# Aplikace analýzy hlavních komponent pro redukci dimenze transportně-reakčního problému*

Jan Šembera
*Ústav nových technologií a aplikované informatiky*
*Fakulta mechatroniky a mezioborových inženýrských studií Technické univerzita v Liberci*
*e-mail: jan.sembera@tul.cz*

**Abstrakt**

Příspěvek se zabývá redukcí dimenze transportně-reakčního problému v této oblasti nestandardním postupem. Je při něm využíván postup hojně aplikovaný v analýze signálu i při řešení dalších technických problémů – analýza hlavních komponent. Příspěvek nepřináší ucelenou metodiku redukce dimenze obecné úlohy s užitím analýzy hlavních komponent, ale na jednom příkladu ukazuje možnosti tohoto postupu a nastoluje řadu otázek, které je třeba řešit pro každou úlohu zvlášť.

Základním problémem transportních úloh zásadně ovlivňovaných chemickými reakcemi je příliš veliká dimenze úlohy. Nechme nyní stranou, jak velká dimenze je příliš veliká a jaká je již přijatelná (závislost těchto pojmů na konkrétní úloze a jejím uspořádání je skutečně významná). Základním postupem, který chemici při tvorbě chemického modelu užívají, je klasifikace složek roztoku na řídící a marginální a vyčlenění hlavních chemických reakcí mezi řídícími složkami, které ovlivňují řešení problému zásadně, od ostatních chemických reakcí, které mohou být v modelu považovány za doplňující, nebo mohou být úplně vynechány. Tímto postupem je často výrazně redukována dimenze řešené úlohy s tím, že hlavní jevy a koncentrace řídících složek v chemickém systému jsou simulovány a ostatní jevy a koncentrace lze v případě potřeby s větší či menší přesností následně odvozovat z výsledku simulace.

Tento postup je velmi výhodný v případě, že sledujeme vývoj koncentrací řídících složek, nebo změny podmínek ovlivněných těmito řídícími složkami. Pokud nás zajímají některé jevy významně ovlivněné konkrétní marginální složkou, můžeme model rozšířit o tuto složku (zařadit ji mezi řídící složky) a zvýšit dimenzi úlohy. Pokud ale studujeme takovou situaci, kdy pro nás všechny složky systému mají přibližně stejný význam, nelze tímto postupem účinně redukovat dimenzi úlohy jinak než na úkor kvality výsledku.

Taková situace nastala při modelování předpovědi dlouhodobého vývoje kontaminace na lokalitě Stráž pod Ralskem s. p. DIAMO po provedení sanace metodou neutralizace in-situ. V podzemí je směs roztoků s 22 měřenými složkami, z nichž některé přímo řídí hlavní chemické děje, proto je model nesmí opomíjet, ale mezi marginálními složkami jsou také nejnebezpečnější kontaminanty. Ty model také nesmí opominout, protože jejich bilanci a šíření je třeba počítat co nejpřesněji.

Je tedy třeba zachovat počet simulovaných složek a zároveň redukovat dimenzi problému. To lze provést postupem vycházejícím z lineární algebry. Pohlížejme na množinu všech provedených analýz roztoků ve sledované lokalitě jako na množinu $M$ vektorů ve 22-rozměrném lineárním vektorovém prostoru $V$, jehož souřadné osy odpovídají koncentraci jednotlivých složek roztoku.

*Tento výsledek byl realizován s podporou státního rozpočtu České republiky prostřednictvím projektu č. 1M0554 Výzkumné centrum Pokročilé sanační technologie a procesy v programu MŠMT Výzkumná centra (PP2-DP01) a s podporou Grantové agentury České republiky, projekt č. 102/06/P450.

Hledejme takový $n$-rozměrný lineární vektorový podprostor $V_n$ prostoru $V$, který bude „nejblíž“ množině $M$, tj. bude minimalizovat chybu průmětu $E_n^2$ definovanou jako součet druhých mocnin vzdáleností všech vektorů z $M$ od jejich průmětů do $V_n$:

$$E_n^2 = \sum_{\mathbf{x} \in M} \|\mathbf{x} - \Pi_{V_n}\mathbf{x}\|^2. \tag{1}$$

Zde $\Pi_{V_n}$ označuje operátor kolmého promítání do prostoru $V_n$.

Pokud se nám podaří najít podprostor dostatečně malé dimenze $n_0$ s dostatečně malou chybou průmětu $E_{n_0}^2$, můžeme zredukovat dimenzi úlohy transportu z původních 22 na $n_0$ a ze simulace chemických reakcí nevyčleňovat žádnou složku.

Pro řešení tohoto problému můžeme s výhodou využít analýzu hlavních komponent (Principal Component Analysis - PCA), což je metoda redukce dimenze s minimální ztrátou informace v datech standardně užívaná pro řešení řady technických problémů, která se také uplatňuje v ekonomických vědách a lékařství. Je založena na transformaci souřadného systému - nalezení speciální ortonormální báze prostoru, ve kterém jsou data umístěna. Vektory hledané ortonormální báze jsou uspořádány tak, že první určuje směr obsahující největší možnou jednorozměrnou informaci v datech a ve směru posledního bázového vektoru je obsah informace v datech minimální. Tento postup se standardně užívá při zpracování signálu k dekorelaci dat.

Postup jsme použili na řadu chemických analýz roztoků odebraných z různých míst lokality Stráž pod Ralskem v různých časech. Následovalo pozorování vlastností průmětů měření do zvoleného podprostoru z hlediska „přijatelnosti“ pro další zpracování. Pojem „přijatelnost“ nebyl nijak předem specifikován a bylo třeba ho definovat. Vzhledem k tomu, že každý vektor, se kterým pracujeme, musí být možno interpretovat jako potenciální chemickou analýzu roztoku, byla přirozeným požadavkem kladnost, přesněji nezápornost, každé složky průmětu. Dále musíme požadovat, aby průmět byl blízko k původnímu měření nejen v $l^2$ normě, ale také z hlediska každé složky, tj. v nějaké konkrétní vážené maximové normě $\|\mathbf{x} - \Pi_{V_n}\mathbf{x}\|_{\vec{\alpha}} = \max_i \alpha_i |x_i - (\Pi_{V_n} x)_i|$, kde $\vec{\alpha}$ vystihuje významnost každé složky.

Z provedených analýz nevyplývá zatím jasný závěr o tom, který z uvedených postupů je vhodnější. Ukazuje se, že provedení analýzy způsobem, který je naznačen v úplné verzi příspěvku publikované na CD, je třeba udělat pro každou konkrétní úlohu podobného typu znovu. Vhodnost konkrétního zvoleného postupu pak závisí na konkrétní definici „přijatelnosti“ výsledků a dalších prioritách.

Nalezení vhodné redukce původního prostoru není posledním krokem k simulaci transportu a chemických reakcí v mnohasložkovém roztoku s užitím redukované dimenze. Kromě identifikace vhodného podprostoru je třeba zvolit jeho vhodnou bázi, do které budeme rozkládat počáteční podmínky, ve které budeme provádět transportní výpočty a kterou budeme používat pro zpětnou rekonstrukci mnohasložkových dat.

Bázové vektory v mnohasložkovém prostoru lze interpretovat stejně jako všechny vektory reprezentující měření jako bázové roztoky. Tedy bázový vektor odpovídá v jistém smyslu chemické analýze nějakého bázového roztoku. V takovém případě je ale na místě požadavek, aby všechny jeho složky byly nezáporné. Souřadnice jednotlivých měřených roztoků v bázi redukovaného prostoru lze interpretovat jako směšovací poměry bázových roztoků. Pak je ale na místě požadovat, aby tyto souřadnice byly nezáporné (a navíc jejich součet nebyl větší než jedna). Uvedené dva požadavky nejsou pro samotný model nutné, ale pro jeho interpretaci jsou významné a je-li možné najít takovou bázi, pak je přínosné ji mít k dispozici.

V úplné verzi příspěvku publikované na CD jsme tento problém rozřešili s užitím algoritmu pro určení konvexního obalu množiny vektorů a přiložili grafickou interpretaci navrženého postupu a motivační příklad dobře provedené redukce dimenze s dobře interpretovatelnou bází.