Distances and their role in robustness

Marianthi Markatou

Dept. of Biostatistics

SUNY, Buffalo

Joint work with the late Bruce G. Lindsay

Acknowledgments

THANKS TO

- Dr. Rositsa Dimova
- Dr. Ayan Basu
- Dr. Abhik Ghosh
- NSF, USA
- SUNY, Buffalo Biostatistics and SBMS

OUTLINE

- Introduction
- Some classes of distances and their properties
- Related recent work
- Discussion and conclusions

DIRECTION WE COME FROM: MINIMUM DISTANCE METHODS

- The work presented relates to minimum distance methods originated by Beran (1977)
- Interested in *robust modeling—much more in the context of misspecification of model in an unanticipated way.*
- That was original thought but we diverged on the way to discovery!

INTRODUCTION

- Distance measures play a ubiquitous role in statistical theory and thinking
- Within the statistical literature this role has been played out mostly behind the scenes with other aspects of the statistical problem being viewed as more central, more important or more interesting.

INTRODUCTION

- **Estimation:** we often use estimators based on minimizing a distance, explicitly or implicitly, but we rarely study how the *properties of a distance determine the properties of the estimators*.
- **Goodness-of-fit:** the usual question we ask is “how powerful a method is against a set of interesting alternatives” not *what aspects of the difference between the hypothetical model and the alternative are we measuring?*

INTRODUCTION

- Statistical distances are defined in a variety of ways. For example, we can define distances by comparing distribution functions, probability mass/density functions, characteristic functions or moment generating functions.
- Discrete and continuous analogues
- A lot of literature from many scientific fields
- We concentrate on “density” based “distances”

ROBUSTNESS IN MODELING

- Let τ denote the unknown probability law that generated data d and let m denote the model used to describe the data. Our goal is to identify what models m are *good descriptors* of τ.
- **Central tenet:** If *M* is a class of models the role of *M* is to serve as a good approximation to τ. The quantity that measures the quality of approximation offered by *M* is the *statistical distance.*

ROBUSTNESS IN MODELING

- Let τ, m be two probability “densities”. Then d(τ, m) is a *statistical distance* iff d(τ, m)≥ 0 with equality when τ=m for all statistical purposes.
- *Interpretation*: d(τ, m) measures “lack of fit” or “model adequacy” in the sense that larger values of d mean that the model element m is a worse fit to τ for our purposes.

ROBUSTNESS

- How does robustness fit into this picture?
- In robustness literature there is a denial of the model’s truth. Following this logic we start with *goodness of fit by identifying a measure that assesses whether the “model” fits the data “adequately”*. Then we examine whether this measure of adequacy is *robust and in what sense.*

EXAMPLES OF DISTANCES

- The class of power divergence measures (Cressie and Read, 1984) is defined as

$$d(\tau, m) = \frac{1}{\lambda(\lambda+1)} \sum \tau(t)\{(\frac{\tau(t)}{m(t)})^{\lambda} - 1\}.$$

For λ= -2 obtain Neyman’s chi-squared

$$d(\tau, m) = \frac{1}{2} \sum \frac{[\tau(t) - m(t)]^2}{\tau(t)}$$

EXAMPLES OF DISTANCES

- For λ= -1 obtain Kullback-Leibler divergence given as

$$KL(\tau, m) = \sum m(t) \ln(\frac{m(t)}{\tau(t)})$$

For λ= -1/2 obtain twice the squared Hellinger distance given as

$$HD(\tau, m) = 2 \sum \{\sqrt{\tau(t)} - \sqrt{m(t)}\}^2$$

EXAMPLES OF DISTANCES

- For λ=0 obtain the likelihood disparity given as

$$LD(\tau, m) = \sum \tau(t) \ln(\frac{\tau(t)}{m(t)})$$

For λ =1 obtain Pearson’s chi-squared divided by 2 and given as:

$$PCS(\tau, m) = \frac{1}{2} \sum \frac{[\tau(t) - m(t)]^2}{m(t)}$$

BLENDED CHI-SQUARED

- In a chi-squared distance one can modify the “weights” given to squared discrepancies. Let 0≤ α ≤ 1, then the blended chi-squared distances (Lindsay, 1994) are defined as:

$$BWCS(\alpha) = \sum \frac{[\tau(t) - m(t)]^2}{2[\alpha\tau(t) + (1-\alpha)m(t)]}$$

α=0 then PCS, α=1 gives NCS and α=1/2 gives symmetric chi-squared.

Blended weighted Hellinger distance

- The blended weighted Hellinger distance is defined as follows:

$$BWHD(\alpha) = \sum \frac{[\tau(t) - m(t)]^2}{2[\alpha\sqrt{\tau(t)} + (1-\alpha)\sqrt{m(t)}]^2}$$

where α=0 we obtain PCS, α=1 gives NCS and α=1/2 obtains Hellinger distance.

RESIDUAL SYSTEMS

- **Pearson Residuals** are defined as $\delta(t) = \frac{\tau(t)}{m(t)}$ - 1, take values in [-1, ∞)**,** and it is called so because PCS = $\sum[\delta(t)]^2 m(t)$**.**

- **Symmetrized Residuals** are defined as $r_{sym}(\mathrm{t}) = \frac{\tau(t) - m(t)}{\tau(t) + m(t)}$, and are defined in [-1, 1].
- Many of the distances above can be expressed in terms of these residuals.

CHI-SQUARED MEASURES

- Define the *generalized chi-squared distances*

$$\sum \frac{[\tau(t)-m(t)]^2}{a(t)},$$

where if $a(t) = m(t)$ we obtain PCS, if $a(t)$ =d(t) we obtain NCS and if $a(t) = 0.5\tau(t) + 0.5m(t)$ we obtain

symmetric chi-squared.

There is *a strong dependence of the properties of the distance and the denominator* $a(t)$.

CHI-SQUARED MEASURES: ROBUSTNESS & INTERPRETATIONS

- We can write the *Pearson chi-squared* statistic as follows:

$$PCS = \ sup_h \frac{[E_\tau\big(h(X)\big) - E_m\big(h(X)\big)]^2}{Var_m(h(X))}$$

$$= \ sup_h \frac{[(\frac{1}{n}) \sum h(x_i) - E_m(h(X))]^2}{Var_m(h(X))}$$

$$= \ \frac{1}{n} sup_h Z^2(h),$$

that is, PCS is the supremum of z-statistics so it is *highly non-robust*.

ROBUSTNESS & INTERPRETATIONS

- Similarly, we can write the *Neyman chi-squared statistic* as

$$NCS = \frac{1}{n} sup_h[t(h)]^2,$$

where t(h) is a t-statistic; thus NCS is *robust*.

Therefore, a small distance means that the *means of the two probability models* τ, m are close on the scale of standard deviation.

ROBUSTNESS & INTERPRETATIONS

- In general, the generalized chi-squared measure can be expressed as:

$$sup_h \frac{[E_\tau\left(h(X)\right) - E_m\left(h(X)\right)]^2}{Var_a(h(X))},$$

with $a(t)$ a general density function. If $a(t) = 0.5\tau(t) + 0.5m(t)$ we obtain symmetric chi-squared, a *robust measure* in the sense of exhibiting 50% BP (Markatou et al, 1998, Lindsay, 1994)

SYMMETRIC CHI-SQUARED

- Furthermore, we can express the symmetric chi-squared distance in terms of the symmetrized residuals as follows:

$$S(\tau, m)^2 = 4\sum b(t)[\tfrac{\tau(t)-m(t)}{\tau(t)+m(t)}]^2$$

$$= 4\sum b(t)[r_{sym}(t)]^2,$$

where $b(t) = 0.5\tau(t) + 0.5m(t)$.

SYMMETRIC CHI-SQUARED: TESTING INTERPRETATION

- The symmetric chi-squared distance can be viewed as a good compromise between the non-robust Pearson chi-squared and the robust Neyman chi-squared. It is a metric, that is it satisfies the triangle inequality and it is symmetric, and has the following testing interpretation.

TESTING INTERPRETATION

- Let ϕ be a test function and consider testing $H_0\colon\ \tau = f\ vs\ H_1\colon\ \tau = g$. Let θ be a random variable taking the value 1 if H_1 is true and taking value 0 if H_0 is true. The solution φ_{opt} of the optimization problem

$$min_{\phi} E_{\pi}[\left(\theta - \phi(X)\right)^2],$$

where $\pi(\theta)$, the prior probability on θ, is given by $\pi(\theta) = \frac{1}{2}\ , \theta = 0\ and\, \frac{1}{2}, \theta = 1$, is not a 0-1

TESTING INTERPRETATION

decision, but equals the posterior expectation of θ given X. That is,

$$\phi(t) = E(\theta|X = t) = P(\theta = 1|X = t)$$

$$= \frac{0.5g(t)}{0.5f(t)+0.5g(t)}$$

the probability that the alternative is correct.

EFFICIENCY & ROBUSTNESS

- These distances introduce a method of estimation that produces first-order efficient estimators—the concept that quantifies robustness is the *residual adjustment function* (Lindsay, 1994). It is defined as an increasing, twice differentiable function A on [-1, ∞), such that $A(0) = 0$ and $A'(0) = 1$. The function $A(\delta) = \delta$ corresponds to the MLE.
- The RAFs introduced by the power divergences are given as follows:

EFFICIENCY & ROBUSTNESS

$$A_\lambda(\delta) = \frac{(1+\delta)^\lambda}{\lambda+1},$$

λ=0 corresponds to MLE,
λ=-1/2 corresponds to Hellinger distance,
λ=1 corresponds to PCS,
λ=-2 corresponds to NCS,
λ=-1 corresponds to KL.

QUDRATIC DISTANCES

Definition: Let ℵ be a sample space. The function K(s, t) is a bounded, symmetric, non-negative definite kernel defined on ℵ x ℵ if the relationship $\iint K(s, t) d\sigma(s) d\sigma(t) \geq 0$ holds for all bounded, signed measures σ.

Definition: Given a non-negative definite kernel function K(s, t) the K-based quadratic distance between two probability measures F and G is defined as :

D(F, G) = $\iint K(s, t) d(F-G)(s) d(F-G)(t)$.

Quadratic distances can be thought of as extensions of the chi-squared distance—by construction exhibit *discretization robustness* as they can be written as follows:

QUADRATIC DISTANCES

$$D(F,G) = \int [f^*(t) - g^*(t)]^2 dt\,,$$

where $$f^*(t) = \int K^{\frac{1}{2}}(t,r)dF(r).$$

The family of chi-squared measures has a simple and attractive interpretation as a loss function. It gets at the very point of *making sure that the model m and the true unknown mechanism τ have the same structure in means, on the standard deviation scale.*

QUADRATIC DISTANCES

- By *discretization robustness* we mean robustness of the distance measure to the difference between continuous distributions and discrete distributions—data are always discrete
- Quadratic distance can be interpreted as a loss function

QUADRATIC DISTANCES

- An unbiased estimator of the distance is provided by the U-statistic

$$U_n = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K_{cen}(x_i, x_j)$$

where $K_{cen}(x, y) = K(x, y) - K(x, G) - K(G, y) + K(G, G), K(x, G) = \int K(x, y) dG(y)$.

OBSERVATIONS

- Outliers can influence chi-squared measures more, because they are based on means. I.e. PCS can be made dramatically large by increasing the amount of data in a cell with small probability under the model. In fact, if there is data in a cell with model probability 0 the distance is infinite.
- The symmetric chi-squared, by averaging data and model in the denominator, avoids such an extreme behavior.

OBSERVATIONS

- The symmetric chi-squared has also two attractive features; 1) it is a metric, and 2) because it is closely linked with the total variation distance, since total variation is robust, this property carries to some degree over to the symmetric chi-squared (see Lindsay, 1994; Markatou et al, 1998, and Markatou, 2000 where they show that the symmetric chi-squared generates highly efficient and robust methods of estimation).

DENSITY POWER DIVERGENCES

- Basu et al (1998, Biometrika)
- Consider a parametric family of models $\{F_\theta\}$, indexed by an unknown parameter θϵΘ possessing densities $\{f_\theta\}$ with respect to the Lebesgue measure, and let *G* be the class of all distributions with densities g. Define the *density power divergences* as follows:

DENSITY POWER DIVERGENCES

$$d_\alpha(g,f) = \int \{f^{1+\alpha}(z) - \left(1+\frac{1}{\alpha}\right) g(z) f^{\alpha}(z) +$$

$$\frac{1}{\alpha} g^{1+\alpha}(z)\} dz \,, \quad \text{for all } \alpha > 0.$$

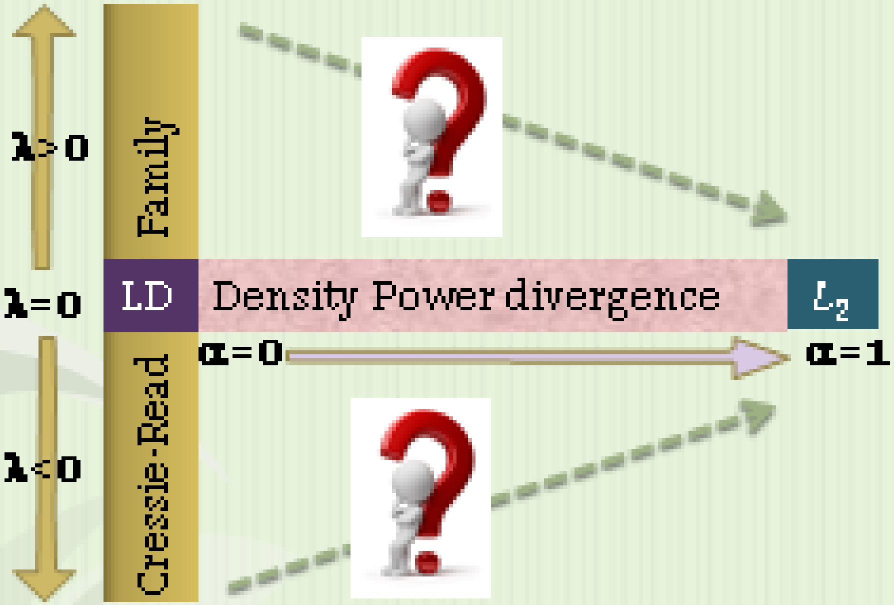
When α = 0, the power divergence equals the Kullback-Leibler distance.

The minimization of the above function provides a method of estimation that balances the trade-off between efficiency and robustness.

DENSITY POWER DIVERGENCES

- When α = 1 the density power divergence becomes the L_2 distance between g, f.
- Other power divergence measures were also developed such as logarithmic density power divergence (Jones et al, 2001), S-divergence family (Ghosh et.al., 2013), etc.
- The first-order influence function *cannot portray* the true robustness properties of the minimum divergence estimators in the S-divergence family.

2. New Divergence Measures and their applications



- DPD family connects one member of the PD family to the L_2-divergence with increasing robustness
- **S - Divergence family** connects the whole PD family to the L_2-divergence

- It is defined in terms of two parameters α and λ as follows:

$$S_{(\alpha,\lambda)}(g,f) = \frac{1}{A}\int f^{1+\alpha} - \frac{1+\alpha}{AB}\int f^{B} g^{A} + \frac{1}{B}\int g^{1+\alpha},$$

where $A = 1 + \lambda(1-\alpha)$ and $B = \alpha - \lambda(1-\alpha)$.

DISCUSSION

- Distances present an approach that is able to address model robustness;
- A lot of work remains to be done to understand the extend to which this approach is useful;
- Density power divergences address the balance between efficiency and robustness through the parameter α