Downloaded from orbit.dtu.dk on: Apr 27, 2024

DTU Library

A new likelihood inequality for models with latent variables

Olsen, Niels Aske Lundtorp

Published in:
Statistics and Probability Letters

Link to article, DOI:
10.1016/j.spl.2023.109998

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Olsen, N. A. L. (2024). A new likelihood inequality for models with latent variables. *Statistics and Probability Letters*, *206*, Article 109998. https://doi.org/10.1016/j.spl.2023.109998

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Statistics and Probability Letters 206 (2024) 109998

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A new likelihood inequality for models with latent variables

Niels Lundtorp Olsen

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

ARTICLE INFO

Keywords:
Statistical Inference
Latent variables
Model selection
Likelihood theory

ABSTRACT

Likelihood-based approaches are central in statistics and its applications, yet often challenging since likelihoods can be intractable. Many methods such as the EM algorithm have been developed to alleviate this.

We present a new likelihood inequality involving posterior distributions of a latent variable that holds under conditions similar to the EM algorithm. Potential scopes of the inequality includes maximum-likelihood estimation, likelihood ratios tests and model selection. We demonstrate the latter by performing selection in a non-linear mixed-model using MCMC.

1. Introduction

Likelihood is arguably the most important concept in statistics, formally introduced and popularized by Fisher (1922). Being the 'inverse probability', likelihood measures the goodness-of-fit for a given statistical model, and is thus central to statistical inference. Maximum-likelihood estimation is arguably the most used principle for statistical inference and is underpinned by much theory. However, due to the fact that solving the associated score equations is often infeasible, auxiliary methods have been developed specifically to facilitate maximization of the likelihood function, most notably the *Expectation–Maximization (EM) Algorithm* (Dempster et al., 1977) and derivatives of this, such as the ECM algorithm, (Meng and Rubin, 1993). We will refer to these as *EM-class algorithms*. An extensive treatise can be found in McLachlan and Krishnan (2007).

Likelihood is also used for *likelihood ratio-tests*, which are widely used in statistics and is theoretically justified by the Neyman–Pearson lemma (Neyman and Pearson, 1933).

Latent variables. A notable challenge in statistical inference is the presence of *latent* or *unobserved* variables. A latent variable W is characterized by the fact that it acts as part of the statistical model, but is not observed. In terms of evaluating the likelihood, this implies the presence of a sum (if W is discrete) or an integral (if W is continuous) in the likelihood expression. Integrals are notoriously difficult to evaluate, so other approaches are often needed.

Two popular approaches to handle latent variables in maximum likelihood estimation (sometimes in combination are Monte Carlo methods and EM-class algorithms. Monte Carlo methods sample from some distribution (typically that of w). By correctly aggregating the results, this approximates the value of the likelihood.

EM-class algorithms differ in sense that they do not compute the actual likelihood value (which is rarely of interest), but points to some value being more optimal, which eventually converges to a stationary point for the likelihood (ideally the maximum). Whereas EM-class algorithms are very useful when it comes to parameter estimation, they are not useful for model selection. Another drawback is the relatively slow convergence rate of the EM algorithm.

There exists a large literature on statistical methods for latent variables, a good overview including modelling and estimation methods can be found in Song (2007).

E-mail address: nalo@dtu.dk.

https://doi.org/10.1016/j.spl.2023.109998
Received 8 March 2023; Received in revised form 12 October 2023; Accepted 21 November 2023
Available online 29 November 2023
0167-7152/© 2023 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contribution of this article. In this article, we present a new inequality for likelihood in models involving latent variables, Similar to the EM algorithm, we do not compute the actual likelihood values, but use the conditional distribution of the latent variable.

There are three scopes of the inequality: Likelihood-based inference, model selection and likelihood ratio (tests). The conditional distributions can be approximated using MCMC methods, making implementation of e.g. model selection easy.

The remainder of the article is organized as follows: In Section 2 we present and proof the new inequality. In Section 3 we apply the theorem to a model selection problem, where posteriors are approximated by MCMC, and finally we discuss the results in Section 4.

2. Theorem

Suppose that we are given a statistical family consisting of an observation $Y \in \mathcal{Y}$, a latent variable $W \in \mathcal{W}$ and an unknown parameter θ in parameter space Θ.

Assume that the joint variable (Y, W) is dominated; that is $P_\theta((Y,W) \in A) = \int_A p_\theta(Y,W) d(\lambda \otimes \mu)(Y,W)$ for a measure λ on $\mathcal{Y}$ and μ on $\mathcal{W}$, and assume that for every θ, $p_\theta(Y,W)$ is non-zero for almost all W, Y.

Let $p_\theta(Y)$ be the marginal density for Y, and let $L(\theta) := p_\theta(Y), L(\theta, W) := p_\theta(Y,W)$ denote marginal and posterior likelihoods, respectively.

Theorem 1. *Let $\theta_1, \theta_2 \in \Theta$ Then $L(\theta_1) < L(\theta_2)$ if and only if the following inequality is true:*

$$\int \min\left(1, \frac{L(\theta_2, W)}{L(\theta_1, W)}\right) \mathrm{d}P_{\theta_1}(W|Y) > \int \min\left(1, \frac{L(\theta_1, W)}{L(\theta_2, W)}\right) \mathrm{d}P_{\theta_2}(W|Y), \tag{1}$$

where $P_\theta(W|Y)$ is the posterior distribution of W under θ given Y.

That is, by integrating the "truncated likelihood-ratios" under the posterior distributions, we can compare the likelihood of y under θ_1 and θ_2.

Proof. Let A denote the subset of W where $L(\theta_2, W) < L(\theta_1, W)$. First consider the left integral:

$$\begin{aligned}
&\int \min\left(1, \frac{L(\theta_2, W)}{L(\theta_1, W)}\right) \mathrm{d}P_{\theta_1}(W|Y) = \int \min\left(1, \frac{L(\theta_2, W)}{L(\theta_1, W)}\right) \frac{L(\theta_1, W)}{L(\theta_1)} \mathrm{d}\mu(W) = \\
&\quad \int_A \frac{L(\theta_2, W)}{L(\theta_1)} \mathrm{d}\mu(W) + \int_{A^c} \frac{L(\theta_1, W)}{L(\theta_1)} \mathrm{d}\mu(W) = \\
&\quad \frac{1}{L(\theta_1)} \int 1_A(W) L(\theta_2, W) + 1_{A^c}(W) L(\theta_1, W) \, \mathrm{d}\mu(W)
\end{aligned} \tag{2}$$

We get a similar result for the right integral with $L(\theta_1)$ replaced by $L(\theta_2)$. Now the theorem follows. □

If given two parameters or statistical models θ_1 and θ_2, we shall refer to the integral

$$\int \min\left(1, \frac{L(\theta_2, W)}{L(\theta_1, W)}\right) \mathrm{d}P_{\theta_1}(W|Y)$$

as the *truncated likelihood-ratio* (integral) of θ_2 wrt. θ_1.

Corollary 2. *From the proof of the theorem it follows that the likelihood-ratio is given by the truncated likelihood-ratios divided by each other:*

$$\frac{L(\theta_1)}{L(\theta_2)} = \frac{\int \min\left(1, \frac{L(\theta_1, W)}{L(\theta_2, W)}\right) \mathrm{d}P_{\theta_2}(W|Y)}{\int \min\left(1, \frac{L(\theta_2, W)}{L(\theta_1, W)}\right) \mathrm{d}P_{\theta_1}(W|Y)} \tag{3}$$

Remarks. Note that the truncated likelihood-ratio is numerically stable due to the upper limit of 1. We have been slightly restrictive for the ease of presentation: the non-zero property of $p_\theta(Y,W)$ can be relaxed somewhat, and we may also include prior probabilities for θ_1 and θ_2 in the inequality.

Scope of the theorem. Apart from its intrinsic value, the greatest benefit of Theorem 1 is the fact that $L(\theta_1, W)$ is typically easy and fast to calculate.

Thus, the presented result can be used as an algorithmic tool in likelihood-based statistics, wherever latent variables are present and we do not wish to evaluate the actual likelihood values. We identify three scopes for application, *maximum likelihood estimation*, *model selection* and *likelihood-ratio testing*.

2.1. Maximum likelihood estimation

For maximum-likelihood estimation, suppose that we have a current parameter estimate θ_0. Let θ_1 be a new proposal. According to Theorem 1 we can determine if $L(\theta_1) > L(\theta_0)$ by evaluating and comparing the truncated likelihoods. Assuming that we can evaluate the truncated likelihood, we can devise a (model-specific) algorithm for likelihood inference without actually computing the likelihood.

As noted in the introduction, there already exists a widely used algorithm, that shares the same basic characteristics, namely the EM algorithm. However, the EM algorithm is restricted by the fact that it cannot move "too far" from the current estimate, often giving slow convergence. There are potentially use cases where an algorithm based on Theorem 1 with suitably tailored parameter proposals might outperform the EM algorithm in terms of convergence. We leave this as future work.

2.2. Model selection

Assume that we have N different parameter values, $\theta_1, \dots, \theta_n$, corresponding to n groups, and a datum $y \in \mathcal{Y}$. We wish to assign y to the group having the highest likelihood, $\arg\max p_{\hat{\theta}_k}(y)$, this is the typical setting of model-based classification common to statistics and machine learning.

Using Theorem 1, we can devise an algorithm that calculates $\arg\max p_{\hat{\theta}_k}(y)$ without knowing the value of $p_{\hat{\theta}_k}(y)$:

Algorithm 1.

1. For $j, k = 1, \dots, N$, calculate the truncated likelihood-ratio d_{jk} by
$$d_{jk} = \int \min\left(1, \frac{L(\theta_k, W)}{L(\theta_j, W)}\right) \mathrm{d}P_{\theta_j}(W|y)$$
2. If it holds that
$$d_{jl} > d_{lj} \quad \text{for all } l$$
for some j, we conclude that θ_j has the highest likelihood, ie. $j = \arg\max p_{\hat{\theta}_k}(y)$.

However, finding the conditional distributions $P_{\theta_1}(W|y)$ often pose a challenge in itself, one solution is to use MCMC methods:

Algorithm 2.

1. Select K, the number of particles in the MCMC algorithm
2. For $j = 1, \dots, N$, run MCMC sampler to obtain K samples $w_{j,1}, \dots, w_{j,K}$ of the conditional distribution $(W|y_{\theta_j})$.
3. For $j, k = 1, \dots, N$, approximate the truncated likelihood-ratio d_{jk} by
$$d_{jk} = \frac{1}{K}\sum_{l=1}^{K} \min\left(1, \frac{p_{\theta_k}(y, w_{jl})}{p_{\theta_j}(y, w_{jl})}\right)$$
4. If it holds that
$$d_{jl} > d_{lj} \quad \text{for all } l$$
for some j, we conclude that θ_j has the highest likelihood, ie. $j = \arg\max p_{\hat{\theta}_k}(y)$.

Since the d_{jk} values are an approximation of the true truncated likelihoods, there is no guarantee that we can find a j with the latter property. However if $p_{\theta_1}(y), \dots, p_{\theta_K}(y)$ are distinct, an "optimal" j will eventually exist for increasing K.

Computational cost. The computational cost for model selection consists of two parts:

- A computational cost for approximating the posterior, this is $O(n)$
- A computational cost for pairwise comparison of likelihoods this is $O(n^2)$

In general, we expect the computational cost of approximating the posterior to dominate the computational cost of pairwise comparisons for small N, as the former typically would rely on potentially costly MCMC (cf. algorithm 2).

2.3. Likelihood ratios

Finally, we can use Corollary 2 to approximate likelihood-ratios and hence likelihood-ratio tests. The approximation by MCMC is

$$\frac{L(\theta_j)}{L(\theta_k)} \approx \frac{\sum_{l=1}^{K} \min\left(1, \frac{p_{\theta_k}(y, w_{jl})}{p_{\theta_j}(y, w_{jl})}\right)}{\sum_{l=1}^{K} \min\left(1, \frac{p_{\theta_j}(y, w_{kl})}{p_{\theta_k}(y, w_{kl})}\right)}$$

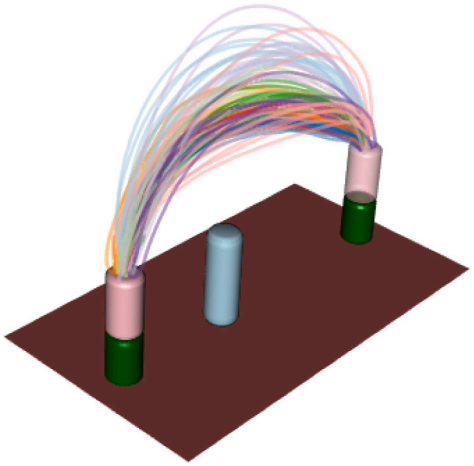

Fig. 1. Trajectories of arm movements. The movement is from left to right.

using the notation of Algorithm 2.

3. Example: Model selection for arm movement data

In this section we will demonstrate the theorem by applying it to a classification problem on a data set of arm movements from functional data analysis (Grimme, 2014). Here, the goal is to detect which of 10 people that performed a specific trajectory. Olsen et al. (2018) outlines the details of the classification experiment and also proposed a non-linear mixed-effects model that were superior compared to other methods. Trajectories for the 10 people can be seen in Fig. 1.

The classification method proposed by Olsen et al. (2018) is essentially a model selection problem.

Statistical model. A datum y consists of discrete observations from a trajectory $y : [0,1] \to \mathbb{R}^3$, which we model by

$$y(t) = \theta_j(v(t,w)) + x_n(t) \tag{4}$$

where x_n is a Gaussian process, θ_j is the "mean curve" of subject j and $v(,w)$ models the temporal deviation as a function of a latent Gaussian variable $w \in \mathbb{R}^7$.

The full-observation likelihood $p_j(y,w)$ can be described as a Gaussian probability:

$$p_j(y,w) = p_j^0(y|w)p_j^1(w)$$

where both p_j^0 and p_j^1 are Gaussian probability densities. However, the highly non-linear model (4) makes the likelihood for the observed data

$$p_j(y) = \int p_j^0(y|w)p_j^1(w)\,\mathrm{d}w$$

intractable, and so other methods are needed for estimation and model selection. Since estimation in (4) is not within the scope of this article, we refer to Olsen et al. (2018) and restrict our focus to model selection.

3.1. Model selection

To perform model selection for a new, unseen datum y, Olsen et al. (2018) used a Laplace approximation to approximate $L_j(y)$.

Here we instead demonstrate model selection using Algorithm 2. This does not require approximating the full likelihood $L_j(y)$, but only the conditional distribution $(w|y;j)$, for which an MCMC sampler can be used.

To do this, we sampled from the posteriors $(w|y_n,\theta_j)$ using an MCMC sampler, and compared these likelihoods using Theorem 1. In detail, let y_{test} be a datum from the test set. Then for each subject $j \in \{1,\dots,10\}$:

- Initialize $w_j = 0 \in \mathbb{R}^7$
- Run MCMC sampler to obtain 60 samples $w_{j,1},\dots,w_{j,60}$ of the posterior distribution $(w|y;p_j)$.
- For each pair of subjects $j,k = 1,\dots,10$, approximate the truncated likelihood-ratio d_{jk} by

$$d_{jk} = \frac{1}{60}\sum_{l=1}^{60} \min\left(1, \frac{p_k^0(y|w_{jl})p_k^1(w_{jl})}{p_j^0(y|w_{jl})p_j^1(w_{jl})}\right)$$

Table 1
Array of log-transformed truncated likelihood ratios $\log(d_{jk})$ when doing model selection for a datum. Bold indicates entries where $d_{jk} \geq d_{kj}$. Bottom row indicates the number of such occurrences for each column.

	1	2	3	4	5	6	7	8	9	10
1	**0.00**	**−0.94**	**0.00**	**0.00**	**−165.9**	**−0.13**	**−116.7**	**0.00**	−91.19	**−1.91**
2	−449.77	**0.00**	**−64.78**	**0.00**	−91.37	**−2.71**	**0.00**	−69.34	−255.8	−92.0
3	−346.0	−87.55	**0.00**	**−4.09**	−303.4	**−73.33**	**−145.1**	−57.96	−330.7	−284.4
4	−484.07	−106.9	−148.2	**0.00**	−178.1	−91.99	−79.58	−149.5	−328.0	−157.3
5	−551.7	**−14.17**	**−201.9**	**0.00**	**0.00**	**−0.42**	**0.00**	**−143.1**	−290.0	**−7.75**
6	−497.5	−46.24	−108.9	**0.00**	−81.10	**0.00**	**−0.79**	−109.6	−276.4	−91.65
7	−812.0	−123.9	−301.0	**−56.34**	−142.4	−94.79	**0.00**	−313.1	−470.2	−207.3
8	−243.7	**−65.24**	**−13.57**	**0.00**	−223.1	**−43.47**	**−134.7**	**0.00**	−236.5	**−115.8**
9	**−64.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
10	−431.3	**−66.7**	**−239.8**	**−0.15**	−52.98	**−23.82**	**−4.09**	−126.0	−291.9	**0.00**
#	2	6	7	**10**	3	8	9	4	1	5

This returns a 10×10 array of values d_{jk} for the selected datum. By comparing these values we can now classify y to the class with the highest likelihood, cf. Algorithm 2.

Results are shown in Table 1: here the datum was classified to category 4. The full "ordering" of models is given by:

$$L_4 > L_7 > L_6 > L_3 > L_2 > L_{10} > L_8 > L_5 > L_1 > L_9$$

where $L_j = p_j(y)$ is the likelihood corresponding to group j.

4. Discussion

With its general setting, the presented theorem is valid in a wide range of models, since latent variables are present in many classes of statistical models. The great benefit of the results lies in the fact that calculating "full" likelihood is often fast and easy to implement. We remark however, that the presented work is not a modelling tool, and we are still subject to robustness, misspecification and other aspects of statistical modelling.

As noted in the introduction, there are similarities to the EM-class algorithms, but also a notable difference: EM-class algorithms use only the posterior distribution of the current estimate, whereas applications of the presented theorem use two posterior distributions. With the presented theorem we are free to choose any proposal θ^* for an updated estimate, but we are not guaranteed a likelihood improvement as in the EM-class algorithms.

We expect the main application of the presented theorem to be model selection as demonstrated in Section 3. Apart from setting up the MCMC, the proposed solution is fairly plug-and-play and does not rely on any integral approximation for the likelihood, which may require considerable work in a practical setting and can be hard to assess

When using MCMC for approximating the posterior, as done in the presented example, this introduces some uncertainty in the comparison of the truncated likelihoods due to the randomness in the MCMC algorithms. Increasing the number of particles in the MCMC sampler would decrease this uncertainty at a price of increased computational cost. In the presented example we did pairwise comparisons of all 10 models, which required the evaluation of 100 truncated likelihood integrals. This was helpful in assuring consistency — the pairwise evaluations gave a consistent ordering of the models.

In the example we used a Metropolis–Hastings (MH) algorithm. The MH algorithm is arguably the most popular MCMC algorithm, but many other strong MCMC algorithms exist. A discussion of pros and cons is left for future work, we refer to Brooks et al. (2011) for a general discussion of MCMC algorithms.

We see this work as an addition to the statistician's toolbox, where it may be combined with a wide range of models and methods. Whereas it may not outperform specialized methods for classification (it is a purely algorithmic tool) or lead to fast and general optimization algorithms, it is easily implemented in an MCMC setting, which is one of the most common and versatile tools in statistics. The truncated likelihood ratio can potentially be combined with various tools of Bayesian methodology such as *INLA* (Rue et al., 2009) and *Approximate Bayesian computation*, we leave this for future work.

Data availability

Data sources can be found in the cited material.

Acknowledgements

I am grateful to Associate Professor Anders Stockmarr (Technical University of Denmark) for comments and inputs to the manuscript.

References

Brooks, S., Gelman, A., Jones, G., Meng, X.-L., 2011. Handbook of Markov Chain Monte Carlo. CRC press.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–22.
Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond. Ser. A 222 (594–604), 309–368.
Grimme, B., 2014. Analysis and Identification of Elementary Invariants as Building Blocks of Human Arm Movements (Ph.D. thesis). International Graduate School of Biosciences, Ruhr-Universität Bochum, (In German).
McLachlan, G.J., Krishnan, T., 2007. The EM Algorithm and Extensions. Vol. 382, John Wiley & Sons.
Meng, X.-L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.
Neyman, J., Pearson, E., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc..
Olsen, N.L., Markussen, B., Raket, L.L., 2018. Simultaneous inference for misaligned multivariate functional data. J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (5), 1147–1176.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc.: Ser. B 71 (2), 319–392.
Song, P., 2007. Correlated Data Analysis: Modeling, Analytics, and Applications. Springer.