Vzorová úloha 4.16 *Postup vícerozměrné kalibrace*

Postup vícerozměrné kalibrace ukážeme na úloze **C4.10** *Vícerozměrný kalibrační model kvality bezolovnatého benzinu*. Dle následujících kroků na základě naměřených NIR spekter sestrojte vícerozměrný kalibrační model pro jednu z charakteristik kvality, tj. koncentraci jedné ze složek bezolovnatého benzinu. Model pak užijte ke kontrole kvality benzinů z kontinuální produkce:

(a) Sestrojte nejlepší jednorozměrný kalibrační model a použijte ho ke srovnání s výsledky vícerozměrné kalibrace.

(b) Prozkoumejte vícerozměrná spektrálních data metodou hlavních komponent (PCA). Interpretujte rozptylové grafy komponentního skóre, grafy zátěží a matici vlastních čísel. Nalezněte i odlehlá spektra.

(c) Sestrojte vícerozměrný kalibrační model metodou PCR a PLS. Interpretujte graf chyby predikce *versus* počet latentních proměnných. Zvolte optimální počet proměnných v modelu. Interpretujte souvislost mezi latentními proměnnými a odlehlými body.

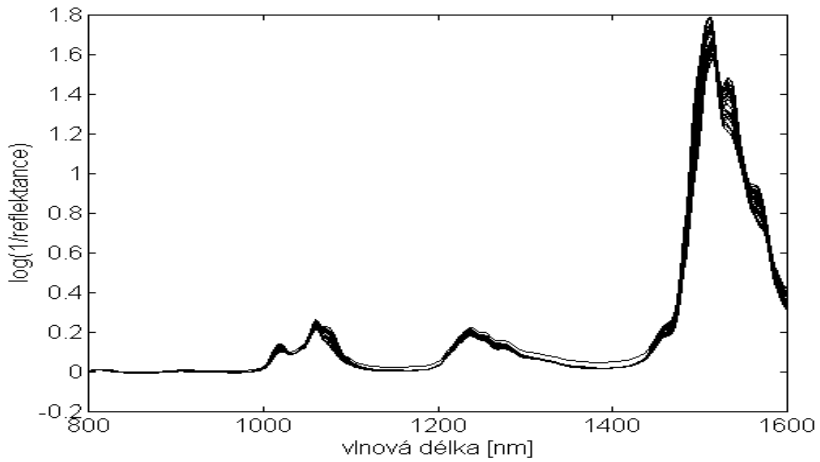
(d) Z absorbanční matice $\boldsymbol{X}$ eliminujte odlehlá spektra, jež prokazatelně zhoršují kvalitu kalibračního modelu. Sestrojte finální PCR a PLS model.

(e) Ke kalibraci aplikujte krokovou vícerozměrnou lineární regresi, sloužící zde jako jako alternativa k metodám s latentními proměnnými. Diskutujte statistickou významnost jednotlivých proměnných v modelu s ohledem na výsledky t-testu. Metodou příčné validace nalezněte na závěr optimální model. Porovnejte takto dosažené výsledky s výsledky z metody regrese na hlavních komponentách PCR a částečnými nejmenšími čtverci PLS.

(f) Zhodnoťte všechny výsledky a doporučte metodiku, která bude aplikována v praxi.

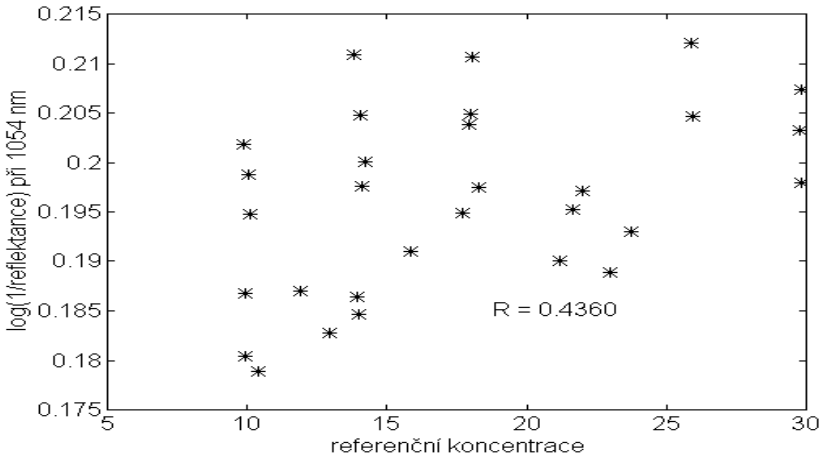
Data: Naměřená NIR spektra analyzovaných bezolovnatých benzinů úlohy C4.10 jsou na obr. 4.25.

	l_1	l_2	l_3	l_4	l_5
vzorek1	0.0033390	0.0047277	0.0062653	0.0077811	0.0090141
...	...	...	...	...	...
...	...	...	...	...	...
vzorek 30	0.0056775	0.0058778	0.0066916	0.0076192	0.0087291



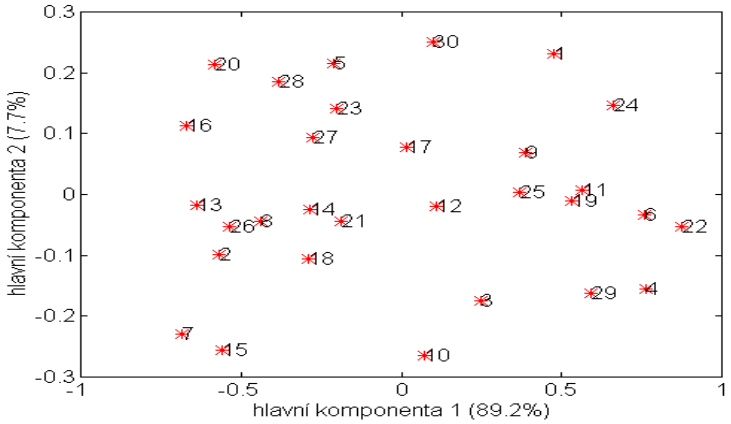
Obr. 4.25 NIR spektra vzorků benzinů naměřená v oblasti 800-1600 nm.

Řešení: Vzhledem k velkému rozsahu dat, a to 700 spekter při 30 vlnových délkách, je v následující tabulce uvedeno pouze prvních 5 hodnot signálu pro první a poslední spektrum datového souboru C4.10.

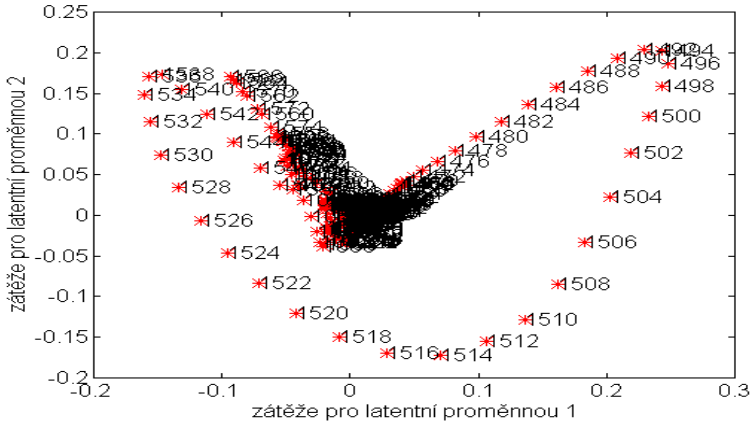


Obr. 4.26 Odezva vzorků při vlnové délce 1054 nm *versus* jeho referenční koncentrace.

Z obrázku obr. 4.25 je patrné, že nejméně jedno spektrum je odlišné od ostatních, a to zejména ve spektrální oblasti 1100-1220 nm a 1250-1470 nm. Toto spektrum by mělo být identifikovatelné v grafu hlavních komponent.



Obr. 4.27 Graf hlavních komponent 1 a 2. Čísla objektů v grafu odpovídají číslům vzorků.

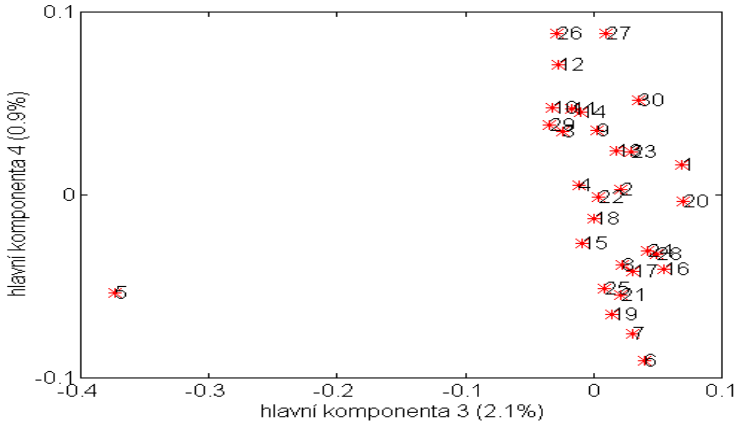


Obr. 4.28 Graf zátěží. Čísla objektů v grafu odpovídají číslům vzorků.

(a) *Jednorozměrná kalibrace:* použitím korelačního koeficientu byla vybrána vlnová délka 1054 nm, která poskytuje nejlepší jednorozměrný model. Jak je patrné z obr. 4.26 tento model není však v laboratoři použitelný. Korelační koeficient mezi měřením a kalibrovanou koncentrací je 0.4360 a vypočtená střední kvadratická chyba predikce RMSEP je 6.00. Pokud by neexistovala možnost použít vícerozměrnou kalibraci, analytický problém by nebyl řešitelný.

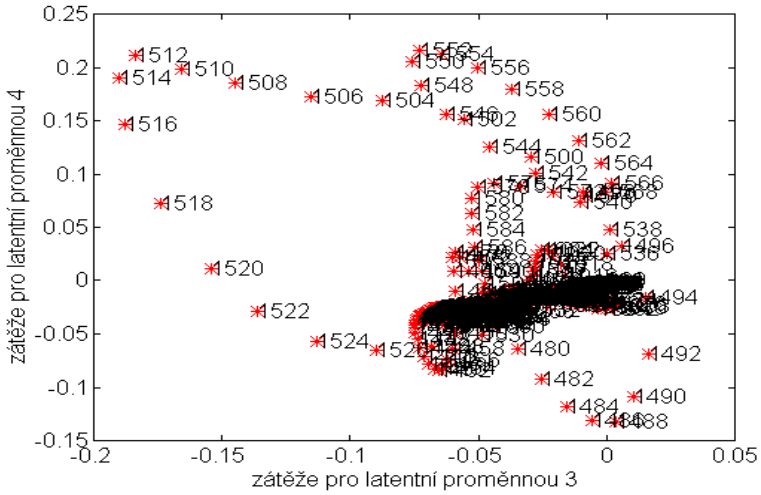
(b) *Analýza hlavních komponent (PCA)*: z matice vlastních čísel plyne, že první hlavní komponenta vysvětluje 89.2% celkového rozptylu zdrojové spektrální matice $\boldsymbol{X}$, druhá hlavní komponenta 7.7 %, třetí hlavní komponenta 2.1% a čtvrtá hlavní komponenta 0.9%. Graf

komponentního skóre na první hlavní komponentě *versus* komponentního skóre na druhé hlavní komponentě je prezentován na obr. 4.27. Graf odpovídajících zátěží je na obr. 4.28. Hodnoty v grafu odpovídají vlnovým délkám.



Obr. 4.29 Graf hlavních komponent 3 a 4. Bod č. 5 je odlehlý na komponentě 3.

Největší váhu v při konstrukci první hlavní komponenty mají originální proměnné kolem vlnové délky 1496 resp. 1536 nm a při konstrukci druhé latentní proměnné originální proměnné okolo vlnových délek 1492 a 1514 nm, obr. 4.28.

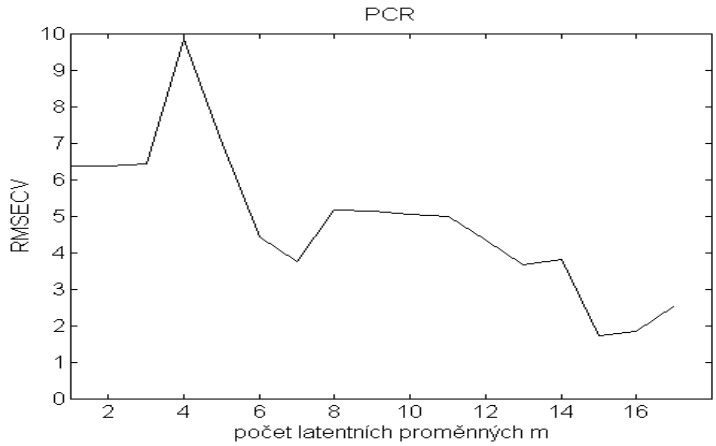


Obr. 4.30 Zátěže pro hlavní komponent 3 a 4.

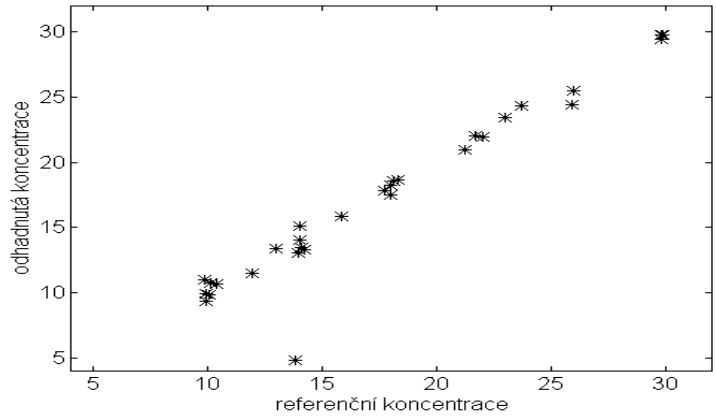
Toto zjištění je v souladu s originálními spektry z obr. 4.25, protože uvedené vlnové délky odpovídají maximům absorpčních pásů a skutečně reprezentují maximum rozptylu v datech. Graf prvních dvou hlavních komponent kromě toho ukazuje rovnoměrné rozložení bodů v prostoru. Žádná abnormalita v grafu nebyla zjištěna. Naproti tomu, rozptylový graf komponentních skóre na třetí hlavní komponentě *versus* čtvrtá hlavní komponenta, obr. 4.29, ukazuje jednoznačnou odlehlost bodu 5 na třetí latentní proměnné.

Odlehlost spektra č. 5 je způsobena hlavně rozdílem v odezvách kolem vlnové délky 1514 nm, obr. 4.25.

(c) *PCR a PLS kalibrace*: obr. 4.31 ukazuje závislost velikosti střední kvadratické chyby predikce na počtu latentních proměnných v modelu. Na křivce jsou patrná dvě minima. První, při 7 latentních proměnných, odpovídá chybě predikce RMSCV > 3.5, což je ještě velká chyba. Druhé, při 15 latentních proměnných, reprezentuje nerobustní řešení, protože tolik proměnných v modelu znamená modelování velké části šumu. Odlehlý bod č. 5 ovlivňuje predikci tak zásadním a negativním způsobem, že musí být vyloučen a model sestrojen znovu.

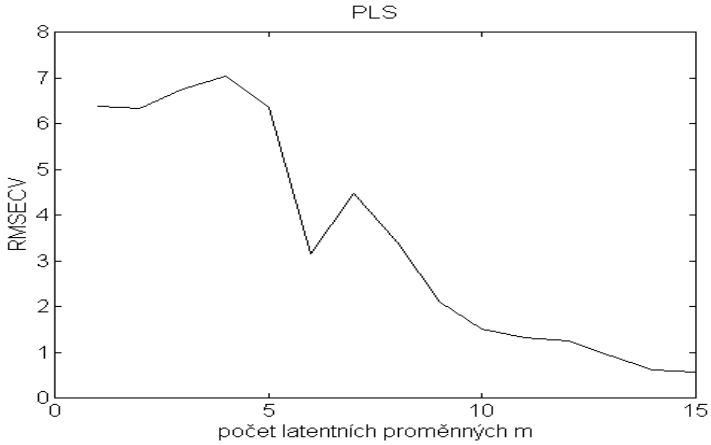


Obr. 4.31 Střední kvadratická chyba predikce RMSECV *versus* počet latentních proměnných.

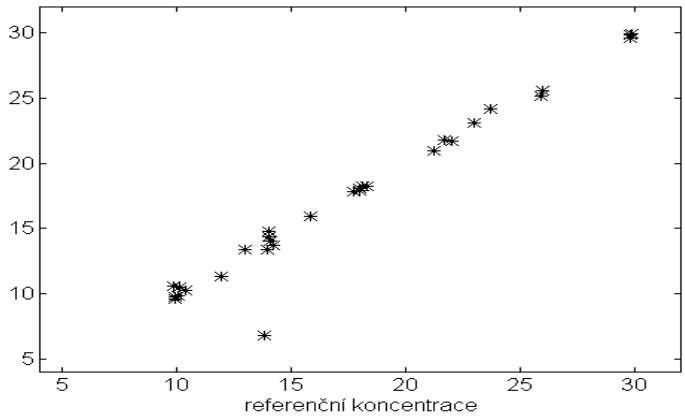


Obr. 4.32 Odhadnutá *versus* referenční koncentrace vzorků. Predikované hodnoty byly stanoveny metodou příčné validace.

Obrázek odhadnutých *versus* referenčních koncentrací, obr. 4.32, potvrzuje uvedený fakt: odhadnutá koncentrace pro bod 5 je výrazně nižší, než je její odpovídající referenční hodnota. Obr. 4.33 ukazuje, že predikce dosažená metodou částečných nejmenších čtverců PLS na datech zahrnujících bod 5 je podobná regresi hlavních komponent PCR.



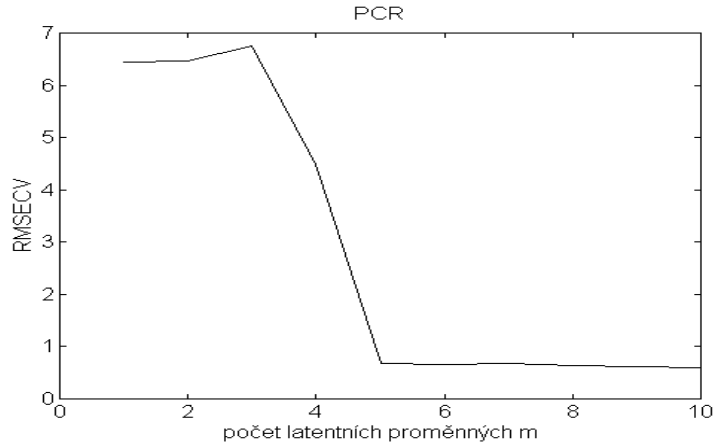
Obr. 4.33 Střední kvadratická chyba predikce RMSECV *versus* počet latentních proměnných.



Obr. 4.34 Graf predikovaných versus referenčních koncentrací kalibračních vzorků. Predikované hodnoty byly stanoveny metodou příčné validace.

Obr. 4.34 potvrzuje, že bod č. 5 by měl být z kalibračních dat eliminován.

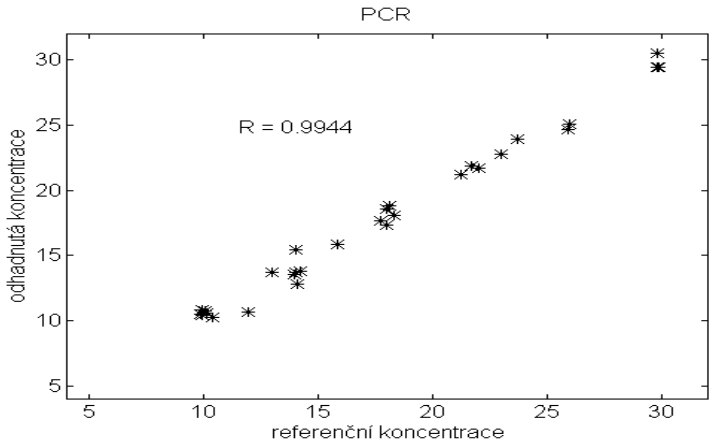
(d) Jelikož absorbanční spektrum č. 5 prokazatelně zhoršuje kvalitu kalibračního modelu, bylo odstraněno z matice ***X***. Obr. 4.35 ukazuje rozptylový graf komponentních skóre na třetí hlavní komponentě *versus* skóre na čtvrté hlavní komponentě po této eliminaci. Výsledkem je téměř rovnoměrné rozložení bodů v prostoru. Jelikož odlehlý bod představoval poměrný silný zdroj rozptylu, po jeho odstranění relativní významnost třetí hlavní komponenty poklesla z 2.1 na 1.0%, viz obr. 4.27 a obr. 4.29.



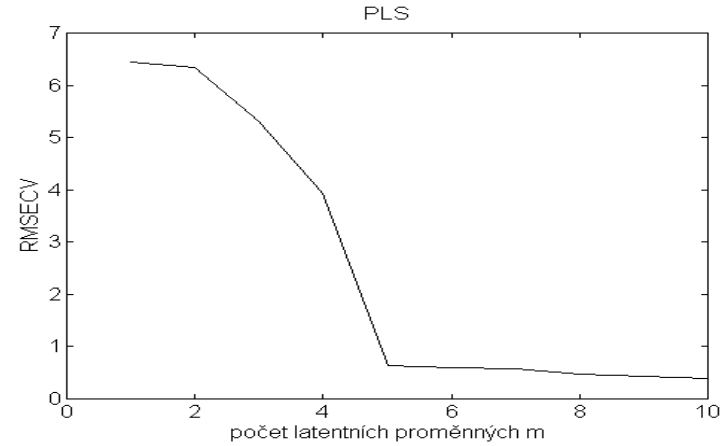
Obr. 4.35 Optimalizace počtu proměnných ve finálním modelu. RMSECV *versus* počet latentních proměnných použitý ke konstrukci modelu.

Po eliminaci odlehlého bodu je třeba zopakovat postup optimalizace počtu latentních proměnných v modelu metodou příčné validace. Graf chyby predikce *versus* počet hlavních komponent v PCR modelu je zobrazen v obr. 4.35. Graf ukazuje přijatelnější průběh než v případě dat s odlehlým měřením, obr. 4.31. Po dosažení počátečního maxima střední kvadratická chyba predikce prudce klesá. Minimum je dosaženo při 5 hlavních komponentách, což je mnohem nižší počet než 15 na obr. 4.31.

Graf predikovaných *versus* referenčních koncentrací dosažených s optimálním PCR modelem, obsahujícím 5 latentních proměnných je prezentován v obr. 4.36. Zmizel odlehlý bod č. 5 a střední kvadratická chyba predikce poklesla na hodnotu 0.67. Korelační koeficient mezi predikovanými a referenčními koncentracemi dosáhl hodnoty 0.9944.



Obr. 4.36 Koncentrace odhadnuté pomocí PCR modelu s 5 hlavními komponentami.



Obr. 4.37 Optimalizace počtu latentních proměnných v PLS modelu.

PLS kalibrace poskytuje podobné výsledky jako PCR, obr. 4.37. Optimální model obsahuje 5 latentních proměnných. Odpovídající chyba predikce je 0.64. Graf částečných nejmenších čtverců odhadnutých PLS *versus* referenčních koncentrací je podobný grafu na obr. 4.36.

Finální modely, určené regresí na hlavních komponentách PCR, resp. metodou částečných nejmenších čtverců PLS jsou sestrojeny použitím 5 latentních proměnných. Oba modely poskytují odhad koncentrace neznámých vzorků, zatížený absolutní chybou menší než 0.7.

(e) *Kalibrace krokovou vícenásobnou lineární regresí:* postupná selekce proměnných byla provedena dle následujících kroků:

1. krok: byla vybrána proměnná č. 127, odpovídající vlnové délce 1052 nm. S použitím Studentova t-testu bylo potvrzeno, že tato proměnná vysvětluje významnou část rozptylu kalibrované koncentrace (t_{exp} = 2.98 je větší než kritický kvantil t_{crit} = 2.05). Dopředná selekce proměnných v modelu proto pokračuje.

2. krok: nejvyšší korelační koeficient s vektorem koncentračních reziduí poskytuje měření č. 259, tj. vlnová délka 1316 nm. Studentův t-test potvrdil, že obě proměnné (č. 127 a 259) jsou významné, protože pro obě je experimentální hodnota $t_{exp} > 15$. Dopředná selekce proměnných v modelu proto pokračuje.

3. krok: další kandidátskou proměnnou je měření 96, tj. vlnová délka 990 nm. I zde byla při aplikaci t-testu prokázána významnost všech 3 proměnných v modelu.

4. krok: byla vybrána proměnná č. 221, odpovídající vlnové délce 1240 nm. Potvrzena významnost všech členů v modelu. Selekce pokračuje.

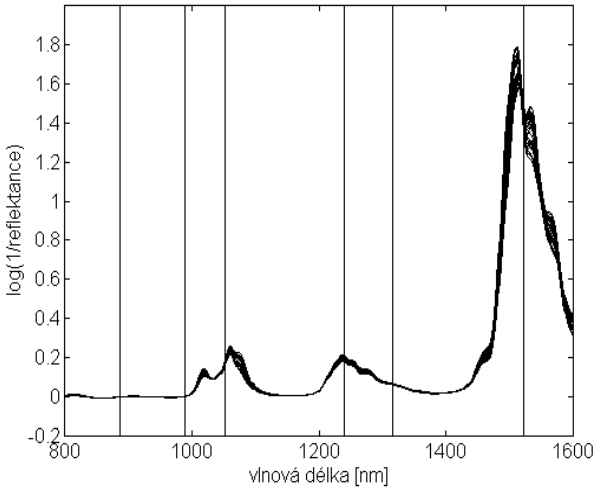
5. krok: byla vybrána proměnná č. 45, tj. vlnová délka 888 nm. Potvrzena významnost všech proměnných v modelu.

6. krok: byla nalezena proměnná č. 362, tj. 1522 nm. Potvrzena významnost členů. Dopředná selekce pokračuje.

7 krok: byla vybrána proměnná č. 8 při 814 nm. Významnost této proměnné nebyla potvrzena. Ukončení algoritmu.

Finální model krokovou vícenásobnou lineární regresí obsahuje proměnné č. 45, 96, 127, 221, 259 a 362, což odpovídá vlnovým délkám 888, 990, 1052, 1240, 1316 a 1522. Poloha proměnných v NIR spektru je zobrazena v obr. 4.38.

Z obr. 4.38 je patrné, že selekce proměnných s pomocí matematických metod je nezastupitelná. Prosté vizuální porovnání spekter analytikovi nedovoluje vybrat kombinaci proměnných, která by vedla ke smysluplným kalibračním výsledkům.



Obr. 4.38 NIR spektra benzinů použitá ke kalibraci s vyznačením 6 vlnových délek vybraných Stepwisovou metodou vícenásobné lineární regrese.

Příčná validace potvrdila, že všech 6 proměnných, vybraných krokovou metodou lineární regrese je významných. Minimální střední kvadratická chyba predikce je RMSECV = 0.446. Kdyby v modelu nebyla obsažena poslední vybraná proměnná, tj. č. 362, střední kvadratická chyba predikce by dosáhla hodnoty 0.67.

(f) *Závěr*: Použitím metody částečných nejmenších čtverců PCA bylo zjištěno, že naměřená blízká infračervená spektra NIR obsahují jedno odlehlé měření. Kalibrační výsledky toto zjištění potvrdily. PCR i PLS model se po odstranění odlehlého spektra výrazně zlepšil a zjednodušil. Počet hlavních komponent klesl z 15 na 5 komponent.

Hlavní komponenty použité ke kalibraci byly analyzovány. Grafy zátěží pro sledované komponenty vykazují maxima při vlnových délkách odpovídajících hlavním absorpčním pásům v původním spektru. To je očekávané zjištění a dokladuje logickou konstrukci hlavních komponent.

K optimalizaci počtu proměnných v modelu metodou regrese hlavních komponent PCR a částečných nejmenších čtverců PLS byla použita metoda příčné validace. Absolutní chyba predikce, obdržená s modelem obsahujícím 5 latentních proměnných, je menší než 0.7. Korelace predikovaných koncentrací s referenčními je velmi dobrá. Korelační koeficient je vyšší než 0.994.

Aplikace krokové metody vícenásobné lineární regrese prokázala, že i kalibrační model založený na podsouboru původních proměnných může dávat dobré výsledky. Ve studovaném případě dokonce lepší než PCR nebo PLS. Z důvodu menší robustnosti takového modelu k posunu vlnových délek, ke změnám v citlivosti přístroje a k linearitě odezvy by se v praxi tento model uplatnil pouze v krátkém časovém horizontu. Pokud by cílem kalibrace bylo používat model v průběhu např. 1 roku, pak by byl upřednostněn robustnější model PCR nebo PLS.