Reduced-Rank-Regression

Daniel Herbst

10. Mai 2021

Zusammenfassung

Nachdem in der vorigen Präsentation das multivariate Regressionsmodell mit festen Inputvariablen vorgestellt wurde und bereits Modelle betrachtet wurden, bei denen die Regressionskoeffizienten gewisse (lineare) Nebenbedingungen erfüllen, wollen wir nun eine ähnliche Situation betrachten, bei der wir die Inputvariablen allerdings als zufällig annehmen und den Rang der Regressionskoeffizientenmatrix einschränken. Hierbei handelt es sich dann um das Reduced-Rank-Regressionsmodell, das unter anderem auch einige Techniken der multivariaten Statistik, etwa zur Dimensionsreduktion, verallgemeinert. Im Wesentlichen folgt der Vortrag [Iz08, Kapitel 6.3].

1 Einleitung

Unsere Ausgangssituation ist die Folgende:

$$\mathbf{X} := (X_1, \dots, X_r)^\top \quad \text{und} \quad \mathbf{Y} := (Y_1, \dots, Y_s)^\top,$$

seien Zufallsvektoren mit gemeinsamer Verteilung $\mathbb{P}^{(\mathbf{X},\mathbf{Y})}$, wobei $r, s \in \mathbb{N}$ mit $s \leq r$, und ferner definieren wir folgende Schreibweisen für die entsprechenden Erwartungwerte und Kovarianzmatrizen:

$$\boldsymbol{\mu}_{\mathbf{X}} := \mathbb{E}\mathbf{X}, \quad \boldsymbol{\mu}_{\mathbf{Y}} := \mathbb{E}\mathbf{Y} \quad \text{und} \quad \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{pmatrix} := \Sigma\left(\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}\right).$$

Abgesehen von $\mathbb{P}^{\mathbf{X}} \ll \lambda^r$ und $\mathbb{P}^{\mathbf{Y}} \ll \lambda^s$, d.h. der Stetigkeit der Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$, möchten wir zunächst keine weiteren Annahmen über die Verteilungen von $\mathbf{X}$ und $\mathbf{Y}$ treffen.

2 Klassisches multivariates Regressionsmodell mit zufälliger Inputvariable

Das klassische multivariate Regressionsmodell mit zufälliger Inputvariable ist von folgender Form: Für $\mathbf{X}$ und $\mathbf{Y}$ gelte

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Theta}\mathbf{X} + \mathcal{E}, \tag{2.1}$$

wobei $\boldsymbol{\mu} \in \mathbb{R}^s$ und $\boldsymbol{\Theta} \in \mathbb{R}^{s \times r}$ unbekannte Parameter seien sowie $\mathcal{E}$ ein (nicht beobachtbarer) s-dimensionaler zufälliger Fehler ist mit Erwartungswert $\mathbb{E}\mathcal{E} = 0$ und Kovarianzmatrix $\boldsymbol{\Sigma}_{\mathcal{E}\mathcal{E}}$. Zudem werden $\mathbf{X}$ und $\mathcal{E}$ als unabhängig vorausgesetzt.

Ausgehend von der Situation aus der Einleitung wollen wir nun optimale $\boldsymbol{\mu}$ und $\boldsymbol{\Theta}$ finden in dem Sinne, dass

$$W(\boldsymbol{\mu}, \boldsymbol{\Theta}) := \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})^\top] \in \mathbb{R}^{s \times s}$$

in der Spektralnorm

$$\|\mathbf{A}\|_2 := \max_{\|\mathbf{v}\|_2 = 1} \|\mathbf{A}\mathbf{v}\|_2, \quad \mathbf{A} \in \mathbb{R}^{s \times s}$$

minimiert wird.

Für symmetrische, positiv semidefinite Matrizen $\mathbf{A}$ lässt sich (etwa mit dem Spektralsatz) zeigen, dass

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{v}\|_2 = 1} \mathbf{v}^\top \mathbf{A}\mathbf{v} = \lambda_{\max}(\mathbf{A}),$$

wobei $\lambda_{\max}(\mathbf{A})$ der größte Eigenvektor von $\mathbf{A}$ sei. Da $W(\boldsymbol{\mu}, \boldsymbol{\Theta})$ für alle möglichen $\boldsymbol{\mu}$, $\boldsymbol{\Theta}$ symmetrisch positiv semidefinit ist, werden wir diese Darstellung der Spektralnorm auch zur Minimierung von $\|W(\boldsymbol{\mu}, \boldsymbol{\Theta})\|_2$ benutzen.

Satz 2.1. *Seien $\mathbf{X}$ und $\mathbf{Y}$ wie in der Einleitung und zudem $\boldsymbol{\Sigma}_{\mathbf{XX}}$ nichtsingulär, so ist*

$$\operatorname*{argmin}_{(\boldsymbol{\mu}, \boldsymbol{\Theta})} \|W(\boldsymbol{\mu}, \boldsymbol{\Theta})\|_2 = (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}) =: (\boldsymbol{\mu}_{\min}, \boldsymbol{\Theta}_{\min})$$

mit

$$W(\boldsymbol{\mu}_{\min}, \boldsymbol{\Theta}_{\min}) = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}.$$

Ferner gilt für $\mathcal{E} := \mathbf{Y} - \boldsymbol{\mu}_{\min} - \boldsymbol{\Theta}_{\min}\mathbf{X}$, dass

$$\mathbb{E}\mathcal{E} = 0 \quad \textit{und} \quad \mathbb{E}[\mathcal{E}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^\top] = 0,$$

d.h. $\mathbf{X}$ und $\mathcal{E}$ sind unkorreliert.

Beweis. Mit $\mathbf{X}_c := \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}$, $\mathbf{Y}_c := \mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}$ gilt für alle $\boldsymbol{\Theta} \in \mathbb{R}^{s \times r}$, $\boldsymbol{\mu} \in \mathbb{R}^s$:

$$\begin{aligned}
W(\boldsymbol{\mu}, \boldsymbol{\Theta}) &= \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})^\top] \\
&= \mathbb{E}[(\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}))(\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}))^\top] \\
&= \mathbb{E}[\mathbf{Y}_c\mathbf{Y}_c^\top - \mathbf{Y}_c\mathbf{X}_c^\top\boldsymbol{\Theta}^\top + \mathbf{Y}_c(\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})^\top \\
&\qquad - \boldsymbol{\Theta}\mathbf{X}_c\mathbf{Y}_c^\top + \boldsymbol{\Theta}\mathbf{X}_c\mathbf{X}_c^\top\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\mathbf{X}_c(\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})^\top \\
&\qquad + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})\mathbf{Y}_c^\top - (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})\mathbf{X}_c^\top\boldsymbol{\Theta}^\top \\
&\qquad + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})(\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})^\top] \\
&= \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{\mathbf{XY}} + \boldsymbol{\Theta}\boldsymbol{\Sigma}_{\mathbf{XX}}\boldsymbol{\Theta}^\top \\
&\qquad + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})(\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})^\top \\
&= \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}} \\
&\qquad + (\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1/2} - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{\mathbf{XX}}^{1/2})(\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1/2} - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{\mathbf{XX}}^{1/2})^\top \\
&\qquad + (\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})(\boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu})^\top,
\end{aligned}$$

und mit

$$A(\boldsymbol{\Theta}) := \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1/2} - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{\mathbf{XX}}^{1/2} \quad , \quad B(\boldsymbol{\mu}, \boldsymbol{\Theta}) := \boldsymbol{\mu}_\mathbf{Y} - \boldsymbol{\Theta}\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}$$

gilt für $\mathbf{v} \in \mathbb{R}^s$, $\|\mathbf{v}\|_2 = 1$:

$$\begin{aligned}\mathbf{v}^\top W(\boldsymbol{\mu}, \boldsymbol{\Theta})\mathbf{v} &= \mathbf{v}^\top(\boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})\mathbf{v} + \mathbf{v}^\top A(\boldsymbol{\Theta})^\top A(\boldsymbol{\Theta})\mathbf{v} + \mathbf{v}^\top B(\boldsymbol{\mu}, \boldsymbol{\Theta})^\top B(\boldsymbol{\mu}, \boldsymbol{\Theta})\mathbf{v} \\ &= \mathbf{v}^\top(\boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})\mathbf{v} + \|A(\boldsymbol{\Theta})\mathbf{v}\|_2 + \|B(\boldsymbol{\mu}, \boldsymbol{\Theta})\mathbf{v}\|_2 \\ &\geq \mathbf{v}^\top(\boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})\mathbf{v}\end{aligned}$$

und daher ist

$$\|W(\boldsymbol{\mu}, \boldsymbol{\Theta})\|_2 \geq \|\boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\|_2,$$

mit Gleichheit falls $A(\boldsymbol{\Theta}) = B(\boldsymbol{\mu}, \boldsymbol{\Theta}) = 0$, was der Fall ist, wenn

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{X}} =: \boldsymbol{\mu}_{\min} \quad \text{und} \quad \boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} =: \boldsymbol{\Theta}_{\min}.$$

Man erhält nun

$$\begin{aligned}\mathcal{E} &= \mathbf{Y} - \boldsymbol{\mu}_{\min} - \boldsymbol{\Theta}_{\min}\mathbf{X} \\ &= \mathbf{Y} - (\boldsymbol{\mu}_{\mathbf{Y}} - (\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1})\boldsymbol{\mu}_{\mathbf{X}}) - (\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1})\mathbf{X} \\ &= \mathbf{Y}_c - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{X}_c\end{aligned}$$

und es gilt offensichtlich $\mathbb{E}\mathcal{E} = 0$ sowie

$$\mathbb{E}[\mathcal{E}\mathbf{X}_c^\top] = \mathbb{E}[(\mathbf{Y}_c - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{X}_c)\mathbf{X}_c^\top] = \boldsymbol{\Sigma}_{\mathbf{YX}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XX}} = 0,$$

d.h. $\mathbf{X}$ und $\mathcal{E}$ sind unkorreliert. □

3 Reduced-Rank-Regressionsmodell

Wir betrachten nun das folgende Regressionsmodell

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{C}\mathbf{X} + \mathcal{E}, \tag{3.1}$$

bei dem $\boldsymbol{\mu} \in \mathbb{R}^s$ und $\mathbf{C} \in \mathbb{R}^{s\times r}$ unbekannte Parameter sind sowie $\mathcal{E}$ ein nicht beobachtbarer, zufälliger Fehler mit $\mathbb{E}\mathcal{E} = 0$ und Kovarianzmatrix $\boldsymbol{\Sigma}_{\mathcal{E}\mathcal{E}}$, wobei $\mathbf{X}$ und $\mathcal{E}$ unabhängig seien.

Im Gegensatz zu (2.1), wo wir über $\boldsymbol{\Theta}$ keine weiteren Annahmen treffen wollten, soll es nun zusätzlich ein gewisses t geben mit

$$\operatorname{rk}\mathbf{C} \leq t \leq s.$$

In diesem Fall gibt es Matrizen $\mathbf{A} \in \mathbb{R}^{s\times t}$ und $\mathbf{B} \in \mathbb{R}^{t\times r}$ mit $\mathbf{C} = \mathbf{AB}$, d.h. das Modell (3.1) lässt sich nun schreiben als

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{ABX} + \mathcal{E}, \tag{3.2}$$

wobei $\boldsymbol{\mu}$, $\mathbf{A}$ und $\mathbf{B}$ nun die unbekannten Parameter sind.

Ein Kleinste-Quadrate-Kriterium

Auch hier möchten wir für $\mathbf{X}$ und $\mathbf{Y}$ wie in Satz 2.1 $\boldsymbol{\mu} \in \mathbb{R}^s$, $\mathbf{A} \in \mathbb{R}^{s\times t}$ und $\mathbf{B} \in \mathbb{R}^{t\times r}$ finden, die $\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX}$ auf eine gewisse Art und Weise minimieren, gehen nun aber etwas anders vor: Für eine symmetrische, positiv definite Gewichtsmatrix $\boldsymbol{\Gamma} \in \mathbb{R}^{s\times s}$ wollen wir nun

$$W_{t,\boldsymbol{\Gamma}}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) := \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})^\top\boldsymbol{\Gamma}(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})]$$

über $\boldsymbol{\mu}$, $\mathbf{A}$ und $\mathbf{B}$ minimieren.

Satz 3.1. *Seien* $\mathbf{X}$ *und* $\mathbf{Y}$ *wie in der Einleitung,* $\mathbf{\Sigma_{XX}}$ *nichtsingulär und* $1 \leq t \leq s$. *Dann ist*

$$\operatorname*{argmin}_{(\boldsymbol{\mu},\mathbf{A},\mathbf{B})} W_{t,\mathbf{\Gamma}}(\boldsymbol{\mu},\mathbf{A},\mathbf{B}) = (\boldsymbol{\mu}^{(t)}_{\min}, \mathbf{A}^{(t)}_{\min}, \mathbf{B}^{(t)}_{\min})$$

mit

$$\begin{aligned}
\mathbf{A}^{(t)}_{\min} &:= \mathbf{\Gamma}^{-1/2}\mathbf{U}_t && \in \mathbb{R}^{s\times t}\\
\mathbf{B}^{(t)}_{\min} &:= \mathbf{U}_t^\top \mathbf{\Gamma}^{1/2}\mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}^{-1}} && \in \mathbb{R}^{t\times r}\\
\boldsymbol{\mu}^{(t)}_{\min} &:= \boldsymbol{\mu_Y} - \mathbf{A}^{(t)}_{\min}\mathbf{B}^{(t)}_{\min}\boldsymbol{\mu_X} && \in \mathbb{R}^{s},
\end{aligned}$$

wobei $\mathbf{U}_t := (\mathbf{u}_1, \dots, \mathbf{u}_t) \in \mathbb{R}^{s\times t}$ *und* $\mathbf{u}_1, \dots, \mathbf{u}_t$ *orthonormale Eigenvektoren zu den Eigenwerten* $\lambda_1 \geq \cdots \geq \lambda_t \geq 0$ *von* $\mathbf{\Gamma}^{1/2}\mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}^{-1}}\mathbf{\Sigma_{XY}}\mathbf{\Gamma}^{1/2}$ *seien. Ferner gilt*

$$W_{t,\mathbf{\Gamma}}(\boldsymbol{\mu}^{(t)}_{\min}, \mathbf{A}^{(t)}_{\min}, \mathbf{B}^{(t)}_{\min}) = \operatorname{tr}(\mathbf{\Sigma_{YY}})\operatorname{tr}(\mathbf{\Gamma}) - \sum_{i=1}^{t}\lambda_i.$$

Für den Beweis benötigen wir zunächst einen Hilfssatz aus der Linearen Algebra:

Lemma 3.2 (Satz von Eckart-Young)**.** *Seien* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ *und* $b := \operatorname{rk}\mathbf{B} \leq \operatorname{rk}\mathbf{A} =: r$, *und sei* $\lambda_j(\mathbf{C})$ *für eine reelle symmetrischen Matrix* $\mathbf{C}$ *deren* j*-größter Eigenwert. Dann gilt:*

$$\lambda_j((\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\top) \geq \lambda_{j+b}(\mathbf{A}\mathbf{A}^\top)$$

mit Gleichheit für

$$\mathbf{A}_b := \sum_{i=1}^{b}\lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^\top,$$

wobei $\lambda_i := \lambda_i(\mathbf{A}\mathbf{A}^\top)$ *und* $\mathbf{u}_i$ *bzw.* $\mathbf{v}_i$ *jeweils orthonormale Eigenvektoren von* $\mathbf{A}\mathbf{A}^\top$ *bzw.* $\mathbf{A}^\top\mathbf{A}$ *zu* λ_i *seien.*

Beweis. Folgt (unter Benutzung der Singulärwertzerlegung von $\mathbf{A}$) etwa direkt aus [BZ21, Satz 4.6]. □

Bemerkung. *Insbesondere zeigt Lemma 3.2, dass das dort definierte* $\mathbf{A}_b$ *den Abstand von* $\mathbf{A}$ *unter allen Rang-b-Matrizen in* $\mathbb{R}^{m\times n}$ *in der Spektralnorm minimiert.*

Beweis von Satz 3.1. Seien $\boldsymbol{\mu} \in \mathbb{R}^s$, $\mathbf{A} \in \mathbb{R}^{s\times t}$, $\mathbf{B} \in \mathbb{R}^{t\times r}$ und setze $\mathbf{X}_c := \mathbf{X} - \boldsymbol{\mu_X}$, $\mathbf{Y}_c = \mathbf{Y} - \boldsymbol{\mu_Y}$, $\mathbf{C} := \mathbf{A}\mathbf{B}$. Dann gilt:

$$\begin{aligned}
W_{t,\mathbf{\Gamma}}(\boldsymbol{\mu},\mathbf{C}) &= \mathbb{E}[(\mathbf{Y}-\boldsymbol{\mu}-\mathbf{C}\mathbf{X})^\top\mathbf{\Gamma}(\mathbf{Y}-\boldsymbol{\mu}-\mathbf{C}\mathbf{X})]\\
&= \mathbb{E}[(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c + (\boldsymbol{\mu_Y} - \mathbf{C}\boldsymbol{\mu_X} - \boldsymbol{\mu}))^\top\mathbf{\Gamma}(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c + (\boldsymbol{\mu_Y} - \mathbf{C}\boldsymbol{\mu_X} - \boldsymbol{\mu}))]\\
&= \mathbb{E}[(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c)^\top\mathbf{\Gamma}(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c)]\\
&\quad + \underbrace{(\boldsymbol{\mu_Y} - \mathbf{C}\boldsymbol{\mu_X} - \boldsymbol{\mu})^\top\mathbf{\Gamma}(\boldsymbol{\mu_Y} - \mathbf{C}\boldsymbol{\mu_X} - \boldsymbol{\mu})}_{\geq 0 \quad (\text{da } \mathbf{\Gamma} \text{ symmetrisch positiv definit})} \qquad (3.3)
\end{aligned}$$

Unser Ziel ist es nun, (3.3) zu minimieren, indem wir zunächst den ersten Term nach $\mathbf{C}$ minimieren. Für das so festgelegte $\mathbf{C}_{\min}$ lässt sich $\boldsymbol{\mu}_{\min} := \boldsymbol{\mu_Y} - \mathbf{C}_{\min}\boldsymbol{\mu_X}$ aber immer noch so

wählen, dass der zweite Term zu 0 wird, was impliziert, dass die so bestimmte Kombination $(\boldsymbol{\mu}_{\min}, \mathbf{C}_{\min})$ tatsächlich $W_{t,\boldsymbol{\Gamma}}(\boldsymbol{\mu}, \mathbf{C})$ minimiert.

Mit $\boldsymbol{\Sigma}^*_{\mathbf{XX}} := \boldsymbol{\Sigma}_{\mathbf{XX}}$, $\boldsymbol{\Sigma}^*_{\mathbf{YY}} := \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YY}}\boldsymbol{\Gamma}^{1/2}$, $\boldsymbol{\Sigma}^*_{\mathbf{YX}} := \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YX}}$, $\boldsymbol{\Sigma}^*_{\mathbf{XY}} := \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Gamma}^{1/2}$ und $\mathbf{C}^* := \boldsymbol{\Gamma}^{1/2}\mathbf{C}$ ist

$$\begin{aligned}
\mathbb{E}[(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c)^\top \boldsymbol{\Gamma}(\mathbf{Y}_c - \mathbf{C}\mathbf{X}_c)] &= \mathbb{E}[\mathbf{Y}_c^\top \boldsymbol{\Gamma}\mathbf{Y}_c - \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma}\mathbf{Y}_c - \mathbf{Y}_c^\top \boldsymbol{\Gamma}\mathbf{C}\mathbf{X}_c + \mathbf{X}_c^\top \mathbf{C}^\top \boldsymbol{\Gamma}\mathbf{C}\mathbf{X}_c] \\
&= \operatorname{tr}(\boldsymbol{\Sigma}^*_{\mathbf{YY}} - \mathbf{C}^*\boldsymbol{\Sigma}^*_{\mathbf{XY}} - \boldsymbol{\Sigma}^*_{\mathbf{YX}}\mathbf{C}^{*\top} + \mathbf{C}^*\boldsymbol{\Sigma}^*_{\mathbf{XX}}\mathbf{C}^{*\top}) \\
&= \operatorname{tr}(\boldsymbol{\Sigma}^*_{\mathbf{YY}} - \boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1}_{\mathbf{XX}}\boldsymbol{\Sigma}^*_{\mathbf{XY}}) \\
&\quad + \operatorname{tr}(\ (\mathbf{C}^*\boldsymbol{\Sigma}^{*1/2}_{\mathbf{XX}} - \boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}}) \\
&\qquad \cdot (\mathbf{C}^*\boldsymbol{\Sigma}^{*1/2}_{\mathbf{XX}} - \boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}})^\top\ ). \qquad (3.4)
\end{aligned}$$

Sei nun

$$k := \operatorname{rk} \boldsymbol{\Sigma}_{\mathbf{YX}} = \operatorname{rk} \boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}}$$

und seien $\mathbf{u}_i$ orthonormale Eigenvektoren zum jeweils i-größten Eigenwert λ_i von

$$(\boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}})(\boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}})^\top = \boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1}_{\mathbf{XX}}\boldsymbol{\Sigma}^*_{\mathbf{XY}} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Gamma}^{1/2}$$

für $1 \le i \le s$ sowie

$$\mathbf{v}_i := \lambda_i^{-1/2}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}}\boldsymbol{\Sigma}^*_{\mathbf{XY}}\mathbf{u}_i = \lambda_i^{-1/2}\boldsymbol{\Sigma}^{-1/2}_{\mathbf{XX}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Gamma}^{1/2}\mathbf{u}_i, \quad 1 \le i \le k$$

zugehörige orthonormale Eigenvektoren von $(\boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}})^\top(\boldsymbol{\Sigma}^*_{\mathbf{YX}}\boldsymbol{\Sigma}^{*-1/2}_{\mathbf{XX}})$, wobei wir für $i > k$ die restlichen $\mathbf{v}_i$ so definieren, dass wir eine Orthonormalbasis von $\mathbb{R}^r$ erhalten. Dann lässt sich der Satz von Eckart-Young (Lemma 3.2) wie folgt anwenden: Der zweite Term von (3.4) wird minimal für

$$\mathbf{C}^*\boldsymbol{\Sigma}^{*1/2}_{\mathbf{XX}} = \boldsymbol{\Gamma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}_{\mathbf{XX}} = \sum_{i=1}^{\min\{k,t\}} \lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^\top$$

(denn die Spur ist als Summe der Eigenwerte monoton steigend in jedem Eigenwert), also folgt, dass

$$\begin{aligned}
\mathbf{C}^{(t)}_{\min} &:= \boldsymbol{\Gamma}^{-1/2}\Bigg(\sum_{i=1}^{\min\{k,t\}} \lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^\top\Bigg)\boldsymbol{\Sigma}^{-1/2}_{\mathbf{XX}} \\
&= \boldsymbol{\Gamma}^{-1/2}\Bigg(\sum_{i=1}^{\min\{k,t\}} \mathbf{u}_i\mathbf{u}_i^\top\Bigg)\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}} \\
&= \boldsymbol{\Gamma}^{-1/2}\Bigg(\sum_{i=1}^{t} \mathbf{u}_i\mathbf{u}_i^\top\Bigg)\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}
\end{aligned}$$

den ersten Summand von (3.3) über alle Matrizen $\mathbf{C} \in \mathbb{R}^{s\times r}$ von Rang $\le t$ minimiert. Daher wird $W_{t,\boldsymbol{\Gamma}}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ nun minimal für

$$\begin{aligned}
\mathbf{A}^{(t)}_{\min} &= \boldsymbol{\Gamma}^{-1/2}\mathbf{U}_t && \in \mathbb{R}^{s\times t} \\
\mathbf{B}^{(t)}_{\min} &= \mathbf{U}_t^\top\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}} && \in \mathbb{R}^{t\times r} \\
\boldsymbol{\mu}^{(t)}_{\min} &= \boldsymbol{\mu}_{\mathbf{Y}} - \mathbf{A}^{(t)}_{\min}\mathbf{B}^{(t)}_{\min}\boldsymbol{\mu}_{\mathbf{X}} && \in \mathbb{R}^{s},
\end{aligned}$$

wobei $\mathbf{U}_t := (\mathbf{u}_1, \ldots, \mathbf{u}_t) \in \mathbb{R}^{s\times t}$. Mit (3.3), (3.4) ist das dadurch angenommene Minimum dann

$$\begin{aligned} W_{t,\mathbf{\Gamma}}(\boldsymbol{\mu}_{\min}^{(t)}, \mathbf{A}_{\min}^{(t)}, \mathbf{B}_{\min}^{(t)}) &= \operatorname{tr}(\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{\mathbf{YY}}\mathbf{\Gamma}^{1/2} - \mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Gamma}^{1/2}) \\ &\quad + \operatorname{tr}((\sum_{i=t+1}^{k} \lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^\top)(\sum_{i=t+1}^{k} \lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^\top)^\top) \\ &= \operatorname{tr}(\mathbf{\Sigma}_{\mathbf{YY}})\operatorname{tr}(\mathbf{\Gamma}) - \operatorname{tr}(\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Gamma}^{1/2}) + \sum_{i=t+1}^{k} \lambda_i \\ &= \operatorname{tr}(\mathbf{\Sigma}_{\mathbf{YY}})\operatorname{tr}(\mathbf{\Gamma}) - \sum_{i=1}^{t} \lambda_i. \end{aligned}$$

□

Bemerkung. *Für $t = s$ ist $\sum_{i=1}^{s} \mathbf{u}_i\mathbf{u}_i^\top = \mathbf{I}_s \in \mathbb{R}^{s\times s}$, also $\mathbf{C}_{\min}^{(t)} = \mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1} = \mathbf{\Theta}$ und damit sind in diesem Fall die Vorgehensweisen aus Satz 2.1 und Satz 3.1 äquivalent.*

Spezialfälle

Für spezielle Wahlen von $\mathbf{\Gamma}$ und $\mathbf{X}, \mathbf{Y}$ können wir nun wie angekündigt mit der Reduced-Rank-Regression einige klassische Methoden der multivariaten Statistik zusammenfassen, die wir teilweise auch noch in diesem Seminar kennenlernen werden:

- Im Fall $\mathbf{X} = \mathbf{Y}$ und $\mathbf{\Gamma} = \mathbf{I}_s$ erhalten wir die *Hauptkomponentenanalyse (principal component analysis, PCA)* (Vortrag 6).
- Für $\mathbf{\Gamma} = \mathbf{\Sigma}_{\mathbf{YY}}^{-1}$ erhalten wir die *kanonische Korrelationsanalyse* (Vortrag 7).
- Nutzen wir das Setup der kanonischen Korrelationsanalyse für einen $\{0,1\}$-wertigen Vektor $\mathbf{Y}$, der Gruppenzugehörigkeiten modelliert, erhalten wir die *lineare Diskriminanzanalyse* (Vortrag 8).

Stichprobenschätzung bei der Reduced-Rank-Regression

Um Satz 3.1 zu verwenden, werden $\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\mu}_{\mathbf{Y}}$ und $\mathbf{\Sigma}_{\mathbf{XX}}.\mathbf{\Sigma}_{\mathbf{XY}}, \mathbf{\Sigma}_{\mathbf{YX}}, \mathbf{\Sigma}_{\mathbf{YY}}$ benötigt, die in der Regel natürlich unbekannt sind. Stattdessen liegt häufig eine Stichprobe von $n \in \mathbb{N}$ Beobachtungen

$$\mathcal{D} := \{(x_j, y_j) \mid 1 \le i \le n\}$$

vor, die wir als unabhängige Realisierungen von r- bzw. s-dimensionalen Zufallsvektoren $\mathbf{X}$ bzw. $\mathbf{Y}$ (wie in der Einleitung) auffassen.

Damit schätzen wir $\boldsymbol{\mu}_{\mathbf{X}}$ und $\boldsymbol{\mu}_{\mathbf{Y}}$ auf naheliegende Weise:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{X}} := \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j = \bar{x}_n, \quad \widehat{\boldsymbol{\mu}}_{\mathbf{Y}} := \frac{1}{n}\sum_{j=1}^{n} \mathbf{y}_j = \bar{y}_n \tag{3.5}$$

Für die Kovarianzmatrizen definieren wir zunächst

$$\mathbf{x}_{cj} := \mathbf{x}_j - \hat{x}_n, \quad \mathbf{y}_{cj} := \mathbf{y}_j - \hat{y}_n, \quad \mathcal{X}_c := (\mathbf{x}_{c1}, \ldots, \mathbf{x}_{cn}) \in \mathbb{R}^{r\times n}, \quad \mathcal{Y}_c := (\mathbf{y}_{c1}, \ldots, \mathbf{y}_{cn}) \in \mathbb{R}^{s\times n}$$

und schätzen diese dann wie folgt:

$$\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} := \frac{1}{n}\mathcal{X}_c\mathcal{X}_c^\top \in \mathbb{R}^{r\times r}, \quad \widehat{\mathbf{\Sigma}}_{\mathbf{YY}} := \frac{1}{n}\mathcal{Y}_c\mathcal{Y}_c^\top \in \mathbb{R}^{s\times s},$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{YX}} := \frac{1}{n}\mathcal{Y}_c\mathcal{X}_c^\top = \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}^\top \in \mathbb{R}^{s\times r}. \tag{3.6}$$

Die entsprechenden Größen aus Satz 3.1 können wir nun einfach schätzen, indem wir die Schätzungen für die Erwartungswerte und Kovarianzmatrizen aus (3.5), (3.6) verwenden:

$$\widehat{\mathbf{C}}_{\min}^{(t)} := \mathbf{\Gamma}^{-1/2}\left(\sum_{i=1}^{t}\widehat{\mathbf{u}}_i\widehat{\mathbf{u}}_i^\top\right)\mathbf{\Gamma}^{1/2}\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1},$$

wobei $\widehat{\mathbf{u}}_i$, $1 \le i \le t$ orthonormale Eigenvektoren zu den je i-größten Eigenwerten λ_i, $1 \le i \le t$ von $\mathbf{\Gamma}^{1/2}\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{XY}}\mathbf{\Gamma}^{1/2}$ seien. $\mathbf{\Theta}_{\min}$ aus Satz 2.1 wird analog durch

$$\widehat{\mathbf{\Theta}} := \widehat{\mathbf{\Sigma}}_{\mathbf{YX}}\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}$$

geschätzt.

Es kann nun natürlich auch passieren, dass $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ nicht invertierbar ist oder $\kappa(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}})$ zumindest so groß ist, dass Invertieren schwierig wird. In diesem Fall kann man etwa stattdessen die Pseudoinverse von $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ verwenden oder, wie [Iz08, Seite 182] nahelegt, durch eine kleine Veränderung der Diagonaleinträge erreichen, dass man eine (besser) invertierbare Matrix erhält:

$$\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{\delta} := \frac{1}{n}(\mathcal{X}_c\mathcal{X}_c^\top + \delta\mathbf{I}_r)$$

für ein $\delta > 0$. In [Iz08, Kapitel 6.3.4] wird etwa erklärt, wie man δ in der Praxis sinnvoll wählen könnte.

Effektive Dimension der Regression

Letztlich bleibt noch die Frage zu beantworten, wie man die sogenannte *effektive Dimension* t bei der Reduced-Rank-Regression geeignet wählen kann. Dafür beachte man, dass man in der Situation von Satz 3.1, wenn der Rang der Koeffizientenmatrix $t = t_0$ zu $t = t_1$, $t_0 < t_1$ vergrößert wird, eine Verringerung von $W_{t,\mathbf{\Gamma}}(\boldsymbol{\mu}_{\min}^{(t)}, \mathbf{A}_{\min}^{(t)}, \mathbf{B}_{\min}^{(t)})$ der Höhe

$$W_{t_0,\mathbf{\Gamma}}(\boldsymbol{\mu}_{\min}^{(t_0)}, \mathbf{A}_{\min}^{(t_0)}, \mathbf{B}_{\min}^{(t_0)}) - W_{t_1,\mathbf{\Gamma}}(\boldsymbol{\mu}_{\min}^{(t_1)}, \mathbf{A}_{\min}^{(t_1)}, \mathbf{B}_{\min}^{(t_1)}) = \sum_{i=t_0+1}^{t_1}\lambda_i$$

bewirkt. Hiermit könnte man in der Praxis (wo man dann natürlich mit $\widehat{\lambda}_i$ arbeitet) untersuchen, bis wann es sich „lohnt“, t zu vergrößern, was sich, wie in [Iz08, Seite 185] beschrieben wird, mit Tests, ob $\operatorname{rk}\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Gamma}^{1/2} = \operatorname{rk}\mathbf{\Sigma}_{\mathbf{YX}}$ kleiner als eine festgelegte Zahl ist, überprüfen ließe. Ferner wird in [Iz08, Kapitel 6.3.4] ein anderer Algorithmus vorgestellt, mit dem die effektive Dimension t bestimmt werden kann.

Literatur

[Iz08] Izenman, Alan Julian. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics, Springer-Verlag New York, 2008.

[BZ21] Bandeira, Afonso S. und Zhivotovskiy, Nikita. *Lecture Notes for Mathematics of Machine Learning*. 2021. `https://metaphor.ethz.ch/x/2021/fs/401-2684-00L/sc/Math_of_ML_Lecture_Notes.pdf`