Oponentský posudek disertační práce Mgr. Ondřeje Vencálka

**Weighted data depth and depth based discrimination**

Práce je rozčleněna do tří rozsáhlých kapitol. V první kapitole je definována funkce hloubky dat (depth function) a je podán přehled takových funkcí zavedených v odborné literatuře. Jsou to zejména poloprostorová hloubka (halfspace depth), simpliciální hloubka (simplicial depth), $L_1$-hloubka ($L_1$-depth), zonoidální hloubka (zonoid depth) a další. Je poukázáno na možné aplikace hloubky dat. Vlastní výsledky doktoranda jsou pak uvedeny v dalších částech práce.

Druhá kapitola je věnována vážené hloubce dat. Jsou uvedeny příklady váhové funkce a základní vlastnosti. Pozornost je věnována otázkám konzistence (podmínkám regularity a striktní bodové konzistenci). Podobně jako v celé práci i zde je věnována velká péče výpočetním aspektům a speciálním případům. Zde se tyto speciální případy týkají symetrických a nesymetrických rozdělení v konvexních i nekonvexních případech.

Poslední kapitola je věnována problému diskriminace řešeného pomocí hloubky dat. Výsledky jsou ilustrovány na simulační studii.

Výzkum prezentovaný ve druhé kapitole byl motivován některými nedostatky metod založených na poloprostorové hloubce. Nově navrhovaná hloubková funkce dovoluje, aby centrální oblasti byly i nekonvexní.

Novým výsledkem v kapitole o diskriminaci je modifikovaný klasifikátor založený na $k$ nejbližších sousedech. Tato metoda se implementuje snáze než srovnatelné klasifikátory založené na jádrových odhadech hustoty.

Práce je velmi rozsáhlá a přitom je napsána s velkou pečlivostí. Samotných překlepů jsem nalezl jen velmi málo. Např. na str. $7^{11}$ má být $\boldsymbol{\theta}$ místo $\theta$, na str. $28_{17}$ by mělo být $HD(\boldsymbol{x})$, na str. $41^{11}$ by mělo být therefore. Z ostatních připomínek uvádím alespoň tyto.

Na řadě míst autor uvádí velikost úhlu dvou vektorů, ale nepíše, v čem tuto velikost měří. Viz str. 31, 35 (může být úhel i záporný?), 38, 40 (podle této definice by se zaváděl i úhel mezi dvěma nulovými vektory).

Na str. $32_8$ by se asi mělo předpokládat, že matice $\boldsymbol{\Sigma}$ je pozitivně definitní.

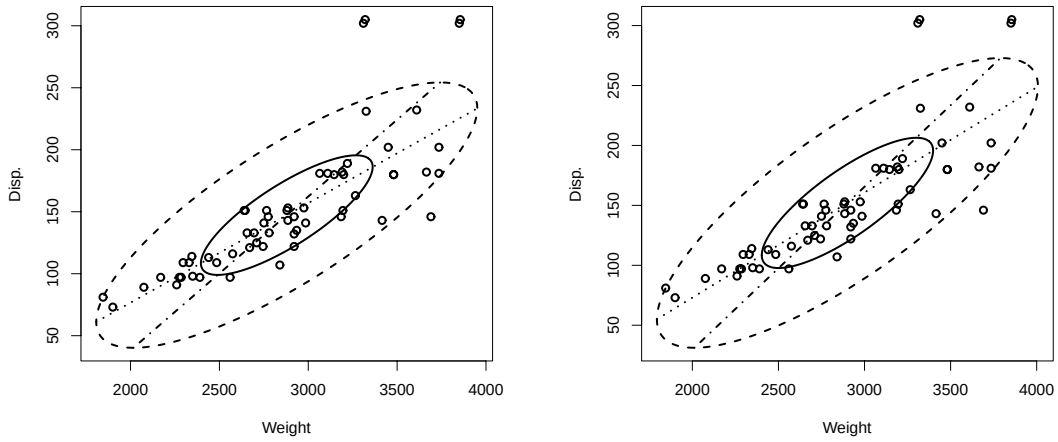
Exponenciální rozdělení lze z jednorozměrné verze do více dimenzí zobecnit více způsoby. Kterou verzi měl autor na mysli na str. 43?

Na přiloženém CD je popsán příkaz `set.seed(23041983)`. Použití příkazu `set.seed` je třeba pochválit, rádo se na to zapomíná. Dotaz je, zda argument 23041983 má nějaké zvlášť dobré vlastnosti, nebo proč byl takto zvolen.

Boxplot lze z jedné dimenze do vyšších dimenzí zobecnit více způsoby. Podle mého názoru statistici spíše než bagplot uvedený na str. 18 používají dvojrozměrný boxplot zavedený v článku Goldberg, Iglewicz (1992). Byl převzat do knihy Everitt (2005), která je poměrně populární. Je mi líto, že o něm v práci není ani zmínka. Proto bych ho stručně připomněl a tímto způsobem zpracoval data znázorněná autorem na obr. 1.2 na str. 19.

Na obr. 1 jsou znázorněny dvojrozměrné boxploty proměnných `Weight` a `Disp.` ze souboru `car.test.frame` (Automobile Data from 'Consumer Reports' 1990), který je k disposici v knihovně `rpart` a je použit k ilustraci na obr. 1.2 na str. 19 v předkládané disertaci. Dvojrozměrný boxplot založený na robustních odhadech polohy, měřítka a korelace je uveden v levé části obrázku 1, boxplot založený na nerobustních odhadech je uveden vpravo. Boxplot se skládá v podstatě ze dvou koncentrických elips. Vniřní elipsa

se nazývá „hinge“ a zahrnuje 50 % dat. Vnější elipsa se nazývá „fence“ a vyčleňuje potenciální problematická odlehlá pozorování. Dále jsou znázorněny regresní přímky $y$ na $x$ a $x$ na $y$. Jejich průsečík znázorňuje dvojrozměrný odhad polohy. Oba obrázky indikují, že 5 dat může být považováno za odlehlá pozorování. To je ve shodě s obr. 1.2 v disertaci.




Obrázek 1: Dvojrozměrný boxplot založený na robustních odhadech (vlevo) a na nerobustních odhadech (vpravo)

Jak jsem již uvedl, práce je napsána velmi pečlivě. Přináší nové výsledky, které jsou důležité pro rozvoj tohoto vědního oboru i pro aplikace. Autor navazuje na nedávno publikované výsledky, což svědčí o aktuálnosti tématiky. Disertační práce jednoznačně prokazuje předpoklady autora k samostatné tvořivé práci. Proto doporučuji, aby na základě této práce byl autorovi udělen titul PhD.

# Reference

[1] Everitt B. S. (2005): An R and S-PLUS Companion to Multivariate Analysis. Springer, London. (pp. 26 - 29)

[2] Goldberg K. M., Iglewicz B. (1992): Bivariate extensions of the boxplot. Technometrics **34**, 307-320.

V Praze dne 12. září 2011

Prof. RNDr. Jiří Anděl, DrSc.