"Honey, I shrunk the parameters": A Comparison of Lasso, Ridge, and Elastic Net Methods

Carolina Alvarez, Edoardo Falchi, and Emily Anne Schwab

Research Module in Econometrics and Statistics

January 18, 2022

Table of Contents

Introduction: Drawbacks to OLS

- Consider the following multivariate linear regression, where $Y = X\beta + \varepsilon$:
  - $Y = (y_1, ..., y_n) \in \mathbb{R}^n$ is the response vector
  - $X \in \mathbb{R}^{n \times p}$ is the predictor variable matrix
  - $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ is the error vector
- When is OLS a sub-optimal model for prediction?
  - Imperfect multicollinearity
    - $cov(X_i, X_j) \neq 0$
    - β_{OLS} still BLUE, but variance and standard errors will increase.
    - Leads to lower t-statistics.
  - High-dimensionality ($p > n$)
    - Rank $(X) < p$ such that OLS does not generate a unique solution.
    - Variance of estimated coefficients becomes infinitely large.

Regularization Methods

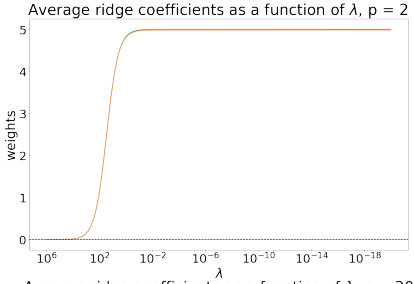
Solution: Use regularization method(s) to reduce the variance of the model fit in exchange for a marginal increase in the bias.

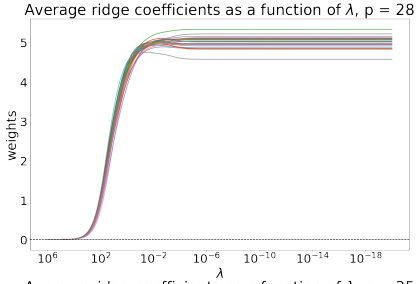
Ridge Regression

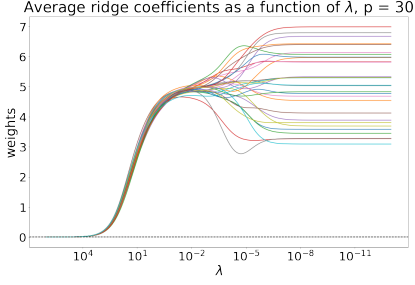
Similar to OLS, ridge minimizes the sum of squared residuals (SSR), but includes an ℓ_2-norm shrinkage penalty with a tuning parameter (λ):

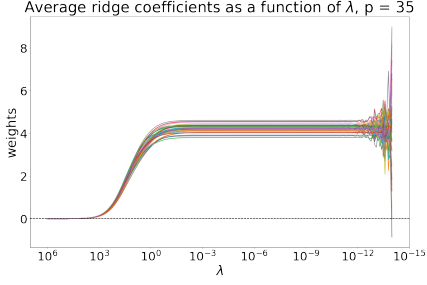
$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{SSR}}+\underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{shrinkage penalty}}. \tag{1}$$

- If $\lambda = 0$, we are left with the OLS estimates.
- As λ increases, the estimated ridge coefficients shrink towards zero.


Average ridge coefficients as a function of λ, p = 2
weights
λ


Average ridge coefficients as a function of λ, p = 28
weights
λ


Average ridge coefficients as a function of λ, p = 30
weights
λ


Average ridge coefficients as a function of λ, p = 35
weights
λ

Ridge Regression

Alternatively,

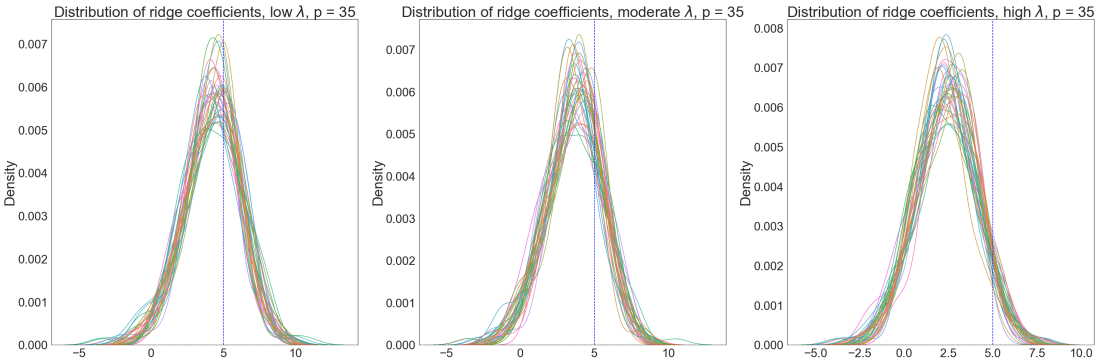
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{2}$$

where decreasing values of t indicate an increasingly restrictive optimization constraint. Ridge regression yields a closed-form solution for β_{ridge}:

$$\hat{\beta}_{\lambda}^{R} = \mathbf{Z}_{\lambda} \hat{\beta} = \left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\mathbf{X}'\mathbf{X} \right) \hat{\beta} \tag{3}$$

and for $var(\beta_{ridge})$:

$$V\left(\hat{\beta}_{\lambda}^{R} \right) = \sigma^2 \left[\mathbf{X}'\mathbf{X} + \lambda \mathbf{I} \right]^{-1} \left(\mathbf{X}'\mathbf{X} \right) \left[\mathbf{X}'\mathbf{X} + \lambda \mathbf{I} \right]^{-1}. \tag{4}$$

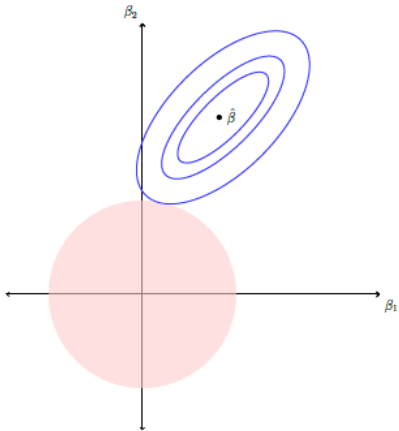

Distribution of ridge coefficients, low λ, p = 35
Distribution of ridge coefficients, moderate λ, p = 35
Distribution of ridge coefficients, high λ, p = 35
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
−5
0
5
10
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Lasso Regression

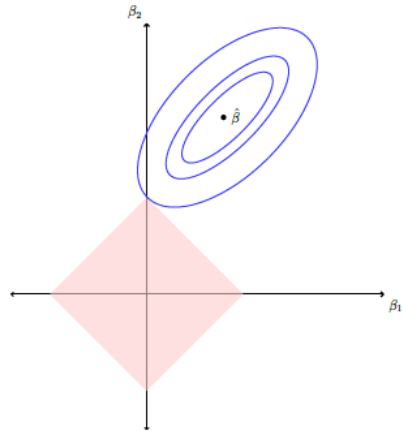
Consider the lasso minimization problem, which uses an ℓ_1-norm constraint:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \left| \beta_j \right| \leq t, \tag{5}$$

- where the $\hat{\beta}_\lambda^L$ solution is *sparce*: the lasso model holds onto relevant (non-zero) coefficients and sets irrelevant coefficients to **zero**
- Unlike ridge, lasso is a quadratic programming problem that does not have a closed-form solution.



(a) Ridge constraint region, $\beta_1^2 + \beta_2^2 \leq t$



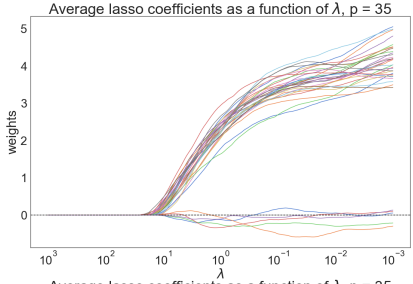
(b) Lasso constraint region, $|\beta_1| + |\beta_2| \leq t$

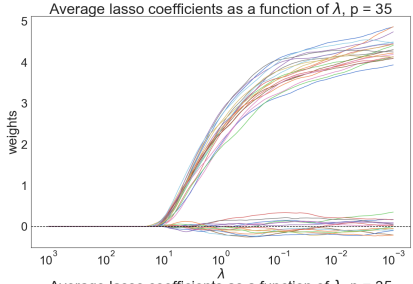
Figure 1: Contours of the SSR (in blue) and the constraint region (in red).

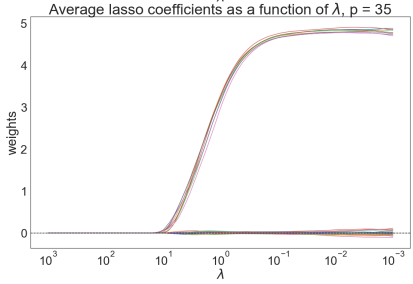
Lasso Regression

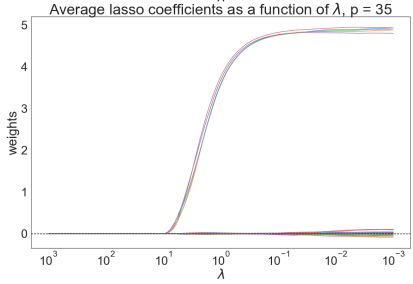
Limitations:

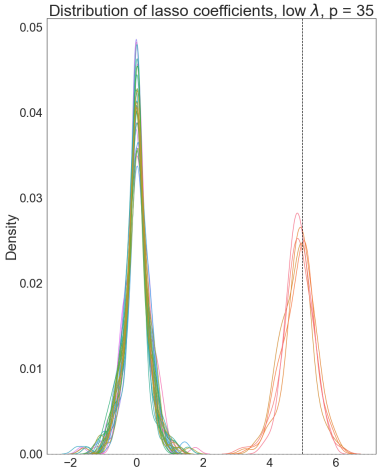
- Degree of sparsity
  - Lasso does poorly when a large subset of regressors are $\{j : \beta_j \neq 0\}$ of $\{1, ..., p\}$
- With high pairwise correlation among predictors, ridge regression outperforms lasso in terms of prediction.
- Lasso can only select up to n regressors when $p > n$.

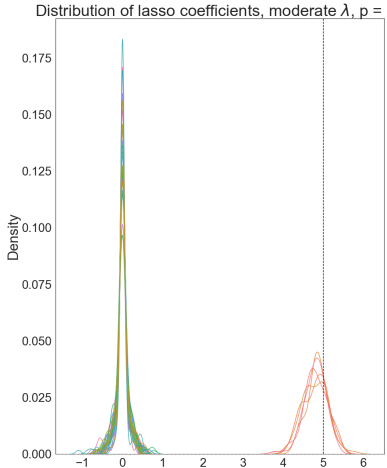

Average lasso coefficients as a function of λ, p = 35
weights
λ

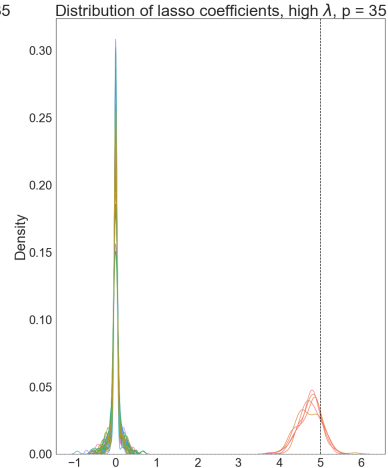

Average lasso coefficients as a function of λ, p = 35
weights
λ


Average lasso coefficients as a function of λ, p = 35
weights
λ


Average lasso coefficients as a function of λ, p = 35
weights
λ


Distribution of lasso coefficients, low λ, p = 35
Density
0.05
0.04
0.03
0.02
0.01
0.00
−2
0
2
4
6


Distribution of lasso coefficients, moderate λ, p = 35
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−1
0
1
2
3
4
5
6


Distribution of lasso coefficients, high λ, p = 35
Density
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−1
0
1
2
3
4
5
6

"Naïve" Elastic Net Regression

Consider the "naïve" elastic net minimization problem, which is subject to a constraint that is represented as a convex combination of the ridge and lasso shrinkage penalties:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t \qquad (6)$$

where α represents the weight assigned to the ridge penalty.

- Contains the best features of ridge and lasso regression (continuous shrinkage and simultaneous variable selection).
  - For all $\alpha \in [0, 1)$, the elastic net penalty function is singular at 0 and strictly convex for all $\alpha > 0$.
  - Strict convexity is what sets elastic net apart from lasso.

"Naïve" Elastic Net Regression

- **Improves over the lasso by**
  - overcoming the saturation of variable selection when $p > n$. **(variable selection)**
  - accounting for highly correlated "grouped" variables. **(continuous shrinkage)**
  - strengthening prediction performance if substantial multicollinearity among predictors exists in cases where $n > p$.
- **Limitations**
  - Double shrinkage does not reduce variance and unnecessarily introduces added bias.
  - Hence, works only when its solution is very close to either ridge or lasso.

Elastic Net Regression

Given the data set (y,X) and the two fixed non-zero Lagrangian parameters, (λ_1,λ_2), derived from the naïve elastic net's optimization problem, *Lemma 1* of Zou et al. 2005 defines the artificial data set ($\mathbf{y}^*$, $\mathbf{X}^*$):

$$\hat{\beta}^* = \arg\min_{\hat{\beta}^*} |\mathbf{y}^* - \mathbf{X}^*\beta^*|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}}|\beta^*|_1 \tag{7}$$

With the corrected elastic net estimates $\hat{\beta}$ defined as

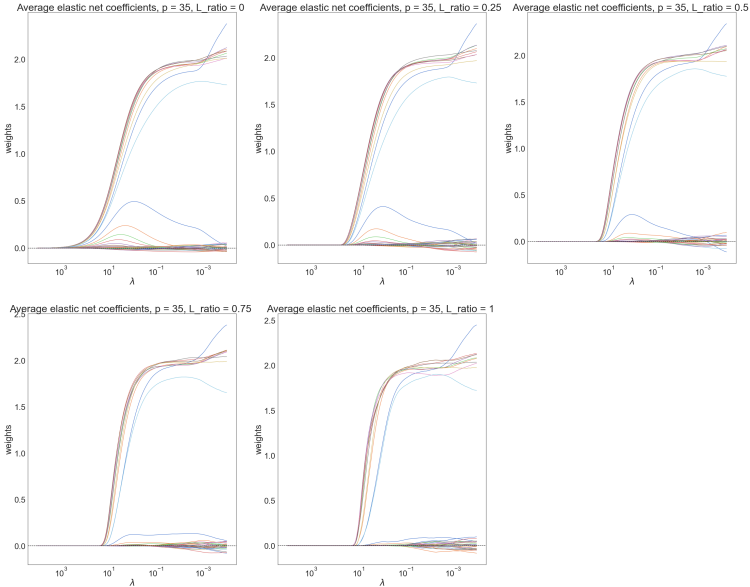
$$\hat{\beta}(\textit{elastic net}) = \sqrt{(1+\lambda_2)}\hat{\beta}^* \tag{8}$$

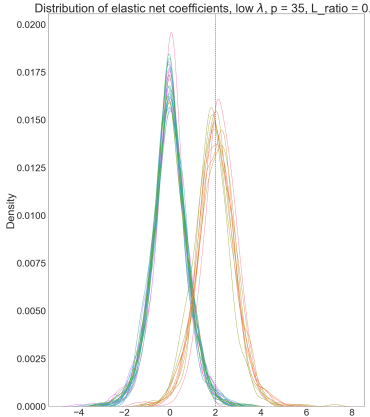
and

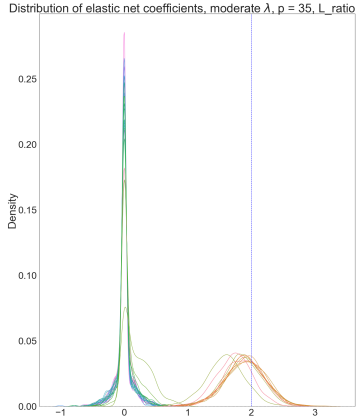
$$\hat{\beta}(\textit{naive elastic net}) = 1/\sqrt{(1+\lambda_2)}\hat{\beta}^* \tag{9}$$

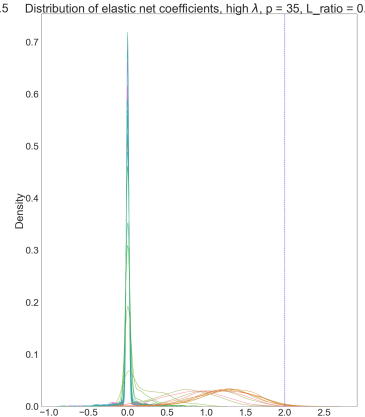
Finally,

$$\hat{\beta}(\textit{elastic net}) = (1+\lambda_2)\hat{\beta}(\textit{naive elastic net}). \tag{10}$$


Average elastic net coefficients, p = 35, L_ratio = 0
Average elastic net coefficients, p = 35, L_ratio = 0.25
Average elastic net coefficients, p = 35, L_ratio = 0.5
Average elastic net coefficients, p = 35, L_ratio = 0.75
Average elastic net coefficients, p = 35, L_ratio = 1
weights
λ


Distribution of elastic net coefficients, low λ, p = 35, L_ratio = 0.5
Density
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−4
−2
0
2
4
6
8


Distribution of elastic net coefficients, moderate λ, p = 35, L_ratio = 0.5
Density
0.25
0.20
0.15
0.10
0.05
0.00
−1
0
1
2
3


Distribution of elastic net coefficients, high λ, p = 35, L_ratio = 0.5
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

Regularization Methods Summarized

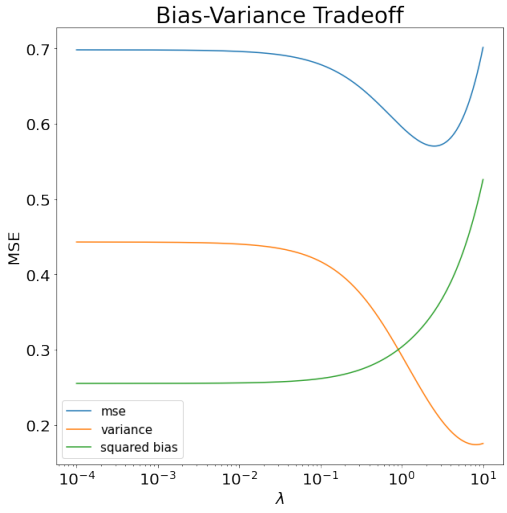
Model	Characteristics	Drawbacks
Ridge	• ℓ_2-norm shrinkage penalty • Performs well in the presence of multicollinearity.	• Does not perform variable selection.
Lasso	• ℓ_1-norm shrinkage penalty • Performs variable selection.	• Does not behave well when sparsity is very low. • Cannot handle high correlation or grouped multicollinearity. • Can only select up to n regressors when $p > n$
Naïve Elastic Net	• convex combination of ℓ_1-norm and ℓ_2-norm shrinkage penalties. • Potentially selects all p in the presence of grouped collinearity.	• Double shrinkage does not contribute to further reduction of the variances and adds unnecessary bias. • Works only when solution is very close to ridge or lasso.
Elastic Net	• Maintains all characteristics from naïve model, but corrects for double shrinkage.	

Measuring Prediction Performance

- Selects the model that best fits the data.
- Bias-Variance Decomposition
  - We carefully follow Hastie et al. 2008.
  - Assume $Y = f(X) + \varepsilon$, where $E[\varepsilon] = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$
  - We can derive the Mean Squared Error (MSE) out of the expected prediction error of $\hat{f}(X)$ at a **fixed** point $X = x_0$

$$
\begin{aligned}
Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] && (11)\\
&= \sigma_\varepsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 && (12)\\
&= \sigma_\varepsilon^2 + \ Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) && (13)\\
&= Irreducible\ error + \underbrace{Bias^2 + Variance}_{\text{MSE}} && (14)
\end{aligned}
$$

MSE as a Measurement of Model Fit



DGP setup

$$y = X\beta + \varepsilon, \quad \text{with} \quad \varepsilon \sim \mathcal{N}_n(0, 1)$$

$X_{nxp} \sim \mathcal{N}(0, 1)$ such that the pairwise correlation between x_i and x_j is expressed as

$$corr(i,j) = \text{correlation factor}^{|i-j|},$$

where the correlation factor takes on a value between 0 and 1.

Minimum Test MSE (Case 1)

Set up

- High dimensionality with varying sparsity, no multicollinearity.
- $n = 30$
- $p = 35$
- All non-zero $\beta = 2$.

Model	High Sparsity (10 non-zero β's)	Med Sparsity (20 non-zero β's)	Low Sparsity (35 non-zero β's)
Ridge	12.544	23.895	36.074
Elnet (Naive), 0.2	10.182	21.360	40.366
Elnet (Naive), 0.5	8.760	21.687	46.961
Elnet (Naive), 0.7	7.421	22.380	52.046
Lasso	5.903	24.090	60.170

Minimum Test MSE (Case 2)

Set up

- Low dimensionality with varying sparsity, moderate to high pairwise correlation (corr. factor = 0.8).
- $n = 30$, $p = 10$
- All non-zero $\beta = 2$.

Model	High Sparsity (3 non-zero β's)	Med Sparsity (7 non-zero β's)	Low Sparsity (10 non-zero β's)
Ridge	0.570	0.270	1.140
Elnet (Naive), 0.2	0.572	0.265	1.149
Elnet (Naive), 0.5	0.577	0.246	1.164
Elnet (Naive), 0.7	0.552	0.218	1.166
Lasso	0.607	0.298	1.166

Minimum Test MSE (Case 3)

Set up

- Low dimensionality with high sparsity, varying degrees of pairwise correlation.
- $n = 20$
- $p = 8$
- $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$

Model	Low Pairwise Corr. (corr. factor = 0.1)	Med Pairwise Corr. (corr. factor = 0.3)	High Pairwise Corr. (corr. factor = 0.7)
Ridge	1.238	2.071	0.905
Elnet (Naive), 0.2	1.160	2.034	0.889
Elnet (Naive), 0.5	0.979	1.897	0.864
Elnet (Naive), 0.7	0.828	1.724	0.851
Lasso	0.600	1.340	0.829

Minimum Test MSE (Case 4)

Set up

- Low dimensionality with no sparsity, varying degrees of pairwise correlation.
- $n = 20$
- $p = 8$
- All $\beta = 0.85$

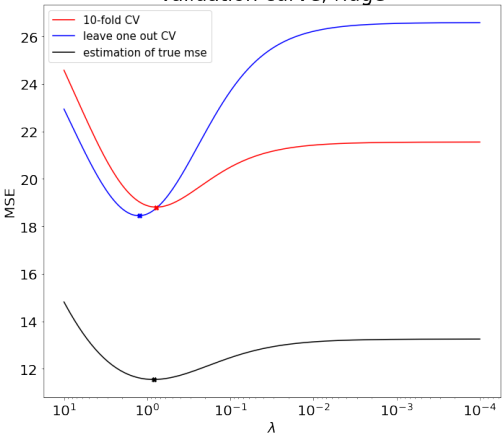
Model	Low Pairwise Corr. (corr. factor = 0.1)	Med Pairwise Corr. (corr. factor = 0.3)	High Pairwise Corr. (corr. factor = 0.7)
Ridge	2.471	0.521	0.736
Elnet (Naive), 0.2	2.506	0.585	0.772
Elnet (Naive), 0.5	2.542	0.882	0.831
Elnet (Naive), 0.7	2.548	1.124	0.867
Lasso	2.548	1.532	0.906

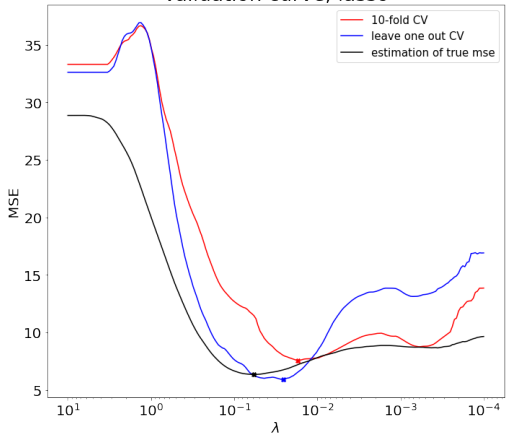
Minimum Test MSE (Case 5)

Set up

- High dimensionality with high sparsity, varying degrees of pairwise correlation.
- $n = 30$
- $p = 35$
- 10 non-zero $\beta = 2$

Model	Low Pairwise Corr. (corr. factor = 0.1)	Med Pairwise Corr. (corr. factor = 0.3)	High Pairwise Corr. (corr. factor = 0.7)
Ridge	9.412	15.714	3.263
Elnet (Naive), 0.2	7.932	11.092	2.659
Elnet (Naive), 0.5	6.049	10.505	1.882
Elnet (Naive), 0.7	5.122	8.570	1.589
Lasso	4.079	6.575	1.716


Validation curve, ridge
10-fold CV
leave one out CV
estimation of true mse
MSE
λ

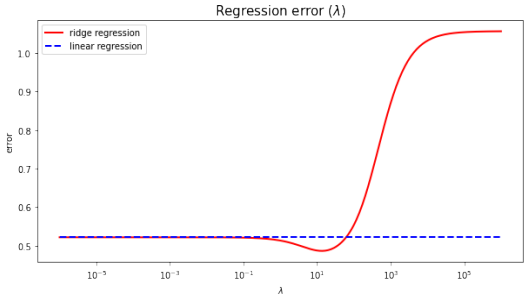

Validation curve, lasso
10-fold CV
leave one out CV
estimation of true mse
MSE
λ

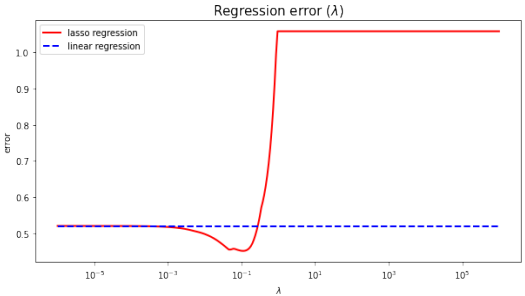
Prostate Cancer Data Set

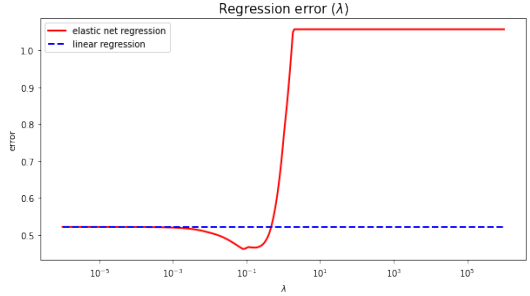
- What is the relationship between log prostate specific antigen (lpsa) levels, which is elevated in men with prostate cancer, and other clinical measures ($p = 8$)
- $n = 97$
- For comparability we use the same train-test split as Zou et al. 2005
- High correlation between
  - lcp *(log of capsular penetration)* and lcavol *(log cancer volume)*
  - svi *(seminal vesicle invasion)* and lcp
  - pgg45 *(percent of Gleason score 4 or 5)* and gleason *(Gleason score)*
  - pgg45 and lcp

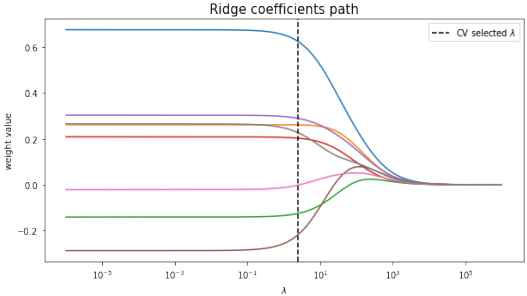
Correlation Matrix

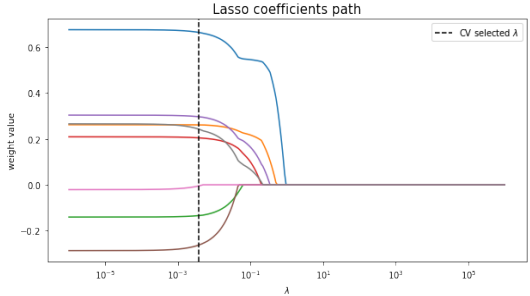
	lpsa	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lpsa	1.000000	0.734460	0.433319	0.169593	0.179809	0.566218	0.548813	0.368987	0.422316
lcavol	0.734460	1.000000	0.280521	0.225000	0.027350	0.538845	0.675310	0.432417	0.433652
lweight	0.433319	0.280521	1.000000	0.347969	0.442264	0.155385	0.164537	0.056882	0.107354
age	0.169593	0.225000	0.347969	1.000000	0.350186	0.117658	0.127668	0.268892	0.276112
lbph	0.179809	0.027350	0.442264	0.350186	1.000000	-0.085843	-0.006999	0.077820	0.078460
svi	0.566218	0.538845	0.155385	0.117658	-0.085843	1.000000	0.673111	0.320412	0.457648
lcp	0.548813	0.675310	0.164537	0.127668	-0.006999	0.673111	1.000000	0.514830	0.631528
gleason	0.368987	0.432417	0.056882	0.268892	0.077820	0.320412	0.514830	1.000000	0.751905
pgg45	0.422316	0.433652	0.107354	0.276112	0.078460	0.457648	0.631528	0.751905	1.000000

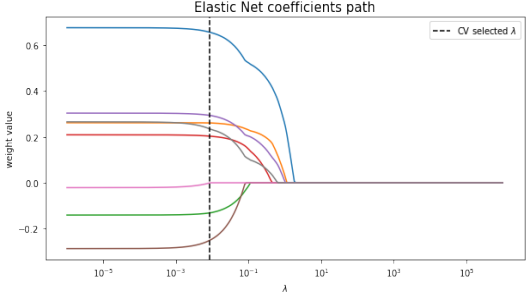

Regression error (λ)
ridge regression
linear regression
error
λ


Regression error (λ)
lasso regression
linear regression
error
λ


Regression error (λ)
elastic net regression
linear regression
error
λ


Ridge coefficients path
CV selected λ
weight value
λ


Lasso coefficients path
CV selected λ
weight value
λ


Elastic Net coefficients path
CV selected λ
weight value
λ

Model Selection

Model	Test MSE (with 10-fold CV)	Test MSE (AIC)	Variable Selection
Linear	0.5213	-	All
Ridge	0.5043	-	All
Lasso	0.5112	-	All
Elastic Net (Naive)	0.5043	-	All
LARS	0.5084	0.5033	lcavol, lweight, lbph, svi, pgg45

Takeaway

- Naive elastic net performs as well as ridge, also reported by Zou et al. 2005
- Lasso performs best if LARS algorithm is implemented and AIC is used for tuning λ. It performs variable selection as in Zou et al. 2005.

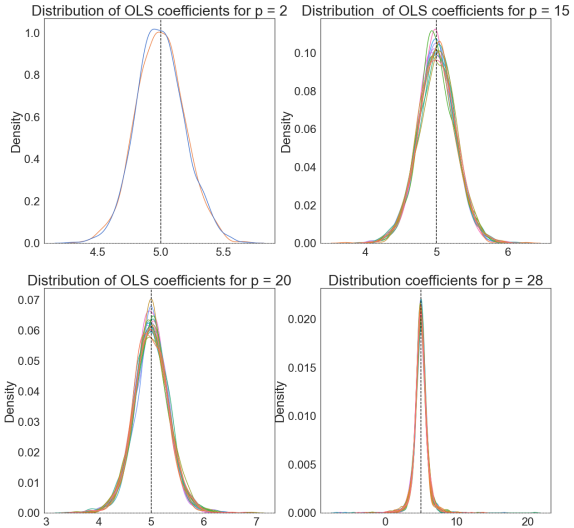
Summary: best model based on simulations

		Regressors' correlation	
		High	**Low**
Coefficients' sparsity	**High**	Elastic Net (high-dimensionality) Lasso (low-dimensionality)	Lasso
	Low	Ridge	Ridge (high-dimensionality) OLS (low-dimensionality)

- The data application is in line with *Case 2* of our simulation study.

Thank you for your attention!

Appendix


Distribution of OLS coefficients for p = 2
Distribution of OLS coefficients for p = 15
Distribution of OLS coefficients for p = 20
Distribution coefficients for p = 28
Density

Appendix

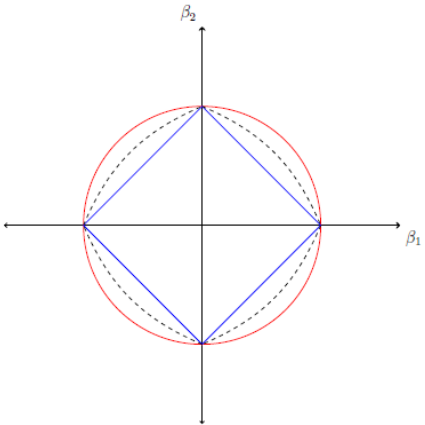
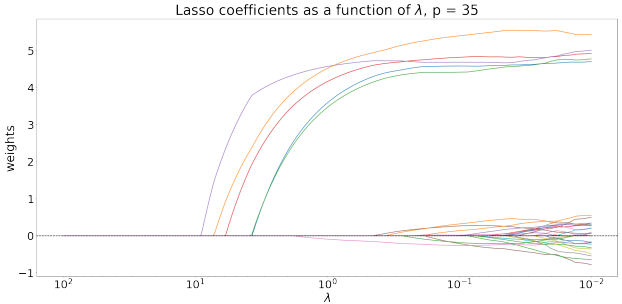


Figure 2: constraint region for Ridge (red) L2-norm, Lasso (blue) L1-norm, Elastic Net (dashed black)

Appendix


Lasso coefficients as a function of λ, p = 35
weights
λ
Lasso coefficients as a function of λ, p = 35
weights

Appendix

- We used a *10-fold cross validation* approach to find the optimal value of the tuning parameter.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{15}$$

- We identify the model with the lowest test error.
- We identify the λ value that minimizes the MSE and then plug it back into the model of interest.

Appendix

Least-angle regression (LARS) (Efron et al. 2003)

- Provides an extremely efficient algorithm for computing the entire lasso path.
- Lasso is a variant of the LARS procedure.
- LARS algorithm:
  1. Standardize predictors.
  2. Find predictor x_j that is most correlated with y.
  3. Move β_j from 0 towards least-squares coefficient until x_k has as much correlation with the current residual as does x_j.
  4. Move β_j and β_k in the direction of the joint least squares coefficient of the current residual on (x_j, x_k), until x_l has as much correlation with the current residual.
  5. Continue until all p have been entered.
  6. *If a non-zero coefficient hits zero, drop the corresponding variable from the active set of variables and recompute the current joint least squares direction.*

References I

Aheto, J. M. K. et al. (2021). “A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare?” In: *Preventive Medicine Reports* 23, p. 101475.

Efron, B. et al. (2003). “Least Angle Regression”. In: *The Annals of Statistics* 32.2, pp. 407–451.

Hastie, T., R. Tibshirani, and J. Friedman (2008). *Elements of Statistical Learning*. Springer Series in Statistics.

Hastie, T., R. Tibshirani, and M. Wainwright (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hoerl, A. E. and R. W. Kennard (1970a). “Ridge regression: applications to nonorthogonal problems”. In: *Technometrics* 12.1, pp. 69–82.

— (1970b). “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1, pp. 55–67.

James, G. et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Kidwell, J. S. and L. H. Brown (1982). “Ridge regression as a technique for analyzing models with multicollinearity”. In: *Journal of Marriage and the Family*, pp. 287–299.

Marquardt, D. W. and R. D. Snee (1975). “Ridge regression in practice”. In: *The American Statistician* 29.1, pp. 3–20.

Melkumova, L. and S. Y. Shatskikh (2017). “Comparing Ridge and LASSO estimators for data analysis”. In: *Procedia engineering* 201, pp. 746–755.

Snee, R. D. (1973). “Some aspects of nonorthogonal data analysis: Part I. Developing prediction equations”. In: *Journal of Quality Technology* 5.2, pp. 67–79.

Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

References II

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.