1. Metode skupljanja (Shrinkage Methods)

Metode izbora podskupa koje smo opisali na prošlom predavanju uključuju upotrebu metode najmanjih kvadrata kako bi se kreirao linearni model koji sadrži podskup prediktora. Kao alternativu, možemo kreirati model koji sadrži svih p prediktora koristeći tehniku koja ograničava ili reguliše procene koeficijenata, ili ekvivalentno tome, koja smanjuje (eng. *shrink*) procene koeficijenta na nulu.

U ovom trenutku deluje nejasno, jer nije očigledno, zašto bi takvo ograničenje trebalo da poboljša prilagođavanje modela podacima, *ali se ispostavilo da skupljanje procena koeficijenta može da značajno smanji varijansu modela*. Dve najpoznatije tehnike za skupljanje koeficijenata regresije ka nuli su grebena regresija (eng. *ridge regression*) i laso (eng. *lasso*).

1.1 Grebena regresija

Za klasični linearnu regresiju sa kojom smo se upoznali ranije važi:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

Grebena regresija je vrlo slična metodi najmanjih kvadrata, osim što se koeficijenti procenjuju minimiziranjem malo drugačije veličine.

Procene koeficijenta grebene regresije $\hat{\beta}^R$ su vrednosti koje minimiziraju sledeći izraz:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

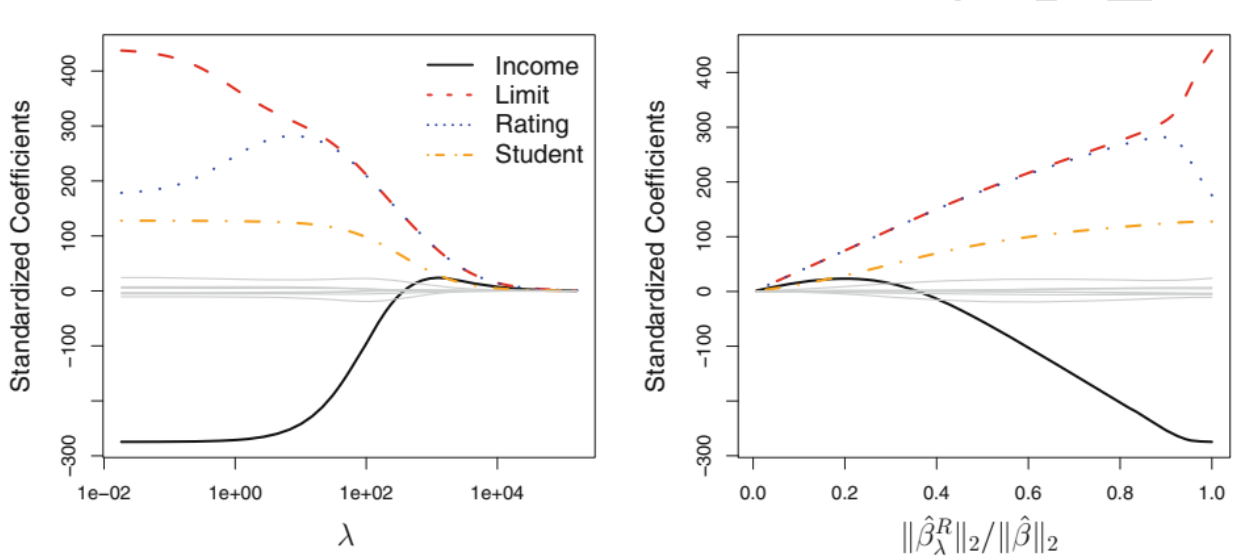
gde je λ dodatni parameter koji treba odrediti posebno. Kao i kod metode najmanjih kvadrata, grebena regresija traži procene koeficijenata koje dobro odgovaraju podacima, čineći *RSS* malim. Međutim, drugi izraz, $\lambda\sum_{j=1}^{p}\beta_j^2$, koji se naziva kazna skupljanja (eng. *shrinkage penalty*), ima malu vrednost kada su $\beta_1, \beta_2, \ldots, \beta_p$ blizu nule, pa to izaziva efekat skupljanja β_j prema nuli.

Parametar podešavanja (eng. *tuning parameter*) λ služi za kontrolu relativnog uticaja ova dva člana na procenu koeficijenata regresije. Šta važi kada je njegova vrednost 0, a šta kada je vrednost velika?

Za razliku od metode najmanjih kvadrata, koji generišu samo jedan skup procena koeficijenata, grebena regresija će proizvesti drugačiji skup koeficijenata procene, $\hat{\beta}_\lambda^R$, za svaku vrednost λ.

BITNO: β_0 nije deo drugog sabirka, jer ne želimo da skupimo *intercept* vrednost.

Primer: *Credit dataset*



Slika 1. Prikazani su standardizovani koeficijenti grebene regresije za skup podataka o kreditima, u funkciji od λ i $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$ (**shrinkage factor**).

scale *equivariant*

$\|\beta\|_2$ je l_2 norma koja se računa kao $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$. l_2 norma meri udaljenost β od nule.

Standardne procene koeficijenta metode najmanjih kvadrata koje smo razmatrali do sada su ekvivarijantne na skali (eng. *scale equivariant*): množenje X_j sa konstantom c dovodi do skaliranja procena koeficijenta najmanjih kvadrata sa faktorom $1/c$. Drugim rečima, bez obzira na to kako je j-ti prediktor skaliran, $X_j\hat{\beta}_j$ će ostati isti. Suprotno tome, procene koeficijenta grebene regresije mogu se bitno promeniti množenjem datog prediktora konstantom. Zbog toga je najbolje primeniti regresiju grebena nakon standardizacija prediktora, koristeći formulu:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}},$$

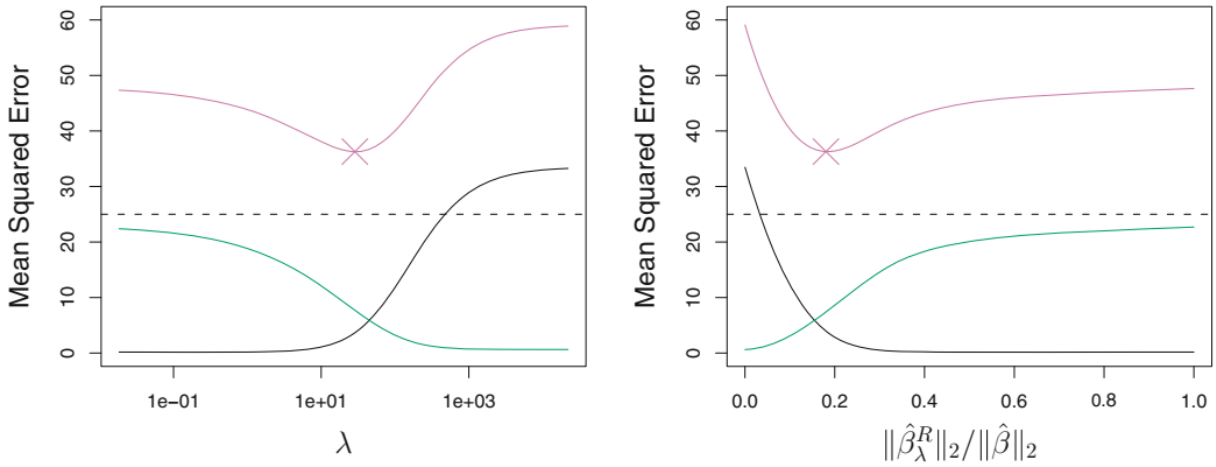
tako da su prediktori na istoj skali. Na slici 1, osa i prikazuje standardizovane procene koeficijenta grebene regresije - to jest, procene koeficijenata koje su rezultat izvršavanja grebene regresije pomoću standardizovanih prediktora

Zašto je ovaj pristup bolji od metode najmanjih kvadrata?

Prednost grebenske regresije u odnosu na metodu najmanjih kvadrata zasnovana je na kompromisu pristrasnosti i varijanse (eng. *bias-variance trade-off*).

Neka su simulirani podaci za koje je p = 45 i n = 50.

Kako se λ povećava, fleksibilnost grebenske regresije opada, što dovodi do smanjenja varijanse, ali i povećane pristrasnosti.



Slika 2. Kvadrirani bias (crna boja), varijansa (zelena boja), i test MSE (ljubičasta) za predikciju grebene regresije na simuliranim podacima, kao funkcija od λ. Isprekidana horizontalna linija predstavlja najmanju moguću MSE.

Podsetnik:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + Var(e),$$

Grebenska regresija takođe ima značajne računske prednosti u odnosu na najbolji odabir podskupa, koji zahteva pretragu kroz 2^p modela. Čak i za umerene vrednosti p, takva pretraga može biti računski neizvodljiva. Suprotno tome, za bilo koju fiksnu vrednost λ, grebena regresija fituje samo jedan model.

1.2 Lasso

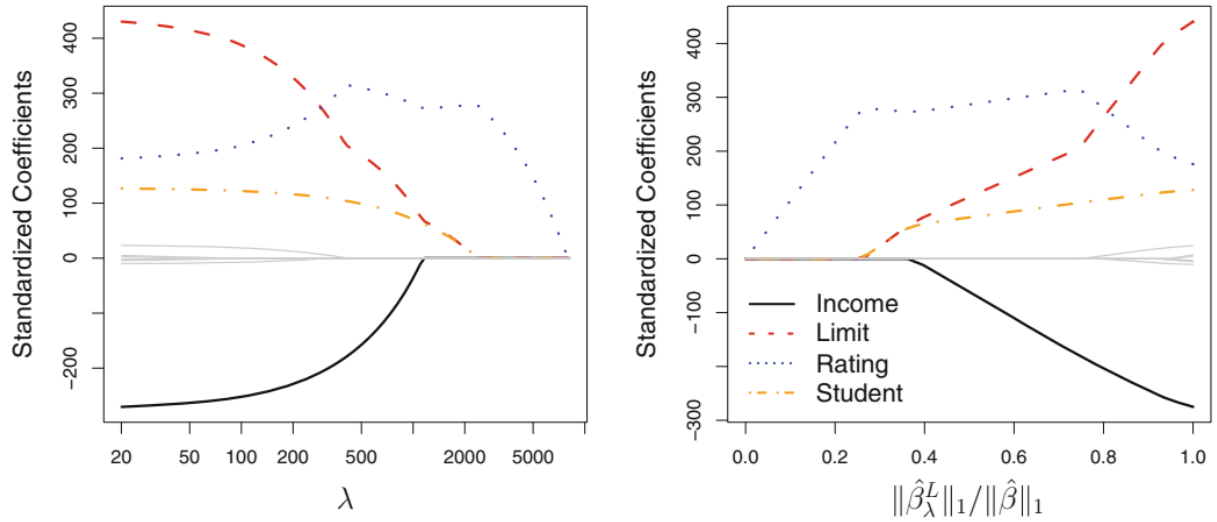
Grebena regresija ima jedan očigledan nedostatak. Za razliku od pristupa izbora najboljeg podskupa, izbor sa korakom unapred i korakom unazad, koji će generalno odabrati modele koji uključuju samo podskup promenljivih, grebena regresija će u konačni model uključiti svih p prediktora.

Laso je relativno nedavna alternativa grebenoj regresiji koja prevazilazi ovaj nedostatak.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

Statističkim jezikom, laso koristi kaznu l_1 umesto kazne l_2. Norma l_1 vektora koeficijenta β data je sa $\|\beta\|_1 = \sum|\beta_j|$.

Kao i kod grebene regresije, i laso smanjuje procene koeficijenta ka nuli. Međutim, u slučaju lasa, kazna l_1 ima za posledicu „prisiljavanja“ nekih procena koeficijenata da budu tačno jednake nuli kada je parametar podešavanja λ dovoljno velik. Slično kao i izbor najboljeg podskupa, i laso vrši izbor prediktora, pa je samim tim i interpretacija modela lakša.



Slika 3. Standardizovani laso koeficijenti na skupu podataka Credit.

1.3 Poređenje dva pristupa

Jasno je da laso ima veliku prednost nad grebenom regresijom, jer proizvodi jednostavnije i razumljivije modele koji uključuju samo podskup prediktora. Međutim, koja metoda dovodi do bolje tačnosti predviđanja? Ni grebenska regresija ni laso neće univerzalno dominirati jedno nad drugim.

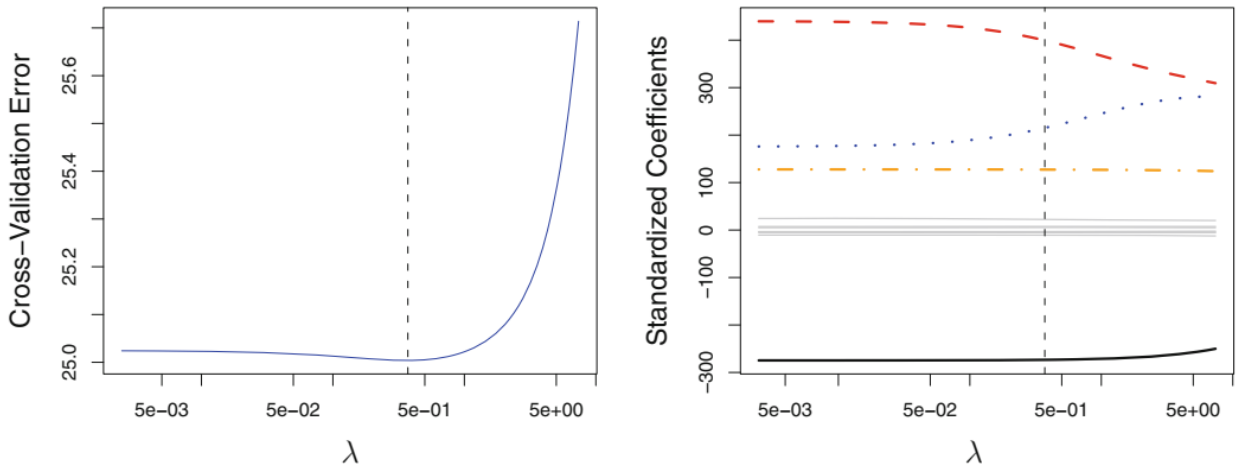
Generalno, moglo bi se očekivati da će laso imati bolji učinak u okruženju u kojem relativno mali broj prediktora ima značajne koeficijente, a preostali prediktori imaju koeficijente koji su vrlo mali ili jednaki nuli. Grebena regresija će bolje raditi kada je output funkcija mnogih prediktora, svi sa koeficijentima približno jednake veličine.

Međutim, broj prediktora koji je povezan sa odgovorom nikada nije poznat *a priori* za stvarne skupove podataka.

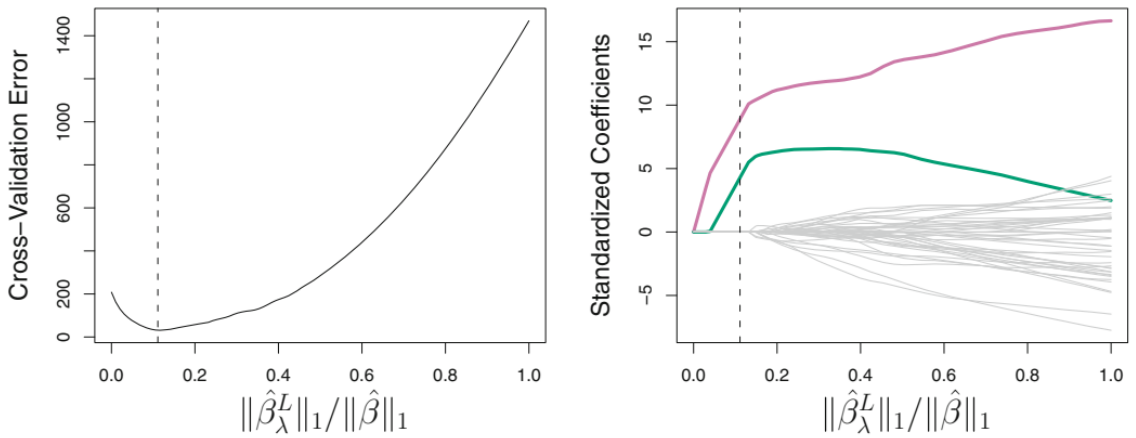
Kao i kod grebene regresije, kada procene metode najmanjih kvadrata imaju prekomerno veliku varijansu, laso rešenje može da dovede do smanjenja varijanse na račun malog povećanja bias-a, i shodno tome može da generiše tačnija predviđanja. Za razliku od grebene regresije, laso vrši selekciju prediktora i otuda rezultira modelima koji su lakši za tumačenje.

1.4 Izbor parametra podešavanja

Unakrsna validacija pruža jednostavan način za rešavanje ovog problema. Biramo mrežu vrednosti λ i izračunavamo grešku unakrsne provere za svaku vrednost λ. Zatim biramo vrednost parametra podešavanja za koju je greška unakrsne provere najmanja.



Slika 4. Levo: Greške unakrsne validacije koje su rezultat primene *grebena regresije* na kreditni skup podataka sa različitom vrednošću λ. **Desno:** Procena koeficijenta u funkciji od λ. Vertikalna isprekidana linije označava vrednost λ izabranog unakrsnom validacijom.



Slika 5. Levo: 10-fold unakrsna validacija. **Desno:** Odgovarajući laso koeficijenti.

2. Smanjenje dimenzija (Dimensionality reduction)

Metode o kojima smo do sada diskutovali su kontrolisale varijansu na dva različita načina, bilo korišćenjem podskupova izvornih promenljivih (feature selection), bilo smanjivanjem njihovih koeficijenata prema nuli (shrinkage methods).

Sada istražujemo klasu pristupa koji transformišu prediktore, a zatim fituju model najmanjih kvadrata koristeći tako transformisane promenljive. Ove tehnike se nazvaju metodama smanjenja dimenzija.

Neka $Z_1, Z_2, \dots, Z_M$ predstavljaju $M < p$ linearne kombinacije naših originalnih p prediktora.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j,$$

za neke konstante $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}$, gde je $m = 1,2, \dots, M$. Sada linearna regresija ima sledeći oblik:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \dots, n$$

Ako se konstante $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}$ biraju na odgovarajući način takvi pristupi smanjenja dimenzija često mogu nadmašiti regresiju zasnovanu na metodi najmanjih kvadrata.

Pojam smanjenje dimenzija dolazi iz činjenice da ovaj pristup smanjuje problem procene $p + 1$ koeficijenta $\beta_0, \beta_1, \dots \beta_p$ do jednostavnijeg problema procene $M + 1$ koeficijenta $\theta_0, \theta_1, \dots \theta_M$, gde je $M < p$.

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

gde je

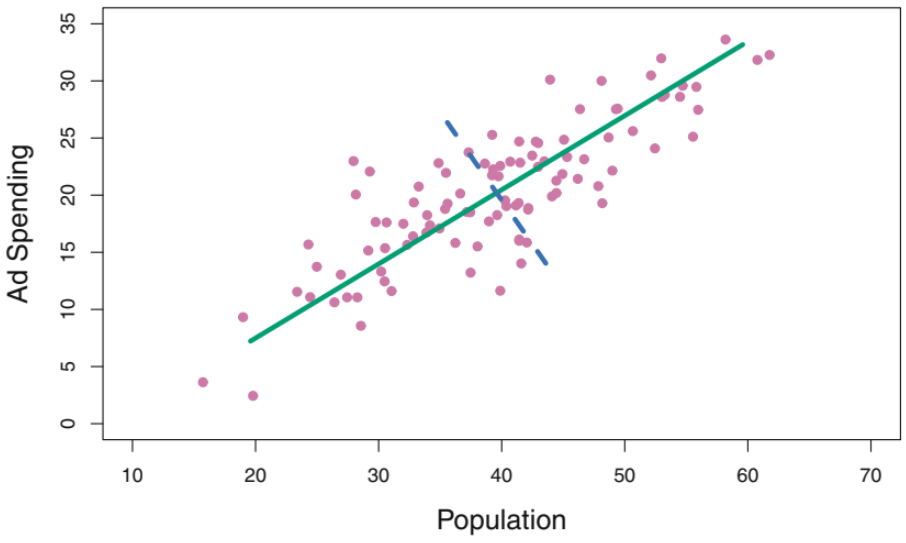
$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Smanjenje dimenzija služi za ograničavanje procenjenih β_j koeficijenata, jer sada oni moraju da imaju navedeni oblik. Ovo ograničenje u obliku koeficijenata ima potencijal da pristrasno (*bias*) proceni koeficijente. Međutim, u situacijama kada je p veliko u odnosu na n, odabirom vrednosti $M \ll p$ može se značajno smanjiti varijansa dobijenih koeficijenata.

2.1 Analiza glavnih komponenti (Principal Components Analysis)

Analiza glavnih komponenata (PCA) je popularan pristup za dobijanje manje-dimenzionalnog skupa prediktora iz velikog skupa promenljivih.

Prvi glavni pravac komponenti podataka je onaj u kojem se observacije najviše razlikuju. Linija koja je najbliža podacima!



Slika 6. Veličina populacije (*pop*) i potrošnja oglašavanja (ad) za 100 različitih gradovi su prikazani kao ljubičaste tačke. Zelena puna linija označava prvu glavnu komponentu, a plava isprekidana linija drugu glavnu komponentu.

Prva glavna komponenta prikazana je grafički na slici 6, ali kako se ona može matematički prikazati? Prva glavna komponenta je data formulom:

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad}).$$

Ideja je da se od svake moguće linearne kombinacije *pop* i *ad* takvih da je $\phi_{11}^2 + \phi_{21}^2 = 1$, dobije konkretna linearna kombinacija koja daje najveću varijansu.

$$\mathrm{Var}(\phi_{11} \times (\mathrm{pop} - \overline{\mathrm{pop}}) + \phi_{21} \times (\mathrm{ad} - \overline{\mathrm{ad}}))$$

Neophodno je razmotriti samo linearne kombinacije oblika $\phi_{11}^2 + \phi_{21}^2 = 1$, jer bismo u suprotnom mogli da povećavamo ϕ_{11} i ϕ_{21} proizvoljno i tako bismo mogli da „naduvamo“ varijansu.

Do sada smo se koncentrisali na prvu glavnu komponentu. Generalno, može se konstruisati do p različitih glavnih komponenti. Druga glavna komponenta Z_2 je linearna kombinacija promenljivih koja nije u korelaciji sa Z_1 i ima najveću varijansu koja podleže ovom ograničenju. Pokazuje se da je nulti uslov korelacije Z_1 sa Z_2 ekvivalentan uslovu da pravac mora biti ortogonalan na prvi smer glavne komponente.

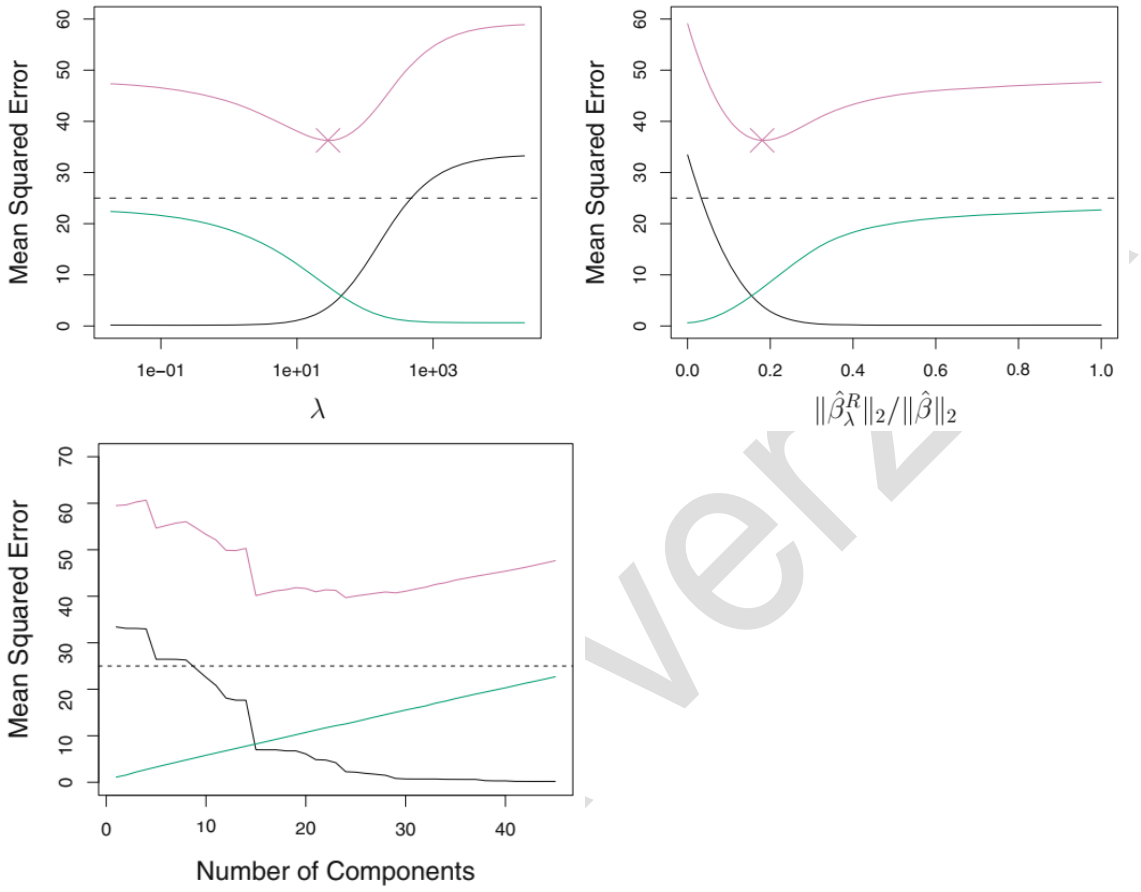
Ako bismo imali druge prediktore, kao što su starost stanovništva, nivo dohotka, obrazovanje itd., onda bi se mogle konstruisati dodatne komponente. Oni bi sukcesivno uvećavali varijansu, prema ograničenju da budu nekorelisani sa prethodnim komponentama.

2.2 Principal Component Regression (PCR)

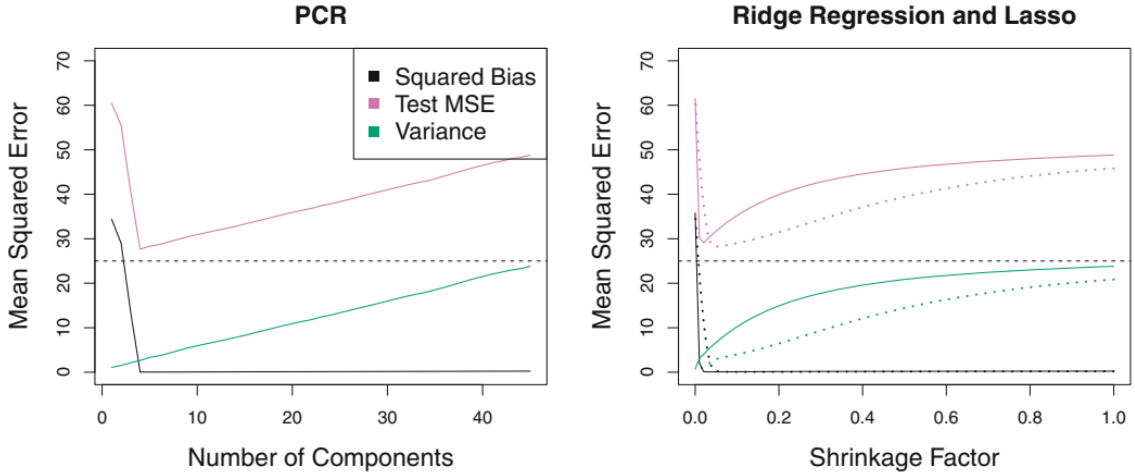
Ključna ideja je da je često mali broj glavnih komponenti dovoljan da objasni većinu varijabilnosti u podacima, kao i odnos sa outputom. Drugim rečima, pretpostavljamo da pravci u kojima $X_1, X_2, \dots, X_p$ pokazuju najviše varijacija su pravci koji su povezani sa Y (*).

Ako važi pretpostavka (*) u osnovi PCR-a, prilagođavanje modela na $Z_1, Z_2, \dots, Z_M$ će dovesti do boljih rezultata od prilagođavanja modela na $X_1, X_2, \dots, X_p$, jer je većina ili sve informacije u podacima koje se odnose na output već sadržane u $Z_1, Z_2, \dots, Z_M$, a procenom samo $M \ll p$ koeficijenata možemo ublažiti *overfitting*.

I PRIMER: grebena regresija i PCR



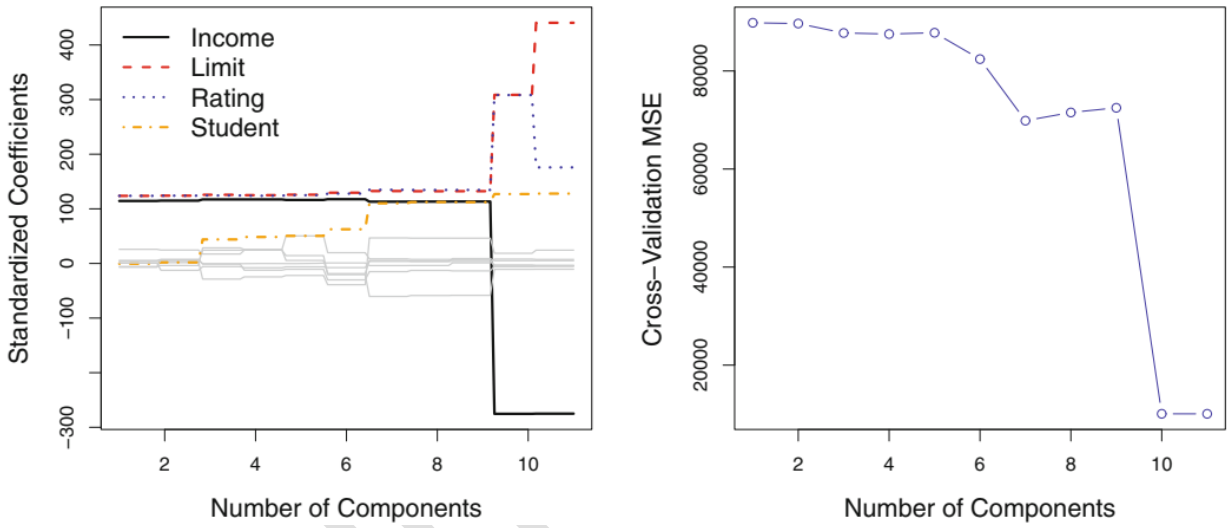
II PRIMER



Slika 7. PCR, grebena regresija i laso primenjeni su na simulirani skup podataka u kojem prvih pet glavnih komponenti X sadrži sve informacije o odgovoru Y. Na svakom panelu, ireducibilna greška je $Var(e)$ prikazana kao vodoravna isprekidana linija. **Levo**: Rezultati za PCR. **Desno**: Rezultati za laso (solid) i grebena regresiju (tačkasto). X-osa prikazuje faktor skupljanja koeficijenata.

PCR nije feature selection metod! To je zato što je svaka od M glavnih komponenti korišćenih u regresiji linearna kombinacija svih p originalnih karakteristika.

U PCR-u, broj glavnih komponenata, M, obično se bira unakrsnom validacijom.



Slika 8. Levo: PCR standardizovane procene koeficijenta na kreditnim podacima u zavisnosti od vrednosti M. **Desno**: MSE vrednost za 10-fold unakrsna validaciju dobijenu primenom PCR-a, u funkciji od M.

Prilikom izvođenja PCR-a, generalno se preporučuje standardizacija svih prediktora pre generisanja glavnih komponenti. Ova standardizacija osigurava da su sve promenljive na istoj skali.

Partial Least Square (PLS) – DOMAĆI!