BAYESIAN LINEAR REGRESSION

EFIM ABRIKOSOV

1. Framework

1.1. Notations.

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

(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 $\mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}\right)$
(5) For unknown error variance, the prior for $\left(\boldsymbol{\beta}, \sigma^2\right) \in \mathbb{R}^{k+1}$ has σ^2 following inverse gamma distribution with parameters (a_0, b_0), and conditional distribution for linear relation weights $f\left(\boldsymbol{\beta} \mid \sigma^2\right) = \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}\right)$

2. Gaussian Prior with Known Variance

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:*

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

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

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

Proposition 2.2 (Posterior predictive distribution with known variance)**.** *For an observation (x, y) and its expectation $(x, \widehat{y} \stackrel{\text{def}}{=} x \cdot \boldsymbol{\beta})$, 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: June 28, 2024.

Proposition 2.3 (Bayesian regression under known vairance and non-informative prior)**.** *If the prior on parameter $\boldsymbol{\beta}$ is non-informative, i.e. $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0^{-1} = 0$, then posterior distribution of model parameters and predictive distributions recover standard OLS formulas:*

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

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

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

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:

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

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 $\boldsymbol{\beta}$:

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

Set $\widetilde{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \boldsymbol{\beta} - \boldsymbol{\mu}_1$ and $\widetilde{y} \stackrel{\text{def}}{=} y - x\boldsymbol{\mu}_1$. The expression 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 $\exp(-(\boldsymbol{\beta} - \boldsymbol{v})^T Q (\boldsymbol{\beta} - \boldsymbol{v}))$ is independent of $\boldsymbol{v} \in \mathbb{R}^k$, we find that up to a multiplicative constant:

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

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. Inverse Gamma Prior for Variance Scale Parameter

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

$$a_1 = a_0 + \frac{n}{2} \tag{3.1}$$

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

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

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

Proposition 3.2 (Posterior predictive distribution). *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 location-scale t-distributions with parameters given below:*

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

$$f(\widehat{y} \mid x, X, Y) = lst\left(x\boldsymbol{\mu}_1, \frac{b_1}{a_1} x\boldsymbol{\Sigma}_1 x^T, 2a_1\right) \tag{3.4}$$

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

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

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

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

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

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

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

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

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

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:

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

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

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

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:

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

The second equality in 3.2 follows from simple algebraic manipulations:

$$\begin{aligned} & Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_1 = Y^T Y + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \left(X^T Y + \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0\right)^T \boldsymbol{\mu}_1 = Y^T\left(Y - X\boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) = \\ & \quad = \left(Y - X\boldsymbol{\mu}_1\right)^T\left(Y - X\boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_1^T X^T\left(Y - X\boldsymbol{\mu}_1\right) + \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^T \boldsymbol{\Lambda}_0\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_0\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) \end{aligned}$$

It remains to show that the sum of the second and fourth terms is zero:

$$\begin{aligned} & \boldsymbol{\mu}_1^T X^T\left(Y - X\boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_0\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) = \boldsymbol{\mu}_1^T X^T\left(Y - X\boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) - \boldsymbol{\mu}_1^T X^T X\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) = \\ & \quad = \boldsymbol{\mu}_1^T X^T Y + \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right) - \boldsymbol{\mu}_1 X^T X \boldsymbol{\mu}_0 = \boldsymbol{\mu}_1^T X^T Y + \boldsymbol{\mu}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\left(X^T Y + \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0\right) - \boldsymbol{\mu}_1 X^T X \boldsymbol{\mu}_0 = \\ & \quad = \boldsymbol{\mu}_1^T\left(\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_0 - X^T X\right)\boldsymbol{\mu}_0 = 0 \end{aligned}$$

□

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

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

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

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

$$=\int\limits_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{(\det\boldsymbol{\Lambda}_1)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1+k}{2}}} \exp\left(-\frac{\widetilde{y}^2}{2\sigma^2}\left(1+x\boldsymbol{\Sigma}_1x^T\right)^{-1}\right)\exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}-\boldsymbol{v})^TQ(\boldsymbol{\beta}-\boldsymbol{v})\right)\right)d\boldsymbol{\beta}=$$

$$=\text{const}\cdot\frac{1}{(\sigma^2)^{\frac{1}{2}}}\exp\left(-\frac{\widetilde{y}^2}{2\sigma^2}\left(1+x\boldsymbol{\Sigma}_1x^T\right)^{-1}\right)$$

where $\boldsymbol{v}\in\mathbb{R}^k$ is a vector and Q is a positive-definite quadratic form, both independent of $\boldsymbol{\beta}$ 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 $s^2=\left(1+x\boldsymbol{\Sigma}_1x^T\right)$. Substituting the above expression in 3.12 and using the explicit form of Inv-$\Gamma(a_1,b_1)$ distribution one gets:

$$f(y\mid x,X,Y)\propto\int\limits_{\sigma^2\in\mathbb{R}_+}\frac{1}{(\sigma^2)^{a_1+1+\frac{1}{2}}}\exp\left(-\frac{1}{\sigma^2}\left(b_1+\frac{\widetilde{y}^2}{2s^2}\right)\right)d\sigma^2=$$

$$=\Gamma\left(a_1+\frac{1}{2}\right)\left(b_1+\frac{\widetilde{y}^2}{2s^2}\right)^{-a_1-\frac{1}{2}}=\frac{\Gamma\left(a_1+\frac{1}{2}\right)}{b_1^{\frac{2a_1+1}{2}}}\left(1+\frac{a_1\widetilde{y}^2}{2a_1b_1s^2}\right)^{-\frac{2a_1+1}{2}}$$

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 $\frac{a_1\widetilde{y}^2}{b_1s^2}$ follows t-distribution with $2a_1$ degrees of freedom which readily implies the proposition. □

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

$$f(\beta)=\int_{\sigma^2\in\mathbb{R}_+} f\left(\beta,\sigma^2\right)d\sigma^2=$$

$$=\int_{\sigma^2\in\mathbb{R}_+}\frac{(\det\boldsymbol{\Lambda})^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{k}{2}}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\cdot\frac{b^a}{\Gamma(a)}\left(\sigma^2\right)^{-a-1}\exp\left(-\frac{b}{\sigma^2}\right)d\sigma^2=$$

$$=\text{const}\cdot\int_{\sigma^2\in\mathbb{R}_+}\left(\sigma^2\right)^{-a-\frac{k}{2}-1}\exp\left(-\frac{1}{\sigma^2}\left(b+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\right)d\sigma^2=$$

$$=\text{const}\cdot\frac{\Gamma\left(a+\frac{k}{2}\right)}{(b+(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\mu}))^{a+\frac{k}{2}}}=\frac{\text{const}\cdot b^{-a-\frac{k}{2}}\Gamma\left(a+\frac{k}{2}\right)}{\left(1+\frac{1}{2a}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\left(\frac{a}{b}\boldsymbol{\Lambda}\right)(\boldsymbol{\beta}-\boldsymbol{\mu})\right)^{\frac{2a+k}{2}}}$$

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

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

$$f\left(\boldsymbol{\beta}, \sigma^2 \mid X, Y\right) \propto \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}\left(Y - X\boldsymbol{\beta}\right)^T\left(Y - X\boldsymbol{\beta}\right)\right) =$$
$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\beta} - \left(X^TX\right)^{-1}X^TY\right)^T X^TX\left(\boldsymbol{\beta} - \left(X^TX\right)^{-1}X^TY\right)\right)\right)\cdot$$
$$\cdot\frac{1}{\left(2\pi\sigma^2\right)^{\frac{n-k}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}\left(Y^TY - Y^TX\left(X^TX\right)^{-1}X^TY\right)\right)$$

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

4. Exponential Moving Average via Bayesian Linear Regression

4.1. **Notations.**

(1) Observed data $(X_t, Y_t) \in \mathbb{R}^{n_t\times k} \times \mathbb{R}^{n_t},\ t \geq 0$
(2) Linear regression weights $\boldsymbol{\beta}_t \in \mathbb{R}^k,\ t \geq 0$
(3) Model parameter distribution mean $\boldsymbol{\mu}_t \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma}_t \in \mathbb{R}^{k\times k}$ for $t \geq 0$
(4) Observation error variance $\sigma_t^2,\ t \geq 0$

Weighting scheme ω for a sequence of observations labelled by $t \geq 0$ is defined in one of the equivalent ways:

(1) $w_t \in \mathbb{R}_{\geq 0},\ t \geq 0$
(2) $w_{t,t} \in \mathbb{R}_{\geq 0},\ t \geq 0$ and $\omega_t \in \mathbb{R}_+,\ t \geq 1$
(3) $w_{t,\tau} \in \mathbb{R}_{\geq 0},\ t \geq 0, \tau = 0, \dots t$ such that $w_{t_1,\tau_1}w_{t_2,\tau_2} = w_{t_1,\tau_2}w_{t_2,\tau_1}$ and $w_{t,\tau_1} = 0 \Leftrightarrow w_{t,\tau_2} = 0$

Given weighting scheme in one of the forms above, one can produce a sequence of weighted *averages* for any series of conforming objects $\overline{o_{w,t}} \stackrel{\text{def}}{=} \left(\sum_{\tau=0}^{t} w_\tau\right)^{-1}\left(\sum_{\tau=0}^{t} w_\tau o_\tau\right)$ defined for any index $t \geq t_0$ corresponding to a nonzero cumulative weight. If all weights are non-zero then this data can be produced using relative weights $0 < r_{w,t} < 1$, $t \geq 1$: $\overline{o_{w,t}} = \left(1 - r_{w,t}\right)o_t + r_{w,t}\overline{o_{w,t-1}}$.

Example 4.1 (Exponential Weighted Moving Average)**.** Fix exponential decay rate $0 < \omega < 1$.

(1) $w_0 = 1,\ w_t = \dfrac{1-\omega}{\omega^t},\ t \geq 1$
(2) $w_{0,0} = 1,\ w_{t,t} = 1 - \omega,\ t \geq 1$ and $\omega_t = \omega,\ t \geq 1$
(3) $w_{t,0} = \omega^t,\ w_{t,\tau} = (1-\omega)\,\omega^{t-\tau},\ \tau = 1, \dots t$
(4) $r_{w,t} = \omega,\ t \geq 1$

Example 4.2 (Adaptive Exponential Weighted Moving Average)**.** Fix exponential decay rate $0 < \omega < 1$.

(1) $w_t = \omega^{-t},\ t \geq 0$
(2) $w_{t,t} = 1,\ t \geq 0$ and $\omega_t = \omega,\ t \geq 1$
(3) $w_{t,\tau} = \omega^{t-\tau},\ t \geq 0,\ \tau = 0, \dots t$
(4) $r_{w,t} = \dfrac{\omega - \omega^{t+1}}{1 - \omega^{t+1}},\ t \geq 1$

Proposition 4.3. *Consider a weighting scheme in the format* $w_{t,t} \in \mathbb{R}_{\geq 0},\ t \geq 0$ *and* $\omega_t \in \mathbb{R}_+,\ t \geq 1$ *with* $w_{t,t} = 1,\ \forall t > 0$. *Assume that before observing* $(X_t, Y_t),\ t > 0$, *the prior "decays" based on the rule:*

$$\mathbf{\Sigma}_{t-1}^{-1} \to \omega \mathbf{\Sigma}_{t-1}^{-1} \tag{4.1}$$

$$a_{t-1} \to \omega a_{t-1} \tag{4.2}$$

$$b_{t-1} \to \omega b_{t-1} \tag{4.3}$$

and assume that the initial prior is given by parameters... TODO

Caveat...

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 $A = a = (a_1, a_2, \ldots, a_k)$ 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 $d\boldsymbol{\beta} = d\beta_1 \boldsymbol{d\beta}_{-1} = d\left(\frac{\widehat{y} - \boldsymbol{a}_{-1}\boldsymbol{\beta}_{-1}}{a_1}\right) d\boldsymbol{\beta}_{-1} = \frac{1}{a_1} d\widehat{y}\boldsymbol{d\beta}{-1}$. 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:

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

We will use that the integral of the exponent of a quadratic 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):

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

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

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

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:*

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

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

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

If matrices A and D are invertible then:

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

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 rewritten further

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

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 $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$. $\square$