian linear models

- VW example

Bayesian variable selection

- Bayesian model comparison
- Bayesian variable selection
- Connection to Ridge Regression and Lasso
- VW example

Factor model

- Typical problem in finance applications:

  estimate dependencies of individual stocks

- Direct correlation estimator too noisy
- Use (linear) factor model; express a vector of returns $(r_1, \ldots, r_p)$ as

$$r_i = \alpha_i + \beta_{i1} F_1 + \beta_{i2} F_2 + \cdots + \beta_{id} F_d + \varepsilon_i, \qquad i = 1, \ldots, p,$$

  with

  - $F_1, \ldots, F_d$ the **common factor (returns)**,
  - $\beta_{i1}, \beta_{i2}, \ldots, \beta_{id}$ **factor loadings**.

Factor model

- The $p \times p$ covariance matrix of the returns $(r_1, \dots, r_p)$ is given by

$$\Sigma \approx B \, \Omega \, B^T,$$

  where

  - B denotes $p \times d$ matrix of factor loadings,
  - Ω denotes the $d \times d$ covariance matrix of the common factors.

VW example

- In this example, we model VW stock returns as a linear function of MSCI (GICS) industry and MSCI country factors.
- The data set consists of daily returns of
  - MSCI stock indices representing 11 industries and 24 countries;
  - individual stock returns
- Time period: 2002-2018

VW example

```
>>> import numpy as np
>>> import scipy as sp
>>> import pandas as pd
>>> import scipy.stats as scs
>>> import statsmodels.api as sm
>>> import math
>>> import matplotlib.pyplot as plt

>>> data = pd.read_csv('returns.csv', index_col=0)

>>> Y = data['VOW3 GY Equity']
>>> X = data[data.columns[0:34]]
>>> X = sm.add_constant(X)
>>> n, p = X.shape

>>> ols = sm.OLS(Y, X)
>>> result=ols.fit()
>>> #print(result.summary()) # results will be shown together with Bayes' analysis

>>> results = pd.DataFrame(result.params.round(4), columns=['OLS coef'])
>>> results['OLS pval'] = result.pvalues.round(4)
```

VW example

- Setting up the posterior distributions:

```
>>> n, p = X.shape
>>> s = 10 # relatively non-informative prior
>>> lam = result.resid.var()
>>> nu = 25
>>> a = nu/2
>>> b = nu * lam / 2
>>> M = s**2 * sp.identity(p)
>>> Minv = sp.linalg.inv(M)
>>> m = np.zeros(p)
>>> tM = sp.linalg.inv((X.transpose()).dot(X) + Minv)
>>> tMinv = (X.transpose().dot(X) + Minv)
>>> tm = tM.dot(Minv.dot(m) + (X.transpose()).dot(Y))
>>> ta = a + n/2
>>> tb = b + 0.5 * ((Y.transpose()).dot(Y) \
...                  + (m.transpose()).dot(Minv.dot(m)) \
...                  - (tm.transpose()).dot(tMinv.dot(tm)))
```

VW example

- Compare the OLS results with the Bayesian results (hpd = highest posterior density):

```
>>> results['bayes_mean'] = tm.round(4)
>>> results['hpd_2.5'] = [scs.t.ppf(0.025, 2*ta, loc = tm[k], \
...       scale = np.sqrt(((tb/ta) * tM)[k,k])).round(4) for k in range(len(tm))]
>>> results['hpd_97.5'] = [scs.t.ppf(0.975, 2*ta, loc = tm[k], \
...       scale = np.sqrt(((tb/ta) * tM)[k,k])).round(4) for k in range(len(tm))]
```

VW example

```
>>> print(results[:10])
                 OLS coef  OLS pval  bayes_mean  hpd_2.5  hpd_97.5
const              0.0002    0.4813      0.0002  -0.0004    0.0009
MXWOOEN Index      0.1265    0.0863      0.0429  -0.0579    0.1437
MXWOOMT Index     -0.0832    0.3206     -0.0873  -0.2233    0.0488
MXWOOIN Index      0.2915    0.0261      0.1961   0.0001    0.3921
MXWOOCD Index      1.1367    0.0000      0.8980   0.7264    1.0695
MXWOOCS Index      0.0677    0.5662     -0.0010  -0.1765    0.1744
MXWOOHC Index      0.0302    0.7665     -0.0600  -0.2001    0.0801
MXWOOFN Index      0.0502    0.7136     -0.0627  -0.2137    0.0882
MXWOOIT Index     -0.1022    0.3036     -0.1466  -0.2710   -0.0222
MXWOOTC Index     -0.2037    0.0017     -0.2137  -0.3277   -0.0998
>>> print(results[-10:])
          OLS coef  OLS pval  bayes_mean  hpd_2.5  hpd_97.5
MSDUCA      0.0387    0.4697      0.0448  -0.0500    0.1396
MXUS       -0.5276    0.1229     -0.0974  -0.3889    0.1942
MXIL       -0.0230    0.4517     -0.0201  -0.0795    0.0393
MSDUSPT    -0.1003    0.0071     -0.0986  -0.1705   -0.0268
MSDUIT      0.1406    0.0071      0.1377   0.0401    0.2354
MSDUHK      0.0854    0.0228      0.0898   0.0182    0.1615
MSDUIE      0.1028    0.0002      0.1012   0.0474    0.1551
MSDUDE     -0.0410    0.2912     -0.0405  -0.1153    0.0344
MSDUNO      0.0384    0.2416      0.0346  -0.0287    0.0979
MSDUGR      1.1242    0.0000      1.0695   0.9574    1.1815
```

Contents

Overview

Bayesian linear models

- VW example

Bayesian variable selection

- Bayesian model comparison
- Bayesian variable selection
- Connection to Ridge Regression and Lasso
- VW example

Bayesian variable selection

- Bayesian variable selection (BVS) refers to Bayesian methods of choosing the variables to include in a model.
- We consider two methods:
- Bayesian model selection compares posterior probabilities of different models.
- Spike and slab priors include an indicator variable for each coefficient and determines the indicator variable's posterior probability of taking value one.
- References are (Koop, 2003; Fahrmeir et al., 2009, 2013).

Contents

Overview

Bayesian linear models

- VW example

Bayesian variable selection

- Bayesian model comparison
- Bayesian variable selection
- Connection to Ridge Regression and Lasso
- VW example

Bayesian model comparison

- Denote candidate models by M_i, $i = 1, \ldots, m$.
- For example in a linear regression setting, each model M_i includes a specific subset of independent variables and excludes the other variables.
- The posterior model probability is $p(M_i|\boldsymbol{y})$.
- Using Bayes rule:

$$p(M_i|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_i)p(M_i)}{p(\boldsymbol{y})}, \tag{1}$$

  where

  - $p(M_i)$ is the prior model probability
  - $p(\boldsymbol{y}|M_i)$ is called the marginal likelihood;
  - $p(\boldsymbol{y}) = \sum_i p(\boldsymbol{y}|M_i)p(M_i)$.

  (see e.g. Appendix B.5.4 of (Fahrmeir et al., 2013))

Bayesian model comparison

- MCMC sampling makes use of $p(M_i|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_i)p(M_i)$.
- A popular model prior (Section 4.4.3 of (Fahrmeir et al., 2013)) is

$$p(M_i) = \theta^{p_i}(1-\theta)^{p-p_i}, \tag{2}$$

  where
  - $\theta \in (0,1)$,
  - p_i is the number of independent variables (equivalently, the number of coefficients to be estimated) in model M_i,
  - p is the full number of variables / coefficients.
- If $\theta = 1/2$ (uninformative), $p(M_i)$ is equal for all models, in which case

$$p(M_i|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_i).$$

Bayesian model comparison*

- Other choices of θ are motivated as follows:
- Let S denote the (unknown) model size.
- Define indicator variables γ_k, $k = 1, \ldots, p$, with $\gamma_k = 1$ if the coefficient β_k is included in the model and zero otherwise (i.e., $\gamma_k = 1$ iff $\beta_k \neq 0$).
- Then:
  - each γ_k follows a Bernoulli distribution, $\gamma_k \sim B(1, \theta)$,
  - S follows a binomial distribution, $S \sim B(p, \theta)$ (assuming independence of $\gamma_1, \ldots, \gamma_p$).
- Hence, $\mathbb{E}(S) = \theta \cdot p$.
- A natural way to specify θ is to choose the expected model size and set $\theta = \mathbb{E}(S)/p$.

Bayesian model comparison

- Given that posterior density for model M_i is

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, M_i) = \frac{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, M_i) p(\boldsymbol{\beta}, \sigma^2 | M_i)}{p(\boldsymbol{y}|M_i)},$$

  re-arranging gives the marginal likelihood as

$$p(\boldsymbol{y}|M_i) = \frac{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, M_i) p(\boldsymbol{\beta}, \sigma^2 | M_i)}{p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, M_i)},$$

  which consists of likelihood, prior and posterior under M_i.

- A closed formula exists for the NIG case above (see e.g. (Koop, 2003)):

$$p(\boldsymbol{y}|M) = \frac{1}{(2\pi)^{n/2}} \sqrt{\frac{\det(\tilde{\boldsymbol{M}})}{\det(\boldsymbol{M})}} \frac{b^a}{\tilde{b}^{\tilde{a}}} \frac{\Gamma(\tilde{a})}{\Gamma(a)}. \tag{3}$$

Marginal likelihood in VW example

```
>>> from python import BVS
>>> # BVS is my class for Bayesian Variable Selection

>>> bm = BVS.BayesModel(np.array(Y), np.array(X), m, 1 * M, a, b)



>>> # The call below produces an overflow error; this happens a lot as
>>> # often very large and very small constants are involved
>>> bm.marginal_likelihood()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/Users/nat/Documents/GitHub/BayesianStatisticsExplorations/03_BVS/python/BVS.py",
    * self.__b**self.__a / self.__tb**self.__ta * math.gamma(self.__ta) / math.gamma(self.
OverflowError: (34, 'Result too large')
```

- We'll use MCMC later to compute this quantity.
- It also helps to take logs...

Bayesian model comparison

- Define indicator variables γ_k, $k = 1, \dots, p$, with $\gamma_k = 1$ if the coefficient β_k is included in the model and zero otherwise (i.e., $\gamma_k = 1$ iff $\beta_k \neq 0$).
- Posteriori probability of γ_k across all models M_γ given by posterior inclusion probabilities (PIP):

$$\mathbf{P}(\gamma_k = 1|\boldsymbol{y}) = \sum_{\beta_k \in M_\gamma} \mathbf{P}(M_\gamma|\boldsymbol{y}). \tag{4}$$

- If number of parameters p is large, then full calculation of 2^p posterior model probabilities is infeasible.
- $\Rightarrow$ Use Monte Carlo simulation or MCMC.
- PIP's are determined as the frequency of visited models including the covariate relative to the total number of visited models (see page 245 of (Fahrmeir et al., 2013)).

Spike and slab prior (I)*

- (Fahrmeir et al., 2013), pp. 250, demonstrate that, instead of choosing a fixed prior parameter θ, a Beta-distributed hyper prior leads to a more uninformative prior.
- Idea is that parameter θ itself is modelled by a $\text{Beta}(a, b)$ distribution.
- This leads to modification of the prior model probability, cf. (2):

$$p(M_i) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+p_i)\Gamma(b+p-p_i)}{\Gamma(a+b+p)}. \tag{5}$$

Spike and slab prior (I)*

- The two approaches can be compared through the expectation of S, the model size, where

$$\mathbb{E}(S) = \theta \cdot p \qquad \text{(fixed prior)}$$
$$\mathbb{E}(S) = \frac{a}{a+b} \cdot p \qquad \text{(Beta hyper prior).}$$

- Choosing $a = 1$ one can infer $b = (p - \mathbb{E}S)/(\mathbb{E}S)$.
- This approach is equal to a simple spike and slab prior, see (Fahrmeir et al., 2013), p. 253.

Contents

Overview

Bayesian linear models

- VW example

Bayesian variable selection

- Bayesian model comparison
- Bayesian variable selection
- Connection to Ridge Regression and Lasso
- VW example

Spike and slab prior (II)

- In this section, we follow (George and McCulloch, 1997).
- The prior in this method is called a spike and slab prior.
- In Bayesian variable selection (BVS), p indicator variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ are added to the model, indicating that β_i is included in the model if $\gamma_i = 1$ and excluded if $\gamma_i = 0$.
- A conjugate prior is given by

$$\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}|\sigma, \boldsymbol{\gamma})\pi(\sigma|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}),$$

with

$$\boldsymbol{\beta}|\sigma, \boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma^2 D_\gamma^* R_\gamma D_\gamma^*),$$

where D_γ^* is a diagonal matrix and R_γ is a correlation matrix. The specification of D_γ^* is given below.

Spike and slab prior (II)

- For the residual variance, using the notation of (George and McCulloch, 1997),

$$\sigma^2 | \boldsymbol{\gamma} \sim \mathrm{IG}(\nu/2, \lambda_\gamma/2).$$

- The parameters have the following interpretation:
  - λ_γ is a prior estimate of σ^2,
  - ν is a prior “sample size” associated with this estimate (in this sense, ν quantifies the uncertainty associated with λ_γ).
- Converting to the previous notation, we have $a = \nu/2$ and $b = \nu \cdot \lambda_\gamma/2$.
- (George and McCulloch, 1997) recommend to choose $\lambda_\gamma = s^2_{\mathrm{L}S}$ where $s^2_{\mathrm{L}S}$ is the least squares estimate of σ^2 in the full model.

Spike and slab prior (II)

- Finally, the prior of γ is conveniently modelled as

  $$\pi(\gamma) = \prod w_i^{\gamma_i}(1 - w_i)^{(1-\gamma_i)} \tag{6}$$

  (although any other prior can be considered as well). If $w_i = \theta$, $i = 1, \ldots, p$, then the model prior corresponds to the model prior in (2).

Spike and slab prior (II)

- The i-th diagonal element of D^*_γ is specified by

$$(D_\gamma^{*2})_{ii} = \begin{cases} v^*_{0\gamma_{(i)}}, & \text{if } \gamma_i = 0, \\ v^*_{1\gamma_{(k)}}, & \text{if } \gamma_i = 1, \end{cases}$$

  where $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(k)}}$ are "small" and "large", respectively.

- More specifically, from

$$\pi(\beta_i|\sigma,\gamma) = (1-\gamma_i)\underbrace{\mathrm{N}(0,\sigma^2 v^*_{0\gamma_{(k)}})}_{\text{spike}} + \gamma_i \underbrace{\mathrm{N}(0,\sigma^2 v^*_{1\gamma_{(k)}})}_{\text{slab}},$$

  one can choose $v^*_{0\gamma_{(k)}}$ and $v^*_{1\gamma_{(k)}}$ proportional to the variance of $\hat\beta$ in the OLS estimate, i.e.,

$$v^*_{\cdot\gamma_{(k)}} = c_\cdot \frac{\sigma^2_{\hat\beta}}{\hat\sigma^2},$$

  with c_0, c_1 constants and where $\hat\sigma^2$ refers to an estimate of σ^2, such as the OLS estimate of the residual variance.

Spike and slab prior (II)

- With this specification, it is possible to calculate

$$\pi(\gamma|\boldsymbol{y}) \propto g(\gamma) = \left|\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right|^{-1/2} \left|D_\gamma^* R_\gamma D_\gamma^*\right|^{-1/2} (\nu\lambda + S_\gamma^2)^{(n+\nu)/2} \pi(\gamma), \tag{7}$$

where

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ (D_\gamma^* R_\gamma D_\gamma^*)^{-1/2} \end{pmatrix}$$

$$\tilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ 0 \end{pmatrix}$$

$$S_\gamma^2 = \tilde{\boldsymbol{Y}}'\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}'\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{Y}}.$$

Contents

Overview

Bayesian linear models

VW example

Bayesian variable selection

Bayesian model comparison

Bayesian variable selection

Connection to Ridge Regression and Lasso

VW example

Connection to ridge regression and Lasso

- See Chapter 6 of (James et al., 2013).
- Let (assuming that X is fixed):

$$p(\beta|X, Y) \propto f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta)$$

- Assume that $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$, with g a density.
- Ridge regression: g normal distribution with mean zero and s.d. a function of the $L2$-constraint λ:
  - $\Rightarrow$ Posterior mode of β equals ridge regression solution
- Lasso: g double-exponential (Laplace) with mean zero and scale parameter a function of the $L1$-constraint λ:
  - $\Rightarrow$ Posterior mode or β equals lasso solution

Contents

Overview

Bayesian linear models

- VW example

Bayesian variable selection

- Bayesian model comparison
- Bayesian variable selection
- Connection to Ridge Regression and Lasso
- VW example

VW example

- Marginal likelihood via MCMC:

```
>>> # simulate marginal likelihood via MCMC
>>> x, sc = bm.simulate_posterior_probability(nsimulations=5000, rel=True, \
...         theta=8/34)
```

- Gamma posterior from (7):

```
>>> v0 = 0.001 * result.bse / result.resid.std()
>>> v1 = 1 * result.bse /result.resid.std()
>>> v0.index=range(p)
>>> v1.index=range(p)
>>> x2, sc2 = bm.simulate_gamma_posterior(nu, lam, v0, v1, theta=8/34, \
...         nsimulations=3000, rel=True)
```

VW example

```
>>> res = None
>>> for k in range(p):
...     if k==0:
...         res = pd.DataFrame([[X.columns[k], \
...                              x[np.where(sc[:,k]==1)[0]].shape[0]/x.shape[0], \
...                              x2[np.where(sc2[:,k]==1)[0]].shape[0]/x2.shape[0], \
...                              result.pvalues.iloc[k]]])
...     else:
...         res = res.append([[X.columns[k], \
...                              x[np.where(sc[:,k]==1)[0]].shape[0]/x.shape[0],\
...                              x2[np.where(sc2[:,k]==1)[0]].shape[0]/x2.shape[0], \
...                              result.pvalues.iloc[k]]])
...
>>> res.index = range(p)
>>> res.columns = ['coef', 'PIP', 'BVS', 'pvalue']
```

VW example

- The median probability model selects those variables that have probability greater than 0.5.
- (Barbieri and Berger, 2004) show that the median probability model is often optimal in terms of prediction.
- (Note that the median probability model depends on the choice of θ, which specifies the probability of including / excluding a variable)

VW example

```
>>> print(res[res['PIP']>0.5].round(4))
                coef     PIP     BVS  pvalue
4   MXW00CD Index  1.0000  1.0000  0.0000
9   MXW00TC Index  0.9848  0.9900  0.0017
10  MXW00UT Index  0.9996  1.0000  0.0000
18        MSDUSZ  0.6788  0.4940  0.0105
19        MSDUAT  0.7998  0.7613  0.0000
34        MSDUGR  1.0000  1.0000  0.0000
```

VW example

```
>>> print(res[res['BVS']>0.5].round(4))
             coef     PIP     BVS  pvalue
3   MXWO0IN Index  0.0890  0.6687  0.0261
4   MXWO0CD Index  1.0000  1.0000  0.0000
9   MXWO0TC Index  0.9848  0.9900  0.0017
10  MXWO0UT Index  0.9996  1.0000  0.0000
11  MXWO0RE Index  0.1344  0.7783  0.0043
19         MSDUAT  0.7998  0.7613  0.0000
23         MSDUJN  0.1216  0.8843  0.0160
28        MSDUSPT  0.2942  0.6920  0.0071
29         MSDUIT  0.1734  0.6517  0.0071
31         MSDUIE  0.3484  0.6263  0.0002
34         MSDUGR  1.0000  1.0000  0.0000
```

VW example

- In application, we estimated PIP's for several hundreds of stocks and across a rolling time window.
- To ensure some minimal stability, prior knowledge about country and industry is included (e.g. Consumer Discretionary and Germany are always included for VW).
- This kind of "fine-tuning" is not possible with Lasso.

References

Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression*. Springer, 2nd edition, 2009.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2013.

Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Gary M Koop. *Bayesian Econometrics*. John Wiley & Sons Inc., 2003.

Thank you!

Prof. Dr. Natalie Packham
Professor of Mathematics and Statistics
Berlin School of Economics and Law
Badensche Str. 52
10825 Berlin

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law