

1 Inequalities

1.1 Arithmetic average is greater than the geometric average

Let $\{a_n\}_{n=1}^N$ be a sequence of nonnegative numbers and define

$$A_n := \frac{\sum_{i=1}^n a_i}{n} \text{ and } G_n := \left(\prod_{i=1}^n a_i \right)^{1/n}.$$

Then $A_n \geq G_n$ for any n . Unless all $\{a_i\}_{i=1}^N$ are equal to each other, the last inequality is strict. For example, when $n = 2$

$$\frac{a_1 + a_2}{2} \geq \sqrt{a_1 a_2}.$$

In other words, $A_2 \geq G_2$. This can be shown by using algebra and starting with $(a_1 - a_2)^2$.

Here is an outline for a recursive proof for $n > 2$. First without loss of generality, assume that numbers $\{a_i\}_{i=1}^N$ are reindexed from smaller to larger. Thus, we have $a_n > A_{n-1}$ unless $a_1 = a_2 = \dots = a_n$. Then we can write A_n in two parts and observe that the second part is positive

$$A_n = A_{n-1} + \frac{a_n - A_{n-1}}{n}.$$

Now take the equality to n th power and cancel terms to obtain

$$A_n^n > a_n A_{n-1}^{n-1}$$

Lastly use the induction hypothesis to finish the proof.

Actually the arithmetic - geometric average inequality is true also for weighted averages

$$a_1^{w_1} a_2^{w_2} \dots a_n^{w_n} \leq w_1 a_1 + w_2 a_2 + \dots + w_n a_n \tag{1}$$

where w_i s are nonnegative weights summing up to 1.

1.2 Hölder's Inequality

Arithmetic - geometric average inequality can be generalized by incorporating L sequences $\{a_n^l\}_{n=1}^N$ where $l = 1 \dots L$.

$$(a_1^1)^{w_1} (a_2^1)^{w_2} \dots (a_n^1)^{w_n} + (a_1^2)^{w_1} (a_2^2)^{w_2} \dots (a_n^2)^{w_n} + \dots + (a_1^L)^{w_1} (a_2^L)^{w_2} \dots (a_n^L)^{w_n} \leq (a_1^1 + a_1^2 + \dots + a_1^L)^{w_1} + (a_2^1 + a_2^2 + \dots + a_2^L)^{w_2} + \dots + (a_n^1 + a_n^2 + \dots + a_n^L)^{w_n}$$

But this inequality is generally known in alternative form for $L = 2$

$$\sum_{i=1}^n a_i^1 a_i^2 \leq \left(\sum_{i=1}^n (a_i^1)^p \right)^{1/p} \left(\sum_{i=1}^n (a_i^2)^q \right)^{1/q}$$

where $1/p + 1/q = 1$. This inequality can be written in a vector form after recalling the norm notation for a vector a^1

$$\|a^1\|_p = \left(\sum_{i=1}^n (a_i^1)^p \right)^{1/p}$$

Then

$$a^1 a^2 \leq \|a^1\|_p \|a^2\|_q$$

With a stretch of imagination, we can replace sequences a^1 and a^2 with functions f and g , and sums with integration

$$\int_X |f(x)g(x)|dx \leq \left(\int_X |f(x)|^p dx \right)^{1/p} \left(\int_X |g(x)|^q dx \right)^{1/q}$$

In the special case of $p = q = 2$, we obtain

$$\int_X |f(x)g(x)|dx \leq \left(\int_X f^2(x)dx \right)^{1/2} \left(\int_X g^2(x)dx \right)^{1/2}$$

This is commonly known as Cauchy-Schwarz inequality. It is often used to establish convexity properties of a function not obtainable in a closed form.

2 Convex sets and functions

We first define a convex set. A set $S \subseteq \mathfrak{R}^n$ is convex if $x_1, x_2 \in S$ implies $\lambda x_1 + (1 - \lambda)x_2 \in S$ for $0 \leq \lambda \leq 1$. Observe that intersections of convex sets are still convex. In particular, intersection of hyperplanes are convex. This observation is used in Linear Programming to establish the convexity of feasible regions.

For every function f defined on a domain D in \mathfrak{R}^n , we can define a set called epigraph of f as

$$\text{epi } f := \{(x, y) \in \mathfrak{R}^{n+1} : x \in D, y \in \mathfrak{R}, y \geq f(x)\}$$

A function is called convex if and only if its epigraph is convex. An equivalent definition of convexity can be given as follows: $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$ for a function f defined on a convex set S (What if S is not convex?). A function f is called concave if $-f$ is convex. Some properties of convex functions follow:

1. Weighted sums of convex functions are convex. If f_i is convex and $a_i \geq 0$, then $\sum_i a_i f_i$ is also convex.
2. Upper envelope of convex functions is convex. If $f(x, \omega)$ is convex for all $\omega \in \Omega$, then $\sup_{\omega \in \Omega} f(x, \omega)$ is convex.
3. Expectation of a convex function is convex. If $f(x, d)$ is convex in x for each d and let D be a random variable, then $E f(x, D)$ is convex in x .
4. Dimension reduction via minimization preserves convexity. Let $f(x, y)$ be a jointly convex function in x and y and define $g(x) := \inf_y f(x, y)$. Then $g(x)$ is convex in x . For example $f(x, y) = xy$ is convex in each variable but it is not jointly convex.
5. Level sets of convex functions are convex. If f is a convex function then $\{x : f(x) \leq \alpha\}$ is a convex set for every α .

6. Existence of a subgradient. If f is a convex function, there exists a vector ∂f which satisfies

$$\text{Subgradient Inequality:} \quad f(z) \geq f(x) + \partial f(z - x)$$

for every z . In general, there could be several (actually infinitely many) subgradients of a function at a given point.

7. Jensen Inequality. If f is a convex function, then $E_D f(D) \geq f(E_D(D))$.

3 Information Updates

3.1 Bayesian update

Instead of presenting general Bayesian statistics, we focus on a model often studied in SN. Let D_t be i.i.d. (independently and identically distributed) $N(\mu, \sigma^2)$

$$f_{D_t}(d|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(d - \mu)^2}{2\sigma^2}\right)$$

We assume that σ^2 is known and suppress it in the notation. In frequentist statistics, after observing a population of $\{d_t\}_{t=1}^T$ parameter μ is estimated by solving

$$\max_{\mu} \prod_{t=1}^T f_{D_t}(d_t|\mu)$$

This estimation method is known as maximum likelihood estimation and the function above is called the (sample) likelihood function.

In Bayesian statistics, a (posterior) distribution is used to estimate parameter μ

$$f(\mu|\{d_t\}_{t=1}^T) = \frac{\prod_{t=1}^T f_{D_t}(d_t|\mu) f(\mu)}{\int \prod_{t=1}^T f_{D_t}(d_t|\mu) f(\mu) d\mu}. \quad (2)$$

Observe that the joint density of $(\{d_t\}_{t=1}^T, \mu)$ appears in the numerator. The terms in this density can be grouped into two: those involving $(\{d_t\}_{t=1}^T, \mu)$ and those involving only $(\{d_t\}_{t=1}^T)$. In a special case, the first group may represent a density. Then the computation of the posterior greatly simplifies.

To illustrate these ideas, we define a prior for the expected value μ . Our prior is also Normal but with expected value m and variance σ^2/ν

$$f_{\mu}(\mu) = \frac{1}{\sqrt{2\pi\sigma^2/\nu}} \exp\left(-\frac{(\mu - m)^2}{2\sigma^2/\nu}\right).$$

With this prior, the joint distribution of $(\{d_t\}_{t=1}^T, \mu)$ becomes

$$\prod_{t=1}^T \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(d_t - \mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2/\nu}} \exp\left(-\frac{(\mu - m)^2}{2\sigma^2/\nu}\right).$$

The joint distribution can be written as (check the algebra) $f(\mu|\{d_t\}_{t=1}^T)f(\{d_t\}_{t=1}^T)$, specifically

$$f(\mu|\{d_t\}_{t=1}^T) = \frac{1}{\sqrt{2\pi\sigma^2/(\nu+T)}} \exp\left(-\frac{(\mu-\bar{m})^2}{2\sigma^2/(\nu+T)}\right) \text{ and}$$

$$f(d = \{d_t\}_{t=1}^T) = \frac{1}{\sqrt{2\pi\sigma^2}^T \sqrt{\det(\mathbf{I}_T + 1_T 1_T'/\nu)}} \exp\left(-\frac{(d - m\mathbf{1}_T)(\mathbf{I}_T + 1_T 1_T'/\nu)^{-1}(d - m\mathbf{1}_T)'}{2\sigma^2}\right) \quad (3)$$

where

$$\bar{m} = \left(\frac{\nu}{\nu+T}\right)m + \left(\frac{T}{\nu+T}\right)\left(\sum_{t=1}^T d_t/T\right)$$

and \mathbf{I}_T is the identity matrix of size $T \times T$, $\mathbf{1}_T$ is a vector of 1's with size T . With these observations, the right hand side of (2) simplifies to the specific form of $f(\mu|\{d_t\}_{t=1}^T)$ given in (3).

We conclude that the posterior distribution of μ conditioned on normally distributed data $\{d_t\}_{t=1}^T$ is $N(\bar{m}, \sigma^2/(\nu+T))$ if the prior distribution of μ is $N(m, \sigma^2/\nu)$. Note that the distribution of D_t remains as $N(\mu, \sigma^2)$ but the distribution of the mean μ changes with the update. On the other hand the marginal distribution of $\{d_t\}_{t=1}^T$ is $N(m\mathbf{1}_T, \sigma^2(\mathbf{I}_T + 1_T 1_T'/\nu))$.

In an SN context, $\{d_t\}_{t=1}^T$ could be the demands for a product observed in the last T periods. The prior can be constructed from a population $\{d_t^o\}_{t=1}^\nu$ of ν observations. This population could be coming from another but similar product for the last ν periods. A logical estimate of μ is the average of $\{d_t^o\}_{t=1}^\nu$. If the other product's demand d_t^o is $N(m, \sigma^2)$, then the average is $N(m, \sigma^2/\nu)$. This explanation motivates our choice of the prior distribution.

This Bayesian update construction is used in V. Iyer and M. E. Bergen, Quick response in manufacturer-retailer channels, *Management Science*, 43, 559-570 (1997).

3.2 Bivariate update

Let us first define the bivariate normal density for demands d_X and d_Y . Suppose that their means are μ_X and μ_Y , variances are σ_X^2 and σ_Y^2 , correlation is ρ . Then the variance-covariance matrix is defined by

$$\Sigma := \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

The density is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\frac{([x \ y] - [\mu_X \ \mu_Y])\Sigma^{-1}([x \ y] - [\mu_X \ \mu_Y])'}{2}\right)$$

Sometimes this distribution is denoted by $N(\mu_x, \mu_y, \sigma_X^2, \sigma_Y^2, \rho)$. When $D_Y = d_y$ is given, the conditional distribution of D_X is

$$[D_X|D_Y = d_y] \sim N(\mu_x + \rho(\sigma_X/\sigma_Y)(d_y - \mu_y), \sigma_X^2(1 - \rho^2))$$

We note that unlike the specific Bayesian update presented above, the Bivariate update mechanism modifies the variance of D_X . It is theoretically possible to update the variance in Bayesian fashion but the algebra

becomes overwhelming. The Bivariate update mechanism provides an easier alternative. However, it requires additional correlation information ρ .

For modelling purposes, both Bayesian and Bivariate updates are acceptable. Bivariate updates are possibly more common in the literature. However, both updates rely on the special properties of the Normal distribution to avoid heavy algebra. To extend these approaches to other distributions is not an easy task. For example, many distributions do not have a convenient bivariate density function as simple as Normal density. It will be very useful to be able to do information update without assuming Normal demands.

3.3 Stepping outside the Normal updates: Resolution of additive uncertainties

We study a two period model but the framework can be extended to longer periods. Suppose that we are studying the demand that will materialize at the end of the second period. Depending on when the demand is studied, it can have two representations

$$D_{0,2} = X_1 + X_2$$

$$D_{1,2} := [D_{0,2}|X_1 = x_1] = x_1 + X_2$$

where X_1 and X_2 are independent (not necessarily identical) random variables. Let $D_{0,2}$ denote the demand when no information is available and let $D_{1,2}$ denote the demand (possibly modified) with the additional information gained in period 1. In general, $D_{s,t}$ can denote the random variable for period t demands when that variable is measured in period $s \leq t$. With this interpretation X_s denotes the update made on to $D_{s-1,t}$ to obtain $D_{s,t}$. This update happens because of the additional information obtained in period s . It is possible to argue under certain conditions that updates are uncorrelated.¹

Before the information update we have demand $D_{0,2}$. This demand evolves into $D_{1,2}$ with the update. Although, we argue for the uncorrelatedness of updates X_1 and X_2 , demands $D_{0,2}$ and $D_{1,2}$ are always correlated. This is because, $D_{0,2}$ and $D_{1,2}$ both contain the same random variable. X_1 and X_2 are both uncertain updates initially but they are resolved (observed) one by one in each period, hence the name of this subsection. It is possible to develop an alternative multiplicative updates framework as opposed to additive but conceptually the difference is limited (because the multiplicative framework can easily be deduced from the additive equations by putting them into the exponential function).

Clearly, additive uncertainties framework does not require X_1 or X_2 be Normal so it can be used in contexts where Normal assumption is questionable. However, it would be nice to manipulate the additions of updates so it is advisable to work with updates whose convolutions are readily available. Examples could be exponential, binomial or Poisson distribution. Applications of additive uncertainties framework are very few in the SN literature.

4 Exercises

1. Is \emptyset a convex set?

¹Define $D_{s,t} = E[D_t|\mathfrak{S}_s]$ where \mathfrak{S}_s information available in period s . Consider the conditional expectation as a projection on to subspace generated by information observations until s . Use the perpendicularity (corresponds to no correlation) of a projection of a vector to the vector itself minus the projection.

2. Either prove that union of two convex sets is convex or disprove with a counterexample.
3. Prove that upper envelope of convex functions is convex.
4. Prove that expectation over a discrete random variable of a convex function is convex.
5. Prove that dimension reduction via minimization preserves convexity.
6. Prove that the level sets of convex functions are convex.
7. Prove the weighted version of the arithmetic and geometric inequality in (1) using the concavity of the logarithm function.
8. Prove the existence of a subgradient using the existence of a supporting hyperplane for every convex set.
9. Prove that if f is convex then $f(Ax + b)$ is convex with a matrix A and vector b .
10. Prove that if $g(x, d)$ is linear for each fixed d and f is convex then, $Ef(g(x, D))$ is convex in x .
11. Prove the Jensen Inequality using the existence of a subgradient.
12. Suppose that D_1 and D_2 are two random variables and $E(D_2) = 0$. Let f be a hypothetical convex cost function for a system.
 - a) Establish that $E_{D_1, D_2} f(D_1 + D_2) \geq E_{D_1} f(D_1)$. Explain if you need D_1 and D_2 to be independent.
 - b) Use the above inequality to intuitively explain that the value of information is always nonnegative.
 - c) Use a) to argue that $D_1 + D_2$ is stochastically more variable than D_1 .
 - d)* Suppose that ξ_1 is stochastically more variable than ξ_2 , does there exists a random variable ξ_3 independent of ξ_2 such that $\xi_1 = \xi_2 + \xi_3$ in distribution.
13. Refer to Bayesian information updates. Consider the posterior distribution for μ . Let us call $\nu/(\nu+T)$ as the relative (to $\{d_t\}_{t=1}^T$ observations) accuracy of the information in the prior. Interpret the following special cases and discuss what happens to the posterior and the relative accuracy of the prior: a) $\nu = 0$. b) $\nu = \infty$. c) $T = 0$. d) $T = \infty$.
14. Read the interpretation of the prior on μ when it is representing the data coming from another product. Note that the mean of the posterior is a weighted average of two means. What are these means and what are the weights? Instead of using the prior distribution for our Bayesian estimation, what happens to our estimation if we directly and indiscriminately use the population $\{d_t^0\}_{t=1}^{\nu}$ for the other product and the population $\{d_t\}_{t=1}^T$ for our product? Do we have the same distribution for the average defined as

$$\frac{\sum_{t=1}^{\nu} d_t^0 + \sum_{t=1}^T d_t}{\nu + T}.$$

If yes, is the Bayesian estimation equivalent to Frequentist estimation?

15. In (3), the marginal distribution of $\{d_t\}_{t=1}^T$ is $N(m\mathbf{1}_T, \sigma^2(\mathbf{I}_T + \mathbf{1}_T\mathbf{1}'_T/\nu))$. Interestingly enough, now demands are correlated! However, initially we assumed independent demands with $N(\mu, \sigma^2)$. Observe that as ν grows correlations decrease. Shall we to conclude that not knowing and estimating μ has caused dependence, explain. (Metin: I am not quite sure as how to interpret this properly, your comments are welcome.)

16. Extend the additive uncertainties framework to a 4 period case. Point out the dependence structure among random variables. List all the random variables by the end of the third period. This exercise just tests the understanding of the notation.
17. Consider the two-period additive uncertainties framework. In addition, suppose that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Let us see if we can make this framework equivalent to Bivariate updates. Suppose that we respectively observe $X_1 = x_1$ and $D_Y = d_Y$ in additive uncertainties and Bivariate frameworks where $x_1 = d_Y$. To establish equivalency, we want to ensure that $[D_{0,2}]$ and $[D_X]$ have the same distribution both before and after the observations x_1 and d_Y . How should D_X and D_Y be chosen (their distribution, means, variances and correlation). Express these parameters in terms of $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$. Check if with your choice of parameter values $\text{Cov}(D_{0,2}, X_1) = \text{Cov}(D_X, D_Y)$.
18. In the additive uncertainties framework, if $E(X_1) = 0$ and $E(X_2) = 0$ then $E(D_{0,2}|X_1 = x_1) = x_1$. A series of similar equalities can be shown for $D_{s,t}$ which leads to establishing that $D_{s,t}$ is a martingale. Explain if this property is harmed when $E(X_s) > 0$, for example do we end up with a sub(super) martingale? Note that many non-Normal distributions whose convolutions are easy have positive expected values.

5 HW1 Solutions

- Prove the Jensen Inequality using the existence of a subgradient.

Take the expected value of the Subgradient Inequality at $E_D(D)$

$$f(D) \geq f(E_D(D)) + \partial f(E_D(D))(D - E_D(D))$$

and observe that the second term becomes zero.

- Suppose that D_1 and D_2 are two random variables and $E(D_2) = 0$. Let f be a hypothetical convex cost function for a system.
 - a) Establish that $E_{D_1, D_2} f(D_1 + D_2) \geq E_{D_1} f(D_1)$. Explain if you need D_1 and D_2 to be independent.

Proceed as

$$E_{D_1, D_2} f(D_1 + D_2) = E_{D_1} E_{D_2} f(D_1 + D_2) \geq E_{D_1} f(D_1 + E_{D_2} D_2) = E_{D_1} f(D_1)$$

where inequality follows from the Jensen inequality. Independence is required. Otherwise, let $f(x) = x^2$ and $D_2 = -D_1$, clearly the inequality does not hold.

- b) Use the above inequality to intuitively explain that the value of information is always nonnegative.

You can visualize $E_{D_1, D_2} f(D_1 + D_2)$ as the cost when D_1 is the base demand and D_2 as a random error. When $D_2 = 0$ is observed at a later, the cost decreases to $E_{D_1} f(D_1)$. Thus the value of observing D_2 is nonnegative because of a).

- c) Use a) to argue that $D_1 + D_2$ is stochastically more variable than D_1 .

Since $E(D_1 + D_2) = E(D_1)$, it suffices to have $E_{D_1, D_2} f(D_1 + D_2) \geq E_{D_1} f(D_1)$ for all convex functions f , which is established in a).

- Consider the two-period additive uncertainties framework. In addition, suppose that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Let us see if we can make this framework equivalent to Bivariate updates. Suppose that we respectively observe $X_1 = x_1$ and $D_Y = d_Y$ in additive uncertainties and Bivariate frameworks where $x_1 = d_Y$. To establish equivalency, we want to ensure that $[D_{0,2}]$ and $[D_X]$ have the same distribution both before and after the observations x_1 and d_Y . How should D_X and D_Y be chosen (their distribution, means, variances and correlation). Express these parameters in terms of $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$. Check if with your choice of parameter values $\text{Cov}(D_{0,2}, X_1) = \text{Cov}(D_X, D_Y)$.

Since we have Normal distributions for all random variables. Make sure that means and variances are the same. We set $x_1 + \mu_2 = \mu_X + \rho(\sigma_X/\sigma_Y)(x_1 - \mu_Y)$, $\sigma_2^2 = \sigma_X^2(1 - \rho^2)$, $\mu_1 + \mu_2 = \mu_X$, $\sigma_1^2 + \sigma_2^2 = \sigma_X^2$. Solving these equations for every possible x_1 yields $\mu_x = \mu_1 + \mu_2$, $\sigma_X = \sqrt{\sigma_1^2 + \sigma_2^2}$, $\mu_Y = \mu_1$, $\sigma_Y = \sigma_1$ and $\rho = \sqrt{\sigma_1^2/(\sigma_1^2 + \sigma_2^2)}$. With these values, covariances are equal.