Discussion: Model Misspecification

Larry Wasserman

Background + Signal

$$X_1, \ldots, X_n \sim (1 - \lambda) p_0(x; \gamma) + \lambda p_1(x; \theta)$$

Background + Signal

$$X_1, \ldots, X_n \sim (1 - \lambda) p_0(x; \gamma) + \lambda p_1(x; \theta)$$

Background $= p_0(x; \gamma)$
Signal $= p_1(x; \theta)$

Background + Signal

$$X_1, \ldots, X_n \sim (1-\lambda)p_0(x;\gamma) + \lambda p_1(x;\theta)$$

Background $= p_0(x;\gamma)$
Signal $= p_1(x;\theta)$

Test $H_0 : \lambda = 0$

Background + Signal

$$X_1, \ldots, X_n \sim (1-\lambda)p_0(x;\gamma) + \lambda p_1(x;\theta)$$

Background $= p_0(x;\gamma)$
Signal $= p_1(x;\theta)$

Test $H_0 : \lambda = 0$

Confidence interval for θ

Systematic Bias

Model misspecification (non-statistical error)

$$\inf_{\gamma} \int (p_0(x;\gamma) - p_{0,true}(x))^2 dx$$

Can cause false positives and low power.

Systematic Bias

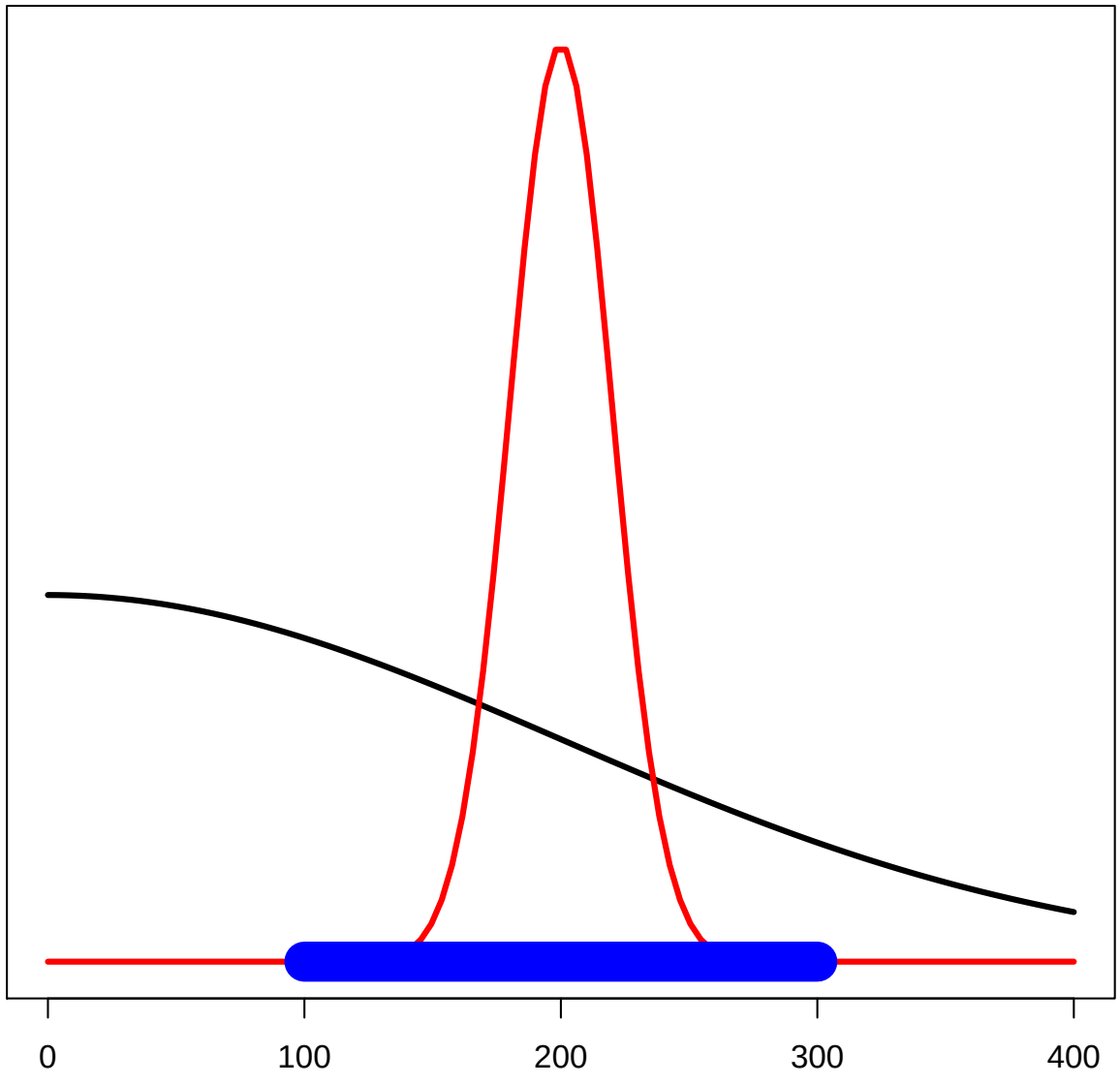
Model misspecification (non-statistical error)

$$\inf_{\gamma} \int (p_0(x;\gamma) - p_{0,true}(x))^2 dx$$

Can cause false positives and low power.

How to include this as part of the uncertainty? Very difficult.

Example



Estimating the Background

$$p_0(x;\gamma) = \sum_{j=1}^{k} \gamma_j \phi_j(x)$$

or

$$p_0(x;\gamma) = \frac{e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}{\int e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}$$

where $\phi_1, \phi_2, \ldots$ are basis functions.

Estimating the Background

$$p_0(x;\gamma) = \sum_{j=1}^{k} \gamma_j \phi_j(x)$$

or

$$p_0(x;\gamma) = \frac{e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}{\int e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}$$

where $\phi_1, \phi_2, \ldots$ are basis functions.

Choosing k involves a bias/variance tradeoff:

Estimating the Background

$$p_0(x;\gamma) = \sum_{j=1}^{k} \gamma_j \phi_j(x)$$

or

$$p_0(x;\gamma) = \frac{e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}{\int e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}$$

where $\phi_1, \phi_2, \ldots$ are basis functions.

Choosing k involves a bias/variance tradeoff:

$$\underbrace{\sum_{j=k+1}^{\infty} \beta_j^2}_{\text{bias}} + \underbrace{\frac{k}{n}}_{\text{variance}}$$

Estimating the Background

$$p_0(x;\gamma) = \sum_{j=1}^{k} \gamma_j \phi_j(x)$$

or

$$p_0(x;\gamma) = \frac{e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}{\int e^{\sum_{j=1}^{k} \gamma_j \phi_j(x)}}$$

where $\phi_1, \phi_2, \ldots$ are basis functions.

Choosing k involves a bias/variance tradeoff:

$$\underbrace{\sum_{j=k+1}^{\infty} \beta_j^2}_{\text{bias}} + \underbrace{\frac{k}{n}}_{\text{variance}}$$

The mixture model is not identified if k is unrestricted.

Approaches

Approaches

1. Use all the data and fit the mixture. (Must restrict k)

Approaches

1. Use all the data and fit the mixture. (Must restrict k)
2. Estimate background using the sideband/control region then extrapolate to signal region. (No restriction on k but we may have extrapolation bias.)

Approaches

1. Use all the data and fit the mixture. (Must restrict k)
2. Estimate background using the sideband/control region then extrapolate to signal region. (No restriction on k but we may have extrapolation bias.)
3. Use other data: $Z_1, \ldots, Z_n \sim q$. Then use an exponential tilt:

$$p(y) \propto q(y) e^{\theta^T \phi(y)}$$

Estimate θ over sideband. Semiparametric: p and q are unrestricted.

Approaches

1. Use all the data and fit the mixture. (Must restrict k)
2. Estimate background using the sideband/control region then extrapolate to signal region. (No restriction on k but we may have extrapolation bias.)
3. Use other data: $Z_1, \ldots, Z_n \sim q$. Then use an exponential tilt:

$$p(y) \propto q(y) e^{\theta^T \phi(y)}$$

Estimate θ over sideband. Semiparametric: p and q are unrestricted.

or use optimal transport (morphing). See Tudor Manole's talk on Wednesday for more on this.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Power Divergence

Minimize

$$S(\gamma) = \int p_0(x;\gamma)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(x;\gamma)^{\alpha}$$

Power Divergence

Minimize

$$S(\gamma) = \int p_0(x;\gamma)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(x;\gamma)^{\alpha}$$

$\alpha \to 0$ gives mle.

Power Divergence

Minimize

$$S(\gamma) = \int p_0(x;\gamma)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(x;\gamma)^{\alpha}$$

$\alpha \to 0$ gives mle.

$\alpha > 0$ much more robust.

Power Divergence

Minimize

$$S(\gamma) = \int p_0(x;\gamma)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(x;\gamma)^{\alpha}$$

$\alpha \to 0$ gives mle.

$\alpha > 0$ much more robust.

$\alpha = 1$ is L_2: $\int (p - p_\gamma)^2$.

Power Divergence

Minimize

$$S(\gamma) = \int p_0(x;\gamma)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(x;\gamma)^{\alpha}$$

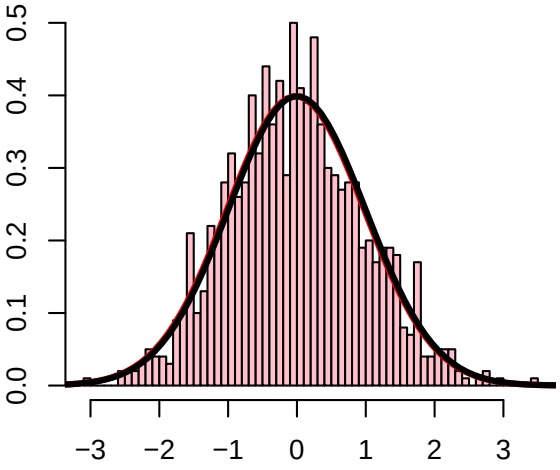
$\alpha \to 0$ gives mle.

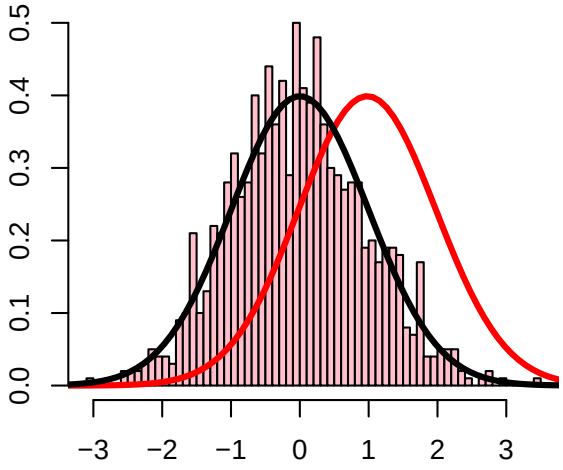
$\alpha > 0$ much more robust.

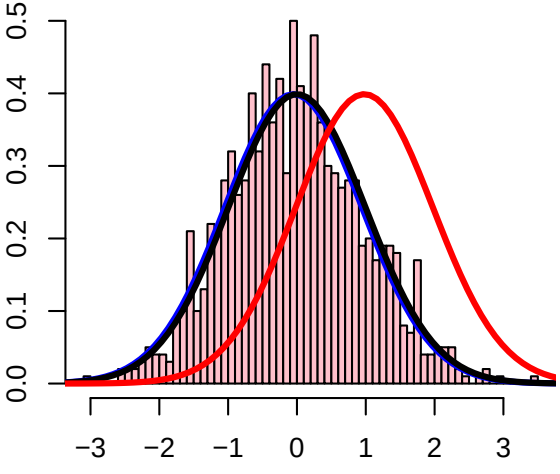
$\alpha = 1$ is L_2: $\int (p - p_\gamma)^2$.

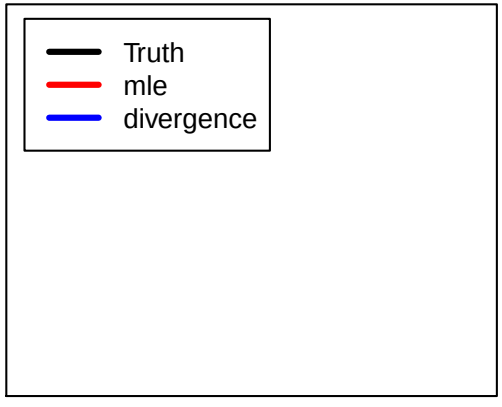
$\alpha \uparrow$: robustness $\uparrow$ and efficiency $\downarrow$

Example


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


Truth
mle
divergence

Conclusion

Conclusion

- Model misspecification is very difficult. Bias not variance.

Conclusion

- Model misspecification is very difficult. Bias not variance.

- Comparing several methods? Does using the difference between methods as a measure of systematic bias make sense?

Conclusion

- Model misspecification is very difficult. Bias not variance.

- Comparing several methods? Does using the difference between methods as a measure of systematic bias make sense?

- Does using robust methods help?

Conclusion

- Model misspecification is very difficult. Bias not variance.

- Comparing several methods? Does using the difference between methods as a measure of systematic bias make sense?

- Does using robust methods help?

- THE END