Consistent Confidence Intervals for Maximum Pseudolikelihood Estimators

Bruce A. Desmarais
Department of Political Science
University of Massachusetts Amherst
Amherst, MA 01003
desmarais@polsci.umass.edu

Skyler J. Cranmer
Department of Political Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
skyler@unc.edu

Abstract

Maximum pseudolikelihood estimation (MPLE) constitutes a computationally efficient and easily implemented alternative to maximum likelihood and simulation-based methods. The MPLE has been shown to be consistent and asymptotically normally distributed in a number of interesting cases. However, the coverage probability of the conventional confidence interval for the MPLE is biased downward. We provide a bootstrap method for consistently estimating confidence intervals for the MPLE. We then apply this method to the U.S. Supreme Court, where Justices' votes on cases are characterized as a fully visible Boltzmann machine.

1 Modeling Dependent Discrete Data

Many problems involve training a model for a p-dimensional binary vector $\boldsymbol{x} \in \{x^-, x^+\}^p$, using a sample of n vectors. A flexible probability model for such an application is given by

$$\mathcal{P}(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\exp\{\boldsymbol{\theta}' \boldsymbol{\Gamma}(\boldsymbol{x}_i)\}}{\sum_{\text{all } \boldsymbol{x}^* \in \boldsymbol{\mathcal{X}}} \exp\{\boldsymbol{\theta}' \boldsymbol{\Gamma}(\boldsymbol{x}^*)\}}, \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter in $\mathbb{R}^k$ and $\boldsymbol{\Gamma}(\cdot)$ is a function that maps $\{x^-, x^+\}^p$ into $\mathbb{R}^k$. Special cases of this distribution have been applied to social network analysis in the form of the temporal exponential random graph model (TERGM) [7], fully visible Boltzmann machines [8], and spatially autoregressive logistic regression [2]. The popularity of this distribution arises from the facts that (1) the $\boldsymbol{\Gamma}(\cdot)$ can be specified to capture nearly any form of dependence among the dimensions of $\boldsymbol{x}$ or dependence of $\boldsymbol{x}$ on exogenous covariates, and (2) many favorable statistical properties are implied by the fact that it is a member of the exponential family [12].

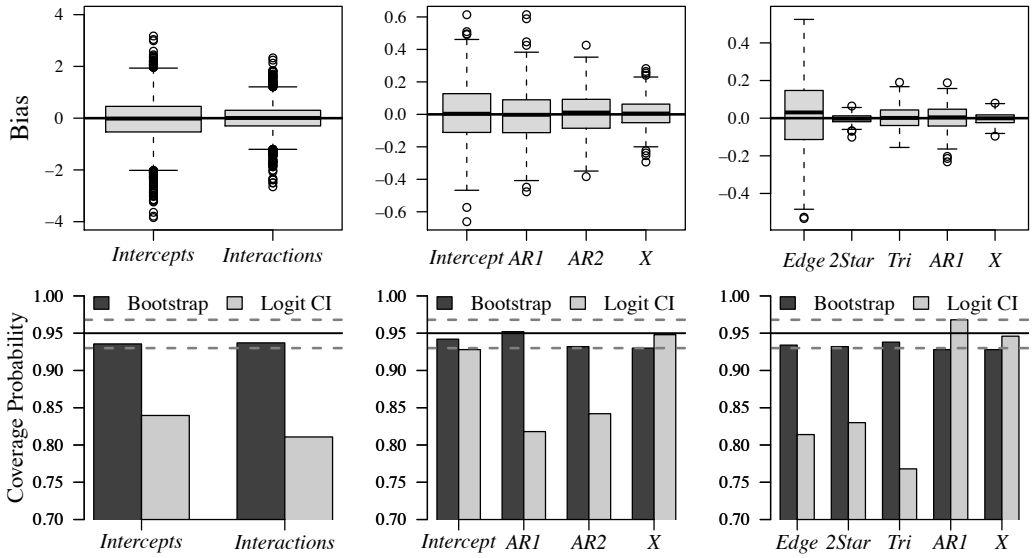
A major obstacle is that the computation of the denominator in (1) requires a summation over 2^p vectors; a prohibitively large number even in moderate-sized applications. As such, the log-likelihood must be approximated. Two general approximation methods exist. First, the denominator can be approximated by summing over a series of $\boldsymbol{x}^*$'s simulated by Markov Chain Monte Carlo [3]. The second approach, Maximum Pseudolikelihood Estimation (MPLE), consists of replacing the joint probability of $\boldsymbol{x}$ with the product over the p conditional probabilities of each element of $\boldsymbol{x}$ given the rest of $\boldsymbol{x}$. The simulation approach is generally more efficient than MPLE, but is much slower than MPLE, can exhibit poor mixing properties in some cases [6], and is not accessible to all practitioners. Conversely, MPLE is fast and convenient in that the estimates can be computed using logistic regression software. Additionally, given standard regularity conditions, the MPLE is consistent and asymptotically normally distributed [1]. Consistency and asymptotic normality of the MPLE have been proven in a number of interesting cases [2, 8, 11]. However, the conventional confidence intervals for the MPLE, derived from the inverse of the observed Fisher information of the logistic regression estimate (Logit CIs), exhibit low coverage probabilities, resulting in a high type-I error rate [12].

2 Consistent Bootstrap Confidence Intervals for Pseudolikeihood

We present a nonparametric bootstrap method that produces valid confidence intervals for the MPLE. Let $\hat{\boldsymbol{\Theta}}_t$ be a sample of t estimates of $\boldsymbol{\theta}$ constructed by computing $\hat{\boldsymbol{\theta}}$ on t samples of n vectors drawn with replacement from $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. It is important to note that vectors are resampled without disturbing the configurations within the vectors. If $\hat{\theta}$ is an M-estimator with respect to the vector-valued random variable $\boldsymbol{x}$, computed by solving for the zeroes of the score of $\sum_{i=1}^{n} h(\boldsymbol{x}_i, \boldsymbol{\theta})$, then $\hat{\boldsymbol{\Theta}}_t$ provides a consistent (in n) estimate of the sampling distribution of $\hat{\theta}$ [9]. We note that the MPLE is such an M-estimator with $h(\boldsymbol{x}, \boldsymbol{\theta}) = \log[\prod_{j=1}^{p} \mathcal{P}(x_j | \boldsymbol{x}_{-j}, \hat{\boldsymbol{\theta}})]$. Thus, the bootstrap sample of MPLEs provides a consistent estimate of confidence intervals for the MPLE.

We conduct a simulation to asses the applicability of the consistency result for the bootstrap method in moderate sample sizes. The models used in the simulation are as follows: (1) a fully visible Boltzmann machine applied to $n = 200$ vectors in $\{-1, 1\}^{10}$, with parameters drawn from a $\mathcal{N}(0, 0.25)$; (2) a second-order autologistic model applied to $n = 50$ 4×4 grids with an intercept, first and second order autoregressive parameters and a standard normally distributed covariate with parameter values of $\{-0.85, 1.0, -0.5, 0.5\}$; (3) a TERGM applied to a series of 25 networks, each with 25 nodes, parametrized with edge, two-star, triangle, edgewise first-order autoregression, and a standard normal dyadic covariate with parameter values $\{-0.25, -0.2, 0.5, 1, 0\}$. In each case 1,000 re-samples are used in the bootstrapping. For each model, the simulation study consists of 500 iterations.[1]

The results are presented in figure 1. We do not find evidence that the MPLE is biased in any of the models under consideration. Also, the coverage probability of the 95% bootstrap confidence intervals, given by the 2.5^{th} and 97.5^{th} percentiles of the bootstrap sample of MPLEs, is very close to 0.95. In most cases the bootstrap technique offers an order-of-magnitude reduction in the bias of the coverage probability relative to the logit CIs. These results provide evidence that, even with moderate n and p, the bootstrap technique offers a way to take advantage of the consistency of the MPLE, while not sacrificing the consistency of hypothesis tests.



(a) Boltzmann machine (b) Spatial Logit (c) TERGM

Figure 1: Monte Carlo Results. Box plots of the iteration-wise bias ($\hat{\theta} - \theta$) over the 500 iterations in the Monte Carlo study are given in the first row. The second row gives the empirical coverage probability of 95% confidence intervals. The dashed-grey lines in the coverage probability plots are placed at the 0.05-level critical values of a two-sided binomial test of the null hypothesis that the true coverage probability is 0.95.

[1] All computations, in both the simulation study and the Supreme Court application, are performed in the R statistical software. Code and data necessary to replicate our results are available upon request.

3 Application: The U.S. Supreme Court as a Boltzmann Machine

A pressing problem in the study of the U.S. Supreme Court is the estimation of the unidimensional, liberal-conservative ideologies of the Justices. The conventional method for deriving these estimates is to posit that the votes cast by the nine Justices on a case are independent and driven by Justice ideologies; item response theory is then used to estimate ideal points (ideologies) based on Justice-vote data [10]. Alternatively, a Boltzmann machine can be used to represent the propensity of each individual dimension in a binary vector, as well as the association *between* each pair of dimensions in that vector [5]. An indicator that codes whether a justice voted in a liberal or conservative direction on a case is available in *The Supreme Court Database* (`http://scdb.wustl.edu/`). We use a Boltzmann machine to characterize voting on the Court in all 178 cases from the 2007 and 2008 terms in which a written opinion was issued. Each case is a vector ($\boldsymbol{v}$) of nine votes coded either 1 (liberal) or -1 (conservative). Let $\boldsymbol{\beta}$ be the ideal points of the nine justices, and $\boldsymbol{\lambda}$ the 36 justice-justice pairwise association parameters. The probability model we fit to the data is

$$\mathcal{P}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{178}\}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} \frac{\exp\{\sum_{j=1}^{9} \beta_j v_{ij} + \frac{1}{2} \sum_{\text{all} j \neq k} \lambda_{jk} v_{ij} v_{ik}\}}{\sum_{\text{all } \boldsymbol{v}^* \in \mathcal{V}} \exp\{\sum_{j=1}^{9} \beta_j v_j^* + \frac{1}{2} \sum_{\text{all} j \neq k} \lambda_{jk} v_j^* v_k^*\}}. \quad (2)$$

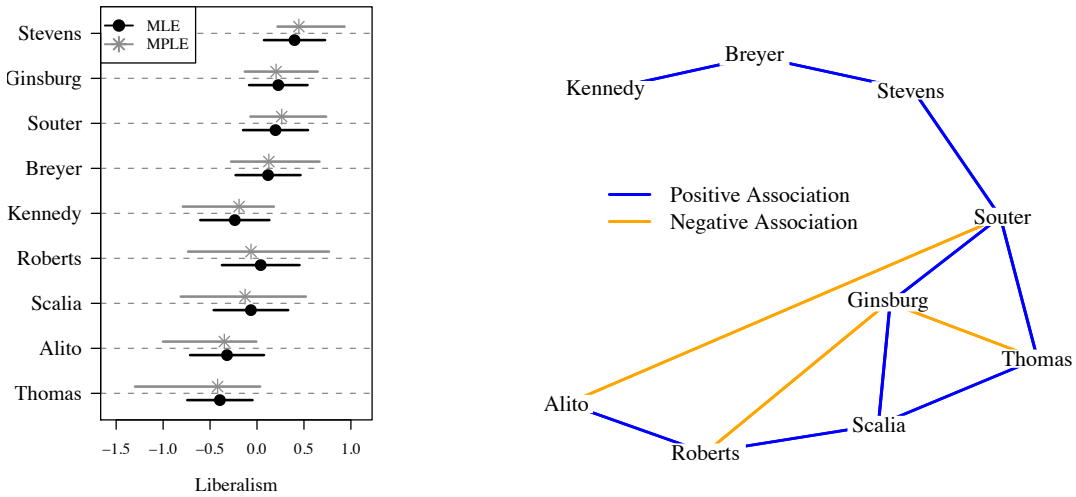
This model allows us to asses the propensity of each Justice to vote liberal and to determine whether the votes of the Justices are related or independent. The estimates are given in Table 3 and Figure 2. Both MLE and MPLE can be applied since 2^9 is only 512 and the normalizing constant can be computed. The MLE and MPLE are very similar. The median value of $|\frac{\text{MPLE}-\text{MLE}}{\text{MLE}}|$ is 0.102, meaning the majority of the 45 MPLE estimates are within a 10% range around the MLE. Moreover, only in 4 out of 36 instances do hypothesis tests at the 0.05 level for association between Justices differ between the bootstrap CI's and the asymptotic variance of the MLE. This divergence is balanced with regards to the result of the test, with 2 instances where the bootstrap rejects the null and the MLE does not, and 2 instances where the MLE rejects and the bootstrap does not. In contrast, tests based on the logit CIs indicate significance 8 more times than does the MLE.

The results indicate that it is inappropriate to assume that the votes of Justices are independent. Both the MLE and MPLE indicate that 11 associations are different from zero at the 0.05 level, and every justice is significantly associated with at least one other. Both the Wald test for the MPLE [4] and the likelihood ratio statistic using the MLE (262,151 and 738 respectively) reject the null hypothesis that all association parameters should be constrained to zero. We can see the impact of interdependence among the justices on the estimates of their ideal points. In panel (a) of figure 2 we see that, without the association parameters, Justice Kennedy (a Reagan appointee) is estimated to be more liberal than Chief Justice Roberts . However, once association is accounted for, Roberts appears more liberal than Kennedy. Kennedy's liberal voting is partly explained through his positive association with the liberal Breyer. Similarly, Roberts' conservatism is partly attributable to positive associations with the conservatives Alito and Scalia and a negative association with the liberal Ginsburg.

Table 1: U.S. Supreme Court Boltzmann Machine: Association Parameters

	Kenn.	Scal.	Thom.	Sout.	Rob.	Stev.	Gins.	Alito	Brey.
Kennedy		0.18	0.11	0.11	$0.39_{\times}$	-0.03	$0.26_{\times}$	$0.38_{\times}$	$0.56^{+}_{\times}$
Scalia	0.20		$0.95^{+}_{\times}$	0.02	$0.79^{+}_{\times}$	-0.10	$0.51^{+}_{\times}$	0.07	-0.27
Thomas	0.09	0.97^{+}		$0.57^{+}_{\times}$	-0.10	-0.11	$-0.35^{+}_{\times}$	$0.44_{\times}$	-0.19
Souter	0.10	0.02	0.51^{+}		$0.37_{\times}$	$0.51^{+}_{\times}$	$0.72^{+}_{\times}$	$-0.64_{\times}$	$0.26_{\times}$
Roberts	0.38	0.80^{+}	-0.14	0.34		0.16	$-0.54_{\times}$	$0.90^{+}_{\times}$	0.15
Stevens	-0.07	-0.07	-0.09	0.51^{+}	0.09		$0.29_{\times}$	0.14	$0.57^{+}_{\times}$
Ginsburg	0.34	0.42	-0.33	0.72^{+}	-0.54^{+}	0.27		0.15	$0.24_{\times}$
Alito	0.38	0.01	0.48^{+}	-0.62^{+}	0.92^{+}	0.15	0.16		$0.34_{\times}$
Breyer	0.58^{+}	-0.24	-0.21	0.24	0.11	0.58^{+}	0.20	0.31	

The upper triangle consists of MPLE estimates, and the lower triangle MLEs. $^{+}$ indicates significantly different from zero at the 0.05 level (two-sided) according to the bootstrap of the MPLE and the estimate of the variance of the MLE, $_{\times}$ indicates significance according to the logistic regression covariance matrix.



(a) Estimates of Justices' Ideal Points

(b) Significant Associations Between Justices

Figure 2: The bars in (a) span 95% confidence intervals. In (b) an edge is drawn if the association parameter has the respective sign in at least 95% of the bootstrap samples.

4 Conclusions

The nonparametric bootstrap can be used to construct consistent confidence intervals for the MPLE. In a Monte Carlo study, we show that this result applies at moderate finite sample sizes. Additionally, we provide an application where inference with the bootstrapped MPLE and MLE lead to practically equivalent conclusions about justices' voting behavior on U.S. Supreme Court cases.

References

[1] B. C. Arnold and D. Strauss. Pseudolikelihood estimation: Some examples. *Sankhya: The Indian Journal of Statistics, Series B*, 53(2):pp. 233–243, 1991.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[3] C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699, 1992.

[4] H. Geys, G. Molenberghs, and L. M. Ryan. Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94(447):734–745, 1999.

[5] A. Gunawardana and C. Meek. Tied boltzmann machines for cold start recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 19–26, 2008.

[6] M. S. Handcock. Assessing Degeneracy in Statistical Models of Social Networks. In *Workshop on Dynamic Social Network Analysis, Washington, DC, November*, 2003.

[7] S. Hanneke and E. P. Xing. Discrete temporal models of social networks. *The Electronic Journal of Statistics*, 4:585–605, 2010.

[8] A. Hyvarinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

[9] S. N. Lahiri. On bootstrapping m-estimators. *Sankhya: The Indian Journal of Statistics, Series A*, 54(2):pp. 157–170, 1992.

[10] A. D. Martin and K. M. Quinn. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 19531999. *Political Analysis*, 10(2):134–153, 2002.

[11] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

[12] M. A. van Duijn, K. J. Gile, and M. S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52 – 62, 2009.