Notes on Probability Theory

Paul F. Roysdon, Ph.D.

I. Definitions

The *expected value*, or *mean*, of a discrete random variable (r.v. hereafter) $\mathbf{X}$:

$$\begin{aligned}\mathbf{m_X} &= \mathrm{E}\left\langle \mathbf{X} \right\rangle \\ &= \sum_{k=1}^{\infty} \boldsymbol{x}_k \ P_{\mathbf{X}}(\boldsymbol{x}_k)\end{aligned}$$

The *variance* of the r.v. $\mathbf{X}$:

$$\begin{aligned}\boldsymbol{\sigma}_{\mathbf{X}}^2 &= \mathrm{var}\left\langle \mathbf{X} \right\rangle \\ &= \mathrm{E}\left\langle (\mathbf{X} - \mathbf{m_X})^2 \right\rangle \\ &= \sum_{k=1}^{\infty} (\boldsymbol{x} - \mathbf{m_X})^2 \ P_{\mathbf{X}}(\boldsymbol{x}_k)\end{aligned}$$

which can also be expressed as

$$\begin{aligned}\mathrm{var}\left\langle \mathbf{X} \right\rangle &= \mathrm{E}\left\langle (\mathbf{X} - \mathbf{m_X})^2 \right\rangle \\ &= \mathrm{E}\left\langle \mathbf{X}^2 - 2\mathbf{m_X}\mathbf{X} + \mathbf{m_X}^2 \right\rangle \\ &= \mathrm{E}\left\langle \mathbf{X}^2 \right\rangle - 2\mathbf{m_X}\mathrm{E}\left\langle \mathbf{X} \right\rangle + \mathbf{m_X}^2 \\ &= \mathrm{E}\left\langle \mathbf{X}^2 \right\rangle - \mathbf{m_X}^2\end{aligned}$$

where $\mathrm{E}\left\langle \mathbf{X}^2 \right\rangle$ is the called the second moment of $\mathbf{X}$.

The *standard deviation* of the r.v. $\mathbf{X}$:

$$\begin{aligned}\boldsymbol{\sigma}_{\mathbf{X}} &= \mathrm{std}\left\langle \mathbf{X} \right\rangle \\ &= \sqrt{\mathrm{var}\left\langle \mathbf{X} \right\rangle}\end{aligned}$$

For the discrete r.v. $\boldsymbol{x} \in \mathbb{R}^{m\times 1}$ and scalar variable c, the variance has the following properties, each resulting in a *scalar random variable*:

1) Adding a constant c, does not affect the variance of $\mathbf{X}$

$$\begin{aligned}\mathrm{var}\left\langle \boldsymbol{x} + c \right\rangle &= \mathrm{E}\left\langle (\boldsymbol{x} + c - (\mathrm{E}\left\langle \boldsymbol{x} \right\rangle + c))^2 \right\rangle \\ &= \mathrm{E}\left\langle (\boldsymbol{x} - \mathrm{E}\left\langle \boldsymbol{x} \right\rangle)^2 \right\rangle \\ &= \mathrm{var}\left\langle \boldsymbol{x} \right\rangle\end{aligned}$$

2) Multiplying by a constant c, scales the variance of $\boldsymbol{x}$ by c^2

$$\begin{aligned}\mathrm{var}\left\langle c\boldsymbol{x} \right\rangle &= \mathrm{E}\left\langle (c\boldsymbol{x} - c\mathrm{E}\left\langle \boldsymbol{x} \right\rangle)^2 \right\rangle \\ &= \mathrm{E}\left\langle c^2(\boldsymbol{x} - \mathrm{E}\left\langle \boldsymbol{x} \right\rangle)^2 \right\rangle \\ &= c^2\mathrm{var}\left\langle \boldsymbol{x} \right\rangle\end{aligned}$$

3) Let $\boldsymbol{x} = c$, then the constant r.v. has zero variance

$$\begin{aligned}\mathrm{var}\left\langle \boldsymbol{x} \right\rangle &= \mathrm{E}\left\langle (\boldsymbol{x} - c)^2 \right\rangle \\ &= \mathrm{E}\left\langle 0 \right\rangle \\ &= 0\end{aligned}$$

For the discrete r.v. $\boldsymbol{x}, \mathbf{y} \in \mathbb{R}^{m\times 1}$, the covariance has the following property, resulting in a *matrix of random variables*:

$$\begin{aligned}\mathrm{cov}\left\langle \boldsymbol{x}, \mathbf{y} \right\rangle &= \mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} - \boldsymbol{x}\mathrm{E}\left\langle \mathbf{y} \right\rangle - \mathbf{y}\mathrm{E}\left\langle \boldsymbol{x} \right\rangle + \mathrm{E}\left\langle \boldsymbol{x} \right\rangle \mathrm{E}\left\langle \mathbf{y} \right\rangle \right\rangle \\ &= \mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} \right\rangle - 2\mathrm{E}\left\langle \boldsymbol{x} \right\rangle \mathrm{E}\left\langle \mathbf{y} \right\rangle + \mathrm{E}\left\langle \boldsymbol{x} \right\rangle \mathrm{E}\left\langle \mathbf{y} \right\rangle \\ &= \mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} \right\rangle - \mathrm{E}\left\langle \boldsymbol{x} \right\rangle \mathrm{E}\left\langle \mathbf{y} \right\rangle\end{aligned}$$

If either $\boldsymbol{x}$ or $\mathbf{y}$ have zero mean, i.e. $\mathrm{E}\left\langle \boldsymbol{x} \right\rangle = 0$ or $\mathrm{E}\left\langle \mathbf{y} \right\rangle = 0$, then $\mathrm{cov}\left\langle \boldsymbol{x}, \mathbf{y} \right\rangle = \mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} \right\rangle$. If the correlation of $\boldsymbol{x}$ and $\mathbf{y}$ is zero, i.e. $\mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} \right\rangle = 0$ then $\boldsymbol{x}$ and $\mathbf{y}$ are orthogonal.

For the discrete r.v. $\boldsymbol{x}, \mathbf{y} \in \mathbb{R}^{m\times 1}$, the correlation coefficient has the following property

$$\begin{aligned}\rho_{\boldsymbol{x},\mathbf{y}} &= \frac{\mathrm{cov}\left\langle \boldsymbol{x}, \mathbf{y} \right\rangle}{\boldsymbol{\sigma}_{\boldsymbol{x}}\boldsymbol{\sigma}_{\mathbf{y}}} \\ &= \frac{\mathrm{E}\left\langle \boldsymbol{x}\mathbf{y} \right\rangle - \mathrm{E}\left\langle \boldsymbol{x} \right\rangle \mathrm{E}\left\langle \mathbf{y} \right\rangle}{\boldsymbol{\sigma}_{\boldsymbol{x}}\boldsymbol{\sigma}_{\mathbf{y}}}\end{aligned}$$

where $\boldsymbol{\sigma}_{\boldsymbol{x}} = \sqrt{\mathrm{var}\left\langle \boldsymbol{x} \right\rangle}$, $\boldsymbol{\sigma}_{\mathbf{y}} = \sqrt{\mathrm{var}\left\langle \mathbf{y} \right\rangle}$, and $-1 \le \rho_{\boldsymbol{x},\mathbf{y}} \le 1$. If $\rho_{\boldsymbol{x},\mathbf{y}} = 0$ then $\boldsymbol{x}$ and $\mathbf{y}$ are said to be *uncorrelated.*

II. Linear Combinations

A. Linear Forms

Assume $\mathbf{X}$ and $\boldsymbol{x}$ to be a matrix and a vector of random variables. Then

$$\begin{aligned}\mathrm{E}\left\langle \mathbf{AXB} + \mathbf{C} \right\rangle &= \mathbf{A} \ \mathrm{E}\left\langle \mathbf{X} \right\rangle \ \mathbf{B} + \mathbf{C} \\ \mathrm{var}\left\langle \mathbf{A}\boldsymbol{x} \right\rangle &= \mathbf{A} \ \mathrm{var}\left\langle \boldsymbol{x} \right\rangle \ \mathbf{A}^{\mathsf{T}} \\ \mathrm{cov}\left\langle \mathbf{A}\boldsymbol{x}, \mathbf{B}\mathbf{y} \right\rangle &= \mathbf{A} \ \mathrm{cov}\left\langle \boldsymbol{x}, \mathbf{y} \right\rangle \ \mathbf{B}^{\mathsf{T}}\end{aligned}$$

Assume $\boldsymbol{x}$ to be a stochastic vector with mean $\mathbf{m}$, then

$$\begin{aligned}\mathrm{E}\left\langle \mathbf{A}\boldsymbol{x} + \mathbf{b} \right\rangle &= \mathbf{Am} + \mathbf{b} \\ \mathrm{E}\left\langle \mathbf{A}\boldsymbol{x} \right\rangle &= \mathbf{Am} \\ \mathrm{E}\left\langle \boldsymbol{x} + \mathbf{b} \right\rangle &= \mathbf{m} + \mathbf{b}\end{aligned}$$

B. Quadratic Forms

Assume $\mathbf{A}$ is symmetric, $\mathbf{c} = \mathrm{E}\left\langle \boldsymbol{x} \right\rangle$ and $\boldsymbol{\Sigma} = \mathrm{var}\left\langle \boldsymbol{x} \right\rangle$. Assume also that all coordinates x_i are independent, have the same central moments $\mu_1, \mu_2, \mu_3, \mu_4$ and denote $\mathbf{a} = \mathrm{diag}(\mathbf{A})$. Then

$$\begin{aligned}\mathrm{E}\left\langle \boldsymbol{x}^{\mathsf{T}}\mathbf{A}\boldsymbol{x} \right\rangle &= \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \mathbf{c}^{\mathsf{T}}\mathbf{A}\mathbf{c} \\ \mathrm{var}\left\langle \boldsymbol{x}^{\mathsf{T}}\mathbf{A}\boldsymbol{x} \right\rangle &= 2\mu_2^2\mathrm{Tr}(\mathbf{A}^2) + 4\mu_2\mathbf{c}^{\mathsf{T}}\mathbf{A}^2\mathbf{c} \\ &\quad + 4\mu_3\mathbf{c}^{\mathsf{T}}\mathbf{A}\mathbf{a} + (\mu_4 - 3\mu_2^2)\mathbf{a}^{\mathsf{T}}\mathbf{a}\end{aligned}$$

Also, assume $\boldsymbol{x}$ to be a stochastic vector with mean $\mathbf{m}$, and covariance $\mathbf{M}$. Then

$$\begin{aligned}
\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x}+\mathbf{a})(\mathbf{B}\boldsymbol{x}+\mathbf{b})^\intercal\right\rangle &= \mathbf{AMB}^\intercal+(\mathbf{Am}+\mathbf{a})(\mathbf{Bm}+\mathbf{b})^\intercal \\
\mathrm{E}\left\langle\boldsymbol{x}\boldsymbol{x}^\intercal\right\rangle &= \mathbf{M}+\mathbf{mm}^\intercal \\
\mathrm{E}\left\langle\boldsymbol{x}\mathbf{a}^\intercal\boldsymbol{x}\right\rangle &= (\mathbf{M}+\mathbf{mm}^\intercal)\mathbf{a} \\
\mathrm{E}\left\langle\boldsymbol{x}^\intercal\mathbf{a}\boldsymbol{x}^\intercal\right\rangle &= \mathbf{a}^\intercal(\mathbf{M}+\mathbf{mm}^\intercal) \\
\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x})(\mathbf{A}\boldsymbol{x})^\intercal\right\rangle &= \mathbf{A}(\mathbf{M}+\mathbf{mm}^\intercal)\mathbf{A}^\intercal \\
\mathrm{E}\left\langle(\boldsymbol{x}+\mathbf{a})(\boldsymbol{x}+\mathbf{a})^\intercal\right\rangle &= \mathbf{M}+(\mathbf{m}+\mathbf{a})(\mathbf{m}+\mathbf{a})^\intercal \\
\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x}+\mathbf{a})^\intercal(\mathbf{B}\boldsymbol{x}+\mathbf{b})\right\rangle &= \mathrm{Tr}(\mathbf{AMB}^\intercal)+(\mathbf{Am}+\mathbf{a})^\intercal(\mathbf{Bm}+\mathbf{b}) \\
\mathrm{E}\left\langle\boldsymbol{x}^\intercal\boldsymbol{x}\right\rangle &= \mathrm{Tr}(\mathbf{M})+\mathbf{m}^\intercal\mathbf{m} \\
\mathrm{E}\left\langle\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x}\right\rangle &= \mathrm{Tr}(\mathbf{AM})+\mathbf{m}^\intercal\mathbf{Am} \\
\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x})^\intercal(\mathbf{A}\boldsymbol{x})\right\rangle &= \mathrm{Tr}(\mathbf{AMA}^\intercal)+(\mathbf{Am})^\intercal(\mathbf{Am}) \\
\mathrm{E}\left\langle(\boldsymbol{x}+\mathbf{a})^\intercal(\boldsymbol{x}+\mathbf{a})\right\rangle &= \mathrm{Tr}(\mathbf{M})+(\mathbf{m}+\mathbf{a})^\intercal(\mathbf{m}+\mathbf{a})
\end{aligned}$$

C. Cubic Forms

Assume $\boldsymbol{x}$ to be a stochastic vector with independent coordinates, mean $\mathbf{m}$, covariance $\mathbf{M}$ and central moments $\mathbf{v}_3=\mathrm{E}\left\langle(\boldsymbol{x}-\mathbf{m})^3\right\rangle$. Then

$$\begin{aligned}
&\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x}+\mathbf{a})(\mathbf{B}\boldsymbol{x}+\mathbf{b})^\intercal(\mathbf{C}\boldsymbol{x}+\mathbf{c})\right\rangle \\
&\quad=\mathbf{A}\ \mathrm{diag}(\mathbf{B}^\intercal\mathbf{C})\mathbf{v}_3 \\
&\qquad+\mathrm{Tr}(\mathbf{BMC}^\intercal)(\mathbf{Am}+\mathbf{a}) \\
&\qquad+\mathbf{AMC}^\intercal(\mathbf{Bm}+\mathbf{b}) \\
&\qquad+\left(\mathbf{AMB}^\intercal+(\mathbf{Am}+\mathbf{a})(\mathbf{Bm}+\mathbf{b})^\intercal\right)(\mathbf{Cm}+\mathbf{c}) \\
&\mathrm{E}\left\langle\boldsymbol{x}\boldsymbol{x}^\intercal\boldsymbol{x}\right\rangle \\
&\quad=\mathbf{v}_3+2\mathbf{Mm}+(\mathrm{Tr}(\mathbf{M})+\mathbf{m}^\intercal\mathbf{m})\mathbf{m} \\
&\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x}+\mathbf{a})(\mathbf{A}\boldsymbol{x}+\mathbf{a})^\intercal(\mathbf{A}\boldsymbol{x}+\mathbf{a})\right\rangle \\
&\quad=\mathbf{A}\ \mathrm{diag}(\mathbf{A}^\intercal\mathbf{A})\mathbf{v}_3 \\
&\qquad+[2\mathbf{AMA}^\intercal+(\mathbf{A}\boldsymbol{x}+\mathbf{a})(\mathbf{A}\boldsymbol{x}+\mathbf{a})^\intercal](\mathbf{Am}+\mathbf{a}) \\
&\qquad+\mathrm{Tr}(\mathbf{AMA}^\intercal)(\mathbf{Am}+\mathbf{a}) \\
&\mathrm{E}\left\langle(\mathbf{A}\boldsymbol{x}+\mathbf{a})\mathbf{b}^\intercal(\mathbf{C}\boldsymbol{x}+\mathbf{c})(\mathbf{D}\boldsymbol{x}+\mathbf{d})^\intercal\right\rangle \\
&\quad=(\mathbf{A}\boldsymbol{x}+\mathbf{a})\mathbf{b}^\intercal(\mathbf{CMD}^\intercal+(\mathbf{Cm}+\mathbf{c})(\mathbf{Dm}+\mathbf{d})^\intercal) \\
&\qquad+(\mathbf{AMC}^\intercal+(\mathbf{Am}+\mathbf{a})(\mathbf{Cm}+\mathbf{c})^\intercal)\mathbf{b}(\mathbf{Dm}+\mathbf{d})^\intercal \\
&\qquad+\mathbf{b}^\intercal(\mathbf{Cm}+\mathbf{c})(\mathbf{AMD}^\intercal-(\mathbf{Am}+\mathbf{a})(\mathbf{Dm}+\mathbf{d})^\intercal)
\end{aligned}$$

III. Gaussians

The following section contains some useful relations from statistics theory.

A. Density and normalization

The density of $\boldsymbol{x}\sim\mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$ is

$$p(\boldsymbol{x})=\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\mathbf{m})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mathbf{m})\right]$$

Note that if $\boldsymbol{x}$ is d-dimensional, then $\det(2\pi\boldsymbol{\Sigma})=(2\pi)^d\det(\boldsymbol{\Sigma})$.

Integration and normalization is given by:

$$\begin{aligned}
&\int\exp\left[-\frac{1}{2}(\boldsymbol{x}-\mathbf{m})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mathbf{m})\right]d\boldsymbol{x} \\
&\quad=\sqrt{\det(2\pi\boldsymbol{\Sigma})} \\
&\int\exp\left[-\frac{1}{2}\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x}+\mathbf{b}^\intercal\boldsymbol{x}\right]d\boldsymbol{x} \\
&\quad=\sqrt{\det(2\pi\mathbf{A}^{-1})}\exp\left[\frac{1}{2}\mathbf{b}^\intercal\mathbf{A}^{-1}\mathbf{b}\right] \\
&\int\exp\left[-\frac{1}{2}\mathrm{Tr}(\mathbf{S}^\intercal\mathbf{AS})+\mathrm{Tr}(\mathbf{B}^\intercal\mathbf{S})\right]d\mathbf{S} \\
&\quad=\sqrt{\det(2\pi\mathbf{A}^{-1})}\exp\left[\frac{1}{2}\mathrm{Tr}(\mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B})\right]
\end{aligned}$$

The derivatives of the density are

$$\begin{aligned}
\frac{\partial p(\boldsymbol{x})}{\partial\boldsymbol{x}}&=-p(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mathbf{m}) \\
\frac{\partial^2 p}{\partial\boldsymbol{x}\partial\boldsymbol{x}^\intercal}&=p(\boldsymbol{x})\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mathbf{m})(\boldsymbol{x}-\mathbf{m})^\intercal\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\right)
\end{aligned}$$

B. Marginal Distribution

Assume $\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where

$$\boldsymbol{x}=\begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix}\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}\quad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_c\\ \boldsymbol{\Sigma}_c^\intercal & \boldsymbol{\Sigma}_b\end{bmatrix}$$

then

$$\begin{aligned}
p(\boldsymbol{x}_a)&=\mathcal{N}_{\boldsymbol{x}_a}(\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a) \\
p(\boldsymbol{x}_b)&=\mathcal{N}_{\boldsymbol{x}_b}(\boldsymbol{\mu}_b,\boldsymbol{\Sigma}_b)
\end{aligned}$$

C. Conditional Distribution

Assume $\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where

$$\boldsymbol{x}=\begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix}\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}\quad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_c\\ \boldsymbol{\Sigma}_c^\intercal & \boldsymbol{\Sigma}_b\end{bmatrix}$$

then

$$\begin{aligned}
p(\boldsymbol{x}_a|\boldsymbol{x}_b)&=\mathcal{N}_{\boldsymbol{x}_a}(\hat{\boldsymbol{\mu}}_a,\hat{\boldsymbol{\Sigma}}_a) \\
p(\boldsymbol{x}_b|\boldsymbol{x}_a)&=\mathcal{N}_{\boldsymbol{x}_b}(\hat{\boldsymbol{\mu}}_b,\hat{\boldsymbol{\Sigma}}_b)
\end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_a&=\boldsymbol{\mu}_a+\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\
\hat{\boldsymbol{\Sigma}}_a&=\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_c^\intercal \\
\hat{\boldsymbol{\mu}}_b&=\boldsymbol{\mu}_b+\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) \\
\hat{\boldsymbol{\Sigma}}_b&=\boldsymbol{\Sigma}_b-\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_c^\intercal
\end{aligned}$$

D. Linear combination

Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}_x, \boldsymbol{\Sigma}_x)$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{m}_y, \boldsymbol{\Sigma}_y)$

$$\begin{aligned}&\mathbf{A}\boldsymbol{x} + \mathbf{B}\mathbf{y} + \mathbf{c}\\ &\quad \sim \mathcal{N}(\mathbf{A}\mathbf{m}_x + \mathbf{B}\mathbf{m}_y + \mathbf{c},\ \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^\intercal + \mathbf{B}\boldsymbol{\Sigma}_y\mathbf{B}^\intercal)\end{aligned}$$

E. Rearranging Means

$$\begin{aligned}\mathcal{N}_{\mathbf{A}\boldsymbol{x}}\left[\mathbf{m}, \boldsymbol{\Sigma}\right] &= \frac{\sqrt{\det(2\pi(\mathbf{A}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1})}}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\\ &\quad \times \mathcal{N}_{\boldsymbol{x}}\left[\mathbf{A}^{-1}\mathbf{m},\ (\mathbf{A}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\right]\end{aligned}$$

F. Rearranging into squared form

If $\mathbf{A}$ is symmetric, then

$$\begin{aligned}&-\frac{1}{2}\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x} + \mathbf{b}^\intercal\boldsymbol{x}\\ &\quad = -\frac{1}{2}(\boldsymbol{x} - \mathbf{A}^{-1}\mathbf{b})^\intercal\mathbf{A}(\boldsymbol{x} - \mathbf{A}^{-1}\mathbf{b})^\intercal + \frac{1}{2}\mathbf{b}^\intercal\mathbf{A}^{-1}\mathbf{b}\\ &-\frac{1}{2}\mathrm{Tr}(\mathbf{X}^\intercal\mathbf{A}\mathbf{X}) + Tr(\mathbf{B}^\intercal\mathbf{X})\\ &\quad = -\frac{1}{2}[(\mathbf{X} - \mathbf{A}^{-1}\mathbf{B})^\intercal\mathbf{A}(\mathbf{X} - \mathbf{A}^{-1}\mathbf{B})] + \frac{1}{2}\mathrm{Tr}(\mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B})\end{aligned}$$

G. Sum of two squared forms

In vector form (assuming $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ are symmetric)

$$\begin{aligned}&-\frac{1}{2}(\boldsymbol{x} - \mathbf{m}_1)^\intercal\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \mathbf{m}_1)\\ &-\frac{1}{2}(\boldsymbol{x} - \mathbf{m}_2)^\intercal\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \mathbf{m}_2)\\ &= -\frac{1}{2}(\boldsymbol{x} - \mathbf{m}_c)^\intercal\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \mathbf{m}_c) + C\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma}_c^{-1} &= \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\\ \mathbf{m}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\mathbf{m}_1 + \boldsymbol{\Sigma}_2^{-1}\mathbf{m}_2)\\ C &= \frac{1}{2}(\mathbf{m}_1^\intercal\boldsymbol{\Sigma}_1^{-1} + \mathbf{m}_2^\intercal\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}\\ &\quad (\boldsymbol{\Sigma}_1^{-1}\mathbf{m}_1 + \boldsymbol{\Sigma}_2^{-1}\mathbf{m}_2) - \frac{1}{2}(\mathbf{m}_1^\intercal\boldsymbol{\Sigma}_1^{-1}\mathbf{m}_1 + \mathbf{m}_2^\intercal\boldsymbol{\Sigma}_2^{-1}\mathbf{m}_2)\end{aligned}$$

In a trace form (assuming $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ are symmetric)

$$\begin{aligned}&-\frac{1}{2}\mathrm{Tr}\left[(\mathbf{X} - \mathbf{M}_1)^\intercal\boldsymbol{\Sigma}_1^{-1}(\mathbf{X} - \mathbf{M}_1)\right]\\ &-\frac{1}{2}\mathrm{Tr}\left[(\mathbf{X} - \mathbf{M}_2)^\intercal\boldsymbol{\Sigma}_2^{-1}(\mathbf{X} - \mathbf{M}_2)\right]\\ &= -\frac{1}{2}\mathrm{Tr}\left[(\mathbf{X} - \mathbf{M}_c)^\intercal\boldsymbol{\Sigma}_c^{-1}(\mathbf{X} - \mathbf{M}_c)\right] + C\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma}_c^{-1} &= \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\\ \mathbf{M}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_1 + \boldsymbol{\Sigma}_2^{-1}\mathbf{M}_2)\\ C &= \frac{1}{2}\mathrm{Tr}[(\mathbf{M}_1^\intercal\boldsymbol{\Sigma}_1^{-1} + \mathbf{M}_2^\intercal\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}\\ &\quad (\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_1 + \boldsymbol{\Sigma}_2^{-1}\mathbf{M}_2)] - \frac{1}{2}(\mathbf{M}_1^\intercal\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_1\\ &\quad + \mathbf{M}_2^\intercal\boldsymbol{\Sigma}_2^{-1}\mathbf{M}_2)\end{aligned}$$

H. Product of Gaussian densities

Let $\mathcal{N}_{\boldsymbol{x}}(\mathbf{m}, \boldsymbol{\Sigma})$ denote a density of $\boldsymbol{x}$, then

$$\mathcal{N}_{\boldsymbol{x}}(\mathbf{m}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}_{\boldsymbol{x}}(\mathbf{m}_2, \boldsymbol{\Sigma}_2) = c_c\mathcal{N}_{\boldsymbol{x}}(\mathbf{m}_c, \boldsymbol{\Sigma}_c)$$

$$\begin{aligned}c_c &= \mathcal{N}_{\mathbf{m}_1}(\mathbf{m}_1, (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2))\\ &= \frac{1}{\sqrt{\det(2\pi(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2))}}\\ &\quad \exp\left[-\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^\intercal(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]\\ \mathbf{m}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\mathbf{m}_1 + \boldsymbol{\Sigma}_2^{-1}\mathbf{m}_2)\\ \boldsymbol{\Sigma}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}\end{aligned}$$

IV. Moments

A. Mean and covariance of linear forms

First and second moments. Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$

$$\mathrm{E}\langle\boldsymbol{x}\rangle = \mathbf{m} \tag{1}$$

then

$$\begin{aligned}\mathrm{Cov}\langle\boldsymbol{x}, \boldsymbol{x}\rangle &= \mathrm{Var}\langle\boldsymbol{x}\rangle\\ &= \boldsymbol{\Sigma}\\ &= \mathrm{E}\langle\boldsymbol{x}\boldsymbol{x}^\intercal\rangle - \mathrm{E}\langle\boldsymbol{x}\rangle\,\mathrm{E}\langle\boldsymbol{x}^\intercal\rangle\\ &= \mathrm{E}\langle\boldsymbol{x}\boldsymbol{x}^\intercal\rangle - \mathbf{m}\mathbf{m}^\intercal\end{aligned}$$

As for any other distribution is holds for gaussians that

$$\begin{aligned}\mathrm{E}\langle\mathbf{A}\boldsymbol{x}\rangle &= \mathbf{A}\mathrm{E}\langle\boldsymbol{x}\rangle\\ \mathrm{Var}\langle\mathbf{A}\boldsymbol{x}\rangle &= \mathbf{A}\mathrm{Var}\langle\boldsymbol{x}\rangle\,\mathbf{A}^\intercal\\ \mathrm{Cov}\langle\mathbf{A}\boldsymbol{x}, \mathbf{B}\mathbf{y}\rangle &= \mathbf{A}\mathrm{Cov}\langle\boldsymbol{x}, \mathbf{y}\rangle\,\mathbf{B}^\intercal\end{aligned}$$

B. Mean and variance of square forms

Mean and variance of square forms. Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$

$$\begin{aligned}\mathrm{E}\langle\boldsymbol{x}\boldsymbol{x}^\intercal\rangle &= \boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^\intercal\\ \mathrm{E}\langle\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x}\rangle &= \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \mathbf{m}^\intercal\mathbf{A}\mathbf{m}\\ \mathrm{Var}\langle\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x}\rangle &= 2\sigma^4\mathrm{Tr}(\mathbf{A}^2) + 4\sigma^2\mathbf{m}^\intercal\mathbf{A}^2\mathbf{m}\\ \mathrm{Cov}\langle(\boldsymbol{x} - \mathbf{m}')^\intercal\mathbf{A}(\boldsymbol{x} - \mathbf{m}')\rangle &= (\boldsymbol{x} - \mathbf{m}')^\intercal\mathbf{A}(\boldsymbol{x} - \mathbf{m}') + \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma})\end{aligned}$$

Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{A}$ and $\mathbf{B}$ to be symmetric, then

$$\mathrm{Cov}\langle\boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x}, \boldsymbol{x}^\intercal\mathbf{B}\boldsymbol{x}\rangle = 2\sigma^4\mathrm{Tr}(\mathbf{A}\mathbf{B})$$

C. Cubic forms

$$\begin{aligned}\mathrm{E}\langle\boldsymbol{x}\mathbf{b}^\intercal\boldsymbol{x}\boldsymbol{x}^\intercal\rangle &= \mathbf{m}\mathbf{b}^\intercal(\mathbf{M} + \mathbf{m}\mathbf{m}^\intercal) + (\mathbf{M} + \mathbf{m}\mathbf{m}^\intercal)\mathbf{b}\mathbf{m}^\intercal\\ &\quad + \mathbf{b}^\intercal\mathbf{m}(\mathbf{M} - \mathbf{m}\mathbf{m}^\intercal)\end{aligned}$$

D. Moments

$$\begin{aligned}\mathrm{E}\langle\boldsymbol{x}\rangle &= \sum_k \rho_k\mathbf{m}_k\\ \mathrm{Cov}\langle\boldsymbol{x}\rangle &= \sum_k\sum_{k'} \rho_k\rho_{k'}(\boldsymbol{\Sigma}_k + \mathbf{m}_k\mathbf{m}_k^\intercal + \mathbf{m}_{k'}\mathbf{m}_{k'}^\intercal)\end{aligned}$$

V. WHITENING

Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, then

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \mathbf{m}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Conversely having $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ one can generate data $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ by setting

$$\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{z} + \mathbf{m} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$$

Note that $\boldsymbol{\Sigma}^{1/2}$ means the matrix which fulfills $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$, and that it exists and is unique since $\boldsymbol{\Sigma}$ is positive definite.

VI. CHI-SQUARE

Assume $\boldsymbol{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ and $\boldsymbol{x}$ to be n dimensional, then

$$z = (\boldsymbol{x} - \mathbf{m})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mathbf{m}) \sim \boldsymbol{\chi}_n^2$$

where $\boldsymbol{\chi}_n^2$ denotes the Chi square distribution with n degrees of freedom.

VII. MIXTURE OF GAUSSIANS

A. Density

The variable $\boldsymbol{x}$ is distributed as a mixture of gaussians if it has the density

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \rho_k \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mathbf{m}_k)^\intercal \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \mathbf{m}_k)\right]$$

where ρ_k sum to 1 and the $\boldsymbol{\Sigma}_k$ all are positive definite.

B. Derivatives

Defining $p(s) = \sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, then

$$\begin{aligned}
\frac{\partial \ln p(s)}{\partial \rho_j} &= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \rho_j} \ln[\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \\
&= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{1}{\rho_j} \\
\frac{\partial \ln p(s)}{\partial \boldsymbol{\mu}_j} &= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\mu}_j} \ln[\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \\
&= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} [-\boldsymbol{\Sigma}_k^{-1}(s - \boldsymbol{\mu}_k)] \\
\frac{\partial \ln p(s)}{\partial \boldsymbol{\Sigma}_j} &= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\Sigma}_j} \ln[\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \\
&= \frac{\rho_j \mathcal{N}_s(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \rho_k \mathcal{N}_s(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \\
&\quad \frac{1}{2}[\boldsymbol{\Sigma}_j^{-\intercal} + \boldsymbol{\Sigma}_j^{-\intercal}(s - \boldsymbol{\mu}_j)(s - \boldsymbol{\mu}_j)^\intercal \boldsymbol{\Sigma}_j^{-\intercal}]
\end{aligned}$$

Note, ρ_k and $\boldsymbol{\Sigma}_k$ must be constrained.