LECTURE 3: MORE STATISTICS AND INTRO TO DATA MODELING

- Summarizing the posterior information: mean or mode, and variance. Typically we are interested in more than mean and variance
- Posterior intervals: e.g. 95% credible interval can be constructed as central (relative to median) or highest posterior density. Typically these agree, but:

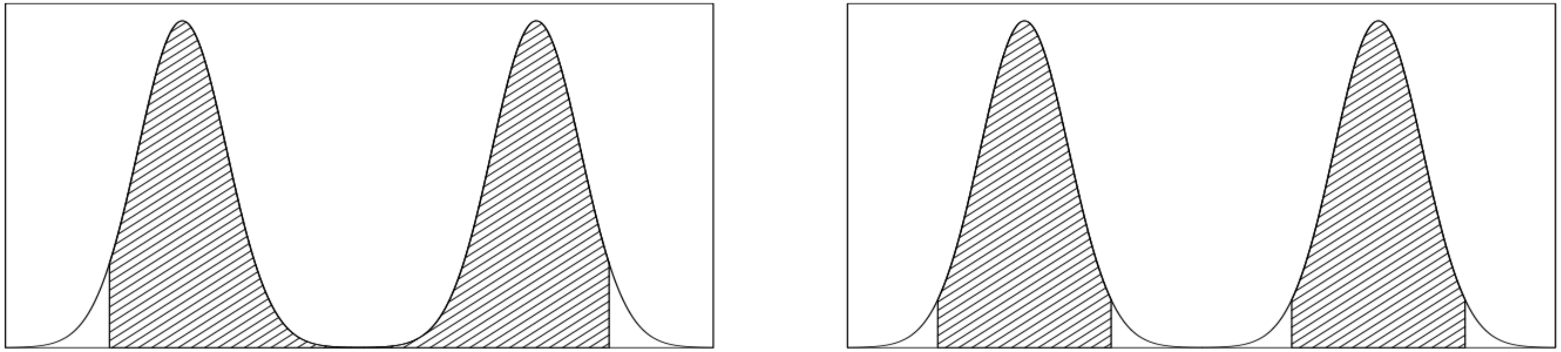


Figure 2.2 *Hypothetical density for which the 95% central interval and 95% highest posterior density region dramatically differ: (a) central posterior interval, (b) highest posterior density region.*

How to choose informative priors?

- Conjugate prior: when posterior takes the same form as the prior it is conjugate to likelihood
- Example: beta distribution is conjugate to binomial (HW 2)
- Can be interpreted as additional data

- For Gaussian with known ***σ*** : $p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$
- Posterior: $p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)$
- Completing the square: $p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Posterior predictive distribution

- Predicting future observation conditional on current data *y*

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)d\theta. \end{aligned}$$

$$\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\mathrm{E}(\tilde{y}|\theta,y)|y) = \mathrm{E}(\theta|y) = \mu_1,$$

and

$$\begin{aligned} \mathrm{var}(\tilde{y}|y) &= \mathrm{E}(\mathrm{var}(\tilde{y}|\theta,y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\theta,y)|y) \\ &= \mathrm{E}(\sigma^2|y) + \mathrm{var}(\theta|y) \\ &= \sigma^2 + \tau_1^2. \end{aligned}$$

Two sources of uncertainty!

Non-informative priors

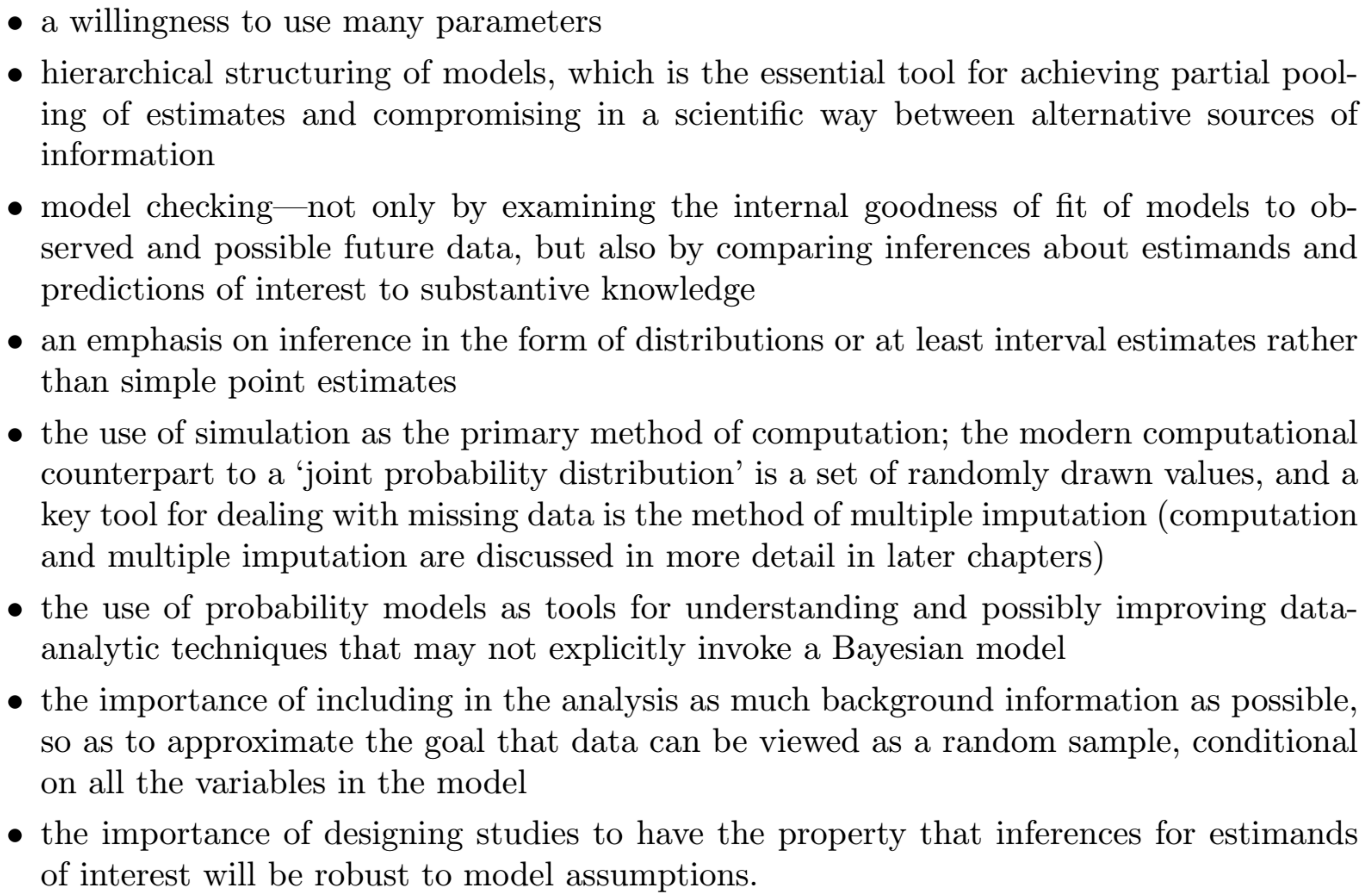
- No prior is truly non-informative, because the transformation of variable changes it
- Priors can be improper: do not integrate to 1. But posteriors must be proper (this must be checked)
- Jeffrey's prior based on minimal Fisher information matrix (to be discussed later): not a universal recipe

Non-informative priors

- Pivotal quantity has distribution independent of y and parameter λ: if this is y-λ then this is a location parameter: uniform prior. E.g. mean of a gaussian
- Scale parameter: pivotal in y/λ. This leads to uniform prior in log λ. E.g. variance of a gaussian
- Prior is rarely an issue in 1-d: either the data are good in which case prior does not matter or are not (so get more data!)
- Priors can become problematic in many dimensions, especially if we have more parameters than needed by the data: posteriors can be a projection of multi-dimensional priors without us knowing it: care must be taken to avoid this (we will discuss further)

Modern statistical methods (Bayesian or not)

Gelman et al., *Bayesian Data Analysis*, 3$^{\text{rd}}$ edition



- a willingness to use many parameters
- hierarchical structuring of models, which is the essential tool for achieving partial pooling of estimates and compromising in a scientific way between alternative sources of information
- model checking—not only by examining the internal goodness of fit of models to observed and possible future data, but also by comparing inferences about estimands and predictions of interest to substantive knowledge
- an emphasis on inference in the form of distributions or at least interval estimates rather than simple point estimates
- the use of simulation as the primary method of computation; the modern computational counterpart to a 'joint probability distribution' is a set of randomly drawn values, and a key tool for dealing with missing data is the method of multiple imputation (computation and multiple imputation are discussed in more detail in later chapters)
- the use of probability models as tools for understanding and possibly improving data-analytic techniques that may not explicitly invoke a Bayesian model
- the importance of including in the analysis as much background information as possible, so as to approximate the goal that data can be viewed as a random sample, conditional on all the variables in the model
- the importance of designing studies to have the property that inferences for estimands of interest will be robust to model assumptions.

INTRO TO MODELING OF DATA

- We are given N number of data measurements (x_i, y_i)
- Each measurement comes with an error estimate σ_i
- We have a parametrized model for the data $y = y(x_i)$
- We think the error probability is Gaussian and the measurements are uncorrelated:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(y(x_i)-y_i)^2}{2\sigma_i^2}}$$

$$p(\vec{y}) = \prod_i p(y_i)$$

We can parametrize the model in terms of M free parameters $y(x_i|a_1,a_2,a_3,\ldots,a_M)$

Bayesian formalism gives us the full posterior information on the parameters of the model

$$p(\vec{y}|\vec{a}) = \prod_i p(y_i|\vec{a}) = \mathcal{L}(\vec{a})$$

$$p(a_1, ..., a_M|\vec{y}) = \frac{\prod_i p(y_i|\vec{a})p(\vec{a})}{p(y_i)}$$

We can assume a flat prior $p(a_1,a_2,a_3,\ldots,a_M) = \text{constant}$
In this case posterior proportional to likelihood

Normalization (evidence, marginal) $p(y_i)$ not needed if we just need relative posterior density

Maximum likelihood estimator (MLE)

- Instead of the full posterior we can ask what is the best fit value of parameters $a_1, a_2, a_3, \ldots, a_M$
- We can define this in different ways: mean, median, mode
- Choosing the mode (peak posterior or peak likelihood) means we want to maximize the likelihood: maximum likelihood estimator (or MAP for non-uniform prior)

$$\text{MLE}: \quad \frac{\partial \mathcal{L}}{\partial \vec{a}} = 0 \quad \text{or} \quad \frac{\partial \ln \mathcal{L}}{\partial \vec{a}} = 0$$

Maximum likelihood estimator

$$\underbrace{-2\ln\mathcal{L} = \sum_i \left\{ \frac{(y_i - y(x_i|a_1, ..., a_M))^2}{\sigma_i^2} + \ln\sigma_i \right\}}_{\chi^2}$$

Since σ_i does not depend on a_i, MLE means minimizing χ^2

$$\frac{\partial\chi^2}{\partial a_k} = 0 \quad \rightarrow \quad \sum_i \frac{y_i - y(x_i)}{\sigma_i^2} \frac{\partial y(x_i)}{\partial a_k} = 0$$

This is a system of M nonlinear equations for M unknowns

Fitting data to a straight line

Linear Regression $y(x) = y(x; a, b) = a + bx$

$$\chi^2(a,b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

Measures how well the model agrees with the data

Minimize χ^2:

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

Define: $S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

$$\begin{array}{ccc} 0 = \dfrac{\partial \chi^2}{\partial a} = -2 \displaystyle\sum_{i=1}^{N} \frac{y_i - a - b x_i}{\sigma_i^2} & & aS + bS_x = S_y \\ & \longrightarrow & \\ 0 = \dfrac{\partial \chi^2}{\partial b} = -2 \displaystyle\sum_{i=1}^{N} \frac{x_i (y_i - a - b x_i)}{\sigma_i^2} & & aS_x + bS_{xx} = S_{xy} \end{array}$$

Matrix Form: $\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$

Solve this with linear algebra

$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

Solution: Define $\Delta \equiv S S_{xx} - (S_x)^2$

$$\hat{a} = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}$$

$$\hat{b} = \frac{S S_{xy} - S_x S_y}{\Delta}$$

This gives best fit $\hat{a}$ & $\hat{b}$

What about the errors?

- We approximate the log posterior around its peak with a quadratic function
- The posterior is thus approximated as a Gaussian
- This goes under name Laplace approximation
- Note that the errors need to be described as a matrix

$$-2 \cdot \ln p(a,b|y_i) = -2 \cdot \ln \mathcal{L}(a,b)$$

Taylor expansion around the peak $(\hat{a}\ \&\ \hat{b})$

Let $a = x_1, b = x_2$

$$-2 \cdot \ln \mathcal{L}(x_1,x_2) = -2 \cdot \ln \mathcal{L}(\hat{x_1},\hat{x_2}) - 2 \cdot \frac{1}{2} \sum_{i,j=1,2} \frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j}\Big|_{x_i=\hat{x}_i} \Delta x_i \Delta x_j$$

where $\Delta x_i = x_i - \hat{x}_i$

$$-\frac{1}{2} \sum_{ij} \Delta x_i C^{-1}_{ij} \Delta x_j$$

Note: $\langle \Delta x_i \Delta x_j \rangle = C_{ij}$

Gaussian posterior approximation

$$-\frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j} \equiv C^{-1}_{ij}$$

$(C^{-1} = \alpha$ is called precision matrix.)

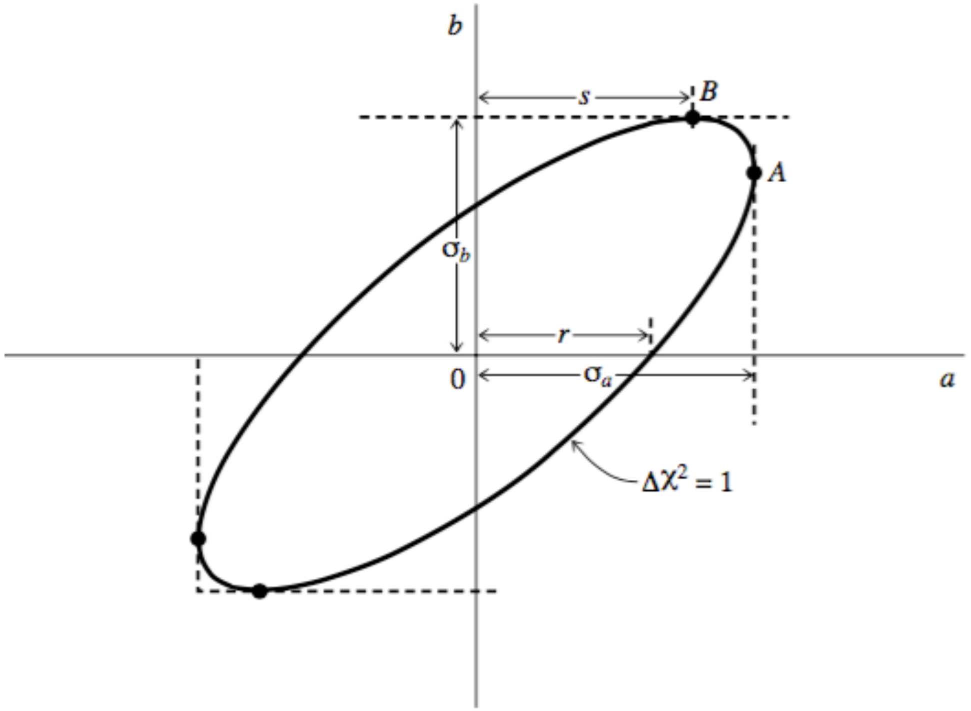
$$\mathcal{L} \ \propto \ e^{-\frac{1}{2}\sum_{ij} \Delta x_i C^{-1}_{ij} \Delta x_j}$$

$$-2 \cdot \ln\mathcal{L} = \chi^2$$

$$\frac{\partial^2\chi^2}{\partial a^2} = 2\sum_i \frac{1}{\sigma_i^2} = 2S$$

$$\frac{\partial^2\chi^2}{\partial b^2} = 2\sum_i \frac{x_i^2}{\sigma_i^2} = 2S_{xx}$$

$$\frac{\partial^2\chi^2}{\partial a \partial b} = 2\sum_i \frac{x_i}{\sigma_i^2} = 2S_x$$



$$C^{-1} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \qquad C = \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

Define $\Delta \equiv SS_{xx} - (S_x)^2$

$$\boxed{\sigma_a^2 = S_{xx}/\Delta \qquad \sigma_b^2 = S/\Delta}$$

Asymptotics theorems

(Le Cam 1953, adopted to Bayesian posteriors)

- Posteriors approach a multi-variate Gaussian in the large *N* limit (*N*: number of data points):

this is because the 2nd order Taylor expansion of $\ln L$ is more and more accurate in this limit, i.e. we can drop 3rd order terms

- The marginalized means approach the true value and the variance approaches the Fisher matrix, defined as ensemble average of precision matrix $<C^{-1}>$

- The likelihood dominates over the prior in large *N* limit

Asymptotics theorems

(Le Cam 1953, adopted to Bayesian posteriors)

- There are counter-examples, e.g. when data are not informative about a parameter or some linear combination of them, when number of parameters M is comparable to N, when posteriors are improper or likleihoods are unbouded… Always exercise care!
- In practice the asymptotic limit is often not achieved for nonlinear models, i.e. we cannot linearize the model across the region of non-zero posterior: this is why we often use sampling to evaluate the posteriors instead of Gaussian

Bayesian View

- The posterior distribution $p(a,b|y_i)$ is described as a 2-d C^{-1} ellipse in (a,b) plane
- At any fixed value of a (or b) the posterior of b (or a) is a gaussian with variance $[C^{-1}{}_{bb(aa)}]^{-1}$
- If we want to know the error on b (or a) independent of a (or b) we need to marginalize over a (or b)
- This marginalization can be done analytically, and leads to $C_{bb(aa)}$ as the variance of b (or a)
- This will increase the error: $C_{bb(aa)}>[C^{-1}{}_{bb(aa)}]^{-1}$

Show

$$\int da \cdot e^{-\frac{1}{2}\left[(a-\hat{a})^2 C_{aa}^{-1} + (a-\hat{a})(b-\hat{b}) C_{ab}^{-1} + (b-\hat{b})^2 C_{bb}^{-1}\right]} \propto e^{-\frac{1}{2}\frac{(b-\hat{b})^2}{C_{bb}}}$$

(Complete the square in *a*)

Show

$$\int da \cdot e^{-\frac{1}{2}\left[(a-\hat{a})^2 C_{aa}^{-1}+(a-\hat{a})(b-\hat{b})C_{ab}^{-1}+(b-\hat{b})^2 C_{bb}^{-1}\right]} \;\propto\; e^{-\frac{1}{2}\frac{(b-\hat{b})^2}{C_{bb}}}$$

(Complete the square in a)

Solution:

$$C_{aa}^{-1}\left[(\delta a+\frac{C_{ab}^{-1}}{C_{aa}^{-1}}\delta b)^2-\frac{C_{ab}^{-2}}{C_{aa}^{-2}}\delta b^2\right]+C_{bb}^{-1}\delta b^2$$

$$\int_{-\infty}^{\infty} da \cdot e^{-\frac{1}{2}C_{aa}^{-1}\left[(\delta a+\frac{C_{ab}^{-1}}{C_{aa}^{-1}}\delta b)^2\right]} = \sqrt{2\pi C_{aa}}$$

$$\propto\; e^{\frac{1}{2}\delta b^2\left[\frac{-C_{ab}^{-2}+C_{bb}^{-1}C_{aa}^{-1}}{C_{aa}^{-1}}\right]} = e^{-\frac{1}{2}\frac{\delta b^2}{C_{bb}}}$$

Multivariate linear least squares

- We can generalize the model to a generic functional form

  $y_i = a_0 X_0(x_i) + a_1 X_1(x_i) + \ldots + a_{M-1} X_{M-1}(x_i)$

- The problem is linear in a_j and can be nonlinear in x_i,

  e.g. $X_j(x_i) = x_i^j$

$$\chi^2 = \sum_{i=0}^{N-1} \left[\frac{y_i - \sum_{k=0}^{M-1} a_k X_k(x_i)}{\sigma_i} \right]^2$$

- We can define design matrix $A_{ij} = X_j(x_i)/\boldsymbol{\sigma}_i$ and

- $b_i = y_i / \boldsymbol{\sigma}_i$

$$\boxed{\chi^2 = |\mathbf{A} \cdot \mathbf{a} - \mathbf{b}|^2}$$

Design matrix

$$\begin{array}{ccc}
 & & \longleftarrow \text{basis functions} \longrightarrow \\
 & & \begin{array}{cccc} X_0(\ ) & X_1(\ ) & \cdots & X_{M-1}(\ ) \end{array} \\
\updownarrow \text{data points} & \begin{array}{c} x_0 \\[2em] x_1 \\[2em] \vdots \\[2em] x_{N-1} \end{array} & \begin{pmatrix}
\dfrac{X_0(x_0)}{\sigma_0} & \dfrac{X_1(x_0)}{\sigma_0} & \cdots & \dfrac{X_{M-1}(x_0)}{\sigma_0} \\
\dfrac{X_0(x_1)}{\sigma_1} & \dfrac{X_1(x_1)}{\sigma_1} & \cdots & \dfrac{X_{M-1}(x_1)}{\sigma_1} \\
\vdots & & & \vdots \\
\vdots & & & \vdots \\
\vdots & & & \vdots \\
\dfrac{X_0(x_{N-1})}{\sigma_{N-1}} & \dfrac{X_1(x_{N-1})}{\sigma_{N-1}} & \cdots & \dfrac{X_{M-1}(x_{N-1})}{\sigma_{N-1}}
\end{pmatrix}
\end{array}$$

Credit: NR, Press et al.

Solution by normal equations

$$0=\sum_{i=0}^{N-1}\frac{1}{\sigma_i^2}\left[y_i-\sum_{j=0}^{M-1}a_jX_j(x_i)\right]X_k(x_i)\qquad k=0,\ldots,M-1 \tag{15.4.6}$$

Interchanging the order of summations, we can write (15.4.6) as the matrix equation

$$\sum_{j=0}^{M-1}\alpha_{kj}a_j=\beta_k \tag{15.4.7}$$

where

$$\alpha_{kj}=\sum_{i=0}^{N-1}\frac{X_j(x_i)X_k(x_i)}{\sigma_i^2}\qquad \text{or, equivalently,}\qquad \boldsymbol{\alpha}=\mathbf{A}^T\cdot\mathbf{A} \tag{15.4.8}$$

an $M\times M$ matrix, and

$$\beta_k=\sum_{i=0}^{N-1}\frac{y_iX_k(x_i)}{\sigma_i^2}\qquad \text{or, equivalently,}\qquad \boldsymbol{\beta}=\mathbf{A}^T\cdot\mathbf{b} \tag{15.4.9}$$

$$\boldsymbol{\alpha}\cdot\mathbf{a}=\boldsymbol{\beta}\qquad \text{or as}\qquad \left(\mathbf{A}^T\cdot\mathbf{A}\right)\cdot\mathbf{a}=\mathbf{A}^T\cdot\mathbf{b} \tag{15.4.10}$$

Gaussian posterior

$$P(\delta\mathbf{a})\,da_0\ldots da_{M-1} = \text{const.} \times \exp\left(-\tfrac{1}{2}\delta\mathbf{a}\cdot\boldsymbol{\alpha}\cdot\delta\mathbf{a}\right)\,da_0\ldots da_{M-1}$$

Marginalization over nuisance parameters

- If we want to know the error on *j*-th parameter we need to marginalize over all other parameters
- In analogy to 2-d case this leads to $\sigma_j^2 = C_{jj}$
- So we need to invert the precision matrix $a = C^{-1}$
- Analytic marginalization is only possible for a multi-variate Gaussian distribution: a great advantage of using a Gaussian
- If the posterior is not Gaussian it may be made more Gaussian by a nonlinear transformation of the variable

What about multi-dimensional projections?

- Suppose we are interested in v components of a, marginalizing over remaining M- v components.

- We take the components of C corresponding to v parameters to create v x v matrix C_{proj}

- Invert the matrix to get precision matrix C_{proj}^{-1}

- Posterior distribution is proportional to

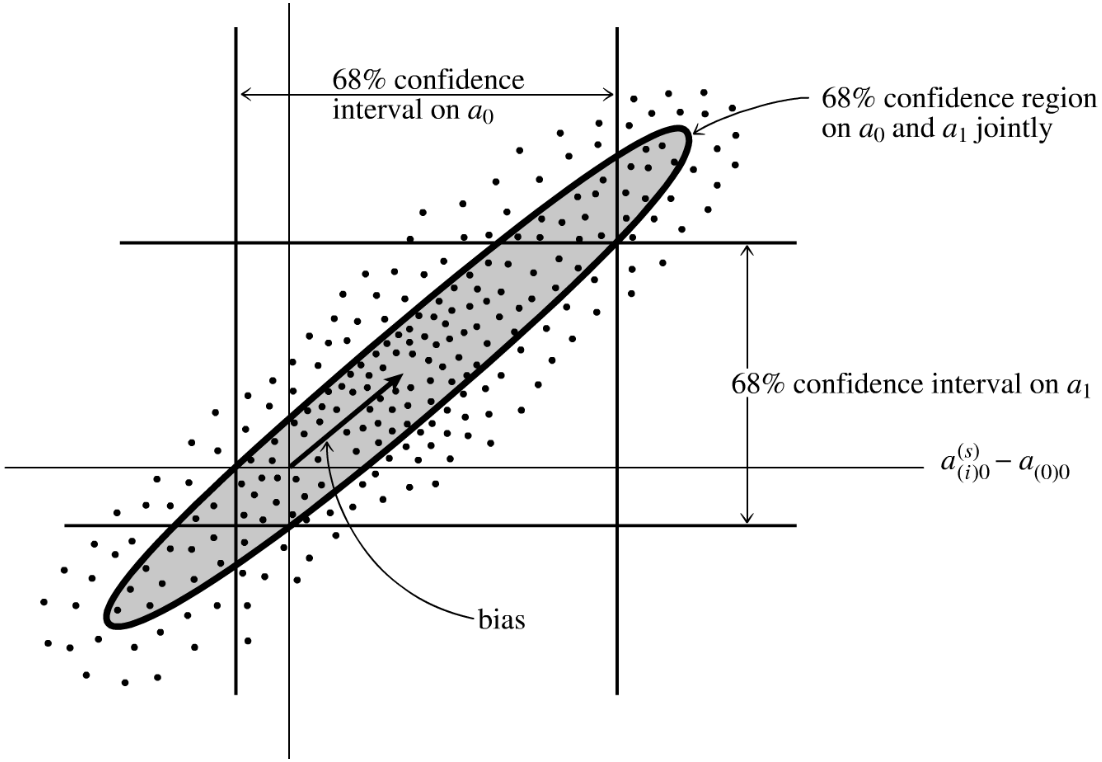
  $\exp(-\delta a_{\mathrm{proj}}^{T} C_{\mathrm{proj}}^{-1}\, \delta a_{\mathrm{proj}}/2)$,

  which is distributed as $\exp(-\Delta\chi^2/2)$,

  i.e. χ^2 with v degrees of freedom

Credible intervals under Gaussian posterior approx.

- We like to quote posteriors in terms of X% credible intervals
- For Gaussian likelihoods most compact posteriors correspond to a constant change $\Delta\chi^2$ relative to MAP/MLE
- The intervals depend on the dimension: example for X=68


68% confidence interval on a_0
68% confidence region on a_0 and a_1 jointly
68% confidence interval on a_1
$a^{(s)}_{(i)0} - a_{(0)0}$
bias

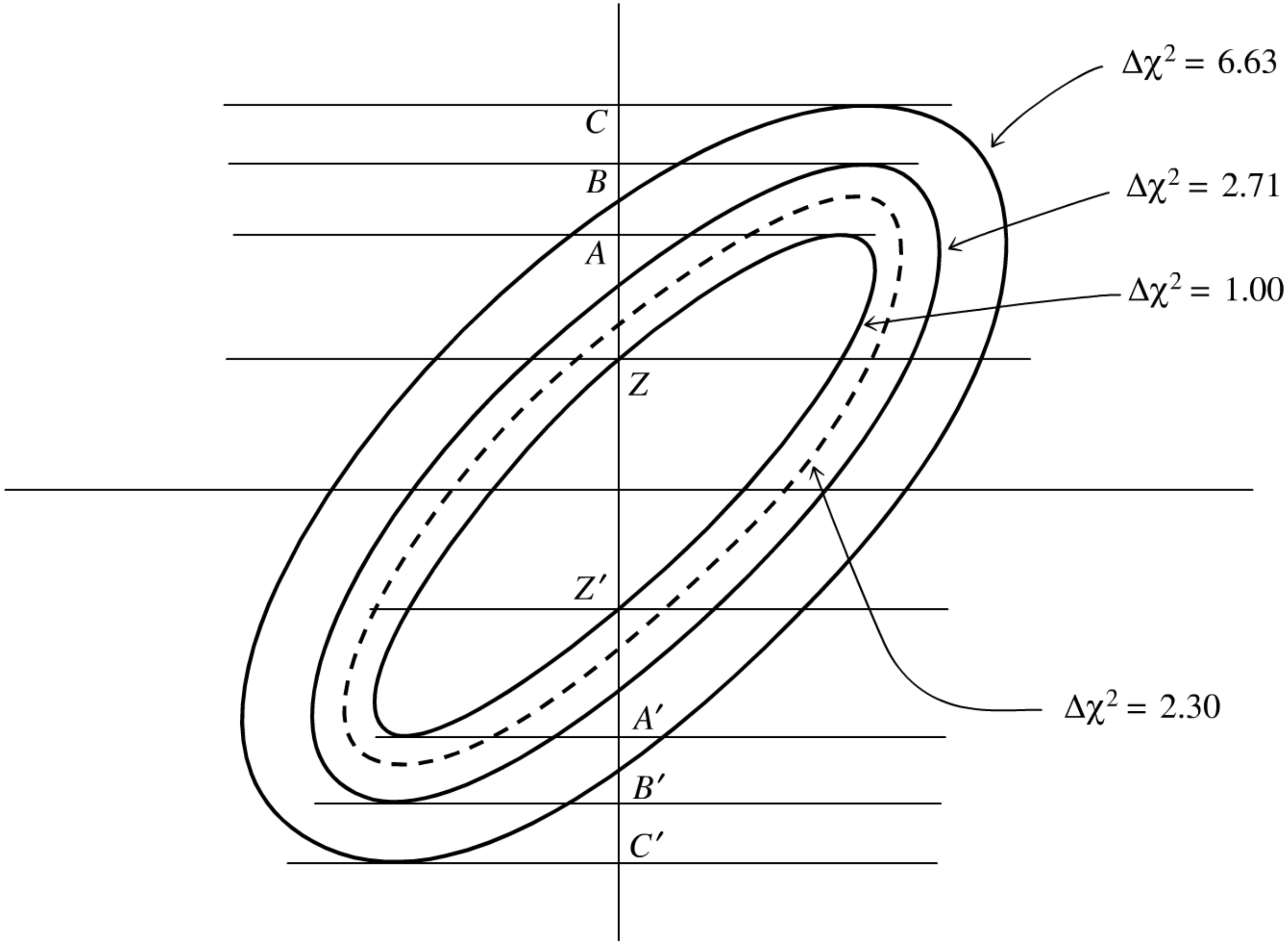
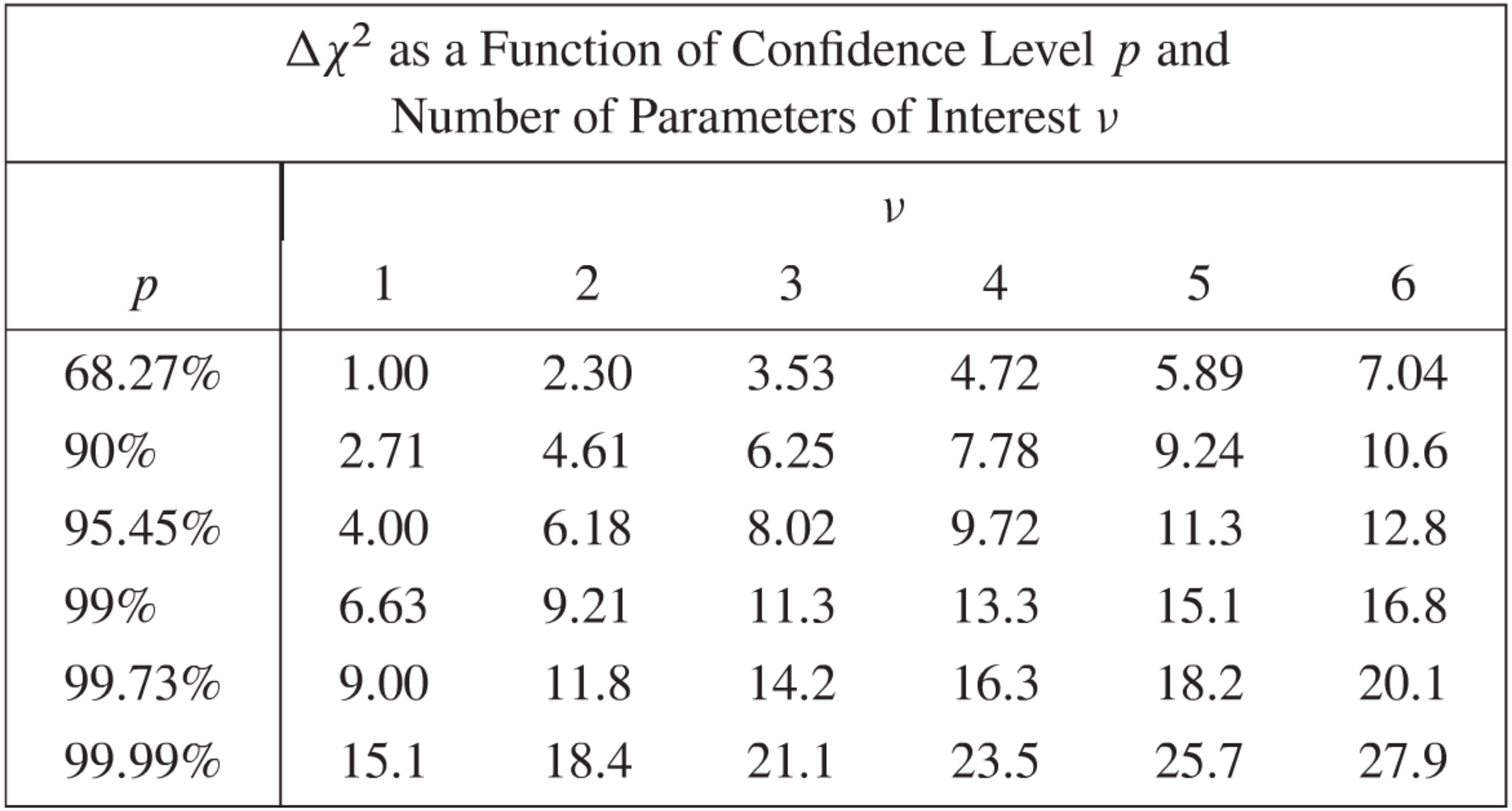


Figure 15.6.4. Confidence region ellipses corresponding to values of chi-square larger than the fitted minimum. The solid curves, with $\Delta\chi^2 = 1.00, 2.71, 6.63$, project onto one-dimensional intervals AA', BB', CC'. These intervals — not the ellipses themselves — contain 68.3%, 90%, and 99% of normally distributed data. The ellipse that contains 68.3% of normally distributed data is shown dashed and has $\Delta\chi^2 = 2.30$. For additional numerical values, see the table on p. 815.



$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν

	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

We rarely go above $v = 2$ dimensions in projections (difficult to visualize)

To solve the normal equations to obtain best fit values and the precision matrix we need to learn linear algebra numerical methods: topic of next lecture

Literature

- *Numerical Recipes*, Press et al., Chapter 15 (http://apps.nrbook.com/c/index.html)
- *Bayesian Data Analysis,* Gelman et al. , Chapter 1-4