Modeling Income and Wealth Distributions II — Fitting and Checking

36-313, Fall 2021

7 September 2021 (Lecture 4)

Contents

In this lecture, we'll look at how we actually fit models, like the log-normal or Pareto, to data, at some ways of checking the fit. We'll hold off on talking (much) about processes which produce distributions like these until a later date.

1 Fitting

There are two big ways of fitting probability distributions to data. One is to try to match some features or aspects of the data: pick some summary statistics, calculate them on the data, and then tune the parameters until they match what we see. The other leading type of estimator is the method of maximum likelihood, where we use the model to calculate the probability of the data set we got at various possible parameters, and adjust the parameters until the data is as "likely" as possible. Matching summary features is older, easier to understand, and widely applicable, but maximum likelihood (usually) gives better estimates, when we can use it. We'll start with feature matching.

1.1 Common Notation

We have a probability model which specifies the probability density $f(x;\theta)$ as a function of the parameter θ. θ will usually be a vector, e.g., for the log-normal distribution $\theta = (\mu, \sigma^2)$ with $\mu = \mathbb{E}(logX)$ and $\sigma^2 = \mathrm{Var}(\log X)$. We'll write d for the number of dimensions or coordinates in θ. ($d = 2$ for the log-normal.) We'll need a symbol to distinguish the correct value of θ from the others; I'll write θ^*.

Our data set consists of measurements of X for n members of the population, which we'll imaginative write as $X_1, X_2, \ldots X_n$. (The capital letters here are reminder that the data are random.) If we need to abbreviate we'll write $X_{1:n}$ for the whole data, $X_{1:3}$ for the first 3 data points, etc.

We'll assume that every individual in the data is statistically independent of every other, so

$$f(x_{1:n};\theta) = \prod_{i=1}^{n} f(xi;\theta) \tag{1}$$

(Independence isn't strictly necessary but it simplifies our math in some places.)

Our *main* goal is to go from $X_1, \ldots X_n$ to an estimate of the true θ, say $\hat{\theta}$. We should really write this as $\hat{\theta}(X_{1:n}$ because it's a function of the data, but we'll suppress that to keep the notation under control. Because the data are random, the estimate $\hat{\theta}$ is also random. But we'd like it to be *close* to θ^*, and we'd like it to get closer as we get more data — we want $\hat{\theta}$ to **converge on** θ^*, $\hat{\theta}(X_{1:n}) \to \theta^*$ as $n \to \infty$. Estimators which converge on the truth are called **consistent**[1].

[1]The name makes more sense historically; see [http://bactra.org/notebooks/consistency-pac.html].

2 Fitting by Matching Features or Summary Statistics

In this approach, we don't work directly with the individual observations, but only with selected **features** or **statistics** which we have calculated from the data (or have had handed to us). These features or statistics are *functions* of the data, and they're supposed to summarize important and informative aspects of it, while being smaller and easier to understand than the full data.

In algebra, we have m statistics $G_1, G_2, \ldots G_m$, each defined as a function of the data:

$$G_k = g_k(X_{1:n}), \; k \in 1:m \tag{2}$$

(Notice that because the data are random, the values of the summary statistics are also random variables, hence the capital letters G_k, while the *functions* used to calculate them are fixed, hence the lower-case g_k.) We can also think of a single m-dimensional vector-valued statistic, $G = g(X_{1:n})$.

Examples of summary statistics include, but are not limited to:

- **Moments**, such as the sample mean $\overline{x}$, the sample mean square $\overline{x^2}$, up to the kth moment $\overline{x^k}$ and beyond
- **Central moments** such as the sample variance $\overline{(x-\overline{x})^2}$ (or the square root of this, the sample standard deviation) or the kth central moment $\overline{(x-\overline{x})^k}$
- **Percentiles** or other quantiles, including the median
- Ratios of or differences of any of these
- etc., etc., etc.: anything we can calculate from the data can, in principle, be used as a summary statistic, e.g., the Gini index.

We have seen all of these as ways of just describing the data.

So far, these summary statistics are *just* functions of the data, and don't involve any probability model at all. But, once we have a probability model, we can work out what these quantities *ought* to be, for the whole population or the distribution. At its most basic, if we consider the sample mean

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

the corresponding theoretical quantity is the expected value,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x;\theta) dx \tag{4}$$

Notice that the right-hand side of that last equation, once we've gone and done the integral, is going to be a function of the parameter θ. For instance, for the log-normal distribution, $\mathbb{E}(X) = e^{\mu+\sigma^2/2}$.

All of the kinds of summary statistics I mentioned have theoretical, model-based counter-parts for the whole distribution. Once we have (say) a model for income distribution, we can use it to calculate what (say) the ratio between the median income and the 99th percentile of income *should* be, in the whole population, as a function of θ.

> *For every choice of summary statistics $g(X_1, \ldots X_n)$ and probability model $f(x;\theta)$, there is a corresponding theoretical value $\gamma(\theta)$*

(Notice I am using the Greek letter γ, which is the counter-part to the Roman letter g, to stand for the theoretical quantity.)

Here is the key idea of the feature-matching approach to estimating parameters:

> Adjust the parameter θ so that the theoretical values $\gamma_1(\theta), \ldots \gamma_m(\theta)$ all match the observed summary statistics $g_1(X_{1:n}), \ldots g_m(X_{1:n})$

One way to think of this is to set up a system of equations:

$$\gamma_1(\theta) = G_1 = g_1(X_1, \ldots X_n) \quad (5)$$
$$\gamma_2(\theta) = G_2 = g_2(X_1, \ldots X_n) \quad (6)$$
$$\vdots \quad (7)$$
$$\gamma_m(\theta) = G_m = g_m(X_1, \ldots X_n) \quad (8)$$

All the right-hand sides are known (or at least stuff we can calculate from the data), and all the left-hand sides are functions of the unknown parameter θ. So we try to solve for θ, and call the result our estimate $\hat{\theta}$.

Remember that θ is a vector, with dimension d, so we've got m equations and d unknowns. And we remember some rules about counting equations and unknowns from algebra:

- If $m < d$, there are fewer equations than unknowns, and there won't be a unique solution. (The system of equations is "under-determined".) That is, we can't possibly hope to get *the* unique θ this way (though we might put some useful constraints on it). We need at least one feature/statistic per parameter.
- If $m = d$, there will (generally[2]) be a unique solution. (I will explain that weasily "hope" in a moment.) — Again, we want at least one feature per parameter.
- If $m > d$, there will generally not be *one* value of θ which can solve *all* the equations. (The system of equations is "over-determined".) We therefore seem to be in trouble if we use *more* information to estimate the parameters rather than less, but that sounds stupid.

Fortunately, about two hundred years ago the Ancestors realized that there was a sensible way to get an estimate of d parameters from $m > d$ observations, even when the equations can't be solved uniquely[3]: this is the method of least squares (Farebrother 1999). For each quantity, we calculate the squared error $(G_k - \gamma_k(\theta))^2$, we add up, and we adjust θ to minimize the sum of squared errors. In modern symbols:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2 \quad (9)$$

Since the sum of squares is necessarily ≥ 0, if there's a θ which *does* exactly match the observed values of the statistics, that's automatically what this method will pick out, but this *also* works even when $m > d$.

(Of course, there is nothing magical about treating all the summary statistics equally. We could also do weighted least squared

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2 w_k \quad (10)$$

with weights $w_k > 0$, and this is often useful in practice, say if one statistic is more important for applications, or better-measured. See Complementary Problem 2.)

2.1 A little worked example with the log-normal distribution

This has all been very abstract so it's good to give a concrete worked example with a distribution we're using a lot, like the log-normal. In Lecture 3, we saw that the expected value of the log-normal is $e^{\mu+\sigma^2/2}$. We also saw there that the theoretical median is just e^{μ}. The sample mean and median are of course among our very first and simplest summary statistics, let's say G_1 and G_2. So *one* way of estimating the parameters of the

[2] I put in the weasle-word "generally" because there can be some exceptions, say where random sampling noise gives a value for one of the G_k which just can't *exactly* match $\gamma_k(\theta)$ for any θ. (This can happen when we have continuous parameters for discrete observations and a small sample size, for example.) This sort of annoying edge case will be handled by the same least-squares trick I'm about to explain for over-determined systems.

[3] Before that, for thousands of years, the best scientific practice was to try to determine which d observations were the most accurate, use them to solve for the parameters, and discard the rest of the data, or at best use it to check later (Farebrother 1999). This was wasteful of data!

log-normal distribution would be to adjust μ and σ^2 until it matched the observed values of the mean and the median:

$$\gamma_1(\mu, \sigma^2) = e^{\mu+\sigma^2/2} = G_1 = \text{sample mean} \quad (11)$$

$$\gamma_2(\mu, \sigma^2 = e^{\mu} = G_2 = \text{sample median} \quad (12)$$

In this case, $d = m$ so there'll be a unique solution, and you can, in fact, work it out explicitly (because the functions involved are very well-behaved):

$$\hat{\mu} = \log G_2 \quad (13)$$

$$\widehat{\sigma^2} = 2\log(G_1/G_2) \quad (14)$$

It is *also* true that, for a log-normal distribution, the Gini index is given by $2\Phi(\sigma/\sqrt{2}) - 1$, where Φ is the CDF of the standard Gaussian distribution. If we had $m = 3$ summary statistics available to us, which were the mean, the median, and the Gini index, we'd need to set up the least-squares problem I mentioned above[4].

2.2 When will feature matching work, and how well?

We pick our features or summary statistics $g_1, \ldots g_m$; we calculate them on data $X_1, \ldots X_n$ to get numbers $G_1, \ldots G_m$; we try to match them to the corresponding theoretical values $\gamma_1(\theta), \ldots \gamma_m(\theta)$. We get a value $\hat{\theta}$ for the parameter, either by exactly solving the system of equations $\gamma_k(\theta) = G_k$, or by least squares. How well does this $\hat{\theta}$ approximate the true θ? Better said, what makes for better or worse estimates from this approach?

There are two *very* important requirements which have to hold if feature-matching is to have any chance of working:

1. **Convergence**: As we get more and more data, our summary statistics or features have to approach their theoretical limits:
$$\lim_{n\to\infty} g_k(X_{1:n}) \to \gamma_k(\theta^*) \quad (15)$$
This is where probability results like the law of large numbers come in: it needs to be the case that increasingly big samples become closer and closer approximations to the whole data-generating distribution. Most of the common summary statistics have this property!
2. **Identifiability**: Differences in the parameters *must* make a difference to the theoretical values of the features. That is, if we have two different parameter vectors $\theta \neq \theta'$, there has to be *some* γ_k where $\gamma_k(\theta) \neq \gamma_k(\theta')$. Said differently the vector-valued function $\gamma(\theta) = (\gamma_1(\theta), \ldots \gamma_m(\theta))$ must be invertible, so γ^{-1} must be well-defined[5]. Checking this is, in principle, just a matter of doing math with our favorite probability model.

If convergence fails, matching summary statistics to theoretical values is senseless. If identifiability or invertibility fails, even perfect matching can't pick out ("identify") the correct parameter value.

So, let's assume that we're using summary statistics which converge, and that the model parameters are identifiable from them. There usually going to be *lots* of different combinations of features with these properties. Some of them will work better than others, i.e., give more accurate and precise estimates. Two aspects of the features will improve the quality of the estimates:

3. **Sensitivity** to the parameters: small changes to the parameter vector should lead to big changes to the theoretical values of the features. More precisely, if we think of θ as a vector $\theta = (\theta_1, \theta_2, \ldots \theta_d)$, we'd like all the partial derivatives $\frac{\partial \gamma_k}{\partial \theta_i}$ to be large. In fact, all else being equal, $\text{Var}\left(\hat{\theta}\right)$ will be proportional

[4] Alternatively, we could pick any two of the three summary statistics, use them to estimate μ and σ^2, and then check that against the predicted value for the third statistic...

[5] We sometimes demand a stronger condition, that γ^{-1} isn't just well-defined, but also "smooth", say continuous, or even differentiable. This is so that small errors in the summaries will "track back" through γ^{-1} to small changes in the parameters. It is genuinely hard to come up with examples of real models and summaries where γ^{-1} exists but *isn't* smooth, however.

to $\left(\frac{\partial \gamma_k}{\partial \theta_i}\right)^{-2}$. In words: we'll estimate θ better if even small changes in θ are easily to detect in the features.

4. **Precision** of the statistics: the smaller the variance of each summary statistic, $\mathrm{Var}(G_k)$, the more precise the estimates will be. In fact, all else being equal, $\mathrm{Var}\left(\hat{\theta}\right)$ will be proportional to $\mathrm{Var}(G_k)$. In words: we'll estimate θ better if we're matching features which aren't themselves very noisy.

These last two points are, I hope, ones which make sense intuitively, but the next sub-section will explain them in some mathematical detail — and cash out my vague "all else being equal" in actual formulas.

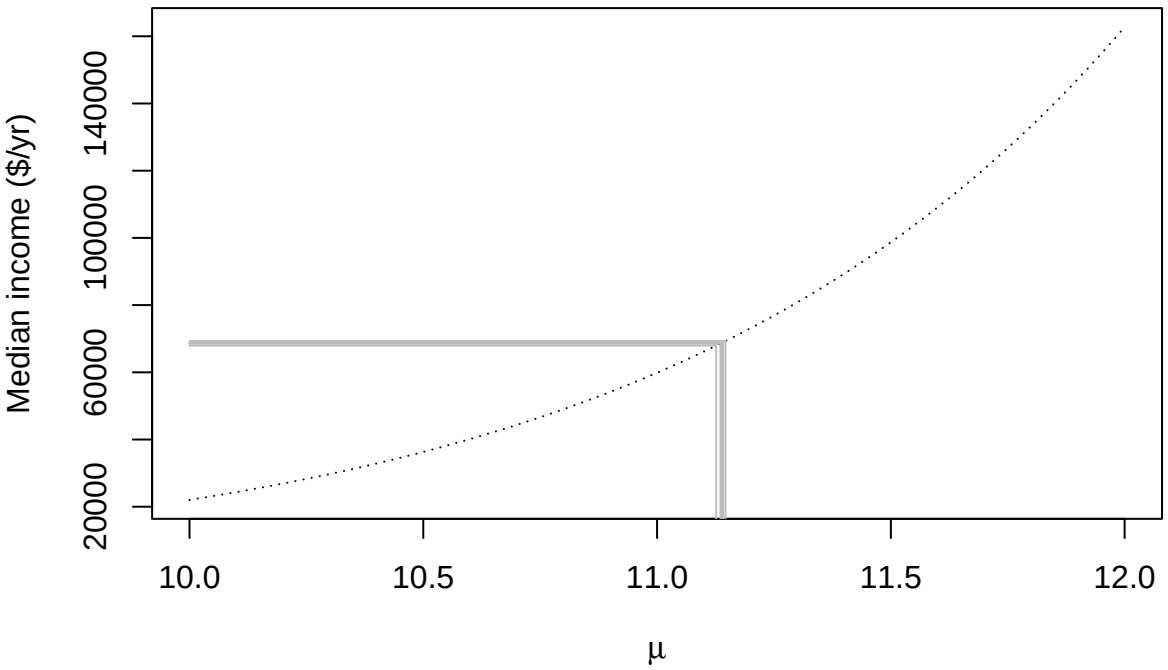

Figure 1: Dotted curve: theoretical median income as a function of the μ parameter in a log-normal model (namely e^{μ}). Black horizontal line: measurement of median income by the Current Population Survey in 2019. Black vertical line: implied value of μ from matching. Grey lines horizontal lines: a random scatter around the point estimate of the median, with the same uncertainty as the reported standard error for the median. Grey vertical lines: scatter of implied μ values, showing the induced uncertainty in this estimate of the parameter. Notice the slope $\frac{\partial \gamma}{\partial \mu}$ (dashed blue line) helps propagate uncertainty in the observed statistic into uncertainty in the estimated parameter.

2.3 Asymptotics for feature matching

This sub-section is a bit more mathematically involved than the rest of this section; it's useful stuff but not quite so crucial to what we'll do in the rest of the course, so you might want to skim it first, then come back to it later.

I said above that we can think of fitting a model by matching features as solving a least squares problem:

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}} \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2 \tag{16}$$

It'll be convenient to have a name for the function we're minimizing, so let's call it M:

$$M(G, \theta) \equiv \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2 \tag{17}$$

M is a sum of squares, so it must be ≥ 0. The only way we can have $M = 0$ is if $G_k = \gamma_k(\theta)$ for all $k \in 1 : m$. That is, the smallest conceivable value of M is attained when we match all the observed features exactly.

2.3.1 Consistency

Let's use the assumption I called "convergence" and numbered (1) above: as $n \to \infty$, $G_k \to \gamma_k(\theta^*)$. (Remember θ^* is the true value of θ, which we're trying to learn.) So

$$M(G, \theta^*) \to 0 \tag{18}$$

For any other $\theta \neq \theta^*$,

$$M(G, \theta) \to \sum_{k=1}^{m} (\gamma_k(\theta) - \gamma_k(\theta^*))^2 \tag{19}$$

Now use the assumption I called "identifiability" or "invertibility" and numbered (2) above: if $\theta \neq \theta^*$, then $\gamma_k(\theta) \neq \gamma_k(\theta^*)$ for (at least) some i. Thus, with this extra assumption, if $\theta \neq \theta^*$,

$$M(G, \theta) \to \sum_{k=1}^{m} (\gamma_k(\theta) - \gamma_k(\theta^*))^2 > 0 \tag{20}$$

Hence, under these assumptions, as $n \to \infty$, θ^* will come closer and closer to being the minimizer of $M(G, \theta)$, and any *other* parameter value will look worse and worse by comparison. Conclusion[6]:

> _Assuming convergence and identifiability, then $\hat{\theta} \to \theta^*$ as $n \to \infty$.

2.3.2 Uncertainties (variances) and the "sandwich variance" formula

We can say more about *how* $\hat{\theta} \to \theta^*$ if we use a little calculus. The first point is that $\hat{\theta}$ minimizes a function, and we know from calculus that the derivative is zero at a minimum[7]. So

$$\left.\frac{\partial M}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \tag{21}$$

Remember that $M(G, \theta) = \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2$, so this becomes

$$2\sum_{k=1}^{m} (G_k - \gamma_k(\hat{\theta})) \left.\frac{\partial \gamma_k}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \tag{22}$$

(I have my reasons for not dropping the factor of 2 here.)

It's annoying to keep track of all the derivatives separately, so we bundle them into a vector, the **gradient** of the function (with respect to M):

$$\nabla M(G, \hat{\theta}) = 0 \tag{23}$$

I will leaving writing this out in terms of the vector $G - \gamma(\hat{\theta})$ and a matrix of partial derivatives as an exercise.

Now we use another idea from basic calculus, that of a Taylor expansion or Taylor approximation: $h(u) \approx h(u_0) + (u - u_0) \left.\frac{dh}{du}\right|_{u=u_0}$. Applied here, and expanding around the true parameter value θ^*,

$$\nabla M(G, \hat{\theta}) \approx \nabla M(G, \theta^*) + (\nabla\nabla M(G, \theta^*))(\hat{\theta} - \theta^*) \tag{24}$$

[6] If you worry that this isn't a totally air-tight proof, you're not wrong. The missing extra assumptions ("regularity conditions") to make it rigorous aren't too onerous, but they do take us too far afield. If you really want to know more, see Vaart (1998), chapter 5.

[7] Assuming the function is smooth, the minimum is located in the interior of the domain, etc., etc.

Here, $\nabla\nabla M(G,\theta^*)$ is the matrix of second partial derivatives (with respect to the coordinates of θ), also called the **Hessian** of M.

Putting these last two equations together,

$$0 \approx \nabla M(G,\theta^*) + (\nabla\nabla M(G,\theta^*))(\hat{\theta} - \theta^*) \tag{25}$$

$$0 \approx (\nabla\nabla M(G,\theta^*))^{-1}\nabla M(G,\theta^*) + (\hat{\theta} - \theta^*) \tag{26}$$

$$\hat{\theta} \approx \theta^* - (\nabla\nabla M(G,\theta^*))^{-1}\nabla M(G,\theta^*) \tag{27}$$

As $n \to \infty$, by convergence, $\nabla M(G,\theta^*) \to 0$ (why?), so this again tells us that $\hat{\theta} \to \theta^*$. It tells us a little more, too. As n grows, $\nabla\nabla M(G,\theta^*)$ approaches a limiting Hessian matrix, say $\mathbf{h}$, and then

$$\hat{\theta} \approx \theta^* - \mathbf{h}^{-1}\nabla M(G,\theta^*) \tag{28}$$

so all the randomness in $\hat{\theta}$ comes from the randomness in $\nabla M(G,\theta^*)$. In fact, using the rules for algebra with variances,

$$\mathrm{Var}\left(\hat{\theta}\right) \approx \mathbf{h}^{-1}\mathrm{Var}\left(\nabla M(G,\theta^*)\right)\mathbf{h}^{-1} \tag{29}$$

This is called the **sandwich variance formula**.

Let's explore the two parts of this, the (limiting) Hessian $\mathbf{h}$ and the variance of the gradient $\mathrm{Var}\left(\nabla M(G,\theta^*)\right)$.

As $n \to \infty$, $G(X_{1:n}) \to \gamma(\theta^*)$, so

$$M(G,\theta) \approx \sum_{k=1}^{m}(\gamma_k(\theta) - \gamma_k(\theta^*))^2 \tag{30}$$

We can just take the second partial derivatives:

$$\frac{\partial^2 M}{\partial\theta_i\partial\theta_k} \approx \frac{\partial}{\partial\theta_k}2\sum_{k=1}^{m}(\gamma_k(\theta) - \gamma_k(\theta^*))\frac{\partial\gamma_k}{\partial\theta_i} \tag{31}$$

$$\approx 2\sum_{k=1}^{m}(\gamma_k(\theta) - \gamma_k(\theta^*))\frac{\partial^2\gamma_k}{\partial\theta_i\partial\theta_k} + \frac{\partial\gamma_k}{\partial\theta_i}\frac{\partial\gamma_k}{\partial\theta_k} \tag{32}$$

Evaluated at $\theta = \theta^*$, we get

$$\mathbf{h}_{jk} = \left.\frac{\partial^2 M}{\partial\theta_i\partial\theta_k}\right|_{\theta=\theta^*} \approx 2\sum_{k=1}^{m}\frac{\partial\gamma_k}{\partial\theta_i}\frac{\partial\gamma_k}{\partial\theta_k} \tag{33}$$

The partial derivatives $\left.\frac{\partial\gamma_k}{\partial\theta_i}\right|_\theta$ form an $m \times d$ matrix; in vector calculus we call this the **Jacobian matrix** of γ, so let's write this matrix $\mathbf{j}(\theta)$. Then we can say

$$\mathbf{h}_{jk} \approx 2\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*) \tag{34}$$

(You should check that this gives a $d \times d$ matrix, as required.)

Turning to the variance of the gradient, we have from above that the jth component, $\nabla M(\theta^*)_i$, is

$$\nabla M(G,\theta^*)_i = 2\sum_{k=1}^{m}(G_k - \gamma_k(\theta^*))\left.\frac{\partial\gamma_k}{\partial\theta_i}\right|_{\theta=\theta^*} \tag{35}$$

In matrix-vector form, this is

$$\nabla M(G,\theta^*) = 2(G - \gamma(\theta^*))\mathbf{j}(\theta^*) = 2G\mathbf{j}(\theta^*) - 2\gamma(\theta^*)\mathbf{j}(\theta^*) \tag{36}$$

The non-random term doesn't matter for the variance, so

$$\mathrm{Var}\left(\nabla M(G,\theta^*)\right) = 4\mathrm{Var}\left(G\mathbf{j}(\theta^*)\right) \tag{37}$$

$$= 4\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*) \tag{38}$$

In reading these formulas, remember that $G = g(X_{1:n})$, so the variance on the right-hand side changes — we hope it shrinks! — as n grows, even though the notation doesn't make that explicit.

Putting everything together,

$$\begin{aligned} \mathrm{Var}\left(\hat{\theta}\right) &\approx (2\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1}4\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*)(2\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1} & (39)\\ &= (\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1}\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*)(\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1} & (40)\end{aligned}$$

Notice that the smaller $\mathrm{Var}\left(G\right)$ is, the smaller $\mathrm{Var}\left(\hat{\theta}\right)$ is — less noise in the features we're matching on translates directly into more precise estimates.

The role of the derivative matrices is harder to grasp, but we can make a *little* more sense of this by considering the special case where if $m = p$, so that $\mathbf{j}$ is a square matrix which could be inverted. Then we have some cancellations on the right-hand side:

$$\begin{aligned} Var\hat{\theta} &\approx (\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1}\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*)(\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*))^{-1} & (41)\\ &= (\mathbf{j}(\theta^*))^{-1}((\mathbf{j}(\theta^*)^T)^{-1}\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*)(\mathbf{j}(\theta^*))^{-1}((\mathbf{j}(\theta^*)^T)^{-1} & (42)\\ &= (\mathbf{j}(\theta^*))^{-1}\mathrm{Var}\left(G\right)((\mathbf{j}(\theta^*)^T)^{-1} & (43)\end{aligned}$$

In this situation, we can say that increasing the derivatives $\frac{\partial \gamma_k}{\partial \theta_i}$ makes $\mathrm{Var}\left(\hat{\theta}\right)$ smaller. That is, making the features more "sensitive" to the parameters we're trying to estimate leads to more precise estimates.

(The last claim might suggest a "cheat" to get very precise estimates: just multiply all the features through by a big number c. This will, after all, increase the derivatives by the same factor c, and so reduce the variance of $\hat{\theta}$ by a factor of c^2. The catch, though, is that this will also increase $\mathrm{Var}\left(G\right)$ by a factor of c^2, so in the end it does nothing. Can you show that scaling all the features doesn't change $\mathrm{Var}\left(\hat{\theta}\right)$ in the general case, where $\mathbf{j}$ isn't invertible?)

Now that you've seen how things work when there are some cancellations, it should be easier to convince yourself that the same idea also applies when $\mathbf{j}$ is non-invertible: the larger the derivatives $\frac{\partial \gamma_k}{\partial \theta_i}$ get, the smaller $\mathrm{Var}\left(\hat{\theta}\right)$ gets.

2.3.2.1 How big is $\mathrm{Var}\left(G\right)$?

$\mathrm{Var}\left(G\right)$ is really $\mathrm{Var}\left(g(X_{1:n})\right)$, i.e., we're really looking at n samples. If those samples are independent of each other, then for a *lot* of summary statistics, $\mathrm{Var}\left(G\right) \propto 1/n$. This is true of sample means, but also (with some caveats) of medians, correlations, and many other summary statistics. It can even remain true if there is *some* dependence among the X's, but not too much. (We could make that precise but it's a different class, namely 36-467.)

2.3.2.2 Practical application of the sandwich variance formula

We've written the sandwich variance formula in two ways:

$$\mathrm{Var}\left(\hat{\theta}\right) \approx \mathbf{h}^{-1}\mathrm{Var}\left(\nabla M(G, \theta^*)\right)\mathbf{h}^{-1} \quad (44)$$

and

$$\mathrm{Var}\left(\hat{\theta}\right) \approx (\mathbf{j}(\theta^*)^T\mathbf{j}(\theta^*))^{-1}\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*)(\mathbf{j}(\theta^*)^T\mathrm{Var}\left(G\right)\mathbf{j}(\theta^*))^{-1} \quad (45)$$

Both involve knowing something about derivatives at the true parameter value θ^*, which of course is what we're trying to figure out!

To actually do calculations, what we usually do is substitute in $\hat{\theta}$ for θ^* throughout. This is justified on the grounds that, after all, $\hat{\theta} \to \theta^*$, and the hope that all the functions are continuous in the parameter. That still leaves the problem of getting $\mathrm{Var}\left(G\right)$, the variance of the summary statistics, but there are lots of ways of tackling this. (You know how to handle it if the summary statistics are means, for instance.)

2.3.2.3 A little worked example with the log-normal

Let's go back to our example with estimating the log-normal by matching the sample mean and sample median. Remember that

$$\gamma_1(\mu, \sigma^2) = e^{\mu+\sigma^2/2} \text{ (theoretical mean)} \tag{46}$$
$$\gamma_2(\mu, \sigma^2 = e^{\mu} \text{ (theoretical median)} \tag{47}$$

So

$$\mathbf{j}(\mu, \sigma^2) = \begin{bmatrix} \frac{\partial \gamma_1}{d\mu} & \frac{\partial \gamma_1}{\partial \sigma^2} \\ \frac{\partial \gamma_2}{\partial \mu} & \frac{\partial \gamma_2}{\partial \sigma^2} \end{bmatrix} \tag{48}$$
$$= \begin{bmatrix} e^{\sigma^2/2}\frac{\partial e^{\mu}}{\partial \mu} & e^{\mu}\frac{\partial e^{\sigma^2/2}}{\partial \sigma^2} \\ \frac{\partial e^{\mu}}{d\mu} & \frac{\partial e^{\mu}}{\partial \sigma^2} \end{bmatrix} \tag{49}$$
$$= \begin{bmatrix} e^{\mu+\sigma^2/2} & \frac{1}{2}e^{\mu+\sigma^2/2} \\ e^{\mu} & 0 \end{bmatrix} \tag{50}$$

We need to invert this, but it's just a 2×2 matrix:

$$\mathbf{j}(\mu, \sigma^2)^{-1} = \frac{1}{-e^{\mu+\sigma^2/2}e^{\mu}/2} \begin{bmatrix} 0 & -e^{\mu+\sigma^2/2}/2 \\ -e^{\mu} & e^{\mu+\sigma^2/2} \end{bmatrix} \tag{51}$$
$$= \begin{bmatrix} 0 & 1/e^{\mu} \\ 2/e^{\mu+\sigma^2/2} & -2/e^{\mu} \end{bmatrix} \tag{52}$$

When we've matched on G_1 and G_2, we get

$$(\mathbf{j})^{-1} = \begin{bmatrix} 01/G_2 & \\ 2/G_1 & 2/G_2 \end{bmatrix} \tag{53}$$

So we have

$$\text{Var}\left(\hat{\theta}\right) \approx \begin{bmatrix} 01/G_2 & \\ 2/G_1 & 2/G_2 \end{bmatrix} \begin{bmatrix} \text{Var}\left(G_1\right) & \text{Cov}\left(G_1, G_2\right) \\ \text{Cov}\left(G_1, G_2\right) & \text{Var}\left(G_2\right) \end{bmatrix} \begin{bmatrix} 01/G_2 & \\ 2/G_1 & 2/G_2 \end{bmatrix} \tag{54}$$

3 Fitting by Maximum Likelihood

Let's recall how the **method of maximum likelihood** works in general.

We have a sample $X_{1:n}$. We also have a model which gives us a probability density for such data sets, and this model contains one or more adjustable parameters, which generically we'll write as θ. The **likelihood** of a particular data set $x_{1:n} = (x_1, x_2, \ldots x_n)$ at a particular parameter value θ is the joint probability density of that data under that parameter value,

$$f(x_{1:n}; \theta) \tag{55}$$

If the data points are independent and identically distributed (which we usually try very hard to arrange!), then

$$f(x_{1:n}; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{56}$$

The **method of maximum likelihood** is to select the parameter value which gives the data the highest possible likelihood.

People usually work with the log-likelihood rather than the likelihood, for several reasons. The first reason is that the θ which maximizes one is always the same as the θ which maximizes the other (because log itself is a monotonically-increasing function); the second reason is that it leads to easier calculations (as we'll see in a moment or two); and the third is that it leads to easier proofs (as in the sketch in the optional section below). So

$$L(\theta) \equiv \sum_{i=1}^{n} \log f(x_i; \theta) \tag{57}$$

The **maximum likelihood estimator** is the value of θ which maximizes $L(theta)$,

$$\hat{\theta} \equiv \operatorname*{argmax}_{\theta} L(\theta) \tag{58}$$

(Notice that subscript here indicates what variable or variables we're maximizing over: here it's just the parameter(s); we treat the data as given, fixed.)

There are, broadly speaking, two ways to actually find $\hat{\theta}$. One is to say that $L(\theta)$ is just a function, which we can calculate at any particular value of θ, and hand the problem over to some general-purpose numerical optimization algorithm. This is *often* what we end up doing in practice, particularly when the models become more complicated.

The other approach is to use our calculus, and remember that generally a maximum is a point where the first derivative is zero and the second derivative is negative. (The exceptions are typically when the maximum is at the boundary of the allowed parameter region.) Taking the first derivatives and setting them to zero would give a system of equations

$$\left.\frac{partial L}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \tag{59}$$

for each $j \in 1 : d$. This becomes

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \log f(x_i; \hat{\theta}) = 0 \tag{60}$$

which, using the chain rule from calculus, can also be written

$$\sum_{i=1}^{n} \frac{1}{f(x_i; \hat{\theta})} \frac{\partial}{\partial \theta_i} f(x_i; \hat{\theta}) = 0 \tag{61}$$

This system of equations are called the **estimating equations** for the model, or sometimes the **first-order conditions** (because they involve only first derivatives). The system of estimating equations will have one

equation for each unknown parameter θ_i. Solving these equations gives us a candidate[8] value for $\hat{\theta}$. Sometimes, especially for very well-behaved models, the equations can be solved explicitly, and we end up writing down nice formulas for $\hat{\theta}$ in terms of the data. Even when there aren't closed-form solutions, however, there are techniques to solve the equations numerically.

(Notice, by the way, that it'd be much harder to do the equivalent calculus if we worked with the likelihood rather than the log-likelihood, because differentiating a product is more complicated than differentiating a sum. This is one of the promised reasons for preferring *log*-likelihood.)

3.1 Why should we use maximum likelihood estimates?

There are three big reasons.

1. Generally, maximum likelihood estimates are consistent, which remember means that as we get more and more data ($n \to \infty$), they will converge on the true value of the parameter[9]. There are some exceptional cases where maximum likelihood estimates don't converge on the truth, but they're genuinely odd.
2. Generally, maximum likelihood estimates are **efficient**. We want our estimates to converge on the true value, and we can measure how far off they are by, say, the expected squared error, $\mathbb{E}\left(|\hat{\theta} - \theta^*|^2\right)$. If we consider any other consistent (probably-approximately-correct) estimator, say $\tilde{\theta}$, it's generally the case that $\mathbb{E}\left(|\tilde{\theta} - \theta^*|^2\right) \geq \mathbb{E}\left(|\hat{\theta} - \theta^*|^2\right)$. Other estimators *usually* make bigger errors[10].
3. We can be precise about the standard errors and confidence sets of the MLE, at least asymptotically.

The next sub-section, which is optional, sketches the argument for consistency. The one after that, which is less optional, states the results about standard errors and confidence sets.

3.1.1 Why the MLE is consistent (optional and sketchy)

Let's define

$$\overline{L}(\theta) = \frac{1}{n} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta) \tag{62}$$

The factor of $1/n$ doesn't change the location of the maximum, so

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \overline{L}(\theta) \tag{63}$$

The point of introducing $\overline{L}$ is that it's a sample average of a function of the data, and we know something about the behavior of sample averages. Since we're supposing the data points X_i are independent and identically distributed[11], the terms $\log f(X_i; \theta)$ are also IID. Now the law of large numbers tells us that, with IID samples, for any function h,

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \to \mathbb{E}\left(h(X)\right) = \int h(x) f(x; \theta^*) dx \tag{64}$$

[8] In principle, we should still check that we've found a maximum rather than a minimum or an inflection point, but it's honestly a little unusual for that to be a serious problem.

[9] The formal definition is this: we can pick any margin of error ϵ, and any error probability δ, and if n is big enough, $\mathbb{P}\left(|\hat{\theta} - \theta| \epsilon\right) \leq \delta$. We also say that $\hat{\theta}$ **converges in probability** to θ^*. In the 1980s, computer scientists re-invented this concept, which statisticians had been using explicitly since the 1920s, under the name of "probably approximately correct". This was a clear case of re-inventing the wheel, but the re-invention did have a much more descriptive and transparent name, which statisticians should probably adopt.

[10] If this doesn't sound like a very crisply stated mathematical theorem, that's because it isn't. The actual result, the Cramer-Rao theorem, is fairly complicated even to state, but what I say in the main text does get the main idea across. If you want the gory details, see, e.g., Pitman (1979).

[11] Maximum likelihood also works if the data points *aren't* IID, but we need to find some replacement or generalization for the law of large numbers that works with whatever kind of dependence or heterogeneity we're dealing with.

That is, sample averages converge on expected values. So

$$\overline{L}(\theta) \to \int \log\left(f(x;\theta)\right) f(x;\theta^*) dx \equiv \lambda(\theta, \theta^*) \tag{65}$$

The function $\lambda(\theta, \theta^*)$ says what *expected* log-likelihood the parameter value θ will get when the true parameter value is θ^*.

Since the random function $\overline{L}(\theta)$ is approaching the deterministic function $\lambda(\theta, \theta^*)$, the MLE $\hat{\theta}$ will approach

$$\operatorname*{argmax}_{\theta} \lambda(\theta, \theta^*) \tag{66}$$

We can say something about what that maximizer is by using a fundamental (if hardly obvious) result from probability theory, sometimes called **Gibbs's inequality**, which says that for any two pdfs f and f',

$$\int \log f'(x) f(x) \leq \int \log f(x) f(x) dx \tag{67}$$

and that the only way to get equality is if $f'(x) = f(x)$ everywhere[12] (Complementary Problem 1). Applied here, we see that the θ which maximizes $\lambda(\theta, \theta^*)$ is in fact nothing other than θ^*.

Conclusion: the MLE $\hat{\theta}$ should converge on the true parameter value θ^* as $n \to \infty$.

3.1.1.1 Going deeper into the weeds: what if the model is wrong?

In the previous argument, there was only one place where I assumed that the probability model was correct. This is when I asserted that the true pdf from which the data are drawn is $f(x;\theta^*)$ for the true-but-unknown θ^*. Suppose the *actual* pdf was instead some $p(x)$, and this pdf is not equal to $f(x;\theta)$ no matter what value of θ we try. In the jargon, we say that the model is then **mis-specified**. What happens to the MLE when the model is mis-specified?

Well, the law of large numbers part of the argument still holds, so

$$\overline{L}(\theta) \to \mathbb{E}\left(\log f(X;\theta)\right) \tag{68}$$

but now

$$\mathbb{E}\left(\log f(X;\theta)\right) == \int \log f(x;\theta) p(x) dx \tag{69}$$

$\hat{\theta}$ will converge to the θ^* which maximizes *this*. What can we say about it?

$$\int \log f(x;\theta) p(x) dx \quad = \quad \int \log p(x) p(x) dx + \int \log\left(\frac{f(x;\theta)}{p(x)}\right) p(x) dx \leq \int \log p(x) p(x) dx \tag{70}$$

This tells us that $\int \log\left(\frac{f(x;\theta)}{p(x)}\right) p(x) dx \leq 0$. It's more common to work with the negative of this quantity,

$$D(p \| f(\cdot;\theta)) = \int \log\left(\frac{p(x)}{f(x;\theta)}\right) p(x) dx \tag{71}$$

is called the **divergence**[13] of the density $f(\cdot;\theta)$ from p. It's the expected log likelihood ratio between the true distribution (p) and the one our model hypothesizes ($f(\cdot;\theta)$). The divergence is ≥ 0, and it's only $= 0$ if the two distributions coincide completely.

[12] More exactly, $f'(x) = f(x)$ "almost everywhere", i.e., except on a set of points of total length ("measure") zero. If you know enough to demand this qualifier, though, you also know enough to insert it for yourself.

[13] Other names for this quantity include the **Kullback-Leibler divergence** (after Kullback and Leibler (1951)), the KL divergence, the KL **information** (or "KL information number" or "KL information criterion"), and the **relative entropy**.

In general, then, when the model is not right, the MLE will converge on the parameter value θ^* which minimizes the divergence from the true distribution_. It will, in this very particular sense, find the best approximation to the true distribution that the model can provide. Econometricians have a wonderful name for θ^*; they call it the "pseudo-truth", or the "pseudo-true parameter value". So when the model is correctly specified, the MLE converges on the truth, and when the model is mis-specified, the MLE converges on the pseudo-truth.

3.1.2 Uncertainty around the MLE (less optional)

Since the MLE is the maximum of the log-likelihood, the first derivative there is zero, and the second derivative tells us how sharply peaked the log-likelihood function is. It should make some sense that the more sharply peaked the the log-likelihood function, the more easily we can locate the maximum, and the less uncertainty there is. This not only makes sense, but is true. The **observed information** is

$$I(\hat{\theta}) \equiv -\frac{d^2 L}{d\theta^2}(\hat{\theta}) \tag{72}$$

and, quite generally,

$$\mathrm{Var}\left(\hat{\theta}\right) \approx I(\hat{\theta})^{-1} \tag{73}$$

when the model is well-specified.

Notice that L is a sum of n terms, so, with independent samples, we can expect it to be $\propto n$. Hence $I(\hat{\theta})^{-1} \propto n^{-1}$. The standard error of the MLE should therefore be $\propto n^{-1/2}$.

If we're not sure that the model is well-specified, that the MLE is well-behaved, etc., a robust, if computationally more expensive, alternative, is to use the bootstrap (lecture 6).

3.1.2.1 What if there's more than one parameter?

We need the $d \times d$ matrix of all partial derivatives,

$$I_{ij}(\hat{\theta}) \equiv -\frac{\partial^2 L}{\partial \theta_i \theta_j} \tag{74}$$

and then

$$\mathrm{Var}\left(\hat{\theta}\right) \approx \mathbf{I}(\hat{\theta})^{-1} \tag{75}$$

The matrix of second partial derivatives of a function is called that function's **Hessian**, so you will often encounter that word in statistical contexts, R documentation, etc.

The argument that this should be $\propto n^{-1}$, and standard errors $\propto n^{-1/2}$, works just as though there was one parameter.

3.1.2.2 Expected information, Fisher information

The **Fisher information** in one observation at parameter value θ is a matrix, with entries

$$F_{ij}(\theta) \equiv \mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta_i \partial \theta_j}\right) \tag{76}$$

The **expected information** in n observations is

$$J_{ij}(\theta) \equiv \mathbb{E}\left(\frac{\partial^2 f(X_{1:n};\theta)}{\partial \theta_i \theta_j}\right) \tag{77}$$

With IID observations, $J_{ij}(\theta) = nF_{ij}(\theta)$. That is, information adds up across independent observations. This is relevant because when the true parameter value is θ^*,

$$\mathrm{Var}\left(\hat{\theta}\right) \to \mathbf{J}(\theta^*) = n\mathbf{F}(\theta^*) \tag{78}$$

More exactly,

$$\frac{1}{n}\mathrm{Var}\left(\hat{\theta}\right) \to \mathbf{F}(\theta^*) \tag{79}$$

We don't know θ^*, but we do have an estimate of it, $\hat{\theta}$, and that's generally consistent, so we can say

$$\frac{1}{n}\mathrm{Var}\left(\hat{\theta}\right) \to \mathbf{F}(\hat{\theta}) \tag{80}$$

Taking the expectation buried in the definition of $\mathbf{F}$ can be hard, but we usually don't have to. Again assuming IID observations, the law of large numbers applies: at any fixed θ,

$$\frac{1}{n}\mathbf{I}(\theta) \to \mathbf{F}(\theta) \tag{81}$$

since $F\mathbf{F}(\theta)$ is the expected value of the observed information $\mathbf{I}$ in a single data point. Since the MLE is consistent, and $\mathbf{F}$ is (usually) continuous in θ,

$$\frac{1}{n}\mathbf{I}(\hat{\theta}) \to \mathbf{F}(\hat{\theta}) \to \mathbf{F}(\theta^*) \tag{82}$$

So we're back to saying that

$$\mathrm{Var}\left(\hat{\theta}\right) \approx \mathbf{I}(\hat{\theta})^{-1} \tag{83}$$

3.1.3 Situations where maximum likelihood gives unreasonable or inconsistent answers

Maximum likelihood *usually* works, but it doesn't always. Some of the places where it fails are carefully-arranged mathematical horror shows. Others are more logical problems. For instance, if we have a collection of n data points, $x_1, x_2, \ldots x_n$, and ask "what pdf, out of all possible pdfs, gives these points the highest likelihood?", we'll find that the answer is always a set of infinitely high, infinitely narrow spikes centered at the data points. (In symbols, $f(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$.) (If we try using merely very tall, very narrow spikes centered at the x_i, we can always increase the likelihood by making them taller and narrower.) This isn't *wrong*, exactly — any probability we put on a point x which isn't one of the x_i has to be taken *from* the x_i — but it seems unreasonable. Maximum likelihood generally does poorly when it faces such "infinite-dimensional" or "nonparametric" problems.

Finally, I should mention that the MLE also does poorly for finding a parameter like $x_{\min}$ in the Pareto distribution — it always suggests using the smallest observed x. This is actually consistent *if* the data come purely from a Pareto distribution. But usually we just want a Pareto *tail*, and are being tactically (or tactfully) vague about the distribution of the body. Since we're not including the distribution of the body in our calculation of the likelihood, we shouldn't be surprised that the MLE does something unreasonable. (Once we have $x_{\min}$, though, using the MLE to get α is definitely the right way to go.) For a more reasonable approach to estimating $x_{\min}$, see Clauset, Shalizi, and Newman (2009).

4 Matching Features vs. Maximizing Likelihood

I have gone on at some length about how to use maximum likelihood because it's a fundamental technique in statistics, and usually the best thing to do *when we can*. The problem, for a lot of social analyses, is with the italicized phrase.

Maximum likelihood presumes not just that *someone* collected a sample of individual data points, but that we have access to those data points, so we can evaluate $\log f(x_i; \theta)$ and add it up across the data set. Often, however, what *we* have access to aren't the individual x_i, but *only* summary statistics $g_1(x_1, \ldots x_n)$, $g_2(x_1, \ldots x_n)$, down to $g_m(x_1, \ldots x_n)$. These are functions of the data, as the notation indicates, but they are *not* the data themselves. We then need to fall back on the technique of matching these summaries[14].

There are various reasons why we might need to do this.

1. *The original data is too large to release.* This can still be true, but is less of an issue today than in, say, 1900, or even 1980, and will presumably be less and less of an issue in the future.
2. *The original data is too sensitive to release.* A lot of our knowledge about income and wealth distributions comes from official records. Part of the reason that people are willing to be (fairly) honest with the authorities is that modern governments are really powerful and intrusive and lying to them too blatantly about money has reliably bad consequences. But part of the way modern governments soften this pill is by promising not to disclose everyone's personal information to anyone with an Internet connection. If you are a trusted employee of the IRS or the Census Bureau, you could probably calculate MLEs on the full tax records of the US population. But the rest of us are never going to see *that* data.
3. *The original data was not preserved.* Especially if we're interested in historical comparisons, it becomes harder and harder to retrieve the original data, but summaries may be easily obtained.

Of these three reasons, the second is probably the one which will be the most important to you in practice.

4.1 Summary matching estimates are usually less efficient than maximum-likelihood estimates

I said above that maximum likelihood estimates are generally *efficient*, meaning that nothing else converges on the truth faster. Summary-matching estimates are "something else", so they will generally converge slower than the MLE. This is disappointing, but we're usually using them because we don't have access to the data which we'd need to calculate the MLE.

4.1.1 Sufficient Statistics

There is a theoretically-important case where summary-matching estimates *are* efficient. This is when we're matching what are called **sufficient statistics**. This only happens when the likelihood can be written in the following special form:

$$f(x_{1:n}; \theta) = t(s(x_{1:n}); \theta)\phi(x_{1:n}) \tag{84}$$

That is, we can factor the whole likelihood into two terms, one of which, $\phi(x_{1:n})$, doesn't involve the parameter at all, and the other, which does involve the parameter, doesn't use the data directly, but only a function of the data $s(x_{1:n})$. That function s is the "sufficient statistic"[15].

[14] We could in principle calculate the joint distribution of all the summary statistics as functions of the parameter θ, and so maximize the likelihood of the summaries. The key phrase in that sentence is "in principle" — working out exact sampling distributions of summary statistics is usually extremely hard. (For an example of the sort of heroic calculations required, see, e.g., Virkar and Clauset (2014), where the model is a power law and the summary statistics are the number of observations in fixed bins or brackets.) For very large n, many sampling distributions become approximately Gaussian, which simplifies this approach (Wood 2010), but also makes it more similar to a least-squares matching estimate...

[15] That definition is based on what's called the "Neyman factorization criterion". There are actually multiple, equivalent ways of defining "sufficient statistic". Another way to do so is to say that a statistic is sufficient when it keeps all the information the data had about the parameter θ. This can be made precise using information theory (Kullback and Leibler 1951).

Most models do *not* have sufficient statistics, but many familiar ones do. For the Gaussian distribution, for instance, the sample mean and the sample variance are sufficient. For the log-normal, the sufficient statistics are the sample mean logarithm and the sample variance of the logarithm. For a Pareto distribution with known minimum, the mean logarithm is a sufficient statistic.

When we can identify a model's sufficient statistics, and they're available to us, we should certainly prefer matching them when we can. In fact, a lot of the time[16], matching sufficient statistics gives us the MLE.

[16]The distribution needs to be (or be put into) what's called **exponential family** form, so $f(x_1, x_2, \ldots x_n; \theta) = h(x)\frac{e^{s(x_{1:n})\cdot\theta}}{z(\theta)}$. (The denominator here, $z(\theta)$, makes sure that this is a valid probability distribution that adds up to 1 across all possible values of $x_{1:n}$. It's called the **normalizing factor** or sometimes the **partition function**.)} If a distribution has statistics and satisfies some additional technical requirements, it *must* be an exponential family, but there are many common distributions which aren't exponential families, like the t distribution, or some of the natural generalizations of the Pareto distribution.

5 Summing Up on Fitting

- If we have individual-level data, the method of maximum likelihood is usually better than matching summary statistics.
  - The MLE will generally converge on the true parameter values ("consistency").
  - The MLE will converge at least as fast as any other consistent estimator ("efficiency").
  - For large samples, the variance of the MLE can be found from the matrix of 2nd derivatives of the log-likelihood. (Parameters which make a big difference to the log-likelihood will be more precisely estimated than the others.)
  - There's usually no exact formula for the MLE, but that's why we have numerical optimization algorithms.
- Matching summary statistics or features is *usually* inferior to the MLE.
  - If statistics converge on population values as n grows, and the features uniquely identify parameters, then matching is consistent.
  - Matching estimates are more precise when the summary statistics have less variance (the statistics we're matching aren't very noisy)
  - Matching estimates are more precise when the features are very sensitive to the parameters (small changes to parameters imply big changes to the features).
  - If we know the variances of the summary statistics, and how sensitive the features are to the parameters, we can calculate (approximate, large-n) variances for our matching estimators
  - Matching estimators will have at least as much variance as the MLE
- There are many situations, especially in social statistics, where we *only* have summary statistics, so we need to rely on matching.

6 Checking Goodness-of-Fit

We fit our model by estimating the parameter(s) θ. The model however may or may not be right — it's responsible to check the over-all **goodness of fit**. (Or, if we're pessimists, the **lack of fit**.) There are two big ways of doing this:

- See if the model, with the estimated parameters, can predict *other* features of the data
- See if the model, with the estimated parameters, gives a distribution that matches the over-all distribution of the data.

If you like, you can think of *both* of these approaches as instances of a more general rule, "see if the model, with the estimated parameters, can predict something beyond those parameters". But the practicalities are somewhat different.

6.1 Checking by Calculating New Features

This works rather like estimating by matching features. On the one hand, we have some summary statistic of the data, let's say $g_{test}(X_1, \ldots X_n)$. This, as the notation indicates, is used as the "test statistic". On the other hand, we can calculate what, theoretically, that should be, through some function $\gamma_{test}(\theta)$. We can then just *compare* the observed value on our actual data, $g_{test}(x_1, \ldots x_n)$, to the theoretically predicted value with our estimated parameters, $\gamma_{test}(\hat{\theta})$. The bigger the gap, the worse the fit. We can do this *whether or not* our estimate comes from matching features, from maximizing likelihood, from a dream, etc. Basically *any* feature of the data can be used as test statistic in this sense, though not all of them will work equally well.

There are some important considerations here.

- *Do not expect a perfect match.* In the first place, the data are random, so $G_{test} = g_{test}(X_1, \ldots X_n)$ is going to be a random variable. Even if we somehow knew θ^*, we shouldn't expect $G_{test} = \gamma_{test}(\theta^*)$ all the time. In the second place, we've only got an estimate $\hat{\theta}$, not θ^*.
  - As with estimating by matching, we prefer test statistics with small variance, and we prefer features which are very sensitive to the underlying parameters.
- *Deciding how big a mis-match is too big requires more work.* We'll look, later in the course, at some methods for working how much the test statistic can differ from the theoretical prediction when the model is right. (One way to do this is called "bootstrapping".) For now, though, we often prefer a more qualitative, touchy-feely evaluation.
- *Some features can be more important than others* We might, in some applications, really *care* about some features more than others. (For instance, we might really want to get the income shares of high percentiles right.) In those situations, it can be legitimate to focus about whether those features get predicted well, and not worry so much about other features which might be missed.

6.1.1 Do Not Use Features Twice

At this point however I need to issue an important warning. If we estimated our model by matching on some features, we cannot check the fit of the model by seeing if it predicts those same features. We've just made sure that it does! At best, this just makes sure we didn't botch any of our calculations. So, for example, if we estimate a log-normal distribution by matching the mean and the median, we learn nothing by seeing that the fitted model successfully predicts the mean and the median — what else could it do? The same goes for any statistic which is a function of the features we matched on (say the ratio of mean to median). What we need is to look at *other* features, which are not implied by what we matched on[17].

[17] You might think this wouldn't be a concern if we used maximum likelihood instead of feature-matching. However, *if* the model as sufficiency statistics (as defined above), especially if it's an "exponential family" model (as defined above), the MLE will automatically match those statistics. So for instance using the MLE to estimate the log-normal will automatically match the mean of the log and the variance of the log, and so anything which is a function of those (like the "coefficient of variation" of the log, $\mathrm{Var}(\log X)/\mathbb{E}(\log X)$).

6.2 More Global Checks of Fit for Distributions

We can also check the fit of a model by comparing the distribution with the estimated parameters to some *other*, less-model-dependent way of estimating the distribution. For instance, we might compare the CDF implied by the model with the "empirical" CDF of the raw data. We can divide the range of X up into bins or brackets, and compare the actual number of samples in each bin to the number we'd expect based on the model. This in turn is very similar to comparing the pdf implied by the model to a histogram of the data, which suggests we can just compare the model's pdf to other estimates of the pdf which don't rely on the model (such as the ones plotted in the last few lectures).

Some of these comparisons can be turned into formal statistical tests, like the Kolmogorov-Smirnov test (which compares the empirical CDF of the data to the CDF implied by the model) or the χ^2 test (which compares the observed counts in bins to those expected under the model). The distribution of the test statistic is usually computed assuming that the theoretical model has no adjustable parameters. This is, of course, rarely the case. If the model *does* have parameters estimated from the data, then the model's fit will be apparently more impressive (the test statistic will be smaller) than the usual calculation suggests, *just because* the model was adjusted to fit *this* particular data set. The simplest way to avoid this problem is **data splitting**: randomly split the data into two distinct parts, fit the model on one part and check the fit on the other. When data splitting is impractical, though, there are other ways to correct for parameter estimation, such as the bootstrap methods we'll look at later in the class.

6.3 Good Fit $\not\Rightarrow$ Truth

Even if we've found a model which seems to fit well, we should be cautious about concluding that it's *true*. There are at least two reasons for this.

1. Other models might also fit the data in the ways we're checking.
2. The model might fail to fit the data in other ways, which we haven't looked at.

There are two ways to deal with the first problem. We can *either* try to report the whole range of models which are (so far as we can tell) compatible with the data, and admit to a certain amount of **model uncertainty**[18]. *Or* we can try to find some *other* aspect of the data, or new data, where the different models make *different* predictions, and test those predictions. (That's science.) Or, of course, we can pursue both approaches at once, if we have the time and the energy.

That even the best-fitting model might fail to accommodate new data, or more refined probes of old data, is just a fact of life: science is fallible, or more precisely corrigible. There are even many situations where we continue to use models we know are not altogether accurate, if the inaccuracies are small and/or un-important for our purposes.

[18] Within a model, the set of all parameter values which are compatible with the data, at a certain level of goodness-of-fit, forms a **confidence interval**, or more generally a **confidence set**. In words: the confidence set contains all the parameter values we *can't* confidently reject. This idea is just extending that same approach to all the *models* we can't reject. (It was advocated long ago by Haavelmo (1944), and quite correctly in my opinion.)

7 Complementary Problems

1. Prove Gibbs's inequality.
2. *Feature weighting.* Consider doing feature matching by solving a *weighted* least squares problem,

$$\hat{\theta}_w = \operatorname*{argmin}_{\theta} \sum_{k=1}^{m} (G_k - \gamma_k(\theta))^2 w_k \tag{85}$$

   with weights $w_k > 0$.
   a. Find a formula for the asymptotic variance of this estimator, similar to the one derived for the unweighted estimator above. Check that when all the weights are equal, your answer matches the formula for the unweighted estimator.
   b. Using your formula from (a), find the weights w_k which make $\mathrm{Var}\left(\hat{\theta}_w\right)$ as small as possible. Is the answer unique? How would you describe it in words?
   c. Convince yourself that using weights w_k is the same as replacing the function $g_k(x_{1:n})$ by $\sqrt{w_k} g_k(x_{1:n})$, and similarly for $\gamma_k(\theta)$. Why then does using weights make any difference?
3. *Propagation of error* Read the handout on "propagation of error" linked to on the class website for HW 1.
   a. Use this technique to find formulas for the standard errors of estimating the log-normal parameters from the sample mean and the sample median.
   b. Do your results from (a) formulas match the ones obtained above, from the "sandwich variance" formula? Should they?
   c. What's the general relationship between the propagation-of-error approach and the sandwich variance? *Hint*: If $f(x)$ has an inverse, how is the derivative of f^{-1} related to the derivative of f?

8 Further reading

Estimation theory is a huge subject, which can be pursued at many levels of mathematical and computational sophistication. Some of the standard references, like Lehmann and Casella (1998), are pretty stiff going.

There is a good discussion of likelihood-based inference with minimal technicalities, in Cox (2006), and, with a *bit* more mathematical demands, in Pitman (1979). The latter gives the simplest proof I've seen of the Cramer-Rao inequality, a key result which says that the Fisher information gives a lower bound on the variance of any estimator. (This result goes back to Cramér (1945) and Rao (1945).)

More abstractly, maximum likelihood and feature-matching are both instances of a general strategy for estimation, called "M-estimation", based on minimizing (or maximizing) some function which says how well the model matches the data. I have tried to present both feature-matching and the MLE in a way which suggests this common structure. In doing so I've been very influenced by the presentation in Vaart (1998) and Geer (2000), but both of these call for a lot more math than this course. I'm not really aware of another presentation of *general* M-estimation at our level.

On the importance of test models, and using summary statistics to do so, see Gelman and Shalizi (2013).

References

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51:661–703. https://doi.org/10.1137/070710111.

Cox, D. R. 2006. *Principles of Statistical Inference*. Cambridge, England: Cambridge University Press.

Cramér, Harald. 1945. *Mathematical Methods of Statistics*. Uppsala: Almqvist; Wiksells.

Farebrother, Richard William. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-0545-6.

Geer, Sara A. van de. 2000. *Empirical Processes in M-Estimation*. Cambridge, England: Cambridge University Press.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66:8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 (supplement):iii–115. https://doi.org/10.2307/1906935.

Kullback, Solomon, and R. A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86. https://doi.org/10.1214/aoms/1177729694.

Lehmann, E. L., and George Casella. 1998. *Theory of Point Estimation*. 2nd ed. Berlin: Springer-Verlag.

Pitman, E. J. G. 1979. *Some Basic Theory for Statistical Inference*. London: Chapman; Hall.

Rao, C. R. 1945. "Information and the Accuracy Attainable in the Estimation of Statistical Parameters." *Bulletin of the Calcutta Mathematical Society* 37:81–91.

Vaart, A. W. van der. 1998. *Asymptotic Statistics*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

Virkar, Yogesh, and Aaron Clauset. 2014. "Power-Law Distributions in Binned Empirical Data." *Annals of Applied Statistics* 8:89–119. https://doi.org/10.1214/13-AOAS710.

Wood, Simon N. 2010. "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems." *Nature* 466:1102–4. https://doi.org/10.1038/nature09319.