LECTURE 7: PRINCIPAL COMPONENTS, SOME STANDARD DISTRIBUTIONS

1. SQUARE ROOTS OF A P.S.D. MATRIX

$$\Sigma \text{ positive semi-definite (p.s.d.)} \Leftrightarrow (\forall c \in \mathbb{R}^n)\, c'\Sigma c \geq 0$$
$$\text{Var}(X) = \Sigma \Rightarrow \Sigma \text{ p.s.d}$$
$$\Sigma \text{ p.s.d.} \Leftrightarrow \Sigma = W'W$$

- W is not unique.
- Upper triangular W is unique if Σ is p.d. (p.s.d. with $|\Sigma| > 0$). This is the **Cholesky decomposition** of Σ.
- With $W' = W$, we have the **symmetric square root** of Σ. (Not often used.)
- $\Sigma = Q\Lambda Q'$, with Λ diagonal, all $\lambda_{ii} \geq 0$, $Q'Q = I$. Columns of Q are the **eigenvectors** of Σ, and corresponding λ_{ii}'s are its **eigenvalues**. $W = \Lambda^{1/2}Q'$ is another W with $W'W = \Sigma$.

2. ORTHOGONAL DECOMPOSITION OF A RANDOM VECTOR

- If $\text{Var}(X) = \Sigma$, $W'W = \Sigma$. Let $Z = W'^{-1}X$. Then $\text{Var}(Z) = I$.
- $X = W'Z$ represents each element of X as a linear combination of mutually uncorrelated, unit-variance, random variables.
- $Y = Q'X$ has $\text{Var}(Y) = \Lambda$. If the elements of Λ are sorted so they decrease in size from left to right, then Y_1 is the first **principal component** of X, Y_2 the second, etc.
- PC decomposition is *not scale-invariant*. Suppose we change the components of X from pounds to ounces, feet to meters, quarts to liters, etc. New random vector $G = DX$, D diagonal. If $Y = Q'X$ are the principal components of X, then rescaling the PC decomposition of X would give us $G = DQ'Y$, which is *an* orthogonal decomposition of G. But applying principal components directly to G and its covariance matrix $D\Sigma D'$ gives us something else.

3. USING PRINCIPAL COMPONENTS

- Each X_i has a representation $X_i = Q_{i\cdot}Y$, so that $\text{Var}(X_i) = \sum_j q_{ij}^2 \lambda_{jj}^2$.
- Often most of the variance of most of the components of X is "accounted for" by the first few principal components. Looking at the structure of the corresponding $q_{i\cdot}$'s and $q_{\cdot j}$'s can be helpful in thinking about how to model the X's.

- To sidestep scale-sensitivity, it is common to standardize all the X_i-variances at 1, so the covariance matrix becomes a **correlation** matrix. The correlation of X_i with X_j is

$$\rho_{ij} = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{Var}(X_i)\operatorname{Var}(X_j)}}.$$

- Principal components is just a way of rewriting a covariance matrix. It is interesting because we hope it gives insight into how to simplify the description of the data distribution, but the PC calculation in itself does not simplify. **Factor analysis** is a class of models that actually do assert that we can simplify the characterization of the distribution on the basis of a principal components decomposition.

4. The Multivariate Normal

- pdf:

$$\varphi(x;\mu,\Sigma) \text{ or } \varphi(x-\mu;\Sigma) \;=\; (2\pi)^{-\frac{n}{2}}\,|\Sigma|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}.$$

- cdf: Φ, not analytic.
- properties:
  - Vector X normal implies that $\operatorname{Var}(X)$ diagonal $\Leftrightarrow$ X mutually independent.
  - $X \sim N(\mu,\Sigma), A, c$ constant $\Leftrightarrow$ $AX + c \sim N(A\mu + c, A\Sigma A')$, i.e. the class of normal vectors is closed under linear transformations.

5. Why so much attention to the normal?

- It's convenient to have the class of distributions closed under linear transformations and to have zero correlation imply independence.
- It's in a certain sense conservative. It's possible to derive a measure of how much uncertainty is embodied in a distribution, and such a measure is called the **entropy** of the distribution. If $EX = \mu$ and $\operatorname{Var}(X) = \Sigma$ are given, the distribution that implies "maximum ignorance" subject to these constraints is $N(0,\Sigma)$. We won't explore this result in this course — consult an information theory text if you are interested in pursuing it — but it is worthwhile knowing that such results exist.
- Central limit theorems. These are theorems that state that simple averages of random variables tend, if the number of terms in the average is large, to a normal form. It can be argued that "error terms" in economic models are often collections of small, unrelated effects, and thus that a CLT justifies treating them as normal.

6. Convergence in Distribution

- A sequence of random vectors $\{X_T\}_{T=1}^{\infty}$ **converges in distribution** to a distribution $\mathcal{L}$ iff

$$(\forall \text{continuous, bounded } f : \mathbb{R}^n \to \mathbb{R})\; E[f(X_T)] \xrightarrow[t\to\infty]{} E[f(Z)],$$

  where Z is any random variable with the distribution $\mathcal{L}$.

- There are many equivalent ways to define convergence in distribution, and even to name it. It is sometimes called **weak convergence** of distributions. (Weak convergence is a concept that can be applied to a more general class of spaces than spaces of probability measures, but for our purposes it is just a synonym for convergence in distribution. See Pollard (2002, Chapter 7) for a complete tabulation of equivalent definitions of convergence in distribution.)
- The traditional definition of convergence in distribution in $\mathbb{R}^1$ is

$$X_T \xrightarrow[T\to\infty]{\mathcal{D}} X \;\Leftrightarrow\; (\forall a \in \mathbb{R}) \left(F_X \text{ continuous at } a \;\Rightarrow\; \left(F_{X_T}(a) \xrightarrow[T\to\infty]{} F_X(a)\right) \right) .$$

- The generalization of this definition to arbitrary probability spaces is

$$X_T \xrightarrow[T\to\infty]{\mathcal{D}} X \;\Leftrightarrow\; (\forall B \in \mathcal{B}) \left(\left(P[X_T \in \partial B] = 0\right) \;\Rightarrow\; \left(P[X_T \in B] \xrightarrow[T\to\infty]{} P[X \in B]\right) \right) ,$$

  where ∂B is the boundary of B (its closure minus its interior) and $\mathcal{B}$ is the Borel σ-field on the space where X and the X_T's take their values.
- Question: Is pointwise convergence of the distribution function at all points of continuity equivalent to convergence in distribution on $\mathbb{R}^k$ for $k > 1$?

7. A Central Limit Theorem

Theorem:

If $\{X_T\}_{T=1}^{\infty}$ is an i.i.d. sequence of random vectors with $EX_T = \mu$ and $\text{Var}(X_T) = \Sigma$, then

$$\frac{1}{\sqrt{T}} \sum_{T=1}^{\infty} X_T \xrightarrow[T\to\infty]{\mathcal{D}} N(\mu, \Sigma)$$

- There are many CLT's.
- The independence assumption can be weakened. The identical distributions assumption can be weakened. But weakening one tends to require strengthening the other, so there is no uniquely most general CLT.

References

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.