Assessing and Improving the Reliability of Models of Molecular Evolution

November 3, 2021

Overview

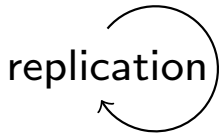
- Ch. 1 - Introduction to Molecular Evolution and Models
- Ch. 2 - Modified Likelihood to Restore LR Test Regularity
- Ch. 3 - Smoothed Bootstrap Aggregation
- Ch. 4 - Difficulties with Branch-Site Models

Models of Molecular Evolution

Model Substitutions between Codons to Detect Positive Selection

- Continuous-time Markov process
- Likelihood framework
- Bayesian classification

The Central Dogma of Molecular Biology



DNA		RNA		Protein
	transcription		translation	
C-G		G		
A-T		U		V
G-C		C		
G-C		C		
T-A		A		Q
T-A		A		
A-T		U		
G-C		C		S
G-C		C		
.		.		.
.		.		.
.		.		.

Genetic Code

First Codon Position	Second Codon Position: U		Second Codon Position: C		Second Codon Position: A		Second Codon Position: G	
U	UUU	F	UCU	S	UAU	Y	UGU	C
	UUC	F	UCC	S	UAC	Y	UGC	C
	UUA	L	UCA	S	UAA	*STOP*	UGA	*STOP*
	UUG	L	UCG	S	UAG	*STOP*	UGG	W
C	CUU	L	CCU	P	CAU	H	CGU	R
	CUC	L	CCC	P	CAC	H	CGC	R
	CUA	L	CCA	P	CAA	Q	CGA	R
	CUG	L	CCG	P	CAG	Q	CGG	R
A	AUU	I	ACU	T	AAU	N	AGU	S
	AUC	I	ACC	T	AAC	N	AGC	S
	AUA	I	ACA	T	AAA	K	AGA	R
	AUG	M/*START*	ACG	T	AAG	K	AGG	R
G	GUU	V	GCU	A	GAU	D	GGU	G
	GUC	V	GCC	A	GAC	D	GGC	G
	GUA	V	GCA	A	GAA	E	GGA	G
	GUG	V	GCG	A	GAG	E	GGG	G

$4^3 = 64$ codons

20 amino acids

Nonsynonymous and Synonymous Substitutions

Mutations and Substitutions

Wild-type Allele *A*

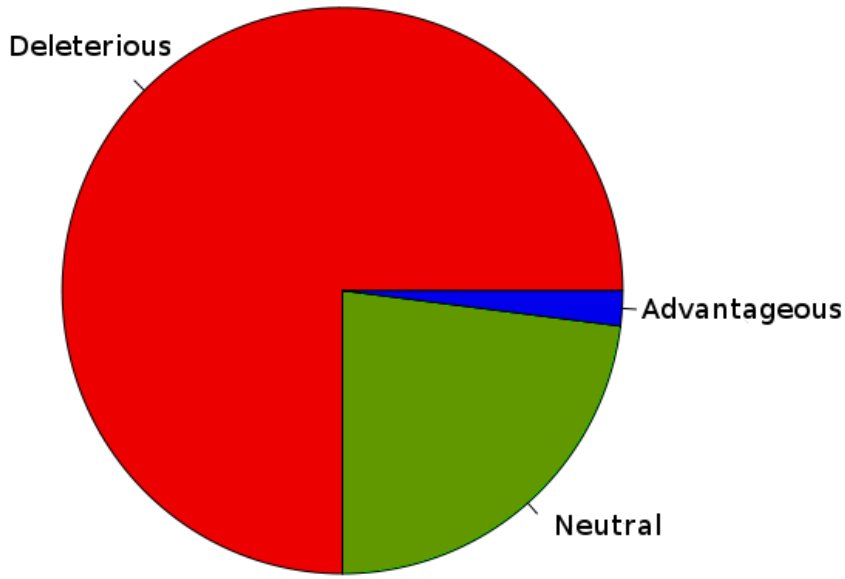
ATG	GTG	CAC	CTG	ACT	CCT	**GAG**	GAG	AAG
Start	Val	His	Leu	Thr	Pro	**Glu**	Glu	Lys

Mutant allele *a*

ATG	GTG	CAC	CTG	ACT	CCT	**GTG**	GAG	AAG
Start	Val	His	Leu	Thr	Pro	**Val**	Glu	Lys

Signature of Positive Selection and Neutral Theory

- $(p_N/p_S)/(p_N^0/p_S^0) = \omega$: ratio of ratios of long-run proportions of nonsynonymous and synonymous substitutions under alternative and null models



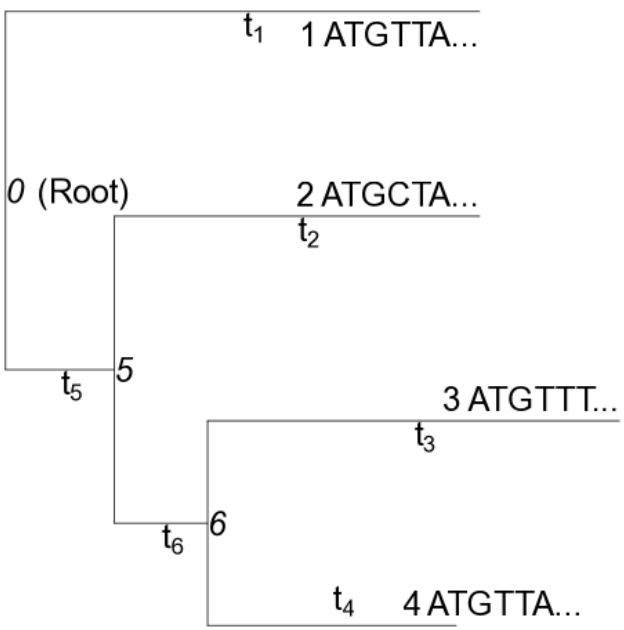
Netural Theory predicts $\omega \leq 1$

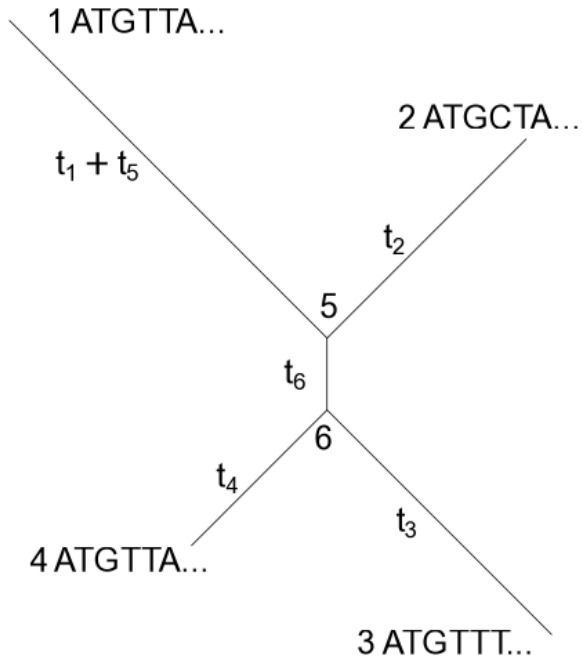
$$2\Delta\ell = 2[log(L_a) - log(L_0)]$$

Data

Taxon — Codon Site

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	A	T	G	T	T	A	T	T	T	A	G	T	T	G	G	T	T	C	A	T	T	T	A	T	A	T	A	A	T	A	A	T	A	A	T	T	T
2	A	T	G	C	T	A	T	T	T	A	G	T	T	G	A	T	T	T	A	G	A	T	A	T	A	T	A	A	T	G	G	T	G	T	T	A	T	T	G	A	T	T	T
3	A	T	G	T	T	T	T	T	T	A	G	T	T	G	A	T	T	T	A	T	T	A	T	A	T	A	T	T	A	A	G	G	A	T	A	A	T	A	G	T	T	T	T
4	A	T	G	T	T	A	T	T	T	A	G	T	T	G	A	T	T	T	A	T	T	A	G	T	A	T	A	G	T	A	G	T	A	A	T	A	G	T	A	T	T	T	T

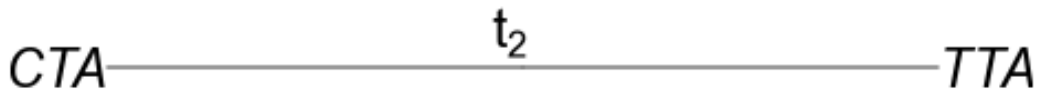




Codon Substitution as a Markov Process

Assumptions:

- Independence across sites
- Independence across branches



$$\ell = \log(L) = \sum_{h=1}^{n} \log\{f(\boldsymbol{x}_h|\theta)\}$$

Codon Substitution as a Markov Process

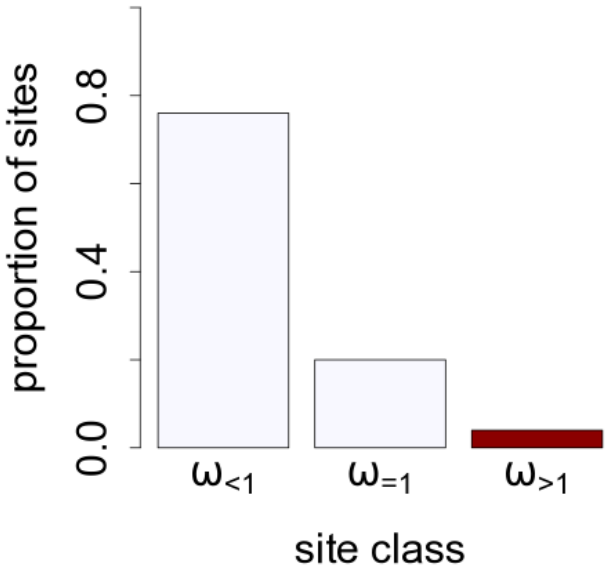
Markov Process with 61 Sense Codons as States

$$q_{ij} = \begin{cases} 0, & i \text{ and } j \text{ differ at two or three codon positions} \\ \pi_j, & i \text{ and } j \text{ differ by a synonymous transversion} \\ \kappa\pi_j, & i \text{ and } j \text{ differ by a synonymous transition} \\ \omega\pi_j, & i \text{ and } j \text{ differ by a nonsynonymous transversion} \\ \omega\kappa\pi_j, & i \text{ and } j \text{ differ by a nonsynonymous transition} \end{cases}$$

$$P(t) = e^{Qt}$$

Mixture Models

- ω averaged over sites ineffective
- Introduce discretized ω distribution
- p_i: proportion of sites in class i



- $f(\boldsymbol{x_h}|\hat{\boldsymbol{\theta}}) = \sum_{i=0}^{k-1} f(\boldsymbol{x_h}|\omega_i;\hat{\boldsymbol{\theta}})p_i$
- $f(\omega_i|\boldsymbol{x_h};\hat{\boldsymbol{\theta}}) = \frac{f(\boldsymbol{x_h}|\omega_i;\hat{\boldsymbol{\theta}})p_i}{\sum_{j=0}^{k-1} f(\boldsymbol{x_h}|\omega_j;\hat{\boldsymbol{\theta}})p_j}$

Chapter 2

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 148(3):929-936.

Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular biology and evolution, 18(8):1585-1592.

Wong, W. S., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics, 168(2):1041-1051.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc, 82(398):605–610.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(1):19-29.

Mingrone J., Susko E., Bielawski J.P. 2018. ModL: exploring and restoring regularity when testing for positive selection. Bioinformatics 35:2545-2554.

LR Distribution under Null Hypothesis

- $p(x;\beta,p_+) = p_0 p(x|\omega<1;\zeta,\lambda)+(1-p_+)(1-p_0)p(x|1;\zeta)+p_+(1-p_0)p(x|\omega_+;\zeta)$
- ζ: parameters common to each ω
- λ: mixture parameters under purifying selection
- $\omega_+ \geq 1$
- Let $\psi = (\zeta^T, \lambda^T, p_0)^T$ be parameters common to H_0 and H_a

$$LR = 2\Delta\ell = 2\{l(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\hat{\psi}_H)\}$$

- Standard likelihood theory: $\sim \chi^2_{\Delta_p}$
- Estimates on boundary of parameter space
- Unidentifiability: $\underline{p_+ = 0 \text{ any } \omega_+}$ or $\underline{\omega_+ = 1 \text{ any } p_+}$ gives null model

LR Distribution under Null Hypothesis

$$LR = 2\{l(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\hat{\psi}_H)\} \tag{1}$$

$$LR_{\omega_+>1} = 2\{l(\hat{p}_+(\omega_+), \omega_+, \hat{\psi}(\omega_+)) - l_H(\hat{\psi}_H)\} \tag{2}$$

- Identifiability with $\omega_+ > 1$ fixed
- Case 5 of Self and Liang (1987): (2) $\sim \chi^2_0/2 + \chi^2_1/2$
- (1) obtained by maximizing (2) over $\omega_+ \geq 1$
  - (1) anti-conservative

Modified Likelihood - Chen et al. (2001)

Similar difficulty when testing heterogeneity of mixture model

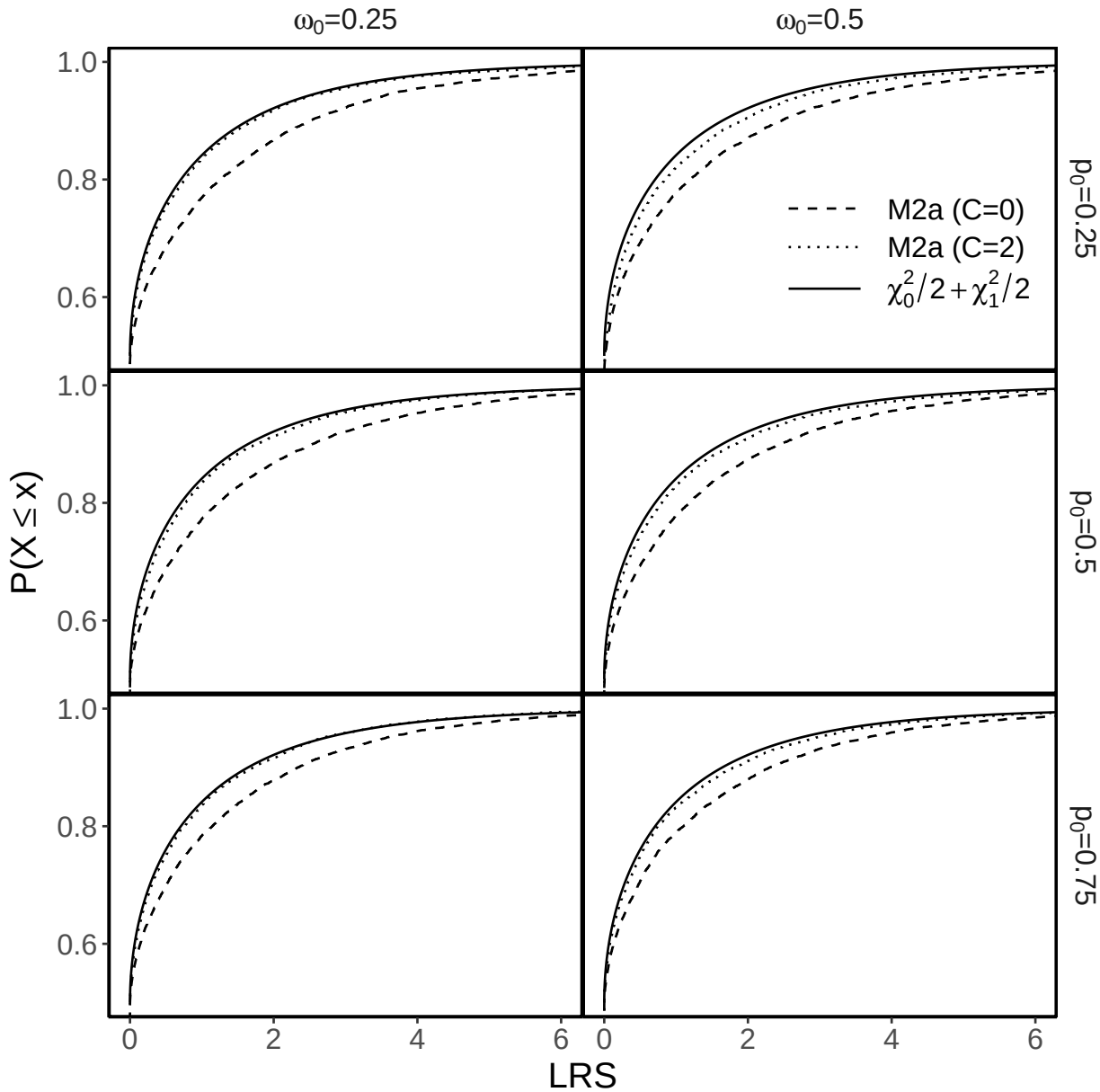
- Data x is $\gamma p(x;\theta_1) + (1-\gamma)p(x;\theta_2)$
- Unidentifiability: $\underline{\gamma = 0 \text{ any } \theta_1}$ or $\underline{\theta_1 = \theta_2 \text{ any } \gamma}$
- Chen et al. (2001)
  - Restore simple limiting distributions while maintaining a test statistic similar to the LR statistic
  - Replace log likelihoods with modified log likelihoods: $C \log[\gamma(1-\gamma)]$ ($C > 0$ is tuning parameter)

Modified Likelihood - Codon Models

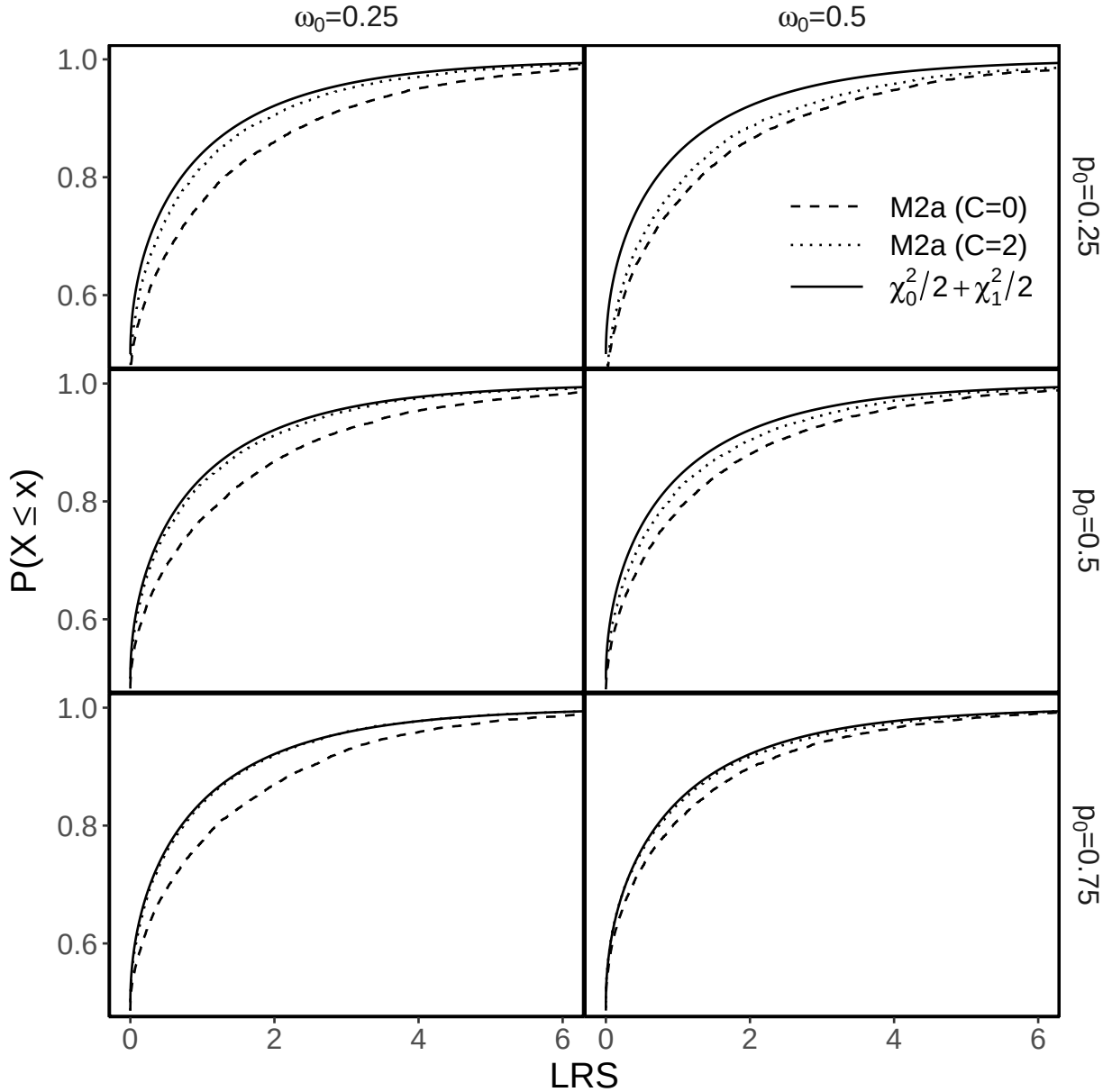
Mingrone, Susko, Bielawski (2019)

- Modified ℓ under H_a: $\tilde{l}(p_+, \omega_+, \psi) = l(p_+, \omega_+, \psi) + C \log(p_+)$
- ModL statistic: $2\{\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\hat{\psi}_H)\} \sim \chi^2_0/2 + \chi^2_1/2$

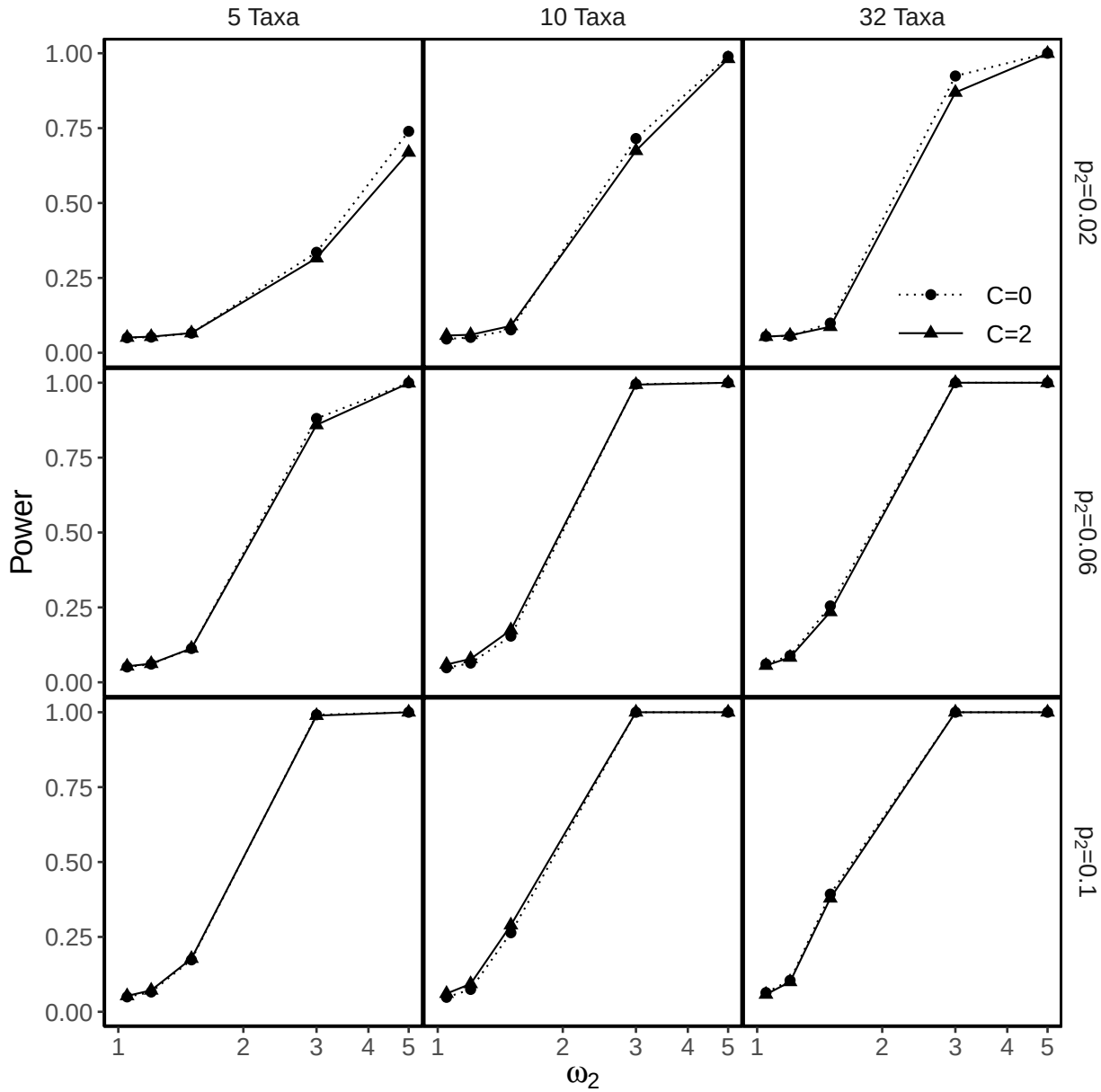
Modified LR Distribution Approximations Accurate for Most Settings



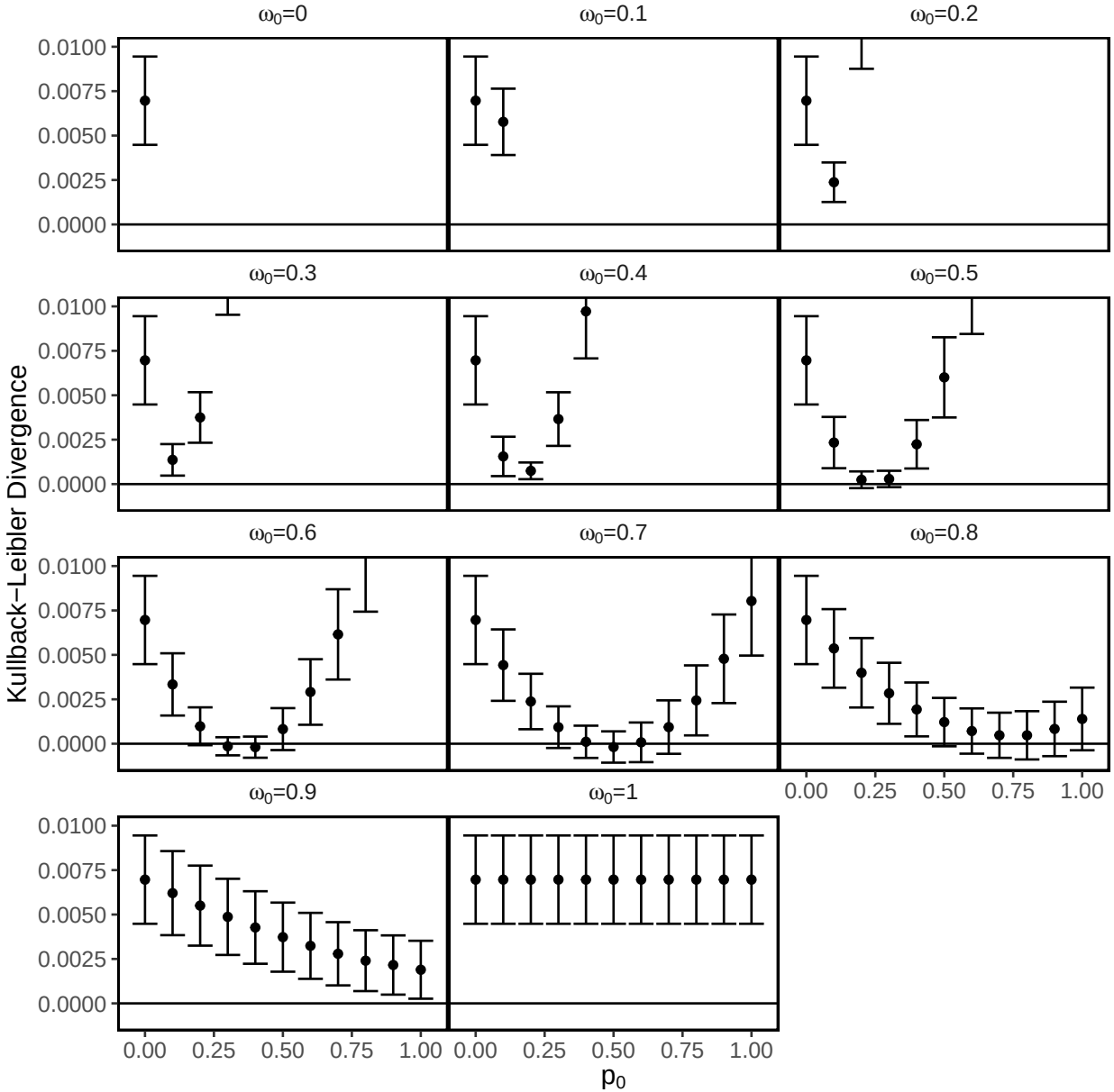
Modified LR Distribution Approximations
Data Poor Setting



Modified LR Distribution
Minimal Impact on Power



Investigation of Problematic Mixing Distributions



Neither LR nor ModLR is well approximated by $\chi^2_0/2 + \chi^2_1/2$ when true mixing distribution is $(p_0, \omega_0) = (0.25, 0.5)$ and data poor setting

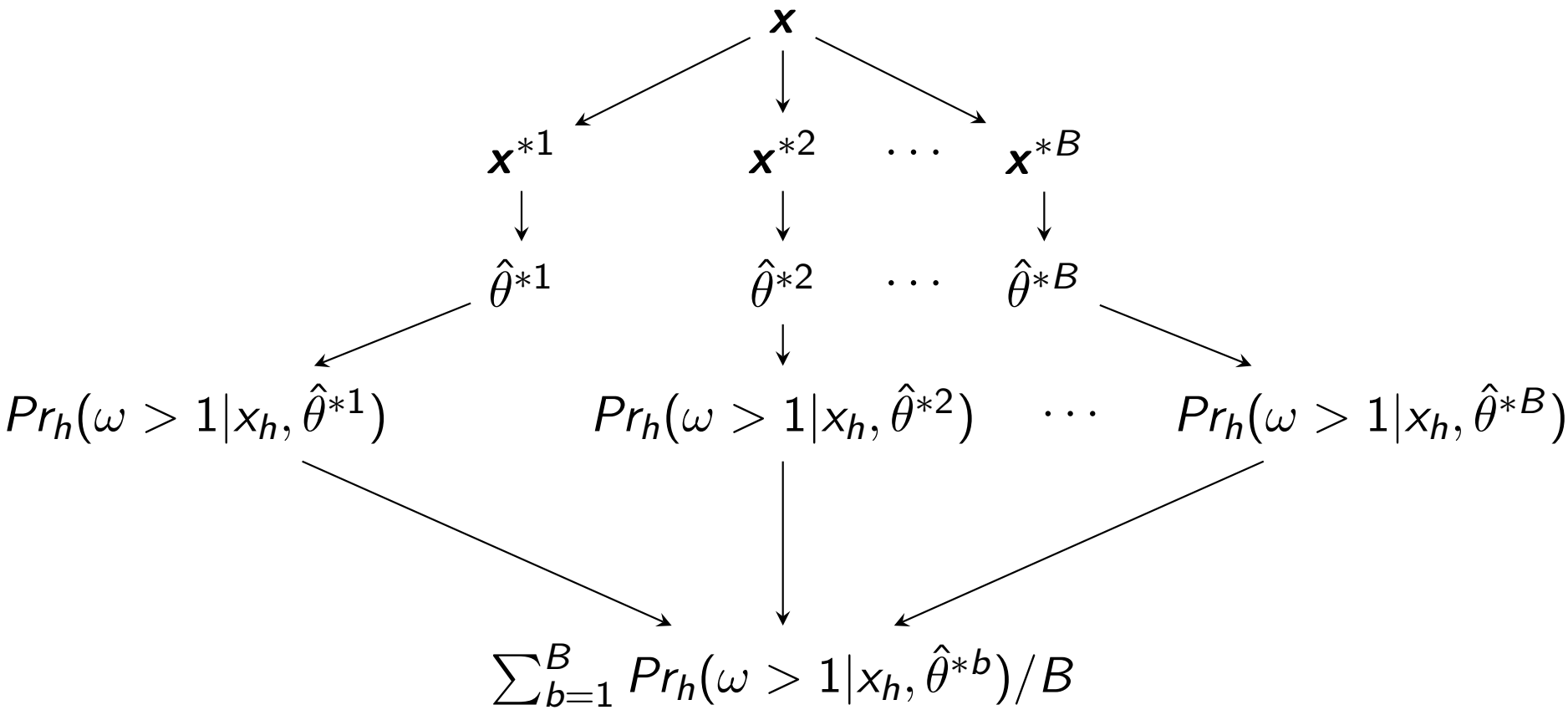
Chapter 3: Smoothed Bootstrap Aggregation (SBA)

Posterior Probability:

$$f(\omega_i|\boldsymbol{x_h};\hat{\boldsymbol{\theta}}) = \frac{f(\boldsymbol{x_h}|\omega_i;\hat{\boldsymbol{\theta}})p_i}{\sum_{j=0}^{k-1} f(\boldsymbol{x_h}|\omega_j;\hat{\boldsymbol{\theta}})p_j} \tag{3}$$

NEB: Pass MLEs directly to (3)
BEB: Assign priors to some MLEs to adjust for uncertainty
SBA: Bootstrap site patterns to obtain many posteriors

SBA


$\boldsymbol{x}$
$\boldsymbol{x}^{*1}$
$\boldsymbol{x}^{*2}$
$\cdots$
$\boldsymbol{x}^{*B}$
$\hat{\theta}^{*1}$
$\hat{\theta}^{*2}$
$\cdots$
$\hat{\theta}^{*B}$
$Pr_h(\omega > 1|x_h, \hat{\theta}^{*1})$
$Pr_h(\omega > 1|x_h, \hat{\theta}^{*2})$
$\cdots$
$Pr_h(\omega > 1|x_h, \hat{\theta}^{*B})$
$\sum_{b=1}^{B} Pr_h(\omega > 1|x_h, \hat{\theta}^{*b})/B$