دانشکده علوم پایه

گروه ریاضی کاربردی و علوم کامپیوتر

پایان‌نامه

کارشناسی ارشد در رشته

ریاضی کاربردی، گرایش آنالیز عددی

عنوان

به‌روزسازی مؤلفه‌های اصلی هسته

استاد راهنما

دکتر مرتضی رحمانی

استاد مشاور

دکتر بهنام زرپاک

نگارش

نسیم خسروی

شهریور ۱۳۸۹

بسم الله الرحمن الرحيم

تقدیم به

آیه سبز مصحف عشق

یا ایّها العزیز مسّنا و اهلنا الضر و جئنا ببضاعة مزجاة فاوف لنا الکیل و تصدق علینا ان الله یجزی المتصدقین

(سوره یوسف آیه ۸۸)

سپاس

پروردگارا

تو را سپاس می گویم که عظمت و قدرت خویش را چنان بر من نموده ای که جز از تو استعانت نخواهم و برای سگرگزاری از تو، خود را رهین منت همه آن هایی بدانم که بزرگوارانه در سامان یافتن این رساله یاریم کردند.

ابتدا دوست دارم از آقای دکتر مرتضی رحمانی به خاطر رهنمودهایشان طی دو سال اخیر تشکر کنم.

استاد گرامی، راهنمایی ها و حمایت های بی دریغ شما همراه با اخلاق پسندیده در بسیاری از زمان های انجام این پژوهش همواره موجب دلگرمی بوده و صمیمانه از شما سپاس گذاری می کنم.

از راهنمایی های ارزنده استاد عالیقدر جناب آقای دکتر بهنام زرپاک که مشاورت پایان نامه را بر عهده داشتند، قدردانی می نمایم.

از اعضای محترم داوران جناب آقای دکتر سیّد حمید حاجی سیّد جوادی و دکتر علی محمد نظری که زحمت مطالعه و داوری پایان نامه را بر عهده داشتند، نهایت تشکر را دارم.

از پدر و مادرم که صبورانه در لحظه های امیدواری و ناامیدی در کنارم بودند و پیوسته یاریم کردند سپاسگزارم.

امیدوارم روزگار مرا به جبران زحمت هایی که برای یکایک عزیزان فراهم آورده ام، توفیق دهد.

چکیده

تحلیل مؤلفه‌های اصلی (PCA)، یک روش مشهور برای تشخیص نمونه، تحلیل داده و کاهش بعد (سیستم) است. این روش یک تبدیل متعامد است که جهت‌های دارای بیشترین واریانس در داده را نشان می‌دهد.
ماتریس تبدیل، دورانی است که محورهای مختصات را به جهت مؤلفه‌های اصلی، مرتب می‌کند. روش آن، بدست‌آوردن ماتریس پایه‌هایی است که ستون‌های آن ماتریس، مؤلفه‌های اصلی هستند.
در این پایان‌نامه، ابتدا PCA و روش‌های هسته‌ای برای PCA غیرخطی ($KPCA$) مطالعه شده است.
به منظور غلبه بر محدودیت‌های $KPCA$، روش محاسباتی $IKPCA$، $KPCA$ افزایشی استفاده شده است.
روش آن کاهش پیچیدگی محاسباتی $KPCA$ و انجام $KPCA$ با ظرفیت به روزرسانی داده‌های قبلی با داده‌های موردنظر است. انجام $IKPCA$ در فضای القایی هسته و به‌کارگیری مجموعه کاهش‌یافته برای ثابت نگه داشتن سرعت به‌روزرسانی مورد مطالعه قرار می‌گیرد. سپس روی SVD افزایشی و ارتباط آن با $IKPCA$ متمرکز می‌شویم.
در ادامه، یک الگوریتم برای بهبودبخشیدن $IKPCA$ به منظور بدست‌آوردن خطای نسبی کمتر بین ماتریس واقعی در فرایند روش $IKPCA$ و ماتریس تقریب حاصل از ضرب ماتریس‌های به روزشده‌ی فرایند PCA و SVD پیشنهاد شده است.
در پایان، الگوریتم پیشنهادی برای بدست‌آوردن نتایج عددی برای ۲ هسته آزمایش کردیم:هسته همانی و هسته گاؤسی بررسی شده است.

واژگان کلیدی: تحلیل مؤلفه‌های اصلی هسته به روش افزایشی، تجزیه مقادیرتکین، بسط مجموعه کاهش‌یافته، کاهش پیچیدگی زمانی

مقدمه

با گسترش روزافزون و همه جانبه عصر ارتباطات و تلاش برای انتقال اطّلاعات در کوتاه‌ترین زمان ممکن، نیاز به یافتن ویژگی‌های فشرده‌سازی داده‌ها، غیر قابل انکار است. متأسفانه واسطه‌های ارتباطی بسیاری از ویژگی‌های موضوعات را حین تبدیل ماشینی آن‌ها به داده‌های عددی از دست می‌دهند. به همین دلیل تحقیقات زیادی در زمینه‌های مرتبط، برای یافتن الگوهایی در بین این حجم عظیم از داده‌ها متمرکز شده است تا با استخراج ویژگی‌های مختص یک موضوع و فشرده‌سازی و پارامتری‌کردن آن‌ها، تحلیل و طبقه‌بندی این موضوعات به آسانی انجام پذیرد. از آن‌جایی‌که به دلیل محدودیت‌های فیزیکی و زمانی به دنبال توصیف کوتاه از موضوعات و مسائل پیچیده هستیم، لذا ضرورت بحث حاضر و پژوهش در آن از جایگاه ویژه‌ای برخوردار است.

داده‌ها زمانی می‌توانند مفید واقع شوند که پردازش شده باشند. در قبال این گسترش سریع حجم داده‌ها، دیدگاه‌های مختلفی برای پردازش آن‌ها وجود دارد: به‌کارگیری روش‌های کلاسیک، طراحی ساختارهای به مراتب قوی محاسباتی مانند محاسبات غیرمتمرکز، پردازش‌های چند هسته‌ای، ابر کامپیوترها و غیره. با توجه به رشد حجم و پیچیدگی داده‌های قابل ذخیره نسبت به رشد قدرت محاسبات - تقریباً معادل دو برابر - لزوم تحقیق و پژوهش در حوزه دست‌یابی به روش‌های کاراتر ولی ساده را محرز می‌نماید.

برای مثال به‌کارگیری زیرفضاهای غالب به‌وسیله تجزیه مقادیرتکین(منفرد، ویژه)[۱] (SVD) [۱۹] به عنوان بهترین مدل برای کاهش پچیدگی داده‌ها و سیستم‌های پیچیده پیشنهاد شده است که در مقایسه با روش‌های مشابه با همان پیچیدگی کمترین خطا را (نسبت به برخی اندازه‌ها) داراست. اما PCA علیرغم ارتباط تنگاتنگی که باSVDدارد، واجد ویژگی‌هایی است که تکنیک‌های مربوط به SVD چنین نیستند. در فصل ۲ به این ویژگی‌ها اشاره خواهد شد. مدل PCA و یا در حالت غیرخطی $KPCA$ بردارهای اصلی یک سیستم،به منظور پردازش، کدنگاری و فشرده‌سازی مشخص می‌شوند.

ارتباط نزدیک و برتری این روش نسبت به تجزیه مقادیر تکین از اصلی‌ترین دلایلی است که توجه متخصصین امر را به منظور به‌کارگیری و ارتقاء این روش سوق داده است. همچنین بدلیل کاربردهای فراوان این روش در انواع سیستم‌های خطی و غیرخطی و کارایی بالایی که برای علوم مختلف دارد، به‌روزرسانی این روش مورد از توجه خاصی برخوردار شده است.

دراین پایان‌نامه چگونگی به‌روزرسانی $KPCA$[۲] به کمک به‌روزرسانی (SVD) مورد بررسی قرار می‌گیرد. این کار با کاهش پیچیدگی الگوریتم $IKPCA$ همراه است.

روش افزایشی به‌روزسازی مؤلفه‌های اصلی ماتریس هسته‌ی داده‌ها[۳] این امکان را فراهم می‌کند که بدون از سرگیری محاسبات پیشین، با استفاده ازاطّلاعات داده‌های اولیه موجود ، تنها با به‌کارگیری داده‌های جدید بردارهای اصلی کل سیستم، بدست آیند.

---

۱ $Singular\ Value\ Decomposition$

۲ $Kernel\ Principal\ Component\ Analysis$

۳ $Incremental\ Kernel\ Principal\ Component\ Analysis$

مسأله تورم و تأثیر آن روی سیاست‌گذاری‌های کلان کشور بسیار حائز اهمیت است. درصد تغییر سالیانه و یا ماهیانه‌ی بهای کالاها و خدمات مصرفی هرکدام سهم خاصی در تعیین شاخص‌های نرخ تورم دارد. تعیین سهم تاثیر گذاری گروه‌های اصلی کالا و خدمات مصرفی می‌تواند در مدیریت، برنامه ریزی و همچنین هزینه‌های تعیین شاخص تورم تأثیرگذار باشد. در این مجموعه، به منظور بدست‌آوردن گروه‌های اصلی شاخص بها در هر دوره، و تعیین تأثیرات تورمی کالاها و خدمات در شاخص کل، از روش $IKPCA$ بهره می‌بریم.

فهرست مطالب

لیست جداول

لیست تصاویر

فصل ۱

یادآوری

۱.۱ تعاریف و قضایای مورد نیاز

در این فصل به بیان تعاریف و مفاهیم اصلی مورد نیاز در پایان‌نامه می‌پردازیم. غالب قضایا و نتایج بدون اثبات بیان شده است.

۱.۱.۱ مفاهیم موردنیاز ماتریس و بردار

تذکر:ماتریس حقیقی $m \times n$ یک جدول m سطری و n ستونی از اعداد حقیقی است. در صورت تساوی تعداد ستون‌ها و سطرهای یک ماتریس به آن ماتریس مربعی اطلاق می‌شود. مجموعه تمام ماتریس‌های حقیقی $m \times n$ با $\mathbb{R}^{m\times n}$ نشان داده می‌شود. در این پایان‌نامه تمام ماتریس‌ها حقیقی هستند و از حروف بزرگ برای نشان دادن آن‌ها استفاده می‌شود.

تعریف ۱.۱. رتبه ماتریس A از مرتبه $m \times n$ بعد زیرفضای تولید شده توسط ستون‌های A تعریف می‌شود:

$$rank(A) = dim(span(A_{:1}, A_{:2}, \ldots, A_{:n})) \leq min(m, n).$$

و یا به عبارتی ماکزیمم تعداد سطرهای مستقل خطی(ماکزیمم تعداد ستون‌های مستقل خطی) رتبه ماتریس[1]نامیده می‌شود.

که در آن A_iها ستون‌های ماتریس A نامیده می‌شوند.

[1] $Matrix\ Rank$

تعریف ۲.۱. ماتریس A نامنفرد[۲] می‌باشد هرگاه ماتریسی مانند B یافت شود به‌طوری‌که

$$AB = BA = I$$

تعریف ۳.۱. هرگاه ماتریس $A_{n\times n}$ نامنفرد نباشد، منفرد[۳]نامیده می‌شود.

تعریف ۴.۱. ماتریس مربعی A، معکوس‌پذیر نامیده می‌شود هرگاه ماتریسی مانند B چنان موجود باشد که:

B معکوس ماتریس A نامیده می‌شود و با $B = A^{-۱}$ نشان می‌دهیم. از آن جایی که تمامی ماتریس‌ها معکوس‌پذیر نیستند، معکوس تعمیم‌یافته (یا معکوس تعمیم‌یافته مور - پنروز)[۴] تعمیمی از آن است که حتی برای ماتریس‌های غیر مربعی نیز تعریف می‌شود. معکوس تعمیم‌یافته منحصر به‌فردA^+ برای ماتریس A دارای چهار شرط زیر است:

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)^T = AA^+ \text{و} \quad (A^+A)^T = A^+A.$$

به‌ویژه اگر A^TA معکوس‌پذیر باشد آن‌گاه $A^+ = (A^TA)^{-۱}A^T$.

تعریف ۵.۱. ماتریس متقارنA معین‌مثبت[۵] است هرگاه: $x^TAx > ۰ \qquad \forall x \neq ۰$
و A نیمه‌معین‌مثبت[۶] است هرگاه: $x^TAx \geq ۰ \qquad \forall x$

تعریف ۶.۱. برای هر i, j *ضرب‌داخلی* دو بردار حقیقی $x, y \in \mathbb{R}^n$ به‌صورت تابع حقیقی زیر تعریف می‌شود:

$$\langle x, y\rangle = \sum_{i=۱}^{n} x_i y_i = x^T y.$$

تعریف ۷.۱. بردارهای ناصفر $x, y \in \mathbb{R}^n$ متعامد نامیده می‌شوند هرگاه حاصل‌ضرب‌داخلی آن‌ها برابر صفر باشد:

$$\langle x, y\rangle = ۰$$

.

[۲] $Nonsingular$
[۳] $Singular$
[۴] $Moore\ Penerose$
[۵] $Positive\ Definite$
[۶] $Positives\ Semi\ definite$

تعریف ۱.۸. حاصل‌ضرب‌داخلی دو ماتریس حقیقی هم‌مرتبه به‌صورت زیر تعریف می‌شود:

$$\langle A,B\rangle = vec(A)^T vec(B) = \sum_{ij} A_{ij}B_{ij} = trace(A^T B)$$

که در آن منظور از $trace(A)$ مجموع عناصر قطری ماتریس A است.

۲.۱.۱ مقدارهای ویژه و بردارهای ویژه

تعریف ۱.۹. مقدارهای ویژه و بردارهای ویژه: یکی از مباحث اصلی در جبر خطی و آنالیز ماتریس، مقدارهای ویژه و بردارهای ویژه ماتریس‌های مربعی هستند که اطّلاعات اساسی درباره ماتریس در اختیار ما قرار می‌دهند. مفهوم مرتبط برای ماتریس‌های مستطیلی مقدارهای تکین و بردارهای تکین نامیده می‌شوند که نقش تعیین کننده‌ای در تقریب رتبه پایین ماتریس بازی می‌کنند و مشخصات مهم ماتریس اصلی را معین می‌کنند.

تعریف ۱.۱۰. اسکالر $\lambda \in \mathbb{C}$ مقدار ویژه ماتریس $A \in \mathbb{C}^{n\times n}$ است، هرگاه بردار ناصفر $x \in \mathbb{C}^n$ موجود باشد به‌طوری که $Ax = \lambda x$. بردار x بردار ویژه متناظر به مقدار ویژه λ نامیده می‌شود.

یک ماتریس مربعی از مرتبه $n\times n$ با احتساب تعداد تکرار، دارای n مقدار ویژه است. مجموعه تمام مقدارهای ویژه با $\sigma(A)$ نشان داده می‌شود. شعاع طیفی A بیشینه مقدار قدر مطلق $\sigma(A)$ است و با $\rho(A)$ نشان داده می‌شود:

$$\rho(A) = max\{|\lambda|;\ \lambda \in \sigma(A)\}.$$

یک ابزار بسیار مفید در آنالیز ماتریس، *تجزیه مقدار ویژه* (*مقدارتکین*) است که در قضیه بعدی تعریف می‌شود:

قضیه ۱.۱. *برای هر ماتریس $A \in \mathbb{R}^{m\times n}$، ماتریس‌های متعامد $U \in \mathbb{R}^{m\times m}$ و $V \in \mathbb{R}^{n\times n}$ چنان وجود دارند که*

$$A = U\Sigma V^T$$

و

$$\Sigma = \begin{pmatrix} \sigma_1 & & 0 & \\ & \ddots & & O_{r\times(n-r)} \\ 0 & & \sigma_r & \\ & O_{(m-r)\times r} & & O_{(m-r)\times(n-r)} \end{pmatrix}$$

که درآن مقدارهای تکین σ_i اسکالرهای حقیقی هستند و

$$\sigma_1 \geq \ldots \sigma_r > 0.$$

ستون‌های U و V به ترتیب بردهای ویژه A^TA و AA^T هستند.[۱۸] *برای اثبات وجود و الگوریتم‌های مربوط به تجزیه مقدارهای تکین به مرجع*[۱۹] *رجوع شود.*

ویژگی‌های تجزیه مقدار ویژه: فرض کنید $A = U\Sigma V^T$ در آن صورت:

$$\begin{aligned} A^T = V\Sigma U^T &\Rightarrow A^TA = V\Sigma U^T U\Sigma V^T = V\Sigma^۲ V^T \\ &\Rightarrow A^TAV = V\Sigma^۲ V^T V = V\Sigma^۲ \end{aligned}$$

از طرف دیگر داریم:

$$\begin{aligned} AA^T &= U\Sigma V^T V\Sigma U^T = U\Sigma^۲ U^T \\ AA^TU &= U\Sigma^۲ U^T U = U\Sigma^۲ \end{aligned}$$

اگر $\mathcal{N}(A)$ نمایشگر فضای پوچ A باشد؛ آنگاه $\mathcal{N}(A^TA) \subseteq \mathcal{N}(A)$ زیرا

$$\begin{aligned} & x \in \mathcal{N}(A) : Ax = ۰ \Rightarrow A^TAx = ۰ \Rightarrow x \in \mathcal{N}(A^TA) \Rightarrow \mathcal{N}(A) \subseteq \mathcal{N}(A^TA) \\ & x \in \mathcal{N}(A^TA) : A^TAx = ۰ \Rightarrow < x, A^TAx >= ۰ \Rightarrow \|Ax\|^۲ = ۰ \Rightarrow \\ & Ax = ۰ : x \in \mathcal{N}(A) \Rightarrow \mathcal{N}(A^TA) \subseteq \mathcal{N}(A) \end{aligned}$$

نتیجه:

- ستون‌های V بردار ویژه A^TA می‌باشند.
- ستون‌های U بردار ویژه AA^T می‌باشند.
- مربع مقادیرتکین، مقدار ویژه A^TA و AA^T می‌باشند.
- فضای پوچ A=فضای پوچ A^TA.

تعریف ۱۱.۱. برای ماتریس حقیقی و متقارن A، اینرسی[۷] با سه تایی (ρ, υ, ζ) تعریف می‌شود که ρ و υ و ζ به ترتیب تعداد مقادیر ویژه‌های مثبت، منفی و صفر ماتریس A می‌باشند[۱۸].

قضیه ۲.۱. *قانون سیلوستر از اینرسی*[۸]
فرض کنید A ماتریس مربعی، متقارن از مرتبه n با درایه‌های حقیقی و S ماتریس نامنفرد باشد به‌طوری‌که ماتریس

[۷] Inertia
[۸] Sylvesters law of Inertia

D با رابطه $D = SAS^T$ یک ماتریس قطری باشد، درآن صورت ماتریس A و D اینرسیای یکسان دارند(ماتریس قطری D تنها درایه‌های $\{۱, ۰, -۱\}$ حول قطر دارد. قانون فوق بیان می‌نماید برای هر ماتریس وارون‌پذیر S، تعداد عناصر مثبت، منفی در ماتریس A با تعداد عناصر مثبت، منفی در ماتریس D یکسان می‌باشند).

۳.۱.۱ نرم‌ها

نرم اسکالری است که به یک بردار یا یک ماتریس نسبت داده می‌شود. یک نرم برداری روی $\mathbb{R}^n$ تابعی حقیقی و خطی مانند $\|.\|$ روی $\mathbb{R}^n$ با خواص زیر است:

$$\begin{aligned}
&\|x\| \geq ۰, \forall x \in \mathbb{R}^n\\
&\|x\| = ۰ \iff x = ۰;\\
&\|\alpha x\| = |\alpha|\|x\|, \forall x \in \mathbb{R}^n \text{ و } \forall \alpha \in \mathbb{R};\\
&\|x+y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n
\end{aligned}$$

به‌همین ترتیب یک نرم ماتریسی مانند $\|.\|$ تابعی حقیقی و خطی روی $\mathbb{R}^{m\times n}$ با خواص زیر می‌باشد:

$$\begin{aligned}
&\|X\| \geq ۰,\ \forall X \in \mathbb{R}^{m\times n};\\
&\|X\| = ۰ \iff X = ۰;\\
&\|\alpha X\| = |\alpha|\|X\|,\ \forall X \in \mathbb{R}^{m\times n} \text{ و } \forall \alpha \in \mathbb{R};\\
&\|X+Y\| \leq \|X\| + \|Y\|,\ \forall X, Y \in \mathbb{R}^{m\times n},\\
&\|XY\| \leq \|X\|\|Y\|,\ \forall X \in \mathbb{R}^{m\times n} \text{ و } \forall Y \in \mathbb{R}^{n\times k}.
\end{aligned}$$

در بسیاری از کاربردها ماتریس‌ها وبردارها توأم با یکدیگر مطرح می‌شوند، به‌همین دلیل خواص نرم ماتریسی همراه با یک شرط کمکی می‌تواند برای تعریف نرم وابسته به نرم برداری تعریف می‌شود. چنین نرمی را نرم ماتریسی سازگار می‌نامیم که به‌صورت زیر تعریف می‌شود:

$$|A| = \underset{|x|>۰}{sup}\frac{|Ax|}{|x|} = \underset{|x|=۱}{sup}|Ax|$$

شرط فوق معادل شرط زیر است:

$$\|Ax\| \leq \|A\||x|.$$

یک نرم متداول برداری یا ماتریسی نرم اقلیدسی یا نرم فربنیوس[۹] است که از ضرب‌داخلی حاصل می‌شود:

$$\|x\|_F = \sqrt{\langle x, x\rangle},$$

که در آن x یک بردار یا ماتریس است. این نرم نقش محوری در مسائل کمترین مربعات بازی می‌کند و به عنوان معیار کاهش خطا در حل این مسائل به‌کار می‌رود.
نرم‌های برداری که بیشتر استفاده می‌شوند مثال‌هایی از نرم معروف هولدر[۱۰] (p - نرم) هستند:

$$\|x\|_p = \Big(\sum_{i=۱}^{n} |x_i|^p\Big)^{۱/p} \qquad , p = ۱, ۲, \ldots$$

این نرم بازای $p = ۱$ ، $p = ۲$ و $p = \infty$ از کاربرد بیشتری برخوردار است:

$$\begin{aligned}
&\|x\|_۱ = |x_۱| + |x_۲| + \ldots + |x_n|,\\
&\|x\|_۲ = \sqrt{|x_۱|^۲ + |x_۲|^۲ + \ldots + |x_n|^۲},\\
&\|x\|_\infty = max_i |x_i|..
\end{aligned}$$

برای بردارها، نرم ۲ همان نرم فربنیوس است. اما p - نرم ماتریسی به‌صورت نرم تبعی از p- نرم برداری حاصل می‌شود:

$$\|A\|_p = \max_{x \neq ۰} \frac{\|Ax\|_p}{\|x\|_p}.$$

به‌راحتی می‌توان اثبات نمود که:

$$\begin{aligned}
&\|A\|_۱ = max_j \sum_i |A_{ij}|\\
&\|A\|_۲ = [\rho(A^T A)]^{۱/۲}\\
&\|A\|_\infty = max_i \sum_j |A_{ij}|.
\end{aligned}$$

تعریف ۱.۱۲. PCA[۱۱] یک تبدیل خطی متعامد است که با انتخاب محورها و یا (عواملی) که بیشترین اهمیت را در سیستم دارند، عوامل اضافی را حذف می‌کند و بعد مسأله را کاهش می‌دهد.

[۹] $Frobenius$
[۱۰] $Holder$
[۱۱] $Principal\ Component\ Analysis$

تعریف ۱.۱۳. ماتریس تباهیده: ماتریس مربعی که دترمینان آن صفر است.

تعریف ۱.۱۴. فاصله ماهالانوبیس [۱۲]: فاصله‌ای برای تعیین میزان شباهت یک نمونه از داده‌ها به یک مجموعه‌ی شناخته‌شده، که به ماتریس کوواریانس داده‌ها وابسته می‌باشد:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}$$
$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$$
$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i-y_i)^2}{\sigma_i^2}}$$

که در این روابط ، ماتریس کوواریانس داده‌ها با S نشان داده شده است.

تعریف ۱.۱۵. کد: تابعی که به هر عنصر از دنباله‌ی هدف یک رشته دودویی را اختصاص می‌دهد. با این فرض که این نمایش منحصراً قابل رمزگشایی است.

تعریف ۱.۱۶. آشفتگی[۱۳](عدم قطعیت): کمیتی که شامل محتوای اطّلاعات می‌باشد و حدود طول رمز بهینه را مشخص می‌کند.

تعریف ۱.۱۷. حشو (زیادی)[۱۴] : یکی از دو نوع آلودگی داده‌ها است که به کمیت کوواریانس داده‌ها مرتبط و شامل داده‌های اضافی سیستم می‌شود.

تعریف ۱.۱۸. اختلال [۱۵]: نوع دیگری از آلودگی داده‌ها که یک پدیده ذاتی در هر سیستم الکترونیکی است. به‌ویژه اگر داده‌های ورودی بسیار کوچک باشند. يك اختلال يك مجموعه از اغتشاشات داده است که باید پایین (کم) باشد یا بیشتر از روش تحلیل سیستم و اطّلاعات سیستم نباشد. يك معیار سنجش اختلال سیستم، اندازه زیر است که مقایسه این کسر با عدد ۱ میزان دقت داده‌ها را نشان می‌دهد:

$$SNR = \frac{\sigma_{signal}^2}{\sigma_{noise}^2}$$
$$SNR >> 1 \qquad \equiv \qquad noise \uparrow$$
$$SNR << 1 \qquad \equiv \qquad noise \downarrow$$

[۱۲] $Mahalanobis$
[۱۳] $Antropy$
[۱۴] $Redundancy$
[۱۵] $Noise$

تعریف ۱.۱۹. واریانس[۱۶]: معیاری برای سنجش میزان پراکندگی داده‌ها از میانگین آن‌ها است و در کاهش آلودگی نوع اول (اختلال) داده‌ها به کار می‌رود.

تعریف ۱.۲۰. واریانس تعدادی داده توسط رابطه زیر محاسبه می‌شود:

$$\sigma^{2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^{2}}{n-1}$$

تعریف ۱.۲۱. کوواریانس[۱۷]: معیاری برای سنجش میزان پراکندگی داده‌ها از میانگین آن‌ها نسبت به یکدیگر است و در کاهش آلودگی نوع دوم داده‌ها (حشو) به کار می‌رود.

تعریف ۱.۲۲. ماتریس کوواریانس یک ماتریس:فرض کنید X یک ماتریس از اندازه$m \times n$باشد

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

آن‌گاه ماتریس کوواریانس آن به این شکل محاسبه می‌شود$S_X = \frac{1}{n-1}XX^T$

$$COV(x,y) = \begin{pmatrix} cov(x_1,y_1) & \dots & \dots \\ cov(x_2,y_1) & \dots & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

که در آن کوواریانس بین هر دو بردار یک ماتریس به شکل زیر محاسبه می‌شود:

$$cov(x_i,y_i) = \frac{\sum_{j=1}^{m}(x_{i_j} - \bar{x_i})(y_{i_j} - \bar{y_i})}{m-1}$$

تعریف ۱.۲۳. تجزیه [۱۸]: دقت و وضوح بالا در تجزیه تصویر بدلیل تقسیمات بسیار ریز تصویر.

تعریف ۱.۲۴. سیگنال[۱۹]: درلغت به معنای علامت و درعلوم ریاضی و مهندسی به معنی هرگونه تغییر وارد شده به محیط است.

[۱۶] $Variance$
[۱۷] $Covariance$
[۱۸] $High\ resulotion$
[۱۹] $Signal$