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

A Flexible Non-Convex Regularization Penalty based on the Triple-Gamma-Prior

by

Lucas Unterweger

(Student-ID: 11913169)

Abstract

Lorem ipsum...

Acknowledgements

I like to acknowledge ...

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

Generally speaking, a major part of statistical learning deals with trying to describe a certain output variable Y with a set of input variables $X_1, \cdots, X_p$ by trying to find a functional form f which uses the given information in the inputs and - ideally - describes the hidden relationship between Y and the inputs X_i as accurately as possible. However, it comes as no surprise that the functional form f depends on the statistical problem at hand. What type of data has been collected? Are we assuming a linear or non-linear relationship between the predictors? How much data is available and can its quality be guaranteed? But more importantly, it is necessary to ask the question whether the goal of the statistical learning method is *inference* or *prediction*.

The first of these two goals - *inference* - aims at understanding the relationship that may or may not exist between the input variables X_i and the output Y. Especially applied sciences like Economics, Psychology and Medicine often try to find a (causal) relationship within their theoretical framework to evaluate a policy, a medication or a new form of therapy. Linear models for example often provide a simple and straightforward framework which provide the scientist with interpretable effects. *Prediction* on the other hand aims at forecasting the output variable Y as accurately as possible and using every bit of information that is available, but disregards interpretability. (**ESL**; **21**) These goals can be summarized by viewing it as a decision between *prediction accuracy* and *model interpretability*, which has thoroughly been explained by **ESL<empty citation>**.

For the purpose of this thesis, the following chapters will restrict itself to the case of linear models, hence where assume that the functional form f is linear in its inputs. The commonly known *least squares estimator* is one way of estimating such a functional form and due to its many desirable properties has gained popularity in various scientific fields.

2.1 Model Complexity and the Problem of Under- and Overfitting

- BIAS VARIANCE TRADE-OFF - DIMENSIONALITY PROBLEM
General problem: Fit vs. Complexity. What to choose and where is the optimal balance?

In general, a linear regression model fitted with least squares on a sufficient amount of data points ($n \gg p$) will produce estimates which both have low bias and low variance and thus tends to perform well out of sample. However, problems arise with this approach arise when the sample size decreases, because the variance in the estimates will increase. The main idea behind this can be easily be visualized by trying to fit a polynomial curve

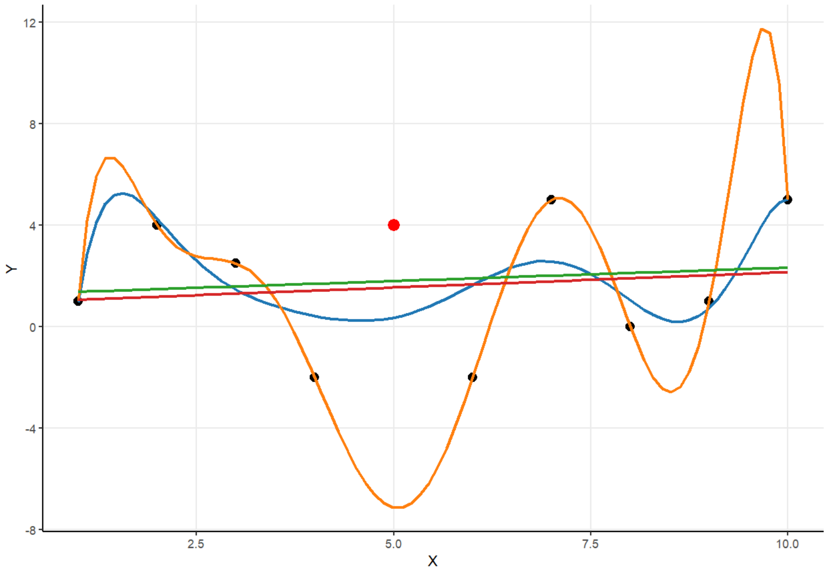


Figure 1: 1st and 8th order polynomial fit to data
(Green & Blue Lines fitted with red data point; Orange & Red Lines fitted without).

on a two dimensional space and then altering the available sample for polynomials of different order.

Such a plot can be found in figure 1. Here, the black points represent the main data sample plotted on a two-dimensional Y-X-space. The red dot plotted at $(5, 4)$ represents the additional data point which is used to alter the respective sample. Polynomials of specific orders can now be used to emulate certain levels of model complexity. For example, a first order polynomial has two coefficients to estimate: the intercept β_0 and the slope β_1. A polynomial of order of order eight has nine coefficients to estimate. What can now be seen in figure 1 is the change in estimates if we include/exclude the additional red data point. A first-order polynomial needs at least two data points and as $n = 9 > p = 2$, the overall fit of the polynomial doesn't change drastically with the additional tenth data point. In the case of the eighth-order polynomial, we need at least nine data points to create a fit as we have nine predictors (one intercept and one coefficient for each of the eight powers). In this case $n = p$ and adding another data point drastically changes the estimates coefficients, which means that the model won't perform well in out-of-sample scenarios. This emphasizes the importance a sufficient sample size when fitting linear models. And this leads us to the core field of study of this thesis: What are ways to improve the generalization of a linear model in scenarios where the sample size is not sufficiently large enough?

- WAYS TO BATTLE COMPLEXITY
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)

2.1.1 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 Vapnik (1991), 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 and neither the targets y_i nor the inputs x_i. 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)*

2.2 Bayesian View on Battling Model Complexity using Shrinkage Priors

Explain how Bayesians battle model complexity.

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. Arguably, this methodological approach to penalized regression started with the publication of two pieces of literature published by Arthur Hoerl and Robert Kennard in 1970 (A. E. Hoerl & Kennard, 1970a, 1970b). With these two papers the authors introduced the widely known *Ridge Regression*, which has been developed from the previously known concept of Ridge analysis. In its core, the authors were trying to tackle the problem of high variances of the regression coefficients in high-dimensional problem settings. This shrinkage estimator, which uses the squared coefficient as a penalty term, "attempt[s] to shrink the coefficients to reduce these variances, while adding some bias." (R. W. Hoerl, 2020) This closely resembles the previously discussed issue of the *Bias-Variance-Tradeoff*, which has been discussed in the *Under- and Overfitting* chapter in section 2 (Roger W. Hoerl, Arthur Hoerl's son, published a historical overview of the development of the concept of *Ridge Regression* in 2020 (R. W. Hoerl, 2020)). The closed from solution of the Ridge estimator is given by

$$\hat{\beta}_{Ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

, which adjusts the OLS estimator by shifting the main diagonal entries of the design matrix by λ ($\lambda \geq 0$). It can be shown that this closed form estimator is equivalent to a Lagrangian problem of the following form (van Wieringen, 2015):

$$\hat{\beta}_{Ridge}(\lambda) = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

This resembles an example of the above defined regularization framework with squared residual loss and a penalty term of the form $\|\beta\|_2^2$, which only depends on the parameter β. In case of λ being equal to zero, this reduces to the *maximum likelihood (ML) estimator*.

The publications of Arthur Hoerl and Robert Kennard have led to further advancements, although it took more than 25 years, in shrinkage estimation or related concepts. One concept which is almost as famous *Ridge Regression* is the *Least Absolute Shrinkage and Selection Operator*, more commonly know as *LASSO*, developed by Tibshirani (1996). He argues that the two at the time most prominent shrinkage methods - Ridge and Subset Selection - both have their drawbacks. Ridge regression on the one hand is an optimization problem which continuously shrinks coefficients towards zero, but doesn't select them in a discrete sense, which makes it hard to interpret these models. Subset Selection on the other hand chooses variables in a discrete sense - a variables either stays within the model or it doesn't - and thus creates easily interpretable models, but "[s]mall changes in the data can result in very different models being selected and this can reduce its prediction accuracy." (Tibshirani, 1996) *LASSO* is trying to combine both methods' advantages by using $\|\beta\|_1$ as a penalty term.

Bit more on LASSO?

LASSO and to some extend *Ridge* can be viewed as a special case of a l_p-norm regularization with corresponding values for p ($p = 1$ for LASSO and $p = 2$ for Ridge) (Frank & Friedman, 1993).

$$\|\beta\|_p = \left(\sum_{i=1}^{p} |\beta_i|^p \right)^{1/p}$$

Work published by researchers in the nineties, like the previously mentioned Frank and Friedman (1993) or Fu (1998), as well as more recent literature like F. Wang et al. (2020) have repeatedly shown there is no go-to-method to tackle regularization problems, as the effectiveness of a specific approach highly depends on the data siuation at hand. Due to this particular situation in the literature, several other methods have been proposed in the recent years and decades. An approach combining *Ridge* and *LASSO* two methods is called *Elastic Net* regularization and has been developed by Zou and Hastie (2005). The authors there elaborate on some of the shortcomings of the LASSO method. For example, in a special case where there are more predictors p than data points n ($p > n$), LASSO only selects up to n variables due to the nature of the convex optimization problem. Should several of the included variables be highly *pairwise* correlated with each other, *LASSO* tends to only select on of these variables. In its core, *Elastic Net Regularization* linearly combines the penalty terms of *Ridge* and *LASSO* regularization, yielding a loss function of the form:

$$J(\beta) = \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\beta\|_1$$

The authors have shown that, especially when it comes to encouraging the aforementioned grouping effects, *Elastic Net* tends to perform better than the *LASSO*.

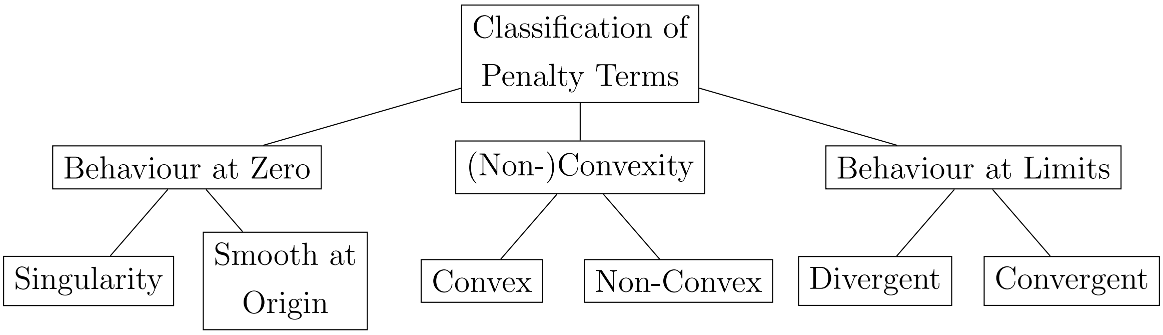
Recent years however have opened up a new subfield of approaches to regularization. Methods like *LASSO, Bridge* (Frank & Friedman, 1993), *Ridge* and *Elastic Net*[1] are convex functions in its parameters and are thus usually classified as *Convex Regularization Penalties*. Recently, the literature has shifted towards penalties which are non-convex functions in its parameters, usually called *Non-Convex Regularization Penalties*[2]. One recent example of such a penalty includes John et al. (2022), who proposed a penalty

[1] Elastic Net, due to its mathematical definition, can be view as a generalization of *Ridge* and *LASSO*.

[2] Note: They are called *non-convex* penalties and not *concave* penalties, because non-convexity does not necessarily imply concavity.

structure called *Gaussian penalty* and which is based on a Gaussian-like function by using $J(\beta) = 1 - e^{-\kappa\beta^2}$. Another method proposed by Y. Wang and Zhu (2016) is called the *Atan penalty* - or *Arctan* penalty - and makes use of the favourable properties of the *Arctan* function by using the penalty $J(\beta,\gamma) = (\gamma + \frac{2}{\pi})\arctan(\frac{|\beta|}{\gamma})$. Several others include *SCAD* (Fan & Li, 2001), *MCP* (Zhang, 2010) or *Laplace* (Trzasko & Manduca, 2009) penalties.

Seeing this vast array of different pieces of literature immediately raises the question on advantages and disadvantages of specific methods and why so many have established itself in this particular field of study. Keeping in mind that the effectiveness of a method still highly depends on the data situation at hand, it is still important to distinguish methods based on its mathematical properties. John et al. (2022) have distinguished methods based on two broad criteria: (1) How does a penalty behave in small neighbourhoods around zero? (Is it smooth or singular at origin?) and (2) Is the underlying function convex or non-convex? To better incorporate the proposed concept of this thesis into this body of literature, I am proposing a third property to distinguish penalty terms: (3) How does the function behave in the limits? (Does it converge or diverge?)



As already mentioned earlier in section 2, a penalty function is a function dependent on the parameter of the model and it does not depend on the data at hand. Thus, changes in your data set does not effect the penalty directly, but only the overall optimization problem through the empirical risk functional R_{emp}. The following table classifies some of the existing concepts based on the three criteria mentioned earlier. A visualisation of them can be found in figure 6 in chapter 5.2.

Depending on what property a penalty has, it can perform specific tasks or provide challenges in application. As thoroughly described by (John et al., 2022), penalties with a singularity at origin tend to be suitable when the goal of the statistical analysis is variable selection but ”could pose theoretical and computational challenges when the focus is on regularization alone without variable selection.” In addition, it has been shown in the literature that ”objective functions which are a sum of a nonconvex, singular-at-origin penalty function coupled with either a smooth nonconvex loss function or a nonsmooth convex loss function, often fails to satisfy a theoretical condition known as ‘Clarke regularity’.” (A result presented in **QiCuiLiuPang2021<empty citation>**) Convex penalties are generally easier to implement using common optimization algorithms than non-convex penalties and, more importantly, penalties which are divergent and its limits, which most convex penalties are, yield biased results. Due to the fact that, when using divergent and convex penalties like LASSO and Ridge, the additional penalty for larger estimates keeps rising with estimates which deviate from zero, estimates are getting artificially pulled towards zero and are thus biased. Non-Convex penalties, which are usually convergent in its limits towards $-\infty$ and $+\infty$, tend to yield unbiased results as the additional penalty becomes zero as soon as a certain threshold is crossed (Again, referring to the visualisation in figure 6).

Penalty		Classification		
Penalty	Reference	Behaviour at Origin	(Non-)Convexity	Limits
LASSO	Tibshirani (1996)	Singular	Convex	Divergent
Ridge	A. E. Hoerl and Kennard (1970b)	Smooth	Convex	Divergent
Gaussian	John et al. (2022)	Smooth	Non-Convex	Convergent
Ar(c)tan	Y. Wang and Zhu (2016)	Singular	Non-Convex	Convergent
Elastic-Net	Zou and Hastie (2005)	Singular	Convex	
Bridge	Frank and Friedman (1993)	Both		
MCAP	Zhang (2010)	Singular	Non-Convex	
SCAD	Fan and Li (2001)	Singular	Non-Convex	
Laplace	Trzasko and Manduca (2009)	Non-Convex		
Triple-Gamma	-	Depending on parameters	Non-Convex	Divergent

Table 1: Classification of several Penalties

4 Bayesian-Frequentist Duality of Ridge and LASSO Regression

The main motivation for this thesis stems from a striking duality that exists between Bayesian shrinkage priors (discussed in chapter 2.2) and the frequentist approach using regularization. Both approaches aim at tackling the issue of model complexity and overfitting by making it necessary for the data to be more convincing that the value of an estimate is statistically significant different from zero. As mentioned, Bayesian statistics use specific prior distributions with a usually a lot of mass around zero and heavy tails, whereas frequentist statisticians usually alter their optimization problems incorporating penalty terms into into their loss functions. On a first glance, these two approaches have no immediate mathematical connection, however this turns out to be wrong.

In a Bayesian setting, one usually assumes that the parameter vector β has a prior distribution $p(\beta$. By multiplying it with the likelihood of the data $f(y|X,\beta)$ and by utilizing Bayes' rule, one retrieves the posterior distribution of the parameter distribution, up to a proportionality constant:

$$p(\beta|y,X) \propto f(y|X,\beta) \times p(\beta)$$

It turns out, as can be seen in Hastie et al. (2009, pp. 248–250) and van Wieringen (2015) among others, that when choosing certain prior distributions and using standard assumptions in Bayesian modelling (ie the individual parameters of the parameter vector are independent a prior and the errors of the standard linear model are drawn from a normal distribution), that the moments of the posterior distributions correspond to the point estimates of a regularization approach in a frequentist setting. As this idea is essential to the entire thesis, a shortened derivation of this duality for the case of *Ridge* regression will now be shown.

Assuming that a response variable y_i follows a linear model with k variables and errors $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then the likelihood of the data is given by

$$\mathcal{L}(y|X,\beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right).$$

Under the previously mentioned assumption that $p(\beta) = \prod_{i=1}^{k}$ and by choosing a Gaussian distribution with mean zero and variance τ^2, we gain the prior distribution

$$p(\beta) = \frac{1}{(2\pi\tau^2)^{p/2}} \exp\left(-\frac{1}{2\tau^2}\left\lVert\beta\right\rVert_2^2\right).$$

Putting both together using Bayes' theorem yields and taking the log of the distribution yields

$$\log(p(\beta|y,X)) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{1}{2\tau^2}\left\lVert\beta\right\rVert_2^2.$$

By viewing $\frac{1}{2\tau^2}$ as the weighting parameter λ, which controls the strength of the regularization. The *maximum-a-priori* estimate $\hat{\beta}_{MAP}$, which can be viewed as trying to find the $\arg\max_{\beta}$ of the log-posterior, can be used to construct the following optimization problem:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{2\tau^2} \|\beta\|_2^2 \right\}$$

This optimization problem constitutes the connection between the Bayesian and the frequentist approach as it can be interpreted as (1) finding the mode of the posterior distribution when using a Gaussian prior distribution and (2) finding the parameters which minimize the loss when using a Ridge regression with $\lambda = 1/(2\tau^2)$. The same connection exists for the case of *LASSO* regression which can be derived using the same procedure and a double-exponential - also called *Laplace* - distribution with mean zero and some scale parameter τ as the prior for the parameter vector β.

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

This opens up a path of possibility to utilize this connection and derive a new regularization penalty by using a different shrinkage prior. Recently, a new development in the are of shrinkage priors has been made by Cadonna et al. (2020), who proposed a new shrinkage prior called the *Triple-Gamma-Prior* which has several advantageous properties and also comes with a closed-form solution of the marginal prior distribution for β, which sets the building block for this thesis' research question:

> Can the closed-form marginal distribution of the Triple-Gamma-Prior be used to derive a new regularization penalty and do its advantages carry over into the frequentist framework?

Yet, before diving into the mathematical core of this thesis which derives said concept, it is necessary to have a look at the Triple-Gamma-Prior in depth to explain why this particular prior might prove itself to provide a useful building block for a novel regularization penalty.

5 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 Theorem 1 (a) and is given by

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

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

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

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

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

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.

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

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**:

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

5.1 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, also 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:

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

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_B as $\phi^{\xi} = (2c^{\xi})/(\kappa_B^2 a^{\xi})$. Here, the restrictions $a^{\xi} > 0.5$ and $0 < c^{\xi} < \infty$ are necessary to ensure that the penalty for a β_j being equal to zero remains finite and not diverges to negative infinity at zero. This results, which has already been presented and proven as part of Theorem 2 in Cadonna et al. (2020, pp. 5–6), ensures that the negative log of the hypergeometric function remains finite and thus does not produce parameter estimates which are zero for every variables. (**BESSER SCHREIBEN?**)

Write more to fill this page

Variations of the Hyperparameter a^{ξ}

The first hyper-parameter which can be adjusted is a^{ξ}. A plot with a set of different values for a^{ξ} can be found in figure 2. As already mentioned earlier, the necessary restriction for for this hyper-parameter is that it has to be strictly greater than 0.5 to guarantee the finiteness of the penalty. To demonstrate the effects of changes in a^{ξ}, the other parameters have been set to $c^{\xi} = 0.1$ and $\kappa_B = 2$. It can easily be seen from the figure that a^{ξ} steers the sharpness of the penalty in small neighbourhoods around $\beta = 0$. As a^{ξ} increases, the penalty because smoother at $\beta = 0$ with it eventually converging a *Gaussian Penalty* like behaviour. From a modelling perspective, this opens up the possibility of steering the degree of variable selection the penalty performs. Nonetheless, the overall structure of the penalty in the tails does not change systematically apart from a parallel shift, which can be readjusted by specifying a different weighting parameter λ or a different value for κ_B (more on effect of κ_B on the penalty can be found later in this chapter).

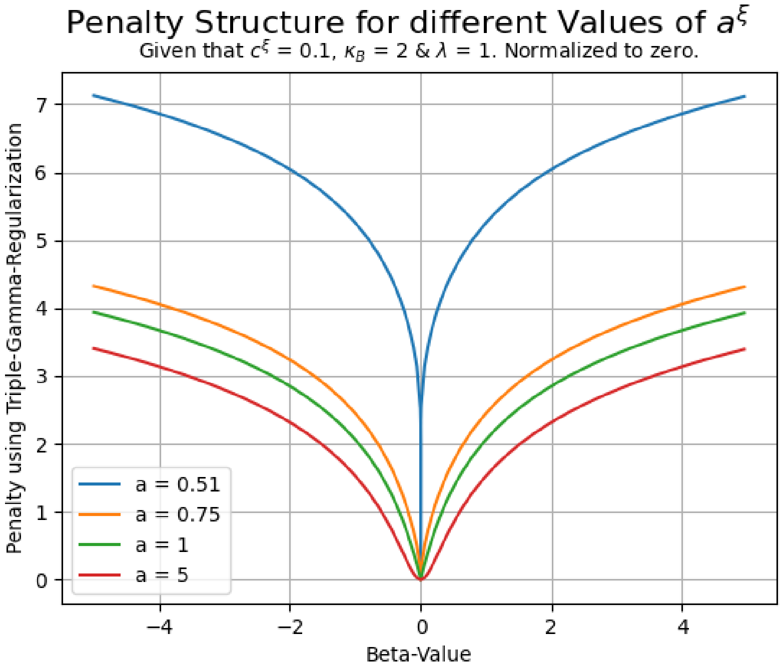


Figure 2: Triple-Gamma-Penalty using different values of a^{ξ}

Seeing this, it is apparent that a value of a^{ξ} close but strictly larger that 0.5 mimics the behaviour of the *Arctan Penalty* by Y. Wang and Zhu (2016) in small neighbourhoods of $\beta = 0$. Similar, large positive values for a^{ξ} lead to a *Gaussian Penalty* like behaviour in small neighbourhoods of $\beta = 0$ as recently proposed by John et al. (2022).

Variations of the Hyperparameter c^{ξ}

In contrast to the hyperparameter a^{ξ}, which mainly affects the behaviour at and around $\beta = 0$, changes in c^{ξ} mainly affect the behaviour in the tails. However, the effect that a change in c^{ξ} has on the penalty structure can be split up in two rough subsets of $(0, \infty)$. The effect of the first subset of values for c^{ξ} which are strictly greater than 0 but less or equal than 0.1 can be found in figure 3 (as already mentioned before, by definition, c^{ξ} has to be strictly greater than zero: $c^{\xi} > 0$). Here, it can be seen that as the values for c^{ξ} become smaller, a shifting effect takes place which generally does not influence the overall structure of the penalty, but increases the amount of penalty which is added to the risk function for β-values which are different from zero (In a sense, this has a similar effect to changes in κ_B, which will be explained later). Or, to put it differently, with values of c^{ξ} closer to zero, the data has be become even more convincing that the value is significantly different from zero.

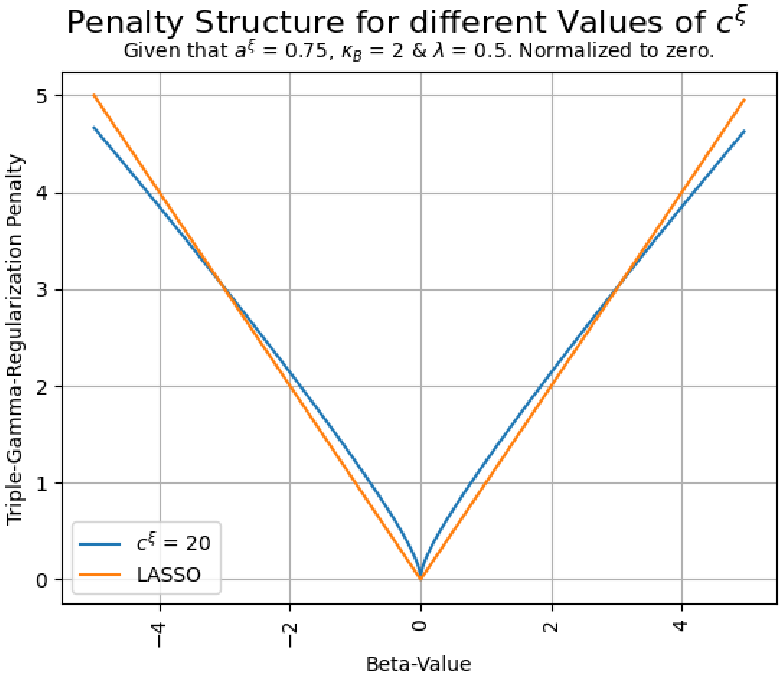


Figure 3: Triple-Gamma-Penalty using different values of c^{ξ} with $0 < c^{\xi} \leq 0.1$

However, the more interesting effect that a change in c^{ξ} has on the penalty structure can be seen for values of c^{ξ} that are greater than 0.1. In figure 4, a plot can be found with the Triple-Gamma-Penalty for larger values of c^{ξ}. Again, starting from the baseline with $c^{\xi} = 0.1$, higher values for this hyper-parameter mainly change the behaviour of the penalty in the tails. A result that has already been shown by Cadonna et al. (2020) in Table 1, where multiple different hyper-parameter settings are presented, is that with an increasing value for c^ξ and with $a^\xi = 1$ **as well as kappaB = ??**, the Triple-Gamma-Prior converges to a *LASSO* like shrinkage behaviour. A property which carries over to the proposed regularization setting when using the proposed hyper-parameter values, thus showing that the Triple-Gamma-Penalty can be used both as a non-convex penalty as well as a *LASSO* penalty, creating increased flexibility in modelling approaches.

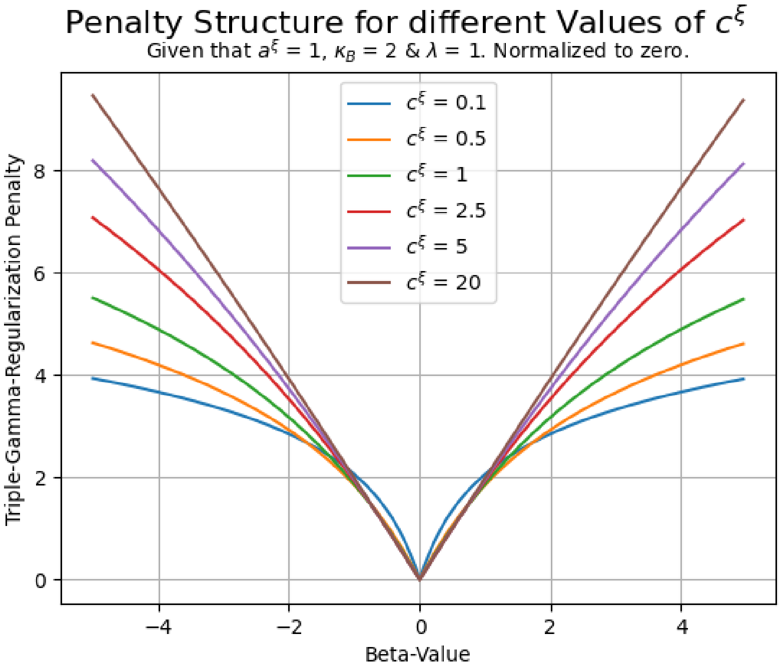


Figure 4: Triple-Gamma-Penalty using different values of c^ξ with $c^\xi \geq 0.1$

Variations of the Hyperparameter κ_B

The third and final hyper-parameter κ_B enters the Triple-Gamma-Penalty $J_{TG}(\beta)$ as part of $\phi^\xi = \frac{2c^\xi}{\kappa_B^2 a^\xi}$ as can be seen from equation 6. Notably, κ_B is squared and thus only the absolute value of κ_B, $|\kappa_B|$, influences the structure of the penalty. The overall third function value is defined as $\frac{\beta_j^2}{2\phi^\xi}$ and by plugging in ϕ^ξ we get $\frac{\beta_j^2 \kappa_B^2 a^\xi}{4c^\xi}$, it can be seen that a value of $\kappa_B = 0$ leads to the entire parameter value being zero for all values of β_j. Hence, a change in β_j won't influence the penalty and furthermore won't have an influence on the overall risk minimization problem, the result being that the optimal set of parameters will only depend on the chosen loss function.

For all values of $\kappa_B \neq 0$, the value of the hyper-parameter will influence the penalty structure. A plot with several different values for the absolute value of κ_B can be found in figure 5.

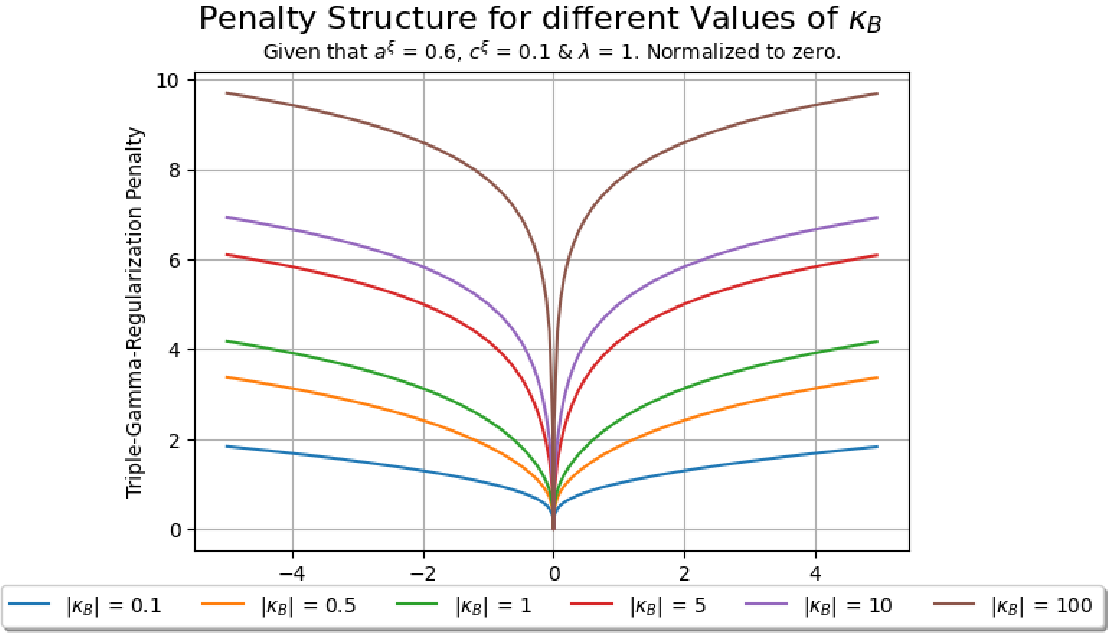


Figure 5: Triple-Gamma-Penalty using different values of κ_B

	Change		Mathematical Properties	
Variable	**Positive** Change	**Negative** Change	Defined Range	Misc.
a^{ξ}	Shifting towards *Gaussian*-like behaviour at $\beta = 0$	Shifting towards singularity at $\beta = 0$; Higher immediate penalty for coefficients $\beta \neq 0$	$(0.5, \infty)$	Expl.
c^{ξ}	Generally, convergence towards convexity and, given certain settings for a^{ξ} and κ_B, LASSO. Higher values increase the additional penalty for higher absolute values of β.	For values smaller than 0.1, similar effect to increase in a^{ξ}	$(0, \infty)$	Expl.
κ_B	15883	5.2e-8	$(-\infty, \infty)/\{0\}$	Expl.

Table 2: Summary of the effects of changes in the hyperparameters a^{ξ}, c^{ξ} and κ_B on the penalty structure

5.2 Comparison to already existing Penalty Terms

As already mentioned in section 3, several other penalty terms have already been widely studied in the literature. Convex penalties like *Ridge* (A. E. Hoerl & Kennard, 1970b) or non-convex penalties like the *Ar(c)tan* (Y. Wang & Zhu, 2016) and *Gaussian* (John et al., 2022) have managed to establish itself as prominent approaches to regularization. It is now certainly of interest to see how the *Triple-Gamma* penalty compares to the established methods. Seeing that, due to its flexibility, there is not *one Triple-Gamma* penalty, three distinct hyper-parameter setting have been chosen to represent the proposed penalty term.

Describe differences

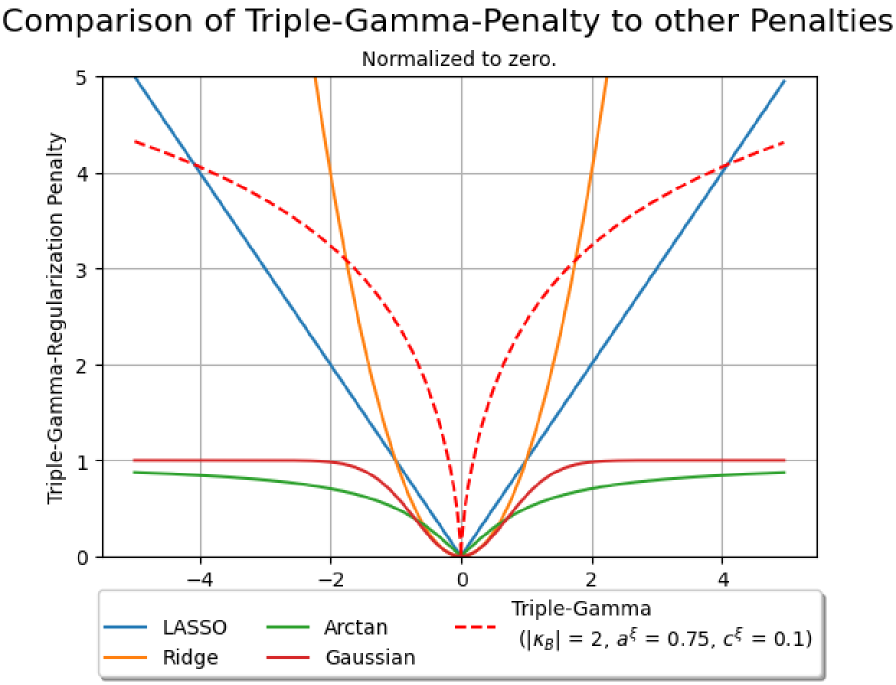


Figure 6: Basic Representation of the Triple-Gamma-Penalty using Hyperparameters $c^{\xi} = 0.1, \kappa_B = 2, a^{\xi} = 0.75$ compared to *LASSO* and *Ridge* Penalties

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

Weglassen?

5.4 Alternative Specification (with +1 to mimic artan)

As previously mentioned, the *Arctan* penalty developed by Y. Wang and Zhu (2016) has a similar structure as the *Triple-Gamma-Penalty* in small neighbourhoods near $\beta = 0$

when using specific hyper-parameters (see section 5.2). The distinctive structural difference between these two penalties emerge when looking at the limits when $\lim_{\beta \to +/-\infty} J_{TG}(\beta)$

Propose different conecpts?

5.5 Approaches to Estimation

6 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?

Ganze Simulation

6.1 Implementation as Python Package

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

6.2 Computational Performance

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

See how computation times change when increasing the size of the data set or when the number of parameter changes. =¿ Use Stochastic Gradient Descent (SGD)

Kurzes Ansprechen von Computations Efficieny

7 Possible Extensions and Criticism

7.1 Implementation as Python Package

Interessante unifying properties, aber langsam implementiert.

8 Conclusion

8.1 Implementation as Python Package

ENDE

9 List of Figures

List of Figures

10 List of Tables

List of Tables

11 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.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of computational and graphical statistics*, *7*(3), 397–416.

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

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoerl, R. W. (2020). Ridge regression: A historical context. *Technometrics*, *62*(4), 420–425.

John, M., Vettam, S., & Wu, Y. (2022). A novel nonconvex, smooth-at-origin penalty for statistical learning. *arXiv preprint arXiv:2204.03123*.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

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

Trzasko, J., & Manduca, A. (2009). Relaxed conditions for sparse signal recovery with general concave priors. *IEEE Transactions on Signal Processing*, *57*(11), 4347–4354.

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Vapnik, V. (1991). Principles of risk minimization for learning theory [Definition of Empirical Risk Minimization]. *Advances in neural information processing systems*, *4*.

Wang, F., Mukherjee, S., Richardson, S., & Hill, S. M. (2020). High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking [Paper with simulation study]. *Statistics and computing*, *30*, 697–719.

Wang, Y., & Zhu, L. (2016). Variable selection and parameter estimation with the atan regularization method. *Journal of Probability and Statistics*, *2016*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.