

BAYESIAN LINEAR REGRESSION

EFIM ABRIKOSOV

1. FRAMEWORK

1.1. Notations.

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

- (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 $\beta \in \mathbb{R}^k$
- (4) Model parameter distribution mean $\mu \in \mathbb{R}^k$ and covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$
- (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 $\beta \in \mathbb{R}^k$ is $\mathcal{N}(\mu, \sigma^2 \Sigma)$
- (5) For unknown error variance, the prior for $(\beta, \sigma^2) \in \mathbb{R}^{k+1}$ has σ^2 following inverse gamma distribution with parameters (a_0, b_0) , and conditional distribution for linear relation weights $f(\beta \mid \sigma^2) = \mathcal{N}(\mu, \sigma^2 \Sigma)$

2. GAUSSIAN PRIOR WITH KNOWN VARIANCE

2.1. Summary of Results.

Proposition 2.1 (Posterior paramer distribution with known variance). *The posterior distribution of model parameters is normal $f(\beta \mid X, Y) = \mathcal{N}(\mu_1, \sigma^2 \Sigma_1)$ with parameters:*

$$(2.1) \quad \Sigma_1^{-1} = \Sigma_0^{-1} + X^T X$$

$$(2.2) \quad \mu_1 = \Sigma_1 \left(X^T X \hat{\beta} + \Sigma_0^{-1} \mu_0 \right)$$

$$(2.3) \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

Proposition 2.2 (Posterior predictive distribution with known variance). *For an observation (x, y) and its expectation $(x, \hat{y} \stackrel{\text{def}}{=} x \cdot \beta)$, posterior conditional distributions of y and \hat{y} are normal with parameters given below*

$$(2.4) \quad f(y \mid x, X, Y) = \mathcal{N}(x \mu_1, \sigma^2 (1 + x \Sigma_1 x^T))$$

$$(2.5) \quad f(\hat{y} \mid x, X, Y) = \mathcal{N}(x \mu_1, \sigma^2 x \Sigma_1 x^T)$$

Proposition 2.3 (Bayesian regression under known variance and uninformative prior). *If the prior parameter β distribution is uninformative, i.e. $\mu_0 = 0$ and $\Sigma_0^{-1} = 0$, then posterior distributions of model parameters and predictive distributions recover standard OLS formulas:*

$$(2.6) \quad \beta \sim \mathcal{N}(\mu_1, \sigma^2 (X^T X)^{-1})$$

$$(2.7) \quad f(y | x, X, Y) = \mathcal{N}(x\mu_1, \sigma^2 (1 + x(X^T X)^{-1}x^T))$$

$$(2.8) \quad f(\hat{y} | x, X, Y) = \mathcal{N}(x\mu_1, \sigma^2 x(X^T X)^{-1}x^T)$$

Proof of Proposition 2.1. Define for convenience $\Lambda_0 \stackrel{\text{def}}{=} \Sigma_0^{-1}$. Using $f(\beta | X, Y) \propto f(X, Y | \beta)f(\beta)$ and taking logarithms one has:

$$\begin{aligned} \ln f(\beta | X, Y) + \text{const} &= - \left(\frac{n}{2} \ln \sigma^2 + \frac{n}{2} \ln 2\pi + \frac{1}{2} \ln \Sigma_0 + \frac{k}{2} \ln 2\pi \right) - \\ &\quad - \frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) - \frac{1}{2\sigma^2} (\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0) = \\ &= - \left(\frac{n}{2} \ln \sigma^2 + \frac{n}{2} \ln 2\pi + \frac{1}{2} \ln \Sigma_0 + \frac{k}{2} \ln 2\pi \right) - \\ &\quad - \frac{1}{2\sigma^2} \left(\beta^T (X^T X + \Lambda_0) \beta - (Y^T X + \mu_0^T \Lambda_0) \beta - \beta^T (X^T Y + \Lambda_0 \mu_0) + Y^T Y + \mu_0^T \Lambda_0 \mu_0 \right) \end{aligned}$$

It suffices to show that this expression is a quadratic form $-\frac{1}{2\sigma^2}(\beta - \mu_1)^T \Lambda_1 (\beta - \mu_1)$ up to an additive term independent of β . Here μ_1 and Λ_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 , linear form L and vector v*

$$v^T Q v - v^T L - L^T v = (v - Q^{-1}L)^T Q (v - Q^{-1}L) - L^T Q^{-1}L$$

□

Proof of Proposition 2.2. To find the posterior predictive distribution 2.4, we integrate out parameter β :

$$\begin{aligned} f(y | x, X, Y) &= \int_{\beta \in \mathbb{R}^k} f(y | x, X, Y, \beta) f(\beta | X, Y) d\beta = \\ &= \int_{\beta \in \mathbb{R}^k} - \frac{(\det \Lambda_1)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1+k}{2}}} \exp \left(-\frac{1}{2\sigma^2} ((y - x\beta)^2 + (\beta - \mu_1)^T \Lambda_1 (\beta - \mu_1)) \right) d\beta \end{aligned}$$

Set $\tilde{\beta} \stackrel{\text{def}}{=} \beta - \mu_1$ and $\tilde{y} \stackrel{\text{def}}{=} y - x\mu_1$. The expression under the exponential can be rewritten as:

$$\begin{aligned} (y - x\beta)^2 + (\beta - \mu_1)^T \Lambda_1 (\beta - \mu_1) &= (\tilde{y} - x\tilde{\beta})^2 + \tilde{\beta}^T \Lambda_1 \tilde{\beta} = \\ &= \tilde{y}^2 - \tilde{y}x(\Lambda_1 + x^T x)^{-1}x^T \tilde{y} + \left(\tilde{\beta} - (\Lambda_1 + x^T x)^{-1}x^T \tilde{y} \right)^T (\Lambda_1 + x^T x) \left(\tilde{\beta} - (\Lambda_1 + x^T x)^{-1}x^T \tilde{y} \right) \end{aligned}$$

Using that for any positive definite form Q the integral $\exp(-(\beta - \mu)^T Q (\beta - \mu))$ is independent of μ , we find that up to a multiplicative constant:

$$f(y | x, X, Y) = \text{const} \cdot \exp \left(-\frac{\tilde{y}^2}{2\sigma^2} \left(1 - x(\Lambda_1 + x^T x)^{-1}x^T \right) \right) = \text{const} \cdot \exp \left(-\frac{\tilde{y}^2}{2\sigma^2} (1 + x^T \Sigma_1 x)^{-1} \right)$$

Consequently, $\tilde{y} = y - x\mu_1$ is normally distributed with variance as prescribed by 2.4. The second equality above follows from Sherman-Morrison formula.

Theorem 2.5 (Sherman-Morrison formula). *Suppose $A \in \mathbb{R}^{k \times k}$ is an invertible matrix and $u, v \in \mathbb{R}^k$ are vectors. Then $A + uv^T$ is invertible iff $1 + v^T A^{-1} u \neq 0$. In this case,*

$$(2.9) \quad (A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}$$

$$(2.10) \quad v^T (A + uv^T)^{-1} u = \frac{v^T A^{-1}u}{1 + v^T A^{-1}u}$$

The proof of 2.5 may be seen as a direct consequence of the general result on convolution of multivariate normal distributions. The result is well-known and is usually demonstrated using Fourier transform. We give an “elementary” proof in Appendix A \square

Proof of Proposition 2.3. Straightforward by substituting $\mu_0 = 0$ and $\Sigma_0^{-1} = 0$ in propositions 2.1 and 2.2 \square

3. VARIANCE SCALE INVERSE GAMMA PRIOR

4. CONJUGATE PRIORS FOR OBSERVATION VARIANCE AND LINEAR WEIGHTS

4.1. Setup.

$$(4.1) \quad f(Y, X \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

$$(4.2) \quad f(\beta \mid \sigma^2) = |2\pi\Sigma_0|^{-1} \exp\left(-\frac{1}{2\sigma^2}(\beta - \beta_0)\Sigma_0^{-1}(\beta - \beta_0)^T\right)$$

$$(4.3) \quad f(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right)$$

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

4.2. Summary of Results. Posterior distribution of β :

$$(4.4) \quad f(\sigma^2 \mid Y, X) = \text{Inv-}\Gamma(a_1, b_1)$$

$$(4.5) \quad a_1 = a_0 + \frac{n}{2}$$

$$(4.6) \quad b_1 = b_0 + \frac{1}{2}\left(Y^T Y + \beta_0 \Sigma_0^{-1} \beta_0^T - \beta_1 \Sigma_1^{-1} \beta_1^T\right)$$

$$(4.7) \quad f(\beta \mid Y, X, \sigma^2) = \mathcal{N}(\beta_1, \sigma^2 \Sigma_1)$$

$$(4.8) \quad \Sigma_1^{-1} = \Sigma_0^{-1} + X^T X$$

$$(4.9) \quad \beta_1 = \Sigma_1 \left(X^T X \hat{\beta} + \Sigma_0^{-1} \beta_0 \right)$$

$$(4.10) \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

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

$$(4.11) \quad f(\hat{y} \mid Y, X, x) \propto \left(1 + \frac{a_1 (y - x\beta_1)^2}{vb_1} \frac{1}{2a_1}\right)^{-\frac{2a_1+1}{2}}$$

$$(4.12) \quad v = \left(1 - x \left(\mathbf{\Sigma}_1 + x^T x\right)^{-1} x^T\right)^{-1}$$

This is Student's t -distribution on $y - x\beta_1$ with scale $\frac{vb_1}{a_1}$ and $2a_1$ degrees of freedom.

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

$$(4.13) \quad f(y | Y, X, x) = ??$$

APPENDIX A. "ELEMENTARY" PROOF OF GAUSSIAN CONVOLUTION

Proposition A.1. *Let $\beta \in \mathbb{R}^k$ be a gaussian vector with distribution $\mathcal{N}(\mu, \Sigma)$ and let A be any $l \times k$ matrix. Then $A\beta \in \mathbb{R}^l$ is a gaussian vector with distribution $\mathcal{N}(A\mu, A\Sigma A^T)$*

Proof. We only consider the case $l = 1$ with $A = a = (a_1, a_2, \dots, a_k)$ is and $A\beta$ is a one dimensional random variable. Without loss of generality we may assume $a_1 \neq 0$. Introduce notations:

$$(A.1) \quad \beta = \left(\beta_1, \beta_{-1}^T\right)^T, \quad \beta_1 \in \mathbb{R}, \quad \beta_{-1} \in \mathbb{R}^{k-1}$$

$$(A.2) \quad \tilde{\beta} = \left(\tilde{\beta}_1, \tilde{\beta}_{-1}^T\right)^T = \beta - \mu$$

$$(A.3) \quad \hat{y} = a\beta$$

$$(A.4) \quad \tilde{y} = \hat{y} - a\mu$$

$$(A.5) \quad \Lambda = \Sigma^{-1}$$

$$(A.6) \quad \Lambda = \begin{pmatrix} \lambda_1 & \lambda_1 \\ \lambda_1^T & \Lambda_{-1} \end{pmatrix}$$

We note that the volume element $d\beta = d\beta_1 d\beta_{-1} = d\left(\frac{\hat{y}-}{a_1}\right) d\beta_{-1}$

□