

Monte Carlo Conditional Analysis for Log-Linear and Logistic Models

Brian S. Caffo

Department of Biostatistics

Johns Hopkins University School of Public Health

bcaffo@jhsph.edu

References

- Caffo and Booth (1999). *An MCMC Algorithm for Approximating Exact Conditional Probabilities*. JCGS.
- Caffo and Booth (To Appear). *Monte Carlo and Markov Chain Monte Carlo Inference for Log-linear and Logistic Models*. SMMR.
- Booth and Butler (1998). *An importance sampling algorithm for exact conditional tests in log-linear models*. Biometrika.
- <http://www.biostat.jhsph.edu/~bcaffo/>

Fisher's exact test for 2×2 tables

Treatment	Successes	Failures	
A	X	$N_A - X$	$\text{Bin}(p_A, N_A)$
B	Y	$N_B - Y$	$\text{Bin}(p_B, N_B)$

- $\text{logit}(p_A) = \beta$
- $\text{logit}(p_B) = \beta + \lambda$

λ is the *parameter of interest*

β is the *nuisance parameter*

$H_o : \lambda = 0$ vs $H_a : \lambda \neq 0$

- h test statistic with observed value h_{obs}

$$P(h \geq h_{obs}; \beta)$$

Conditional Inference

- Eliminate β by conditioning on its sufficient statistic
- Preserves the nominal type I error rate unconditionally
- Many standard tests are conditional tests

Criticisms of Conditional Inference

- In the cases we consider today, the sufficient statistic for β is not ancillary for λ .

For example, in the two sample binomial example $X + Y$ is not ancillary for λ

- The conditional distributions can be very discrete or even degenerate
- Can lead to overly conservative tests
- Often hard/impossible to do because of the complexity or size of the support of the conditional distribution

Conditional analysis for log-linear models

- $\mathbf{y} = (y_1, \dots, y_n)^t \sim \text{Poisson}(\boldsymbol{\mu})$
- Alternative model: $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\lambda}$
- Null model: $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\lambda}_0$
- Sufficient statistics for $\boldsymbol{\beta}$ under H_0 are
 $\mathbf{s}_{\boldsymbol{\beta}} = \mathbf{X}^t \mathbf{y}$
- We can eliminate $\boldsymbol{\beta}$ from null distribution by conditioning on $\mathbf{s}_{\boldsymbol{\beta}}$

Conditional Distribution Under H_0

$$f(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{\prod_{i=1}^n y_i!} \exp(\boldsymbol{\beta}' \mathbf{s}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{y} - \mu_+)$$

$$f(\mathbf{s}_{\boldsymbol{\beta}}; \boldsymbol{\beta}) = c \exp(\boldsymbol{\beta}' \mathbf{s}_{\boldsymbol{\beta}} - \mu_+) \sum_{\mathbf{v} \in \Gamma} \exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{v})$$

$$f(\mathbf{y} | \mathbf{s}_{\boldsymbol{\beta}}) = \frac{1}{c \prod_{i=1}^n y_i!} \times \frac{\exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{y})}{\sum_{\mathbf{v} \in \Gamma} \exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{v})}$$

For $\mathbf{y} \in \Gamma = \{\mathbf{v} | \mathbf{X}' \mathbf{v} = \mathbf{s}_{\boldsymbol{\beta}}\}$

Γ is referred to as *the reference set*

Exact Conditional P-value

$$\sum_{\mathbf{y} \in \Gamma} \frac{I(h(\mathbf{y}) \geq h_{obs})}{\prod_{i=1}^n y_i!} \left(\sum_{\mathbf{y} \in \Gamma} \frac{1}{\prod_{i=1}^n y_i!} \right)^{-1}$$

Example

Pathologist A	Pathologist B				
	1	2	3	4	5
1	22	2	2	0	0
2	5	7	14	0	0
3	0	2	36	0	0
4	0	1	14	7	0
5	0	0	3	0	3

- Uniform Association model (Agresti 1990)

$$\log(\mu_{ij}) = \beta_0 + \beta_i^X + \beta_j^Y + \gamma_{ij}$$

- Sufficient statistics for the β 's are the margins of the table
- Sufficient statistic for γ is $\sum_{ij} y_{ij} ij$
- 12 billion tables with the same margins
- Only 34,000 tables with the same margins and $\sum_{ij} y_{ij} ij$

Caffo and Booth Algorithm

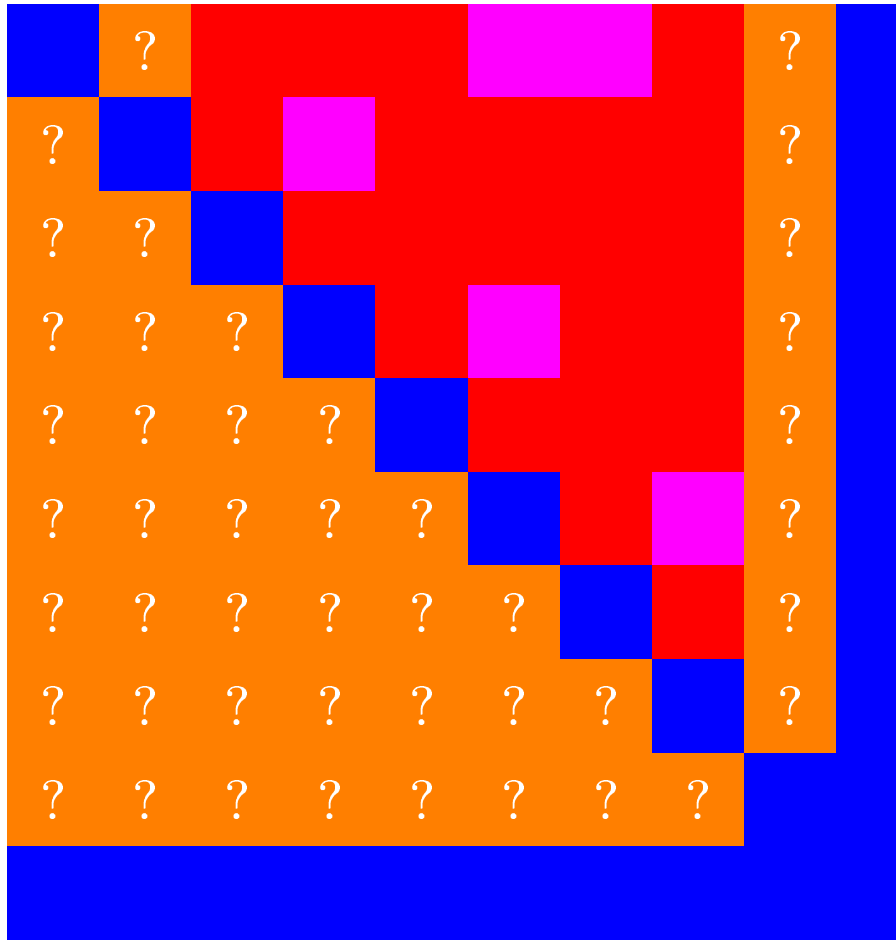
- Estimate conditional P-values by simulating samples from $f(\mathbf{y}|\mathbf{s}_\beta)$
- Modify Booth and Butler's normal candidate to perform "local updates" to increase the number of data sets and models that can be analyzed
- Create a Markov chain that updates a portion of the table while leaving the remainder fixed
- Provide this chain is irreducible and aperiodic, we can use the Metropolis/Hastings/Green algorithm to guarantee the correct invariant distribution

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	7	7	2	3	19
	F	2	8	3	7	20
	V	1	5	4	9	19
	A	2	8	9	14	33
		12	28	18	33	91

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	7	7	2	3	19
	F	2	8	3	7	20
	V	1	5	4	9	19
	A	2	8	9	14	33
		12	28	18	33	91

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	?	7	2	?	19
	F	2	8	3	?	20
	V	?	5	?	?	19
	A	?	?	?	?	33
		12	28	18	33	91

Quasi-symmetry model for a 9×9 table



\mathbf{y} is Poisson with mean vector $\boldsymbol{\mu}$

$\mathbf{s}_\beta = \mathbf{X}'\mathbf{y}$ Sufficient statistic (blue area)

\mathbf{y}_1 Free area given \mathbf{s}_β (red area)

$\mathbf{X}' = [\mathbf{X}'_1 \ \mathbf{X}'_2]$ where \mathbf{X}_2 is invertible

Constructing a candidate

- \mathbf{y} is approximately multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $D(\boldsymbol{\mu})$ ($D = \text{diagonal}$)

- Then it follows that

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{s}_\beta \end{bmatrix} = \begin{bmatrix} I & 0 \\ \mathbf{X}'_1 & \mathbf{X}'_2 \end{bmatrix} \mathbf{y}$$

is approximately multivariate normal

- $\mathbf{y}_1 | \mathbf{s}_\beta$ is then also approximately multivariate normal

- $P\mathbf{y}_1 | \mathbf{s}_\beta$ for permutation matrix P is also multivariate normal
- Then $P_1\mathbf{y}_1 | \mathbf{s}_\beta, P_2\mathbf{y}_1$ where $P' = [P'_1 P'_2]$ is also multivariate normal
- Generate from this approximation *sequentially*
“round as you go”
Easy to specify the probability of the candidate given the current state
- Solve for the **orange cells** (linear system)
- Automatically reject tables with negative entries

- Choose the number of cells in the **magenta area** as a binomial random with success probability that is treated as a tuning parameter.
- If we update few cells on average
 - The more valid tables with non-negative entries we obtain
 - The slower the mixing of the chain
- The probability generating the current table given the candidate can be calculated along with sequential means
- Use a t distribution instead of a normal
- Computing tricks ...

Benefits

- Allows for an arbitrary number of table entries to be updated at each iteration in a random order.
- t candidate attempts to model covariance structure of $\mathbf{y} | \mathbf{s}_\beta$.
- Algorithm (theoretically) works for any log linear model.
- Algorithm guaranteed to produce an aperiodic irreducible chain.

exactLoglinTest

```
> install.packages(exactLoglinTest)
> library(exactLoglinTest)
> data(pathologist.dat)
> mcx <- mcexact(y ~ factor(A) + factor(B) + I(A * B),
+               data = pathologist.dat)
> mcx
```

	deviance	Pearson
observed.stat	16.214	14.729
pvalue	0.044	0.128
mcse	0.014	0.030

```
> mcx <- update(mcx, nosim = 10000)
> mcx
```

	deviance	Pearson
observed.stat	16.2144	14.729
pvalue	0.0429	0.132
mcse	0.0075	0.011

```
> pchisq(c(16.21, 14.79), 15, lower.tail = FALSE)
[1] 0.37 0.47
```

Titanic data

Surv	Sex	Age	Class			
			Crew	First	Second	Third
no	F	Child	0	0	0	17
		Adult	3	4	13	89
	M	Child	0	0	0	35
		Adult	670	118	154	387
yes	F	Child	0	1	13	14
		Adult	20	140	80	76
	M	Child	0	5	11	13
		Adult	192	57	14	75

Source <http://www.cytel.com>

- Model: survival is binary with logit success probability depending on age, class and gender.
- Focus on testing existence of a gender effect

- It is straightforward to write the null model as a Poisson log-linear model

$$\log \mu_{ijkl} = \beta_l^1 + \beta_{il}^2 + \beta_{jl}^3 + \beta_{ijk}^4$$

- (ITC + AGE + CLASS) * SURV + AGE : CLASS : SEX
- Score statistic for the gender effect is $\sum_{ij} y_{ij}^2$.
- After 100,000 simulations, P-value estimate is very nearly 0

Czech Auto Workers Data

				B	1	0		
F	E	D	C	A	1	0	1	0
1	1	1	1		44	40	112	67
			0		129	145	12	23
		0	1		35	12	80	33
			0		109	67	7	9
	0	1	1		23	32	70	66
			0		50	80	7	13
		0	1		24	25	73	57
			0		51	63	7	16
0	1	1	1		5	7	21	9
			0		9	17	1	4
		0	1		4	3	11	8
			0		14	17	5	2
	0	1	1		7	3	14	14
			0		9	16	2	3
		0	1		4	0	13	11
			0		5	14	4	4

Source Dobra et al. (2002) originally Edwards et al. (1985).

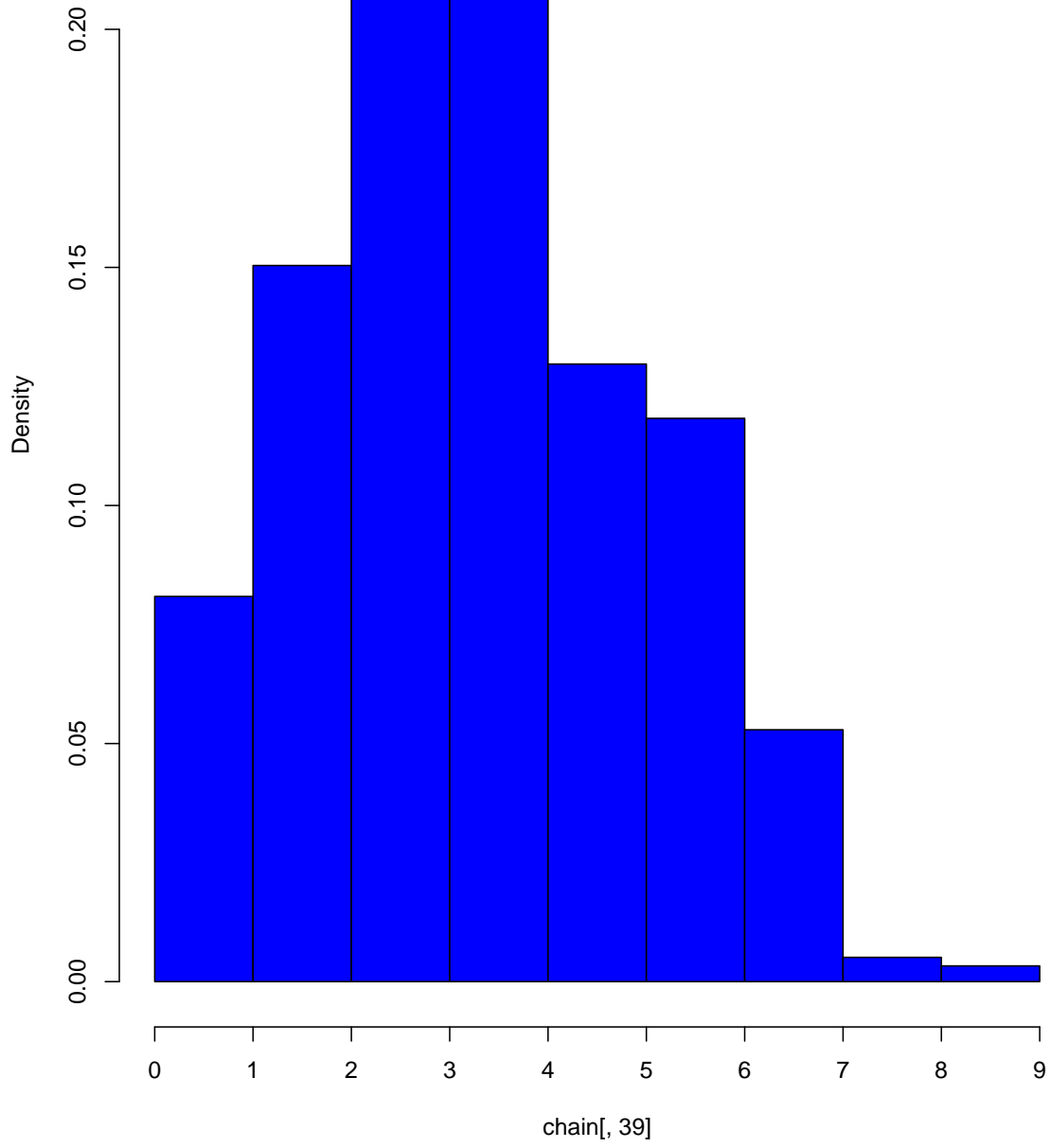
Disclosure Risk in Releasing all Two-way Marginals

- Simulate tables from the hypergeometric distribution obtained by conditioning on all two-way marginals.

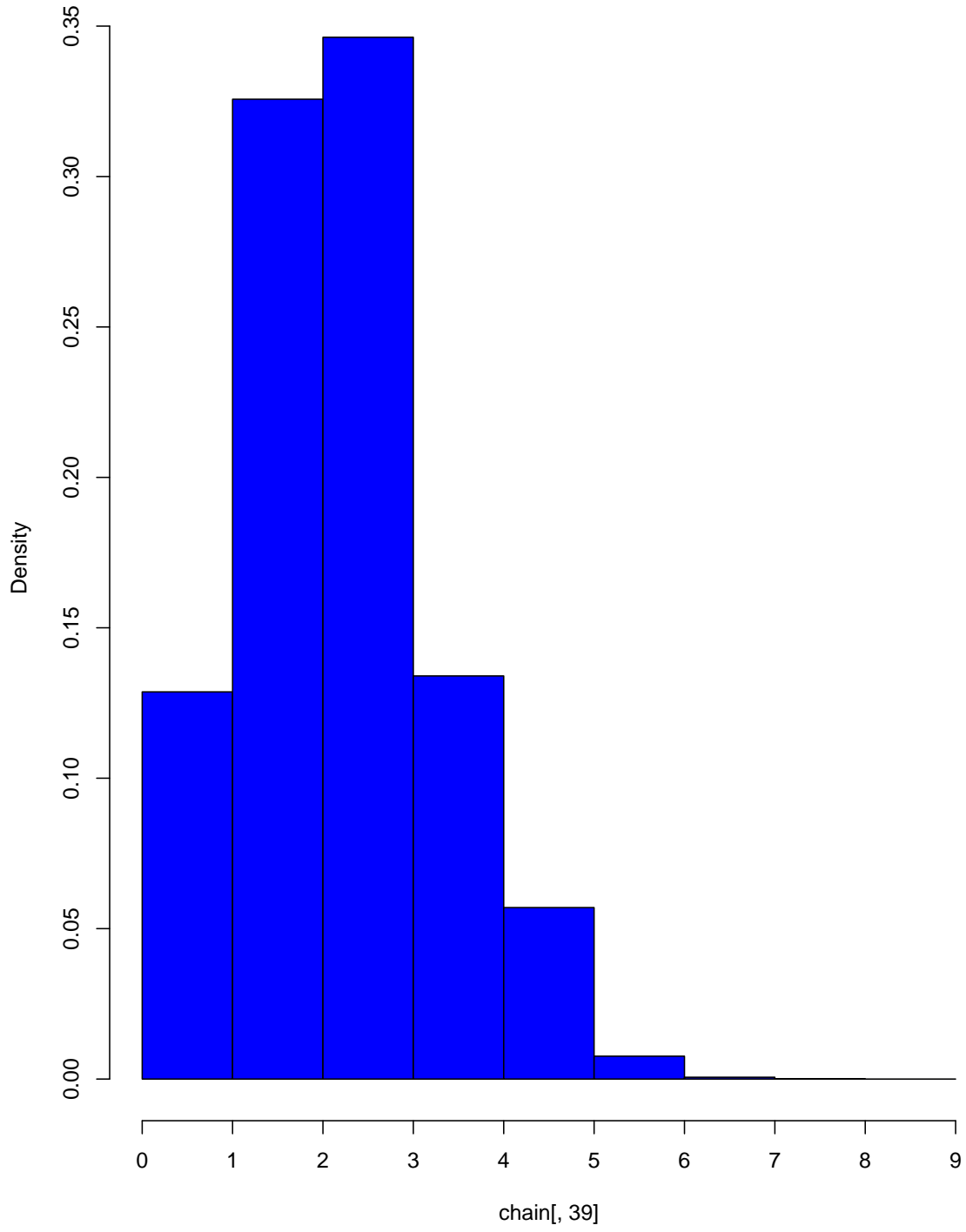
```
> data(czech.dat)
> chain <- simulate.conditional(y ~ . ^ 2,
+                               data = czech.dat,
+                               method = "cab",
+                               nosim = 10 ^ 4,
+                               p = .1)
> hist(chain[,39])
```

(Note `simulate.conditional` has not yet implemented on the version of `exactLoglinTest` on CRAN)

Histogram of chain[, 39]



Histogram of chain[, 39]



Summary

- Benefits of normal approximation
 - Applies very generally
 - Easy to implement
 - Can be very fast
- Drawbacks
 - Not as fast as hit-and-run Gibbs samplers targeted at particular models. Forster McDonald and Smith (JRSSB 1996), Diaconis and Sturmfels (Annals 1998).
- Future Work
 - Conditional likelihood calculations