Partial Least Squares regrese (PLS-R)

Menu:	QCExpert	Prediktivní metody	Partial Least Squares

Modul PLS regrese poskytuje uživateli jednu z nejvýkonnějších současných výpočetních nástrojů pro vyhodnocování dvojice vícerozměrných proměnných, mezi nimiž se předpokládá možná lineární závislost jak unitř jedné, či druhé vícerozměrné proměnné, tak mezi oběma proměnnými navzájem. Tato výpočetně poměrně náročná metodika umožňuje vysvětlovat a predikovat jednu skupinu proměnných pomocí jiné skupiny proměnných. Velké množství aplikací nalezne metoda PLS regrese v řízení a plánování jakosti, ve výrobních technologiích, designu a optimalizace vlastností produktů, při vývoji nových produktů, marketingových studiích, ve výzkumu při vyhodnocování experimentů, při klinických studiích. Příkladem může být modelování vztahů mezi technologickými parametry při výrobě a parametry kvality produktu, nebo mezi chemickým složením a fyzikálními či biologickými vlastnostmi. Mezi typické otázky z technologické praxe, na něž často mohou postupy PLS odpovědět patří například:

Má čistota suroviny vliv na pevnost produktu?
Co se stane, zvýší-li se teplota při zpracování?
Dá se zvýšit stabilita produktu snížením otáček?
Které procesní parametry nejvíce ovlivňují pevnost?
Na jaké hodnoty nastavit procesní parametry, aby bylo dosaženo požadovaných vlastností?
Co způsobilo pokles parametru?
Čím a jak se od sebe liší jednotlivé výrobní šarže?
Jak zvýšit stabilitu/kvalitu?
Jak zvýšit pevnost/hodnotu/konkurenceschopnost?
Které vstupní parametry jsou rozhodující pro kvalitu?
Které procesní parametry jsou rozhodující pro kvalitu?

Matematická podstata metody PLS regrese

Tabulku naměřených hodnot *p* veličin (sloupců) s *n* řádky označme jako matici $\mathbf{X}(n \times p)$, a odpovídající tabulku se stejným počtem řádků *n* ale s *q* veličinami označme $\mathbf{Y}(n \times q)$, sloupce **X** a **Y** se vycentrují (odečtou se sloupcové průměry). Abychom extrahovali maximum informace z *p*- a *q*-rozměrných matic do prostoru s nižší dimenzí *k*, rozložíme **X** a **Y** na součin ortogonální matice **T** ($n \times k$) a **U** ($n \times k$), s koeficienty **P** a **Q**

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$$

při současné maximalizci korelace mezi **T** a **U**. Požadovanou dimenzi *k*, $1 < k \le \min(p, q)$ zvolíme sami, například na základě grafu poklesu součtu čtverců (scree plot) tak, že zvolíme takové *k*, nad nímž již součet čtverců výrazně neklesá, viz dále. Příliš vysoká hodnota *k* je obvykle zbytečná: snižuje stabilitu predikce modelu a přitom nezískáme mnoho užitečné informace. Příliš nízká hodnota vede sice ke stabilnímu, ale příliš hrubému modelu a ztrátě užitečné informace obsažené v datech. Více o výběru *k* viz dále. Šum a irelevantní informační smetí obsažené v každých měřených datech se uklidí jako do popelnice do reziduálních matic **E** a **F**. Rozkladem **U** = **TB** (kde **B** je čtvercová diagonální matice) získáme nástroj pro výpočet (odhad) **Y** z **X** ale zároveň **X** z **Y**, protože model PLS-R je symetrický, stačí formálně zaměnit obě tabulky.

$$\hat{\mathbf{Y}} = \mathbf{TBQ}^T,$$

T se zde vypočítá z nových dat **X**, $\mathbf{T} = \mathbf{XP}^-$ (zde $\mathbf{P}^-$ značí zobecněnou, např. Moore-Penroseovu pseudoinverzi obdélníkové matice **P**), případně z nových specifikací, parametrů nové suroviny, a podobně, pro něž potřebujeme predikci všech parametrů **Y**. Navíc je zde vidět vnitřní vazbu mezi **X** a **Y**. Označíme-li $\mathbf{W} = \mathbf{BQ}^T$, můžeme původní dvojici vztahů přepsat do tvaru

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{T}\mathbf{W} + \mathbf{F}$$

takže data **X** a **Y** jsou svázány pomocí společné matice skórů **T**, což je vlastně ortogonalizovaná původní matice **X** v obecně obecně menším počtu dimenzí s maximálním informačním obsahem původně obsaženým v **X**, odstraněným šumem (ten je v matici **E**), a současně maximální relevancí (kovariancí) s daty v matici **Y**. Pomocí vztahu

$$\mathbf{A} = \mathbf{P}^{-}\,\mathbf{B}\,\mathbf{Q}^T$$

lze rekonstruovat koeficienty klasického regresního modelu s vícerozměrnou odezvou **Y** = **AX**. Sloupce $\mathbf{a}_i$ matice **A** pak obsahují lineární koeficienty (absolutní členy jsou nulové) modelů $\mathbf{y}_i = \mathbf{X}\mathbf{a}_i$, kde $\mathbf{y}_i$ je *i*-tý sloupec matice **Y**. Hodnoty koeficientů nebudou obvykle číselně zcela totožné s koeficienty získanými klasickou lineární regresí. Budou vychýlené, avšak zároveň zkrácené (shrinked), s nižšími rozptyly, a tedy obecně stabilnější.

Jak bylo řečeno, tato metoda hledá vztah mezi dvěma fenomény popsanými vícerozměrnými číselnými vektory, měly by proto podle toho být vybírány i sloupce matic **X** a **Y**. Typickým příkladem je tedy matice **X** obsahující procesní, technologické parametry při výrobě jednotlivých kusů nebo šarží produktu a matice **Y** obsahující fyzikální parametry odpovídajících hotových produktů, jejich odchylky od specifikací apod. Jiným příkladem je matice **X** obsahující klimatické a chemické popisy prostředí různých lokalit a matice **Y** s biologickými parametry mikroorganismů, vegetace či fauny v těchto lokalitách. Pomocí predikce je možné získat odhady neznámých veličin matice **Y** na základě známých hodnot **X**.

Validace modelu

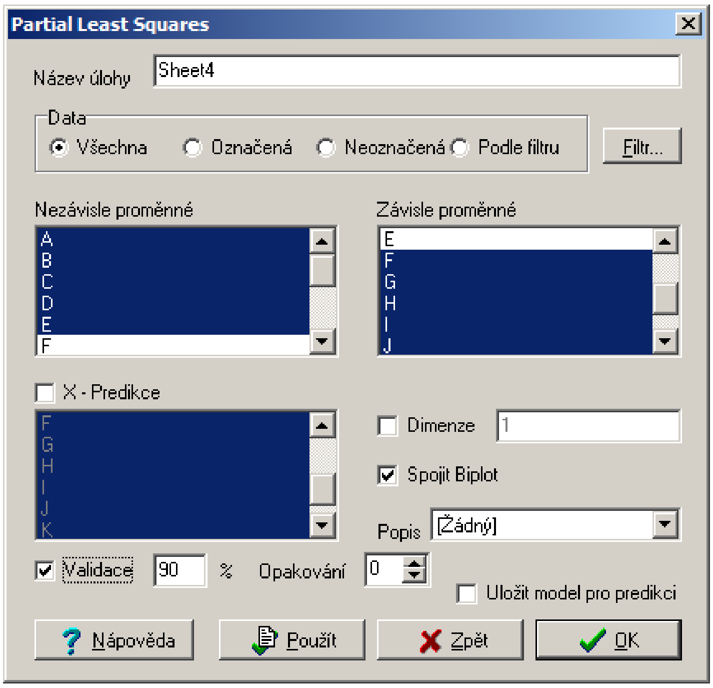
Kvalitu konkrétního modelu PLS lze posoudit na základě jeho schopnosti predikovat hodnoty **y** pro dané hodnoty **x**. Toho se využívá při různých validačních postupech, nekdy nazývaných cross-validace. Princip validace modelu je stejný jako u neuronových sítí. spočívá v tom, že před výpočtem „zatajíme“ část dat, které označíme jako testovací nebo validační. Ze zbytku dat, označených jako trénovací, vypočítáme parametry modelu PLS. Pak validační data „odtajníme“ a posoudíme, zda model dobře popisuje i tato validační data. Validační která by měla mít stejnou povahu, fyzikální mechanismus vzniku, rozsah hodnot **x**, a tedy i stejný model jako data trénovací. Pro validační data se pak konstruují diagnostické grafy, z nichž jednoduše usuzujeme, zda je model vhodný pro všechna data. Pokud model popisuje dobře jem trénovací data a nikoliv data validační, obvykle to znamená, že máme málo dat (řádků), nebo že jsme zvolili příliš velký podíl validačních dat. Obvykle volíme podíl validačních dat mezi 10 a 40%. Tento podíl není ovšem závazný, bývá užitečné ověřit predikci i pro různé podíly validačních dat.

Samozřejmě ani elegantní PLS regrese není zázračná metoda a má určitá omezení, která spočívají hlavně v předpokladu přibližné linearity a normality chyb. Spolu s možností predikce a grafických výstupů však poskytuje velmi výkonný nástroj pro analýzu a predikci vícerozměrných dat. Především díky predikčním možnostem je PLS regrese ideálním nástrojem pro plánování jakosti, design produktů, optimalizaci technologií a aplikovaný výzkum.

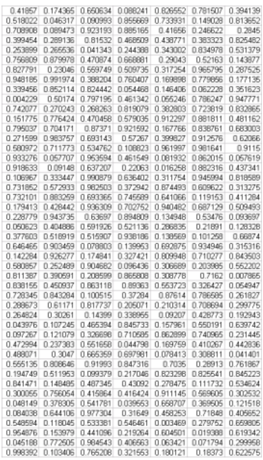
Data a parametry

Do modulu PLS regrese vstupují dvě matice **X** s *p* sloupci a **Y** s *q* sloupci vybrané prostřednictvím sloupců datové tabulky v poli Matice **X** a Matice **Y**. Vybrané sloupce musí obsahovat číselná data, počet řádků matice **X** a **Y** musí být stejný, hodnoty v jednotlivých řádcích si mají odpovídat (tj. vztahovat se k témuž výrobku, objektu, času, experimentu, a podobně). Každý výběr musí obsahovat nejméně dva sloupce. Sloupce matice **X** by se neměly objevit v matici **Y**. Po zaškrtnutí políčka *Dimenze* lze omezit dimenzi *k*. Není-li políčka *Dimenze* zaškrtnuto, provádí se PLS v maximální možné dimenzi, tedy $k = \min(p, q)$. Je možné nejdříve povést PLS v maximální dimenzi, pak určit vhodnou hodnotu *k* pomocí grafu scree-plot a s tímto *k* pak výpočet opakovat. Zaškrtneme-li políčko *Spojit Biplot*, budou jednotlivé body v grafech Biplot spojeny podle pořadí dat ve spreadsheetu. Tím se může projevit případná trajektorie procesu ve vícerozměrném prostoru. Zaškrtneme-li nepovinné políčko *Predikce*, je nutno vybrat v příslušném poli seznam proměných, které odpovídají proměnným matice **X** včetně pořadí. Pro tyto proměnné budou pak pomocí modelu PLS-R vypočítány predikované hodnoty **Y** odpovídající zadaným hodnotám **X**. Hodnoty **X** pro predikci musí mít stejný počet sloupců jako matice **X**, mohou mít ale jiný počet řádků (minimálně 2 řádky).

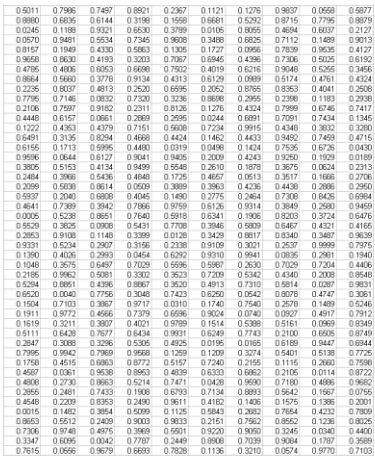
Vypočítaný model PLS lze uložit do souboru pro pozdější použití modulem *Predikce*, zaškrtnutím políčka *Uložit model pro predikci*. Zaškrtnutím políčka *Validace* a zadáním procent použitých pro výpočet modelu se náhodně vybere zadaný podíl dat pro výpočet modelu a zbytek je využit pro validaci modelu.



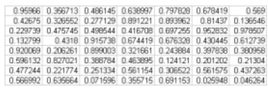
Obrázek 1 Základní nastavení dialogového panelu PLS



X (n x p)



Y (n x q)



X pro predikci (n_1 x p)

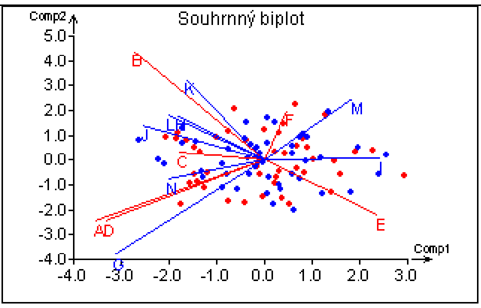
Obrázek 2 Příklad vstupních dat pro PLS

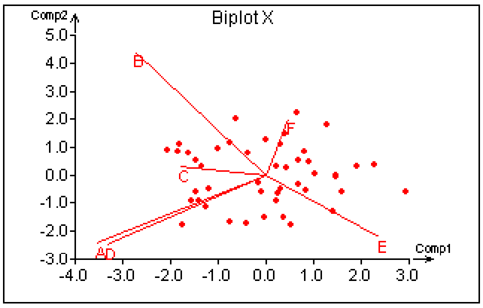
Protokol

Vstupní data	Specifikace vstupních dat
Počet řádků	Počet platných analyzovaných řádků.

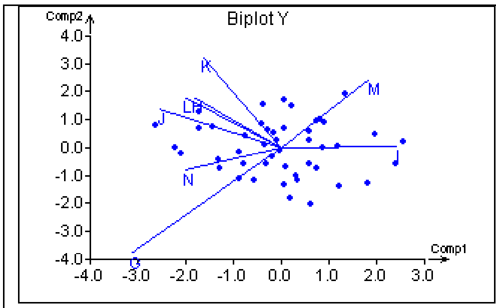
Počet sloupců	Počet sloupců matic **X** a **Y**.
Sloupce	Názvy jednotlivých sloupců obou vstupních matic.
Zvolená dimenze	Dimenze výpočtu PLS podle zadání v dialogovém okně. K určení vhodného *k* je možno využít grafu Scree-plot, viz dále.
Koeficienty PLS, B	Diagonální prvky matice **B**.
Vysvětlený součet čtverců	Tabulka postupného snižování součtu čtverců odchylek predikce od naměřených hodnot s postupným přidáváním komponent (dimenzí).
Počet komponent	Počet komponent použitý v modelu PLS-R, odpovídá postupnému zvyšování *k*.
RSČ	Reziduální součet čtverců. Při nulovém počtu komponent odpovídá celkovému součtu čtverců bez modelu.
Procent	Procentuální vyjádření RSČ.
Vysvět. procent	Procentuální vyjádření (100 – %RSČ), tedy procento kvadratické chyby, které se podařilo vysvětlit modelem PLS-R v dimenzi dané počtem použitých komponent.
Zátěže (Loadings) X, P	Prvky matice zátěží **P**.
Zátěže (Loadings) Y, Q	Prvky matice zátěží **Q**.
Regresní koeficienty, A	Regresní koeficienty přibližně odpovídající vícenásobnému regresnímu modelu **Y** = **XA**. Hodnoty koeficientů nebudou obvykle číselně zcela totožné s koeficienty získanými klasickou lineární regresí. Budou vychýlené, avšak zároveň zkrácené (shrinked), s nižšími rozptyly, a tedy obecně stabilnější.
Predikce	Predikované hodnoty **Y** odpovídající zvoleným hodnotám prediktoru v poli *X-Predikce*. Není-li zaškrtnuto políčko *X-Predikce*, odstavec Predikce se nevytváří.

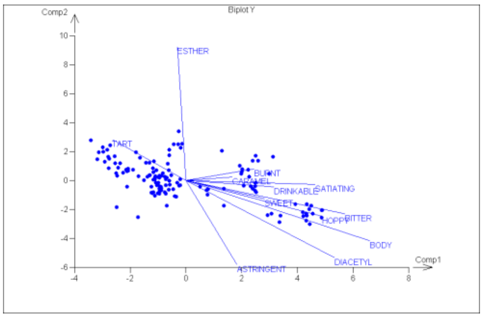
Grafy

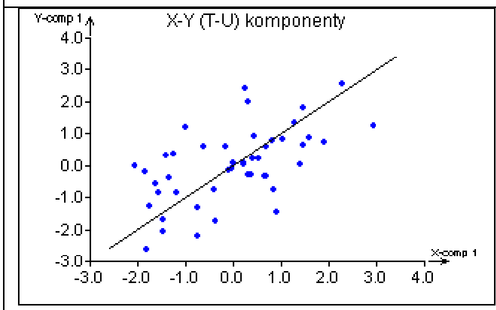




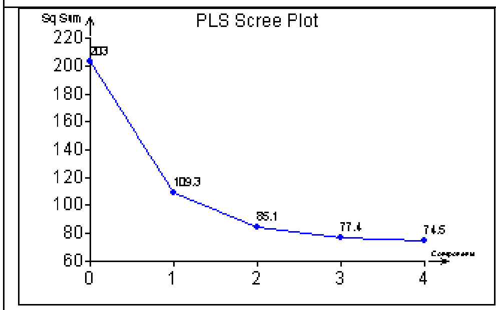
Bi-plot pro matici **X** i **Y** v jednom grafu, pro matici **X** a pro matici **Y** zvlášť. Biplot je projekce vícerozměrných dat do plochy (optimální z hlediska nejmenších čtverců). Body reprezentují řádky, paprsky odpovídají sloupcům.
K identifikaci řádků lze použít označení bodů v interaktivním grafu. Při interpretaci grafu se vychází z toho, že aproximace původních dat úměrná vektorovému součinu jednotlivých bodů a úseček (bod reprezentuje konec vektoru s počátkem v bodě (0,0)). Z toho plyne, že blízké vektory řádků (body) nebo sloupců (paprsky) budou zřejmě vzájemně korelované. Vektory řádků, ležící ve směru některého vektoru sloupce budou mít v tomto sloupci vyšší, resp. nižší hodnoty. Znaménko (smysl vektoru) při tom nehraje roli. Je třeba brát v úvahu, že vzhledem k drastickému snížení počtu rozměrů jsou zvláště pro větší *m* tyto informace orientační a slouží spíše ke globálnímu posouzení struktury a možných souvislostí v datech. Bylo-li vybráno v zadání úlohy *Spojit Biplot*, jsou body v grafu spojeny úsečkami v chronologickém pořadí, což někdy umožní identifikovat nenáhodné časové trendy v procesech.



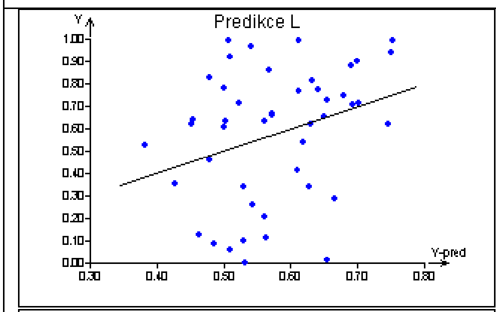


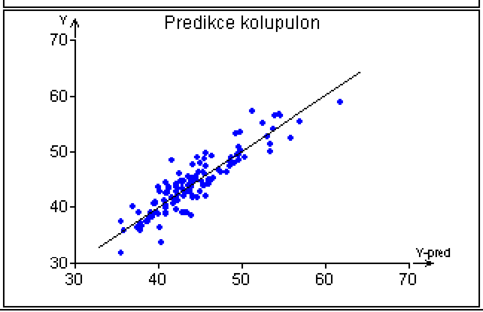


Graf dosažené shody mezi sloupci **T** a **U**. Tento soubor grafů svědčí o globální úspěšnosti a vhodnosti modelu. Čím více se body blíží přímce, tím je model PLS úspěšnější.

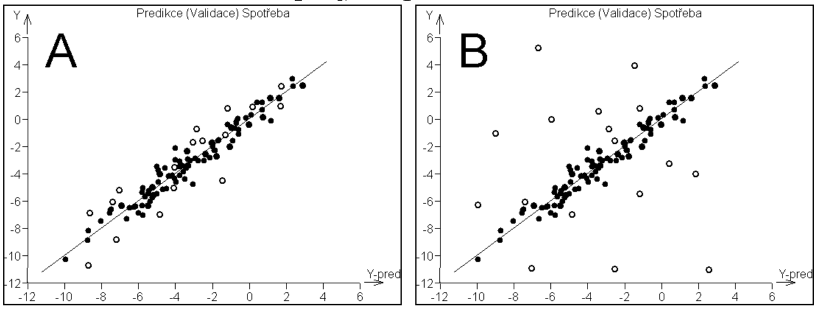
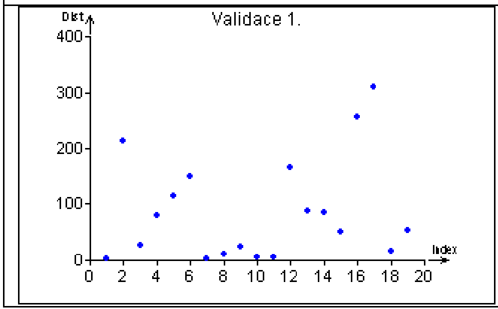


Efektivita modelu vyjádřena pomocí snížení nevysvětleného (reziduálního) součtu čtverců v závislosti na počtu zahrnutých faktorů (sloupců matic **T** a **U**).





Graf Y-Predikce. Graf shody naměřených (zadaných) hodnot závisle proměnné a vypočítaných hodnot predikce. Čím více se body blíží přímce, tím lepší je fit, tedy predikce dobře odpovídá skutečným hodnotám. Tento graf se vytváří pro každou závisle proměnnou. Některé proměnné může PLS predikovat lépe, jiné hůře. Nemá-li graf tvar zřetelného trendu, nebyl nalezen model, který by dobře predikoval tuto nezávisle proměnnou. Byla-li zvolena *Validace*, je v nadpisu grafu v závorce slovo *Validace*, validační body jsou pak v grafu označeny červeně (zde označeno prázdnými kroužky). Jsou-li validovaná data v souhlasu s trénovacími daty (v grafu modře, zde plné kroužky), tak jako na grafu A, je model PLS stabilní a poskytuje spolehlivou predikci. Jsou-li však validační data ve výrazném nesouhlasu s ostatními daty, jako na grafu B, je model PLS možná přeurčený, popisuje dobře pouze data, z nichž byl vypočítán a zřejmě se nehodí pro predikci nových

<table>
<tr>
<td></td>
<td>hodnot. Doporučuje se zkusit snížit dimenzi modelu zadáním čísla menšího než min(p, q) do políčka Dimenze, Obrázek 1.
</td>
</tr>
<tr>
<td></td>
<td>Graf Validace slouží k posouzení kvality predikce validačních dat. Značně odlehlé body mohou představovat vybočující měření. Na ose Y jsou Eukleidovské vzdálenosti naměřených dat od modelu. Porovnání grafů validace pro různé dimenzionality modelu s týmiž daty, nebo stejné modely s různými daty může sloužit jako vodítko pro posouzení kvality modelu a dat. Menší hodnoty vzdálenosti odpovídají lepší predikci pro validační data.</td>
</tr>
</table>