Aerial Robotic Autonomy: Methods and Systems

Primer on Probability Theory

Kostas Alexis

Autonomous Robots Lab
Norwegian University of Science and Technology

Table of Contents

Table of Contents

Table of Contents

Definitions

Let x be a random variable distributed over a Probability Density Function (PDF) $p(x)$ over the interval $[a; b]$. Then we write:

$$\int_a^b p(x)dx = 1 \tag{1}$$

Thus satisfying the axiom of total probability.

For any two $c; d$ within $[a; b]$ the probability that x lies within c and d, $Pr(c \leq x \leq d)$ takes the form:

$$Pr(c \leq x \leq d) = \int_c^d p(x)dx \tag{2}$$

Definitions

To introduce a conditional variable, let $p(x|y)$ be a PDF over $x \in [a; b]$ conditioned over $y \in [r; s]$ such that:

$$(\forall y) \int_a^b p(x|y)dx = 1 \qquad (3)$$

For the case of N-dimensional continous variables let $\mathbf{x} = (x_1; x_2; ...; x_N)$ with $x_i \in [a_i; b_i]$. We then denote $p(\mathbf{x})$ or $p(x_1; x_2; ...; x_N)$. The axiom of total probability requires:

$$\int_{\mathbf{a}}^{\mathbf{b}} p(\mathbf{x})d\mathbf{x} = \int_{a_N}^{b_N} ... \int_{a_2}^{b_2}\int_{a_1}^{b_1} p(x_1; x_2; ...; x_N)dx_1dx_2...dx_N = 1 \qquad (4)$$

where $\mathbf{a} = (a_1; a_; ...; a_N)$ and $\mathbf{b} = (b_1; b_2; ...; b_N)$.

Bayes' Rule and Inference

The Bayes' rule

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \tag{5}$$

Allows us to infer the posterior or likelihood of the state given the measurements, $p(\mathbf{x}|\mathbf{y})$, *if* we have a prior PDF over the state $p(\mathbf{x})$ and the sensor model $p(\mathbf{y}|\mathbf{x})$.

The denominator is computed by marginalization

$$p(\mathbf{y}) = p(\mathbf{y})\int p(\mathbf{x}|\mathbf{y})d\mathbf{x} = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{x} = \int p(\mathbf{x},\mathbf{y})d\mathbf{x} = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{6}$$

We thus write:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}} \tag{7}$$

Moments

Moments of PDFs represent most notable characteristics.

The *zeroth* probability moment is always 1 (axiom of total probability).

The *rst* probability moment is known as the *mean*

$$= E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \tag{8}$$

where $E[\ ]$ denotes the expectation operator. For a general matrix function $\mathbf{F}(\mathbf{x})$ we write

$$E[\mathbf{F}(\mathbf{x})] = \int \mathbf{F}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{9}$$

Moments

The *second* probability moment is called the *covariance matrix*

$$= E[(\mathbf{x} \qquad)(\mathbf{x} \qquad)^T] \tag{10}$$

The next two moments are called *skewness* and *kurtosis*.

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Sample Mean and Covariance

Let $\mathbf{x}$ be a random variable with PDF $p(\mathbf{x})$. We can draw samples from this density

$$\mathbf{x}_{meas} \quad p(\mathbf{x}) \tag{11}$$

Taking N such samples allow us to derive the *sample mean* and *sample covariance*

$$_{meas} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i,meas}$$

$$_{meas} = \frac{1}{N \quad 1} \sum_{i=1}^{N} (\mathbf{x}_{i,meas} \quad _{meas})(\mathbf{x}_{i,meas} \quad _{meas})^T \tag{12}$$

- Normalization term $N \quad 1$, rather than N, is called the *Bessel's correction* and reflects the fact that the sample covariance uses the difference of the measurements against the sample mean.

Statistically Independent, Uncorrelated

Let $\mathbf{x}; \mathbf{y}$ be two random variables.

We say that $\mathbf{x}; \mathbf{y}$ are statistically independent if

$$p(\mathbf{x}; \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{13}$$

We say that $\mathbf{x}; \mathbf{y}$ are uncorrelated if

$$E[\mathbf{x}\mathbf{y}^T] = E[\mathbf{x}]E[\mathbf{y}]^T \tag{14}$$

- If the variables are statistically independent *this implies* that they are also uncorrelated.
- The reverse *is not always* true.

Normalized Product

If $p_1(\mathbf{x}), p_2(\mathbf{x})$ are two PDFs of $\mathbf{x}$, the normalized product $p(\mathbf{x})$ is formulated as

$$p(\mathbf{x}) = \; p_1(\mathbf{x})p_2(\mathbf{x}), \quad \; = \left(\int p_1(\mathbf{x})p_2(\mathbf{x})d\mathbf{x} \right)^{1} \tag{15}$$

with normalization constant ensuring that $p(\mathbf{x})$ satisfies the axiom of total probability.

The normalized product can be used to fuse independent estimates of a variable under the assumption of a uniform prior

$$p(\mathbf{x}|\mathbf{y}_1, \mathbf{y}_2) = \; p(\mathbf{x}|\mathbf{y}_1)p(\mathbf{x}|\mathbf{y}_2), \quad \; = \frac{p(\mathbf{y}_1)p(\mathbf{y}_2)}{p(\mathbf{y}_1, \mathbf{y}_2)p(\mathbf{x})} \tag{16}$$

Note: If we let the prior $p(\mathbf{x})$ be uniform for all values of $\mathbf{x}$ then is also a constant.

Shannon and Mutal Information

To assess our confidence for the estimate of the mean of a PDF we can use the *negative entropy* or *Shannon information H*

$$H(\mathbf{x}) = \quad E[\ln p(\mathbf{x})] = \quad \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \tag{17}$$

To measure how much knowing one of the variables, reduces the uncertainty for the other we can use the *mutual information* $I(\mathbf{x};\mathbf{y})$ between two random variables $\mathbf{x};\mathbf{y}$

$$I(\mathbf{x};\mathbf{y}) = E \quad \ln \quad \frac{p(\mathbf{x};\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \quad = \iint p(\mathbf{x};\mathbf{y}) \ln \quad \frac{p(\mathbf{x};\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \quad d\mathbf{x}d\mathbf{y} \tag{18}$$

- When $\mathbf{x};\mathbf{y}$ are statistically independent, $I(\mathbf{x};\mathbf{y}) = 0$.
- When $\mathbf{x};\mathbf{y}$ are statistically dependent, $I(\mathbf{x};\mathbf{y}) \quad 0$ and $I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) \quad H(\mathbf{x};\mathbf{y})$

Considering a deterministic parameter θ that influences a random variable $\mathbf{x}$, $p(\mathbf{x}|\theta)$ and a sample $\mathbf{x}_{meas} \sim p(\mathbf{x}|\theta)$, then the *Cramer-Rao lower bound (CRLB)* says that the covariance of any *unbiased estimate* $\hat{\theta}$ (based on $\mathbf{x}_{meas}$) of the deterministic parameter θ is bounded by the *Fisher Information Matrix* $\mathbf{I}(\mathbf{x}|\theta)$

$$\text{cov}(\hat{\theta}|\mathbf{x}_{meas}) = E\left[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\right] \geq \mathbf{I}^{-1}(\mathbf{x}|\theta) \quad (19)$$

The Fisher Information Matrix takes the form

$$\mathbf{I}(\mathbf{x}|\theta) = E\left[\left(\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}\right)^T \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}\right] \quad (20)$$

Thus, CRLB sets a fundamental limit on how certain we can be about an estimate of a parameter, given our measurements.

Table of Contents

Definitions

Throughout most of our work in state estimation for robotics we will be working with Gaussian PDFs.
One-dimensional Gaussian PDF over random variable $x \; 2 \; \mathrm{R}$

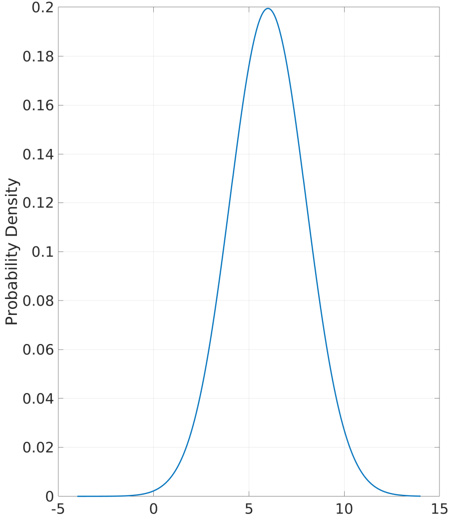
$$p(x j \quad ; \quad ^2) = \frac{1}{p \overline{2 \quad ^2}} \exp \quad \frac{1}{2} \frac{(x \qquad)^2}{\quad ^2} \tag{21}$$

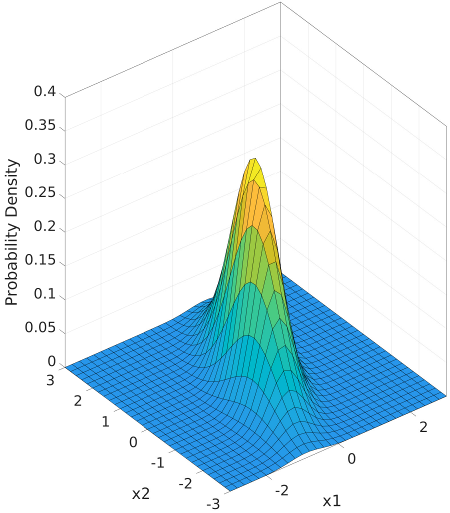
where $\; ; \; ^2$ are the mean and variance respectively (is called the standard deviation).
Multi-variate Gaussian PDF over random variable $x \; 2 \; \mathrm{R}^N$

$$p(\mathbf{x} j \quad ; \quad) = \frac{1}{p \overline{(2 \quad)^N \det}} \exp \quad \frac{1}{2} (\mathbf{x} \qquad)^T \quad ^1 (\mathbf{x} \qquad) \tag{22}$$

where $\; 2 \; \mathrm{R}^N; \quad 2 \; \mathrm{R}^{N \quad N}$ are the mean and covariance matrix respectively (symmetric, positive-definite).

Definitions





Definitions

Accordingly,

$$= E[\mathbf{x}] = \int_{\infty}^{\infty} \mathbf{x} \frac{1}{\sqrt{(2\ )^N \det}} \exp \left(\frac{1}{2}(\mathbf{x}\ \ )^T\ \ ^{1}(\mathbf{x}\ \ ) \right) d\mathbf{x} \quad (23)$$

$$= E[(\mathbf{x}\ \ )(\mathbf{x}\ \ )^T] = \int_{\infty}^{\infty} (\mathbf{x}\ \ )(\mathbf{x}\ \ )^T \frac{1}{\sqrt{(2\ )^N \det}} \exp \left(\frac{1}{2}(\mathbf{x}\ \ )^T\ \ ^{1}(\mathbf{x}\ \ ) \right) d\mathbf{x} \quad (24)$$

We also write that **x** is *normally* distributed

$$\mathbf{x}\quad N(\ ;\ ) \quad (25)$$

We also say that a random vaeriable is *standard normally distributed* if

$$\mathbf{x}\quad N(\mathbf{0}; \mathbf{1}) \quad (26)$$

where **1** is an $N \quad N$ identity matrix.

Isserli's Theorem

Isserli's theorem allows to compute certain **higher-order moments of a Gaussian random variable** $\mathbf{x} = (x_1, x_2, \ldots, x_{2M}) \in \mathbb{R}^{2M}$. In general, it says that

$$E[x_1 x_2 x_3 \ldots x_{2M}] = \sum \prod E[x_i x_j], \tag{27}$$

where this implies summing over all distinct ways of partitioning into a product of M pairs. There are $(2M)!/(2^M M!)$ terms in the sum.

With four variables we write

$$E[x_i x_j x_k x_\ell] = E[x_i x_j]E[x_k x_\ell] + E[x_i x_k]E[x_j x_\ell] + E[x_i x_\ell]E[x_j x_k] \tag{28}$$

Assuming $\mathbf{x} \quad N(\mathbf{0}; \quad) \; 2 \; \mathrm{R}^N$ we will have occasion to compute expressions of the form

$$E[\mathbf{x}(\mathbf{x}^T\mathbf{x})^p\mathbf{x}^T] \tag{29}$$

where p non-negative integer.

- For $p = 0$, it holds that $E[\mathbf{x}\mathbf{x}^T] =$
- For $p = 1$, it holds that $E[\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T] = \quad (\mathrm{tr}(\quad)\mathbf{1} + 2 \quad)$

Note that for the scalar case, we have $x \quad N(0; \; ^2$ and thus $E[x^4] = \; ^2(\; ^2 + 2 \; ^2) = 3 \; ^4$.

Let us also consider the case where

$$\mathbf{x} = \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{matrix} \quad N \quad \mathbf{0} \quad \begin{matrix} {}_{11} & {}_{12} \\ {}_{21} & {}_{22} \end{matrix} \quad ; \quad \dim(\mathbf{x}_1) = N_1; \ \dim(\mathbf{x}_2) = N_2 \tag{30}$$

Throughout our work in state estimation, we will need to compute expressions of the form

$$E[\mathbf{x}(\mathbf{x}_1^T \mathbf{x}_1)^p \mathbf{x}^T]; \quad p \ 2 \ \mathrm{N}_0 \tag{31}$$

For $p = 0$, $E[\mathbf{x}\mathbf{x}^T] =$
For $p = 1$ it holds that

$$E[\mathbf{x}\mathbf{x}_1^T \mathbf{x}_1 \mathbf{x}^T] = \quad \mathrm{tr}(\ {}_{11})\mathbf{1} + 2 \ \begin{matrix} {}_{11} & {}_{12} \\ \mathbf{0} & \mathbf{0} \end{matrix} \tag{32}$$

Similarly,

$$E[\mathbf{x}\mathbf{x}_2^T \mathbf{x}_2 \mathbf{x}^T] = \quad \mathrm{tr}(\ {}_{22})\mathbf{1} + 2 \ \begin{matrix} \mathbf{0} & \mathbf{0} \\ {}^{T}_{12} & {}_{22} \end{matrix} \tag{33}$$

Accordingly, we have the final check

$$E[\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T] = E[\mathbf{x}(\mathbf{x}_1^T\mathbf{x}_1 + \mathbf{x}_2^T\mathbf{x}_2)\mathbf{x}^T] = E[\mathbf{x}\mathbf{x}_1^T\mathbf{x}_1\mathbf{x}^T] + E[\mathbf{x}\mathbf{x}_2^T\mathbf{x}_2\mathbf{x}^T] \quad (34)$$

Furthermore, it holds that

$$E[\mathbf{x}\mathbf{x}^T\mathbf{A}\mathbf{x}\mathbf{x}^T] = \quad (\mathrm{tr}(\mathbf{A} \quad)\mathbf{1} + \mathbf{A} \quad + \mathbf{A}^T \quad) \quad (35)$$

where $\mathbf{A}$ is a compatible square matrix.

Joint Gaussian PDFs, their Factors, and Inference

We can have a joint Gaussian over a pair of variables $(\mathbf{x}; \mathbf{y})$

$$p(\mathbf{x}; \mathbf{y}) = N \quad x \;\; ; \;\; xx \quad xy \qquad (36)$$
$$y \qquad\quad yx \quad yy$$

with $yx = {}^{T}_{xy}$.

Schur complement

Suppose $\mathbf{A}; \mathbf{B}; \mathbf{C}; \mathbf{D}$ are respectively $p \quad p$, $p \quad q$, $q \quad p$, $q \quad q$, $p; q \; 2 \; \mathrm{N}_0$ matrices of complex numbers. Let matrix $\mathbf{M}$ with dimensions $(p+q) \quad (p+q)$

$$\mathbf{M} = \begin{matrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{matrix} \qquad (37)$$

the Schur complements of $\mathbf{A}$ in M are the matrices of the form $\mathbf{S} = \mathbf{D} \quad \mathbf{CaB}$, where $\mathbf{a}$ is the generalized inverse of $\mathbf{A}$. If $\mathbf{A}$ is invertible, this is $\mathbf{S} = \mathbf{D} \quad \mathbf{CA} \;\; {}^{1}\mathbf{B}$

Breaking a joint density into the product of two factors $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$ we can work out the details for the joint Gaussian using the Schur complement. It turns out that:

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \tag{38}$$

$$p(\mathbf{x}|\mathbf{y}) = N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) \tag{39}$$

$$p(\mathbf{y}) = (\mu_y, \Sigma_{yy}) \tag{40}$$

with both $p(\mathbf{x}|\mathbf{y}), p(\mathbf{y})$ are Gaussian PDFs.

If we know the value of $\mathbf{y}$ (i.e., it is measured), then we can work out the likelihood of $\mathbf{x}$ by computing $p(\mathbf{x}|\mathbf{y})$ using Eq. 39. This is in fact the *cornerstone of Gaussian Inference*: We start with a prior about our state, $\mathbf{x} \sim N(\mu_x, \Sigma_{xx})$ and then narrow down based on some measurements $\mathbf{y}_{meas}$. In Eq. 39 we see that both the mean μ_x and the covariance Σ_{xx} are adjusted.

Statistically Independent, Uncorrelated

For Gaussian PDFs, statistically independent variables are also uncorrelated **and** uncorrelated variables are also statistically independent.
Assuming statistical independence, $p(\mathbf{x};\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ and so $p(\mathbf{x}j\mathbf{y}) = p(\mathbf{x}) = N(\ _x;\ \ _{xx})$. This implies

$$_{xy}\ \ _{yy}^{1}(\mathbf{y}\ \ \ _y) = \mathbf{0} \quad (41)$$

$$_{xy}\ \ _{yy}^{1}\ \ _{yx} = \mathbf{0} \quad (42)$$

thus, also $\ _{xy} = \mathbf{0}$. Furthermore, since

$$_{xy} = E[(\mathbf{x}\ \ \ _x)(\mathbf{y}\ \ \ _y)^T] = E[\mathbf{x}\mathbf{y}^T]\ \ E[\mathbf{x}]E[\mathbf{y}]^T \quad (43)$$

we conclude the uncorrelated condition

$$E[\mathbf{x}\mathbf{y}^T] = E[\mathbf{x}]E[\mathbf{y}]^T \quad (44)$$

Linear Change of Variables

Suppose we have a Gaussian random variable $\mathbf{x} \in \mathbb{R}^N \sim \mathcal{N}(\mu_x, \Sigma_{xx})$ and that we have a second random variable $\mathbf{y} \in \mathbb{R}^M$ related to $\mathbf{x}$ through the linear map

$$\mathbf{y} = \mathbf{G}\mathbf{x}, \quad \mathbf{G} \in \mathbb{R}^{M \times N} \ (const) \tag{45}$$

With respect to the statistical properties of $\mathbf{y}$ it holds that

$$\mu_y = E[\mathbf{y}] = E[\mathbf{G}\mathbf{x}] = \mathbf{G}E[\mathbf{x}] = \mathbf{G}\mu_x \tag{46}$$

$$\Sigma_{yy} = E[(\mathbf{y} - \mu_y)(\mathbf{y} - \mu_y)^T] = \mathbf{G}E[(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T]\mathbf{G}^T = \mathbf{G}\Sigma_{xx}\mathbf{G}^T \tag{47}$$

and thus we write that $\mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_{yy}) = \mathcal{N}(\mathbf{G}\mu_x, \mathbf{G}\Sigma_{xx}\mathbf{G}^T)$.

Linear Change of Variables

An alternative way to conclude the same is through a change of variables. We assume that the linear map is *injective* (that is that two values of $\mathbf{x}$ cannot map to a single $\mathbf{y}$ value) and in fact $\boldsymbol{G}$ is *invertible* (thus also $M = N$). By the axiom of total probability

$$\int_{-\infty}^{\infty} p(\mathbf{x})d\mathbf{x} = 1 \tag{48}$$

A small volume of $\mathbf{x}$ is related to a small volume $\mathbf{y}$ by $d\mathbf{y} = |\det \mathbf{G}| d\mathbf{x}$. We can then make a substitution of variables to have

$$1 = \int_{-\infty}^{\infty} p(\mathbf{x})d\mathbf{x} = \ldots = \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{G}\Sigma_{xx}\mathbf{G}^T)}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{G}\mu_x)^T(\mathbf{G}\Sigma_{xx}\mathbf{G}^T)^{-1}(\mathbf{y} - \mathbf{G}\mu_x)\right) d\mathbf{y} \tag{49}$$

where we have $\mu_y = \mathbf{G}\mu_x, \quad \Sigma_{yy} = \mathbf{G}\Sigma_{xx}\mathbf{G}^T$ as derived before.

Likewise, we can derive the statistics of **x** from **y** given that this linear mapping is invertible. This however is a bit trickier as the resulting covariance of **x** will blow up since we are dilating to a larger space. To overcome this problem, we switch to *information form*. Let

$$\mathbf{u} = \ {}_{yy}^{\ 1}\mathbf{y} \tag{50}$$

we have

$$\mathbf{u} \quad N(\ {}_{yy}^{\ 1} \ {}_{y}; \ {}_{yy}^{\ 1}) \tag{51}$$

Likewise, let

$$\mathbf{v} = \ {}_{xx}^{\ 1}\mathbf{x} \tag{52}$$

we have

$$\mathbf{v} \quad N(\ {}_{xx}^{\ 1} \ {}_{x}; \ {}_{xx}^{\ 1}) \tag{53}$$

Linear Change of Variables

Since the mapping from **y** to **x** is not unique, we need to specify what we want to do. One choice is

$$\mathbf{v} = \mathbf{G}^T \mathbf{u} \quad , \qquad {}_{xx}^{\;1}\mathbf{x} = \mathbf{G}^T \;\; {}_{yy}^{\;1}\mathbf{y} \tag{54}$$

then we take the expectations

$$ {}_{xx}^{\;1} \;\; {}_{x} = E[\mathbf{v}] = \mathbf{G}^T E[\mathbf{u}] = \mathbf{G}^T \;\; {}_{yy}^{\;1} \;\; {}_{y} \tag{55}$$

$$ {}_{xx}^{\;1} = E[(\mathbf{v} \qquad {}_{xx}^{\;1} \;\; {}_{x})(\qquad {}_{xx}^{\;1} \;\; {}_{x})^T] = \mathbf{G}^T \;\; {}_{yy}^{\;1}\mathbf{G} \tag{56}$$

Note: if ${}_{xx}^{\;1}$ is not full rank then we cannot recover ${}_{xx};\;\; {}_{x}$ and must keep them in information form. However, multiple such estimates can be fused together.

Normalized Product of Gaussians

The normalized product of K Gaussian PDFs is *also a Gaussian PDF*

$$\exp \quad \frac{1}{2}(\mathbf{x} \qquad)^T \quad {}^{1}(\mathbf{x} \qquad) \qquad \prod_{k=1}^{K} \exp \quad \frac{1}{2}(\mathbf{x} \qquad {}_k)^T \quad {}_k{}^{1}(\mathbf{x} \qquad {}_k) \tag{57}$$

where $\;^{1} = \sum_{k=1}^{K} \;{}_k{}^{1}$, $\;^{1} \; = \sum_{k=1}^{N} \;{}_k{}^{1} \;{}_k$, and is a normalization constant to enforce the axiom of total probability.

- The normalized product of Gaussians comes up when fusing multiple estimates together.

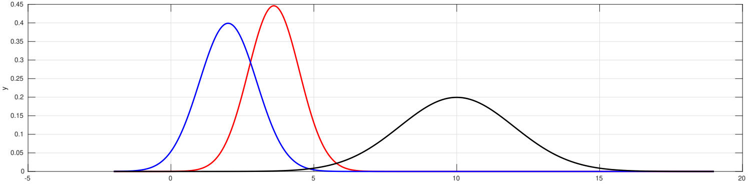


Figure: Illustrational example

We also have that

$$\exp \quad \frac{1}{2}(\mathbf{x} \qquad)^T \quad ^{1}(\mathbf{x} \qquad) \qquad \prod_{k=1}^{K} \exp \quad \frac{1}{2}(\mathbf{G}_k\mathbf{x} \qquad _k)^T \quad _k{}^{1}(\mathbf{G}_k\mathbf{x} \qquad _k) \tag{58}$$

where

$$^{1} = \sum_{k=1}^{K} \mathbf{G}_k^T \quad _k{}^{1}\mathbf{G}_k \tag{59}$$

$$^{1} \quad = \sum_{k=1}^{K} \mathbf{G}_k^T \quad _k{}^{1} \quad _k \tag{60}$$

in the case that the matrices, $\mathbf{G}_k$ 2 $\mathrm{R}^{M_k \quad N}$, are present, with $M_k \quad N$ and is again a normalization constant.

Sherman-Morrison-Woodbury (SMW) Identity

In state estimation, the Sherman-Morrison-Woodbury (SMW) *matrix identity(ies)* (also called the *matrix inversion lemma*) is commonly used. SMW is in fact four identities coming from the same derivation.

A matrix can be factored into either a *lower-diagonal-upper (LDU)* or *upper-diagonal-lower (UDL)* form:

$$\begin{bmatrix} \mathbf{A}^{1} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{CA} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} + \mathbf{CAB} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{AB} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \quad \text{(LDU)} \qquad (61)$$

$$\begin{bmatrix} \mathbf{A}^{1} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{BD}^{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{1} + \mathbf{BD}^{1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{D}^{1}\mathbf{C} & \mathbf{1} \end{bmatrix} \quad \text{(UDL)} \qquad (62)$$

We then invert each of these forms. For the LDU

$$\begin{matrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{matrix}^{1} = \begin{matrix} \mathbf{1} & \mathbf{AB} \\ \mathbf{0} & \mathbf{1} \end{matrix} \begin{matrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & (\mathbf{D}+\mathbf{CAB})^{1} \end{matrix} \begin{matrix} \mathbf{1} & \mathbf{0} \\ \mathbf{CA} & \mathbf{1} \end{matrix}) \tag{63}$$

$$\begin{matrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{matrix}^{1} = \begin{matrix} \mathbf{A} \quad \mathbf{AB}(\mathbf{D}+\mathbf{CAB})^{1}\mathbf{CA} & \mathbf{AB}(\mathbf{D}+\mathbf{CAB})^{1} \\ (\mathbf{D}+\mathbf{CAB})^{1}\mathbf{CA} & (\mathbf{D}+\mathbf{CAB})^{1} \end{matrix} \tag{64}$$

For the UDL case

$$\begin{matrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{matrix}^{1} = \begin{matrix} (\mathbf{A}^{1}+\mathbf{BD}^{1}\mathbf{C})^{1} & (\mathbf{A}^{1}+\mathbf{BD}^{1}\mathbf{C})^{1}\mathbf{BD}^{1} \\ \mathbf{D}^{1}\mathbf{C}(\mathbf{A}^{1}+\mathbf{BD}^{1}\mathbf{C})^{1} & \mathbf{D}^{1} \quad \mathbf{D}^{1}\mathbf{C}(\mathbf{A}^{1}+\mathbf{BD}^{1}\mathbf{C})^{1}\mathbf{BD}^{1} \end{matrix} \tag{65}$$

Sherman-Morrison-Woodbury (SMW) Identity

Comparing from Eq. 63 and 64 we conclude the SMW identities:

SMW Identities

$$(\mathbf{A}\ ^{1} + \mathbf{BD}\ ^{1}\mathbf{C})\ ^{1} \quad \mathbf{A} \quad \mathbf{AB}(\mathbf{D} + \mathbf{CAB})\ ^{1}\mathbf{CA} \tag{66}$$

$$(\mathbf{D} + \mathbf{CAB})\ ^{1} \quad \mathbf{D}\ ^{1} \quad \mathbf{D}\ ^{1}\mathbf{C}(\mathbf{A}\ ^{1} + \mathbf{BD}\ ^{1}\mathbf{C})\ ^{1}\mathbf{BD}\ ^{1} \tag{67}$$

$$\mathbf{AB}(\mathbf{D} + \mathbf{CAB})\ ^{1} \quad (\mathbf{A}\ ^{1} + \mathbf{BD}\ ^{1}\mathbf{C})\ ^{1}\mathbf{BD}\ ^{1} \tag{68}$$

$$(\mathbf{D} + \mathbf{CAB})\ ^{1}\mathbf{CA} \quad \mathbf{D}\ ^{1}\mathbf{C}(\mathbf{A}\ ^{1} + \mathbf{BD}\ ^{1}\mathbf{C})\ ^{1} \tag{69}$$

The SMW identities are frequently used when manipulating expressions involving covariance matrices of Gaussian PDFs.

Passing a Gaussian through a Nonlinearity

We can examine the process of passing a Gaussian PDF through a *stochastic nonlinearity*, namely computing

$$p(\mathbf{y}) = \int_{-\infty}^{\infty} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \quad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{g}(\mathbf{x}), \mathbf{R}), \quad p(\mathbf{x}) = \mathcal{N}(\mu_x, \Sigma_{xx}) \tag{70}$$

and $\mathbf{g}(\cdot), :\mathbf{x} \to \mathbf{y}$ is a nonlinear map that is then corrupted by zero-mean Gaussian noise with covariance $\mathbf{R}$.

- We shall require this type of stochastic nonlinearity when modeling sensors.
- Passing a Gaussian through this type of function is required when performing full Bayesian inference.

Passing a Gaussian through a Nonlinearity

Scalar Deterministic Case via Change of Variables

Let a Gaussian random variable $x \in \mathbb{R}$, $x \sim N(0, \sigma^2)$, i.e., $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right)$.
Consider the nonlinear mapping

$$y = \exp(x) \quad \rightarrow \quad x = \ln(y) \tag{71}$$

The infinitesimal integration volumes for x, y are then related by

$$dy = \exp(x)dx \text{ or } dx = \frac{1}{y}dy \tag{72}$$

According to the axiom of total probability

$$1 = \int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) dx \tag{73}$$

Passing a Gaussian through a Nonlinearity

$$= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\ln(y))^2}{\sigma^2}\right)\frac{1}{y}dy = \int_0^\infty p(y)dy \tag{74}$$

giving the exact expression for $p(y)$ which is visualized in the Figure below.

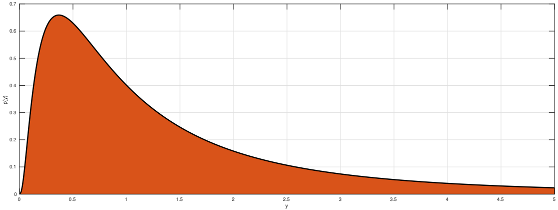


Figure: The PDF, $p(y)$, resulting from $p(x)$ being Gaussian PDF and passing through the nonlinearity $y = \exp(x)$.

Note that $p(y)$ is no longer Gaussian owing to the nonlinear change of variables.

General Case via Linearization: Such analytical calculation -as before - is not generally possible. Moreover, when the nonlinearity is stochastic (i.e., $\mathbf{R} > 0$), our mapping will never be invertible due to the extra input coming from the noise. One way to solve for processing the nonlinear effect is through *linearization*. We linearize the nonlinear map such that

$$\begin{aligned} \mathbf{g}(\mathbf{x}) \quad & {}_y + \mathbf{G}(\mathbf{x} \quad {}_x) \\ \mathbf{G} &= \frac{@\mathbf{g}(\mathbf{x})}{@\mathbf{x}}\bigg|_{\mathbf{x}= \ {}_x} \\ {}_y &= \mathbf{g}(\ {}_x) \end{aligned} \tag{75}$$

where $\mathbf{G}$ is the *Jacobian* pf $\mathbf{g}(\ )$ with respect to $\mathbf{x}$. This process allows us to pass the Gaussian through the linearized function in closed form and it is thus an approximation that works well for mildly nonlinear maps.

Returning to

$$p(\mathbf{y}) = \int_{-\infty}^{\infty} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{76}$$

we have that

$$\begin{aligned} p(\mathbf{y}) = \int_{-\infty}^{\infty} & \exp\left(-\frac{1}{2}(\mathbf{y} - (\mu_y + \mathbf{G}(\mathbf{x} - \mu_x)))^T \mathbf{R}^{-1}(\mathbf{y} - (\mu_y + \mathbf{G}(\mathbf{x} - \mu_x)))\right) \\ & \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_x)^T \Sigma_{xx}^{-1}(\mathbf{x} - \mu_x)\right) d\mathbf{x} \end{aligned} \tag{77}$$

$$\begin{aligned} = \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_y)^T \mathbf{R}^{-1}(\mathbf{y} - \mu_y)\right) \int_{-\infty}^{\infty} & \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_x)^T (\Sigma_{xx}^{-1} + \mathbf{G}^T \mathbf{R}^{-1}\mathbf{G})(\mathbf{x} - \mu_x)\right) \\ & \exp\left((\mathbf{y} - \mu_y)^T \mathbf{R}^{-1}\mathbf{G}(\mathbf{x} - \mu_x)\right) d\mathbf{x} \end{aligned} \tag{78}$$

Passing a Gaussian through a Nonlinearity

After manipulation

$$p(\mathbf{y}) = \quad \exp \quad \frac{1}{2}(\mathbf{y} \qquad {}_{y})^{T}(\mathbf{R}^{\ 1} \quad \mathbf{R}^{\ 1}\mathbf{G}(\mathbf{G}^{T}\mathbf{R}^{\ 1}\mathbf{G} + \quad {}_{xx}^{\ 1})^{\ 1}\mathbf{G}^{T}\mathbf{R}^{\ 1})(\mathbf{y} \qquad {}_{y}) \tag{79}$$

By exploiting the SMW inequalities it turns out

$$p(\mathbf{y}) = \qquad \frac{1}{2}(\mathbf{y} \qquad {}_{y})^{T}(\mathbf{R} + \mathbf{G} \quad {}_{xx}\mathbf{G}^{T})^{\ 1}(\mathbf{y} \qquad {}_{y}) \tag{80}$$

wherre   is the new normalization constant. Accordingly we write

$$\mathbf{y} \quad N(\quad {}_{y}; \quad {}_{yy}) = N(\mathbf{g}(\quad {}_{x}); \mathbf{R} + \mathbf{G} \quad {}_{xx}\mathbf{G}^{T}) \tag{81}$$

Shannon Information of a Gaussian

Shannon information of a Gaussian PDF

$$H(\mathbf{x}) = \int_{\infty}^{\infty} p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$
$$= \frac{1}{2} \ln \left((2 \quad)^N \det \quad \right) + \frac{1}{2} E \left[(\mathbf{x} \quad)^T \quad ^{1} (\mathbf{x} \quad) \right] \qquad (82)$$

where the second term is written as an expectation and in fact corresponds to a squared *Mahalanobis distance*.

The Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936.

Shannon Information of a Gaussian

Mahalanobis distance

For an observation **x** from a set of observations with mean   and covariance matrix   the Mahalanobis distance is defined as

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} \quad)^T \quad ^{1}(\mathbf{x} \quad)} \tag{83}$$

The Mahalanobis distance can also be defined as the dissemilarity between two vectors $\mathbf{x}; \mathbf{y}$, of the same distribution with covariance   , as

$$d_M(\mathbf{x}; \mathbf{y}) = \sqrt{(\mathbf{x} \quad \mathbf{y})^T \quad ^{1}(\mathbf{x} \quad \mathbf{y})} \tag{84}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called a standardized Euclidean distance.

Shannon Information of a Gaussian

The quadratic function inside the expectation in $H(\mathbf{x})$ can be written as

$$(\mathbf{x} \qquad)^T \quad {}^1(\mathbf{x} \qquad) = \mathrm{tr}(\quad {}^1(\mathbf{x} \qquad)(\mathbf{x} \qquad)^T) \tag{85}$$

Accordingly it can be shown that

$$E[(\mathbf{x} \qquad)^T \quad {}^1(\mathbf{x} \qquad)] = \mathrm{tr}(E[\quad {}^1(\mathbf{x} \qquad)(\mathbf{x} \qquad)^T] = N \tag{86}$$

Substituting back to into the expression for Shannon information

$$H(\mathbf{x}) = \frac{1}{2} \ln \quad (2 \quad e)^N \det \tag{87}$$

which is purely a function of the covariance matrix   of the Gaussian PDF.

Geometric interpretation: $p\overline{\det \quad}$ is proportional to the *volume of the uncertainty ellipsoid* formed by the Gaussian.

Mutual Information of a Joint Gaussian PDF

Let the joint Gaussian for variables $\mathbf{x}\ 2\ \mathrm{R}^N; \mathbf{y}\ 2\ \mathrm{R}^M$

$$p(\mathbf{x};\mathbf{y}) = N(\quad;\quad) = N \quad \begin{matrix} x \\ y \end{matrix} \; ; \; \begin{matrix} xx & xy \\ yx & yy \end{matrix} \tag{88}$$

It can be shown that the *mutual information for the joint Gaussian* is

$$I(\mathbf{x};\mathbf{y}) = \frac{1}{2} \ln \frac{\det}{\det \quad_{xx} \det \quad_{yy}} \tag{89}$$

And by further processing it turns out that

$$I(\mathbf{x};\mathbf{y}) = \frac{1}{2} \ln \det(\mathbf{1} \quad {}_{xx}^{1} \quad {}_{xy} \quad {}_{yy}^{1} \quad {}_{yx}) = \tag{90}$$

$$= \frac{1}{2} \ln \det(\mathbf{1} \quad {}_{yy}^{1} \quad {}_{yx} \quad {}_{xx}^{1} \quad {}_{xy}) \tag{91}$$

The two forms above are equivalent based on *Sylvester's determinant theorem* which states that $\det(\mathbf{1} \quad \mathbf{AB}) = \det(\mathbf{1} \quad \mathbf{BA})$ even when $\mathbf{A};\mathbf{B}$ are not square.

Cramér-Rao Lower Bound for Gaussian PDFs

Let K samples (measurements) $\mathbf{x}_{meas;k} \in \mathrm{R}^N$ drawn from a Gaussian PDF. The K *statistically independent random variables* associated with these measurements are

$$(\forall)\mathbf{x}_k \quad N(\quad;\quad) \tag{92}$$

Due to statistical independence, it holds that $E(\mathbf{x}_k \quad)(\mathbf{x}_\ell \quad)^T] = 0;\ k \neq \ell$. It can be shown that the *Fisher Information Matrix* takes the form

$$I(\mathbf{x}|\quad) = K \quad {}^{1} \tag{93}$$

Thus, the CRLB says

$$\mathrm{cov}(\hat{\mathbf{x}}|\mathbf{x}_{meas;1}, \ldots, \mathbf{x}_{meas;K}) \quad \frac{1}{K} \tag{94}$$

i.e., the more measurements, the smaller the lower limit of the uncertainty in the estimate.

Cramér-Rao Lower Bound for Gaussian PDFs

Note that in computing the CRLB there was no need to specify the form of the unbiased estimator. *The CRLB is the lower bound for* **any** *unbiased estimator.* An estimator that performs exactly at the CRLB can be found

$$\hat{\mathbf{x}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{meas;k} \quad (95)$$

where

$$E[\hat{\mathbf{x}}] = E\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k\right] = \frac{1}{K}\sum_{k=1}^{K}E[\mathbf{x}_k] = \frac{1}{K}\sum_{k=1}^{K} \quad = \quad (96)$$

$$\mathrm{cov}(\hat{\mathbf{x}}|\mathbf{x}_{meas;1},\ldots,\mathbf{x}_{meas;K}) = E[(\hat{\mathbf{x}} \quad)(\hat{\mathbf{x}} \quad)^T] = \quad (97)$$

$$= E\left[\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k \quad \right)\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k \quad \right)^T\right] = \frac{1}{K}$$

which is exactly at the CRLB.

Table of Contents

Gaussian Processes

We denote a Gaussian random variable $\mathbf{x}$ *2* R^N

$$\mathbf{x} \quad N(\;\; ; \;\;) \tag{98}$$

we use this type of random variable extensively to represent discrete-time quantities.

To represent state quantities that are continuous functions of time, t, we introduce *Gaussian Processes (GPs)*. For GPs we say that there is a *mean function*, (t), and a *covariance function*, $(t; t^0)$.

- The whole trajectory is viewed as a single variable belonging to a class of functions.
- The closer a function is to the mean function, the more likely it is.
- The covariance controls how smooth the function is by describing the correlation between two times t and t^0.

$$\mathbf{x}(t) \quad GP(\quad (t); \quad (t; t^{0}))$$

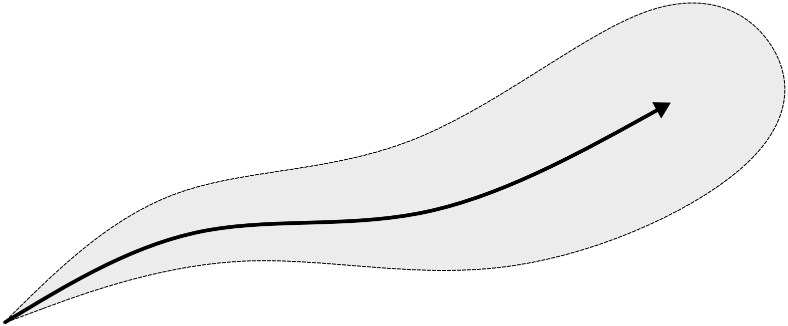


Figure: Continuous-time trajectories can be represented using Gaussian processes, which have a mean function (dark line) and a covariance function (shaded area).

Gaussian Processes

To indicate that a continuous-time trajectory is a *Gaussian Process (GP)*, we write

$$\mathbf{x}(t) \quad GP(\quad (t); \quad (t;t^{0})) \tag{99}$$

If we consider a variable at a single particular time of interest, , we write

$$\mathbf{x}(\ ) \quad N(\quad (\ ); \quad (\ ; \ )) \tag{100}$$

where $(\ ; \ )$ is a simple covariance matrix. We have marginalized out all of the other instants of time, leaving $\mathbf{x}(\ )$ as a usual Gaussian variable.

Zero-mean, White noise process

$$\mathbf{w}(t) \quad GP(\mathbf{0}; \mathbf{Q} \ (t \quad t^{0})) \tag{101}$$

where (t) is Dirac's delta and $\mathbf{Q}$ is a *power spectral density*. This is a *stationary* noise process as it depends only on the difference $t \quad t^{0}$.

Thank you

Q&A

Assignments

Other matters