4 L_1, L_2 and Elastic Net Regularizations

Given the underlying asset pricing theory explained above, linear models hold considerable relevance in the context of empirical asset pricing. I use them not only for theoretical reasons but also for practical applications. In fact, linear models are Taylor expansions of non linear parametric models. This is useful in a setting as complex as return forecasting: By acting as the most lenient approximation in a highly intricate setting, linear models emerge as optimal amidst unknown models. In the context of high dimensional factor modeling however, standard linear models are not adequate. Therefore, in this section, I introduce regularization methods tailored for linear models, allowing their application in high-dimensional settings. Regularization serves two main purposes: Preventing over fitting, and solving ill-posed problems. Both issues are fundamental in Returns forecasting given that we are searching for the model that generalizes best to unseen data using limited data [25] . I will tackle both approaches to regularization and propose three regularization methods Ridge, Lasso and Elastic Net to remedy the problem.

4.1 Regularization and Overfitting

As explained, finding the best predictive model is equivalent to finding the hypothesis space (or strategy) that generalizes best out of sample. Regularization introduces the idea of constraining the hypothesis space. Essentially, it defines a complexity measure $\Omega : F \longrightarrow [0, \infty)$ and restricts the hypothesis space to some level of complexity defined by $\Omega(F)$. Formally, For some hypothesis space F And a complexity measure $\Omega(F)$, we reduce the hypothesis place to a subset of hypothesis $F = \{f \in F \mid \Omega(F) \leqslant r\}$ For some level r of complexity. Complexity measures are defined as L_p Norms, such that

$$\begin{cases} l_0 \text{ complexity: } \Omega(f) = \text{ \# of Non zero coefficients} \\ l_1 \text{ 'Lasso' complexity: } \Omega(f) = \sum_{i=1}^{p} |w_i| \\ \text{Squared } l_2 \text{ 'Ridge' complexity: } \Omega(f) = \sum_{i=1}^{p} {w_i}^2 \end{cases}$$

Complexity, given a linear regression, is defined as the degrees of freedom of parameters. The more free to vary are the parameters, the more complex is the model and vice versa. Hence, constraining the hypothesis space by some complexity measure reduces the models complexity. Having defined the framework, regularization is defined as an Empirical Risk Minimization problem restricted by a subset of model's hypothesis space. Ivanov regularization (IR) Formalizes clearly this idea: For some complexity measure defined as an L_p norm and a tuning parameter $r > 0$ defining the complexity level. We define Ivanov regularization as : $\min_{f \in F} \left[\frac{1}{N} \sum_{i=1}^{N} l\left(y_i, f\left(x_i\right)\right)\right]$ such that $\Omega(F) \leq r$. By considering r as a hyper parameter, IR incorporates the models complexity to the minimization prob-

[25] Limited data will restrict our problem to an ill-posed one

lem: Not only empirical risk is minimized, but complexity is also adjusted so that model generalizes best on some validation set. A more common regularization is the Tikonov regularization (TR) : For some complexity $\Omega(F)$ defined by an L_p norm and a tuning parameter $\lambda \geq 0$, TR is defined as : $\operatorname{Min}_{f \in F} \frac{1}{N} \sum_{i=1}^{N} l\left(y_i, f\left(x_i\right)\right)+\lambda \Omega(f)$. Tikhonov regularization is equivalent to Ivanov regularization for certain loss functions and for certain complexity measures. Importantly, equivalence holds for both Ridge and Lasso regressions(Oneto et al. [20]).IR gives a clear idea of what's happening under the hood, while, TR is an easier to solve, unconstrained, minimization problem.

In summary, Ridge and Lasso regressions represented by Tikhonov regularization, express the idea clearly formalized by Ivanov regularization: That is, finding the best empirical risk minimized model by cutting down the hypothesis space from $\mathcal{F}$ to $\mathcal{F}_r$; i.e. by reducing the complexity level. We now have a clear picture of how regularization controls overfitting:

Recall the approximation error $r\left(s^*\right)-r_0$ which is inversely proportional to changes in complexity; the estimation error $r(\hat{s})-r\left(s^*\right)$ which, along the generalization error $\|\hat{r}(\hat{s})-r(\hat{s})\|_p$ have positively proportional bounds with respect to complexity; by cutting down the hypothesis space and hence by controlling model's complexity, regularization "controls" overfitting by trading off generalization and estimation error for approximation error (Overfitting can thus never be completely eliminated) [26]

In forecasting returns and in predictive models in general, overfitting constitutes a fundamental issue; hence, regularization appears as a natural alternative to Simple ordinary least squares Factor modeling used traditionally.

4.2 Regularization and Ill-posedness

In addition to over fitting, ill-posedness is a customary issue in Factor modeling, i.e we almost always need to solve a problem where either the solution does not exist or the number of solutions is infinite. In fact, factor modelling is fundamentally an $Ax = b$ problem whereby A is generally not well determined. [27] Ill-posed systems are either due to over determined systems, those are typically the case in classical Factor modeling with P factors and N Returns such that P¡¡N; or due to under-determined systems: Those are typical in high dimensional machine learning applications whereby the number of features surpasses the number of observations; P¿¿N. In addition, Ill-posedness may also be the result of (Multi)-Collinearity or Near-(Multi)collinearity in the design matrix when two or more columns (or rows) vectors are linearly dependent (or nearly linearly dependent). Note: an

[26] Along this conclusion we also deduce from statistical learning theory that as the number of observations N increase we want less and less regularization. In fact, regularization increases approximation error and decreases estimation error; the speed at which approximation errors changes is independent of $N \Rightarrow O(1)$ while the estimation and generalization error bound depends on N (Assuming Rademacher complexity), we find that the bound $\leq O\left(\sqrt{\frac{\lg n}{n}}\right)$ Thus, one should regularize less as N increases.

[27] Note, in the context of factor modeling A represents the design Matrix - or 'the factor Matrix- , b the returns vector and X the coefficients.

underdetermined system implies multicollinearity in the design matrix [28] ; however, Multicollinearity does not strictly reduce to an underdetermined system it also may result in an overdetermined one [29]. Importantly: Multicollinearity implies ill-posedness of the system (See appendix for more information about multicollinearity evaluation measures).

Classical asset pricing is mainly concerned with the issue of over determined systems, this paper however deals with both over and under determined scenarios as high dimensional Factor models are explored.

Typically, from a classical linear algebra point of view, ill-posed problems are solved by finding the x that minimizes$||Ax-b||_2$. Laub (2005) [12] provides the general solution to this least squares problem as such: For $A \in \mathbb{R}^{N\times P}\&b \in \mathbb{R}^{N\times K}$, the solution to the least squares problem is:

$$X_{LS} = A^{+}b + \left(I - A^{+}A\right)Y \tag{3}$$

for some arbitrary vector $Y \in \mathbb{R}^{\mathbb{P}\times\mathbb{K}}$ and with A^{+} the Moore-Penrose Pseudo Inverse of A. Hence, we can always find a solution x_{LS} For any $Ax = b$ problem as a function of the pseudo inverse. The pseudo inverse is a generalization of the two-sided inverse [30]that applies on any Matrix whether singular or rectangular or Multicollinear etc... It is not exactly an inverse but its properties resemble that of an inverse [31] and is defined as $A^{+} = V \begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^{\top}$ - which is obtained by exploiting the Singular Value decomposition of A [32] . By employing the pseudo inverse one can always find a unique (sometimes approximate) solution for ill-posed problems: In fact, using 3 we will either get a unique solution. Or infinitely many solutions and accordingly choose the minimum norm solution. For instance for some under determined system $Ax = b$ we get $X_{LS} = A^{+}b + (I - A^{+}A)\,Y$ infinitely many solutions, we can chose however a unique approximate solution: the minimum-norm solution by choosing $X_{LS} = A^{+}b$ such that $Y = 0$ [33] In this sense, by utilizing the Pseudo inverse, one can always get, an approximate unique solution no matter how ill-posed the system is: This is the standard

[28] If A $(N \times P)$ has $P >> N$, this implies multicollinearity by the rank nullity theorem. See Multicollinearity section in appendix for proof

[29] $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ is an example of multicollinearity that reduces to an underdetermined system and $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ is an example of multicollinearity that reduces to an overdetermined system

[30] The two-sided inverse, is actually what we refer as the "inverse" in general. We define the 2-sided inverse of some matrix A as $G = A^{-1} : GA = AG = I$

[31] For some matrix A, those are, for instance, some of the main "inverse-like" properties of the Pseudo Inverse G=A^{+}:
$\begin{cases} AGA = A \\ GAG = G \\ (AG)^{T} = AG \\ (GA)^{T} = GA \end{cases}$ ......

[32] There are also closed forms for A^{+}.For example,for A having full row rank $A^{+} = A^{\top}\left(AA^{\top}\right)^{-1}$ and for A having full column rank $A^{\dagger} = \left(AA^{\top}\right)^{-1} A^{\top}$)

[33] See Least Squares Solutionappendix for graphical representation

approach to ill-posedness in linear algebra.

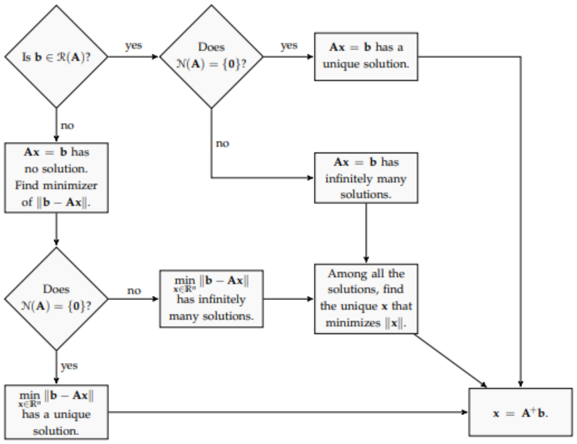
Is $\mathbf{b} \in \mathcal{R}(\mathbf{A})$?
yes
Does $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$?
yes
$\mathbf{Ax} = \mathbf{b}$ has a unique solution.
no
$\mathbf{Ax} = \mathbf{b}$ has no solution. Find minimizer of $\|\mathbf{b} - \mathbf{Ax}\|$.
no
$\mathbf{Ax} = \mathbf{b}$ has infinitely many solutions.
Does $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$?
no
$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{Ax}\|$ has infinitely many solutions.
Among all the solutions, find the unique $\mathbf{x}$ that minimizes $\|\mathbf{x}\|$.
yes
$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{Ax}\|$ has a unique solution.
$\mathbf{x} = \mathbf{A}^{+}\mathbf{b}$.

Figure 2: This Chart shows how using A^+ can always result in a unique (approximate) solution. This representation builds upon Strang's four fundamental subspaces of linear algebra to represent different reactions to ill-posedness.

Regularization as an alternative :

This "standard" method for solving ill-posed systems is ubiquitous [34] ; it however suffers from two major drawbacks. From the one hand, using the A^+ to compute the solution, does not directly address the issue that is causing ill-posedness (Multicollinearity for instance); but rather, solves the problem in a generic manner. Second, and most importantly, computing the Pseudo inverse can be computationally expensive for large matrices, as it requires computing the eigenvectors/eigenvalues of the matrix and the reciprocals of its singular values. For a large and sparse design matrix A, as it is in this paper, this may be practically very challenging. Hence, for interpretabilty and computational reasons, one ought to find alternatives to Pseudo-inverse; Regularization emerges here as a natural solution.

Instead of exploiting the singular value decomposition of A, Tikhonov regularization solves ill-posed problems numerically by controlling the norm of $||x||$ while minimizing $||Ax - b||_2$;and is formally defined as $\text{ArgMin}_x \, ||Ax - b||_2 + \lambda|||x||_p$. This method has also its limitations (which I discuss in the end of this section) but it is much less computationally expensive and more natural than the pseudo inverse solution. I will present three different expressions of the Tikhonov regularization and explain how I will use them in my paper:

First, the Ridge Regression (RR) is expressed as a Tikhonov regulatization with a squared l_2 normed penalty. Explicitly, Ridge is defined as: $\text{ArgMin}_x \, ||Ax - b||_2 + \lambda|||x||_2$, with λ a tuning parameter > 0

[34] Notably, many programming libraries, do incorporate this logic into their estimation algorithm for linear models. Hence, when fitting a linear model - with some ill-posedness resulting in under determined systems (for instance with a mutlicollinear matrix A , or using a design matrix with more features than observations...) we can still get a result. This solution is actually the approximate solution found by using the Moore Penrose pseudoinverse and picking the least norm solution.

,which translates in our linear regression framework into :

$$\hat{\beta}_{\text{ridge }}=\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_{i}-\beta_{0}-\sum_{j=1}^{p} x_{i j} \beta_{j}\right)^{2}+\lambda \sum_{j=1}^{p} \beta_{j}^{2}\right\}$$

We have already seen in the previous section that - since Ivanov and Tikhonov regularizations are equivalent under ridge regression - we can express Ridge in a more "intuitive and interpretable form" :

$$\begin{gathered}\hat{\beta}_{\text{ridge }}=\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_{i}-\beta_{0}-\sum_{j=1}^{p} x_{i j} \beta_{j}\right)^{2} \\ \text{subject to } \sum_{j=1}^{p} \beta_{j}^{2} \leq r\end{gathered}$$

We can now clearly see how Ridge regularization essentially constrains the coefficients to some squared norm. Ridge regularization requires that the predictors are standardized in order to prevent an unfair shrinking between them. In fact, Ridge constrains the size of the coefficients using a squared l_2 norm: s.t. $\beta_1^2+\ldots+\beta_p^2<r$; if one (or more) of these variables is scaled differently than the others, this will be reflected in the magnitude of βs and induce unfair penalisation; we mitigate this problem by standardizing all factors. Consequently, the intercept is not penalized by ridge [35]. The estimation is a two steps procedure : After standardizing the predictors, one should first set the intercept as $\bar{y}$ then estimate the other coefficients by ridge penalization. This regularization method has a closed form [36] : $\hat{\beta}_R=\left(\mathbf{X}^{\top} \mathbf{X}+\lambda \mathbf{I}\right)^{-1} \mathbf{X}^{\top} \mathbf{y}$. We can clearly see from the closed form solution how ridge mitigates ill posedness by adding to the Gramian matrix a diagonal matrix. For ill-posed systems due to multicollinearity, the Gramian matrix is singular, adding to it a diagonal element transforms it to an invertible matrix (Non-singular).

As explained Ridge solves ill-posedness, accordingly, one can prove algebraically, using the Singular value decomposition of X, that the Ridge regression is nothing more than some scaled version of the pseudo inverse solution which I introduced previously. Concretely, Setting aside algebraic manipulations [37] , ridge regression estimators can be written as $\hat{\beta}_{ridge}=\mathbf{V}\left(\mathbf{\Sigma}^{2}+\lambda \mathbf{I}_{n}\right)^{-1} \mathbf{\Sigma} \mathbf{U}^{\top} \mathbf{Y}$; while, the "pseudoinverse" solution in the context of a linear regression is $\beta_{\text{pseudoinverse}}=X^{+} Y=V \Sigma^{-1} U^{T} Y=$

[35] For some $\beta_0+\sum \beta_j X_j$ regression model with standardized (more precisely centered) predictors we get $E[Y]=\beta_0$. Thus for estimation purposes, we do not penalize the intercept and we require it to be $\bar{Y}$

[36] The closed form solution of ridge regression is found by simply deriving with respect to β the objective function . This is feasible as the function is convex and involves a simple quadratic function

[37] Using Singular Value decomposition and simple algebraic and matrix manipulations; one can prove:

$\mathbf{V}\left(\Sigma^2\right)^{-1}\Sigma\mathbf{U}^\top\mathbf{Y}$ [38] . Comparing both formulae we clearly see that the ridge estimator is equivalent to a scaled pseudoinverse estimator by $\frac{\Sigma^2}{\Sigma^2+\lambda\mathbf{I}_n} \in [0,1]$. For $\lambda = 0$ both ridge and pseudoinverse coefficients are equivalent, and as the penalty term λ increases, the ridge coefficient converges to 0. In addition, expanding the last ridge estimator formula, $\hat{\beta}_{ridge} = \frac{\Sigma^2}{\Sigma^2+\lambda\mathbf{I}_n}\beta_{pseudoinv} = \left\{\frac{s_i^2}{s_i^2+\lambda}\right\}\beta_{pseudoinv}$ [39] one can infer the dynamics of the ridge regression:Ridge estimator shrinks as the squared singular values increase, and since the Since $\Sigma^2 = \left\{\frac{s_i^2}{s_i^2+\lambda}\right\}$ is the covariance matrix of the demeaned predictor X [40], and because its columns are indicative of the amount of variance within each principal component; Ridge regression is essentially, a regularization method that penalizes smoothly [41] low-variance Principal components directions.

The lasso regularization is a Tikhonov regularization method with a $l1$ normed penalization; which is defined as $\text{ArgMin}_x \, ||Ax - b||_2 + \lambda|||x||_1$. In the particular context of linear regression, lasso is formalized as : $\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\left\{\frac{1}{2}\sum_{i=1}^N\left(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^p|\beta_j|\right\}$. Or equivalently as an Ivanov regularization :

$$\begin{aligned}\hat{\beta}_{\text{lasso}} =& \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^N\left(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j\right)^2\\ &\text{subject to } \sum^p|\beta_j| \le t.\end{aligned}$$

The same rules of estimation apply on Lasso Regression. The estimation procedure is done in two steps : First we compute the intercept, then estimate the other coefficients. However, Lasso does not have a closed form solution, as the objective function is not differentiable. There are different numerical methods that solves this optimization problem, I propose, the accelerated proximal gradient descent method which I explain in details in the Numerical Methods chapter. Lasso Regression is widely used in asset pricing modeling primarily due to its inherent feature selection capability [42] making the model

$$\begin{aligned}\hat{\beta}_{ridge} &= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{Y}\\ &= \left(\mathbf{V\Sigma U}^\top\mathbf{U\Sigma V}^\top + \lambda\mathbf{I}_p\right)^{-1}\mathbf{V\Sigma U}^\top\mathbf{Y}\\ &= \left(\mathbf{V\Sigma}^2\mathbf{V}^\top + \lambda\mathbf{I}_p\right)^{-1}\mathbf{V\Sigma U}^\top\mathbf{Y}\\ &= \left(\mathbf{V\Sigma}^2\mathbf{V}^\top + \lambda\mathbf{VV}^\top\right)^{-1}\mathbf{V\Sigma U}^\top\mathbf{Y}\\ &= \mathbf{V}\left(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n\right)^{-1}\mathbf{V}^\top\mathbf{V\Sigma U}^\top\mathbf{Y}\\ &= \mathbf{V}\left(\Sigma^2 + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma U}^\top\mathbf{Y}\end{aligned}$$

[38] We have discussed the pseudoinverse previously from a pure linear algebra point of view using Ax=b. In the context of linear algebra, nothing changes, but instead of writing A we write X,our factor matrix, and instead of x we write β the coefficients matrix and finally instead of b we write Y. Thus $x = A^+b$ is written as $\beta = X^+y$ (Refer to Figure R1 - for more information)

[39] For s_i^2 the singular values the "stretching" Sigma matrix

[40] Using the SVD of X ,X^TX/N, the covariance matrix is equal to $\mathbf{VD}^2\mathbf{V}^T$ .Since, the covariance matrix is symmetric, then the V and D^2 matrices correspond to the covariance matrix's eigenvectors/values. This is an important result in Principal component analysis; whereby the eigenvector and eigenvalues of the covariance matrix of the demeaned X are indicatives of the PCs and their relevance; it is reviewed in the following chapters

[41] Penalty is smoothed by the tuning parameter λ

[42] Feature selection involves the elimination of certain features while retaining others

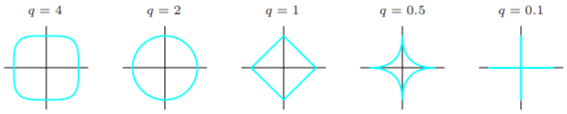

Figure 3: Illustration of different penalty contours $\sum_{j=1}^{p} |\beta_j|^q \leq 1$ for different values of q on a 2 dimensional parameter space. As q decreases the contour of the penalty promotes increasing sparsity in the coefficients.

more interpretable.

Specifically, Lasso Regression employs a least squares approach, effectively nullifying "irrelevant" coefficients[43]. One can interpret the sparsity of the result both algebraically and graphically. In fact, assuming a linear model with an l_1 penalization,the objective function of our problem is $\min y^T y - 2y^T x\hat{\beta} + \hat{\beta} x^T x\hat{\beta} + 2\lambda|\hat{\beta}|$. For $\hat{\beta} > 0$, the solution has a closed form $\hat{\beta} = \left(y^T x - \lambda\right) / \left(x^T x\right)$ which decreases as l_1 regularization tuning parameter λ increases until reaching $\hat{\beta} = 0$; and for $\hat{\beta} < 0$, the solution has also a closed form, $\hat{\beta} = (y^T x + \lambda)/(x^T x)$ which increases as the lasso regularization parameter increases, until $\hat{\beta} = 0$ is reached. We hence, ultimately expect sparsity when penalizing using lasso. This is contrary to the squared squared $l2$ normed ridge penalty which has the following closed form solution $\hat{\beta} = y^T x / \left(x^T x + \lambda\right)$ for both, $\beta <$ and > 0 ; the $\hat{\beta}$ here , however, does not decrease specifically to zero as regularization grows (Note: The coefficient does decrease, but not to zero as was the case for l_1 regularization). In addition, one can see that the sparsity of lasso coefficients arises from the distinctive shape of the l_1 penalty. This characteristic is a direct result of the contours of the l_1-normed penalty, where decreasing the degree of the l_p norm is associated with an increased expectation of sparsity.

By controlling the size of the coefficients, lasso regression does solve ill-posedness. However, it becomes unstable under perfect collinearity for two identical vectors in the design matrix. One can illustrate this result graphically in a two dimensional parameter space: Specifically, in the case of an ill-posed problem arising from multicollinearity, leading to an underdetermined system, the objective function is depicted as an "infinite" ellipsoid.

[43] Subtlety: In this context, "Irrelevance" does not imply "uninformative" or "statistically independent." Instead, it is defined within the framework of Lasso regression. Lasso regression may eliminate informative features if it perceives them as irrelevant to the least squares minimization objective.

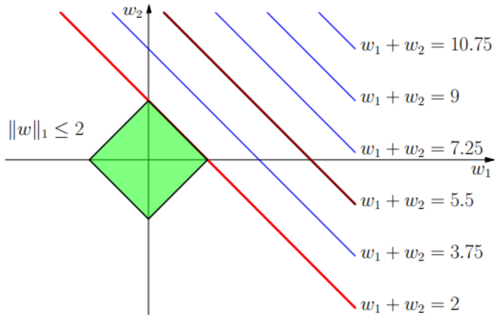

Figure 4: Mutlicollinearity might induce loss function's ellipsoid to degenerate wherein each point on the ellipsoid is now represented by a line

Parallel lines represent the loss function; which due to collinearity result in a "degenerate" ellipsoid (One can imagine a horizontally squished and infinitely elongated ellipsoid where each point is a now a line). Importantly this representation depicts instability under multicollinearity because the lasso penalty represented by a square , cannot solve for loss functions that are parallel to its contours (There are infinitely many solutions). This is coherent, since, for equivalent features, any variable selection method will arbitrarily select a matrix.

To resume, ridge regression shrinks the coefficients without performing feature selection while lasso regression results in variable selection, but may fail to provide stable solutions under particular cases of perfect multicollinearity. Hence, Elastic Net Regularization emerges as a middle ground solution between ridge and lasso. This method is a l_p normed regularization that combines both L_1 and L_2 penalties in an additive manner. Elastic Net regression is formally represented as: $\hat{\beta}_{\text{EN}} = \underset{\beta}{\operatorname{argmin}}\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left(\rho\beta_j^2+(1-\rho)\left|\beta_j\right|\right)$, with ρ a tuning parameter $\in [0,1]$. Intuitively, the dynamics of this method become apparent when conjecturing from the contour of the elastic net penalty in a two-dimensional parameter space representation (see Figure 5). These contours promote sparsity as they look like the l_1 squared penalty, yet, the slight curvature along their sides helps alleviate instability caused by multicollinearity, preventing the degenerate loss function from being parallel to its contours.

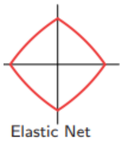

Figure 5: In a two dimensional coefficients space, the restricted Elastic Net penalty takes this form (Refer to Ivanov Regularization for an explanation of "restricted" penalties).

Thus from one side, by introducing absolute values, this penalization introduces sparsity and is a feature selection method; and from the other, the squared penalty assures the stability of the solution, even under perfect multicollinearity with equal features. Like for lasso, and for the same reasons, elastic net regularization has no closed form solutions. I use Accelerated Proximal Gradient Descent to estimate the coefficients; I discuss the details of this method in the ”Numerical Methods” chapter.