BAYESIAN LINEAR REGRESSION

EFIM ABRIKOSOV

1. Framework

1.1. Notations.

$$y = x \cdot \boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1.1}$$

(1) Individual observations $(x, y) \in \mathbb{R}^k \times \mathbb{R}$
(2) Observed data $(X, Y) \in \mathbb{R}^{n \times k} \times \mathbb{R}^n$
(3) Linear regression weights $\boldsymbol{\beta} \in \mathbb{R}^k$
(4) Model parameter distribution mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$
(5) Observation error variance σ^2

1.2. Model Assumptions.

(1) Observations (x, y) satisfy linear relation 1.1
(2) Observation errors are independent, normally distributed with mean zero and variance σ^2
(3) For posterior estimation, observations X must have full rank
(4) For known error variance σ^2, the prior on the space of parameters $\boldsymbol{\beta} \in \mathbb{R}^k$ is $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$
(5) For unknown error variance, the prior for $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^{k+1}$ has σ^2 following inverse gamma distribution with parameters (a_0, b_0), and conditional distribution for linear relation weights $f(\boldsymbol{\beta} \mid \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$

2. Known Observation Variance with Linear Weights Prior

2.1. Summary of Results.

Proposition 2.1 (Posterior Paramer Distribution with Known Variance)**.** *The posterior distribution of model parameters is normal $f(\boldsymbol{\beta} \mid X, Y) = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2 \boldsymbol{\Sigma}_1)$ with parameters:*

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^T X \tag{2.1}$$

$$\mu_1 = \boldsymbol{\Sigma}_1 \left(X^T X \widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1} \mu_0 \right) \tag{2.2}$$

$$\widehat{\boldsymbol{\beta}} = \left(X^T X \right)^{-1} X^T Y \tag{2.3}$$

Proposition 2.2 (Posterior Predictive Distribution with Known Variance)**.** *For an observation (x, y) and its expectation $(x, \widehat{y} \overset{\text{def}}{=} x \cdot \boldsymbol{\beta})$, posterior conditional distributions of y and $\widehat{y}$ are normal with parameters given below*

$$f(y \mid x, X, Y) = \mathcal{N}\left(x\boldsymbol{\mu}_1, \sigma^2 \left(1 + x \left(X^T X \right)^{-1} x^T \right) \right) \tag{2.4}$$

$$f(\widehat{y} \mid x, X, Y) = \mathcal{N}\left(x\boldsymbol{\mu}_1, \sigma^2 x \left(X^T X \right)^{-1} x^T \right) \tag{2.5}$$

In particular, the variance of predictive distributions does not depend on $\boldsymbol{\Sigma}_0$.

Proposition 2.3 (Bayesian Regresssion under Known Vairance and Uninformative Prior)**.** *If the prior parameter $\boldsymbol{\beta}$ distribution is uninformative, i.e. $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0^{-1} = 0$, then posterior distribution of model parameters recovers standard OLS formulas:*

$$\begin{aligned} &\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \sigma^2 \left(X^T X\right)^{-1}\right) \\ &\boldsymbol{\mu}_1 = \left(X^T X\right)^{-1} X^T Y \end{aligned} \tag{2.6}$$

In addition, posterior predictive distributions 2.4 and 2.5 coincide with standard OLS formulas.

Proof. Define for convenience $\boldsymbol{\Lambda}_0 \stackrel{\text{def}}{=} \boldsymbol{\Sigma}_0^{-1}$. Using $f(\beta \mid X, Y) \propto f(X, Y \mid \boldsymbol{\beta}) f(\boldsymbol{\beta})$ and taking logarithms one has:

$$\begin{aligned} \ln f(\boldsymbol{\beta} \mid X, Y) + \text{const} = &-\left(\frac{n}{2}\ln\sigma^2 + \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln\boldsymbol{\Sigma}_0 + \frac{k}{2}\ln 2\pi\right) - \\ &-\frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta}) - \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0(\boldsymbol{\beta} - \boldsymbol{\mu}_0) = \\ = &-\left(\frac{n}{2}\ln\sigma^2 + \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln\boldsymbol{\Sigma}_0 + \frac{k}{2}\ln 2\pi\right) - \\ -\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^T\left(X^TX + \boldsymbol{\Lambda}_0\right)\boldsymbol{\beta} - (Y^TX + \boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0)\boldsymbol{\beta} - \boldsymbol{\beta}^T\left(X^TY + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right) + Y^TY + \boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right)& \end{aligned}$$

It suffices to show that this expression is a quadratic form $-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_1(\boldsymbol{\beta} - \boldsymbol{\mu}_1)$ up to an additive term independent of $\boldsymbol{\beta}$. Here $\boldsymbol{\mu}_1$ and $\boldsymbol{\Lambda}_1$ are as in 2.1-2.3. This follows immediately from the lemma on completing squares:

Lemma 2.4. *For any symmetric quadratic form Q, liner form L and vector v*

$$v^TQv - v^TL - L^Tv = (v - Q^{-1}L)^TQ(v - Q^{-1}L) - L^TQ^{-1}L$$

To find the posterior predictive distribution 2.4, we integrate out parameter $\boldsymbol{\beta}$:

$$\int\limits_{\boldsymbol{\beta}\in\mathbb{R}^k} f(y \mid x, X, Y, \boldsymbol{\beta}) f(\boldsymbol{\beta} \mid X, Y) d\boldsymbol{\beta} = \int\limits_{\boldsymbol{\beta}\in\mathbb{R}^k} -\frac{\det\boldsymbol{\Lambda}_1}{(2\pi\sigma^2)^{\frac{1+k}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left((y - x\beta)^2 + (\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_1(\boldsymbol{\beta} - \boldsymbol{\mu}_1)\right)\right) d\boldsymbol{\beta}$$

□

Posterior distribution $f(\boldsymbol{\beta} \mid Y, X) = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\boldsymbol{\Sigma}_1)$:

$$f(\boldsymbol{\beta} \mid Y, X) = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\boldsymbol{\Sigma}_1) \tag{2.7}$$

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^TX \tag{2.8}$$

$$\mu_1 = \boldsymbol{\Sigma}_1\left(X^TX\widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1}\mu_0\right) \tag{2.9}$$

$$\widehat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^TY \tag{2.10}$$

Posterior prediction distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta}$:

$$f(\widehat{y} \mid Y, X, x) = \mathcal{N}\left(x\boldsymbol{\beta}_1, \sigma^2 x\left(X^TX\right)^{-1}x^T\right) \tag{2.11}$$

Posterior observation distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta} + e$:

$$f(y \mid Y, X, x) = \mathcal{N}\left(x\boldsymbol{\beta}_1, \sigma^2\left(1 + x\left(X^TX\right)^{-1}x^T\right)\right) \tag{2.12}$$

[Meaning?][Confidence interval, for a fixed value $\underline{\boldsymbol{\beta}}$ and a linear constraint $\boldsymbol{c} \in \mathbb{R}^k$:

$$\frac{\boldsymbol{c}(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)}{\sigma\sqrt{(\boldsymbol{c}\boldsymbol{\Sigma}_1\boldsymbol{c}^T)}} \sim \mathcal{N}(0,1) \tag{2.13}$$

Joint f-test for a set of linear constraints $\boldsymbol{C} \in \mathbb{R}^{l\times k}$]

2.2. **Uninformative Prior.** With the prior $\Lambda_0 \equiv \boldsymbol{\Sigma}_0^{-1} = 0$, the posterior 2.7 reduces to:

$$f(\boldsymbol{\beta} \mid Y, X) \sim \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Sigma}_1) \tag{2.14}$$

$$\boldsymbol{\Sigma}_1 = \left(X^T X\right)^{-1} \tag{2.15}$$

$$\boldsymbol{\beta}_1 = \left(X^T X\right)^{-1} X^T Y \tag{2.16}$$

which is the standard result obtained in classical OLS set up.

The standard prediction interval for $\widehat{y}(x)$ and the confidence interval for an observation $y(x)$ follow from normal distributions in 2.11 and 2.12 respectively.

3. Conjugate Priors For Observation Variance and Linear Weights

3.1. **Setup.**

$$f(Y, X \mid \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta})\right) \tag{3.1}$$

$$f(\boldsymbol{\beta} \mid \sigma^2) = |2\pi\boldsymbol{\Sigma}_0|^{-1} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\right) \tag{3.2}$$

$$f(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right) \tag{3.3}$$

Alternatively $f(\sigma^2)$ can be written as scaled inverse chi-squared distribution with parameters $\left(\nu_0, \tau_0^2\right) = \left(2a_0, \frac{b_0}{a_0}\right)$

3.2. **Summary of Results.** Posterior distribution of $\boldsymbol{\beta}$:

$$f(\sigma^2 \mid Y, X) = \text{Inv-}\Gamma(a_1, b_1) \tag{3.4}$$

$$a_1 = a_0 + \frac{n}{2} \tag{3.5}$$

$$b_1 = b_0 + \frac{1}{2}\left(Y^T Y + \boldsymbol{\beta}_0\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0^T - \boldsymbol{\beta}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}_1^T\right) \tag{3.6}$$

$$f(\boldsymbol{\beta} \mid Y, X, \sigma^2) = \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Sigma}_1) \tag{3.7}$$

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^T X \tag{3.8}$$

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_1\left(X^T X\widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right) \tag{3.9}$$

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y \tag{3.10}$$

Posterior prediction distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta}$:

$$f(\widehat{y} \mid Y, X, x) \propto \left(1 + \frac{a_1\left(y - x\boldsymbol{\beta}_1\right)^2}{vb_1}\frac{1}{2a_1}\right)^{-\frac{2a_1+1}{2}} \tag{3.11}$$

$$v = \left(1 - x\left(\mathbf{\Sigma}_1 + x^T x\right)^{-1} x^T\right)^{-1} \tag{3.12}$$

This is Student's t-distribution on $y - x\boldsymbol{\beta}_1$ with scale $\frac{vb_1}{a_1}$ and $2a_1$ degrees of freedom.

Posterior observation distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta} + e$:

$$f(y \mid Y, X, x) = ?? \tag{3.13}$$