Notes on Multivariate Analysis

Niels Richard Hansen
April 17, 2008

1 Multivariate analysis with the normal distribution

1.1 Introduction

In multivariate analysis the focus in on multivariate observations and in particular the statistical modeling and inference of the dependence structure between variables. For quite a large number of practical statistical problems the data can be organized conveniently in a $n \times p$ matrix X. We will refer to p as the number of *features* and n as the number of *replications*. Depending on the objectives of the analysis, the nature of the data and the data collection process a range of relevant statistical models present themselves. From earlier statistics courses you may think of linear normal models where we have a one-dimensional response variable measured independently n times. The experimental design in terms of continues covariates and/or factor levels for each observation are also registered, and together with the response variable form the p features. But the statistical models are models of the one-dimensional response variable – conditionally on the covariates. In multivariate analysis we extend the setup to focus on two or more response variables and their joint distribution. Still the joint distribution may then be considered conditionally on other variables and/or an experimental design.

It is the general assumption in multivariate analysis that the p features are dependent. Or at least it is partly the objective in multivariate analysis to study the dependence structure between the features. We may think of the features as different measurements on a single individual. This could be clinical measurements on a patient or test scores for one individual from an exam or IQ-test. But it could also be repeated measurements on the same individual over time. In the latter case one often focus on parametric models that are more specific than those we consider in multivariate analysis, but the models are submodels (often non-linear) of the general models considered. The features could also be the prices or returns for p assets or other financial products. In portfolio management it is crucial to acknowledge that such prices or returns are dependent. However, observations made over time can introduce another dependence structure. The features are not just dependent within each row but also within each column. The problem therefore belongs to the world of multivariate time series analysis. On the other hand, returns may actually be found to be empirically uncorrelated over time, and the statistical inference for the so-called capital asset pricing model (CAPM) from financial econometrics builds on the assumptions of temporally independent returns and a multivariate normal distribution of the p asset returns. In multivariate analysis we do generally consider the n rows in the matrix to represent independent observations of the p features. Therefore, assuming temporally independent returns, the statistical analysis of the CAPM model falls within the framework of multivariate analysis anyway.

For linear normal models we are required to build models on the vector space $\mathbb{R}^n$ of column vectors. Multivariate analysis requires that we can build models on $M(n, p)$ of $n \times p$ matrices. This is a vector space and in many ways it is simply $\mathbb{R}^{np}$, but the organization of data in the matrix form does actually mean something. Some prefer to consequently map everything into $\mathbb{R}^{np}$ and work with the theory for the normal distribution, say, on $\mathbb{R}^{np}$. Others, including this author, prefer to work with $M(n, p)$ as a vector space directly. This implies basically that we need to discuss some general vector space theory. Moreover, the models considered in these notes are all given in terms of the normal distribution. Thus we need to understand what the normal distribution on a general vector space is.

1.2 Vector spaces

It is assumed that the reader has a basic knowledge of finite dimensional vector space theory. For instance, that for any finite dimensional vector space V we can choose a basis, and that all such bases have the same number of elements. This number, d, say, is the dimension of V. For another finite dimensional vector space W of dimension d' we can likewise choose a basis, and any linear map from V to W can be represented in the chosen bases using a $d' \times d$ matrix.

Two vector spaces V and W are called isomorphic if there is a bijective, linear map from V to W. One then shows that two vector spaces are isomorphic if and only if they have the same dimension. We know one example of a d dimensional vector space; $\mathbb{R}^d$. Thus all vector spaces are essentially (up to isomorphism) equal to $\mathbb{R}^d$. However, for the vector spaces we will consider in multivariate analysis, the concrete representation of the vector space plays an important role, were the reduction of the space to $\mathbb{R}^d$ for some d is not always a good idea.

The next level of abstraction – when we get beyond studying a single linear map between two vector spaces, say – is to study a range of sets of linear maps. These sets also turn out to have a linear structure themselves and thus be vector spaces too. For any vector space V we define the *dual space* as

$$V^* = \{z : V \to \mathbb{R} \mid z \text{ is a linear function}\},$$

which is the set of linear functions on V into the real line. For two vector spaces, V and W, we define

$$\mathcal{L}(V, W) = \{A : V \to W \mid A \text{ is a linear map}\}$$

of linear maps from V to W. And finally we also define

$$\mathcal{B}(V,W) = \{B : V \times W \to \mathbb{R} \mid B \text{ is a bilinear function}\}$$

of bilinear functions on $V \times W$ into $\mathbb{R}$. That $B : V \times W \to \mathbb{R}$ is bilinear means that $y \mapsto B(x,y)$ is linear for all $x \in V$ and that $x \mapsto B(x,y)$ is linear for all $y \in W$. All three sets can be equipped with pointwise addition and scalar multiplication, which can easily be seen to make the sets into vector spaces.

Lemma 1.2.1. *There is an isomorphism*

$$\Phi : \mathcal{B}(V,W) \mapsto \mathcal{L}(V,W^*)$$

determined by the

$$\Phi(B)(x)(y) = B(x,y)$$

for any $B \in \mathcal{B}(V,W)$*,* $x \in V$ *and* $y \in W$*.*

Proof: It follows by bilinearity of B that $\Phi(B)(x)$ is a linear function on W and that $\Phi(B)$ is a linear map then from V to W^*. It's obvious that Φ is a linear map. To show that it is an isomorphism we just need to find the inverse. For A a linear map from V to W^* we define Φ^{-1} by

$$\Phi^{-1}(A)(x,y) = A(x)(y),$$

which is obviously a bilinear function on $V \times W$ by linearity of $A(x)$ and A. We observe that

$$\Phi(\Phi^{-1}(A))(x)(y) = A(x)(y),$$

which shows that Φ^{-1} is the inverse of Φ. □

Since V^* is a vector space it has a dual space, which we can denote V^{**} – the second dual of V. This space is in a very fundamental way isomorphic to V. If $x \in V$ we can define the linear function $\varphi(x) \in V^{**}$ by

$$\varphi(x)(z) := z(x).$$

Obviously $\varphi(x) \in V^{**}$ and by linearity of z, φ is a linear map from V to V^{**}. Moreover, for $x \neq y$ there is a $z \in V^*$ with $z(x) \neq z(y)$, cf. Exercise 1.3.3, and this shows that $\varphi(x) \neq \varphi(y)$. Thus φ is one-to-one and the image of φ is a d-dimensional subspace of V^{**} since V has dimension d. But V^* also has dimension d, cf. Exercise 1.3.1, and so does V^{**} and therefore

$$\varphi(V) = V^{**}.$$

The isomorphism φ is of such a fundamental nature that we will typically suppress any reference to it and regard V and V^{**} as the same space. It is important though to recognize that the identification is not just an arbitrary isomorphism based on equal dimensions but the very specific identification that tells exactly how $x \in V$ should be regarded as a linear function on V^*.

The two spaces V and V^* have the same dimension and are thus isomorphic, but there is no abstract, natural isomorphism between them. If we choose a basis $e_1, \dots, e_d$ for V we have, however, automatically introduced a natural identification of V and V^*. With $e^i(x)$ denoting the i'th coordinate of x in the basis the function $x \mapsto e^i(x)$ is linear, and we denote it by e^i. For any $y \in V$ we get the function

$$x \mapsto \sum_{i=1}^{d} e^i(y)e^i(x).$$

We let $\psi : V \to V^*$ be given by

$$\psi(y) = \sum_{i=1}^{d} e^i(y)e^i,$$

and then ψ is a linear map from V to V^*, cf. also Exercise 1.3.1. In the exercise $e^1, \dots, e^d$ is shown to be a basis for V^*, and it is called the *dual basis* of V^*. Once we have introduced a basis for V we should always choose the dual basis for V^*. We can show that ψ is one-to-one and thus an isomorphism. Again we may be sloppy and simply write $y(x)$ instead of $\psi(y)(x)$ for the linear map $\psi(y)$ evaluated on $x \in V$. However, in this case the identification of the vector spaces is basis dependent. You may note that

$$y(x) = \psi(y)(x) = \sum_{i=1}^{d} e^i(y)e^i(x) = \psi(x)(y) = x(y),$$

which is a symmetry that holds for this particular type of isomorphism.

The choice of a basis is actually also equivalent to choosing an inner product on V. A bilinear function $\Lambda \in \mathcal{B}(V, V)$ is an inner product if it is symmetric and positive definite. That is, if $\Lambda(v, w) = \Lambda(w, v)$ for all $v, w \in V$ and if $\Lambda(v, v) \geq 0$ and $= 0$ if and only if $v = 0$. It defines an orthogonality concept on V;

$$v \perp_\Lambda w \Leftrightarrow \Lambda(v, w) = 0$$

and a norm on V;

$$||v||_\Lambda = \sqrt{\Lambda(v, v)}.$$

If $e_1, \dots, e_d$ is a basis, the function

$$I(z, x) = \sum_{i=1}^{d} e^i(z)e^i(x)$$

is bilinear and positive definite. Moreover, in this inner product the basis is an orthonormal basis meaning that $||e_i||_I = 1$ and $I(e_i, e_j) = 0$ if $i \neq j$. For a given basis we refer to this inner product as the standard inner product – defined in terms of the basis. On the other hand, if we have an inner product Λ on V then all linear functions on V can be written as

$$x \mapsto \Lambda(x, y)$$

for some $y \in V$. This also provides an identification of V and V^*. If we choose any Λ-*orthonormal basis* for V the resulting basis isomorphism ψ above coincides with the inner product identification of V and V^*.

Lemma 1.2.2. *Given a basis for W and the corresponding basis induced isomorphism $\psi : W \to W^*$ then there is an isomorphism $\Psi : \mathcal{L}(V, W) \to \mathcal{L}(V, W^*)$ defined by*

$$\Psi(B)(x) = \psi(B(x))$$

for $B \in \mathcal{L}(V, W)$ and $x \in V$. Thus

$$\Phi_\psi := \Psi^{-1} \circ \Phi : \mathcal{B}(V, W) \to \mathcal{L}(V, W).$$

is an isomorphism.

Proof: Obviously $\psi(B(x)) \in W^*$ and linearity of B and ψ assures that $\Psi(B) \in \mathcal{L}(V, W^*)$. Moreover, linearity of ψ assures the Ψ is linear. The inverse is clearly

$$\Psi^{-1}(B)(x) = \psi^{-1}(B(x))$$

for $B \in \mathcal{L}(V, W^*)$ and $x \in V$. Since the composition of two isomorphisms is an isomorphism the last result follows. □

There is a nice interpretation of Φ_ψ for $V = \mathbb{R}^d$ and $W = \mathbb{R}^{d'}$ endowed with the standard bases. For these vector spaces of column vectors, the set of bilinear functions on $V \times W$ can be identified with the set of $d \times d'$ matrices $M(d, d')$, such that $B \in M(d, d')$ is identified with the bilinear function

$$(x, y) \mapsto x^t B y.$$

In this case the dual $(\mathbb{R}^{d'})^*$ is interpreted as the space of row vectors and $\psi(x) = x^t$ for $x \in \mathbb{R}^{d'}$ defines the isomorphism of $\mathbb{R}^{d'}$ and $(\mathbb{R}^{d'})^*$ using standard bases. Then for $B \in M(d, d')$

$$\Phi_\psi(B)(x) = B^t x.$$

Thus the linear map corresponding to the bilinear function given by the matrix B is given by the transposed matrix B^t.

In an abstract vector space we can add elements and we can multiply elements by real scalars, but there is no general "multiplication" or "composition" of elements. The vector space $\mathcal{B}(V, V)$ does not have an obvious composition, but given a basis for V and thus the identification Φ_ψ from Lemma 1.2.2 of $\mathcal{B}(V, V)$ and $\mathcal{L}(V, V)$ we can introduce a composition. For $A, B \in \mathcal{B}(V, V)$ we define the composition by

$$A \circ B = \Phi_\psi^{-1}([\Phi_\psi(A)] \circ [\Phi_\psi(B)]).$$

For $B \in \mathcal{B}(V, V)$ the linear map $\Phi_\psi(B)$ may be invertible. In that case we define the inverse of the bilinear map B by the formula

$$B^{-1} := \Phi_\psi^{-1}(\Phi_\psi(B)^{-1}).$$

Obviously then $B \circ B^{-1} = B^{-1} \circ B = I$ where I refers to the standard inner product.

For $V = \mathbb{R}^d$ and representing the bilinear functions by matrices the composition is matrix multiplication and the inverse is given by the matrix inverse. However, we must be very careful here. If B is a bilinear function regarded as a $d \times d$ matrix, the bilinear function determined by B^{-1} is in fact given as

$$(x, y) \mapsto x^t (B^{-1})^t y = y^t B^{-1} x.$$

Thus the inverse bilinear function equals $(x, y) \mapsto x^t B^{-1} y$ if and only if B and thus B^{-1} are symmetric.

1.3 Exercises

Exercise 1.3.1. Show that if $e_1, \ldots, e_d$ is a basis for V, we can define the functions $e^i : V \to \mathbb{R}$ by $e^i(x)$ is the i'the coordinate of x in this basis. Show that the functions $e^1, \ldots, e^d$ are linear, thus elements in V^*, and linearly independent. Then show that they span V^* and thus form a basis, which makes V^* into a d dimensional vector space too.

Exercise 1.3.2. If $x, y \in V$ and $x \neq y$ show that there exists $z \in V^*$ such that $z(x) \neq z(y)$.

Exercise 1.3.3. Show that for a bilinear function B and $(x, y) \in V \times W$

$$B(x, y) = \Phi(B)(x)(y).$$

Since $\Phi(B)$ is a linear map from V to W^* it may have an inverse, $\Phi(B)^{-1}$, from W^* to V. The inverse is regarded as a bilinear function on $V^* \times W^*$ by the definition

$$(x, y) \mapsto x(\Phi(B)^{-1} y).$$

Show that if $V = W$ then

$$B(\Phi(B)^{-1} x, \Phi(B)^{-1} y) = x(\Phi(B)^{-1} y).$$

for $(x, y) \in V^* \times V^*$, and show that if B is symmetric, then

$$x(\Phi(B)^{-1} y) = y(\Phi(B)^{-1} x).$$

We could write $\Phi(B)^{-1} = B \circ (\Phi(B)^{-1} \times \Phi(B)^{-1})$ where

$$(\Phi(B)^{-1} \times \Phi(B)^{-1})(x, y) = (\Phi(B)^{-1} x, \Phi(B)^{-1} y)$$

for $(x, y) \in V^* \times V^*$. We have shown that for $V = W$ then $\Phi(B)^{-1}$ is symmetric, regarded as a bilinear function on $V^* \times V^*$, if B is symmetric.

If B is a symmetric bilinear function on $V \times V$ and we let $\psi \times \psi$ be defined by

$$(\psi \times \psi)(x, y) = (\psi(x), \psi(y))$$

then $\Phi(B)^{-1} \circ (\psi \times \psi)$ is a symmetric bilinear function on $V \times V$. Show that in this case

$$B^{-1} = \Phi(B)^{-1} \circ (\psi \times \psi).$$

Hint: Show first the formula $B^{-1}(x, y) = \psi(\Phi(B)^{-1}(\psi(x)))(y)$.

1.4 The normal distribution on a finite dimensional vector space

We consider a d-dimensional, real vector space V endowed with the usual Borel σ-algebra and a Lebesgue measure. The Lebesgue measure is unique up to a proportionality constant. The regular normal distribution on V is given in terms of a symmetric, positive definite bilinear function Λ on V and a location parameter $\mu \in V$.

The set of symmetric, positive definite bilinear functions (the inner products) on V is denoted $\mathrm{S}_+(V)$, and in the language of the normal distribution an element $\Lambda \in \mathrm{S}_+(V)$ is also called a *precision*. For a precision parameter $\Lambda \in \mathrm{S}_+(V)$ and a location parameter $\mu \in V$ the regular normal distribution on V has density $f_{\Lambda,\mu}$ w.r.t. to the Lebesgue measure where

$$f_{\Lambda,\mu}(x) = \frac{1}{\varphi_d(\Lambda)} \exp\left(-\frac{1}{2}\Lambda(x-\mu, x-\mu)\right). \tag{1.1}$$

Here $\varphi_d(\Lambda)$ is a normalization constant that depends on the dimension d and on Λ – but not on μ. This is because the Lebesgue measure on V is translation invariant. We can write the normalization constant as

$$\varphi_d(\Lambda) = \int \exp\left(-\frac{1}{2}\Lambda(x,x)\right) \mathrm{d}x$$

where the integration is w.r.t. the chosen Lebesgue measure on V.

An affine subspace of V is a set $\mu + U$ where $U \subseteq V$ is a linear subspace of V and $\mu \in V$. It is possible to define a normal distribution, whose support is an affine subspace. Since U is a subspace of V and $\mu \in V$ we can define the embedding $\tau_\mu : U \to V$ by

$$\tau_\mu(u) = \mu + u.$$

The image measure under τ_μ of the regular normal distribution on U with location parameter 0 and precision Λ is called the normal distribution with location μ, precision Λ and support $\mu + U$. If $U \neq V$ we call this a singular normal distribution. If $V = U$ the definition coincides with the definition of the regular normal distribution above. In terms of random variables we say that Y follows a normal distribution on V with location μ, precision Λ and support $\mu + U$ if

$$Y = \mu + X$$

where the distribution of X is a normal distribution on U with location 0 and precision Λ.

If we choose a fixed basis $e_1, \ldots, e_d$ for V we use $e^i(x) \in \mathbb{R}$ to denote the i'th coordinate of $x \in V$. We can identify x with the column vector $(e^1(x), \ldots, e^d(x))^t \in \mathbb{R}^d$ and

$$x = \sum_{i=1}^{d} e^i(x) e_i.$$

If X is a random variable on V let $\xi_i = \mathbb{E}e^i(X)$. Then the mean value of X is defined as

$$\xi = \sum_{i=1}^{d} \xi_i e_i,$$

which can be identified with the column vector $(\xi_1, \ldots, \xi_d)^t$. Moreover, since we have chosen a basis we identify V with V^* such that for $z \in V$ with $z = \sum_{i=1}^{d} e^i(z) e_i$ the corresponding linear function in V^* is

$$z(x) = \sum_{i=1}^{d} e^i(x) e^i(z).$$

We identify z with the column vector $(e^1(z), \ldots, e^d(z))^t$ so that using matrix multiplication we may simply write $z(x) = z^t x$. If X is a random variable with values in V and $z \in V$ we find that

$$\mathbb{E}z(X) = z^t \xi.$$

To introduce the covariance of X we define the map $\Sigma : V \times V \to \mathbb{R}$ by

$$\Sigma(z, w) = \operatorname{cov}(z(X), w(X))$$

for any $(z, w) \in V \times V$. We may show that this is a symmetric and positive semidefinite bilinear function on V. We may identify this bilinear function with a matrix, which we with abuse of notation call Σ too, and whose entries are given by

$$\Sigma_{ij} = \operatorname{cov}(e^i(X), e^j(X)) = \mathbb{E}(e^i(X) - \xi_i)(e^j(X) - \xi_j).$$

By bilinearity

$$\Sigma(z, w) = \sum_{i,j=1}^{d} e^i(z) \Sigma_{ij} e^j(w).$$

Thus in terms of matrix multiplication – and identifying z and w with column vectors – this equals $z^t \Sigma w$. One can show that for a regular normal distribution with location parameter μ and precision Λ the mean is μ and the covariance is the inverse of Λ, that is, $\Sigma = \Lambda^{-1}$.

The following results are of great importance. The proofs are skipped.

Theorem 1.4.1. *For any $\xi \in V$ and symmetric, positive semidefinite bilinear function Σ on V there is one and only one normal distribution on V with mean ξ and covariance Σ. We denote this distribution by $N(\xi, \Sigma)$. Moreover, it is a regular normal distribution if and only if Σ is positive definite, and in this case the precision is $\Lambda = \Sigma^{-1}$.*

If X is a random variable with distribution $N(\xi, \Sigma)$ we also write $X \sim N(\xi, \Sigma)$.

If $A : V \to W$ is a linear map and $f_1, \ldots, f_{d'}$ is a basis of W, then we can represent the map as a $d' \times d$ matrix, which we also call A. The transpose, A^t, represents a map $W \to V$, which we also denote A^t. For a bilinear function Σ on V we can define the bilinear function $A\Sigma A^t$ on W by

$$A\Sigma A^t(x, y) := \Sigma(A^t x, A^t y).$$

If we identify x and y with d'-dimensional column vectors, $A^t x$ and $A^t y$ are d-dimensional column vectors and

$$\Sigma(A^t x, A^t y) = (A^t x)^t \Sigma(A^t y) = (x^t A)\Sigma(A^t y) = x^t(A\Sigma A^t)y.$$

Hence the notation, $A\Sigma A^t$ for the bilinear map. Because Σ is symmetric the linear map from W to W corresponding to the the bilinear function $A\Sigma A^t$ on $W \times W$ is precisely the map

$$x \mapsto A(\Sigma(A^t(x)))$$

where Σ is identified as a linear map $\Sigma : V \to V$. It is easy to see that if Σ is positive semidefinite so is $A\Sigma A^t$.

Theorem 1.4.2. *If $X \sim N(\xi, \Sigma)$, if $A : V \to W$ is a linear map and if $\nu \in W$ then $\nu + AX \sim N(\nu + A\xi, A\Sigma A^t)$.*

The choice of basis introduces an inner product, which makes the basis into an orthonormal basis. Using this inner product we can for a subspace $U \subseteq V$ define the orthogonal complement $U^\perp$ as the subspace of vectors in V that are orthogonal to U. Let P_1 denote the orthogonal projection on U and $P_2 = I - P_1$ the orthogonal projection onto $U^\perp$. Assume that $X \sim N(\xi, \Sigma)$ and define $X_1 = P_1 X$, $X_2 = P_2 X$, $\xi_1 = P_1\xi$ and $\xi_2 = P_2\xi$. Define for $i, j = 1, 2$ the bilinear functions

$$\Sigma_{i,j} = P_i \Sigma P_j^t.$$

From Theorem 1.4.2 $X_2 \sim N(\xi_2, \Sigma_{2,2})$. In addition, we have a result on the conditional distribution of X_1 given X_2.

Theorem 1.4.3. *If $\Sigma_{2,2}$ has an inverse, then*

$$X_1 | X_2 \sim N(\xi_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(X_2 - \xi_2), \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}).$$

2

Tensor products

2.1 Introduction

"Unfortunately no one can be told what *the matrix* is. You have to see it for yourself". This is what Morpheus says to Neo just before he takes *the red pill*.

I can tell you what *a matrix* is. It's a set of real numbers arranged in an square array – but maybe that is not what Morpheus had in mind. However, in mathematics one may rephrase the citation as

> "Unfortunately no one can be told what *a tensor* is. You have to see it for yourself".

The problem is that I don't have a red pill, and it may require equal parts of hard work and simple acceptance of my authority to gain confidence in working with tensors and tensor products.

What we are aiming at is quite down to earth; the organization of linear and bilinear maps so that they can be manipulated algebraically without tedious references to coordinates and choices of basis in the underlying vector space(s). And even if we choose a basis we need to organize the set of real coefficients in a suitable choice of *scheme* – like the linear map representation using a matrix.

Why are matrices not always sufficiently good schemes? For multivariate statistical analysis the answer is that the data are naturally in a matrix form. This means that the data vector is written down as a matrix, but this is *not* something we think of as a linear map but as an element in a vector space of matrices. On this vector space we need to consider linear maps and we find ourself in a mess of matrices at different levels if we insist on the matrix organization for linear maps. However, the abstract and coordinate free theory for tensor products on vector spaces is – well, abstract. It does not offer the opportunity to exploit the concrete structure on the vector space of real $n \times p$ matrices, which is of most importance to multivariate analysis. Therefore we choose a concrete approach to tensor products.

2.2 Concrete tensor products

We introduce tensor products in this section using constructions that are based on the explicit structure of the vector spaces – in particular the vector space $\mathbb{R}^d$ and the space $M(n,p)$ of real $n \times p$ matrices. In both cases we take advantage of standard bases and inner products and the explicit identification of the vector space with it's dual. For the matrix vector space we can also exploit matrix algebra to obtain convenient algebraic expressions for tensor product constructions and computations.

2.2.1 The space $\mathbb{R}^d$

Consider the vector space $\mathbb{R}^d$. Elements in this space are thought of as column vectors and we write $x = (x_1, \ldots, x_d)^t$ for the typical element in $\mathbb{R}^d$. Here the "^{t}" means to transpose, that is, physically rotate the column vector 90 degrees. The vector space has a standard basis, $e_1, \ldots, e_d$, where e_i denotes the column vector with a 1 at the i'th coordinate and 0 elsewhere, and we can then write the typical element as

$$x = \sum_{i=1}^{d} x_i e_i.$$

The dual of $\mathbb{R}^d$ is identified with $\mathbb{R}^d$ and $z \in \mathbb{R}^d$ defines the linear function

$$x \mapsto z^t x = \sum_{i=1}^{d} z_i x_i$$

on $\mathbb{R}^d$. Therefore we often think of and represent the dual of $\mathbb{R}^d$ as row vectors and thus for $z \in \mathbb{R}^d$ the corresponding dual element is z^t. This identification of $\mathbb{R}^d$ and it's dual is equivalent to introducing the standard inner product on $\mathbb{R}^d$ given by

$$(x, z) \mapsto z^t x,$$

which makes the standard basis into an orthonormal basis.

Definition 2.2.1. *The tensor product $\mathbb{R}^d \otimes \mathbb{R}^{d'}$ is defined as the vector space $\mathcal{B}(\mathbb{R}^d, \mathbb{R}^{d'})$.*

For any two $x \in \mathbb{R}^d$ and $z \in \mathbb{R}^{d'}$ we define the *elementary tensor* $x \otimes z$ as the bilinear function

$$x \otimes z(a, b) = x^t a z^t b \tag{2.1}$$

for $(a, b) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. If $f_1, \ldots, f_d$ is the standard basis on $\mathbb{R}^{d'}$ we can consider the set of elementary tensors $(e_i \otimes f_j)_{i=1,\ldots,d, j=1,\ldots,d'}$.

Lemma 2.2.2. *The set $(e_i \otimes f_j)_{i=1,\ldots,d, j=1,\ldots,d'}$ is a basis for $\mathbb{R}^d \otimes \mathbb{R}^{d'}$. Any $B \in \mathbb{R}^d \otimes \mathbb{R}^{d'}$ can therefore be written as*

$$B = \sum_{i,j} B_{ij} e_i \otimes f_j$$

and it is natural to organize the coefficients for B in a $d \times d'$ matrix and simply write

$$B = \begin{pmatrix} B_{11} & \cdots & B_{1d'} \\ \vdots & & \vdots \\ B_{d1} & \cdots & B_{dd'} \end{pmatrix}.$$

This provides an identification of $\mathbb{R}^d \otimes \mathbb{R}^{d'}$ and the set $M(d, d')$ of $d \times d'$ matrices.

Proof: Define $B_{ij} = B(e_i, f_j)$. Then for any $x = \sum_{i=1}^{d} x_i e_i \in V$ and $y = \sum_{j=1}^{d'} y_j f_j \in W$ we have by bilinearity of B that

$$B(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{d'} x_i y_j B(e_i, f_j) = \sum_{i=1}^{d} \sum_{j=1}^{d'} B_{ij} x_i y_j.$$

By definition $e_i \otimes f_j(x, y) = x_i y_j$ it follows that

$$B = \sum_{i,j} B_{ij} e_i \otimes f_j,$$

and we have shown that $(e_i \otimes f_j)_{i=1,\ldots,d, j=1,\ldots,d'}$ span $\mathbb{R}^d \otimes \mathbb{R}^{d'}$. Since

$$e_i \otimes f_j(e_k, f_l) = \delta_{ik} \delta_{jl}$$

where δ_{ik} is 1 if and only if $i = k$ it follows that $(e_i \otimes f_j)_{i=1,\ldots,d, j=1,\ldots,d'}$ are also linearly independent (no element can be written as a linear combination of the others due to the identity above, which is 1 if and only if $i = k$ and $j = l$). This shows that $(e_i \otimes f_j)_{i=1,\ldots,d, j=1,\ldots,d'}$ forms a basis for $\mathbb{R}^d \otimes \mathbb{R}^{d'}$. $\square$

Note that the lemma also shows that the dimension of $\mathbb{R}^d \otimes \mathbb{R}^{d'}$ is dd'.

2.2.2 The space $M(n, p)$

The vector space $M(n, p)$ of real $n \times p$ matrices is obviously an np-dimensional real vector space and thus isomorphic to $\mathbb{R}^{np}$. We have a standard basis $(e_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$ where e_{ij} is the matrix with a 1 at position (i, j) and 0 elsewhere, which can be used to make the identification. However, this vector space has more structure than $\mathbb{R}^{np}$. Trying to understand the tensor product $M(n, p) \otimes M(n, p)$ we may refer to the former section and regard it as the set of $M(np, np)$ matrices, but this is not a very good idea.

The first important structure on matrices that we will exploit is that for square matrices we have a so-called trace. The trace is the map $\mathrm{tr} : M(d, d) \to \mathbb{R}$ defined by

$$\mathrm{tr}(A) = \sum_{i=1}^{d} A_{ii}.$$

It has the following important properties for two matrices $A \in M(n,p)$ and $B \in M(p,n)$ that

$$\begin{aligned} \operatorname{tr}(AB) &= \operatorname{tr}(BA). && (2.2)\\ \operatorname{tr}(\lambda A + \gamma B) &= \lambda \operatorname{tr}(A) + \gamma \operatorname{tr}(B), \quad \text{for } \lambda, \gamma \in \mathbb{R}. && (2.3)\\ \operatorname{tr}(A) &= \operatorname{tr}(A^t), \quad \text{if } n = p && (2.4) \end{aligned}$$

To identify $M(n,p)$ with it's own dual using the standard basis – or to introduce the standard inner product on $M(n,p)$ – we can use the trace together with transposition. For $A \in M(n,p)$ we can define the corresponding linear function $\operatorname{tr}(A^t \cdot)$, that is, the function

$$B \mapsto \operatorname{tr}(A^t B).$$

The inner product is

$$\operatorname{tr}(A^t B) = \sum_{i=1}^{p} \left(\sum_{j=1}^{n} A_{ji} B_{ji} \right) = \sum_{i,j} A_{ij} B_{ij}.$$

Note that it is an algebraic trick to write the double sum as a matrix multiplication followed by a trace but by no means a computational trick. There is a large amount of computations involved in $A^t B$ that is not needed for computing the trace afterwards. To be precise, $A^t B$ requires $p^2 n$ multiplications and $p^2(n-1)$ additions. The following computation of the trace requires then in addition $p-1$ additions. But the double sum can be computed using pn multiplications and $pn - 1$ additions.

Definition 2.2.3. *The tensor product $M(m,n) \otimes M(p,q)$ is defined as $\mathcal{B}(M(m,n), M(p,q))$, that is, as the set of bilinear functions from $M(m,n) \times M(p,q)$ into $\mathbb{R}$.*

For two matrices $A \in M(m,n)$ and $B \in M(p,q)$ we have the elementary tensor $A \otimes B$, which is given as the bilinear function

$$A \otimes B(C,D) = \operatorname{tr}(A^t C)\operatorname{tr}(B^t D)$$

for $C \in M(m,n)$ and $D \in M(p,q)$. Moreover, we have the basis $(e_{ir} \otimes e_{js})_{i,j=1,\ldots,n,r,s=1,\ldots,p}$ for $M(n,p) \otimes M(n,p)$. For the elementary tensor $A \otimes B$ the coordinates in this basis are

$$(A \otimes B)_{irjs} = A_{ir} B_{js}.$$

There is, however, another fruitful construction of tensors that rely heavily on matrix algebra. For $A \in M(n,n)$ and $B \in M(p,p)$ we define the *kronecker* tensor product $A \otimes B \in M(n,p) \otimes M(n,p)$ as the bilinear function

$$(A \otimes B)(C,D) := \operatorname{tr}(B C^t A^t D). \tag{2.5}$$

We choose (2.5) as the equation that defines the tensor product $A \otimes B$ as a bilinear function. By the general correspondence in Lemma 1.2.2 we may also regard $A \otimes B$ as a linear map from $M(n,p)$ to $M(n,p)$, which is formally written as $\Phi_\psi(A \otimes B)$. We see that

$$\Phi_\psi(A \otimes B)(C) = ACB^t.$$

Thus dropping the isomorphism Φ_ψ we simply write $A \otimes B$ for the linear map also, that is,

$$(A \otimes B)(C) = ACB^t \tag{2.6}$$

for $C \in M(n,p)$. The coordinates in the standard basis are

$$(A \otimes B)_{irjs} = A_{ij} B_{rs}.$$

The composition of two elements in $M(n,p) \otimes M(n,p)$ is defined abstractly using the isomorphism Φ_ψ. We find that regarded as linear maps

$$(A \otimes B) \circ (C \otimes D)(E) = (A \otimes B)(CED^t) = ACED^t B^t = (AC)E(BD)^t = (AC \otimes BD)(E).$$

Thus we have proved the general formula for the kronecker tensor product;

$$(A \otimes B) \circ (C \otimes D) = AC \otimes BD$$

for $A, C \in M(n,n)$ and $B, D \in M(p,p)$.

In principle there is a chance of confusion. The notation for the kronecker tensor product of two matrices could also be taken to mean an elementary tensor in $M(n,n) \otimes M(p,p)$. Especially if $n = p$ we could be in serious trouble. Does $A \otimes B$ mean an elementary tensor or the kronecker tensor product for two $p \times p$ matrices A and B?

Example 2.2.4. Let X_0 and Z be independent, real valued p-dimensional random variables with second moment and define[1]

$$\begin{aligned} X_1 &= \Omega_1^{1/2} Z \\ \Omega_1 &= \Omega + A X_0 X_0^t A^t \end{aligned}$$

for two matrices $\Omega, A \in M(p,p)$, where Ω is symmetric and positive definite. We assume that Z has mean 0 and the identity matrix I as covariance matrix.

We ask if there is a distribution on X_0 such that $X_1 \overset{\mathcal{D}}{=} X_0$. A more modest question is if we can find a mean value and covariance matrix Σ of X_0 such that X_1 has the same mean value and covariance matrix[2]

[1] the formula is used in particular in econometrics to define recursively a class of stochastic processes known as ARCH processes

[2] the interest is whether there is a time stationary version of the process, or perhaps a so-called weakly stationary version where just the fist and second moments do not change with time

We compute the mean in the defining equation and find that $\mathbb{E}X_1 = 0$, thus X_0 should have mean 0. Then we compute the variance of X_1 to be

$$\mathbb{V}(X_1) = \mathbb{E}\Omega_1 = \Omega + A\Sigma A^t$$

where we have used that if $\mathbb{E}X_0 = 0$ then $\mathbb{E}X_0X_0^t = \Sigma$. Thus to get the same covariance matrix for X_1 we need that Σ fulfills the equation

$$\Sigma - A\Sigma A^t = \Omega.$$

The interest is whether there is a matrix solution in Σ to this equation and whether the solution if it exists is symmetric and positive definite.

By relying on the definition of the tensor product we see that in reality we are trying to solve the linear equation

$$(I - A \otimes A)\Sigma = \Omega$$

where I denotes the identity map from $M(p,p)$ to $M(p,p)$. The linear map $(I - A \otimes A)$ is invertible if the eigenvalues of $A \otimes A$ are < 1. In this case the inverse is given by the Neumann series

$$(I - A \otimes A)^{-1} = \sum_{k=0}^{\infty} (A \otimes A)^k,$$

which converges and $(I - A \otimes A)^{-1}\Omega$ is positive definite for all positive definite Ω. This shows that if the eigenvalues of $A \otimes A$ are < 1 then

$$\Sigma = \sum_{k=0}^{\infty} (A \otimes A)^k \Omega = \sum_{k=0}^{\infty} A^k \Omega (A^k)^t$$

is the unique positive definite solution.

2.3 Exercises

Exercise 2.3.1. Lemma 2.2.2 allows us to regard elements in $\mathbb{R}^n \otimes \mathbb{R}^p$ as matrices in $M(n,p)$. Show the the matrix representation of the elementary tensor $x \otimes y$ for $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$ is xy^t.

In Lemma 1.2.2 an isomorphism Φ_ψ from $\mathbb{R}^n \otimes \mathbb{R}^p = \mathcal{B}(\mathbb{R}^n, \mathbb{R}^p)$ to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$ is established if we choose the standard bases for $\mathbb{R}^n$ and $\mathbb{R}^p$. Show that the matrix representation of the linear map $\Phi_\psi(x \otimes y)$ is yx^t.

Exercise 2.3.2. For $A \in M(n,n)$ and $B \in M(p,p)$ the kronecker tensor product $A \otimes B$ can by (2.6) be regarded as a linear map on $M(n,p) \to M(n,p)$. Using Lemma 2.2.2 we identify $M(n,p)$ by $\mathbb{R}^n \otimes \mathbb{R}^p$, and thus we can regard $A \otimes B$ as a linear map on $\mathbb{R}^n \otimes \mathbb{R}^p \to \mathbb{R}^n \otimes \mathbb{R}^p$. Writing $(A \otimes B)(C)$ for $C \in \mathbb{R}^n \otimes \mathbb{R}^p$ show that

$$(A \otimes B)(x \otimes y) = (Ax) \otimes (By)$$

for $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$.