WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Student-ID:	11913169
Degree Program:	Master of Science in Economics (Science Track)
Examiner:	Peter Knaus, PhD
Submission date:	TBA

Triple-Gamma-Regularization *or* **On the Duality of Frequentist Point Estimates and Bayesian Shrinkage Priors**

An Extension based on the Triple-Gamma-Prior

by

Lucas Unterweger GitHub

(Student-ID: 11913169)

Abstract

Lorem ipsum...

Contents

1 Introduction

Willam of Ockham, born in Ockham, Surrey, probably lived between 1287 and 1348 and is nowadays recognized as a pre-eminent philosopher of the middle ages. Although his name itself is no common knowledge, a principle carrying his name is: *Ockham's Razor*. Interestingly, the main formulation of the principle (*Entia non sunt multiplicanda praeter necessitatem* [plurality should not be posited without necessity]) can not be traced back to Ockham directly, but variations of it can be found in Ockham's writings. Since then nonetheless, the principle has long been used by statisticians and other researchers as a a scientific credo to capture the notion that ”the simpler of two explanations is to be preferred” (Lazar, 2010).

2 Literature Review

2.1 Bayesian View

2.2 Frequentist View

2.3 Machine Learning Literature combating Model Complexity

3 Theoretical Section

3.1 Model Complexity

3.2 Under- and Overfitting

3.2.1 Regularization

3.2.2 Shrinkage Priors

3.3 Duality of the Ridge Regression

4 Model Setup and Derivation

Coming to the theoretical framework of the *triple-gamma-regularization*, let's assume we have a response variable y and p predictors along with n data points. More formally, let $y = [y_1 \quad y_2 \cdots y_n]^T$ and $x_i = [x_{i1} \quad x_{i2} \cdots x_{in}]^T$ with $\forall i \in \{1, ..., n\} : y_i, x_i \in \mathbb{R}$. Here, x_i is the i-th predictor, thus resulting in the design matrix $X = [x_1 \quad x_2 \cdots x_p]$.

Starting from the Bayesian framework, the standard linear regression model is given by

$$y_i = x_i^T \cdot \beta + \varepsilon_i \quad i \in \{1, ..., n\}$$

with the assumed distribution of $\varepsilon_i \sim N(0, \sigma^2)$. Thus it follows that $y \sim N_n(X\beta, \sigma^2 I)$. The posterior distribution of the parameter vector β, according to Bayes' Rule, is then proportional to the product of the likelihood of the data and the prior distribution, which can be seen in equation 1.

$$p(\beta|y, X, \sigma^2) \propto \mathcal{L}(y|\beta, \sigma^2, X) \times p(\beta) \tag{1}$$

As stated above, each data point y_i is assumed to be identically and independently drawn from a normal distribution with mean $X\beta$ and variance σ_i^2, thus:

$$\begin{aligned}\mathcal{L}(\mathbf{y}|\beta, \sigma^2, X) &= \prod_i^n p(y_i|\beta, \sigma^2, X_i)\\ &= \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)\\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right) \end{aligned} \tag{2}$$

The log of the likelihood function is then given by

$$\begin{aligned}\log \mathcal{L}(\mathbf{y}|\beta, \sigma^2, \mathbf{X}) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right)\right)\\ &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) + \log\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right)\\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\\ &\propto -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) = -\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2\end{aligned}$$

The marginal prior distribution for the parameter vector β stems from the Triple-Gamma-Prior constructed in Cadonna et al. (2020) given in formula Theorem 1 (a) and is given by

$$\begin{aligned} p(\sqrt{\beta_j}|\phi^\xi, a^\xi, c^\xi) &= \frac{\Gamma(c^\xi + \frac{1}{2})}{\sqrt{2\pi\phi^\xi} \cdot B(a^\xi, c^\xi)} \cdot U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right) \\ &\propto U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right) \end{aligned}$$

As this prior is specified for the parameter $\sqrt{\beta_j}$, we transform the prior by squaring the parameter to gain

$$p(\beta_j|\phi^\xi, a^\xi, c^\xi) \propto U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j^2}{2\phi^\xi}\right)$$

Now, assuming that the parameters are independent a priori, the prior distribution is given by

$$\begin{aligned} p(\beta) &= \prod_j^p p(\beta_j) \\ &= \prod_j^p p(\beta_j|\phi^\xi, a^\xi, c^\xi) \\ &\propto \prod_j^p U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j^2}{2\phi^\xi}\right) \\ &= \prod_j^p \frac{1}{\Gamma(c^\xi + \frac{1}{2})} \int_0^\infty e^{-(\frac{\beta_j^2}{2\phi^\xi})t} t^{c^\xi + \frac{1}{2} - 1} (1+t)^{\frac{3}{2} - a^\xi - c^\xi + \frac{1}{2} - 1} dt \\ &\propto \prod_j^p \int_0^\infty \exp\left(-\frac{\beta_j^2}{2\phi^\xi} t\right) t^{c^\xi - \frac{1}{2}} (1+t)^{1 - a^\xi - c^\xi} dt \end{aligned} \tag{3}$$

Here, in line 1 the assumption of independence between the parameters has been used to describe the distribution of the parameter vector as the product of its individual parameter distributions. In line 2, the marginal prior from Cadonna et al. (2020) has been used as the prior distribution for each individual parameter β_j. In line 3, scaling parameters have been removed by using the proportionality assumption. The last two lines of the derivation insert the definition of the confluent hyper-geometric function of the second kind and again apply proportionality.

Taking the log of the prior distribution and using the properties of the logarithmic function yields the general result

$$\log(p(\beta)) = \log(\prod_j^p p(\beta_j|\phi^\xi, a^\xi, c^\xi)) = \sum_j^p \log(p(\beta_j|\phi^\xi, a^\xi, c^\xi))$$

A common approach to estimation in regularization settings is the *maximum a posteriori probability (MAP)* estimator **(Missing Citation)**,which is defined as

$$\hat{\beta}_{MAP}(x) = \underset{\beta \in \mathbb{R}^p}{\arg\max} \{f(x|\beta)g(\beta)\}$$

where $f(x|\beta)$ describes the the probability density function of a variable x, which is parametrized by the parameter vector β. The second function $g(\beta)$ incorporates our prior information about the parameter vector β into the optimization problem.

Returning to our specific problem at hand, the posterior distribution of our parameter vector β can be retrieved by applying Bayes' theorem and the previously gained results in equations 3 and 2. Thus, the posterior distribution of the parameter vector β is proportional to

$$\begin{aligned} p(\beta|y, X, \sigma^2) &\propto p(y|X, \beta, \sigma) \times p(\beta) \\ &\propto \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)} \times \prod_j^p U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right) \end{aligned} \tag{4}$$

Making use of the monotonicity of the logarithmic function and seeing that it is easier to optimize the log-posterior, Taking the log of the posterior probability distribution, we take the log of result 4.

$$\begin{aligned} \log(\beta|X, y, \sigma^2) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) - \frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2 + \log\left(\prod_j^p U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right)\right) \\ &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) - \frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2 + \sum_j^p \log\left(U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right)\right) \\ &\propto -\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2 + \sum_j^p \log\left(U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right)\right) \end{aligned} \tag{5}$$

By minimizing the negative log-posterior, we can retrieve the *maximum a posteriori probability (MAP)* estimator using **Triple-Gamma-Regularization**:

$$\hat{\beta}_{MAP} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\left(\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2 - \sum_j^p \log\left(U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j}{2\phi^\xi}\right)\right)\right) \tag{6}$$

BREAK

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1} dt \tag{7}$$

5 Simulation Section

Use the earlier derivation, code up the functions and simulate data to check behaviour for different datasets. Compare to base OLS, Ridge and Lasso Regression?

5.1 Computational Performance

5.2 Implementation as R Package?

6 Small Applied Section

Here I am planning to apply the TGP using Peter's Bayesian Package and the self coded frequentist code on a small dataset.

7 Conclusion

8 Possible Extensions and Criticism

9 References

References

Cadonna, A., Frühwirth-Schnatter, S., & Knaus, P. (2020). Triple the gamma—a unifying shrinkage prior for variance and variable selection in sparse state space and tvp models. *Econometrics*, *8*(2), 20.

Lazar, N. (2010). Ockham's razor. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 243–246.