Chapter 14

Multivariate Distributions

14.1 Review of Definitions

Let's review some definitions from basic probability. When we have a random vector $\vec{X}$ with p different components, $X_1, X_2, \ldots X_p$, the **joint cumulative distribution function** is

$$F(\vec{a}) = F(a_1, a_2, \ldots a_p) = \Pr\left(X_1 \leq a_1, X_2 \leq a_2, \ldots X_p \leq a_p\right) \tag{14.1}$$

Thus

$$F(\vec{b}) - F(\vec{a}) = \Pr\left(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \ldots a_p < X_p \leq b_p\right) \tag{14.2}$$

This is the probability that X is in a (hyper-)rectangle, rather than just in an interval.

The **joint probability density function** is

$$p(\vec{x}) = p(x_1, x_2, \ldots x_p) = \left.\frac{\partial^p F(a_1, \ldots a_p)}{\partial a_1 \ldots \partial a_p}\right|_{\vec{a}=\vec{x}} \tag{14.3}$$

Of course,

$$F(\vec{a}) = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \ldots \int_{-\infty}^{a_p} p(x_1, x_2, \ldots x_p) dx_p \ldots dx_2 dx_1 \tag{14.4}$$

(In this case, the order of integration doesn't matter. Why?)

From these, and especially from the joint PDF, we can recover the marginal PDF of any group of variables, say those numbered 1 through q,

$$p(x_1, x_2, \ldots x_q) = \int p(x_1, x_2, \ldots x_p) dx_{q+} 1 dx_{q+2} \ldots dx_p \tag{14.5}$$

(What are the limits of integration here?) Then the conditional pdf for some variables given the others — say, use variables 1 through q to condition those numbered $q+1$

through p — just comes from division:

$$p(x_{q+1}, x_{q+2}, \ldots x_p | X_1 = x_1, \ldots X_q = x_q) = \frac{p(x_1, x_2, \ldots x_p)}{p(x_1, x_2, \ldots x_q)} \tag{14.6}$$

These two tricks can be iterated, so, for instance,

$$p(x_3|x_1) = \int p(x_3, x_2|x_1) dx_2 \tag{14.7}$$

14.2 Multivariate Gaussians

The multivariate Gaussian is just the generalization of the ordinary Gaussian to vectors. Scalar Gaussians are parameterized by a mean μ and a variance σ^2, which we symbolize by writing $X \sim \mathcal{N}(\mu, \sigma^2)$. Multivariate Gaussians, likewise, are parameterized by a mean vector $\vec{\mu}$, and a variance-covariance matrix $\mathbf{\Sigma}$, written $\vec{X} \sim \mathcal{MVN}(\vec{\mu}, \mathbf{\Sigma})$. The components of $\vec{\mu}$ are the means of the different components of $\vec{X}$. The i, j^{th} component of $\mathbf{\Sigma}$ is the covariance between X_i and X_j (so the diagonal of $\mathbf{\Sigma}$ gives the component variances).

Just as the probability density of scalar Gaussian is

$$p(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \tag{14.8}$$

the probability density of the multivariate Gaussian is

$$p(\vec{x}) = (2\pi \det \mathbf{\Sigma})^{-p/2} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}) \cdot \mathbf{\Sigma}^{-1}(\vec{x} - \vec{\mu})\right\} \tag{14.9}$$

Finally, remember that the parameters of a Gaussian change along with linear transformations

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2) \tag{14.10}$$

and we can use this to "standardize" any Gaussian to having mean 0 and variance 1 (by looking at $\frac{X-\mu}{\sigma}$). Likewise, if

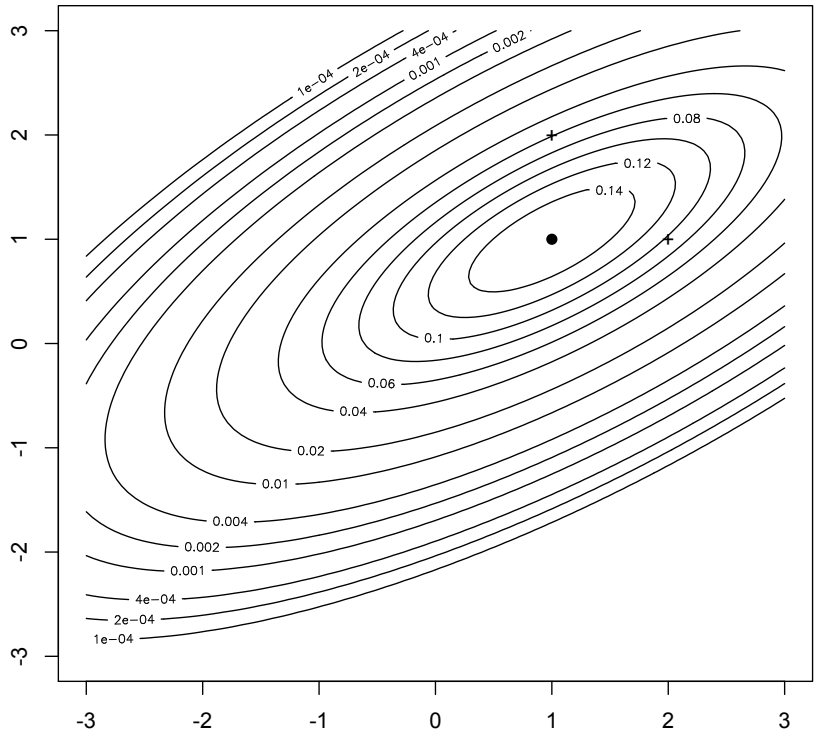
$$\vec{X} \sim \mathcal{MVN}(\vec{\mu}, \mathbf{\Sigma}) \tag{14.11}$$

then

$$\mathbf{a}\vec{X} + \vec{b} \sim \mathcal{MVN}(\mathbf{a}\vec{\mu} + \vec{b}, \mathbf{a}\mathbf{\Sigma}\mathbf{a}^T) \tag{14.12}$$

In fact, the analogy between the ordinary and the multivariate Gaussian is so complete that it is very common to not really distinguish the two, and write $\mathcal{N}$ for both.

The multivariate Gaussian density is most easily visualized when $p = 2$, as in Figure 14.1. The probability contours are ellipses. The density changes comparatively slowly along the major axis, and quickly along the minor axis. The two points marked + in the figure have equal geometric distance from $\vec{\mu}$, but the one to its right lies on a higher probability contour than the one above it, because of the directions of their displacements from the mean.




```
library(mvtnorm)
x.points <- seq(-3,3,length.out=100)
y.points <- x.points
z <- matrix(0,nrow=100,ncol=100)
mu <- c(1,1)
sigma <- matrix(c(2,1,1,1),nrow=2)
for (i in 1:100) {
  for (j in 1:100) {
     z[i,j] <- dmvnorm(c(x.points[i],y.points[j]),mean=mu,sigma=sigma)
  }
}
contour(x.points,y.points,z)
```

Figure 14.1: Probability density contours for a two-dimensional multivariate Gaussian, with mean $\vec{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ (solid dot), and variance matrix $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. Using `expand.grid`, as in Chapter 4 and later, would be more elegant coding than this double `for` loop.

14.2.1 Linear Algebra and the Covariance Matrix

We can use some facts from linear algebra to understand the general pattern here, for arbitrary multivariate Gaussians in an arbitrary number of dimensions. The covariance matrix $\mathbf{\Sigma}$ is symmetric and positive-definite, so we know from matrix algebra that it can be written in terms of its eigenvalues and eigenvectors:

$$\mathbf{\Sigma} = \mathbf{v}^T \mathbf{d} \mathbf{v} \tag{14.13}$$

where $\mathbf{d}$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}$, and $\mathbf{v}$ is the matrix whose columns are the eigenvectors of $\mathbf{\Sigma}$. (Conventionally, we put the eigenvalues in $\mathbf{d}$ in order of decreasing size, and the eigenvectors in $\mathbf{v}$ likewise, but it doesn't matter so long as we're consistent about the ordering.) Because the eigenvectors are all of length 1, and they are all perpendicular to each other, it is easy to check that $\mathbf{v}^T\mathbf{v} = \mathbf{I}$, so $\mathbf{v}^{-1} = \mathbf{v}^T$ and $\mathbf{v}$ is an orthogonal matrix. What actually shows up in the equation for the multivariate Gaussian density is Σ^{-1}, which is

$$(\mathbf{v}^T \mathbf{d} \mathbf{v})^{-1} = \mathbf{v}^{-1} \mathbf{d}^{-1} \left(\mathbf{v}^T\right)^{-1} = \mathbf{v}^T \mathbf{d}^{-1} \mathbf{v} \tag{14.14}$$

Geometrically, orthogonal matrices represent rotations. Multiplying by $\mathbf{v}$ rotates the coordinate axes so that they are parallel to the eigenvectors of $\mathbf{\Sigma}$. Probabilistically, this tells us that the axes of the probability-contour ellipse are parallel to those eigenvectors. The radii of those axes are proportional to the square roots of the eigenvalues. To see *that*, look carefully at the math. Fix a level for the probability density whose contour we want, say f_0. Then we have

$$\begin{aligned}
f_0 &= (2\pi \det \mathbf{\Sigma})^{-p/2} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}) \cdot \mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right\} && (14.15)\\
c &= (\vec{x}-\vec{\mu}) \cdot \mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu}) && (14.16)\\
&= (\vec{x}-\vec{\mu})^T \mathbf{v}^T \mathbf{d}^{-1} \mathbf{v} (\vec{x}-\vec{\mu}) && (14.17)\\
&= (\vec{x}-\vec{\mu})^T \mathbf{v}^T \mathbf{d}^{-1/2} \mathbf{d}^{-1/2} \mathbf{v} (\vec{x}-\vec{\mu}) && (14.18)\\
&= \left(\mathbf{d}^{-1/2}\mathbf{v}(\vec{x}-\vec{\mu})\right)^T \left(\mathbf{d}^{-1/2}\mathbf{v}(\vec{x}-\vec{\mu})\right) && (14.19)\\
&= \left\|\mathbf{d}^{-1/2}\mathbf{v}(\vec{x}-\vec{\mu})\right\|^2 && (14.20)
\end{aligned}$$

where c combines f_0 and all the other constant factors, and $\mathbf{d}^{-1/2}$ is the diagonal matrix whose entries are one over the square roots of the eigenvalues of $\mathbf{\Sigma}$. The $\mathbf{v}(\vec{x}-\vec{\mu})$ term takes the displacement of $\vec{x}$ from the mean, $\vec{\mu}$, and replaces the components of that vector with its projection on to the eigenvectors. Multiplying by $\mathbf{d}^{-1/2}$ then scales those projections, and so the radii have to be proportional to the square roots of the eigenvalues.[1]

[1] If you know about principal components analysis and you find all this manipulation of eigenvectors and eigenvalues of the covariance matrix very reminiscent of principal components analysis, you're right; this was one of the ways in which PCA was originally discovered. But PCA does not require any distributional assumptions. If you do not know about PCA, wait for Chapter 18.

14.2.2 Conditional Distributions and Least Squares

Suppose that $\vec{X}$ is bivariate, so $p = 2$, with mean vector $\vec{mu} = (\mu_1, \mu_2)$, and variance matrix $\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. One can show (exercise!) that the conditional distribution of X_2 given X_1 is Gaussian, and in fact

$$X_2|X_1 = x_1 \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \tag{14.21}$$

To understand what is going on here, remember from Chapter 1 that the optimal slope for linearly regressing X_2 on X_1 would be $\mathrm{Cov}[X_2, X_1]/\mathrm{Var}[X_1]$. This is *precisely* the same as $\Sigma_{21}\Sigma_{11}^{-1}$. So in the bivariate Gaussian case, the best linear regression and the optimal regression are exactly the same — there is no need to consider nonlinear regressions. Moreover, we get the same conditional variance for each value of x_1, so the regression of X_2 on X_1 is homoskedastic, with independent Gaussian noise. This is, in short, exactly the situation which all the standard regression formulas aim at.

More generally, if $X_1, X_2, \ldots X_p$ are multivariate Gaussian, then conditioning on $X_1, \ldots X_q$ gives the remaining variables $X_{q+1}, \ldots X_p$ a Gaussian distribution as well. If we say that $\vec{\mu} = (\vec{\mu}_A, \vec{\mu}_B)$ and $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AB} \\ \mathbf{\Sigma}_{BA} & \mathbf{\Sigma}_{BB} \end{bmatrix}$, where A stands for the conditioning variables and B for the conditioned, then

$$\vec{X}_B|\vec{X}_A = \vec{x}_a \sim \mathcal{MVN}(\vec{\mu}_B + \mathbf{\Sigma}_{BA}\mathbf{\Sigma}_{AA}^{-1}(\vec{x}_A - \vec{\mu}_A), \mathbf{\Sigma}_{BB} - \mathbf{\Sigma}_{BA}\mathbf{\Sigma}_{AA}^{-1}\mathbf{\Sigma}_{AB}) \tag{14.22}$$

(Remember that here $\mathbf{\Sigma}_{BA} = \mathbf{\Sigma}_{AB}^T$ [Why?].) This, too, is just doing a linear regression of $\vec{X}_B$ on $\vec{X}_A$.

14.2.3 Projections of Multivariate Gaussians

A useful fact about multivariate Gaussians is that all their univariate projections are also Gaussian. That is, if $\vec{X} \sim \mathcal{MVN}(\vec{\mu}, \mathbf{\Sigma})$, and we fix any unit vector $\vec{w}$, then $\vec{w} \cdot \vec{X}$ has a Gaussian distribution. This is easy to see if $\mathbf{\Sigma}$ is diagonal: then $\vec{w} \cdot \vec{X}$ reduces to a sum of independent Gaussians, which we know from basic probability is also Gaussian. But we can use the eigen-decomposition of $\mathbf{\Sigma}$ to check that this holds more generally.

One can also show that the converse is true: if $\vec{w} \cdot \vec{X}$ is a univariate Gaussian for *every* choice of $\vec{w}$, then $\vec{X}$ must be multivariate Gaussian. This fact is more useful for probability theory than for data analysis[2], but it's still worth knowing.

14.2.4 Computing with Multivariate Gaussians

Computationally, it is not hard to write functions to calculate the multivariate Gaussian density, or to generate multivariate Gaussian random vectors. Unfortunately,

[2]It's a special case of a result called the **Cramér-Wold theorem**, or the **Cramér-Wold device**, which asserts that two random vectors $\vec{X}$ and $\vec{Y}$ have the same distribution if and only if $\vec{w} \cdot \vec{X}$ and $\vec{w} \cdot \vec{Y}$ have the same distribution for every $\vec{w}$.

no one seems to have thought to put a standard set of such functions in the basic set of R packages, so you have to use a different library. The `MASS` library contains a function, `mvrnorm`, for generating multivariate Gaussian random vectors. The `mvtnorm` contains functions for calculating the density, cumulative distribution and quantiles of the multivariate Gaussian, as well as generating random vectors[3] The package `mixtools`, which will use in Chapter 20 for mixture models, includes functions for the multivariate Gaussian density and for random-vector generation.

14.3 Inference with Multivariate Distributions

As with univariate distributions, there are several ways of doing statistical inference for multivariate distributions. Here I will focus on parametric inference, since non-parametric inference is covered in the next chapter.

14.3.1 Estimation

The oldest method of estimating parametric distributions is **moment-matching** or the **method of moments**. If there are q unknown parameters of the distribution, one picks q expectation values — means, variances, and covariances are popular — and finds algebraic expressions for them in terms of the parameters. One then sets these equal to the sample moments, and solves for the corresponding parameters. This method can fail if you happen to chose algebraically redundant moments, since then you really have fewer equations than unknowns[4]. Perhaps more importantly, it quickly becomes very awkward to set up and solve all the necessary equations, and anyway this neglects a lot of information the data.

The approach which has generally replaced the method of moments is simply the method of maximum likelihood. The likelihood is defined in *exactly* the same way for multivariate distributions as for univariate ones. If the observations $\vec{x}_i$ are assumed to be independent, and θ stands for all the parameters bundled together, then

$$L(\theta) = \prod_{i=1}^{n} p(\vec{x}_i; \theta) \tag{14.23}$$

and the maximum likelihood estimate (MLE) is

$$\widehat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} L(\theta) \tag{14.24}$$

Again, as in the univariate case, it is usually simpler and more stable to use the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log p(\vec{x}_i; \theta) \tag{14.25}$$

[3] It also has such functions for multivariate t distributions, which are to multivariate Gaussians exactly as ordinary t distributions are to univariate Gaussians.

[4] For instance, you can't use variances, covariances *and* correlations, since knowing variances and covariances fixes the correlations.

making use of the fact that

$$\operatorname*{argmax}_{\theta} L(\theta) = \operatorname*{argmax}_{\theta} \ell(\theta) \tag{14.26}$$

The simplest possible case for this is the multivariate Gaussian, where the MLE *is* the sample mean vector and the sample covariance matrix. Generally, however, the maximum likelihood estimate and the moment-matching estimate will not coincide.

Of course, for inference, we generally need more than just a point estimate like $\widehat{\theta}_{MLE}$, we need some idea of uncertainty. We can get that pretty generically from maximum likelihood. *Very* informally, since we are maximizing the log-likelihood, the precision with which we estimate the parameter depends on how sharp that maximum is — the bigger the second derivative, the more precise our estimate. In fact, one can show (Wasserman, 2003, §9.7 and 9.10) that

$$\widehat{\theta}_{MLE} \rightsquigarrow \mathcal{MVN}(\theta_0, -\mathbf{H}^{-1}(\theta_0)) \tag{14.27}$$

where θ_0 is the true parameter value, and $\mathbf{H}$ is the **Hessian** of the log-likelihood, its matrix of second partial derivatives,

$$H_{jk}(\theta) = \left.\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right|_{\theta} \tag{14.28}$$

In turn,

$$\frac{1}{n} H_{jk}(\theta_0) \to \mathbf{E}\left[\frac{\partial^2 \log p(X;\theta_0)}{\partial \theta_j \partial \theta_k}\right] \equiv -I_{jk}(\theta_0) \tag{14.29}$$

which defines the **Fisher information matrix I**. One can therefore get (approximate) confidence regions by assuming that $\widehat{\theta}_{MLE}$ has a Gaussian distribution with covariance matrix $n^{-1}\mathbf{I}^{-1}(\widehat{\theta}_{MLE})$, or, somewhat more accurately, $-\mathbf{H}^{-1}(\widehat{\theta}_{MLE})$. We thus get that $\mathrm{Var}\left[\widehat{\theta}_{MLE}\right] = O(n^{-1})$, and $\widehat{\theta}_{MLE} - \theta_0 = O(n^{-1/2})$.

Note that Eq. 14.27 is *only* valid as $n \to \infty$, and further assumes that (i) the model is well-specified, (ii) the true parameter value θ_0 is in the interior of the parameter space, and (iii) the Hessian matrix is strictly positive. If these conditions fail, then the distribution of the MLE need not be Gaussian, or controlled by the Fisher information matrix, etc.

An alternative to the asymptotic formula, Eq. 14.27, is simply parametric or nonparametric bootstrapping.

14.3.2 Model Comparison

Out of sample, models can be compared on log-likelihood. When a strict out-of-sample comparison is not possible, we can use cross-validation.

In sample, a likelihood ratio test can be used. This has two forms, depending on the relationship between the models. Suppose that there is a large or wide model, with parameter Θ, and a narrow or small model, with parameter θ, which we get

by fixing some of the components of Θ. Thus the dimension of Θ is q and that of θ is $r < q$. Since every distribution we can get from the narrow model we can also get from the wide model, in-sample the likelihood of the wide model must always be larger. Thus

$$\ell(\widehat{\Theta}) - \ell(\widehat{\theta}) \geq 0 \tag{14.30}$$

Here we have a clear null hypothesis, which is that the data comes from the narrower, smaller model. Under this null hypothesis, as $n \to \infty$,

$$2[\ell(\widehat{\Theta}) - \ell(\widehat{\theta})] \rightsquigarrow \chi^2_{q-r} \tag{14.31}$$

provided that the restriction imposed by the small model doesn't place it on the boundary of the parameter space of Θ. (See Appendix B.)

For instance, suppose that $\vec{X}$ is bivariate, and the larger model is an unrestricted Gaussian, so $\Theta = \left\{(\mu_1, \mu_2), \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right\}$. A possible narrow model might impose the assumption that the components of $\vec{X}$ are uncorrelated, so $\theta = \left\{(\mu_1, \mu_2), \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right\}$. This is a restriction on the broader model, but not one which is on the boundary of the parameter space, so the large-sample χ^2 distribution should apply. A restriction which *would* be on the boundary would be to insist that X_2 was constant, so $\Sigma_{22} = 0$. (This would also force $\Sigma_{12} = 0$.)

If, on the other hand, that we have two models, with parameters θ and ψ, and they are completely non-nested, meaning there are no parameter combinations where

$$p(\cdot; \theta) = p(\cdot; \psi) \tag{14.32}$$

then in many ways things become easier. For *fixed* parameter values θ_0, ψ_0, the mean log-likelihood ratio is just an average of IID terms:

$$\frac{1}{n}[\ell(\theta_0) - \ell(\psi_0)] \equiv \frac{1}{n}\sum_{i=1}^{n} \Lambda_i \tag{14.33}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i; \theta_0)}{p(x_i; \psi_0)} \tag{14.34}$$

By the law of large numbers, then, the mean log-likelihood ratio converges to an expected value $\mathbf{E}[\Lambda]$. This is positive if θ_0 has a higher expected log-likelihood than ψ_0, and negative the other way around. Furthermore, by the central limit theorem, as n grows, the fluctuations around this expected value are Gaussian, with variance σ^2_Λ / n. We can estimate σ^2_Λ by the sample variance of $\log \frac{p(x_i;\theta_0)}{p(x_i;\psi_0)}$.

Ordinarily, we don't have just a single parameter value for each model, but also ordinarily, $\widehat{\theta}_{MLE}$ and $\widehat{\psi}_{MLE}$ both converge to limits, which we can call θ_0 and psi_0. At the cost of some fancy probability theory, one can show that, in the non-nested case,

$$\frac{\sqrt{n}}{n} \frac{\ell(\widehat{\theta}) - \ell(\widehat{\psi})}{\sigma^2_\Lambda} \rightsquigarrow \mathcal{N}(\mathbf{E}[\Lambda], 1) \tag{14.35}$$

and that we can consistently estimate $\mathbf{E}[\Lambda]$ and σ^2_Λ by "plugging in" $\widehat{\theta}$ and $\widehat{\psi}$ in place of θ_0 and ψ_0. This gives the **Vuong test** for comparing the two models Vuong (1989). The null hypothesis in the Vuong test is that the two models are equally good (and neither is exactly true). In this case,

$$V = \frac{1}{\sqrt{n}} \frac{\ell(\widehat{\theta}) - \ell(\widehat{\psi})}{\widehat{\sigma}_\Lambda} \rightsquigarrow \mathcal{N}(0,1) \tag{14.36}$$

If V is significantly positive, we have evidence in favor of the θ model being better (though not necessarily *true*), while if it is significantly negative we have evidence in favor of the ψ model being better.

The cases where two models *partially* overlap is complicated; see Vuong (1989) for the gory details[5]

14.3.3 Goodness-of-Fit

For univariate distributions, we often assess goodness-of-fit through the Kolmogorov-Smirnov (KS) test[6], where the test statistic is

$$d_{KS} = \max_a |\widehat{F}_n(a) - F(a)| \tag{14.37}$$

with $\widehat{F}_n$ being the empirical CDF, and F its theoretical counterpart. The null hypothesis here is that the data were drawn IID from F, and what Kolmogorov and Smirnov did was to work out the distribution of d_{KS} under this null hypothesis, and show it was the same for all F (at least for large n). This lets us actually calculate p values.

We could use such a test statistic for multivariate data, where we'd just take the maximum over vectors a, rather than scalars. But the problem is that we do not know its sampling distribution under the null hypothesis in the multivariate case — Kolmogorov and Smirnov's arguments don't work there — so we don't know whether a given value of d_{KS} is large or small or what.

There is however a fairly simple approximate way of turning univariate tests into multivariate ones. Suppose our data consists of vectors $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_n$. Pick a unit vector $\vec{w}$, and set $z_i = \vec{w} \cdot \vec{x}_i$. Geometrically, this is just the projection of the data along the direction $\vec{w}$, but these projections are *univariate* random variables. If the $\vec{x}_i$ were drawn from F, then the z_i must have be drawn from the corresponding projection of F, call it $F_{\vec{w}}$. If we can work out the latter distribution, then we can apply our favorite univariate test to the z_i. If the fit is bad, then we know that the $\vec{x}_i$ can't have come from F. If the fit is good for the z_i, then the fit is also good for the $\vec{x}_i$ — at least

[5] If you are curious about why this central-limit-theorem argument doesn't work in the nested case, notice that when we have nested models, and the null hypothesis is true, then $\widehat{\Theta} \to \widehat{\theta}$, so the numerator in the Vuong test statistic, $[\ell(\widehat{\theta}) - \ell(\widehat{\psi})]/n$, is converging to zero, but so is the denominator σ^2_Λ. Since $0/0$ is undefined, we need to use a stochastic version of L'Hoptial's rule, which gives us back Eq. 14.31. See, yet again, Vuong (1989).

[6] I discuss the KS test here for concreteness. Much the same ideas apply to the Anderson-Darling test, the Cramér-von Mises test, and others which, not being such good ideas, were only invented once.

along the direction $\vec{w}$. Now, we can either carefully pick $\vec{w}$ to be a direction which we care about for some reason, or we can chose it *randomly*. If the projection of the $\vec{x}_i$ along several random directions matches that of F, it becomes rather unlikely that they fail to match over-all[7].

To summarize:

1. Chose a random unit vector $\vec{W}$. (For instance, let $\vec{U} \sim \mathcal{MVN}(0, \mathbf{I}_p)$, and $\vec{W} = \vec{U}/||\vec{U}||$.)
2. Calculate $Z_i = \vec{W} \cdot \vec{x}_i$.
3. Calculate the corresponding projection of the theoretical distribution F, call it $F_{\vec{W}}$.
4. Apply your favorite univariate goodness-of-fit test to $\vec{Z}_i$ and $F_{\vec{W}}$.
5. Repeat (1)–(4) multiple times, with Bonferroni correction for multiple testing.

14.4 Exercises

To think through, not to hand in.

1. Write a function to calculate the density of a multivariate Gaussian with a given mean vector and covariance matrix. Check it against an existing function from one of the packages mentioned in §14.2.4.
2. Write a function to generate multivariate Gaussian random vectors, using `rnorm`.
3. If $\vec{X}$ has mean $\vec{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}$, and $\vec{w}$ is a fixed, non-random vector, find the mean and variance of $w \cdot X$.
4. If $\vec{X} \sim \mathcal{MVN}(\vec{\mu}, \mathbf{\Sigma})$, and $\mathbf{b}$ and $\mathbf{c}$ are two non-random matrices, find the covariance matrix of $\mathbf{b}\vec{X}$ and $\mathbf{c}\vec{X}$.

[7] Theoretically, we appeal to the Cramér-Wold device again: the random vectors $\vec{X}$ and $\vec{Y}$ have the same distribution if and only if $\vec{w} \cdot \vec{X}$ and $\vec{w} \cdot \vec{Y}$ have the same distribution for every $\vec{w}$. Failing to match for any $\vec{w}$ implies that $\vec{X}$ and $\vec{Y}$ have different distributions. Conversely, if $\vec{X}$ and $\vec{Y}$ differ in distribution at all, $\vec{w} \cdot \vec{X}$ must differ in distribution from $\vec{w} \cdot \vec{Y}$ for *some* choice of $\vec{w}$. Randomizing the choice of $\vec{w}$ gives us power to detect a lot of differences in distribution.