

# Candidate Sampling Schemes and Some Important Applications

Brian S. Caffo

Chair: James G. Booth

## Dissertation outline

- Introduction
- Review of Monte Carlo
- Review of conditional inference
- An MCMC algorithm for approximating conditional probabilities
- ESUP accept/reject sampling
- MCEM algorithm
- Discussion

## Candidate sampling schemes

- Target distribution  $F$  with density  $f$
- Candidate distribution  $G$  with density  $g$
- $\text{Supp}(F) \subset \text{Supp}(G)$

$F$  and  $G$  are the same type

$$F \ll G$$

- Accept/reject sampling
- Independence metropolis algorithm
- Metropolis Hastings algorithm, allows  $G$  to depend on previously generated variable

## Metropolis Hastings algorithm

- Current state  $x_i$
- Target density  $f(\cdot)$
- Candidate transition density  $g(\cdot|x_i)$
- Generate  $Y \sim g(y|x_i)$
- Accept  $Y$  as the next state with probability

$$\min \left( \frac{f(Y)g(x_i|Y)}{f(x_i)g(Y|x_i)}, 1 \right)$$

otherwise next state is  $x_i$

- Markov chain with  $f$  as stationary density
- $g_\theta(\cdot|x_i)$  can be selected at random

Select  $\theta$  at random from  $\pi(\theta)$

Candidate is  $g_\theta(\cdot|x_i)$

## 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  $H_0 : \boldsymbol{\lambda} = \mathbf{0}$

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

Sufficient statistics for  $\boldsymbol{\beta}$  under  $H_0$  are  $\mathbf{S}_{\boldsymbol{\beta}} = \mathbf{X}^t \mathbf{Y}$

Test fit of the null model using conditional distribution

$$f(\mathbf{y} | \mathbf{S}_{\boldsymbol{\beta}}) \propto \prod_{i=1}^n y_i!^{-1}$$

where  $\mathbf{Y}$  is in  $\Gamma = \{\mathbf{y} | \mathbf{X}^t \mathbf{y} = \mathbf{X}^t \mathbf{y}_{obs}\}$

## Benefits of conditional inference

- Eliminates  $\beta$
- Inferences are “exact”
- Induce correlation
- Avoid Neyman/Scott
- Most standard tests are conditional tests

## Criticisms of conditional inference

- $S_{\beta}$  not ancillary for  $\lambda$
- Often too conservative
- Degenerate conditional distribution
- Often hard/impossible to do

Complexity or size of  $\Gamma$  make calculating all of its members impossible

## Our algorithm for Monte Carlo conditional inference

- Use the MH algorithm to approximate expectations from  $f(\mathbf{y}|\mathbf{S}_\beta)$

- Poisson rv's are approximately normal

- Sampling normal random variables

$\mathbf{Y}$  subject to  $\mathbf{X}^t \mathbf{Y} = \mathbf{s}_\beta$  is easy

- Sampling normal random variables

$$\mathbf{Y} = [ \mathbf{Y}_1^t \quad \mathbf{Y}_2^t ]^t$$

subject to  $\mathbf{X}^t \mathbf{Y} = \mathbf{s}_\beta, \mathbf{Y}_2 = \mathbf{y}_2$

is nearly as easy

- Update a few random cells each iteration
- Round in a clever way
- Specify  $g(\mathbf{y}_{new}|\mathbf{y}_{old})$
- Specify  $g(\mathbf{y}_{old}|\mathbf{y}_{new})$
- Irreducibility

		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
	Tot	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
	Tot	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

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



## Accept/reject sampling

- Target distribution  $F$  density  $f$
- Candidate distribution  $G$  density  $g$
- $C \equiv \sup_x \frac{f(x)}{g(x)} < \infty$  (for now)
- Simulates  $G$  variates and accepts those most consistent with being  $F$  variates

---

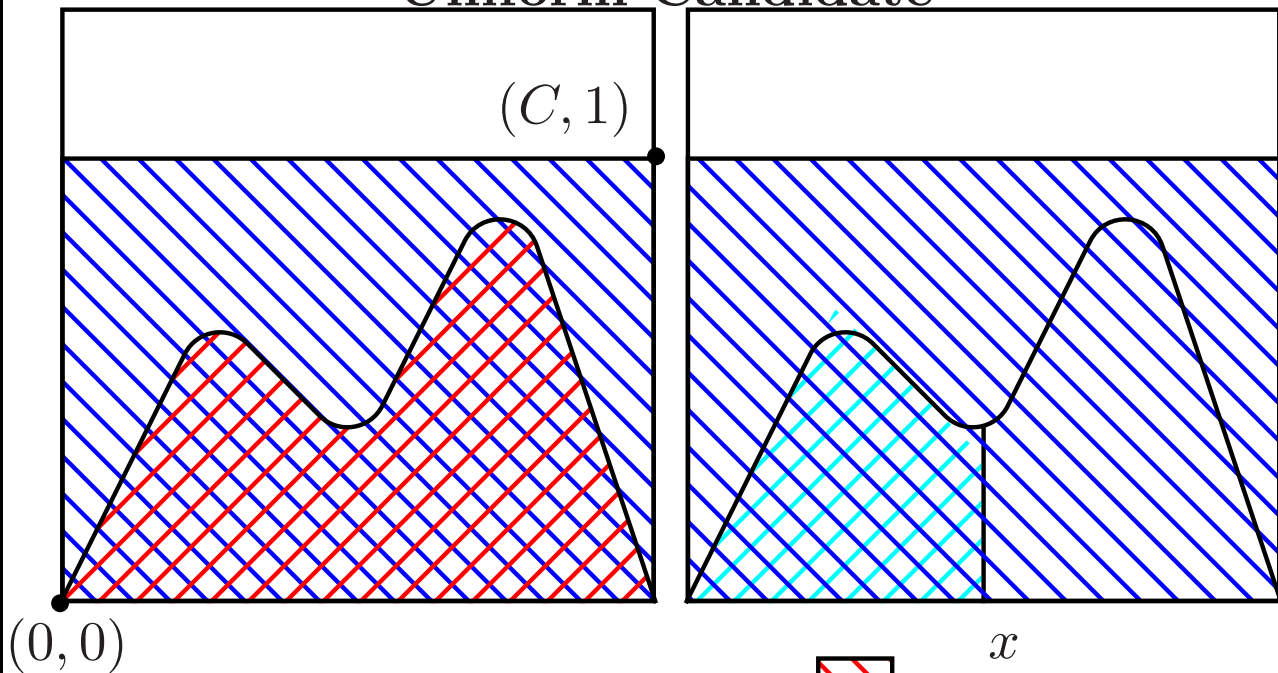
## Accept/reject algorithm

---

- 1 Generate  $X \sim G$
  - 2 Generate  $U \sim U(0,1)$
  - 3 Accept  $X$  if  $U \leq \frac{f(X)}{Cg(X)}$
- 

- Accepted  $X$ s have distribution  $F$
- Acceptance rate is  $1/C$
- Only have to know  $f$  and  $g$  up to constants of proportionality
- Any upper bound on  $C$  works

# Accept/reject sampling Uniform Candidate



$P(\text{Point accepted})$

$$= \frac{\text{[Red shaded area]}}{\text{[Blue shaded area]}} = 1/C$$

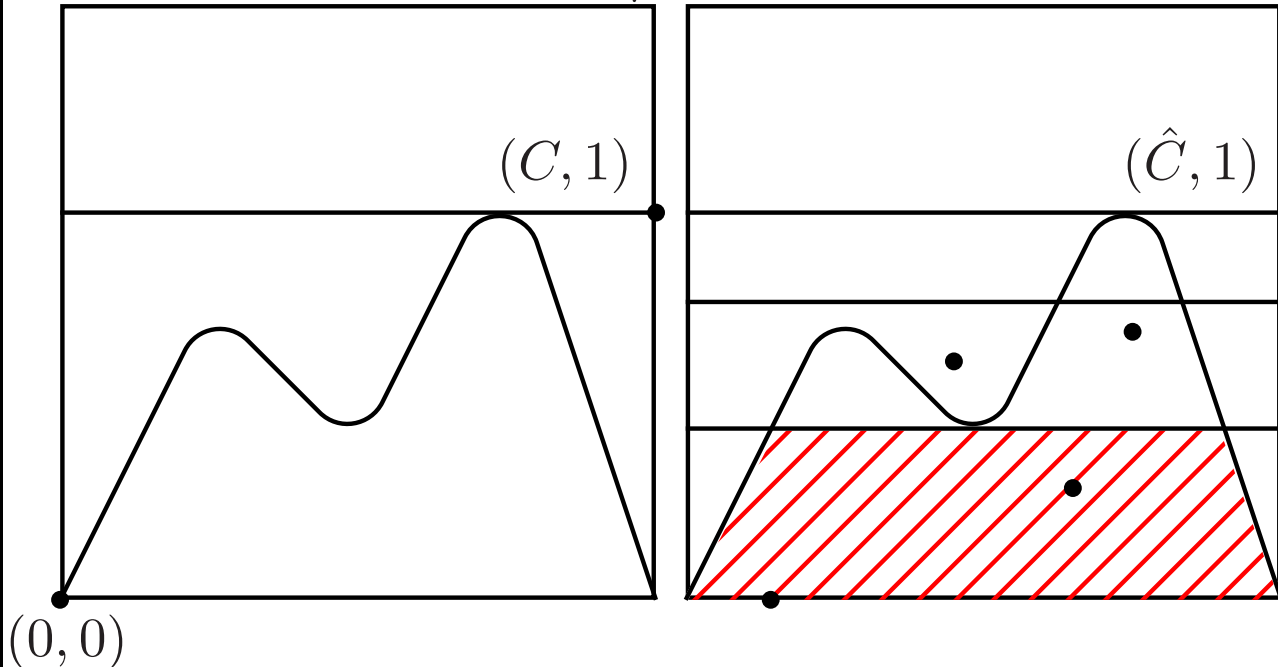
$P(\text{Pt acpt and } X < x)$

$$= \frac{\text{[Cyan shaded area]}}{\text{[Blue shaded area]}} = F(x)/C$$

$P(X < x | \text{Pt acpt})$

$$= \frac{\text{[Cyan shaded area]}}{\text{[Red shaded area]}} = F(x)$$

## ESUP Accept/reject sampling



- Estimate  $C$  with the largest observed value of  $f(X_i)/g(X_i)$
- Sequence of accepted  $X$ s are no longer independent or identically distributed
- Conditional on the value of  $\hat{C}$  really simulating from  $\min(f, \hat{C}g)$

---

## ESUP Accept/reject algorithm

---

- 1 Generate  $X \sim G$
  - 2 Generate  $U \sim U(0, 1)$
  - 3 Accept  $X$  if  $U \leq \frac{f(X)}{\hat{C}g(X)}$
  - 4 Update  $\hat{C} = \max(\hat{C}, \frac{f(X)}{g(X)})$
- 

- Prove everything about ESUP accept/reject by comparing the candidates it accepts with the candidates KSUP accepts
- As  $\hat{C} < C$  ESUP accept/reject always accepts candidates that KSUP accept/reject accepts
- For convenience we assume contrary to the algorithm above that  $\hat{C}$  is updated only once for every accepted candidate

## Notation

- Let  $\{X_{ij}\} \sim G$
- Let  $\{U_{ij}\} \sim \text{Uniform}(0, 1)$
- $Y_i = X_{i\tau_i}$  where
 
$$\tau_i = \min\{j | U_{ij} \leq f(X_{ij})/Cg(X_{ij})\}$$
- $\tilde{Y}_i = X_{i\tilde{\tau}_i}$  where
 
$$\tilde{\tau}_i = \min\{j | U_{ij} \leq f(X_{ij})/\hat{C}_i g(X_{ij})\}$$
- Assume  $C = f(x)/g(x)$  for some  $x$  in the support of  $F$
- Note if  $\sum_{i=1}^{\infty} P(Y_i \neq \tilde{Y}_i) < \infty$  it doesn't matter whether we use ESUP or KSUP accept/reject sampling

$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$
$U_{11}$	$U_{12}$	$U_{13}$	$U_{14}$	$U_{15}$
	$\tilde{Y}_1 = X_{12}$			$Y_1 = X_{15}$
	ESUP			KSUP

**Theorem** *If the support of  $F$  is discrete then*

$$\sum_{i=1}^{\infty} P(Y_i \neq \tilde{Y}_i) < \infty$$

*The sequences are the same for all but finitely many  $i$*

Argument

- For each candidate generated there is a positive probability  $\hat{C} = C$
- Notice then the number of iterations until  $\hat{C} = C$  is finite with probability one

$$P(Y_i \neq \tilde{Y}_i) \leq P(\text{A geometric random variable} \geq i)$$

- It is then easy to show the right hand side of the inequality sums.

In a sense, this theorem also covers continuous cases

## Continuous Case

**Theorem**  $P(Y_i \neq \tilde{Y}_i) \leq E[C/\hat{C}_i - 1]$

**Theorem**  $E[C/\hat{C}_i - 1] = \mathcal{O}(i^{-1})$

Argument

Note  $Z_i = \hat{C}_i/C$  is a max of i.i.d. rvs bdd by 1

$P(Z_i \leq z) = F_Z^i(x)$  for distribution function  $F_Z$

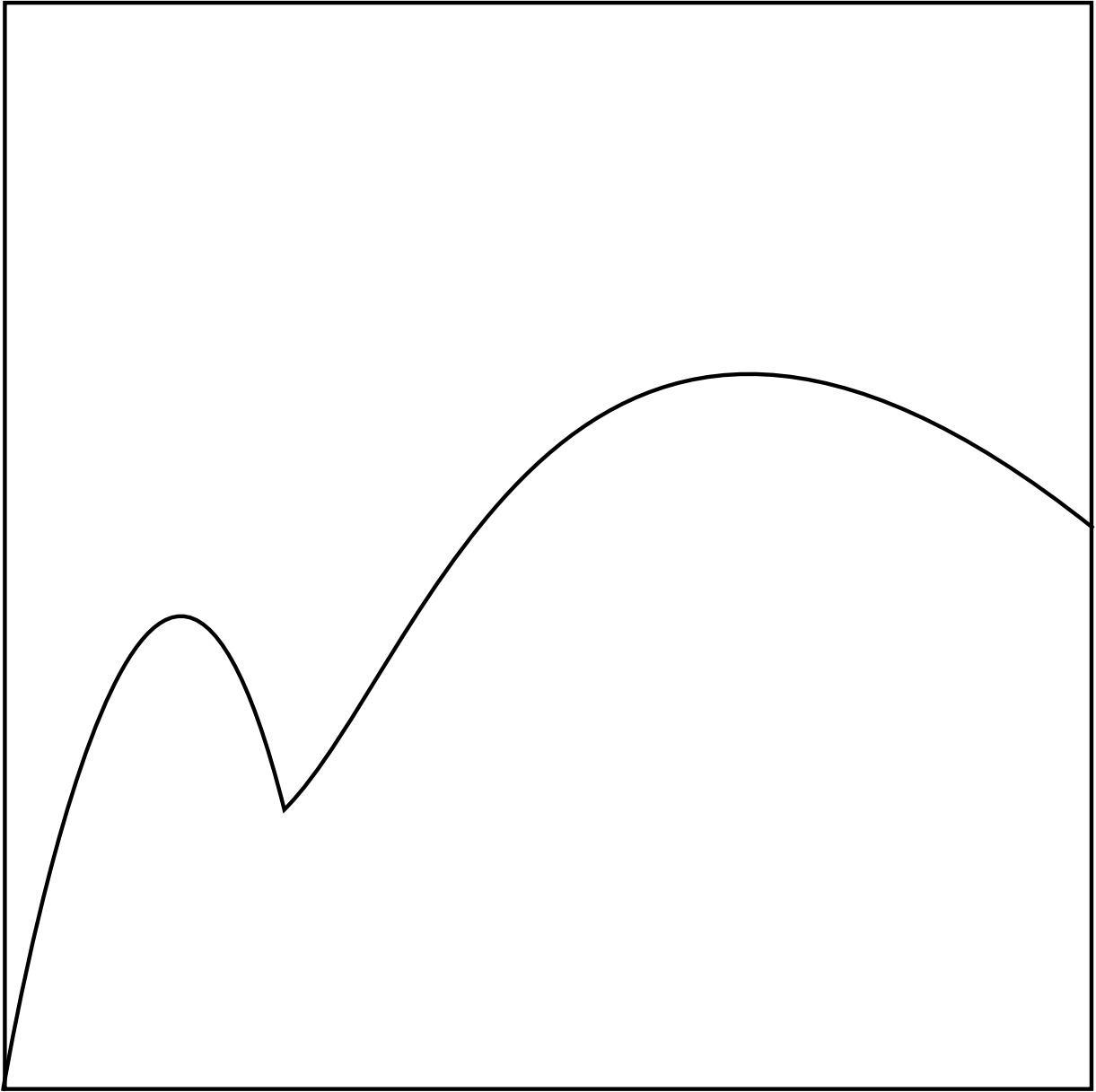
$$\begin{aligned} E[1 - Z_i] &= 1 - \int_0^1 (1 - F_Z^i(x)) dx \\ &\leq F_Z^i(\epsilon) + \int_\epsilon^1 F_Z^i(x) dx \end{aligned}$$

Need to show  $F_Z^i(x) \leq x^{pi}$  for some  $0 < p \leq 1$  and  $\epsilon \geq x \leq 1$

Equivalently  $f_Z(x) \leq px^{p-1}$

Always true for some  $p$  if  $f_Z(1) > 0$

Assumption that  $C = f(x)/g(x)$  for some  $x$  in the support of  $F$



## Main Theorem

- $\{Y_i\}$  the i.i.d. sequence of  $F$  variates from the KSUP algorithm
- $\{\tilde{Y}_i\}$  the sequence from the ESUP algorithm
- If  $Y$  is an  $F$  variate, let the mean of  $h(Y) = \mu_h$  and the variance of  $h(Y)$  be  $\sigma_h^2$

$$(i.) \frac{\sum_{i=1}^n h(Y_i)}{n} \rightarrow \mu_h$$

$$(ii.) \frac{\sqrt{n} \left( \frac{\sum_{i=1}^n h(Y_i)}{n} - \mu_h \right)}{\sigma_h} \xrightarrow{\mathcal{D}} \mathbf{N}(0, 1)$$

**Theorem** If  $h$  is continuous and  $E|h(Y)|^{2+\delta} < \infty$  for some  $\delta > 0$  then (i.) and (ii.) hold with  $Y_i$  replaced with  $\tilde{Y}_i$

Proof  $P(Y_i \neq \tilde{Y}_i) = \mathcal{O}(i^{-1})$  and Holder inequality

## Notes

- All theorems apply using any estimates of  $C$  that exceed the  $\hat{C}_i$
- Under smoothness assumptions large sample behavior of  $\hat{C}_i$  shows that

$$P(Y_i \neq \tilde{Y}_i) = \mathcal{O}(i^{-2})$$

- Infinite value of  $C$  can be diagnosed using large sample behavior of *exceedances*  
Results in a p-value based on *Greenwood's* statistic
- Convergence of the ESUP is independent of dimension of  $x$

## Question

Gender	1	2	3	Count
Male	yes	yes	yes	342
Male	yes	yes	no	26
Male	yes	no	yes	11
Male	yes	no	no	32
Male	no	yes	yes	6
Male	no	yes	no	21
Male	no	no	yes	19
Male	no	no	no	356
Female	yes	yes	yes	440
Female	yes	yes	no	25
Female	yes	no	yes	14
Female	yes	no	no	47
Female	no	yes	yes	14
Female	no	yes	no	18
Female	no	no	yes	22
Female	no	no	no	457

## Example: MCEM for a Logit/Normal Model

- Person  $i$  question  $j$
- Response  $Y_{ij}|U_i$  are independent Bernoulli( $\pi_{ij}$ )

$$\begin{aligned}\log \frac{\pi_{ij}}{1 - \pi_{ij}} &= \text{Intc} + \text{Sex} + \text{Question} + \text{Person} \\ &= \alpha + \gamma I(\text{person } i \text{ is female}) + \beta_j + U_i\end{aligned}$$

- $U_i$  are independent Normal( $0, \sigma^2$ )
- Perhaps not the best model for this data

## MCEM Algorithm

- Let  $\boldsymbol{\theta} = (\alpha, \gamma, \beta_1, \beta_2, \sigma)^t$
- Let  $\boldsymbol{\theta}_t$  be the current estimate of  $\hat{\boldsymbol{\theta}}$
- Let  $c_i$  be the count of people in group  $i$
- EM algorithm obtains  $\boldsymbol{\theta}_{t+1}$  by maximizing the expected complete data log-likelihood

$$Q_t = \sum_{i=1}^{16} c_i E_t^* [\log f(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\theta})]$$

- Where  $E_t^*$  denotes expectation with respect to  $U_i | \mathbf{y}_i, \boldsymbol{\theta}_t$
- MCEM maximizes

$$\hat{Q}_t = \sum_{i=1}^{16} \frac{c_i}{m_i} \sum_{k=1}^{m_i} \log f(\mathbf{y}_i, \mathbf{U}_{ik}; \boldsymbol{\theta})$$

where  $U_{ik}$  are i.i.d. from  $U_i | \mathbf{y}_i, \boldsymbol{\theta}_t$

## Details

- To minimize the Monte Carlo variance of  $\hat{Q}_t$  we should set

$$d_i = c_i (\text{Var} [\log f(\mathbf{y}_i, \mathbf{U}_{ik}; \boldsymbol{\theta})])^{1/2}$$

$$m_i = M \frac{d_i}{\sum_{i=1}^{16} d_i}$$

Usually the counts dominate this estimate, we can just set  $d_i = c_i$

- ESUP accept/reject sampling
- Use a shifted and scaled  $t_3$  distribution as the candidate distribution
- Location and scale parameters are second order Taylor approximations to the moments of  $U_i | \mathbf{y}_i, \boldsymbol{\theta}_t$
- Difficult to calculate the exact  $C$  for this problem

## Performance of ESUP for one iteration and one cluster in EM algorithm

Table 1: Average number of differences (AND) and acceptance rate (AR) for marginal and Laplace candidates with  $z/n = 1/3$  for  $M = 1,000$ .

$n$	$z$	Marginal		Laplace/ $t$	
		AND	AR	AND	AR
9	3	20.56	0.11	0.28	0.85
12	4	18.359	0.07	0.43	0.85
15	5	17.03	0.05	0.27	0.85
18	6	16.09	0.04	0.25	0.86
21	7	14.429	0.03	0.31	0.86
24	8	13.703	0.02	0.28	0.86
27	9	13.134	0.02	0.32	0.86
30	10	11.999	0.02	0.25	0.86

## Future research

- Standard errors for MCMC exact conditional inference

Bounding the mixing time

Perfect sampling

- Extensions to the conditional saddlepoint approximation

- Groebner bases ?

- Completely automated accept/reject sampler

Rao-Blackwellization