STSCI 4780
Assigning direct probabilities: Sampling distributions and priors

Tom Loredo, CCAPS & SDS, Cornell University

Well-Posed Problems

The rules (BT, LTP, . . .) express desired probabilities in terms of other probabilities—they comprise a kind of *grammar* for inference

To get a numerical value *out*, at some point we have to put numerical values *in*—we need a *vocabulary*

Direct probabilities are probabilities with numerical values determined directly by premises/conditioning info (via modeling assumptions, symmetry arguments, previous calculations, desperate presumption . . .)

An inference problem is *well posed* only if all the needed direct probabilities are assignable. We may need to add new assumptions as we see what needs to be assigned. We may not be entirely comfortable with what we need to assume—remember Euclid's fifth postulate!

Should explore how results depend on convenient/defeasible assumptions ("robustness")

Lec04 recap: Essential contextual information

We can only be uncertain about a proposition, A, if there are alternatives (at least $\overline{A}$!); what they are will bear on our uncertainty. *We must explicitly specify relevant alternatives.*

Hypothesis space: The set of alternative hypotheses of interest (and auxiliary hypotheses needed to predict the data, e.g., for LTP)

Data/sample space: The set of possible data we may have predicted before learning of the observed data

Predictive model: Information specifying the likelihood function (e.g., the conditional predictive dist'n/sampling dist'n)

Other prior information: Any further information available or necessary to assume to make the problem *well posed*

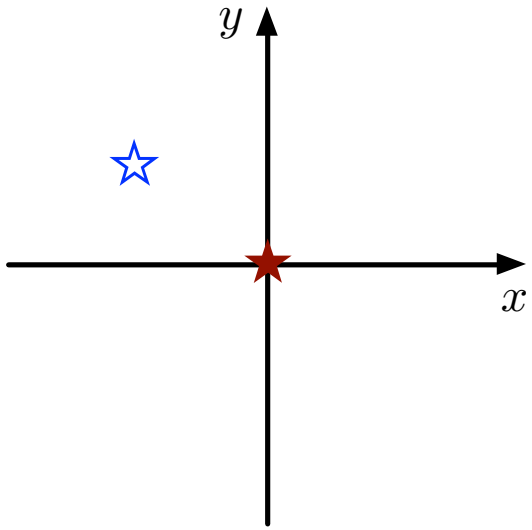
Where do predictive models (sampling dist'ns) come from? Seek patterns & approaches that may inform how we assign priors.

Directly assigned sampling distributions

Some examples of reasoning leading to sampling distributions:

- Binomial distribution:
  - Ansatz: Probability for a Bernoulli trial, α
  - LTP $\Rightarrow$ binomial for n successes in N trials
- Poisson distribution:
  - Ansatz: $P(\text{event in } dt|\lambda) \propto \lambda dt$; probabilities for events in disjoint intervals independent
  - Product & sum rules $\Rightarrow$ Poisson for n in T
- Gaussian distribution:
  - CLT: Probability theory for sum of many quantities with independent, finite-variance PDFs
  - Sufficiency (Gauss): Seek distribution with sample mean as sufficient statistic (also sample variance)
  - Asymptotic limits: large$n$ Binomial, Poisson
  - Others: Herschel's invariance argument (2-D), maximum entropy...

Herschel-Maxwell derivation of 2-D normal



- Knowledge of x tells us nothing about y:

$$\rho(x,y)dx\,dy = f(x)dx \times g(y)dy$$

- Same distribution in x and y:

$$\rho(x,y)dx\,dy = f(x)dx \times f(y)dy$$

- Express in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$:

$$\rho(x,y)dx\,dy = f(r\cos\theta)\,f(r\sin\theta)\,r\,dr\,d\theta$$

- Distribution independent of angle:

$$f(r\cos\theta)f(r\sin\theta) = g(r)$$
$$\Rightarrow f(x)f(y) = g\left(\sqrt{x^2+y^2}\right)$$

Solve this *functional equation*:

$$f(x)f(y) = g\left(\sqrt{x^2 + y^2}\right)$$

For $y = 0$, $f(x)f(0) = g(x)$; replacing $g(\cdot)$ gives

$$f(x)f(y) = f\left(\sqrt{x^2 + y^2}\right)\ f(0)$$

$$\ln\left[\frac{f(x)}{f(0)}\right] + \ln\left[\frac{f(y)}{f(0)}\right] = \ln\left[\frac{f\left(\sqrt{x^2 + y^2}\right)}{f(0)}\right]$$

Requires a function of x plus a function y that's a function only of $x^2 + y^2$:

$$\ln\left[\frac{f(x)}{f(0)}\right] = ax^2$$

Normalization requires $a < 0$ and determines the normalization constant $f(0)$:

$$f(x) = \sqrt{\frac{\alpha}{\pi}} e^{-\alpha x^2}, \quad \alpha > 0,$$

$$\rho(x, y) = \frac{\alpha}{\pi} e^{-\alpha(x^2+y^2)}$$

Maxwell extended argument to 3-D (velocities) $\Rightarrow$ Maxwellian velocity distribution in statistical mechanics

Theme/pattern: Qualitative requirements (typically from symmetry, invariance) $\Rightarrow$ constraints in the form of functional equations $\Rightarrow$ a specific family of dist'ns from solving the eqn's

Assigning priors

Sources of prior information

- *Discovery chains*: Analysis of previous experimental or observational data (but begs the question of what prior to use for the first such analysis)
- *Subjective priors*: Elicit a prior from an expert in the problem domain, e.g., via ranges, moments, quantiles, histograms (more radical *subjective Bayesians* assert an agent's priors need only express an opinion that agrees with bets the agent is willing to make)
- *Population priors*: When it's meaningful to pool observations, we potentially can *learn* a shared prior—hierarchical/graphical/multilevel models do this

"Non-informative" priors

- Seek a prior that in some sense (TBD!) expresses a lack of information prior to considering the data
- No universal solution—this notion must be problem-specific
- *Objective Bayes (OBayes)*—Bayesian inference using priors that are largely or entirely determined from specification of the parameters and likelihood function

Imagine a Stan++ package that uses the data, parameter, and model blocks to derive a prior when one isn't specified, purely from the problem specification—how might it work?

Discrete uniform prior

Seek an *algorithm* that assigns a discrete PMF $p_i = P(A_i|\mathcal{C})$ from symbolic expression of a problem

Problem 1, $\mathcal{C}_1$ with a suite of N propositions denoted A_i, assigns

$$p_i = P(A_i|\mathcal{C}_1)$$

Problem 1, $\mathcal{C}_2$ with the *same* suite of N propositions, but with different labels/symbols, B_i, assigning

$$p'_i = P(B_i|\mathcal{C}_2)$$

Suppose *semantically* we can identify equivalent propositions that happen to have different labels, e.g.,

$$B_1 \equiv A_2; \quad B_2 \equiv A_1$$
$$B_k \equiv A_k \text{ for } k = 3 \text{ to } N$$

Transformation equations: equate probabilities that refer to equivalent propositions (reflecting *semantic* equivalence),

$$\Rightarrow p_1 = p'_2,\ p_2 = p'_1,\ p_k = p'_k \text{ for } k > 2$$

Now suppose neither $\mathcal{C}_1$ and $\mathcal{C}_2$ express information distinguishing among each problem's N propositions; *symbolically/syntactically*, the two problems look equivalent (differing only in labels)

A rule that assigns a probability p_i to $p(A_i|\mathcal{C}_1)$ based solely on the pattern of symbols in the problem definition must assign the *same* value to $p(B_i|\mathcal{C}_2)$

$$P(A_i|\mathcal{C}_1) = P(B_i|\mathcal{C}_2)$$
$$\Rightarrow p_1 = p'_1,\ p_2 = p'_2, \ldots$$

Symmetry equations: Reflect *consistency requirement*—a formal rule should assign the same probabilities to problems with equivalent symbolic expressions (*syntactic* equivalence)

Combine transformation & symmetry:

$$p_1 = p_2$$

Consider further problems that differ from $\mathcal{C}_1$ by different permutations of the labels for the propositions $\Rightarrow$

$$p_i = p_j \text{ for all } i, j$$

Imposing normalization:

$$p_i = \frac{1}{N}$$

for a suite of N propositions with no information in $\mathcal{C}$ distinguishing between them

Known as the *principle of insufficient reason* (adopted without name by Bernoulli, Laplace), aka *principle of indifference* (Keynes, Jaynes)

Continuous uniform prior can't be universal

"Method of inverse probability"—Bayes's theorem with uniform/flat prior PDFs for *continuous* parameters (adopted for expedience)

Inverse probability was heavily used by Laplace and subsequently dominated statistical practice until late 19th/early 20th century

Criticism (Boole, Venn, others):

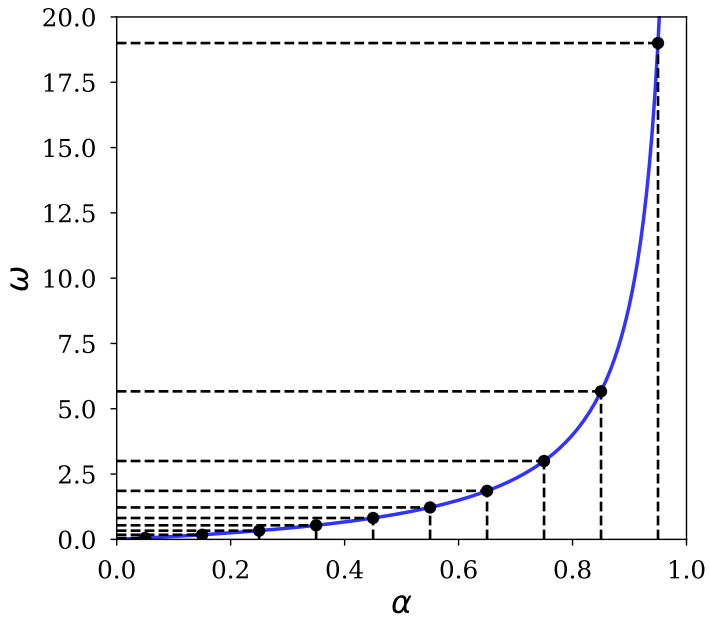
- Investigator 1 analyzes data using a model with parameter θ, with prior PDF $\pi_1(\theta) = C_1$
- Investigator 2 analyzes the *same* data using the *same* model, but parameterized in terms of $\phi = \Phi(\theta)$; assigns the prior PDF $\pi_2(\phi) = C_2$
- To an interval $d\phi$ corresponding to models with $\theta \in d\theta$, investigator 2 assigns probability

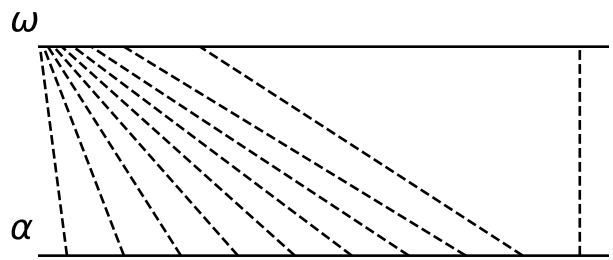
$$\pi_2(\phi)d\phi = C_2 d\phi = C_2 \Phi'(\theta) d\theta$$

This won't be constant wrt θ in general $\Rightarrow$ investigators adopting uniform priors with different parameterizations may disagree in inferences with the same model and data

Consider two parameterizations of a binomial sampling dist'n:

- Success probability, α: $p(n|\alpha) \propto \alpha^n(1-\alpha)^{N-1}$
- Odds, $\omega \equiv \alpha/(1-\alpha)$: $p(n|\alpha) \propto \left(\frac{\omega}{1+\omega}\right)^n \left(\frac{1}{1+\omega}\right)^{N-1}$





Uniform over α is inconsistent with uniform over ω!

Is there a "natural" way to pick intervals we should consider equally probable a priori?

Non-informative priors from transformation groups

The principle of indifference could not be justified in this case because the two versions of the problem have different symbolic structure (parameterizations) $\rightarrow$ there is no obvious counterpart to the *symmetry equations*

Sometimes a problem with continuous parameters *does* have a symmetry that identifies symmetry equations—particular transformations that make different parameterizations of the problem symbolically equivalent

Poisson rate (for discrete data)

$\mathcal{C}_1$: Estimate a Poisson rate, using a parameter r in counts/sec, from n counts observed in time δ sec, based on

$$p(n|r,\mathcal{C}_1) = \frac{(r\delta)^n}{n!}e^{-r\delta}, \qquad p(r|\mathcal{C}_1) = f(r)$$

$\mathcal{C}_2$: Estimate the same a Poisson rate, using a parameter R in counts/hr, from n counts observed in time Δ hr, based on

$$p(n|R,\mathcal{C}_2) = \frac{(R\Delta)^n}{n!}e^{-R\Delta}, \qquad p(R|\mathcal{C}_2) = g(R)$$

Since they are describing the same situation, R and r, and Δ and δ must be related; for $\alpha \equiv 1/3600$:

$$\Delta = \alpha\delta, \qquad R = r/\alpha$$

which ensures that $r\delta = R\Delta$, so they assign the same probabilities to n in equivalent situations

Repeating:

$$\Delta = \alpha\delta, \qquad R = r/\alpha$$

Since they are describing the same situation, the priors must be related via change-of-variables (transformation eq'ns):

$$g(R)\,\mathrm{d}R = f(r)\,\mathrm{d}r \quad \Rightarrow \quad g(R) = f(\alpha R) \times \alpha$$

This particular scale transformation makes the two formulations of the problem look equivalent. If we have no prior information distinguishing the formulations—we have no information assigning a particular time scale to the phenomenon—we should assign priors of the same form (symmetry eq'ns):

$$g(\cdot) = f(\cdot)$$

Repeating:

$$g(R) = f(\alpha R) \times \alpha; \qquad g(\cdot) = f(\cdot)$$

$$\rightarrow f(R) = \alpha f(\alpha R)$$

This holds for any (R, α). Substitute $R = 1$:

$$f(1) = \alpha f(\alpha)$$

$$f(\alpha) = \frac{C}{\alpha}$$

For Poisson rate inference, $\pi(r) \propto 1/r$ expresses ignorance of the time scale for the phenomenon

Note that this prior is *improper*—a generic feature of these types of priors when the parameter space is infinite

Location parameter (for continuous data)

For a PDF $p(x|\mu)$ for x with parameter μ, if it is of the form $p(x|\mu) = h(x - \mu)$, then μ is called a *location parameter*

$\mathcal{C}_1$: Estimate a location parameter, μ, using a sampling distribution for data x_i of the form

$$p(x|\mu, \mathcal{C}_1) = h(x - \mu);$$

assign a prior $P(\mu \in d\mu|\mathcal{C}_1) = f(\mu)d\mu$

$\mathcal{C}_2$: In a coordinate system shifted by Δ, use data $x'_i = x_i + \Delta$ to estimate $\mu' = \mu + \Delta$, using a sampling distribution of the same form,

$$p(x'|\mu', \mathcal{C}_2) = h(x' - \mu');$$

assign a prior $P(\mu' \in d\mu'|\mathcal{C}_2) = g(\mu')d\mu'$

Symmetry: Note that $h(x' - \mu') = h(x - \mu)$; the two problems both look the same (same $h(\cdot)$) and assign the same probability density to equivalent data

Provided there is no information in $\mathcal{C}_1$ or $\mathcal{C}_2$ identifying a special location, a formal rule should assign the same functions as priors:

$$f(u) = g(u)$$

Transformation: The shift in coordinates, $\mu' = \mu + \Delta$, implies that a consistent assignment of prior probabilities must obey

$$f(\mu)d\mu = f(\mu' - \Delta)d\mu' = g(\mu')d\mu'$$

Since symmetry implies $f = g$,

$$f(u - \Delta) = f(u)$$

Only a *constant PDF* $f(u) = C$ satisfies this functional eq'n

Scale parameter

For a PDF $p(x|\sigma)$ for x with parameter σ, if it is of the form $p(x|\sigma) = h(x/\sigma)/s$, then σ is called a *scale parameter*

$\mathcal{C}_1$: Estimate a scale parameter, σ, using a sampling distribution for data x_i of the form

$$p(x|\sigma, \mathcal{C}_1) = \frac{1}{\sigma} h(x/\sigma);$$

assign a prior $P(\sigma \in d\sigma|\mathcal{C}_1) = f(\sigma)d\sigma$

$\mathcal{C}_2$: In a coordinate system rescaled by s (e.g., changing units), use data $x'_i = sx_i$ to estimate $\sigma' = s\sigma$, using a sampling distribution of the same form,

$$p(x'|\sigma', \mathcal{C}_2) = \frac{1}{\sigma'} h(x'/\sigma');$$

assign a prior $P(\sigma' \in d\sigma'|\mathcal{C}_2) = g(\sigma')d\sigma'$

Symmetry: Note that $h(x'/\sigma') = h(x/\sigma)$; the two problems both look the same (same $h(\cdot)$) and assign the same probability density to equivalent data

Provided there is no information in $\mathcal{C}_1$ or $\mathcal{C}_2$ identifying a special scale, a formal rule should assign the same functions as priors:

$$f(u) = g(u)$$

Transformation: The shift in scale, $\sigma' = s\sigma$, implies that a consistent assignment of prior probabilities must obey

$$f(\sigma)d\sigma = f(\sigma'/s)d\sigma'/s = g(\sigma')d\sigma'$$

Since symmetry implies $f = g$,

$$f\left(\frac{u}{s}\right) = sf(u)$$

Only $f(u) = C/u$ satisfies this; this PDF is flat in $\log(u)$

Priors derived from the likelihood function

Few common problems beyond location/scale problems admit a transformation group argument $\rightarrow$ we need a more general approach to formal assignment of priors that express "ignorance" or are "noncommittal" in some sense

There is no universal consensus on how to do this (yet? ever?)

A common underlying idea: The same $\mathcal{C}$ appears in the prior, $p(\theta|\mathcal{C})$, and the sampling dist'n (likelihood), $p(D|\theta,\mathcal{C})$—the prior "knows" about the form of the likelihood function, although it doesn't know what data values will be plugged into it

Jeffreys priors: Use Fisher information to define a (parameter-dependent) scale defining a prior; parameterization invariant, but strange behavior in multivariate problems

Reference priors: Use information theory to define a prior that (asymptotically) has the least effect on the posterior; complicated algorithm; gives good frequentist behavior to Bayesian inferences