Marginal likelihood

Patrick Breheny

November 17, 2025

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

- The classical approach to marginal regression is rather similar to conditional likelihood
- As in the previous lecture, suppose we can derive statistics v and w such that the likelihood can be factored into a marginal distribution of w and a conditional distribution of $v \mid w$
- However, now it will be the marginal distribution that is free of nuisance parameters:

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

  the first term, $L_m(\boldsymbol{\theta}) = p(w \mid \boldsymbol{\theta})$, 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

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

- As with conditional likelihood, there is the possibility that we are losing information by ignoring the first part of the likelihood... with a single sample, however, it is hard to see how $\bar{x}$ could tell us much about σ^2

REML

- Another example: when fitting an ordinary linear regression model, the MLE for σ^2, RSS/n, is biased
- Alternatively, we could apply the transformation

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

- The marginal distribution is

$$\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top])$$

  - Free of $\boldsymbol{\beta}$
  - Yields the MLE $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$
- This is known the "restricted maximum likelihood" (REML) estimate, although the name is slightly misleading

Marginalization as a general technique

- Unlike conditional likelihood, however, marginal likelihood can be applied widely, even in settings without a factorization based on sufficient statistics
- In probability, we routinely eliminate random variables from joint distributions through integration:

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

- In likelihood theory, we can eliminate nuisance parameters in the same way; this is known as *marginal likelihood* or *integrated likelihood*

Marginalization and Bayesian statistics

- As we remarked in an earlier lecture, if the nuisance parameters have a distribution (as they do in Bayesian statistics), they can always be integrated out
- This is a major advantage of the Bayesian approach to inference
- Our primary focus today will be on extending these ideas to frequentist inference, but integrated likelihoods are important in Bayesian inference as well – integrating out nuisance parameters reduces the dimension of MCMC and often greatly improves efficiency

Mixed models

- Marginal likelihoods are also useful outside the Bayesian framework if we are willing to treat nuisance parameters as unobserved random variables, not fixed constants
- To do so, the nuisance parameters must be supplied with a distribution (note that this adds a layer of assumptions to our model)
- 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 more comprehensively in other courses, 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 $\boldsymbol{\beta}$ 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
- **Conditional:** See previous lecture (note that this approach eliminates nuisance parameters, but does not make additional distributional assumptions about $\{\alpha_i\}_{i=1}^n$; this works nicely in the paired setting but doesn't generalize easily)

Results

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

method	BetaAvg	BetaRMSE	Variance
Profile	1.02	0.35	0.49
Mixed	1.01	0.30	1.00
Oracle	1.01	0.24	0.99
Naive	1.00	0.36	1.99
Conditional	1.02	0.35	0.99

Remarks

- In terms of estimating β, all methods produce reasonable estimates
- However, the marginal likelihood mixed model results in the most accurate estimate (except for the oracle)
- With respect to estimating σ^2:
  - The profile likelihood approach substantially underestimates (we've seen this already)
  - The naïve approach substantially overestimates (this makes sense)
  - 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:

method	BetaAvg	BetaRMSE	Variance
Profile	1.00	0.10	0.50
Mixed	1.44	0.45	1.20
Oracle	1.00	0.05	0.99
Naive	1.50	0.51	1.50
Conditional	1.00	0.10	0.99

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

- There are many ways to do this, which you may be familiar with from Bayesian statistics
- 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'll 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:

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

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 ($N = 100,000$)	0.1367
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 $\alpha_i = \tau a_i$, $a_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, the marginal likelihood is

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

Approximate marginal likelihood

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

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

- 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

Gaussian quadrature in software

- Quadrature is the most accurate method for integrated likelihood when the random-effects dimension is small (typically 1–3)
- Unfortunately, the number of points required increases exponentially with dimension (K^d), so it doesn't scale well to high-dimensional random effects (Laplace approximations are used here instead)
- For example, `lme4::glmer()` uses adaptive Gaussian quadrature (AGQ) for random-intercept models, but falls back to Laplace when multiple random effects appear

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 ($n = 100$):

method	Mean	RMSE
Naive	0.63	0.40
Profile	2.24	1.60
Conditional	1.12	0.52
Marginal	0.96	0.31

Data were simulated with $\beta = 1; \tau^2 = 4$; 1,000 independent replications

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
- 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)