Konsistenzeigenschaften des (Adaptive) Lasso

Hauptseminar Erweiterungen des linearen Regressionsmodells und genomische Anwendungen in der Biomedizin
WS 2014/2015

Kristina Kaucher

1. Dezember 2014

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Einleitung

- Zugrunde liegendes lineares Modell: $\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$
- Mit Ergebnisvektor $\mathbf{Y} \in \mathbb{R}^{n\times 1}$, Designmatrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ und dem Vektor der Messfehler $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$.
- $\mathbf{X}$ fest vorgegeben und $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\, \boldsymbol{I})$ unb $\boldsymbol{\beta}^0$ bezeichnet den wahren Koeffizientenvektor.
- Was versteht man unter einer „Oracle-Ungleichung“?

- Annahme $p \leq n$ und $\mathbf{X}$ hat vollen Rang p
  - Kleinster Quadrate Schätzer $\hat{\mathrm{b}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ gibt den Fehler $\| \mathbf{X}(\hat{\mathrm{b}} - \boldsymbol{\beta}^0)\|_2^2/\sigma^2 \sim \chi_p^2$, also $\frac{\mathbb{E}\left[\| \mathbf{X}(\hat{\mathrm{b}} - \boldsymbol{\beta}^0)\|_2^2\right]}{n} = \frac{\sigma^2}{n} p$.
  - Reparametrisierung zu orthonormalem Design $\Rightarrow$ jeder Parameter $\boldsymbol{\beta}_j^0$ wird mit quadratischer Genauigeit $\frac{\sigma^2}{n}$ geschätzt.
- Was ist jedoch im Fall $p > n$?
  - Wir müssen annehmen, dass nur ein paar unserer $\boldsymbol{\beta}_j^0$ ungleich Null sind, dazu sei $\mathrm{S}_0 := \{j : \boldsymbol{\beta}_j^0 \neq 0, j = 1, \ldots, p\}$ mit $|\,\mathrm{S}_0\,| := s_0$ die aktive Menge der Koeffizienten.
  - Wüssten wir wie S_0 aussieht, dann wüssten wir auch welche $\mathbf{X}^{(j)}$ für unser Modell irrelevant sind und hätten eine quadratische Genauigkeit von $\frac{\sigma^2}{n} s_0$ für den Schätzer von $\boldsymbol{\beta}^0$.
  - S_0 ist jedoch unbekannt $\Rightarrow$ Regularisierung der $\boldsymbol{\beta}_j$ benötigt.

- Regularisierung mit der l_1-Norm über den Lasso-Schätzer:

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \left\{ \frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta} \|_2^2}{n} + \lambda \| \boldsymbol{\beta} \|_1 \right\}$$

mit dem Regularisierungsparameter $\lambda > 0$

- Wir werden sehen: Mit geeigneter Wahl von λ (der Ordnung $\sigma\sqrt{\log \frac{p}{n}}$) erfüllt der Lasso mit großer Wahrscheinlichkeit die Oracle-Ungleichung

$$\frac{\| \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \|_2^2}{n} \leq c \frac{\sigma^2 \log(p)}{n} s_0.$$

- $c > 0$ kann dabei explizit in Abhängigkeit von p und/oder n durch $\hat{\Sigma} := \mathbf{X}^\top \mathbf{X} / n$ angegeben werden, $\log(p)$ ist der Preis, den wir zahlen S_0 nicht zu kennen.

Oracle Eigenschaften des Lasso

Lemma 2.1 (Basis-Ungleichung)

Für den Lasso Schätzer $\hat{\beta}$ gilt die folgende Ungleichung

$$\frac{\| \mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda \| \hat{\beta} \|_1 \leq \frac{2}{n} \boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\beta} - \beta^0) + \lambda \| \beta^0 \|_1$$

Beweis.

Es gilt

$$\frac{\| \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} \|_1 \leq \frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0 \|_2^2}{n} + \lambda \| \boldsymbol{\beta}^0 \|_1 .$$

Die Addition von $\frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0 \|_2^2}{n}$ *auf beiden Seiten bringt*

$$\frac{\| \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \|_2^2}{n} + \frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0 \|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} \|_1 \leq 2\frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0 \|_2^2}{n} + \lambda \| \boldsymbol{\beta}^0 \|_1 .$$

Nach Ausnutzung der Dreiecksungleichung und der Eigenschaft $\mathbf{Y} = \mathbf{X}\,\hat{\boldsymbol{\beta}}$ *erhalten wir für die linke Seite der Ungleichung*

$$\begin{aligned} &\frac{\| \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \|_2^2}{n} + \frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0 \|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} \|_1 \\ &\qquad \geq \frac{\| \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} + \mathbf{Y} - \mathbf{Y}\,\boldsymbol{\beta}^0 \|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} \|_1 \end{aligned}$$

$$= \frac{\| \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} \|_1.$$

Die rechte Seite der Ungleichung ergibt mit $\mathbf{Y} = \mathbf{X}\,\beta^0 + \varepsilon$ und $\mathbf{Y} = \mathbf{X}\,\hat{\beta}$

$$2\frac{\| \mathbf{Y} - \mathbf{X}\,\beta^0 \|_2^2}{n} + \lambda \| \boldsymbol{\beta}^0 \|_1 = \frac{2}{n}(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0)^T(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}^0)$$
$$= \frac{2}{n}\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda \| \boldsymbol{\beta}^0 \|_1.$$

Womit die Aussage bewiesen ist. $\square$

- $\frac{2}{n}\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$ ist der empirische Verfahrensteil des Problems, er lässt sich hinsichtlich der $l1$-Norm einfach einschränken mit

$$2|\,\boldsymbol{\varepsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)| \leq \left(\max_{1\leq j\leq p} 2|\,\boldsymbol{\varepsilon}^\top \mathbf{X}^{(j)}\,|\right) \|\,\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}^0\,\|_1.$$

- Im empirischen Verfahrensteil spielt der Zufall noch eine Rolle $\Rightarrow$ Regularisierung um ihn zu überstimmen wird benötigt, dazu sei

$$\mathcal{J} := \left\{\max_{1\leq j\leq p} \frac{2}{n}|\,\boldsymbol{\varepsilon}^\top \mathbf{X}^{(j)}\,| \leq \lambda_0\right\}.$$

- Unter der Annahme, dass $\lambda \geq \lambda_0$ können wir sicherstellen, dass wir auf $\mathcal{J}$ den zufälligen Teil des Problems loswerden können.

- Die Diagonalelemente der Grammatrix $\hat{\Sigma}$ bezeichnen wir im folgenden mit $\hat{\sigma}_j := \hat{\Sigma}_{jj} \; \forall j = 1, \ldots, p$

Lemma 2.2

Sei $\hat{\sigma}_j^2 = 1 \; \forall j = 1, \ldots, p$. *Dann gilt für* $t > 0$ *und für* $\lambda_0 := 2\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}$

$$\mathbb{P}[\mathcal{J}] \geq 1 - 2\exp\left(\frac{-t^2}{2}\right).$$

Beweis.

Siehe [BvdG11] Kapitel 6, Beweis zu Lemma 6.2. □

- Aus den beiden Lemmata 2.1 und 2.2 gewinnen wir eine Konsistenzeigenschaft des Lasso

Korollar 2.3 (Konsistenzeigenschaft des Lasso.)

Sei $\hat{\sigma}_j^2 = 1\ \forall j = 1, \ldots, p$. *Für* $t > 0$ *sei der Regularisierungsparameter*

$$\lambda_0 := 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log(p)}{n}},$$

wobei $\hat{\sigma}$ *ein Schätzer für* σ *ist. Dann gilt mit einer Wahrscheinlichkeit von mindestens* $1-\alpha$ *für*

$$\alpha := 2\exp\left(\frac{-t^2}{2}\right) + \mathbb{P}[\hat{\sigma} \leq \sigma],$$

dass

$$\frac{2}{n}\| \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \leq 3\lambda\| \boldsymbol{\beta}^0 \|_1.$$

- Wir schließen, dass die Wahl von λ mit Ordnung $\sqrt{\log(p)/n}$ und die Annahme, dass die $l1$-Norm des wahren $\boldsymbol{\beta}^0$ von kleinerer Ordnung als $\sqrt{n/\log(p)}$ ist, in der Konsistenz des Lasso endet.
- Dabei ist die geeignete Wahl des Schätzers $\hat{\sigma}$ von σ (nicht zu klein aber auch nich zu groß) wichtig, wir denken an den Schätzer $\hat{\sigma} := \mathbf{Y}^\top \mathbf{Y}/n$.
- Mit dem Signal-Rausch-Verhältnis $SNR := \|\mathbf{X}\boldsymbol{\beta}^0\|_2/\sqrt{n}\sigma$ unterliegt der Schätzer $\hat{\sigma}$ der Ungleichung $\sigma \leq \hat{\sigma} \leq c\sigma$, wobei sich die Konstante c gut kontrollieren lässt.
- Wir definieren

$$\boldsymbol{\beta}_{j,S} := \boldsymbol{\beta}_j \mathbb{1}_{\{j \in S\}} \text{ und } \boldsymbol{\beta}_{j,S^c} := \boldsymbol{\beta}_j \mathbb{1}_{\{j \notin S\}}$$

dann gilt $\boldsymbol{\beta} = \boldsymbol{\beta}_S + \boldsymbol{\beta}_{S^c}$.

Lemma 2.4

Auf dem Ereignis $\mathcal{J}$ *gilt für* $\lambda \geq 2\lambda_0$

$$\frac{2}{n}\| \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + \lambda\| \hat{\boldsymbol{\beta}}_{\mathrm{S_0}^c} \|_1 \leq 3\lambda\| \hat{\boldsymbol{\beta}}_{\mathrm{S_0}} - \boldsymbol{\beta}^0_{\mathrm{S_0}} \|_1.$$

Beweis.

Siehe Tafelanschrieb. $\square$

- Es werden weitere Bedingungen an die Designmatrix $\mathbf{X}$ benötigt $\Rightarrow$ Kompatibilität zwischen der l_1-Norm und der l_2-Norm

Definition 2.5 (Kompatibilitätsbedingung)

Wir sagen, dass die Kompatibilitätsbedingung für die Menge S_0 erfüllt ist, wenn für ein $\phi_0 > 0$ und alle $\boldsymbol{\beta}$, die der Ungleichung $\| \boldsymbol{\beta}_{\mathrm{S}_0{}^c} \|_1 \leq 3 \| \boldsymbol{\beta}_{\mathrm{S}_0} \|_1$ genügen, gilt

$$\| \boldsymbol{\beta}_{\mathrm{S}_0} \|_1^2 \leq (\boldsymbol{\beta}^\top \hat{\Sigma} \, \boldsymbol{\beta}) \frac{s_0}{\phi_0^2}.$$

Satz 2.6

Angenommen die aktive Menge S_0 *erfüllt die Ungleichung (2.5), dann gilt auf dem Ereignis* $\mathcal{J}$ *mit* $\lambda \geq 2\lambda_0$

$$\frac{\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \|_1 \leq \frac{4s_0}{\phi_0^2}\lambda^2.$$

Beweis.

Siehe Tafelanschrieb. □

- Kombination des Satzes 2.6 und dem Lemma 2.2 sichert unter bestimmten Bedingungen die Konsistenz des Lasso:

Korollar 2.7

Es sei $\hat{\sigma}_j = 1$ *für alle* $j = 1, \dots, p$ *und die Kompatibilitäsbedingung (2.5) gelte für* S_0. *Für* $t > 0$ *sei der Regularisierungsparameter* $\lambda := 4\hat{\sigma}\sqrt{\frac{t^2+2\log(p)}{n}}$, *dann gilt mit einer Wahrscheinlicheit von mindestens* $1-\alpha$, *wobei*

$$\alpha := 2\exp\left(\frac{-t^2}{2}\right) + \mathbb{P}[\hat{\sigma} \leq \sigma],$$

dass

$$\frac{\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda \| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \|_1 \leq \frac{4s_0}{\phi_0^2}\lambda^2.$$

- Frage: Wann versagt der Lasso?
- Betrachte die Grammatrix $\hat{\Sigma}$ die wir mit Hilfe von Blockmatritzen $\hat{\Sigma}_{ij}$ $i,j = 1,2$ folgendermaßen darstellen können

$$\hat{\Sigma} = \mathbf{X}^\top \mathbf{X}/n = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix},$$

wobei $\hat{\Sigma}_{11} \in \mathbb{R}^{s_0 \times s_0}$.

Satz 2.8 (Inkonsistenz der Variablenauswahl.)

Angenommen $s_0 = 2m + 1 \geq 3$ für ein $m \in \mathbb{N}$ und $p = s_0 + 1$, sodass es genau einen irrelevanten Koeffizienten β_j^0 gibt. Sei $\hat{\Sigma}_{11} = (1 - \rho_1)\boldsymbol{I}_{s_0 \times s_0} + \rho_1 \boldsymbol{J}_{s_0 \times s_0}$ wobei $\boldsymbol{J}_{s_0 \times s_0}$ die $s_0 \times s_0$ Matrix mit nur 1en ist. Sei $\hat{\Sigma}_{12} = \rho_2 \boldsymbol{I}_{s_0 \times 1}$ und $\hat{\Sigma}_{22} = 1$. Gilt

$$-\frac{1}{s_0 - 1} < \rho_1 < -\frac{1}{s_0} \quad \textit{und} \quad 1 + (s_0 - 1)\rho_1 < |\rho_2| < \sqrt{\frac{1 + (s_0 - 1)\rho_1}{s_0}},$$

dann kann die Variablenauswahl des Lasso nicht konsistent sein.

Der Adaptive Lasso

- $\hat{\beta}$ erfüllt nur unter sehr starken Voraussetzungen mit hoher Wahrscheinlichkeit die Oracle-Ungleichung.
- $\hat{\beta}$ regularisiert jeden Koeffizient $\boldsymbol{\beta}_j$ genau gleich, die Beobachtungen $\mathbf{X}^{(j)}$ können jedoch unterschiedlich starken Einfluss haben $\Rightarrow$ unterschiedlich starke Regularisierung der Koeffizienten $\boldsymbol{\beta}_j$ benötigt.
- Der Adaptive Lasso realisiert diese Regularisierung:

$$\hat{\boldsymbol{\beta}}^{(n)}_{\text{adap}} := \arg\min_{\boldsymbol{\beta}} \left\{ \frac{\| \mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta} \|_2^2}{n} + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\, \boldsymbol{\beta}_j \,| \right\},$$

  mit $\tilde{w}_j := 1/|\tilde{\boldsymbol{\beta}}_j|^\gamma$, $\gamma > 0$ wobei $\tilde{\boldsymbol{\beta}}$ ein initialer ($\sqrt{n}$-konsistenter) Schätzer für $\boldsymbol{\beta}^0$ ist (z.B. kleinste Quadrate Schätzer).
- $\tilde{w}$ sollte in Abhängigkeit von den Daten clever gewählt werden und ist $\tilde{\boldsymbol{\beta}}_j = 0$ so setze auch $\hat{\boldsymbol{\beta}}_{\text{adap},j} = 0$.
- $\hat{S}_{\text{adap}} := \{j : \hat{\boldsymbol{\beta}}_{\text{adap},j} \neq 0, j = 1, \ldots, p\}$ ist die aktive Menge des Adaptive Lasso.

- Für geeigneten Wahl von λ_n besitzt der Adaptive Lasso die Oracle Eigenschafte.

Satz 3.1

Gilt $\lambda_n/\sqrt{n} \to 0$ *und* $\lambda_n n^{(\gamma-1)/2} \to \infty$, *dann erfüllt der Adaptive Lasso* $\hat{\beta}^{(n)}_{adap}$ *die folgenden Eigenschaften*

1. *Konsistenz in der Variablenauswahl, d.h.* $\lim_{n\to\infty} \mathbb{P}[\hat{S}_{adap} = \mathrm{S}_0] = 1$
2. *Asymptotische Normailität:* $\sqrt{n}(\hat{\beta}^{(n)}_{adap,\mathrm{S}_0} - \beta^0_{\mathrm{S}_0}) \to \mathcal{N}(0, \sigma^2 \hat{\Sigma}^{-1}_{11})$,

wobei $\hat{\Sigma}_{11} \in \mathbb{R}^{s_0 \times s_0}$.

Beweis.

Siehe [Z06] Kapitel „APPENDIX: PROOFS“. □

Bemerkung 3.2

(1) Der initiale Schätzer $\tilde{\boldsymbol{\beta}}$ muss nicht unbedingt $\sqrt{n}$-konsistent sein, damit der Adaptive Lasso seine guten Eigenschaften behält. Angenommen es existiert eine aufsteigende Folge $\{a_n\}_{n\in\mathbb{N}}$, sodass $a_n \to \infty$ für $n \to \infty$ und $a_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = O_p(1)$, dann gelten die Oracle Eigenschaften aus Satz 3.1 auch noch wenn $\lambda_n = o(\sqrt{n})$ und $a_n^\gamma \lambda_n/\sqrt{n} \to \infty$.

(2) Das von den Daten abhängige Regularisierungsgewicht $\tilde{w}$ ist der Schlüssel in Satz 3.1. Wenn der Stichprobenumfang wächst werden die Gewichte der $\hat{\beta}^{(n)}_{\text{adap},j} = 0$ zu unendlicher Größe aufgeblasen, während die Gewichte der $\hat{\beta}^{(n)}_{\text{adap},j} \neq 0$ gegen eine feste Konstante konvergieren. Also

$$\tilde{w}_j \to \infty \text{ für } j \notin \hat{S}_{\text{adap}},$$
$$\tilde{w}_j \to c_j < \infty \text{ für } j \in \hat{S}_{\text{adap}}.$$

Simulation des Lasso und Adaptive Lasso

- Jetzt geht es weiter mit Simulationen, die die vorgestellten Eigenschaften des Lasso und des Adaptive Lasso veranschaulichen und vergleichen.

Literaturverzeichnis

[BvdG11] PETER BÜHLMANN, SARA VAN DE GEER. *„Statistics for High-Dimensional Data“, Springer*, 2011.

[Z06] HUI ZOU. *„The adaptive lasso and its oracle properties“, Journal of the American Statistical Association,101(476), 1418-1429*, 2006.