Astrostatistics Notes

Autumn Mapes

Fall 2025

1 Fundamentals

Definition of CDF

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

Card ID: 1775594407540

Easiest way to find the distribution of the sum of two random variables

Use the property that the MGF of a sum of RVs is the product of their MGFs:

$$\phi_{X+Y}(t) = \phi_X(t) * \phi_Y(t)$$

Card ID: 1770679535305

MGF of the Univariate Normal RV

$$\phi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$$

Card ID: 1770679535300

Def: Characteristic Function (Statistics)

$$\varphi_X(t) = \mathbb{E}(e^{itX})$$

Card ID: 1775594407544

Def: Precision

The inverse of variance, often σ^{-2} or Σ^{-1}

Card ID: 1778015739639

Def: Posterior of two multivariate normals, one determining the mean of the other

(same as proportional product of two normal pdfs)

Given $N(\vec{\mu}_1, \Sigma_1)$ and $N(\vec{\mu}_2, \Sigma_2)$, the product of their pdfs is a new $N(\mu^*, \Sigma^*)$ with

$$\Sigma^* = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$$

$$\mu^* = \Sigma^* \left(\Sigma_1^{-1}\vec{\mu}_1 + \Sigma_2^{-1}\vec{\mu}_2\right)$$

For the posterior result, just use $\mu_2 := A$.

The precisions just sum. The mean is the precision-weighted average.

Card ID: 1778015739641

Def: Marginalization

$$P(x) = \int P(x, y)dy = \int P(x \mid y)P(y)dy = \mathbb{E}_y[P(x \mid y)]$$

Card ID: 1779828180693

2 Basic Stats Review

Def: The Delta Method

Just a second-order Taylor approximation with a statistics coat of paint:

$$\mathbb{E}[g(X)] \approx g(\mathbb{E}[X]) + \frac{1}{2}g''(\mathbb{E}[X])\text{Var}(X)$$

Card ID: 1775594407546

Variance of the sample mean

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

Card ID: 1770679535308

Def: Fisher Information Matrix

$$\big[\mathcal{I}(\theta)\big]_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\;\middle|\;\theta\right]$$

Or, equivalently, given some regularity conditions,

$$\big[\mathcal{I}(\theta)\big]_{i,j} = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X;\theta)\;\middle|\;\theta\right]$$

Card ID: 1770679535310

Def: Univariate Unbiased Cramer-Rao Lower Bound

Given an unbiased estimator T for θ, we have:

$$\mathrm{Var}(T) \geq \frac{1}{I(\theta)}$$

(this is achieved by unbiased MLEs asymptotically)

Card ID: 1770679535312

Def: Univariate General Cramer-Rao Lower Bound

Given any estimator T for θ, we have:

$$\mathrm{Var}(T) \geq \frac{\left(1 + \frac{d}{d\theta}\mathrm{Bias}(T)\right)^2}{I(\theta)}$$

Card ID: 1775594407549

Def: Multivariate Cramer-Rao Lower Bound

For any unbiased estimator $\vec{T}$ of $\vec{\theta}$,

$$\text{Cov}(\vec{T}) - I(\theta)^{-1}$$

must be positive-semi-definite. In particular, this means that

$$\text{Var}(T_i) \geq [I(\vec{\theta})^{-1}]_{ii}$$

for any index i.

Card ID: 1770679535313

How do you solve a problem with selection effects?

Use Bayes' Theorem conditioning on the selection function (usually a step function), should end up with a normal CDF normalizing factor

Card ID: 1775594407550

Def: Split Multivariate Gaussian Marginals

Suppose we have a partioned multivariate Gaussian like

$$\boldsymbol{f} = \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_V \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UV} \\ \boldsymbol{\Sigma}_{VU} & \boldsymbol{\Sigma}_V \end{bmatrix} \right)$$

Then its marginals are exactly what we'd expect:

$$P(U) = \int P(U,V)dV = N(U \,|\, \mu_U, \Sigma_U)$$

$$P(V) = \int P(U,V)dU = N(V \,|\, \mu_V, \Sigma_V)$$

Card ID: 1779828180698

Def: Split Multivariate Gaussian Conditional Probabilities

$$U \mid V \sim N(\mathbb{E}[U \mid V], \mathrm{Var}[U \mid V])$$

$$\mathbb{E}[U \mid V] = \mu_U + \Sigma_{UV}\Sigma_V^{-1}(V - \mu_V)$$

$$\mathrm{Var}[U \mid V] = \Sigma_U - \Sigma_{UV}\Sigma_V^{-1}\Sigma_{VU}$$

Card ID: 1779828180699

How can we construct a joint probability from

$$V \sim N(V_0, \Sigma_V)$$

and

$$U \mid V \sim N(U_0 + XV, \Sigma_{U \mid V})$$

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{U}_0 + \boldsymbol{X}\boldsymbol{V}_0 \\ \boldsymbol{V}_0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{X}\boldsymbol{\Sigma}_V\boldsymbol{X}^T + \boldsymbol{\Sigma}_{U|V} & \boldsymbol{X}\boldsymbol{\Sigma}_V \\ \boldsymbol{\Sigma}_V\boldsymbol{X}^T & \boldsymbol{\Sigma}_V \end{pmatrix}\right)$$

Card ID: 1779828180701

Def: MLE and Unbiased Estimator for Normal RV σ

MLE:

$$\hat{\sigma}^2 = \frac{(x_i - \hat{\mu})^2}{N}$$

Unbiased Estimator:

$$\hat{\sigma}^2 = \frac{(x_i - \hat{\mu})^2}{N - 1}$$

where $\hat{\mu} = \bar{x}$.

Card ID: 1779828180703

Def: Inverse of a 2x2 Matrix

For a 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

its inverse is:

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

To invert, swap the entries on the diagonal, negate the off-diagonal, and divide everything by the determinant.

Card ID: 1780146565460

3 Bayesian Statistics and MCMC

3.1 Conceptuals

State the Frequentist model of statistics (in the context of a regression)

We have some true variables we want to compute, but we can only view data that is confounded with some source of error (following a distribution)

So the underlying (latent) variables we seek to understand are deterministic, there's just a bunch of stochastic error complicating matters.

Card ID: 1775594407552

State the Bayesian model of statistics (in the context of a regression)

Interpret probability as the degree of certainty in an event rather than its long-run chance. We start with some guess as to how we think the parameter is distributed, and use the data to update that prior into a more accurate posterior.

The parameters are themselves random variables! We estimate their distribution, not just a single value.

Card ID: 1775594407553

3.2 Conditioning Probabilities

TODO

3.3 Bayesian Approximation

Def: Monte Carlo Integration

for integrating over some interval I

$$\hat{I} = \frac{1}{m} \sum_{i=1}^{m} f(\theta_i)$$

where $f(\theta)$ is some data processing to estimate some statistic (mean, variance, etc)

Card ID: 1775594407554

Monte Carlo integrator for the posterior mean

$$f(\theta) = \theta$$

Card ID: 1775594407555

Monte Carlo integrator for the posterior variance

$$f(\theta) = (\theta - \mathbb{E}[\theta \mid D])^2$$

Card ID: 1775594407556

Monte Carlo integrator in an interval $[a, b]$

$f(\theta) = I_{[a,b]}(\theta)$

Card ID: 1775594407557

Variance of a Monte Carlo Integrator

$$\mathrm{Var}(\hat{I}) = \frac{1}{m}\mathrm{Var}[f(\theta)]$$

Note that this is independent of the dimension of θ!
The Monte Carlo error is just the square root of this (the standard deviation)

Note that this can also be used to estimate the sample variance:

$$\hat{\mathrm{Var}}(\{f(\theta_i)\}) = \frac{1}{m-1}\sum_{i=1}^{m}(f(\theta_i) - \hat{I})$$

Card ID: 1775594407559

Def: Direct Sampling method for estimating a posterior distribution

In the case where a posterior distribution can be broken down into the product of named distributions, simply draw from one distribution, feed into the next, etc, and do this to get a bunch of samples.

Note that it's easy to get the marginals this way: just ignore all but one value!

Card ID: 1775594407560

Def: Kernel Density Estimation

Creates a smooth histogram from a bunch of samples θ_i:

$$\hat{P}(\theta \mid D) = \frac{1}{m}\sum_{i=1}^{m} N(\theta \mid D_i, b_w^2)$$

where b_w is the bandwidth.

Card ID: 1775594407561

Silverman's Rule of Thumb

$$b_w = \left(\frac{4\hat{\sigma}^5}{3m}\right)^{1/5}$$

where σ is the estimated sample standard deviation, $\hat{\sigma}^2 = \hat{\mathrm{Var}}(\{\theta_i\})$

Card ID: 1775594407562

3.4 Markov Chain Monte Carlo

Def: The Metropolis Algorithm

1. Pick some μ_0
2. Choose a new $\mu_{\text{prop}} \sim \text{N}(\mu_{\text{prev}}, \tau^2)$

Note: the jump distribution has to be symmetric: $J(a|b) = J(b|a)$

3. Accept with probability $\min(1, r)$, $r = \frac{\mathbb{P}(\mu_{\text{prop}} \mid \vec{y})}{\mathbb{P}(\mu_{\text{prev}} \mid \vec{y})}$
4. Repeat 2 and 3 until you have enough samples.

This works for higher dimensions as well, just draw from a multivariate Gaussian jump distribution.

Card ID: 1775594407563

State two post-processing strategies to properly format MCMC results

---

1. Get rid of burn-in: throw out about the first 20 percent of your samples
2. Thinning: Throw out every other sample

Card ID: 1775594407565

Def: Stationary Distribution

Given some distribution P and some random process kernel T:

$$\sum_x P(x)T(x \to x') = P(x')$$

Card ID: 1775594407566

Def: Detailed Balance

and what it implies

Given some distribution P and some random process kernel T:

$$P(x)T(x \to x') = P(x')T(x' \to x)$$

This implies that P is a stationary distribution: just sum both sides over x!

Card ID: 1775594407568

Def: Metropolis-Hastings

1. Pick some μ_0
2. Choose a new $\mu_{\text{prop}} \sim J(\mu_{\text{prev}}, \tau^2)$

Note: the jump distribution need not be symmetric: $J(a|b) \neq J(b|a)$

3. Accept with probability $\min(1, r)$, $r = \frac{\mathbb{P}(\mu_{\text{prop}} \mid \vec{y}) / J(\mu_{\text{prop}} \mid \mu_{\text{prev}})}{\mathbb{P}(\mu_{\text{prev}} \mid \vec{y}) / J(\mu_{\text{prev}} \mid \mu_{\text{prop}})}$
4. Repeat 2 and 3 until you have enough samples.

Card ID: 1775594407569

Def: Gibbs Sampling

Assumes that we have marginal distributions along every parameter alone.

1. Pick some $\vec{\mu}_0$
2. Run one Gibbs cycle:

for each $i \in [d]$, set $\mu_{next,i} \sim \mathbb{P}(\mu_{prev,i} \mid \mu_{prev,-i}, y)$

4. Repeat 2 until you have enough samples.

Card ID: 1775594407571

Def: Metropolis-Within-Gibbs

If we can't express the conditional distributions necessary to run Gibbs sampling, estimate them by attempting to take a Metropolis step

Card ID: 1775594407573

How do we choose a good jump function when tuning the Metropolis algorithm?

We can use Laplace approximation to show that the best choice is $c^2 A^{-1}$, where A is the Hessian of the log-posterior and $c \approx \frac{2.4}{\sqrt{\dim \theta}}$

Card ID: 1775594407574

What acceptance rate should we aim for when using the Metropolis algorithm?

---

44 percent in one dimension, 23 percent in dimensions greater than 5

Card ID: 1775594407576

How can we compare the convergence of two different MCMC methods?

Check that your Gelman-Rubin ratio is about 1.
Check the autocorrelation timescale, compare effective sample sizes.

Card ID: 1775594407577

Explain mixed samplers and parameter blocking

Mixed samplers describe the process of choosing different means of sampling different parameters of the function we want to sample from.

Parameter blocking refers to updating some parameters together (in blocks), especially ones that are highly correlated (to prevent zig-zagging during Gibbs, for instance).

Card ID: 1775594407579

4 Gaussian Processes

Def: Gaussian Process

Time-dependent Gaussian distribution

$$f(t) \sim GP(m(t), K_{A,\tau^2}(t, t'))$$

Card ID: 1779828180715

Def: Squared Exponential Kernel

$$k(t, t') = A^2 \exp(-|t - t'|^2/\tau^2)$$

Card ID: 1779828180716

Def: Stationary Gaussian Process

$$K(t, t') = K(t + a, t' + a)$$

for all a

Card ID: 1779828180717

5 Graphical Modelling and Hierarchical Bayes

Def: How to draw probabilistic graphical models

Open circle / square nodes: latent parameter, unobserved data
Shaded nodes: Something we condition on (ie data)
Filled dot: Known constant
Plate: iid replications of what's inside

Card ID: 1779828180719

TODO some stuff on PGMs and deriving conditional independence expressions?

Def: Shrinkage Estimator

The idea is to bias the individual estimates towards the population mean to reduce the final MSE.
This is equivalent to just using Hierarchical Bayes with τ^2, where you estimate τ^2 given data first, then consider that for your other estimators.

Card ID: 1779648617570

6 Model Comparison

Def: Posterior Odds Ratio

and the Bayes factor

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)}{P(D \mid M_2)} \frac{P(M_1)}{P(M_2)}$$

M_1 and M_2 are our models. The first multiplicand is the Bayes factor, the second multiplicand is the prior odds of one model versus the other.

Card ID: 1779648617572

Def: Evidence

$$P(D \mid M_1) = \int_\Theta P(D \mid \theta, M_1) P(\theta \mid M_1) d\theta$$

also called the marginal likelihood.

Note that $P(\theta \mid M_1)$ needs to be a proper prior!

Card ID: 1779648617573

Def: The Jefferies Scale

Let α be a placeholder for the natural log of the Bayes factor.
$\alpha < 1$: Terrible
$1 < \alpha < 2.5$: Weak but significant
$2.5 < \alpha < 5$: Significant
$5 < \alpha$: Decisive

Card ID: 1779648617574

Def: Harmonic Mean Estimator

$$P(D) \approx \left[\frac{1}{M} \sum_{i=1}^{M} P(D \mid \theta_i)^{-1} \right]^{-1}$$

This is unstable in practice

Card ID: 1779648617575

Def: Occam Factor

and how to derive it

Define $g(\theta; D, M) = g(\theta) = P(\theta \mid D, M)P(D \mid M) = P(D \mid \theta, M)P(\theta \mid M)$.
Find a MAP estimate $\theta_0 = \underset{\Theta}{\text{argmax}} \ln(g(\theta))$, use a second order Taylor expansion for $\ln(g(\theta))$ then exponentiate to get:

$$g(\theta) \approx g(\theta) \exp(-\frac{1}{2}(\theta - \theta_0)^{\mathrm{T}} A(\theta - \theta_0))$$

where A is the Hessian of the log of g at the MAP
From there see that

$$\int g(\theta) d\theta = \int P(\theta \mid D, M)P(D \mid M) d\theta = P(D \mid M) \int P(\theta \mid D, M) d\theta = P(D \mid M)$$

So we have:

$$P(D|M) \approx g(\theta_{MAP}) \det(A/2\pi)^{-1/2} = P(D \mid \theta_0)P(\theta_0) \det(A/2\pi)^{-1/2}$$

$P(\theta_0) \det(A/2\pi)^{-1/2}$ is called the Occam factor.

Card ID: 1779828180723

Def: Savage-Dickey Ratio

When we are comparing nested models, it's possible to simplify the Bayes factor to

$$\frac{P(\psi \mid D, M_1)}{P(\psi \mid M_1)}$$

where $\psi = 0$.

Card ID: 1779648617578

Def: Bayesian Model Averaging

Bayesian model comparison between many models, not just two candidates.

Card ID: 1779648617580

Goal of Nested Sampling

We need to be able to evaluate the likelihood $L(\theta) := P(D \mid \theta, M)$ and the prior $\pi(\theta) := P(\theta \mid M)$, we want to approximate the integral

$$Z = \int P(D \mid \theta, M) P(\theta | M) d\theta = \int L(\theta) \pi(\theta) d\theta$$

(remember the prior is diffuse, and the likelihood is peaked, which makes this problem difficult)

Card ID: 1780146565496

Def: Nested Sampling

1. Take N_{live} points and their likelihoods, start with $Z = 0$.
2. Kill the point with the smallest likelihood, call it L_i^* for step i
3. (trivial) Compute the shrinkage factor $t_i = \frac{X_i}{X_{i-1}} \approx e^{-1/N_{live}}$ (approximate with the mean of the beta-distribution Beta($\mathrm{N_{live}}, 1$))
4. Accumulate evidence:

$$\Delta Z = L_i^*(X_{i-1} - X_i) = L_i^*(1 - t_i)X_{i-1}$$

5. Sample new point with likelihood greater than L_i^*
6. Repeat steps 2-5 until convergence
7. Add information about remaining points to the evidence:

$$\Delta Z = \bar{L}X_{end} = \bar{L}e^{-m/N_{live}}$$

where $\bar{L}$ is the average likelihood of remaining points, and m is the number of steps taken so far.
8. Compute the weighted posterior samples:

$$w_i = \frac{L_i^*(1 - t_i)X_{i-1}}{Z}$$

Card ID: 1780146565497

Def: Variational Inference

Idea: approximate the evidence with a distribution that is 'close' (measured by Kullback-Leibler divergence)

Card ID: 1780146565500

Def: Kullback-Leibler Divergence

$$KL(q(\theta) \,||\, p(\theta)) = -\int q(\theta) \ln\left(\frac{p(\theta)}{q(\theta)}\right) d\theta$$

Nonnegative and zero only when $p(\theta) = q(\theta)$, but NOT symmetric and so not a metric!

Card ID: 1780146565502

Goal of ELBO

We want to find a q that minimizes

$$KL(q(\theta) \,||\, P(\theta \,|\, D))$$

(approximate our evidence with a 'close' q distribution)

Card ID: 1780945172290

Def: ELBO

and how to derive it

Assume we want to minimize $KL(q(\theta) \,||\, P(\theta \,|\, D))$. Break down the KL divergence to extract the $p(D)$ piece, yielding

$$KL(q(\theta) \,||\, P(\theta \,|\, D)) + \mathbb{E}_q[\ln(q(\theta)) - \ln(P(D \,|\, \theta)P(\theta))] = \ln(P(D))$$

Call the expectation bit the Evidence (Log) Lower Bound (ELBO). If we want to minimize the KL divergence, we have to minimize it (can't ever do better than $\ln(P(D))$)

Card ID: 1780146565504

7 Appendix: Some Important Astronomy Equations

Def: The Apparent Magnitude Equation

$$m = M + \mu$$

where m is the true apparent magnitude, M is the absolute magnitude, and μ is the distance modulus $\mu = 25 + 5\log_{10}(d)$.

Card ID: 1780667434372

Def: Parallax Equation

$$\frac{\omega}{\text{arcsec}} = \frac{\text{parsec}}{d}$$

where ω is the true parallax angle and d is the distance to the star.
(note this uses small angle approximations, is not valid for large ω)

Card ID: 1780667434384