نویسندگان
فریده باتقوا[۱و۳*]
هادی اصفهانی[۲و۳]
*fbataghva@yahoo.com

معرفی کمومتریکس

(بخش اول)

چکیده

با ورود به دنیای فن‌آوری‌های جدید و پیشرفت‌های شگرف در علوم مختلف، علم شیمی هم دچار تحولات زیادی به‌خصوص در زمینه شیمی تجزیه شده‌است. ورود دستگاه‌هایی با توانمندی‌های بالا به‌منظور تولید داده‌ها، مشکل به‌دست آوردن اطلاعات که روزگاری نه چندان دور، معضلی برای شیمیدانان تلقی می‌شد را آسان کرده است. اما سوالی که مطرح می‌شود این است که چگونه می‌توان بیشترین اطلاعات لازم را از این طیف وسیع داده‌ها استخراج کرد؟ کمومتریکس به‌عنوان یک علم و ابزار ریاضی در شیمی، می‌تواند تا حدودی پاسخگوی این سوال باشد. این علم جدید با بهره‌گیری از روش‌های ریاضی و آمار کمک شایانی به حل این مسئله کرده است.

واژه‌های کلیدی

کمومتریکس، روش‌های آماری، آنالیز چند متغیره، PCA.

مقدمه

واژه کمومتریکس در اوائل دهه ۱۹۷۰ توسط اسوانت ولد[۴] و کوالسکی[۵] مطرح شد [۱]. این واژه مشابه واژه‌های بیومتریکس[۶] و اکانومتریکس[۷] است که قبلاً در علوم زیست‌شناسی و اقتصاد مطرح شده بودند. با تاسیس انجمن بین‌المللی کمومتریکس[۸] در سال ۱۹۷۱، رشد و توسعه کمومتریکس سرعتی بیش از پیش به خود گرفت و امروزه به‌طور گسترده‌ای در زمینه‌های مختلف شیمی به‌خصوص شیمی تجزیه کاربرد دارد. در سال ۱۹۷۴ اسوانت ولد با طرح سوال زیر تعریفی کلی برای کمومتریکس ارائه کرد: چگونه می‌توان اطلاعات شیمیایی را از داده‌های اندازه‌گیری شده به‌دست آورد و آن‌ها را ارائه و نمایش داد؟ او نتیجه گرفت که کمومتریکس مجموعه‌ای از روش‌های ریاضی، اصول آماری و دیگر روش‌های مبتنی بر منطق است که به استخراج اطلاعات شیمیایی کمک می‌کند. بنابراین، کمومتریکس یک توسعه طبیعی از شیمی تجزیه محسوب می‌شود. در واقع کمومتریکس یک زمینه بین رشته‌ای است که شامل آمارهای چند متغیره[۹] مدل کردن‌های ریاضی، علم کامپیوتر و شیمی تجزیه است. در اوائل دهه ۱۹۸۰ هاوری[۱۰] و هیرش[۱۱] [۲] توسعه کمومتریکس را به دو مرحله تقسیم‌بندی کردند:

- مرحله اول که به قبل از سال ۱۹۷۰ بر می‌گردد، شامل توسعه یک سری روش‌های ریاضی در زمینه‌های مختلف ریاضی، علوم رفتاری و علوم مهندسی بود. در این دوره، شیمیدانان اساساً خود را به آنالیز داده‌ها، شامل محاسبه عوامل آماری مانند میانگین، انحراف استاندارد و غیره محدود کرده بودند. در این سال‌ها بود که تحقیقات هاوری و هیرش به‌منظور برقراری ارتباط بین داده‌های شیمیایی با خواص مولکولی مناسب باعث تشکیل اساس زمینه‌ای مهم در کمومتریکس به نام ارتباط کمی ساختار فعالیت[۱۲] شد.

- مرحله دوم رشد کمومتریکس به دهه ۱۹۷۰ بر می‌گردد که طی آن این زمینه جدید توجه شیمیدانان و به‌خصوص شیمیدانان تجزیه را به خود جلب کرد و طی آن، شیمیدانان علاوه‌بر روش‌های موجود مورد استفاده، روش‌های جدیدی را به‌منظور رفع نیازهایشان توسعه دادند. تکامل کمومتریکس توسط هاوری و هیرش و همچنین بعدها توسط براون[۱۳] [۲و۳] انجام شده‌است. در این دوره، توجه پژوهشگران به‌طور عمومی روی روش‌های چند متغیره متمرکز بود. از آن جا که جهان اطراف ما ذاتاً چند متغیره است، بنابراین عملیات آنالیز داده‌ها در اندازه‌گیری‌های چند گانه قابل درک می‌شود. برای مثال، وقتی جذب UV یک محلول اندازه‌گیری می‌شود، اندازه‌گیری طیف کامل آن نسبت به حالتی که جذب آن در یک طول موج خاص اندازه‌گیری می‌شود با نوفه کمتر و به‌طور سریع امکان‌پذیر است. با بررسی و در نظر گرفتن مناسب توزیع متغیرهای چندگانه به‌طور هم‌زمان، می‌توان اطلاعات بیشتری نسبت به حالتی که هر متغیر به‌طور مجزا بررسی می‌شود، به‌دست آورد. این مزیتی است که مزیت چند متغیره[۱۴] نامیده می‌شود.

به‌طور کلی زمینه‌های موجود در کمومتریکس شامل موارد زیر هستند:
۱. روش‌های تفکیک چند متغیره[۱۵]؛
۲. روش‌های کالیبراسیون چند متغیره[۱۶]؛
۳. روش‌های تشخیص الگو[۱۷]؛
۴. روش‌های ارتباط کمی ساختار فعالیت؛
۵. روش‌های آنالیز تصویری[۱۸]؛
۶. روش‌های طراحی آزمایش[۱۹]؛
۷. روش‌های کالیبراسیون مرتبه دوم[۲۰].

از میان این زمینه‌ها، با توجه به نیاز همیشگی در شیمی تجزیه به‌منظور به‌دست آوردن اطلاعات کیفی و کمی گونه‌های هدف در بافت‌های پیچیده، نقش روش‌های تفکیک چند متغیره و کالیبراسیون مرتبه دوم از همان ابتدای تولد کمومتریکس مورد توجه شیمیدانان تجزیه بوده است. افزایش خیره کننده تعداد مقالات چاپ شده طی سال‌های اخیر، تایید کننده این مدعاست. لذا تحقیق و توسعه در این راستا می‌تواند کمک شایانی به شیمیدانان کند.

آنالیز فاکتور

آنالیز فاکتور، اولین بار توسط دانشمندان علوم رفتاری[۲۱] به دنیای علم معرفی شد و هارمن[۲۲] به‌عنوان اولین کسی که در این زمینه مطالعاتی را انجام داد، شناخته شده‌است. اولین مقاله در زمینه آنالیز فاکتور در سال ۱۹۰۱ توسط پیرسون[۲۳] به چاپ رسیده است. اولین توسعه واقعی آنالیز فاکتور در شیمی توسط هوتلینگ[۲۴] در سال ۱۹۳۳ صورت گرفته است. اما تا دهه ۱۹۷۰ که به‌عنوان دهه تولد کمومتریکس شناخته می‌شود، توجه چندانی به آن، به‌منظور حل مسائل در شیمی نشد. با ورود رایانه، پیشرفت و توسعه این زمینه به شدت تحت تاثیر قرار گرفت و با توجه به این که امروزه بیشتر شیمیدانان با رایانه، ریاضیات و آمار آشنا هستند، شاهد پیشرفت بسیار سریعی در این زمینه هستیم. مالینوسکی[۲۵] تعریفی جامع از آنالیز فاکتور، بیان کرده است [۴]. آنالیز فاکتور روشی چند متغیره است که به‌منظور کاهش دادن حجم فضای ماتریس داده‌ها به ماتریس‌های با ابعاد کمتر با استفاده از فضای تعریف شده با فاکتورهای ریاضی عمود بر هم[۲۶] به کار می‌رود و در نتیجه تبدیل‌هایی ارائه خواهد کرد که منجر به تولید فاکتورهای شیمیایی قابل فهم خواهد شد. آنالیز فاکتور بیشتر به ما این اجازه را می‌دهد که اساسی‌ترین سوالات در یک مسئله شیمیایی را پاسخ دهیم. سوالاتی از این قبیل که: چه تعداد فاکتور پدیده قابل مشاهده ما را تحت تاثیر قرار می‌دهد؟ ماهیت فیزیکی عوامل چه هستند؟

آنالیز فاکتور شامل مراحل زیر است: در مرحله آمادی‌سازی[۲۷]، داده‌هایی که می‌خواهند آنالیز شوند، انتخاب می‌شوند. این انتخاب براساس یک سری از معیارها انجام و سپس پیش پردازش ریاضی، روی داده‌ها انجام می‌شود. در مرحله بازسازی مجدد[۲۸]، حل ریاضی براساس فضای فاکتورهای محدود شده روی داده‌ها انجام می‌شود و فاکتورهای اولیه ریاضی که مفهوم شیمیایی ندارند، به‌دست می‌آیند. دستورالعمل برای محاسبه فاکتورهای اولیه ریاضی را آنالیز ویژه[۲۹] می‌نامند که منجر به تولید مقادیر و بردارهای ویژه می‌شود. دو مرحله آماده‌سازی و بازسازی مجدد در تمامی آنالیزهای فاکتور مشترک هستند. به‌منظور مفهوم دادن به فاکتورهای اولیه ریاضی، باید یک تبدیل[۳۰] ریاضی انجام شود.

سه نوع از تبدیل‌های ریاضی برای این منظور استفاده می‌شود: آزمون‌های هدف[۳۱]، روش‌های ویژه[۳۲] و چرخش اولیه[۳۳]. روش آزمون هدف، روشی خاص به‌منظور تعیین فاکتورهای واقعی مجزاست. روش‌های ویژه از مزیت اطلاعات شیمیایی شناخته شده با قرار دادن محدودیت‌های[۳۴] مهم روی تبدیل ریاضی استفاده می‌کند. این روش‌ها زمانی استفاده می‌شوند که اهداف به‌طور دقیق یا مناسب نمی‌توانند فرموله‌بندی شوند. چرخش، فاکتورهای خام اولیه را به فاکتورهایی که آسان‌تر قابل تفسیر هستند، تبدیل می‌کند. بعد از این مرحله، به مرحله ترکیب کردن[۳۵] می‌رسیم که طی آن مدلی کامل، متشکل از فاکتورهای واقعی می‌تواند برای پیش‌بینی داده‌های جدید در مرحله پیش‌بینی استفاده شود. بنابراین، براساس آن چه گفته شد، قلب روش‌های آنالیز فاکتوری، تجزیه ماتریس اولیه است و به‌طور عمده برای انجام چنین کاری، روش آنالیز جزء اصلی[۳۶] مورد استفاده قرار می‌گیرد [۵] که در ادامه توضیح داده خواهد شد.

آنالیز جزء اصلی (PCA)

آنالیز جزء اصلی، تبدیلی در فضای برداری است که بیشتر برای کاهش ابعاد مجموعه داده‌ها مورد استفاده قرار می‌گیرند. در PCA، آنالیز داده‌ها برای ساخت مدل‌های خطی چند متغیره برای مجموعه داده‌های پیچیده انجام می‌شود [۶].

مدل‌های دوتایی خطی[۳۷] چند متغیره PCA با استفاده از بردارهای اصلی متعامد (بردارهای ویژه[۳۸]) به‌وجود آمده‌اند، که عموما اجزاء اصلی نامیده می‌شوند. اجزا اصلی به همان خوبی که خطای تصادفی مورد اندازه‌گیری را مدل می‌کنند، تغییرات عمده در مجموعه داده‌ها را به‌طور آماری مدل می‌کنند. یکی از اهداف PCA را می‌توان حذف اجزاء اصلی که با نوفه در ارتباط است، نام برد؛ در نتیجه مشکلات پیچیده مربوط به ابعاد بالای داده‌ها کاهش می‌یابد و اثرات خطای مورد اندازه‌گیری را به کوچکترین مقدار خود می‌رساند.

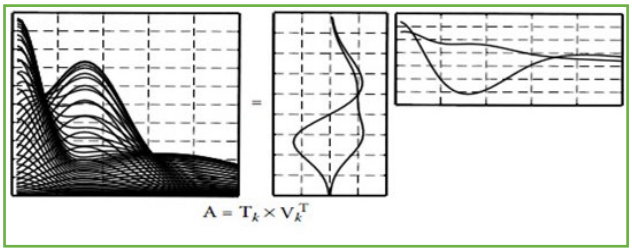
◈ مدل جزء اصلی

خوشبختانه، با استفاده از PCA، این امکان وجود دارد که مجموعه‌های مجزای بردارهای اصلی که فضای عمده ماتریس داده‌ای مثل A را توصیف می‌کنند، بدون هیچ دانش قبلی محاسبه شوند. با استفاده از PCA، این امکان وجود دارد که مدلی ریاضی - تجربی از داده‌ها که با معادله (۱) نشان داده شده را ساخت:

$$A = T_k V_k^T + E \qquad (1)$$

که در آن: $T_k(n\times k)$ ماتریس اسکورهای[۳۹] جزء اصلی و $V_k(n\times k)$ ماتریس لودینگ‌های جزء اصلی نامیده می‌شوند. بردارهای ویژه نرمالیزه موجود در V_k می‌توانند برای تشکیل مجموعه‌ای از بردارهای اصلی ردیفی متعامد برای A مورد استفاده قرار گیرند. این بردارهای ویژه لودینگ‌ها[۴۰]، فاکتورهای چکیده[۴۱] و یا طیف‌های ویژه[۴۲] نامیده می‌شوند که نشان دهنده این هستند که وقتی بردارها، مجموعه اصلی برای فضای ردیف‌های A تشکیل دادند، تفسیر فیزیکی بردارها همیشه ممکن نیست. در شکل (۱) تفکیک جزء اصلی برای مجموعه داده‌های کروماتوگرافی- طیف‌سنجی شامل دو جزء

شیمیایی نشان داده می‌شود.



شکل (۱): نمایی از مدل جزء اصلی برای مجموعه داده‌های کروماتوگرافی – طیف‌سنجی

ستون‌های Tk، اسکور نامیده می‌شوند که متعامد بوده ولی نرمالیزه نیستند. آنها می‌توانند برای تشکیل مجموعه بردارهای اصلی ستون‌های A به کار روند.

برای هر منبع مجزای تغییرات در داده‌ها، یک جزء اصلی مجزا (بردار ویژه) در مدل مورد انتظار است. اولین ستون اسکورها و اولین بردار ویژه مربوط به اولین فاکتور است. اولین بردار ویژه، مربوط به جزیی است که دارای بالاترین مقدار ویژه می‌شود. این قابل مشاهده است که اولین فاکتور بیشترین مقدار ممکن از واریانس را در داده اصلی توصیف می‌کند. فاکتور دوم، فاکتور مهم بعدی است که مربوط به دومین ستون از اسکورها و دومین مقدار ویژه است. این فاکتور بیشترین مقدار واریانس باقیمانده ماتریس در داده اصلی است. به‌طور خلاصه، ماتریس داده‌ها می‌تواند به‌عنوان مجموعه‌ای از K ماتریس داده با ابعاد n×m با مرتبه یک تعریف شود (شکل (۲)).

$$A = t_1 v_1^T + t_2 v_2^T + \ldots\ldots + t_K v_K^T + E \qquad (۲)$$

حاصل ضرب خارجی[۴۳] بردارهای $t_1 v_1^T$ تغییرات توصیف شده با استفاده از اولین فاکتور است [۷].

$$A = t_1^{V_1^t} + t_2^{V_2^t} + \cdots + E$$

شکل (۲): نمایی از تفکیک جزء اصلی برای K جزء با نمایش آنها به‌صورت اجزاء دوتایی خطی در دو بعد.

◈ مرتبه[۴۴] و شبه مرتبه[۴۵] ماتریس

برای تعیین تعداد عوامل اصلی در یک ماتریس، نیازمند به دانستن مفهوم «مرتبه» هستیم. مرتبه یک ماتریس مفهومی ریاضی است که مربوط به تعداد اجزاء اصلی موجود در مجموعه داده‌ها[۴۶] می‌شود که در مفهوم شیمیایی مربوط به تعداد ترکیبات موجود در یک مخلوط است. به‌طور مثال، اگر شش ترکیب در یک کروماتوگرام وجود داشته باشد، مرتبه ماتریس داده‌های کروماتوگرام به‌طور ایده‌آل باید شش باشد. به هر حال همیشه به این سادگی هم نیست. آنچه که رخ می‌دهد این است که نوفه این حالت ایده‌آل را از بین می‌برد؛ بنابراین، حتی اگر شش ترکیب وجود داشته باشد، ممکن است مرتبه سیستم ۱۰ یا بیشتر باشد و یا حتی ممکن است مرتبه سیستم کاهش یابد. این در صورتی است که الگوها[۴۷] برای ترکیبات مشخصی از نوفه قابل تشخیص نباشد [۸]. مفهوم شیمیایی مرتبه به شبه مرتبه معروف است. اصطلاح شبه مرتبه، تخمین مرتبه داده‌های واقعی را توصیف می‌کند [۹] که چندین دستورالعمل برای تعیین آن پیشنهاد شده‌است. به‌عنوان مثال، در معادله (۳) اگرشبه مرتبه ماتریس A برابر با ۳ باشد، تعداد اصلی ۳ عامل یا جزء اصلی برای بازسازی ماتریس A لازم است به این صورت:

$$A_{m\times n} = T_{m\times f}\ V^T_{f\times n} + E_{m\times n} \qquad (۳)$$

به‌عبارت دیگر، با توجه به کاهش اهمیت اجزاء اصلی فراتر از f، ماتریس‌های تفکیک شده از (۱+f) به بعد با یکدیگر جمع شده و در ماتریس باقی مانده‌های (E) نمایان می‌شود.

روش‌های مدل‌سازی

از آنجا که بیشتر روش‌های آماری که در کمومتریکس به کار می‌رود براساس روش‌های مدل‌سازی است به همین دلیل به روش‌های مدل‌سازی و انواع مدل‌ها پرداخته می‌شود.

در گذشته مدل‌سازی متداول براساس اصول اولیه[۴۸] بوده است که به آن مدل‌های سخت[۴۹] هم گفته می‌شود. یک مدل سخت، مدلی است که سیستم را براساس روابط ریاضی حاکم بر اندازه‌گیری متغیرها و متغیرهای وابسته توضیح می‌دهد و خروجی آن نیز روابط بین متغیرها است. از آنجا که سیستم‌های شیمیایی بسیار پیچیده‌تر بود، روش‌ها و قواعد خاصی را می‌طلبید که حل سیستم معادلات را بسیار سخت و در بسیاری موارد غیر ممکن می‌ساخت.

به همین دلیل، روش‌های مدل‌سازی نرم[۵۰] در شیمی ابداع شد. در این روش‌ها بر خلاف روش‌های مدل‌سازی سخت از فرضیات نرم‌تری در روابط حاکم بین داده‌ها و توصیف واریانس داده‌ها استفاده می‌شود. بیشتر این مدل‌های نرم براساس روش‌های ریاضی صورت گرفته و شناسایی تعداد منابع ایجاد کننده واریانس در داده‌ها، تخمین‌های کیفی و در نهایت کمی را ممکن می‌سازد. نتایج روش‌های مدل‌سازی نرم برای ارزیابی روش‌های مدل‌سازی سخت، به‌خصوص برای حل سیستم‌های پیچیده بسیار مفید است.

نتیجه‌گیری

کمومتریکس مجموعه‌ای از روش‌های ریاضی، اصول آماری و دیگر روش‌های مبتنی بر منطق است که به استخراج اطلاعات شیمیایی کمک می‌کند. بنابراین، کمومتریکس یک توسعه طبیعی از شیمی تجزیه محسوب می‌شود. به‌منظور به‌دست آوردن اطلاعات کیفی و کمی گونه‌های هدف در بافت‌های پیچیده، نقش روش‌های تفکیک چند متغیره و کالیبراسیون مرتبه دوم از کمومتریکس مورد توجه است که بر این اساس، مجموعه داده‌های پیچیده در فضای ابعادی دیگر با استفاده از مدل ریاضی تبیین می‌شوند.

پی‌نوشت

۱. کارشناسی ارشد شیمی تجزیه، پژوهشگاه شیمی و مهندسی شیمی
۲. دانشجوی دکتری شیمی تجزیه، پژوهشکده کیمیازی
۳. عضو کارگروه تخصصی کروماتوگرافی

4. Svante Wold
5. Bruce R.Kowalski
6. Biometrics
7. Econometrics
8. International chemometrics society
9. Multivariate statistic
10. Howery
11. Hirsch
12. Quantitative structure-activity relationship
13. Brown
14. Multivariate Advantage
15. Multivariate resolution
16. Multivariate calibration
17. Pattern recognition
18. Image analysis
19. Design of experiment
20. Second-order calibration
21. Behavioral Scientists
22. Harman
23. Pearson
24. Hotelling
25. Malinowski
26. Orthogonal
27. Preparation
28. Reproduction
29. Eigen analysis
30. Transformation
31. Target testing
32. Special method
33. Abstract rotation
34. Constraint
35. Combination
36. Principal component analysis (PCA)
37. Bilinear
38. Eigenvectors
39. Scores
40. Loadings
41. Abstract Factors
42. Eigen spectra
43. Outer Product
44. Rank
45. Pseudo-Rank
46. Data Sets
47. Profiles
48. First- Principles
49. Hard Models
50. Soft Models

مراجع

[1] http://www.emsl.pnl.gov:2080/docs/incinc/homepage .html, in, 2011.

[2] Gemperline, P. Practical Guide to Chemometrics, 2nd ed, Taylor & Francis, New York, 2006.

[3] Hassan, S.; Hussain, S.; Ansari, M.T. J. Zool. 2011, 43, 909. [[4] Malinowski, E.R. Factor Analysis in chemistry, 3rded, John Wiley & Sons, New York, 1991.

[5] Wold, S.; Esbensen, K.; Geladi, P. chemom. Intell. Lab. Syst. 1987, 2, 37.

[6] Meyers, R.A, Encyclopedia of .Analytical Chemistry Application Theory and Instrumentation.

[7] Danzer, K. J. Chemom. 1998, 2, 247.

[8] Ho, C.H. Anal Chem. 1978, 50, 1108.

[9] Comas, C.; Gimeno, R. J. Vhrom A. 2004, 1035, 192.