Stat 8931 (Exponential Families) Lecture Notes

The Eggplant that Ate Chicago (the Notes that Got out of Hand)

Charles J. Geyer

November 2, 2016

1 Extended Real Numbers

The extended real number system $\overline{\mathbb{R}}$ consists of the real numbers plus two additional points $+\infty$ and $-\infty$. The topology is obvious for the most part. $\overline{\mathbb{R}}$ is a compact set (every sequence has a convergent subsequence). So is the arithmetic, the only non-obvious operations being $0 \cdot \infty$, which we define to be zero (like they do in measure theory), and $\infty - \infty$, which we leave undefined (like division by zero).

An extended-real-valued function f on a metric space is lower semicontinuous (LSC) if

$$\liminf_{n \to \infty} f(x_n) \geq f(x), \qquad \text{whenever } x_n \to x.$$

Theorem 1. *An LSC extended-real-valued function achieves its minimum over any nonempty compact metric space.*

Proof. If $f(x) = -\infty$ for any x, the minimum is achieved.

If $f(x) = +\infty$ for every x, the minimum is achieved.

Otherwise, let K be the compact metric space that is the domain of f and define

$$\alpha = \inf_{x \in K} f(x).$$

Then we have two cases, either $\alpha = -\infty$ or $\alpha \in \mathbb{R}$.

In the first of these, choose x_n such that $f(x_n) \leq -n$, for all n. By the Bolzano-Weierstrass theorem, this has a convergent subsequence, say $x_{n_k} \to x$. By LSC we must have $f(x) = -\infty$, contrary to the "otherwise" above. So this case ($\alpha = -\infty$ but $f(x) \neq -\infty$ for all x) cannot happen.

In the second of these, choose x_n such that $f(x_n) \leq \alpha + 1/n$, for all n. By the Bolzano-Weierstrass theorem, this has a convergent subsequence, say $x_{n_k} \to x$. By LSC we must have $f(x) = \alpha$, and the minimum is achieved. □

2 Convex Sets and Functions

A subset S of a vector space E is *convex* if

$$tx + (1-t)y \in S, \qquad \text{whenever } x, y \in S \text{ and } 0 < t < 1.$$

An extended-real-valued function f on a vector space E is *convex* if

$$f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y), \\ \text{whenever } x, y \in \operatorname{dom} f \text{ and } 0 < t < 1, \quad (1)$$

where

$$\operatorname{dom} f = \{\, x \in E : \operatorname{dom} f < +\infty \,\}.$$

The set $\operatorname{dom} f$ is called the *effective domain* of f when f is a convex function (of course, the *domain* of f is E by definition).

The point of the restriction $x, y \in \operatorname{dom} f$ in (1) is to avoid $\infty - \infty$. The inequality (1) is called the *convexity inequality.*

The point of allowing $+\infty$ as a possible value is to allow for constraints. If we want to solve the optimization problem minimize g subject to the constraint that the solution lie in the set K, this is the same as the unconstrained problem of minimizing the function f defined

$$f(x) = \begin{cases} g(x), & x \in K \\ +\infty, & x \notin K \end{cases}$$

(the constraint set is incorporated in the function itself).

We are mostly not interested in functions that take the value $-\infty$ (this value is allowed only to make $\overline{\mathbb{R}}$ a compact set). Nor are we interested in functions that are everywhere $+\infty$. We rule out such functions with the term "proper." An extended-real-valued convex function is *proper* if it is nowhere $-\infty$ and somewhere finite (so not everywhere $+\infty$).

A function f has a *local minimum* at a point x (and this point is called a *local minimizer*) if there exists a neighborhood W of x such that f achieves its minimum over W at x. We know from calculus that a necessary condition for this is that the gradient vector of f be zero at x (assuming f is differentiable at x). We also know from calculus that a sufficient condition for this is that the Hessian matrix of f be positive definite at x (assuming f is twice differentiable at x).

For contrast with the term "local minimum" we say a function f has a *global minimum* at a point x (and this point is called a *global minimizer*) if f achieves its minimum (over its whole domain) at x. Calculus is no help for finding global minima.

Theorem 2. *Every local minimizer of an extended-real-valued convex function is a global minimizer so long as the local minimum is finite.*

Proof. Let f be the convex function in question. Proof by contradiction. Suppose x is a local minimizer of f and minimizes it over a neighborhood W of x. And suppose x fails to be a global minimizer of f, that is, there exists a point y such that $f(y) < f(x)$, and (by assumption) $f(x) < \infty$. By the convexity inequality, for $0 < t < 1$

$$f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y) < f(x)$$

and for t sufficiently close to 1 we have $tx + (1-t)y \in W$, which contradicts x being a local minimizer. □

Theorem 3. *For an extended-real-valued proper convex function, every point where the gradient vector is zero is a global minimizer.*

Proof. Let f be the convex function in question. Proof by contradiction. Suppose $\nabla f(x) = 0$. And suppose x fails to be a global minimizer of f, that is, there exists a point y such that $f(y) < f(x)$, and (by definition of differentiability) $f(x) < \infty$. By the convexity inequality, for $0 < t < 1$

$$f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y). \tag{2}$$

Define g on $\mathbb{R}$ by

$$g(t) = f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y)$$

then g is differentiable at 1 where the derivative is zero by the assumption that $\nabla f(x) = 0$. But calculating a one-sided derivative of g at 1 using (2) gives

$$\begin{aligned}
\lim_{t \uparrow 1} \frac{g(t) - g(1)}{1-t} &\le \lim_{t \uparrow 1} \frac{tf(x) + (1-t)f(y) - f(x)}{1-t} \\
&= \lim_{t \uparrow 1} [f(y) - f(x)] \\
&= \lim_{t \uparrow 1} [f(y) - f(x)] f(y) - f(x) \\
&< 0
\end{aligned}$$

the contradiction. □

A proper extended-real-valued convex function is *strictly convex* if the convexity inequality holds with strict inequality, that is,

$$f\bigl(tx + (1-t)y\bigr) < tf(x) + (1-t)f(y), \qquad \text{whenever } x, y \in \operatorname{dom} f \text{ and } 0 < t < 1, \tag{3}$$

Theorem 4. *For a strictly convex proper convex function, the global minimizer is unique if it exists.*

Proof. Let f be the convex function in question. Proof by contradiction. Suppose x is a global minimizer of f but is not unique, that is, there exists a different global minimizer y. Both $f(x)$ and $f(y)$ are finite because f is proper, and $f(x) = f(y)$ because both are global minimizers. By the strict convexity inequality, for $0 < t < 1$

$$f\bigl(tx + (1-t)y\bigr) < tf(x) + (1-t)f(y) < f(x) = f(x) = f(y)$$

which contradicts x and y being global minimizers. □

3 Directions of Recession and Constancy

Suppose f is an LSC, proper, extended-real-valued convex function on a finite-dimensional vector space E. A direction δ in E is called a *direction of recession* of f if there exists an $x \in E$ such that

$$\limsup_{s \to \infty} f(x + s\delta) < \infty. \tag{4}$$

Theorem 5. *The vector δ is a direction of recession of the LSC proper convex function f if and only if the function $s \mapsto f(y + s\delta)$ is actually nonincreasing on $\mathbb{R}$ for all $y \in E$.*

Proof. Suppose there is an $x \in E$ such that (4) holds. Then we claim the function $s \mapsto f(x + s\delta)$ is actually nonincreasing. We show this by proof by contradiction. Suppose there exist s_1 and s_2 such that $s_1 < s_2$ and $f(x + s_1\delta) < f(x + s_2\delta)$. This implies $f(x + s_1\delta)$ is finite. If $f(x + s_2\delta) = \infty$, then convexity requires $f(x + s_3\delta) = \infty$, whenever $s_2 < s_3$. But that would violate (4). Hence $f(x + s_2\delta)$ is also finite. Then the convexity inequality implies for $s_3 > s_2$ that

$$f(x + s_2\delta) \le \frac{s_3 - s_2}{s_3 - s_1} f(x + s_1\delta) + \frac{s_2 - s_1}{s_3 - s_1} f(x + s_3\delta)$$

so

$$
\begin{aligned}
f(x + s_3\delta) &\geq \frac{s_3 - s_1}{s_2 - s_1} \left[f(x + s_2\delta) - \frac{s_3 - s_2}{s_3 - s_1} f(x + s_1\delta) \right] \\
&= \frac{s_3 - s_1}{s_2 - s_1} f(x + s_2\delta) - \frac{s_3 - s_2}{s_2 - s_1} f(x + s_1\delta) \\
&= f(x + s_2\delta) + \frac{s_3 - s_1}{s_2 - s_1} \bigl[f(x + s_2\delta) - f(x + s_1\delta) \bigr]
\end{aligned}
$$

and, since we are assuming the term in square brackets is strictly positive, this goes to infinity as $s_3 \to \infty$ contradicting (4). This completes the proof that $s \mapsto f(x + s\delta)$ is nonincreasing.

Now consider another point $y \in E$ and real numbers s_1 and s_2 such that $s_1 < s_2$. We need to show that $f(y + s_1\delta) \geq f(y + s_2\delta)$. If $f(y + s_1\delta) = \infty$, there is nothing to prove. So assume $f(y + s_1\delta)$ is finite. We also know from the first part of this proof that there exists s_0 such that $f(x + s_0\delta)$ is finite (otherwise (4) would not hold) and $f(x + s\delta) \leq f(x + s_0\delta)$ for all $s \geq s_0$.

So consider the point

$$
(1 - u)(y + s_1\delta) + u(x + s\delta). \tag{5}
$$

Choose u as a function of s so that this converges to $y + s_2\delta$ as $s \to \infty$. Clearly we need $u \to 0$ and $us \to s_2 - s_1$. So we can choose $u = (s_2 - s_1)/(s - s_0)$ for $s > s_0$. The convexity inequality says

$$
\begin{aligned}
f\bigl((1 - u)(y + s_1\delta) + u(x + s\delta)\bigr) &\leq (1 - u) f(y + s_1\delta) + u f(x + s\delta) \\
&\leq (1 - u) f(y + s_1\delta) + u f(x + s_0\delta) \\
&\to f(y + s_1\delta)
\end{aligned}
$$

as $s \to \infty$ (and $u \to 0$ and (5) converges to $y + s_2\delta$). Then f being LSC implies

$$
\liminf_{s \to \infty} f\bigl((1 - u)(y + s_1\delta) + u(x + s\delta)\bigr) \geq f(y + s_2\delta)
$$

so $f(y + s_1\delta) \geq f(y + s_2\delta)$, and this finishes the proof that $s \mapsto f(y + s\delta)$ is nonincreasing for all y. □

Theorem 6. *An LSC proper convex function having no nonzero directions of recession achieves its infimal value.*

Proof. Let f be the function in question. Choose a point x_0 such that $f(x_0)$ is finite (there is one since f is proper). Let B_R denote the closed ball of radius R centered at x_0, and let S_R denote its boundary. This implies we impose a norm on E, but it does not matter which one since all norms are equivalent (induce the same convergent sequences). Since S_R is a compact set by the Heine-Borel theorem, f achieves its minimum over S_R by LSC, say at x_R, so

$$f(x_R) = \inf\{\, f(x) : x \in S_R \,\}.$$

We claim $f(x_R) \to \infty$ as $R \to \infty$. Again we use proof by contradiction. Assume there exists a subsequence R_n such that $R_n \to \infty$ but $f(x_{R_n})$ has a finite upper bound, say M. Define $v_R = (x_R - x_0)/\|x_R - x_0\|$, where $\|\cdot\|$ denotes the chosen norm for E. Since $v_R \in S_1$ for all R, by the Bolzano-Weierstrass theorem v_{R_n} has a convergent subsequence

$$v_{R_{n_k}} \to v.$$

For $s > 0$, the point

$$x_0 + s\frac{x_{R_{n_k}} - x_0}{\|x_{R_{n_k}} - x_0\|}$$

converges to $x_0 + sv$ and is a convex combination $(1-u)x_0 + ux_{R_{n_k}}$ with $u = s/\|x_{R_{n_k}} - x_0\|$ when k is large enough so that $u < 1$. Hence

$$f\bigl((1-u)x_0 + ux_{R_{n_k}}\bigr) \le (1-u)f(x_0) + uf(x_{R_{n_k}}) \le f(x_0) + M$$

so by f being LSC

$$f(x_0 + sv) \le \liminf_{k\to\infty} f\bigl((1-u)x_0 + ux_{R_{n_k}}\bigr) \le f(x_0) + M$$

holds for all $s > 0$ and this implies that v is a direction of recession, which contradicts the assumption of the theorem. So this proves $f(x_R) \to \infty$ as $R \to \infty$.

Hence there exists R such that $f(x_R) > f(x_0)$. Consider any $y \in E$ in the exterior of B_R. Define

$$z = x_0 + R\frac{y - x_0}{\|y - x_0\|}$$

the point in S_R on the line segment between x_0 and y. Let $t = R/\|y - x_0\|$ so $z = (1-t)x_0 + ty$ and the convexity inequality implies $f(z) \le (1-t)f(x_0) + tf(y)$ hence

$$f(y) \ge f(x_0) + \frac{f(z) - f(x_0)}{t} \ge f(x_0) + \frac{f(x_R) - f(x_0)}{t} \ge f(x_R)$$

so no point outside B_R can minimize f. Hence the minimum of f is the minimum of f over the compact set B_R, which is achieved by LSC. □

A direction δ in E is called a *direction of constancy* of an LSC proper convex function f if the function $s \mapsto f(x + s\delta)$ is a constant function on $\mathbb{R}$ for all $x \in E$ (a different constant function for each x). Clearly, every direction of recession is a direction of constancy, and δ is a direction of constancy if and only if both δ and $-\delta$ are directions of recession. Clearly every linear combination of directions of constancy is another direction of constancy, so the set of all directions of constancy is a vector subspace of E called the *constancy space* of f.

Theorem 7. *An LSC proper convex function for which every direction of recession is a direction of constancy achieves its infimal value.*

Proof. Let C be the constancy space of f and decompose E as the direct sum $C \oplus D$. The restriction of f to D has no directions of recession and hence achieves its infimum by Theorem 6. By definition of direct sum any $x \in E$ can be uniquely written $y + z$ with $y \in C$ and $z \in D$. Since C is the constancy space, we have $f(x) = f(z)$. So if f achieves its minimum over D at z, then it also achieves its minimum over E there. □

4 Concave Functions

Turning everything upside down gives concave functions. A function f is *concave* if and only if $-f$ is convex. It is *strictly concave* if and only if $-f$ is strictly convex. it is *proper* if and only if $-f$ is proper. If f is concave, then its *effective domain* is $\operatorname{dom} f = \operatorname{dom}(-f)$. A vector δ is a *direction of recession* (resp., direction of constancy) of the concave function f if and only if it is a direction of recession (resp., direction of constancy) of the convex function $-f$ This makes all of our theorems about minimizing convex functions have obvious analogs about maximizing concave functions.

5 Exponential Families

Let E be a finite-dimensional vector space and E^* its dual space and let $\langle \,\cdot\,, \,\cdot\, \rangle$ be the bilinear form placing these spaces in duality, defined by

$$\langle y, \theta \rangle = \theta(y), \qquad y \in E \text{ and } \theta \in E^*. \tag{6}$$

Let λ be a positive Borel measure on E ("positive" meaning $\lambda(A) \geq 0$ for all measurable sets A and $\lambda(E) > 0$). We are getting a little bit ahead of ourselves in that we assume we know what the Borel sigma-field of E is, when we will not deal with this formally until Section 8.14 below. We just take it for granted that we do know for now. The *log Laplace transform* of λ is the function $c : E^* \to \overline{\mathbb{R}}$ given by

$$c(\theta) = \log \int e^{\langle y, \theta \rangle} \, \lambda(dy), \qquad \theta \in E^*.$$

Since the integrand is strictly positive, the integral cannot be zero, so a log Laplace transform never takes the value $-\infty$. We define it to have the value $+\infty$ whenever the integral does not exist.

Theorem 8. *A log Laplace transform is a proper convex function unless it is the constant function always equal to $+\infty$. When proper, it is LSC. When proper, it is strictly convex unless λ is concentrated on a hyperplane.*

Proof. Convexity follows from Hölder's inequality, LSC from Fatou's lemma. Strict convexity follows from the conditions for equality in Hölder's inequality. □

Let c be the log Laplace transform of the positive Borel measure λ on the finite-dimensional vector space E. If c is proper, then its effective domain

$$\Theta = \operatorname{dom} f$$

is nonempty, and for each $\theta \in \Theta$

$$f_\theta(y) = e^{\langle y, \theta \rangle - c(\theta)}, \qquad y \in E \tag{7}$$

defines a probability density with respect to λ. The set of all such probability densities is called the *full standard exponential family* (of densities) *generated by* λ.

A proper subfamily would be a (non-full) standard exponential family.

A family of densities with respect to a positive measure μ on an arbitrary measurable space having densities of the form

$$p_\theta(\omega) = e^{\langle Y(\omega), \theta \rangle - c(\theta)}$$

is a (not standard) exponential family. Change of variable, so that y is the variable, gives a standard family. The details are in Chapter 1 of my thesis (Geyer, 1990).

6 Convex Support

The *convex support* of a Borel measure on a finite-dimensional vector space is the smallest closed convex set that supports it (perhaps this should be called "closed convex" support, but just "convex" is the accepted usage). Before we use this concept, we have to assure ourselves that it exists.

It is easily shown that the intersection of convex sets is convex, and it is an axiom of topology that the intersection of closed sets is closed. Hence for any Borel measure λ on a finite-dimensional vector space E the intersection K of all closed convex sets that support λ exists and is closed and convex. The remaining question is whether K supports λ. To prove this we use the fact that E is a second countable topological space (every open set is a union of open sets in a countable family $\mathcal{O}$ called the "countable basis" of the topology for E; for the countable basis one can take all open balls whose radii are rational and whose centers are vectors having rational coordinates). Let C denote the intersection of all closed sets that support λ. The complement of C is open and is the union of a countable number of sets that have λ measure zero. Hence $\lambda(C^c) = 0$. But $C \subset K$, so we also have $\lambda(K^c) = 0$.

The *support function* of the set K is the function $\sigma_K : E \to \overline{\mathbb{R}}$ defined by

$$\sigma_K(\eta) = \sup_{x \in K} \langle x, \eta \rangle.$$

The "support" in "support function" has nothing to do with K being a support of λ. Rather it has to do with the notion of "supporting hyperplane" in convex geometry. Hyperplanes of the form

$$\{\, y \in E : \langle y, \eta \rangle = \sigma_K(\eta) \,\}$$

are the supporting hyperplanes in question.

7 Maximum Likelihood in Exponential Families

Now we use what we have learned about convex minimization (and concave maximization) to understand maximum likelihood in exponential families. For a concave function l we say δ is a direction of recession of l if and only if δ is a direction of recession of the convex function $-l$.

Lemma 9. *Consider a positive Borel measure λ on a finite-dimensional vector space E having log Laplace transform c not identically $+\infty$ and having* *convex support K. Then for any $\theta \in \operatorname{dom} c$, any $\eta \in E^*$, and any $a \in \mathbb{R}$*

$$c(\theta + s\eta) - as \to \begin{cases} -\infty, & a > \sigma_K(\eta) \\ c_H(\theta), & a = \sigma_K(\eta) \\ +\infty, & a < \sigma_K(\eta) \end{cases}$$

as $s \to \infty$, where

$$H = \{\, y \in E : \langle y, \delta \rangle = \sigma_K(\eta) \,\}$$

and

$$c_H(\theta) = \log \int_H e^{\langle y, \theta \rangle} \, \lambda(dy)$$

(if the integral is zero, then $c_H(\theta) = -\infty$; if the integral does not exist, then $c_H(\theta) = +\infty$).

Note that c_H is the log Laplace transform of the measure obtained by restricting λ to H.

Proof. First we do the case where $\sigma_K(\eta)$ is finite and $\lambda(H) > 0$. Then λ is concentrated on the half space

$$A = \{\, y \in E : \langle y, \eta \rangle \le \sigma_K(\eta) \,\},$$

and, if $a = \sigma_K(\eta)$, then

$$\begin{aligned} c(\theta + s\eta) - as &= \log \int e^{\langle y, \theta + s\eta \rangle - as} \, \lambda(dy) \\ &= \log \int_A e^{\langle y, \theta \rangle + s[\langle y, \eta \rangle - a]} \, \lambda(dy) \end{aligned}$$

and the integrand is a decreasing function of s, which converges to $e^{\langle y, \theta \rangle} I_H(y)$ as $s \to \infty$. Hence

$$c(\theta + s\eta) - as \to c_H(\theta) \tag{8}$$

as asserted by the theorem, and, moreover, $c_H(\theta)$ is finite because of the assumption $\lambda(H) > 0$ and because of $c_H(\theta) \le c(\theta)$. The other two cases of the limit asserted by the theorem follow immediately from this case.

Second we do the case where $\sigma_K(\eta)$ is finite and $\lambda(H) = 0$. The argument proceeds as above until we obtain (8), but now $c_H(\theta) = -\infty$. Now we can get only one other case of the limit asserted by the theorem: that (8) also holds when $a > \sigma_K(\eta)$.

In case $a < \sigma_K(\eta)$, we have

$$e^{\langle y, \theta + s\eta \rangle - as} \to \begin{cases} 0, & \langle y, \eta \rangle < a \\ e^{\langle y, \theta \rangle}, & \langle y, \eta \rangle = a \\ +\infty, & \langle y, \eta \rangle > a \end{cases}$$

and, since this is infinite on a set of positive λ measure,

$$c(\theta + s\eta) - as = \log \int e^{\langle y, \theta + s\eta \rangle - as} \, \lambda(dy) \to +\infty$$

by monotone convergence (the convergence is increasing on some set, decreasing on another set, and constant on yet another set).

Third we do the case where $\sigma_K(\eta) = +\infty$. Now there is only one case in the limit asserted by the theorem. We can only have $a < \sigma_K(\eta)$. The proof of this case proceeds the same as the proof of the $a < \sigma_K(\eta)$ case in the second part above.

Lastly, we note that $\sigma_K(\eta) = -\infty$ is not possible so long as K is nonempty, which is required by the assumption that λ is a positive measure. □

Theorem 10. *Consider a full exponential family with canonical statistic Y and observed value of the canonical statistic y. A direction η in the parameter space is a direction of recession of the log likelihood if and only if $\langle Y - y, \eta \rangle \le 0$ almost surely.*

Before starting the proof, we note that "almost surely" refers to any distribution in the family (they all have the same support, which is the support of the generating measure λ, because all of the densities (7) are everywhere positive). We also note that the condition $\langle Y - y, \eta \rangle \le 0$ almost surely can be rewritten using the notation introduced in the preceding section as follows. We have $Y \in K$ almost surely hence $\langle Y, \eta \rangle \le \sigma_K(\eta)$ almost surely. Hence we have $\langle Y - y, \eta \rangle \le 0$ almost surely if and only if we have $\sigma_K(\eta) \le \langle y, \eta \rangle$.

Proof. The log likelihood is

$$l(\theta) = \langle y, \theta \rangle - c(\theta)$$

so

$$l(\theta + s\eta) = \langle y, \theta \rangle + s\langle y, \eta \rangle - c(\theta + s\eta).$$

Applying Lemma 9 we obtain

$$l(\theta + s\eta) \to \begin{cases} +\infty, & \langle y, \eta \rangle > \sigma_K(\eta) \\ \langle y, \theta \rangle - c_H(\theta), & \langle y, \eta \rangle = \sigma_K(\eta) \\ -\infty, & \langle y, \eta \rangle < \sigma_K(\eta) \end{cases}$$

Applying the definition of direction of recession for concave functions (the limit is not $-\infty$), we see that η is a direction of recession in case $\langle y, \eta \rangle > \sigma_K(\eta)$ or in case $\langle y, \eta \rangle = \sigma_K(\eta)$. The latter is because of $c_H(\theta) \le c(\theta)$ and because $c(\theta) < \infty$ by assumption. So we cannot have $c_H(\theta) = +\infty$ and the limit $-\infty$ in the middle case.

Thus we do indeed have η is a direction of recession if and only if $\langle y, \eta \rangle \geq \sigma_K(\eta)$, which, as the note preceding the proof says, is the same as $\langle Y - y, \eta \rangle \leq 0$ almost surely. □

Corollary 11. *Consider a full exponential family with canonical statistic Y and observed value of the canonical statistic y. A direction η in the parameter space is a direction of constancy of the log likelihood if and only if $\langle Y - y, \eta \rangle = 0$ almost surely.*

Theorem 12. *Consider a full exponential family with canonical parameter θ. A direction η in the parameter space is a direction of constancy of the log likelihood if for some θ in the canonical parameter space Θ and some $s \neq 0$, the parameter values θ and $\theta + s\eta$ correspond to the same distribution, in which case this is true for every $\theta \in \Theta$ and every $s \in \mathbb{R}$.*

Proof. Suppose θ and $\theta + s\eta$ correspond to the same distribution and $s \neq 0$. Then we must have $f_\theta(y) = f_{\theta + s\eta}(y)$ for almost all y with respect to the generating measure λ. That is, we must have

$$\langle y, \theta \rangle - c(\theta) = \langle y, \theta + s\eta \rangle - c(\theta + s\eta),$$

for almost all y, where c is the cumulant function, or, equivalently,

$$s\langle y, \eta \rangle = c(\theta + s\eta) - c(\theta)$$

for almost all y. This says $\langle Y, \eta \rangle$ is almost surely constant. And that constant must be $\langle y, \eta \rangle$, where y is the observed value of the canonical statistic.

Conversely, assume $\langle Y, \eta \rangle = a$ almost surely for some constant a. Then

$$\begin{aligned} c(\theta + s\eta) &= \log \int e^{\langle y, \theta + s\eta \rangle} \, \lambda(dy) \\ &= as + \log \int e^{\langle y, \theta \rangle} \, \lambda(dy) \\ &= as + c(\theta) \end{aligned}$$

holds for all $\theta \in \Theta$ and all $s \in \mathbb{R}$. But then

$$\begin{aligned} f_{\theta + s\eta}(y) &= e^{\langle y, \theta + s\eta \rangle - c(\theta + s\eta)} \\ &= e^{\langle y, \theta \rangle + as - [as + c(\theta)]} \\ &= e^{\langle y, \theta \rangle - c(\theta)} \\ &= f_\theta(y) \end{aligned}$$

holds for all $\theta \in \Theta$, for all $s \in \mathbb{R}$, and for almost all y with respect to λ. □

Because of Theorem 12 we can always eliminate directions of constancy by reparameterization. In fact if we choose a "minimal" canonical parameterization, there will be no directions of constancy. But, as we shall see, there are many reasons for allowing directions of constancy. And we have just seen they do no harm. They may be a computational nuisance, but not a theoretical nuisance. Directions of constancy are the only kind of non-identifiability an exponential family can have. And they are completely understood theoretically.

Theorem 13. *For a full exponential family, if η is a direction of recession of the log likelihood l this is not a direction of constancy of l, then for every θ in the canonical parameter space the function $s \mapsto l(\theta + s\eta)$ is strictly increasing on the interval where it is finite.*

Proof. Proof by contradiction. Assume that for some θ in the canonical parameter space $s \mapsto l(\theta + s\eta)$ is not strictly increasing on the interval where it is finite. This means there exist s_1 and s_2 such that $s_1 < s_2$

$$l(\theta + s_1\eta) = l(\theta + s_2\eta) < \infty.$$

By concavity we must have

$$l(\theta + s\eta) = l(\theta + s_1\eta) = l(\theta + s_2\eta), \qquad s_1 < s < s_2. \tag{9}$$

Consider the one-parameter submodel given by the parameterization $\theta + s\eta$, so s is the submodel canonical parameter and $\langle Y, \eta \rangle$ is the submodel canonical statistic. Let c be the cumulant function of the full model. Then the submodel cumulant function is defined by

$$c_{\text{sub}}(s) = c(\theta + s\eta),$$

and the submodel log likelihood is defined by

$$l_{\text{sub}}(s) = l(\theta + s\eta).$$

We know from moment generating function theory that

$$\begin{aligned} c'_{\text{sub}}(s) &= E_{\theta + s\eta}\{\langle Y, \eta \rangle\} \\ c''_{\text{sub}}(s) &= \operatorname{var}_{\theta + s\eta}\{\langle Y, \eta \rangle\} \end{aligned}$$

hold for all s in the interior of the submodel canonical parameter space, which includes the interval (s_1, s_2). On this interval, (9) implies l''_{sub} is zero, hence c''_{sub} is zero, hence $\operatorname{var}_{\theta + s\eta}\{\langle Y, \eta \rangle\}$ is zero, hence $\langle Y, \eta \rangle$ is constant almost surely, which contradicts η not being a direction of constancy by Corollary 11. □

Theorem 14. *For a full exponential family, the maximum likelihood estimate for the canonical parameter exists if and only if every direction of recession of the log likelihood is a direction of constancy of the log likelihood.*

Proof. One direction is an obvious corollary of (the concavity analog of) Theorem 7. The other direction is Theorem 13. (A strictly increasing function can never achieve its maximum. If l is the log likelihood and η is a direction of recession that is not a direction of constancy, then $l(\theta) < l(\theta + \eta)$ for every θ in the canonical parameter space, so no θ can maximize the log likelihood.) □

Note that one direction has nothing to do with any property of exponential families other than that the log likelihood is proper, upper semicontinuous, and concave. It is all convexity theory. But the other direction uses more about the theory of exponential families. It does not follow merely from the log likelihood being proper, upper semicontinuous, and concave.

Theorem 15. *If the maximum likelihood estimate for the canonical parameter of a full exponential family is not unique, then all maximum likelihood estimates correspond to the same probability distribution.*

Thus we have uniqueness where it really counts.

Proof. The proof is almost the same as the proof of Theorem 13. Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates (MLE). Then the concavity inequality says

$$l\bigl(s\hat{\theta}_1 + (1 - s)\hat{\theta}_2\bigr) \geq s l(\hat{\theta}_1) + (1 - s) l(\hat{\theta}_2), \qquad 0 < s < 1, \tag{10}$$

but $\hat{\theta}_1$ and $\hat{\theta}_2$ being MLE requires that we have equality in (10). Then differentiating with respect to s gives

$$\begin{aligned} 0 &= \frac{d^2}{ds^2} l\bigl(s\hat{\theta}_1 + (1-s)\hat{\theta}_2\bigr) \\ &= -\operatorname{var}_{s\hat{\theta}_1+(1-s)\hat{\theta}_2}\{\langle Y, \hat{\theta}_1 - \hat{\theta}_2\rangle\} \end{aligned}$$

holds when $0 < s < 1$ and in the variance $\hat{\theta}_1$ and $\hat{\theta}_2$ are being considered as constants rather than as functions of the data (only Y is random). But this implies $\langle Y, \hat{\theta}_1 - \hat{\theta}_2\rangle$ is constant almost surely, hence $\hat{\theta}_1 - \hat{\theta}_2$ is a direction of constancy, hence that $\hat{\theta}_1$ and $\hat{\theta}_2$ correspond to the same distribution by Corollary 11 and Theorem 12. □

7.1 Counterexample

This finishes our discussion of the theory of the existence of maximum likelihood estimators for the canonical parameter of a full exponential family. Note that this theory is good even for non-regular families. The MLE that are guaranteed to exist by the theory above may occur on the boundary of the canonical parameter space.

Let us cook up a pathological example to illustrate this phenomenon. Examples interesting in practice are given by Geyer and Møller (1994) and Geyer (1999). A one-dimensional example will suffice.

Beyond regular exponential families are steep exponential families, which are defined in Barndorff-Nielsen (1978) in the text just before Theorem 8.2, which says every regular family is steep. Steep families have the observed equals expected property of maximum likelihood estimation (Barndorff-Nielsen, 1978, Theorem 9.14, the text preceding it, and Corollary 9.6). Thus they are almost as nice as regular exponential families. Thus to get an example of a really badly behaving exponential family, we want a non-steep family.

One example of a non-steep exponential family is given by Barndorff-Nielsen (1978, Example 9.10), but it is continuous. We want a discrete example.

Steep families have the property that the mean goes to infinity as the canonical parameter approaches the boundary of the canonical parameter space so we want a generating measure that has a mean but does not have a moment generating function. For our example take the measure on the positive integers having mass function

$$\lambda(x) = x^{-\alpha}, \qquad x = 1, 2, 3, \dots.$$

For $\alpha > 1$ this measure is proportional to a probability distribution. For $\alpha > 2$ this probability distribution has a mean. For $\alpha > 3$ this probability distribution has a variance. So we take $\alpha = 4$. For future reference the mean and standard deviation of this distribution are approximately

```
> x <- seq(1, 1e7)
> p <- 1 / x^4
> p <- p / sum(p)
> mu <- sum(x * p)
> sigma.sq <- sum((x - mu)^2 * p)
> sigma <- sqrt(sigma.sq)
> mu

[1] 1.110627

> sigma

[1] 0.5350947
```

(and third moments do not exist).

The Laplace transform of this measure is given by

$$C(\theta) = \sum_{n=1}^{\infty} n^{-\alpha} e^{n\theta}$$

is finite if and only if $\theta \leq 0$, so

$$\Theta = \{\, \theta \in \mathbb{R} : \theta \leq 0 \,\}$$

is the canonical parameter space of the full exponential family generated by this measure.

Because the full canonical parameter space is a closed set, this exponential family is non-regular. Because the mean exists for the distribution for the parameter value on the boundary (0), the family is also non-steep.

So how does maximum likelihood work? From what we know about directions of recession, if the observed data is $x = 1$, then (because this x is on the boundary of the convex support), the MLE is "at infinity" and is concentrated at this point ($x = 1$). If the observed data is a possible mean value then the MLE is the corresponding canonical parameter value. But the mean values range from 1 to 1.1106, so no data can match a possible mean value except $x = 1$.

We show that for observed data $x > 1$ the MLE is 0 (that is, the boundary distribution). We have $l'(\theta) = x - E_\theta(X)$ for all $\theta < 0$, and, since we are considering data $x \geq 2$ and means $E_\theta(X) < 1.1106$, we have $l'(\theta) > 0$ for all $\theta < 0$, hence the likelihood is strictly increasing and must be maximized at the boundary where $\theta = 0$.

Now consider data $X_1, \ldots, X_m$ that are IID from some distribution in this family. We again have an exponential family, but now the canonical statistic is

$$Y = \sum_{i=1}^{m} X_i,$$

the canonical parameter is the same as before, and the cumulant function is mc, where c is the cumulant function for sample size 1. Now $E_\theta(Y)$ ranges from m to $1.1106 \cdot m$, and the possible data values that are also possible mean values (so we can have observed equals expected) are any integer in this range. If the observed data is y, then we again have a direction of recession -1 if $y = m$, and the MLE distribution is concentrated at m. If the observed data y satisfies $m < y < 1.1106 \cdot m$, then we are in the "regular" case where there is a θ such that $y = E_\theta(Y)$ and this θ is the MLE. And if the observed data y satisfies $y \geq 1.1106 \cdot m$, then we are in the "boundary" case where the MLE is $\theta = 0$.

Notice that the non-steep non-regularity is at the opposite end of the sample space and parameter space from the directions of recession and limits. We have pathology (directions of recession) when we are on the boundary of the convex support ($y = m$ in the example), and we have the other kind of pathology (observed equals expected fails) when we are on the boundary of the full canonical parameter space.

7.2 Mean Value Parameterization of Regular Families

This section is not immediately relevant, but the math is almost the same as some of the other theorems about directions of constancy. So we put it here.

For a regular exponential family, the distribution corresponding to the canonical parameter value θ has a moment generating function

$$t \mapsto e^{c(\theta + t) - c(\theta)},$$

where c is the cumulant function of the family, and cumulant generating function

$$t \mapsto c(\theta + t) - c(\theta).$$

And every distribution has moments and cumulants of the canonical statistic of all orders given by derivatives of these functions evaluated at zero.

In particular the mean and variance are given by

$$
\begin{aligned}
E_\theta(Y) &= \nabla c(\theta) \\
\operatorname{var}_\theta(Y) &= \nabla^2 c(\theta)
\end{aligned}
$$

We learned in the 8112 notes (Geyer, 2013, Lemma 9) that regular exponential families can also be parameterized by mean values, although those notes restrict to the identifiable case. Here we allow non-identifiability (directions of constancy) and see that they do no harm.

If Θ is the canonical parameter space (a subset of E^*) then the function $g : \Theta \to E$ defined by

$$
g(\theta) = \nabla c(\theta), \qquad \theta \in \Theta, \tag{11}
$$

maps the canonical parameter θ to the *mean value parameter* $\mu = g(\theta)$. Lemma 9 of the 8112 notes says that in the case where the canonical parameterization is identifiable (there are no nonzero directions of constancy) that both g and its inverse mapping are infinitely differentiable and Lemma 9 give the formula for the first derivative of g^{-1} that comes from the inverse function theorem of real analysis. But here we drop the identifiability requirement, which means the inverse does not exist.

Theorem 16. *The mean value parameterization of a regular full exponential family is identifiable. If θ_1 and θ_2 are canonical parameter values corresponding to the same mean value parameter value, then $\theta_1 - \theta_2$ is a direction of constancy of the log likelihood of the family.*

Proof. As in the proofs of Theorem 13 and Theorem 15 consider the one-parameter submodel having canonical parameter s and canonical statistic $\langle y, \theta_1 - \theta_2 \rangle$ that is embedded into the full model by the mapping $\theta = s\theta_1 + (1 - s)\theta_2$ for $0 < s < 1$. As in the proof of Theorem 13 we have

$$
\begin{aligned}
c'_{\text{sub}}(s) &= E_{s\theta_1 + (1-s)\theta_2}\{\langle Y, \theta_1 - \theta_2 \rangle\} \\
c''_{\text{sub}}(s) &= \operatorname{var}_{s\theta_1 + (1-s)\theta_2}\{\langle Y, \theta_1 - \theta_2 \rangle\}
\end{aligned}
$$

where

$$
c_{\text{sub}}(s) = c\bigl(s\theta_1 + (1 - s)\theta_2\bigr)
$$

defines the submodel cumulant function (the full model cumulant function being c). Because variances are nonnegative, we see that c'_{sub} is a nondecreasing function. Hence, since the theorem statement assumes the mean value parameter values for corresponding to θ_1 and θ_2 are the same, so must be the submodel parameter values for $s = 0$ and $s = 1$, but then c'_{sub} being nondecreasing implies

$$c'_{\text{sub}}(s) = c'_{\text{sub}}(0) = c'_{\text{sub}}(1), \qquad 0 < s < 1.$$

Differentiating this gives $c''_{\text{sub}}(s) = 0$ for $0 < s < 1$, and this implies $\langle Y, \theta_1 - \theta_2 \rangle$ is constant almost surely, hence $\theta_1 - \theta_2$ is a direction of constancy by Corollary 11. □

8 Affine Spaces

8.1 Introduction

A nonempty subset of a vector space is an *affine subspace* if it is the translate of a vector subspace, a set of the form

$$x + V = \{\, x + v : v \in V \,\}$$

for some vector subspace V.

For technical reasons the empty set is also defined to be an affine subspace. We know that the intersection of closed sets is a closed set, the intersection of convex sets is a convex set, and the intersection of vector subspaces is a vector subspace. We want the analogous property to hold for affine subspaces. But the intersection of affine spaces can be empty (consider parallel lines). So that is why the empty set is considered an affine subspace.

It is also possible to consider affine spaces a mathematical objects in their own right (not as subspaces of some other mathematical object). But vector spaces play a role in their definition in much the same way that fields play a role in the definition of vector spaces.

If we consider affine subspaces as subspaces, we note that the vector space operations do not work for affine subspaces.

If A is an affine subspace and x and y are points in A, then $x+y$ need not be a point of A. If $A = z + V$ for some vector z and some vector subspace V, then $x = z + v_1$ and $y = z + v_2$ for v_1 and v_2 in V. If $x + y \in V$, then $2z + v_1 + v_2 \in V$, which implies $z \in V$, which implies $A = V$, so A is actually a vector subspace rather than a general affine space. (Every vector subspace is an affine subspace too, but an affine subspace of vector space is itself a vector space if and only if it contains the zero vector).

Also, if s is a scalar and x a point in an affine space A, then sx need not be a point of A. If $A = z + V$ as before, so $x = z + v_1$ as before, then $sx \in A$ if and only if $sz \in V$, which implies $z \in V$ and $A = V$ unless $s = 0$, which is not the general case.

This means that we cannot consider the vector space operations to be affine space operations. Some very bright people have thought about this and come up with the following idea. If we have two points x and y in an affine space, it makes sense to subtract them resulting in the vector $y - x$ that is the vector from x to y. Clearly, if $x = z + v_1$ and $y = z + v_2$, as before, then $y - x = v_2 - v_1 \in V$. So $y - x$ is a vector in the vector subspace V that A is a translate of (it is not an element of the affine space). If we write $v = y - x$, then we also write $y = x + v$.

8.2 Formal Definition

Every affine space A comes with a vector space V called its *translation space*. We say A is "over" V in the same way that every vector space is over a field. To be careful about terminology, we always call elements of the affine space "points," elements of its translation space "vectors," and elements of the field the translation space is over "scalars."

The following operations involving points and vectors are defined,

- subtraction of points x and y yields a vector $y - x$,
- addition of a point x and a vector v yields a point $x + v$,

Since addition is usually thought of as commutative we allow ourselves to write $v + x$ instead of $x + v$.

Three axioms govern these operators

(a) If x and y are points and v is a vector, then

$$v = y - x \text{ if and only if } y = x + v.$$

(b) If x is a point and v and w are vectors, then

$$(x + v) + w = x + (v + w).$$

(c) If x is a point and 0 is the zero vector, then

$$x + 0 = x.$$

Every field can be thought of as a vector space over itself. Just let $x + y$ and sx denote the usual field operations. Then the field axioms imply the vector space axioms for these operations.

Similarly every vector space can be thought of as an affine space over itself. Just let $y-x$ and $x+v$ denote the usual vector space operations. Then the vector space axioms imply the affine space axioms for these operations.

It is clear that, if A is an affine space and V is its translation space, then $y \mapsto y - x$ is an invertible map $A \to V$ (its inverse is $v \mapsto x + v$). And we can think of these maps as "identifying" the two spaces. A colloquial way to say this is an affine space is a vector space when you have forgotten where zero is, and an affine space can be turned into a vector space by choosing an arbitrary point to serve as the origin.

But rather than build up affine space theory from the axioms, we shall just consider affine subspaces as subsets of vector spaces with the affine space operations inherited from the vector space ones. This will give us a firm knowledge of the theory without a lot of work.

The only concessions to abstract affine space theory is that when we talk about A as an abstract affine space we won't mention an enclosing vector space explicitly. All operations will involve only A and its translation space

$$V = \{ y - x : x, y \in A \}.$$

For example, when we talk about convex sets in an affine space and convex functions on an affine space, we will not write $tx + (1 - t)y$, $0 < t < 1$ for the elements of the line segment with endpoints x and y. The reason is that scalar multiplication and vector addition are not affine space operations. However, in a vector space we have

$$tx + (1 - t)y = y + t(x - y) \tag{12}$$

and the right-hand side does make sense in an abstract affine space (x and y are points, hence $x - y$ is a vector, hence $t(x - y)$ is a vector (scalar multiple of a vector), hence $y + t(x - y)$ is a point). So our main concession to abstract affine space theory is to always (in the future) write the right-hand side of (12), never the left-hand side.

8.3 Affine Functions

A structure-preserving mapping between objects having a certain mathematical structure is the key idea in modern mathematics.

In linear algebra the structure-preserving mappings are the linear functions, structure-preserving meaning

$$f(x + y) = f(x) + f(y) \tag{13a}$$

and

$$f(tx) = tf(x) \tag{13b}$$

and

$$f(0) = 0 \tag{13c}$$

(it preserves operations and constants). Of course, (13c) is the special case of (13b) when $t = 0$, so it need not be checked in practice, but it is important in theory.

In measure theory and probability theory, the structure-preserving mappings are the measurable maps. There the set-to-set inverse of a map $f : A \to B$ defined by

$$f^{-1}(C) = \{\, x \in A : f(x) \in C \,\} \tag{14}$$

takes measurable subsets of B to measurable subsets of A.

In topology, the structure-preserving mappings are the continuous functions. There the set-to-set inverse of a map $f : A \to B$ defined by (14) takes open subsets of B to open subsets of A. And f itself maps convergent sequences to convergence sequences: $x_n \to x$ implies $f(x_n) \to f(x)$.

It is not immediately clear how this should apply to affine spaces because the operations mix points and vectors. Perhaps we should want to write

$$f(x - y) = f(x) - f(y)$$

and

$$f(x + v) = f(x) + f(v)$$

but this is confusing because sometimes the argument is a point and sometimes it is a vector. We want points to map to points and vectors to map to vectors, so let us use different letters for the two parts of the map.

If A and B are affine spaces having translation spaces U and V, respectively, we say $f : A \to B$ is an *affine function* if there exists a linear function $g : U \to V$ such that

$$g(x - y) = f(x) - f(y), \qquad x \in A \text{ and } y \in A,$$

and

$$f(x + v) = f(x) + g(v), \qquad x \in A \text{ and } v \in U,$$

both hold. It should be obvious that each of these implies the other by the affine space axioms. So we only need to check one.

If A and B are affine spaces having translation spaces U and V, respectively, we say $f : A \to B$ is and *affine function* if for any $y \in A$ the map $g : U \to V$ defined by

$$g(v) = f(y + v) - f(y), \qquad v \in U, \tag{15}$$

is a linear function. (The reason we insist on linearity is because we want to preserve not only the affine space operations but also the vector space operations on the translation spaces.)

Corollary 17. *A function between vector spaces is affine if and only if it is a linear function plus a constant function.*

Proof. Write the function $f : U \to V$ and consider U and V to be their own translation spaces. Then, taking $y = 0$ in the characterization above, we have

$$f(v) = g(v) + f(0), \qquad v \in U,$$

where g is linear, and of course $v \mapsto f(0)$ is a constant function. □

So that gives us the definition of affine function that most people who use the term mean (since they are usually mapping between vector spaces rather than abstract affine subspaces). But, of course, most people call affine functions “linear” functions. It is only in linear algebra that calling f “linear” means $f(0) = 0$.

But if we want to consider $f : A \to B$ to be a function between affine spaces having translation spaces U and V, respectively, (perhaps affine subspaces of vector spaces, perhaps abstract affine spaces) we need the more general definition: f is affine if and only if (15) defines a linear function $U \to V$.

One last thing about affine functions. Recall that the empty set is an affine space. If A is empty, and B is an affine space, then there is exactly one function $A \to B$, the empty function, which has no argument-value pairs, that is, its graph is the empty set. We consider this an affine function, even though we cannot define a linear function from it.

Empty sets are affine spaces, but weird ones. They do not have translation spaces. We cannot consider the empty set its own translation space, because, by definition, every vector space has at least one element, the zero vector.

8.4 Category Theory

The notion of structure-preserving mappings has led to a branch of mathematics called category theory that, so far, has had little influence on statistic (only a few papers). It generalizes the notion of mathematical structure to a very abstract level. A category consists of some things called “objects” and some other things called either “morphisms” or “arrows” that are related as follows. For any two objects A and B there may be some morphisms (arrows) $f : A \to B$ also denoted in pictures

$$A \xrightarrow{f} B \tag{16}$$

The reason for the name “arrow” is obvious. The reason for the name “morphism” is less obvious. It comes by back formation from homomorphism, homeomorphism, isomorphism, and so forth, which are names for various kinds of structure-preserving mappings in various areas.

In (16) A is called the domain and B is called the codomain of f, just as if f is an ordinary function. But (as we shall see — very briefly — later) not all arrows (morphisms) have to be functions.

There is one operation on arrows. Given arrows and objects

$$A \xrightarrow{f} B \xrightarrow{g} C$$

there is an arrow $A \to C$ called the *composition* of f and g, written $g \circ f$. And this operation has to satisfy the *associative law:* given arrows and objects

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D$$

we must have

$$(h \circ g) \circ f = h \circ (g \circ f)$$

(composition is associative).

There is also for each object A an *identity arrow* (identity morphism) denoted

$$A \xrightarrow{\mathrm{id}_A} A$$

And these have to satisfy the *identity laws* in composition: for every arrow $f : A \to B$ we must have

$$f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f$$

(the identities “do nothing” in composition).

It is an easy exercise to see that identity arrows are unique.

When the objects of a category are sets (perhaps with some additional structure) and the arrows are functions between these sets (perhaps preserving the additional structure, if any), the category axioms always hold, so long as the composition of structure-preserving functions is structure-preserving. Composition as defined for ordinary functions

$$(g \circ f)(x) = g\bigl(f(x)\bigr), \qquad \text{for all } x \text{ in the domain of } f$$

is easily seen to obey the associative law. And identity arrows defined to be identity functions $x \mapsto x$ are easily seen to satisfy the identity laws.

8.4.1 Isomorphism

Another kind of morphism is an *isomorphism* (called that even if you prefer to call morphisms "arrows"). An arrow $f : A \to B$ is called an *isomorphism* if there exists an arrow $g : B \to A$ such that

$$\begin{aligned} g \circ f &= \mathrm{id}_A \\ f \circ g &= \mathrm{id}_B \end{aligned}$$

If f is a function and g is its inverse function, then the above always holds. The only question is whether the inverse of a structure-preserving mapping (however defined in the category being studied) is itself structure-preserving.

Objects connected by an isomorphism are said to be *isomorphic*. Because isomorphisms preserve structure going both ways, isomorphic objects are the "same" as far as the mathematical structure "preserved" is concerned.

8.4.2 Linear Category

Theorem 18. *There is a category where the objects are vector spaces and the arrows are linear functions.*

Proof. We need to show that (i) identity functions are linear functions and (ii) the composition of linear functions is a linear function.

If f and g are linear functions, and $h = g \circ f$, then

$$\begin{aligned} h(x + y) &= g\bigl(f(x + y)\bigr) \\ &= g\bigl(f(x) + f(y)\bigr) \\ &= g\bigl(f(x)\bigr) + g\bigl(f(y)\bigr) \\ &= h(x) + h(y) \end{aligned}$$

and

$$\begin{aligned} h(tx) &= g\big(f(tx)\big) \\ &= g\big(tf(x)\big) \\ &= tg\big(f(x)\big) \\ &= th(x) \end{aligned}$$

That proves (i). If id is an identity function on a vector space, then

$$\begin{aligned} \mathrm{id}(x+y) &= x+y \\ &= \mathrm{id}(x)+\mathrm{id}(y) \\ \mathrm{id}(tx) &= tx \\ &= t\mathrm{id}(x) \end{aligned}$$

That proves (ii). □

Theorem 19. *An invertible linear function is an isomorphism (of linear category).*

Proof. Suppose $f : A \to B$ is invertible with its inverse being $g : B \to A$, and suppose f is linear. What is to be shown is that g is also linear.

Start with addition

$$g(x+y) = g\big(f(u)+f(v)\big)$$

where $u = g(x)$ and $y = g(v)$ by invertibility of f. And

$$g\big(f(u)+f(v)\big) = g\big(f(u+v)\big) = u+v$$

by linearity and invertibility of f. Reading end-to-end gives

$$g(x+y) = g(x)+g(y).$$

Next scalar multiplication, for scalar t and vector x

$$g(tx) = g\big(tf(u)\big)$$

where $u = g(x)$. And

$$g\big(tf(u)\big) = g\big(f(tu)\big) = tu$$

by linearity and invertibility of f. Reading end-to-end gives

$$g(tx) = tg(x).$$

□

8.4.3 Preorder

Arrows do not have to be functions. Consider a category in which between any two objects A and B there is at most one arrow $A \to B$. Such a category is called a *preorder*. You can think of the arrow as indicating a relation. If we write $A \lesssim B$ instead of $A \to B$, then this becomes more obvious. The composition rule applied to a preorder says $A \lesssim B$ and $B \lesssim C$ implies $A \lesssim C$ (this property of a binary relation $\lesssim$ is called *transitivity*) and the identity rule says $A \lesssim A$ for every object A (this property of a binary relation $\lesssim$ is called *reflexivity*). Thus a preorder (as defined using this category-theoretic woof) is "just" a transitive, reflexive relation. And the arrows are not functions.

8.4.4 Homotopy

Another very important instance where arrows are not functions arises in algebraic topology where there are categories in which the objects are topological spaces and the arrows are homotopy equivalence classes of continuous functions, a *homotopy* being a continuous deformations of a continuous function, which we won't bother to explain — too complicated and too off-topic.

8.4.5 Affine Category

So we finally arrive at the category in which the objects are abstract affine spaces (including the empty space $\varnothing$) and the arrows are affine functions (including empty functions $\varnothing \to A$ with A an affine space). We still have to prove this is a category.

In aid of category-theoretic woof about affine spaces and affine functions, we introduce a major tool of category theory: the commutative diagram. This is a graph whose nodes are objects and whose edges are arrows, and such that different paths connecting the same nodes (different compositions of arrows) are asserted to be equal — that's what the word "commutative" is doing in there.

We turn our definition of affine function into a commutative diagram. For an affine space A having translation space U, define the functions $\text{sub}_x : A \to U$ and $\text{add}_x : U \to A$ by $y \mapsto y - x$ and $v \mapsto x + v$, respectively. These are the functions that the first axiom of affine spaces says are inverses of each other. Let B be another affine space having translation space V.

Then our definition of affine function says that $f : A \to B$ is an affine function if and only if there is a linear function $g : U \to V$ such that the following diagram commutes

$$\begin{tikzcd}
A \arrow[r, "f"] & B \arrow[d, "\mathrm{sub}_{f(x)}"] \\
U \arrow[u, "\mathrm{add}_x"] \arrow[r, "g"] & V
\end{tikzcd} \tag{17}$$

Here "commutes" means

$$g = \mathrm{sub}_{f(x)} \circ f \circ \mathrm{add}_x \tag{18}$$

or, decoding, the meaning of the "add" and "sub" functions

$$g(v) = f(x+v) - f(x), \tag{19}$$

which is just what we said before.

So why the abstraction? Just to make pictures? Just to make the simple complicated? No! The abstraction makes the complicated simple.

Theorem 20. *The "add" and "sub" functions $v \to x + v$ and $y \mapsto y - x$ are invertible affine functions whose inverses are also affine.*

Proof. We already know they are inverses of each other. So it is enough to prove both are affine. As usual, we consider A to be an affine subspace of some enclosing vector space, which we do not explicitly mention.

Let A be an affine space, V its translation space, and note that V is its own translation space (when V is considered as an affine space).

First $\mathrm{sub}_x : A \to V$. We must show that $g : V \to V$ defined by

$$g(v) = \mathrm{sub}_x(x + v) - \mathrm{sub}_x(x) = [(x + v) - x] - [x - x] = v$$

is linear. It is clearly the identity function on V, hence linear.

Second $\mathrm{add}_x : V \to A$. Now we must show that $g : V \to V$ defined by

$$g(v) = \mathrm{add}_x(x + v) - \mathrm{add}_x(x) = [(x + v) + v] - [x + v] = v$$

is linear. It is clearly the identity function on V, hence linear. □

Theorem 21. *There is a category where the objects are affine spaces (including the empty space) and the arrows are affine functions (including empty functions).*

Proof. First we deal with nonempty spaces. Consider the following commutative diagram

$$\begin{tikzcd}
A \arrow[r, "f"] & B \arrow[r, "g"] \arrow[d, "\mathrm{sub}_{f(x)}"] & C \arrow[d, "\mathrm{sub}_{g(f(x))}"] \\
U \arrow[u, "\mathrm{add}_x"] \arrow[r, "h"] & V \arrow[r, "j"] & W
\end{tikzcd}$$

where A, B, and C are affine spaces, U, V, and W, respectively, are their translation spaces, f and g are affine functions, which happens if and only if h and j are linear functions. Reading around the outer rectangle shows that $g \circ f$ is affine because $j \circ h$ is linear by Theorem 18.

Now consider the commutative diagram

$$\begin{tikzcd}
A \arrow[r, "\mathrm{id}_A"] & A \arrow[d, "\mathrm{sub}_x"] \\
U \arrow[u, "\mathrm{add}_x"] \arrow[r, "\mathrm{id}_U"] & U
\end{tikzcd}$$

This shows identity functions on affine spaces are affine functions, because identity functions on vector spaces are linear functions.

Now we have to deal with empty morphisms. Trivially, the composition of an empty morphism with any other morphism is empty, hence affine by definition. Also the empty mapping is, again trivially, the identity mapping on the empty set (it takes any element of the empty set — there are none — to itself). □

We see that commutative diagrams show the "big picture" immediately.

Corollary 22. *The "add" and "sub" functions described by Theorem 20 are isomorphisms of affine category.*

Proof. This is just what Theorem 20 says, but in different language. □

Theorem 23. *An invertible affine function (including the empty function $\varnothing \to \varnothing$) is an isomorphism (of affine category).*

Proof. Consider the commutative diagram (17). We know f is affine if and only if g is linear. We assume f is affine, so g is linear. We also assume f is invertible. But, since the vertical arrows are also invertible (Corollary 22), that proves that g being the composition of invertible functions (18) is invertible. But then that proves f is invertible with

$$f^{-1} = \mathrm{add}_x \circ g^{-1} \circ \mathrm{sub}_{f(x)}$$

and we are done (the inverse of every invertible affine function is affine).

Except that we still have to do the empty function $\varnothing \to \varnothing$, which is obviously its own inverse and affine by definition. (An empty function $\varnothing \to A$ with A nonempty, cannot be an isomorphism because there is no function $A \to \varnothing$). □

8.5 Affine Subspaces

If A is an affine space having translation space U, then B is an *affine subspace* of A if and only if B is an affine space when it inherits its operations from A. Since the operations involve both points and vectors, we also need to know the translation space of B. We know that the operation sub_x defined in the preceding section maps an affine space onto its translation space. Hence the translation space of B must be

$$V = B - x = \{ \, y - x : y \in B \, \},$$

where x is any point in B. Thus the translation space of B is a vector subspace of U. Conversely, we have

$$B = x + V = \{ \, x + v : v \in V \, \},$$

so that is our criterion for B being a nonempty affine subspace of A. It has the form $x + V$, where V is a vector subspace of the translation space of A.

Of course, the empty set is also an affine subspace of A, but does not have this form because the empty affine space has no translation space.

Theorem 24. *The image of a vector subspace under a linear map is a vector subspace. Same with preimage instead of image.*

Proof. Suppose $f : U \to V$ is a linear function between vector spaces.

First suppose W is a vector subspace of U. We are to show that the image (also called the range) of W under f

$$f(W) = \{ \, f(w) : w \in W \, \}$$

is a vector subspace of V. Consider points z_1 and z_2 in $f(W)$. They must have the form $z_i = f(w_i)$ for points w_1 and w_2 in W. By linearity, we have $z_1 + z_2 = f(w_1) + f(w_2) = f(w_1 + w_2)$, so $z_1 + z_2 \in f(W)$. And consider a point $z \in f(W)$ and a scalar t. We must have $z = f(w)$ for some $w \in W$. By linearity, we have $tz = tf(w) = f(tw)$, so $tz \in f(W)$. That proves $f(W)$ is a vector subspace of V.

Second suppose W is a vector subspace of V. We are to show that the preimage of W under f

$$f^{-1}(W) = \{\, u \in U : f(u) \in W \,\}$$

is a vector subspace of U. Consider points z_1 and z_2 in $f^{-1}(W)$. They must satisfy $f(z_i) = w_i$, where $w_i \in W$. By linearity, we have $f(z_1 + z_2) = f(z_1) + f(z_2) = w_1 + w_2 \in W$, so $z_1 + z_2 \in f^{-1}(W)$. And consider a point $z \in f^{-1}(W)$ and a scalar t. We must have $f(z) = w$ for some $w \in W$. By linearity, we have $f(tz) = tf(z) = tw \in W$, so $tz \in f^{-1}(W)$. That proves $f^{-1}(W)$ is a vector subspace of U. □

Theorem 25. *The image of an affine subspace under an affine map is an affine subspace. Same with preimage instead of image.*

Proof. Suppose A and B are affine spaces having translation spaces U and V, respectively, and suppose $f : A \to B$ is an affine function.

First, suppose C is an affine subspace of A. If $C = \varnothing$, then the image $f(C)$ is also empty, hence an affine space, by definition. If nonempty, C has the form $C = x + W$, where W is a vector subspace of U. If g is the linear function defined by (17), (18), or (19), then we see that

$$f = \operatorname{add}_{f(x)} \circ g \circ \operatorname{sub}_x$$

so

$$f(C) = \operatorname{add}_{f(x)}\Big(g\big(\operatorname{sub}_x(C)\big)\Big) = g(C - x) + f(x)$$

Since we know that $C - x$ is the translation space of C, and hence that $g(C - x)$ is a vector subspace of U (by Theorem 24), this proves $f(C)$ is an affine subspace of V.

Second, suppose C is an affine subspace of B. If B is empty, then its preimage is also empty, and is an affine space, by definition. If nonempty, C has the form $C = f(x) + W$, where W is a vector subspace of V. So we know by Theorem 24 that $g^{-1}(W)$ is a vector subspace of U. We claim that for any $x \in f^{-1}(C)$ that

$$f^{-1}(C) = x + g^{-1}\big(C - f(x)\big), \tag{20}$$

which makes it an affine subspace of A. So it only remains to check (20). We have $a \in f^{-1}(C)$ if and only if $f(a) \in C$, which happens if and only if $f(a) - f(x) \in C - f(x)$. But, by definition of g, we have $f(a) - f(x) = g(a - x)$ So, continuing our chain of equivalences, we have $a \in f^{-1}(C)$ if and only if

$g(a - x) \in C - f(x)$, which happens if and only if $a - x \in g^{-1}\big(C - f(x)\big)$, which happens if and only if $a \in x + g^{-1}\big(C - f(x)\big)$. This proves (20).

Finally, we have to consider empty affine functions $\varnothing \to A$. Clearly every image or preimage (of any set, not just affine ones) under such a function is empty, hence affine. □

8.6 Convexity, Convex Hull, Affine Hull

All of our discussion of convexity and concavity can be moved from the setting of vector spaces to the setting of abstract affine spaces with only modest rewriting, mostly, as noted above, replacing the expression $tx + (1 - t)y$, which makes no sense in an affine space, by the expression $y + t(x - y)$, which does make sense.

The *convex hull* of an arbitrary set S in an affine space is the intersection of the set of all convex sets containing A. It is denoted $\operatorname{con} S$. Of course, the convex hull of the empty set is empty.

The *affine hull* of an arbitrary set S in an affine space is the intersection of the set of all affine subspaces containing A. It is denoted $\operatorname{aff} S$. Of course, the affine hull of the empty set is empty.

The *span* (also called *linear hull*) of an arbitrary set S in a vector space is the intersection of the set of all vector subspaces containing A. It is denoted $\operatorname{span} S$. Since every vector space contains the zero vector, span of the empty set is the trivial vector space, which contains exactly one element, the zero vector.

For any nonempty set S in an affine space and any $x \in S$ we can write

$$\operatorname{aff} S = x + \operatorname{span}(S - x).$$

8.7 Combinations

For any finite set of vectors $x_1, \ldots, x_n$ in a vector space and any set of scalars $t_1, \ldots, t_n$,

$$\sum_{i=1}^{n} t_i x_i \tag{21}$$

is called a *linear combination* of the vectors. Under the restriction that all of the scalars are nonnegative, it is called a *nonnegative combination* or a *conical combination.* Under the restriction that the scalars sum to one, it is called an *affine combination.* Under both restrictions, it is called a *convex combination.* (A linear combination is a convex combination if it is both a nonnegative combination and an affine combination.)

It follows by mathematical induction from the definitions that a set is convex if and only if it contains all convex combinations of its points and a set is a vector space if and only if it contains all linear combinations of its points.

Theorem 26. *A subset of a vector space is affine if and only if it contains all affine combinations of its points.*

Proof. If empty, it is affine by definition, and it contains all affine combinations of its points vacuously (there are none).

Otherwise, let x_0, …, x_n be points in the set, and let t_0, …, t_n be scalars that sum to one. Then

$$\sum_{i=0}^{n} t_i x_i = \sum_{i=1}^{n} t_i (x_i - x_0) + x_0 \left(\sum_{i=0}^{n} t_i \right) = x_0 + \sum_{i=1}^{n} t_i (x_i - x_0) \tag{22}$$

If the set is affine, then it has the form $A = x_0 + V$, where V a vector subspace. Then any affine combination is in A because the last sum in (22) is a linear combination of vectors, hence an element of V. Conversely, if A contains all of its affine combinations, then the set $V = A - x_0$ contains all of its linear combinations, hence is a vector space. So A is an affine space. □

8.8 Positive Hull

We have four types of combinations but only three types of hulls. To fill out the pattern we should have a fourth type of hull, and there is one, but it doesn't have a simple name.

In Rockafellar and Wets (1998) a *cone* is defined to be a subset of a vector space that is closed under multiplication by nonnegative scalars. They define the *positive hull* of the subset S of a vector space V to be

$$\operatorname{pos} S = \{\, tv : t \geq 0 \text{ and } v \in S \,\}.$$

But this is not the "hull" that is closed under nonnegative combinations. That is either $\operatorname{con}(\operatorname{pos} S)$ or $\operatorname{pos}(\operatorname{con} S)$. So we don't have a simple name or a simple notation for the "hull" that goes with nonnegative combinations. But we do have the concept. The smallest convex cone that contains S, which is $\operatorname{con}(\operatorname{pos} S)$, is the set of all nonnegative combinations of points of S.

8.9 Hulls in Affine Spaces

The notions of linear hulls (spans) and positive hulls make no sense in abstract spaces (obvious from the fact that both must contain the zero vector, and abstract affine spaces don't have a zero vector).

Affine and convex hulls do make sense in affine spaces if redefined. We have already seen that we can redefine an affine combination so that it makes sense in an affine space. The left-hand side of (22) makes no sense in an affine space, but the right-hand side does make sense. So we take the right-hand side of (22) to define an affine combination of the points x_0, $\ldots$, x_n. The scalars t_1, $\ldots$, t_n can be any real numbers (they do not have to sum to one or satisfy any other restriction).

Similarly, we can also take the right-hand side of (22) to define a convex combination of the points x_0, $\ldots$, x_n. The scalars t_1, $\ldots$, t_n now have to be nonnegative real numbers that sum to less than or equal to one (because on the left-hand side of (22) they had to be nonnegative scalars that sum to one when t_0 is included, so when we omit t_0 the sum is less than or equal to one).

8.10 Linear and Affine Independence

A set of vectors is *linearly independent* if no linear combination of them is equal to zero except when all of the scalars are zero. Equivalently, they are linearly independent if no one of them is a linear combination of the others.

A set of points is *affinely independent* if no one of them is an affine combination of the others. Equivalently (from (22)), a set of points x_0, $\ldots$, x_n is affinely independent if and only if the set of vectors $x_1 - x_0$, $\ldots$, $x_n - x_0$ is linearly independent.

8.11 Dimension

The dimension of a vector space is the size of its largest linearly independent subset (any such subset is called a *basis*). The dimension of a nonempty affine space is the dimension of its translation space. Clearly, the dimension of an affine set is d if and only if the size of its largest affinely independent set is $d + 1$. An empty affine space has no dimension.

Vector spaces and affine spaces can be infinite-dimensional, but only finite-dimensional vector spaces and affine spaces can have exponential families on them.

Theorem 27. *Every finite-dimensional vector space is isomorphic to $\mathbb{R}^d$ for some d.*

Proof. Let x_1, $\ldots$, x_d be a basis. Then every vector can be written as a linear combination of this basis in exactly one way, say

$$y = \sum_{i=1}^{d} f_i(y) x_i \tag{23}$$

because if we could also write

$$y = \sum_{i=1}^{d} \tilde{f}_i(y) x_i$$

with $f_i(y) \neq \tilde{f}_i(y)$ for some i, this would violate linear linear independence of the basis

$$0 = \sum_{i=1}^{d} \bigl[\tilde{f}_i(y) - f_i(y) \bigr] x_i.$$

The functions f_i are linear because

$$\begin{gathered} y_1 + y_2 = \left(\sum_{i=1}^{d} f_i(y_1) x_i \right) + \left(\sum_{i=1}^{d} f_i(y_2) x_i \right) = \sum_{i=1}^{d} f_i(y_1 + y_2) x_i \\ ty = t \left(\sum_{i=1}^{d} f_i(y) x_i \right) = \sum_{i=1}^{d} f_i(ty) x_i \end{gathered}$$

by uniqueness of definition of the f_i. Hence we have $f_i(y_1 + y_2) = f_i(y_1) + f_i(y_2)$ and $f_i(ty) = t f_i(y)$ because this makes the above equations correct (so it can be correct) and the f_i are uniquely defined (so it must be correct).

Now we claim that the function f that maps y to the element of $\mathbb{R}^d$ having components $f_i(y)$ is a vector space isomorphism. Having already shown it is linear, it is enough to exhibit its inverse mapping, and we have already done so, (23) is it. □

Corollary 28. *Finite-dimensional vector spaces are isomorphic if and only if they have the same dimension.*

Proof. If U and V are vector spaces having dimension d, then by the theorem we have isomorphisms

$$U \longrightarrow \mathbb{R}^d \longrightarrow V$$

and the composition of isomorphisms is an isomorphism. Conversely, if U has dimension d and V has dimension e with $d < e$ but U and V were isomorphic, we would have isomorphisms

$$\mathbb{R}^d \longrightarrow U \longrightarrow V \longrightarrow \mathbb{R}^e$$

which would make $\mathbb{R}^d$ and $\mathbb{R}^e$ isomorphic. But this is impossible. Proof by contradiction. Assume $f : V \to U$ is a linear isomorphism and $x_1, \ldots, x_e$ is a basis for V. Then $f(x_1), \ldots, f(x_e)$ cannot be a basis for U, so we can write

$$f(x_i) = \sum_{\substack{j=1 \\ j \neq i}}^{e} t_j f(x_j) = f\left(\sum_{\substack{j=1 \\ j \neq i}}^{e} t_j x_j \right)$$

but since f is an isomorphism, this implies x_i is a linear combination of the rest of the x_j, which contradicts the assumption that $x_1, \ldots, x_e$ is a basis. □

8.12 Functions as Vectors

The set of all functions from an arbitrary set to a vector space can be thought of as a vector space with addition and scalar multiplication of functions defined in the obvious way

$$(f + g)(x) = f(x) + g(x) \tag{24a}$$

$$(tf)(x) = tf(x) \tag{24b}$$

The same is true when the functions are restricted in some way provided that the restriction holds for sums and scalar multiples. For example all continuous functions on a topological space form a vector space, and all linear functions on a vector space form a vector space. The vector space consisting of all linear functions from a vector space U to a vector space V is denoted $L(U, V)$.

8.13 Dual Spaces

The construction in the preceding section, when specialized to linear functions whose codomain is the scalar field (that is, $L(U, \mathbb{R})$ when $\mathbb{R}$ is the scalar field) is called the *dual space* of U. It is denoted U^*. Its elements are called *linear functionals* ("functional" rather than "function" means the codomain is the scalar field). We already used this concept in Section 5. Now we say a bit more.

As in Section 5 and everywhere else, we write $\langle x, f\rangle$ instead of $f(x)$. The idea is that this indicates that not only is $\langle\,\cdot\,, f\rangle$ a linear functional (on E) for each f but also $\langle x,\,\cdot\,\rangle$ is a linear functional (on E^*) for each x. This follows from the very definition of addition and scalar multiplication of functions given in the preceding section.

Theorem 29. *Every finite-dimensional vector space has the same dimension as its dual space.*

Proof. Let U and U^* be the spaces in question. We suppose U has dimension d. In case $d = 0$, the theorem is obvious. The only linear functional on the zero-dimensional vector space $\{0\}$ is the zero functional, because we always have $f(0) = 0$ for any linear function f. When $d > 0$, let $x_1, \ldots, x_d$ be a basis for U. Then we know that any $u \in U$ can be written in exactly one way as a linear combination of $x_1, \ldots, x_d$, say

$$u = \sum_{i=1}^{d} f_i(u)x_i. \tag{25}$$

We claim the f_i are linear functionals that form a basis for U^*. They have already been shown to be linear functionals in the proof of Theorem 27. To see that they are linearly independent, observe that when when we write

$$x_i = \sum_{j=1}^{d} f_j(x_i)x_j$$

that we can have $f_i(x_i) = 1$ and $f_j(x_i) = 0$, $j \neq i$, so we must have that by uniqueness. And this shows that none of the f_i can be written as a linear combination of the others.

To see that every linear functional can be written as a linear combination of the f_i, let g be a linear functional. Then

$$\begin{aligned} g(y) &= g\left(\sum_{i=1}^{d} f_i(y)x_i\right) \\ &= \sum_{i=1}^{d} f_i(y)g(x_i) \end{aligned}$$

by linearity. So

$$g = \sum_{i=1}^{d} g(x_i)f_i.$$

□

Although a finite-dimensional vector space and its dual are both isomorphic to $\mathbb{R}^d$ for some d, there are many different isomorphisms. A cure for thinking that we want to always interpret $\langle x, \theta \rangle$ as $x^T \theta$ when we think of vectors as $d \times 1$ matrices, is to understand that we could just as well think of it as $x^T A \theta$, where A is any invertible matrix (fixed throughout the discussion). There is no "natural" way to identify a vector space and its dual. (There is for inner product spaces, but we are explicitly not assuming an inner product.)

The situation is different for a space and its double dual. If E is a finite-dimensional vector space, then E^* is the set of all linear functionals on E, and E^{**} is the set of all linear functionals on E^*. If $\langle \,\cdot\,, \,\cdot\, \rangle$ is the canonical bilinear form placing E and E^* in duality defined by (6), then every $\langle \,\cdot\,, \theta \rangle$ is a linear functional on E, hence an element of E^* ($\langle \,\cdot\,, \theta \rangle$ and θ are different notations for the same linear functional, just like f and $f(\,\cdot\,)$ mean the same thing for any function f). And every $\langle x, \,\cdot\, \rangle$ is a linear functional on E^* hence an element of E^{**}.

Theorem 30. *If E is a finite-dimensional vector space, then $x \mapsto \langle x, \,\cdot\, \rangle$ is a linear isomorphism $E \to E^{**}$.*

Proof. Let F denote the function $x \mapsto \langle x, \,\cdot\, \rangle$. It is a linear function because $\langle \,\cdot\,, \,\cdot\, \rangle$ is a bilinear form. We claim it is a one-to-one function. Suppose $\langle x_1, \theta \rangle = \langle x_2, \theta \rangle$ for all $\theta \in E^*$. This implies $\langle x_1 - x_2, \theta \rangle = 0$, for all $\theta \in E^*$. Let $f_1, \ldots, f_d$ be the basis for E^* constructed in the proof of Theorem 29. Then we have $f_i(x_1 - x_2) = 0$ for all i, but this implies $x_1 - x_2 = 0$ by (25). This shows that if $\langle x_1, \,\cdot\, \rangle$ and $\langle x_2, \,\cdot\, \rangle$ are the same linear functional on E^*, then $x_1 = x_2$, and that shows that F is one-to-one.

By linearity and Theorem 24 $F(E)$ is a vector subspace of E^{**}, by Theorem 29, $F(E)$ and E^{**} have the same dimension, hence they must be the same space. □

Theorem 30 does not carry over to infinite dimensions. Neither part of the proof works. The mapping in the theorem need not be one-to-one and need not be onto. (Infinite dimensionality is different.)

The isomorphism defined in Theorem 30 is called the *natural isomorphism* of a finite-dimensional vector space and its double dual.

8.14 Topology

We don't assume anyone in the class has had general topology, and we don't want to cover that, which would be a semester course in itself.

Suffice it to say that topology is the concept in which the open sets of a space are specified directly (not in terms of a metric or a norm or anything else). Then all other topological concepts, such as closed sets, closure of sets, compact sets, limits of sequences, and continuous functions, are specified in terms of open sets. For example, a function f from one topological space to another is *continuous* if the preimage of every open set (in the codomain) is an open set (in the domain). When both topologies (domain and codomain) are specified in terms of metrics, this is equivalent to the usual epsilon-delta definition. Otherwise, this is a more general concept.

A *topological vector space* is a vector space equipped with a *vector topology*, which is a topology satisfying three axioms: (Rudin, 1991, Section 1.6).

- Vector addition is a continuous operation $V \times V \to V$.
- Scalar multiplication is a continuous operation $\mathbb{R} \times V \to V$.
- Points are closed sets.

An infinite-dimensional vector space may be equipped with many different vector topologies. It is commonplace in the study of infinite-dimensional topological vector spaces, which is called *functional analysis*, to use more than one vector topology for the same space, sometimes in the same argument.

For finite-dimensional vector spaces the situation is different. A morphism of the category of topological vector spaces is a function that is both linear and continuous. A finite-dimensional vector space V has exactly one vector topology, and any linear isomorphism $V \to \mathbb{R}^d$ is a topological vector space isomorphism (Rudin, 1991, Theorem 1.21, this theorem is stated for complex rather than real scalars, but the comment following the theorem statement says the proof is also valid for real scalars).

The topology for $\mathbb{R}^d$ is the "usual" topology, which is the product topology (open sets are unions of open boxes) inherited from the "usual" topology for $\mathbb{R}$, which is the order topology (open sets are unions of open intervals).

Every linear function from one abstract finite-dimensional vector space to another is continuous (Rudin, 1991, Theorems 1.18 and 1.21). It follows that every invertible linear function between abstract finite-dimensional topological vector spaces is a topological vector space isomorphism.

This is why considering all finite-dimensional vector spaces to be $\mathbb{R}^d$ for some d entails no loss of generality. Every finite-dimensional vector space "is" $\mathbb{R}^d$ for some d, up to isomorphism. It does entail loss of elegance, but no loss of generality.

We want to extend these results to affine spaces, but then we need a notion of *affine topology.* It seems this should require subtraction of points and addition of a point and a vector to be continuous operations. And it would also seem that we want the topology of the translation space to be a vector topology (hence for the finite-dimensional case the only possible vector topology). And this seems to say that the functions add_x and sub_x defined in Section 8.4.5, which we already know are affine space isomorphisms (Corollary 22) are also topological affine space isomorphisms. Thus the open sets of a *topological affine space* are defined to be either the images of open sets of its translation space under add_x (for any point x) or the preimages of open sets of its translation space under sub_x (for any point x).

We don't give a proof of that this is the only possible topology that makes the operations continuous, because it would involve general topology. We just take it as a definition.

Theorem 31. *Any invertible linear function between finite-dimensional vector spaces is an isomorphism of the category of topological vector spaces. Any invertible affine function between finite-dimensional affine spaces is an isomorphism of the category of topological affine spaces.*

Proof. We already know that invertible linear functions are isomorphisms of the category of vector spaces. We only need to show that they are continuous both ways and the aforementioned Theorems 1.18 and 1.21 in Rudin (1991) prove this.

We already know that invertible affine functions are isomorphisms of the category of affine spaces (Theorem 23). We only need to show that they are continuous both ways. We already know that linear functions are continuous and that the functions add_x and sub_x are continuous for any x (we took that at the definition of the topology of finite-dimensional affine space). Now the fact that every affine function f can be expressed as the composition as in (17)

$$f = \mathrm{add}_{f(x)} \circ g \circ \mathrm{sub}_x$$

where g is linear and hence every affine is the composition of linear functions, which is linear (and that latter fact is obvious from the notion that there is a category where the objects are topological spaces and the morphisms are continuous functions or from the definition of continuous function as one for which preimages of open sets are open sets). □

The convex hull of an affinely independent finite set of points is called a *simplex.*

Corollary 32. *A simplex of maximal dimension in an affine space has nonempty interior.*

Maximal dimension means if the dimension is d, the simplex is the convex hull of $d+1$ points.

Proof. So consider d-dimensional affine space A and let the simplex be the convex hull of the points $x_0, \ldots, x_d$. We claim there is a topological affine space isomorphism that takes x_0 to the origin of $\mathbb{R}^d$ and x_i to the i-th vector of the "standard basis" for $\mathbb{R}^d$ (the vector that has all components equal to zero except for the i-th, which is equal to one).

We know that for any point y the set $x_0, \ldots, x_d, y$ is not affinely independent (because we assumed A has dimension d) so y can be written as an affine combination of $x_0, \ldots, x_d$ and can be so written in exactly one way. We show the latter by proof by contradiction. Assume

$$\begin{aligned} y &= x_0 + \sum_{i=1}^{d} t_i(x_i - x_0) \\ &= x_0 + \sum_{i=1}^{d} t_i^*(x_i - x_0) \end{aligned}$$

with $t_i \neq t_i^*$ for some i. Then

$$0 = \sum_{i=1}^{d} (t_i - t_i^*)(x_i - x_0)$$

contradicting the assumption that the vectors $x_i - x_0$ are linearly independent. The map in question now is seen to take y to $(t_1, \ldots, t_d)$. And this is an invertible affine function. The image of the convex hull of $x_0, \ldots, x_d$ under this mapping is the set

$$S = \left\{ x \in \mathbb{R}^d : x_i \geq 0, \text{ for all } i \text{ and } \sum_{i=1}^{d} x_i \leq 1 \right\},$$

the interior of which is obviously

$$\operatorname{int} S = \left\{ x \in \mathbb{R}^d : x_i > 0, \text{ for all } i \text{ and } \sum_{i=1}^{d} x_i < 1 \right\}.$$

(For any point in $x \in \operatorname{int} S$ there is a $\varepsilon > 0$ such that the box

$$\left\{ y \in \mathbb{R}^d : |y_i - x_i| < \varepsilon, \text{ for all } i \text{ and} \right\}$$

is contained in $\operatorname{int} S$. Take ε to be smaller than $(1 - x_1 - \cdots - x_d)/d$ and smaller than x_i for all i.)

Clearly the preimage of $\operatorname{int} S$ under the claimed map is

$$\left\{ x_0 + \sum_{i=1}^{d} t_i (x_i - x_0) : t \in \operatorname{int} S \right\}$$

and this is the largest open set contained in the convex hull of x_0, $\ldots$, x_d, hence the interior of this convex hull. □

The dimension of a convex set S is the dimension of $\operatorname{aff} S$. It is also the dimension of the largest affinely independent subset of S, because $\operatorname{aff} S$ is the set of all affine combinations of that affinely independent subset.

Corollary 33. *Every convex set having full dimension (the same dimension as the enclosing affine space) has nonempty interior.*

8.15 Relative Interior

The *relative interior* of a convex set S in an affine space is its interior relative to its affine hull (the interior of S considered as a subset of $\operatorname{aff} S$). The *relative boundary* is the closure minus the relative interior. The relative interior of S is denoted $\operatorname{ri} S$.

A one-dimensional affine space is called a *line*. Its translation space is a one-dimensional vector space, which has the form

$$\{\, tv : t \in \mathbb{R} \,\}$$

for some nonzero vector v (which is the only element of a basis for the translation space. Thus a line has the form

$$\{\, x + tv : t \in \mathbb{R} \,\}$$

for some point x in the affine space and some nonzero vector v in its translation space.

Theorem 34. *Any nonempty convex set has a nonempty relative interior.*

Proof. This is Corollary 33 stated in different words. □

Theorem 35. *For any convex set S in an affine space we have $x \in \operatorname{ri} S$ if and only if for every $y \in S$ such that $y \neq x$ there exists $z \in S$ such that x is in the relative interior of the line segment having endpoints y and z, that is there exists t such that $0 < t < 1$ and $x = y + t(z - y)$.*

Proof. Suppose $x \in \operatorname{ri} S$. If $S = \{x\}$, then $\operatorname{ri} S = \{x\}$, and there are no points y to check so x passes the test vacuously. Otherwise, consider y as in the theorem statement. Consider the line L determined by x and y. This line is contained in $\operatorname{aff} S$, so in order for x to be a relative interior point of S there must be a neighborhood W of x in $\operatorname{aff} S$ such that x is in the relative interior of $S \cap L$. But this implies the existence of a z as in the theorem statement.

Conversely, if S is empty, there is nothing to prove, and if S is a singleton set, then it is its own relative interior, and the test of the theorem statement is passed vacuously because there are no such points y. Otherwise, consider an affine isomorphism $f : \operatorname{aff} S \to \mathbb{R}^d$ for some d. We have $d \geq 1$ because otherwise $\mathbb{R}^d$ would be a singleton set and so would $\operatorname{aff} S$, hence S. The image $f(S)$ has nonempty interior (because it has full dimension in $\mathbb{R}^d$), the the boundary B of $f(S)$ is a closed set. Define

$$\alpha = \inf_{y \in B} \|f(x) - y\|,$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$ (the reason for the mapping f is that an abstract affine space doesn't have a given norm, so we map to $\mathbb{R}^d$, which does have a default one). We cannot have $\alpha = 0$ because then for $\varepsilon > 0$, the set

$$K = \{\, y \in B : \|f(x) - y\| \leq \varepsilon \,\}$$

is compact, so the infimum is achived, but this would contradict $f(x)$ being an interior point of $f(S)$. But then we see that the line determined by $f(y)$ and $f(x)$ has points in $\operatorname{int} f(S)$ on both sides of $f(x)$. Hence the line determined by y and x has points in $\operatorname{ri} S$ on both sides of x. □

Theorem 36. *Suppose K is a nonempty convex set in a finite-dimensional vector space V and σ_K is its support function. Then $x \in \operatorname{ri} K$ if and only if for every $\eta \in V^*$ either*

$$\langle x, \eta \rangle < \sigma_K(\eta) \tag{26a}$$

or

$$\langle y, \eta \rangle = \sigma_K(\eta), \qquad y \in K. \tag{26b}$$

Proof. First suppose $x \in \operatorname{ri} K$. For η such that (26b) does not hold we have $y \in K$ such that $\langle y, \eta \rangle < \sigma_K(\eta)$. But then by Theorem 35 there is a $z \in K$ such that $z = x + t(y - x)$ with $t > 1$, and we must have $\langle z, \eta \rangle \leq \sigma_K(\eta)$. It follows that $\langle x, \eta \rangle < \sigma_K(\eta)$. So x passes the test for this theorem.

Conversely, suppose $x \in K$ and x passes the test for this theorem. □

9 Expectations and Moments on Vector Spaces

9.1 Random Vectors

In any topological vector space V, a random vector is a random object described by a Borel probability measure on V. Thus this concept also extends far beyond finite-dimensional vector spaces.

9.2 Ordinary Moments

In $\mathbb{R}^d$ the mean of a random vector X is the vector whose components are the means of the components of X. In an abstract vector space, vectors have no components (they can be given components by a choice of basis, but every different choice of basis leads to a different notion of components).

The good analog of components in $\mathbb{R}^d$ in an abstract vector space V is all of the $\langle X, \eta \rangle$ for all $\eta \in V^*$. This gives us an infinite number of "components" $\langle X, \eta \rangle$, but, of course, since V^* is finite-dimensional, this infinite number of "components" are determined by $\langle X, \eta_i \rangle$ where η_1, $\ldots$, η_d are a basis for V^*. This analogy also works the other way. In $\mathbb{R}^d$ the map $y \mapsto y_i$, where y_i denotes a component of y, is a linear functional (a vector-to-scalar linear function), hence an element of the dual space of $\mathbb{R}^d$ hence a $\langle \,\cdot\,, \eta \rangle$ for some η in the dual space.

Having got the right analogy, it is fairly obvious how moments should be defined (and we get confirmation when we consider moment generating functions in Section 9.13 below).

If X is a random vector in an abstract finite-dimensional vector space V, then the first ordinary moment is the unilinear form $V^* \to \mathbb{R}$ defined by

$$\alpha_1(\eta) = E(\langle X, \eta \rangle), \qquad \eta \in V^*$$

("unilinear form" is not a term we use often, it means the same thing as "linear functional" but goes with the terms we use for higher moments). The second ordinary moment is the symmetric bilinear form $V^* \to V^* \to \mathbb{R}$ defined by

$$\alpha_2(\eta_1)(\eta_2) = E(\langle X, \eta_1 \rangle \langle X, \eta_2 \rangle), \qquad \eta_1, \eta_2 \in V^*.$$

The third ordinary moment is the symmetric trilinear form $V^* \to V^* \to V^* \to \mathbb{R}$ defined by

$$\alpha_3(\eta_1)(\eta_2)(\eta_3) = E(\langle X, \eta_1 \rangle \langle X, \eta_2 \rangle \langle X, \eta_3 \rangle), \qquad \eta_1, \eta_2, \eta_3 \in V^*.$$

And so on (you get the general idea, we hope, although we actually won't be interested in higher than second moments in this document).

The "ordinary" in "ordinary moment" is not a widespread usage. Your humble author uses it to contrast ordinary moments and central moments (Section 9.5 below). Most people say "moment" instead of "ordinary moment."

9.3 Mean

The first ordinary moment is also called the *mean.* So we, like everybody else, also use the notation μ for mean instead of α_1 for first ordinary moment.

As we said in the preceding section, if X is a random vector in V, then its mean μ is a unilinear form on V^*, which is the same thing as saying μ is a linear functional on V^*, which is the same thing as saying $\mu \in V^{**}$.

When V is finite-dimensional, we have the natural isomorphism that allows us to identify V and V^{**}, and that allows us to consider μ an element of V, the same space where X lives.

9.4 Type Theory

It is convenient to steal some notation from type theory as used in functional programming (computer languages like Haskell and also the twenty-first century's new foundations of mathematics, homotopy type theory). A function of two variables $f(x, y)$ can also be thought of a function of one variable whose value is a function (of the other variable) $f(x)(y)$. This conversion (or rethinking) is called *currying.* The programming language Haskell is named after the logician Haskell Curry (1900–1982), and currying is also named after him.

We particularly want to consider currying of multilinear forms. A bilinear form $V \times V \to \mathbb{R}$, can also be curried to be considered a function $V \to (V \to \mathbb{R})$. In type theory the parentheses are always considered redundant, the arrow denoting a function (or morphism) is always considered to associate to the right. So $U \to V \to W$ always means $U \to (V \to W)$.

This ways we can always consider $L(U, L(V, W))$ to be the same as $L(U \times V, W)$. The notation with arrows does not imply that the functions in question are linear like the "L" in $L(U, V)$ does. So we will just have to remember when we are talking about linear functions that we are doing so. But $U \to V \to W$ is simpler notation than $L(U, L(V, W))$ because it reads left to right instead of inside out.

There is another difference. $L(U, L(V, W)$ denotes a set. But $U \to V \to W$ denotes a type, which is considered different from a set in type theory. But we won't bother to fuss about this distinction.

9.5 Central Moments

If X is a random vector in an abstract finite-dimensional vector space V having mean μ, then the first central moment is the unilinear function $V^* \to \mathbb{R}$ defined by

$$\mu_1(\eta) = E(\langle X - \mu, \eta \rangle), \qquad \eta \in V^*,$$

the second central moment is the symmetric bilinear form $V^* \to V^* \to \mathbb{R}$ defined by

$$\mu_2(\eta_1)(\eta_2) = E(\langle X - \mu, \eta_1 \rangle \langle X - \mu, \eta_2 \rangle), \qquad \eta_1, \eta_2 \in V^*,$$

the third central moment is the symmetric trilinear form $V^* \to V^* \to V^* \to \mathbb{R}$ defined by

$$\mu_3(\eta_1)(\eta_2)(\eta_3) = E(\langle X - \mu, \eta_1 \rangle \langle X - \mu, \eta_2 \rangle \langle X - \mu, \eta_3 \rangle), \qquad \eta_1, \eta_2, \eta_3 \in V^*,$$

and so on.

It is customary to use μ without subscripts for the mean and μ with subscripts for central moments, which can be confusing, but in this document we will have little possibility of confusion because we will only be interested in the second central moment and will give it a special name and notation.

The first central moment μ_1 is weird because, by linearity of expectation,

$$\mu_1(\eta) = E(\langle X - \mu, \eta \rangle) = E(\langle X, \eta \rangle) - \langle \mu, \eta \rangle = \langle \mu, \eta \rangle - \langle \mu, \eta \rangle = 0$$

so μ_1 is just another name for the zero unilinear form. We could have started the definitions with μ_2, but chose not to because it is sometimes nice to use the notation μ_1 to preserve symmetry of formulas (but we won't see that in this document).

It might appear at first sight that because of the appearance of $X - \mu$ in the formulas for central moments that this concept relies on the representation $V^{**} = V$ for the double dual (unless $\mu \in V$ we cannot do the subtraction $X - \mu$). But appearances are deceiving because

$$\langle X - \mu, \eta \rangle = \langle X, \eta \rangle - \langle \mu, \eta \rangle$$

and we can always consider the right-hand side to be well defined even if we consider μ to be an element of V^{**}. In that case, the two canonical bilinear forms on the right-hand side are different. In $\langle X, \eta \rangle$, it is the one placing V and V^* in duality. In $\langle \mu, \eta \rangle$, which we should perhaps now write $\langle \eta, \mu \rangle$, it is the the one placing V^* and V^{**} in duality. Thus we could write

$$\mu_2(\eta_1)(\eta_2) = E([\langle X, \eta_1 \rangle - \langle \eta_1, \mu \rangle][\langle X, \eta_2 \rangle - \langle \eta_2, \mu \rangle])$$

and so forth to avoid reliance on $V^{**} = V$. This seems a little too pedantic for us, so we won't.

9.6 Adjoints

Suppose U and V are finite-dimensional abstract vector spaces, and suppose $f : U \to V$ is a linear function. Then there is a unique $f^* : V^* \to U^*$ satisfying

$$\langle f(x), y \rangle = \langle x, f^*(y) \rangle, \qquad y \in V^*, \ x \in U, \tag{27}$$

and f^* is called the *adjoint* of f. Note that the two canonical bilinear forms in (27) are different; on the left-hand side we have the canonical bilinear form placing V and V^* in duality, but on the right-hand side we have the canonical bilinear form placing U and U^* in duality.

Not using the bracket notation and remembering that elements of dual spaces are actually linear functionals as well as vectors makes the existence of the adjoint trivial, because (27) becomes

$$y\bigl(f(x)\bigr) = f^*(y)(x), \qquad y \in V^*, \ x \in U,$$

which says

$$f^*(y) = y \circ f. \tag{28}$$

If we specialize to $U = U^* = \mathbb{R}^d$ and $V = V^* = \mathbb{R}^e$, let the canonical bilinear forms be $\langle x, y \rangle = x^T y$, and confuse linear functions with the matrices representing them, then the adjoint is just the matrix transpose. If M is an $e \times d$ matrix, then the adjoint of the linear function $x \mapsto Mx$ is $y \mapsto M^T y$. But this depends on this particular choice of canonical bilinear forms. It is neither abstract nor general.

9.7 Variance

The second central moment is also called the *variance*, and we also use the notation Σ for variance instead of μ_2 for second central moment.

As we said in the preceding section, if X is a random vector in V, then its variance Σ is a symmetric bilinear form on V^*, which is the same thing as saying Σ has type $V^* \to V^* \to \mathbb{R}$, but every linear function $V^* \to \mathbb{R}$ is a linear functional on V^*, hence an element of V^{**}, so the type can also be written $V^* \to V^{**}$. When we are using the representation $V^{**} = V$, the type is also $V^* \to V$.

Thus we can consider variance to be either a symmetric bilinear form on V^* or a linear function $V^* \to V$. Considered as a bilinear form, it is

$$\Sigma(\eta_1)(\eta_2) = \operatorname{cov}\{\langle X, \eta_1 \rangle, \langle X, \eta_2 \rangle\}, \qquad \eta_1, \eta_2 \in V^*. \tag{29}$$

Considered as a linear function, it maps $\eta_1 \in V^*$ to the unique $x \in V$ such that $\langle x, \,\cdot\, \rangle$ is the linear function on V^* that is also denoted $\Sigma(\eta_1)$. If Σ denotes this linear function, then we can also write the bilinear form

$$(\eta_1, \eta_2) \mapsto \langle \Sigma(\eta_1), \eta_2 \rangle. \tag{30}$$

If we take V and V^* to both be $\mathbb{R}^d$ and the canonical bilinear form to be $\langle x, y \rangle = x^T y$, the usual conventions for probability theory on $\mathbb{R}^d$, then variance is usually defined to be a matrix M having components

$$m_{ij} = \operatorname{cov}(X_i, X_j) \tag{31}$$

In this case, the corresponding bilinear form is $(x, y) \mapsto x^T M y$, and the corresponding linear function is $x \mapsto Mx$. A matrix can represent either a bilinear form or a linear function.

Many people do not like the term "variance matrix" for the matrix having components (31) because it involves covariances. Some call it the "covariance matrix" but that is a really bad term, because what then do you call the covariance of two random vectors? Others call it the "variance-covariance matrix." Others call it the "dispersion matrix." But your humble author always uses "variance matrix" on the grounds that it is the vector analogue of the variance of a random scalar. This is seen in the formulas for the change of mean and variance under a linear transformation (Section 9.9 below), in the central limit theorem and the delta method (Sections 9.18 and 9.19 below), and in many other places. Hence we also call Σ defined by (29) the "variance" (considered as either a symmetric bilinear form or as a linear function).

A variance matrix is symmetric and positive semidefinite and every symmetric and positive semidefinite matrix is a variance matrix (there is, for example, a normal random vector with that variance matrix). Moreover, the variance matrix fails to be positive definite if and only if its random vector is concentrated on a hyperplane. Moreover, the variance matrix fails to be positive definite if and only if it is not invertible. What are the analogs of these properties in the abstract picture?

First we consider variance as a bilinear form, in which case symmetric refers to the bilinear form. We have

$$\Sigma(\eta)(\eta) = E\{\langle X - \mu, \eta \rangle^2\} \geq 0,$$

and this is the positive semidefiniteness property. A symmetric bilinear form $\Sigma : V^* \times V^* \to \mathbb{R}$ is *positive semidefinite* if

$$\Sigma(\eta)(\eta) \geq 0, \qquad \eta \in V. \tag{32}$$

It is *positive definite* if

$$\Sigma(\eta)(\eta) > 0, \qquad \eta \in V, \ \eta \neq 0. \tag{33}$$

Positive definiteness fails if there is an $\eta \neq 0$ such that

$$\Sigma(\eta)(\eta) = E\{\langle X - \mu, \eta \rangle^2\} = 0,$$

which happens if and only if $\langle X - \mu, \eta \rangle = 0$ almost surely, which is the same as saying that X is concentrated on the hyperplane

$$H = \{\, x \in V : \langle x - \mu, \eta \rangle = 0 \,\}. \tag{34}$$

When we think of Σ as a linear function $V^* \to V$ the bilinear form is then (30) and symmetry of this bilinear form says

$$\langle \eta_1, \Sigma^*(\eta_2) \rangle = \langle \Sigma(\eta_1), \eta_2 \rangle = \langle \Sigma(\eta_2), \eta_1 \rangle$$

which seems to say Σ is its own adjoint, but this is because of our using the representation $V^{**} = V$ (if we were not using this representation Σ would type $V^* \to V^{**}$ and Σ^* would have type $V^{**} \to V^{***}$, so they could not be the same).

It is true that variance considered as a bilinear form is positive definite if and only if variance considered as a linear function is invertible. We will derive this by transfer from probability theory on $\mathbb{R}^d$ in the following section.

If we treat V and V^* as both being $\mathbb{R}^d$ for some d and take the canonical bilinear form to be $\langle x, \eta \rangle = x^T A \eta$ for some (fixed throughout the discussion) symmetric positive definite matrix A, then the matrix corresponding to the linear function Σ will be a symmetric positive semidefinite matrix. But, as we said in Section 8.13 above, any invertible matrix A will do here, not necessarily either symmetric or positive definite. And then the matrix corresponding to Σ need not be symmetric or positive semidefinite. Of course, for practical calculations, one wouldn't do that because it would be confusing. But one could do that, which shows that (beating a dead horse here) abstract vector spaces aren't really $\mathbb{R}^d$ for some d, and linear functions between them aren't really matrices, and there is no unique way to associate matrices with them.

9.8 Elementary Properties of Expectation

Theorem 37. *Suppose X and Y are random vectors on the same finite-dimensional vector space. Then $E(X+Y) = E(X)+E(Y)$, provided X and Y have expectations.*

For any random vector X on a finite-dimensional vector space and constant scalar a (in the field that vector space is over) $E(aX) = aE(X)$, provided X has expectation.

If X is a constant random vector, that is, $X = \mu$ almost surely for some constant vector μ, then $E(X) = \mu$.

Proof. Write μ_U for the expectation of the random vector U. Then

$$\begin{aligned} \mu_{X+Y}(\eta) &= E\{\langle X+Y, \eta\rangle\} \\ &= E\{\langle X, \eta\rangle\} + E\{\langle Y, \eta\rangle\} \\ &= \mu_X(\eta) + \mu_Y(\eta) \end{aligned}$$

for all η, and this proves the first assertion. Also

$$\begin{aligned} \mu_{aX}(\eta) &= E\{\langle aX, \eta\rangle\} \\ &= aE\{\langle X, \eta\rangle\} \\ &= a\mu_X(\eta) \end{aligned}$$

for all η, and this proves the second assertion. If $X = \mu$ almost surely, then

$$\begin{aligned} \mu_X(\eta) &= E\{\langle X, \eta\rangle\} \\ &= E\{\langle \mu, \eta\rangle\} \\ &= \langle \mu, \eta\rangle \end{aligned}$$

for all η, and this proves the third assertion. □

We see that all these elementary properties of the expectation (also called mean) of random vectors follow from bilinearity of $\langle\,\cdot\,,\,\cdot\,\rangle$ and from the natural isomorphism of finite-dimensional vector spaces with their double duals.

9.9 Change of Mean and Variance under Linear Functions

Suppose U and V are finite-dimensional abstract vector spaces and f is a linear function $U \to V$. Suppose X is a random vector in U and $Y = f(X)$. What are the mean and variance of Y in terms of the mean and variance of X?

Denote these means by μ_X and μ_Y and these variances by Σ_X and Σ_Y. Then

$$\begin{aligned} \mu_Y(\eta) &= E(\langle f(X), \eta \rangle) \\ &= E(\langle X, f^*(\eta) \rangle) \\ &= \mu_X\bigl(f^*(\eta)\bigr) \end{aligned} \tag{35}$$

the last step using (28), and, similarly,

$$\begin{aligned} \Sigma_Y(\eta_1)(\eta_2) &= \operatorname{cov}\{\langle f(X), \eta_1 \rangle, \langle f(X), \eta_2 \rangle\} \\ &= \operatorname{cov}\{\langle X, f^*(\eta_1) \rangle, \langle X, f^*(\eta_1) \rangle\} \\ &= \Sigma_X\bigl(f^*(\eta_1)\bigr)\bigl(f^*(\eta_2)\bigr) \end{aligned} \tag{36}$$

Equation (35) says $\mu_Y = \mu_X \circ f^*$ which makes perfect sense when μ_X and μ_Y are considered as unilinear forms. It says the following diagram

$$\begin{tikzcd} & \mathbb{R} \\ V^* \arrow[r, "f^*"] \arrow[ur, "\mu_Y"] & U^* \arrow[u, "\mu_X"'] \end{tikzcd}$$

commutes. But (35) makes no sense when we want to consider μ_X an element of U and μ_Y an element of V, in which case we should have $\mu_Y = f(\mu_X)$ mimicking $Y = f(X)$. To see that, we rewrite (35) as

$$\langle \mu_Y, \eta \rangle = \langle \mu_X, f^*(\eta) \rangle = \langle f(\mu_X), \eta \rangle$$

and this seems to say $\mu_Y = f(\mu_X)$, and this is correct, but only because of our using the representations $U^{**} = U$ and $V^{**} = V$.

Rewriting (36) using the canonical bilinear form we get

$$\bigl\langle \Sigma_Y(\eta_1), \eta_2 \bigr\rangle = \bigl\langle \Sigma_X\bigl(f^*(\eta_1)\bigr), f^*(\eta_2) \bigr\rangle = \Bigl\langle f\Bigl(\Sigma_X\bigl(f^*(\eta_1)\bigr)\Bigr), \eta_2 \Bigr\rangle$$

and this says

$$\Sigma_Y = f \circ \Sigma_X \circ f^* \tag{37}$$

meaning the following diagram

$$\begin{tikzcd} V^* \arrow[r, "f^*"] \arrow[rrr, bend right, "\Sigma_Y"'] & U^* \arrow[r, "\Sigma_X"] & U \arrow[r, "f"] & V \end{tikzcd} \tag{38}$$

commutes. Both (37) and (38) depend on our using the representations $U^{**} = U$ and $V^{**} = V$.

If we take $U = V = V^* = U^* = \mathbb{R}^d$ and $\langle x, y \rangle = x^T y$ then (37) says

$$M_Y = B M_X B^T$$

where M_X and M_Y are the symmetric positive semidefinite matrices that represent the variance functions and B is the matrix that represents the linear transformation f. So (37) does generalize what we know from multivariate probability theory.

9.10 Differentiation

9.10.1 Definition

The abstract theory of differentiation is less familiar to statisticians than the abstract theory of integration, but it can also be found in functional analysis. Here we follow Lang (1993). The reader must excuse the appearance of Banach spaces (possibly infinite-dimensional complete normed vector spaces). We will specialize to the finite-dimensional special case as soon as the definitions are finished. The reason we introduce the functional analysis definitions is to show that coordinates (isomorphism to $\mathbb{R}^d$) play no essential role in differentiation, contrary to what one learns in multivariable calculus.

Let U and V be Banach spaces. Then $L(U, V)$ denotes the set of all continuous linear maps $U \to V$, which is itself a vector space, the operations being given by (24a) and (24b). It is also a Banach space (Lang, 1993, pp. 65–66) when given the norm defined by

$$\|f\| = \sup_{\substack{x \in U \\ \|x\| \le 1}} \|f(x)\| \tag{39}$$

in which the $\| \cdot \|$ notation refers to three different norms: the expression $\|x\|$ refers to the norm of U, the expression $\|f(x)\|$ refers to the norm of V, and the expression $\|f\|$ refers to the norm for $L(U, V)$ which (39) defines.

Let O be open in U and let $f : O \to V$ be a map. Then f is *differentiable* at a point $x \in O$ if there exists $g \in L(U, V)$ such that

$$\lim_{h \to 0} \frac{f(x + h) - f(x) - g(h)}{\|h\|} = 0. \tag{40}$$

in which case g is the unique element of $L(U, V)$ having this property (Lang, 1993, p. 334) and we say that g is the *derivative of f at x* and write $f'(x) = g$.

If f is differentiable at every point of O, then it defines a map $x \mapsto f'(x)$ from O to $L(U, V)$. If this map is continuous, we say f is *continuously differentiable*.

When this differentiation theory on Banach spaces is specialized to abstract finite-dimensional vector spaces, it becomes simpler. On a finite-dimensional vector space, *every* linear function is continuous (as discussed in Section 8.14 above), so $L(U, V)$ consists of *all* linear functions $U \to V$ (one doesn't need to say "continuous linear function" in the finite-dimensional case). Moreover, on a finite-dimensional vector space, all norms are equivalent (Lang, 1993, Corollary 3.14 of Chapter II), meaning for any norms $\| \cdot \|_1$ and $\| \cdot \|_2$ there exist constants c_1 and c_2 such that $\|x\|_1 \le c_2 \|x\|_2$ and $\|x\|_2 \le c_1 \|x\|_1$ for all x, so the choice of norm does not affect the derivative. This also means that every norm on a finite-dimensional vector space induces the same topology, but we already knew that. Norms induce vector topologies and there is only one vector topology a finite-dimensional vector space can have (Section 8.14 above).

9.10.2 Philosophy

This is very different from the conceptualization of the derivative one gets from calculus. There the derivative of a scalar-to-scalar function is just a number; here it is a linear function. The correspondence is that the slope of the linear function is the derivative in the ordinary calculus sense.

In multivariate calculus the derivative of a vector-to-vector function $f : \mathbb{R}^d \to \mathbb{R}^e$ is the matrix of partial derivatives $\partial f_i(x) / \partial x_j$ (the so-called Jacobian matrix); here it is a linear function $f'(x) : \mathbb{R}^d \to \mathbb{R}^e$, the linear function represented by the Jacobian matrix (that is, if J is the Jacobian matrix, then the linear function is $x \mapsto Jx$).

On an abstract finite-dimensional vector space there are no coordinates hence no Jacobian matrix. We can introduce coordinates, but there are many ways to do so, and each gives a different Jacobian matrix. But the abstract derivative is a unique linear function, which does not depend on coordinates.

9.10.3 Higher Order Derivatives

Second and higher derivatives are just derivatives of derivatives. If the map $f' : O \to L(U, V)$ where O is open in U is differentiable at x we write its derivative as $f''(x)$. It is, by definition, an element of $L\bigl(U, L(U, V)\bigr)$. Its value at some point $h_1 \in U$, written $f''(x)(h_1)$ is an element of $L(U, V)$. And in turn, the value of this at some point $h_2 \in U$, written $f''(x)(h_1)(h_2)$ is an element of V.

The map $(h_1, h_2) \mapsto f''(x)(h_1)(h_2)$ is bilinear (linear in both arguments) and continuous $U \times U \to V$. Thus we can also consider $f''(x)$ a continuous bilinear form on U (Lang, 1993, p. 343, ff.). If f is twice continuously differentiable, meaning the map $x \mapsto f''(x)$ is continuous from some neighborhood of x in U to $L\bigl(U, L(U, V)\bigr)$, then this bilinear form is symmetric, meaning

$$f''(x)(h_1)(h_2) = f''(x)(h_2)(h_1), \qquad h_1, h_2 \in U.$$

Similarly, a continuous third derivative can be identified with a symmetric trilinear form, a continuous fourth derivative with a symmetric tetralinear form, and so forth.

The language of type theory comes in handy here too. The "official" type of a second derivative is $U \to U \to V$, a linear function $U \to L(U, V)$, but we also want to consider the uncurried form $U \times U \to V$, a bilinear form.

9.10.4 More Philosophy

So higher derivatives are even more different from the conceptualization of higher derivatives one gets from calculus. Instead of second, third, fourth, etc. derivatives of scalar-to-scalar functions being just numbers, they are now symmetric bilinear, trilinear, tetralinear, etc. forms. That seems crazy, but for vector-to-vector functions it is not so crazy. Once the number of indices gets to more than two, so partial derivatives can no longer be laid out in a matrix, as with second derivatives of a vector-to-vector function $\partial^2 f_i(x) / \partial x_j \partial x_k$, the conventions of multivariable calculus become inconvenient too.

Crazy or not, we will use the PhD level real analysis theory of derivatives as linear functions or multilinear forms in the rest of this document.

9.10.5 The Chain Rule

Let U, V, and W be Banach spaces, let O be open in U and P be open in V, and let $f : O \to P$ and $g : P \to W$ be maps. Then the *chain rule* says that if f is differentiable at x and g is differentiable at $f(x)$, then the composition $h = g \circ f$ is differentiable at x and its derivative is given by

$$h'(x) = g'\bigl(f(x)\bigr) \circ f'(x)$$

(Lang, 1993, p. 337). This says that $h'(x)$ is the composition of linear functions

$$\begin{tikzcd} U \arrow[r, "f'(x)"] \arrow[dr, swap, "h'(x)"] & V \arrow[d, "g'\bigl(f(x)\bigr)"] \\ & W \end{tikzcd}$$

9.10.6 Linearity of Differentiation

As in ordinary calculus, and as is obvious from the definition of differentiation (and uniqueness of the derivative), the derivative of a constant function is zero (the zero linear function) and the derivative of a linear function is itself.

In detail, let U and V be vector spaces and let O be open in U. If $f : O \to V$ is a constant function, then $f'(x)$ is the zero function $U \to V$. If $f : U \to V$ is a linear function, then $f'(x) = f$ for all $x \in O$. And then, since f' is a constant function, $f''(x)$ is the zero function $U \to U \to V$.

As in ordinary calculus, and as is obvious from the definition of differentiation (and uniqueness of the derivative), the derivative of a sum is the sum of the derivatives. If $h = f + g$, then $h'(x) = f'(x) + g'(x)$. In particular, the derivative of an affine function, the sum of a constant function and a linear function, is that linear function.

9.10.7 Transfer for Differentiation

If $f : \mathbb{R}^d \to \mathbb{R}^e$ is differentiable, then the derivative is the linear function represented by the matrix of partial derivatives (Browder, 1996, Theorem 8.21). The converse statement is false (Browder, 1996, Example 8.22), but if the partial derivatives are continuous functions, then f is continuously differentiable (Browder, 1996, Theorem 8.23). So, as long as we restrict our attention to continuously differentiable functions, there is no difference for $\mathbb{R}^d \to \mathbb{R}^e$ functions, between abstract differentiability and what we know from multivariable calculus.

Now suppose U and V are abstract finite-dimensional vector spaces of dimensions d and e, respectively, suppose O is open in U, and suppose $f : O \to V$ is continuously differentiable. If we want to calculate using multivariable calculus, we need isomorphisms $g : U \to \mathbb{R}^d$ and $h : V \to \mathbb{R}^e$. Then the map $j : \mathbb{R}^d \to \mathbb{R}^e$ defined by $j = h \circ f \circ g^{-1}$ "represents" f in multivariable calculus

$$\begin{tikzcd}
\mathbb{R}^d \arrow[r, "j"] & \mathbb{R}^e \\
U \arrow[u, "g"] \arrow[r, "f"] & V \arrow[u, "h"']
\end{tikzcd}$$

Having gotten ahold of j, we can differentiate it by multivariable calculus (represented by the matrix of partial derivatives) and, since $f = h^{-1} \circ j \circ g$, the chain rule and the linear function rule gives the derivative

$$f'(x) = h^{-1} \circ j'\bigl(g(x)\bigr) \circ g$$

that is, letting $y = g(x)$, the diagram

$$\begin{tikzcd}
\mathbb{R}^d \arrow[r, "j'(y)"] & \mathbb{R}^e \\
U \arrow[u, "g"] \arrow[r, "f'(x)"] & V \arrow[u, "h"']
\end{tikzcd}$$

is commutative. This is "transfer" for differentiation.

9.11 Integration

If we have a topology, then we know the open sets and the Borel sigma-algebra (the smallest sigma-algebra containing the open sets). So we can identify Borel measures and Borel-measurable functions and integrals of real-valued Borel-measurable functions with respect to Borel measures.

Transfer of integrals by linear isomorphism from $\mathbb{R}^d$ to any abstract finite-dimensional vector space is accomplished by the change-of-variable theorem for abstract integration.

For any measurable function f from a measurable space $(A, \mathcal{A})$ to a measurable space $(B, \mathcal{B})$, any measure μ on A induces a measure ν on B defined by

$$\nu(C) = \mu\bigl(f^{-1}(C)\bigr), \qquad C \in \mathcal{B},$$

where

$$f^{-1}(C) = \{\, x \in A : f(x) \in C \,\}$$

defines the set-to-set inverse of f. This ν is called the *image* of μ under f, and this operation is denoted $\nu = \mu \circ f^{-1}$. Moreover, when μ and ν have this relation, a real-valued function g on B is integrable with respect to ν if and only if $g \circ f$ is integrable with respect to μ, in which case

$$\int (g \circ f) \, d\mu = \int g \, d\nu = \int g \, d(\mu \circ f^{-1})$$

(Billingsley, 1979, Theorem 16.12).

The other change-of-variable theorem, the one for densities with respect to Lebesgue measure (the one involving Jacobian determinants), which involves differentiation, is the next section.

9.12 Lebesgue Measure

Suppose $f : \mathbb{R}^d \to V$ is a vector space isomorphism, λ is Lebesgue measure on $\mathbb{R}^d$, and $\mu = \lambda \circ f^{-1}$ (notation defined in Section 9.11 above). Then μ is a translation-invariant measure meaning

$$\mu(x + B) = \mu(B), \qquad \text{for all } x \in V \text{ and all Borel subsets } B \text{ of } V,$$

where

$$x + B = \{\, x + y : y \in B \,\}. \tag{41}$$

Moreover, every nonempty open set has positive μ measure. So μ has most of the properties of Lebesgue measure.

The only important property that is lacking is uniqueness. Different isomorphisms f may induce different measures μ. Lebesgue measure on $\mathbb{R}^d$ assigns hypervolume one to the unit cube. In an abstract vector space there is no notion of hypervolume because there is no preferred basis or, what is the same thing considered another way, because there is no metric. We can compare lengths of parallel vectors: it is clear that $2x$ and $-2x$ are twice as long as x. But if x is not a scalar multiple of y, there is no way to compare lengths of x and y.

To see what is happening, consider another vector space isomorphism $g : \mathbb{R}^d \to V$, and let $\nu = \lambda \circ g^{-1}$. What is the relationship between μ and ν?

Let $\rho = \nu \circ f = \lambda \circ g^{-1} \circ f$. Since ρ is a measure on $\mathbb{R}^d$ we know how to relate it to λ. By the change-of-variable theorem for integrals with respect to Lebesgue measure on $\mathbb{R}^d$ (the theorem about Jacobians) $d\rho$ is $d\lambda$ times the Jacobian determinant of the matrix representing the linear function $f^{-1} \circ g$. Since every linear function has a constant derivative, the Jacobian determinant is a scalar constant. Thus we see that ρ is just a constant times λ. And, since $\mu = \lambda \circ f^{-1}$ and $\nu = \rho \circ f^{-1}$, it follows that μ and ν are also constant multiples of each other.

Thus Lebesgue measure on a finite-dimensional vector space is uniquely defined up to multiplication by positive scalars. But that is as unique as it gets. We say each measure like μ and ν defined above is a *version* of Lebesgue measure, and we consider no version in any way special.

If we want to define densities with respect to Lebesgue measure, they have to be unnormalized densities (because no version of Lebesgue measure is special). We can say let h be an unnormalized probability density with respect to Lebesgue measure on V, and this means h is a nonnegative function such that $\int h \, d\mu$ is nonzero and finite, where μ is any version of Lebesgue measure. When we want to evaluate an expectation, it has the form

$$E_h\{g(X)\} = \frac{\int g(x)h(x)\mu(dx)}{\int h(x)\,\mu(dx)}$$

where μ is any version of Lebesgue measure (the arbitrary unknown normalization cancels so the expectation is unique).

Suppose X is a random vector in an abstract finite-dimensional vector space U, suppose g is a one-to-one map from an open subset O of U that is a support of X to another abstract finite-dimensional vector space V, define $Y = g(X)$, and suppose X has an unnormalized density h_X with respect to Lebesgue measure on U, then we expect the usual change-of-variable formula for one-to-one functions of continuous random vectors

$$h_Y(y) = h_X\bigl(g^{-1}(y)\bigr) \cdot J(y)$$

defines an unnormalized density with respect to Lebesgue measure on V for Y, where g^{-1} is the inverse of the codomain restriction of g, that is, the inverse of g considered as a function $O \mapsto g(O)$, and where $J(y)$ is the "Jacobian" whatever that means. Transfer (Section 9.10.7 above) tells us what it has to mean. Let f_X be an isomorphism $U \to \mathbb{R}^d$ and f_Y be an isomorphism $V \to \mathbb{R}^d$. Define $W = f_Y(g(O))$, and define a map $W \to \mathbb{R}^d$ by $j = f_X \circ g^{-1} \circ f_Y^{-1}$, where f_Y^{-1} here means the restriction of f_Y^{-1} to W. Let $J(y)$ be the absolute value of the determinant of $j'(y)$. This we can calculate because j is a map from an open subset of $\mathbb{R}^d$ to $\mathbb{R}^d$. Of course, $J(y)$ is only determined up to an unknown constant because its value depends on the choice of isomorphisms f_X and f_Y, but since h_Y is only supposed to be an unnormalized density anyway, this does not matter.

9.13 Moment Generating Functions

The moment generating function (MGF) of a random vector X in an abstract finite-dimensional vector space V, if X has an MGF, is the nonlinear function M on V^* defined by

$$M(\eta) = E\bigl\{e^{\langle X, \eta \rangle}\bigr\}, \qquad \eta \in V^*$$

(we define $M(\eta) = \infty$ at η such that the expectation does not exist). We say that M is the MGF of X if M is finite on a neighborhood of zero. Otherwise we say that X does not have an MGF.

The reason for the name is that derivatives of M evaluated at zero are the ordinary moments of X, that is,

$$\begin{aligned} M'(0) &= \alpha_1 \\ M''(0) &= \alpha_2 \\ M'''(0) &= \alpha_3 \end{aligned}$$

and so forth (in the first line both sides are unilinear forms on V^*, in the second line both sides are symmetric bilinear forms on V^*, and so forth). If M is an MGF, then it is infinitely differentiable, and X has moments of all orders.

All of this is provable by transfer. Let X be a random vector in an abstract finite-dimensional vector space V, let $f : \mathbb{R}^d \to V$ be a vector space isomorphism, and let $Y = f^{-1}(X)$. Letting M_X and M_Y denote the MGF's of X and Y, if they exist, we have

$$\begin{aligned} M_X(\eta) &= E\bigl\{e^{\langle f(Y), \eta \rangle}\bigr\} \\ &= E\bigl\{e^{\langle Y, f^*(\eta) \rangle}\bigr\} \\ &= M_Y\bigl(f^*(\eta)\bigr) \end{aligned}$$

so $M_X = M_Y \circ f^*$. Hence X has an MGF if and only if Y does, and if they do, the derivatives of M_X are given by

$$\begin{aligned} M_X'(\eta)(\zeta) &= M_Y'\bigl(f^*(\eta)\bigr)\bigl(f^*(\zeta)\bigr) \\ M_X''(\eta)(\zeta_1)(\zeta_2) &= M_Y''\bigl(f^*(\eta)\bigr)\bigl(f^*(\zeta_1)\bigr)\bigl(f^*(\zeta_2)\bigr) \end{aligned}$$

and so forth. So

$$\begin{aligned} M_X'(0)(\zeta) &= M_Y'(0)\bigl(f^*(\zeta)\bigr) \\ &= E\bigl\{\langle Y, f^*(\zeta) \rangle\bigr\} \\ &= E\bigl\{\langle f(Y), \zeta \rangle\bigr\} \\ &= E\bigl\{\langle X, \zeta \rangle\bigr\} \\ M_X''(0)(\zeta_1)(\zeta_2) &= M_Y''(0)\bigl(f^*(\zeta_1)\bigr)\bigl(f^*(\zeta_2)\bigr) \\ &= E\bigl\{\langle Y, f^*(\zeta_1) \rangle \langle Y, f^*(\zeta_2) \rangle\bigr\} \\ &= E\bigl\{\langle X, \zeta_1 \rangle \langle X, \zeta_2 \rangle\bigr\} \end{aligned}$$

and so forth.

9.14 Cumulant Generating Functions

The log of the MGF of X, if it exists, is called the *cumulant generating function* (CGF) of X. Derivatives of the CGF evaluated at zero are called the *cumulants* of X.

Cumulants of order m are polynomial functions of the ordinary moments up to order m and vice versa (Cramér, 1951, Section 15.10). Here we will only be interested in the first two cumulants, which are the mean and the variance. If M is the MGF and K is the CGF, then

$$\begin{aligned} K'(\eta)(\zeta) &= \frac{M'(\eta)(\zeta)}{M(\eta)} \\ K''(\eta)(\zeta_1)(\zeta_2) &= \frac{M''(\eta)(\zeta_1)(\zeta_2)}{M(\eta)} - \frac{M'(\eta)(\zeta_1)}{M(\eta)} \cdot \frac{M'(\eta)(\zeta_2)}{M(\eta)} \end{aligned}$$

so, since $M(0) = 1$,

$$\begin{aligned} K'(0)(\zeta) &= M'(0)(\zeta) \\ K''(0)(\zeta_1)(\zeta_2) &= M''(0)(\zeta_1)(\zeta_2) - M'(0)(\zeta_1) \cdot M'(0)(\zeta_2) \end{aligned}$$

and this together with

$$\mu_2(\zeta_1)(\zeta_2) = \alpha_2(\zeta_1)(\zeta_2) - \alpha_1(\zeta_1) \cdot \alpha_1(\zeta_2),$$

which is the vector analog of $\operatorname{var}(X) = E(X^2) - E(X)^2$ and is proved the same way, by linearity of expectation, finishes the proof that the first two cumulants are mean and variance.

9.15 Law of Large Numbers

Transfer for the law of large numbers (LLN) is obvious. If X_1, X_2, ... is a sequence of IID random vectors having mean μ in an abstract finite-dimensional vector space V, if $f : V \to \mathbb{R}^d$ is a linear isomorphism, and we define $Y_i = f(X_i)$ for all i, then we know Y_1, Y_2, ... is a sequence of IID random vectors having mean $f(\mu)$. Define

$$\begin{aligned} \overline{X}_n &= \frac{1}{n} \sum_{i=1}^{n} X_i \\ \overline{Y}_n &= \frac{1}{n} \sum_{i=1}^{n} Y_i \end{aligned} \tag{42}$$

then we know $\overline{Y}_n = f(\overline{X}_n)$ by linearity. The LLN for $\mathbb{R}^d$ tells us

$$\overline{Y}_n \xrightarrow{\text{a. s.}} f(\mu),$$

and applying f^{-1} to both sides gives

$$\overline{X}_n \xrightarrow{\text{a. s.}} \mu.$$

9.16 Normal Distributions

The *standard normal* distribution on $\mathbb{R}$ is the continuous distribution having unnormalized probability density with respect to Lebesgue measure $z \mapsto \exp(-z^2/2)$.

The *standard normal* distribution on $\mathbb{R}^d$ is the continuous distribution of a random vector having IID standard normal components. It has unnormalized density with respect to Lebesgue measure $z \mapsto \exp(-z^T z/2)$.

A general *normal distribution* on an abstract finite-dimensional vector space V is the image of a standard normal distribution on $\mathbb{R}^d$ for some d under an affine function $\mathbb{R}^d \to V$

Like general normal distributions on $\mathbb{R}^d$, general normal distributions on an abstract finite-dimensional vector space come in two kinds. *Degenerate* normal distributions are concentrated on hyperplanes. They cannot have unnormalized densities with respect to Lebesgue measure. They are the ones whose variances, considered as bilinear forms, are not positive definite (Section 9.7 above). *Nondegenerate* normal distributions are not concentrated on any proper affine subspace. They give positive probability to any open set and do have unnormalized densities with respect to Lebesgue measure.

We know the unnormalized density of a general normal distribution on $\mathbb{R}^d$, written in matrix notation, is

$$x \mapsto \exp\bigl(-(x-\mu)^T \Sigma^{-1} (x-\mu)/2\bigr).$$

Our notation is

$$(x-\mu)^T \Sigma^{-1} (x-\mu)$$

when we think of x, μ, and Σ as matrices. Our notation is

$$\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle$$

when we think of x and μ as vectors and Σ as a linear function whose inverse linear function is Σ^{-1}. Our notation is

$$\Sigma^{-1}(x-\mu)(x-\mu)$$

when we think of x and μ as vectors and Σ^{-1} as the bilinear form that corresponds to Σ^{-1} as a linear function.

Now suppose X is a nondegenerate normal random vector on an abstract finite-dimensional vector space V, and suppose $f : \mathbb{R}^d \to V$ is an isomorphism. Then $Y = f^{-1}(X)$ is also a nondegenerate normal random vector, and the variances of X and Y are related by (37). By $X = f(Y)$ and the change of variable theorem in Section 9.12 above we get that the unnormalized density of X is given by

$$h_X(x) = h_Y(f(x)) = \exp\bigl(-\langle f(x) - \mu_Y, \Sigma_Y^{-1}(f(x) - \mu_Y)\rangle/2\bigr)$$

and

$$\begin{aligned} \langle f(x) - \mu_Y, \Sigma_Y^{-1}(f(x) - \mu_Y)\rangle &= \langle f(x) - f(\mu_X), \Sigma_Y^{-1}[f(x) - f(\mu_Y)]\rangle \\ &= \langle f(x - \mu_X), \Sigma_Y^{-1} f(x - \mu_Y)\rangle \\ &= \langle x - \mu_X, (f^* \circ \Sigma_Y^{-1} \circ f)(x - \mu_Y)\rangle \end{aligned}$$

the second equality being linearity of f, and

$$\begin{aligned} (f^* \circ \Sigma_Y^{-1} \circ f)^{-1} &= f^{-1} \circ \Sigma_Y \circ (f^*)^{-1} \\ &= f^{-1} \circ \Sigma_Y \circ (f^{-1})^* \\ &= \Sigma_X \end{aligned}$$

the second equality being the inverse of an adjoint is the adjoint of the inverse (Halmos, 1974, Section 44) and (37). Hence

$$h(x) = \exp\bigl(-\Sigma_X^{-1}(x - \mu_X)(x - \mu_X)/2\bigr)$$

gives an unnormalized probability density with respect to Lebesgue measure on an abstract finite-dimensional vector space. (This is transfer for densities with respect to Lebesgue measure.) Densities have the same form on abstract finite-dimensional vector spaces as on $\mathbb{R}^d$. We just have to write them using the correct notation.

9.17 Convergence in Distribution

If X_1, X_2, $\ldots$ is a sequence of random vectors in a finite-dimensional vector space V and X is another random vector in V, then we say the sequence *converges in distribution* to Y if

$$E\{f(X_n)\} \to E\{f(X)\}$$

for every bounded continuous function $f : V \to \mathbb{R}$, and to indicate this we write

$$X_n \xrightarrow{\mathcal{D}} Y.$$

Convergence in distribution is also called *convergence in law* and *weak convergence.*

Some readers may not recognize this definition, being instead familiar with a definition involving convergence of distribution functions (Ferguson, 1996, Definition 1 of Chapter 1). However, our definition is equivalent to that one by a result known as the Helly-Bray theorem (Ferguson, 1996, Theorem 3 of Chapter 3). The definition adopted here is more general being the one always used in general complete separable metric spaces (Billingsley, 1999, Chapter 1).

One might wonder why we are using this definition for abstract finite-dimensional vector spaces, which don't have a metric. But they can be given a metric. There is just no unique way to do so. But the definition of convergence in distribution only depends on the topology (which is unique, Section 8.14 above), because the topology determines which functions are continuous (those for which inverse images of open sets are open). So the definition of convergence in distribution only depends on the topology not on the metric.

Our basic tool for working with convergence in distribution is the continuous mapping theorem (Billingsley, 1999, Theorem 2.7), which says that if $X_n \xrightarrow{\mathcal{D}} X$ and g is a continuous function, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$. (Actually the theorem says more than that. It is enough for g to be continuous on a set having probability one under the distribution of X. But we will not need this refinement.)

9.18 The Central Limit Theorem

Suppose X_1, X_2, $\ldots$ is a sequence of IID random vectors in an abstract finite-dimensional vector space V having mean μ_X and variance Σ_Y. Then the central limit theorem (CLT) says

$$\sqrt{n}(\overline{X}_n - \mu_X) \xrightarrow{\mathcal{D}} \text{Normal}(0, \Sigma_X), \tag{43}$$

where $\overline{X}_n$ is defined in (42).

This is proved by transfer, which is just like transfer for the LLN *mutatis mutandis.* Suppose $f : V \to \mathbb{R}^d$ is a linear isomorphism, and define $Y_i = f(X_i)$ for all i. Then Y_1, Y_2, $\ldots$ is a sequence of IID random vectors having mean $\mu_Y = f(\mu_X)$ and variance Σ_Y given by (37). Define $\overline{X}_n$ and $\overline{Y}_n$ by (42). The CLT for $\mathbb{R}^d$ tells us

$$\sqrt{n}(\overline{Y}_n - \mu_Y) \xrightarrow{\mathcal{D}} Z,$$

where Z has a normal distribution with mean zero and variance Σ_Y. By linearity of f

$$\begin{aligned}
\sqrt{n}(\overline{Y}_n - \mu_Y) &= \sqrt{n} \left(\left[\frac{1}{n} \sum_{i=1}^n f(X_i) \right] - f(\mu_X) \right) \\
&= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n [f(X_i) - f(\mu_X)] \right) \\
&= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n f(X_i - \mu_X) \right) \\
&= f \left(\sqrt{n} \left\{ \left[\frac{1}{n} \sum_{i=1}^n X_i \right] - \mu_X \right\} \right) \\
&= f \left(\sqrt{n} \left[\overline{X}_n - \mu_X \right] \right)
\end{aligned}$$

so

$$\sqrt{n}(\overline{X}_n - \mu_X) = f^{-1}\bigl(\sqrt{n}(\overline{Y}_n - \mu_Y)\bigr) \xrightarrow{\mathcal{D}} f^{-1}(Z)$$

by the continuous mapping theorem. And $f^{-1}(Z)$ is a linear function of multivariate normal is multivariate normal. Its mean is $f^{-1}(0) = 0$. and its variance is by (37)

$$\begin{aligned}
f^{-1} \circ \Sigma_Y \circ (f^{-1})^* &= f^{-1} \circ f \circ \Sigma_X \circ f^* \circ (f^{-1})^* \\
&= f^{-1} \circ f \circ \Sigma_X \circ f^* \circ (f^*)^{-1} \\
&= \Sigma_X
\end{aligned}$$

and thus we get the usual form of the CLT in abstract finite dimensional vector spaces (43).

9.19 The Delta Method

The delta method says that if

$$\sqrt{n}(X_n - \xi) \xrightarrow{\mathcal{D}} Y \tag{44}$$

and g is a function that is differentiable at ξ, then

$$\sqrt{n}\bigl(g(X_n) - g(\xi)\bigr) \xrightarrow{\mathcal{D}} g'(\xi)(Y). \tag{45}$$

The delta method is usually only proved for the $\mathbb{R}^d$ case (Ferguson, 1996, Theorem 7 of Chapter 7). We need to transfer that result to the abstract finite-dimensional vector space case.

We again adopt the notation used in the Section 9.15. Let $f : \mathbb{R}^d \to V$ be a linear isomorphism, $W_n = f^{-1}(X_n)$, and $Z = f^{-1}(Y)$. Also write $\omega = f^{-1}(\xi)$. Then, by assumption (44),

$$\sqrt{n}\bigl(f(W_n) - f(\omega)\bigr) = \sqrt{n}(X_n - \xi) \xrightarrow{\mathcal{D}} Y = f(Z).$$

Write $h = g \circ f$. Then by the $\mathbb{R}^d$ case of the delta method we have

$$\sqrt{n}\bigl(g(X_n) - g(\xi)\bigr) = \sqrt{n}\bigl(h(W_n) - h(\omega)\bigr) \xrightarrow{\mathcal{D}} h'(\omega)(Z)$$

By the chain rule $h'(\omega) = g'(\xi) \circ f$, so

$$h'(\omega)(Z) = g'(\xi)\bigl(f(Z)\bigr) = g'(\xi)\bigl(Y\bigr)$$

10 Expectations and Moments on Affine Spaces

Now we want to redo everything in Section 9 changing vector spaces to affine spaces. Fortunately much of the redo is trivial.

A random point of an affine space is just like a random vector of a vector space, described by a Borel probability measure on the space.

Ordinary moments of random points in an affine space, at least as they are defined in Section 9.2, make no sense because an affine space has no dual space or canonical bilinear form. So ordinary moments cannot be defined as in Section 9.2.

10.1 Mean

We try to define means, avoiding dual and double dual spaces. We take a hint from functional analysis, where the following sort of definition is used to define expectation of infinite-dimensional random vectors.

Suppose X is a random point in a topological affine space A, the *mean* of X, also called the *expectation* of X, is the point μ satisfying

$$E\{f(X)\} = f(\mu), \qquad \text{for all affine functions } f : A \to \mathbb{R}, \tag{46}$$

provided all of these expectations exist (if any do not exist, we say the mean of X does not exist or that X has no expectation).

Theorem 38. *The mean of a random vector as defined in Sections 9.2 and 9.3 can also be characterized by the definition given just above.*

Proof. We have already learned in Section 9.3 that $E\{g(X)\} = g(\mu)$ for any random vector X and any linear function g and that this depends on identifying the vector space where X lives with its double dual. And we alreay know that an affine function on a vector space is a linear function plus a constant function (Corollary 17). Thus what we need to show is

$$E\{a + g(X)\} = a + g(\mu)$$

for any linear function g and any constant a (which is a vector in the codomain of g). But that follows from the elementary properties of expectation (Theorem 37). We have now proved that (46) holds for the mean defined as in Sections 9.2 and 9.3.

It remains to be shown that if (46) holds, then there is at most one vector μ that satisfies it. But this holds by definition. Among the affine functions $A \to \mathbb{R}$ are the linear functionals (when A is a vector space), and if we just restrict to them, (46) says

$$E\{\langle X, \eta \rangle\} = \langle \mu, \eta \rangle, \qquad \text{for all } \eta \in A^*,$$

and this is just the identification of the mean considered as a linear functional on A^* and as an element of A using the natural isomorphism $A \to A^{**}$. □

So now we know (at least as far as vector spaces go), that (46) is compatible without old definition of mean (also called expectation). In order for it to be a good extension of our definition to affine spaces, we need another theorem that is essentially the last part of the preceding theorem extended to affine spaces.

Theorem 39. *If A is a finite-dimensional affine space, X is a random point of A, and all of the expectations in* (46) *exist, then there is exactly one point $\mu \in A$ such that* (46) *holds.*

Before proving the theorem, we prove two trivial lemmas.

Lemma 40. *If A is an affine space containing distinct points x and y, then there exists an affine function $f : A \to \mathbb{R}$ such that $f(x) \neq f(y)$.*

The short way to state the assertion of the lemma is “affine functions separate points of an affine space.”

Proof. We already know the corresponding fact for vector spaces (linear functionals separate points of a vector space); this is part of the assertion of Theorem 30, as is clear from its proof.

By the definition of affine function in Section 8.3, a function $f : A \to \mathbb{R}$ is affine if and only if the function $g : V \to \mathbb{R}$ defined by

$$g(v) = f(x + v) - f(x), \qquad v \in V,$$

is linear. Write $y = x + v$, so $v = y - x$. Then

$$g(y - x) = f(y) - f(x).$$

So all we need is that whenever $v \neq 0$, there is a linear functional g such that $g(v) \neq 0$, and that is guaranteed by Theorem 30. □

Lemma 41. *The expectation of a constant point in an affine space is that constant.*

Proof. Suppose $X = \mu$ almost surely, where X is a random point in the affine space A. Then for any affine function $f : A \to \mathbb{R}$ we have $f(X) = f(\mu)$ almost surely. Hence $E\{f(X)\} = f(\mu)$. So (46) is satisfied.

Conversely, if $\nu \neq \mu$ is a point that satisfies (46) with μ replaced by ν, then we would have $f(\mu) = f(\nu)$ and this implies $\mu = \nu$ by Lemma 40. □

Proof of Theorem 39. Fix $z \in A$. Continue with the notation that μ_W is the mean of the random vector W or random point W as the case may be. We claim that

$$\mu_X = z + \mu_{X-z} \tag{47}$$

is the unique μ satisfying (46), where μ_{X-z} (being the mean of a random vector) is defined by either our old or new definition.

Let V be the translation space of A. By the definition of affine function in Section 8.3, a function $f : A \to \mathbb{R}$ is affine if and only if the function $g : V \to \mathbb{R}$ defined by

$$g(v) = f(z + v) - f(z), \qquad v \in V,$$

is linear. And linearity of g implies (by Theorem 38) that

$$E\{g(X - z)\} = g(E\{X - z\}) = g(\mu_{X-z}).$$

Thus we have

$$E\{f(X) - f(z)\} = g(\mu_{X-z}) = f(z + \mu_{X-z}) - f(z).$$

But also

$$E\{f(X) - f(z)\} = E\{f(X)\} - f(z)$$

by the usual rules of probability for random scalars. Hence

$$E\{f(X)\} = f(z + \mu_{X-z})$$

holds for all affine functions $f : A \to \mathbb{R}$. And that says (47) is one μ that satisfies (46).

Uniqueness now follows from Lemma 40. □

Corollary 42. *If A and B are finite-dimensional affine spaces, X is a random point in A, and $f : A \to B$ is an affine function, then*

$$E\{f(X)\} = f(E\{X\}). \tag{48}$$

Proof. We already know that (48) holds by definition when $B = \mathbb{R}$. Hence if $g : B \to \mathbb{R}$ is an affine function, then we know.

$$E\{(g \circ f)(X)\} = g\bigl(f(E\{X\})\bigr)$$

and we also know (again by definition of mean for affine spaces) we can take g out of the left-hand side obtaining

$$g\bigl(E\{f(X)\}\bigr) = g\bigl(f(E\{X\})\bigr).$$

Now by Lemma 40, this holding for all affine functions $g : B \to \mathbb{R}$ implies (48). □

10.2 Sample Mean

If $x_1, \ldots, x_n$ are points in a finite-dimensional affine space, it makes no sense to define the sample mean by

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

because neither addition of points nor multiplication of points by scalars is defined.

However, we can define the empirical distribution P_n by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_i).$$

This is the distribution that results from taking a random point from the "population" $x_1, \ldots, x_n$.

Now we can define the sample mean using the theory in Section 10.1, which says there is a unique μ that satisfies

$$\frac{1}{n} \sum_{i=1}^n f(x_i) = f(\mu), \qquad \text{for all affine functions } f : A \to \mathbb{R},$$

and we denote this μ by $\bar{x}_n$.

And now we make everything random. Suppose $X_1, \ldots, X_n$ are random points in A, then there is for each ω in the underlying probability space a unique μ that satisfies

$$\frac{1}{n} \sum_{i=1}^n f\bigl(X_i(\omega)\bigr) = f(\mu), \qquad \text{for all affine functions } f : A \to \mathbb{R},$$

and we denote this μ by $\overline{X}_n(\omega)$, and this defines a random point $\overline{X}_n$.

10.3 Law of Large Numbers

The proofs in Section 9.15 require almost no change to apply to affine spaces when the sample and population means are defined as in the preceding two sections. If X_1, X_2, $\ldots$ is a sequence of IID random points in a finite-dimensional affine space A, then A must be nonempty. And for any affine isomorphism $f : A \to \mathbb{R}^d$, we have

$$f(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

by Corollary 42 applied to the empirical distribution. Then the LLN for random scalars, says this converges almost surely to

$$E\{f(X_i)\} = f(\mu)$$

by Corollary 42 applied to the population distribution. But since f is a topological isomorphism, this implies

$$\overline{X}_n \xrightarrow{\text{a. s.}} \mu.$$

10.4 Central Limit Theorem

The proofs in Section 9.18 require almost no change to apply to affine spaces when the sample and population means are defined as in the preceding two sections. If X_1, X_2, $\ldots$ is a sequence of IID random points in a finite-dimensional affine space A, and $E(X_i) = \mu$, then $X_i - \mu$ is a vector (in the translation space of A) and so is $\overline{X}_n - \mu$. Thus the CLT is about vectors rather than points. Also note that the variance in the CLT is the variance of these vectors. We do not need a definition of variance on affine spaces.

10.5 Differentiation

Let A and B be finite-dimensional vector spaces having translation spaces U and V, respectively. Let O be open in U and let $f : O \to B$ be a function. Then f is *differentiable* at a point $x \in O$ if

$$\lim_{h \to 0} \frac{f(x + h) - f(x) - g(h)}{\|h\|} = 0.$$

holds. This is an exact copy of (40), which was used to define differentiation on vector spaces (Section 9.10.1 above), but it may require some re-interpretation in this context.

In order for $f(x)$ to make sense, we must have $x \in A$. In order for $f(x + h)$ to make sense, we must have $x + h \in A$. But that makes $h \in U$. We must have $f(x)$ and $f(x+h)$ in B. But that makes $f(x+h) - f(x) \in V$. Since subtraction of a point from a vector is undefined, $g(h)$ must be a vector that can be added to $f(x + h) - f(X)$, that is, $g(h)$ is also in V. So $g : U \to V$. As in Section 9.10.1 we assume g is a linear function. Also as in Section 9.10.1 we can take $\| \cdot \|$ to be any norm on U because all norms on a finite-dimensional vector space are equivalent.

Again, as in Section 9.10.1 the function g is unique if it exists and we call it the *derivative of f at x* and write $f'(x) = g$.

Then everything is the same as in Section 9.10. If $f'(\,\cdot\,)$ is a continuous function on O, we say f is *continuously differentiable* on O. If $f'(\,\cdot\,)$ is differentiable at x, we say its derivative, denoted $f''(x)$ is the *second derivative of f at x*. And so forth.

So even though our functions map between affine spaces, their derivatives are still linear functions that map between vector spaces.

Those who have been exposed to differential geometry, will know that this sort of construction goes beyond affine spaces to so-called *manifolds*, which are topological spaces that are locally topologically isomorphic to finite-dimensional vector spaces (like smooth curved surfaces in Euclidean space, the surface of a sphere or torus, for example). Then the derivative of a function from one manifold to another is defined to be a linear function between vector spaces, which are called the *tangent spaces* to the manifolds in question at the points in question. If we are differentiating f at x, then the tangent spaces in question are those at x and $f(x)$.

Unlike the cases discussed in this section, where the manifolds in question are affine spaces, for general manifolds the tangent spaces change as x and $f(x)$ change.

But we are not interested in differential geometry except as it applies to affine spaces. So the only change we need to our theory of differentiation to make it match up with differential geometry is to change the names of the spaces U and V from *translation space* to *tangent space.*

10.6 The Delta Method

Now we see that nothing needs to change from Section 9.19 to apply the delta method to random points of finite-dimensional affine spaces. We still have

$$\sqrt{n}(X_n - \xi) \xrightarrow{\mathcal{D}} Y \tag{49a}$$

implies

$$\sqrt{n}\bigl(g(X_n) - g(\xi)\bigr) \xrightarrow{\mathcal{D}} g'(\xi)(Y) \tag{49b}$$

(these are exact copies of equations (44) and (45) in Section 9.19), but they may require some re-interpretation in this context.

Let A and B be affine spaces having translation spaces (a. k. a., tangent spaces) U and V, respectively. Suppose X_1, X_2, ... are a sequence of random points of A and ξ is a nonrandom point of A. Then (49a) is interpreted the same as in Section 9.17 because $X_n - \xi$ is a vector (an element of U), so we only need the theory of convergence in distribution for random vectors.

Similarly, in (49b), $g(X_n)$ and $g(\xi)$ are elements of B (one random, the other nonrandom) so their difference is a random vector in V. On the right-hand side of this equation Y is a random vector in U and $g'(\xi)$ is a (nonrandom) linear function $U \to V$, so the right-hand side is a random vector in V (like the left-hand side). Everything works with the theory we already have for random vectors.

10.7 Moments, Cumulants, and Generating Functions

We have seen in the preceding several sections that most of what we need for statistical theory carries over to the setting of affine spaces. But not everything does. General ordinary moments have no analog for random points in affine spaces. Central moments do (when we use the definition of mean given in Section 10.1) because the central moments of the random point X are the ordinary moments of the random vector $X - \mu$, where $\mu = E(X)$, and we do have a theory of moments of random vectors.

Similarly, moment and cumulant generating functions make no sense for random points in affine spaces. The best we can do is use moment and cumulant functions for some random vector. If X is a random point in a finite dimensional affine space A having translation space U, and $a \in A$ is a constant point, then the moment or cumulant generating function of the random vector $X - a$ tells us most of what moments can tell us about X. And anything it doesn't tell us, we cannot have.

10.8 Exponential Families

We have the elegant definition of exponential families on affine spaces (Geyer, 1990, Section 1.4): a family of probability densities with respect to a positive Borel measure on a finite-dimensional affine space is a *standard exponential family* if the log densities are affine functions, and a family of probability densities with respect to an arbitrary positive sigma-finite measure (on an arbitrary measurable space) is an *exponential family* if it has a sufficient statistic that is a random point in a finite-dimensional affine space and the densities are affine functions of that sufficient statistic (see Section 1.4 in Geyer, 1990, for details).

We will eventually see that this definition has a lot of virtues. But it seems weird to most statisticians in that it discusses a parametric statistical model without discussing parameters.

By reduction by sufficiency, it is enough to consider standard families. So let λ be a positive Borel measure on a finite-dimensional affine space A having translation space V. Let $\mathcal{H}$ denote the set of all affine functions $h : A \to \mathbb{R}$ such that

$$\int e^h \, d\lambda = \int e^{h(x)} \, \lambda(dx) = 1.$$

This is the family of log densities of the *full* standard exponential family *generated* by λ.

If we compare this characterization with that given in Section 5, we see that we want

$$h(x) = \langle x, \theta \rangle - c(\theta), \tag{50}$$

and the right-hand side is indeed an affine function of x. And we also see that we can recover the canonical parameterization by differentiating (with respect to x, not θ)

$$\theta = \nabla h(x) \tag{51}$$

(h, being an affine function, has constant derivative, that is, not a function of x, so the left-hand side is not a function of x, despite appearances).

But this only makes sense when X is a random vector, not a random point. When X is a random point (in an affine space), (50) makes no sense, although (51) does make sense and give a parameterization of the model.

But if we want to write the densities as a function of the parameter, it seems we have to give in an move to vector spaces. We can validly write (even when X is a random points)

$$h_\theta(x) = \langle X - z, \theta \rangle - c(\theta),$$

where z is an arbitrary (nonrandom) point in A and

$$c(\theta) = \log\left(\int e^{x - z, \theta} \, \lambda(dx) \right), \tag{52}$$

so the log likelihood is

$$l(\theta) = \langle X - z, \theta \rangle - c(\theta), \qquad \theta \in V^*.$$

We still define the full family to be *regular* if

$$\{ \, \nabla h(x) : h \in \mathcal{H} \, \}$$

is an open subset of V^*. That definition need not mention cumulant functions at all.

11 Mean Value Parameterization Again

The mean value parameterization was defined back in Section 7.2. Specifically, equation (11) defines a function g that maps the canonical parameter to the mean value parameter.

That was for exponential families on abstract vector spaces. For exponential families on abstract affine spaces, as defined in the preceding section, if $\mathcal{H}$ is the family of log densities with respect to the generating measure λ of the family, then, of course, we can define mean values by

$$\mu(h) = E_h(X) = \int x e^{h(x)} \, \lambda(dx), \qquad h \in \mathcal{H}.$$

Theorem 43. *The mean value parameterization for a regular full exponential family on an affine space exists (each distribution has a mean) and is identifiable (different distributions have different means).*

Proof. We already know this for exponential families on vector spaces (Theorem 16). Consider a standard exponential family on a finite-dimensional affine space A having translation space V, and let z be an arbitrary, fixed, nonrandom point of A. If X is the canonical statistic of the family, then $X - z$ is the canonical statistic of a standard exponential family on V, and this family is also full and regular (both families have the same canonical parameter space). Hence we know from Theorem 16 that each distribution has a mean for $X - z$, hence for X, and different distributions have different means for $X - z$, hence for X. □

REVISED DOWN TO HERE

Theorem 44. *The mean value parameterization of a regular full exponential family maps its canonical parameter space to the relative interior of its convex support.*

Proof. The proof is almost the same as the proof of Theorem 14. The only difference is that in the log likelihood we replace the observed value of the canonical statistic y with an arbitrary μ in the affine space where the family lives. Let us call the result

$$q(\theta) = \langle \mu - z, \theta \rangle - c(\theta),$$

where the cumulant function c is given by (52).

This is a concave function having η as a direction of recession if and only if $\langle Y - \mu, \eta \rangle \le 0$ almost surely, where Y is the canonical statistic. The proof of this is the same as the proof of Theorem 10 with q replacing the log likelihood. Now following the proof of Theorem 14 with q replacing l we get that the maximum of q is achieved if and only if every direction of recession of q is a direction of constancy of q. That is,

$$\langle y - \mu, \eta \rangle \le 0, \qquad \text{for all } y \in K$$

implies

$$\langle y - \mu, \eta \rangle = 0, \qquad \text{for all } y \in K.$$

Or, what is the same thing, if

$$\langle \mu, \eta \rangle \le \sigma_K(\eta) \tag{53a}$$

implies

$$\langle \mu, \eta \rangle = \sigma_K(\eta). \tag{53b}$$

If $\mu \in \operatorname{ri} K$, and

The convex support is nonempty. Hence it has a nonempty relative interior (Theorem 34 □

References

Barndorff-Nielsen, O. (1978). *Information and Exponential Families.* Wiley, Chichester.

Billingsley, P. (1979). *Probability and Measure.* Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, second edition. Wiley, New York.

Browder, A. (1996). *Mathematical Analysis: An Introduction.* Springer-Verlag, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, CA.

Cramér, H. (1951). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman & Hall, London.

Geyer, C. J. (1990). Likelihood and Exponential Families. PhD thesis, University of Washington. `http://purl.umn.edu/56330`.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In *Stochastic Geometry: Likelihood and Computation*, W. Kendall, O. Barndorff-Nielsen and M. N. M. van Lieshout, eds. Chapman and Hall/CRC, London, 141–172.

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, C. J. (2013). Stat 8112 Lecture Notes. Asymptotics of exponential families. `http://www.stat.umn.edu/geyer/8112/notes/expfam.pdf`.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Halmos, P. (1974). *Finite Dimensional Vector Spaces.* Springer-Verlag, New York. Reprint of second edition (1958), originally published by Van Nostrand

Lang, S. (1993). *Real and Functional Analysis*, third edition. Springer-Verlag, New York.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, NJ.

Rockafellar, R. T., and Wets, R. J.-B. (1998). *Variational Analysis.* Springer-Verlag, Berlin. (The corrected printings contain extensive changes. We used the third corrected printing, 2010.)

Rudin, W. (1991). *Functional Analysis*, second edition. McGraw-Hill, Boston.