Projekt 1: Signalrekonstruktion

Version: 21. November 2022

1 Anleitung

In diesem Projekt verwenden wir Methoden und Erkentnisse aus der numerischen linearen Algebra, um ein herausforderndes statistisches Problem zu lösen. Die Aufgaben weiter unten führen dabei durch die wichtigsten Arbeitsschritte.

Abgabe

- Die Abgabe erfolgt per Dateiupload auf Moodle und besteht aus
  1. schriftlicher Report als `.pdf`-Datei,
  2. `.R` oder `.Rmd` Datei, die alle Ergebnisse/Graphen reproduziert.
- Abgabetermin (vorläufig): Mittwoch 30. November, 11:59 Uhr.

Form

- Das Endresultat soll ein kohärenter, wissenschaftlicher Report über die wichtigsten Methoden, Erkentnisse und Ergebnisse sein. Nicht erwünscht ist eine einfache Punkt-für-Punkt-Antwort auf die gestellten Aufgaben.
- Teile den Report in (Unter-)Abschnitte mit informativen Überschriften. Versehe alle Abbildungen mit Bildunterschriften, die das gezeigte erklären.
- Fasse dich kurz und präzise. Als Richtlinie: Ein Student im nächsten Jahr sollte nur anhand des Reports und den Vorlesungsunterlagen verstehen können, was vor sich geht.
- Der Report kann in deutscher oder englischer Sprache verfasst sein.

Bewertung

Bewertet wird in erster Linie, wie vollständig und gut die einzelnen Aufgaben bearbeitet wurden. 10% der Note entfallen auf die äußere Form und Reproduzierbarkeit (ca. 0,4 in der Endnote).

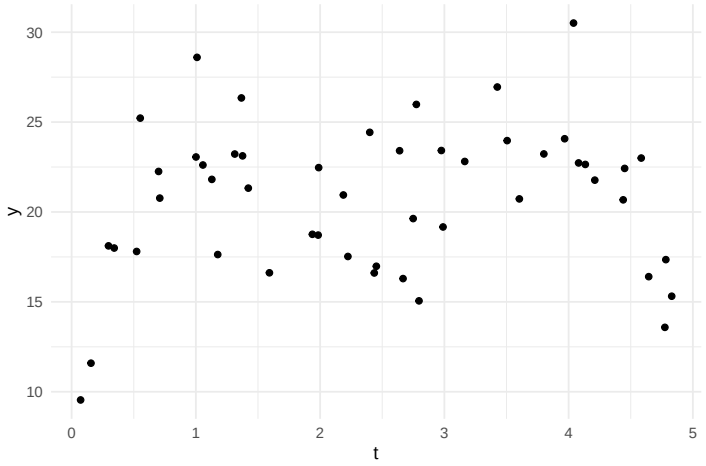

Abbildung 1: Daten aus dem physikalischen Experiment.

Gruppenarbeit

Es ist ausdrücklich erlaubt, in Gruppen an den Aufgaben zu arbeiten — mit der Einschränkung, dass jeder am Ende eine eigenstïndige Lösung schreibt und nicht Kommilitonen plagiiert. Im Zweifelsfall ist eine mündliche Nachprüfung möglich. Bitte gebt am Anfang eures Reports die Namen eurer Gruppenmitglieder an.

RMarkdown

Ein paar nützliche Optionen, falls ihr den Report mit RMarkdown erstellt:

- `fig.cap="Eine Bildunterschrift"` fügt einer Abbildung eine Unterschrift hinzu.
- `include=FALSE` versteckt Code und Output.
- `echo=FALSE` versteckt nur den Code, aber nicht den Output.

Siehe auch: https://bookdown.org/yihui/rmarkdown-cookbook.

2 Das Problem

Kontext

Ein verzweifelter Physiker braucht unsere Hilfe. Er hat ein aufwändiges Experiment durchgeführt, um die Auswirkungen eines elektromagnetischen Impulses zu erforschen. Er hatte dabei eine schöne, glatte Kurve erwartet. Zu seinem Entsetzen erweist sich das Messgerät aber als sehr ungenau und die tatsächlichen Messungen sind stark verrauscht.

Die Daten sind in Abbildung 1 zu sehen und auf moodle als `messung.csv` verfügbar. Wir wollen ihm helfen, die Messfehler zu korrigieren und die glatte Kurve zu rekonstruieren.

Mathematisches Modell

Wir stellen zunächst ein passendes Mathematisches Modell auf. Wir beschreiben die Daten $(Y_i, t_i)_{i=1}^n$ durch

$$Y_i = f(t_i) + \varepsilon_i,$$

wobei f die zu rekonstruierende, glatte Funktion und $\varepsilon_1, \dots, \varepsilon_n$ Messfehler sind. Um die Notation zu vereinfachen, schreiben wir die Gleichung oben auch als $\boldsymbol{Y} = f(\boldsymbol{t}) + \boldsymbol{\epsilon}$. Die Messzeitpunkte $\boldsymbol{t} = (t_1, \dots, t_n)$ betrachten wir als gegeben (nicht zufällig). Für unsere Rekonstruktionsmethode treffen wir die folgenden Annahmen:

- Für alle $t_1', \dots, t_m' \in [0, 5]$ gilt $f(\boldsymbol{t}') \sim \mathcal{N}(\boldsymbol{0}, K_{\boldsymbol{t}',\boldsymbol{t}'})$, wobei für alle $\boldsymbol{t} \in \mathbb{R}^n, \boldsymbol{s} \in \mathbb{R}^m$

$$K_{\boldsymbol{t},\boldsymbol{s}} := \begin{pmatrix} k(|t_1 - s_1|) & \cdots & k(|t_1 - s_m|) \\ \vdots & \ddots & \vdots \\ k(|t_n - s_1|) & \cdots & k(|t_n - s_m|), \end{pmatrix}$$

  mit *Kernfunktion* $k(s) = \exp(-s^2/\gamma)$.

- Die Fehler $\varepsilon_1, \dots, \varepsilon_n$ sind unabhängig, identisch $\mathcal{N}(0, \sigma^2)$ verteilt und unabhängig von $f(\boldsymbol{t})$.

Die Parameter γ und σ^2 können wir selbst frei wählen.

Rekonstruktionsmethode

Aus den Annahmen folgt, dass

$$\begin{pmatrix} \boldsymbol{Y} \\ f(\boldsymbol{t}') \end{pmatrix} \sim \mathcal{N}(\boldsymbol{0}, \Sigma), \qquad \Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix} = \begin{pmatrix} K_{\boldsymbol{t},\boldsymbol{t}} + \sigma^2 I_n & K_{\boldsymbol{t},\boldsymbol{t}'} \\ K_{\boldsymbol{t},\boldsymbol{t}'}^\top & K_{\boldsymbol{t}',\boldsymbol{t}'} \end{pmatrix}.$$

Die multivariate Normalverteilung hat viele praktische Eigenschaften. Für uns besonders wichtig ist, dass bedingte Verteilungen auch normal sind. Insbesondere gilt

$$f(\boldsymbol{t}') \mid \boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}\Big(\boldsymbol{\mu}_{2|1}, \Sigma_{2|1}\Big),$$

wobei

$$\boldsymbol{\mu}_{2|1} = \Sigma_{2,1}\Sigma_{1,1}^{-1}\boldsymbol{y}, \qquad \Sigma_{2|1} = \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}.$$

Sind $\boldsymbol{y}$ unsere tatsächlich beobachteten Daten, verwenden wir $\widehat{f}(\boldsymbol{t}') = \boldsymbol{\mu}_{2|1} = \mathbb{E}[f(\boldsymbol{t}') \mid \boldsymbol{Y} = \boldsymbol{y}]$ als Rekonstruktion von $f(\boldsymbol{t}')$.

3 Aufgaben

1. Beschreibe das Problem und Modell kurz in eigenen Worten. Formuliere die Berechnung von $\boldsymbol{\mu}_{2|1}$ als numerisches Problem bezüglich $\boldsymbol{y}$, $\Sigma_{2,1}$ und $\Sigma_{1,1}$.

2. Untersuche die (relative) Konditionierung der Teilprobleme für die Inputs $\boldsymbol{y}, \Sigma_{2,1}$ und $\Sigma_{1,1}$. Leite eine allgemeine Schranke für den relativen Outputfehler bezüglich aller relativen Inputfehler her. Formuliere die Hauptresultate als Theorem(e) mit formalem Beweis.

3. Untersuche den Einfluss der Hyperparameter σ und γ auf die Konditionierung des Problems. Argumentiere welche Werte den best und worst case darstellen und verifiziere die Erkentnisse numerisch. Tipp: Verwende die Funktion `svd()` zur Berechnung von Singulärwerten.

4. Schreibe einen möglichst stabilen und effizienten Algorithmus zur Berechnung von $\boldsymbol{\mu}_{2|1}$ für beliebiges $\boldsymbol{t}' \in \mathbb{R}^m$. Standardalgorithmen (z.B. für Multiplikation, Inverse und Faktorisierung von Matrizen) dürfen dabei ohne weitere Details als Subroutinen verwendet werden. Erkläre aber die Wahl der verwendeten Subroutinen.

5. Leite die Laufzeit-Komplexität des Algorithmus bezüglich m und n her und interpretiere das Resultat.

6. Schreibe einen alternativen, schlechten Algorithmus, der zwar theoretisch das richtige Resultat gibt, aber weniger stabil ist. Erkläre welchen Fehler du dabei gemacht hast.

7. Wende den guten Algorithmus jetzt auf die Daten an. Erzeuge dazu ein Gittter $0 = t'_1 < \cdots < t'_m = 5$, berechne $\boldsymbol{\mu}_{2|1}$ und stelle $\widehat{f}$ als Graphen gemeinsam mit den Originaldaten dar. Wenn alles richtig ist (und γ, σ sinnvoll gewählt), sollte jetzt eine glatte Kurve durch die Datenwolke erscheinen.

8. Untersuche den Einfluss der Parameter σ und γ auf die Rekonstruktion. Kannst du Parameter finden, die den Algorithmus numerisch versagen lassen? Wenn ja, warum? Gibt es Parameter, die nur beim schlechten Algorithmus zu Problemen führt?

9. Fasse kurz die wichtigsten Erkentnisse zusammen.