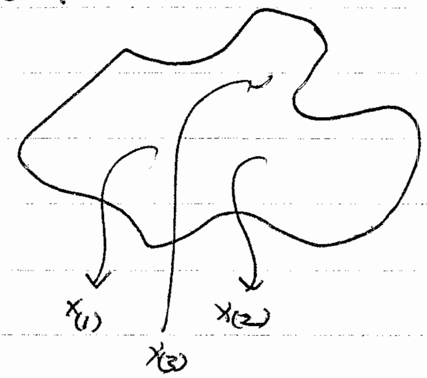
II Ensembles, Estimators, Gaussians

To model "noise" or variability, we postulate an ensemble Ω of possible observations, & consider each observation as a random draw from Ω.



For univariate quantities, we can characterize Ω by $p(x)$, where $p(x)dx$ is the probability that a random draw is between x & $x + dx$.

[because it is too hard]

Typically we are not interested in determining $p(x)$, but rather "statistics" (functionals) of $p(x)$, e.g.,

$$\text{mean} = \int_\Omega x\,p(x)\,dx = \langle x \rangle_\Omega$$

$$\text{variance} = \int_\Omega (x - \langle x \rangle)^2 p(x)\,dx = \langle (x - \langle x \rangle_\Omega)^2 \rangle_\Omega$$

$$\text{entropy} = \int_\Omega -p(x) \ln p(x)\,dx = -\langle \ln p(x) \rangle_\Omega$$

Similarly, for x multivariate ($\vec{x} = (x_1, x_2, \cdots, x_N)$):

$$\text{covariance}_{jk} = \int_\Omega (x_j - \langle x_j \rangle)(x_k - \langle x_k \rangle)\,d\vec{x}.$$

Time series can be considered as multivariate quantities

$$s(t) = (s(t_1), s(t_2), s(t_3), \cdots)$$

An *estimator* for a statistic Θ is a procedure for taking a set of observations $x_{(1)}, x_{(2)}, \cdots, x_{(K)}$ into an estimate Θ^{EST}.

We expect that $\Theta^{EST}(x_{(1)}, \cdots, x_{(K)})$ will not yield the true value $\hat{\Theta} = \hat{\Theta}(\Omega)$, but it should be "close".

Two kinds of errors

$$\underline{\text{BIAS}} = \left\langle \Theta^{EST} - \hat{\Theta} \right\rangle_{\text{all draws of } K \text{ samples}}$$

$$\underline{\text{VARIANCE}} = \left\langle \left(\Theta^{EST} - \left\langle \Theta^{EST} \right\rangle_{K \text{ draws}} \right)^2 \right\rangle_{K \text{ draws}}$$

$$\underline{\text{VARIANCE} + (\text{BIAS})^2} = \left\langle \left(\Theta^{EST} - \hat{\Theta} \right)^2 \right\rangle_{K \text{ draws}}$$

The "plug-in" estimator replaces $\int_\Omega$ by $\frac{1}{K}\sum$.

Plug-in estimator for mean:

$$\Theta^{PI} = \frac{1}{K}\sum_{j=1}^{K} x_{(j)}.$$

This is unbiased, since $\left\langle \Theta^{PI} \right\rangle_{K \text{ draws}} = \frac{1}{K}\sum_{j=1}^{K} \left\langle x_{(j)} \right\rangle_\Omega$

So is the plug-in estimator for any statistic that is linear in $\vec{x}$, e.g.

$$\sum a_r x_r.$$

3

The plug-in estimator for the variance is BIASED

$$\Theta^{PI} = \frac{1}{K}\sum_{j=1}^{K}\left(X_{(j)} - \frac{1}{K}\sum_{m=1}^{k} X_{(m)}\right)^2$$

$$= \frac{1}{K}\sum_{j=1}^{K} X_{(j)}^2 - \frac{1}{K^2}\sum_{j=1}^{K}\sum_{m=1}^{K} X_{(j)}X_{(m)}$$

$$= \frac{1}{K}\sum_{j=1}^{K} X_{(j)}^2 - \left(\frac{1}{K}\sum_{j=1}^{K} X_{(j)}\right)^2$$

If Ω has variance 1, & mean 0, then $\left\langle \sum_{j=1}^{k} X_{(j)}^2 \right\rangle = K$

$$\left\langle \left(\sum_{j=1}^{k} X_{(j)}\right)^2 \right\rangle = K$$

since $\langle X_{(j)} X_{(m)} \rangle = 0$ on separate draws

so $\Theta^{PI} = \frac{1}{K}\cdot K - \frac{1}{K^2}\cdot K$

$= 1 - 1/K$, not 1.

So we typically use $\frac{\Theta^{PI}}{K-1} = \frac{1}{K-1}\sum_{j=1}^{k}\left(X_{(j)} - \frac{1}{K}\sum_{m=1}^{k} X_{(m)}\right)^2$

as an unbiased estimator of the variance.

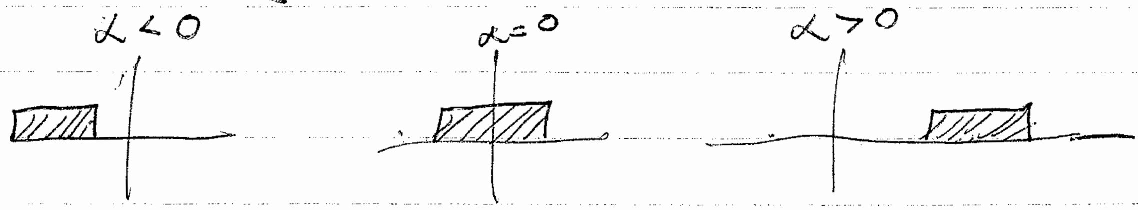
The above analyses were "lucky" – bias could be calculated independent of Ω. But this is not typical. And we also don't know Ω (otherwise, we wouldn't need to measure $\vec{X}_{(1)}, \ldots$ to estimate Θ).

So we need a more general setup:

Assume Ω is part of a parametric family $\Omega(\vec{\alpha})$.

E.g., [univariate case]

$$\Omega(\alpha) = \begin{cases} 1, & |x-\alpha| \le \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$



α can be multivariate.

In the above case, the "plug-in" estimator for the mean is unbiased, but has unnecessarily high variance.

A better estimator: choose the highest observation x_{max}

" " lowest " x_{min}

take $\theta = \frac{1}{2}(x_{min} + x_{max})$.

Similarly, if distrib looks like

it is best to discard the extremes before averaging.

Moral: plug-in estimator neither typically unbiased nor least variance.

The above set-up leads to 3 generic kinds of estimators (at least):

A. Maximum likelihood estimator

Given the observations $\vec{x}_{(1)}, \cdots, \vec{x}_{(K)}$, calculate, for each $\vec{\alpha}$, the likelihood of this set of observations

$$P(\vec{\alpha}) = \prod_{m=1}^{K} p\left(\vec{x}_{(m)} \text{ drawn from } \Omega(\vec{\alpha})\right)$$

Find the $\vec{\alpha}_{MLE}$ that maximizes $P(\vec{\alpha})$.

Then $\Theta^{MLE} = \hat{\Theta}(\Omega(\vec{\alpha}_{MLE}))$.

B. Bayesian estimators

Assume some prior distribution for $\vec{\alpha}$, say, $Q(\vec{\alpha})$.
Given the observations $\vec{x}_{(1)}, \cdots, \vec{x}_{(K)}$, the a priori distribution is modified to an a posteriori distribution

$$Q^{post}(\vec{\alpha}) = \frac{P(\vec{\alpha})\, Q(\vec{\alpha})}{\int P(\vec{\alpha}')\, Q(\vec{\alpha}')\, d\vec{\alpha}'}$$

Can take $\Theta^{BAYES,1} = \int \hat{\Theta}(\Omega(\vec{\alpha}))\, Q^{post}(\vec{\alpha})\, d\vec{\alpha}$

or

$$\Theta^{BAYES,\infty} = \hat{\Theta}(\Omega(\vec{\beta})) \text{ where } Q^{post}(\vec{\beta}) \text{ is max}$$

or $$\Theta^{BAYES,gen} = \frac{\int \hat{\Theta}(\Omega(\vec{\alpha}))\, g(Q^{post}(\vec{\alpha}))\, d\vec{\alpha}}{\int g(Q^{post}(\vec{\alpha}))\, d\vec{\alpha}}$$

C. "Best" estimator(s).
use the procedure that minimizes

$$\left\langle \left(\Theta - \hat{\Theta}(\alpha)\right)^2 \right\rangle_\alpha$$

or, use the procedure that minimizes $\max_\alpha \langle (\Theta - \hat{\Theta}(\alpha))^2 \rangle$

All of the above depends on choosing $\Omega(\vec{\alpha})$. Usually Gaussians.
Two reasons:

- Central Limit Thm
- Maximum Entropy property of Gaussians

For a Gaussian, the plug-in estimator for the mean, and the corrected plug-in estimator for the variance, are the MLE's.

Bayesian + "best" estimators may differ; $\Omega(\vec{\alpha})$ could be bizarre.

Central Limit Theorem

Informally: Say x_1 is drawn from p_1
x_2 " " " p_2
⋮
x_k " " " p_k and
no one of these dominate. Then,

as $k \to \infty$, $\frac{1}{K}\sum_{j=1}^{k} x_j$ is distributed "like" a Gaussian.

The plausibility argument will tell us how to make the conditions more precise.

Let $y_2 = x_1 + x_2$. The p.d.f. for y_2 is

$$q_2(y_2) = \int_{-\infty}^{\infty} p_1(y_2 - x)\, p_2(x)\, dx$$

So, with $\hat{p}(\omega) = \int e^{-i\omega y} p(y)\, dy$, etc,

$$\hat{q}_2(\omega) = \hat{p}_1(\omega)\,\hat{p}_2(\omega).$$

Similarly for $y_k = x_1 + \cdots + x_k$, $\hat{q}_k(\omega) = \prod_{m=1}^{k} \hat{p}_m(\omega)$.

Let $z_k = y_k / k$, + say $r_k(z_k)$ is the p.d.f. for z_k

$$\hat{r}(\omega) = \int e^{-i\omega z} r_k(z)\, dz$$

$$= \int e^{-i\omega z} q_k(kz) \cdot k\, dz$$

$$\hat{r}(\omega) = \int e^{-i\omega y/k} q_k(y)\, dy = \hat{q}_k\left(\frac{\omega}{k}\right) = \prod_{m=1}^{k} \hat{p}_m\left(\frac{\omega}{k}\right)$$

$$\log \hat{r}(\omega) = \sum_{m=1}^{k} \log\left(\hat{p}_m\left(\frac{\omega}{k}\right)\right)$$

8

For a p.d.f. $p(x)$, $\log(\hat{p}(\omega))$ is called the "characteristic function" $C_p(\omega)$, r obeys

$$C_{p*q}(\omega) = C_p(\omega) + C_q(\omega).$$

Write

$$C_p(\omega) = \sum_{s=0}^{\infty} A_s \frac{(-i)^s}{s!} \omega^s, \quad \text{(Interpret the } A\text{'s later)}$$

$$\hat{p}(\omega) = \int e^{-i\omega x} p(x)\,dx, \text{ so } \quad \begin{aligned} \hat{p}(0) &= 1 \\ \hat{p}'(0) &= -i\langle x\rangle \\ \hat{p}''(0) &= -\langle x^2\rangle \\ \frac{d^s}{d\omega^s}\hat{p}(0) &= (-i)^s\langle x^s\rangle. \end{aligned}$$

So $$\hat{p}(\omega) = \sum_{s=0}^{\infty} \frac{\omega^s}{s!}(-i)^s\langle x^s\rangle$$ [Fine print: the Taylor series must exist.]

$$\hat{p}(\omega) = 1 - i\omega\langle x\rangle - \frac{\omega^2}{2}\langle x^2\rangle + \frac{i\omega^3}{6}\langle x^3\rangle \cdots$$

Let's assume that each p_m has mean 0. Then

$$\hat{p}_m\left(\frac{\omega}{k}\right) = 1 - \frac{\omega^2}{2k^2}\langle x_m^2\rangle + \frac{i\omega^3}{6k^3}\langle x_m^3\rangle \cdots$$

As $k \to \infty$, since $\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots$, we can approximate $\log \hat{p}_m\left(\frac{\omega}{k}\right)$ as:

$$\log \hat{p}_m\left(\frac{\omega}{k}\right) = -\frac{\omega^2}{2k^2}\langle x_m^2\rangle + \frac{i\omega^3}{6k^3}\langle x_m^3\rangle \cdots$$
$$-\frac{1}{2}\left(-\frac{\omega^2}{2k^2}\langle x_m^2\rangle + \frac{i\omega^3}{6k^3}\langle x_m^3\rangle\cdots\right)$$
$$+\frac{1}{3}\left(\cdots\right)^3\cdots$$

Only one term that is $O\left(\frac{1}{k^2}\right)$. So

$$\log \hat{r}(\omega) = \sum_{m=1}^{k} \log \hat{p}_m\left(\frac{\omega}{k}\right) = -\frac{\omega^2}{2k^2}\sum_{m=1}^{k}\langle x_m^2\rangle + k\,O\left(\frac{\omega^3}{k^3}\right) + \cdots$$

[fine point: need $\sum_{m=1}^{k}\langle x_m^2\rangle$ to grow in proportion to k to be sure that term dominates, similarly for $\langle x_m^3\rangle$.]

$$\ast \quad \hat{r}(\omega) \approx e^{-\frac{\omega^2}{2k^2}\sum_{m=1}^{k}\langle x_m^2\rangle} \approx e^{-\frac{\omega^2}{2k}V}$$

where V is the average variance of the x_m's.

Can now estimate $r(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega z}\,\hat{r}(\omega)\,d\omega$

$$\approx \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{V/k}}\, e^{-\frac{1}{2}z^2/(V/k)}$$

(we'll do these integrals later) a Gaussian of variance V/k.

[More fine point: $\approx$ in $\ast$ means: second moment well approximated, higher moments not well-approximated.]

Interpretation of the A's. $A_k(p*q) = A_k(p) + A_k(q)$

"semi-invariants"
"cumulants"

$C_p(\omega) = \log(\hat{p}(\omega)) =$

$$-i\omega\langle x\rangle - \frac{\omega^2}{2}\langle x^2\rangle + \frac{i\omega^3}{6}\langle x^3\rangle \cdots$$

$$-\frac{1}{2}\left(-i\omega\langle x\rangle - \frac{\omega^2}{2}\langle x^2\rangle + \frac{i\omega^3}{6}\langle x^3\rangle \cdots\right)^2$$

$$+\frac{1}{3}\left(-i\omega\langle x\rangle - \frac{\omega^2}{2}\langle x^2\rangle + \frac{i\omega^3}{6}\langle x^3\rangle \cdots\right)^3 \cdots$$

$$\sum_{s=0}^{\infty} A_s \frac{(-i)^s}{s!}\omega^s = -i\omega\langle x\rangle - \frac{\omega^2}{2}\langle x^2\rangle + \frac{1}{2}\omega^2\langle x\rangle^2$$

$$+\frac{i\omega^3}{6}\langle x^3\rangle - \frac{1}{2}\left(\frac{i\omega^3}{2}\right)2\langle x\rangle\langle x^2\rangle + \frac{1}{3}(-i)^3\langle x\rangle^3$$

$$+ O(\omega^4)$$

$$-iA_1 = -i\langle x\rangle \Rightarrow A_1 = \langle x\rangle$$

$$-\frac{A_2}{2!} = -\frac{1}{2}\langle x^2\rangle + \frac{1}{2}\langle x\rangle^2 \Rightarrow A_2 = \langle x^2\rangle - \langle x\rangle^2$$

$$(-i)^3\frac{A_3}{3!} = \frac{i}{6}\langle x^3\rangle - \frac{i}{2}\langle x\rangle\langle x^2\rangle + (-i)^3\frac{\langle x\rangle^3}{3}$$

$$A_3 = \langle x^3\rangle - 3\langle x\rangle\langle x^2\rangle + 2\langle x\rangle^3$$

A_1 = mean, A_2 = variance, $\frac{A_3}{(A_2)^{3/2}}$ = skewness, $\frac{A_4}{(A_2)^2}$ = kurtosis

11

Some Gaussian integrals

$$I = \int \cdots \int e^{-(\vec{x}-\vec{a})M(\vec{x}-\vec{a})^T/2} e^{-\vec{x}\cdot\vec{b}} \, d\vec{x}$$

where M is a $K \times K$ matrix, symmetric

$$(\vec{x}-\vec{a})M(\vec{x}-\vec{a})^T = \sum_{j,m=1}^{K} (x_j - a_j)(x_m - a_m) M_{jm}$$

Say we can write $M = RDR^{-1}$ $= RDR^T$ $\begin{cases} R & \text{real + unitary} \\ D & \text{diagonal, positive eigenvalues} \end{cases}$

$$(\vec{x}-\vec{a})M(\vec{x}-\vec{a})^T = \vec{y} D \vec{y}^T, \quad \vec{y} = (\vec{x}-\vec{a})R$$

[if D had an eigenvalue ≤ 0, then the integral would not $\to 0$ as $\vec{y} = \alpha \cdot$ corresp. eigenvector]

Say $D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{pmatrix}$, put $z_1 = y_1\sqrt{\lambda_1}$, $\vdots$, $z_K = y_K\sqrt{\lambda_K}$

$$\vec{z} = \vec{y}\sqrt{D} = (\vec{x}-\vec{a})R\sqrt{D}$$

$$(\vec{x}-\vec{a})M(\vec{x}-\vec{a}) = \vec{z}\vec{z}^T.$$

$$\vec{x} = \vec{z}\sqrt{D}^{-1}R^T + \vec{a}$$

$$d\vec{x} = \det\left(\sqrt{D}^{-1}R\right) d\vec{z} = \frac{1}{\sqrt{\det M}} d\vec{z} \cdot \quad (\det R = 1).$$

12

So $I = \int \cdots \int e^{-\vec{z}\vec{z}^T/2} \, e^{(\vec{z}\sqrt{D}^{-1}R^T + \vec{a})\cdot\vec{b}} \dfrac{d\vec{z}}{\sqrt{\det M}}$

$$= \frac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \int \cdots \int e^{-\vec{z}\vec{z}^T/2} \, e^{\vec{z}\sqrt{D}^{-1}R^T \cdot \vec{b}} \, d\vec{z}$$

Now $\vec{z}\vec{z}^T = z_1^2 + z_2^2 + \cdots + z_k^2$

and

$\vec{z}\sqrt{D}^{-1}R^T \cdot \vec{b} = \sum c_m z_m$ for some $\vec{c}^T = \sqrt{D}^{-1}R^T b^T$

So $I = \dfrac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \int \cdots \int e^{-(z_1^2 - z_2^2 \cdots z_k^2)/2} \, e^{\sum c_m z_m} \, dz_1 \cdots dz_k$

$$= \frac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \prod_{m=1}^{K} \int_{-\infty}^{\infty} e^{-\frac{z_m^2}{2} + c_m z_m} \, dz_m$$

$$= \frac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \prod_{m=1}^{K} e^{+c_m^2/2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z_m^2 - 2c_m z_m + c_m^2)} \, dz_m$$

$$= \frac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \, e^{\sum c_m^2/2} \left(\sqrt{2\pi}\right)^K \qquad \left[\int_{-\infty}^{\infty} e^{-\omega^2/2} \, d\omega = \sqrt{2\pi}\right]$$

$$\sum c_m^2 = |\vec{c}_m|^2 = c\,c^T = (b R \sqrt{D}^{-1})(\sqrt{D}^{-1} R^T b^T) = b M^{-1} b^T$$

$$I = \frac{e^{\vec{a}\cdot\vec{b}}}{\sqrt{\det M}} \cdot (2\pi)^{K/2} \cdot e^{b M^{-1} b^T/2}$$

Say $M = V^{-1}$; $\vec{a} = 0$, $\vec{b} = 0$:

$$\int \cdots \int e^{-\vec{x} V^{-1} \vec{x}^T/2} \, d\vec{x} = \frac{(2\pi)^{k/2}}{\sqrt{\det M}} = \sqrt{\det V}\,(2\pi)^{k/2}$$

or

$$\frac{1}{\sqrt{2\pi}^k} \frac{1}{\sqrt{\det V}} \int \cdots \int e^{-\vec{x} V^{-1} \vec{x}^T/2} \, d\vec{x} = 1.$$

Similar argument* ($V_{km} = V_{mk}$)

$$\frac{1}{(\sqrt{2\pi})^k} \frac{1}{\sqrt{\det V}} \int \cdots \int x_k x_m \, e^{-\vec{x} V^{-1} \vec{x}^T/2} \, d\vec{x} = V_{km} \qquad (\dagger)$$

$$\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-u^2/2} \, du = 1 \right]$$

This establishes $\dfrac{1}{\sqrt{(2\pi)}^k} \dfrac{1}{\sqrt{\det V}} e^{-\vec{x} V^{-1} \vec{x}^T/2}$ as the p.d.f. of a Gaussian with variances & covariances given by $V = \{V_{km}\}$.

---

* Similar argument: $\vec{x}^T \vec{x} = (\vec{z} \sqrt{D}^{-1} R^T)^T \vec{z} \sqrt{D}^{-1} R^T$

$= R \sqrt{D}^{-1} \vec{z}^T \vec{z} \sqrt{D}^{-1} R^T$

So $\dfrac{1}{\sqrt{2\pi}^k} \dfrac{1}{\sqrt{\det V}} \int \cdots \int \vec{x}^T \vec{x} \, e^{-\vec{x} V^{-1} \vec{x}^T/2} \, d\vec{x} =$

$$R \sqrt{D}^{-1} \cdot \left(\frac{1}{\sqrt{2\pi}^k} \frac{1}{\sqrt{\det V}} \int \cdots \int \vec{z}^T \vec{z} \, e^{-\vec{z}\vec{z}^T/2} \sqrt{\det V} \, d\vec{z} \right) \sqrt{D}^{-1} R^T$$

$$= R D^{-1} R^T = M^{-1} = V$$

If $p(\vec{x}) = \frac{1}{(\sqrt{2\pi})^k} \frac{1}{\sqrt{\det V}} e^{-\vec{x} V^{-1} \vec{x}^T/2}$,

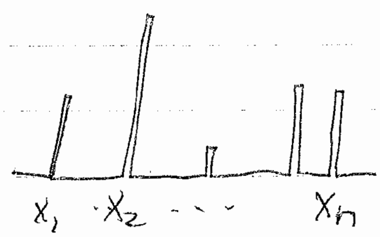
$$\tilde{p}(\vec{\omega}) = \int \cdots \int p(\vec{x})\, e^{-i\vec{\omega}\cdot\vec{x}}\, d\vec{x}$$

$$= e^{-\omega V \omega^T}. \quad [\text{See pg } \boxed{9}]$$

$$\left[\begin{array}{l} \vec{a} = 0 \\ M = V^{-1} \\ b = i\omega \\ b^T = -i\omega^T \end{array}\right.$$

Maximum Entropy Distribution

It would be nice to specify $\Omega(\vec{x})$ with only a few parameters -- in particular, would like Not to have to specify all moments, or, all $p(\vec{x})$!.

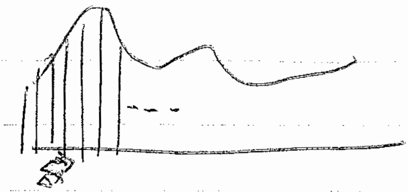
Entropy of a discrete distribution $p(x) =$ 

$$p(x) = \sum_{r=1}^{n} p_r\, \delta(x - x_r)$$

given by $H(p) = -\sum p_r \log p_r$

if $\log = \log_2$, $H(p) =$ # of yes-no questions, on average, needed to determine r, if each obs. occurs with frequency p_r.

For a continuous distribution:



? $$H(p) = \lim_{\Delta x \to 0} \sum_{x_i} \left(p(x_i)\Delta x\right) \log p(x_i \Delta x)$$

$$= \lim_{\Delta x \to 0} \sum_{x_i} p(x_i)\,\Delta x \left(\log p(x_i) + \log \Delta x\right)$$

$$= \lim_{\Delta x \to 0} \left[\sum_{x_i} p(x_i) \log p(x_i)\,\Delta x + (\log \Delta x) \sum_{x_i} p(x_i)\Delta x \right]$$

$$= \underbrace{\int p(x) \log p(x)\,dx}_{\text{"differential entropy" of } p} + \lim_{\Delta x \to 0} \underbrace{(\log \Delta x)}_{\text{nuisance}}$$

"Nuisance" is irrelevant for comparing entropies.

Plan: Specify only a few moments of p, by $\vec{\alpha}$.

Then, find $\Omega(\vec{\alpha})$ as the maximum differential entropy ensemble with moments $\vec{\alpha}$.

How to find a maxent distribution, given some constraints?

Discrete case:

$$\text{Entropy} = -\sum p_r \log(p_r)$$

Typical constraints: Mean $= \mu \Rightarrow \sum p_r x_r = \mu$

Variance $= \sigma^2 \Rightarrow \sum p_r (x_r - \mu)^2 = \sigma^2$

The constraints are linear in p.

k^{th} constraint: $\sum_{r=1}^{n} p_r C_{kr} = A_k$

[We even have a 0^{th} constraint: $\sum_{r=1}^{n} p_r = 1$, i.e., $C_{0r} = 1$]

Method of Lagrange Multipliers:

Say $f = f(u_1, u_2, \dots, u_n)$ is to be maximized, subject to constraints $c_k(u_1, \dots, u_n) = a_k$.

Let $\mathcal{F} = f(u_1, \dots, u_n) + \sum_{k=1}^{K} \lambda_k c_k(u_1, \dots, u_n)$.

The constrained extremum of $\mathcal{F}$ occurs where $\dfrac{\partial \mathcal{F}}{\partial u_j} = 0$.

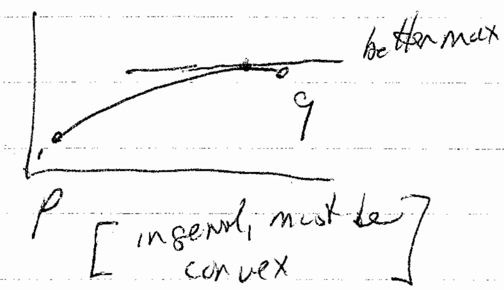
This typically leads to equations $u_j = U_j(\lambda_1, \dots, \lambda_K)$, from which the values of the λ's must be determined.

For entropy, the constrained maximum, if it exists, must be unique.

Say p + q are both local maxima that satisfy the constraints, with $H(p) = H(q)$ but $p \neq q$.

$$H((1-m)p + mq) > H(p) \text{ and } H(q)$$

[Mixing property of entropy].



∴ neither could be the max.

Apply Lagrange Multipliers to $-\sum_{r=1}^{n} p_r \log p_r$, & $\sum_{r=1}^{n} p_r C_{kr}$ as constraints

$$J = -\sum_{r=1}^{n} p_r \log p_r + \sum_{k=0}^{K} \lambda_k \sum_{r=1}^{n} p_r C_{kr}.$$

$$\frac{\partial J}{\partial p_j} = -1 + \log p_j + \sum_{k=0}^{K} \lambda_k C_{kj}$$

$$p_j = B\, e^{-\sum_{k=1}^{K} \lambda_k C_{kj}}$$

Note $C_{0r} = 1$

So, if we constrain variances + means, the C_{kj} will be:

! (replacing variance by 2^{nd} moment)

$$C_{1r} = x_r$$

$$C_{2r} = x_r^2.$$

Continuous univariate case:

$$\mathcal{J} = - \int p(x) \log p(x) dx + \sum_{k=0}^{K} \lambda_k \int p(x) C_k(x) dx$$

$C_{kr} \rightarrow C_k(x)$. want $\frac{\partial \mathcal{J}}{\partial [p(x)]}$ [formally, a "Calculus of Variations" problem]

$$\frac{\partial \mathcal{J}}{\partial [p(x)]} = -1 - \log p(x) + \sum_{k=0}^{K} \lambda_k C_k(x)$$

and again

$$p(x) = B\, e^{\sum_{k=1}^{K} \lambda_k C_k(x)} .$$

Mean: $C_1(x) = x$
Variance: $C_2(x) = x^2$.

So $p(x)$ is a Gaussian, + we need to choose λ_1 + λ_2 .

Easiest to first shift the problem so that $\mu = 0$, and then use (‡) p. $\boxed{13}$ to get $\lambda_2 = \frac{1}{2} V^{-1}$.

Multivariate case: one constraint for each variable's mean
" " " " " variance
" " " " covariance.

Same strategy.

Conclusion: The multivariate Gaussian is the maxent distribution for a constrained mean + variance.