E Finite-dimensional probability theory

This appendix provides a short survey of the primary tools of classical probability theory with an emphasis on Fourier-domain techniques. It also includes a complete characterization of multivariate Gaussian probabilities.

E.1 Background: The probabilistic formalism

Generally speaking, a random variable is a variable whose possible values are the numerical outcomes of some random phenomenon or experiment. In probability theory and mathematical statistics, it is customary to represent such a random variable as a *measurable* function X (see terminology in Appendix F) that maps the outcome of an experiment $\omega \in \Omega$—a point in a conceptual sample space Ω of possible outcomes—into a concrete numerical output or state $X(\omega) \in \mathcal{X}$ where $\mathcal{X}$ is the so-called state space.

It is convenient to think of the elements of the outcome space as all the different possibilities that could happen, and of a realization $X(\omega)$ as the value that the random variable X attains when one of the possibilities did happen. To indicate the distinction between the random generation process as a whole and an actual outcome, we use straight/capitalized roman letters to denote random variables $\big($e.g., X or $\boldsymbol{X} = (X_1, \ldots, X_N)\big)$ and an ordinary mathematical font for the corresponding domain or output variables (e.g., $x \in \mathbb{R}$ or $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$). The latter span the space of possible outcomes and also serve as index of the underlying probability density function $\big($e.g., $p_{\boldsymbol{X}} : \boldsymbol{x} \mapsto p_{\boldsymbol{X}}(\boldsymbol{x})\big)$ considered in Section E.2.

At the more abstract level of probability theory where Ω and $\mathcal{X}$ can be arbitrary sets, the complete statistical information available on X is encoded in the probability measure $\mathscr{P}_X$, which provides the probability of any event $E \in \Sigma(\mathcal{X})$ (the σ-field associated with $\mathcal{X}$, see Appendix F, Definition 42):

$$0 \leq \mathscr{P}_X(E) = \text{Prob}(X \in E) \leq 1$$

where $E \subseteq \mathcal{X}$ is a subset of the state space describing the configurations of interest. For instance, if $\mathcal{X} = \mathbb{R}$ and $E = (-\infty, x_0]$, then $\mathscr{P}_X(E)$ returns the probability that the random variable X takes a value $x = X(\omega)$ smaller or equal to x_0.

The formal basis for this description is the classical notion of a probability space $(\Omega, \Sigma, \mathscr{P})$, which has three fundamental components:

1. The sample space Ω, which is the set of all possible outcomes of an experiment.

2. The σ-field $\Sigma = \Sigma(\Omega)$, which is a collection of subsets Ω that satisfies the completeness properties of Definition 42.

3. The probability measure $\mathscr{P}$ on Σ, which is a map that associates a consistent probability $0 \leq \mathscr{P}(E) \leq 1$ to any event $E \in \Sigma$.

Definition 38 (Probability measure)**.** *Let Ω be a sample space with corresponding σ-field $\Sigma(\Omega)$. The map $\mathscr{P} : \Sigma(\Omega) \to [0,1]$ is said to be a* probalility measure *on $\Sigma(\Omega)$ if it satisfies the three consistency conditions*

- $\mathscr{P}(\emptyset) = 0$
- $\mathscr{P}(\Omega) = 1$
- *Countable additivity: For any countable collection $\{E_i\}_{i\in I} \subseteq \Sigma(\Omega)$ of pairwise disjoint sets: $\mathscr{P}\Big(\bigcup_{i\in I} E_i\Big) = \sum_{i\in I} \mathscr{P}(E_i)$.*

In the formal representation of a random variable as a measurable function $X : \Omega \to \mathcal{X}$, the role of the sample space Ω is primarily notational, as it provides us with an indexing mechanism to describe specific realizations of X. The only constraint is that Ω should be rich enough to map into the chosen state space $\mathcal{X}$. The assumption that the map $\omega \mapsto X(\omega)$ is measurable implies that $X^{-1}(E) = \{\omega : X(\omega) \in E\}$, the preimage of $E \in \Sigma(\mathcal{X})$, is included in $\Sigma(\Omega)$ for any admissible event E in state space. Consequently, if the underlying probability space is $(\Omega, \Sigma, \mathscr{P})$, then the probability measure $\mathscr{P}_X : \Sigma(\mathcal{X}) \to [0,1]$ that is induced on the random variable X is such that

$$\mathscr{P}_X(E) = \mathscr{P}(\{\omega : X(\omega) \in E\}).$$

An obvious choice is to simply take $(\Omega = \mathcal{X}, \Sigma = \Sigma(\mathcal{X}), \mathscr{P} = \mathscr{P}_X)$, which ensures that all the compatibility conditions are met.

The bottom line is that the outcome of a random experiment yields a realization $X(\omega) = x \in \mathcal{X}$ of the random variable X and that all the statistical information is condensed into the probability measure $\mathscr{P}_X$.

Example 6 (Binary variable)**.** *Here the sample space is $\Omega_{\text{binary}} =$ {False, True}, while the corresponding σ-field is $\Sigma = \{\emptyset, E_0, E_1, \Omega_{\text{binary}}\}$ with $E_0 =$ {False} and $E_1 =$ {True}. The discrete probability measure on Σ associated with equiprobable outcomes is specified as $\mathscr{P}(\emptyset) = 0$, $\mathscr{P}(E_0) = \frac{1}{2}$, $\mathscr{P}(E_1) = \frac{1}{2}$, and $\mathscr{P}(\Omega_{\text{binary}}) = 1$. Finally, the binary random variable $X : \Omega_{\text{binary}} \to \{0,1\}$ is defined as $X(\text{False}) = 0$ and $X(\text{True}) = 1$.*

Example 7 (Scalar random variable)**.** *Here the sample space is $\Omega_X = \mathbb{R}$ which goes hand-in-hand with the Borel algebra $B(\mathbb{R})$.*

Example 8 (Random vector)**.** *Here the sample space is* $\Omega_{\boldsymbol{X}} = \mathbb{R}^N$ *and the standard choice of* σ*-field is the Borel* σ*-algebra* $B(\mathbb{R}^N)$*. The state space is* $\mathbb{R}^N$ *as well so that the random vector* $\boldsymbol{X}$ *specifies the map*

$$\omega \mapsto \boldsymbol{X}(\omega) = \big(X_1(\omega), \ldots, X_N(\omega)\big)$$

In this configuration, each component X_n *of* $\boldsymbol{X}$ *is a scalar random variable on its own right.*

E.2 Probability density functions and expectations

Our generic random variable $\boldsymbol{X} = (X_1, \ldots, X_N)$ is multivariate because it is composed of N scalar component random variables $X_1, \ldots, X_N$. The outcome of an experiment yields a realization (or observed value) $\boldsymbol{x} = \boldsymbol{X}(\omega) = \big(X_1(\omega), \ldots, X_N(\omega)\big)$ that takes some fixed value in $\mathbb{R}^N$ (the state space of possible outcomes). We recall that the random aspect of this probabilistic model is the mechanism that produces ω (outcome of the experiment) in accordance with the underlying probability law.

Since $\mathbb{R}^N$ is a finite-dimensional vector space, it is convenient to characterize the statistical distribution of $\boldsymbol{X}$ by its (joint) probability density function (pdf) $p_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{R}^+$. The only constraint here is that $p_{\boldsymbol{X}}$, in addition to being positive, should be Borel-measurable on $\mathbb{R}^N$ and such that $\int_{\mathbb{R}^N} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. The corresponding probability measure $\mathscr{P}_{\boldsymbol{X}} : B(\mathbb{R}^N) \to [0, 1]$ (see Section E.1) is then given by

$$\mathscr{P}_{\boldsymbol{X}}(E) = \int_E p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathrm{Prob}\{\boldsymbol{X} \in E\} \tag{131}$$

where $E \subseteq \mathbb{R}^N$ is any Borelian subset of $\mathbb{R}^N$. In particular, the positivity condition $p_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0$ ensures that $\mathscr{P}_{\boldsymbol{X}}(E) \geq 0$ for all $E \in B(\mathbb{R}^N)$, while the normalization constraint $\int_{\mathbb{R}^N} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$ gives $\mathrm{Prob}(\boldsymbol{X} \in \mathbb{R}^N) = \mathscr{P}_{\boldsymbol{X}}(\mathbb{R}^N) = 1$.

Definition 39 (Statistical independence)**.** *The random variables* $\boldsymbol{X}_1 \in \mathbb{R}^{N_1}$ *and* $\boldsymbol{X}_2 \in \mathbb{R}^{N_2}$ *with respective pdfs* $p_{\boldsymbol{X}_1}$ *and* $p_{\boldsymbol{X}_2}$ *are independent if their joint probability density function* $p_{(\boldsymbol{X}_1, \boldsymbol{X}_2)} : \mathbb{R}^{N_1+N_2} \to \mathbb{R}^+$ *can be factorized as*

$$p_{(\boldsymbol{X}_1, \boldsymbol{X}_2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = p_{\boldsymbol{X}_1}(\boldsymbol{x}_1) p_{\boldsymbol{X}_2}(\boldsymbol{x}_2).$$

The pdfs $p_{\boldsymbol{X}_1}$ and $p_{\boldsymbol{X}_2}$ in Definition 39 are also called the marginals of

$p_{(\boldsymbol{X}_1,\boldsymbol{X}_2)}$ and are such that

$$p_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \int_{\mathbb{R}^{N_2}} p_{(\boldsymbol{X}_1,\boldsymbol{X}_2)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_2$$
$$p_{\boldsymbol{X}_2}(\boldsymbol{x}_2) = \int_{\mathbb{R}^{N_1}} p_{(\boldsymbol{X}_1,\boldsymbol{X}_2)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1.$$

In particular, when the components X_n of the random variable $\boldsymbol{X}$ are i.i.d. with common pdf p_X, then the multivariate (or joint) pdf of $\boldsymbol{X}$ is separable with

$$p_{\boldsymbol{X}}(x_1,\ldots,x_N) = \prod_{n=1}^{N} p_X(x_n).$$

Basic examples of separable pdfs are the shifted Dirac distribution

$$p_{\mathrm{Const}}(\boldsymbol{x}) = \delta(\boldsymbol{x}-\boldsymbol{\mu}) = \prod_{n=1}^{N} \delta(x_n-\mu_n)$$

and the multivariate standardized Gaussian distribution

$$p_{\mathrm{Gauss}}(\boldsymbol{x}) = \prod_{n=1}^{N} \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x_n^2} = (2\pi)^{-N/2}\mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}\|^2}. \tag{132}$$

The first example represents the probability law of a constant—i.e, $\boldsymbol{X}=\boldsymbol{\mu}$ with probability 1—while the second specifies a standardized white Gaussian noise whose components are i.i.d. Gaussian with $X_n \sim \mathcal{N}(\mu=0,\sigma^2=1)$.

Let $f : \boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x}) = \big(f_1(\boldsymbol{x}),\ldots,f_M(\boldsymbol{x})\big)$ be a multivariate function $\mathbb{R}^N \to \mathbb{R}^M$ with measurable component functions f_m. Then, the *expected value* of $\boldsymbol{f}(\boldsymbol{X})$, where $\boldsymbol{X}$ is a random vector with pdf $p_{\boldsymbol{X}}$, is

$$\mathbb{E}\{\boldsymbol{f}(\boldsymbol{X})\} = \int_{\mathbb{R}^N} \boldsymbol{f}(\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

In particular, the mean of $\boldsymbol{X}$ is defined as

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}\{\boldsymbol{X}\} = \int_{\mathbb{R}^N} \boldsymbol{x}p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = (\mu_1,\ldots,\mu_N)$$

with $\mu_n = \mathbb{E}\{x_n\}$. The second-order dependencies of $\boldsymbol{X}$ are conveniently summarised by the $N\times N$ covariance matrix

$$\mathbf{C}_{\boldsymbol{X}} = \mathbb{E}\left\{(\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{X}})^T\right\},$$

which is symmetric and positive-definite. The entries of $\mathbf{C}_{\boldsymbol{X}}$ are the centered correlations $[\mathbf{C}_{\boldsymbol{X}}]_{m,n} = \mathbb{E}\{(x_m-\mu_m)(x_n-\mu_n)\}$ with the diagonal terms providing the variance of the component variables x_n; i.e., $[\mathbf{C}_{\boldsymbol{X}}]_{n,n} = \mathrm{Var}\{x_n\}$.

E.3 Characteristic function

The characteristic function (cf) $\hat{p}_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{C}$ of $\boldsymbol{X}$ is the (conjugate) N-dimensional Fourier transform of $p_{\boldsymbol{X}}$. Specifically,

$$\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathbb{E}\{\mathrm{e}^{\mathrm{j}\langle \boldsymbol{\xi}, \boldsymbol{x}\rangle}\} = \int_{\mathbb{R}^N} p_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{e}^{\mathrm{j}\langle \boldsymbol{\xi}, \boldsymbol{x}\rangle} \mathrm{d}\boldsymbol{x} = \mathcal{F}^*\{p_{\boldsymbol{X}}\}(\boldsymbol{\xi}),$$

where $\mathcal{F}^*$ is the adjoint[6] of the conventional Fourier operator $\mathcal{F}$ with the property $\mathcal{F}^*\{f\}(\boldsymbol{\xi}) = \mathcal{F}\{f\}(-\boldsymbol{\xi})$. The Fourier transform being invertible, the cf is in one-to-one correspondence with the pdf.

Theorem 34. *The characteristic function $\hat{p}_{\boldsymbol{X}} = \mathcal{F}^*\{p_{\boldsymbol{X}}\} : \mathbb{R}^N \to \mathbb{C}$ enjoys the following properties:*

1. *$\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi})$ is continuous, bounded (i.e. $|\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi})| \leq 1$), Hermitian-symmetric (i.e., $\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \overline{\hat{p}_{\boldsymbol{X}}(-\boldsymbol{\xi})}$) and such that $\hat{p}_{\boldsymbol{X}}(\mathbf{0}) = 1$.*

2. *Invertibility:*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{F}^{*-1}\{\hat{p}_{\boldsymbol{X}}\}(\boldsymbol{x}) = \int_{\mathbb{R}^N} \hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) \mathrm{e}^{-\mathrm{j}\langle \boldsymbol{\xi}, \boldsymbol{x}\rangle} \frac{\mathrm{d}\boldsymbol{\xi}}{(2\pi)^N}.$$

3. *Preservation of separability (or joint cf of a collection of independent random variables). Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ with $p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{(\boldsymbol{X}_1, \boldsymbol{X}_2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = p_{\boldsymbol{X}_1}(\boldsymbol{x}_1) p_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$. Then,*

$$\hat{p}_{(\boldsymbol{X}_1, \boldsymbol{X}_2)}(\boldsymbol{\xi}) = \hat{p}_{\boldsymbol{X}_1}(\boldsymbol{\xi}_1) \hat{p}_{\boldsymbol{X}_2}(\boldsymbol{\xi}_2)$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$. In particular, if the components x_n are i.i.d. with common cf $\hat{p}_X$, then

$$\hat{p}_{\boldsymbol{X}}(\omega_1, \ldots, \omega_N) = \prod_{n=1}^{N} \hat{p}_X(\omega_n).$$

The converse is also true: the separability of the cf implies the separability of the pdf and hence independence.

[6] This slight inconvenience results from the convention of statisticians which is to use j rather than $-\mathrm{j}$ in the definition of their Fourier transform (cf). Since $p_{\boldsymbol{X}}(\boldsymbol{x})$ is real-valued, we also have that $\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \overline{\mathcal{F}\{p_{\boldsymbol{X}}\}(\boldsymbol{\xi})}$.

4. *Linear transformation: Let* $\mathbf{H} \in \mathbb{R}^{M \times N}$ *be an arbitrary transformation matrix and* $\mathbf{b} \in \mathbb{R}^M$ *some constant offset vector. Then, the characteristic function of the transformed variable* $\boldsymbol{Y} = \mathbf{H}\boldsymbol{X} + \mathbf{b} \in \mathbb{R}^M$ *is*

$$\hat{p}_{\boldsymbol{Y}}(\boldsymbol{\xi}) = \hat{p}_{\mathbf{H}\boldsymbol{X}+\mathbf{b}}(\boldsymbol{\xi}) = \hat{p}_{\boldsymbol{X}}(\mathbf{H}^T\boldsymbol{\xi})\mathrm{e}^{\mathrm{j}\mathbf{b}^T\boldsymbol{\xi}}$$

with Fourier-domain variable $\boldsymbol{\xi} \in \mathbb{R}^M$ *and* $\mathbf{H}^T\boldsymbol{\xi} \in \mathbb{R}^N$.

5. *Sum of independent random variables: Let* $\boldsymbol{X}_1 \in \mathbb{R}^N$ *and* $\boldsymbol{X}_2 \in \mathbb{R}^N$ *be two independent random variable with cfs* $\hat{p}_{\boldsymbol{X}_1}$ *and* $\hat{p}_{\boldsymbol{X}_1}$*, respectively. Then, the characteristic function of* $\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2$ *is*

$$\hat{p}_{\boldsymbol{X}_1+\boldsymbol{X}_2}(\boldsymbol{\xi}) = \hat{p}_{\boldsymbol{X}_1}(\boldsymbol{\xi})\hat{p}_{\boldsymbol{X}_2}(\boldsymbol{\xi}).$$

Proof. Property 1 is a slight variation of the Riemann-Lebesgue lemma which states that the Fourier transform of a function $f \in L_1(\mathbb{R}^N)$ is bounded, continuous and decaying at infinity. The relevant bound here is

$$|\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi})| \leq \int_{\mathbb{R}^N} |p_{\boldsymbol{X}}(\boldsymbol{x})|\,|\mathrm{e}^{\mathrm{j}\langle\boldsymbol{\xi},\boldsymbol{x}\rangle}|\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^N} p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \hat{p}_{\boldsymbol{X}}(\mathbf{0}) = 1.$$

Likewise, $p_{\boldsymbol{X}}(\boldsymbol{x})\,|1 - \mathrm{e}^{-\mathrm{j}\langle\boldsymbol{h},\boldsymbol{x}\rangle}|$ is upper bounded by the measurable function $2p_{\boldsymbol{X}}(\boldsymbol{x})$. This allows us to apply Lebesgue's dominated convergence theorem to show that

$$\begin{aligned}\lim_{\boldsymbol{h}\to\mathbf{0}} \big|\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) - \hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}-\boldsymbol{h})\big| &\leq \lim_{\boldsymbol{h}\to\mathbf{0}} \int_{\mathbb{R}^N} p_{\boldsymbol{X}}(\boldsymbol{x})\big|1 - \mathrm{e}^{-\mathrm{j}\langle\boldsymbol{h},\boldsymbol{x}\rangle}\big|\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^N} p_{\boldsymbol{X}}(\boldsymbol{x}) \lim_{\boldsymbol{h}\to\mathbf{0}} \big|1 - \mathrm{e}^{-\mathrm{j}\langle\boldsymbol{h},\boldsymbol{x}\rangle}\big|\mathrm{d}\boldsymbol{x} = 0,\end{aligned}$$

which establishes the continuity of $\hat{p}_{\boldsymbol{X}}$. Finally, $\hat{p}_{\boldsymbol{X}}$ is Hermitian-symmetric simply because it is the Fourier transform of a real-valued function.

Property 2 is the standard Fourier inversion formula with $\boldsymbol{\xi}$ being substituted by $-\boldsymbol{\xi}$.

Property 3 is a direct consequence of the separability of the Fourier kernel if one sets $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ and $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$:

$$\mathrm{e}^{\pm\mathrm{j}\left\langle(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2),(\boldsymbol{x}_1,\boldsymbol{x}_2)\right\rangle} = \mathrm{e}^{\pm\mathrm{j}\left(\langle\boldsymbol{\xi}_1,\boldsymbol{x}_1\rangle+\langle\boldsymbol{\xi}_2,\boldsymbol{x}_2\rangle\right)} = \mathrm{e}^{\pm\mathrm{j}\langle\boldsymbol{\xi}_1,\boldsymbol{x}_1\rangle}\mathrm{e}^{\pm\mathrm{j}\langle\boldsymbol{\xi}_2,\boldsymbol{x}_2\rangle}.$$

A concise derivation of Property 4 is

$$\hat{p}_{\boldsymbol{Y}}(\boldsymbol{\xi}) = \mathbb{E}\{\mathrm{e}^{\mathrm{j}\langle\boldsymbol{\xi},\mathbf{H}\boldsymbol{X}+\mathbf{b}\rangle}\} = \mathbb{E}\{\mathrm{e}^{\mathrm{j}\langle\mathbf{H}^T\boldsymbol{\xi},\boldsymbol{X}\rangle}\mathrm{e}^{\mathrm{j}\langle\boldsymbol{\xi},\mathbf{b}\rangle}\} = \hat{p}_{\boldsymbol{X}}(\mathbf{H}^T\boldsymbol{\xi})\mathrm{e}^{\mathrm{j}\langle\boldsymbol{\xi},\mathbf{b}\rangle}.$$

To prove Property 5, we consider the pooled variable $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ whose joint cf is simply $\hat{p}_{(\boldsymbol{X}_1,\boldsymbol{X}_2)}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \hat{p}_{\boldsymbol{X}_1}(\boldsymbol{\xi}_1)\hat{p}_{\boldsymbol{X}_2}(\boldsymbol{\xi}_2)$ (by Property 3). We then apply the transformation matrix $\mathbf{H} = [\mathbf{I}_M \;\; \mathbf{I}_M] \in \mathbb{R}^{N\times 2N}$ and use Property 4 to get the desired result.

□

The 1D Fourier transform pairs that are relevant to our two introductory examples are $\mathcal{F}\{\delta\}(\xi) = 1$ and $\mathcal{F}\{\mathrm{e}^{-\frac{1}{2}x^2}\}(\xi) = \sqrt{2\pi}\mathrm{e}^{-\frac{1}{2}\xi^2}$. The application of Properties 3 and 4 in Theorem 34 then yields

$$\begin{aligned}\hat{p}_{\mathrm{Const}}(\boldsymbol{\xi}) &= \mathrm{e}^{\mathrm{j}\langle \boldsymbol{\xi}, \boldsymbol{\mu}\rangle} \\ \hat{p}_{\mathrm{Gauss}}(\boldsymbol{\xi}) &= \prod_{n=1}^{N} \mathrm{e}^{-\frac{1}{2}\xi_n^2} = \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{\xi}\|^2}. \end{aligned} \tag{133}$$

Another nice property of characteristic functions is that they are guaranteed to be positive-definite.

Definition 40 (Positive-definite function)**.** *A function $f : \mathbb{R}^N \to \mathbb{C}$ is said to be positive-semidefinite (positive-definite, for short) if*

$$\sum_{m=1}^{M}\sum_{m'=1}^{M} z_m f(\boldsymbol{x}_m - \boldsymbol{x}_{m'})\overline{z}_{m'} \geq 0$$

for every possible choice of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^M$, $z_1, \ldots, z_M \in \mathbb{C}$, and $M \in \mathbb{N}^+$.

Let us note the similarity with the definition of a positive-definite kernel which has two variables instead of one (Definition 2). In fact, we reconcile the two definitions by specifying the kernel $h(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x} - \boldsymbol{y})$ where f is a real-valued, symmetric, positive-definite function (see Appendix A). As it turns out, the latter can be specified as the Fourier transform (or characteristic function) of some symmetric pdf (up to some normalization factor $f(\mathbf{0}) \neq 0$ since $p_{\boldsymbol{X}}(\mathbf{0}) = 1$).

Indeed, if $\hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathcal{F}^*\{p_{\boldsymbol{X}}\}(\boldsymbol{\xi})$ is a valid characteristic function, then

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{n=1}^{M} z_m \overline{z}_n \hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi}_m - \boldsymbol{\xi}_n) \\
&\qquad = \sum_{m=1}^{M}\sum_{n=1}^{M} z_m \overline{z}_n \int_{\mathbb{R}^N} \mathrm{e}^{\mathrm{j}\langle \boldsymbol{\xi}_m - \boldsymbol{\xi}_n, \boldsymbol{x}\rangle} p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_{\mathbb{R}^N} \sum_{m=1}^{M} z_m \mathrm{e}^{-\mathrm{j}\langle \boldsymbol{\xi}_m, \boldsymbol{x}\rangle} \sum_{n=1}^{M} \overline{z}_n \mathrm{e}^{+\mathrm{j}\langle \boldsymbol{\xi}_n, \boldsymbol{x}\rangle}\, p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_{\mathbb{R}} \underbrace{\left| \sum_{m=1}^{M} z_m \mathrm{e}^{-\mathrm{j}\langle \boldsymbol{\xi}_m, \boldsymbol{x}\rangle} \right|^2}_{\geq 0} \underbrace{p_{\boldsymbol{X}}(\boldsymbol{x})}_{\geq 0}\, \mathrm{d}\boldsymbol{x} \geq 0,
\end{aligned}$$

which proves that $\hat{p}_{\boldsymbol{X}}$ is necessarily positive-definite. What is more remarkable is that positive definiteness is also necessary when combined with the analytical properties in the first statement of Theorem 34.

Theorem 35 (Bochner's theorem)**.** *The function $\hat{p}_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{C}$ is a valid characteristic function if and only if it is continuous, Hermitian-symmetric, positive-definite and such that $\hat{p}_{\boldsymbol{X}}(\mathbf{0}) = 1$. This is equivalent to the existence of a unique real-valued function $p_{\boldsymbol{X}} = \mathcal{F}^{*-1}\{\hat{p}_{\boldsymbol{X}}\}$ (the pdf of $\boldsymbol{X}$) such that $p_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^N$ and $\int_{\mathbb{R}^N} p_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$.*

The interest of this theorem is that it provides a comprehensive Fourier-domain criterion for the construction of valid characteristic functions. The second foundational result for the Fourier-based formulation of probability theory is as follows.

Theorem 36 (Lévy's continuity theorem)**.** *Consider a series of N-dimensional random vectors $\boldsymbol{X}_n$ with respective characteristic functions $\hat{p}_{\boldsymbol{X}_n} : \mathbb{R}^N \to \mathbb{C}$. If there exists a function $\hat{p}_{\boldsymbol{X}}$ such that*

$$\lim_{n\to\infty} \hat{p}_{\boldsymbol{X}_n}(\boldsymbol{\xi}) = \hat{p}_{\boldsymbol{X}}(\boldsymbol{\xi})$$

pointwise on $\mathbb{R}^N$ and if, in addition, $\hat{p}_{\boldsymbol{X}}$ is continuous at $\boldsymbol{\xi} = \mathbf{0}$, then $\hat{p}_{\boldsymbol{X}}$ is the characteristic function of some random vector $\boldsymbol{X}$ with pdf $p_{\boldsymbol{X}} = \mathcal{F}^{*-1}\{\hat{p}_{\boldsymbol{X}}\}$. Moreover, $\boldsymbol{X}_n$ converges to $\boldsymbol{X}$ in law, meaning that, for any continuous function $f : \mathbb{R}^N \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\{f(\boldsymbol{X}_n)\} = \mathbb{E}\{f(\boldsymbol{X})\},$$

which is equivalent to the weak convergence of $p_{\boldsymbol{X}_n}$ to $p_{\boldsymbol{X}}$.

In particular, Theorem 36 implies that two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with identical cfs are undistinguishable in law, which justifies the use of Fourier analysis methods. A remarkable aspect in the statement of Theorem 36 is that the continuity of $\hat{p}_{\boldsymbol{X}}$ at $\boldsymbol{\xi} = \mathbf{0}$ is sufficient to ensure continuity on $\mathbb{R}^N$ (see Property 1 in Theorem 34).

E.4 Multivariate Gaussian distributions

An effective way of introducing the multivariate Gaussian distribution is through the construction of its characteristic function. The idea is to generalize (133) by replacing $\|\boldsymbol{\xi}\|^2$ by the more general quadratic form $\boldsymbol{\xi}^T\mathbf{C}\boldsymbol{\xi}$ (where $\mathbf{C}$ is some arbitrary symmetric, positive-definite matrix) and by including an additional offset term.

Definition 41 (Multivariate Gaussian)**.** *The characteristic function of a multivariate Gaussian random variable of dimension N with mean $\boldsymbol{\mu} \in \mathbb{R}^N$ and symmetric positive-definite covariance matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ is*

$$\hat{p}_{\text{Gauss}}(\boldsymbol{\xi}|\boldsymbol{\mu}, \mathbf{C}) = \exp\left(-\tfrac{1}{2}\boldsymbol{\xi}^T\mathbf{C}\boldsymbol{\xi} + \mathrm{j}\boldsymbol{\mu}^T\boldsymbol{\xi}\right). \tag{134}$$

Henceforth, we shall denote the property that a random vector $\boldsymbol{X}$ is multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ by $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$.

The simplest instance of (134) is $\exp(-\frac{1}{2}\|\boldsymbol{\xi}\|^2) = \hat{p}_{\text{Gauss}}(\boldsymbol{\xi}|\mathbf{0}, \mathbf{I}_N)$, which specifies the standardized white Gaussian noise $\boldsymbol{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ of our introductory example. Let us now pick an arbitrary linear transformation matrix $\mathbf{H} \in \mathbb{R}^{N\times N}$. By defining $\boldsymbol{X} = \mathbf{H}\boldsymbol{G} + \boldsymbol{\mu}$ and applying Property 4 in Theorem 34, we obtain a specific instance of (134) with $\mathbf{C} = \mathbf{H}\mathbf{H}^T$. Likewise, we find that the mean vector of $\boldsymbol{X}$ is simply

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}\{\mathbf{H}\boldsymbol{G} + \boldsymbol{\mu}\} = \mathbf{H}\,\mathbb{E}\{\boldsymbol{G}\} + \mathbb{E}\{\boldsymbol{\mu}\} = \boldsymbol{\mu},$$

while its covariance matrix is

$$\mathbf{C}_{\boldsymbol{X}} = \mathbb{E}\{\mathbf{H}\boldsymbol{G}\boldsymbol{G}^T\mathbf{H}^T\} = \mathbf{H}\,\mathbb{E}\{\boldsymbol{G}\boldsymbol{G}^T\}\mathbf{H}^T = \mathbf{H}\mathbf{I}_N\mathbf{H}^T = \mathbf{H}\mathbf{H}^T = \mathbf{C},$$

which confirms the original claim in Definition 41. The described filtering procedure is actually universal since it is possible to factorize any (symmetric) covariance matrix as $\mathbf{H}\mathbf{H}^T = \mathbf{C}$.

In the event where $\mathbf{C}$ is invertible, we can also reverse the procedure by computing $\boldsymbol{G} = \mathbf{H}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, which returns a "whitened" version of $\boldsymbol{X}$. We then exploit the one-to-one linear correspondence between $\boldsymbol{X}$ and $\boldsymbol{G}$, whose pdf is given by (132), to determine the explicit form of $p_{\text{Gauss}}(\boldsymbol{x}|\boldsymbol{\mu}, \mathbf{C})$ by

simple change of variable. Ultimately, we find that the generic form of a N-dimensional multivariate Gaussian probability density function is

$$p_{\text{Gauss}}(\boldsymbol{X}|\boldsymbol{\mu}, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^N |\det(\mathbf{C})|}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right), \tag{135}$$

a formula that obviously requires the covariance matrix $\mathbf{C}$ to be invertible. Likewise, one easily deduces that the component variables X_n of $\boldsymbol{X}$ are univariate Gaussian with mean $\mu_{X_n} = [\boldsymbol{\mu}]_n$ and variance $\text{Var}\{X_n\} = [\mathbf{C}]_{n,n}$.

The bottom line is that the Gaussianity property is invariant to general affine transformations:

Proposition 27. *Consider some fixed matrix $\mathbf{H} \in \mathbb{R}^{N_2 \times N_1}$, an offset vector $\mathbf{b} \in \mathbb{R}^{N_1}$ and some N_1-dimensional Gaussian random vector $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$. Then, $\boldsymbol{X}_2 = \mathbf{H}\boldsymbol{X}_1 + \mathbf{b} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ is multivariate Gaussian as well with*

$$\boldsymbol{\mu}_2 = \mathbf{H}\boldsymbol{\mu}_1 + \mathbf{b}$$
$$\mathbf{C}_2 = \mathbf{H}\mathbf{C}_1\mathbf{H}^T.$$

In particular, for any $\mathbf{u} \in \mathbb{R}^N$, the scalar "projection" $Y = \mathbf{u}^T\boldsymbol{X} = \langle \mathbf{u}, \boldsymbol{X} \rangle$ of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ is a Gaussian random variable with mean $\mu_Y = \mathbf{u}^T\boldsymbol{\mu} = \langle \mathbf{u}, \boldsymbol{\mu} \rangle$ and variance $\sigma_Y^2 = \mathbf{u}^T\mathbf{C}\mathbf{u}$. Conversely, the Gaussianity of $Y = \langle \mathbf{u}, \boldsymbol{X} \rangle$ for any $\mathbf{u} \in \mathbb{R}^N$ implies that $\boldsymbol{X}$ is multivariate Gaussian, which is a high-level property that can also serve as the definition of this class of distributions.

E.5 Gaussian conditional probabilities

To investigate conditional dependencies, we split the Gaussian random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$$

and rewrite the mean and covariance as

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{bmatrix}.$$

The joint pdf $p_{\boldsymbol{X}} = p_{(\boldsymbol{X}_1, \boldsymbol{X}_2)}$ is multivariate Gaussian by assumption. By integrating $p_{\boldsymbol{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ over $\boldsymbol{x}_2$ (resp. $\boldsymbol{x}_1$)—or, equivalently, by using Property 4 in Theorem 34 with $\mathbf{H} = [\mathbf{I}_{N_1} \;\; \mathbf{0}_{N_2}]$ (resp., $\mathbf{H} = [\mathbf{0}_{N_1} \;\; \mathbf{I}_{N_2}]$)—we find

that the marginal distributions $p_{\boldsymbol{X}_1}$ and $p_{\boldsymbol{X}_2}$ are multivariate Gaussian as well, with $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_{11})$ and $\boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_{22})$.

To determine the conditional probabilities, we use Bayes' rule $p(\boldsymbol{x}_1|\boldsymbol{x}_2) = p_{\boldsymbol{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)/p_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$ and perform the relevant but somewhat lengthy algebra to show that the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ is multivariate Gaussian with mean

$$\mathbb{E}\{\boldsymbol{X}_1|\boldsymbol{X}_2\} = \boldsymbol{\mu}_1 + \mathbf{C}_{12}\mathbf{C}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2) \tag{136}$$

and covariance matrix

$$\mathbf{C}_{\boldsymbol{X}_1|\boldsymbol{X}_2} = \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{12}^T.$$

F Basic definitions from measure theory

A measurable space is a set considered together with a corresponding σ-algebra (also called σ-field).

Definition 42 (σ-field)**.** *Let Ω be a set. A collection $\Sigma = \Sigma(\Omega)$ of subsets of Ω is said to be a σ-field (or σ-algebra) of Ω if it satisfies the following properties*

- *Full coverage: $\Omega \in \Sigma$*
- *Closure under complementation: If $E \in \Sigma$, then $\overline{E} = \Omega \backslash E \in \Sigma$*
- *Closure under countable unions: if $E_i \in \Sigma$ with $i \in I$, then $\bigcup_{i\in I} E_i \in \Sigma$ for any countable index set I.*

The axioms in Definition 42 also imply that Σ includes the empty set $\emptyset$ and that it is closed under countable intersections; i.e., $\bigcap_{i\in I} E_i \in \Sigma$.

A subset E of Ω is said to be measurable if it is included in the underlying σ-field. One can therefore also refer to $\Sigma(\Omega)$ as the collection of all measurable sets of Ω.

When the set Ω is a vector space such as $\mathbb{R}$, $\mathbb{R}^N$ or, by extension, some topological function space $\mathcal{X}$, there is a systematic way of of specifying a σ-field by taking $\Sigma(\mathcal{X}) = B(\mathcal{X})$ where $B(\mathcal{X})$ is the Borel algebra of $\mathcal{X}$; that is, the smallest σ-algebra that contains all the open sets (or, equivalently, the closed sets) of $\mathcal{X}$.

A function between two measurable spaces, say Ω and $\mathcal{X}$, is said to be *measurable* if the preimage of each measurable set of $\mathcal{X}$ is measurable.

A measure on Ω is a systematic rule that assigns a non-negative number to each suitable subset E of that set. It is typically used for measuring the size of E or for assigning some probability to it.

Definition 43 (Measure)**.** *Let (Ω, Σ) be a measurable space. Then, the map $\mu : \Sigma \to \mathbb{R}^+$ is called a measure if it satisfies the following properties:*

- *Non-negativity: $\mu(E_k) \geq 0$ for all $E_k \in \Sigma(\Omega)$*
- *Null empty set: $\mu(\emptyset) = 0$*
- *Countable additivity: For all countable collections $\{E_k\}_{k \in S}$ of pairwise disjoint sets of Σ,*

$$\mu\left(\bigcup_{k \in S} E_k\right) = \sum_{k \in S} \mu(E_k)$$

A measure that is defined on the Borel sets $B(\mathcal{X})$ of a topological vector space $\mathcal{X}$ is called a Borel measure. It is a probability measure if $\mu(\mathcal{X}) = 1$.

In the context of probability theory, the set Ω is called the *sample space*. The elements $\omega \in \Omega$ represent the possible outcomes of an experiment. The members $E \subseteq \Omega$ of the σ-field $\Sigma(\Omega)$ are called *events*; these are measurable by definition. The individual points of the sample space are elementary events, which are typically also part of the σ-field $\Sigma(\Omega)$. The final ingredient is the probability measure $\mathscr{P} : \Sigma(\Omega) \to [0, 1]$ which provides the probability law for the random generation mechanism that produces ω (outcome of the experiment): $\text{Prob}(\omega \in E) = \mathscr{P}(E)$.

Bibliography

References

[1] Philippe G Ciarlet. *Linear and nonlinear functional analysis with applications*, volume 130. SIAM, 2013.

[2] S. Johansen. An application of extreme point methods to the representation of infinitely divisible distributions. *Probability Theory and Related Fields*, 5:304–316, 1966.

[3] B. L. S. Prakasa Rao. Infinitely divisible characteristic functionals on locally convex topological vector spaces. *Pacific Journal of Mathematics*, 35(1):221–225, 1970.

[4] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):pp. 522–536, 1938.