PRINCIPLE COMPONENTS; MODELS, PARAMETERS AND LIKELIHOOD; GAUSSIAN REGRESSION

1. Analyzing covariance and correlation matrices

- Obviously the diagonal elements of Σ, being variances of individual random variables, are a measure of the "spread" of their distributions, as in the univariate case.
- The off-diagonal elements of the covariance matrix are a measure of the strength of pairwise relations among the variables.
- But there can be strong multivariate relations among variables that don't show up in pairwise correlations.

2. Characteristics of covariance matrices

- A matrix Σ is **positive semi-definite** (p.s.d.) if and only if for every conformable vector c, $c'\Sigma c \geq 0$.
- The covariance matrix Σ of a random vector X must be p.s.d. because (as you should be able to verify for yourself) $c'\Sigma c = \text{Var}(c'X)$, and a variance cannot be negative.
- Note also that Σ is symmetric, meaning $\Sigma = \Sigma'$. This follows from the fact that $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$.
- Any symmetric, p.s.d. matrix can be a covariance matrix.

3. Eigenvalue decomposition

- Any symmetric, p.s.d., $n \times n$ matrix Σ can be decomposed as

$$\Sigma = \sum_{i=1}^{n} v_i \lambda_i v_i' = V \Lambda V' ,$$

  where λ_i, $i = 1, \ldots, n$ are non-negative real numbers, v_i are $n \times 1$ vectors, V is a $n \times n$ matrix with the v_i as columns, and Λ is a diagonal matrix with the λ_i down the diagonal.
- V satisfies $V'V = I$, which is to say that V is an **orthonormal** matrix. The columns of V are the right **eigenvectors** of Σ (because $\Sigma V = V\Lambda$ and the λ_i are the **eigenvalues**.
- Matlab, R or Rats will find V and Λ for you with a single command.

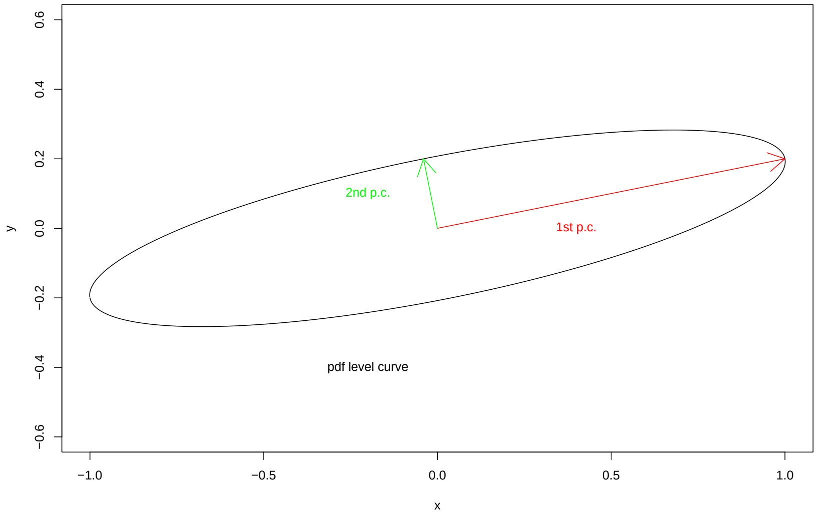
4. From eigenvalue decomposition of Σ to components of X

- The X vector can be respresented as
$$X = \sum_{i=1}^{n} \lambda_i z_i v_i \,,$$
where z_i, $i = 1, \dots, n$ are i.i.d. with mean zero and variance 1.
- The z_i, or sometimes the vector random vectors $\lambda_i z_i v_i$, are known as the **principal components** of X. The v_i's associated with large λ_i's correspond to directions in $\mathbb{R}^n$ in which the X vector varies a lot, while those with small λ_i's correspond to directions with very little variation.
- We could calculate principal components of the correlation matrix also. Even though the correlation matrix is just a rescaling of the Σ matrix, its principal components will be different. That is, the correlation matrix is $D^{-\frac{1}{2}}\Sigma D^{-\frac{1}{2}}$, where D is a diagonal matrix with $\sqrt{\mathrm{Var}(X_i)}$ on the diagonal — this is what we mean by saying the correlation matrix is a rescaling of the covariance matrix. But if we take the eigenvalue decomposition $R = WMW'$ of the correlation matrix R, we will find $V \neq D^{\frac{1}{2}}W$ and there is no simple correspondence between the eigenvalues of R and Σ.
- This reflects a general fact about principal component decompositions — they are not scale invariant. Change the units of measurement of some of the components of X and you will change the principle components decomposition.
- Sensitivity to scaling is not the only pitfall to look out for in using the results of a principal components analysis. If a large number of closely related variables are added to the X vector, the first principal component will eventually reflect mainly the common component of those variables.
- This might be desired behavior. But often we are tempted to use p.c. analysis when we have several imperfect measures of two or more concepts and we are looking for relations between the concepts. For example we have three measures of education and 3 of income, each of them imperfect. Principal components on this vector will give a different answer with the full 6-dimensional X than what we get if we leave out any element of the X vector.
- Nonetheless principal components decompositions of Σ and/or R are useful descriptive devices in many cases.
- Not in all cases. As with any function of a distribution we might use to summarize its shape, whether the summary is useful or not depends on the class of distributions we have in mind and what our uncertainties about the distribution are. As with variances, covariance matrices may not exist. Even

when they do exist, they may be misleading as measures of spread or of dependence between variables.

5. 2-DIMENSIONAL GEOMETRIC INTERPRETATION

- If the joint pdf of X, Y has same-shaped elliptical level curves centered at zero, i.e. if the pdf can be written as $p(ax^2 + bxy + y^2)$ with $b^2 < 4ac$, or equivalently as $p([x, y]M[x, y]')$ with M positive definite (meaning p.s.d. but with all eigenvalues strictly positive), then if X, Y have a finite covariance matrix, it is proportional to M^{-1}, the eigenvectors of the covariance matrix (and of M) are the principal axes of the ellipses, and the lengths of the principal axes are proportional to the square roots of the eigenvalues of the covariance matrix.



6. THE MULTIVARIATE NORMAL DISTRIBUTION

$\underset{1\times n}{X}$ is jointly normal with mean μ and variance matrix Σ implies

$$\text{pdf:} \qquad \phi(x \mid \mu, \Sigma) = (2\pi)^{-n/2} \, |\Sigma|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)'}$$

Suppose $X \sim N(\mu, \Sigma)$ has two component vectors X_1 and X_2 with means μ_1 and μ_2 and

$$\mathrm{Var}(X) = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} .$$

- X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

- Conditional distributions are normal and conditional means linear:

$$X_2 \mid X_1 \sim N(X_1\beta, \Sigma_{2|1}),$$
$$\beta = \Sigma_{11}^{-1}\Sigma_{12} \qquad \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

- This means we can write

$$X_2 = X_1\beta + \varepsilon$$

  with $\{\varepsilon \mid X_1, \beta\} \sim N(0, \Sigma_{2|1})$.

7. Why is a normality assumption so common?

- Central limit theorems: sums of many more or less equally sized, more or less independent, random variables or vectors with finite covariance matrices and common means have approximately normal distributions.
- Information theory — the kind that measures the capacity of your high-speed internet link in bits or bytes per second — suggests a well-defined way to define an amount of information. Being told that X is distributed as $N(\mu, \Sigma)$ is being given less information than being told it has any other pdf with the same mean and variance. In this sense a normality assumption is "conservative".
- Often it turns out that in large samples inference based on the normality assumption is nearly the same as inference based on any other assumption.

8. The SNLM

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\beta + \underset{T\times 1}{\varepsilon}$$
$$\left\{\varepsilon \mid X, \beta, \sigma^2\right\} \sim N(0, \sigma^2 I)$$
$$\text{or}$$
$$\left\{y \mid X, \beta, \sigma^2\right\} \sim N(X\beta, \sigma^2 I)$$

9. Types of random variables in a model

exogenous: X . The model conditions on it, makes no assertion about its distribution, and it will be observed. (This is sometimes called *strict* exogeneity. What other kind there might be we will see later.)

parameters: β, σ^2 . The model conditions on them, makes no assertion about their distribution. and they will not be observed.

nuisance parameter: σ^2, often. We don't care about it directly in many cases, but we have to include it in the list of parameters to complete the model.

disturbances or shocks: ε They will not be observed and the model makes assertions about their joint distribution.

10. Which Greek letters get distributions?

- Why couldn't we give β a distribution, say $N(\bar{\beta}, \Omega)$, and condition on ε? Then the model becomes
$$\{y \mid X, \varepsilon, \bar{\beta}, \Omega\} \sim N(X\bar{\beta} + \varepsilon, X\Omega, X') .$$
- Is there something wrong with interpreting a regression equation this way?
- There are situations where something like this inversion of roles for β and ε makes sense. But usually it doesn't.
- ε, though it can't be observed, can be precisely estimated if we have precise estimates of β, and since T is usually fairly large, we can check the validity of claims about the distribution of the i.i.d. ε_t's.
- There is only one β in the model. As we shall see, if T is fairly large (or more precisely if $X'X$ is big), we can get a precise estimate of β. But the pre-sample distribution of β can generally not be proved right or wrong. If you think β is $N(0, I)$ before you see y, while I think it's $N(0, 2I)$, you may be more surprised than I if the sample information implies that $\beta = (3, 3, 3)'$ is likely, but this doesn't prove that you or I wrongly characterized our pre-sample uncertainty. Since there are no repeated observations on the distribution of β, it can't be proved right or wrong.

11. Criteria for parameters and disturbances

- Parameters usually do not have distributions from which we observe repeated draws. Only distributions that describe our presample uncertainty, and post-sample distributions conditional on the data.
- Disturbances have distributions whose validity can be checked by observing data, because there are many draws from their distributions reflected in the data.
- Parameters are quantities which are in this sense "subjective", while disturbances have "objective" uncertainty.

12. Blurry lines between parameters and disturbances

An example where the dividing line is not clear. N industries, T years in each industry,
$$y_{it} = \alpha + X_{it}\beta + \nu_i + \varepsilon_{it} .$$
Is ν_i a parameter or a disturbance? If $i = 1, 2$ indexes durables and nondurables, then it is hard to argue that ν_i is repeatedly observed and thus has an objective

distribution. But if $T = 2$ and $N = 100$ (which happens in actual data sets), then ν_i starts to look like a disturbance.

This is the simplest version of a panel data model, and we will come back to it.

13. The likelihood principle, sufficiency

LHP: If our model gives the data Y the conditional pdf $p(y \mid \beta)$, then all we need to know about the data in order to form the conditional pdf $q(\beta \mid y)$ is the **likelihood function** $p(y \mid \cdot)$, and we need to know it only up to a factor of proportionality. This follows from Bayes'rule.

Sufficiency: If there is a (possibly vector-valued) function $S(y)$ such that

$$p(y \mid \beta) = p_1\big(S(y), \beta\big)\, p_2(y)\,,$$

then inference about β depends on the data only through $S(y)$. It can be shown under fairly general conditions that decision rules $\delta(y)$ that cannot be written as $\delta(S(y))$ are suboptimal.

14. Inference in the SNLM

- The likelihood function:

$$(2\pi\sigma^2)^{-T/2} \exp\big(-\tfrac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\big)\,.$$

Sufficient statistics: $y'y$, $y'X$, $X'X$(?). Note that these are generally a much smaller set of numbers than the data themselves, y, X.