Indirect prior elicitation for generalised linear models with R package indirect

Geoffrey R. Hosack
Commonwealth Scientific and Industrial Research Organisation

Abstract

The R package **indirect** supports the elicitation of multivariate normal priors for generalised linear models from domain experts. The software can be applied to indirect elicitation for a generalised linear model that is linear in the parameters. That is, the linear predictor can admit interactions, polynomial functions of the covariates or other choice of basis functions. The package is designed such that the facilitator executes functions within the R console during the elicitation session to provide graphical and numerical feedback at each design point. Various methodologies for eliciting fractiles (equivalently, percentiles or quantiles) are supported. For example, experts may be asked to provide central credible intervals that correspond to a certain probability. Or experts may be allowed to vary the probability allocated to the central credible interval for each design point. Additionally, a median may or may not be elicited. The package provides automatic document generation that summarises the elicitation session for the participating expert at the conclusion of the session.

Keywords: expert opinion, generalised linear models, subjective probability, elicitation, R.

1. Introduction

The R package **indirect** is introduced to support the elicitation of multivariate normal priors for generalised linear models. Key guidance for the general elicitation of subjective probability distributions from domain experts is given by Garthwaite, Kadane, and O'Hagan (2005) and O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley, and Rakow (2006). These references cover the importance of preparing and educating experts prior to an elicitation session, identify the need to clearly elucidate the targets of an elicitation session, and compare the advantages and disadvantages of choices of elicitation protocols.

A subcategory of elicitation procedures focuses on targeting potentially observable quantities that arise from a parametric model instead of targeting elicitation on the parameter directly. This form of elicitation may be referred to as "indirect" elicitation (Winkler 1967). For example, rather than eliciting a distribution for a binomial probability parameter directly, an expert may instead be asked to consider hypothetical observations, which are then used to infer a subjective probability distribution for the probability parameter.

An important application area of indirect elicitation is the prior elicitation for regression models (Low Choy, O'Leary, and Mengersen 2009). Generally, it is thought that indirect elicitation is an easier task compared to an attempted direct assessment of multi-dimensional probability distributions for the unknown parameters (Kadane, Dickey, Winkler, Smith, and Peters

1980; O'Hagan *et al.* 2006). For binomial regression models, software is available from the author (James, Low Choy, and Mengersen 2010). Low Choy, Murray, James, and Mengersen (2010) describe an extension of this software that implements the Conditional Mean Prior (CMP) approach of Bedrick, Christensen, and Johnson (1996) for binomial regression models in a way that can apply to other generalised linear models (Low-Choy, James, Murray, and Mengersen 2012). Garthwaite, Al-Awadhi, Elfadaly, and Jenkinson (2013) also use the CMP approach to elicit multivariate normal priors for generalised linear models, which has associated software available for download (`http://statistics.open.ac.uk/elicitation`). This approach elicits quantiles for the expected response of a generalised linear model. A drawback is that no interactions among the covariates are allowed to influence the response. Elfadaly and Garthwaite (2015) discuss an extension of the software to gamma regression and the normal linear model. At this time, no `R` packages other than `indirect` exist for indirect elicitation of generalised linear models.

The scope of the package `indirect` is focused on supporting the elicitation session, recording of results and reporting summaries for indirect elicitation of multivariate normal priors for generalised linear models. All functions in package **indirect** are implemented using the R system for statistical computing R Core Team (2017). R is available from the comprehensive R archive network (CRAN, `http://CRAN.R-project.org/`), which is distributed under the terms of the GNU General Public License, either Version 2 (GPL-2) or Version 3 (GPL-3). The **indirect** package is available from CRAN under the GPL-3 license (Hosack 2018).

2. Generalised linear model

The generalised linear model (GLM) has three components (McCullagh and Nelder 1989):

1. An *observation model*, $p(y_i|\theta_i, \xi)$, for data y_i conditional on the expected response $\mathrm{E}[y_i] = \theta_i$ at each design point, $i = 1, \ldots, n$. The observation model is chosen from the exponential family and may include additional parameters ξ.

2. The *linear predictor*,
$$\eta_i = x_i^\top \beta, \tag{1}$$
where the $p \times 1$ vector x_i may encode continuous or categorical covariates at the i^{th} design point, and β is the $p \times 1$ vector of unknown parameters.

3. An invertible *link function*, $g(\theta_i) = \eta_i$, that models the relationship between the expected response and the linear predictor.

3. Independent conditional mean priors

A proper prior for the unknown parameters, $p(\beta)$ is sought, which can take various forms. However, direct elicitation of the parameters β would be exceedingly difficult for experts. An alternative approach indirectly elicits the prior $p(\beta)$ given subjective probability distributions elicited on an interpretable scale. The mean response θ_i usually is accessible to experts in terms of units and definition. For example, the response θ_i may be a percentage, a probability, an abundance, or a density given known covariates $x_i^\top$. The task is then to elicit independent conditional mean priors for the mean response θ_i at each design point or scenario $x_i^\top$,

$i = 1, \ldots, n$. In particular, specifying a normal prior for β induces a class of independent conditional mean priors within the generalised linear model framework (Bedrick *et al.* 1996). In many statistical applications a normal prior is specified (Garthwaite *et al.* 2013; Hosack, Hayes, and Barry 2017), $p(\beta) = N(\mu, \Sigma)$, and this is the basic assumption used in package **indirect**.

3.1. Specifying the design matrix

The design points (scenarios) $x_i^\top$ for $i = 1, \ldots, n$ compose the rows of the $n \times p$ design matrix X. The matrix X is assumed to have full column rank $p \leq n$. Given a normal prior fixed for β, the independence property for the conditional mean priors does not hold for all possible choices of the design matrix X. Nevertheless, the independence assumption is often reasonable if the design points are spread out in a certain sense (Bedrick *et al.* 1996). Optimal design, such as using balanced designs, can be used to assist with this objective (Hosack *et al.* 2017). In general, the design matrix may be arbitrary and include interactions or basis functions.

A suggested diagnostic for general X is the condition number of a rescaled design matrix X_s where each column of X is scaled to unit length (Bedrick *et al.* 1996). This diagnostic is implemented with function `CNdiag` in package **indirect**. A large condition number $\kappa(X_s)$ may suggest, but does not necessarily indicate, dependency in the design matrix (Belsley, Kuh, and Welsch 2005). For the linear system

$$\eta = X\beta, \tag{2}$$

Thisted (1988) notes that if X and η "are 'good to t decimal places', then the solution to the linear system $[\beta]$ may only be good to $t - \log_{10}(\kappa(X))$ decimal places". Note that this interpretation should here be applied on the scale of the linear predictor η.

For example, consider a balanced design that specifies one design point to each of three categorical variables. This produces a low condition number.

```
R> X <- matrix(c(rep(1, 3), c(0, 1, 0), c(0, 0, 1)), nrow = 3,
+               dimnames = list(designPt = 1:3,paste0("covar", 1:3)))
R> X

designPt covar1 covar2 covar3
       1      1      0      0
       2      1      1      0
       3      1      0      1

R> indirect::CNdiag(X)

[1] 4.294698
```

The above example is also D-optimal if the second and third covariates are instead continuous. Contrast the above result with a suboptimal design, where the second column (covariate) of the design matrix, now continuous, has been adjusted so that the second and third design points (i.e., the second and third rows of X) are very close to each other in the design space.

```
R> X <- matrix(c(rep(1, 3), c(0, 0.1, 0.9), c(0, 0, 1)), nrow = 3,
+               dimnames = list(designPt = 1:3, paste0("covar", 1:3)))
R> X

designPt covar1 covar2 covar3
       1      1    0.0      0
       2      1    0.1      0
       3      1    0.9      1

R> indirect::CNdiag(X)

[1] 34.40988
```

The condition number diagnostic has increased. The relatively high condition number indicates that the design points may not be sufficiently spread out in the latter example.

3.2. Eliciting independent conditional mean priors

Conditional on a given scenario described by the design point $x_i^\top$, the elicitation exercise seeks to elicit from the expert a subjective probability distribution for the expected response θ_i. Again the elicitation target θ_i typically is chosen to represent an interpretable quantity to an expert and is so defined on a scale familiar to the expert, e.g., in units of proportion, probability, abundance or density (Hosack *et al.* 2017). Generally, the advice for efficient elicitation of a subjective probability distribution supports the elicitation of fractiles (equivalently, quantiles or percentiles) from experts instead of most likely estimates or moments such as means and variances (Garthwaite *et al.* 2005; O'Hagan *et al.* 2006). Garthwaite *et al.* (2013) elicits fractiles from experts in a conditional mean prior approach. Fractiles are also the elicited quantities for the elicitation target θ_i in package **indirect**.

For an arbitrary distribution function $F(t)$, the q^{th} fractile is defined as $f = F^{-1}(q)$. At each scenario, a finite set of K fractiles is elicited from the expert[1]. Form the vector $f = [f_1, \ldots, f_K]^\top$ with associated probabilities $q = [q_1, \ldots, q_K]^\top$, where $q_k = F(f_k)$ for $k = 1, \ldots, K$.[2] These fractiles are used to bound $K + 1$ bins, B_k, $k = 1, \ldots, K + 1$, where each B_k is a real interval and Lebesgue measurable. The bins have bounds $(-\infty, f_1]$ for B_1, bounds $(f_{k-1}, f_k]$ for $1 < k \leq K$, and bounds (f_K, ∞) for $k = K + 1$. The collection of bins $\{B_k, k = 1, \ldots, K + 1\}$ forms a discrete set. The elicited distribution P_e is approximated by assigning probability $p_1 = q_1$ to bin B_1, probability $p_k = q_k - q_{k-1}$ to B_k for $1 < k \leq K$ and probability $p_{K+1} = 1 - q_K$ to bin B_{K+1}. The elicited distribution is thus essentially approximated by a histogram, with the bins defined by the support of the target distribution and the bounds of the elicited credible intervals.

The goal is to derive a normal prior for the unknown coefficients β. A normal distribution is therefore elicited on the linear predictor scale. This process begins by transforming the fractiles f through the monotonic link function $g(\cdot)$. For a given normal distribution $P_s(\eta_i)$ with

[1] Typically K is a small number. Many strategies for choosing the set of fractiles to elicit have been proposed in the literature. Several of these approaches are supported by package **indirect**; further discussion is postponed until Section 3.3. For the moment, assume that a set of K fractiles have been elicited.

[2] The dependence of the fractiles f and associated probabilities q on the i^{th} design point is suppressed here to simplify notation.

mean m_i and variance v_i, an approximation to the elicited probability intervals is constructed. This normal distribution assigns probability $\rho_1 = \int_{-\infty}^{g(f_1)} N(t|m_i, v_i)dt$ to B_1, probability $\rho_k = \int_{g(f_{k-1})}^{g(f_k)} N(t|m_i, v_i)dt$ to B_k for $1 < k \leq K$ and probability $\rho_{K+1} = \int_{g(f_K)}^{\infty} N(t|m_i, v_i)dt$ to B_{K+1}.

Normal distributions have been fitted to elicited credible intervals using various techniques (O'Hagan *et al.* 2006). An optimisation of the parameters m_i and v_i requires the specification of an objective function. One possibility is least squares (O'Hagan *et al.* 2006), which corresponds to choosing m_i and v_i such that the sum of squares

$$\sum_{k}^{K+1} (p_k - \rho_k)^2 \tag{3}$$

is minimised[3]. This objective function is supported by **indirect**.

Another possibility is to minimise the Kullback-Leibler divergence from the parametric subjective probability distribution P_s to the unknown elicited distribution P_e, which is described only by the raw elicited fractiles (Hosack *et al.* 2017). This approach seeks to minimise the loss from reporting P_s if P_e is true under a logarithmic utility function. The objective function is given by a discretised approximation to the Kullback–Leibler divergence,

$$KL(P_e : P_s) = \int \log \frac{dP_e}{dP_s} dP_e \approx \sum_{k}^{K+1} \log\left(\frac{p_k}{\rho_k}\right) p_k. \tag{4}$$

The discretised approximation generally results in a loss of information (Kullback 1959). Subject to regularity conditions, the information loss can be reduced to an arbitrarily small amount by further partitioning (Kale 1964), that is, increasing the number of elicited fractiles K. This approximate Kullback-Leibler divergence objective function was implemented by Hosack *et al.* (2017) and is also supported by **indirect**.

3.3. Which fractiles to elicit?

In practice, only a small number of fractiles are pragmatic to elicit. The following strategies are supported:

- Any arbitrary central credible interval, which may either be preset by the facilitator or chosen by the expert. The probability associated with the central credible interval is allowed to vary by design point.
- Any central credible interval and also the median. This allows the inclusion of the median as a central point estimate, which is used for example by the method of bisection (Garthwaite *et al.* 2005), see also Hosack *et al.* (2017) for an example using indirect elicitation.

The latter method elicits more data than free parameters. O'Hagan *et al.* (2006) call this process "overfitting", and argues that overfitting allows the expert to more critically assess an approximating parametric distribution. The elicited fractiles f are a step towards this goal, and may well be adjusted several times until the expert judges the distribution $P_s(\theta)$, which

[3]The dependence of ρ_k on m_i and v_i is suppressed to simplify notation.

is the distribution $P_s(\eta)$ transformed by the inverse link function $g^{-1}(\cdot)$, to be an acceptable representation of their belief. **Always remember that the subjective probability distribution P_s, which is a normal distribution on the linear predictor scale determined by the invertible link function $g(\cdot)$, is ultimately the elicited "data".** The package **indirect** provides both graphical and numerical feedback to the expert to facilitate this process of constructing an acceptable subjective probability distribution (see Section 4 for illustrations).

3.4. The induced prior

The independent conditional mean prior is normally distributed on the linear predictor scale, $p(\eta) = N(\eta|m, V)$, with location vector $m = [m_1, \ldots, m_N]^\top$ and diagonal covariance matrix $V = \text{diag}[v_1, \ldots, v_N]$. Conditional on the elicited data and a design matrix X of full column rank, the probability distribution of η is given by,

$$\begin{aligned}\prod_{i=1}^{n} p\left(\eta_i|m_i, v_i\right) &= \prod_{i=1}^{n} p\left(x_i^\top\beta\middle|\, m_i, v_i\right)\\ &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} v_i^{-1}(x_i^\top\beta - m_i)^2\right\}\\ &= \exp\left\{-\frac{1}{2}\text{Tr}\left[V^{-1}(X\beta - m)(X\beta - m)^\top\right]\right\}\\ &= \exp\left\{-\frac{1}{2}(X\beta - m)^\top V^{-1}(X\beta - m)\right\}, \end{aligned} \tag{5}$$

which is proportional to the exponential of a quadratic form in β.

The distribution for the unknown β conditional on m and V is proportional to the multivariate normal distribution,

$$\begin{aligned} p(\beta|m, V) \propto &\exp\left\{-\frac{1}{2}\left[\beta^\top X^\top V^{-1} X\beta - 2m^\top V^{-1} X\beta\right]\right\}\\ = &\exp\left\{-\frac{1}{2}\left(\beta-\mu\right)^\top \Sigma^{-1}\left(\beta-\mu\right) + \frac{1}{2}\, X^\top V^{-1} m\left(X^\top V^{-1} X\right)^{-1} m^\top V^{-1} X\right\}. \end{aligned}$$

Given the assumptions of a normally distributed independent conditional mean prior, the induced normal prior on the unknown coefficients of the generalised linear model is given by,

$$p(\beta) = N(\beta|\mu, \Sigma) \tag{6}$$

where $\mu = (X^\top V^{-1} X)^{-1} X^\top V^{-1} m$ and $\Sigma = (X^\top V^{-1} X)^{-1}$ (Bedrick *et al.* 1996; Hosack *et al.* 2017). Given the proper prior $p(\beta)$, the Bayesian update

$$p(\beta|y_1, y_2, \ldots, y_L) \propto p(y_L|y_1, \ldots, y_{L-1}, \beta) p(y_{L-1}|y_1, \ldots, y_{L-2}, \beta) \ldots p(y_1|\beta) p(\beta), \tag{7}$$

can now be obtained for future empirical observations y_l, $l = 1, \ldots, L$.

4. Illustrative example

There are 3 categories of functions in package **indirect**:

1. Elicitation functions that do the following:
   - specify the problem structure, for example, the design points and link function, and
   - assimilate expert statements in the form of fractiles, percentiles or quantiles (note that all of these terms are equivalent) into this problem structure.
2. Fitting functions that map expert statements into models of the expert opinion; many of these are helper functions that typically do not to be accessed by the user.
3. Plotting functions that provide graphical and numerical feedback to expert(s) during the course of the elicitation session.

An example illustration is given here. The demonstration creates an artificial expert that understands the system perfectly. That is, the expert believes in the true model and is able to specify the distribution of β, independent design points X and the correct link function. Obviously this will not happen in nature and this example is intended to simply illustrate the proof of concept.

```
R> set.seed(100)
R> # number of covariates
R> p <- 5
R> # mean  beta
R> mu <- rnorm(p)
R> # simulate covariance matrix from inverse Wishart
R> # diagonal scale matrix and p + 5 d.f. nu
R> alpha <- MASS::mvrnorm(p + 5, mu = rep(0, p), Sigma = diag(p)*50)
R> initial.icov <- t(alpha[1, , drop = FALSE])%*%alpha[1, , drop = FALSE]
R> for (i in 2:ncol(alpha)) {
+     initial.icov <- initial.icov +
+       + t(alpha[i, , drop = FALSE])%*%alpha[i, , drop = FALSE]
+   }
R> Sigma <- chol2inv(chol(initial.icov))
R> # Design with independence priors:
R> # the following choice of design matrix produces
R> # independent conditional mean priors.
R> # Of course, in a real elicitation session the prior
R> # for beta is unknown and so this example is only for illustration.
R> # This implements an Independent Conditional Mean prior as
R> # defined by Bedrick et al. (1996), p. 1458.
R> P <- diag(p) # identity matrix used (could use any orthogonal transformation)
R> X <- P%*%solve(t(chol(Sigma)))
R> D <- diag(1/rnorm(p, -X%*%mu, 0.5)) # arbitrary diagonal matrix
R> X <- round(D%*%X, digits = 6)
R> rownames(X) <- paste("DesignPt", 1:nrow(X))
R> colnames(X) <- paste("Covariate", 1:ncol(X))
R> X
```

```
           Covariate 1 Covariate 2 Covariate 3 Covariate 4 Covariate 5
DesignPt 1    0.582579    0.000000    0.000000    0.000000    0.000000
DesignPt 2   -1.459825   -7.083247    0.000000    0.000000    0.000000
DesignPt 3    3.633226    1.865657   -2.552306    0.000000    0.000000
DesignPt 4    0.551245   -0.225478   -0.216412   -0.837665    0.000000
DesignPt 5    3.520089   -1.759519    2.083022    0.801938    2.798532
```

```
R> # elicited moments and quartiles
R> g.m <- X%*%mu
R> g.V <- X%*%Sigma%*%t(X)
R> g.theta.median <- qnorm(0.5, g.m, sqrt(diag(g.V)))
R> g.theta.lower <- qnorm(0.25, g.m, sqrt(diag(g.V)))
R> g.theta.upper <- qnorm(0.75, g.m, sqrt(diag(g.V)))
R> # The "perfect" elicitations are stored in the following matrix
R> # perfect expert has cloglog link function
R> perfect.elicitations <- 1 - exp(-exp(cbind(g.theta.lower,
+                                             g.theta.median, g.theta.upper)))
R> colnames(perfect.elicitations) <- c("lower", "median", "upper")
R> perfect.elicitations
```

```
               lower    median     upper
DesignPt 1 0.2457165 0.5259040 0.8612915
DesignPt 2 0.3151375 0.5595308 0.8306763
DesignPt 3 0.1736253 0.2228651 0.2834920
DesignPt 4 0.2918991 0.2996736 0.3076072
DesignPt 5 0.2331577 0.2772015 0.3276358
```

In the above, the elicitations are now recorded in the object `perfect.elicitations`. Of course, this is an artificial situation. In a real session, this elicited information could only be obtained by an exchange between the facilitator and the expert. The package `indirect` facilitates this exchange with a combination of iterative graphical and numerical feedback.

Prior to the start of the elicitation session, it is a good idea to write out a `R` script that will serve as a reproducible transcript of the session. The elicited data and comments contributed by the expert will then be edited into this `R` script. There are also functions to store elicitation `R` objects created during the `R` session and, at the end of the session, share a summary of the session for the expert's own records. In this way, the facilitator is cautiously using multiple mechanisms to document the valuable data created during the elicitation session.

The `R` transcript begins with a creation of an empty elicitation record using the function `designLink`. There is the opportunity to add any introductory comments that may pertain to the session. The facilitator will later have the option of producing a session report. This report will be processed using `Sweave`. The comments will be printed with a call to `Sexpr`, and so it is recommended that the comments only use ASCII text and avoid special characters.

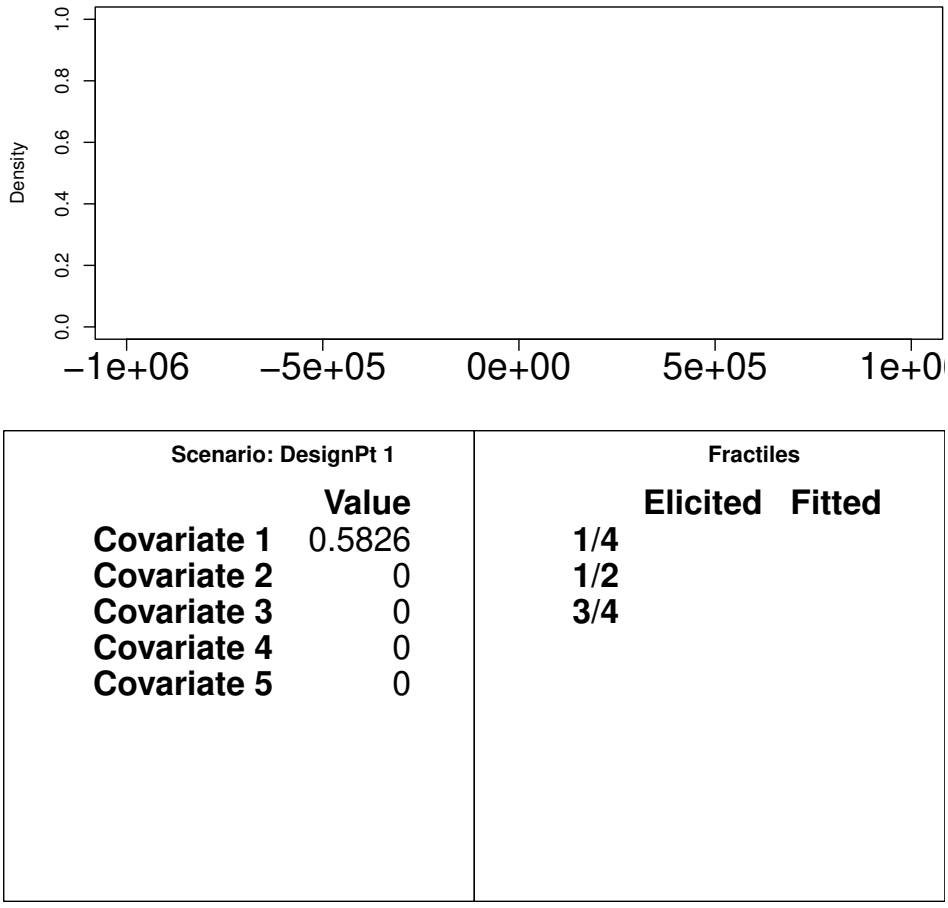
```
R> # Initialise list with elicitation session information.
R> # Here design is the same as X but not usually the case, that is,
R> # the covariates presented to the expert may differ from
```

```
R> # the model design due to transformations, contrasts and coding.
R> # Setting CI.prob = 1/2 specifies that 0.5 probability is allocated to the
R> # central credible interval; the upper and lower bounds
R> # of the central CI are then the upper and lower quartiles.
R> Z <- indirect::designLink(design = X, link = "cloglog",
+     target = "Target", CI.prob = 1/2,
+     intro.comments = "This is a record of the elicitation session.",
+     expertID = "Expert", facilitator = "Facilitator",
+     rapporteur = "none")
R> Z

$design
           Covariate 1 Covariate 2 Covariate 3 Covariate 4 Covariate 5
DesignPt 1    0.582579    0.000000    0.000000    0.000000    0.000000
DesignPt 2   -1.459825   -7.083247    0.000000    0.000000    0.000000
DesignPt 3    3.633226    1.865657   -2.552306    0.000000    0.000000
DesignPt 4    0.551245   -0.225478   -0.216412   -0.837665    0.000000
DesignPt 5    3.520089   -1.759519    2.083022    0.801938    2.798532

$theta
     lower median upper CI_prob
[1,]    NA     NA    NA     0.5
[2,]    NA     NA    NA     0.5
[3,]    NA     NA    NA     0.5
[4,]    NA     NA    NA     0.5
[5,]    NA     NA    NA     0.5

$link
[1] "cloglog"

$target
[1] "Target"

$expertID
[1] "Expert"

$facilitator
[1] "Facilitator"

$rapporteur
[1] "none"

$intro.comments
[1] "This is a record of the elicitation session."

$comments
[1] " " " " " " " " " " " "
```

```
$fit.method
[1] "KL"
```

Now have a look at a plot for the first design point without any elicitations included. The plot will go to the current device, which may require resizing. Usually a plot with the default dimensions (7 inches for both width and height) is sufficient[4].

```
R> # elicitations
R> # design point 1
R> indirect::plotDesignPoint(Z, design.pt = 1)
```

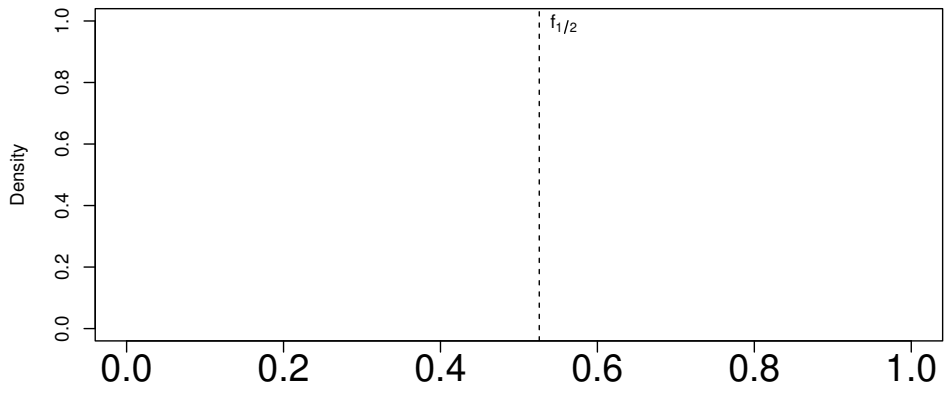


An example elicitation at the first design point is presented. This example applies the approach of Hosack *et al.* (2017), which uses the method of bisection (Garthwaite *et al.* 2005) followed by graphical and numerical feedback. This process iterates until the expert accepts the parametric distribution as an adequate representation of their beliefs. The process begins by restricting the support of the plot to $(0, 1)$, which is appropriate given the complementary log log link, and eliciting the median, which was previously stored in the matrix

[4]RStudio is a good (free) IDE that supports convenient switching among script, R console and the graphical device (https://www.rstudio.com/products/rstudio/download/).

`perfect.elicitations` above. The median is the value that the expert believes gives a 50/50 chance (equivalently, a 1/2 chance, equal odds or probability 0.5) of being above or below the target θ_i.

```
R> # Example elicited fractiles are stored in perfect.elicitations
R> # In a real application, median would be entered as a numeric scalar that was
R> # contributed by the expert.
R> # CI.prob was initially set by designLink
R> Z <- indirect::elicitPt(Z, design.pt = 1,
+                           lower.CI.bound = NA,
+                           median = perfect.elicitations[1, "median"],
+                           upper.CI.bound = NA,
+                           CI.prob = NULL)
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1))
```

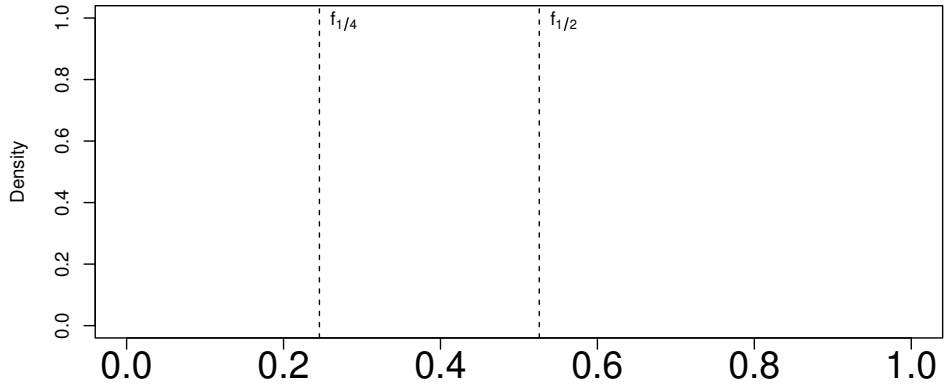


Scenario: DesignPt 1	
	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

Fractiles		
	Elicited	**Fitted**
1/4		
1/2	0.526	
3/4		

Next, the target θ_i is assumed to be below the median. Given this assumption, the expert is asked to provide the value that gives a 50/50 chance that the target is above or below; this value is equivalent to the lower quartile, $f_{1/4}$.

```
R> Z <- indirect::elicitPt(Z, design.pt = 1,
+                          lower.CI.bound = perfect.elicitations[1, "lower"],
+                          median = perfect.elicitations[1, "median"],
+                          upper.CI.bound = NA)
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1))
```



Scenario: DesignPt 1

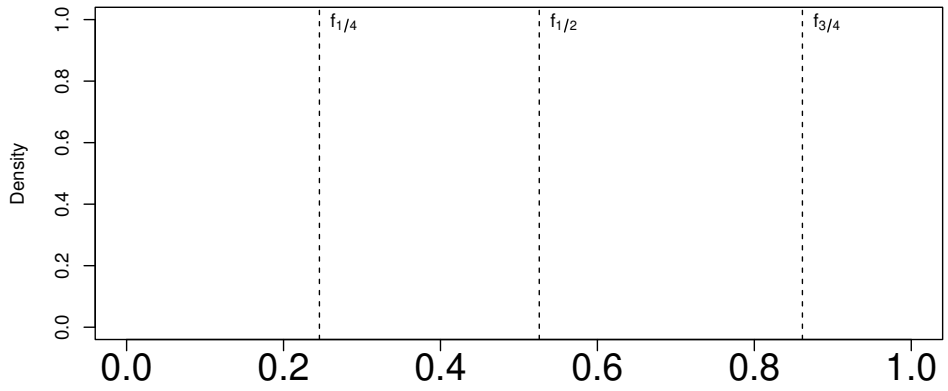
	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

Fractiles

	Elicited	Fitted
1/4	0.246	
1/2	0.526	
3/4		

Next, the target θ_i is assumed to be above the median. Given this assumption, the expert is asked to provide the value that gives a 50/50 chance that the target is above or below; this value is equivalent to the upper quartile, $f_{3/4}$.

```
R> Z <- indirect::elicitPt(Z, design.pt = 1,
+                          lower.CI.bound = perfect.elicitations[1, "lower"],
+                          median = perfect.elicitations[1, "median"],
+                          upper.CI.bound = perfect.elicitations[1, "upper"],
+                          comment = "No major comments.")
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1))
```



Scenario: DesignPt 1

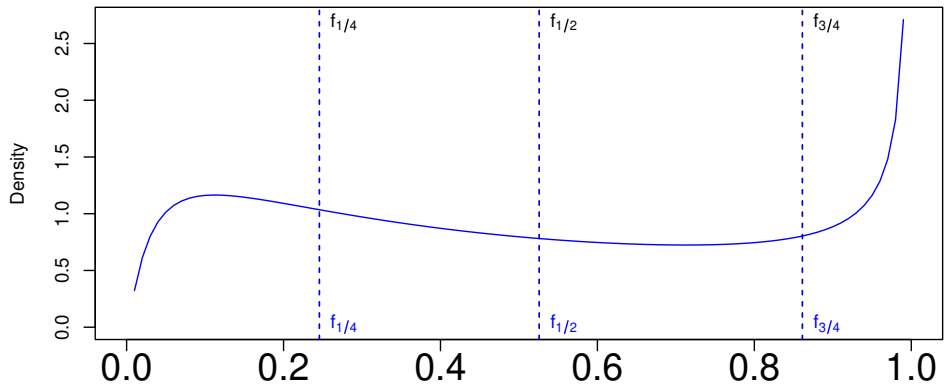
	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

Fractiles

	Elicited	Fitted
1/4	0.246	
1/2	0.526	
3/4	0.861	

These raw elicited fractiles are then compared to the fitted conditional normal that minimises the Kullback–Leibler divergence with partitioning based on the elicited fractiles. The approximation is exact in this example because the expert believes in the true model and reports their beliefs accurately. The subjective probability density function of the conditional normal is also plotted.

```
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1),
+     fitted.fractiles = TRUE, fitted.curve = TRUE)
```



Scenario: DesignPt 1

	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

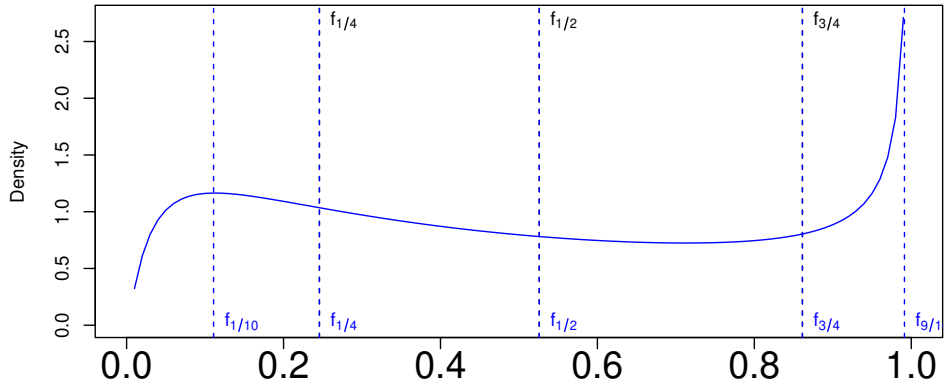
Fractiles

	Elicited	Fitted
1/4	0.246	0.246
1/2	0.526	0.526
3/4	0.861	0.861

Typically the parametric distribution matches the original elicited fractiles f inexactly. In the overfitting process (O'Hagan *et al.* 2006), the fractiles f are then iteratively adjusted until the fitted distribution and fractiles are acceptable to the expert as an adequate representation of their beliefs.

The model is then used to predict out to the extreme deciles, that is, $f_{1/10}$ and $f_{9/10}$. This provides another check in the overfitting process.

```
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1),
+     fitted.fractiles = c(1/10, 1/4, 1/2, 3/4, 9/10), fitted.curve = TRUE)
```



Scenario: DesignPt 1

	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

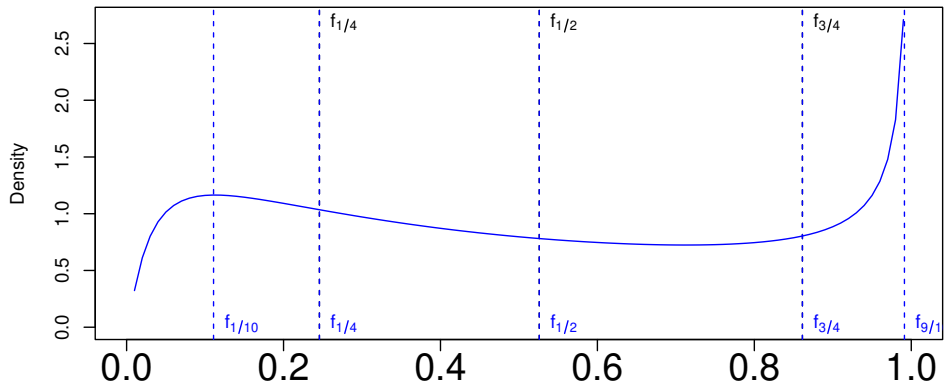
Fractiles

	Elicited	Fitted
1/10		0.111
1/4	0.246	0.246
1/2	0.526	0.526
3/4	0.861	0.861
9/10		0.991

Sometimes the estimated (fitted) cumulative probabilities for different values of the target θ_i are of interest. For example, the cumulative probabilities $P_s(\theta_i = 1/3)$ and $P_s(\theta_i = 1/2)$ can be estimated using the `estimated.probs` argument as follows.

```
R> indirect::plotDesignPoint(Z, design.pt = 1,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1),
+     fitted.fractiles = c(1/10, 1/4, 1/2, 3/4, 9/10), fitted.curve = TRUE,
+     estimated.probs = c(1/3, 0.5))
```

```
P(x <= 0.333333333333333)                P(x <= 0.5)
                0.3362133                   0.4795661
```



Scenario: DesignPt 1

	Value
Covariate 1	0.5826
Covariate 2	0
Covariate 3	0
Covariate 4	0
Covariate 5	0

Fractiles

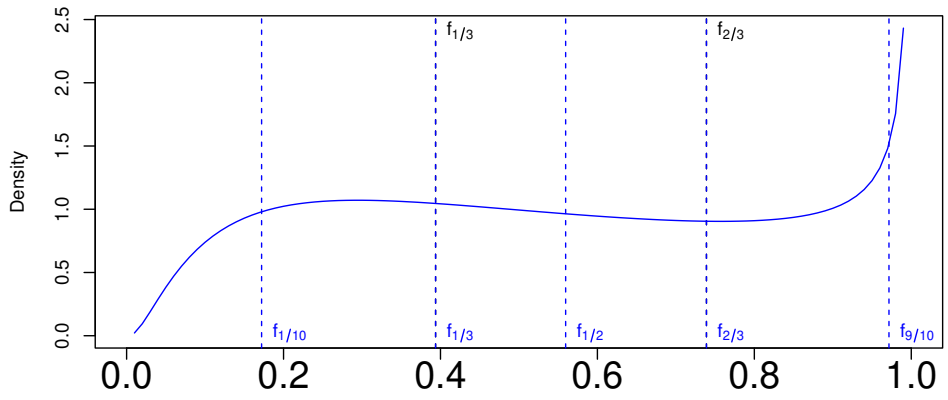
	Elicited	Fitted
1/10		0.111
1/4	0.246	0.246
1/2	0.526	0.526
3/4	0.861	0.861
9/10		0.991

The estimated probabilities are printed to the R console. The raw fractiles may require further adjustment at this point. Once the fitted subjective probability distribution is deemed acceptable by the expert then the elicitation proceeds to the next design point.

At the second design point, say the expert suddenly wished to switch to an alternative method with reference to tertiles, $f_{1/3}$ and lower $f_{2/3}$, rather than quartiles. Further, the expert wished to contribute only the upper and lower tertiles without explicit reference to the median. With this approach, the expert will contribute two fractiles that divide the support of the target into intervals of equal probability or odds (each probability interval above, below, and between the elicited fractiles will have probability 1/3). Suppose that these changes were accepted by the elicitation protocol. The changes can be accommodated in the following way.

```
R> # Pefect elicitations now moving to tertiles for the second design point
R> g.tertiles.d2 <- qnorm(c(1/3, 2/3), g.m[2], sqrt(g.V[2, 2]))
R> # inverse link function
R> theta.tertiles.d2 <- 1 - exp(-exp(c(g.tertiles.d2)))
R> # tertiles only elicited without median
R> Z <- indirect::elicitPt(Z, design.pt = 2, CI.prob = 1/3,
+                          lower.CI.bound = theta.tertiles.d2[1],
+                          upper.CI.bound = theta.tertiles.d2[2],
+        comment = "Switched to tertile method without median.")
```

```
R> indirect::plotDesignPoint(Z, design.pt = 2,
+     elicited.fractiles = TRUE, theta.bounds = c(0, 1),
+     fitted.fractiles = c(1/10, 1/3, 1/2, 2/3, 9/10), fitted.curve = TRUE)
```



Scenario: DesignPt 2

	Value
Covariate 1	−1.46
Covariate 2	−7.083
Covariate 3	0
Covariate 4	0
Covariate 5	0

Fractiles

	Elicited	Fitted
1/10		0.172
1/3	0.394	0.394
1/2		0.56
2/3	0.739	0.739
9/10		0.972

Only the current design point has the central credible interval set to probability 1/3.

```
R> Z$theta
```

```
         lower   median     upper   CI_prob
[1,] 0.2457165 0.525904 0.8612915 0.5000000
[2,] 0.3937816       NA 0.7389746 0.3333333
[3,]        NA       NA        NA 0.5000000
[4,]        NA       NA        NA 0.5000000
[5,]        NA       NA        NA 0.5000000
```

Once this subjective probability distribution is deemed acceptable by the expert then the elicitation proceeds to the next design point, and so on for each design point $x_i^\top$, $i = 1, \ldots, n$.

```
R> # All remaining elicitations are entered into the record
R> # for this artificial elicitation example
```

```
R> Z$theta[3:nrow(perfect.elicitations), 1:3] <-
+     perfect.elicitations[3:nrow(perfect.elicitations), ]
```

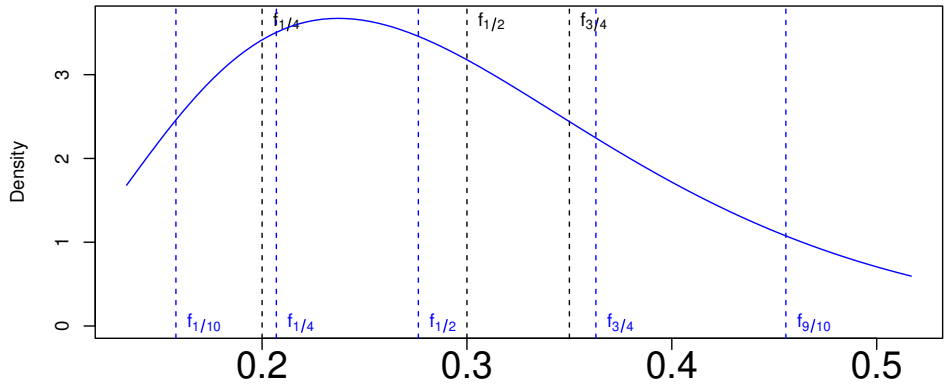
The elicited prior for this particular model can then be obtained with the function `muSigma`.

```
R> prior <- indirect::muSigma(Z, X = Z$design)
R> prior

$mu
                   [,1]
Covariate 1 -0.50219235
Covariate 2  0.13153117
Covariate 3 -0.07891709
Covariate 4  0.88678481
Covariate 5  0.11697127

$Sigma
            Covariate 1 Covariate 2 Covariate 3 Covariate 4 Covariate 5
Covariate 1    6.135611  -1.2645215    7.809760   2.3603956  -15.002022
Covariate 2   -1.264521   0.2867828   -1.590426  -0.4984535    3.097499
Covariate 3    7.809760  -1.5904260    9.981010   2.9888879  -19.108948
Covariate 4    2.360396  -0.4984535    2.988888   0.9184058   -5.770263
Covariate 5  -15.002022   3.0974987  -19.108948  -5.7702627   36.705706

$log.like
           [,1]
[1,] 0.05858989

R> # compare with original prior parameters defined above
R> all.equal(as.numeric(prior$mu), mu)

[1] TRUE

R> all.equal(prior$Sigma, Sigma, check.attributes = FALSE) # a small number

[1] "Mean relative difference: 2.45204e-07"
```

Alternative models may also be considered so long as the information coded in the model matrix X matches with what was presented to the expert via the argument `design` in the function `designLink`.

This would conclude the elicitation session for the perfect expert, which recovers the correct prior. In reality, such a situation is unobtainable. An example is given where the perfect expert elicitation is jittered, so that it is no longer exactly correct. By setting `theta.bounds = NULL`, the plot bounds are allowed to automatically adjust to the credible interval of the elicited subjective probability distribution. The bounds of the credible interval are specified by `cumul.prob.bounds`, with default interval given by $(0.05, 0.95)$.

```
R> # perfect elicitions for design point 5
R> perfect.elicitations[5, ]

    lower    median     upper
0.2331577 0.2772015 0.3276358

R> # jittered elicitations
R> d5.jittered <- c(0.2, 0.3, 0.35)
R> # plot jittered elicitation
R> Z <- indirect::elicitPt(Z, design.pt = 5,
+                          lower.CI.bound = d5.jittered[1],
+                          median = d5.jittered[2],
+                          upper.CI.bound = d5.jittered[3],
+                          comment = "Jittered elicitation.")
R> indirect::plotDesignPoint(Z, design.pt = 5,
+     elicited.fractiles = TRUE, theta.bounds = NULL,
+     cumul.prob.bounds = c(0.05, 0.95),
+     fitted.fractiles = c(1/10, 1/4, 1/2, 3/4, 9/10),
+     fitted.curve = TRUE
+   )
```



Scenario: DesignPt 5

	Value
Covariate 1	3.52
Covariate 2	−1.76
Covariate 3	2.083
Covariate 4	0.8019
Covariate 5	2.799

Fractiles

	Elicited	Fitted
1/10		0.158
1/4	0.2	0.207
1/2	0.3	0.276
3/4	0.35	0.363
9/10		0.456

To conclude the session, the facilitator should ask the expert if there are further questions. After confirmation with the expert, the record can be saved using the function `saveRecord`. This will save the record as a `RDS` object using the base `R` function `saveRDS`.

```
R> # Not run:
R> # for this example, create a temporary directory to store record
R> tmp.rds <- tempfile(pattern = "record", fileext =".rds")
R> # save record to this directory
R> indirect::saveRecord(Z,
+     conclusion.comments = "This concludes the elicitation record.",
+     file = tmp.rds)
```

A summary record of the elicitation session can be shared with the expert by function `makeSweave` that automatically generates a pdf document in the current working directory. The `makeSweave` function sources the elicitation record from the saved `rds` file and creates a `.Rnw` file. This latter file may be processed by the `utils::Sweave` and `tools::texi2pdf` functions to create a `.pdf` document.

```
R> # Not run:
R> tmpReport <- tempfile(pattern = "SessionSummary")
R> indirect::makeSweave(filename.rds = tmp.rds,
+              reportname = tmpReport,
+              title = "Elicitation session record",
+              contact.details = "contact at email address",
+              fitted.fractiles = c(1/10, 1/4, 1/2, 3/4, 9/10))
R> # change working directory to where the record RDS object was stored
R> setwd(tempdir())
R> utils::Sweave(paste0(tmpReport, ".Rnw"))
R> tools::texi2pdf(paste0(tmpReport, ".tex"))
```

The function `makeSweave` saves pdf files of all figures and the summary report document into the current working directory.

5. Summary

This introduction to the R package **indirect** (Hosack 2018) describes the motivation and methods of the functionality provided to support the indirect prior elicitation of multivariate normal priors for generalised linear models. Several approaches to elicitation of central credible intervals within a generalised linear model framework are supported, including versions of the approach of Hosack *et al.* (2017). The goal is to elicit subjective probabilities conditional on different combinations of covariate values at specified design points, or "scenarios". These subjective probabilities subsequently induce a multivariate normal prior for a generalised linear model. Alternative choices of design matrix may be explored outside of the elicitation session without violating the elicitation protocol, as long as the alternative models agree with the information provided to the expert during the session for each design point and the link function is unaltered. Currently the identity, logit, complementary log log, and

log link functions are supported. The basic options for this indirect elictation approach are described here with examples of code usage.

Acknowledgments

The author thanks Keith Hayes and Adrien Ickowicz for their helpful feedback.

References

Bedrick EJ, Christensen R, Johnson W (1996). "A new perspective on priors for generalized linear models." *Journal of the American Statistical Association*, **91**, 1450–1460.

Belsley DA, Kuh E, Welsch RE (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, volume 571. John Wiley and Sons.

Elfadaly FG, Garthwaite PH (2015). "Eliciting prior distributions for extra parameters in some generalized linear models." *Statistical Modelling*, **15**(4), 345–365.

Garthwaite PH, Al-Awadhi SA, Elfadaly FG, Jenkinson DJ (2013). "Prior distribution elicitation for generalized linear and piecewise-linear models." *Journal of Applied Statistics*, **40**(1), 59–75.

Garthwaite PH, Kadane JB, O'Hagan A (2005). "Statistical methods for eliciting probability distributions." *Journal of the American Statistical Association*, **100**(470), 680–701.

Hosack GR (2018). *indirect: Elicitation of Independent Conditional Means Priors for Generalised Linear Models*. R package version 0.1.1.

Hosack GR, Hayes KR, Barry SC (2017). "Prior elicitation for Bayesian generalised linear models with application to risk control option assessment." *Reliability Engineering and System Safety*, **167**, 351 – 361. doi:10.1016/j.ress.2017.06.011.

James A, Low Choy S, Mengersen K (2010). "Elicitator: An expert elicitation tool for regression in ecology." *Environmental Modelling and Software*, **25**, 129–145.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). "Interactive elicitation of opinion for a normal linear model." *Journal of the American Statistical Association*, **75**(372), 845–854.

Kale B (1964). "A note on the loss of information due to grouping of observations." *Biometrika*, pp. 495–497.

Kullback S (1959). *Information Theory and Statistics*. Wiley, New York. Republication by Dover, 1997.

Low-Choy S, James A, Murray J, Mengersen K (2012). "Elicitator: A User-Friendly, Interactive Tool to Support Scenario-Based Elicitation of Expert Knowledge." In AH Perera, CA Drew, CJ Johnson (eds.), *Expert Knowledge and Its Application in Landscape Ecology*, pp. 39–67. Springer New York, New York, NY.

Low Choy S, Murray J, James A, Mengersen KL (2010). "Indirect elicitation from ecological experts: from methods and software to habitat modelling and rock-wallabies." In *The Oxford Handbook of Applied Bayesian Analysis*, pp. 511–544. Oxford University Press.

Low Choy S, O'Leary R, Mengersen K (2009). "Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models." *Ecology*, **90**, 265–277.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Florida USA.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Thisted RA (1988). *Elements of Statistical Computing*. Chapman and Hall.

Winkler RL (1967). "The assessment of prior distributions in Bayesian analysis." *Journal of the American Statistical Association*, **62**, 776–800.

Affiliation:

Geoffrey R. Hosack
Commonwealth Scientific and Industrial Research Organisation
Hobart, Tasmania, Australia 7001
E-mail: geoff.hosack@csiro.au
URL: http://people.csiro.au/H/G/Geoff-Hosack