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 Theoretical Section

2.1 Model Complexity

Overfitting is a problem. Especially when dimensionality issues arise. For example, 50 observations and 20 features. Model will perform well/almost perfect on training data, but will do terrible on training data.

2.2 Under- and Overfitting

General problem: Fit vs. Complexity. What to choose and where is the optimal balance?

Solutions (find source, seen in LMU course): (1) More data (Not always feasible) (2) Better data (Same issue) (3) Reduce Model Complexity (Regularization) (4) Less aggressive optimization (Maybe short review; Idea: Stop optimization when generalization error degrades)

2.2.1 Regularization

Focusing on option three to battle model complexity, regularization methods have been studied thoroughly since the 1990s. However, the emergence of data science - and especially machine learning - as a standalone field of study has led to a broader meaning of the term *regularization*. This phenomenon has been discussed by Kukačka et al. (2017), where the authors establish a taxonomy to distinguish between multiple different definitions. In the traditional sense, as can be seen in Hastie et al. (2009, pp. 167–170), *regularization* refers to a general class of problems of the form

$$\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right\}$$

where $L(.)$ refers to a loss function defined as some function of the true values and the predicted values and $J(f)$ is penalty based on the chosen functional from a space of functions $\mathcal{H}$. In the context of penalized linear regression, this is equivalent to finding the set of risk minimizing coefficients $\hat{\beta}$ from the set of all possible combinations of coefficients $\boldsymbol{\beta}$. **(Missing Citation)**Thus, resulting in the general class of regularization problems of the form:

$$\min_{\beta \in \boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} L(y_i, f_\beta(x_i)) + \lambda J(\beta) \right\}$$

where $f_\beta(x_i)$ is a linear function of the inputs x_i parametrized by the coefficients β. Hence, in this setting, regularization deals with penalizing the risk function based on the value of the chosen set of coefficients.
However, this only describes a subset of *regularization* methods as stated by Kukačka et al. (2017). The authors use a more general definition of regularization:

Defintion 1. Regularization is any supplementary technique that aims at making the model generalize better, i.e. produce better results on the test set.

Building on that, they split up the majority of *regularization* methods into (1) methods applied to the data set like transformations or modifications of the inputs, (2) methods altering the selected model family , (3) methods applied to the error/loss function $L(y_i, f_\beta(x_i))$, (4) methods applied to regularization/penalty term as described above and (5) alterations of the optimization procedure itself.

Vielleicht noch genauer auf die Taxonomy eingehen?

Unsurprisingly, this thesis is concerned with the fourth group of regularization methods, which add a penalty/regularizer term $J(\beta)$ into the risk function, but before advancing to literature that deals with this kind of problem, it is necessary to establish a terminology which will be used throughout this thesis. **TODO**

Let $\mathcal{D}$ be a training data set with $n \in \mathbb{N}$ observations, where every consists of a target variable $y_i \in \mathbb{R}$ along with a number of corresponding inputs $x_i \in \mathbb{R}$. Given a linear function $f_\beta(x_i)$ of the inputs parametrized by coefficients $\beta \in \boldsymbol{\beta}$, $L(y_i f_\beta(x_i))$ is the **Loss** function measuring the discrepancy between the actual target y_i and the output of the linear function $f_\beta(x_i)$. According to **Vapnik1991<empty citation>**, the **Empirical Risk Functional** is then

$$R_{emp}(\beta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_\beta(x_i)).$$

Regularization in this thesis' context refers then to adding some penalty function $J(\beta)$ dependent on the set of parameters β, multiplied by some weighting parameter λ, to the

empirical risk functional $R_{emp}(\beta)$. Thus, $R_{reg} = R_{emp}(\beta) + \lambda \cdot J(\beta)$. This results in the overall optimization problem

$$\begin{aligned}
&\arg\min_{\beta \in \boldsymbol{\beta}} \{R_{reg}(\beta)\} \\
&= \arg\min_{\beta \in \boldsymbol{\beta}} \{R_{emp}(\beta) + \lambda \cdot J(\beta)\} \\
&= \arg\min_{\beta \in \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} L(y_i, f_\beta(x_i)) + \lambda \cdot J(\beta) \right\}
\end{aligned}$$

It is important to note here that as $J(\beta)$ is only a function of the coefficients β, it only affects the generalization error of the model, not the training error given by the empirical risk functional $R_{emp}(\beta)$.

Ad Solution (3) because it is relatively easy to implement: Simple approach =¿ Start with simplest model and iteratively add one feature OR iteratively get rid of feature by feature. (Problem: Very arbitrary and hard to reasonably do) Thus, adjust risk function minimization. Important: Regularization adjusts generalization error, not training error!
Short break: Talk about Taxonomy by Kukacka et al 2017 on the differing use of the term "regularization"
What is needed:

$$R_{reg}(\beta) = R_{emp}(\beta) + \lambda \cdot J(\beta)$$

where $R_{emp}(\beta) = \sum_i^n L(y_i, f(x_i, \beta))$. *Note: Regularization term does not depend on data, just on parametrization.* λ *controls the strength of regularization. Thus,* $\lambda = 0$ *means simple MSE optimization and* $\lambda \to \infty$ *chooses simplest model. As* λ *is set manually, this also bears some problems. However, typical solution is cross-validation.*
Literature *(Kukačka et al., 2017)*

3 Literature Review

The concept of a penalized regression has been around for quite some time and been studied widely in various fields of scientific research.

3.0.1 Short Deviation to Bayesian Shrinkage Priors

3.1 Duality of the Ridge Regression

3.2 Triple-Gamma-Prior by Cadonna et al. (2020)

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

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

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

by

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

Here, $U(a, b, z)$ refers the confluent hyper-geometric function of the second kind which was introduced by Tricomi (1947). 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 integral representation of the confluent hyper-geometric function of the second kind, $U(a, b, z)$, which is valid in the case of a positive real part for the first parameter ($\mathfrak{Re}(a) > 0$) 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

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

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

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

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.

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

To align with the general specification structure of regularization problems, which can be seen from equation **missEQ**, a parameter λ will be multiplicatively added in front of the penalty term, which makes it possible to adjust the strength of the influence that the penalty has on the chosen parameters. By minimizing the negative log-posterior adjusted with λ, we can retrieve the *maximum a posteriori probability (MAP)* estimator using **Triple-Gamma-Regularization**:

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

4.1 Varying the Hyperparameters

After closer inspection of equation 6, it can easily be seen that this resembles the general penalized regression already seen in **ESLpage398<empty citation>** and in section 2 as $R(\beta) + \lambda \cdot J(\beta)$. The first term, alos called the empirical loss in machine learning literature, is the widely known residual sum of squares:

$$R(\beta) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

The second part of the optimization problem can be viewed as a penalty imposed on the total risk based on the size of the estimates:

$$J(\beta) = \sum_j^p - \log\left(U\left(c^\xi + \frac{1}{2}, \frac{3}{2} - a^\xi, \frac{\beta_j^2}{2\phi^\xi}\right)\right)$$

In contrast to the *LASSO* penalty, which uses the the absolute value of the coefficient, or the *Ridge* penalty, which uses the square of the estimate, this penalty derived from Cadonna et al. (2020) is based on the log of the confluent hyper-geometric of the second kind. Notably, this penalty term has three additional hyper-parameters: c^ξ, a^ξ and κ as $\phi^\xi = (2c^\xi)/(\kappa_B^2 a^\xi)$.

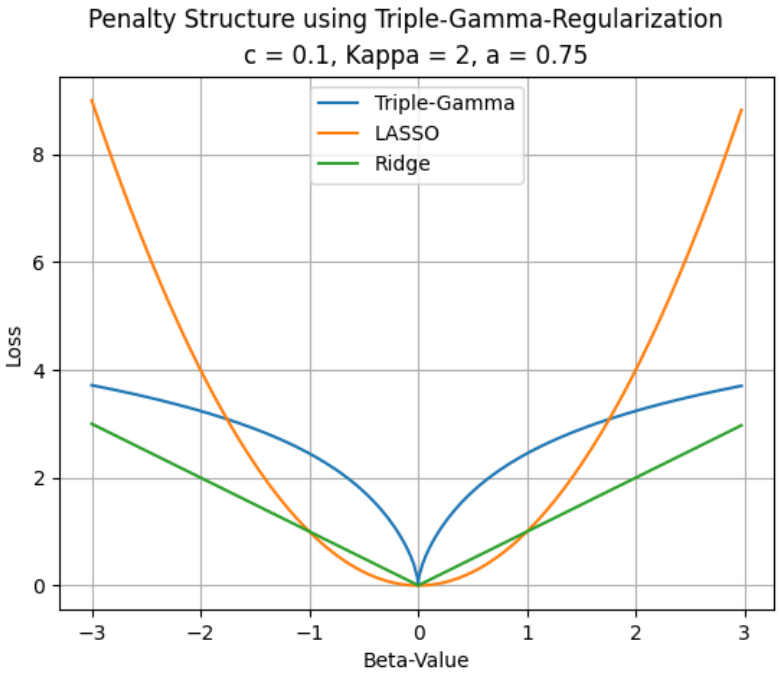

Figure 1: Basic Representation of the Triple-Gamma-Penalty using Hyperparameters $c^\xi = 0.1, \kappa_B = 2, a^\xi = 0.75$ compared to *LASSO* and *Ridge* Penalties

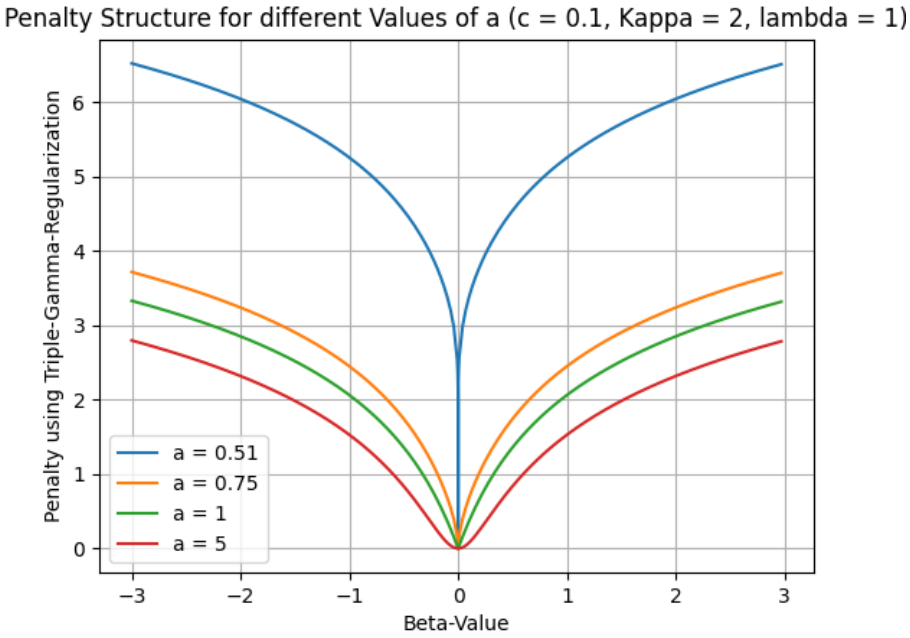

Figure 2: Triple-Gamma-Penalty using different values of a^{ξ}

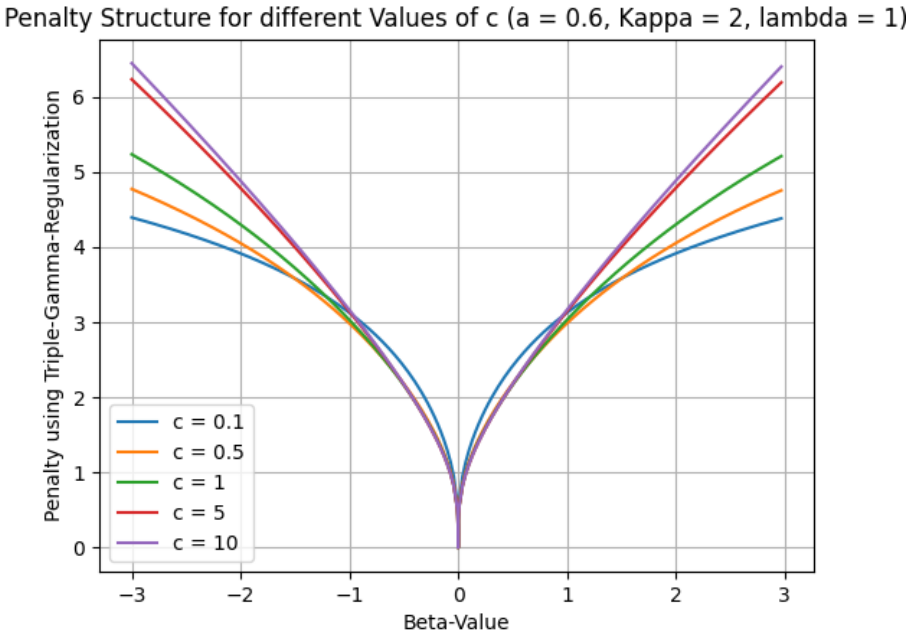

Figure 3: Triple-Gamma-Penalty using different values of c^{ξ}

4.2 Restricted Strong Convexity (RSC) of the Triple-Gamma-Regularization

4.3 Alternative Specification (with +1 to mimic artan)

4.4 Approaches to Estimation

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?

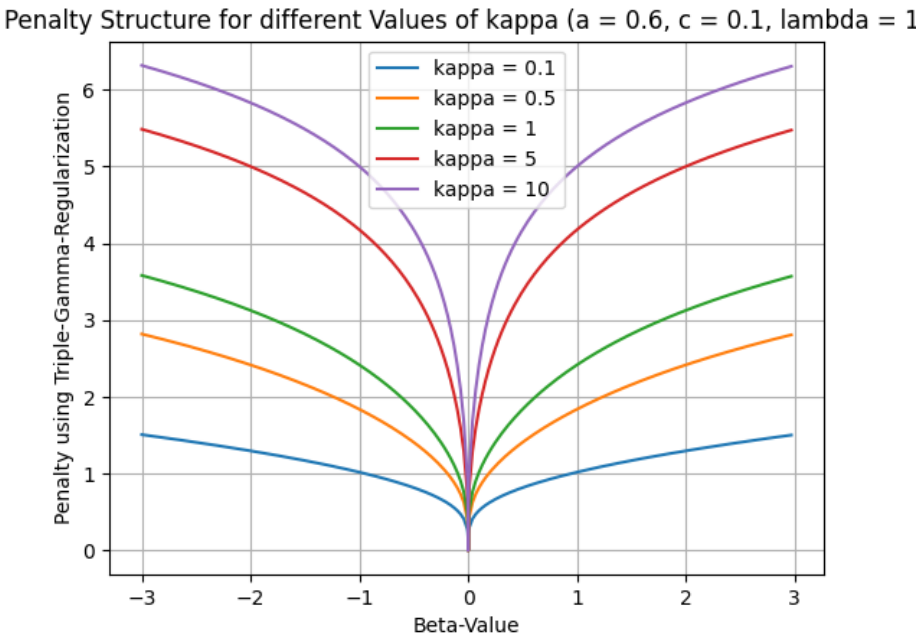

Figure 4: Triple-Gamma-Penalty using different values of κ_B

5.1 Computational Performance

Talking about Gradient Descent Methods in more depth. Why Gradient Clipping. Maybe more modern estimation techniques?

5.2 Implementation as Python Library

How to use it. Explanation of Functions. Input - Output Tables

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.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Kukačka, J., Golkov, V., & Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.

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

Tricomi, F. (1947). Sulle funzioni ipergeometriche confluenti [Paper for hypergeometric function of second kind.]. *Annali di Matematica Pura ed Applicata*, *26*(1), 141–175. https://doi.org/10.1007/BF02415375