Marginal likelihood

Patrick Breheny

November 17

Introduction

- In our previous lecture, we introduced the idea of conditioning in order to obtain a distribution free of nuisance parameters
- Today, our goal will also be to create a distribution free of nuisance parameters, although this time, we will be accomplishing that goal by (in one way or another) constructing a marginal distribution without nuisance parameters

Definition

- As in the previous lecture, suppose we can transform the data x into v and w
- We will again be factoring the likelihood, only this time it will be the marginal distribution that is free of nuisance parameters:

$$p(x|\boldsymbol{\theta}, \boldsymbol{\eta}) = p(v|\boldsymbol{\theta})p(w|v, \boldsymbol{\theta}, \boldsymbol{\eta});$$

  the first term, $L(\boldsymbol{\theta}) = p(v|\boldsymbol{\theta})$, is known as the *marginal likelihood*
- Note that this term is free of nuisance parameters and that, like the conditional likelihood, is a true likelihood, corresponding to an actual distribution of observed data

Example: Normal distribution

- As an example, suppose $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
- We have already seen that the (profile) MLE, $\frac{1}{n}\sum_i (x_i - \bar{x})^2$, is biased
- Consider instead the transformation

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- From ordinary normal distribution theory, we know that

$$(n-1)s^2 \sim \sigma^2 \chi^2_{n-1}$$

Example: Normal distribution (cont'd)

- This marginal likelihood is

$$\ell(\sigma^2) = -\frac{n-1}{2}\log\sigma^2 - \frac{(n-1)s^2}{2\sigma^2};$$

  thus $\hat{\sigma}^2 = s^2$, an unbiased estimate

- Note that $\bar{x} \sim \mathrm{N}(\mu, \sigma^2/n)$ and $\bar{x} \perp\!\!\!\perp s^2$, so in terms of likelihood, we have

$$L(\mu, \sigma^2) = L(\mu, \sigma^2|\bar{x})L(\sigma^2|s^2)$$

- As with conditional likelihood, there is the possibility that we are losing information by ignoring the first part of the likelihood

Remarks

- In this scenario, are we losing information? Does $\bar{x}$ contain any information about σ^2?
- Certainly, if we had a repeated sample with several means, this would tell us something about σ^2
- With a single sample, however, it is hard to see how $\bar{x}$ could tell us anything about σ^2

Neyman-Scott problem

- As another example, consider the Neyman-Scott problem: $Y_{i1}, Y_{i2} \sim \mathrm{N}(\mu_i, \sigma^2)$
- If we apply the transformation

$$v_i = (y_{i1} - y_{i2})/\sqrt{2},$$

  then $v_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, a marginal distribution that is free of the nuisance parameters μ_i
- The marginal log-likelihood is therefore

$$\ell(\sigma^2) \propto -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i v_i^2$$

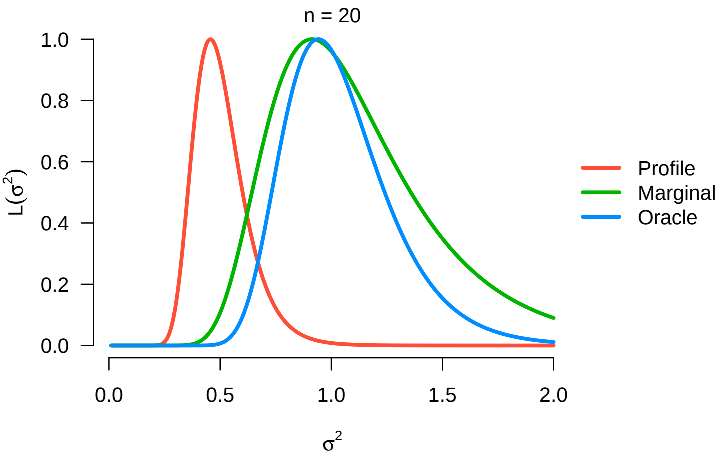
Marginal likelihood MLE

- The marginal likelihood therefore yields the estimate

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i v_i^2$$

- This is equal to RSS/n, the unbiased estimator from a classical ANOVA analysis
- Again, recall that the (profile) MLE was $\mathrm{RSS}/(2n)$, not only biased but inconsistent

Illustration



Information loss

- As the figure indicates, we are certainly losing information (compared to the oracle) by not knowing the μ_i parameters; indeed, the information loss is 50%
- A more fair comparison can be made between this marginal likelihood and a mixed model (more on these later) assuming that $\mu_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$
- In this case, it can be shown that the proportion of information lost is

$$\frac{1}{1 + (1 + 2\tau^2/\sigma^2)^2};$$

when $\tau^2 = \sigma^2$, this loss is 10%

REML

- Lastly, suppose we are fitting an ordinary linear regression model; as we have seen, the MLE for σ^2, RSS/n, is biased
- An alternative approach using marginal likelihood is to apply the transformation

$$\mathbf{v} = [\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y}$$

- The transformed data $\mathbf{v}$ has distribution $\mathrm{N}(\mathbf{0}, \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top])$, which is
    - Free of $\boldsymbol{\beta}$
    - Yields the marginal likelihood MLE

$$\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$$

- This is known as "restricted maximum likelihood" (REML)

Marginalization as a general technique

- Although possible to apply marginal likelihood in standard settings (as we have just done), its most common use is in "mixed" models
- Deriving marginal distributions from joint distributions is of course a standard tool in statistics:

$$p(x) = \int p(x, y)\, dy$$

- What we are attempting to do here, however, is to eliminate nuisance *parameters* by marginalizing

Marginalization and Bayesian statistics

- As we remarked in an earlier lecture, if the nuisance parameters have a distribution (as they do in Bayesian statistics), then standard tools apply
- Again, this is a major advantage of the Bayesian approach to inference . . . can it be applied outside of purely Bayesian frameworks?
- Indeed it can, if we are willing to treat the nuisance parameters not as parameters in the traditional frequentist sense, but as unobserved random variables

Mixed models

- In doing so, these unobserved random variables must be supplied with a distribution
- Obviously, this adds a layer of assumptions to our model, but without it, there is no way to integrate out the nuisance parameters
- Such a model, in which certain parameters are treated as unobserved random variables and others as unknown constants, is known as a "mixed" model

Motivating example

- Mixed models will be covered much more comprehensively in longitudinal data analysis (BIOS 7310), but we'll take a brief look at them here in order to see how marginal likelihood can be applied in general modeling settings
- Let's consider the model

$$y_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(\alpha_i + x_{ij}\beta, \sigma^2),$$

  and assume we are interested in estimating both β and σ
- Such a model might arise if there were repeated measurements on a subject, within a family, etc.
- As in the Neyman-Scott problem, the number of parameters is increasing with the sample size, which poses a challenge to maximum likelihood

Marginal likelihood

- How can we proceed with a marginal likelihood approach?
- In the case of linear models, we can use known properties of the multivariate normal distribution to work everything out in closed form
- Specifically, if we are willing to assume that $\alpha_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2)$, with $\{\alpha_i\}$ and the residual errors mutually independent, then we can write our model as

$$y_{ij} = \mu + x_{ij}\beta + \varepsilon_{ij},$$

where ε_{ij} has mean zero and variance $\sigma^2 + \tau^2$, as it incorporates both the between-group variability (from α_i) and the within-group variability

Correlation structure

- The ε_{ij} terms, however, are not independent, as the α_i term is shared across multiple observations
- This gives rise to the following correlation structure (assuming consecutive observations are paired):

$$\mathbb{V}\boldsymbol{\varepsilon} = \begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & 0 & 0 & \dots \\ \tau^2 & \sigma^2+\tau^2 & 0 & 0 & \dots \\ 0 & 0 & \sigma^2+\tau^2 & \tau^2 & \dots \\ 0 & 0 & \tau^2 & \sigma^2+\tau^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

- Marginally, we have $\mathbf{y} \sim \mathrm{N}(\mu + \mathbf{x}\beta, \mathbf{V})$, where $\mathbf{V} = \mathbb{V}\boldsymbol{\varepsilon}$

Estimation

- As we've seen in our homework assignment, however, we can estimate β in closed form regardless of what structure the variance has:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

  where $\mathbf{W} = \mathbf{V}^{-1}$

- This, of course, assumes that $\mathbf{V}$ is known
- In our case, the *structure* of $\mathbf{V}$ is known (or at least assumed), but the values of σ^2 and τ^2 are not
- Thus, in order to fit this model, we will need to proceed in an iterative fashion, updating $\boldsymbol{\beta}$ given τ^2 and σ^2, then updating τ^2 and σ^2 given $\boldsymbol{\beta}$, and so on

Competitors

- So, how well does this approach work?
- Let's introduce some competing ideas for how to analyze this data
- **Naïve:** Simply regress $\mathbf{y}$ on $\mathbf{x}$, don't even worry about α_i
- **Profile:** Ordinary least squares with all $n+2$ parameters ($\{\alpha_i\}_{i=1}^n$, β, and σ)
- **Oracle:** Gets to use the true $\{\alpha_i\}_{i=1}^n$ values
- **Differencing:** Analyze $v_i = y_{i1} - y_{i2}$, which causes the α_i term to cancel; note that this is also a marginal likelihood approach, but doesn't make any distributional assumptions about $\{\alpha_i\}_{i=1}^n$ (note that this is not so easily extended beyond the paired setting)

Results

I simulated $n = 100$ pairs of observations, with $\sigma^2 = \tau^2 = \beta = 1$:

	$\widehat{\beta}$	SE($\widehat{\beta}$)	$\hat{\sigma}^2$
Oracle	1.00	0.23	0.93
Marginal	0.89	0.29	0.98
Differencing	1.14	0.34	0.97
Profile	1.14	0.34	0.48
Naïve	0.66	0.33	1.89

Remarks

- In terms of estimating $\boldsymbol{\beta}$, all methods produce reasonable estimates (the naïve approach looks bad in this particular simulation, but it isn't biased)
- However, the marginal likelihood mixed model results in the most accurate (lowest SE) estimate, except for the oracle
- As we have seen, the profile likelihood approach substantially underestimates σ^2
- As we might expect, the naïve approach substantially overestimates σ^2; all other methods produce reasonable estimates

Changing the data generating process

- This looks very good for marginal likelihood – and indeed, it is a very effective and widely used approach in situations like this
- However, it is important to keep in mind that it comes at the expense of added assumptions that may or may not be true
- For example, we have assumed that the distribution of α_i is independent of x_{ij}
- However, what if $x_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(\alpha_i, 1)$?

Results, part 2

In this case, the mixed model's assumptions are wrong and the resulting coefficient estimate is biased (here, $n = 1,000$):

	$\widehat{\beta}$	SE($\widehat{\beta}$)	$\hat{\sigma}^2$
Oracle	1.00	0.02	0.98
Marginal	1.42	0.02	1.08
Differencing	1.04	0.03	0.94
Profile	1.04	0.03	0.47
Naïve	1.49	0.02	1.44

Introduction to nonlinear mixed models

- This same idea can be extended to nonlinear models as well
- The big difference, however, is that without the nice properties of the multivariate normal distribution, we cannot simply derive the marginal distribution in closed form
- Instead, we will have to rely on a numeric algorithm to approximate the integral

Non-quadrature approaches

- You should be somewhat familiar with this idea from Bayesian methods, as numeric integration is ubiquitous in Bayesian analysis
- Monte Carlo approaches are indeed one way to integrate out the random effects
- Another approach is the trapezoid rule, approximating the integral by breaking it up into a large number of little trapezoids

Gaussian quadrature

- However, a more widely used method for mixed models is something called Gaussian quadrature
- The basic idea of Gaussian quadrature is to approximate an integral with a weighted sum:

$$\int_a^b f(x)p(x)\,dx \approx \sum_{k=1}^{K} w_k f(z_k)$$

- The cleverness of Gaussian quadrature is to choose the weights $\{w_k\}$ and focal points (or "abscissas") $\{z_k\}$ so that this approximation is as accurate as possible

Brief theory of quadrature

- The theory of Gaussian quadrature, while rather elegant, is beyond the scope of this course
- Nevertheless, I'll share the result of one theorem (without proof) so that you can get a sense of how well it works
- **Theorem:** For any absolutely continuous distribution, there exist positive weights $\{w_k\}_{k=1}^K$ and points $\{z_k\}_{k=1}^K$ such that the quadrature formula is exact whenever f is a polynomial of degree $2K+1$ or lower.

Computation of points and weights

- Solving for these points and weights, of course, is not trivial, but for common probability distributions $p(x)$, the problem has already been solved by long-dead brilliant mathematicians
- Gauss-Legendre quadrature gives the points and weights for the uniform distribution, Gauss-Laguerre for the gamma distributions, Gauss-Jacobi the beta distribution, and so on
- The most widely used in statistics are the Gauss-Hermite polynomials, which correspond to the normal distribution
- Several R packages provide these points and weights; I tend to use `GHrule` from the `lme4` package

Example: Variance of the median

- If $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, with n odd, the sample median has density

$$p(x) = \frac{n!}{m!m!}\Phi(x)^m\{1-\Phi(x)\}^m\phi(x),$$

  where $m = (n-1)/2$
- By symmetry, the expected value of the median is zero, but the variance is not easy to calculate
- This is therefore a natural candidate for a numerical method such as quadrature:

$$\begin{aligned}\mathbb{V}X_{(m+1)} &= \int x^2 p(x)\,dx = \int f(x)\phi(x)\,dx \\ &\approx \sum_{k=1}^{K} w_k f(x_k)\end{aligned}$$

Results

- We could also approximate this result with Monte Carlo integration (simulate a sample of normal variables, take the median, repeat thousands of times, and calculate the variance) or with asymptotic theory, which says that the variance should be about $\pi/(2n)$
- Results for $n = 11$:

	Variance
Monte Carlo ($N = 100,000$)	0.1368
Asymptotic	0.1428
Gauss-Hermite ($K = 20$)	0.1476
Gauss-Hermite ($K = 100$)	0.1372

A mixed effects logistic regression

- To see how this works in statistical modeling, let's consider the binary analog of our earlier model:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \mu + x_{ij}\beta + \alpha_i,$$

where again we will assume that $\alpha_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$

- Letting $\alpha_i = \tau w_i$, $w_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, the marginal likelihood is

$$\begin{aligned} L(\beta, \mu, \tau^2) &= \prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{m_i} p(y_{ij}|x_{ij}, \alpha_i, \beta, \mu) \right\} p(\alpha_i|\tau^2)\, d\alpha_i \\ &= \prod_{i=1}^{n} \int \exp\left\{ \sum_{j=1}^{m_i} \log p(y_{ij}|x_{ij}, \tau w_i, \beta, \mu) \right\} \phi(w_i)\, dw_i \end{aligned}$$

Approximate marginal likelihood

- Having now written the integral in the form $\int f(x)\phi(x)\,dx$, we can apply Gauss-Hermite quadrature:

$$L(\beta,\mu,\tau^2) \approx \prod_{i=1}^{n} \sum_{k=1}^{K} w_k \exp\left\{ \sum_{j=1}^{m_i} \log p(y_{ij}|x_{ij},\tau z_k,\beta,\mu) \right\}$$

- We now have the likelihood in a form that, while not necessarily simple, is at least manageable in terms of taking gradients to find the score and information
- This method is implemented in various software packages such as `glmer` in R and `PROC GLIMMIX` in SAS, although there are a variety of other numeric approximations available

Simulation case study

- As we did with the linear models, let's compare this marginal likelihood approach with some other plausible ways of analyzing this data
- **Naïve:** As before, ignore the α_i effects completely and just fit a standard logistic regression
- **Profile:** As before, fit a standard logistic regression with $n + 1$ parameters
- **Conditional:** The method we derived in the previous lecture, where we form a conditional likelihood from pairs such that $y_{i1} + y_{i2} = 1$

Results

Simulation case study results ($n = 1,000$):

	$\widehat{\beta}$	SE
Naïve	0.82	0.05
Profile	2.17	0.16
Conditional	1.09	0.11
Marginal	0.93	0.07

As before, the data were simulated with $\beta = \tau^2 = 1$

Remarks

- As we would expect from our earlier analytical look at this problem, the profile MLE is biased upwards, while the naïve MLE is biased downward (somewhat)
- The conditional and marginal likelihood approaches both look reasonable, although as before, the marginal likelihood mixed model has a somewhat smaller SE (primarily due to making stronger assumptions, of course)