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

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

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

Date: April 8, 2024.

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 distributions of model parameters and predictive distributions recover standard OLS formulas:*

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

$$f(y \mid x, X, Y) = \mathcal{N}\left(x\boldsymbol{\mu}_1, \sigma^2 \left(1 + x\left(X^T X\right)^{-1} x^T\right)\right) \tag{2.7}$$

$$f(\widehat{y} \mid x, X, Y) = \mathcal{N}\left(x\boldsymbol{\mu}_1, \sigma^2 x\left(X^T X\right)^{-1} x^T\right) \tag{2.8}$$

Proof of Proposition 2.1. 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:

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

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:*

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

□

Proof of Proposition 2.2. To find the posterior predictive distribution 2.4, we integrate out parameter $\boldsymbol{\beta}$:

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

Set $\widetilde{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \boldsymbol{\beta} - \boldsymbol{\mu}_1$ and $\widetilde{y} \stackrel{\text{def}}{=} y - x\boldsymbol{\mu}_1$. The expession under the exponential can be rewritten as:

$$\begin{aligned}
(y - x\beta)^2 + (\boldsymbol{\beta} - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_1 (\boldsymbol{\beta} - \boldsymbol{\mu}_1) &= \left(\widetilde{y} - x\widetilde{\boldsymbol{\beta}}\right)^2 + \widetilde{\boldsymbol{\beta}}^T \boldsymbol{\Lambda}_1 \widetilde{\boldsymbol{\beta}} = \\
= \widetilde{y}^2 - \widetilde{y}x\left(\boldsymbol{\Lambda}_1 + x^T x\right)^{-1} x^T \widetilde{y} &+ \left(\widetilde{\boldsymbol{\beta}} - \left(\boldsymbol{\Lambda}_1 + x^T x\right)^{-1} x^T \widetilde{y}\right)^T \left(\boldsymbol{\Lambda}_1 + x^T x\right)\left(\widetilde{\boldsymbol{\beta}} - \left(\boldsymbol{\Lambda}_1 + x^T x\right)^{-1} x^T \widetilde{y}\right)
\end{aligned}$$

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

$$f(y \mid x, X, Y) = \text{const} \cdot \exp\left(-\frac{\widetilde{y}^2}{2\sigma^2}\left(1 - x\left(\boldsymbol{\Lambda}_1 + x^T x\right)^{-1} x^T\right)\right) = \text{const} \cdot \exp\left(-\frac{\widetilde{y}^2}{2\sigma^2}\left(1 + x^T \boldsymbol{\Sigma}_1 x^T\right)^{-1}\right)$$

Consequently, $\widetilde{y} = y - x\boldsymbol{\mu}_1$ 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,*

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

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

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 □

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 □

3. Variance Scale Inverse Gamma Prior

4. Conjugate Priors For Observation Variance and Linear Weights

4.1. Setup.

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

$$f(\boldsymbol{\beta} \mid \sigma^2) = |2\pi\boldsymbol{\Sigma}_0|^{-1} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\right) \tag{4.2}$$

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

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

4.2. Summary of Results.

Posterior distribution of $\boldsymbol{\beta}$:

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

$$a_1 = a_0 + \frac{n}{2} \tag{4.5}$$

$$b_1 = b_0 + \frac{1}{2}\left(Y^T Y + \boldsymbol{\beta}_0\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0^T - \boldsymbol{\beta}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}_1^T\right) \tag{4.6}$$

$$f(\boldsymbol{\beta} \mid Y, X, \sigma^2) = \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Sigma}_1) \tag{4.7}$$

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^T X \tag{4.8}$$

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_1\left(X^T X\widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right) \tag{4.9}$$

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y \tag{4.10}$$

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

$$f(\widehat{y} \mid Y, X, x) \propto \left(1 + \frac{a_1\left(y - x\boldsymbol{\beta}_1\right)^2}{vb_1}\frac{1}{2a_1}\right)^{-\frac{2a_1+1}{2}} \tag{4.11}$$

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

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

$$f(y \mid Y, X, x) = ?? \tag{4.13}$$

Appendix A. "Elementary" Proof of Gaussian Convolution

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

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

$$\widehat{y} = a\boldsymbol{\beta} \tag{A.1}$$

$$\widetilde{y} = \widehat{y} - a\boldsymbol{\mu} \tag{A.2}$$

$$a = \left(a_1,\ \boldsymbol{a}_{-1}\right),\ a_1 \in \mathbb{R}^k,\ \boldsymbol{a}_{-1} \in \mathbb{R}^{k-1} \tag{A.3}$$

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

$$\widetilde{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{\beta}}_1,\ \widetilde{\boldsymbol{\beta}}_{-1}^T\right)^T = \boldsymbol{\beta} - \boldsymbol{\mu} \tag{A.5}$$

$$\mathbf{\Lambda} = \mathbf{\Sigma}^{-1} \tag{A.6}$$

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

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

$$\left(\frac{1}{a_1}\left(\widehat{y} - \boldsymbol{a}_{-1}\boldsymbol{\beta}_{-1}\right) - \boldsymbol{\mu}_1,\ \boldsymbol{\beta}_{-1}^T - \boldsymbol{\mu}_{-1}^T\right) = \left(\frac{1}{a_1}\left(\widehat{y} - a_1\boldsymbol{\mu}_1 - \boldsymbol{a}_{-1}\left(\boldsymbol{\beta}_{-1} - \boldsymbol{\mu}_{-1}\right) - \boldsymbol{a}_{-1}\boldsymbol{\mu}_{-1}\right),\ \boldsymbol{\beta}_{-1}^T - \boldsymbol{\mu}_{-1}^T\right) =$$
$$= \left(\frac{1}{a_1}\left(\widetilde{y} - \boldsymbol{a}_{-1}\widetilde{\boldsymbol{\beta}}_{-1}\right) - \boldsymbol{\mu}_1,\ \widetilde{\boldsymbol{\beta}}_{-1}^T\right)$$

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

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

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

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

The next step is to apply 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:*

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

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

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

If matrices A and D are invertible then:

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

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

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

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