W&J Cap 6: Modelamento de dados e estimativa de parâmetros

- Grande atenção é dada nos livros-texto de estatística a métodos de estimativa de parâmetros e o cálculo de seus intervalos de confiança. Por exemplo, se os seus N dados Z_i seguem uma distribuição gaussiana

$$prob(z) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]$$

então a estatística

$$m = \frac{1}{N}\sum_i Z_i$$

é um bom estimador de μ e tem uma distribuição conhecida (uma gaussiana novamente). Ou, usando o método bayesiano, pode-se calcular a distribuição de probabilidade de μ dados os dados.

- Qualquer procedimento de modelamento de dados é apenas uma versão mais elaborada disso, supondo que as distribuições são conhecidas.

- Suponha que os dados Z_i foram medidos em vários valores de uma variável independente X_i e acreditamos que esses estão espalhados, com erros gaussianos, ao redor da relação funcional subjacente

$$\mu = \mu(x, \alpha_1, \alpha_2, ...)$$

na qual $\alpha_1, \alpha_2, ...$ são os parâmetros desconhecidos a relação.

W&J Cap 6: Modelamento de dados e estimativa de parâmetros

- Temos então que:

$$prob(z|, \alpha_1, \alpha_2, ...) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(z - \mu(x, \alpha_1, \alpha_2, ...))^2}{2\sigma^2}\right],$$

  e, pelo teorema da Bayes temos que a distribuição da probabilidade posterior para os parâmetro é dada por

$$prob(\alpha_1, \alpha_2, ...|Z_i, \mu) \propto \prod_i \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(z - \mu(x, \alpha_1, \alpha_2, ...))^2}{2\sigma^2}\right] prob(\alpha_1, \alpha_2, ...)$$

  incluindo, como sempre, a informação *a priori*.

- Nesse exemplo μ foi incluído como um dos dados para enfatizar que tudo depende desse ser o modelo correto.

- Do ponto de vita formal, isso encerra a tarefa. Temos a distribuição de probabilidade dos parâmetros do modelo, dados os dados observacionais ou experimentais.

W&J Cap 6: Modelamento de dados e estimativa de parâmetros

- Essa é uma abordagem bastante geral. No caso limite dos *a priori* não-informativos ou difusos, se parece muito com o método da máxima verossimilhança. Se a distribuição dos resíduos do modelo for de fato gaussiana, se aproxima muito do método dos mínimos quadrados.

- Ademais, com essa expressão, é muito fácil atualizar os valores dos parâmetros à medida que chegam novos dados. O posterior de um estágio de experimentação torna-se o *a priori* para o seguinte.

- Como já vimos, é muito fácil lidar com parâmetros indesejáveis através da marginalização. O resultado desse processo é a distribuição da variável de interesse levando em consideração *toda* a gama de valores plausíveis dos parâmetros indesejados.

W&J Cap 6: Modelamento de dados e estimativa de parâmetros

- O modelamento pode ser um processo muito caro em qualquer pesquisa. Não é a toa que aproximações analíticas foram desenvolvidas no passado.

- O processo de modelamento envolve a determinação de máximos ou mínimos se alguma função de mérito. Sem a ajuda das expressões analíticas isso significa calcular a função, e talvez também suas derivadas, muitas vezes.

- Como o modelo, ele mesmo, pode ser complexo, por exemplo se resultante de computação intensiva, calcula-lo repetidas vezes sobre um intervalo de valores de parâmetros pode ser proibitivo.

W&J Cap 6: Modelamento de dados e estimativa de parâmetros

- Outra dificuldade da abordagem bayesiana é a integração numérica. Problemas interessantes tendem a ter vários parâmetros. Operações como a marginalização ou as necessárias para discriminar modelos requerem integrações multidimensionais.

- Essas são muito custosas em termos de tempo e recursos computacionais e difíceis de serem checadas. Nesses casos qualquer aproximação analítica é de grande valia. Mais a frente veremos que teoremas poderosos permitem grandes simplificações.

- Convém enfatizar que esses podem estar errados. Nesse caso também estarão os parâmetros bem como seus erros, o que passa uma falsa sensação de segurança. Um exemplo clássico é a suposição sobre a distribuição dos resíduos de um modelo.

- É muito importante ter uma gama de modelos à disposição e sempre testar os modelos optimizados em função dos dados procurando por pontos discrepantes ou grupos de valores positivos ou negativos. Os testes de qui-quadrado e dos *runs* são muito úteis nesses casos.

O método da máxima verossimilhança

- O método da máxima verossimilhança é um antigo conhecido e já vem sendo usado desde o século XVI. Sua formulação atual foi feita por Fischer em 1922.
- A verossimilhança já foi descrita anteriormente como $prob(X_i|\alpha)$, onde X_i são os N dados e α um vetor de parâmetros. Supondo que os dados são distribuídos de acordo com uma dada função f, e que cada ponto é independente dos demais temos que a verossimilhança é dada por:

$$\mathcal{L}(X_1, X_2, ..., X_N) = \prod_{i=1}^{N} f(X_i|\alpha)$$

- Num ponto de vista clássico essa e'a probabilidade, dado α de se obter os dados. No modo bayesiano isso é proporcional à probabilidade de α, dados os dados e supondo *a prioris* difusos, o que para efeitos práticos significa que variam pouco dentro da região do pico da verossimilhança.
- Encontrar a constante que transforma a proporcionalidade em igualdade frequentemente envolve as integrações trabalhosas já mencionadas.

O método da máxima verossimilhança

- Sob qualquer ponto de vista o valor do pico de $\mathcal{L}$ é provavelmente uma escolha útil para a obtenção dos 'melhores' valores de α.

- Formalmente o MLE (*Maximum Likelihood Estimator*) de α é $\hat{\alpha}$, que é o valor de α que maximiza $\mathcal{L}(\alpha)$. Frequentemente isso pode ser econtrado via

$$\frac{\partial}{\partial \alpha} \ln \mathcal{L}(\alpha)|_{\alpha=\hat{\alpha}} = 0$$

  embora algumas vezes isso não seja possível, como veremos num exemplo à frente.

- Notem que estamos maximizando o *logaritmo* ao invés da grandeza em si. Isso é feito apenas por conveniência numérica e algébrica e não gera viéses.

- O MLE é uma *estatística* pois depende apenas dos dados e não dos parâmetros.

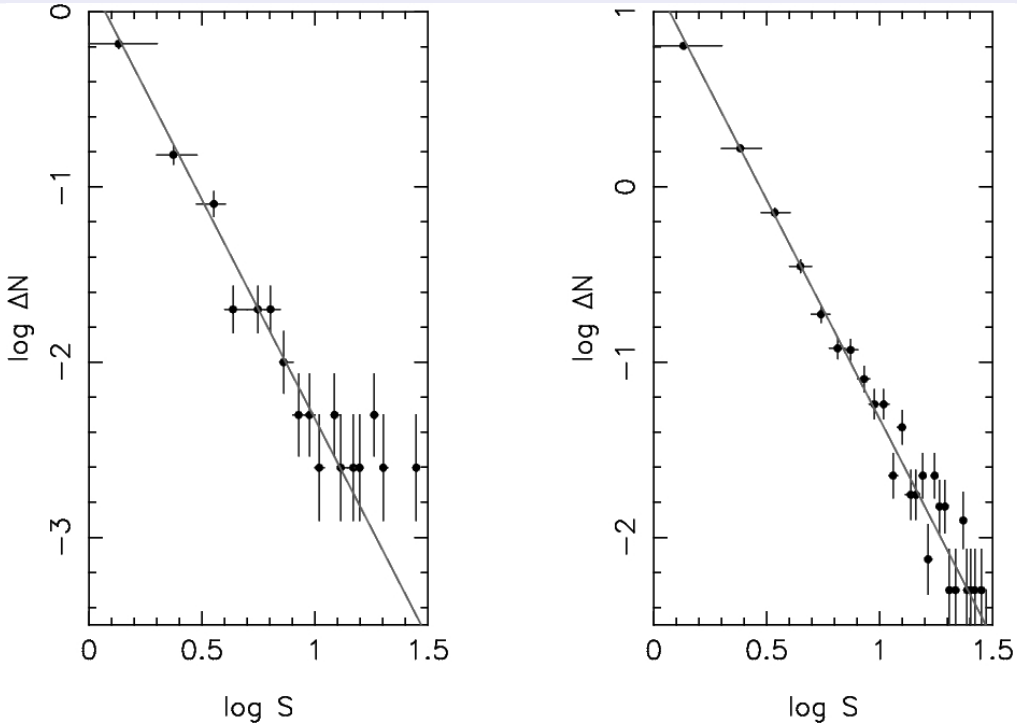
- Exemplo W&J P. 129

O método da máxima verossimilhança

- Com o exemplo anterior derivamos as fórmulas para o ajuste por mínimos quadrados ordinário (OLS), onde y é estimado em função da variável 'independente' x.
- Isso ocorre quando consideramos que Y tinha uma distribuição Normal em função do modelo, cujo espalhamento foi descrito por um valor apenas: σ. Obviamente, também se supos que o modelo da linha reta era o correto.
- Não é necessário que seja dessa forma. Pode haver um valor de σ_i associado a cada Y_i; a distribuição de y em relação À linha pode não ser gaussiana, pode ser uniforme ou depender de $|y_i - modelo|$ ao invés de $(y_i - modelo)^2$.
- A formulação é idêntica, ainda que a álgebra pode resultar não tão elegante quanto na regressão linear OLS.
- Por fim existe a uma vantagem do MLE que é que ele pode ser estimado numericamente, sem o uso de álgebra.

O método da máxima verossimilhança

Exemplo W&J P. 130



As figuras mostram uma contagem diferencial de fontes geradas via o método de monte Carlo com um desvio inicial uniforme que obedece uma lei de contagens do tipo $N(>S) = kS^{-1.5}$. A linhas representam a inclinação esperada de -2.5. *a*) $k = 1.0$ e 400 objetos; *b*) $k = 10.0$ e 4000 objetos. Os resultados obtidos via o MLE para as inclinações são respectivamente -2.52 ± 0.09 e -2.49 ± 0.03, onde as incertezas são dadas pelo ponto onde o *log* da verossimilhança diminui por um fator 2 em relação ao máximo. Os erros esperados para as inclinações são dados por $|\text{inclinação}|/\sqrt{\text{objetos}}$ e são 0.075 e 0.024

O método da máxima verossimilhança

- Uma vez que uma estimativa de MLE foi feita é essencial efetuar-se uma verificação final para saber se o ajuste é razoável. Em caso negativo ou há problemas nos dados ou no modelo.

- Há vários modos de se fazer essa verificação, tais como o teste de qui-quadrado e o teste K-S.

- No caso particular onde a distribuição dos resíduos entre o melhor ajuste do modelo e os dados é gaussiana o log da verossimilhança se torna a soma dos quadrados dos resíduos e caímos no famoso método dos mínimos quadrados.

O método da máxima verossimilhança: teoremas

- A razão mais forte para se usar o MLE é que esse apresenta uma mínima variância quando comparado com quaisquer outros métodos e está distribuído assintoticamente ao redor do valor verdadeiro. Por outro lado o MLE não é sempre 'não-tendencioso'.
- Se estimamos o vetor $\hat{\vec{\alpha}}$ pelo método da máxima verossimilhança, então os componentes do vetor estão distribuídos assintoticamente ao redor dos valores verdadeiros, como uma gaussiana multidimensional.
- No caso "assintoticamente" refere-se ao caso onde temos muitos dados. Estritamente falando, infinitos.
- A matriz de covariança que descreve essa gaussiana pode ser derivada pelas segundas derivadas da verossimilhança em relação aos parâmetros. Isso define a matriz chamada de *Hessiano*, que é

$$\mathcal{H} = \begin{bmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1 \partial \alpha_2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1 \partial \alpha_3} & \cdots \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2 \partial \alpha_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2 \partial \alpha_3} & \cdots \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3 \partial \alpha_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3 \partial \alpha_2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3^2} & \cdots \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix}$$

O método da máxima verossimilhança

- A matriz *Hessiana* depende claramente dos dados. Tomando seu valor esperado (o valor 'médio' de cada componente da matriz, ou $E[\mathcal{H}]$ para simplificar), temos uma expressão para simples para a matriz de covariança da distribuição gaussiana multiconjugada dos MLEs dos parâmetros:

$$C = (E[\mathcal{H}])^{-1}$$

  onde $(...)^{-1}$ significa a matriz inversa.

- A média (negativa) do Hessiano é importante o suficiente para ter um nome: matriz de informação de Fisher. Ela é importante porque descreve a largura da função de verossimilhança e, portanto, o espalhamento nos estimadores de máxima verossimilhança.

- A matriz de Fisher pode ser calculada para vários desenhos experimentais como uma medida do quão bem sucedido será esse experimento.

- A distribuição de probabilidades dos nossos N MLE's $\hat{\vec{\alpha}}$ é então

$$prob(\hat{\alpha}_1, \hat{\alpha}_2, ...) = \frac{1}{\sqrt{(2\pi)^N |\det C|}} exp\left[-\frac{1}{2}(\hat{\alpha} - \alpha) \cdot C^{-1} \cdot (\hat{\alpha} - \alpha)^\intercal\right]$$

- Assim sendo os MLE ($\hat{\vec{\alpha}}$) se distribuem ao redor do valor verdadeiro $\vec{\alpha}$ com um espalhamento descrito pela matriz de covariância C ou, equivalentemente, pela matriz de Fisher.

O método da máxima verossimilhança

- Porque razão as MLRs obedecem a esse teorema? Analisemos um caso simples de uma gaussiana de média μ e variância σ^2. Ignorando as constantes, dados N dados X_i, o logaritmo da verossimilhança é:

$$\ln \mathcal{L} = \frac{-1}{2\sigma^2} \sum_i (X_i - \mu)^2 - N \ln \sigma$$

e

$$\frac{-\partial^2 \ln \mathcal{L}}{\partial \mu^2} = \frac{N}{\sigma^2}$$

- Essa é a "matriz" Hessiana do problema. É simples obter o inverso da expectativa nesse caso e o resultado nos dá a variância na estimativa da média como σ^2/N, como já poderíamos antecipar.

- Esse exemplo ajudar a que se aceite o teorema. Os exercícios trazem outros exemplos, não tão simples que ajudarão mais nesse processo.

-

Exemplo W&J P. 133

Fim da aula 9