

Physics 403

Parameter Estimation, Correlations, and Error Bars

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Announcements

Midterm:

- ▶ In class, Monday Feb. 29
- ▶ The exam will count for 20% of your final grade
- ▶ Material: basic rules of probability, common distributions, PDF/likelihood maximization and estimators
- ▶ Properties of PDFs: marginalization, transformation rules, etc.
- ▶ No numerical component, obviously, but you may get a conceptual question about it

Reading:

- ▶ Sivia, Ch. 3.2 and 3.3

Table of Contents

- 1 Review of Last Class
 - Best Estimates and Reliability
 - Properties of a Good Estimator
- 2 Parameter Estimation in Multiple Dimensions
 - Return of the Quadratic Approximation
 - The Hessian Matrix and its Geometrical Interpretation
 - Maximum of the Quadratic Form
 - Covariance
- 3 Multidimensional Estimators
 - Gaussian Mean and Width
 - Student- t Distribution
 - χ^2 Distribution

Best Estimates and Reliability

- ▶ Identify the best estimator \hat{x} of a PDF by **maximizing** $p(x|D, I)$:

$$\left. \frac{dp}{dx} \right|_{\hat{x}} = 0, \quad \left. \frac{d^2p}{dx^2} \right|_{\hat{x}} < 0$$

- ▶ We assessed the reliability of the estimator by **Taylor expanding** $L = \ln p$ about the best value and found that

$$\hat{\sigma}^2 = \left(- \left. \frac{d^2L}{dx^2} \right|_{\hat{x}} \right)^{-1}$$

- ▶ This only works when the **quadratic approximation** is reasonable
- ▶ For an **asymmetric PDF**, it's better to use a confidence interval when reporting the reliability of an estimate
- ▶ For a **multimodal PDF**, summarizing the PDF with an estimator is not very well defined or useful

Example Estimators from Last Class

- ▶ Best estimator of binomial probability p (n successes in N trials):

$$\hat{p} = \frac{n}{N}, \quad \hat{\sigma}^2 = \frac{n(N-n)}{N^3} = \frac{\hat{p}(1-\hat{p})}{N}$$

- ▶ **Arithmetic mean**: best estimator of Gaussian with known variance σ^2 :

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \hat{\sigma}^2 = \frac{\sigma^2}{N}$$

- ▶ **Weighted mean**: best estimator of Gaussian with different error bars:

$$\hat{\mu} = \frac{\sum_{i=1}^N x_i w_i}{\sum_{i=1}^N w_i}, \quad \hat{\sigma}^2 = \frac{1}{\sum_{i=1}^N w_i}, \quad w_i = 1/\sigma_i^2$$

Caution: don't confuse width of distribution with uncertainty on the mean

Properties of a Good Estimator (for Frequentists)

A good estimator should be:

1. **Consistent.** The estimate tends toward the **true value** with more data:

$$\lim_{N \rightarrow \infty} \hat{\theta} = \theta$$

2. **Unbiased.** The expectation value is equal to the true value:

$$b = \langle \hat{\theta} \rangle - \theta = \int d\mathbf{x} p(\mathbf{x}|\theta) \hat{\theta}(\mathbf{x}) - \theta = 0$$

3. **Efficient.** The variance of the estimator is as small as possible (minimum variance bound, to be discussed):

$$\begin{aligned} \text{var}(\hat{\theta}) &= \int d\mathbf{x} p(\mathbf{x}|\theta) (\hat{\theta}(\mathbf{x}) - \hat{\theta})^2 \\ \text{MSE} &= \langle (\hat{\theta} - \theta)^2 \rangle = \text{var}(\hat{\theta}) + b^2 \end{aligned}$$

Table of Contents

- 1 Review of Last Class
 - Best Estimates and Reliability
 - Properties of a Good Estimator
- 2 Parameter Estimation in Multiple Dimensions
 - Return of the Quadratic Approximation
 - The Hessian Matrix and its Geometrical Interpretation
 - Maximum of the Quadratic Form
 - Covariance
- 3 Multidimensional Estimators
 - Gaussian Mean and Width
 - Student- t Distribution
 - χ^2 Distribution

Parameter Estimation in Higher Dimensions

- ▶ Moving to more dimensions:

$$x \rightarrow \mathbf{x}, \quad p(x|D, I) \rightarrow p(\mathbf{x}|D, I)$$

- ▶ As in the 1D case, the posterior PDF still encodes all the information we need to get the best estimator.
- ▶ The maximum of the PDF gives the best estimate of the quantities $\mathbf{x} = \{x_j\}$.
- ▶ We solve the set of **simultaneous equations**

$$\left. \frac{\partial p}{\partial x_i} \right|_{\{\hat{x}_j\}} = 0$$

- ▶ **Question:** how to we make sure that we're at the maximum and not a minimum or a **saddle point**?

The Quadratic Approximation Revisited

- ▶ It's easier to deal with $L = \ln p(\{x_j\} | D, I)$, so let's do that. Let's also simplify to 2D, without loss of generality, so that $\mathbf{x} = (x, y)$.
- ▶ The maximum of the posterior satisfies

$$\left. \frac{\partial L}{\partial x} \right|_{\hat{x}, \hat{y}} = 0 \quad \text{and} \quad \left. \frac{\partial L}{\partial y} \right|_{\hat{x}, \hat{y}} = 0$$

- ▶ Look at the behavior of L about the maximum using its **Taylor expansion**:

$$\begin{aligned} L = L(\hat{x}, \hat{y}) &+ \frac{1}{2} \left. \frac{\partial^2 L}{\partial x^2} \right|_{\hat{x}, \hat{y}} (x - \hat{x})^2 + \frac{1}{2} \left. \frac{\partial^2 L}{\partial y^2} \right|_{\hat{x}, \hat{y}} (y - \hat{y})^2 \\ &+ \left. \frac{\partial^2 L}{\partial x \partial y} \right|_{\hat{x}, \hat{y}} (x - \hat{x})(y - \hat{y}) + \dots \end{aligned}$$

where the linear terms are zero because we're at the maximum.

The Hessian Matrix

- ▶ As in the 1D case, the quadratic terms in the expansion dominate the behavior near the maximum.
- ▶ **Insight:** rewrite the quadratic terms in **matrix notation**:

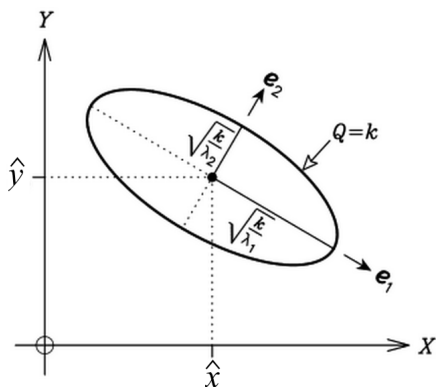
$$\begin{aligned} Q &= \frac{1}{2} (x - \hat{x} \quad y - \hat{y}) \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} x - \hat{x} \\ y - \hat{y} \end{pmatrix} \\ &= \frac{1}{2} (\mathbf{x} - \hat{\mathbf{x}})^\top \mathbf{H}(\hat{\mathbf{x}}) (\mathbf{x} - \hat{\mathbf{x}}) \end{aligned}$$

where $\mathbf{H}(\hat{\mathbf{x}})$ is a 2×2 **real symmetric matrix** with components

$$A = \left. \frac{\partial^2 L}{\partial x^2} \right|_{\hat{x}, \hat{y}}, \quad B = \left. \frac{\partial^2 L}{\partial y^2} \right|_{\hat{x}, \hat{y}}, \quad C = \left. \frac{\partial^2 L}{\partial x \partial y} \right|_{\hat{x}, \hat{y}}$$

- ▶ Note: $\mathbf{H}(\hat{\mathbf{x}})$, the matrix of second derivatives, is called the **Hessian matrix** of L .

Geometrical Interpretation



- ▶ Contour of Q in xy plane is an **ellipse** centered at (\hat{x}, \hat{y})
- ▶ Orientation and eccentricity are determined by the values of A , B , and C
- ▶ Principal axes correspond to the **eigenvectors** of \mathbf{H} . I.e., if we solve

$$\mathbf{H}\mathbf{x} = \lambda\mathbf{x}$$

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

we get **two eigenvalues** λ_1 and λ_2 which are inversely related to the square of the semi-major and semi-minor axes of the ellipse

Condition for a Maximum

- ▶ $L(\hat{\mathbf{x}})$ is a maximum if the **quadratic form** $Q(\mathbf{x} - \hat{\mathbf{x}}) = Q(\Delta\mathbf{x}) < 0 \forall \mathbf{x}$.
- ▶ \mathbf{H} is real and symmetric, so there exists an orthogonal matrix $\mathbf{O} = (\mathbf{e}_1 \ \mathbf{e}_2)$ such that

$$\mathbf{O}^\top \mathbf{H} \mathbf{O} = \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

where \mathbf{e}_1 and \mathbf{e}_2 are the eigenvectors of \mathbf{H} .

- ▶ Therefore, $\mathbf{H} = \mathbf{O} \mathbf{D} \mathbf{O}^\top$, and we can express Q as

$$\begin{aligned} Q &\propto \Delta\mathbf{x}^\top \mathbf{H} \Delta\mathbf{x} \\ &= \Delta\mathbf{x}^\top (\mathbf{O} \mathbf{D} \mathbf{O}^\top) \Delta\mathbf{x} \\ &= (\mathbf{O}^\top \Delta\mathbf{x})^\top \mathbf{D} (\mathbf{O}^\top \Delta\mathbf{x}) = \Delta\mathbf{x}'^\top \mathbf{D} \Delta\mathbf{x}' \\ &= \lambda_1(x - \hat{x})^2 + \lambda_2(y - \hat{y})^2 \end{aligned}$$

- ▶ $\therefore Q < 0$ iff λ_1 and λ_2 are both **negative**.

Condition for a Maximum

- ▶ The eigenvalues of \mathbf{H} are given by

$$\lambda_{1(2)} = \frac{1}{2} \text{Tr } \mathbf{H} + (-) \sqrt{(\text{Tr } \mathbf{H})^2/4 - \det \mathbf{H}}$$

where

$$\text{Tr } \mathbf{H} = A + B, \quad \det \mathbf{H} = AB - C^2$$

- ▶ **Intuition:** what happens if the cross term $C = 0$? Then the principal axes of the ellipse defined by Q are **aligned with the x and y axes** and the eigenvalues reduce to

$$\lambda_1 = A, \quad \lambda_2 = B$$

- ▶ Analogous to the 1D case, we can associate the “error bars” on \hat{x} and \hat{y} as the inverse root of the **diagonal terms of the Hessian**, or

$$\hat{\sigma}_x^2 = |\lambda_1|^{-1} = \left(-\frac{\partial^2 L}{\partial x^2} \Big|_{\hat{x}, \hat{y}} \right)^{-1}, \quad \hat{\sigma}_y^2 = |\lambda_2|^{-1} = \left(-\frac{\partial^2 L}{\partial y^2} \Big|_{\hat{x}, \hat{y}} \right)^{-1}$$

General Case: $C \neq 0$

- ▶ What happens when the off-diagonal term of \mathbf{H} is nonzero?
- ▶ Let's work in 2D. If we were only interested in the reliability of \hat{x} , then we would evaluate the behavior of the **marginal distribution**

$$p(x|D, I) = \int_{-\infty}^{\infty} p(x, y|D, I) dy$$

about the maximum

- ▶ Using our **quadratic approximation**, $p(x, y|D, I) = \exp L \propto \exp Q$:

$$\begin{aligned} p(x|D, I) &\approx \int_{-\infty}^{\infty} \exp\left(\frac{1}{2}\Delta\mathbf{x}^\top \mathbf{H}\Delta\mathbf{x}\right) dy \\ &= \int_{-\infty}^{\infty} \exp\left(\frac{1}{2}(Ax^2 + By^2 + 2Cxy)\right) dy, \end{aligned}$$

where (without loss of generality) we set $\hat{x} = \hat{y} = 0$.

General Case: $C \neq 0$

Solving the Gaussian Integral

Factor out terms in x , and explicitly change signs because we know that $Q < 0$:

$$\begin{aligned} p(x|D, I) &= \int_{-\infty}^{\infty} e^{-\frac{1}{2}(Ax^2 + By^2 + 2Cxy)} dy \\ &= e^{-\frac{1}{2}Ax^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(By^2 + 2Cxy)} dy \\ &= e^{-\frac{1}{2}\left(A + \frac{C^2}{B}\right)x^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}B\left(y + \frac{Cx}{B}\right)^2} dy \end{aligned}$$

where we **completed the square**: $By^2 + 2Cxy = B\left(y + Cx/B\right)^2 - C^2x^2/B$, allowing us to rearrange the xy cross term.

The remaining integral is a **Gaussian integral** of form

$$\int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2\sigma^2}\right) du = \sigma\sqrt{2\pi}$$

General Case: $C \neq 0$

Expressions for σ_x and σ_y

- ▶ Therefore, the marginal distribution becomes

$$\begin{aligned} p(x|D, I) &= \sqrt{\frac{2\pi}{B}} \exp\left(-\frac{1}{2} \frac{AB - C^2}{B} x^2\right) \\ &= \sqrt{\frac{2\pi}{B}} \exp\left(-\frac{x^2}{2\sigma_x^2}\right), \end{aligned}$$

where

$$\sigma_x^2 = \frac{-B}{AB - C^2} = \frac{-H_{yy}}{\det \mathbf{H}}$$

- ▶ Similarly, if we solve instead for $p(y|D, I)$, we'll find that

$$\sigma_y^2 = \frac{-A}{AB - C^2} = \frac{-H_{xx}}{\det \mathbf{H}}$$

- ▶ Note: we **absorbed a negative sign** back into A and B to match the properties of the Hessian.

Connection to Variance and Covariance

- ▶ Recall the **definition of variance** for a 1D PDF:

$$\text{var}(x) = \langle (x - \mu)^2 \rangle = \int dx (x - \mu)^2 p(x|D, I)$$

- ▶ This can be extended using the 2D PDF

$$\sigma_x^2 = \langle (x - \hat{x})^2 \rangle = \iint dx dy (x - \hat{x})^2 p(x, y|D, I)$$

- ▶ If we use the quadratic approximation for $p(x, y|D, I)$, we find

$$\sigma_x^2 = \frac{-B}{AB - C^2} = \frac{-H_{yy}}{\det \mathbf{H}}$$

and similarly,

$$\sigma_y^2 = \langle (y - \hat{y})^2 \rangle = \frac{-A}{AB - C^2} = \frac{-H_{xx}}{\det \mathbf{H}},$$

the same expressions we just derived (convince yourself).

Connection to Variance and Covariance

- ▶ Also recall the **definition of covariance**:

$$\begin{aligned}\sigma_{xy}^2 &= \langle (x - \hat{x})(y - \hat{y}) \rangle \\ &= \iint_C dx dy (x - \hat{x})(y - \hat{y}) p(x, y|D, I) \\ &= \frac{C}{AB - C^2} \\ &= \frac{H_{xy}}{\det \mathbf{H}}\end{aligned}$$

if we use the quadratic expansion of $p(x, y|D, I)$.

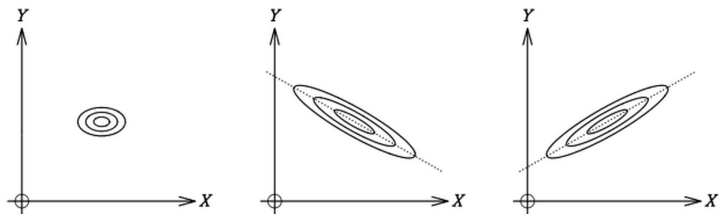
- ▶ Putting it all together: the **covariance matrix**, defined a couple of weeks ago, is the **negative inverse of the Hessian matrix**:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \frac{1}{AB - C^2} \begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} = -\mathbf{H}^{-1}(\hat{\mathbf{x}})$$

Covariance Matrix

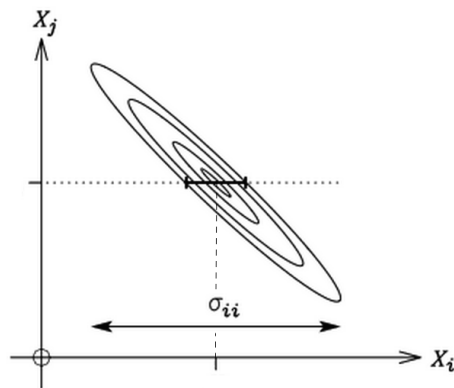
Geometric Interpretation

- ▶ $C = 0$ implies x and y are **completely uncorrelated**. The contours of the posterior PDF are symmetric



- ▶ As C increases, the PDF becomes more and more **elongated**
- ▶ For $C = \pm\sqrt{AB}$, the contours are infinitely wide in one direction (though the prior on x or y could vanish somewhere)
- ▶ Also, while $C = \pm\sqrt{AB}$ implies \hat{x} and \hat{y} are totally unreliable, the **linear correlation** $y = \pm mx$ (with $m = \sqrt{AB}$) can still be inferred

Caution: Using the Correct Error Bar



- ▶ Be careful about calculating the uncertainty on a parameter in a multidimensional PDF
- ▶ **Right:** $\sigma_{ii}^2 = -H_{ii}^{-1}$, from **marginalization** of $p(\mathbf{x}|D, I)$
- ▶ **Wrong:** get σ_{ii}^2 by holding parameters $x_{j \neq i}$ **fixed** at their *optimal values* (underestimate!)
- ▶ See difference in error bars from two procedures at left
- ▶ **Reason:** when using the Hessian, don't confuse the inverse of the diagonals of \mathbf{H} for the diagonals of \mathbf{H}^{-1}

Table of Contents

- 1 Review of Last Class
 - Best Estimates and Reliability
 - Properties of a Good Estimator
- 2 Parameter Estimation in Multiple Dimensions
 - Return of the Quadratic Approximation
 - The Hessian Matrix and its Geometrical Interpretation
 - Maximum of the Quadratic Form
 - Covariance
- 3 Multidimensional Estimators
 - Gaussian Mean and Width
 - Student- t Distribution
 - χ^2 Distribution

Gaussian PDF: Both μ and σ^2 Unknown

- ▶ Last time we derived best estimators for a Gaussian distribution using

$$p(\mu|\sigma, D, I),$$

i.e., σ was given. Now we have the tools to calculate

$$p(\mu|D, I) = \int_0^\infty p(\mu, \sigma|D, I) d\sigma.$$

i.e., we can calculate the best estimator for σ^2 not known *a priori*.

- ▶ First we have to express the **joint posterior PDF** to a likelihood and prior using Bayes' Theorem:

$$p(\mu, \sigma|D, I) \propto p(D|\mu, \sigma, I) p(\mu, \sigma|I)$$

- ▶ If the data are independent, then by the **product rule**

$$p(D|\mu, \sigma, I) = (2\pi\sigma^2)^{-N/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 \right]$$

Gaussian PDF: Priors on μ , σ

- ▶ Now we need to define the prior $p(\mu, \sigma | I)$. Let's assume the **priors for μ and σ are independent**:

$$p(\mu, \sigma | I) = p(\mu | I) p(\sigma | I)$$

- ▶ Since μ is a **location parameter** it makes sense to choose a uniform prior

$$p(\mu | I) = \frac{1}{\mu_{\max} - \mu_{\min}}$$

- ▶ Since σ is a **scale parameter** we'll use a Jeffreys prior:

$$p(\sigma | I) = \frac{1}{\sigma \ln(\sigma_{\max}/\sigma_{\min})}$$

- ▶ Let's also assume the prior ranges on μ and σ are large and don't cut off the integration in a weird way

Aside: Parameterization of σ

- ▶ Note that we parameterized our width prior in terms of σ , not the variance σ^2 . **Does the parameterization make a difference?**
- ▶ For the Jeffreys prior in σ ,

$$p(\sigma|I) d\sigma = k \frac{d\sigma}{\sigma}$$

where k depends on the limits of σ .

- ▶ Now convert to variance ν . Since $\sigma = \sqrt{\nu}$,

$$d\sigma = \frac{d\nu}{2\sqrt{\nu}}$$

- ▶ Therefore,

$$p(\sigma|I) d\sigma = p(\nu|I) d\nu = k \frac{d\nu}{2\nu} = k' \frac{d\nu}{\nu}$$

- ▶ So the Jeffreys prior **has the same form** if we work in terms of σ or σ^2 .

Posterior PDF of μ

- ▶ Substitute the likelihood and prior into our expression for $p(\mu|D, I)$:

$$\begin{aligned} p(\mu|D, I) &\propto \int_0^\infty p(D|\mu, \sigma, I) p(\mu|I) p(\sigma|I) d\sigma \\ &= \frac{(2\pi)^{-N/2}}{\Delta\mu \ln(\sigma_{\max}/\sigma_{\min})} \int_{\sigma_{\min}}^{\sigma_{\max}} \sigma^{-(N+1)} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2} d\sigma \end{aligned}$$

- ▶ Let $\sigma = 1/t$ so that $d\sigma = -dt/t^2$:

$$p(\mu|D, I) \propto \int_{t_{\min}}^{t_{\max}} t^{N-1} e^{-t^2 \sum_{i=1}^N (x_i - \mu)^2} dt$$

- ▶ Change variables again so that $\tau = t\sqrt{\sum (x_i - \mu)^2}$:

$$p(\mu|D, I) \propto \left[\sum_{i=1}^N (x_i - \mu)^2 \right]^{-N/2}$$

Best Estimator and Reliability

- ▶ As in past calculations, we maximize $L = \ln p$:

$$L = -\frac{N}{2} \ln \left[\sum_{i=1}^N (x_i - \mu)^2 \right]$$
$$\left. \frac{dL}{d\mu} \right|_{\hat{\mu}} = \frac{N \sum_{i=1}^N (x_i - \hat{\mu})}{\sum_{i=1}^N (x_i - \hat{\mu})^2} = 0$$

- ▶ This can only be satisfied if the **numerator is zero**, so

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- ▶ In other words, the best estimate of the PDF is still just the **arithmetic mean** of the measurements x_i

Best Estimator and Reliability

- ▶ The second derivative gives the estimate of the width:

$$\left. \frac{d^2 L}{d\mu^2} \right|_{\hat{\mu}} = - \frac{N^2}{\sum_{i=1}^N (x_i - \hat{\mu})^2}$$

- ▶ Therefore, setting $\hat{\sigma}^2 = -(d^2 L/d\mu^2)^{-1}$ we find that

$$\mu = \hat{\mu} \pm \frac{S}{\sqrt{N}},$$

where we define

$$S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

- ▶ This is almost the usual definition of **sample variance** but it's narrower because we divide by $1/N$ instead of $1/(N-1)$.

Aside: Uniform Distribution in σ

- ▶ Suppose at the beginning of this problem we didn't choose a Jeffreys prior for σ , but a **uniform prior** such that

$$p(\sigma|I) = \begin{cases} \text{constant} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

- ▶ In this case, the posterior PDF would have been

$$p(\mu|D, I) \propto \left[\sum_{i=1}^N (x_i - \mu)^2 \right]^{-(N-1)/2}$$

and the width estimator would have been the usual **sample variance**

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

- ▶ In other words, the Jeffreys prior gives us a **narrower constraint** on $\hat{\mu}$!

Student- t Distribution



- ▶ What is the shape of the PDF with unknown σ
 $p(\mu|D, I) \propto \left[\sum_{i=1}^N (x_i - \mu)^2 \right]^{-N/2}$? First, write

$$\sum_{i=1}^N (x_i - \mu)^2 = N(\bar{x} - \mu)^2 + V,$$

where

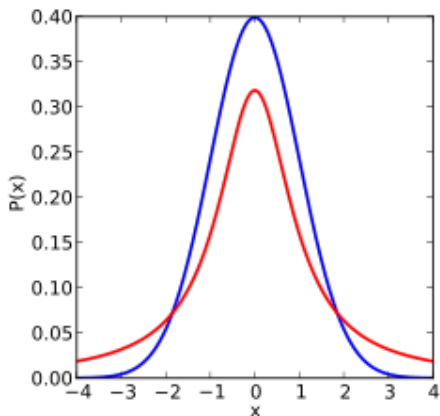
$$V = \sum_{i=1}^N (x_i - \bar{x})^2$$

- ▶ Substituting into the PDF gives

$$p(\mu|D, I) \propto [N(\bar{x} - \mu)^2 + V]^{-N/2}$$

- ▶ This is the heavy-tailed **Student- t distribution**, used for estimating μ when σ is unknown and N is small

Student- t Distribution



- ▶ Published pseudonymously by William S. Gosset of Guinness Brewery in 1908 [1]
- ▶ t -distributions describe small samples drawn from a normally distributed population
- ▶ Used to estimate the error on a mean when only a few samples N are available, σ unknown
- ▶ Basis of the frequentist t -test to compare two data sets
- ▶ As $N \rightarrow$ large, the tails of the distribution are killed off (Central Limit Theorem)

Best Estimate of σ

- ▶ Now that we've calculate the best estimate of a mean, what's the **best estimate of σ** given a set of measurements?
- ▶ Start with the posterior PDF $p(\sigma|D, I)$:

$$\begin{aligned} p(\sigma|D, I) &= \int_{-\infty}^{\infty} p(\mu, \sigma|D, I) d\mu \\ &= \int_{-\infty}^{\infty} p(D|\mu, \sigma, I) p(\mu|I) p(\sigma|I) d\mu \end{aligned}$$

- ▶ Plugging in our likelihood and priors gives

$$\begin{aligned} p(\sigma|D, I) &= \frac{(2\pi)^{-N/2}}{\Delta\mu \ln(\sigma_{\max}/\sigma_{\min})} \sigma^{-(N+1)} \int_{\mu_{\min}}^{\mu_{\max}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2} d\mu \\ &\propto \sigma^{-(N+1)} e^{-\frac{V}{2\sigma^2}} \int_{\mu_{\min}}^{\mu_{\max}} e^{-\frac{N(\bar{x} - \mu)^2}{2\sigma^2}} d\mu \end{aligned}$$

χ^2 Distribution

- ▶ Ignoring all constant terms (including the integral over μ) leaves

$$p(\sigma|D, I) \propto \sigma^{-N} \exp\left(-\frac{V}{2\sigma^2}\right)$$

- ▶ Note that if we had used a **uniform prior for σ** we would have

$$p(\sigma|D, I) \propto \sigma^{-(N-1)} \exp\left(-\frac{V}{2\sigma^2}\right)$$

Let's maximize this expression:

$$\begin{aligned} L &= \ln p = -(N-1) \ln \sigma - \frac{V}{2\sigma^2} \\ \left. \frac{dL}{d\sigma} \right|_{\hat{\sigma}} &= \frac{-(N-1)}{\sigma} + \frac{V}{\sigma^3} = 0 \\ \therefore \hat{\sigma}^2 &= \frac{V}{N-1} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = s^2 \end{aligned}$$

χ^2 Distribution

- ▶ Taking the second derivative of L gives

$$\begin{aligned}\frac{d^2 L}{d\sigma^2} \Big|_{\hat{\sigma}} &= \frac{N-1}{\hat{\sigma}^2} - \frac{3V}{\hat{\sigma}^4} \\ &= \frac{(N-1)\hat{\sigma}^2}{\hat{\sigma}^4} - \frac{3(N-1)\hat{\sigma}^2}{\hat{\sigma}^4} \\ &= -\frac{2(N-1)}{\hat{\sigma}^2}\end{aligned}$$

- ▶ Therefore, the **optimal value of the width** is

$$\sigma = \hat{\sigma} \pm \frac{\hat{\sigma}}{\sqrt{2(N-1)}}$$

- ▶ Note: with the **change of variables** $X = V/\sigma^2$, we see that

$$p(\sigma|D, I) \propto \sigma^{-(N-1)} \exp\left(-\frac{X}{2}\right)$$

is the χ^2_ν distribution with $\nu = 2(N-1)$.

Summary

- ▶ We related the width of a multidimensional distribution — the **Hessian matrix \mathbf{H}** — to the **covariance matrix** via

$$[\sigma^2]_{ij} = [-\mathbf{H}^{-1}]_{ij}$$

- ▶ **Caution:** the right way to get the uncertainty on a parameter from a multidimensional distribution is to marginalize $p(x, y, \dots | D, I)$
- ▶ The **wrong way** to get the uncertainty on a parameter from such a distribution is to fix parameters y, z, \dots at the optimal values and find the uncertainty on x
- ▶ When marginalizing σ in a Gaussian distribution, we obtain **the Student- t distribution**
- ▶ When marginalizing μ in a Gaussian distribution, we obtain the **$\chi^2_{2(N-1)}$ distribution**

References I

- [1] “Student” (W.S. Gosset). “The Probable Error of a Mean”. In: *Biometrika* 6 (1908), pp. 1–25.