ESE 524: Detection and Estimation Theory

Recitation 4

Washington University in St. Louis

Outline

- Bayesian inference
  - We will use Bayesian inference address two practical problems: searching for ship and plane wrecks, and to estimate cancer mortality rates.
  - Then we will derive an important analytical result.

Useful Formulas

- Sequential Bayes estimation:

$$p(\boldsymbol{\theta}|x_1, x_2) \propto p(x_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1)$$

- Jeffrey's prior:

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\mathcal{I}(\boldsymbol{\theta})}$$

- Classical MSE:

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \int (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$$

- Bayesian MSE:

$$\mathrm{BMSE}(\hat{\boldsymbol{\theta}}) = \int \int (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{x} d\boldsymbol{\theta}$$

Useful Formulas (Cont.)

- Bayesian MMSE:

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\boldsymbol{x})$$

- Gaussian linear model:

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}, \text{ where } \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \quad \boldsymbol{C}_{\boldsymbol{w}})$$

$$\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{x}) = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{x} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}})$$

$$\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}$$

- Maximum aposteriori estimation (MAP) :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x})$$

Useful Formulas (Cont.)

- Asymptotic normality of MAP:

$$\lim_{N\to\infty} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, N\mathcal{I}(\boldsymbol{\theta}))$$

Example 1: Bayesian Search Theory

- Ships and planes sometimes sink and are lost in the sea. Finding them is very challenging since their last known location usually differs from the sinking location and water currents can move them.
- Many submarines have been lost, particularly during the cold war.
- The U.S. Navy misplaced the nuclear submarine USS *Scorpion* in 1968 and a B-52 bomber crashed in 1966 (along with it's payload, a hydrogen bomb).
- The K-129 soviet submarine was lost in 1968 and recovered from the sea ground by the CIA.
- More recent examples include the Malaysian Airlines plane (2014) and the Argentinian ARA San Juan submarine (2017).
- How are these wrecks searched and eventually found?
- Actually, just using the idea of sequential Bayesian estimation from the lecture slides!

Scorpion Itinerary

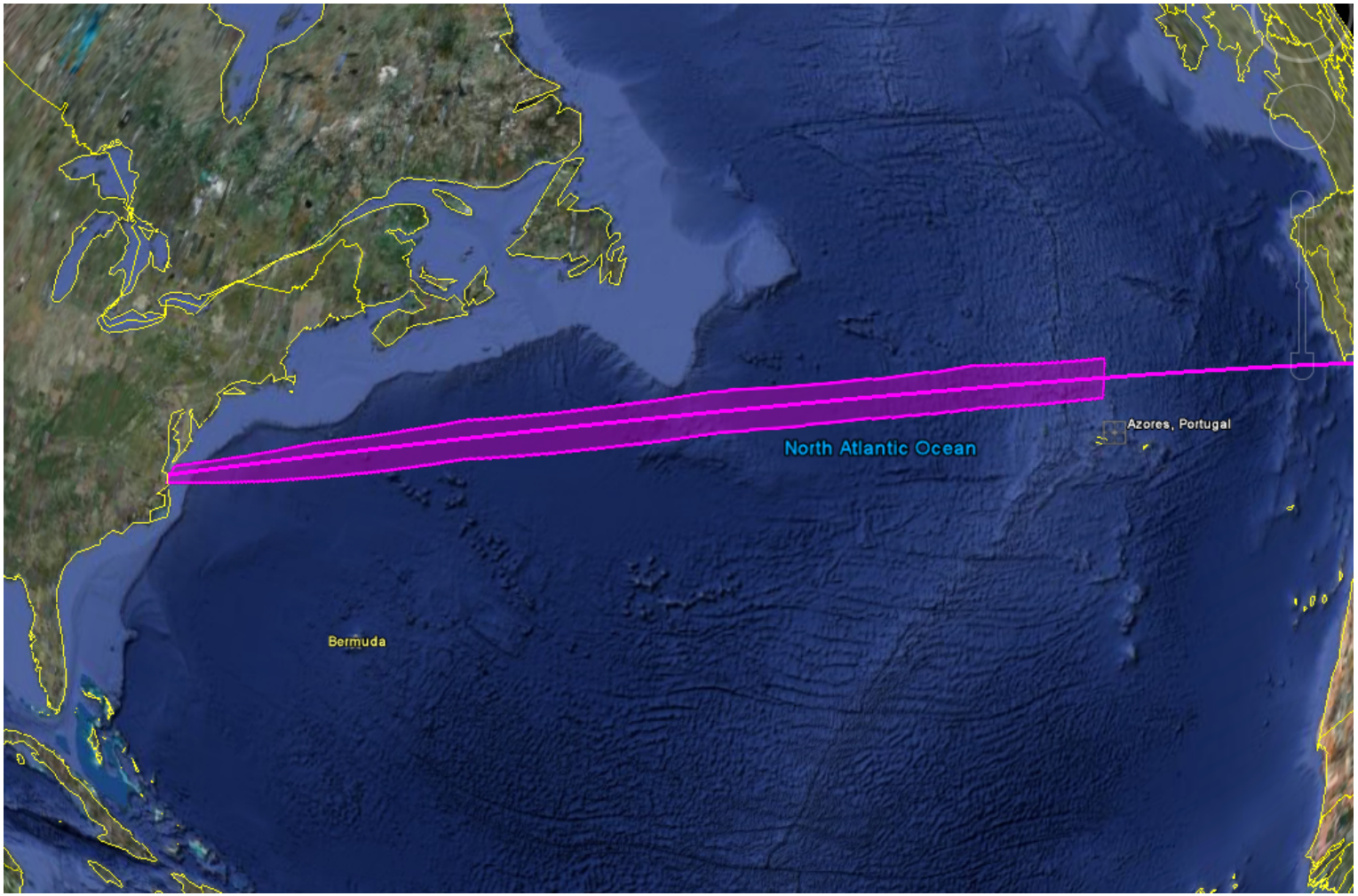

Figure 1: Itinerary and uncertainty region for the USS *Scorpion*.

Mathematical Model for the Search

- Divide up the search area into N cells, denoted by $i = 1, 2, ...N$.
- Let $\theta_i \in \{0, 1\}$ denote whether the target is in cell i.
- Define $p_i = p(\theta_i)$ as the **prior distribution** modeling the belief of the target's location.
- Define $x_i^j \in \{0, 1\}$ as the result of the j^{th} search of cell i.
- Define $q = p(x_i^j = 1 | \theta_i = 1)$ as the probability of detecting the target in any cell. I.e have the same probability of successfully searching every cell.

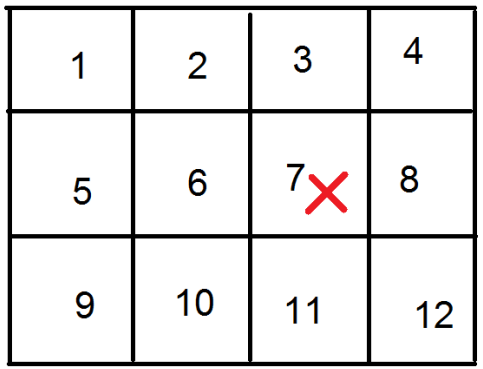

Figure 2: Example of a search grid with the target in one cell.

The Original *Scorpion* Prior Distribution

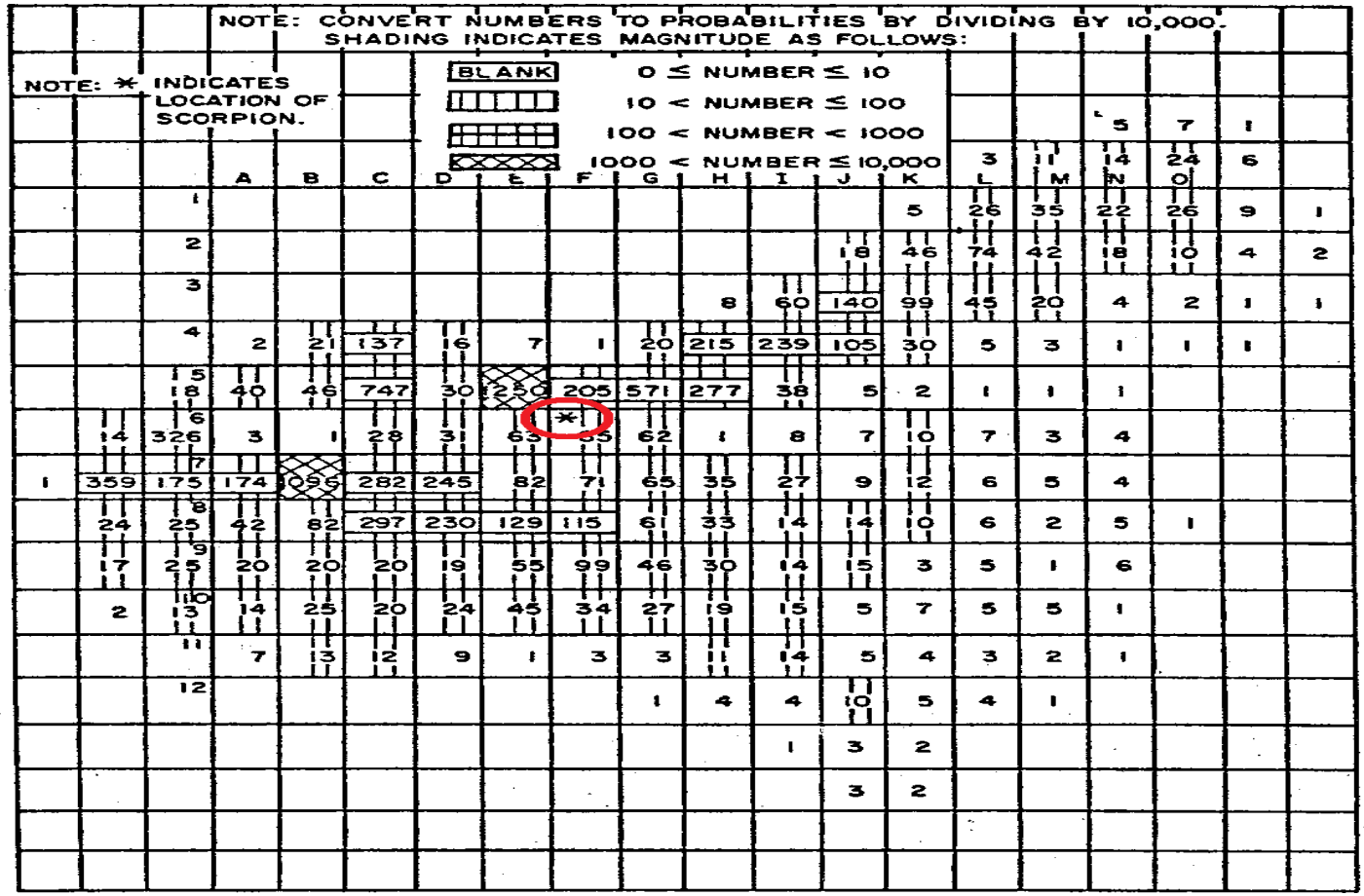

FIGURE 2. Overall A Priori distribution for *Scorpion* search

Figure 3: Original Prior Distribution for the *Scorpion* search. The sub was found 200 yards from the original highest probability cell!

The First Update

- Even if we search a cell, the wreck could still be there.
- Use Bayes theorem to update the probability:

$$p(\theta_i = 1|\text{Search did not detect}) = \frac{p(\text{Search did not detect in cell } i|\theta_i{=}1)p(\theta_i{=}1)}{p(\text{Search did not detect in cell } i)}$$

$$= p(\theta_i = 1|x_i = 0) = \frac{p(x_i^1{=}0|\theta_i{=}1)p(\theta_i{=}1)}{p(\theta_i{=}1)p(x_i{=}0|\theta_i{=}1)+p(\theta_i{=}0)p(x_i{=}0|\theta_i{=}0)}$$

$$= \frac{(1-q)p_i}{p_i(1-q)+(1-p_i)(1)} = p_i\frac{1-q}{1-p_iq}$$

- This changes every other cell as well.

The Other Cells

- Not finding the target in a cell should increase the probability that the target is in other cells.
- Again apply Bayes theorem:

$$p(\theta_j = 1 | \text{Search did not detect in in cell } i)$$

$$= \frac{p(\text{Search did not detect in cell } i | \theta_j = 1) p(\theta_j = 1)}{p(\text{Search did not detect in cell } i)}$$

- But $p(\text{Search did not detect in cell } i | \theta_j = 1) = 1$! so the update for the other cells is:

$$= p(\theta_j = 1 | x_i = 0) = \frac{p_j}{p_i(1-q) + (1-p_i)}$$

Searching the Next Cell

- Create a new prior over every cell:

$$p_j = p(\theta_j = 1 | x_i = 0)$$

- Then apply the same formulas as before.
- The key ingredient for a successful search is an accurate prior.
- For shipwrecks priors are usually given by a combination of a normal distribution based on the itinerary/last known points and the output of fluid flow simulations to approximate drift.

Visual Example - Air France Flight 447

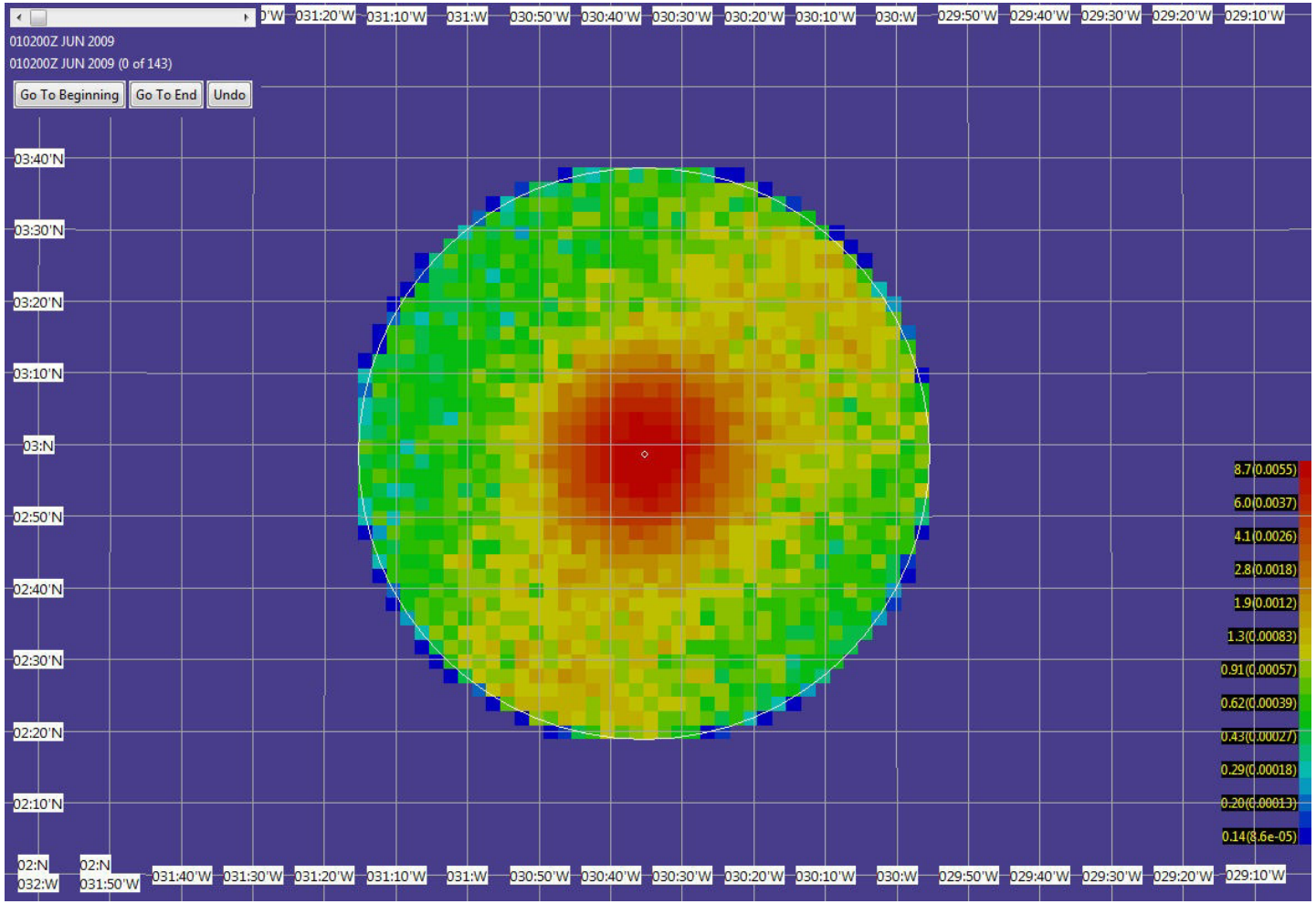

Figure 4: The prior distribution for the Air France crash generated by Metron.

Visual Example - Air France Flight 447

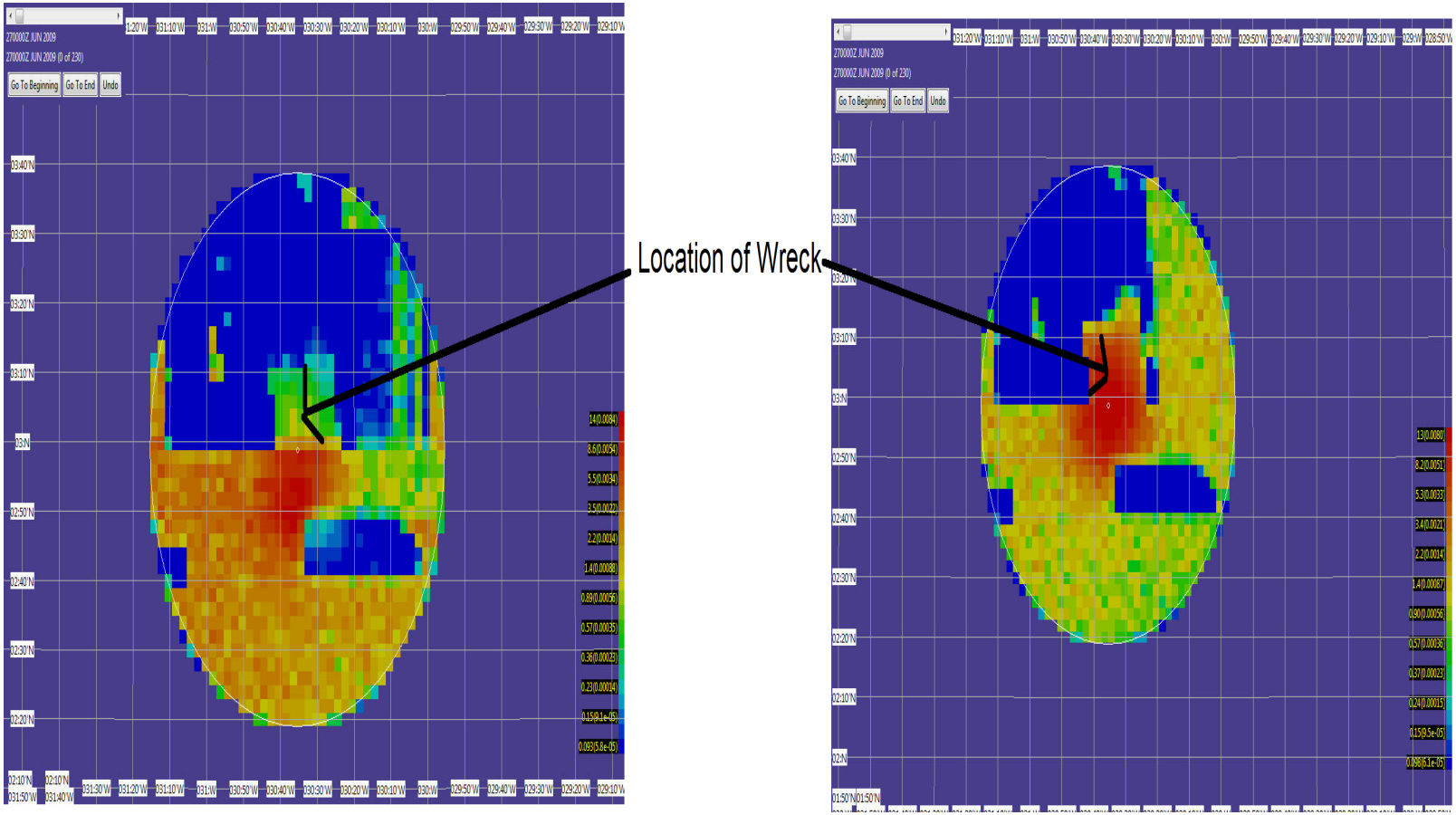

Figure 5: Sequential posterior distributions generated after searching for a while. Note that it becomes "better" to search some locations over again as probabilities are updated.

Comments

- How do you choose where to search? How much time do you spend in each cell?
- How do you make a prior?
- Most searches incorporate several different teams and search methods, so the probability equations become more complex.
- In the Malaysian Air case, there is not much good prior information available, but searchers did use Bayesian approaches.

Example 2: Bayesian Methods for Kidney Cancer Mortality Rates

- Knowing the incidence of different diseases is useful to allocate resources efficiently.
- Cancer treatments require expensive equipment and large personnel.
- We wish to have sufficient resources to treat all the patients but not to pay for resources that won't be used.
- In this example we will estimate the mortality rates of kidney cancer in the US.
- This example can be extended to other diseases (e.g. COVID-19).

Bayesian Methods for Kidney Cancer Mortality Rates (Cont.)

- This example is example 2.7 in *Bayesian Data Analysis*, by Andrew Gelman et al.
- The example is interested in predicting kidney cancer death rates per county.

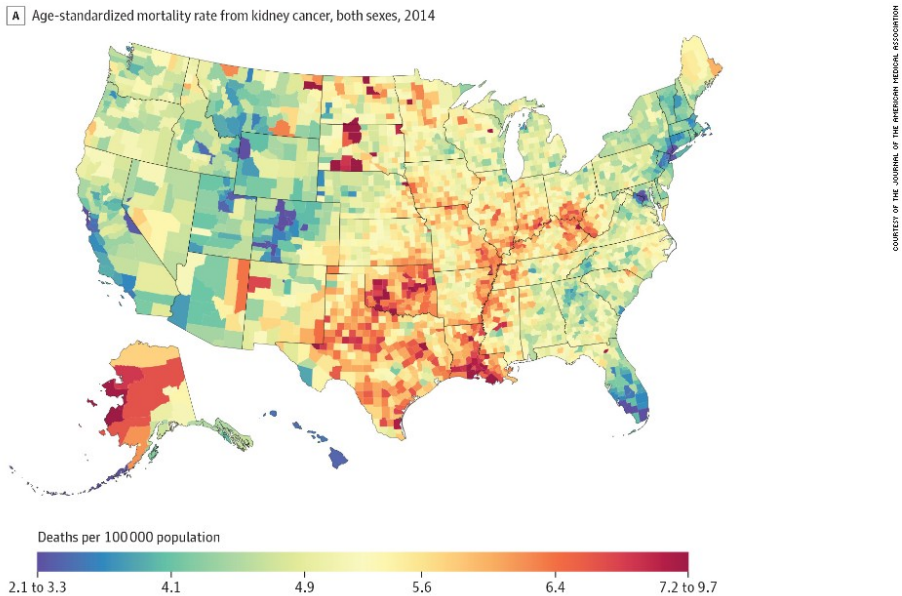

Figure 6: Age-adjusted motrality rates for kidney cancer in 2014. Highest counties are in Kentucky and the south.

Setting Up the Models

- Let y_j be the number of deaths in county j due to kidney cancer.
- Let n_j be the population of county j
- Let θ_j be the underlying "true" kidney cancer mortality rate.
- Since y_j is a counting type object, model it with a Poisson distribution, $p(y_j|\theta_j) \sim \text{Poisson}(10n_j\theta_j)$.
- For mathematical convenience choose a conjugate prior $p(\theta_j) \sim \text{Gamma}(\alpha, \beta)$, where, α and β are parameters TBD.

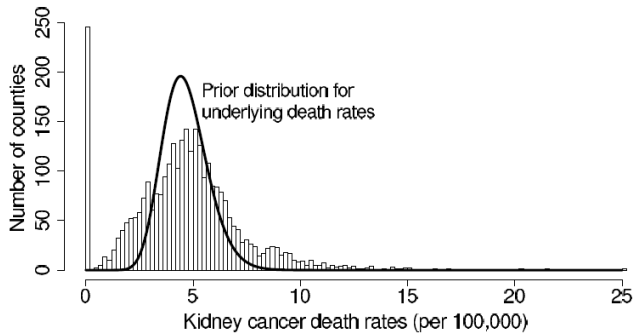

Figure 7: We will estimate the prior from the empirical data in a later slide.

Conjugate Priors

- Why did we choose to model the prior distribution $p(\theta_j)$ using a Gamma distribution?
- Because the Gamma distribution is the conjugate prior of the Poisson distribution.
- If we multiply a likelihood distribution by a conjugate prior distribution, then the posterior distribution will belong to the same family of distributions as the prior distribution.
- This simplifies the calculations.

Finding the Posterior

- Since the prior has been chosen as conjugate, we know the posterior will also be a gamma distribution.
- $p(\theta_j|y_j) = \frac{p(y_j|\theta_j)p(\theta_j)}{p(y)}$

$$\begin{aligned}
\propto p(y_j|\theta_j)p(\theta_j) &= \frac{(10n_j\theta_j)^{y_j}e^{-10n_j\theta_j}}{y_j!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_j^{\alpha-1}e^{-\beta\theta_j} \\
&\propto \theta_j^{y_j}e^{-10n_j\theta_j}\theta_j^{\alpha-1}e^{-\beta\theta_j} \\
&\propto \theta_j^{y_j+\alpha-1}e^{-(10n_j+\beta)\theta_j} \\
&\propto \text{Gamma}(\alpha+y_j, \beta+10n_j)
\end{aligned}$$

- From this, we can see that α is an "average" mortality rate and β is an "average" (scaled) county population in the prior.

Finding the Right Prior Parameters

- In general, we would assign a prior to the parameters α, and β and compute something called a hierarchical model, but that is beyond the scope of this class so we will take a different approach to figure out the prior.
- The marginal distribution of y_j is $p(y_j) = \frac{p(y_j|\theta_j)p(\theta_j)}{p(\theta_j|y_j)}$, by Bayes theorem
- We know all these distributions!

$$
\begin{aligned}
p(y_j) &= \frac{\text{Poisson}(10n_j\theta_j)\text{Gamma}(\alpha,\beta)}{\text{Gamma}(\alpha+y_j,\beta+10n_j)} \\
&= \frac{\Gamma(\alpha+y_j)\beta^{\alpha}}{\Gamma(\alpha)y_j!(10n_j+\beta)^{\alpha+y_j}} \\
&= \binom{\alpha+y_j-1}{y_j}\left(\frac{\beta}{\beta+10n_j}\right)^{\alpha}\left(\frac{1}{\beta+10n_j}\right)^{y_j} \\
&\sim \text{Neg. Binomial}\left(\alpha, \frac{\beta}{10n_j}\right)
\end{aligned}
$$

Finding the Right Prior Parameters Cont.

- To find α and β, use the expectation and variance of y_j.
- $\mathrm{E}(y_j) = 10n_j \frac{\alpha}{\beta}$
- $\mathrm{var}(y_j) = 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2}$
- Setting these equal to the sample mean and variance yields $\alpha = 20$, $\beta = 430,000$.
- Then the estimates of mortality rates are given by

$$\mathrm{E}(\theta_j|y_j) = \frac{20 + y_j}{430000 + 10n_j}$$

$$\mathrm{var}(\theta_j|y_j) = \frac{20 + y_j}{(430000 + 10n_j)^2}$$

- As the county population increases, the mortality rate goes down. As the number of recorded deaths increases, the mortality rate increases.

Model Expansions

- We used a "poor man's hierarchical model" to estimate the prior distribution. A better way to find the parameters is to assign them their own prior and apply bayesian inference again.
- Figure 6 clearly shows some spatial correlation. A more involved model could make use of something called a variogram.
- To do this idea model mortality rates as a function of location, $y_j = y_j(\text{county } j \text{ location})$.
- The variogram is given by $v(y_i, y_j) = \frac{1}{2}\mathrm{E}(y_i(\text{county } i \text{ location}) - y_j(\text{county } j \text{ location})^2)$
- Then describe the likelihood as non-independent poisson random variables with covariances given by the variogram, pick a prior, and repeat the process.
- Other models might include information about other risk factors determined by doctors.

Example 3: Bayesian Linear Models

- This example is a precursor/derivation for Theorem 2 on slide 52.
- We know that the minimum MSE estimator for a single parameter in the Bayesian case is given by the average value of the posterior distribution:

$$\mathrm{E}_{\theta|x} = [\theta|x]$$

- So, as long as we know the posterior distribution, we can gain an estimate.
- However, analytic solutions don't usually exist, especially as models and priors get more complicated/realistic.
- For linear models with Gaussian priors we can find a solution!

Setting Up the Model

- Let $\boldsymbol{x} = [x[1], \ldots, x[n]]^{\mathrm{T}}$ be a vector of samples modeled by :

$$\boldsymbol{x} = \boldsymbol{H\theta} + \boldsymbol{w}$$

  where $\boldsymbol{w} = [w[1], \ldots, w[n]]$ are i.i.d. samples of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C_w})$, and $\boldsymbol{\theta}$ is a vector of parameters to be estimated.
- Then the likelihood function is $p(x|\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{H\theta}, \boldsymbol{C_w})$
- Let the prior distribution be $\pi(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{C_\theta})$ be independent of $\boldsymbol{w}$.
- Then the normal bayesian approach is

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- We know from "l4.pdf" pages 11-16 that multiplying Gaussian likelihood with a Gaussian prior should yield a Gaussian result. However, this is a messy computation, so we will us a different approach to find the posterior.

Using Independence

- We know that $\boldsymbol{w}$ and θ are independent of each other, and because they are Gaussian, this means that their joint distribution is also Gaussian. Define:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H} & \mathrm{I} \\ \mathrm{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \theta \\ \boldsymbol{w} \end{bmatrix}$$

- Then the expectations are:

$$\mathrm{E}(\boldsymbol{z}) = \mathrm{E}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\theta} \end{bmatrix} \right) = \begin{bmatrix} \mathrm{E}(\boldsymbol{H\theta} + \boldsymbol{w}) \\ \mathrm{E}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}\mathrm{E}(\boldsymbol{\theta}) + \mathbf{0} \\ \boldsymbol{\mu_\theta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H\mu_\theta} \\ \boldsymbol{\mu_\theta} \end{bmatrix}$$

- The joint distribution is given by

$$p(\boldsymbol{x}, \boldsymbol{\theta}) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{H\mu_\theta} \\ \boldsymbol{\mu_\theta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{C_{xx}} & \boldsymbol{C_{x\theta}} \\ \boldsymbol{C_{x\theta}} & \boldsymbol{C_{\theta\theta}} \end{bmatrix} \right)$$

Covariance Matrices

- $\boldsymbol{C_{\theta\theta}}$ is easy, since that is just the covariance of theta $\boldsymbol{C_\theta}$.
- The covariance of $\boldsymbol{x}$ is influenced by the prior pdf:

$$
\begin{aligned}
\boldsymbol{C_{xx}} &= \mathrm{E}[(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x})(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))^{\mathrm{T}}] \\
&= \mathrm{E}(\boldsymbol{H\theta} + \boldsymbol{w} - \boldsymbol{H\mu_\theta})(\boldsymbol{H}\theta + \boldsymbol{w} - \boldsymbol{H\mu_\theta})^{\mathrm{T}} \\
&= \mathrm{E}(\boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\mu_\theta}) + \boldsymbol{w})(\boldsymbol{H}(\theta - \boldsymbol{\mu_\theta}) + \boldsymbol{w})^{\mathrm{T}} \\
&= \boldsymbol{H}\mathrm{E}[(\boldsymbol{\theta} - \boldsymbol{\mu_\theta})(\boldsymbol{\theta} - \boldsymbol{\mu_\theta})^{\mathrm{T}}]\boldsymbol{H^{\mathrm{T}}} + \mathbf{0} + \mathbf{0} + \mathrm{E}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) \\
&= \boldsymbol{H}\boldsymbol{C_\theta}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C_w}
\end{aligned}
$$

Covariance Matrices (Cont.)

- The cross covariance is given by

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{x}\boldsymbol{\theta}} &= \mathrm{E}[(\boldsymbol{\theta} - \mathrm{E}(\boldsymbol{\theta})(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))^{\mathrm{T}} \\
&= \mathrm{E}(\theta - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{H}\theta + \boldsymbol{w} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}})^{\mathrm{T}} \\
&= \mathrm{E}(\theta - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{H}(\theta - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \boldsymbol{w})^{\mathrm{T}} \\
&= \mathrm{E}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}))^{\mathrm{T}} + \mathbf{0} \\
&= \boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^{\mathrm{T}}
\end{aligned}$$

Finding the Actual Estimator

- Now that we have the covariance matrices, we can find the posterior distribution using the conditional Gaussian formula from lecture 1!
- $p(\boldsymbol{\theta}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu_\theta} + \boldsymbol{C_{x\theta}}\boldsymbol{C_{xx}^{-1}}(\boldsymbol{x} - \boldsymbol{H\mu_\theta}), \boldsymbol{C_{\theta\theta}} - \boldsymbol{C_x\theta}\boldsymbol{C_{xx}^{-1}}\boldsymbol{C_{x\theta}})$
- So our MMSE estimator is:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu_\theta} + \boldsymbol{C_\theta}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{C_\theta}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C_w})^{-1}(\boldsymbol{x} - \boldsymbol{H\mu_\theta})$$

- Compare this to the maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = (\boldsymbol{H}^{\mathbf{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathbf{T}}\boldsymbol{x}$$

- The bayesian case looks more complicated, but sometimes $\boldsymbol{H}^{\mathbf{T}}\boldsymbol{H}$ is not invertible, and by choosing the right covariance matrices we can fix that issue.
- In this case $\boldsymbol{C_\theta}$ "fixes" $\boldsymbol{H}$.

Example: Fourier Transform

- Recall the linear example about the Fourier series:

$$x[n] = \Sigma_{k=1}^{M} a_k \cos(\frac{2\pi kn}{N}) + b_k \sin(\frac{2\pi kn}{N})_+ w[n]$$

- We constructed a linear model by creating the matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ \cos(\frac{2\pi}{N}) & \dots & \cos(\frac{2\pi M}{N}) & \sin(\frac{2\pi}{N}) & \dots & \sin(\frac{2\pi M}{N}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos(\frac{2\pi(N-1)}{N}) & \dots & \cos(\frac{2\pi M(N-1)}{N}) & \sin(\frac{2\pi(N_1)}{N}) & \dots & \sin(\frac{2\pi M(N-1)}{N}) \end{bmatrix}$$

- For simplicity set

$$\boldsymbol{\mu_\theta} = \mathbf{0}$$
$$\boldsymbol{C_\theta} = \sigma_\theta^2 I$$
$$\boldsymbol{C_w} = \sigma_w^2 I$$

Bayesian Estimator of the Fourier Transform

- Using our formula, the Bayesian least squares estimator is

$$\hat{\boldsymbol{\theta}} = \mathbf{0} + \sigma_\theta^2 \boldsymbol{H}^{\mathrm{T}} (\sigma_\theta^2 \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} + \sigma_w^2 I)^{-1} (\boldsymbol{x})$$

- But $\boldsymbol{H}^{\mathbf{T}}\boldsymbol{H} = \frac{N}{2}\mathbb{I}$, which means we can simplify further.

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \sigma_\theta^2 (\frac{\sigma_\theta^2 N}{2} + \sigma_w^2) I)^{-1} \boldsymbol{H}^{\mathrm{T}} \boldsymbol{x} \\ &= \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2} \boldsymbol{H}^{\mathbf{T}} \boldsymbol{x} \end{aligned}$$

- From the last time we looked at this example,

$$\boldsymbol{H}^{\mathbf{T}}\boldsymbol{x} = \Sigma_{n=0}^{N-1} \cos(\frac{2\pi k n}{N}) x[n] \quad \text{or} \quad \Sigma_{n=0}^{N-1} \sin(\frac{2\pi k n}{N}) x[n]$$

- Our "almost" Fourier coefficients are:

$$\hat{a_k} = \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2} \Sigma_{n=0}^{N-1} \cos(\frac{2\pi kn}{N}) x[n]$$

$$\hat{b_k} = \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2} \Sigma_{n=0}^{N-1} \sin(\frac{2\pi kn}{N}) x[n]$$

- These look like Fourier transform coefficients.
- The prior variance here is important, as is it is relative size to the noise variance.