Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2016/2017

Lesson 10 15 March 2017

The Normal Density Function

Bibliographical sources:
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	A variable
X	A random variable (unpredictable value). an observation.
M	The number of possible values for X
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
C_k	The class k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $X \in C_k$
$P(\omega_k) = P(X \in C_k)$	Probability that the observation X is a member of the class k.
M_k	Number of examples for the class k.
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{\vec{X}_m\}$	A set of training samples
$\{y_m\}$	A set of indicator vectors for the training samples in $\{\vec{X}_m\}$
$p(X)$	Probability density function for a continuous value X
$p(\vec{X})$	Probability density function for continuous $\vec{X}$
$p(\vec{X} \mid \omega_k)$	Probability density for $\vec{X}$ give the class k. $\omega_k = X \in C_k$.

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

Bayesian Classification

Our problem is to build a box that maps a set of features $\vec{X}$ from an observation, X to a class C_k from a set of K possible classes.

Let ω_k be the proposition that the event belongs to class k: $\omega_k = \vec{X} \in C_k$

In order to minimize the number of mistakes, we will maximize the probability that $\omega_k \equiv X \in C_k$

$$\hat{\omega}_k = \arg\max_{\omega_k} \left\{ P(\omega_k \mid \vec{X}) \right\}$$

Our primary tool for this is Baye's Rule : $P(\omega_k \mid \vec{X}) = \frac{P(\vec{X} \mid \omega_k) P(\omega_k)}{P(\vec{X})} = \frac{P(\vec{X} \mid \omega_k)}{\sum_{k=1}^{K} P(\vec{X} \mid \omega_k)} P(\omega_k)$

To apply Baye's rule, we require a representation for the probabilities $P(\vec{X} \mid \omega_k)$, $P(\vec{X})$, and $p(\omega_k)$. Today we will discus use of the Normal Density function.

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

and

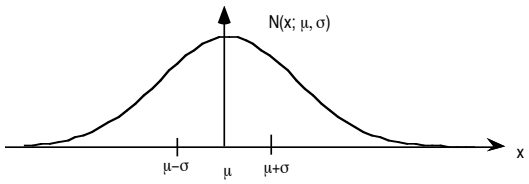
$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

The Normal Density Function

Univariate Normal

Whenever a random variable is determined by a sequence of independent random events, the outcome will be a Normal or Gaussian density function. This is known as the Central Limit Theorem. The essence of the derivation is that repeated random events are modeled as repeated convolutions of density functions, and for any finite density function will tend asymptotically to a Gaussian (or normal) function.

$$p(X) = \mathcal{N}(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



The parameters of $\mathcal{N}(x;\ \mu,\ \sigma)$ are the first and second moments, μ and σ of the function.

According to the Central Limit theorem, for any real density $p(X)$:

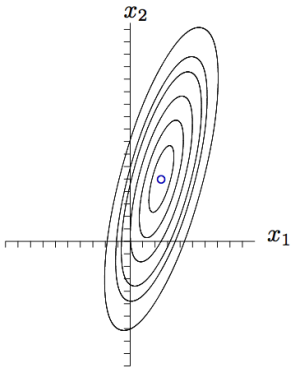
$$as\ M \to \infty \quad p(X)^{*M} \to \mathcal{N}(x;\ \mu,\ \sigma)$$

Multivariate Normal

For a vector of D features, $\vec{X}$, the Normal density has the form:

$$p(\vec{X}) = \mathcal{N}(\vec{X};\vec{\mu},\ \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\ \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T\ \Sigma^{-1}(\vec{X}-\vec{\mu})}$$

The result can be visualized as a set of ellipses representing contours of probability.



Ellipses for 99%, 95%, 90%, 75%, 50%, and 20% of the mass

There are 3 parts to $\mathcal{N}(\vec{X};\vec{\mu},\Sigma)$:

$$\frac{1}{(2\pi)^{\frac{D}{2}}\det(\Sigma)^{\frac{1}{2}}}, \quad e, \quad \text{and} \quad d^2 = -\frac{1}{2}\left(\vec{X}-\vec{\mu}\right)^T \Sigma^2 \left(\vec{X}-\vec{\mu}\right)$$

1) $e = 2.7818281828\ldots$ Euler's Constant : $\int e^x dx = e^x$. Used to simplify the algebra.

2) The term $(2\pi)^{\frac{D}{2}}\det(\Sigma)^{\frac{1}{2}}$ is a normalization factor.

$$(2\pi)^{\frac{D}{2}}\det(\Sigma)^{\frac{1}{2}} = \int\int\ldots\int e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1}(\vec{X}-\vec{\mu})} dx_1 dx_2 \ldots dx_D$$

The determinant, det(**Σ**) is an operation that gives the volume of **Σ**.

for D=2 $\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = a\cdot d - b\cdot c$

for D > 2 this continues recursively.

ex: D=3 $\det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} + b\begin{vmatrix} f & d \\ i & g \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$

$$a(e\cdot i - f\cdot h) + b(f\cdot g - d\cdot i) + c(d\cdot h - e\cdot g)$$

3) The Mahalanobis distance. This is where the action is.

The Mahalanobis Distance

The exponent of $\mathcal{N}(\vec{X};\vec{\mu}, \Sigma)$ is known as the "Mahalanobis Distance" named for a famous Indian statistician.

$$d^2 = -\frac{1}{2}\left(\vec{X}-\vec{\mu}\right)^T \Sigma^2 \left(\vec{X}-\vec{\mu}\right)$$

This is a distance normalized by the covariance.
It is positive and 2nd order (quadratic).

The covariance is provides a distance metric that can be used when the features of $\vec{X}$ have different units (for example with height (cm), weight (kg) and age (y). In this case, the features are compared to the standard deviation of a population. The unit of distance can be said to be the Standard deviation (std).

Expected Values and Moments

The average value is the first moment of the samples

A training set of M samples $\{\vec{X}_m\}$ an be used to calculate moments

For M samples of a numerical feature value $\{X_m\}$, the "expected value" E{X} is defined as the average or the mean:

$$E\{X\} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

$\mu_x = E\{X\}$ is the first moment (or center of gravity) of the values of $\{X_m\}$.
$\mu_x = E\{X\}$ is also the first moment (or center of gravity) of the resulting pdf.

$$E\{X\} = \frac{1}{M}\sum_{m=1}^{M} X_m = \int_{-\infty}^{\infty} p(x)\cdot x\, dx$$

This is also true for histograms.
The mass of the histogram is M $\quad M = \sum_{x=1}^{N} h(x)$

Which is also the number of samples used to compute h(x).

The expected value of $\{X_m\}$ is the 1st moment of h(x).

$$\mu = E\{X_m\} = \frac{1}{M}\sum_{m=1}^{M} X_m = \frac{1}{M}\sum_{x=1}^{N} h(x)\cdot x$$

For D dimensions, the center of gravity is a vector

$$\vec{\mu} = E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M} \vec{X} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{pmatrix} = \begin{pmatrix} E\{X_1\} \\ E\{X_2\} \\ \dots \\ E\{X_D\} \end{pmatrix}$$

The vector $\vec{\mu}$ the vector of averages for the components, X_d of $\vec{X}$.
It is also center of gravity (first moment) of the normal density function.

$$\mu_d = E\{X_{d,m}\} = \frac{1}{M}\sum_{m=1}^{M} X_{d,m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} p(x_1, x_2, \dots x_D)\cdot x_d \, dx_1, dx_2, \dots, dx_D$$

$$\vec{\mu} = E\{\vec{X}\} = \begin{pmatrix} E\{X_1\} \\ E\{X_2\} \\ \dots \\ E\{X_D\} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{pmatrix}$$

The variance σ^2 is the square of the expected deviation from the average

$$\sigma^2 = E\{(X-\mu)^2\} = E\{X^2\} - \mu^2 = E\{X^2\} - E\{X\}^2$$

This is also the second moment of the pdf

$$\sigma^2 = E\{(X-\mu)^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu)^2 = \int_{-\infty}^{\infty} p(x)\cdot(x-\mu\ )^2 \, dx$$

and the second moment of a histogram for discrete features.

$$\sigma^2 = E\{(X - E\{X\})^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu)^2 = \frac{1}{M}\sum_{x=1}^{N} h(x)\cdot(x-\mu)^2$$

For D dimensions, the second moment is a co-variance matrix composed of D^2 terms:

$$\sigma_{ij}^2 = E\{(X_i - \mu_i)(X_j - \mu_j)\} = \frac{1}{M}\sum_{m=1}^{M}(X_{i,m} - \mu_i)(X_{j,m} - \mu_j)$$

This is can be written as $\Sigma = E\{(\vec{X} - E\{\vec{X}\})(\vec{X} - E\{\vec{X}\})^T\}$

and gives $$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \ldots & \sigma_{2D}^2 \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{D1}^2 & \sigma_{D2}^2 & \ldots & \sigma_{DD}^2 \end{pmatrix}$$

If x_i and x_j are statistically independent, then $\sigma_{ij}^2 = 0$

For positive values of σ_{ij}^2, x_i and x_j vary together.
For negative values of σ_{ij}^2, x_i and x_j vary in opposite directions.

For example, consider features x_1 = height *(meters)* and x_2 = weight *(kg)*

In most people height and weight vary together and so σ_{12}^2 would be positive

This provides the parameters for $$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

Quadratic Discrimination

The classification function can be decomposed into two parts: d() and $g_k()$:

$$\hat{\omega}_k = d\left(g_k\left(\vec{X}\right)\right)$$

$g(\vec{X})$: A discriminant function : $R^D \rightarrow R^K$
d() : a decision function $R^K \rightarrow \{\omega_K\}$

The discriminant is a vector of functions:

$$\vec{g}(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \vdots \\ g_K(\vec{X}) \end{pmatrix}$$

Quadratic discrimination functions can be derived directly from $p(\omega_k \mid \vec{X})$

$$P(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k)P(\omega_k)}{p(\vec{X})}$$

To minimize the number of errors, we will choose k such that

$$\hat{\omega}_k = \underset{\omega_k}{\arg\max}\{\frac{p(\vec{X} \mid \omega_k)P(\omega_k)}{p(\vec{X})}\}$$

but because $p(\vec{X})$ is identical for all k, it can be eliminated.

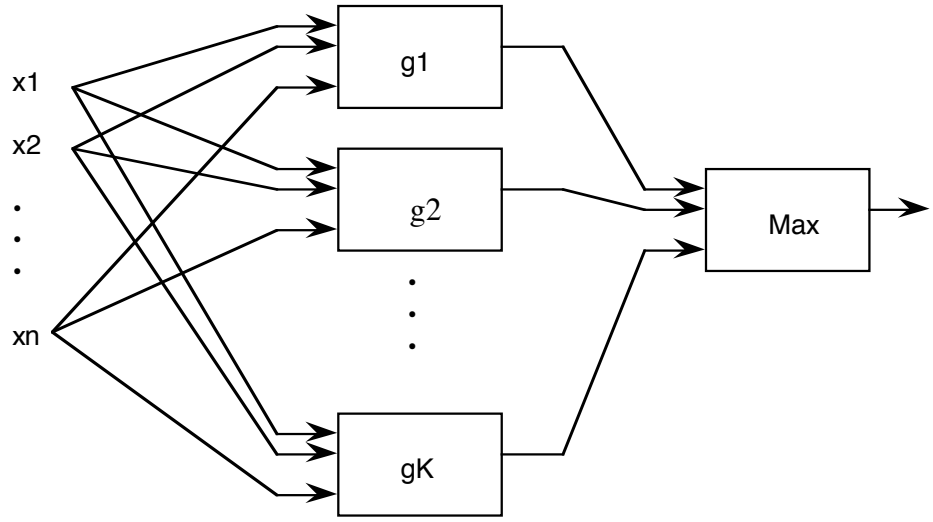
$$\hat{\omega}_k = \underset{\omega_k}{\arg\max}\{P(\vec{X} \mid \omega_k)p(\omega_k)\}$$

This is called a "Maximum Likelihood" classifier.

Warning: Maximum likelihood can sometimes give a highly improbable answer. Remember that confidence in the choice $\hat{\omega}_k$ is provided by the full probability:

$$CF_{\hat{\omega}_k} = p(\hat{\omega}_k \mid \vec{X}) = \frac{P(\vec{X} \mid \hat{\omega}_k)p(\hat{\omega}_k)}{P(\vec{X})}$$

The maximum likelihood classifier can be organized as a set of parallel discriminant functions. This gives a sort of parallel machine that resembles:



The functions $g_k(X)$ are commonly constructed by from the Log of the likelihood:

$$g_k(X) = Log\{p(\vec{X} \mid \omega_k)P(\omega_k)\}$$

This gives a simple function in the case where the pdf is a multivariate norm:

$$P(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

or even to a Gaussian Mixture Model

$$P(\vec{X} \mid \omega_k) = \sum_{n=1}^{M} \alpha_n \mathcal{N}(\vec{X}; \vec{\mu}_n, \Sigma_n)$$

Discrimination using Log Likelihood

As a simple example, let D=1

$$p(X) = \mathcal{N}(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

The Likelihood function takes the form:

$$L_k(X) = p(X \mid \omega_k) \cdot P(\omega_k) = P(\omega_k) \cdot \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}$$

We can simplify the math by working with the Logarithm of the likelihood.
We note that:

$$\hat{k} = \arg\max_k \{L_k(X)\} = \arg\max_k \{Log\{L_k(X)\}\}$$

because Log{} is a monotonic function. In this case we define $g_k()$ to be the Log-likelihood.

$$g_k(X) = Log\{\mathcal{N}(X;\mu_k,\sigma_k) \cdot P(\omega_k)\}$$

$$g_k(X) = Log\{\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}\} + Log\{P(\omega_k)\}$$

$$g_k(X) = Log\{\frac{1}{\sqrt{2\pi}\sigma_k}\} + Log\{e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}\} + Log\{P(\omega_k)\}$$

$$g_k(X) = -Log\{\sqrt{2\pi}\} - Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

$$g_k(X) = -Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

and

$$\hat{k} = \arg\max_k \{g_k(X)\}$$

Example for K > 2 and D > 1

In general, it is more effective to work with D>1 features.
In this case:

$$g_k(\vec{X}) = Log\{p(\omega_k \mid \vec{X})P(\omega_k)\}$$

Thus the classifier is a machine that calculates K functions $g_k(\vec{X})$
Followed by a maximum selection.

The discrimination function is $g_k(\vec{X}) = Log\{p(\omega_k \mid \vec{X})P(\omega_k)\}$

For a Gaussian (Normal) density function

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

$$Log(p(\vec{X} \mid \omega_k)) = Log(\frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k)})$$

$$Log(p(\vec{X} \mid \omega_k)) = -\frac{D}{2}Log(2\pi) - \frac{1}{2}Log\{Det(\Sigma_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k)$$

We can observe that $-\frac{D}{2}Log(2\pi)$ can be ignored because it is constant for all k.

The discrimination function becomes:

$$g_k(\vec{X}) = -\frac{1}{2}Log\{\det(\Sigma_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k) + Log\{p(\omega_k)\}$$

Different families of Bayesian classifiers can be defined by variations of this formula.
This becomes more evident if we reduce the equation to a quadratic polynomial.

Canonical Form for the discrimination function

The quadratic discriminant can be reduced to a standard (canonical) form.

$$g_k(\vec{X}) = -\frac{1}{2} Log\{\det(\Sigma_k)\} - \frac{1}{2}(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k) + Log\{p(\omega_k)\}$$

Let us start with the term $(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k)$.
This can be rewritten as:

$$(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k) = \vec{X}^T \Sigma_k^{-1} \vec{X} - \vec{X}^T \Sigma_k^{-1} \vec{\mu}_k - \vec{\mu}_k{}^T \Sigma_k^{-1} \vec{X} + \vec{\mu}_k{}^T \Sigma_k^{-1} \vec{\mu}_k$$

We note that $\vec{X}^T \Sigma_k^{-1} \vec{\mu}_k = \vec{\mu}_k{}^T \Sigma_k^{-1} \vec{X}$
and thus : $-\vec{X}^T \Sigma_k^{-1} \vec{\mu}_k - \vec{\mu}_k{}^T \Sigma_k^{-1} \vec{X} = -2\vec{\mu}_k{}^T \Sigma_k^{-1} \vec{X}$

we define: $\vec{W}_k = \frac{1}{2}\left(-2\vec{\mu}_k{}^T \Sigma_k^{-1}\right) = -\vec{\mu}_k{}^T \Sigma_k^{-1}$
Let us also define $D_k = -\frac{1}{2}\Sigma_k^{-1}$

The remaining terms are constant. Let us defined the constant

$$b_k = -\frac{1}{2}\vec{\mu}_k{}^T \Sigma_k^{-1} \vec{\mu}_k - \frac{1}{2} Log\{\det(\Sigma_k)\} + Log\{p(\omega_k)\}$$

which gives a quadratic polynomial

$$\boxed{g_k(\vec{X}) = \vec{X}^T D_k \vec{X} + \vec{W}_k^T \vec{X} + b_k}$$

where: $D_k = -\frac{1}{2} C_k^{-1}$

$\vec{W}_k = -\vec{\mu}_k{}^T \Sigma_k^{-1}$

and $b_k = -\frac{1}{2}\vec{\mu}_k{}^T \Sigma_k^{-1} \vec{\mu}_k - \frac{1}{2} Log\{\det(\Sigma_k)\} + Log\{p(\omega_k)\}$

A set of K discrimination functions $g_k(\vec{X})$ partitions the space $\vec{X}$ into a disjoint set of regions with quadratic boundaries. The boundaries are the functions $g_i(\vec{X}) - g_j(\vec{X}) = 0$

The boundaries are defined by points for which
$g_i(\vec{X}) = g_j(\vec{X}) \geq g_k(\vec{X}) \; \forall k \neq i, j$

More about Normal Density Functions

Biased and Unbiased Variance

Note that this is a "Biased" variance. The unbiased variance would be

$$\tilde{\sigma}^2 = \frac{1}{M-1}\sum_{m=1}^{M}(X_m - \mu)^2$$

If we draw a random sample $\{X_m\}$ of M random variables from a Normal density with parameters (μ, σ)

$$\{X_m\} \leftarrow \mathcal{N}(x;\mu,\tilde{\sigma})$$

Then we compute the moments, we obtain.

$$\mu = E\{X_m\} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

and

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu)^2 \qquad \text{Where } \tilde{\sigma}^2 = \frac{M}{M-1}\hat{\sigma}^2$$

Note the notation: ~ means "true", ^ means estimated.

The expectation underestimates the variance by 1/M.

The RMS error for estimating *p(X)* from M samples $\{X_m\}$ is the difference between a biased and unbiased error.

Linear Transforms of the Normal Multivariate Density

The Normal (Gaussian) function is a defined only by its moments.
It is thus invariant to transformations of its moments, that is affine transformations.
The affine transformations include rotation, translation, scale changes and other linear transformations.

For example consider a cosine vector $\vec{R}$, such that $\|R\| = 1$

$$\vec{R} = \begin{pmatrix} \cos(\alpha_1) \\ \cos(\alpha_2) \\ \cdots \\ \cos(\alpha_D) \end{pmatrix}$$

R is "direction vector" . A vector $\vec{X}$ may be projected onto R to give $\vec{Y}$

$$\vec{Y} = \vec{R}^T \vec{X}$$

The Normal (Gaussian) function is defined only by its moments.
It is thus invariant to affine transformations of its moments.
The affine transformations include all linear transformations such as rotation, translation, scale changes and sheer.

For a projection onto a 1D vector Y, R is D x 1 : $\vec{Y} = \vec{R}^T \vec{X}$

$$\mu_y = \vec{R}^T \vec{\mu}_x \qquad \sigma_y^2 = \vec{R}^T \Sigma_x \vec{R}$$

Note that for the Covariance, projection requires pre- and post- multiplication by $\vec{R}$.
We can demonstrate this with a linear algebraic expression of the moments.

Linear Algebraic Form for Moment Calculation

Calculation of the mean and covariance are often expressed using linear algebra. Such expressions are widely used in machine learning.

As before, assume a set of M training examples $\{X_m\}$

Recall $$\vec{\mu} = E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix}$$

We can compose a matrix with M columns and D rows from $\{X_m\}$.

$$X = \begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1M}\\ x_{21} & x_{22} & \cdots & x_{2M}\\ \cdots & \cdots & & \cdots\\ x_{D1} & x_{D2} & \cdots & x_{DM}\end{pmatrix} \qquad \text{Let us define the unit vector : } \vec{u} = \begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix}$$

Then $$\vec{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1N}\\ x_{21} & x_{22} & \cdots & x_{2N}\\ \cdots & \cdots & \ddots & \cdots\\ x_{D1} & x_{D2} & \cdots & x_{DN}\end{pmatrix}\begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix} = X\cdot\vec{u}$$

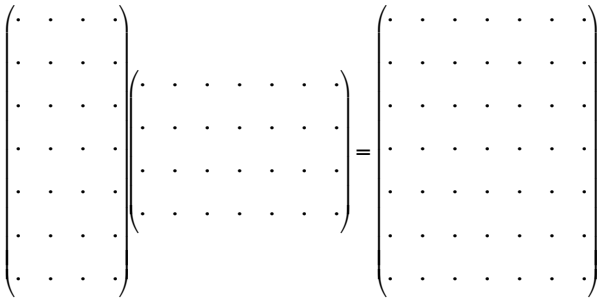
Let us define $\vec{V}_m = \vec{X}_m - E\{\vec{X}_m\} = \vec{X}_m - \vec{\mu}$.

We can compose a matrix with M columns and D rows from $\{V_m\}$.

$$V = \begin{pmatrix}v_{11} & v_{12} & \cdots & v_{1M}\\ v_{21} & v_{22} & \cdots & v_{2M}\\ \cdots & \cdots & \ddots & \cdots\\ v_{D1} & v_{D2} & \cdots & v_{DM}\end{pmatrix}$$

From this: $\Sigma_x = E\{\vec{V}\,\vec{V}^T\}$ can be computed as a vector product.

$\Sigma_x \equiv V\ V^T$ is a D x D matrix that captures the "co-variance" of the elements of *i*,*j* of the vectors in $\{X_m\}$

This can be seen as $\Sigma_X = \boldsymbol{V}\boldsymbol{V}^T =$ 

We can use this to show that projection of a covariance requires pre and post multiplication:

suppose that we have $y = \vec{R}^T\vec{V}$
Then for a set of M vectors, $\{X_m\}$, we can produce a vector of M values of y

$$Y = R^T V$$

Each element is the projection of a V in the direction R.

Thus $\Sigma_y = YY^T$

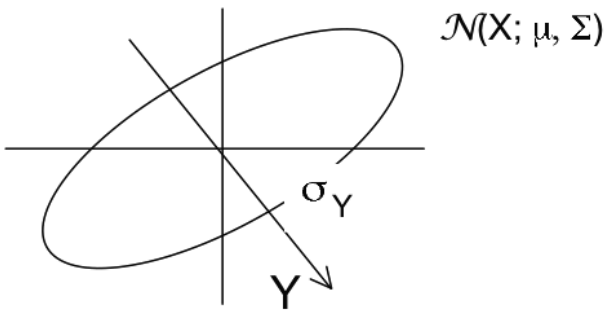
$$\Sigma_y = (\vec{R}^T V)(\vec{R}^T V)^T$$

Note that $(\vec{R}^T V)^T = (V^T\vec{R})$

$$\Sigma_y = (\vec{R}^T V)(V^T\vec{R})$$
$$\Sigma_y = \vec{R}^T(VV^T)\vec{R}$$
$$\Sigma_y = \vec{R}^T\Sigma_X\vec{R}$$

Thus projection of a covariance requires pre and post multiplication by $\vec{R}$.

In the case of projection to a 1D vector Y: $\sigma_y^2 = \vec{R}^T\Sigma_X\vec{R}$ the result is a variance for a 1-D Normal in the direction R, that is a cross section of Σ_X



Incremental Estimation of a Gaussian

It is possible to incrementally estimate the parameters for a Normal Density, as new data arrives. For this we need to have the number of observations, N, used for the prior estimate. The new observation is taken as the N+1 sample.

Incremental Update for the mean

$$\mu_M = \frac{1}{M}\sum_{m=1}^{M} X_m$$

$$M \cdot \mu_M = \sum_{m=1}^{M} X_m$$

$$(M+1)\cdot \mu_M = X_{M+1} + \sum_{m=1}^{M} X_m = X_{M+1} + M \cdot \mu$$

$$\mu_{M+1} = \frac{1}{(M+1)}\left(X_{M+1} + M \cdot \mu\right) = \frac{1}{(M+1)} X_{M+1} + \frac{M}{(M+1)}\mu$$

We can generalize with $\alpha = \frac{M}{(M+1)}$ then $\mu_{M+1} = (1-\alpha)\cdot X_{M+1} + \alpha \cdot \mu$

Incremental Update for the Variance

$$\sigma_M^2 = E\{X_M^2\} - E\{X_M\}^2 = \frac{1}{M}\sum_{m=1}^{M} X_m^2 - (\frac{1}{M}\sum_{m=1}^{M} X_m)^2$$

Substitute $\frac{1}{M}\sum_{m=1}^{M} X_m = \mu_M$

$$\sigma_M^2 = \frac{1}{M}\sum_{m=1}^{M} X_m^2 - \mu_M^2$$

$$\sigma_M^2 + \mu_M^2 = \frac{1}{M}\sum_{m=1}^{M} X_m^2$$

$$M(\sigma_M^2 + \mu_M^2) = \sum_{m=1}^{M} X_m^2$$

$$(M+1)(\sigma_{M+1}^2 + \mu_{M+1}^2) = \sum_{m=1}^{M} X_m^2 + X_{M+1}^2$$

substitute : $\sum_{m=1}^{M} X_m^2 = M(\sigma_M^2 + \mu_M^2)$

$$(M+1)(\sigma_{M+1}^2+\mu_{M+1}^2)=M(\sigma_M^2+\mu_M^2)+X_{M+1}^2$$

$$(\sigma_{M+1}^2+\mu_{M+1}^2)=\frac{M}{(M+1)}(\sigma_M^2+\mu_M^2)+\frac{1}{(M+1)}X_{M+1}^2$$

$$\sigma_{M+1}^2=\frac{M}{(M+1)}(\sigma_M^2+\mu_M^2)+\frac{1}{(M+1)}X_{M+1}^2-\mu_{M+1}^2$$

$$\sigma_{M+1}^2=\alpha(\sigma_M^2+\mu_M^2)+(1-\alpha)X_{M+1}^2-\mu_{M+1}^2$$

Combining Multiple Density Functions.

We can use a similar formula to combine (or average) two Gaussian densities
$\mathcal{N}(X;\mu_1,\sigma_1)$estimated from M_1 observations $\{X_m^{(1)}\}$ and
$\mathcal{N}(X;\mu_2,\sigma_2)$ estimated from M_2 observations $\{X_m^{(2)}\}$

We note that : $\mu_1 = \frac{1}{M_1}\sum_{m=1}^{M_1} X_m^{(1)} \quad \mu_2 = \frac{1}{M_2}\sum_{m=1}^{M_2} X_m^{(2)}$

note (1) and (2) are indices and we do not need to know that actual sample sets. We will simply use the formula for the derivation.

$$M \cdot \mu_M = \sum_{m=1}^{M} X_m$$

$$N\mu_N = \sum_{n=1}^{N} X_n$$

$$(M_1 + M_2)\mu_3 = \sum_{m=1}^{M_1} X_m^{(1)} + \sum_{m=1}^{M_2} X_m^{(2)} = M_1\mu_1 + M_2\mu_2$$

so

$$\mu_3 = \frac{M_1\mu_1 + M_2\mu_2}{M_1 + M_2}$$

Variance:

$$\sigma^2 = E\{X_m^2\} - E\{X_m\}^2 = \frac{1}{M}\sum_{m=1}^{M} X_m^2 - (\frac{1}{M}\sum_{m=1}^{M} X_m)^2$$

$$\sigma^2 = \frac{1}{M}\sum_{m=1}^{M} X_m^2 - \mu^2$$

$$M(\sigma^2 + \mu^2) = \sum_{m=1}^{M} X_m^2$$

$$M_1(\sigma_1^2 + \mu_1^2) = \sum_{m=1}^{M_1} X_m^{(1)^2} \qquad M_2(\sigma_2^2 + \mu_2^2) = \sum_{m=1}^{M_1} X_m^{(2)^2}$$

$$M_3(\sigma_3^2 + \mu_3^2) = \sum_{m=1}^{M_3} X_m^{(3)^2} = \sum_{m=1}^{M_1} X_m^{(1)^2} + \sum_{m=1}^{M_2} X_m^{(2)^2}$$

Substitute $M(\sigma^2 + \mu^2) = \sum_{m=1}^{M} X_m^2$

$$N_3(\sigma_{N_3}^2 + \mu_{N_3}^2) = N_1(\sigma_1^2 + \mu_1^2) + N_2(\sigma_2^2 + \mu_2^2)$$

$$\sigma_{N_3}^2 + \mu_{N_3}^2 = \frac{N_1}{N_3}(\sigma_1^2 + \mu_1^2) + \frac{N_2}{N_3}(\sigma_2^2 + \mu_2^2)$$

$$\sigma_{N_3}^2 + \mu_{N_3}^2 = \frac{N_1}{N_3}(\sigma_1^2 + \mu_1^2) + \frac{N_2}{N_3}(\sigma_2^2 + \mu_2^2)$$

$$\sigma_{N_3}^2 + \mu_{N_3}^2 = \frac{N_1}{N_1 + N_2}(\sigma_1^2 + \mu_1^2) + \frac{N_2}{N_1 + N_2}(\sigma_2^2 + \mu_2^2)$$

$$\sigma_{N_3}^2 = \frac{N_1}{N_1 + N_2}(\sigma_1^2 + \mu_1^2) + \frac{N_2}{N_1 + N_2}(\sigma_2^2 + \mu_2^2) - \mu_{N_3}^2$$

We can use this for the case where each observation has some intrinsic imprecision. The observation is represented as a Gaussian, centered on X with precision σ.

$$\mathcal{N}(X;\sigma)$$