A generalization of marginal likelihood

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington

Marginal likelihood

$$Y \sim p(y|\theta, \psi)$$

- θ is the parameter of interest
- ψ is the nuisance parameter, possibly high dimensional

Suppose we have a statistic $t()$ such that

$$p(t(y)|\theta, \psi) = p(t(y)|\theta)$$

Then

$$p(y|\theta, \psi) = p(t(y), y|\theta, \psi)$$

Marginal likelihood

$$Y \sim p(y|\theta, \psi)$$

- θ is the parameter of interest
- ψ is the nuisance parameter, possibly high dimensional

Suppose we have a statistic $t()$ such that

$$p(t(y)|\theta, \psi) = p(t(y)|\theta)$$

Then

$$\begin{aligned} p(y|\theta, \psi) &= p(t(y), y|\theta, \psi) \\ &= p(t(y)|\theta, \psi) \times p(y|t(y), \theta, \psi) \end{aligned}$$

Marginal likelihood

$$Y \sim p(y|\theta, \psi)$$

- θ is the parameter of interest
- ψ is the nuisance parameter, possibly high dimensional

Suppose we have a statistic $t()$ such that

$$p(t(y)|\theta, \psi) = p(t(y)|\theta)$$

Then

$$\begin{aligned} p(y|\theta, \psi) &= p(t(y), y|\theta, \psi) \\ &= p(t(y)|\theta, \psi) \times p(y|t(y), \theta, \psi) \\ &= p(t(y)|\theta) \times p(y|t(y), \theta, \psi) \end{aligned}$$

A marginal likelihood estimate of θ can be obtained from $p(t(y)|\theta)$. Specification or estimation of ψ is not necessary.

In this talk I will discuss a generalization of marginal likelihood.

Outline

Multivariate data

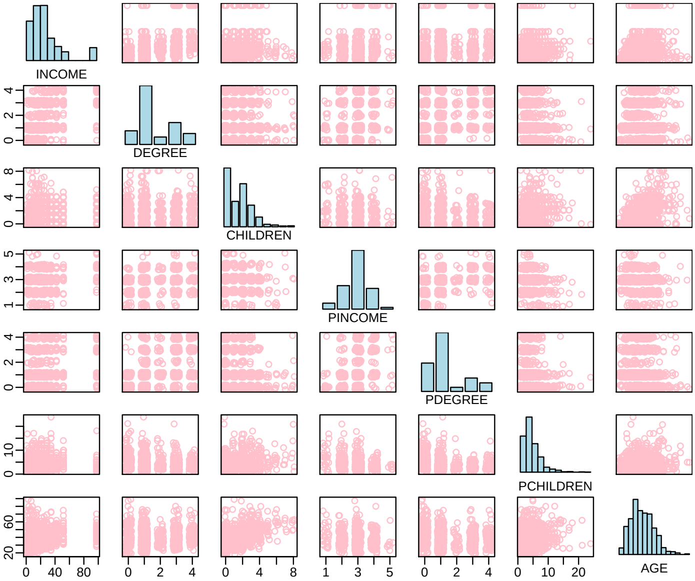
Survey data often yield multivariate data of varied types.

Hypothetical survey data: A vector of responses $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})$ for each person i in a sample of survey respondents, $i \in \{1, \ldots, n\}$.

- $y_{i,1}=$ income
- $y_{i,2}=$ education level
- $y_{i,3}=$ number of children
- $y_{i,4}=$ age
- $y_{i,5}=$ attitude (Likert scale)

A mix of continuous and discrete ordinal data.

GSS data


INCOME
DEGREE
CHILDREN
PINCOME
PDEGREE
PCHILDREN
AGE
0 40 80
0 2 4
0 4 8
1 3 5
0 2 4
0 10 20
20 60
0 10
0 4 8
1 3 5

Measures of association

Often of interest are the potential associations among these variables.

"Pearson's ρ": Measures the linear association between two data vectors, or more precisely, the angle between the data vectors:

$$\hat{\rho} = \frac{\sum(y_{i,1} - \bar{y}_{\cdot,1})(y_{i,2} - \bar{y}_{\cdot,2})}{\sqrt{\sum(y_{i,1} - \bar{y}_{\cdot,1})^2 \sum(y_{i,2} - \bar{y}_{\cdot,2})^2}}$$

"Spearman's ρ": Let $r_{i,j}$ be the rank of $y_{i,j}$ among responses $\{y_{1,j}, \ldots, y_{n,j}\}$, $i = \{1, \ldots, n\}$, $j \in \{1, 2\}$.

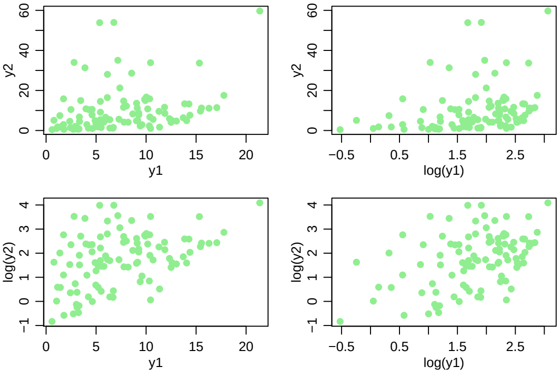
$$\hat{\rho} = \mathrm{Cor}[(r_{1,1}, \ldots, r_{n,1}), (r_{1,2}, \ldots, r_{n,2})]$$

"Kendall's τ": $(y_{i,1}, y_{i,2})$ and $(y_{j,1}, y_{j,2})$ are a concordant pair if $(y_{i,1} - y_{j,1}) \times (y_{i,2} - y_{j,2}) > 0$, otherwise they are discordant.

$$\hat{\tau} = \frac{1}{\binom{n}{2}}(c - d)$$

All are between -1 and +1. The latter two are invariant to monotone transformations, and so are "scale free". The moment correlation is not.

Monotone transformations



variables	moment	rank	concordance
y_1, y_2	.28	.39	.27
$\log y_1, y_2$	.26	.39	.27
$y_1, \log y_2$	.42	.39	.27
$\log y_1, \log y_2$	.44	.39	.27

Conditional models

Interest is typically in the conditional relationship between pairs of variables, accounting for heterogeneity in other variables of less interest. Standard bivariate rank-based methods are inappropriate.

Model 1

$$\mathrm{INC}_i = \beta_0 + \beta_1\mathrm{CHILD}_i + \beta_2\mathrm{DEG}_i + \beta_3\mathrm{AGE}_i + \beta_4\mathrm{PCHILD}_i + \beta_5\mathrm{PINC}_i + \beta_6\mathrm{PDEG}_i + \epsilon_i$$

p-value for β_1 is 0.11: "not strong evidence" that $\beta_1 \neq 0$

Model 2

$$\mathrm{CHILD}_i \sim \mathrm{Pois}(\exp\{\beta_0+\beta_1\mathrm{INC}_i+\beta_2\mathrm{DEG}_i+\beta_3\mathrm{AGE}_i+\beta_4\mathrm{PCHILD}_i+\beta_5\mathrm{PINC}_i+\beta_6\mathrm{PDEG}_i\})$$

p-value for β_1 is 0.01: "strong evidence" that $\beta_1 \neq 0$.

	Predictor						
Response	INC	CHILD	DEG	AGE	PCHILD	PINC	PDEG
INC	NA	**1.10 (.11)**	7.03 ($<$.01)	.34 ($<$.01)	4.07 ($<$.01)	.28 (.41)	1.40 (.12)
CHILD	**.01 (.01)**	NA	-.07 (.06)	.04 ($<$.01)	-.06 (.20)	.02 (.08)	-.05 (.20)

Inverse normal model

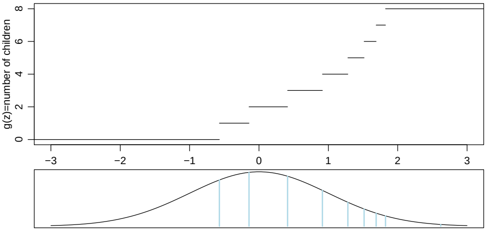
One possibility would be to transform the data to have normal marginals, then fit a multivariate normal model. This cannot be done for discrete data, but such data can be viewed as a function of normal data.

If F is a distribution there exists a nondecreasing function g such that

1. if $Z \sim$ normal(0,1),
2. and $Y = g(Z)$,

then $Y \sim F$.

If F is continuous then $g(z) = F^{-1}(\Phi(z))$, g^{-1} is a function and $g^{-1}(Y)$ is standard normal. If F is not continuous then g^{-1} maps to a set (this includes probit models, for example).



Multivariate normal copula model

This idea motivates the following "latent variable" model:

$$\begin{aligned}(Z_1, \ldots, Z_p) &\sim \text{multivariate normal}(\mathbf{0}, \Sigma) \\ (Y_1, \ldots, Y_p) &= (g_1(Z_1), \ldots, g_p(Z_p))\end{aligned}$$

Σ parameterizes the dependence, $g_1, \ldots, g_p$ the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

Estimation strategies:

- estimation of Σ conditional on plug-in estimates of $g_1, \ldots, g_p$; (procedures for continuous data gives inconsistent results for discrete data)
- joint estimation of Σ and $g_1, \ldots, g_p$;
  (parametric models of g too simple, nonparametric too complex)
- marginal likelihood estimation.
  (how would that work?)

Rank likelihood

Semiparametric Gaussian copula model:

$$\begin{aligned} \mathbf{Z}_1, \ldots, \mathbf{Z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{\Sigma}) \\ Y_{i,j} &= g_j(Z_{i,j}) \end{aligned}$$

- Σ is the parameter of interest
- $g_1, \ldots, g_p$ are high-dimensional nuisance parameters

For continuous data, let $r_{i,j} =$ rank of $y_{i,j}$ among $y_{1,j}, \ldots, y_{n,j}$. Then

$$p(\mathbf{y}|\mathbf{\Sigma}, \mathbf{g}) = p(\mathbf{r}, \mathbf{y}|\mathbf{\Sigma}, \mathbf{g})$$

Rank likelihood

Semiparametric Gaussian copula model:

$$\begin{aligned} \mathbf{Z}_1, \ldots, \mathbf{Z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{\Sigma}) \\ Y_{i,j} &= g_j(Z_{i,j}) \end{aligned}$$

- Σ is the parameter of interest
- $g_1, \ldots, g_p$ are high-dimensional nuisance parameters

For continuous data, let $r_{i,j}$ = rank of $y_{i,j}$ among $y_{1,j}, \ldots, y_{n,j}$. Then

$$\begin{aligned} p(\mathbf{y}|\mathbf{\Sigma}, \mathbf{g}) &= p(\mathbf{r}, \mathbf{y}|\mathbf{\Sigma}, \mathbf{g}) \\ &= p(\mathbf{r}|\mathbf{\Sigma}, \mathbf{g}) \times p(\mathbf{y}|\mathbf{r}, \mathbf{\Sigma}, \mathbf{g}) \end{aligned}$$

Rank likelihood

Semiparametric Gaussian copula model:

$$\begin{aligned} \mathbf{Z}_1, \ldots, \mathbf{Z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \boldsymbol{\Sigma}) \\ Y_{i,j} &= g_j(Z_{i,j}) \end{aligned}$$

- Σ is the parameter of interest
- $g_1, \ldots, g_p$ are high-dimensional nuisance parameters

For continuous data, let $r_{i,j}$ = rank of $y_{i,j}$ among $y_{1,j}, \ldots, y_{n,j}$. Then

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) &= p(\mathbf{r}, \mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) \\ &= p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g}) \times p(\mathbf{y}|\mathbf{r}, \boldsymbol{\Sigma}, \mathbf{g}) \\ &= p(\mathbf{r}|\boldsymbol{\Sigma}) \times p(\mathbf{y}|\mathbf{r}, \boldsymbol{\Sigma}, \mathbf{g}) \end{aligned}$$

Will this work for discrete data?

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g})$ depends on $\mathbf{g}$.

So the rank likelihood depends on $\mathbf{g}$.

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g})$ depends on $\mathbf{g}$.

So the rank likelihood depends on $\mathbf{g}$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $\mathbf{Y} = \mathbf{y}$ we do know

$$\mathbf{Z} \in A(\mathbf{y}) = \{\mathbf{z} : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g})$ depends on $\mathbf{g}$.

So the rank likelihood depends on $\mathbf{g}$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $\mathbf{Y} = \mathbf{y}$ we do know

$$\mathbf{Z} \in A(\mathbf{y}) = \{\mathbf{z} : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

We can construct the following marginal likelihood:

$$p(\mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) \quad = \quad p(\mathbf{Z} \in A(\mathbf{y}), \mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g})$$

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g})$ depends on $\mathbf{g}$.

So the rank likelihood depends on $\mathbf{g}$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $\mathbf{Y} = \mathbf{y}$ we do know

$$\mathbf{Z} \in A(\mathbf{y}) = \{\mathbf{z} : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

We can construct the following marginal likelihood:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) &= p(\mathbf{Z} \in A(\mathbf{y}), \mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) \\ &= \Pr(\mathbf{Z} \in A(\mathbf{y})|\boldsymbol{\Sigma}, \mathbf{g}) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \boldsymbol{\Sigma}, \mathbf{g}) \end{aligned}$$

Extending the rank likelihood

If g_j is not strictly increasing then

- variable j has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(\mathbf{r}|\boldsymbol{\Sigma}, \mathbf{g})$ depends on $\mathbf{g}$.

So the rank likelihood depends on $\mathbf{g}$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $\mathbf{Y} = \mathbf{y}$ we do know

$$\mathbf{Z} \in A(\mathbf{y}) = \{\mathbf{z} : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

We can construct the following marginal likelihood:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) &= p(\mathbf{Z} \in A(\mathbf{y}), \mathbf{y}|\boldsymbol{\Sigma}, \mathbf{g}) \\ &= \Pr(\mathbf{Z} \in A(\mathbf{y})|\boldsymbol{\Sigma}, \mathbf{g}) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \boldsymbol{\Sigma}, \mathbf{g}) \\ &= \Pr(\mathbf{Z} \in A(\mathbf{y})|\boldsymbol{\Sigma}) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \boldsymbol{\Sigma}, \mathbf{g}) \end{aligned}$$

$$\Pr(\mathbf{Z} \in A(\mathbf{y})|\Sigma) = \int_{A(\mathbf{y})} \prod p(\mathbf{z}_i|\boldsymbol{\Sigma}) \, d\mathbf{z}_i$$

If g_j's are continuous, then $\Pr(\mathbf{Z} \in A(\mathbf{y})|\Sigma) = \Pr(\mathbf{R} = \mathbf{r}|\Sigma)$.

Estimation

Bayesian estimates are easy to obtain.

Given a prior distribution $p(\boldsymbol{\Sigma})$, we iterate the following steps:

1. for each i, j, sample $Z_{i,j} \sim p(Z_{i,j} | \boldsymbol{\Sigma}, \mathbf{Z}_{-(i,j)}, \mathbf{Z} \in A(\mathbf{y}))$,
2. sample $\boldsymbol{\Sigma} \sim p(\boldsymbol{\Sigma} | \mathbf{Z}, \mathbf{Z} \in A(\mathbf{y})) = p(\boldsymbol{\Sigma} | \mathbf{Z})$.

This generates a Markov chain $\{\boldsymbol{\Sigma}^{(1)}, \boldsymbol{\Sigma}^{(2)}, \ldots\}$ such that

$$\boldsymbol{\Sigma}^{(s)} \xrightarrow{d} p(\boldsymbol{\Sigma} | \mathbf{Z} \in A(\mathbf{y})).$$

The actual R-code

Given `{Z,S}` and `{Ranks,n,p,S0,n0}`:

```
#### update S
S<-solve(rwish(solve(S0*n0+t(Z)%*%Z),n0+n))
####

#### update Z
for(j in 1:p) {

   Sjc<- S[j,-j]%*%solve(S[-j,-j])
   sdj<- sqrt( S[j,j] -S[j,-j]%*%solve(S[-j,-j])%*%S[-j,j]  )
   muj<- Z[,-j]%*%t(Sjc)

   for(r in unique(Ranks[,j])){

      ir<- (1:n)[Ranks[,j]==r & !is.na(Ranks[,j])]
      lb<-suppressWarnings(max( Z[ Ranks[,j]==r-1,j],na.rm=TRUE ))
      ub<-suppressWarnings(min( Z[ Ranks[,j]==r+1,j],na.rm=TRUE ))
      Z[ir,j]<-qnorm(runif(length(ir),
               pnorm(lb,muj[ir],sdj),pnorm(ub,muj[ir],sdj)),muj[ir],sdj)
                              }

   ir<-(1:n)[is.na(Ranks[,j])]
   Z[ir,j]<-rnorm(length(ir),muj[ir],sdj)
            }
####
```

GSS Example

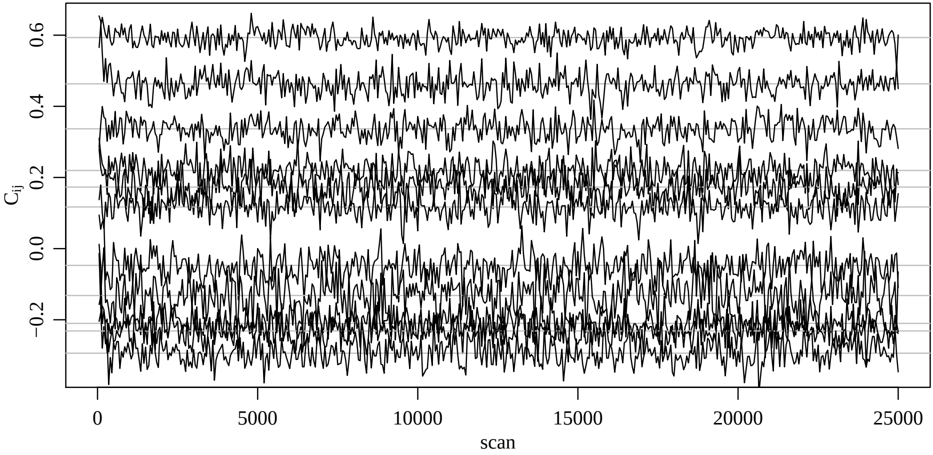
Data on 1002 male respondents to the 1994 GSS.

- `INC` : income of respondent
- `DEG` : highest degree obtained
- `CHILD` : number of children
- `PINC` : income category of parents
- `PDEG` : maximum of mother's and father's highest degree
- `PCHILD` : number of siblings plus one
- `AGE` : age in years

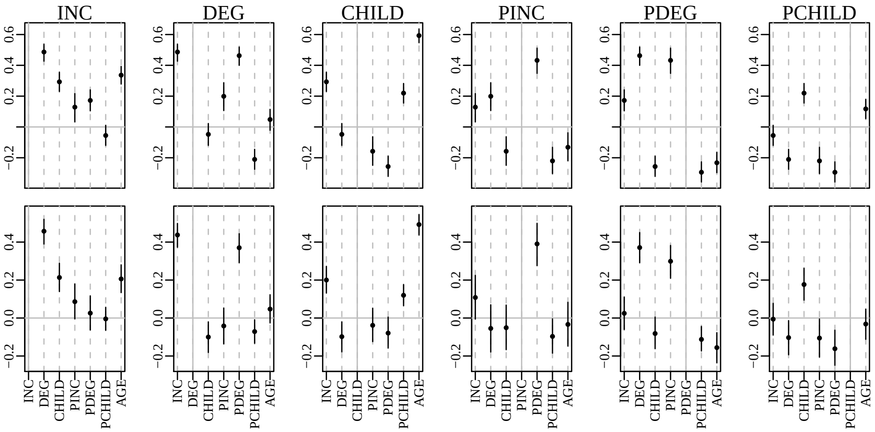
Using MCMC integration, we estimate

- Σ, the correlation matrix, and
- $\Sigma_{[j,-j]}\Sigma^{-1}_{[-j,-j]}$, the regression coefficients.

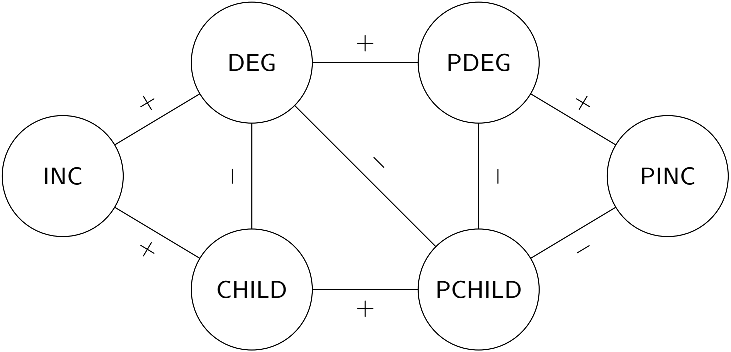
MCMC diagnostics



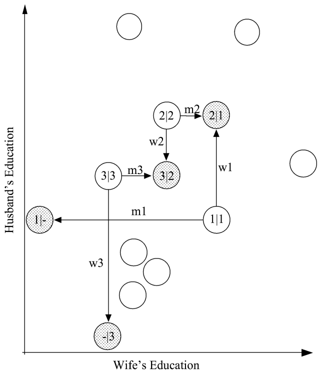
Correlations and regressions



Correlations and regressions



Two-sided matching



What characteristics do men and women prefer in their marriage partners?

- $x_1, \ldots, x_n$ are characteristics of females
- $y_1, \ldots, y_m$ are characteristics of males

- $h_j =$ index of husband of woman j, $h_j = 0$ if she is single
- $w_i =$ index of wife of man i, $w_i = 0$ if he is single

Can we ascertain preferences for characteristics from these data?

We treat characteristics $\{\mathbf{X}, \mathbf{Y}\}$ as fixed and the matching $\{\mathbf{w}, \mathbf{h}\}$ as random.

Assumptions about the matching process

- $U_{i,j}$ = man i's utility for woman j, $U_{i,0}$ = utility for being single
- $V_{j,i}$ = woman j's utility for man i, $V_{j,0}$ = utility for being single

Assumptions about the matching process

- $U_{i,j}$ = man i's utility for woman j, $U_{i,0}$ = utility for being single
- $V_{j,i}$ = woman j's utility for man i, $V_{j,0}$ = utility for being single

The matching process is that

1. members of the population meet,
2. make proposals to and marry each other,
3. the resulting matching is observed.

Assumptions about the matching process

- $U_{i,j}$ = man i's utility for woman j, $U_{i,0}$ = utility for being single
- $V_{j,i}$ = woman j's utility for man i, $V_{j,0}$ = utility for being single

The matching process is that

1. members of the population meet,
2. make proposals to and marry each other,
3. the resulting matching is observed.

It is assumed that the matching is *stable*, meaning

- not that it is unchanging over time, but that
- matches are voluntary, so that

$$\mathbf{U} \in \{\mathbf{u} : u_{i,w_i} > u_{i,j} \ \forall j : v_{j,i} > v_{j,h_j}\}$$

$$\mathbf{V} \in \{\mathbf{v} : v_{j,h_j} > v_{j,i} \ \forall i : u_{i,j} > u_{i,w_i}\}$$

A parametric marriage model

Goal: relate observed characteristics $\{\mathbf{X}, \mathbf{Y}\}$ to utilities $\{\mathbf{U}, \mathbf{V}\}$.

1. utilities are generated: $\mathbf{U} \sim p(\mathbf{u}|\boldsymbol{\alpha}, \mathbf{X})$, $\mathbf{V} \sim p(\mathbf{v}|\boldsymbol{\beta}, \mathbf{Y})$

A parametric marriage model

Goal: relate observed characteristics $\{\mathbf{X}, \mathbf{Y}\}$ to utilities $\{\mathbf{U}, \mathbf{V}\}$.

1. utilities are generated: $\mathbf{U} \sim p(\mathbf{u}|\alpha, \mathbf{X})$, $\mathbf{V} \sim p(\mathbf{v}|\beta, \mathbf{Y})$
2. marriages result from an unknown matching process: $\{\mathbf{h}, \mathbf{w}\} = g(\mathbf{U}, \mathbf{V})$

A parametric marriage model

Goal: relate observed characteristics $\{\mathbf{X}, \mathbf{Y}\}$ to utilities $\{\mathbf{U}, \mathbf{V}\}$.

1. utilities are generated: $\mathbf{U} \sim p(\mathbf{u}|\alpha, \mathbf{X})$, $\mathbf{V} \sim p(\mathbf{v}|\beta, \mathbf{Y})$
2. marriages result from an unknown matching process: $\{\mathbf{h}, \mathbf{w}\} = g(\mathbf{U}, \mathbf{V})$
3. the characteristics $\{\mathbf{X}, \mathbf{Y}\}$ and the matching $\{\mathbf{h}, \mathbf{w}\}$ are observed.

A parametric marriage model

Goal: relate observed characteristics $\{\mathbf{X}, \mathbf{Y}\}$ to utilities $\{\mathbf{U}, \mathbf{V}\}$.

1. utilities are generated: $\mathbf{U} \sim p(\mathbf{u}|\alpha, \mathbf{X})$, $\mathbf{V} \sim p(\mathbf{v}|\beta, \mathbf{Y})$
2. marriages result from an unknown matching process: $\{\mathbf{h}, \mathbf{w}\} = g(\mathbf{U}, \mathbf{V})$
3. the characteristics $\{\mathbf{X}, \mathbf{Y}\}$ and the matching $\{\mathbf{h}, \mathbf{w}\}$ are observed.

It is assumed that the matching we observe is stable:

$$\mathbf{U} \in \{u_{i,j} : u_{i,w_i} > u_{i,j} \ \forall j : v_{j,i} > v_{j,h_j}\}$$

$$\mathbf{V} \in \{v_{j,i} : v_{j,h_j} > v_{j,i} \ \forall i : u_{i,j} > u_{i,w_i}\}$$

Thus observing the matching $\{\mathbf{h}, \mathbf{w}\}$ implies that

$$\begin{aligned} \{\mathbf{U}, \mathbf{V}\} &\in A(\{\mathbf{h}, \mathbf{w}\}) \\ \mathbf{Z} &\in A(\mathbf{y}) \end{aligned}$$

Marginal likelihood estimation

- $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, the parameters of interest
- $\mathbf{Z} = \{\mathbf{U}, \mathbf{V}\}$, the unobserved utilities
- $\mathbf{y} = \{\mathbf{h}, \mathbf{w}\} = g(\mathbf{Z})$, the observed matching.

Observing $\mathbf{Y} = \mathbf{y}$ tells us

1. $\mathbf{y}$ is a stable matching, so $\mathbf{Z} \in A(\mathbf{y})$
2. $\mathbf{y}$ is the actual observed matching resulting from a marriage process.

Using information in 2 requires estimation/specification of the marriage process.

Using information in 1 does not.

$$p(\mathbf{y}|\theta, g) \quad = \quad p(\mathbf{Z} \in A(\mathbf{Y}), \mathbf{y}|\theta, g)$$

Marginal likelihood estimation

- $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, the parameters of interest
- $\mathbf{Z} = \{\mathbf{U}, \mathbf{V}\}$, the unobserved utilities
- $\mathbf{y} = \{\mathbf{h}, \mathbf{w}\} = g(\mathbf{Z})$, the observed matching.

Observing $\mathbf{Y} = \mathbf{y}$ tells us

1. $\mathbf{y}$ is a stable matching, so $\mathbf{Z} \in A(\mathbf{y})$
2. $\mathbf{y}$ is the actual observed matching resulting from a marriage process.

Using information in 2 requires estimation/specification of the marriage process.

Using information in 1 does not.

$$
\begin{aligned}
p(\mathbf{y}|\theta, g) &= p(\mathbf{Z} \in A(\mathbf{Y}), \mathbf{y}|\theta, g) \\
&= \Pr(\mathbf{Z} \in A(\mathbf{y})|\theta, g) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \theta, g)
\end{aligned}
$$

Marginal likelihood estimation

- $\theta = \{\alpha, \beta\}$, the parameters of interest
- $\mathbf{Z} = \{\mathbf{U}, \mathbf{V}\}$, the unobserved utilities
- $\mathbf{y} = \{\mathbf{h}, \mathbf{w}\} = g(\mathbf{Z})$, the observed matching.

Observing $\mathbf{Y} = \mathbf{y}$ tells us

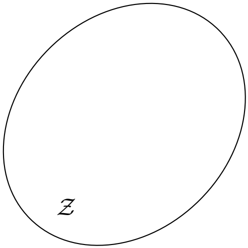
1. $\mathbf{y}$ is a stable matching, so $\mathbf{Z} \in A(\mathbf{y})$
2. $\mathbf{y}$ is the actual observed matching resulting from a marriage process.

Using information in 2 requires estimation/specification of the marriage process.

Using information in 1 does not.

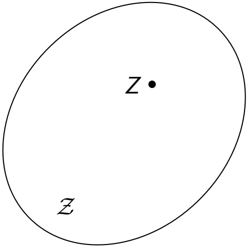
$$\begin{aligned} p(\mathbf{y}|\theta, g) &= p(\mathbf{Z} \in A(\mathbf{Y}), \mathbf{y}|\theta, g) \\ &= \Pr(\mathbf{Z} \in A(\mathbf{y})|\theta, g) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \theta, g) \\ &= \Pr(\mathbf{Z} \in A(\mathbf{y})|\theta) \times p(\mathbf{y}|\mathbf{Z} \in A(\mathbf{y}), \theta, g) \end{aligned}$$

The general transformation model



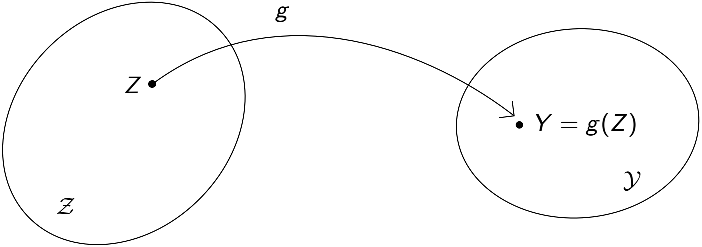
The general transformation model

$Z \sim p(z|\theta)$



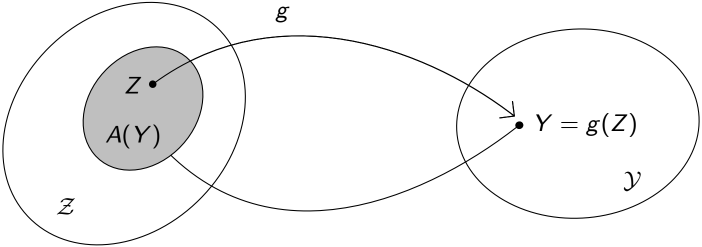
The general transformation model

$Z \sim p(z|\theta)$ $\quad\quad$ $Y = g(Z)$



The general transformation model

$Z \sim p(z|\theta)$ $\quad\quad$ $Y = g(Z)$ $\quad\quad$ $Z \in A(Y)$



Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$\begin{aligned} g^{-1}(y) &\subset A(y)\ \forall y, g, \quad \text{or equivalently,} \\ z &\in A(g(z))\ \forall z, g, \end{aligned}$$

Then $\Pr(Z \in A(Y)|\theta, g) = 1$, so

$$\Pr(Y = y|\theta, g) = \Pr(Z \in A(Y), Y = y|\theta, g)$$

Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$\begin{aligned} g^{-1}(y) &\subset A(y)\ \forall y, g, \quad \text{or equivalently,} \\ z &\in A(g(z))\ \forall z, g, \end{aligned}$$

Then $\Pr(Z \in A(Y)|\theta, g) = 1$, so

$$\begin{aligned} \Pr(Y = y|\theta, g) &= \Pr(Z \in A(Y), Y = y|\theta, g) \\ &= \Pr(Z \in A(y), Y = y|\theta, g) \end{aligned}$$

Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$\begin{aligned} g^{-1}(y) &\subset A(y) \ \forall y, g, \quad \text{or equivalently,} \\ z &\in A(g(z)) \ \forall z, g, \end{aligned}$$

Then $\Pr(Z \in A(Y)|\theta, g) = 1$, so

$$\begin{aligned} \Pr(Y = y|\theta, g) &= \Pr(Z \in A(Y), Y = y|\theta, g) \\ &= \Pr(Z \in A(y), Y = y|\theta, g) \\ &= \Pr(Z \in A(y)|\theta, g) \times \Pr(Y = y|Z \in A(y), \theta, g) \end{aligned}$$

Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \to \sigma(\mathcal{Z})$ such that

$$\begin{aligned} g^{-1}(y) &\subset A(y)\ \forall y, g, \quad \text{or equivalently,} \\ z &\in A(g(z))\ \forall z, g, \end{aligned}$$

Then $\Pr(Z \in A(Y)|\theta, g) = 1$, so

$$\begin{aligned} \Pr(Y = y|\theta, g) &= \Pr(Z \in A(Y), Y = y|\theta, g) \\ &= \Pr(Z \in A(y), Y = y|\theta, g) \\ &= \Pr(Z \in A(y)|\theta, g) \times \Pr(Y = y|Z \in A(y), \theta, g) \\ &= \Pr(Z \in A(y)|\theta) \times \Pr(Y = y|Z \in A(y), \theta, g) \end{aligned}$$

Idea: estimate θ using only the marginal likelihood $\Pr(Z \in A(y)|\theta)$

Most informative sets

Which set-valued function is most informative?

Consider the class of functions

$$\mathcal{A} = \{A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z}) \ , \ z \in A(g(z)) \ \ \forall z, g\}.$$

A marginal set likelihood could be based on any element of $\mathcal{A}$.
Intuitively, we want to use the “smallest” such function $\tilde{A}()$.

Most informative sets

Which set-valued function is most informative?

Consider the class of functions

$$\mathcal{A} = \{A() : \mathcal{Y} \to \sigma(\mathcal{Z}) \ , \ z \in A(g(z)) \ \ \forall z, g\}.$$

A marginal set likelihood could be based on any element of $\mathcal{A}$.
Intuitively, we want to use the "smallest" such function $\tilde{A}()$.

Lemma: For each y, let $\tilde{A}(y) = \cap_{\mathcal{A}} A(y)$. Then

- $\tilde{A} \in \mathcal{A}$
- $\tilde{A}(y) = \{z : y = g(z) \text{ for some } g\}$.

Most informative sets

Which set-valued function is most informative?

Consider the class of functions

$$\mathcal{A} = \{A() : \mathcal{Y} \to \sigma(\mathcal{Z}) \ , \ z \in A(g(z)) \ \ \forall z, g\}.$$

A marginal set likelihood could be based on any element of $\mathcal{A}$. Intuitively, we want to use the "smallest" such function $\tilde{A}()$.

Lemma: For each y, let $\tilde{A}(y) = \cap_{\mathcal{A}} A(y)$. Then

- $\tilde{A} \in \mathcal{A}$
- $\tilde{A}(y) = \{z : y = g(z) \text{ for some } g\}$.

Lemma:

- for the copula model, $\tilde{A}(y) = \{\mathbf{z} : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$
- for the marriage model, $\tilde{A}(y) = \{\mathbf{u}, \mathbf{v} : y \text{ is a stable match}\}$

What is a statistic?

Any statistic can be defined in terms of a set function:

$$A(y) = \{z : t(g(z)) = t(y)\}$$

Can be used for marginal likelihood if $\Pr(Z \in A(y)|\theta, g) = \Pr(Z \in A(y)|\theta)$.

What is a statistic?

Any statistic can be defined in terms of a set function:

$$A(y) = \{z : t(g(z)) = t(y)\}$$

Can be used for marginal likelihood if $\Pr(Z \in A(y)|\theta, g) = \Pr(Z \in A(y)|\theta)$.

Example (rank likelihood for regression):

$$\begin{aligned} Z_i &= \beta x_i + \epsilon_i, \;\; Y_i = g(Z_i), \;\; g \text{ nondecreasing} \\ R(\mathbf{y}) &= \text{ranks}(y_1, \ldots, y_n) \\ A(\mathbf{y}) &= \{\mathbf{z} : z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\} \end{aligned}$$

What is a statistic?

Any statistic can be defined in terms of a set function:

$$A(y) = \{z : t(g(z)) = t(y)\}$$

Can be used for marginal likelihood if $\Pr(Z \in A(y)|\theta, g) = \Pr(Z \in A(y)|\theta)$.

Example (rank likelihood for regression):

$$\begin{aligned} Z_i &= \beta x_i + \epsilon_i,\ \ Y_i = g(Z_i),\ \ g \text{ nondecreasing} \\ R(\mathbf{y}) &= \text{ranks}(y_1, \ldots, y_n) \\ A(\mathbf{y}) &= \{\mathbf{z} : z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\} \end{aligned}$$

If g is strictly increasing,

$$\mathbf{Z} \in A(\mathbf{y}) \Leftrightarrow R(g(\mathbf{Z})) = R(\mathbf{y})\ \forall g \Leftrightarrow A(\mathbf{Y}) = A(\mathbf{y})$$

What is a statistic?

Any statistic can be defined in terms of a set function:

$$A(y) = \{z : t(g(z)) = t(y)\}$$

Can be used for marginal likelihood if $\Pr(Z \in A(y)|\theta, g) = \Pr(Z \in A(y)|\theta)$.

Example (rank likelihood for regression):

$$\begin{aligned} Z_i &= \beta x_i + \epsilon_i, \;\; Y_i = g(Z_i), \;\; g \text{ nondecreasing} \\ R(\mathbf{y}) &= \text{ranks}(y_1, \ldots, y_n) \\ A(\mathbf{y}) &= \{\mathbf{z} : z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\} \end{aligned}$$

If g is strictly increasing,

$$\mathbf{Z} \in A(\mathbf{y}) \Leftrightarrow R(g(\mathbf{Z})) = R(\mathbf{y}) \; \forall g \Leftrightarrow A(\mathbf{Y}) = A(\mathbf{y})$$

If g is not strictly increasing, then

$$A(\mathbf{Y}) = A(\mathbf{y}) \Rightarrow \mathbf{Z} \in A(\mathbf{y}) \;\; \text{but} \;\; \mathbf{Z} \in A(\mathbf{y}) \not\Rightarrow A(\mathbf{Y}) = A(\mathbf{y})$$

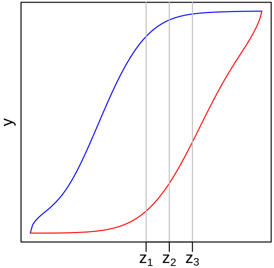
- $\{\mathbf{z} : R(g(\mathbf{z})) = R(\mathbf{y})\} \subset A(\mathbf{y})$
- $\Pr(R(g(\mathbf{Z})) = R(\mathbf{y})|\theta, g)$ depends on g
- $\Pr(\mathbf{Z} \in A(\mathbf{y})|\theta, g)$ does not depend on g

Example: rank likelihood

$$A(\mathbf{y}) = \{z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\}$$

Suppose $\mathbf{Z} \in a = \{\mathbf{z} : z_1 < z_2 < z_3\}$

g strictly increasing



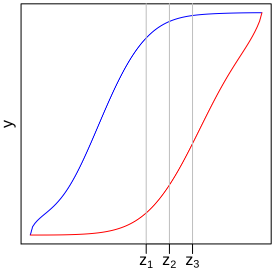
$$\begin{aligned} A(\mathbf{Y}) &= \{z_1 < z_2 < z_3\} \ \forall g \\ \mathbf{Z} \in a &\Rightarrow A(\mathbf{Y}) = a \end{aligned}$$

Example: rank likelihood

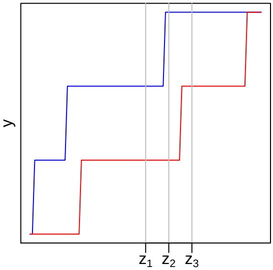
$$A(\mathbf{y}) = \{z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\}$$

Suppose $\mathbf{Z} \in a = \{\mathbf{z} : z_1 < z_2 < z_3\}$

g strictly increasing



g not strictly increasing



$$\begin{aligned} A(\mathbf{Y}) &= \{z_1 < z_2 < z_3\} \; \forall g \\ \mathbf{Z} \in a &\Rightarrow A(\mathbf{Y}) = a \end{aligned}$$

$$\begin{aligned} A(\mathbf{Y}) &= \{z_1 < (z_2, z_3)\} \text{ if } g = g_1 \\ A(\mathbf{Y}) &= \{(z_1, z_2) < z_3\} \text{ if } g = g_2 \\ \mathbf{Z} \in a &\not\Rightarrow A(\mathbf{Y}) = a \end{aligned}$$

What is a statistic?

Can the event $\{Z \in A(y)\}$ be written as $\{t(g(Z)) = t(y)\}$ for some statistic t?

What is a statistic?

Can the event $\{Z \in A(y)\}$ be written as $\{t(g(Z)) = t(y)\}$ for some statistic t?

If a marginal likelihood is based on a statistic $t(y)$, then

$$\begin{aligned} \Pr(t(Y) = t(y)|\theta) &= \Pr(Z \in A(y)|\theta) \text{ where} \\ A(y) &\equiv \{z : t(g(z)) = t(y)\} \end{aligned}$$

What is a statistic?

Can the event $\{Z \in A(y)\}$ be written as $\{t(g(Z)) = t(y)\}$ for some statistic t?

If a marginal likelihood is based on a statistic $t(y)$, then

$$\begin{aligned} \Pr(t(Y) = t(y)|\theta) &= \Pr(Z \in A(y)|\theta) \ \text{ where} \\ A(y) &\equiv \{z : t(g(z)) = t(y)\} \end{aligned}$$

So for a statistic-based likelihood,

$$\begin{aligned} Z \in A(y) &\Leftrightarrow t(Y) \equiv t(g(Z)) = t(y) \\ Z \in A(y) &\Leftrightarrow A(Y) = A(y) \end{aligned}$$

What is a statistic?

Can the event $\{Z \in A(y)\}$ be written as $\{t(g(Z)) = t(y)\}$ for some statistic t?

If a marginal likelihood is based on a statistic $t(y)$, then

$$\begin{aligned} \Pr(t(Y) = t(y)|\theta) &= \Pr(Z \in A(y)|\theta) \text{ where} \\ A(y) &\equiv \{z : t(g(z)) = t(y)\} \end{aligned}$$

So for a statistic-based likelihood,

$$\begin{aligned} Z \in A(y) &\Leftrightarrow t(Y) \equiv t(g(Z)) = t(y) \\ Z \in A(y) &\Leftrightarrow A(Y) = A(y) \end{aligned}$$

But for some problems,

$$Z \in A(y) \not\Leftrightarrow A(Y) = A(y)$$

Not all set-based likelihoods can be expressed as statistic-based likelihoods.

Likelihood derivatives

Does the distinction matter?

Statistic-based likelihoods:

$$\mathrm{E}[\frac{d \log p(t|\theta)}{d\theta}|\theta] \quad = \quad \int \frac{p'(t|\theta)}{p(t|\theta)} p(t|\theta) dt$$

Likelihood derivatives

Does the distinction matter?

Statistic-based likelihoods:

$$\begin{aligned}
\mathrm{E}[\frac{d \log p(t|\theta)}{d\theta}|\theta] &= \int \frac{p'(t|\theta)}{p(t|\theta)} p(t|\theta) dt \\
&= \frac{d}{d\theta} \int p(t|\theta) dt = 0
\end{aligned}$$

Likelihood derivatives

Does the distinction matter?

Statistic-based likelihoods:

$$
\begin{aligned}
\mathrm{E}[\frac{d \log p(t|\theta)}{d\theta}|\theta] &= \int \frac{p'(t|\theta)}{p(t|\theta)} p(t|\theta) dt \\
&= \frac{d}{d\theta} \int p(t|\theta) dt = 0
\end{aligned}
$$

Set-based likelihoods:

$$
\mathrm{E}[\frac{d \log \Pr(Z \in A(y)|\theta)}{d\theta}|\theta] = \sum_a \frac{\Pr'(Z \in a|\theta)}{\Pr(Z \in a|\theta)} p(A(Y) = a|\theta, g)
$$

Likelihood derivatives

Does the distinction matter?

Statistic-based likelihoods:

$$\begin{aligned} \mathrm{E}[\frac{d \log p(t|\theta)}{d\theta}|\theta] &= \int \frac{p'(t|\theta)}{p(t|\theta)} p(t|\theta) dt \\ &= \frac{d}{d\theta} \int p(t|\theta) dt = 0 \end{aligned}$$

Set-based likelihoods:

$$\begin{aligned} \mathrm{E}[\frac{d \log \Pr(Z \in A(y)|\theta)}{d\theta}|\theta] &= \sum_a \frac{\Pr'(Z \in a|\theta)}{\Pr(Z \in a|\theta)} p(A(Y) = a|\theta, g) \\ &= \sum_a \Pr'(Z \in a|\theta) \Pr(A(Y) = a|\theta, g, Z \in a) = \ ? \end{aligned}$$

Proper by coverage

Let $\pi(\theta)$ be a prior, $L(\theta|y)$ some positive function and define

$$p_L(\theta|y) \propto \pi(\theta) \times L(\theta|y).$$

Proper by coverage

Let $\pi(\theta)$ be a prior, $L(\theta|y)$ some positive function and define

$$p_L(\theta|y) \propto \pi(\theta) \times L(\theta|y).$$

An α-level confidence set based on p_L is a set $C(y)$ such that

$$\int_{C(y)} p_L(\theta|y)\ d\theta = \alpha \ \ \forall y$$

so $C(y)$ has the property that

$$\text{if } \tilde{\theta} \sim p_L(\tilde{\theta}|y), \ \ \text{then } \Pr(\tilde{\theta} \in C(y)|y) = \alpha, \ \ \text{for every } y.$$

Proper by coverage

Let $\pi(\theta)$ be a prior, $L(\theta|y)$ some positive function and define

$$p_L(\theta|y) \propto \pi(\theta) \times L(\theta|y).$$

An α-level confidence set based on p_L is a set $C(y)$ such that

$$\int_{C(y)} p_L(\theta|y) \; d\theta = \alpha \;\; \forall y$$

so $C(y)$ has the property that

$$\text{if } \tilde{\theta} \sim p_L(\tilde{\theta}|y), \;\; \text{then } \Pr(\tilde{\theta} \in C(y)|y) = \alpha, \;\; \text{for every } y.$$

The "likelihood" function $L(\theta|\mathbf{y})$ is *proper by coverage* (Monahan and Boos, 1992) for a model $p(y|\theta)$ if

$$\text{when } \theta \sim \pi(\theta) \text{ and } Y \sim p(y|\theta), \;\; \text{then } \Pr(\theta \in C(Y)) = \alpha$$

Proper by coverage

If $L(\theta, g|y) = p(y|\theta, g)$ then $L(\theta, g|y)$ is proper by coverage

Proper by coverage

If $L(\theta, g|y) = p(y|\theta, g)$ then $L(\theta, g|y)$ is proper by coverage

If $L(\theta|y) = p(t|\theta)$ for $t = t(y)$ then $L(\theta|y)$ is proper by coverage

Proper by coverage

If $L(\theta, g|y) = p(y|\theta, g)$ then $L(\theta, g|y)$ is proper by coverage

If $L(\theta|y) = p(t|\theta)$ for $t = t(y)$ then $L(\theta|y)$ is proper by coverage

What about if $L(\theta|y) = \Pr(Z \in A(y)|\theta)$?

Proposition:

- For some priors on g, $\Pr(Z \in A(y)|\theta)$ will be proper by coverage, but
- For other priors on g, it won't be.

Proper by coverage

If $L(\theta, g|y) = p(y|\theta, g)$ then $L(\theta, g|y)$ is proper by coverage

If $L(\theta|y) = p(t|\theta)$ for $t = t(y)$ then $L(\theta|y)$ is proper by coverage

What about if $L(\theta|y) = \Pr(Z \in A(y)|\theta)$?

Proposition:

- For some priors on g, $\Pr(Z \in A(y)|\theta)$ will be proper by coverage, but
- For other priors on g, it won't be.

For rank regression, the set likelihood will be proper by coverage if $\pi(g)$ makes $p(A(g(Z)) = a|Z)$ uniform over possible sets a.

Summary

Observing $Y = y$ can tell us that some event $A(y)$ is true.

Summary

Observing $Y = y$ can tell us that some event $A(y)$ is true.

The probability that $A(y)$ is true might be independent of g.

- If so, $\{A(y) \text{ is true}\}$ can be used to construct a likelihood.

Summary

Observing $Y = y$ can tell us that some event $A(y)$ is true.

The probability that $A(y)$ is true might be independent of g.

- If so, $\{A(y)$ is true$\}$ can be used to construct a likelihood.

The fact that $A(y)$ is true may imply that we learn that it is true

- If it does, our likelihood is statistic-based.
- If it does not, then our likelihood is not statistic-based.

Summary

Observing $Y = y$ can tell us that some event $A(y)$ is true.

The probability that $A(y)$ is true might be independent of g.

- If so, $\{A(y) \text{ is true}\}$ can be used to construct a likelihood.

The fact that $A(y)$ is true may imply that we learn that it is true

- If it does, our likelihood is statistic-based.
- If it does not, then our likelihood is not statistic-based.

Questions:

- Are there other applications of set-based likelihoods?
- What are the general properties of set-based likelihoods?
  - asymptotics
  - Bayesian propriety
- How can one identify the optimal $A(y)$ in a given problem?