Regularizace a robustnost při klasifikační analýze genetických dat

Jan Kalina

Odd. medicínské informatiky a biostatistiky
Ústav informatiky AV ČR, v.v.i.

Regularizace a robustnost při klasifikační analýze genetických dat

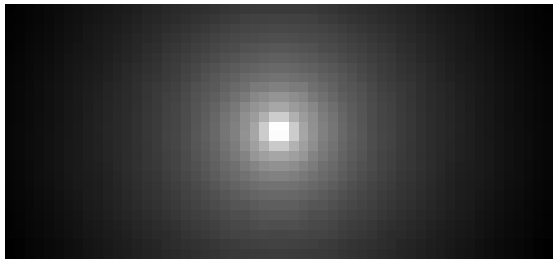
- **Optimální šablony**
- Klasifikační analýza
- Regrese
- Selekce proměnných

Vážený korelační koeficient

$r_W(\mathbf{x}, \mathbf{y}; \mathbf{w})$ = vážený korelační koeficient mezi $\mathbf{x}, \mathbf{y}$ s vahami $\mathbf{w}$

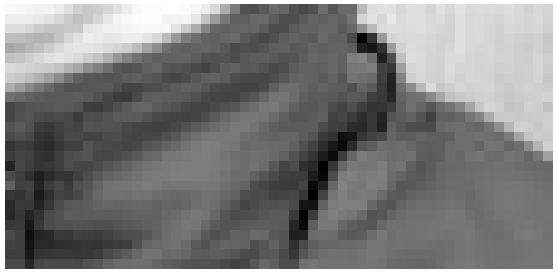
Šablona:

Radiální váhy:



Ústa:

Neústa:



	Váhy	
	Shodné	Radiální
r_W(šablona, ústa; $\mathbf{w}$)	0,48	0,66
r_W(šablona, neústa; $\mathbf{w}$)	0,52	0,38

Motivace

Šablona jako průměr?

Cíl: Modifikovat váhy, zachovat šablonu.

Separace mezi konkrétními ústy a konkrétními neústy:

$$\frac{r_W^F(\text{šablona, ústa; } \mathbf{w})}{r_W^F(\text{šablona, neústa; } \mathbf{w})}$$

Fisherova transformace $r_W(\mathbf{x}, \mathbf{y}; \mathbf{w})$ zdůrazní extrémní hodnoty:

$$r_W^F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \frac{1}{2} \log \frac{1 + r_W(\mathbf{x}, \mathbf{y}; \mathbf{w})}{1 - r_W(\mathbf{x}, \mathbf{y}; \mathbf{w})}$$

Optimalizační kritérium

Přístup minimaxu (uvažuje se separace mezi každými ústy a neústy, v každém obraze).

Maximum přes váhy
Minimimum přes obrazy
Minimum přes neústa
Maximum přes pozice úst (drobné posunutí)

$$\frac{r_W^F(\text{šablona, ústa}; \mathbf{w})}{r_W^F(\text{šablona, neústa}; \mathbf{w})}$$

Optimalizační algoritmy (mohou však skončit jen v lokálním extrému):

- Lineární aproximace
- Genetická optimalizace (“hrubý aproximativní algoritmus”)

Lineární aproximace

Pro konkrétní ústa, neústa a šablonu označme pomocí $f(w_1, \ldots, w_n)$ separaci mezi ústy a neústy. Zde $n = 26 \times 56 = 1456$. Taylorův rozvoj 1. řádu:

$$f(w_1 + \delta_1, \ldots, w_n + \delta_n) \approx f(w_1, \ldots, w_n) + \sum_{i=1}^{n} \delta_i \frac{\partial f(w_1, \ldots, w_n)}{\partial w_i}$$

pro malé hodnoty $\delta_1, \ldots, \delta_n$.

Optimalizační úloha: lineární problém

$$\max_{\delta_1, \ldots, \delta_n \in \mathbb{R}} \sum_{i=1}^{n} \delta_i \frac{\partial f(w_1, \ldots, w_n)}{\partial w_i}$$

za podmínek

- $0 \leq w_i + \delta_i \leq c$ (pro určité c), $\quad i = 1, \ldots, n$
- $\sum_{i=1}^{n} \delta_i = 0$
- podmínka na symetrii vah
- event. $\sum_{i=1}^{n} \delta_i \frac{\partial f(w_1, \ldots, w_n)}{\partial w_i} = \sum_{i=1}^{n} \delta_i \frac{\partial f^*(w_1, \ldots, w_n)}{\partial w_i}$
  pro jiný nejhorší případ se separací $f^*(w_1, \ldots, w_n)$

Optimalizace šablony i vah

Počáteční šablona:

Počáteční váhy:

0,78

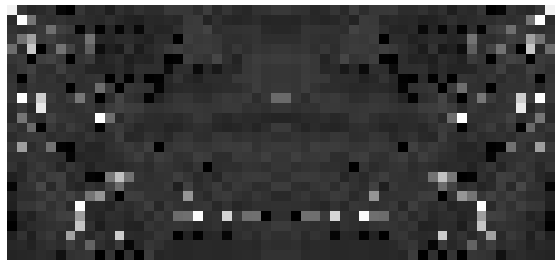
Optimální šablona:

Počáteční váhy:

Nejhorší separace:
2,12

Optimální šablona:

Optimální váhy:

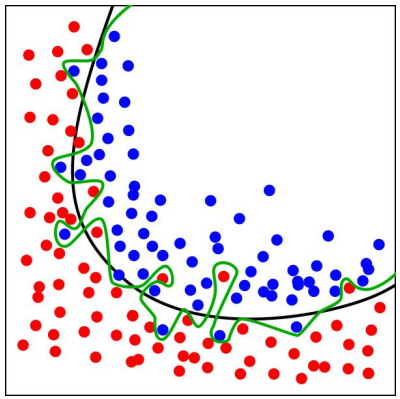


2,29

Optimalizace vyžaduje symetrii a regularizační podmínky.

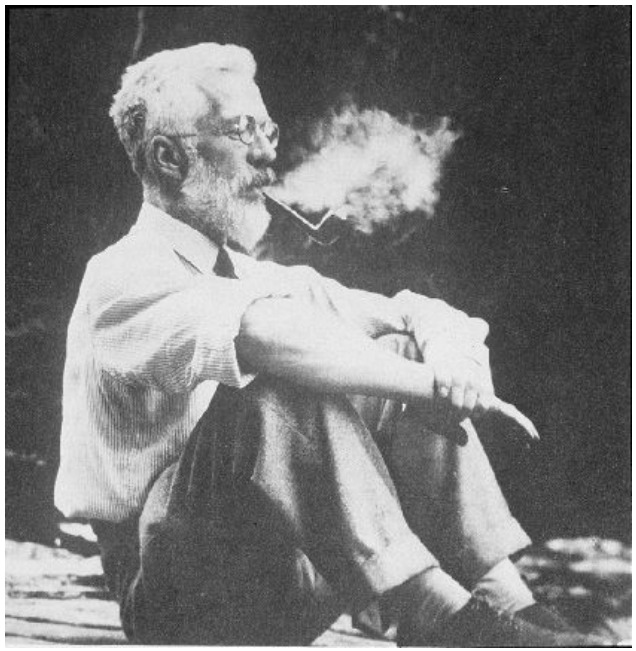
Regularizace

- Numerická lineární algebra: $\mathbf{Ax} = \mathbf{b}$
  - Modifikace úlohy vedoucí k potlačení vlivu šumu ve vektoru $\mathbf{b}$ na spočítané řešení a k ošetření špatné podmíněnosti $\mathbf{A}$ (Duintjer Tebbens a kol., 2012)
- Analýza obrazu
  - Odstranění (vyhlazení) šumu z obrazu pomocí *wavelet shrinkage* (Donoho & Johnstone, 1994)
- Teorie aproximací
  - Apriorní představa o hladkém chování aproximované funkce, kontrola její komplexnosti (Hastie *et al.*, 2009)
- Strojové učení
  - Dodatečná informace pro vyřešení špatně podmíněných problémů nebo prevence přeučení



Regularizace a robustnost při klasifikační analýze genetických dat

- Optimální šablony
- **Klasifikační analýza**
- Regrese
- Selekce proměnných



Ronald A. Fisher (1890–1962)

Genetická studie v Centru biomedicínské informatiky

Cíl studie: Které geny vedou k závažným onemocněním (resp. jeho těžké formě)?

Data o pacientech (interní nebo ortopedické odd. Městské nemocnice Čáslav):

1. AIM = akutní infarkt myokardu ($n = 98$). Kontrola po 6 měsících.
2. CMP = cévní mozková příhoda ($n = 46$).
3. Kontroly ($n = 169$).

Párování pacientů. Kontrola starší o 0 až 5 roků. Shoda v rizkových (klinických) faktorech: pohlaví, HN, kouření. Výhody párového designu.

Měřená data:
Osobní údaje. Klinické a biochemické veličiny. **Exprese** (=aktivita) všech genů ve vzorku periferní krve.

Medicínský význam studie.
Specifika české populace - v ČR hlavní příčina úmrtnosti.

Kvadratická diskriminační analýza (QDA)

K skupin mnohorozměrných dat (navzájem nezávislé náhodné výběry).

Mnohorozměrná normalita (odlišné vektory středních hodnot, odlišné varianční matice).

Nové pozorování $\mathbf{Z}$ je klasifikováno do k-té skupiny ($k = 1, \ldots, K$), pokud k je rovno

$$\arg \max_{k=1,\ldots,K} \left[p_k (2\pi)^{-p/2} |\mathbf{S}_k|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{Z} - \bar{\mathbf{X}}_k)^T \mathbf{S}_k^{-1} (\mathbf{Z} - \bar{\mathbf{X}}_k) \right\} \right],$$

kde

- $\bar{\mathbf{X}}_k$ = průměr dat v k-té skupině ($k = 1, \ldots, K$),
- $\mathbf{S}_k$ = odhad varianční matice v k-té skupině.
- p_k jsou apriorní pravděpodobnosti toho, že pozorujeme data z k-té skupiny.

Lineární diskriminační analýza (LDA)

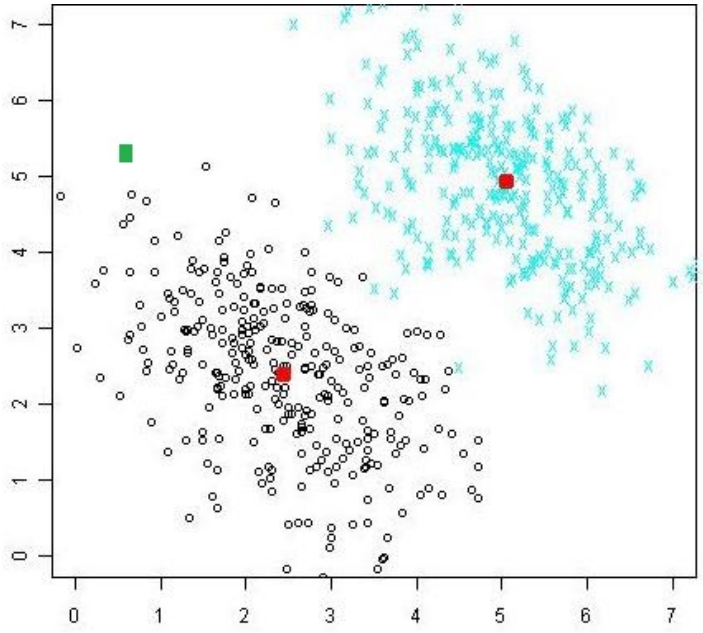
Lineární diskriminační analýza: společná varianční matice.

Nové pozorování $\mathbf{Z}$ je klasifikováno do k-té skupiny, která minimalizuje

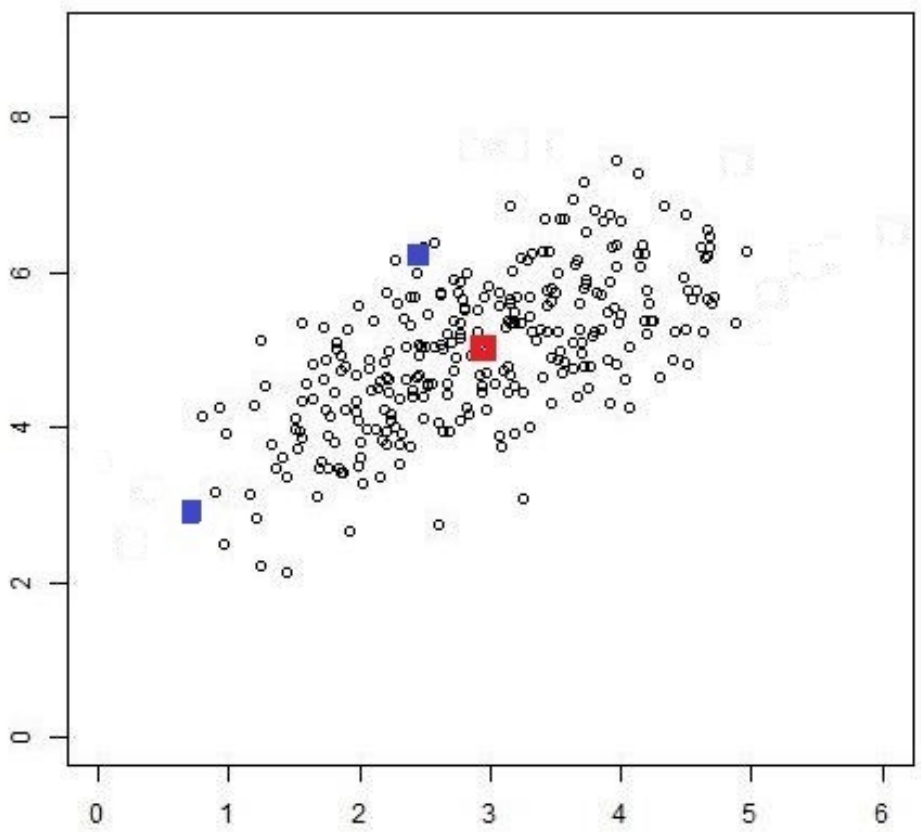
$$(\bar{\mathbf{X}}_k - \mathbf{Z})^T \mathbf{S}^{-1} (\bar{\mathbf{X}}_k - \mathbf{Z}),$$

kde

- $\bar{\mathbf{X}}_k$ = je průměr k-té skupiny,
- $\mathbf{S}$ = odhad varianční matice.



Mahalanobisova vzdálenost

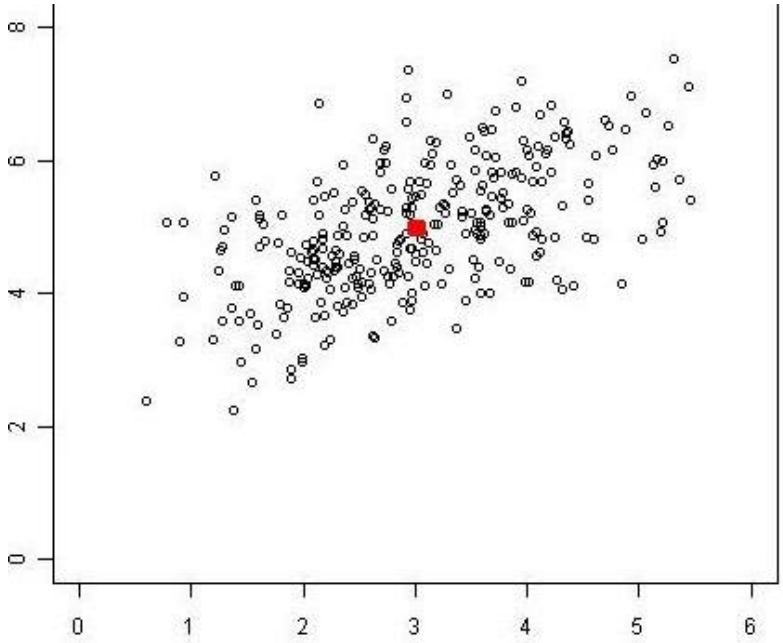


Prokletí dimenzionality: výpočet LDA nelze provést pro $n < p$!

$$d(\mathbf{Z}, \bar{\mathbf{X}}) = (\mathbf{Z} - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{Z} - \bar{\mathbf{X}})$$

Smrštěný odhad průměru

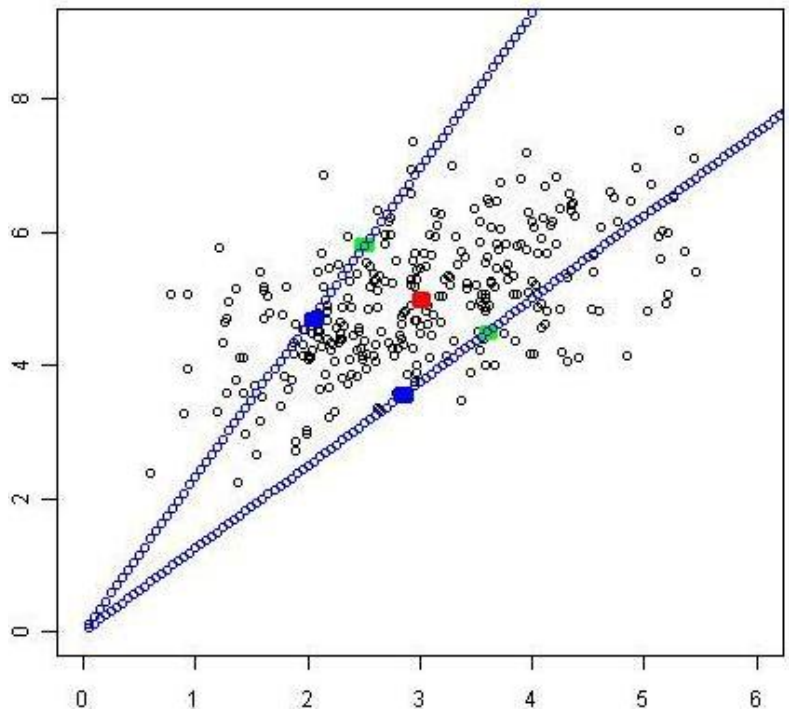
Data pocházející z p-rozměrného normálního rozdělení:



Cíl: vychýlený odhad $\mathrm{E}\mathbf{X} = \mathrm{E}\,(X_1, \ldots, X_p)^T$.

Tím se neříká, že by byl přesnější odhad $\mathrm{E}X_i$ pro pevné i.

Smrštěný odhad průměru



Smrštěný (**shrinkage**) odhad (Stein, 1956) směrem k nule nebo k libovolnému pevnému bodu:

- Nepřípustnost průměru pro mnohorozměrné normální rozdělení pro $p > 2$.
- Vychýlený odhad s menším kvadratickým rizikem než průměr.

Steinův smrštěný odhad

Smrštěné odhady (menší riziko za cenu vychýlení) (Stein, 1956; Bock, 1975):

- Za předpokladu $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\mathcal{I}}_p)$, odhad $\boldsymbol{\mu}$ ve tvaru

$$\left(1 - \frac{p-2}{||\mathbf{X}||^2}\right)\mathbf{X}$$

  dominuje výběrový průměr. Jde o odhad, který **smršťuje X** směrem k nule.

Paradoxní chování:

- Přirovnání: při hodu na terč je lepší mířit mimo střed.
- Smrštění k libovolnému bodu.
- Smrštěný odhad je výhodný i pro jediné pozorování.

Friedman (1989): Regularizovaná diskriminační analýza

Steinův paradox lze uplatnit i na odhad varianční matice, tj. pro regularizaci.

Modifikace LDA.

Dvojí regularizace varianční matice:

- Smrštění odhadu varianční matice v každé skupině $\mathbf{S}_k$ ke sdružené kovarianční matici $\mathbf{S}$, tj. smrštění QDA k LDA (parametr $\lambda \in [0, 1]$).
- Smrštění odhadu varianční matice (v každé skupině) k diagonální matici (parametr $\gamma \in [0, 1]$).

Odhad varianční matice v k-té skupině:

$$\hat{\mathbf{S}}_k(\lambda) = \lambda \mathbf{S}_k + (1 - \lambda)\mathbf{S},$$

$$\hat{\mathbf{S}}_k(\lambda, \gamma) = \gamma \hat{\mathbf{S}}_k(\lambda) + (1 - \gamma) \frac{\text{tr}\left(\hat{\mathbf{S}}_k(\lambda)\right)}{p} \mathcal{I},$$

kde

- $k = 1, ..., K$
- $\mathcal{I}$ = jednotková matice
- tr = stopa matice.

Prediction Analysis for Microarrays (PAM)

Klasická LDA pro nové pozorování $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ počítá

$$\arg \min_{k=1,\ldots,K} \left[(\mathbf{Z} - \bar{\mathbf{X}}_k)^T \mathbf{S}^{-1} (\mathbf{Z} - \bar{\mathbf{X}}_k) - 2 \log p_k \right],$$

kde p_k je apriorní pravděpodobnost, že pozorování pochází z k-té skupiny.

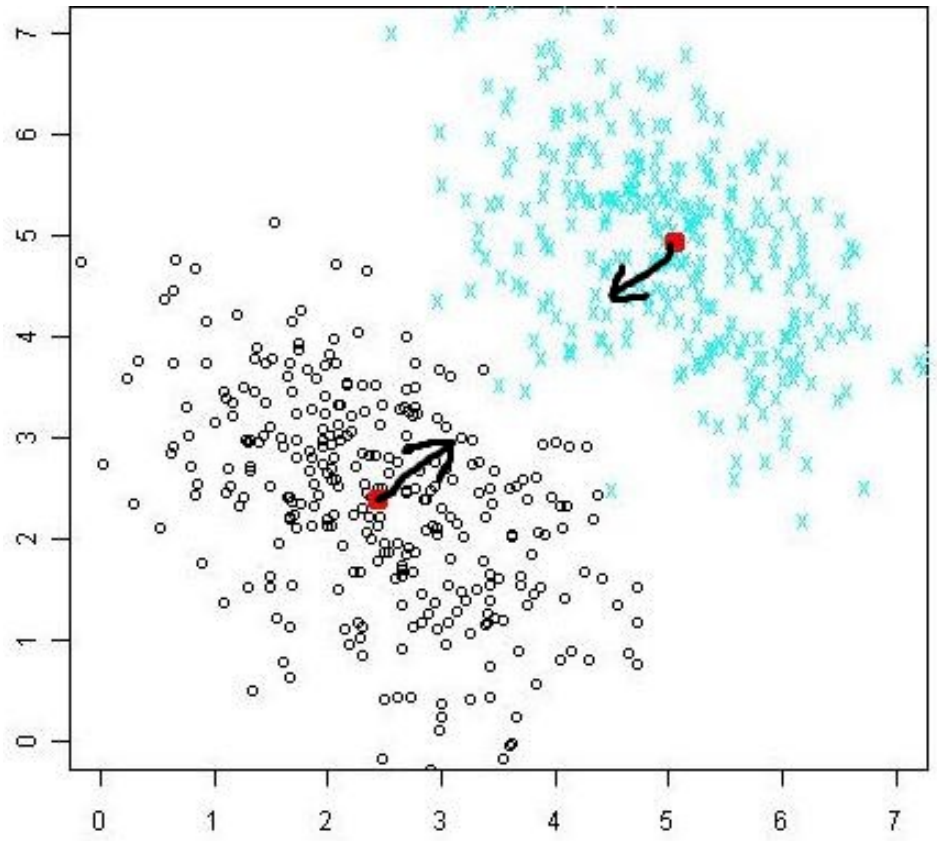
- Indexy pro geny $j = 1, \ldots, p$.
- Indexy pro pacienty $i = 1, \ldots, n$.
- Hodnoty genových expresí X_{ji}.
- Skupiny $k = 1, \ldots, K$.

Prediction Analysis for Microarrays (PAM; Tibshirani *et al.*, 2002):

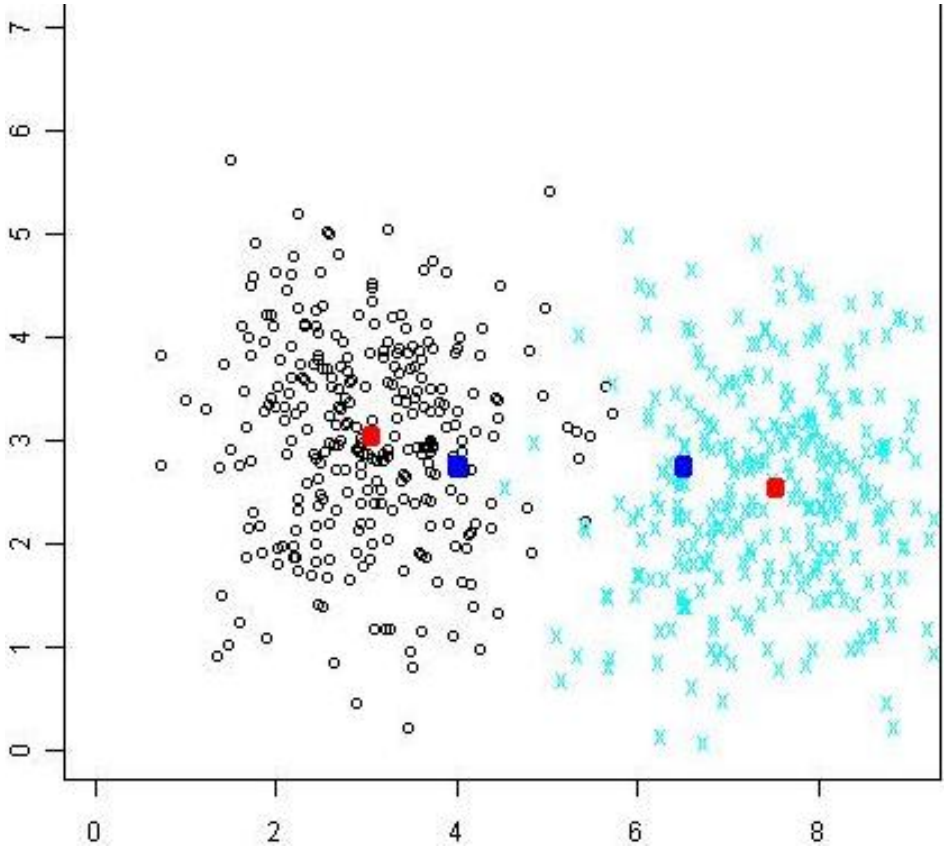
$$\arg \min_{k=1,\ldots,K} \left[\sum_{j=1}^{p} \frac{(Z_j - \bar{X}'_{jk})^2}{(s_j + s_0)^2} - 2 \log p_k \right],$$

kde s_j^2 je odhad rozptylu pro j-tý gen, s_0 je kladná konstanta, $\bar{X}'_{jk}$ je modifikovaná průměrná exprese j-tého genu ve skupině k.

PAM: Motivace pro smrštěné odhady průměrů



PAM: Motivace pro smrštěné odhady průměrů



Jde o redukci dimenze?

Prediction Analysis for Microarrays (PAM)

Smrštěný průměr (centroid) $\bar{X}'_{jk}$ exprese j-tého genu v k-té skupině:

$$\bar{X}'_{jk} = \bar{X}_j + \lambda c, \quad \text{pokud} \quad \bar{X}_{jk} - \bar{X}_j < -\lambda c,$$

$$\bar{X}'_{jk} = \bar{X}_j, \quad \text{pokud} \quad -\lambda c \leq \bar{X}_{jk} - \bar{X}_j \leq \lambda c,$$

$$\bar{X}'_{jk} = \bar{X}_j - \lambda c, \quad \text{pokud} \quad \lambda c < \bar{X}_{jk} + \bar{X}_j,$$

kde

- $\bar{X}_{jk}$ je průměrná exprese j-tého genu ve skupině k,
- $\bar{X}_j$ je průměrná exprese j-tého genu (přes skupiny),
- $$c = (s_j + s_0)\sqrt{\frac{1}{n_k} + \frac{1}{n}},$$
- s_j^2 je odhad rozptylu pro j-tý gen,
- s_0 je kladná konstanta,
- $\lambda \geq 0$ je parametr smrštění.

Výpočet: software R, knihovna pamr.

Prediction Analysis for Microarrays (PAM)

Výhody PAM:

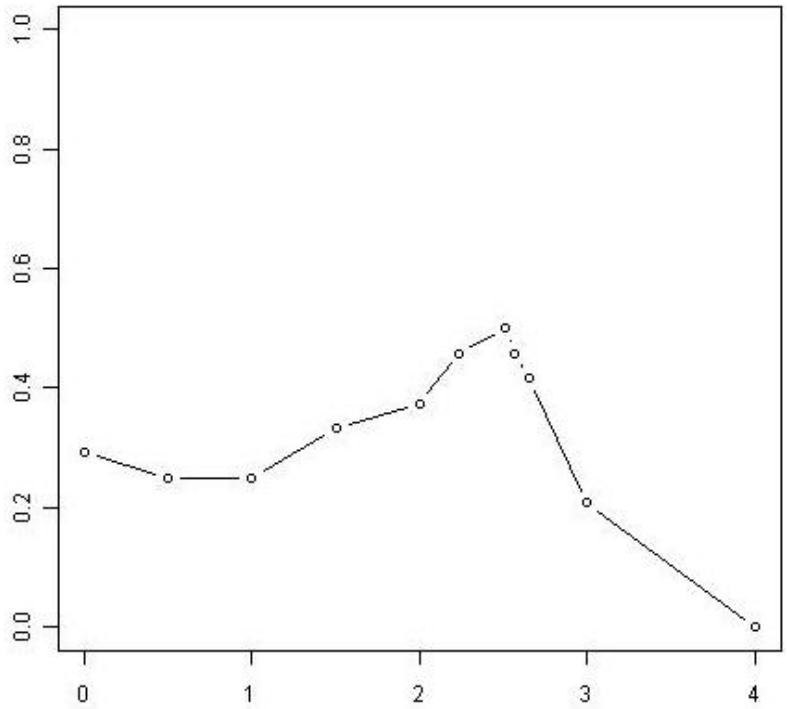
- Používá smrštěné průměry (namísto klasických).
- Prediktivita. Parametr smrštění se volí tak, aby byla minimální klasifikační chyba pro nezávislá data. (Desetinásobná křížová validace).
- Odhady průměrů lze označit za robustní.

Nevýhody PAM:

- Varianční matice se odhaduje jako diagonální.
- Rozptyly jsou odhadnuty nerobustně.

Výsledky získané metodou PAM

Hledání optimální hodnoty parametru smrštění:



Výsledky získané metodou PAM

Youdenův index v závislosti na parametru smrštění:

Parametr smrštění	Počet genů	Youdenův index
0.0	38590	0.292
0.5	14615	0.250
1.0	4100	0.250
1.5	884	0.333
2.0	140	0.375
2.233	45	0.458
2.509	20	0.500
2.567	15	0.458
2.650	10	0.417
3.0	2	0.208
4.0	0	0.000

Senzitivita = pravděpodobnost, že test bude pozitivní u nemocného pacienta.
Specificita = pravděpodobnost, že test bude negativní u zdravého pacienta.
Youdenův index = senzitivita + specificita -1 = charakteristika klasifikačního pravidla.

Výsledky získané metodou PAM

Studie akutního infarktu myokardu:

- SE = senzitivita = pravděpodobnost, že test bude pozitivní u nemocného pacienta.
- SP = Specificita = pravděpodobnost, že test bude negativní u zdravého pacienta.
- **Youdenův index** = SE + SP −1.

Porovnávané skupiny	SE	SP	**Youden**	Počet prediktivních genů (PAM)	Exprese (+/-)
AIM & kontroly	0.79	0.85	**0.64**	343	(151/192)
AIM6 & kontroly	1.00	0.87	**0.87**	45	(41/4)
AIM6 & AIM	1.00	1.00	**1.00**	17	(11/6)

Smrštěný odhad varianční matice

$\mathbf{S}^*$ invertovatelná i pro $n \ll p$.

Další možnost odhadu $\mathbf{\Sigma}$ s parametrem smrštění $\lambda \in [0, 1]$:

- $\mathbf{S}^* = \lambda \mathbf{S} + (1 - \lambda)\mathcal{I}$
- $\mathbf{S}^* = \lambda \mathbf{S} + (1 - \lambda) s \mathcal{I}, \quad s = \sum_{i=1}^{p} S_{ii}/p$
- $\mathbf{S}^* = \left(S_{ij}^*\right)_{i,j=1}^{p}, \quad S_{ij}^* = R_{ij}^* \sqrt{S_{ii} S_{jj}},$

$$R_{ij}^* = \begin{cases} 1, & \text{if } i = j \\ \lambda R_{ij} + (1 - \lambda), & \text{if } i \neq j \end{cases}$$

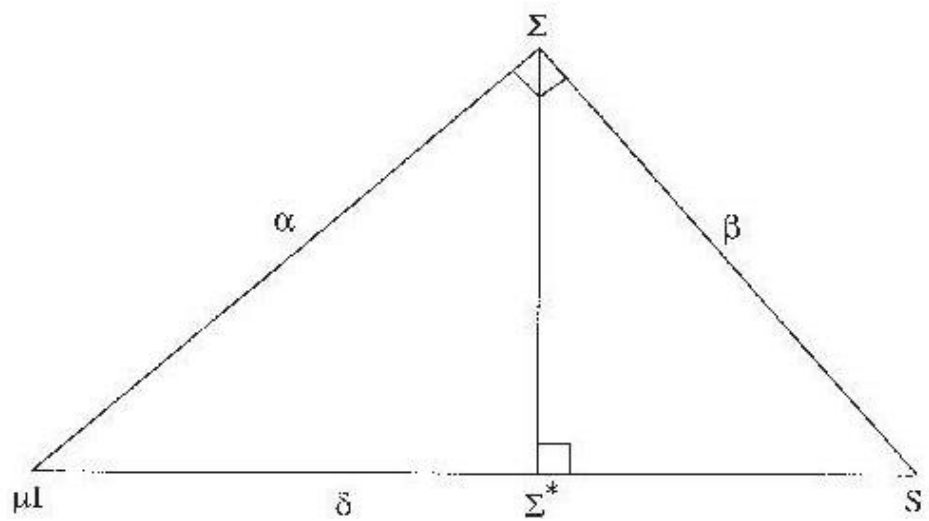
- $\mathbf{S}^* = \left(S_{ij}^*\right)_{i,j=1}^{p}, \quad S_{ij}^* = R_{ij}^* \sqrt{S_{ii} S_{jj}},$

$$R_{ij}^* = \begin{cases} 1, & \text{if } i = j \\ \lambda R_{ij} + (1 - \lambda) r, & \text{if } i \neq j \end{cases}, \quad r = \frac{1}{\frac{(p-1)p}{2}} \sum_{i=2}^{p} \sum_{j=1}^{i-1} r_{ij}$$

Smrštěný odhad varianční matice - interpretace

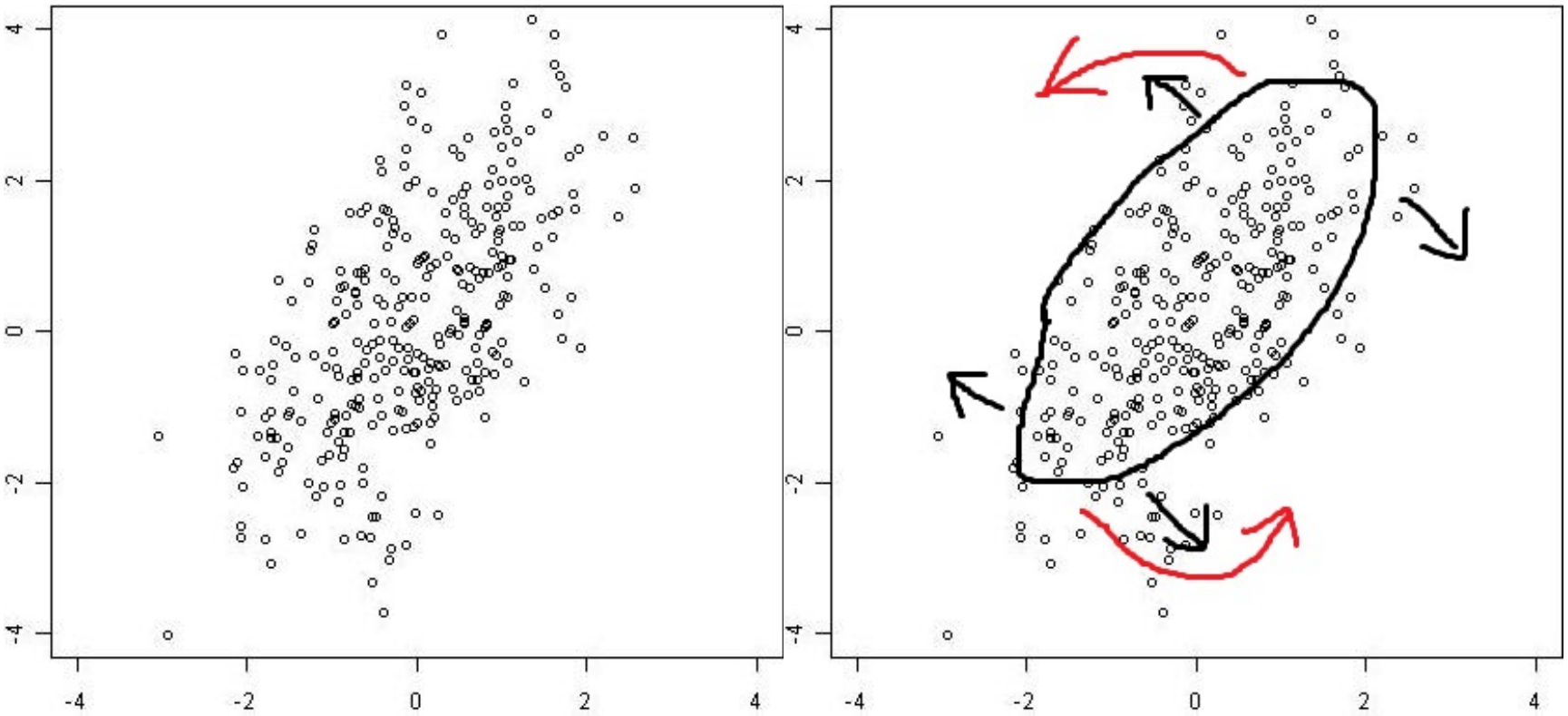
$$\mathbf{S}^* = \lambda \mathbf{S} + (1 - \lambda)\boldsymbol{\mathcal{I}}, \quad \lambda \in [0, 1]$$

- Minimalizace kvadratického rizika.
- Bayesův odhad.
- Korekce $\mathbf{S}$ na konečné n.
- Projekce v Hilbertově prostoru symetrických matic $p \times p$.
- Korekce na odhad vlastních čísel: smrštěná vlastní čísla.



Převzato: Ledoit & Wolf (2004).

Smrštěný odhad varianční matice - interpretace



Smrštěný odhad varianční matice

Parametr smrštění λ:

- Jde o parametr regularizace.
- Asymptoticky optimální hodnota (Ledoit & Wolf (2004), Schäfer & Strimmer (2005))
- V situaci

$$\mathbf{S}^* = \hat{\lambda}\mathbf{S} + (1-\hat{\lambda})\mathbf{T}, \quad T_{ij} = \begin{cases} S_{ii}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

  má optimální λ hodnotu

$$\hat{\lambda} = \frac{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}\widehat{\text{var}}(S_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{p}(S_{ij}-T_{ij})^2} = \frac{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}\widehat{\text{var}}(S_{ij})}{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}(S_{ij})^2}$$

- Pomalý výpočet.
- Využití při klasifikační analýze.

LDA*, QDA*

Smrštěná lineární diskriminační analýzu LDA*.
Pro nové pozorování $\mathbf{Z}$ uvažujeme diskriminačních skóry

$$l_k^* = \bar{\mathbf{X}}_k^T (\mathbf{S}^*)^{-1} \mathbf{Z} - \frac{1}{2} \bar{\mathbf{X}}_k^T (\mathbf{S}^*)^{-1} \bar{\mathbf{X}}_k + \log p_k, \quad k = 1, \ldots, K.$$

Nové pozorování $\mathbf{Z}$ je zařazeno do k-té skupiny, pokud $l_k^* > l_j^*$ pro všechna $j \neq k$.

Smrštěná kvadratická diskriminační analýza QDA* je založena na skórech

$$q_k^* = \bar{\mathbf{X}}_k^T (\mathbf{S}_k^*)^{-1} \bar{\mathbf{Z}} - \frac{1}{2} \bar{\mathbf{X}}_k^T (\mathbf{S}_k^*)^{-1} \bar{\mathbf{X}}_k - \frac{1}{2} \bar{\mathbf{Z}}^T (\mathbf{S}_k^*)^{-1} \bar{\mathbf{Z}} + \frac{1}{2} \log |\mathbf{S}_k^*| + \log p_k,$$

kde $\mathbf{S}_k^*$ je smrštěný odhad varianční matice k-té skupiny.
Nové pozorování $\mathbf{Z}$ je zařazeno do k-té skupiny, pokud $q_k^* > q_j^*$ pro všechna $j \neq k$.

Vše jsou speciální případy metody SCRDA (Guo, 2007).

Rychlý algoritmus #1 pro LDA*: Obecná situace

K skupin p-rozměrných dat

1. Vypočítej $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)\mathbf{T}$.
2. Spektrální rozklad
$$\mathbf{S}^* = \mathbf{Q}^*\mathbf{D}^*\mathbf{Q}^{*T}$$
3. Vypočítej
$$\mathbf{S}^{*-1} = \mathbf{Q}^*\mathbf{D}^{*-1}\mathbf{Q}^{*T}.$$
4. Klasifikuj $\mathbf{Z}$ do skupiny k, pokud
$$(\bar{\mathbf{X}}_k - \mathbf{Z})^T\mathbf{S}^{*-1}(\bar{\mathbf{X}}_k - \mathbf{Z}) = \arg\min_{j=1,\ldots,K}\left\{(\bar{\mathbf{X}}_j - \mathbf{Z})^T\mathbf{S}^{*-1}(\bar{\mathbf{X}}_j - \mathbf{Z})\right\}.$$
5. Opakuj pro různé hodnoty λ, najdi optimální klasifikační pravidlo.

Rychlý algoritmus #2 pro LDA*: Speciální cílové matice

K skupin p-rozměrných dat

1. Spektrální rozklad $\mathbf{S} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T, \quad \mathbf{D} = \mathrm{diag}\{\theta_1, \ldots, \theta_p\}$.
2. Pro pevné $\lambda \in [0, 1]$
   - $$\mathbf{D}^* = \mathrm{diag}\{\lambda\theta_1 + (1-\lambda), \ldots, \lambda\theta_p + (1-\lambda)\}.$$
     To odpovídá modelu $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)\boldsymbol{\mathcal{I}}, \ \lambda \in [0, 1]$.
   - $$\mathbf{D}^* = \mathrm{diag}\{\lambda\theta_1 + (1-\lambda)s, \ldots, \lambda\theta_p + (1-\lambda)s\}, \quad s = \sum_{i=1}^{p} S_{ii}/p,$$
     To odpovídá modelu $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)s\boldsymbol{\mathcal{I}}, \ \lambda \in [0, 1]$.
3. $$\mathbf{S}^{*-1} = \mathbf{Q}\mathbf{D}^{*-1}\mathbf{Q}^T$$
4. Klasifikuj $\mathbf{Z}$ do skupiny k, pokud
$$(\bar{\mathbf{X}}_k - \mathbf{Z})^T\mathbf{S}^{*-1}(\bar{\mathbf{X}}_k - \mathbf{Z}) = \arg\min_{j=1,\ldots,K}\left\{(\bar{\mathbf{X}}_j - \mathbf{Z})^T\mathbf{S}^{*-1}(\bar{\mathbf{X}}_j - \mathbf{Z})\right\}.$$
5. Najdi optimální λ, které dá minimální klasifikační chybu.

Guo et al. (2005): LDA**

Klasifikuj $\mathbf{Z}$ do skupiny k, pokud

$$(\bar{\mathbf{X}}'_k - \mathbf{Z})^T \mathbf{S}^{*-1} (\bar{\mathbf{X}}'_k - \mathbf{Z}) = \arg \min_{j=1,\ldots,K} \left\{ (\bar{\mathbf{X}}'_j - \mathbf{Z})^T \mathbf{S}^{*-1} (\bar{\mathbf{X}}'_j - \mathbf{Z}) \right\}.$$

- Smrštěné odhady **průměrů**. V k-té skupině:

$$\bar{\mathbf{X}}_k = \operatorname{sgn}(\bar{\mathbf{X}}_k) \left(|\bar{\mathbf{X}}_k| - \Delta \right)_+$$

- Smrštěný odhad **varianční matice**:
  - Posílení hlavní diagonály:

$$\mathbf{S}^* = \lambda \mathbf{S} + (1 - \lambda) \boldsymbol{\mathcal{I}}, \quad \lambda \in [0, 1],$$

    kde $\mathbf{S}$ je empirická varianční matice.
  - Ekvivalentně:

Parametry smrštění Δ, λ. Křížová validace. Lepší klasifikační výsledky než PAM či LDA*.

Princip smrštění: další aplikace

Další **smrštěné** verze statistických metod:

- Korelační koeficient: Yao *et al.* (2008)
- Shluková analýza: Gao & Hitchcock (2010)
- Dvouvýběrový test: Shen *et al.* (2011)
- Mnohorozměrná analýza rozptylu: Tsai & Chen (2009)

Regularizovaná Mahalanobisova vzdálenost.

- Průměr $\Longrightarrow$ regularizovaný průměr.
- Varianční matice $\Longrightarrow$ regularizovaná varianční matice.

Popis genetické studie v Centru biomedicínské informatiky

Měření genových expresí přes celý genom ($p = 38\,590$ genových transkriptů):

- Cévní mozková příhoda (CMP): 24 pacientů.
- Kontroly (osoby bez kardiovaskulárního onemocnění): 24 pacientů.

Klasifikace do 2 skupin: predikce rizika budoucí mrtvice.

Příklad #1

Klasifikace: pacienti s CMP (24) vs. kontroly (24).
Exprese 38 590 genových transkriptů.

1. Redukce dimenze
   1. Prediction Analysis for Microarrays (PAM)
   2. Linear Models for Microarray Data (limma)
   3. Analýza hlavních komponent (PCA)
2. Klasifikační analýza
   1. Prediction Analysis for Microarrays (PAM)
   2. Prediction Analysis for Microarrays bez smrštění (PAM†)
   3. Lineární diskriminační analýza (LDA, LDA*)
   4. Kvadratická diskriminační analýza (QDA, QDA*)
   5. Logistická regrese (LR)

- Valenta Z., Kalina J., Kolář M., Zvárová J. (2013+): Exploring shrinkage approach in analyzing gene expression data. Submitted.

Příklad #1

Redukce dimenze	Klasif. metoda	Youdenův index 45 genů	20 genů	15 genů	10 genů
PAM	PAM	0.675	0.635	0.680	0.638
	LDA	0.306	0.595	0.629	0.704
	LDA*	0.399	0.658	0.670	0.735
	QDA	-	-	0.135	0.408
	QDA*	0.420	0.588	0.610	0.694
	LR	-	0.481	0.516	0.609
limma	PAM	0.591	0.588	0.603	0.579
	PAM†	0.629	0.681	0.679	0.629
	LDA	0.458	0.579	0.581	0.608
	LDA*	0.648	0.634	0.603	0.614
	QDA	-	-	0.101	0.329
	QDA*	0.610	0.683	0.661	0.618
	LR	-	0.501	0.564	0.560
PCA	PAM	-	0.231	0.245	0.216
	PAM†	-	0.230	0.243	0.220
	LDA	-	0.165	0.171	0.136
	LDA*	-	0.180	0.179	0.159
	QDA	-	-	0.076	0.154
	QDA*	-	0.133	0.118	0.101
	LR	-	0.145	0.106	0.143

Příklad #2

Senzitivita = pravděpodobnost pozitivního testu u nemocného pacienta.
Specificita = pravděpodobnost negativního testu u kontrolní osoby.
Youdenův index = senzitivita + specificita -1.

Křížová validace (*leave-one-out*):

Metoda	Youdenův index
LDA	Infeasible
PAM	0.833
LDA*	1.000
LDA**	1.000
PCA $\Longrightarrow$ LDA	0.542
PCA $\Longrightarrow$ LDA*	0.625
PCA $\Longrightarrow$ LDA**	0.708

- PCA používá 20 hlavních komponent.
- LDA*, LDA ** používá $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)\boldsymbol{\mathcal{I}}$.

Příklad #3

Data o metabolitech, které souvisí s rakovinou prostaty: $p = 518$ proměnných.

- Sreekumar *et al.* (2009): Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457** (7231), 910 – 914.

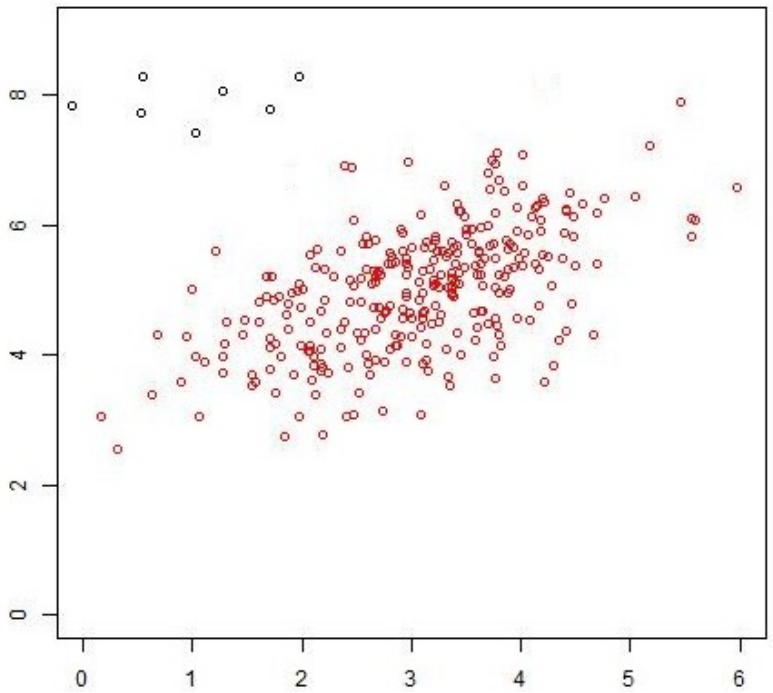
Klasifikace do 2 skupin:

- Benigní rakovina prostaty: 16 pacientů
- Ostatní: 26 pacientů

Metoda	Youdenův index
LDA	Nelze
PAM	0.822
LDA*	1.000
LDA**	1.000
PCA $\Longrightarrow$ LDA	0.899
PCA $\Longrightarrow$ LDA*	0.808
PCA $\Longrightarrow$ LDA**	0.923

20 hlavních komponent, $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)\boldsymbol{\mathcal{I}}$.

Robustní klasifikační analýza



Robustní statistické metody

- Motivace
- Bod selhání
- Odhady střední hodnoty a varianční matice

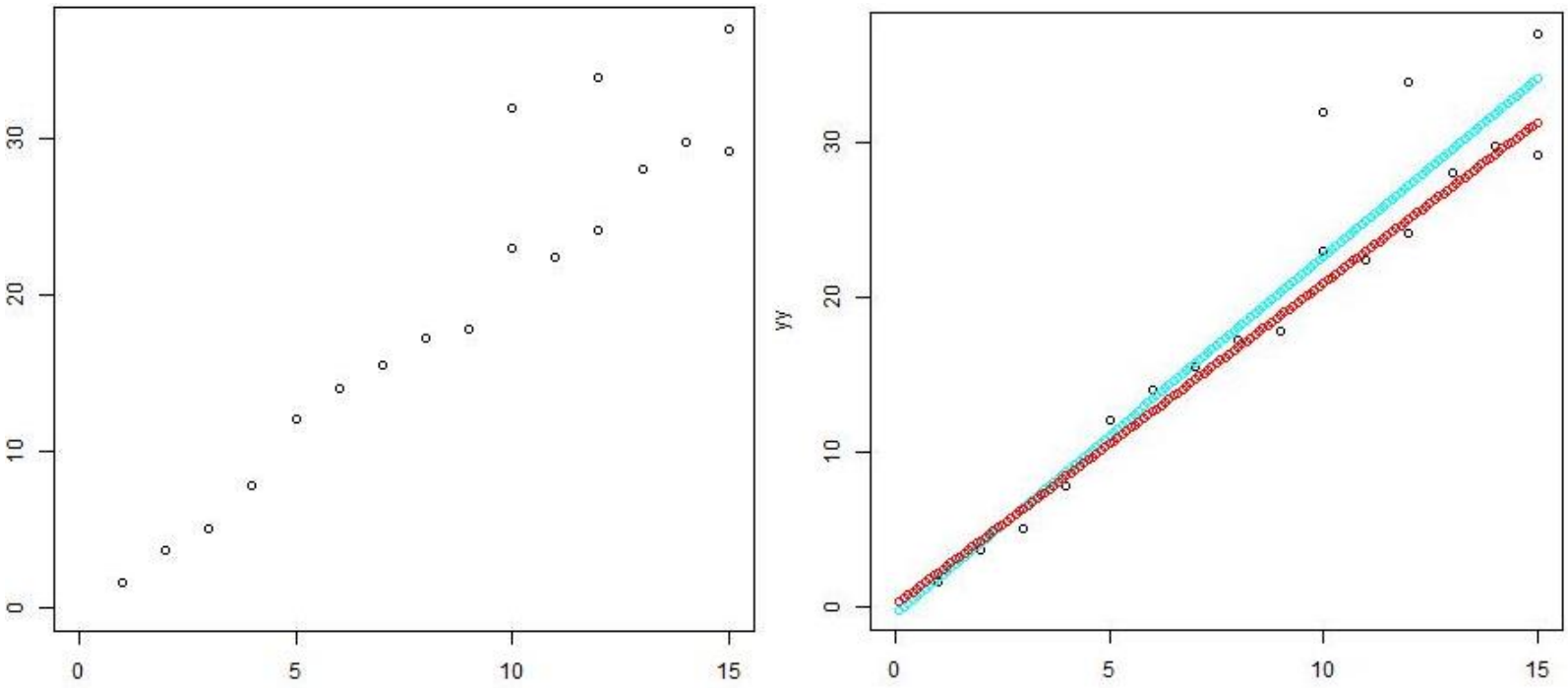
Regularizace a robustnost při klasifikační analýze genetických dat

- Optimální šablony
- Klasifikační analýza
- **Regrese**
- Selekce proměnných

1. Čížek P. (2011): Semiparametrically weighted robust estimation of regression models. *Computational Statistics and Data Analysis* **55**, 774 – 788.
2. Víšek J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika* **47**, 179 – 206.

Lineární regrese

Odhad metodou nejmenších čtverců.
Odhad metodou nejmenších vážených čtverců.



Metoda nejmenších čtverců: nerobustní, nevhodné pro velkou dimenzi.
Robustní metody: selhávají pro velkou dimenzi.
Metody pro velkou dimenzi: nerobustní.

Lineární regrese: Robustní odhad

Lineární regresní model $Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n.$

Nejmenší čtverce: kritika (citlivost k normalitě, odlehlým pozorováním).

Rezidua pro pevnou hodnotu $\mathbf{b} = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$:

$$u_i(\mathbf{b}) = y_i - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \ldots, n.$$

Uspořádáme druhé mocniny reziduí podle velikosti:

$$u^2_{(1)}(\mathbf{b}) \leq u^2_{(2)}(\mathbf{b}) \leq \cdots \leq u^2_{(n)}(\mathbf{b}).$$

Víšek (2002), Čížek (2011):
Odhad metodou **nejmenších vážených čverců** (*least weighted squares*, **LWS**):

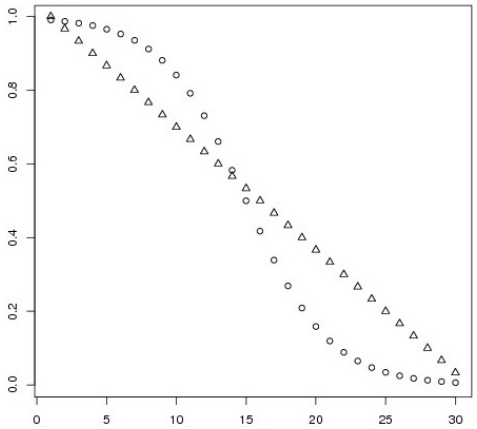
$$\mathbf{b}_{LWS} = \arg\min \sum_{i=1}^{n} w_i u^2_{(i)}(\mathbf{b}) \quad \text{přes} \quad \mathbf{b} = (b_1, \ldots, b_p)^T \in \mathbb{R}^p,$$

kde $w_1, \ldots, w_n$ jsou adaptivní váhy (spočítané na základě dat).

Lineární regrese: Robustní odhad

Příklady vah:

- Adaptivní (závislé na datech)
- Lineární
- Logistické



Lineární regrese: Regularizovaný odhad (LASSO)

- $$\arg \min_{\beta_1,\ldots,\beta_p} \left[\sum_{i=1}^{n} u_i^2\right] \quad \text{za podmínky} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\iff \arg \min_{\beta_1,\ldots,\beta_p} \left[\sum_{i=1}^{n} u_i^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right]$$

- $$\hat{\beta}_j = \text{sgn}(b_j^{LS}) \left(|b_j^{LS}| - \lambda\right)_+, \quad j = 1, \ldots, p,$$

  - $\mathbf{b}_{LS} = (b_1^{LS}, \ldots, b_p^{LS})^T$ je odhad $\boldsymbol{\beta}$ metodou nejmenších čtverců
  - $(x)_+$ označuje kladnou část x
  - Lagrangeův multiplikátor $\lambda > 0$

Hřebenová regrese

Jde o hřebenovou regularizaci matice $\mathbf{X}^T\mathbf{X}$.

$$\arg \min_{\beta_1,\ldots,\beta_p} \left[\sum_{i=1}^{n} u_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \quad \text{pro pevné } \lambda$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{\mathcal{I}})^{-1}\mathbf{X}^T\mathbf{Y},$$

kde $\boldsymbol{\mathcal{I}}$ je jednotková matice.

Vhodná volba $\lambda \geq 0$ závisí na t a v praxi se určuje křížovou validací.

Odhad je vhodný za multikolinearity, ale nerobustní.

Regularizovaný a robustní odhad v regresi

Při multikolinearitě: robustní odhady nejsou schopny správně odhalit odlehlé hodnoty.

Jurczyk (2010, 2012): **ridge least weighted squares** (RLWS)

$$\min \left(\sum_{i=1}^{n} w_i u_{(i)}^2(\mathbf{b}) + \lambda \sum_{j=1}^{p} b_j^2 \right)$$

Regularizace a robustnost při klasifikační analýze genetických dat

- Optimální šablony
- Klasifikační analýza
- Regrese
- **Selekce proměnných**

1. Auffarth B., López M., Cerquides J. (2010): Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images. In: Advances in Data Mining, Applications and Theoretical Aspects. *Lecture Notes in Computer Science* 6171, Springer, Berlin, 248 – 262.
2. Liu X., Krishnan A., Modry A. (2005): An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* **6**, Article 76.
3. Peng H., Long F., Ding C. (2005): Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions of Pattern Analysis and Machine Intelligence* **27** (8), 1226 – 1238.
4. Schäfer J., Strimmer K. (2005): A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4** (1), Article 32, 1 – 30.

Typy metod pro redukci dimenze

Redukce dimenze:

- Obvyklá (ne však nutná) procedura při analýze vysoce rozměrných dat
- V praxi se často provede nejprve redukce dimenze (mimo klasifikační kontext), až pak klasifikace: slabé!
- Redukce dimenze by měla být ušitá na míru pro klasifikaci.
- Oslabuje následnou klasifikační analýzu.

Různé metody - obecné typy:

- Statistické testování hypotéz (jde o uspořádání genů, ne samotné p-hodnoty).
- Redukce dimenze "obalená" kolem klasifikační metody, která se používá jako černá skříňka. Výpočetní složitost. Křížová validace. Hrozí přeučení. (*Wrappers.*)
- Redukce dimenze je "vložená" přímo do klasifikace. Vychází z vlastností klasifikátoru, který řídí proces hledání vhodných genů. (*Embedded methods.*)

Principal component analysis (PCA)

Principal component analysis (Pearson, 1901):

- p-dimensional data $\mathbf{X}_1, \ldots \mathbf{X}_n$
- $\mathbf{S}$ = covariance matrix
- A new observation $\mathbf{Z}$ is replaced by $\mathbf{a}_1^T\mathbf{Z}, \ldots, \mathbf{a}_p^T\mathbf{Z}$, where $\mathbf{a}_1, \ldots, \mathbf{a}_p$ are eigenvectors of $\mathbf{S}$.
- Numerically stable.
- For $n \ll p$ regularization needed.
- Data in groups: criticism of PCA (Dai *et al.*, 2006).

Regularized PCA: PCA*

- A new observation $\mathbf{Z}$ is replaced by $\mathbf{a}_1^{*T}\mathbf{Z}, \ldots, \mathbf{a}_p^{*T}\mathbf{Z}$, where $\mathbf{a}_1^*, \ldots, \mathbf{a}_p^*$ are eigenvectors of $\lambda\mathbf{S} + (1-\lambda)\mathbf{T}$, where $\lambda \geq [0, 1]$ and $\mathbf{T}$ is a target matrix.

Principal component analysis (PCA)

Special shrinkage with specific targets:

Theorem

In a special case with

- $\mathbf{T} = \mathcal{I}$, *or*
- $\mathbf{T} = s\mathcal{I}$ *with* $s = \sum_{i=1}^{p} S_{ii}/p$,

it holds

$$\mathbf{a}_1^* = \mathbf{a}_1, \ldots, \mathbf{a}_p^* = \mathbf{a}_p.$$

In these special cases, regularized PCA = classical PCA = non-robust PCA.

Variable selection

- Common procedure in the analysis of gene expression measurements.
- Tendency to pick highly correlated (redundant) genes.

Classification analysis.

- Curse of dimensionality.
- Dimension reduction: weaker classification performance.

Solution: **Minimum Redundancy Maximum Relevance** (**MRMR**) criterion.

- Gene set redundancy minimized.
- Too sensitive to noise or presence of outlying measurements.

Our aim: **improvement** of the MRMR dimension reduction,
exploiting modern shrinkage and robust statistical methodology.

MRMR criterion

Forward search: Genes are iteratively added to the set of selected genes.

- $\mathbf{Y}$ = response (labels of samples, correct classification result).
- S = set of selected genes.
- X_k $(k \in S)$ = expressions of the k-th gene (across patients).
- $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ = expressions of a candidate gene to be added to S.

Usual **relevance** criteria:

- Mutual information $I(Z, Y)$.
- F-test statistic of the analysis of variance $(Y \sim Z)$.
- Spearman rank correlation coefficient $|r_S(Z, Y)|$.

Usual **redundancy** criteria:

$$\frac{1}{|S|} \sum_{k \in S} |R_2(X_k, Z)|, \quad R_2 = \begin{cases} \text{Mutual information} \\ \text{Statistic of the Kolmogorov-Smirnov test} \\ \text{Statistic of the sign test} \end{cases}$$

MRMR criterion

The optimal gene set maximizes the **MRMR criterion**, which combines relevance and redundancy.
Usual MRMR criteria:

- $$\max \frac{\text{Relevance}}{\text{Redundancy}}$$
- $$\max\{\text{Relevance} - \text{Redundancy}\}$$
- $$\max\{\text{Relevance} - \beta \cdot \text{Redundancy}\},$$

  where the maximization is computed over all possible gene sets for a fixed (known) $\beta \in [0, 1]$.

MRMR criterion

Our combination of relevance and redundancy:
$\mathbf{Y}$= response (labels of samples, correct classification result).
S = gene set (relevant genes selected so far).
$\mathbf{X}_k$= expressions of the k-th gene in S across patients.
$\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ expressions of a candidate gene.

$$\max \left[|\text{Relevance}(\mathbf{Y}, \mathbf{Z})| - \beta \sum_{k \in S} |\text{Redundancy}(\mathbf{X}_k, \mathbf{Z})| \right],$$

where the maximization is computed over all possible gene sets **and** over $\beta \geq 0$.

Novel measures of relevance and redundancy:

- **Relevance:** robust correlation coefficient.
- **Redundancy:** shrinkage coefficient of multiple correlation.

Shrinkage approach to redundancy

Redundancy of a gene set: **multivariate** measure of association within a gene set.

Shrinkage correlation matrix (Schäfer and Strimmer, 2005):

$$\mathbf{R}^* = \lambda \mathbf{R} + (1 - \lambda)\boldsymbol{\mathcal{I}} \quad \text{with a shrinkage parameter } \lambda \in [0, 1],$$

where

- $\mathbf{R}$ = classical (unconstrained, unbiased) estimate of the correlation matrix,
- $\boldsymbol{\mathcal{I}}$ =unit matrix.

Analytical solution for optimal λ.

Shrinkage approach to redundancy

- S = gene set
- $\mathbf{X} = (X_{ij})_{i,j}$ = expressions of genes in S (patients $i = 1, \ldots, n$; genes $j = 1, \ldots, p$)
- $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ = expressions of a candidate gene
- $\mathbf{R_{ZX}} = (\mathrm{cor}(\mathbf{Z}, \mathbf{X}_1), \ldots, \mathrm{cor}(\mathbf{Z}, \mathbf{X}_p))^T$
- $\mathbf{R_{XX}} = \mathrm{cor}(\mathbf{X}, \mathbf{X})$

Coefficient of multiple correlation = measure of association between a particular gene and a set of genes:

$$\tilde{r}(\mathbf{Z}, \mathbf{X}) = \sqrt{\mathbf{R}_{\mathbf{ZX}}^T \mathbf{R}_{\mathbf{XX}}^{-1} \mathbf{R}_{\mathbf{ZX}}}.$$

Shrinkage approach to redundancy

Coefficient of multiple correlation = measure of association between a particular gene and a set of genes:

$$\tilde{r}(\mathbf{Z}, \mathbf{X}) = \sqrt{\mathbf{R}_{\mathbf{ZX}}^T \mathbf{R}_{\mathbf{XX}}^{-1} \mathbf{R}_{\mathbf{ZX}}}.$$

Shrinkage coefficient of multiple correlation:

$$\tilde{r}^*(\mathbf{Z}, \mathbf{X}) = \sqrt{\left(\mathbf{R}_{\mathbf{ZX}}^*\right)^T \left(\mathbf{R}_{\mathbf{XX}}^*\right)^{-1} \mathbf{R}_{\mathbf{ZX}}^*},$$

where $\mathbf{R}_{\mathbf{XX}}^*$ and $\mathbf{R}_{\mathbf{ZX}}^*$ are obtained as components of the shrinkage correlation matrix of the data

$$\begin{pmatrix} X_{11} & \cdots & X_{1p} & Z_1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{n1} & \cdots & X_{np} & Z_n \end{pmatrix}.$$

Robust approach to relevance

Aim: Highly robust correlation coefficient (based on the LWS).

For the data $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, let $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_n)^T$ denote the (normalized) weights obtained by the LWS estimator in the model

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, \ldots, n.$$

The robust correlation coefficient $r_{LWS}(\mathbf{X}, \mathbf{Y})$ is defined as the weighted

$$r_W(\mathbf{X}, \mathbf{Y}; \tilde{\mathbf{w}}) = \frac{\sum_{i=1}^n \tilde{w}_i (X_i - \bar{X}_{\tilde{\mathbf{w}}})(Y_i - \bar{Y}_{\tilde{\mathbf{w}}})}{\sqrt{\sum_{i=1}^n [\tilde{w}_i (X_i - \bar{X}_{\tilde{\mathbf{w}}})^2] \sum_{j=1}^n [\tilde{w}_j (Y_j - \bar{Y}_{\tilde{\mathbf{w}}})^2]}},$$

where $\bar{X}_{\tilde{\mathbf{w}}} = \sum_{i=1}^n \tilde{w}_i X_i$ and $\bar{Y}_{\tilde{\mathbf{w}}} = \sum_{i=1}^n \tilde{w}_i Y_i$.

- High breakdown point (= high robustness) for contaminated normal distribution.
- High efficiency for normal distribution.

Methods

- Mutual information (for discretized data, based on comparison with the mean)
- r = Pearson correlation coefficient
- r_S = Spearman rank correlation coefficient
- r_{LWS} = robust correlation coefficient based on the least weighted squares estimator
  (weights: adaptive; linear; logistic)
- K-S = Kolmogorov-Smirnov test (p-value)
- Sign test (p-value)
- Mult. $|\tilde{r}|$ = coefficient of multiple correlation
- Shrinkage mult. $|\tilde{r}^*|$ = shrinkage coefficient of multiple correlation

Example # 4

Results of cross-validation (patients vs. kontroly, $n = 48$):

Relevance	Redundancy	Youden's index
Mutual info.	Mutual info.	0.92
$\|r\|$	$\|r\|$	1.00
$\|r_S\|$	$\|r_S\|$	0.96
$\|r\|$	K-S test	0.84
$\|r\|$	Sign test	0.84
$\|r\|$	Mult. $\|\tilde{r}\|$	1.00
$\|r\|$	Shrinkage mult. $\|\tilde{r}^*\|$	1.00
$\|r_{LWS}\|$ (linear weights)	Shrinkage mult. $\|\tilde{r}^*\|$	1.00
$\|r_{LWS}\|$ (logistic weights)	Shrinkage mult. $\|\tilde{r}^*\|$	1.00
$\|r_{LWS}\|$ (adaptive weights)	Shrinkage mult. $\|\tilde{r}^*\|$	1.00

Classification results obtained by LDA over 10 genes selected by various MRMR criteria:
Sensitivity = probability of a positive test for a patient with disease.
Specificity = probability of a negative test for a control.
Youden's index = sensitivity + specificity -1.

Sensitivity study

Data (average gene expressions) **contaminated** by noise under various distributional models. (Noise independent on gene and patient.)

Noise 1: **Normal** $\mathrm{N}(0, \sigma^2 = 0.1)$.

Noise 2: **Contaminated normal**

$$\Delta F + (1 - \Delta) G,$$

where $\Delta = 0.85$, $F \sim \mathrm{N}(0, \sigma^2 = 0.01)$, $G \sim \mathrm{N}(0, \sigma^2 = 1)$.

Noise 3: **Cauchy** with probability density function

$$f(x) = \frac{c}{\pi(x^2 + c^2)}, \quad x \in \mathbb{R}, \quad c = 0.002.$$

Example # 4: Sensitivity study

MRMR (gene selection)

$\Longrightarrow$ linear discriminant analysis.

Leave-one-out cross validation, average **Youden's index**:

Relevance	Redundancy	Noise 1 Normal	Noise 2 Contam. normal	Noise 3 Cauchy
Mutual info.	Mutual info.	0.58	0.75	0.83
$\lvert r\rvert$	$\lvert r\rvert$	0.83	0.71	0.92
$\lvert r_S\rvert$	$\lvert r_S\rvert$	0.83	0.83	0.92
$\lvert r\rvert$	K-S	0.79	0.67	0.79
$\lvert r\rvert$	Sign test	0.67	0.83	0.75
$\lvert r\rvert$	Mult. $\lvert\tilde{r}\rvert$	0.71	0.75	0.92
$\lvert r\rvert$	Shrinkage mult. $\lvert\tilde{r}^*\rvert$	0.79	0.71	0.88
$\lvert r_{LWS}\rvert$ (linear weights)	Shrinkage mult. $\lvert\tilde{r}^*\rvert$	1.00	1.00	0.96
$\lvert r_{LWS}\rvert$ (logistic weights)	Shrinkage mult. $\lvert\tilde{r}^*\rvert$	1.00	1.00	0.96
$\lvert r_{LWS}\rvert$ (adaptive weights)	Shrinkage mult. $\lvert\tilde{r}^*\rvert$	1.00	1.00	1.00

Discussion of results

Results: arguments in favor of the novel relevance and redundancy measures.

Best method:
Minimum Shrinkage Redundancy Maximum Robust Relevance (MSRMRR)

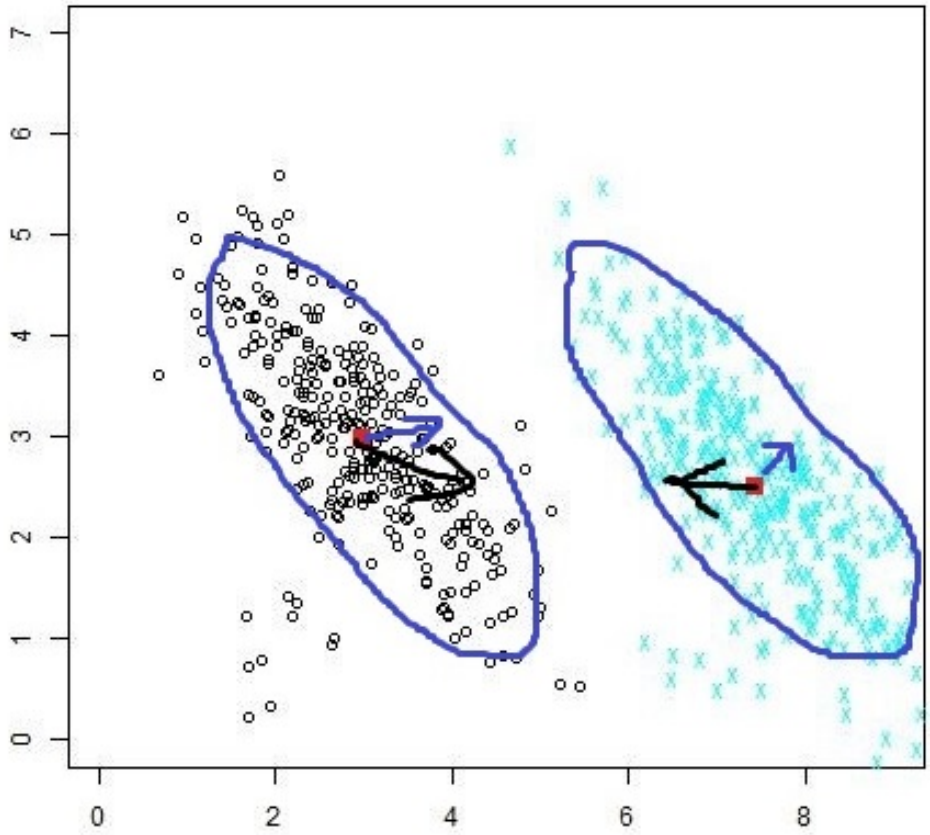
- Relevance: robust correlation coefficient r_{LWS}.
- Redundancy: shrinkage coefficient of multiple correlation.

This is a first MRMR criterion based on robust and shrinkage statistics.
Advantages:

- Suitable for high-dimensional data.
- Robustness properties.
- Especially preferable for gene expression measurements contaminated by noise or outliers.

Analogous results are obtained with other classification methods.

Budoucí výzkum: Robustní regularizovaná klasifikační analýza



Deformace Mahalanobisovy vzdálenosti:

- Černě: regularizace.
- Modře: robustní postup.

Regularizace a robustnost při klasifikační analýze genetických dat

$\Longrightarrow$ DĚKUJI ZA POZORNOST $\Longleftarrow$