On **models** in Bruno's work, and model selection

Yosi Rinott
The Hebrew U of Jerusalem and LUISS, Rome

Giornata in ricordo di Bruno Bassan, October 2014

In this talk:

1. On some **models** in Bruno's work.

2. Let's think together about a model I was asked about last week. I think Bruno might have liked the story.

3. Some comments related to my recent interest in model selection in Statistics.

Bruno was mostly an Applied Probabilist. His range of application was very wide: Stochastic Control Theory, Mathematical Finance, Game Theory, Reliability Theory, Dependence and Stochastic Orderings, Statistics, and more.

Applied probability most often starts with a model.

What is a model?

Go to Google ...

Bruno was mostly an Applied Probabilist. His range of application was very wide: Stochastic Control Theory, Mathematical Finance, Game Theory, Reliability Theory, Dependence and Stochastic Orderings, Statistics, and more.

Applied probability most often starts with a model.

What is a model?

Go to Google ...

Oxford Online Dictionary: “*A model is a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions*"

Oxford Online Dictionary: “*A model is a simplified description, especially a mathematical one, of a system or process, to assist calculations* [prove theorems, do numerical calculation/examples, and understand the system] *and predictions*"

Oxford Online Dictionary: “*A model is a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions*"

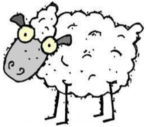
G. Box : “*all models are wrong, but some are useful*"

Stanford Encyclopedia of Philosophy: “*A ‘good’ scientific theory, Popper thus argued, has a higher level of verisimilitude [verosimiglianza, likelihood] than its rivals*"

SOME EXAMPLES. Bassan and Scarsini 1998

Once upon a time there was a village of shepherds who pastured their flocks. One day the grazing grounds become parched, and the chief of the village decides that all the shepherds should move to a different area. The chief and the shepherds know that wolves dwell in all but one of the paths ... The chief faces two contrasting tendencies as she releases more information [about where the wolves might be, with some uncertainty or probability]: the more each shepherd is informed, the more likely he is to make the good decision, and this is socially desirable; on the other hand, the more the shepherds know, the less likely they are to diversify their behavior. This, in view of the requirement that at least one shepherd survive, could be socially detrimental.

The model that we study in this paper involves a central planner, who can release information to agents whose behavior affects the social welfare of a community. Each of the agents has a utility function ...



The model that we study in this paper involves a central planner, who can release information to agents whose behavior affects the social welfare of a community. Each of the agents has a utility function ...

2. The model

Consider a measurable space $(\Omega, \mathscr{F})$ endowed with a filtration $\{\mathscr{F}_t \mid t \in T \subset \mathbb{N}\}$. Let $\mathscr{P}(\Omega, \mathscr{F})$ be the set of σ-additive probability measures on $(\Omega, \mathscr{F})$... At time t each agent adopts the decision that maximizes his own expected utility...

Bassan, Bruno and Natoli, Giuseppina 2004

We present here a toy **model** for pricing options in terms of game-theoretic concepts, without resorting explicitly to no-arbitrage or hedging. In particular, Aumann's Theorem about the impossibility of 'agreeing to disagree' is invoked. In our simple setup, there are only one investor and one financial intermediary, the bank, who are considering the trade of a share of a stock and of one option issued on it. We consider here perpetual American options...

In finance, a perpetual American option is a contract which gives the buyer the right, but not the obligation, to buy at any future time an underlying asset or instrument at a specified strike price.

Bassan, Rinott, Vardi

A simple problem in an unfinished paper with Yehuda Vardi and me: let X_i denote random service times of customers who arrive one after the other from a queue to a server. We assume that they are independent, from a common distribution.

```
--- |----------  X1  ---------- | -- X2 --| -----------  X3------------| ---- X4 ----| -----
                                          |----- Z ------|  --- Y --- |
-------|--------------------------|----------|---------------t------------ |--------------|------------
```

We observe the situation at a random time t. Let Z denote service time until t and Y the remaining service time, so $Z + Y = X$. Yehuda looked at many independent such systems (servers), and noticed on some data, that the Y's tend to be larger than the X's. If a person before you in a queue is already being served for a long time, he is likely to take much time to finish.

Bassan, Rinott, Vardi

A simple problem in an unfinished paper with Yehuda Vardi and me: let X_i denote random service times of customers who arrive one after the other from a queue to a server. We assume that they are independent, from a common distribution.

```
         --- |----------  X1  ---------- | -- X2 --| -----------  X3------------| ---- X4 ----| -----
                                                   |-----  Z  ------|  ---  Y --- |
       -------|--------------------------|----------|---------------t------------ |--------------|------------
```

We observe the situation at a random time t. Let Z denote service time until t and Y the remaining service time, so $Z + Y = X$. Yehuda looked at many independent such systems (servers), and noticed on some data, that the Y's tend to be larger than the X's. If a person before you in a queue is already being served for a long time, he is likely to take much time to finish.

Bassan, Rinott, Vardi

A simple problem in an unfinished paper with Yehuda Vardi and me: let X_i denote random service times of customers who arrive one after the other from a queue to a server. We assume that they are independent, from a common distribution.

```
--- |----------  X1  ---------- | -- X2 --| -----------  X3------------| ---- X4 ----| -----
                                            |-----  Z  ------|  ---  Y --- |
-------|--------------------------|----------|---------------t------------ |--------------|------------
```

We observe the situation at a random time t. Let Z denote service time until t and Y the remaining service time, so $Z + Y = X$. Yehuda looked at many independent such systems (servers), and noticed on some data, that the Y's tend to be larger than the X's. If a person before you in a queue is already being served for a long time, he is likely to take much time to finish.

Let's think of a model (Avrahami and Kareev)

Suppose that in town there exist **k** restaurants. You have tried them all, and the probability of choosing this or that restaurant for your next dinner is a **function** (which?) of your (recent) past experience. For simplicity assume two kinds of restaurants, bad (0) and good (1). You may come out of a bad restaurant with a good impression with some (small) probability, and vice versa.

Does more choice lead to better dining? for example, if there are **two bad and two good restaurants** in town, are you going to eat better than in the case of **one bad and one good restaurants**?

We need to construct a 'consistent' decision function for varying **k**'s.

Let's think of a model (Avrahami and Kareev)

Suppose that in town there exist **k** restaurants. You have tried them all, and the probability of choosing this or that restaurant for your next dinner is a **function** (which?) of your (recent) past experience. For simplicity assume two kinds of restaurants, bad (0) and good (1). You may come out of a bad restaurant with a good impression with some (small) probability, and vice versa.

Does more choice lead to better dining? for example, if there are **two bad and two good restaurants** in town, are you going to eat better than in the case of **one bad and one good restaurants**?

We need to construct a 'consistent' decision function for varying k's.

Let's think of a model (Avrahami and Kareev)

Suppose that in town there exist **k** restaurants. You have tried them all, and the probability of choosing this or that restaurant for your next dinner is a **function** (which?) of your (recent) past experience. For simplicity assume two kinds of restaurants, bad (0) and good (1). You may come out of a bad restaurant with a good impression with some (small) probability, and vice versa.

Does more choice lead to better dining? for example, if there are **two bad and two good restaurants** in town, are you going to eat better than in the case of **one bad and one good restaurants**?

We need to construct a 'consistent' decision function for varying **k**'s.

Suggestion 1. $\mathbf{k} = 2$: suppose, for example that the true state of affairs is (0,1), that is, one bad one good, and we use only most recent experience. If it is (0,0) or (1,1) choose at random, if (0,1), say, choose the second with probability $p = 0.95$, say, so the probability ratio is 19 to 1.

For $\mathbf{k} = 4$: keep the ratio, so for example if the current experience is (0,0,1,1) choose either the 3rd or the 4th restaurant with probability 19/(19+19+1+1). Good model? How is it affected by more choice? If the true and current state is (0,0,...,0,1) then instead of choosing the one good restaurant, you will go with high probability to one of the bad ones.

Suggestion 2. Choose good with probability 0.95 and then at random among the good ones, and bad with probability 0.05, and then at random.

Once 'such' a model is chosen, it is a finite Markov chain, and one can compute a stationary distribution, etc.

Suggestion 1. **k** $= 2$: suppose, for example that the true state of affairs is (0,1), that is, one bad one good, and we use only most recent experience. If it is (0,0) or (1,1) choose at random, if (0,1), say, choose the second with probability $p = 0.95$, say, so the probability ratio is 19 to 1.
For **k** $= 4$: keep the ratio, so for example if the current experience is (0,0,1,1) choose either the 3rd or the 4th restaurant with probability 19/(19+19+1+1). Good model? How is it affected by more choice?
If the true and current state is (0,0,...,0,1) then instead of choosing the one good restaurant, you will go with high probability to one of the bad ones.

Suggestion 2. Choose good with probability 0.95 and then at random among the good ones, and bad with probability 0.05, and then at random.

Once 'such' a model is chosen, it is a finite Markov chain, and one can compute a stationary distribution, etc.

Suggestion 1. **k** $= 2$: suppose, for example that the true state of affairs is (0,1), that is, one bad one good, and we use only most recent experience. If it is (0,0) or (1,1) choose at random, if (0,1), say, choose the second with probability $p = 0.95$, say, so the probability ratio is 19 to 1.
For **k** $= 4$: keep the ratio, so for example if the current experience is (0,0,1,1) choose either the 3rd or the 4th restaurant with probability 19/(19+19+1+1). Good model? How is it affected by more choice?
If the true and current state is (0,0,...,0,1) then instead of choosing the one good restaurant, you will go with high probability to one of the bad ones.

Suggestion 2. Choose good with probability 0.95 and then at random among the good ones, and bad with probability 0.05, and then at random.

Once 'such' a model is chosen, it is a finite Markov chain, and one can compute a stationary distribution, etc.

Suggestion 1. $\mathbf{k} = 2$: suppose, for example that the true state of affairs is (0,1), that is, one bad one good, and we use only most recent experience. If it is (0,0) or (1,1) choose at random, if (0,1), say, choose the second with probability $p = 0.95$, say, so the probability ratio is 19 to 1.

For $\mathbf{k} = 4$: keep the ratio, so for example if the current experience is (0,0,1,1) choose either the 3rd or the 4th restaurant with probability 19/(19+19+1+1). Good model? How is it affected by more choice? If the true and current state is (0,0,...,0,1) then instead of choosing the one good restaurant, you will go with high probability to one of the bad ones.

Suggestion 2. Choose good with probability 0.95 and then at random among the good ones, and bad with probability 0.05, and then at random.

Once 'such' a model is chosen, it is a finite Markov chain, and one can compute a stationary distribution, etc.

Suggestion 1. $\mathbf{k} = 2$: suppose, for example that the true state of affairs is (0,1), that is, one bad one good, and we use only most recent experience. If it is (0,0) or (1,1) choose at random, if (0,1), say, choose the second with probability $p = 0.95$, say, so the probability ratio is 19 to 1.
For $\mathbf{k} = 4$: keep the ratio, so for example if the current experience is (0,0,1,1) choose either the 3rd or the 4th restaurant with probability 19/(19+19+1+1). Good model? How is it affected by more choice?
If the true and current state is (0,0,...,0,1) then instead of choosing the one good restaurant, you will go with high probability to one of the bad ones.

Suggestion 2. Choose good with probability 0.95 and then at random among the good ones, and bad with probability 0.05, and then at random.

Once 'such' a model is chosen, it is a finite Markov chain, and one can compute a stationary distribution, etc.

Exersice: suppose your satisfaction value from each restaurant is a continuous random variable (say Uniform(0,1), but it does not matter). You always go to the restaurant that was best the last time you visited it, and replace its last value by the current one. How does this process behave in time?

1. What is the distribution of the smallest, second smallest, etc. value? Partial Answer: they are all decreasing to zero, except for the largest. In other words, the worst restaurant - as you value them- is getting worse, and so is the second, and so on up to the $k-1$st.

2. Suppose you look at your last value from a particular restaurant. Is it decreasing?

Statistics - choosing a model for given data:

Data $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n) \sim g(\boldsymbol{y}^{(n)})$, g is the true generating model. We assume it is unknown, and in general g will not be in our list of candidate models.

We have a list of candidate models $\{f_\alpha(\boldsymbol{y}^{(n)}|\boldsymbol{\theta})\}$, where $\boldsymbol{\theta} \in \Theta_\alpha$ with finite dimension d_α, $\alpha \in A$, a finite list of models.

Goal: to choose from our list the “best" model for the data.

$\boldsymbol{Y}^{(n)}$ could be n independent observations, or some n values from some stochastic process, or some other joint distribution.

Statistics - choosing a model for given data:

Data $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n) \sim g(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)})$, g is the true generating model. We assume it is unknown, and in general g will not be in our list of candidate models.

We have a list of candidate models $\{f_\alpha(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\theta})\}$, where $\boldsymbol{\theta} \in \Theta_\alpha$ with finite dimension d_α, $\alpha \in A$, a finite list of models.

Goal: to choose from our list the “best" model for the data.

$\boldsymbol{Y}^{(n)}$ could be n independent observations, or some n values from some stochastic process, or some other joint distribution.

Examples of models:

- $Y_i = r_\alpha(\boldsymbol{x}_i; \theta) + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$, and the **regression** function $r_\alpha(\boldsymbol{x}_i; \theta)$ can be a polynomial of degree α with a vector of coefficients θ, or α could indicate a subset of the covariates. Everything we do for iid Y_i's will apply if $(Y_i, \boldsymbol{X}_i)$ are iid.

- $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)$ is a realization of a k-step Markov chain, such as some autoregressive process.

Example (Seneta 2004) P_t price of asset

$$P_t = P_0 e^{ct + \theta T_t + \sigma W(T_t)}$$

where $T_t \geq 0$ increasing process, independent of the BM $W(t)$. Therefore

$$Y_t = \log P_t - \log P_{t-1} = c + \theta(T_t - T_{t-1}) + \sigma(W(T_t) - W(T_{t-1})).$$

• $T_t = t \quad \Rightarrow \quad$ Geometric BM.
• T_t has independent (or iid) increments $\quad \Rightarrow \quad Y_t$ are independent (or iid).
• $T_t - T_{t-1} \sim Exp(\eta)$ or Gamma(α, β) or ...

Which model “fits" the data?

Example (Seneta 2004) P_t price of asset

$$P_t = P_0 e^{ct+\theta T_t+\sigma W(T_t)}$$

where $T_t \geq 0$ increasing process, independent of the BM $W(t)$. Therefore

$$Y_t = \log P_t - \log P_{t-1} = c + \theta(T_t - T_{t-1}) + \sigma(W(T_t) - W(T_{t-1})).$$

- $T_t = t \quad \Rightarrow \quad$ Geometric BM.
- T_t has independent (or iid) increments $\quad \Rightarrow \quad$ Y_t are independent (or iid).
- $T_t - T_{t-1} \sim Exp(\eta)$ or Gamma(α, β) or ...

Which model "fits" the data?

Example (Seneta 2004) P_t price of asset

$$P_t = P_0 e^{ct+\theta T_t + \sigma W(T_t)}$$

where $T_t \geq 0$ increasing process, independent of the BM $W(t)$. Therefore

$$Y_t = \log P_t - \log P_{t-1} = c + \theta(T_t - T_{t-1}) + \sigma(W(T_t) - W(T_{t-1})).$$

- $T_t = t \quad \Rightarrow \quad$ Geometric BM.
- T_t has independent (or iid) increments $\quad \Rightarrow \quad Y_t$ are independent (or iid).
- $T_t - T_{t-1} \sim Exp(\eta)$ or Gamma(α, β) or ...

Which model “fits" the data?

More examples

$\boldsymbol{Y} = (Y_1, \ldots, Y_m) \sim \text{Multinomial}_m(n, \boldsymbol{p})$, where $\boldsymbol{p} = \boldsymbol{p}(\theta)$.

Hardy-Weinberg: $m = 3$, $p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$,

so $\boldsymbol{p} \in$ one-dimensional curve in the 3-simplex

$\{(p_1, p_2, p_3) \geq 0 \ : \ p_1 + p_2 + p_3\}$.

$\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)$ a sample (iid) from $\text{Exp}(\theta)$ or $\Gamma(\alpha, \theta)$, or a mixture $\sum_i w_i \Gamma(\alpha_i, \theta_i)$ or ...

More examples

$\boldsymbol{Y} = (Y_1, \ldots, Y_m) \sim \text{Multinomial}_m(n, \boldsymbol{p})$, where $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$.

Hardy-Weinberg: $m = 3$, $p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$,

so $\boldsymbol{p} \in$ one-dimensional curve in the 3-simplex

$\{(p_1, p_2, p_3) \geq 0 \; : \; p_1 + p_2 + p_3\}$.

$\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)$ a sample (iid) from $\text{Exp}(\theta)$ or $\Gamma(\alpha, \theta)$, or a mixture $\sum_i w_i \Gamma(\alpha_i, \theta_i)$ or ...

More examples

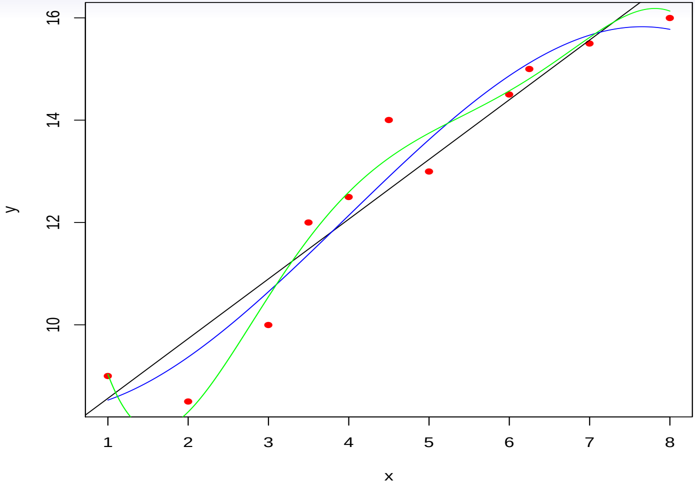
$\boldsymbol{Y} = (Y_1, \ldots, Y_m) \sim \text{Multinomial}_m(n, \boldsymbol{p})$, where $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$.

Hardy-Weinberg: $m = 3$, $p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$,

so $\boldsymbol{p} \in$ one-dimensional curve in the 3-simplex

$\{(p_1, p_2, p_3) \geq 0 \;:\; p_1 + p_2 + p_3\}$.

$\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)$ a sample (iid) from $\text{Exp}(\theta)$ or $\Gamma(\alpha, \theta)$, or a mixture $\sum_i w_i \Gamma(\alpha_i, \theta_i)$ or ...



Data: 11 data points (Red). Black line : linear regression,
Blue : 3rd degree poly, Green : 5th degree

Regression: well known model selection methods

Forward/backward selection, Mallows C_p, AIC, BIC, Lasso.

Penalty based methods , regularization

$$\min_{\beta}\{\|\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}\beta\|^2\},$$

Regression: well known model selection methods

Forward/backward selection, Mallows C_p, AIC, BIC, Lasso.

Penalty based methods , regularization

$$\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}\boldsymbol{\beta}\|^2\},$$

Regression: well known model selection methods

Forward/backward selection, Mallows C_p, AIC, BIC, Lasso.

Penalty based methods , regularization

$$\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}\boldsymbol{\beta}\|^2 + \lambda \, |||\boldsymbol{\beta}\,|||\},$$

where

$$|||\boldsymbol{\beta}\,||| = \begin{cases} \sum_j \beta_j^2 & \text{Ridge regession} \\ \sum_j |\beta_j| & \text{Lasso} \ \text{\scriptsize least absolute shrinkage and selection operator} \\ \#\text{non-zero } \beta_j' s & \text{AIC, BIC} \end{cases}$$

Regression: well known model selection methods

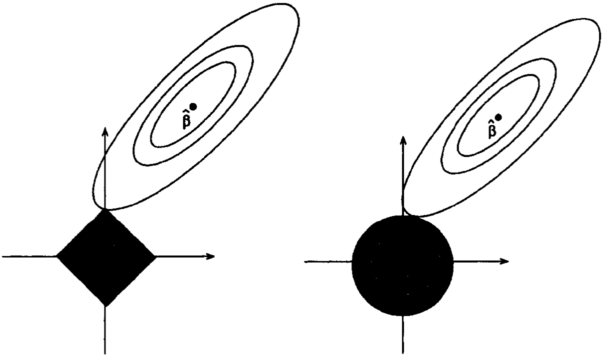
Forward/backward selection, Mallows C_p, AIC, BIC, Lasso.

Penalty based methods , regularization

$$\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}\beta\|^2 + \lambda\, |||\boldsymbol{\beta}\,|||\}, \quad \text{or} \quad \min_{\||\boldsymbol{\beta}\|\leq t} \|\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}\beta\|^2$$

where

$$|||\boldsymbol{\beta}\,||| = \begin{cases} \sum_j \beta_j^2 & \text{Ridge regession} \\ \sum_j |\beta_j| & \text{Lasso} \ \text{\small least absolute shrinkage and selection operator} \\ \#\text{non-zero } \beta_j' s & \text{AIC, BIC} \end{cases}$$



$\hat{\beta}$
$\hat{\beta}$

What do we mean by: Which model “fits" the data?

Regression : Suppose we fit a model through **Least Squares**:

$$(LS) \qquad \min_\theta \textstyle\sum_i (y_i - r_\alpha(\boldsymbol{x}_i;\theta))^2,$$

where $r_\alpha(\boldsymbol{x}_i;\theta)$ is a polynomial of degree d_α, say.

Assuming normality the minimizer of (LS) $\widehat{\boldsymbol{\theta}}_\alpha \in \Theta_\alpha$ is the Maximum Likelihood Estimator, and (LS) becomes
$\sum_i \log f_\alpha(y_i|x_i,\widehat{\boldsymbol{\theta}}_\alpha)$,
and the best fit of the data is obtained by

$$(LK) \qquad \max_\alpha \textstyle\sum_i \log f_\alpha(y_i|x_i,\widehat{\boldsymbol{\theta}}_\alpha).$$

If we now choose α to minimize $\sum_i (y_i - r_\alpha(\boldsymbol{x}_i; \widehat{\theta}_\alpha))^2$ then clearly the "largest" model will always be chosen. It will provide the best fit for the given data. But it will **overfit** the data.

A criterion for the predictive value of the model: Suppose that a hypothetical new sample of y_i^*'s $\sim g$ arising from the same $\mathbf{x}_i$ were available. We would like to choose the model attaining

$$(*) \qquad \min_\alpha \sum_i (y_i^* - r_\alpha(\boldsymbol{x}_i; \widehat{\theta}_\alpha))^2.$$

Assuming normality a model that minimizes $(*)$ over α maximizes the log-likelihood of the hypothetical data, and the criterion becomes

$$(**) \qquad \max_\alpha \sum_i \log f_\alpha(y_i^* | x_i, \widehat{\theta}_\alpha).$$

If we now choose α to minimize $\sum_i (y_i - r_\alpha(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_\alpha))^2$ then clearly the "largest" model will always be chosen. It will provide the best fit for the given data. But it will **overfit** the data.

A criterion for the predictive value of the model: Suppose that a hypothetical new sample of y_i^*'s $\sim g$ arising from the same $\mathbf{x}_i$ were available. We would like to choose the model attaining

$$(*) \qquad \min_\alpha \sum_i (y_i^* - r_\alpha(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_\alpha))^2.$$

Assuming normality a model that minimizes $(*)$ over α maximizes the log-likelihood of the hypothetical data, and the criterion becomes

$$(**) \qquad \max_\alpha \sum_i \log f_\alpha(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_\alpha).$$

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_{i} \log f_{\alpha}(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_{\alpha})$$

we need to estimate $\sum_i \log f_{\alpha}(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_{\alpha})$. But y_i^*'s are not available. set $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\alpha}$.

Specializing on the iid case write $\sum_{i} \log f_{\alpha}(y_i^* | \widehat{\theta})$

Y_i and $Y_i^* \sim g$, and therefore

it is tempting to estimate the above expression by

$$\frac{1}{n} \sum_{i} \log f_{\alpha}(y_i | \widehat{\theta})$$

This will clearly be an over-estimate.

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_i \log f_\alpha(y_i^*|x_i, \widehat{\theta}_\alpha)$$

we need to estimate $\sum_i \log f_\alpha(y_i^*|x_i, \widehat{\theta}_\alpha)$. But y_i^*'s are not available. set $\widehat{\theta} = \widehat{\theta}_\alpha$.

Specializing on the iid case write $\sum_i \log f_\alpha(y_i^*|x_i, \widehat{\theta})$

Y_i and $Y_i^* \sim g$, and therefore $\Uparrow$

it is tempting to estimate the above expression by

$$\frac{1}{n}\sum_i \log f_\alpha(y_i|\widehat{\theta})$$

This will clearly be an over-estimate.

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_i \log f_\alpha(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_\alpha)$$

we need to estimate $\sum_i \log f_\alpha(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_\alpha)$. But y_i^*'s are not available. set $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\alpha$.

Specializing on the iid case write $\displaystyle\sum_i \log f_\alpha(y_i^* | \widehat{\boldsymbol{\theta}})$

Y_i and $Y_i^* \sim g$, and therefore $\Uparrow$

it is tempting to estimate the above expression by

$$\frac{1}{n} \sum_i \log f_\alpha(y_i | \widehat{\boldsymbol{\theta}})$$

This will clearly be an over-estimate.

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_i \log f_\alpha(y_i^*|x_i, \widehat{\boldsymbol{\theta}}_\alpha)$$

we need to estimate $\sum_i \log f_\alpha(y_i^*|x_i, \widehat{\boldsymbol{\theta}}_\alpha)$. But y_i^*'s are not available. set $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\alpha$.

Specializing on the iid case write $\displaystyle \frac{1}{n}\sum_i \log f_\alpha(y_i^*|\widehat{\boldsymbol{\theta}})$

Y_i and $Y_i^* \sim g$, and therefore

it is tempting to estimate the above expression by

$$\frac{1}{n}\sum_i \log f_\alpha(y_i|\widehat{\boldsymbol{\theta}})$$

This will clearly be an over-estimate.

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_i \log f_\alpha(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_\alpha)$$

we need to estimate $\sum_i \log f_\alpha(y_i^* | x_i, \widehat{\boldsymbol{\theta}}_\alpha)$. But y_i^*'s are not available. set $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\alpha$.

Specializing on the iid case write $\displaystyle \frac{1}{n} \sum_i \log f_\alpha(y_i^* | \widehat{\boldsymbol{\theta}})$

Y_i and $Y_i^* \sim g$, and therefore $\Uparrow$

it is tempting to estimate the above expression by

$$\frac{1}{n} \sum_i \log f_\alpha(y_i | \widehat{\boldsymbol{\theta}})$$

This will clearly be an over-estimate.

In order to find the maximizer in

$$(**) \qquad \max_{\alpha} \sum_i \log f_\alpha(y_i^* | x_i, \widehat{\theta}_\alpha)$$

we need to estimate $\sum_i \log f_\alpha(y_i^* | x_i, \widehat{\theta}_\alpha)$. But y_i^*'s are not available. set $\widehat{\theta} = \widehat{\theta}_\alpha$.

Specializing on the iid case write $\frac{1}{n} \sum_i \log f_\alpha(y_i^* | \widehat{\theta})$

Y_i and $Y_i^* \sim g$, and therefore $\Uparrow$

it is tempting to estimate the above expression by

$$\frac{1}{n} \sum_i \log f_\alpha(y_i | \widehat{\theta})$$

This will clearly be an over-estimate.

$\frac{1}{n}\sum_i \log f_\alpha(y_i^*|\widehat{\theta})$ is a random variable. So we want to estimate its **expectation** under the true (unknown) model $Y_i^* \sim g$, and in the iid case we have

$$E_{y^*} \log f_\alpha(y_i^*|\widehat{\theta}) = \int \log f_\alpha(y^*|\widehat{\theta}) g(y^*) dy^*.$$

Maximizing the latter over α is 'equivalent' to minimizing the Kullback-Leibler divergence

$$D(g(y^*)\,||\,f_\alpha(y_i^*|\widehat{\theta})) = \int g(y^*) \log \frac{g(y^*)}{f_\alpha(y_i^*|\widehat{\theta})} dy^*$$

$\frac{1}{n}\sum_i \log f_{\alpha}(y_i^*|\widehat{\theta})$ is a random variable. So we want to estimate its **expectation** under the true (unknown) model $Y_i^* \sim g$, and in the iid case we have

$$E_{y^*} \log f_{\alpha}(y_i^*|\widehat{\theta}) = \int \log f_{\alpha}(y^*|\widehat{\theta}) g(y^*) dy^*.$$

Maximizing the latter over α is 'equivalent' to minimizing the Kullback-Leibler divergence

$$D(g(y^*) \,||\, f_{\alpha}(y_i^*|\widehat{\theta})) = \int g(y^*) \log \frac{g(y^*)}{f_{\alpha}(y_i^*|\widehat{\theta})} dy^*$$

Following various asymptotic expansions and “approximations",

the **AIC** (Akaike 1974) estimates (♣) by

$$\mathbf{AIC}_{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \log f_{\alpha}(y_i|\widehat{\theta}) - d_{\alpha}/n,$$

where d_{α} is the dimension of the parameter space of the αth

model.

A Bayesian approach

Schwarz BIC (1978): $\mathbf{Y} = Y_1, \ldots, Y_n$ sample (iid) from some

$f_k(y|\theta)$, $\theta \in \Theta_k$. Model k is true with prior probability w_k.

Prior: $\theta \sim \sum_k w_k \mu_k$, where μ_k are prior measures on Θ_k,

$\sum_k w_k = 1$.

$P(k|\mathbf{Y})$ = posterior probability of k.

Theorem: $\frac{1}{n}\log P(k|\mathbf{Y}) = \frac{1}{n}\sum_i \log f_k(y_i|\widehat{\theta}) - \mathbf{\log(n)}\frac{d_k}{2n} + \frac{R_n}{n} + D_n$,

where d_k is the dimension of the parameter space of the kth

model Θ_k, and R_n is a bounded r.v., D_n does not depend on k.

A Bayesian approach

Schwarz BIC (1978): $\mathbf{Y} = Y_1, \ldots, Y_n$ sample (iid) from some

$f_k(y|\theta)$, $\theta \in \Theta_k$. Model k is true with prior probability w_k.

Prior: $\theta \sim \sum_k w_k \mu_k$, where μ_k are prior measures on Θ_k,

$\sum_k w_k = 1$.

$P(k|\mathbf{Y})$ = posterior probability of k.

Theorem: $\frac{1}{n} \log P(k|\mathbf{Y}) = \frac{1}{n} \sum_i \log f_k(y_i|\widehat{\theta}) - \mathbf{log(n)} \frac{d_k}{2n} + \frac{R_n}{n} + D_n$,

where d_k is the dimension of the parameter space of the kth

model Θ_k, and R_n is a bounded r.v., D_n does not depend on k.

A Bayesian approach

Schwarz BIC (1978): $\mathbf{Y} = Y_1, \ldots, Y_n$ sample (iid) from some $f_k(y|\theta)$, $\theta \in \Theta_k$. Model k is true with prior probability w_k.

Prior: $\theta \sim \sum_k w_k \mu_k$, where μ_k are prior measures on Θ_k, $\sum_k w_k = 1$.

$P(k|\mathbf{Y})$ = posterior probability of k.

Theorem: $\frac{1}{n} \log P(k|\mathbf{Y}) = \frac{1}{n} \sum_i \log f_k(y_i|\widehat{\theta}) - \mathbf{log(n)} \frac{d_k}{2n} + \frac{R_n}{n} + D_n$,

where d_k is the dimension of the parameter space of the kth model Θ_k, and R_n is a bounded r.v., D_n does not depend on k.

A Bayesian approach

Schwarz BIC (1978): $\mathbf{Y} = Y_1, \ldots, Y_n$ sample (iid) from some

$f_k(y|\theta)$, $\theta \in \Theta_k$. Model k is true with prior probability w_k.

Prior: $\theta \sim \sum_k w_k \mu_k$, where μ_k are prior measures on Θ_k,

$\sum_k w_k = 1$.

$P(k|\mathbf{Y})$ = posterior probability of k.

Theorem: $\frac{1}{n} \log P(k|\mathbf{Y}) = \frac{1}{n} \sum_i \log f_k(y_i|\widehat{\theta}) - \mathbf{log(n)} \frac{d_k}{2n} + \frac{R_n}{n} + D_n$,

where d_k is the dimension of the parameter space of the kth

model Θ_k, and R_n is a bounded r.v., D_n does not depend on k.

Proof:

$$D_n P(k \,|\, Y) = \int_{\Theta_k} e^{\sum_{j=1}^n \log f(Y_j \,|\, k, \boldsymbol{\theta})} w_k \, \mu_k(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$= \int_{\Theta_k} e^{n \sum_{j=1}^n \log f(Y_j \,|\, k, \boldsymbol{\theta})/n} w_k \mu_k(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$\approx e^{\sum_{j=1}^n \log f(Y_j \,|\, k, \widehat{\boldsymbol{\theta}})} w_k \mu_k(\widehat{\boldsymbol{\theta}}) (2\pi)^{d_k/2} |\Sigma_k|^{-1/2} n^{-d_k/2}.$$

$\approx$ uses Laplace approximation method (saddle point),

where $\Sigma_k = -[\frac{\partial^2}{\partial\theta_i \partial\theta_j} n^{-1} \sum_{j=1}^n \log f(Y_j \,|\, k, \widehat{\boldsymbol{\theta}})]$.

Comments: BIC makes sense when we assume that the true model is in the list of candidate models, and has a positive prior probability.

Is this ever a realistic assumption? Are there "**true models**"? is it an oxymoron?

Comments: BIC makes sense when we assume that the true model is in the list of candidate models, and has a positive prior probability.

In this case, BIC is consistent, **unlike AIC** it will converge to the true model (consistency).

Is this ever a realistic assumption? Are there "true **models**"? is it an oxymoron?

Comments: BIC makes sense when we assume that the true model is in the list of candidate models, and has a positive prior probability.

Is this ever a realistic assumption? Are there "**true models**"? is it an oxymoron?

Comments: BIC makes sense when we assume that the true model is in the list of candidate models, and has a positive prior probability.

Is this ever a realistic assumption? Are there "**true models**"? is it an oxymoron?

The bias correction d_α was an approximation. Define

$$K = E\left\{[\frac{\partial}{\partial\theta_i}f(Y_i\,;\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_j}f(Y_i\,;\boldsymbol{\theta}_0)]/f^2(Y_i\,;\boldsymbol{\theta}_0)\right\}, \qquad (1)$$

where $Y_i \sim g$, and $f = f_\alpha$, and $\theta_0 = \arg\min_\theta D(g||f(\cdot|\theta))$.

$$\begin{aligned} J &= -E\left\{[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y_k\,;\boldsymbol{\theta}_0)]\right\} \\ &= -E\left\{[\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(Y_k\,;\boldsymbol{\theta}_0)]/f(Y_k\,;\boldsymbol{\theta}_0)\right\} + K, \qquad (2) \\ &\Uparrow \end{aligned}$$

which is obtained by straightforward differentiation under the expectation sign. When $g(y) = f(y\,;\boldsymbol{\theta}_0)$, the first term on the right-hand side of (2) vanishes, and we obtain the well known Fisher information identity $J = K$. This is the case in standard MLE theory.

K and J are $d_\alpha \times d_\alpha$ matrices.

The bias correction is $\mathrm{Trace}(J^{-1}K)$, which becomes $\mathrm{Trace}(I_{d_\alpha}) = d_\alpha$ in the case $K = J$ described above.

It is not easy to estimate $\mathrm{Trace}(J^{-1}K)$, and d_α is proposed as an approximation, which is good for models that are close to the true one, and exact for a model that contains the true one.

How do these K and J arise?

K and J are $d_\alpha \times d_\alpha$ matrices.

The bias correction is Trace($J^{-1}K$), which becomes Trace(I_{d_α}) $= d_\alpha$ in the case $K = J$ described above.

It is not easy to estimate Trace($J^{-1}K$), and d_α is proposed as an approximation, which is good for models that are close to the true one, and exact for a model that contains the true one.

How do these K and J arise?

We compute a Taylor expansion around the MLE $\hat{\theta}$, with the first derivative vanishing at the MLE, and obtain

$$\frac{1}{n}\sum_{k=1}^{n}\log f(Y_k\,;\hat{\boldsymbol{\theta}}) - \frac{1}{n}\sum_{k=1}^{n}\log f(Y_k\,;\boldsymbol{\theta}_0)$$
$$\approx -\frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T[\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta_i\partial\theta_k}\log f(Y_k\,;\hat{\boldsymbol{\theta}})](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})...$$

My plan (with David Azriel): given a data set of a large size *N*, of a certain type, assume that different researchers (will) have different data sets of different (smaller than *N*) sizes of this type. The model they should use depends on the size of their sample which they use to estimate the parameters for their own case. On the basis of the large data set, we want to determine the model to be used by a researcher with data of size *n*, that is, determine $\alpha = \alpha(n)$, where as before, α is the index of models, and to quantify the value of the chosen model.

Motivation: experimental game theory (economics).

My plan (with David Azriel): given a data set of a large size *N*, of a certain type, assume that different researchers (will) have different data sets of different (smaller than *N*) sizes of this type. The model they should use depends on the size of their sample which they use to estimate the parameters for their own case. On the basis of the large data set, we want to determine the model to be used by a researcher with data of size *n*, that is, determine $\alpha = \alpha(n)$, where as before, α is the index of models, and to quantify the value of the chosen model.

Motivation: experimental game theory (economics).