Triple-Gamma-Regularization

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

Unterweger Lucas Paul

Vienna University of Econonomics and Business (WU), Department of Economics

21st of June 2024

Overview

Motivating Problem

Back to contents

Overfitting I

- **Sparse Data Settings**: Scenarios where data is sparse, meaning the number of data points is limited relative to the number of features.
- **High Dimensionality Problem**: Model is at risk of overfitting because the model can fit the noise than the underlying pattern.
- **Ill-posed Problems in Regression**: Solution to the problem becomes sensitive to small changes in the data, resulting in large variances in the estimated parameters $\implies$ model's predictions may generalize poorly.
- **Bias-Variance Tradeoff**: High variance models (*overfitting*) capture noise and fluctuations in training data $\implies$ poor generalization. High bias models (*underfitting*) fail to capture the underlying trend. Hard to find balance!

Overfitting II

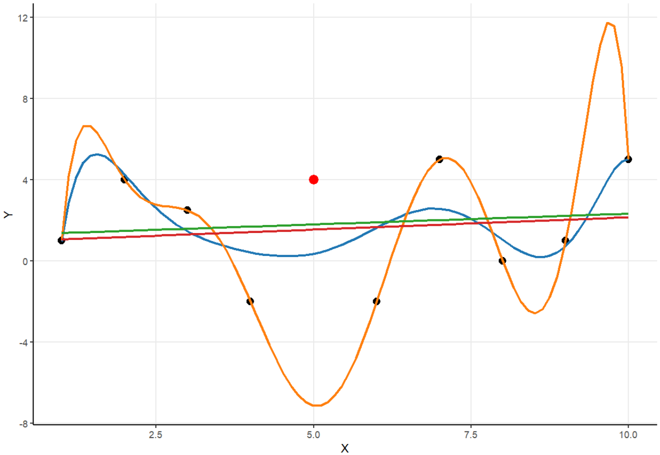


Figure 1: 1st and 8th order polynomial fit to data (Green & Blue Lines fitted with extra red data point; Orange & Red Lines fitted without).

Some Solutions to Ill-posed Problems

1. **Cross-Validation** Estimate model based on subsets of the data to make estimates more robust. However, often not feasible due to data availablity issues.
2. **Feature Selection** Selecting a subset of relevant features, but often reliant on strict assumptions.
3. **Data Augmentation and Acquisition** Gather more data, use stochastic approaches to estimates your models, but similar issue as with *Cross-Validation*.
4. **Ensemble Methods** Combining the predictions of multiple models (e.g., bagging, boosting).
5. **Regularization using Loss Penalty** Induce shrinkage on esimates and penalize too-complex models by altering the Loss-function

Regularization

This thesis focuses on the 5th approach to *regularization*, which adds a penalty term to the risk minimization problem.

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

where $L(.)$ refers to a loss function defined as some function of the true values y_i and the predicted values $f(x_i)$[1] and $J(f)$ is a penalty based on the chosen functional from a space of functions $\mathcal{H}$.

[1] In the setting of OLS, this would be the *sum of squared residuals (SSR)*

Existing Concepts

Back to contents

Proposed Concepts

Name	Penalty	Reference
Ridge	$\|\|\beta\|\|^2$	(Hoerl & Kennard, 1970)
LASSO	$\|\|\beta\|\|$	(Tibshirani, 1996)
Elastic Net	$\lambda_1 \|\|\beta\|\|^2 + \lambda_2 \|\|\beta\|\|_1$	(Zou & Hastie, 2005)
Arctan	$\frac{2}{\pi} \arctan(\|\beta\|)$	(Y. Wang & Zhu, 2016)
Gaussian	$1 - e^{-\beta^2}$	(John et al., 2022)

Table 1: Established Regularization Penalties

- others are *SCAD*, *MCP*, *SILO*, *Dantzig Selector*, ...
- F. Wang et al. (2020) find no "go-to" method which suits a broad range of problems

Motivating Duality of Ridge and LASSO

- An interesting mathematical connection can be found when looking at this problem from a Bayesian point of view.
- In Bayesian regression, the *posterior distribution* (up to a proportionality constant) takes the form

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

  with β being a coefficient vector, Y being the target vector and X being a matrix of features.
- Choosing specific prior distributions lead to posterior distributions which have moments that correspond to point estimates from regularization (*Gaussian prior*: Ridge; *double-exponential prior*: LASSO)

The Triple-Gamma-Prior

- Cadonna et al. (2020) developed a new prior distribution which has unifying properties and provides a general form for several shrinkage effects. Morevoer, it is given by a closed-form solution:

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

Hypothesis/Aim of this thesis:

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?

Mathematical Derivation

Back to contents

Likelihood

Let's assume we have n data points of a response variable $\mathbf{y}$ and and a set of features $\mathbf{X}$. Assuming a standard linear model with a parameter vector β and standard normal i.i.d. errors, the *likelihood* is

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

Prior

Assuming the individual parameters are independent a priori and using the *Triple-Gamma-Prior* from Cadonna et al. (2020), the *prior* distribution of the parameter vector β is

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

Posterior

Using Bayes' Theorem, the posterior distribution (up to a proportionality constant) $p(\beta|\mathbf{X},\mathbf{y},\sigma^2)$ is then given by

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

Taking the $\log$ yields

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

Triple-Gamma-Regularization

Finally, the Triple-Gamma-Regularization can be retrieved from looking at the *maximum-a-posteriori* estimate, which minimizes the **negative**(!) log-posterior:

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

Thus, the **Triple-Gamma-Penalty** is given by:

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

Properties of the Triple-Gamma Penalty

Back to contents

Comparison to other Penalties

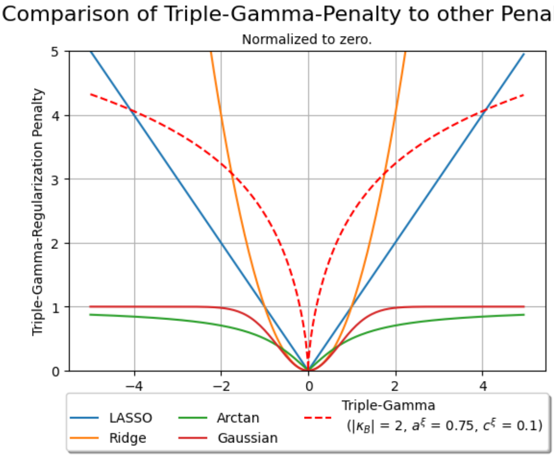
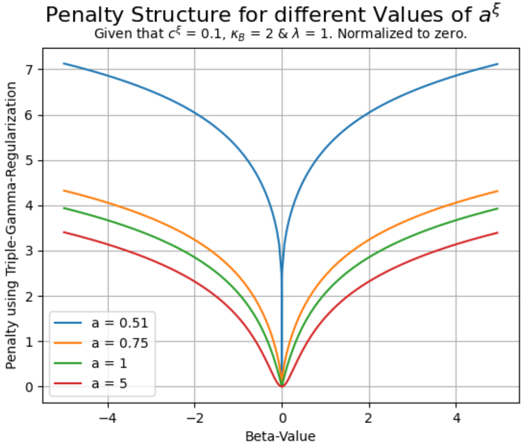


Figure 2: Comparison of one form of the Triple-Gamma-Penalty to existing penalties

Varying a^{ξ}


Penalty Structure for different Values of a^{ξ}
Given that $c^{\xi} = 0.1$, $\kappa_B = 2$ & $\lambda = 1$. Normalized to zero.
Penalty using Triple-Gamma-Regularization
Beta-Value
a = 0.51
a = 0.75
a = 1
a = 5
0
1
2
3
4
5
6
7
−4
−2
0
2
4

Varying c^ξ

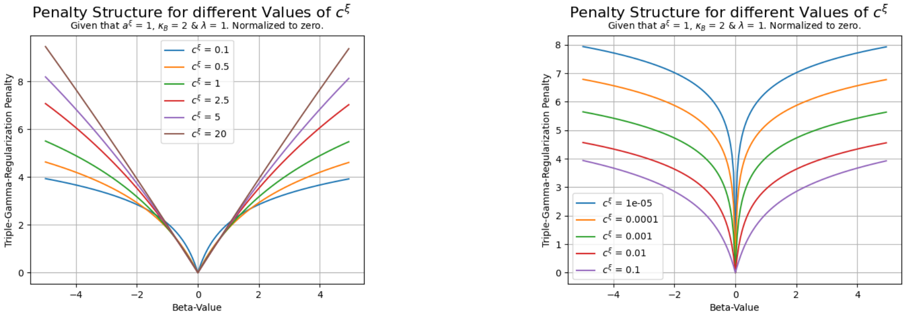
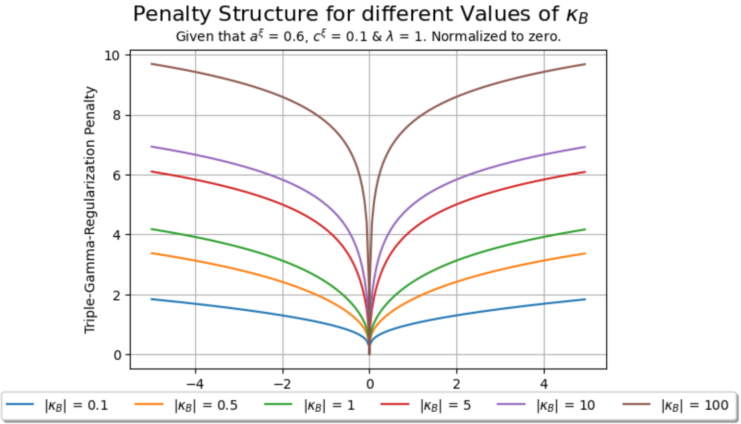


Figure 3: (Left) Varying $c^\xi \to \infty$, (Right) Varying $c^\xi \to 0$

Varying κ_B[2]



[2]Note that per definition $\phi^\xi = (2c^\xi)/(\kappa_2^\xi a^\xi)$

Simulation Study

Back to contents

Sparse Data Setting

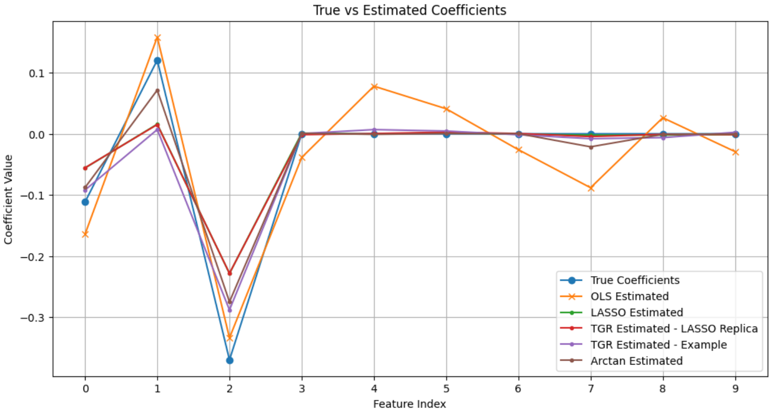


Figure 4: Comparison of Estimates from several Regularization Approaches $(n = 100, p = 10)$

Simulation Study

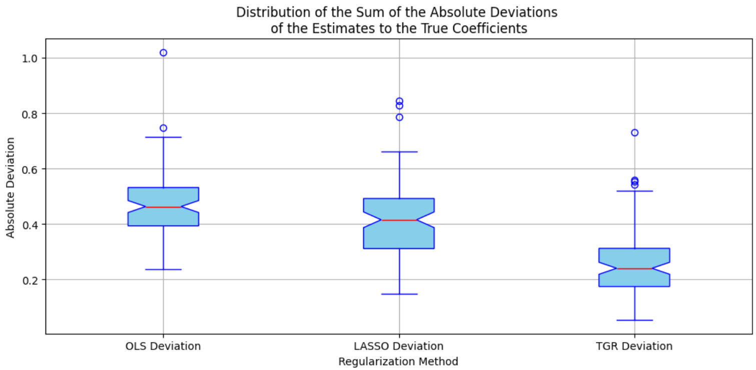


Figure 5: Distribution of the Sum of the Absolute Deviations of the Estimates to the True Coefficients (200 runs)

Shortcomings and Discussion

Back to contents

Summary, Shortcomings & Potential Extensions

- Implementation in Python still relatively slow compared to similar approaches
- Cannot reproduce effects of converging non-convex penalites like *Arctan*, *Gaussian*, *etc*

BUT

- It can replicate results from e.g. *LASSO* regression or induce its own form of shrinkage
- *Triple-Gamma-Penalty* is a flexible regularization penalty that corresponds with the Bayesian *Triple-Gamma-Prior*

References I

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.

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

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

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

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. *Statistics and computing*, *30*, 697–719.

References II

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

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.

Appendix

Back to contents

Unifying Property of the Triple-Gamma-Prior

Table 1. Priors on $\sqrt{\theta_j}$ which are equivalent to (top) or special cases of (bottom) the triple gamma prior.

Prior for $\sqrt{\theta_j}$		a^ξ	c^ξ	κ_B^2	ϕ^ξ
$\mathcal{N}\left(0,\psi_j^2\right), \psi_j^2 \sim \mathrm{GG}\left(a^\xi, c^\xi, \phi^\xi\right)$	normal-gamma-gamma	a^ξ	c^ξ	$\frac{2c^\xi}{\phi^\xi a^\xi}$	ϕ^ξ
$\mathcal{N}\left(0,\frac{1}{\kappa_j}-1\right), \kappa_j \sim \mathcal{TPB}\left(a^\xi, c^\xi, \phi^\xi\right)$	generalized beta mixture	a^ξ	c^ξ	$\frac{2c^\xi}{\phi^\xi a^\xi}$	ϕ^ξ
$\mathcal{N}\left(0,\psi_j^2\right), \psi_j^2 \sim \mathrm{SBeta2}\left(a^\xi, c^\xi, \phi^\xi\right)$	hierarchical scaled beta2	a^ξ	c^ξ	$\frac{2c^\xi}{\phi^\xi a^\xi}$	ϕ^ξ
$\mathcal{DE}\left(0,\sqrt{2}\psi_j\right), \psi_j^2 \sim \mathcal{G}\left(c^\xi, \frac{1}{\lambda^2}\right)$	normal-exponential-gamma	1	c^ξ	$2\lambda^2 c^\xi$	$\frac{1}{\lambda^2}$
$\mathcal{N}\left(0,\tau^2\psi_j^2\right), \psi_j \sim t_1$	Horseshoe	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{\tau^2}$	τ^2
$\mathcal{N}\left(0,\frac{1}{\kappa_j}-1\right), \kappa_j \sim \mathcal{B}(1/2, 1)$	Strawderman-Berger	$\frac{1}{2}$	1	4	1
$\mathcal{N}\left(0,\tau^2\tilde{\xi}_j\right), \tilde{\xi}_j \sim \mathcal{G}\left(a^\xi, a^\xi\right)$	double gamma	a^ξ	∞	$\frac{2}{\tau^2}$	-
$\mathcal{N}\left(0,\tau^2\tilde{\xi}_j\right), \tilde{\xi}_j \sim \mathcal{E}(1)$	Lasso	1	∞	$\frac{2}{\tau^2}$	-
$t_\nu\left(0,\tau^2\right)$	half-t	∞	$\frac{\nu}{2}$	$\frac{2}{\tau^2}$	-
$t_1\left(0,\tau^2\right)$	half-Cauchy	∞	$\frac{1}{2}$	$\frac{2}{\tau^2}$	-
$\mathcal{N}(0, B_0)$	normal	∞	∞	$\frac{2}{B_0}$	-

Figure 6: Table 1 from Cadonna et al. (2020)