CHARACTERIZING DISTRIBUTIONS, CONTINUED; BAYES RULE; MODELS

CHRISTOPHER A. SIMS
PRINCETON UNIVERSITY
SIMS@PRINCETON.EDU

1. Derivation of the Change of Variables Rule

- Let $\ell(a,b,c,d)$ denote the parallelogram in $\mathbb{R}^2$ defined by the two vectors (a,b) and (c,d). See the graph in Figure 1. The area of such a parallelogram can be found (recall high school geometry) as "base times height", where here the base can be taken as the length of the vector (c,d) and the height as the length of the line labeled "h", which runs from (a,b) to meet (c,d) in a 90° angle. It is perhaps somewhat surprising that this turns out to be exactly the same as
$$ad - bc = \begin{vmatrix} a & b \\ c & d \end{vmatrix},$$
but you should be able to verify this for yourself.
- If X, Y have joint pdf $p(x,y)$, the probability of a small square with one corner x, y and the other $x+\delta, y+\delta$ is approximately $p(x,y)\delta^2$. Suppose $U = f(X,Y)$, $V = g(X,Y)$. With f and g differentiable, our square in x,y-space gets mapped into a parallelogram in u,v-space of the form
$$\big(f(x,y), g(x,y)\big) + \ell\big(D_1 f(x,y)\delta, D_1 g(x,y)\delta, D_2 f(x,y)\delta, D_2 g(x,y)\delta\big).$$

Date: October 4, 2001.

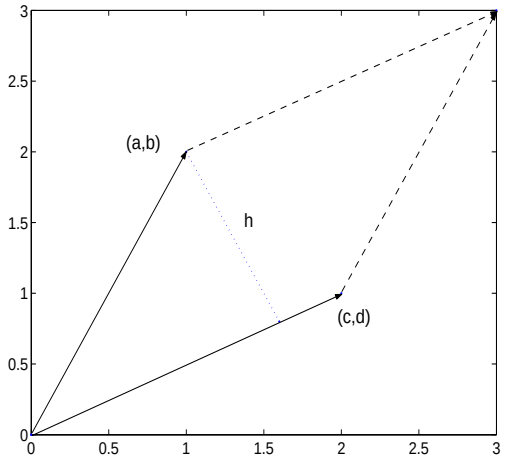

FIGURE 1. $\ell(a,b,c,d)$

That is, a parallelogram with one corner at $(u, v) = (f(x, y), g(x, y))$ and defined by the two vectors starting from (u, v) and going to $(u, v) + (D_1 f(x, y)\delta, D_1 g(x, y)\delta)$ and $(u, v) + (D_2 f(x, y)\delta, D_2 g(x, y)\delta$. Note that the square we started with in x, y-space is itself $(x, y) + \ell(\delta, 0, 0, \delta)$. If $g(u, v)$ is the pdf of U, V, then for small δ it must be true that the probability of this parallelogram is approximately its area times $g(u, v)$, i.e.

$$\big(D_1 f(x, y) D_2 g(x, y) - D_1 g(x, y) D_2 f(x, y)\big)\delta^2 g(u, v) = \left|\frac{\partial(u, v)}{\partial(x, y)}\right| \delta^2 g(u, v) \, .$$

But this must also be the probability of the small square in x, y-space that we started with, which allows us to conclude that

$$g(u, v) \left|\frac{\partial(u, v)}{\partial(x, y)}\right| = f(x, y) \, ,$$

which is the change of variables formula.

- It should be clear that this argument generalizes to $\mathbb{R}^k$ as soon as we verify that the volume of a k-dimensional parallelogram defined by k vectors a_{ij}, $i = 1 \ldots k$, $j = 1 \ldots k$ in $\mathbb{R}^k$ is $\left|a_{ij}\right|$. This task is left to the skeptical student.

2. Quantiles

- $p'th$ quantiles: $x_p \in \mathbb{R}$ such that $P[X < x_p] = p$.
- median, quartiles, interquartile range. Common substitutes for mean, standard deviation (square root of variance).
- Insensitive to tail behavior.
- Not linear. So mean and variance usually allow a better estimate of the median and interquartile range of a sum of independent, identically distributed, (i.i.d) variables than do median and interquartile range of the individual variables themselves.

3. Taylor expansion of $\log p$

- Recap of Taylor series:

$$f(x) \doteq f(x_0) + Df(x_0)(x - x_0) + \tfrac{1}{2!}(x - x_0)' D^2 f(x_0)(x - x_0) + ...$$

- This is not going to work well if applied to a pdf directly, because $p > 0$, $p(x) \to 0$ as $x \to \infty$. Polynomials can't behave that way.
- But e raised to a polynomial exponent does behave that way, so long as the degree of the polynomial is even and the largest power appearing in it has a negative coefficient.
- A very widely applied strategy for summarizing a distribution, especially when it is high-dimensional: a second order Taylor expansion of $\log p$.

- Usually $x_0 = \text{argmax}\{p(x)\}$ or some point close to it. Accuracy of the expansion is greatest near x_0, and accuracy of $\log p$ approximation matters little for probability purposes in regions where $p \doteq 0$.
- So in the univariate case we are approximating $p(x)$ by

$$\hat{p}(x) = e^{a_2(x-x_0)^2 + a_1(x-x_0) + a_0}.$$

- If $p(x_0)$ is the maximum of p, then $Dp(x_0) = 0$, so $a_1 = 0$.
- a_0 just scales $\hat{p}$, and most often we are free to choose it so that $\hat{p}$ integrates to 1.
- $a_2 = \frac{1}{2}D^2 \log p(x_0)$, which has to be negative for this to work.
- A pdf of this form, with a_0 chosen so it integrates to 1, is called a $N(x_0, -1/(2a_2))$ distribution. It has mean zero and variance $-1/(2a_2) = -1/D^2 f(x_0)$. The standard way to write it is as the pdf of a $N(\mu, \sigma^2)$ distribution:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The Taylor series approximation we are discussing, then, approximates p by a normal distribution with mean $\mu = \text{argmax}(p)$ and variance $-1/D^2 \log p(\mu)$.

4. The Multivariate Normal

- This kind of approximation can be done as well when X is a vector. Then the approximating pdf takes the form

$$e^{(x-x_0)' a_0 (x-x_0) + a_1(x-x_0) + a_0}.$$

  As before, if we expand around the peak of the pdf, $a_1 = 0$.
- The multivariate normal pdf has the form

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

  This is a pdf, and it has mean μ and variance matrix $\text{Var}(X) = \Sigma$. Our approximation is then a multivariate normal with $\mu = \text{argmax}(p(x))$ and $\Sigma = -\left(D^2 \log p(\mu)\right)^{-1}$.
- The definition of $\text{Var}(X)$:

$$\underset{n\times n}{\mu_2} = E[XX'], \quad \mu_1 = EX, \quad \text{Var}(X) = \mu_2 - \mu_1\mu_1'$$

5. Properties of the multivariate normal

- Consider any family of pdfs consisting of all pdf's of the form $p(x) = g(x'a_2x + a_1x + a_0)$, where g is fixed and the a_i's vary over the family, and where a_2 is positive definite. If $z = Ax$ and A is square with $|A| \neq 0$, the pdf of z remains within the same family (i.e. can be written with the same g, just different a's. As

a special case, a linear transformation of a normal random vector is itself normal. Follows from application of the change of variables rule.

- If X is jointly normal, all marginal and conditional distributions of individual X's or groups of X's are also normal. This is not too hard to see — they all end up as pdf's whose logs are quadratic.
- A multivariate random vector X is independent if and only if it has a diagonal Σ matrix, i.e. if and only if all the X_i's have zero **covariances** with each other. (The covariance of X_i with X_j, written $\mathrm{Cov}(X_i, X_j)$ is the i, j'th element of $\mathrm{Var}(X)$, i.e. $E[X_i X_j] - E[X_i]E[X_j]$.) This means also that if we have a jointly normal random vector, and all its elements are pairwise independent, then the full vector is independent. As we noted earlier, this is not true in general for variables that are not jointly normal.
- If X, Y are independent, then $\mathrm{Cov}(X, Y) = 0$. The reverse implication is not true in general.

6. Normal conditional densities: details

Note that some of the expressions on the board during the lecture that covered this contained mistakes in signs, subscript order, etc. The formulas below should be right.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

We introduce the notation

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$
$$\Sigma^{-1} \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix}$$

Therefore the joint pdf is (where X_i is $n_i \times 1$)

$$p(x_1, x_2) = (2\pi)^{-(n_1+n_2)/2} |\Sigma|^{-\frac{1}{2}} \cdot$$
$$\exp\Big(-\tfrac{1}{2}\big((x_1-\mu_1)'\Sigma^{11}(x_1-\mu_1) + 2(x_1-\mu_1)'\Sigma^{12}(x_2-\mu_2) + (x_2-\mu_2)'\Sigma^{22}(x_2-\mu_2)\big)\Big) .$$

In forming the conditional pdf of $X_2 \mid X_1$, we care only about how this expression behaves as a function of x_2, and we would like to cast the exponent, which is the only part that depends on x_2, in the standard form of a $N(\mu, \Sigma)$ pdf. It is a standard exercise in the algebra of "completing the square" to determine that the argument of exp can be written as

$$(x_2 - \mu_2 - \beta(x_1 - \mu_1))'\Sigma^{22}(x_2 - \mu_2 - \beta(x_1 - \mu_1)) + \text{ a function of } x_1 ,$$

where

$$\beta = -\left(\Sigma^{22}\right)^{-1}\Sigma^{21} .$$

From this it is clear that, once normalized to integrate to one, the distribution will be of the form

$$N\Big(\mu_2 + \beta(x_1 - \mu_1), \big(\Sigma^{22}\big)^{-1}\Big)$$

Notice that this means

- The conditional variance of $X_2 \mid X_1$ does not depend at all on X_1.
- The conditional mean depends linearly on X_1, and is the same as the unconditional mean μ_2 when $X_1 = \mu_1$.
- When $\Sigma^{12} = 0$, which is true if and only if $\Sigma_{12} = 0$, The conditional distribution is the same as the unconditional distribution and does not depend on X_1. I.e., in this case X_1, X_2 are independent of each other.

These formulas are more useful if we rewrite them in terms of the four blocks in the original Σ matrix rather than the blocks of Σ^{-1}. Note that, from the fact that $\Sigma^{-1}\Sigma = I$,

$$\Sigma^{21}\Sigma_{11} + \Sigma^{22}\Sigma_{21} = 0\,.$$

From this we can see immediately that

$$\beta = -\big(\Sigma^{22}\big)^{-1}\Sigma^{21} = \Sigma_{21}\Sigma_{11}^{-1}\,.$$

We also know that

$$\Sigma^{21}\Sigma_{12} + \Sigma^{22}\Sigma_{22} = \Sigma^{22}(\beta\Sigma_{12} + \Sigma_{22}) = I\,.$$

From this and our previous expression for β it is easy to get

$$\big(\Sigma^{22}\big)^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\,.$$

So an alternative expression for the pdf of $X_2 \mid X_1$ is

$$N\big(\mu_2 + \beta(x_1 - \mu), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\big)\,, \quad \text{where } \beta = \Sigma_{21}\Sigma_{11}^{-1}\,.$$

7. Principal Components

- $\mathrm{Var}(X)$ must be a positive semi-definite (p.s.d.) matrix. If c is a $1 \times n$ vector of constants and $Z = cX$, then linearity of E implies $\mathrm{Var}(Z) = c\,\mathrm{Var}(X)c'$. But this must be non-negative, then, for any c, and $c\,\mathrm{Var}(X)c' \geq 0$ for every c is the definition of $\mathrm{Var}(X)$ p.s.d.
- $$\mathrm{Var}(X) = \Sigma\,, \quad \underset{k\times n}{Z} = cX\,, \quad \Rightarrow \quad \mathrm{Var}(Z) = c\Sigma c'\,.$$
- Any p.s.d. matrix Σ can be written as
  $$\Sigma = Q'DQ\,,$$
  with $Q'Q = I$ (Q **orthonormal**) and D diagonal, with all its diagonal elements non-negative.
- Therefore $X \sim N(\mu, \Sigma)$ implies $Z = QX \sim N(Q\mu, D)$ and $X = Q'Z$. This means X can be represented as a linear combination of independent normal variables.

- The Z variables are known as the **principal components** of X. It sometimes happens that a few diagonal elements of D, say the first q, are much larger than the others. This means that the corresponding Z's have much bigger variance than the others, so that

$$\sum_{j=1}^{q} Q_j Z_j$$

  is a good approximation to X. (X is exactly this sum, if we let q be n. It is approximately this sum with $q < n$ if the variances of the Z_j's are all small for $j > q$.)
- So with a large covariance matrix, reporting the first q rows of Q and the corresponding first q diagonal elements of D, may give a good summary of the structure. Of course it may also turn out that the diagonal elements of D are rather similar, so this strategy for summarizing does not work.
- Principal components analysis is sensitive to the units in which data are measured. If some of our data are distances, measured in feet, and others are weights, measured in pounds, we might consider changing everything to meters and kilograms. The calculated Q and D would then be different, and not just by the rescaling effects we would expect from the change in units. The resulting metric Z vector will be different, not just a rescaling of the original English system Z vector. Hence it is not a good idea to work too hard at giving substantive interpretations to principal components.

8. Choleski decomposition

Principal components correspond to just one way of writing $\Sigma = W'W$. With such a representation, we can always define $Z = (W')^{-1}X$, which makes $\mathrm{Var}(Z) = I$, and then in turn write $X = W'Z$, giving us a representation of X as a linear combination of independent $N(0,1)$ random variables. Principal components amounts to taking $W = D^{1/2}Q$. We get a good summary if only a few rows of W contain large coefficients (as if only a few diagonal elements of D are large, in the case of principal components).

We can always, for p.s.d. Σ, find a W that is upper triangular, meaning its elements ω_{ij} are zero for $i > j$, and that satisfies $\Sigma = W'W$. Such a factoring of Σ is called a **Choleski decomposition** of Σ. It, too, may provide a good summary of the data if only a few rows of W are large. Notice that there are two extreme cases: Only the first few rows of W are large, or only the last few rows of W are large. If only the last few rows of W are large, then because it is upper triangular, only the lower right corner is large. That means that all the elements of X are small except the last few. While this does provide a summary statement about the nature of Σ, we didn't need a decomposition to describe it. More interesting is the case where the only first few rows of W are large. This implies that all the X's depend mainly on the first few Z's only. Furthermore, because of the triangular structure, the first few Z's are linear combinations of the first few X's alone. So with this

structure, the first few elements of X can be thought of as "determining" the remaining ones.

The Choleski decomposition, unlike principal components, is scale-invariant. That is, premultiplying X by a positive definite diagonal matrix will leave the Z's implied by the Choleski decomposition unchanged. On the other hand, the Choleski decomposition, unlike principal components, obviously does depend on how the elements of X are ordered. There may be substantive considerations that suggest that some small subset of variables are naturally thought of as "determining" the rest. In this case, a Choleski decomposition with those variables at the top of the X vector is appealing, as it has some substantive interpretation as well as some chance of showing us a simple structure for Σ.

9. INEQUALITIES

Markov:

$$P[X \geq 0] = 1 \text{ and } E[X] = c \;\Rightarrow\; P[X \geq b] \leq \frac{c}{b}\,.$$

Chebyshev:

$$E[X^2] = c \;\Rightarrow\; P[|X| \geq b] \leq \frac{c}{b^2}$$

Jensen's:

$$f \text{ concave}, E[X] < \infty \;\Rightarrow E[f(X)] \leq f(E[X])$$

Observe that if we let $b \to \infty$ in the Markov inequality, we conclude that the cdf of X must converge to 1 at the rate c/b, i.e. "harmonically" in b. Many distributions decline faster than this, but if the distribution has finite expectation it must decline at least this fast. If we let $b \to \infty$ in the Chebyshev inequality, we get a similar bound on the rate of convergence of the cdf to 0 and 1 as its argument goes to $\pm\infty$, though this bound is more stringent, since it implies convergence at the rate $1/b^2$.