

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 parameter 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 non-informative prior). *If the prior on parameter β is non-informative, i.e. $\mu_0 = 0$ and $\Sigma_0^{-1} = 0$, then posterior distribution of model parameters and predictive distributions recover standard OLS formulas:*

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

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

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

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 , liner 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 :

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

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 $\exp(-(\beta - v)^T Q (\beta - v))$ is independent of $v \in \mathbb{R}^k$, we find that up to a multiplicative constant:

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

Consequently, $\tilde{y} = y - x\boldsymbol{\mu}_1$ 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 $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0^{-1} = 0$ in propositions 2.1 and 2.2 \square

3. INVERSE GAMMA PRIOR FOR VARIANCE SCALE PARAMETER

Proposition 3.1 (Posterior parameter distribution). *Under assumptions with unknown variance 1.2, the posterior distribution decomposes as $f(\boldsymbol{\beta}, \sigma^2 | X, Y) = f(\boldsymbol{\beta} | X, Y, \sigma^2) \cdot f(\sigma^2 | X, Y)$ where the conditional posterior distribution $f(\boldsymbol{\beta} | X, Y, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2 \boldsymbol{\Sigma}_1)$ is normal with parameters as in 2.1-2.3 and $f(\sigma^2 | X, Y) = \text{Inv-}\Gamma(a_1, b_1)$ with parameters:*

$$(3.1) \quad a_1 = a_0 + \frac{n}{2}$$

$$(3.2) \quad b_1 = b_0 + \frac{1}{2} (Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1)$$

In particular, for $\boldsymbol{\Lambda}_0 = 0$ the increment in the update of b_0 is the residual sum of squares for OLS regression:

$$b_1 = b_0 + \frac{1}{2} (Y^T Y - Y^T X (X^T X)^{-1} X^T Y)$$

Proposition 3.2 (Posterior predictive distribution). *For an observation (x, y) and its expectation $(x, \hat{y} \stackrel{\text{def}}{=} x \cdot \boldsymbol{\beta})$, posterior conditional distributions of y and \hat{y} are location-scale t -distributions with parameters given below:*

$$(3.3) \quad f(y | x, X, Y) = \text{lst} \left(x\boldsymbol{\mu}_1, \frac{b_1}{a_1} (1 + x\boldsymbol{\Sigma}_1 x^T), 2a_1 \right)$$

$$(3.4) \quad f(\hat{y} | x, X, Y) = \text{lst} \left(x\boldsymbol{\mu}_1, \frac{b_1}{a_1} x\boldsymbol{\Sigma}_1 x^T, 2a_1 \right)$$

Proposition 3.3 (Bayesian regression and non-informative prior). *Assume that prior on parameters has form $f(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$. Then posterior distribution of model parameters and predictive distributions recover standard OLS formulas:*

$$(3.5) \quad \widehat{\sigma^2} = \frac{1}{n-k} (Y^T Y - Y^T X (X^T X)^{-1} X^T Y)$$

$$(3.6) \quad \boldsymbol{\beta} \sim t_{n-k} \left((X^T X)^{-1} X^T Y, \widehat{\sigma^2} (X^T X)^{-1} \right)$$

$$(3.7) \quad f(y | x, X, Y) = t_{n-k} \left(x (X^T X)^{-1} X^T Y, \widehat{\sigma^2} (1 + x (X^T X)^{-1} x^T) \right)$$

$$(3.8) \quad f(\hat{y} | x, X, Y) = t_{n-k} \left(x (X^T X)^{-1} X^T Y, \widehat{\sigma^2} x (X^T X)^{-1} x^T \right)$$

More specifically, the posterior parameter distribution can be decomposed as in Proposition 3.1 with:

$$(3.9) \quad f(\boldsymbol{\beta} \mid X, Y, \sigma^2) = \mathcal{N}\left((X^T X)^{-1} X^T Y, \sigma^2 (X^T X)^{-1}\right)$$

$$(3.10) \quad f(\sigma^2 \mid X, Y) = \text{Inv-}\Gamma\left(\frac{n-k}{2}, \frac{1}{2} \left(Y^T Y - Y^T X (X^T X)^{-1} X^T Y\right)\right)$$

Lemma 3.4. Suppose the distribution of $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+$ is $f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta} \mid \sigma^2) \cdot f(\sigma^2)$ with $f(\boldsymbol{\beta} \mid \sigma^2)$ normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f(\sigma^2) = \text{Inv-}\Gamma(a, b)$. Then the marginal distribution of $\boldsymbol{\beta}$ is multivariate t -distribution with $2a$ degrees of freedom: $f(\boldsymbol{\beta}) = t_{2a}(\boldsymbol{\mu}, \frac{b}{a} \boldsymbol{\Sigma})$.

Proof of Proposition 3.1. Using $f(\boldsymbol{\beta}, \sigma^2 \mid X, Y) \propto f(X, Y \mid \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta} \mid \sigma^2) f(\sigma^2)$ and substituting assumptions for model distributions one gets:

$$(3.11) \quad \begin{aligned} f(\boldsymbol{\beta}, \sigma^2 \mid X, Y) &\propto \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\boldsymbol{\beta})^T (Y - X\boldsymbol{\beta})\right) \cdot \\ &\cdot \frac{(\det \boldsymbol{\Lambda}_0)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right) \frac{b_0^{a_0}}{\Gamma(a_0)} (\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right) \end{aligned}$$

Following the proof of Proposition 2.1, completing the squares under the exponential gives:

$$\begin{aligned} &-\frac{1}{2\sigma^2} \left(\boldsymbol{\beta}^T \boldsymbol{\Lambda}_1 \boldsymbol{\beta} - \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{\beta} - \boldsymbol{\beta}^T \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_1 + Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 \right) = \\ &= -\frac{1}{2\sigma^2} \left((\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_1 (\boldsymbol{\beta} - \boldsymbol{\mu}_1) + Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_1 \right) \end{aligned}$$

It follows that up to a constant independent of $\boldsymbol{\beta}$ and σ^2 , the posterior $f(\boldsymbol{\beta}, \sigma^2 \mid X, Y)$ can be written as:

$$\frac{1}{(2\pi\sigma^2)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_1 (\boldsymbol{\beta} - \boldsymbol{\mu}_1)\right) (\sigma^2)^{-a_0-\frac{n}{2}-1} \exp\left(-\frac{1}{\sigma^2} \left(b_0 + \frac{1}{2} (Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_1)\right)\right)$$

□

Proof of Proposition 3.2. We only give the proof for predictive distribution of y , the result for \hat{y} is shown analogously. By definition,

$$(3.12) \quad f(y \mid x, X, Y) = \int_{\sigma^2 \in \mathbb{R}_+} \left[\int_{\boldsymbol{\beta} \in \mathbb{R}^k} f(y \mid x, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta} \mid X, Y, \sigma^2) d\boldsymbol{\beta} \right] f(\sigma^2 \mid X, Y) d\sigma^2$$

Following the proof and reusing notations of Proposition 2.2, the inner integral has the form:

$$\begin{aligned} &\int_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{(\det \boldsymbol{\Lambda}_1)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1+k}{2}}} \exp\left(-\frac{1}{2\sigma^2} ((y - x\boldsymbol{\beta})^2 + (\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_1 (\boldsymbol{\beta} - \boldsymbol{\mu}_1))\right) d\boldsymbol{\beta} = \\ &= \int_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{(\det \boldsymbol{\Lambda}_1)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1+k}{2}}} \exp\left(-\frac{\tilde{y}^2}{2\sigma^2} (1 + x\boldsymbol{\Sigma}_1 x^T)^{-1}\right) \exp\left(-\frac{1}{2\sigma^2} ((\boldsymbol{\beta} - \mathbf{v})^T Q (\boldsymbol{\beta} - \mathbf{v}))\right) d\boldsymbol{\beta} = \\ &= \text{const} \cdot \frac{1}{(\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{\tilde{y}^2}{2\sigma^2} (1 + x\boldsymbol{\Sigma}_1 x^T)^{-1}\right) \end{aligned}$$

where $\mathbf{v} \in \mathbb{R}^k$ is a vector and Q is a positive-definite quadratic form, both independent of β and σ^2 . In the second equality we used that the k -th power of σ^2 cancels out after integrating the second exponential (k -dimensional volume scale).

Denote $s^2 = (1 + x\mathbf{\Sigma}_1 x^T)$. Substituting the above expression in 3.12 and using the explicit form of $\text{Inv-}\Gamma(a_1, b_1)$ distribution one gets:

$$\begin{aligned} f(y | x, X, Y) &\propto \int_{\sigma^2 \in \mathbb{R}_+} \frac{1}{(\sigma^2)^{a_1+1+\frac{1}{2}}} \exp\left(-\frac{1}{\sigma^2} \left(b_1 + \frac{\tilde{y}^2}{2s^2}\right)\right) d\sigma^2 = \\ &= \Gamma\left(a_1 + \frac{1}{2}\right) \left(b_1 + \frac{\tilde{y}^2}{2s^2}\right)^{-a_1-\frac{1}{2}} = \frac{\Gamma\left(a_1 + \frac{1}{2}\right)}{b_1^{\frac{2a_1+1}{2}}} \left(1 + \frac{a_1 \tilde{y}^2}{2a_1 b_1 s^2}\right)^{-\frac{2a_1+1}{2}} \end{aligned}$$

Where the first step follows from $\int_{\mathbb{R}_+} r^{-A-1} \exp\left(-\frac{B}{r}\right) dr = \frac{\Gamma(A)}{B^A}$ for any $A > 0, B > 0$. Hence, the variable $\frac{a_1 \tilde{y}^2}{b_1 s^2}$ follows t -distribution with $2a_1$ degrees of freedom which readily implies the proposition. \square

Proof of Lemma 3.4. The proof is similar to the last part of proof of Proposition 3.2. Denote for convenience $\mathbf{\Lambda} \stackrel{\text{def}}{=} \mathbf{\Sigma}^{-1}$.

$$\begin{aligned} f(\beta) &= \int_{\sigma^2 \in \mathbb{R}_+} f(\beta, \sigma^2) d\sigma^2 = \\ &= \int_{\sigma^2 \in \mathbb{R}_+} \frac{(\det \mathbf{\Lambda})^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\sigma^2} (\beta - \mu)^T \mathbf{\Lambda} (\beta - \mu)\right) \cdot \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right) d\sigma^2 = \\ &= \text{const} \cdot \int_{\sigma^2 \in \mathbb{R}_+} (\sigma^2)^{-a-\frac{k}{2}-1} \exp\left(-\frac{1}{\sigma^2} \left(b + \frac{1}{2} (\beta - \mu)^T \mathbf{\Lambda} (\beta - \mu)\right)\right) d\sigma^2 = \\ &= \text{const} \cdot \frac{\Gamma\left(a + \frac{k}{2}\right)}{(b + (\beta - \mu)^T \mathbf{\Lambda} (\beta - \mu))^{a+\frac{k}{2}}} = \frac{\text{const} \cdot b^{-a-\frac{k}{2}} \Gamma\left(a + \frac{k}{2}\right)}{\left(1 + \frac{1}{2a} (\beta - \mu)^T \left(\frac{a}{b} \mathbf{\Lambda}\right) (\beta - \mu)\right)^{\frac{2a+k}{2}}} \end{aligned}$$

It remains to recognize the expression above as the standard representation of $t_{2a}(\mu, \frac{b}{a} \mathbf{\Sigma})$. \square

Proof of Proposition 3.3. Use $f(\sigma^2) = \frac{1}{\sigma^2}$ in 3.11:

$$\begin{aligned} f(\beta, \sigma^2 | X, Y) &\propto \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)\right) = \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left((\beta - (X^T X)^{-1} X^T Y)^T X^T X (\beta - (X^T X)^{-1} X^T Y)\right)\right) \cdot \\ &\quad \cdot \frac{1}{(2\pi\sigma^2)^{\frac{n-k}{2}-1}} \exp\left(-\frac{1}{2\sigma^2} (Y^T Y - Y^T X (X^T X)^{-1} X^T Y)\right) \end{aligned}$$

It follows that $f(\beta, \sigma^2 | X, Y)$ has form as in Lemma 3.4 with $a = \frac{n-k}{2}$ and $b = \frac{1}{2} (Y^T Y - Y^T X (X^T X)^{-1} X^T Y)$. Note that the non-informative prior is uniquely determined by the requirement that the corresponding differential form $\frac{1}{\sigma^2} d\beta d\sigma^2$ is preserved by the action of group $\mathbb{R}^k \rtimes \mathbb{R}_+$. \square

APPENDIX A. “ELEMENTARY” PROOF OF GAUSSIAN CONVOLUTION

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

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

$$(A.1) \quad \hat{y} = a\boldsymbol{\beta}$$

$$(A.2) \quad \tilde{y} = \hat{y} - a\boldsymbol{\mu}$$

$$(A.3) \quad a = (a_1, \mathbf{a}_{-1}), \quad a_1 \in \mathbb{R}^k, \quad \mathbf{a}_{-1} \in \mathbb{R}^{k-1}$$

$$(A.4) \quad \boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1}^T)^T, \quad \beta_1 \in \mathbb{R}^k, \quad \boldsymbol{\beta}_{-1} \in \mathbb{R}^{k-1}$$

$$(A.5) \quad \tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\boldsymbol{\beta}}_{-1}^T)^T = \boldsymbol{\beta} - \boldsymbol{\mu}$$

$$(A.6) \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$$

$$(A.7) \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_1^T & \boldsymbol{\Lambda}_{-1} \end{pmatrix}$$

We note that the volume element $d\boldsymbol{\beta} = d\beta_1 d\boldsymbol{\beta}_{-1} = d\left(\frac{\hat{y} - \mathbf{a}_{-1}\boldsymbol{\beta}_{-1}}{a_1}\right) d\boldsymbol{\beta}_{-1} = \frac{1}{a_1} d\hat{y} d\boldsymbol{\beta}_{-1}$. Thus, to get the density of \hat{y} it suffices to integrate out $d\boldsymbol{\beta}_{-1}$. It's easy to re-center variables for convenience using:

$$\begin{aligned} \left(\frac{1}{a_1} (\hat{y} - \mathbf{a}_{-1}\boldsymbol{\beta}_{-1}) - \mu_1, \boldsymbol{\beta}_{-1}^T - \boldsymbol{\mu}_{-1}^T\right) &= \left(\frac{1}{a_1} (\hat{y} - a_1\mu_1 - \mathbf{a}_{-1}(\boldsymbol{\beta}_{-1} - \boldsymbol{\mu}_{-1}) - \mathbf{a}_{-1}\boldsymbol{\mu}_{-1}), \boldsymbol{\beta}_{-1}^T - \boldsymbol{\mu}_{-1}^T\right) = \\ &= \left(\frac{1}{a_1} (\tilde{y} - \mathbf{a}_{-1}\tilde{\boldsymbol{\beta}}_{-1}) - \mu_1, \tilde{\boldsymbol{\beta}}_{-1}^T\right) \end{aligned}$$

We will use that the integral of the exponent of a quadratic form in a shift of $\tilde{\boldsymbol{\beta}}$ gives a multiplicative constant that does not depend on \tilde{y} (same idea as in the proof of Proposition 2.2, see 2.4):

$$\begin{aligned} \left(\frac{1}{a_1} (\tilde{y} - \mathbf{a}_{-1}\tilde{\boldsymbol{\beta}}_{-1})\right)_{\tilde{\boldsymbol{\beta}}_{-1}}^T \boldsymbol{\Lambda} \left(\frac{1}{a_1} (\tilde{y} - \mathbf{a}_{-1}\tilde{\boldsymbol{\beta}}_{-1})\right)_{\tilde{\boldsymbol{\beta}}_{-1}} &= \\ &= \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}}^T \boldsymbol{\Lambda} \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}} + \left(\frac{\tilde{y}}{\mathbf{0}}\right)^T \boldsymbol{\Lambda} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right) \tilde{\boldsymbol{\beta}}_{-1} + \tilde{\boldsymbol{\beta}}_{-1}^T \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)^T \boldsymbol{\Lambda} \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}} + \tilde{\boldsymbol{\beta}}_{-1}^T \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)^T \boldsymbol{\Lambda} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right) \tilde{\boldsymbol{\beta}}_{-1} \end{aligned}$$

Completing squares in $\tilde{\boldsymbol{\beta}}_{-1}$ and integrating it out, we're left with an exponential of the following expression in \tilde{y} :

$$\begin{aligned} \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}}^T \boldsymbol{\Lambda} \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}} - \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}}^T \boldsymbol{\Lambda} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right) \left(\left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)^T \boldsymbol{\Lambda} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)\right)^{-1} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)^T \boldsymbol{\Lambda} \left(\frac{\tilde{y}}{a_1}\right)_{\mathbf{0}} &= \\ (A.8) \quad &= \frac{\tilde{y}^2}{a_1^2} \lambda_{11} - \left(-\frac{\tilde{y}}{a_1^2} \lambda_{11} \mathbf{a}_{-1} + \frac{\tilde{y}}{a_1} \boldsymbol{\lambda}_1\right) \left(\left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)^T \boldsymbol{\Lambda} \left(-\frac{\mathbf{a}_{-1}}{\mathbf{I}}\right)\right)^{-1} \left(-\frac{\tilde{y}}{a_1^2} \lambda_{11} \mathbf{a}_{-1}^T + \frac{\tilde{y}}{a_1} \boldsymbol{\lambda}_1^T\right) = \\ &= \frac{\tilde{y}^2}{a_1^2} \lambda_{11} - \left(-\frac{\tilde{y}}{a_1^2} \lambda_{11} \mathbf{a}_{-1} + \frac{\tilde{y}}{a_1} \boldsymbol{\lambda}_1\right) \left(\frac{1}{a_1^2} \mathbf{a}_{-1}^T \mathbf{a}_{-1} - \frac{1}{a_1} \boldsymbol{\lambda}_{-1}^T \mathbf{a}_{-1} - \frac{1}{a_1} \mathbf{a}_{-1}^T \boldsymbol{\lambda}_{-1} + \boldsymbol{\Lambda}_{-1}\right)^{-1} \left(-\frac{\tilde{y}}{a_1^2} \lambda_{11} \mathbf{a}_{-1}^T + \frac{\tilde{y}}{a_1} \boldsymbol{\lambda}_1^T\right) \end{aligned}$$

The next step is to recognize the above expression as the inverse of a matrix using two versions of the block-matrix inverse:

Lemma A.2. *If matrices A and $D - CA^{-1}B$ are invertible then:*

$$(A.9) \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

If matrices D and $A - BD^{-1}C$ are invertible then:

$$(A.10) \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}$$

If matrices A and D are invertible then:

$$(A.11) \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & \mathbf{0} \\ \mathbf{0} & (D - CA^{-1}B)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -BD^{-1} \\ -CA^{-1} & \mathbf{I} \end{pmatrix}$$

Applying the above to $A = \frac{a_1^2}{\lambda_{11}}$, $B = \mathbf{a}_{-1} - \frac{a_1}{\lambda_{11}}\boldsymbol{\lambda}_1$, $C = -B^T$, $D = \boldsymbol{\Lambda}_{-1} - \frac{1}{\lambda_{11}}\boldsymbol{\lambda}_1^T\boldsymbol{\lambda}_1$ we see that the right hand side of A.8 can be rewritten further

$$(A.12) \quad \begin{aligned} & \tilde{y}^2 \left(\frac{a_1^2}{\lambda_{11}} + \left(-\mathbf{a}_{-1} + \frac{a_1}{\lambda_{11}}\boldsymbol{\lambda}_1 \right) \left(\boldsymbol{\Lambda}_{-1} - \frac{1}{\lambda_{11}}\boldsymbol{\lambda}_1^T\boldsymbol{\lambda}_1 \right)^{-1} \left(-\mathbf{a}_{-1}^T + \frac{a_1}{\lambda_{11}}\boldsymbol{\lambda}_1^T \right) \right)^{-1} = \\ & = \tilde{y}^2 + \left(\mathbf{a} \left(\begin{pmatrix} \frac{1}{\lambda_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \frac{1}{\lambda_{11}}\boldsymbol{\lambda}_1 \\ -\mathbf{I} \end{pmatrix} \left(\boldsymbol{\Lambda}_{-1} - \frac{1}{\lambda_{11}}\boldsymbol{\lambda}_1^T\boldsymbol{\lambda}_1 \right)^{-1} \begin{pmatrix} \frac{1}{\lambda_{11}}\boldsymbol{\lambda}_1^T & -\mathbf{I} \end{pmatrix} \right) \mathbf{a}^T \right)^{-1} \end{aligned}$$

It suffices to observe that the quadratic form in \mathbf{a} in the expression above equals $\boldsymbol{\Sigma}$. Indeed, this follows from the formula A.9 applied to $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$. \square