Das elastic net und Gruppierung korrelierter Prädiktoren

Bettina Wiebe*

8. Dezember 2014

1 Wiederholung

Im folgenden Kapitel, welches nicht vorgetragen wird, wird kurz an das vorliegende Modell sowie einige grundsätzlich als bekannt vorausgesetzte Schätzer und ihre Eigenschaften erinnert, welche im Weiteren häufiger auftreten.

1.1 Lineares Modell

Wir betrachten das folgende lineare Regressionsmodell

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

wobei $\mathbf{X} = (x_1|\ldots|x_p)$ die Designmatrix, $\mathbf{y} = (y_1, \ldots, y_n)^T$ der Responsevektor der Beobachtungen, $\beta = (\beta_1, \ldots, \beta_p)^T$ der Parametervektor und $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ der Störungsvektor sind. Dabei besteht die Designmatrix aus p Vorhersagevektoren (Kovariablenvektor) $x_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, die entweder deterministisch feste Werte oder Zufallsvariablen sind. Die Beobachtungen werden als unabhängig angenommen und die $\varepsilon_1, \ldots, \varepsilon_n$ sind unabhängig und identisch verteilt sowie unabhängig von der Menge $\{x_i, i = 1, \ldots, n\}$ und mit $\mathbb{E}[\varepsilon_i] = 0$. Wir beschränken uns hier auf den Fall, dass die Störgrößen einer Normalverteilung folgen, also $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Nach einer Orts- und Maßstabstransformation können wir ohne Einschränkung davon ausgehen, dass die Responsevariable zentriert und die Vorhersagevariablen standardisiert sind, das heißt es gilt

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{und} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{für } j = 1, \ldots, p.$$

Somit werden die Beobachtungen der Variablen derart transformiert, dass sie den Mittelwert null und eine Varianz von eins haben. Dies ist nötig, da der ridge-regression-, der lasso- und der elastic net-Schätzer von der Skalierung der Einflussgrößen, sowie von der Wahl des Ursprungs abhängen, was sich auf die von ihnen vorgenommene Parameterschrumpfung zurückführen lässt.

*Der Vortrag basiert auf der Veröffentlichung „Regularization and variable selection via the elastic net“ von Hui Zuo und Trevor Hastie und ist Teil des Hauptseminars Erweiterungen des linearen Regressionsmodells und genomische Anwendungen in der Biomedizin im Wintersemester 2014/15.

1.2 Kleinste-Quadrate-Schätzer

Der Kleinste-Quadrate-Schätzer $\hat{\beta}_{\text{KQ}}$ ist gegeben durch

$$\hat{\beta}_{\text{KQ}} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 .$$

Falls die Designmatrix $\mathbf{X}$ vollen Rang hat und somit $(\mathbf{X}^T\mathbf{X})^{-1}$ existiert, gilt

$$\hat{\beta}_{\text{KQ}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Es lässt sich sogar zeigen, dass der Kleinste-Quadrate-Schätzer der beste erwartungstreue Schätzer ist im Sinne der kleinsten Varianz. In manchen Situationen ist es allerdings sinnvoll, auf die Erwartungstreue zu verzichten und sich auch nach verzerrten Schätzern umzusehen, da der Kleinste-Quadrate-Schätzer im Fall $p > n$ nicht mehr eindeutig ist. Die Betrachtung von verzerrten Schätzern kann auch von Vorteil sein, um den Mean-Squared-Error (MSE), der sich aus der Varianz und dem Bias eines Schätzers zusammensetzt, zu verringern.

1.3 Ridge-regression-Schätzer

Der ridge-regression-Schätzer $\hat{\beta}_{\text{ridge}}(\lambda)$ (Hoerl und Kennard, 1988) hat zusätlich zu dem Term des Kleinste-Quadrate-Schätzers eine l_2-Regularisierung

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

wobei $\lambda \geq 0$ ein deterministischer Skalar ist. Um den Tuningparameter λ zu bestimmen, bei dem der MSE minimal ist, bietet sich Kreuzvalidierung an. Bei dem Verfahren der Kreuzvalidierung geht man so vor, dass man seine Datenmenge in k Teilmengen einteilt und dann k Durchläufe startet. Bei dem i-ten Durchlauf dient die i-te Teilmenge als Testmenge zur Fehlerbestimmung und die restlichen $(k-1)$-Teilmengen werden als Trainingsmengen zur Parameterbestimmung genutzt.

Ridge-regression hat zwar eine bessere Vorhersagegenauigkeit, aber der Nachteil ist, dass es keine Variablenselektierung betreibt, sondern alle Vorhersagen im Modell behält und man somit kein „einfaches“ Modell erhält.

1.4 Lasso-Schätzer

Ein Verfahren, das Variablenselektierung sowie Schrumpfung der Variablen gleichzeitig betreibt und damit sowohl ein einfach zu interpretierendes Modell, als auch eine gute Vorhersagegenauigkeit erreicht, ist das lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right),$$

das von Tibshirani (1996) vorgeschlagen wurde. Hier wird die Kleinste-Quadrate-Methode durch eine l_1-Regulierung der Regressionskoeffizienten verbessert und der Tuningparamter λ wird wie bei ridge regression meist durch Kreuzvalidierung bestimmt.

Der Schätzer erreicht Variablenselektierung, indem $\hat{\beta}^{(j)}_{\text{lasso}}(\lambda)$ abhängig von dem gewählten λ für einige j gleich null geschätzt wird.

2 Einleitung

In diesem Vortrag wird das Verhalten des lasso-Schätzers bei stark korrelierten Prädiktoren diskutiert und damit die Vorstellung des elastic net-Schätzers motiviert, der in diesem Fall die Schätzwerte der Koeffizienten gleichmäßiger auf die Gruppen verteilt. Des Weiteren werden einige Eigenschaften des elastic net-Schätzers aufgeführt sowie bewiesen.

3 Lasso bei stark korrelierten Prädiktoren

Obwohl der lasso-Schätzer oft gute Ergebnisse liefert, stößt er in einigen Situationen an seine Grenzen:

- Im Fall $p > n$ kann der lasso-Schätzer höchstens n Variablen auswählen.
- Er ist nicht wohldefiniert, sofern die l_1-Norm der Koeffizienten nicht kleiner als ein bestimmter Wert ist.
- Ridge regression liefert bessere Ergebnisse als der lasso-Schätzer im Fall von $n > p$ und starken Korrelationen zwischen den Vorhersagevariablen.
- Bei einer Gruppe von stark korrelierten Variablen neigt der lasso-Schätzer dazu, nur eine Variable auszuwählen, ohne dabei zu berücksichtigen welche.

Wir veranschaulichen die genannten Punkte am Beispiel eines Gen-Selektierungsproblem bei einer Microarray-Datenmenge. Eine solch typische Datenmenge hat mehr als 1000 Vorhersagevariablen (Gene) und meistens weniger als 100 Proben, somit befinden wir uns in dem Fall $p \gg n$. Gene können stark korreliert sein, falls sie den gleichen biologischen „Pfad“ gemeinsam haben. Wir stellen uns vor, dass diese Gene eine Gruppe bilden. Die optimale Genselektierungsmethode sollte die nichtssagenden Gene eliminieren und ganze Gruppen automatisch einbinden, falls ein Gen dieser Gruppe ausgewählt wurde.

Für diesen Fall von $p \gg n$ mit gruppierten Variablen ist das lasso wegen den zu Beginn genannten Beobachtungen keine ideale Methode. Wir suchen also einen neuen Schätzer, der genauso gut ist wie der lasso bei dessen optimalen Situationen und besser bei den oben genannten Problemfällen.

4 Naive elastic net

4.1 Definition

Für feste und nichtnegative λ_1 und λ_2 definieren wir das naive elastic net-Kriterium

$$L(\lambda_1, \lambda_2, \beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 .$$

Der naive elastic net-Schätzer ist dann gegeben durch

$$\hat{\beta} = \arg\min_{\beta}\{L(\lambda_1, \lambda_2, \beta)\} \tag{1}$$

und kann als regulierter Kleinste-Quadrate-Schäzer angesehen werden. Sei $\alpha := \frac{\lambda_2}{\lambda_1+\lambda_2}$, dann ist (1) äquivalent zu

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{unter der Nebenbedingung} \quad (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t,$$

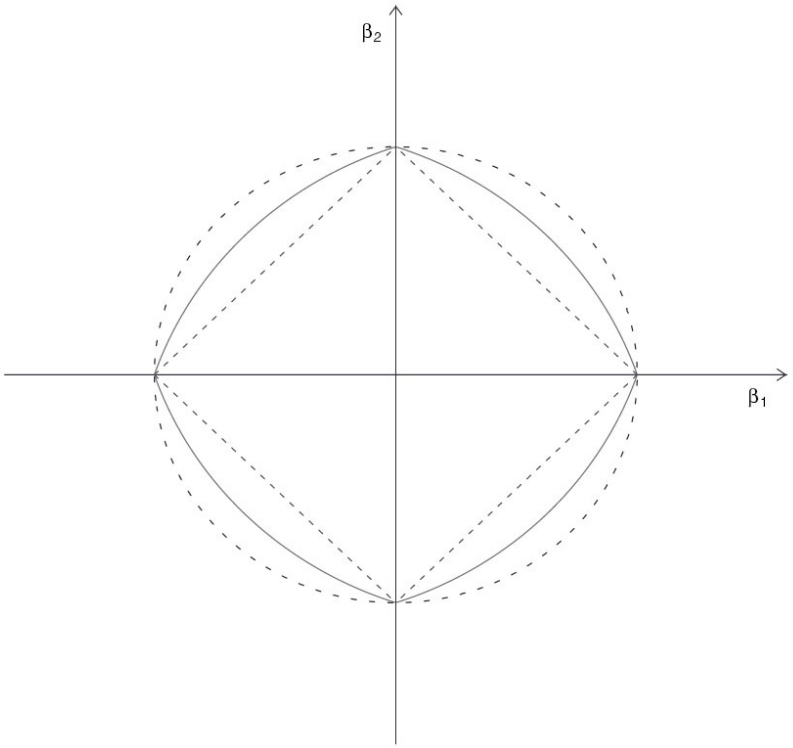

Abbildung 1: Zweidimensionale Konturlinien der (von außen nach innen) ridge-Regularisierung, elastic net-Regularisierung mit $\alpha = 0,5$ und der lasso-Regularisierung.

für ein $t := t(\lambda_1+\lambda_2, \alpha)$. Wir nennen $(1-\alpha)\|\beta\|_1+\alpha\|\beta\|_2^2$ die elastic net-Regularisierungsfunktion, die eine Konvexkombination der lasso- und der ridge-Regularisierung ist. Es ist leicht zu sehen, dass man im Fall $\alpha = 1$ den ridge regression-Schätzer erhält. Im Weiteren betrachten wir nur $\alpha < 1$. Für diese α ist die elastic net-Regularisierungsfunktion singulär in 0 und strikt konvex für alle $\alpha > 0$ und hat somit die Eigenschaften von lasso und ridge regression, was man in Abb.1 sehen kann.

4.2 Lösung

Im Folgenden zeigen wir, dass (1) zu einem lasso-Optimierungsproblem äquivalent ist.

Lemma 4.1 *Gegeben sei die Datenmenge* $(\boldsymbol{y}, \boldsymbol{X})$ *und* (λ_1, λ_2). *Definiere eine neue künstlich erzeugte Datenmenge* $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ *mit*

$$\boldsymbol{X}^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2} I \end{pmatrix} \quad \textit{und} \quad \boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix},$$

wobei $\boldsymbol{X}^*$ *eine* $((n+p)\times p)$*-Matrix ist und* $\boldsymbol{y}^*$ *ein Vektor mit* $(n+p)$ *Spalten. Sei* $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$ *und* $\beta^* = \sqrt{1+\lambda_2}\beta$, *dann kann das naive elastic net-Kriterium geschrieben werden als*

$$L(\gamma, \beta) = L^*(\gamma, \beta^*) = \|\boldsymbol{y}^* - \boldsymbol{X}^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1 .$$

Sei $\hat{\beta}^* = \arg\min_{\beta^*}\{L^*(\gamma, \beta^*)\}$, *dann ist auch*

$$\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*.$$

Beweis: Zeige erst, dass $L(\gamma, \beta) = L^*(\gamma, \beta^*)$ ist.

$$\begin{aligned}
L^*(\gamma, \beta^*) &= \|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \gamma \|\beta^*\|_1 \\
&= \left\| \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} - \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} I \end{pmatrix} \sqrt{1+\lambda_2}\beta \right\|_2^2 + \gamma \left\| \sqrt{1+\lambda_2}\beta \right\|_1 \\
&= \left\| \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} I \end{pmatrix} \beta \right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \sqrt{1+\lambda_2} \|\beta\|_1 \\
&= \left\| \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{X}\beta \\ \sqrt{\lambda_2}\beta \end{pmatrix} \right\|_2^2 + \lambda_1 \|\beta\|_1 \\
&= \left\| \begin{pmatrix} \mathbf{y} - \mathbf{X}\beta \\ -\sqrt{\lambda_2}\beta \end{pmatrix} \right\|_2^2 + \lambda_1 \|\beta\|_1 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \\
&= L(\gamma, \beta).
\end{aligned}$$

Sei nun $\hat{\beta}^* = \arg\min_{\beta^*}\{L^*(\gamma, \beta^*)\}$, dann ist auch

$$\sqrt{1+\lambda_2}\hat{\beta} = \arg \min_{\sqrt{1+\lambda_2}\beta} \{L(\gamma, \beta)\} = \arg \min_{\beta^*} \{L^*(\gamma, \beta^*)\} = \hat{\beta}^*.$$

QED

Somit kann das naive elastic net-Problem in ein äquivalentes lasso-Problem mit vergrößerter Datenmenge umgewandelt werden. Das Lemma zeigt auch, dass das naive elastic net ähnlich zu dem lasso automatische Variablenselektierung betreibt.

4.3 Gruppierungseffekt

In dem Fall $p \gg n$ ist die Situation mit gruppierten Variablen von großem Belang, deswegen studieren wir in diesem Abschnitt den Gruppierungseffekt einiger Verfahren.

Wir betrachten hier die generische Regularisierungsmethode

$$\hat{\beta} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda J(\beta), \tag{2}$$

wobei $J(\cdot)$ positive Werte für $\beta \neq 0$ annimmt.

Eine Regressionsmethode sollte einen Gruppierungseffekt aufweisen, falls die Regressionskoeffizienten einer Gruppe von stark korrelierten Variablen dazu neigen, gleich zu sein (bis auf Vorzeichen). Besonders in dem Fall, wenn einige Variablen gleich sind, sollte die Methode identischen Variablen identische Koeffizienten zuweisen.

Lemma 4.2 *Es gibt $i, j \in \{1, \ldots, p\}$ mit $x_i = x_j$.*

(a) Falls $J(\cdot)$ strikt konvex und symmetrisch ist, gilt $\hat{\beta}_i = \hat{\beta}_j$ für alle $\lambda > 0$.

(b) Falls $J(\beta) = \|\beta\|_1$, dann ist $\hat{\beta}_i\hat{\beta}_j \geq 0$ und $\tilde{\hat{\beta}}^*$ ist ein weiterer Minimierer von (2), wobei

$$\tilde{\hat{\beta}}^* = \begin{cases} \hat{\beta}_k & \text{, falls } k \neq i \text{ und } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j)s & \text{, falls } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j)(1-s) & \text{, falls } k = j, \end{cases}$$

mit $s \in [0, 1]$.

Beweis:

(a) Sei $J(\cdot)$ strikt konvex und $\lambda > 0$. Angenommen $\hat{\beta}_i \neq \hat{\beta}_j$.

$\hat{\beta}^*$ sei definiert wie folgt

$$\hat{\beta}^* = \begin{cases} \hat{\beta}_k & \text{, falls } k \neq i \text{ und } k \neq j, \\ \frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) & \text{, falls } k = i \text{ oder } k = j. \end{cases}$$

Da $x_i = x_j$, ist auch $\mathbf{X}\hat{\beta}^* = \mathbf{X}\hat{\beta}$ und somit auch

$$\left\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\right\|_2^2 = \left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|_2^2.$$

Weil $J(\cdot)$ strikt konvex ist, erhalten wir

$$\begin{aligned} J(\hat{\beta}^*) &= J\left(\frac{1}{2}\left(\hat{\beta}_1, \ldots, \hat{\beta}_i, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_p\right)^T + \frac{1}{2}\left(\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_i, \ldots, \hat{\beta}_p\right)^T\right) \\ &< \frac{1}{2}J(\hat{\beta}) + \frac{1}{2}J\left(\left(\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_i, \ldots, \hat{\beta}_p\right)^T\right). \end{aligned}$$

Da $J(\cdot)$ symmetrisch ist, gilt $J(\left(\hat{\beta}_1, \cdots, \hat{\beta}_j, \cdots, \hat{\beta}_i, \cdots, \hat{\beta}_p\right)^T) = J(\hat{\beta})$ und somit erhalten wir $J(\hat{\beta}^*) < J(\hat{\beta})$. Dies ist aber ein Widerspruch zur Minimalität von $\hat{\beta}$ in (2) und daher folgt $\hat{\beta}_i = \hat{\beta}_j$.

(b) Sei $J(\beta) = \|\beta\|_1$, also erhalten wir in (2) die lasso-Regularisierungsfunktion. Angenommen $\hat{\beta}_i\hat{\beta}_j < 0$ und somit auch $\text{sgn}\{\hat{\beta}_i\} \neq \text{sgn}\{\hat{\beta}_j\}$. Damit folgt $|\hat{\beta}_i + \hat{\beta}_j| < |\hat{\beta}_i| + |\hat{\beta}_j|$. Betrachte nun dasselbe $\hat{\beta}^*$ wie oben, also auch $\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|_2^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2$. Dann gilt

$$\begin{aligned} \|\hat{\beta}^*\|_2^2 &= |\hat{\beta}_1| + \ldots + |\hat{\beta}_{i-1}| + |\frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j)| + |\hat{\beta}_{i+1}| + \ldots + |\hat{\beta}_{j-1}| \\ &\quad + |\frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j)| + |\hat{\beta}_{j+1}| + \ldots + |\hat{\beta}_p| \\ &< \|\hat{\beta}\|_2^2, \end{aligned}$$

da $\hat{\beta}_i\hat{\beta}_j < 0$. Somit erhalten wir einen Widerspruch zu der Minimalität des lasso-Schätzers $\hat{\beta}$ und es folgt $\hat{\beta}_i\hat{\beta}_j \geq 0$ und damit auch $\text{sgn}\{\hat{\beta}_i\} = \text{sgn}\{\hat{\beta}_j\}$.

Es bleibt zu zeigen, dass $\tilde{\hat{\beta}}^*$ auch ein Minimierer von (2) ist. Mit dem gleichen Argument wie in Teil (a) gilt

$$\left\|\mathbf{y} - \mathbf{X}\tilde{\hat{\beta}}^*\right\|_2^2 = \left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|_2^2.$$

Es ist aber auch $\|\tilde{\hat{\beta}}^*\|_1 = \|\hat{\beta}\|_1$, da

$$\begin{aligned} \|\tilde{\hat{\beta}}^*\|_1 &= |\hat{\beta}_1| + \ldots + |\hat{\beta}_{i-1}| + |(\hat{\beta}_i + \hat{\beta}_j)s| + |\hat{\beta}_{i+1}| + \ldots + |\hat{\beta}_{j-1}| \\ &\quad + |(\hat{\beta}_i + \hat{\beta}_j)(1-s)| + |\hat{\beta}_{j+1}| + \ldots + |\hat{\beta}_p| = \|\hat{\beta}\|_1. \end{aligned}$$

QED

Lemma 4.2 zeigt den Unterschied zwischen strikt konvexen Regularisierungsfunktionen, also insbesondere dem elastic net mit $\lambda_2 > 0$, und der lasso-Regularisierung. Die strikte Konvexität garantiert einen Gruppierungseffekt in dem Fall, dass es identische Vorhersagevariablen gibt. Das lasso hat in diesem Fall nicht einmal eine eindeutige Lösung.

Satz 4.3 *Gegeben sei die Datenmenge* $(\boldsymbol{y}, \boldsymbol{X})$ *und die Parameter* (λ_1, λ_2). $\hat{\beta}(\lambda_1, \lambda_2)$ *sei der naive elastic net-Schätzer und wir nehmen an, dass* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ *ist. Definiere*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{\|\boldsymbol{y}\|_2}\left|\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)\right|,$$

dann ist

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)},$$

wobei $\rho = x_i^T x_j$ *die Stichprobenkorrelation ist.*

Beweis: Da wir vorausgesetzt haben, dass $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, sind $\hat{\beta}_i(\lambda_1, \lambda_2)$ und $\hat{\beta}_j(\lambda_1, \lambda_2)$ nicht 0 und haben das gleiche Vorzeichen, also $\operatorname{sgn}\{\hat{\beta}_i(\lambda_1, \lambda_2)\} = \operatorname{sgn}\{\hat{\beta}_j(\lambda_1, \lambda_2)\}$.

Wegen $\hat{\beta} = \arg\min_\beta\{L(\lambda_1, \lambda_2, \beta)\}$, gilt

$$\frac{\partial L(\lambda_1,\lambda_2,\beta)}{\partial \beta_k}\,|_{\beta=\hat{\beta}(\lambda_1,\lambda_2)} = 0, \quad \text{wobei}$$

$$\begin{aligned}
\frac{\partial L(\lambda_1,\lambda_2,\beta)}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k}\left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right) \\
&= \frac{\partial}{\partial \beta_k}\left(\sum_{i=1}^{n}[y_i - (x_1^{(i)}\beta_1 + \cdots + x_p^{(i)}\beta_p)]^2 + \lambda_2\left(\sum_{j=1}^{p}\beta_j^2\right) + \lambda_1\left(\sum_{j=1}^{p}|\beta_j|\right)\right) \\
&= 2\left(\sum_{i=1}^{n}(-x_k^{(i)})\left(y_i - (x_1^{(i)}\beta_1 + \cdots + x_p^{(i)}\beta_p)\right)\right) + 2\lambda_2\beta_k + \lambda_1\operatorname{sgn}\{\beta_k\} \\
&= -2x_k^T(\mathbf{y} - \mathbf{X}\beta) + 2\lambda_2\beta_k + \lambda_1\operatorname{sgn}\{\beta_k\},
\end{aligned}$$

falls $\hat{\beta}_k(\lambda_1, \lambda_2) \neq 0$, da die Abbildung $\beta \mapsto \|\beta\|_1$ nicht differenzierbar in der 0 ist. Da $\hat{\beta}_i(\lambda_1, \lambda_2)$ und $\hat{\beta}_j(\lambda_1, \lambda_2)$ ja gerade ungleich 0 sind, erhalten wir

$$\begin{aligned}
-2x_i^T(\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1,\lambda_2)) + 2\lambda_2\hat{\beta}_i(\lambda_1,\lambda_2) + \lambda_1\operatorname{sgn}\{\hat{\beta}_i(\lambda_1,\lambda_2)\} = 0 \quad \text{und} \\
-2x_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1,\lambda_2)) + 2\lambda_2\hat{\beta}_j(\lambda_1,\lambda_2) + \lambda_1\operatorname{sgn}\{\hat{\beta}_j(\lambda_1,\lambda_2)\} = 0.
\end{aligned}$$

Subtrahieren ergibt

$$(x_j^T - x_i^T)(\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1,\lambda_2)) + \lambda_2(\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)) = 0,$$

was äquivalent ist zu

$$\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2) = \frac{1}{\lambda_2}(x_i^T - x_j^T)\hat{r}(\lambda_1,\lambda_2), \tag{3}$$

wobei $\hat{r}(\lambda_1, \lambda_2) = \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1, \lambda_2)$ der Residuumsvektor ist. Da $\mathbf{X}$ standardisiert ist, ist

$$\begin{aligned}
\|x_i - x_j\|_2^2 &= \sum_{k=1}^{n}(x_i^{(k)} - x_j^{(k)})^2 = \sum_{k=1}^{n}({x_i^{(k)}}^2 - 2x_i^{(k)}x_j^{(k)} + {x_j^{(k)}}^2) \\
&= \|x_i\|_2^2 - 2x_i^T x_j + \|x_j\|_2^2 = 2 - 2x_i^T x_j = 2(1-\rho),
\end{aligned}$$

wobei $\rho = x_i^T x_j$.

Weil $\hat{\beta}(\lambda_1, \lambda_2)$ der Minimierer von $L(\lambda_1, \lambda_2, \beta)$ ist, gilt außerdem $L(\lambda_1, \lambda_2, \hat{\beta}(\lambda_1, \lambda_2)) \leq L(\lambda_1, \lambda_2, \beta = 0)$, das heißt

$$\|\hat{r}(\lambda_1, \lambda_2)\|_2^2 + \lambda_2 \left\|\hat{\beta}(\lambda_1, \lambda_2)\right\|_2^2 + \lambda_1 \left\|\hat{\beta}(\lambda_1, \lambda_2)\right\|_1 \leq \|\mathbf{y}\|_2^2 = L(\lambda_1, \lambda_2, \beta = 0).$$

Somit ist auch $\|\hat{r}(\lambda_1, \lambda_2)\|_2^2 \leq \|\mathbf{y}\|_2^2$ und mit (3) ergibt dies

$$\begin{aligned} D_{\lambda_1,\lambda_2}(i,j) &= \frac{1}{\|y\|_2} \left|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)\right| = \frac{1}{\|y\|_2} \left|\frac{1}{\lambda_2}(x_i^T - x_j^T)\hat{r}(\lambda_1, \lambda_2)\right| \\ &\overset{\text{Cauchy-Schwarz}}{\leq} \frac{1}{\lambda_2} \frac{\|\hat{r}(\lambda_1, \lambda_2)\|_2}{\|y\|_2} \|x_i - x_j\|_2 \leq \frac{1}{\lambda_2} \|x_i - x_j\|_2 = \frac{1}{\lambda_2} \sqrt{2(1-\rho)}. \end{aligned}$$

QED

Die einheitslose Quantität $D_{\lambda_1,\lambda_2}(i,j)$ beschreibt die Differenz der Koeffizientenpfade der Vorhersagen i und j. Falls x_i und x_j stark korreliert sind, d.h. $\rho \approx 1$, ist nach Satz 4.3 die Differenz der Koeffizientenpfade von i und j fast 0. Die obere Schranke der Ungleichung ist eine quantitative Beschreibung für den Gruppierungseffekt des naiven elastic net.

4.4 Defizit

Empirische Belege zeigen, dass das naive elastic net nicht zufriedenstellend ist, wenn es nicht nahe an ridge regression oder dem lasso ist, was auch der Grund für die Bezeichnung „naiv“ ist.

Das naive elastic net ist ein zweistufiges Verfahren: Für jedes feste λ_2 finden wir zuerst die ridge regression Koeffizienten und betreiben dann die lasso-Schrumpfung entlang der Lösungspfade der lasso-Koeffizienten. Somit entsteht eine doppelte Schrumpfung, die nicht sehr dabei hilft, die Varianz zu verringern und unnötiges zusätzliches Bias verglichen mit reiner lasso- oder ridge-Schrumpfung verursacht.

5 Elastic net

In diesem Abschnitt wollen wir die Vorhersageleistung des naiven elastic net verbessern, indem wir die doppelte Schrumpfung korregieren.

Gegeben seien die Datenmenge $(\mathbf{y}, \mathbf{X})$, Regularisierungsparameter (λ_1, λ_2) und die vergrößerte Datenmenge $(\mathbf{y}^*, \mathbf{X}^*)$. Der naive elastic net-Schätzer löst das Regressionsmodell in der lasso-Schreibweise

$$\hat{\beta}^* = \arg\min_{\beta^*} \|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\beta^*\|_1 . \tag{4}$$

Der korrigierte elastic net-Schätzer $\hat{\beta}$ ist definiert durch

$$\hat{\beta}_{\text{elastic net}} := \sqrt{1+\lambda_2}\hat{\beta}^*.$$

In Lemma 4.1 haben wir gezeigt, dass

$$\hat{\beta}_{\text{naive elastic net}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$$

und somit ist dann

$$\hat{\beta}_{\text{elastic net}} = (1+\lambda_2)\hat{\beta}_{\text{naive elastic net}}.$$

Daher sind die elastic net-Koeffizienten umskalierte naive elastic net-Koeffizienten, also gelten auch alle zuvor beschriebenen Eigenschaften für das elastic net. Empirisch wurde festgestellt, dass das elastic net im Vergleich zu lasso und ridge regression sehr gute Ergebnisse liefert.

Diese Transformation des naiven elastic net erhält die Eigenschaft der Variablenselektierung und ist der einfachste Weg, die Schrumpfung rückgängig zu machen.

Wir wollen an dieser Stelle eine weitere Motivation für die Umskalierung liefern. Hierzu betrachten wir den orthogonalen Fall, in dem die Kovarianz der Spalte der Designmatrix 0 ist. Dann ist die ridge-regression-Lösung mit Parameter λ_2 gegeben durch

$$\hat{\beta}_{\text{ridge}} = \frac{\hat{\beta}_{\text{KQ}}}{(1+\lambda_2)}$$

und die lasso-Lösung mit Parameter λ_1 durch

$$\hat{\beta}^{(i)}_{\text{lasso}} = \left(\left|\hat{\beta}^{(i)}_{\text{KQ}}\right| - \frac{\lambda_1}{2}\right)_+ \text{sgn}\{\hat{\beta}^{(i)}_{\text{KQ}}\},$$

wobei $\hat{\beta}_{\text{KQ}} = \mathbf{X}^T\mathbf{y}$ und z_+ den positiven Anteil bezeichnet. In diesem orthogonalen Fall erhält man als Lösung für das naive elastic net mit Parametern (λ_1, λ_2), unter der Beachtung von

$$\hat{\beta}^*_{\text{KQ}} = \frac{1}{\sqrt{1+\lambda_2}}\mathbf{X}^{*T}\mathbf{y}^* = \frac{1}{\sqrt{1+\lambda_2}}(\mathbf{X}^T|\sqrt{\lambda_2}I)\begin{pmatrix}\mathbf{y}\\0\end{pmatrix} = \frac{1}{\sqrt{1+\lambda_2}}\mathbf{X}^T\mathbf{y} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}_{\text{KQ}}$$

die Gleichung

$$\begin{aligned}
\hat{\beta}^{(i)}_{\text{naive elastic net}} &= \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^{(i)}_{\text{lasso}} = \frac{1}{\sqrt{1+\lambda_2}}\left(\left|\hat{\beta}^{*}_{\text{KQ}}{}^{(i)}\right| - \frac{\gamma}{2}\right)_+ \text{sgn}\{\hat{\beta}^{*}_{\text{KQ}}{}^{(i)}\}\\
&= \frac{1}{\sqrt{1+\lambda_2}}\left(\frac{1}{\sqrt{1+\lambda_2}}\left|\hat{\beta}^{(i)}_{\text{KQ}}\right| - \frac{\lambda_1}{2\sqrt{1+\lambda_2}}\right)_+ \text{sgn}\{\frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^{(i)}_{\text{KQ}}\}\\
&= \frac{1}{1+\lambda_2}\left(\left|\hat{\beta}^{(i)}_{\text{KQ}}\right| - \frac{\lambda_1}{2}\right)_+ \text{sgn}\{\hat{\beta}^{(i)}_{\text{KQ}}\}.
\end{aligned}$$

Das naive elastic net kann also als eine zweistufige Methode angesehen werden, eine Art ridge-Schrumpfung gefolgt von einer Art lasso-thresholding. Da wir aber wissen, dass der lasso-Schätzer in dem orthogonalen Fall optimal schätzt, ist die direkte Schrumpfung $\frac{1}{1+\lambda_2}$ des naiven elastic net nicht nötig und wird durch die Reskalierung bei dem elastic net behoben. Obwohl ridge-regression diese Schrumpfung benötigt, um die Varianz zu kontrollieren, verlassen wir uns bei dem elastic net auf die lasso-Schrumpfung, um die Varianz zu bändigen.

Im Weiteren bezeichnet $\hat{\beta}$ nun immer den elastic net-Schätzer $\hat{\beta}_{\text{elastic net}}$.

Satz 5.1 *Gegeben sei der Datensatz* $(\boldsymbol{y}, \boldsymbol{X})$ *und* (λ_1, λ_2). *Dann ist der elastic net-Schätzer* $\hat{\beta}$ *gegeben durch*

$$\hat{\beta} = \arg\min_{\beta}\left(\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_2 I}{1+\lambda_2}\right)\beta - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \lambda_1\left\|\beta\right\|_1.$$

und somit dann auch

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\beta^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)\beta - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \lambda_1\left\|\beta\right\|_1.$$

Beweis: Sei $\hat{\beta}$ der elastic net-Schätzer. Nach Definition und (4) erhalten wir

$$\begin{aligned}\hat{\beta}^* &= \arg\min_{\beta^*} \|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\beta^*\|_1 ,\\ \hat{\beta} &= \sqrt{1+\lambda_2}\hat{\beta}^*\end{aligned}$$

und somit gilt

$$\begin{aligned}\hat{\beta} &= \sqrt{1+\lambda_2}\hat{\beta}^* = \sqrt{1+\lambda_2}\arg\min_{\beta^*} \|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\beta^*\|_1\\ &= \arg\min_{\beta} \left\|\mathbf{y}^* - \mathbf{X}^*\frac{\beta}{\sqrt{1+\lambda_2}}\right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left\|\frac{\beta}{\sqrt{1+\lambda_2}}\right\|_1\\ &= \arg\min_{\beta} \sum_{i=1}^{n} (\mathbf{y}_i^*)^2 - 2\left(\mathbf{y}_i^*(\mathbf{X}^*\beta)_i \frac{1}{\sqrt{1+\lambda_2}}\right) + \frac{(\mathbf{X}^*\beta)_i^2}{1+\lambda_2} + \frac{\lambda_1}{1+\lambda_2}\|\beta\|_1\\ &= \arg\min_{\beta} \mathbf{y}^{*T}\mathbf{y}^* - 2\frac{\mathbf{y}^{*T}\mathbf{X}^*\beta}{\sqrt{1+\lambda_2}} + \beta^T\left(\frac{\mathbf{X}^{*T}\mathbf{X}^*}{1+\lambda_2}\right)\beta + \frac{\lambda_1}{1+\lambda_2}\|\beta\|_1 .\end{aligned}$$

Substituiert man $\mathbf{X}^* = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}I\end{pmatrix}$ und $\mathbf{y}^* = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}$, folgt damit

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \mathbf{y}^T\mathbf{y} + \frac{1}{1+\lambda_2}\left(-2\mathbf{y}^T\mathbf{X}\beta + \beta^T\left(\frac{\mathbf{X}^T\mathbf{X}+\lambda_2 I}{1+\lambda_2}\right)\beta + \lambda_1\|\beta\|_1\right)\\ &= \arg\min_{\beta} -2\mathbf{y}^T\mathbf{X}\beta + \beta^T\left(\frac{\mathbf{X}^T\mathbf{X}+\lambda_2 I}{1+\lambda_2}\right)\beta + \lambda_1\|\beta\|_1 .\end{aligned}$$

Dabei ist der letzte Schritt möglich, da das minimierende Argument durch Skalierung und Verschiebung der zu minimierenden Funktion nicht verändert wird.

Es bleibt zu zeigen, dass $\hat{\beta}_{\text{lasso}} = \arg\min_\beta \beta^T\left(\mathbf{X}^T\mathbf{X}\right)\beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1\|\beta\|_1$ gilt. Da beim lasso aber gerade $\lambda_2 = 0$ ist, folgt die Gleichung direkt aus der Formel für den elastic net-Schätzer. QED

Somit kann man das elastic net nach Satz 5.1 als numerisch stabilere Version des lasso betrachten. Man erhält also, dass die Reskalierung nach der elastic net Regularisierung mathematisch äquivalent ist zu dem Ersetzen von $\mathbf{X}^T\mathbf{X}$ mit dessen geschrumpter Version im lasso.