فهرست

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه‌ای بر کاربرد کمومتریکس در مطالعات روابط کمی بین ساختار و فعالیّت / خاصیّت

۱. ۱ مقدمه

۱. ۱. ۱ شیمی

شیمی[۱] شاخه‌ای از علوم فیزیکی[۲] است که وظیفه‌ی مطالعه‌ی ترکیب، خواص و رفتار ماده را به عهده دارد. گاهاً شیمی را به این علت که رابطه‌ای برای فیزیک و دیگر علوم طبیعی مانند زمین‌شناسی و زیست-شناسی است به نام "علم مرکزی[۳]" نیز می‌خوانند. اگرچه شیمی شاخه‌ای از علوم فیزیکی است، اما از علم فیزیک جداست [۱].

اگر شخصی در پی ریشه‌یابی کلمه‌ی شیمی باشد پس از سفرهای فراوان در زمان‌های قدیم و چالش کردن با تمدن‌هایی همچون اعراب، یونانیان، مصریان، چینی‌ها و یهودیان باستان و همچنین برخورد با کلماتی از قبیل آلکمی[۴]، ألکیمیا[۵]و کیمی[۶] باز هم به درستی به منشاء این کلمه نخواهد رسید. با تمام این قضایا در قرن ۱۸ میلادی واژه‌ی شیمی برای علم امروزی انتخاب شد.

[1] Chemistry
[2] Physical Sciences
[3] Central Science
[4] alchemy
[5] Al-kimiya
[6] kimi

در قرن ۲۰ میلادی و با گسترش علوم و نفوذ و همپوشانی آنها در برخی از موارد باعث شد که شاخه‌های جدیدی از علم و به نام‌های ترکیبی به علوم افزوده شود. در شیمی این شاخه‌ها با ترکیب‌شدن سایر واژه‌ها با دگر شکل‌هایی از واژه‌ی شیمی مانند کمو[۱]، کمی[۲] و کم[۳] ، به شکل کموسفر[۴]، کموتراپی[۵]، کمیسورپشن[۶] و کمینفورماتیکس[۷]، بوجود آمدند.

در این زمان بود که چندین رشته‌ی علمی با نام‌های ترکیبی که انتهای آنها به متری (مانند ژئومتری[۸]) و متریکس (مانند بایومتریکس[۹]) که از یک واژه‌ی یونانی به معنای اندازه‌گیری گرفته‌شده بود. از واژه‌های مشابه که برای نام‌گذاری علوم جدید استفاده شده، می‌توان به سایکومتری[۱۰] و سایکومتریکس[۱۱] در رشته‌ی روانشناسی و اکونومتری[۱۲] و اکونومتریکس[۱۳] در اقتصاد اشاره کرد.

۱. ۱. ۲ کمومتریکس

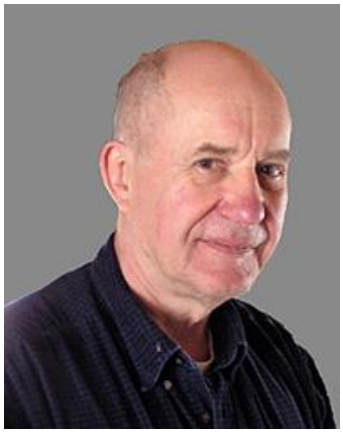
در سال ۱۹۷۱ میلادی بود که برای اولین بار واژه ی کمومتری[۱۴] در زبان سوئدی و کمومتریکس[۱۵] در زبان انگلیسی به عنوان ترکیبی از شیمی و اندازه‌گیری توسط یک شیمیدان آلی به نام پروفسور سوانته ولد[۱۶] ارائه شد.

پروفسور سوانته ولد کمومتریکس را در سال ۱۹۷٤ به عنوان هنری برای بیرون کشیدن اطلاعات موردنظر شیمیایی از داده‌هایی که در آزمایش‌های شیمیایی تهیه شده‌اند، تعریف کرد. ایشان در بیست سال پس از این تعریف، در مقاله‌ای [۲] تعریف کمومتریکس را به این شکل اصلاح کردند:

---

[1] Chemo
[2] Chemi
[3] Chem
[4] Chemosphere
[5] Chemotherapy
[6] Chemisorption
[7] Cheminformatics
[8] Geometry
[9] Biometrics
[10] Psychometry
[11] Psychometrics
[12] Econometry
[13] Econometrics
[14] Kemometri
[15] Chemometrics
[16] Svante Wold

چگونه اطلاعات موردنظر را از داده‌های شیمیایی اندازه‌گیری شده خارج کنیم، چگونه آن‌ها را نمایش و ارائه دهیم و در نهایت چگونه آن‌ها را به داده‌ها وارد کنیم.



شکل (۱-۱) : پروفسور سوانته ولد

پس از آن تا به امروز دانشمندان کمومتریکس تعریف‌های بسیاری را برای عرصه‌ی کمومتریکس بیان کرده‌اند که برایندی از آنها را می توان به شکل زیر بیان کرد [۳]:

- *کمومتریکس به عنوان زیرشاخه‌ای از شیمی، رهیافتی مبتنی پایه‌های علم شیمی و قوانین ریاضی و آمار است که الف) وظیفه‌ی طراحی و یا انتخاب روش اندازه-گیری و آزمایش و ب) فراهم نمودن بیشترین اطلاعات مورد نیاز از سیستم مورد نظر به وسیله‌ی تحلیل داده‌های شیمیایی را به عهده دارد.*

با طرح یک مثال می‌توان چگونگی برخورد یک فرد را با کمومتریکس و استفاده‌ی از آن بیان کرد: در ابتدا شیمیدان با استفاده از یک دانش پایه‌ای شیمی در زمینه‌ی مورد نظر خود فرضیه‌ای را تعریف می‌کند. سپس برای اینکه بتوانند این فرضیه را ارزیابی کند و اعتبار این فرض را نشان بدهد به داده‌های آزمایشگاهی نیاز دارد. راهکارهای پیش روی دانشمندان کمومتریکس به این صورت است که با کمک‌گرفتن از روش‌های ریاضی و آمار و ارتباط

دادن آن‌ها با علم شیمی دستورکار بهینه‌ای را برای بدست‌آوردن داده‌های مورد نیاز و به مفیدترین شکل آن ارائه دهد. معمولا این مورد با استفاده از روش‌های طراحی آزمایش[1] انجام می‌شود [٤].

از آنجاییکه معمولا نمی‌توان از داده‌های به اصطلاح خام بدست آمده از آزمایش‌ها اطلاعات موردنظر را در مورد سیستم مورد بررسی بدست‌آورد، وظیفه‌ی بعدی و اصلی کمومتریکس در اینجا بیان می‌شود. دانشمندان عرصه‌ی کمومتریکس با استناد به قوانین، اصول و روش‌های آمار و ریاضی، در پی بهبود روش‌های موجود و طرح روش‌های جدیدتر به تحلیل داده‌های خام می‌پردازند. سپس شیمیدان می‌تواند با استفاده‌ی از این اطلاعات و دانش شیمیایی خود، دانستنی‌های بیش‌تری را در مورد سیستم موردنظر مطالعه بدست‌آورد.

همانطور که می‌دانیم امروزه با توسعه‌ی دستگاه‌های آزمایشگاهی برای هر آزمایش سیل عظیمی از داده بدست‌می آید، این امر سبب شده‌است که بیش‌تر شیمیدانان تجزیه نقش توسعه‌دهنده‌ی روش‌های کمومتریکس و همچنین پیداکردن کاربردهای جدید آن‌ها در علم شیمی را داشته‌باشند.

نکته‌ای را که پروفسور سوانته ولد بنیانگذار کمومتریکس و دیگر دانشمندان عرصه‌ی کمومتریکس همواره بدان اشاره داشتند این است که کاربران روش‌های کمومتریکس می‌بایست در خاطر داشته باشند که شیمی به عنوان مرکز و پایه‌ی کمومتریکس است و رهیافت‌های کمومتریکس بدون توجه به اصول شیمی و بدون مورد قبول بودن در حیطه‌ی شیمی، هیچ گونه ارزشی ندارد [٢].

جدا از کاربردهای بهینه‌سازی آزمایش توسط روش‌های کمومتریکس، با توجه به هدف کاربر از روش‌های کمومتریکس، این روش‌ها می‌توانند به شکل‌های توصیفی[2] و یا پیشگویانه[3] بکار روند. در شکل توصیفی خواص سیستم‌های شیمیایی به تنها منظور شناخت ساختار نهفته در داده و همچنین رابطه‌ی بین نمونه‌ها، مدل می‌شوند. در شکل پیشگویانه مدل‌سازی به منظور شناخت ساختار، رفتار و خواص نمونه‌های جدید ایجاد می‌شود.

[1] Experimental Design
[2] Descriptive
[3] Predictive

۱. ۱. ۳ طراحی دارو

با توجه به آنچه که برای تعریف کمومتریکس در بالا بدان اشاره‌شد، می‌توان به بالا بودن جایگاه بزرگ کمومتریکس در علم شیمی پی برد. امروزه با تعامل هرچه بیشتر شاخه‌های مختلف علوم و مخصوصا در بین شیمی و علوم زیستی، کمومتریکس توانسته‌است که مشکلات زیادی را از پیش روی مطالعات زیستی بردارد. یکی از مواردی که اخیرا به وسیله‌ی دانشمندان کمومتریکس توسعه‌ی چشمگیری داشته‌است، مبحث طراحی دارو[1] و دسته‌ی QSAR/QSPR (مطالعه‌ی رابطه‌ی کمی فعالیت/خاصیت – ساختار) می‌باشد.

شیمی محاسباتی پتانسیل قابل توجهی برای طراحی دارو دارد، به منظور بهبود شرایط و استفاده‌ی منطقی از شیمی محاسباتی بهتر است که مدل سازی مولکولی با تحلیل ریاضی رابطه‌ی بین پاسخ‌های آشکار شده توسط آزمایش های زیستی[2] و مجموعه ای مناسب از خواص مولکولی که از روش های محاسباتی بدست آمده اند، همراه باشد. یک راه معمولی قبول شده برای غلبه بر نبود داده‌های آزمایشگاهی در پدیده‌های شیمیایی پیچیده نظیر طراحی دارو، آنالیز بر اساس رابطه های کمی بین ساختار و فعالیت (خواص) است.

۱. ۲ کمومتریکس در QSAR/QSPR

اکنون با استفاده از تعریف کمومتریکس به چگونگی کاربرد کمومتریکس در QSAR می‌پردازیم:

- فرض اولیه‌ی بکار رفته در QSAR بر پایه‌ی دانش‌های شیمی و زیستی حدود بیش از یک قرن و نیم پیش در سال ۱۸۴۱ میلادی توسط براون[3] و فرزر[4] به این صورت بیان شد که فعالیت زیستی یک ماده به ساختار (ترکیب) آن وابسته است [۵]. در مرحله‌ی بعد تعدادی مولکول همجنس[5] انتخاب می‌شوند، سپس میزان فعالیت زیستی آنها به وسیله‌ی یک روش معتبر که می‌تواند با روش طراحی آزمایش بهینه شده باشد، به طور یکسان برای تمام مولکول‌ها اندازه‌گیری می شود

[1] Drug Design
[2] Bioassays
[3] A. Crum Brown
[4] Thomas R. Fraser
[5] Congeneric

همچنین رمزی کردن[1] اطلاعات مربوط به ساختار مولکول‌ها که توسط روش‌های ریاضی انجام می‌شود. در مرحله‌ی دوم است که با استفاده از روش‌های گوناگون مدل‌سازی، سعی بر ایجاد یک مدل بهینه و معتبر از نقطه نظر شیمیایی و آماری و با کیفیت از لحاظ پیش‌پیش بینی می‌شود. کیفیت مدل ایجاد شده در مرحله‌ی دوم نه تنها به روش‌های بکاررفته در مدل‌سازی، بلکه به صحت و درستی داده‌های بدست‌آمده از آزمایش در مرحله‌ی اول نیز یستگی دارد. سپس از روی مدل ایجاد شده، سعی در تفسیر رفتار مولکول‌ها بر اساس ساختار آنها می‌شود.

با توجه به موارد گفته‌شده در بالا، نمودار گردشی[2] در شکل (۱-۲) شمایی از خط مشی[3] QSAR را نمایش می‌دهد.

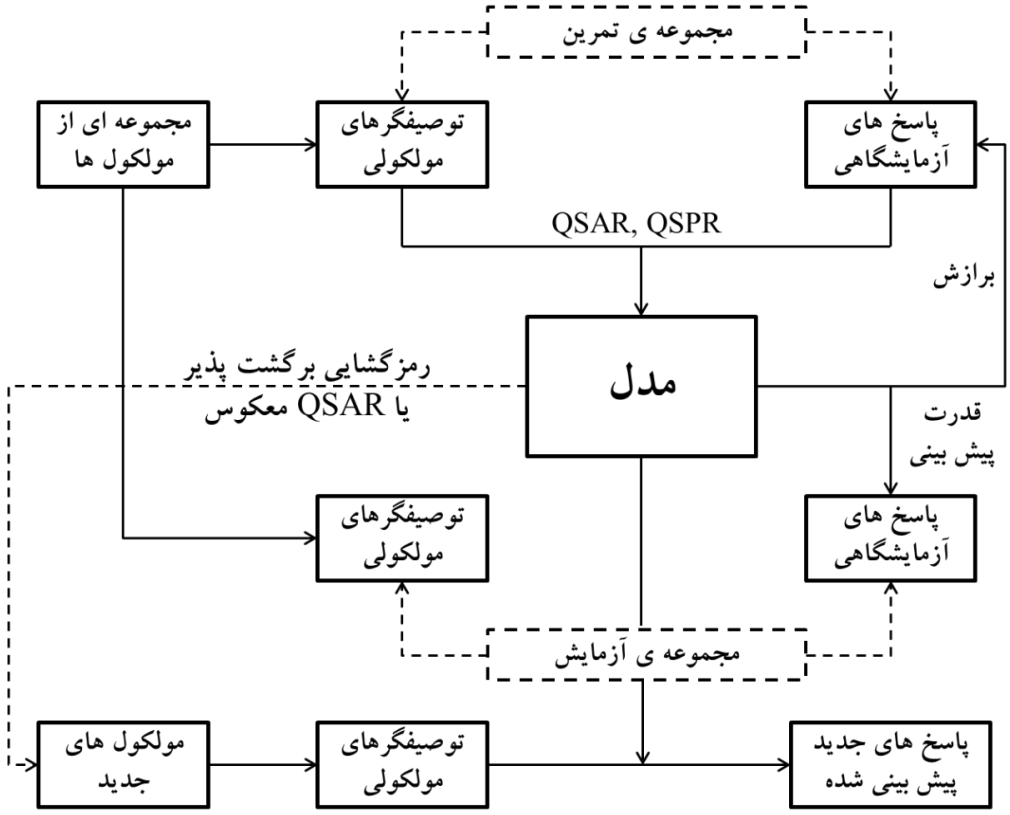
یکی از محدودیت‌های مدل‌های مورد مطالعه‌ی QSAR، کیفیت داده‌های آزمایشگاهی می‌باشد. کیفیت و صحت مدل ایجاد شده و در نهایت فعالیت پیش‌بینی شده توسط مدل، نمی تواند از بیش‌تر از صحت مربوط به داده‌های آزمایشگاهی باشد. همچنین امکان دارد که داده‌های بکاررفته در مدل از منابع مختلف همچون سایر مقالات جمع‌آوری شده‌باشد، بنابراین در این شرایط نیز باید سعی شود برای جلوگیری از گوناگونی و تفاوت در نوع نمونه‌ها، داده‌ها از یک منبع مشخص گرفته شود [٦].

مرحله‌ی نهایی که معمولا هدف اکثر مطالعات QSAR می‌باشد، انتخاب روشی برای ارزیابی و اعتباربخشی[4] مدل است که علاوه بر ارزیابی عملکرد برازش[5] مدل، اجازه‌ی ارزیابی قدرت پیش‌بینی مدل را می‌دهد. معمولا مورد دوم مهمترین خصوصیت برای یک مدل QSAR قابل قبول مطرح می‌شود.

میزان قدرت پیش‌بینی مدل به وسیله‌ی تقسیم کردن مجموعه‌ی داده‌ها به مجموعه‌ی تمرین که مدل از روی آن محاسبه می‌شود و مجموعه‌ی آزمایش که قدرت پیش‌بینی مدل را ارزیابی می‌کند [٦].

[1] Encoding
[2] Flowchart
[3] Strategy
[4] Validation
[5] Fitting

در حدود نزدیک به ۵۰ سال پیش بود که اولین نمونه‌هایی از کاربرد QSAR (مطالعه‌ی رابطه‌ی کمی ساختار-فعالیت) در مطالعات مربوط به شیمی زراعی، شیمی دارویی، سم‌شناسی و سرانجام در بیش‌تر شاخه‌های شیمی پیداشد [۷].



شکل (۱-۲) : شمای عمومی خط مشی QSAR/QSPR [۶]

افزایش چشمگیر کاربرد و توانایی آن را می‌توان از دو جهت مورد بررسی قرار داد [۸]:

- اهمیت فرض اولیه‌ای که نشان می‌دهد فعالیت زیستی مولکول‌ها به صورت تابعی از ساختاری است که به وسیله‌ی ویژگی‌های الکترونیکی، آب‌گریزی و خصوصیات فضایی بیان می شود.
- رشد و توسعه‌ی سریع در روش‌ها و ابزارهای محاسباتی که به دنبال تعیین و تصحیح‌کردن تعداد زیادی از متغیرهایی که مولکول‌ها را توصیف می‌کنند، بوجودآمده است.

با توجه به آنچه که تا به اینجا در مورد QSAR/QSPR توضیح داده شد می‌توان بیان کرد که این دو رهیافت بر این اصل استوار هستند که ترکیب‌های متفاوت شیمیایی می‌توانند به صورت ریاضی گونه‌ای از خواص مولکولی و یا توصیف‌گرهای مولکولی رمزگذاری شوند. بنابراین یک روش مدل‌سازی آماری قادر خواهد بود تا به جستجوی وابستگی موجود در بین توصیف‌گرهای مولکولی و خصوصیت هدف تعریف شده پرداخته و از آن برای پیش‌بینی صحیح خصوصیت متناظر ترکیب‌های جدید استفاده کند. در اینجا اگر هدف نهایی مورد بررسی و پیش‌بینی (فعالیت موردنظر مولکول) به صورت مجموعه‌ای از اعداد پیوسته باشد، قالب رگرسیون برای مدل-سازی انتخاب می‌شود. در حالیکه اگر هدف از رمزگذاری مولکول‌ها شناسایی دسته‌های مختلف در ترکیبات باشد از روش‌های طبقه‌بندی مجزا[1] بدین منظور استفاده می‌شود. (در این مطالعه مدل‌سازی صرفا برای پیش‌بینی فعالیت مولکول‌های مورد بررسی انجام خواهد گرفت).

در هر رهیافت مبتنی بر QSAR/QSPR، چند مرحله‌ی کلیدی با عناوین جمع‌آوری داده، محاسبه‌ی توصیف-گرهای مولکولی مناسب از روی ساختارهای شیمیایی متناظر، در صورت ضرورت انتخاب آنها، ساخت مدل و در پایان ارزیابی درونی و بیرونی مدل ساخته‌شده وجود دارد[۹].

در اینجا سعی شده‌است تا با ترتیب مراحل بیان‌شده در بالا، داده‌های مورد استفاده، روش‌های انتخاب توصیفگر، روش‌های مختلف مدل‌سازی برای داده‌های چند متغیره، روش‌های ارزیابی مدل و به طور کلی فرایند پیش روی این مطالعه توضیح داده شود. از آنجاییکه در اکثر موارد، روش‌های انتخاب متغیر با مدل‌سازی همراه است، در ادامه ابتدا در مورد روش‌های مدل‌سازی و سپس روش‌های انتخاب متغیر بحث خواهدشد.

۱. ۳ داده‌های مورد استفاده در این مطالعه

داده‌هایی که در این مطالعه از آن‌ها استفاده شده شامل سه مورد گابا[2]، فلور[3] و ایدز[4] است که به وسیله‌ی پژوهشگران تعدادی توصیفگر برای مولکول‌های مورد نظر محاسبه شده‌است.

[1] Discrete Classification
[2] GABA
[3] FLUOR
[4] HIV

FLUOR : شامل ۱۱٦ ترکیب آلی که توسط ۱۲٦۸ توصیفگر نظری بیان شده‌است. متغیر معیار در این داده میزان فلوئور دوستی این ۱۱٦ ترکیب آلی است [۱۰].

GABA : این داده شامل ۷۸ مشتق از ترکیب فلاون است که به همراه ۱۱۸۷ توصیفگر مولکولی میزان تمایل لیگانهای فلاونوئید را برای قرارگرفتن در محل‌های بنزودی‌آزپین پذیرنده‌های گابا [۱۱] مورد بررسی قرار می‌هد [۱۰].

HIV : ۱۰۷ مشتق از داروی HEPT به همراه ۱٦۰ توصیفگر مولکولی این داده را تشکیل می‌هند و فعالیت مورد بررسی در این مطالعه نیمه‌ی بیش‌ترین غلظت موثر از این دارو برای حفاظت از سلول‌های MT-4 در برابر بیماری HIV-1 است [۱۲].

۱. ٤ روش‌های مدل‌سازی

داده‌های اندازه‌گیری شده همان اطلاعات نیستند. بنابراین یک مورد مهم در تمام علوم تجربی که وظیفه‌ی اندازه-گیری داده‌ها را به عهده دارد، این است که چگونه اطلاعات مناسب برای هدف آزمایش از داده‌ها آشکار شود [۱۳]. بدست‌آوردن اطلاعات دقیق‌تر که نتیجه‌ی استفاده‌ی از داده‌های چندمتغیره نسبت به تک متغیره است [۱٤] و همچنین توسعه‌ی روزافزون دستگاه‌های تجزیه‌ای باعث شده تا بیان یک نمونه با صدها و یا هزاران متغیر اندازه‌گیری شده (مانند طول موج، زمان بازداری، نسبت جرم به بار و یا توصیفگرهای شیمیایی) رواج یابد. بنابراین به منظور بدست‌آوردن اطلاعات مناسب و با دقت کافی در مورد سیستم مورد نظر، می‌بایست از روش‌های مدل‌سازی چند متغیره استفاده کرد. بیشتر مسائل مربوط به مدل‌سازی در علوم کاربردی می‌تواند در قالب رگرسیون بیان شود [۱۵]. از این رو یک مدل رگرسیون به منظور ارتباط دادن مجموعه‌ای از متغیرهای توصیفگر (متغیرهای مستقل[1]) به یک متغیر معیار (متغیر وابسته[2]) استفاده می شود.

[1] Independent variables
[2] Dependent variable

به طور کلی روش‌های رگرسیون برای رسیدن به دو هدف استفاده می‌شود که در زیر با مثالی بیان شده‌اند:

- کاربرد رگرسیون در فرایندهای صنعتی شیمیایی منحصرا به منظور پیش‌بینی است. از آنجاییکه اندازه‌گیری تمامی متغیرهای درگیر در سیستم مورد بررسی در فرایندهای صنعتی آسان نبوده و همراه با پیچیدگی‌های خاصی همراه است، مدل رگرسیون بکار رفته در اینجا به منظور ارتباط دادن متغیرهای اندازه‌گیری شده به متغیری است که تعیین میزان آن اهمیت زیادی در سیستم داشته و اندازه‌گیری مستقیم آن امکان‌پذیر نبوده و یا پیچیده است. بنابراین مدل رگرسیون با ارتباط بین این دو دسته قاعده‌ی پیش‌بینی را بنا می نهد و اندازه‌گیری متغیرهای پیچیده را امکان می‌سازد.
- کاربرد دیگر آن را می‌توان در زمینه‌ی فناوری زیستی[1] مانند QSAR/QSPR مثال زد، به طوری که در اینجا هدف از مدل‌سازی بر اساس رگرسیون پی بردن به چگونگی تاثیر هر کدام از توصیفگرهای ساختاری بر فعالیت زیستی مولکول مورد نظر ویا تخمین فعالیت مولکول‌های جدید است.

۱. ٤. ۱ رگرسیون خطی چندتایی[2]:

هدف یک روش رگرسیون ارتباط دادن مجموعه‌ای از متغیرهای مستقل (توصیفگرهای شیمیایی) به متغیر وابسته (متغیر معیار) است که عموما به شکل خطی زیر بیان می‌شود:

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i \ , \qquad i = 1, \dots, n \qquad (۱-۱)$$

در اینجا y_i به عنوان i امین متغیر وابسته (مقدار فعالیت و یا خاصیت)، $\boldsymbol{x_i}$، i امین سطر از ماتریس داده که رمزی شده‌ی یک مولکول به شکل عددی است. $\boldsymbol{\beta}$ بردار ضرایب رگرسیون و هم بعد با بردار $\boldsymbol{x_i}$ و ε_i مقدار خطای مدل در پیش‌بینی مربوط به فعالیت i امین مولکول است.

همانطور که می‌دانیم بردار $\boldsymbol{\beta}$ در جهتی بدست می‌آید که مربع تفاوت میزان پیش‌بینی شده‌ی برای بردار $\boldsymbol{y}$ توسط مدل و میزان حقیقی آن کمترین مقدار شود که توسط تابع فقدان[3] کمترین مربعات بیان می‌شود:

[1] Biotechnology
[2] Multiple Linear Regression
[3] Loss function

$$f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 \qquad (۲-۱)$$

اگر ماتریس $\boldsymbol{X'X}$ غیرمنفرد[1] باشد، بردار $\boldsymbol{\beta}$ به صورت منحصر بفرد[2] و با استفاده از رابطه‌ی زیر بدست می‌آید.

$$\boldsymbol{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \qquad (۳-۱)$$

اما اگر در شرایطی ماتریس $\boldsymbol{X'X}$ منفرد[3] باشد (وجود وابستگی در بین متغیرها و ترکیب خطی بودن متغیرها از یکدیگر)، مجموعه‌ای از بردارهای $\boldsymbol{\beta}$ بدست خواهد آمد که تمامی آنها برای معادله‌ی (۱-۲) جواب هستند و از رابطه‌ی زیر بدست می‌آیند:

$$\boldsymbol{\beta} = \boldsymbol{X^+y} + \boldsymbol{NV} \qquad (٤-۱)$$

که در این رابطه بالانشان + نماد معکوس مور- پنروز[4] است، $\boldsymbol{N}$ ماتریس متمم عمود[5] $\boldsymbol{X'}$ است به طوریکه $\boldsymbol{XN} = 0$ و $\boldsymbol{V}$ ماتریسی دلخواه با اندازه‌ی مناسب است [۱۵]. بنابراین با انتخاب‌های متفاوت از $\boldsymbol{V}$ ، بردارهای متفاوتی از $\boldsymbol{\beta}$ بدست خواهد آمد که تمامی آنها معیار معادله‌ی (۱-۲) را دارا هستند. در این مورد یک راه معمول برای انتخاب از بین این مجموعه‌ی بردارها این است که مقدار $\boldsymbol{V} = 0$ را انتخاب کنیم.

در واقع می‌توان رابطه‌ی (۱-۳) را نیز به شکل $\boldsymbol{X^+y}$ بیان نمود، زیرا هنگامی که $\boldsymbol{X'X}$ غیرمنفرد باشد آنگاه $\boldsymbol{X^+} = \boldsymbol{(X'X)^{-1}X'}$ است. هنگامی که ماتریس $\boldsymbol{X'X}$ منفرد بدست می‌آید و یا به عبارتی بعضی از مقادیر تکین[6] آن بسیار کوچک و نزدیک صفر است، رویه‌ی معکوس مور- پنروز سبب می‌شود که معکوس آن مقادیر تکین و به تبع از آن عناصر $\boldsymbol{\beta}$ بسیار بزرگ شوند و جواب‌های بسیاری نیز برای $\boldsymbol{\beta}$ وجود داشته‌باشد. به طور معمول دیده شده‌است که بزرگ‌شدن عناصر بردار $\boldsymbol{\beta}$ باعث کاهش عمومیت این بردار برای استفاده‌ی در مجموعه-های دیگری از داده می‌شود [۱۵].

[1] Nonsingular
[2] Unique
[3] Singular
[4] Moore-Penrose Inverse
[5] Orthocomplement (Orthogonal Complement)
[6] Singular Values

۱. ٤. ۲ رگرسیون مرزی[1]:

یک روش معمول که برای جلوگیری از بزرگ شدن اندازه‌ی بردار $\boldsymbol{\beta}$ استفاده از انقباض[2] است که به صورت زیر در هنگامی که ماتریس $\boldsymbol{X'X}$ منفرد بدست می‌آید، بکار می‌رود:

$$g(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \qquad (۵-۱)$$

فرایند بکار رفته در رابطه‌ی (۱-۵) به رگرسیون مرزی معروف است. عدد λ بیان کننده‌ی میزان جریمه[3] برای کاهش اندازه‌ی بردار $\boldsymbol{\beta}$ است. کمترین مقدار برای رابطه‌ی (۱-۵) از رابطه‌ی زیر بدست می‌آید:

$$\boldsymbol{\beta} = (\boldsymbol{X'X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X'y} \qquad (٦-۱)$$

توجه داشته باشید که تمامی مقادیر منفرد ماتریس معکوس در این رابطه همواره غیر صفر خواهند بود.

با این حال انقباض صورت گرفته بر روی بردار $\boldsymbol{\beta}$ با روش رگرسیون مرزی جوابگوی بهترین جواب برای مدل‌سازی بر اساس رگرسیون نیست و نمی‌تواند پشتوانه‌ای برای مجموعه‌ی دیگری از داده باشد. رگرسیون مرزی نوع دیگر از یک روش کلی‌تر به نام تنظیم تیخونوف[4] است [١٦]. رهیافت دیگری که به منظور بهینه‌کردن فرایند رگرسیون به نظر می‌رسد، استفاده از خلاصه‌ی متغیرهای موجود در ماتریس $\boldsymbol{X}$ است.

۱. ٤. ۳ روش‌های رگرسیون چندمتغیره مبتنی بر متغیرهای پنهان

معمول‌ترین روش برای بدست‌آوردن خلاصه‌ای از متغیرها، استفاده از روش تحلیل اجزای اصلی[5] ($\boldsymbol{PCA}$) است. ماتریس $\boldsymbol{X}$ با استفاده از $\boldsymbol{PCA}$ به دو زیرماتریس امتیاز[6] ($\boldsymbol{T}$) و ظرفیت[7] ($\boldsymbol{P}$) تجزیه می‌شود و هر کدام شامل اجزای اصلی ماتریس $\boldsymbol{X}$ هستند که به صورت ترکیب خطی ستون‌ها در زیرماتریس امتیاز و ترکیب خطی سطرها

[1] Ridge Regression
[2] Shrinkage
[3] Penalty
[4] Tikhonov rigularization
[5] Principal Components Analysis
[6] Score
[7] Loading

در وزن آورده شده‌است و به اصطلاح به آنها متغیرهای پنهان[1] نیز می‌گویند. حال که متغیرهای موجود در ماتریس $\boldsymbol{X}$ توسط روش $\boldsymbol{PCA}$ به تعدادی محدود از متغیرهای پنهان خلاصه‌شد، می‌توان از این متغیرها به عنوان متغیرهای جدیدی که متغیر معیار را توصیف می‌کنند در یک مدل رگرسیون استفاده کرد که روش‌های رگرسیون اجزای اصلی[2] ($\boldsymbol{PCR}$) و کمترین مربعات جزئی[3] ($\boldsymbol{PLS}$) بر مبنای این متغیرها عمل می‌کنند.

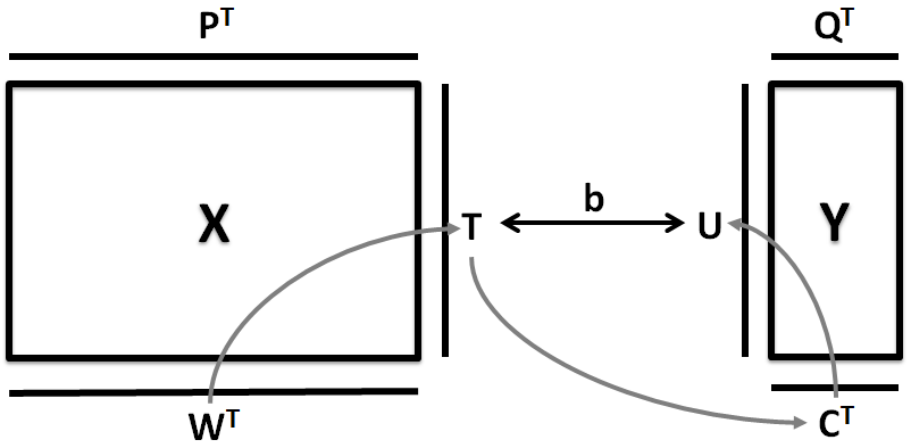
هر دوی این روش‌ها با استفاده از مختصات جدیدی که توسط متغیرهای پنهان برای سطر و ستون ماتریس $\boldsymbol{X}$ توصیف می‌کنند، سعی در یافتن بهترین جواب برای $\boldsymbol{\beta}$ هستند. با توجه به محدودبودن تعداد متغیرهای پنهان (ماتریس $\boldsymbol{T}$) و عمود بودن آنها، ماتریس $\boldsymbol{T'T}$ غیرمنفرد بدست می‌آید و تنها یک جواب بهینه برای $\boldsymbol{\beta}$ وجود خواهد داشت. تفاوت این دو روش در چگونگی بدست‌آوردن ضرایب رگرسیون و یا به عبارتی بدست‌آوردن متغیرهای پنهان (ماتریس $\boldsymbol{T}$) است. در روش $\boldsymbol{PCR}$ این متغیرها با استفاده از روش $\boldsymbol{PCA}$ بدست می‌آیند. یکی از مواردی که علاوه بر ورود متغیرهای پنهان به مدل در مقایسه با متغیرهای اصلی به بهبود مدل ساخته شده کمک می‌کند، خلاصه شدن ماتریس داده در تعداد محدودی متغیر پنهان است که از ورود خطای مربوط به داده‌های اندازه‌گیری شده در مدل رگرسیون تا حدود زیادی جلوگیری می‌کند.

در روش $\boldsymbol{PLS}$ علاوه بر وجود تعداد کمتر متغیرهای پنهان بکارفته در مدل و بهینه بودن ضرایب رگرسیون، دو ویژگی دیگر وجود دارد که باعث برتری آن به روش $\boldsymbol{PCR}$ می‌شود. از آنجاییکه در روش $\boldsymbol{PCR}$ همانطور که در بالا اشاره شد تا حدودی مدل ساخته‌شده از خطای همراه با داده‌ی اندازه‌گیری شده در امان است، در روش $\boldsymbol{PLS}$ به این علت که برای هر دو مجموعه‌ی متغیرهای مستقل و وابسته مجموعه‌ی متغیرهای پنهان بدست می‌آید، کاهش خطا را در هر دو ماتریس خواهیم داشت. در ضمن در روش $\boldsymbol{PCR}$ از آنجاییکه هیچ‌گونه وابستگی بین خطای همراه با ماتریس داده‌های مستقل و وابسته وجود ندارد، متغیرهای پنهان بدست‌آمده برای ماتریس مستقل در جهتی تصادفی و غیر از جهت متغیرهای پنهان ماتریس وابسته بدست می‌آیند. این تفاوت در جهت منجر به زاویه‌دارشدن دو بردار نسبت بهم می‌شود و مدل رگرسیون قابل قبولی را ایجاد نمی‌کند. در روش $\boldsymbol{PLS}$ متغیرهای پنهان در دو ماتریس داده‌ی وابسته و مستقل طوری در فضای پنهان چرخیده می‌شوند که این زاویه به

[1] Latent Variables
[2] Principle Components Regression
[3] Partial least Squares

کمترین مقدار خود برسد و به عبارتی فرایند حذف خطا به مراتب خیلی بهتر از روش ***PCR*** انجام می‌شود. شمایی از چگونگی رویه‌ی روش ***PLS*** در شکل (۱-۳) آورده شده‌است.

بردارهای $\mathbf{t}$ و $\mathbf{u}$ متغیرهای پنهانی هستند که از رویه‌ی **PLS** به ترتیب برای ماتریس‌های $\mathbf{X}$ و $\mathbf{y}$ بدست می‌آیند، اطلاعات تمام ستون‌های $\mathbf{X}$ و $\mathbf{y}$ را دارا هستند و همچنین مقداری از پراکنش[1] غیرمشترک در بین $\mathbf{X}$ و $\mathbf{y}$ را حذف کرده‌اند. به ازای تعداد متغیر پنهان در نظر گرفته برای مدل **PLS** از بردارهای $\mathbf{t}$ و $\mathbf{u}$ بدست می‌آید و در نهایت یک رگرسیون خطی چندمتغیره بین ماتریس‌های $\mathbf{T}$ و $\mathbf{U}$ به منظور مدل‌سازی اطلاعات بین $\mathbf{X}$ و $\mathbf{y}$ انجام می‌شود. همانطور که پیش‌تر بدان اشاره‌شد، متغیرهای پنهان بکاررفته در روش **PLS** طوری چرخیده می‌شوند که بیشترین همبستگی بین $\mathbf{T}$ و $\mathbf{U}$ بدست‌آید. این چرخیدن به این صورت انجام می‌شود که ماتریس $\mathbf{T}$ از ترکیب خطی ستون‌های $\mathbf{X}$ و $\mathbf{w}$ بدست می‌آید که خود بردار وزن $\mathbf{w}$ از ترکیب خطی ستون‌های ماتریس $\mathbf{X}$ و بردار $\mathbf{y}$ بدست می‌آید.



شکل (۱-۳) : شمای کلی رابطه‌ی بین ماتریس‌های $\mathbf{X}$ و $\mathbf{y}$ در رویه‌ی PLS

اگرچه روش‌های بر پایه‌ی متغیرهای پنهان در ابتدا در روانشناسی معرفی شدند، امروزه در بسیاری از زمینه‌ها مانند علوم اجتماعی، علوم طبیعی و اقتصاد استفاده می‌شوند. ورود این روش‌ها را از روانشناسی به دیگر رشته‌ها را می‌توان از اواخر دهه‌ی پنجاه میلادی تخمین زد.

[1] Variance