Coordinate-free characterization of the multi-variate Gaussian distribution

Steve Cheng

July 3, 2025

1 Introduction

This article introduces the theory of the multi-variate Gaussian (normal) probability distribution. In contrast to most textbooks[1] on probability and statistics (elementary or not), we strive to work in a coordinate-free manner, using the full power of modern linear algebra.

In my own past experience as a university student, linear algebra presented as a theory of vector spaces gets introduced early in the curriculum for a mathematics major — starting in the first year, finishing in the second year of undergraduate study at the latest — yet the teaching of the applications (e.g. physics, statistics) is usually geared to other students not majoring in mathematics, and so do not use the more abstract theory, preferring to work with (list) vectors and matrices.

I have always found that dissatisfying — and partly for that reason had learned a lot of statistics for a long time by re-deriving key results myself. Nowadays, I would argue that even in applications, the geometric perspective underlying vector-space theory is indispensable, in reasoning through situations more complicated than the toy problems of textbooks. Without a rigorous theory, geometric intuition might just be figments of the mind that may lead us astray. And, when we want to compute results, invariably programming a computer to do so, I think it is useful to understand at a deep level the meanings of quantities we want to compute, and to derive the analytics in the most generic form as is possible and practical — thus helping with code re-use — *before* assigning indices and coordinates for the ultimate vector and matrix representations.

This is an article I wish I could have read when I was an undergraduate. Not only because the applications are explained with more elegant techniques (in my not-so-humble opinion), but also it might help to impart intuition behind concepts like adjoints (from linear algebra), or the connection between orthogonality (in certain inner product spaces) and stochastic independence — which often go under-explained in pure-mathematics textbooks too.

Besides linear algebra, the background needed to read this article include basic single-variable calculus, and elementary knowledge of probability theory. Unfortunately, strictly speaking, "elementary" probability theory is not enough, as we depend on certain well-known fundamental facts, like the unique association of probability measures with their characteristic functions, that require serious learning of real analysis and the Kolmogorov axiomatization of probability theory[2] for their development. The reader without this knowledge will just have to treat such facts as given — but should still be able to follow the prose and algebraic manipulations.

[1] [Wichura] is a wonderful exception that has inspired this work.

[2] A rapid yet rigorous development can be found in [Rosenthal], which should be approachable by an undergraduate who has studied rigorous calculus.

2 Basic definition

Let V be a real finite-dimensional vector space[3].

A random variable Z taking values in V is said to be *normal*, or *Gaussian*, if for every linear functional φ, the real-valued random variable φZ has the one-dimensional Gaussian distribution, including the de-generate case that it is a constant.

Thus Z is Gaussian if and only if, for all linear functionals φ, the characteristic function of φZ has the form:

$$f_{\varphi Z}(t) = \mathbb{E}\big[\exp(it\varphi Z)\big] = \exp(i\mu t - \tfrac{1}{2}\sigma^2 t^2)\,.$$

Since any characteristic function f of a real-valued, square-integrable random variable X satisfies:

$$\begin{aligned} f(0) &= 1\,, \\ f'(0) &= \left.\frac{d}{dt}\right|_{t=0} \mathbb{E}\big[\exp(itX)\big] = \mathbb{E}\big[iX\exp(itX)\big] \;|_{t=0} = i\mathbb{E}X\,, \\ f''(0) &= \left.\frac{d^2}{dt^2}\right|_{t=0} \mathbb{E}\big[\exp(itX)\big] = \mathbb{E}\big[(iX)^2\exp(itX)\big] \;|_{t=0} = -\mathbb{E}X^2\,, \end{aligned}$$

by differentiating $f_{\varphi Z}$ we see that its parameters must necessarily satisfy:

$$\begin{aligned} \mu &= \mathbb{E}\big[\varphi Z\big]\,, \\ \mu^2 + \sigma^2 &= \mathbb{E}\big[(\varphi Z)^2\big]\,, \quad \text{i.e. } \sigma^2 = \mathrm{Var}(\varphi Z)\,. \end{aligned}$$

Many elementary works define an n-dimensional Gaussian random variable by taking n independent one-dimensional standard[4] Gaussian variables and assigning each to a coordinate — the resulting (list) vector is then said to be the standard n-dimensional Gaussian. A non-standard Gaussian random variable, i.e. with some non-trivial variance/covariance, would have to defined indirectly by applying a linear transformation to a standard one. But in applications we often just want to specify a random variable to follow a Gaussian law with some variance/covariance. It is awkward to be forced to think that any desired random variable must be driven by some standard Gaussian every time, supposedly constructed inside whatever the probability space is at hand.

Yet other elementary works define an n-dimensional Gaussian through its n-dimensional probability density function! Not only is the formula sometimes an unneeded complication — we do not use it once in this article — that approach totally fails if the Gaussian is "degenerate", i.e. its variance-covariance matrix is singular. Not that degeneracy happens in many applications, but requiring separate analytics for that special case surely is inconvenient.

3 Variance and covariance

In the following, we further assume that V is a real inner product space. Recall that the inner product $\langle , \rangle$ induces a natural isomorphism between V and its dual space V^*: every $\varphi \in V^*$ is of the form $\varphi(u) = \langle u, v\rangle$ for some $v \in V$.

The *covariance* of two random variables X and Y taking values in V is defined to be the symmetric bilinear form:

$$\mathcal{C}_{XY}(u, v) = \mathrm{Cov}\big[\langle u, X\rangle, \langle v, Y\rangle\big]\,, \quad u, v \in V\,.$$

This bilinear form returns the covariance of two scalar-valued variables $\langle u, X\rangle$ and $\langle v, Y\rangle$. Thus it encapsulates the same data as a covariance matrix in the usual non-coordinate-free approach to vector-valued random variables.

[3] The definition in this section need not be restricted to finite-dimensional vector spaces, but all the following sections require finite-dimensionality to work.

[4] Having a mean of 0 and a variance of 1.

The definition of covariance clearly depends on the inner product. However, in finite-dimensional spaces, any two inner products can be related by an invertible linear transformation, so changing the inner product only amounts to a "change of variables". If V is identified with $\mathbb{R}^n$ (being set exactly to $\mathbb{R}^n$ or being isomorphic to $\mathbb{R}^n$), then the inner product usually taken on V is, naturally, the standard inner product[5] of $\mathbb{R}^n$.

The *variance* of a random variable Z is, of course, the covariance of Z versus Z. For brevity, it is often re-cast as a quadratic form with no additional comment:

$$\mathcal{C}(u) = \mathcal{C}_{ZZ}(u,u) = \mathrm{Var}\big[\langle u, Z\rangle\big]\,, \quad u \in V\,.$$

The bi-linear and quadratic forms make explicit the "quadratic" nature of the concepts of covariance and variance.

We have more related definitions. By the natural isomorphism between V and V^*, there exists a vector Cv such that

$$\mathcal{C}_{XY}(u,v) = \langle u, Cv\rangle\,.$$

Since $\mathcal{C}_{XY}(u,v)$ is linear in v, Cv is also linear in v. So, there always exists a linear operator $C\colon V \to V$ such that

$$\mathrm{Cov}\big[\langle u, X\rangle, \langle v, Y\rangle\big] = \langle u, Cv\rangle\,.$$

We call this C the *covariance operator*; its matrix representation under the standard basis of V is the usual "covariance matrix".

In the case that $X = Y = Z$ is the same random variable, C is also called the *variance operator* or *dispersion operator*. Since $\mathcal{C}_{ZZ}(u,v)$ is symmetric in u and v, the variance operator is self-adjoint: $C^* = C$, where C^* denotes the adjoint of C. Also, C is positive semi-definite, for $\mathcal{C}_{ZZ}(u,u) \geq 0$ for all $u \in V$.

In matrix notation, the adjoint is realized as the transpose operation[6], but we will consistently refer to adjoints instead of transposes in this article, because we will be making extensive use of theory from linear algebra, that under a more general context applies to adjoints only but not the plain old transpose.

There is some ambiguity in the terminology, in particular as it relates to the matrix representations. One rarely hears the term "variance matrix" to refer to the matrix representation of the variance operator, since the matrix still contains entries for the covariances of individual components of Z against each other. Instead it is usually referred to as a "covariance matrix" or "variance-covariance matrix". On the other hand, in the coordinate-free setting, the covariances of the components are considered part of the data for the variance/dispersion of a vector-valued random variable.

We make another comment on notation. Many other texts will freely apply the notations Cov and Var on vector-valued random variables directly, writing:

$$\mathrm{Cov}(X,Y)\,, \quad \mathrm{Var}(Z)\,.$$

We will *not* do so in this article, not only to avoid ambiguity, but also because the laws on manipulating vector-valued Cov and Var are not clear[7] at this point. We write Cov and Var only for scalar arguments — in particular, always using the inner product to intermediate vector-valued random variables appearing inside Cov and Var. We do allow vector-valued expectations $\mathbb{E}$ however, since it is easily defined (component-wise), and its "linear" behavior is obvious and easy to understand.

Finally, with regards to covariance, we will occasionally require X and Y to have distinct co-domains, say vector spaces V and W. The coordinate-free definition of covariance is easily re-cast to such situations: the inner product on X and the inner product on Y, appearing as arguments to Cov, will just work on V and W respectively. However, the corresponding concept of the covariance operator C needs a small tweak: it obviously becomes a linear transformation $C : W \to V$. Since the

[5] This matters for matrix and coordinate representations. For example, if the variance operator (defined below) C of a $\mathbb{R}^n$-valued random variable Z is the identity, under the standard inner product that means the coordinate components Z_k, $k = 1, \ldots, n$ have variance 1, and are uncorrelated with each other — rather than weird combinations like $2Z_1 - 5Z_2$.

[6] Only when using an orthonormal basis, and when the vector space is real, not complex.

[7] The rules may be clear to the expert user, but perhaps not to the student for whom the text is intended!

term "(linear) operator" usually refers to a transformation with the co-domain being the same as the domain, we might call $C : W \to V$ instead a *covariance homomorphism*[8].

4 Induced inner product from the variance operator

Given the variance $\mathcal{C}$ of a Gaussian random variable Z, we may derive the exact form of the characteristic function of Z as follows. Recall that such a characteristic function (over $\mathbb{R}^n$) is defined as:

$$f_Z(\xi) = \mathbb{E}\big[\exp(i\langle \xi, Z\rangle)\big]\,, \quad \xi \in \mathbb{R}^n.$$

First assume $\mathbb{E}[Z] = 0$, so that $\mathbb{E}[\varphi Z] = 0$ for all $\varphi \in V^*$. Now consider the characteristic function of φZ where $u \in V$ is the corresponding element to $\varphi \in V^*$. Since Z is Gaussian, by definition $\varphi Z = \langle u, Z\rangle$ is Gaussian, and

$$f_{\varphi Z}(t) = \mathbb{E}\big[\exp(it\langle u, Z\rangle)\big] = \exp(-\tfrac{1}{2}\sigma_u^2 t^2)\,, \quad \sigma_u^2 = \mathbb{E}\big[\langle u, Z\rangle^2\big] = \mathcal{C}(u)\,.$$

In particular set $t = 1$ and $u = \xi$, and we obtain this formula for the characteristic function:

$$f_Z(\xi) = \exp\big(-\tfrac{1}{2}\mathcal{C}(\xi)\big)\,.$$

If $\mathbb{E}[Z] = z_0 \neq 0$, applying the preceding special case to the zero-mean random variable $Z - z_0$ reveals:

$$\begin{aligned} f_Z(\xi) &= \mathbb{E}\big[\exp(i\langle \xi, Z - z_0 + z_0\rangle)\big] = \exp\big(i\langle \xi, z_0\rangle\big)\mathbb{E}[\exp(i\langle \xi, Z - z_0\rangle)\big] \\ &= \exp\big(i\langle \xi, z_0\rangle - \tfrac{1}{2}\mathcal{C}(\xi)\big)\,. \end{aligned}$$

In the preceding section, we remarked that the variance operator C corresponding to $\mathcal{C}$ is self-adjoint and positive semi-definite. These properties let us define:

$$\|\xi\|_C = \sqrt{\langle \xi, C\xi\rangle}\,, \quad \xi \in V\,,$$

as a semi-norm on V, determined by the "covariance structure" $\mathcal{C}$ of Z. With this semi-norm $\|\cdot\|_C$, the characteristic function may be re-written as:

$$f_Z(\xi) = \exp\big(i\langle \xi, z_0\rangle - \tfrac{1}{2}\langle \xi, C\xi\rangle\big) = \exp\big(i\langle \xi, z_0\rangle - \tfrac{1}{2}\|\xi\|_C^2\big)\,.$$

If C is non-singular, $\mathcal{C}(u, v) = \langle u, Cv\rangle$ defines another (positive-definite) inner product on V, and $\|\cdot\|_C$ becomes a norm.

5 Linear transformations

We may also consider the random variable LZ defined by an arbitrary linear transformation $L\colon V \to W$, where $W = \mathbb{R}^m$ with its standard inner product. The characteristic function of LZ would be:

$$\begin{aligned} f_{LZ}(\xi) = \mathbb{E}\big[\exp(-i\langle \xi, LZ\rangle)\big] = \mathbb{E}\big[\exp(-i\langle L^*\xi, Z\rangle)\big] &= f_Z(L^*\xi) \\ &= \exp\big(i\langle L^*\xi, z_0\rangle - \tfrac{1}{2}\|L^*\xi\|_C^2\big) \\ &= \exp\big(i\langle \xi, Lz_0\rangle - \tfrac{1}{2}\langle \xi, (LCL^*)\xi\rangle\big)\,, \end{aligned}$$

where $L^*\colon W \to V$ is the adjoint of L.

Evidently, the variance operator of LZ is LCL^*, which, of course, could have been derived directly from the definition of variance. The transformation of the mean from $\mathbb{E}Z = z_0$ to $\mathbb{E}[LZ] = Lz_0$ is obvious.

[8] *Homomorphism* in the context of linear vector spaces means a linear transformation. For reference, the term *endomorphism* is sometimes heard to refer to a linear operator.

Having an explicit formula for $L^*\colon W \to V$, in coordinate-free form, will be helpful in the following sections. To derive one, let $L\colon V \to W$ be expressed in the form:

$$Lz = \sum_{k=1}^{m} \langle z, v_k \rangle_V \, e_k \,, \quad z \in V \,,$$

where $e_1, \ldots, e_m$ form any orthonormal[9] basis of W, and $v_1, \ldots, v_m \in V$ are determined uniquely by:

$$\langle z, v_k \rangle_V = \langle Lz, e_k \rangle_W \,, \quad \text{for all } z \in V.$$

We distinguish the inner products on V and W by subscripts here, for clarity.

Applying the definition of the adjoint to the preceding equation:

$$\langle z, v_k \rangle_V = \langle z, L^* e_k \rangle_V \,, \quad \text{for all } z \in V,$$

we obtain immediately:

$$L^* e_k = v_k \,. \tag{5.1}$$

And then, more generally, for any $w \in W$:

$$L^* w = L^* \left(\sum_{k=1}^{n} \langle w, e_k \rangle_W \, e_k \right) = \sum_{k=1}^{n} \langle w, e_k \rangle_W \, v_k \,.$$

We may more easily remember this formula by seeing that the vectors v_k and e_j "switch places" in going from L to L^*.

6 Lack of correlation implies independence

The following sections describe two important transformations that generate such orthogonal, independent components.

7 Orthogonal components from spectral decomposition

Since C is self-adjoint, by the spectral theorem (in finite dimensions), it has a complete set of orthonormal eigenvectors $v_1, \ldots, v_n \in V$ (with respect to the standard inner product), with corresponding eigenvalues $\lambda_k \in \mathbb{R}$:

$$C v_k = \lambda_k v_k \,.$$

Also, the eigenvalues are non-negative since C is positive semi-definite.

Assume again that $\mathbb{E}Z = z_0 = 0$. If we next consider the random variables

$$X_k = \langle v_k, Z \rangle \,,$$

[9] We note this derivation works with any inner products assigned to V and W, although we are only concerned with the standard inner products here.

their n-dimensional joint distribution has the following characteristic function:

$$\begin{aligned} f_X(\xi) = \mathbb{E}\left[\exp\left(i\sum_{k=1}^n \xi_k X_k\right)\right] = \mathbb{E}\left[\exp\left(i\sum_{k=1}^n \langle \xi_k v_k, Z\rangle\right)\right] &= f_Z\left(\sum_{k=1}^n \xi_k v_k\right) \\ &= \exp\left(-\frac{1}{2}\left\langle \sum_{j=1}^n \xi_j v_j, C\sum_{k=1}^n \xi_k v_k\right\rangle\right) \\ &= \exp\left(-\frac{1}{2}\sum_{k=1}^n \lambda_k \xi_k^2\right) \\ &= \prod_{k=1}^n \exp\left(-\tfrac{1}{2}\lambda_k \xi_k^2\right). \end{aligned}$$

Because f_X factors into a product of characteristic functions for n instances of the Gaussian distribution, $X_1, \ldots, X_n$ are independent random variables. Their respective variances are

$$\lambda_k = \mathcal{C}(v_k) = \langle v_k, Cv_k\rangle .$$

8 Orthogonal components from inverse mapping, scalar form

There exists another transformation of Z whose components are independent. It will become useful in §10.

We start with a technical set up in case the variance operator C is singular. Intuitively, we just restrict it to a subspace where it is invertible.

To that end, let $CV = W \subseteq V$ be the range of C; if C is singular then $W \neq V$. Restricting the domain of C to W makes it invertible, because W equals the orthogonal complement (with respect to the standard inner product) of the null space of C. The last assertion comes from that fact, for *any* transformation $T : V \to V'$ between finite-dimensional inner product spaces:

$$T^*V = \left(T^{-1}\{0\}\right)^\perp . \tag{8.1}$$

And we substitute in $T = C = T^*$.

The subspace W thus has an inner product induced by C^{-1} restricted to W:

$$\langle v, w\rangle_{C^{-1}} = \langle v, C^{-1}w\rangle , \quad v, w \in W .$$

Without change of notation, we extend the linear transformation $C^{-1}\colon W \to V$ to $C^{-1}\colon V \to V$ by setting C^{-1} to be zero on $W^\perp$. Note that C^{-1} extended[10] this way remains self-adjoint. We continue to write $\langle v, w\rangle_{C^{-1}}$ for vectors $v, w \in V \setminus W$ even though $\langle , \rangle_{C^{-1}}$ may be only a positive semi-definite bilinear form over all of V.

Let $w_1, \ldots, w_m \in W$ be an orthonormal basis, with respect to $\langle , \rangle_{C^{-1}}$. Consider the random variables

$$Y_k = \langle w_k, Z\rangle_{C^{-1}} = \langle C^{-1}w_k, Z\rangle ,$$

Set $L\colon V \to W$ by:

$$Lz = \sum_{k=1}^m \langle C^{-1}w_k, z\rangle e_k ,$$

which implies

$$L^* e_k = C^{-1}w_k ,$$

[10] This extension is just the pseudo-inverse, but we hardly need to invoke the whole theory of pseudo-inverses here.

from eq. 5.1 on page 5.

Assuming $\mathbb{E}Z = 0$, we may expand the characteristic function of the m-dimensional joint distribution of Y_k, in the same manner we did in §7:

$$\begin{aligned} f_Y(\xi) = \mathbb{E}\left[\exp\left(i\sum_{k=1}^m \xi_k Y_k\right)\right] &= \mathbb{E}\left[\exp\left(i\sum_{k=1}^m \langle \xi_k C^{-1} w_k, Z\rangle\right)\right] \\ &= f_Z\left(\sum_{k=1}^m \xi_k C^{-1} w_k\right) \\ &= \exp\left(-\frac{1}{2}\left\langle \sum_{j=1}^m \xi_j C^{-1} w_j, C\sum_{k=1}^m \xi_k C^{-1} w_k\right\rangle\right) \\ &= \exp\left(-\frac{1}{2}\left\langle \sum_{j=1}^m \xi_j w_j, \sum_{k=1}^m \xi_k w_k\right\rangle_{C^{-1}}\right) \\ &= \prod_{k=1}^m \exp\left(-\tfrac{1}{2}\xi_k^2\right). \end{aligned}$$

Again we have a factorization of the characteristic function, except this time the components have unit variance — there is no λ_k in the above equation.

We have shown the random variables $Y_1, \ldots, Y_m$ to be independent.

9 Orthogonal components from inverse mapping, vector form

The "inverse mapping" from the preceding section can be phrased in a more abstract way, more convenient for some analytical work.

Let W be orthogonally decomposed into subspaces M_ℓ, for $1 \le \ell \le p$, with the M_k being chosen in any way as long as they are mutually orthogonal under $\langle , \rangle_{C^{-1}}$:

$$W = M_1 \oplus M_2 \oplus \cdots \oplus M_p\,.$$

If P_{M_ℓ} is the $\langle , \rangle_{C^{-1}}$-orthogonal projection to M_ℓ, then the summands in:

$$Z = P_{M_1}(Z) + P_{M_2}(Z) + \cdots + P_{M_p}(Z)$$

ought to be independent, because they "live in" orthogonal linear subspaces, even if not aligned to the coordinate axes.

In the preceding proof of independence (from §8), we had to work with scalar random variables, i.e. Y_k, so that we can stack them up to form their joint probability law under $\mathbb{R}^m$. And reducing everything to the standard inner product was essential because characteristic functions, i.e. the Fourier transform of probability laws, rely on the Euclidean structure of $\mathbb{R}^n$. Showing the vector-valued random variables $P_{M_\ell}(Z)$ are independent just requires making some transformations. There is nothing deep in the following demonstration, only mildly tedious bookkeeping.

Let the orthonormal basis $w_1, \ldots, w_m \in W$, from §8, be arranged so that the basis vectors for M_1 come first, followed by those for M_2, and so on up to M_p. Formally:

$$w_k \in M_\ell \quad \text{whenever } m_{\ell-1} < k \le m_\ell\,, \quad m_\ell = \sum_{s=1}^{\ell} \dim M_s\,.$$

Define the linear transformation $T\colon \mathbb{R}^m \to W$ by $Te_k = w_k$ for $1 \le k \le m$, where $e_1, \ldots, e_m \in \mathbb{R}^m$ form the orthonormal basis under the standard inner product. Then

$$T(\overline{Y}_\ell) = \sum_{k=m_{\ell-1}}^{m_\ell} \langle w_k, Z\rangle_{C^{-1}}\, w_k = P_{M_\ell}(Z) \in M_\ell\,, \quad \text{for} \quad \overline{Y}_\ell = \sum_{k=m_{\ell-1}}^{m_\ell} Y_k\, e_k\,.$$

With $Y_1, \ldots, Y_m$ being independent, it is clear (from the basic definition[11] of independence of random variables), that $Y_1 e_1, \ldots, Y_m e_m$ are independent also, and so are the vector sums $\overline{Y}_1, \ldots, \overline{Y}_p$ whose summands are disjoint. Then a transformation T applied to each of $\overline{Y}_1, \ldots, \overline{Y}_p$ preserves independence, and, of course, that means $P_{M_1}(Z), \ldots, P_{M_p}(Z)$ are independent.

The reader may be a little puzzled, as this author had been, that the inner product with respect to C^{-1} rather than to C is involved throughout here and in §8 . It is essential; mapping through $\langle , \rangle_C$ does not work. To convince ourselves of that we shall demonstrate a certain relation with the C-orthogonal projections.

We first write down this obvious relation:

$$\langle w_j, w_k \rangle_{C^{-1}} = \langle w_j, C^{-1} w_k \rangle = \langle C^{-1} w_j, C C^{-1} w_k \rangle = \langle v_j, C v_k \rangle = \langle v_j, v_k \rangle_C \,, \quad w_k = C v_k \,,$$

so that the vectors $w_1, \ldots, w_m$ are C^{-1}-orthonormal if and only if $v_1, \ldots, v_m$ are C-orthonormal. We are led to define the map the subspaces by:

$$M'_\ell = C^{-1} M_\ell = \operatorname{span} \left\{ v_k \colon m_{\ell-1} < k \leq m_\ell \right\} ,$$

and take the orthogonal projections $Q_{M'_\ell} \colon W \to W$ to M'_k under inner product $\langle , \rangle_C$.

Let us try to write P_{M_ℓ} in terms of $Q_{M'_\ell}$:

$$P_{M_\ell}(z) = \sum_{k=m_{\ell-1}}^{m_\ell} \langle z, w_k \rangle_{C^{-1}} \, w_k = \sum_{k=m_{\ell-1}}^{m_\ell} \left\langle z, C^{-1} C v_k \right\rangle C v_k = \sum_{k=m_{\ell-1}}^{m_\ell} \left\langle C^{-1} z, v_k \right\rangle_C C v_k \,.$$

More concisely:

$$P_{M_\ell} = C Q_{M'_\ell} C^{-1} \,.$$

This strange formula happens to have an interpretation, which requires a digression. We know that any orthogonal projection is self-adjoint — when the adjoint is based on the same inner product as on the projection, of course. What if we take an adjoint based on a different inner product? Then we need a "change of variables" formula for the adjoint. We derive it now for an arbitrary transformation $T \colon V \to U$ between inner product spaces, which will prove useful later too.

Let $T^* \colon U \to V$ and $T^{\circledast} \colon U \to V$ denote the adjoint of T with respect to the inner products $\langle , \rangle$ and $\langle , \rangle_C$ respectively. Then starting from the definition of adjoints, for all $u \in U$ and $v \in V$ we have:

$$\langle T^{\circledast} u, v \rangle_C = \langle u, T v \rangle_C = \langle u, C T v \rangle = \langle C u, T v \rangle = \langle T^* C u, v \rangle = \langle C^{-1} T^* C u, v \rangle_C \,.$$

So we find that:

$$T^{\circledast} = C^{-1} T^* C \,, \quad \text{or} \quad T^* = C T^{\circledast} C^{-1} \,. \tag{9.1}$$

Setting $T = Q_{M'_\ell} = T^{\circledast}$, we find that:

$$P_{M_\ell} = Q^*_{M'_\ell} \,.$$

And the random variable Z can be decomposed as:

$$Z = Q^*_{M'_1}(Z) + Q^*_{M'_2}(Z) + \cdots + Q^*_{M'_p}(Z) \,,$$

with the p random variables on the right being stochastically independent.

There is no real advantage in decomposing with $Q^*_{M'_\ell}$ versus P_{M_ℓ}, since they are exactly the same transformation. We display $Q^*_{M'_\ell}$ only as a theoretical curiosity.

[11] $\Pr(Y_1 \in B_1 \,, Y_2 \in B_2, \cdots, Y_m \in B_m) = \prod_{k=}^{m} \Pr(Y_k \in B_k)$ for all Borel sets B_k on the corresponding codomain of Y_k.

10 Application: linear regression models

As an application, consider the linear regression model, operating on data represented as a random point in V:

$$Y = K\beta + \sigma\varepsilon\,.$$

We take ε to be a V-valued Gaussian random variable with zero mean, and with non-singular variance operator $C\colon V \to V$. The parameter vector β lives in an m-dimensional real vector space W, while $K\colon W \to V$ is an injective linear transformation, and $\sigma > 0$ is a real scalar.

Y is called the *dependent variable* in statistics. Typically, the components of Y represent individual data points, coming from experiments or observations. Any independent "inputs" or variable that Y may or is thought to depend on are encoded in linear transformation K. Its matrix representation is often called the *design matrix*. The standard deviation σ is separated out from the "correlation structure" C, so that the former can be estimated as part of doing the linear regression, while the latter must simply be assumed as part of the model specification. In most applications, $C = I$ is the identity operator, i.e. the observations are assumed to be independent.

Given the observed data, i.e. a realization of the random variable Y, we would like to assume the above parametric model governing Y, though the parameters β and σ are unknown and must be estimated.

A usual choice is to estimate β by minimizing the "sum of squares":

$$\big\|Y - K\widetilde{\beta}\big\|^2_{C^{-1}}\,, \quad \text{over } \widetilde{\beta} \text{ taking values in } W.$$

Abstractly, the solution for $K\widetilde{\beta}$ is the orthogonal projection P_M of Y onto the image $M = KW$, with respect to the inner product induced by C^{-1}. In most applications, C is the identity operator, so $\|Y - K\widetilde{\beta}\|^2_{C^{-1}}$ weights each squared residual equally. If C is not the identity, then $\|Y - K\widetilde{\beta}\|^2_{C^{-1}}$ weights the principal components in inverse proportion to their intrinsic variances, so each component contributes "fairly".

Observe that $I - P_M$ is the orthogonal projection to the orthogonal complement $M^\perp$ to M, and

$$Y = P_M Y + (I - P_M)Y$$

for all values of Y. Defining the random variable $\widehat{\beta} = K^{-1}P_M Y$, we find:

$$\begin{aligned}
\big\|Y - K\widetilde{\beta}\big\|^2_{C^{-1}} &= \big\|P_M Y - K\widetilde{\beta} + (I - P_M)Y\big\|^2_{C^{-1}} \\
&= \big\|P_M(Y - K\widetilde{\beta}) + (I - P_M)Y\big\|^2_{C^{-1}} \\
&= \big\|P_M(Y - K\widetilde{\beta})\big\|^2_{C^{-1}} + \big\|(I - P_M)Y\big\|^2_{C^{-1}} \\
&= \big\|K(\widehat{\beta} - \widetilde{\beta})\big\|^2_{C^{-1}} + \big\|(I - P_M)Y\big\|^2_{C^{-1}} \\
&\geq \big\|(I - P_M)Y\big\|^2_{C^{-1}}\,,
\end{aligned}$$

with equality occurring if and only if $\widetilde{\beta} = \widehat{\beta}$. This justifies $\widehat{\beta}$ being called the "least-squares estimator" of β. It is an unbiased estimator, for:

$$\mathbb{E}\big[K\widehat{\beta}\big] = \mathbb{E}\big[P_M Y\big] = P_M(\mathbb{E}Y) = P_M(K\beta) = K\beta\,.$$

Also, $\sigma^{-2}\big\|Y - K\widehat{\beta}\big\|^2_{C^{-1}}$ is stochastically independent of $\widehat{\beta}$, and has the χ^2 (chi-squared) distribution with $n - m$ degrees of freedom. To see the first assertion, let $v_1, \ldots, v_m$ be an orthonormal basis (with respect to C^{-1}) for the subspace $M \subseteq V$, and $v_{m+1}, \ldots, v_n$ be an orthonormal basis for $M^\perp \subseteq V$. We write the orthogonal projections explicitly:

$$K\widehat{\beta} = P_M Y = \sum_{k=1}^{m} \langle v_k, Y\rangle_{C^{-1}} v_k\,, \quad Y - K\widehat{\beta} = (I - P_M)Y = \sum_{k=m+1}^{n} \langle v_k, Y\rangle_{C^{-1}} v_k\,.$$

The Gaussian random variables $\langle \sigma^{-1} v_k, Y\rangle_{C^{-1}}$ are independent by the result of §8, applied to the random variable Y which has covariance $\sigma^2 C$. So, $K\widehat{\beta}$ and $Y - K\widehat{\beta}$, being weighted sums on disjoint $\langle v_k, Y\rangle_{C^{-1}}$, must be independent too.

Additionally, by the same result of §8,

$$\frac{1}{\sigma^2}\|Y - K\widehat{\beta}\|^2_{C^{-1}} = \sum_{k=m+1}^{n} \langle \sigma^{-1} v_k, Y\rangle^2_{C^{-1}}$$

is a sum of $n - m$ independent Gaussian random variables of variance 1. Provided that the means of those Gaussians are zero, the sum is distributed as $\chi^2(n-m)$. That $\langle \sigma^{-1} v_k, Y\rangle_{C^{-1}}$ does have zero mean[12] for $k > m$ comes from:

$$\begin{aligned}\mathbb{E}\big[\langle v_k, Y\rangle_{C^{-1}}\big] &= \mathbb{E}\left[\big\langle v_k, K\widehat{\beta}\big\rangle_{C^{-1}} + \big\langle v_k, Y - K\widehat{\beta}\big\rangle_{C^{-1}}\right] \\ &= \mathbb{E}\left[\big\langle v_k, P_M Y\big\rangle_{C^{-1}}\right] + \big\langle v_k, \mathbb{E}\big[Y - K\widehat{\beta}\big]\big\rangle_{C^{-1}} = 0 + 0\,.\end{aligned}$$

Finally, the estimate of σ^2 from the data is, naturally:

$$\widehat{\sigma}^2 = \frac{1}{n-m}\|Y - K\widehat{\beta}\|^2_{C^{-1}}\,.$$

It follows immediately from the mean of a $\chi^2(n-m)$ distribution, which is simply $n - m$, that this estimator is unbiased: $\mathbb{E}[\widehat{\sigma}^2] = \sigma^2$. We remark that many elementary texts on linear regression fail to explain intuitively why the "sum of squares" needs to be divided by $n - m$. In our geometric approach, the subtraction of m clearly comes from the m "degrees of freedom" (number of dimensions of randomness) being moved from the span $M^\perp$ of the residuals, into the span M of the parameter estimators.

We will end this section by noting that the mean and variance of $\widehat{\sigma}^2$ could be computed from the covariance C directly, without assuming ε has a normal distribution. Also the estimators $K\widehat{\beta}$ and $\widehat{\sigma}^2$ remain uncorrelated if ε is not normal. So the linear regression model could still be applied (in a more limited way) to observations with non-normal errors.

11 Application: expectation conditional on a linear transform

Let X be a V-valued Gaussian variable, and $A\colon V \to W$ be a surjective linear transformation to a (smaller) vector space W. We wish to compute the conditional expectation:

$$\mathbb{E}[X \mid Y]\,, \quad Y = AX\,.$$

To tackle this problem, imagine we want to "solve" for $x \in V$ in the equation $Ax = y$, given $y \in W$. The infinitely many solutions can be characterized as belonging to one certain coset of the null space $A^{-1}\{0\}$.

As in the derivation of the linear regression model, we want to split X into two components, one component being essentially the same as AX, and the other component orthogonal to the first. Then the expectation of the second component conditional on AX should be equal to the unconditional expectation, since the two components are independent.

[12] In applications to empirical data, the design matrix K is almost always required to have a column of ones, i.e. the image of K spans the vector $(1, \ldots, 1)$, so that one of the parameters being estimated (inside the parameter vector β) is the sample mean. Should the sample mean be not included, then the "residual sum of squares" would not be an unbiased estimator of the variance. There is no contradiction to what we are proving here, because the necessary hypothesis is already implicitly embedded in our abstract model specification — when we say that ε must have zero mean, that implies $\mathbb{E}Y = K\beta$.

In light of these observations, and eq. 8.1 on page 6, we consider the linear subspace M to be the range of the adjoint of A, not of A itself[13]. Since X may have a non-trivial variance operator C, the adjoint of A should be taken with respect to the $\langle , \rangle_{C^{-1}}$ inner product, assuming C is non-singular. We will denote such an adjoint by $A^{\circledast}$, to disambiguate it from the adjoint A^* with respect to the standard inner product; we will be working with both in this section.

So V is decomposed as the direct sum

$$V = M \oplus M^{\perp}\,, \quad M = A^{\circledast}W\,, \quad M^{\perp} = A^{-1}\{0\}\,,$$

with the orthogonal complement taken with respect to $\langle , \rangle_{C^{-1}}$. Then by the result of §9, our Gaussian random variable is also decomposed:

$$X = P_M X + P_{M^{\perp}} X = P_M X + (I - P_M)X$$

into independent V-valued random variables, generated through the orthogonal projections P_M and $P_{M^{\perp}}$, to M and $M^{\perp}$ respectively.

Now consider any $x \in V$. It has the representation:

$$x = P_M(x) + P_{M^{\perp}}(x) = A^{\circledast}w + z\,, \quad \text{for some } w \in W,\ z \in V \text{ such that } Az = 0\,.$$

We multiply this equation by A on the left to relate it to (a given value of) $Ax = y$:

$$Ax = AA^{\circledast}w\,.$$

The operator $AA^{\circledast}\colon W \to W$ is invertible: for $A^{\circledast}W$ is orthogonal to $A^{-1}\{0\}$, so for any $w \in W$, the vector $v = A^{\circledast}w \in A^{\circledast}W$ either has $Av \neq 0$ or $v = 0$. The latter case occurs only if $w = 0$ because $A^{\circledast}$ is injective, which in turn follows from applying eq. 8.1 on page 6 with $T = A^{\circledast}$, together with our assumption that A is surjective.

Hence, we may indeed solve:

$$\begin{aligned} w &= (AA^{\circledast})^{-1}Ax\,. \\ P_M(x) = A^{\circledast}w &= A^{\circledast}(AA^{\circledast})^{-1}Ax. \end{aligned} \tag{11.1}$$

Substituting $x = X$, we obtain:

$$P_M X = A^{\circledast}(AA^{\circledast})^{-1}AX = A^{\circledast}(AA^{\circledast})^{-1}Y\,.$$

This equation shows $P_M X$ is a function of Y. Multiplying the same equation by A on the left, we have $AP_M X = Y$, thereby showing, in the other direction, that Y is a function of $P_M X$. So conditioning an expectation on Y is the same[14] as conditioning on $P_M X$. We can thus calculate:

$$\begin{aligned} \mathbb{E}[X \mid Y] = \mathbb{E}\big[X \mid P_M X\big] &= \mathbb{E}\big[P_M X + P_{M^{\perp}} X \mid P_M X\big] \\ &= P_M X + \mathbb{E}\big[P_{M^{\perp}} X\big] \\ &= P_M X + P_{M^{\perp}}\big(\mathbb{E}X\big) \\ &= P_M X + \big(I - P_M\big)\big(\mathbb{E}X\big) \\ &= \mathbb{E}X + P_M\big(X - \mathbb{E}X\big) \\ &= \mathbb{E}X + A^{\circledast}(AA^{\circledast})^{-1}(Y - \mathbb{E}Y)\,. \end{aligned}$$

[13] Using the range of $A\colon V \to W$ would not even make any sense, since the values of X live in V, not in W. On the other hand, the adjoint $A^{\circledast}\colon W \to V$ does have the correct co-domain. (In the matrix formulation of this and similar problems, the author used to often get confused as to which transformation matrix should be transposed: the one that multiplies the covariance matrix on the left, or the one on the right? If we assign *distinct* vector spaces to the different roles played by the vectors, even if they are the same dimension, then there will always be only one answer that fits — making the formulas a little easier to remember.)

[14] Y and $P_M X$ generate the same σ-algebra in the Kolmogorov definition of conditional probability.

In practical computations, we can use the "change of variables" formula, eq. 9.1 on page 8, to express $A^{\circledast}$ in terms of A^* and C. It reads[15] : $A^{\circledast} = CA^*C^{-1}$. So:

$$\begin{aligned}\mathbb{E}[X \mid Y] &= \mathbb{E}X + CA^*C^{-1}(ACA^*C^{-1})^{-1}(Y - \mathbb{E}Y) \\ &= \mathbb{E}X + CA^*(ACA^*)^{-1}(Y - \mathbb{E}Y)\,, \quad Y = AX\,.\end{aligned}$$

This formula is easier to remember if we observe that ACA^* is the variance operator of Y, while CA^* is the covariance homomorphism of Y versus X. In an obvious notation:

$$\begin{aligned}\mathbb{E}[X \mid Y] &= \mathbb{E}X + C_{X,Y}\,C_Y^{-1}(Y - \mathbb{E}Y)\,. \\ C_Y &: W \to W \quad \text{defined by } \langle w, C_Y w\rangle = \operatorname{Var}\langle w, Y\rangle\,. \\ C_{X,Y} &: W \to V \quad \text{defined by } \langle v, C_{X,Y} w\rangle = \operatorname{Cov}\big[\langle v, X\rangle, \langle w, Y\rangle\big]\,.\end{aligned} \tag{11.2}$$

These formulas may be familiar when Y is scalar-valued, meaning C_Y^{-1} is just scalar division by $\operatorname{Var} Y$.

If X is not assumed normal, we get a weaker result that is still useful to know. Without loss of generality[16], assume $\mathbb{E}X = 0$ and $\mathbb{E}Y = 0$. Let $B_0 = C_{X,Y}\,C_Y^{-1}$ be the linear transformation for "predicting" X from the input condition Y. This predictor is linear even though, in general, $\mathbb{E}[X \mid Y]$ need not be linear in Y. But we can demonstrate, among all *linear*[17] predictors $B : W \to V$, the predictor B_0 is the best at minimizing the variance of the error in prediction, in every direction $v \in V$. Stated in this abstract form, the random variable Y can be arbitrary (as long as it has finite variance) and does not have to be a linear transformation of X!

Fix $v \in V$ and let $B : W \to V$ vary. The variance of error in direction v can be expanded like so:

$$\begin{aligned}\operatorname{Var}\langle v, X - BY\rangle &= \operatorname{Cov}\big[\langle v, X - BY\rangle, \langle v, X - BY\rangle\big] \\ &= \operatorname{Var}\langle v, X\rangle - 2\operatorname{Cov}\big[\langle v, X\rangle, \langle v, BY\rangle\big] + \operatorname{Var}\langle v, BY\rangle \\ &= \operatorname{Var}\langle v, X\rangle - 2\operatorname{Cov}\big[\langle v, X\rangle, \langle B^*v, Y\rangle\big] + \operatorname{Var}\langle B^*v, Y\rangle \\ &= \langle v, Cv\rangle - 2\langle v, C_{X,Y}B^*v\rangle + \langle B^*v, C_Y B^*v\rangle \\ &= \langle v, v\rangle_C - 2\langle C_Y^{-1}C_{X,Y}^* v, B^*v\rangle_{C_Y} + \langle B^*v, B^*v\rangle_{C_Y} \\ &= \big\|v\big\|_C^2 - 2\langle B_0^*v, B^*v\rangle_{C_Y} + \big\|B^*v\big\|_{C_Y}^2\,.\end{aligned}$$

The last expression is a quadratic form in B^*v. We can "complete the square" on it in analogy to scalar quadratic polynomials:

$$\operatorname{Var}\langle v, X - BY\rangle = \big\|v\big\|_C^2 + \big\|B^*v - B_0^*v\big\|_{C_Y}^2 - \big\|B_0^*v\big\|_{C_Y}^2 \geq \big\|v\big\|_C^2 - \big\|B_0^*v\big\|_{C_Y}^2\,.$$

The lower bound is attained for all $v \in V$ when $B = B_0$, as claimed.

12 Application: Kalman filtering

The *Kalman filter* iteratively estimates a sequence of random variables X_1,X_2,..., taking values in a real vector space, given a known affine recursive relation between X_j and X_{j+1}, along with a separate sequence of observations $Y_1, Y_2, \ldots$ whose individual elements X_j are linearly derived from the corresponding X_j. The random variables X_j are interpreted as representing the hidden state of some

[15] In applying eq. 9.1 for $T = A$ here, C must swap places with C^{-1} there, because $T^{\circledast}$ there denotes the adjoint taken with C, while the adjoint is taken with C^{-1} here.

[16] This assumption gets rid of the additive constants involving $\mathbb{E}X$ and $\mathbb{E}Y$ in the predictors B and B_0. The constants can always be added back, since doing so does not affect $\operatorname{Var}\langle v, X - BY\rangle$ below.

[17] For any random variable X in L^2 (i.e. with finite variance), $\mathbb{E}[X \mid Y]$ has the much stronger property that it minimizes the variance of prediction error across *all* L^2 functions $g(Y)$ of Y. That immediately follows from the well-known identity: $\operatorname{Var} Z = \operatorname{Var}\big(\mathbb{E}[Z \mid Y]\big) + \mathbb{E}[\operatorname{Var}(Z \mid Y)]$ for real-valued Z, and substituting $Z = \langle v, X - g(Y)\rangle$. Actually, $\mathbb{E}[X \mid Y]$ can be realized as a certain orthogonal projection of Z, in an infinite-dimensional Hilbert space.

system at times $j = 1, 2, \ldots$. We can only observe some transformation or projection Y_j and must try to find out what the realizations X_j are. At the same time, the evolution of both X_j and Y_j are subject to random "noise" which may be modelled as Gaussian[18].

Let us begin by establishing the mathematical notation more precisely.

- Let $X_1, X_2, \ldots$ be random variables taking values in a finite-dimensional inner product space V. They represent the *hidden state* of the system at increasing time points.
- Let $X_0 = x_0 \in V$ be some known or assumed initial state of the system.
- Let X_j evolve from X_{j-1} according to:

  $$X_j = \Psi_j X_{j-1} + \mu_j + \delta_j\,, \quad j = 1, 2, \ldots,$$

  where $\Psi_j \colon V \to V$ is a known linear operator, $\mu_j \in V$ is a known perturbation or introduced force on the system, and δ_j is V-valued Gaussian noise with zero mean and known variance operator $Q_j \colon V \to V$. The random variables δ_j model the *deviation or noise in the underlying processes* driving X_j. The transformation Ψ_j models how the hidden state evolves as time goes by.
- Let Y_j be random variables taking values in a finite-dimensional inner product space W. They represent the *observed state* of the system.
- Y_j are defined by:

  $$Y_j = A_j X_j + \varepsilon_j\,,$$

  for known linear transformations $A_j \colon V \to W$, and W-valued Gaussian noise ε_j with zero mean and known variance operator $R_j \colon W \to W$. The random variables ε_j model the *error or noise in observing or measuring* the state of the system.
- To concisely refer to what is known about the system at time j, we define $\mathcal{F}_j$ to be the σ-algebra generated by $Y_1, \ldots, Y_j$. For $\mathcal{F}_0$, set it to the σ-algebra consisting only of $\emptyset$ and its complement, representing the trivial information known at the start. These σ-algebras are increasing with time: $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$.
- The random noises δ_j and ε_j shall be mutually independent. Thus, in particular δ_j and ε_j are independent of $\mathcal{F}_{j-1}$.

To give a concrete example, Kalman filtering might be applied in *navigation by dead-reckoning*. Imagine we have a vehicle we must navigate. We know our spatial position & velocity x_0 at the start, but must continually estimate our position & velocity X_j as we make navigational inputs, with incremental effect μ_j on the vehicle, while we might only be able to observe our current velocity Y_j (so A_j is a projection of X_j that drops the component of absolute position). The random variables δ_j represent error or deviations in physically controlling the vehicle or other unaccounted forces (e.g. friction) acting on the vehicle, while ε_j represent measurement error of the velocity of the vehicle.

We note that the Kalman filter can be generalized slightly to allow quantities like u_j to be random, or to be dependent on preceding observations, and for Y_j to incorporate random/affine perturbations as well. If their uncertainty can be fully resolved by $\mathcal{F}_{j-1}$ (at time $j-1$), we can treat them as known without affecting anything else. For notational simplicity, we will not detail such generalization further in this section.

Under our more abstract presentation of the mathematics, we aim to explicitly display a formula for the conditional expectation

$$\widehat{X}_j = \mathbb{E}\big[X_j \mid \mathcal{F}_j\big]\,, \quad j = 1, 2, \ldots,$$

[18] Actually, normality need not be assumed. The results we use from §11 have been shown, towards the end of that section, to apply to non-Gaussian probabilities as well. Without assuming a Gaussian law of errors, the Kalman filter computes a best least-squares estimator.

which we take as the *(best) (linear) estimator* of X_j given $\mathcal{F}_j$. The derivation turns out to be a straightforward, if notationally-heavy, application of the results of §11.

To begin, we fix a value for j and consider the random variables $X_1, X_2, \ldots, X_j$ and $Y_1, \ldots, Y_j$ under the probability law[19] *after conditioning under* $\mathcal{F}_{j-1}$. Under such a probability law, conditioning any expectations, variances and covariances further on the random variable Y_j will be equivalent to conditioning on $\mathcal{F}_j$ originally. With this set-up, eq. 11.2 on page 12 can be applied to expand:

$$\widehat{X}_j = \mathbb{E}\big[X_j \mid \mathcal{F}_j\big] = \mathbb{E}\big[X_j \mid \mathcal{F}_{j-1}\big] + C_{XY,j} C_{Y\,j}^{-1} \big(Y_j - \mathbb{E}\big[Y_j \mid \mathcal{F}_{j-1}\big]\big)$$

with:

- $C_{Y,j}(\omega)\colon W \to W$ being the variance operator[20] of Y_j conditional on $\mathcal{F}_{j-1}$, defined on test vectors $w \in W$ by
$$\langle w, C_{Y,j} w\rangle = \mathrm{Var}\big[\langle w, Y_j\rangle \mid \mathcal{F}_{j-1}\big]\,; \quad \text{and}$$
- $C_{XY,j}(\omega)\colon W \to V$ being the covariance homomorphism of X_j versus Y_j conditional on $\mathcal{F}_{j-1}$, defined on test vectors $v \in V$, $w \in W$ by
$$\langle v, C_{XY,j} w\rangle = \mathrm{Cov}\big[\langle v, X_j\rangle, \langle w, Y_j\rangle \mid \mathcal{F}_{j-1}\big]\,.$$

Like near the end of §11, we give a short name, B_j, to the predictive transformation:

$$B_j = C_{XY,j} C_{Y,j}^{-1}\,, \tag{12.1}$$

which is often called the *Kalman gain*[21].

Then, continuing to expand on $\widehat{X}_j$, we have:

$$\begin{aligned}
\widehat{X}_j &= \mathbb{E}\big[X_j \mid \mathcal{F}_{j-1}\big] + B_j\big(Y_j - \mathbb{E}\big[A_j X_j + \varepsilon_j \mid \mathcal{F}_{j-1}\big]\big) \\
&= \mathbb{E}\big[X_j \mid \mathcal{F}_{j-1}\big] + B_j\big(Y_j - A_j \mathbb{E}\big[X_j \mid \mathcal{F}_{j-1}\big]\big) \\
&= \big(I - B_j A_j\big)\mathbb{E}\big[X_j \mid \mathcal{F}_{j-1}\big] + B_j Y_j \\
&= \big(I - B_j A_j\big)\mathbb{E}\big[\Psi_j X_{j-1} + \mu_j + \delta_j \mid \mathcal{F}_{j-1}\big] + B_j Y_j \\
&= \big(I - B_j A_j\big)\big(\Psi_j \widehat{X}_{j-1} + \mu_j\big) + B_j Y_j\,.
\end{aligned} \tag{12.2}$$

It remains to exhibit computable formulas for $C_{Y,j}$ and $C_{XY,j}$.

We tackle $C_{Y,j}$ first:

$$\begin{aligned}
\langle w, C_{Y,j} w\rangle &= \mathrm{Var}\big[\langle w, Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \mathrm{Var}\big[\langle w, A_j X_j + \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \mathrm{Var}\big[\langle w, A_j X_j\rangle \mid \mathcal{F}_{j-1}\big] + \mathrm{Var}\big[\langle w, \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big]\,.
\end{aligned} \tag{12.3}$$

The last term on eq. 12.3 is easy to recognize. $\mathrm{Var}\big[\langle w, \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big]$ is the same as $\mathrm{Var}\langle w, \varepsilon_j\rangle$ because ε_j is completely independent of $\mathcal{F}_{j-1}$. And $\mathrm{Var}\langle w, \varepsilon_j\rangle$, by definition, is the variance of ε_j evaluated (as a quadratic form) at test vector w. Expressed in terms of the variance operator R_j we have:

$$\mathrm{Var}\big[\langle w, \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big] = \mathrm{Var}\langle w, \varepsilon_j\rangle = \langle w, R_j w\rangle\,.$$

[19] Since $\mathcal{F}_{j-1}$ is generated by a finite set of random variables taking values in a finite-dimensional space, a conditional probability measure consistent with the Kolmogorov theory of probability can be readily constructed, without the complications that commonly ensue from filtrations that cannot be finitely generated.

[20] Conditional variances are *random variables* (measurable with respect to $\mathcal{F}_{j-1}$ or $\mathcal{F}_j$ in our case) under the Kolmogorov theory of probability, so $C_{Y,j}$ is really a random variable whose realization $C_{Y,j}(\omega)$, for each sample point ω in the probability space, is a linear operator $W \to W$. Thus we would be formally wrong to write the intuitive notation $C_{Y,j}\colon W \to W$. However, as with most works dealing with probability theory, ω is mostly irrelevant and will be suppressed in our notation. These comments apply similarly to the objects $C_{XY,j}$, S_j and $\widehat{S}_j$ introduced later in this section.

[21] Because $B_j(Y_j - \mathbb{E}[Y_j \mid \mathcal{F}_{j-1}])$ is added to the "predicted" value $\mathbb{E}[X_j \mid \mathcal{F}_{j-1}]$ (from before observing Y_j) to obtain the new estimate $\widehat{X}_j$. See immediately below.

On the other hand, the first term on eq. 12.3 requires expanding X_j by its recurrence relation with X_{j-1}, like inside $\widehat{X}_j$ earlier:

$$\begin{aligned}\operatorname{Var}\big[\langle w, A_j X_j\rangle \mid \mathcal{F}_{j-1}\big] &= \operatorname{Var}\big[\langle w, A_j(\Psi_j X_{j-1} + \mu_j + \delta_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Var}\big[\langle w, A_j\Psi_j X_{j-1}\rangle \mid \mathcal{F}_{j-1}\big] + \operatorname{Var}\big[\langle w, A_j\delta_j\rangle \mid \mathcal{F}_{j-1}\big]\,.\end{aligned}$$

The last term above may be recognized as analogous to $\operatorname{Var}\big[\langle w, \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big]$, but with $A_j\delta_j$ replacing ε_j; it is obviously related to the variance operator Q_j.

Unfortunately, our insistence in this document on only passing scalar arguments to $\operatorname{Var}[\cdot]$ obscures the simple concept behind the last equation: what happens to the variance (operator) of a vector-valued random variable after applying linear transformations. We already know the answer from §5 — compose the original variance operator on the left by the transformation, and on the right by its adjoint — and do not need to repeat the derivations in detail.

We pause our analysis on eq. 12.3 to introduce the following simplifying notation for the conditional variances of the hidden state — which already have been seen lurking so far.

- For $j = 1, 2, \ldots$, let the (conditional variance) operator $S_j(\omega)\colon V \to V$ be defined by:

  $$\langle v, S_j v\rangle = \operatorname{Var}\big[\langle v, X_j\rangle \mid \mathcal{F}_j\big]\,, \quad v \in V\,.$$

  The special case S_0 is defined as the identically zero transformation, which is consistent with the above formula with j set to 0.

- For $j = 1, 2, \ldots$, let the (conditional variance) operator $\widehat{S}_j(\omega)\colon V \to V$ be defined by:

  $$\langle v, \widehat{S}_j v\rangle = \operatorname{Var}\big[\langle v, X_j\rangle \mid \mathcal{F}_{j-1}\big]\,, \quad v \in V\,.$$

The latter, $\widehat{S}_j$, exactly represents the variance of X_j conditional on $\mathcal{F}_{j-1}$, i.e. letting X_j evolve from X_{j-1} according to the established law but without having observed Y_j yet. We may find a simple formula for it, proceeding as follows:

$$\begin{aligned}\langle v, \widehat{S}_j v\rangle &= \operatorname{Var}\big[\langle v, \Psi_j X_{j-1} + \mu_j + \delta_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Var}\big[\langle v, \Psi_j X_{j-1} + \mu_j\rangle \mid \mathcal{F}_{j-1}\big] + \operatorname{Var}\big[\langle v, \delta_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Var}\big[\langle v, \Psi_j X_{j-1}\rangle \mid \mathcal{F}_{j-1}\big] + \operatorname{Var}\langle v, \delta_j\rangle\,.\end{aligned}$$

The first line involves the recurrence relation of X_j. The second line follows because δ_j is independent of $\Psi_j X_{j-1} + \mu_j$ — even unconditionally, but all the more true under $\mathcal{F}_{j-1}$. Then the third line follows because μ_j can be treated as a constant under $\operatorname{Var}[\cdot \mid \mathcal{F}_{j-1}]$, while $\operatorname{Var}[\langle v, \delta_j\rangle \mid \mathcal{F}_{j-1}]$ simply equals $\operatorname{Var}\langle v, \delta_j\rangle$.

Notice that the left term on the third line represents the conditional variance of X_{j-1} after transformation by Ψ_j. Dropping the test vectors $v \in V$, we thus recognize this equality of linear operators:

$$\widehat{S}_j = \Psi_j S_{j-1} \Psi_j^* + Q_j\,.$$

Returning to eq. 12.3, we may argue with the same reasoning to arrive at:

$$C_{Y,j} = A_j \widehat{S}_j A_j^* + R_j\,,$$

after dropping out the test vectors $w \in W$.

Next we attack the covariance homomorphism $C_{XY,j}(\omega)\colon W \to V$, which turns out to be very easy:

$$\begin{aligned}\langle v, C_{XY,j} w\rangle &= \operatorname{Cov}\big[\langle v, X_j\rangle, \langle w, Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Cov}\big[\langle v, X_j\rangle, \langle w, A_j X_j + \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Cov}\big[\langle v, X_j\rangle, \langle w, A_j X_j\rangle \mid \mathcal{F}_{j-1}\big] + \operatorname{Cov}\big[\langle v, X_j\rangle, \langle w, \varepsilon_j\rangle \mid \mathcal{F}_{j-1}\big] \\ &= \operatorname{Cov}\big[\langle v, X_j\rangle, \langle A_j^* w, X_j\rangle \mid \mathcal{F}_{j-1}\big] + 0 \\ &= \langle v, \widehat{S}_j A_j^* w\rangle\,.\end{aligned}$$

The fourth line follows from ε_j being independent of X_j. Dropping the test vectors $v \in V$ and $w \in W$, we therefore see:

$$C_{XY,j} = \widehat{S}_j A_j^*. \tag{12.4}$$

To complete the specification of the Kalman filter as an algorithm, we must have a formula for S_j, which represents the conditional variance of the hidden state X_j given the observations known so far, up to time j.

As one might expect, it can be computed recursively. Actually, we have essentially computed it in §11 already, under somewhat different notation. For clarity, we will repeat those arguments, adapted to the present situation. The main principle at work is the orthogonal decomposition of the (conditional) variance of the random variable X_j as the sum of the variance of its least-squares predictor $\widehat{X}_j$ plus the variance of the residual $X_j - \widehat{X}_j$.

Consider:

$$\begin{aligned}
\langle v, S_j v\rangle &= \mathrm{Var}\big[\langle v, X_j\rangle \mid \mathcal{F}_j\big] \\
&= \mathrm{Var}\big[\langle v, (X_j - \widehat{X}_j) + \widehat{X}_j\rangle \mid \mathcal{F}_j\big] \\
&= \mathrm{Var}\big[\langle v, X_j - \widehat{X}_j\rangle \mid \mathcal{F}_j\big] \\
&= \mathrm{Var}\big[\langle v, X_j - \widehat{X}_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \mathrm{Var}\big[\langle v, X_j - B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big]\,.
\end{aligned}$$

The third line follows because the predictor $\widehat{X}_j$ is a $\mathcal{F}_j$-measurable function (i.e. a function of $Y_1, \ldots, Y_j$), so it can be treated like a constant when conditioning under $\mathcal{F}_j$; the variance of a constant is zero. The fourth line simply removes the variable Y_j from the conditioning, which is allowed because the residual $X_j - \widehat{X}_j$ (under the conditional probability law of $\mathcal{F}_{j-1}$) is independent of Y_j, as proven[22] in §11. Finally, the last line comes from a casual observation of eq. 12.2: $\widehat{X}_j$ is the sum of a $\mathcal{F}_{j-1}$-measurable function plus $B_j Y_j$; the former has zero variance when conditioning under $\mathcal{F}_{j-1}$.

Continuing from the last line, we have:

$$\begin{aligned}
\langle v, S_j v\rangle &= \mathrm{Cov}\big[\langle v, X_j - B_j Y_j\rangle, \langle v, X_j - B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \mathrm{Var}\big[\langle v, X_j\rangle \mid \mathcal{F}_{j-1}\big] - 2\,\mathrm{Cov}\big[\langle v, X_j\rangle, \langle v, B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] + \mathrm{Var}\big[\langle v, B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \mathrm{Var}\big[\langle v, X_j\rangle \mid \mathcal{F}_{j-1}\big] - 2\,\mathrm{Cov}\big[\langle v, X_j\rangle, \langle B_j^* v, Y_j\rangle \mid \mathcal{F}_{j-1}\big] + \mathrm{Var}\big[\langle B_j^* v, Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \langle v, \widehat{S}_j v\rangle - 2\langle v, C_{XY,j} B_j^* v\rangle + \langle v, B_j C_{Y,j} B_j^* v\rangle \\
&= \langle v, \widehat{S}_j v\rangle - \langle v, B_j C_{XY,j}^* v\rangle \\
&= \langle v, \widehat{S}_j v\rangle - \langle v, B_j (A_j \widehat{S}_j) v\rangle\,.
\end{aligned}$$

The fifth line follows from substituting the definition $B_j = C_{XY,j} C_{Y,j}^{-1}$, from eq. 12.1, and simplifying. The last line follows from substituting in eq. 12.4.

Finally, we observe that $B_j(A_j \widehat{S}_j) = \widehat{S}_j A_j^* C_{Y,j}^{-1} A_j \widehat{S}_j$ is self-adjoint, which allows us to conclude:

$$S_j = (I - B_j A_j)\widehat{S}_j\,. \tag{12.5}$$

(Here we are relying on the fact[23] that $\langle Tv, v\rangle = 0$ for $v \in V$ implies $T = 0$ *for self-adjoint* $T\colon V \to V$. But $T = S_j - (\widehat{S}_j - B_j A_j \widehat{S}_j)$ being self-adjoint is a hypothesis that must be proven first.)

[22] Because $Y_j = AX_j + \varepsilon_j$ has an error term ε_j that was not present in §11, we detail the precise set-up. Consider the input random variable to be the $V \times V$-valued $Z = (X_j, \varepsilon_j)$. Let M be the range of the adjoint of the linear transformation $(x, e) \mapsto Ax + e$, with respect to the inner product induced by the conditional variance under probability law $\mathcal{F}_{j-1}$. Then §9 says the random variables $P_M Z$ and $P_{M^\perp} Z = Z - P_M Z$ are independent. The former is an invertible function of Y_j, while the latter is the residual whose first component is exactly $X_j - \widehat{X}_j$.

[23] We actually used this fact twice earlier, but did not call out the self-adjointness hypothesis. It was obvious enough in those instances.

In applications, S_j is often computed not with eq. 12.5, but with a longer formula called the *Joseph form*:

$$S_j = (I - B_j A_j)\widehat{S}_j(I - B_j A_j)^* + B_j R_j B_j^* \,. \tag{12.6}$$

It has the advantage that the terms making it up are self-adjoint (their matrices are symmetric) even in the presence of round-off error in floating-point calculations, which is not true of eq. 12.5. It is also robust against B_j not being equal exactly to the optimal transformation, due to model error or round-off error, as it does not assume B_j is the Kalman gain. That is, $B_j(\omega)\colon W \to V$ can be arbitrary like at the end of §11, when we compared the optimal predictor against arbitrary predictors. (Note that the definition of B_j from eq. 12.1, was only used in the last steps in deriving eq. 12.5, and nowhere else earlier in this section.)

To derive the Joseph form, we cannot assume self-adjointness of $B_j A_j \widehat{S}_j$, so we must introduce test vectors $u, v \in V$ and expand fully as covariances[24]:

$$\begin{aligned}
\langle u, S_j v\rangle &= \operatorname{Cov}\big[\langle u, X_j - B_j Y_j\rangle, \langle v, X_j - B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \operatorname{Cov}\big[\langle u, X_j\rangle, \langle v, X_j\rangle \mid \mathcal{F}_{j-1}\big] - \operatorname{Cov}\big[\langle u, X_j\rangle, \langle v, B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&\quad - \operatorname{Cov}\big[\langle u, B_j Y_j\rangle, \langle v, X_j\rangle \mid \mathcal{F}_{j-1}\big] + \operatorname{Cov}\big[\langle u, B_j Y_j\rangle, \langle v, B_j Y_j\rangle \mid \mathcal{F}_{j-1}\big] \\
&= \langle u, \widehat{S}_j v\rangle - \langle u, C_{XY,j} B_j^* v\rangle - \langle v, C_{XY,j} B_j^* u\rangle + \langle B_j^* u, C_{Y,j} B_j^* v\rangle \\
&= \langle u, \widehat{S}_j v\rangle - \langle u, C_{XY,j} B_j^* v\rangle - \langle u, B_j C_{XY,j}^* v\rangle + \langle u, B_j C_{Y,j} B_j^* v\rangle.
\end{aligned}$$

The above equality holds true for all u, v; in turn, this equality of linear operators holds:

$$\begin{aligned}
S_j &= \widehat{S}_j - C_{XY,j} B_j^* - B_j C_{XY,j}^* + B_j C_{Y,j} B_j^* \\
&= \widehat{S}_j - \widehat{S}_j A_j^* B_j^* - B_j A_j \widehat{S}_j + B_j (A_j \widehat{S}_j A_j^* + R_j) B_j^* \,,
\end{aligned}$$

after substituting for $C_{XY,j}$ and $C_{Y,j}$. The last expression is equal to the right-hand side of eq. 12.6, after re-arranging terms.

13 Bibliography

[Axler] Sheldon Axler. *Linear Algebra Done Right*. Fourth Edition. Springer, 2024. https://linear.axler.net/

[Billingsley] Patrick Billingsley. *Probability and Measure*. *Second Edition*. John Wiley & Sons, 1986.

[Feller] William Feller. *An Introduction to Probability Theory and Its Applications*, *Volume II*. John Wiley & Sons, 1970.

[Folland] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. *Second Edition*. John Wiley & Sons, 1999.

[Friedberg] Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence. *Linear Algebra. Third Edition*. Prentice Hall, 1997.

[Halmos] Paul R. Halmos. *Finite-Dimensional Vector Spaces*. Springer, 1987.

[24] We could have done this right off the bat and thus derive eq. 12.5 and eq. 12.6 together in a very concise manner. But the calculation would look rather unmotivated without the benefit of seeing the conceptually simpler version first.

[Kálmán] Rudolf E. Kálmán. "A New Approach to Linear Filtering and Prediction Problems". *Transactions of the ASME – Journal of Basic Engineering*, volume 82, series D, pages 35–45, 1960. https://www.cs.unc.edu/∼welch/kalman/kalmanPaper.html

[Kolmogorov] Andrey N. Kolmogorov. Translated by Nathan Morrison. *Foundations of the Theory of Probability*. Chelsea, 1956.

[Labbe] Roger R. Labbe Jr. *Kalman and Bayesian Filters in Python*. Last version accessed: July 15, 2024. https://github.com/rlabbe/Kalman-and-Bayesian-Filters-in-Python

[Øksendal] Bernt Øksendal. Stochastic Differential Equations: An Introduction with Applications. Fifth Edition, Corrected Printing. Springer-Verlag, 2000.

[Rohatgi] Vijay K. Rohatgi, A. K. MD. Ehsanes Saleh. *An Introduction to Probability and Statistics. Second Edition*. Prentice Hall, 2001.

[Rosenthal] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, 2003. https://probability.ca/jeff/grprobbook.html

[Wichura] Michael J. Wichura. *The Coordinate-Free Approach to Linear Models*. Cambridge University Press, 2006.