Factorization Theorem

Ilan Shlain

Introduction

Whenever statisticians are interested in estimating an unknown variable (referred to as parameter), data is their bread and butter. When data has valuable information related to our parameter we are able to better estimate it and often provide actionable insights, so it is only natural that statisticians would like to get as much data as possible to work with. However, there is often times a distinction between useful and unnecessary information related to whatever we want to estimate. A lot of times, when data is collected in large quantities, a heap of unnecessary info is brought along side it. This unnecessary information can be an obstacle when trying to conduct data analysis and estimation, because it requires more time to analyze more data; since analysis of this unnecessary data doesn't provide insight to our parameter, statisticians would like to avoid its incorporation. This is where the idea of a **Sufficient Statistic** comes in.

Firstly, a statistic is some value that has been extracted from provided data. For example, if we collect the height of 100 people, the average height calculated by $\frac{\textbf{total height of everyone}}{100}$, would be a statistic of the data. A sufficient statistic is a statistic that doesn't lose any information relevant to a given parameter. For example, if you were trying to estimate the rate at which a light bulb factory made defective products (per 100), you wouldn't need to include all 100 light-bulbs in your data analysis; all you would need is to count the defective ones and then you would be able to estimate the true rate. This statistic of counting the defective light-bulbs is sufficient because it still gives you all of the information related to the defective rate (assuming you know the sample size of 100), but now you only need to deal with a single number for your analysis rather than 100 data points related to each light-bulb. The idea of a **sufficient statistic** can be condensed into a simple sentiment: we want to reduce the data we were given while maintaining all of the relevant information for our goal.

Sufficient statistics are great, but how do we get them or even know if a statistic can be considered sufficient? This is where the **Factorization Theorem** comes in. Commonly known as the Fisher-Neyman Factorization Theorem (or criterion), this theorem gives us a rigorous mathematical definition of sufficiency. Put simply, when we consider a sample of data points who come from the same distribution and want to focus on some parameter of that distribution, if an expression describing how likely our data is can be broken down into two parts-one that depends only on the parameter and a chosen statistic, and another that depends only on the data- then that statistic contains all of the relevant information and is considered to be sufficient for our parameter.

In this paper, we will cover the definitions and extensions of the Factorization Theorem, explore the history and significance of the theorem's development, recall how the Factorization Theorem relates to the coursework in UIC's STAT 411 class, we will look at some recent real world applications of the Factorization Theorem, and finally we will explore the potential applications of the Factorization Theorem to new or existing problems.

Definition of the Factorization Theorem

Fisher, Neyman, Fisher-Neyman are all different ways to refer to the same theorem. For simplicity, we will refer to our theorem as the **Factorization Theorem**, and then later look at the contributions of different mathematicians who are mentioned alongside it.

Firstly, let's define a key term related to the Factorization theorem: Sufficiency. Given a sample $\vec{X} = \{X_1, ..., X_n\}$ and a parameter θ, a statistic $T(\vec{X})$ is sufficient for θ if it captures all of the information pertaining to it within the sample. In other words, the statistic $T(\vec{X})$ reduces the data without losing any information relevant to the estimation of θ. There are two main ways of determining if a statistic is sufficient. The first is checking to see if the conditional distribution of $\vec{X}$ given $T(\vec{X}) = t$ is independent of θ. This can often become tedious and difficult as conditional distributions become more complicated, so we turn to the second method of utilizing the Factorization Theorem.

The general intuition behind the Factorization Theorem is as follows: If you can factor the joint probability density of a sample into two distinct parts, one dependent on a statistic $T(\vec{X})$ as well as θ and the other dependent on only $\vec{X}$, then $T(\vec{X})$ is sufficient for θ.

Here is a more formal definition,

Factorization Theorem. Given a sample $\vec{X} = X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta), \theta \in \Omega$. The statistic $T(\vec{X})$ is sufficient for θ iff there exist two non-negative functions k_1 and k_2 such that,

$$\prod_{i=1}^{n} f(x_i;\theta) = f(x_1;\theta)f(x_2;\theta), \ldots, f(x_n;\theta) = k_1(T(\vec{X});\theta)k_2(x_1, \ldots, x_n)$$

where $k_2(x_1, \ldots, x_n)$ doesn't depend on θ.

Let us look at an example of how the Factorization Theorem can be used to determine the sufficiency of a statistic.

Example 1. (Poisson Distribution). Consider the following sample $X_1, \ldots, X_n \overset{iid}{\sim} poisson(\theta)$, and consider the statistic $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$. We will utilize the factorization theorem to show that $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

$$\prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \underbrace{e^{-n\theta}\theta^{\sum x_i}}_{k_1(T(X_1,\ldots,X_n,;\theta))} \underbrace{\prod_{i=1}^{n} \frac{1}{x_i!}}_{k_2(X_1,\ldots,X_n)}$$

We have now factored our joint pdf into the product of a function k_1 dependent on T and θ, and a function k_2 dependent only on $X_1, \ldots, X_n$. So by the factorization theorem, $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

The factorization theorem also extends to the multivariate case, where multiple statistics can be jointly sufficient for one or more parameters. If we now define $\vec{T} = (T_1, \ldots, T_m)$ to be a vector of m statistics and $\vec{\theta} = (\theta_1, \ldots, \theta_p)$ to be a vector containing p parameters, then we have the following definition for the factorization theorem.

Factorization Theorem (Multivariate Case) $\vec{T}$ is considered to be jointly sufficient for our parameters $\vec{\theta}$ iff there exist two non-negative functions k_1 and k_2 such that

$$\prod_{i=1}^{n} f(x_i; \vec{\theta}) = k_1(t; \vec{\theta}) k_2(x_1, \ldots, x_n)$$

where $k_2(x_1, \ldots, x_n)$ doesn't depend on $\vec{\theta}$.

This definition is really not much different from our univariate case, and this is a big reason the factorization theorem becomes incredibly useful; whenever utilizing the conditional probability definition of sufficiency, as we work with more variables and more complex distributions, it becomes more desirable to factor out a giant product than to work with potentially messy conditional distributions.

For both the univariate and multivariate case, the factorization theorem provides a very intuitive understanding of sufficiency. When you think of the joint distribution of the samples (i.e the likelihood) as all of the data available, you can think of the factorization process as splitting up the data into one section k_1 that has meaningful information about your parameter θ and another section k_2 that has no meaningful data related to θ. When you look at the meaningful section k_1, you will see that the only information related to θ or than θ itself is the statistic t which k_1 is dependent on; or in other words (to repeat our definition of sufficiency) t contains all of the available information for θ.

History of the Factorization Theorem

In the early 1920's, Ronald A. Fisher, in his work titled "On the Mathematical Foundations of Theoretical Statistics", introduced what was called the *Criterion of Sufficiency*; which was simply stated to be that for any given sample, "That the statistic chosen should summarize the whole of the relevant information supplied by the sample". Fisher's idea that a single statistic can drastically reduce the amount of data while maintaining all relevant information was revolutionary and incredibly promising for the field of mathematical statistics. However, as Fisher himself mentioned in his work, he had not yet found a method of systematically identifying a statistic that satisfied the criterion of sufficiency; more specifically, Fisher was "not satisfied as to the mathematical rigor of any proof which I can put forward to that effect".

This is where we introduce Jerzy Neyman, who in the 1930's, attached a more rigorous mathematical definition to Fisher's idea. Firstly, he formalized Fishers idea's regarding sufficiency and the conditional distribution.

$$T(x) \text{ is sufficient for } \theta \iff \text{the conditional distribution of } X \mid T(X) = t \text{ does not depend on } \theta$$

Neyman was also the one to introduce the **factorization criterion** which is at the center of this paper. The introduction of the factorization theorem as presented by Neyman gave us a more straightforward way of determining sufficient statistics.

The importance of Neyman's contribution is that it turned sufficiency from a problem of intuition and conditional probability into something that could be tested through the likelihood function. In most cases, the Factorization Theorem proved to be much more convenient than picking an ideal candidate for a sufficient statistic and then testing it through the conditional distribution. The process of factorization often gives insight into a possible sufficient statistic, even if there was no candidate to begin with.

Naturally, Neyman's Factorization Theorem was seen as the go to method of attaining sufficient statistics and was used commonly in the application of the Rao-Blackwell Theorem which was introduced in the 1940's.

Then in their 1949 paper titled "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics", Paul R. Halmos and Leonard J. Savage, gave a more general version of the Factorization Theorem using measure-theory and probability. This development was important for the modern applications of the Factorization Theorem, because prior to Halmos and Savage, the Factorization Theorem was limited to more textbook like settings such as continuous and discrete distributions; with their contributions, the Factorization Theorem could be used in more general probability spaces, which was crucial for modern applications which utilize models with more complex distributions

Overall the Factorization Theorem was developed from Fisher's initial idea of sufficiency, and through collaboration and different branches of mathematics turned into one of statistics' most valuable tools.

Stat 411 and the Factorization Theorem

The factorization theorem is at the core of our coursework in STAT 411 at UIC. We utilize the factorization theorem whenever there is a problem that requires students to find the sufficient statistic, joint sufficient statistic, joint minimally sufficient statistic, etc. This is the case because Neyman's formalization of sufficiency through the factorization theorem gives us a relatively simple process by which we can determine if a statistic is sufficient. Most of the time it is simpler to factor the joint distribution (as it is essentially a giant product) than to derive and observe a distribution conditioned on a statistic.

Consider our previous example with the Poisson distribution, and how we would solve it with conditional probability.

Example 2. (Poisson Distribution) Recall that we want to show that the statistic $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ is sufficient for our sample of *iid* Poisson random variables.

$$T = \sum_{i=1}^{n} x_i \sim Poisson(n\theta), \qquad f_T(t;\theta) = \frac{e^{-n\theta}(n\theta)^t}{t!}, \quad t = 0, 1, 2, \ldots$$

We then calculate the conditional probability

$$\frac{f(x_1, \ldots, x_n; \theta)}{f_T(t;\theta)} = \frac{e^{-n\theta}\theta^{\sum x_i} \prod_{i=1}^{n} \frac{1}{x_i!}}{\frac{e^{-n\theta}(n\theta)^t}{t!}} = \frac{t!}{n^t \prod_{i=1}^{n} x_i!} = H(x_1, \ldots, x_n)$$

As we see $H(x_1, \ldots, x_n)$ does not depend on θ. So by the definition of sufficient statistic, $T(x_1, \ldots, x_n)$ is sufficient for θ.

Now although this has proven to be quite computable, this was still noticeably more complicated than simply looking at the joint distribution and trying to factor. Though both methods have their place, with more complicated distributions and even for most examples we deal with in STAT 411, the factorization theorem proves to be much simpler to utilize than the conditional distribution method.

Extending to the multivariate case, whenever we encounter a task that requires joint sufficiency we often find ourselves using the factorization theorem because of its relative simplicity. Consider the following example with the normal distribution, which will not only show how the factorization theorem is applied in the multivariate case, but also provide an example of it deals with continuous random variables.

Example 3. (Multivariate Normal) Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta_1, \theta_2), \quad -\infty < \theta_1 < \infty$ and $\theta_2 > 0$. Find a statistic $\vec{T}$ that is jointly sufficient for θ_1 and θ_2.

$$\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) = \underbrace{exp\left\{\frac{-1}{2\theta_2}\sum_{i=1}^{n} x_i^2 + \frac{\theta_1}{\theta_2}\sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2\theta_2} - n\log(\sqrt{2\pi\theta_2})\right\}}_{k_1(T(\vec{X});\theta_1,\theta_2)} * \underbrace{1}_{k_2(X_1,\ldots,X_n)}$$

Where

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = \sum_{i=1}^{n} X_i^2, \quad \text{and} \quad T(\vec{X}) = \{T_1, T_2\}$$

So by the Factorization Theorem, the statistics $\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ are jointly sufficient for (θ_1, θ_2).

Since sufficiency is tied to many of the other properties and processes we do on class, it is also the case that the factorization theorem comes up. For example consider Rao-Blackwellization, where we condition on an unbiased estimator on a sufficient statistic; most of the time when doing this process, we will utilize the factorization theorem to attain the sufficient statistic.

Recent Developments and Applications

The factorization theorem and sufficiency go hand in hand, so it is only natural to find its applications present whenever sufficiency of a statistic is necessary. Consider a recent development in the estimation of traffic velocity in heterogeneous networks presented by Tiwari et al. (2025), who employed probabilistic modeling and the factorization theorem to derive a sufficient statistic for velocity estimation. Essentially, as it relates to our topic , Tiwari et al. modeled the amount of handoffs (i.e the number of times a mobile device switches its network connection) as a Rayleigh distribution and applied the factorization theorem to get a sufficient statistic for velocity estimation.

The reason that the factorization theorem was so crucial here, was because their whole goal was estimation. Naturally, they wanted to find the MVUE via Rao-Blackwellization; this requires a sufficient statistic to condition an unbiased estimator on, which is where our trusty theorem appears! Though the distribution and some terms from this paper may be unfamiliar to us, the underlying process of factoring a joint probability mass function is still the same. Here lies an invaluable aspect of the factorization theorem in which we see a retention of simplicity even in more complex circumstances.

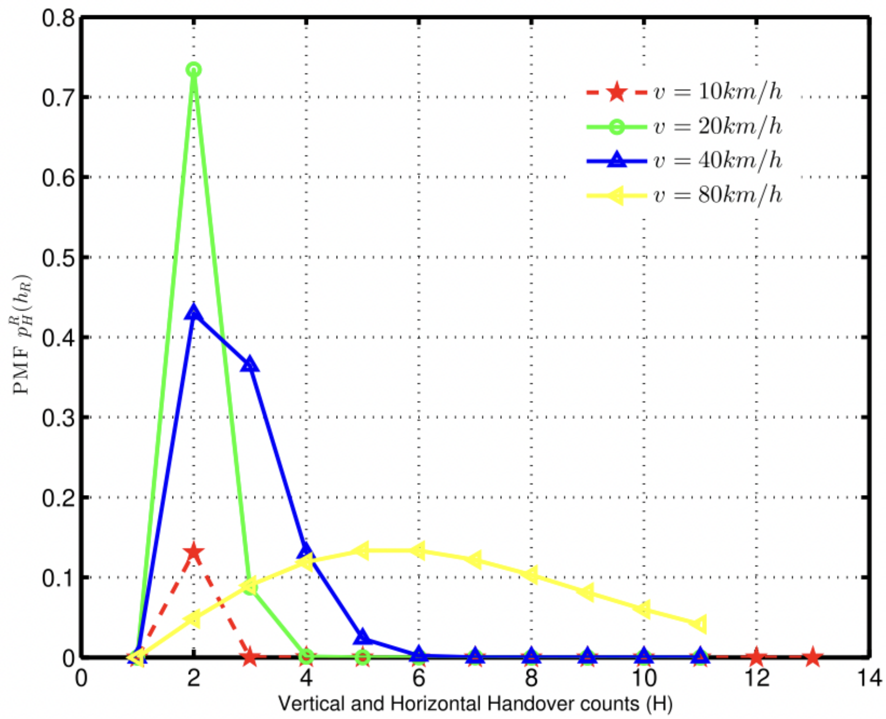

Figure 1: Rayleigh distribution approximating the probability mass function of handover counts for varying user velocities, as seen in Tiwari.

Another example of the factorization theorem being used appears in recent work related to estimating and tracking the location of multiple speakers using microphone data. Weisberg et al. (2019) consider a statistical model relating to signals coming from microphones and utilize maximum likelihood methods in order to estimate certain unknown parameters related to the speakers being monitored. A key aspect of this paper is that the researchers didn't want to work with **all** of the data that was collected, as much of it was considered to be unnecessary information unrelated to the parameters. To deal with this, they utilized the factorization theorem to conclude that their signal processing algorithm the (MVDR-BF) was "a sufficient statistics for estimating the speech PSD" (Weisberg et al. 2019), where speech PSD was one of the unknown parameters they were trying to estimate.

We see in this paper, that although the factorization theorem seems like it is a trivial tool used in classrooms for sufficiency problems, it is actually the case that the Factorization Theorem is still being actively used in tasks as complicated as signal processing for data reduction. This again shows how the Factorization Theorem retains some levels of simplicity even when applied to tasks of significant difficulty.

A Potential Application of the Factorization Theorem

Let us now look at a potential application of the Factorization Theorem today. Due to the nature of the Factorization Theorem and its relationship to sufficient statistics, it is natural to find it wherever sufficiency may prove useful. Consider the field of cyber-security, in which monitoring information is crucial to maintaining user safety. When monitoring login attempts on a network, it would be impractical to analyze every single attempt and algorithmically make a decision on whether or not a login should be interrupted. Instead we should strive to find a sufficient statistic such that we can make an informed decision without having to look at **every** single point. This is where the Factorization Theorem may come in handy.

We could model the problem with a Poisson distribution with parameter θ. Where $X_1, \ldots, X_n$ represent the number of suspicious login attempts during n equal time intervals. We've seen in *Example 1* that via the Factorization Theorem our sufficient statistic T has the form $\sum_{i=1}^{n} X_i$. This result may seem underwhelming because of its relatively standard result, but the potential we want to focus on is the robustness of the Factorization Theorem. In reality suspicious login attempts can be better modeled by more complex models and distributions such as a Poisson regression model, but even then, the Factorization Theorem may have utility in helping find a jointly sufficient statistic vector.

Overall, the potential of the Factorization Theorem goes as far as the potential of sufficient statistics, as it is an indispensable tool or rather definition that allows statisticians to efficiently explore data reduction using likelihood. Whether it be potential network monitoring or real world location estimation via sound waves, the Factorization Theorem, a pivotal result, has incredible breadth and depth in its use.

In Conclusion

The Factorization Theorem, on the surface may seem like a formality students must participate in to get to the result they actually want, but that overlooks the true beauty behind the theorem that Fisher, Neyman, Halmos, and Savage, so carefully contributed to. The idea that the likelihood encodes valuable information which can potentially be simplified through factoring is simple yet profound; despite the increase in model complexity over the years, the intuition behind the Factorization Theorem and its relationship to the likelihood retains its strength and continues to be a cornerstone of statistical sufficiency and data reduction.

References

[1] Fisher, R. A. (1922). *On the mathematical foundations of theoretical statistics*. Philosophical Transactions of the Royal Society of London. Series A, 222, 309–368.

[2] Halmos, P. R., & Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20(2), 225–241.

[3] STAT 411 Course Notes. (2025). *Sufficient Statistics and the Factorization Theorem*. University of Illinois Chicago.

[4] Tiwari, R., Gahlaut, V., Kaushik, M., Rani, P., & others. (2025). An enhanced approach to minimum variance unbiased velocity estimation, incorporating horizontal and vertical handoff in HetNets. *Scientific Reports*.

[5] Weisberg, K., Gannot, S., & Schwartz, O. (2019). An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 656–660). IEEE. https://doi.org/10.1109/ICASSP.2019.8682659