Risk and Portfolio Management with Econometrics, Courant Institute, Fall 2009
http://www.math.nyu.edu/faculty/goodman/teaching/RPME09/index.html
Jonathan Goodman

Section 2, Basic statistics and econometrics

The mean variance analysis of Section 1 assumes that the investor knows μ and Σ. This Section discusses how one might estimate them from market data. We think of μ and Σ as parameters in a model of market returns. Abstract and general statistical discussions usually use θ to represent one or several model parameters to be estimated from data. We will look at ways to estimate parameters, and the uncertainties in parameter estimates.

These notes have a long discussion of Gaussian random variables and the distributions of t and χ^2 random variables. This may not be discussed explicitly in class, and is background for the reader.

1 Statistical parameter estimation

We begin the discussion of statistics with material that is naive by modern standards. As in Section 1, this provides a simple context for some of the fundamental ideas of statistics. Many modern statistical methods are descendants of those presented here.

Let $f(x, \theta)$ be a probability density as a function of x that depends on a parameter (or collection of parameters) θ. Suppose we have n independent *samples* of the population, $f(\cdot, \theta)$. That means that the X_k are independent random variables each with the probability density $f(\cdot, \theta)$. The goal of parameter estimation is to estimate θ using the samples X_k. The parameter estimate is some function of the data, which is written

$$\theta \approx \widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n) \, . \tag{1}$$

A *statistic* is a function of the data. Thus, the estimate of θ is a statistic. Some statistics, sample means or quantiles, are simple functions of the data. Others are complicated and hard to compute.

For example, suppose $f = \mathcal{N}(\mu, 1)$. That is, the samples come from a normal with unknown mean but variance $\sigma^2 = 1$. If we use the sample mean to estimate the true mean, the estimator is

$$\widehat{\mu}(X_1, \ldots, X_n) = \frac{1}{n} \left(\sum_{k=1}^{n} X_k \right) \, . \tag{2}$$

The notation simplifies if we refer to the whole *dataset*, $\vec{X} = (X_1, \ldots, X_n)$ simply as "the data", or even "the sample". Then a general estimator of θ is $\widehat{\theta}_n(\vec{X})$.

One measure of the accuracy of a statistic, or a statistical estimate, is its *mean square error*

$$E\left[\left(\theta - \widehat{\theta}(\vec{X})\right)^2\right] . \tag{3}$$

As was mentioned in Section 1, it may be more important to measure the error of a statistical estimate by the consequences of the error in our application. For example, there may be no cost for an error less than ϵ and a large cost otherwise. Then the appropriate measure of accuracy is

$$\Pr\left[\left|\theta - \widehat{\theta}\right| \ \geq \ \epsilon\right] .$$

1.1 Frequentist and Bayesian statistics

There are two major "schools" of statistical estimation. The *frequentist* school (sometimes called *Fisherian*, after Fisher) thinks of θ as a fixed but unknown parameter. In particular, θ is not random. Frequentists do not use a probability distribution to model their uncertainty over the value of θ. The mean square error (3) is considered to be a function of θ. A *uniformly best* estimator in the least squares sense is a $\widehat{\theta}(\vec{x})$ that minimizes (3) no matter what θ happens to be. An annoying paradox of frequentist statistics is that this cannot exist, because the ignorant estimator $\widehat{\theta} = \theta_0$ for all $\widehat{X}$ will have mean square error equal to zero if θ happens to be equal to θ_0. Most discussions of statistics (including those here) are frequentist unless they explicitly state otherwise.

The *Bayesian* school (after Bayes) models θ itself as a random variable. In Bayesian terminology, the probability density of θ is the *prior*, called $f(\theta)$. The density for X given θ, is written[1] $f(x \mid \theta)$. This is the same as the function called $f(x, \theta)$ before. The notation here indicates conditional probability, which is the Bayesian understanding of what it is. A Bayesian *experiment* is to first choose $\theta \sim f(\theta)$, then choose n independent samples of X from the conditional density. The resulting joint density of $\vec{X}$ and θ is given by *Bayes' rule*

$$F(\vec{x}, \theta) \ = \ f(\theta) \prod_{k=1}^{n} f(x_k \,|\, \theta) \ .$$

The central object in Bayesian statistics is the *posterior density*. This is the probability density for θ conditional on the observed data:

$$f_p(\theta \,|\, \vec{X}) \ = \ f(\theta \,|\, \vec{X}) \ = \ \frac{f(\theta \text{ and } \vec{X})}{f(\vec{X})} \ , \tag{4}$$

where $f(\vec{x})$ is the marginal density

$$f(\vec{x}) \ = \ \int F(\vec{x}, \theta) \, d\theta \ = \ \int \left(\prod_{k=1}^{n} f(x_k \,|\, \theta) \right) f(\theta) \, d\theta \ . \tag{5}$$

[1] They use f for all probability densities. Different densities are distinguished by the names of their arguments. Thus, $f(\theta)$ is the density of θ, $f(x)$ is the density of x, etc. The functions are both called f, but they are not the same. This is something like operator overloading in `C++`.

General Bayesian statistics (4) and (5) would be impractical without modern Monte Carlo sampling methods.

One advantage of Bayesian statistics over frequentist statistics is that Bayesians have a mechanism for putting in prior information that they might have. For example, in the problem of estimating the expected return on a stock, we might have a range of returns of stocks of similar companies. Another advantage of the Bayesian point of view is that the output is not a single number $\widehat{\theta}$ (called a *point estimate* in this context), but the whole posterior distribution. That is, we get the posterior uncertainty in θ along with the posterior mean. Finally, sometimes a prior distribution on θ serves only to make the estimate robust against *outliers* (wild data points) in the data. This is particularly useful in cases where θ is more complex than just a few numbers and we want to constrain its qualitative behavior.

1.2 Maximum likelihood, one parameter

We stay in the frequentist world for a while. Suppose we have a dataset $\vec{X}$ and a formula $f(x, \theta)$ and that we want to estimate θ. For now, suppose that θ is a single parameter. The multi-parameter version is below. The probability density of the dataset is (because the samples are independent)

$$F(\vec{x}, \theta) \;=\; F(x_1, \ldots, x_n, \theta) \;=\; \prod_{k=1}^{n} f(x_k, \theta) \; .$$

The *likelihood function* is this probability density as a function of θ given the dataset:

$$L(\theta) \;=\; F(\vec{X}, \theta) \;=\; \prod_{k=1}^{n} f(X_k, \theta) \; . \tag{6}$$

The *maximum likelihood* estimator of θ is[2]

$$\widehat{\theta} \;=\; \arg\max_{\theta} L(\theta) \; . \tag{7}$$

The notation "arg max" means to take the argument that maximizes the function. That is, choose $\widehat{\theta}$ so that $\max_\theta L(\theta) = L(\widehat{\theta})$. Of course, $\widehat{\theta}$ depends $\vec{X}$ because L depends on $\vec{X}$.

The maximum likelihood estimator (7) is a heuristic. It does not represent the most likely value of θ. For one thing, θ is not random and has only one value (though we do not know this value). Furthermore, the function $L(\theta)$ is not a probability density as a function of θ. It is positive, but $\int L(\theta)\, d\theta \neq 1$ except in accidental coincidental cases. The term "likelihood" is supposed to sound like "probability" but not actually be probability. What maximum likelihood does is find the parameter value that maximizes the probability density for the data we have.

[2] We persist in saying *the* maximum likelihood estimate even though the maximizer may not be unique or even exist.

Maximum likelihood estimation has important advantages. It is a simple general recipe that applies in a wide range of problems. It also has theoretical properties of being, in very general circumstances, a *consistent* estimator that is *asymptotically optimal* among approximately *unbiased* estimators in the least squares sense (3).

Here are the definitions. The *bias* of an estimator is (writing $\widehat{\theta}$ for $\widehat{\theta}_n$ when the value of n is not important)

$$\text{bias} \;=\; E\Big[\,\widehat{\theta}\,\Big] \;-\; \theta\,. \tag{8}$$

The variance is simply the variance of $\widehat{\theta}$ as a random variable. It is easy to check the error decomposition

$$E\Big[\,\Big(\widehat{\theta}-\theta\Big)^2\,\Big] \;=\; \text{bias}^2 \;+\; \text{var}\Big[\,\widehat{\theta}\,\Big]\;. \tag{9}$$

Let $\widehat{\theta}_n = \widehat{\theta}_n(x_1,\ldots,x_n)$ be a family of estimators. The family is *consistent* if, for any $\epsilon > 0$,

$$\Pr\Big[\,\Big|\,\widehat{\theta}_n - \theta\,\Big| \;\geq\; \epsilon\,\Big] \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty\,. \tag{10}$$

Informally, we say that $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \to \theta$ as $n \to \infty$. In fact, (10) is the definition of convergence *in probability*. If an estimator is consistent, the usual reason is that bias $\Big[\widehat{\theta}_n\Big] \to 0$ and var $\Big[\widehat{\theta}_n\Big] \to 0$ as $n \to \infty$. Though neither condition is technically necessary for (10), there are no consistent estimators used in practice that do not have vanishing bias and variance as $n \to \infty$.

In the decomposition (9), the bias is usually much smaller than the variance (which often is somewhat improperly called *statistical error*). Maximum likelihood estimators, for example, have, for large n,

$$\text{bias}\,\Big[\widehat{\theta}_n\Big] \;\approx\; \frac{C_b}{n}\,. \tag{11}$$

and

$$\text{var}\,\Big[\widehat{\theta}_n\Big] \;\approx\; \frac{C_v}{n}\,. \tag{12}$$

Thus, the bias contributes $O1/n^2)$ to the right side of (9) while the variance contributes $O(1/n)$. (We will see later a trick called *shrinkage* that lowers mean square error by decreasing the variance at the expense of more bias.) Maximum likelihood is *asymptotically optimal* in the mean square sense that among all estimators that satisfy (11) and (12), maximum likelihood has the smallest C_v – the smallest variance and therefore the smallest mean square estimation error.

1.3 Maximum likelihood for univariate normals

The univeriate normal density is

$$f(x,\mu,\sigma) \;=\; \mathcal{N}(\mu,\sigma^2)(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}\,. \tag{13}$$

Suppose, at first, that we have n independent samples of f and that we want to estimate μ with σ being known (admittedly an unlikely scenario). The maximum likelihood estimate comes from maximizing the likelihood function (6) over μ. It is convenient to use the log of the likelihood function (see (13))

$$l(\mu) \;=\; \ln\left(L(\mu)\right) \;=\; -\sum_{k=1}^{n} \ln\left(f(X_k,\mu,\sigma)\right) \;=\; \frac{-1}{2\sigma^2}\sum_{k=1}^{n}\left(X_k-\mu\right)^2 \;+\; C\;,$$

where C is independent of μ and the X_k. Maximizing L is the same as maximizing $l = \ln(L)$. In the present case, maximizing l over μ is the same as minimizing the sum of squares

$$\mathrm{SS} \;=\; \sum_{k=1}^{n}\left(X_k-\mu\right)^2 \;. \tag{14}$$

A simple calculation (reader: do it) shows that the minimizer is the *sample mean* (42).

Maximum likelihood estimation with Gaussian models often leads to least squares. This is true here, and also in linear regression. Least squares is a convenient computational procedure, and it's theoretical justification is Gaussian models. We come back this theme often.

The maximum likelihood estimate of the mean and variance comes from maximizing L or l over both μ and σ. Maximizing over μ leads to (42) regardless of σ. Maximizing over σ is the same as maximizing over $v = \sigma^2$. For this purpose the log likelihood may be written:

$$l(\mu,v) \;=\; \frac{-1}{2v}\,\mathrm{SS} \;-\; \frac{n}{2}\ln(v) \;+\; C\;,$$

where now C is independent of the data and both parameters. Setting $\partial_v l(\mu,v) = 0$ gives

$$\widehat{v}_{\mathrm{ml}} \;=\; \widehat{\sigma}^2_{\mathrm{ml}} \;=\; \frac{1}{n}\sum_{k=1}^{n}\left(\,X_k-\widehat{\mu}\,\right)^2 \;. \tag{15}$$

The formulas (42) and (15) maximize $L(\vec{X},\mu,\sigma)$ over the parameters σ and μ.

As the general theory predicts, the estimator (15) has bias of order $1/n$. It turns out (reader: this is simple algebra) that the estimator

$$\widehat{\sigma}^2_{\mathrm{ub}} \;=\; s^2 \;=\; \frac{1}{n-1}\sum_{k=1}^{n}\left(\,X_k-\widehat{\mu}\,\right)^2 \tag{16}$$

is *unbiased* in the sense that in the Gaussian model

$$E\left[\widehat{\sigma}^2_{\mathrm{ub}}\right] \;=\; \sigma^2\;.$$

Some people prefer the unbiased estimator (16) to the maximum likelihood estimator (15). It is a very rare situation where the difference matters.

1.4 Hypothesis testing, and Student's t distribution

Hypothesis testing is the problem of answering a yes/no question using statistical data. A positive answer takes the form:

$$\text{"Yes" with probability } 1-p,$$

where p is a small probability. The *confidence* of the statement is $1-p$. Traditional confidence levels are 95% and 99%.

A traditional hypothesis test has two *hypotheses*, the *null hypothesis* and the *alternative hypothesis*, denoted H_0 and H_1 respectively. For example, suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and only σ is known (unlikely). Say H_0 is the hypothesis $\mu = 0$ and H_1 is the hypothesis $\mu > 0$. Rejecting the null hypothesis means stating, with a satisfying level of confidence, that $\mu > 0$. If the data do not allow us to reject the null hypothesis, we may be reluctant to say that we accept H_0, only that the data or the test did not allow us to reject it. The *power* of a statistical hypothesis test is it's ability to reject the null hypothesis when H_0 is in fact false. Statistical tests are developed first by fixing a desired level of confidence, then trying to increase the power.

Suppose we have n independent normal samples with known variance but unknown mean, and we hope to state with high confidence that $\mu > 0$. The reader should check (reader: please) that under H_0

$$Z \;=\; \frac{1}{\sigma}\,\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k \tag{17}$$

has a standard normal distribution $Z \sim \mathcal{N}(0,1)$. If H_1 is true, this variable is likely to be significantly positive. If H_0 is true, it is not likely to be very far from zero. This is quantified using the *cumulative normal* distribution function $N(z) = \Pr[Z \le z]$. Because the Gaussian distribution is symmetric, for confidence level $1-p$, we may define the cutoff, z_p by

$$\Pr[Z \ge z_p] \;=\; N(-z_p) \;=\; p\,. \tag{18}$$

The 95% and 99% levels are $z_{.05} = 1.96$ and $z_{.01} = 2.8$. The statistical test is: compute the *z statistic* from the data using (17). If $Z \ge z_p$, say with confidence $1-p$ that $\mu > 0$. If $Z < z_p$, report that you failed to reject the hull hypothesis. This procedure – using the statistic (17) and the cutoff (18) – is called the Z *test*.

We would like to modify this procedure for the more likely situation of unknown μ and σ. A natural idea would be to replace σ with the estimate (16) $s = \widehat{\sigma} = \sqrt{\widehat{\sigma}^2_{\rm ub}}$ in (17). The resulting statistic (recall that any function of the data is a statistic) is called the *Student's*[3] t statistic:

$$t \;=\; \frac{1}{s}\,\frac{1}{\sqrt{n}}\,\widehat{\mu}\,. \tag{19}$$

[3]Student is a pseudonym for the early twentyeth century statistician William Gosset.

This defines t as a function of $\vec{X} = (X_1, \ldots, X_n)$, using (42) and (16). Student noticed that the probability density of t depends on n but not on the parameters μ and σ. More precisely, if the $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then the distribution of t does not depend on μ and σ.

This theorem of Student may be proven using the fact that if

$$X \;=\; \mu \;+\; \sigma Z \;, \tag{20}$$

then $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $Z \sim \mathcal{N}(0, 1)$. If $X_k = \mu + \sigma Z_k$ and the Z_k, then $\widehat{\mu}$ and s^2 (from (42) and (16)) both are independent of μ. Furthermore, we see that (reader: check this)

$$t \;=\; \sqrt{n-1}\,\frac{\overline{Z}}{\sqrt{\sum_{k=1}^{n} \left(Z_k - \overline{Z}\right)^2}} \tag{21}$$

This proves Student's theorem by expressing t in a way that is independent of μ and σ. It is easy to use this definition of t to find $t_{n,p}$ so that

$$\Pr\left[\, t \;>\; t_{n,p} \,\right] \;=\; 1 - p \;.$$

The numbers $t_{n,p}$ are tabulated in books and available in any decent statistical software. Hypothesis testing with Gaussians of unknown μ and σ is: compute $\widehat{\mu}$ from (42) and s from (16), then t from (19). If $t > t_{n,p}$ you have confidence $1 - p$ in asserting $\mu > 0$. Otherwise, you have nothing. This Student's $t-$test.

1.5 The distributions χ^2 and t

The t random variable (21), and the related χ^2 are interesting in their own right. The have many uses as examples in statistics. We begin with χ^2. The definition is that a χ^2 (pronounced "chi squared") random variable with n *degrees of freedom* has the distribution

$$\chi^2 \;=\; \sum_{k=1}^{n} Z_k^2 \;, \tag{22}$$

where the $Z_k \sim \mathcal{N}(0, 1)$ are independent standard normals. This definition resembles the denominator of (21), a resemblence that soon will get stronger. Although the definition (22) involves an n dimensional standard normal, χ^2 itself is one dimensional.

2 Gaussian random variables

Gaussian random variables have a remarkable combination of properties that anyone doing statistics must be aware of. Some of these are related to the *central limit theorem*. Others are related to the connection between linear algebra and

transformation of multi-variate normals. This section is a quick review of some properties of Gaussians.

It is easy to forget how special Gaussians are. Here is a list of warnings of mistakes that can be made in forgetting that other random variables are not like Gaussians

1. Maximum likelihood estimation is equivalent to least squares only for Gaussians.
2. Un-correlation implies independence only for Gaussians.
3. Non-gaussians can have much larger tail probabilities than Gaussians
4. The mean and variance describe the random variable completely only for Gaussians (and some other simple families).
5. Arbitrary linear transformations and combinations yield random variables in the same class mostly for Gaussians.
6. The Cholesky algorithm for generating correlated Gaussians applies only to Gaussians.

The univariate normal with mean μ and variance σ^2 has density (13). The multi-variate random variable $X = (X_1, \ldots, X_n)$ is *multi-variate normal* with mean $\mu = (\mu_1, \ldots, \mu_n)$ and covariance Σ (an $n \times n$ positive definite symmetric matrix) if its probability density is given by

$$f(x) \;=\; \frac{1}{(2\pi)^{n/2}} \, \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(\frac{-1}{2} x^t \Sigma^{-1} x \right) . \tag{23}$$

The notation for the multi-variate and univariate cases are slightly inconsistent. If we take $n = 1$ in the multi-variate formula above, the 1×1 matrix Σ would have as its single entry the number σ^2.

2.1 Limit Theorems

Suppose random variables Y_k are independent samples of a distribution f. Suppose they have mean μ and variance σ^2. Then the sample means converge to μ:

$$\overline{Y}_n \;=\; \frac{1}{n} \sum_{k=1}^{n} Y_k \;\to\; \mu \qquad \text{as} \quad n \;\to\; \infty \, .$$

This fact, called the *law of large numbers*, is easy or hard to prove depending on which version you take. The *weak* law of large numbers states that this limit is taken in probability, the kind of convergence we used in the definition of a consistent statistic. That is, for every $\epsilon > 0$ and $\delta > 0$ there should be an N so that if $n \geq N$ then

$$\Pr\left[\left| \overline{Y}_n - \mu \right| \;\geq \epsilon \right] \;\leq\; \delta \, . \tag{24}$$

The *strong law* of large numbers states that with probability one (called *almost surely* in probability)

$$\lim_{n\to\infty} \overline{Y}_n \;\to\; \mu \qquad \text{as } n \;\to\; \infty \,. \tag{25}$$

That is, there is zero probability that the limit does not exist or is not equal to μ is zero. The theorem that (25) holds with probability one under the hypothesis that $E\,[|Y|] < \infty$ is the *Kolmogorov* strong law. It is one of the hardest theorems in the first year PhD level probability class at Courant.

The χ^2 distribution (22) is a good example of the use of the law of large numbers. Take $Y_k = Z_k^2$ (so that $\mu = E\,[Y_k] = 1$), and get

$$\frac{1}{n}\,\chi^2 \;=\; \frac{1}{n}\left(\sum_{k=1}^{n} Y_k\right) \to 1 \quad \text{as } n \to \infty \,.$$

This says that the χ^2 random variable with n degrees of freedom is, to a first approximation, equal to n. We will see shortly that this simplifies the t statistic (21).

The error in the law of large numbers is given by the *Central Limit Theorem* (capitalized by tradition). A simple calculation shows that

$$\mathrm{var}\big[\,\overline{Y}_n - \mu\,\big] \;=\; \frac{\sigma_Y^2}{n} \,,$$

where $\sigma_Y^2 = \mathrm{var}[Y]$. This suggests that $\overline{Y}_n$ is on the order of $1/\sqrt{n}$, so that

$$R_n \;=\; \sqrt{n}\,\left(\,\overline{Y}_n \,-\, \mu\,\right) \tag{26}$$

has some kind of limit as $n \to \infty$. The numbers R_n themselves do not converge[4], but the distribution of the random variable R_n converges to a Gaussian with mean zero and variance σ_Y^2. A Basic Probability class has many technical statements of this convergence theorem under different technical hypotheses. One simple one is that, for any a, ($Z \sim \mathcal{N}(0,1)$ is a standard normal, $\sigma_Y Z \sim \mathcal{N}(0, \sigma_Y^2)$)

$$\Pr[\,R_n \,\leq\, a\,] \;\to\; \Pr[\,\sigma_Y Z \,\leq\, a\,] \;=\; N(\sigma_Y a) \qquad \text{as } n \to \infty \,. \tag{27}$$

Warning: the accuracy of the approximation $\Pr\,[R_n \leq a] \approx N(\sigma_Y a)$ depends on a as well as n. In particular, the tails of R_n may be much larger than the normal tails (tails of the Gaussian). See the exercises for some examples of this. It is not safe to use the Central Limit Theorem approximation to estimate the probabilities of rare events "in the tails" unless n is very large.

The limit theorems hold also for multivariate random variables. The law of large numbers is exactly as (24) and (25). The only difference is that Y and μ

[4]Stochastic Calculus studies the relation between different numbers R_n for large n. The limit is related to Brownian motion.

are vectors with several components, and $\left|\overline{Y}_n - \mu\right|$ is the norm of the vector in a vector space of the appropriate dimension.

The Central Limit Theorem also holds for multivariate random variables. Use the following terminology: Y is a multivariate random variable with m components $(Y_1, \ldots, Y_m)$. The covariance matrix of Y, here denoted Σ_Y, has entries

$$\sigma_{\alpha\beta} \;=\; \mathrm{cov}\,[Y_\alpha, Y_\beta] \;=\; E\,[(Y_\alpha - \mu_\alpha)(Y_\alpha - \mu_\alpha)] \;.$$

This definition applies also to the diagonal entries $\sigma_{\alpha\alpha} = \mathrm{var}\,[Y_\alpha]$. Let $f(y)$, for $y \in R^n$, be the probability density of Y. Let $g_n(r)$ be the probability density for the m component random variable R_n defined by (26). Let $g(r)$ be the multivariate normal with mean zero and covariance Σ_Y

$$g(r) \;=\; \frac{1}{(2\pi)^{m/2}}\,\frac{1}{\sqrt{\det\,(\Sigma_Y)}}\,\exp\left(-\,r^t\Sigma_Y^{-1}r/2\right) \;.$$

The multi-variate Central Limit Theorem states that $g_n(r) \to g(r)$ as $n \to \infty$. For example, if $A \subseteq R^m$ is an open set, then

$$\lim_{n\to\infty} \Pr[\,R_n \;\in\; A\,] \;=\; \int_A g(r)\,dr \;.$$

The warning above about the univariate Central Limit Theorem applies with at least as much force here.

The multi-variate central limit theorem has many interesting applications. For example, suppose $U \in [-1, 1]$ is uniformly distributed, and consider the two component random variable $Y = (U, U^3)$. The covariance matrix of this Y is easy to calculate:

$$\begin{aligned}
\sigma_{11} \;&=\; \mathrm{var}\,[U] \;=\; E\left[U^2\right] \;=\; \frac{1}{2}\int_{-1}^{1} u^2\,du \;=\; \frac{1}{3} \\
\sigma_{12} \;&=\; \mathrm{cov}\left[U, U^3\right] \;=\; E\left[U^4\right] \;=\; \frac{1}{5} \\
\sigma_{22} \;&=\; \mathrm{var}\left[U^3\right] \;=\; E\left[U^6\right] \;=\; \frac{1}{7} \;.
\end{aligned}$$

2.2 Properties of the normal distribution

Gaussian random variables have many distinctive properties. Many of these may be understood as consequences of the Central Limit Theorem. The distribution of a Gsussian random variable is completely determined by the mean and variance (covariance matrix if multi-component). Therefore, any formula about a Gaussian random variable must depend only on the mean and variance. In particular, if the mean is zero, it must depend only on the variance.

For univariate normals, many formulas follow from the representation $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0, 1)$ is a standard normal. A normal $X \sim \mathcal{N}(\mu, \sigma^2)$ may be represented in this way. Such a representation also is possible for multivariate normals using the *Cholesky* factorization of Σ, see below.

For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$, there is a formula of the form

$$E\big[\, e^X \,\big] \;=\; L(\mu, \sigma) \;.$$

We know this because whatever the left side is, it can depend only on the parameters μ and σ. Representing X in terms of a standard normal, we have

$$L(\mu, \sigma) \;=\; E\big[\, e^{\mu} e^{\sigma Z} \,\big] \;=\; e^{\mu}\, E\big[\, e^{\sigma Z} \,\big] \;=\; \frac{e^{\mu}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\sigma z} e^{-z^2/2}\, dz \;.$$

The integral on the right is done by completing the square in the exponent: $\sigma z - z^2/2 = \sigma^2/2 - (z-\sigma)^2/2$. Altogether (reader: check this)

$$L(\mu, \sigma) \;=\; e^{\mu + \sigma^2/2} \;. \tag{28}$$

This formula also may be found as a direct consequence of the Central Limit Theorem, done here (without loss of generality) for $\mu = 0$. Suppose Y is a univariate random variable with $E[Y] = 0$ and $\mathrm{var}[Y] = \sigma^2$. For example, we could take $Y \sim \mathcal{N}(0, \sigma^2)$. Then $X_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k$ (the Y_k being independent samples of Y) approximately (or possibly exactly) has the distribution $\mathcal{N}(0, \sigma^2)$. Therefore,

$$L(0, \sigma) \;=\; \lim_{n \to \infty} E\left[\exp\left(\sum_{k=1}^{n} Y_k / \sqrt{n} \right) \right] \;.$$

The trick is to notice that the random variables $Y_k/\sqrt{n}$ are independent, so the exponential of the sum is the product of the exponentials. Moreover, the $Y_k/\sqrt{n}$ have the same distribution, so the numbers $E\left[\exp(Y_k/\sqrt{n})\right]$ are all the same. Therefore,

$$L(0, \sigma) \;=\; \lim_{n \to \infty} \left(E\big[\, \exp\left(Y_k/\sqrt{n}\right) \,\big] \right)^n \tag{29}$$

Since $Y_k/\sqrt{n}$ is small, we use the Taylor series expansion

$$\exp(Y_k/\sqrt{n}) \;=\; 1 \;+\; \frac{1}{\sqrt{n}} Y_k \;+\; \frac{1}{2n} Y_k^2 \;+\; O(n^{-3/2}) \;,$$

and

$$E\left[\exp(Y_k/\sqrt{n})\right] \;=\; 1 \;+\; \frac{\sigma^2}{2n} \;+\; O(n^{-3/2}) \;=\; e^{\sigma^2/2n \,+\, O(n^{-3/2})} \;.$$

Putting this back into (29) gives (after some simplification)

$$L(0, \sigma) \;=\; \lim_{n \to \infty} e^{\sigma^2/2 \,+\, O(n^{1/2})} \;=\; e^{\sigma^2/2} \;,$$

as before. This is not a simpler way to derive (28), but it is more fundamental. It explains (28) as a direct consequence of the Central Limit Theorem.

2.3 Linear changes of variables, χ^2, and t, again

Suppose A is an $m \times n$ matrix, so that $Ax \in R^m$ if $x \in R^n$. If $Y = AX + b$ and X is multivariate normal, then Y also is multivariate normal. This theorem has to be interpreted carefully if the row rank of A is less than m (i.e. if $m > n$, or $m = n$ and A is singular, or $m < n$ and $\text{rank}(A) < m$), because in that case the set of all possible Y values forms a proper subspace of R^m, and Y has no proper probability density. We ignore this case at least for a while, and probably forever. The fact that AX is Gaussian when X is Gaussian may be calculated by integration and linear algebra, but it also is obvious from the Central Limit Theorem. If X is an average of a large number of i.i.d. (independent and identically distributed) random variables, then AX also is such an average. Linear algebra is an essential part of all Gaussian analysis.

One application is to the χ^2 random variable. Suppose $Z_1, \ldots, Z_n$ are n independent standard normals, and consider the random variable

$$X \;=\; \sum_{k=1}^{n} \left(Z_k - \overline{Z} \right)^2 \quad , \qquad \text{where} \quad \overline{Z} \;=\; \frac{1}{n} \sum_{j=1}^{n} Z_j \; . \tag{30}$$

This is the denominator of (21). Linear algebra helps us determine the distribution of X. Let $M \subset R^n$ be the linear subspace of all vectors $z \in R^n$ with mean zero: $\sum_{j=1}^{n} z_j = 0$. There is an orthogonal projection P_M that projects any $z \in R^n$ to the closest point in M. If $w = P_M z$ then w minimizes

$$\|w - z\|_2^2 \;=\; \sum_{j=1}^{n} (z_j - w_j)^2$$

over all $w \in M$. It is "easy to see" (i.e. you should be able to check, with the background you have, in less than an hour) using the method of Lagrange multipliers, that the w that minimizes $\|w - z\|_2^2$ with the constraint $\sum_{j=1}^{n} w_j = \mathbf{1}^t w = 0$ is given by

$$w_j \;=\; z_j \;-\; \overline{z} \quad , \qquad \text{where} \quad \overline{z} \;=\; \frac{1}{n} \sum_{j=1}^{n} z_j \; .$$

There, (30) is equivalent to

$$X \;=\; \|P_M Z\|_2^2 \; . \tag{31}$$

So, the distribution of X depends on the distribution of $\|W\|_2 = \|P_M Z\|_2$, when $Z \in R^n$ is a normal with mean zero and covariance $\Sigma_Z = I$, and P_M is orthogonal projection onto the subspace M. We can figure that out using an orthonormal basis of R^n so that the first m vectors in the basis are an orthonormal basis of M ($m < n$ is the dimension of M, which need not be $n-1$ for this part of the argument). If these vectors are v_k, then

$$\|P_M Z\|_2^2 \;=\; \sum_{j=1}^{m} \left(v_k^t Z \right)^2 \; .$$

Now, the numbers $Y_k = v_k^t Z$ are Gaussians with mean zero, variance one and all covariances equal to zero. Therefore,

$$\|P_M Z\|_2^2 \;=\; \sum_{j=1}^{m} Y_k^2 \;,$$

which is the sum of squares of m independent standard normals. The conclusion is that the random variable X in (30) has the same distribution as the sum of $n-1$ squares of independent standard normals. That is, $X \sim \chi^2_{n-1}$.

Return to the t random variable (21). We have understood the denominator. But our analysis also shows that the numerator, $\overline{Z}$ is independent of the denominator. Indeed, the numerator (in the terminology of the previous paragraph) is a multiple of $v_n^t Z$, which is a Gaussian independent of the others. The conclusion is that the random variable (21) has the same distribution as

$$t \;=\; \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \;, \tag{32}$$

where Z is a univariate standard normal. We put $\chi^2_{n-1}/(n-1)$ in the denominator to emphasize that it is a random variable with mean 1, for any n. The number of terms in the χ^2 distribution (22) is called the number of *degrees of freedom*. The random variable that arises in the t test (19) involves a χ^2 with $n-1$ degrees of freedom when there are n data points. This is because one of the n degrees of freedom is lost when we use the sample mean instead of the true mean to estimate σ^2.

2.4 Fat tailed t

Besides its use in hypothesis testing, the t random variable (32) often is used simply as an example of a random variable with *fat tails*. For any random variable, X, the *tails* refer to the probability of large positive or negative values. The term is a little vague and is used in related but not entirely consistent ways. For example, X is said to have *power law* tails if $\Pr[|X| > r] \sim r^{-p}$ for large x. A number x is in the tails of a random variable distribution if $\Pr(X > x)$ is small and if its distribution is well described by some simple large x approximation at that point.

A random variable has *fat tails*, or *heavy tails*, if its tail probabilities are larger than are convenient. How big is considered fat depends on the observer and the application. Gaussian and exponential random variables have tails that are exponentially thin: $\Pr(X > x) < e^{-Cx^2}$ (different C for different Gaussians), or $\Pr(X > x) < e^{-Cx}$ (for exponentials). The χ^2 random variables also have thin tails. Some people consider any power law tails to be fat, particularly if the power is not very big. Some think a tail is not fat until $\mathrm{var}[X] = \infty$.

Many random variables in life have fat tails: the sizes of earthquakes, wealths of rich people, the sizes of requests to web sites, and sized of insurance claims are among the many examples. Daily stock market returns are observed to have

power law tails with[5] $p \approx 3$. Gaussian models often are adequate for market returns in the central regions (not the tails), but are terrible models for the tails.

The t random variable has power law tails, with the power $p = n - 1$. It sometimes is used to model returns or other random variables that are expected to have fat tails. There seems to be no quantitative argument for this, nothing like the Central Limit Theorem argument for Gaussians. The power law may be understood directly from formulas for the probability density of a t random variable (see e.g. Chapter 1 of Attilio Meucci's book), but here we give a less formal argument.

By far the most likely way to have a large t in (32) is to have a small χ^2, Z having exponentially thin tails. We figure out this probability from the fact (see (22)), that if $R \sim \sqrt{\chi^2_n}$, then ($f(\vec{z})$ being the probability density of $\vec{Z} = (Z_1, \ldots, Z_n)^t$)

$$\begin{aligned} \Pr[R < r] &= \Pr\left[\sqrt{Z_1^2 + \cdots + Z_n^2} < r\right] \\ &= \int_{|\vec{z}|<r} f(\vec{z})\, dz \\ &= C_n \int_{s=0}^{r} e^{-s^2/2}\, s^{n-1}\, ds \\ &\sim C'_n r^n \qquad \text{for small } r. \end{aligned}$$

For the tail probabilities for t, we have (note that (32) has $n-1$ degrees of freedom)

$$\Pr[\, t > t_0\,] \;\sim\; \Pr[\, R \leq 1/t_0\,] \;\sim\; C'_n t_0^{-(n-1)}\,. \tag{33}$$

This is the tail behavior of the t random variable. The number n appears as a parameter in the formulas for the probability densities of t and χ^2. This allows them to be defined when n is not an integer. This would be silly for hypothesis testing, but it is useful if we are using t as a generic fat tailed random variable. It allows us to have power law tails where the power is not an integer.

3 Errors in parameter estimation

There are several ways to assess the accuracy of a statistical estimate of something. The mean square error (3) is one of the simplest, both in theory and in practice.

[5]The power $p \approx 3$ is empirical, and is discussed at length in the interesting book *Introduction to Econophysics: Correlations & Complexity in Finance*, by well known theoretical physicists R. N. Mantegna and H. E. Stanley.

3.1 Maximum likelihood

There is a simple error analysis for maximum likelihood estimation that involves the *Fisher information*

$$I(\theta) \;=\; E_\theta\left[\left(\frac{\partial_\theta f(X,\theta)}{f(X,\theta)}\right)^2\right] \;. \tag{34}$$

This is a measure of how strongly the distribution of X depends on θ. Clearly, it is impossible to estimate θ from observations of X if the distribution of X does not depend on θ. The quantity in (34) involves the derivative with respect to θ of the log likelihood: $\partial_\theta \log(f(x,\theta))$. The result is that if $\widehat{\theta}_n$ is the maximum estimate and θ is the true parameter value, then, for large n:

$$\mathrm{var}\left[\widehat{\theta}_n\right] \;\approx\; \frac{1}{n}\,\frac{1}{I(\theta)} \;\; . \tag{35}$$

A clever argument, called the *Cramer Rao* bound, shows that any statistical estimator of θ (except trivial ones) has mean square error at least as large as (35), approximately. This is described in any good statistics book, but not here.

The derivation of (35) helps us to understand many other parameter estimation methods. We adopt the Bayesian practice of distinguishing between different functions by the names and number of arguments, starting with the expected log likelihood function

$$l(\theta) \;=\; E_{\theta_0}[\,\ln(f(X,\theta))\,] \;=\; \int \ln(f(x,\theta))f(x,\theta_0)\,dx \;.$$

Here, and for the next few paragraphs, we call the exact parameter value θ_0. The derivative of l with respect to θ, with θ_0 fixed, is

$$m(\theta) \;=\; \partial_\theta\, E_{\theta_0}[\,\ln(f(X,\theta))\,] \;=\; \int \frac{\partial_\theta f(x,\theta)}{f(x,\theta)} f(x,\theta_0)\,dx \;.$$

In particular, since $\int f(x,\theta)dx = 1$ for all θ, we have[6]

$$m(\theta_0) \;=\; \int \partial_\theta f(x,\theta)\,dx \;=\; 0 \;,$$

and (with some algebra)

$$m'(\theta_0) \;=\; -\int \left(\frac{\partial_\theta f(x,\theta_0)}{f(x,\theta_0)}\right)^2 f(x,\theta_0)\,dx \;=\; -\,E_{\theta_0}\left[\left(\frac{\partial_\theta f(X,\theta_0)}{f(X,\theta_0)}\right)^2\right] \;.$$

This is the negative of the Fisher information (34). This shows that θ_0 is at least a local maximizer of $l(\theta)$, since $l''(\theta_0) < 0$.

[6] This has an important interpretation in terms of information and relative entropy that we skip over for now.

With this, we can understand the difference between θ_0 and $\widehat{\theta}$ using the Central Limit Theorem and Taylor series. We find $\widehat{\theta}$ by maximizing the log likelihood with respect to θ. This is the same as setting the derivative of the log likelihood function to zero. In formulas, (writing $M_k = \partial_\theta \ln(f(X_k, \theta))$)

$$0 \;=\; \partial_\theta \sum_{k=1}^{n} l(X_k, \widehat{\theta}) \;=\; \sum_{k=1}^{n} M_k(\widehat{\theta}) \; .$$

Now (and this is the main point), we already have calculated that $E_{\theta_0}[M_k(\theta)] = m(\theta)$ and, in particular, that

$$E_{\theta_0}[\, M_k(\theta_0)\,] \;=\; 0 \; .$$

Also, (this seems like a coincidence, but it is not really)

$$\mathrm{var}[\, M_k(\theta_0)\,] \;=\; E_{\theta_0}\big[\, M_k(\theta_0)^2\,\big] \;=\; E_{\theta_0}\left[\left(\frac{\partial_\theta f(X,\theta_0)}{f(x,\theta_0)}\right)^2\right] \;=\; I(\theta_0) \; .$$

Putting these facts together,

$$0 \;=\; \sum_{k=1}^{n} M_k(\widehat{\theta}) \;\approx\; \sum_{k=1}^{n} M_k(\theta_0) \;+\; \left(\sum_{k=1}^{n} M_k'(\theta_0)\right)\Big(\,\widehat{\theta} - \theta_0\,\Big) \; .$$

Therefore

$$\widehat{\theta} - \theta_0 \;\approx\; \frac{\dfrac{-1}{n}\displaystyle\sum_{k=1}^{n} M_k(\theta_0)}{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} M_k'(\theta_0)} \; ,$$

The numerator is approximately (Central Limit Theorem) a normal with mean zero and variance $\frac{1}{n} I(\theta_0)$. The denominator is approximately (law of large numbers) $I(\theta_0)$. Therefore, $\widehat{\theta} - \theta_0$ is approximately normal with mean zero, and variance approximately

$$\frac{1}{n}\,\frac{I(\theta_0)}{I(\theta_0)^2} \;=\; \frac{1}{n}\,\frac{1}{I(\theta_0)} \; ,$$

as claimed. A longer calculation of this kind shows that

$$\mathrm{bias}\Big(\,\widehat{\theta}_n\,\Big) \;=\; E_{\theta_0}\Big[\,\widehat{\theta}_n \;-\; \theta_0\,\Big] \;=\; O\Big(\frac{1}{n}\Big) \; .$$

The error coming from the variance is much larger than the error coming from the bias.

The multivariate case $\theta = (\theta_1, \ldots, \theta_m)$ is just about the same (see the GMM section below). Calculations just like those above show

$$\mathrm{cov}\left[\widehat{\theta}\right] \;\approx\; \frac{1}{n}\, I^{-1}(\theta) \; . \tag{36}$$

Here I is the information matrix whose entries are

$$I_{jk} \;=\; E_\theta\left[\frac{\partial_{\theta_j} f(X,\theta)\partial_{\theta_k} f(X,\theta)}{f(X,\theta)^2}\right] \; .$$

3.2 Maximum likelihood and multivariate normals

Suppose that $X \in R^n$ is multivariate normal and that we have independent observations X_k, for $k = 1, \ldots, n$. The maximum likelihood estimate of the mean is the sample mean as before

$$\widehat{\mu}_{\text{ml}} \;=\; \overline{X} \;=\; \frac{1}{n} \sum_{k=1}^{n} X_k \;. \tag{37}$$

The maximum likelihood estimate of the covariance matrix is the sample covariance

$$\widehat{\Sigma}_{\text{ml}} \;=\; \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \overline{X}\right) \left(X_k - \overline{X}\right)^t \quad . \tag{38}$$

We give a derivation of these formulas, mostly as practice doing matrix algebra. What we called θ in the abstract theory is now the mean and covariance:

$$\theta \;=\; (\mu, \Sigma) \;.$$

Let d be the number of components of X (as opposed to n, the number of samples). The number of parameters in θ is d for the components of μ, and $d(d+1)/2$ for the unknown entries of the symmetric matrix Σ. There are n^2 entries in Σ, but the entries below the diagonal are the same as the corresponding entries above. Altogether, there are $d(d+3)/2$ unknown parameters. This is a very large number, if d is large. Estimating a large number of parameters either is unreliable or requires a very large amount of data. This makes estimating general Gaussian models problematic in many component situations.

For the multivariate normal, the probability density is

$$f(x, \theta) \;=\; f(x, \mu, \Sigma) \;=\; \frac{1}{(2\pi)^{d/2}} \, \frac{1}{\sqrt{\det(\Sigma)}} \, \exp\left(\frac{-1}{2} \, (x-\mu)^t \, \Sigma^{-1} \, (x-\mu) \right) \;.$$

Taking the log gives (C is independent of the data and the parameters.)

$$l(x, \mu, \Sigma) \;=\; C \;-\; \frac{1}{2} \ln\left(\det(\Sigma)\right) \;-\; \frac{1}{2} \, (x-\mu)^t \, \Sigma^{-1} \, (x-\mu) \;. \tag{39}$$

We find $\widehat{\mu}$ by minimizing over μ. The derivative is

$$m(x, \mu, \Sigma) \;=\; \nabla_\mu l(x, \mu, \Sigma) \;=\; \Sigma^{-1} \, (x-\mu) \;.$$

As long as Σ is non-signular,

$$\sum_{k=1}^{n} m(X_k, \widehat{\mu}, \Sigma) \;=\; 0 \quad \Longrightarrow \quad \sum_{k=1}^{n} X_k \;-\; n\widehat{\mu} \;=\; 0 \;,$$

which is the formula (37).

Differentiating with respect to Σ is more complicated. You have to do it carefully or the mess of indices will be overwhelming. The method of *virtual*

perturbations is one convenient approach. Suppose $G(A)$ is a function of a symmetric matrix, A. Suppose $\dot{A}$ is another symmetric matrix, which we call the virtual perturbation. The corresponding virtual perturbation of G is

$$\dot{G}(A, \dot{A}) \;=\; \lim_{h \to 0} \frac{1}{h} \Big(G(A + h\dot{A}) - G(A) \Big) \;.$$

This is just the directional derivative of G in the direction $\dot{A}$. We can maximize a scalar function of A by setting $\dot{G} = 0$ for all $\dot{A}$. The method works for symmetric or non-symmetric matrices, but the application we have calls for symmetric.

The virtual perturbation method makes is easy to differentiate using the product rule and implicit differentiation. The product rule is clear:

$$\dot{(FG)} \;=\; \dot{F}G \,+\, F\dot{G} \;.$$

This is true even if G and H are matrices or vectors, as the reader should check. We compute the derivative of the matrix inverse function as an example of the technique in action. Write $A^{-1} = B$ in the form $AB = I$. Then differentiate, use the product rule, and note that $\dot{I} = 0$:

$$\dot{A}B \,+\, A\dot{B} \;=\; 0 \quad \Longrightarrow \quad \dot{B} \;=\; -A^{-1}\dot{A}B \;=\; -A^{-1}\dot{A}A^{-1} \;.$$

This is the matrix version of the formula

$$\frac{d}{dt}x^{-1} = x^{-2}\frac{dx}{dt} \;,$$

but in the matrix version, the $\dot{x}$ term goes between the two x^{-1} terms.

The target application (39) asks us to differentiate $\det(\Sigma)$. For this there is the elegant formula (the left side is $\dot{G}$, where $G(A) = \ln(\det(A))$.)

$$\dot{\ln\left(\det(A)\right)} \;=\; \mathrm{Tr}\left(A^{-1}\dot{A}\right) \;. \tag{40}$$

On the right side is the *trace*: $\mathrm{Tr}(B) = \sum_{k=1}^{d} B_{kk}$. The proof of (40) is below.

We now have the tools to differentiate (39):

$$\dot{l} \;=\; -\frac{1}{2}\mathrm{Tr}\Big(\Sigma^{-1}\dot{\Sigma}\Big) \,+\, \frac{1}{2}\,(x-\mu)^t\,\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}\,(x-\mu) \;.$$

Now again we assume that Σ is invertible, so we can use the change of variables $\Sigma^{-1}\dot{\Sigma}\Sigma^{-1} = \dot{M}$, so that $\dot{\Sigma} = \Sigma\dot{M}\Sigma$, and allowing all possible $\dot{M}$ is the same as allowing all possible Σ. With this,

$$\dot{l} \;=\; -\frac{1}{2}\mathrm{Tr}\Big(\dot{M}\Sigma\Big) \,+\, \frac{1}{2}\,(x-\mu)^t\,\dot{M}\,(x-\mu) \;.$$

We use this to get information about Σ by taking $\dot{M}$ in a special way. In particular, take $\dot{M}$ to be non-zero only in entries[7] (α, β) and (β, α), with $\dot{M}_{\alpha\beta} =$

[7] We use Greek letters here to distinguish from Latin letters what index the sample. In general, Greek letters from from 1 to d, while Latin letters run from 1 to n.

$\dot{M}_{\beta\alpha} = 1$. In view of the formula

$$\mathrm{Tr}\,(AB) \;=\; \sum_{\alpha=1}^{d}\sum_{k=0}^{d} A_\beta B_{\beta\alpha} \;=\; \sum_{j=1}^{d}\sum_{k=0}^{d} A_{\alpha\beta} B_{\alpha\beta}\ ,$$

(the last only if B is symmetric) we have

$$\dot{l} \;=\; -\sigma_{\alpha\beta} \,+\, (x_\alpha - \mu_\alpha)(x_\beta - \mu_\beta)\ .$$

The factors of $\frac{1}{2}$ disappeared in case because there were two equal terms, one from (α, β) and the other from (β, α). It would help the reader to check that this formula also is true in the case $\alpha = \beta$.

3.3 GMM, Generalized Method of Moments

This is less systematic than maximum likelihood but simpler to apply in many situations. Suppose you have parameters $\theta = (\theta_1, \ldots, \theta_m)$ and a family of probability distributions depending on θ. You also have n independent samples of the θ distribution and want an estimate, $\widehat{\theta}$. The idea is to have m (or more, see below) functions $g_1(x), \ldots, g_m(x)$ whose moment functions are known:

$$M(\theta) \;=\; E_\theta\left[g(X)\right]\ . \tag{41}$$

These are *moments* if $g_k(x) = x^k$, and *generalized moments* otherwise. The *generalized method of moments*, or *GMM*, is to estimate

$$\widehat{M}_k \;=\; \frac{1}{n}\sum_{j=1}^{n} g(X_j)\ , \tag{42}$$

then choose $\widehat{\theta}$ by solving the equations

$$M\Big(\widehat{\theta}\Big) \;=\; \widehat{M}\ . \tag{43}$$

We can estimate the error in the GMM using the ideas we just used for maximum likelihood. In fact, maximum likelihood may be viewed[8] as a special case of the GMM. For large n, $\widehat{M} = M + \frac{1}{\sqrt{n}}X$, where X is a mean zero finite variance (i.e. variance independent of n) random variable that is approximately normal. We then have, as before

$$M'(\theta)\left(\widehat{\theta} - \theta\right) \;\approx\; \frac{1}{\sqrt{n}}X \qquad\Longrightarrow\qquad \widehat{\theta} \;\approx\; \theta \,+\, \frac{1}{\sqrt{n}}\left(M'(\theta)\right)^{-1} X\ .$$

[8]This does not make GMM better than maximum likelihood. In fact, we showed that for large n, maximum likelihood achieves the Cramer Rao bound. Therefore, the best another GMM method (based on other moments than l) could do is to be as good as maximum likelihood.

This implies that the estimation error is approximately normal for large n (as most things are in theory), and that the error is determined by $(M'(\theta))^{-1}$ and the covariance matrix for X. Let us call this matrix $G = \mathrm{cov}_\theta\,[g(X)]$. Then

$$\mathrm{cov}\Big[\,\widehat{\theta}\,\Big] \;\approx\; (M'(\theta))^{-1}\, G(\theta)\, (M'(\theta))^{-t} \;. \tag{44}$$

Here we use write A^{-t} for the transpose of the inverse of A, which is the same as the inverse of the transpose of A (that is, $A^{-t} = \left(A^{-1}\right)^t = (A^t)^{-1}$).

The maximum likelihood case has two special properties that make its formula (36) simpler than the more general (44). One is that the m moment functions $M(\theta)$ are derivatives with respect to the components of θ, so the entries of $m \times m$ martix M' are second derivatives. This makes the matrix M' symmetric, which is why we did not see the transpose before (44). The other reason is that $(M')^{-1}\, G = I$, which seems to be a coincidence related to the log function.

An example for the method of moments is a Gaussian *mixture* model. This assumes that there are two Gaussians and a probability $p \in (0,1)$. You first choose one of the Gaussians with probability p, then choose X using that Gaussian. For simplicity, suppose both normals have mean zero, so we only have variances σ_1^2 and σ_2^2. Then, with probability p we take $X = \sigma_1 Z$ and with probability $1-p$ we take $X = \sigma_2 Z$. Here, $Z \sim \mathcal{N}(0,1)$ is a standard normal. The probability density for X is

$$f(x) \;=\; p\,\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-x^2/2\sigma_1^2} \;+\; (1-p)\,\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-x^2/2\sigma_2^2} \;.$$

The three parameters are σ_1^2, σ_2^2, and p. It is not easy to estimate them by maximum likelihood, but it is easier to use moments.