دانشگاه صنعتی شریف
دانشکده‌ی مهندسی برق

گزارش پروژه‌ی درس تئوری یادگیری ماشین

حل مسئله‌ی کاهش بُعد غیرخطی تفسیرپذیر

محمّدرضا رحمانی بهراد منیری

استاد

دکتر محمّدعلی مدّاح‌علی

زمستان ۱۳۹۸

فهرست مطالب

نمادگذاری

ماتریس	X
اثر ماتریس X	$\mathrm{tr}(X)$
ترانهاده‌ی ماتریس X	$X^\top$
وارون ماتریس X	X^{-1}
ضرب هادامارد دو ماتریس X و Y	$X \odot Y$
ماتریس همانی از مرتبه‌ی n	I_n
بردار	$\mathbf{x}$
نرم p بردار $\mathbf{x}$	$\|\mathbf{x}\|_p$
بردار تمام‌یک n تایی	$\mathbf{1}_n$
مجموعه‌ی اعداد حقیقی	$\mathbb{R}$
مجموعه‌ی اعداد طبیعی	$\mathbb{N}$
فضای توزیع‌های احتمال	$\mathcal{P}$
RKHS ساخته‌شده توسط کرنل k	$\mathcal{H}_k$
تابع علامت	sign
گرادیان نسبت به W	∇_W
هسین نسبت به W	∇_{WW}

فصل ۱

مقدمه

در این پروژه، به بررسی مقاله‌ی [۱] می‌پردازیم. مسئله‌ی اصلی مطرح‌شده در این مقاله، مسئله‌ی کاهش بعد است. کاهش بعد به معنای معرّفی نگاشتی است که داده‌ها را از فضای اصلی، به فضایی با بعد کمتر بنگارد، به گونه‌ای که ویژگی‌های اصلی داده‌ها که می‌توانند داده‌ها را از یکدیگر متمایز کنند، در فضای کاهش بعد یافته حضور داشته باشند. هدف اصلی کاهش بعد، صرفه‌جویی در حافظه‌ی موردنیاز برای ذخیره‌سازی داده‌ها و همچنین پردازش آسان‌تر و سریع‌تر داده‌ها است.

یکی از روش‌های متداول و معروف برای کاهش بعد، روش Principal Component Analysis (PCA) است. این روش برای اوّلین بار توسّط Pearson و در سال ۱۹۰۱ معرّفی شد [۲]. در این روش اگر فرض کنیم داده‌ها به صورت بردارهای $\mathbf{x}_i \in \mathbb{R}^d$ داده شده‌اند، هدف آن است که یک ماتریس $W \in \mathbb{R}^{n\times d}$ بیابیم، به صورتی که $\mathbf{x}'_i = W\mathbf{x}_i \in \mathbb{R}^n$ ویژگی‌های اصلی داده‌ها را حفظ کند. مزیّت اصلی روش PCA در تفسیرپذیر (interpretable) بودن ویژگی‌ها در فضای کاهش بعد یافته است. به تعبیر دیگر با داشتن ماتریس W می‌توان دریافت که ویژگی‌های داده‌ها در فضای اصلی، به چه صورت باهم ترکیب شده‌اند و فضای کاهش بعد یافته را ساخته‌اند. ضعف اصلی روش PCA در آنست که این روش، تنها می‌تواند روابط خطّی بین ویژگی‌ها را استخراج کند و از استخراج روابط غیرخطّی بین ویژگی‌ها ناتوان است.

برای استخراج روابط غیرخطّی در ویژگی‌ها، باید از کرنل‌ها استفاده کنیم. به این معنا که ابتدا با کمک یک کرنل مناسب، داده‌ها را به فضایی با بعد بالاتر بنگاریم و سپس در آن فضا، الگوریتم PCA را اجرا کنیم تا بتوانیم بعد داده‌ها را کم کرده و روابط غیرخطّی بین ویژگی‌ها را نیز شناسایی کنیم. [۳] این الگوریتم را Kernel PCA (KPCA) می‌نامند. [۴، ۵] الگوریتم KPCA بسیار قدرتمند است، ولی دو ضعف دارد:

۱. اگر داده‌ها برچسب داشته باشند، این الگوریتم نمی‌تواند از برچسب‌های آن‌ها برای کاهش بعد دقیق‌تر استفاده‌ای کند

۲. چون PCA در یک فضای با بعد بالاتر از ابعاد داده‌ها اجرا می‌شود، نمی‌توان به دقّت گفت که هرکدام از ویژگی‌ها در فضای کاهش بعد یافته، چگونه به ویژگی‌های اصلی داده‌ها مربوط می‌شوند.

این دو مشکل توسّط الگوریتم Interpretable Kernel Dimension Reduction (IKDR) حل می‌شوند. در این الگوریتم، برخلاف KPCA، داده‌ها ابتدا کاهش بعد داده می‌شوند و بعد از آن کرنل روی آن‌ها عمل می‌کند. همچنین این الگوریتم از برچسب‌های داده‌ها هم استفاده می‌کند و ماتریس Wای را پیدا می‌کند که XW و Y تا حدّ امکان به هم وابسته شوند. برای تعیین میزان وابستگی هم از معیار Hilbert Schmidt Independence Criterion (HSIC) [۶] استفاده می‌کند. [۷، ۸، ۵، ۹، ۴، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴] در حقیقت هدف الگوریتم IKDR حل‌کردن این مسئله است:

$$\max_{X} \ \mathrm{HSIC}(XW, Y) \qquad \text{s.t. } W^\top W = I \qquad (۱.۱)$$

مشکلی که وجود دارد، آن است که این مسئله از نظر محاسباتی پیچیده است، زیرا محدّب نیست و به شدّت غیرخطّی است.

با توجّه به شرط $W^\top W = I$، این مسئله یک مسئله‌ی بهینه‌سازی روی رویه است. در نتیجه قید را می‌توان به صورت یک Stiefel Manifold یا Grassmann manifold در نظر گرفت. [۱۵، ۱۶، ۱۷]

راه‌حل‌های مختلفی برای حلّ این مسئله پیشنهاد شده است. در [۱۸] مسئله‌ی مشابهی روی یک Grassmann manifold در نظر گرفته شده و از روش‌های مرتبه‌ی اوّل و دوم ناحیه اطمینان ریمانی برای حلّ آن استفاده شده است. در [۱۹] از روش ناحیه اطمینان برای کمینه‌کردن یک تابع هزینه روی یک Stiefel Manifold استفاده شده است. روش به کار رفته در [۲۰]، به این صورت است که Stiefel Manifold باز می‌شود و به صورت یک صفحه‌ی تخت درمی‌آید و مسئله‌ی بهینه‌سازی روی این صفحه حل می‌شود. روش‌های مبتنی بر manifold، با تعداد داده‌های کم و هم‌چنین با تعداد ویژگی‌های کم به خوبی جواب می‌دهند، ولی با افزایش تعداد ویژگی‌ها یا افزایش تعداد داده‌ها به شدّت ناکارآمد می‌شوند.

درکنار روش‌های مبتنی بر manifold، [۲۱] روش Dimension Growth (DG) را برای اجرای الگوریتم کاهش گرادیان با یک روش greedy معرّفی می‌کند.

راه‌حل‌های قبلی همگی کند هستند و پیاده‌سازی آن‌ها دشوار است. بهترین راه‌حل برای حلّ این مسئله، الگوریتم Interactive Spectral Method (ISM) است. در [۷] از این الگوریتم برای حلّ مسئله‌ی خوشه‌بندی جایگزین استفاده شده است و مجموعه‌داده‌ای که خوشه‌بندی آن‌ها با روش DG، حدود ۲ روز زمان می‌برد، در ۲ ثانیه خوشه‌بندی شده است. این الگوریتم، به جای جستجو برای یافتن مقدار ویژه‌ها و بردار ویژه‌های ماتریس kernel، به دنبال مقدار ویژه‌ها و بردار ویژه‌های یک ماتریس جایگزین و کوچک‌تر Φ می‌گردد. در نتیجه ISM می‌تواند این مسئله را سریع‌تر حل کند و پیاده‌سازی آسان‌تری هم دارد. ولی از لحاظ تئوری، فقط درمورد کرنل‌های گاوسی همگرایی و اعتبار این الگوریتم تضمین شده است [۷].

در این مقاله، تضمین‌های موجود درباره‌ی اجرای الگوریتم ISM روی کرنل‌های گاوسی، به خانواده‌ی بزرگ‌تری از کرنل‌ها (که ISM Family نامیده شده‌اند) تعمیم داده شده است، ماتریس Φ برای هر عضو این خانواده محاسبه شده است و همچنین این تضمین‌ها درباره‌ی ترکیب‌های خطّی با ضرایب مثبت اعضای این خانواده اثبات شده‌اند.

هم‌چنین چون تضمین‌های تئوری ISM برای همه‌ی کرنل‌های خانواده‌ی ISM اثبات شده است، می‌توان از الگوریتم ISM برای حلّ مسائل گوناگون یادگیری استفاده کرد، مانند کاهش بعد نظارت‌شده [۸، ۵، ۹]، کاهش بعد بدون بی‌نظارت [۴، ۱۰]، کاهش بعد با نظارت ناقص، [۱۱، ۱۲] و مسائل دسته‌بندی جایگزین [۷، ۱۳، ۲۱].

فصل ۲

فرمول‌بندی مسئله

در ابتدا، به تعریف معیار استقلال هیلبرت- اشمیت (HSIC) می‌پردازیم.

۱.۲ معیار استقلال هیلبرت- اشمیت (HSIC)

Hilbert-Schmidt Independence Criterion یا HSIC معیاری برای سنجیدن استقلال دو متغیر تصادفی است. این معیار مشابه اطّلاعات متقابل، فاصله‌ای بین توزیع‌های P_{XY} و $P_X P_Y$ است. اطّلاعات متقابل از فاصله‌ی KL Divergence استفاده می‌کند، در مقابل، HSIC از فاصله Maximum Mean Discrepancy (MMD) بهره می‌برد. ایده‌ی اصلی این فاصله آن است که در ابتدا، نگاشتی بر مبنای یک کرنل از فضای توزیع‌های احتمال به یک RKHS و به شکل زیر ساخته می‌شود:

$$\begin{cases} \mathcal{F} : \mathcal{P} \to \mathcal{H}_k \\ \mathcal{F}(P) = \mathbb{E}_P \left[k(.,X) \right] \end{cases} \tag{۱.۲}$$

در ادامه نیز از فاصله‌ی فضای هیلبرت ساخته شده توسط کرنل، به عنوان فاصله‌ی بین دو توزیع استفاده می‌کند.

$$\mathrm{MMD}(P,Q) = \left\| \mathbb{E}_P \left[k(.,X) \right] - \mathbb{E}_Q \left[k(.,X) \right] \right\|_{\mathcal{H}} \tag{۲.۲}$$

اگر نگاشت بین فضای توزیع‌ها و فضای هیلبرت، نگاشتی یک به یک باشد، در این نگاشت هیچ اطلاعاتی از توزیع از بین نمی‌رود و به راحتی می‌توان نشان داد که عبارت فوق، تمام خواص یک فاصله را دارا می‌باشد. به کرنل‌هایی که در آن‌ها نگاشت مذکور یک به یک است، کرنل مشخصه می‌گویند. کرنل گوسی مشهورترین کرنل مشخصه است. برای یک کرنل مشخصه‌ی $k(.,.)$، تعریف می‌کنیم:

$$\mathrm{HSIC}_k(X,Y) \triangleq \mathrm{MMD}_k(P_X P_Y, P_{XY}) \tag{۳.۲}$$

برای دو متغیّر تصادفی X و Y، $\text{HSIC}_k(X,Y)$ برابر با صفر می‌شود، اگر و تنها اگر X و Y مستقل باشند. بزرگ بودن $\text{HSIC}_k(X,Y)$ نیز به معنای وابستگی این دو متغیر است.

۱.۱.۲ تخمین HSIC از روی داده

حال فرض کنید مجموعه‌ی $\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ با توزیع مشترک P_{XY} داده شده است و قصد داریم بررسی کنیم که آیا X و Y مستقل هستند یا خیر. فرض کنید $X \in \mathbb{R}^{n\times d}$ و $Y \in \mathbb{R}^{n\times c}$ بردارهای مشاهدات باشند. ماتریس‌های گرام مربوط به X و Y را $K_X \in \mathbb{R}^{N\times N}$ و $K_Y \in \mathbb{R}^{N\times N}$ می‌نامیم. ماتریس H را نیز ماتریس centering بگیرید، $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. در این صورت تخمین‌گیر زیر، یک تخمین‌گر نااریب از HSIC است:

$$\text{HSIC}(X,Y) = \frac{1}{(n-1)^2}\text{tr}(K_X H K_Y H) \tag{۴.۲}$$

فرض کنید $X \in \mathbb{R}^{n\times d}$ مجموعه‌ی داده‌ها با d ویژگی و n نمونه و $Y \in \mathbb{R}^{n\times k}$ برچسب‌های این نمونه‌ها (k تعداد دسته‌ها است) باشند. با استفاده از دو کرنل k_x و k_y داده‌شده، دو ماتریس گرام $K_X,\ K_Y \in \mathbb{R}^{n\times n}$ ساخته شده‌اند. در مسئله‌ی IKDR قصد داریم ماتریس $W \in \mathbb{R}^{d\times q}$ را به نحوی بسازیم که $\text{HSIC}(XW,Y)$ بیشینه شود (توجه کنید که $q < d$). دلیل این امر آن است که می خواهیم ماتریس کاهش‌بعدی را بیابیم که بعد از اعمال آن، ویژگی‌های کاهش‌یافته بیشترین اطلاعات ممکن در مورد برچسب‌ها را داشته باشند.

$$\begin{aligned}
\text{HSIC}(XW,Y) &= \frac{1}{(n-1)^2}\text{tr}(K_{XW} H K_Y H)\\
&= \frac{1}{(n-1)^2}\text{tr}(H K_Y H K_{XW})\\
&= \frac{1}{(n-1)^2}\text{tr}(\Gamma K_{XW})
\end{aligned}$$

که در آن $\Gamma = HK_YH$ است.

در نتیجه، مسئله‌ی IKDR را می‌توان به صورت زیر فرمول‌بندی کرد:

$$\max_W \quad \text{tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^\top W = I \tag{۵.۲}$$

شرط $W^\top W = I$ برای اجتناب از جواب‌های نامحدود و حذف ابهام مقیاس افزوده شده است. این مسئله در حالت کلی، مسئله‌ای شدیداً غیر محدب است. هدف اصلی این مقاله و مقالات مشابه، ارائه‌ی الگوریتمی کارا و با تضمین همگرایی، برای حل این مسئله است.

۲.۲ فرمول‌بندی مسائل مختلف یادگیری

۱.۲.۲ کاهش بعد نظارت‌شده

در مسئله‌ی کاهش بُعد نظارت‌شده [۵، ۹]، ماتریس داده‌ها (X) و ماتریس برچسب‌ها (Y) هردو معلوم هستند. در نتیجه کافیست که یک ماتریس W بیابیم که XW و Y تا حدّ امکان به هم وابسته شوند. در نتیجه باید $\mathrm{HSIC}(XW, Y)$ تا حدّ امکان بزرگ شود. از رابطه‌ی (۴.۲) می‌دانیم که بیشینه شدن $\mathrm{HSIC}(XW, Y)$ معادل است با بیشینه شدن $\mathrm{tr}(K_X H K_Y H)$. و از آن‌جا که برچسب‌ها را داریم، ماتریس $\Gamma = HK_Y H$ تماماً معلوم است. پس این مسئله به صورت زیر فرمول‌بندی می‌شود:

$$\max_{W} \ \mathrm{tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^\top W = I \tag{۶.۲}$$

۲.۲.۲ کاهش بعد بدون نظارت

در مسئله‌ی کاهش بعد بدون نظارت، برچسب‌های داده‌ها معلوم نیستند. در نتیجه باید هم W و هم Y یاد گرفته شوند. اگر فرض کنیم $K_Y = YY^\top$، در این صورت می‌توانیم با یک الگوریتم بازگشتی، در هر مرحله Y را تخمین بزنیم و سپس بر مبنای Y تخمین زده شده، W را به گونه‌ای بیابیم که $\mathrm{HSIC}(XW, Y)$ بیشینه شود. امّا سؤال آنست که با فرض داشتن W، چگونه Y را به صورت مناسبی تخمین بزنیم؟ پاسخ این سؤال در [۱۰] بیان شده است. در این مرجع، روشی برای خوشه‌بندی بر مبنای فرمول‌بندی HSIC مطرح شده‌است. هنگامی که Y مشخّص شود، یافتن W مانند بخش قبل است.

۳.۲.۲ کاهش بعد با نظارت ناقص

در مسائل کاهش بعد با نظارت ناقص [۱۲]، برخی از برچسب‌ها برای همه‌ی داده‌ها داده شده‌اند و برخی دیگر نه، یعنی ماتریس برچسب‌ها به صورت $\hat{Y} \in \mathbb{R}^{n\times r}$ است. هم‌چنین فرض می‌شود که دو نمونه‌ی مشابه، برچسب مشابه دارند. در این مسائل، هدف آنست که داده‌ها را با کمک‌گرفتن از برچسب‌های داده‌شده خوشه‌بندی کنیم. خوشه‌بندی را می‌توان با استفاده از خوشه‌بندی طیفی [۲۲] انجام داد و معیار HSIC می‌تواند اطّلاعات برچسب‌هایی که داریم را هم وارد خوشه‌بندی کند. در نتیجه برای آن‌که هم‌زمان کیفیت خوشه‌بندی طیفی خوب باشد و $\mathrm{HSIC}(XW, \hat{Y})$ هم زیاد باشد، مسئله به این صورت فرمول‌بندی می‌شود:

$$\max_{W,Y} \ \mathrm{tr}(Y^T \mathcal{L}_W Y) + \mu \mathrm{tr}(K_{XW} H K_{\hat{Y}} H), \tag{۷.۲}$$

$$\text{s.t.} \quad \mathcal{L}_W = D^{-\frac{1}{2}} K_{XW} D^{-\frac{1}{2}}, \ W^\top W = I, \ Y^\top Y = I \tag{۸.۲}$$

که μ ثابتی است که سهم جمله‌ی اوّل و دوم در بهینه‌سازی را مشخّص می‌کند و $D \in \mathbb{R}^{n\times n}$ ماتریس درجه‌ی K_{XW} است. یعنی یک ماتریس قطری که برای درایه‌های روی قطر آن داریم: $D_{diag} = K_{XW}\mathbf{1}_n$.

مانند مسئله‌ی کاهش بعد بدون نظارت، بهینه‌سازی به صورت بازگشتی و روی W, Y انجام می‌شود. از آن‌جا که جمله‌ی دوم به Y وابسته نیست، وقتی که W معلوم باشد، مسئله به یک خوشه‌بندی طیفی تقلیل می‌یابد و می‌توان Y را محاسبه کرد.

وقتی در هر دور از الگوریتم، Y محاسبه شود، تعریف می‌کنیم:

$$\Psi = HK_{\hat{Y}}H, \qquad \Omega = D^{-\frac{1}{2}}YY^{\top}D^{-\frac{1}{2}}$$

و مسئله‌ی بهینه‌سازی ۷.۲ به مسئله‌ی زیر تبدیل می‌شود:

$$\max_{W,Y} \operatorname{tr}[(\Omega + \mu\Psi)K_{XW}] \qquad \text{s.t. } W^{\top}W = I$$

و اگر تعریف کنیم $\Gamma = \Omega + \mu\Psi$ این بخش از مسئله به مسئله‌ی ۵.۲ تبدیل می‌شود.

۴.۲.۲ دسته‌بندی جایگزین

در مسائل دسته‌بندی جایگزین [۲۱]، یک مجموعه برچسب کامل $\hat{Y} \in \mathbb{R}^{n\times k}$ وجود دارد. این برچسب‌ها را «برچسب‌های اصلی» می‌نامیم.

در این دسته از مسائل، هدف آنست که یک مجموعه برچسب جایگزین بیابیم که کیفیت دسته‌بندی با این برچسب‌ها بالا باشد، ولی برچسب‌های جایگزین تا حدّ امکان با برچسب‌های اصلی متفاوت باشند. به بیان دیگر، هدف آنست که داده‌ها را از منظری دیگر دسته‌بندی کنیم. این مسئله مانند مسئله‌ی کاهش بعد با نظارت ناقص است، با این تفاوت که در آن مسئله هدف این بود که برچسب‌گذاری نهایی تا حدّ امکان با برچسب‌های موجود هم‌خوانی داشته باشد، ولی در این مسئله هدف اینست که برچسب‌گذاری نهایی با برچسب‌های موجود بیشترین فاصله را داشته باشد. در نتیجه فرمول‌بندی این مسئله هم مانند مسئله‌ی کاهش بعد با نظارت ناقص خواهد بود:

$$\max_{W,Y} \ \operatorname{tr}(Y^{\top}\mathcal{L}_W Y) - \mu\operatorname{tr}(K_{XW}HK_{\hat{Y}}H), \tag{۹.۲}$$

$$\text{s.t.} \quad \mathcal{L}_W = D^{-\frac{1}{2}}K_{XW}D^{-\frac{1}{2}},\ W^{\top}W = I,\ Y^{\top}Y = I. \tag{۱۰.۲}$$

مشاهده می‌شود که تنها تغییر، علامت قبل از μ است. روش حلّ این مسئله کاملاً مشابه مسئله‌ی دسته‌بندی با نظارت ناقص است.

فصل ۳

Iterative Spectral Method

در فصل قبل، بیان صورت کلّی مسئله‌ی کاهش بعد غیرخطی تفسیرپذیر پرداختیم. در این فصل، به معرفی الگوریتم Iterative Spectral Method برای حل این مسئله می‌پردازیم. در ابتدا، به معرفی خانواده‌ی بزرگی از کرنل‌ها، به نام کرنل‌های ISM می‌پردازیم و گارانتی‌هایی بر عملکرد الگوریتم Iterative Spectral Method برای کرنل‌های ISM ارائه می‌کنیم.

۱.۳ کرنل‌های ISM

تعریف ۱.۱.۳. (کرنل ISM) کرنل متقارن و مثبت معین $k(.,.)$، یک کرنل ISM است، اگر دو بار مشتق‌پذیر باشد و برای آن داشته باشیم:

$$\forall \mathbf{x}_i,\ \mathbf{x}_j \in \mathbb{R}^d \ \ k(\mathbf{x}_i W, \mathbf{x}_j W) = f(\beta_{ij}) \tag{۱.۳}$$

که در آن

$$\beta_{ij} = a(\mathbf{x}_i, \mathbf{x}_j)^\top W W^\top b(\mathbf{x}_i, \mathbf{x}_j). \tag{۲.۳}$$

دو نمونه‌ی $\mathbf{x}_i$ و $\mathbf{x}_j$ را در نظر می‌گیریم. اگر با کمک ماتریس W بعد این نمونه‌ها را کاهش دهیم، به بردارهای $W^\top \mathbf{x}_i$ و $W^\top \mathbf{x}_j$ می‌رسیم. از طرف دیگر دو تابع برداری $a : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ و $b\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ را در نظر می‌گیریم. اگر این دو تابع را به دو نمونه‌ی $\mathbf{x}_i$ و $\mathbf{x}_j$ اعمال کنیم و خروجی‌های آن‌ها را با ماتریس W کاهش بعد دهیم، به بردارهای $W^\top a(\mathbf{x}_i, \mathbf{x}_j)$ و $W^\top b(\mathbf{x}_i, \mathbf{x}_j)$ می‌رسیم. ضرب داخلی این دو بردار را β_{ij} می‌نامیم.

$$\beta_{ij} = \langle W^\top a(\mathbf{x}_i, \mathbf{x}_j), W^\top b(\mathbf{x}_i, \mathbf{x}_j) \rangle = a(\mathbf{x}_i, \mathbf{x}_j)^\top W W^\top b(\mathbf{x}_i, \mathbf{x}_j)$$

پس تعریف کرنل ISM را می‌توان به این صورت نوشت:

$$k(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) = f\left(\langle W^\top a(\mathbf{x}_i, \mathbf{x}_j), W^\top b(\mathbf{x}_i, \mathbf{x}_j)\rangle\right)$$

در نتیجه کرنل ISM، کرنلی است که اعمال آن بر روی دو نمونه در فضای کاهش بعد یافته، معادل باشد با تابعی از ضرب داخلی دو تابع برداری از آن نمونه‌ها در فضای کاهش بعد یافته.

۲.۳ شرایط لازم مرتبه‌ی اول و مرتبه‌ی دوم

در ابتدا به بیان قضیه‌ی معروف شرایط لازم، در بهینه‌سازی غیرمحدب می‌پردازیم.

قضیه ۱.۲.۳. یک مسئله‌ی بهینه‌سازی مقید غیرمحدّب را به فرم $\min_{h(W)=0} f(W)$ را در نظر بگیرید، که در آن $f : \mathbb{R}^{d\times q} \to \mathbb{R}$ و $h : \mathbb{R}^{d\times q} \to \mathbb{R}^{q\times q}$ توابعی دو بار مشتق‌پذیر با مشتق پیوسته بوده و $\mathcal{L}$ لاگرانژین این مسئله‌ی بهینه‌سازی باشد. آن‌گاه، برای هر بهینه‌ی محلّی، ماتریس Λ^* وجود دارد که در شرایط زیر صدق می‌کند:

- شرط KKT:

$$\nabla_W \mathcal{L}(W^*, \Lambda^*) = 0 \qquad \nabla_\Lambda \mathcal{L}(W^*, \Lambda^*) = 0 \tag{۳.۳}$$

- شرایط لازم مرتبه‌ی دوم:

$$\mathrm{tr}(Z^\top \nabla^2_{WW} \mathcal{L}(W^*, \Lambda^*) Z) \geq 0 \ \ \forall Z \neq 0 \quad \text{with} \quad \nabla h(W^*)^\top Z = 0 \tag{۴.۳}$$

اثبات. اثبات این قضیه در اکثر کتاب‌های بهینه‌سازی غیرمحدّب موجود است. به عنوان مثال [۲۳] را ببینید. □

۳.۳ جواب مسئله‌ی IKDR برای کرنل‌های ISM

در این بخش قصد داریم به بررسی شرایط لازم مرتبه‌ی اول و مرتبه‌ی دوم برای مسئله‌ی بهینه‌سازی IKDR بپردازیم و الگوریتمی ارائه کنیم که نقاط ثابت آن، در این شرایط صدق کنند.

۱.۳.۳ شرایط لازم مرتبه‌ی اول

در فصل قبل، مسئله‌ی IKDR را بدین صورت تعریف کردیم:

$$\max_W \ \mathrm{tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^\top W = I \tag{۵.۳}$$

می‌دانیم $\mathrm{tr}(\Gamma K_{XW}) = \sum_{i,j} \Gamma_{i,j} K_{XW_{i,j}}$. تعریف می‌کنیم $\mathbf{a} = a(x_i, x_j)$ و $\mathbf{b} = b(x_i, x_j)$. با این تعریف، لاگرانژین مسئله برابر می‌شود با:

$$\mathcal{L}(W, \Lambda) = -\sum_{ij} \Gamma_{ij} f(\mathbf{a}^T W W^T \mathbf{b}) - \mathrm{tr}[\Lambda(W^T W - I)]. \tag{۶.۳}$$

با مشتق‌گیری از این عبارت نسبت به W داریم:

$$\nabla_W \mathcal{L}(W, \Lambda) = -\sum_{ij} \Gamma_{ij} f'(\mathbf{a}^T W W^T \mathbf{b})(\mathbf{b}\mathbf{a}^\top + \mathbf{a}\mathbf{b}^\top) W - 2W\Lambda. \tag{۷.۳}$$

اگر تعریف کنیم $A_{i,j} = \mathbf{b}\mathbf{a}^\top + \mathbf{a}\mathbf{b}^\top$ و مشتق لاگرانژین را برابر با صفر قرار دهیم، به دست می‌آید:

$$0 = \left[-\frac{1}{2} \sum_{ij} \Gamma_{ij} f'(\mathbf{a}^T W W^T \mathbf{b}) A_{i,j} \right] W - W\Lambda. \tag{۸.۳}$$

با تعریف $\Psi_{i,j} = -\frac{1}{2}\Gamma_{i,j} f'(\mathbf{a}^T W W^T \mathbf{b})$، معادله‌ی فوق به صورت زیر ساده می‌شود:

$$\left[\sum_{ij} \Psi_{ij} A_{i,j} \right] W = W\Lambda. \tag{۹.۳}$$

اگر تعریف کنیم $\Phi = [\sum_{i,j} \Psi_{ij} A_{i,j}]$ داریم $\Phi W = W\Lambda$. این بدین معناست که بردارهای ویژه‌ی ماتریس Φ در شرایط لازم مرتبه‌ی اول (KKT) صدق می‌کنند. از آن‌جا که بردارهای ویژه‌ی Φ یکه‌ی متعامد هستند، شرط

$$\nabla_\Lambda \mathcal{L} = W^\top W - I = 0 \tag{۱۰.۳}$$

نیز ارضا می‌شود.

۲.۳.۳ شرایط لازم مرتبه‌ی دوم

با در نظر گرفتن قید $h(W) = W^\top W - I = 0$، ابتدا جهت‌های $\nabla h(W^*)^\top Z = 0$ را به دست می‌آوریم. می‌دانیم:

$$\nabla h(W^*)^\top Z = \lim_{t \to 0} \frac{\partial}{\partial t} h(W + tZ), \tag{۱۱.۳}$$

در نتیجه می‌توان شرط را به صورت زیر بازنویسی کرد:

$$\begin{aligned} \nabla h(W^*)^\top Z = 0 &= \lim_{t\to 0} \tfrac{\partial}{\partial t}[(W + tZ)^\top (W + tZ) - I], \\ 0 &= \lim_{t\to 0} \tfrac{\partial}{\partial t}[(W^\top W + tW^\top Z + tZ^\top W + t^2 Z^\top Z) - I], \\ 0 &= \lim_{t\to 0} W^\top Z + Z^\top W + 2tZ^\top Z. \end{aligned} \tag{۱۲.۳}$$

با صفر قرار دادن این عبارت، به دست می‌آید:

$$W^\top Z + Z^\top W = 0 \tag{۱۳.۳}$$

از آن‌جا که Φ یک ماتریس متقارن است، مقادیر ویژه‌ی آن کل فضا را span می‌کنند. دو ماتریس W و $\bar{W}$ را به ترتیب مقادیر ویژه‌ای می‌گیریم که در الگوریتم انتخاب شده و انتخاب نشده‌اند. دو ماتریس جایگذاری (که ستون‌ها را در جای خود بگذارند) B و $\bar{B}$ وجود دارند که:

$$Z = WB + \bar{W}\bar{B} \tag{۱۴.۳}$$

از آن‌جا که بردارهای ویژه‌ی متفاوتی در W و $\bar{W}$ قرار دارند و به واسطه‌ی تعامد بردارهای ویژه، $W^\top \bar{W} = 0$ است. با جای‌گذاری Z در معادله‌ی (۱۳.۳) داریم:

$$\begin{aligned} 0 &= W^\top(WB + \bar{W}\bar{B}) + (WB + \bar{W}\bar{B})^\top W \\ 0 &= B + B^\top. \end{aligned} \tag{۱۵.۳}$$

باید مشتق دوم لاگرانژین را محاسبه کنیم. مجدداً می‌دانیم:

$$\nabla^2_{WW}\mathcal{L}(W,\Lambda)Z = \lim_{t\to 0}\frac{\partial}{\partial t}\nabla\mathcal{L}(W+tZ). \tag{۱۶.۳}$$

در بخش قبل، برای گرادیان لاگرانژین داشتیم:

$$\nabla_W\mathcal{L}(W) = -\frac{1}{2}\left[\sum_{i,j}\Gamma_{i,j}f'(\beta)A_{i,j}\right]W - W\Lambda. \tag{۱۷.۳}$$

با تغییر W به $W+tZ$ در تعریف βW داریم:

$$\begin{aligned} \beta(W+tZ) &= \mathbf{a}(W+tZ)(W+tZ)^\top\mathbf{b}, \\ &= \mathbf{a}^\top WW^\top\mathbf{b} + [\mathbf{a}^\top(WZ^\top + ZW^\top)\mathbf{b}]t + [\mathbf{a}^\top ZZ^\top\mathbf{b}]t^2, \\ &= \beta + c_1 t + c_2 t^2, \end{aligned} \tag{۱۸.۳}$$

حال، با این دانسته‌ها می‌توانیم به سادگی مشتق دوم لاگرانژین را محاسبه کنیم:

$$\nabla^2_{WW}\mathcal{L}(W,\Lambda)Z = \lim_{t\to 0}\frac{\partial}{\partial t}\left[-\frac{1}{2}\sum_{i,j}\Gamma_{i,j}f'(\beta + c_1 t + c_2 t^2)A_{i,j}\right](W+tZ) - (W+tZ)\Lambda. \tag{۱۹.۳}$$

با مشتق‌گیری و محاسه‌ی حد عبارت در $t\to 0$:

$$\nabla^2_{WW}\mathcal{L}(W,\Lambda)Z = \left[-\frac{1}{2}\sum_{i,j}\Gamma_{i,j}f''(\beta)c_1A_{i,j}\right]W + \left[-\frac{1}{2}\sum_{i,j}\Gamma_{i,j}f'(\beta)A_{i,j}\right]Z - Z\Lambda. \tag{۲۰.۳}$$

حال به مسئله‌ی اصلی که محاسبه‌ی $\mathrm{tr}(Z^\top \nabla^2_{WW}\mathcal{L}(W,\Lambda)Z)$ بود باز می‌گردیم. با توجه به معادله‌ی (۲۰.۳)، این عبارت را می‌توان به صورت جمع سه جمله نوشت:

$$\mathrm{tr}(Z^\top \nabla^2_{WW}\mathcal{L}(W,\Lambda)Z) = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 \tag{۲۱.۳}$$

که در آن:

$$\mathcal{T}_1 = \mathrm{tr}\left(Z^\top \left[-\frac{1}{2}\sum_{i,j}\Gamma_{i,j} f''(\beta) c_1 A_{i,j} \right] W \right), \tag{۲۲.۳}$$

$$\mathcal{T}_2 = \mathrm{tr}(Z^\top \Phi Z), \tag{۲۳.۳}$$

$$\mathcal{T}_3 = -\mathrm{tr}(Z^\top Z\Lambda). \tag{۲۴.۳}$$

فرض کنید Λ و $\bar{\Lambda}$ به ترتیب مقادیر ویژه‌ی ماتریس‌های W و $\bar{W}$ باشند. می‌توان عبارت مربوط به $\mathcal{T}_2$ را می‌توان با جای‌گذاری Z ساده کرد.

$$\begin{aligned}
\mathrm{tr}(Z^\top \Phi Z) &= \mathrm{tr}((WB+\bar{W}\bar{B})^\top \Phi (WB+\bar{W}\bar{B})) \\
&= \mathrm{tr}(B^\top W^\top \Phi WB + \bar{B}^\top \bar{W}^\top \Phi WB + B^\top W^\top \Phi \bar{W}\bar{B} + \bar{B}^\top \bar{W}^\top \Phi \bar{W}\bar{B}) \\
&= \mathrm{tr}(B^\top W^\top W\Lambda B + \bar{B}^\top \bar{W}^\top W\Lambda B + B^\top W^\top \bar{W}\bar{\Lambda}\bar{B} + \bar{B}^\top \bar{W}^\top \bar{W}\bar{\Lambda}\bar{B}) \\
&= \mathrm{tr}(B^\top \Lambda B + 0 + 0 + \bar{B}^\top \bar{\Lambda}\bar{B}) \\
&= \mathrm{tr}(B^\top \Lambda B + \bar{B}^\top \bar{\Lambda}\bar{B}).
\end{aligned}$$

با جایگذاری Z در عبارت $\mathcal{T}_3$ نیز به صورت مشابه بالا به دست می‌آید:

$$\begin{aligned}
\mathrm{tr}(Z^\top Z\Lambda) &= -\mathrm{tr}((WB+\bar{W}\bar{B})^\top (WB+\bar{W}\bar{B})\Lambda) \\
&= -\mathrm{tr}(B^\top W^\top WB\Lambda + \bar{B}^\top \bar{W}^\top WB\Lambda + B^\top W^\top \bar{W}\bar{B}\Lambda + \bar{B}^\top \bar{W}^\top \bar{W}\bar{B}\Lambda) \\
&= -\mathrm{tr}(B^\top B\Lambda + 0 + 0 + \bar{B}^\top \bar{B}\Lambda) \\
&= -\mathrm{tr}(B\Lambda B^\top + \bar{B}\Lambda\bar{B}^\top).
\end{aligned}$$

با این وجود، می‌توان شرط لازم مرتبه‌ی دوم را بدین صورت بازنویسی کرد:

$$\mathrm{tr}(Z^\top \nabla^2_{WW}\mathcal{L}(W,\Lambda)Z) = \mathrm{tr}(B^\top \Lambda B) + \mathrm{tr}(\bar{B}^\top \bar{\Lambda}\bar{B}) - \mathrm{tr}(B\Lambda B^\top) - \mathrm{tr}(\bar{B}\Lambda\bar{B}^\top) + \mathcal{T}_1 \geq 0. \tag{۲۵.۳}$$

اما در معادله‌ی (۱۵.۳) نشان دادیم که ماتریس B پاد متقارن است. بنابراین داریم:

$$\mathrm{tr} B\Lambda B^\top = \mathrm{tr}(B^\top \Lambda B)$$

یعنی:

$$\mathrm{tr}(\bar{B}^\top \bar{\Lambda}\bar{B}) - \mathrm{tr}(\bar{B}\Lambda\bar{B}^\top) + \mathcal{T}_1 \geq 0. \tag{۲۶.۳}$$

مقاله در این بخش اثبات دچار اشتباهاتی شده بود که بعد از تماس ما با نویسنده‌ی مقاله و بیان مشکل موجود در اثبات، او آن را اصلاح و نسخه‌ای اصلاح‌شده را برای ما ارسال نمود. ادامه‌ی این اثبات، در مقاله‌ی موجود در سایت NIPS غلط است.

از طرفی، اگر در فرم مربعی $\mathrm{tr}(\bar{B}^\top\bar{\Lambda}\bar{B})$، به ازای هر درایه‌ی عضو $\bar{\Lambda}$، کمینه‌ی درایه‌های آن (یعنی $\min_i \bar{\Lambda}_i$) را بگذاریم، عبارت کوچک‌تر می‌شود. بنابراین:

$$\mathrm{tr}(\bar{B}^\top\bar{\Lambda}\bar{B}) \geq (\min_i \bar{\Lambda}_i)\ \mathrm{tr}(\bar{B}\bar{B}^\top)$$

مشابه عبارت قبل، اگر در $\mathrm{tr}(\bar{B}\Lambda\bar{B}^\top)$، تمام درایه‌ها را با بزرگترین درایه‌ی Λ_j جایگزین کنیم، این عبارت بزرگ‌تر می‌شود:

$$\mathrm{tr}(\bar{B}\Lambda\bar{B}^\top) \leq (\max_j \Lambda_j)\ \mathrm{tr}(\bar{B}^\top\bar{B})$$

در نتیجه، شرط لازم مرتبه‌ی دوم به صورت زیر در می آید:

$$\mathrm{tr}(\bar{B}^\top\bar{\Lambda}\bar{B}) - \mathrm{tr}(\bar{B}\Lambda\bar{B}^\top) + \mathcal{T}_1 \geq (\min_i \bar{\Lambda}_i)\ \mathrm{tr}(\bar{B}\bar{B}^\top) - (\max_j \Lambda_j)\ \mathrm{tr}(\bar{B}^\top\bar{B}) + \mathcal{T}_1 \qquad (۲۷.۳)$$

تعریف کنید $\alpha = \mathrm{tr}(\bar{B}\bar{B}^\top) = \mathrm{tr}(\bar{B}^\top\bar{B})$. بنابراین، برای برقراری شرط لازم مرتبه‌ی دوم (معادل مثبت بودن عبارت بالا)، کافی است داشته باشیم:

$$\min_i(\bar{\Lambda}_i) - \max_j(\Lambda_j) \geq \frac{1}{\alpha}\mathcal{T}_1 = \mathcal{C} \qquad (۲۸.۳)$$

این شرط بدان معناست که فاصله‌ی بین کوچکترین مقدار ویژه‌ی مربوط به بردار ویژه‌های انتخاب‌نشده در الگوریتم، با بزرگترین مقدار ویژه‌ی انتخاب شده، بزرگ‌تر از حدی باشد. یعنی اگر قصد داشته باشیم q بردار ویژه‌ی ماتریس Φ را به عنوان W انتخاب کنیم، باید q بردار ویژه‌ی کوچک را انتخاب نماییم. از آن‌جایی که با توجه به روش‌های مثل Power Method برای محاسبه‌ی مقادیر و بردارهای ویژه، بهتر است که مقادیر ویژه‌ی بزرگ یک ماتریس را پیدا کنیم، فلذا در الگوریتم، ماتریس $-\Phi$ را به عنوان Φ در نظر می‌گیریم و مقادیر ویژه‌ی بزرگ آن را محاسبه می‌کنیم.

۳.۳.۳ الگوریتم ISM

حال، با توجه به نتایج بخش‌های قبل، تلاش می‌کنیم الگوریتمی ارائه دهیم که نقاط ثابت این الگوریتم، در شرایط لازم مرتبه‌ی اول و مرتبه‌ی دوم صدق کنند. با توجه به شرایط لازم مرتبه‌ی اول، برای یک کرنل ISM داده‌شده، به تعریف ماتریس Φ می‌پردازیم.

تعریف ۳.۱.۳. کرنل *ISM* تعریف شده در تعریف (۱.۱.۳) را در نظر بگیرید. ماتریس Φ مربوط به این کرنل و مسئله‌ی بهینه‌سازی

$$\max_W \mathrm{tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^\top W = I \qquad (۲۹.۳)$$

را بدین صورت تعریف می‌کنیم. توجه کنید که همان‌طور که در انتهای بخش قبل شاره شد، علامت منفی به این دلیل اضافه شده که در الگوریتم نهایی، به دنبال بردار ویژه‌های مربوط به بزرگترین مقادیر ویژه باشیم:

$$\Phi = -\frac{1}{2}\sum_{i,j}\Gamma_{i,j}f'(\beta)\Big(b(x_i,x_j)a(x_i,x_j)^\top + a(x_i,x_j)b(x_i,x_j)^\top\Big) \qquad (۳۰.۳)$$

با توجه به تعریف Ψ در معادله‌ی (۹.۳) به طور معادل می‌توان نوشت:

$$\Phi = \sum_{i,j}\Psi_{i,j}A_{i,j} \qquad (۳۱.۳)$$

حال، الگوریتم زیر را در نظر بگیرید:

Algorithm 1 ISM Algorithm

Input : Data X, kernel, Subspace Dimension q
Output : Projected subspace W
Initialization : Initialize Φ_0.
Set W_0 to Dominant Eigenvectors of of Φ_0
While $||\Lambda_i - \Lambda_{i-1}||_2/||\Lambda_i||_2 < \delta$
Compute Φ with W_{k-1}
Set W_k to Dominant Eigenvectors of Φ
End

با توجه به قضایای بالا، این الگوریتم، بسیار طبیعی است! در ابتدا، یک ماتریس Φ اولیه به نام Φ_0 انتخاب کرده‌ایم. در ادامه در مورد انتخاب Φ_0 توضیحاتی روشی ارائه خواهیم داد. سپس q بردار ویژه‌ی غالب Φ_0 را به عنوان W_0 انتخاب می‌کنیم. حال به صورت تکرار شونده، در هر مرحله، به کمک W مرحله‌ی قبل، Φ مربوط به کرنل انتخاب شده در ابتدای الگوریتم را ساخته و W را برابر q بردار ویژه‌ی بزرگ آن قرار می‌دهیم. دلیل این انتخاب، معادله‌ی (۲۸.۳) است.

به طور خلاصه، تا کنون در این فصل نشان داده‌ایم که مقاط ثابت الگوریتم ISM با هر کرنل ISM دلخواه، در شرایط لازم مرتبه‌ی اول و دوم مسئله‌ی بهینه‌سازی IKDR صدق می‌کنند. حال نشان می‌دهیم که این الگوریتم همگراست.

قضیه ۳.۱.۳. دنباله‌ی $\{W_kW_k^\top\}$ تولید شده توسط الگوریتم *ISM* دارای یک زیردنباله‌ی همگراست.

۴.۳ ماتریس Φ_0 و شرایط اولیه‌ی الگوریتم

الگوریتم فوق، در قدم اول نیاز به یک ماتریس Φ_0 به عنوان نقطه‌ی شروع دارد. در حالت کلی، برای یک کرنل ISM، ماتریس Φ می‌تواند تابعی از ماتریس W باشد، بنابراین لازم است به دنبال نقطه‌ی شروعی برای الگوریتم تکرارشونده‌ی ISM باشیم که تابعی از W نباشد و در تکرار اول قابل

محاسبه باشد. برای کرنل‌هایی که در آن Φ تابع W نیست، جواب مسئله‌ی IKDR به وضوح دارای یک فرم بسته است که این فرم بسته، مشابه PCA بزرگ‌ترین بردارهای ویژه‌ی Φ هستند.

از آن‌جایی که مسئله‌ی IKDR مسئله‌ای به شدت غیرمحدب است، انتخاب نقطه‌ی اولیه در آن از اهمیت بسیار بالایی برخوردار است.

قضیه ۴.۱.۳. هر کرنل خانواده‌ی *ISM* را می‌توان به صورت زیر تقریب زد. این تقریب، مستقل از W است.

$$\Phi \approx \text{sign}\Big((\nabla_\beta f(0))\Big) \sum_{i,j} \Gamma_{i,j} A_{i,j} \tag{۳۲.۳}$$

که در آن $\mu = \nabla_\beta f(0)$.

اثبات. ابتدا، تابع $f(\beta(W))$ مربوط به کرنل ISM را حول $W = 0$ بسط تیلور می‌دهیم. تا مرتبه‌ی دوم داریم:

$$f(\beta(W)) \approx f(0) + \frac{1}{2!}\text{tr}(W^\top f''(0)W) \tag{۳۳.۳}$$

در نتیجه، می‌توان لاگرانژی مسئله‌ی IKDR را به صورت زیر بازنویسی کرد:

$$\mathcal{L} = -\sum_{i,j} \Gamma_{i,j} \Big[f(0) + \frac{1}{2!}\text{tr}\big(W^T f''(0)W\big)\Big] - \text{tr}\Big[\Lambda(W^\top W - I)\Big] \tag{۳۴.۳}$$

مشابه بخش (۱.۳.۳)، شرط لازم مرتبه‌ی اول را برای این تقریب از مسئله‌ی بهینه‌سازی می‌نویسیم. برای این کار باید هسین تابع $\beta(W)$ را بنویسیم: $\beta(W) = a^T W W^T b = \text{tr}(W^T b a^T W)$. با دو بار مشتق‌گیری، به دست می آید:

$$\nabla_{W,W}\beta(W) = [ba^T + ab^T], \tag{۳۵.۳}$$
$$\nabla_{W,W}\beta(W = 0) = [ba^T + ab^T]. \tag{۳۶.۳}$$

حال می‌توانیم هسین و گرادیان $f(\beta(W))$ را محاسبه کنیم:

$$f(\beta(W)) = f(a^T W W^T b) = f(\text{tr}(W^T b a^T W)), \tag{۳۷.۳}$$
$$f'(\beta(W)) = \nabla_\beta f(\beta(W))[ba^T + ab^T]W = \nabla_\beta f(\beta(W))\nabla_W \beta(W) \tag{۳۸.۳}$$
$$f''(\beta(W) = \nabla_{\beta,\beta} f(\beta(W))[ba^T + ab^T]W(...) + \nabla_\beta f(\beta(W))[ba^T + ab^T] \tag{۳۹.۳}$$
$$f''(\beta(W = 0)) = 0 + \nabla_\beta f(\beta(W))\nabla_{W,W}\beta(W = 0) \tag{۴۰.۳}$$
$$f''(\beta(W = 0)) = \nabla_\beta f(\beta(W))\nabla_{W,W}\beta(W = 0) \tag{۴۱.۳}$$
$$f''(0) = \text{sign}\Big((\nabla_\beta f(0))\Big) A_{i,j}. \tag{۴۲.۳}$$

در نتیجه گرادیان وهسین لاگرانژی به فرم زیر است:

$$\nabla_W \mathcal{L} \approx -\sum_{i,j} \Gamma_{i,j} f''(0) W - 2W\Lambda, \tag{۴۳.۳}$$

$$\nabla_W \mathcal{L} \approx -\text{sign}\Big((\nabla_\beta f(0))\Big) \sum_{i,j} \Gamma_{i,j} A_{i,j} W - 2W\Lambda. \tag{۴۴.۳}$$

با صفر قرار دادن گرادیان لاگرانژي، به معادله‌ی

$$\left[-\text{sign}\big((\nabla_\beta f(0))\big) \sum_{i,j} \Gamma_{i,j} A_{i,j}\right] W = W\Lambda. \tag{۴۵.۳}$$

می‌رسیم. در نتیجه

$$\Phi \approx \text{sign}\Big((\nabla_\beta f(0))\Big) \sum_{i,j} \Gamma_{i,j} A_{i,j} \tag{۴۶.۳}$$

تقریب مرتبه دوم Φ است. $\square$

در الگوریتم ISM از Φ_0 قضیه‌ی فوق به عنوان شرط اولیه استفاده می‌شود. در بخش بعد ماتریس‌های Φ_0 کرنل‌های معروف ISM محاسبه شده‌اند.

۵.۳ برخی از اعضای معروف خانواده‌ی ISM

در این بخش، بعضی کرنل‌های معروف را در نظر می‌گیریم و نشان می‌دهیم تعریف کرنل ISM درباره‌ی آن‌ها صدق می‌کند و توابع $a(\mathbf{x}_i, \mathbf{x}_j)$، $b(\mathbf{x}_i, \mathbf{x}_j)$ و $f(\beta_{ij})$ را برای آن‌ها محاسبه می‌کنیم.

۱.۵.۳ Linear Kernel

تابع $K(.,.)$ برای کرنل خطّی به این صورت است:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle + c$$

در نتیجه داریم:

$$K(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) = \langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j \rangle + c$$

در نتیجه کرنل خطّی، یک کرنل ISM است و داریم:

$$a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \qquad b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_j$$

$$\beta_{ij} = \langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j \rangle$$

$$f(\beta_{ij}) = \beta_{ij} + c$$

۲.۵.۳ Polynomial Kernel

تابع $K(.,.)$ برای کرنل چندجمله‌ای به این صورت است:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\alpha\langle \mathbf{x}_i, \mathbf{x}_j\rangle + c)^p$$

در نتیجه داریم:

$$K(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) = (\alpha\langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j\rangle + c)^p$$

در نتیجه کرنل چندجمله‌ای، یک کرنل ISM است و داریم:

$$a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \qquad\qquad b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_j$$

$$\beta_{ij} = \langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j\rangle$$
$$f(\beta_{ij}) = (\alpha\beta_{ij} + c)^p$$

۳.۵.۳ Gaussian Kernel

تابع $K(.,.)$ برای کرنل گاوسی به این صورت است:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right)$$

در نتیجه داریم:

$$\begin{aligned} K(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) &= \exp\left(-\frac{\left\|W^\top(\mathbf{x}_i - \mathbf{x}_j)\right\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\langle W^\top(\mathbf{x}_i - \mathbf{x}_j), W^\top(\mathbf{x}_i - \mathbf{x}_j)\rangle}{2\sigma^2}\right) \end{aligned}$$

در نتیجه کرنل گاوسی، یک کرنل ISM است و داریم:

$$a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j \qquad\qquad b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j$$

$$\beta_{ij} = \langle W^\top(\mathbf{x}_i - \mathbf{x}_j), W^\top(\mathbf{x}_i - \mathbf{x}_j)\rangle$$
$$f(\beta_{ij}) = e^{-\frac{\beta_{ij}}{2\sigma^2}}$$

۳.۵.۴ Squared Kernel

تابع $K(.,.)$ برای کرنل مربّعی به این صورت است:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2$$

در نتیجه داریم:

$$\begin{aligned} K(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) &= \left\|W^\top (\mathbf{x}_i - \mathbf{x}_j)\right\|^2 \\ &= \langle W^\top (\mathbf{x}_i - \mathbf{x}_j), W^\top (\mathbf{x}_i - \mathbf{x}_j)\rangle \end{aligned}$$

در نتیجه کرنل مربّعی، یک کرنل ISM است و داریم:

$$a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j \qquad b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j$$

$$\beta_{ij} = \langle W^\top (\mathbf{x}_i - \mathbf{x}_j), W^\top (\mathbf{x}_i - \mathbf{x}_j)\rangle$$

$$f(\beta_{ij}) = \beta_{ij}$$

۳.۵.۵ Multiquadratic Kernel

تابع $K(.,.)$ برای کرنل multiquadratic به این صورت است:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 + c^2}$$

در نتیجه داریم:

$$\begin{aligned} K(W^\top \mathbf{x}_i, W^\top \mathbf{x}_j) &= \sqrt{\left\|W^\top (\mathbf{x}_i - \mathbf{x}_j)\right\|^2 + c^2} \\ &= \sqrt{\langle W^\top (\mathbf{x}_i - \mathbf{x}_j), W^\top (\mathbf{x}_i - \mathbf{x}_j)\rangle + c^2} \end{aligned}$$

در نتیجه کرنل multiquadratic، یک کرنل ISM است و داریم:

$$a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j \qquad b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j$$

$$\beta_{ij} = \langle W^\top (\mathbf{x}_i - \mathbf{x}_j), W^\top (\mathbf{x}_i - \mathbf{x}_j)\rangle$$

$$f(\beta_{ij}) = \sqrt{\beta_{ij} + c^2}$$

۶.۳ محاسبه‌ی ماتریس‌های Φ و Φ_0 برای هر کرنل معروف خانواده‌ی ISM

از تعریف ماتریس‌های Φ و Φ_0 می‌دانیم:

$$\Phi = \sum_{i,j} \Psi_{i,j} A_{i,j} = \frac{1}{2} \sum_{i,j} \Gamma_{i,j} f'(\beta_{ij})(\mathbf{b}\mathbf{a}^\top + \mathbf{a}\mathbf{b}^\top) \tag{۴۷.۳}$$

$$\Phi_0 = \mathbf{sign}(\mu) \sum_{i,j} \Gamma_{i,j} A_{i,j} \tag{۴۸.۳}$$

در بخش قبل مشاهده کردیم که در کرنل‌های معروف خانواده‌ی ISM، بردارهای $\mathbf{a}$ و $\mathbf{b}$ یا به صورت

$$\mathbf{a} = a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i, \qquad \mathbf{b} = b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_j$$

هستند و یا به صورت

$$\mathbf{a} = a(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j, \qquad \mathbf{b} = b(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i - \mathbf{x}_j$$

.

$A_{i,j}$ در حالت اوّل به صورت $A_{i,j} = \mathbf{x}_i\mathbf{x}_j^\top + \mathbf{x}_j\mathbf{x}_i^\top$ و در حالت دوم به صورت $A_{i,j} = 2(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$ درمی‌آید. در این دو حالت می‌توانیم حاصل $\Phi = \sum_{i,j} \Psi_{i,j} A_{i,j}$ را مقداری ساده‌تر کنیم.

۱.۶.۳ محاسبه‌ی Φ در حالتی که $A_{i,j} = \mathbf{x}_i\mathbf{x}_j^\top + \mathbf{x}_j\mathbf{x}_i^\top$

در این حالت داریم:

$$\Gamma = HK_YH$$

$$\beta_{ij} = \langle W^\top \mathbf{a}, W^\top \mathbf{b} \rangle = \langle W^\top \mathbf{x}_i, W^\top \mathbf{x}_j \rangle$$

و چون $\Psi_{i,j} = \Gamma_{i,j} f'(\beta_{ij})$، ماتریس Ψ یک ماتریس متقارن است. در نتیجه با آنکه $\mathbf{x}_i\mathbf{x}_j^T \neq \mathbf{x}_j\mathbf{x}_i^T$، داریم:

$$\sum_{i,j} \Psi_{i,j} \mathbf{x}_i\mathbf{x}_j^T = \sum_{i,j} \Psi_{i,j} \mathbf{x}_j\mathbf{x}_i^T. \tag{۴۹.۳}$$

در نتیجه می‌توان نوشت:

$$\sum_{i,j} \Psi_{i,j} A_{i,j} = 2 \sum_{i,j} \Psi_{i,j} \mathbf{x}_i\mathbf{x}_j^T.$$

حال مجموع را برای $i = 1$ بسط می‌دهیم:

$$\begin{aligned}\sum_{j=1}^{n} \Psi_{1,j}\mathbf{x}_1\mathbf{x}_j^T &= [\Psi_{1,1}\mathbf{x}_1\mathbf{x}_1^T + \ldots + \Psi_{1,n}\mathbf{x}_1\mathbf{x}_n^T] \\ &= \mathbf{x}_1[\Psi_{1,1}\mathbf{x}_1^T + \ldots + \Psi_{1,n}\mathbf{x}_n^T] \\ &= \mathbf{x}_1\left[\begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}\begin{bmatrix} \Psi_{1,1} \\ \vdots \\ \Psi_{1,n} \end{bmatrix}\right]^T \\ &= \mathbf{x}_1\left[\begin{bmatrix} \Psi_{1,1} & \ldots & \Psi_{1,n} \end{bmatrix}\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}\right].\end{aligned}$$

و حال روی همه‌ی iها جمع می‌زنیم:

$$\begin{aligned}\sum_{i,j} \Psi_{i,j}\mathbf{x}_i\mathbf{x}_j^T &= x_1\left[\begin{bmatrix} \Psi_{1,1} & \ldots & \Psi_{1,n} \end{bmatrix}\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}\right] + \ldots + \mathbf{x}_n\left[\begin{bmatrix} \Psi_{n,1} & \ldots & \Psi_{n,n} \end{bmatrix}\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}\right] \\ &= \left[\mathbf{x}_1\begin{bmatrix} \Psi_{1,1} & \ldots & \Psi_{1,n} \end{bmatrix} + \ldots + \mathbf{x}_n\begin{bmatrix} \Psi_{n,1} & \ldots & \Psi_{n,n} \end{bmatrix}\right]\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \\ &= \left[\begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}\begin{bmatrix} \Psi_{1,1} \\ \vdots \\ \Psi_{n,1} \end{bmatrix} + \ldots + \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}\begin{bmatrix} \Psi_{1,n} \\ \vdots \\ \Psi_{n,n} \end{bmatrix}\right]\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}\left[\begin{bmatrix} \Psi_{1,1} \\ \vdots \\ \Psi_{n,1} \end{bmatrix} \; \ldots \; \begin{bmatrix} \Psi_{1,n} \\ \vdots \\ \Psi_{n,n} \end{bmatrix}\right]\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}.\end{aligned}$$

و اگر تعریف کنیم $X = \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}^T$، داریم:

$$\sum_{i,j} \Psi_{i,j}A_{i,j} = 2\sum_{i,j} \Psi_{i,j}\mathbf{x}_i\mathbf{x}_j^T = 2X^T\Psi X. \qquad (۵۰.۳)$$

۲.۶.۳ محاسبه‌ی Φ در حالتی که $A_{i,j} = 2(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$

در این حالت هم مشابه حالت قبلی، ماتریس Ψ متقارن است و داریم:

$$\begin{aligned}
\sum_{i,j} \Psi_{i,j} A_{i,j} &= 2\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \\
&= 2\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}_j\mathbf{x}_i^T - \mathbf{x}_i\mathbf{x}_j^T + \mathbf{x}_j\mathbf{x}_j^T) \\
&= 4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}_j\mathbf{x}_i^T) \\
&= \left[4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T)\right] - \left[4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_j\mathbf{x}_i^T)\right].
\end{aligned}$$

اگر مجموع اوّل را در $i = 1$ محاسبه کنیم، داریم:

$$\sum_{j}^{n} \Psi_{1,j}(\mathbf{x}_1\mathbf{x}_1^T) = \Psi_{1,1}(\mathbf{x}_1\mathbf{x}_1^T) + \ldots + \Psi_{1,n}(\mathbf{x}_1\mathbf{x}_1^T) = \left[\sum_{i=1,j}^{n} \Psi_{1,j}\right] \mathbf{x}_1\mathbf{x}_1^T.$$

توجّه کنید که $\left[\sum_{i=1,j}^{n} \Psi_{1,j}\right]$ به معنای جمع سطر اوّل ماتریس Ψ است. اگر ماتریس D_Ψ را اینگونه تعریف کنیم که جمع سطر iام ماتریس Ψ را در درایه‌ی $i.i$ ماتریس D_Ψ بگذاریم و بقیه‌ی درایه‌ها را با صفر پر کنیم، داریم:

$$\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T) = D_{\Psi 1,1}\mathbf{x}_1\mathbf{x}_1^T + \ldots + D_{\Psi n,n}\mathbf{x}_n\mathbf{x}_n^T.$$

در نتیجه:

$$4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T) = 4X^T D_\Psi X.$$

و داریم:

$$\begin{aligned}
\sum_{i,j} \Psi_{i,j} A_{i,j} &= 4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_i\mathbf{x}_i^T) - 4\sum_{i,j} \Psi_{i,j}(\mathbf{x}_j\mathbf{x}_i^T) \\
&= 4X^T D_\Psi X - 4X^T \Psi X \\
&= 4X^T[D_\Psi - \Psi]X.
\end{aligned}$$

در نتیجه:

$$\sum_{i,j} \Psi_{i,j} A_{i,j} = 4X^T[D_\Psi - \Psi]X. \tag{۵۱.۳}$$

۳.۶.۳ محاسبه‌ی Φ_0

بنا به تعریف Φ_0 در (۴.۱.۳) داریم:

$$\Phi_0 = \text{sign}(\mu) \sum_{i,j} \Gamma_{i,j} A_{i,j} \tag{۵۲.۳}$$

در نتیجه، کاملاً مشابه بخش قبل، می‌توان در دو حالت زیر، ماتریس Φ_0 را محاسبه کرد:

- **a** و **b** هر کدام به صورت $x_i - x_j$ باشند:

$$\Phi_0 = \text{sign}(4\mu) X^\top (D_\Gamma - \Gamma) X \tag{۵۳.۳}$$

- (**a**، **b**) به صورت (x_i, x_j) باشد:

$$\Phi_0 = \text{sign}(2\mu) X^\top \Gamma X \tag{۵۴.۳}$$

۴.۶.۳ Linear Kernel

در کرنل خطّی داریم $(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{a}, \mathbf{b})$ و $f(\beta_{ij}) = \beta_{ij} + c$، در نتیجه:

$$\Phi = \frac{1}{2} \sum_{i,j} \Gamma_{i,j} [\nabla_\beta f(\beta_{ij})] A_{i,j} = \frac{1}{2} \sum_{i,j} \Gamma_{i,j} A_{i,j}. \tag{۵۵.۳}$$

و اگر بخواهیم بردارهای ویژه‌ی ماتریس Φ را محاسبه کنیم، تنها علامت ضرایب مؤثّر است و نه مقدار آن‌ها. در نتیجه با توجّه به (۵۰.۳) داریم:

$$\Phi = \text{sign}(2\frac{1}{2}) X^T \Gamma X = X^T \Gamma X. \tag{۵۶.۳}$$

همچنین:

$$\text{sign}(2\nabla_\beta f(\beta)) = \text{sign}(2) = 1. \tag{۵۷.۳}$$

لذا بر طبق (۵۳.۳):

$$\Phi_0 = X^T \Gamma X. \tag{۵۸.۳}$$

۵.۶.۳ Polynomial Kernel

در کرنل چندجمله‌ای داریم $(\mathbf{a},\mathbf{b}) = (\mathbf{x}_i,\mathbf{x}_j)$ و $f(\beta_{ij}) = (\alpha\beta_{ij}+c)^p$، در نتیجه:

$$\Phi = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\nabla_\beta f(\beta_{ij})]A_{i,j} = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\alpha p(\alpha\beta_{ij}+c)^{p-1}]A_{i,j}. \qquad (۵۹.۳)$$

از آن‌جا که α و p ثابتند و $K_{XW,p-1} = (\alpha\beta_{ij}+c)^{p-1}$ خود یک کرنل چندجمله‌ای با توان $p-1$ است، داریم:

$$\Psi = \Gamma \odot K_{XW,p-1} \qquad (۶۰.۳)$$

و از آن‌جا که معمولاً α و p مثبت هستند، با توجّه به (۵۰.۳) داریم:

$$\Phi = \text{sign}(\alpha p)X^T\Psi X = X^T\Psi X = X^T\Gamma\odot K_{XW,p-1}X \qquad (۶۱.۳)$$

همچنین داریم:

$$\text{sign}(2\nabla_\beta f(\beta)) = \text{sign}(2p(\beta+c)^{p-1}) = 1. \qquad (۶۲.۳)$$

بنابراین بر طبق (۵۳.۳) داریم:

$$\Phi_0 = X^T\Gamma X. \qquad (۶۳.۳)$$

۶.۶.۳ Gaussian Kernel

در کرنل گاوسی داریم $(\mathbf{a},\mathbf{b}) = (\mathbf{x}_i-\mathbf{x}_j,\mathbf{x}_i-\mathbf{x}_j)$ و $f(\beta_{ij}) = \exp(\frac{-\beta_{ij}}{2\sigma^2})$، در نتیجه:

$$\begin{aligned}\Phi = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\nabla_\beta f(\beta_{ij})]A_{i,j} &= \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[-\frac{1}{2\sigma^2}e^{-\frac{\beta_{ij}}{2\sigma^2}}]A_{i,j}\\ &= -\frac{1}{4\sigma^2}\sum_{i,j}\Gamma_{i,j}[K_{XW}]_{i,j}A_{i,j}. \qquad (۶۴.۳)\end{aligned}$$

در نتیجه داریم

$$\Psi = \Gamma\odot K_{XW} \qquad (۶۵.۳)$$

و با توجّه به (۵۱.۳) داریم:

$$\Phi = \text{sign}(-\frac{2}{4\sigma^2})X^T(D_\Psi-\Psi)X = -X^T(D_\Psi-\Psi)X. \qquad (۶۶.۳)$$

همچنین داریم:

$$\text{sign}(4\nabla_\beta f(\beta)) = \text{sign}(-\frac{4}{2\sigma^2}e^{-\frac{\beta}{2\sigma^2}}) = -1. \qquad (۶۷.۳)$$

بنابراین بر طبق (۵۳.۳) داریم:

$$\Phi_0 = -X^T(D_\Gamma - \Gamma)X. \qquad (۶۸.۳)$$

۷.۶.۳ Squared Kernel

در کرنل مربّعی داریم $(\mathbf{a}, \mathbf{b}) = (\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j)$ و $f(\beta_{ij}) = \beta_{ij}$، در نتیجه:

$$\Phi = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\nabla_\beta f(\beta_{ij})]A_{i,j} = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}A_{i,j}. \qquad (۶۹.۳)$$

در نتیجه با توجّه به (۵۱.۳) داریم:

$$\Phi = \text{sign}(1)X^T(D_\Gamma - \Gamma)X = X^T(D_\Gamma - \Gamma)X. \qquad (۷۰.۳)$$

$$\text{sign}(4\nabla_\beta f(\beta)) = \text{sign}(\frac{4}{2}(\beta + c^2)^{-1/2}) = 1. \qquad (۷۱.۳)$$

بنابراین بر طبق (۵۲.۳) داریم:

$$\Phi_0 = X^T(D_\Gamma - \Gamma)X. \qquad (۷۲.۳)$$

۸.۶.۳ Multiquadratic Kernel

در کرنل multiquadratic داریم $(\mathbf{a}, \mathbf{b}) = (\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j)$ و $f(\beta_{ij}) = \sqrt{\beta_{ij} + c^2}$، در نتیجه:

$$\begin{aligned}\Phi = \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\nabla_\beta f(\beta_{ij})]A_{i,j} &= \frac{1}{2}\sum_{i,j}\Gamma_{i,j}[\frac{1}{2}(\beta_{ij} + c^2)^{-1/2}]A_{i,j} \\ &= \frac{1}{4}\sum_{i,j}\Gamma_{i,j}[K_{XW}]_{i,j}^{(-1)}A_{i,j}. \qquad (۷۳.۳)\end{aligned}$$

در نتیجه:

$$\Psi = \Gamma \odot K_{XW}^{(-1)} \qquad (۷۴.۳)$$

و با توجّه به (۵۱.۳) داریم:

$$\Phi = \text{sign}(\frac{1}{4})X^T(D_\Psi - \Psi)X = X^T(D_\Psi - \Psi)X. \qquad (۷۵.۳)$$

همچنین داریم:

$$\text{sign}(4\nabla_\beta f(\beta)) = \text{sign}(\frac{4}{2}(\beta + c^2)^{-1/2}) = 1. \qquad (۷۶.۳)$$

بنابراین بر طبق (۵۲.۳)

$$\Phi_0 = X^T(D_\Gamma - \Gamma)X. \qquad (۷۷.۳)$$

۷.۳ ترکیب کرنل‌های ISM

گزاره ۷.۱.۳. ترکیب خطی کرنل‌های *ISM* با ضرایب مثبت، خود یک کرنل *ISM* است.

اثبات. مسئله‌ی بهینه‌سازی

$$\max_W \ \text{tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^\top W = I \qquad (۷۸.۳)$$

با کرنلی که ترکیب خطی m کرنل است، به صورت زیر قابل بازنویسی است:

$$\max_W -\text{tr}(\Gamma\left[\mu_1 K_1 + \mu_2 K_2 + \dots \mu_m K_m)\right] \quad \text{s.t.} \quad W^\top W = I \qquad (۷۹.۳)$$

لاگرانژین این مسئله برابر است با:

$$\mathcal{L} = -\text{tr}(\mu_1 \Gamma K_1) - \text{tr}(\mu_2 \Gamma K_2) - \text{tr}(\mu_m \Gamma K_m) - m\text{tr}(\Lambda[W^\top W - I]) \quad (۸۰.۳)$$

مشابه بخش (۱.۳.۳) می‌توان گرادیان لاگرانژین را محاسبه کرد:

$$\nabla_W \mathcal{L} = [-\mu_1\Phi_1 - \mu_2\Phi_2 - \dots - \mu_m\Phi_m]W - mW\Lambda \qquad (۸۱.۳)$$

که در آن Φ_i ماتریس مربوط به کرنل i ام است. با صفر قرار دادن این گرادیان، داریم:

$$\frac{1}{m}[-\mu_1\Phi_1 - \mu_2\Phi_2 - \dots - \mu_m\Phi_m]W = W\Lambda \qquad (۸۲.۳)$$

این بدین معناست که ترکیب خطی با ضرایب مثبت تعدادی کرنل ISM کرنلی ISM است و ماتریس Φ_S مربوط به این کرنل، ترکیب خطی ماتریس‌های Φ مربوط به کرنل‌های جمع‌شده است:

$$\Phi_S = \mu_1\Phi_1 + \mu_2\Phi_2 + \dots + \mu_m\Phi_m \qquad (۸۳.۳)$$

□

۸.۳ جمع‌بندی

نتایج این بخش را می‌توان به صورت خلاصه در دو جدول زیر مشاهده کرد.

جدول (۱.۳) شامل تعریف کرنل‌های معروف ISM و تابع f هر یک از آن‌هاست.

کرنل	$f(\beta)$	$a(x_i, x_j)$	$b(x_i, x_j)$
Linear	β	x_i	x_j
Squared	β	$x_i - x_j$	$x_i - x_j$
Polynomial	$(\beta + c)^p$	x_i	x_j
Gaussian	$e^{\frac{-\beta}{2\sigma^2}}$	$x_i - x_j$	$x_i - x_j$
Multiquadratic	$\sqrt{\beta + c^2}$	$x_i - x_j$	$x_i - x_j$

جدول ۱.۳: $f(\beta)$ مربوط به کرنل‌های معروف

جدول (۲.۳) شامل ماتریس‌های Φ هر یک از کرنل‌های معرفی شده است.

کرنل	ماتریس Φ
Linear	$\Phi = X^T \Gamma X$
Squared	$\Phi = X^T \mathcal{L}_\Gamma X$
Polynomial	$\Phi = X^T \Psi X$ ، $\Psi = \Gamma \odot K_{XW,p-1}$
Gaussian	$\Phi = -X^T \mathcal{L}_\Psi X$ ، $\Psi = \Gamma \odot K_{XW}$
Multiquadratic	$\Phi = X^T \mathcal{L}_\Psi X$ ، $\Psi = \Gamma \odot K_{XW}^{(-1)}$

جدول ۲.۳: ماتریس‌های Φ برای کرنل‌های معروف

جدول (۳.۳) شامل ماتریس‌های Φ_0 هرکدام از کرنل‌های معرفی شده است.

کرنل	تقریب ماتریس Φ
Linear	$\Phi_0 = X^T \Gamma X$
Squared	$\Phi_0 = X^T \mathcal{L}_\Gamma X$
Polynomial	$\Phi_0 = X^T \Gamma X$
Gaussian	$\Phi_0 = -X^T \mathcal{L}_\Gamma X$
Multiquadratic	$\Phi_0 = X^T \mathcal{L}_\Gamma X$

جدول ۳.۳: ماتریس‌های Φ_0 برای کرنل‌های معروف

فصل ۴

نتایج تجربی

۱.۴ بررسی توابع پیاده‌سازی شده

در این بخش به بررسی کد موجود برای الگوریتم ISM که توسط نویسندگان ارائه شده است می‌پردازیم.

- **تابع `./sdr.py`**

این تابع، مربوط به اجرای تنظیمات اولیه است. با فراخوانی این تابع، دیتاست مورد نظر در db قرار گرفته و توسط تمام توابع دیگر قابل دسترسی می‌شود. آستانه‌ی مورد نظر برای پایان الگوریتم (δ) نیز در این تابع تنظیم می‌شود.

- **تابع `./optimizer/ism.py`**

این تابع، پیاده‌سازی الگوریتم ISM است. این تابع، ماتریس Φ_0 را عنوان ورودی دریافت کرده و ماتریس Φ نهایی را باز می‌گرداند.

- **توابع فولدر `./kernels/`**

در این فولدر، برای هر کرنل یک فایل وجود دارد که در هر فایل، تابعی برای محاسبه‌ی Φ_0 و تابعی برای محاسبه‌ی Φ هر کرنل در آن قرار گرفته است.

۲.۴ نتایج تجربی

۱.۲.۴ توضیح اجمالی در مورد داده‌ها

در مجموعه داده‌ی Wine ویژگی‌ها پیوسته هستند در حالی که در مجموعه‌ی داده‌های Cancer، داده‌ها گسسته می‌باشند. مجموعه‌ی داده‌های MNIST نیز شامل تصاویری از اعداد انگلیسی به صورت دست‌نوشته است.

۲.۲.۴ شرح آزمایش‌ها

کاهش بعد نظارت‌شده

در این آزمایش، ابتدا کاهش بعد انجام می‌شود و سپس بر روی داده‌های کاهش بعد یافته، یک SVM آموزش داده‌شده و درصد صحت در 10-fold cross validation گزارش می‌شود. در هر fold، ماتریس W به کمک الگوریتم ISM و تنها با استفاده از داده‌های یادگیری در آن fold انجام می‌شود.

کاهش‌بعد نظارت‌نشده

بعد از یادگیری W، بر روی داده‌های کاهش‌بعد یافته، یعنی XW، الگوریتم Spectral Clustering را اجرا می‌کنیم. برای سنجش «خوبی» الگوریتم خوشه‌بندی، از معیار NMI استفاده می‌شود. NMI بین دو خوشه‌بندی مختلف به صورت

$$\mathrm{NMI}(L, U) = \frac{I(L;U)}{\sqrt{H(L)H(U)}} \tag{۱.۴}$$

تعریف می‌شود که در آن I اطلاعات متقابل دو لیبل و H آنتروپی هر لیبل است.

دسته‌بندی جایگزین (Alternative Clustering)

برای سنجش عملکرد الگوریتم نیز از تنها یک تصویر استفاده شده و Alternative Clustering به عنوان روشی برای بخش‌بندی تصویر مورد استفاده قرار گرفته است.

۳.۲.۴ تنظیم پارامترها

در آزمایش‌هایی که از کرنل گوسی در آن استفاده شده است، انحراف‌معیار این کرنل برابر میانه‌ی فاصله‌ی دو به دوی نقاط در نظر گرفته شده است. درجه‌ی تمام کرنل‌های چندجمله‌ای ۳ است و بعد فضا بعد از کاهش‌بعد، برابر تعداد لیبل‌های موجود در داده فرض‌شده است. از $\delta = 0.01$ نیز به عنوان شرط توقف ISM استفاده شده است.

۴.۲.۴ نتایج آزمایش‌ها

کاهش بعد نظارت‌شده

در جدول (۲.۴) نتایج آزمایش مربوط به کاهش بعد نظارت شده آمده است. تمام نتایج مربوط به الگوریتم ISM توسط کد ارائه شده توسط نویسندگان مقاله بررسی شده است. در این مقاله، نتابج ISM با روش‌های پیشین SM، GM و GD مقایسه شده‌اند. نویسندگان نسخه‌ی پیاده‌سازی‌شده‌ی روش‌های

پیشین را ارائه نکرده‌اند، بنابراین در جدول فوق، زمان اجرا برای الگوریتم ISM در هر آزمایش، همان زمان ارائه‌شده در مقاله باقی گذاشته شده است که با زمان روش‌های دیگر قابل مقایسه باقی بماند.

با مقایسه‌ی نتایج دیده می‌شود که در کاهش‌بعد نظارت‌شده، در همه‌ی دیتاست‌ها به جز دیتاست Cancer الگوریتم ISM دقت بیشتری داشته است. این بهتر بودن عملکرد، در دیتاست MNIST بهتر از دیگر دیتاست‌ها دیده می‌شود، در این دیتاست، دو روش DG و GM بیش از سه روز زمان نیاز داشتند، در حالی که ISM جواب را در 13.8 ثانیه به دست آورد. بهتر بودن عملکرد الگوریتم ISM تنها محدود به کرنل‌های گاوسی نبوده است و در نتایج هر دو کرنل گاوسی و چندجمله‌ای دیده می‌شود. در دیتاست Caner نیز با وجود عملکرد بهتر الگوریتم GM از نظر دقت، زمان اجرای ISM در آن به مراتب کمتر است.

کاهش بعد نظارت‌نشده

در (۱.۴)، نتایج آزمایش کاهش‌بعد نظارت نشده آمده است. مشابه بخش قبل، تمام نتایج مربوط به الگوریتم ISM توسط کد ارائه شده توسط نویسندگان مقاله بررسی شده است. در این مقاله، نتایج ISM با روش‌های پیشین SM، GM و GD مقایسه شده‌اند. نویسندگان نسخه‌ی پیاده‌سازی‌شده‌ی روش‌های پیشین را ارائه نکرده‌اند، بنابراین در جدول فوق، زمان اجرا برای الگوریتم ISM در هر آزمایش، همان زمان ارائه‌شده در مقاله باقی گذاشته شده است که با زمان روش‌های دیگر قابل مقایسه باقی بماند.

Unsupervised		**Gaussian**				**polynomial**			
		ISM	DG	SM	GM	ISM	DG	SM	GM
Wine	**Time**	**0.01s**	9.9s	0.6s	16.7m	**0.02s**	14.4s	2.9s	33.5m
	Cost	**-27.4**	-25.2	-27.3	-27.3	**-1600**	-1582	-1598	-1496
	NMI	**0.86**	**0.86**	**0.86**	**0.86**	**0.84**	**0.84**	**0.84**	0.83
Cancer	**Time**	**0.57s**	4.3m	3.9s	44m	**0.5s**	8.0m	8.8m	41m
	Cost	**-243**	-133	-146	-142	**-15804**	-14094	-15749	-11985
	NMI	**0.8**	0.79	**0.8**	0.79	0.79	**0.80**	0.79	**0.80**
Face	**Time**	**0.3s**	1.3d	5.3s	55.9m	**1.0s**	> 3d	22m	1.6d
	Cost	**-169.3**	-167.7	-168.9	-37	**-368**	NA	-348	-321
	NMI	0.94	**0.95**	0.93	0.89	**0.94**	N/A	0.89	0.89
MNIST	**Time**	**1.8h**	> 3d	1.3d	> 3d	**8.3m**	> 3d	0.9d	> 3d
	Cost	**-2105**	N/A	-2001	N/A	**-51358**	N/A	-51129	N/A
	NMI	**0.47**	N/A	0.46	N/A	**0.32**	N/A	**0.32**	N/A

جدول ۱.۴: مقایسه‌ی الگوریتم‌های ISM، DG، SM و GM از نظر زمان اجرا و NMI در کاهش بعد نظارت نشده

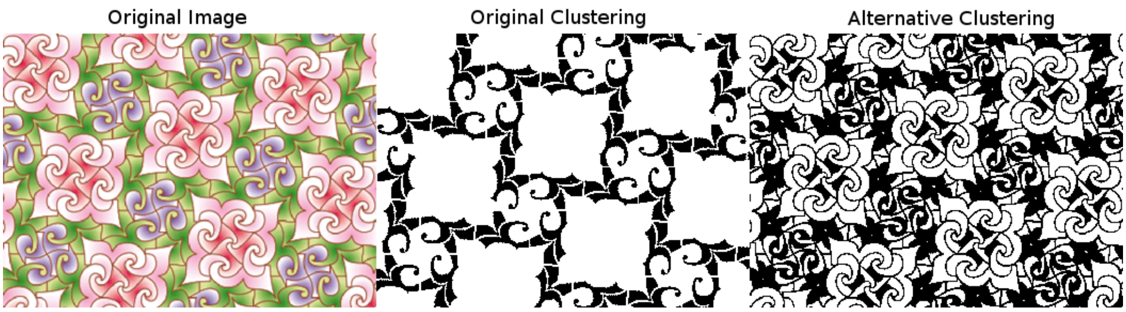
مشاهده می‌شود که با وجود این‌که زمان اجرا در حالت نظارت‌نشده به مراتب بیش از حالت نظارت شده است (که به دلیل نیاز به آپدیت لیبل‌ها در هر تکرار طبیعی است)، هنوز الگوریتم ISM از نظر زمانی از دیگر الگوریتم‌ها عملکرد بهتری دارد. از نظر دقّت (که با NMI سنجیده‌شده است) نیز این الگوریتم عملکردی برابر و یا بهتر از سایر روش‌ها داشته است.

Supervised		**Gaussian**				**polynomial**			
		ISM	DG	SM	GM	ISM	DG	SM	GM
Wine	**Time**	**0.02s ± 0.01s**	7.9s ± 2.9s	1.7s ± 0.7s	16.8m ± 3.4s	**0.02s ± 0.0s**	13.2s ± 6.2s	14.77s ± 0.6s	16.82m ± 3.6s
	Cost	**-1311 ± 26**	-1201 ± 25	-1310 ± 26	-1307 ± 25	**-114608 ± 1752**	-112440 ± 1719	-111339 ± 1652	-108892 ± 1590
	Accuracy	**95.0% ± 5%**	93.2% ± 5.5%	**95% ± 4.2%**	**95% ± 6%**	**97.2% ± 3.7%**	93.8% ± 3.9%	96.6% ± 3.7%	96.6% ± 2.7%
Cancer	**Time**	**0.08s ± 0.0s**	4.5m ± 103s	17s ± 12s	17.8m ± 80s	**0.13s ± 0.0s**	4m ± 1.2m	3.3m ± 3s	17.5m ± 1.1m
	Cost	**-32249 ± 338**	-30302 ± 2297	-31996 ± 499	-30998 ± 560	**-1894 ± 47**	-1882 ± 47	-1737 ± 84	-1690 ± 108
	Accuracy	97.3%± 0.3%	97.3%± 0.3%	97.3%± 0.2%	**97.4%± 0.4%**	**97.4%± 0.3%**	97.3% ± 0.3%	**97.4% ± 0.3%**	97.3% ± 0.3%
Face	**Time**	**0.99s ± 0.1s**	1.92d ± 11h	10s ± 5s	22.7m ± 18s	**0.7s ± 0.03s**	2.1d ± 13.9h	5.0m ± 5.7s	21.5m ± 9.8s
	Cost	**-3754 ± 31**	-3431 ± 32	-3749 ± 33	-771 ± 28	**-82407 ± 1670**	-78845 ± 1503	-37907 ± 15958	-3257 ± 517
	Accuracy	**100% ± 0%**	**100% ± 0%**	**100% ± 0%**	99.2% ± 0.2%	**100% ± 0%**	**100% ± 0%**	**100% ± 0%**	99.8% ± 0.2%
MNIST	**Time**	**13.8s ± 2.3s**	> 3d	2.5m ± 1.0s	> 3d	**12.1s ± 1.4s**	> 3d	2.1m ± 3s	> 3d
	Cost	**-639 ± 2.3**	N/A	-621 ± 5.1	N/A	**-639 ± 2**	N/A	-620 ± 5.1	N/A
	Accuracy	**99% ± 0%**	N/A	98.5% ± 0.4%	N/A	**99% ± 0%**	N/A	**99% ± 0%**	N/A

جدول ۲.۴: مقایسه‌ی الگوریتم‌های ISM، DG، SM و GM از نظر زمان اجرا و دقّت در کاهش بعد نظارت‌شده

دسته‌بندی جایگزین

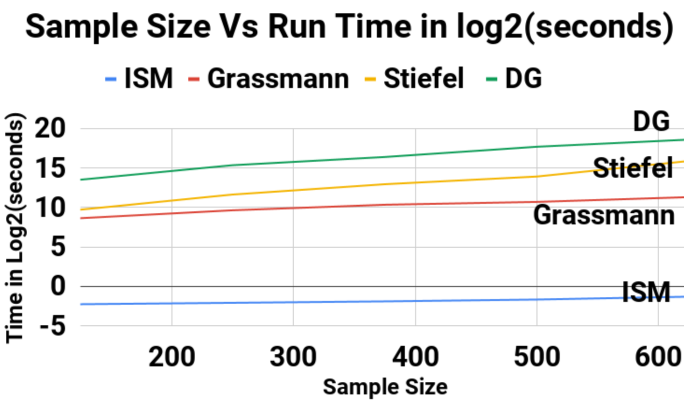
در شکل (۱.۴)، یک نمونه از دسته‌بندی جایگزین به کمک الگوریتم ISM دیده می‌شود. عملکرد یک الگوریتم دسته‌بندی جایگزین را نمی‌توان به راحتی با یک معیار سنجید، ولی در شکل دیده می‌شود که دسته‌بندی جایگزین معرفی‌شده توسط الگوریتم، توانسته پترن دیگری از عکس را به دست آورد که با دسته‌بندی اول متفاوت است.



شکل ۱.۴: دسته‌بندی جایگزین برای یک پترن مخصوص

بررسی زمان اجرا

در این بخش، به بیان نتایج مقایسه‌ی الگوریتم ISM با سایر الگوریتم‌ها، از نظر زمان اجرا می‌پردازیم.



شکل ۲.۴: زمان اجرای الگوریتم‌های مختلف بر جسب تعداد نمونه

مطابق انتظار، خم مربوط به الگوریتم ISM همواره پایین‌تر از منحنی سایر الگوریتم‌ها قرار دارد.

فصل ۵

پیشنهادها

برای ادامه‌ی کار این مقاله، می‌توان چند پیشنهاد به صورت زیر داد:

۱. مقاله‌ی حاضر، در محاسبه‌ی W^*، به ارائه‌ی یک شرط کافی (معادله‌ی ۲۸.۳) برای احراز شرایط مرتبه‌ی دوم مسأله‌ی بهینه‌سازی بسنده کرده و شرایط مطرح شده، لازم نیستند. بهتر بود اگر شرایط لازم برای برقراری شرایط مرتبه‌ی دوم مسأله‌ی بهینه‌سازی بررسی و مطرح می‌شدند. ممکن است این شرایط به الگوریتم دیگری برای محاسبه‌ی نقطه‌ی بهینه منتهی شوند.

۲. در [۲۴]، روشی برای محاسبه‌ی میزان وابستگی دو متغیّر تصادفی به شرط یک متغیّر تصادفی دیگر، بر مبنای متر HSIC بیان شده است. می‌توانیم از این روش استفاده کرده و مسأله‌ای مشابه مسأله‌ی این مقاله، در حوزه‌ی transfer learning مطرح کرد. مسئله به این صورت است که فرض کنیم داده‌های X_1, Y_1 با توزیع $p_1(x, y)$ داریم. حال می‌خواهیم داده‌های X_2, Y_2 با توزیع $p_2(x, y)$ را کاهش بعد دهیم و در این فرآیند از اطّلاعات موجود در داده‌های قبلی هم استفاده کنیم. مسأله می‌تواند به صورت زیر فرمول‌بندی شود:

$$\max_{W} \ \mathrm{HSIC}(X_2 W, Y_2 | X_1, Y_1) \qquad \text{s.t.} \ \ W^\top W = I$$

به عنوان مثال، فرض کنید X_1 داده‌های مربوط به خریدهای افراد مختلف از یک فروشگاه اینترنتی ایرانی باشد و Y_1 سنّ این افراد باشد. همچنین فرض کنید که X_2 اطّلاعات خریدهای افراد دیگری در یک فروشگاه اینترنتی غیرایرانی باشد و Y_2 سنّ این افراد باشد. هدف آنست که با توجّه به داده‌های فروشگاه ایرانی و همچنین برچسب‌های داده‌ها، ماتریس Wای بیابیم که داده‌ها را به نحو مناسبی کاهش بعد دهد.

۳. استفاده از ایده‌های روش‌های عددی بهینه‌سازی غیرمحدب برای حل IKDR

در بهینه‌سازی غیرمحدب، روش‌های زیادی برای فرار از بهینه‌های محلی معرفی شده‌اند. یکی از این روش‌ها، روش Gradual Non-Convexity است. در این روش، ابتدا تخمینی محدب از تابع هدف بهینه‌سازی زده می‌شود، سپس مسئله‌ی بهینه‌سازی برای این تابع هدف

حل شده و جواب آن به دست می‌آید. سپس تخمینی دقیق‌تر (و غیرمحدب‌تر) ارائه می‌شود و الگوریتم بهینه‌سازی از نقطه‌ی نهایی الگوریتم قبل، شروع به حرکت در جهت کمینه می‌کند. این عملیات تکرار می‌شود و در هر مرحله مقداری عدم تحدب به مسئله افزوده شده و از نقطه‌ی پایان مرحله قبل برای شروع بهینه‌سازی استفاده می‌شود. روش‌هایی مبتنی بر Gradual Non-Convexity تا کنون برای حل دسته‌های وسیعی از مسائل کاربرد داشته‌اند، به طور مثال می‌توان به مسئله‌ی بازسازی تنک و کمینه کردن نرم صفر با قید خطی اشاره کرد [۲۵]. بررسی عملکرد الگوریتم‌های مبتنی بر Gradual Non-Convexity و سایر ایده‌های موجود در بهینه‌سازی غیرمحدب برای حل این مسئله نیز یکی از راه‌های پیش‌رو برای توسعه و بهبود روش‌های حل مسئله‌ی IKDR است.

منابع

[1] C. Wu, J. Miller, Y. Chang, M. Sznaier, and J. Dy, "Solving interpretable kernel dimensionality reduction," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019.

[2] F. Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series*, vol.6, no.2, p.11, 1901.

[3] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*, pp.583–588, Springer, 1997.

[4] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol.10, no.5, pp.1299–1319, 1998.

[5] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol.44, no.7, pp.1357–1371, 2011.

[6] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *International conference on algorithmic learning theory*, pp.63–77, Springer, 2005.

[7] C. Wu, S. Ioannidis, M. Sznaier, X. Li, D. Kaeli, and J. Dy, "Iterative spectral method for alternative clustering," in *International Conference on Artificial Intelligence and Statistics*, pp.115–123, 2018.

[8] K. Fukumizu, F. R. Bach, M. I. Jordan, *et al.*, "Kernel dimension reduction in regression," *The Annals of Statistics*, vol.37, no.4, pp.1871–1905, 2009.

[9] M. Masaeli, J. G. Dy, and G. M. Fung, "From transformation-based dimensionality reduction to feature selection," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp.751–758, 2010.

[10] D. Niu, J. Dy, and M. Jordan, "Dimensionality reduction for spectral clustering," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp.552–560, 2011.

[11] M. J. Gangeh, S. M. Bedawi, A. Ghodsi, and F. Karray, "Semi-supervised dictionary learning based on hilbert-schmidt independence criterion," in *International Conference Image Analysis and Recognition*, pp.12–19, Springer, 2016.

[12] Y. Chang, J. Chen, M. H. Cho, P. J. Castaidi, E. K. Silverman, and J. G. Dy, "Clustering with domain-specific usefulness scores," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp.207–215, SIAM, 2017.

[13] D. Niu, J. G. Dy, and M. I. Jordan, "Multiple non-redundant spectral clustering views," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp.831–838, 2010.

[14] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *Journal of Machine Learning Research*, vol.13, no.May, pp.1393–1434, 2012.

[15] I. M. James. *The topology of Stiefel manifolds*, vol.24. Cambridge University Press, 1976.

[16] Y. Nishimori and S. Akaho, "Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold," *Neurocomputing*, vol.67, pp.106–135, 2005.

[17] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM journal on Matrix Analysis and Applications*, vol.20, no.2, pp.303–353, 1998.

[18] N. Boumal and P.-a. Absil, "Rtrmc: A riemannian trust-region method for low-rank matrix completion," in *Advances in neural information processing systems*, pp.406–414, 2011.

[19] F. J. Theis, T. P. Cason, and P.-A. Absil, "Soft dimension reduction for ica by joint diagonalization on the stiefel manifold," in *International Conference on Independent Component Analysis and Signal Separation*, pp.354–361, Springer, 2009.

[20] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol.142, no.1-2, pp.397–434, 2013.

[21] D. Niu, J. G. Dy, and M. I. Jordan, "Iterative discovery of multiple alternativeclustering views," *IEEE transactions on pattern analysis and machine intelligence*, vol.36, no.7, pp.1340–1353, 2014.

[22] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol.17, no.4, pp.395–416, 2007.

[23] J. Nocedal and S. Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

[24] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *arXiv preprint arXiv:1202.3775*, 2012.

[25] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smooth l_0 norm," *IEEE Transactions on Signal Processing*, vol.57, no.1, pp.289–301, 2008.