Anwendungen von (orthogonaler) Projektion in Statistik

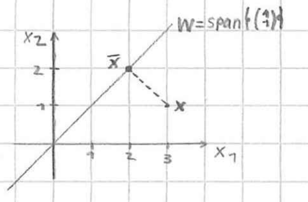
Approximation in Subspace

aus geometrischer Sicht

I. Mittelwert-Berechnung

Grob gesagt kommen orthogon. Projektionen vor allem bei linearen Schätzern vor. Erwartungswerte sind meistens orth. Projektionen auf Raum von ZV oder messbaren Fkt. Einfachster Fall ist MW-Berechnung

Bsp. 2D-Vektor mit zwei Werten $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, $\bar{x} = proj_W(x) = \frac{\langle x, (1,1) \rangle}{\langle (1,1),(1,1) \rangle} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{3 \cdot 1 + 1 \cdot 1}{1^2 + 1^2} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \overset{\text{arithm. Rechnung}}{=} \text{für } x_1, x_2 \ \frac{1}{2}(3+1) = 2$



MW ist quasi bester linearer Schätzer für jede Komponente von x

Durch $\perp$ Projektion auf Winkelhalb. $W = span\{(1,1)\}$ wird jeder x-Komponente selber neuer Wert (MW) zugewiesen

$span\{\mathbb{1}_n\}$ ist allgem. Menge aller konstanten Vektoren in $\mathbb{R}^n$ mit identisch Einträgen in allen Dimensionen

Mittelwert von $x \in \mathbb{R}^n$ ist Projektion auf diesen UVR $\hat{=}$ beste konstante Approximation des Vektors x

Statistisch als Optimierungsproblem: $\bar{x} = \underset{x}{argmin} \underbrace{\sum_{i=1}^{n} (x_i - x)^2}_{\text{Least squares}}$ eng verknüpft mit niedrigstem Abstand: $\bar{x} = proj_W(x) = argmin_{w \in W} \|x - w\|$

II. Kleinste Quadrate Methode (KQ, (O)LS "(Ordinary) Least Squares")

Problem: Daten $(y_1, x_1), \dots, (y_n, x_n) \in \mathbb{R}^{p+1}$ — n Beobachtungen von $(y_i; x_i)$ wobei p-verschiedene x-Variablen, also p+1 erklärende Variablen bzw. p + Intercept

Daten folgen Modell mit unbekanntem Parameter β, also $y_i = x_i^T \beta + \underbrace{\varepsilon_i}_{\text{Zufallsfehler, sonst unrealistisch}}$ $i = 1, \dots, n$

$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$ Jede Zeile ist ein Beobachtungsvektor

Das LGS $X\beta = y$ soll uns den passenden Parametervektor β liefern, aber es gibt i.d.R. keine Lösung — deutlich mehr Beobachtungen als unbekannte Parameter (Zeilen ≫ Spalten von X)

$\beta = X^{-1} y$ ↯ weil X nicht invertierbar

$$n \text{ Zeilen} \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}}_{p+1 \text{ Spalten}} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta \in \mathbb{R}^{p+1}} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Denn:
Wir können nur den Abstand zwischen $X\beta$ und y möglichst klein machen, um beste Approximation von $X\beta$ an die Beobachtungen (y) zu finden

$\hat{\beta} = argmin_\beta \|X\beta - y\|$ $\|y - X\hat{\beta}\| \leq \|y - X\beta\|$ für alle mögl. Vektoren β

Äquivalent ist Minimierung von $\|X\beta - y\|^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \dots \varepsilon_n^2$ $(\varepsilon = X\beta - y)$

Ableiten v. Absolutbetrag ist schwierig

$\|\ \|$ Norm bzw. Betrag in 1D, weil $\sum_i (y_i - \hat{y}_i)$ kann = 0 sein trotz krassen Abweichungen nach oben u. unten, die sich in Summe canceln

Lösung: Suchen $\hat{y} = X\hat{\beta} \in col(X)$ mit minimalem Abstand zu y. Dies ist die eindeutige $\perp$ Projektion von y auf $col(X)$. Ermitteln dann $\hat{\beta}$ s.d. $X\hat{\beta} = \hat{y}$

$\hat{y} = proj_{col(X)} y$ $\quad \hat{y} \in \mathbb{R}^n, \hat{\beta} \in \mathbb{R}^{p+1}$

i) 1. Projiziere y orthogonal auf $col(X)$ 2. Löse das LGS $X\hat{\beta} = \hat{y}$ nach $\hat{\beta}$ Achtung! $\hat{y}$ ist eindeutig, $\hat{\beta}$ aber leider nicht zwingend

X muss orthogonale Spalten haben. Falls nein → Gram-Schmidt vorher (X muss aber linear unabh. Spalten haben!)

ii) Projektion durchzuführen ist sogar optional, denn Lsg.menge von KQ-Problem $\hat{=}$ Lösung(en) der Normalgleichungen $(X^T X)\hat{\beta} = X^T y$

Spalten von X müssen hierfür "nur" linear unabh. sein, denn genau dann ist $X^T X$ invertierbar $\Leftrightarrow \hat{\beta} = (X^T X)^{-1} X^T y$

$X^T X$ inv.bar $\Rightarrow X^T X \vec{v} = 0$ hat nur triviale Lsg. $\vec{v} = \vec{0}$, $X^T X v = 0 \Leftrightarrow (v^T X^T) X v = v^T 0 = 0 \Leftrightarrow \|Xv\|^2 = 0 \Rightarrow Xv = 0$ $X^T X$ immer quadr. & symmetr.

↳ X inv.bar, voller Rang, $Xv = 0$ hat nur triviale Lsg.

Falls Spalten linear abh. (oder bei starker Multikollinearität annähernd abh., was zu numerisch instabiler Inverse $(X^T X)^{-1}$ führt), so braucht es andere Ansätze (alle mit ihren Problemchen) wie etwa SVD-basierte Pseudo-Inverse, Regularisierung via Ridge Regression (!), PCA zur Datenreduktion auf orthogonale Dimensionen, Variablen-Entfernung etc.

iii) Weitere äquivalente Lösungen, die aber siehe MLE verzerrte Schätzer für Fehlervarianz liefern, oder aus anderen Gründe selten zum Einsatz kommen

Warum geht es eigentlich? Bewertung der Genauigkeit von Vorhersage-/Erklärmodellen, insbes. bei Modellierung linearer Zusammenhänge

Modellqualität wird typischerweise mit "Mean Squared Error" (MSE) bzw. dessen Wurzel "Root" (RMSE) beurteilt

$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Diesen mittleren quadratischen Fehler gilt es zu minimieren!

$= \frac{1}{n} \|y - \hat{y}\|^2$

Quadr. Abweichung sorgt für höhere Gewichtung hoher Abweichungen vom Vorhersagewert, aber es ist mathematisch viel handlicher als Betrag (nicht überall diffbar) und kann einfach abgeleitet werden.

$= \frac{1}{n} \langle y - \hat{y}, y - \hat{y} \rangle = \frac{1}{n} \left((y_1 - \hat{y}_1)^2 + \dots + (y_n - \hat{y}_n)^2 \right)$