Harvard Department of Government 2017
Chapter 4, Bayesian Priors

JEFF GILL

Visiting Professor, Spring 2025

A Prior Discussion of Priors

- Developing Bayesian models *requires* specifying prior distributions for unknown parameters.

- Central to the Bayesian philosophy is the recognition that not only do the data, $\mathbf{X}$, possess a distribution, but so do unknown parameters, θ, by assumption.

- This is the key source of Bayesian/non-Bayesian conflict.

- Recall the role of the prior: $\pi(\theta|\mathbf{x}) \propto p(\theta)L(\theta|\mathbf{x})$.

Different Flavors of Bayesians

- Diaconis and Ylvisaker (1985) identify three distinct Bayesian philosophies regarding the selection of priors:
  - *classical*: the prior is a necessary inconvenience and one should specify a “flat” prior so as to interject the least amount of prior knowledge as possible.
  - *modern parametric*: specify priors possessing deliberate characteristics, such as conjugacy, and assign prior parameter values according to specific criteria unrelated to nonstatistical knowledge.
  - *subjective*: elicit prior distributions according to preexisting scientific knowledge in the substantive field.
- But these are not necessarily mutually exclusive definitions.

Bayesian Shrinkage

- The prior distribution works to move the posterior away from the likelihood and toward its own position.
- In cases where sharply defined priors are specified in the form of distributions with small variance, than Bayesian estimates will have lower variance then corresponding classical likelihood-based estimates.
- The greater the correlation between coefficients in a given model, the greater the extent of the "shrinkage" toward the prior mean.
- Hierarchal specifications (multilevels of priors) display more shrinkage due to correlations between parameters.

Bayesian Shrinkage (cont.)

- Consider the normal model in with prior mean m and variance s^2 for μ and σ_0 known and therefore substituted back in.

- Start with the posterior mean from the conjugate model:

$$\hat{\mu} = \left(\frac{m}{s^2} + \frac{n\bar{x}}{\sigma_0^2}\right) \Bigg/ \left(\frac{1}{s^2} + \frac{n}{\sigma_0^2}\right)$$

  and replace σ_0 with σ as if we knew its value.

- Then

$$\hat{\mu} = \left(\frac{m}{s^2} + \frac{n\bar{x}}{\sigma^2}\right) \Bigg/ \left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right) \qquad = \frac{m/s^2}{\frac{1}{s^2} + \frac{n}{\sigma^2}} + \frac{n\bar{x}/\sigma^2}{\frac{1}{s^2} + \frac{n}{\sigma^2}}$$

$$= \frac{m}{1 + \frac{ns^2}{\sigma^2}} + \frac{n\bar{x}}{\frac{\sigma^2}{s^2} + n} \qquad = \underbrace{\frac{\sigma^2}{\sigma^2 + ns^2}}_{S_f} m + \underbrace{\frac{ns^2}{\sigma^2 + ns^2}}_{1-S_f} \bar{x}$$

Bayesian Shrinkage (cont.)

- Here

$$S_f = \frac{\sigma^2}{(\sigma^2 + ns^2)}$$

is the shrinkage factor that is necessarily bounded by $[0{:}1]$.

- So the shrinkage factor gives the *proportional* distance that the posterior mean has shrunk back to the prior mean away from the classical maximum likelihood estimate $\bar{x}$.

- Question #1: how does a large data variance, s^2, effect shrinkage?

- Question #2: how does large data size, n, effect shrinkage?

Bayesian Shrinkage (cont.)

- The posterior variance in the normal-normal model can also be rewritten in similar fashion:

$$\hat{\sigma}^2 = \left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right)^{-1} = (S_f\sigma^2)^{1/2}\left((1-S_f)\frac{s^2}{n}\right)^{1/2} = S_f^{1/2}\sigma \quad \times \quad \left(\frac{1-S_f}{n}\right)^{1/2} s.$$

- This shows that the posterior variance is a product of the square root of the prior variance weighted by the shrinkage factor and the square root of the data variance weighted by the complement of the shrinkage factor.

- We see specifically how the shrinkage factor determines the compromise between prior uncertainty and data uncertainty.

Bayesian Shrinkage (cont.)

- Return to the Beta-Binomial conjugate setup from Day 1:

$$Y \sim \mathfrak{Bin}(n, p) \qquad\qquad p \sim \mathfrak{Beta}(A, B)$$

- So:

$$p|y \sim \mathfrak{Beta}(y + A, n - y + B)$$

- With:

$$\hat{p} = \frac{(y + A)}{(y + A) + (n - y + B)} = \left[\frac{n}{A + B + n}\right]\left(\frac{y}{n}\right) + \left[\frac{A + B}{A + B + n}\right]\left(\frac{A}{A + B}\right)$$

- So $\frac{A+B}{A+B+n}$ is the *shrinkage estimate* where the degree of shrinkage is determined by the magnitude of $A + B$ relative to n.

Conjugate Priors

- Usually attributed to Raiffa and Schlaifer (1961).
- The posterior distribution of the $\boldsymbol{\theta}$ vector might not have an analytically tractable form particularly in higher dimensions.
- Conjugacy is a joint property of the prior and the likelihood function that provides a posterior from the same distributional family as the prior.
- In other words, the mathematical form of the prior distribution "passes through" the data-conditioning phase and endures in the posterior: closure under sampling.

Another Example: Conjugacy in Exponential Specifications

- A very simple to way to model the time that something endures (wars, lifetimes, regimes, bull markets, marriages, etc.) is to use the exponential PDF. This has the form:

$$\mathcal{E}(X|\theta) = \theta \text{exp}[-\theta X], \quad 0 \leqslant X, 0 < \theta.$$

- Specify a gamma prior for θ in the exponential PDF:

$$p(\theta|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\theta^{\alpha-1}\text{exp}[-\beta\theta], \qquad \theta,\alpha,\beta > 0.$$

- We observe $x_1, x_2, \ldots, x_n \sim \text{ iid}\,\mathcal{E}(X|\theta)$ and produce the likelihood function:

$$L(\theta|\mathbf{x}) = \Pi_{i=1}^{n}\theta e^{-\theta x_i} = \theta^{n}\text{exp}\left[-\theta\sum x_i\right].$$

- Note that $\sum x_i$ is a sufficient statistic for θ.

Another Example: Conjugacy in Exponential Specifications (cont.)

- The posterior distribution is produced as follows:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto L(\theta|\mathbf{x})p(\theta|\alpha,\beta)\\ &= \theta^n \exp\left[-\theta\sum x_i\right]\frac{1}{\Gamma(\alpha)}\beta^\alpha\theta^{\alpha-1}\exp[-\beta\theta]\\ &\propto \theta^{(\alpha+n)-1}\exp\left[-\theta\left(\sum x_i+\beta\right)\right].\end{aligned}$$

- So this is another gamma distribution with parameters $\alpha+n$ and $\sum x_i+\beta$.

The Exponential Family Form of Conjugacy

- Form: $f(x|\theta) = \exp\big[t(x)u(\theta)\big]r(x)s(\theta)$, where r and t are real-valued functions of x that do not depend on θ, and s and u are real-valued functions of θ that do not depend on x, and $r(x) > 0, s(\theta) > 0 \ \forall x, \theta$.

- Because of the properties of logs and exponentiation, this form can always be rewritten as:

$$f(x|\theta) = \exp\Big[\ \underbrace{t(x)u(\theta)}_{\substack{\text{interaction}\\ \text{component}}} + \underbrace{\log(r(x)) + \log(s(\theta))}_{\text{additive component}}\ \Big],$$

- Common functional forms that can be expressed in exponential family form include: normal, binomial, gamma, Poisson, and negative binomial.

The Exponential Family Form of Conjugacy (cont.)

- Example: normal PDF.

$$
\begin{aligned}
f(y|\mu,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp\left[\Big(\underbrace{y\mu}_{y\theta} - \underbrace{\frac{\mu^2}{2}}_{b(\theta)}\Big)/\underbrace{\sigma^2}_{\phi} + \underbrace{\frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)}_{c(y,\phi)}\right].
\end{aligned}
$$

- This form produces useful quantities, for instance:

$$E[y] = \frac{\partial}{\partial\theta}b(\theta) = \frac{\partial}{\partial\mu}\frac{\mu^2}{2} = \mu.$$

The Exponential Family Form of Conjugacy (cont.)

- As a counter example the Weibull PDF (useful for modeling general failure times):

  $$f(x|\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}\text{exp}(-x^{\gamma}/\beta)$$

  for $x \geqslant 0, \gamma, \beta > 0$ *cannot* be put in exponential family form. The term $-\frac{1}{\beta}x^{\gamma}$ in the exponent disqualifies this PDF from the exponential family classification since it cannot be expressed in the additive or multiplicative form necessary.

- Recall that the structure of the exponential family form is preserved under sampling so the joint density function of iid random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ is simply:

  $$f(\mathbf{x}|\theta) = \exp\left[u(\theta)\sum_{i=1}^{n} t(x_i) + \sum_{i=1}^{n}\text{log}(r(x_i)) + n\text{log}(s(\theta))\right].$$

The Exponential Family Form of Conjugacy (cont.)

- Start with a corresponding conjugate prior form in generalized notation:

$$p(\theta|k,\gamma) = c(k,\gamma)\text{exp}[ku(\theta)\gamma + k\text{log}(s(\theta))].$$

- Now calculate the posterior from multiplying the likelihood function and the prior:

$$\begin{aligned}\pi(\theta|\mathbf{x},k,\gamma) &\propto f(\mathbf{x}|\theta)p(\theta|k,\gamma)\\ &\propto \exp\left[u(\theta)\left(\sum_{i=1}^{n} t(x_i) + k\gamma\right) + (n+k)\text{log}(s(\theta))\right]\\ &= \exp\left[u(\theta)(n+k)\left(\frac{\sum_{i=1}^{n} t(x_i) + k\gamma}{n+k}\right) + (n+k)\text{log}(s(\theta))\right].\end{aligned}$$

- So the posterior distribution is an exponential family form like the prior with parameters: $k' = (n+k), \gamma' = \frac{\sum_{i=1}^{n} t(x_i)+k\gamma}{n+k}$, giving a criteria for exponential family form in the Bayesian context.

The Exponential Family Form of Conjugacy (cont.)

Likelihood Form	**Conjugate Prior Distribution**	**Hyperparameters**
Bernoulli	Beta	$\alpha > 0, \beta > 0$
Binomial	Beta	$\alpha > 0, \beta > 0$
Multinomial	Dirichlet	$\theta_j > 0, \Sigma\theta_j = 1$
Negative Binomial	Beta	$\alpha > 0, \beta > 0$
Poisson	Gamma	$\alpha > 0, \beta > 0$
Exponential	Gamma	$\alpha > 0, \beta > 0$
Gamma (incl. χ^2)	Gamma	$\alpha > 0, \beta > 0$
Normal for μ	Normal	$\mu \in \mathbb{R}, \sigma^2 > 0$
Normal for σ^2	Inverse Gamma	$\alpha > 0, \beta > 0$
Pareto for α	Gamma	$\alpha > 0, \beta > 0$
Pareto for β	Pareto	$\alpha > 0, \beta > 0$
Uniform	Pareto	$\alpha > 0, \beta > 0$

Uninformative Priors

- Recall that a uninformative prior is one in which little new explanatory power about the unknown parameter is provided by intention.

- Uninformative priors are very useful from the perspective of traditional Bayesianism that sought to mitigate frequentist criticisms of intentional subjectivity.

- Fisher (1930, p. 531) was characteristically negative on the subject: "…how are we to avoid the staggering falsity of saying that however extensive our knowledge of the values of x may be, yet we know nothing and can know nothing about the values of θ?"

Uniform Priors

- An obvious and easy choice for the uninformative prior is the uniform distribution, but there are concerns.

- Uniform priors are particularly easy to specify in the case of a parameter with bounded support.

- *Proper* uniform priors can be specified for parameters defined over unbounded space if we are willing to impose prior restrictions.

- Thus if it reasonable to restrict the range of values for a variance parameter in a normal model, instead of specifying it over $[0\!:\!\infty]$, we restrict it to $[0\!:\!\nu]$ and can now articulate it as $p(\sigma) = 1/\nu,\ 0 \leqslant \theta \leqslant \nu$.

Uniform Priors (cont.)

- *Improper* uniform priors that do not possess bounded integrals and surprisingly, these can result in fully proper posteriors under some circumstances (although this is far from guaranteed).

- Specifically:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &= \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}} \\ &= \frac{cg(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})}{c\int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}} \\ &= \frac{g(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})}{\int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}} \end{aligned}$$

- Consider the common case of a uninformative uniform prior for the mean of a normal distribution. It would necessarily have uniform mass over the interval: $p(\theta) = c,\ [-\infty \geqslant \theta \geqslant \infty]$. Therefore to give *any* nonzero probability to values on this support, $p(\theta) = \epsilon > 0$, would lead to a prior with infinite density: $\int_{-\infty}^{\infty} p(\theta)d\theta = \infty$.

Problems with Uniform Priors

- The uniform prior is not invariant under transformation: simple transformations of the uniform prior produce a reexpression that is not uniform and loses whatever sense of uninformedness that the equiprobability characteristic of the uniform gives.

- For example, suppose again that we are interested in developing a uninformative prior for a normal model variance term and specify the improper uniform prior: $p(\sigma) = c,\ 0 \leqslant \sigma < \infty$.

- A simple transformation that provides a parameter space over the entire real line is given by: $\tau = \log(\sigma)$, and the new PDF is given by applying the transformation with Jacobian ($J = |\frac{\partial}{\partial\tau}g^{-1}(\tau)|$):

$$\tau = g(\sigma) = \log(\sigma) \longrightarrow g^{-1}(\tau) = \sigma = e^{\tau}$$

$$p(\tau) = p(g^{-1}(\tau))\left|\frac{\partial}{\partial\tau}g^{-1}(\tau)\right| = (c)\left|\frac{\partial}{\partial\tau}e^{\tau}\right| \propto e^{\tau}.$$

This resulting prior clearly violates even the vaguest sense of uninformed-ness and makes a strong statement about values that are *a priori* more likely than others.

Problems with Uniform Priors (cont.)

- Consider a proper uniform prior on a Bernoulli probability parameter: $f(p) = 1,\ 0 \leqslant p \leqslant 1$. If we change from the probability metric to the odds ratio metric (fairly common), then we impose the transformation: $q = \frac{p}{1-p}$ and the new distribution is given by:

$$q = g(p) = \frac{p}{1-p} \longrightarrow g^{-1}(q) = \frac{q}{1+q}$$

$$f(q) = f(g^{-1}(q)) \left| \frac{\partial}{\partial q} g^{-1}(q) \right| = (1) \left| \frac{\partial}{\partial q} \frac{q}{1+q} \right| = (1+q)^{-2}.$$

- Answer to frequent question:

$$f(g^{-1}(q)) = f\left(\frac{q}{1+q}\right) = f\left(\frac{p/(1-p)}{(1+p/(1-p))}\right) = f\left(\frac{p}{(1-p)+p}\right) = f(p) = 1$$

- Also, uniform priors have an inherent bias against the endpoints of the specified interval, and therefore do not necessarily provide the coverage that the researcher desires (particularly if these endpoints are of theoretical importance as might be the case for the $\mathcal{U}(0,1)$ specification).

Jeffreys Prior

- Jeffreys (1961, p. 181) addresses the problems associated with uniform priors by suggesting a prior that is invariant under transformation:

- Specifically:

$$p_{\text{Jeffreys}}(\alpha) = p_{\text{Jeffreys}}(\alpha') \left| \frac{\partial \alpha'}{\partial \alpha} \right|.$$

- *Jeffreys' Rule* states that any prior that is proportional to the Jeffreys prior is uninformative in the sense that it interjects as little subjective information into the posterior as possible.

- The most useful aspect of the Jeffreys prior is that it comes from a very mechanical process, but almost always produces a uninformative form with the invariance property:

- While the Jeffreys prior is straightforward in one dimension, unfortunately it can be quite difficult in multiparameter models.

Jeffreys Prior

- The Jeffreys prior for a single parameter, θ, is created by:

$$p(\theta) = \left[-E_{\mathbf{X}|\theta} \left(\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x}|\theta) \right) \right]^{\frac{1}{2}}.$$

  This is the square root of the determinant of the familiar negative expected Fisher information matrix.

- Note that the expectation here is taken over $f(\mathbf{X}|\theta)$.

- The expression given by Jeffreys prior is given in the single-dimension case, for a parameter vector, $\boldsymbol{\theta}$, Jeffreys prior is:

$$p(\theta) = \left[E_{\mathbf{X}|\theta} \left(\left[\frac{\partial}{\partial\theta} \log f(\mathbf{x}|\boldsymbol{\theta}) \right]' \left[\frac{\partial}{\partial\theta} \log f(\mathbf{x}|\boldsymbol{\theta}) \right] \right) \right]^{1/2}.$$

Jeffreys Prior (cont.)

- **Example: Bernoulli Trials and Jeffreys Prior**

  - Consider a repeated Bernoulli trial with x successes out of n attempts in which we are interested in obtaining a posterior distribution for the unknown probability of success p.

  - The binomial PMF, for the sum of the trials, is given by:

$$\mathfrak{Bin}(x|n,p) = \binom{n}{p} p^{x}(1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \quad 0 \geqslant p \geqslant 1,$$

  - and the log likelihood is given by:

$$\ell(p|n,x) = \log\binom{n}{x} + x\log(p) + (n-x)\log(1-p).$$

Jeffreys Prior (cont.)

▻ The first and second derivatives are given by:

$$\begin{aligned}\frac{\partial}{\partial p}\ell(p|n,x) &= \frac{x}{p} + \frac{n-x}{1-p}(-1)\\ &= xp^{-1} - (n-x)(1-p)^{-1},\\ \frac{\partial^2}{\partial p^2}\ell(p|n,x) &= -xp^{-2} - (n-x)(1-p)^{-2}(-1)(-1)\\ &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.\end{aligned}$$

▻ Since $E[x] = np$, the last stage is trivial:

$$J = \left(E[-\frac{\partial^2}{\partial p^2}\ell(p|n,x)]\right)^{\frac{1}{2}} = \left(\frac{np}{p^2} + \frac{n-np}{(1-p)^2}\right)^{\frac{1}{2}} = \left(\frac{n}{p(1-p)}\right)^{\frac{1}{2}},$$

▻ which suggests the prior $p(p) = p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}$, a $\mathfrak{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$ distribution.

Jeffreys Prior for a Poisson Model

- Now we have $f(x|\lambda) = e^{-\lambda}\lambda^x/x!$, so the log-likelihood function is:

$$\ell(\lambda|\mathbf{x}) = -n\lambda + \left(\sum \mathbf{x}_i\right) \log\lambda - \log\sum(\mathbf{x}_i)$$

- Taking derivatives:

$$\frac{\partial}{\partial\lambda}\ell(\lambda|\mathbf{x}) = -n + \lambda^{-1}\sum \mathbf{x}_i$$

$$\frac{\partial^2}{\partial^2\lambda}\ell(\lambda|\mathbf{x}) = -\lambda^{-2}\sum \mathbf{x}_i$$

- Now take the expectation over $\mathbf{x}$:

$$E_{\mathbf{x}|\lambda}\left(-\frac{\partial^2}{\partial^2\lambda}\ell(\lambda|\mathbf{x})\right) = E_{\mathbf{x}|\lambda}\left(\lambda^{-2}\sum \mathbf{x}_i)\right) = (n\lambda)\lambda^{-2}\propto \lambda^{-1}$$

- Finally, the square root produces $p_{Jeffreys}(\lambda) = \lambda^{-\frac{1}{2}}$

Jeffreys Prior for a "Stylized Normal" Model

- Suppose we have $x \sim \mathcal{N}(\mu, 1)$, so the likelihood function is:

$$L(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\sum \mathbf{x}_i^2 - 2n\bar{x}\mu + \mu^2\right)\right]$$

- The log-likelihood is therefore:

$$\ell(\mu|\mathbf{x}) \propto -\frac{1}{2}\sum \mathbf{x}_i^2 + n\bar{x}\mu - \frac{1}{2}\mu^2$$

- Making the first and second derivatives very easy:

$$\frac{\partial}{\partial\mu}\ell(\mu|\mathbf{x}) = n\bar{\mathbf{x}} - \mu \qquad\qquad \frac{\partial^2}{\partial^2\mu}\ell(\mu|\mathbf{x}) = -1$$

- So the square root of the negative of this is trivial and the Jeffreys prior is simply $p_{Jeffreys}(\mu) = 1$ over $-\infty < \mu < \infty$.

Reference Priors

- A reference prior is a, not necessarily flat, prior distribution such that for the given problem (only), the likelihood is data translated (imposed on the posterior).

- The distinction between a reference prior and a uninformative prior is murky and author-dependent.

- Also, a number of authors refer to reference priors as "automatic priors."

- Box and Taio (p. 23) define a reference prior as "a prior which it is convenient to use as a standard" and is "dominated by the likelihood." Thus it need not be uninformative.

- Reference priors can be enormously helpful in dealing with nuisance parameters

- Kass and Wasserman (1995) identify two distinct interpretations of reference priors: as an expression of ignorance or as a socially agreed upon standard (model specific) alternative to subjective priors, and they proceed to identify a litany of associated problems with the use of reference priors.

Reference Priors

- The original idea for the modern definition of a reference prior is from Bernardo (1979), who introduces the useful notion that a difficult parameter vector (perhaps difficult in the sense that Jeffreys prior does not work well), can be segmented into two components, parameters of high interest and nuisance parameters.

- A reference prior for the parameters of high interest is found by maximizing the distance between the chosen prior and the posterior according to some criteria like the Kullback-Leibler distance.

- This could be called a *dominant likelihood prior*, since it is a prior that is dominated by the likelihood function over the region of interest.

- *Example*: a very diffuse normal distribution centered somewhere near the expected mode of the posterior.

- *Example*: The Zellner-Siow (1980) prior is similar but more dispersed: a Cauchy distribution restricted to the range of interest.

Maximum Entropy Priors

- Jaynes introduces the idea of an *entropy prior* to describe relative levels of uncertainty about the distribution of prior parameters.
- One advantage to the entropy approach is that within the same framework, uncertainty ranging from that provided by the uniform prior to that provided by absolute certainty given by a degenerate (single point) distribution, can be modeled in the same way.
- Unfortunately, however, entropy priors are not invariant to reparameterization and therefore have somewhat limited applicability.

Maximum Entropy Priors (cont.)

- The core idea of entropy is the quantification of uncertainty of a transmission or observation (Shannon 1948), and this can be interpreted as uncertainty in a PDF or PMF (Rosenkranz 1977).

- Assume that we are interested in a discrete parameter θ with entropy of θ for a given parametric form, $p(\theta)$, as:

$$H(\theta) = -\sum_{\Theta} p(\theta_i) \log[p(\theta_i)],$$

  where the sum is taken over the categories of θ.

Maximum Entropy Priors (cont.)

- In the case of discrete distributions, we have a wide selection of parametric forms for $p(\theta)$, but consider two very different varieties for k possible values in the sample space: $\Theta = [\theta_1, \theta_2, \ldots, \theta_k]$.

- if we assign each outcome a uniform prior probability of occurrence: $1/k$, then the entropy of this prior is at its maximum:

$$H(\theta) = -\sum_{\Theta} \frac{1}{k} \log \left[\frac{1}{k} \right] = \log[k].$$

- Prior uncertainty increases logarithmically with increases in the number of discrete alternatives.

- So what happens with $p(\theta_i) = 1$, and $p(\theta_{\neg i}) = 0$?

$$H(\theta) = -\sum_{\Theta[-i]} 0 \log[0] + 1 \log[1] = 0,$$

where this calculation requires the assumption that $0\log[0] = 0$.

Maximum Entropy Priors (cont.)

- Suppose that we could stipulate the first two moments as constraints:

$$E[\theta] = \sum_{\Theta} p(\theta)\theta = \mu_1$$
$$E[\theta^2] = \sum_{\Theta} p(\theta)\theta^2 = \mu_2.$$

Adding the further constraint that $p(\theta)$ is proper, gives the prior density:

$$\tilde{p}(\theta_i) = \frac{\exp\left[\lambda_1\theta_i + \lambda_2(\theta_i - \mu_1)^2\right]}{\sum_j \exp\left[\lambda_1\theta_j + \lambda_2(\theta_j - \mu_1)^2\right]},$$

where the constants, λ_1 and λ_2, are determined from the constraints.

Maximum Entropy Priors (cont.)

- The continuous case
$$H(\theta) = -\int_{\Theta} p(\theta)\log[p(\theta)]d\theta,$$
leads to different answers depending on alternative definitions of the underlying reference measures.

- Essentially this just means that mathematical requirements for getting a usable prior are more difficult to obtain and in some circumstances are actually impossible.

- For the M moment estimation case, the constraining statement is now:
$$E[g_m(\theta)] = \int_{\Theta} p(\theta)g_m(\theta)d\theta = \mu_m, \qquad m = 1, \ldots, M.$$

- *Example*: the support of θ is $[0\!:\!\infty]$ and we specify the constraint that the expected value is equal to some constant: $E[\theta] = c$. If we further specify that the prior have the least informative possible exponential PDF form, then the continuous form of the entropy constraint specifies the prior: $f(\theta|c) = c\exp(-c\theta)$.

Power Priors

- An informed prior that explicitly uses data from previous studies.
- The idea is to weight data from earlier work as input for the prior used in the current model.
- Define $\mathbf{x}_0$ as this older data and $\mathbf{x}$ as the current data.
- Our interest centers on the unknown parameter θ, which is studied in both periods.
- Specify a regular prior for θ, $p(\theta)$ that would have been used un-modified if the previous data were not included.
- This can be a diffuse prior if desired, although it will become informed through this process.

Power Priors

- An elementary power prior is created by updating the regular prior with a likelihood function from the previous data, which is scaled by a scalar $a_0 \in [0{:}1]$:
$$p(\theta|\mathbf{x}_0, a_0) \propto p(\theta)[L(\theta|\mathbf{x}_0)]^{a_0}.$$

- This is still a *prior* form and the regular process follows wherein the posterior is obtained by conditioning this distribution on the data through the likelihood function based on the current data:
$$\pi(\theta|\mathbf{x}, \mathbf{x}_0, a_0) \propto p(\theta|\mathbf{x}_0, a_0) L(\theta|\mathbf{x}).$$

- a_0 scales our confidence in the similarity or applicability of the previous data for current inferences (lower means less confidence).

Power Priors

- To reduce the influence of a single choice for a_0, we specify a mixture of these priors using a specified distribution for this parameter, $p(a_0|.)$:

$$\begin{aligned} p(\theta|\mathbf{x}_0) &= \int_0^1 p(\theta|\mathbf{x}_0, a_0)p(a_0|.)da_0 \\ &= \int_0^1 p(\theta)[L(\theta|\mathbf{x}_0)]^{a_0}p(a_0|.)da_0. \end{aligned}$$

- The mixture specification has the effect of inducing heavier tails in the marginal distribution of θ and thus represents a more conservative choice of prior.

- A convenient choice for $p(a_0|.)$ is the beta distribution.

Spike and Slab Priors for Linear Models

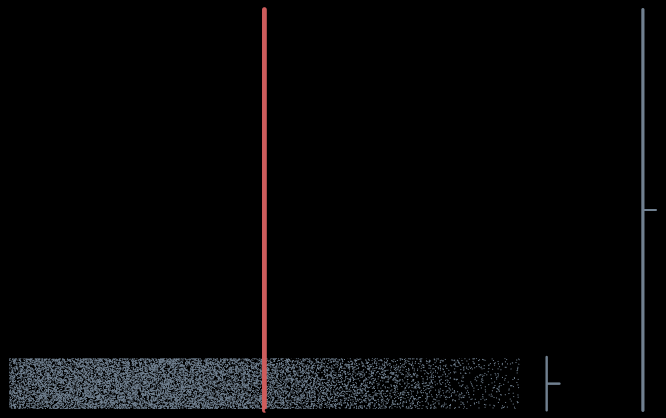
- Mitchell and Beauchamp (1988) introduce a species of priors designed to facilitate explanatory variable selection in linear regression.
- The basic idea is to specify a reasonably skeptical prior distribution by stipulating density spike at zero surrounded symmetrically by a flat slab with specified boundaries.
- Thus the prior makes a pretty strong claim about zero effect for some regression parameter, but admits the possibility of non-zero effects.

Spike and Slab Priors for Linear Models

- Suppose we are seeking a prior for the jth regression coefficient (from $j \in 1, \ldots, k$) in a standard linear model: $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \epsilon$, with all of the usual Gauss-Markov assumptions, $k < n$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- The prior is then:

$$\begin{aligned} &\text{spike:} \quad p(\beta_j = 0) = h_{0j} \\ &\text{slab:} \quad p(\beta_j \in [-f_j : f_j], \beta_j \neq 0) = 2h_{1j}f_j, \quad -f_j < \beta_j < f_j, \end{aligned}$$

where $h_{0j} + 2h_{1j}f_j = 1$.



Spike and Slab Priors for Linear Models

- So obviously h_{0j} is the height of the spike and h_{1j} is the height of the slab, and varying degrees skepticism about the veracity of the effect of $\mathbf{x}_j$ can be modeled by altering the ratio:

$$\gamma_j = \frac{h_{0j}}{h_{1j}} = 2f_j \frac{h_{0j}}{1 - h_{0j}}.$$

- For terms that should be included in the model with probability one, just set $h_{0j} = 0$.
- Since all prior density not accounted for by the spike falls to the slab, the prior for the mth model, A_m, is just the product given by:

$$p(A_m) = \prod_{j \in A_m} (1 - h_{0j}) \prod_{j \notin A_m} (h_{0j}),$$

which is the product of the slab density for those coefficients in model A_m, times the product of the spike density for those coefficients not in A_m.

Spike and Slab Priors for Linear Models

- If we put a prior on σ such that $\log(\sigma)$ is uniform between $-\log(\sigma_0)$ and $\log(\sigma_0)$, for some large value of σ_0.

- The resulting posterior for model A_m with k_m number of coefficients (including the constant) in the matrix $\mathbf{X}_m$:

$$\pi(A_m|\mathbf{X},\mathbf{y}) = \pi^{k_m/2}\Gamma\left(\frac{n-k_m}{2}\right)\left|\mathbf{X}_m'\mathbf{X}_m\right|^{-\frac{1}{2}}(S_m^2)^{-(n-k_m)/2}\prod_{j\notin A_m}\gamma_j$$

  where $S_m^2 = (\mathbf{y}-\mathbf{X}_m\hat{\boldsymbol{\beta}}_m)'(\mathbf{y}-\mathbf{X}_m\hat{\boldsymbol{\beta}}_m)$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}$.

- On each MCMC iteration the Bayesian linear estimation of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X}+\boldsymbol{\beta}_0^{-1})^{-1}\mathbf{X}'\mathbf{y}$, where $\boldsymbol{\beta}_0$ is the variance matrix from a multivariate normal prior $N_K(\mathbf{0},\boldsymbol{\beta}_0)$ for $\boldsymbol{\beta}$.

- Thus we can see how the data updates our model priors across all of the proposed specifications.

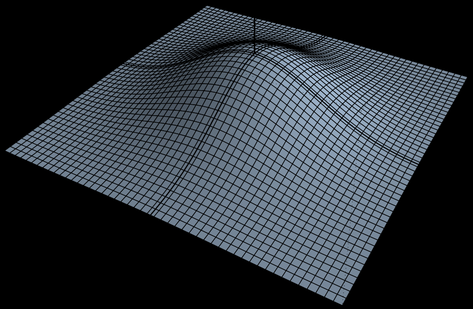
Spike and Slab Priors for Linear Models

- This setup also facilitates coefficient quality assessment as well since we can average coefficient probabilities across model space.

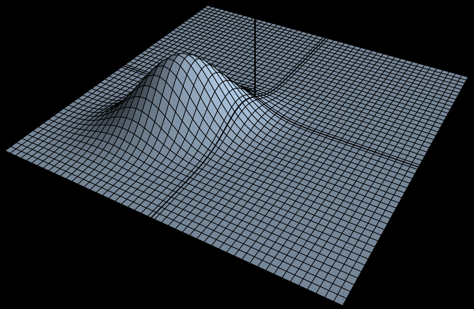
- So the joint posterior distribution of all k of the coefficients is given by:

$$\pi(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \sum_m \pi(A_m|\mathbf{X},\mathbf{y})\pi(\boldsymbol{\beta}|\boldsymbol{\alpha}_m,\mathbf{X},\mathbf{y})$$

‘ where $\pi(\boldsymbol{\beta}|\boldsymbol{\alpha}_m,\mathbf{X},\mathbf{y})$ is multivariate t-distribution centered at $\hat{\boldsymbol{\beta}}$.



Bivariate Spike and Slab Prior



Hypothetical Resulting Posterior

Types of Elicited Priors

- **Clinical Priors:** elicited from substantive experts who are taking part in the research project.
- **Skeptical Priors:** built with the assumption that the hypothesized effect does not actually exist and are typically operationalized through a probability function (PMF or PDF) with mean zero.
- **Enthusiastic Priors:** the opposite of the sceptical prior, built around the positions of partisan experts or advocates and generally assuming the existence of the hypothesized effect.
- **Reference Priors:** produced from expert opinion as a way to express informational uncertainty, but they are somewhat misguided in this context since the purpose of elicitation is to glean information that can be described formally. Somewhat misguided.

Particular Process Phases

- **Deterministic Phase.** The problem is codified and operationalized into specified variables and definitions that produce individual questions.
- **Probabilistic Phase.** Experts are interviewed to assign probabilities or values associated with hypothetical or historical events.
- **Informational Phase.** The assessor assigned probabilities or values are tested for inconsistencies and response completeness is verified.

Deterministic Phase

- Focus here on:
  - specifying explanatory variables in the model and possibly the assumed prior parametric form for their associated coefficients,
  - determining the relevant data collection processes,
  - selecting the number of experts to query,
  - and planning how to evaluate the reliability of their contributions.
- Some of this work is difficult: experts might need to be trained before elicitation, variable selection can be influenced by the difficulty of elicitation, and cost estimates may be uncertain.

Probabilistic Phase

- Assessors can be asked fixed value questions with probability responses ("P-methods"), fixed probability questions with value responses ("V-methods"), or, questions to be answered on probability and value scales simultaneously ("PV-methods").
- P-methods determine interesting levels of explanatory variables (or perhaps a range of values in the case of interval measurement) and require the assessor to provide the probability of occurrence for levels of the outcome variable.
- V-methods ask the more challenging question of determining explanatory variable levels associated with some given probability value.
- PV-methods are even more demanding because they require that the assessor pick cumulative distribution points and associated levels as a pair without prompting on either.
- Asking open-ended questions and coding the responses can be a very informative alternative method to these approaches, although unstructured answers are more difficult to translate onto a probability metric.

Informational Phase

- Includes testing elicitation responses for internal consistency, calibrating these responses with reliable references, and sometimes weighting the assessors relative to each other.

- Consistency is an important issue since assessors can differ in their familiarity with the details of the studied effect.

- Not surprisingly, inexperienced assessors tend to give more internal inconsistencies, especially with continuous rather than discrete choices.

- Consistency specifically means that an individual’s (or a group’s) answers do not self-contradict.

- One study found participants predicted the occurrence of more words ending in “-*ing*” rather than “-_*n*_” even though the former is included in the latter (Tversky and Kahneman 1983).

- Relatedly, we are also concerned with coherence: producing priors that do not violate the standard probability axioms.

Constructing Elicited Priors

- The key challenge is translating verbal or written opinions into specific probability statements.
- This process ranges from informal assignments to detailed elicitation plans and even regression analysis across multiple experts
- Commonly this process is automated so that assessors fill-in numerical values or select menu options on a terminal, see the subsequent posterior effects, and then make adjustments to their original input.
- A common strategy is to query experts about outcome variable quantiles for given (hypothetical) levels of specific explanatory variables.

Constructing Elicited Priors (cont.)

- For example, in one well-known study an emergency room physician is asked to estimate the survival probabilities of hypothetical patients with specified injury types, injury severity scores, trauma scores, and ages, set by the researchers (Bedrick, Christensen, and Johnson 1997).

- This provides survival probabilities at various explanatory variable levels, so given an assumed distributional form for each of the priors, it is now possible to solve "backwards" for the needed prior parameters.

- Assessors are first asked to give mean outcome variable responses or probability quantiles corresponding to a set of differing explanatory variable levels: $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_m$ (called *design points*).

- The elicited outcome variable values now correspond to the "spread" of the full *design matrix* created by stacking these $\tilde{\mathbf{X}}_i$ vectors as if they were empirical cases. Prior parameter determination can be carried-out via some accepted parametric model (Garthwaite and Dickey 1988), or conversely by scatterplot smoothing.

Simple Elicitation Strategy

- The (V-method procedure) question asked is what would be an expected low value in the form of a 0.25 quantile (x_1) and an expected high value in the form of a 0.75 quantile (x_2).

- These values then help specify a normal distribution for this event (other prior forms can be specified in similar fashion).

- The two supplied quantile values, x_1 and x_2 corresponding to $z_1 = 0.25$ and $z_2 = 0.75$, exactly specify the shape of a normal PDF since there are two unknown parameters and two equations, according to:

$$z_1 = \frac{x_1 - \alpha}{\beta} \qquad\qquad z_2 = \frac{x_2 - \alpha}{\beta},$$

where α and β are the mean and standard deviation parameters of the normal form:

$$f(x|\alpha, \beta) = (2\pi\beta^2)^{-\frac{1}{2}} \exp\left[-(x-\alpha)^2/2\beta^2\right].$$

Therefore when we solve for α and β we have a fully defined prior distribution from the elicitation.

Simple Elicitation Strategy

- One expert is often insufficient so we now query experts $1, 2, \ldots, J$, producing an over-specified series of equations since there are $J \times 2$ equations and only two unknowns ($J = 10$ for example).

- We assume for now that these experts are exchangeable in the sense that they provide equal quality information.

- Given the cost of interviewing, we are much more likely to ask each expert for more than just two quantiles, and it is always helpful to have more assessed points if these are deemed reliable.

- So each assessor is asked to give five quantile values at $m = [0.01, 0.25, 0.5, 0.75, 0.99]$ corresponding to standard normal points z_m. Now the normal form can be re-expressed for the quantile level m given by assessor j:

$$x_{jm} = \alpha + \beta z_{jm}.$$

Simple Elicitation Using Linear Regression

- Therefore the total amount of expert-elicited information constitutes the following over-specification ($J \times 5$ equations and 2 unknowns) of a normal distribution:

$$\begin{array}{lllll}
x_{11} = \alpha + \beta z_{11} & x_{21} = \alpha + \beta z_{21} & \ldots & x_{(J-1)1} = \alpha + \beta z_{(J-1)1} & x_{J1} = \alpha + \beta z_{J1} \\
x_{12} = \alpha + \beta z_{12} & x_{22} = \alpha + \beta z_{22} & \ldots & x_{(J-1)2} = \alpha + \beta z_{(J-1)2} & x_{J2} = \alpha + \beta z_{J2} \\
x_{13} = \alpha + \beta z_{13} & x_{23} = \alpha + \beta z_{23} & \ldots & x_{(J-1)3} = \alpha + \beta z_{(J-1)3} & x_{J3} = \alpha + \beta z_{J3} \\
x_{14} = \alpha + \beta z_{14} & x_{24} = \alpha + \beta z_{24} & \ldots & x_{(J-1)4} = \alpha + \beta z_{(J-1)4} & x_{J4} = \alpha + \beta z_{J4} \\
x_{15} = \alpha + \beta z_{15} & x_{25} = \alpha + \beta z_{25} & \ldots & x_{(J-1)5} = \alpha + \beta z_{(J-1)5} & x_{J5} = \alpha + \beta z_{J5}
\end{array}$$

- The solution suggested by this setup is to run a simple linear regression with α as the outcome variable and β as the explanatory variable.

Simple Elicitation Using Linear Regression (cont.)

- Suppose we are interested in eliciting a prior distribution for expected campaign contributions received by major-party candidates in the upcoming election for contested U.S. Senate seats as part of a larger Bayesian model.

- Elicit opinions from election experts on what contribution levels they might expect to see. This is a good example of the value of elicitation since many people expert in campaign contributions are practitioners or analysts outside of academic political science.

- Eight experts are queried for three quantiles at levels $m = [0.1, 0.5, 0.9]$, and they provide the following values reflecting the national range of expected total intake by Senate candidates (in thousands):

$x_{11} = 400$	$x_{21} = 150$	$x_{31} = 300$	$x_{41} = 250$	$x_{51} = 450$	$x_{61} = 100$	$x_{71} = 500$	$x_{81} = 300$
$x_{12} = 2500$	$x_{22} = 1000$	$x_{32} = 900$	$x_{42} = 1200$	$x_{52} = 1800$	$x_{62} = 1000$	$x_{72} = 2100$	$x_{82} = 1200$
$x_{13} = 4000$	$x_{23} = 2500$	$x_{33} = 1800$	$x_{43} = 2000$	$x_{53} = 3000$	$x_{63} = 2500$	$x_{73} = 4200$	$x_{83} = 2000$

(recall, for instance, that x_{83} indicates expert eight's third quantile).

Simple Elicitation Using Linear Regression (cont.)

- Since none of the experts have supplied quantile values out of logical order, these results are *consistent* (more on this shortly).

- Using these "data" we regress x on z to obtain the intercept and slope values: $\alpha = 1506$, $\beta = 953$.

Variance Components Elicitation

- A pervasive problem with direct quantile elicitation procedures is that assessors tend to misjudge the occurrence of unusual values because it is more difficult to visualize and estimate tail behavior than to estimate means or medians.

- When assessors are asked to estimate spread by providing high probability coverage intervals such as at 99%, then there is an inclination to perceive this as near-certainty coverage and overstate the bounds.

- Conversely, in other settings people tend to think of rare events in the tails of distributions as more likely than they really are.

- O'Hagan (1998) suggests improving elicited estimates of spread by separately requiring assessors to consider two types of uncertainty: uncertainty about an estimate relative to an assumed known summary statistic, and the uncertainty of this summary.

Variance Components Elicitation (cont.)

- First the assessor gives a (modal) point estimate for the explanatory variable coefficient: τ.

- Then they are asked: "given your estimate of τ, what is the middle 50% probability interval around τ?"

- This V-method specifies a density estimate centered at the assessors modal point, and if the form of the distribution is assumed or known, then the exact value for the variance can be backed out mathematically.

- O'Hagan (1998) prefers asking for the middle 66% of the density (he calls this the "two-to-one interval" since the middle coverage is twice that of the combined tails).

Variance Components Elicitation (cont.)

- If a normal prior is assumed then this interval quickly yields a value for the standard deviation since it covers approximately two of them (it should actually be multiplied by $\frac{68}{66}$ but analysts typically do not worry about the difference).

- Once the assessor gives this interval, the researcher calculates the implied variance and shows the assessor credible intervals at familiar $(1-\alpha)$-levels, such as 50% or 99%, so that the assessor can see the general implications of their assigned spread. If these are deemed to be too large or too small, then the process is repeated.

- Suppose that the purpose of elicitation is to obtain prior distributions for unknown values τ_i across n cases, with unknown total $T = \sum_{i=1}^{n} \tau_i$.

Variance Components Elicitation (cont.)

- The assessor first provides point estimates for each case: $x_1, x_1, \ldots, x_n$, so that the estimated total is given by $x_T = \sum_{i=1}^{n} x_i$.

- The individual deviance of the ith estimate from its true value can be rewritten algebraically according to:

  $$\tau_i - x_i = \left(\tau_i - \frac{x_i}{x_T}T\right) + \frac{x_i}{x_T}\left(T - x_T\right).$$

  The first quantity on the right-hand-side is the deviance of τ_i from an estimate that would be provided if we knew T for a fact:

  $$E[\tau_i|T] = \frac{x_i}{x_T}T,$$

  which can be considered as between-case deviance. The second quantity on the right-hand-side is the weighted deviation of T, i.e. uncertainty about the true total.

Variance Components Elicitation (cont.)

- This expected value form helps us obtain the variance of τ_i:

$$\begin{aligned}
\text{Var}(\tau_i) &= E\left[\text{Var}(\tau_i|T)\right] + \text{Var}\left(E[\tau_i|T]\right) \\
&= E\left[\tau_i - \frac{x_i}{x_T}T\right]^2 + \left(E\left[E[\tau_i|T]^2\right] - \left(E\left[E[\tau_i|T]\right]\right)^2\right) \\
&= E\left[\tau_i - \frac{x_i}{x_T}T\right]^2 + E\left[\left(\frac{x_i}{x_T}T\right)^2\right] - \left[E\left(\frac{x_i}{x_T}T\right)\right]^2 \\
&= E\left[\tau_i - \frac{x_i}{x_T}T\right]^2 + \left(\frac{x_i}{x_T}\right)^2 \text{Var}(T),
\end{aligned}$$

  which shows the general form of the two variance components.

- Also, a more natural form for elicitation is achieved by dividing both sides of this equation by x_i^2:

$$\text{Var}\left(\frac{\tau_i}{x_i}\right) = \text{Var}\left(\frac{\tau_i}{x_i} - \frac{T}{x_T}\right) + \text{Var}\left(\frac{T}{x_T}\right).$$

Variance Components Elicitation (cont.)

- Now assessors can be queried about the middle spread around the two quantities separately.

- First, they are asked to give an estimate of middle spread around $\frac{T}{x_T}$, assuming accuracy of the sum x_T as an estimate of T.

- Second, they are then asked for the middle spread around each $\frac{\tau_i}{x_i}$ temporarily assuming again that $\frac{T}{x_T} = 1$ so there is no second component to the variance to consider.

- Once the individual means and variances are elicited, these $\frac{x_i}{x_T}$ values can be plugged into an assumed distribution defined over $[0\!:\!1]$ (since they are proportions) to create a complete prior distribution specification.

- Two direct forms are the normal CDF and the beta distribution.

- If $\frac{\tau_i}{T}$ is assumed to be distributed beta, then we can readily solve for the parameters with two moment equations: $E\left[\frac{\tau_i}{T}\right] = \frac{\alpha}{\alpha+\beta}$, $\text{Var}\left(\frac{\tau_i}{T}\right) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Variance Components Elicitation Example

- An expert on minority electoral participation is asked to estimate upcoming Hispanic turnout for n precincts in a given district: $\tau_1, \tau_2, \ldots, \tau_n$, with total Hispanic turnout in the district equal to T.
- The expert first gives estimates $x_1, x_2, \ldots, x_n$ for each precinct which produces a district turnout estimate of T by summing, x_T.
- The expert is then asked to provide the two-to-one interval for $\frac{T}{x_T}$, which is $[0.7\!:\!1.3]$, meaning they believe that the summed estimate of Hispanic turnout is correct to plus or minus 30% with probability 0.66 (from the two-to-one interval).
- To test the expert's conviction about this variance, the value $\sigma_T = 0.3$ is plugged into the appropriate normal CDF at levels to give the credible interval summaries:

$$\begin{aligned} 50\%\, CI &= [\Phi_{\mu=1,\sigma=0.3}(0.25)\!:\!\Phi_{\mu=1,\sigma=0.3}(0.75)] \\ &= [0.798\!:\!1.202] \end{aligned} \qquad \begin{aligned} 99\%\, CI &= [\Phi_{\mu=1,\sigma=0.3}(0.01)\!:\!\Phi_{\mu=1,\sigma=0.3}(0.99)] \\ &= [0.302\!:\!1.698] \end{aligned}$$

  which are then read back to the expert.
- If they agree that these are reasonable summaries then the variance is set to $\sigma_T^2 = (0.3)^2 = 0.09$.

Variance Components Elicitation Example

- Next the expert is asked to repeat this process for each of the x_i estimates under the temporary assumption that $x_T = T$.
- This "certainty" means that the right-hand-side reduces to the variance of $\frac{\tau_i}{x_i}$ and the expert can perform the same interval process as was done with $\frac{T}{x_T}$ for each of the n precincts.
- Suppose that two-to-one interval for the estimate of Hispanic turnout at the first precinct ($x_1 = 0.2$) is given as $[0.5 : 1.5]$, meaning that the estimate is believed to be correct to plus or minus 50% with probability 0.66.
- This implies a variance of $\sigma_1 = (0.5)^2 = 0.25$, and we will assume that the subsequent 50% and 99% credible interval summaries are approved by the expert.
- The total elicited variance for the first precinct is where x_1^2 is moved back to the right-hand-side:

$$\text{Var}(\tau_1) = x_1^2(0.25 + 0.09) = 0.0136.$$

Notice, incidentally, the dependency here on the modal estimate x_1.

Elicited Prior Coherence

- If P-methods are used to elicit prior information from assessors then a primary concern is *coherence*, ensuring that the resulting probability statements produce valid probability functions.
- For one country consider the events: L for losing a war, T for settling a war by treaty, O concluding a war by another means (victory, armistice, temporary cessation), and W^c for no war.
- The elicitations provided by the expert are:

$$p(L) = 0.33 \qquad p(T) = 0.27 \qquad p(O) = 0.23 \qquad p(W^c) = 0.12$$

$$p(L|W) = 0.41 \qquad p(T|W) = 0.31 \qquad p(O|W) = 0.28.$$

- Incoherency: $p(L) + p(T) + p(O) + p(W^c) = 0.95$, even though $p(L|W) + p(T|W) + p(O|W) = 1.00$.
- More subtly, the conditional probability ratios do not coincide with the unconditional probability ratios: $p(L|W)/p(T|W) = 1.32$, but $p(L)/p(T) = 1.22$, which is a logical disagreement since L and T are both subsets of the event W.

Elicited Prior Coherence

- V-methods are generally safer from incoherence because researchers control the probability levels, and ask for values.

- The only necessary caution stems from the possibility of substantively illogical directional responses. For instance, the following paired level responses are incoherent:

Question	Response
If there is 60% probability that the war lasts an additional 2 years, how many total casualties do you anticipate?	5,000
If there is 20% probability that the war lasts an additional 2 years, how many total casualties do you anticipate?	8,000

Elicited Prior Consistency

- Consistency means that the assessor applies the same subjective personal criteria across events.

- Therefore consistency is a property *within* each assessor that is violated when the individual produces probabilities or levels that are not directly comparable.

- Consider the infamous "Russian judge" stereotype for grading ice skaters where the defining characteristic is that he or she gives low grades for *all* skaters relative to the other judges.

- This is not considered a problem since ice skating contests are won by relative scores and as long as the Russian judge is *consistent* in their harshness, the contest is considered "fair."

Elicited Prior Consistency

- Unfortunately, consistency measured between assessors is more complex.

- Whenever information is acquired from multiple assessors, the resulting prior is a combination of these elicitations.

- Probability or level statements can be averaged, or averaged with weighting.

- If the variable is discrete and it is essential to preserve this discreteness in the prior, then the elicitations can be considered "votes" and the prior is the histogram of the resulting values.

- Since a histogram is actually a density estimate, prior uncertainty can be obtained parametrically or nonparametrically from the variance of this histogram.

Elicited Prior Calibration

- Seidenfeld (1985) formalizes criteria for *calibrated* subjective interpretation of probability statements as a means of verifying good elicitation:
  - **Coherence.** Assessors have "belief-states" that are modeled by well-behaved conditional probability statements.
  - **Total Evidence.** Assessors include all relevant information known at the time of elicitation, including background knowledge as well as the information directly queried by the researcher.
  - **Conditionalization.** Assessors update their knowledge when new evidence is observed in the same way that Bayesian prior information is updated by conditioning on observed data provided through the likelihood function.
- These criteria imply that the process of elicitation is inherently Bayesian in nature, even before the formal calculation of the posterior (for a counter-argument, see Kahneman and Tversky [1982]).

Elicited Prior Calibration (cont.)

- Calibration is also defined in terms of inter-rater consistency across elicitations.

- Lindley, Tversky, and Brown (1979) point out that calibration is a characteristic of multiple assessments compared with past accuracy.

- To use their example, a well-calibrated meteorologist is one whose record of prediction is correct on average: rain occurs 2/3 of the time that they make a 2/3 prediction of rain (although variances could differ as well, giving a different view of reliability).

- So calibration differs from consistency since it implies a direct comparison with subsequently observed events that are confirming or disconfirming.

Bias Issues with Elicited Priors

- Expert bias is a problem because it can harm the quality of the results by altering inferences in a direction away from the target population values, even with reasonably large sample size.
- *Motivational bias* occurs when the expert responds to social group pressure within the project or research environment generally.
- *Cognitive bias* comes from: inconsistency (discussed above), "anchoring," memory limitations, and underestimation of uncertainty.
- Anchoring is a psychological phenomenon in assessors whereby they fix an early response as a reference point for subsequent responses, whether this is appropriate or not.

Bias Issues with Elicited Priors (cont.)

- Bias also comes from a (familiar) third source: sample selection.
- The assessor or group of assessors picked for the study could have some systematic bias about probabilities of occurrence based on their backgrounds.
- For instance, querying Defense Department experts and State Department experts about the probability of a war can lead to radically different answers due to their respective professional orientations (trained to fight wars versus trained to avoid wars).
- Elicitor selection bias can often be detected by picking a broad range of assessors and contrasting their responses to key questions.

Bias Issues with Elicited Priors (cont.)

- Meyer and Booker (2001) give some specific advice for reducing potential bias in a given study:
  - ▻ Anticipate potential biases and test for their existence through inter- and intra-assessor comparisons.
  - ▻ Design, and possibly redesign, the study to minimize anticipated or observed biases.
  - ▻ Minimize bias through pre-training of assessors.
  - ▻ Monitor the elicitation process and check the full set of elicitations.
- In other words: be careful.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- In 2002, we (i.e., Lee Walker) administered a general survey in person to individuals in twenty community-based organizations across two cities in Nicaragua ($n = 226$).
- The outcome variable is a dichotomous measure of trust in justice system fairness from the survey.
- The explanatory variables include measures of: political attitudes, ideology, occupation, and religion.

- Stipulate a conventional logit likelihood function for $L(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y})$, but elicit expert-based priors for $p(\boldsymbol{\beta})$.
- The posteriors were produced by conditioning this prior on the observed data, $\pi(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) \propto p(\boldsymbol{\beta})L(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y})$, for the coefficient vector $\boldsymbol{\beta}$ and the data $\mathbf{X},\mathbf{Y}$.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- The prior distributions were produced according to a P-method procedure modified from the (linear) method of Kadane, *et al.* (1980)
- Respondents asked for probability of trust, $P(\tilde{\mathbf{Y}}_i = 1|\tilde{\mathbf{X}})$ corresponding to chosen $\tilde{\mathbf{X}}$ levels.
- They were then used as probability outcomes to produce point estimates with a standard logit link function as if the design matrix constituted real data.
- 11 community elites consented to participate in the investigation: 2 judicial elites, 3 leaders of groups associated with the political right, and 6 leaders of groups associated with the political left.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- Establish m design points of the explanatory variable vector: $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_m$, representing interesting cases or values spanning the range of the k variables.

- The assessors are then asked to study each of the $\tilde{\mathbf{X}}_i$ scenarios and produce $P(\tilde{\mathbf{Y}}_i = 1|\tilde{\mathbf{X}}_i)$.

- These values then represent "typical" responses to the hypothesized design points specified in the $\tilde{\mathbf{X}}_i$ and could be labeled $\mathbf{Y}_{i,0.50}$ with the bounded support $[0{:}1]$.

- Reexpress as a log-odds model by moving the link function to the LHS

$$\log\left[\frac{P(\tilde{\mathbf{Y}}_i = 1|\tilde{\mathbf{X}}_i)}{P(\tilde{\mathbf{Y}}_i = 0|\tilde{\mathbf{X}}_i)}\right] = \log\left[\frac{P(\tilde{\mathbf{Y}}_i = 1|\tilde{\mathbf{X}}_i)}{1 - P(\tilde{\mathbf{Y}}_i = 1|\tilde{\mathbf{X}}_i)}\right] = \tilde{\mathbf{X}}_i\boldsymbol{\beta},$$

  and run a linear model for $\boldsymbol{\beta}$ point estimates.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- Assume that the prior distribution for $\boldsymbol{\beta}$ is students-t around the estimated points with greater than 2 degrees of freedom.

- Thus we are actually specifying a somewhat conservative prior since large data size under weak conditions leads to Bayesian posterior normality of linear model coefficients, and t-distributed forms with smaller data size (i.e., Berger 1985, 224; Lindley and Smith 1972).

- Unfortunately there is no direct guidance about setting the degrees of freedom for this t-distribution since the m value was established arbitrarily by the researchers, and it is not generally helpful to elicit a degrees of freedom parameter directly from subject matter experts. Obtaining it from the data here is not helpful either because the elicitation process is supposed to take place before conditioning on the observations.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- To solve this problem Kadane, *et al.* suggest a further continuation of questioning.

- After eliciting $\mathbf{Y}_{i,0.50}$ for each $\tilde{\mathbf{X}}_i$, also elicit $\mathbf{Y}_{i,0.75}$ by asking for the median point above the median point just provided.

- Elicit now according to this scheme two more times in the same direction for $\mathbf{Y}_{i,0.875}$, and $\mathbf{Y}_{i,0.9375}$.

- Then for each of the m assessments calculate the ratio:

$$a(\tilde{\mathbf{X}}) = (\mathbf{Y}_{i,0.9375} - \mathbf{Y}_{i,0.50})/(\mathbf{Y}_{i,0.75} - \mathbf{Y}_{i,0.50}),$$

  where the subtraction makes the numerator and denominator independent of the center.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

- This ratio uniquely describes tail behavior for some t-distribution because it is the relative “drop-off” in quantiles. Kadane, *et al.* tabulate degrees of freedom against values of this ratio:

df	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	∞
$a(\tilde{\mathbf{X}})$	2.76	2.62	2.53	2.48	2.45	2.42	2.40	2.39	2.37	2.36	2.35	2.34	2.33	2.31	2.31	2.30	2.27.

- Values greater than 2.76 indicate that the assessor should reevaluate their responses, and values less than 2.27 imply that a standard normal prior centered at the (original) $\boldsymbol{\beta}_{0.50}$ can be used.

Elicited Priors Application: Trust in the Nicaraguan Judicial System

	Posterior	Prior Means			Pooled
Respondent	Mean	Judicial	Left of	Right of	Standard
Characteristic	1997	Group	Center	Center	Error
Politicized	-0.996	-1.609	-1.386	-0.916	(0.668)
Center	-0.567	-1.204	-0.728	-1.235	(0.872)
Left	-0.785	0.150	-0.883	-0.192	(0.720)
Blue Collar	0.383	0.357	0.246	0.313	(1.072)
White Collar	0.270	-0.051	-0.296	0.214	(0.825)
Catholic	-1.175	-1.609	0.514	-0.273	(0.744)

Elicited Priors Application: Trust in the Nicaraguan Judicial System

Respondent	Judicial Prior		Left Prior		Right Prior		Uniform Prior	
Characteristic	Coef.	90% CI	Coef.	90% CI	Coef.	90% CI	Coef.	90% CI
(Intercept)	0.752	[0.09: 1.41]	0.582	[-0.05: 1.21]	0.492	[-0.15: 1.14]	0.623	[-0.27: 1.51]
Politicized	-1.573	[-2.11:-1.03]	-1.397	[-1.90:-0.90]	-1.408	[-1.92:-0.90]	-1.638	[-2.38:-0.90]
Center	-0.554	[-1.13: 0.02]	-0.569	[-1.13:-0.01]	-0.597	[-1.17:-0.02]	-0.573	[-1.29: 0.14]
Left	-0.740	[-1.32:-0.16]	-0.859	[-1.42:-0.30]	-0.823	[-1.39:-0.25]	-0.983	[-1.76:-0.20]
Blue Collar	0.912	[0.32: 1.50]	0.768	[0.20: 1.34]	1.053	[0.47: 1.64]	1.143	[0.36: 1.93]
White Collar	0.590	[0.02: 1.16]	0.439	[-0.12: 1.00]	0.702	[0.13: 1.27]	0.783	[0.06: 1.51]
Catholic	-0.732	[-1.26:-0.20]	-0.354	[-0.85: 0.15]	-0.475	[-1.00: 0.03]	-0.577	[-1.24: 0.09]