Principal Component Analysis and Factor Models: An Exact Exposition of the Core Theory

Prepared for J. Keogh

August 2026

Abstract

This document develops the core theory of principal component analysis (PCA) and of factor models from first principles. Every term is defined before use. Every claim is proved, with two declared exceptions: the spectral theorem for real symmetric matrices and the singular value decomposition, which are stated as known results (Section 2). Existence is *proved*, not assumed, for both PCA (Theorem 3.2) and factor models (Theorem 5.4). A section on cross-sectional (fundamental) factor models covers the estimation architecture of commercial equity risk models. A further section itemises every modelling choice in that architecture that is imposed rather than estimated. A final section maps the conflicting vocabularies of the PCA, factor-analysis, and finance literatures onto the exact objects defined here.

Contents

1 Probabilistic foundations

Throughout, p, k, n are positive integers. Vectors in $\mathbb{R}^p$ are column vectors. $\langle u, v\rangle = u^\top v$ and $\|u\|_2 = \sqrt{u^\top u}$ denote the Euclidean inner product and norm. $e_1, \dots, e_p$ is the standard basis of $\mathbb{R}^p$, and I_p the identity matrix.

Definition 1.1 (measure; finite measure; probability measure)**.** Let $(\Omega, \mathcal{F})$ be a *measurable space*: a set Ω together with a σ-algebra $\mathcal{F}$ of subsets of Ω. A *measure* is a function $\nu : \mathcal{F} \to [0, \infty]$ with $\nu(\emptyset) = 0$ that is countably additive: for every sequence $A_1, A_2, \dots \in \mathcal{F}$ of pairwise disjoint sets, $\nu\big(\bigcup_{i\geq 1} A_i\big) = \sum_{i\geq 1} \nu(A_i)$. The measure ν is *finite* if $\nu(\Omega) < \infty$, and is a *probability measure* if $\nu(\Omega) = 1$. Every probability measure is a finite measure. (Contrast: Lebesgue measure on $\mathbb{R}$ is not finite.)

Definition 1.2 (random vector)**.** Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, held fixed for the entire document. A *random vector in* $\mathbb{R}^p$ is a measurable map $X : \Omega \to \mathbb{R}^p$, where $\mathbb{R}^p$ carries its Borel σ-algebra. Writing $X = (X_1, \dots, X_p)^\top$, each coordinate $X_i : \Omega \to \mathbb{R}$ is a real random variable, and *all coordinates live on the same probability space.*

Remark 1.3 (one joint law, many marginals). A random vector induces exactly one probability measure on $\mathbb{R}^p$, the *joint distribution* $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$. Each coordinate X_i has a *marginal distribution*, the pushforward of $\mathbb{P}_X$ under the coordinate projection $x \mapsto x_i$. The marginals are in general all different from one another, and no independence between coordinates is assumed anywhere in this document. Questions of the form "do the coordinates have the same distribution?" have no forced answer: the theory below constrains only second moments of the joint law.

Assumption 1.4 (finite second moments)**.** $\mathbb{E}[X_i^2] < \infty$ for every $i \in \{1, \dots, p\}$; equivalently $\mathbb{E}\|X\|_2^2 < \infty$, since $\|X\|_2^2 = \sum_{i=1}^p X_i^2$ and expectation is linear and monotone. Every random vector in this document satisfies Assumption 1.4. In the language of function spaces: each X_i lies in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, the space of square-integrable random variables.

Assumption 1.4 is not decoration; it is exactly what makes the following objects exist.

Lemma 1.5 (existence of mean and covariance)**.** *Under Assumption 1.4, the quantities*

$$\mu_i := \mathbb{E}[X_i], \qquad \Sigma_{ij} := \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big]$$

exist and are finite for all $i, j \in \{1, \dots, p\}$.

Proof. Since $\mathbb{P}$ is a finite measure (Definition 1.1), the constant function 1 is square-integrable with $\mathbb{E}[1^2] = \mathbb{P}(\Omega) = 1$, and the Cauchy–Schwarz inequality gives

$$\mathbb{E}|X_i| = \mathbb{E}\big[\,|X_i| \cdot 1\,\big] \leq \big(\mathbb{E}[X_i^2]\big)^{1/2}\big(\mathbb{E}[1^2]\big)^{1/2} = \big(\mathbb{E}[X_i^2]\big)^{1/2} < \infty,$$

so μ_i exists and is finite. This is the inclusion $L^2 \subseteq L^1$, and it is where finiteness of the measure is used: on an infinite-measure space the inclusion fails — on $\mathbb{R}$ with Lebesgue measure, $f(x) = (1 + |x|)^{-1}$ has $\int f^2 < \infty$ but $\int f = \infty$. Then $X_i - \mu_i \in L^2$ because L^2 is a vector space containing the constants. Cauchy–Schwarz again: $\mathbb{E}\big|(X_i - \mu_i)(X_j - \mu_j)\big| \leq \|X_i - \mu_i\|_{L^2}\,\|X_j - \mu_j\|_{L^2} < \infty$, so Σ_{ij} exists and is finite. □

Definition 1.6 (mean vector, covariance matrix)**.** Under Assumption 1.4, the *mean vector* is $\mu := \mathbb{E}[X] = (\mu_1, \dots, \mu_p)^\top$ and the *covariance matrix* is

$$\Sigma := \mathrm{Cov}(X) := \mathbb{E}\big[(X - \mu)(X - \mu)^\top\big] \in \mathbb{R}^{p\times p},$$

the matrix with entries Σ_{ij} of Lemma 1.5. For two random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ on the same probability space, both satisfying Assumption 1.4, the *cross-covariance matrix* is

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)^\top\big] \in \mathbb{R}^{p\times q},$$

which exists by the same Cauchy–Schwarz argument. Note $\mathrm{Cov}(X) = \mathrm{Cov}(X, X)$ and $\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)^\top$.

Lemma 1.7 (bilinearity of covariance)**.** *For constant matrices $A \in \mathbb{R}^{m\times p}$, $B \in \mathbb{R}^{l\times q}$, constant vectors $a \in \mathbb{R}^m$, $b \in \mathbb{R}^l$, and random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ as above:*

$$\operatorname{Cov}(AX + a,\ BY + b) = A\,\operatorname{Cov}(X, Y)\,B^\top.$$

In particular $\operatorname{Cov}(AX + a) = A\Sigma A^\top$ *and, for* $w \in \mathbb{R}^p$, $\operatorname{Var}(w^\top X) := \operatorname{Cov}(w^\top X) = w^\top \Sigma w$.

Proof. $\mathbb{E}[AX + a] = A\mu_X + a$ by linearity of $\mathbb{E}$, so $(AX + a) - \mathbb{E}[AX + a] = A(X - \mu_X)$, and likewise for Y. Then $\operatorname{Cov}(AX+a, BY+b) = \mathbb{E}\big[A(X-\mu_X)(Y-\mu_Y)^\top B^\top\big] = A\,\mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)^\top\big]B^\top$, pulling constant matrices out of the (entrywise) expectation by linearity. □

Lemma 1.8 (Σ is symmetric PSD)**.** $\Sigma = \Sigma^\top$, *and* $w^\top \Sigma w \geq 0$ *for all* $w \in \mathbb{R}^p$ *(*positive semidefinite*, PSD).*

Proof. Symmetry: $\Sigma_{ij} = \Sigma_{ji}$ directly from Lemma 1.5. PSD: by Lemma 1.7, $w^\top \Sigma w = \operatorname{Var}(w^\top X) = \mathbb{E}\big[(w^\top X - w^\top \mu)^2\big] \geq 0$. □

Definition 1.9 (Pearson correlation; correlation matrix)**.** Whenever the word *correlation* appears in this document, it means the Pearson correlation: for random variables $U, V \in L^2$ with $\operatorname{Var}(U) > 0$ and $\operatorname{Var}(V) > 0$,

$$\operatorname{Corr}(U, V) := \frac{\operatorname{Cov}(U, V)}{\sqrt{\operatorname{Var}(U)\,\operatorname{Var}(V)}}.$$

If $\Sigma_{ii} > 0$ for all i, the *correlation matrix* of X is $R := D^{-1/2}\,\Sigma\,D^{-1/2}$ with $D := \operatorname{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$, so that $R_{ij} = \operatorname{Corr}(X_i, X_j)$ and $R_{ii} = 1$. When the phrase *cross-correlation* between vectors X, Y appears, it means the matrix $\operatorname{Cov}(X, Y)$ of Definition 1.6, or its entrywise Pearson-normalised version; each use below states which.

Lemma 1.10 (centering changes nothing PCA will look at)**.** *Let X be a random vector in $\mathbb{R}^p$ with mean μ, and set $Y := X - \mu$. Then:*

(a) Y is a random vector in $\mathbb{R}^p$ — the same *p. The map $T(x) = x - \mu$ is a measurable bijection of $\mathbb{R}^p$ with measurable inverse; no dimension is created or destroyed.*

(b) $\mathbb{E}[Y] = 0$ and $\operatorname{Cov}(Y) = \operatorname{Cov}(X) = \Sigma$.

(c) For every $w \in \mathbb{R}^p$, $\operatorname{Var}(w^\top Y) = \operatorname{Var}(w^\top X)$.

Consequently any optimisation problem whose objective and constraints are built only from $\{\operatorname{Var}(w^\top \cdot) : w \in \mathbb{R}^p\}$ has identical feasible sets, objective values, maximisers and maxima for X and for Y.

Proof. (a) Y is a composition of measurable maps; $T^{-1}(y) = y + \mu$. (b) $\mathbb{E}[Y] = \mu - \mu = 0$; apply Lemma 1.7 with $A = I_p$, $a = -\mu$. (c) Lemma 1.7: both sides equal $w^\top \Sigma w$. □

Convention 1.11. By Lemma 1.10 we assume $\mu = 0$ from Section 3 onward, without loss of generality. (A genuine dimension phenomenon connected with *sample* centering exists; it is a rank statement, not an ambient-dimension statement, and is proved as Lemma 4.2.)

2 Linear-algebra toolkit

Two classical results are used without proof; these are the only unproved statements in the document.

Theorem 2.1 (spectral theorem; stated without proof)**.** *Every symmetric $A \in \mathbb{R}^{p\times p}$ admits a* spectral decomposition *$A = Q\Lambda Q^\top$, where $Q \in \mathbb{R}^{p\times p}$ is orthogonal ($Q^\top Q = I_p$) and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p$ real. The columns $q_1, \ldots, q_p$ of Q form an orthonormal basis of $\mathbb{R}^p$ with $Aq_j = \lambda_j q_j$; the λ_j are the eigenvalues of A, listed with multiplicity. The ordering $\lambda_1 \geq \cdots \geq \lambda_p$ is fixed for the whole document.*

Theorem 2.2 (singular value decomposition; stated without proof)**.** *Every $M \in \mathbb{R}^{n\times p}$ of rank r admits a decomposition $M = UDV^\top$ where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{p\times r}$ have orthonormal columns and $D = \operatorname{diag}(d_1, \dots, d_r)$ with $d_1 \geq \cdots \geq d_r > 0$. The d_j are the* singular values*; columns of U and V are left and right* singular vectors*.*

Three consequences are proved, because the main theorems consume them.

Lemma 2.3 (eigenvalues of a PSD matrix)**.** *If A is symmetric PSD with spectral decomposition $A = Q\Lambda Q^\top$, then $\lambda_j \geq 0$ for every j; and $\operatorname{tr}(A) = \sum_{j=1}^{p} \lambda_j$.*

Proof. $\lambda_j = q_j^\top A q_j \geq 0$ by PSD-ness. For the trace, $\operatorname{tr}(A) = \operatorname{tr}(Q\Lambda Q^\top) = \operatorname{tr}(\Lambda Q^\top Q) = \operatorname{tr}(\Lambda) = \sum_j \lambda_j$, using the cyclic property $\operatorname{tr}(BC) = \operatorname{tr}(CB)$, which holds since both sides equal $\sum_{i,j} B_{ij}C_{ji}$. □

Lemma 2.4 (eigenbasis completion)**.** *Let $A \in \mathbb{R}^{p\times p}$ be symmetric with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ (with multiplicity), and let $w_1, \dots, w_m$ ($m < p$) be orthonormal eigenvectors with $Aw_j = \lambda_j w_j$ for $j = 1, \dots, m$. Then there exist vectors $w_{m+1}, \dots, w_p$ such that $w_1, \dots, w_p$ is an orthonormal basis of $\mathbb{R}^p$ of eigenvectors of A with $Aw_j = \lambda_j w_j$ for* all *j; that is, the completion carries exactly the remaining eigenvalues $\lambda_{m+1}, \dots, \lambda_p$, with multiplicity.*

Proof. Let $W := \operatorname{span}(w_1, \dots, w_m)$ and let $W^\perp$ be its orthogonal complement, of dimension $p - m$.

Step 1: $W^\perp$ is A-invariant. If $v \in W^\perp$, then for each $j \leq m$, $\langle Av, w_j\rangle = \langle v, Aw_j\rangle = \lambda_j\langle v, w_j\rangle = 0$ (symmetry of A), so $Av \in W^\perp$.

Step 2: diagonalise the restriction. Fix any orthonormal basis $b_1, \dots, b_{p-m}$ of $W^\perp$ and let $B = [\, b_1 \cdots b_{p-m}\,] \in \mathbb{R}^{p\times(p-m)}$. The matrix $A' := B^\top AB \in \mathbb{R}^{(p-m)\times(p-m)}$ is symmetric, so by Theorem 2.1 it has an orthonormal eigenbasis $u_1, \dots, u_{p-m}$ with eigenvalues $\nu_1 \geq \cdots \geq \nu_{p-m}$. Set $w_{m+i} := Bu_i \in W^\perp$. These are orthonormal ($\langle Bu_i, Bu_j\rangle = u_i^\top B^\top Bu_j = u_i^\top u_j$) and are eigenvectors of A: since $ABu_i \in W^\perp$ by Step 1 and $BB^\top$ acts as the identity on $W^\perp$,

$$Aw_{m+i} = ABu_i = BB^\top ABu_i = BA'u_i = \nu_i Bu_i = \nu_i\, w_{m+i}.$$

Together with $w_1, \dots, w_m$ this gives an orthonormal basis of $\mathbb{R}^p$ consisting of eigenvectors of A.

Step 3: the completion carries exactly the remaining eigenvalues. In the orthonormal basis $w_1, \dots, w_p$, A is represented by $\operatorname{diag}(\lambda_1, \dots, \lambda_m, \nu_1, \dots, \nu_{p-m})$. The characteristic polynomial is basis-invariant, so the multiset $\{\lambda_1, \dots, \lambda_m\} \cup \{\nu_1, \dots, \nu_{p-m}\}$ equals the multiset $\{\lambda_1, \dots, \lambda_p\}$ of all eigenvalues of A. Cancelling the shared multiset $\{\lambda_1, \dots, \lambda_m\}$ leaves $\{\nu_1, \dots, \nu_{p-m}\} = \{\lambda_{m+1}, \dots, \lambda_p\}$ as multisets; since both lists are sorted decreasingly, $\nu_i = \lambda_{m+i}$ for each i. □

Lemma 2.5 (Ky Fan trace bound)**.** *Let A be symmetric PSD with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and spectral decomposition $A = Q\Lambda Q^\top$. Let $P \in \mathbb{R}^{p\times p}$ be an orthogonal projection (i.e. $P = P^\top = P^2$) with $\operatorname{rank}(P) \leq k$. Then*

$$\operatorname{tr}(AP) \;\leq\; \sum_{j=1}^{k} \lambda_j,$$

with equality when $P = \sum_{j=1}^{k} q_j q_j^\top$.

Proof. Set $c_i := q_i^\top P q_i$. Each $c_i \in [0, 1]$: indeed $c_i = q_i^\top P^\top P q_i = \|Pq_i\|_2^2 \geq 0$, and $\|Pq_i\|_2 \leq \|q_i\|_2 = 1$ because an orthogonal projection is a contraction ($\|v\|^2 = \|Pv\|^2 + \|(I - P)v\|^2$ by the Pythagorean identity, both terms nonnegative). Moreover $\sum_{i=1}^{p} c_i = \operatorname{tr}(Q^\top PQ) =$

$\mathrm{tr}(P) = \mathrm{rank}(P) =: r \le k$, using that the trace of an orthogonal projection equals its rank (in an orthonormal basis adapted to its range, P is $\mathrm{diag}(1,\dots,1,0,\dots,0)$ with r ones). Then, by cyclicity of trace,

$$\mathrm{tr}(AP) = \mathrm{tr}(Q\Lambda Q^\top P) = \mathrm{tr}(\Lambda\, Q^\top P Q) = \sum_{i=1}^{p} \lambda_i c_i \;\le\; \sum_{j=1}^{r} \lambda_j \;\le\; \sum_{j=1}^{k} \lambda_j,$$

where the first inequality holds because, over the polytope $\{c \in [0,1]^p : \sum_i c_i = r\}$, the linear functional $c \mapsto \sum_i \lambda_i c_i$ with decreasing weights is maximised by $c = (1,\dots,1,0,\dots,0)$ (r ones): any feasible c can be transformed into this point by repeatedly moving mass from a lower-weight coordinate to a higher-weight one, and no such move decreases the objective. The second inequality uses $\lambda_j \ge 0$. Equality at $P = \sum_{j\le k} q_j q_j^\top$: then $c = (1,\dots,1,0,\dots,0)$ with k ones and $\mathrm{tr}(AP) = \sum_{j\le k} \lambda_j$. □

3 Principal component analysis: population version

Fix a random vector X in $\mathbb{R}^p$ satisfying Assumption 1.4, centered per Convention 1.11, with covariance matrix Σ and spectral decomposition $\Sigma = Q\Lambda Q^\top$, $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$ (nonnegativity by Lemmas 1.8 and 2.3).

Definition 3.1 (principal directions and components)**.** A *first principal direction* is any solution of

$$w_1 \in \operatorname*{arg\,max}_{\|w\|_2=1} \mathrm{Var}(w^\top X) = \operatorname*{arg\,max}_{\|w\|_2=1} w^\top \Sigma w.$$

Recursively, given principal directions $w_1,\dots,w_{m-1}$, an *m-th principal direction* is any solution of

$$w_m \in \operatorname*{arg\,max}_{\substack{\|w\|_2=1 \\ w\perp w_1,\dots,w_{m-1}}} w^\top \Sigma w.$$

The scalar random variable $Z_m := w_m^\top X$ is the *m-th principal component*. (Terminology is not uniform in the literature: some sources call the direction w_m "the principal component". This document reserves *component* for the random variable Z_m and *direction* for the vector w_m.)

The definition is a sequential optimisation problem; that solutions exist at every stage is a theorem, and the theorem also identifies all solutions.

Theorem 3.2 (existence and characterisation of PCA)**.** *For every $m \in \{1,\dots,p\}$:*

(a) the m-th maximisation problem of Definition 3.1 attains its maximum, and the maximum value is λ_m;

(b) the choice $w_m = q_m$ is valid; conversely, in every solution sequence, w_m is a unit eigenvector of Σ with eigenvalue λ_m;

(c) $\mathrm{Var}(Z_m) = \lambda_m$, and $\mathrm{Cov}(Z_j, Z_m) = 0$ for $j \ne m$.

Proof. Attainment. The feasible set of the m-th problem is the intersection of the unit sphere with the subspace $\{w : w \perp w_1,\dots,w_{m-1}\}$, which has dimension $p - m + 1 \ge 1$ and hence contains unit vectors. The feasible set is closed and bounded in $\mathbb{R}^p$, therefore compact (Heine–Borel), and nonempty; the objective $w \mapsto w^\top \Sigma w$ is a polynomial, hence continuous; a continuous function on a nonempty compact set attains its maximum. This proves existence at every stage. It remains to compute the value and identify the maximisers; we induct on m.

Base case $m = 1$. Substitute $y := Q^\top w$; since Q is orthogonal, $\|y\|_2 = \|w\|_2$ and the substitution is a bijection of the unit sphere. Then

$$w^\top \Sigma w = y^\top \Lambda y = \sum_{i=1}^{p} \lambda_i y_i^2 \;\le\; \lambda_1 \sum_{i=1}^{p} y_i^2 = \lambda_1,$$

with equality if and only if $y_i = 0$ for every i with $\lambda_i < \lambda_1$, i.e. if and only if w is a unit vector in the eigenspace of λ_1. In particular $w = q_1$ attains the value λ_1.

Inductive step. Suppose $w_1, \dots, w_{m-1}$ are orthonormal eigenvectors with $\Sigma w_j = \lambda_j w_j$, $j < m$, each attaining its stage's maximum. By Lemma 2.4, complete them to an orthonormal eigenbasis $\tilde{q}_1 := w_1, \;\dots,\; \tilde{q}_{m-1} := w_{m-1}, \; \tilde{q}_m, \dots, \tilde{q}_p$ of Σ with $\Sigma \tilde{q}_j = \lambda_j \tilde{q}_j$ for all j. Let $\tilde{Q} := [\tilde{q}_1 \cdots \tilde{q}_p]$ and substitute $y := \tilde{Q}^\top w$ as before. The constraints $w \perp w_1, \dots, w_{m-1}$ read $y_1 = \cdots = y_{m-1} = 0$, so on the feasible set

$$w^\top \Sigma w = \sum_{i=m}^{p} \lambda_i y_i^2 \;\le\; \lambda_m \sum_{i=m}^{p} y_i^2 = \lambda_m,$$

with equality if and only if $y_i = 0$ whenever $\lambda_i < \lambda_m$ (for $i \ge m$), i.e. if and only if w is a unit vector in the span of those $\tilde{q}_i$, $i \ge m$, whose eigenvalue equals λ_m — all of which are eigenvectors of Σ with eigenvalue λ_m. The maximum λ_m is attained, e.g. at $w = \tilde{q}_m$, and only at such eigenvectors. This proves (a) and (b).

(c). By Lemma 1.7, $\mathrm{Var}(Z_m) = w_m^\top \Sigma w_m = \lambda_m$ and, for $j \ne m$, $\mathrm{Cov}(Z_j, Z_m) = w_j^\top \Sigma w_m = \lambda_m \, w_j^\top w_m = 0$ by orthogonality. □

Remark 3.3 (uniqueness). The direction w_m is unique up to sign if and only if, at stage m, the relevant eigenvalue is simple within the remaining spectrum. Repeated eigenvalues produce whole eigenspaces of valid choices. Two software packages may therefore legitimately return different "PCA loadings" on the same input when eigenvalues are (nearly) tied.

Definition 3.4 (variance explained)**.** By Lemma 2.3, $\mathrm{tr}(\Sigma) = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \sum_{j=1}^{p} \lambda_j$. The quantity $\lambda_m \big/ \sum_{j=1}^{p} \lambda_j$ is the *proportion of variance explained* by the m-th component. "Total variance" means $\mathrm{tr}(\Sigma)$ — nothing else.

Theorem 3.5 (reconstruction optimality)**.** *Let* $W_k := [\, w_1 \cdots w_k \,]$ *be a matrix of principal directions from Theorem 3.2 and let* $P_k := W_k W_k^\top$*, the orthogonal projection onto* $\mathrm{span}(w_1, \dots, w_k)$*. Then for every orthogonal projection* P *of rank at most* k*,*

$$\mathbb{E}\,\|X - P_k X\|_2^2 \;\le\; \mathbb{E}\,\|X - PX\|_2^2,$$

and the common minimum value is $\sum_{j=k+1}^{p} \lambda_j$*. Thus "maximise retained variance" and "minimise expected squared reconstruction error over rank-*k *orthogonal projections" are the same problem. The random vector* $\hat{X} := P_k X$ *is the* rank-k PCA reconstruction *of* X*.*

Proof. For any orthogonal projection P, the matrix $I - P$ is also an orthogonal projection ($(I-P)^\top = I - P$ and $(I-P)^2 = I - 2P + P^2 = I - P$). Using $\mathbb{E}[X] = 0$, $\|v\|_2^2 = \mathrm{tr}(vv^\top)$, and linearity of trace and expectation,

$$\begin{aligned}\mathbb{E}\|(I-P)X\|_2^2 &= \mathbb{E}\,\mathrm{tr}\big((I-P)XX^\top(I-P)\big) = \mathrm{tr}\big((I-P)\,\Sigma\,(I-P)\big) \\ &= \mathrm{tr}\big(\Sigma(I-P)^2\big) = \mathrm{tr}(\Sigma) - \mathrm{tr}(\Sigma P),\end{aligned}$$

where the third equality is cyclicity of trace and the fourth is idempotence. Minimising $\mathbb{E}\|X - PX\|_2^2$ over rank-$\le k$ orthogonal projections is therefore equivalent to maximising $\mathrm{tr}(\Sigma P)$. By Lemma 2.5, $\mathrm{tr}(\Sigma P) \le \sum_{j \le k} \lambda_j$ with equality at $P = \sum_{j \le k} q_j q_j^\top$; and P_k itself achieves equality because each w_j is a unit eigenvector for λ_j (Theorem 3.2(b)), so $\mathrm{tr}(\Sigma P_k) = \sum_{j \le k} w_j^\top \Sigma w_j = \sum_{j \le k} \lambda_j$. The minimum value is $\mathrm{tr}(\Sigma) - \sum_{j \le k} \lambda_j = \sum_{j > k} \lambda_j$ by Lemma 2.3. □

Definition 3.6 (the word "loading" in PCA, pinned down)**.** The PCA literature uses *loading* in two incompatible ways:

(L1) *weight convention*: the loading of variable i on component m is $(w_m)_i$, the i-th entry of the unit-norm direction; the *loading matrix* is then Q (orthonormal columns);

(L2) *scaled convention*: the loading is $(w_m)_i\sqrt{\lambda_m}$; the loading matrix is then $Q\Lambda^{1/2}$, where $\Lambda^{1/2} := \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$.

Proposition 3.7 (exact content of the scaled convention)**.** *For each variable i and component m,*

$$\mathrm{Cov}(X_i, Z_m) = \lambda_m (w_m)_i, \qquad \text{and, if } \Sigma_{ii} > 0 \text{ and } \lambda_m > 0, \qquad \mathrm{Corr}(X_i, Z_m) = \frac{(w_m)_i \sqrt{\lambda_m}}{\sqrt{\Sigma_{ii}}}.$$

Hence: if *the variables are standardised ($\Sigma_{ii} = 1$ for all i, i.e. PCA is run on the correlation matrix R of Definition 1.9), the scaled loading (L2)* equals *the Pearson correlation between variable i and component m. Under any other circumstances, "loadings are correlations" is false.*

Proof. By Lemma 1.7 with $A = e_i^\top$, $B = w_m^\top$: $\mathrm{Cov}(X_i, Z_m) = e_i^\top \Sigma w_m = \lambda_m e_i^\top w_m = \lambda_m (w_m)_i$. Divide by $\sqrt{\Sigma_{ii}}\sqrt{\mathrm{Var}(Z_m)} = \sqrt{\Sigma_{ii}}\sqrt{\lambda_m}$ (Theorem 3.2(c)) to get the correlation formula; substitute $\Sigma_{ii} = 1$ for the final claim. □

Warning 3.8 (covariance PCA vs correlation PCA). PCA applied to Σ and PCA applied to R are different problems with different eigenvectors; no formula maps one solution to the other in general. PCA is not invariant under rescaling of individual coordinates: replacing X_i by $c\,X_i$ changes Σ by a non-orthogonal congruence and changes the maximisers. Choosing R is choosing to force every variable to variance 1 first. Always determine which matrix a given source diagonalised.

4 Sample PCA

Definition 4.1 (data matrix; sample covariance)**.** A *data matrix* is $\mathbf{X} \in \mathbb{R}^{n \times p}$, row i being observation i. Define the columnwise mean $\bar{x} := \frac{1}{n}\mathbf{X}^\top \mathbf{1}_n \in \mathbb{R}^p$ (where $\mathbf{1}_n$ is the all-ones vector), the *centered data matrix* $\mathbf{X}_c := \mathbf{X} - \mathbf{1}_n \bar{x}^\top$, and the *sample covariance matrix*

$$S := \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{X}_c \in \mathbb{R}^{p \times p}.$$

(Some authors divide by n; this scales every eigenvalue by the constant $\frac{n-1}{n}$ and changes no eigenvector.) *Sample PCA* is Definition 3.1 with Σ replaced by S; since S is symmetric and PSD ($w^\top S w = \frac{1}{n-1}\|\mathbf{X}_c w\|_2^2 \geq 0$), Theorem 3.2 applies verbatim.

Lemma 4.2 (centering matrix; the true rank-drop fact)**.** *Let $C := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Then:*

(a) $\mathbf{X}_c = C\mathbf{X}$*;*

(b) C is the orthogonal projection of $\mathbb{R}^n$ onto $\{v \in \mathbb{R}^n : \mathbf{1}_n^\top v = 0\}$, a subspace of dimension $n-1$;

(c) consequently every column of $\mathbf{X}_c$ lies in that subspace, so $\mathrm{rank}(\mathbf{X}_c) \leq \min(n-1, p)$ and $\mathrm{rank}(S) \leq n-1$; when $p \geq n$, S has at least $p-(n-1)$ zero eigenvalues.

Sample centering thus pins one linear degree of freedom of the n observations in $\mathbb{R}^n$; it does not reduce the ambient dimension p of the variable space, and it has no population analogue (Lemma 1.10).

Proof. (a) $C\mathbf{X} = \mathbf{X} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n^\top \mathbf{X}) = \mathbf{X} - \mathbf{1}_n\bar{x}^\top$. (b) $C^\top = C$, and $C^2 = I - \frac{2}{n}\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}\mathbf{1}(\mathbf{1}^\top\mathbf{1})\mathbf{1}^\top = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top = C$, so C is an orthogonal projection; $C\mathbf{1} = 0$ and $Cv = v$ for $v \perp \mathbf{1}$, so its range is $\{\mathbf{1}\}^\perp$, of dimension $n-1$. (c) Columns of $C\mathbf{X}$ lie in the range of C; a matrix whose columns lie in an $(n-1)$-dimensional subspace has rank at most $n-1$, and rank at most p trivially. $\operatorname{rank}(S) = \operatorname{rank}(\mathbf{X}_c^\top\mathbf{X}_c) = \operatorname{rank}(\mathbf{X}_c)$, since $\mathbf{X}_c^\top\mathbf{X}_c v = 0 \Rightarrow \|\mathbf{X}_c v\|^2 = v^\top\mathbf{X}_c^\top\mathbf{X}_c v = 0 \Rightarrow \mathbf{X}_c v = 0$, so the two matrices have the same null space. A symmetric PSD $p \times p$ matrix of rank r has $p - r$ zero eigenvalues by Theorem 2.1. □

Proposition 4.3 (the SVD computes sample PCA)**.** *Let $\mathbf{X}_c = UDV^\top$ be an SVD (Theorem 2.2) of rank r. Then*

$$S = \frac{1}{n-1} V D^2 V^\top,$$

which exhibits the nonzero part of a spectral decomposition of S. Hence: the sample principal directions with positive variance are the columns of V; the corresponding sample eigenvalues are $\hat{\lambda}_j = d_j^2/(n-1)$; and the matrix of scores *— entry (i, j) being the value of the j-th sample principal component at observation i — is*

$$\mathbf{X}_c V = UD.$$

Proof. $\mathbf{X}_c^\top\mathbf{X}_c = VDU^\top UDV^\top = VD^2V^\top$ since $U^\top U = I_r$; the columns of V are orthonormal and D^2 is diagonal with decreasing positive entries, and any orthonormal completion of V (Lemma 2.4 applied to S) carries eigenvalue 0. Finally $\mathbf{X}_c V = UDV^\top V = UD$. □

5 Factor models: latent version

A factor model is *not* an optimisation problem. It is a structural hypothesis about the distribution of X. This is the single most important contrast with PCA.

Definition 5.1 (k-factor model)**.** A random vector X in $\mathbb{R}^p$ (Assumption 1.4) *admits a k-factor model* if there exist, on the same probability space:

- a random vector $F \in \mathbb{R}^k$ (the *common factors*) with $\mathbb{E}[F] = 0$ and $\operatorname{Cov}(F) = I_k$;
- a random vector $\varepsilon \in \mathbb{R}^p$ (the *idiosyncratic errors*, or *specific factors*) with $\mathbb{E}[\varepsilon] = 0$ and $\operatorname{Cov}(\varepsilon) = \Psi$ *diagonal*, Ψ PSD;
- the orthogonality condition $\operatorname{Cov}(F, \varepsilon) = 0$ (the $k \times p$ cross-covariance matrix of Definition 1.6 is the zero matrix);
- a constant matrix $\Lambda \in \mathbb{R}^{p\times k}$ (the *factor loading matrix*);

such that

$$X = \mu + \Lambda F + \varepsilon.$$

Three normalisations are built in, and they are conventions, not laws: $\operatorname{Cov}(F) = I_k$ (the *orthogonal factor model*); diagonality of Ψ (the substantive assumption: idiosyncratic errors are mutually uncorrelated); and uncorrelatedness of F and ε.

Proposition 5.2 (what "loading" means here, exactly)**.** *Under Definition 5.1,*

$$\operatorname{Cov}(X, F) = \Lambda, \qquad \textit{i.e.} \qquad \Lambda_{ij} = \operatorname{Cov}(X_i, F_j),$$

and if $\Sigma_{ii} > 0$ then $\operatorname{Corr}(X_i, F_j) = \Lambda_{ij}/\sqrt{\Sigma_{ii}}$ (each F_j has variance 1 by assumption). In factor analysis, "loading" therefore unambiguously means covariance-with-factor — and correlation-with-factor when the variables are standardised. This matches the scaled *PCA convention (L2) of Definition 3.6, not the weight convention (L1); the collision of these conventions is one root of the terminology confusion between the fields.*

Proof. By Lemma 1.7, $\operatorname{Cov}(X, F) = \operatorname{Cov}(\Lambda F + \varepsilon,\ F) = \Lambda \operatorname{Cov}(F) + \operatorname{Cov}(\varepsilon, F) = \Lambda I_k + 0 = \Lambda$. Divide entry (i, j) by $\sqrt{\Sigma_{ii} \cdot 1}$ for the correlation statement. □

Proposition 5.3 (implied covariance structure)**.** *Under Definition 5.1,*

$$\Sigma = \Lambda\Lambda^\top + \Psi.$$

In particular $\operatorname{Var}(X_i) = \sum_{j=1}^{k} \Lambda_{ij}^2 + \Psi_{ii}$. *The number* $h_i^2 := \sum_{j=1}^{k} \Lambda_{ij}^2$ *is the* communality *of variable i (variance attributable to the common factors) and* Ψ_{ii} *its* uniqueness *(idiosyncratic variance).*

Proof. By Lemma 1.7 expanded on the sum $\Lambda F + \varepsilon$:

$$\operatorname{Cov}(\Lambda F + \varepsilon) = \Lambda \operatorname{Cov}(F)\Lambda^\top + \Lambda \operatorname{Cov}(F, \varepsilon) + \operatorname{Cov}(\varepsilon, F)\Lambda^\top + \operatorname{Cov}(\varepsilon) = \Lambda\Lambda^\top + 0 + 0 + \Psi. \quad \square$$

Theorem 5.4 (existence of factor models)**.** *Let $\Sigma \in \mathbb{R}^{p\times p}$ be symmetric PSD and $\mu \in \mathbb{R}^p$. Then:*

(a) There exists a random vector X with mean μ and covariance Σ admitting a k-factor model ***if and only if*** *there exists a diagonal PSD matrix Ψ with*

$$\Sigma - \Psi \ \textit{PSD} \qquad \textit{and} \qquad \operatorname{rank}(\Sigma - \Psi) \le k.$$

(b) Consequently every Σ admits a p-factor model (take $\Psi = 0$), while for $k < p$ existence is a genuine restriction on Σ.

Proof. ($\Rightarrow$) If $X = \mu + \Lambda F + \varepsilon$ as in Definition 5.1 with $\Lambda \in \mathbb{R}^{p\times k}$, then $\Sigma - \Psi = \Lambda\Lambda^\top$ by Proposition 5.3. The matrix $\Lambda\Lambda^\top$ is PSD ($w^\top\Lambda\Lambda^\top w = \|\Lambda^\top w\|_2^2 \ge 0$) and has rank at most k ($\operatorname{rank}(\Lambda\Lambda^\top) \le \operatorname{rank}(\Lambda) \le k$).

($\Leftarrow$) Suppose such a Ψ exists. Set $A := \Sigma - \Psi$, a PSD matrix of rank $r \le k$, with spectral decomposition $A = \sum_{j=1}^{r} \nu_j\, u_j u_j^\top$, $\nu_1 \ge \cdots \ge \nu_r > 0$ (Theorem 2.1, Lemma 2.3). Define

$$\Lambda := \left[\ \sqrt{\nu_1}\, u_1 \ \ \cdots \ \ \sqrt{\nu_r}\, u_r \ \ \mathbf{0} \ \ \cdots \ \ \mathbf{0}\ \right] \in \mathbb{R}^{p\times k}, \qquad \text{so } \Lambda\Lambda^\top = \sum_{j=1}^{r} \nu_j\, u_j u_j^\top = A.$$

Now *construct* the random variables. On a suitable probability space take $F \sim \mathcal{N}(0, I_k)$ and $\varepsilon \sim \mathcal{N}(0, \Psi)$ *independent* (a product of Gaussian measures on $\mathbb{R}^k \times \mathbb{R}^p$ exists and supplies such a pair; any other pair of independent laws with these first two moments works equally well). Independence implies $\operatorname{Cov}(F, \varepsilon) = 0$: each entry is $\mathbb{E}[F_j\varepsilon_i] - \mathbb{E}[F_j]\mathbb{E}[\varepsilon_i] = \mathbb{E}[F_j]\mathbb{E}[\varepsilon_i] - 0 = 0$, using that independence factorises the expectation of the product. Define $X := \mu + \Lambda F + \varepsilon$. Then $\mathbb{E}[X] = \mu$, and by Proposition 5.3, $\operatorname{Cov}(X) = \Lambda\Lambda^\top + \Psi = A + \Psi = \Sigma$. All clauses of Definition 5.1 hold, so X admits a k-factor model.

(b) With $\Psi = 0$ and $k = p$: $\Sigma - 0 = \Sigma$ is PSD with $\operatorname{rank}(\Sigma) \le p$, so the condition in (a) holds. □

Remark 5.5. Theorem 5.4 answers: *does a k-factor distribution with prescribed (μ, Σ) exist?* That is the operative existence question, because in practice one only ever fits the covariance restriction $\Sigma = \Lambda\Lambda^\top + \Psi$. Whether a *given* joint law decomposes as in Definition 5.1 is a strictly stronger question, not addressed here.

Proposition 5.6 (rotational non-identifiability)**.** *If $(\Lambda, F, \varepsilon)$ realises a k-factor model for X, then so does $(\Lambda G,\ G^\top F,\ \varepsilon)$ for* every *orthogonal $G \in \mathbb{R}^{k\times k}$.*

Proof. $(\Lambda G)(G^\top F) = \Lambda(GG^\top)F = \Lambda F$, so the defining equation for X is unchanged. Moreover $\mathbb{E}[G^\top F] = 0$; $\operatorname{Cov}(G^\top F) = G^\top I_k G = I_k$ (Lemma 1.7); and $\operatorname{Cov}(G^\top F, \varepsilon) = G^\top \operatorname{Cov}(F, \varepsilon) = 0$. All clauses of Definition 5.1 hold for the new triple. □

Remark 5.7. So factor loadings are determined at best up to right multiplication by an orthogonal matrix. This is why factor-analysis software offers "rotations" (varimax and relatives): they are rules for selecting one representative from this orbit. PCA has no such freedom beyond signs and eigenvalue ties (Remark 3.3), because PCA is defined by an *ordered* optimisation, not by an unordered covariance decomposition.

Remark 5.8 (PCA versus factor model: the actual difference). Both frameworks write $\Sigma \approx$ (low rank)+(remainder). PCA takes the optimal low-rank part in the sense of Theorem 3.5 and places *no constraint* on the remainder; the factor model demands the remainder be *diagonal*. The truncation $\Sigma \approx \sum_{j \le k} \lambda_j q_j q_j^\top$ generally leaves a non-diagonal residual $\sum_{j>k} \lambda_j q_j q_j^\top$, so truncated PCA does not, in general, satisfy Definition 5.1, and eigenvectors are not factor loadings. The two constructions become close in special regimes — e.g. $\Psi = \sigma^2 I_p$, or asymptotics with $p \to \infty$ and pervasive factors, which is what licenses "PCA as factor estimation" in large panels — but as exact finite-dimensional statements they are different objects.

6 Factor models in finance: observed factors

Finance usage adds a twist: the factors are frequently *observed* time series, and "loading" silently becomes "regression coefficient".

Definition 6.1 (observed-factor model)**.** Let $r_t \in \mathbb{R}^p$ denote asset (excess) returns at time t and $f_t \in \mathbb{R}^k$ observed factor returns (e.g. market, value, momentum portfolio returns), with the pair (r_t, f_t) jointly stationary and satisfying Assumption 1.4 at each t; stationarity makes all second moments below time-invariant. The model asserts constants $\alpha \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p \times k}$ with

$$r_t = \alpha + B f_t + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t] = 0, \qquad \operatorname{Cov}(f_t, \varepsilon_t) = 0.$$

B is the *loading* or *beta* matrix; row i contains asset i's *betas*. Note what is *not* assumed: $\operatorname{Cov}(f_t)$ need not be I_k, and $\operatorname{Cov}(\varepsilon_t)$ need not be diagonal.

Proposition 6.2 (betas are population regression coefficients)**.** *Under Definition 6.1, with* $\Sigma_f := \operatorname{Cov}(f_t)$ *invertible,*

$$B = \operatorname{Cov}(r_t, f_t)\, \Sigma_f^{-1},$$

where $\operatorname{Cov}(r_t, f_t) \in \mathbb{R}^{p \times k}$ *is the cross-covariance matrix of Definition 1.6. In the one-factor case* $k = 1$*:* $\beta_i = \operatorname{Cov}(r_{i,t}, f_t) / \operatorname{Var}(f_t)$ *— the CAPM beta formula.*

Proof. $\operatorname{Cov}(r_t, f_t) = \operatorname{Cov}(B f_t + \varepsilon_t,\ f_t) = B\,\Sigma_f + \operatorname{Cov}(\varepsilon_t, f_t) = B\Sigma_f$ by Lemma 1.7 (the constant α drops out); right-multiply by Σ_f^{-1}. □

Remark 6.3. Contrast with Proposition 5.2: there, $\operatorname{Cov}(F) = I_k$ was imposed, so loading = covariance-with-factor; here Σ_f is whatever the market delivers, so loading = covariance-with-factor *times* Σ_f^{-1} — an ordinary-least-squares–type coefficient. Same word, two formulas, reconciled exactly when the factors are standardised to identity covariance.

Proposition 6.4 (risk decomposition)**.** *Under Definition 6.1, with* $\Sigma_\varepsilon := \operatorname{Cov}(\varepsilon_t)$*,*

$$\operatorname{Cov}(r_t) = B\,\Sigma_f\,B^\top + \Sigma_\varepsilon.$$

The first term is the systematic *(factor) covariance, the second the* idiosyncratic *covariance. For a portfolio with weight vector* $w \in \mathbb{R}^p$*,*

$$\operatorname{Var}(w^\top r_t) = (B^\top w)^\top\, \Sigma_f\, (B^\top w) \;+\; w^\top \Sigma_\varepsilon w,$$

and the vector $B^\top w \in \mathbb{R}^k$ *is the portfolio's* factor exposure.

Proof. Identical to the proof of Proposition 5.3, expanding $\mathrm{Cov}(Bf_t + \varepsilon_t)$ by Lemma 1.7 and using $\mathrm{Cov}(f_t, \varepsilon_t) = 0$; then apply Lemma 1.7 once more with $A = w^\top$. □

Definition 6.5 (strict vs approximate)**.** The observed-factor model is called *strict* if additionally Σ_ε is diagonal — which, after standardising the factors as in Remark 6.3, reproduces Definition 5.1. The *approximate factor model* of the econometrics literature relaxes diagonality to a bound on the eigenvalues of Σ_ε as $p \to \infty$. Latent (statistical) factor models in finance are exactly Section 5 applied to returns, usually *estimated* by PCA on the sample covariance of returns — a procedure justified by the asymptotic regimes of Remark 5.8, not by any finite-p identity between PCA and the factor model.

7 Cross-sectional (fundamental) factor models

Commercial equity risk models of the Barra/Axioma "fundamental" type invert the information structure of Section 6: there, the factors f_t were observed and the loadings B were unknown; here the *loadings are observed data* and the *factor returns are estimated*, by one regression per period run *across assets*. The word *cross-sectional* means exactly this: at a fixed time t, the statistical sample is the collection of assets $i = 1, \dots, p$, not the collection of time periods. To stay inside the toolkit of Section 1, the analysis below fixes a single period and treats it in the population; the multi-period assembly is described at the end (Remark 7.16).

Definition 7.1 (characteristic; exposure; standardisation)**.** A *characteristic* of asset i is a number computed from observable data about that asset — for example the logarithm of its market capitalisation (*size*), its book value divided by its market value (*value*), its trailing return (*momentum*), or an *industry dummy*: the indicator equal to 1 if the asset belongs to a given industry and 0 otherwise. Continuous characteristics are *standardised* across the asset universe before use: given raw values $c_1, \dots, c_p$,

(i) *winsorisation at level* α: each c_i is replaced by $\min\{\max\{c_i, q_\alpha\}, q_{1-\alpha}\}$, where q_α and $q_{1-\alpha}$ are the empirical α- and $(1-\alpha)$-quantiles of $c_1, \dots, c_p$ — values beyond the quantiles are clamped to them;

(ii) *z-scoring*: the winsorised values $\tilde{c}_i$ are replaced by $z_i := (\tilde{c}_i - m)/s$, where m is a cross-sectional mean of the $\tilde{c}_i$ (equal-weighted or capitalisation-weighted — a modelling choice) and s is their cross-sectional standard deviation, assumed nonzero.

The resulting number z_i is asset i's *exposure* to the characteristic. The *exposure matrix* is $B \in \mathbb{R}^{p \times k}$, row i holding asset i's exposures to all k factors.

Assumption 7.2 (cross-sectional model, single period)**.** The following are the model's assumptions, stated in full:

(C1) *Known exposures.* $B \in \mathbb{R}^{p \times k}$ is a constant (non-random) matrix, computed from data observable *before* the period's returns realise.

(C2) *Return structure.* The return vector $r \in \mathbb{R}^p$ is a random vector (Assumption 1.4) satisfying $r = Bf + \varepsilon$, where $f \in \mathbb{R}^k$ (the *factor returns*) and $\varepsilon \in \mathbb{R}^p$ (the *specific returns*) are random vectors with $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(f, \varepsilon) = 0$, and $\Sigma_\varepsilon := \mathrm{Cov}(\varepsilon)$ diagonal.

(C3) *Full column rank.* $\mathrm{rank}(B) = k$; that is, the columns of B are linearly independent. (A matrix has *full column rank* when its rank equals its number of columns; equivalently $Bg = 0 \Rightarrow g = 0$.) When (C3) fails — and for realistic B it does fail; see Remark 7.15 — a constraint is added.

(C4) *Regression weights.* A matrix $\Omega \in \mathbb{R}^{p \times p}$ is chosen, diagonal with strictly positive diagonal entries; hence *positive definite* (PD): $v^\top \Omega v > 0$ for all $v \neq 0$. (A PSD matrix, Lemma 1.8, permits $= 0$; PD forbids it.) Common choices: $\Omega_{ii} = \sqrt{\text{market cap}_i}$, or $\Omega_{ii} = 1/\hat{\sigma}^2_{\varepsilon,i}$. Ω is a modelling choice, not an estimated quantity, in what follows.

Neither f nor ε is observed; only r realises, and B, Ω are known. Note what is *not* assumed: $\mathrm{Cov}(f) = I_k$ (contrast Definition 5.1).

Definition 7.3 (weighted least squares)**.** The *weighted least squares (WLS) estimator* of f is the minimiser, over $g \in \mathbb{R}^k$, of the criterion

$$J(g) := (r - Bg)^\top \Omega\, (r - Bg),$$

i.e. the weighted sum of squared residuals, weight Ω_{ii} on asset i. "Estimator" here means: a function of the observable r intended to approximate the unobservable f.

Proposition 7.4 (WLS: existence, uniqueness, closed form)**.** *Under Assumption 7.2 (C3)–(C4), the matrix $B^\top \Omega B \in \mathbb{R}^{k\times k}$ is invertible, and J has the unique minimiser*

$$\hat{f} := (B^\top \Omega B)^{-1} B^\top \Omega\, r.$$

Proof. Invertibility. $B^\top \Omega B$ is square, so it suffices to show its null space is trivial. If $B^\top \Omega B g = 0$ then $0 = g^\top B^\top \Omega B g = (Bg)^\top \Omega (Bg)$, which by positive definiteness (C4) forces $Bg = 0$, and then $g = 0$ by full column rank (C3).

Optimality. Let $\hat{f}$ be as displayed and $g \in \mathbb{R}^k$ arbitrary; write $g = \hat{f} + h$. Expanding J and using $B^\top \Omega (r - B\hat{f}) = B^\top \Omega r - (B^\top \Omega B)(B^\top \Omega B)^{-1} B^\top \Omega r = 0$ (the *normal equations*):

$$\begin{aligned} J(g) &= \big(r - B\hat{f} - Bh\big)^\top \Omega \big(r - B\hat{f} - Bh\big) \\ &= J(\hat{f}) - 2\, h^\top \underbrace{B^\top \Omega (r - B\hat{f})}_{=\,0} + h^\top B^\top \Omega B\, h \;=\; J(\hat{f}) + (Bh)^\top \Omega (Bh). \end{aligned}$$

The added term is ≥ 0, and $= 0$ only if $Bh = 0$, i.e. (by C3) $h = 0$. So $\hat{f}$ is the unique minimiser. □

Definition 7.5 (portfolio; pure factor portfolio)**.** A *portfolio* is a vector $\pi \in \mathbb{R}^p$ of *weights*; its return is the scalar random variable $\pi^\top r$, and its *exposure vector* is $B^\top \pi \in \mathbb{R}^k$ (as in Proposition 6.4). Define

$$\Pi := (B^\top \Omega B)^{-1} B^\top \Omega \;\in\; \mathbb{R}^{k\times p},$$

so that $\hat{f} = \Pi r$, and write $\pi_j^\top$ for the j-th row of Π. A portfolio is a *pure factor portfolio* for factor j if its exposure vector is e_j: exposure 1 to factor j and 0 to every other factor.

Proposition 7.6 (each estimated factor return is the return of a pure factor portfolio)**.** *$\Pi B = I_k$. Equivalently, for each j: $B^\top \pi_j = e_j$, so π_j is a pure factor portfolio for factor j, and the j-th estimated factor return $\hat{f}_j = \pi_j^\top r$ is the realised return of that portfolio. Moreover $\hat{f} = f + \Pi\varepsilon$: the estimator equals the true factor return plus a weighted aggregate of specific returns.*

Proof. $\Pi B = (B^\top \Omega B)^{-1}(B^\top \Omega B) = I_k$. Row j of this identity reads $\pi_j^\top B = e_j^\top$, i.e. $B^\top \pi_j = e_j$. For the last claim: $\hat{f} = \Pi r = \Pi(Bf + \varepsilon) = (\Pi B) f + \Pi\varepsilon = f + \Pi\varepsilon$. □

Definition 7.7 (industry partition)**.** An *industry partition* of the asset universe $\{1, \dots, p\}$ into m industries is a family of sets $I_1, \dots, I_m \subseteq \{1, \dots, p\}$, pairwise disjoint, each nonempty, with union $\{1, \dots, p\}$: every asset belongs to exactly one industry. Write $\mathrm{ind}(i)$ for the unique j with $i \in I_j$. The associated *dummy matrix* is $D \in \{0,1\}^{p\times m}$ with $D_{ij} = 1$ if $i \in I_j$ and $D_{ij} = 0$ otherwise; each row of D contains exactly one 1.

Example 7.8 (a four-asset running example)**.** Let $p = 4$, with assets labelled 1 (a large technology company), 2 (a small technology company), 3 (a large bank), 4 (a small bank); industries $I_{\mathrm{T}} = \{1, 2\}$ (technology) and $I_{\mathrm{B}} = \{3, 4\}$ (banking); one style, size, with exposures already standardised per Definition 7.1. Take $k = 3$ and

$$B \;=\; \begin{pmatrix} 1 & 0 & +1.2 \\ 1 & 0 & -0.3 \\ 0 & 1 & +0.1 \\ 0 & 1 & -1.0 \end{pmatrix},$$

columns ordered (technology dummy, banking dummy, size). Row i is asset i's exposure vector. Under Assumption 7.2(C2), asset 1's return satisfies $r_1 = f_{\mathrm{T}} + 1.2\, f_{\mathrm{S}} + \varepsilon_1$, where $f_{\mathrm{T}}, f_{\mathrm{B}}, f_{\mathrm{S}}$ denote the three coordinates of f: the dummy column contributes the *same* term f_{T} to every technology asset (its coefficient is 1 for each of them) and nothing to the others. This B has full column rank, so (C3) holds; the collinearity of Remark 7.15 would appear only if a market column $\mathbf{1}_4$ were added alongside both dummies.

Proposition 7.9 (dummies-only WLS estimates weighted industry means)**.** *Suppose the exposure matrix consists of the industry dummies alone: $B = D$ for an industry partition as in Definition 7.7, so $k = m$. Let Ω satisfy (C4). Then (C3) holds automatically, and the WLS estimator of Proposition 7.4 is, coordinate by coordinate,*

$$\hat{f}_j \;=\; \frac{\sum_{i \in I_j} \Omega_{ii}\, r_i}{\sum_{i \in I_j} \Omega_{ii}}, \qquad j = 1, \dots, m,$$

the Ω-weighted mean of the returns in industry j. In particular, for $\Omega = I_p$, $\hat{f}_j = \frac{1}{|I_j|} \sum_{i \in I_j} r_i$: the arithmetic mean of industry j's returns. In Example 7.8 with the size column deleted and $\Omega = I_4$: $\hat{f}_{\mathrm{T}} = \frac{1}{2}(r_1 + r_2)$ and $\hat{f}_{\mathrm{B}} = \frac{1}{2}(r_3 + r_4)$.

Proof. Compute the $k \times k$ matrix $D^{\top}\Omega D$ entrywise. Since Ω is diagonal,

$$(D^{\top}\Omega D)_{jl} = \sum_{i=1}^{p} D_{ij}\, \Omega_{ii}\, D_{il}.$$

Each row of D has exactly one nonzero entry, so $D_{ij} D_{il} = 0$ whenever $j \neq l$; hence the off-diagonal entries vanish, and $(D^{\top}\Omega D)_{jj} = \sum_{i \in I_j} \Omega_{ii} > 0$, positive because $I_j \neq \emptyset$ and $\Omega_{ii} > 0$ by (C4). Thus $D^{\top}\Omega D = \operatorname{diag}\big(\sum_{i \in I_1} \Omega_{ii}, \dots\big)$ is invertible; invertibility of $B^{\top}\Omega B$ forces $\operatorname{rank}(B) = k$ (if $Bg = 0$ then $B^{\top}\Omega B g = 0$, so $g = 0$), which is (C3). Similarly $(D^{\top}\Omega r)_j = \sum_{i \in I_j} \Omega_{ii} r_i$. Proposition 7.4 then gives $\hat{f}_j = \big(\sum_{i \in I_j} \Omega_{ii}\big)^{-1} \sum_{i \in I_j} \Omega_{ii} r_i$, as claimed; the $\Omega = I_p$ and Example 7.8 statements are the special cases $\Omega_{ii} \equiv 1$. □

Proposition 7.10 (raw industry averages are contaminated portfolios)**.** *Return to a general exposure matrix B containing an industry partition's dummy columns together with style columns, as in Example 7.8, and let $\pi \in \mathbb{R}^p$ be any portfolio* supported on *industry I_j — meaning $\pi_i = 0$ for all $i \notin I_j$ — with weights summing to one, $\mathbf{1}_p^{\top}\pi = 1$. Then the exposure vector $B^{\top}\pi$ of Definition 7.5 has: entry 1 in the coordinate of dummy j; entry 0 in the coordinate of every other dummy; and, in the coordinate of each style s, the π-weighted mean $\sum_{i \in I_j} \pi_i B_{is}$ of that style's exposures over the industry. Consequently π is a pure factor portfolio for factor j (Definition 7.5)* if and only if *every style's π-weighted industry mean vanishes; generically it does not, and the raw industry-average return $\pi^{\top} r$ is* not *$\hat{f}_j$. In Example 7.8, the equal-weighted technology average $\pi = (\frac{1}{2}, \frac{1}{2}, 0, 0)^{\top}$ has exposure vector $B^{\top}\pi = (1,\, 0,\, +0.45)^{\top}$: it carries a size exposure of $+0.45$, so it earns $f_{\mathrm{T}} + 0.45\, f_{\mathrm{S}}$ plus averaged specific returns, not f_{T} alone. The pure factor portfolio π_j of Proposition 7.6 is exactly the modification of the industry average that shorts enough size tilt to cancel this term.*

Proof. Coordinate l of $B^\top \pi$ is $\sum_{i=1}^{p} B_{il}\,\pi_i = \sum_{i \in I_j} B_{il}\,\pi_i$, using the support condition. If column l is dummy j, then $B_{il} = 1$ for all $i \in I_j$, giving $\sum_{i \in I_j} \pi_i = \mathbf{1}_p^\top \pi = 1$. If column l is a different dummy $j' \neq j$, then $B_{il} = 0$ for all $i \in I_j$ (disjointness of the partition), giving 0. If column l is a style s, the sum is the stated weighted mean. The pure-factor characterisation is then immediate from Definition 7.5, and the numerical claim is $\frac{1}{2}(1.2) + \frac{1}{2}(-0.3) = 0.45$. □

Lemma 7.11 (within-model subtraction identity)**.** *Under Assumption 7.2(C2), for every asset i,*

$$r_i \;-\; \sum_{l\,:\,l \neq \mathrm{ind}(i)\text{-}dummy} B_{il}\, f_l \;=\; f_{\mathrm{ind}(i)} + \varepsilon_i,$$

where the sum runs over all factor coordinates other than asset i's own industry dummy. That is: after subtracting all other factor contributions, every asset in a given industry sits at the same *level $f_{\mathrm{ind}(i)}$, displaced by its own ε_i; and under (C2) the displacements are mean-zero and mutually uncorrelated across assets.*

Proof. Row i of $r = Bf + \varepsilon$ reads $r_i = \sum_{l=1}^{k} B_{il} f_l + \varepsilon_i$; the coefficient on asset i's own industry dummy is 1 and on every other dummy is 0 (Definition 7.7). Move all terms except $B_{i,\mathrm{ind}(i)} f_{\mathrm{ind}(i)} = f_{\mathrm{ind}(i)}$ to the left. The properties of the ε_i restate (C2). □

Proposition 7.12 (omitted-variable formula)**.** *Partition the factors into two blocks: $B = [\,B_1 \mid B_2\,]$ with $B_1 \in \mathbb{R}^{p \times k_1}$, $B_2 \in \mathbb{R}^{p \times k_2}$, and correspondingly $f = (f_1^\top, f_2^\top)^\top$, so that (C2) reads $r = B_1 f_1 + B_2 f_2 + \varepsilon$. Suppose (C3)–(C4) hold for the full B. Define the* misspecified estimator *as the WLS estimator computed* as if *B_1 were the whole exposure matrix: $\tilde{f}_1 := (B_1^\top \Omega B_1)^{-1} B_1^\top \Omega\, r$ (well defined, since full column rank of B implies full column rank of B_1, and the invertibility argument of Proposition 7.4 applies). Then*

$$\tilde{f}_1 \;=\; f_1 \;+\; \Gamma f_2 \;+\; (B_1^\top \Omega B_1)^{-1} B_1^\top \Omega\, \varepsilon, \qquad \Gamma := (B_1^\top \Omega B_1)^{-1} B_1^\top \Omega B_2 \in \mathbb{R}^{k_1 \times k_2}.$$

The systematic contamination Γf_2 — called omitted-variable bias *— vanishes for all realisations of f_2 if and only if $B_1^\top \Omega B_2 = 0$, i.e. the retained and omitted exposure columns are orthogonal in the inner product $\langle u, v \rangle_\Omega := u^\top \Omega v$. In Example 7.8, estimating the two industry factors while omitting the size column ($B_1 = D$, $B_2 =$ size column, $\Omega = I_4$) gives, by Proposition 7.9, $\Gamma = \binom{+0.45}{-0.45}$: the misspecified technology factor is $f_{\mathrm{T}} + 0.45 f_{\mathrm{S}}$ plus noise — exactly the contaminated industry average of Proposition 7.10.*

Proof. Substitute the full model into the misspecified formula and use linearity:

$$\tilde{f}_1 = (B_1^\top \Omega B_1)^{-1} B_1^\top \Omega\,(B_1 f_1 + B_2 f_2 + \varepsilon) = f_1 + \Gamma f_2 + (B_1^\top \Omega B_1)^{-1} B_1^\top \Omega\, \varepsilon,$$

since $(B_1^\top \Omega B_1)^{-1} B_1^\top \Omega B_1 = I_{k_1}$. The term Γf_2 is zero for every value of f_2 iff $\Gamma = 0$ iff $B_1^\top \Omega B_2 = 0$, the invertible prefactor being harmless. For the example: $\Gamma = (D^\top D)^{-1} D^\top B_2 = \mathrm{diag}(\frac{1}{2}, \frac{1}{2}) \binom{1.2-0.3}{0.1-1.0} = \binom{+0.45}{-0.45}$, reading off $D^\top B_2$ as the industry sums of size exposures. □

Remark 7.13 (joint, not sequential, estimation). Propositions 7.9–7.12 settle a natural question: are the style factors estimated first, or the industry factors? Neither. All k coordinates of $\hat{f}$ are produced by the single joint regression of Proposition 7.4. A sequential scheme — fit one block of factors against raw returns, subtract, fit the remaining block on residuals — makes its first stage exactly the misspecified estimator of Proposition 7.12, contaminated by Γf_2 unless the blocks happen to be Ω-orthogonal. By contrast, "removing" other factors by *subtraction inside the joint model* (Lemma 7.11) is exact. The joint fit is also what delivers Proposition 7.6 in full strength: $\Pi B = I_k$ makes each pure factor portfolio simultaneously neutral to all $k - 1$ other factors — the industry average, corrected for its style tilts (Proposition 7.10), in

one step. (Deliberately hierarchical designs — e.g. a market or world factor estimated first and finer factors fitted to residuals — do exist in commercial and academic practice; they accept the structure described by Proposition 7.12 knowingly, and are a different model, not a different algorithm for this one.)

Proposition 7.14 (the regression identity of Proposition 6.2 fails here)**.** *Under Assumption 7.2, with* $\Sigma_f := \mathrm{Cov}(f)$*, suppose* $\mathrm{Cov}(\hat{f})$ *is invertible. Then*

$$\mathrm{Cov}(r, \hat{f})\ \mathrm{Cov}(\hat{f})^{-1} \;=\; \big(B\Sigma_f + \Sigma_\varepsilon \Pi^\top\big)\big(\Sigma_f + \Pi\Sigma_\varepsilon\Pi^\top\big)^{-1},$$

and this equals B *if and only if* $(I_p - B\Pi)\,\Sigma_\varepsilon\Pi^\top = 0$*. The condition holds in the special case* $\Omega = I_p$*,* $\Sigma_\varepsilon = \sigma^2 I_p$*, but fails for generic diagonal* Σ_ε*. Hence, in the cross-sectional model,* B *is exogenous data and is* not *recoverable as the population regression coefficient* $\mathrm{Cov}(r, \hat{f})\,\mathrm{Cov}(\hat{f})^{-1}$*: the estimated factors* $\hat{f}$ *are themselves contaminated by* ε *(Proposition 7.6), which correlates the "regressor" with the "error".*

Proof. By Proposition 7.6, $\hat{f} = f + \Pi\varepsilon$. Lemma 1.7 and $\mathrm{Cov}(f, \varepsilon) = 0$ give

$$\mathrm{Cov}(r, \hat{f}) = \mathrm{Cov}(Bf + \varepsilon,\ f + \Pi\varepsilon) = B\Sigma_f + \Sigma_\varepsilon\Pi^\top, \qquad \mathrm{Cov}(\hat{f}) = \Sigma_f + \Pi\Sigma_\varepsilon\Pi^\top.$$

Then $\mathrm{Cov}(r, \hat{f})\,\mathrm{Cov}(\hat{f})^{-1} = B$ holds iff $B\Sigma_f + \Sigma_\varepsilon\Pi^\top = B\Sigma_f + B\Pi\Sigma_\varepsilon\Pi^\top$, i.e. iff $(I_p - B\Pi)\Sigma_\varepsilon\Pi^\top = 0$. Special case: with $\Omega = I_p$, $\Pi^\top = B(B^\top B)^{-1}$ and $B\Pi = B(B^\top B)^{-1}B^\top$, which is the orthogonal projection onto the column space of B (it is symmetric and idempotent by the arguments of Section 3); with $\Sigma_\varepsilon = \sigma^2 I_p$ the condition becomes $(I - B\Pi)\,\sigma^2 B(B^\top B)^{-1} = 0$, which holds because $(I - B\Pi)B = B - B = 0$. For general diagonal Σ_ε, the condition requires every column of $\Sigma_\varepsilon\Pi^\top$ to lie in the column space of B (the null space of $I - B\Pi$ is exactly that column space, since $B\Pi$ acts as the identity on it and $\mathrm{rank}(B\Pi) = k$), which is a nongeneric coincidence. □

Remark 7.15 (collinearity of the exposure matrix, and constraints). Realistic exposure matrices violate (C3) exactly. Suppose $B = [\,\mathbf{1}_p \mid D \mid S\,]$ where $\mathbf{1}_p$ is the *market* column (every asset has exposure 1), $D \in \{0,1\}^{p\times m}$ holds industry dummies with each asset in exactly one industry, and S holds style exposures. "Each asset in exactly one industry" means each row of D sums to 1: $D\mathbf{1}_m = \mathbf{1}_p$. Then

$$B\begin{pmatrix} 1 \\ -\mathbf{1}_m \\ 0 \end{pmatrix} = \mathbf{1}_p - D\mathbf{1}_m = 0,$$

so the columns of B are linearly dependent, $\mathrm{rank}(B) \le k - 1$, and by the failure of (C3) the WLS minimiser is not unique (adding any multiple of the displayed null vector to g leaves Bg, hence $J(g)$, unchanged). The remedy is a *linear constraint*: minimise J over $\{g : c^\top g = 0\}$ for a chosen $c \in \mathbb{R}^k$ (e.g. capitalisation-weighted industry factor returns sum to zero). If c is chosen with $c^\top v \neq 0$ for the null vector v above — i.e. the constraint hyperplane is transverse to the null direction — then B restricted to the constraint set is injective, and the constrained problem inherits existence and uniqueness from Proposition 7.4 by parametrising the constraint set with a basis of $\{c\}^\perp$ and applying the proposition to the reparametrised full-column-rank design. The estimator remains linear in r, so Proposition 7.6's portfolio interpretation survives with e_j replaced by the appropriate constrained target.

Remark 7.16 (multi-period assembly; the full assumption tower). The production risk model repeats the single-period construction each period t: exposures B_{t-1} are recomputed and re-standardised (Definition 7.1); $\hat{f}_t = \Pi_{t-1} r_t$ is estimated (Proposition 7.4); specific variances

$\hat{\sigma}^2_{\varepsilon,i}$ are estimated from the residuals $r_t - B_{t-1}\hat{f}_t$. The factor covariance $\hat{\Sigma}_f$ is then estimated from the *time series* $\{\hat{f}_t\}$ — typically with exponential downweighting of old observations and adjustments for serial correlation — and the delivered asset covariance is

$$\hat{\Sigma}_r \;=\; B\,\hat{\Sigma}_f\,B^\top + \hat{\Delta}, \qquad \hat{\Delta} := \operatorname{diag}(\hat{\sigma}^2_{\varepsilon,1}, \ldots, \hat{\sigma}^2_{\varepsilon,p}),$$

the sample analogue of Proposition 6.4. Each layer adds assumptions beyond Assumption 7.2: stationarity across periods for the time-series estimation of $\hat{\Sigma}_f$; the identification of $\hat{f}_t$ (which by Proposition 7.6 carries an $\Pi\varepsilon_t$ contamination) with f_t; and the diagonality of $\hat{\Delta}$, inherited from (C2). This document proves the single-period algebra; the time-series estimation theory is beyond its scope and is flagged, not proved.

8 What is imposed, what is estimated, what is identified

Every quantity in this document belongs to exactly one of three classes, and factor-model practice systematically blurs them. This section states the classification and then itemises, with references, every imposition in the cross-sectional model of Section 7 — the class of model sold commercially (Barra/Axioma type). The purpose is that a reader of any vendor's documentation can locate each of its choices in this list.

Definition 8.1 (imposed; estimated; identified)**.** Relative to a model, a quantity is:

(i) *imposed* if it is fixed by the modeller and is not the solution of any optimisation or estimation criterion stated by the model — changing it changes the model, and no realisation of the data can contradict the choice from within the model;

(ii) *estimated* if it is a stated function of realised data, computed conditionally on the impositions;

(iii) *identified* if it is uniquely determined by the joint distribution of the observables, given the impositions — i.e. two parameter values producing the same distribution of observables must be equal.

An estimated quantity can target an unidentified one; Proposition 5.6 below is the standing example.

Remark 8.2 (the impositions of the cross-sectional model, itemised). Each item states the choice, where it enters the formalism, and what varies with it. None is derived from data; each is fixed before the period's returns realise.

(I1) **Universe and frequency.** Which assets constitute $\{1, \ldots, p\}$, and the length of a period. Every subsequent object — the cross-sectional quantiles and moments of Definition 7.1, hence B, hence $\hat{f}$ — is relative to this choice.

(I2) **Factor identity.** The list of k characteristics forming the columns of B (Definition 7.1, Assumption 7.2(C1)). No criterion inside the model selects this list; two vendors with different lists produce different $\hat{f}_t$, different $\hat{\Sigma}_f$, and different risk numbers for the same portfolio, and no theorem in this document adjudicates between them.

(I3) **Functional form.** Linearity of returns in exposures, and exposures entering as the current period's standardised characteristics — the equation of (C2) itself. Nonlinear response to a characteristic, or dependence on its history rather than its current z-score, is excluded by fiat, not by test.

(I4) **The standardisation recipe.** The winsorisation level α, the choice of equal- versus capitalisation-weighted cross-sectional mean, and per-period re-scoring (Definition 7.1(i)–(ii)). Since $\hat{f}$ is, for fixed r, a function of B (Proposition 7.4), every one of these choices propagates into the estimated factor returns.

(I5) **Diagonality of Σ_ε** (Assumption 7.2(C2)). This is the substantive imposition: it *declares* that all cross-asset covariance flows through the chosen factors — assets covary through shared characteristics, plus nothing else. It is routinely false at the level of asset pairs (any economic linkage not represented by a column of B violates it) and is imposed nonetheless; the approximate-factor relaxation of Definition 6.5 is the admission, made rigorous asymptotically, that it holds only "on average" as $p \to \infty$.

(I6) **The weight matrix Ω** (Assumption 7.2(C4)). Chosen, not estimated, within the single-period model. It changes the estimator (Proposition 7.4), the pure factor portfolios Π (Definition 7.5), and hence which weighted industry mean the dummies-only estimator computes (Proposition 7.9). (Variants that set $\Omega_{ii} = 1/\hat{\sigma}^2_{\varepsilon,i}$ estimate Ω from *prior* periods; the functional form of the weighting remains imposed.)

(I7) **The constraint** (Remark 7.15). The vector c resolving the market/industry collinearity is a convention, and different conventions reallocate the same fitted systematic return between the market factor and the industry factors: the fitted values $B\hat{f}$ are invariant, the coordinates of $\hat{f}$ are not. A cross-vendor comparison of "the market factor return" compares numbers defined under different conventions.

(I8) **Rotation pinning, in the latent variant.** When the factors are latent (Section 5, used by statistical model variants), Proposition 5.6 shows the data identify at most the pair $(\Lambda\Lambda^\top, \Psi)$ — the split of covariance into common and specific — never the coordinates: (Λ, F) and $(\Lambda G, G^\top F)$ are indistinguishable for every orthogonal G. Any individuated factor ("the second statistical factor") is a representative chosen by an imposed rule (an ordering, a rotation criterion), in the sense of Definition 8.1(iii): estimated, but targeting an unidentified object. The fundamental model avoids this only by imposition (I2): fixing B exogenously pins the coordinates by fiat.

(I9) **Covariance assembly.** The exponential halflife, serial-correlation adjustments, and any shrinkage used to turn $\{\hat{f}_t\}$ into $\hat{\Sigma}_f$, and residuals into $\hat{\Delta}$ (Remark 7.16). These live outside the single-period model entirely and are flagged there as assumed, not proved.

Remark 8.3 (consequences for reading practitioner language). Two corollaries of Remark 8.2, stated bluntly. First: a sentence of the form "factor j returned x this period" is, by Proposition 7.6, the statement that one specific portfolio — π_j, whose composition depends on (I1)–(I4), (I6) and (I7), since Π is a function of B and Ω and of the constraint alone — realised the return x. It is a true statement about a constructed portfolio, not the observation of a feature of nature; and the same-named factor from a different vendor is a *different* portfolio, so cross-vendor agreement or disagreement of factor returns carries no evidential weight about anything except the overlap of the two vendors' choices. Second: what the data do supply — and it is the sole empirical motivation standing under the whole construction — is that equity return covariance exhibits low-dimensional common structure: a few sample eigenvalues of the return correlation matrix stand far above the remainder, reproducibly. That fact motivates *some* decomposition of the form $\Sigma = (\text{low rank}) + (\text{remainder})$; it selects *none* of (I1)–(I9). The gap between "a low-dimensional structure exists" and "the factors are these ones" is filled entirely by imposition, and every use of a factor model inherits the full list above. This document proves what follows from the impositions; it cannot, and no document can, prove the impositions.

9 The terminology decoder

Word	PCA meaning	Factor-analysis meaning	Finance meaning
loading	$(w_m)_i$ *or* $(w_m)_i\sqrt{\lambda_m}$ — source-dependent (Def. 3.6)	$\Lambda_{ij} = \mathrm{Cov}(X_i, F_j)$ (Prop. 5.2)	B_{ij}, a regression coefficient: $B = \mathrm{Cov}(r, f)\Sigma_f^{-1}$ from Prop. 6.2
component / factor	$Z_m = w_m^\top X$: a *constructed* random variable	F_j: a *postulated latent* random variable	$f_{j,t}$: an *observed* series (or latent, if statistical)
variance explained	$\lambda_m / \operatorname{tr}\Sigma$ (Def. 3.4)	communality h_i^2, per variable (Prop. 5.3)	systematic share $\frac{(B\Sigma_f B^\top)_{ii}}{\Sigma_{ii}}$: the regression R^2 of asset i
identification	unique up to signs, eigenvalue ties aside (Rem. 3.3)	up to orthogonal rotation (Prop. 5.6)	fully identified given observed f_t and invertible Σ_f

Section 7 adds a *fourth* meaning of "loading", used by fundamental risk models (Barra/Axioma type): an *observed standardised characteristic* B_{ij} (Definition 7.1, Assumption 7.2(C1)) — data, not an estimated or derived quantity, and not recoverable as a regression coefficient (Proposition 7.14).

Reading protocol for any source in this area: (i) determine whether the setting is population or sample (Sections 1 vs 4); (ii) determine which matrix is being diagonalised, Σ or R (Warning 3.8); (iii) determine which of the four column-meanings above the word "loading" carries; and (iv) locate the source's choices in the imposition list of Remark 8.2: whatever is not so located is being presented as data-derived when it is not. Almost every hand-wavy statement in the applied literature is a true statement about exactly one cell of this table, quoted outside it.

10 Exercises

Difficulty is marked E (easy: tests the definitions and theorem statements), M (moderate: applies the machinery to concrete problems), H (hard: requires constructing proofs that extend the text).

1. **(E)** Let Σ be a covariance matrix with $\Sigma_{ii} > 0$ for all i, and $R = D^{-1/2}\Sigma D^{-1/2}$ its correlation matrix (Definition 1.9). Prove directly from the definitions: (a) $R_{ii} = 1$ for every i; (b) $|R_{ij}| \le 1$ for all i, j (use the Cauchy–Schwarz inequality on L^2, as in Lemma 1.5); (c) R is PSD.

2. **(E)** Let $\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$. Compute both eigenvalues and an orthonormal pair of eigenvectors; write down the principal directions, the variance explained by each component (Definition 3.4), and the loading matrix under *both* conventions of Definition 3.6. Verify Proposition 3.7 numerically for variable 1 and component 1.

3. **(E)** Exhibit a covariance matrix $\Sigma \in \mathbb{R}^{3\times 3}$ for which the first principal direction is not unique up to sign, and describe the *complete* set of valid first principal directions. Which part of Theorem 3.2(b) does your description instantiate? What is the smallest number of distinct eigenvalues a 3×3 covariance matrix can have while still possessing a first principal direction that *is* unique up to sign?

4. **(M)** Let $\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 2 & 5 \end{pmatrix}$ ($n = 3$ observations, $p = 2$ variables). Compute $\bar{x}$, $\mathbf{X}_c$, and S (Definition 4.1); find the sample principal directions and sample eigenvalues; and compute the 3×2 score matrix. Verify Lemma 4.2(c) on your numbers: what is $\operatorname{rank}(S)$, and what is the bound $\min(n-1, p)$?

5. **(M)** Let $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Using Theorem 5.4(a), determine *all* diagonal PSD matrices Ψ such that Σ is the covariance of a 1-factor model, and for one such Ψ exhibit an explicit loading matrix $\Lambda \in \mathbb{R}^{2\times 1}$. Then show that for $p = 2$ the 1-factor decomposition is never unique: the set of valid Ψ contains a continuum.

6. **(M)** In Example 7.8, delete the size column, keep the two industry dummies, and take $\Omega = \operatorname{diag}(4, 1, 4, 1)$. Compute $\hat{f}_{\mathrm{T}}$ and $\hat{f}_{\mathrm{B}}$ as explicit weighted means (Proposition 7.9), then compute the full matrix Π (Definition 7.5) and verify $\Pi B = I_2$ by hand. Which asset dominates each pure factor portfolio, and why?

7. **(M)** In Example 7.8 (all three columns, so $B \in \mathbb{R}^{4\times 3}$), suppose $\Sigma_f = \operatorname{diag}(0.04,\ 0.04,\ 0.01)$ and $\Sigma_\varepsilon = \operatorname{diag}(0.09,\ 0.16,\ 0.09,\ 0.16)$, in per-period variance units. For the long–short portfolio $w = (1, -1, 0, 0)^\top$: compute the exposure vector $B^\top w$, the systematic variance $w^\top B \Sigma_f B^\top w$, and the specific variance $w^\top \Sigma_\varepsilon w$ (Proposition 6.4). Which factor dominates the systematic term, and what feature of w explains why the industry factors contribute nothing?

8. **(H)** Suppose $\lambda_k > \lambda_{k+1}$ in the spectral decomposition of Σ. Prove that although the individual directions $w_1, \ldots, w_k$ need not be unique (Remark 3.3), the *subspace* $\operatorname{span}(w_1, \ldots, w_k)$ is the same for every solution sequence of Definition 3.1: it equals the span of the eigenspaces of $\lambda_1, \ldots, \lambda_k$. Conclude that the projection P_k of Theorem 3.5, and hence the rank-k reconstruction $\hat{X}$, is unique whenever $\lambda_k > \lambda_{k+1}$.

9. **(H)** (Identification of the one-factor model at $p = 3$.) Let $\Sigma \in \mathbb{R}^{3\times 3}$ admit a 1-factor decomposition $\Sigma = \Lambda\Lambda^\top + \Psi$ (Definition 5.1) with $\Lambda = (\Lambda_1, \Lambda_2, \Lambda_3)^\top$ and all $\Lambda_i \neq 0$. (a) Show $\Lambda_1^2 = \Sigma_{12}\Sigma_{13}/\Sigma_{23}$ (and cyclically), so that (Λ, Ψ) is determined by Σ up to the sign ambiguity $\Lambda \mapsto -\Lambda$; reconcile this with Proposition 5.6 (what are the orthogonal matrices in $\mathbb{R}^{1\times 1}$?). (b) Show identification fails when some $\Sigma_{jk} = 0$ with $j \neq k$: exhibit one Σ with two essentially different decompositions. (c) Contrast with Exercise 5: what changes between $p = 2$ and $p = 3$?

10. **(H)** (PCA recovers the factor subspace under isotropic noise.) Let $\Sigma = \Lambda\Lambda^\top + \sigma^2 I_p$ with $\Lambda \in \mathbb{R}^{p\times k}$ of full column rank $k < p$ and $\sigma^2 > 0$ — the factor model of Definition 5.1 with $\Psi = \sigma^2 I_p$. Prove: (a) the eigenvalues of Σ are $\nu_1 + \sigma^2 \geq \cdots \geq \nu_k + \sigma^2 > \sigma^2 = \cdots = \sigma^2$, where $\nu_1 \geq \cdots \geq \nu_k > 0$ are the nonzero eigenvalues of $\Lambda\Lambda^\top$; (b) the span of the first k principal directions of Σ equals the column space of Λ (use Exercise 8); (c) the individual principal directions do *not* in general equal the columns of Λ, even up to sign and order — identify the additional condition on $\Lambda^\top\Lambda$ under which they do. This makes precise the claim of Remark 5.8 that PCA and the factor model coincide "in the special regime $\Psi = \sigma^2 I$": what is recovered is the subspace, never the rotation.

11. **(H)** (Orthogonality is load-bearing in the reconstruction theorem.) Theorem 3.5 minimises $\mathbb{E}\|X - PX\|_2^2$ over *orthogonal* projections of rank at most k. An *oblique projection* is a matrix P with $P^2 = P$ but $P^\top \neq P$: it projects onto its range *along* a complement

that is not the orthogonal one. (a) Show that the key identity of the proof of Theorem 3.5, $\mathbb{E}\|X - PX\|_2^2 = \operatorname{tr}(\Sigma) - \operatorname{tr}(\Sigma P)$, uses symmetry and fails for oblique P by deriving the correct general formula $\mathbb{E}\|X - PX\|_2^2 = \operatorname{tr}\left((I-P)\Sigma(I-P)^\top\right)$. (b) Show by explicit example in $\mathbb{R}^2$ — e.g. $\Sigma = I_2$ and $P = \left(\begin{smallmatrix} 1 & c \\ 0 & 0 \end{smallmatrix}\right)$, $c \neq 0$ — that an oblique projection onto the *same* subspace as an orthogonal one has strictly larger expected squared reconstruction error, and compute the excess as a function of c. (c) Where exactly does Lemma 2.5 consume $P = P^\top$? Identify the first line of its proof that fails for oblique P.

12. (H) (The regression identity and its failure, quantitatively.) Work in the observed-factor model of Definition 6.1 and the cross-sectional model of Section 7 side by side. (a) In the setting of Proposition 6.2, let $k = 1$, $p = 2$, $\operatorname{Var}(f_t) = 0.04$, $\operatorname{Cov}(r_{1,t}, f_t) = 0.048$, $\operatorname{Cov}(r_{2,t}, f_t) = 0.02$: compute B and interpret each entry. (b) Now pass to the cross-sectional model with this B as the (exogenous) exposure matrix, $\Omega = I_2$, $\Sigma_f = \operatorname{Var}(f) = 0.04$ (scalar), and $\Sigma_\varepsilon = \operatorname{diag}(0.01,\, 0.09)$. Compute Π, then $\operatorname{Cov}(r, \hat{f})$ and $\operatorname{Cov}(\hat{f})$ from the proof of Proposition 7.14, and evaluate $\operatorname{Cov}(r, \hat{f})\operatorname{Cov}(\hat{f})^{-1}$ numerically. By how much, in relative terms, does each entry miss B? (c) Verify the iff criterion of Proposition 7.14 on your numbers: compute $(I_2 - B\Pi)\Sigma_\varepsilon\Pi^\top$ and confirm it is nonzero; then exhibit a diagonal Σ_ε (other than a multiple of I_2) for which it *is* zero for this B and Ω, or prove none exists.

13. (H) (Constrained WLS, completed.) This exercise fills in the proof sketched in Remark 7.15. Let $B \in \mathbb{R}^{p\times k}$ have $\operatorname{rank}(B) = k - 1$ with one-dimensional null space spanned by $v \neq 0$, let Ω satisfy (C4), and let $c \in \mathbb{R}^k$ satisfy $c^\top v \neq 0$. (a) Construct a matrix $N \in \mathbb{R}^{k\times(k-1)}$ whose columns form a basis of $\{c\}^\perp := \{g : c^\top g = 0\}$, and show that $BN \in \mathbb{R}^{p\times(k-1)}$ has full column rank $k - 1$ (this is where $c^\top v \neq 0$ enters: show $\ker(BN) = \{h : Nh \in \operatorname{span}(v)\} = \{0\}$). (b) Deduce from Proposition 7.4, applied to the design BN, that $J(g) = (r - Bg)^\top\Omega(r - Bg)$ has a unique minimiser over $\{g : c^\top g = 0\}$, and write it in closed form. (c) Show the constrained estimator remains linear in r, exhibit the resulting $k \times p$ matrix Π_c with $\hat{f} = \Pi_c\, r$, and prove the modified pure-factor identity $\Pi_c BN = N^\ddagger$ for a matrix $N^\ddagger$ you should identify — then explain, in one sentence tied to imposition (I7) of Remark 8.2, why the rows of Π_c deserve the name "pure factor portfolios" only relative to the constraint convention. (d) Verify the whole construction numerically on the collinear 4×4 design $B = [\mathbf{1}_4 \mid D]$ of Example 7.8's industries with $\Omega = I_4$ and $c = (0, 1, 1)^\top$ (coordinates ordered market, technology, banking), so that the constraint $c^\top g = 0$ makes the two industry factor returns sum to zero: recover the closed form $\hat{f}_{\text{mkt}} = \frac{1}{4}\sum_i r_i$, $\hat{f}_{\text{T}} = \frac{1}{2}(r_1{+}r_2) - \hat{f}_{\text{mkt}}$, $\hat{f}_{\text{B}} = \frac{1}{2}(r_3{+}r_4) - \hat{f}_{\text{mkt}}$.