Hasselt, 15 grudnia 2017

Tomasz Burzykowski
Profesor Uniwersytetu Hasselt
Centrum Statystyczne

Recenzja rozprawy doktorskiej

Tytuł rozprawy: „Computational and Statistical Methods for Mass Spectrometry Data Analysis”
Autor: Mateusz Krzysztof Łącki

Recenzowana rozprawa doktorska dotyczy metod analizy danych uzyskanych przy pomocy spektrometrii mas (MS). W szczególności, rozważa metody wyznaczania struktury izotopowej (rozdział 2), szacowania prawdopodobieństw i kinetyki reakcji fragmentacji jonów (rozdział 3 i 4) oraz „rozplatania” (deconvolution) nałożonych na siebie w obserwowanym spektrum sygnałów uzyskanych dla różnych cząsteczek (rozdział 5).

W recenzowanej rozprawie Autor posługuje się metodami modelowania matematycznego, statystycznego oraz technikami informatycznymi, wykazując szeroką wiedzę w tym zakresie. Użycie modeli matematycznych i statystycznych pozwoliło Autorowie na opracowanie nowatorskich metod analizy danych MS. Na pochwałę zasługuje fakt implementacji proponowanych metod w postaci publicznie dostępnych narzędzi informatycznych, które umożliwiają wykorzystanie metod w praktyce.

Materiał teroretyczny prezentowany w rozprawie został uwzględniony w pięciu artykułach naukowych, z których trzy zostały już opublikowane, jeden jest zaakceptowany do publikacji, a jeden jest w trakcie oceny przez czasopismo naukowe. Publikacje dotyczą czasopism liczących się w dziedzinie MS.

Poniżej przedstawiam uwagi dotyczące aspektów niektórych z proponowanych przez Autora metod.

W rozdziale drugim autor rozważa zagadnienie wyznaczania struktury izotopowej. Interesującą i nowatorską propozycją jest użycie w tym kontekście wielowymiarowego rozkładu normalnego jako asymptotycznego dla rozkładu wielomianowego. Przyznam, że nie potrafiłem w tekście rozprawy znaleźć odpowiedzi na pytanie w jaki sposób uwzględniana w obliczeniach jest kwestia niedokładnych przybliżeń uzyskanych przy użyciu rozkładu normalnego dla „małych” zliczeń atomów.

Rozdział trzeci rozważa zagadnienie szacowania prawdopodobieństw reakcji fragmentacji jonów. Na stronie 51 w problemie minimizacyjnym postulowane jest użycie funkcji kary, motywowane następująco: „To minimize the risk of numerical instability and perform model selection”. Wartości współczynników funkcji ustalone są na 0.001. Powstaje pytanie dlaczego właśnie takie wartości? Czy ich zmiana wpływa na uzyskiwane wyniki? Zwraca uwagę, że postulowana forma funkcji kary odpowiada metodzie „elastic net” (Zou & Hastie, JRSSB 2005). Metoda ta została zaproponowana w kontekście w budowie modeli predykcyjnych przy dużej liczbie predyktorów (większej niż liczba obserwacji). W tym kontekście ważna jest

zdolność predykcyjna modelu, a nie nieobciążone szacowanie wartości współczynników model. W przypadku przedstawionym w rozprawie celem jest jednak nie predykcja, ale właśnie szacowanie intensywności α. Z tego punktu widzenia, uzyskanie niebciążonych oszacowań byłoby prorytetem. Nie jest też dla mnie jasne, że w praktycznych zastosowaniach mamy do czynienia z liczba predyktorów większą od liczby obserwacji. Czy w takim razie wybór metody jest słuszny?

Na stronie 52 pojawia się definicja funkcji wiarogodności L. Z tekstu nie wynika jednoznacznie, czego ona dokładnie dotyczy? Jej postać zakłada, że każdy z N_i jonów przeszedł dokładnie N^i_{PTR} reakcji PTR i N^i_{ETnoD} reakcji EtnoD. Jest to dla mnie zaskakujące założenie (i nie wynika jednoznacznie z tekstu), a w tekście nie znajduję jego motywacji. Postać estymatorów największej wiarygodności wskazuje na użycie ogrnaiczenia $p_{PTR} + p_{ETnoD} = 1$, które nigdzie w tekście *explicite* sie nie pojawia.

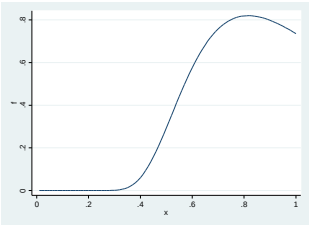
Rozdział trzeci rozważa zagadnienie szacowania prawdopodobieństw reakcji fragmentacji jonów. Na stronie 51 w problemie minimizacyjnym postulowane jest użycie funkcji kary, motywowane następująco: „To minimize the risk of numerical instability and perform model selection". Wartości współczynników funkcji ustalone są na 0.001. Powstaje pytanie dlaczego właśnie takie wartości? Czy ich zmiana wpływa na uzyskiwane wyniki? Zwraca uwagę, że postulowana forma funkcji kary odpowiada metodzie „elastic net" (Zou & Hastie 2005). Metoda ta została zaproponowana w kontekście w budowie modeli predykcyjnych przy dużej liczbie predyktorów (większej niż liczba obserwacji). W tym kontekście ważna jest zdolność predykcyjna modelu, a nie nieobciążone szacowanie wartości współczynników model. W przypadku przedstawionym w rozprawie celem jest jednak nie predykcja, ale właśnie szacowanie intensywności α. Z tego punktu widzenia, uzyskanie niebciążonych oszacowań byłoby prorytetem. Nie jest też dla mnie jasne, że w praktycznych zastosowaniach mamy do czynienia z liczba predyktorów większą od liczby obserwacji. Czy w takim razie wybór metody jest słuszny?

Na stronie 52 pojawia się definicja funkcji wiarogodności L. Z tekstu nie wynika jednoznacznie, czego ona dokładnie dotyczy? Jej postać zakłada, że każdy z N_i jonów przeszedł dokładnie N^i_{PTR} reakcji PTR i N^i_{ETnoD} reakcji EtnoD. Jest to dla mnie zaskakujące założenie (i nie wynika jednoznacznie z tekstu), a w tekście nie znajduję jego motywacji. Postać estymatorów największej wiarygodności wskazuje na użycie ogrnaiczenia $p_{PTR} + p_{ETnoD} = 1$, które nigdzie w tekście *explicite* sie nie pojawia.

W *rozdziale czwartym* kontynuowane są rozważania dotyczące reakcji fragmentacji jonów. W szczególności konstruowany jest model wykorzystujący równania opisujące kinetykę reakcji. W celu dopasowania modelu do danych (str. 81) postulowane jest użycie metody najmniejszych kwadratów (z opcjonalną funkcją kary) funkcji kary. Ponieważ obserwowane i oczekiwane wartości odnoszą się do frakcji, powstaje pytanie czy nie należałoby uwzględnić sum kwadratów ważonych przez wartości oczekiwane frakcji? Różnica rzędu 0.01 dla frakcji wynoszącej np. 0.5 jest mniej „znacząca" niż dla frakcji wynoszącej np. 0.05.

Rozdział piąty prezentuje podstawy teoretyczne Bayesowskiej metody „rozplatania" (deconvolution) nałożonych na siebie w obserwowanym spektrum sygnałów uzyskanych dla różnych cząsteczek. Jak rozumiem, metoda będzie dopiero obiektem badań, na co wskazuje informacja na stronie 104 („More test will be carried soon"). W badaniach tych warto poświęcić uwagę np. wyborom rozkładów a priori. W rozprawie preferowane są rozkłady sprzężone (conjugate): np. „It is natural to choose (…) the gamma distribution, as it is the

conjugate distribution to the Poisson distribution" (str. 93), "Gaussian choice is motivated by the ease of drawing from the conjugate distribution" (str. 96). W obecnej chwili, wobec dostępności szerokiego zakresu (zaimplementowanych, np. WinBUGS) metod MCMC, ograniczanie się do tych rozkładów nie jest konieczne. Wybór powinien raczej dokonywany na podstawie analizy pożądanych cech rozkładów. Z tego punktu widzenia niektóre decyzje Autora zasługiwałyby na pełniejsze wyjaśnienie i/lub modyfikację. Dla przykładu, na str. 96 jako rozkład a priori dla ν^{-2} użyty jest rozkład gamma(1,1). Warto zwrócić uwagę, że taki wybór implikuje dla ν tzw. rozkład root-inverse-gamma, który – dla gamma(1,1) – oznacza eliminację, a priori, wartości ν poniżej 0.3 (zob. wykres gęstości przedstawiony poniżej). Czy wykluczenie tego zakresu wartości jest rzeczywiście uzasadnione? Jeśli tak, to na podstawie jakich argumentów?



Kluczowe dla metody równanie (5.4), opisujące rozkład a posteriori, zawiera funkcje gęstości dla rozkładu gamma(a_m,b_m). Jednakże na stronie 93 możemy przeczytać, że: „The gamma distribution has density $\gamma_{a,b}(x)$ proportional to $x^{\alpha-1}e^{-\beta x}$. It depends upon two hyper-parameters that need to be chosen in advance. It is natural to assume α=1, as this results in an *a priori* distribution with a mode at o, as we should expect *a priori* that a molecular species is absent." Pomijając problemy notacyjne (α użyte zamiast *a*, o zamiast 0) przedstawionym fragmencie, sugeruje on, że pierwszy parametr rozkładu gamma powinien być przyjęty jako 1. W równaniu (5.4) natomiast parametr ten jest (najwyraźniej) traktowany jako nieznany.

W równaniu dla I_w (str. 95) postulowany jest model z addytywnym błędem resztowym o stałej wariancji. Założenie homoskedastyczności nie jest oczywiste – w literaturze nt. MS istnieja doniesienia o możliwej heteroskedastyczności błędów losowych, w której wariancja tych ostatnich zależy od funkcji potęgowej wartości średniej (Zhu & Burzykowski , JASMS 2011).

Pozytywne wrażenie, jakie odnosiłem w trakcie zapoznawania się z merytoryczną zawartością rozprawy, było, niestety, zakłócane problemami wynikającymi z redakcji rozprawy. Tekst zawiera błędy językowe, notacyjne i redakcyjne. Jest ich na tyle dużo, że przeszkadzają w lekturze. Nie sposób równiez z tego powodu ich wszystkich wymienić. Poniżej podaję jedynie niektóre, wybrane przykłady:

- błędy gramatyczne: str. 10, „Experimental findings does not..." zamiast "do not"; str. 12: "that relative few configurations" zamiast "relatively few"; str. 13: "peak heights truly results" zamiast "result"; str. 14: "each carrying a radicals – an electron" zamiast "a radical"; str. 22: "Fig. 2.2 suggest" zamiast "suggests";
- błędy językowe i stylistyczne: str. 11, „By far, it is also by far"; str. 12: "It results, that the number…" – raczej "It follows"; "as while performing" zamiast "when performing"; str. 14: "inability to tell apart" zamiast "to distinguish between"; str. 21: "threshold can be precised" zamiast "can be specified"; "str. 27: „truly appauling" zamiast „appalling"; str. 45: "to make that idea more clear" zamiast "to clarify the idea"; str. 53: "they undergone" zamiast "underwent";

- mylące odniesienia do rycin, tabel, równań: str. 21, „As demonstrated in Fig. 2.2" – powinno być Fig. 2.1; str. 58: odniesienia do Fig 3.9 i 3.10 pojawiają sie przed odniesieniem do Fig 3.8; str. 66: „Figure 4.1 presents the considered set" – powinno być Table 4.1; str. 68: „see Figure 4.1" – powinno być Table 4.1; str. 31: "It results from Eq. (2)" – powinno być Eq. (2.2);

- niekonsekwentna notacja: np. macierze czasem oznaczane są wielkimi literami greckiego alfabetu (str. 27: Σ_e, Δ_e), czasem małymi literami alfabetu łacińskiego (str. 27: d(p)), czasem pogrubionymi wielkimi literami alfabetu łacińskiego (str. 96: **N**);

- brak oznaczeń osi wykresów (Fig. 2.4 – oznaczenia pojawiają sie dopiero na Fig. 2.7) lub mylący opis rycin (Fig. 2.7: „Then, $p > 99.99\%$, but removing more isotopologues would bring joint probability below 99.99%.", ale celem było chyba 99.9%?);

- Table 4.1: znaki plus w nawiasach kwadratowych po prawej stronie strzałek zostały przeniesione do superskryptu i w efekcie są mylące, bo sugerują dodatni ładunek;

- po przeczytaniu – merytorycznego – rozdziału 1 byłem zaskoczony pojawieniem sie listy publikacji i personalnymi podziękowaniami Autora rozprawy. Uważam, że te fragmenty tekstu powinny zostać umieszczone na początku rozprawy, przed rozdziałem 1.

Pomimo usterek redakcyjnych, zawartość merytoryczną ocenianej rozprawy uważam za interesującą i w wielu aspektach nowatorską. Pokazuje ona jasno, że Autor dogłębnie zapoznał się z problematyką analizy danych MS. Potwierdza również, że Autor potrafi zaproponować nowe, interesujące metody analizy oraz skutecznie je zaimplementować w postaci narzędzi informatycznych, które pozwalają na użycie metod w praktyce. Wobec powyższego stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie magistra Mateusza Krzysztofa Łąckiego do dalszych etapów przewodu doktorskiego.